

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/256492530>

# Modeling and Simulation for Automatic Control

Book · January 2002

---

CITATIONS

284

READS

3,985

2 authors:



Olav Egeland

Norwegian University of Science and Technology

256 PUBLICATIONS 5,710 CITATIONS

[SEE PROFILE](#)



Jan Tommy Gravdahl

Norwegian University of Science and Technology

317 PUBLICATIONS 4,416 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Revealing nanomechanical properties of soft samples in atomic force microscopy [View project](#)



Geometric Algebra in Robotics and Computer Vision [View project](#)

# 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>

Copyright © 2002 by Marine Cybernetics AS.

All rights reserved.

**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

No parts of this publication may be reproduced by any means, transmitted, or translated into machine language without the written permission of the author. Requests for permission to reproduce parts of the book should be addressed directly to Professor Olav Egeland, Department of Engineering Cybernetics, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway; E-mail: Olav.Egeland@itk.ntnu.no, fax: [+47] 73594399.

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

<b>I Modeling</b>	<b>1</b>
<b>1 Model representation</b>	<b>3</b>
1.1 Introduction . . . . .	3
1.2 State space methods . . . . .	4
1.2.1 State space models . . . . .	4
1.2.2 Second order models of mechanical systems . . . . .	5
1.2.3 Linearization of state space models . . . . .	5
1.2.4 Linearization of second order systems . . . . .	7
1.2.5 Stability with zero input . . . . .	8
1.2.6 Stability of linear systems . . . . .	9
1.2.7 Stability analysis using a linearized model . . . . .	9
1.3 Transfer function models . . . . .	10
1.3.1 Introduction . . . . .	10
1.3.2 The transfer function of a state-space model . . . . .	10
1.3.3 Rational transfer functions . . . . .	11
1.3.4 Impulse response and step response . . . . .	12
1.3.5 Loop transfer function . . . . .	14
1.3.6 Example: Actuator with dynamic compensation . . . . .	15
1.3.7 Stability of transfer functions . . . . .	16
1.3.8 Stability of closed loop systems . . . . .	17
1.3.9 Partial differential equations . . . . .	17
1.4 Network description . . . . .	19
1.4.1 Introduction . . . . .	19
1.4.2 Background . . . . .	20
1.4.3 Multiport . . . . .	21
1.4.4 Example: DC motor with flexible load . . . . .	21
1.4.5 Example: Voltage controlled DC motor . . . . .	22
1.4.6 Example: Diesel engine with turbocharger . . . . .	23
1.4.7 Assigning computational inputs and outputs . . . . .	24
1.4.8 Bond graphs . . . . .	26
1.5 Linear network theory . . . . .	26
1.5.1 Driving point impedance . . . . .	26
1.5.2 Linear two-ports . . . . .	28
1.5.3 Impedance of two-port with termination . . . . .	28
1.5.4 Example: Passive mechanical two-port . . . . .	29
1.5.5 Mechanical analog of PD controller . . . . .	31
1.6 Example: Transmission line model . . . . .	33
1.6.1 Introduction . . . . .	33

1.6.2	Introductory example . . . . .	33
1.6.3	Effort and flow model . . . . .	34
1.6.4	Transfer functions . . . . .	35
1.6.5	Transfer function for terminated transmission line . . . . .	36
1.6.6	Wave variables . . . . .	37
1.6.7	Lossless transmission line . . . . .	38
1.6.8	Line termination . . . . .	38
<b>2</b>	<b>Model analysis tools</b>	<b>41</b>
2.1	Frequency response methods . . . . .	41
2.1.1	The frequency response of a system . . . . .	41
2.1.2	Second order oscillatory system . . . . .	42
2.1.3	Performance of a closed loop system . . . . .	43
2.1.4	Stability margins . . . . .	44
2.2	Elimination of fast dynamics . . . . .	45
2.2.1	Example: The electrical time constant in a DC motor . . . . .	45
2.2.2	Nonlinear system . . . . .	46
2.3	Energy-based methods . . . . .	46
2.3.1	Introduction . . . . .	46
2.3.2	The energy function . . . . .	47
2.3.3	Second-order systems . . . . .	47
2.3.4	Example: Mass-spring-damper . . . . .	48
2.3.5	Lyapunov methods . . . . .	49
2.3.6	Contraction . . . . .	50
2.3.7	Energy flow in a turbocharged diesel engine . . . . .	51
2.4	Passivity . . . . .	52
2.4.1	Introduction . . . . .	52
2.4.2	Definition . . . . .	53
2.4.3	Examples . . . . .	53
2.4.4	Energy considerations . . . . .	55
2.4.5	Positive real transfer functions . . . . .	56
2.4.6	Positive real rational transfer functions . . . . .	56
2.4.7	Positive realness of irrational transfer functions . . . . .	58
2.4.8	Passivity and positive real transfer functions . . . . .	59
2.4.9	No poles on the imaginary axis . . . . .	60
2.4.10	Single poles at the imaginary axis . . . . .	60
2.4.11	Bounded real transfer functions . . . . .	61
2.4.12	Passivity of PID controllers . . . . .	62
2.4.13	Closed loop stability of positive real systems . . . . .	62
2.4.14	Storage function formulation . . . . .	63
2.4.15	Interconnections of passive systems . . . . .	64
2.4.16	Storage function for PID controller . . . . .	65
2.4.17	Passive plant with PID controller . . . . .	65
2.4.18	Example: Control of mass-spring-damper system . . . . .	66
2.4.19	Example: Active vibration damping . . . . .	66
2.4.20	Passive electrical one-port . . . . .	67
2.4.21	Electrical analog of PID controller . . . . .	68
2.4.22	Passive electrical two-port . . . . .	69
2.4.23	Termination of electrical two-port . . . . .	69
2.4.24	Passive electrical n-ports . . . . .	70

2.4.25 Example: Telemanipulation . . . . .	70
2.4.26 Passivity and gain . . . . .	73
2.5 Uncertainty in modeling . . . . .	74
2.5.1 General state space models . . . . .	74
2.5.2 Exact kinematic models . . . . .	75
2.5.3 Balance equations . . . . .	75
2.5.4 Passivity . . . . .	76
<b>II Motors and actuators</b>	<b>77</b>
<b>3 Electromechanical systems</b>	<b>79</b>
3.1 Introduction . . . . .	79
3.2 Electrical motors . . . . .	79
3.2.1 Introduction . . . . .	79
3.2.2 Basic equations . . . . .	80
3.2.3 Gear model . . . . .	80
3.2.4 Motor and gear . . . . .	81
3.2.5 Transformation of rotation to translation . . . . .	82
3.2.6 Torque characteristics . . . . .	83
3.2.7 The four quadrants of the motor . . . . .	84
3.3 The DC motor with constant field . . . . .	85
3.3.1 Introduction . . . . .	85
3.3.2 Model . . . . .	85
3.3.3 Energy function . . . . .	86
3.3.4 Laplace transformed model . . . . .	87
3.4 DC motor control . . . . .	88
3.4.1 Introduction . . . . .	88
3.4.2 Current controlled DC motor . . . . .	89
3.4.3 Velocity controlled DC motor . . . . .	90
3.4.4 Position controlled DC motor . . . . .	90
3.5 Motor and load with elastic transmission . . . . .	91
3.5.1 Introduction . . . . .	91
3.5.2 Equations of motion . . . . .	91
3.5.3 Transfer functions . . . . .	92
3.5.4 Zeros of the transfer function . . . . .	95
3.5.5 Energy analysis . . . . .	95
3.5.6 Motor with several resonances in the load . . . . .	95
3.5.7 Two motors driving an elastic load . . . . .	96
3.5.8 Energy analysis of two motors and load . . . . .	97
3.6 Motor and load with deadzone in the gear . . . . .	97
3.6.1 Introduction . . . . .	97
3.6.2 Elastic gear with deadzone . . . . .	98
3.6.3 Rigid gear with deadzone . . . . .	98
3.6.4 Two motors with deadzone and load . . . . .	99
3.7 Electromechanical energy conversion . . . . .	100
3.7.1 Introduction . . . . .	100
3.7.2 Inductive circuit elements . . . . .	101
3.7.3 Capacitive circuit elements . . . . .	102
3.7.4 Magnetic energy of a linear inductive element . . . . .	103

3.7.5	Stored energy of a linear capacitive element . . . . .	103
3.7.6	Energy and coenergy . . . . .	103
3.7.7	Electromechanical two-port with inductive element . . . . .	105
3.7.8	Electromechanical two-port with linear flux linkage . . . . .	107
3.7.9	Magnetic levitation . . . . .	107
3.7.10	Voice coil . . . . .	109
3.7.11	Electromagnetic three-port . . . . .	110
3.7.12	Electromechanical capacitive element . . . . .	111
3.7.13	Electromechanical two-port with linear charge . . . . .	112
3.7.14	Example: Capacitive microphone . . . . .	113
3.7.15	Piezoelectric actuator . . . . .	114
3.7.16	Actuator configuration . . . . .	115
3.8	DC motor with externally controlled field . . . . .	116
3.8.1	Model . . . . .	116
3.8.2	Network description . . . . .	118
3.8.3	DC motor with field weakening . . . . .	119
3.9	Dynamic model of the general AC motor . . . . .	120
3.9.1	Introduction . . . . .	120
3.9.2	Notation . . . . .	120
3.9.3	Dynamic model . . . . .	121
3.10	Induction motors . . . . .	126
3.10.1	Basic dynamic model . . . . .	126
3.10.2	Induction motor model in stator frame . . . . .	126
3.10.3	Dynamic model in the flux frame . . . . .	128
3.11	Lagrangian description of electromechanical systems . . . . .	131
3.11.1	Generalized coordinates . . . . .	131
3.11.2	Energy and coenergy . . . . .	131
3.11.3	Analogy of electrical and mechanical systems . . . . .	132
3.11.4	The Lagrangian . . . . .	133
3.11.5	Electromechanical systems . . . . .	134
3.11.6	Lagrange formulation of general AC motor . . . . .	136
3.11.7	Lagrange formulation of induction motor . . . . .	138
3.11.8	Lagrange formulation of DC motor . . . . .	138
<b>4</b>	<b>Hydraulic motors</b>	<b>141</b>
4.1	Introduction . . . . .	141
4.2	Valves . . . . .	141
4.2.1	Introduction . . . . .	141
4.2.2	Flow through a restriction . . . . .	141
4.2.3	Regularization of turbulent orifice flow . . . . .	142
4.2.4	Four-way valve . . . . .	144
4.2.5	Matched and symmetrical four-way valve . . . . .	145
4.2.6	Symmetric motor and valve with critical spool . . . . .	145
4.2.7	Symmetric motor and valve with open spool . . . . .	148
4.2.8	Flow control using pressure compensated valves . . . . .	148
4.2.9	Balance valve . . . . .	150
4.3	Motor models . . . . .	151
4.3.1	Mass balance . . . . .	151
4.3.2	Rotational motors . . . . .	152
4.3.3	Elastic modes in the load . . . . .	154

4.3.4	Hydraulic cylinder . . . . .	155
4.4	Models for transfer function analysis . . . . .	156
4.4.1	Matched and symmetric valve and symmetric motor . . . . .	156
4.4.2	Valve controlled motor: Transfer function . . . . .	157
4.4.3	Hydraulic motor with P controller . . . . .	159
4.4.4	Symmetric cylinder with matched and symmetric valve . . . . .	161
4.4.5	Pump controlled hydraulic drive with P controller . . . . .	162
4.4.6	Transfer functions for elastic modes . . . . .	162
4.4.7	Mechanical analog . . . . .	164
4.5	Hydraulic transmission lines . . . . .	165
4.5.1	Introduction . . . . .	165
4.5.2	PDE Model . . . . .	166
4.5.3	Laplace transformed model . . . . .	167
4.5.4	Lossless model . . . . .	168
4.5.5	Linear friction . . . . .	168
4.5.6	Nonlinear friction . . . . .	169
4.5.7	Wave variables . . . . .	169
4.5.8	Example: Lossless pipe . . . . .	171
4.5.9	Linear network models of transmission lines . . . . .	172
4.5.10	Rational approximations of transfer function models . . . . .	172
4.5.11	Rational series expansion of impedance model . . . . .	173
4.5.12	Rational series expansion of admittance model . . . . .	174
4.5.13	Galerkin derivation of impedance model . . . . .	175
4.5.14	Galerkin derivation of the admittance model . . . . .	176
4.5.15	Galerkin derivation of the hybrid model . . . . .	177
4.5.16	Rational simulation models . . . . .	178
4.6	Lumped parameter model of hydraulic line . . . . .	180
4.6.1	Introduction . . . . .	180
4.6.2	Helmholtz resonator model . . . . .	181
4.6.3	Model formulation . . . . .	181
4.6.4	Admittance model . . . . .	182
4.6.5	Impedance model . . . . .	183
4.6.6	Hybrid model . . . . .	183
4.6.7	Natural frequencies . . . . .	184
4.7	Object oriented simulation models . . . . .	185
4.7.1	Introduction . . . . .	185
4.7.2	Pump controlled hydraulic motor . . . . .	185
4.7.3	Cylinder with balance valve . . . . .	188
<b>5</b>	<b>Friction</b>	<b>191</b>
5.1	Introduction . . . . .	191
5.1.1	Background . . . . .	191
5.2	Static friction models . . . . .	192
5.2.1	Models for the individual phenomena . . . . .	192
5.2.2	Combination of individual models . . . . .	195
5.2.3	Problems with the static models . . . . .	196
5.2.4	Problems with signum terms at zero velocity . . . . .	197
5.2.5	Karnopp's model of Coulomb friction . . . . .	198
5.2.6	More on Karnopp's friction model . . . . .	198
5.2.7	Passivity of static models . . . . .	199

5.3	Dynamic friction models . . . . .	200
5.3.1	Introduction . . . . .	200
5.3.2	The Dahl model . . . . .	200
5.3.3	Passivity of the Dahl model . . . . .	202
5.3.4	The Bristle and LuGre model . . . . .	202
5.3.5	Passivity of the LuGre model . . . . .	204
5.3.6	The Elasto-Plastic model . . . . .	205
5.3.7	Passivity of the Elasto-Plastic model . . . . .	206
<b>III</b>	<b>Dynamics</b>	<b>207</b>
<b>6</b>	<b>Rigid body kinematics</b>	<b>209</b>
6.1	Introduction . . . . .	209
6.2	Vectors . . . . .	209
6.2.1	Vector description . . . . .	209
6.2.2	The scalar product . . . . .	210
6.2.3	The vector cross product . . . . .	211
6.3	Dyadics . . . . .	213
6.3.1	Introduction . . . . .	213
6.3.2	Introductory example: The inertia dyadic . . . . .	214
6.3.3	Matrix representation of dyadics . . . . .	215
6.4	The rotation matrix . . . . .	218
6.4.1	Coordinate transformations for vectors . . . . .	218
6.4.2	Properties of the rotation matrix . . . . .	219
6.4.3	Composite rotations . . . . .	220
6.4.4	Simple rotations . . . . .	221
6.4.5	Coordinate transformations for dyadics . . . . .	222
6.4.6	Homogeneous transformation matrices . . . . .	223
6.5	Euler angles . . . . .	224
6.5.1	Introduction . . . . .	224
6.5.2	Roll-pitch-yaw . . . . .	225
6.5.3	Classical Euler angles . . . . .	226
6.6	Angle-axis description of rotation . . . . .	226
6.6.1	Introduction . . . . .	226
6.6.2	Angle-axis parameters . . . . .	227
6.6.3	Derivation of rotation dyadic . . . . .	227
6.6.4	The rotation dyadic . . . . .	228
6.6.5	Rotation matrix . . . . .	229
6.7	Euler parameters . . . . .	231
6.7.1	Definition . . . . .	231
6.7.2	Quaternions . . . . .	232
6.7.3	Unit quaternions . . . . .	233
6.7.4	The quaternion product for unit quaternions . . . . .	234
6.7.5	Rotation by the quaternion product . . . . .	235
6.7.6	Euler parameters from the rotation matrix . . . . .	236
6.7.7	The Euler rotation vector . . . . .	237
6.7.8	Euler-Rodrigues parameters . . . . .	238
6.8	Angular velocity . . . . .	239
6.8.1	Introduction . . . . .	239

6.8.2	Definition . . . . .	240
6.8.3	Simple rotations . . . . .	240
6.8.4	Composite rotations . . . . .	241
6.8.5	Differentiation of coordinate vectors . . . . .	242
6.8.6	Differentiation of vectors . . . . .	242
6.9	Kinematic differential equations . . . . .	244
6.9.1	Introduction . . . . .	244
6.9.2	Attitude deviation . . . . .	244
6.9.3	Homogeneous transformation matrices . . . . .	245
6.9.4	Euler angles . . . . .	246
6.9.5	Euler parameters . . . . .	247
6.9.6	Normalization for numerical integration . . . . .	249
6.9.7	Euler rotation . . . . .	250
6.9.8	Euler-Rodrigues parameters . . . . .	250
6.9.9	Passivity of kinematic differential equations . . . . .	251
6.9.10	Angle-axis representation . . . . .	252
6.10	The Serret-Frenet frame . . . . .	253
6.10.1	Kinematics . . . . .	253
6.10.2	Control deviation . . . . .	255
6.11	Navigational kinematics . . . . .	255
6.11.1	Introduction . . . . .	255
6.11.2	Coordinate frames . . . . .	256
6.11.3	Acceleration . . . . .	258
6.12	Kinematics of a rigid body . . . . .	259
6.12.1	Configuration . . . . .	259
6.12.2	Velocity . . . . .	259
6.12.3	Acceleration . . . . .	259
6.13	The center of mass . . . . .	261
6.13.1	System of particles . . . . .	261
6.13.2	Rigid body . . . . .	261
<b>7</b>	<b>Newton-Euler equations of motion</b>	<b>263</b>
7.1	Introduction . . . . .	263
7.2	Forces and torques . . . . .	263
7.2.1	Resultant force . . . . .	263
7.2.2	Torque . . . . .	265
7.2.3	Equivalent force and torque . . . . .	265
7.2.4	Forces and torques on a rigid body . . . . .	266
7.2.5	Example: Robotic link . . . . .	268
7.3	Newton-Euler equations for rigid bodies . . . . .	268
7.3.1	Equations of motion for a system of particles . . . . .	268
7.3.2	Equations of motion for a rigid body . . . . .	269
7.3.3	Equations of motion about a point . . . . .	271
7.3.4	The inertia dyadic . . . . .	272
7.3.5	The inertia matrix . . . . .	274
7.3.6	Expressions for the inertia matrix . . . . .	275
7.3.7	The parallel axes theorem . . . . .	275
7.3.8	The equations of motion for a rigid body . . . . .	276
7.3.9	Satellite attitude dynamics . . . . .	277
7.4	Example: Ball and beam dynamics . . . . .	278

7.5	Example: Inverted pendulum . . . . .	281
7.5.1	Equations of motion . . . . .	281
7.5.2	Double inverted pendulum . . . . .	284
7.6	Example: The Furuta pendulum . . . . .	285
7.7	Principle of virtual work . . . . .	288
7.7.1	Introduction . . . . .	288
7.7.2	Generalized coordinates . . . . .	289
7.7.3	Virtual displacements . . . . .	290
7.7.4	d'Alembert's principle . . . . .	290
7.8	Principle of virtual work for a rigid body . . . . .	293
7.8.1	Virtual displacements for a rigid body . . . . .	293
7.8.2	Force and torque of constraint . . . . .	294
7.9	Multi-body dynamics and virtual work . . . . .	295
7.9.1	Introduction . . . . .	295
7.9.2	Equations of motion . . . . .	295
7.9.3	Equations of motion about a point . . . . .	297
7.9.4	Ball and beam . . . . .	298
7.9.5	Single and double inverted pendulum . . . . .	300
7.9.6	Furuta pendulum . . . . .	302
7.9.7	Planar two-link manipulator: Derivation 1 . . . . .	302
7.9.8	Planar two-link manipulator: Derivation 2 . . . . .	304
7.9.9	Kane's computational scheme for two-link manipulator . . . . .	306
7.9.10	Manipulator dynamics in coordinate form . . . . .	308
7.9.11	Spacecraft and manipulator . . . . .	309
7.10	Recursive Newton-Euler . . . . .	310
7.10.1	Inverse dynamics . . . . .	310
7.10.2	Simulation . . . . .	311
<b>8</b>	<b>Analytical mechanics</b>	<b>313</b>
8.1	Introduction . . . . .	313
8.2	Lagrangian dynamics . . . . .	313
8.2.1	Introduction . . . . .	313
8.2.2	Lagrange's equation of motion . . . . .	314
8.2.3	Generalized coordinates and generalized forces . . . . .	316
8.2.4	Pendulum . . . . .	316
8.2.5	Mass-spring system . . . . .	317
8.2.6	Ball and beam . . . . .	317
8.2.7	Furuta pendulum . . . . .	318
8.2.8	Manipulator . . . . .	319
8.2.9	Passivity of the manipulator dynamics . . . . .	321
8.2.10	Example: Planar two-link manipulator 1 . . . . .	322
8.2.11	Example: Planar two-link manipulator 2 . . . . .	323
8.2.12	Limitations of Lagrange's equation of motion . . . . .	324
8.3	Calculus of variations . . . . .	324
8.3.1	Introduction . . . . .	324
8.3.2	Variations versus differentials . . . . .	325
8.3.3	The variation of a function . . . . .	325
8.3.4	The Euler-Lagrange equation for a general integral . . . . .	327
8.3.5	The variation of the rotation matrix . . . . .	328
8.3.6	The variation of the homogeneous transformation matrix . . . . .	330

8.4	The adjoint formulation . . . . .	331
8.4.1	Introduction . . . . .	331
8.4.2	Rotations . . . . .	332
8.4.3	Rigid motion . . . . .	333
8.5	The Euler-Poincaré equation . . . . .	334
8.5.1	A central equation . . . . .	334
8.5.2	Rotating rigid body . . . . .	336
8.5.3	Free-floating rigid body . . . . .	337
8.5.4	Mechanism with $n$ degrees of freedom . . . . .	339
8.6	Hamilton's principle . . . . .	340
8.6.1	Introduction . . . . .	340
8.6.2	The extended Hamilton principle . . . . .	340
8.6.3	Derivation of Lagrange's equation of motion . . . . .	341
8.6.4	Hamilton's principle . . . . .	342
8.6.5	Rotations with the Euler-Poincaré equation . . . . .	343
8.6.6	Rigid motion with the Euler-Poincaré equation . . . . .	343
8.7	Lagrangian dynamics for PDE's . . . . .	344
8.7.1	Flexible beam dynamics . . . . .	344
8.7.2	Euler-Bernoulli beam . . . . .	346
8.7.3	Lateral vibrations in a string . . . . .	346
8.8	Hamilton's equations of motion . . . . .	347
8.8.1	Introduction . . . . .	347
8.8.2	Hamilton's equation of motion . . . . .	347
8.8.3	The energy function . . . . .	349
8.8.4	Change of generalized coordinates . . . . .	350
8.9	Control aspects . . . . .	351
8.9.1	Passivity of Hamilton's equation of motion . . . . .	351
8.9.2	Example: Manipulator dynamics . . . . .	352
8.9.3	Example: The restricted three-body problem . . . . .	353
8.9.4	Example: Attitude dynamics for a satellite . . . . .	355
8.9.5	Example: Gravity gradient stabilization . . . . .	356
8.10	The Hamilton-Jacobi equation . . . . .	358
<b>9</b>	<b>Mechanical vibrations</b>	<b>361</b>
9.1	Introduction . . . . .	361
9.2	Lumped elastic two-ports . . . . .	362
9.2.1	Hybrid two-port . . . . .	362
9.2.2	Displacement two-port . . . . .	362
9.2.3	Three masses in the hybrid formulation . . . . .	363
9.2.4	Three masses in the displacement formulation . . . . .	364
9.2.5	Four masses . . . . .	365
9.3	Vibrating string . . . . .	365
9.3.1	Linearized model . . . . .	365
9.3.2	Orthogonal shape functions . . . . .	366
9.3.3	Galerkin's method for orthogonal shape functions . . . . .	367
9.3.4	Finite element shape functions . . . . .	368
9.3.5	String element . . . . .	369
9.3.6	Assembling string elements . . . . .	370
9.4	Nonlinear string dynamics . . . . .	371
9.4.1	Kirchhoff's nonlinear string model . . . . .	371

9.4.2	Marine cables . . . . .	371
9.5	Euler Bernoulli beam . . . . .	373
9.5.1	Model . . . . .	373
9.5.2	Boundary conditions . . . . .	375
9.5.3	Energy . . . . .	376
9.5.4	Orthogonal shape functions . . . . .	376
9.5.5	Clamped-free Euler Bernoulli beam . . . . .	378
9.5.6	Beam fixed to an inertia and a mass . . . . .	380
9.5.7	Orthogonality of the eigenfunctions . . . . .	381
9.5.8	Galerkin's method for orthogonal mode shapes . . . . .	382
9.6	Finite element model of Euler Bernoulli beam . . . . .	385
9.6.1	Introduction . . . . .	385
9.6.2	Beam element . . . . .	385
9.6.3	Assembling a structure . . . . .	386
9.6.4	Finite element model and Galerkin's method . . . . .	388
9.7	Motor and Euler Bernoulli beam . . . . .	390
9.7.1	Equations of motion . . . . .	390
9.7.2	Assumed mode shapes . . . . .	391
9.7.3	Finite elements . . . . .	392
9.8	Irrational transfer functions for beam dynamics . . . . .	393
9.8.1	Introduction . . . . .	393
9.8.2	Clamped-free beam . . . . .	394
9.8.3	Motor and beam . . . . .	395
<b>IV</b>	<b>Balance equations</b>	<b>399</b>
<b>10</b>	<b>Kinematics of Flow</b>	<b>401</b>
10.1	Introduction . . . . .	401
10.2	Kinematics . . . . .	401
10.2.1	The material derivative . . . . .	401
10.2.2	The nabla operator . . . . .	402
10.2.3	Divergence . . . . .	403
10.2.4	Curl . . . . .	404
10.2.5	Material coordinates . . . . .	406
10.2.6	The dilation . . . . .	406
10.3	Orthogonal curvilinear coordinates . . . . .	408
10.3.1	General results . . . . .	408
10.3.2	Cylindrical coordinates . . . . .	411
10.4	Reynolds' transport theorem . . . . .	413
10.4.1	Introduction . . . . .	413
10.4.2	Basic transport theorem . . . . .	413
10.4.3	The transport theorem for a material volume . . . . .	414
10.4.4	The transport theorem and balance laws . . . . .	414
<b>11</b>	<b>Mass, momentum and energy balances</b>	<b>417</b>
11.1	The mass balance . . . . .	417
11.1.1	Differential form . . . . .	417
11.1.2	Integral form . . . . .	418
11.1.3	Control volume with compressible fluid . . . . .	418

11.1.4 Mass flow through a pipe . . . . .	419
11.1.5 Continuity equation and Reynolds' transport theorem . . . . .	420
11.1.6 Multi-component systems . . . . .	422
<b>11.2 The momentum balance . . . . .</b>	<b>423</b>
11.2.1 Euler's equation of motion . . . . .	423
11.2.2 The momentum equation for a control volume . . . . .	425
11.2.3 Example: Waterjet . . . . .	426
11.2.4 Example: Sand dispenser and conveyor . . . . .	426
11.2.5 Irrotational Bernoulli equation . . . . .	427
11.2.6 Bernoulli's equation along a streamline . . . . .	428
11.2.7 Example: Transmission line . . . . .	429
11.2.8 Liquid mass flow through a restriction . . . . .	431
11.2.9 Example: Water turbine . . . . .	432
11.2.10 Example: Waterhammer . . . . .	436
<b>11.3 Angular momentum balance . . . . .</b>	<b>437</b>
11.3.1 General expression . . . . .	437
11.3.2 Centrifugal pump with radial blades . . . . .	437
11.3.3 Euler's turbomachinery equation . . . . .	439
11.3.4 Pump instability . . . . .	439
<b>11.4 The energy balance . . . . .</b>	<b>441</b>
11.4.1 Material volume . . . . .	441
11.4.2 Fixed volume . . . . .	442
11.4.3 General control volume . . . . .	445
11.4.4 The heat equation . . . . .	446
11.4.5 Transfer function for the heat equation . . . . .	447
<b>11.5 Viscous flow . . . . .</b>	<b>448</b>
11.5.1 Introduction . . . . .	448
11.5.2 Tensor notation . . . . .	448
11.5.3 The velocity gradient tensor . . . . .	451
11.5.4 Example: The velocity gradient for a rigid body . . . . .	452
11.5.5 The stress tensor . . . . .	453
11.5.6 Cauchy's equation of motion . . . . .	454
11.5.7 Newtonian fluids . . . . .	456
11.5.8 The Navier-Stokes equation . . . . .	458
11.5.9 The Reynolds number . . . . .	459
11.5.10 The equation of kinetic energy . . . . .	460
11.5.11 The energy balance for a viscous fluid . . . . .	462
11.5.12 Fixed volume . . . . .	463
11.5.13 General control volume . . . . .	463
<b>12 Gas dynamics</b>	<b>465</b>
12.1 Introduction . . . . .	465
12.2 Energy, enthalpy and entropy . . . . .	465
12.2.1 Energy . . . . .	465
12.2.2 Enthalpy . . . . .	465
12.2.3 Specific heats . . . . .	466
12.2.4 Entropy . . . . .	467
12.2.5 The entropy equation . . . . .	467
12.2.6 Internal energy equation in terms of temperature . . . . .	470
12.2.7 Energy balance in terms of pressure . . . . .	471

12.2.8	Piston motion . . . . .	472
12.3	Isentropic conditions . . . . .	472
12.3.1	Isentropic processes . . . . .	472
12.3.2	Stagnation state . . . . .	474
12.3.3	Energy balance for isentropic processes . . . . .	474
12.3.4	The speed of sound . . . . .	475
12.3.5	Helmholtz resonator . . . . .	476
12.4	Acoustic resonances in pipes . . . . .	477
12.4.1	Dynamic model . . . . .	477
12.4.2	Pipe closed at both ends . . . . .	478
12.4.3	Pipe closed at one end . . . . .	479
12.4.4	Pressure measurement in diesel engine cylinder . . . . .	480
12.5	Gas flow . . . . .	480
12.5.1	Gas flow through a restriction . . . . .	480
12.5.2	Example: Discharge of gas from tank . . . . .	482
12.5.3	The Euler equation around sonic speed . . . . .	483
<b>13</b>	<b>Compressor dynamics</b>	<b>485</b>
13.1	Introduction . . . . .	485
13.1.1	Compressors . . . . .	485
13.1.2	Surge and rotating stall . . . . .	486
13.2	Centrifugal Compressors . . . . .	486
13.2.1	Introduction . . . . .	486
13.2.2	Shaft dynamics . . . . .	487
13.2.3	Compressor system . . . . .	489
13.2.4	Mass balance . . . . .	489
13.2.5	Momentum equation . . . . .	490
13.3	Compressor characteristic . . . . .	491
13.3.1	Derivation . . . . .	491
13.3.2	The compressor characteristic at zero mass flow . . . . .	494
13.4	Compressor surge . . . . .	497
13.4.1	The Greitzer surge model . . . . .	497
13.4.2	Linearization . . . . .	499
13.4.3	Passivity of the Greitzer surge model . . . . .	500
13.4.4	Curvefitting of compressor characteristic . . . . .	501
13.4.5	Compression systems with recycle . . . . .	504
<b>V</b>	<b>Simulation</b>	<b>507</b>
<b>14</b>	<b>Simulation</b>	<b>509</b>
14.1	Introduction . . . . .	509
14.1.1	The use of simulation in automatic control . . . . .	509
14.1.2	The Moore Greitzer model . . . . .	510
14.1.3	The restricted three-body problem . . . . .	513
14.1.4	Mass balance of chemical reactor . . . . .	516
14.2	Preliminaries . . . . .	517
14.2.1	Notation . . . . .	517
14.2.2	Computation error . . . . .	517
14.2.3	The order of a one-step method . . . . .	517

14.2.4 Linearization . . . . .	518
14.2.5 The linear test function . . . . .	520
14.3 Euler methods . . . . .	521
14.3.1 Euler's method . . . . .	521
14.3.2 The improved Euler method . . . . .	523
14.3.3 The modified Euler method . . . . .	525
14.4 Explicit Runge-Kutta methods . . . . .	526
14.4.1 Introduction . . . . .	526
14.4.2 Numerical scheme . . . . .	526
14.4.3 Order conditions . . . . .	527
14.4.4 Some explicit Runge-Kutta methods . . . . .	528
14.4.5 Case study: Pneumatic spring . . . . .	528
14.4.6 Stability function . . . . .	531
14.4.7 FSAL methods . . . . .	534
14.5 Implicit Runge-Kutta methods . . . . .	534
14.5.1 Stiff systems . . . . .	534
14.5.2 Implicit Runge-Kutta methods . . . . .	535
14.5.3 Implicit Euler method . . . . .	535
14.5.4 Trapezoidal rule . . . . .	536
14.5.5 Implicit midpoint rule . . . . .	537
14.5.6 The theta method . . . . .	538
14.5.7 Stability function . . . . .	538
14.5.8 Some implicit Runge-Kutta methods . . . . .	539
14.5.9 Case study: Pneumatic spring revisited . . . . .	540
14.6 Stability of Runge-Kutta methods . . . . .	544
14.6.1 Aliasing . . . . .	544
14.6.2 A-stability, L-stability . . . . .	544
14.6.3 Stiffly accurate methods . . . . .	546
14.6.4 Padé approximations . . . . .	548
14.6.5 Stability for Padé approximations . . . . .	550
14.6.6 Example: Mechanical vibrations . . . . .	551
14.6.7 Frequency response . . . . .	551
14.6.8 AN-stability . . . . .	555
14.6.9 B-stability . . . . .	557
14.6.10 Algebraic stability . . . . .	558
14.6.11 Properties of Runge-Kutta methods . . . . .	560
14.7 Automatic adjustment of step size . . . . .	560
14.7.1 Estimation of the local error for Runge-Kutta methods . . . . .	560
14.7.2 Adjustment algorithm . . . . .	563
14.8 Implementation aspects . . . . .	563
14.8.1 Solution of implicit equations . . . . .	563
14.8.2 Dense outputs . . . . .	565
14.8.3 Event detection . . . . .	566
14.8.4 Systems with inertia matrix . . . . .	566
14.9 Invariants . . . . .	567
14.9.1 Introduction . . . . .	567
14.9.2 Linear invariants . . . . .	567
14.9.3 Quadratic functions . . . . .	568
14.9.4 Quadratic invariants . . . . .	569

14.9.5 Symplectic Runge-Kutta methods . . . . .	571
14.10 Rosenbrock methods . . . . .	574
14.11 Multistep methods . . . . .	576
14.11.1 Explicit Adams methods . . . . .	576
14.11.2 Implicit Adams methods . . . . .	578
14.11.3 Predictor-Corrector implementation . . . . .	579
14.11.4 Backwards differentiation methods . . . . .	580
14.11.5 Linear stability analysis . . . . .	581
14.11.6 Stability of Adams methods . . . . .	582
14.11.7 Stability of BDF methods . . . . .	583
14.11.8 Frequency response . . . . .	583
14.11.9 Adams methods . . . . .	585
14.11.10 BDF methods . . . . .	585
14.12 Differential-algebraic equations . . . . .	585
14.12.1 Implicit Runge-Kutta methods for index 1 problems . . . . .	587
14.12.2 Multistep methods for index 1 problems . . . . .	589
<b>15 Computational fluid dynamics</b>	<b>591</b>
15.1 Introduction . . . . .	591
15.2 Governing equations . . . . .	591
15.3 Classification . . . . .	592
15.3.1 Hyperbolic equations . . . . .	595
15.3.2 Parabolic equations . . . . .	596
15.3.3 Elliptic equations . . . . .	597
15.4 Diffusion . . . . .	599
15.4.1 Introduction . . . . .	599
15.4.2 Finite volume method for stationary diffusion . . . . .	599
15.5 Solution of equations . . . . .	603
15.5.1 Worked example on stationary diffusion . . . . .	603
15.6 Stability issues . . . . .	604
15.7 Finite volume method for diffusion dynamics . . . . .	605
15.8 Finite volumes for Convection-Diffusion . . . . .	610
15.8.1 Introduction . . . . .	610
15.8.2 Finite volume method for 1D diffusion and convection dynamics . . . . .	611
15.9 Pressure-velocity coupling . . . . .	614
15.9.1 Introduction . . . . .	614
15.9.2 The staggered grid . . . . .	615
15.9.3 The momentum equations . . . . .	617
15.9.4 The transient SIMPLE algorithm . . . . .	618
15.10 Von Neuman stability method . . . . .	620

# **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*

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

$$\vdots \quad (1.2)$$

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

which in vector form is written

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

is a set of first order differential equations describing the dynamics of the *state vector*  $\mathbf{x} = (x_1, \dots, x_n)^T$  under the action of the *control* or *input vector*  $\mathbf{u} = (u_1, \dots, u_p)^T$ . The *measurement* or *output vector*  $\mathbf{y}$  is often included in the model formulation, and the state-space model is written

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

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

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

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{Ax} + \mathbf{Bu} \\ \mathbf{y} &= \mathbf{Cx} + \mathbf{Du} \end{aligned} \quad (1.7)$$

A block diagram is shown in Figure 1.1.

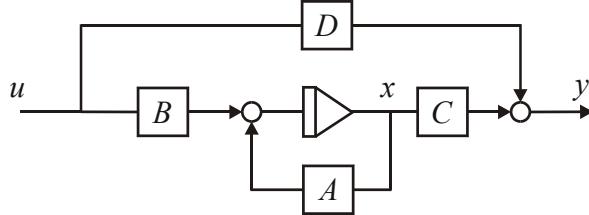


Figure 1.1: Linear time-invariant state space model

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

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

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

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

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

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

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)

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

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,  $\mathbf{q} = (q_1, \dots, q_6)$  is the vector of generalized coordinates, and  $\boldsymbol{\tau} = (\tau_1, \dots, \tau_6)$  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

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

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

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

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

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

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

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

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

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

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

where the matrices appearing from the differentiations have elements given by

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

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

$$\Delta \dot{\mathbf{x}} = \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \Big|_{\mathbf{x}_0(t), \mathbf{u}_0(t)} \Delta \mathbf{x} + \frac{\partial \mathbf{f}}{\partial \mathbf{u}} \Big|_{\mathbf{x}_0(t), \mathbf{u}_0(t)} \Delta \mathbf{u} \quad (1.20)$$

$$\Delta \dot{\mathbf{y}} = \frac{\partial \mathbf{h}}{\partial \mathbf{x}} \Big|_{\mathbf{x}_0(t), \mathbf{u}_0(t)} \Delta \mathbf{x} + \frac{\partial \mathbf{h}}{\partial \mathbf{u}} \Big|_{\mathbf{x}_0(t), \mathbf{u}_0(t)} \Delta \mathbf{u} \quad (1.21)$$

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

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

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

**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

$$m\dot{v} = -\frac{1}{2}C_D\rho Av^2 + K_t u \quad (1.24)$$

where  $v$  is the velocity and  $m$  is the mass of the car,  $C_D$  is the drag coefficient,  $\rho$  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  $\dot{v}_0 = 0$  in the model, which gives

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

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

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

**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

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

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

$$i_d = \sqrt{\frac{mg}{C}} z_d \quad (1.28)$$

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

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

### 1.2.4 Linearization of second order systems

A second order system

$$\ddot{\mathbf{q}} = \mathbf{f}(\mathbf{q}, \dot{\mathbf{q}}, \mathbf{u}) \quad (1.30)$$

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

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

Taylor series expansion of the model around the solution gives

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

and combination with (1.31) gives the linearized model

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

**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  $\theta$  is set to zero when the pendulum is hanging downwards. The equation of motion for the pendulum is

$$mL^2\ddot{\theta} + mLg \sin \theta = 0 \quad (1.34)$$

which can be written

$$\ddot{\theta} = -\frac{g}{L} \sin \theta \quad (1.35)$$

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

$$\ddot{\theta} = -\frac{g}{L} \theta \quad (1.36)$$

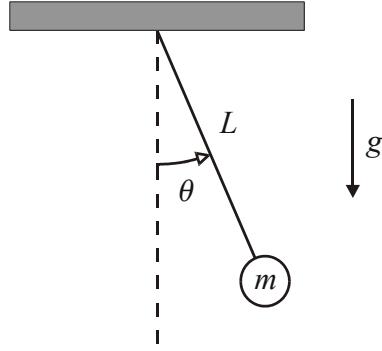


Figure 1.2: Pendulum

### 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

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

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

$$\mathbf{x} = \begin{pmatrix} \theta \\ \dot{\theta} \end{pmatrix} \quad (1.38)$$

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

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

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

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu} \quad (1.40)$$

$$y = \mathbf{Cx} + \mathbf{Du} \quad (1.41)$$

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

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

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

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

where  $\mathbf{K}_i$ ,  $i = 1, \dots, n$  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

$$\begin{aligned} \operatorname{Re}[\lambda_i] &\leq 0, & \lambda_i &\text{ simple eigenvalue} \\ \operatorname{Re}[\lambda_i] &< 0, & \lambda_i &\text{ multiple eigenvalue} \end{aligned} \quad (1.44)$$

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

$$\operatorname{Re}[\lambda_i] < 0 \quad i = 1, \dots, n \quad (1.45)$$

### 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

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

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

$$\ddot{\theta} + 2\zeta\omega_0\dot{\theta} + \omega_0^2\theta = 0 \quad (1.47)$$

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

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

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

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

$$y(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}u(t) \quad (1.50)$$

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

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

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

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

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

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

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

$$s\mathbf{x}(s) = \mathbf{Ax}(s) + \mathbf{Bu}(s) \quad (1.54)$$

$$y(s) = \mathbf{Cx}(s) + \mathbf{Du}(s) \quad (1.55)$$

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

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

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

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

and write

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

A block diagram is shown in Figure 1.3.

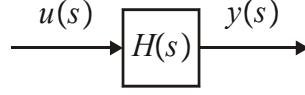


Figure 1.3: Transfer function representation of system

### 1.3.3 Rational transfer functions

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

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

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

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

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

$$y(s) = H(s) \cdot 1 = H(s) \quad (1.61)$$

which corresponds to the time function

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

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

We define the *unit step function*

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

which has the Laplace transform

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

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

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

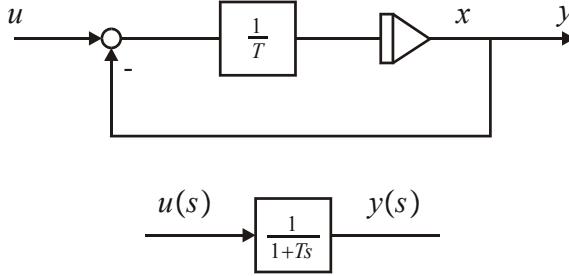


Figure 1.4: A time constant

**Example 8** *The dynamic system*

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

$$y = x \quad (1.67)$$

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

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

The impulse response of the system is

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

while the step response of the system is

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

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

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

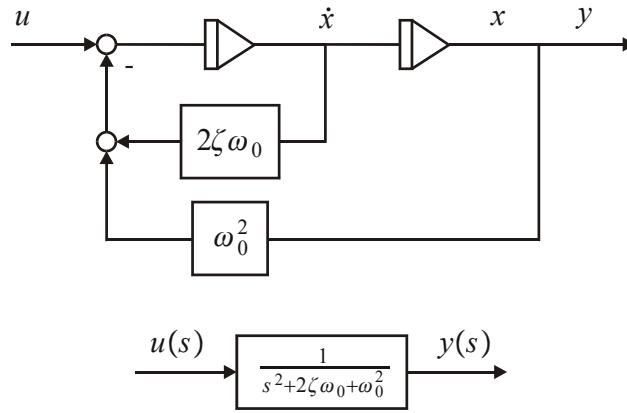


Figure 1.5: Second order oscillatory system

**Example 10** The model

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

$$y(t) = x(t) \quad (1.73)$$

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

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

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

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

The transfer function is

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

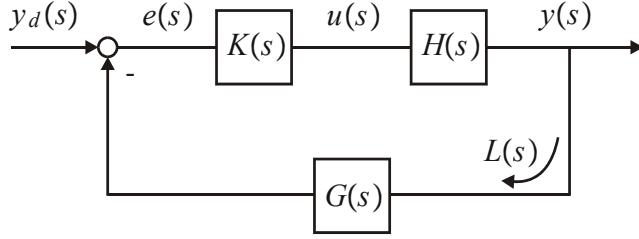


Figure 1.6: Plant  $H(s)$  with a series compensation controller  $K(s)$  in feedback compensation controller  $G(s)$

### 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

$$y(s) = H(s)u(s) \quad (1.77)$$

and the controller is given by

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

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

$$L(s) = K(s)H(s)G(s). \quad (1.79)$$

From

$$e(s) = y_d(s) - G(s)y(s) = y_d(s) - G(s)H(s)K(s)e(s) \quad (1.80)$$

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

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

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

$$y(s) = K(s)H(s)e(s) = K(s)H(s)[y_d(s) - G(s)y(s)] \quad (1.82)$$

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

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

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

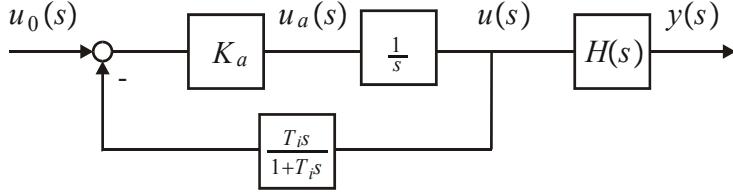


Figure 1.7: System with dynamic feedback control of the actuator.

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

$$y(s) = H(s) u(s) \quad (1.84)$$

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

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

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

$$u_a(s) = K_a [u_0(s) - G_a(s) u(s)] \quad (1.86)$$

where the dynamic feedback is the high-pass filter

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

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

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

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

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

where

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

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

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

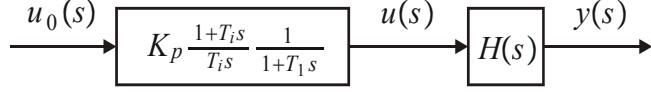


Figure 1.8: System with the closed loop dynamics of the actuator with controller.

### 1.3.7 Stability of transfer functions

Consider the system

$$y(s) = H(s)u(s) \quad (1.92)$$

where  $H(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}$  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  $\lambda_i$  of  $H(s)$  have real parts that are less than zero, that is,  $\text{Re } \lambda_i < 0$ ,  $i = 1, \dots, n$ .

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

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

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

$$y(t) = \int_0^\infty h(\tau) u(t - \tau) d\tau \quad (1.94)$$

Taking the absolute values on both sides, we find that

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

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

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

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

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

We find that

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

which shows that when  $\operatorname{Re}[\lambda_i] < 0$ , 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  $y(t) \propto t \rightarrow \infty$ . 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  $y(t) \propto t \cos \omega_0 t \rightarrow \infty$  where  $\propto$  denotes proportional to. If there are poles with real part larger than zero, then  $y(t)$  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  $\delta$  to the course angle  $\psi$  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)

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

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  $\psi$ . The transfer function has a pole at  $s = 1/T_2$ . An autopilot with a PD controller

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

gives the characteristic equation

$$s(1 + 0.1 T_1 s)(T_2 s - 1) + K K_p (1 + T_a s) = 0 \quad (1.101)$$

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

$$c \frac{\partial v(x, t)}{\partial x} = -\frac{\partial v(x, t)}{\partial t}, \quad v(0, t) = v_1(t) \quad (1.102)$$

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  $v(x, s) = \mathcal{L}\{v(x, t)\}$ , and the time derivative has the transform

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

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

$$c \frac{\partial v(x, s)}{\partial x} = -sv(x, s), \quad v(0, s) = v_1(s) \quad (1.104)$$

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

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

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

$$\frac{v_2}{v_1}(s) = e^{-Ts} \quad (1.106)$$

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  $P_k^k(-Ts)$  of order  $k$  where (Golub and van Loan 1989, p. 557)

$$P_k^k(s) = \frac{Q_{kk}(s)}{Q_{kk}(-s)} \quad (1.107)$$

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

A third order Padé approximation is found to be given by

$$e^{-Ts} \approx P_3^3(-Ts) = \frac{1 - \frac{Ts}{2} + \frac{(Ts)^2}{10} - \frac{(Ts)^3}{120}}{1 + \frac{Ts}{2} + \frac{(Ts)^2}{10} + \frac{(Ts)^3}{120}} \quad (1.109)$$

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

$$\tanh s = \frac{\sinh s}{\cosh s} \quad (1.110)$$

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

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

This occurs for

$$s = jk\pi \quad (1.112)$$

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

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

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

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

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  $-90^\circ$  and  $+90^\circ$ .

## 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

	Translation	Rotation	Thermal flow	Hydrostatic flow	Electrical
Effort $e$	$F$ (N)	$\tau$ (Nm)	$h$ (J/kg)	$p$ (N/m <sup>2</sup> )	$u$ (V)
Flow $f$	$v$ (m/s)	$\omega$ (1/s)	$w$ (kg/s)	$q$ (m <sup>3</sup> /s)	$i$ (A)
Power $P$	$Fv$ (W)	$\tau\omega$ (W)	$wh$ (W)	$pq$ (W)	$vi$ (W)

Table 1.1: Efforts and flows for physical systems.

leads naturally to object-oriented modeling, which is of great use in simulation. Energy-flow models can be assigned signal flow directions so that the model can be described in state-space, or with transfer functions. This means that the use of an energy-flow description in the modeling phase goes well together with the use of signal-flow methods in the controller design and analysis phase.

Object-oriented modeling 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. To make this interconnection possible, it is necessary that a suitable interface is defined between the subsystem models. As mentioned above, such an interface has been established in the form of energy-flow variables. The main advantage of this approach is that a library of models can be developed for different physical subsystems, and models for a total system can then be established by simply interconnecting these library models. Moreover, additional subsystems can easily be attached to the systems, and subsystems can be upgraded or changed by simply changing the relevant library module. This leads to re-use of models and straightforward updating of subsystem models. This approach is very useful in the development of simulation systems.

This chapter will present methods for this modeling technique. In particular we will focus on the network description that to a large extent originates from electrical circuit theory. We will also comment on the bond-graph formulation, which was developed for energy-flow modeling of systems consisting of subsystems like electrical circuits, electrical motors, hydraulic motors and mechanical parts. We will show that object-oriented modeling leads to models that are well suited for use in control techniques based on state space and transfer functions.

### 1.4.2 Background

The network description has been very successful in the analysis and design of electrical circuits (Anderson and Vongpanitlerd 1973), (Nilsson 1983). The underlying properties of an electrical circuit that make the network description efficient are seen also in other types of physical systems like mechanical translation, mechanical rotation, thermal flow and hydrostatic flow. To make the discussion general it is advantageous to define flow variables  $f$  and effort variables  $e$  for these systems. The flow  $f$  corresponds to the current  $i$  in an electrical network, while the effort variable  $e$  corresponds to the voltage  $u$  in an electrical network. The efforts and flows for typical physical systems is shown in Table 1.1.

### 1.4.3 Multiport

A *multiport*, which is also called an  $n$ -port, is a system with  $n$  ports. To port has an effort  $e_k$  and a flow  $f_k$  so that the net power flowing into the  $n$ -port is given by

$$P = \sum_{k=1}^n e_k f_k \quad (1.115)$$

At each port of the  $n$ -port it is possible to connect a port of another  $m$ -port system as long as the effort and flow variables are compatible, that is, an electrical port can be connected with an electrical port, a translational port can be connected to a translational port and so on.

### 1.4.4 Example: DC motor with flexible load

Transfer function models give a representation of a control system which is modular with respect to the signal flow. It gives a clear model structure with well defined inputs and outputs of the physical system and the controller. Therefore, when the signal-flow description is used, it is convenient to connect the plant output to a controller, and to connect the output of the controller as the input to the plant. Moreover, it is straightforward to change controller type and controller structure in this setting.

However, signal flow models may be cumbersome to modify when the physical system is modified. To illustrate this we may consider a DC motor with inertia  $J_m$  with no load. We let the output be the motor velocity  $\omega_m$ , while the input is the motor torque  $T$ . The transfer function is found from the equation of motion

$$J_m \dot{\omega}_m(t) = T(t) \quad (1.116)$$

which gives the transfer function model

$$\frac{\omega_m}{T}(s) = \frac{1}{J_m s} \quad (1.117)$$

Suppose that there is a need to modify the model to account for a load with inertia  $J_1$  that is driven through a transmission with a spring and a damper. Then the elasticity of the transmission has to be included in the transfer function. The transfer function changes to

$$\frac{\omega_m}{T}(s) = \frac{1}{J_s} \frac{1 + 2\zeta_a \frac{s}{\omega_a} + \left(\frac{s}{\omega_a}\right)^2}{1 + 2\zeta_1 \frac{s}{\omega_1} + \left(\frac{s}{\omega_1}\right)^2} \quad (1.118)$$

where the parameters are given by

$$J_e = \frac{J_m J_1}{J} \quad \text{and} \quad J = J_m + J_1 \quad (1.119)$$

$$\omega_1 = \sqrt{\frac{K}{J_e}}, \quad \zeta_1 = \frac{D}{2} \frac{1}{\sqrt{J_e K}}, \quad \omega_a = \sqrt{\frac{J_m}{J}} \omega_1 \quad \text{and} \quad \zeta_a = \sqrt{\frac{J_m}{J}} \zeta_1 \quad (1.120)$$

The derivation of this transfer function requires some work, but it is possible to do by hand. A critical observation is that the model parameters are functions of both the motor parameters and the load parameters.

If one more elastic transmission and an inertia  $J_2$  is added, then the transfer function becomes

$$\frac{\omega_m}{T}(s) = \frac{1}{(J_m + J_1 + J_2)s} \frac{\left(1 + 2\zeta_{2a}\frac{s}{\omega_{2a}} + \left(\frac{s}{\omega_{2a}}\right)^2\right)}{\left(1 + 2\zeta_{21}\frac{s}{\omega_{21}} + \left(\frac{s}{\omega_{21}}\right)^2\right)} \frac{\left(1 + 2\zeta_{2b}\frac{s}{\omega_{2b}} + \left(\frac{s}{\omega_{2b}}\right)^2\right)}{\left(1 + 2\zeta_{22}\frac{s}{\omega_{22}} + \left(\frac{s}{\omega_{22}}\right)^2\right)} \quad (1.121)$$

At this level the modeling becomes quite complicated, and the inclusion of even more degrees of freedom will lead to very extensive modeling efforts. We conclude that the inclusion of new physical objects in the model may be quite complicated to account for in the transfer function setting, and that the complexity increases dramatically when the order of the system increases. In particular, we note that in the present formulation we do not have a transfer function library of motors and elastic transmissions that can easily be combined.

In contrast to this the network approach to modeling makes it possible to assemble the model from the models of the physical objects of the system. In that case the modules are connected through the ports. The motor model is

$$J_m \dot{\omega}_m = T_m - T_L \quad (1.122)$$

If the motor is running alone, then  $T_L$  is set to zero. The inertia  $J_1$  and the flexible transmission is described by

$$J_1 \dot{\omega}_1 = T_L - T_1 \quad (1.123)$$

$$\frac{d}{dt}(\theta_m - \theta_1) = (\omega_m - \omega_1) \quad (1.124)$$

$$T_L = D_1(\omega_m - \omega_1) + K_1(\theta_m - \theta_1) \quad (1.125)$$

The motor model (1.122) and the load model (1.123–1.125) are then connected through the port variables  $T_L$  and  $\omega_m$ , while  $T_1$  is set to zero. Moreover, a second flexibility described by

$$J_2 \dot{\omega}_2 = T_1 - T_2 \quad (1.126)$$

$$\frac{d}{dt}(\theta_1 - \theta_2) = (\omega_1 - \omega_2) \quad (1.127)$$

$$T_1 = D_2(\omega_1 - \omega_2) + K_2(\theta_1 - \theta_2) \quad (1.128)$$

can be connected to the model with the port variables  $T_1$  and  $\omega_m$  with  $T_2 = 0$ . We see that this leads to what may be called a object-oriented approach where each physical object has a model, and where a model of an interconnection of physical objects is obtained by simply interconnecting the models of the objects through port variables. This approach scales well in the sense that the complexity of the modeling does not increase with the order of the model. The key to this is the careful selection of the interconnection variables. Note that once the model has been established, a state-space model is available, and a signal-flow model can be obtained by selecting input and output. This makes it possible to apply controller design based on signal flow representation.

#### 1.4.5 Example: Voltage controlled DC motor

To illustrate the combination of electrical and mechanical ports in the network setting we discuss the model structure of a voltage controlled DC electrical motor with an inertial load that is driven over an elastic shaft. The port interconnections are shown in

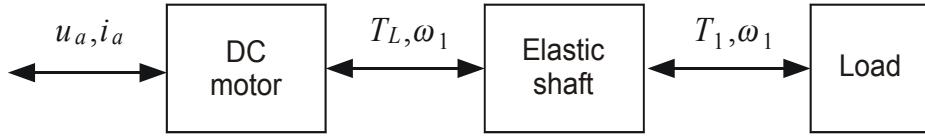


Figure 1.9: DC motor with elastic shaft and inertial load

Figure 1.9. The DC motor with constant field is connected to the electric supply by the terminals of the armature circuit and the motor shaft. The motor may be described as a two-port where one port is electrical and one port is mechanical. The electrical port is the armature port where the effort variable is the armature voltage  $u_a$  and the flow variable is the armature current  $i_a$ . The mechanical port is the motor shaft where the effort is the load torque  $T_L$  and the flow is the angular velocity  $\omega_1$ . The dynamic model of the motor is given in state-space form as

$$L_a \frac{di_a}{dt} = -R_a i_a - K_E \omega_m + u_a \quad (1.129)$$

$$J_m \dot{\omega}_m = K_T i_a - T_L \quad (1.130)$$

$$\dot{\theta}_m = \omega_m \quad (1.131)$$

Suppose that the motor shaft is connected to a mechanical two-port describing a spring and a damper. The port on the motor side has port variables  $T_L$  and  $\omega_m$ , and the port on the load side has variables  $T_1$  and  $\omega_1$ . The model is

$$\frac{d}{dt} (\theta_m - \theta_1) = \omega_m - \omega_1 \quad (1.132)$$

$$T_L = K(\theta_m - \theta_1) + B(\omega_m - \omega_1) \quad (1.133)$$

$$T_1 = T_L \quad (1.134)$$

Finally, the spring and the damper is connected to an inertial load described as a one-port with port variables  $T_1$  and  $\omega_1$ . The model is

$$J_m \dot{\omega}_1 = T_L \quad (1.135)$$

$$\dot{\theta}_1 = \omega_1 \quad (1.136)$$

#### 1.4.6 Example: Diesel engine with turbocharger

A model of a diesel engine with turbocharger can be described as a system of three multiports, namely, the compressor, the turbine, and the diesel engine. The port interconnections are shown in Figure 1.10. The turbine and the compressor have a common shaft, and it is convenient to include the inertia of the shaft in the turbine model.

The compressor has one port for the air intake, one port for the air outlet, and one for the turbocharger shaft. The port variables for the air intake is the specific enthalpy  $h_{ci}$  of the air, which is the effort, and the mass flow  $w_{ci}$ , which is the flow. The port variables of the air outlet is the specific enthalpy  $h_{co}$  and the mass flow  $w_{co}$ . The port variables for the turbocharger shaft is the compressor torque  $T_c$  and the angular velocity  $\omega_{tc}$ . The turbine has an air intake port with port variables  $h_{ti}$  and  $w_{ti}$ , and air outlet port with port variables  $h_{to}$  and  $w_{to}$ , and the turbine shaft has a port with variables  $T_c$  and the angular velocity  $\omega_{tc}$ . The diesel engine has an air intake port connected to the air outlet

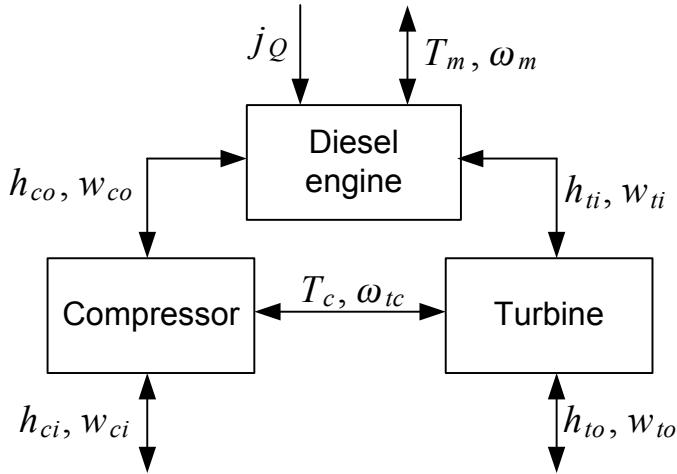


Figure 1.10: Diesel engine with turbocharger in network description

port of the compressor with port variables  $h_{co}$  and the mass flow  $w_{co}$ . The engine has an exhaust outlet port connected to the air intake port of the turbine with port variables  $h_{ti}$  and  $w_{ti}$ . In addition, the motor drives the motor shaft which can be described as a port with variables being the motor torque  $T_m$  and the motor shaft speed  $\omega_m$ . We note that the power delivered from the engine is  $P_m = T_m \omega_m$ . Finally, the engine has a fuel injector where the power added is the heat rate  $j_Q$  of the fuel. However, this is not a port in the usual sense as there is no bilateral energy flow.

#### 1.4.7 Assigning computational inputs and outputs

In the network description using multiports the signal flow directions are not specified. This agrees with our intuition that energy flows both ways, and that the multiports interact through the port interconnections. However, in simulation systems where the multiport models are used in computations, it must be specified which of the port variables that is the input to the computation and which of the port variables that is the output from the computation. This means that a signal flow structure is assigned to the model. This must be done so that:

1. Differential equations can be evaluated by integration and not differentiation.
2. Computational inputs and computational outputs must be compatible when port interconnections are made so that signal flow directions agree.

**Example 15** Signal directions can be assigned according to Figure 1.11 for the DC motor with load shown in Figure 1.9 so that integrations are used in the computations. This is done by using  $u_a$  and  $T_L$  as inputs to the DC motor so that  $di_a/dt$ ,  $\dot{\omega}_m$  and  $\dot{\theta}_m$  can be evaluated from (1.129–1.131). For the spring  $\omega_m$  and  $\omega_1$  are used as inputs so that  $T_L$  and  $T_1$  can be computed from (1.132–1.134), and for the load the input is  $T_1$  so that  $\dot{\omega}_1$  and  $\dot{\theta}_1$  can be computed from (1.135–1.136).

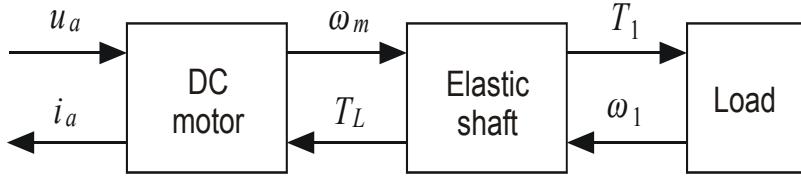


Figure 1.11: DC motor and load in network descriptions with signal directions assigned

**Example 16** Signal directions are indicated in Figure 1.12 for the diesel engine with turbocharger as shown in Figure 1.10. Given the turbocharger speed  $\omega_{tc}$  and the specific enthalpies  $h_{ci}$  and  $h_{co}$  the compressor will produce a mass flow  $w_c$  to the engine, and a compressor torque  $T_c$  on the shaft. Given the compressor torque  $T_c$  and the specific enthalpies  $h_{ti}$  and  $h_{to}$ , the turbine dynamics will give the turbocharger speed  $\omega_{tc}$  and the exhaust mass flow  $w_t$ . The gas contained in the diesel engine will have as inputs the mass flow  $w_c$  into the engine, the mass flow  $w_t$  out of the engine, the motor shaft speed  $\omega_m$ , and the heat flow  $j_Q$ . The resulting outputs are the specific enthalpies  $h_{co}$  and  $h_{ti}$ , and the motor torque  $T_m$ .

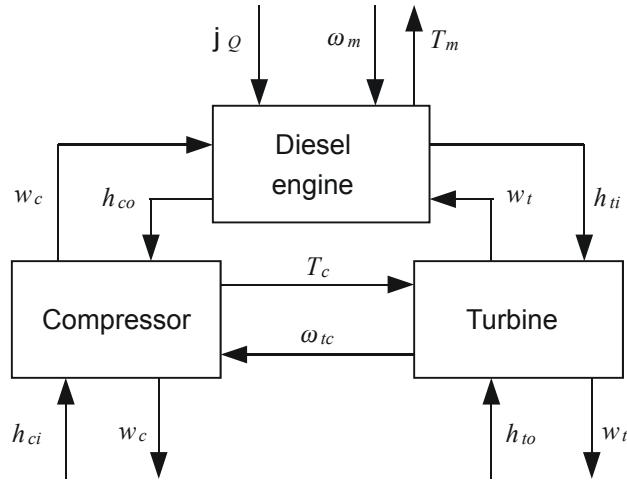


Figure 1.12: Turbocharged diesel engine with signal directions assigned for the port variables

**Example 17** Karnopp's friction model extends the basic Coulomb friction model for dry friction to be valid also for zero velocity (Karnopp 1985). Karnopp's model can be explained by considering a mass  $m$  with velocity  $v$  that is pushed on a flat surface with an actuator force  $F_a$ . The friction force on the mass is  $F_f$  so that the equation of motion is

$$m\dot{v} = F_a - F_f \quad (1.137)$$

The Karnopp model can be formulated as a two-port where the input port has effort  $F_a$

and flow  $v$ , while the output port has effort  $F_a - F_f$  and flow  $v$ , and where

$$F_f = \begin{cases} F_a & v = 0 \text{ and } |F_a| \leq F_c \\ F_c \operatorname{sgn}(v) & \text{else} \end{cases} \quad (1.138)$$

Note that the computational input of the model at the input port is  $F_a$  under the condition  $v = 0$  and  $|F_a| \leq F_c$ , and that the computational input changes to  $v$  when the condition does no longer hold. Further discussion on this model is found in Chapter 5.

### 1.4.8 Bond graphs

In the basic definition of network models there is no assignment of the direction of signal flow. However, to use network models in simulation the signal flow directions must be specified. This gives a network with specified signal flow structure. By adding a special graph representation of the network and the of signal flow, a formalism called *bond graphs* is obtained (Karnopp, Margolis and Rosenberg 2000). Bond graph theory provides us with interesting tools for connecting network modules when the signal flow structure is important, as is the case if the model is to be used for simulations. Note that bond-graph models combines well with state-space formulations of networks, where the bond graph techniques can be used to assign signal-flow directions. The terminology used in bond graphs is easily related to the network terminology. In particular, a bond is the same as a port, and *computational causality* is related to the selection of inputs and outputs to a computation.

In the bond graph setting the interconnection between multiports are given by *bonds*. Each bond transmits an effort signal  $e$  and a flow signal  $f$ . The effort and flow are selected so that the product  $ef$  has dimension power. We note that the effort and flow for a bond corresponds to the port variables in the network description. A bond is assigned a direction, which is the direction of positive power flow. From the outset we are free to choose the direction of the effort signal for a bond, although in interconnections of bonds there are certain rules that must be obeyed. The topic of bond graphs will not be further discussed in this book, but the methods that are used in this book are closely related to the concept of bond graphs.

## 1.5 Linear network theory

### 1.5.1 Driving point impedance

An electrical multiport is an electrical circuit which can be connected to the outside via  $n$  ports. Each port has one positive and one negative electrical terminal. The voltage over the terminals of port  $k$  is  $u_k$ , while the current into the positive terminal and out of the negative terminal is  $i_k$ . This means that the power flowing into the circuit from port  $k$  is  $P_k = u_k i_k$ . The port of one electrical multiport can be connected to a port of another electrical multiport by connecting the terminals of the two ports.

A linear time-invariant electrical one-port with voltage  $u$  and current  $i$  can be described by its *driving-point impedance*  $Z(s)$  which is the impedance of the port so that

$$u(s) = Z(s) i(s) \quad (1.139)$$

In the same way the *driving-point admittance*  $Y(s) = Z(s)^{-1}$  satisfies

$$i(s) = Y(s) u(s) \quad (1.140)$$

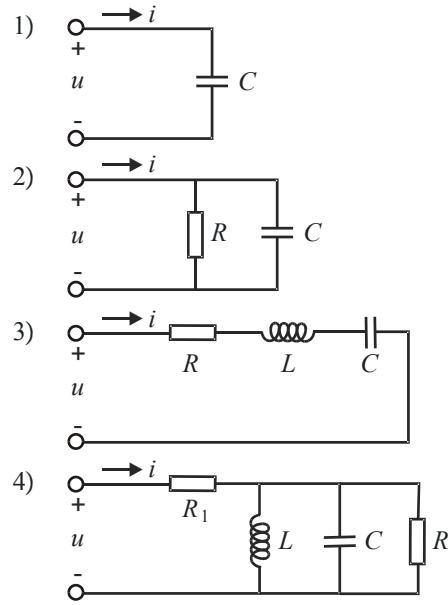


Figure 1.13: Passive eletrical one-ports

To become more familiar with the notion of the driving point impedance of passive electrical one-ports we present the driving-point impedance of the one-ports in Figure 1.13. Circuit 1 is a capacitor, circuit 2 is a parallel interconnection of a resistor and a capacitor, circuit 3 is a RLC series connection, and circuit 4 is a resistor in series with a parallel RLC interconnection. The driving point impedances are

$$Z_1(s) = \frac{1}{Cs} \quad (1.141)$$

$$Z_2(s) = \frac{R}{1 + RCs} \quad (1.142)$$

$$Z_3(s) = \frac{1 + RCs + LCs^2}{Cs} \quad (1.143)$$

$$Z_4(s) = R_1 \frac{1 + \left(\frac{L}{R_1} + \frac{L}{R}\right)s + LCs^2}{1 + \frac{L}{R}s + LCs^2} \quad (1.144)$$

We see that  $Z_1(s)$  and  $Z_2(s)$  are strictly proper,  $Z_3(s)$  is not proper, and  $Z_4(s)$  is proper. It is straightforward to verify that the poles of the impedances appear in the left half plane, or as simple poles on the imaginary axis, and that the phase of the impedances is between  $-90^\circ$  and  $+90^\circ$ .

We define the driving point impedance  $Z(s)$  and the driving point admittance  $Y(s)$  of a general one-port with effort  $e$  and flow  $f$  according to

$$e(s) = Z(s)f(s), \quad f(s) = Y(s)e(s) \quad (1.145)$$

### 1.5.2 Linear two-ports

A linear two-port can be described with a transfer function matrix. The convention is then that the flows are positive when they are directed into the two-port. The transfer function matrices give the relation between two input variables and two output variables. Different transfer function matrices result depending on which variables that are selected as inputs and outputs variables. In the impedance description the flows are inputs and the efforts are outputs, and the transfer function model of the two-port is written

$$\begin{pmatrix} e_1(s) \\ e_2(s) \end{pmatrix} = \begin{pmatrix} z_{11}(s) & z_{12}(s) \\ z_{21}(s) & z_{22}(s) \end{pmatrix} \begin{pmatrix} f_1(s) \\ f_2(s) \end{pmatrix} \quad (1.146)$$

In the admittance form the efforts are inputs and the flows are outputs, and the model is written

$$\begin{pmatrix} f_1(s) \\ f_2(s) \end{pmatrix} = \begin{pmatrix} y_{11}(s) & y_{12}(s) \\ y_{21}(s) & y_{22}(s) \end{pmatrix} \begin{pmatrix} e_1(s) \\ e_2(s) \end{pmatrix} \quad (1.147)$$

A cascade formulation can be used where the variables of port two are inputs and the variables of port one are outputs, or vice versa, which leads to the two model representations

$$\begin{pmatrix} e_1(s) \\ f_1(s) \end{pmatrix} = \begin{pmatrix} a_{11}(s) & a_{12}(s) \\ a_{21}(s) & a_{22}(s) \end{pmatrix} \begin{pmatrix} e_2(s) \\ -f_2(s) \end{pmatrix} \quad (1.148)$$

and

$$\begin{pmatrix} e_2(s) \\ f_2(s) \end{pmatrix} = \begin{pmatrix} b_{11}(s) & b_{12}(s) \\ b_{21}(s) & b_{22}(s) \end{pmatrix} \begin{pmatrix} e_1(s) \\ -f_1(s) \end{pmatrix} \quad (1.149)$$

The hybrid formulation is based on having the effort of one port and the flow of the other port as inputs, and then having the remaining flow and the remaining effort as outputs. This gives the two alternative model formulations

$$\begin{pmatrix} e_1(s) \\ f_2(s) \end{pmatrix} = \begin{pmatrix} h_{11}(s) & h_{12}(s) \\ h_{21}(s) & h_{22}(s) \end{pmatrix} \begin{pmatrix} e_2(s) \\ f_1(s) \end{pmatrix} \quad (1.150)$$

and

$$\begin{pmatrix} f_1(s) \\ e_2(s) \end{pmatrix} = \begin{pmatrix} g_{11}(s) & g_{12}(s) \\ g_{21}(s) & g_{22}(s) \end{pmatrix} \begin{pmatrix} f_2(s) \\ e_1(s) \end{pmatrix} \quad (1.151)$$

In (Nilsson 1983) this is discussed in further detail, and formulas for transforming the parameters of one formulation to the parameters of another formulation are given.

### 1.5.3 Impedance of two-port with termination

Consider the case where a linear two-port is terminated with the impedance  $Z_L(s)$  at port 2. This means that port two is connected to a one-port

$$e_L(s) = Z_L(s) f_L(s) \quad (1.152)$$

with impedance  $Z_L(s)$ . The interconnection is ensured with the conditions

$$e_2 = e_L, \quad f_2 = -f_L \quad (1.153)$$

From the impedance formulation (1.146) of the two-port it is seen that the termination gives

$$\begin{pmatrix} e_1(s) \\ -Z_L(s) f_2 \end{pmatrix} = \begin{pmatrix} z_{11}(s) & z_{12}(s) \\ z_{21}(s) & z_{22}(s) \end{pmatrix} \begin{pmatrix} f_1(s) \\ f_2(s) \end{pmatrix} \quad (1.154)$$

This set of equations can be solved to find the driving-point impedance description

$$e_1(s) = Z_1(s)f_1(s) \quad (1.155)$$

of the resulting one-port. From the second row of (1.154) it is found that

$$f_2 = -\frac{z_{21}}{z_{22} + Z_L}f_1 \quad (1.156)$$

Insertion in the first row of (1.154) gives

$$e_1 = \left( z_{11} - \frac{z_{12}z_{21}}{z_{22} + Z_L} \right) f_1 \quad (1.157)$$

Therefore, the terminated one-port has driving-point impedance

$$Z_1 = z_{11} - \frac{z_{12}z_{21}}{z_{22} + Z_L} \quad (1.158)$$

#### 1.5.4 Example: Passive mechanical two-port

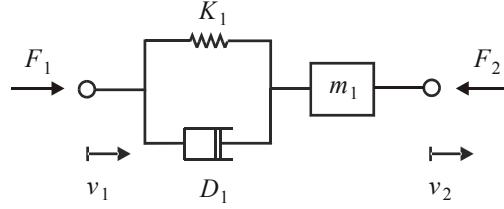


Figure 1.14: Mechanical two-port

The mechanical system in Figure 1.14 can be described as a two-port. A mass \$m\_1\$ is connected to a parallel interconnection of a spring with stiffness \$K\_1\$ and a damper with coefficient \$D\_1\$. Port 1 is connected to the spring and the damper, and has flow equal to the velocity \$v\_1\$ and effort equal to the force \$F\_1\$. Port 2 is connected to the mass, and has flow \$v\_2\$ and effort \$F\_2\$. The equation of motion for the mass is given by Newton's law to be

$$m_1 \ddot{x}_2 = F_1 - F_2 \quad (1.159)$$

where the force from the spring and the damper is

$$F_1 = K_1(x_1 - x_2) + D_1(v_1 - v_2) \quad (1.160)$$

Laplace transformation gives

$$m_1 s v_2(s) = F_1(s) - F_2(s) \quad (1.161)$$

and

$$F_1(s) = \frac{K_1 + D_1 s}{s} [v_1(s) - v_2(s)] \quad (1.162)$$

as  $sx_1(s) = v_1(s)$  and  $sx_2(s) = v_2(s)$ . By solving for  $F_2(s)$  and  $v_2(s)$  we get the cascade description

$$\begin{pmatrix} F_2(s) \\ v_2(s) \end{pmatrix} = \begin{pmatrix} \frac{m_1 s^2 + D_1 s + K_1}{D_1 s + K_1} & -m_1 s \\ -\frac{s}{D_1 s + K_1} & 1 \end{pmatrix} \begin{pmatrix} F_1(s) \\ v_1(s) \end{pmatrix} \quad (1.163)$$

Alternatively, we may develop the impedance description

$$\begin{pmatrix} F_1(s) \\ F_2(s) \end{pmatrix} = \begin{pmatrix} \frac{K_1 + D_1 s}{s} & \frac{K_1 + D_1 s}{s} \\ \frac{K_1 + D_1 s}{s} & \frac{m_1 s^2 + D_1 s + K_1}{s} \end{pmatrix} \begin{pmatrix} v_1(s) \\ -v_2(s) \end{pmatrix} \quad (1.164)$$

or the hybrid description

$$\begin{pmatrix} v_1(s) \\ F_2(s) \end{pmatrix} = \begin{pmatrix} \frac{s}{K_1 + D_1 s} & -1 \\ 1 & m_1 s \end{pmatrix} \begin{pmatrix} F_1(s) \\ -v_2(s) \end{pmatrix} \quad (1.165)$$

We note that the mechanical two-port has an electrical analog as shown in Figure 1.15.

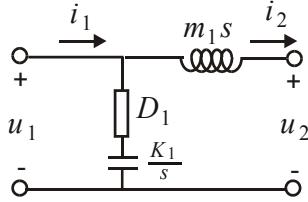


Figure 1.15: Electrical analog of mechanical two-port. The voltages  $u_1$  and  $u_2$  are the analogs of the forces  $F_1$  and  $F_2$ , while the currents  $i_1$  and  $i_2$  are the analogs of the velocities  $v_1$  and  $v_2$ .

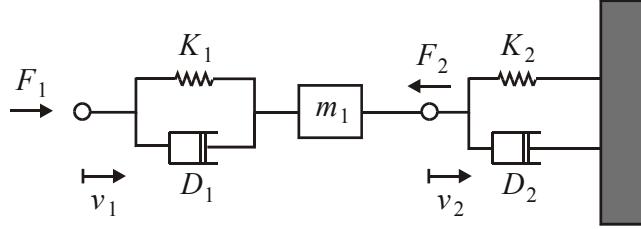


Figure 1.16: Termination of mechanical two-port with mechanical one-port with spring and damper.

Suppose that the two-port of the previous section is terminated at port 2 by a mechanical one-port consisting of a parallel interconnection of a spring with stiffness  $K_2$  and a damper with coefficient  $D_2$  connected to a fixed point. The resulting mechanical one-port is shown in Figure 1.16. The impedance of the termination is

$$Z_L(s) = \frac{K_2 + D_2 s}{s} = K_2 \frac{1 + \frac{D_2}{K_2} s}{s} \quad (1.166)$$

so that

$$F_2(s) = Z_L(s)v_2(s) \quad (1.167)$$

This gives the impedance

$$\frac{F_1}{v_1}(s) = Z_1 = z_{11} - \frac{z_{12}z_{21}}{z_{22} + Z_L} \quad (1.168)$$

After some calculation we find that

$$\frac{F_1}{v_1}(s) = \frac{K_1 + D_1 s}{s} \frac{K_2}{K_1 + K_2} \frac{(1 + \frac{D_2}{K_2} s + \frac{m_1}{K_2} s^2)}{(1 + \frac{D_1 + D_2}{K_1 + K_2} s + \frac{m_1}{K_1 + K_2} s^2)} \quad (1.169)$$

### 1.5.5 Mechanical analog of PD controller

We will now present mechanical analogs for PD controllers for position control when the control input is the applied force. Note that this is a PI controller from velocity.

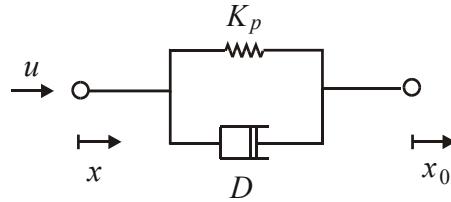


Figure 1.17: Mechanical analog of PD controller.

We consider a mass  $m$  with position  $x$  and velocity  $v = \dot{x}$ . The equation of motion is  $m\ddot{x} = u$  where the applied force  $u$  is the control input. The desired position is  $x_d$ , and the desired velocity is  $v_d = \dot{x}_d$ . A PD controller  $u = K(1 + T_d s)(x_d - x) + D(v_d - v)$  is used. The control law can be written

$$u = K(x_d - x) + D(v_d - v) \quad (1.170)$$

where  $D = KT_d$ . The mechanical analog is found from the observation that the controller force is the same force as the force that would appear if the mass  $m$  at position  $x$  was connected to a point of position  $x_d$  with a parallel interconnection of a spring with stiffness  $K$  and a damper with coefficient  $D$ , as shown in Figure 1.17.

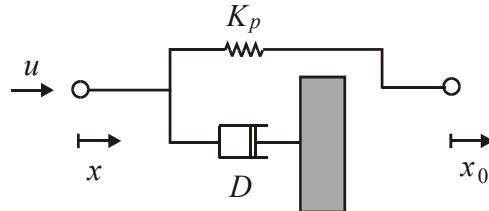


Figure 1.18: Mechanical analog of PD controller without velocity reference.

An alternative control law which is used when  $v_d$  is not available is

$$u = K_p(x_0 - x) - Dv \quad (1.171)$$

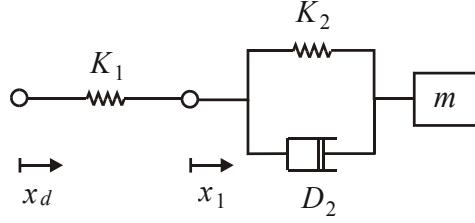


Figure 1.19: Mechanical analog mass controlled with lead-lag controller.

This is the force that would result if the mass  $m$  was connected by a spring of stiffness  $K$  to a point  $x_d$  and by a damper to a fixed point as shown in Figure 1.18.

If the velocity is not measured it is possible to use a PD controller with limited derivative action. Then the control is

$$u(s) = K_p \frac{1 + T_d s}{1 + \alpha T_d s} [x_0(s) - x(s)] \quad (1.172)$$

where  $0 \leq \alpha \leq 1$ . This control law gives the same force as a mechanical analog where the mass  $m$  is connected to a point with position  $x_d$  by a spring of stiffness  $K_1$  in series with a parallel interconnection of a spring of stiffness  $K_2$  and a damper with coefficient  $D_2$  as shown in Figure 1.19.

This is shown by letting  $x_1$  be the position of the interconnection point between the spring  $K_1$  and the parallel interconnection. Then the spring force is  $u = K_1(x_1 - x)$ , and Laplace transformation gives  $x_1(s) = x(s) + u(s)/K_1$ . As there is no mass in the point  $x_1$ , this force must be equal to the force from the parallel interconnection, so that

$$u(s) = K_2[x_d(s) - x_1(s)] + D_2[v_d(s) - v_1(s)] = (K_2 + D_2 s)[x_d(s) - x_1(s)] \quad (1.173)$$

Insertion of  $x_1(s)$  gives

$$u(s) = (K_2 + D_2 s)[x_0(s) - x(s) - \frac{1}{K_1} u(s)] \quad (1.174)$$

We solve for  $u(s)$  and get

$$\begin{aligned} u(s) &= K_1 \frac{K_2 + D_2 s}{K_1 + K_2 + D_2 s} [x_0(s) - x(s)] \\ &= \frac{K_1 K_2}{K_1 + K_2} \frac{1 + \frac{D_2}{K_2} s}{1 + \frac{K_2}{K_1 + K_2} \frac{D_2}{K_2} s} [x_0(s) - x(s)] \end{aligned} \quad (1.175)$$

We see that this is a PD controller with limited derivative action where

$$K_p = \frac{K_1 K_2}{K_1 + K_2} \quad (1.176)$$

$$T_d = \frac{D_2}{K_2} \quad (1.177)$$

$$\alpha = \frac{K_2}{K_1 + K_2} \in [0, 1] \quad (1.178)$$

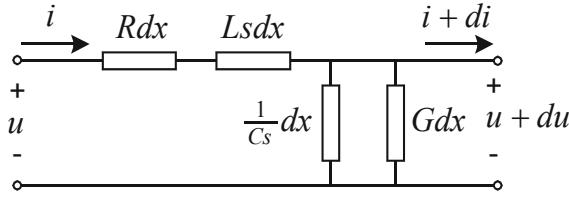


Figure 1.20: Length element of electric transmission line.

## 1.6 Example: Transmission line model

### 1.6.1 Introduction

The dynamics of transmission lines are important in several applications including electrical lines for signal transmission, water pipes for hydroelectric power plants, gas and oil pipelines, and systems with hydraulic drives. It is by no means obvious that these systems have dynamic properties in common, however, it turns out that a comprehensive theory of transmission line dynamics can be developed to describe important dynamic properties for all these systems in spite of their diverse nature. We will first derive the partial differential equations describing a transmission line, and we will then study techniques of analysis.

### 1.6.2 Introductory example

In the following we will present the model for a general transmission line of length  $L$  which is described as a two-port where one port is at the the input side with effort  $e_1$  and flow  $f_1$ , and the other port at the output side with effort  $e_2$  and the flow  $f_2$ . The model will be presented in the form of a partial differential equation with boundary conditions depending on the port variables. Then this model will be used to derive transfer functions. Also wave variables and impedance matching be discussed. Then, in a later chapter the results will be specialized for hydraulic transmission lines will, where the effort variable is the pressure and the flow variable is the volumetric flow.

To make the ideas of the following presentation clear we start with the equations of an electrical transmission line as an introductory example. To derive the model of an electric transmission line we consider a length element  $dx$  of the line which is described with the length coordinate  $x$ . The length element is modelled as by a series impedance consisting of a resistor  $Rdx$  and an inductor  $Mdx$ , and a parallel admittance consisting of an admittance  $Gdx$  and a capacitor  $Cdx$ . The voltage and current laws give

$$u(x + dx, t) - u(x, t) = -Rdx i(x, t) - Mdx \frac{\partial i}{\partial t}(x, t) \quad (1.179)$$

$$i(x + dx, t) - i(x, t) = -Gdx u(x + dx, t) - Cdx \frac{\partial u}{\partial t}(x + dx, t) \quad (1.180)$$

Dividing by  $dx$  we get

$$\frac{\partial u}{\partial x}(x, t) = -Ri(x, t) - M \frac{\partial i}{\partial t}(x, t) \quad (1.181)$$

$$\frac{\partial i}{\partial x}(x, t) = -Gu(x, t) - C \frac{\partial u}{\partial t}(x, t) \quad (1.182)$$

Laplace transformation of (1.181) and (1.182) gives the transmission line equations

$$\frac{\partial u}{\partial x}(x, s) = -(R + Ms)i(x, s) \quad (1.183)$$

$$\frac{\partial i}{\partial x}(x, s) = -(G + Cs)u(x, s) \quad (1.184)$$

Note that these equations have the form

$$\frac{\partial u}{\partial x}(x, s) = -Z(s)i(x, s) \quad (1.185)$$

$$\frac{\partial i}{\partial x}(x, s) = -Y(s)u(x, s) \quad (1.186)$$

where  $Z(s) = R + Ms$  is the series impedance, and  $Y(s) = G + Cs$  is the parallel admittance.

### 1.6.3 Effort and flow model

A general transmission line is characterized in terms of the line length  $L$ , the series impedance  $Z(s)$ , the parallel admittance  $Y(s)$ , and the two first-order differential equations

$$\frac{\partial e}{\partial x}(x, s) = -Z(s)f(x, s) \quad (1.187)$$

$$\frac{\partial f}{\partial x}(x, s) = -Y(s)e(x, s) \quad (1.188)$$

in the Laplace domain, where  $e$  is the effort variable and  $f$  is the flow variable.

From this characterization we arrive at the following second order differential equation for the effort

$$L^2 \frac{\partial^2 e}{\partial x^2}(x, s) = \Gamma^2(s)e(x, s) \quad (1.189)$$

where we have introduced the propagation operator  $\Gamma(s)$  defined by

$$\Gamma(s) = L\sqrt{Z(s)Y(s)} \quad (1.190)$$

and the additional requirement that  $\text{Re}[\Gamma(s)] \geq 0$ . The solution of this differential equation is

$$e(x, s) = C_1 \cosh\left(\Gamma \frac{x}{L}\right) + C_2 \sinh\left(\Gamma \frac{x}{L}\right) \quad (1.191)$$

where  $C_1$  and  $C_2$  are given by the boundary conditions. From (1.191) we see that the flow is given by

$$f(x, s) = -\frac{1}{Z_c(s)} \left( C_1 \sinh\left(\Gamma \frac{x}{L}\right) + C_2 \cosh\left(\Gamma \frac{x}{L}\right) \right) \quad (1.192)$$

where the *characteristic impedance*  $Z_c(s)$  is defined by

$$Z_c(s) = \sqrt{\frac{Z(s)}{Y(s)}} \quad (1.193)$$

### 1.6.4 Transfer functions

A transmission line of length  $L$  can be regarded as a two-port with the input port at  $x = 0$  and the output port at  $x = L$ . At the input port the effort is  $e(0, t)$  and the flow is  $f(0, t)$ , while at the output port the effort is  $e(L, t)$  and the flow is  $f(L, t)$ . The transfer functions between the port variables can be derived from (1.191, 1.192). We denote the Laplace transformed port variables  $e(0, s) = e_1(s)$ ,  $f(0, s) = f_1(s)$ ,  $e(L, s) = e_2(s)$  and  $f(L, s) = f_2(s)$ . From (1.191, 1.192) it is seen that we may then write the solutions in the form

$$\begin{pmatrix} e_1(s) \\ Z_c f_1(s) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} C_1(s) \\ C_2(s) \end{pmatrix} \quad (1.194)$$

$$\begin{pmatrix} e_2(s) \\ Z_c f_2(s) \end{pmatrix} = \begin{pmatrix} \cosh \Gamma & \sinh \Gamma \\ -\sinh \Gamma & -\cosh \Gamma \end{pmatrix} \begin{pmatrix} C_1(s) \\ C_2(s) \end{pmatrix} \quad (1.195)$$

From these two equations we can eliminate the constants  $C_1$   $C_2$  and find the *cascade form* of the transfer functions.

The *cascade form* of the transfer functions is given by

$$\begin{pmatrix} e_2(s) \\ f_2(s) \end{pmatrix} = \begin{pmatrix} \cosh \Gamma & -Z_c \sinh \Gamma \\ -\frac{\sinh \Gamma}{Z_c} & \cosh \Gamma \end{pmatrix} \begin{pmatrix} e_1(s) \\ f_1(s) \end{pmatrix} \quad (1.196)$$

$$\begin{pmatrix} e_1(s) \\ f_1(s) \end{pmatrix} = \begin{pmatrix} \cosh \Gamma & Z_c \sinh \Gamma \\ \frac{\sinh \Gamma}{Z_c} & \cosh \Gamma \end{pmatrix} \begin{pmatrix} e_2(s) \\ f_2(s) \end{pmatrix} \quad (1.197)$$

The cascade form can be used for analysis purposes like the derivation transfer functions when the line is terminated with a load impedance. However, the cascade form is an ill-posed boundary value problem, and numerical models derived from the cascade form may be ill-conditioned (Mäkinen, Piché and Ellman 2000).

By straightforward manipulation of the equations (1.194, 1.195) we can find the transfer functions between different combinations of port variables. First we consider the admittance forms which are the impedance form and the admittance form. These forms are found from

$$\underbrace{\begin{pmatrix} e_1(s) \\ e_2(s) \end{pmatrix}}_{\mathbf{e}} = \underbrace{\begin{pmatrix} 1 & 0 \\ \cosh \Gamma & \sinh \Gamma \end{pmatrix}}_{\mathbf{G}(s)} \underbrace{\begin{pmatrix} C_1(s) \\ C_2(s) \end{pmatrix}}_{\mathbf{c}} \quad (1.198)$$

$$Z_c \underbrace{\begin{pmatrix} f_1(s) \\ -f_2(s) \end{pmatrix}}_{\mathbf{f}} = \underbrace{\begin{pmatrix} 0 & -1 \\ \sinh \Gamma & \cosh \Gamma \end{pmatrix}}_{\mathbf{H}(s)} \underbrace{\begin{pmatrix} C_1(s) \\ C_2(s) \end{pmatrix}}_{\mathbf{c}} \quad (1.199)$$

Elimination of the vector  $\mathbf{c}$  gives the expressions

$$\mathbf{e} = Z_c \mathbf{G}(s) \mathbf{H}(s)^{-1} \mathbf{f}, \quad \mathbf{f} = \frac{1}{Z_c} \mathbf{H}(s) \mathbf{G}(s)^{-1} \mathbf{e} \quad (1.200)$$

This gives the *impedance form*

$$\begin{pmatrix} e_1(s) \\ e_2(s) \end{pmatrix} = Z_c \begin{pmatrix} \frac{\cosh \Gamma}{\sinh \Gamma} & \frac{1}{\sinh \Gamma} \\ \frac{1}{\sinh \Gamma} & \frac{\cosh \Gamma}{\sinh \Gamma} \end{pmatrix} \begin{pmatrix} f_1(s) \\ -f_2(s) \end{pmatrix} \quad (1.201)$$

and the *admittance form*

$$\begin{pmatrix} f_1(s) \\ -f_2(s) \end{pmatrix} = \frac{1}{Z_c} \begin{pmatrix} \frac{\cosh \Gamma}{\sinh \Gamma} & -\frac{1}{\sinh \Gamma} \\ -\frac{1}{\sinh \Gamma} & \frac{\cosh \Gamma}{\sinh \Gamma} \end{pmatrix} \begin{pmatrix} e_1(s) \\ e_2(s) \end{pmatrix} \quad (1.202)$$

The *hybrid form* of the transfer functions are found from

$$\underbrace{\begin{pmatrix} e_1(s) \\ -f_2(s) \end{pmatrix}}_{\mathbf{u}} = \underbrace{\begin{pmatrix} 1 \\ \frac{\sinh \Gamma}{Z_c} \end{pmatrix}}_{\mathbf{E}(s)} \underbrace{\begin{pmatrix} C_1(s) \\ C_2(s) \end{pmatrix}}_{\mathbf{c}} \quad (1.203)$$

$$\underbrace{\begin{pmatrix} f_1(s) \\ e_2(s) \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 0 & -\frac{1}{Z_c} \\ \cosh \Gamma & \sinh \Gamma \end{pmatrix}}_{\mathbf{F}(s)} \underbrace{\begin{pmatrix} C_1(s) \\ C_2(s) \end{pmatrix}}_{\mathbf{c}} \quad (1.204)$$

Proceeding as in the impedance and admittance case we eliminate the  $\mathbf{c}$  vector and get

$$\mathbf{y} = \mathbf{F}(s)\mathbf{E}(s)^{-1}\mathbf{u}, \quad \mathbf{u} = \mathbf{E}(s)\mathbf{F}(s)^{-1}\mathbf{y} \quad (1.205)$$

The hybrid forms are

$$\begin{pmatrix} f_1(s) \\ e_2(s) \end{pmatrix} = \begin{pmatrix} \frac{1}{Z_c} \tanh \Gamma & -\frac{1}{\cosh \Gamma} \\ \frac{1}{\cosh \Gamma} & Z_c \tanh \Gamma \end{pmatrix} \begin{pmatrix} e_1(s) \\ -f_2(s) \end{pmatrix} \quad (1.206)$$

$$\begin{pmatrix} e_1(s) \\ -f_2(s) \end{pmatrix} = \begin{pmatrix} Z_c \tanh \Gamma & \frac{1}{\cosh \Gamma} \\ -\frac{1}{\cosh \Gamma} & \frac{1}{Z_c} \tanh \Gamma \end{pmatrix} \begin{pmatrix} f_1(s) \\ e_2(s) \end{pmatrix} \quad (1.207)$$

### 1.6.5 Transfer function for terminated transmission line

Suppose that a termination in the form of an impedance  $Z_L(s)$  is used at port 2, which means that the effort at port 2 is given by

$$e_2 = Z_L f_2 \quad (1.208)$$

Insertion of the termination equation (1.208) in the transmission form (1.197) gives

$$\begin{pmatrix} e_1(s) \\ f_1(s) \end{pmatrix} = \begin{pmatrix} \cosh \Gamma & Z_c \sinh \Gamma \\ \frac{\sinh \Gamma}{Z_c} & \cosh \Gamma \end{pmatrix} \begin{pmatrix} 1 \\ \frac{1}{Z_L} \end{pmatrix} e_2(s) \quad (1.209)$$

and we find the transfer functions

$$\frac{e_2}{e_1}(s) = \frac{1}{\frac{Z_c}{Z_L} \sinh \Gamma + \cosh \Gamma} \quad (1.210)$$

$$\frac{e_2}{f_1}(s) = Z_c \frac{1}{\frac{Z_c}{Z_L} \cosh \Gamma + \sinh \Gamma} \quad (1.211)$$

which can be combined to give

$$\frac{f_1}{e_1}(s) = \frac{1}{Z_c} \frac{\frac{Z_c}{Z_L} \cosh \Gamma + \sinh \Gamma}{\frac{Z_c}{Z_L} \sinh \Gamma + \cosh \Gamma} \quad (1.212)$$

### 1.6.6 Wave variables

The system (1.187, 1.188) can be written

$$\frac{\partial}{\partial x} \begin{pmatrix} e(x, s) \\ f(x, s) \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & -Z(s) \\ -Y(s) & 0 \end{pmatrix}}_{\mathbf{A}} \begin{pmatrix} e(x, s) \\ f(x, s) \end{pmatrix} \quad (1.213)$$

This system can be made diagonal by finding the eigenvalues and the eigenvectors of the matrix  $\mathbf{A}$ . The eigenvalues of  $\mathbf{A}$  are found to be  $\pm \Gamma(s)/L$ , and the system is made diagonal with the following change of variables:

$$e(x, s) = \frac{1}{2} (a(x, s) + b(x, s)) \quad (1.214)$$

$$f(x, s) = \frac{1}{2Z_c(s)} (a(x, s) - b(x, s)) \quad (1.215)$$

which correspond to

$$a(x, s) = e(x, s) + Z_c(s)f(x, s) \quad (1.216)$$

$$b(x, s) = e(x, s) - Z_c(s)f(x, s) \quad (1.217)$$

This leads to

$$\frac{\partial a}{\partial x}(x, s) = -\frac{\Gamma(s)}{L} a(x, s) \quad (1.218)$$

$$\frac{\partial b}{\partial x}(x, s) = \frac{\Gamma(s)}{L} b(x, s) \quad (1.219)$$

The variables  $a$  and  $b$  are called the *wave variables* of the transmission line. The wave variables have the solutions

$$a(x, s) = a(0, s) \exp \left[ -\Gamma(s) \frac{x}{L} \right] \quad (1.220)$$

$$b(x, s) = b(L, s) \exp \left[ -\Gamma(s) \frac{(L-x)}{L} \right] \quad (1.221)$$

We define the wave variables at the end-points of the line to be  $a_1(s) = a(0, s)$ ,  $a_2 = a(L, s)$ ,  $b_1(s) = b(L, s)$  and  $b_2(s) = b(L, s)$ . Then we get

$$\frac{a_2(s)}{a_1(s)} = \exp [-\Gamma(s)] \quad (1.222)$$

$$\frac{b_1(s)}{b_2(s)} = \exp [-\Gamma(s)] \quad (1.223)$$

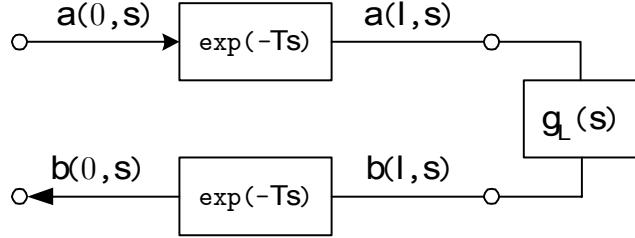


Figure 1.21: Scattering representation of lossless transmission line with load  $g_L(s)$  corresponding to load impedance  $z_L(s)$  at  $x = \ell$ .

### 1.6.7 Lossless transmission line

The transmission line is said to be *lossless* if the impedance  $Z(s)$  is purely inductive and the admittance  $Y(s)$  is purely capacitive. This means that there are real and positive constants  $M$  and  $C$  so that  $Z(s) = Ms$  and  $Y(s) = Cs$ . In this case the characteristic impedance is real and given by  $Z_c = Z_0 := \sqrt{M/C}$  while the propagation operator  $\Gamma(s)$  becomes

$$\Gamma(s) = L\sqrt{MC}s = Ts \quad (1.224)$$

where

$$c := \frac{1}{\sqrt{MC}} \quad (1.225)$$

is the wave propagation velocity and  $T = L/c$  is the propagation time. This gives the transfer functions

$$\frac{a_2(s)}{a_1(s)} = \exp(-Ts) \quad (1.226)$$

$$\frac{b_1(s)}{b_2(s)} = \exp(-Ts) \quad (1.227)$$

which are pure time delays of  $T = L/c$ . It is seen that the  $a(x, t)$  propagates in the positive direction with velocity  $c$ , while  $b(x, t)$  propagates in the negative direction with velocity  $-c$ . Because of this the variables  $a$  and  $b$  can be seen as wave variables.

$$\frac{b_1(s)}{a_1(s)} = \exp(-2Ts) G_L(s) \quad (1.228)$$

### 1.6.8 Line termination

Consider a transmission line with a load impedance  $Z_L$  connected at  $x = L$ . Then

$$e(L, s) = Z_L(s)f(L, s) \quad (1.229)$$

and

$$\frac{b_2(s)}{a_2(s)} = \frac{Z_L(s) - Z_c}{Z_L(s) + Z_c} =: G_L(s) \quad (1.230)$$

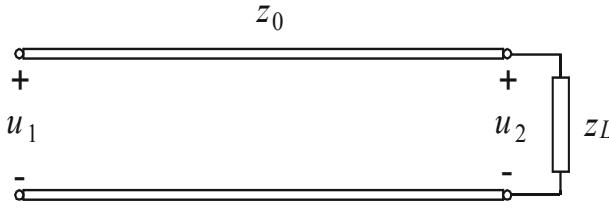


Figure 1.22: Electrical transmission line with load impedance  $Z_L(s)$  at  $x = \ell$ .

The boundary conditions in terms of the wave variables  $a$  and  $b$  are

$$a(0, s) = a_1(s) \quad (1.231)$$

$$b(L, s) = G_L(s)a(L, s) \quad (1.232)$$

The transfer function from  $a_1(s)$  to  $b_1(s)$  is given by

$$\frac{b_1(s)}{a_1(s)} = \frac{b_1(s)}{b_2(s)} \frac{b_2(s)}{a_2(s)} \frac{a_2(s)}{a_1(s)} = \exp(-2\Gamma(s)) G_L(s) \quad (1.233)$$

**Example 18** Consider a short circuit as the load, so that  $z_L = 0$ . Then  $g_L = -1$ , and

$$\frac{b_1(s)}{a_1(s)} = -\exp(-2\Gamma(s)) \quad (1.234)$$

which in the lossless case is a pure time delay of  $2T$  together with a change of sign. The effort transfer function is obviously

$$\frac{e(L, s)}{e(0, s)} = 0 \quad (1.235)$$

**Example 19** Consider an open circuit as the load. In this case  $i(L, s) = 0$ ,  $z_L = \infty$ , and  $g_L = 1$ . The transfer function becomes

$$\frac{b_1(s)}{a_1(s)} = \exp(-2\Gamma(s)) \quad (1.236)$$

which in the lossless case is a pure time delay of  $2T$ . The effort transfer function is

$$\frac{e(L, s)}{e(0, s)} = \frac{2}{[\exp(\Gamma(s)) + \exp(-\Gamma(s))]} = \frac{1}{\cosh(\Gamma(s))} \quad (1.237)$$

**Example 20** Impedance matching is achieved with  $z_L = Z_c$ . Then  $g_L = 0$ , and

$$\frac{b_1(s)}{a_1(s)} = 0 \quad (1.238)$$

which means that no wave is reflected. The voltage transfer function is

$$\frac{e(L, s)}{e(0, s)} = \exp(-\Gamma(s)) \quad (1.239)$$

which in the lossless case is a pure time delay.

**Example 21** An electrical transmission line is lossless if  $R = 0$  and  $G = 0$ . Then the series impedance is  $Z(s) = Ms$  is the series impedance, and the parallel admittance is  $Y(s) = Cs$ . In this case the characteristic impedance is real and given by  $Z_c = \sqrt{M/C}$  while the function  $\Gamma(s)$  becomes

$$\Gamma(s) = L\sqrt{MC}s = sT \quad (1.240)$$

Here we have used the wave propagation time  $T = L/c$ , and the wave propagation velocity

$$c := \frac{1}{\sqrt{MC}} \quad (1.241)$$

For a standard coaxial cable the wave velocity is 75 % of the speed of light, which gives  $c = 2.25 \times 10^8$  m/s.

# 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

$$y(s) = H(s)u(s) \quad (2.1)$$

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

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

will exist as

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

Moreover, we see that the Fourier transform is given by

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

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

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

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

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

with magnitude and phase given by

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

### 2.1.2 Second order oscillatory system

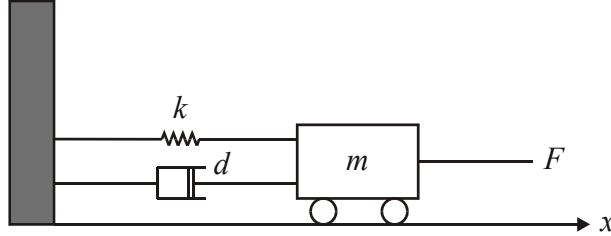


Figure 2.1: Mass-spring-damper system

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

$$m\ddot{x} + d\dot{x} + kx = F \quad (2.8)$$

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

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

where the *undamped natural frequency* is

$$\omega_0 = \sqrt{\frac{k}{m}} \quad (2.10)$$

and the relative damping is

$$\zeta = \frac{1}{2\omega_0} \frac{d}{m} = \frac{d}{2\sqrt{km}} \quad (2.11)$$

The transfer function is found by inserting

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

which leads to

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

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

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

The frequency response is

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

In particular we find that

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

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

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

which is inversely proportional to the relative damping  $\zeta$ .

### 2.1.3 Performance of a closed loop system

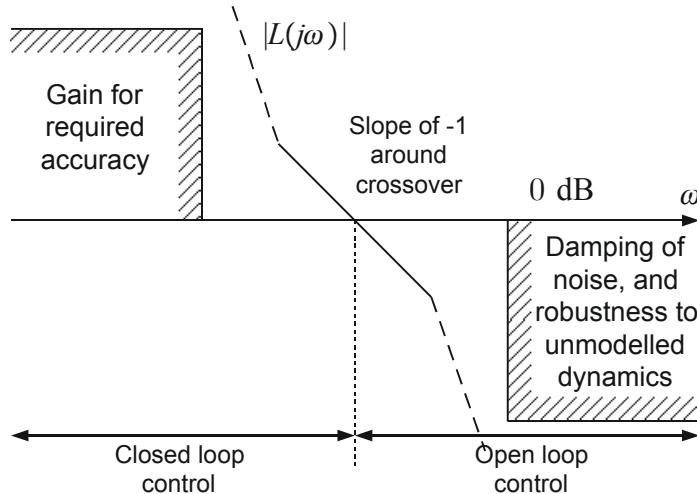


Figure 2.2: Performance requirements on the loop transfer function in a Bode diagram.

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

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

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

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

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  $\omega_c$ , 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

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

This can be expressed as conditions on the *gain margin*  $\Delta K$  and the *phase margin*  $\phi$  as

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

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

where  $\omega_c$  is the crossover frequency defined by  $|L(j\omega_c)| = 1$ , and  $\omega_{180}$  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

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

We also note that for any system we have

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

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

$$|y(j\omega)| = |T(j\omega)| |y_d(j\omega)| \quad (2.25)$$

we see that we will have an amplification of the desired value by a factor of 1.3 close to the crossover frequency  $\omega_c$ . 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:

$$T_m \frac{d\omega_m}{dt} = -\frac{R_a}{K} i_a \quad (2.26)$$

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

where  $\omega_m$  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

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

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

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

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

to be

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

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

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

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

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

which is obtained using the approximation

$$1 + T_a s \approx 1 \quad (2.34)$$

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

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

This give the approximated state-space model

$$T_m \frac{d\omega_m}{dt} = -\frac{R_a}{K} i_a \quad (2.36)$$

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

We can use the second equation to eliminate  $i$  using

$$i_a = -\frac{K}{R_a}\omega_m + \frac{1}{R_a}u_a \quad (2.38)$$

and the first equation becomes

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

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

$$\dot{x} = f(x, z, t, \epsilon) \quad (2.40)$$

$$\epsilon \dot{z} = g(x, z, t, \epsilon) \quad (2.41)$$

If  $\epsilon$  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

$$\dot{x} = f(x, z, t, 0) \quad (2.42)$$

$$0 = g(x, z, t, 0) \quad (2.43)$$

which is a differential-algebraic system.

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

$$z = z(x, t) \quad (2.44)$$

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

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

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

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

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

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

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

$$\dot{x}_1 = x_2 \quad (2.48)$$

$$\dot{x}_2 = -\omega_0^2 x_1 - 2\zeta\omega_0 x_2 \quad (2.49)$$

and the energy function

$$V = \frac{1}{2}\omega_0^2 x_1^2 + \frac{1}{2}x_2^2 \quad (2.50)$$

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

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

which gives

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

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

### 2.3.3 Second-order systems

If a system is given as a second order system

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

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

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

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

$$\ddot{x} = -\omega_0^2 x - 2\zeta\omega_0 \dot{x} \quad (2.55)$$

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

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

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

$$m\ddot{x} + dx + kx = 0 \quad (2.57)$$

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

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

The time derivative of the energy function is

$$\dot{V} = m\dot{x}\ddot{x} + kx\dot{x} \quad (2.59)$$

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

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

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

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

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

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

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

$$m\ddot{x} + d\dot{x} + kx = -F_f \quad (2.63)$$

Then the time derivative of the energy function is

$$\dot{V} = -F_f \dot{x} - d\dot{x}^2 \quad (2.64)$$

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

$$F_f \dot{x} \geq 0 \quad (2.65)$$

and, accordingly,

$$\dot{V} \leq -d\dot{x}^2 \quad (2.66)$$

### External force

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

$$m\ddot{x} + d\dot{x} + kx = F \quad (2.67)$$

The time derivative of the energy function is found to be

$$\dot{V} = F\dot{x} - d\dot{x}^2 \quad (2.68)$$

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

$$F = -K_d \dot{x} \quad (2.69)$$

will give

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

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

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}) \quad (2.71)$$

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

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

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

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

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  $\frac{1}{2}k_p x^2$ . 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

$$u = -k_p x - k_d \dot{x} \quad (2.74)$$

and get the closed loop dynamics

$$m\ddot{x} + k_d \dot{x} + k_p x = 0. \quad (2.75)$$

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:

$$V = \frac{1}{2}m\dot{x}^2 + \frac{1}{2}k_p x^2 \quad (2.76)$$

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

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

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

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

where

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

We see that whenever

$$\operatorname{Re}[\lambda_i(\mathbf{J} + \mathbf{J}^T)] < 0 \quad (2.80)$$

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

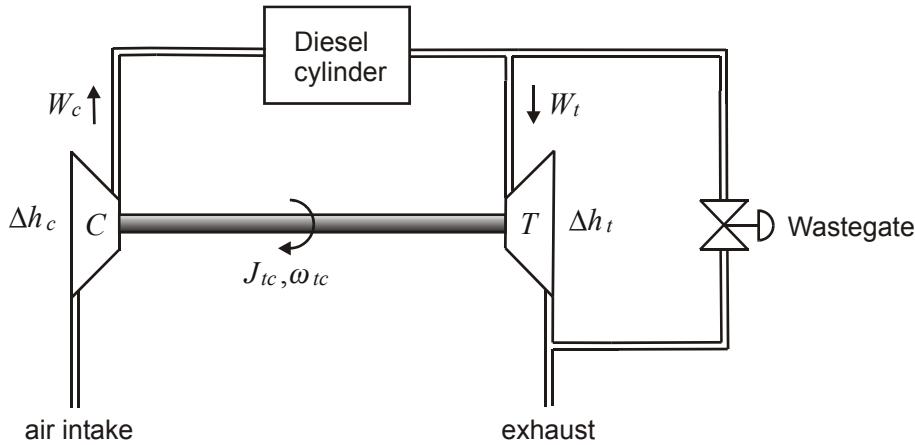


Figure 2.3: Diesel engine with turbocharger

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  $\tau$  is torque and  $\omega$  is angular velocity. The speed  $\omega_{tc}$  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

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

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,  $\Delta h_t$  is the reduction in specific enthalpy over the turbine,  $w_c$  is the mass flow through the compressor, and  $\Delta h_c$  is the increase in specific enthalpy over the compressor.

If the energy carried by the exhaust becomes too 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 guide vanes 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  $\Delta h_t$  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  $\Delta h_t$ . 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 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

$$\int_0^T y(t)u(t) dt \geq -E_0 \quad (2.82)$$

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

$$\dot{y} = -ay + u \quad (2.83)$$

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

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

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

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

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

$$m\ddot{x} + B\dot{x} + Kx = F \quad (2.86)$$

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

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

$$\geq -E_0 + B \int_0^T v^2 dt \quad (2.88)$$

where

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

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

$$\int_0^T F\dot{x}dt \geq -E_0 \quad (2.90)$$

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

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

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

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

where

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

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 taken 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  $\text{Re}[s] > 0$ .
2.  $H(s)$  is real for all positive and real  $s$ .
3.  $\text{Re}[H(s)] \geq 0$  for all  $\text{Re}[s] > 0$ .

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  $\text{Re}[s] > 0$ , 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

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

This implies that  $\text{Re}[H(j\omega)] \geq 0$  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.  $\text{Re } H(j\omega) \geq 0$  for all  $\omega$  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

$$\text{Res}_{s=j\omega_0}[H(s)] = \lim_{s \rightarrow j\omega_0} (s - j\omega_0)H(s) \quad (2.95)$$

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

$$R_\infty = \lim_{\omega \rightarrow \infty} \frac{H(j\omega)}{j\omega} \quad (2.96)$$

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

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

The frequency response is

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

and we see that

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

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

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

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

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

We find that if  $c \leq a + b$ , then  $\operatorname{Re}[h_2(j\omega)] > 0$  for all  $\omega$ , and the transfer function is positive real. If  $c > a + b$ , then  $h_2(s)$  is not positive real as  $\operatorname{Re}[h_2(j\omega)] < 0$  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  $\operatorname{Re}[H(j\omega)] = 0$ . The transfer function has only one pole, which is at infinity. As

$$R_\infty = \lim_{\omega \rightarrow \infty} \frac{j\omega L}{j\omega} = L \quad (2.102)$$

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

**Example 30** Consider the transfer function

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

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

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

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

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

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

**Example 31** Consider the transfer function

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\sinh(\sigma + j\omega) = \sinh \sigma \cos \omega + j \cosh \sigma \sin \omega \quad (2.113)$$

$$\cosh(\sigma + j\omega) = \cosh \sigma \cos \omega + j \sinh \sigma \sin \omega. \quad (2.114)$$

we find that

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

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

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

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

$$\text{Re}\left[\frac{1}{\cosh s}\right] = \frac{\cosh \sigma \cos \omega}{|\cosh s|^2} \quad (2.117)$$

It follows that for all  $\omega$  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

$$y(s) = H(s)u(s) \quad (2.118)$$

with rational and strictly proper transfer function

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

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

$$u(t) = U \sin \omega_0 t \quad (2.120)$$

Then the output is

$$y(t) = U |H(j\omega_0)| \sin [\omega_0 t + \angle H(j\omega_0)] + y_t(t) \quad (2.121)$$

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

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

Integration gives

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

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

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

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

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

for all  $T \geq 0$ . From this result it is seen that the system is passive if and only if  $\text{Re } H(j\omega) \geq 0$  for all  $\omega$ . The if part is obvious. Concerning the only if part, it is seen that if  $\text{Re } H(j\omega_0) < 0$  for some  $\omega_0$ , then there is no lower bound on  $\int_0^T y(t)u(t)dt$  as  $|U^2 T \text{Re } H(j\omega_0)/2|$  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

$$H(s) = \frac{\text{Res}_{s=\pm ja} H(s)}{s + ja} + \frac{\text{Res}_{s=\pm ja} H(s)}{s - ja} + G(s) \quad (2.126)$$

$$= 2 \frac{s \text{Res}_{s=\pm ja} H(s)}{s^2 + a^2} + G(s) \quad (2.127)$$

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

$$y(s) = H(s)u(s) = 2 \frac{s\omega_0 \text{Res}_{s=\pm ja} H(s)}{(s^2 + \omega_0^2)^2} + \dots \quad (2.128)$$

which corresponds to the time function

$$y(t) = t \sin(\omega_0 t) \operatorname{Res}_{s=\pm j\omega_0} H(s) + \dots \quad (2.129)$$

and it follows that

$$\int_0^T y(t) u(t) dt = \int_0^T t U \sin^2(\omega_0 t) \operatorname{Res}_{s=\pm j\omega_0} H(s) dt + \dots \quad (2.130)$$

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

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

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

$$y(s) = H(s)u(s) = \frac{U \operatorname{Res}_{s=\pm j\omega_0} H(s)}{s^2} + \dots \quad (2.132)$$

and the time function is

$$y(t) = t \operatorname{Res}_{s=0} H(s) + \dots \quad (2.133)$$

This gives

$$\int_0^T y(t) u(t) dt = \int_0^T t^2 \operatorname{Res}_{s=0} H(s) dt + \dots \quad (2.134)$$

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  $\operatorname{Res}_{s=j\omega_0}[H(s)]$  is real and positive.

#### 2.4.11 Bounded real transfer functions

**Definition 3** *The rational or irrational function  $B(s)$  is bounded real if*

1.  $B(s)$  is analytic for all  $\operatorname{Re}[s] > 0$ .
2.  $|B(s)| \leq 1$  for all positive and real  $s$

The transfer function

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

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

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

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

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

$$\operatorname{Re} H(s) = \frac{1}{2} [H(s) + H^*(s)] = \frac{1}{2} \frac{1 + B(s)}{1 - B(s)} + \frac{1}{2} \frac{1 + B^*(s)}{1 - B^*(s)} \quad (2.137)$$

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

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

**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 possible to conclude from the maximum modulus theorem that  $B(s) \neq -1$  in  $\operatorname{Re}[s] > 0$ , and we may express the inverse of  $H(s)$  as

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

which is analytic in  $\operatorname{Re}[s] > 0$  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

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

#### 2.4.12 Passivity of PID controllers

A PID controller

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

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

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

From this equation it is seen that the phase must satisfy

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

It follows that  $\operatorname{Re}[H(j\omega)] \geq 0$  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

$$\operatorname{Res}_{s=0} H_r(s) = \lim_{s \rightarrow 0} [s H_r(s)] = \frac{K}{T_i} \quad (2.144)$$

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

$$y(s) = H(s)u(s) \quad (2.145)$$

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

$$u(s) = G(s) [y_d(s) - y(s)] \quad (2.146)$$

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  $90^\circ$ . This implies that  $L(j\omega)$  has phase that is greater than  $-180^\circ$ . 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

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}) \quad (2.147)$$

$$\mathbf{y} = \mathbf{h}(\mathbf{x}) \quad (2.148)$$

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

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

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

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

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

$$m\ddot{x} + d\dot{x} + kx = F \quad (2.151)$$

The total energy

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

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

$$\dot{V} = F\dot{x} - d\dot{x}^2 \quad (2.153)$$

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

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

for some constant  $E_g \geq 0$ .

### 2.4.15 Interconnections of passive systems

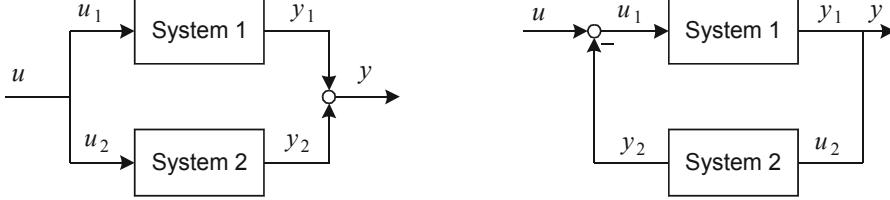


Figure 2.4: Parallel interconnection and feedback interconnection of two passive systems.

We now consider two systems

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

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

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

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

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

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

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

implies that the function  $V$  defined by

$$V := V_1 + V_2 \geq 0 \quad (2.160)$$

satisfies

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

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

A feedback interconnection

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

implies that

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

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

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

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

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

can be written in state-space form as

$$\dot{z} = \frac{e}{T_i} \quad (2.166)$$

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

Consider the nonnegative function

$$V_{pid} = \frac{1}{2} K T_i z^2 + \frac{1}{2} K T_d e^2 \quad (2.168)$$

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

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

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

$$\dot{V}_p = y u - g_p \quad (2.170)$$

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

$$V_{cl} = V_p + V_{pid} \quad (2.171)$$

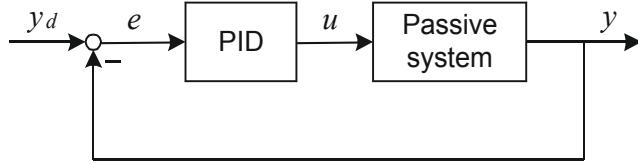


Figure 2.5: Passive plant with PID controller.

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

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

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

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

#### 2.4.18 Example: Control of mass-spring-damper system

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

$$\dot{V} = F\dot{x} - d\dot{x}^2 \quad (2.174)$$

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

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

Suppose that the PID controller

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

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

$$\dot{V}_{cl} = \dot{V} + \dot{V}_{pid} \quad (2.177)$$

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

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<sup>2</sup> 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

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

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,  $\mathbf{D}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}}$  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

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

is the kinetic energy, and

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

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

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

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

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

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

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

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

$$\dot{V} = ui - P_r \quad (2.185)$$

Integration of this equation gives

$$\int_0^T i(t) u(t) dt = V(T) - V(0) + \int_0^T P_r(t) dt \quad (2.186)$$

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

$$\int_0^T i(t) u(t) dt \geq -V(0) \quad (2.187)$$

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

$$i = \frac{1}{R}u + C\dot{u} \quad (2.188)$$

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

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

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

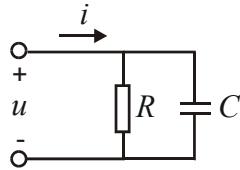


Figure 2.6: Passive electrical one-port.

#### 2.4.21 Electrical analog of PID controller

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

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

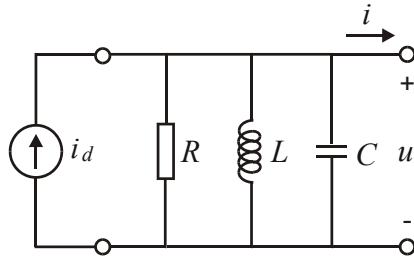


Figure 2.7: Electrical analog of PID controller for an electrical one-port where the current  $i$  is controlled with the voltage  $u$ .

where  $i_d$  is the desired current. This is an electrical one-port where a current source with current  $i_d$  is placed in parallel to a series connection with a resistor  $R = K(1 + T_d/T_i)$ , a capacitor  $C = T_i/K$ , and an inductor  $L = T_dK$  (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

$$P(t) = u_1(t)i_1(t) + u_2(t)i_2(t) \quad (2.191)$$

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

$$\int_0^T [u_1(t)i_1(t) + u_2(t)i_2(t)] dt \geq -V_2(0) \quad (2.192)$$

#### 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

$$u_2 = u, \quad i_2 = -i \quad (2.193)$$

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

$$\int_0^T [u_1(t)i_1(t) + u_2(t)i_2(t)] dt \geq -V_2(0) \quad (2.194)$$

$$\int_0^T i(t)u(t) dt \geq -V_1(0) \quad (2.195)$$

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

$$\int_0^T i_1(t)u_1(t) dt \geq -[V_1(0) + V_2(0)] \quad (2.196)$$

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

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

where  $\mathbf{u} = (u_1, \dots, u_n)^T$  and  $\mathbf{i} = (i_1, \dots, i_n)^T$ . 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

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

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

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

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. This 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

$$m_m \dot{v}_m = F_h - F_m \quad (2.200)$$

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

$$m_s \dot{v}_s = F_s - F_e \quad (2.201)$$

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

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

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

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

The physical interpretation of this is that the energy that is returned to the operator through the handle on the master is less than 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

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

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

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

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

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

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

$$m_s \dot{v}_s = F_s - F_e \quad (2.207)$$

$$F_s = K_s(x_m - x_s) + D_s(v_m - v_s) \quad (2.208)$$

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:

$$m_m \dot{v}_m = F_h - F_s \quad (2.209)$$

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  $D_s$ .

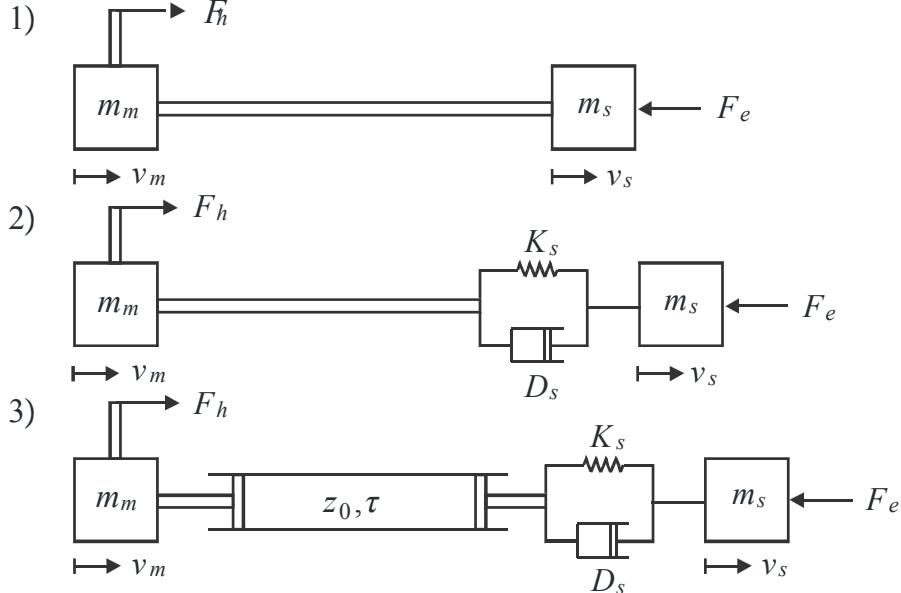


Figure 2.8: Mechanical analogs of telemanipulation systems with force reflection. The master is represented by a mass  $m_m$  that is moved with a force  $F_h$  from the human operator. The slave is represented by a mass  $m_s$  which is exposed to a force  $F_e$  from the environment. In 1) there is a direct mechanical coupling between the master and the slave. 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  $\tau$  in the electronic transmission between master and slave. Early attempts involved using the same solution as presented above, but with transmission given by

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

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 were 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  $\tau$  of the transmission line. We therefore introduce the wave variables

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

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

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

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

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

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

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

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

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

$$F_s = K_s(x_d - x_s) + D_s(v_d - v_s) \quad (2.216)$$

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

$$r = u + \lambda y \quad (2.217)$$

Then

$$\int_0^T r^2 dt = \int_0^T u^2 dt + 2\lambda \int_0^T uy dt + \lambda^2 \int_0^T y^2 dt \quad (2.218)$$

From this equation it is seen that

$$\int_0^T uy dt \geq -E_0 \quad (2.219)$$

is equivalent for all  $\lambda > 0$  to

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

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*

$$\mathbf{r} = \boldsymbol{\omega} + \lambda \boldsymbol{\epsilon} \quad (2.221)$$

*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*

$$\mathbf{r} = \dot{\mathbf{q}} + \lambda \mathbf{q} \quad (2.222)$$

*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

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, u, t) \quad (2.223)$$

$$y = h(\mathbf{x}, t) \quad (2.224)$$

or, in the case of linear models,

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u \quad (2.225)$$

$$y = \mathbf{C}\mathbf{x} + \mathbf{D}u \quad (2.226)$$

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.

### 2.5.2 Exact kinematic models

Some types of models or parts of models are exact. In particular this applies to kinematics, which is the geometric description of motion. Kinematics, which is discussed in great detail in Chapter 6, play an important role in the models describing the motion of rigid bodies such as planes, ships, robots, cars and mechanisms. The main role of kinematics in this context is in the kinematic differential equations that give the time derivative of the configuration as a function of velocity and angular velocity. It may be of great importance that the kinematic part of a model is exact because the kinematics are often nonlinear, and even complicated. Therefore it may be important to keep in mind which part of the model that is kinematic and therefore exact.

**Example 39** *In inertial navigation the position of a plane is calculated from inertial sensors (Titterton and Weston 1997). The inertial sensors are gyroscopes, which measure the angular velocity of the plane, and accelerometers, which measure the acceleration. The velocity and position are calculated by integrating the acceleration measurements, while the angular velocity measurements are used to calculate the direction of the axes of the accelerometers. The direction of the accelerometers are usually described in terms of Euler parameters, also known as quaternions. When the angular velocity vector  $\omega^b$  of the plane is given, the Euler parameters may be calculated by numerical integration of the kinematic differential equations*

$$\dot{\eta} = -\frac{1}{2}\epsilon^T \omega^b \quad (2.227)$$

$$\dot{\epsilon} = \frac{1}{2}(\eta \mathbf{I} + \epsilon^\times) \omega^b \quad (2.228)$$

where  $\epsilon^\times \omega^b$  denotes the vector cross product between  $\epsilon$  and  $\omega^b$ . The rotation matrix is  $\mathbf{R}_{\eta, \epsilon} = \mathbf{I} + 2\eta\epsilon^\times + 2\epsilon^\times\epsilon^\times$ , and the acceleration in a star-fixed coordinate frame is

$$\mathbf{a} = (\mathbf{I} + 2\eta\epsilon^\times + 2\epsilon^\times\epsilon^\times) \mathbf{a}^b \quad (2.229)$$

where  $\mathbf{a}^b$  are the measured accelerations. We see that the equations for calculating the acceleration  $\mathbf{a}$  are relatively complicated and highly nonlinear. However, it is interesting to note that there are no approximations or uncertainty involved in the development of this model.

In addition, kinematics play an important role in fluid mechanics, where the concepts of divergence and curl are kinematic, and moreover, the Laplace equation

$$\nabla^2 \phi = 0 \quad (2.230)$$

in potential flow is purely kinematic. Also Reynolds' transport theorem is a kinematic result.

### 2.5.3 Balance equations

Balance equations are based on the physical phenomenon that certain properties like mass, momentum and energy are conserved. Therefore, the balance equations are exact models from the outset. As an example of this, the Cauchy equation of motion for a fluid is exact. This equation, which is discussed in detail in Section 11.2.1, is written

$$\rho \frac{Dv_i}{Dt} = T_{ji,j} + \rho f_i \quad (2.231)$$

where  $\rho$  is density,  $v_i$  is the velocity in direction  $i$ ,  $T_{ji,j}$  is the gradient of the stress tensor, and  $f_i$  is the body force. The modeling assumptions enter in the constitutive equations where the assumptions on the functional dependence of the stress tensor on fluid motion enters. The mass balance is important in many physical systems, and perhaps even more important is the energy balance. The model

$$\begin{aligned} \frac{d}{dt} [\text{total energy in a } V] &= [\text{energy flow into } V] \\ &\quad - [\text{energy flow out of } V] \\ &\quad + [\text{energy generated in } V] \end{aligned} \quad (2.232)$$

where  $V$  is a fixed volume, is exact, and is not influenced by changes in parameter values. Such insight may be especially interesting in relation to energy-based methods for controller design.

#### 2.5.4 Passivity

In several important applications the plant may have passivity as a physical property for a given combination of control variable and output variable. This means that the plant will be passive when plant parameters undergo large variations in numerical values, and even under changes in model order. Then with a properly selected PID controller the closed loop system will always be stable although there may be a high degree of uncertainty in numerical parameters and detailed modeling.

**Example 40** A DC motor is used to control a joint in a robot arm (Slotine 1991), (Lozano et al. 2000). The DC motor is current controlled, and delivers a specified torque  $u$ . The robot arm has several mechanical resonances, and it may pick up loads of unknown inertia. Moreover, the robot may come into contact with a rigid environment. However, in spite of all the uncertainty, the arm seen from the motor will be passive because

$$\frac{d}{dt} [\text{total energy in robot}] = \dot{q}u - [\text{energy dissipated by friction}] \quad (2.233)$$

where  $\dot{q}$  is the shaft velocity of the robot joint, and  $u$  is the torque from the DC motor. Therefore, the closed loop system is stable as long as the controller is passive. This is the case if a PI controller from  $\dot{q}$  is used, which is the same as a PD controller from the motor angle  $q$ .

## **Part II**

# **Motors and actuators**



# 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  $\omega_m$ , and it has some device for setting up a motor torque  $T$  so that the motor shaft has the following equation of motion:

$$J_m \dot{\omega}_m = T - T_L \quad (3.1)$$

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

$$P_m = T \omega_m \quad (3.2)$$

while the mechanical power delivered to the load is

$$P_L = T_L \omega_m \quad (3.3)$$

The motor shaft dynamics can be described as a two-port with effort  $T$  and flow  $\omega_m$  at the input port, and effort  $T_L$  and flow  $\omega_m$  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 or 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

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

### 3.2.3 Gear model

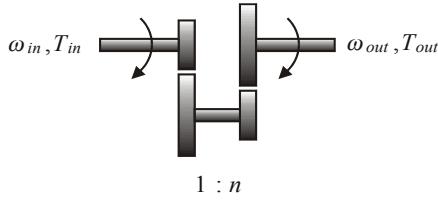


Figure 3.1: Reduction gear

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

$$\omega_{out} = n \omega_{in} \quad (3.5)$$

where  $\omega_{in}$  is the angular velocity of the shaft on the input side of the gear, and  $\omega_{out}$  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,

$$T_{in}\omega_{in} = T_{out}\omega_{out} \quad (3.6)$$

Inserting the expression for  $\omega_{out}$  we find that

$$T_{out} = \frac{1}{n}T_{in} \quad (3.7)$$

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

$$\omega_{out} = n\omega_{in} \quad (3.8)$$

$$T_{out} = \frac{1}{n}T_{in} \quad (3.9)$$

with variables  $T_{in}$  and  $\omega_{in}$  at the input port, and variables  $T_{out}$  and  $\omega_{out}$  at the output port.

### 3.2.4 Motor and gear

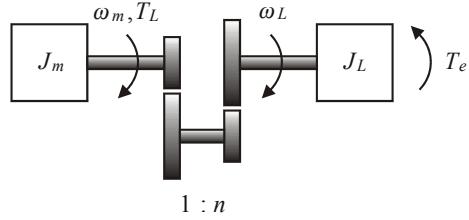


Figure 3.2: Motor and gear.

Consider a motor with equation of motion

$$J_m\dot{\omega}_m = T - T_L \quad (3.10)$$

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

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

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

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

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

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

### 3.2.5 Transformation of rotation to translation

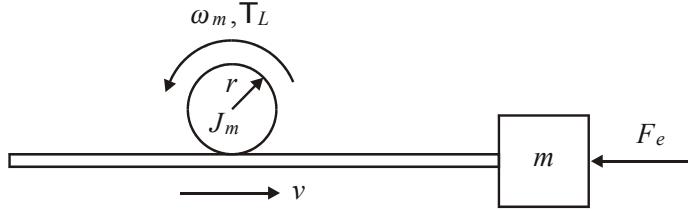


Figure 3.3: Transmission from rotation to translation.

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  $\omega_m$ , 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

$$v = r\omega_m \quad (3.14)$$

$$F = \frac{1}{r} T_L \quad (3.15)$$

with variables  $T_L$  and  $\omega_m$  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

$$m\dot{v} = F - F_e \quad (3.16)$$

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

$$J_m \dot{\omega}_m = T - T_L \quad (3.17)$$

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

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

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

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

### 3.2.6 Torque characteristics

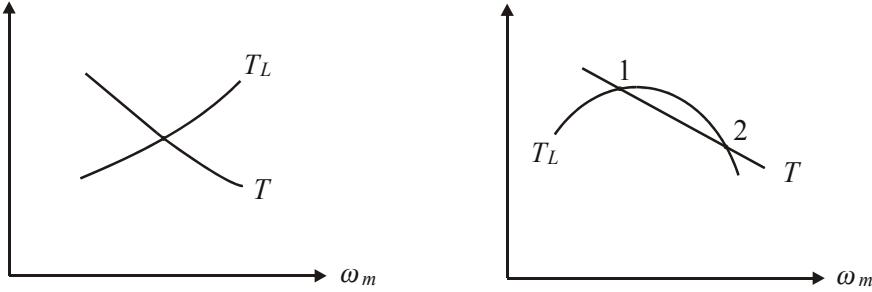


Figure 3.4: To the left is shown a stable system where the load torque  $T_L$  is increasing for increasing motor velocity  $\omega_m$ . To the right is shown a system with two equilibrium points. Equilibrium 1 is stable, while equilibrium 2 is unstable as the load torque  $T_L$  decreases faster than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

$$J_m\Delta\dot{\omega}_m = k\Delta\omega_m \quad (3.20)$$

where

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

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

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

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

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

The constant  $\omega_s$  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

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

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

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

This shows that the system is unstable for constant motor torque  $T$  at the speed  $\omega_m$  if

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

### 3.2.7 The four quadrants of the motor

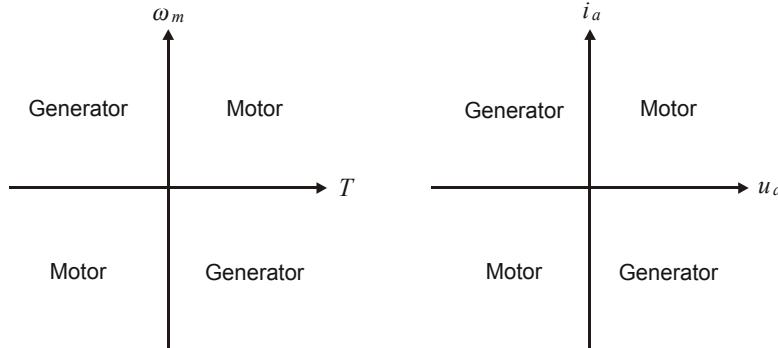


Figure 3.5: The four quadrants of the motor (to the left), and the four quadrants of the power amplifier (to the right).

A motor delivers the mechanical power  $T\omega_m$  through the motor shaft, where  $T$  is the motor torque and  $\omega_m$  is the motor speed. If  $T$  and  $\omega_m$  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  $\omega_m$  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

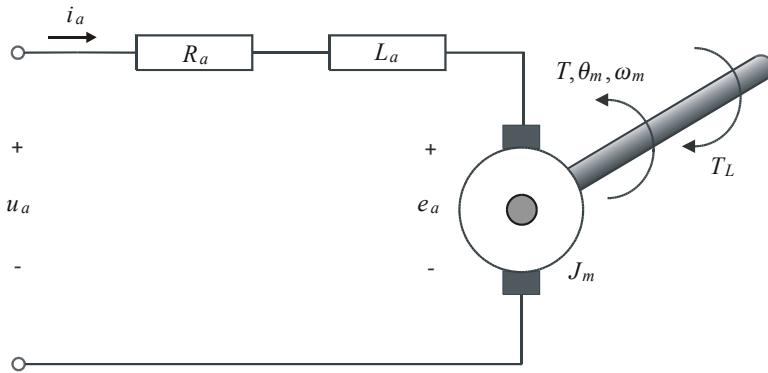


Figure 3.6: Armature circuit of DC motor with constant field.

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  $\omega_m$  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

$$T = K_T i_a \quad (3.27)$$

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  $\omega_m$ . Finally, the third two-port is the motor shaft with input port with variables  $T$  and  $\omega_m$  and output port with variables  $T_L$  and  $\theta_m$ . 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

$$e_a i_a = T \omega_m = K_T i_a \omega_m \Rightarrow e_a = K_E \omega_m \quad (3.28)$$

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

$$L_a \frac{d}{dt} i_a = -R_a i_a - K_E \omega_m + u_a \quad (3.29)$$

$$J_m \dot{\omega}_m = K_T i_a - T_L \quad (3.30)$$

$$\dot{\theta}_m = \omega_m \quad (3.31)$$

The block diagram is shown in Figure 3.7.

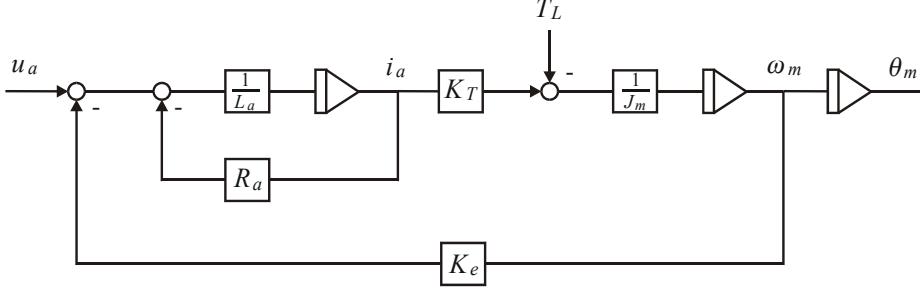


Figure 3.7: Voltage controlled DC motor.

### 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

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

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

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

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

$$\dot{V}_L \leq \omega_m T_L \quad (3.34)$$

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

$$\dot{V}_t \leq i_a u_a \quad (3.35)$$

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  $\omega_m$  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  $\omega_m$  and output  $T_L$  is passive with storage function  $V_L = 0$  as this gives

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

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  $\theta_L$  connected to the motor shaft by a spring with torque

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

The equation of motion for the inertia is

$$J_L \ddot{\theta}_L = T_L \quad (3.38)$$

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

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

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

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

### 3.3.4 Laplace transformed model

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

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

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

$$s \theta_m(s) = \omega_m(s) \quad (3.43)$$

The equation of motion gives

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

while the armature equation gives

$$(L_a s + R_a)i_a(s) = -K_E s \theta_m(s) + u_a(s) \quad (3.45)$$

Insertion of (3.45) in (3.44) gives

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

and finally

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

This can be written

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

where

$$T_a = \frac{L_a}{R_a} \quad (3.49)$$

is the electrical time constant of the motor, and

$$T_m = \frac{J_m R_a}{K_E K_T} \quad (3.50)$$

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

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

This leads to the following result:

The transfer function from the input  $u_a$  to the angle  $\theta_m$  is

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

## 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 be 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

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

From the Laplace transformed model we see that

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

and we find that

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

where it is used that  $K_E = K_T$ . The following current controller is used:

$$u_a = K_i (i_d - i_a) \quad (3.56)$$

This gives the closed loop dynamics

$$\frac{i_a}{i_d}(s) = \frac{K_i H_a(s)}{1 + K_i H_a(s)} \quad (3.57)$$

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

$$i_a(s) = i_d(s) \quad (3.58)$$

Insertion in (3.44) gives the following result:

The model of a current controlled DC motor is given by the double integrator model

$$\theta_m(s) = \frac{1}{J_m s^2} [K_T i_d(s) - T_L(s)] \quad (3.59)$$

where the input is the desired current  $i_d$ .

The block diagram is shown in Figure 3.8.

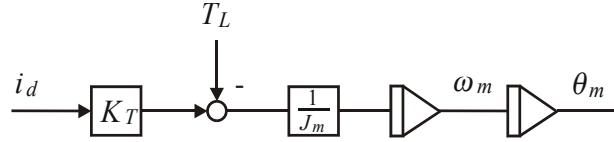


Figure 3.8: Current controlled DC motor.

**Example 44** Consider a PI controller

$$u_a = K_i T_i \frac{1 + T_i s}{T_i s} (i_0 - i_a) \quad (3.60)$$

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

$$L(s) = K_i \frac{J_m}{K_E K_T} \frac{1}{1 + T_a s} \quad (3.61)$$

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  $\omega_m$  of the motor satisfies  $s\theta_m = \omega_m$ , and it follows that

$$\frac{\omega_m}{i_d}(s) = \frac{\omega_m}{\theta_m}(s) \frac{\theta_m}{i_d}(s) = \frac{K_T}{J_m s} \quad (3.62)$$

This transfer function is the product of a gain  $K_T/J_m$  and an integrator  $1/s$ . The velocity controller

$$i_d = K_\omega (\omega_d - \omega_m) \quad (3.63)$$

gives the closed-loop dynamics

$$\frac{\omega_m}{\omega_d}(s) = \frac{\frac{K_a}{s}}{1 + \frac{K_a}{s}} = \frac{1}{1 + \frac{s}{K_a}} \quad (3.64)$$

where

$$K_a = \frac{K_T K_\omega}{J_m} \quad (3.65)$$

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  $\omega_m$  to the angle  $\theta_m$  is an integrator, which leads to

$$\frac{\theta_m}{\omega_d}(s) = \frac{1}{s \left( 1 + \frac{s}{K_a} \right)} \quad (3.66)$$

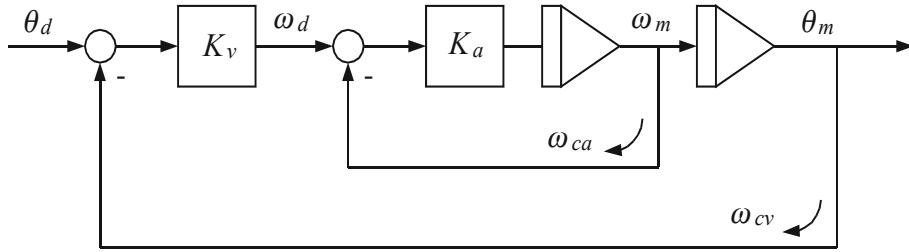


Figure 3.9: Current controlled DC motor with velocity loop and position loop. The crossover frequency of the velocity loop is  $\omega_{ca} = K_a$ , and the crossover frequency of the position loop is  $\omega_{cv} = K_v$  provided that  $K_a \gg K_v$ .

The position controller

$$\omega_d = K_v (\theta_d - \theta_m) \quad (3.67)$$

where  $K_v$  is the velocity constant gives the closed-loop dynamics

$$\frac{\theta_m}{\theta_d} (s) = \frac{1}{1 + \frac{1}{K_v}s + \frac{1}{K_v K_a}s^2} \quad (3.68)$$

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

$$\frac{\theta_m}{\theta_d} (s) = \frac{1}{\left(1 + \frac{s}{K_v}\right)\left(1 + \frac{s}{K_a}\right)} \quad (3.69)$$

## 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

$$J_m \ddot{\theta}_m = T_m - T_L \quad (3.70)$$

$$J_L \ddot{\theta}_L = T_L \quad (3.71)$$

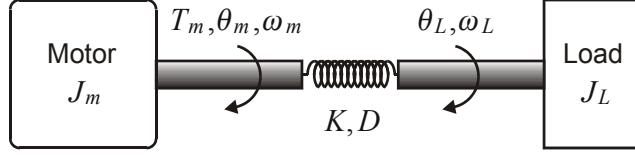


Figure 3.10: Motor with elastic transmission

where  $\theta_m$  is the motor shaft angle,  $\theta_L$  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

$$T_L = -K\theta_e - D\dot{\theta}_e \quad (3.72)$$

where

$$\theta_e = \theta_L - \theta_m \quad (3.73)$$

is the elastic deflection of the the transmission. In the derivation of the transfer functions it is helpful to introduce the variable

$$\theta_r = \theta_m + \frac{J_L}{J_m}\theta_L. \quad (3.74)$$

and derive the model in terms of the variables  $\theta_e$  and  $\theta_r$ .

The equations of motion for  $\theta_e$  and  $\theta_r$  are found by combining the equations (3.70–3.72). This gives

$$\ddot{\theta}_e + \frac{D}{J_e}\dot{\theta}_e + \frac{K}{J_e}\theta_e = -\frac{1}{J_m}T_m \quad (3.75)$$

$$\ddot{\theta}_r = \frac{T_m}{J_m} \quad (3.76)$$

where

$$J_e = \frac{J_m J_L}{J}, \quad J = J_m + J_L \quad (3.77)$$

### 3.5.3 Transfer functions

We see from (3.75) and (3.76) that the dynamic model of the elastic deflection  $\theta_e$  is a second order oscillatory system, while the rigid motion  $\theta_r$  results from a double integrator from the motor torque  $T_m$ . The transfer functions from the input  $T_m$  to  $\theta_e$  and  $\theta_r$  are found to given by

$$\frac{\theta_e}{T_m}(s) = -\frac{1}{J_m} \frac{\left(\frac{1}{\omega_1}\right)^2}{1 + 2\zeta_1 \frac{s}{\omega_1} + \left(\frac{s}{\omega_1}\right)^2} \quad (3.78)$$

$$\frac{\theta_r}{T_m}(s) = \frac{1}{J_m s^2} \quad (3.79)$$

where

$$\omega_1 = \sqrt{\frac{K}{J_e}} \quad \text{and} \quad \zeta_1 = \frac{D}{2} \frac{1}{\sqrt{J_e K}} \quad (3.80)$$

The transfer functions for the original variables are found by solving (3.73) and (3.74), which gives

$$\theta_m = \frac{J_m}{J} \left( \theta_r - \frac{J_L}{J_m} \theta_e \right) \quad (3.81)$$

$$\theta_L = \frac{J_m}{J} (\theta_r + \theta_e) \quad (3.82)$$

This gives

$$\begin{aligned} \frac{\theta_m}{T_m}(s) &= \frac{J_m}{J} \left[ \frac{\theta_r}{T_m}(s) - \frac{J_L}{J_m} \frac{\theta_e}{T_m}(s) \right] \\ &= \frac{1}{J} \left[ \frac{1}{s^2} + \frac{\frac{J_L}{J_m} \left( \frac{1}{\omega_1} \right)^2}{1 + 2\zeta_1 \frac{s}{\omega_1} + \left( \frac{s}{\omega_1} \right)^2} \right] \end{aligned} \quad (3.83)$$

and

$$\begin{aligned} \frac{\theta_L}{T_m}(s) &= \frac{J_m}{J} \left[ \frac{\theta_r}{T_m}(s) + \frac{\theta_e}{T_m}(s) \right] \\ &= \frac{J_m}{J} \left[ \frac{1}{J_m s^2} - \frac{1}{J_m} \frac{\left( \frac{1}{\omega_1} \right)^2}{1 + 2\zeta_1 \frac{s}{\omega_1} + \left( \frac{s}{\omega_1} \right)^2} \right] \end{aligned} \quad (3.84)$$

After some work the following result is found:

The motor and elastic load with elastic transmission is described by the two transfer functions

$$H_{\theta m}(s) : = \frac{\theta_m}{T_m}(s) = \frac{1}{J s^2} \frac{1 + 2\zeta_a \frac{s}{\omega_a} + \left( \frac{s}{\omega_a} \right)^2}{1 + 2\zeta_1 \frac{s}{\omega_1} + \left( \frac{s}{\omega_1} \right)^2} \quad (3.85)$$

$$H_{\theta L}(s) : = \frac{\theta_L}{T_m}(s) = \frac{1}{J s^2} \frac{1 + 2\zeta_1 \frac{s}{\omega_1}}{1 + 2\zeta_1 \frac{s}{\omega_1} + \left( \frac{s}{\omega_1} \right)^2} \quad (3.86)$$

where the parameters are given by

$$\zeta_1 = \frac{D}{2} \frac{1}{\sqrt{J_e K}}, \quad \omega_1 = \sqrt{\frac{K}{J_e}} \quad (3.87)$$

$$\zeta_a = \sqrt{\frac{J_m}{J}} \zeta_1, \quad \omega_a = \sqrt{\frac{J_m}{J}} \omega_1 < \omega_1 \quad (3.88)$$

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

$$H_{\omega m}(j\omega) : = \frac{\omega_m}{T_m}(s) = \frac{1}{Js} \frac{1 + 2\zeta_a \frac{s}{\omega_a} + \left(\frac{s}{\omega_a}\right)^2}{1 + 2\zeta_1 \frac{s}{\omega_1} + \left(\frac{s}{\omega_1}\right)^2} \quad (3.89)$$

$$H_{\omega L}(j\omega) : = \frac{\omega_L}{T_m}(s) = \frac{1}{Js} \frac{1 + 2\zeta_1 \frac{s}{\omega_1}}{1 + 2\zeta_1 \frac{s}{\omega_1} + \left(\frac{s}{\omega_1}\right)^2} \quad (3.90)$$

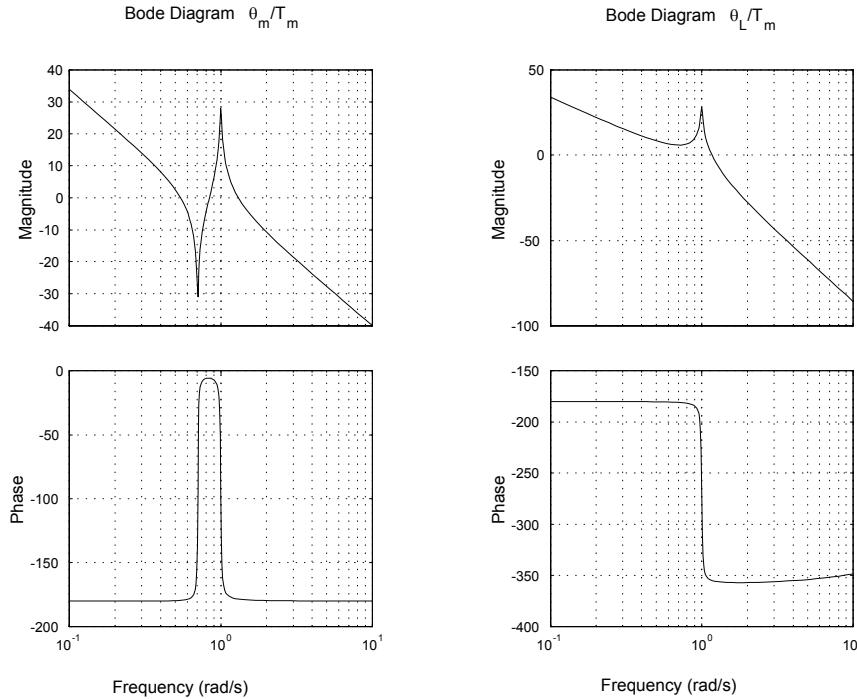


Figure 3.11: Frequency response from the motor torque  $T_m$  to the motor angle  $\theta_m$  (to the left), and frequency response from the motor torque  $T_m$  to the load angle  $\theta_L$  (to the right).

An important observation is that  $\omega_a < \omega_1$ , which means that the break frequency of the zeros in  $H_{\theta m}(j\omega)$  is smaller than the break frequency of the poles in  $H_{\theta L}(j\omega)$ . 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  $180^\circ$  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  $\omega_1$ . In contrast to this, the crossover may be selected above  $\omega_1$  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  $\omega_1$ .

### 3.5.4 Zeros of the transfer function

The zeros of the transfer function  $H_{\theta m}(s)$  are the roots of

$$1 + 2\zeta_a \frac{s}{\omega_a} + \left(\frac{s}{\omega_a}\right)^2 = 0 \quad (3.91)$$

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  $\omega_a$  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

$$\frac{\theta_L}{\theta_m}(s) = \frac{\theta_L}{T_m}(s) \frac{T_m}{\theta_m}(s) = \frac{1 + 2\zeta_1 \frac{s}{\omega_1}}{1 + 2\zeta_a \frac{s}{\omega_a} + \left(\frac{s}{\omega_a}\right)^2} \quad (3.92)$$

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

$$V = \frac{1}{2}J_m\omega_m^2 + \frac{1}{2}J_L\omega_L^2 + \frac{1}{2}K\theta_e^2 \quad (3.93)$$

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

$$\dot{V} = \dot{\theta}_m T_m - D\dot{\theta}_e^2 \quad (3.94)$$

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

$$\text{Re}[H_{\omega m}(j\omega)] \geq 0 \quad \text{for all } \omega. \quad (3.95)$$

Thus, from simple energy arguments we can establish that

$$|\angle H_{\omega m}(j\omega)| \leq 90^\circ \quad (3.96)$$

which implies that

$$-180^\circ \leq \angle H_{\theta m}(j\omega) \leq 0^\circ. \quad (3.97)$$

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

$$J_L \dot{\omega}_L = T_L - T_1 \quad (3.98)$$

and by adding the mechanical two-port

$$J_2 \dot{\omega}_2 = T_1 - T_2 \quad (3.99)$$

$$\frac{d}{dt} (\theta_1 - \theta_2) = (\omega_1 - \omega_2) \quad (3.100)$$

$$T_1 = D_1 (\omega_1 - \omega_2) + K_1 (\theta_1 - \theta_2). \quad (3.101)$$

where  $\omega_2$  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  $\omega_1$ , while the output port has effort  $T_2$  and flow  $\omega_2$ . We may add on any number of additional degrees of freedom as two-ports

$$J_i \dot{\omega}_i = T_{i-1} - T_i \quad (3.102)$$

$$\frac{d}{dt} (\theta_{i-1} - \theta_i) = (\omega_{i-1} - \omega_i) \quad (3.103)$$

$$T_{i-1} = D_{i-1} (\omega_{i-1} - \omega_i) + K_{i-1} (\theta_{i-1} - \theta_i) \quad (3.104)$$

with port variables  $T_{i-1}$  and  $\omega_{i-1}$  at the input and  $T_i$  and  $\omega_i$  at the output. In a computational setting the inputs are  $\omega_{i-1}$  and  $T_i$ , while the outputs are  $\omega_i$  and  $T_{i-1}$ .

The sum of kinetic and potential energy for a motor with  $n$  degrees of freedom in the load is

$$V = \frac{1}{2} J_m \omega_m^2 + \sum_{i=1}^n \frac{1}{2} J_i \omega_i^2 + \frac{1}{2} K \theta_e^2 + \sum_{i=1}^{n-1} \frac{1}{2} K_i (\theta_i - \theta_{i+1})^2 \quad (3.105)$$

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

$$\dot{V} = \omega_m T_m - D \dot{\theta}_e^2 - \sum_{i=1}^{n-1} D_i (\omega_i - \omega_{i+1})^2 \quad (3.106)$$

This implies that the system with input  $T_m$  and output  $\omega_m$  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  $\theta_L$  that is driven by two motors. Motor 1 has shaft angle  $\theta_1$ , inertia  $J_1$  and motor torque  $T_{m1}$ , while motor 2 has shaft angle  $\theta_2$ , 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

$$J_1 \ddot{\theta}_1 = T_{m1} - T_{g1} \quad (3.107)$$

$$J_2 \ddot{\theta}_2 = T_{m2} - T_{g2} \quad (3.108)$$

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

$$J_L \ddot{\theta}_L = \frac{1}{n} (T_{g1} + T_{g2}) - T_e \quad (3.109)$$

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

$$T_{g1} = K_1\phi_1 + D_1\dot{\phi}_1, \quad T_{g2} = K_2\phi_2 + D_2\dot{\phi}_2 \quad (3.110)$$

### 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

$$V = \frac{1}{2} \left( J_1\dot{\theta}_1^2 + J_2\dot{\theta}_2^2 + J_L\dot{\theta}_L^2 \right) + \frac{1}{2} (K_1\phi_1^2 + K_2\phi_2^2) \geq 0 \quad (3.111)$$

The time derivative of the energy will be the power supplied by the motor torques minus the power dissipated in the dampers. This gives

$$\dot{V} = T_{m1}\dot{\theta}_1 + T_{m2}\dot{\theta}_2 - D_1\dot{\phi}_1^2 - D_2\dot{\phi}_2^2 \quad (3.112)$$

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

$$T_{m2} = -K_{p2}\theta_2 - K_{d2}\dot{\theta}_2 \quad (3.113)$$

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

$$V_a = \frac{1}{2} \left( J_1\dot{\theta}_1^2 + J_2\dot{\theta}_2^2 + J_L\dot{\theta}_L^2 \right) + \frac{1}{2} (K_1\phi_1^2 + K_2\phi_2^2 + K_{p2}\theta_2^2) \geq 0 \quad (3.114)$$

which will have time derivative along the solutions of the system given by

$$\dot{V}_a = T_{m1}\dot{\theta}_1 - D_1\dot{\phi}_1^2 - D_2\dot{\phi}_2^2 - K_{d2}\dot{\theta}_2^2 \quad (3.115)$$

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

$$J_m \ddot{\theta}_m = T_m - T_{gm} \quad (3.116)$$

$$J_L \ddot{\theta}_L = T_{gL} \quad (3.117)$$

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

$$\phi = \theta_m - \frac{1}{n} \theta_L \quad (3.118)$$

The gear has a deadzone  $\delta$ . 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  $\phi$  according to

$$T_{gm}(\phi) = \begin{cases} K(\phi + \delta), & \phi \leq -\delta \\ 0, & -\delta \leq \phi \leq \delta \\ K(\phi - \delta), & \delta \leq \phi \end{cases}, \quad T_{gL}(\phi) = \frac{1}{n} T_{gm}(\phi) \quad (3.119)$$

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

$$\left. \begin{aligned} T_{gm}(\phi) &= 0 \\ T_{gL}(\phi) &= 0 \end{aligned} \right\}, \quad |\phi| < \delta \quad (3.120)$$

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

$$\left. \begin{aligned} \dot{\theta}_L &= n \dot{\theta}_m \\ T_{gm} &= n T_{gL} \end{aligned} \right\}, \quad |\phi| = \delta \quad (3.121)$$

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 system 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

$$\left. \begin{array}{l} (J_m + n^2 J_L) \ddot{\theta}_m = T_m \\ \theta_L = n(\theta_m + \delta) \end{array} \right\} \quad \text{when } \phi = -\delta \text{ and } T_m < 0 \quad (3.122)$$

$$\left. \begin{array}{l} (J_m + n^2 J_L) \ddot{\theta}_m = T_m \\ \theta_L = n(\theta_m - \delta) \end{array} \right\} \quad \text{when } \phi = \delta \text{ and } T_m > 0 \quad (3.123)$$

$$\left. \begin{array}{l} J_m \ddot{\theta}_m = T_m \\ J_L \ddot{\theta}_L = 0 \end{array} \right\} \quad \text{otherwise} \quad (3.124)$$

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

$$J_1 \ddot{\theta}_1 = T_{m1} - T_{g1} \quad (3.125)$$

$$J_2 \ddot{\theta}_2 = T_{m2} - T_{g2} \quad (3.126)$$

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

$$J_L \ddot{\theta}_L = \frac{1}{n}(T_{g1} + T_{g2}) - T_e \quad (3.127)$$

where  $T_e$  is an external disturbance torque.

The gears are supposed to have a spring constant  $K$  and a deadzone  $\delta$ . 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

$$T_{gi} = \begin{cases} K(\phi_i + \delta), & \phi_i \leq -\delta \\ 0, & -\delta \leq \phi_i \leq \delta \\ K(\phi_i - \delta), & \delta \leq \phi_i \end{cases}, \quad i = 1, 2 \quad (3.128)$$

The motors can then be connected to the load with the gear equations. The system is shown in Figure 3.12.

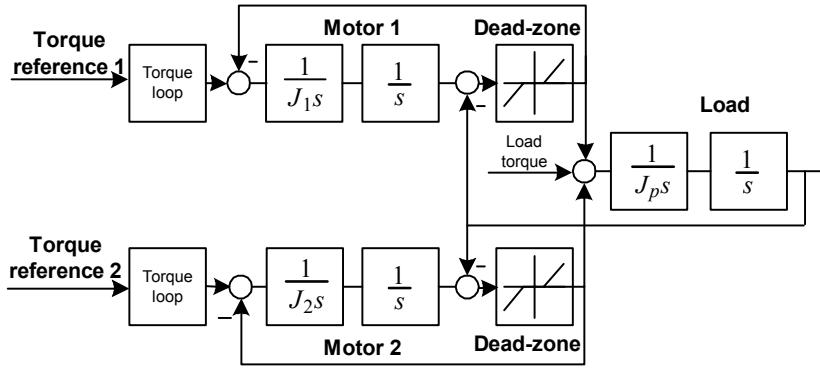


Figure 3.12: Block diagram of two motors driving a load through elastic gears with deadzone.

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  $\dot{\phi}_1$  and  $\dot{\phi}_2$  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.

### 3.7.2 Inductive circuit elements

The magnetic field intensity  $\vec{H}$  and the magnet flux density  $\vec{B}$  satisfy the equations

$$\nabla \times \vec{H} = \vec{J} \quad (3.129)$$

$$\nabla \cdot \vec{B} = 0 \quad (3.130)$$

where  $\vec{J}$  is the current density. The relation between the two vector fields  $\vec{H}$  and  $\vec{B}$  is given by

$$\vec{B} = \mu \vec{H} \quad (3.131)$$

where the permeability  $\mu$  is a material constant given by the constitutive law

$$\mu = (1 + \chi_m) \mu_0 \quad (3.132)$$

of the magnetic material, where  $\mu_0$  is the permeability of free space.

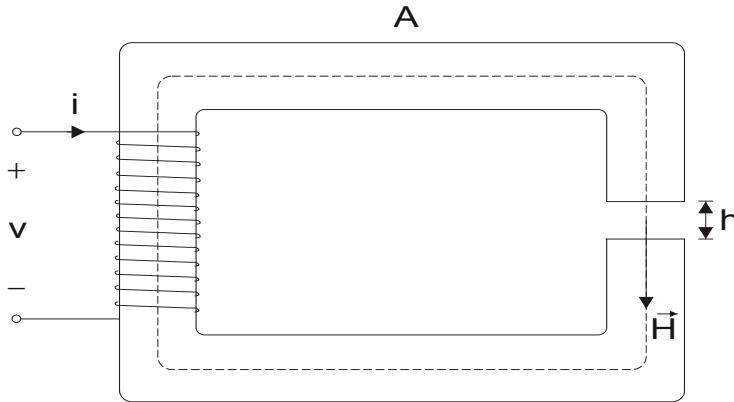


Figure 3.13: Magnetic circuit with airgap

We consider the magnetic circuit in Figure 3.13. The magnetic circuit has an iron core of cross section  $A$  and length  $\ell$ . The iron core has a small airgap of length  $h$  and a winding of  $N$  turns with current  $i$ . Then the integral form of (3.129) is

$$\oint \vec{H} \cdot d\vec{s} = Ni =: \mathcal{F} \quad (3.133)$$

where  $\mathcal{F}$  is the *magnetomotive force* (mmf), which is also called the *ampere-turns*, and  $d\vec{s}$  is a differential length which is tangent to the path of integration. The magnetic flux  $\phi$  of the circuit is

$$\phi = AB \quad (3.134)$$

The iron core has much higher magnetic permeability than the surrounding air, and as the airgap is small, the iron core and the airgap form a closed path for the magnetic flux  $\phi$  of length  $\ell$  in the core and length  $h$  in air. The magnetic flux density  $\vec{B}$  has zero divergence, which means that the flux  $\phi$  is the same for all cross sections of the iron core and for the airgap. Therefore we may consider the iron core and the airgap to form a magnetic circuit where the magnetomotive force  $Ni$  sets up a magnetic flux  $\phi$  that flows

through the circuit. The magnitude of the magnetic field intensity is  $H_c = B/\mu$  in the core and  $H_a = B/\mu_0$  in the airgap. Then (3.133) gives

$$Ni = H_c \ell + H_a h = \left( \frac{\ell}{\mu} + \frac{h}{\mu_0} \right) B = \left( \frac{\ell}{\mu} + \frac{h}{\mu_0} \right) \frac{\phi}{A} \quad (3.135)$$

The magnetic circuit has a reluctance  $\mathcal{R}$  defined so that the relation between the magnetomotive force and the flux is

$$Ni = \mathcal{R}\phi \quad (3.136)$$

We see that the reluctance of the magnetic circuit is

$$\mathcal{R} = \frac{Ni}{\phi} = \frac{1}{A} \left( \frac{\ell}{\mu} + \frac{h}{\mu_0} \right) \approx \frac{h}{A\mu_0} \quad (3.137)$$

where the approximation can be done as the magnetic permeability  $\mu_0$  in air is much smaller than the permeability  $\mu$  in iron.

We now turn to the electrical circuit with the coil. Define the magnetic *flux linkage*  $\lambda$  of the coil which is defined by

$$\lambda := Ni \quad (3.138)$$

In the example in this section, the flux linkage is a linear function of the current, which is given by

$$\lambda = Li \quad (3.139)$$

where

$$L = \frac{\mu_0 N^2 A}{h} \quad (3.140)$$

is the inductance of the coil.

The current is given by the constitutive equation

$$i = i(\lambda) \quad (3.141)$$

The voltage set up in the winding is the time derivative of the state  $\lambda$  by Faraday's law:

$$u = \dot{\lambda} \quad (3.142)$$

### 3.7.3 Capacitive circuit elements

The state of a capacitive circuit element is given by the electrical charge  $q$ , and the current is defined by

$$i = \dot{q} \quad (3.143)$$

The voltage is given by the constitutive equation

$$u = u(q) \quad (3.144)$$

In the case of the linear element the charge is given by

$$q = Cu \quad (3.145)$$

where  $C$  is the capacitance.

### 3.7.4 Magnetic energy of a linear inductive element

In this section we will derive an expression for the stored energy of a linear inductive element with current  $i$ , flux linkage  $\lambda = Li$ , and voltage  $u$ . The state of the element is given by the flux linkage  $\lambda$ . The voltage is given by Faraday's law to be  $u = \dot{\lambda}$ . The power supplied to the inductive element is  $P = iu$ , and the stored energy has the differential

$$dW_m = Pdt = iudt = i \frac{d\lambda}{dt} dt = id\lambda \quad (3.146)$$

We can then calculate the stored energy corresponding to the state  $\lambda$  from

$$W_m(\lambda) = \int_0^\lambda i(\lambda') d\lambda' = \int_0^\lambda \frac{\lambda'}{L} d\lambda' = \frac{\lambda^2}{2L} \quad (3.147)$$

In this expression the variable  $\lambda'$  is introduced as a integration variable because  $\lambda$  is the upper limit of the integration, and at the same time the current  $i(\lambda)$  is a function of  $\lambda$ .

### 3.7.5 Stored energy of a linear capacitive element

Consider a linear capacitive element with charge  $q$ , current  $i$  and voltage  $u$ . The state of the capacitive element can be given by the charge  $q$ . The constitutive equation for the element is  $q = Cu$  where  $C$  is the capacitance of the element. The current is by definition given by  $i = \dot{q}$ . Proceeding as for the inductive element we find that the stored energy  $W_c$  has differential

$$dW_c = uidt = u \frac{dq}{dt} dt = u dq \quad (3.148)$$

The stored energy corresponding to the state  $q$  is found to be

$$W_c(q) = \int_0^q u(q') dq' = \int_0^q \frac{q'}{C} dq' = \frac{q^2}{2C} \quad (3.149)$$

### 3.7.6 Energy and coenergy

We have seen that the stored energy of an inductive element is given as a function of the flux linkage  $\lambda$ , while the stored energy of a capacitive element can be given as a function of the charge  $q$ . In the derivation of models for electromechanical systems it is convenient to have energy functions given by the current  $i$  and the voltage  $u$ , and this is the motivation for introducing the concept of the *coenergy*. It is important to notice the coenergy does not have a clear physical interpretation, it is merely a mathematical tool that is useful in the to calculate the forces acting in electromechanical systems.

The energy of an inductive element is

$$W_m(\lambda) = \int_0^\lambda i(\lambda') d\lambda' \quad (3.150)$$

where the energy is given as a function of the flux linkage  $\lambda$ . To change the variable to the current  $i$  we define the *coenergy* of an inductive element by

$$W_m^*(i) = \int_0^i \lambda(i') di' \quad (3.151)$$

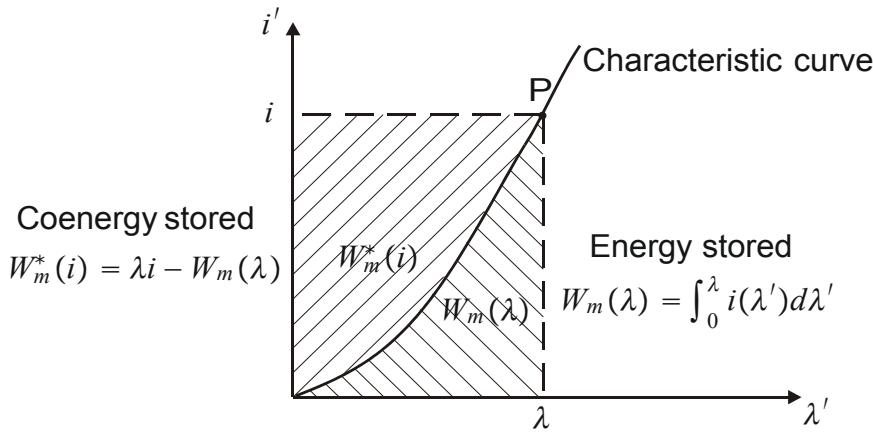


Figure 3.14: Energy and coenergy of an inductive element.

From Figure 3.14 it is seen that

$$W_m(\lambda) + W_m^*(i) = \lambda i \quad (3.152)$$

which means that the coenergy is found from the energy through a Legendre transformation

$$W_m^*(i) = \lambda i - W_m(\lambda) \quad (3.153)$$

Again we stress that the coenergy is not necessarily a meaningful physical quantity, but rather a mathematical tool that is useful in the modeling of electromechanical systems. The differentials of the energy and the coenergy is given by

$$dW_m(\lambda) = id\lambda, \quad dW_m^*(i) = id\lambda + \lambda di - dW_m(\lambda) = \lambda di \quad (3.154)$$

This implies

$$\frac{\partial W_m(\lambda)}{\partial \lambda} = i \quad \text{and} \quad \frac{\partial W_m^*(i)}{\partial i} = \lambda \quad (3.155)$$

**Example 46** If the flux linkage is given by  $\lambda = Li$  where the impedance  $L$  is a constant, then the coenergy is

$$W_m^*(i) = \lambda i - \frac{\lambda^2}{2L} = \frac{1}{2}Li^2 \quad (3.156)$$

Then the numerical value of the energy  $W_m(\lambda)$  and the coenergy  $W_m^*(i)$  is the same, note however, that the energy is a function of  $\lambda$  while the coenergy is a function of  $i$ .

The energy of a capacitive element is

$$W_c(q) = \int_0^q u(q') dq' \quad (3.157)$$

We note that the energy is given in terms of the stored charge  $q$ . It may be desirable to use the voltage  $u$  as a free variable instead of  $q$ . We therefore define the *coenergy* by

$$W_c^*(u) = \int_0^u q(u') du' \quad (3.158)$$

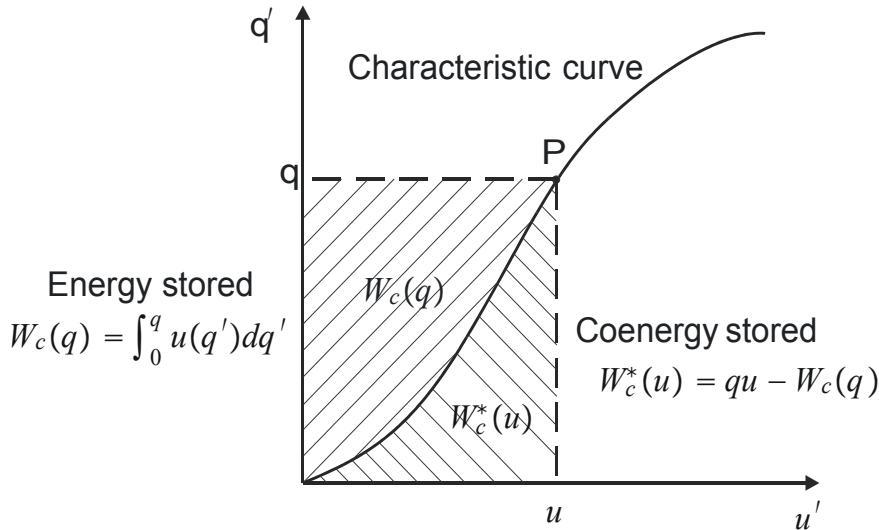


Figure 3.15: Energy and coenergy of a capacitive element.

From Figure 3.15 it is seen that

$$W_c(q) + W_c^*(u) = qu \quad (3.159)$$

it may be concluded that the coenergy is found from the energy through the Legendre transformation

$$W_c^*(u) = qu - W_c(q) \quad (3.160)$$

The differential of the energy is

$$dW_c(q) = u dq \quad (3.161)$$

while the differential of coenergy is

$$dW_c^*(u) = q du + u dq - dW_c(q) = q du \quad (3.162)$$

We see that

$$\frac{\partial W_c(q)}{\partial q} = u \quad \text{and} \quad \frac{\partial W_c^*(u)}{\partial u} = q \quad (3.163)$$

**Example 47** If  $q = Cu$  and the capacitance  $C$  is a constant, then

$$W_c^*(u) = qu - \frac{q^2}{2C} = \frac{1}{2}Cu^2 \quad (3.164)$$

and we see that the numerical value of the energy and the coenergy is the same. In this case the characteristic curve of Figure 3.15 is a straight line.

### 3.7.7 Electromechanical two-port with inductive element

In this section we will consider an important energy-conversion system which will be described as a two-port where input port is electrical with voltage  $u$  and current  $i$ , and the output port is mechanical with effort  $F$  and flow  $\dot{x}$ . The conversion from electrical energy to mechanical energy is done through an inductive element with flux linkage  $\lambda$ .

The electrical port has current  $i$  and voltage given by  $u = \dot{\lambda}$ . The constitutive equations are assumed to be given by

$$i = i(\lambda, x) \quad (3.165)$$

$$F = F(\lambda, x) \quad (3.166)$$

The power supplied to the two-port is

$$P = i\dot{\lambda} - F\dot{x} \quad (3.167)$$

The energy stored in the magnetic system is assumed to be given by  $W_m(\lambda, x)$ , and we get

$$dW_m(\lambda, x) = Pdt = id\lambda - Fdx \quad (3.168)$$

Note that  $dW_m(\lambda, x)$  is the absolute differential of the function  $W_m(\lambda, x)$ . The absolute differential may also be written

$$dW_m(\lambda, x) = \frac{\partial W_m(\lambda, x)}{\partial \lambda} d\lambda + \frac{\partial W_m(\lambda, x)}{\partial x} dx \quad (3.169)$$

Comparing the two expressions for the absolute differential we find that

$$i(\lambda, x) = \frac{\partial W_m(\lambda, x)}{\partial \lambda} \quad (3.170)$$

$$F(\lambda, x) = -\frac{\partial W_m(\lambda, x)}{\partial x} \quad (3.171)$$

The coenergy of the system is

$$W_m^*(i, x) = \lambda i - W_m(\lambda, x) \quad (3.172)$$

The absolute differential of the coenergy is

$$dW_m^*(i, x) = id\lambda + \lambda di - dW_m(\lambda, x) = \lambda di + Fdx \quad (3.173)$$

Then, from the general expression

$$dW_m^*(i, x) = \frac{\partial W_m^*(i, x)}{\partial i} di + \frac{\partial W_m^*(i, x)}{\partial x} dx \quad (3.174)$$

for the absolute differential it follows that

The flux linkage  $\lambda$  and the force  $F$  of an electromechanical inductive element are given from the coenergy  $W_m^*(i, x)$  by

$$\lambda(i, x) = \frac{\partial W_m^*(i, x)}{\partial i} \quad (3.175)$$

$$F(i, x) = \frac{\partial W_m^*(i, x)}{\partial x} \quad (3.176)$$

### 3.7.8 Electromechanical two-port with linear flux linkage

If the flux linkage is linear in the current so that  $\lambda = L(x)i$ , and if the force is zero when the flux linkage is zero so that  $F(0, x) = 0$ , then the energy  $W_m(\lambda, x)$  can be found by first integrating (3.168) from  $(0, 0)$  to  $(0, x)$ , and then from  $(0, x)$  to  $(\lambda, x)$ . This gives

$$W_m(\lambda, x) = - \int_0^x F(0, x') dx' + \int_0^\lambda i(\lambda', x) d\lambda' = \int_0^\lambda \frac{\lambda'}{L(x)} d\lambda' = \frac{1}{2} \frac{\lambda^2}{L(x)} \quad (3.177)$$

The coenergy is then given by

$$W_m^*(i, x) = \lambda i - \frac{\lambda^2}{2L(x)} = \frac{1}{2} L(x) i^2 \quad (3.178)$$

An electromechanical two-port with linear flux linkage  $\lambda = L(x)i$  has force given by

$$F(\lambda, x) = - \frac{\partial W_m(\lambda, x)}{\partial x} = \frac{1}{2} \frac{\partial L(x)}{\partial x} \frac{\lambda^2}{L(x)^2} \quad (3.179)$$

or alternatively, by

$$F(i, x) = \frac{\partial W_m^*(i, x)}{\partial x} = \frac{1}{2} \frac{\partial L(x)}{\partial x} i^2 \quad (3.180)$$

The voltage is given by

$$u = \dot{\lambda} = \frac{d}{dt} [L(x) i] = L \frac{di}{dt} + \frac{\partial L(x)}{\partial x} \dot{x} i \quad (3.181)$$

A convenient way of describing this is to write

$$u = L \frac{di}{dt} + e \quad (3.182)$$

where the first term is the self inductance of the coil, and

$$e := i \frac{\partial L(x)}{\partial x} \dot{x} \quad (3.183)$$

is the voltage induced because of the velocity  $\dot{x}$ . The system is then regarded as a two-port with input port variables  $u$  and  $i$  and output port variables  $e$  and  $\dot{x}$ . The output port is connected to a two-port describing the electromechanical conversion. This two-port has input port with variables  $e$  and  $i$ , and output variables  $F$  and  $\dot{x}$ . This electromechanical conversion unit does not store energy as  $ei = F\dot{x}$ , which means the electrical power  $ei$  is equal to mechanical power  $F\dot{x}$  out.

### 3.7.9 Magnetic levitation

In this section we will derive the model of the magnetic levitation experiment which is used for teaching in many control laboratories. An iron ball of radius  $R$  is lifted by a magnet with a coil of  $N$  turns and a current  $i$  around a core of length  $l_c$  and cross section  $A = \pi R^2$ . The vertical position of the ball is  $z$ , which is positive in the downwards vertical direction. This coil has a magnetic circuit with magnetomotive force  $Ni$ . The flux  $\phi$  flows through the iron core, then over the airgap, through the ball, and finally along the return path through the open air as shown in Figure 3.16. This gives

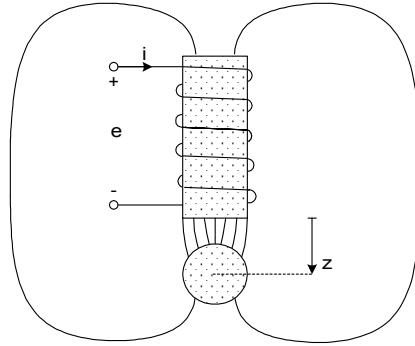


Figure 3.16: Magnetic levitation experiment.

$$Ni = \phi (\mathcal{R}_a + \mathcal{R}_c + \mathcal{R}_b + \mathcal{R}_r) \quad (3.184)$$

where

$$\mathcal{R}_a = \frac{z}{A\mu_0} \quad (3.185)$$

is the reluctance of the airgap,  $\mathcal{R}_c$  is the reluctance of the core,  $\mathcal{R}_b$  is the reluctance of the ball, and  $\mathcal{R}_r$  is the reluctance of the return path through the open air. The reluctances  $\mathcal{R}_c$  and  $\mathcal{R}_b$  are negligible. Moreover, if the ball is sufficiently close to the core, then the return path through the open air will not change significantly as the ball moves, and we may assume that  $\mathcal{R}_r$  is a constant. Therefore we may write

$$Ni = \phi \frac{z + z_0}{A\mu_0} \quad (3.186)$$

where  $z_0 = \mathcal{R}_r A\mu_0$  is a constant. The flux linkage is

$$\lambda = N\phi = \frac{N^2 A\mu_0 i}{z + z_0}, \quad (3.187)$$

and we find that the inductance is

$$L(z) = \frac{N^2 A\mu_0}{z + z_0} \quad (3.188)$$

The magnetic force on the ball is found from (3.180) to be

$$F = \frac{i^2}{2} \frac{\partial L(z)}{\partial z} = -\frac{\mu_0 A N^2}{2} \frac{i^2}{(z + z_0)^2} \quad (3.189)$$

The dynamics of the electrical circuit are given by

$$u = Ri + \dot{\lambda} = Ri + \left( \frac{d}{dt} L(z) \right) i + L(z) \frac{di}{dt} \quad (3.190)$$

The model for the magnetic levitation experiment is then found from the equation of motion and the circuit dynamics to be

$$m\ddot{z} = -\frac{1}{2}AN^2\mu_0 \frac{i^2}{(z+z_0)^2} + mg \quad (3.191)$$

$$L\frac{di}{dt} = -Ri + \frac{\mu_0 AN^2}{(z+z_0)^2} \frac{dz}{dt} i + u \quad (3.192)$$

Let  $z_d$  be the constant desired position of the ball. The solution  $(z_d, i_d)$  is found from

$$0 = m\ddot{z}_d = -\frac{1}{2}AN^2\mu_0 \frac{i_d^2}{(z_d+z_0)^2} + mg \quad (3.193)$$

which gives the constant current

$$i_d = \sqrt{\frac{2mg}{AN^2\mu_0}} (z_d + z_0) \quad (3.194)$$

corresponding to  $z_d$ . We define  $\Delta z = z - z_d$  and  $\Delta i = i - i_d$  and get the linearized model around  $z = z_d$  in the form

$$\begin{aligned} m\Delta\ddot{z} &= \frac{AN^2\mu_0 i_d^2}{(z_d+z_0)^3} \Delta z - AN^2\mu_0 \frac{i_d}{(z_d+z_0)^2} \Delta i \\ &= \frac{2mg}{z_d+z_0} \Delta z - \frac{\sqrt{2AN^2\mu_0 mg}}{z_d+z_0} \Delta i \end{aligned} \quad (3.195)$$

### 3.7.10 Voice coil

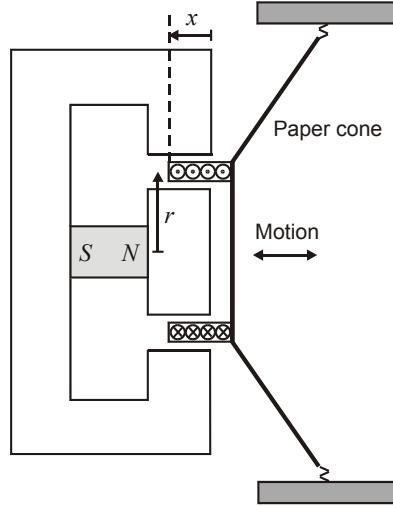


Figure 3.17: Voice coil actuator for loudspeaker.

In a loudspeaker the sound is created with a voice coil motor by setting up motion in a membrane as shown in Figure 3.17. The voice coil is an inductive motor with translational motion. In this section we will develop a simplified model for the loudspeaker dynamics that gives some insight into the dynamics at low frequencies (Meisel 1966), (Crandall, Karnopp, E.F. Kurtz and Pridmore-Brown 1968). More accurate models will depend on acoustic properties of the loudspeaker, and on the modes of vibration of the membrane. The membrane has mass  $m$ , the air resistance is modeled as a damper with coefficient  $b$ , and the cone suspension is modeled as a spring with stiffness  $k$ . The electrical circuit has input voltage  $u$ , resistance  $R$  and inductance  $L$  and induced voltage  $e$ .

The flux density  $B$  in the airgap is constant and set up by a permanent magnet. The magnetic flux from the permanent magnet through the winding is

$$\phi = 2\pi r B x \quad (3.196)$$

where  $x$  is the position of winding which is wound around a cylinder that is fixed to the membrane. The flux linkage for the coil with  $N$  windings can then be written

$$\lambda(i, x) = Li + K_e x \quad (3.197)$$

where  $L$  is the self inductance of the coil and  $K_e = 2\pi r N B$ . The magnetic coenergy is found by integrating (3.173) from  $(0, 0)$  to  $(0, x)$ , and then from  $(0, x)$  to  $(i, x)$ . Under the assumption that the force is zero when  $i = 0$ , this gives

$$W_m^*(i, x) = \int_0^i \lambda(i', x) di' = \frac{1}{2} L i^2 + K_e x i \quad (3.198)$$

We find the equations

$$F = K_e i, \quad e = K_e \dot{x} \quad (3.199)$$

from  $e = \dot{\lambda}$  and  $F = \partial W_m^*(i, x) / \partial x$ .

The equation of motion of the membrane and the voltage law for the circuit give the model

$$m \ddot{x} + b \dot{x} + kx = K_e i \quad (3.200)$$

$$L \frac{di}{dt} = -Ri - K_e \dot{x} + u \quad (3.201)$$

### 3.7.11 Electromagnetic three-port

Electrical motors may have two electrical ports and one mechanical port. In this section we will study the energy flow of this type of electromechanical three-port.

We consider a three-port where port one has voltage  $u_1 = \dot{\lambda}_1$  and current  $i_1$ , port two has voltage  $u_2 = \dot{\lambda}_2$  and current  $i_2$ , while the mechanical port has effort  $T$  and flow  $\dot{\theta}$ . The power flowing into the system is

$$P = \mathbf{i}^T \dot{\boldsymbol{\lambda}} - T \dot{\theta} \quad (3.202)$$

where  $\mathbf{i} = (i_1, i_2)^T$  and  $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T$ . The absolute differential of the energy  $W_m(\boldsymbol{\lambda}, \theta)$  stored in the system is found from  $dW = P dt$  to be

$$dW_m(\boldsymbol{\lambda}, \theta) = \mathbf{i}^T d\boldsymbol{\lambda} - T d\theta \quad (3.203)$$

and by comparing this to the defining expression for the absolute differential we find that

$$\mathbf{i} = \frac{\partial W_m(\boldsymbol{\lambda}, \theta)^T}{\partial \boldsymbol{\lambda}}, \quad T = -\frac{\partial W_m(\boldsymbol{\lambda}, \theta)}{\partial \theta} \quad (3.204)$$

If we assume that the flux linkage vector is a linear function of the current vector, then the flux linkage can be written

$$\boldsymbol{\lambda} = \mathbf{M}(\theta)\mathbf{i} \quad (3.205)$$

where  $\mathbf{M}(\theta)$  is the inductance matrix. The matrix  $\mathbf{M}(\theta)$  can be shown to be positive definite and symmetric. We may then integrate  $dW_m(\boldsymbol{\lambda}, \theta)$  to get the energy function

$$W_m(\boldsymbol{\lambda}, \theta) = \int_0^{\boldsymbol{\lambda}} (\boldsymbol{\lambda}')^T \mathbf{M}^{-1}(\theta) d\boldsymbol{\lambda}' = \frac{1}{2} \boldsymbol{\lambda}^T \mathbf{M}^{-1}(\theta) \boldsymbol{\lambda} \quad (3.206)$$

The coenergy is found to be

$$W_m^*(\mathbf{i}, \theta) = \boldsymbol{\lambda}^T \mathbf{i} - \frac{1}{2} \boldsymbol{\lambda}^T \mathbf{M}^{-1}(\theta) \boldsymbol{\lambda} = \mathbf{i}^T \mathbf{M}(\theta) \mathbf{i} - \frac{1}{2} \mathbf{i}^T \mathbf{M}^T(\theta) \mathbf{M}^{-1}(\theta) \mathbf{M}(\theta) \mathbf{i} \quad (3.207)$$

which gives

$$W_m^*(\mathbf{i}, \theta) = \frac{1}{2} \mathbf{i}^T \mathbf{M}(\theta) \mathbf{i} \quad (3.208)$$

It follows that the torque is

$$T = \frac{\partial W_m^*(\mathbf{i}, \theta)}{\partial \theta} = \frac{1}{2} \mathbf{i}^T \frac{\partial \mathbf{M}(\theta)}{\partial \theta} \mathbf{i} \quad (3.209)$$

### 3.7.12 Electromechanical capacitive element

In this section we will study an electromechanical system with a capacitive element. The system is described as a two-port where the input port is electrical with voltage  $u$  and current  $i$ , and the output port is mechanical with effort  $F$  and flow  $\dot{x}$ . We assume that the constitutive equations are given by

$$u = u(q, x) \quad (3.210)$$

$$F = F(q, x) \quad (3.211)$$

The power supplied to the two-port is

$$P = u\dot{q} - F\dot{x} \quad (3.212)$$

The energy stored in the magnetic system is  $W_c(q, x)$ . The rate of change of the stored energy must be equal to the energy supplied due to the power  $P$ . This may be expressed

$$\frac{dW_c(q, x)}{dt} = P \Rightarrow dW_c(q, x) = Pdt = udq - Fdx \quad (3.213)$$

Note that  $dW_c(q, x)$  is the absolute differential of the function  $W_c(q, x)$ . The absolute differential may also be written

$$dW_c(q, x) = \frac{\partial W_c(q, x)}{\partial q} dq + \frac{\partial W_c(q, x)}{\partial x} dx \quad (3.214)$$

Comparing the two expressions for the absolute differential we find that

$$u(q, x) = \frac{\partial W_c(q, x)}{\partial q} \quad (3.215)$$

$$F(q, x) = -\frac{\partial W_c(q, x)}{\partial x} \quad (3.216)$$

The coenergy of the system is defined by

$$W_c^*(u, x) = uq - W_c(q, x) \quad (3.217)$$

The absolute differential of the coenergy is

$$dW_c^*(u, x) = udq + qdu - dW_c(q, x) = qdu + Fdx \quad (3.218)$$

The charge  $q(u, x)$  and the force  $F(u, x)$  of an electromechanical capacitive element with constitutive equations (3.210, 3.211) are given from the coenergy  $W_c^*(u, x)$  by

$$q(u, x) = \frac{\partial W_c^*(u, x)}{\partial u} \quad (3.219)$$

$$F(u, x) = \frac{\partial W_c^*(u, x)}{\partial x} \quad (3.220)$$

### 3.7.13 Electromechanical two-port with linear charge

If the charge  $q$  is linear in the voltage  $u$  so that  $q = C(x)u$ , and if the force is zero when the charge is zero so that  $F(0, x) = 0$ , then the energy  $W_c(q, x)$  can be found by first integrating (3.213) from  $(0, 0)$  to  $(0, x)$ , and then from  $(0, x)$  to  $(q, x)$ . This gives

$$W_c(q, x) = - \int_0^x F(0, x') dx' + \int_0^q u(q', x) dq' = \int_0^q \frac{q'}{C(x)} dq' = \frac{1}{2} \frac{q^2}{C(x)} \quad (3.221)$$

The coenergy is then from (3.217) to be

$$W_c^*(v, x) = C(x)u^2 - \frac{1}{2} \frac{C(x)^2 u^2}{C(x)} = \frac{1}{2} C(x)u^2 \quad (3.222)$$

An electromechanical system with linear charge equation  $q = C(x)u$  has force given by the energy according to (3.216), which gives

$$F(q, x) = -\frac{1}{2} \frac{d}{dx} \left( \frac{1}{C(x)} \right) q^2 \quad (3.223)$$

Alternatively, the force is found from the coenergy expression (3.220) to be

$$F(v, x) = \frac{1}{2} \frac{dC(x)}{dx} u^2 \quad (3.224)$$

### 3.7.14 Example: Capacitive microphone

A capacitive microphone can be designed with a membrane that moves a capacitive element so that capacitance is altered. The capacitive element is charged with a voltage  $E_0$ , and the change in capacitance can then be picked up by measuring the capacitor current using an amplifier with resistance  $R$ . In this way the sound that excites the membrane can be recorded (Crandall et al. 1968).

The constitutive equation for the charge over the capacitor is  $q = C(x)u$  where the capacitance  $C(x)$  can be described by the relation

$$C(x) = C_0 \frac{d_0}{d_0 + x} \quad (3.225)$$

Here  $d_0$  is the nominal position of the moving plate with mass  $m$  and position  $x$ . Moreover, the force over the capacitor is zero when the charge is zero. The force over the capacitor is then found from (3.223) to be

$$F(q, x) = -\frac{1}{2} \frac{d}{dx} \left( \frac{d_0 + x}{C_0 d_0} \right) q^2 \quad (3.226)$$

and the constitutive equations are found to be

$$u(q, x) = \frac{q}{C(x)}, \quad F(q, x) = -\frac{q^2}{2C_0 d_0} \quad (3.227)$$

The dynamic model is then

$$m\ddot{x} + b\dot{x} + kx + \frac{q^2}{2C_0 d_0} = -F_a \quad (3.228)$$

$$R\dot{q} + \frac{q(d_0 + x)}{C_0 d_0} = u_0 \quad (3.229)$$

Here (3.228) is the equation of motion for the moving plate where  $b$  is the friction coefficient,  $k$  is the spring stiffness, the last term on the left side is the force from the capacitor, and  $F_a$  is the acoustic excitation force due to the sound. The stationary value of this force is zero. Equation (3.229) is the voltage law for the electric circuit with a resistor  $R$  in series with the capacitor, and with a constant driving voltage  $u_0$ . We note that the dynamics of the system are nonlinear. Before linearization the equilibrium points are investigated. By setting the time derivatives of  $x$  and  $q$  to zero and eliminating  $q$  we find the equilibrium values  $x_0$  and  $q_0$ , which are related to  $u_0$  by

$$kx_0 + \frac{q_0^2}{2C_0 d_0} = 0, \quad \frac{q_0(d_0 + x_0)}{C_0 d_0} = u_0 \quad (3.230)$$

Elimination of  $q_0$  by inserting the second equation of (3.230) into the first equation gives

$$kx_0 + \frac{C_0 d_0 u_0^2}{2(d_0 + x_0)^2} = 0, \quad -d_0 \leq x_0 \leq 0 \quad (3.231)$$

where the first term is the spring force of the mechanical spring and the second term can be regarded as a nonlinear electrical spring. There are two equilibrium points where one is stable, and the other is unstable.

Linearization and Laplace transformation of (3.228) and (3.229) around  $x = x_0$ ,  $q = q_0$  and  $F_a = 0$  gives

$$(ms^2 + bs + k) \Delta x(s) + \frac{q_0}{C_0 d_0} \Delta q(s) = -F_a(s) \quad (3.232)$$

$$q_0 \Delta x(s) + (C_0 d_0 R s + d_0 + x_0) \Delta q(s) = 0 \quad (3.233)$$

By inserting  $\Delta x$  from the second equation into the first equation we get

$$\left[ (ms^2 + bs + k) (C_0 d_0 R s + d_0 + x_0) - \frac{q_0^2}{C_0 d_0} \right] \frac{\Delta q(s)}{q_0} = F_a(s) \quad (3.234)$$

The output from the microphone is the increment in voltage  $\Delta u_R$  over the resistor  $R$ , which is  $\Delta u_R = R \Delta i = R s \Delta q$ . The transfer function from the acoustic force to the measurement is found from (3.234) to be

$$\frac{\Delta u_R(s)(s)}{F_a(s)} = \frac{R s \Delta q(s)}{F_a(s)} = \frac{R s q_0}{(ms^2 + bs + k) (C_0 d_0 R s + d_0 + x_0) - \frac{q_0^2}{C_0 d_0}} \quad (3.235)$$

For large  $R$  this expression can be approximated with

$$\frac{\Delta u_R(s)}{F_a}(s) = \frac{q_0}{(ms^2 + bs + k) C_0 d_0} = \frac{\frac{q_0}{k C_0 d_0}}{\left(1 + 2\zeta_0 \frac{s}{\omega_0} + \left(\frac{s}{\omega_0}\right)^2\right)} \quad (3.236)$$

where

$$\omega_0^2 = \frac{k}{m}, \quad \zeta_0 = \frac{b}{2\sqrt{km}} \quad (3.237)$$

We see that the measured voltage  $u_0$  is proportional to the acoustic excitation force  $F_a$  in the frequency range up to  $\omega_0$ .

**Example 48** The condition for the equilibrium point to be stable is given by

$$k(d_0 + x_0) - \frac{q_0^2}{C_0 d_0} \geq 0 \quad (3.238)$$

which is simplified to

$$x_0 \geq -\frac{d_0}{3} \quad (3.239)$$

by inserting the equation (3.230) for the equilibrium point.

### 3.7.15 Piezoelectric actuator

Piezoelectric sensors or actuators are electromechanical systems where the energy conversion is due to capacitive effects. An example of this is the axial piezoelectric stack actuator (IEEE 1987), (Fuller, Elliott and Nilsen 1996), which is a translational capacitive actuator with small displacements that can exert large forces. These actuators are of particular interest for vibration damping.

The constitutive equations for a piezoelectric actuator are given by

$$q(u, x) = C_p x + C u \quad (3.240)$$

$$F_{pe}(u, x) = -K x + C_p u \quad (3.241)$$

where  $x$  is the elongation of the piezoelectric material,  $F_{pe}$  is the actuator force,  $q$  is the charge,  $u$  is the voltage,  $C$  is the capacitance,  $C_p$  is the piezoelectric constant, and  $K$  is the mechanical stiffness. The coenergy  $W_c^*(u, x)$  of the actuator is found by integrating  $dW_c^* = qdu + F_{pe}dx$  from  $(0, 0)$  to  $(0, x)$ , and then from  $(0, x)$  to  $(u, x)$ . This gives

$$W_c^*(u, x) = \int_0^x F_{pe}(0, x') dx' + \int_0^u q(u', x) du' \quad (3.242)$$

$$= - \int_0^x Kx' dx' + \int_0^q (C_p x + C u') du' \quad (3.243)$$

The coenergy is then found to be

$$W_c^*(u, x) = -\frac{1}{2}Kx^2 + C_p x u + \frac{1}{2}C u^2 \quad (3.244)$$

and it can be verified that

$$F_{pe} = \frac{\partial W_c^*(u, x)}{\partial x}, \quad q = \frac{\partial W_c^*(u, x)}{\partial u} \quad (3.245)$$

Moreover, the energy stored in the actuator is found to be

$$W_c(q, x) = qu - W_c^*(u, x) = \frac{1}{2}Kx^2 + \frac{1}{2}C u^2(q, x) \quad (3.246)$$

### 3.7.16 Actuator configuration

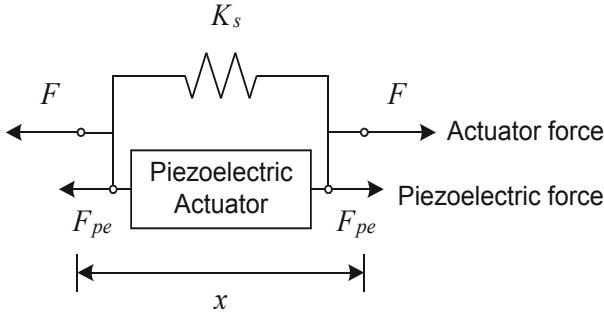


Figure 3.18: Piezoelectric actuator pretensioned by a spring

A typical actuator configuration would be to have the actuator pretensioned by a spring with stiffness  $K_s$  and equilibrium  $x_s$  (Figure 3.18), and to input the voltage  $u$  to the actuator. Then the actuator force will be  $F = F_{pe} - K_s(x - x_s)$ , and the actuator is described by the equation

$$F = -(K + K_s)(x - x_0) + C_p u \quad (3.247)$$

where  $x_0 = x_s K_s / (K + K_s)$ . In this case we view the piezoelectric actuator as a force actuator where the actuator force  $F$  is controlled with the voltage  $u$ . Then  $F$  is seen as an output while the deflection  $x$  is an input to the actuator (Figure 3.19). Alternatively, the actuator can be regarded as a displacement actuator controlled by the voltage  $u$  according to

$$x = x_0 + \frac{C_p u - F}{K + K_s} \quad (3.248)$$

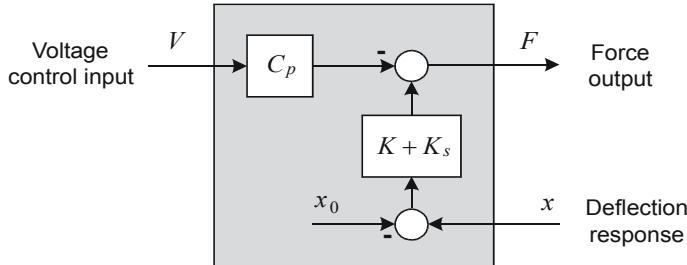


Figure 3.19: Piezoelectric actuator seen as a force actuator

In this case the deflection  $x$  is treated as an output while the force  $F$  on the actuator is an input (Figure 3.20. The decision on whether to treat the piezoelectric actuator as

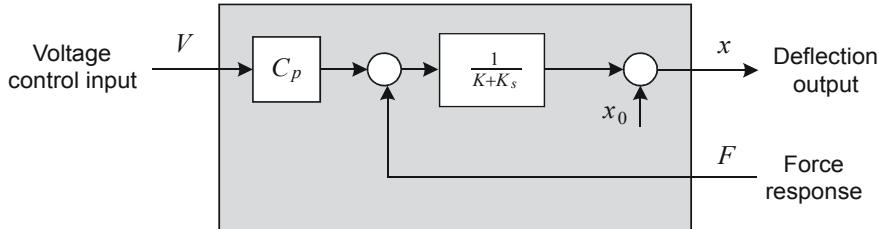


Figure 3.20: Piezoelectric actuator as a deflection actuator

a force or deflection actuator depends on how the actuator is connected to the system to be controlled. If the actuator is connected to a spring, then deflection should be the output of the actuator. If the actuator is connected to mass, then force should be the actuator output.

## 3.8 DC motor with externally controlled field

### 3.8.1 Model

In this section the model for DC motor with externally controlled field is presented (Figure 3.21). The field is assumed to be set up by a field winding. The field winding is an electrical circuit which has a conductor that is wound around an iron core in the stator. The rotor has a winding, which is called the armature winding, around an iron core in the rotor. The current in the field windings in the stator is used to set up a magnetic flux density  $\vec{B}_e$  through the rotor, and a Lorentz force  $\vec{F} \sim \vec{i}_a \times \vec{B}_e$  on the rotor result when a current  $\vec{i}_a$  is run through a conductor in the rotor windings.

Kirchhoff's voltage law for the field circuit in the stator gives

$$R_e i_e + \dot{\lambda}_e = u_e \quad (3.249)$$

where  $i_e$  is the field current,  $R_e$  is the field resistance,  $N_e$  is the number of windings in the field circuit,  $\lambda_e = \lambda_e(i_e)$  is the flux linkage of the field circuit, and  $u_e$  is the field

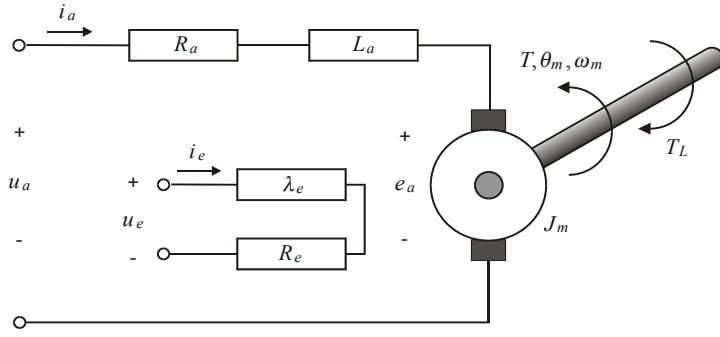


Figure 3.21: Armature circuit and field circuit of DC motor.

circuit voltage. The voltage law for the armature circuit in the rotor gives

$$R_a i_a + L_a \frac{d}{dt} i_a + e_a = u_a \quad (3.250)$$

where  $i_a$  is the armature current,  $R_a$  is the armature resistance,  $L_a$  is the armature inductance,  $u_a$  is the armature voltage,  $e_a$  is the induced voltage,  $\omega_m$  is the angular velocity of the motor shaft and  $k_f$  is a constant. The main feature of a DC motor is that the motor is designed so that the field is set up so that the induced voltage

$$e_a = k_e \lambda_e \omega_m \quad (3.251)$$

is proportional to the flux linkage  $\lambda_e$  and the shaft speed  $\omega_m$ . The electrical power  $e_a i_a$  and the mechanical power  $T \omega_m$  of the electromechanical conversion unit must be equal, and this implies that the motor torque

$$T = k_T \lambda_e i_a \quad (3.252)$$

where  $k_T = k_e$  is a constant, is proportional to the product of the magnetizing flux linkage  $\lambda_e$  and the armature current  $i_a$ . Then the armature part of the DC motor may be described by the two-port

$$R_a i_a + L_a \frac{d}{dt} i_a + k_e \lambda_e \omega_m = u_a \quad (3.253)$$

$$T = k_T \lambda_e i_a \quad (3.254)$$

The input port variables are  $u_a$  and  $i_a$  and the output port has variables  $T$  and  $\omega_m$ .

The equation of motion for the motor shaft is

$$J_m \dot{\omega}_m = T - T_L \quad (3.255)$$

where  $J_m$  is the inertia of the motor shaft, and  $T_L$  is the torque that acts on the motor shaft from the load.

The dynamic model of a DC motor with externally controlled field is given by

$$\dot{\lambda}_e = -R_e i_e + u_e \quad (3.256)$$

$$L_a \frac{di_a}{dt} = -R_a i_a - k_e \lambda_e \omega_m + u_a \quad (3.257)$$

$$J_m \dot{\omega}_m = k_T \lambda_e i_a - T_L \quad (3.258)$$

$$\dot{\theta}_m = \omega_m \quad (3.259)$$

The input variables to the DC motor are the field voltage  $u_e$  and the armature voltage  $u_a$ . Note that the model is nonlinear due to the product  $\lambda_e \omega_m$  in the armature equation and the product  $\lambda_e i_a$  in the field equation.

### 3.8.2 Network description

The energy of the armature circuit and the motor shaft is

$$V = \frac{1}{2} L_a i_a^2 + \frac{1}{2} J \omega_m^2 \quad (3.260)$$

The time derivative along solution trajectories is

$$\begin{aligned} \dot{V} &= i_a L_a \frac{di_a}{dt} + \omega_m J_m \frac{d\omega_m}{dt} \\ &= i_a (-R_a i_a - k_e \lambda_e \omega_m + u_a) + \omega_m (k_T \lambda_e i_a - T_L) \\ &= i_a u_a - \omega_m T_L - R_a i_a^2 \end{aligned} \quad (3.261)$$

This shows that if the load is passive, then the DC motor with input  $u_a$  and output  $i_a$  is passive. Note in particular that this result does not rely on any assumptions on how the field circuit is controlled. This means that a passive current controller like a PI controller will give a stable armature current control independently of the field circuit.

Moreover, the energy of the field circuit is

$$V_e = \frac{1}{2} L_e i_e^2 \quad (3.262)$$

with time derivative

$$\dot{V}_e = i_e L_e \frac{di_e}{dt} = i_e (-R_e i_e + u_e) \quad (3.263)$$

$$= i_e u_e - R_e i_e^2 \quad (3.264)$$

along the solutions of the system. These calculations show that the field circuit is passive with input  $u_e$  and output  $i_e$ . Moreover, it is seen that there is no energy exchange between the field circuit and the rest of the system. This means that the field current  $i_e$  can be controlled with a passive controller of the PI type independently of how the armature circuit is controlled.

In terms of a network description the field circuit is a passive electrical one-port with input port with voltage  $u_e$  and current  $i_e$ . The dynamic model of the armature circuit and the motor shaft can be described as in the case of a DC motor with constant field with interconnection of three passive two-ports, namely, the armature circuit, the electromechanical energy conversion unit, and the motor shaft.

### 3.8.3 DC motor with field weakening

Stationary conditions for a DC motor are found by setting the time derivatives to zero. Then the following equations result:

$$R_e i_e = u_e \quad (3.265)$$

$$R_a i_a + k_f \lambda_e \omega_m = u_a \quad (3.266)$$

$$k_T \lambda_e i_a = T_L. \quad (3.267)$$

The armature equation (3.266) gives an expression for the stationary velocity

$$\omega_m = \frac{u_a - R_a i_a}{k_f \lambda_e} = \frac{u_a}{k_f \lambda_e} - \frac{R_a T_L}{k_T k_f \lambda_e^2} \quad (3.268)$$

where the torque equation (3.267) was used to eliminate  $i_a$ . Suppose that the armature voltage is limited by

$$-U \leq u_a \leq U \quad (3.269)$$

while the flux linkage is limited by

$$0 \leq \lambda_e \leq \bar{\lambda} \quad (3.270)$$

Then, with a constant maximum flux linkage  $\lambda_e = \bar{\lambda}$  the maximum velocity will be

$$\omega_m = \frac{U}{k_f \bar{\lambda}} - \frac{R_a T_L}{k_T k_f \bar{\lambda}^2}. \quad (3.271)$$

It turns out that the speed of the the motor may be increased over this value by reducing the flux linkage  $\lambda_e$ . This is referred to as *field weakening*. To see this we note that for  $u_a = U$  then

$$\frac{\partial \omega_m}{\partial \lambda_e} = -\frac{U}{k_f \bar{\lambda}^2} + 2 \frac{R_a T_L}{k_T k_f \bar{\lambda}^3} < 0 \quad \text{if } \lambda_e > \lambda_{\min} = \frac{2R_a T_L}{k_T U}. \quad (3.272)$$

This means that when the armature voltage  $u_a$  has reached its maximum value  $U$ , an additional increase in velocity can be achieved by weakening the field as long as the load torque  $T_L$  is sufficiently small, so that  $\lambda_e > \lambda_{\min}$ . Field weakening is typically implemented so that the induced voltage  $e_a = k_f \lambda_e \omega_m$  is kept constant, that is, with a desired field is set to

$$\lambda_{ed} = \begin{cases} \bar{\lambda} & |e_a| < e_{a,\max} \\ \left| \frac{e_{a,\max}}{k_f \omega_m} \right| & \text{otherwise} \end{cases} \quad (3.273)$$

The induced voltage can be computed from the armature voltage and armature current measurements using

$$e_a = u_a - R_a i_a - L_a \frac{di_a}{dt} \quad (3.274)$$

and a feedback controller can be used to achieve the desired flux linkage  $\lambda_{ed}$ . In this case the desired flux linkage  $\lambda_{ed}$  is considered as an input to the system, and, assuming that the field circuit is sufficiently fast, the model is

$$L_a \frac{di_a}{dt} = -R_a i_a - k_f \lambda_{ed} \omega_m + u_a \quad (3.275)$$

$$J_m \dot{\omega}_m = k_T \lambda_{ed} i_a - T_L \quad (3.276)$$

$$\dot{\theta}_m = \omega_m \quad (3.277)$$

Typically, the armature current  $i_a$  is controlled with a PI controller where the armature voltage  $u_a$  is the input. Stability of the armature current control loop is ensured as the system with input  $u_a$  and output  $i_a$  is passive independently of the dynamics of the field circuit.

## 3.9 Dynamic model of the general AC motor

### 3.9.1 Introduction

DC motors with a constant field have been the usual electrical motors in control applications. The reason for this is that it is straightforward to control the armature current of this type of motor, and as the motor torque of a DC motor is proportional to the armature current when the field is constant, it has been possible to control the motor torque. This is certainly an ideal situation in control applications. In contrast to this, AC motors have mainly been used in steady-state drives where the transient performance has not been the main objective. There is a large literature on steady-state dynamics for AC motors. However, in the last two decades there has been a strong development in power electronics (Mohan, Undeland and Robbins 1989), and it is now possible to achieve servo control of AC motors (Leonhard 1996). In the context of servo control there is a need for dynamic models of AC motors that account for the transient dynamics. The inclusion of transient dynamics may at first be unfamiliar for readers with a background in the steady-state dynamics of AC motors, but it is needed for servo designs for AC motors. The dynamic models presented in the following will start with the general AC motor with a rotor that has a circular cross section. Then, this result will be specialized for induction motors, which are of great interest in control systems due to their rugged and reliable design.

### 3.9.2 Notation

For the modeling of the general AC motor we will represent voltages, currents and flux linkages in terms of two-dimensional coordinate vectors in the stator-fixed frame  $s$ , the rotor-fixed frame  $r$ , and in the rotor flux frame  $f$ . The stator-fixed frame  $s$  is also referred to as the  $ab$  frame, while the flux frame  $f$  is referred to as the  $dq$  frame. The tradition in the literature of electrical machines is to write the coordinate vectors as complex numbers. A voltage vector given in the coordinates of the stator frame  $s$  is written as the complex number

$$u^s = u_a + j u_b. \quad (3.278)$$

The same vector in the flux frame is written

$$u^f = u_d + j u_q. \quad (3.279)$$

The frame  $f$  is obtained by rotation of frame  $s$  by an angle  $\rho$ . The coordinate transformation is written

$$u^s = u^f e^{j\rho}. \quad (3.280)$$

The time derivative of the vector  $u^s$  in the  $s$  frame can be expressed by the time derivative in the  $f$  frame according to

$$\frac{du^s}{dt} = \frac{d}{dt} (u^f e^{j\rho}) = \left( \frac{du^f}{dt} + j\dot{\rho}u^f \right) e^{j\rho} \quad (3.281)$$

In the same way the vector can be given in the rotor frame  $r$  as  $u^r$ .

The rotational angle from the stator to the rotor is denoted  $\theta$ , so that

$$u^s = u^r e^{j\theta} \quad (3.282)$$

and

$$\frac{du^s}{dt} = \left( \frac{du^r}{dt} + j\dot{\theta}u^r \right) e^{j\theta} \quad (3.283)$$

### 3.9.3 Dynamic model

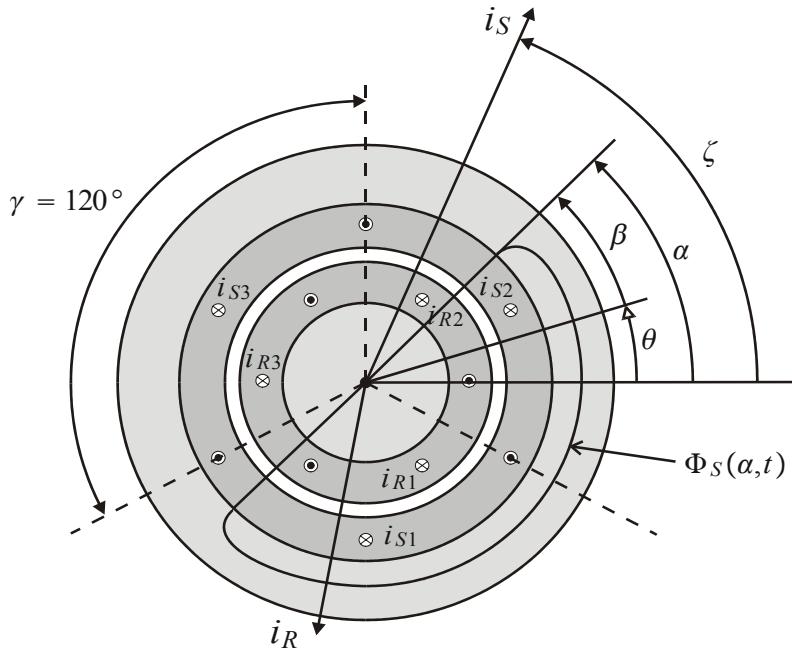


Figure 3.22: General AC machine.

We will here develop the dynamic model of a general alternating current (AC) motor based on the presentation in (Leonhard 1996) where complex numbers are used to represent vectors in the plane. An alternative formulation based on vector notation is found in (Vas 1990). The results of this section will later be specialized to the dynamic model of the induction motor. The general AC motor has a stationary part called the stator, which is assumed to be an iron cylinder containing a concentric rotor, which is the moving part of the motor. Both the rotor and the stator are assumed to have circular cross sections separated by a constant airgap  $h$ . Both stator and rotor have symmetrical three-phase windings close to the airgap. Each stator phase has  $N_S$  windings, while each rotor phase has  $N_R$  windings. The magnetic permeability of the fully laminated stator and rotor iron is assumed to be infinite, and iron losses are left out of the analysis. A motor with two poles is considered. For details on this consult (Leonhard 1996) for control

issues, and (Fitzgerald et al. 1983), which is a basic textbook on electrical machines.

The geometry of the AC motor is given in Figure 3.22. The angle  $\theta$  is the rotor angle,  $\alpha$  is the angular coordinate of the stator, and  $\beta = \alpha - \theta$  is the corresponding angular coordinate of the rotor. The currents of the stator phases 1, 2 and 3 are denoted  $i_{S1}$ ,  $i_{S2}$  and  $i_{S3}$ , respectively. The stator currents are balanced so that

$$i_{S1}(t) + i_{S2}(t) + i_{S3}(t) = 0 \quad (3.284)$$

Note that the currents  $i_{S1}(t)$ ,  $i_{S2}(t)$  and  $i_{S3}(t)$  are real, and, moreover, they are general time functions that need not have any specified wave-form. The radial magnetomotive force (mmf) distribution from the stator windings at the angular coordinate  $\alpha$  is by design sinusoidal and given by

$$\mathcal{F}_S(\alpha, t) = N_S [i_{S1} \cos \alpha + i_{S2} \cos(\alpha - \gamma) + i_{S3} \cos(\alpha - 2\gamma)], \quad \gamma = 120^\circ \quad (3.285)$$

From the trigonometric identity

$$\cos(x - y) = \cos x \cos y + \sin x \sin y \quad (3.286)$$

we get

$$\mathcal{F}_S(\alpha, t) = N_S [\cos \alpha (i_{S1} + i_{S2} \cos \gamma + i_{S3} \cos 2\gamma)] \quad (3.287)$$

$$+ \sin \alpha (i_{S1} + i_{S2} \sin \gamma + i_{S3} \sin 2\gamma). \quad (3.288)$$

Complex notation

$$\cos \alpha = \frac{1}{2} (e^{j\alpha} + e^{-i\alpha}) \quad (3.289)$$

$$\sin \alpha = \frac{1}{2j} (e^{j\alpha} - e^{-j\alpha}) \quad (3.290)$$

is introduced, and this makes it possible to write the magnetomotive force in the form

$$\mathcal{F}_S(\alpha, t) = \frac{1}{2} N_S [i_S^s(t) e^{-j\alpha} + i_S^{s*}(t) e^{j\alpha}] = N_S \operatorname{Re}[i_S^s(t) e^{-j\alpha}] \quad (3.291)$$

where the complex variable  $i_S^s$  and its complex conjugate  $i_S^{s*}$  are defined by

$$i_S^s = i_{S1} + i_{S2} e^{j\gamma} + i_{S3} e^{j2\gamma} \quad (3.292)$$

$$i_S^{s*} = i_{S1} + i_{S2} e^{-j\gamma} + i_{S3} e^{-j2\gamma}. \quad (3.293)$$

To explain the introduction of the complex current  $i_S^s$  further we note that in the generation of the magnetomotive force  $\mathcal{F}_S(\alpha, t)$  the three stator phases are rotated in the stator by an angle  $\gamma = 120^\circ$  relative to each other. Because of this, the influence of a current  $i_{S2}(t)$  in the second phase is the same as the influence of a current  $i_{S2}(t) e^{j\gamma}$  in the first phase. In the same way, the influence of a current  $i_{S3}(t)$  in the third phase is the same as the influence of a current  $i_{S3}(t) e^{j2\gamma}$  in the first phase. The combined influence of the three phases  $i_{S1}(t)$ ,  $i_{S2}(t)$  and  $i_{S3}(t)$  is therefore the same as the single complex current  $i_S^s$ .

The polar form of  $i_S^s$  is written

$$i_S^s = |i_S| e^{j\zeta} \quad (3.294)$$

where  $|i_S| = \sqrt{i_S^s i_S^{s*}}$  is the magnitude of  $i_S^s$  and  $\zeta$  is the angle of  $i_S^s$ . Using the polar form of the stator current we may write the magnetomotive force at the angular coordinate  $\alpha$  as

$$\mathcal{F}_S(\alpha, t) = \frac{1}{2} N_S \left[ |i_S| e^{j(\zeta-\alpha)} + |i_S| e^{-j(\zeta-\alpha)} \right] = N_S |i_S| \cos(\zeta - \alpha). \quad (3.295)$$

The currents of the rotor phases 1, 2 and 3 are denoted  $i_{R1}$ ,  $i_{R2}$  and  $i_{R3}$ , respectively. The radial magnetomotive force from the three-phase rotor windings is given at the angular coordinate  $\beta$  in the rotor as

$$\mathcal{F}_R(\beta, t) = N_R [i_{R1} \cos \beta + i_{R2} \cos(\beta - \gamma) + i_{R3} \cos(\beta - 2\gamma)] \quad (3.296)$$

where

$$\beta = \alpha - \theta. \quad (3.297)$$

The rotor currents  $i_{R1}$ ,  $i_{R2}$  and  $i_{R3}$  are represented by the complex variable  $i_R^r$  and its complex conjugate  $i_R^{r*}$  defined by

$$i_R^r = i_{R1} + i_{R2} e^{j\gamma} + i_{R3} e^{j2\gamma} = |i_R| e^{j\xi} \quad (3.298)$$

$$i_R^{r*} = i_{R1} + i_{R2} e^{-j\gamma} + i_{R3} e^{-j2\gamma} = |i_R| e^{-j\xi} \quad (3.299)$$

which gives

$$\mathcal{F}_R(\alpha, \theta, t) = \frac{1}{2} N_R \left[ i_R^r e^{-j(\alpha-\theta)} + i_R^{r*} e^{j(\alpha-\theta)} \right]. \quad (3.300)$$

The total radial magnetomotive force is then given by the sum

$$\mathcal{F} = \mathcal{F}_S(\alpha, t) + \mathcal{F}_R(\alpha, \theta, t) \quad (3.301)$$

As the magnetic permeability of the iron is much higher than the magnetic permeability for air, the magnetomotive force will be over the airgaps. The corresponding magnetic flux density at the airgap is

$$B_S(\alpha, \theta, t) = \frac{\mu_0}{2h} [\mathcal{F}_S(\alpha, t) + \mathcal{F}_R(\alpha, \theta, t)] \quad (3.302)$$

where  $\mu_0$  is the magnetic permeability of air.

The flux linkage of stator winding 1 is

$$\lambda_{S1}(t) = \frac{1}{2} N_S \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos \psi \int_{\psi - \frac{\pi}{2}}^{\psi + \frac{\pi}{2}} lr B_S(\alpha, \theta, t) d\alpha d\psi. \quad (3.303)$$

Integration gives, after some work,

$$\lambda_{S1}(t) = \frac{1}{3} L_S [i_S^s(t) + i_S^{s*}(t)] + \frac{1}{3} M [i_R^r(t) + i_R^{r*}(t)] \quad (3.304)$$

where

$$L_S = 3 \frac{N_S^2 lr \pi \mu_0}{8h}, \quad M = 3 \frac{N_S N_R lr \pi \mu_0}{8h} \quad (3.305)$$

Proceeding in the same way with the flux linkages  $\lambda_{S2}$  and  $\lambda_{S3}$  it is found that

$$\lambda_{S2}(t) = \frac{1}{3} L_S [i_S^s(t) e^{-j\gamma} + i_S^{s*}(t) e^{j\gamma}] + \frac{1}{3} M [i_R^r(t) e^{-j\gamma} + i_R^{r*}(t) e^{j\gamma}]$$

$$\lambda_{S3}(t) = \frac{1}{3} L_S [i_S^s(t) e^{-j2\gamma} + i_S^{s*}(t) e^{j2\gamma}] + \frac{1}{3} M [i_R^r(t) e^{-j2\gamma} + i_R^{r*}(t) e^{j2\gamma}]$$

Then, we may define the complex stator flux linkage  $\lambda_S^s$  by

$$\lambda_S^s = \lambda_{S1} + \lambda_{S2}e^{j\gamma} + \lambda_{S3}e^{j2\gamma}, \quad (3.306)$$

and we find that the complex stator flux linkage  $\lambda_S^s$  is given by the complex stator current  $i_S^s$  and the complex rotor current  $i_R^s$  according to

$$\lambda_S^s = L_S i_S^s + M i_R^s \quad (3.307)$$

Note that the complex rotor current  $i_R^s = i_R^r e^{j\theta}$  is given in the stator frame  $s$ .

In the same way the rotor flux linkages  $\lambda_{R1}$ ,  $\lambda_{R2}$  and  $\lambda_{R3}$  of the rotor phases 1, 2 and 3 are represented by the complex rotor flux variable  $\lambda_R^r$  defined by

$$\lambda_R^r = \lambda_{R1} + \lambda_{R2}e^{j\gamma} + \lambda_{R3}e^{j2\gamma} \quad (3.308)$$

The complex rotor flux linkage  $\lambda_R^r$  is given by the complex current variables  $i_R^r$  and  $i_S^r$  according to

$$\lambda_R^r = L_R i_R^r + M i_S^r \quad (3.309)$$

where

$$L_S = 3 \frac{N_R^2 lr \pi \mu_0}{8h} \quad (3.310)$$

and the complex stator current  $i_S^r = i_S^s e^{-j\theta}$  is given in the rotor frame  $r$ . The stator

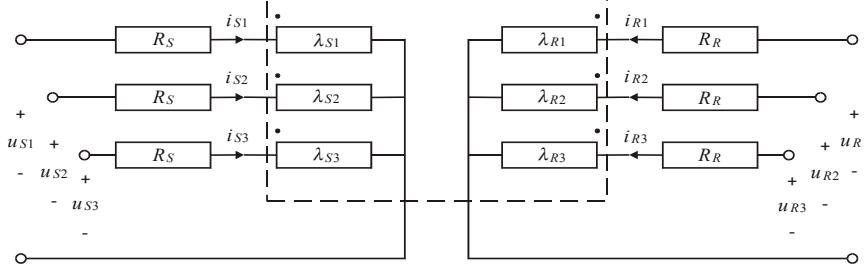


Figure 3.23: Stator and rotor circuits

and rotor circuits are shown in Figure 3.23. For each of the stator phases 1, 2 and 3 the voltage law leads to

$$R_S i_{S1} + \frac{d\lambda_{S1}}{dt} = u_{S1} \quad (3.311)$$

$$R_S i_{S2} + \frac{d\lambda_{S2}}{dt} = u_{S2} \quad (3.312)$$

$$R_S i_{S3} + \frac{d\lambda_{S3}}{dt} = u_{S3} \quad (3.313)$$

Likewise, the voltage law for each of the rotor phases gives

$$R_R i_{R1} + \frac{d\lambda_{R1}}{dt} = u_{R1} \quad (3.314)$$

$$R_R i_{R2} + \frac{d\lambda_{R2}}{dt} = u_{R2} \quad (3.315)$$

$$R_R i_{R3} + \frac{d\lambda_{R3}}{dt} = u_{R3} \quad (3.316)$$

Define the complex stator voltage

$$u_S^s = u_{S1} + u_{S2}e^{j\gamma} + u_{S3}e^{j2\gamma} \quad (3.317)$$

and the complex rotor voltage

$$u_R^r = u_{R1} + u_{R2}e^{j\gamma} + u_{R3}e^{j2\gamma} \quad (3.318)$$

The voltage laws for stator and rotor in terms of complex variables can be written

$$R_S i_S^s + \frac{d\lambda_S^s}{dt} = u_S^s \quad (3.319)$$

$$R_R i_R^r + \frac{d\lambda_R^r}{dt} = u_R^r. \quad (3.320)$$

where

$$\lambda_S^s = L_S i_S^s + M i_R^r, \quad \lambda_R^r = L_R i_R^r + M i_S^s \quad (3.321)$$

Note that the dynamics of the stator circuits are given in the stator-fixed frame  $s$ , while the dynamics of the rotor circuits are given in the rotor fixed frame  $r$ .

The motor torque is due to the Lorentz force on the rotor, which is caused by the rotor current flowing in the magnetic flux from the stator. The magnetic flux from the stator is

$$B_{RS}(\beta, \theta, t) = \frac{N_S \mu_0}{4h} \left[ i_S^s(t) e^{-j(\beta+\theta)} + i_S^{s*}(t) e^{j(\beta+\theta)} \right] \quad (3.322)$$

while the current density in the rotor is given by

$$a_R(\beta, t) = \frac{1}{2r} \frac{\partial \mathcal{F}_R}{\partial \beta} = -j \frac{N_R}{4r} [i_R^r e^{-j\beta} - i_R^{r*} e^{j\beta}] \quad (3.323)$$

The resulting torque is

$$dT = -r B_{RS} a_R l r d\beta \quad (3.324)$$

$$= \frac{j}{2\pi} \frac{M}{3} \left[ i_S^s e^{-j(\beta+\theta)} + i_S^{s*} e^{j(\beta+\theta)} \right] [i_R^r e^{-j\beta} - i_R^{r*} e^{j\beta}] d\beta \quad (3.325)$$

$$= \frac{j}{2\pi} \frac{M}{3} \left[ i_S^{s*} i_R^r e^{j\theta} - i_S^s i_R^{r*} e^{-j\theta} + i_S^s i_R^r e^{-j(2\beta+\theta)} - i_S^{s*} i_R^{r*} e^{j(2\beta+\theta)} \right] d\beta \quad (3.326)$$

which is integrated to

$$T = \frac{M}{3} \left[ \frac{i_S^s i_R^{r*} e^{-j\theta} - i_S^{s*}(t) i_R^r e^{j\theta}}{2j} \right] = \frac{2}{3} M \operatorname{Im} [i_S^s i_R^{s*}] \quad (3.327)$$

The dynamic model of a general AC motor is given by

$$R_S i_S^s + L_S \frac{di_S^s}{dt} + M \frac{d}{dt} (i_R^r e^{j\theta}) = u_S^s \quad (3.328)$$

$$R_R i_R^r + L_R \frac{di_R^r}{dt} + M \frac{d}{dt} (i_S^s e^{-j\theta}) = u_R^r \quad (3.329)$$

$$J \frac{d\omega_m}{dt} = \frac{2}{3} M \operatorname{Im} [i_S^s (i_R^r e^{j\theta})^*] - T_L \quad (3.330)$$

$$\frac{d\theta}{dt} = \omega_m \quad (3.331)$$

The general AC motor can be described as an electromechanical three-port that is connected to a mechanical rotary two-port. The electromechanical three-port has a stator port with effort  $u_S^s$  and flow  $i_S^s$ , a rotor port with effort  $u_R^r$  and flow  $i_R^r$ , and a mechanical port with effort  $T$  and flow  $\omega_m$ . The mechanical port is connected to the mechanical port which has effort  $T$  and flow  $\omega_m$  at the input port and effort  $T_L$  and flow  $-\omega_m$  at the input port.

## 3.10 Induction motors

### 3.10.1 Basic dynamic model

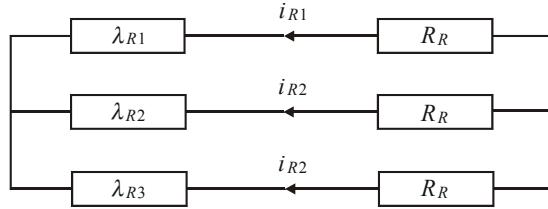


Figure 3.24: Rotor circuit of induction motor.

A type of induction motor that is widely used is the squirrel cage induction motor, which is designed so that the rotor circuits are short-circuited (Figure 3.24). The dynamic model is obtained by setting the rotor voltage  $u_R^r$  to zero in the general AC motor model.

The model for an induction motor with short-circuited rotor is give by

$$R_S i_S^s + L_S \frac{di_S^s}{dt} + M \frac{d}{dt} (i_R^r e^{j\theta}) = u_S^s \quad (3.332)$$

$$R_R i_R^r + L_R \frac{di_R^r}{dt} + M \frac{d}{dt} (i_S^s e^{-j\theta}) = 0 \quad (3.333)$$

$$J \frac{d\omega_m}{dt} = \frac{2}{3} M \operatorname{Im} [i_S^s (i_R^r e^{j\theta})^*] - T_L \quad (3.334)$$

$$\frac{d\theta}{dt} = \omega_m \quad (3.335)$$

In a network description the induction motor model is obtained from the general AC motor model by introducing a short circuit of the rotor port.

### 3.10.2 Induction motor model in stator frame

An alternative model formulation that is often seen for induction motors is obtained by a change of variables from current  $i_R^r$  to flux linkage  $\lambda_R^s$  in the stator equation (3.332) and the rotor equation (3.333). The resulting model, which is called the *ab*-model is given in the stator frame in terms of stator currents  $i_S^s$  and rotor flux  $\lambda_R^s$ . The model is derived from (3.319) and (3.320). First it is noted that (3.320) implies that

$$R_R i_R^s + \frac{d}{dt} (\lambda_R^s e^{-j\theta}) e^{j\theta} = 0 \quad (3.336)$$

where the equation is expressed in the  $s$  frame, and where it is used that  $u_R^r = 0$ . Differentiation gives

$$R_R i_R^s + \frac{d\lambda_R^s}{dt} - j\omega_m \lambda_R^s = 0 \quad (3.337)$$

Next, insertion of (3.307) into (3.319) gives

$$R_S i_S^s + \frac{d}{dt} (L_S i_S^s + M i_R^s) = u_S^s \quad (3.338)$$

From (3.309) it follows that

$$i_R^s = i_R^r e^{j\theta} = \frac{1}{L_R} (\lambda_R^s - M i_S^s) \quad (3.339)$$

and, combining this with (3.337), it is found that

$$\begin{aligned} \frac{di_R^s}{dt} &= \frac{1}{L_R} \left( \frac{d\lambda_R^s}{dt} - M \frac{di_S^s}{dt} \right) = \frac{1}{L_R} \left( j\omega_m \lambda_R^s - R_R i_R^s - M \frac{di_S^s}{dt} \right) \\ &= \frac{1}{L_R} \left[ \left( j\omega_m - \frac{R_R}{L_R} \right) \lambda_R^s + \frac{R_R}{L_R} M i_S^s - M \frac{di_S^s}{dt} \right] \end{aligned} \quad (3.340)$$

Insertion of these results into (3.338) and (3.337) gives

$$R_S i_S^s + L_S \frac{di_S^s}{dt} + \frac{M}{L_R} \left[ \left( j\omega_m - \frac{R_R}{L_R} \right) \lambda_R^s + \frac{R_R}{L_R} M i_S^s - M \frac{di_S^s}{dt} \right] = u_S^s \quad (3.341)$$

$$\frac{R_R}{L_R} (\lambda_R^s - M i_S^s) + \frac{d\lambda_R^s}{dt} - j\omega_m \lambda_R^s = 0 \quad (3.342)$$

which is simplified to

$$L_S \sigma \frac{di_S^s}{dt} = - \left( R_S + \frac{R_R M^2}{L_R^2} \right) i_S^s - \frac{M}{L_R} \left( j\omega_m - \frac{R_R}{L_R} \right) \lambda_R^s + u_S^s \quad (3.343)$$

$$\frac{d\lambda_R^s}{dt} = - \frac{R_R}{L_R} \lambda_R^s + j\omega_m \lambda_R^s + \frac{R_R M}{L_R} i_S^s \quad (3.344)$$

where

$$\sigma = 1 - \frac{M^2}{L_S L_R} \quad (3.345)$$

The torque can be expressed

$$\begin{aligned} T &= \frac{2}{3} M \operatorname{Im} [i_S^s (i_R^r e^{j\theta})^*] = \frac{2}{3} \frac{M}{L_R} \operatorname{Im} [i_S^s (\lambda_R^r e^{j\theta})^* + M i_S^s (i_S^s)^*] \\ &= \frac{2}{3} \frac{M}{L_R} \operatorname{Im} [i_S^s (\lambda_R^r e^{j\theta})^*] = \frac{2}{3} \frac{M}{L_R} \operatorname{Im} [i_S^s \lambda_R^s]^* \end{aligned} \quad (3.346)$$

where it is used that  $\operatorname{Im} [i_S^s (i_S^s)^*] = 0$ . The equation of motion is then

$$J \frac{d\omega_m}{dt} = \frac{2}{3} \frac{M}{L_R} \operatorname{Im} [i_S^s (\lambda_R^r e^{j\theta})^*] - T_L$$

The coordinate form of this model is found by inserting

$$i_S^s = i_a + j i_b, \quad u_S^s = u_a + j u_b, \quad \lambda_R^s = \lambda_a + j \lambda_b, \quad (3.347)$$

This gives the *ab*-model for an induction motor:

$$\frac{di_a}{dt} = -\frac{L_R^2 R_S + M^2 R_R}{L_R^2 L_S \sigma} i_a + \frac{R_R \kappa}{L_R} \lambda_a + \kappa \omega_m \lambda_b + \frac{1}{L_S \sigma} u_a \quad (3.348)$$

$$\frac{di_b}{dt} = -\frac{L_R^2 R_S + M^2 R_R}{L_R^2 L_S \sigma} i_b + \frac{R_r \kappa}{L_R} \lambda_b - \kappa \omega_m \lambda_a + \frac{1}{L_S \sigma} u_b \quad (3.349)$$

$$\frac{d\lambda_a}{dt} = -\frac{R_R}{L_R} \lambda_a - \omega_m \lambda_b + \frac{R_R M}{L_R} i_a \quad (3.350)$$

$$\frac{d\lambda_b}{dt} = -\frac{R_R}{L_R} \lambda_b + \omega_m \lambda_a + \frac{R_R M}{L_R} i_b \quad (3.351)$$

$$J_m \dot{\omega}_m = \mu (\lambda_a i_b - \lambda_b i_a) - T_L \quad (3.352)$$

where

$$\sigma = 1 - \frac{M^2}{L_S L_R}, \quad \kappa = \frac{M}{\sigma L_S L_R}, \quad \mu = \frac{2}{3} \frac{M}{L_R} \quad (3.353)$$

This model has no trigonometric terms.

### 3.10.3 Dynamic model in the flux frame

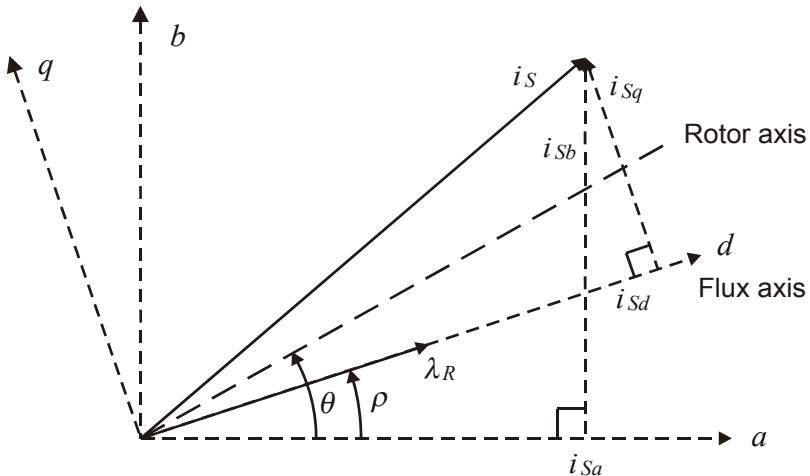


Figure 3.25: The flux frame  $dq$ .

A model formulation that is of great practical importance is the  $dq$ -model which is the basis for field-oriented control of induction motors. This model has strong similarities with the dynamic model of the DC motor, and is particularly suited for the design of servo controllers for induction motors. To arrive at the  $dq$ -model the rotor flux linkage is written in polar form

$$\lambda_R^s = \lambda_R^r e^{j\theta} = |\lambda_R| e^{j\rho} \quad (3.354)$$

where  $|\lambda_R|$  is the magnitude of the rotor flux linkage. We define the rotor flux frame  $f$  as the frame rotated by an angle  $\rho$  relative to the stator frame  $s$  (Figure 3.25). This frame

is also referred to as the  $dq$  frame. Here  $d$  refers to the direct axis, which is the  $x_f$  axis, and  $q$  refers to the quadrature axis, which is the  $y_f$  axis. The rotor flux linkage in the  $dq$  frame is then purely real, which is seen from

$$\lambda_R^f = \lambda_R^r e^{j\theta} e^{-j\rho} = |\lambda_R| \quad (3.355)$$

The stator current and voltage in the  $dq$  frame are written

$$i_S^f = i_S^s e^{-j\rho} = i_{Sd} + j i_{Sq} \quad (3.356)$$

$$u_S^f = u_S^s e^{-j\rho} = u_{Sd} + j u_{Sq} \quad (3.357)$$

The torque is then

$$T = \frac{2}{3} \frac{M}{L_R} \operatorname{Im} \left[ i_S^f (\lambda_R^f)^* \right] = \frac{2}{3} \frac{M}{L_R} |\lambda_R| i_{Sq}. \quad (3.358)$$

This expression for the torque is very interesting as it resembles the torque expression for a DC motor. In particular, it is seen that if the field represented by the rotor flux linkage  $|\lambda_R|$  can be controlled to a constant, say  $|\lambda_R| = \Lambda$ , then the torque will be proportional to the stator current component  $i_{Sq}$  along the quadrature axis  $y_f$ . To investigate this further we develop a model of the induction motor in the  $dq$  frame. The rotor flux linkage can be expressed by the rotor and stator currents by rotating (3.309) to the flux frame  $f$ , which gives

$$\lambda_R^f = L_R i_R^f + M i_S^f \quad (3.359)$$

This expression is combined with the voltage law

$$R_R i_R^r + \frac{d\lambda_R^r}{dt} = 0 \quad (3.360)$$

for the rotor windings. In the  $f$  frame this gives

$$\frac{d(\lambda_R^f e^{-j(\theta-\rho)})}{dt} e^{j(\theta-\rho)} + \frac{R_R}{L_R} (\lambda_R^f - M i_S^f) = 0, \quad (3.361)$$

which leads to the following differential equation for the rotor flux linkage:

$$\frac{d\lambda_R^f}{dt} + j(\rho - \omega_m) \lambda_R^f + \frac{R_R}{L_R} (\lambda_R^f - M i_S^f) = 0. \quad (3.362)$$

The real and imaginary part of this equation gives

$$\frac{L_R}{R_R} \frac{d|\lambda_R|}{dt} + |\lambda_R| = M i_{Sd} \quad (3.363)$$

$$\dot{\rho} = \omega_m + \frac{MR_R}{L_R} \frac{i_{Sq}}{|\lambda_R|} \quad (3.364)$$

Note that the second equation is singular when the rotor flux is zero, that is, for  $|\lambda_R| = 0$ .

The stator flux equation is

$$R_S i_S^s + \frac{d}{dt} (L_S i_S^s + M i_R^s) = u_S^s \quad (3.365)$$

The rotor current is eliminated, which gives

$$R_S i_S^s + \frac{d}{dt} \left[ L_S \sigma i_S^s + \frac{M}{L_R} \lambda_R^s \right] = u_S^s \quad (3.366)$$

where

$$\sigma = 1 - \frac{M^2}{L_S L_R} \quad (3.367)$$

In the  $f$  frame this gives

$$R_S i_S^f + L_S \sigma \left( \frac{di_S^f}{dt} + j \dot{\rho} i_S^f \right) + \frac{M}{L_R} \frac{d\lambda_R^f}{dt} + j \dot{\rho} \frac{M}{L_R} \lambda_R^f = u_S^f \quad (3.368)$$

The real and imaginary part of this equation is

$$L_S \sigma \frac{di_{Sd}}{dt} = -R_S i_{Sd} + u_{Sd} - \frac{M}{L_R} \frac{d|\lambda_R|}{dt} + L_S \sigma \dot{\rho} i_{Sq} \quad (3.369)$$

$$L_S \sigma \frac{di_{Sq}}{dt} = -R_S i_{Sq} + u_{Sq} - \frac{M}{L_R} \dot{\rho} |\lambda_R| - L_S \sigma \dot{\rho} i_{Sd} \quad (3.370)$$

A reasonable assumption is that the rotor flux linkage is slowly changing compared to the stator current, and this expression is therefore not developed further by inserting expressions for the derivative of the rotor flux linkage.

The  $dq$  model of an induction motor is given by

$$L_S \sigma \frac{di_{Sd}}{dt} = -R_S i_{Sd} + u_{Sd} - \frac{M}{L_R} \frac{d|\lambda_R|}{dt} + L_S \sigma \dot{\rho} i_{Sq} \quad (3.371)$$

$$L_S \sigma \frac{di_{Sq}}{dt} = -R_S i_{Sq} + u_{Sq} - \frac{M}{L_R} \dot{\rho} |\lambda_R| - L_S \sigma \dot{\rho} i_{Sd} \quad (3.372)$$

$$\frac{L_R}{R_R} \frac{d|\lambda_R|}{dt} + |\lambda_R| = M i_{Sd} \quad (3.373)$$

$$\dot{\rho} = \omega_m + \frac{MR_R}{L_R} \frac{i_{Sq}}{|\lambda_R|} \quad (3.374)$$

$$J_m \dot{\omega}_m = \mu |\lambda_R| i_{Sq} - T_L \quad (3.375)$$

We can interpret this in the following way: The stator current component  $i_{Sd}$  along the direct axis is controlled by the voltage  $u_{Sd}$ , and is used to magnetize the motor, that is, to control the flux linkage  $|\lambda_R|$ . The stator current component  $i_{Sq}$  along the quadrature axis is controlled by the quadrature voltage component  $u_{Sq}$ , and is used to control the torque  $T_m$ . This is the underlying principle of *field-oriented control*. In field-oriented control the magnitude  $|\lambda_R|$  of the rotor flux is controlled to a constant value  $\Lambda$  with the controllers

$$i_{Sd,ref} = K_{id}(s) [\Lambda - |\lambda_R|(s)] \quad (3.376)$$

$$u_{Sd} = K_{ud}(u) [i_{Sd,ref} - i_{Sd}] \quad (3.377)$$

Then the torque can be controlled to a desired value  $\tau_d$  by controlling the quadrature current

$$u_{Sq} = K_{uq}(u) [i_{Sq,ref} - i_{Sq}] \quad (3.378)$$

where  $i_{Sq,ref} = \tau_d / (\mu \Lambda)$ . In addition, the flux angle  $\rho$  must be calculated in some way. Solutions to this problem are discussed in great detail in (Leonhard 1996).

**Example 49** The dq model of the induction motor can be described as a electromechanical three-port with a quadrature axis port with voltage  $u_{Sq}$  and current  $i_{Sq}$ , a direct axis port with voltage  $u_{Sd}$  and current  $i_{Sd}$ , and a mechanical rotary port with effort  $T$  and flow  $\omega_m$ . The direct port is used to control the flux in the machine, while the quadrature port is used to control the torque.

## 3.11 Lagrangian description of electromechanical systems

### 3.11.1 Generalized coordinates

The dynamics of electrical motors involves the dynamics of an electrical system and a mechanical system. The mechanical system can be modeled with the Lagrange's equation of motion, and this is quite useful for analysis and controller design. It turns out that also the electrical system can be modelled using a Lagrange formalism, and this description can be integrated with the Lagrangian description of the mechanical system. A detailed discussion of this is found in (Meisel 1966) and (Crandall et al. 1968), and the material in this section is based on these two references.

Lagrangian models for mechanical systems (Chapter 8 relies on the definition of generalized coordinates  $q_i$  that may be positions or angles. In addition the generalized momenta

$$p_i = \frac{\partial L}{\partial \dot{q}_i} \quad (3.379)$$

are used, where  $L$  is the Lagrangian. In electrical systems there are two alternatives for generalized coordinates. One is the electrical charge  $q$  of a capacitive element, and the other is the flux linkage  $\lambda$  of an inductive element. The time derivative of the electrical charge is a current, and the time derivative of a flux linkage is a voltage, so we may write

$$\dot{q} = i \quad \text{and} \quad \dot{\lambda} = u \quad (3.380)$$

### 3.11.2 Energy and coenergy

The power input to a circuit of  $n$  elements can be written

$$P = \sum_{k=1}^n i_k u_k = \dot{q}_k \dot{\lambda}_k \quad (3.381)$$

where  $\dot{q}_k = i_k$  is the current through element  $k$ , and  $\dot{\lambda}_k = u_k$  is the voltage over element  $k$ . The energy stored in the circuit is

$$W = \sum_{k=1}^n \int_{t_0}^t \dot{q}_k \dot{\lambda}_k dt \quad (3.382)$$

Suppose that circuit element  $k$  is capacitive, and that the voltage  $\dot{\lambda}_k$  can be given as a function of the charge  $q_k$  according to

$$\dot{\lambda}_k = \dot{\lambda}_k(q_k) \quad (3.383)$$

Then, using  $\dot{q}_k dt = dq_k$ , the energy of element  $k$  can be written

$$W_{ck}(q_k) = \int_{q_k(t_0)}^{q_k} \dot{\lambda}_k(q'_k) dq'_k \quad (3.384)$$

where  $q'_k$  is the dummy variable used in the integration. The coenergy of element  $k$  is

$$W_{ck}^* (\dot{\lambda}_k) = \int_{\dot{\lambda}_k(t_0)}^{\dot{\lambda}_k} q_k (\dot{\lambda}'_k) d\dot{\lambda}'_k \quad (3.385)$$

where  $\dot{\lambda}'_k$  is the dummy variable in the integration. Then the following expressions hold

$$\frac{\partial W_{ck}(q_k)}{\partial q_k} = \dot{\lambda}_k \quad \text{and} \quad \frac{\partial W_{ck}^*(\dot{\lambda}_k)}{\partial \dot{\lambda}_k} = q_k \quad (3.386)$$

Next, suppose that circuit element  $i$  is inductive, and that the current can be written as a function

$$\dot{q}_i = \dot{q}_i(\lambda_i) \quad (3.387)$$

of the flux linkage  $\lambda_i$ . Then the energy is of an inductive element is

$$W_{mi}(\lambda_i) = \int_{\lambda_i(t_0)}^{\lambda_i} \dot{q}_i(\lambda'_i) d\lambda'_i \quad (3.388)$$

Define the *coenergy* of an inductive element by

$$W_{mi}^*(\dot{q}_i) = \int_{\dot{q}_i(t_0)}^{\dot{q}_i} \lambda_i(\dot{q}'_i) d\dot{q}'_i \quad (3.389)$$

The flux linkage and the current are then given by

$$\frac{\partial W_{mi}(\lambda_i)}{\partial \dot{q}_i} = \lambda_i \quad \text{and} \quad \frac{\partial W_{mi}^*(\dot{q}_i)}{\partial \lambda_i} = \dot{q}_i \quad (3.390)$$

The energy  $W_c(\mathbf{q})$  and the coenergy  $W_c^*(\dot{\lambda})$  of the capacitive elements of the circuit are given by

$$W_c(\mathbf{q}) = \sum_k W_{ck}(q_k), \quad W_c^*(\dot{\lambda}) = \sum_k W_{ck}^*(\dot{\lambda}_k) \quad (3.391)$$

where the summation is done over the capacitive elements of the circuit. The energy  $W_m(\boldsymbol{\lambda})$  and the coenergy  $W_m^*(\dot{\mathbf{q}})$  of the circuit are given by

$$W_m(\boldsymbol{\lambda}) = \sum_i W_{mi}(\lambda_i), \quad W_m^*(\dot{\mathbf{q}}) = \sum_i W_{mi}^*(\dot{q}_i) \quad (3.392)$$

where the summation is done over the inductive elements of the circuit.

### 3.11.3 Analogy of electrical and mechanical systems

In a translational mechanical system the effort is the force  $F$  and the flow is the velocity  $\dot{x}$ . In an electrical circuit the usual analog of a mechanical system is obtained using the voltage  $u$  as the effort variable, and the current  $\dot{q}$  as the flow variable. This gives electrical analog 1 where a mass corresponds to an inductance, a viscous damper corresponds to a resistor, and a spring corresponds to a capacitor as in Table 3.1.

An alternative electrical analog appears if the current is used as the effort variable, and the voltage is taken as the flow variable. This gives electrical analog 2 where a mass corresponds to a capacitor, a viscous damper corresponds to a resistor, and a spring corresponds to an inductance as in Table 3.2.

	Translational		Electrical analog 1	
	Effort	Energy	Effort	Energy
Mass	$F = m\ddot{x}$	$K = \frac{1}{2}m\dot{x}^2$	$u = L\ddot{q}$	$W_m^* = \frac{1}{2}L\dot{q}^2$
Damper	$F = B\dot{x}$	0	$u = R\dot{q}$	0
Spring	$F = Kx$	$V = \frac{1}{2}Kx^2$	$u = Cq$	$W_c = \frac{1}{2C}q^2$

Table 3.1: Electrical analog 1.

	Translational		Electrical analog 2	
	Effort	Energy	Effort	Energy
Mass	$F = m\ddot{x}$	$K = \frac{1}{2}m\dot{x}^2$	$i = \frac{1}{C}\ddot{\lambda}$	$W_c^* = \frac{1}{2}C\dot{\lambda}^2$
Damper	$F = B\dot{x}$	0	$i = \frac{1}{R}\dot{\lambda}$	0
Spring	$F = Kx$	$V = \frac{1}{2}Kx^2$	$i = \frac{1}{L}\lambda$	$W_m = \frac{1}{2L}\lambda^2$

Table 3.2: Electrical analog 2.

Suppose that electrical analog 1 is used, which means that the charge  $q$  is taken as a generalized coordinate of the electrical system. Then the energy stored in an inductance is analog to the kinetic energy of a mass, and the energy stored in a capacitor is the analog of the potential energy in a spring. On the equation level there is no distinction between mechanical components and electrical components, and therefore Lagrangian dynamics can be formulated also for electrical systems by using the analogies that have been pointed out.

### 3.11.4 The Lagrangian

With electrical analog 1 the voltages are given as functions of the charge and current. Then the generalized coordinate vector is the charge vector  $\mathbf{q}$ , and the Lagrangian can be defined by

$$L_e(\mathbf{q}, \dot{\mathbf{q}}) = W_m^*(\dot{\mathbf{q}}) - W_c(\mathbf{q}), \quad (3.393)$$

which is the coenergy of the inductive elements minus the energy of the capacitive elements. When electrical analog 2 is used the currents are given as functions of the flux linkage and the voltages. Then the generalized coordinate vector is the flux linkage vector  $\boldsymbol{\lambda}$ , and the Lagrangian is

$$L_e(\boldsymbol{\lambda}, \dot{\boldsymbol{\lambda}}) = W_c^*(\dot{\boldsymbol{\lambda}}) - W_m(\boldsymbol{\lambda}), \quad (3.394)$$

which is the coenergy of the capacitive elements minus the energy of the inductive elements.

The generalized forces  $Q_i$  corresponding to the generalized coordinate in the form of an electrical charge  $q_i$  is found from a power consideration. The power a generalized force  $Q_i$  is  $P_i = Q_i \dot{q}_i$ , which shows that the generalized force must be the voltage over element  $i$ , that is,

$$Q_i = u_i. \quad (3.395)$$

The generalized force  $\Lambda_k$  corresponding to a generalized coordinate in the form of a flux linkage  $\lambda_k$  is found from  $P_k = \Lambda_k \dot{\lambda}_k$ , and it follows that the generalized force  $\Lambda_k$  must be the current through element  $k$ , that is

$$\Lambda_k = i_k. \quad (3.396)$$

### 3.11.5 Electromechanical systems

For an electromechanical system the Lagrangian  $L$  for the complete system turns out to be the sum of the Lagrangian for the electrical system and the Lagrangian of the mechanical system. Suppose that the charge vector  $\mathbf{q}_e$  is the generalized coordinate vector of the electrical system. The generalized coordinate vector  $\mathbf{q}$  and the generalized force vector  $\mathbf{Q}$  for the total system are then taken to be

$$\mathbf{q} = \begin{pmatrix} \mathbf{q}_e \\ \mathbf{q}_m \end{pmatrix}, \quad \mathbf{Q} = \begin{pmatrix} \mathbf{Q}_e \\ \boldsymbol{\tau} \end{pmatrix} \quad (3.397)$$

where  $\mathbf{Q}_e$  are the generalized forces corresponding to the electrical coordinates  $\mathbf{q}_e$ , and  $\boldsymbol{\tau}$  are the generalized forces corresponding to the mechanical coordinates  $\mathbf{q}_m$ . We define the kinetic energy term to be

$$T(\mathbf{q}, \dot{\mathbf{q}}) = T_m(\mathbf{q}_m, \dot{\mathbf{q}}_m) + W_m^*(\dot{\mathbf{q}}_e, \mathbf{q}_m) \quad (3.398)$$

which is the sum of the kinetic energy of the mechanical system and the coenergy of the inductive elements of the electrical system. The potential energy term in the Lagrangian is defined to be

$$V(\mathbf{q}) = V_m(\mathbf{q}_m) + W_c(\mathbf{q}_e, \mathbf{q}_m) \quad (3.399)$$

which is the sum of the potential energy of the mechanical system and the potential energy of the capacitive elements of the electrical system. Then the Lagrangian for the total system is

$$L(\mathbf{q}, \dot{\mathbf{q}}) = T(\mathbf{q}, \dot{\mathbf{q}}) - V(\mathbf{q}) \quad (3.400)$$

The dynamics are given by

$$\frac{d}{dt} \frac{\partial L}{\partial \dot{\mathbf{q}}_e}^T - \frac{\partial L}{\partial \mathbf{q}_e} = \mathbf{Q} \quad (3.401)$$

$$\frac{d}{dt} \frac{\partial L}{\partial \dot{\mathbf{q}}_m}^T - \frac{\partial L}{\partial \mathbf{q}_m} = \boldsymbol{\tau} \quad (3.402)$$

These expressions can be further developed into

$$\frac{d}{dt} \left( \frac{\partial W_m^*(\dot{\mathbf{q}}_e, \mathbf{q}_m)}{\partial \dot{\mathbf{q}}_e} \right)^T + \frac{\partial W_c(\mathbf{q}_e, \mathbf{q}_m)}{\partial \mathbf{q}_e}^T = \mathbf{Q} \quad (3.403)$$

and

$$\begin{aligned} \frac{d}{dt} \left( \frac{\partial T_m(\mathbf{q}_m, \dot{\mathbf{q}}_m)}{\partial \dot{\mathbf{q}}_m} \right)^T - \frac{\partial T_m(\mathbf{q}_m, \dot{\mathbf{q}}_m)}{\partial \mathbf{q}_m}^T - \frac{\partial W_m^*(\dot{\mathbf{q}}_e, \mathbf{q}_m)}{\partial \mathbf{q}_m}^T \\ + \frac{\partial W_c(\mathbf{q}_e, \mathbf{q}_m)}{\partial \mathbf{q}_m}^T + \frac{\partial V_m(\mathbf{q}_m)}{\partial \mathbf{q}_m}^T = \boldsymbol{\tau} \end{aligned} \quad (3.404)$$

The term  $\partial W_m^*/\partial \mathbf{q}_m$  is the change in the electrical coenergy  $W_m^*$  due to a change in the mechanical configuration  $\mathbf{q}_m$ . This is the generalized interaction force on the mechanical system due to electromagnetic forces.

**Example 50** We consider an electrical motor with shaft angle  $q_m$  and inertia  $J_m$ . The kinetic energy of the shaft is given by  $T_m(\dot{q}_m) = \frac{1}{2}J_m\dot{q}_m^2$ , and the potential energy is  $V(\mathbf{q}) = 0$ . The coenergy of the inductive elements of the motor is given by

$$W_m^*(\dot{\mathbf{q}}_e, q_m) = \frac{1}{2}\dot{\mathbf{q}}_e^T \mathbf{M}_e(q_m) \dot{\mathbf{q}}_e \quad (3.405)$$

and the Lagrangian is  $L = T_m(\dot{q}_m) + W_m^*(\dot{\mathbf{q}}_e, q_m)$ . Then the dynamics of the motor is given by

$$\mathbf{M}_e(q_m) \ddot{\mathbf{q}}_e + \dot{q}_m \frac{\partial \mathbf{M}_e(q_m)}{\partial q_m} \dot{\mathbf{q}}_e = \mathbf{u} \quad (3.406)$$

$$J_m \ddot{q}_m - \frac{1}{2} \dot{\mathbf{q}}_e^T \frac{\partial \mathbf{M}_e(q_m)}{\partial q_m} \dot{\mathbf{q}}_e = \tau \quad (3.407)$$

where the input voltage  $\mathbf{u}$  is the generalized force corresponding to the generalized coordinates, and  $\tau$  is the external torque acting on the motor. The coupling between the electrical part and the mechanical part gives rise to a torque

$$\tau_m = \frac{1}{2} \dot{\mathbf{q}}_e^T \frac{\partial \mathbf{M}_e(q_m)}{\partial q_m} \dot{\mathbf{q}}_e \quad (3.408)$$

delivered from the motor, and an induced voltage

$$\mathbf{e} = \dot{q}_m \frac{\partial \mathbf{M}_e(q_m)}{\partial q_m} \dot{\mathbf{q}}_e \quad (3.409)$$

In terms of the energy  $T = T_m(\dot{q}_m) + W_m^*(\dot{\mathbf{q}}_e, q_m)$ , the coupling terms balance the energy exchange between the electrical part and the mechanical part along the solutions of the system. This is seen from

$$\begin{aligned} \dot{T} &= \dot{q}_m J_m \ddot{q}_m + \dot{\mathbf{q}}_e^T \mathbf{M}_e(q_m) \ddot{\mathbf{q}}_e + \frac{1}{2} \dot{q}_m \dot{\mathbf{q}}_e^T \frac{\partial \mathbf{M}_e(q_m)}{\partial q_m} \dot{\mathbf{q}}_e \\ &= \dot{q}_m \tau + \dot{\mathbf{q}}_e^T \mathbf{u} \end{aligned} \quad (3.410)$$

and we may conclude that if the mechanical port with effort  $\tau_m$  and flow  $\dot{q}_m$  is terminated with a passive mechanical one-port, then the system with input  $\mathbf{u}$  and output  $\dot{\mathbf{q}}_e$  is passive.

**Example 51** The electrical motor in the previous example can be described as a passive electrical two-port in series with a passive electromechanical two-port. The input port of the electrical two-port has variables  $\mathbf{u}$  and  $\dot{\mathbf{q}}_e$  and output port with variables  $\mathbf{e}$  and  $-\dot{\mathbf{q}}_e$ . The electromechanical two-port has input port with variables  $\mathbf{e}$  and  $-\dot{\mathbf{q}}_e$  and output port with variables  $\tau_m$  and  $\dot{q}_m$ .

**Example 52** The flux linkage for the motor in the previous examples is given by

$$\lambda = \frac{\partial W_m^*(\dot{\mathbf{q}}_e, q_m)}{\partial \dot{\mathbf{q}}_e} = \mathbf{M}_e(q_m) \dot{\mathbf{q}}_e \quad (3.411)$$

**Example 53** The Lagrangian of a mechanical system is actually defined in (Meisel 1966) and (Crandall et al. 1968) as  $L = T^* - V$  where  $T^*$  is the kinetic coenergy. This is based on the application of Newton's law in the form  $F = \dot{p}$  where  $p = p(v)$  is the momentum.

The power is  $P = Fv = \dot{p}v$ , the differential of the kinetic energy is  $dT = Pdt = vdp$ , and the kinetic energy is

$$T(p) = \int_0^p v(p') dp' \quad (3.412)$$

The kinetic coenergy is then

$$T^*(v) = \int_0^v p(v') dv' \quad (3.413)$$

In the linear case where  $p = mv$ , the kinetic energy and the kinetic coenergy have same numerical values as

$$T(p) = \frac{1}{2} \frac{p^2}{m} = \frac{1}{2} mv^2 = T^*(v) \quad (3.414)$$

Therefore, the kinetic coenergy  $T^*(v)$  is referred to as the kinetic energy and is denoted by  $T(v)$ . However, in a relativistic setting where  $m(v) = m_0/\sqrt{1 - (v^2/c^2)}$  there is a difference between kinetic energy and kinetic coenergy.

### 3.11.6 Lagrange formulation of general AC motor

In this section we will derive the model of the general AC using Lagrange's equation of motion in vector notation. The charge vector  $\mathbf{q}_e$ , the current vector  $\dot{\mathbf{q}}_e$  and the voltage vector  $\mathbf{u}$  are given by

$$\mathbf{q}_e = \begin{pmatrix} \mathbf{q}_S^s \\ \mathbf{q}_R^r \end{pmatrix}, \quad \dot{\mathbf{q}}_e = \begin{pmatrix} \mathbf{i}_S^s \\ \mathbf{i}_R^r \end{pmatrix} \quad \text{and} \quad \mathbf{u} = \begin{pmatrix} \mathbf{u}_S^s \\ \mathbf{u}_R^r \end{pmatrix} \quad (3.415)$$

Note that the stator charge vector  $\mathbf{q}_S^s$  is given in the stator frame  $s$ , and that the rotor charge vector  $\mathbf{q}_R^r$  is given in the rotor frame  $r$ . The reason for this is that the stator charge is in the stator, while the rotor charge rotates with the rotor. If the definition of current as the time derivative of charge is to be meaningful, it is necessary to take the time derivative of the charge in the frame where the charge is flowing. Therefore the currents must be defined by

$$\mathbf{i}_S^s = \frac{d}{dt}(\mathbf{q}_S^s) = \dot{\mathbf{q}}_S^s, \quad \mathbf{i}_R^r = \frac{d}{dt}(\mathbf{q}_R^r) =: \dot{\mathbf{q}}_R^r \quad (3.416)$$

which is in agreement with (3.415).

We consider an electrical motor with shaft angle  $q_m$  and inertia  $J_m$ . The kinetic energy of the shaft is given by  $T_m(\dot{q}_m) = \frac{1}{2} J_m \dot{q}_m^2$ . The coenergy of the inductive elements of the motor is given by

$$W_m^*(\dot{\mathbf{q}}_e, q_m) = \frac{1}{2} \dot{\mathbf{q}}_e^T \mathbf{M}_e(q_m) \dot{\mathbf{q}}_e \quad (3.417)$$

where the inductance matrix  $\mathbf{M}_e(q_m)$  is

$$\mathbf{M}_e(q_m) = \begin{pmatrix} L_S \mathbf{I}_2 & M \mathbf{R}_2(q_m) \\ M \mathbf{R}_2(-q_m) & L_R \mathbf{I}_2 \end{pmatrix} \quad (3.418)$$

and

$$\mathbf{R}_2(q_m) = \begin{pmatrix} \cos q_m & -\sin q_m \\ \sin q_m & \cos q_m \end{pmatrix} \quad (3.419)$$

is the  $2 \times 2$  rotation matrix corresponding to a rotation by an angle  $q_m$ . We note that

$$\frac{\partial \mathbf{R}_2(q_m)}{\partial q_m} = \mathbf{S} \mathbf{R}_2(q_m), \quad \dot{\mathbf{R}}_2(q_m) = \dot{q}_m \mathbf{S} \mathbf{R}_2(q_m) \quad (3.420)$$

where

$$\mathbf{S} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad (3.421)$$

The partial derivative of the inductance matrix with respect to the shaft angle is

$$\mathbf{P}(q_m) := \frac{\partial \mathbf{M}_e(q_m)}{\partial q_m} = \begin{pmatrix} \mathbf{0} & M\mathbf{S}\mathbf{R}_2(q_m) \\ -M\mathbf{S}\mathbf{R}_2(-q_m) & \mathbf{0} \end{pmatrix} \quad (3.422)$$

The Lagrangian is equal to the sum of kinetic energy and electrical coenergy:

$$L = T(\dot{\mathbf{q}}_e, q_m, \dot{q}_m) = T_m(q_m) + W_m^*(\dot{\mathbf{q}}_e, q_m) \quad (3.423)$$

Then the dynamics of the motor is given by

$$\mathbf{M}_e(q_m)\ddot{\mathbf{q}}_e + \mathbf{P}(q_m)\dot{q}_m\dot{\mathbf{q}}_e + \mathbf{R}_e\dot{\mathbf{q}}_e = \mathbf{u} \quad (3.424)$$

$$J\ddot{q}_m - \frac{1}{2}\dot{\mathbf{q}}_e^T \mathbf{P}(q_m)\dot{\mathbf{q}}_e = -\tau_L \quad (3.425)$$

**Example 54** The dynamic equations for the electrical part can be written out as

$$L_S \frac{d\mathbf{i}_S^s}{dt} + M\mathbf{R}_2(q_m) \frac{d\mathbf{i}_S^r}{dt} + M\dot{q}_m \mathbf{S}\mathbf{R}_2(q_m) \mathbf{i}_R^r + R_S \mathbf{i}_S^s = \mathbf{u}_S^s \quad (3.426)$$

$$M\mathbf{R}_2(-q_m) \frac{d\mathbf{i}_S^s}{dt} + L_R \frac{d\mathbf{i}_R^r}{dt} - M\dot{q}_m \mathbf{S}\mathbf{R}_2(-q_m) \mathbf{i}_S^s + R_R \mathbf{i}_R^r = \mathbf{u}_R^r \quad (3.427)$$

This leads to the result

$$L_S \frac{d\mathbf{i}_S^s}{dt} + M \frac{d}{dt} [\mathbf{R}_2(q_m) \mathbf{i}_R^r] + R_S \mathbf{i}_S^s = \mathbf{u}_S^s \quad (3.428)$$

$$L_R \frac{d\mathbf{i}_R^r}{dt} + M \frac{d}{dt} [\mathbf{R}_2(-q_m) \mathbf{i}_S^s] + R_R \mathbf{i}_R^r = \mathbf{u}_R^r \quad (3.429)$$

$$J \frac{d\dot{q}_m}{dt} = \tau_m - \tau_L \quad (3.430)$$

where the motor torque is given by

$$\tau_m = M(\mathbf{i}_S^s)^T \mathbf{S}\mathbf{R}_2(q_m) \mathbf{i}_R^r \quad (3.431)$$

This result corresponds to the model (3.328–3.331) of the general AC motor except for the factor 2/3 in the torque expression which is due to the three phase assumption used in the previous derivation.

**Example 55** The time derivative of the energy function  $T$  for solutions of the general AC motor is given by

$$\begin{aligned} \dot{T} &= \frac{\partial T}{\partial \mathbf{q}} \dot{\mathbf{q}} + \frac{\partial T}{\partial \dot{\mathbf{q}}} \ddot{\mathbf{q}} \\ &= \left( \begin{array}{c} \mathbf{0} \\ \frac{1}{2}\dot{\mathbf{q}}_e^T \mathbf{P}(q_m) \dot{\mathbf{q}}_e \end{array} \right)^T \dot{\mathbf{q}} + \left( \begin{array}{c} \mathbf{M}_e(q_m) \dot{\mathbf{q}}_e \\ J\dot{q}_m \end{array} \right)^T \ddot{\mathbf{q}} \\ &= \frac{1}{2}\dot{\mathbf{q}}_e^T \mathbf{P}(q_m) \dot{\mathbf{q}}_e \dot{q}_m + \dot{\mathbf{q}}_e^T \mathbf{M}_e(q_m) \ddot{\mathbf{q}}_e + \dot{q}_m J \ddot{q}_m \\ &= \frac{1}{2}\dot{\mathbf{q}}_e^T \mathbf{P}(q_m) \dot{\mathbf{q}}_e \dot{q}_m + \dot{\mathbf{q}}_e^T [-\mathbf{P}(q_m) \dot{q}_m \dot{\mathbf{q}}_e - \mathbf{R}_e \dot{\mathbf{q}}_e - \mathbf{u}_e] \\ &\quad + \dot{q}_m \left[ \frac{1}{2}\dot{\mathbf{q}}_e^T \mathbf{P}(q_m) \dot{\mathbf{q}}_e - \tau_L \right] \\ &= \dot{\mathbf{q}}_e^T \mathbf{u}_e - \dot{q}_m \tau_L - \dot{\mathbf{q}}_e^T \mathbf{R}_e \dot{\mathbf{q}}_e \\ &= \mathbf{i}_S^{sT} \mathbf{u}_S^s + \mathbf{i}_R^{rT} \mathbf{u}_R^r - \dot{q}_m \tau_L - \dot{\mathbf{q}}_e^T \mathbf{R}_e \dot{\mathbf{q}}_e \end{aligned}$$

### 3.11.7 Lagrange formulation of induction motor

The vector model of the induction motor is given by

$$\mathbf{M}_e(q_m)\ddot{\mathbf{q}}_e + \mathbf{P}(q_m)\dot{q}_m\dot{\mathbf{q}}_e + \mathbf{R}_e\dot{\mathbf{q}}_e = \begin{pmatrix} \mathbf{u}_S^s \\ \mathbf{0} \end{pmatrix} \quad (3.432)$$

$$J\ddot{q}_m - \frac{1}{2}\dot{\mathbf{q}}_e^T \mathbf{P}(q_m) \dot{\mathbf{q}}_e = -\tau_L \quad (3.433)$$

The total energy is

$$V = \frac{1}{2}\dot{\mathbf{q}}_e^T \mathbf{M}_e(q_m) \dot{\mathbf{q}}_e + \frac{1}{2}J\dot{q}_m^2 \quad (3.434)$$

and the time derivative of the energy is

$$\dot{V} = \mathbf{i}_S^{sT} \mathbf{u}_S^s - \dot{q}_m \tau_L - \dot{\mathbf{q}}_e^T \mathbf{R}_e \dot{\mathbf{q}}_e \quad (3.435)$$

It follows that if the load is passive, then the stator circuit dynamics with input  $\mathbf{u}_S^s$  and output  $\mathbf{i}_S^s$  is passive. The passivity of this model was used for controller design in (Nicklasson 1996) and (Ortega, Loria, Nicklasson and Sira-Ramirez 1998).

### 3.11.8 Lagrange formulation of DC motor

In this section the model for a DC motor with external field will be derived from the model (3.428–3.430) for the general AC motor. In a DC motor with externally controlled field the current vectors are given by

$$\mathbf{i}_S^s = \begin{pmatrix} i_e \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{i}_R^s = \mathbf{R}_2(q_m) \mathbf{i}_R^r = \begin{pmatrix} 0 \\ -i_a \end{pmatrix} \quad (3.436)$$

This means that the field current is along the  $a$  axis of the stator frame  $s$ , while the armature current in the rotor is along the negative  $b$  axis. Note that the angle between the stator current vector and the rotor current vector is  $90^\circ$ . In the same way the voltage vectors are given by

$$\mathbf{u}_S^s = \begin{pmatrix} u_e \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{u}_R^s = \mathbf{R}_2(q_m) \mathbf{u}_R^r = \begin{pmatrix} 0 \\ -u_a \end{pmatrix} \quad (3.437)$$

The dynamics of the motor is then

$$L_S \frac{d\mathbf{i}_S^s}{dt} + M \frac{d\mathbf{i}_R^s}{dt} + R_S \mathbf{i}_S^s = \mathbf{u}_S^s \quad (3.438)$$

$$L_R \frac{d}{dt} [\mathbf{R}_2(-q_m) \mathbf{i}_R^s] + M \frac{d}{dt} [\mathbf{R}_2(-q_m) \mathbf{i}_S^s] + R_R \mathbf{i}_R^s = \mathbf{u}_R^r \quad (3.439)$$

$$J \frac{dq_m}{dt} = M (\mathbf{i}_S^s)^T \mathbf{S} \mathbf{i}_R^s - \tau_L \quad (3.440)$$

The rotor equation is transformed to the stator frame, and the product rule of differentiation is used to arrive at

$$L_R \frac{d\mathbf{i}_R^s}{dt} - L_R \dot{q}_m \mathbf{S} \mathbf{i}_R^s + M \frac{d\mathbf{i}_S^s}{dt} - M \dot{q}_m \mathbf{S} \mathbf{i}_S^s + R_R \mathbf{i}_R^s = \mathbf{u}_R^s \quad (3.441)$$

The stator flux linkage is

$$\lambda_S^s = L_S \mathbf{i}_S^s + M \mathbf{i}_R^s = \begin{pmatrix} L_e i_e \\ -M i_a \end{pmatrix} = \begin{pmatrix} \lambda_e \\ -M i_a \end{pmatrix} \quad (3.442)$$

where the field flux linkage  $\lambda_e = L_S i_e$ . In the torque expression (3.431) the stator current can be expressed in terms of the stator flux linkage according to

$$\mathbf{i}_S^s = \frac{1}{L_S} (\boldsymbol{\lambda}_S^s - M \mathbf{i}_R^s) \quad (3.443)$$

which gives the torque expression

$$\tau_m = \frac{M}{L_S} (\boldsymbol{\lambda}_S^s)^T \mathbf{S} \mathbf{i}_R^s = \frac{M}{L_S} \lambda_e i_a \quad (3.444)$$

where it is used that  $(\mathbf{i}_S^s)^T \mathbf{S} \mathbf{i}_R^s = 0$ .

Then the model is found from the first row of the stator equation (3.438), the second row of the rotor equation (3.441), and the equation of motion (3.440). This gives

$$L_S \frac{di_e}{dt} + R_S i_e = u_e \quad (3.445)$$

$$L_R \frac{di_a}{dt} + \frac{M}{L_S} \dot{q}_m \lambda_e + R_R i_a = u_a \quad (3.446)$$

$$J \frac{d\dot{q}_m}{dt} = \frac{M}{L_S} \lambda_e i_a - \tau_L \quad (3.447)$$



# 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<sup>2</sup> = 6897 Pa = 0.068 bar
- 1 Pa = 1 N/m<sup>2</sup>

### 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

$$q = C_d A \sqrt{\frac{2}{\rho} \Delta p} \quad (4.1)$$

where  $A$  is the cross section of the orifice, and  $\Delta p$  is the pressure drop over the orifice, and  $\rho$  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

$$\text{Re} = \frac{D}{A\nu} q \quad (4.2)$$

where  $D$  is the diameter of the restriction,  $A$  is the cross sectional area of the flow, and  $\nu$  is the kinematic viscosity. For hydraulic oil the kinematic viscosity is approximately  $\nu \approx 30 \times 10^{-6} \text{ m}^2/\text{s}$ . 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

$$q = C_l \Delta p \quad (4.3)$$

where  $C_l$  is a constant and  $\Delta p$  is the pressure drop. This is the case for leakage flows through narrow openings, and for typical restrictions in pressure feedback channels.

When  $\text{Re} > 1000$  the flow through a restriction will be turbulent and proportional to the square root of the pressure difference according to (4.1). When  $\text{Re} < 10$  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):*

$$C_l = \frac{r^2}{8\mu L} A \quad (4.4)$$

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  $\text{Re}_{tr}$  for the Reynolds number by

$$\text{Re}_{tr} = 2 \frac{C_d^2 D A}{C_l \mu} \quad (4.5)$$

and by expressing the laminar flow characteristic (4.3) according to

$$q = C_d \sqrt{\frac{\text{Re}}{\text{Re}_{tr}}} A \sqrt{\frac{2}{\rho} \Delta p} \quad (4.6)$$

This result is verified by squaring (4.6) and inserting the Reynolds number from (4.2), which gives

$$q = \frac{C_d^2}{\text{Re}_{tr}} \frac{2DA}{\mu} \Delta p \quad (4.7)$$

Then (4.3) is recovered when  $\text{Re}_{tr}$  is defined by (4.5).

This means that we should seek a flow characteristic that satisfies

$$q = \begin{cases} C_d \sqrt{\frac{\text{Re}}{\text{Re}_{tr}}} A \sqrt{\frac{2}{\rho} \Delta p} & \text{Re} \ll \text{Re}_{tr} \\ C_d A \sqrt{\frac{2}{\rho} \Delta p} & \text{Re}_{tr} \ll \text{Re} \end{cases} \quad (4.8)$$

To obtain a solution which is defined for all  $\Delta p$  a smooth transition between the laminar and turbulent regimes was introduced in (Ellman and Piché 1999) using

$$q = \begin{cases} \frac{3\nu \text{Re}_{tr}}{4} \frac{A}{D} \frac{\Delta p}{p_{tr}} \left( 3 - \frac{\Delta p}{p_{tr}} \right) & \Delta p \leq p_{tr} \\ C_d A \sqrt{\frac{2}{\rho} \Delta p} & p_{tr} \leq \Delta p \end{cases} \quad (4.9)$$

Here the threshold pressure

$$p_{tr} = \frac{9 \text{Re}_{tr}^2 \rho \nu^2}{8C_d^2} \frac{1}{D^2} \quad (4.10)$$

corresponds to a given threshold  $\text{Re}_{tr}$  for the Reynolds number. Assuming a circular orifice with diameter  $D$ , we have

$$A = \frac{\pi D^2}{4} \Rightarrow D^2 = \frac{4A}{\pi} \Rightarrow \frac{A}{D} = \frac{\sqrt{\pi}}{2} \sqrt{A} \quad (4.11)$$

Moreover, we define the constant

$$F_{tr} = p_{tr} A = p_{tr} \frac{\pi D^2}{4} = \frac{9 \text{Re}_{tr}^2 \rho \nu^2 \pi}{8C_d^2} \frac{1}{4} \quad (4.12)$$

Then the following result has been established:

The flow through a restriction can be described by the regularized flow characteristic

$$q(A, \Delta p) = \begin{cases} \frac{3\nu \text{Re}_{tr}}{4} \frac{\sqrt{\pi}}{2} \sqrt{A} \frac{A \Delta p}{F_{tr}} \left( 3 - \frac{A \Delta p}{F_{tr}} \right) & A \Delta p \leq F_{tr} \\ C_d A \sqrt{\frac{2}{\rho} \Delta p} & F_{tr} \leq A \Delta p \end{cases} \quad (4.13)$$

where

$$F_{tr} = \frac{9 \text{Re}_{tr}^2 \rho \nu^2 \pi}{8C_d^2} \frac{1}{4} \quad (4.14)$$

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):  $\text{Re}_{tr} = 1000$ ,  $\rho = 900 \text{ kg/m}^3$ ,  $\nu = 30 \times 10^{-6} \text{ m}^2/\text{s} = 30 \text{ cSt}$ ,  $C_d = 0.6$ . The regularized characteristic (4.13) is well suited for numerical simulation as it is physically justified, and it eliminates the problems that are experienced with the turbulent characteristic (4.1).

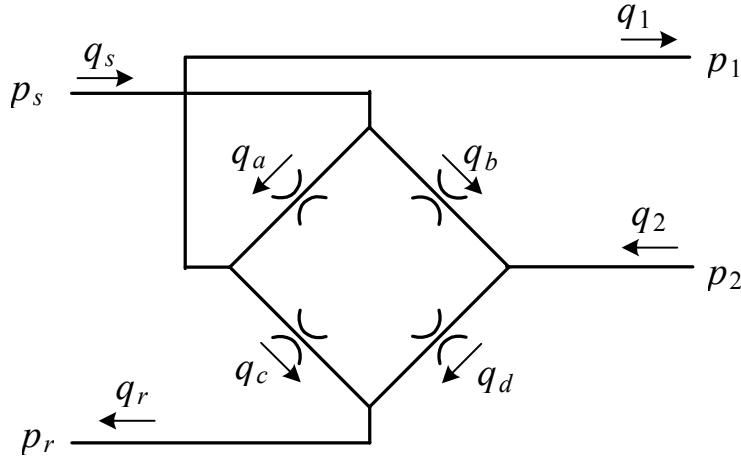


Figure 4.1: Four-way valve

#### 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

$$\begin{aligned} q_a &= C_d A_a(x_v) \sqrt{\frac{2}{\rho} (p_s - p_1)} \\ q_b &= C_d A_b(x_v) \sqrt{\frac{2}{\rho} (p_s - p_2)} \\ q_c &= C_d A_c(x_v) \sqrt{\frac{2}{\rho} (p_1 - p_r)} \\ q_d &= C_d A_d(x_v) \sqrt{\frac{2}{\rho} (p_2 - p_r)} \end{aligned} \quad (4.15)$$

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

$$q_s = q_a + q_b, \quad q_r = q_c + q_d \quad (4.16)$$

$$q_1 = q_a - q_c, \quad q_2 = q_d - q_b \quad (4.17)$$

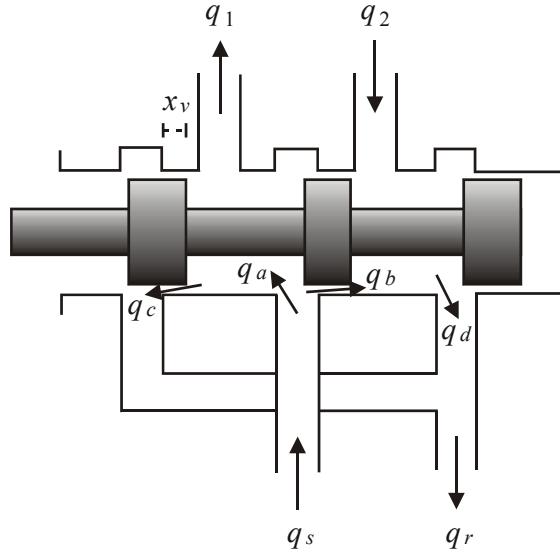


Figure 4.2: A matched and symmetric four-way valve.

#### 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

$$A_a(x_v) = A_d(x_v) = A_b(-x_v) = A_c(-x_v) \quad (4.18)$$

If a matched and symmetric valve is equipped with a critical spool and rectangular orifices, then the port areas are given by

$$A_a(x_v) = A_d(x_v) = \begin{cases} 0, & x_v \leq 0 \\ bx_v, & x_v \geq 0 \end{cases} \quad (4.19)$$

$$A_b(x_v) = A_c(x_v) = \begin{cases} -bx_v, & x_v \leq 0 \\ 0, & x_v \geq 0 \end{cases} \quad (4.20)$$

A matched and symmetric valve with open center spool and rectangular orifices has port openings given by

$$A_a(x_v) = A_d(x_v) = b(U + x_v), \quad |x_v| \leq U \quad (4.21)$$

$$A_b(x_v) = A_c(x_v) = b(U - x_v), \quad |x_v| \leq U \quad (4.22)$$

#### 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

$$q_1 = q_2 \quad (4.23)$$

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 be 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 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

$$q_a = q_d, \quad q_b = q_c \quad (4.24)$$

which in turn imply that

$$p_s + p_r = p_1 + p_2 \quad (4.25)$$

In the symmetric load case it is convenient to define the load pressure

$$p_L = p_1 - p_2 \quad (4.26)$$

and the load flow

$$q_L = \frac{1}{2}(q_1 + q_2) \quad (4.27)$$

We then find that the pressures  $p_1$  and  $p_2$  can be expressed as

$$p_1 = \frac{p_s + p_L}{2}, \quad p_2 = \frac{p_s - p_L}{2} \quad (4.28)$$

and the load flow can be found to be

$$q_L = C_d A_a(x_v) \sqrt{\frac{1}{\rho} (p_s - p_L)} - C_d A_b(x_v) \sqrt{\frac{1}{\rho} (p_s + p_L)} \quad (4.29)$$

If a valve with a critical spool and rectangular ports is connected to a symmetric load, then the port areas are given by

$$A_a(x_v) = \begin{cases} 0 & x_v \leq 0 \\ bx_v & x_v \geq 0 \end{cases}, \quad A_b(x_v) = \begin{cases} -bx_v & x_v \leq 0 \\ 0 & x_v \geq 0 \end{cases} \quad (4.30)$$

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

$$q_L = C_d b x_v \sqrt{\frac{1}{\rho} (p_s - \text{sgn}(x_v)p_L)} \quad (4.31)$$

The valve characteristic (4.31) is usually written in the nondimensional form

$$\frac{q_L}{C_d b x_{v \max} \sqrt{p_s / \rho}} = \frac{x_v}{x_{v \max}} \sqrt{1 - \text{sgn}(x_v) \frac{p_L}{p_s}} \quad (4.32)$$

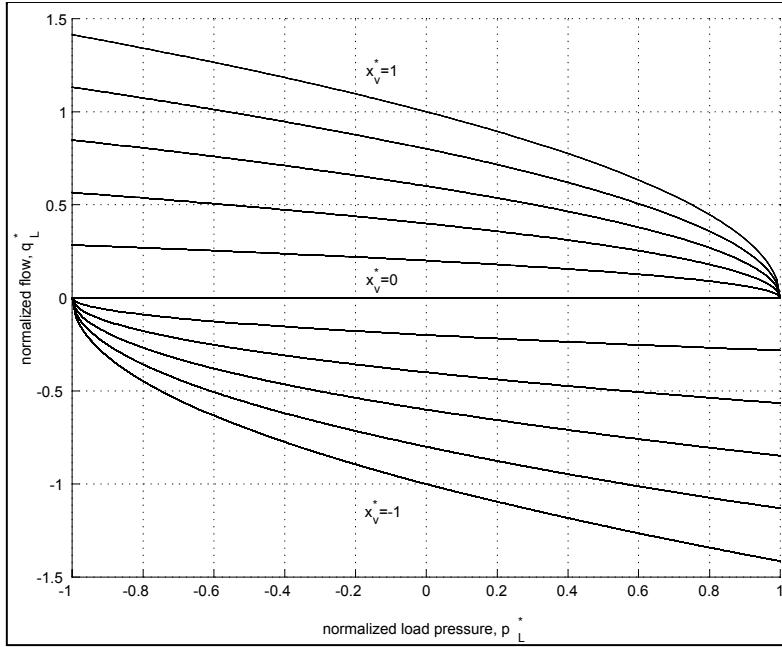


Figure 4.3: Valve characteristic

which is commonly plotted using pressure-flow curves as in (Figure 4.3), where the normalized values are

$$q_L^* = \frac{q_L}{C_d b x_{v \max} \sqrt{\frac{p_s}{\rho}}}, \quad x_v^* = \frac{x_v}{x_{v \max}}, \quad p_L^* = \frac{p_L}{p_s} \quad (4.33)$$

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

$$q_L = K_q x_v - K_c p_L \quad (4.34)$$

where

$$K_q = \frac{\partial q_L}{\partial x_v} = C_d b \sqrt{\frac{1}{\rho} (p_s - \text{sgn}(x_v) p_L)} \quad (4.35)$$

and

$$K_c = -\frac{\partial q_L}{\partial p_L} = \frac{C_d b x_v \sqrt{(1/\rho)(p_s - \text{sgn}(x_v)p_L)}}{2(p_s - \text{sgn}(x_v)p_L)} \quad (4.36)$$

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

$$K_{q0} = C_d b \sqrt{\frac{p_s}{\rho}} \quad (4.37)$$

$$K_{c0} = 0 \quad (4.38)$$

If the system is designed so that the load pressure satisfies the condition  $|p_L| < \frac{2}{3}p_s$ , then

$$|p_L| < \frac{2}{3}p_s \quad \Rightarrow \quad 0.577K_{q0} \leq K_q \leq 1.29K_{q0} \quad (4.39)$$

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

$$\tilde{p} := p_s - \text{sgn}(x_v)p_L \quad (4.40)$$

and inserting  $\Delta p = \tilde{p}/2$  into the expression of (4.13). This gives

$$q_L = \begin{cases} \frac{3\nu}{4} \text{Re}_{tr} \frac{\sqrt{\pi}}{2} \sqrt{A} \frac{A\tilde{p}}{2F_{tr}} \left( 3 - \frac{A\tilde{p}}{2F_{tr}} \right) & A\tilde{p} \leq 2F_{tr} \\ C_d A_v \sqrt{\frac{1}{\rho}\tilde{p}} & 2F_{tr} \leq A\tilde{p} \end{cases} \quad (4.41)$$

where

$$F_{tr} = \frac{9 \text{Re}_{tr}^2 \rho \nu^2 \pi}{8C_d^2} \frac{1}{4}, \quad A = bx_v, \quad \tilde{p} = p_s - \text{sgn}(x_v)p_L \quad (4.42)$$

#### 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

$$\frac{q_L}{C_d b U \sqrt{p_s/\rho}} = \left( 1 + \frac{x_v}{U} \right) \sqrt{1 - \frac{p_L}{p}} - \left( 1 - \frac{x_v}{U} \right) \sqrt{1 + \frac{p_L}{p}}, \quad |x_v| \leq U \quad (4.43)$$

#### 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

$$F_c = -K_c x_c + F_{c0} \quad (4.44)$$

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

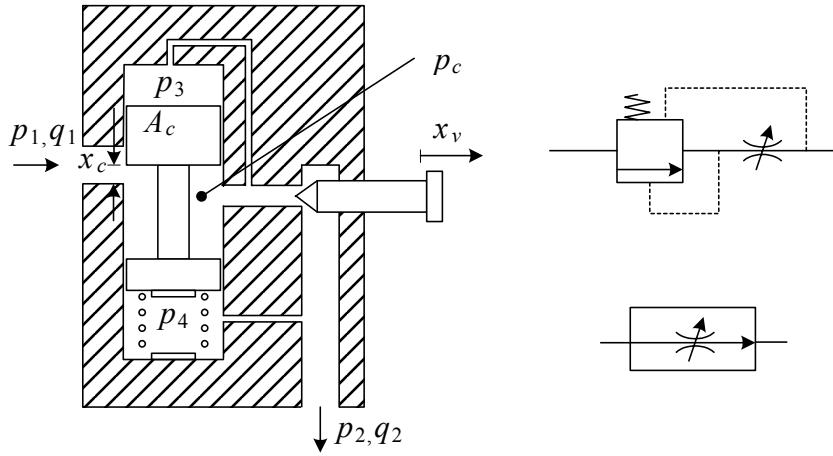


Figure 4.4: The figure shows the mechanical design and symbols for a pressure compensated valve for volume flow control. The compensation spool with position  $x_c$  which is positive in the upwards direction will be positioned so that the pressure difference  $p_2 - p_c$  over the main spool is approximately 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

$$q_1 = C_d b c x_c \sqrt{\frac{2}{\rho} (p_1 - p_c)} \quad (4.45)$$

$$q_2 = C_d b x_v \sqrt{\frac{2}{\rho} (p_c - p_2)} \quad (4.46)$$

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

$$m_c s^2 x_c(s) + K_c x_c(s) = A_c [p_4(s) - p_3(s)] + F_{c0} \quad (4.47)$$

and the pressure dynamics of the compensation chambers

$$\frac{V_3}{\beta} \dot{p}_3 = +A_c \dot{x}_c + C_3 (p_c - p_3) \quad (4.48)$$

$$\frac{V_4}{\beta} \dot{p}_4 = -A_c \dot{x}_c + C_4 (p_2 - p_4) \quad (4.49)$$

The Laplace transformed pressure equations are

$$C_3 \left( 1 + \frac{V_3}{C_3 \beta} s \right) p_3 = +A_c s x_c + C_3 p_c \quad (4.50)$$

$$C_4 \left( 1 + \frac{V_4}{C_4 \beta} s \right) p_4 = -A_c s x_c + C_4 p_2 \quad (4.51)$$

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

$$C_3 p_3 = +A_c s x_c + C_3 p_c \Rightarrow p_3 = p_c + \frac{A_c}{C_3} s x_c \quad (4.52)$$

$$C_4 p_4 = -A_c s x_c + C_4 p_2 \Rightarrow p_4 = p_2 - \frac{A_c}{C_4} s x_c \quad (4.53)$$

Insertion in the equation of motion gives

$$m_c s^2 x_c(s) + A_c \frac{C_3 + C_4}{C_3 C_4} s x_c(s) + K_c x_c(s) = A_c [p_2(s) - p_c(s)] + F_{c0} \quad (4.54)$$

which is Laplace transformed and rearranged to

$$\begin{aligned} p_c(s) - p_2(s) &= +\frac{F_{c0}}{A_c} - \frac{K_c}{A_c} \left( \frac{m_c}{K_c} s^2 + \frac{A_c^2}{K_c} \frac{C_3 + C_4}{C_3 C_4} s + 1 \right) x_c(s) \\ &= +\frac{F_{c0}}{A_c} - \frac{K_c}{A_c} \left( \frac{s^2}{\omega_c^2} + 2\zeta_c \frac{s}{\omega_c} + 1 \right) x_c(s) \end{aligned} \quad (4.55)$$

where  $\omega_c^2 = K_c/m_c$ . It can be seen that for frequencies well below  $\omega_c$ , the compensation spool dynamics will satisfy

$$p_c - p_2 = \frac{F_{c0}}{A_c} - \frac{K_c x_c}{A_c} \quad (4.56)$$

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

$$q_1 = q_2 = C_d \sqrt{\frac{2 F_{c0}}{\rho A_c} b x_v} \quad (4.57)$$

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  $b x_v$  of the main spool.

#### 4.2.9 Balance valve

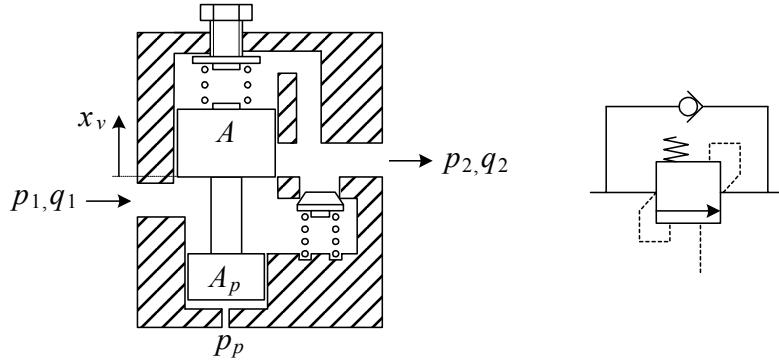


Figure 4.5: Balance valve: Mechanical design and symbol

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

$$A_r = A - A_p \quad (4.58)$$

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

$$m_v \ddot{x}_v = p_p A_p + (p_1 - p_0) A_r - K x_v - p_2 A \quad (4.59)$$

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

$$x_v = \frac{A_r}{K} [p_1 - p_0 + R p_p - p_2 (R + 1)], \quad 0 \leq x_v \leq x_{v,\max}$$

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 ratio  $R$  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

$$q_1 = \begin{cases} C_d x_v b \sqrt{\frac{2}{\rho} (p_1 - p_2)} & p_1 > p_2 \\ -C_d A_c \sqrt{\frac{2}{\rho} (p_2 - p_1)} & p_1 < p_2 \end{cases} \quad (4.60)$$

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  $\rho$  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

$$\frac{d\rho}{\rho} = \frac{dp}{\beta} \quad (4.61)$$

where  $\beta$  is the *bulk modulus*.

We see that the bulk modulus  $\beta$  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

$$\frac{d}{dt}(\rho V) = w_{in} - w_{out} \quad (4.62)$$

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

$$\dot{\rho}V + \rho\dot{V} = \rho(q_{in} - q_{out}) \quad (4.63)$$

and insertion of the expression (4.61) leads to the following result:

The mass balance of a hydraulic volume  $V$  is

$$\frac{V}{\beta}\dot{p} + \dot{V} = q_{in} - q_{out} \quad (4.64)$$

**Example 58** The differential pressure work on a volume  $V$  of constant mass  $m$  is

$$pdV = pd\left(\frac{m}{\rho}\right) = -pV\frac{d\rho}{\rho} = -\frac{pV}{\beta}dp \quad (4.65)$$

This means that the stored energy in a volume  $V$  due to a pressure  $p$  is

$$W_p = \int_0^p \frac{V}{\beta} p' dp' = \frac{1}{2} \frac{V}{\beta} p^2 \quad (4.66)$$

### 4.3.2 Rotational motors

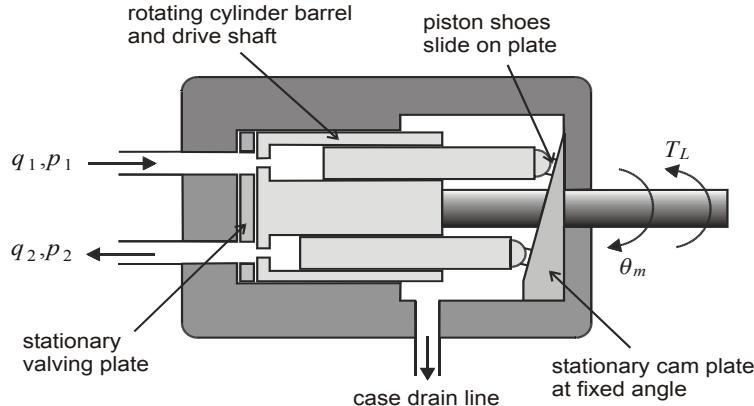


Figure 4.6: Hydraulic motor

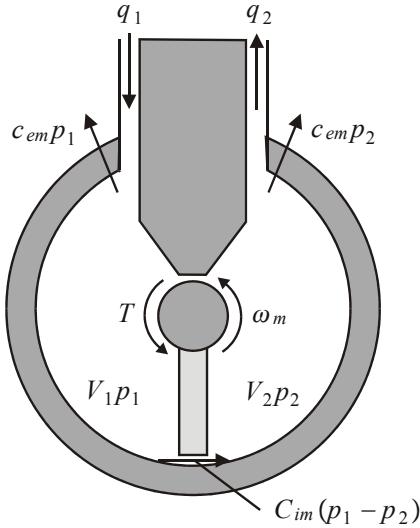


Figure 4.7: Rotational hydraulic motor of the single vane type with limited travel.

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  $180^\circ$  or slightly less than  $360^\circ$ , 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

$$\dot{V}_1 + \frac{V_1}{\beta} \dot{p}_1 = -C_{im}(p_1 - p_2) - C_{em}p_1 + q_1 \quad (4.67)$$

$$\dot{V}_2 + \frac{V_2}{\beta} \dot{p}_2 = -C_{im}(p_2 - p_1) - C_{em}p_2 - q_2 \quad (4.68)$$

where  $C_{im}$  is the coefficient for the internal leakage and  $C_{em}$  is the coefficient for leakage out of the motor.  $\beta$  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  $\omega_m$  of the motor:

$$\dot{V}_1 = -\dot{V}_2 = D_m \omega_m \quad (4.69)$$

The constant  $D_m$  is called the displacement. The shaft angle is denoted  $\theta_m$ .

The motor torque  $T$  is proportional to the pressure difference, and by equating the

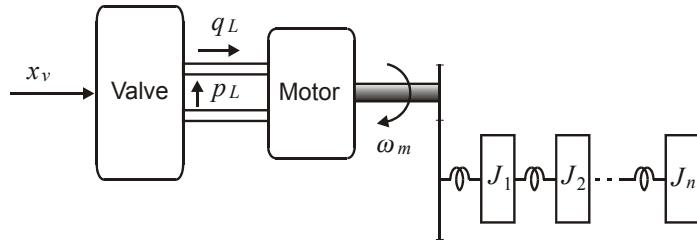


Figure 4.8: Valve controlled motor with elastic modes in the load.

power of the motor torque with the power of the working fluid we find that

$$T\omega_m = p_1 \dot{V}_1 + p_2 \dot{V}_2 = (p_1 - p_2) D_m \omega_m \quad (4.70)$$

It follows that the motor torque is

$$T = D_m (p_1 - p_2) \quad (4.71)$$

The equation of motion for the motor shaft is therefore

$$J_t \dot{\omega}_m = -B_m \omega_m + D_m (p_1 - p_2) - T_L \quad (4.72)$$

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

$$\frac{V_1}{\beta} \dot{p}_1 = -C_{im}(p_1 - p_2) - C_{em}p_1 - D_m \omega_m + q_1 \quad (4.73)$$

$$\frac{V_2}{\beta} \dot{p}_2 = -C_{im}(p_2 - p_1) - C_{em}p_2 + D_m \omega_m - q_2 \quad (4.74)$$

$$J_t \dot{\omega}_m = -B_m \omega_m + D_m (p_1 - p_2) - T_L \quad (4.75)$$

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  $\omega_m$ . 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  $\omega_m$ . The shaft model has one port with effort  $T_1$  and flow  $\omega_m$ , one port with effort  $-T_2$  and flow  $\omega_m$ , and one port with effort  $T_L$  and flow  $\omega_m$ . In terms of computation the systems can be interconnected if the variables  $q_1$  and  $\omega_m$  are inputs to chamber 1,  $q_2$  and  $\omega_m$  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

$$J_1 \dot{\omega}_1 = T_L - T_1 \quad (4.76)$$

$$\dot{\theta}_1 = \omega_1 \quad (4.77)$$

$$T_L = D_1 (\omega_m - \omega_1) + K_1 (\theta_m - \theta_1) \quad (4.78)$$

where the input port has been connected to the motor shaft. The inputs to the two-port are  $\omega_m$  and  $T_1$ , while  $T_L$  and  $\omega_1$  are outputs. We may add on any number of additional degrees of freedom as two-ports

$$J_i \dot{\omega}_i = T_{i-1} - T_i \quad (4.79)$$

$$T_i = D_i (\omega_{i-1} - \omega_i) + K_i (\theta_{i-1} - \theta_i) \quad (4.80)$$

with port variables  $T_{i-1}$  and  $\omega_{i-1}$  at the input and  $T_i$  and  $\omega_i$  at the output.

#### 4.3.4 Hydraulic cylinder

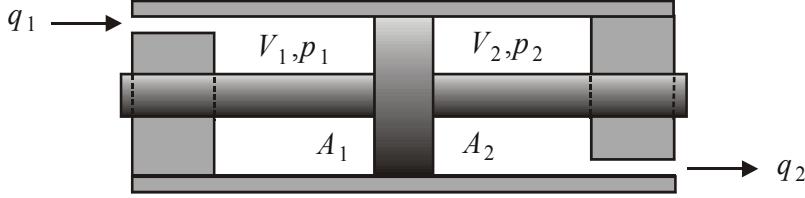


Figure 4.9: Symmetric hydraulic cylinder

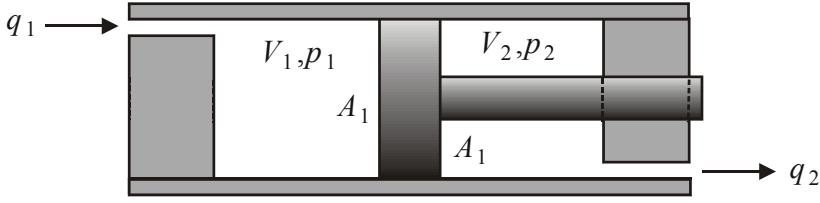


Figure 4.10: Single-rod hydraulic piston

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

$$A_1 = A_2 = A_p - A_r \quad (4.81)$$

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

$$A_1 = A_p, \quad A_2 = A_p - A_r \quad (4.82)$$

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

$$\frac{V_{10} + A_1 x_p}{\beta} \dot{p}_1 = -C_{im}(p_1 - p_2) - C_{em}p_1 - A_1 \dot{x}_p + q_1 \quad (4.83)$$

$$\frac{V_{20} - A_2 x_p}{\beta} \dot{p}_2 = -C_{im}(p_2 - p_1) - C_{em}p_2 + A_2 \dot{x}_p - q_2 \quad (4.84)$$

$$m_t \ddot{x}_p = -B_p \dot{x}_p + A_1 p_1 - A_2 p_2 - F_L \quad (4.85)$$

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

$$V_1 = V_0 + D_m \theta_m, \quad V_2 = V_0 - D_m \theta_m \quad (4.86)$$

Subtraction of the mass balance (4.74) for chamber 2 from the mass balance (4.73) for chamber 1 gives

$$2D_m \omega_m + \frac{V_0}{\beta} (\dot{p}_1 - \dot{p}_2) + \frac{D_m \theta_m}{\beta} (\dot{p}_1 + \dot{p}_2) = q_1 + q_2 - 2C_{im}(p_1 - p_2) - C_{em}(p_1 - p_2) \quad (4.87)$$

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

$$\frac{V_t}{4\beta} \dot{p}_L = -C_{tm} p_L - D_m \omega_m + q_L \quad (4.88)$$

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

$$\frac{V_t}{4\beta} \dot{p}_L = -C_{tm} p_L - D_m \omega_m + q_L \quad (4.89)$$

$$J_t \dot{\omega}_m = -B_m \omega_m + D_m p_L - T_L \quad (4.90)$$

**Example 59** An energy function of the motor is

$$V = \frac{1}{2} J_t \omega_m^2 + \frac{1}{2} \frac{V_t}{4\beta} p_L^2 \quad (4.91)$$

The time derivative is

$$\begin{aligned} \dot{V} &= \omega_m J_t \dot{\omega}_m + p_L \frac{V_t}{4\beta} \dot{p}_L \\ &= -B_m \omega_m^2 - \omega_m T_L + \omega_m D_m p_L - C_{tm} p_L^2 - p_L D_m \omega_m + p_L q_L \\ &= p_L q_L - \omega_m T_L - B_m \omega_m^2 - C_{tm} p_L^2 \end{aligned} \quad (4.92)$$

We see that 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

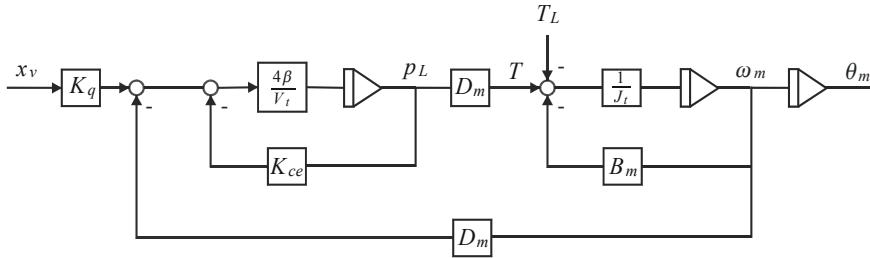


Figure 4.11: Valve controlled hydraulic motor.

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

$$\frac{V_t}{4\beta} \dot{p}_L = -K_{ce} p_L - D_m \omega_m + K_q x_v \quad (4.93)$$

$$J_t \dot{\omega}_m = -B_m \omega_m + D_m p_L - T_L \quad (4.94)$$

$$\dot{\theta}_m = \omega_m \quad (4.95)$$

where  $K_{ce} = K_c + C_{tm}$  is the leakage coefficient for motor and valve,  $B_m$  is the viscous friction coefficient, while  $\theta_m$  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

$$K_{ce} \left( 1 + \frac{V_t}{4\beta K_{ce}} s \right) p_L = (-D_m s \theta_m + K_q x_v) \quad (4.96)$$

$$(J_t s^2 + B_m s) \theta_m = D_m p_L - T_L \quad (4.97)$$

Insertion of the mass balance (4.96) into the equation of motion (4.97) gives

$$\begin{aligned} K_{ce} \left( 1 + \frac{V_t}{4\beta K_{ce}} s \right) (J_t s^2 + B_m s) \theta_m &= -D_m^2 s \theta_m + D_m K_q x_v \\ &\quad - K_{ce} \left( 1 + \frac{V_t}{4\beta K_{ce}} s \right) T_L \end{aligned} \quad (4.98)$$

which can be rearranged as

$$\theta_m(s) = \frac{\frac{K_q}{D_m} x_v(s) - \frac{K_{ce}}{D_m^2} \left( 1 + \frac{V_t}{4\beta K_{ce}} s \right) T_L(s)}{s \left[ \frac{V_t J_t}{4\beta D_m^2} s^2 + \left( \frac{K_{ce} J_t}{D_m^2} + \frac{B_m V_t}{4\beta D_m^2} \right) s + \left( 1 + \frac{B_m K_{ce}}{D_m^2} \right) \right]} \quad (4.99)$$

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

$$\theta_m(s) = \frac{\frac{K_q}{D_m} x_v(s) - \frac{K_{ce}}{D_m^2} \left( 1 + \frac{s}{\omega_t} \right) T_L(s)}{s \left( 1 + 2\zeta_h \frac{s}{\omega_h} + \frac{s^2}{\omega_h^2} \right)} \quad (4.100)$$

where  $\omega_h$  the hydraulic undamped natural frequency,  $\zeta_h$  is the relative damping, and  $\omega_t$  is the break frequency of the pressure dynamics defined by

$$\omega_h^2 = \frac{4\beta D_m^2}{V_t J_t}, \quad \zeta_h = \frac{K_{ce}}{D_m} \sqrt{\frac{\beta J_t}{V_t}}, \quad \omega_t = \frac{4\beta K_{ce}}{V_t} \quad (4.101)$$

We note that

$$2\zeta_h \omega_h = \frac{4\beta K_{ce}}{V_t} = \omega_t \quad (4.102)$$

The transfer function from the spool position  $x_v$  to the shaft angle  $\theta_m$  is given by

$$H_m(s) = \frac{\theta_m}{x_v}(s) = \frac{\frac{K_q}{D_m}}{s \left( 1 + 2\zeta_h \frac{s}{\omega_h} + \frac{s^2}{\omega_h^2} \right)} \quad (4.103)$$

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

$$(1 + 2\zeta_h \frac{s}{\omega_h} + \frac{s^2}{\omega_h^2}) \approx 1 \quad (4.104)$$

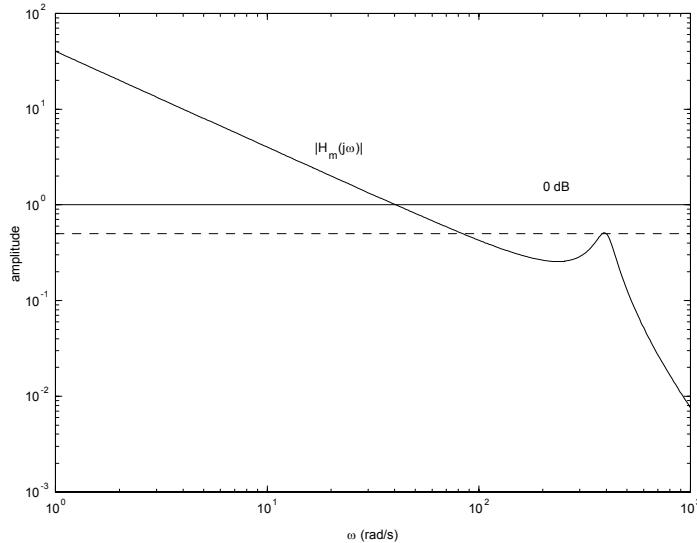


Figure 4.12: Magnitude of the frequency response from the valve spool position  $x_v$  to the motor angle  $\theta$ . Numerical values are  $K_q/D_m = 40$ ,  $\omega_h = 400$  rad/s and  $\zeta_h = 0.1$ . The dashed line is drawn at -6 dB.

the motor velocity  $\omega_m$  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  $\omega_h$  is an important parameter in the design of electrohydraulic servomechanisms. The undamped natural frequency is given by  $\beta$ ,  $D_m$  and  $J_t$ . The parameters  $J_t$  and  $D_m$  can be found with high accuracy, while the bulk modulus  $\beta$  may vary. However, the numerical value  $\beta = 7.0 \cdot 10^8$  Pa (= 10<sup>5</sup> 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  $\theta_m$  is

$$\frac{\theta_m}{T_L}(s) = \frac{-\frac{K_{ce}}{D_m^2} \left(1 + \frac{s}{\omega_t}\right)}{s \left(1 + 2\zeta_h \frac{s}{\omega_h} + \frac{s^2}{\omega_h^2}\right)} \quad (4.105)$$

#### 4.4.3 Hydraulic motor with P controller

With a proportional controller

$$x_v = K_p(\theta_d - \theta_m) \quad (4.106)$$

the loop transfer function for a hydraulic motor is

$$L(s) = K_p H_m(s) = \frac{K_v}{s \left( 1 + 2\zeta_h \frac{s}{\omega_h} + \frac{s^2}{\omega_h^2} \right)} \quad (4.107)$$

where

$$K_v = \frac{K_p K_q}{D_m} \quad (4.108)$$

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  $\omega_{180}$ , which is the frequency where the frequency response  $L(j\omega_h)$  has a phase of  $180^\circ$ . A closer inspection of the loop transfer function  $L(s)$  reveals that  $\omega_{180} = \omega_h$ , and that

$$|L(j\omega_h)| = \frac{K_v}{2\zeta_h \omega_h}, \quad \angle L(j\omega_h) = -180^\circ \quad (4.109)$$

Using the expression for  $\zeta_h$  in (4.101) we find that

$$|L(j\omega_{180})| = \frac{K_v}{2\zeta_h \omega_h} \quad (4.110)$$

Thus, a gain margin of  $\Delta K = 6$  dB, which occurs for  $|L(j\omega_{180})| = 1/2$ , is achieved with

$$K_v = \zeta_h \omega_h \Rightarrow K_p = \frac{D_m}{K_q} \zeta_h \omega_h \quad (4.111)$$

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

$$K_v = \frac{K_p K_q}{D_m} \leq 2\zeta_h \omega_h \Rightarrow K_p \leq 2 \frac{D_m}{K_q} \zeta_h \omega_h \quad (4.112)$$

A gain margin of 6 dB is achieved with

$$K_v = \zeta_h \omega_h \Rightarrow K_p = \frac{D_m}{K_q} \zeta_h \omega_h \quad (4.113)$$

**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  $\beta$  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

$$2\zeta_h \omega_h = \left( \frac{K_{ce} J_t}{D_m^2} + \frac{B_m V_t}{4\beta D_m^2} \right) \frac{4\beta D_m^2}{V_t J_t} = \frac{4\beta K_{ce}}{V_t} + \frac{B_m}{J_t} \quad (4.114)$$

#### 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

$$\frac{V_t}{4\beta} \dot{p}_L = -C_{tp} p_L - A_p \dot{x}_p + q_L \quad (4.115)$$

$$m_t \ddot{x}_p = -B_p \dot{x}_p + A_p p_L - F_L \quad (4.116)$$

The Laplace transform of a symmetric cylinder with matched and symmetric valve is

$$x_p(s) = \frac{\frac{K_q}{A_p} x_v(s) - \frac{K_{ce}}{A_p^2} \left( 1 + \frac{s}{\omega_t} \right) F_L(s)}{s \left( 1 + 2\zeta_h \frac{s}{\omega_h} + \frac{s^2}{\omega_h^2} \right)} \quad (4.117)$$

where

$$\omega_h^2 = \frac{4\beta A_p^2}{V_t m_t}, \quad \zeta_h = \frac{K_{ce}}{A_p} \sqrt{\frac{\beta m_t}{V_t}}, \quad \omega_t = \frac{4\beta K_{ce}}{V_t} \quad (4.118)$$

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

$$K_p = \frac{A_p}{K_q} \zeta_h \omega_h \quad (4.119)$$

**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  $\omega_s$  as defined by the crossover frequency. Then the bandwidth of the system with volume  $V_t$  will be

$$\omega_c = \omega_s \frac{V_s}{V_t} \quad (4.120)$$

#### 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  $\omega_p$  is

$$\theta_m(s) = \frac{\frac{k_p \omega_p}{D_m} \phi_p(s) - \frac{C_t}{D_m^2} \left(1 + \frac{s}{\omega_0}\right) T_L(s)}{s \left(1 + 2\zeta_h \frac{s}{\omega_h} + \frac{s^2}{\omega_h^2}\right)} \quad (4.121)$$

where

$$\omega_h^2 = \frac{\beta D_m^2}{V_0 J_t}, \quad \zeta_h = \frac{C_t}{2D_m} \sqrt{\frac{\beta J_t}{V_0}}, \quad \omega_0 = \frac{\beta C_t}{V_0} \quad (4.122)$$

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  $\omega_h^2$  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  $\theta_m$ :

$$u_p = K_p (\theta_d - \theta_m) \quad (4.123)$$

The loop transfer function  $L(s)$  is seen to be

$$L(s) = \frac{K_v}{s \left(1 + 2\zeta_h \frac{s}{\omega_h} + \frac{s^2}{\omega_h^2}\right)} \quad (4.124)$$

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

$$K_v \leq 2\zeta_h \omega_h \quad (4.125)$$

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

$$K_v = \zeta_h \omega_h \Rightarrow K_p = \frac{D_m}{k_p \omega_p} \zeta_h \omega_h \quad (4.126)$$

#### 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.

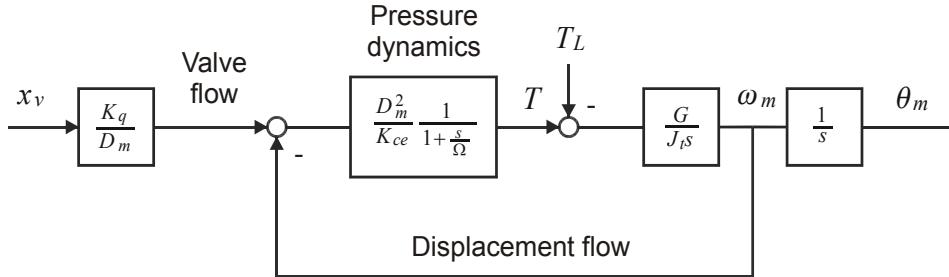


Figure 4.13: Block diagram for valve controlled motor with elastic modes in the load.

Then the transfer function from the motor torque  $T$  to the motor angle  $\theta_m$  is given by (3.85) to be

$$Js^2\theta_m(s) = G(s)T(s) \quad (4.127)$$

where

$$G(s) = \frac{1 + 2\zeta_a \frac{s}{\omega_a} + \left(\frac{s}{\omega_a}\right)^2}{1 + 2\zeta_1 \frac{s}{\omega_1} + \left(\frac{s}{\omega_1}\right)^2}, \quad \omega_a < \omega_1 \quad (4.128)$$

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

$$K_{ce} \left(1 + \frac{V_t}{4\beta K_{ce}} s\right) p_L = (-D_m s \theta_m + K_q x_v) \quad (4.129)$$

$$J_t s^2 \theta_m = G(s) D_m p_L \quad (4.130)$$

Insertion of the first equation into the second gives

$$K_{ce} \left(1 + \frac{V_t}{4\beta K_{ce}} s\right) J_t s^2 \theta_m = G(s) (-D_m^2 s \theta_m + D_m K_q x_v) \quad (4.131)$$

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

$$H_e(s) = \frac{\theta_m}{x_v}(s) = \frac{G(s) \frac{K_q}{D_m}}{s \left(G(s) + 2\zeta_h \frac{s}{\omega_h} + \frac{s^2}{\omega_h^2}\right)} \quad (4.132)$$

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  $\omega_a$  and above  $\omega_1$  the frequency response is the same for the rigid and the elastic case.
2. If  $\omega_1 \ll \omega_h$ , then

$$H_e(s) \approx \frac{\frac{K_q}{D_m}}{s \left(1 + 2\zeta_h \frac{s}{\omega_h} + \frac{s^2}{\omega_h^2}\right)} \quad (4.133)$$

3. If  $\omega_h \ll \omega_a$ , then

$$H_e(s) \approx \frac{G(s) \frac{K_q}{D_m}}{s \left( 1 + 2\zeta_h \frac{s}{\omega_h} + \frac{s^2}{\omega_h^2} \right)} \quad (4.134)$$

#### 4.4.7 Mechanical analog

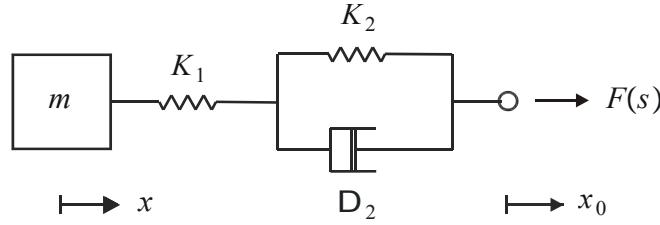


Figure 4.14: Mechanical analog of hydraulic motor with proportional position controller.

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

$$F_1(s) = K_p \frac{\left(1 + \frac{s}{\omega_1}\right)}{\left(1 + \frac{s}{\omega_2}\right)} [x_0(s) - x(s)] \quad (4.135)$$

which clearly shows that this corresponds to a PD controller with limited derivative action, and where the constants are given by.

$$K_p = \frac{K_1 K_2}{K_1 + K_2}, \quad \omega_1 = \frac{K_2}{D_2}, \quad \omega_2 = \frac{K_1 + K_2}{D_2} \quad (4.136)$$

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

$$(m s^2 + B s) x(s) = K_p \frac{\left(1 + \frac{s}{\omega_1}\right)}{\left(1 + \frac{s}{\omega_2}\right)} [x_0(s) - x(s)] - F_L \quad (4.137)$$

Consider a hydraulic motor with equation of motion given by (4.98)

$$(J_t s^2 + B_m s) \theta_m = \frac{D_m^2}{K_{ce}} \frac{\frac{K_q}{D_m} x_v - s \theta_m}{1 + \frac{s}{\omega_t}} - T_L$$

where

$$\omega_t = \frac{4\beta K_{ce}}{V_t} \quad (4.138)$$

Then, with proportional feedback  $x_v = K_p(\theta_0 - \theta_m)$  this becomes

$$(J_t s^2 + B_m s) \theta_m = -K_v \frac{D_m^2}{K_{ce}} \frac{1 + \frac{s}{K_v}}{1 + \frac{s}{\omega_t}} \theta_m + K_v \frac{D_m^2}{K_{ce}} \frac{1}{1 + \frac{s}{\omega_t}} \theta_0 - T_L \quad (4.139)$$

where  $K_v = K_p K_q / D_m$  is the velocity constant. If we introduce the variable  $\theta_d$  defined by

$$\theta_0(s) = \frac{1}{1 + \frac{s}{K_v}} \theta_d(s) \quad (4.140)$$

then equation (4.139) can be written

$$(J_t s^2 + B_m s) \theta_m = K_v \frac{D_m^2}{K_{ce}} \frac{1 + \frac{s}{K_v}}{1 + \frac{s}{\omega_t}} [\theta_d(s) - \theta_m(s)] - T_L$$

and we find that the dynamics are the same as for the mechanical analog if the constants satisfy

$$K_v \frac{D_m^2}{K_{ce}} = \frac{K_1 K_2}{K_1 + K_2}, \quad K_v = \frac{K_2}{D_2}, \quad \omega_t = \frac{K_1 + K_2}{D_2} \quad (4.141)$$

We may solve for the parameters of the mechanical analog, which are found to be

$$K_1 = \frac{4\beta D_m^2}{V_t}, \quad K_2 = \frac{K_v}{\omega_t - K_v} \frac{4\beta D_m^2}{V_t}, \quad D_2 = \frac{1}{\omega_t - K_v} \frac{4\beta D_m^2}{V_t} \quad (4.142)$$

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 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

$$\frac{V}{\beta} \dot{p} + \dot{V} = q_{in} - q_{out} \quad (4.143)$$

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

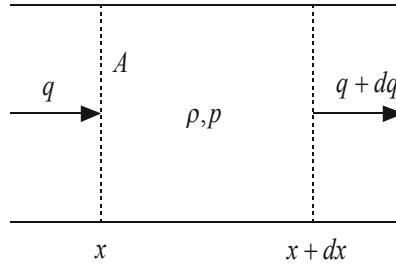


Figure 4.15: Volume element for hydraulic transmission line.

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  $\beta$ .

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  $\rho_0$ , 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

$$\frac{\partial p(x, t)}{\partial t} = -cZ_0 \frac{\partial q(x, t)}{\partial x} \quad (4.144)$$

$$\frac{\partial q(x, t)}{\partial t} = -\frac{c}{Z_0} \frac{\partial p(x, t)}{\partial x} - \frac{F[q(x, t)]}{\rho_0} \quad (4.145)$$

where the sonic velocity  $c$  and the line impedance  $Z_0$  are defined by

$$c = \sqrt{\frac{\beta}{\rho_0}}, \quad Z_0 = \frac{\rho_0 c}{A} = \frac{\sqrt{\rho_0 \beta}}{A} \quad (4.146)$$

### 4.5.3 Laplace transformed model

The PDE model (4.144, 4.145) can be Laplace transformed to give

$$\frac{\partial q(x, s)}{\partial x} = -\frac{s}{cZ_0} p(x, s) \quad (4.147)$$

$$\frac{\partial p(x, s)}{\partial x} = -\frac{Z_0 s}{c} q(x, s) - \frac{Z_0 F[q(x, s)]}{c\rho_0} \quad (4.148)$$

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

$$\frac{Z_0 \Gamma(s)^2}{LTs} q(x, s) = \frac{Z_0 s}{c} q(x, s) + \frac{Z_0 F[q(x, s)]}{c\rho_0} \quad (4.149)$$

where  $T = L/c$  is the propagation time.

The transmission line model can be written

$$\frac{\partial q(x, s)}{\partial x} = -\frac{Ts}{LZ_0} p(x, s) \quad (4.150)$$

$$\frac{\partial p(x, s)}{\partial x} = -\frac{Z_0 \Gamma(s)^2}{LTs} q(x, s) \quad (4.151)$$

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

$$L^2 \frac{\partial^2 p(x, s)}{\partial x^2} - \Gamma^2 p(x, s) = 0 \quad (4.152)$$

$$L^2 \frac{\partial^2 q(x, s)}{\partial x^2} - \Gamma^2 q(x, s) = 0 \quad (4.153)$$

**Example 65** The series impedance  $X(s)$  and the parallel admittance  $Y(s)$  are given by

$$X(s) = \frac{Z_0 \Gamma(s)^2}{LTs}, \quad Y(s) = \frac{Ts}{L(s) Z_0} \quad (4.154)$$

The characteristic impedance  $Z_c(s)$  is then found to be

$$Z_c(s) = \sqrt{\frac{X(s)}{Y(s)}} = Z_0 \frac{\Gamma(s)}{Ts} \quad (4.155)$$

#### 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

$$\frac{\partial q(x, s)}{\partial x} = -\frac{s}{cZ_0} p(x, s) \quad (4.156)$$

$$\frac{\partial p(x, s)}{\partial x} = -\frac{Z_0 s}{c} q(x, s) \quad (4.157)$$

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

$$\Gamma(s) = Ts, \quad Z_c(s) = Z_0 \quad (4.158)$$

#### 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

$$F = \rho_0 B q \quad (4.159)$$

where the friction coefficient  $B$  is

$$B = \frac{8\nu_0}{r_0^2} \quad (4.160)$$

where  $r_0$  is the radius of the pipe, and  $\nu_0$  is the kinematic viscosity. The model (4.147, 4.148) becomes

$$\frac{\partial q(x, s)}{\partial x} = -\frac{s}{cZ_0} p(x, s) \quad (4.161)$$

$$\frac{\partial p(x, s)}{\partial x} = -\frac{Z_0}{c} (s + B) q(x, s) \quad (4.162)$$

Then the propagation operator can be found according to (4.150) to be given by

$$\Gamma^2 = T^2 s (s + B) \quad (4.163)$$

From this result and (4.155) it is seen that:

With linear friction the propagation operator and the characteristic impedance are

$$\Gamma = Ts \sqrt{\frac{s+B}{s}}, \quad Z_c = Z_0 \sqrt{\frac{s+B}{s}} \quad (4.164)$$

In this case the wave equation from the lossless case is modified to

$$\frac{\partial^2 p}{\partial t^2} + B \frac{\partial p}{\partial t} - c^2 \frac{\partial^2 p(x, s)}{\partial x^2} = 0 \quad (4.165)$$

or, using the Laplace transform,

$$L^2 \frac{\partial^2 p(x, s)}{\partial x^2} = T^2 s (s + B) p(x, s) \quad (4.166)$$

#### 4.5.6 Nonlinear friction

In the case of nonlinear friction the PDE model is (Goodson and Leonard 1972)

$$\frac{\partial q}{\partial t} = -\frac{A}{\rho_0} \frac{\partial p}{\partial x} + \frac{\mu_0}{\rho_0} \left( \frac{\partial^2 q}{\partial r^2} + \frac{1}{r} \frac{\partial q}{\partial r} \right) \quad (4.167)$$

$$\frac{\partial p}{\partial t} = -\frac{\rho_0}{A} \frac{\partial q}{\partial x} - \rho_0 \left( \frac{\partial v}{\partial r} + \frac{v}{r} \right) \quad (4.168)$$

$$\frac{\partial T}{\partial t} = \alpha_0 \left( \frac{\partial^2 T}{\partial r^2} + \frac{1}{r} \frac{\partial T}{\partial r} \right) \quad (4.169)$$

In this case viscosity and heat transfer effects has been added to the model. Without further explanation we state that the propagation operator is

$$\Gamma^2 = (Ts)^2 \frac{1}{N(r\sqrt{\frac{s}{\nu}})} \quad (4.170)$$

where the function  $N$  is given by the two alternative expressions

$$N(z) = 1 - \frac{2J_1(jz)}{jzJ_0(jz)} = \frac{I_2(z)}{I_0(z)} \quad (4.171)$$

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

$$\frac{\partial}{\partial x} \begin{pmatrix} q(x, s) \\ p(x, s) \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & -\frac{Ts}{L(s)Z_0} \\ -\frac{Z_0\Gamma(s)^2}{LTs} & 0 \end{pmatrix}}_{\mathbf{A}} \begin{pmatrix} q(x, s) \\ p(x, s) \end{pmatrix} \quad (4.172)$$

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

$$a(x, s) = p(x, s) + Z_c(s)q(x, s) \quad (4.173)$$

$$b(x, s) = p(x, s) - Z_c(s)q(x, s) \quad (4.174)$$

Here

$$Z_c = Z_0 \frac{\Gamma}{Ts} \quad (4.175)$$

is the characteristic impedance of the transmission line. From 4.172) the model for the wave variables is found to be

$$\frac{\partial a(x, s)}{\partial x} = -\frac{\Gamma}{L} a(x, s) \quad (4.176)$$

$$\frac{\partial b(x, s)}{\partial x} = \frac{\Gamma}{L} b(x, s) \quad (4.177)$$

The solutions for the wave variables are given by

$$a(x, s) = \exp\left(-\Gamma \frac{x}{L}\right) a(0, s) \quad (4.178)$$

$$b(x, s) = \exp\left(-\Gamma \frac{L-x}{L}\right) b(L, s) \quad (4.179)$$

In the lossless case

$$\Gamma = Ts; \quad Z_c = Z_0 = \frac{\rho_0 c}{A} \quad (4.180)$$

The solutions are then

$$a(x, s) = \exp\left(-\frac{x}{L} Ts\right) a(0, s) \quad (4.181)$$

$$b(x, s) = \exp\left(-\frac{L-x}{L} Ts\right) b(L, s) \quad (4.182)$$

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

$$p(x, s) = \frac{1}{2} [a(x, s) + b(x, s)] \quad (4.183)$$

$$q((x, s)) = \frac{1}{2Z_c(s)} [a(x, s) - b(x, s)] \quad (4.184)$$

Suppose that the transmission line is terminated with an impedance  $Z_L(s)$  so that

$$p(L, s) = Z_L(s) q(L, s) \quad (4.185)$$

The boundary conditions for the wave variables are given by

$$a(0, s) = a_1(s) \quad (4.186)$$

$$b(L, s) = G_L(s) a(L, s) \quad (4.187)$$

where

$$G_L(s) = \frac{Z_L(s) - Z_c}{Z_L(s) + Z_c} \quad (4.188)$$

The transfer function from  $a(0, s)$  to  $b(0, s)$  is then found to be

$$\frac{b(0, s)}{a(0, s)} = \exp(-2\Gamma) G_L(s) \quad (4.189)$$

In the lossless case this gives the transfer function

$$\frac{b(0, s)}{a(0, s)} = \exp(-2Ts) G_L(s) \quad (4.190)$$

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

$$\frac{p(L, s)}{p(0, s)} = \frac{a(L, s) + b(L, s)}{a(0, s) + b(0, s)} = \frac{[1 + G_L(s)] \exp(-Ts)}{[1 + \exp(-2Ts) G_L(s)]} \quad (4.191)$$

while the transfer function from  $q(0, s)$  to  $p(0, s)$  is given by

$$\begin{aligned} \frac{p(0, s)}{q(0, s)} &= Z_c \frac{a(0, s) + b(0, s)}{a(0, s) - b(0, s)} = Z_c \frac{1 + \frac{b(0, s)}{a(0, s)}}{1 - \frac{b(0, s)}{a(0, s)}} = Z_c \frac{\exp(Ts) + \exp(-Ts) G_L(s)}{\exp(Ts) - \exp(-Ts) G_L(s)} \\ &= Z_c \frac{Z_L \cosh Ts + Z_c \sinh Ts}{Z_c \cosh Ts + Z_L \sinh Ts} \end{aligned} \quad (4.192)$$

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

$$\frac{b(0, s)}{a(0, s)} = -\exp(-2Ts) \quad (4.193)$$

while

$$\frac{p(L, s)}{p(0, s)} = 0 \quad (4.194)$$

and

$$\frac{p(0, s)}{q(0, s)} = Z_c \tanh Ts \quad (4.195)$$

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

$$\frac{b(0, s)}{a(0, s)} = \exp(-2Ts) \quad (4.196)$$

and

$$\frac{p(L, s)}{p(0, s)} = \frac{1}{\cosh(Ts)} \quad (4.197)$$

and

$$\frac{p(0, s)}{q(0, s)} = Z_c \frac{1}{\tanh Ts} \quad (4.198)$$

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

$$\frac{b(0, s)}{a(0, s)} = 0 \quad (4.199)$$

$$\frac{p(L, s)}{p(0, s)} = \exp(-Ts) \quad (4.200)$$

and

$$\frac{p(0, s)}{q(0, s)} = Z_c \quad (4.201)$$

#### 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

$$\begin{pmatrix} p_1(s) \\ p_2(s) \end{pmatrix} = \mathbf{Z}(s) \begin{pmatrix} q_1(s) \\ -q_2(s) \end{pmatrix} \quad (4.202)$$

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

$$\begin{pmatrix} q_1(s) \\ -q_2(s) \end{pmatrix} = \mathbf{Y}(s) \begin{pmatrix} p_1(s) \\ p_2(s) \end{pmatrix} \quad (4.203)$$

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

$$\begin{pmatrix} p_1(s) \\ q_2(s) \end{pmatrix} = \mathbf{H}(s) \begin{pmatrix} q_1(s) \\ p_2(s) \end{pmatrix} \quad (4.204)$$

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

$$\begin{pmatrix} p_2(s) \\ q_2(s) \end{pmatrix} = \mathbf{B}(s) \begin{pmatrix} p_1(s) \\ q_1(s) \end{pmatrix} \quad (4.205)$$

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

$$\begin{pmatrix} p_1(s) \\ p_2(s) \end{pmatrix} = Z_c(s) \begin{pmatrix} \frac{\cosh \Gamma}{\sinh \Gamma} & \frac{1}{\sinh \Gamma} \\ \frac{1}{\sinh \Gamma} & \frac{\cosh \Gamma}{\sinh \Gamma} \end{pmatrix} \begin{pmatrix} q_1(s) \\ -q_2(s) \end{pmatrix} \quad (4.206)$$

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

$$q_s = \frac{1}{2}(q_1 - q_2), \quad p_s = \frac{1}{2}(p_1 + p_2) \quad (4.207)$$

$$q_a = \frac{1}{2}(q_1 + q_2), \quad p_a = \frac{1}{2}(p_1 - p_2) \quad (4.208)$$

The transfer function model of a hydraulic transmission line can be written in the impedance form

$$p_s(s) = Z_s(s)q_s(s) \quad (4.209)$$

$$p_a(s) = Z_a(s)q_a(s) \quad (4.210)$$

where the transfer functions or impedance functions are found from (4.206–4.208) to be given by

$$Z_s(s) = \frac{Z_0 \Gamma(s)}{Ts} \left( \frac{\cosh \Gamma(s) + 1}{\sinh \Gamma(s)} \right) \quad (4.211)$$

$$Z_a(s) = \frac{Z_0 \Gamma(s)}{Ts} \left( \frac{\cosh \Gamma(s) - 1}{\sinh \Gamma(s)} \right) \quad (4.212)$$

The transfer functions  $Z_s(s)$  and  $Z_a(s)$  both have singularities for

$$\sinh \Gamma = 0 \Leftrightarrow e^{2\Gamma} = 1 \Leftrightarrow \Gamma = j\omega_{sk} \quad (4.213)$$

where the natural frequencies are

$$\omega_{sk} = k\pi, \quad k = 0, \pm 1, \pm 2, \dots \quad (4.214)$$

Note that there are infinitely many singularities with an even spacing of  $\pi$  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:

$$Z_s(s) = \frac{2Z_0}{Ts} + \sum_{k=2,4,\dots}^{\infty} \frac{4Z_0 \Gamma^2}{Ts(\Gamma^2 + \omega_{sk}^2)}, \quad Z_a(s) = \sum_{k=1,3,\dots}^{\infty} \frac{4Z_0 \Gamma^2}{Ts(\Gamma^2 + \omega_{sk}^2)} \quad (4.215)$$

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

$$\frac{f(s)}{g(s)} = \frac{A_1}{s - a_1} + \frac{A_2}{s - a_2} + \dots \Rightarrow A_i = \frac{f(a_i)}{g'(a_i)} \quad (4.216)$$

We can then find the coefficients of the partial fraction expansions from

$$\frac{\cosh \Gamma + 1}{\cosh \Gamma} = \frac{e^{2\Gamma} + 2e^\Gamma + 1}{e^{2\Gamma} + 1} = \begin{cases} 0 & \Gamma = j\omega_{sk}, k = \text{odd} \\ 2 & \Gamma = j\omega_{sk}, k = \text{even} \end{cases} \quad (4.217)$$

The coefficient for the second order terms are found for even  $k$  from

$$\frac{2}{\Gamma + j\omega_{sk}} + \frac{2}{\Gamma - j\omega_{sk}} = \frac{4\Gamma}{\Gamma^2 + \omega_{sk}^2} \quad (4.218)$$

In the same way we calculate

$$\frac{\cosh \Gamma - 1}{\cosh \Gamma} = \frac{e^{2\Gamma} - 2e^\Gamma + 1}{e^{2\Gamma} + 1} = \begin{cases} 2 & \Gamma = j\omega_{sk}, k = \text{odd} \\ 0 & \Gamma = j\omega_{sk}, k = \text{even} \end{cases} \quad (4.219)$$

and find that for odd  $k$  we have

$$\frac{2}{\Gamma + j\omega_{sk}} + \frac{2}{\Gamma - j\omega_{sk}} = \frac{4\Gamma}{\Gamma^2 + \omega_{sk}^2} \quad (4.220)$$

#### 4.5.12 Rational series expansion of admittance model

The admittance for of the transmission line model is given by (1.202) as

$$\begin{pmatrix} q_1(s) \\ -q_2(s) \end{pmatrix} = \frac{1}{Z_c} \begin{pmatrix} \frac{\cosh \Gamma}{\sinh \Gamma} & -\frac{1}{\sinh \Gamma} \\ -\frac{1}{\sinh \Gamma} & \frac{\cosh \Gamma}{\sinh \Gamma} \end{pmatrix} \begin{pmatrix} p_1(s) \\ p_2(s) \end{pmatrix} \quad (4.221)$$

Again the model is simplified by using symmetric and asymmetric variables defined in (4.207) and (4.208). Then the transfer functions become

$$q_s(s) = Y_s(s)q_s(s) \quad (4.222)$$

$$q_a(s) = Y_a(s)q_a(s) \quad (4.223)$$

where the admittances are

$$Y_s(s) = \frac{Ts}{Z_0 \Gamma} \frac{\cosh \Gamma + 1}{\sinh \Gamma} \quad (4.224)$$

$$Y_a(s) = \frac{Ts}{Z_0 \Gamma} \left( \frac{\cosh \Gamma - 1}{\sinh \Gamma} \right) \quad (4.225)$$

Using partial fraction expansion in the same way as for the impedance model we find the following rational representation of the infinite-dimensional admittances:

$$Y_s(s) = \sum_{k=1,3,\dots}^{\infty} \frac{4Ts}{Z_0 (\Gamma^2 + \omega_{sk}^2)}, \quad Y_a(s) = \frac{2Ts}{Z_0 \Gamma^2} + \sum_{k=2,4,\dots}^{\infty} \frac{4Ts}{Z_0 (\Gamma^2 + \omega_{sk}^2)} \quad (4.226)$$

### 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

$$\bar{p}(s, x) = \sum_{k=0}^N P_k(s) \phi_k(x) \quad (4.227)$$

This is used in combination with the transmission line model (4.157) which is

$$\Gamma^2 p(s) - L^2 \frac{d^2 p(s)}{dx^2} = 0 \quad (4.228)$$

The boundary conditions are supposed to be

$$\frac{\partial p(0, s)}{\partial x} = -\frac{Z_0 \Gamma(s)^2}{L T s} q(0, s), \quad \frac{\partial p(L, s)}{\partial x} = -\frac{Z_0 \Gamma(s)^2}{L T s} q(L, s) \quad (4.229)$$

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

$$\phi_k(x) = \cos\left(\frac{k\pi}{L}x\right) \quad (4.230)$$

which is a well-established result for the wave equation. These shape functions are orthogonal in the sense that

$$\int_0^L \phi_k(x) \phi_j(x) dx = \int_0^L \cos\left(\frac{k\pi}{L}x\right) \cos\left(\frac{j\pi}{L}x\right) dx = \frac{L}{2} \delta_{kj} \quad (4.231)$$

Moreover, the derivatives of the shape functions are orthogonal and satisfies

$$\int_0^L \phi'_k(x) \phi'_j(x) dx = \int_0^L \sin\left(\frac{k\pi}{L}x\right) \sin\left(\frac{j\pi}{L}x\right) dx = \frac{(k\pi)^2}{2} \delta_{kj} \quad (4.232)$$

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

$$I := \int_0^L \phi_k(x) \left( \Gamma \bar{p}(s, x) - L^2 \frac{d^2 \bar{p}(s, x)}{dx^2} \right) dx = 0 \quad (4.233)$$

As usual in the Galerkin approach the expression for the integral is developed using partial integration:

$$\begin{aligned} I &= \int_0^L \phi_k \left( \Gamma^2 \phi_k(x) \bar{p}(s, x) + L^2 \frac{d\phi_k(x)}{dx} \frac{d\bar{p}(s, x)}{dx} \right) dx \\ &\quad + \phi_k(x) L^2 \frac{\partial \bar{p}(s, x)}{\partial x} \Big|_0^L \\ &= \int_0^L \left( \Gamma^2 \phi_k(x) \bar{p}(s, x) + L^2 \frac{d\phi_k(x)}{dx} \frac{d\bar{p}(s, x)}{dx} \right) dx \\ &\quad + \frac{z_0 \Gamma^2 L}{T s} [\phi_k(L) q(L) - \phi_k(0) q(0)] \end{aligned} \quad (4.234)$$

Using the orthogonality of  $\phi_k(x)$  and  $\phi'_k(x)$  as stated in (4.231) and (4.232) we find that

$$P_k(s) \frac{L}{2} \left( \Gamma^2 + (k\pi)^2 \right) - \frac{z_0 \Gamma^2 L}{Ts} [(-1)^k q(L) + q(0)] = 0 \quad (4.235)$$

which means that the pressure coefficients are given by

$$P_k(s) = \frac{2z_0 \Gamma^2}{Ts \left( \Gamma^2 + (k\pi)^2 \right)} [q(0) + (-1)^k q(L)] \quad (4.236)$$

Then the impedance functions  $Z_s(s)$  and  $Z_a(s)$  in the model

$$p_s(s) = Z_s(s) q_s(s) \quad (4.237)$$

$$p_a(s) = Z_a(s) q_a(s) \quad (4.238)$$

for the symmetric variables are found to be

$$Z_s(s) = \frac{2Z_0}{Ts} + \sum_{k=2,4,\dots}^{\infty} \frac{4Z_0 \Gamma^2}{Ts \left( \Gamma^2 + (k\pi)^2 \right)}, \quad Z_a(s) = \sum_{k=1,3,\dots}^{\infty} \frac{4Z_0 \Gamma^2}{Ts \left( \Gamma^2 + (k\pi)^2 \right)} \quad (4.239)$$

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

$$\bar{q}(s, x) = \sum_{k=0}^{\infty} Q_k(s) \phi_k(x) \quad (4.240)$$

and the transmission line model is given by (4.157) as

$$\Gamma^2 q(s) - L^2 \frac{d^2 q(s)}{dx^2} = 0 \quad (4.241)$$

with boundary conditions

$$\frac{\partial q(0, s)}{\partial x} = -\frac{s}{cZ_0} p(0, s), \quad \frac{\partial q(L, s)}{\partial x} = -\frac{s}{cZ_0} p(L, s) \quad (4.242)$$

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

$$\phi_k(x) = \cos \left( \frac{k\pi}{L} x \right) \quad (4.243)$$

The Galerkin approach gives

$$I := \int_0^L \phi_k(x) \left( \Gamma^2 \bar{q}(x) - L^2 \frac{d^2 \bar{q}(x)}{dx^2} \right) dx = 0 \quad (4.244)$$

and

$$\begin{aligned}
 I &= \int_0^L \phi_k \left( \Gamma^2 \phi_k(x) \bar{q}(s) + L^2 \frac{d\phi_k(x)}{dx} \frac{d\bar{q}(x)}{dx} \right) dx \\
 &\quad + \phi_k(x) L^2 \frac{\partial q(x)}{\partial x} \Big|_0^L \\
 &= \int_0^L \left( \Gamma^2 \phi_k(x) \bar{q}(x) + L^2 \frac{d\phi_k(x)}{dx} \frac{d\bar{q}(x)}{dx} \right) dx \\
 &\quad + \frac{LTs}{z_0} [\phi_k(L)p(L) - \phi_k(0)p(0)]
 \end{aligned} \tag{4.245}$$

and due to the orthogonality of  $\phi_k(x)$  and  $\phi'_k(x)$ , it follows that

$$Q_k(s) \frac{L}{2} (\Gamma^2 + (k\pi)^2) = \frac{LTs}{z_0} [p(0) + (-1)^k p(L)] \tag{4.246}$$

so that

$$Q_k(s) = \frac{2Ts}{z_0 (\Gamma^2 + (k\pi)^2)} [p(0) + (-1)^k p(L)] \tag{4.247}$$

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

$$Y_s(s) = \sum_{k=1,3,\dots}^{\infty} \frac{4Ts}{Z_0 (\Gamma^2 + (k\pi)^2)}, \quad Y_a(s) = \frac{2Ts}{Z_0 \Gamma^2} + \sum_{k=2,4,\dots}^{\infty} \frac{4Ts}{Z_0 (\Gamma^2 + (k\pi)^2)} \tag{4.250}$$

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

$$I := \int_0^L \phi_k(x) \left( \Gamma \bar{p}(s, x) - L^2 \frac{d^2 \bar{p}(s, x)}{dx^2} \right) dx = 0 \quad (4.255)$$

and

$$\begin{aligned} I &= \int_0^L \phi_k \left( \Gamma^2 \phi_k(x) \bar{p}(s, x) + L^2 \frac{d\phi_k(x)}{dx} \frac{d\bar{p}(s, x)}{dx} \right) dx \\ &\quad + \phi_k(x) L^2 \frac{\partial \bar{p}(s, x)}{\partial x} \Big|_0^L \\ &= \int_0^L \left( \Gamma^2 \phi_k(x) \bar{p}(s, x) + L^2 \frac{d\phi_k(x)}{dx} \frac{d\bar{p}(s, x)}{dx} \right) dx \\ &\quad + \frac{Z_0 \Gamma^2 L}{Ts} [(-1)^k q(L)] \end{aligned} \quad (4.256)$$

and orthogonality of the shape functions gives

$$P_k(s) \frac{L}{2} \left( \Gamma^2 + \left[ \left( k - \frac{1}{2} \right) \pi \right]^2 \right) + \frac{\Gamma^2 L}{(k - \frac{1}{2}) \pi} p_1 - \frac{Z_0 \Gamma^2 L}{Ts} [(-1)^k q(L)] = 0 \quad (4.257)$$

and the pressure coefficients are found to be

$$P_k(s) = -\frac{2\Gamma^2}{\Gamma^2 + [(k - \frac{1}{2}) \pi]^2} \left( \frac{1}{(k - \frac{1}{2}) \pi} p_1 + (-1)^k \frac{Z_0}{Ts} q_2 \right) \quad (4.258)$$

The output variables  $q_1(s)$  and  $p_2(s)$  are then found to be

$$p_2(s) = p_1 + \sum_{k=1}^{\infty} (-1)^{k+1} P_k(s) \quad (4.259)$$

$$q_1(s) = -\frac{Ts}{Z_0 \Gamma^2} \sum_{k=1}^{\infty} \left( k - \frac{1}{2} \right) \pi P_k(s) \quad (4.260)$$

#### 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

$$p_s(s) = Z_s(s) q_s(s) \quad (4.261)$$

$$p_a(s) = Z_a(s) q_a(s) \quad (4.262)$$

where the impedances are given by the truncated versions

$$Z_s = \frac{2Z_0}{Ts} + \sum_{k=2,4,\dots}^N \frac{4Z_0 \Gamma^2}{Ts(\Gamma^2 + \omega_{sk}^2)}, \quad Z_a = \sum_{k=1,3,\dots}^{N-1} \frac{4Z_0 \Gamma^2}{Ts(\Gamma^2 + \omega_{sk}^2)} \quad (4.263)$$

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

$$\frac{4Z_0 \Gamma^2}{Ts(\Gamma^2 + \omega_{sk}^2)} = \frac{4Z_0 Ts}{T^2 s^2 + \omega_{sk}^2} \quad (4.264)$$

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

$$\frac{4Z_0\Gamma^2}{Ts(\Gamma^2 + \omega_{sk}^2)} = \frac{4Z_0T(s + B)}{T^2s^2 + BT^2s + \omega_{sk}^2} \quad (4.265)$$

which is also a second-order system.

With nonlinear friction it is necessary to introduce a rational approximation of  $\Gamma^2$ . An approximation due to (Woods 1983) is

$$\Gamma^2 = \frac{(Ts)^2}{1 - (1 + 2\frac{Ts}{B})^{-1/2}} \quad (4.266)$$

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

$$\frac{4Z_0\Gamma^2}{Ts(\Gamma^2 + \omega_{sk}^2)} = \frac{4T(s + B)}{(Ts)^2 + B_kT^2s + \omega_k^2} \quad (4.267)$$

where

$$B_k = \frac{1}{2}\sqrt{\omega_{sk}B} + \frac{B}{8}, \quad \omega_k = \omega_{sk} - \frac{B_k}{2} \quad (4.268)$$

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

$$Z_s = \frac{2Z_0}{Ts} + \sum_{k=2,4,\dots}^N \frac{4Z_0\Gamma^2\sigma_k}{Ts(\Gamma^2 + \omega_{sk}^2)}, \quad Z_a = \sum_{k=1,3,\dots}^{N-1} \frac{4Z_0\Gamma^2\sigma_k}{Ts(\Gamma^2 + \omega_{sk}^2)} \quad (4.269)$$

where

$$\sigma_k = \frac{\sin \beta_k}{\beta_k}, \quad \beta_k = \frac{\omega_{sk}}{N+1} \quad (4.270)$$

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

$$p_2 = p_1 - \varepsilon Z_0 q \quad (4.271)$$

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

$$b_N = \left( 8 \sum_{k=1,3,\dots}^{N-1} \frac{\sigma_k}{\omega_k^2} \right)^{-1} \quad (4.272)$$

$\Gamma^2$	$Z_s(s)/Z_0$	$Z_a(s)/Z_0$
$(Ts)^2$	$\frac{2}{Ts} + \sum_{k=2,4,\dots}^N \frac{4\sigma_k Ts}{(Ts)^2 + (k\pi)^2}$	$\sum_{k=1,3,\dots}^{N-1} \frac{4\sigma_k Ts}{(Ts)^2 + (k\pi)^2}$
$T^2 s (s + B)$	$\frac{2}{Ts} + \sum_{k=2,4,\dots}^N \frac{4\sigma_k T(s+B)}{(Ts)^2 + BT^2 s + (k\pi)^2}$	$\sum_{k=1,3,\dots}^{N-1} \frac{4\sigma_k T(s+b_N B)}{(Ts)^2 + BT^2 s + (k\pi)^2}$
$(Ts)^2 \frac{1}{N(r\sqrt{\frac{s}{\nu}})}$	$\frac{2}{Ts} + \sum_{k=2,4,\dots}^N \frac{4\sigma_k T(s+B)}{(Ts)^2 + B_k T^2 s + \omega_k^2}$	$\sum_{k=1,3,\dots}^{N-1} \frac{4\sigma_k T(s+b_N B)}{(Ts)^2 + B_k T^2 s + \omega_k^2}$

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

$$T = \frac{L}{c}, \quad B = \frac{8\nu_0}{r_0^2}, \quad B_k = \frac{1}{2}\sqrt{k\pi B} + \frac{B}{8}, \quad \omega_k^2 = k\pi - \frac{B_k}{2} \quad (4.273)$$

$$\sigma_k = \frac{\sin \beta_k}{\beta_k}, \quad \beta_k = \frac{\omega_{sk}}{N+1}, \quad b_N = \left( 8 \sum_{k=1,3,\dots}^{N-1} \frac{\sigma_k}{\omega_k^2} \right)^{-1} \quad (4.274)$$

Numerical values for a transmission line is  $L = 20$  m,  $\rho = 870$  kg/m<sup>3</sup>,  $c = 1400$  m/s,  $\nu_0 = 8 \times 10^{-5}$  m<sup>2</sup>/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

$$\frac{V}{\beta} \dot{p} = q \quad (4.275)$$

while the momentum balance of the pipe is

$$h\rho_0 \dot{q} = -Ap \quad (4.276)$$

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

$$\ddot{p} + \omega_H^2 p = 0 \quad (4.277)$$

is obtained, where

$$\omega_H^2 = \frac{A\beta}{Vh\rho_0} = \frac{Ac_0^2}{Vh} \quad (4.278)$$

is the Helmholtz frequency. Here  $c_0^2 = \beta/\rho_0$  is the sonic speed corresponding to the constant density  $\rho_0$ .

### 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

$$\frac{Ah}{\beta} \dot{p}_i = q_{i-1} - q_i \quad (4.279)$$

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

$$h\rho_0 \dot{q}_{i-1} = A(p_{i-1} - p_i) - Fh \quad (4.280)$$

for an incompressible fluid with density  $\rho_0$  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

$$\dot{p}_i = \frac{c^2 \rho_0}{Ah} (q_{i-1} - q_i) \quad (4.281)$$

$$\dot{q}_{i-1} = \frac{A}{h\rho_0} (p_{i-1} - p_i) - \frac{F}{\rho_0} \quad (4.282)$$

Note that when  $h$  tends to zero, then the model will converge to the transmission line model

$$\frac{\partial p}{\partial t} = -\frac{c^2 \rho_0}{A} \frac{\partial q}{\partial x} \quad (4.283)$$

$$\frac{\partial q}{\partial t} = -\frac{A}{\rho_0} \frac{\partial p}{\partial x} - \frac{F}{\rho_0} \quad (4.284)$$

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  $i = 1, \dots, N$ . This is shown in Figure 4.16 for  $N = 3$ . The model has  $N$  volumes so that volume  $i$

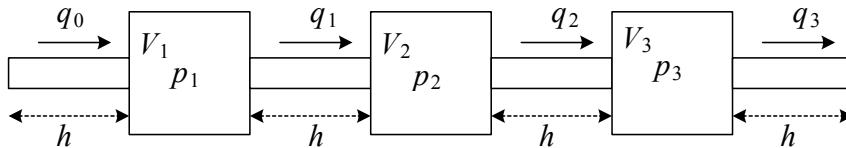


Figure 4.16: A chain of interconnected Helmholtz resonators with ducts of length  $h$  representing a transmission line in the admittance form.

is centered at  $x_i = ih$  for  $i = 1, \dots, N$ , and there are  $(N + 1)$  pipes where pipe  $i$  is centered at  $x_{i+1/2} = (i + 1/2)h$  for  $i = 0, \dots, N$  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

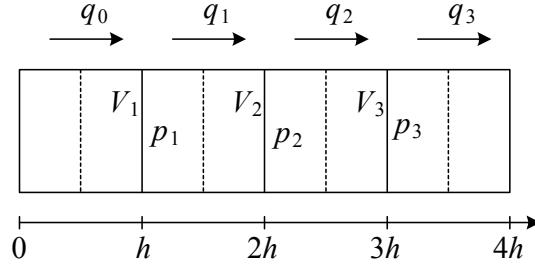


Figure 4.17: Spatial discretization of a transmission line with pressure inputs at both sides to get a description in the form of a chain of Helmholtz resonators.

when  $h$  tends to zero. The model is

$$\dot{p}_i = \frac{c^2 \rho_0}{Ah} (q_{i-1} - q_i), \quad i = 1, \dots, N \quad (4.285)$$

$$\dot{q}_{i-1} = \frac{A}{h \rho_0} (p_{i-1} - p_i) - \frac{F}{\rho_0}, \quad i = 1, \dots, N+1 \quad (4.286)$$

$$p_0 = p_{in}, \quad p_{N+1} = p_{out} \quad (4.287)$$

#### 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  $i = 1, \dots, N$ , 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  $i = 1, \dots, N - 1$ . We note that  $V = NV_h$  and  $L = Nh$ . The model is

$$\dot{p}_i = \frac{c^2 \rho_0}{Ah} (q_{i-1} - q_i), \quad i = 1, \dots, N \quad (4.288)$$

$$\dot{q}_{i-1} = \frac{A}{h \rho_0} (p_{i-1} - p_i) - \frac{F}{\rho_0}, \quad i = 2, \dots, N \quad (4.289)$$

$$q_0 = q_{in}, \quad q_N = q_{out} \quad (4.290)$$

#### 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  $i = 1, \dots, N$  with model

$$\dot{p}_i = \frac{c^2 \rho_0}{Ah} (q_{i-1} - q_i), \quad i = 1, \dots, N \quad (4.291)$$

$$\dot{q}_{i-1} = \frac{A}{h \rho_0} (p_{i-1} - p_i) - \frac{F}{\rho_0}, \quad i = 1, \dots, N \quad (4.292)$$

$$p_0 = p_{in}, \quad q_N = q_{out} \quad (4.293)$$

In this case volume  $i$  will be centered at  $x_i = ih$  for  $i = 1, \dots, N$ , and there are  $N$  pipes where pipe  $i$  is centered at  $x_{i+1/2} = (i + 1/2)h$  for  $i = 0, \dots, N - 1$ . 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

$$\begin{pmatrix} q_{i-1}(s) \\ p_i(s) \end{pmatrix} = \begin{pmatrix} \frac{A}{h\rho_0\omega_H^2} \frac{s}{1+\frac{s^2}{\omega_H^2}} & -\frac{1}{1+\frac{s^2}{\omega_H^2}} \\ \frac{1}{1+\frac{s^2}{\omega_H^2}} & \frac{h\rho_0}{A} \frac{s}{1+\frac{s^2}{\omega_H^2}} \end{pmatrix} \begin{pmatrix} p_{i-1}(s) \\ -q_i(s) \end{pmatrix} \quad (4.294)$$

where the Helmholtz frequency  $\omega_H$  is given by

$$\omega_H^2 = \frac{Ac^2}{V_h h} \Rightarrow \omega_H = \frac{c}{h} \quad (4.295)$$

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

$$\dot{p}_1 = -\frac{c^2 \rho_0}{Ah} q \quad (4.296)$$

$$\dot{p}_2 = \frac{c^2 \rho_0}{Ah} q \quad (4.297)$$

$$\dot{q} = \frac{A}{\rho_0 h} (p_1 - p_2) \quad (4.298)$$

Laplace transformation of the model, and insertion of the pressure equations in the mass flow equation leads to

$$\left( s^2 + 2\frac{c^2}{h^2} \right) q(s) = 0 \quad (4.299)$$

This system has undamped natural frequency

$$\omega_0 = \sqrt{2} \frac{c}{h} = 2\sqrt{2} \frac{c}{L} = 2.82 \frac{c}{L} \quad (4.300)$$

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

$$\omega_1 = \frac{2\pi}{T_1} = \pi \frac{c}{L} = 3.14 \frac{c}{L} \quad (4.301)$$

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

$$Ap_i = \omega_H^2 (h\rho_0 q_{i-1} - h\rho_0 q_i), \quad i = 1, 2, 3 \quad (4.302)$$

$$h\rho_0 \dot{q}_{i-1} = (Ap_{i-1} - Ap_i), \quad i = 2, 3 \quad (4.303)$$

$$\frac{d}{dt} \begin{pmatrix} Ap_1 \\ h\rho_0 q_1 \\ Ap_2 \\ h\rho_0 q_2 \\ Ap_3 \end{pmatrix} = \begin{pmatrix} 0 & -\omega_H^2 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 0 & \omega_H^2 & 0 & -\omega_H^2 & 0 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & \omega_H^2 & 0 \end{pmatrix} \begin{pmatrix} Ap_1 \\ h\rho_0 q_1 \\ Ap_2 \\ h\rho_0 q_2 \\ Ap_3 \end{pmatrix} \quad (4.304)$$

The resonance frequencies of the system matrix are

$$\omega_H = \frac{c}{h} = 3\frac{c}{L} \quad \text{and} \quad \sqrt{3}\omega_H = \frac{3\sqrt{3}c}{L} = 5.2\frac{c}{L} \quad (4.305)$$

while the first and second resonance of the exact model are

$$\omega_1 = \pi \frac{c}{L} = 3.14 \frac{c}{L} \quad \text{and} \quad \omega_2 = 2\pi \frac{c}{L} = 6.28 \frac{c}{L} \quad (4.306)$$

**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

$$\frac{d}{dt} \begin{pmatrix} Ap_1 \\ h\rho_0 q_1 \end{pmatrix} = \begin{pmatrix} 0 & -\omega_H^2 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} Ap_1 \\ h\rho_0 q_1 \end{pmatrix} \quad (4.307)$$

The resonance is at

$$\omega_1 = \frac{3}{2} \frac{c}{L} = 1.5 \frac{c}{L} \quad (4.308)$$

while the exact value for the first resonance is

$$\omega = \frac{\pi}{2} \frac{c}{L} = 1.57 \frac{c}{L} \quad (4.309)$$

## 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

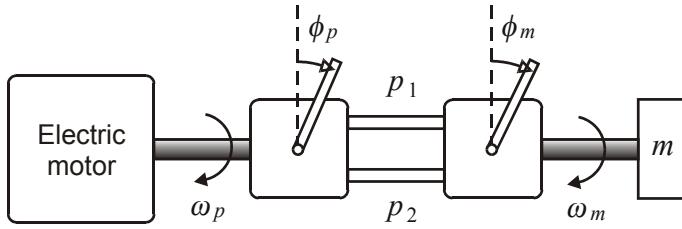


Figure 4.18: Pump controlled hydraulic motor with variable displacement in the pump and the motor. The arrangement is known as a hydrostatic gear.

rotation of the rotor can then be controlled with the displacement of the pump which is varied by the stroke angle  $\phi_p$  of the pump. The stroke of the pump is controlled with a small servomechanism, which may use a valve controlled hydraulic motor or an electrical actuator.

The dynamic model of the pump is given by (4.73–4.75). This is also the case for the hydraulic motor. In this type of system the high pressure side with pressure  $p_1$  will be driving the motor, while the pressure  $p_2$  of the return pipe from the motor to the pump will be zero. Therefore, only the mass balances of the high pressure side are included in the model. Then the model for the pump is

$$\frac{V_{1p}}{\beta} \dot{p}_{1p} = D_p \omega_p - C_{ip} p_{1p} - C_{ep} p_{1p} - q_{1p} \quad (4.310)$$

$$J_p \ddot{\omega}_p = -B_p \omega_p - D_p p_{1p} + T_{em} \quad (4.311)$$

where  $D_p$  is the pump displacement,  $\omega_p$  is the angular velocity of the pump,  $V_{1p}$  is the volume of chamber 1 of the pump,  $C_{ip}$  and  $C_{ep}$  are leakage coefficients,  $J_p$  is the inertia of the pump,  $T_{em}$  is the torque from the electrical motor, and  $B_m$  is the friction coefficient of the pump. The model for the hydraulic motor is

$$\frac{V_{1m}}{\beta} \dot{p}_{1m} = -D_m \omega_m - C_{im} p_{1m} - C_{em} p_{1m} + q_{1m} \quad (4.312)$$

$$J_m \ddot{\omega}_m = -B_m \omega_m + D_m p_{1m} - T_L \quad (4.313)$$

where  $V_{1m}$  is the volume of chamber 1 of the motor,  $C_{im}$  and  $C_{em}$  are leakage coefficients, and  $B_m$  is the friction coefficient of the motor. The displacement of the pump and motor are assumed to be given by

$$D_p = k_p \phi_p \quad (4.314)$$

$$D_m = k_m \phi_m \quad (4.315)$$

where  $\phi_p$  is the stroke angle of pump,  $\phi_m$  is the stroke angle of motor, and  $k_p$  and  $k_m$  are displacement coefficients.

In contrast to the model of the valve controlled motor it is not possible to connect the outputs of the pump model to the inputs of the hydraulic motor and vice versa. The reason for this is the model are interconnected by imposing the conditions  $q_1 = q_{1p} = q_{1m}$  and  $p_1 = p_{1p} = p_{1m}$ . However,  $q_1$  will then be input to both models, and  $p_1$  will be output from both models. The solution to this problem is to combine the two mass balances

(4.310) and (4.312) by adding them. This gives the model

$$\frac{V_1}{\beta} \dot{p}_1 = -D_m \omega_m + D_p \omega_p - C_1 p_1 \quad (4.316)$$

$$J_p \dot{\omega}_p = T_{em} - B_p \omega_p - D_p p_1 \quad (4.317)$$

$$J_m \dot{\omega}_m = -B_m \omega_m + D_m p_1 - T_L \quad (4.318)$$

where  $V_1 = V_{1p} + V_{1m}$ ,  $C_{it} = C_{ip} + C_{im}$  and  $C_1 = C_{it} + C_{ep} + C_{em}$ .

**Example 70** Suppose that the motor displacement  $D_m$  and the pump speed  $\omega_p$  are constants, and that the pump displacement is  $D_p = k_p \phi_p$  where the stroke angle  $\phi_p$  of the pump is the control input. Then the model is

$$\frac{V_1}{\beta} \dot{p}_1 = -D_m \omega_m - C_1 p_1 + k_p \omega_p \phi_p \quad (4.319)$$

$$J_m \dot{\omega}_m = -B_m \omega_m + D_m p_1 - T_L \quad (4.320)$$

**Example 71** If the pump and the hydraulic motor are connected with a transmission line on the high pressure side, then there will be independent mass balances for the pump and the motor. The high pressure port of the pump would then have pressure  $p_p$  and flow  $q_p$  and the high pressure port of the motor would have pressure  $p_m$  and flow  $q_m$ . The models of the pump and the motor would be

$$\frac{V_{1p}}{\beta} \dot{p}_{1p} = D_p \omega_p - C_{ip} p_{1p} - C_{ep} p_{1p} - q_{1p} \quad (4.321)$$

$$J_p \dot{\omega}_p = -B_p \omega_p - D_p p_{1p} + T_{em} \quad (4.322)$$

and

$$\frac{V_{1m}}{\beta} \dot{p}_{1m} = -D_m \omega_m - C_{im} p_{1m} - C_{em} p_{1m} + q_{1m} \quad (4.323)$$

$$J_m \dot{\omega}_m = -B_m \omega_m + D_m p_{1m} - T_L \quad (4.324)$$

These model have flows as inputs and pressures as outputs. The transmission line model of the high pressure line would be an admittance model

$$\begin{pmatrix} q_{1p}(s) \\ q_{1m}(s) \end{pmatrix} = \frac{1}{Z_c} \begin{pmatrix} \frac{\cosh \Gamma}{\sinh \Gamma} & -\frac{1}{\sinh \Gamma} \\ \frac{1}{\sinh \Gamma} & -\frac{\cosh \Gamma}{\sinh \Gamma} \end{pmatrix} \begin{pmatrix} p_{1p}(s) \\ p_{1m}(s) \end{pmatrix} \quad (4.325)$$

with the pressures  $p_{1p}$  and  $p_{1m}$  as inputs and the flows  $q_{1p}$  and  $q_{1m}$  as outputs. The admittance model is irrational due to the transcendental functions  $\cosh \Gamma$  and  $\sinh \Gamma$ . Rational approximation of this irrational model must be used to develop a simulation model.

**Example 72** A load with inertia  $J_1$  and angular velocity  $\omega_1$  with elastic transmission can be connected to the motor by formulating the load with a port with the motor speed  $\omega_m$  as the input variable, and the load torque  $T_L$  as the output variable. This is done by the load model

$$J_1 \dot{\omega}_1 = T_L - T_1 \quad (4.326)$$

$$\dot{\theta}_1 = \omega_1 \quad (4.327)$$

$$T_L = D_1 (\omega_m - \omega_1) + K_1 (\theta_m - \theta_1) \quad (4.328)$$

We may add on any number of additional degrees of freedom as two-ports

$$J_i \dot{\omega}_i = T_{i-1} - T_i \quad (4.329)$$

$$\dot{\theta}_i = \omega_i \quad (4.330)$$

$$T_i = D_i (\omega_{i-1} - \omega_i) + K_i (\theta_{i-1} - \theta_i) \quad (4.331)$$

with port variables  $T_{i-1}$  and  $\omega_{i-1}$  at the input and  $T_i$  and  $\omega_i$  at the output.

**Example 73** Consider the connection of a load with inertia  $J_1$  with a stiff connection to the motor so that the load has the same shaft speed as the motor. Then there would not be an extra state because of the load, and the load is included in the model by adding the load inertia to the motor inertia in the equation of motion for the motor shaft, which gives

$$(J_m + J_1) \dot{\omega}_m = -B_m \omega_m + D_m p_m - T_L \quad (4.332)$$

#### 4.7.3 Cylinder with balance valve

Hydraulics for heavy lifting with a crane will often be implemented with a hydraulic cylinder with a single-rod piston. To prevent problems with hanging loads the cylinder may be connected to a balance valve on the lifting side of the piston. In this section this type of system will be controlled with a valve. To establish a object oriented model we first define the ports of the components, and then the ports will be connected, and finally we present the models for each of the components.

The inlet port of the cylinder has pressure  $p_1$  and flow  $q_1$  into the chamber, the outlet port has pressure  $p_2$  and flow  $q_2$  out of the chamber. The balance valve has input port pressure  $p_{BV1}$  and flow  $q_{BV}$  into the valve, and outlet port with pressure  $p_{BV2}$  and flow  $q_{BV}$  out of the valve. The control valve has one port  $A$  with pressure  $p_A$  and flow  $q_A$  out of the valve, and one port  $B$  with pressure  $p_B$  and flow  $q_B$  into the valve. The pipes from the control valve to the motor and balance valve are denoted  $A$  and  $B$ . The volume of pipe  $A$  is  $V_A$  and the volume of pipe  $B$  is  $V_B$ . The inlet of pipe  $A$  has pressure  $p_A$  and flow  $q_{A1}$  into the pipe, and the outlet port has pressure  $p_A$  and flow  $q_{A2}$  out of the pipe. The inlet of pipe  $B$  has pressure  $p_B$  and flow  $q_{B1}$  into the pipe, and the outlet port has pressure  $p_B$  and flow  $q_{B2}$  out of the pipe.

The control valve port  $A$  is connected to pipe  $A$ , and the valve port  $B$  is connected to pipe  $B$ . Pipe  $A$  is connected to the inlet port of the balance valve so that  $p_{BV1} = p_A$ , and pipe  $B$  is connected to the outlet port of the motor so that  $p_B = p_2$ . The outlet port of the balance valve is connected to the inlet port of the motor so that  $p_{BV2} = p_1$  and  $q_{BV} = q_1$ .

The flows  $q_A$  and  $q_B$  are found from the orifice equations

$$\begin{aligned} q_a &= C_d A_a(x_v) \sqrt{\frac{2}{\rho} (p_s - p_A)} \\ q_b &= C_d A_b(x_v) \sqrt{\frac{2}{\rho} (p_s - p_B)} \\ q_c &= C_d A_c(x_v) \sqrt{\frac{2}{\rho} (p_A - p_r)} \\ q_d &= C_d A_d(x_v) \sqrt{\frac{2}{\rho} (p_B - p_r)} \end{aligned} \quad (4.333)$$

and

$$q_A = q_a - q_c, \quad q_B = q_d - q_b \quad (4.334)$$

where  $x_v$  is an input to the model.

The pressures  $p_A$  is found from the pipe model that is given by the mass balance

$$\frac{V_A}{\beta} \dot{p}_A = q_A - q_1 \quad (4.335)$$

Note that there is no model for pipe  $B$  because pipe  $B$  is connected to the outlet chamber of the motor so that  $p_B = p_2$ . Instead, the volume of pipe  $B$  is included in the chamber volume of chamber 2 in the motor.

The volumetric flow  $q_1$  is found from the flow characteristic of the the balance valve, which is

$$q_1 = \begin{cases} C_d x_{bv1} b \sqrt{\frac{2}{\rho} (p_1 - p_A)} & p_A > p_1 \\ -C_d A_c \sqrt{\frac{2}{\rho} (p_A - p_1)} & p_1 < p_A \end{cases} \quad (4.336)$$

where the spool position of the balance valve is

$$x_{bv1} = \frac{A_r}{K} [p_1 - p_{01} + Rp_2 - p_A (R + 1)], \quad 0 \leq x_{bv1} \leq x_{bv1,\max}$$

The piston moves in the vertical direction, and the position of the piston is denoted by  $x_p$  which is positive in the upwards direction. The cylinder has one chamber with area  $A_1$ , pressure  $p_1$  and volume  $V_1 = V_{10} + A_1 x_p$ . The other chamber has area  $A_2 = A_1 - A_r$  where  $A_r$  is the area of the rod, pressure  $p_2$  and volume  $V_2 = V_{20} - A_2 x_p$ . The model for the cylinder is then

$$\frac{V_{10} + A_1 x_p}{\beta} \dot{p}_1 = -C_{im}(p_1 - p_2) - C_{em} p_1 - A_1 \dot{x}_p + q_1 \quad (4.337)$$

$$\frac{V_B + V_{20} - A_2 x_p}{\beta} \dot{p}_2 = -C_{im}(p_2 - p_1) - C_{em} p_2 + A_2 \dot{x}_p - q_2 \quad (4.338)$$

$$m_t \ddot{x}_p = -B_p \dot{x}_p + A_1 p_1 - A_2 p_2 - F_L \quad (4.339)$$

where  $F_L$  is an input to the model.

**Example 74** If a balance valve had been added at the outlet port of the motor, then pressure  $p_B$  would have to be computed from a line model

$$\frac{V_B}{\beta} \dot{p}_B = q_2 - q_B \quad (4.340)$$

and the mass balance of chamber 2 of the motor would have been

$$\frac{V_{20} - A_2 x_p}{\beta} \dot{p}_2 = -C_{im}(p_2 - p_1) - C_{em} p_2 + A_2 \dot{x}_p - q_2 \quad (4.341)$$

The flow  $q_2$  would then be found from the balance valve characteristic of the balance valve at port 2 as

$$q_2 = \begin{cases} -C_d x_{bv2} b \sqrt{\frac{2}{\rho} (p_2 - p_B)} & p_B > p_2 \\ C_d A_c \sqrt{\frac{2}{\rho} (p_B - p_2)} & p_2 < p_B \end{cases} \quad (4.342)$$

where the spool position of the balance valve is

$$x_{bv2} = \frac{A_r}{K} [p_2 - p_{02} + Rp_1 - p_B (R + 1)], \quad 0 \leq x_{bv2} \leq x_{bv2,\max}$$



# 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  $5^\circ - 10^\circ$ , and the width is typical  $10 \mu\text{m}$ . 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*.

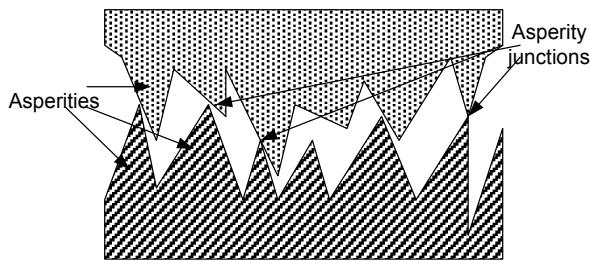


Figure 5.1: The asperities and junctions of two bodies in contact, viewed at a microscopical scale

The tangential force  $F_t(x)$  can in this regime be approximated by

$$F_t(x) = -k_t x \quad (5.1)$$

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

$$k_t = \frac{F_b}{x_b} \quad (5.2)$$

where  $x_b$  is the break-away displacement. Before break-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 four regimes:

**I. Static friction** Elastic deformation of the asperities, the Dahl effect.

**II. Boundary lubrication** For very low velocities, no fluid lubrication occurs, 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

$$F_f = F_c \operatorname{sgn}(v), \quad v \neq 0 \quad (5.3)$$

where the Coulomb force  $F_c$  is given by

$$F_c = \mu F_N \quad (5.4)$$

Here  $\mu$  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

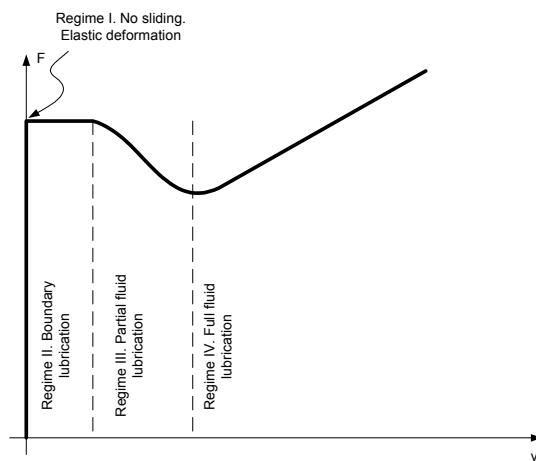


Figure 5.2: The generalized Stribeck curve, showing friction as a function of velocity for low velocities, (Armstrong-Hélovry et al. 1994).

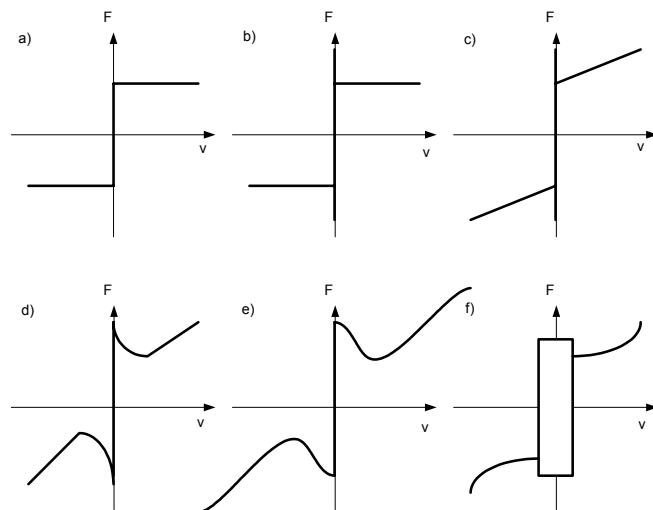


Figure 5.3: Static friction models: a) Coulomb friction b) Coulomb+stiction c) Coulomb+stiction+viscous d) Stribeck effect e) Hess and Soom; Armstrong f) Karnopp model

be defined as the shear strength of the asperity junction areas, and the friction force is proportional to the true area of contact  $A_c$

$$F_f = A_c f_s \quad (5.5)$$

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

$$A_c = \frac{F_N}{p_y} \quad (5.6)$$

where  $F_N$  is the load, and  $p_y$  is the yield pressure, a constant material property. Combining (5.6) with (5.5) gives

$$F_f = \frac{F_N}{p_y} f_s = \mu F_N \quad (5.7)$$

where the friction coefficient  $\mu$  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-Hélouvy 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

$$F_{fv} = F_v v \quad (5.8)$$

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

$$F_{fv} = F_v |v|^{\delta_v} \operatorname{sgn}(v) \quad (5.9)$$

where  $0 < \delta_v \leq 1$  is a constant.

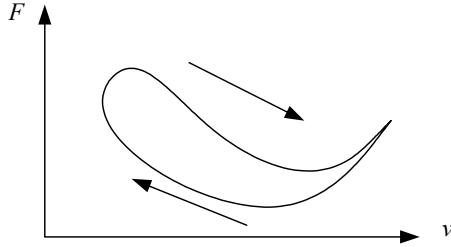


Figure 5.4: Frictional lag

### 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-Hélouvy (1990) propose to use

$$F_f = \left[ F_c + (F_s - F_c)e^{-(v/v_s)^2} \right] \operatorname{sgn}(v), \quad v \neq 0 \quad (5.10)$$

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  $|F_f| \rightarrow |F_c|$  when the velocity tends to zero. Hess and Soom (1990) propose the expression

$$F_f = \left[ F_c + \frac{(F_s - F_c)}{1 + (v/v_s)^2} \right] \operatorname{sgn}(v), \quad v \neq 0 \quad (5.11)$$

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

$$F = F_c \operatorname{sgn}(v) + F_v v \quad (5.12)$$

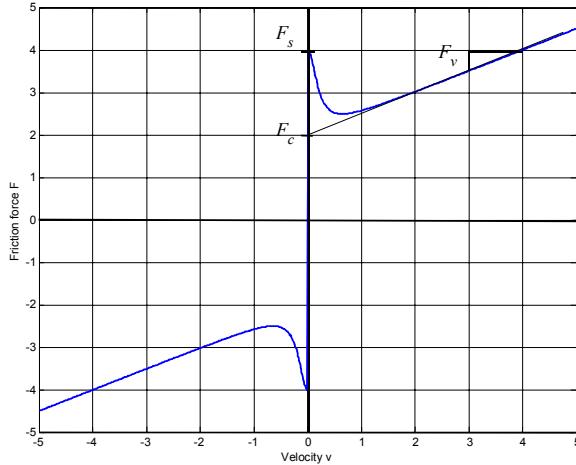


Figure 5.5: An example of a friction map

The Stribeck effect can also be included using (5.11), which gives (Hess and Soom 1990)

$$F = \left( F_c + \frac{(F_s - F_c)}{1 + (v/v_s)^2} \right) \operatorname{sgn}(v) + F_v v \quad (5.13)$$

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:

$$F_f(v, F) = \begin{cases} F_f(v), & |v| \geq \eta \\ F_f(F), & |v| < \eta \end{cases} \quad (5.14)$$

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-Hélouvy 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:

$$F(x) = \sigma_0 x, \quad (5.15)$$

when sticking, and

$$F = \left( F_c + F_s(\gamma, t_d) \frac{1}{1 + (v(t - \tau_l)/v_s)^2} \right) \text{sgn}(v) + F_v v \quad (5.16)$$

where

$$F_s(\gamma, t_d) = F_{s,a} + \left( F_{s,\infty} - F_{s,a} \frac{t_d}{t_d + \gamma} \right) \quad (5.17)$$

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  $F_f = F_c \text{sgn}(v)$ . 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

$$F_c \text{sgn}(v) = \begin{cases} -F_c, & v < 0 \\ 0, & v = 0 \\ F_c, & 0 < v \end{cases} \quad (5.18)$$

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

$$m\dot{v} = F_a - F_f \quad (5.19)$$

Now, suppose that  $v = 0$ . If the friction force is given by  $F_f = F_c \text{sgn}(v)$  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

$$m\dot{v} = \begin{cases} F_a + F_c, & v < 0 \\ F_a - F_c, & v > 0 \end{cases} \quad (5.20)$$

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

$$m\dot{v} = 0, \quad v = 0 \text{ and } |F_a| \leq F_c \quad (5.21)$$

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

$$F_f = \begin{cases} \text{sat}(F_a, F_c) & \text{when } v = 0 \\ F_c\text{sgn}(v) & \text{else} \end{cases} \quad (5.22)$$

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

$$F_f = \begin{cases} \text{sat}(F_a, F_c) & \text{when } v = 0 \\ \left[ F_c + (F_s - F_c)e^{-(v/v_s)^2} \right] \text{sgn}(v) & \text{else} \end{cases} \quad (5.23)$$

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

$$F_f = \begin{cases} \text{sat}(F_a, F_c) & \text{when } |v| \leq \delta \\ F_c \text{sgn}(v) & \text{else} \end{cases} \quad (5.24)$$

where the magnitude  $\delta$  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

$$F_f(v) \in \text{sector}(k_1, k_2) \iff k_1 v^2 < F_f(v)v < k_2 v^2 \quad (5.25)$$

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

$$F_f(v) \in \text{sector}[k_1, \infty), \text{ where } 0 < k_1 \leq F_v \quad (5.26)$$

Calculating the power  $F_f(v)v$  for a static friction model where  $F_f(v)$  satisfies (5.26), and integrating, we get

$$\int_0^T F_f(v)v dt > \int_0^T k_1 v^2 dt, \text{ where } 0 < k_1 \leq F_v \quad (5.27)$$

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:

$$\frac{dF}{dt} = \frac{\partial F}{\partial t} + \frac{\partial F}{\partial x} \frac{dx}{dt} \quad (5.28)$$

Then if  $\partial F/\partial t = 0$ , the expression

$$\frac{dF}{dt} = \frac{dF}{dx} \frac{dx}{dt} \quad (5.29)$$

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:

$$x \ll 1 \Rightarrow |F| = |\sigma x| \ll F_c \quad (5.30)$$

with  $\sigma$  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

$$\begin{aligned} \frac{dF}{dx} &= \sigma \left| 1 - \frac{F}{F_c} \operatorname{sgn} \frac{dx}{dt} \right|^{\alpha} \operatorname{sgn} \left( 1 - \frac{F}{F_c} \operatorname{sgn} \frac{dx}{dt} \right) \\ &= \sigma \left( 1 - \frac{F}{F_c} \operatorname{sgn} \frac{dx}{dt} \right)^{\alpha} \operatorname{sgn}^{\alpha+1} \left( 1 - \frac{F}{F_c} \operatorname{sgn} \frac{dx}{dt} \right) \end{aligned} \quad (5.31)$$

By using the fact that  $F < F_c$  it follows that  $\operatorname{sgn} \left( 1 - \frac{F}{F_c} \operatorname{sgn} \frac{dx}{dt} \right) > 0$ , and (5.31) simplifies to

$$\frac{dF}{dx} = \sigma \left( 1 - \frac{F}{F_c} \operatorname{sgn} \frac{dx}{dt} \right)^{\alpha} \quad (5.32)$$

The constant  $\alpha$  depends on the material of the solid,  $\alpha \geq 1$  describes ductile type materials, while  $\alpha < 1$  describes brittle type materials. Applications of the model, however, typically employ  $\alpha = 1$ , so that

$$\frac{dF}{dx} = \sigma \left( 1 - \frac{F}{F_c} \operatorname{sgn} \frac{dx}{dt} \right) \quad (5.33)$$

It follows from (5.33) that in steady state

$$F_{ss} = F_c \operatorname{sgn} \frac{dx}{dt} = F_c \operatorname{sgn}(v) \quad (5.34)$$

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  $\sigma$  being the spring stiffness. For large displacements the friction force 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

$$\frac{dF}{dt} = \sigma \left( v - |v| \frac{F}{F_c} \right) \quad (5.35)$$

The Dahl model can be written in the form

$$\frac{dF}{dt} = \sigma \frac{|v|}{F_c} (F_f - F) \quad (5.36)$$

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

$$-\frac{F_c}{\sigma} \ln \left( \sigma - \frac{\sigma}{F_c} F \right) = x + C' \quad (5.37)$$

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

$$F = F_c \left( 1 - e^{-\frac{\sigma}{F_c} x} \right) \quad (5.38)$$

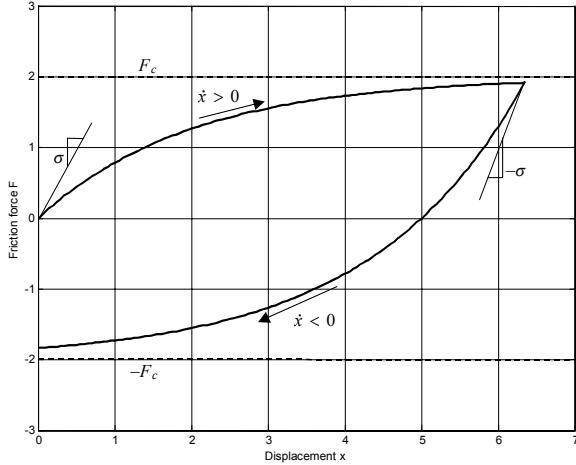


Figure 5.6: The Dahl friction model for positive and negative velocity.

By doing the same calculations for  $\dot{x} < 0$ , we find that

$$F = F_c \operatorname{sgn}(v) \left( 1 - e^{-\frac{\sigma}{F_c} x \operatorname{sign}(v)} \right) \quad (5.39)$$

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

$$\dot{V} = \frac{1}{\sigma_0} F \dot{F} = Fv - \frac{F^2 |v|}{F_c} \quad (5.40)$$

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

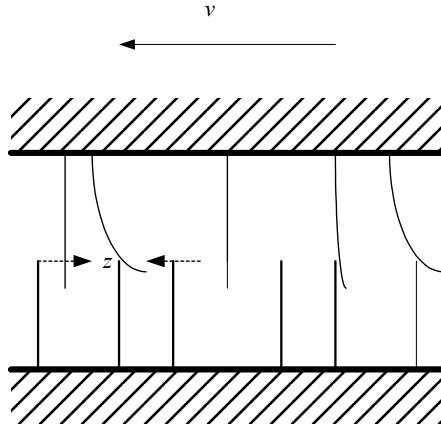


Figure 5.7: The asperity junctions of two bodies in contact are modeled as elastic bristles. For simplicity, only the upper body is moving in this figure, and the bristles of the lower body are assumed rigid.

other the strain in the bond increases and the bristles acts as springs giving rise to the friction force. The force is given by

$$F = \sum_{i=1}^N \sigma_0 (x_i - b_i) \quad (5.41)$$

where  $N$  is the number of bristles,  $\sigma_0$  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  $\sigma_0$  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

$$\dot{z} = v - \sigma_0 \frac{|v|}{g(v)} z \quad (5.42)$$

where the function  $g(v)$  is selected to be

$$\sigma_0 g(v) = F_c + (F_s - F_c) e^{-(v/v_s)^2} \quad (5.43)$$

to account for stiction and the Stribeck effect as in (5.10). The friction force is given by

$$F = \sigma_0 z + \sigma_1 \dot{z} + \sigma_2 v. \quad (5.44)$$

The following property of the model is noted: It is seen from (5.42) that the model for  $z$  can be written

$$\dot{z} = \frac{|v|}{g(v)} [g(v)\text{sgn}(v) - \sigma_0 z] \quad (5.45)$$

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

$$|\sigma_0 z(t)| \leq g_{\max} \quad (5.46)$$

It is also noted that

$$F_c \leq \sigma_0 g(v) \leq F_s \quad (5.47)$$

It is seen that the stationary solution of (5.42) is

$$z_{ss} = g(v)\text{sgn}(v). \quad (5.48)$$

This shows that the term  $\sigma_2 v$  in (5.44) will tend to represent stiction and the Stribeck effect, while  $\sigma_2 v$  term will represent viscous friction. This is characterized by the six parameters  $\sigma_0$ ,  $\sigma_1$ ,  $\sigma_2$ ,  $v_s$ ,  $F_s$  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

$$\int_0^T v F dt = \int_0^T \sigma_0 v z dt + \int_0^T v \left( \sigma_1 \frac{dz}{dt} + \sigma_2 v \right) dt \quad (5.49)$$

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

$$\begin{aligned} \int_0^T \sigma_0 v z dt &= \int_0^T \sigma_0 z \left( \dot{z} + \sigma_0 \frac{|v|}{g(v)} z \right) dt \\ &= \frac{\sigma_0}{2} [z^2(T) - z^2(0)] + \int_0^T z^2 \sigma_0 \frac{|v|}{g(v)} dt \\ &\geq -\frac{\sigma_0}{2} z^2(0) \end{aligned} \quad (5.50)$$

The second term on the right side of (5.49) is somewhat more involved. We find that

$$\int_0^T v \left( \sigma_1 \frac{dz}{dt} + \sigma_2 v \right) dt = \int_0^T v \left( (\sigma_1 + \sigma_2) v - \sigma_1 \frac{|v|}{g(v)} \sigma_0 z \right) dt \quad (5.51)$$

Using (5.46) we find that

$$\int_0^T v \left( \sigma_1 \frac{dz}{dt} + \sigma_2 v \right) dt \geq \int_0^T v^2 \left( (\sigma_1 + \sigma_2) - \sigma_1 \frac{g_{\max}}{g_{\min}} \sigma_0 z \right) dt \quad (5.52)$$

This implies that the LuGre model is passive from  $v$  to  $F$  if

$$(\sigma_1 + \sigma_2) - \sigma_1 \frac{g_{\max}}{g_{\min}} \geq 0 \Rightarrow \sigma_1 \leq \sigma_2 \frac{g_{\min}}{g_{\max} - g_{\min}} \quad (5.53)$$

Insertion of the maximum and minimum values of  $g$  according to (5.47) leads to the conclusion that the LuGre model is passive if

$$\sigma_1 \leq \sigma_2 \frac{F_c}{F_s - F_c} \quad (5.54)$$

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

$$\dot{z} = v \left( 1 - \alpha(z, v) \frac{\sigma_0 \operatorname{sgn}(v)}{g(v)} z \right)^i \quad (5.55)$$

$$F = \sigma_0 z + \sigma_1 \frac{dz}{dt} + \sigma_2 v \quad (5.56)$$

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

$$\alpha(z, v) = \begin{cases} \begin{cases} 0 & |z| \leq z_b \\ 0 < \alpha < 1 & z_b < |z| < z_{\max}(v) \\ 1 & |z| \geq z_{\max}(v) \end{cases}, \operatorname{sgn}(v) = \operatorname{sgn}(z) \\ 0, \operatorname{sgn}(v) \neq \operatorname{sgn}(z) \end{cases}, \quad (5.57)$$

where

$$0 < z_b < z_{\max}(v) = \frac{g(v)}{\sigma_0}, \forall v \in \mathbb{R}. \quad (5.58)$$

An example of the term  $\alpha(z, v)$  is

$$\alpha(z, v) = \begin{cases} \begin{cases} 0 & |z| \leq z_b \\ \frac{1}{2} \sin \left( \pi \frac{z - \left( \frac{z_{\max} + z_b}{2} \right)}{z_{\max} - z_b} \right) + \frac{1}{2} & z_b < |z| < z_{\max}(v) \\ 1 & |z| \geq z_{\max}(v) \end{cases}, \operatorname{sgn}(v) = \operatorname{sgn}(z) \\ 0, \operatorname{sgn}(v) \neq \operatorname{sgn}(z) \end{cases}$$

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

$$\sigma_1 \leq \sigma_2 \frac{F_c}{F_s - F_c} \quad (5.59)$$

# **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

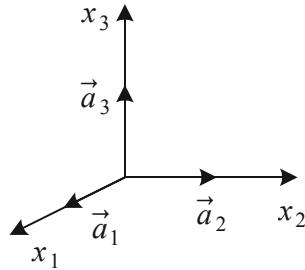
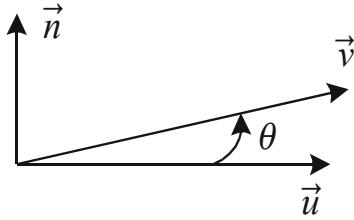


Figure 6.1: The coordinate frame  $a$ .

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

$$\vec{u} = u_1 \vec{a}_1 + u_2 \vec{a}_2 + u_3 \vec{a}_3 \quad (6.1)$$

where

$$u_i = \vec{u} \cdot \vec{a}_i, \quad i \in \{1, 2, 3\} \quad (6.2)$$

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

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \quad (6.3)$$

### 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

$$\vec{u} \cdot \vec{v} = |\vec{u}| |\vec{v}| \cos \theta \quad (6.4)$$

where  $\theta$  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

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \quad (6.5)$$

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

$$\begin{aligned} \vec{u} \cdot \vec{v} &= (u_1 \vec{a}_1 + u_2 \vec{a}_2 + u_3 \vec{a}_3) \cdot (v_1 \vec{a}_1 + v_2 \vec{a}_2 + v_3 \vec{a}_3) \\ &= u_1 v_1 + u_2 v_2 + u_3 v_3 \\ &= \mathbf{u}^T \mathbf{v} \end{aligned}$$

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

$$\vec{u} \cdot \vec{v} = \sum_{i=1}^3 u_i v_i = \mathbf{u}^T \mathbf{v} \quad (6.6)$$

### 6.2.3 The vector cross product

The *vector cross product* is given in the coordinate-free form by

$$\vec{u} \times \vec{v} = \vec{n} |\vec{u}| |\vec{v}| \sin \theta \quad (6.7)$$

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

$$\vec{w} = \vec{u} \times \vec{v} = \begin{vmatrix} \vec{a}_1 & \vec{a}_2 & \vec{a}_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} \quad (6.8)$$

In component form this may alternatively be expressed by introducing the *permutation symbol*

$$\varepsilon_{ijk} = \begin{cases} 1 & \text{when } i, j, k \text{ is a cyclic permutation} \\ -1 & \text{when } i, j, k \text{ is not a cyclic permutation} \\ 0 & \text{when } i = j, i = k \text{ or } j = k \end{cases} \quad (6.9)$$

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

$$\varepsilon_{ijk} = -\varepsilon_{jik} = -\varepsilon_{ikj} \quad (6.10)$$

$$\varepsilon_{ijk} = \varepsilon_{jki} = \varepsilon_{kij} \quad (6.11)$$

Then the components of  $\vec{w} = \vec{u} \times \vec{v}$  are given by

$$w_i = \sum_{j=1}^3 \sum_{k=1}^3 \varepsilon_{ijk} u_j v_k \quad (6.12)$$

and the vector may be written

$$\vec{w} = \sum_{i=1}^3 \sum_{j=1}^3 \sum_{k=1}^3 \varepsilon_{ijk} \vec{a}_i u_j v_k \quad (6.13)$$

In coordinate vector notation we introduce the *skew-symmetric form* of the coordinate vector  $\mathbf{u}$  defined by

$$\mathbf{u}^\times := \begin{pmatrix} 0 & -u_3 & u_2 \\ u_3 & 0 & -u_1 \\ -u_2 & u_1 & 0 \end{pmatrix} \quad (6.14)$$

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:

$$\vec{w} = \vec{u} \times \vec{v} \Leftrightarrow w_i = \sum_{i=1}^3 \sum_{j=1}^3 \sum_{k=1}^3 \varepsilon_{ijk} u_j v_k \Leftrightarrow \mathbf{w} = \mathbf{u}^\times \mathbf{v} \quad (6.15)$$

**Example 76** Orthogonal unit vectors  $\vec{a}_1, \vec{a}_2, \vec{a}_3$  satisfy

$$\vec{a}_i \cdot \vec{a}_j = \delta_{ij} = \begin{cases} 1, & \text{when } i = j \\ 0, & \text{when } i \neq j \end{cases} \quad (6.16)$$

and

$$\vec{a}_1 \times \vec{a}_2 = \vec{a}_3, \quad \vec{a}_1 \times \vec{a}_3 = -\vec{a}_2 \quad (6.17)$$

$$\vec{a}_2 \times \vec{a}_3 = \vec{a}_1, \quad \vec{a}_2 \times \vec{a}_1 = -\vec{a}_3 \quad (6.18)$$

$$\vec{a}_3 \times \vec{a}_1 = \vec{a}_2, \quad \vec{a}_3 \times \vec{a}_2 = -\vec{a}_1 \quad (6.19)$$

**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

$$(\mathbf{u}^\times)_{ik} = \varepsilon_{ijk} u_j \quad (6.20)$$

$$u_j = \frac{1}{2} \varepsilon_{ijk} (\mathbf{u}^\times)_{ik} \quad (6.21)$$

**Example 78** For three arbitrary vectors  $\vec{a}, \vec{b}, \vec{c}$  the vector cross product satisfies

$$\vec{a} \times (\vec{b} \times \vec{c}) = \vec{b} \vec{a} \cdot \vec{c} - \vec{a} \cdot \vec{b} \vec{c} \quad (6.22)$$

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

$$\mathbf{a}^\times \mathbf{b}^\times \mathbf{c} = \mathbf{b} \mathbf{a}^T \mathbf{c} - \mathbf{a}^T \mathbf{b} \mathbf{c} = (\mathbf{b} \mathbf{a}^T - \mathbf{a}^T \mathbf{b} \mathbf{I}) \mathbf{c} \quad (6.23)$$

which implies

$$\mathbf{a}^\times \mathbf{b}^\times = \mathbf{b} \mathbf{a}^T - \mathbf{a}^T \mathbf{b} \mathbf{I} \quad (6.24)$$

In particular we note that

$$\mathbf{a}^\times \mathbf{a}^\times = \mathbf{a} \mathbf{a}^T - \mathbf{a}^T \mathbf{a} \mathbf{I}. \quad (6.25)$$

**Example 79** From (6.25) it follows that

$$\mathbf{a}^\times \mathbf{a}^\times \mathbf{a}^\times = \mathbf{a}^\times (\mathbf{a} \mathbf{a}^T - \mathbf{a}^T \mathbf{a} \mathbf{I}) = -(\mathbf{a}^T \mathbf{a}) \mathbf{a}^\times \quad (6.26)$$

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

$$\mathbf{k}^\times \mathbf{k}^\times \mathbf{k}^\times = -\mathbf{k}^\times \quad (6.27)$$

**Example 80** Let  $\vec{a}$ ,  $\vec{b}$  and  $\vec{c}$  be three arbitrary vectors. The Jacobi identity is written

$$\vec{a} \times (\vec{b} \times \vec{c}) + \vec{b} \times (\vec{c} \times \vec{a}) + \vec{c} \times (\vec{a} \times \vec{b}) = \vec{0} \quad (6.28)$$

This identity is established from (6.22) which gives

$$\vec{b}\vec{a} \cdot \vec{c} - \vec{a} \cdot \vec{b}\vec{c} + \vec{c}\vec{b} \cdot \vec{a} - \vec{b} \cdot \vec{c}\vec{a} + \vec{a}\vec{c} \cdot \vec{b} - \vec{c} \cdot \vec{a}\vec{b} = \vec{0} \quad (6.29)$$

The coordinate form of the Jacobi identity is

$$\mathbf{a}^\times \mathbf{b}^\times \mathbf{c} + \mathbf{b}^\times \mathbf{c}^\times \mathbf{a} + \mathbf{c}^\times \mathbf{a}^\times \mathbf{b} = \mathbf{0} \quad (6.30)$$

**Example 81** The Jacobi identity implies that

$$(\vec{a} \times \vec{b}) \times \vec{c} = \vec{a} \times (\vec{b} \times \vec{c}) - \vec{b} \times (\vec{a} \times \vec{c}) \quad (6.31)$$

In coordinate form this is written

$$(\mathbf{a}^\times \mathbf{b})^\times \mathbf{c} = \mathbf{a}^\times \mathbf{b}^\times \mathbf{c} - \mathbf{b}^\times \mathbf{a}^\times \mathbf{c} \quad (6.32)$$

which implies that

$$(\mathbf{a}^\times \mathbf{b})^\times = \mathbf{a}^\times \mathbf{b}^\times - \mathbf{b}^\times \mathbf{a}^\times \quad (6.33)$$

**Example 82** The following problem is investigated: To what extent can the vector  $\vec{v}$  be determined when

$$\vec{w} = \vec{u} \times \vec{v} \quad (6.34)$$

and  $\vec{w}$  and  $\vec{u}$  are given? In coordinate form this is written

$$\mathbf{w} = \mathbf{u}^\times \mathbf{v} \quad (6.35)$$

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

$$\vec{w} = \frac{\vec{w}}{|\vec{w}|} \sin \theta |\vec{u}| |\vec{v}| \Rightarrow |\vec{v}| = \frac{|\vec{w}|}{|\vec{u}| \sin \theta} \quad (6.36)$$

This shows that if the angle  $\theta$  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

$$\vec{h} = \sum_{i=1}^3 h_i \vec{a}_i, \quad \mathbf{h} = \begin{pmatrix} h_1 \\ h_2 \\ h_3 \end{pmatrix} \quad (6.37)$$

Likewise the angular velocity can be represented by a vector  $\vec{\omega}$ , by a column vector  $\boldsymbol{\omega}$ , or by the generic component  $\omega_i$ , where

$$\vec{\omega} = \sum_{i=1}^3 \omega_i \vec{a}_i, \quad \boldsymbol{\omega} = \begin{pmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{pmatrix} \quad (6.38)$$

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

$$\mathbf{h} = \mathbf{M}\boldsymbol{\omega}, \quad h_i = \sum_{j=1}^3 m_{ij} \omega_j \quad (6.39)$$

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  $\omega_j$ . 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

$$\vec{M} := \sum_{i=1}^3 \sum_{j=1}^3 m_{ij} \vec{a}_i \vec{a}_j \quad (6.40)$$

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:

$$\begin{aligned} \vec{M} \cdot \vec{\omega} &= \sum_{i=1}^3 \sum_{j=1}^3 m_{ij} \vec{a}_i \vec{a}_j \cdot \sum_{k=1}^3 \omega_k \vec{a}_k \\ &= \sum_{i=1}^3 \sum_{j=1}^3 \sum_{k=1}^3 m_{ij} \vec{a}_i (\vec{a}_j \cdot \omega_k \vec{a}_k) \\ &= \sum_{i=1}^3 \sum_{j=1}^3 m_{ij} \omega_j \vec{a}_i \end{aligned} \quad (6.41)$$

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

$$\vec{h} = \vec{M} \cdot \vec{\omega} \quad (6.42)$$

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

$$\vec{h} = \vec{M} \cdot \vec{\omega} \Leftrightarrow \mathbf{h} = \mathbf{M}\boldsymbol{\omega} \Leftrightarrow h_i = \sum_{j=1}^3 m_{ij}\omega_j \quad (6.43)$$

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

$$\vec{D} = \sum_{i=1}^3 \sum_{j=1}^3 d_{ij} \vec{a}_i \vec{a}_j \quad (6.44)$$

where

$$d_{ij} = \vec{a}_i \cdot \vec{D} \cdot \vec{a}_j \quad (6.45)$$

are the components of the dyadic  $\vec{D}$  in frame  $a$ . The matrix

$$\mathbf{D} = \{d_{ij}\} \quad (6.46)$$

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

$$\begin{aligned} \vec{w} &= \vec{u} \cdot \vec{D} = \sum_{k=1}^3 u_k \vec{a}_k \cdot \sum_{i=1}^3 \sum_{j=1}^3 d_{ij} \vec{a}_i \vec{a}_j \\ &= \sum_{i=1}^3 \sum_{j=1}^3 d_{ij} u_i \vec{a}_j \end{aligned} \quad (6.47)$$

Scalar postmultiplication with a vector, which is the scalar product of a dyadic with a vector gives the vector

$$\vec{z} = \vec{D} \cdot \vec{u} = \sum_{i=1}^3 \sum_{j=1}^3 d_{ij} \vec{a}_i \vec{a}_j \cdot \sum_{k=1}^3 u_k \vec{a}_k \quad (6.48)$$

$$= \sum_{i=1}^3 \sum_{j=1}^3 d_{ij} u_j \vec{a}_i \quad (6.49)$$

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

$$\vec{w} = \vec{u} \cdot \vec{D} \Leftrightarrow \mathbf{w}^T = \mathbf{u}^T \mathbf{D} \quad (6.50)$$

$$\vec{z} = \vec{D} \cdot \vec{u} \Leftrightarrow \mathbf{z} = \mathbf{D}\mathbf{u} \quad (6.51)$$

The *identity dyadic* is defined by

$$\vec{I} := \sum_{i=1}^3 \sum_{j=1}^3 \delta_{ij} \vec{a}_i \vec{a}_j = \vec{a}_1 \vec{a}_1 + \vec{a}_2 \vec{a}_2 + \vec{a}_3 \vec{a}_3 \quad (6.52)$$

where  $\delta_{ij}$  is equal to unity when  $i = j$ , and zero otherwise. This implies that for any vector  $\vec{u}$

$$\vec{I} \cdot \vec{u} = \vec{u} \cdot \vec{I} = \vec{u} \quad (6.53)$$

and for any dyadic  $\vec{D}$  we have

$$\vec{I} \cdot \vec{D} = \vec{D} \cdot \vec{I} = \vec{D} \quad (6.54)$$

The equivalent matrix form of these equations are

$$\mathbf{I}\mathbf{u} = (\mathbf{u}^T \mathbf{I})^T = \mathbf{u} \quad (6.55)$$

$$\mathbf{I}\mathbf{D} = \mathbf{D}\mathbf{I} = \mathbf{D} \quad (6.56)$$

**Example 83** Let  $\vec{\omega}$  be a vector and let  $\vec{M}$  be a dyadic. Define the scalar

$$K = \frac{1}{2} \vec{\omega} \cdot \vec{M} \cdot \vec{\omega} \quad (6.57)$$

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

$$\vec{M} = \sum_{i=1}^3 \sum_{j=1}^3 m_{ij} \vec{a}_i \vec{a}_j. \quad (6.58)$$

Then the quadratic form is found to be

$$K = \frac{1}{2} \sum_{i=1}^3 \sum_{j=1}^3 \omega_i \omega_j m_{ij} \quad (6.59)$$

The corresponding representation in matrix form is given by

$$K = \frac{1}{2} \boldsymbol{\omega}^T \mathbf{M} \boldsymbol{\omega} \quad (6.60)$$

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

$$\vec{w} \cdot \vec{K} = (\vec{w} \cdot \vec{k}) \vec{k} \quad \text{and} \quad \vec{K} \cdot \vec{w} = \vec{k} (\vec{k} \cdot \vec{w}) \quad (6.61)$$

The coordinate form is

$$\mathbf{w}^T \mathbf{K} = \mathbf{w}^T \mathbf{k} \mathbf{k}^T \quad \text{and} \quad \mathbf{K} \mathbf{w} = \mathbf{k} \mathbf{k}^T \mathbf{w} \quad (6.62)$$

It follows that the matrix form of  $\vec{K} = \vec{k}\vec{k}$  is

$$\mathbf{K} = \mathbf{k} \mathbf{k}^T \quad (6.63)$$

**Example 85** The dyadic form of the vector cross product is

$$\vec{u} \times \vec{v} = \vec{u}^\times \cdot \vec{v} = \vec{u} \cdot \vec{v}^\times \quad (6.64)$$

where

$$\vec{u}^\times = \sum_{i=1}^3 \sum_{j=1}^3 \sum_{k=1}^3 \varepsilon_{ijk} u_j \vec{a}_i \vec{a}_k \quad (6.65)$$

is the dyadic form of the skew symmetric form  $\mathbf{u}^\times$ , and

$$\vec{v}^\times = \sum_{i=1}^3 \sum_{j=1}^3 \sum_{k=1}^3 \varepsilon_{ijk} v_j \vec{a}_i \vec{a}_k \quad (6.66)$$

We may then check that

$$\vec{u}^\times \cdot \vec{v} = \sum_{i=1}^3 \sum_{j=1}^3 \sum_{k=1}^3 \varepsilon_{ijk} u_j \vec{a}_i \vec{a}_k \cdot \sum_{p=1}^3 v_p \vec{a}_p = \sum_{i=1}^3 \sum_{j=1}^3 \sum_{k=1}^3 \varepsilon_{ijk} \vec{a}_i u_j v_k = \vec{u} \times \vec{v} \quad (6.67)$$

$$\begin{aligned} \vec{u} \cdot \vec{v}^\times &= \sum_{p=1}^3 u_p \vec{a}_p \cdot \sum_{i=1}^3 \sum_{j=1}^3 \sum_{k=1}^3 \varepsilon_{ijk} v_j \vec{a}_i \vec{a}_k = \sum_{i=1}^3 \sum_{j=1}^3 \sum_{k=1}^3 \varepsilon_{ijk} \vec{a}_k u_i v_j \end{aligned} \quad (6.68)$$

$$\begin{aligned} &= \sum_{i=1}^3 \sum_{j=1}^3 \sum_{k=1}^3 \varepsilon_{kij} \vec{a}_k u_i v_j = \sum_{i=1}^3 \sum_{j=1}^3 \sum_{k=1}^3 \varepsilon_{ijk} \vec{a}_i u_j v_k = \vec{u} \times \vec{v} \end{aligned} \quad (6.69)$$

The skew symmetric form used in the coordinate vector form is consistent with (6.64) as

$$\mathbf{u}^\times \mathbf{v} = -\mathbf{v}^\times \mathbf{u} = (\mathbf{v}^\times)^T \mathbf{u} = (\mathbf{u}^T \mathbf{v}^\times)^T \quad (6.70)$$

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

$$\vec{a}^\times \cdot \vec{b}^\times \cdot \vec{c} = [\vec{b}\vec{a} - (\vec{a} \cdot \vec{b})\vec{I}] \cdot \vec{c} \quad (6.71)$$

and it follows that

$$\vec{a}^\times \cdot \vec{b}^\times = \vec{b}\vec{a} - \vec{a} \cdot \vec{b}\vec{I} \quad (6.72)$$

In particular, we note that

$$\vec{a}^\times \cdot \vec{a}^\times = \vec{a}\vec{a} - \vec{a} \cdot \vec{a}\vec{I} \quad (6.73)$$

**Example 87** The triple scalar product satisfies

$$(\vec{d} \times \vec{a}) \cdot \vec{w} = \vec{d} \cdot (\vec{a} \times \vec{w})$$

for any vectors  $\vec{d}$ ,  $\vec{a}$  and  $\vec{w}$ , and it follows from (6.71) that

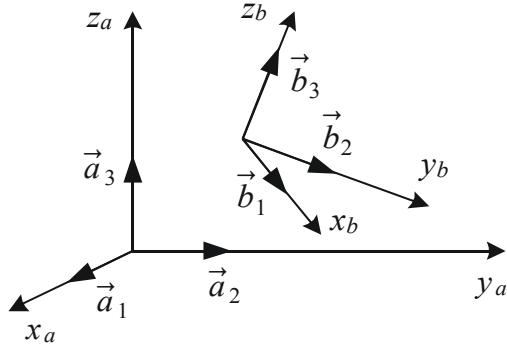
$$(\vec{d} \times \vec{a}) \cdot (\vec{b} \times \vec{c}) = \vec{d} \cdot (\vec{a} \times (\vec{b} \times \vec{c})) = \vec{d} \cdot (\vec{a} \cdot \vec{c}\vec{I} - \vec{c}\vec{a}) \cdot \vec{b} \quad (6.74)$$

The same result could have been obtained using

$$(\vec{d} \times \vec{a}) \cdot (\vec{b} \times \vec{c}) = -(\vec{d} \times \vec{a}) \cdot (\vec{c} \times \vec{b}) = -\vec{d} \cdot \vec{a}^\times \cdot \vec{c}^\times \cdot \vec{b} = -\vec{d} \cdot (\vec{a}^\times \cdot \vec{c}^\times) \cdot \vec{b} \quad (6.75)$$

In the development of the kinetic energy for a rigid body the following special case will be used:

$$(\vec{\omega} \times \vec{r}) \cdot (\vec{\omega} \times \vec{r}) = \vec{\omega} \cdot (\vec{r} \cdot \vec{r}\vec{I} - \vec{r}\vec{r}) \cdot \vec{\omega} = -\vec{\omega} \cdot (\vec{r}^\times \cdot \vec{r}^\times) \cdot \vec{\omega} \quad (6.76)$$

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

$$\vec{v} = \sum_{i=1}^3 v_i^a \vec{a}_i \quad \text{and} \quad \vec{v} = \sum_{i=1}^3 v_i^b \vec{b}_i \quad (6.77)$$

where

$$v_i^a = \vec{v} \cdot \vec{a}_i \quad (6.78)$$

are the coordinates of  $\vec{v}$  in  $a$ , and

$$v_i^b = \vec{v} \cdot \vec{b}_i \quad (6.79)$$

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

$$\mathbf{v}^a = \begin{pmatrix} v_1^a \\ v_2^a \\ v_3^a \end{pmatrix} \quad \text{and} \quad \mathbf{v}^b = \begin{pmatrix} v_1^b \\ v_2^b \\ v_3^b \end{pmatrix} \quad (6.80)$$

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:

$$\begin{aligned} v_i^a &= \vec{v} \cdot \vec{a}_i = (v_1^b \vec{b}_1 + v_2^b \vec{b}_2 + v_3^b \vec{b}_3) \cdot \vec{a}_i \\ &= \sum_{j=1}^3 v_j^b (\vec{a}_i \cdot \vec{b}_j) \end{aligned} \quad (6.81)$$

This leads to the following result:

The coordinate transformation from frame  $b$  to frame  $a$  is given by

$$\mathbf{v}^a = \mathbf{R}_b^a \mathbf{v}^b \quad (6.82)$$

where

$$\mathbf{R}_b^a = \{\vec{a}_i \cdot \vec{b}_j\} \quad (6.83)$$

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

$$\mathbf{R}_a^b = \{\vec{b}_i \cdot \vec{a}_j\} \quad (6.84)$$

For all  $\mathbf{v}^b$  we have

$$\mathbf{v}^b = \mathbf{R}_a^b \mathbf{v}^a = \mathbf{R}_a^b \mathbf{R}_b^a \mathbf{v}^b \quad (6.85)$$

This implies that

$$\mathbf{R}_a^b \mathbf{R}_b^a = \mathbf{I}, \quad (6.86)$$

and it follows that

$$\mathbf{R}_a^b = (\mathbf{R}_b^a)^{-1} \quad (6.87)$$

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

$$\mathbf{R}_a^b = (\mathbf{R}_b^a)^{-1} = (\mathbf{R}_b^a)^T \quad (6.88)$$

Consider a vector  $\vec{p}$  with coordinate vector  $\mathbf{p}^a$  in frame  $a$ . Define the vector  $\vec{q}$  defined by its coordinate vector

$$\mathbf{q}^a = \mathbf{R}_b^a \mathbf{p}^a \quad (6.89)$$

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

$$\mathbf{q}^b = \mathbf{R}_a^b \mathbf{q}^a = \mathbf{R}_a^b \mathbf{R}_b^a \mathbf{p}^a = \mathbf{p}^a \quad (6.90)$$

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:

- 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

$$\mathbf{v}^a = \mathbf{R}_b^a \mathbf{v}^b \quad (6.91)$$

In this equation  $\mathbf{R}_b^a$  acts as a coordinate transformation matrix.

- 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

$$\mathbf{q}^a = \mathbf{R}_b^a \mathbf{p}^a \quad (6.92)$$

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

$$\mathbf{a}_1^a = \mathbf{b}_1^b = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{a}_2^a = \mathbf{b}_2^b = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{a}_3^a = \mathbf{b}_3^b = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad (6.93)$$

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

$$\mathbf{R}_b^a = \begin{pmatrix} \mathbf{b}_1^a & \mathbf{b}_2^a & \mathbf{b}_3^a \end{pmatrix} \quad (6.94)$$

which is the third result.

The determinant of the rotation matrix  $\mathbf{R}_b^a$  is found by direct calculation to be

$$\begin{aligned} \det \mathbf{R}_b^a &= r_{11}(r_{22}r_{33} - r_{32}r_{23}) + r_{21}(r_{32}r_{13} - r_{12}r_{33}) + r_{31}(r_{12}r_{23} - r_{22}r_{13}) \\ &= (\mathbf{b}_1^a)^T [(\mathbf{b}_2^a)^\times \mathbf{b}_3^a] = (\mathbf{b}_1^a)^T \mathbf{b}_1^a = 1 \end{aligned}$$

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

$$\det \mathbf{R}_b^a = 1 \quad (6.95)$$

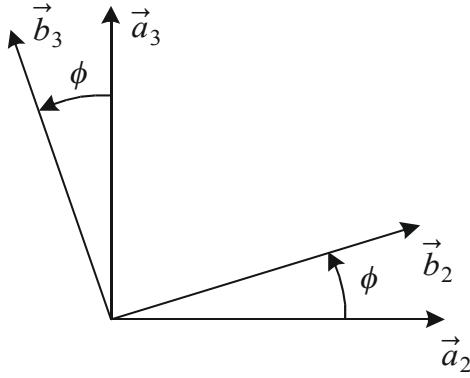
Finally, the set  $SO(3)$  is defined. We have established that the rotation matrix is orthogonal and has a determinant equal to unity. The set of all matrices that are orthogonal and with a determinant equal to unity is denoted by  $SO(3)$ , that is,

$$SO(3) = \{\mathbf{R} | \mathbf{R} \in R^{3 \times 3}, \quad \mathbf{R}^T \mathbf{R} = \mathbf{I} \quad \text{and} \quad \det \mathbf{R} = 1\} \quad (6.96)$$

Here  $R^{3 \times 3}$  is the set of all  $3 \times 3$  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 as 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  $\phi$  around  $\vec{a}_1$ .

to  $b$  and to  $a$  is given by

$$\begin{aligned}\mathbf{v}^b &= \mathbf{R}_c^b \mathbf{v}^c \\ \mathbf{v}^a &= \mathbf{R}_c^a \mathbf{v}^c\end{aligned}$$

Combining these two equations we get

$$\mathbf{v}^a = \mathbf{R}_b^a \mathbf{v}^b = \mathbf{R}_b^a \mathbf{R}_c^b \mathbf{v}^c$$

This shows that:

The rotation matrix of a composite rotation is the product of the rotation matrices:

$$\mathbf{R}_c^a = \mathbf{R}_b^a \mathbf{R}_c^b$$

This shows that the rotation matrix for the composite rotation  $\mathbf{R}_c^a$  is simply the product of the rotation matrices  $\mathbf{R}_b^a$  from  $a$  to  $b$  and  $\mathbf{R}_c^b$  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

$$\mathbf{R}_d^a = \mathbf{R}_b^a \mathbf{R}_c^b \mathbf{R}_d^c \quad (6.97)$$

#### 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  $\phi$  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  $\theta$  about the  $y$  axis, and  $\mathbf{R}_z(\psi)$  to be the rotation by an angle  $\psi$  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

$$\vec{a}_1 \cdot \vec{b}_2 = \vec{a}_1 \cdot \vec{b}_3 = \vec{a}_2 \cdot \vec{b}_1 = \vec{a}_3 \cdot \vec{b}_1 = 0 \quad (6.98)$$

$$\vec{a}_2 \cdot \vec{b}_2 = \cos \phi, \quad \vec{a}_3 \cdot \vec{b}_3 = \cos \phi \quad (6.99)$$

$$\vec{a}_3 \cdot \vec{b}_2 = \sin \phi, \quad \vec{a}_2 \cdot \vec{b}_3 = -\sin \phi \quad (6.100)$$

In the same way we can find the elements of the matrices  $\mathbf{R}_y(\theta)$  and  $\mathbf{R}_z(\psi)$ . This results in

$$\mathbf{R}_x(\phi) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos \phi & -\sin \phi \\ 0 & \sin \phi & \cos \phi \end{pmatrix} \quad (6.101)$$

$$\mathbf{R}_y(\theta) = \begin{pmatrix} \cos \theta & 0 & \sin \theta \\ 0 & 1 & 0 \\ -\sin \theta & 0 & \cos \theta \end{pmatrix} \quad (6.102)$$

$$\mathbf{R}_z(\psi) = \begin{pmatrix} \cos \psi & -\sin \psi & 0 \\ \sin \psi & \cos \psi & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (6.103)$$

#### 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

$$\vec{D} = \sum_{i=1}^3 \sum_{j=1}^3 d_{ij}^a \vec{a}_i \vec{a}_j = \sum_{p=1}^3 \sum_{q=1}^3 d_{pq}^b \vec{b}_p \vec{b}_q \quad (6.104)$$

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

$$\mathbf{D}^a = \{d_{ij}^a\}, \quad \mathbf{D}^b = \{d_{ij}^b\} \quad (6.105)$$

Let the vector  $\vec{z}$  be given by

$$\vec{z} = \vec{D} \cdot \vec{u} \quad (6.106)$$

Then, in frames  $a$  and  $b$  this may be written in matrix form as

$$\mathbf{z}^a = \mathbf{D}^a \mathbf{u}^a, \quad \mathbf{z}^b = \mathbf{D}^b \mathbf{u}^b \quad (6.107)$$

We then find that

$$\mathbf{D}^a \mathbf{u}^a = \mathbf{z}^a = \mathbf{R}_b^a \mathbf{z}^b = \mathbf{R}_b^a \mathbf{D}^b \mathbf{u}^b = \mathbf{R}_b^a \mathbf{D}^b \mathbf{R}_a^b \mathbf{u}^a \quad (6.108)$$

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

$$\mathbf{D}^a = \mathbf{R}_b^a \mathbf{D}^b \mathbf{R}_a^b \quad (6.109)$$

**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

$$\vec{M} = \sum_{i=1}^3 \sum_{j=1}^3 m_{ij}^b \vec{b}_i \vec{b}_j \quad (6.110)$$

and the corresponding matrix representation is

$$\mathbf{M}^b = \{m_{ij}^b\} \quad (6.111)$$

An important result in rigid body dynamics is that when frame  $b$  is fixed in the rigid body, and therefore moves with the rigid body, 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

$$\mathbf{M}^a = \mathbf{R}_b^a \mathbf{M}^b \mathbf{R}_a^b \quad (6.112)$$

**Example 89** The relation between the skew symmetric forms of a vector is given by

$$(\mathbf{u}^b)^\times \mathbf{v}^b = \mathbf{w}^b = \mathbf{R}_a^b \mathbf{w}^a = \mathbf{R}_a^b (\mathbf{u}^a)^\times \mathbf{v}^a = \mathbf{R}_a^b (\mathbf{u}^a)^\times \mathbf{R}_b^a \mathbf{v}^b \quad (6.113)$$

which implies that

$$(\mathbf{u}^b)^\times = \mathbf{R}_a^b (\mathbf{u}^a)^\times \mathbf{R}_b^a \quad (6.114)$$

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  $(\mathbf{u}^a)^\times$  is the matrix representation of the dyadic  $\bar{u}^\times$ .

#### 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

$$\mathbf{T}_b^a = \begin{pmatrix} \mathbf{R}_b^a & \mathbf{r}_{ab}^a \\ \mathbf{0}^T & 1 \end{pmatrix} \in SE(3) \quad (6.115)$$

Here the set  $SE(3)$  is the Special Euclidean Group of dimension 3 defined by

$$SE(3) = \left\{ \mathbf{T} \mid \mathbf{T} = \begin{pmatrix} \mathbf{R} & \mathbf{r} \\ \mathbf{0}^T & 1 \end{pmatrix}, \mathbf{R} \in SO(3), \mathbf{r} \in R^3 \right\} \quad (6.116)$$

The inverse of  $\mathbf{T}_b^a$  is found by matrix inversion to be

$$(\mathbf{T}_b^a)^{-1} = \begin{pmatrix} (\mathbf{R}_b^a)^T & -(\mathbf{R}_b^a)^T \mathbf{r}_{ab}^a \\ \mathbf{0}^T & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R}_a^b & \mathbf{r}_{ba}^b \\ \mathbf{0}^T & 1 \end{pmatrix} = \mathbf{T}_a^b \quad (6.117)$$

This means that

$$(\mathbf{T}_b^a)^{-1} = \mathbf{T}_a^b \quad (6.118)$$

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) through 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

$$\mathbf{T}_6^0 = \mathbf{T}_1^0 \mathbf{T}_2^1 \dots \mathbf{T}_6^5. \tag{6.123}$$

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  $\theta_i$  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  $3 \times 3$  matrix with nine elements. The orthogonality of the

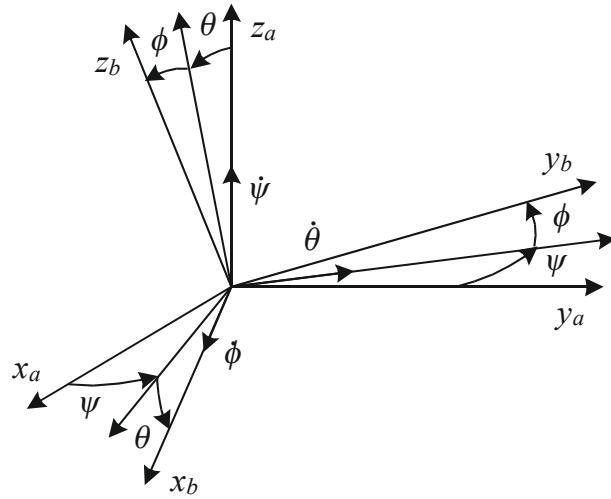


Figure 6.5: Roll-pitch-yaw Euler angles.

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  $\psi$  about the  $z_a$  axis, then a rotation  $\theta$  about the current (rotated)  $y$  axis, and finally a rotation  $\phi$  about the current (rotated)  $x$  axis as shown in Figure 6.5. The resulting rotation matrix is

$$\mathbf{R}_b^a = \mathbf{R}_z(\psi)\mathbf{R}_y(\theta)\mathbf{R}_x(\phi) \quad (6.125)$$

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

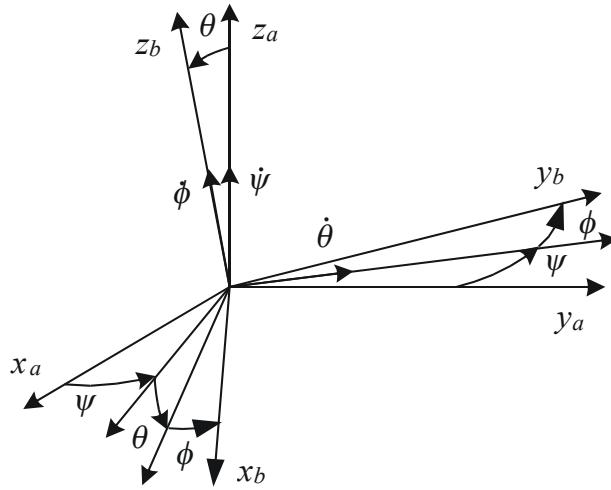


Figure 6.6: Classical Euler angles.

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  $\psi$  about the  $z_a$  axis, then a rotation  $\theta$  about the current (rotated)  $y$  axis, and finally a rotation  $\phi$  about the current (rotated)  $z$  axis as shown in Figure 6.6. The resulting rotation matrix is

$$\mathbf{R}_b^a = \mathbf{R}_z(\psi) \mathbf{R}_y(\theta) \mathbf{R}_x(\phi) \quad (6.126)$$

## 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

$$\mathbf{R}_b^a \mathbf{k} = \mathbf{k} \quad (6.127)$$

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

$$\mathbf{k}^a = \mathbf{k} \quad (6.128)$$

in frame  $a$ . The transformation rule

$$\mathbf{k}^a = \mathbf{R}_b^a \mathbf{k}^b \quad (6.129)$$

then implies that

$$\mathbf{k}^a = \mathbf{k}^b = \mathbf{k} \quad (6.130)$$

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  $\theta$  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  $\theta$  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  $\theta$  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  $\theta$  and  $\vec{k}$ . We will do this by deriving an expression where  $\vec{q}$  is given by  $\vec{p}$ ,  $\theta$  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  $\theta$  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  $\vec{p}$ ,  $\theta$  and  $\vec{k}$ .

The vectors  $\vec{p}$  and  $\vec{q}$  can be written

$$\vec{p} = x\vec{c}_1 + y\vec{c}_2 + z\vec{k} \quad (6.131)$$

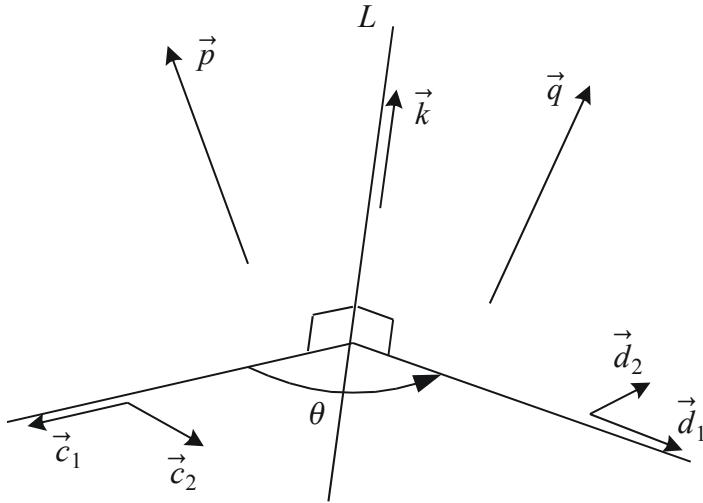


Figure 6.7: Rotation of the vector  $\vec{p}$  by an angle  $\theta$  around the  $\vec{k}$  vector.

and

$$\vec{q} = x\vec{d}_1 + y\vec{d}_2 + z\vec{k} \quad (6.132)$$

The unit vectors  $\vec{d}_1$  and  $\vec{d}_2$  can be written

$$\vec{d}_1 = \cos \theta \vec{c}_1 + \sin \theta \vec{c}_2 \quad (6.133)$$

$$\vec{d}_2 = -\sin \theta \vec{c}_1 + \cos \theta \vec{c}_2 \quad (6.134)$$

and insertion into (6.132) gives

$$\vec{q} = (x \cos \theta - y \sin \theta) \vec{c}_1 + (x \sin \theta + y \cos \theta) \vec{c}_2 + z \vec{k} \quad (6.135)$$

Consider the calculation

$$\begin{aligned} & \cos \theta \vec{p} + \sin \theta \vec{k} \times \vec{p} + (1 - \cos \theta) \vec{k} \vec{k} \cdot \vec{p} \\ &= \cos \theta (x \vec{c}_1 + y \vec{c}_2 + z \vec{k}) - \sin \theta (-x \vec{c}_1 + y \vec{c}_2) + (1 - \cos \theta) z \vec{k} \\ &= (x \cos \theta - y \sin \theta) \vec{c}_1 + (x \sin \theta + y \cos \theta) \vec{c}_2 + z \vec{k} \end{aligned} \quad (6.136)$$

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

$$\begin{aligned} \vec{q} &= \cos \theta \vec{p} + \sin \theta \vec{k} \times \vec{p} + (1 - \cos \theta) \vec{k} \vec{k} \cdot \vec{p} \\ &= (\cos \theta \vec{I} + \sin \theta \vec{k}^\times + (1 - \cos \theta) \vec{k} \vec{k}) \cdot \vec{p} \end{aligned} \quad (6.137)$$

#### 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

$$\vec{q} = \vec{R}_{k,\theta} \cdot \vec{p} \quad (6.138)$$

where

$$\vec{R}_{k,\theta} = \cos \theta \vec{I} + \sin \theta \vec{k}^\times + (1 - \cos \theta) \vec{k} \vec{k} \quad (6.139)$$

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  $\theta$  about a unit vector  $\vec{k}$  where  $\mathbf{R}_b^a$  is given by

$$\mathbf{R}_b^a = \cos \theta \mathbf{I} + \sin \theta (\mathbf{k}^a)^\times + (1 - \cos \theta) \mathbf{k}^a (\mathbf{k}^a)^T \quad (6.140)$$

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

$$\mathbf{k}^a = \mathbf{R}_b^a \mathbf{k}^b = \mathbf{k}^b \quad (6.141)$$

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

$$\mathbf{R}_b^a = \begin{pmatrix} k_x^2 v_\theta + c_\theta & k_x k_y v_\theta - k_z s_\theta & k_x k_z v_\theta + k_y s_\theta \\ k_x k_y v_\theta + k_z s_\theta & k_y^2 v_\theta + c_\theta & k_y k_z v_\theta - k_x s_\theta \\ k_x k_z v_\theta - k_y s_\theta & k_y k_z v_\theta + k_x s_\theta & k_z^2 v_\theta + c_\theta \end{pmatrix} \quad (6.142)$$

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

$$\mathbf{R}_b^a = \begin{pmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{pmatrix} = \mathbf{R}_{z,\theta} \quad (6.143)$$

which is to be expected as this is a rotation by an angle  $\theta$  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

$$\mathbf{R}_b^a = \mathbf{R}_c^a \mathbf{R}_{z,\theta} \mathbf{R}_c^c \quad (6.144)$$

### 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

$$\mathbf{R}_{k,\theta} := \cos \theta \mathbf{I} + \mathbf{k}^\times \sin \theta + \mathbf{k} \mathbf{k}^T (1 - \cos \theta) \quad (6.145)$$

An alternative expression is found by inserting the identity

$$\mathbf{k}^\times \mathbf{k}^\times = \mathbf{k} \mathbf{k}^T - \mathbf{k}^T \mathbf{k} \mathbf{I} = \mathbf{k} \mathbf{k}^T - \mathbf{I}. \quad (6.146)$$

which gives

$$\mathbf{R}_{k,\theta} = \mathbf{I} + \mathbf{k}^\times \sin \theta + \mathbf{k}^\times \mathbf{k}^\times (1 - \cos \theta) \quad (6.147)$$

The inverse to the rotation matrix is

$$(\mathbf{R}_b^a)^T = \mathbf{R}_a^b = \mathbf{R}_{k,-\theta}. \quad (6.148)$$

**Example 94** The derivative of  $\mathbf{R}_{k,\theta}$  with respect to  $\theta$  is found from (6.147) to be

$$\frac{d\mathbf{R}_{k,\theta}}{d\theta} = \mathbf{k}^\times \cos \theta + \mathbf{k}^\times \mathbf{k}^\times \sin \theta \quad (6.149)$$

Using (6.27) we find that

$$\mathbf{k}^\times \mathbf{R}_{k,\theta} = \mathbf{k}^\times + \mathbf{k}^\times \mathbf{k}^\times \sin \theta - \mathbf{k}^\times (1 - \cos \theta) \quad (6.150)$$

and we may conclude that

$$\frac{d}{d\theta} \mathbf{R}_{k,\theta} = \mathbf{k}^\times \mathbf{R}_{k,\theta} \quad (6.151)$$

**Example 95** It is known that the differential equation  $\frac{d}{dt}\mathbf{x} = \mathbf{Ax}$  has the solution  $\mathbf{x}(t) = \mathbf{x}(0)\exp(\mathbf{At})$  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

$$\mathbf{R}_{k,\theta} = \exp[\mathbf{k}^\times \theta] \quad (6.152)$$

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

$$\exp(\mathbf{A}) = \mathbf{I} + \mathbf{A} + \frac{1}{2!} \mathbf{A}^2 + \frac{1}{3!} \mathbf{A}^3 \dots \quad (6.153)$$

The result (6.152) can be derived directly from (6.147). First we use (6.27) to establish the identity

$$(\mathbf{k}^\times)^{2n+1} = (-1)^n \mathbf{k}^\times \quad (6.154)$$

by induction. This is done by noting that (6.27) implies that (6.154) is true for  $n = 1$ , and moreover, for  $n = 1, 2, \dots$  we have

$$(\mathbf{k}^\times)^{2n+1} = (-1)^n \mathbf{k}^\times \Rightarrow (\mathbf{k}^\times)^{2(n+1)+1} = (-1)^n \mathbf{k}^\times (\mathbf{k}^\times)^2 = (-1)^n (-1) \mathbf{k}^\times. \quad (6.155)$$

Then we may evaluate  $\exp[\mathbf{k}^\times \theta]$  directly from the definition (6.153), and find that

$$\begin{aligned} \exp[\mathbf{k}^\times \theta] &= \mathbf{I} + \mathbf{k}^\times \theta + (\mathbf{k}^\times)^2 \frac{\theta^2}{2!} + (\mathbf{k}^\times)^3 \frac{\theta^3}{3!} + (\mathbf{k}^\times)^4 \frac{\theta^4}{4!} + (\mathbf{k}^\times)^5 \frac{\theta^5}{5!} + (\mathbf{k}^\times)^6 \frac{\theta^6}{6!} \dots \\ &= \mathbf{I} + \mathbf{k}^\times [\theta + \mathbf{k}^\times \frac{\theta^2}{2!}] + (\mathbf{k}^\times)^3 [\frac{\theta^3}{3!} + \mathbf{k}^\times \frac{\theta^4}{4!}] + (\mathbf{k}^\times)^5 [\frac{\theta^5}{5!} + \mathbf{k}^\times \frac{\theta^6}{6!}] \dots \\ &= \mathbf{I} + \mathbf{k}^\times [\theta + \mathbf{k}^\times \frac{\theta^2}{2!}] - \mathbf{k}^\times [\frac{\theta^3}{3!} + \mathbf{k}^\times \frac{\theta^4}{4!}] + \mathbf{k}^\times [\frac{\theta^5}{5!} + \mathbf{k}^\times \frac{\theta^6}{6!}] \dots \\ &= \mathbf{I} + \mathbf{k}^\times [\theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} \dots] - (\mathbf{k}^\times)^2 [\frac{\theta^2}{2!} - \frac{\theta^4}{4!} + \frac{\theta^6}{6!} \dots] \\ &= \mathbf{I} + \mathbf{k}^\times \sin \theta + (\mathbf{k}^\times)^2 (1 - \cos \theta) \end{aligned} \quad (6.156)$$

This shows in view of (6.147) that

$$\mathbf{R}_{k,\theta} = \exp[\mathbf{k}^\times \theta] \quad (6.157)$$

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 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  $\theta$  and  $\vec{k}$ , and are given by the scalar  $\eta$  and the vector  $\vec{\epsilon}$  defined by

$$\eta = \cos \frac{\theta}{2}, \quad \vec{\epsilon} = \vec{k} \sin \frac{\theta}{2} \quad (6.158)$$

In coordinate form this is written

$$\eta = \cos \frac{\theta}{2}, \quad \boldsymbol{\epsilon} = \mathbf{k} \sin \frac{\theta}{2} \quad (6.159)$$

We note that

$$\eta^2 + \vec{\epsilon} \cdot \vec{\epsilon} = \eta^2 + \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = \cos^2 \frac{\theta}{2} + \sin^2 \frac{\theta}{2} = 1 \quad (6.160)$$

Insertion of the trigonometric identities

$$\sin \theta = 2 \sin \frac{\theta}{2} \cos \frac{\theta}{2} \quad (6.161)$$

$$\cos \theta = \cos^2 \frac{\theta}{2} - \sin^2 \frac{\theta}{2} = 2 \cos^2 \frac{\theta}{2} - 1 = 1 - 2 \sin^2 \frac{\theta}{2} \quad (6.162)$$

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

$$\mathbf{R}_e(\eta, \boldsymbol{\epsilon}) = (\eta^2 - \boldsymbol{\epsilon}^T \boldsymbol{\epsilon}) \mathbf{I} + 2\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T + 2\eta\boldsymbol{\epsilon}^\times \quad (6.163)$$

$$= (2\eta^2 - 1) \mathbf{I} + 2\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T + 2\eta\boldsymbol{\epsilon}^\times \quad (6.164)$$

$$= (1 - 2\boldsymbol{\epsilon}^T \boldsymbol{\epsilon}) \mathbf{I} + 2\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T + 2\eta\boldsymbol{\epsilon}^\times \quad (6.165)$$

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 (6.164) an alternative form of the rotation matrix is found.

The rotation matrix is given by the corresponding Euler parameters according to

$$\mathbf{R}_e(\eta, \boldsymbol{\epsilon}) = \mathbf{I} + 2\eta\boldsymbol{\epsilon}^\times + 2\boldsymbol{\epsilon}^\times\boldsymbol{\epsilon}^\times \quad (6.166)$$

A given rotation will correspond to two sets of Euler parameters  $(\eta, \boldsymbol{\epsilon})$  and  $(-\eta, -\boldsymbol{\epsilon})$  with opposite signs as

$$\mathbf{R}_e(-\eta, -\boldsymbol{\epsilon}) = \mathbf{R}_e(\eta, \boldsymbol{\epsilon}) \quad (6.167)$$

The inverse of  $\mathbf{R}_e(\eta, \boldsymbol{\epsilon})$  given by

$$\mathbf{R}_e(\eta, \boldsymbol{\epsilon})^T = \mathbf{R}_e(\eta, -\boldsymbol{\epsilon}) \quad (6.168)$$

corresponds to the Euler parameters  $(\eta, -\boldsymbol{\epsilon})$ .

**Example 97** From (6.165) and (6.160) we find that

$$\text{Trace}\mathbf{R} = 3(\eta^2 - \boldsymbol{\epsilon}^T \boldsymbol{\epsilon}) + 2\boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = 4\eta^2 - 1 \quad (6.169)$$

### 6.7.2 Quaternions

The vector

$$\mathbf{p} = \begin{pmatrix} \eta \\ \boldsymbol{\epsilon} \end{pmatrix} \quad (6.170)$$

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

$$\mathbf{q} = \begin{pmatrix} \alpha \\ \boldsymbol{\beta} \end{pmatrix} \quad (6.171)$$

of dimension 4 where  $\alpha$  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  $\mathbf{q}_1 = (\alpha_1 \ \ \boldsymbol{\beta}_1^T)^T$  and  $\mathbf{q}_2 = (\alpha_2 \ \ \boldsymbol{\beta}_2^T)^T$  in  $R^4$  is defined by

$$\begin{pmatrix} \alpha_1 \\ \boldsymbol{\beta}_1 \end{pmatrix} \otimes \begin{pmatrix} \alpha_2 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \begin{pmatrix} \alpha_1\alpha_2 - \boldsymbol{\beta}_1^T \boldsymbol{\beta}_2 \\ \alpha_1\boldsymbol{\beta}_2 + \alpha_2\boldsymbol{\beta}_1 + \boldsymbol{\beta}_1^\times \boldsymbol{\beta}_2 \end{pmatrix} \quad (6.172)$$

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

$$\begin{pmatrix} \alpha_1 \\ \boldsymbol{\beta}_1 \end{pmatrix} \otimes \begin{pmatrix} \alpha_2 \\ \boldsymbol{\beta}_2 \end{pmatrix} - \begin{pmatrix} \alpha_2 \\ \boldsymbol{\beta}_2 \end{pmatrix} \otimes \begin{pmatrix} \alpha_1 \\ \boldsymbol{\beta}_1 \end{pmatrix} = 2 \begin{pmatrix} 0 \\ \boldsymbol{\beta}_1 \end{pmatrix} \otimes \begin{pmatrix} 0 \\ \boldsymbol{\beta}_2 \end{pmatrix} \quad (6.173)$$

where it is used that  $\boldsymbol{\beta}_1^\times \boldsymbol{\beta}_2 = -\boldsymbol{\beta}_2^\times \boldsymbol{\beta}_1$ .

**Example 99** Define the matrices

$$\mathbf{F}(\mathbf{q}) = \begin{pmatrix} \alpha & -\boldsymbol{\beta}^T \\ \boldsymbol{\beta} & \alpha\mathbf{I} + \boldsymbol{\beta}^\times \end{pmatrix} \in R^{4 \times 4} \quad (6.174)$$

$$\mathbf{E}(\mathbf{q}) = \begin{pmatrix} \alpha & -\boldsymbol{\beta}^T \\ \boldsymbol{\beta} & \alpha\mathbf{I} - \boldsymbol{\beta}^\times \end{pmatrix} \in R^{4 \times 4} \quad (6.175)$$

The matrix  $\mathbf{F}(\mathbf{q})$  represents quaternion pre-multiplication with  $\mathbf{q}$  in the sense that for any  $\mathbf{u} \in R^4$

$$\mathbf{q} \otimes \mathbf{u} = \mathbf{F}(\mathbf{q})\mathbf{u} \quad (6.176)$$

while  $\mathbf{E}(\mathbf{q})$  represents quaternion post-multiplication with  $\mathbf{q}$  in the sense that

$$\mathbf{u} \otimes \mathbf{q} = \mathbf{E}(\mathbf{q})\mathbf{u} \quad (6.177)$$

**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

$$\mathbf{q} = \alpha + i\beta_1 + j\beta_2 + k\beta_3 \quad (6.178)$$

where  $i$ ,  $j$  and  $k$  are imaginary units satisfying

$$i^2 = j^2 = k^2 = -1 \quad (6.179)$$

$$ij = -ji = k, \quad jk = -kj = i, \quad ki = -ik = j \quad (6.180)$$

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)

$$\mathbf{q}_1 = \alpha_1 + i\beta_{11} + j\beta_{12} + k\beta_{13} \quad (6.181)$$

$$\mathbf{q}_2 = \alpha_2 + i\beta_{21} + j\beta_{22} + k\beta_{23} \quad (6.182)$$

Then the product of  $\mathbf{q}_1$  and  $\mathbf{q}_2$  is found to be

$$\begin{aligned} \mathbf{q}_1 \mathbf{q}_2 &= (\alpha_1 + i\beta_{11} + j\beta_{12} + k\beta_{13})(\alpha_2 + i\beta_{21} + j\beta_{22} + k\beta_{23}) \\ &= \alpha_1 \alpha_2 - \beta_{11} \beta_{21} - \beta_{12} \beta_{22} - \beta_{13} \beta_{23} \\ &\quad + i(\alpha_1 \beta_{21} + \alpha_2 \beta_{11} + \beta_{12} \beta_{23} - \beta_{13} \beta_{22}) \\ &\quad + j(\alpha_1 \beta_{22} + \alpha_2 \beta_{12} + \beta_{13} \beta_{21} - \beta_{11} \beta_{23}) \\ &\quad + k(\alpha_1 \beta_{23} + \alpha_2 \beta_{13} + \beta_{11} \beta_{22} - \beta_{12} \beta_{21}) \end{aligned} \quad (6.183)$$

We see that this product of quaternion numbers satisfy

$$\mathbf{q} = \mathbf{q}_1 \mathbf{q}_2 \quad (6.184)$$

where

$$\mathbf{q} = \alpha + i\beta_1 + j\beta_2 + k\beta_3 \quad (6.185)$$

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

$$\mathbf{p} = \begin{pmatrix} \eta \\ \epsilon \end{pmatrix} \quad (6.186)$$

is a quaternion with unit length, that is, a quaternion that satisfies

$$\mathbf{p}^T \mathbf{p} = \eta^2 + \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = 1 \quad (6.187)$$

We see that if  $\eta$  and  $\boldsymbol{\epsilon}$  are Euler parameters, then  $\mathbf{p}$  is a unit quaternion corresponding to the rotation matrix  $\mathbf{R}_{\eta, \boldsymbol{\epsilon}}$ . The unit quaternion corresponding to  $\mathbf{R}_{\eta, \boldsymbol{\epsilon}}^{-1} = \mathbf{R}_{\eta, -\boldsymbol{\epsilon}}$  is the *inverse unit quaternion*  $\bar{\mathbf{p}}$  defined by

$$\bar{\mathbf{p}} = \begin{pmatrix} \eta \\ -\boldsymbol{\epsilon} \end{pmatrix} \quad (6.188)$$

The unit quaternion corresponding to the identity matrix  $\mathbf{R}_{1,0} = \mathbf{I}$  is the identity quaternion  $\mathbf{p}_{id}$  defined by

$$\mathbf{p}_{id} = \begin{pmatrix} 1 \\ \mathbf{0} \end{pmatrix} \quad (6.189)$$

#### 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

$$\mathbf{p} := \mathbf{p}_1 \otimes \mathbf{p}_2 = \begin{pmatrix} \eta_1 \eta_2 - \boldsymbol{\epsilon}_1^T \boldsymbol{\epsilon}_2 \\ \eta_1 \boldsymbol{\epsilon}_2 + \eta_2 \boldsymbol{\epsilon}_1 + \boldsymbol{\epsilon}_1^\times \boldsymbol{\epsilon}_2 \end{pmatrix} \quad (6.190)$$

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  $\eta_2^2 + \boldsymbol{\epsilon}_2^T \boldsymbol{\epsilon}_2 = 1$ .

It is straightforward to check that the quaternion product of  $\mathbf{p}$  and the inverse  $\bar{\mathbf{p}}$  is the identity quaternion, that is,

$$\mathbf{p} \otimes \bar{\mathbf{p}} = \bar{\mathbf{p}} \otimes \mathbf{p} = \mathbf{p}_{id} \quad (6.191)$$

This follows from

$$\mathbf{p} \otimes \bar{\mathbf{p}} = \begin{pmatrix} \eta \\ \boldsymbol{\epsilon} \end{pmatrix} \otimes \begin{pmatrix} \eta \\ -\boldsymbol{\epsilon} \end{pmatrix} = \begin{pmatrix} \eta^2 + \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} \\ \eta \boldsymbol{\epsilon} - \eta \boldsymbol{\epsilon} - \boldsymbol{\epsilon}^\times \boldsymbol{\epsilon} \end{pmatrix} = \begin{pmatrix} 1 \\ \mathbf{0} \end{pmatrix} = \mathbf{p}_{id} \quad (6.192)$$

In the same way it can be shown that  $\bar{\mathbf{p}} \otimes \mathbf{p} = \mathbf{p}_{id}$ .

We may also verify that

$$\mathbf{p} \otimes \mathbf{p}_{id} = \mathbf{p}_{id} \otimes \mathbf{p} = \mathbf{p} \quad (6.193)$$

**Example 101** The unit quaternion satisfies

$$\dot{\mathbf{p}} \otimes \bar{\mathbf{p}} = \begin{pmatrix} \dot{\eta} \\ \dot{\boldsymbol{\epsilon}} \end{pmatrix} \otimes \begin{pmatrix} \eta \\ -\boldsymbol{\epsilon} \end{pmatrix} = \begin{pmatrix} 0 \\ -\dot{\eta} \boldsymbol{\epsilon} + \eta \dot{\boldsymbol{\epsilon}} + \boldsymbol{\epsilon}^\times \dot{\boldsymbol{\epsilon}} \end{pmatrix} \quad (6.194)$$

where we have used

$$\dot{\eta}\eta + \dot{\boldsymbol{\epsilon}}^T \boldsymbol{\epsilon} = \frac{1}{2} \frac{d}{dt} (\eta^2 + \boldsymbol{\epsilon}^T \boldsymbol{\epsilon}) = 0 \quad (6.195)$$

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

$$\mathbf{p} \otimes \bar{\mathbf{p}} = \mathbf{p}_{id} \Rightarrow \dot{\mathbf{p}} \otimes \bar{\mathbf{p}} + \mathbf{p} \otimes \dot{\bar{\mathbf{p}}} = \mathbf{0} \quad (6.196)$$

which implies

$$\dot{\bar{\mathbf{p}}} = -\bar{\mathbf{p}} \otimes \dot{\mathbf{p}} \otimes \bar{\mathbf{p}} \quad (6.197)$$

### 6.7.5 Rotation by the quaternion product

Let  $\mathbf{R} := \mathbf{R}_e(\eta, \epsilon)$  be the rotation matrix corresponding to the Euler parameters  $\eta$  and  $\epsilon$ . Let  $\mathbf{v} \in R^3$  be an arbitrary vector. We are already familiar with the notion that  $\mathbf{R}\mathbf{v}$  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  $\mathbf{R}\mathbf{v}$  can be achieved with the Euler parameters and the quaternion product according to

$$\begin{pmatrix} 0 \\ \mathbf{R}\mathbf{v} \end{pmatrix} = \begin{pmatrix} \eta \\ \epsilon \end{pmatrix} \otimes \begin{pmatrix} 0 \\ \mathbf{v} \end{pmatrix} \otimes \begin{pmatrix} \eta \\ -\epsilon \end{pmatrix} \quad (6.198)$$

This is shown by direct computation of the quaternion products:

$$\begin{aligned} \begin{pmatrix} \eta \\ \epsilon \end{pmatrix} \otimes \begin{pmatrix} 0 \\ \mathbf{v} \end{pmatrix} \otimes \begin{pmatrix} \eta \\ -\epsilon \end{pmatrix} &= \begin{pmatrix} \eta\epsilon^T\mathbf{v} - \eta\epsilon^T\mathbf{v} - \epsilon^T\epsilon^\times\mathbf{v} \\ \eta^2\mathbf{v} + 2\eta\epsilon^\times\mathbf{v} + \epsilon\epsilon^T\mathbf{v} + \epsilon^\times\epsilon^\times\mathbf{v} \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ (\mathbf{I} + 2\eta\epsilon^\times + 2\epsilon^\times\epsilon^\times)\mathbf{v} \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ \mathbf{R}\mathbf{v} \end{pmatrix} \end{aligned} \quad (6.199)$$

where we have used  $(\epsilon^\times)^2 = \epsilon\epsilon^T - \epsilon^T\epsilon\mathbf{I}$  and  $\eta^2 + \epsilon^T\epsilon = 1$ .

We will now see how composite rotations can be expressed in terms of unit quaternions. Let

$$\mathbf{R}_1 = \mathbf{R}_e(\eta_1, \epsilon_1) \quad \text{and} \quad \mathbf{R}_2 = \mathbf{R}_e(\eta_2, \epsilon_2) \quad (6.200)$$

and let

$$\mathbf{R} = \mathbf{R}_1\mathbf{R}_2 = \mathbf{R}_e(\eta, \epsilon). \quad (6.201)$$

be the composite rotation where  $\mathbf{p} = (\eta \ \epsilon^T)^T$ ,  $\mathbf{p}_1 = (\eta_1 \ \epsilon_1^T)^T$  and  $\mathbf{p}_2 = (\eta_2 \ \epsilon_2^T)^T$ .

Let  $\mathbf{u}$  be an arbitrary vector, and define  $\mathbf{v} := \mathbf{R}_2\mathbf{u}$  and  $\mathbf{w} := \mathbf{R}_1\mathbf{v} = \mathbf{R}\mathbf{u}$ . Then

$$\begin{pmatrix} 0 \\ \mathbf{v} \end{pmatrix} = \mathbf{p}_2 \otimes \begin{pmatrix} 0 \\ \mathbf{u} \end{pmatrix} \otimes \bar{\mathbf{p}}_2 \quad (6.202)$$

$$\begin{pmatrix} 0 \\ \mathbf{w} \end{pmatrix} = \mathbf{p}_1 \otimes \begin{pmatrix} 0 \\ \mathbf{v} \end{pmatrix} \otimes \bar{\mathbf{p}}_1 = \mathbf{p}_1 \otimes \mathbf{p}_2 \otimes \begin{pmatrix} 0 \\ \mathbf{u} \end{pmatrix} \otimes \bar{\mathbf{p}}_2 \otimes \bar{\mathbf{p}}_1 \quad (6.203)$$

At the same time we have  $\mathbf{w} = \mathbf{R}_1\mathbf{R}_2\mathbf{u} = \mathbf{R}\mathbf{u}$  which gives

$$\begin{pmatrix} 0 \\ \mathbf{w} \end{pmatrix} = \mathbf{p} \otimes \begin{pmatrix} 0 \\ \mathbf{u} \end{pmatrix} \otimes \bar{\mathbf{p}} \quad (6.204)$$

Comparing these results we find that the Euler parameters  $\eta, \epsilon$  corresponding to the composite rotation  $\mathbf{R}$  is given by

$$\mathbf{p} = \mathbf{p}_1 \otimes \mathbf{p}_2 \quad (6.205)$$

which can also be written

$$\eta = \eta_1\eta_2 - \epsilon_1^T\epsilon_2 \quad (6.206)$$

$$\epsilon = \eta_1\epsilon_2 + \eta_2\epsilon_1 + \epsilon_1^\times\epsilon_2 \quad (6.207)$$

**Example 103** We see that

$$\mathbf{R}\mathbf{R}^T = \mathbf{I} \quad (6.208)$$

is consistent with

$$\mathbf{p} \otimes \bar{\mathbf{p}} = \mathbf{p}_{id} \quad (6.209)$$

Moreover,

$$\mathbf{R}\mathbf{I} = \mathbf{I}\mathbf{R} = \mathbf{R} \quad (6.210)$$

is seen to be consistent with

$$\mathbf{p} \otimes \mathbf{p}_{id} = \mathbf{p}_{id} \otimes \mathbf{p} = \mathbf{p}_{id} \quad (6.211)$$

**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

$$\begin{pmatrix} 0 \\ \mathbf{R}\mathbf{v} \end{pmatrix} = \mathbf{p} \otimes \begin{pmatrix} 0 \\ \mathbf{v} \end{pmatrix} \otimes \bar{\mathbf{p}} \quad (6.212)$$

can be written

$$\begin{pmatrix} 0 \\ \mathbf{R}\mathbf{v} \end{pmatrix} = \mathbf{F}(\mathbf{p}) \begin{pmatrix} 0 \\ \mathbf{v} \end{pmatrix} \otimes \bar{\mathbf{p}} = \mathbf{F}(\mathbf{p})\mathbf{E}(\bar{\mathbf{p}}) \begin{pmatrix} 0 \\ \mathbf{v} \end{pmatrix} \quad (6.213)$$

This leads to one more formula for the rotation matrix:

$$\mathbf{R} = \begin{pmatrix} -\epsilon & \eta\mathbf{I} + \boldsymbol{\epsilon}^\times \end{pmatrix} \begin{pmatrix} -\epsilon & \eta\mathbf{I} - \boldsymbol{\epsilon}^\times \end{pmatrix}^T \quad (6.214)$$

### 6.7.6 Euler parameters from the rotation matrix

The problem to be solved in this section is how to find the Euler parameters  $\eta, \boldsymbol{\epsilon}$  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):

$$\mathbf{R} = \mathbf{R}_e(\eta, \boldsymbol{\epsilon}) = \begin{pmatrix} \eta^2 + \epsilon_1^2 - \epsilon_2^2 - \epsilon_3^2 & 2(\epsilon_1\epsilon_2 - \eta\epsilon_3) & 2(\epsilon_1\epsilon_3 + \eta\epsilon_2) \\ 2(\epsilon_1\epsilon_2 + \eta\epsilon_3) & \eta^2 - \epsilon_1^2 + \epsilon_2^2 - \epsilon_3^2 & 2(\epsilon_2\epsilon_3 - \eta\epsilon_1) \\ 2(\epsilon_1\epsilon_3 - \eta\epsilon_2) & 2(\epsilon_2\epsilon_3 + \eta\epsilon_1) & \eta^2 - \epsilon_1^2 - \epsilon_2^2 + \epsilon_3^2 \end{pmatrix} \quad (6.215)$$

In addition,  $\eta^2 + \epsilon_1^2 + \epsilon_2^2 + \epsilon_3^2 = 1$ . The following notation is introduced to simplify the algorithms:

$$\mathbf{z} = \begin{pmatrix} z_0 \\ z_1 \\ z_2 \\ z_3 \end{pmatrix} := 2 \begin{pmatrix} \eta \\ \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix} \quad (6.216)$$

$$T := r_{11} + r_{22} + r_{33} = \text{Trace}\mathbf{R} \quad (6.217)$$

and

$$r_{00} := T \quad (6.218)$$

This gives the symmetric set of equations

$$z_0^2 = 1 + 2r_{00} - T \quad (6.219)$$

$$z_1^2 = 1 + 2r_{11} - T \quad (6.220)$$

$$z_2^2 = 1 + 2r_{22} - T \quad (6.221)$$

$$z_3^2 = 1 + 2r_{33} - T \quad (6.222)$$

that appear from the diagonal elements of  $\mathbf{R}$ , while the off-diagonal terms give the equations

$$z_0 z_1 = r_{32} - r_{23} \quad (6.223)$$

$$z_0 z_2 = r_{13} - r_{31} \quad (6.224)$$

$$z_0 z_3 = r_{21} - r_{12} \quad (6.225)$$

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
- $$|z_i| = \sqrt{1 + 2r_{ii} - T} \quad (6.226)$$
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

$$z_1 = (r_{32} - r_{23})/z_0 \quad (6.227)$$

$$z_2 = (r_{13} - r_{31})/z_0 \quad (6.228)$$

$$z_3 = (r_{21} - r_{12})/z_0 \quad (6.229)$$

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

$$\mathbf{e} = \mathbf{k} \sin \theta \in R^3 \quad (6.230)$$

is defined from the angle-axis parameters  $(\mathbf{k}, \theta)$ . From (6.145) it is seen that the rotation matrix  $\mathbf{R}_{k,\theta}$  and its transpose  $\mathbf{R}_{k,\theta}^T$  are given by

$$\mathbf{R}_{k,\theta} = \mathbf{e}^\times + \cos \theta \mathbf{I} + \mathbf{k} \mathbf{k}^T (1 - \cos \theta) \quad (6.231)$$

$$\mathbf{R}_{k,\theta}^T = -\mathbf{e}^\times + \cos \theta \mathbf{I} + \mathbf{k} \mathbf{k}^T (1 - \cos \theta) \quad (6.232)$$

which implies that

$$\mathbf{e}^\times = \frac{1}{2} (\mathbf{R}_{k,\theta} - \mathbf{R}_{k,\theta}^T) \quad (6.233)$$

From this we see that if  $\mathbf{R}_{k,\theta} = \{r_{ij}\}$ , then the Euler rotation vector can be found from

$$\mathbf{e} = \frac{1}{2} \begin{pmatrix} r_{32} - r_{23} \\ r_{13} - r_{31} \\ r_{21} - r_{12} \end{pmatrix} \quad (6.234)$$

We note that if  $\mathbf{R}_{k,\theta} = \mathbf{R}_b^a$ , then

$$\mathbf{e} = \mathbf{e}^a = \mathbf{e}^b \quad (6.235)$$

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

$$\mathbf{R}_d = \begin{pmatrix} \mathbf{n}_d & \mathbf{s}_d & \mathbf{a}_d \end{pmatrix} \in SO(3) \quad (6.236)$$

Suppose that the actual orientation of the robot hand is

$$\mathbf{R} = \begin{pmatrix} \mathbf{n} & \mathbf{s} & \mathbf{a} \end{pmatrix} \in SO(3) \quad (6.237)$$

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  $\tilde{\mathbf{R}} = \{\tilde{r}_{ij}\}$  which is defined by

$$\tilde{\mathbf{R}} := \mathbf{R}\mathbf{R}_d^T \quad \Rightarrow \quad \mathbf{R} = \tilde{\mathbf{R}}\mathbf{R}_d \quad (6.238)$$

The component form of this equation is

$$\tilde{r}_{ij} = n_i n_{dj} + s_i s_{dj} + a_i a_{dj} \quad (6.239)$$

If an angle-axis parameters of  $\tilde{\mathbf{R}}$  are  $(\tilde{\mathbf{k}}, \tilde{\theta})$  and  $\tilde{\mathbf{e}} = \tilde{\mathbf{k}} \sin \tilde{\theta}$  is the associated Euler rotation vector, then (6.234) gives

$$\begin{aligned} \tilde{\mathbf{e}} &= \frac{1}{2} \begin{pmatrix} \tilde{r}_{32} - \tilde{r}_{23} \\ \tilde{r}_{13} - \tilde{r}_{31} \\ \tilde{r}_{21} - \tilde{r}_{12} \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} n_3 n_{d2} - n_2 n_{d3} \\ n_1 n_{d3} - n_3 n_{d1} \\ n_2 n_{d1} - n_1 n_{d2} \end{pmatrix} + \frac{1}{2} \begin{pmatrix} s_3 s_{d2} - s_2 s_{d3} \\ s_1 s_{d3} - s_3 s_{d1} \\ s_2 s_{d1} - s_1 s_{d2} \end{pmatrix} + \frac{1}{2} \begin{pmatrix} a_3 a_{d2} - a_2 a_{d3} \\ a_1 a_{d3} - a_3 a_{d1} \\ a_2 a_{d1} - a_1 a_{d2} \end{pmatrix} \end{aligned}$$

Using the definition of the vector cross product the Euler rotation vector  $\tilde{\mathbf{e}}$  corresponding to the deviation  $\tilde{\mathbf{R}}$  can be written

$$\tilde{\mathbf{e}} = \frac{1}{2}(\mathbf{n}_d^\times \mathbf{n} + \mathbf{s}_d^\times \mathbf{s} + \mathbf{a}_d^\times \mathbf{a}) \quad (6.240)$$

### 6.7.8 Euler-Rodrigues parameters

The Euler-Rodrigues parameters are defined by (Hughes 1986)

$$\boldsymbol{\rho} = \mathbf{k} \tan \frac{\theta}{2} \quad (6.241)$$

This can be expressed in terms of the Euler parameters according to

$$\boldsymbol{\rho} = \frac{\boldsymbol{\epsilon}}{\eta} \quad (6.242)$$

It is evident that the Euler-Rodrigues parameters are undefined when  $\eta = 0 \Leftrightarrow \theta = \pi + 2k\pi$  where  $k \in \{\dots -1, 0, 1 \dots\}$ .

The derivations that follows use the relation

$$1 = \eta^2 + \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = \eta^2 (1 + \boldsymbol{\rho}^T \boldsymbol{\rho}) \quad (6.243)$$

which implies

$$\eta^2 = \frac{1}{1 + \boldsymbol{\rho}^T \boldsymbol{\rho}} \quad (6.244)$$

Then

$$\mathbf{R} = \mathbf{I} + \frac{2}{1 + \boldsymbol{\rho}^T \boldsymbol{\rho}} [\boldsymbol{\rho}^\times + \boldsymbol{\rho}^\times \boldsymbol{\rho}^\times] \quad (6.245)$$

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

$$\boldsymbol{\rho} = \frac{\boldsymbol{\epsilon}}{\eta} = \frac{\mathbf{e}}{2\eta^2} = \frac{1}{\text{Trace}\mathbf{R} + 1} \begin{pmatrix} r_{32} - r_{23} \\ r_{13} - r_{31} \\ r_{21} - r_{12} \end{pmatrix} \quad (6.246)$$

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

$$\boldsymbol{\rho}^\times \boldsymbol{\rho}^\times \boldsymbol{\rho}^\times = -(\boldsymbol{\rho}^T \boldsymbol{\rho}) \boldsymbol{\rho}^\times \quad (6.247)$$

Using this result and (6.245) we find that

$$\mathbf{R} (\mathbf{I} - \boldsymbol{\rho}^\times) = \left[ \mathbf{I} + \frac{2}{1 + \boldsymbol{\rho}^T \boldsymbol{\rho}} (\boldsymbol{\rho}^\times + \boldsymbol{\rho}^\times \boldsymbol{\rho}^\times) \right] (\mathbf{I} - \boldsymbol{\rho}^\times) = \mathbf{I} + \boldsymbol{\rho}^\times \quad (6.248)$$

This leads to the following result:

The rotation matrix can be given by Cayley's formula

$$\mathbf{R} = (\mathbf{I} + \boldsymbol{\rho}^\times) (\mathbf{I} - \boldsymbol{\rho}^\times)^{-1} \quad (6.249)$$

where  $\boldsymbol{\rho}$  is the vector of Euler-Rodrigues parameters corresponding to  $\mathbf{R}$ .

The *Cayley transformation*  $\text{cay}(\mathbf{u}) \in SO(3)$  maps a three-dimensional vector  $\mathbf{u}$  into a rotation matrix according to

$$\text{cay}(\mathbf{u}) := \left[ \mathbf{I} + \frac{1}{2} \mathbf{u}^\times \right] \left[ \mathbf{I} - \frac{1}{2} \mathbf{u}^\times \right]^{-1} \in SO(3) \quad (6.250)$$

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

$$\mathbf{R} = \text{cay}(2\boldsymbol{\rho}) \quad (6.251)$$

and that

$$\text{cay}(\mathbf{k}\theta) \approx \mathbf{R}_{k,\theta}, \quad \theta \text{ small} \quad (6.252)$$

## 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

$$\mathbf{R}_b^a(\mathbf{R}_b^a)^T = \mathbf{I} \quad (6.253)$$

Time differentiation of the matrix product gives

$$\frac{d}{dt} [\mathbf{R}_b^a(\mathbf{R}_b^a)^T] = \dot{\mathbf{R}}_b^a(\mathbf{R}_b^a)^T + \mathbf{R}_b^a(\dot{\mathbf{R}}_b^a)^T = \mathbf{0} \quad (6.254)$$

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  $3 \times 3$  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

$$(\boldsymbol{\omega}_{ab}^a)^\times = \dot{\mathbf{R}}_b^a(\mathbf{R}_b^a)^T \quad (6.255)$$

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

$$\dot{\mathbf{R}}_b^a = (\boldsymbol{\omega}_{ab}^a)^\times \mathbf{R}_b^a \quad (6.256)$$

$$\dot{\mathbf{R}}_b^a = \mathbf{R}_b^a(\boldsymbol{\omega}_{ab}^b)^\times \quad (6.257)$$

### 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

$$[\boldsymbol{\omega}_x(\dot{\phi})]^\times : = \dot{\mathbf{R}}_x(\phi)\mathbf{R}_x^T(\phi) \quad (6.258)$$

$$[\boldsymbol{\omega}_y(\dot{\theta})]^\times : = \dot{\mathbf{R}}_y(\theta)\mathbf{R}_y^T(\theta) \quad (6.259)$$

$$[\boldsymbol{\omega}_z(\dot{\psi})]^\times : = \dot{\mathbf{R}}_z(\psi)\mathbf{R}_z^T(\psi) \quad (6.260)$$

From the definitions it is clear that  $\omega_x(\dot{\phi})$  is the angular velocity of a rotation by an angular rate  $\dot{\phi}$  about the  $x$  axis,  $\omega_y(\dot{\theta})$  is the angular velocity of a rotation by an angular rate  $\dot{\theta}$  about the  $y$  axis, and  $\omega_z(\dot{\psi})$  is the angular velocity of a rotation by an angular rate  $\dot{\psi}$  about the  $z$  axis. From (6.101) we find that

$$[\omega_x(\dot{\phi})]^\times = \dot{\phi} \begin{pmatrix} 0 & 0 & 0 \\ 0 & -\sin \phi & -\cos \phi \\ 0 & \cos \phi & -\sin \phi \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos \phi & \sin \phi \\ 0 & -\sin \phi & \cos \phi \end{pmatrix} \quad (6.261)$$

$$= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\dot{\phi} \\ 0 & \dot{\phi} & 0 \end{pmatrix} \quad (6.262)$$

In the same way we may compute  $[\omega_y(\dot{\theta})]^\times$  and  $[\omega_z(\dot{\psi})]^\times$ . This results in

$$\omega_x(\dot{\phi}) = \begin{pmatrix} \dot{\phi} \\ 0 \\ 0 \end{pmatrix}, \quad \omega_y(\dot{\theta}) = \begin{pmatrix} 0 \\ \dot{\theta} \\ 0 \end{pmatrix}, \quad \text{and} \quad \omega_z(\dot{\psi}) = \begin{pmatrix} 0 \\ 0 \\ \dot{\psi} \end{pmatrix} \quad (6.263)$$

Consider the angle-axis parameterization when  $\mathbf{k} = \mathbf{k}^a$  is a constant vector and

$$\mathbf{R}_b^a = \mathbf{R}_{k,\theta} = \mathbf{I} + \mathbf{k}^\times \sin \theta + \mathbf{k}^\times \mathbf{k}^\times (1 - \cos \theta) \quad (6.264)$$

Then the angular velocity is

$$\begin{aligned} (\omega_{ab}^a)^\times &= \dot{\theta} (\mathbf{k}^\times \cos \theta + \mathbf{k}^\times \mathbf{k}^\times \sin \theta) (\mathbf{I} - \mathbf{k}^\times \sin \theta + \mathbf{k}^\times \mathbf{k}^\times (1 - \cos \theta)) \\ &= \dot{\theta} [\mathbf{k}^\times \cos \theta + \mathbf{k}^\times \mathbf{k}^\times \sin \theta - \mathbf{k}^\times \mathbf{k}^\times \cos \theta \sin \theta + \mathbf{k}^\times \sin^2 \theta \\ &\quad - \mathbf{k}^\times (\cos \theta - \cos^2 \theta) + \mathbf{k}^\times \mathbf{k}^\times \cos \theta \sin \theta - \mathbf{k}^\times \mathbf{k}^\times \sin \theta] \\ &= \dot{\theta} \mathbf{k}^\times \end{aligned} \quad (6.265)$$

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

$$\vec{\omega}_{ab} = \dot{\theta} \vec{k} \quad (6.266)$$

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  $\mathbf{R}_d^a = \mathbf{R}_b^a \mathbf{R}_c^b \mathbf{R}_d^c$ . The time derivative of  $\mathbf{R}_d^a$  is, according to the product rule,

$$\dot{\mathbf{R}}_d^a = \dot{\mathbf{R}}_b^a \mathbf{R}_c^b \mathbf{R}_d^c + \mathbf{R}_b^a \dot{\mathbf{R}}_c^b \mathbf{R}_d^c + \mathbf{R}_b^a \mathbf{R}_c^b \dot{\mathbf{R}}_d^c \quad (6.267)$$

and the transpose is  $(\mathbf{R}_d^a)^T = (\mathbf{R}_d^c)^T (\mathbf{R}_c^b)^T (\mathbf{R}_b^a)^T$ . Then the angular velocity of the composite rotation is

$$\begin{aligned} (\omega_{ad}^a)^\times &= \dot{\mathbf{R}}_d^a (\mathbf{R}_d^a)^T = \left( \dot{\mathbf{R}}_b^a \mathbf{R}_c^b \mathbf{R}_d^c + \mathbf{R}_b^a \dot{\mathbf{R}}_c^b \mathbf{R}_d^c + \mathbf{R}_b^a \mathbf{R}_c^b \dot{\mathbf{R}}_d^c \right) (\mathbf{R}_d^c)^T (\mathbf{R}_c^b)^T (\mathbf{R}_b^a)^T \\ &= \dot{\mathbf{R}}_b^a (\mathbf{R}_b^a)^T + \mathbf{R}_b^a \dot{\mathbf{R}}_c^b (\mathbf{R}_c^b)^T (\mathbf{R}_b^a)^T + \mathbf{R}_c^a \dot{\mathbf{R}}_d^c (\mathbf{R}_d^c)^T (\mathbf{R}_c^a)^T \\ &= (\omega_{ab}^a)^\times + \mathbf{R}_b^a (\omega_{bc}^b)^\times (\mathbf{R}_b^a)^T + \mathbf{R}_c^a (\omega_{cd}^c)^\times (\mathbf{R}_c^a)^T \\ &= (\omega_{ab}^a)^\times + (\omega_{bc}^a)^\times + (\omega_{cd}^a)^\times \end{aligned} \quad (6.268)$$

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

$$\vec{\omega}_{ad} = \vec{\omega}_{ab} + \vec{\omega}_{bc} + \vec{\omega}_{cd} \quad (6.269)$$

**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

$$\mathbf{R}_p^b = \mathbf{R}_z(\psi) \mathbf{R}_y(\theta) \mathbf{R}_x(\phi) \quad (6.270)$$

which corresponds to the angular velocity vector

$$\boldsymbol{\omega}_{bp}^b = \boldsymbol{\omega}_z(\dot{\psi}) + \mathbf{R}_z(\psi) \boldsymbol{\omega}_y(\dot{\theta}) + \mathbf{R}_z(\psi) \mathbf{R}_y(\theta) \boldsymbol{\omega}_x(\dot{\phi}) \quad (6.271)$$

### 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:

$$\dot{\mathbf{u}}^a := \frac{d}{dt}(\mathbf{u}^a) = \frac{d}{dt} \begin{pmatrix} u_1^a \\ u_2^a \\ u_3^a \end{pmatrix} = \begin{pmatrix} \dot{u}_1^a \\ \dot{u}_2^a \\ \dot{u}_3^a \end{pmatrix} \quad (6.272)$$

The relation between the time derivative in frame  $a$  and the time derivative in frame  $b$  is found by differentiating the equation

$$\mathbf{u}^a = \mathbf{R}_b^a \mathbf{u}^b \quad (6.273)$$

which gives

$$\dot{\mathbf{u}}^a = \mathbf{R}_b^a \dot{\mathbf{u}}^b + \dot{\mathbf{R}}_b^a \mathbf{u}^b \quad (6.274)$$

Insertion of  $\dot{\mathbf{R}}_b^a = \mathbf{R}_b^a (\boldsymbol{\omega}_{ab}^b)^\times$  gives the relation

$$\dot{\mathbf{u}}^a = \mathbf{R}_b^a [\dot{\mathbf{u}}^b + (\boldsymbol{\omega}_{ab}^b)^\times \mathbf{u}^b] \quad (6.275)$$

### 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

$$\overset{a}{\frac{d}{dt}} \vec{u} := \dot{u}_1^a \vec{a}_1 + \dot{u}_2^a \vec{a}_2 + \dot{u}_3^a \vec{a}_3 \quad (6.276)$$

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

$$\overset{b}{\frac{d}{dt}} \vec{u} = \dot{u}_1^b \vec{b}_1 + \dot{u}_2^b \vec{b}_2 + \dot{u}_3^b \vec{b}_3 \quad (6.277)$$

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

$$\dot{\mathbf{u}}^a = \begin{pmatrix} \dot{u}_1^a \\ \dot{u}_2^a \\ \dot{u}_3^a \end{pmatrix}, \quad \dot{\mathbf{u}}^b = \begin{pmatrix} \dot{u}_1^b \\ \dot{u}_2^b \\ \dot{u}_3^b \end{pmatrix} \quad (6.278)$$

From (6.275) we find that

$$\frac{^a d}{dt} \vec{u} = \frac{^b d}{dt} \vec{u} + \vec{\omega}_{ab} \times \vec{u} \quad (6.279)$$

where

$$\frac{^a d}{dt} \vec{u} = \dot{u}_1^a \vec{a}_1 + \dot{u}_2^a \vec{a}_2 + \dot{u}_3^a \vec{a}_3 \quad (6.280)$$

$$\frac{^b d}{dt} \vec{u} = \dot{u}_1^b \vec{b}_1 + \dot{u}_2^b \vec{b}_2 + \dot{u}_3^b \vec{b}_3 \quad (6.281)$$

**Example 107** In the same way partial differentiation with respect to some variable  $q$  in frame  $a$  and  $b$  is defined by

$$\frac{^a \partial \vec{u}}{\partial q} : = \frac{\partial u_1^a}{\partial q} \vec{a}_1 + \frac{\partial u_2^a}{\partial q} \vec{a}_2 + \frac{\partial u_3^a}{\partial q} \vec{a}_3 \quad (6.282)$$

$$\frac{^b \partial \vec{u}}{\partial q} : = \frac{\partial u_1^b}{\partial q} \vec{b}_1 + \frac{\partial u_2^b}{\partial q} \vec{b}_2 + \frac{\partial u_3^b}{\partial q} \vec{b}_3 \quad (6.283)$$

**Example 108** An alternative definition of the angular velocity vector is used in (Kane and Levinson 1985):

$$\vec{\omega}_{ab} = \vec{b}_1 \left( \frac{^a d \vec{b}_2}{dt} \cdot \vec{b}_3 \right) + \vec{b}_2 \left( \frac{^a d \vec{b}_3}{dt} \cdot \vec{b}_1 \right) + \vec{b}_3 \left( \frac{^a d \vec{b}_1}{dt} \cdot \vec{b}_2 \right) \quad (6.284)$$

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

$$\mathbf{R}_b^a = (\mathbf{b}_1^a \quad \mathbf{b}_2^a \quad \mathbf{b}_3^a) \quad (6.285)$$

and from the definition of  $\boldsymbol{\omega}_{ab}^b$  in (6.255) we get

$$(\boldsymbol{\omega}_{ab}^b)^\times = \mathbf{R}^T \dot{\mathbf{R}} = \begin{pmatrix} 0 & \mathbf{b}_1^{aT} \dot{\mathbf{b}}_2^a & \mathbf{b}_1^{aT} \dot{\mathbf{b}}_3^a \\ \mathbf{b}_2^{aT} \dot{\mathbf{b}}_1^a & 0 & \mathbf{b}_2^{aT} \dot{\mathbf{b}}_3^a \\ \mathbf{b}_3^{aT} \dot{\mathbf{b}}_1^a & \mathbf{b}_3^{aT} \dot{\mathbf{b}}_2^a & 0 \end{pmatrix} \quad (6.286)$$

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

$$\boldsymbol{\omega}_{ab}^b = \begin{pmatrix} \mathbf{b}_3^{aT} \dot{\mathbf{b}}_2^a \\ \mathbf{b}_1^{aT} \dot{\mathbf{b}}_3^a \\ \mathbf{b}_2^{aT} \dot{\mathbf{b}}_1^a \end{pmatrix} = \begin{pmatrix} \vec{b}_3 \cdot \frac{^a d \vec{b}_2}{dt} \\ \vec{b}_1 \cdot \frac{^a d \vec{b}_3}{dt} \\ \vec{b}_2 \cdot \frac{^a d \vec{b}_1}{dt} \end{pmatrix} \quad (6.287)$$

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  $\tilde{\mathbf{R}}_a \in SO(3)$  defined by

$$\tilde{\mathbf{R}}_a := \mathbf{R}\mathbf{R}_d^T \quad \Rightarrow \quad \mathbf{R} = \tilde{\mathbf{R}}_a\mathbf{R}_d \quad (6.288)$$

We see that the rotation matrix  $\mathbf{R}$  is described as the composite rotation defined by the rotation matrices  $\tilde{\mathbf{R}}_a$  and  $\mathbf{R}_d$ . To make this clear we introduce the intermediate frame  $c$  so that

$$\tilde{\mathbf{R}}_a = \mathbf{R}_c^a, \quad \mathbf{R}_d = \mathbf{R}_b^c \quad (6.289)$$

and

$$\frac{d}{dt}\tilde{\mathbf{R}}_a = \dot{\mathbf{R}}_c^a = (\boldsymbol{\omega}_{ac}^a)^\times \mathbf{R}_c^a \quad (6.290)$$

$$\dot{\mathbf{R}}_d = \dot{\mathbf{R}}_b^c = \mathbf{R}_b^c(\boldsymbol{\omega}_{cb}^b)^\times \quad (6.291)$$

Then, if we define the angular velocity vectors

$$\boldsymbol{\omega}^a = \boldsymbol{\omega}_{ab}^a, \quad \tilde{\boldsymbol{\omega}}^a := \boldsymbol{\omega}_{ac}^a, \quad \boldsymbol{\omega}_d^b := \boldsymbol{\omega}_{cb}^b \quad (6.292)$$

we find that the kinematic differential equations for  $\mathbf{R}$ ,  $\tilde{\mathbf{R}}_a$  and  $\mathbf{R}_d$  are given by

$$\dot{\mathbf{R}} = (\boldsymbol{\omega}^a)^\times \mathbf{R}, \quad \frac{d}{dt}\tilde{\mathbf{R}}_a = (\tilde{\boldsymbol{\omega}}^a)^\times \tilde{\mathbf{R}}_a, \quad \dot{\mathbf{R}}_d = \mathbf{R}_d(\boldsymbol{\omega}_d^b)^\times \quad (6.293)$$

It follows from

$$\boldsymbol{\omega}_{ab}^a = \boldsymbol{\omega}_{ac}^a + \boldsymbol{\omega}_{cb}^a \quad (6.294)$$

that  $\tilde{\boldsymbol{\omega}}^a = \boldsymbol{\omega}^a - \boldsymbol{\omega}_d^a$ , and we may sum up that the kinematic differential equations for the attitude deviation is

$$\tilde{\mathbf{R}}_a := \mathbf{R}\mathbf{R}_d^T \quad (6.295)$$

$$\tilde{\boldsymbol{\omega}}^a = \boldsymbol{\omega}^a - \boldsymbol{\omega}_d^a \quad (6.296)$$

$$\frac{d}{dt}\tilde{\mathbf{R}}_a = (\tilde{\boldsymbol{\omega}}^a)^\times \tilde{\mathbf{R}}_a \quad (6.297)$$

We define an alternative representation of the deviation between the two rotations using the rotation matrix  $\tilde{\mathbf{R}}_b \in SO(3)$  defined by

$$\tilde{\mathbf{R}}_b := \mathbf{R}_d^T \mathbf{R} \Rightarrow \mathbf{R} = \mathbf{R}_d \tilde{\mathbf{R}}_b \quad (6.298)$$

The desired angular velocity  $\boldsymbol{\omega}_d^a$  is in this case defined by

$$\mathbf{R}_d = (\boldsymbol{\omega}_d^a)^\times \mathbf{R}_d \quad (6.299)$$

Then, by introducing an intermediate frame as in the case above, we find that the kinematic differential equations referred to the  $b$  frame is

$$\tilde{\mathbf{R}}_b := \mathbf{R}_d^T \mathbf{R} \quad (6.300)$$

$$\tilde{\boldsymbol{\omega}}^b = \boldsymbol{\omega}^b - \boldsymbol{\omega}_d^b \quad (6.301)$$

$$\frac{d}{dt}\tilde{\mathbf{R}}_b = \tilde{\mathbf{R}}_b(\tilde{\boldsymbol{\omega}}^b)^\times \quad (6.302)$$

### 6.9.3 Homogeneous transformation matrices

The time derivative of the homogeneous transformation matrix

$$\mathbf{T}_b^a = \begin{pmatrix} \mathbf{R}_b^a & \mathbf{r}_{ab}^a \\ \mathbf{0}^T & 1 \end{pmatrix} \in SE(3) \quad (6.303)$$

is found to be

$$\begin{aligned} \dot{\mathbf{T}}_b^a &= \begin{pmatrix} \mathbf{R}_b^a(\boldsymbol{\omega}_{ab}^b)^\times & \dot{\mathbf{r}}_{ab}^a \\ \mathbf{0}^T & 0 \end{pmatrix} = \begin{pmatrix} \mathbf{R}_b^a & \mathbf{r}_{ab}^a \\ \mathbf{0}^T & 1 \end{pmatrix} \begin{pmatrix} (\boldsymbol{\omega}_{ab}^b)^\times & \mathbf{v}_{ab}^b \\ \mathbf{0}^T & 0 \end{pmatrix} \\ &= \mathbf{T}_b^a \begin{pmatrix} (\boldsymbol{\omega}_{ab}^b)^\times & \mathbf{v}_{ab}^b \\ \mathbf{0}^T & 0 \end{pmatrix} \end{aligned} \quad (6.304)$$

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

$$\mathbf{w} = \begin{pmatrix} \mathbf{v}_{ab}^b \\ \boldsymbol{\omega}_{ab}^b \end{pmatrix} \quad (6.305)$$

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  $\omega_{ab}^b$  and its matrix form  $(\omega_{ab}^b)^\times$ , the twist vector  $\mathbf{w}$  has a matrix form in the set  $se(3)$  which is

$$\hat{\mathbf{w}} = \begin{pmatrix} (\omega_{ab}^b)^\times & \mathbf{v}_{ab}^b \\ \mathbf{0}^T & 0 \end{pmatrix} \in se(3) \quad (6.306)$$

The time derivative of the homogeneous transformation matrix is given by

$$\dot{\mathbf{T}}_b^a = \mathbf{T}_b^a \hat{\mathbf{w}} \quad (6.307)$$

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:*

$$\begin{aligned} \dot{\mathbf{T}}_b^a &= \begin{pmatrix} (\omega_{ab}^a)^\times \mathbf{R}_b^a & \mathbf{v}_{ab}^a \\ \mathbf{0}^T & 0 \end{pmatrix} \\ &= \begin{pmatrix} (\omega_{ab}^a)^\times & \mathbf{v}_{ab}^a - (\omega_{ab}^a)^\times \mathbf{r}_{ab}^a \\ \mathbf{0}^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{R}_b^a & \mathbf{r}_{ab}^a \\ \mathbf{0}^T & 1 \end{pmatrix} \end{aligned} \quad (6.308)$$

The physical interpretation of the velocity term  $\mathbf{v}_{ab}^a - (\omega_{ab}^a)^\times \mathbf{r}_{ab}^a$  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

$$\mathbf{R}_b^a = \mathbf{R}_z(\psi), \quad \mathbf{R}_c^b = \mathbf{R}_y(\theta) \quad \text{and} \quad \mathbf{R}_d^c = \mathbf{R}_x(\phi) \quad (6.309)$$

We see that the angular velocities associated with the simple rotations are

$$\omega_{ab}^a = \begin{pmatrix} 0 \\ 0 \\ \dot{\psi} \end{pmatrix}, \quad \omega_{bc}^b = \begin{pmatrix} 0 \\ \dot{\theta} \\ 0 \end{pmatrix} \quad \text{and} \quad \omega_{cd}^c = \begin{pmatrix} \dot{\phi} \\ 0 \\ 0 \end{pmatrix} \quad (6.310)$$

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  $\psi$ ,  $\theta$  and  $\phi$ :

$$\vec{\omega}_{ad} = \vec{\omega}_{ab} + \vec{\omega}_{bc} + \vec{\omega}_{cd} \quad (6.311)$$

In the  $a$  frame this gives:

$$\begin{aligned} \omega_{ad}^a &= \begin{pmatrix} 0 \\ 0 \\ \dot{\psi} \end{pmatrix} + \mathbf{R}_{z,\psi} \begin{pmatrix} 0 \\ \dot{\theta} \\ 0 \end{pmatrix} + \mathbf{R}_{z,\psi} \mathbf{R}_{y,\theta} \begin{pmatrix} \dot{\phi} \\ 0 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} -\sin \psi \dot{\theta} + \cos \psi \cos \theta \dot{\phi} \\ \cos \psi \dot{\theta} + \sin \psi \cos \theta \dot{\phi} \\ \dot{\psi} - \sin \theta \dot{\phi} \end{pmatrix} \end{aligned} \quad (6.312)$$

In the  $d$  frame we find

$$\boldsymbol{\omega}_{ad}^d = \mathbf{R}_{x,-\phi} \mathbf{R}_{y,-\theta} \begin{pmatrix} 0 \\ 0 \\ \dot{\psi} \end{pmatrix} + \mathbf{R}_{x,-\phi} \begin{pmatrix} 0 \\ \dot{\theta} \\ 0 \end{pmatrix} + \begin{pmatrix} \dot{\phi} \\ 0 \\ 0 \end{pmatrix} \quad (6.313)$$

$$= \begin{pmatrix} -\sin \theta \dot{\psi} + \dot{\phi} \\ \sin \phi \cos \theta \dot{\psi} + \cos \phi \dot{\theta} \\ \cos \phi \cos \theta \dot{\psi} - \sin \phi \dot{\theta} \end{pmatrix} \quad (6.314)$$

Define the vector  $\boldsymbol{\phi} = (\phi, \theta, \psi)^T$ . We can then write

$$\boldsymbol{\omega}_{ad}^a = \mathbf{E}_a(\boldsymbol{\phi}) \dot{\boldsymbol{\phi}} = \begin{pmatrix} \cos \psi \cos \theta & -\sin \psi & 0 \\ \sin \psi \cos \theta & \cos \psi & 0 \\ -\sin \theta & 0 & 1 \end{pmatrix} \dot{\boldsymbol{\phi}} \quad (6.315)$$

and

$$\boldsymbol{\omega}_{ad}^d = \mathbf{E}_d(\boldsymbol{\phi}) \dot{\boldsymbol{\phi}} = \begin{pmatrix} 1 & 0 & -\sin \theta \\ 0 & \cos \phi & \sin \phi \cos \theta \\ 0 & -\sin \phi & \cos \phi \cos \theta \end{pmatrix} \dot{\boldsymbol{\phi}}. \quad (6.316)$$

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

$$\dot{\boldsymbol{\phi}} = \mathbf{E}_a(\boldsymbol{\phi})^{-1} \boldsymbol{\omega}_{ad}^a = \frac{1}{\cos \theta} \begin{pmatrix} \cos \psi & \sin \psi & 0 \\ -\sin \psi \cos \theta & \cos \psi \cos \theta & 0 \\ \cos \psi \sin \theta & \sin \psi \sin \theta & \cos \theta \end{pmatrix} \boldsymbol{\omega}_{ad}^a \quad (6.317)$$

and

$$\dot{\boldsymbol{\phi}} = \mathbf{E}_d(\boldsymbol{\phi})^{-1} \boldsymbol{\omega}_{ad}^d = \frac{1}{\cos \theta} \begin{pmatrix} \cos \theta & \sin \phi \sin \theta & \cos \phi \sin \theta \\ 0 & \cos \phi \cos \theta & -\sin \phi \cos \theta \\ 0 & \sin \phi & \cos \phi \end{pmatrix} \boldsymbol{\omega}_{ad}^d \quad (6.318)$$

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  $\mathbf{R} = \mathbf{R}_e(\eta, \epsilon)$  and  $\mathbf{p} = (\eta \ \epsilon^T)^T$ . We then have

$$\dot{\mathbf{R}} = (\boldsymbol{\omega}^a)^\times \mathbf{R} = \mathbf{R}(\boldsymbol{\omega}^b)^\times \quad (6.319)$$

The derivation is based on the coordinate transformation rule using the quaternion product. For an arbitrary vector  $\mathbf{u} \in R^3$  we have

$$\begin{pmatrix} 0 \\ \mathbf{R}\mathbf{u} \end{pmatrix} = \mathbf{p} \otimes \begin{pmatrix} 0 \\ \mathbf{u} \end{pmatrix} \otimes \bar{\mathbf{p}} \quad (6.320)$$

We take the time derivative of both sides and get

$$\begin{pmatrix} 0 \\ \dot{\mathbf{R}}\mathbf{u} \end{pmatrix} + \begin{pmatrix} 0 \\ \mathbf{R}\dot{\mathbf{u}} \end{pmatrix} = \dot{\mathbf{p}} \otimes \begin{pmatrix} 0 \\ \mathbf{u} \end{pmatrix} \otimes \bar{\mathbf{p}} + \mathbf{p} \otimes \begin{pmatrix} 0 \\ \dot{\mathbf{u}} \end{pmatrix} \otimes \bar{\mathbf{p}} + \mathbf{p} \otimes \begin{pmatrix} 0 \\ \mathbf{u} \end{pmatrix} \otimes \dot{\bar{\mathbf{p}}} \quad (6.321)$$

Then, because the transformation rule in (6.320) is valid for any vector it is also valid for  $\dot{\mathbf{u}}$ . This implies that

$$\begin{aligned} \begin{pmatrix} 0 \\ \dot{\mathbf{R}}\mathbf{u} \end{pmatrix} &= \dot{\mathbf{p}} \otimes \begin{pmatrix} 0 \\ \mathbf{u} \end{pmatrix} \otimes \bar{\mathbf{p}} + \mathbf{p} \otimes \begin{pmatrix} 0 \\ \mathbf{u} \end{pmatrix} \otimes \dot{\bar{\mathbf{p}}} \\ &= \dot{\mathbf{p}} \otimes \bar{\mathbf{p}} \otimes \mathbf{p} \otimes \begin{pmatrix} 0 \\ \mathbf{u} \end{pmatrix} \otimes \bar{\mathbf{p}} - \mathbf{p} \otimes \begin{pmatrix} 0 \\ \mathbf{u} \end{pmatrix} \otimes \bar{\mathbf{p}} \otimes \dot{\mathbf{p}} \otimes \bar{\mathbf{p}} \\ &= (\dot{\mathbf{p}} \otimes \bar{\mathbf{p}}) \otimes \begin{pmatrix} 0 \\ \mathbf{R}\mathbf{u} \end{pmatrix} - \begin{pmatrix} 0 \\ \mathbf{R}\mathbf{u} \end{pmatrix} \otimes (\dot{\mathbf{p}} \otimes \bar{\mathbf{p}}) \\ &= 2 \begin{pmatrix} 0 \\ (\eta\dot{\epsilon} - \dot{\eta}\epsilon + \epsilon^\times\dot{\epsilon})^\times \mathbf{R}\mathbf{u} \end{pmatrix} \end{aligned} \quad (6.322)$$

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

$$\begin{pmatrix} 0 \\ \dot{\mathbf{R}}\mathbf{u} \end{pmatrix} = \begin{pmatrix} 0 \\ (\boldsymbol{\omega}^a)^\times \mathbf{R}\mathbf{u} \end{pmatrix} \quad (6.323)$$

Comparing this with (6.322) we find that the angular velocity  $\boldsymbol{\omega}^a$  is given by

$$\boldsymbol{\omega}^a = 2[\eta\dot{\epsilon} - \dot{\eta}\epsilon + \epsilon^\times\dot{\epsilon}] \quad (6.324)$$

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

$$\begin{pmatrix} 0 \\ \boldsymbol{\omega}^a \end{pmatrix} = 2\dot{\mathbf{p}} \otimes \bar{\mathbf{p}}, \quad \begin{pmatrix} 0 \\ \boldsymbol{\omega}^b \end{pmatrix} = 2\bar{\mathbf{p}} \otimes \dot{\mathbf{p}} \quad (6.325)$$

and the kinematic differential equation for the quaternion vector is

$$\dot{\mathbf{p}} = \frac{1}{2} \begin{pmatrix} 0 \\ \boldsymbol{\omega}^a \end{pmatrix} \otimes \mathbf{p}, \quad \dot{\bar{\mathbf{p}}} = \frac{1}{2} \mathbf{p} \otimes \begin{pmatrix} 0 \\ \boldsymbol{\omega}^b \end{pmatrix} \quad (6.326)$$

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  $\bar{\mathbf{p}}$  for the expression in the *b* frame.

The component form of these last four equations gives the result

$$\boldsymbol{\omega}^b = 2[\eta\dot{\epsilon} - \dot{\eta}\epsilon - \epsilon^\times\dot{\epsilon}] \quad (6.327)$$

$$\boldsymbol{\omega}^a = 2[\eta\dot{\epsilon} - \dot{\eta}\epsilon + \epsilon^\times\dot{\epsilon}] \quad (6.328)$$

$$\dot{\eta} = -\frac{1}{2}\boldsymbol{\epsilon}^T \boldsymbol{\omega}^b \quad (6.329)$$

$$\dot{\boldsymbol{\epsilon}} = \frac{1}{2}[\eta \mathbf{I} + \boldsymbol{\epsilon}^\times] \boldsymbol{\omega}^b \quad (6.330)$$

and

$$\dot{\eta} = -\frac{1}{2}\boldsymbol{\epsilon}^T \boldsymbol{\omega}^a \quad (6.331)$$

$$\dot{\boldsymbol{\epsilon}} = \frac{1}{2}[\eta \mathbf{I} - \boldsymbol{\epsilon}^\times] \boldsymbol{\omega}^a \quad (6.332)$$

**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

$$\dot{\mathbf{p}} = \frac{1}{2} \begin{pmatrix} 0 & -(\boldsymbol{\omega}^a)^T \\ \boldsymbol{\omega}^a & (\boldsymbol{\omega}^a)^\times \end{pmatrix} \mathbf{p} = \frac{1}{2} \begin{pmatrix} 0 & -(\boldsymbol{\omega}^b)^T \\ \boldsymbol{\omega}^b & -(\boldsymbol{\omega}^b)^\times \end{pmatrix} \mathbf{p} \quad (6.333)$$

or

$$\dot{\mathbf{p}} = \frac{1}{2} \begin{pmatrix} \eta & -\boldsymbol{\epsilon}^T \\ \boldsymbol{\epsilon} & \eta \mathbf{I} + \boldsymbol{\epsilon}^\times \end{pmatrix} \begin{pmatrix} 0 \\ \boldsymbol{\omega}^a \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \eta & -\boldsymbol{\epsilon}^T \\ \boldsymbol{\epsilon} & \eta \mathbf{I} - \boldsymbol{\epsilon}^\times \end{pmatrix} \begin{pmatrix} 0 \\ \boldsymbol{\omega}^a \end{pmatrix} \quad (6.334)$$

### 6.9.6 Normalization for numerical integration

From (6.334) it is seen that

$$\frac{d}{dt} (\mathbf{p}^T \mathbf{p}) = \mathbf{p}^T \begin{pmatrix} \eta & -\boldsymbol{\epsilon}^T \\ \boldsymbol{\epsilon} & \eta \mathbf{I} + \boldsymbol{\epsilon}^\times \end{pmatrix} \begin{pmatrix} 0 \\ \boldsymbol{\omega}^a \end{pmatrix} = 0 \quad (6.335)$$

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:

$$\dot{\mathbf{p}} = \frac{1}{2} \begin{pmatrix} \eta & -\boldsymbol{\epsilon}^T \\ \boldsymbol{\epsilon} & \eta \mathbf{I} + \boldsymbol{\epsilon}^\times \end{pmatrix} \begin{pmatrix} 0 \\ \boldsymbol{\omega}^a \end{pmatrix} + \frac{\lambda}{2}(1 - \mathbf{p}^T \mathbf{p})\mathbf{p} \quad (6.336)$$

Here  $\lambda$  is a positive gain. Then

$$\frac{d}{dt} (\mathbf{p}^T \mathbf{p}) = \frac{\lambda}{2}(1 - \mathbf{p}^T \mathbf{p})\mathbf{p}^T \mathbf{p} \quad (6.337)$$

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

$$\mathbf{p} := \frac{\mathbf{p}}{\sqrt{\mathbf{p}^T \mathbf{p}}} \quad (6.338)$$

### 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

$$\mathbf{e} = 2\eta\boldsymbol{\epsilon}. \quad (6.339)$$

The kinematic differential equation is then found from

$$\dot{\mathbf{e}} = 2(\dot{\eta}\boldsymbol{\epsilon} + \eta\dot{\boldsymbol{\epsilon}}). \quad (6.340)$$

This gives

$$\dot{\mathbf{e}} = [\eta^2 \mathbf{I} - \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T - \eta\boldsymbol{\epsilon}^\times]\boldsymbol{\omega}^a \quad (6.341)$$

$$\dot{\mathbf{e}} = [\eta^2 \mathbf{I} - \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T + \eta\boldsymbol{\epsilon}^\times]\boldsymbol{\omega}^b \quad (6.342)$$

where (6.331) and (6.332) are used.

Alternative expressions are found from

$$\mathbf{R} = (2\eta^2 - 1)\mathbf{I} + 2\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T + 2\eta\boldsymbol{\epsilon}^\times \quad (6.343)$$

and

$$\mathbf{R}^T = (2\eta^2 - 1)\mathbf{I} + 2\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T - 2\eta\boldsymbol{\epsilon}^\times \quad (6.344)$$

which leads to

$$\dot{\mathbf{e}} = \frac{1}{2}[\text{Trace}(\mathbf{R})\mathbf{I} - \mathbf{R}]\boldsymbol{\omega}^a \quad (6.345)$$

$$\dot{\mathbf{e}} = \frac{1}{2}[\text{Trace}(\mathbf{R}^T)\mathbf{I} - \mathbf{R}^T]\boldsymbol{\omega}^b \quad (6.346)$$

where we have used that  $4\eta^2 - 1 = \text{Trace}(\mathbf{R})$ .

Note that for  $\theta = 0$  we have

$$\dot{\mathbf{e}}|_{\theta=0} = \boldsymbol{\omega}^a = \boldsymbol{\omega}^b \quad (6.347)$$

### 6.9.8 Euler-Rodrigues parameters

The kinematic differential equation is derived from

$$\dot{\boldsymbol{\rho}} = \frac{d}{dt} \frac{\boldsymbol{\epsilon}}{\eta} = \frac{\eta\dot{\boldsymbol{\epsilon}} - \dot{\eta}\boldsymbol{\epsilon}}{\eta^2} \quad (6.348)$$

which gives

$$\dot{\boldsymbol{\rho}} = \frac{1}{2}[\mathbf{I} + \boldsymbol{\rho}^\times + \boldsymbol{\rho}\boldsymbol{\rho}^T]\boldsymbol{\omega}^b. \quad (6.349)$$

An equation for the angular velocity is found from

$$\boldsymbol{\rho}^\times \dot{\boldsymbol{\rho}} = \frac{1}{\eta}\boldsymbol{\epsilon}^\times \frac{\eta\dot{\boldsymbol{\epsilon}} - \dot{\eta}\boldsymbol{\epsilon}}{\eta^2} = \frac{1}{\eta^2}\boldsymbol{\epsilon}^\times \dot{\boldsymbol{\epsilon}} \quad (6.350)$$

as  $\boldsymbol{\epsilon}^\times \boldsymbol{\epsilon} = \mathbf{0}$ . From (6.327) it is seen that

$$\boldsymbol{\omega}^b = 2\eta^2 \left( \frac{\eta\dot{\boldsymbol{\epsilon}} - \dot{\eta}\boldsymbol{\epsilon}}{\eta^2} - \frac{1}{\eta^2}\boldsymbol{\epsilon}^\times \dot{\boldsymbol{\epsilon}} \right) \quad (6.351)$$

Insertion of (6.244), (6.348) and (6.350) gives

$$\boldsymbol{\omega}^b = \frac{2}{1 + \boldsymbol{\rho}^T \boldsymbol{\rho}} [\mathbf{I} - \boldsymbol{\rho}^\times] \dot{\boldsymbol{\rho}} \quad (6.352)$$

**Example 111** Equation (6.245) can be written

$$\mathbf{R} = \frac{1}{1 + \boldsymbol{\rho}^T \boldsymbol{\rho}} [2\mathbf{I} + 2\boldsymbol{\rho}^\times + 2\boldsymbol{\rho}\boldsymbol{\rho}^T - (1 + \boldsymbol{\rho}^T \boldsymbol{\rho})\mathbf{I}] \quad (6.353)$$

which implies that

$$\mathbf{I} + \boldsymbol{\rho}^\times + \boldsymbol{\rho}\boldsymbol{\rho}^T = \frac{1 + \boldsymbol{\rho}^T \boldsymbol{\rho}}{2} (\mathbf{R} + \mathbf{I}) \quad (6.354)$$

Then (6.349) can be written

$$\dot{\boldsymbol{\rho}} = \frac{1 + \boldsymbol{\rho}^T \boldsymbol{\rho}}{4} (\mathbf{R} + \mathbf{I}) \boldsymbol{\omega}^b \quad (6.355)$$

We recall from (6.169) that

$$\text{Trace}\mathbf{R} + 1 = 4\eta^2 \quad (6.356)$$

and, using (6.244), we arrive at

$$\dot{\boldsymbol{\rho}} = \frac{1}{\text{Trace}\mathbf{R} + 1} (\mathbf{R} + \mathbf{I}) \boldsymbol{\omega}^b \quad (6.357)$$

### 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

$$\dot{V}_x = \frac{\partial V_x}{\partial \mathbf{x}} \dot{\mathbf{x}} = \mathbf{x}^T \mathbf{v} \quad (6.358)$$

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

$$\dot{\eta} = -\frac{1}{2} \boldsymbol{\epsilon}^T \boldsymbol{\omega} \quad (6.359)$$

where  $|\eta| = |\cos \frac{\theta}{2}| \leq 1$ . Define

$$V_\epsilon = 2(1 - \eta) \geq 0 \quad (6.360)$$

The time derivative for solutions of the kinematic differential equations of the Euler parameters is

$$\dot{V}_\epsilon = -2\dot{\eta} = \boldsymbol{\epsilon}^T \boldsymbol{\omega} \quad (6.361)$$

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  $\eta$ , we get

$$\eta\dot{\eta} = -\frac{1}{2}\eta\boldsymbol{\epsilon}^T \boldsymbol{\omega} = -\frac{1}{4}\mathbf{e}^T \boldsymbol{\omega} \quad (6.362)$$

where  $\mathbf{e} = 2\eta\boldsymbol{\epsilon}$  is the Euler rotation vector. We are then lead to the function

$$V_e = 2(1 - \eta^2) \geq 0 \quad (6.363)$$

which has time derivative

$$\dot{V}_e = -4\eta\dot{\eta} = \mathbf{e}^T \boldsymbol{\omega} \quad (6.364)$$

and we have shown that the kinematic system with input  $\boldsymbol{\omega}$  and output  $\mathbf{e}$  is passive.

Finally we note that

$$\frac{\dot{\eta}}{\eta} = -\frac{1}{2} \frac{\mathbf{e}^T}{\eta} \boldsymbol{\omega} = -\frac{1}{2} \boldsymbol{\rho}^T \boldsymbol{\omega} \quad (6.365)$$

Define

$$V_\rho = -2 \ln |\eta| \geq 0, \quad \eta \neq 0 \quad (6.366)$$

which is defined for all  $\eta$  except for  $\eta = 0$  where also  $\boldsymbol{\rho}$  is undefined. Then

$$\dot{V}_\rho = -2 \frac{\dot{\eta}}{\eta} = \boldsymbol{\rho}^T \boldsymbol{\omega} \quad (6.367)$$

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

$$\boldsymbol{\epsilon}^T \boldsymbol{\epsilon} + (1 - \eta)^2 = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} + \eta^2 - 2\eta + 1 = 2(1 - \eta) = V_\epsilon \quad (6.368)$$

and that

$$2\boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = 2(1 - \eta^2) = V_e \quad (6.369)$$

**Example 113** Consider a system with equation of motion

$$\mathbf{M}\dot{\boldsymbol{\omega}} + \boldsymbol{\omega}^\times \mathbf{M}\boldsymbol{\omega} = \boldsymbol{\tau} \quad (6.370)$$

where  $\mathbf{M}$  is a constant, symmetric and positive definite matrix, and where the input is selected to be

$$\boldsymbol{\tau} = -\mathbf{K}_d \boldsymbol{\omega} - k_p \boldsymbol{\epsilon} \quad (6.371)$$

where  $\mathbf{K}_d$  is a constant, symmetric and positive definite matrix, and  $k_p$  is a positive constant. The energy function

$$V = \frac{1}{2} \boldsymbol{\omega}^T \mathbf{M} \boldsymbol{\omega} + 2k_p(1 - \eta) \geq 0 \quad (6.372)$$

has time derivative

$$\begin{aligned} \dot{V} &= \boldsymbol{\omega}^T (-\boldsymbol{\omega}^\times \mathbf{M} \boldsymbol{\omega} - \mathbf{K}_d \boldsymbol{\omega} - k_p \boldsymbol{\epsilon}) + k_p \boldsymbol{\epsilon}^T \boldsymbol{\omega} \\ &= -\boldsymbol{\omega}^T \mathbf{K}_d \boldsymbol{\omega} \end{aligned} \quad (6.373)$$

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  $\theta$  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

$$\dot{\theta} = \mathbf{k}^T \boldsymbol{\omega} \quad (6.374)$$

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

$$\frac{1}{2} [\eta \mathbf{I} + \boldsymbol{\epsilon}^\times] \boldsymbol{\omega} = \frac{1}{2} \cos \left( \frac{\theta}{2} \right) \dot{\theta} \mathbf{k} + \sin \frac{\theta}{2} \dot{\mathbf{k}}$$

which gives

$$\begin{aligned} 2 \sin \frac{\theta}{2} \dot{\mathbf{k}} &= \eta (\mathbf{I} - \mathbf{k} \mathbf{k}^T) \boldsymbol{\omega} + \boldsymbol{\epsilon}^\times \boldsymbol{\omega} \\ &= \cos \frac{\theta}{2} (\mathbf{I} - \mathbf{k} \mathbf{k}^T) \boldsymbol{\omega} + \sin \frac{\theta}{2} \mathbf{k}^\times \boldsymbol{\omega} \end{aligned}$$

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

$$\dot{\mathbf{k}} = \frac{1}{2} \left[ \mathbf{k}^\times - \mathbf{k}^\times \mathbf{k}^\times \cot \frac{\theta}{2} \right] \boldsymbol{\omega} \quad (6.375)$$

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

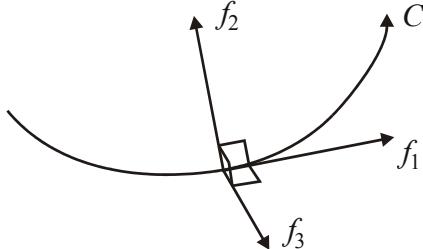


Figure 6.8: The Serret-Frenet frame for a curve  $C$ .

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.

We let the curve  $C$  be given by  $\vec{r}(s)$  where  $s$  is the length along the curve. We define the unit tangent

$$\vec{f}_1 = \frac{i d\vec{r}}{ds} \quad (6.376)$$

The principal unit normal of the curve is defined by the unit vector

$$\vec{f}_2 = \frac{1}{\kappa} \frac{i d\vec{f}_1}{ds} \quad (6.377)$$

where

$$\kappa = \left| \frac{i d\vec{f}_1}{ds} \right| = \left| \frac{i d^2 \vec{r}}{ds^2} \right| \quad (6.378)$$

is the curvature of the curve. Finally the unit binormal vector is defined by the unit vector

$$\vec{f}_3 = \vec{f}_1 \times \vec{f}_2 \quad (6.379)$$

so that  $\vec{f}_1, \vec{f}_2, \vec{f}_3$  forms a set of orthogonal unit vectors. The plane defined by  $\vec{f}_1$  and  $\vec{f}_2$  is called the *osculating plane*, the plane defined by  $\vec{f}_2$  and  $\vec{f}_3$  called the *normal plane*, while the plane defined by  $\vec{f}_3$  and  $\vec{f}_1$  is called the *rectifying plane*.

From  $\vec{f}_3 \cdot \vec{f}_3 = 1$  it follows that

$$0 = \frac{d}{ds} (\vec{f}_3 \cdot \vec{f}_3) = 2 \frac{i d\vec{f}_3}{ds} \cdot \vec{f}_3 \quad (6.380)$$

while (6.377) and  $\vec{f}_3 \cdot \vec{f}_2 = 0$  implies that

$$\vec{f}_3 \cdot \frac{i d\vec{f}_1}{ds} = \vec{f}_3 \cdot \kappa \vec{f}_2 = 0 \quad (6.381)$$

Then, from  $\vec{f}_3 \cdot \vec{f}_1 = 0$  it follows that

$$0 = \frac{d}{ds} (\vec{f}_3 \cdot \vec{f}_1) = \frac{i d\vec{f}_3}{ds} \cdot \vec{f}_1 + \vec{f}_3 \cdot \kappa \vec{f}_2 = \frac{i d\vec{f}_3}{ds} \cdot \vec{f}_1 \quad (6.382)$$

From (6.380) and (6.382) it is seen that  $i d\vec{f}_3/ds$  is along  $\vec{f}_2$ . This makes it possible to write

$$\frac{i d\vec{f}_3}{ds} = \tau \vec{f}_2 \quad (6.383)$$

where

$$\tau = \vec{f}_2 \cdot \frac{i d\vec{f}_3}{ds} \quad (6.384)$$

is the *torsion* of the curve. From  $\vec{f}_2 = \vec{f}_3 \times \vec{f}_1$  we find that

$$\frac{i d\vec{f}_2}{ds} = \frac{i d\vec{f}_3}{ds} \times \vec{f}_1 + \vec{f}_3 \times \frac{i d\vec{f}_1}{ds} = \tau \vec{f}_2 \times \vec{f}_1 + \vec{f}_3 \times \kappa \vec{f}_2 = -\tau \vec{f}_3 - \kappa \vec{f}_1 \quad (6.385)$$

The angular velocity  $\vec{\omega}_{if}$  of the Serret-Frenet frame  $f$  relative to the frame  $i$  is seen from (6.284) to be

$$\vec{\omega}_{if} = \left( \vec{f}_3 \cdot \frac{i d\vec{f}_2}{dt} \right) \vec{f}_1 + \left( \vec{f}_1 \cdot \frac{i d\vec{f}_3}{dt} \right) \vec{f}_2 + \left( \vec{f}_2 \cdot \frac{i d\vec{f}_1}{dt} \right) \vec{f}_3 = \dot{s} (\tau \vec{f}_1 + \kappa \vec{f}_3) \quad (6.386)$$

To sum up the results of this section:

The unit vectors  $\vec{f}_1, \vec{f}_2, \vec{f}_3$  of the Serret-Frenet frame satisfies the kinematic differential equations

$$\frac{i d \vec{f}_1}{ds} = \kappa \vec{f}_2, \quad \frac{i d \vec{f}_2}{ds} = -\kappa \vec{f}_1 - \tau \vec{f}_3, \quad \frac{i d \vec{f}_3}{ds} = -\tau \vec{f}_2 \quad (6.387)$$

where  $\kappa$  is the curvature and  $\tau$  is the torsion of the curve. The angular velocity  $\vec{\omega}_{if}$  of the Serret-Frenet frame  $f$  relative to the frame  $i$ , and the velocity  $\vec{v}_f$  of the origin of frame  $f$  are given by

$$\vec{\omega}_{if} = \dot{s} (\tau \vec{f}_1 + \kappa \vec{f}_3), \quad \vec{v}_f = \dot{s} \vec{f}_1 \quad (6.388)$$

### 6.10.2 Control deviation

Suppose that a curve is to be followed by some body  $b$ , and that the origin of the Serret-Frenet frame  $f$  is placed so that the origin of the body-fixed frame  $b$  has position

$$\vec{r}_{bf} = y \vec{f}_2 + z \vec{f}_3 \quad (6.389)$$

or in other words, the frame  $f$  is placed so that the origin of frame  $b$  is in the normal plane of  $f$ . Then the velocity of the origin of  $b$  is

$$\begin{aligned} \vec{v}_b &= \vec{v}_f + \frac{i d}{dt} \vec{r}_{bf} \\ &= \vec{v}_f + \frac{f d}{dt} \vec{r}_{bf} + \vec{\omega}_{if} \times \vec{r}_{bf} \end{aligned} \quad (6.390)$$

which in coordinate form in frame  $f$  gives

$$\mathbf{v}_b^f = \begin{pmatrix} \dot{s} \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \dot{y} \\ \dot{z} \end{pmatrix} + \begin{pmatrix} 0 & -\dot{s}\kappa & 0 \\ \dot{s}\kappa & 0 & -\dot{s}\tau \\ 0 & \dot{s}\tau & 0 \end{pmatrix} \begin{pmatrix} 0 \\ y \\ z \end{pmatrix} \quad (6.391)$$

This gives the following relation between the time derivatives  $\dot{s}, \dot{y}, \dot{z}$  of the parameters  $s, y, z$  and the body velocity  $\mathbf{v}_b^b$

$$\begin{pmatrix} \dot{s} - \dot{s}\kappa y \\ \dot{y} - \dot{s}\tau z \\ \dot{z} + \dot{s}\tau y \end{pmatrix} = \mathbf{R}_b^f \mathbf{v}_b^b \quad (6.392)$$

while the relative angular velocity between the frames  $b$  and  $f$  is given by

$$\boldsymbol{\omega}_{bf}^f = \boldsymbol{\omega}_{if}^f - \boldsymbol{\omega}_{ib}^f = \begin{pmatrix} \dot{s}\tau \\ 0 \\ \dot{s}\kappa \end{pmatrix} - \mathbf{R}_b^f \boldsymbol{\omega}_{ib}^b \quad (6.393)$$

## 6.11 Navigational kinematics

### 6.11.1 Introduction

Navigation systems rely on GPS navigation using satellite signals, and on the use of inertial navigation (Titterton and Weston 1997), (Farell and Barth 1999). In inertial navigation the position and rotation of a ship, aeroplane or some other vehicle is computed

using measurements from gyroscopes and accelerometers. The accelerometers measure the acceleration, and the gyroscopes measure the angular velocity  $\omega$  of the gyroscopes so the rotation matrix  $\mathbf{R}$  of the accelerometers can be computed by integration of the kinematic differential equation  $\dot{\mathbf{R}} = \mathbf{R}\boldsymbol{\omega}^\times$  provided that the initial rotation of the accelerometers is known. In this section the kinematics of inertial navigation is presented.

### 6.11.2 Coordinate frames

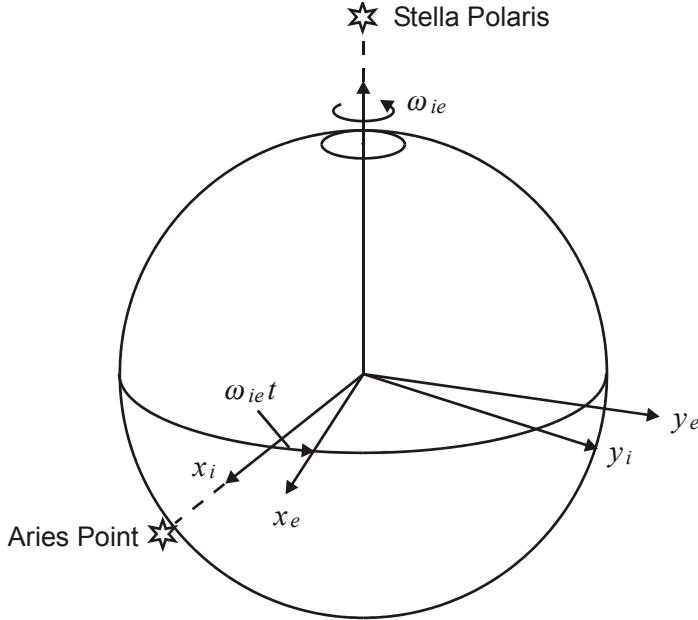


Figure 6.9: The earth-centered star-fixed frame  $i$  and the earth-centered earth-fixed frame  $e$ .

The frame  $i$  is star-fixed and earth-centered, and is considered to be an inertial frame where Newton's law is valid. The frame  $i$  has its origin in the center of the earth, and has axes  $(x_i, y_i, z_i)$  that point towards certain fixed stars. In particular, the  $z_i$  axis points towards Stella Polaris, while the  $x_i$  axis is pointing towards the Aries point in the Vernal Equinox direction. The earth-fixed frame is denoted by  $e$ , and has its origin in the center of the earth, and has axes  $(x_e, y_e, z_e)$  that rotate with the earth. The  $z_e$  axis points towards Stella Polaris through the North Pole. Frames  $i$  and  $e$  are shown in Figure 6.9. The locally horizontal frame or geographic frame is denoted  $n$ . This frame has its origin at some point with latitude  $L$  and longitude  $l$  at the surface of the earth, and has axes  $(N, E, D)$  pointing north, east and down as shown in Figure 6.10. A spherical earth with radius  $r_e$  is assumed. In high precision system the earth is described by an ellipsoid, but to avoid a too complicated presentation this will not be done here.

The rotation of the earth frame  $e$  relative to the inertial frame  $i$  is given by the rotation matrix

$$\mathbf{R}_e^i = \mathbf{R}(z, \omega_{ie}t)$$

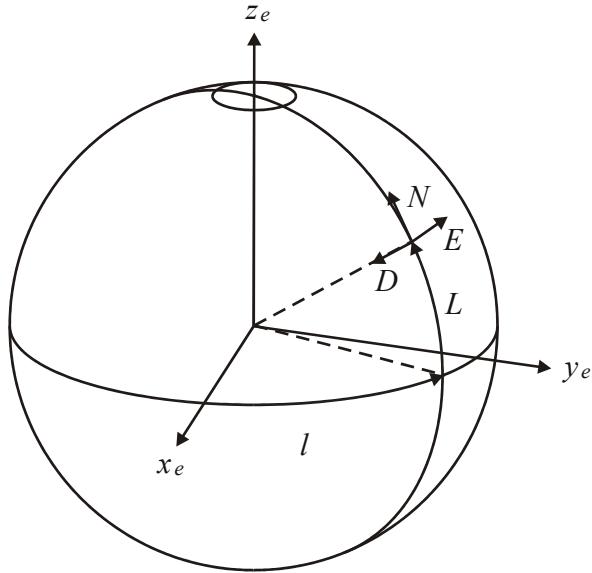


Figure 6.10: The local horizontal frame  $n$  with axes  $N, E, D$  pointing north, east and down.

where the angular velocity is

$$\boldsymbol{\omega}_{ie}^i = \boldsymbol{\omega}_{ie}^e = \begin{pmatrix} 0 \\ 0 \\ \omega_{ie} \end{pmatrix}$$

The scalar  $\omega_{ie}$  corresponds to  $360^\circ$  in 24 hours, which gives  $\omega_{ie} = 7.27 \times 10^{-5}$  rad/s.

The rotation from the inertial frame  $i$  to the geographic frame  $n$  is given by

$$\begin{aligned} \mathbf{R}_n^i &= \mathbf{R}(z, \lambda) \mathbf{R}\left[y, -\left(L + \frac{\pi}{2}\right)\right] \\ &= \begin{pmatrix} -\sin L \cos \lambda & -\sin \lambda & -\cos L \cos \lambda \\ -\sin L \sin \lambda & \cos \lambda & -\cos L \sin \lambda \\ \cos L & 0 & -\sin L \end{pmatrix} \end{aligned}$$

where  $\lambda = l + \omega_{iet}t$ . The angular velocity of frame  $n$  is

$$\boldsymbol{\omega}_{in}^i = \begin{pmatrix} 0 \\ 0 \\ \dot{\lambda} \end{pmatrix} + \mathbf{R}(z, \lambda) \begin{pmatrix} 0 \\ -\dot{L} \\ 0 \end{pmatrix} = \begin{pmatrix} \dot{L} \sin \lambda \\ -\dot{L} \cos \lambda \\ \dot{\lambda} \end{pmatrix}$$

or in the geographic frame

$$\boldsymbol{\omega}_{in}^n = \begin{pmatrix} 0 \\ -\dot{L} \\ 0 \end{pmatrix} + \mathbf{R}(y, 90^\circ) \mathbf{R}(y, L) \begin{pmatrix} 0 \\ 0 \\ \dot{\lambda} \end{pmatrix} = \begin{pmatrix} \dot{\lambda} \cos L \\ -\dot{L} \\ -\dot{\lambda} \sin L \end{pmatrix}$$

### 6.11.3 Acceleration

An accelerometer measures the *specific force*

$$\vec{f} = \frac{i}{dt^2} \vec{r} - \vec{g}_0$$

where  $\vec{g}_0$  is the acceleration of gravity. The gravitational field  $\vec{g}$  of the earth is the combined effect of the acceleration of gravity and the centripetal acceleration due to the rotation of the earth:

$$\vec{g} = \vec{g}_0 - \vec{\omega}_{ie} \times (\vec{\omega}_{ie} \times \vec{r})$$

The velocity

$$\vec{v} := \frac{e}{dt} \vec{r} = \frac{i}{dt} \vec{r} - \vec{\omega}_{ie} \times \vec{r} \quad (6.394)$$

is defined as the time derivative of the position vector in the earth frame. Note that  $\vec{v}$  is not defined as  $\frac{i}{dt} \vec{r}$ . In navigation algorithms the time derivative of the velocity  $\vec{v}$  in the geographic frame is needed. This is the position vector  $\vec{r}$  differentiated first in the  $e$  frame and then in the  $n$  frame. This gives

$$\begin{aligned} \frac{n}{dt} \vec{v} &= \frac{i}{dt} \vec{v} - \vec{\omega}_{in} \times \vec{v} \\ &= \frac{i}{dt} \left( \frac{i}{dt} \vec{r} - \vec{\omega}_{ie} \times \vec{r} \right) - \vec{\omega}_{in} \times \vec{v} \\ &= \frac{i}{dt^2} \vec{r} - \frac{i}{dt} \vec{\omega}_{ie} \times \vec{r} - \vec{\omega}_{ie} \times \frac{i}{dt} \vec{r} - \vec{\omega}_{in} \times \vec{v} \\ &= \frac{i}{dt^2} \vec{r} - (\vec{\omega}_{ie} + \vec{\omega}_{in}) \times \vec{v} - \vec{\omega}_{ie} \times (\vec{\omega}_{ie} \times \vec{r}) \\ &= \frac{i}{dt^2} \vec{r} - (2\vec{\omega}_{ie} + \vec{\omega}_{en}) \times \vec{v} - \vec{\omega}_{ie} \times (\vec{\omega}_{ie} \times \vec{r}) \end{aligned} \quad (6.395)$$

where it is used that  $\vec{\omega}_{ie}$  is constant in the  $i$  frame, and that  $\vec{\omega}_{in} = \vec{\omega}_{ie} + \vec{\omega}_{en}$ . The specific force  $\vec{f}$  is therefore

$$\vec{f} = \frac{n}{dt} \vec{v} + (2\vec{\omega}_{ie} + \vec{\omega}_{en}) \times \vec{v} - \vec{g}$$

In the  $n$  frame the velocity  $\mathbf{v}^n$  is found to be given by  $L, l, h$  and their derivatives:

$$\mathbf{v}^n = \begin{pmatrix} v_N \\ v_E \\ v_D \end{pmatrix} = \begin{pmatrix} (r_e + h) \dot{L} \\ (r_e + h) \dot{l} \cos L \\ -\dot{h} \end{pmatrix}$$

where  $h$  is the height above the surface of the earth. The vector of gravity is assumed to be

$$\mathbf{g}^n = (0, 0, g)^T$$

Then

$$\mathbf{f}^n = \begin{pmatrix} f_N \\ f_E \\ f_D \end{pmatrix} = \begin{pmatrix} \dot{v}_N \\ \dot{v}_E \\ \dot{v}_D \end{pmatrix} + \begin{pmatrix} (l + 2\omega_{ie}) v_E \sin L - \dot{L} v_D \\ -(l + 2\omega_{ie}) (v_N \sin L + v_D \cos L) \\ (l + 2\omega_{ie}) v_E \cos L + \dot{L} v_N \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ g \end{pmatrix} \quad (6.396)$$

As  $v_N, v_E, v_D$  are function of  $L, l, h$  and their derivatives this clearly shows that  $\mathbf{f}^n$  is a function of  $L, l, h$  and their derivatives.

## 6.12 Kinematics of a rigid body

### 6.12.1 Configuration

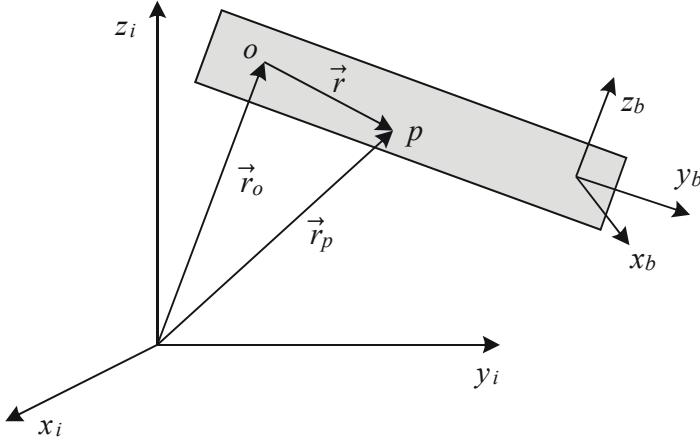


Figure 6.11: Rigid body  $b$  with the fixed frame  $b$  and the fixed points  $o$  and  $p$ .

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

$$\vec{r}_p = \vec{r}_o + \vec{r} \quad (6.397)$$

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

$$\mathbf{r}^i = \mathbf{R}_b^i \mathbf{r}^b \quad (6.398)$$

### 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

$$\vec{v}_o := \frac{^i d}{dt} \vec{r}_o, \quad \vec{v}_p := \frac{^i d}{dt} \vec{r}_p \quad (6.399)$$

From (6.397) and the rule for differentiation in moving frames it is seen that the velocity of  $p$  can be expressed as

$$\vec{v}_p = \vec{v}_o + \frac{^b d}{dt} \vec{r} + \vec{\omega}_{ib} \times \vec{r} \quad (6.400)$$

### 6.12.3 Acceleration

The acceleration vectors are defined by

$$\vec{a}_p := \frac{^i d^2}{dt^2} \vec{r}_p, \quad \vec{a}_o := \frac{^i d^2}{dt^2} \vec{r}_o \quad (6.401)$$

while the angular acceleration vector is defined by

$$\vec{\alpha}_{ib} := \frac{i}{dt} \vec{\omega}_{ib} = \frac{b}{dt} \vec{\omega}_{ib} \quad (6.402)$$

where the second equality is a consequence of

$$\frac{i}{dt} \vec{\omega}_{ib} = \frac{b}{dt} \vec{\omega}_{ib} + \vec{\omega}_{ib} \times \vec{\omega}_{ib} = \frac{b}{dt} \vec{\omega}_{ib} \quad (6.403)$$

In the formulation of the equations of motion for rigid bodies, we need the following result:

$$\begin{aligned} \frac{i}{dt^2} \vec{r}_p &= \frac{i}{dt^2} \vec{r}_o + \frac{i}{dt} \left( \frac{i}{dt} \vec{r} \right) = \frac{i}{dt^2} \vec{r}_o + \frac{i}{dt} \left( \frac{b}{dt} \vec{r} + \vec{\omega}_{ib} \times \vec{r} \right) \\ &= \frac{i}{dt^2} \vec{r}_o + \frac{b}{dt} \left( \frac{b}{dt} \vec{r} + \vec{\omega}_{ib} \times \vec{r} \right) + \vec{\omega}_{ib} \times \left( \frac{b}{dt} \vec{r} + \vec{\omega}_{ib} \times \vec{r} \right) \\ &= \frac{i}{dt^2} \vec{r}_o + \frac{b^2}{dt^2} \vec{r} + 2\vec{\omega}_{ib} \times \frac{b}{dt} \vec{r} + \left( \frac{b}{dt} \vec{\omega}_{ib} \right) \times \vec{r} + \vec{\omega}_{ib} \times (\vec{\omega}_{ib} \times \vec{r}) \end{aligned} \quad (6.404)$$

In terms of acceleration, angular acceleration and velocities this is written

$\underbrace{\vec{a}_p}$ Acceleration of $p$	$=$	$\underbrace{\vec{a}_o}$ Acceleration of $o$	+	$\underbrace{\frac{b}{dt^2} \vec{r}}$ Second derivative of $\vec{r}$ in $b$
		$+ \underbrace{2\vec{\omega}_{ib} \times \frac{b}{dt} \vec{r}}$ Coriolis acceleration	$+ \underbrace{\vec{\alpha}_{ib} \times \vec{r}}$ Transversal acceleration	$+ \underbrace{\vec{\omega}_{ib} \times (\vec{\omega}_{ib} \times \vec{r})}$ Centripetal acceleration
				(6.405)

An alternative formulation is obtained by inserting the expression

$$\vec{a}_o = \frac{i}{dt} \vec{v}_o = \frac{b}{dt} \vec{v}_o + \vec{\omega}_{ib} \times \vec{v}_o \quad (6.406)$$

which gives

$$\vec{a}_p = \frac{b}{dt} \vec{v}_o + \vec{\omega}_{ib} \times \vec{v}_o + \frac{b}{dt^2} \vec{r} + 2\vec{\omega}_{ib} \times \frac{b}{dt} \vec{r} + \vec{\alpha}_{ib} \times \vec{r} + \vec{\omega}_{ib} \times (\vec{\omega}_{ib} \times \vec{r}) \quad (6.407)$$

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  $2\vec{\omega}_{ib} \times \frac{b}{dt} \vec{r}$  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

$$\frac{b}{dt} \vec{r} = \vec{0} \Rightarrow \frac{i}{dt} \vec{r} = \vec{\omega}_{ib} \times \vec{r}, \quad \vec{r} \text{ fixed in } b \quad (6.408)$$

For  $\vec{v}_p$  this gives

$$\vec{v}_p = \vec{v}_o + \vec{\omega}_{ib} \times \vec{r}, \quad \vec{r} \text{ fixed in } b \quad (6.409)$$

The acceleration is found to be

$$\vec{a}_p = \vec{a}_o + \vec{\alpha}_{ib} \times \vec{r} + \vec{\omega}_{ib} \times (\vec{\omega}_{ib} \times \vec{r}), \quad \vec{r} \text{ fixed in } b \quad (6.410)$$

We see that in this case there is no Coriolis acceleration. Using (6.406) the acceleration can be written

$$\vec{a}_p = \frac{^b d}{dt} \vec{v}_o + \vec{\omega}_{ib} \times \vec{v}_o + \vec{\alpha}_{ib} \times \vec{r} + \vec{\omega}_{ib} \times (\vec{\omega}_{ib} \times \vec{r}), \quad \vec{r} \text{ fixed in } b \quad (6.411)$$

## 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

$$m\vec{r}_c = \sum_{k=1}^N m_k \vec{r}_k \quad (6.412)$$

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

$$m\vec{v}_c = m \frac{^i d\vec{r}_c}{dt} = \sum_{k=1}^N m_k \vec{v}_k \quad (6.413)$$

$$m\vec{a}_c = m \frac{^i d^2\vec{r}_c}{dt^2} = \sum_{k=1}^N m_k \vec{a}_k \quad (6.414)$$

### 6.13.2 Rigid body

The position  $\vec{r}_c$  of the center of mass of a rigid body  $b$  is defined by

$$m\vec{r}_c = \int_b \vec{r}_p dm \quad (6.415)$$

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

$$\vec{r}_p = \vec{r}_c + \vec{r} \quad (6.416)$$

as shown in Figure 6.12. From the definition of the center of mass we see that

$$\int_b \vec{r} dm = \int_b \vec{r}_p dm - m\vec{r}_c = \vec{0} \quad (6.417)$$

The velocity  $\vec{v}_c$  of the center of mass is given by

$$m\vec{v}_c = m \frac{^i d\vec{r}_c}{dt} = \frac{^i d}{dt} \int_b \vec{r}_p dm = \int_b \frac{^i d\vec{r}_p}{dt} dm = \int_b \vec{v}_p dm \quad (6.418)$$

while the acceleration  $\vec{a}_c$  of the center of mass is found in the same way. We conclude that

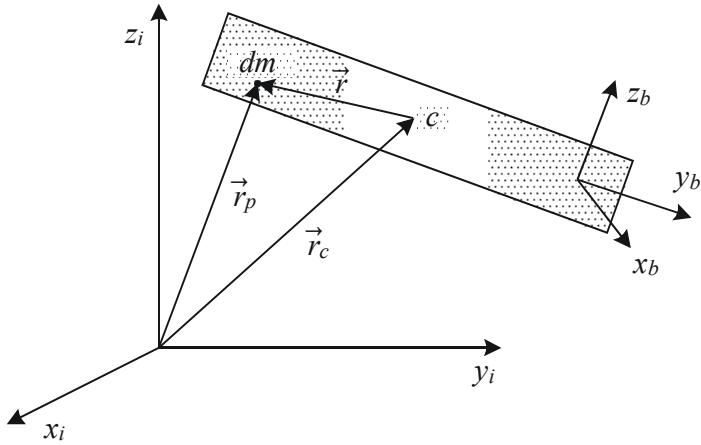


Figure 6.12: Mass element  $dm$  relative to the center of mass  $c$ .

The motion of the mass center in the rigid body  $b$  satisfies the equations

$$m\vec{r}_c = \int_b \vec{r}_p dm, \quad m\vec{v}_c = \int_b \vec{v}_p dm, \quad m\vec{a}_c = \int_b \vec{a}_p dm \quad (6.419)$$

# 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

$$\vec{F}_S^{(r)} = \sum_{j=1}^{n_F} \vec{F}_j \quad (7.1)$$

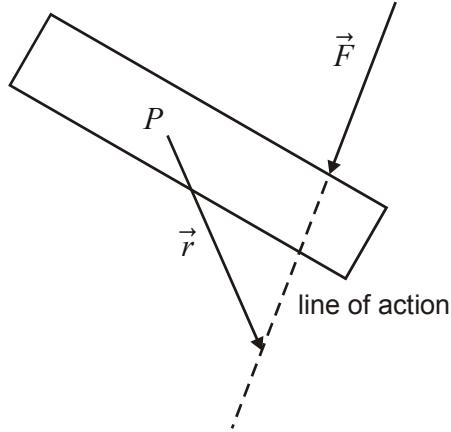


Figure 7.1: A force  $\vec{F}$  acting on a rigid body. The line of action of the force is indicated as a dashed line, and the distance  $\vec{r}$  from a point  $P$  is shown.

while the moment about  $P$  of the set  $S$  of forces is

$$\vec{N}_{S/P} = \sum_{j=1}^{n_F} \vec{r}_{Pj} \times \vec{F}_j \quad (7.2)$$

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

$$\begin{aligned} \vec{N}_{S/Q} &= \sum_{j=1}^{n_F} \vec{r}_{Qj} \times \vec{F}_j = \sum_{j=1}^{n_F} (\vec{r}_{Pj} + \vec{r}_{QP}) \times \vec{F}_j \\ &= \sum_{j=1}^{n_F} \vec{r}_{Pj} \times \vec{F}_j + \vec{r}_{QP} \times \sum_{j=1}^{n_F} \vec{F}_j \end{aligned} \quad (7.3)$$

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  $P$ :

$$\vec{N}_{S/Q} = \vec{N}_{S/P} + \vec{r}_{QP} \times \vec{F}_S^{(r)} \quad (7.4)$$

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

$$\vec{T}_C := \vec{N}_{C/P} = \vec{N}_{C/Q} \quad (7.5)$$

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

$$\vec{T} = \vec{r}_1 \times \vec{F}_1 + \vec{r}_2 \times \vec{F}_2 = \vec{r}_1 \times \vec{F}_1 - (\vec{r}_1 - \vec{r}_{21}) \times \vec{F}_1 = \vec{r}_{21} \times \vec{F}_1 \quad (7.6)$$

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

$$\vec{F}^{(r)} = \sum_{j=1}^6 \vec{F}_j, \quad \vec{N}_c = \sum_{j=1}^3 \vec{T}_j + \sum_{j=1}^6 \vec{r}_j \times \vec{F}_j \quad (7.7)$$

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

$$\vec{T}_c = \sum_{j=1}^3 \vec{T}_j + \sum_{j=1}^3 2\vec{r}_j \times \vec{F}_j \quad (7.8)$$

### 7.2.3 Equivalent force and torque

Two sets  $S$  and  $\Sigma$  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  $\Sigma$  of forces can then be defined with a single force and a torque due to a couple. To do this we let the set  $\Sigma$  to be the force  $\vec{F}_\Sigma$  with line of action through the point  $P$ , and the torque  $\vec{T}_\Sigma$  so that

$$\vec{F}_\Sigma = \vec{F}_S^{(r)}, \quad \vec{T}_\Sigma = \vec{N}_{S/P} \quad (7.9)$$

To see that the sets  $S$  and  $\Sigma$  will be equivalent we observe that the set  $\Sigma$  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,

$$\vec{N}_{\Sigma/Q} = \vec{T}_{\Sigma} + \vec{r}_{QP} \times \vec{F}_{\Sigma} \quad (7.10)$$

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  $\Sigma$  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  $\Sigma$  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

$$\vec{G} = \int_b \vec{g}dm = m\vec{g} \quad (7.11)$$

The moment of the gravity force about the origin of frame  $i$  is

$$\vec{N}_{G/i} = \int_b \vec{r}_p \times \vec{g}dm = \int_b \vec{r}_p dm \times \vec{g} = \vec{r}_c \times m\vec{g} = \vec{r}_c \times \vec{G} \quad (7.12)$$

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

$$\vec{F}_b^{(r)} = \vec{G} + \sum_{j=1}^{n_F} \vec{F}_j \quad (7.13)$$

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

$$\vec{N}_{b/c} = \vec{T}_b + \sum_{j=1}^{n_F} \vec{r}_{cj} \times \vec{F}_j \quad (7.14)$$

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

$$\vec{N}_{b/o} = \vec{N}_{b/c} + \vec{r}_g \times \vec{F}_b^{(r)} \quad (7.15)$$

where

$$\vec{r}_g := \vec{r}_{oc} \quad (7.16)$$

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

$$\vec{F}_{bo} = \vec{F}_{bc} \quad (7.17)$$

$$\vec{T}_{bo} = \vec{T}_{bc} + \vec{r}_g \times \vec{F}_{bc} \quad (7.18)$$

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

$$\vec{N}_{b/o} = \vec{T}_{bc} + \vec{r}_g \times \vec{F}_{bc} \quad (7.19)$$

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

$$\vec{N}_{b/c} = \vec{T}_{bo} - \vec{r}_g \times \vec{F}_{bo} \quad (7.20)$$

### 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  $k$

$$\vec{F}_k^{(r)} = \vec{F}_{k-1,k} + \vec{F}_{k+1,k} + \vec{G}_k \quad (7.21)$$

The moment about the center of mass  $k_c$  with position  $\vec{r}_{k_c}$  is

$$\vec{N}_{k/k_c} = \vec{T}_{k-1,k} + \vec{T}_{k+1,k} + (\vec{r}_{k-1} - \vec{r}_{k_c}) \times \vec{F}_{k-1,k} + (\vec{r}_k - \vec{r}_{k_c}) \times \vec{F}_{k+1,k} \quad (7.22)$$

An equivalent description is possible with the force

$$\vec{F}_{kc} := \vec{F}_k^{(r)} \quad (7.23)$$

with line of action through the center of mass, and the torque

$$\vec{T}_{kc} := \vec{N}_{k/k_c} \quad (7.24)$$

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

$$\vec{N}_{k/k-1} = \vec{T}_{k-1,k} + \vec{T}_{k+1,k} + (\vec{r}_k - \vec{r}_{k-1}) \times \vec{F}_{k+1,k} + (\vec{r}_{k_c} - \vec{r}_{k-1}) \times \vec{G}_k \quad (7.25)$$

## 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)

$$m\vec{a}_c = \vec{F}^{(r)} \quad (7.26)$$

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

$$\vec{h}_{k/c} = (\vec{r}_k - \vec{r}_c) \times m_k \vec{v}_k \quad (7.27)$$

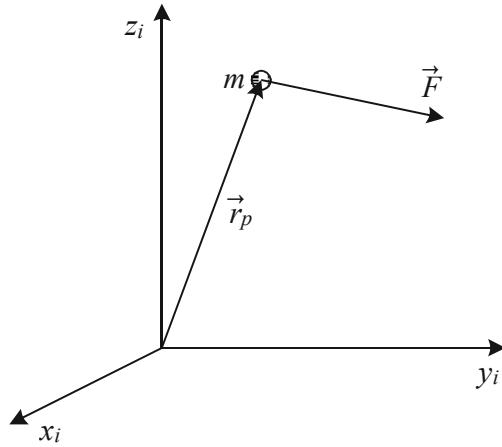


Figure 7.2: Mass point subject to a force  $\vec{F}$ .

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

$$\begin{aligned} \frac{^i d}{dt} \vec{h}_{k/c} &= (\vec{v}_k - \vec{v}_c) \times m_k \vec{v}_k + \vec{r}_{ck} \times m_k \vec{a}_k \\ &= -\vec{v}_c \times m_k \vec{v}_k + \vec{r}_{ck} \times m_k \vec{a}_k \end{aligned} \quad (7.28)$$

Summation over all particles leads to

$$\frac{^i d}{dt} \vec{h}_c = \vec{N}_c \quad (7.29)$$

where (6.413) and (7.26) is used, and where

$$\vec{h}_c = \sum_{k=1}^N \vec{r}_{ck} \times m_k \vec{v}_k \quad (7.30)$$

is the angular momentum of the system about the center of mass, and

$$\vec{N}_c = \sum_{k=1}^N \vec{r}_{ck} \times \vec{F}_k \quad (7.31)$$

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

$$\vec{F}_{bc} = m\vec{a}_c \quad (7.32)$$

$$\vec{T}_{bc} = \frac{i}{dt}\vec{h}_{b/c} \quad (7.33)$$

where

$$\vec{h}_{b/c} = \int_b \vec{r} \times \vec{v}_p dm \quad (7.34)$$

is the angular momentum of the body  $b$  about the center of mass, and

$$\vec{r} = \vec{r}_p - \vec{r}_c \quad (7.35)$$

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

$$\begin{aligned} \vec{h}_{b/c} &= \int_b \vec{r} dm \times \vec{v}_c + \int_b \vec{r} \times (\vec{\omega}_{ib} \times \vec{r}) dm \\ &= \int_b \vec{r} \times (\vec{\omega}_{ib} \times \vec{r}) dm \\ &= - \int_b \vec{r} \times (\vec{r} \times \vec{\omega}_{ib}) dm \end{aligned} \quad (7.36)$$

where we have used (6.417). By introducing the dyadic representation of the vector cross product we may write this in the form

$$\vec{h}_{b/c} = - \int_b \vec{r}^\times \cdot (\vec{r}^\times \cdot \vec{\omega}_{ib}) dm = - \int_b \vec{r}^\times \cdot \vec{r}^\times dm \cdot \vec{\omega}_{ib} \quad (7.37)$$

This expression motivates the definition of the *inertia dyadic* of  $b$  about  $c$  as

$$\vec{M}_{b/c} = - \int_b \vec{r}^\times \cdot \vec{r}^\times dm \quad (7.38)$$

The angular momentum about  $c$  can then be written

$$\vec{h}_{b/c} = \vec{M}_{b/c} \cdot \vec{\omega}_{ib} \quad (7.39)$$

Insertion in (7.33) gives

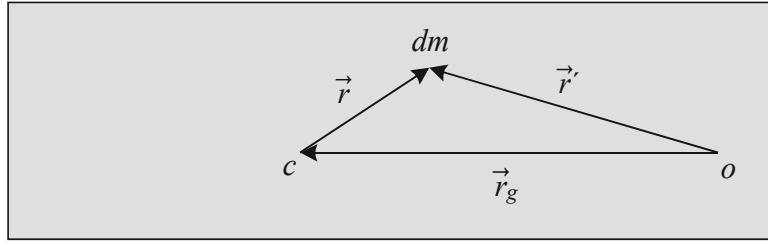
$$\vec{T}_{bc} = \frac{i}{dt} (\vec{M}_{b/c} \cdot \vec{\omega}_{ib}) = \frac{b}{dt} (\vec{M}_{b/c} \cdot \vec{\omega}_{ib}) + \vec{\omega}_{ib} \times (\vec{M}_{b/c} \cdot \vec{\omega}_{ib}) \quad (7.40)$$

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

$$\vec{F}_{bc} = m\vec{a}_c \quad (7.41)$$

$$\vec{T}_{bc} = \vec{M}_{b/c} \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_{b/c} \cdot \vec{\omega}_{ib}) \quad (7.42)$$

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

$$\vec{T}_{bo} = \vec{T}_{bc} + \vec{r}_g \times \vec{F}_{bc} \quad (7.43)$$

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

$$\vec{F}_{bo} = m\vec{a}_c \quad (7.44)$$

$$\vec{T}_{bo} = \vec{r}_g \times m\vec{a}_c + \vec{M}_{b/c} \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_{b/c} \cdot \vec{\omega}_{ib}) \quad (7.45)$$

Then the force equation can be referenced to the point  $o$  by combining (7.32) and (6.410). This gives

$$\vec{F}_{bo} = m [\vec{a}_o + \vec{\alpha}_{ib} \times \vec{r}_g + \vec{\omega}_{ib} \times (\vec{\omega}_{ib} \times \vec{r}_g)] . \quad (7.46)$$

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

$$\vec{h}_{b/o} = \int_b \vec{r}' \times \vec{v}_p dm \quad (7.47)$$

where

$$\vec{r}_p = \vec{r}_o + \vec{r}' \quad (7.48)$$

$$\vec{v}_p = \vec{v}_o + \vec{\omega}_{ib} \times \vec{r}' \quad (7.49)$$

Insertion of (7.49) into (7.47) leads to

$$\begin{aligned}\vec{h}_{b/o} &= \int_b \vec{r}' dm \times \vec{v}_o + \int_b \vec{r}' \times (\vec{\omega}_{ib} \times \vec{r}') dm \\ &= \int_b (\vec{r}_p - \vec{r}_o) dm \times \vec{v}_o - \int_b \vec{r}' \times (\vec{r}' \times \vec{\omega}_{ib}) dm \\ &= \vec{r}_g \times m\vec{v}_o + \vec{M}_{b/o} \cdot \vec{\omega}_{ib}\end{aligned}\quad (7.50)$$

where the inertia dyadic  $\vec{M}_{b/o}$  of  $b$  about  $o$  is defined by

$$\vec{M}_{b/o} = - \int_b (\vec{r}')^\times \cdot (\vec{r}')^\times dm \quad (7.51)$$

Note that  $\vec{M}_{b/o}$  is constant in frame  $b$ . Time differentiation with respect to reference to frame  $i$  gives

$$\begin{aligned}\frac{i}{dt} \vec{h}_{b/o} &= (\vec{v}_c - \vec{v}_o) \times m\vec{v}_o + \vec{r}_g \times m\vec{a}_o + \frac{b}{dt} (\vec{M}_{b/o} \cdot \vec{\omega}_{ib}) + \vec{\omega}_{ib} \times (\vec{M}_{b/o} \cdot \vec{\omega}_{ib}) \\ &= \vec{v}_c \times m\vec{v}_o + \vec{r}_g \times m\vec{a}_o + \vec{M}_{b/o} \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_{b/o} \cdot \vec{\omega}_{ib})\end{aligned}\quad (7.52)$$

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

$$\vec{h}_{b/o} = \int_b (\vec{r} + \vec{r}_g) \times \vec{v}_p dm = \vec{h}_{b/c} + \int_b \vec{r}_g \times \vec{v}_p dm \quad (7.53)$$

$$= \vec{h}_{b/c} + \vec{r}_g \times m\vec{v}_c \quad (7.54)$$

which implies that

$$\frac{i}{dt} \vec{h}_{b/o} = \frac{i}{dt} \vec{h}_{b/c} + \vec{r}_g \times m\vec{a}_c - \vec{v}_o \times m\vec{v}_c \quad (7.55)$$

From this equation and (7.52) it follows that

$$\frac{i}{dt} \vec{h}_{b/c} = \vec{r}_g \times m(\vec{a}_o - \vec{a}_c) + \vec{M}_{b/o} \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_{b/o} \cdot \vec{\omega}_{ib}) \quad (7.56)$$

This result in combination with equations (7.32), (7.33) and (7.43) gives

$$\vec{T}_{bo} = \vec{r}_g \times m\vec{a}_o + \vec{M}_{b/o} \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_{b/o} \cdot \vec{\omega}_{ib}) \quad (7.57)$$

With reference to a point  $o$  the equation of motion for a rigid body can be written

$$\vec{F}_{bo} = m[\vec{a}_o + \vec{\alpha}_{ib} \times \vec{r}_g + \vec{\omega}_{ib} \times (\vec{\omega}_{ib} \times \vec{r}_g)] \quad (7.58)$$

$$\vec{T}_{bo} = \vec{r}_g \times m\vec{a}_o + \vec{M}_{b/o} \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_{b/o} \cdot \vec{\omega}_{ib}) \quad (7.59)$$

### 7.3.4 The inertia dyadic

The inertia dyadic of  $b$  about the center of the mass was defined as

$$\vec{M}_{b/c} = - \int_b \vec{r}^\times \cdot \vec{r}^\times dm \quad (7.60)$$

From (6.73) the alternative expression

$$\vec{M}_{b/c} = \int_b (\vec{r}^2 \vec{I} - \vec{r}\vec{r}) dm \quad (7.61)$$

is found. The dyadic can be evaluated in the  $b$  frame and is written

$$\vec{M}_{b/c} = \sum_{i=1}^3 \sum_{j=1}^3 m_{ij}^b \vec{b}_i \vec{b}_j \quad (7.62)$$

**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

$$\vec{M}_{b/c} = m_{11} \vec{b}_1 \vec{b}_1 + m_{22} \vec{b}_2 \vec{b}_2 + m_{33} \vec{b}_3 \vec{b}_3. \quad (7.63)$$

where  $m_{11}$ ,  $m_{22}$  and  $m_{33}$  are constants. The angular velocity is written

$$\vec{\omega}_{ib} = \omega_1 \vec{b}_1 + \omega_2 \vec{b}_2 + \omega_3 \vec{b}_3 \quad (7.64)$$

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

$$\vec{h}_{b/c} = (m_{11} \vec{b}_1 \vec{b}_1 + m_{22} \vec{b}_2 \vec{b}_2 + m_{33} \vec{b}_3 \vec{b}_3) \cdot (\omega_1 \vec{b}_1 + \omega_2 \vec{b}_2 + \omega_3 \vec{b}_3). \quad (7.65)$$

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

$$\vec{h}_{b/c} = m_{11} \omega_1 \vec{b}_1 + m_{22} \omega_2 \vec{b}_2 + m_{33} \omega_3 \vec{b}_3 \quad (7.66)$$

**Example 118** The kinetic energy of a rigid body is

$$K = \frac{1}{2} \int_b \vec{v}_p \cdot \vec{v}_p dm. \quad (7.67)$$

Insertion of  $\vec{v}_p = \vec{v}_c + \vec{\omega}_{ib} \times \vec{r}$  gives

$$K = \frac{1}{2} m \vec{v}_c^2 + \frac{1}{2} \int_b (\vec{\omega}_{ib} \times \vec{r}) \cdot (\vec{\omega}_{ib} \times \vec{r}) dm \quad (7.68)$$

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

$$\begin{aligned} \frac{1}{2} \int_b (\vec{\omega}_{ib} \times \vec{r}) \cdot (\vec{\omega}_{ib} \times \vec{r}) dm &= -\frac{1}{2} \int_b (\vec{\omega}_{ib} \times \vec{r}) \cdot (\vec{r} \times \vec{\omega}_{ib}) dm \\ &= -\frac{1}{2} \int_b \vec{\omega}_{ib} \cdot \vec{r}^\times \cdot \vec{r}^\times \cdot \vec{\omega}_{ib} dm \\ &= -\frac{1}{2} \vec{\omega}_{ib} \cdot \int_b \vec{r}^\times \cdot \vec{r}^\times dm \cdot \vec{\omega}_{ib} \\ &= \frac{1}{2} \vec{\omega}_{ib} \cdot \vec{M}_{b/c} \cdot \vec{\omega}_{ib} \end{aligned} \quad (7.69)$$

This leads to the following expression for the kinetic energy:

$$K = \frac{1}{2} m \vec{v}_c^2 + \frac{1}{2} \vec{\omega}_{ib} \cdot \vec{M}_{b/c} \cdot \vec{\omega}_{ib} \quad (7.70)$$

**Example 119** Using (7.49) the kinetic energy is found from the computation

$$\begin{aligned} K &= \frac{1}{2} \int_b \vec{v}_p \cdot \vec{v}_p dm \\ &= \frac{1}{2} m \vec{v}_o \cdot \vec{v}_o + \vec{v}_o \cdot \left( \vec{\omega}_{ib} \times \int_b \vec{r}' dm \right) + \frac{1}{2} \int_b (\vec{\omega}_{ib} \times \vec{r}') \cdot (\vec{\omega}_{ib} \times \vec{r}') dm \\ &= \frac{1}{2} m \vec{v}_o \cdot \vec{v}_o + \vec{v}_o \cdot (\vec{\omega}_{ib} \times m \vec{r}_g) + \frac{1}{2} \vec{\omega}_{ib} \cdot \vec{M}_{b/o} \cdot \vec{\omega}_{ib} \end{aligned} \quad (7.71)$$

to have the form

$$K = \frac{1}{2} m \vec{v}_o \cdot \vec{v}_o - \vec{v}_o \cdot m \vec{r}_g^\times \cdot \vec{\omega}_{ib} + \frac{1}{2} \vec{\omega}_{ib} \cdot \vec{M}_{b/o} \cdot \vec{\omega}_{ib} \quad (7.72)$$

### 7.3.5 The inertia matrix

The matrix representation of the inertia dyadic  $\vec{M}_{b/c}$  in frame  $b$  is the *inertia matrix*

$$\mathbf{M}_{b/c}^b = - \int_b (\mathbf{r}^b)^\times (\mathbf{r}^b)^\times dm \quad (7.73)$$

From (6.25) the more usual expression

$$\mathbf{M}_{b/c}^b = \int_b \left[ (\mathbf{r}^b)^T \mathbf{r}^b \mathbf{I} - \mathbf{r}^b (\mathbf{r}^b)^T \right] dm \quad (7.74)$$

is found. The angular momentum of  $b$  about  $c$  can then be written in column vector form as

$$\mathbf{h}_{b/c}^b = \mathbf{M}_{b/c}^b \boldsymbol{\omega}_{ib}^b \quad (7.75)$$

Equation (7.75) can be transformed to frame  $i$  using the transformation rule:

$$\mathbf{R}_b^i \mathbf{h}_{ib}^b = \mathbf{R}_b^i \mathbf{M}_{b/c}^b \boldsymbol{\omega}_{ib}^b \quad (7.76)$$

Insertion of  $\boldsymbol{\omega}_{ib}^b = \mathbf{R}_b^i \boldsymbol{\omega}_{ib}^i$  gives

$$\mathbf{h}_{ib}^i = \mathbf{R}_b^i \mathbf{M}_{b/c}^b \mathbf{R}_b^i \boldsymbol{\omega}_{ib}^i \quad (7.77)$$

Moreover, the inertia dyadic can also be represented by a matrix frame  $i$  using

$$\mathbf{M}_{b/c}^i = \int_b \left[ (\mathbf{r}^i)^T \mathbf{r}^i \mathbf{I} - \mathbf{r}^i (\mathbf{r}^i)^T \right] dm \quad (7.78)$$

which satisfies

$$\mathbf{h}_{b/c}^i = \mathbf{M}_{b/c}^i \boldsymbol{\omega}_{ib}^i \quad (7.79)$$

Comparison of the two expressions (7.77) and (7.79) leads to the conclusion

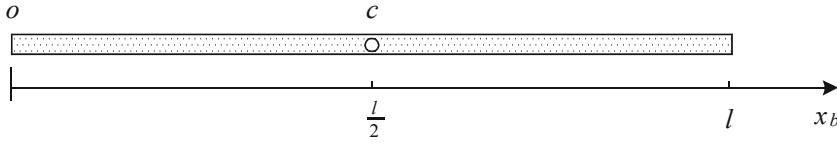
$$\mathbf{M}_{b/c}^i = \mathbf{R}_b^i \mathbf{M}_{b/c}^b \mathbf{R}_i^b \quad (7.80)$$

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

$$K = \frac{1}{2} m (\mathbf{v}_c^b)^T \mathbf{v}_c^b + \frac{1}{2} (\boldsymbol{\omega}_{ib}^b)^T \mathbf{M}_{b/c}^b \boldsymbol{\omega}_{ib}^b \quad (7.81)$$

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

$$\mathbf{M}_{b/c}^b = \int_b [(\mathbf{r}^b)^2 \mathbf{I} - \mathbf{r}^b (\mathbf{r}^b)^T] dm \quad (7.82)$$

in the body-fixed frame  $b$ . Let  $\mathbf{r}^b = (x, y, z)^T$ . The the inertia matrix is then found to be

$$\mathbf{M}_{b/c}^b = \int_b \begin{pmatrix} y^2 + z^2 & -xy & -xz \\ -xy & x^2 + z^2 & -yz \\ -xz & -yz & x^2 + y^2 \end{pmatrix} dm \quad (7.83)$$

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

$$\vec{M}_{b/o} = \vec{M}_{b/c} - m \vec{r}_g^\times \vec{r}_g^\times = \vec{M}_{b/c} + m \left[ (\vec{r}_g \cdot \vec{r}_g) \vec{I} - \vec{r}_g \vec{r}_g \right] \quad (7.84)$$

The corresponding matrix expressions is

$$\mathbf{M}_{b/o}^b = \mathbf{M}_{b/c}^b - m (\mathbf{r}_g^b)^\times (\mathbf{r}_g^b)^\times = \mathbf{M}_{b/c}^b + m [(\mathbf{r}_g^b)^2 \mathbf{I} - \mathbf{r}_g^b (\mathbf{r}_g^b)^T] \quad (7.85)$$

This is *the parallel axes theorem*.

**Example 121** A slender beam has length  $\ell$  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

$$\mathbf{M}_{b/c}^b = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{m\ell^2}{12} & 0 \\ 0 & 0 & \frac{m\ell^2}{12} \end{pmatrix} \quad (7.86)$$

The inertia matrix about the endpoint  $o$  of the beam is found from the parallel axes theorem to be

$$\mathbf{M}_{b/o}^b = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{m\ell^2}{12} & 0 \\ 0 & 0 & \frac{m\ell^2}{12} \end{pmatrix} + m \left(\frac{\ell}{2}\right)^2 \mathbf{I} - \begin{pmatrix} m\left(\frac{\ell}{2}\right)^2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{m\ell^2}{3} & 0 \\ 0 & 0 & \frac{m\ell^2}{3} \end{pmatrix}$$

**Example 122** The kinetic energy of a rigid body can be expressed with reference to a point  $o$  by

$$\begin{aligned} K &= \frac{1}{2}m(\mathbf{v}_c^b)^T \mathbf{v}_c^b + \frac{1}{2}(\boldsymbol{\omega}_{ib}^b)^T \mathbf{M}_{b/c}^b \boldsymbol{\omega}_{ib}^b \\ &= \frac{1}{2}m(\mathbf{v}_o^b + (\boldsymbol{\omega}_{ib}^b)^\times \mathbf{r}_g^b)^T (\mathbf{v}_o^b + (\boldsymbol{\omega}_{ib}^b)^\times \mathbf{r}_g^b) + \frac{1}{2}(\boldsymbol{\omega}_{ib}^b)^T \mathbf{M}_{b/c}^b \boldsymbol{\omega}_{ib}^b \\ &= \frac{1}{2}m(\mathbf{v}_c^b)^T \mathbf{v}_c^b + m(\mathbf{v}_o^b)^T (\mathbf{r}_g^b)^\times T \boldsymbol{\omega}_{ib}^b + m((\boldsymbol{\omega}_{ib}^b)^\times \mathbf{r}_g^b)^T \mathbf{v}_o^b \\ &\quad + \frac{1}{2}(\boldsymbol{\omega}_{ib}^b)^T (\mathbf{M}_{b/c}^b - m(\mathbf{r}_g^b)^\times (\mathbf{r}_g^b)^\times) \boldsymbol{\omega}_{ib}^b \\ &= \frac{1}{2} \begin{pmatrix} \mathbf{v}_o^b \\ \boldsymbol{\omega}_{ib}^b \end{pmatrix}^T \begin{pmatrix} m\mathbf{I} & m(\mathbf{r}_g^b)^\times T \\ m(\mathbf{r}_g^b)^\times & \mathbf{M}_{b/o}^b \end{pmatrix} \begin{pmatrix} \mathbf{v}_o^b \\ \boldsymbol{\omega}_{ib}^b \end{pmatrix} \end{aligned} \quad (7.87)$$

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  $\vec{F}_b^{(r)}$  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  $\vec{F}_{bc} = \vec{F}_b^{(r)}$  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

$$\vec{F}_{bc} = m\vec{a}_c \quad (7.88)$$

$$\vec{T}_{bc} = \vec{M}_{b/c} \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_{b/c} \cdot \vec{\omega}_{ib}) \quad (7.89)$$

In the  $b$  frame the coordinate form is written in matrix form as

$$\begin{pmatrix} m\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{b/c}^b \end{pmatrix} \begin{pmatrix} \mathbf{a}_c^b \\ \boldsymbol{\alpha}_{ib}^b \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ (\boldsymbol{\omega}_{ib}^b)^\times \mathbf{M}_{b/c}^b \boldsymbol{\omega}_{ib}^b \end{pmatrix} = \begin{pmatrix} \mathbf{F}_{bc}^b \\ \mathbf{T}_{bc}^b \end{pmatrix} \quad (7.90)$$

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

$$\vec{F}_{bc} = m \frac{^b d}{dt} \vec{v}_c + m \vec{\omega}_{ib} \times \vec{v}_c \quad (7.91)$$

$$\vec{T}_{bc} = \vec{M}_{b/c} \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_{b/c} \cdot \vec{\omega}_{ib}) \quad (7.92)$$

or

$$\begin{pmatrix} m\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{b/c}^b \end{pmatrix} \begin{pmatrix} \dot{\mathbf{v}}_c^b \\ \boldsymbol{\alpha}_{ib}^b \end{pmatrix} + \begin{pmatrix} m(\boldsymbol{\omega}_{ib}^b)^\times \mathbf{v}_c^b \\ (\boldsymbol{\omega}_{ib}^b)^\times \mathbf{M}_{b/c}^b \boldsymbol{\omega}_{ib}^b \end{pmatrix} = \begin{pmatrix} \mathbf{F}_{bc}^b \\ \mathbf{T}_{bc}^b \end{pmatrix} \quad (7.93)$$

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

$$\vec{F}_{bo} = m[\vec{a}_o + \vec{\alpha}_{ib} \times \vec{r}_g + \vec{\omega}_{ib} \times (\vec{\omega}_{ib} \times \vec{r}_g)] \quad (7.94)$$

$$\vec{T}_{bo} = \vec{r}_g \times m\vec{a}_o + \vec{M}_{b/o} \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_{b/o} \cdot \vec{\omega}_{ib}) \quad (7.95)$$

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

$$\begin{pmatrix} m\mathbf{I} & m(\mathbf{r}_g^b)^\times T \\ m(\mathbf{r}_g^b)^\times & \mathbf{M}_{b/o}^b \end{pmatrix} \begin{pmatrix} \mathbf{a}_o^b \\ \boldsymbol{\alpha}_{ib}^b \end{pmatrix} + \begin{pmatrix} m(\boldsymbol{\omega}_{ib}^b)^\times (\boldsymbol{\omega}_{ib}^b)^\times \mathbf{r}_g^b \\ (\boldsymbol{\omega}_{ib}^b)^\times \mathbf{M}_{b/o}^b \boldsymbol{\omega}_{ib}^b \end{pmatrix} = \begin{pmatrix} \mathbf{F}_{bo}^b \\ \mathbf{T}_{bo}^b \end{pmatrix} \quad (7.96)$$

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  $(\cdot)^\times = -[(\cdot)^\times]^T$ . 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

$$\vec{F}_{bo} = m \left[ \frac{^b d}{dt} \vec{v}_o + \vec{\omega}_{ib} \times \vec{v}_o + \vec{\alpha}_{ib} \times \vec{r}_g + \vec{\omega}_{ib} \times (\vec{\omega}_{ib} \times \vec{r}_g) \right] \quad (7.97)$$

$$\vec{T}_{bo} = m\vec{r}_g \times \frac{^b d}{dt} \vec{v}_o + m\vec{r}_g \times (\vec{\omega}_{ib} \times \vec{v}_o) + \vec{M}_{b/o} \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_{b/o} \cdot \vec{\omega}_{ib}) \quad (7.98)$$

with matrix form

$$\begin{pmatrix} m\mathbf{I} & m(\mathbf{r}_g^b)^\times T \\ m(\mathbf{r}_g^b)^\times & \mathbf{M}_{b/o}^b \end{pmatrix} \begin{pmatrix} \dot{\mathbf{v}}_o^b \\ \boldsymbol{\alpha}_{ib}^b \end{pmatrix} + \begin{pmatrix} m(\boldsymbol{\omega}_{ib}^b)^\times [(\boldsymbol{\omega}_{ib}^b)^\times \mathbf{r}_g^b + \mathbf{v}_o^b] \\ (\boldsymbol{\omega}_{ib}^b)^\times \mathbf{M}_{b/o}^b \boldsymbol{\omega}_{ib}^b + m(\mathbf{r}_g^b)^\times (\boldsymbol{\omega}_{ib}^b)^\times \mathbf{v}_o^b \end{pmatrix} = \begin{pmatrix} \mathbf{F}_{bo}^b \\ \mathbf{T}_{bo}^b \end{pmatrix}$$

### 7.3.9 Satellite attitude dynamics

Suppose that the inertia matrix in the body-fixed frame  $b$  is

$$\mathbf{M}_{b/c}^b = \text{diag}(m_{11}, m_{22}, m_{33}). \quad (7.99)$$

Then the angular momentum is

$$\mathbf{h}_{b/c}^b = \mathbf{M}_{b/c}^b \boldsymbol{\omega}_{ib}^b = \begin{pmatrix} m_{11}\omega_1 \\ m_{22}\omega_2 \\ m_{33}\omega_3 \end{pmatrix}. \quad (7.100)$$

The torque  $\mathbf{T}_{bc}^b = (T_1, T_2, T_3)^T$  is acting on the body. The torque law is then

$$\vec{T}_{bc} = \frac{^i d}{dt} \vec{h}_{b/c} = \frac{^i d}{dt} (\vec{M}_{b/c} \cdot \vec{\omega}_{ib}) = \frac{^b d}{dt} (\vec{M}_{b/c} \cdot \vec{\omega}_{ib}) + \vec{\omega}_{ib} \times (\vec{M}_{b/c} \cdot \vec{\omega}_{ib}) \quad (7.101)$$

The inertia dyadic is constant in the  $b$  frame, therefore

$$\vec{M}_{b/c} \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_{b/c} \cdot \vec{\omega}_{ib}) = \vec{T}_{bc} \quad (7.102)$$

where (6.402) is used. With coordinate vectors this is written

$$\mathbf{M}_{b/c}^b \dot{\omega}_{ib}^b + (\omega_{ib}^b)^\times \mathbf{M}_{b/c}^b \omega_{ib}^b = \mathbf{T}_{bc}^b. \quad (7.103)$$

Written out in components the model is

$$m_{11}\dot{\omega}_1 + (m_{33} - m_{22})\omega_2\omega_3 = T_1 \quad (7.104)$$

$$m_{22}\dot{\omega}_2 + (m_{11} - m_{33})\omega_3\omega_1 = T_2 \quad (7.105)$$

$$m_{33}\dot{\omega}_3 + (m_{22} - m_{11})\omega_1\omega_2 = T_3 \quad (7.106)$$

## 7.4 Example: Ball and beam dynamics

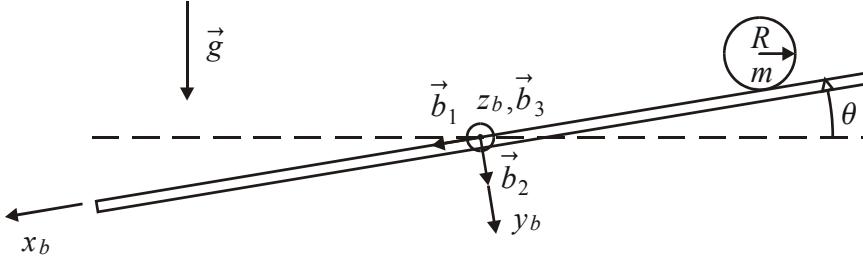


Figure 7.5: Ball and beam system.

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

$$\vec{n}_1 \cdot \vec{b}_1 = \cos \theta, \quad \vec{n}_1 \cdot \vec{b}_2 = -\sin \theta \quad (7.107)$$

$$\vec{n}_2 \cdot \vec{b}_1 = \sin \theta, \quad \vec{n}_2 \cdot \vec{b}_2 = \cos \theta \quad (7.108)$$

$$\vec{n}_3 \cdot \vec{b}_1 = \vec{n}_3 \cdot \vec{b}_2 = 0, \quad \vec{n}_3 \cdot \vec{b}_3 = 1 \quad (7.109)$$

which implies that

$$\vec{n}_1 = \cos \theta \vec{b}_1 - \sin \theta \vec{b}_2 \quad (7.110)$$

$$\vec{n}_2 = \sin \theta \vec{b}_1 + \cos \theta \vec{b}_2, \quad (7.111)$$

$$\vec{n}_3 = \vec{b}_3 \quad (7.112)$$

We note that

$$\vec{n}_1 \times \vec{b}_1 = \sin \theta \vec{b}_3, \quad \vec{n}_1 \times \vec{b}_2 = \cos \theta \vec{b}_3 \quad (7.113)$$

$$\vec{n}_2 \times \vec{b}_1 = -\cos \theta \vec{b}_3, \quad \vec{n}_2 \times \vec{b}_2 = \sin \theta \vec{b}_3 \quad (7.114)$$

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  $\theta$  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

$$\vec{v}_2 = \frac{^b d}{dt} \vec{r}_2 + \vec{\omega}_1 \times \vec{r} = \dot{x}\vec{b}_1 + \dot{\theta}\vec{b}_3 \times (x\vec{b}_1 - R\vec{b}_2) = (\dot{x} + \dot{\theta}R)\vec{b}_1 + \dot{\theta}x\vec{b}_2, \quad (7.115)$$

and the acceleration is

$$\begin{aligned} \vec{a}_2 &= \frac{^b d}{dt} \vec{v}_2 + \vec{\omega}_1 \times \vec{v}_2 \\ &= (\ddot{x} + \ddot{\theta}R)\vec{b}_1 + (\ddot{\theta}x + \dot{\theta}\dot{x})\vec{b}_2 + \dot{\theta}\vec{b}_3 \times [(\dot{x} + \dot{\theta}R)\vec{b}_1 + \dot{\theta}x\vec{b}_2] \\ &= (\ddot{x} + \ddot{\theta}R - \dot{\theta}^2 x)\vec{b}_1 + (\ddot{\theta}x + 2\dot{\theta}\dot{x} + \dot{\theta}^2 R)\vec{b}_2. \end{aligned} \quad (7.116)$$

It follows that the ball rolls along the track with an angular velocity given by

$$\vec{\omega}_2 = \left( \dot{\theta} + \frac{\dot{x}}{R} \right) \vec{b}_3 \quad (7.117)$$

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

$$J_2 = \frac{2}{5}m_2 R^2 \quad (7.118)$$

which is tabulated in textbooks on dynamics. The contact force acting from the beam on the ball is

$$\vec{F} = F_x \vec{b}_1 + F_y \vec{b}_2 \quad (7.119)$$

while the gravitational force on the ball is

$$\vec{G} = m_2 g \vec{n}_2 = m_2 g (\sin \theta \vec{b}_1 + \cos \theta \vec{b}_2). \quad (7.120)$$

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)

$$\vec{T}_{2c} = J_2 \frac{^n d}{dt} \vec{\omega}_2 = J_2 \left( \ddot{\theta} + \frac{\ddot{x}}{R} \right) \vec{b}_3 \quad (7.121)$$

It is convenient to use the moment

$$\vec{N}_{2/o} = \left( -R\vec{b}_2 \right) \times \vec{G}_2 = m_2 R g \sin \theta \vec{b}_3 \quad (7.122)$$

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

$$\vec{N}_{2/o} = \vec{T}_{2c} + \left( -R\vec{b}_2 \right) \times m_2 \vec{a} \quad (7.123)$$

which gives

$$m_2 R g \sin \theta = J_2 \left( \ddot{\theta} + \frac{\ddot{x}}{R} \right) + m_2 R \left( \ddot{x} + \ddot{\theta} R - \dot{\theta}^2 x \right) \quad (7.124)$$

$$= (J_2 + m_2 R^2) \ddot{\theta} + \left( \frac{J_2}{R} + m_2 R \right) \ddot{x} - m_2 R x \dot{\theta}^2 \quad (7.125)$$

This is written

$$(J_2 + m_2 R^2) \ddot{\theta} + \frac{1}{R} (J_2 + m_2 R^2) \ddot{x} = m_2 R x \dot{\theta}^2 + R m_2 g \sin \theta \quad (7.126)$$

By inserting the value of  $J_2$  from (7.118), we get

$$(J_2 + m_2 R^2) = \frac{7}{5} m_2 R^2 \quad (7.127)$$

The Newton's law for the ball is

$$m_2 \vec{a}_2 = \vec{F} + \vec{G}_2 \quad (7.128)$$

In the  $y_b$  direction this gives

$$m_2 \left( \ddot{\theta} x + 2\dot{\theta} \dot{x} + \dot{\theta}^2 R \right) = F_y + m_2 g \cos \theta \quad (7.129)$$

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

$$J_1 \ddot{\theta} \vec{b}_3 = x \vec{b}_1 \times (-F_y \vec{b}_2) + T \vec{b}_3 \quad (7.130)$$

which leads to

$$J_1 \ddot{\theta} = -x F_y + T \quad (7.131)$$

This equation is combined with (7.129), and the result is

$$(J_1 + m_2 x^2) \ddot{\theta} = T + m_2 g x \cos \theta - 2m_2 x \dot{\theta} \dot{x} - m_2 \dot{\theta}^2 x R \quad (7.132)$$

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

$$(J_1 + m_2 x^2) \ddot{\theta} = T + m_2 g x \cos \theta - 2m_2 x \dot{\theta} \dot{x} - m_2 \dot{\theta}^2 x R \quad (7.133)$$

$$(J_2 + m_2 R^2) \ddot{\theta} + \frac{1}{R} (J_2 + m_2 R^2) \ddot{x} = m_2 R x \dot{\theta}^2 + R m_2 g \sin \theta \quad (7.134)$$

**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

$$\begin{aligned} V &= \frac{1}{2}J_1\vec{\omega}_1 \cdot \vec{\omega}_1 + \frac{1}{2}J_2\vec{\omega}_2 \cdot \vec{\omega}_2 + \frac{1}{2}m_2\vec{v}_2 \cdot \vec{v}_2 + m_2g(-x\sin\theta + R\cos\theta) \\ &= \frac{1}{2}J_1\dot{\theta}^2 + \frac{1}{2}J_2\left(\dot{\theta} + \frac{\dot{x}}{R}\right)^2 + \frac{1}{2}m_2\left(\left(\dot{x} + \dot{\theta}R\right)^2 + \left(\dot{\theta}x\right)^2\right) \\ &\quad + m_2g(-x\sin\theta + R\cos\theta) \end{aligned} \quad (7.135)$$

The time derivative along the solutions of the system is

$$\begin{aligned} \dot{V} &= \dot{\theta}J_1\ddot{\theta} + \left(\dot{\theta} + \frac{\dot{x}}{R}\right)J_2\left(\ddot{\theta} + \frac{\ddot{x}}{R}\right) + \left(\dot{x} + \dot{\theta}R\right)m_2\left(\ddot{x} + \ddot{\theta}R\right) \\ &\quad + \dot{\theta}xm_2\ddot{\theta}x + \dot{\theta}xm_2\dot{\theta}\dot{x} - m_2g\left(\dot{x}\sin\theta + x\dot{\theta}\cos\theta + R\dot{\theta}\sin\theta\right) \\ &= \dot{\theta}(J_1 + m_2x^2)\ddot{\theta} + \left(\dot{\theta} + \frac{\dot{x}}{R}\right)(J_2 + m_2R^2)\left(\ddot{\theta} + \frac{\ddot{x}}{R}\right) + m_2x\dot{x}\dot{\theta}^2 \\ &\quad - m_2g\left(\dot{x}\sin\theta + x\dot{\theta}\cos\theta + R\dot{\theta}\sin\theta\right) \\ &= \dot{\theta}T \end{aligned} \quad (7.136)$$

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:

$$(J_1 + m_2x^2)\ddot{\theta} = T + m_2gx\cos\theta - 2m_2x\dot{\theta}\dot{x} - m_2\dot{\theta}^2xR \quad (7.137)$$

$$\frac{7}{5}m_2R^2\dot{\omega}_2 = Rm_2\dot{\theta}^2x + Rm_2g\sin\theta \quad (7.138)$$

**Example 125** If the radius of the ball becomes small, that is, when  $R \rightarrow 0$ , then the model becomes

$$(J_1 + mvx^2)\ddot{\theta} = T + m_2gx\cos\theta - 2mvx\dot{\theta}\dot{x} \quad (7.139)$$

$$\frac{7}{5}m_2\ddot{x} = m_2\dot{\theta}^2x + m_2g\sin\theta \quad (7.140)$$

**Example 126** Linearization about  $\dot{\theta} = 0$ ,  $\theta = 0$ ,  $\dot{x} = 0$  and  $x = 0$  gives

$$J_1\ddot{\theta} = T + m_2gx \quad (7.141)$$

$$\ddot{x} = \frac{5}{7}g\theta \quad (7.142)$$

which gives

$$\frac{d^4}{dt^4}x = \frac{1}{J_1}m_2gx + \frac{5}{7}\frac{g}{J_1}T \quad (7.143)$$

## 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

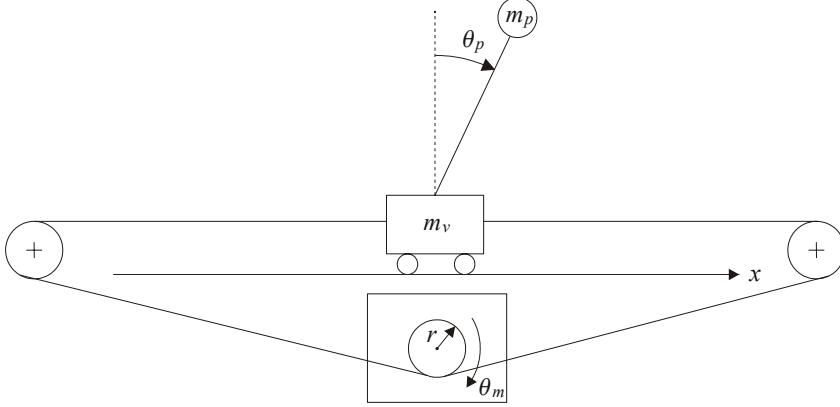


Figure 7.6: Inverted pendulum.

$m_b$  at the end of a massless rod of length  $L_b$ . The angle of the pendulum is denoted  $\theta_b$ , 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  $\theta_b$  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

$$\vec{r}_b = x\vec{n}_1 - L_b\vec{b}_1 \quad (7.144)$$

The velocity is found to be

$$\vec{v}_b = \frac{^n d\vec{r}_b}{dt} = \dot{x}\vec{n}_1 - \dot{\theta}_b \vec{b}_3 \times L_b \vec{b}_1 = \dot{x}\vec{n}_1 + \dot{\theta}_b L_b \vec{b}_2 \quad (7.145)$$

and acceleration is

$$\vec{a}_b = \ddot{x}\vec{n}_1 + \ddot{\theta}_b L_b \vec{b}_1 + \dot{\theta}_b \vec{b}_3 \times \dot{\theta}_b L_b \vec{b}_1 = \ddot{x}\vec{n}_1 + \ddot{\theta}_b L_b \vec{b}_1 + \dot{\theta}_b^2 L_b \vec{b}_2 \quad (7.146)$$

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

$$m_b \vec{a}_b = \vec{F}_b + mg \vec{n}_2 \quad (7.147)$$

where  $g$  is the acceleration of gravity. In the  $\vec{n}_1$  direction this gives

$$m_b (\ddot{x} + \ddot{\theta}_b L_b \cos \theta_b - \dot{\theta}_b^2 L_b \sin \theta_b) = \vec{F}_b \cdot \vec{n}_1 \quad (7.148)$$

Newton's law for the cart gives

$$m_v \ddot{x} = F - \vec{F}_b \cdot \vec{n}_1 \quad (7.149)$$

Combination of the two equations gives

$$(m_v + m_b) \ddot{x} + m_b L_b \ddot{\theta}_b \cos \theta_b - m_b \dot{\theta}_b^2 L_b \sin \theta_b = F \quad (7.150)$$

The torque law for the pendulum about the connection point is according to (7.45)

$$-L_b \vec{b}_2 \times m_b g \vec{n}_2 = -L_b \vec{b}_2 \times m_b (\ddot{x} \vec{n}_1 + \ddot{\theta}_b L_b \vec{b}_1 - \dot{\theta}_b L_b \vec{b}_2) \quad (7.151)$$

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

$$m_b L_b g \sin \theta_b = m_b L_b \ddot{x} \cos \theta_b + m_b L_b^2 \ddot{\theta}_b \quad (7.152)$$

The model for the cart and pendulum has then been found to be

$$(m_v + m_b) \ddot{x} + m_b L_b \ddot{\theta}_b \cos \theta_b = m_b \dot{\theta}_b^2 L_b \sin \theta_b + F \quad (7.153)$$

$$m_b L_b^2 \ddot{\theta}_b + m_b L_b \ddot{x} \cos \theta_b = m_b L_b g \sin \theta_b \quad (7.154)$$

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*

$$\begin{aligned} V &= \frac{1}{2} m_v \dot{x}^2 + \frac{1}{2} m_b \vec{v}_b \cdot \vec{v}_b + m_b g L_b \cos \theta_b \\ &= \frac{1}{2} m_v \dot{x}^2 + \frac{1}{2} m_b (\dot{x}^2 + 2L_b \cos \theta_b \dot{x} \dot{\theta}_b + \dot{\theta}_b^2 L_b^2) + m_b g L_b \cos \theta_b \end{aligned} \quad (7.155)$$

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

$$\begin{aligned} \dot{V} &= \dot{x} [(m_v + m_b) \ddot{x} + m_b L_b \cos \theta_b \ddot{\theta}_b] + \dot{\theta}_b (m_b L_b^2 \ddot{\theta}_b + m_b L_b \cos \theta_b \ddot{x}) \\ &\quad - L_b m_b \sin \theta_b \dot{x} \dot{\theta}_b^2 - m_b g L_b \dot{\theta}_b \sin \theta_b \\ &= \dot{x} (m_b \dot{\theta}_b^2 L_b \sin \theta_b + F) + \dot{\theta}_b m_b L_b g \sin \theta_b - L_b m_b \sin \theta_b \dot{x} \dot{\theta}_b^2 - m_b g L_b \dot{\theta}_b \sin \theta_b \\ &= F \dot{x} \end{aligned} \quad (7.156)$$

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

$$J_m \ddot{\theta}_m = K_T u - T_L \quad (7.157)$$

where  $\theta_m$  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

$$T_L = rF, \quad \dot{x} = r\dot{\theta}_m \quad (7.158)$$

which gives

$$\frac{J_m}{r^2} \ddot{x} = \frac{K_T}{r} u - F \quad (7.159)$$

It is observed that the cart and pendulum is driven by the motor through a port with effort  $F$  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

$$(m + m_b)\ddot{x} + m_b L_b \cos \theta_b \ddot{\theta}_b - m_b \dot{\theta}_b^2 L_b \sin \theta_b = \frac{K_T}{r} u \quad (7.160)$$

where  $m = m_v + J_m/r^2$ .

The model of cart, pendulum and motor is

$$(m + m_b)\ddot{x} + m_b L_b \cos \theta_b \ddot{\theta}_b - m_b \dot{\theta}_b^2 L_b \sin \theta_b = \frac{K_T}{r} u \quad (7.161)$$

$$m_b L_b^2 \ddot{\theta}_b + m_b L_b \ddot{x} \cos \theta_b = m_b L_b g \sin \theta_b \quad (7.162)$$

### 7.5.2 Double inverted pendulum

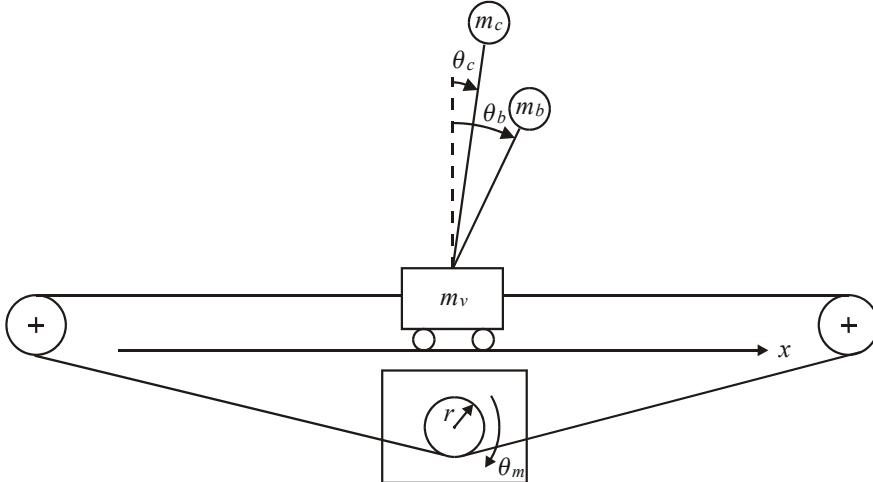


Figure 7.7: Double inverted pendulum.

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

$$\vec{r}_c = x \vec{n}_1 - L_c \vec{c}_2 \quad (7.163)$$

The velocity is

$$\vec{v}_c = \dot{x} \vec{n}_1 + \dot{\theta}_c L_c \vec{c}_1 \quad (7.164)$$

and acceleration is

$$\vec{a}_c = \ddot{x}\vec{n}_1 + \ddot{\theta}_c L_c \vec{c}_1 + \dot{\theta}_c^2 L_c \vec{c}_2 \quad (7.165)$$

Newton's law for the point mass of the second pendulum gives

$$m_c (\ddot{x} + \ddot{\theta}_c L_c \cos \theta_c - \dot{\theta}_c^2 L_c \sin \theta_c) = \vec{F}_c \cdot \vec{n}_1 \quad (7.166)$$

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

$$m_v \ddot{x} = F - \vec{F}_b \cdot \vec{n}_1 - \vec{F}_c \cdot \vec{n}_1 \quad (7.167)$$

The torque law for the second pendulum about the connection point is

$$m_c L_c g \sin \theta_c = m_c L_c \ddot{x} \cos \theta_c + m_c L_c^2 \ddot{\theta}_c \quad (7.168)$$

The model for a cart and two pendulums then is found to be

$$(m_v + m_b + m_c) \ddot{x} + m_b L_b \ddot{\theta}_b \cos \theta_b + m_c L_c \ddot{\theta}_c \cos \theta_c - m_b \dot{\theta}_b^2 L_b \sin \theta_b - m_c \dot{\theta}_c^2 L_c \sin \theta_c = F \quad (7.169)$$

$$m_b L_b^2 \ddot{\theta}_b + m_b L_b \ddot{x} \cos \theta_b = m_b L_b g \sin \theta_b \quad (7.170)$$

$$m_c L_c^2 \ddot{\theta}_c + m_c L_c \ddot{x} \cos \theta_c = m_c L_c g \sin \theta_c \quad (7.171)$$

The motor model is included by inserting

$$F = \frac{K_T}{r} u - \frac{J_m}{r^2} \ddot{x} \quad (7.172)$$

## 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  $\theta_1$  about the  $\vec{n}_3$  vector, and the frame  $c$  is obtained by a rotation  $\theta_2$  about the  $\vec{b}_2$  axis (Figure 7.8). According to (6.103) the frames  $n$  and  $b$  have direction cosines

$$\vec{n}_1 \cdot \vec{b}_1 = \cos \theta_1, \quad \vec{n}_1 \cdot \vec{b}_2 = -\sin \theta_1 \quad (7.173)$$

$$\vec{n}_2 \cdot \vec{b}_1 = \sin \theta_1, \quad \vec{n}_2 \cdot \vec{b}_2 = \cos \theta_1 \quad (7.174)$$

$$\vec{n}_1 \cdot \vec{b}_3 = \vec{n}_2 \cdot \vec{b}_3 = \vec{n}_3 \cdot \vec{b}_1 = \vec{n}_3 \cdot \vec{b}_2 = 0, \quad \vec{n}_3 \cdot \vec{b}_3 = 1 \quad (7.175)$$

and the unit vectors of frame  $b$  and frame  $c$  have direction cosines

$$\vec{b}_1 \cdot \vec{c}_1 = \cos \theta_2, \quad \vec{b}_1 \cdot \vec{c}_3 = \sin \theta_2 \quad (7.176)$$

$$\vec{b}_1 \cdot \vec{c}_2 = \vec{b}_3 \cdot \vec{c}_2 = \vec{b}_2 \cdot \vec{c}_1 = \vec{b}_2 \cdot \vec{c}_3 = 0, \quad \vec{b}_2 \cdot \vec{c}_2 = 1 \quad (7.177)$$

$$\vec{b}_3 \cdot \vec{c}_1 = -\sin \theta_2, \quad \vec{b}_3 \cdot \vec{c}_3 = \cos \theta_2 \quad (7.178)$$

It is noted that

$$\vec{b}_1 = \cos \theta_2 \vec{c}_1 + \sin \theta_2 \vec{c}_3 \quad (7.179)$$

$$\vec{b}_3 = -\sin \theta_2 \vec{c}_1 + \cos \theta_2 \vec{c}_3 \quad (7.180)$$

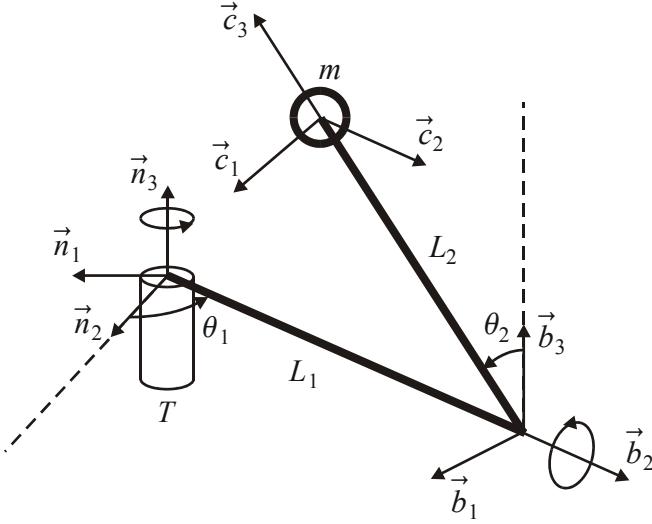


Figure 7.8: Coordinate frames  $n$ ,  $b$  and  $c$  used in the description of the Furuta pendulum.

and that

$$\vec{b}_1 \times \vec{c}_3 = -\cos \theta_2 \vec{c}_2 \quad (7.181)$$

$$\vec{b}_3 \times \vec{c}_1 = \cos \theta_2 \vec{c}_2, \quad \vec{b}_3 \times \vec{c}_2 = -\cos \theta_2 \vec{c}_1 - \sin \theta_2 \vec{c}_3, \quad \vec{b}_3 \times \vec{c}_3 = \sin \theta_2 \vec{c}_2 \quad (7.182)$$

The acceleration of gravity is  $\vec{g} = -g\vec{n}_3$ . The first link is rotated with angular velocity

$$\vec{\omega}_1 = \dot{\theta}_1 \vec{b}_3 \quad (7.183)$$

and the second link is rotated with angular velocity

$$\vec{\omega}_2 = \dot{\theta}_1 \vec{b}_3 + \dot{\theta}_2 \vec{c}_2 \quad (7.184)$$

The position of the mass

$$\vec{r} = L_1 \vec{b}_2 + L_2 \vec{c}_3$$

The velocity is

$$\begin{aligned} \vec{v} &= L_1 \dot{\theta}_1 \vec{b}_3 \times \vec{b}_2 + L_2 (\dot{\theta}_1 \vec{b}_3 + \dot{\theta}_2 \vec{c}_2) \times \vec{c}_3 \\ &= -L_1 \dot{\theta}_1 \vec{b}_1 + L_2 \dot{\theta}_1 \sin \theta_2 \vec{c}_2 + L_2 \dot{\theta}_2 \vec{c}_1 \end{aligned} \quad (7.185)$$

and the acceleration is

$$\begin{aligned} \vec{a} &= -L_1 \ddot{\theta}_1 \vec{b}_1 + L_2 \ddot{\theta}_1 \sin \theta_2 \vec{c}_2 + L_2 \dot{\theta}_1 \dot{\theta}_2 \cos \theta_2 \vec{c}_2 + L_2 \ddot{\theta}_2 \vec{c}_1 \\ &\quad -L_1 \dot{\theta}_1 \dot{\theta}_1 \vec{b}_3 \times \vec{b}_1 + (\dot{\theta}_1 \vec{b}_3 + \dot{\theta}_2 \vec{c}_2) \times (L_2 \dot{\theta}_1 \sin \theta_2 \vec{c}_2 + L_2 \dot{\theta}_2 \vec{c}_1) \\ &= -L_1 \ddot{\theta}_1 \vec{b}_1 + L_2 \ddot{\theta}_1 \sin \theta_2 \vec{c}_2 + L_2 \dot{\theta}_1 \dot{\theta}_2 \cos \theta_2 \vec{c}_2 + L_2 \ddot{\theta}_2 \vec{c}_1 \\ &\quad -L_1 \dot{\theta}_1^2 \vec{b}_2 - L_2 \dot{\theta}_1^2 \sin \theta_2 (\cos \theta_2 \vec{c}_1 + \sin \theta_2 \vec{c}_3) + L_2 \dot{\theta}_1 \dot{\theta}_2 \cos \theta_2 \vec{c}_2 - L_2 \dot{\theta}_2^2 \vec{c}_3 \end{aligned}$$

which gives

$$\vec{a} = -L_1\ddot{\theta}_1\vec{b}_1 + \left( L_2\ddot{\theta}_1 \sin \theta_2 + 2L_2\dot{\theta}_1\dot{\theta}_2 \cos \theta_2 - L_1\dot{\theta}_1^2 \right) \vec{b}_2 \\ + \left( L_2\ddot{\theta}_2 - L_2\dot{\theta}_1^2 \sin \theta_2 \cos \theta_2 \right) \vec{c}_1 - \left( L_2\dot{\theta}_1^2 \sin^2 \theta_2 + L_2\dot{\theta}_2^2 \right) \vec{c}_3 \quad (7.186)$$

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

$$T = J_1\ddot{\theta}_1 + (\vec{r} \times m\vec{a}) \cdot \vec{b}_3 \quad (7.187)$$

After some relative extensive vector calculations we may find that

$$(\vec{r} \times m\vec{a}) \cdot \vec{b}_3 = \left[ \left( L_1\vec{b}_2 + L_2\vec{c}_3 \right) \times m\vec{a} \right] \cdot \vec{b}_3 \\ = mL_1^2\ddot{\theta}_1 - mL_1L_2 \cos \theta_2\ddot{\theta}_2 + mL_1L_2\dot{\theta}_2^2 \sin \theta_2 \\ + mL_2^2\ddot{\theta}_1 \sin^2 \theta_2 + 2mL_2^2\dot{\theta}_1\dot{\theta}_2 \sin \theta_2 \cos \theta_2 \quad (7.188)$$

This gives the equation of motion

$$J_1\ddot{\theta}_1 + mL_1^2\ddot{\theta}_1 + mL_2^2 \sin^2 \theta_2 \ddot{\theta}_1 - mL_1L_2 \cos \theta_2 \ddot{\theta}_2 \\ = T - mL_1L_2\dot{\theta}_2^2 \sin \theta_2 - 2mL_2^2\dot{\theta}_1\dot{\theta}_2 \sin \theta_2 \cos \theta_2 \quad (7.189)$$

The equation of motion for the second link is found from

$$L_2\vec{c}_3 \times (-mg\vec{b}_3) = L_2\vec{c}_3 \times m\vec{a} \quad (7.190)$$

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

$$L_2mg \sin \theta_2 = mL_2\vec{c}_3 \times \left( -L_1\ddot{\theta}_1\vec{b}_1 + L_2\ddot{\theta}_2\vec{c}_1 - L_2\dot{\theta}_1^2 \sin \theta_2 \cos \theta_2 \vec{c}_1 \right) \cdot \vec{c}_2 \quad (7.191)$$

which is simplified to

$$L_2mg \sin \theta_2 = mL_2^2\ddot{\theta}_2 - mL_1L_2 \cos \theta_2\ddot{\theta}_1 - mL_2^2\dot{\theta}_1^2 \sin \theta_2 \cos \theta_2 \quad (7.192)$$

We may then conclude as follows:

The dynamic model of the Furuta pendulum is

$$(J_1 + mL_1^2 + mL_2^2 \sin^2 \theta_2)\ddot{\theta}_1 - mL_1L_2 \cos \theta_2\ddot{\theta}_2 \\ = T - mL_1L_2\dot{\theta}_2^2 \sin \theta_2 - 2mL_2^2\dot{\theta}_1\dot{\theta}_2 \sin \theta_2 \cos \theta_2 \quad (7.193)$$

$$-mL_1L_2 \cos \theta_2\ddot{\theta}_1 + mL_2^2\ddot{\theta}_2 = mL_2^2\dot{\theta}_1^2 \sin \theta_2 \cos \theta_2 + mL_2g \sin \theta_2 \quad (7.194)$$

**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

$$\begin{aligned} V &= \frac{1}{2} J_1 \dot{\theta}_1^2 + \frac{1}{2} m \vec{v} \cdot \vec{v} + mgL_2 \cos \theta_2 \\ &= \frac{1}{2} J_1 \dot{\theta}_1^2 + \frac{1}{2} m \left( L_1^2 \dot{\theta}_1^2 + L_2^2 \sin^2 \theta_2 \dot{\theta}_1^2 + L_2^2 \dot{\theta}_2^2 - 2L_1 L_2 \dot{\theta}_1 \dot{\theta}_2 \cos \theta_2 \right) + mgL_2 \cos \theta_2 \\ &= \frac{1}{2} (J_1 + mL_1^2 + mL_2^2 \sin^2 \theta_2) \dot{\theta}_1^2 + \frac{1}{2} mL_2^2 \dot{\theta}_2^2 \\ &\quad - mL_1 L_2 \dot{\theta}_1 \dot{\theta}_2 \cos \theta_2 + mgL_2 \cos \theta_2 \end{aligned} \tag{7.195}$$

The time derivative for the solutions of the system is

$$\begin{aligned} \dot{V} &= \dot{\theta}_1 \left[ (J_1 + mL_1^2 + mL_2^2 \sin^2 \theta_2) \ddot{\theta}_1 - mL_1 L_2 \ddot{\theta}_2 \cos \theta_2 \right] \\ &\quad + \dot{\theta}_2 \left( mL_2^2 \ddot{\theta}_2 - mL_1 L_2 \ddot{\theta}_1 \cos \theta_2 \right) \\ &\quad + mL_1 L_2 \dot{\theta}_1 \dot{\theta}_2^2 \sin \theta_2 + mL_2^2 \dot{\theta}_1^2 \dot{\theta}_2 \sin \theta_2 \cos \theta_2 - \dot{\theta}_2 mgL_2 \sin \theta_2 \\ &= \dot{\theta}_1 \left( T - mL_1 L_2 \dot{\theta}_2^2 \sin \theta_2 - 2mL_2^2 \dot{\theta}_1 \dot{\theta}_2 \sin \theta_2 \cos \theta_2 \right) \\ &\quad + \dot{\theta}_2 \left( mL_2^2 \dot{\theta}_1^2 \sin \theta_2 \cos \theta_2 + mL_2 g \sin \theta_2 \right) \\ &\quad + mL_1 L_2 \dot{\theta}_1 \dot{\theta}_2^2 \sin \theta_2 + mL_2^2 \dot{\theta}_1^2 \dot{\theta}_2 \sin \theta_2 \cos \theta_2 - \dot{\theta}_2 mgL_2 \sin \theta_2 \\ &= \dot{\theta}_1 T \end{aligned} \tag{7.196}$$

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 joint 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  $k = 1, \dots, N$ . Each particle is of mass  $m_k$  and has position

$$\vec{r}_k = x_k \vec{i}_1 + y_k \vec{i}_2 + z_k \vec{i}_3 \quad (7.197)$$

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

$$m_k \frac{d^2}{dt^2} \vec{r}_k = \vec{F}_k^{(r)} \quad (7.198)$$

Note that all differentiations of vectors are done in the Newtonian frame  $i$  in this section. Adding over all particles gives

$$\sum_{k=1}^N m_k \frac{d^2}{dt^2} \vec{r}_k = \sum_{k=1}^N \vec{F}_k^{(r)} \quad (7.199)$$

Suppose that there is an  $n$ -dimensional column vector  $\mathbf{q} = (q_1, \dots, q_n)^T$  so that the position  $\vec{r}_k$  of all particles are given as functions of  $\mathbf{q}$  and  $t$ , that is,

$$\vec{r}_k = \vec{r}_k [\mathbf{q}(t), t] \quad (7.200)$$

Then the variables  $q_1, \dots, q_n$  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  $q_1, \dots, q_n$  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

$$\vec{v}_k = \frac{d}{dt} \vec{r}_k = \sum_{i=1}^n \frac{\partial \vec{r}_k}{\partial q_i} \dot{q}_i + \frac{\partial \vec{r}_k}{\partial t} \quad (7.201)$$

**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

$$\frac{\partial \vec{v}_k}{\partial \dot{q}_i} = \frac{\partial \vec{r}_k}{\partial q_i} \quad (7.202)$$

Moreover, we find that by interchanging derivation with respect to  $q_i$  and  $t$  that

$$\frac{\partial \vec{v}_k}{\partial q_i} = \frac{\partial}{\partial q_i} \frac{d \vec{r}_k}{dt} = \frac{d}{dt} \frac{\partial \vec{r}_k}{\partial q_i} \quad (7.203)$$

### 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

$$\delta \vec{r}_k = \sum_{i=1}^n \frac{\partial \vec{r}_k}{\partial q_i} \delta q_i \quad (7.204)$$

where  $\delta q_i$  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  $\delta q_i$  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

$$\mathbf{A}(\mathbf{q}) \dot{\mathbf{q}} = \mathbf{0} \quad (7.205)$$

then the virtual displacements of the generalized coordinates will satisfy

$$\mathbf{A}(\mathbf{q}) \delta \mathbf{q} = \mathbf{0} \quad (7.206)$$

where  $\delta \mathbf{q} = (\delta q_1, \dots, \delta q_n)^T$ . 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

$$\sum_{k=1}^N \delta \vec{r}_k \cdot \vec{F}_k^{(c)} = 0 \quad (7.207)$$

Here  $\vec{F}_k^{(c)}$  is the force of constraint acting on particle  $k$ . Then the resultant force on particle  $k$  is given by

$$\vec{F}_k^{(r)} = \vec{F}_k^{(c)} + \vec{F}_k \quad (7.208)$$

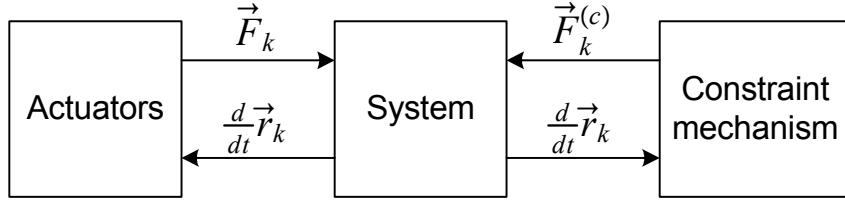


Figure 7.9: A mechanical system is driven by actuators that set up active forces  $\vec{F}_k$  acting on the system. In addition there is a constraint mechanism that set up constraint forces  $\vec{F}_k^{(c)}$  that ensures that the system does not break in parts. It simplifies the equation of motion 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

$$\begin{aligned} \sum_{k=1}^N \delta \vec{r}_k \cdot m_k \frac{d^2 \vec{r}_k}{dt^2} &= \sum_{k=1}^N \delta \vec{r}_k \cdot \vec{F}_k^{(c)} + \sum_{k=1}^N \delta \vec{r}_k \cdot \vec{F}_k \\ &= \sum_{k=1}^N \delta \vec{r}_k \cdot \vec{F}_k \end{aligned} \quad (7.209)$$

and we arrive at the following formulation of the equation of motion

$$\sum_{k=1}^N \delta \vec{r}_k \cdot \left( m_k \frac{d^2 \vec{r}_k}{dt^2} - \vec{F}_k \right) = 0 \quad (7.210)$$

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

$$\sum_{i=1}^n \delta q_i \sum_{k=1}^N \frac{\partial \vec{r}_k}{\partial q_i} \cdot \left( m_k \frac{d^2 \vec{r}_k}{dt^2} - \vec{F}_k \right) = 0 \quad (7.211)$$

If the virtual displacements  $\delta q_i$  in the generalized coordinates are independent, then d'Alembert's principle can be written

$$\sum_{k=1}^N \frac{\partial \vec{r}_k}{\partial q_i} \cdot \left( m_k \frac{d^2 \vec{r}_k}{dt^2} - \vec{F}_k \right) = 0 \quad (7.212)$$

**Example 130** A train is running along a railway. The generalized coordinate of the

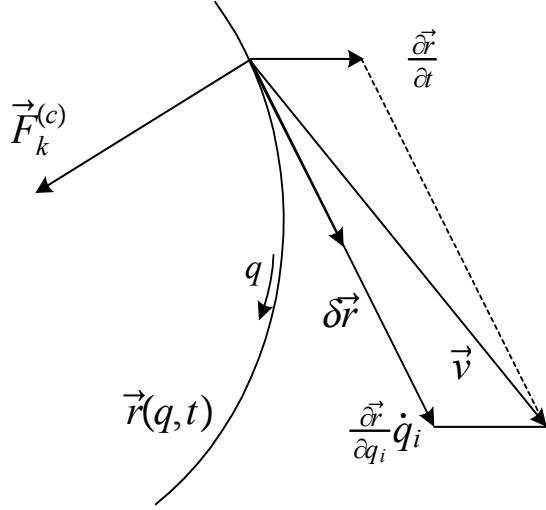


Figure 7.10: Train running along a track  $\mathbf{r}(q, t)$  where the track has a velocity  $\partial\mathbf{r}/\partial t$  due to the rotation of the earth. The virtual displacement  $\delta\mathbf{r}$  is along the track, and the force of constraint  $\mathbf{F}^{(c)}$  is perpendicular to the track 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*

$$\vec{v} = \frac{d\vec{r}}{dt} = \frac{\partial\vec{r}}{\partial q}\dot{q} + \frac{\partial\vec{r}}{\partial t} \quad (7.213)$$

*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*

$$\delta\vec{r} = \frac{\partial\vec{r}}{\partial q}\delta q \quad (7.214)$$

*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*

$$\delta\vec{r} \cdot \vec{F}^{(c)} = 0 \quad (7.215)$$

*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

$$\delta\mathbf{r}_k = \mathbf{J}_k \delta\mathbf{q} \quad (7.216)$$

The principle of virtual work states that for all  $\delta\mathbf{q}$  we have

$$0 = \sum_{k=1}^N \delta\mathbf{r}_k^T \mathbf{F}_k^{(c)} = \delta\mathbf{q}^T \sum_{k=1}^N \mathbf{J}_k^T \mathbf{F}_k^{(c)} \quad (7.217)$$

This means that the constraint force  $\mathbf{F}_k^{(c)}$  satisfies

$$\sum_{k=1}^N \mathbf{J}_k^T \mathbf{F}_k^{(c)} = \mathbf{0} \quad (7.218)$$

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

$$\vec{F}_{bc} = \vec{F}_{bc}^{(a)} + \vec{F}_{bc}^{(c)}, \quad \vec{T}_{bc} = \vec{T}_{bc}^{(a)} + \vec{T}_{bc}^{(c)} \quad (7.219)$$

The constraint force  $\vec{F}_{bc}^{(c)}$  and the constraint torque  $\vec{T}_{bc}^{(c)}$  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

$$\vec{v}_c = \sum_{j=1}^n \vec{v}_{c,j} \dot{q}_j + \vec{v}_t \quad (7.220)$$

$$\vec{\omega}_{ib} = \sum_{j=1}^n \vec{\omega}_{ib,j} \dot{q}_j + \vec{\omega}_t \quad (7.221)$$

where

$$\vec{v}_{c,j} = \frac{\partial \vec{r}_c}{\partial q_j} = \frac{\partial \vec{v}_c}{\partial \dot{q}_j}, \quad \vec{\omega}_{ib,j} = \frac{\partial \vec{\omega}_{ib}}{\partial \dot{q}_j} \quad (7.222)$$

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

$$\delta \vec{r}_c = \sum_{j=1}^n \vec{v}_{c,j} \delta q_j \quad (7.223)$$

$$\vec{\sigma}_{ib} = \sum_{j=1}^n \vec{\omega}_{ib,j} \delta q_j \quad (7.224)$$

**Example 132** If the configuration is given partly in terms of a rotation matrix, then it may not be convenient to use generalized coordinates. In this case the use of generalized velocities  $u_j$  will serve our purpose. Then the velocity and angular velocity is given by

$$\vec{v}_c = \sum_{j=1}^n \vec{v}_{c,j} u_j + \vec{v}_t \quad (7.225)$$

$$\vec{\omega}_{ib} = \sum_{j=1}^n \vec{\omega}_{ib,j} u_j + \vec{\omega}_t \quad (7.226)$$

where the partial velocities and the partial angular velocities are given by

$$\vec{v}_{c,j} = \frac{\partial \vec{v}_c}{\partial u_j}, \quad \vec{\omega}_{ib,j} = \frac{\partial \vec{\omega}_{ib}}{\partial u_j} \quad (7.227)$$

In this case the virtual displacements can be defined by

$$\delta \vec{r}_c = \sum_{j=1}^n \vec{v}_{c,j} \zeta_j \quad (7.228)$$

$$\vec{\sigma}_{ib} = \sum_{j=1}^n \vec{\omega}_{ib,j} \zeta_j \quad (7.229)$$

where  $\zeta_j$  are independent variables.

**Example 133** In a description using the velocity  $\vec{v}_o$  of a point  $o$  in the rigid body the partial velocities  $\vec{v}_{o,j} = \partial \vec{v}_o / \partial \dot{q}_j$  will be needed in the virtual displacements

$$\delta \vec{r}_o = \sum_{j=1}^n \vec{v}_{o,j} \delta \dot{q}_j \quad (7.230)$$

of the point  $o$ . The virtual displacements in rotation are the same as when the velocity of the center of mass is used.

### 7.8.2 Force and torque of constraint

The mechanical power of a mass force  $\vec{f} dm$  acting on the point  $p$  in a rigid body  $b$  is  $dP_m = \vec{v}_p \cdot d\vec{f}$ . The resulting power  $P_m$  from the mass forces on the body  $b$  is found by integrating  $dP_m$  over the body  $b$  using the expression  $\vec{v}_p = \vec{v}_c + \vec{\omega}_{ib} \times \vec{r}$ . The result is

$$P_m = \int_b \vec{v}_p \cdot d\vec{f} = \int_b (\vec{v}_c + \vec{\omega}_{ib} \times \vec{r}) \cdot d\vec{f} = \vec{v}_c \cdot \vec{F}^{(r)} + \vec{\omega}_{ib} \cdot \vec{N}_{b/c} \quad (7.231)$$

where  $\vec{F}^{(r)}$  is the resultant force on the body  $b$ , and  $\vec{N}_{b/c}$  is the moment on the body  $b$  about the center of the mass. As usual we will represent  $\vec{F}^{(r)}$  and  $\vec{N}_{b/c}$  by the equivalent description where  $\vec{F}_{bc}$  is a force vector equal to the resultant force and that has line of action through the center of mass, and  $\vec{T}_{bc}$  is a torque that is equal to the moment of the forces about the center of mass. The power supplied to a rigid body is then given by

$$P = \vec{v}_c \cdot \vec{F}_{bc} + \vec{\omega}_{ib} \cdot \vec{T}_{bc} \quad (7.232)$$

The force  $\vec{F}_{bc}$  is given as a sum of an active force  $\vec{F}_{bc}^{(a)}$  and a constraint force  $\vec{F}_{bc}^{(c)}$ , and in the same way the torque  $\vec{T}_{bc}$  is the sum of an active torque  $\vec{T}_{bc}^{(a)}$  and a constraint torque  $\vec{T}_{bc}^{(c)}$ . This is written

$$\vec{F}_{bc} = \vec{F}_{bc}^{(a)} + \vec{F}_{bc}^{(c)}, \quad \vec{T}_{bc} = \vec{T}_{bc}^{(a)} + \vec{T}_{bc}^{(c)} \quad (7.233)$$

According to the principle of virtual work the force of constraint  $\vec{F}_{bc}^{(c)}$  and the torque of constraint  $\vec{T}_{bc}^{(c)}$  on a rigid body satisfy

$$\delta\vec{r}_c \cdot \vec{F}_{bc}^{(c)} + \vec{\sigma}_{ib} \cdot \vec{T}_{bc}^{(c)} = 0 \quad (7.234)$$

where  $\delta\vec{r}_c$  and  $\vec{\sigma}_{ib}$  are the virtual displacements defined by (7.223 and (7.224)).

## 7.9 Multi-body dynamics and virtual work

### 7.9.1 Introduction

The principle of virtual work can be used to derive the equations of motion for a wide range of mechanical systems. The method is more systematic than a straightforward application of the Newton-Euler equations as it provides a systematic procedure for eliminating the forces and torques of constraint. The physical interpretation of the method is that the equations of motion are projected into the directions associated with the generalized speeds of the system. This may give significant simplification in the derivation of the equations of motion as it leaves out expressions pertaining to the directions associated with the forces of constraint. In the following the principle of virtual work will be used to derive the equations of motion for multi-body systems. This includes manipulators, the ball and beam system, inverted pendulums and the Furuta pendulum.

The principle of virtual work in the derivation of equations of motion has a long tradition in mechanics that stems back to d'Alembert and Euler. Recently the this method has been treated extensively in (Kane et al. 1983) and (Kane and Levinson 1985), and the method presented in this section is based on these references.

### 7.9.2 Equations of motion

In this section we will investigate the dynamics of a multi-body system in the form of  $n_b$  interconnected rigid bodies. The system has  $n$  generalized coordinates  $q_1, \dots, q_n$  that may be angles or translations. The generalized coordinates are assumed to be independent of each other, which implies that the virtual displacements  $\delta q_j$  are independent. Frame  $k$  is assumed to be fixed in rigid body  $k$  of the system. Rigid body  $k$  has mass  $m_k$ , inertia dyadic  $\vec{M}_{k/c}$  about the center of mass, and angular velocity  $\vec{\omega}_{0k}$  relative to the stationary base frame 0. The center of mass in body  $k$  has velocity  $\vec{v}_{ck}$  and acceleration  $\vec{a}_{ck}$ .

The velocity and the angular velocity of body  $k$  are given in terms of the generalized coordinates by

$$\vec{v}_{ck} = \sum_{j=1}^n \vec{v}_{ck,j} \dot{q}_j, \quad \vec{\omega}_{0k} = \sum_{j=1}^n \vec{\omega}_{0k,j} \dot{q}_j \quad (7.235)$$

and the virtual displacements are then given by

$$\delta\vec{r}_{ck} = \sum_{j=1}^n \vec{v}_{ck,j} \delta q_j, \quad \vec{\sigma}_{ok} = \sum_{j=1}^n \vec{\omega}_{0k,j} \delta q_j \quad (7.236)$$

where  $\delta q_j$  are independent variables that can be selected arbitrarily.

The forces and moments acting on body  $k$  are represented by the equivalent description with a force  $\vec{F}_{kc}$  acting through the center of mass, and a torque  $\vec{T}_{kc}$ . The force and the torque are split into an active part and a constraint part according to

$$\vec{F}_{kc} = \vec{F}_{bc}^{(a)} + \vec{F}_{bc}^{(c)}, \quad \vec{T}_{kc} = \vec{T}_{bc}^{(a)} + \vec{T}_{bc}^{(c)} \quad (7.237)$$

The equations of motion for the rigid body can then be written

$$m_k \vec{a}_{ck} - \vec{F}_{kc}^{(a)} - \vec{F}_{kc}^{(c)} = \vec{0} \quad (7.238)$$

$$\vec{M}_{k/c} \cdot \vec{\alpha}_{0k} + \vec{\omega}_{0k} \times (\vec{M}_{k/c} \cdot \vec{\omega}_{0k}) - \vec{T}_{kc}^{(a)} - \vec{T}_{kc}^{(c)} = \vec{0} \quad (7.239)$$

From the equations of motion it follows that

$$\begin{aligned} 0 &= \delta \vec{r}_{ck} \cdot \left( m \vec{a}_{ck} - \vec{F}_{kc}^{(a)} - \vec{F}_{kc}^{(c)} \right) \\ &\quad + \vec{\sigma}_{0k} \cdot \left( \vec{M}_{k/c} \cdot \vec{\alpha}_{0k} + \vec{\omega}_{0k} \times (\vec{M}_{k/c} \cdot \vec{\omega}_{0k}) - \vec{T}_{kc}^{(a)} - \vec{T}_{kc}^{(c)} \right) \end{aligned} \quad (7.240)$$

for body  $k$ .

The power supplied to the multi-body system from the forces  $\vec{F}_{kc}$  and torques  $\vec{T}_{kc}$  is

$$P = \sum_{k=1}^n \left( \vec{v}_{ck} \cdot \vec{F}_{kc} + \vec{\omega}_{0k} \cdot \vec{T}_{kc} \right) \quad (7.241)$$

where the result is obtained by summing up the power supplied to each rigid body. According to the principle of virtual work the forces and torques of constraint satisfy

$$\sum_{k=1}^n \left( \delta \vec{r}_{ck} \cdot \vec{F}_{kc}^{(c)} + \vec{\sigma}_{0k} \cdot \vec{T}_{kc}^{(c)} \right) = 0 \quad (7.242)$$

From this result it is seen that the forces and torques of constraint can be eliminated by summing up the equations (7.240) for all  $k$ , which gives

$$\sum_{k=1}^n \left[ \delta \vec{r}_{ck} \cdot \left( m_k \vec{a}_{ck} - \vec{F}_{kc}^{(a)} \right) + \vec{\sigma}_{0k} \cdot \left( \vec{M}_{k/c} \cdot \vec{\alpha}_{0k} + \vec{\omega}_{0k} \times (\vec{M}_{k/c} \cdot \vec{\omega}_{0k}) - \vec{T}_{kc}^{(a)} \right) \right] = 0$$

Insertion of the expression for the virtual displacements  $\delta \vec{r}_{ck}$  and  $\vec{\sigma}_{0k}$  from (7.236) gives

$$\sum_{k=1}^n \left[ \vec{v}_{ck,j} \cdot \left( m_k \vec{a}_{ck} - \vec{F}_{kc}^{(a)} \right) + \vec{\omega}_{0k,j} \cdot \left( \vec{M}_{k/c} \cdot \vec{\alpha}_{0k} + \vec{\omega}_{0k} \times (\vec{M}_{k/c} \cdot \vec{\omega}_{0k}) - \vec{T}_{kc}^{(a)} \right) \right] \delta q_j = 0$$

Note that this is a sum over the bodies numbered by  $k$ , while the index  $j$  of the generalized coordinate is held constant. To complete the derivation we define the generalized forces

$$\tau_j = \sum_{k=1}^n \vec{v}_{ck,j} \cdot \vec{F}_{kc}^{(a)} + \vec{\omega}_{0k,j} \cdot \vec{T}_{kc}^{(a)} \quad (7.243)$$

and note that the variations  $\delta q_j$  of the generalized coordinates are assumed to be arbitrary. Then the following result appears:

The equations of motion for a multi-body system with generalized coordinates  $q_j$  can be written

$$\sum_{k=1}^n \left[ \vec{v}_{ck,j} \cdot m_k \vec{a}_{ck} + \vec{\omega}_{0k,j} \cdot (\vec{M}_{k/c} \cdot \vec{\alpha}_{0k} + \vec{\omega}_{0k} \times (\vec{M}_{k/c} \cdot \vec{\omega}_{0k})) \right] = \tau_j \quad (7.244)$$

where

$$\tau_j = \sum_{k=1}^n \vec{v}_{ck,j} \cdot \vec{F}_{kc}^{(a)} + \vec{\omega}_{0k,j} \cdot \vec{T}_{kc}^{(a)} \quad (7.245)$$

This is Kane's formulation of the equation of motion for a multi-body system.

### 7.9.3 Equations of motion about a point

In the same way as in the previous section the equations of motion can be found using a description where the velocity of a point  $o$  is used. In this case the force  $\vec{F}_{bo}$  is assumed to have line of action through  $o$ , and the torque is  $\vec{T}_{bo} = \vec{T}_{bc} + \vec{r}_g \times \vec{F}_{bc}$  where  $\vec{r}_g$  is the vector from  $o$  to  $c$ . The equations of motion are in this case given by

$$\vec{F}_{bo} = m \vec{a}_c \quad (7.246)$$

$$\vec{T}_{bo} = \vec{r}_g \times m \vec{a}_o + \vec{M}_{b/o} \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_{b/o} \cdot \vec{\omega}_{ib}). \quad (7.247)$$

The force and torque are written as the sum of the active and constraint part according to  $\vec{F}_{bo} = \vec{F}_{bo}^{(a)} + \vec{F}_{bo}^{(c)}$  and  $\vec{T}_{bo} = \vec{T}_{bo}^{(a)} + \vec{T}_{bo}^{(c)}$ . Again the constrain force and torque is eliminated with the principle of virtual work, which gives

$$\sum_{k=1}^n \left[ \delta \vec{r}_{ok} \cdot \left( m_k \vec{a}_{ck} - \vec{F}_{ko}^{(a)} \right) + \vec{\sigma}_{0k} \cdot \left( \vec{r}_{gk} \times m_k \vec{a}_{ok} + \vec{M}_{k/o} \cdot \vec{\alpha}_{0k} + \vec{\omega}_{0k} \times (\vec{M}_{k/o} \cdot \vec{\omega}_{0k}) - \vec{T}_{ko}^{(a)} \right) \right] = 0 \quad (7.248)$$

$$\tau_j = \sum_{k=1}^n \vec{v}_{ok,j} \cdot \vec{F}_{ko}^{(a)} + \vec{\omega}_{0k,j} \cdot \vec{T}_{ko}^{(a)} \quad (7.249)$$

The equations of motion for a multi-body system with generalized coordinates  $q_j$  can be written

$$\sum_{k=1}^n \left[ \vec{v}_{ok,j} \cdot m_k \vec{a}_{ok} + \vec{\omega}_{0k,j} \cdot \left( \vec{r}_{gk} \times m_k \vec{a}_{ok} + \vec{M}_{k/o} \cdot \vec{\alpha}_{0k} + \vec{\omega}_{0k} \times (\vec{M}_{k/o} \cdot \vec{\omega}_{0k}) \right) \right] = \tau_j \quad (7.250)$$

where

$$\tau_j = \sum_{k=1}^n \vec{v}_{ok,j} \cdot \vec{F}_{ko}^{(a)} + \vec{\omega}_{0k,j} \cdot \vec{T}_{ko}^{(a)} \quad (7.251)$$

This is Kane's formulation about a point  $o$ .

**Example 134** A simple example is presented to illustrate the use of the equation of motion in the form (7.250). It is clear that this formulation is overly complicated for

this system, however, it may be a worthwhile exercise to understand the concept of this section. Consider a thin beam with homogeneous mass distribution and length  $L$ . The end of the beam denoted  $o$ . The beam is driven by a motor that is attached at  $o$  and rotates the beam by a angle  $\theta$  and acts with a torque  $\vec{T} = \tau \vec{b}_3$  on the beam. The velocity of  $o$  is zero. A frame  $b$  is fixed to the beam so that  $\vec{b}_1$  is along the length of the beam, and  $\vec{b}_3$  is the axis of rotation of the motor. A fixed frame  $i$  has axes that coincide with the axes of  $b$  when  $\theta = 0$ . The rotation matrix, the angular velocity and the angular acceleration from  $i$  to  $b$  are

$$\mathbf{R}_b^i = \mathbf{R}_{z,\theta}, \quad \vec{\omega}_{ib} = \dot{\theta} \vec{b}_3, \quad \vec{\alpha}_{ib} = \ddot{\theta} \vec{b}_3 \quad (7.252)$$

The position of the point  $o$  is  $\vec{r}_o = \vec{0}$ , and it follows that  $\vec{v}_o = \vec{0}$  and  $\vec{a}_o = \vec{0}$ . The inertia dyadic about  $o$  is

$$\vec{M}_o = \frac{mL^2}{3} (\vec{b}_2 \vec{b}_2 + \vec{b}_3 \vec{b}_3) \quad (7.253)$$

The angle  $\theta$  is selected as the generalized coordinate of the system. The associated general force is  $\tau$ . Then

$$\vec{v}_{o,1} = \vec{0}, \quad \vec{\omega}_{ib,1} = \vec{b}_3 \quad (7.254)$$

The principle of virtual work gives

$$0 = \vec{\omega}_{ib,1} \cdot (\vec{M}_o \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_o \cdot \vec{\omega}_{ib}) - \vec{T}) \quad (7.255)$$

and insertion of the relevant expressions gives

$$\begin{aligned} \tau &= \vec{b}_3 \cdot \frac{mL^2}{3} (\vec{b}_2 \vec{b}_2 + \vec{b}_3 \vec{b}_3) \cdot \ddot{\theta} \vec{b}_3 + \vec{b}_3 \cdot \dot{\theta} \vec{b}_3 \times \left( \frac{mL^2}{3} (\vec{b}_2 \vec{b}_2 + \vec{b}_3 \vec{b}_3) \cdot \dot{\theta} \vec{b}_3 \right) \\ &= \frac{mL^2}{3} \ddot{\theta} \end{aligned} \quad (7.256)$$

This leads to the equation of motion, which is simply

$$\frac{mL^2}{3} \ddot{\theta} = \tau \quad (7.257)$$

#### 7.9.4 Ball and beam

In this section the equations of motion will be derived for the ball and beam shown in Figure 7.5 with Kane's equations (7.244). In this problem the force of constraint is the contact force  $F_x \vec{b}_1 + F_y \vec{b}_2$  from the beam on the ball. The normal component  $F_y$  will be the contact force from the beam on the ball, while  $F_x$  is the friction force that makes the ball roll without slipping. The magnitude of these two forces need not be known, and it is convenient that the principle of virtual work cancel out these forces.

The generalized coordinates are  $q_1 = \theta$  and  $q_2 = x$ . The angular velocities and the velocity of the ball is given by

$$\vec{\omega}_1 = \dot{\theta} \vec{b}_3 \quad (7.258)$$

$$\vec{\omega}_2 = \dot{\theta} \vec{b}_3 + \dot{x} \frac{\vec{b}_3}{R} \quad (7.259)$$

$$\vec{v}_2 = \dot{\theta} (R \vec{b}_1 + x \vec{b}_2) + \dot{x} \vec{b}_1 \quad (7.260)$$

The acceleration of the center of mass of the ball is

$$\vec{a}_2 = \left( \ddot{x} + \ddot{\theta}R - \dot{\theta}^2 x \right) \vec{b}_1 + \left( \ddot{\theta}x + 2\dot{\theta}\dot{x} + \dot{\theta}^2 R \right) \vec{b}_2 \quad (7.261)$$

The angular accelerations are

$$\vec{\alpha}_1 = \ddot{\theta} \vec{b}_3 \quad (7.262)$$

$$\vec{\alpha}_2 = \left( \ddot{\theta} + \frac{\ddot{x}}{R} \right) \vec{b}_3 \quad (7.263)$$

The partial angular velocities and partial velocities are then found to be

$$\vec{v}_{2,1} = R\vec{b}_1 + x\vec{b}_2, \quad \vec{v}_{2,2} = \vec{b}_1 \quad (7.264)$$

$$\vec{\omega}_{1,1} = \vec{b}_3, \quad \vec{\omega}_{2,1} = \vec{b}_3, \quad \vec{\omega}_{2,2} = \frac{\vec{b}_3}{R} \quad (7.265)$$

The inertia dyadics are

$$\vec{M}_{1/c} = J_1 \left( \vec{b}_2 \vec{b}_2 + \vec{b}_3 \vec{b}_3 \right), \quad \vec{M}_{2/c} = J_2 \left( \vec{b}_1 \vec{b}_1 + \vec{b}_2 \vec{b}_2 + \vec{b}_3 \vec{b}_3 \right) \quad (7.266)$$

The forces and torques acting on the beam and ball are represented by the forces and torques

$$\vec{F}_1 = \vec{0} \quad (7.267)$$

$$\vec{F}_2 = F_x \vec{b}_1 + F_y \vec{b}_2 + m_2 g \vec{n}_2 \quad (7.268)$$

$$\vec{T}_{1/c} = T \vec{b}_3 + x \vec{b}_1 \times (-\vec{F}) = T \vec{b}_3 - x F_y \vec{b}_3 \quad (7.269)$$

$$\vec{T}_{2/c} = R \vec{b}_2 \times \vec{F}_2 = -R F_x \vec{b}_3 \quad (7.270)$$

Here the constraint forces  $F_x$  and  $F_y$  should have been left out according to the procedure of the virtual work, however, but we keep them in the derivation to see that they are actually cancelled.

The equation of motion (7.244) gives

$$\begin{aligned} \vec{v}_{2,1} \cdot m_2 \vec{a}_2 + \vec{\omega}_{1,1} \cdot \vec{M}_{1/c} \cdot \vec{\alpha}_1 + \vec{\omega}_{2,1} \cdot \vec{M}_{2/c} \cdot \vec{\alpha}_2 &= \vec{v}_{2,1} \cdot \vec{F}_2 + \vec{\omega}_{1,1} \cdot \vec{T}_{1/c} + \vec{\omega}_{2,1} \cdot \vec{T}_{2/c} \\ \vec{v}_{2,2} \cdot m_2 \vec{a}_2 + \vec{\omega}_{2,2} \cdot \vec{M}_{2/c} \cdot \vec{\alpha}_2 &= \vec{v}_{2,2} \cdot \vec{F}_2 + \vec{\omega}_{2,2} \cdot \vec{T}_{2/c} \end{aligned}$$

which give

$$\begin{aligned} & \left( R \vec{b}_1 + x \vec{b}_2 \right) \cdot m_2 \left( \left( \ddot{x} + \ddot{\theta}R - \dot{\theta}^2 x \right) \vec{b}_1 + \left( \ddot{\theta}x + 2\dot{\theta}\dot{x} + \dot{\theta}^2 R \right) \vec{b}_2 \right) \\ & + \vec{b}_3 \cdot J_1 \ddot{\theta} \vec{b}_3 + \vec{b}_3 \cdot J_2 \left( \ddot{\theta} + \frac{\ddot{x}}{R} \right) \vec{b}_3 \\ &= \left( R \vec{b}_1 + x \vec{b}_2 \right) \cdot \left( F_x \vec{b}_1 + F_y \vec{b}_2 + m_2 g \vec{n}_2 \right) + \vec{b}_3 \cdot \left( T \vec{b}_3 - x F_y \vec{b}_3 \right) - \vec{b}_3 \cdot R F_x \vec{b}_3 \end{aligned}$$

and

$$\begin{aligned} & \vec{b}_1 \cdot m_2 \left( \left( \ddot{x} + \ddot{\theta}R - \dot{\theta}^2 x \right) \vec{b}_1 + \left( \ddot{\theta}x + 2\dot{\theta}\dot{x} + \dot{\theta}^2 R \right) \vec{b}_2 \right) + \frac{\vec{b}_3}{R} \cdot J_2 \left( \ddot{\theta} + \frac{\ddot{x}}{R} \right) \vec{b}_3 \\ &= \vec{b}_1 \cdot \left( F_x \vec{b}_1 + F_y \vec{b}_2 + m_2 g \vec{n}_2 \right) + \frac{\vec{b}_3}{R} \cdot \left( -R F_x \vec{b}_3 \right) \end{aligned}$$

Note that at this point the forces of constraint  $F_x$  and  $F_y$  are eliminated from the equations. Evaluation of the scalar products gives

$$\begin{aligned} Rm_2 \left( \ddot{x} + \ddot{\theta}R - \dot{\theta}^2 x \right) + xm_2 \left( \ddot{\theta}x + 2\dot{\theta}\dot{x} + \dot{\theta}^2 R \right) \\ + J_1 \ddot{\theta} + J_2 \left( \ddot{\theta} + \frac{\ddot{x}}{R} \right) = m_2 g (R \sin \theta + x \cos \theta) + T \end{aligned}$$

and

$$m_2 \left( \ddot{x} + \ddot{\theta}R - \dot{\theta}^2 x \right) + \frac{J_2}{R} \left( \ddot{\theta} + \frac{\ddot{x}}{R} \right) = m_2 g \sin \theta$$

By rearranging the equations we find that

The equations for a ball and beam system can be written

$$\begin{aligned} [J_1 + J_2 + m_2 (x^2 + R^2)] \ddot{\theta} + \frac{1}{R} (J_2 + m_2 R^2) \ddot{x} + 2m_2 x \dot{x} \dot{\theta} \\ = T + m_2 g (R \sin \theta + x \cos \theta) \end{aligned} \quad (7.271)$$

$$\frac{1}{R} (J_2 + m_2 R^2) \ddot{\theta} + \frac{1}{R^2} (J_2 + m_2 R^2) \ddot{x} - m_2 \dot{\theta}^2 x = m_2 g \sin \theta \quad (7.272)$$

This model is found to be equivalent to the model (7.133, 7.134) derived with the Newton-Euler method as (7.272) is equal to (7.134), and (7.271) is obtained by adding 7.134 to (7.133).

### 7.9.5 Single and double inverted pendulum

In this section the equations of motion will be derived using Kane's equations (7.244) for a single and double inverted pendulum as shown in Figures 7.6 and 7.7. First the single pendulum will be treated. In this problem the force of constraint is the contact force  $\vec{F}_b$  between the pendulum and the cart. The velocity and acceleration of the cart are denoted

$$\vec{v}_v = \dot{x} \vec{n}_1, \quad \vec{a}_v = \ddot{x} \vec{n}_1 \quad (7.273)$$

while the velocity and the acceleration of the point mass at the end of the pendulum is

$$\vec{v}_b = \dot{x} \vec{n}_1 + \dot{\theta}_1 L_b \vec{b}_1 \quad (7.274)$$

$$\vec{a}_b = \ddot{x} \vec{n}_1 + \ddot{\theta}_1 L_b \vec{b}_1 + \dot{\theta}_1^2 L_b \vec{b}_2 \quad (7.275)$$

The angular velocity and angular acceleration of the pendulum are

$$\vec{\omega}_b = \dot{\theta}_1 \vec{b}_3, \quad \vec{\alpha}_b = \ddot{\theta}_1 \vec{b}_3 \quad (7.276)$$

The force of the cart is  $F \vec{n}_1$ , while the force on the pendulum is the gravity force  $m_b g \vec{n}_2$ . The pendulum is assumed to be made with a massless rod and a point mass, and it follows that the inertia dyadic is zero.

The nonzero partial velocities and partial angular velocities are

$$\vec{v}_{v,x} = \vec{n}_1, \quad \vec{v}_{b,x} = \vec{n}_1 \quad (7.277)$$

$$\vec{v}_{b,\theta_b} = L_b \vec{b}_1, \quad \vec{\omega}_{b,\theta_b} = \vec{b}_3 \quad (7.278)$$

Kane's equations (7.244, 7.245) give

$$\vec{n}_1 \cdot m_v \ddot{x} \vec{n}_1 + \vec{n}_1 \cdot m_b \left( \ddot{x} \vec{n}_1 + \ddot{\theta}_b L_b \vec{b}_1 + \dot{\theta}_b^2 L_b \vec{b}_2 \right) = F \quad (7.279)$$

$$L_b \vec{b}_1 \cdot m_b \left( \ddot{x} \vec{n}_1 + \ddot{\theta}_b L_b \vec{b}_1 + \dot{\theta}_b^2 L_b \vec{b}_2 \right) = L_b \vec{b}_1 \cdot g m_b \vec{n}_2 \quad (7.280)$$

where we have left out the force of constraint  $\vec{F}_b$  that would have been cancelled anyway if it had been included. By evaluating the scalar products, we find that the equations of motion are given by

$$(m_v + m_b) \ddot{x} + m_b L_b \cos \theta_b \ddot{\theta}_b - m_b L_b \sin \theta_b \dot{\theta}_b^2 = F \quad (7.281)$$

$$m_b L_b \cos \theta_b \ddot{x} + m_b L_b^2 \ddot{\theta}_b = m_b L_b g \sin \theta_b \quad (7.282)$$

which is the same result as (7.153, 7.154).

Next we add one more pendulum with point mass  $m_c$  at the end of a rod of length  $L_c$ . The velocity and acceleration of the point mass of the second pendulum are

$$\vec{v}_c = \dot{x} \vec{n}_1 + \dot{\theta}_1 L_c \vec{c}_1 \quad (7.283)$$

$$\vec{a}_c = \ddot{x} \vec{n}_1 + \ddot{\theta}_c L_c \vec{c}_1 + \dot{\theta}_c^2 L_c \vec{c}_2 \quad (7.284)$$

while the angular velocity and angular acceleration are

$$\vec{\omega}_c = \dot{\theta}_b \vec{b}_3, \quad \vec{\alpha}_c = \ddot{\theta}_b \vec{b}_3 \quad (7.285)$$

The gravity force on the second pendulum is  $m_c g \vec{n}_2$ . The nonzero partial velocities and partial angular velocities are

$$\vec{v}_{v,x} = \vec{n}_1, \quad \vec{v}_{b,x} = \vec{n}_1, \quad \vec{v}_{c,x} = \vec{n}_1 \quad (7.286)$$

$$\vec{v}_{b,\theta_b} = L_b \vec{b}_1, \quad \vec{\omega}_{b,\theta_b} = \vec{b}_3 \quad (7.287)$$

$$\vec{v}_{c,\theta_c} = L_c \vec{c}_1, \quad \vec{\omega}_{c,\theta_c} = \vec{b}_3 \quad (7.288)$$

Kane's equations give

$$\begin{aligned} \vec{n}_1 \cdot m_v \ddot{x} \vec{n}_1 + \vec{n}_1 \cdot m_b \left( \ddot{x} \vec{n}_1 + \ddot{\theta}_b L_b \vec{b}_1 + \dot{\theta}_b^2 L_b \vec{b}_2 \right) \\ + \vec{n}_1 \cdot m_c \left( \ddot{x} \vec{n}_1 + \ddot{\theta}_c L_c \vec{c}_1 + \dot{\theta}_c^2 L_c \vec{c}_2 \right) = F \end{aligned} \quad (7.289)$$

$$L_b \vec{b}_1 \cdot m_b \left( \ddot{x} \vec{n}_1 + \ddot{\theta}_b L_b \vec{b}_1 + \dot{\theta}_b^2 L_b \vec{b}_2 \right) = L_b \vec{b}_1 \cdot g m_b \vec{n}_2 \quad (7.290)$$

$$L_c \vec{c}_1 \cdot m_c \left( \ddot{x} \vec{n}_1 + \ddot{\theta}_c L_c \vec{c}_1 + \dot{\theta}_c^2 L_c \vec{c}_2 \right) = L_c \vec{c}_1 \cdot g m_c \vec{n}_2 \quad (7.291)$$

and the equations of motion are found by evaluation of the scalar products to be

$$\begin{aligned} (m_v + m_b + m_c) \ddot{x} + m_b L_b \cos \theta_b \ddot{\theta}_b + m_c L_c \cos \theta_c \ddot{\theta}_c \\ - m_b L_b \sin \theta_b \dot{\theta}_b^2 - m_c L_c \sin \theta_c \dot{\theta}_c^2 = F \end{aligned} \quad (7.292)$$

$$m_b L_b \cos \theta_b \ddot{x} + m_b L_b^2 \ddot{\theta}_b = m_b L_b g \sin \theta_b \quad (7.293)$$

$$m_c L_c \cos \theta_c \ddot{x} + m_c L_c^2 \ddot{\theta}_c = m_c L_c g \sin \theta_c \quad (7.294)$$

which is in agreement with (7.169–7.171).

### 7.9.6 Furuta pendulum

In this section we show how Kane's equations of motion can be used for the Furuta pendulum. The equations of motion were derived using the Newton-Euler formulation in Section 7.6.

With reference to Section 7.6 the angular velocities are

$$\vec{\omega}_1 = \dot{\theta}_1 \vec{b}_3 \quad (7.295)$$

$$\vec{\omega}_2 = \dot{\theta}_1 \vec{b}_3 + \dot{\theta}_2 \vec{c}_2 \quad (7.296)$$

The accelerations are

$$\vec{a}_{c1} = \vec{0} \quad (7.297)$$

$$\begin{aligned} \vec{a}_{c2} &= -L_1 \ddot{\theta}_1 \vec{b}_1 + \left( L_2 \ddot{\theta}_1 \sin \theta_2 + 2L_2 \dot{\theta}_1 \dot{\theta}_2 \cos \theta_2 - L_1 \dot{\theta}_1^2 \right) \vec{b}_2 \\ &\quad + \left( L_2 \ddot{\theta}_2 - L_2 \dot{\theta}_1^2 \sin \theta_2 \cos \theta_2 \right) \vec{c}_1 - \left( L_2 \dot{\theta}_1^2 \sin^2 \theta_2 + L_2 \dot{\theta}_2^2 \right) \vec{c}_3 \end{aligned} \quad (7.298)$$

while the partial velocities and partial angular velocities are

$$\vec{v}_{c2,1} = -L_1 \vec{b}_1 + L_2 \sin \theta_2 \vec{c}_2, \quad \vec{\omega}_{1,1} = \vec{b}_3, \quad \vec{\omega}_{2,1} = \vec{b}_3 \quad (7.299)$$

$$\vec{v}_{c2,2} = L_2 \vec{c}_1, \quad \vec{\omega}_{2,2} = \vec{c}_2 \quad (7.300)$$

The inertia dyadics are  $\vec{M}_{1/c} = J_1(\vec{n}_2 \vec{n}_2 + \vec{n}_3 \vec{n}_3)$  and  $\vec{M}_{c2} = \vec{0}$ . Then Kane's equations of motion give

$$\vec{v}_{c2,1} \cdot m_2 \vec{a}_{c2} + \vec{\omega}_{1,1} \cdot (\vec{M}_{1/c} \cdot \vec{\alpha}_1) = \tau_1 \quad (7.301)$$

$$\vec{v}_{c2,2} \cdot m_2 \vec{a}_{c2} = \vec{v}_{c2,2} \cdot (-m_2 g \vec{n}_3) \quad (7.302)$$

These two equations are identical to equations (7.187) and (7.190) that were used in the development of the model with the Newton-Euler approach. The rest of the derivation is therefore as in Section 7.6.

### 7.9.7 Planar two-link manipulator: Derivation 1

To demonstrate the use of Kane's equations of motion we consider the planar two-link manipulator in Figure 7.11. Frame 0 with orthogonal unit vectors  $\vec{i}_0, \vec{j}_0, \vec{k}_0$  is a fixed Newtonian frame with  $\vec{i}_0$  along the  $x_0$  axis,  $\vec{j}_0$  along the  $y_0$  axis, and  $\vec{k}_0$  along the  $z_0$  axis. Frame 1 with orthogonal unit vectors  $\vec{i}_1, \vec{j}_1, \vec{k}_1$  is fixed in link 1, and is obtained from frame 0 by a rotation  $q_1$  about  $\vec{k}_0$ . Frame 2 with orthogonal unit vectors  $\vec{i}_2, \vec{j}_2, \vec{k}_2$  is fixed in link 2, and is obtained by a rotation  $q_2$  about  $\vec{k}_1$ . This means that  $\vec{k}_0 = \vec{k}_1 = \vec{k}_2$ . The direction cosines are given by

$$\vec{i}_0 \cdot \vec{i}_1 = \cos q_1, \quad \vec{i}_0 \cdot \vec{j}_1 = -\sin q_1 \quad (7.303)$$

$$\vec{j}_0 \cdot \vec{i}_1 = \sin q_1, \quad \vec{j}_0 \cdot \vec{j}_1 = \cos q_1 \quad (7.304)$$

and

$$\vec{i}_1 \cdot \vec{i}_2 = \cos q_2, \quad \vec{i}_1 \cdot \vec{j}_2 = -\sin q_2 \quad (7.305)$$

$$\vec{i}_1 \cdot \vec{j}_2 = \sin q_2, \quad \vec{i}_1 \cdot \vec{j}_2 = \cos q_2 \quad (7.306)$$

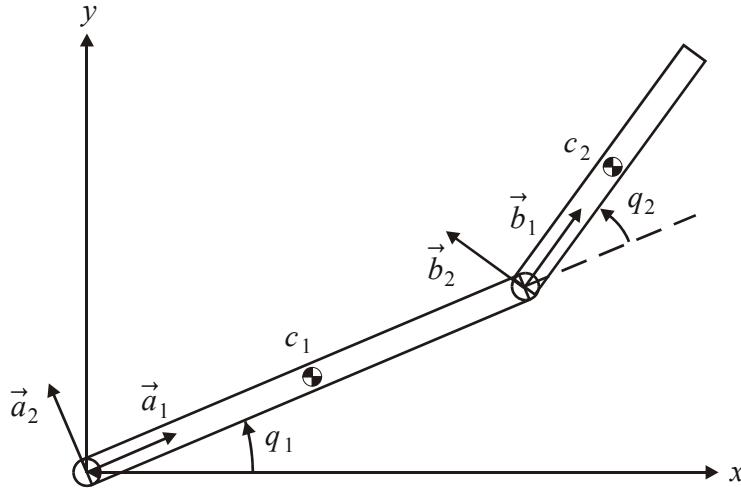


Figure 7.11: Planar two-link manipulator.

The angular velocity of the two links are

$$\vec{\omega}_1 = \dot{q}_1 \vec{k}_1 \quad (7.307)$$

$$\vec{\omega}_2 = (\dot{q}_1 + \dot{q}_2) \vec{k}_2 \quad (7.308)$$

and the angular acceleration is

$$\vec{\alpha}_1 = \ddot{q}_1 \vec{k}_1 \quad (7.309)$$

$$\vec{\alpha}_2 = (\ddot{q}_1 + \ddot{q}_2) \vec{k}_2 \quad (7.310)$$

The position of the centers of mass are

$$\vec{r}_{c1} = L_{c1} \vec{i}_1 \quad (7.311)$$

$$r_{c2} = L_1 \vec{i}_1 + L_{c2} \vec{i}_2 \quad (7.312)$$

The velocity of the centers of mass are

$$\begin{aligned} \vec{v}_{c1} &= L_{c1} \dot{q}_1 \vec{k}_1 \times \vec{i}_1 \\ &= \dot{q}_1 L_{c1} \vec{j}_1 \end{aligned} \quad (7.313)$$

$$\begin{aligned} \vec{v}_{c2} &= L_1 \dot{q}_1 \vec{k}_1 \times \vec{i}_1 + (\dot{q}_1 + \dot{q}_2) \vec{k}_2 \times L_{c2} \vec{i}_2 \\ &= \dot{q}_1 L_1 \vec{j}_1 + (\dot{q}_1 + \dot{q}_2) L_{c2} \vec{j}_2 \end{aligned} \quad (7.314)$$

and the acceleration of the centers of mass are

$$\begin{aligned} \vec{a}_{c1} &= \ddot{q}_1 L_{c1} \vec{j}_1 + \dot{q}_1 L_{c1} \vec{\omega}_1 \times \vec{j}_1 \\ &= \ddot{q}_1 L_{c1} \vec{j}_1 - \dot{q}_1^2 L_{c1} \vec{i}_1 \end{aligned} \quad (7.315)$$

$$\begin{aligned} \vec{a}_{c2} &= \ddot{q}_1 L_1 \vec{j}_1 + (\ddot{q}_1 + \ddot{q}_2) L_{c2} \vec{j}_2 \\ &\quad + \dot{q}_1 L_1 \vec{\omega}_1 \times \vec{j}_1 + (\dot{q}_1 + \dot{q}_2) L_{c2} \vec{\omega}_2 \times \vec{j}_2 \\ &= \ddot{q}_1 L_1 \vec{j}_1 + (\ddot{q}_1 + \ddot{q}_2) L_{c2} \vec{j}_2 - \dot{q}_1^2 L_1 \vec{i}_1 - (\dot{q}_1 + \dot{q}_2)^2 L_{c2} \vec{i}_2 \end{aligned} \quad (7.316)$$

This gives the following partial velocities and partial angular velocities:

$$\vec{v}_{c1,1} = L_{c1}\vec{j}_1, \quad \vec{v}_{c2,1} = L_1\vec{j}_1 + L_{c2}\vec{j}_2, \quad \vec{\omega}_{1,1} = \vec{k}_1, \quad \vec{\omega}_{2,1} = \vec{k}_2 \quad (7.317)$$

$$\vec{v}_{c1,2} = \vec{0}, \quad \vec{v}_{c2,2} = L_{c2}\vec{j}_2, \quad \vec{\omega}_{1,2} = \vec{0}, \quad \vec{\omega}_{2,2} = \vec{k}_2 \quad (7.318)$$

The mass of link 1 is  $m_1$ , the mass of link 2 is  $m_2$ , and the inertia dyadics about the centers of mass are

$$\vec{M}_{1/c} = I_{1x}\vec{i}_1\vec{i}_1 + I_{1y}\vec{j}_1\vec{j}_1 + I_{1z}\vec{k}_1\vec{k}_1 \quad (7.319)$$

$$\vec{M}_{2/c} = I_{2x}\vec{i}_2\vec{i}_2 + I_{2y}\vec{j}_2\vec{j}_2 + I_{2z}\vec{k}_2\vec{k}_2 \quad (7.320)$$

Then the equation of motion (7.244) gives

$$\begin{aligned} & \vec{v}_{c1,1} \cdot m_1 \vec{a}_{c1} + \vec{v}_{c2,1} \cdot m_2 \vec{a}_{c2} \\ & + \vec{\omega}_{1,1} \cdot (\vec{M}_{1/c} \cdot \vec{\alpha}_1 + \vec{\omega}_1 \times (\vec{M}_{1/c} \cdot \vec{\omega}_1)) \\ & + \vec{\omega}_{2,1} \cdot (\vec{M}_{2/c} \cdot \vec{\alpha}_2 + \vec{\omega}_2 \times (\vec{M}_{2/c} \cdot \vec{\omega}_2)) = \tau_1 \end{aligned} \quad (7.321)$$

$$\vec{v}_{c2,2} \cdot m_2 \vec{a}_{c2} + \vec{\omega}_{2,2} \cdot (\vec{M}_{2/c} \cdot \vec{\alpha}_2 + \vec{\omega}_2 \times (\vec{M}_{2/c} \cdot \vec{\omega}_2)) = \tau_2 \quad (7.322)$$

This gives

$$\begin{aligned} & L_{c1}\vec{j}_1 \cdot m_1 (\ddot{q}_1 L_{c1}\vec{j}_1 - \dot{q}_1^2 L_{c1}\vec{i}_1) \\ & + (L_1\vec{j}_1 + L_{c2}\vec{j}_2) \cdot m_2 (\ddot{q}_1 L_1\vec{j}_1 + (\ddot{q}_1 + \ddot{q}_2) L_{c2}\vec{j}_2 \\ & - \dot{q}_1^2 L_1\vec{i}_1 - (\dot{q}_1 + \dot{q}_2)^2 L_{c2}\vec{i}_2) \\ & + \vec{k}_1 \cdot I_{1z}\ddot{q}_1\vec{k}_1 + \vec{k}_2 \cdot I_{2z}(\ddot{q}_1 + \ddot{q}_2)\vec{k}_2 = \tau_1 \end{aligned} \quad (7.323)$$

$$\begin{aligned} & L_{c2}\vec{j}_2 \cdot m_2 (\ddot{q}_1 L_1\vec{j}_1 + (\ddot{q}_1 + \ddot{q}_2) L_{c2}\vec{j}_2 - \dot{q}_1^2 L_1\vec{i}_1 - (\dot{q}_1 + \dot{q}_2)^2 L_{c2}\vec{i}_2) \\ & + \vec{k}_2 \cdot I_{b3}(\ddot{q}_1 + \ddot{q}_2)\vec{k}_2 = \tau_2 \end{aligned} \quad (7.324)$$

which leads to the following conclusion:

The equations of motions for a planar two-link manipulator are given by

$$\begin{aligned} & [I_{1z} + I_{2z} + m_1 L_{c1}^2 + m_2 (L_{c2}^2 + L_1^2 + 2L_1 L_{c2} \cos q_2)] \ddot{q}_1 \\ & + [I_{2z} + m_2 (L_{c2}^2 + L_1 L_{c2} \cos q_2)] \ddot{q}_2 - m_2 L_1 L_{c2} \sin q_2 [2\dot{q}_1 \dot{q}_2 + \dot{q}_2^2] = \tau_1 \end{aligned} \quad (7.325)$$

$$[I_{2z} + m_2 (L_{c2}^2 + L_1 L_{c2} \cos q_2)] \ddot{q}_1 + (I_{2z} + m_2 L_{c2}^2) \ddot{q}_2 + m_2 L_1 L_{c2} \sin q_2 \dot{q}_1^2 = \tau_2 \quad (7.326)$$

### 7.9.8 Planar two-link manipulator: Derivation 2

In this section the equations of motion for the planar two-link manipulator in Figure 7.11 will be derived about the origins of the coordinate frames. The direction cosines, the

angular velocities and the angular accelerations are as in the previous section. The velocity of the origin of frame 1 and 2 are given by

$$\vec{v}_{o1} = \vec{0} \quad (7.327)$$

$$\vec{v}_{o2} = \dot{q}_1 L_1 \vec{j}_1 \quad (7.328)$$

This gives the following partial velocities and partial angular velocities:

$$\vec{v}_{o1,1} = \vec{0}, \quad \vec{v}_{o2,1} = L_1 \vec{j}_1, \quad \vec{\omega}_{1,1} = \vec{k}_1, \quad \vec{\omega}_{2,1} = \vec{k}_2 \quad (7.329)$$

$$\vec{v}_{o1,2} = \vec{0}, \quad \vec{v}_{o2,2} = \vec{0}, \quad \vec{\omega}_{1,2} = \vec{0}, \quad \vec{\omega}_{2,2} = \vec{k}_2 \quad (7.330)$$

The accelerations are

$$\vec{a}_{o1} = \vec{0} \quad (7.331)$$

$$\begin{aligned} \vec{a}_{o2} &= \ddot{q}_1 L_1 \vec{j}_1 + \dot{q}_1 L_1 \vec{\omega}_1 \times \vec{j}_1 \\ &= \ddot{q}_1 L_1 \vec{j}_1 - \dot{q}_1^2 L_1 \vec{i}_1 \end{aligned} \quad (7.332)$$

The inertia dyadic about the origins are found from the parallel axes theorem (7.84) to be

$$\vec{M}_{1/o} = I_{1x} \vec{i}_1 \vec{i}_1 + (I_{1y} + m_1 L_{c1}^2) \vec{j}_1 \vec{j}_1 + (I_{1z} + m_1 L_{c1}^2) \vec{k}_1 \vec{k}_1 \quad (7.333)$$

$$\vec{M}_{2/o} = I_{2x} \vec{i}_2 \vec{i}_2 + (I_{2y} + m_2 L_{c2}^2) \vec{j}_2 \vec{j}_2 + (I_{2z} + m_2 L_{c2}^2) \vec{k}_2 \vec{k}_2 \quad (7.334)$$

Then the equation of motion (7.250) gives

$$\begin{aligned} &\vec{v}_{o2,1} \cdot m_2 \vec{a}_{c2} \\ &+ \vec{\omega}_{1,1} \cdot (\vec{M}_{1/o} \cdot \vec{\alpha}_1 + \vec{\omega}_1 \times (\vec{M}_{1/o} \cdot \vec{\omega}_1)) \\ &+ \vec{\omega}_{2,1} \cdot (\vec{r}_{g2} \times m \vec{a}_{o2} + \vec{M}_{2/o} \cdot \vec{\alpha}_2 + \vec{\omega}_2 \times (\vec{M}_{2/o} \cdot \vec{\omega}_2)) = \tau_1 \end{aligned} \quad (7.335)$$

$$\vec{\omega}_{2,2} \cdot (\vec{r}_{g2} \times m \vec{a}_2 + \vec{M}_{2/o} \cdot \vec{\alpha}_2 + \vec{\omega}_2 \times (\vec{M}_{2/o} \cdot \vec{\omega}_2)) = \tau_2 \quad (7.336)$$

which gives

$$\begin{aligned} &L_1 \vec{j}_1 \cdot m_2 (\ddot{q}_1 L_1 \vec{j}_1 + (\ddot{q}_1 + \ddot{q}_2) L_{c2} \vec{j}_2 - \dot{q}_1^2 L_1 \vec{i}_1 - (\dot{q}_1 + \dot{q}_2)^2 L_{c2} \vec{i}_2) \\ &+ (I_{1z} + m_1 L_{c1}^2) \ddot{q}_1 + \vec{k}_2 \cdot L_{c2} \vec{i}_2 \times m_2 (\ddot{q}_1 L_1 \vec{j}_1 - \dot{q}_1^2 L_1 \vec{i}_1) \\ &+ (I_{2z} + m_2 L_{c2}^2) (\ddot{q}_1 + \ddot{q}_2) = \tau_1 \end{aligned} \quad (7.337)$$

$$\vec{k}_2 \cdot L_{c2} \vec{i}_2 \times m_2 (\ddot{q}_1 L_1 \vec{j}_1 - \dot{q}_1^2 L_1 \vec{i}_1) + (I_{2z} + m_2 L_{c2}^2) (\ddot{q}_1 + \ddot{q}_2) \vec{k}_2 = \tau_2 \quad (7.338)$$

and finally the equations of motions are found to be

$$\begin{aligned} &(I_{1z} + I_{2z} + m_1 L_{c1}^2 + m_2 L_{c2}^2 + m_2 L_1^2 + 2m_2 L_1 L_{c2} \cos q_2) \ddot{q}_1 \\ &+ (I_{2z} + m_2 L_{c2}^2 + m_2 L_1 L_{c2} \cos q_2) \ddot{q}_2 \\ &- m_2 L_1 L_{c2} \sin q_2 [2\dot{q}_1 \dot{q}_2 + \dot{q}_2^2] = \tau_1 \end{aligned} \quad (7.339)$$

$$[I_{2z} + m_2 (L_{c2}^2 + L_1 L_{c2} \cos q_2)] \ddot{q}_1 + (I_{2z} + m_2 L_{c2}^2) \ddot{q}_2 + m_2 L_1 L_{c2} \sin q_2 \dot{q}_1^2 = \tau_2 \quad (7.340)$$

This is the same result as the result of the previous section.

### 7.9.9 Kane's computational scheme for two-link manipulator

In this section we will see that a more efficient derivation is possible by using the method proposed in (Kane and Levinson 1983), where the equations of motion for a six-link robot were derived. We will use the equation of motion in the form (7.244), but we use a change of coordinates so that the generalized speeds are

$$u_1 = \dot{q}_1, \quad u_2 = \dot{q}_1 + \dot{q}_2 \quad (7.341)$$

The corresponding generalized forces are

$$K_1 = \tau_1 - \tau_2, \quad K_2 = \tau_2 \quad (7.342)$$

We note that this implies that

$$K_1 u_1 + K_2 u_2 = (\tau_1 - \tau_2) \dot{q}_1 + \tau_2 (\dot{q}_1 + \dot{q}_2) = \tau_1 \dot{q}_1 + \tau_2 \dot{q}_2 \quad (7.343)$$

The equations of motion for the two link manipulator are then

$$\begin{aligned} K_1 &= \vec{v}_{c1,u1} \cdot m_1 \vec{a}_{c1} + \vec{v}_{c2,1} \cdot m_2 \vec{a}_{c2} \\ &\quad + \vec{\omega}_{1,u1} \cdot (\vec{M}_{1/c} \cdot \vec{\alpha}_1 + \vec{\omega}_1 \times (\vec{M}_{1/c} \cdot \vec{\omega}_1)) \\ &\quad + \vec{\omega}_{2,u1} \cdot (\vec{M}_{2/c} \cdot \vec{\alpha}_2 + \vec{\omega}_2 \times (\vec{M}_{2/c} \cdot \vec{\omega}_2)) \end{aligned} \quad (7.344)$$

$$\begin{aligned} K_2 &= \vec{v}_{c1,u1} \cdot m_1 \vec{a}_{c1} + \vec{v}_{c2,u2} \cdot m_2 \vec{a}_{c2} \\ &\quad + \vec{\omega}_{1,u2} \cdot (\vec{M}_{1/c} \cdot \vec{\alpha}_1 + \vec{\omega}_1 \times (\vec{M}_{1/c} \cdot \vec{\omega}_1)) \\ &\quad + \vec{\omega}_{2,u2} \cdot (\vec{M}_{2/c} \cdot \vec{\alpha}_2 + \vec{\omega}_2 \times (\vec{M}_{2/c} \cdot \vec{\omega}_2)) \end{aligned} \quad (7.345)$$

where the partial velocities are referenced to the generalized speeds.

The expressions for the angular velocities are

$$\vec{\omega}_1 = u_1 \vec{k}_1 \quad (7.346)$$

$$\vec{\omega}_2 = u_2 \vec{k}_2 \quad (7.347)$$

Following the method of (Kane and Levinson 1983) we introduce intermediate variables  $Z_i$  for simplifying the derivation. The velocities are written

$$\vec{v}_{c1} = Z_1 u_1 \vec{j}_1 \quad (7.348)$$

$$\vec{v}_{c2} = Z_2 u_1 \vec{j}_1 + Z_3 u_2 \vec{j}_2 \quad (7.349)$$

and

$$\vec{v}_{c1} = Z_4 \vec{j}_1 \quad (7.350)$$

$$\vec{v}_{c2} = Z_5 \vec{j}_1 + Z_6 \vec{j}_2 \quad (7.351)$$

where the intermediate  $Z_i$  variables are defined by

$$Z_1 = L_{c1}, \quad Z_2 = L_1, \quad Z_3 = L_{c2} \quad (7.352)$$

$$Z_4 = u_1 Z_1, \quad Z_5 = u_1 Z_2, \quad Z_6 = u_2 Z_3 \quad (7.353)$$

The nonzero partial velocities and partial angular velocities with respect to  $u_1$  and  $u_2$  are

$$\vec{v}_{c1,u1} = Z_1 \vec{j}_1, \quad \vec{v}_{c2,u1} = Z_2 \vec{j}_1, \quad \vec{\omega}_{1,u1} = \vec{k}_1 \quad (7.354)$$

$$\vec{v}_{c2,u2} = Z_3 \vec{j}_2, \quad \vec{\omega}_{2,u2} = \vec{k}_2 \quad (7.355)$$

The angular accelerations are

$$\vec{\alpha}_1 = \dot{u}_1 \vec{k}_1 \quad (7.356)$$

$$\vec{\alpha}_2 = \dot{u}_2 \vec{k}_2 \quad (7.357)$$

and the accelerations of the centers of mass are found to be

$$\begin{aligned} \vec{a}_{c1} &= Z_1 \dot{u}_1 \vec{j}_1 + Z_1 u_1 \vec{\omega}_1 \times \vec{j}_1 \\ &= Z_1 \dot{u}_1 \vec{j}_1 - Z_1 u_1^2 \vec{i}_1 \end{aligned} \quad (7.358)$$

$$\begin{aligned} \vec{a}_{c2} &= Z_2 \dot{u}_1 \vec{j}_1 + Z_3 \dot{u}_2 \vec{j}_2 + Z_2 u_1 \vec{\omega}_1 \times \vec{j}_1 + Z_3 u_2 \vec{\omega}_2 \times \vec{j}_2 \\ &= Z_2 \dot{u}_1 \vec{j}_1 + Z_3 \dot{u}_2 \vec{j}_2 - Z_2 u_1^2 \vec{i}_1 - Z_3 u_2^2 \vec{i}_2 \end{aligned} \quad (7.359)$$

Then the equation of motion (7.344) and (7.345) are found to be

$$K_1 = Z_1 \vec{j}_1 \cdot m_1 \left( Z_1 \dot{u}_1 \vec{j}_1 - Z_1 u_1^2 \vec{i}_1 \right) \quad (7.360)$$

$$+ \left( Z_2 \vec{j}_1 + Z_3 \vec{j}_2 \right) \cdot m_2 \left( Z_2 \dot{u}_1 \vec{j}_1 + Z_3 \dot{u}_2 \vec{j}_2 - Z_2 u_1^2 \vec{i}_1 - Z_3 u_2^2 \vec{i}_2 \right) \quad (7.361)$$

$$+ \vec{k}_1 \cdot I_{a3} \dot{u}_1 \vec{k}_1 \quad (7.362)$$

$$K_2 = Z_3 \vec{j}_2 \cdot m_2 \left( Z_2 \dot{u}_1 \vec{j}_1 + Z_3 \dot{u}_2 \vec{j}_2 - Z_2 u_1^2 \vec{i}_1 - Z_3 u_2^2 \vec{i}_2 \right) \quad (7.363)$$

$$+ \vec{k}_2 \cdot I_{b3} \dot{u}_2 \vec{k}_2 \quad (7.364)$$

This can be written

$$K_1 = X_{11} \dot{u}_1 + X_{12} \dot{u}_2 + Z_7 \quad (7.365)$$

$$K_2 = X_{21} \dot{u}_1 + X_{22} \dot{u}_2 + Z_8 \quad (7.366)$$

where the coefficients are given by

$$X_{11} = m_1 Z_1^2 + m_2 Z_2^2 + I_{a3} \quad (7.367)$$

$$X_{12} = X_{21} = m_2 Z_2 Z_3 \cos q_1 \quad (7.368)$$

$$X_{22} = m_2 Z_3^2 + I_{b3} \quad (7.369)$$

$$Z_7 = -m_2 Z_2 Z_3 \sin q_1 u_1^2 \quad (7.370)$$

$$Z_8 = m_2 Z_2 Z_3 \sin q_1 u_1^2 \quad (7.371)$$

To obtain the original variables  $q_i$  and  $\tau_i$  we use that

$$\dot{q}_1 = u_1, \quad \dot{q}_2 = u_2 - u_1 \quad (7.372)$$

$$\tau_1 = K_1 + K_2, \quad \tau_2 = K_2 \quad (7.373)$$

To have a model in a form suitable for simulation the model is written

$$\dot{u}_1 = Y_{11} (K_1 - Z_7) + Y_{12} (K_2 - Z_8) \quad (7.374)$$

$$\dot{u}_2 = Y_{21} (K_1 - Z_7) + Y_{22} (K_2 - Z_8) \quad (7.375)$$

where

$$\begin{pmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix}^{-1} \quad (7.376)$$

**Example 135** Insertion of the  $Z_i$  and  $X_{ij}$  constants gives the equations of motion on the form

$$(m_1 L_{c1}^2 + m_2 L_1^2 + I_{1z}) \dot{u}_1 + (m_2 L_1 L_{c2} \cos q_1) \dot{u}_2 - m_2 L_1 L_{c2} \sin q_1 u_2^2 = K_1 \quad (7.377)$$

$$m_2 L_1 L_{c2} \cos q_1 \dot{u}_1 + (m_2 L_{c2}^2 + I_{2z}) \dot{u}_2 + m_2 L_1 L_{c2} \sin q_1 u_1^2 = K_2 \quad (7.378)$$

A change of variables by insertion of (7.341) and (7.342), and the addition of the second equation to the first equation gives the equations of motion as given by (7.325) and (7.326). Note that the equations of motion have a simpler form when the variables  $u_j$  and  $K_j$  are used.

### 7.9.10 Manipulator dynamics in coordinate form

In this section Kane's formulation (7.244) of the equations of motion for a multi-body system will be used to derive coordinate vector form of the equations of motion for a manipulator. The manipulator has  $n$  joints that connect the  $n$  links of the manipulator. The generalized coordinates in the case of rotational joints are the joint angles  $q_j$ . It is assumed that the joint angles are independent.

In coordinate vector form the equations of motion for link  $k$  are written

$$\begin{pmatrix} m_k \mathbf{a}_{ck}^k - (\mathbf{F}_k^k)^{(a)} - (\mathbf{F}_k^k)^{(c)} \\ \mathbf{M}_{k/c}^k \boldsymbol{\alpha}_{0k}^k + (\boldsymbol{\omega}_{0k}^k)^\times \mathbf{M}_{k/c}^k \boldsymbol{\omega}_{0k}^k - (\mathbf{T}_{kc}^k)^{(a)} - (\mathbf{T}_{kc}^k)^{(c)} \end{pmatrix} = \mathbf{0} \quad (7.379)$$

The velocity and angular velocity of link  $k$  are given by

$$\begin{pmatrix} \mathbf{v}_{ck}^k \\ \boldsymbol{\omega}_{0k}^k \end{pmatrix} = \mathbf{J}_k(\mathbf{q}) \dot{\mathbf{q}}. \quad (7.380)$$

where  $\mathbf{J}_k(\mathbf{q})$  is the Jacobian of link  $k$ . The virtual displacements of rigid body  $k$  are given by

$$\begin{pmatrix} \delta \mathbf{r}_k \\ \boldsymbol{\sigma}_k \end{pmatrix} = \mathbf{J}_k(\mathbf{q}) \delta \mathbf{q}. \quad (7.381)$$

Note that the velocities  $\dot{q}_j$  are assumed to be independent, and this implies that the virtual displacements  $\delta q_j$  are arbitrary.

The principle of virtual work can be written

$$0 = \sum_{k=1}^n \begin{pmatrix} \delta \mathbf{r}_k \\ \boldsymbol{\sigma}_k \end{pmatrix}^T \begin{pmatrix} (\mathbf{F}_k^k)^{(c)} \\ (\mathbf{T}_{kc}^k)^{(c)} \end{pmatrix} = \sum_{k=1}^n \delta \mathbf{q}^T \mathbf{J}_k^T \begin{pmatrix} (\mathbf{F}_k^k)^{(c)} \\ (\mathbf{T}_{kc}^k)^{(c)} \end{pmatrix} \quad (7.382)$$

The virtual displacements  $\delta \mathbf{q}$  are independent, and it follows that

$$\sum_{k=1}^n \mathbf{J}_k^T \begin{pmatrix} (\mathbf{F}_k^k)^{(c)} \\ (\mathbf{T}_{kc}^k)^{(c)} \end{pmatrix} = \mathbf{0} \quad (7.383)$$

The input generalized forces are defined by

$$\boldsymbol{\tau} = \sum_{k=1}^n \mathbf{J}_k^T \begin{pmatrix} \mathbf{F}_k^k \\ \mathbf{T}_{kc}^k \end{pmatrix} \quad (7.384)$$

which are the joint torque in the case of rotary joints. Then the equation of motion can be written

$$\sum_{k=1}^n \mathbf{J}_k^T \left( \mathbf{M}_{k/c}^k \boldsymbol{\alpha}_{0k}^k + (\boldsymbol{\omega}_{0k}^k)^\times \mathbf{M}_{k/c}^k \boldsymbol{\omega}_{0k}^k \right) = \boldsymbol{\tau} \quad (7.385)$$

From this equation it is possible to see that row  $j$  of the equations of motion is the sum of the projections of the equations of motion for link  $k$  along column  $j$  of the Jacobian  $\mathbf{J}_k$ .

### 7.9.11 Spacecraft and manipulator

In this section we will derive the equations of motion for a spacecraft with a manipulator. The spacecraft is described as link 0, and the manipulator links are denoted by link 1 to link 6. The mass of link  $k$  is  $m_k$ , the inertia dyadic of link  $k$  about the center of mass of link  $k$  is  $\bar{M}_{k/c}$ , the position is  $\vec{r}_{ck}$ , the velocity is  $\vec{v}_{ck}$ , and the acceleration is  $\vec{a}_{ck}$ , the angular velocity is with respect to an inertial frame  $i$  is  $\vec{\omega}_{ik}$ , and the angular acceleration is  $\vec{\alpha}_{ik}$ . The links of the manipulator are connected with rotational joints so that joint  $k$  with joint angle  $q_k$  connects link  $k - 1$  and  $k$ . The motor torque applied at joint  $k$  is  $\tau_k$ . The control forces and torques applied to the spacecraft are represented by a force  $\vec{F}_{0c}$  with line of action through the center of mass, and a torque  $\vec{T}_{0c}$ .

The position of the center of mass of link  $k$ ,  $k \geq 1$ , is

$$\vec{r}_{ck} = \vec{r}_{c0} + \sum_{j=1}^k \vec{d}_j - \vec{d}_{k,k_c} \quad (7.386)$$

where  $\vec{d}_j$  is the position of the origin of frame  $k$ ,  $k \geq 1$ , relative to the origin of frame 0, and  $\vec{d}_{k,k_c}$  is the position of the center of mass of link  $k$  relative to the origin of frame  $k$ .

The generalized speed vector and the generalized force vector are defined by

$$\vec{v}_{c0} = u_1 \vec{i}_0 + u_2 \vec{j}_0 + u_3 \vec{k}_0 \quad (7.387)$$

$$\vec{\omega}_{ik} = u_4 \vec{i}_0 + u_5 \vec{j}_0 + u_6 \vec{k}_0 \quad (7.388)$$

$$u_{k+6} = \dot{q}_k \quad (7.389)$$

and the generalized forces are

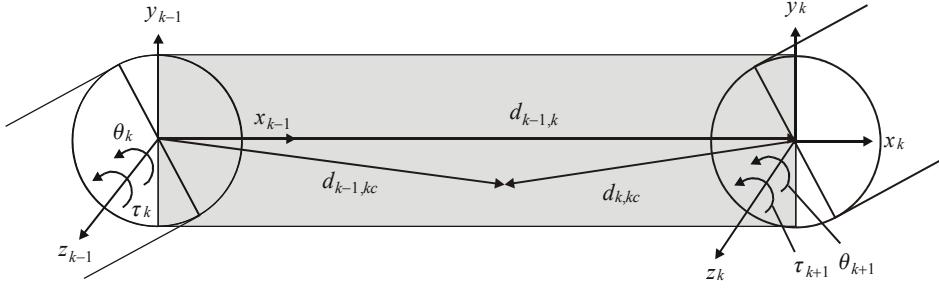
$$\vec{F}_{0c} = F_1 \vec{i}_0 + F_2 \vec{j}_0 + F_3 \vec{k}_0 \quad (7.390)$$

$$\vec{T}_{0c} = F_4 \vec{i}_0 + F_5 \vec{j}_0 + F_6 \vec{k}_0 \quad (7.391)$$

$$F_{k+6} = \tau_k \quad (7.392)$$

The partial velocities  $\vec{v}_{ck,j}$  and partial angular velocities  $\vec{\omega}_{ik,j}$  are then defined by

$$\vec{v}_{ck} = \sum_{k=0}^{12} \vec{v}_{ck,j} u_j, \quad \vec{\omega}_{ik} = \sum_{k=0}^{12} \vec{\omega}_{ik,j} u_j \quad (7.393)$$

Figure 7.12: Link  $k$ .

The equation of motion is then given by

$$\sum_{k=0}^{12} \left[ \vec{v}_{ck,j} \cdot m_k \vec{a}_{ck} + \vec{\omega}_{ik,j} \cdot (\vec{M}_{k/c} \cdot \vec{\alpha}_k + \vec{\omega}_k \times (\vec{M}_{k/c} \cdot \vec{\omega}_k)) \right] = F_j \quad (7.394)$$

Note that this formulation allows us to develop the equations of motion without introducing generalized coordinates in the form of Euler angles for the spacecraft.

## 7.10 Recursive Newton-Euler

### 7.10.1 Inverse dynamics

The inverse dynamics problem for manipulators is the problem of computing  $\tau$  given  $\mathbf{q}$ ,  $\dot{\mathbf{q}}$  and  $\ddot{\mathbf{q}}$ . A computational algorithm for this is the recursive Newton-Euler scheme of (Luh, Walker and Paul 1980). The main idea of the recursive Newton-Euler scheme is to compute velocities, angular velocities, accelerations and angular accelerations recursively from the base to the tip of the arm. Then the equations of motion are used for each link to compute the required resultant forces and torques on each link, and finally the contact forces and torques between the links are found by recursive computation from the tip of the arm to the base. Finally, the motor torques are found by projecting the contact torque onto the rotational axis of the joint.

Consider the link shown in Figure 7.12. The acceleration of the origin of Denavit-Hartenberg frame  $k$  is denoted  $\vec{a}_k$ . The contact force and torque from link  $(k - 1)$  on link  $k$  are denoted  $\vec{F}_{k-1,k}$  and  $\vec{T}_{k-1,k}$ , respectively. The distance from the origin of frame  $(k - 1)$  to the origin of frame  $k$  is denoted  $\vec{d}_{k-1,k}$ . The distance from the origin in frame  $(k - 1)$  to the mass center  $k_c$  is denoted  $\vec{d}_{k-1,k_c}$ , while  $\vec{d}_{k,k_c} = \vec{d}_{k-1,k_c} - \vec{d}_{k-1,k}$  is the distance from the origin of frame  $k$  to  $k_c$ . We note that  $\vec{d}_{k-1,k}$ ,  $\vec{d}_{k-1,k_c}$  and  $\vec{d}_{k,k_c}$  are constant vectors in frame  $k$ . We define  $\mathbf{z} = (0, 0, 1)^T$  as the unit vector in the  $z$  direction. Note that in accordance with the Denavit-Hartenberg convention the rotational axis of the joint between links  $k$  and  $(k + 1)$  is the  $z$  axis of frame  $k$ .

The forces and moments acting on link  $k$  are represented by the force  $\mathbf{F}_{kc}^k$  with magnitude and direction equal to the resultant force and line of action through the center of mass  $k_c$ , and by the torque  $\mathbf{T}_{kc}^k$  which is equal in magnitude and direction to the moment about the center of mass. The acceleration of frame 0 is set to be minus the acceleration of gravity as gravity always enters in the expressions together with acceleration as  $\vec{a}_k - \vec{g}$ .

**Recursive Newton-Euler:**

Initialization:  $\boldsymbol{\omega}_0^0 = \mathbf{0}$ ,  $\boldsymbol{\alpha}_0^0 = \mathbf{0}$ ,  $\mathbf{a}_0^0 = -\mathbf{g}^0$  where  $\mathbf{g}^0$  is the acceleration of gravity.

Outwards recursion ( $k = 1 \dots 6$ ):

$$\boldsymbol{\omega}_k^k = \mathbf{R}_{k-1}^k (\boldsymbol{\omega}_{k-1}^{k-1} + \mathbf{z}\dot{\theta}_k) \quad (7.395)$$

$$\boldsymbol{\alpha}_k^k = \mathbf{R}_{k-1}^k [\boldsymbol{\alpha}_{k-1}^{k-1} + (\boldsymbol{\omega}_{k-1}^{k-1})^\times \mathbf{z}\dot{\theta}_k + \mathbf{z}\ddot{\theta}_k] \quad (7.396)$$

$$\begin{aligned} \mathbf{a}_k^k &= \mathbf{R}_{k-1}^k \mathbf{a}_{k-1}^{k-1} \\ &+ (\boldsymbol{\alpha}_k^k)^\times \mathbf{d}_{k-1,k}^k + (\boldsymbol{\omega}_k^k)^\times (\boldsymbol{\omega}_k^k)^\times \mathbf{d}_{k-1,k}^k \end{aligned} \quad (7.397)$$

The equations of motion referred to the centers of mass ( $k = 1 \dots 6$ ):

$$\mathbf{F}_{kc}^k = m_k [\mathbf{a}_k^k + (\boldsymbol{\alpha}_k^k)^\times \mathbf{d}_{k,kc}^k + (\boldsymbol{\omega}_k^k)^\times (\boldsymbol{\omega}_k^k)^\times \mathbf{d}_{k,kc}^k] \quad (7.398)$$

$$\mathbf{T}_{kc}^k = \mathbf{M}_{k/c}^k \boldsymbol{\alpha}_k^k + (\boldsymbol{\omega}_k^k)^\times \mathbf{M}_{k/c}^k \boldsymbol{\omega}_k^k \quad (7.399)$$

Inwards recursion ( $k = 5 \dots 0$ ):

$$\mathbf{F}_{k-1,k}^k = \mathbf{R}_{k+1}^k \mathbf{F}_{k,k+1}^{k+1} + \mathbf{F}_{kc}^k \quad (7.400)$$

$$\begin{aligned} \mathbf{t}_{k-1,k}^k &= \mathbf{R}_{k+1}^k \mathbf{t}_{k,k+1}^{k+1} \\ &+ (\mathbf{d}_{k-1,k}^k)^\times \mathbf{R}_{k+1}^k \mathbf{F}_{k,k+1}^{k+1} + (\mathbf{d}_{k-1,kc}^k)^\times \mathbf{F}_{kc}^k + \mathbf{T}_{kc}^k \end{aligned} \quad (7.401)$$

Motor torques ( $k = 1 \dots 6$ ):

$$\tau_k = \mathbf{z}^T \mathbf{R}_k^{k-1} \mathbf{t}_{k-1,k}^k \quad (7.402)$$

If joint ( $k+1$ ) is prismatic, then  $\vec{\omega}_k = \vec{\omega}_{k-1}$  and  $\dot{\mathbf{d}}_{k,k}^{k-1} = \dot{d}_k \mathbf{z}$ , where  $d_k$  is the Denavit-Hartenberg parameter that specifies translation along the  $z$  axis. Then the outwards recursion is

$$\boldsymbol{\omega}_k^k = \mathbf{R}_{k-1}^k \boldsymbol{\omega}_{k-1}^{k-1} \quad (7.403)$$

$$\boldsymbol{\alpha}_k^k = \mathbf{R}_{k-1}^k \boldsymbol{\alpha}_{k-1}^{k-1} \quad (7.404)$$

$$\begin{aligned} \mathbf{a}_k^k &= \mathbf{R}_{k-1}^k (\mathbf{a}_{k-1}^{k-1} + \mathbf{z}\ddot{d}_k) + (\boldsymbol{\alpha}_k^k)^\times \mathbf{d}_{k-1,k}^k \\ &+ (\boldsymbol{\omega}_k^k)^\times (\boldsymbol{\omega}_k^k)^\times \mathbf{d}_{k-1,k}^k + 2(\boldsymbol{\omega}_k^k)^\times \mathbf{R}_{k-1}^k \mathbf{z} \dot{d}_k \end{aligned} \quad (7.405)$$

while the motor force is

$$\tau_k = \mathbf{z}^T \mathbf{R}_k^{k-1} \mathbf{F}_{k-1,k}^k \quad (7.406)$$

The algorithm requires  $117n - 24$  multiplications and  $103n - 21$  additions for a manipulator with  $n$  rotational joints, which gives 678 multiplications and 597 additions for a manipulator with six rotational joints.

### 7.10.2 Simulation

The recursive Newton-Euler scheme computes the required generalized forces  $\boldsymbol{\tau}(t)$  when  $\mathbf{q}(t)$ ,  $\dot{\mathbf{q}}(t)$  and  $\ddot{\mathbf{q}}(t)$  are given. In contrast to this the simulation problem is to calculate the state vector given by  $[\mathbf{q}(t), \dot{\mathbf{q}}(t)]$  when the initial state  $[\mathbf{q}(0), \dot{\mathbf{q}}(0)]$  and the generalized forces  $\boldsymbol{\tau}(t)$  are given. This is done by numerical integration of the acceleration  $\ddot{\mathbf{q}}(t)$ . Therefore, in the simulation of manipulator dynamics the acceleration  $\ddot{\mathbf{q}}(t)$  must be

computed when  $[\mathbf{q}(t), \dot{\mathbf{q}}(t)]$  and  $\boldsymbol{\tau}(t)$  are given. In the following a method based on the use of the recursive Newton-Euler algorithm to establish the model in the Lagrangian form. This is a convenient solution when the RNE algorithm is available.

The simulation problem involves the computation of  $\ddot{\mathbf{q}}$  given  $\mathbf{q}$ ,  $\dot{\mathbf{q}}$  and  $\boldsymbol{\tau}$ . This can be done using recursive Newton-Euler with the method of Walker and Orin (Walker and Orin 1982), (Sciavicco and Siciliano 2000). The method is based on the fact that the equation of motion for a manipulator can be written

$$\mathbf{M}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}} + \mathbf{g}(\mathbf{q}) = \boldsymbol{\tau} \quad (7.407)$$

This result will be derived in Section 8.2.8.

We denote the recursive Newton-Euler as a function  $\text{RNE}(\cdot)$  which takes  $\mathbf{q}$ ,  $\dot{\mathbf{q}}$  and  $\ddot{\mathbf{q}}$  as inputs and outputs  $\boldsymbol{\tau}$ . This is written

$$\boldsymbol{\tau} = \text{RNE}(\mathbf{q}, \dot{\mathbf{q}}, \ddot{\mathbf{q}}) \quad (7.408)$$

#### Simulation using the recursive Newton-Euler scheme:

1. Compute column  $j$  of  $\mathbf{M}(\mathbf{q}) = (\mathbf{m}_1 \ \mathbf{m}_2 \ \dots \ \mathbf{m}_6)$  ( $j = 1 \dots 6$ ):

$$\mathbf{m}_j = \text{RNE}(\mathbf{q}, \mathbf{0}, \mathbf{e}_j) \quad (7.409)$$

with  ${}^0\mathbf{a}_0 = \mathbf{0}$  where  $\mathbf{e}_1 = (1, 0, \dots, 0)^T$ ,  $\mathbf{e}_2 = (0, 1, \dots, 0)^T$ ,  $\dots$ ,  $\mathbf{e}_6 = (0, 0, \dots, 1)^T$ .

2. Compute

$$\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}} + \mathbf{g}(\mathbf{q}) = \text{RNE}(\mathbf{q}, \dot{\mathbf{q}}, \mathbf{0}) \quad (7.410)$$

3. Compute the accelerations from

$$\ddot{\mathbf{q}} = \mathbf{M}^{-1}(\mathbf{q})(-\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}} - \mathbf{g}(\mathbf{q}) + \boldsymbol{\tau}) \quad (7.411)$$

We see that the  $\text{RNE}(\cdot)$  function is used  $n+1$  times, and that a Gaussian elimination is required to compute  $\ddot{\mathbf{q}}$ . This means that the computational requirements for the simulation is higher than for the inverse dynamics problem.

# 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  $\vec{r}_k(q_1, \dots, q_n, t)$ , where  $q_1, \dots, q_n$  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)

$$\sum_{i=1}^n \left[ \sum_{k=1}^N \frac{\partial \vec{r}_k}{\partial q_i} \cdot (m_k \vec{a}_k - \vec{F}_k) \right] \delta q_i = 0 \quad (8.1)$$

To proceed we introduce the kinetic energy  $T$  of the system, which is

$$T = \sum_{k=1}^N \frac{1}{2} m_k \vec{v}_k \cdot \vec{v}_k \quad (8.2)$$

We find that

$$\frac{\partial T}{\partial \dot{q}_i} = \frac{\partial}{\partial \dot{q}_i} \left( \sum_{k=1}^N \frac{1}{2} m_k \vec{v}_k \cdot \vec{v}_k \right) = \sum_{k=1}^N \frac{\partial \vec{v}_k}{\partial \dot{q}_i} \cdot m_k \vec{v}_k = \sum_{k=1}^N \frac{\partial \vec{r}_k}{\partial q_i} \cdot m_k \vec{v}_k \quad (8.3)$$

$$\frac{\partial T}{\partial q_i} = \frac{\partial}{\partial q_i} \left( \sum_{k=1}^N \frac{1}{2} m_k \vec{v}_k \cdot \vec{v}_k \right) = \sum_{k=1}^N \frac{\partial \vec{v}_k}{\partial q_i} \cdot m_k \vec{v}_k = \sum_{k=1}^N \frac{d}{dt} \frac{\partial \vec{r}_k}{\partial q_i} \cdot m_k \vec{v}_k \quad (8.4)$$

where (7.202) and (7.203) are used. The following calculation can then be done:

$$\begin{aligned} \frac{d}{dt} \frac{\partial T}{\partial \dot{q}_i} &= \sum_{k=1}^N \frac{d}{dt} \left( \frac{\partial \vec{r}_k}{\partial q_i} \cdot m_k \vec{v}_k \right) = \sum_{k=1}^N \left( \frac{d}{dt} \frac{\partial \vec{r}_k}{\partial q_i} \cdot m_k \vec{v}_k + \frac{\partial \vec{r}_k}{\partial q_i} \cdot m_k \vec{a}_k \right) \\ &= \frac{\partial T}{\partial q_i} + \sum_{k=1}^N \frac{\partial \vec{r}_k}{\partial q_i} \cdot m_k \vec{a}_k \end{aligned} \quad (8.5)$$

This result combined with (8.1) leads to

$$\sum_{i=1}^n \left[ \frac{d}{dt} \left( \frac{\partial T}{\partial \dot{q}_i} \right) - \frac{\partial T}{\partial q_i} - \sum_{k=1}^N \frac{\partial \vec{r}_k}{\partial q_i} \cdot \vec{F}_k \right] \delta q_i = 0 \quad (8.6)$$

The third term in the bracket is defined to be the *generalized force*

$$Q_i := \sum_{k=1}^N \frac{\partial \vec{r}_k}{\partial q_i} \cdot \vec{F}_k \quad (8.7)$$

associated with the generalized coordinate  $q_i$ . This gives

$$\sum_{i=1}^n \left[ \frac{d}{dt} \left( \frac{\partial T}{\partial \dot{q}_i} \right) - \frac{\partial T}{\partial q_i} - Q_i \right] \delta q_i = 0 \quad (8.8)$$

Then, under the assumption that the time derivatives  $\dot{q}_i$  of the generalized coordinates are independent, the virtual displacements  $\delta q_i$  are arbitrary, and it follows that

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \dot{q}_i} \right) - \frac{\partial T}{\partial q_i} = Q_i \quad (8.9)$$

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  $\tau_i$ . This is written

$$Q_i = -\frac{\partial U}{\partial q_i} + \tau_i \quad (8.10)$$

Then the equation of motion becomes

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \dot{q}_i} \right) - \frac{\partial T}{\partial q_i} + \frac{\partial U}{\partial q_i} = \tau_i \quad (8.11)$$

From this result, Lagrange's equation of motion is found:

Lagrange's equation of motion is formulated using the *Lagrangian*

$$L(\mathbf{q}, \dot{\mathbf{q}}, t) = T(\mathbf{q}, \dot{\mathbf{q}}, t) - U(\mathbf{q}) \quad (8.12)$$

The equation of motion is

$$\frac{d}{dt} \left( \frac{\partial L}{\partial \dot{q}_i} \right) - \frac{\partial L}{\partial q_i} = \tau_i \quad (8.13)$$

**Example 136** For use in the part on Hamiltonian dynamics we derive the following result: Time differentiation of the Lagrangian gives

$$\begin{aligned} \frac{dL(\mathbf{q}, \dot{\mathbf{q}}, t)}{dt} &= \frac{\partial L(\mathbf{q}, \dot{\mathbf{q}}, t)}{\partial \mathbf{q}} \dot{\mathbf{q}} + \frac{\partial L(\mathbf{q}, \dot{\mathbf{q}}, t)}{\partial \dot{\mathbf{q}}} \ddot{\mathbf{q}} + \frac{\partial L(\mathbf{q}, \dot{\mathbf{q}}, t)}{\partial t} \\ &= \left( \frac{d}{dt} \frac{\partial L(\mathbf{q}, \dot{\mathbf{q}}, t)}{\partial \dot{\mathbf{q}}} - \boldsymbol{\tau}^T \right) \dot{\mathbf{q}} + \frac{\partial L(\mathbf{q}, \dot{\mathbf{q}}, t)}{\partial \dot{\mathbf{q}}} \ddot{\mathbf{q}} + \frac{\partial L(\mathbf{q}, \dot{\mathbf{q}}, t)}{\partial t} \end{aligned} \quad (8.14)$$

where Lagrange's equation of motion (8.13) has been inserted. This gives

$$\frac{dL(\mathbf{q}, \dot{\mathbf{q}}, t)}{dt} = \frac{d}{dt} \left( \frac{\partial L(\mathbf{q}, \dot{\mathbf{q}}, t)}{\partial \dot{\mathbf{q}}} \dot{\mathbf{q}} \right) + \frac{\partial L(\mathbf{q}, \dot{\mathbf{q}}, t)}{\partial t} - \boldsymbol{\tau}^T \dot{\mathbf{q}} \quad (8.15)$$

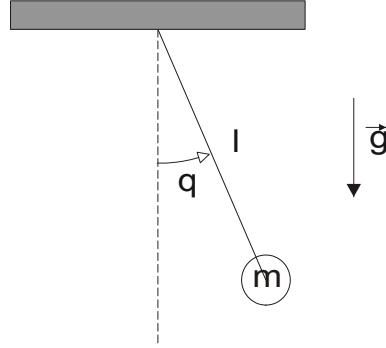


Figure 8.1: Mathematical pendulum.

### 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

$$\begin{aligned} \sum_{k=1}^N \frac{d\vec{r}_k}{dt} \cdot \vec{F}_k &= \sum_{k=1}^N \left( \sum_{i=1}^n \frac{\partial \vec{r}_k}{\partial q_i} \dot{q}_i \cdot \vec{F}_k \right) = \sum_{i=1}^n \left( \sum_{k=1}^N \frac{\partial \vec{r}_k}{\partial q_i} \cdot \vec{F}_k \right) \dot{q}_i \\ &= \sum_{i=1}^n Q_i \dot{q}_i \end{aligned} \quad (8.16)$$

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

$$T = \frac{1}{2}mv^2 = \frac{1}{2}m\ell^2\dot{q}^2 \quad (8.17)$$

The potential energy is

$$U = mgl(1 - \cos q) \quad (8.18)$$

The resulting Lagrangian is

$$L = \frac{1}{2}m\ell^2\dot{q}^2 - mgl(1 - \cos q) \quad (8.19)$$

and the equation of motion is

$$\frac{d}{dt}(m\ell^2\dot{q}) + mgl\sin q = 0 \quad (8.20)$$

which gives

$$\ddot{q} + \omega_0^2 \sin q = 0, \quad \omega_0 = \sqrt{\frac{g}{\ell}} \quad (8.21)$$

### 8.2.5 Mass-spring system

A mass-spring system with mass  $m$  and spring stiffness  $k$  will have Lagrangian

$$L = T - U = \frac{1}{2}m\dot{q}^2 - \frac{1}{2}kq^2 \quad (8.22)$$

Lagrange's equation of motion is then found to be

$$\frac{d}{dt}(m\dot{q}) + kq = \tau \quad (8.23)$$

which can be written in the familiar form

$$m\ddot{q} + kq = \tau \quad (8.24)$$

### 8.2.6 Ball and beam

The ball and beam system presented in Section 7.4 has kinetic energy

$$\begin{aligned} T &= \frac{1}{2}J_1\dot{\theta}^2 + \frac{1}{2}J_2\left(\dot{\theta} + \frac{\dot{x}}{R}\right)^2 + \frac{1}{2}m\left[\left(\dot{x} + \dot{\theta}R\right)^2 + \left(\dot{\theta}x\right)^2\right] \\ &= \frac{1}{2}\begin{pmatrix} \dot{\theta} \\ \dot{x} \end{pmatrix}^T \begin{pmatrix} J_1 + J_2 + m(x^2 + R^2) & \frac{1}{R}(J_2 + mR^2) \\ \frac{1}{R}(J_2 + mR^2) & m + \frac{J_2}{R^2} \end{pmatrix} \begin{pmatrix} \dot{\theta} \\ \dot{x} \end{pmatrix} \end{aligned} \quad (8.25)$$

and potential energy

$$U = mg(R\cos\theta - x\sin\theta) \quad (8.26)$$

The generalized coordinates are selected as

$$q_1 = \theta \quad \text{and} \quad q_2 = x \quad (8.27)$$

Then, with  $L = T - U$ , we have the following partial derivatives

$$\frac{\partial L}{\partial \dot{\theta}} = J_1\dot{\theta} + J_2\left(\dot{\theta} + \frac{\dot{x}}{R}\right) + m\left[\left(\dot{x} + \dot{\theta}R\right)R + \left(\dot{\theta}x\right)x\right] \quad (8.28)$$

$$\frac{\partial L}{\partial \dot{x}} = J_2\left(\dot{\theta} + \frac{\dot{x}}{R}\right)\frac{1}{R} + m\left(\dot{x} + \dot{\theta}R\right) \quad (8.29)$$

$$\frac{\partial L}{\partial \theta} = mg(R\sin\theta + x\cos\theta) \quad (8.30)$$

$$\frac{\partial L}{\partial x} = m\dot{\theta}^2x + mg\sin\theta \quad (8.31)$$

and the equations of motion can be written

$$\begin{aligned} [J_1 + J_2 + m(x^2 + R^2)]\ddot{\theta} \\ + \frac{1}{R}(J_2 + mR^2)\ddot{x} + 2mx\dot{x}\dot{\theta} &= mg(R\sin\theta + x\cos\theta) + \tau \end{aligned} \quad (8.32)$$

$$\frac{1}{R}(J_2 + mR^2)\ddot{\theta} + \left(m + \frac{J_2}{R^2}\right)\ddot{x} - m\dot{\theta}^2x = mg\sin\theta \quad (8.33)$$

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

$$\mathbf{M} \begin{pmatrix} \ddot{\theta} \\ \ddot{x} \end{pmatrix} = \begin{pmatrix} -2mx\dot{x}\dot{\theta} + mg(R\sin\theta + x\cos\theta) \\ m\dot{\theta}^2x + mg\sin\theta \end{pmatrix} + \begin{pmatrix} \tau \\ 0 \end{pmatrix} \quad (8.34)$$

has a positive definite and symmetric mass matrix

$$\mathbf{M} = \begin{pmatrix} J_1 + J_2 + m(x^2 + R^2) & \frac{1}{R}(J_2 + mR^2) \\ \frac{1}{R}(J_2 + mR^2) & m + \frac{J_2}{R^2} \end{pmatrix} \quad (8.35)$$

### 8.2.7 Furuta pendulum

The kinetic energy  $T$  and the potential energy  $U$  of the Furuta pendulum are given by

$$T = \frac{1}{2}(J_1 + mL_1^2 + mL_2^2 \sin^2 \theta_2)\dot{\theta}_1^2 + \frac{1}{2}mL_2^2\dot{\theta}_2^2 - mL_1L_2\dot{\theta}_1\dot{\theta}_2 \cos \theta_2 \quad (8.36)$$

$$U = mgL_2 \cos \theta_2 \quad (8.37)$$

which give the Lagrangian

$$L = \frac{1}{2}(J_1 + mL_1^2 + mL_2^2 \sin^2 \theta_2)\dot{\theta}_1^2 + \frac{1}{2}mL_2^2\dot{\theta}_2^2 - mL_1L_2\dot{\theta}_1\dot{\theta}_2 \cos \theta_2 - mgL_2 \cos \theta_2 \quad (8.38)$$

The partial derivatives are

$$\frac{\partial L}{\partial \dot{\theta}_1} = (J_1 + mL_1^2 + mL_2^2 \sin^2 \theta_2)\dot{\theta}_1 - mL_1L_2\dot{\theta}_2 \cos \theta_2 \quad (8.39)$$

$$\frac{\partial L}{\partial \dot{\theta}_2} = mL_2^2\dot{\theta}_2 - mL_1L_2\dot{\theta}_1 \cos \theta_2 \quad (8.40)$$

$$\frac{\partial L}{\partial \theta_1} = 0 \quad (8.41)$$

$$\frac{\partial L}{\partial \theta_2} = mL_2^2 \sin \theta_2 \cos \theta_2 \dot{\theta}_1^2 + mL_1L_2\dot{\theta}_1\dot{\theta}_2 \sin \theta_2 + mgL_2 \sin \theta_2 \quad (8.42)$$

and the equations of motion are found by evaluation

$$\begin{aligned} \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{\theta}_1} \right) - \frac{\partial L}{\partial \theta_1} &= (J_1 + mL_1^2 + mL_2^2 \sin^2 \theta_2)\ddot{\theta}_1 - mL_1L_2\ddot{\theta}_2 \cos \theta_2 \\ &\quad + 2mL_2^2\dot{\theta}_1\dot{\theta}_2 \sin \theta_2 \cos \theta_2 + mL_1L_2\dot{\theta}_2^2 \sin \theta_2 \end{aligned} \quad (8.43)$$

$$\begin{aligned} \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{\theta}_2} \right) - \frac{\partial L}{\partial \theta_2} &= mL_2^2\ddot{\theta}_2 - mL_1L_2\dot{\theta}_1 \cos \theta_2 + mL_1L_2\dot{\theta}_1\dot{\theta}_2 \sin \theta_2 \\ &\quad - mL_2^2 \sin \theta_2 \cos \theta_2 \dot{\theta}_1^2 - mL_1L_2\dot{\theta}_1\dot{\theta}_2 \sin \theta_2 \\ &\quad - mgL_2 \sin \theta_2 \end{aligned} \quad (8.44)$$

The equations of motion of the Furuta pendulum are

$$(J_1 + mL_1^2 + mL_2^2 \sin^2 \theta_2)\ddot{\theta}_1 - mL_1L_2\ddot{\theta}_2 \cos \theta_2 + 2mL_2^2\dot{\theta}_1\dot{\theta}_2 \sin \theta_2 \cos \theta_2 + mL_1L_2\dot{\theta}_2^2 \sin \theta_2 = \tau \quad (8.45)$$

$$mL_2^2\ddot{\theta}_2 - mL_1L_2\dot{\theta}_1 \cos \theta_2 - mL_2^2 \sin \theta_2 \cos \theta_2 \dot{\theta}_1^2 - mgL_2 \sin \theta_2 = 0 \quad (8.46)$$

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  $\tau_i$  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

$$T_i = \frac{1}{2}m_i(\mathbf{v}_{ci}^i)^T(\mathbf{v}_{ci}^i) + \frac{1}{2}(\boldsymbol{\omega}_{0i}^i)^T\mathbf{M}_{ci}^i\boldsymbol{\omega}_{0i}^i \quad (8.47)$$

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

$$\mathbf{v}_{ci}^i = \sum_{j=1}^i \mathbf{v}_{ci,j}^i(\mathbf{q}) \dot{q}_j = \mathbf{J}_{v_{ci}}(\mathbf{q}) \dot{\mathbf{q}} \quad (8.48)$$

$$\boldsymbol{\omega}_{0i}^i = \sum_{j=1}^i \boldsymbol{\omega}_{0i,j}^i(\mathbf{q}) \dot{q}_j = \mathbf{J}_{\omega_{0i}}(\mathbf{q}) \dot{\mathbf{q}} \quad (8.49)$$

Then the kinetic energy of link  $i$  can be written

$$T_i = \frac{1}{2}m_i \dot{\mathbf{q}}^T \mathbf{J}_{v_{ci}}^T(\mathbf{q}) \mathbf{J}_{v_{ci}}(\mathbf{q}) \dot{\mathbf{q}} + \frac{1}{2} \dot{\mathbf{q}}^T \mathbf{J}_{\omega_{0i}}^T(\mathbf{q}) \mathbf{M}_{ci}^i \mathbf{J}_{\omega_{0i}}(\mathbf{q}) \dot{\mathbf{q}} \quad (8.50)$$

and the total kinetic energy for the manipulator is

$$T = \frac{1}{2} \dot{\mathbf{q}}^T \sum_{i=1}^n [m_i \mathbf{J}_{v_{ci}}^T(\mathbf{q}) \mathbf{J}_{v_{ci}}(\mathbf{q}) + \mathbf{J}_{\omega_i}^T(\mathbf{q}) \mathbf{M}_{ci}^i \mathbf{J}_{\omega_{0i}}(\mathbf{q})] \quad (8.51)$$

This shows that the kinetic energy of the manipulator can be written as the quadratic form

$$T = \frac{1}{2} \dot{\mathbf{q}}^T \mathbf{M}(\mathbf{q}) \dot{\mathbf{q}} \quad (8.52)$$

where the  $n \times n$  mass matrix  $\mathbf{M}(\mathbf{q})$  given by

$$\mathbf{M}(\mathbf{q}) = \sum_{i=1}^n [m_i \mathbf{J}_{v_{ci}}^T(\mathbf{q}) \mathbf{J}_{v_{ci}}(\mathbf{q}) + \mathbf{J}_{\omega_i}^T(\mathbf{q}) \mathbf{M}_{ci}^i \mathbf{J}_{\omega_{0i}}(\mathbf{q})] \quad (8.53)$$

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

$$U(\mathbf{q}) = \sum_{i=1}^n U_i(\mathbf{q}) = \sum_{i=1}^n m_i \mathbf{g}^T \mathbf{r}_{ci}(\mathbf{q}) \quad (8.54)$$

The Lagrangian of the manipulator is therefore

$$L = \frac{1}{2} \dot{\mathbf{q}}^T \mathbf{M}(\mathbf{q}) \dot{\mathbf{q}} - U(\mathbf{q}) \quad (8.55)$$

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

$$\mathbf{M}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}} + \mathbf{g}(\mathbf{q}) = \boldsymbol{\tau} \quad (8.56)$$

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

$$\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}}) = \{c_{kj}\} = \left\{ \sum_{i=1}^n c_{ijk} \dot{q}_i \right\} \quad (8.57)$$

where

$$c_{ijk} := \frac{1}{2} \left( \frac{\partial m_{kj}}{\partial q_i} + \frac{\partial m_{ik}}{\partial q_j} - \frac{\partial m_{ij}}{\partial q_k} \right) \quad (8.58)$$

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

$$T = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n m_{ij}(\mathbf{q}) \dot{q}_i \dot{q}_j \quad (8.59)$$

for the kinetic energy, which gives the Lagrangian

$$L = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n m_{ij}(\mathbf{q}) \dot{q}_i \dot{q}_j - U(\mathbf{q}) \quad (8.60)$$

We find that

$$\begin{aligned} \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_k} &= \frac{d}{dt} \left( \frac{1}{2} \sum_{j=1}^n m_{kj} \dot{q}_j + \frac{1}{2} \sum_{i=1}^n m_{ik} \dot{q}_i \right) \\ &= \sum_{j=1}^n m_{kj}(\mathbf{q}) \ddot{q}_j + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \left( \frac{\partial m_{kj}}{\partial q_i} + \frac{\partial m_{ik}}{\partial q_j} \right) \dot{q}_i \dot{q}_j \end{aligned} \quad (8.61)$$

and that

$$\frac{\partial L}{\partial q_k} = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \frac{\partial m_{ij}}{\partial q_k} \dot{q}_i \dot{q}_j - \frac{\partial U}{\partial q_k} \quad (8.62)$$

The resulting equation of motion is

$$\sum_{j=1}^n m_{kj}(\mathbf{q}) \ddot{q}_j + \sum_{i=1}^n \sum_{j=1}^n c_{ijk}(\mathbf{q}) \dot{q}_i \dot{q}_j + g_k(\mathbf{q}) = \tau_k \quad (8.63)$$

where  $c_{ijk}$  are the Christoffel symbols of the first kind as defined by (8.58), and

$$g_k := \frac{\partial U}{\partial q_k} \quad (8.64)$$

Then the equation of motion (8.56) appears by defining the matrix

$$\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}}) = \{c_{kj}(\mathbf{q}, \dot{\mathbf{q}})\}, \quad c_{kj}(\mathbf{q}, \dot{\mathbf{q}}) = \sum_{i=1}^n c_{ijk}(\mathbf{q}) \dot{q}_i \quad (8.65)$$

and the gravity vector

$$\mathbf{g}(\mathbf{q}) = \frac{\partial U}{\partial \mathbf{q}} \quad (8.66)$$

Finally, we will show that the matrix

$$\mathbf{N} = \dot{\mathbf{M}} - 2\mathbf{C} \quad (8.67)$$

is skew symmetric. This is shown by considering element

$$n_{kj} = \dot{m}_{kj} - 2c_{kj} \quad (8.68)$$

of the matrix. We find that

$$\begin{aligned} \dot{m}_{kj} - 2c_{kj} &= \sum_{i=1}^n \left( \frac{\partial m_{kj}}{\partial q_i} - \frac{\partial m_{kj}}{\partial q_i} - \frac{\partial m_{ik}}{\partial q_j} + \frac{\partial m_{ij}}{\partial q_k} \right) \dot{q}_i \\ &= \sum_{i=1}^n \left( \frac{\partial m_{ij}}{\partial q_k} - \frac{\partial m_{ik}}{\partial q_j} \right) \dot{q}_i \end{aligned} \quad (8.69)$$

This implies

$$n_{kj} = -n_{jk} \quad (8.70)$$

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

$$\begin{aligned} \dot{E}(\mathbf{q}, \dot{\mathbf{q}}) &= \frac{d}{dt} \left( \frac{1}{2} \dot{\mathbf{q}}^T \mathbf{M}(\mathbf{q}) \dot{\mathbf{q}} \right) + \frac{\partial U}{\partial \mathbf{q}} \dot{\mathbf{q}} \\ &= \dot{\mathbf{q}}^T \mathbf{M}(\mathbf{q}) \ddot{\mathbf{q}} + \frac{1}{2} \dot{\mathbf{q}}^T \dot{\mathbf{M}}(\mathbf{q}) \dot{\mathbf{q}} + \frac{\partial U}{\partial \mathbf{q}} \dot{\mathbf{q}} \end{aligned} \quad (8.71)$$

The time derivative along the solutions of the system is found by inserting the equation of motion (8.56) and (8.66). This gives

$$\begin{aligned} \dot{E}(\mathbf{q}) &= \dot{\mathbf{q}}^T [-\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}}) \dot{\mathbf{q}} - \mathbf{g}(\mathbf{q}) + \boldsymbol{\tau}] + \frac{1}{2} \dot{\mathbf{q}}^T \dot{\mathbf{M}}(\mathbf{q}) \dot{\mathbf{q}} + \mathbf{g}(\mathbf{q})^T \dot{\mathbf{q}} \\ &= \dot{\mathbf{q}}^T \boldsymbol{\tau} + \frac{1}{2} \dot{\mathbf{q}}^T [\mathbf{M}(\mathbf{q}) - 2\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})] \dot{\mathbf{q}} \end{aligned} \quad (8.72)$$

Finally, the skew symmetry of  $\dot{\mathbf{M}} - 2\mathbf{C}$  gives the result

$$\dot{E}(\mathbf{q}) = \dot{\mathbf{q}}^T \boldsymbol{\tau} \quad (8.73)$$

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

$$V = \dot{\mathbf{q}}^T \boldsymbol{\tau} \quad (8.74)$$

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

$$T = \frac{1}{2}m_1\vec{v}_{c1} \cdot \vec{v}_{c1} + \frac{1}{2}m_2\vec{v}_{c2} \cdot \vec{v}_{c2} + \frac{1}{2}\vec{\omega}_1 \cdot \vec{M}_{1/c} \cdot \vec{\omega}_1 + \frac{1}{2}\vec{\omega}_2 \cdot \vec{M}_{2/c} \cdot \vec{\omega}_2 \quad (8.75)$$

This can be written

$$T = \frac{1}{2}m_{11}\dot{q}_1^2 + m_{12}\dot{q}_1\dot{q}_2 + \frac{1}{2}m_{22}\dot{q}_2^2 \quad (8.76)$$

where

$$m_{11} = I_{1z} + I_{2z} + m_1L_{c1}^2 + m_2(L_1^2 + L_{c2}^2 + 2L_1L_{c2}\cos q_2) \quad (8.77)$$

$$m_{12} = m_{21} = I_{2z} + m_2L_{c2}^2 + m_2L_1L_{c2}\cos q_2 \quad (8.78)$$

$$m_{22} = I_{2z} + m_2L_{c2}^2 \quad (8.79)$$

are the elements of the inertia matrix. The potential energy is

$$U = (m_1gL_{c1} + m_2gL_1)\sin q_1 + m_2gL_{c2}\sin(q_1 + q_2) \quad (8.80)$$

Then, from  $L = T - U$  the partial derivatives are found to be

$$\frac{\partial L}{\partial \dot{q}_1} = \frac{\partial T}{\partial \dot{q}_1} = m_{11}\dot{q}_1 + m_{12}\dot{q}_2 \quad (8.81)$$

$$\frac{\partial L}{\partial \dot{q}_2} = \frac{\partial T}{\partial \dot{q}_2} = m_{21}\dot{q}_1 + m_{22}\dot{q}_2 \quad (8.82)$$

$$\frac{\partial L}{\partial q_1} = \frac{\partial T}{\partial q_1} - \frac{\partial U}{\partial q_1} = -(m_1L_{c1} + m_2L_1)g\cos q_1 - m_2L_{c2}g\cos(q_1 + q_2) \quad (8.83)$$

$$\frac{\partial L}{\partial q_2} = \frac{\partial T}{\partial q_2} - \frac{\partial U}{\partial q_2} = \frac{1}{2}\frac{\partial m_{11}}{\partial q_2}\dot{q}_1^2 + \frac{\partial m_{21}}{\partial q_2}\dot{q}_1\dot{q}_2 - m_2gL_{c2}g\cos(q_1 + q_2) \quad (8.84)$$

The equations of motion are then found from (8.13) to be

$$m_{11}\ddot{q}_1 + m_{12}\ddot{q}_2 + \left(\frac{\partial m_{11}}{\partial q_2}\dot{q}_2\right)\dot{q}_1 + \left(\frac{\partial m_{12}}{\partial q_2}\dot{q}_2\right)\dot{q}_2 + \frac{\partial U}{\partial q_1} = \tau_1 \quad (8.85)$$

$$m_{21}\ddot{q}_1 + m_{22}\ddot{q}_2 + \left(\frac{\partial m_{21}}{\partial q_2}\dot{q}_2\right)\dot{q}_1 - \frac{1}{2}\frac{\partial m_{11}}{\partial q_2}\dot{q}_1^2 - \frac{\partial m_{21}}{\partial q_2}\dot{q}_1\dot{q}_2 + \frac{\partial U}{\partial q_2} = \tau_2 \quad (8.86)$$

which gives the equations of motion in the form

$$\begin{aligned} & (I_{1z} + I_{2z} + m_1L_{c1}^2 + m_2(L_1^2 + L_{c2}^2 + 2L_1L_{c2}\cos q_2))\ddot{q}_1 \\ & + (I_{2z} + m_2L_{c2}^2 + m_2L_1L_{c2}\cos q_2)\ddot{q}_2 \\ & - m_2L_1L_{c2}\sin q_2(2\dot{q}_1\dot{q}_2 + \dot{q}_2^2) \\ & + (m_1L_{c1} + m_2L_1)g\cos q_1 + m_2gL_{c2}\cos(q_1 + q_2) = \tau_1 \end{aligned} \quad (8.87)$$

$$(I_{2z} + m_2 L_{c2}^2 + m_2 L_1 L_{c2} \cos q_2) \ddot{q}_1 + (I_{2z} + m_2 L_{c2}^2) \ddot{q}_2 \\ + m_2 L_1 L_{c2} \dot{q}_1^2 \sin q_2 + m_2 L_{c2} g \cos(q_1 + q_2) = \tau_2 \quad (8.88)$$

### 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

$$\phi = \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} = \mathbf{A}\mathbf{q} \quad (8.89)$$

with associated generalized forces

$$\mathbf{K} = \begin{pmatrix} K_1 \\ K_2 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \tau_1 \\ \tau_2 \end{pmatrix} = \mathbf{A}^{-T} \boldsymbol{\tau} \quad (8.90)$$

as this gives

$$\mathbf{K}^T \dot{\phi} = \boldsymbol{\tau}^T \mathbf{A}^{-1} \mathbf{A} \dot{\mathbf{q}} = \boldsymbol{\tau}^T \dot{\mathbf{q}} \quad (8.91)$$

Note that  $\dot{\phi}_1 = \omega_1$  and  $\dot{\phi}_2 = \omega_2$ .

With the new set of generalized coordinates the kinetic energy is

$$T = \frac{1}{2} \dot{\mathbf{q}}^T \mathbf{M}(\mathbf{q}) \dot{\mathbf{q}} = \frac{1}{2} \dot{\phi}^T \mathbf{A}^{-T} \mathbf{M}(\mathbf{q}) \mathbf{A}^{-1} \dot{\phi} = \frac{1}{2} \dot{\phi}^T \mathbf{D}(\phi) \dot{\phi} \quad (8.92)$$

where the mass matrix  $\mathbf{D}(\phi) = \{d_{ij}(\phi)\}$  corresponding to the new coordinates  $\phi$  is found to be

$$\begin{aligned} \mathbf{D}(\phi) &= \mathbf{A}^{-T} \mathbf{M}(\mathbf{q}) \mathbf{A}^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \\ &= \begin{pmatrix} m_{11} - 2m_{12} + m_{22} & m_{12} - m_{22} \\ m_{12} - m_{22} & m_{22} \end{pmatrix} \end{aligned} \quad (8.93)$$

which gives

$$\mathbf{D}(\phi) = \begin{pmatrix} I_{1z} + m_1 L_{c1}^2 + m_2 L_1^2 & m_2 L_1 L_{c2} \cos(\phi_2 - \phi_1) \\ m_2 L_1 L_{c2} \cos(\phi_2 - \phi_1) & I_{2z} + m_2 L_{c2}^2 \end{pmatrix} \quad (8.94)$$

The equations of motion are then found from (8.13) to be

$$\begin{aligned} d_{11} \ddot{\phi}_1 + d_{22} \ddot{\phi}_2 + \left( \frac{\partial d_{12}}{\partial \phi_1} \dot{\phi}_1 + \frac{\partial d_{12}}{\partial \phi_2} \dot{\phi}_2 \right) \dot{\phi}_2 - \left( \frac{\partial d_{12}}{\partial \phi_1} \dot{\phi}_1 \dot{\phi}_2 + \frac{\partial U}{\partial \phi_1} \right) &= K_1 \\ d_{21} \ddot{\phi}_1 + d_{22} \ddot{\phi}_2 + \left( \frac{\partial d_{21}}{\partial \phi_1} \dot{\phi}_1 + \frac{\partial d_{21}}{\partial \phi_2} \dot{\phi}_2 \right) \dot{\phi}_1 - \left( \frac{\partial d_{21}}{\partial \phi_2} \dot{\phi}_1 \dot{\phi}_2 + \frac{\partial U}{\partial \phi_2} \right) &= K_2 \end{aligned}$$

to be

$$\begin{aligned} (I_{1z} + m_1 L_{c1}^2 + m_2 L_1^2) \ddot{\phi}_1 + m_2 L_1 L_{c2} \cos(\phi_2 - \phi_1) \ddot{\phi}_2 \\ - m_2 L_1 L_{c2} \sin(\phi_2 - \phi_1) \dot{\phi}_2^2 + (m_1 L_{c1} + m_2 L_1) g \cos \phi_1 &= K_1 \quad (8.95) \\ m_2 L_1 L_{c2} \cos(\phi_2 - \phi_1) \ddot{\phi}_1 + (I_{2z} + m_2 L_{c2}^2) \ddot{\phi}_2 \\ + m_2 L_1 L_{c2} \sin(\phi_2 - \phi_1) \dot{\phi}_1^2 + m_2 L_{c2} g \cos(\phi_2) &= K_2 \quad (8.96) \end{aligned}$$

### 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

$$T = \frac{1}{2} \boldsymbol{\omega}^T \mathbf{M} \boldsymbol{\omega} \quad (8.97)$$

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  $\psi$ ,  $\theta$  and  $\phi$ , that is,  $\mathbf{q} = (\phi, \theta, \psi)^T$ . Then the kinetic energy is found to be

$$T = \frac{1}{2} \dot{\mathbf{q}}^T \mathbf{E}_d^T(\mathbf{q}) \mathbf{M} \mathbf{E}_d(\mathbf{q}) \dot{\mathbf{q}} \quad (8.98)$$

where

$$\mathbf{E}_d(\mathbf{q}) = \begin{pmatrix} 1 & 0 & -\sin \theta \\ 0 & \cos \phi & \sin \phi \cos \theta \\ 0 & -\sin \phi & \cos \phi \cos \theta \end{pmatrix} \quad (8.99)$$

Lagrange's equation of motion is

$$\frac{d}{dt} (\mathbf{E}_d^T(\mathbf{q}) \mathbf{M} \mathbf{E}_d(\mathbf{q}) \dot{\mathbf{q}}) - \frac{\partial}{\partial \mathbf{q}} \left( \frac{1}{2} \dot{\mathbf{q}}^T \mathbf{E}_d^T(\mathbf{q}) \mathbf{M} \mathbf{E}_d(\mathbf{q}) \dot{\mathbf{q}} \right) = \mathbf{0} \quad (8.100)$$

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

$$\mathbf{M} \ddot{\boldsymbol{\omega}} + \boldsymbol{\omega}^\times \mathbf{M} \boldsymbol{\omega} = \mathbf{0} \quad (8.101)$$

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.

### 8.3.2 Variations versus differentials

We consider a continuous and differentiable function  $f(x_1, x_2, x_3)$ , and we would like to investigate the extremal points of the function, which are the points  $(x_1, x_2, x_3)$  where  $f$  has its minima or maxima. This will be done by finding the *stationary values* of the function  $f$ , which are the values  $f(x_1, x_2, x_3)$  at points  $(x_1, x_2, x_3)$  where the rate of change of  $f$  is zero. The usual technique in calculus is to find the differential

$$df = \frac{\partial f}{\partial x_1} dx_1 + \frac{\partial f}{\partial x_2} dx_2 + \frac{\partial f}{\partial x_3} dx_3 \quad (8.102)$$

and locate the stationary points as the points where  $df = 0$ . In mechanical systems it is customary to associate the differentials  $dx_i$  with the actual infinitesimal change of the variables  $x_i$ , and in accordance with this  $df$  denotes the infinitesimal change in the function  $f$  due to the infinitesimal change  $dx_i$ . The time derivative of  $f$  is found by dividing with  $dt$ , giving

$$\frac{df}{dt} = \frac{\partial f}{\partial x_1} \dot{x}_1 + \frac{\partial f}{\partial x_2} \dot{x}_2 + \frac{\partial f}{\partial x_3} \dot{x}_3 \quad (8.103)$$

where  $\dot{x}_i = \frac{dx_i}{dt}$  are the velocities.

As opposed to this, Lagrange introduced the concept *variations* or *virtual changes*  $\delta x_i$  in the variables  $x_i$  (Lanczos 1986), (Lovelock and Rund 1989). The variation  $\delta x_i$  is to be considered as a mathematical experiment without any change in the physical variable  $x_i$ . This makes sense as it should be possible to decide if a point  $(x_1, x_2, x_3)$  is an extremal point without moving the system around. The variation in  $f$  associated with the variation  $\delta x_i$  is

$$\delta f = \frac{\partial f}{\partial x_1} \delta x_1 + \frac{\partial f}{\partial x_2} \delta x_2 + \frac{\partial f}{\partial x_3} \delta x_3 \quad (8.104)$$

which reflects an infinitesimal change in  $f$  due to the mathematical experiment  $\delta x_i$  without any change being done in the variables  $x_i$ . Stationary values are then found when  $\delta f = 0$ , as would be expected. This distinction between the operator  $d$  and  $\delta$  may seem strange at first, but it turns out that this provides us with a number of very useful techniques for analysis of mechanical systems. Prominent examples of such techniques are d'Alembert's principle and Hamilton's principle. Extensive discussion on the concept of variations in dynamics is found in (Lanczos 1986).

### 8.3.3 The variation of a function

In this section we will present a number of useful results on variations. We start with the following definition:

Consider a continuous and differentiable function  $f(x)$ . Define the perturbed function

$$f(x, \alpha) = f(x) + \alpha \phi(x) \quad (8.105)$$

where  $\phi(x)$  is an arbitrary continuous and differentiable function. The variation of  $f$  at  $x$  is then defined as

$$\delta f(x) = \left. \frac{df(x, \alpha)}{d\alpha} \right|_{\alpha=0} = \phi(x) \quad (8.106)$$

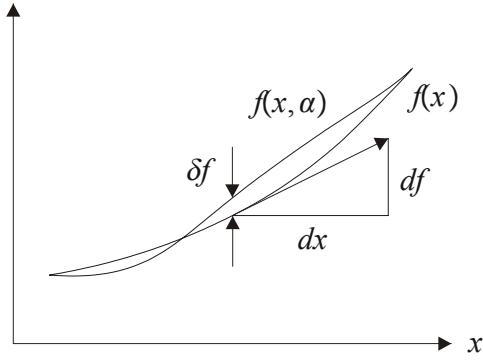


Figure 8.2: The variation  $\delta f$  and the differential  $df$  of a function  $f(x)$ .

The difference between  $df$  and  $\delta f(x)$ , which are both infinitesimal changes in  $f$ , is clear in this case as  $df$  is the infinitesimal change in the function  $f(x)$  due to an infinitesimal change  $dx$ , while  $\delta f(x)$  is the infinitesimal change due to the infinitesimal changes in  $f(x)$  to  $f(x, \epsilon)$  for the same  $x$ . This is shown in Figure 8.2.

Using the definition (8.106) it is straightforward to establish several results for the variation of a function. The variation of the derivative of the function is found from the derivative of  $f(x, \alpha)$ . Consider the function

$$f'(x, \alpha) = f'(x) + \alpha \phi'(x) \quad (8.107)$$

where  $(\cdot)'$  denotes the derivative with respect to  $x$ . The variation is

$$\delta f'(x) = \frac{df'(x, \alpha)}{d\alpha} \Big|_{\alpha=0} = \phi'(x) \quad (8.108)$$

We see that

$$\frac{d}{dx} [\delta f(x)] = \delta \left[ \frac{d}{dx} f(x) \right] \quad (8.109)$$

that is, the derivative of the variation is equal to the variation of the derivative.

The variation of the definite integral

$$I = \int_a^b f(x) dx \quad (8.110)$$

where  $a$  and  $b$  are constants is found from the function

$$I(\alpha) = \int_a^b f(x, \alpha) dx \quad (8.111)$$

The variation is given by

$$\begin{aligned} \delta I &= \frac{dI(\alpha)}{d\alpha} \Big|_{\alpha=0} = \frac{d}{d\alpha} \int_a^b f(x, \alpha) dx \Big|_{\alpha=0} = \int_a^b \frac{df(x, \alpha)}{d\alpha} \Big|_{\alpha=0} dx \\ &= \int_a^b \delta f(x) dx \end{aligned} \quad (8.112)$$

which shows that the variation of the integral is equal to the integral of the variation. We sum up that

The variation operation commutes with differentiation and integration in the sense that

$$\frac{d}{dx} [\delta f(x)] = \delta \left[ \frac{d}{dx} f(x) \right] \quad (8.113)$$

$$\delta \int_a^b f(x, \alpha) dx = \int_a^b \delta f(x) dx \quad (8.114)$$

### 8.3.4 The Euler-Lagrange equation for a general integral

To derive the Euler-Lagrange equation we consider the definite integral

$$I = \int_a^b f(y, y', x) dx \quad (8.115)$$

where

$$y = y(x), \quad y' = \frac{dy}{dx} \quad (8.116)$$

with boundary conditions

$$y(a) = y_a, \quad y(b) = y_b \quad (8.117)$$

At this point this is a purely mathematical exercise without any physical interpretation.

The integral  $I$  will depend on the curve that is defined by the function  $y = y(x)$ . We want to find the path where the integral has a stationary value with respect to curves that result from infinitesimal changes in the function  $y(x)$ . To this end we define the function

$$y(x, \alpha) = y(x) + \alpha\phi(x) \quad (8.118)$$

where  $\phi(x)$  is an arbitrary function that vanishes at the end-points, that is,  $\phi(a) = \phi(b) = 0$ . The variation in  $y$  is then defined as

$$\delta y(x) = \frac{dy(x, \alpha)}{d\alpha} \Big|_{\alpha=0} = \phi(x) \quad (8.119)$$

We note that derivation and variation commute, that is,

$$\frac{d}{dx} (\delta y) = \delta \left( \frac{dy}{dx} \right) \quad (8.120)$$

There is no point in introducing a variation in the variable  $x$ . Therefore  $\delta x = 0$  will always be used in this type of problem.

The variation of the definite integral is

$$\begin{aligned} \delta I &= \delta \int_a^b f(y, y', x) dx \\ &= \frac{d}{d\alpha} \int_a^b f[y(x, \alpha), y'(x, \alpha), x] dx \Big|_{\alpha=0} \\ &= \int_a^b \delta f [y(x, \alpha), y'(x, \alpha), x] dx \end{aligned} \quad (8.121)$$

where

$$\begin{aligned}\delta f[y(x), y'(x), x] &= \frac{d}{d\alpha} f[y(x, \epsilon), y'(x, \epsilon), x] \Big|_{\alpha=0} \\ &= \frac{\partial f}{\partial y} \delta y + \frac{\partial f}{\partial y'} \delta y'\end{aligned}\quad (8.122)$$

This gives

$$\delta I = \int_a^b \left( \frac{\partial f}{\partial y} \delta y + \frac{\partial f}{\partial y'} \delta y' \right) dx \quad (8.123)$$

Then the standard technique is to use partial integration of the second term in the integrand. We see that

$$\int_a^b \frac{\partial f}{\partial y'} \delta y' dx = \left( \frac{\partial f}{\partial y'} \right) \delta y \Big|_a^b - \int_a^b \frac{d}{dx} \left( \frac{\partial f}{\partial y'} \right) \delta y dx = - \int_a^b \frac{d}{dx} \left( \frac{\partial f}{\partial y'} \right) \delta y dx \quad (8.124)$$

because the variation vanishes at the end-points. This gives

$$\delta I = \int_a^b \left[ \frac{\partial f}{\partial y} - \frac{d}{dx} \left( \frac{\partial f}{\partial y'} \right) \right] \delta y dx \quad (8.125)$$

Since  $\delta y(x)$  is arbitrary for all  $x$ , we see that  $\delta I = 0$  implies the Euler-Lagrange equation

$$\frac{d}{dx} \left( \frac{\partial f}{\partial y'} \right) - \frac{\partial f}{\partial y} = 0 \quad (8.126)$$

### 8.3.5 The variation of the rotation matrix

It is not immediately obvious how to define the variation of a rotation matrix. However, we will see in this section that the definition can be formulated in analogy with the definition (8.106) for a function. A rigid body has orientation given by  $\mathbf{R} \in SO(3)$ . The time derivative of  $\mathbf{R}$  is

$$\dot{\mathbf{R}} = \mathbf{R} \boldsymbol{\omega}^\times \quad (8.127)$$

where  $\boldsymbol{\omega}$  is the angular velocity in body coordinates. The differential of the rotation matrix can be written

$$d\mathbf{R} = \mathbf{R} \boldsymbol{\omega}^\times dt \quad (8.128)$$

The variation  $\delta\mathbf{R}$  of the rotation matrix is found by considering the perturbation (Marsden and Ratiu 1994, p. 390)

$$\mathbf{R}(\alpha) = \mathbf{R} \exp(\alpha \boldsymbol{\sigma}^\times) \quad (8.129)$$

where  $\boldsymbol{\sigma}$  is an arbitrary three-dimensional vector in body coordinates, and  $\boldsymbol{\sigma}^\times$  is the corresponding skew-symmetric form. This means that  $\mathbf{R}(\alpha)$  is the composite rotation of  $\mathbf{R}$  and a rotation about the axis defined by  $\boldsymbol{\sigma}$ . The variation in  $\mathbf{R}$  is then defined as

$$\delta\mathbf{R} = \frac{d}{d\alpha} \mathbf{R} \exp(\alpha \boldsymbol{\sigma}^\times) \Big|_{\alpha=0} \quad (8.130)$$

which gives the result

$$\delta\mathbf{R} = \mathbf{R} \boldsymbol{\sigma}^\times \quad (8.131)$$

We note that variation and time differentiation commutes as

$$\frac{d}{dt}(\delta\mathbf{R}) = \frac{d}{dt} \left[ \frac{d}{d\alpha} [\mathbf{R} \exp(\alpha\boldsymbol{\sigma}^\times)] \Big|_{\alpha=0} \right] = \frac{d}{d\alpha} \left[ \frac{d}{dt} (\mathbf{R} \exp(\alpha\boldsymbol{\sigma}^\times)) \right] \Big|_{\alpha=0} = \delta\dot{\mathbf{R}} \quad (8.132)$$

We will now derive the relation between the variation  $\delta\boldsymbol{\omega}^\times$  of the angular velocity  $\boldsymbol{\omega}^\times = \mathbf{R}^T \dot{\mathbf{R}}$  and  $\boldsymbol{\sigma}^\times = \mathbf{R}^T \delta\mathbf{R}$ . First we note that since  $\mathbf{R}^T \mathbf{R} = \mathbf{I}$  we have

$$\mathbf{0} = \delta\mathbf{I} = \delta(\mathbf{R}^T \mathbf{R}) = \delta(\mathbf{R}^T) \mathbf{R} + \mathbf{R}^T \delta\mathbf{R} \Rightarrow \delta(\mathbf{R}^T) = -\mathbf{R}^T \delta\mathbf{R} \mathbf{R}^T \quad (8.133)$$

In the same way we find that

$$\frac{d}{dt}(\mathbf{R}^T) = -\mathbf{R}^T \dot{\mathbf{R}} \mathbf{R}^T \quad (8.134)$$

Consider

$$\delta\boldsymbol{\omega}^\times = \delta(\mathbf{R}^T) \dot{\mathbf{R}} + \mathbf{R}^T \delta\dot{\mathbf{R}} = -\mathbf{R}^T \delta\mathbf{R} \mathbf{R}^T \dot{\mathbf{R}} + \mathbf{R}^T \delta\dot{\mathbf{R}} = -\boldsymbol{\sigma}^\times \boldsymbol{\omega}^\times + \mathbf{R}^T \delta\dot{\mathbf{R}}$$

and

$$\frac{d}{dt}\boldsymbol{\sigma}^\times = \frac{d}{dt}(\mathbf{R}^T) \delta\mathbf{R} + \mathbf{R}^T \frac{d}{dt}(\delta\mathbf{R}) = -\mathbf{R}^T \dot{\mathbf{R}} \mathbf{R}^T \delta\mathbf{R} + \mathbf{R}^T \delta\dot{\mathbf{R}} = -\boldsymbol{\omega}^\times \boldsymbol{\sigma}^\times + \mathbf{R}^T \delta\dot{\mathbf{R}}$$

This gives

$$\delta\boldsymbol{\omega}^\times = \frac{d}{dt}\boldsymbol{\sigma}^\times + \boldsymbol{\omega}^\times \boldsymbol{\sigma}^\times - \boldsymbol{\sigma}^\times \boldsymbol{\omega}^\times = \frac{d}{dt}\boldsymbol{\sigma}^\times + (\boldsymbol{\omega}^\times \boldsymbol{\sigma})^\times \quad (8.135)$$

where (6.33) is used. The vector form of this is

$$\delta\boldsymbol{\omega} = \frac{d}{dt}\boldsymbol{\sigma} + \boldsymbol{\omega}^\times \boldsymbol{\sigma} \quad (8.136)$$

To sum up:

The variation of the rotation matrix can be defined by

$$\delta\mathbf{R} = \mathbf{R}\boldsymbol{\sigma}^\times \quad (8.137)$$

where  $\boldsymbol{\sigma}^\times$  is the skew symmetric form of a vector  $\boldsymbol{\sigma}$  in body coordinates. The variation of the angular velocity in body coordinates is

$$\delta\boldsymbol{\omega} = \frac{d}{dt}\boldsymbol{\sigma} + \boldsymbol{\omega}^\times \boldsymbol{\sigma} \quad (8.138)$$

**Example 137** The time derivative of the rotation matrix  $\mathbf{R}$  at the identity  $\mathbf{I}$  is

$$\frac{d\mathbf{R}}{dt} \Big|_{\mathbf{R}=\mathbf{I}} = \boldsymbol{\omega}^\times \in so(3) \quad (8.139)$$

The set  $so(3)$  of skew symmetric matrices is the Lie algebra of  $SO(3)$ . Thus the matrix forms  $\boldsymbol{\omega}^\times$  and  $\boldsymbol{\sigma}^\times$  are both in the Lie algebra  $so(3)$ . The Lie bracket in  $so(3)$  is (Arnold 1989), (Marsden and Ratiu 1994), (Murray et al. 1994)

$$[\boldsymbol{\omega}^\times, \boldsymbol{\sigma}^\times] := \boldsymbol{\omega}^\times \boldsymbol{\sigma}^\times - \boldsymbol{\sigma}^\times \boldsymbol{\omega}^\times = (\boldsymbol{\omega}^\times \boldsymbol{\sigma})^\times \quad (8.140)$$

The matrix  $\mathbf{ad}_\omega$  is defined from

$$[\boldsymbol{\omega}^\times, \boldsymbol{\sigma}^\times] = (\mathbf{ad}_\omega \boldsymbol{\sigma})^\times \quad (8.141)$$

From these two equations it is seen that

$$\mathbf{ad}_\omega = \boldsymbol{\omega}^\times \quad (8.142)$$

which gives

$$\delta\boldsymbol{\omega} = \frac{d}{dt}\boldsymbol{\sigma} + \mathbf{ad}_\omega \boldsymbol{\sigma} \quad (8.143)$$

### 8.3.6 The variation of the homogeneous transformation matrix

A rigid body has configuration given by

$$\mathbf{T} = \begin{pmatrix} \mathbf{R} & \mathbf{r} \\ 0 & 1 \end{pmatrix} \in SE(3) \quad (8.144)$$

The time derivative of  $\mathbf{T}$  can be written in the form

$$\dot{\mathbf{T}} = \mathbf{T}\hat{\mathbf{w}} \quad (8.145)$$

where

$$\hat{\mathbf{w}} = \begin{pmatrix} \boldsymbol{\omega}^\times & \mathbf{v} \\ 0 & 0 \end{pmatrix} \quad (8.146)$$

is the  $4 \times 4$  matrix representation of the vector  $\mathbf{w} = (\mathbf{v}^T, \boldsymbol{\omega}^T)^T$ .

The variation of  $\mathbf{T}$  is

$$\delta\mathbf{T} = \begin{pmatrix} \delta\mathbf{R} & \delta\mathbf{r} \\ 0 & 0 \end{pmatrix} \quad (8.147)$$

This can be written

$$\delta\mathbf{T} = \mathbf{T}\hat{\boldsymbol{\eta}} \quad (8.148)$$

where

$$\hat{\boldsymbol{\eta}} = \begin{pmatrix} \boldsymbol{\sigma}^\times & \delta\mathbf{r} \\ 0 & 0 \end{pmatrix} \quad (8.149)$$

is the matrix form of the vector  $\boldsymbol{\eta} = (\delta\mathbf{r}^T, \boldsymbol{\sigma}^T)^T$ .

In analogy with the derivation for  $SO(3)$  we find that

$$\delta\hat{\mathbf{w}} = \delta(\mathbf{T}^{-1})\dot{\mathbf{T}} + \mathbf{T}^{-1}\delta\dot{\mathbf{T}} = -\mathbf{T}^{-1}\delta\mathbf{T}\mathbf{T}^{-1}\dot{\mathbf{T}} + \mathbf{T}^{-1}\delta\dot{\mathbf{T}} = -\hat{\boldsymbol{\eta}}\hat{\mathbf{w}} + \mathbf{T}^{-1}\delta\dot{\mathbf{T}}$$

and

$$\frac{d}{dt}\hat{\boldsymbol{\eta}} = \frac{d}{dt}(\mathbf{T}^{-1})\delta\mathbf{T} + \mathbf{T}^{-1}\frac{d}{dt}(\delta\mathbf{T}) = -\mathbf{T}^{-1}\dot{\mathbf{T}}\mathbf{T}^{-1}\delta\mathbf{T} + \mathbf{T}^{-1}\delta\dot{\mathbf{T}} = -\hat{\mathbf{w}}\hat{\boldsymbol{\eta}} + \mathbf{T}^{-1}\delta\dot{\mathbf{T}}$$

and we may conclude that the variation of the six-dimensional velocity vector  $\mathbf{w}$  is related to the time derivative of the variation  $\boldsymbol{\eta}$  by

$$\delta\hat{\mathbf{w}} = \frac{d}{dt}\hat{\boldsymbol{\eta}} + \hat{\mathbf{w}}\hat{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}\hat{\mathbf{w}} \quad (8.150)$$

The last two terms are evaluated by the computation

$$\begin{aligned}
 \hat{\mathbf{w}}\hat{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}\hat{\mathbf{w}} &= \begin{pmatrix} \boldsymbol{\omega}^\times & \mathbf{v} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\sigma}^\times & \delta\mathbf{r} \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} \boldsymbol{\sigma}^\times & \delta\mathbf{r} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\omega}^\times & \mathbf{v} \\ 0 & 0 \end{pmatrix} \\
 &= \begin{pmatrix} \boldsymbol{\omega}^\times\boldsymbol{\sigma}^\times - \boldsymbol{\sigma}^\times\boldsymbol{\omega}^\times & \boldsymbol{\omega}^\times\delta\mathbf{r} - \boldsymbol{\sigma}^\times\mathbf{v} \\ 0 & 0 \end{pmatrix} \\
 &= \begin{pmatrix} (\boldsymbol{\omega}^\times\boldsymbol{\sigma})^\times & \boldsymbol{\omega}^\times\delta\mathbf{r} + \mathbf{v}^\times\boldsymbol{\sigma} \\ 0 & 0 \end{pmatrix} \\
 &= \widehat{\mathbf{ad}_w\boldsymbol{\eta}}
 \end{aligned} \tag{8.151}$$

where  $\mathbf{ad}_w$  is the matrix

$$\mathbf{ad}_w := \begin{pmatrix} \boldsymbol{\omega}^\times & \mathbf{v}^\times \\ 0 & \boldsymbol{\omega}^\times \end{pmatrix} \tag{8.152}$$

The vector form corresponding to (8.150) is therefore

$$\delta\mathbf{w} = \frac{d}{dt}\boldsymbol{\eta} + \mathbf{ad}_w\boldsymbol{\eta} \tag{8.153}$$

The variation of the homogeneous transformation matrix can be defined by

$$\delta\mathbf{T} = \mathbf{T}\hat{\boldsymbol{\eta}} \tag{8.154}$$

where the vector  $\boldsymbol{\eta}$  is in body coordinates. Then the variation of the velocity  $\mathbf{w}$  in body coordinates is

$$\delta\mathbf{w} = \frac{d}{dt}\boldsymbol{\eta} + \mathbf{ad}_w\boldsymbol{\eta} \tag{8.155}$$

**Remark 2** The time derivative of the homogeneous transformation matrix  $\mathbf{T}$  at the identity  $\mathbf{I}$  is

$$\left. \frac{d\mathbf{T}}{dt} \right|_{\mathbf{T}=\mathbf{I}} = \hat{\mathbf{w}} \in se(3) \tag{8.156}$$

The set  $se(3)$  is the set of matrices of the form  $\hat{\mathbf{w}}$  as defined in (8.146). The set  $se(3)$  is the Lie algebra of  $SE(3)$ . The matrix forms  $\hat{\mathbf{w}}$  and  $\hat{\boldsymbol{\eta}}$  are in the Lie algebra  $se(3)$ . The Lie bracket in  $se(3)$  is (Arnold 1989), (Marsden and Ratiu 1994), (Murray et al. 1994)

$$[\hat{\mathbf{w}}, \hat{\boldsymbol{\eta}}] = \hat{\mathbf{w}}\hat{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}\hat{\mathbf{w}} \tag{8.157}$$

The matrix  $\mathbf{ad}_w$  is defined by

$$[\hat{\mathbf{w}}, \hat{\boldsymbol{\eta}}] = \widehat{\mathbf{ad}_w\boldsymbol{\eta}} \tag{8.158}$$

We see from (8.150) that this agrees with the expression (8.152) for  $\mathbf{ad}_w$ .

## 8.4 The adjoint formulation

### 8.4.1 Introduction

In the previous section the **ad** operator was used in  $SO(3)$  and  $SE(3)$ . To make the presentation complete we include a brief presentation of the **Ad** and **ad** operators and the Lie bracket on  $SO(3)$  and  $SE(3)$  in this section.

### 8.4.2 Rotations

The configuration in  $SO(3)$  is given by  $\mathbf{R} \in SO(3)$ . The time derivative is  $\dot{\mathbf{R}} = \mathbf{R}\boldsymbol{\omega}^\times = \boldsymbol{\Omega}^\times\mathbf{R}$  where  $\boldsymbol{\Omega} = \mathbf{R}\boldsymbol{\omega}$ . The skew-symmetric forms of  $\boldsymbol{\omega}$  and  $\boldsymbol{\Omega}$  are related by

$$\boldsymbol{\Omega}^\times = \mathbf{R}\boldsymbol{\omega}^\times\mathbf{R}^T \quad (8.159)$$

The mapping from  $\boldsymbol{\omega}$  to  $\boldsymbol{\Omega}$  can also be written in terms of vectors in the adjoint representation

$$\boldsymbol{\Omega} = \mathbf{Ad}_R\boldsymbol{\omega} \quad (8.160)$$

where the matrix  $\mathbf{Ad}_R$  is the *adjoint transformation* on  $SO(3)$ . It is clear that

$$\mathbf{Ad}_R = \mathbf{R} \quad (8.161)$$

Consider a vector  $\mathbf{v}$  in the moving coordinate frame, and let  $\mathbf{V} = \mathbf{R}\mathbf{v}$  be the vector in the fixed frame. Then

$$\mathbf{V}^\times = \mathbf{R}\mathbf{v}^\times\mathbf{R}^T, \quad \mathbf{V} = \mathbf{Ad}_R\mathbf{v} \quad (8.162)$$

The time derivative is

$$\begin{aligned} \frac{d\mathbf{V}^\times}{dt} &= \mathbf{R} \left( \frac{d\mathbf{v}^\times}{dt} \right) \mathbf{R}^T + \dot{\mathbf{R}}\mathbf{v}^\times\mathbf{R}^T + \mathbf{R}\mathbf{v}^\times\dot{\mathbf{R}}^T \\ &= \mathbf{R} \left( \frac{d\mathbf{v}^\times}{dt} \right) \mathbf{R}^T + \mathbf{R}\boldsymbol{\omega}^\times\mathbf{v}^\times\mathbf{R}^T - \mathbf{R}\mathbf{v}^\times\boldsymbol{\omega}^\times\dot{\mathbf{R}}^T \\ &= \mathbf{R} \left( \frac{d\mathbf{v}^\times}{dt} \right) \mathbf{R}^T + \mathbf{R} (\boldsymbol{\omega}^\times\mathbf{v}^\times - \mathbf{v}^\times\boldsymbol{\omega}^\times) \dot{\mathbf{R}}^T \\ &= \mathbf{R} \left( \frac{d\mathbf{v}^\times}{dt} \right) \mathbf{R}^T + \mathbf{R} [\boldsymbol{\omega}^\times, \mathbf{v}^\times] \dot{\mathbf{R}}^T \end{aligned} \quad (8.163)$$

where

$$[\boldsymbol{\omega}^\times, \mathbf{v}^\times] = \boldsymbol{\omega}^\times\mathbf{v}^\times - \mathbf{v}^\times\boldsymbol{\omega}^\times = (\boldsymbol{\omega}^\times\mathbf{v})^\times \quad (8.164)$$

which is the *Lie bracket* on  $SO(3)$ . In vector form the equation (8.163) is written

$$\frac{d\mathbf{V}}{dt} = \mathbf{R} \frac{d\mathbf{v}}{dt} + \mathbf{R}\boldsymbol{\omega}^\times\mathbf{v} \quad (8.165)$$

However, time differentiation of  $\mathbf{V} = \mathbf{Ad}_R\mathbf{v}$  gives

$$\frac{d\mathbf{V}}{dt} = \mathbf{Ad}_R \frac{d\mathbf{v}}{dt} + \frac{d(\mathbf{Ad}_R)}{dt} \mathbf{v} \quad (8.166)$$

which implies

$$\frac{d(\mathbf{Ad}_R)}{dt} = \mathbf{R}\boldsymbol{\omega}^\times = \mathbf{Ad}_R\boldsymbol{\omega}^\times \quad (8.167)$$

We now define the operator  $\mathbf{ad}_\omega$  according to

$$\frac{d(\mathbf{Ad}_R)}{dt} = \mathbf{Ad}_R \mathbf{ad}_\omega \quad (8.168)$$

and it follows that in  $SO(3)$  we have

$$\mathbf{ad}_\omega = \boldsymbol{\omega}^\times \quad (8.169)$$

This implies that

$$(\mathbf{ad}_\omega\mathbf{v})^\times = [\boldsymbol{\omega}^\times, \mathbf{v}^\times] \quad (8.170)$$

which means that  $\mathbf{ad}_\omega\mathbf{v}$  is a vector representation of the Lie bracket  $[\boldsymbol{\omega}^\times, \mathbf{v}^\times]$ . Intuitively we can think of  $\mathbf{ad}_\omega\mathbf{v}$  as the rate of change in  $\mathbf{v}$  due to the motion due to the angular velocity  $\boldsymbol{\omega}$ , which is the directional derivative of  $\mathbf{v}$  in the direction of  $\boldsymbol{\omega}$ .

### 8.4.3 Rigid motion

The configuration in  $SE(3)$  is given by

$$\mathbf{T} = \begin{pmatrix} \mathbf{R} & \mathbf{r} \\ 0 & 1 \end{pmatrix} \in SE(3) \quad (8.171)$$

with inverse

$$\mathbf{T}^{-1} = \begin{pmatrix} \mathbf{R}^T & -\mathbf{R}^T \mathbf{r} \\ 0 & 1 \end{pmatrix} \in SE(3) \quad (8.172)$$

Consider the vector

$$\mathbf{W} = \begin{pmatrix} \mathbf{V} + \mathbf{r}^\times \boldsymbol{\Omega} \\ \boldsymbol{\Omega} \end{pmatrix} \quad (8.173)$$

in inertial coordinates and

$$\mathbf{w} = \begin{pmatrix} \mathbf{v} \\ \boldsymbol{\omega} \end{pmatrix} \quad (8.174)$$

in body-fixed coordinates. The corresponding matrix forms in  $se(3)$  are given by

$$\hat{\mathbf{W}} = \begin{pmatrix} \boldsymbol{\Omega}^\times & \mathbf{V} + \mathbf{r}^\times \boldsymbol{\Omega} \\ 0 & 0 \end{pmatrix} \in se(3), \quad \hat{\mathbf{w}} = \begin{pmatrix} \boldsymbol{\omega}^\times & \mathbf{v} \\ 0 & 0 \end{pmatrix} \in se(3) \quad (8.175)$$

The derivative of  $\mathbf{T}$  is

$$\dot{\mathbf{T}} = \begin{pmatrix} \mathbf{R}\boldsymbol{\omega}^\times & \mathbf{R}\mathbf{v} \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \mathbf{R} & \mathbf{r} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{\omega}^\times & \mathbf{v} \\ 0 & 0 \end{pmatrix} = \mathbf{T}\hat{\mathbf{w}} \quad (8.176)$$

or

$$\dot{\mathbf{T}} = \begin{pmatrix} \boldsymbol{\Omega}^\times \mathbf{R} & \mathbf{v} \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Omega}^\times & \mathbf{V} + \mathbf{r}^\times \boldsymbol{\Omega} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{R} & \mathbf{r} \\ 0 & 1 \end{pmatrix} = \hat{\mathbf{W}}\mathbf{T} \quad (8.177)$$

We see that

$$\hat{\mathbf{W}} = \mathbf{T}\hat{\mathbf{w}}\mathbf{T}^{-1} \quad (8.178)$$

In vector form we get

$$\mathbf{W} = \mathbf{Ad}_T \mathbf{w} \quad (8.179)$$

where the adjoint transformation on  $SE(3)$  is given by

$$\mathbf{Ad}_T = \begin{pmatrix} \mathbf{R} & \mathbf{r}^\times \mathbf{R} \\ 0 & \mathbf{R} \end{pmatrix} \quad (8.180)$$

with inverse

$$(\mathbf{Ad}_T)^{-1} = \begin{pmatrix} \mathbf{R}^T & -\mathbf{R}^T \mathbf{r}^\times \\ 0 & \mathbf{R}^T \end{pmatrix} \quad (8.181)$$

The time derivative of  $\mathbf{Ad}_T$  is

$$\begin{aligned} \frac{d}{dt} (\mathbf{Ad}_T) &= \begin{pmatrix} \mathbf{R}\boldsymbol{\omega}^\times & \mathbf{R}\mathbf{v}^\times + \mathbf{r}^\times \mathbf{R}\boldsymbol{\omega}^\times \\ 0 & \mathbf{R}\boldsymbol{\omega}^\times \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{R} & \mathbf{r}^\times \mathbf{R} \\ 0 & \mathbf{R} \end{pmatrix} \begin{pmatrix} \boldsymbol{\omega}^\times & \mathbf{v}^\times \\ 0 & \boldsymbol{\omega}^\times \end{pmatrix} \\ &= \mathbf{Ad}_T \mathbf{ad}_w \end{aligned} \quad (8.182)$$

where  $\mathbf{ad}_w$  is defined by

$$\frac{d}{dt} (\mathbf{Ad}_T) = \mathbf{Ad}_T \mathbf{ad}_w \quad (8.183)$$

We see that

$$\mathbf{ad}_w = \begin{pmatrix} \boldsymbol{\omega}^\times & \mathbf{v}^\times \\ 0 & \boldsymbol{\omega}^\times \end{pmatrix} \quad (8.184)$$

Let  $\mathbf{w}$  be an a six dimensional vector, and let  $\mathbf{W} = \mathbf{Ad}_T \mathbf{w}$ . Then

$$\hat{\mathbf{W}} = \mathbf{T} \hat{\mathbf{w}} \mathbf{T}^{-1} \quad (8.185)$$

and time differentiation gives

$$\begin{aligned} \frac{d\hat{\mathbf{W}}}{dt} &= \dot{\mathbf{T}} \frac{d\hat{\mathbf{w}}}{dt} \mathbf{T}^{-1} + \dot{\mathbf{T}} \hat{\mathbf{w}} \mathbf{T}^{-1} + \mathbf{T} \hat{\mathbf{w}} \dot{\mathbf{T}}^{-1} \\ &= \dot{\mathbf{T}} \frac{d\hat{\mathbf{w}}}{dt} \mathbf{T}^{-1} + \mathbf{T} \hat{\mathbf{u}} \hat{\mathbf{w}} \mathbf{T}^{-1} - \mathbf{T} \hat{\mathbf{w}} \hat{\mathbf{u}} \mathbf{T}^{-1} \\ &= \dot{\mathbf{T}} \frac{d\hat{\mathbf{w}}}{dt} \mathbf{T}^{-1} + \mathbf{T} (\hat{\mathbf{u}} \hat{\mathbf{w}} - \hat{\mathbf{w}} \hat{\mathbf{u}}) \mathbf{T}^{-1} \\ &= \dot{\mathbf{T}} \frac{d\hat{\mathbf{w}}}{dt} \mathbf{T}^{-1} + \mathbf{T} [\hat{\mathbf{u}}, \hat{\mathbf{w}}] \mathbf{T}^{-1} \end{aligned} \quad (8.186)$$

where

$$[\hat{\mathbf{u}}, \hat{\mathbf{w}}] = \hat{\mathbf{u}} \hat{\mathbf{w}} - \hat{\mathbf{w}} \hat{\mathbf{u}} \quad (8.187)$$

is the Lie bracket in  $se(3)$ . In adjoint form the time derivative is

$$\frac{d\mathbf{W}}{dt} = \mathbf{Ad}_T \left( \frac{d\mathbf{W}}{dt} + \mathbf{ad}_u \mathbf{w} \right) \quad (8.188)$$

Comparing the two expressions, we find that

$$\widehat{\mathbf{ad}_u \mathbf{w}} = [\hat{\mathbf{u}}, \hat{\mathbf{w}}] \quad (8.189)$$

As in  $SO(3)$  we see that  $\mathbf{ad}_u \mathbf{w}$  is a vector form of the Lie bracket  $[\hat{\mathbf{u}}, \hat{\mathbf{w}}]$ . Much in the same way as in  $SO(3)$  the intuitive interpretation of  $\mathbf{ad}_u \mathbf{w}$  is that it is the rate of change in  $\mathbf{w}$  due to the motion induced by the velocity vector  $\mathbf{u}$ .

**Remark 3** *The physical interpretation of  $\mathbf{U} = \mathbf{Ad}_T \mathbf{u}$  where  $\mathbf{u} = (\mathbf{v} \ \boldsymbol{\omega})^T$  is not quite straightforward. However, this is not a problem as this vector, which is called the spatial velocity vector, is not widely used. We see from  $\mathbf{U} = (\mathbf{R}\mathbf{v} - \boldsymbol{\Omega}^\times \mathbf{r} \ \mathbf{R}\boldsymbol{\omega})^T$  that the velocity  $\mathbf{R}\mathbf{v} - \boldsymbol{\Omega}^\times \mathbf{r}$  is the velocity of a point of the rigid body which is at the origin of the fixed frame. The angular velocity is simply the angular velocity of the rigid body in the coordinates of the fixed frame.*

## 8.5 The Euler-Poincaré equation

### 8.5.1 A central equation

Consider a rigid body  $b$  where the mass element  $dm$  has position  $\mathbf{r}$  in inertial coordinates. The externally applied force on mass element  $dm$  is  $d\mathbf{f}$ , and the force of constraint on the mass element is denoted  $d\mathbf{f}^{(c)}$ . Newton's law for a mass element  $dm$  is

$$\ddot{\mathbf{r}} dm - d\mathbf{f} - d\mathbf{f}^{(c)} = \mathbf{0} \quad (8.190)$$

The virtual displacement of the mass element  $dm$  is denoted  $\delta\mathbf{r}$ . We take the scalar product between the virtual displacement  $\delta\mathbf{r}$  and Newton's law, and integrate the result over the rigid body. This gives the following equation of motion:

$$\int_b \left( \ddot{\mathbf{r}} dm - d\mathbf{f} - d\mathbf{f}^{(c)} \right)^T \delta\mathbf{r} = 0 \quad (8.191)$$

The total kinetic energy of body  $b$  is

$$T = \frac{1}{2} \int_b \dot{\mathbf{r}}^T dm \dot{\mathbf{r}} \quad (8.192)$$

We note that the variation of the kinetic energy expressed in terms of  $\dot{\mathbf{r}}$  is

$$\delta T = \int_b \dot{\mathbf{r}}^T dm \delta\dot{\mathbf{r}} \quad (8.193)$$

Moreover, we define the virtual work

$$W_\delta := \int_b d\mathbf{f}^T \delta\mathbf{r} \quad (8.194)$$

To simplify the expression the product rule for differentiation is used to arrive at

$$\int_b \ddot{\mathbf{r}}^T dm \delta\mathbf{r} = \frac{d}{dt} \int_b \dot{\mathbf{r}}^T dm \delta\mathbf{r} - \int_b \dot{\mathbf{r}}^T dm \delta\dot{\mathbf{r}} = \frac{d}{dt} \int_b \dot{\mathbf{r}}^T dm \delta\mathbf{r} - \delta T \quad (8.195)$$

This results in the equation of motion in the form

$$\frac{d}{dt} \int_b \dot{\mathbf{r}}^T dm \delta\mathbf{r} - \delta T - W_\delta - \int_b \left( d\mathbf{f}^{(c)} \right)^T \delta\mathbf{r} = 0 \quad (8.196)$$

where the forces of constraint  $d\mathbf{f}^{(c)}$  still appears.

Next a change of variables is introduced. The main difference from the usual Lagrange formulation is that we do not necessarily use generalized coordinates. Instead a generalized speed vector  $\mathbf{u}$  is introduced, where the velocity  $\dot{\mathbf{r}}$  is affine in the components of a vector  $\mathbf{u}$ . The motivation for this is that this allows us to work with the rotation matrices and the angular velocity in  $SO(3)$  and  $SE(3)$ . The velocity is expressed by

$$\dot{\mathbf{r}} = \frac{\partial \dot{\mathbf{r}}}{\partial \mathbf{u}} \mathbf{u} + \frac{\partial \mathbf{r}}{\partial t} \quad (8.197)$$

We define the variation associated with  $\mathbf{u}$  to be  $\xi$ , so that the virtual displacement  $\delta\mathbf{r}$  is given by

$$\delta\mathbf{r} = \frac{\partial \dot{\mathbf{r}}}{\partial \mathbf{u}} \xi \quad (8.198)$$

The first term on the left side of (8.196) can then be written

$$\begin{aligned} \frac{d}{dt} \int_b \dot{\mathbf{r}}^T dm \delta\mathbf{r} &= \frac{d}{dt} \int_b \frac{\partial}{\partial \dot{\mathbf{r}}} \left( \frac{\dot{\mathbf{r}}^T dm \dot{\mathbf{r}}}{2} \right) \frac{\partial \dot{\mathbf{r}}}{\partial \mathbf{u}} \xi \\ &= \frac{d}{dt} \left[ \frac{\partial T}{\partial \mathbf{u}} \xi \right] \end{aligned} \quad (8.199)$$

This leads to the following result:

The equation of motion can be written

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \mathbf{u}} \boldsymbol{\xi} \right) - \delta T - W_\delta - \int_b \left( d\mathbf{f}^{(c)} \right)^T \delta \mathbf{r} = 0 \quad (8.200)$$

Suppose that  $T = T(\mathbf{u})$ ,  $W_\delta = \boldsymbol{\tau}^T \boldsymbol{\xi}$  and that  $\int_b (d\mathbf{f}^{(c)})^T \delta \mathbf{r} = 0$ . Then  $\delta T = (\partial T / \partial \mathbf{u}) \delta \mathbf{u}$ , and the equation of motion is found to be

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \mathbf{u}} \right) \boldsymbol{\xi} - \frac{\partial T}{\partial \mathbf{u}} \left( \delta \mathbf{u} - \dot{\boldsymbol{\xi}} \right) - \boldsymbol{\tau}^T \boldsymbol{\xi} = 0 \quad (8.201)$$

The equation of motion (8.200) was presented in (Bremer 1988) where it was termed a *central equation* as it forms a basis from which related results like Lagrange's equation of motion, Hamel-Boltzmann's equation and the Euler-Poincaré equation can be derived with a reasonable effort. In the following we will use the equation of motion in the form (8.201) to derive the Euler-Poincaré equation in  $SO(3)$  and  $SE(3)$ .

**Example 138** It is noted that if generalized coordinates are available so that  $\mathbf{u} = \dot{\mathbf{q}}$ , then the usual equations

$$\dot{\mathbf{r}} = \frac{\partial \mathbf{r}}{\partial \mathbf{q}} \dot{\mathbf{q}} + \frac{\partial \mathbf{r}}{\partial t}, \quad \delta \mathbf{r} = \frac{\partial \mathbf{r}}{\partial \mathbf{q}} \delta \mathbf{q} \quad (8.202)$$

are recovered in place of (8.197) and (8.198). Moreover, if  $T = T(\mathbf{q}, \dot{\mathbf{q}})$ ,  $W_\delta = \boldsymbol{\tau}^T \boldsymbol{\xi}$  and  $\int_b (d\mathbf{f}^{(c)})^T \delta \mathbf{r} = 0$ , then  $\boldsymbol{\xi} = \delta \mathbf{q}$  and

$$\delta T = \frac{\partial T}{\partial \mathbf{q}} \delta \mathbf{q} + \frac{\partial T}{\partial \dot{\mathbf{q}}} \delta \dot{\mathbf{q}} \quad (8.203)$$

This gives the familiar result

$$\left[ \frac{d}{dt} \left( \frac{\partial T}{\partial \dot{\mathbf{q}}} \right) - \frac{\partial T}{\partial \mathbf{q}} - \boldsymbol{\tau} \right]^T \delta \mathbf{q} = 0 \quad (8.204)$$

which gives Lagrange's equation of motion if the elements of  $\delta \mathbf{q}$  are independent.

### 8.5.2 Rotating rigid body

A rotating rigid body  $b$  has configuration  $\mathbf{R}$  and angular velocity  $\boldsymbol{\omega}$  in the body-fixed  $b$  frame. The generalized speed is taken to be  $\mathbf{u} = \boldsymbol{\omega}^\times$  where  $\boldsymbol{\omega}^\times = \mathbf{R}^T \dot{\mathbf{R}}$ . The corresponding variation vector is  $\boldsymbol{\sigma}^\times = \mathbf{R}^T \delta \mathbf{R}$ . The kinetic energy is

$$T = \frac{1}{2} \boldsymbol{\omega}^T \mathbf{M} \boldsymbol{\omega} \quad (8.205)$$

where  $\mathbf{M}$  is constant, positive definite and symmetric. Moreover, suppose that the generalized forces are denoted  $\boldsymbol{\tau}$ , so that the virtual work is  $W_\delta = \boldsymbol{\tau}^T \boldsymbol{\sigma}$ . Then

$$\frac{\partial T}{\partial \boldsymbol{\omega}} = \boldsymbol{\omega}^T \mathbf{M} \quad (8.206)$$

and the equation of motion can then be found from (8.201) to be

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \boldsymbol{\omega}} \right) \boldsymbol{\sigma} - \frac{\partial T}{\partial \boldsymbol{\omega}} (\delta \boldsymbol{\omega} - \dot{\boldsymbol{\sigma}}) - \boldsymbol{\tau}^T \boldsymbol{\sigma} = 0 \quad (8.207)$$

Application of (8.138) gives

$$\left( \frac{d}{dt} \left( \frac{\partial T}{\partial \boldsymbol{\omega}} \right) - \frac{\partial T}{\partial \boldsymbol{\omega}} \boldsymbol{\omega}^\times - \boldsymbol{\tau}^T \right) \boldsymbol{\sigma} = 0 \quad (8.208)$$

If the rigid body undergoes a free rotation, then the components of  $\boldsymbol{\sigma}$  are independent. This leads to:

The Euler-Poincaré equation for a rotating rigid body is

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \boldsymbol{\omega}} \right)^T + \boldsymbol{\omega}^\times \left( \frac{\partial T}{\partial \boldsymbol{\omega}} \right)^T = \boldsymbol{\tau} \quad (8.209)$$

Insertion of (8.206) gives the familiar equation

$$\mathbf{M}\dot{\boldsymbol{\omega}} + \boldsymbol{\omega}^\times (\mathbf{M}\boldsymbol{\omega}) = \boldsymbol{\tau} \quad (8.210)$$

### 8.5.3 Free-floating rigid body

A free-floating rigid body  $b$  has configuration given by the homogeneous transformation matrix

$$\mathbf{T} = \begin{pmatrix} \mathbf{R} & \mathbf{r}_p \\ 0 & 1 \end{pmatrix} \in SE(3) \quad (8.211)$$

where  $\mathbf{R}$  is the rotation matrix and  $\mathbf{r}_p$  is the position of some fixed point  $p$  in  $b$ . The generalized speed is selected to be

$$\mathbf{u} = \mathbf{w} = \begin{pmatrix} \mathbf{v}_p^b \\ \boldsymbol{\omega}^b \end{pmatrix} \quad (8.212)$$

which is given in the body-fixed frame  $b$ , and which has a  $4 \times 4$  matrix form  $\hat{\mathbf{w}} = \mathbf{T}^{-1}\dot{\mathbf{T}}$ . The associated variation vector is  $\boldsymbol{\eta}$  which is defined by its  $4 \times 4$  matrix form  $\hat{\boldsymbol{\eta}} = \mathbf{T}^{-1}\delta\mathbf{T}$

To find an expression for the kinetic energy we need to find expressions for

$$\dot{\mathbf{r}} = \frac{\partial \dot{\mathbf{r}}}{\partial \mathbf{w}} \mathbf{w}, \quad \delta \mathbf{r} = \frac{\partial \dot{\mathbf{r}}}{\partial \mathbf{w}} \boldsymbol{\eta} \quad (8.213)$$

This is found by observing that the velocity  $\dot{\mathbf{r}}$  of the mass element is  $\dot{\mathbf{r}} = \mathbf{R}(\mathbf{v}_p^b + \boldsymbol{\omega}^{b\times} \mathbf{r}_{pm}^b)$  where  $\mathbf{r} = \mathbf{r}_p + \mathbf{r}_{pm}$  and  $\dot{\mathbf{r}}_p = \mathbf{R}\mathbf{v}_p^b$ . This gives

$$\frac{\partial \dot{\mathbf{r}}}{\partial \mathbf{w}} = \mathbf{R} \begin{pmatrix} \mathbf{I} & : & -\mathbf{r}_{pq}^{b\times} \end{pmatrix} \quad (8.214)$$

The kinetic energy is then found to be

$$T = \frac{1}{2} \int_b \dot{\mathbf{r}}^T d\mathbf{m} \dot{\mathbf{r}} \quad (8.215)$$

$$= \frac{1}{2} \int_b \mathbf{w}^T \begin{pmatrix} \mathbf{I} \\ \mathbf{r}_{pq}^{b\times} \end{pmatrix} \mathbf{R}^T \mathbf{R} \begin{pmatrix} \mathbf{I} & : & -\mathbf{r}_{pq}^{b\times} \end{pmatrix} \mathbf{w} d\mathbf{m} \quad (8.216)$$

$$= \frac{1}{2} \mathbf{w}^T \mathbf{D}_p^b \mathbf{w} \quad (8.217)$$

where

$$\mathbf{D}_p^b = \begin{pmatrix} m\mathbf{I} & -\int_b \mathbf{r}_{pq}^{b\times} dm \\ \int_b \mathbf{r}_{pq}^{b\times} dm & -\int_b \mathbf{r}_{pq}^{b\times} \mathbf{r}_{pq}^{b\times} dm \end{pmatrix} = \begin{pmatrix} m\mathbf{I} & m\mathbf{r}_g^{b\times} \\ -m\mathbf{r}_g^{b\times} & \mathbf{M}_p^b \end{pmatrix} \quad (8.218)$$

For a free-floating rigid body the principle of virtual work states that the forces of constraint does no virtual work, that is,

$$\int_b \left( d\mathbf{f}^{(c)} \right)^T \delta \mathbf{r} = 0 \quad (8.219)$$

The virtual work is

$$W_\delta = \int_b d\mathbf{f}^T \delta \mathbf{r} = \begin{pmatrix} \mathbf{F}^b \\ \mathbf{L}_p^b \end{pmatrix}^T \boldsymbol{\eta} \quad (8.220)$$

where  $\mathbf{F}^b$  is the total force on the rigid body with line of action through the point  $P$ , and  $\mathbf{L}_p^b$  is the total torque. The vectors  $\mathbf{F}^b$  and  $\mathbf{L}_p^b$  are given in the frame  $b$ .

In this setting, the kinetic energy will be a function of velocity and angular velocity, which is written  $T = T(\mathbf{w})$ . The equation of motion is then found from (8.201) to be

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \mathbf{w}} \right) \boldsymbol{\eta} - \frac{\partial T}{\partial \mathbf{w}} (\delta \mathbf{w} - \dot{\boldsymbol{\eta}}) - \begin{pmatrix} \mathbf{F}^b \\ \mathbf{L}_p^b \end{pmatrix}^T \boldsymbol{\eta} = 0 \quad (8.221)$$

Then, by noting from (8.153) that  $\delta \mathbf{w} - \dot{\boldsymbol{\eta}} = \mathbf{ad}_w \boldsymbol{\eta}$  we find that

$$\left[ \frac{d}{dt} \left( \frac{\partial T}{\partial \mathbf{w}} \right) - \frac{\partial T}{\partial \mathbf{w}} \mathbf{ad}_w - \begin{pmatrix} \mathbf{F}^b \\ \mathbf{L}_p^b \end{pmatrix}^T \right] \boldsymbol{\eta} = 0 \quad (8.222)$$

Since  $\boldsymbol{\eta}$  is arbitrary this gives

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \mathbf{w}} \right)^T - \mathbf{ad}_w^T \left( \frac{\partial T}{\partial \mathbf{w}} \right)^T = \begin{pmatrix} \mathbf{F}^b \\ \mathbf{L}_p^b \end{pmatrix} \quad (8.223)$$

and using (8.152) we find that

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \mathbf{w}} \right)^T + \begin{pmatrix} \boldsymbol{\omega}^\times & 0 \\ \mathbf{v}^\times & \boldsymbol{\omega}^\times \end{pmatrix} \left( \frac{\partial T}{\partial \mathbf{w}} \right)^T = \begin{pmatrix} \mathbf{F}^b \\ \mathbf{L}_p^b \end{pmatrix} \quad (8.224)$$

This equation can be expanded to Kirchhoff's equation of motion:

Euler-Poincaré's equation for a free-floating rigid body gives Kirchhoff's equations of motion

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \mathbf{v}} \right)^T + \boldsymbol{\omega}^\times \left( \frac{\partial T}{\partial \mathbf{v}} \right)^T = \mathbf{F}^b \quad (8.225)$$

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \mathbf{w}} \right)^T + \mathbf{v}^\times \left( \frac{\partial T}{\partial \mathbf{v}} \right)^T + \boldsymbol{\omega}^\times \left( \frac{\partial T}{\partial \boldsymbol{\omega}} \right)^T = \mathbf{L}_p^b \quad (8.226)$$

Kirchhoff's equations of motion are important in the modeling of ship motion, where also the added inertia effects may be represented in this setting (Lamb 1945), (Sagatun and Fossen 1991), (Leonard 1997), (Fossen 2002).

**Example 139** If it is been assumed that  $P$  is the center of mass of the rigid body, then

$$\left( \frac{\partial T}{\partial \mathbf{w}} \right)^T = \begin{pmatrix} m\mathbf{v}_c^b \\ \mathbf{M}_c^b \omega_c^b \end{pmatrix} \quad (8.227)$$

and the equations of motion are found to be

$$m\dot{\mathbf{v}}^b + \boldsymbol{\omega}^{b \times} m\mathbf{v}_c^b = \mathbf{F}^b \quad (8.228)$$

$$\mathbf{M}_c^b \dot{\boldsymbol{\omega}}^b + \boldsymbol{\omega}^{b \times} \mathbf{M}_c^b \boldsymbol{\omega}^b = \mathbf{L}_c^b \quad (8.229)$$

### 8.5.4 Mechanism with $n$ degrees of freedom

We will now study a mechanism with  $n$  degrees of freedom. The velocity vectors of the rigid body  $k$  are then

$$\mathbf{w}_k = \frac{\partial \mathbf{w}_k}{\partial \mathbf{u}} \mathbf{u}, \quad \boldsymbol{\eta}_k = \frac{\partial \mathbf{w}_k}{\partial \mathbf{u}} \boldsymbol{\eta}, \quad \frac{\partial \mathbf{w}_k}{\partial \mathbf{u}} = \begin{pmatrix} \frac{\partial \mathbf{v}_{kp}^k}{\partial \mathbf{u}} \\ \frac{\partial \mathbf{u}_k}{\partial \mathbf{u}} \end{pmatrix} \quad (8.230)$$

where  $\mathbf{u} = (u_1, \dots, u_n)^T$  is the vector of generalized velocities. Typically, for a robot arm we will have  $\mathbf{u} = \dot{\mathbf{q}}$  where  $\mathbf{q}$  is the  $n$ -dimensional vector of generalized coordinates. In this case the principle of virtual work is used to eliminate the forces of constraint. The principle of virtual work states that the total virtual work of the constraint forces is zero. Therefore we need to sum up the virtual work done by the constraint forces for the whole system to eliminate the constraint forces from the equation of motion. Note that the constraint forces includes two types of constraint forces: Internal constraint forces in each body which makes the body rigid, and interconnecting constraint forces that hold the mechanism together.

The principle of virtual work for a mechanism with  $k$  interconnected rigid bodies can be written

$$\sum_{k=1}^n \int_{b_k} \left( d\mathbf{f}^{(c)} \right)^T \delta \mathbf{r} = 0 \quad (8.231)$$

where  $b_k$  denotes body  $k$ . Therefore the forces of constraint can be eliminated by summing up the equations of motions in the form (8.200). This gives

$$\sum_{k=1}^n \left[ \frac{d}{dt} \left( \frac{\partial T_k}{\partial \mathbf{w}_k} \boldsymbol{\eta}_k \right) - \delta T_k - W_{k\delta} - \int_b \left( d\mathbf{f}^{(c)} \right)^T \delta \mathbf{r} \right] = 0 \quad (8.232)$$

we are able to eliminate the constraint forces, and get

$$\sum_{k=1}^n \left[ \frac{d}{dt} \left( \frac{\partial T_k}{\partial \mathbf{w}_k} \boldsymbol{\eta}_k \right) - \delta T_k - W_{k\delta} \right] = 0 \quad (8.233)$$

If we proceed as for the free-floating rigid body, but keep the virtual displacement  $\boldsymbol{\eta}_k$  in the expression we arrive at

$$\begin{aligned} 0 &= \sum_{k=1}^n \left[ \boldsymbol{\eta}_k^T \left( \begin{array}{l} m_k \dot{\mathbf{v}}_{kc}^k + \boldsymbol{\omega}_k^{k \times} m_k \mathbf{v}_{kc}^k - \mathbf{F}_k^k \\ \mathbf{M}_k^k \dot{\boldsymbol{\omega}}_k^k + \boldsymbol{\omega}_k^{k \times} \mathbf{M}_k^k \boldsymbol{\omega}_k^k - \mathbf{L}_k^k \end{array} \right) \right] \\ &= \sum_{k=1}^n \left[ \boldsymbol{\eta}^T \left( \begin{array}{l} \frac{\partial \mathbf{v}_{kp}^k}{\partial \mathbf{u}} \\ \frac{\partial \mathbf{u}_k}{\partial \mathbf{u}} \end{array} \right)^T \left( \begin{array}{l} m_k \dot{\mathbf{v}}_{kc}^k + \boldsymbol{\omega}_k^{k \times} m_k \mathbf{v}_{kc}^k - \mathbf{F}_k^k \\ \mathbf{M}_k^k \dot{\boldsymbol{\omega}}_k^k + \boldsymbol{\omega}_k^{k \times} \mathbf{M}_k^k \boldsymbol{\omega}_k^k - \mathbf{L}_{kc}^k \end{array} \right) \right] \end{aligned} \quad (8.234)$$

Here  $\boldsymbol{\eta}$  is arbitrary, and it follows that

$$\sum_{k=1}^n \left[ \left( \frac{\partial \mathbf{v}_{kc}^k}{\partial \mathbf{u}} \right)^T (m \dot{\mathbf{v}}_{kc}^k + \boldsymbol{\omega}_k^k \times m \mathbf{v}_{kc}^k - \mathbf{F}_k^k) + \left( \frac{\partial \boldsymbol{\omega}_k^k}{\partial \mathbf{u}} \right)^T (\mathbf{M}_{kc} \dot{\boldsymbol{\omega}}_k^k + \boldsymbol{\omega}_k^{k \times} \mathbf{M}_{kc} \boldsymbol{\omega}_k^k - \mathbf{L}_{kc}^k) \right] = 0 \quad (8.235)$$

We note that  $\mathbf{F}_k^k$  is the applied force to body  $k$ , and  $\mathbf{L}_{kc}$  is the applied torque to body  $k$  around point center of mass of body  $k$ , and that the forces of constraint has been eliminated in the derivation. This form of the equation of motion was called the Newton-Euler equation of motion with eliminated constraint forces in (Bremer 1988). Written out in component form it was called Kane's equation of motion in (Kane and Levinson 1985).

## 8.6 Hamilton's principle

### 8.6.1 Introduction

Hamilton's principle is based on the use of the time integral of certain energy functions. Hamilton's principle can be used to derive Lagrange's equation of motion for a system described by  $n$  generalized coordinates  $q_1, \dots, q_n$ . The motivation for introducing Hamilton's principle is that it is the starting point for the Hamilton-Jacobi equation, and that it is used for systems described by partial differential equations. Moreover, it can be used to derive the Euler-Poincaré equation. The Euler-Lagrange equation for the integral of a function is the starting point for the development.

### 8.6.2 The extended Hamilton principle

The presentation starts with the extended Hamilton principle, which will be derived in the following. Consider a system with  $N$  particles, where particle  $k$  has mass  $m_k$  and position  $\vec{r}_k(q_1, \dots, q_n, t)$  where  $q_i$  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  $\vec{a}_k = d\vec{v}_k/dt$ . The starting point is again d'Alembert's principle

$$\sum_{k=1}^N \left( m_k \frac{d\vec{v}_k}{dt} - \vec{F}_k \right) \cdot \delta \vec{r}_k = 0 \quad (8.236)$$

The virtual work of the forces  $\vec{F}_k$  satisfy

$$\sum_{k=1}^N \vec{F}_k \cdot \delta \vec{r}_k = \widetilde{W}_\delta \quad (8.237)$$

where the function  $\widetilde{W}_\delta$  is defined by

$$\widetilde{W}_\delta = \sum_{j=1}^n \left( \tau_j - \frac{\partial U}{\partial q_j} \right) \delta q_j = W_\delta - \delta U \quad (8.238)$$

Here

$$W_\delta = \sum_{j=1}^n \tau_j \delta q_j \quad (8.239)$$

is the virtual work of the active generalized forces  $\tau_j$ , and

$$\delta U = \sum_{j=1}^n \frac{\partial U}{\partial q_j} \delta q_j \quad (8.240)$$

is the variation of the potential energy  $U$ . The kinetic energy

$$T = \frac{1}{2} \sum_{k=1}^N m_k \vec{v}_k \cdot \vec{v}_k \quad (8.241)$$

has the variation

$$\begin{aligned} \delta T &= \delta \left( \frac{1}{2} \sum_{k=1}^N m_k \vec{v}_k \cdot \vec{v}_k \right) = \sum_{k=1}^N m_k \vec{v}_k \cdot \delta \vec{v}_k \\ &= \sum_{k=1}^N \frac{d}{dt} (m_k \vec{v}_k \cdot \delta \vec{r}_k) - \sum_{k=1}^N m_k \frac{d \vec{v}_k}{dt} \cdot \delta \vec{r}_k \end{aligned} \quad (8.242)$$

This result in combination with (8.236) and (8.237) leads to the equation

$$\delta T - \widetilde{W}_\delta - \sum_{k=1}^N \frac{d}{dt} (m_k \vec{v}_k \cdot \delta \vec{r}_k) = 0 \quad (8.243)$$

A critical observation for the next step in the derivation is the fact that if  $\vec{r}_k(t_1)$  and  $\vec{r}_k(t_2)$  are fixed, then

$$\int_{t_1}^{t_2} \sum_{k=1}^N \frac{d}{dt} (m_k \vec{v}_k \cdot \delta \vec{r}_k) dt = \sum_{k=1}^N (m_k \vec{v}_k \cdot \delta \vec{r}_k) \Big|_{t=t_1}^{t=t_2} = 0 \quad (8.244)$$

This means that we can eliminate the last term of (8.243) by integrating the expression in (8.243) from  $t_1$  to  $t_2$ . This leads to the following result:

The extended Hamilton principle is given by

$$\int_{t_1}^{t_2} (\delta T + \widetilde{W}_\delta) dt = 0 \quad (8.245)$$

where the endpoints are fixed and  $\widetilde{W}_\delta = W_\delta - \delta U$ , or, alternatively, by

$$\int_{t_1}^{t_2} (\delta L + W_\delta) dt = 0 \quad (8.246)$$

where the endpoints are fixed.

### 8.6.3 Derivation of Lagrange's equation of motion

We consider a mechanical system with generalized coordinates  $\mathbf{q}$  and Lagrangian  $L$ . We study a trajectory  $C$  given by  $\mathbf{q}(t)$  on the time interval  $t_1 \leq t \leq t_2$  with the boundary conditions that  $\mathbf{q}(t_1)$  and  $\mathbf{q}(t_2)$  are given. A variation  $\delta \mathbf{q}(t) = \boldsymbol{\psi}(t)$ ,  $\delta \dot{\mathbf{q}}(t) = \dot{\boldsymbol{\psi}}(t)$  is

considered for the trajectory  $C$ , where  $\psi(t)$  is an arbitrary function. The boundary conditions imply that  $\psi(t_1) = \psi(t_2) = 0$ . The corresponding variation  $\delta L$  in the Lagrangian is

$$\delta L = \frac{\partial L}{\partial \mathbf{q}} \delta \mathbf{q} + \frac{\partial L}{\partial \dot{\mathbf{q}}} \delta \dot{\mathbf{q}} = \left( \frac{\partial L}{\partial \mathbf{q}} \psi + \frac{\partial L}{\partial \dot{\mathbf{q}}} \dot{\psi} \right) \quad (8.247)$$

The extended Hamilton principle (8.246) gives

$$\int_{t_1}^{t_2} (\delta L + W_\delta) dt = \int_{t_1}^{t_2} \left( \frac{\partial L}{\partial \mathbf{q}} \psi + \frac{\partial L}{\partial \dot{\mathbf{q}}} \dot{\psi} + \boldsymbol{\tau}^T \psi \right) dt = 0 \quad (8.248)$$

Partial integration gives

$$\int_{t_1}^{t_2} \frac{\partial L}{\partial \dot{\mathbf{q}}} \dot{\psi} dt = \left[ \frac{\partial L}{\partial \dot{\mathbf{q}}} \psi \right]_{t_1}^{t_2} - \int_{t_1}^{t_2} \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{\mathbf{q}}} \right) \psi dt = - \int_{t_1}^{t_2} \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{\mathbf{q}}} \right) \psi dt \quad (8.249)$$

where it is used that  $\psi(t_1) = \psi(t_2) = 0$ . The variation of the integral is then found to be

$$\int_{t_1}^{t_2} \left[ \frac{\partial L}{\partial \mathbf{q}} - \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{\mathbf{q}}} \right) + \boldsymbol{\tau}^T \right] \psi dt = 0 \quad (8.250)$$

Since  $\psi(t)$  is arbitrary, this implies that

$$\frac{d}{dt} \left( \frac{\partial L}{\partial \dot{\mathbf{q}}} \right)^T - \left( \frac{\partial L}{\partial \mathbf{q}} \right)^T = \boldsymbol{\tau} \quad (8.251)$$

which is Lagrange's equation of motion.

#### 8.6.4 Hamilton's principle

Suppose that  $W_\delta = 0$ , which means that there are no active forces  $\boldsymbol{\tau}$  acting on the system. In this case the extended Hamilton's principle gives

$$\int_{t_1}^{t_2} \delta L dt = 0 \quad (8.252)$$

which is known as Hamilton's principle. The system will then follow some trajectory  $(\mathbf{q}, \dot{\mathbf{q}})$  which is denoted  $C$ , where the trajectory  $C$  depends on the initial conditions. Define the *action integral* of a trajectory  $C$  by

$$A(C) = \int_{t_1}^{t_2} L dt \quad (8.253)$$

which is the integral of the Lagrangian. The action integral does not have a clear physical interpretation, it is merely a mathematical tool. Hamilton's principle can then be reformulated to state that the variation of the action integral is zero for the trajectory  $C$ , that is,

$$\delta A(C) = 0 \quad (8.254)$$

The variation of the action integral is

$$\delta A(C) = \delta \int_{t_1}^{t_2} L dt = \int_{t_1}^{t_2} \delta L dt \quad (8.255)$$

and from the derivation in the previous section it is seen that Hamilton's principle implies that

$$\frac{d}{dt} \left( \frac{\partial L}{\partial \dot{\mathbf{q}}} \right) - \frac{\partial L}{\partial \mathbf{q}} = 0 \quad (8.256)$$

which is Lagrange's equation of motion when the applied force is zero

### 8.6.5 Rotations with the Euler-Poincaré equation

Consider a rigid body with rotation matrix  $\mathbf{R}$ , angular velocity  $\boldsymbol{\omega}$  in body-fixed coordinates defined by  $\dot{\mathbf{R}} = \mathbf{R}\boldsymbol{\omega}^\times$ , and variation  $\delta\mathbf{R} = \mathbf{R}\boldsymbol{\sigma}^\times$  where  $\boldsymbol{\sigma}$  is an arbitrary vector. Suppose that the kinetic energy can be written

$$T = \frac{1}{2}\boldsymbol{\omega}^T \mathbf{M} \boldsymbol{\omega} \quad (8.257)$$

where  $\mathbf{M}$  is a constant matrix, and that the generalized force vector acting on the body is  $\boldsymbol{\tau}$  so that  $W_\delta = \boldsymbol{\tau}^T \boldsymbol{\sigma}$ . Assume that the potential energy is zero. Then

$$\begin{aligned} \int_{t_1}^{t_2} (\delta T + W_\delta) dt &= \int_{t_1}^{t_2} \left( \frac{\partial T}{\partial \boldsymbol{\omega}} \delta \boldsymbol{\omega} + \boldsymbol{\tau}^T \boldsymbol{\sigma} \right) dt \\ &= \int_{t_1}^{t_2} \left[ \frac{\partial T}{\partial \boldsymbol{\omega}} (\dot{\boldsymbol{\sigma}} + \boldsymbol{\omega}^\times \boldsymbol{\sigma}) + \boldsymbol{\tau}^T \boldsymbol{\sigma} \right] dt \\ &= \int_{t_1}^{t_2} \left[ -\frac{d}{dt} \left( \frac{\partial T}{\partial \boldsymbol{\omega}} \right) + \frac{\partial T}{\partial \boldsymbol{\omega}} \boldsymbol{\omega}^\times + \boldsymbol{\tau}^T \right] \boldsymbol{\sigma} dt \end{aligned} \quad (8.258)$$

where (8.138) and partial integration was used. Then as the components  $\boldsymbol{\sigma}$  are independent, Hamilton's extended principle (8.245) gives the following result

The equation of motion for a rigid body can be written

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \boldsymbol{\omega}} \right)^T + \boldsymbol{\omega}^\times \left( \frac{\partial T}{\partial \boldsymbol{\omega}} \right)^T = \boldsymbol{\tau} \quad (8.259)$$

which is the Euler-Poincaré equation in  $SO(3)$ .

The Euler-Poincaré can be written out as Euler's equation

$$\mathbf{M} \dot{\boldsymbol{\omega}} + \boldsymbol{\omega}^\times \mathbf{M} \boldsymbol{\omega} = \boldsymbol{\tau} \quad (8.260)$$

### 8.6.6 Rigid motion with the Euler-Poincaré equation

Consider a rigid body with homogeneous transformation matrix  $\mathbf{T}$  and velocity vector  $\mathbf{w} = (\mathbf{v}, \boldsymbol{\omega})^T$  in body coordinates so that  $\dot{\mathbf{T}} = \mathbf{T}\hat{\mathbf{w}}$  where  $\hat{\mathbf{w}}$  is defined in (8.146). Let the variation of  $\mathbf{T}$  be given by  $\delta\mathbf{T} = \mathbf{T}\hat{\boldsymbol{\eta}}$ . Suppose that the kinetic energy is given by

$$T = \frac{1}{2}\mathbf{w}^T \mathbf{M} \mathbf{w} \quad (8.261)$$

where  $\mathbf{M}$  is constant. Assume that the virtual work is  $W_\delta = \boldsymbol{\tau}^T \boldsymbol{\eta}$ , and that the potential energy is  $U = 0$ . Then

$$\int_{t_1}^{t_2} (\delta T + W_\delta) dt = \int_{t_1}^{t_2} \left( \frac{\partial T}{\partial \mathbf{w}} \delta \mathbf{w} + \boldsymbol{\tau}^T \boldsymbol{\eta} \right) dt \quad (8.262)$$

The variation  $\delta\mathbf{w}$  satisfies

$$\delta \mathbf{w} = \dot{\boldsymbol{\eta}} + \text{ad}_w \boldsymbol{\eta} \quad (8.263)$$

$$\boldsymbol{\eta}(t_1) = \boldsymbol{\eta}(t_2) = \mathbf{0} \quad (8.264)$$

where  $\mathbf{ad}_w$  is given by (8.152). This gives

$$\int_{t_1}^{t_2} (\delta T + W_\delta) dt = \int_{t_1}^{t_2} \frac{\partial T}{\partial \mathbf{w}} (\dot{\boldsymbol{\eta}} + \mathbf{ad}_w \boldsymbol{\eta} + \boldsymbol{\tau}^T \boldsymbol{\eta}) dt \quad (8.265)$$

$$= \int_{t_1}^{t_2} \left[ -\frac{d}{dt} \left( \frac{\partial T}{\partial \mathbf{w}} \right) + \frac{\partial T}{\partial \mathbf{w}} \mathbf{ad}_w + \boldsymbol{\tau}^T \right] \boldsymbol{\eta} dt \quad (8.266)$$

and, since  $\boldsymbol{\eta}$  is arbitrary, Hamilton's extended principle (8.245) gives

The equation of motion for rigid motion can be written in the form of a Euler-Poincaré equation as

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \mathbf{w}} \right)^T - \mathbf{ad}_w^T \left( \frac{\partial T}{\partial \mathbf{w}} \right)^T = \boldsymbol{\tau} \quad (8.267)$$

where

$$\mathbf{ad}_w = \begin{pmatrix} \boldsymbol{\omega}^\times & \mathbf{v}^\times \\ 0 & \boldsymbol{\omega}^\times \end{pmatrix} \quad (8.268)$$

The Euler-Poincaré equation with  $\boldsymbol{\tau} = (\mathbf{F}^T, \mathbf{L}^T)^T$  gives the equations

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \mathbf{v}} \right)^T + \boldsymbol{\omega}^\times \left( \frac{\partial T}{\partial \mathbf{v}} \right)^T = \mathbf{F} \quad (8.269)$$

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \boldsymbol{\omega}} \right)^T + \mathbf{v}^\times \left( \frac{\partial T}{\partial \boldsymbol{\omega}} \right)^T + \boldsymbol{\omega}^\times \left( \frac{\partial T}{\partial \boldsymbol{\omega}} \right)^T = \mathbf{L} \quad (8.270)$$

which are known as Kirchhoff's equations.

## 8.7 Lagrangian dynamics for PDE's

### 8.7.1 Flexible beam dynamics

Lagrange's equation of motion can also be used for systems described by partial differential equations. To illustrate this we will derive Lagrange's equation of motion for an Euler-Bernoulli beam (Meirovitch 1980). The beam is of length  $L$ , and the undeformed beam is along the  $x$  axis. The elastic displacement in the  $z$  direction is denoted by  $w(x, t)$ . The kinetic energy is written

$$T(t) = \int_0^L \hat{T} [\dot{w}(x, t), \dot{w}'(x, t)] dx \quad (8.271)$$

where  $\hat{T}dx$  is the kinetic energy of the length element  $dx$  of the beam. The potential energy is

$$U(t) = \int_0^L \hat{U} [w(x, t), w'(x, t), w''(x, t)] dx \quad (8.272)$$

where  $\hat{U}dx$  is the potential energy of the length element  $dx$ . The Lagrangian can then be defined as

$$L(t) = \int_0^L \hat{L} [w(x, t), w'(x, t), w''(x, t), \dot{w}(x, t), \dot{w}'(x, t)] dx \quad (8.273)$$

where  $\hat{L} = \hat{T} - \hat{U}$ . The virtual work on  $dx$  due to nonconservative forces is

$$\hat{W}_\delta(x, t) = f(x, t)\delta w(x, t) \quad (8.274)$$

and the virtual work from nonconservative forces on the beam is therefore

$$W_\delta(t) = \int_0^L \hat{W}_\delta(x, t) dx \quad (8.275)$$

The extended Hamilton principle for this system gives

$$\int_{t_1}^{t_2} (\delta L + W_\delta) dt = \int_{t_1}^{t_2} \int_0^L (\delta \hat{L} + \hat{W}_\delta) dx dt = 0 \quad (8.276)$$

The variation of the Lagrangian density is

$$\delta \hat{L} = \frac{\partial \hat{L}}{\partial w} \delta w + \frac{\partial \hat{L}}{\partial w'} \delta w' + \frac{\partial \hat{L}}{\partial w''} \delta w'' + \frac{\partial \hat{L}}{\partial \dot{w}} \delta \dot{w} + \frac{\partial \hat{L}}{\partial \dot{w}'} \delta \dot{w}' \quad (8.277)$$

and the extended Hamilton principle is therefore

$$\int_{t_1}^{t_2} \int_0^L \left( \frac{\partial \hat{L}}{\partial w} \delta w + \frac{\partial \hat{L}}{\partial w'} \delta w' + \frac{\partial \hat{L}}{\partial w''} \delta w'' + \frac{\partial \hat{L}}{\partial \dot{w}} \delta \dot{w} + \frac{\partial \hat{L}}{\partial \dot{w}'} \delta \dot{w}' + f(x, t) \delta w(x, t) \right) dx dt = 0$$

Using partial integration and that  $\delta w = 0$  and  $\delta w' = 0$  at  $t = t_1$  and  $t = t_2$  it is possible, with some patience, to reach the following result:

$$\begin{aligned} & \int_{t_1}^{t_2} \int_0^L \left[ \frac{\partial \hat{L}}{\partial w} - \frac{\partial}{\partial x} \left( \frac{\partial \hat{L}}{\partial w'} \right) + \frac{\partial^2}{\partial x^2} \left( \frac{\partial \hat{L}}{\partial w''} \right) - \frac{\partial}{\partial t} \left( \frac{\partial \hat{L}}{\partial \dot{w}} \right) \right. \\ & \quad \left. + \frac{\partial^2}{\partial x \partial t} \left( \frac{\partial \hat{L}}{\partial \dot{w}'} \right) + f(x, t) \right] \delta w dx dt \\ & + \int_{t_1}^{t_2} \left\{ \left[ \frac{\partial \hat{L}}{\partial w'} - \frac{\partial}{\partial x} \left( \frac{\partial \hat{L}}{\partial w''} \right) + \frac{\partial}{\partial t} \left( \frac{\partial \hat{L}}{\partial \dot{w}'} \right) \right] \delta w \Big|_0^L + \frac{\partial \hat{L}}{\partial w''} \delta w' \Big|_0^L \right\} dt = 0 \end{aligned} \quad (8.278)$$

As  $\delta w$  and  $\delta w'$  are arbitrary for  $t_1 < t < t_2$  this implies the Lagrangian equation of motion in the form

$$\frac{\partial \hat{L}}{\partial w} - \frac{\partial}{\partial x} \left( \frac{\partial \hat{L}}{\partial w'} \right) + \frac{\partial^2}{\partial x^2} \left( \frac{\partial \hat{L}}{\partial w''} \right) - \frac{\partial}{\partial t} \left( \frac{\partial \hat{L}}{\partial \dot{w}} \right) + \frac{\partial^2}{\partial x \partial t} \left( \frac{\partial \hat{L}}{\partial \dot{w}'} \right) + f(x, t) = 0 \quad (8.279)$$

with boundary conditions

$$\left[ \frac{\partial \hat{L}}{\partial w'} - \frac{\partial}{\partial x} \left( \frac{\partial \hat{L}}{\partial w''} \right) + \frac{\partial}{\partial t} \left( \frac{\partial \hat{L}}{\partial \dot{w}'} \right) \right] \delta w \Big|_0^L = 0 \quad (8.280)$$

$$\frac{\partial \hat{L}}{\partial w''} \delta w' \Big|_0^L = 0 \quad (8.281)$$

**Example 140** To reach the result (8.278) the following partial integrations are used.

$$\begin{aligned}
 \int_0^L \frac{\partial \hat{L}}{\partial w'} \delta w' dx &= \left. \frac{\partial \hat{L}}{\partial w'} \delta w \right|_0^L - \int_0^L \frac{\partial}{\partial x} \left( \frac{\partial \hat{L}}{\partial w'} \right) \delta w dx \\
 \int_0^L \frac{\partial \hat{L}}{\partial w''} \delta w'' dx &= \left. \frac{\partial \hat{L}}{\partial w''} \delta w' \right|_0^L - \frac{\partial}{\partial x} \left( \frac{\partial \hat{L}}{\partial w''} \right) \delta w \Big|_0^L + \int_0^L \frac{\partial^2}{\partial x^2} \left( \frac{\partial \hat{L}}{\partial w''} \right) \delta w dx \\
 \int_{t_1}^{t_2} \frac{\partial \hat{L}}{\partial \dot{w}} \delta \dot{w} dx &= - \int_{t_1}^{t_2} \frac{\partial}{\partial t} \left( \frac{\partial \hat{L}}{\partial \dot{w}} \right) \delta \dot{w} dx \\
 \int_{t_1}^{t_2} \int_0^L \frac{\partial \hat{L}}{\partial \dot{w}'} \delta \dot{w}' dx dt &= \int_{t_1}^{t_2} \left[ \left. \frac{\partial \hat{L}}{\partial \dot{w}'} \delta \dot{w}' \right|_0^L - \int_0^L \frac{\partial}{\partial x} \left( \frac{\partial \hat{L}}{\partial \dot{w}'} \right) \delta \dot{w}' dx \right] dt \\
 &= \int_{t_1}^{t_2} \left[ \frac{\partial}{\partial t} \left( \left. \frac{\partial \hat{L}}{\partial \dot{w}'} \delta w \right|_0^L \right) - \int_0^L \frac{\partial^2}{\partial x \partial t} \left( \frac{\partial \hat{L}}{\partial \dot{w}'} \right) \delta w dx \right] dt
 \end{aligned}$$

### 8.7.2 Euler-Bernoulli beam

For an Euler Bernoulli beam the Lagrangian density is

$$\hat{L} = \frac{1}{2} \rho(x) [\dot{w}(x, t)]^2 - \frac{1}{2} EI(x) [w''(x)]^2 \quad (8.282)$$

The Lagrangian equation of motion (8.279) is in this case

$$\frac{\partial^2}{\partial x^2} \left( \frac{\partial \hat{L}}{\partial w''} \right) - \frac{\partial}{\partial t} \left( \frac{\partial \hat{L}}{\partial \dot{w}} \right) + f(x, t) = 0 \quad (8.283)$$

which is evaluated to be

$$\rho(x) \ddot{w}(x, t) + [EI(x) w''(x)]'' = f \quad (8.284)$$

### 8.7.3 Lateral vibrations in a string

The kinetic energy for lateral vibrations in a string is

$$T = \frac{1}{2} \int_0^L \rho(x) [\dot{w}(x, t)]^2 dx \quad (8.285)$$

while the potential energy is

$$U = \frac{1}{2} \int_0^L P [w'(x)]^2 dx \quad (8.286)$$

The displacement of the string is  $w$ . The string is displaced by a force  $Pw'$ . A change in slope  $dw'$  requires the work  $Pw'dw'$  which integrates to

$$\hat{U} = \int_0^{w'} Pw'dw' = \frac{1}{2} P [w'(x)]^2 \quad (8.287)$$

The Lagrangian density is

$$\hat{L} = \frac{1}{2}\rho(x) [\dot{w}(x, t)]^2 - \frac{1}{2}P [w'(x)]^2 \quad (8.288)$$

The Lagrangian equation of motion (8.279) is in this case

$$-\frac{\partial}{\partial x} \left( \frac{\partial \hat{L}}{\partial w'} \right) - \frac{\partial}{\partial t} \left( \frac{\partial \hat{L}}{\partial \dot{w}} \right) + f(x, t) = 0 \quad (8.289)$$

which gives the equation of motion

$$\rho(x)\ddot{w}(x, t) = [Pw'(x)]' = f \quad (8.290)$$

## 8.8 Hamilton's equations of motion

### 8.8.1 Introduction

Hamilton's equations of motion are strongly related to Lagrange's equation of motion, and are based on energy expressions and generalized coordinates. In addition, the concept of a generalized momentum vector is introduced. Hamilton's equation of motion can be used to establish physical properties that are important in controller design and in simulation. In particular, this formulation is useful to establish energy functions that are invariant with zero control input. This can be used to find Lyapunov function candidates, and for checking the accuracy of numerical simulations. There are even specialized simulation methods for Hamiltonian systems. In addition, the Hamiltonian formulation leads to the Hamilton-Jacobi equation which is an important tool in optimal control theory. Basic references for this section are (Lovelock and Rund 1989) and (Goldstein 1980).

### 8.8.2 Hamilton's equation of motion

We consider a system with generalized coordinates  $\mathbf{q}$  and Lagrangian

$$L(\mathbf{q}, \dot{\mathbf{q}}, t) = T(\mathbf{q}, \dot{\mathbf{q}}, t) - U(\mathbf{q}) \quad (8.291)$$

The *momentum vector* is defined by

$$\mathbf{p}(\mathbf{q}, \dot{\mathbf{q}}, t) = \frac{\partial L(\mathbf{q}, \dot{\mathbf{q}}, t)}{\partial \dot{\mathbf{q}}}^T \quad (8.292)$$

We note that Lagrange's equation of motion can be written

$$\dot{\mathbf{p}}(\mathbf{q}, \dot{\mathbf{q}}, t) - \frac{\partial L(\mathbf{q}, \dot{\mathbf{q}}, t)}{\partial \mathbf{q}}^T = \boldsymbol{\tau} \quad (8.293)$$

To define the Hamiltonian  $H$  from the Lagrangian  $L$  a change of variables from the Lagrangian variables  $(\mathbf{q}, \dot{\mathbf{q}})$  to the Hamiltonian variables  $(\mathbf{q}, \mathbf{p})$  is required. The velocity vector  $\dot{\mathbf{q}}$  is then regarded to be a function

$$\dot{\mathbf{q}} = \phi(\mathbf{q}, \mathbf{p}, t) \quad (8.294)$$

of the Hamiltonian variables  $(\mathbf{q}, \mathbf{p})$ .

The Hamiltonian is defined by

$$H(\mathbf{q}, \mathbf{p}, t) = \mathbf{p}^T \boldsymbol{\phi}(\mathbf{q}, \mathbf{p}, t) - L(\mathbf{q}, \boldsymbol{\phi}, t) \quad (8.295)$$

Partial differentiation of (8.295) gives

$$\begin{aligned} \frac{\partial H(\mathbf{q}, \mathbf{p}, t)}{\partial \mathbf{p}} &= \boldsymbol{\phi}^T + \left( \mathbf{p}^T - \frac{\partial L(\mathbf{q}, \dot{\mathbf{q}}, t)}{\partial \dot{\mathbf{q}}} \right) \frac{\partial \boldsymbol{\phi}(\mathbf{q}, \mathbf{p}, t)}{\partial \mathbf{p}} \\ \frac{\partial H(\mathbf{q}, \mathbf{p}, t)}{\partial \mathbf{q}} &= \left( \mathbf{p}^T - \frac{\partial L(\mathbf{q}, \dot{\mathbf{q}}, t)}{\partial \dot{\mathbf{q}}} \right) \frac{\partial \boldsymbol{\phi}(\mathbf{q}, \mathbf{p}, t)}{\partial \mathbf{q}} - \frac{\partial L(\mathbf{q}, \boldsymbol{\phi}, t)}{\partial \mathbf{q}} \\ \frac{\partial H(\mathbf{q}, \mathbf{p}, t)}{\partial t} &= \left( \mathbf{p}^T - \frac{\partial L(\mathbf{q}, \dot{\mathbf{q}}, t)}{\partial \dot{\mathbf{q}}} \right) \frac{\partial \boldsymbol{\phi}(\mathbf{q}, \mathbf{p}, t)}{\partial t} - \frac{\partial L(\mathbf{q}, \boldsymbol{\phi}, t)}{\partial t} \end{aligned}$$

where the definition (8.292) has been used. It follows from the definition (8.292) of the momentum vector that

$$\frac{\partial H(\mathbf{q}, \mathbf{p}, t)}{\partial \mathbf{p}} = \boldsymbol{\phi}^T \quad (8.296)$$

$$\frac{\partial H(\mathbf{q}, \mathbf{p}, t)}{\partial \mathbf{q}} = -\frac{\partial L(\mathbf{q}, \boldsymbol{\phi}, t)}{\partial \mathbf{q}} \quad (8.297)$$

$$\frac{\partial H(\mathbf{q}, \mathbf{p}, t)}{\partial t} = -\frac{\partial L(\mathbf{q}, \boldsymbol{\phi}, t)}{\partial t} \quad (8.298)$$

Insertion of Lagrange's equation of motion and  $\dot{\mathbf{q}} = \boldsymbol{\phi}(\mathbf{q}, \mathbf{p}, t)$  leads to the result:

Hamilton's equations of motion are given by

$$\dot{\mathbf{q}} = \frac{\partial H(\mathbf{q}, \mathbf{p}, t)}{\partial \mathbf{p}} \quad (8.299)$$

$$\dot{\mathbf{p}} = -\frac{\partial H(\mathbf{q}, \mathbf{p}, t)}{\partial \mathbf{q}} + \boldsymbol{\tau} \quad (8.300)$$

The time derivative of the Hamiltonian is found from the chain rule to be

$$\frac{dH}{dt} = \frac{\partial H}{\partial \mathbf{p}} \dot{\mathbf{p}} + \frac{\partial H}{\partial \mathbf{q}} \dot{\mathbf{q}} + \frac{\partial H}{\partial t} \quad (8.301)$$

By inserting Hamilton's equations of motion (8.299) and (8.300) we find that

$$\frac{dH}{dt} = \dot{\mathbf{q}}^T \left( -\frac{\partial H}{\partial \mathbf{q}}^T + \boldsymbol{\tau} \right) + \frac{\partial H}{\partial \mathbf{q}} \dot{\mathbf{q}} + \frac{\partial H}{\partial t} = \dot{\mathbf{q}}^T \boldsymbol{\tau} + \frac{\partial H}{\partial t} \quad (8.302)$$

which leads to:

The time derivative of the Hamiltonian is

$$\frac{dH(\mathbf{q}, \mathbf{p}, t)}{dt} = \dot{\mathbf{q}}^T \boldsymbol{\tau} + \frac{\partial H(\mathbf{q}, \mathbf{p}, t)}{\partial t} \quad (8.303)$$

The following result is useful:

If the Hamiltonian does not depend on time  $t$ , that is, if  $H = H(\mathbf{q}, \mathbf{p})$ , and if the system is unactuated so that  $\boldsymbol{\tau} = \mathbf{0}$ , then

$$\frac{dH(\mathbf{q}, \mathbf{p})}{dt} = 0 \quad (8.304)$$

### 8.8.3 The energy function

Define the *energy function*  $h(\mathbf{q}, \dot{\mathbf{q}}, t)$  by

$$h(\mathbf{q}, \dot{\mathbf{q}}, t) = \frac{\partial L(\mathbf{q}, \dot{\mathbf{q}}, t)}{\partial \dot{\mathbf{q}}} \dot{\mathbf{q}} - L(\mathbf{q}, \dot{\mathbf{q}}, t) \quad (8.305)$$

The time derivative of the energy function is found from the definition (8.305) to be

$$\frac{dh(\mathbf{q}, \dot{\mathbf{q}}, t)}{dt} = \frac{d}{dt} \left( \frac{\partial L(\mathbf{q}, \dot{\mathbf{q}}, t)}{\partial \dot{\mathbf{q}}} \dot{\mathbf{q}} \right) - \frac{dL}{dt}(\mathbf{q}, \dot{\mathbf{q}}, t) \quad (8.306)$$

Insertion of (8.15) gives the result

$$\frac{dh(\mathbf{q}, \dot{\mathbf{q}}, t)}{dt} = -\frac{\partial L(\mathbf{q}, \dot{\mathbf{q}}, t)}{\partial t} + \boldsymbol{\tau}^T \dot{\mathbf{q}} \quad (8.307)$$

From (8.292), (8.294), (8.295), and (8.305) it is possible to see that the Hamiltonian  $H(\mathbf{q}, \mathbf{p}, t)$ , has the same numerical value as the energy function  $h(\mathbf{q}, \dot{\mathbf{q}}, t)$ , that is,

$$H(\mathbf{q}, \mathbf{p}, t) = h(\mathbf{q}, \dot{\mathbf{q}}, t) \quad (8.308)$$

However, the two functions have different arguments, and should not be confused with each other.

Suppose that the kinetic energy is quadratic in the velocity, that is,

$$T(\mathbf{q}, \dot{\mathbf{q}}, t) = \frac{1}{2} \dot{\mathbf{q}}^T \mathbf{M}(\mathbf{q}, t) \dot{\mathbf{q}} \quad (8.309)$$

Then the the energy function becomes

$$h(\mathbf{q}, \dot{\mathbf{q}}, t) = T(\mathbf{q}, \dot{\mathbf{q}}, t) + U(\mathbf{q}) \quad (8.310)$$

which is the sum of the kinetic and potential energy. This explains the name energy function.

The kinetic energy can in general be written in the form

$$T(\mathbf{q}, \dot{\mathbf{q}}, t) = \frac{1}{2} \dot{\mathbf{q}}^T \mathbf{M}(\mathbf{q}, t) \dot{\mathbf{q}} + \boldsymbol{\alpha}(\mathbf{q}, t)^T \dot{\mathbf{q}} + T_0(\mathbf{q}, t) \quad (8.311)$$

$$= T_2(\mathbf{q}, \dot{\mathbf{q}}, t) + T_1(\mathbf{q}, \dot{\mathbf{q}}, t) + T_0(\mathbf{q}, t) \quad (8.312)$$

where  $T_2(\mathbf{q}, \dot{\mathbf{q}}, t)$  is quadratic in the velocities,  $T_1(\mathbf{q}, \dot{\mathbf{q}}, t)$  is linear in the velocities and  $T_0(\mathbf{q}, t)$  is independent of the velocity. We find that

$$\left[ \frac{\partial L}{\partial \dot{\mathbf{q}}} \right] \dot{\mathbf{q}} = \dot{\mathbf{q}}^T \mathbf{M} \dot{\mathbf{q}} + \boldsymbol{\alpha}^T \dot{\mathbf{q}} = 2T_2 + T_1 \quad (8.313)$$

and using (8.305) and (8.307) we may state the following result:

The energy function is given by

$$h(\mathbf{q}, \dot{\mathbf{q}}, t) = T_2(\mathbf{q}, \dot{\mathbf{q}}, t) + U_a(\mathbf{q}, t) \quad (8.314)$$

where  $U_a(\mathbf{q}, t) = U(\mathbf{q}, t) - T_0(\mathbf{q}, t)$  may be considered to be an apparent potential. If  $h = h(\mathbf{q}, \dot{\mathbf{q}})$  and the system is unactuated, then

$$\frac{dh(\mathbf{q}, \dot{\mathbf{q}})}{dt} = 0 \quad (8.315)$$

### 8.8.4 Change of generalized coordinates

Consider a system with Lagrangian  $L(\mathbf{q}, \dot{\mathbf{q}}, t) = T(\mathbf{q}, \dot{\mathbf{q}}, t) - U(\mathbf{q})$ , where

$$T(\mathbf{q}, \dot{\mathbf{q}}, t) = \frac{1}{2} \dot{\mathbf{q}}^T \mathbf{M}(\mathbf{q}, t) \dot{\mathbf{q}} \quad (8.316)$$

The energy function is

$$h(\mathbf{q}, \dot{\mathbf{q}}, t) = T(\mathbf{q}, \dot{\mathbf{q}}, t) + U(\mathbf{q}) \quad (8.317)$$

while the momentum vector is

$$\mathbf{p}(\mathbf{q}, \dot{\mathbf{q}}, t) = \frac{\partial L(\mathbf{q}, \dot{\mathbf{q}}, t)}{\partial \dot{\mathbf{q}}}^T = \mathbf{M}(\mathbf{q}, t) \dot{\mathbf{q}} \quad (8.318)$$

and the Hamiltonian is

$$H(\mathbf{q}, \mathbf{p}, t) = \mathbf{p}^T \phi - L(\mathbf{q}, \phi, t) \quad (8.319)$$

where  $\phi(\mathbf{q}, \mathbf{p}, t) = \dot{\mathbf{q}}$ .

A change in coordinates

$$\mathbf{q} = \mathbf{q}_0 + \mathbf{Q} \quad (8.320)$$

from  $\mathbf{q}$  to  $\mathbf{Q}$  gives

$$T_Q(\mathbf{Q}, \dot{\mathbf{Q}}, t) = \frac{1}{2} \dot{\mathbf{Q}}^T \mathbf{M}_Q(\mathbf{Q}, t) \dot{\mathbf{Q}} + \dot{\mathbf{q}}_0^T \mathbf{M}_Q(\mathbf{Q}, t) \dot{\mathbf{Q}} + \frac{1}{2} \dot{\mathbf{q}}_0^T \mathbf{M}_Q(\mathbf{Q}, t) \dot{\mathbf{q}}_0 \quad (8.321)$$

$$U_Q(\mathbf{Q}) = U(\mathbf{q}) \quad (8.322)$$

The Lagrangian in the new coordinates is

$$L_Q(\mathbf{Q}, \dot{\mathbf{Q}}, t) = T_Q(\mathbf{Q}, \dot{\mathbf{Q}}, t) - U_Q(\mathbf{Q}) = L(\mathbf{q}, \dot{\mathbf{q}}, t) \quad (8.323)$$

which means that the Lagrangian has the same numerical value after the change in generalized coordinates. The momentum vector is

$$\mathbf{P}(\mathbf{Q}, \dot{\mathbf{Q}}, t) = \frac{\partial L_Q(\mathbf{Q}, \dot{\mathbf{Q}}, t)}{\partial \dot{\mathbf{Q}}}^T = \mathbf{M}_Q(\mathbf{Q}, t) (\dot{\mathbf{Q}} + \dot{\mathbf{q}}_0) = \mathbf{p}(\mathbf{q}, \dot{\mathbf{q}}, t) \quad (8.324)$$

The energy function becomes

$$\begin{aligned} h_Q(\mathbf{Q}, \dot{\mathbf{Q}}, t) &= \frac{1}{2} \dot{\mathbf{Q}}^T \mathbf{M}_Q(\mathbf{Q}, t) \dot{\mathbf{Q}} + U_Q(\mathbf{Q}) - \frac{1}{2} \dot{\mathbf{q}}_0^T \mathbf{M}_Q(\mathbf{Q}, t) \dot{\mathbf{q}}_0 \\ &= h(\mathbf{q}, \dot{\mathbf{q}}, t) - [\dot{\mathbf{q}}_0^T \mathbf{M}_Q(\mathbf{Q}, t) \dot{\mathbf{Q}} + \dot{\mathbf{q}}_0^T \mathbf{M}_Q(\mathbf{Q}, t) \dot{\mathbf{q}}_0] \end{aligned} \quad (8.325)$$

while the Hamiltonian is

$$H_Q(\mathbf{Q}, \mathbf{P}, t) = \mathbf{P}^T \Phi - L_Q(\mathbf{Q}, \Phi, t) = H(\mathbf{q}, \mathbf{p}, t) - \mathbf{p}^T \dot{\mathbf{q}}_0 \quad (8.326)$$

where  $\Phi = \dot{\mathbf{Q}}$ . We see that the Lagrangian and the canonical momentum vector has the same numerical value after a change of coordinates, while the numerical value of the energy function and the Hamiltonian changes when the coordinate is changed from  $\mathbf{q}$  to  $\mathbf{Q}$ .

**Example 141** We will demonstrate the effect of a change of generalized coordinates on the Lagrangian and the Hamiltonian with an example. Consider a satellite modeled as a mass point moving about the earth in a nominally circular orbit. Let the position of the satellite be

$$\mathbf{r} = \mathbf{R} + \mathbf{q} \quad (8.327)$$

where  $\mathbf{R}$  is the nominal circular motion with radius  $R = |\mathbf{R}|$  and a constant angular velocity  $\omega_c$  around the earth. The kinetic and potential energy of the satellite are

$$T_r = \frac{1}{2}m\dot{\mathbf{r}}^T \dot{\mathbf{r}}, \quad U_r = \mu \frac{m}{r} \quad (8.328)$$

where  $r = |\mathbf{r}|$ . The energy function

$$h_r(\mathbf{r}, \dot{\mathbf{r}}) = \frac{1}{2}m\dot{\mathbf{r}}^T \dot{\mathbf{r}} + \mu \frac{m}{r} \quad (8.329)$$

is constant. Suppose that the velocity is given by

$$\mathbf{v} = \omega_c R \mathbf{c}_1 + \dot{\mathbf{q}} \quad (8.330)$$

where  $\mathbf{c}_1$  is the unit vector along the tangent of the nominal orbit. The kinetic energy may be written

$$T_q = \frac{1}{2}m\dot{\mathbf{q}}^T \dot{\mathbf{q}} + \omega_c R \mathbf{c}_1^T \dot{\mathbf{q}} + \frac{1}{2}m\omega_c^2 R^2 \quad (8.331)$$

With these variables the energy function is

$$h_q(\mathbf{q}, \dot{\mathbf{q}}) = \frac{1}{2}m\dot{\mathbf{q}}^T \dot{\mathbf{q}} + U_a(\mathbf{q}) \quad (8.332)$$

where

$$U_a(\mathbf{q}) = \mu \frac{m}{|\mathbf{R} + \mathbf{q}|} - \frac{1}{2}m\omega_c^2 R^2 \quad (8.333)$$

is the apparent potential.

## 8.9 Control aspects

### 8.9.1 Passivity of Hamilton's equation of motion

Suppose that the Hamiltonian does not depend on time, so that  $H = H(\mathbf{q}, \mathbf{p})$ . Then from (8.303) the time derivative of  $H$  along the solutions of Hamilton's equation of motion (8.299, 8.300) is

$$\frac{dH(\mathbf{q}, \mathbf{p})}{dt} = \dot{\mathbf{q}}^T \boldsymbol{\tau} \quad (8.334)$$

Next, suppose that the Hamiltonian is bounded from below, which means that there is a constant  $H_0$  so that  $H \geq H_0$ . Then it is possible to define a nonnegative storage function  $V = H - H_0$  so that the time derivative of  $V$  along the solution of the system is

$$\dot{V} = \dot{\mathbf{q}}^T \boldsymbol{\tau} \quad (8.335)$$

From the results of Section 2.4.14 we see that this leads to the following conclusion:

If the Hamiltonian does not depend on time  $t$ , and if there is a constant  $H_0$  so that  $H(\mathbf{q}, \mathbf{p}) \geq H_0$ , then the system given by Hamilton's equation of motion (8.299, 8.300) is passive with input  $\boldsymbol{\tau}$  and output  $\dot{\mathbf{q}}$ , and  $V = H - H_0$  is a storage function.

**Example 142** Suppose that the Lagrangian  $L$  does not depend on time  $t$ . Then it is seen from (8.298) and the definition (8.305) of the energy function  $h$  that the Hamiltonian  $H$  and the energy function  $h$  will not depend on  $t$ : This gives

$$\frac{dH(\mathbf{q}, \mathbf{p})}{dt} = \frac{dh(\mathbf{q}, \dot{\mathbf{q}})}{dt} = \dot{\mathbf{q}}^T \boldsymbol{\tau} \quad (8.336)$$

If the actuator force is set to zero, then this implies that  $H(\mathbf{q}, \mathbf{p})$  and  $h(\mathbf{q}, \dot{\mathbf{q}})$  are constants for solutions of the system. In the terminology of Hamiltonian dynamics  $H$  and  $h$  are said to be invariants of motion.

**Example 143** Suppose that (8.336) holds. Velocity feedback in the form  $\boldsymbol{\tau} = -K\dot{\mathbf{q}}$  will then lead to

$$\frac{dH(\mathbf{q}, \mathbf{p})}{dt} = \frac{dh(\mathbf{q}, \dot{\mathbf{q}})}{dt} = -K\dot{\mathbf{q}}^T \dot{\mathbf{q}} \leq 0 \quad (8.337)$$

which means that the energy  $h(\mathbf{q}, \dot{\mathbf{q}}) = H(\mathbf{q}, \mathbf{p})$  will be nonincreasing, and that the energy will decrease whenever the velocity is nonzero.

### 8.9.2 Example: Manipulator dynamics

Consider a robotic manipulator with Lagrangian equation of motion

$$\mathbf{M}(\mathbf{q}) \ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}}) \dot{\mathbf{q}} + \mathbf{g}(\mathbf{q}) = \boldsymbol{\tau} \quad (8.338)$$

where  $\mathbf{M}$  is symmetric and positive definite, which implies that  $\mathbf{M}$  is nonsingular, and where  $\dot{\mathbf{M}} - 2\mathbf{C}$  is skew symmetric. The Lagrangian is

$$L = \frac{1}{2} \dot{\mathbf{q}}^T \mathbf{M}(\mathbf{q}) \dot{\mathbf{q}} - U(\mathbf{q}) \quad (8.339)$$

and the momentum vector is

$$\mathbf{p} = \frac{\partial L(\mathbf{q}, \dot{\mathbf{q}}, t)}{\partial \dot{\mathbf{q}}}^T = \mathbf{M}(\mathbf{q}) \dot{\mathbf{q}} \quad (8.340)$$

As  $\mathbf{M}$  is nonsingular the velocity can be written

$$\dot{\mathbf{q}} = \mathbf{M}(\mathbf{q})^{-1} \mathbf{p} \quad (8.341)$$

The Hamiltonian is found from the definition (8.295) to be

$$H = \mathbf{p}^T \mathbf{M}(\mathbf{q})^{-1} \mathbf{p} - \frac{1}{2} \mathbf{p}^T \mathbf{M}(\mathbf{q})^{-1} \mathbf{M}(\mathbf{q}) \mathbf{M}(\mathbf{q})^{-1} \mathbf{p} + U(\mathbf{q}) \quad (8.342)$$

which simplifies to

$$H(\mathbf{q}, \mathbf{p}) = \frac{1}{2} \mathbf{p}^T \mathbf{M}(\mathbf{q})^{-1} \mathbf{p} + U(\mathbf{q}) \quad (8.343)$$

where the first term is the kinetic energy. Then, if  $U(\mathbf{q})$  is lower bounded, passivity of the manipulator dynamics with input  $\boldsymbol{\tau}$  and output  $\dot{\mathbf{q}}$  can be concluded from the general result in Section 8.9.1. This agrees with the passivity analysis based on the Lagrangian dynamics in Section 8.2.9. Note that it is easier to establish passivity from  $\boldsymbol{\tau}$  to  $\dot{\mathbf{q}}$  in the Hamiltonian formulation as it is only necessary to check that  $H$  does not depend on  $t$  and that  $U$  is lower bounded.

### 8.9.3 Example: The restricted three-body problem

The restricted three-body problem (Szebehely 1967) is a classical problem in celestial mechanics that has been adopted as a benchmark for numerical integrators (Hairer and Wanner 1996), (Shampine, Allen and Preuss 1997). The reason for this is that there are periodic solutions that are tabulated, and these solutions are very sensitive to changes in initial conditions. Therefore, the accuracy of a numerical integrator can be investigated by solving the system equations with initial conditions that correspond to a periodic solution, and then check if the numerically computed solution is periodic. This is done in Section 14.1.3.

The system includes three masses moving in a plane. The primary body is the earth, the secondary body is the moon, and the third body is a satellite. The earth has mass  $m_1$ , the moon has mass  $m_2$ , and the satellite has mass  $m_3$ , which is much smaller than  $m_1$  and  $m_2$ . In the formulation of the problem it is assumed that the moon and the earth interact in a gravitational field without being influenced by the satellite. A coordinate frame  $b$  has the unit vector  $\vec{b}_1$  along the axis from the earth to the moon, while the  $\vec{b}_3$  vector is along the axis of rotation of the earth-moon system. According to the law of gravitation the gravity force  $\vec{F}_1$  on the earth from the moon, and the gravity force  $\vec{F}_2$  on the moon from the earth are given by

$$\vec{F}_1 = -\vec{F}_2 = k^2 \frac{m_1 m_2}{L^2} \vec{b}_1 \quad (8.344)$$

where  $k$  is the Gaussian constant of gravitation and  $L$  is the distance between the two bodies. The vector from the center of the earth to the center of the moon rotates with an angular velocity  $\vec{\omega} = \omega \vec{b}_3$ . The earth has position  $\vec{R}_1 = -x_1 \vec{b}_1$  and the moon has position  $\vec{R}_2 = x_2 \vec{b}_1$ , which implies that  $L = x_1 + x_2$ . The accelerations are according to (6.405)

$$\vec{a}_1 = \vec{\omega} \times (\vec{\omega} \times \vec{R}_1) = \omega^2 x_1 \vec{b}_1 \quad (8.345)$$

$$\vec{a}_2 = \vec{\omega} \times (\vec{\omega} \times \vec{R}_2) = -\omega^2 x_2 \vec{b}_1 \quad (8.346)$$

Balance between the centrifugal forces and the gravitational forces imply that

$$k^2 \frac{m_1 m_2}{L^2} = m_1 x_1 \omega^2 = m_2 x_2 \omega^2 \quad (8.347)$$

From this we get Kepler's third law:

$$\omega^2 = \frac{k^2 M}{L^3} \quad (8.348)$$

where  $M = m_1 + m_2$ .

The satellite moves in the rotating gravitational field set up by the earth and the moon. The position of the satellite is

$$\vec{r} = x\vec{b}_1 + y\vec{b}_2 \quad (8.349)$$

the velocity is

$$\vec{v} = \frac{^b d}{dt} \vec{r} + \vec{\omega}_{ib} \times \vec{r} = \dot{x}\vec{b}_1 + \dot{y}\vec{b}_2 + \omega \left( x\vec{b}_2 - y\vec{b}_1 \right) \quad (8.350)$$

and from (6.405) the acceleration is

$$\begin{aligned} \vec{a} &= \frac{^b d^2}{dt^2} \vec{r} + 2\vec{\omega}_{ib} \times \frac{^b d}{dt} \vec{r} + \vec{\alpha}_{ib} \times \vec{r} + \vec{\omega}_{ib} \times (\vec{\omega}_{ib} \times \vec{r}) \\ &= \ddot{x}\vec{b}_1 + \ddot{y}\vec{b}_2 + 2\omega \left( \dot{x}\vec{b}_2 - \dot{y}\vec{b}_1 \right) - \omega^2 \left( x\vec{b}_1 + y\vec{b}_2 \right) \end{aligned} \quad (8.351)$$

The gravitational force on the satellite is the sum of the gravitational forces from the earth and the moon, which gives

$$\vec{F}_3 = -k^2 \frac{m_1 m_3}{r_1^3} \left[ (x + x_1)\vec{b}_1 + y\vec{b}_2 \right] - k^2 \frac{m_2 m_3}{r_2^3} \left[ (x - x_2)\vec{b}_1 + y\vec{b}_2 \right] \quad (8.352)$$

where

$$r_1 = \sqrt{(x + x_1)^2 + y^2}, \quad r_2 = \sqrt{(x - x_2)^2 + y^2} \quad (8.353)$$

Newton's law in the  $x$  and  $y$  directions gives

$$\ddot{x} - 2\omega\dot{y} - \omega^2 x = -k^2 \left[ \frac{m_1}{r_1^3} (x + x_1) + \frac{m_2}{r_2^3} (x - x_2) \right] \quad (8.354)$$

$$\ddot{y} + 2\omega\dot{x} - \omega^2 y = -k^2 \left( \frac{m_1}{r_1^3} + \frac{m_2}{r_2^3} \right) \quad (8.355)$$

This model is normalized by introducing

$$\xi = \frac{x}{L}, \quad \eta = \frac{y}{L}, \quad \tau = \omega t \quad (8.356)$$

$$\rho_1 = \frac{r_1}{L}, \quad \rho_2 = \frac{r_2}{L} \quad (8.357)$$

$$\mu_1 = \frac{m_1}{M} = \frac{x_2}{L}, \quad \mu_2 = \frac{m_2}{M} = \frac{x_1}{L} \quad (8.358)$$

This gives the normalized model for the restricted three-body problem:

$$\frac{d^2\xi}{d\tau^2} - 2\frac{d\eta}{d\tau} - \xi = - \left[ \frac{\mu_1(\xi + \mu_2)}{\rho_1^3} + \frac{\mu_2(\xi - \mu_1)}{\rho_2^3} \right] \quad (8.359)$$

$$\frac{d^2\eta}{d\tau^2} + 2\frac{d\xi}{d\tau} - \eta = - \left( \frac{\mu_1\eta}{\rho_1^3} + \frac{\mu_2\eta}{\rho_2^3} \right) \quad (8.360)$$

A constant energy function of the system is found from the kinetic energy

$$\begin{aligned} T &= \frac{1}{2} m_3 \vec{v} \cdot \vec{v} \\ &= \frac{1}{2} m_3 \left[ \dot{x}\vec{b}_1 + \dot{y}\vec{b}_2 + \omega \left( x\vec{b}_2 - y\vec{b}_1 \right) \right] \cdot \left[ \dot{x}\vec{b}_1 + \dot{y}\vec{b}_2 + \omega \left( x\vec{b}_2 - y\vec{b}_1 \right) \right] \\ &= \underbrace{\frac{1}{2} m_3 (\dot{x}^2 + \dot{y}^2)}_{T_2} + \underbrace{m_3 \omega (-\dot{x}\dot{y} + \dot{y}\dot{x})}_{T_1} + \underbrace{\frac{1}{2} m_3 \omega^2 (x^2 + y^2)}_{T_0} \end{aligned} \quad (8.361)$$

and the potential energy

$$U = -k^2 m_3 \left( \frac{m_1}{r_1} + \frac{m_2}{r_2} \right) \quad (8.362)$$

Then (8.314) and (8.315) imply that an invariant energy function is given by

$$h = \frac{1}{2} m_3 [\dot{x}^2 + \dot{y}^2 - \omega^2 (x^2 + y^2)] - k^2 m_3 \left( \frac{m_1}{r_1} + \frac{m_2}{r_2} \right) \quad (8.363)$$

In the normalized form the invariant energy function is

$$\kappa = \frac{1}{2} \left[ \left( \frac{d\xi}{d\tau} \right)^2 + \left( \frac{d\eta}{d\tau} \right)^2 - \xi^2 - \eta^2 \right] - \frac{\mu_1}{\rho_1} - \frac{\mu_2}{\rho_2} \quad (8.364)$$

#### 8.9.4 Example: Attitude dynamics for a satellite

Consider a satellite that moves in a circular orbit about the earth with radius  $R_c$ . The satellite has mass  $m$  and inertia dyadic  $\vec{M}_c$  about the center of mass. The center of mass has position  $\vec{R}_c$  relative to the origin of frame  $i$ . Frame  $b$  is fixed in the satellite, and frame  $i$  is a Newtonian frame with origin in the center of the earth and with axes pointing at certain fixed stars. An orbital frame  $c$  is defined so that  $\vec{c}_1$  is along the tangent of the orbit in the positive velocity direction,  $\vec{c}_2$  is perpendicular to the orbit, and  $\vec{c}_3$  is locally vertical and pointing downwards. Then  $\vec{R}_c = -R_c \vec{c}_3$ . The attitude dynamics are assumed to be given by

$$\vec{M}_c \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_c \cdot \vec{\omega}_{ib}) = \vec{\tau}_c \quad (8.365)$$

Here  $\vec{\omega}_{ib}$  is the angular velocity of frame  $b$  relative to frame  $i$ ,  $\vec{\alpha}_{ib}$  is the angular acceleration, and  $\vec{\tau}_c$  is the actuator torque.

A satellite moving in a circular orbit will have velocity  $\vec{v} = \omega_c R_c \vec{c}_1$  where

$$\omega_c = \sqrt{\frac{\mu}{R_c^3}} \quad (8.366)$$

is the orbital frequency in the sense that  $T = 2\pi/\omega_c$  is the period of one orbit. In passing we mention that this equation and the fact that the radius of the earth is 6,378 km make it possible to compute the altitude of geostationary orbits as 35,863 km, because this gives an orbital period of 24 hours. Also a low earth orbit with altitude 1,200 km corresponds to a radius  $R_c = 7578$  km and an orbital period of 109 min.

The kinetic energy of the satellite is

$$T_i = \frac{1}{2} \vec{\omega}_{ib} \cdot \vec{M}_c \cdot \vec{\omega}_{ib} + \frac{1}{2} m \omega_c^2 R_c^2 \quad (8.367)$$

while the potential energy is

$$U = -\frac{\mu m}{R_c} \quad (8.368)$$

Then the energy function

$$\begin{aligned} h_i &= T_{i2} - T_{i0} + U \\ &= \frac{1}{2} \vec{\omega}_{ib} \cdot \vec{M}_c \cdot \vec{\omega}_{ib} - \frac{1}{2} m \omega_c^2 R_c^2 - \frac{\mu m}{R_c} \end{aligned} \quad (8.369)$$

is constant as long as the system is not actuated. The last two terms are constants, so that the function

$$V_i = \frac{1}{2} \vec{\omega}_{ib} \cdot \vec{M}_c \cdot \vec{\omega}_{ib} \quad (8.370)$$

will be a constant for the unactuated system.

The description based on the angular velocity  $\vec{\omega}_{ib}$  gives the rotation of the satellite relative to a star-fixed coordinate frame  $i$ . For the stabilization of the attitude in an orbit it is better to study the dynamics in terms of the angular velocity  $\vec{\omega}_{cb}$  of the satellite relative to the orbit frame. The change of variables is done using

$$\vec{\omega}_{ib} = \vec{\omega}_{cb} + \vec{\omega}_{ic} \quad (8.371)$$

$$= \vec{\omega}_{cb} + \omega_c \vec{c}_2 \quad (8.372)$$

This gives the expression

$$T_c = \frac{1}{2} \vec{\omega}_{cb} \cdot \vec{M}_c \cdot \vec{\omega}_{cb} + \omega_c \vec{\omega}_{cb} \cdot \vec{M}_c \cdot \vec{c}_2 + \frac{1}{2} \omega_c^2 \vec{c}_2 \cdot \vec{M}_c \cdot \vec{c}_2 + \frac{1}{2} m \omega_c^2 R_c^2 \quad (8.373)$$

for the kinetic energy. Then the energy function, which is found from

$$h_c = T_{c2} - T_{c0} + U \quad (8.374)$$

$$= \frac{1}{2} \vec{\omega}_{cb} \cdot \vec{M}_c \cdot \vec{\omega}_{cb} - \frac{1}{2} \omega_c^2 \vec{c}_2 \cdot \vec{M}_c \cdot \vec{c}_2 - \frac{1}{2} m \omega_c^2 R_c^2 - \frac{\mu m}{R_c} \quad (8.375)$$

is constant for the unactuated satellite. The last two terms on the right side are constants, and this shows that

$$V_c = \frac{1}{2} \vec{\omega}_{cb} \cdot \vec{M}_c \cdot \vec{\omega}_{cb} - \frac{1}{2} \omega_c^2 \vec{c}_2 \cdot \vec{M}_c \cdot \vec{c}_2 \quad (8.376)$$

is a constant function for the unactuated satellite.

### 8.9.5 Example: Gravity gradient stabilization

In this section we introduce gravity gradient stabilization of the satellite in the previous section. The material is adopted from (Hughes 1986). The gravity force acting on a mass element is then

$$d\vec{f} = -\mu \frac{\vec{R}}{R^3} dm \quad (8.377)$$

where  $R^3$  denotes  $|\vec{R}|^3$ . Note that the gravity force has a gradient in the radial direction, and this creates a torque about the center of mass. This torque is known as the gravity gradient torque, and is given by

$$\vec{g}_c = -\mu \int_b \frac{\vec{r} \times \vec{R}}{R^3} dm \quad (8.378)$$

where  $\vec{r} = \vec{R} - \vec{R}_c$  is the vector from the center of mass to the mass element  $dm$ . The gravitational potential is

$$U = -\mu \int_b \frac{dm}{R} \quad (8.379)$$

The expressions for the gravity gradient  $\vec{g}_c$  and the gravitational potential are difficult to use in their present form, but can be approximated with more suitable expressions using the binomial series

$$\frac{1}{R^3} = \frac{1}{R_c^3} \left( 1 - 3 \frac{\vec{r} \cdot \vec{R}_c}{R_c^2} + \dots \right) \quad (8.380)$$

and

$$\frac{1}{R} = \frac{1}{R_c} \left( 1 - \frac{\vec{r} \cdot \vec{R}_c}{R_c^2} - \frac{1}{2} \frac{r^2}{R_c^2} + \frac{3}{2} \frac{(\vec{r} \cdot \vec{R}_c)^2}{R_c^4} + \dots \right) \quad (8.381)$$

Then, using  $\vec{r} \times \vec{R} = r \times \vec{R}_c$  the gravitational torque can be approximated by

$$\begin{aligned} \vec{g}_c &= -\frac{\mu}{R_c^3} \int_b \vec{r} dm \times \vec{R}_c - \frac{3\mu}{R_c^5} \vec{R}_c \cdot \int_b \vec{r} \vec{r} dm \times \vec{R}_c \\ &= -\frac{3\mu}{R_c^3} \vec{c}_3 \times \int_b \vec{r} \vec{r} dm \cdot \vec{c}_3 \end{aligned} \quad (8.382)$$

where  $\int_b \vec{r} dm = \vec{0}$  is used. The inertia dyadic is

$$\vec{M}_c = \int_b (r^2 \vec{I} - \vec{r} \vec{r}) dm \quad (8.383)$$

It is found that

$$\vec{c}_3 \times \int_b \vec{r} \vec{r} dm \cdot \vec{c}_3 = \vec{c}_3 \times \int_b r^2 \vec{I} dm \cdot \vec{c}_3 + \vec{c}_3 \times \vec{M}_c \cdot \vec{c}_3 \quad (8.384)$$

$$= \vec{c}_3 \times \vec{c}_3 \int_b r^2 dm + \vec{c}_3 \times \vec{M}_c \cdot \vec{c}_3 = \vec{c}_3 \times \vec{M}_c \cdot \vec{c}_3 \quad (8.385)$$

which gives

$$\vec{g}_c = 3\omega_c^2 \vec{c}_3 \times \vec{M}_c \cdot \vec{c}_3 \quad (8.386)$$

The equation of motion can then be written

$$\vec{M}_c \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_c \cdot \vec{\omega}_{ib}) = 3\omega_c^2 \vec{c}_3 \times \vec{M}_c \cdot \vec{c}_3 + \vec{r}_c \quad (8.387)$$

From the binomial series the approximation

$$U = U_0 + \frac{3}{2} \omega_c^2 \vec{c}_3 \cdot \vec{M}_c \cdot \vec{c}_3 \quad (8.388)$$

is found for the potential energy where  $U_0$  is a constant given by

$$U_0 = -\frac{\mu m}{R_c} - \frac{1}{2} \omega_c^2 \text{Trace} \vec{M}_c \quad (8.389)$$

The kinetic energy is

$$T_i = \frac{1}{2} \vec{\omega}_{ib} \cdot \vec{M}_c \cdot \vec{\omega}_{ib} + \frac{1}{2} m \omega_c^2 R_c^2 \quad (8.390)$$

where the velocity has been assumed to be  $\vec{v} = \omega_c R_c \vec{c}_1$ . Then the energy function

$$h_i = T_{i2} - T_{i0} + U \quad (8.391)$$

$$= \frac{1}{2} \vec{\omega}_{ib} \cdot \vec{M}_c \cdot \vec{\omega}_{ib} + \frac{3}{2} \omega_c^2 \vec{c}_3 \cdot \vec{M}_c \cdot \vec{c}_3 + U_0 - \frac{1}{2} m \omega_c^2 R_c^2 \quad (8.392)$$

is constant as long as the system is not actuated, and as the last two terms of the energy function are constants, the function

$$V_i = \frac{1}{2} \vec{\omega}_{ib} \cdot \vec{M}_c \cdot \vec{\omega}_{ib} + \frac{3}{2} \omega_c^2 \vec{c}_3 \cdot \vec{M}_c \cdot \vec{c}_3 \quad (8.393)$$

will be a constant of motion.

A change of variables using

$$\vec{\omega}_{cb} = \vec{\omega}_{ib} - \vec{\omega}_{ic} \quad (8.394)$$

$$= \vec{\omega}_{ib} - \omega_c \vec{c}_2 \quad (8.395)$$

will give the following expression for the kinetic energy

$$T = \underbrace{\frac{1}{2} \vec{\omega}_{ib} \cdot \vec{M}_c \cdot \vec{\omega}_{ib}}_{T_2} - \underbrace{\omega_c \vec{\omega}_{ib} \cdot \vec{M}_c \cdot \vec{c}_2}_{T_1} + \underbrace{\frac{1}{2} \omega_c^2 \vec{c}_2 \cdot \vec{M}_c \cdot \vec{c}_2 + \frac{1}{2} m \omega_c^2 R_c^2}_{T_0} \quad (8.396)$$

Then the energy function is found to be

$$\begin{aligned} h &= T_2 - T_0 + U \\ &= \frac{1}{2} \vec{\omega}_{cb} \cdot \vec{M}_c \cdot \vec{\omega}_{cb} + \frac{3}{2} \omega_c^2 \vec{c}_3 \cdot \vec{M}_c \cdot \vec{c}_3 - \frac{1}{2} \omega_c^2 \vec{c}_2 \cdot \vec{M}_c \cdot \vec{c}_2 + U_0 - \frac{1}{2} m \omega_c^2 R_c^2 \end{aligned}$$

Again, the last two terms of the energy function are constants, and it follows that the function

$$V_c = \frac{1}{2} \vec{\omega}_{cb} \cdot \vec{M}_c \cdot \vec{\omega}_{cb} + \frac{3}{2} \omega_c^2 \vec{c}_3 \cdot \vec{M}_c \cdot \vec{c}_3 - \frac{1}{2} \omega_c^2 \vec{c}_2 \cdot \vec{M}_c \cdot \vec{c}_2 \quad (8.397)$$

is a constant for the unactuated system.

The function  $V_c$  was used for Lyapunov analysis of gravity gradient stabilization of a satellite in (Hughes 1986). What can be learned from this example is that nontrivial energy functions for Lyapunov analysis can be derived using Hamilton theory, and in particular, that this approach offer a systematic way of changing coordinates in the description.

## 8.10 The Hamilton-Jacobi equation

In this section the Hamilton-Jacobi equation will be developed from Hamilton's principle (Lovelock and Rund 1989). The main idea is that the dynamics of a system with zero input forces are found by minimization of the action integral  $A(C)$ . Then, by introducing a modified Lagrangian  $L^*$ , it is possible to define the velocity of the system as a velocity field that satisfies a partial differential equation known as the Hamilton-Jacobi equation. This method is the underlying idea for the use of the Hamilton-Jacobi equation in optimal control.

Let the Lagrangian  $L(\mathbf{q}, \dot{\mathbf{q}}, t)$  and the associated momentum vector

$$\mathbf{p} = \frac{\partial L(\mathbf{q}, \dot{\mathbf{q}}, t)}{\partial \dot{\mathbf{q}}}^T \quad (8.398)$$

be given. Suppose that the velocity can be given by the function

$$\dot{\mathbf{q}} = \phi(\mathbf{q}, \mathbf{p}, t) \quad (8.399)$$

The Hamiltonian is then

$$H(\mathbf{q}, \mathbf{p}, t) = \mathbf{p} \phi(\mathbf{q}, \mathbf{p}, t) - L(\mathbf{q}, \phi(\mathbf{q}, \mathbf{p}, t), t) \quad (8.400)$$

The action integral  $A(C)$  as defined in Section 8.6.4 satisfies  $\delta A(C) = 0$  for a trajectory  $C$  with fixed end points, and this implies Lagrange's equation of motion. We

introduce a function  $S(\mathbf{q}, t)$ , which is yet to be specified, and define the *alternative Lagrangian* function

$$L^*(\mathbf{q}, \dot{\mathbf{q}}, t) = L(\mathbf{q}, \dot{\mathbf{q}}, t) - \frac{dS(\mathbf{q}, t)}{dt} \quad (8.401)$$

From the chain rule it follows that

$$\frac{dS}{dt} = \frac{\partial S}{\partial t} + \frac{\partial S}{\partial \mathbf{q}} \dot{\mathbf{q}} \quad (8.402)$$

Define the *alternative action integral*  $A^*(C)$  by

$$A^*(C) = \int_{t_1}^{t_2} L^* dt = \int_{t_1}^{t_2} \left( L - \frac{dS}{dt} \right) dt = \int_{t_1}^{t_2} L dt - \int_{t_1}^{t_2} dS \quad (8.403)$$

This implies that

$$A^*(C) = A(C) - (S_2 - S_1) \quad (8.404)$$

where  $S_1$  and  $S_2$  are the values of  $S$  at the end points. As the difference  $S_2 - S_1$  is independent of the curve  $C$ , it follows that minimum value for  $A^*(C)$  is found for the same trajectory as the minimum for  $A(C)$ . Therefore the system can be analyzed in terms of  $A^*(C)$  and the function  $L^*(\mathbf{q}, \dot{\mathbf{q}}, t)$  instead of  $A(C)$  and  $L(\mathbf{q}, \dot{\mathbf{q}}, t)$ .

Suppose that a velocity field  $\psi(\mathbf{q}, t)$  defined according to

$$\dot{\mathbf{q}} = \phi(\mathbf{q}, \mathbf{p}, t) = \psi(\mathbf{q}, t) \quad (8.405)$$

so that for a suitably selected function  $S$  the alternative Lagrangian satisfies

$$L^*(\mathbf{q}, \psi(\mathbf{q}, t), t) = 0 \quad (8.406)$$

$$L^*(\mathbf{q}, \dot{\mathbf{q}}, t) > 0 \text{ when } \dot{\mathbf{q}} \neq \psi(\mathbf{q}, t) \quad (8.407)$$

Then the field  $\psi(\mathbf{q}, t)$  is called the *geodesic field*. Note that on the geodesic field  $L^* = 0$  and (8.401) implies that

$$\frac{dS}{dt} = L \quad (8.408)$$

We will now solve our problem under the assumption that a geodesic field exists. The geodesic field  $\dot{\mathbf{q}} = \psi(\mathbf{q}, t)$  may be integrated with respect to the time  $t$  to give a family of curves. We let  $C$  denote one of the curves of this family with initial time  $t_1$  corresponding to a point  $P_1$ , and final time  $t_2$  corresponding to a point  $P_2$ . Then  $\dot{\mathbf{q}} = \psi(\mathbf{q}, t)$  at each point along the curve. Moreover, along the curve we have  $A^*(C) = 0$ , whereas  $A^*(K) > 0$  for any other curve  $K$  between the points  $P_1$  and  $P_2$ . This means that the curve  $C$  minimizes the alternative action integral  $A^*$ , which implies that  $C$  is the solution to the minimization of the action integral  $A$ . The curve  $C$  is therefore the solution to Lagrange's equation of motion with the Lagrangian  $L$ . This means that if we are able to find the velocity field  $\psi(\mathbf{q}, t)$ , then we have the solution of the equation of motion for any initial condition.

By introducing the function  $S$  and the geodesic field  $\psi(\mathbf{q}, t)$  we have changed the problem of minimizing the action integral  $A(C)$  with respect to  $C$  over the time interval  $t_1 \leq t \leq t_2$ , into the problem of minimizing the function  $L^*(\mathbf{q}, \dot{\mathbf{q}}, t)$  with respect to  $\dot{\mathbf{q}}$  for each  $\mathbf{q}$  and each  $t$ . This is simply a minimization problem for a function. In analogy with the result from basic calculus that the minimum for  $f(x)$  is found for  $df/dx = 0$ , the minimum of  $L^*(\mathbf{q}, \dot{\mathbf{q}}, t)$  for fixed  $\mathbf{q}$  and  $t$  is found for

$$\frac{\partial L^*}{\partial \dot{\mathbf{q}}} = \mathbf{0} \quad (8.409)$$

This gives

$$\mathbf{0} = \frac{\partial L^*}{\partial \dot{\mathbf{q}}}^T = \frac{\partial L}{\partial \dot{\mathbf{q}}}^T - \frac{\partial S}{\partial \mathbf{q}}^T = \mathbf{p} - \frac{\partial S}{\partial \mathbf{q}}^T \quad (8.410)$$

which implies that

$$\mathbf{p} = \frac{\partial S}{\partial \mathbf{q}} \quad (8.411)$$

on the geodesic field. Moreover, from (8.402) and (8.408) it follows that

$$L[\mathbf{q}, \psi(\mathbf{q}, t), t] = \frac{dS}{dt} = \frac{\partial S}{\partial t} + \mathbf{p}^T \psi(\mathbf{q}, t) \quad (8.412)$$

This can be combined with

$$H(\mathbf{q}, \mathbf{p}, t) = \mathbf{p}^T \psi(\mathbf{q}, t) - L(\mathbf{q}, \dot{\mathbf{q}}, t) \quad (8.413)$$

and the following result is found:

The momentum vector  $\mathbf{p}$  can be found as a vector field

$$\mathbf{p}(\mathbf{q}, t) = \frac{\partial S(\mathbf{q}, t)}{\partial \mathbf{q}}^T \quad (8.414)$$

where  $S(\mathbf{q}, t)$  is the solution of the Hamilton-Jacobi equation

$$\frac{\partial S(\mathbf{q}, t)}{\partial t} + H\left[\mathbf{q}, \frac{\partial S(\mathbf{q}, t)}{\partial \mathbf{q}}, t\right] = 0 \quad (8.415)$$

which is a partial differential equation in  $S(\mathbf{q}, t)$ .

**Example 144** The time derivative of  $\mathbf{p}$  is found to be

$$\dot{\mathbf{p}} = \frac{d}{dt} \frac{\partial S(\mathbf{q}, t)}{\partial \mathbf{q}}^T = \frac{\partial^2 S(\mathbf{q}, t)}{\partial t \partial \mathbf{q}}^T + \frac{\partial^2 S(\mathbf{q}, t)}{\partial \mathbf{q} \partial \mathbf{q}} \dot{\mathbf{q}} \quad (8.416)$$

Partial differentiation of the Hamilton-Jacobi equation gives

$$\frac{\partial^2 S}{\partial t \partial \mathbf{q}}^T + \frac{\partial H}{\partial \mathbf{q}}^T + \frac{\partial H}{\partial \mathbf{p}} \frac{\partial^2 S}{\partial \mathbf{q} \partial \mathbf{q}} = 0 \quad (8.417)$$

Insertion of  $\dot{\mathbf{q}} = (\partial H / \partial \mathbf{p})^T$  gives

$$\frac{\partial^2 S}{\partial t \partial \mathbf{q}}^T + \frac{\partial^2 S}{\partial \mathbf{q} \partial \mathbf{q}} \dot{\mathbf{q}} = -\frac{\partial H}{\partial \mathbf{q}} \quad (8.418)$$

This shows that

$$\dot{\mathbf{p}} = -\frac{\partial H}{\partial \mathbf{q}}^T \quad (8.419)$$

and it has been established that the solution of the Hamilton-Jacobi equation is consistent with the Hamilton's equations of motion

$$\dot{\mathbf{q}} = \frac{\partial H}{\mathbf{p}}^T \quad (8.420)$$

$$\dot{\mathbf{p}} = -\frac{\partial H}{\partial \mathbf{q}}^T \quad (8.421)$$

# Chapter 9

## Mechanical vibrations

### 9.1 Introduction

Active damping of mechanical vibrations has been important for space structures where there is little damping and where the structures are designed for low weight. As new inexpensive sensors and actuators are becoming available, active vibration damping is being more used also in civil engineering, crane systems, and transportation. In this chapter vibration models will be developed for the string model and the Euler Bernoulli beam model. These models represent important properties that are seen for models of mechanical vibrations. Models are developed using assumed modes, mostly in the form of orthogonal modes, and finite-element models. The Galerkin method is used to illustrate similarities between the assumed mode method and the finite-element approach. Also irrational transfer functions are derived, and examples with positive realness and nonminimum phase dynamics are discussed. The chapter starts with systems with lumped components, and progresses with distributed parameter models.

Elastic systems consisting of rigid bodies connected with springs and dampers are described by ordinary differential equations, and are said to be lumped parameter systems. In contrast to this, systems with elastic bodies are described by partial differential equations, and are said to be distributed parameter systems. In this chapter we will first present results on elastic systems with lumped parameters and then proceed with results on distributed parameter systems.

Energy considerations and passivity properties are important in the analysis and controller design for elastic systems with distributed parameters. We will see that if the applied force at a point is the input, and the velocity at the same point is the output, then the input and output are said to be *collocated*, and the system is passive. A system where force is input and velocity is output will not be passive if the input and output are *noncollocated*, that is, if the velocity is not measured at the point where the force is applied. Such systems may even be nonminimum phase, which may cause severe restrictions on the performance of the closed loop system. In this section we will study these phenomena closer. The material is adopted from (Meirovitch 1967), (Weaver, Timoshenko and Young 1990) and (Rao 1990).

## 9.2 Lumped elastic two-ports

### 9.2.1 Hybrid two-port

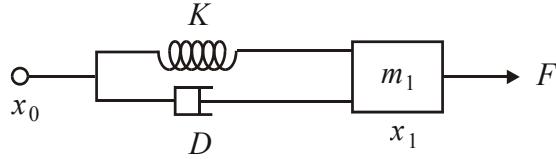


Figure 9.1: Mechanical two-port of the hybrid type with inputs  $\dot{x}_0$  and  $F$ .

The equation of motion for a mass  $m$  that is connected to a point  $x_0$  with a spring with stiffness  $K$  and a damper with coefficient  $D$  is given by

$$m\ddot{x} + D(\dot{x} - \dot{x}_0) + K(x - x_0) = F \quad (9.1)$$

Here  $x$  is the position of the mass, and  $F$  is a force acting on the mass. The system is shown in Figure 9.1. In the following this simple system will be used as a building block in some of the models in this section. In this connection it is useful to have a two-port description of the system where one port has input  $\dot{x}_0$  and output  $F_0 = D(\dot{x} - \dot{x}_0) + K(x - x_0)$ , and the other port has input  $F$  and output  $\dot{x}$ . This will be termed a hybrid formulation as one port has the force as output, and the other has velocity as output. The ports are therefore compatible for a serial interconnection.

### 9.2.2 Displacement two-port

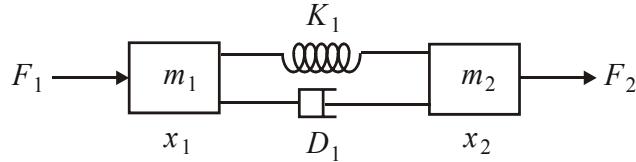


Figure 9.2: Mechanical two-port of the displacement type with inputs  $F_1$  and  $F_2$ .

Another possible building block for models of lumped elastic systems is two masses connected by a spring and a damper. The system has two masses  $m_1$  and  $m_2$  connected by a spring  $K_1$  and a damper  $D_1$ . An externally applied force  $F_1$  is acting on mass  $m_1$ , and a force  $F_2$  is acting on  $m_2$ . The position of mass  $i$  is denoted  $x_i$ . The system is shown in Figure 9.2, and the equations of motion for the system are

$$m_1\ddot{x}_1 + D_1(\dot{x}_1 - \dot{x}_2) + K_1(x_1 - x_2) = F_1 \quad (9.2)$$

$$m_2\ddot{x}_2 + D_1(\dot{x}_2 - \dot{x}_1) + K_1(x_2 - x_1) = F_2 \quad (9.3)$$

This can be written in vector form as

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{D}\dot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \mathbf{F} \quad (9.4)$$

where  $\mathbf{x} = (x_1, x_2)^T$ ,  $\mathbf{F} = (F_1, F_2)^T$ , and the mass matrix  $\mathbf{M}$ , the damping matrix  $\mathbf{D}$  and the stiffness matrix  $\mathbf{K}$  are given by

$$\mathbf{M} = \begin{pmatrix} m_1 & 0 \\ 0 & m_2 \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} D_1 & -D_1 \\ -D_1 & D_1 \end{pmatrix}, \quad \mathbf{K} = \begin{pmatrix} K_1 & -K_1 \\ -K_1 & K_1 \end{pmatrix} \quad (9.5)$$

In a network setting the system described by (9.2) and (9.3) is a two-port where port 1 has input  $F_1$  and output  $\dot{x}_1$ , and port 2 has input force  $F_2$  and output  $\dot{x}_2$ . This type of two-port will be said to be in a displacement formulation as both ports have force as input and velocity as output.

**Example 145** *The ports of a displacement two-port are compatible with connections to springs and dampers, which have velocity as input and force as output. To demonstrate this we consider the case where mass  $m_1$  is connected to a fixed point with a spring  $K_0$  and a damper  $D_0$ , and mass  $m_2$  is connected to a fixed point with a spring  $K_2$  and a damper  $D_2$ . This means that the input port is connected to the one-port*

$$F_1 = -(K_0 x_1 + D_0 \dot{x}_1) \quad (9.6)$$

and that the output port is connected to the one-port

$$F_2 = -(K_2 x_2 + D_2 \dot{x}_2) \quad (9.7)$$

Then the damping and stiffness matrices for the interconnected system is obtained by inserting  $F_1$  and  $F_2$  into (9.2) and (9.3) which gives

$$\mathbf{D} = \begin{pmatrix} D_0 + D_1 & -D_1 \\ -D_1 & D_1 + D_2 \end{pmatrix}, \quad \mathbf{K} = \begin{pmatrix} K_0 + K_1 & -K_1 \\ -K_1 & K_1 + K_2 \end{pmatrix} \quad (9.8)$$

### 9.2.3 Three masses in the hybrid formulation

Consider the three hybrid two-ports

$$m_1 \ddot{x}_1 = F_1 \quad (9.9)$$

$$m_2 \ddot{x}_2 + D_1 (\dot{x}_2 - \dot{x}_{20}) + K_1 (x_2 - x_{20}) = F_2 \quad (9.10)$$

$$m_3 \ddot{x}_3 + D_2 (\dot{x}_3 - \dot{x}_{30}) + K_2 (x_3 - x_{30}) = F_3 \quad (9.11)$$

The two-ports have compatible ports variables, and can be connected by assigning the input forces to be

$$F_1 = D_1 (\dot{x}_2 - \dot{x}_1) + K_1 (x_2 - x_1) + \tau_1 \quad (9.12)$$

$$F_2 = D_2 (\dot{x}_3 - \dot{x}_2) + K_2 (x_3 - x_2) + \tau_2 \quad (9.13)$$

$$F_3 = \tau_3 \quad (9.14)$$

and the input displacements to be

$$x_{20} = x_1, \quad x_{30} = x_2 \quad (9.15)$$

This gives the total system

$$m_1 \ddot{x}_1 + D_1 (\dot{x}_1 - \dot{x}_2) + K_1 (x_1 - x_2) = \tau_1 \quad (9.16)$$

$$m_2 \ddot{x}_2 + D_1 (\dot{x}_2 - \dot{x}_1) + K_1 (x_2 - x_1) \quad (9.17)$$

$$+ D_2 (\dot{x}_2 - \dot{x}_3) + K_2 (x_2 - x_3) = \tau_2 \quad (9.18)$$

$$m_3 \ddot{x}_3 + D_2 (\dot{x}_3 - \dot{x}_2) + K_2 (x_3 - x_2) = \tau_3 \quad (9.19)$$

In vector form this is written

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{D}\dot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \boldsymbol{\tau} \quad (9.20)$$

where  $\mathbf{x} = (x_1, x_2, x_3)^T$ ,  $\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3)^T$ , and

$$\mathbf{M} = \begin{pmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_2 \end{pmatrix} \quad (9.21)$$

$$\mathbf{D} = \begin{pmatrix} D_1 & -D_1 & 0 \\ -D_1 & D_1 + D_2 & -D_2 \\ 0 & -D_2 & D_2 \end{pmatrix} \quad (9.22)$$

$$\mathbf{K} = \begin{pmatrix} K_1 & -K_1 & 0 \\ -K_1 & K_1 + K_2 & -K_2 \\ 0 & -K_2 & K_2 \end{pmatrix} \quad (9.23)$$

### 9.2.4 Three masses in the displacement formulation

In this section we will derive the result of the previous section using displacement two-ports. This procedure is of great interest as it resembles the method that is used to interconnect elements in the displacement formulation of the finite element method. This will be done by connecting two systems given as the displacement two-ports

$$m_1\ddot{x}_1 + D_1(\dot{x}_1 - \dot{x}_2) + K_1(x_1 - x_2) = \tau_1 \quad (9.24)$$

$$\frac{m_2}{2}\ddot{x}_2 + D_1(\dot{x}_2 - \dot{x}_1) + K_1(x_2 - x_1) = F_{12} \quad (9.25)$$

and

$$\frac{m_2}{2}\ddot{x}_2 + D_2(\dot{x}_2 - \dot{x}_3) + K_2(x_2 - x_3) = F_{21} \quad (9.26)$$

$$m_3\ddot{x}_3 + D_2(\dot{x}_3 - \dot{x}_2) + K_2(x_3 - x_2) = \tau_3 \quad (9.27)$$

The interconnection is done by requiring that the second mass of system 1 has the same position as the first mass of system 2, that is, by requiring that  $\dot{x}_2$  is the same for the two systems. In this case the port variables are not compatible, so the equations of motion must be combined. To see how the equations must be combined it is noted that to ensure that the masses are interconnected, there must be a constraint force  $F^{(c)}$  so that the forces acting on the two masses are given by

$$F_{12} = F^{(c)} + \frac{1}{2}\tau_2, \quad F_{21} = -F^{(c)} + \frac{1}{2}\tau_2 \quad (9.28)$$

To derive the equations of motion of the interconnected system it is necessary to eliminate the constraint force  $F^{(c)}$ . This can be done by simply adding the equations (9.25) and (9.26). This gives the equations of motion

$$m_1\ddot{x}_1 + D_1(\dot{x}_1 - \dot{x}_2) + K_1(x_1 - x_2) = \tau_1 \quad (9.29)$$

$$m_2\ddot{x}_2 + D_1(\dot{x}_2 - \dot{x}_1) + K_1(x_2 - x_1) + D_2(\dot{x}_2 - \dot{x}_3) + K_2(x_2 - x_3) = \tau_2 \quad (9.30)$$

$$+ D_2(\dot{x}_3 - \dot{x}_2) + K_2(x_3 - x_2) = \tau_3 \quad (9.31)$$

$$m_3\ddot{x}_3 + D_2(\dot{x}_3 - \dot{x}_2) + K_2(x_3 - x_2) = \tau_3 \quad (9.32)$$

which is the same result as the one derived with the hybrid formulation in the previous section.

### 9.2.5 Four masses

It is straightforward to connect one more mass to the system to have four interconnected masses  $m_1, m_2, m_3$  and  $m_4$ . In addition, springs can be connected to terminate the ports of mass 1 and 4. The equation of motion for the system is found to be

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{D}\dot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \boldsymbol{\tau} \quad (9.33)$$

Here  $\mathbf{x} = (x_1, x_2, x_3, x_4)^T$ ,  $\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3, \tau_4)^T$ ,  $\mathbf{M} = \text{diag}(m_1, m_2, m_3, m_4)$  and the damping and stiffness matrices are

$$\mathbf{D} = \begin{pmatrix} D_0 + D_1 & -D_1 & 0 & 0 \\ -D_1 & D_1 + D_2 & -D_2 & 0 \\ 0 & -D_2 & D_2 + D_3 & -D_3 \\ 0 & 0 & -D_3 & D_3 + D_4 \end{pmatrix} \quad (9.34)$$

$$\mathbf{K} = \begin{pmatrix} K_0 + K_1 & -K_1 & 0 & 0 \\ -K_1 & K_1 + K_2 & -K_2 & 0 \\ 0 & -K_2 & K_2 + K_3 & -K_3 \\ 0 & 0 & -K_3 & K_3 + K_4 \end{pmatrix} \quad (9.35)$$

The total energy of the system is

$$V = \sum_{i=1}^4 \frac{1}{2} m_i \dot{x}_i^2 + \sum_{i=1}^2 \frac{1}{2} K_i (x_i - x_{i+1})^2 + \frac{1}{2} K_0 x_0^2 + \frac{1}{2} K_4 x_4^2 \quad (9.36)$$

The time derivative of the energy for the solutions of the system will be the power  $\sum_{i=1}^4 \tau_i v_i$  supplied by the inputs  $\tau_i$  minus the power dissipated in the dampers. This is written

$$\dot{V} = \sum_{i=1}^4 \tau_i v_i - \sum_{i=1}^3 D_i (\dot{x}_i - \dot{x}_{i+1}) - \frac{1}{2} D_0 \dot{x}_0^2 - \frac{1}{2} D_4 \dot{x}_4^2 \quad (9.37)$$

Suppose that  $\tau_2 = \tau_3 = \tau_4 = 0$ . Then the system with input  $\tau_1$  and output  $v_1$  will be passive.

## 9.3 Vibrating string

### 9.3.1 Linearized model

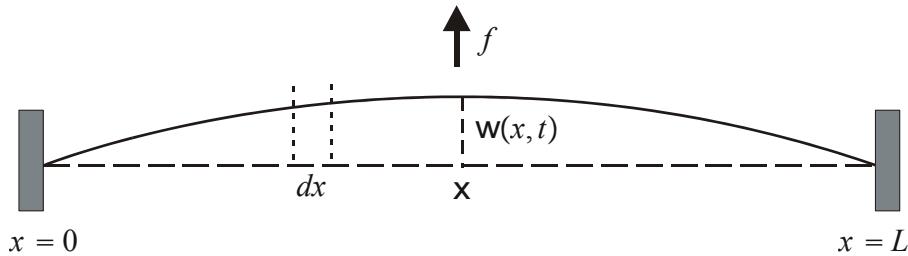


Figure 9.3: Vibrating string of length  $L$ .

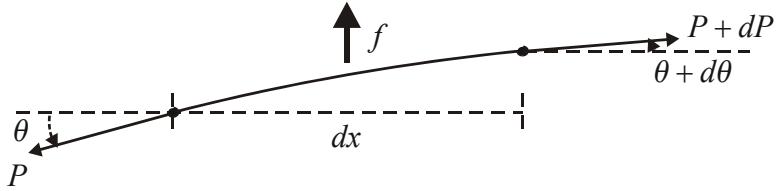


Figure 9.4: Differential string element.

A stretched string with tension  $P$  is fixed at  $x = 0$  and  $x = L$  (Figure 9.3). The string is supposed to have small transverse displacements  $w(x, t)$ , and is excited by a transverse force  $f(x, t)$ . The equation of motion for a differential element of the string as shown in Figure 9.4 is

$$\rho dx \frac{\partial^2 w}{\partial t^2} = (P + dP) \sin(\theta + d\theta) - P \sin \theta + f dx \quad (9.38)$$

where  $\rho dx$  is the mass of the element, and  $\theta$  is the slope of the string. Division with  $dx$  leads to the partial differential equation

$$\rho \frac{\partial^2 w}{\partial t^2} = \frac{\partial}{\partial x} (P \sin \theta) + f \quad (9.39)$$

For small angles we may approximate the sine function by

$$\sin \theta = \frac{\partial w}{\partial x} \quad (9.40)$$

which gives

$$\rho \frac{\partial^2 w}{\partial t^2} = \frac{\partial}{\partial x} \left( P \frac{\partial w}{\partial x} \right) + f \quad (9.41)$$

If the tension is constant along the string, then  $P$  is a constant and the model is

$$\rho \frac{\partial^2 w}{\partial t^2} = P \frac{\partial^2 w}{\partial x^2} + f \quad (9.42)$$

The homogeneous form

$$\frac{\partial^2 w}{\partial t^2} = c^2 \frac{\partial^2 w}{\partial x^2} \quad (9.43)$$

of (9.42) is called the wave equation where

$$c = \sqrt{\frac{P}{\rho}} \quad (9.44)$$

$c$  is the propagation speed of the waves.

### 9.3.2 Orthogonal shape functions

The homogeneous problem as given by the wave equation (9.43) can be solved by separation of variables  $w(x, t) = q(t)\phi(x)$ . Then, according to basic textbooks on mathematics, the following expression is found

$$\frac{\ddot{q}(t)}{q(t)} = \frac{c^2 \phi''(x)}{\phi(x)} = -\omega^2 \quad (9.45)$$

where  $\omega$  is a constant to be determined. This leads to the differential equations

$$\ddot{q}(t) + \omega^2 q(t) = 0 \quad (9.46)$$

$$\phi''(x) + \frac{\omega^2}{c^2} \phi(x) = 0 \quad (9.47)$$

We investigate the solution of the equation of the vibrating string when the end-points are fixed, that is, when  $w(0, t) = w(L, t) = 0$ , which implies that  $\phi(0) = \phi(L) = 0$ . Then there are infinitely many solutions

$$\phi_k(x) = \sqrt{\frac{2}{\rho L}} \sin \frac{k\pi}{L} x, \quad k = 1, 2, \dots \quad (9.48)$$

and each solution  $\phi_k(x)$  corresponds to the value

$$\omega_k = \frac{k\pi}{L} c \quad (9.49)$$

and the solution  $q_k$  of

$$\ddot{q}_k + \omega_k^2 q_k = 0 \quad (9.50)$$

The solution of the wave equation can then be written

$$w(x, t) = \sum_{k=1}^{\infty} q_k(t) \phi_k(x) \quad (9.51)$$

Note that the shape functions are orthogonal in the sense that

$$\int_0^L \rho \phi_j(x) \phi_k(x) dx = \delta_{jk} \quad (9.52)$$

Moreover, the derivatives of the mode shapes are orthogonal, so that

$$\int_0^L \phi'_j(x) \phi'_k(x) dx = \left( \frac{k\pi}{L} \right)^2 \delta_{jk} \quad (9.53)$$

### 9.3.3 Galerkin's method for orthogonal shape functions

The method of separation of variables works well for the wave equation when there is no forcing term  $f$ . If there is a forcing term, then Galerkin's method can be used (Joshi 1989). In this method a solution  $w(x, t) = \sum_{j=1}^{\infty} q_j(t) \phi_j(x)$  is assumed where  $\phi_j(x)$  belong to some set of shape functions. Orthogonal shape functions as found by the separation of variables will be assumed. The equation of motion is obtained by multiplying the wave equation by a shape function  $\phi_i(s)$  and then integrating over the interval. This gives

$$\int_0^L \phi_i(x) \sum_{k=1}^{\infty} \rho [\ddot{q}_j(t) \phi_j(x) - c^2 q_j(t) \phi''_j(x)] dx = \int_0^L \phi_i(x) f(x, t) dx \quad (9.54)$$

It is standard procedure to use partial integration for the term

$$\begin{aligned} \int_0^L \phi_i(x) \phi''_j(x) dx &= - \int_0^L \phi'_i(x) \phi'_j(x) dx + \phi_i(x) \phi'_j(x) \Big|_0^\ell \\ &= - \int_0^L \phi'_i(x) \phi'_j(x) dx \end{aligned} \quad (9.55)$$

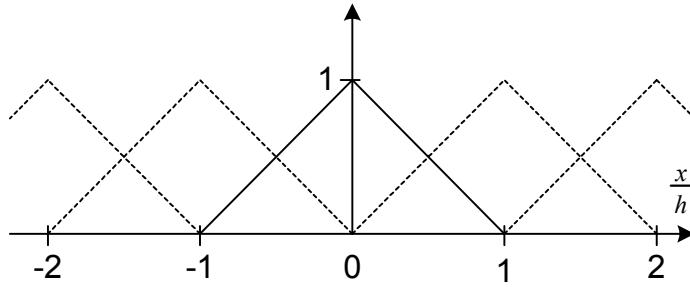


Figure 9.5: Triangular shape functions  $\psi_j(x)$ .

Due to the orthogonality of the shape functions and the derivatives this simplifies to

$$\ddot{q}_i(t) + \omega_i^2 q_i(t) = \int_0^L \phi_i(x) f(x, t) dx \quad (9.56)$$

**Example 146** A point force  $F(t)$  at  $x_F$  can be modelled as  $f(x) = F(t)\delta(x - x_F)$ . In that case the equation of motion becomes

$$\ddot{q}_i(t) + \omega_i^2 q_i(t) = \phi_i(x_F) F(t) \quad (9.57)$$

We see that if  $\phi_i(x_F) = 0$  for some  $i$ , then the force  $F(t)$  will have no influence on  $q_i(t)$ . In control terminology, this means that  $q_i$  is not controllable when  $F$  is the control input.

### 9.3.4 Finite element shape functions

Instead of the orthogonal shape functions another set of shape functions will be introduced in this section. This is done by using shape functions  $\psi_j(x)$  leading to a finite-element model. First we define  $N$  points along the string that are called nodes. The position of node  $j$  is denoted  $x_j$ . The distance between the nodes is set to  $h$ . The  $N$  piece-wise linear shape functions

$$\psi_j(x) = \begin{cases} \frac{x-x_{j-1}}{h}, & x_{j-1} \leq x \leq x_j \\ \frac{x_{j+1}-x}{h}, & x_j \leq x \leq x_{j+1} \\ 0, & \text{otherwise} \end{cases}, \quad j = 1, 2, \dots, N \quad (9.58)$$

that are shown in Figure 9.5 are used, and the solution is approximated by

$$w(x, t) = \sum_{j=1}^N q_j(t) \psi_j(x) \quad (9.59)$$

Note that

$$\psi_j(x_k) = \delta_{jk} \quad (9.60)$$

which implies that the displacement at node  $k$  is given by

$$w(x_k, t) = q_k(t) \quad (9.61)$$

We insert the approximation into the wave equation (9.42) and get

$$\sum_{k=1}^N [\rho \ddot{q}_j(t) \psi_j(x) - P q_j(t) \psi_j''(x)] = f(x, t) \quad (9.62)$$

We apply Galerkin's method, which in this case involves the multiplication with  $\psi_i(x)$  and integration over the interval. This gives

$$\int_0^L \psi_i(x) \sum_{k=1}^N [\rho \ddot{q}_j(t) \psi_j(x) - P q_j(t) \psi_j''(x)] dx = \int_0^L \psi_i(x) f(x, t) dx \quad (9.63)$$

Partial integration gives

$$\sum_{k=1}^N \int_0^L \rho \psi_i(x) \psi_j(x) dx \ddot{q}_j(t) + \sum_{k=1}^N P \int_0^L \psi_i'(x) \psi_j'(x) dx q_j(t) = \int_0^L \psi_i(x) f(x, t) dx \quad (9.64)$$

This may be written in matrix form as

$$\mathbf{M} \ddot{\mathbf{q}} + \mathbf{K} \mathbf{q} = \mathbf{F} \quad (9.65)$$

where  $\mathbf{q} = (q_1, q_2, \dots, q_N)^T$ ,  $\mathbf{M} = \{m_{ij}\}$ ,  $\mathbf{K} = \{k_{ij}\}$  and  $\mathbf{F} = (f_1, f_2, \dots, f_N)^T$ . The components are given by

$$m_{ij} = \rho \int_0^L \psi_i(x) \psi_j(x) dx \quad (9.66)$$

$$k_{ij} = P \int_0^L \psi_i'(x) \psi_j'(x) dx \quad (9.67)$$

$$f_i = \int_0^L \psi_i(x) f(x, t) dx \quad (9.68)$$

### 9.3.5 String element

The usual way of establishing a finite element model for this system is to define a string element, and then to generate a model for the string by assembling string elements. The string element is a model of a string of length  $h$  between two nodes. The elements of the mass matrix  $\mathbf{M}_e$  of the element is then

$$m_{11} = \rho \int_0^h \left(1 - \frac{x}{h}\right)^2 dx = \frac{\rho h}{3}, \quad m_{22} = m_{11} \quad (9.69)$$

$$m_{12} = m_{21} = \rho \int_0^h \left(1 - \frac{x}{h}\right) \frac{x}{h} dx = \frac{\rho h}{6} \quad (9.70)$$

while the stiffness matrix  $\mathbf{K}_e$  of the element has elements

$$k_{11} = P \int_0^h \left(\frac{-1}{h}\right)^2 dx = \frac{P}{h}, \quad k_{22} = k_{11} \quad (9.71)$$

$$k_{1,2} = k_{2,1} = P \int_0^h \frac{-1}{h^2} dx = -\frac{P}{h} \quad (9.72)$$

The model for the string element is found to be

$$\mathbf{M}_e \ddot{\mathbf{q}}_e + \mathbf{K}_e \mathbf{q}_e = \mathbf{F}_e \quad (9.73)$$

where

$$\mathbf{q}_e = \begin{pmatrix} q_{e1} \\ q_{e2} \end{pmatrix}, \quad \mathbf{F}_e = \begin{pmatrix} F_{e1} \\ F_{e2} \end{pmatrix} \quad (9.74)$$

$$\mathbf{M}_e = \frac{\rho h}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{K}_e = \frac{P}{h} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \quad (9.75)$$

### 9.3.6 Assembling string elements

Two string elements can be assembled by specifying that the second coordinate of the first element equals the first coordinate of the second element. This is done by assigning the coordinates to be  $q_1$  and  $q_2$  for the first element and  $q_2$  and  $q_3$  for the second element. To keep the two elements together there must be a constraint force  $F_2^{(c)}$ . The models for the two elements are

$$\mathbf{M}_e \begin{pmatrix} \ddot{q}_1 \\ \ddot{q}_2 \end{pmatrix} + \mathbf{K}_e \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} = \begin{pmatrix} F_1 \\ \frac{1}{2}F_2 + F_2^{(c)} \end{pmatrix} \quad (9.76)$$

$$\mathbf{M}_e \begin{pmatrix} \ddot{q}_2 \\ \ddot{q}_3 \end{pmatrix} + \mathbf{K}_e \begin{pmatrix} q_2 \\ q_3 \end{pmatrix} = \begin{pmatrix} \frac{1}{2}F_2 - F_2^{(c)} \\ F_3 \end{pmatrix} \quad (9.77)$$

The models of the elements are assembled by cancelling the constraint force, which is done by adding the last line of (9.76) with the first line of (9.77). This gives

$$\mathbf{M} \begin{pmatrix} \ddot{q}_1 \\ \ddot{q}_2 \\ \ddot{q}_3 \end{pmatrix} + \mathbf{K} \begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix} = \begin{pmatrix} F_1 \\ F_2 \\ F_3 \end{pmatrix} \quad (9.78)$$

$$\mathbf{M} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 2 \end{pmatrix}, \quad \mathbf{K} = \frac{P}{h} \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix} \quad (9.79)$$

Note that the mass matrix is obtained by adding the mass matrices of the elements in the sense that

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_e & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix} + \begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{M}_e \end{pmatrix} \quad (9.80)$$

In the same way the stiffness matrix is obtained from

$$\mathbf{K} = \begin{pmatrix} \mathbf{K}_e & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix} + \begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{K}_e \end{pmatrix} \quad (9.81)$$

This procedure can be repeated to assemble more elements. For 5 nodes the mass and stiffness matrix will be

$$\mathbf{M} = \frac{\rho h}{6} \begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 \\ 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix}, \quad \mathbf{K} = \frac{P}{h} \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{pmatrix} \quad (9.82)$$

**Example 147** The mass matrix as given by (9.66) is tridiagonal. It is called the consistent mass matrix as it is derived using the shape functions. It is possible to have a simpler model by using a lumped mass model where the mass is lumped at the nodes. This leads to a diagonal mass matrix, which is called the nonconsistent mass matrix. In the case of 5 nodes the nonconsistent mass matrix is

$$\mathbf{M} = \frac{\rho}{h^2} \begin{pmatrix} \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} \end{pmatrix} \quad (9.83)$$

Then the model  $\mathbf{M}\ddot{\mathbf{q}} + \mathbf{K}\mathbf{q} = \mathbf{F}$  describes the dynamics of a mass spring damper arrangement consisting of  $N$  masses that are interconnected by springs in a serial arrangement. When the consistent mass matrix is used there will in general be spring connections between all masses.

## 9.4 Nonlinear string dynamics

### 9.4.1 Kirchhoff's nonlinear string model

If the elastic deformation  $w$  is sufficiently large, then the tension will depend on the deformation. Let the tension be constant along the string and given by  $P = P_0 + EI\Delta x$  where

$$\Delta x = \int_0^L \sqrt{1 + \left(\frac{\partial w}{\partial x}\right)^2} dx - L \approx \frac{1}{2} \int_0^L \left(\frac{\partial w}{\partial x}\right)^2 dx \quad (9.84)$$

is the stretching of the string due to  $w$ . Then the homogeneous string model (9.43) becomes

$$\rho \frac{\partial^2 w}{\partial t^2} = P_0 \left( 1 + EI \int_0^L \left(\frac{\partial w}{\partial x}\right)^2 dx \right) \frac{\partial^2 w}{\partial x^2} \quad (9.85)$$

This is Kirchhoff's nonlinear string model (Shahruz 1999).

### 9.4.2 Marine cables

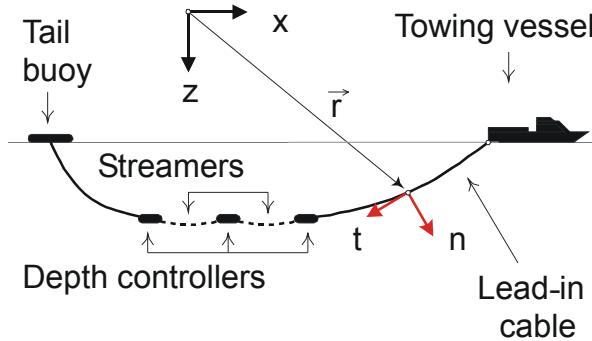


Figure 9.6: Towing arrangement for a marine seismic cable.

In this section the equation of motion for a towed marine cable will be presented. The derivation relies results that will be presented in Section 10.2. In particular this involves the definition of the material derivative  $D/Dt$  and the concept of material coordinates. Towed marine cables are used in marine seismic operation to map oil and gas reservoirs in an arrangement as indicated in Figure 9.6. Moreover, the model that will be presented is also valid for the dynamics of anchor lines.

The position of a point on a cable is described by the spatial length  $p$  to the point along the stretched cable, by the material length  $s$  to the point along the undeformed cable, and the spatial position  $\mathbf{r}(s, t)$  given in a spatial frame  $(x, y, z)$ . The length  $p$  is referred to as the length in spatial coordinates, while the length  $s$  is said to be the length in material coordinates. The displacement of a point from its undeformed position is denoted  $u = p - s$ . The material strain  $\eta$  is defined by

$$\eta = \frac{du}{ds} = \frac{dp}{ds} - 1 \Rightarrow \frac{dp}{ds} = \eta + 1 \quad (9.86)$$

We consider the cable element from  $s$  to  $s+ds$  of the unstretched cable. We call this a material cable element as it contains the same material points as the cable is moved and stretched. In the stretched case this cable element is from  $p$  to  $p+dp$ . The material cable element is of constant length  $ds$  in material coordinates, and of length  $dp = (1 + \eta)ds$  in spatial coordinates. The unit tangent vector of the stretched cable is

$$\mathbf{t} = \frac{\partial \mathbf{r}}{\partial p} = \frac{\partial \mathbf{r}}{\partial s} \frac{ds}{dp} = \frac{1}{1 + \eta} \frac{\partial \mathbf{r}}{\partial s} \quad (9.87)$$

As the tangent is a unit vector it can be written

$$\mathbf{t} = \frac{\partial \mathbf{r}}{\partial s} \left| \frac{\partial \mathbf{r}}{\partial s} \right|^{-1} \quad (9.88)$$

This shows that the material strain  $\eta$  is given by

$$\eta = \left| \frac{\partial \mathbf{r}}{\partial s} \right| - 1 \quad (9.89)$$

The volume of a material cable element is  $A_0 ds = Adp$  where  $A_0$  is the cross section of the unstretched cable, and  $A$  is the cross section of the stretched cable. The mass of a material cable element is constant and is given by  $dm = \rho_0 ds$  where  $\rho_0$  is the mass per unit length of the unstretched cable. A detailed discussion on material and spatial coordinates for this problem is found in (Lin and Segel 1974). The velocity  $\mathbf{v}(s, t)$  and the acceleration  $\mathbf{a}(s, t)$  of a point on the cable are given by

$$\mathbf{v}(s, t) = \frac{D\mathbf{r}(s, t)}{Dt} = \frac{\partial \mathbf{r}(s, t)}{\partial t}, \quad \mathbf{a}(s, t) = \frac{D\mathbf{v}(s, t)}{Dt} = \frac{\partial^2 \mathbf{r}(s, t)}{\partial t^2} \quad (9.90)$$

The equation of motion for the material cable element  $ds$  is given by

$$\frac{D}{Dt} (\mathbf{v} \rho_0 ds) = \rho_0 ds \frac{D\mathbf{v}}{Dt} = \rho_0 ds \frac{\partial^2 \mathbf{r}(s, t)}{\partial t^2} = (P + dP)\mathbf{t}(t, s+ds) - P\mathbf{t}(t, s) + \mathbf{f}(t, s)dp \quad (9.91)$$

where it is used that  $\rho_0 ds$  is a constant. Here  $P$  is the tension,  $\mathbf{t}$  is the tangent vector along the cable, and  $\mathbf{f}$  is the force per unit length of the stretched cable. Dividing by  $ds$  we get

$$\rho_0 \frac{\partial^2 \mathbf{r}(s, t)}{\partial t^2} = \frac{\partial}{\partial s} (P\mathbf{t}) + \mathbf{f} \frac{dp}{ds} \quad (9.92)$$

According to Hooke's law the tension in the cable is  $P = EA_0\eta$  where  $E$  is the Young's modulus, and  $A_0$  is the cross section of the unstretched cable. The force due to the tension  $P$  in the cable can be separated into a force along the tangent  $\mathbf{t}$  and one force orthogonal to the tangent according to

$$\rho_0 \frac{\partial^2 \mathbf{r}(s, t)}{\partial t^2} = \frac{\partial P}{\partial s} \mathbf{t} + P \frac{\partial \mathbf{t}}{\partial s} + \mathbf{f} \frac{dp}{ds} = EA_0 \frac{\partial \eta}{\partial s} \mathbf{t} + P \frac{\partial \mathbf{t}}{\partial p} \frac{dp}{ds} + \mathbf{f} \frac{dp}{ds} \quad (9.93)$$

This gives

$$\rho_0 \frac{\partial^2 \mathbf{r}(s, t)}{\partial t^2} = EA_0 \frac{\partial^2 u}{\partial s^2} \mathbf{t} + P(1 + \eta) \frac{\partial \mathbf{t}}{\partial p} + \mathbf{f}(1 + \eta) \quad (9.94)$$

**Example 148** In (Aamo and Fossen 2000) the formulation

$$\rho_0 \frac{\partial^2 \mathbf{r}}{\partial t^2} = \frac{\partial}{\partial s} \left( EA_0 \frac{\eta}{1 + \eta} \frac{\partial \mathbf{r}}{\partial s} \right) + \mathbf{f}(1 + \eta) \quad (9.95)$$

due to (Triantafyllou 1990) was used to develop a finite-element model of anchor lines for moored offshore vessels using Galerkin's method with

$$\mathbf{r}(t, s) = \sum_{k=1}^N \mathbf{q}_k(t) \psi_k(s) \quad (9.96)$$

The shape functions  $\psi_k(s)$  were selected as the hat functions shown in Figure 9.5.

## 9.5 Euler Bernoulli beam

### 9.5.1 Model

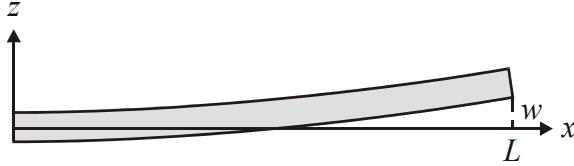


Figure 9.7: Euler Bernoulli beam.

An *Euler Bernoulli beam* is a mathematical model of lateral elastic deformations in a slender beam. The length coordinate along the beam is denoted  $x$ , and the elastic deformation in the  $z$  direction is denoted  $w(x, t)$  as shown in Figure 9.7. The model is derived from the equations of motion for a volume element of length  $dx$ . The bending moment is denoted  $M(x, t)$ , and the shear force is denoted  $V(x, t)$  (Figure 9.8). The momentum balance for the Euler Bernoulli beam is given by the two equations

$$-(V + dV) + f dx + V = \rho dx \frac{\partial^2 w}{\partial t^2} \quad (9.97)$$

$$(M + dM) - M - (V + dV) dx = 0 \quad (9.98)$$

where  $f$  is the external force per unit length of the beam. We see that the moment of inertia is set to zero in the moment equation. This is due to the assumption that the

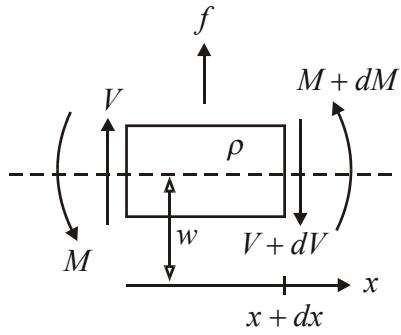


Figure 9.8: Differential element of beam.

beam is slender. A more elaborate model for thick beams is the *Timoshenko beam* which includes a nonzero moment of inertia. Division of the equations by  $dx$  leads to

$$-\frac{\partial V}{\partial x} + f = \rho \frac{\partial^2 w}{\partial t^2} \quad (9.99)$$

$$\frac{\partial M}{\partial x} = V \quad (9.100)$$

Combining these two equations we get

$$-\frac{\partial^2 M}{\partial x^2} + f = \rho \frac{\partial^2 w}{\partial t^2} \quad (9.101)$$

For the Euler Bernoulli beam the bending moment is given by the constitutive equation

$$M(x, t) = EI(x) \frac{\partial^2 w}{\partial x^2} \quad (9.102)$$

where  $E$  is Young's modulus, and  $I(x)$  is the moment of inertia about the  $y$  axis. It is noted that (9.102) implies that the shear force is

$$V = EI(x) \frac{\partial^3 w}{\partial x^3} \quad (9.103)$$

Combination of (9.101) and (9.102) gives the partial differential equation

$$\rho(x) \frac{\partial^2 w(x, t)}{\partial t^2} + \frac{\partial^2}{\partial x^2} \left[ EI(x) \frac{\partial^2 w(x, t)}{\partial x^2} \right] = f(x, t) \quad (9.104)$$

For the case where the moment of inertia  $I(x)$  is a constant and the external force  $f(x, t)$  is zero the following result is obtained:

A homogeneous Euler Bernoulli beam is described by the partial differential equation

$$\frac{\partial^2 w}{\partial t^2} + c^2 \frac{\partial^4 w}{\partial x^4} = 0 \quad (9.105)$$

where

$$c^2 = \frac{EI}{\rho} \quad (9.106)$$

### 9.5.2 Boundary conditions

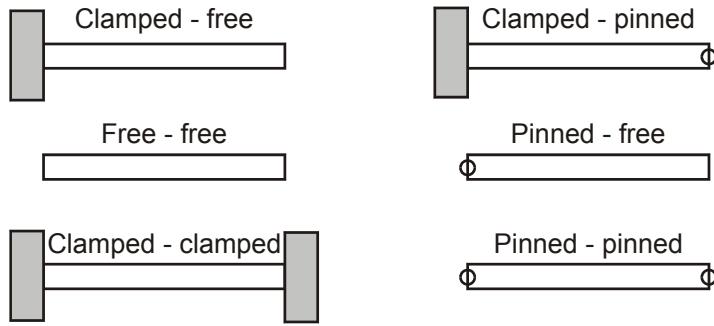


Figure 9.9: Boundary conditions for an Euler Bernoulli beam.

Typical boundary conditions for the Euler Bernoulli beam as shown in Figure 9.9 are:

1. A *clamped end* is defined to have zero elastic deformation and zero elastic angle. This means that

$$w = 0 \text{ and } \frac{\partial w}{\partial x} = 0 \quad (9.107)$$

2. A *free end* has zero bending moment and zero shear force. From (9.102) and (9.103) this is seen to imply that

$$\frac{\partial^2 w}{\partial x^2} = 0 \text{ and } \frac{\partial^3 w}{\partial x^3} = 0 \quad (9.108)$$

3. A *pinned end* has zero elastic deformation and zero bending moment, so that

$$w = 0 \text{ and } \frac{\partial^2 w}{\partial x^2} = 0 \quad (9.109)$$

4. An end with a point mass  $m$  will have zero bending moment, and a shear force

$$V = m \frac{\partial^2 w}{\partial t^2} \quad (9.110)$$

This gives the boundary conditions

$$\frac{\partial^2 w}{\partial x^2} = 0 \text{ and } EI \frac{\partial^3 w}{\partial x^3} = m \frac{\partial^2 w}{\partial t^2} \quad (9.111)$$

5. An end point clamped to a rigid body with mass  $m$  and moment of inertia  $J$  will have a shear force and bending moment given by

$$V = m \frac{\partial^2 w}{\partial t^2}, \quad M = -J \frac{\partial^2}{\partial t^2} \frac{\partial w}{\partial x} \quad (9.112)$$

From (9.102) and (9.103) this is seen to imply that

$$EI \frac{\partial^2 w}{\partial x^2} = -J \frac{\partial^2}{\partial t^2} \frac{\partial w}{\partial x} \text{ and } EI \frac{\partial^3 w}{\partial x^3} = m \frac{\partial^2 w}{\partial t^2} \quad (9.113)$$

### 9.5.3 Energy

For an Euler Bernoulli beam the kinetic energy is

$$T = \frac{1}{2} \int_0^L \rho(x) \dot{w}^2 dx \quad (9.114)$$

while the potential energy is

$$U = \frac{1}{2} \int_0^L EI(x) (w'')^2 dx \quad (9.115)$$

The total energy is then

$$W = \frac{1}{2} \int_0^L \rho(x) \dot{w}^2 dx + \frac{1}{2} \int_0^L EI (w'')^2 dx \quad (9.116)$$

and the time derivative of the energy along the solutions of the system is found to be

$$\begin{aligned} \dot{W} &= \int_0^L (\dot{w} \rho \ddot{w} + w'' E I \dot{w}'') dx \\ &= EI \int_0^L (-\dot{w} w''' + \dot{w} f + w'' \dot{w}'') dx \\ &= -\dot{w} EI w'''|_0^L + EI \int_0^L (\dot{w} f + \dot{w}' w''' + w'' \dot{w}'') dx \\ &= -\dot{w} EI w'''|_0^L + \dot{w}' EI w''|_0^L + EI \int_0^L (\dot{w} f - \dot{w}'' w'' + w'' \dot{w}'') dx \\ &= -\dot{w} V|_0^L + \dot{w}' M|_0^L + EI \int_0^L \dot{w} f dx \end{aligned} \quad (9.117)$$

The total energy of the Euler Bernoulli beam is

$$W = \frac{1}{2} \int_0^L \rho(x) \dot{w}^2 dx + \frac{1}{2} \int_0^L EI (w'')^2 dx \quad (9.118)$$

The time derivative along the solutions of the system is

$$\dot{W} = -\dot{w} V|_0^L + \dot{w}' M|_0^L + EI \int_0^L \dot{w} f dx \quad (9.119)$$

### 9.5.4 Orthogonal shape functions

For a homogeneous Euler Bernoulli beam with zero external force the method of separation of variables leads to a very useful description based on orthogonal mode shapes. This description will be developed in the following, and it will serve as a starting point to explain transfer function models and finite element models that will be presented at a later stage.

In the method of separation of variables the expression

$$w(x, t) = \phi(x)q(t) \quad (9.120)$$

is inserted into the partial differential equation (9.105). This gives

$$\phi(x)\ddot{q}(t) + c^2\phi''''(x)q(t) = 0 \quad (9.121)$$

Following the standard procedure this equation is reformulated as

$$\frac{c^2\phi''''(x)}{\phi(x)} = -\frac{\ddot{q}(t)}{q(t)} = C \quad (9.122)$$

where it is seen that  $C$  must be a constant as it is a function of  $t$  alone and at the same time a function of  $x$  alone. The only nontrivial solutions for  $\phi$  are found for  $C > 0$ . Therefore the constant  $\omega$  is introduced so that  $C = \omega^2$ . This gives the two ordinary differential equations

$$\ddot{q}(t) + \omega^2 q(t) = 0 \quad (9.123)$$

$$\phi''''(x) - \beta^4 \phi(x) = 0 \quad (9.124)$$

where the constant

$$\beta^4 = \frac{\omega^2}{c^2} \quad (9.125)$$

has been introduced.

The first equation (9.123) is recognized as a harmonic oscillator. The solution of the second equation (9.124) is given by

$$\phi(x) = C_1 \cos \beta x + C_2 \sin \beta x + C_3 \cosh \beta x + C_4 \sinh \beta x \quad (9.126)$$

which has derivatives

$$\phi'(x) = \beta(-C_1 \sin \beta x + C_2 \cos \beta x + C_3 \sinh \beta x + C_4 \cosh \beta x) \quad (9.127)$$

$$\phi''(x) = \beta^2(-C_1 \cos \beta x - C_2 \sin \beta x + C_3 \cosh \beta x + C_4 \sinh \beta x) \quad (9.128)$$

$$\phi'''(x) = \beta^3(C_1 \sin \beta x - C_2 \cos \beta x + C_3 \sinh \beta x + C_4 \cosh \beta x) \quad (9.129)$$

Depending on which boundary conditions that apply for the beam, there will be conditions on  $\phi$  and its derivatives at  $x = 0$  and  $x = L$ . The boundary conditions can then be used to determine the constants  $C_i$  through the equations

$$\begin{pmatrix} \phi(0) \\ \frac{1}{\beta}\phi'(0) \\ \frac{1}{\beta^2}\phi''(0) \\ \frac{1}{\beta^3}\phi'''(0) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \\ C_3 \\ C_4 \end{pmatrix} \quad (9.130)$$

and

$$\begin{pmatrix} \phi(L) \\ \frac{1}{\beta}\phi'(L) \\ \frac{1}{\beta^2}\phi''(L) \\ \frac{1}{\beta^3}\phi'''(L) \end{pmatrix} = \begin{pmatrix} \cos \beta L & \sin \beta L & \cosh \beta L & \sinh \beta L \\ -\sin \beta L & \sin \beta L & \cosh \beta L & \sinh \beta L \\ -\cos \beta L & -\sin \beta L & \cosh \beta L & \sinh \beta L \\ \sin \beta L & -\cos \beta L & \sinh \beta L & \cosh \beta L \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \\ C_3 \\ C_4 \end{pmatrix} \quad (9.131)$$

For a given set of boundary conditions there will be a set of *shape functions*  $\phi_i(x)$  with associated constants  $\beta_i$  and natural frequencies  $\omega_i = c\beta_i^2$  so that the solution is given by

$$w(x, t) = \sum_{i=1}^{\infty} \phi_i(x)q_i(t) \quad (9.132)$$

Type of end	Boundary Condition 1	Boundary Condition 2
Clamped	$\phi = 0$	$\phi' = 0$
Free	$\phi'' = 0$	$\phi''' = 0$
Pinned	$\phi = 0$	$\phi'' = 0$
Mass $m$	$\phi'' = 0$	$\phi''' = -\frac{m}{\rho}\beta^4\phi$
Mass $m$ and inertia $J$	$\phi'' = -\frac{J}{\rho}\beta^4\phi'$	$\phi''' = -\frac{m}{\rho}\beta^4\phi$

Table 9.1: Boundary conditions for Euler Bernoulli beam.

Here  $\phi_i(x)$  and  $q_i(t)$  satisfy

$$\ddot{q}_i(t) + \omega_i^2 q_i(t) = 0 \quad (9.133)$$

$$\phi_i'''(x) - \beta_i^4 \phi_i(x) = 0 \quad (9.134)$$

The boundary conditions on the deflection  $w(x, t)$  imply the boundary conditions given in Table 9.1 on the shape functions. Given the boundary conditions, the constants  $C_i$  are found by formulating equations for the boundary conditions according to

$$\mathbf{B}\mathbf{c} = \mathbf{0} \quad (9.135)$$

where  $\mathbf{c} = (C_1, C_2, C_3, C_4)^T$ . Then, to have nontrivial solutions for the constants in the vector  $\mathbf{c}$ , it is necessary that

$$\det \mathbf{B} = 0 \quad (9.136)$$

### 9.5.5 Clamped-free Euler Bernoulli beam

We will derive the solution for an Euler Bernoulli beam which is clamped at the end at  $x = 0$ , and free at  $x = L$ . The boundary conditions are

$$\phi(0) = 0, \quad \phi'(0) = 0, \quad (9.137)$$

$$\phi''(L) = 0, \quad \phi'''(L) = 0 \quad (9.138)$$

This can be written

$$\mathbf{0} = \begin{pmatrix} \phi(0) \\ \frac{1}{\beta}\phi'(0) \\ \frac{1}{\beta^2}\phi''(L) \\ \frac{1}{\beta^3}\phi'''(L) \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ -\cos\beta L & -\sin\beta L & \cosh\beta L & \sinh\beta L \\ \sin\beta L & -\cos\beta L & \sinh\beta L & \cosh\beta L \end{pmatrix}}_{\mathbf{B}_{cf}} \underbrace{\begin{pmatrix} C_1 \\ C_2 \\ C_3 \\ C_4 \end{pmatrix}}_{\mathbf{c}} \quad (9.139)$$

Nontrivial solutions are found when  $\det \mathbf{B}_{cf} = 0$ , which occurs when

$$1 + \cos\beta L \cosh\beta L = 0 \quad (9.140)$$

This transcendental equation holds for infinitely many discrete values of  $\beta$ , and must be solved numerically. The solutions are denoted  $\beta_i$ , and for each  $\beta_i$  there corresponds resonance frequency  $\omega_i = \beta_i^2 c$  in agreement with (9.125), and a set of constants  $C_{1i}$ ,

Type of beam	Characteristic equation
Clamped-free	$\cos \beta L \cosh \beta L + 1 = 0$
Free-free and clamped-clamped	$\cos \beta L \cosh \beta L - 1 = 0$
Clamped-pinned and pinned-free	$\cos \beta L \sinh \beta L - \sin \beta L \cosh \beta L = 0$
Pinned-pinned	$\sin \beta L = 0$

Table 9.2: Characteristic equations for Euler Bernoulli beam.

$C_{2i}$ ,  $C_{3i}$ ,  $C_{4i}$ , and one shape function  $\phi_i(x)$ . We note from the first and second row of (9.139) that the constants are related by  $C_{3i} = -C_{1i}$  and  $C_{4i} = -C_{2i}$ , and that the third and fourth row of (9.139) implies that  $C_{2i}$  can be expressed by  $C_{1i}$  according to

$$C_{2i} = \alpha_i C_{1i} \text{ where } \alpha_i = \frac{\cos \beta_i L + \cosh \beta_i L}{\sin \beta_i L + \sinh \beta_i L} \quad (9.141)$$

This shows that the shape functions are given by

$$\phi_i(x) = C_{1i} [(\cos \beta_i x - \cosh \beta_i x) + \alpha_i (\sin \beta_i x - \sinh \beta_i x)] \quad (9.142)$$

Usually the constant  $C_{1i}$  is normalized so that

$$\int_0^\ell \rho [\phi_i(x)]^2 dx = 1 \quad (9.143)$$

The solution is then

$$w(x, t) = \sum_{i=1}^{\infty} q_i(t) \phi_i(x) \quad (9.144)$$

where the generalized coordinate  $q_i$  satisfy the differential equation

$$\ddot{q}_i(t) + \omega_i^2 q_i(t) = 0 \quad (9.145)$$

Numerical values for the first modes are tabulated in textbooks like (Rao 1990), and for the first three modes we have

$$\beta_1 \ell = 1.875104, \quad \alpha_1 = 0.7341 \quad (9.146)$$

$$\beta_2 \ell = 4.694091, \quad \alpha_2 = 1.0185 \quad (9.147)$$

$$\beta_3 \ell = 7.854757, \quad \alpha_3 = 0.9992 \quad (9.148)$$

In this example

$$\frac{\omega_2}{\omega_1} = \frac{\beta_2^2}{\beta_1^2} = 6.25$$

Numerical values for  $\beta$  are tabulated in many textbooks on vibrations for simple cases like pinned-pinned, free-free, fixed-fixed, fixed-free, fixed-pinned, and pinned-free.

The characteristic equation for different beams are given in Table 9.2.

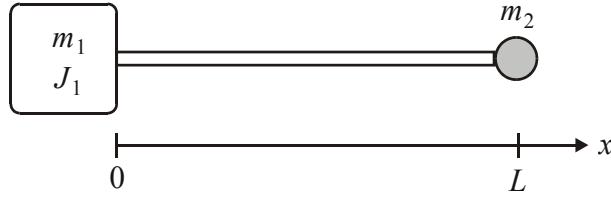


Figure 9.10: Satellite antenna.

### 9.5.6 Beam fixed to an inertia and a mass

The antenna boom on a small satellite is modelled as an Euler Bernoulli beam. The beam is modelled to be clamped at the satellite end at  $x = 0$  to an inertial load with mass  $m_1$  and moment of inertia  $J_1$ , and fixed to a mass  $m_2$  at the end of the boom at  $x = L$  (Figure 9.10). The boundary conditions are then

$$\phi''(0) = -\frac{J_1}{\rho} \beta^4 \phi'(0), \quad \phi'''(0) = -\frac{m_1}{\rho} \beta^4 \phi(0) \quad (9.149)$$

$$\phi''(L) = 0, \quad \phi'''(L) = -\frac{m_2}{\rho} \beta^4 \phi(L) \quad (9.150)$$

which can be written

$$\begin{pmatrix} \frac{1}{\beta^2} \phi''(0) + \frac{J_1}{\rho} \beta^2 \phi'(0) \\ \frac{1}{\beta^3} \phi'''(0) + \frac{m_1}{\rho} \beta \phi(0) \\ \frac{1}{\beta^2} \phi''(L) \\ \frac{1}{\beta^3} \phi'''(L) + \frac{m_2}{\rho} \beta \phi(L) \end{pmatrix} = \mathbf{B}_a \begin{pmatrix} C_1 \\ C_2 \\ C_3 \\ C_4 \end{pmatrix} = \mathbf{0} \quad (9.151)$$

Nontrivial solutions are found for

$$\det \mathbf{B}_a = 0 \quad (9.152)$$

In the study of satellite dynamics the concept of constrained modes are often used. These are the modes that occur when the satellite is assumed to have infinite inertia and mass. Therefore the constrained modes for this antenna is found when  $m_1$  and  $J$  are assumed to approach infinity. The condition for nontrivial solutions in this case is

$$1 + \cos \beta \ell \cosh \beta \ell + \frac{m_2}{\rho} \ell (\cos \beta \ell - \sin \beta \ell \sinh \beta \ell) = 0 \quad (9.153)$$

In contrast to this the unconstrained modes are the modes are found for finite  $m_1$  and  $J$ . For a small satellite we assume that  $m_1 = 78$  kg,  $J = 4$  kg m<sup>2</sup>,  $\ell = 6$  m,  $m_2 = 4$  kg,  $EI = 28.69$  N·m<sup>2</sup> (10,000 lbf·inch<sup>2</sup>),  $\rho = 1$  kg/6 m = 0.17 kg/m. Then the numerical values were found for the constrained case and the unconstrained case using Maple. The results are shown in Table 9.3. It was found that the spacecraft with gravity boom will have resonances with periods 41.0 s, 1.14 s, 0.34 s, ... The computations based on the assumption of constrained modes predicts periods 20.2 s, 1.10 s, 0.34 s. It is seen that the lowest natural frequency computed for the constrained modes is a factor of 2 greater than the natural frequency associated with the unconstrained mode for this example.

Mode #	Constrained modes			Unconstrained modes		
	$\beta$	$\frac{\omega}{2\pi}$ (Hz)	$\frac{1}{f}$ (s)	$\beta$	$\frac{\omega}{2\pi}$ (Hz)	$\frac{1}{f}$ (s)
0	-	-	-	0	0	$\infty$
1	0.1529	0.0488	20.2	0.1082	0.0244	41.0
2	0.6592	0.9074	1.10	0.6487	0.8787	1.14
3	1.1810	2.9125	0.34	1.1782	2.8987	0.34
4	1.7037	6.0610	0.17	1.7025	6.0525	0.17
5	2.2268	10.354	0.10	2.2262	10.3487	0.10

(9.154)

Table 9.3: Natural frequencies for satellite antenna using constrained modes and unconstrained modes.

### 9.5.7 Orthogonality of the eigenfunctions

For specified boundary conditions there will be a set of solutions  $\phi_i(x)$  of (9.124), where each solution  $\phi_i(x)$  corresponds to a value  $\beta_i$  for  $\beta$  so that

$$\phi_i'''(x) = \beta_i^4 \phi_i(x) \quad (9.155)$$

The solutions  $\phi_i(x)$  are called the eigenfunctions (9.124), and the associated values  $\lambda_i = \beta_i^4$  are the eigenvalues of the system. It is assumed in the following that the eigenvalues are distinct. Note that for each eigenfunction  $\phi_i(x)$  there is one natural frequency

$$\omega_i = c\beta_i^2 \quad (9.156)$$

of the harmonic oscillator (9.123).

Consider the eigenfunction  $\phi_i(x)$  with eigenvalue  $\beta_i^4$  and the eigenfunction  $\phi_j(x)$  with eigenvalue  $\beta_j^4$ . Then

$$\int_0^\ell \phi_i(x) \phi_j'''(x) dx = \beta_j^4 \int_0^\ell \phi_i(x) \phi_j(x) dx \quad (9.157)$$

The integral on the left side can also be evaluated by partial integration twice to be

$$\int_0^\ell \phi_i(x) \phi_j'''(x) dx = \phi_i(x) \phi_j'''(x) \Big|_0^\ell - \phi_i'(x) \phi_j''(x) \Big|_0^\ell + \int_0^\ell \phi_i''(x) \phi_j''(x) dx \quad (9.158)$$

Combination of (9.157) and (9.158) gives

$$\beta_j^4 \int_0^\ell \phi_i(x) \phi_j(x) dx = \phi_i(x) \phi_j'''(x) \Big|_0^\ell - \phi_i'(x) \phi_j''(x) \Big|_0^\ell + \int_0^\ell \phi_i''(x) \phi_j''(x) dx \quad (9.159)$$

In the same way, by interchanging  $i$  and  $j$  in the expression the result

$$\beta_i^4 \int_0^\ell \phi_i(x) \phi_j(x) dx = \phi_j(x) \phi_i'''(x) \Big|_0^\ell - \phi_j'(x) \phi_i''(x) \Big|_0^\ell + \int_0^\ell \phi_j''(x) \phi_i''(x) dx \quad (9.160)$$

appears. From (9.159) and (9.160) it is seen that if

$$\phi_i(x) \phi_j'''(x) \Big|_0^\ell - \phi_i'(x) \phi_j''(x) \Big|_0^\ell = \phi_j(x) \phi_i'''(x) \Big|_0^\ell - \phi_j'(x) \phi_i''(x) \Big|_0^\ell \quad (9.161)$$

the eigenfunctions will satisfy

$$(\beta_j^4 - \beta_i^4) \int_0^\ell \phi_i(x) \phi_j(x) dx = 0 \quad (9.162)$$

Due to the assumption  $\beta_i \neq \beta_j$ , this implies that

$$\int_0^\ell \phi_i(x) \phi_j(x) dx = 0, \quad i \neq j \quad (9.163)$$

Moreover, from (9.159) and (9.160) it follows that

$$\int_0^\ell \phi_i''(x) \phi_j''(x) dx = 0, \quad i \neq j \quad (9.164)$$

Usually the eigenfunctions are normalized so that the following result is valid:

The eigenfunctions of (9.124) are orthogonal shape functions in the sense that they satisfy

$$\int_0^\ell \rho \phi_i(x) \phi_j(x) dx = \delta_{ij} \quad (9.165)$$

In addition they satisfy

$$\int_0^\ell \rho \phi_i''(x) \phi_j''(x) dx = \beta_i^4 \delta_{ij} \quad (9.166)$$

where  $\delta_{ij} = 1$  when  $i = j$  and  $\delta_{ij} = 0$  when  $i \neq j$ .

### 9.5.8 Galerkin's method for orthogonal mode shapes

We will now introduce a control force  $u$  in our model of a Euler Bernoulli beam. The control force  $u$  is assumed to be perpendicular to the beam at position  $x_u$ . This can be modelled with the Dirac delta  $\delta(x)$  in the partial differential equation:

$$\rho c^2 \frac{\partial^4 w}{\partial x^4}(x, t) + \rho \frac{\partial^2 w}{\partial t^2}(x, t) = \delta(x - x_u) u \quad (9.167)$$

Due to the forcing term  $\delta(x - x_u) u$  the method of separation of variables cannot be used directly.

The solution  $w(x, t)$  is assumed to be a linear combination of the eigenfunctions  $\phi_j(x)$ ,  $j \in \{1, 2, \dots\}$ , and we may write

$$w(x, t) = \sum_{j=1}^{\infty} q_j(t) \phi_j(x) \quad (9.168)$$

The partial differential equations can then be written

$$\sum_{j=1}^{\infty} [\rho c^2 q_j(t) \frac{\partial^4 \phi_j(x)}{\partial x^4} + \rho \frac{\partial^2 q_j(t)}{\partial t^2} \phi_j(x)] = \delta(x - x_u) u(t) \quad (9.169)$$

The partial differential equation is reformulated by insertion of (9.134) and  $\omega_j^2 = c^2 \beta_j^4$ , which gives

$$\sum_{j=1}^{\infty} \rho \phi_j(x) [\omega_j^2 q_j(t) + \ddot{q}_j(t)] = \delta(x - x_u) u(t) \quad (9.170)$$

At this point Galerkin's method is used. This involves the multiplication of equation (9.170) with  $\phi_i(x)$  and the integration of the result over the interval  $x \in [0, \ell]$ . This gives

$$\int_0^\ell \phi_i(x) \sum_{j=1}^{\infty} \rho \phi_j(x) dx [\omega_j^2 q_j(t) + \ddot{q}_j(t)] = \int_0^\ell \phi_i(x) \delta(x - x_u) dx u(t) \quad (9.171)$$

For any function  $f(x)$  the Dirac delta gives  $\int_0^\ell f(x) \delta(x - x_u) dx = f(x_u)$ . This together with the orthogonality of the eigenfunctions, see (9.165), lead to

$$\omega_i^2 q_i(t) + \ddot{q}_i(t) = \phi_i(x_u) u(t) \quad (9.172)$$

which has the Laplace transform

$$q_i(s) = \frac{\phi_i(x_u)}{\omega_i^2 + s^2} u(s) \quad (9.173)$$

We assume that the measurement  $y(t)$  is the velocity  $\dot{w}(x_y, t)$  of the elastic deflection at position  $x_y$ , that is,

$$y(t) = \dot{w}(x_y, t) \quad (9.174)$$

Then the measurement can be written

$$y(t) = \sum_{i=1}^{\infty} \dot{q}_i(t) \phi_i(x_y) \quad (9.175)$$

and the transfer function from  $u$  to  $y$  is seen to be

$$\frac{y(s)}{u(s)} = \sum_{i=1}^{\infty} \frac{s \phi_i(x_y) \phi_i(x_u)}{\omega_i^2 + s^2} \quad (9.176)$$

The following observations are important:

1. If input and output are collocated, which is the case whenever  $x_u = x_y = x_0$ , then the transfer function from  $u$  to  $y$  is passive because

$$\frac{Y(s)}{U(s)} = \sum_{i=1}^{\infty} \frac{s \phi_i^2(x_0)}{\omega_i^2 + s^2} \quad (9.177)$$

which is a parallel interconnection of passive systems. Alternatively passivity may be established from energy considerations. This result agrees with an energy argument where  $V$  is the sum of kinetic and potential energy. Then

$$\dot{V}(t) = y(t) u(t) - d(t) \quad (9.178)$$

where  $\int_0^T d(t) dt \geq 0$  is the dissipated energy in the system. It follows that the system with input  $u$  and output  $y$  is passive.

2. If measurement and control are not collocated, then  $x_u \neq x_y$ , and it may be that  $\phi_i(x_y)$  and  $\phi_i(x_u)$  have opposite signs for certain  $i$ . In this case the transfer function  $y(s)/u(s)$  will not be positive real. This may cause difficulties in designing a controller to damp out vibrations.

3. If  $\phi_i(x_u) = 0$ , the control  $u$  will have no influence on mode  $i$ . In the state-space terminology this means that mode  $i$  is not controllable with the control  $u$ .
4. If  $\phi_i(x_y) = 0$ , then mode  $i$  will not be noticeable in the measurement  $y$ . This means that mode  $i$  is not observable from the measurement  $y$ .

**Example 149** Consider a homogeneous beam of aluminium with a rectangular cross section, length  $\ell = 2$  m, width  $b = 0.05$  m, height  $h = 0.01$  m, density  $\rho = b \cdot h \cdot 2700$  kg/m<sup>3</sup> = 1.35 kg/m, Young's modulus  $E = 70 \cdot 10^9$  N/m<sup>2</sup> and moment of inertia  $I = bh^3/12 = 4.167 \cdot 10^{-9}$  m<sup>4</sup>. The beam is clamped at  $x = 0$  and free at  $x = \ell$ .

The shape functions can be found to be

$$\phi_1(x) = -0.6086 \cdot \{\cos(\beta_1 x) - \cosh(\beta_1 x) - 0.7341 \cdot [\sin(\beta_1 x) - \sinh(\beta_1 x)]\} \quad (9.179)$$

and

$$\phi_2(x) = -0.6086 \cdot \{\cos(\beta_2 x) - \cosh(\beta_2 x) - 1.0185 \cdot [\sin(\beta_2 x) - \sinh(\beta_2 x)]\} \quad (9.180)$$

First collocation is tried with  $x_u = x_y = 2$  m. Then

$$\phi_1(x_u) = \phi_1(x_y) = 1.22, \quad \phi_2(x_u) = \phi_2(x_y) = -1.22$$

and the transfer function

$$\frac{y}{u}(s) = \frac{1.5s}{12.8^2 + s^2} + \frac{1.5s}{80.1^2 + s^2} = 3.0 \frac{s(57.4^2 + s^2)}{(12.8^2 + s^2)(80.1^2 + s^2)} \quad (9.181)$$

results, which is passive. Note that the complex conjugated zeros at  $s = \pm j57.4$  is located between the poles in  $s = \pm j12.8$  and  $s = \pm j80.1$ . A simple P controller

$$u = -K_p y \quad (9.182)$$

will give stability, with a power dissipation of  $u(t)y(t) = -K_p y(t)^2$ . The gain  $K_p$  is only limited by noise, quantization and discretization effects.

Next noncollocation is tried with  $x_u = 0.5$  m and  $x_y = 2$  m. Then

$$\phi_1(x_u) = 0.12 \quad \phi_1(x_y) = 1.22$$

$$\phi_2(x_u) = 0.51 \quad \phi_2(x_y) = -1.22$$

and the transfer function is

$$\frac{y}{u}(s) = \frac{0.15s}{12.8^2 + s^2} - \frac{0.62s}{80.1^2 + s^2} = -0.47 \frac{s(47.6^2 - s^2)}{(12.8^2 + s^2)(80.1^2 + s^2)} \quad (9.183)$$

This transfer function has a zero in the right half plane at  $s = 47.6$ . This limits the bandwidth of the system. Alternatively, we see that the transfer function is the sum of two transfer functions that are not very different, except that they have opposite signs. Thus if a P controller is used in a negative feedback, this will give stabilization and power dissipation for mode 1, while it will give destabilization and added power to mode 2. In fact the only possibility for stabilization is that the mode with positive feedback has gain less than unity which ensures stability according to Bode-Nyquist stability theory.

## 9.6 Finite element model of Euler Bernoulli beam

### 9.6.1 Introduction

An alternative technique for analyzing the Euler Bernoulli beam is to use the finite-element method. The finite-element method can be seen as a model formulation based on the Galerkin method, where special set of shape functions are used. The characteristic feature of the finite-element method is that the shape functions are locally defined in the sense that they are nonzero only in short intervals of the beam. This is in contrast to the orthogonal shape functions, which are global function on the beam. An alternative way of seeing the finite-element method is that the beam is divided into beam elements. The equations of motion are then derived for the beam element using a cubic shape function, and then the beam model is obtained by connecting the beam element models using multiport techniques. The presentation that follows will focus on the formulation using beam elements, but the Galerkin point of view will also be presented.

### 9.6.2 Beam element

In a finite-element model of an Euler Bernoulli beam the basic building block of the model is a beam element of length  $h$ . The element is defined for the interval  $0 \leq x \leq h$ . At  $x = 0$  the shear force is  $V_1$  and the bending moment is  $M_1$ , the elastic displacement is  $w_1$ , and the elastic angle is  $w'_1$ . This can be seen as one port with effort  $V_1$  and flow  $\dot{w}_1$ , and one port with effort  $M_1$  and flow  $\dot{w}'_1$ . At  $x = h$  the shear force is  $V_2$ , the bending moment is  $M_2$ , the elastic deflection is  $w_2$ , and the elastic angle is this is  $w'_2$ . This is described as a port with effort  $V_2$  and flow  $\dot{w}_2$ , and one port with effort  $M_2$  and flow  $\dot{w}'_2$ . The usual finite-element model of the Euler Bernoulli beam is based on the displacement formulation where the inputs to the model are the forces and torques, and the outputs are the displacements and the displacement angles. In the multiport terminology this is an admittance model where the efforts are input and the flows are outputs.

The displacement in the element is modeled as the cubic expression

$$w(x, t) = c_0(t) + c_1(t)x + c_2(t)x^2 + c_3(t)x^3 \quad (9.184)$$

The motivation for using this expression is that in the stationary case the displacement satisfies  $w''' = 0$ , which has solution (9.184). The generalized coordinates  $a_i(t)$  of the beam element are defined as

$$a_1(t) = w_1(t) \quad a_2(t) = w'_1(t) \quad (9.185)$$

$$a_3(t) = w_2(t), \quad a_4(t) = w'_2(t) \quad (9.186)$$

Combination of (9.184), (9.185) and (9.186) leads to

$$w(x, t) = \sum_{i=1}^4 \alpha_i(x)a_i(t) \quad (9.187)$$

where the shape functions  $\alpha_i(x)$  are given by

$$\alpha_1(x) = 1 - 3\left(\frac{x}{h}\right)^2 + 3\left(\frac{x}{h}\right)^3 \quad (9.188)$$

$$\alpha_2(x) = h\left[\left(\frac{x}{h}\right) - 2\left(\frac{x}{h}\right)^2 + \left(\frac{x}{h}\right)^3\right] \quad (9.189)$$

$$\alpha_3(x) = 3\left(\frac{x}{h}\right)^2 - 2\left(\frac{x}{h}\right)^3 \quad (9.190)$$

$$\alpha_4(x) = h\left[-\left(\frac{x}{h}\right)^2 + \left(\frac{x}{h}\right)^3\right] \quad (9.191)$$

These cubic shape functions are called the Hermitian shape functions.

Galerkin's method for the beam element leads to

$$\mathbf{M}_e \ddot{\mathbf{a}} + \mathbf{K}_e \mathbf{a} = \mathbf{f} \quad (9.192)$$

where the mass matrix of the element is given by

$$\mathbf{M}_e = \int_0^h \rho \boldsymbol{\alpha} \boldsymbol{\alpha}^T dx = \frac{\rho h}{420} \begin{pmatrix} 156 & 22h & 54 & -13h \\ 22h & 4h^2 & 13h & -3h^2 \\ 54 & 13h & 156 & -22h \\ -13h & -3h^2 & -22h & 4h^2 \end{pmatrix} \quad (9.193)$$

and the stiffness matrix of the element is given by

$$\mathbf{K}_e = \int_0^h \rho \boldsymbol{\alpha}'' (\boldsymbol{\alpha}'')^T dx = \frac{2c^2\rho}{h^3} \begin{pmatrix} 6 & 3h & -6 & 3h \\ 3h & 2h^2 & -3h & h^2 \\ -6 & -3h & 6 & -3h \\ 3h & h^2 & -3h & 2h^2 \end{pmatrix} \quad (9.194)$$

and  $\mathbf{f} = (f_1, f_2, f_3, f_4)^T$  where

$$f_i = \int_0^h \alpha_i(x) f(x) \quad (9.195)$$

**Example 150** To simplify the model the mass is sometimes lumped to have a diagonal mass matrix (Rao 1990). For translational degrees of freedom this is simply done by lumping all mass at the nodes, while for rotational inertia the inertia can be computed by having uniform mass distribution for half a beam on each side of the node. This gives in this example

$$\mathbf{M}_e = \frac{\rho h}{2} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{h^2}{12} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \frac{h^2}{12} \end{pmatrix} \quad (9.196)$$

which is the lumped mass matrix.

### 9.6.3 Assembling a structure

To establish the model for a beam of length  $L$  where  $L = Nh$  it is necessary to connect  $N$  beam elements. Elements  $k$  and  $k+1$  can be connected by requiring that the end-point variables satisfy  $a_{k,3} = a_{k+1,1}$  and  $a_{k,4} = a_{k+1,2}$ . Then, there must be forces and torques

of constraints to hold the two element together, and the equations of motion for elements  $k$  and  $k + 1$  are given by

$$\mathbf{M}_e \frac{d^2}{dt^2} \begin{pmatrix} a_{k,1} \\ a_{k,2} \\ a_{k,3} \\ a_{k,4} \end{pmatrix} + \mathbf{K}_e \begin{pmatrix} a_{k,1} \\ a_{k,2} \\ a_{k,3} \\ a_{k,4} \end{pmatrix} = \begin{pmatrix} f_{k,1} \\ f_{k,2} \\ f_{k,3} + f_3^{(c)} \\ f_{k,4} + f_4^{(c)} \end{pmatrix} \quad (9.197)$$

$$\mathbf{M}_e \frac{d^2}{dt^2} \begin{pmatrix} a_{k+1,1} \\ a_{k+1,2} \\ a_{k+1,3} \\ a_{k+1,4} \end{pmatrix} + \mathbf{K}_e \begin{pmatrix} a_{k+1,1} \\ a_{k+1,2} \\ a_{k+1,3} \\ a_{k+1,4} \end{pmatrix} = \begin{pmatrix} f_{k+1,1} - f_3^{(c)} \\ f_{k+1,2} - f_4^{(c)} \\ f_{k+1,3} \\ f_{k+1,4} \end{pmatrix} \quad (9.198)$$

These forces and torques of constraint are eliminated by adding rows 3 and 4 of element  $k$  to rows 1 and 2 of element  $k + 1$ . This gives the model

$$\mathbf{M}\ddot{\mathbf{q}} + \mathbf{K}\mathbf{q} = \mathbf{u} \quad (9.199)$$

where  $\mathbf{q} = (a_{k,1}, a_{k,2}, a_{k+1,1}, a_{k+1,2}, a_{k+1,3}, a_{k+1,4})$ . The mass matrix is obtained from

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_e & \mathbf{0}_{4,2} \\ \mathbf{0}_{2,4} & \mathbf{0}_{2,2} \end{pmatrix} + \begin{pmatrix} \mathbf{0}_{2,2} & \mathbf{0}_{2,4} \\ \mathbf{0}_{4,2} & \mathbf{M}_e \end{pmatrix} \quad (9.200)$$

In the same way the stiffness matrix is obtained from

$$\mathbf{K} = \begin{pmatrix} \mathbf{K}_e & \mathbf{0}_{4,2} \\ \mathbf{0}_{2,4} & \mathbf{0}_{2,2} \end{pmatrix} + \begin{pmatrix} \mathbf{0}_{2,2} & \mathbf{0}_{2,4} \\ \mathbf{0}_{4,2} & \mathbf{K}_e \end{pmatrix} \quad (9.201)$$

Alternatively, the model of the two elements can be written

$$\bar{\mathbf{M}} \frac{d^2}{dt^2} \bar{\mathbf{a}} + \bar{\mathbf{K}} \bar{\mathbf{a}} = \bar{\mathbf{f}} \quad (9.202)$$

$$\bar{\mathbf{a}} = (\mathbf{a}_1, \dots, \mathbf{a}_p)^T \quad (9.203)$$

$$\bar{\mathbf{M}} = \text{block diag}(\mathbf{M}_{e1}, \dots, \mathbf{M}_{ep}), \quad \bar{\mathbf{K}} = \text{block diag}(\mathbf{K}_{e1}, \dots, \mathbf{K}_{ep}) \quad (9.204)$$

where the connection of the elements is obtained by requiring

$$\bar{\mathbf{a}} = \mathbf{C}\mathbf{q}, \quad \mathbf{u} = \mathbf{C}^T \bar{\mathbf{f}} \quad (9.205)$$

where

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad (9.206)$$

when  $N = 2$ . Then the mass matrix and the stiffness matrix are found from

$$\mathbf{M} = \mathbf{C}^T \bar{\mathbf{M}} \mathbf{C}, \quad \mathbf{K} = \mathbf{C}^T \bar{\mathbf{K}} \mathbf{C} \quad (9.207)$$

to be

$$\mathbf{M} = \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} & 0 & 0 \\ m_{21} & m_{22} & m_{23} & m_{24} & 0 & 0 \\ m_{31} & m_{32} & m_{33} + m_{11} & m_{34} + m_{12} & m_{13} & m_{14} \\ m_{41} & m_{42} & m_{43} + m_{21} & m_{44} + m_{22} & m_{23} & m_{24} \\ 0 & 0 & m_{31} & m_{32} & m_{33} & m_{34} \\ 0 & 0 & m_{41} & m_{42} & m_{43} & m_{44} \end{pmatrix} \quad (9.208)$$

$$\mathbf{K} = \begin{pmatrix} k_{11} & k_{12} & k_{13} & k_{14} & 0 & 0 \\ k_{21} & k_{22} & k_{23} & k_{24} & 0 & 0 \\ k_{31} & k_{32} & k_{33} + k_{11} & k_{34} + k_{12} & k_{13} & k_{14} \\ k_{41} & k_{42} & k_{43} + k_{21} & k_{44} + k_{22} & k_{23} & k_{24} \\ 0 & 0 & k_{31} & k_{32} & k_{33} & k_{34} \\ 0 & 0 & k_{41} & k_{42} & k_{43} & k_{44} \end{pmatrix} \quad (9.209)$$

and the resulting model is

$$\mathbf{M}\ddot{\mathbf{q}} + \mathbf{K}\mathbf{q} = \mathbf{u} \quad (9.210)$$

#### 9.6.4 Finite element model and Galerkin's method

A finite-element model for an Euler Bernoulli beam can alternatively be established by applying Galerkin's method with shape functions  $\psi_i(x)$  based on the element shape functions in (9.188–9.191). For the Euler Bernoulli beam,  $N$  nodes are defined at  $x_1 < x_2 < \dots < x_N$ , and the deflection is described by

$$w(x, t) = \sum_{j=1}^N [\alpha_{j,1}(x)a_{j,1}(t) + \alpha_{j,2}(x)a_{j,2}(t)] \quad (9.211)$$

which is expressed in the form

$$w(x, t) = \sum_{j=1}^{2N} \psi_j(x)q_j(t) \quad (9.212)$$

where the generalized coordinates are  $\mathbf{q} = (a_{1,1}, a_{1,2}, \dots, a_{N,1}, a_{N,2})^T$  and the mode shape vector is  $\psi = (\psi_{1,1}, \psi_{1,2}, \dots, \psi_{N,1}, \psi_{N,2})^T$ . The shape functions  $\alpha_{j,1}(x)$  and  $\alpha_{j,2}(x)$  for the Euler Bernoulli beam are selected in agreement with (9.188–9.191) as the Hermitian shape functions

$$\alpha_{i,1}(x) = \begin{cases} 1 - 3\frac{(x-x_i)^2}{\ell_i^2} + 2\frac{(x-x_i)^3}{\ell_i^3}, & x_i \leq x \leq x_{i+1} \\ 3\frac{(x-x_{i-1})^2}{\ell_{i-1}^2} - 2\frac{(x-x_{i-1})^3}{\ell_{i-1}^3}, & x_{i-1} \leq x \leq x_i \\ 0, & \text{otherwise} \end{cases} \quad (9.213)$$

$$\alpha_{i,2}(x) = \begin{cases} x - 2\frac{(x-x_i)^2}{\ell_i^2} + \frac{(x-x_i)^3}{\ell_i^3}, & x_i \leq x \leq x_{i+1} \\ -\frac{(x-x_{i-1})^2}{\ell_{i-1}^2} + \frac{(x-x_{i-1})^3}{\ell_{i-1}^3}, & x_{i-1} \leq x \leq x_i \\ 0, & \text{otherwise} \end{cases} \quad (9.214)$$

These shape functions satisfy

$$\psi_{2k-1} = \alpha_{j,1}(x_k) = \delta_{jk}, \quad \psi'_{2k-1} = \alpha'_{j,1}(x_k) = 0 \quad (9.215)$$

$$\psi_{2k} = \alpha_{j,2}(x_k) = 0, \quad \psi'_{2k} = \alpha'_{j,2}(x_k) = \delta_{jk} \quad (9.216)$$

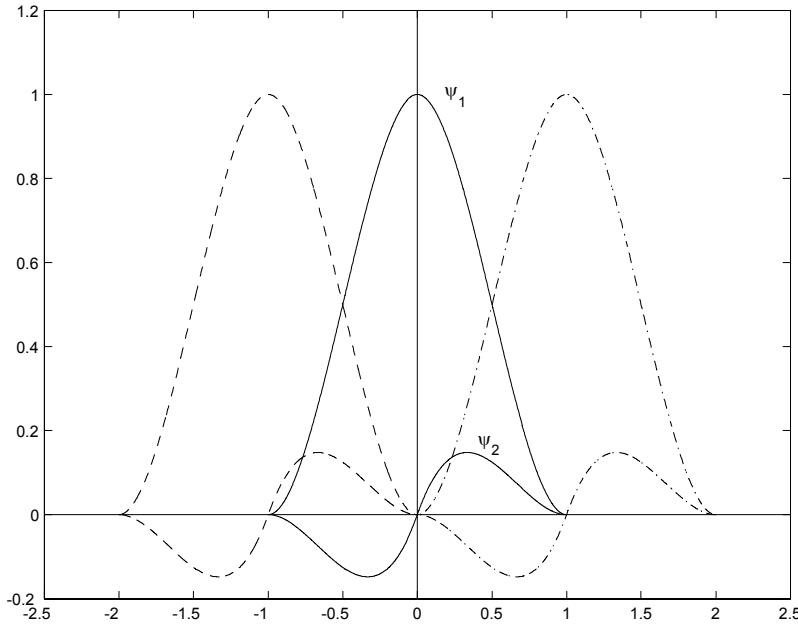


Figure 9.11: Shape functions for Euler-Bernoulli beam.

This gives the following physical interpretation of the generalized coordinates  $q_{2k-1} = a_{k,1}(t)$  and  $q_{2k} = a_{k,2}(t)$ :

$$q_{2k-1} = a_{k,1}(t) = w(x_k, t) \quad (9.217)$$

$$q_{2k} = a_{k,2}(t) = w'(x_k, t) \quad (9.218)$$

Insertion of (9.212) gives

$$\sum_{i=1}^{2N} [\rho \ddot{q}_j(t) \psi_j(x) + \rho c^2 q_j(t) \psi_j'''(x)] = b(x) u(t) \quad (9.219)$$

In the Galerkin method the equation of motion is premultiplied by  $\psi_i(x)$  and integrated over the interval  $x \in [0, \ell]$ . This gives the expression

$$\int_0^\ell \psi_i(x) \sum_{i=1}^{2N} [\rho c^2 q_j(t) \psi_j'''(x) + \rho \ddot{q}_j(t) \psi_j(x)] dx = \int_0^\ell \psi_i(x) b(x) u(t) dx \quad (9.220)$$

Partial integration twice gives

$$\int_0^\ell \psi_i(x) \psi_j'''(x) dx = \psi_i(x) \psi_j'''(x)|_0^\ell - \psi_i'(x) \psi_j''(x)|_0^\ell + \int_0^\ell \psi_i''(x) \psi_j''(x) dx \quad (9.221)$$

and, if we assume that

$$\psi_i(x) \psi_j'''(x)|_0^\ell - \psi_i'(x) \psi_j''(x)|_0^\ell = 0 \quad (9.222)$$

then

$$\sum_{i=1}^N \left[ \int_0^\ell \rho \psi_i(x) \psi_j(x) dx \ddot{q}_i(t) + \int_0^\ell c^2 \rho \psi_i''(x) \psi_j''(x) dx q_j(t) \right] = \int_0^\ell \psi_i(x) b(x) dx u(t)$$

This can be written in vector form as

$$\mathbf{M} \ddot{\mathbf{q}} + \mathbf{K} \mathbf{q} = \mathbf{b} u \quad (9.223)$$

Here  $\mathbf{q} = (q_1, \dots, q_{2N})^T$  is the vector of generalized coordinates,  $\mathbf{M} = \{m_{ij}\}$  is the mass matrix and  $\mathbf{K} = \{k_{ij}\}$  is the stiffness matrix, and  $\mathbf{b} = (b_1, \dots, b_N)^T$  with elements given by

$$m_{ij} = \int_0^\ell \rho \psi_j(x) \psi_i(x) dx \quad (9.224)$$

$$k_{ij} = \int_0^\ell c^2 \rho \psi_j''(x) \psi_i''(x) dx \quad (9.225)$$

$$b_i = \int_0^\ell \psi_i(x) b(x) dx \quad (9.226)$$

Note that

1. The mass matrix  $\mathbf{M}$  and the stiffness matrix  $\mathbf{K}$  are symmetric. Moreover, the  $\mathbf{M}$  and  $\mathbf{K}$  matrices will typically be positive definite.
2. If the orthogonal mode shapes are used, then the mass matrix  $\mathbf{M}$  and the stiffness matrix  $\mathbf{K}$  will be diagonal matrices.

## 9.7 Motor and Euler Bernoulli beam

### 9.7.1 Equations of motion

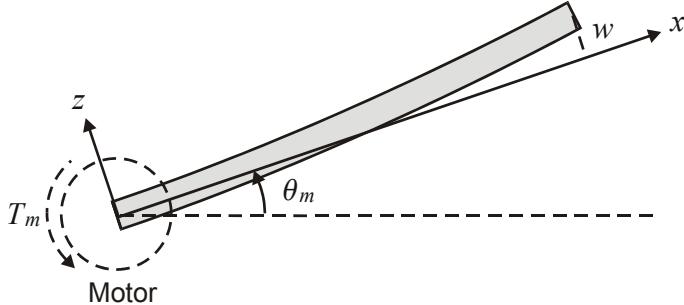


Figure 9.12: Motor with Euler-Bernoulli beam

In this section we will study the dynamics of an Euler Bernoulli beam that is rotated by a motor. This can be seen as a simplified model of a robotic joint with an elastic link. The motor has moment of inertia  $J_m$  and joint angle  $\theta_m$ , and the Euler Bernoulli beam

has length  $\ell$  as shown in Figure 9.12. The position of a point  $x$  along the beam is given by

$$\eta(x, t) = x\theta_m(t) + w(x, t) \quad (9.227)$$

where  $\eta(x, t)$  is the arc length from the reference position along a circle of radius  $x$ , and  $w(x, t)$  is the elastic deflection.

For this system the dynamics are the same as for the usual Euler Bernoulli model, except that the acceleration is  $\partial^2\eta/\partial t^2$  in place of  $\partial^2w/\partial t^2$ . This gives

$$c^2 \frac{\partial^4 w}{\partial x^4}(x, t) + \frac{\partial^2 \eta}{\partial t^2}(x, t) = 0 \quad (9.228)$$

The equation of motion for the motor shaft can be found from

$$\dot{h}(t) = T_m(t) \quad (9.229)$$

where

$$h(t) = J_m \dot{\theta}_m + \int_0^\ell [x \dot{\theta}_m + \dot{w}(x, t)] \rho x dx \quad (9.230)$$

is the angular momentum of the motor and beam. The equation of motion is found to be

$$J_t \ddot{\theta}_m(t) + \int_0^\ell \rho x \dot{w}(x, t) dx = T_m(t) \quad (9.231)$$

where

$$J_t = J_m + \int_0^\ell \rho x^2 dx \quad (9.232)$$

is the total inertia seen from the motor.

### 9.7.2 Assumed mode shapes

An elastic beam is fixed to a moving base, and the elastic deflection of the beam is described by

$$w(x, t) = \sum_{j=1}^{\infty} \phi_j(x) q_i(t) \quad (9.233)$$

where  $\phi_j(x)$  is a set of shape functions. Typically, the orthogonal shape functions may be available from previous analysis, or even from textbooks, and a widely used method is to approximate the solution by assuming that the orthogonal modes of the beam itself is a sufficiently accurate approximation in (9.233). To demonstrate how this can be done we will use the orthogonal shape functions  $\phi_j(x)$  of a pinned Euler Bernoulli beam, which is an accurate approximation if the inertia of the motor is large compared to the inertia of the beam.

The angular momentum of the motor axis and beam is

$$h(t) = J_m \dot{\theta}_m + \int_0^\ell [x \dot{\theta}_m + \sum_{j=1}^{\infty} \phi_j(x) \dot{q}_j(t)] \rho x dx \quad (9.234)$$

which gives the equation of motion

$$J_t \ddot{\theta}_m(t) + \sum_{j=1}^{\infty} h_j \ddot{q}_j(t) = T_m(t) \quad (9.235)$$

where

$$h_j = \int_0^\ell \rho \phi_j(x) x dx \quad (9.236)$$

is the angular momentum coefficient for mode shape  $j$ .

Insertion of (9.227) and (9.233) into (9.228) gives

$$\sum_{j=1}^{\infty} [c^2 \phi_j'''(x) q_j(t) + \phi_j(x) \ddot{q}_j(t)] + x \ddot{\theta}_m(t) = 0 \quad (9.237)$$

As the mode shapes  $\phi_j$  are eigenfunctions that satisfy (9.124), this can be written

$$\sum_{i=1}^{\infty} \phi_i(x) [\omega_i^2 q_i(t) + \ddot{q}_i(t)] + x \ddot{\theta}_m(t) = 0 \quad (9.238)$$

where the natural frequencies are given by  $\omega_j^2 = c^2 \beta_j^4$ . Finally, we may find the differential equations for the generalized coordinates  $q_i$  by multiplying with  $\phi_i(x)$  and integrating over the interval  $x \in [0, \ell]$ . This gives

$$\int_0^\ell \rho \phi_i(x) \left\{ \sum_{j=1}^{\infty} \phi_j(x) [\omega_j^2 q_j(t) + \ddot{q}_j(t)] + x \ddot{\theta}_m(t) \right\} dx = 0 \quad (9.239)$$

Then, by accounting for the orthogonality of the shape functions as expressed in (9.165) we arrive at the differential equations

$$\omega_i^2 q_i(t) + \ddot{q}_i(t) + h_i \ddot{\theta}_m(t) = 0 \quad (9.240)$$

where  $i$  has been inserted for  $j$ , and the angular momentum coefficients  $h_i$  are given by

$$h_i = \int_0^\ell \rho \phi_i(x) x dx \quad (9.241)$$

The model can be written

$$\begin{pmatrix} J_t & h_1 & \dots & h_i & \dots \\ h_1 & 1 & \dots & 0 & \dots \\ \vdots & \vdots & \ddots & \vdots & \dots \\ h_i & 0 & \dots & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} \ddot{\theta}_m \\ \ddot{q}_1 \\ \vdots \\ \ddot{q}_i \\ \vdots \end{pmatrix} + \begin{pmatrix} 0 & 0 & \dots & 0 & \dots \\ 0 & \omega_1^2 & \dots & 0 & \dots \\ \vdots & \vdots & \ddots & \vdots & \dots \\ 0 & 0 & \dots & \omega_i^2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} \theta_m \\ q_1 \\ \vdots \\ q_i \\ \vdots \end{pmatrix} = \begin{pmatrix} T_m \\ 0 \\ \vdots \\ 0 \\ \vdots \end{pmatrix}$$

This type of model formulation is used to simulate the dynamics of flexible antennas mounted on satellites (Hughes 1974).

### 9.7.3 Finite elements

A model for a beam mounted on a motor axis can also be derived with the finite-element method. Again, the position of a point  $x$  along the beam is given by  $\eta(x, t) = x \theta_m(t) + w(x, t)$ , and the elastic deflection of the beam is described by

$$w(x, t) = \sum_{j=1}^{2N} \psi_j(x) q_j(t) \quad (9.242)$$

where  $\psi_j(x)$  are the Hermitian shape functions used in the finite-element method. The partial differential equation (9.228) is then approximated by

$$c^2 \boldsymbol{\psi}'''(x)^T \mathbf{q}(t) + \boldsymbol{\psi}(x)^T \ddot{\mathbf{q}}_i(t) + x \ddot{\theta}_m(t) = 0 \quad (9.243)$$

and Galerkin's method leads to

$$\int_0^\ell \rho \boldsymbol{\psi}(x) \left[ c^2 \boldsymbol{\psi}'''(x)^T \mathbf{q}(t) + \boldsymbol{\psi}(x)^T \ddot{\mathbf{q}}_i(t) + x \ddot{\theta}_m(t) \right] dx = 0 \quad (9.244)$$

Insertion of (9.221) gives

$$\int_0^\ell \left[ c^2 \rho \boldsymbol{\psi}''(x)^T \boldsymbol{\psi}''(x) q_i(t) + \rho \boldsymbol{\psi}(x) \boldsymbol{\psi}(x)^T \ddot{q}_i(t) + \rho \boldsymbol{\psi}(x) x \ddot{\theta}_m(t) \right] dx = 0 \quad (9.245)$$

The equation of motion for the motor is approximated by

$$J_t \ddot{\theta}_m(t) + \sum_{j=1}^{2N} h_j \ddot{q}_j(t) = T_m(t) \quad (9.246)$$

where

$$J_t = J_m + \int_0^\ell \rho x^2 dx \quad (9.247)$$

is the total inertia seen from the motor, and

$$h_j = \int_0^\ell \rho \psi_j(x) x dx \quad (9.248)$$

are the influence coefficients. This leads to the model

$$\mathbf{M} = \begin{pmatrix} J_t & \mathbf{h}^T \\ \mathbf{h} & \mathbf{M}_{fe} \end{pmatrix}, \quad \mathbf{K} = \begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{K}_{fe} \end{pmatrix} \quad (9.249)$$

$$\mathbf{q} = \begin{pmatrix} \theta_m \\ q_1 \\ \vdots \\ q_N \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \mathbf{h} = \begin{pmatrix} h_1 \\ \vdots \\ h_N \end{pmatrix} \quad (9.250)$$

where the mass matrix  $\mathbf{M}_{fe}$  and the stiffness matrix  $\mathbf{K}_{fe}$  of the finite-element method are given by

$$\mathbf{M}_{fe} = \int_0^\ell \rho \boldsymbol{\psi} \boldsymbol{\psi}^T dx, \quad \mathbf{K}_{fe} = \int_0^\ell \rho \boldsymbol{\psi}'' (\boldsymbol{\psi}'')^T dx \quad (9.251)$$

## 9.8 Irrational transfer functions for beam dynamics

### 9.8.1 Introduction

So far the dynamics of the Euler Bernoulli beam has been described using a series expansion of a solution based on shape functions. In this section it will be shown that transfer functions for the beam dynamics can be derived directly from the partial differential equation. This is of great interest in itself, but it also gives useful information on the dynamics of the system as there are no approximations involved. In particular, the singularities and zeros of the dynamics can be found.

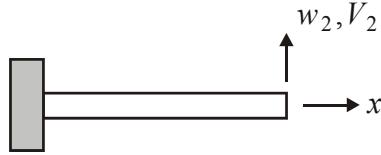


Figure 9.13: Clamped beam with excitation force  $V_2$  and deflection  $w_2$  on the tip.

### 9.8.2 Clamped-free beam

Consider a clamped-free beam which is clamped at  $x = 0$  and free at  $x = L$  as shown in Figure 9.13. The beam is excited with a force  $V_2$  at the end, and we will in the following derive the transfer function from the tip force  $V_2$  to the tip deflection  $w_2$ . Laplace transformation of the partial differential equation (9.105) gives

$$\frac{\partial^4}{\partial x^4}w(x, s) + \frac{s^2}{c^2}w(x, s) = 0 \quad (9.252)$$

Define the complex variable  $\gamma(s)$  by the relation

$$\gamma^4 = -\frac{s^2}{c^2} \quad (9.253)$$

Then the Laplace transformation of the partial differential equation (9.266) gives the ordinary differential equation

$$\frac{\partial^4}{\partial x^4}w(x, s) - \gamma^4w(x, s) = 0 \quad (9.254)$$

The solution to the Laplace transformed model (9.254) is

$$w(x, s) = C_1 \cos \gamma x + C_2 \sin \gamma x + C_3 \cosh \gamma x + C_4 \sinh \gamma x \quad (9.255)$$

while the derivatives are

$$w'(x, s) = \gamma(-C_1 \sin \gamma x + C_2 \cos \gamma x + C_3 \sinh \gamma x + C_4 \cosh \gamma x) \quad (9.256)$$

$$w''(x, s) = \gamma^2(-C_1 \cos \gamma x - C_2 \sin \gamma x + C_3 \cosh \gamma x + C_4 \sinh \gamma x) \quad (9.257)$$

$$w'''(x, s) = \gamma^3(C_1 \sin \gamma x - C_2 \cos \gamma x + C_3 \sinh \gamma x + C_4 \cosh \gamma x) \quad (9.258)$$

For the clamped-free beam the boundary conditions imply that  $w_1 = 0$ ,  $w'_1 = \theta_1 = 0$ , and that  $M_2$  and  $V_2$  are functions of time only. This means that  $w_1$ ,  $\theta_1$ ,  $M_2$  and  $V_2$  should be considered as input to the system, whereas  $w_2$ ,  $\theta_2$ ,  $M_1$  and  $V_1$  can be considered as outputs of the system dynamics. Note that it is critical in the method that is presented here that the appropriate variables are selected as inputs and outputs. From (9.255–9.258) we find that the input variables can be expressed in terms of the constant  $C_i$  as

$$\underbrace{\begin{pmatrix} w_1(s) \\ \frac{1}{\gamma}\theta_1(s) \\ \frac{1}{\gamma^2 EI}M_2(s) \\ \frac{1}{\gamma^3 EI}V_2(s) \end{pmatrix}}_{\mathbf{u}(s)} = \underbrace{\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ -\cos \gamma L & -\sin \gamma L & \cosh \gamma L & \sinh \gamma L \\ \sin \gamma L & -\cos \gamma L & \sinh \gamma L & \cosh \gamma L \end{pmatrix}}_{\mathbf{G}(s)} \underbrace{\begin{pmatrix} C_1 \\ C_2 \\ C_3 \\ C_4 \end{pmatrix}}_{\mathbf{c}} \quad (9.259)$$

while the output variables can be expressed by

$$\underbrace{\begin{pmatrix} w_2(s) \\ \frac{1}{\gamma}\theta_2(s) \\ \frac{1}{\gamma^2 EI}M_1(s) \\ \frac{1}{\gamma^3 EI}V_1(s) \end{pmatrix}}_{\mathbf{y}(s)} = \underbrace{\begin{pmatrix} \cos \gamma L & \sin \gamma L & \cosh \gamma L & \sinh \gamma L \\ -\sin \gamma L & \cos \gamma L & \sinh \gamma L & \cosh \gamma L \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}}_{\mathbf{K}(s)} \underbrace{\begin{pmatrix} C_1 \\ C_2 \\ C_3 \\ C_4 \end{pmatrix}}_{\mathbf{c}} \quad (9.260)$$

Then, as  $\mathbf{G}(s)$  is nonsingular, we can find the transfer function matrix  $\mathbf{H}(s)$  from the expression

$$\mathbf{y}(s) = \mathbf{H}(s)\mathbf{u}(s) = \mathbf{K}(s)\mathbf{G}(s)^{-1}\mathbf{u}(s) \quad (9.261)$$

where  $\mathbf{u}(s)$ ,  $\mathbf{y}(s)$ ,  $\mathbf{G}(s)$ , and  $\mathbf{K}(s)$  are defined in (9.259) and (9.260). Using a symbolic program like MATLAB or MAPLE, we find that the transfer function from the tip force  $V_2$  to the tip deflection  $w_2$  is given by

$$\frac{w_2(s)}{V_2(s)} = \frac{1}{\gamma^3 EI} \frac{\cos \gamma L \sinh \gamma L - \sin \gamma L \cosh \gamma L}{1 + \cos \gamma L \cosh \gamma L} \quad (9.262)$$

The singularities of the transfer function are found by equating the denominator to zero. This gives

$$1 + \cos \gamma L \cosh \gamma L = 0 \quad (9.263)$$

which is the characteristic equation for the clamped-free beam. This means that the natural frequencies of the irrational transfer functions are given by the natural frequencies found from the analysis based on the orthogonal mode shapes. This makes sense, as both methods are exact. The zeros of the transfer function are found from

$$\cos \gamma L \sinh \gamma L - \sin \gamma L \cosh \gamma L = 0 \quad (9.264)$$

which is the characteristic equation for the clamped-pinned beam. This can be explained as the dynamics associated with the zeros gives small amplification from the force  $V_2(s)$  to the deflection  $w_2(s)$ . The clamped-pinned beam has zero deflection  $w_2(s)$ , and it seems reasonable that this is reflected in the location of the zeros.

### 9.8.3 Motor and beam

In this section transfer functions for a motor and an Euler Bernoulli beam of length  $L$  will be derived. The results are based on (Wie and Bryson 1987) and (Gevarter 1970). The starting point for the derivation is the definition of the variable

$$\eta(x, t) = w(x, t) + x\theta(t) \quad (9.265)$$

which leads to the partial differential equation

$$\frac{\partial^4}{\partial x^4}\eta(x, t) + \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\eta(x, t) = 0 \quad (9.266)$$

which is the same as for the beam model. The position variable and the angles at the end-points are given by

$$\eta_1(s) = \eta(0, s), \quad \eta_2(s) = \eta(L, s) \quad (9.267)$$

$$\theta_1(s) = \eta'(0, s), \quad \theta_2(s) = \eta'(L, s) \quad (9.268)$$

while the shear forces and moments at the end-points are given by

$$\frac{1}{EI}V_1(s) = \eta'''(0, s), \quad \frac{1}{EI}V_2(s) = \eta'''(L, s) \quad (9.269)$$

$$\frac{1}{EI}M_1(s) = \eta''(0, s), \quad \frac{1}{EI}M_2(s) = \eta''(L, s) \quad (9.270)$$

The beam is pinned to the motor at  $x = 0$  and is free at  $x = L$ , which leads to the boundary conditions

$$\eta(0, t) = \eta''(L, t) = \eta'''(L, t) = 0 \quad (9.271)$$

$$\eta''(0, t) = \frac{1}{EI}M_1(t) = T_L(t) \quad (9.272)$$

where  $T_L(t)$  is the torque from the motor shaft on the beam.

We will now develop a transfer function model of the form  $\mathbf{y}(s) = \mathbf{H}(s)\mathbf{u}(s)$  of the motor and beam. First we have to select the input and output variables of the model. The input and output variables are  $\eta_1$ ,  $\eta_2$ ,  $\theta_1$ ,  $\theta_2$ ,  $V_1$ ,  $V_2$ ,  $M_1$  and  $M_2$ . The boundary conditions (9.271) and (9.272) imply that  $\eta_1$ ,  $M_2$  and  $V_2$  are zero, and that  $M_1$  is a function of time only. Therefore these variables must be in the input vector  $\mathbf{u}$  of the transfer function model. The remaining four variables  $\eta_2$ ,  $\theta_1$ ,  $\theta_2$ , and  $V_1$  are in the output vector  $\mathbf{y}$ .

The partial differential equation (9.266) for  $\eta$  is the same as the partial differential equation (9.105) for  $w$ . Therefore, the solution  $\eta(x, s)$  and its derivatives with respect to  $x$  will be given by (9.255–9.258), and the input and output variables can be expressed by

$$\underbrace{\begin{pmatrix} \frac{1}{\gamma^2 EI} M_1(s) \\ \eta_1(s) \\ \frac{1}{\gamma^3 EI} V_2(s) \\ \frac{1}{\gamma^2 EI} M_2(s) \end{pmatrix}}_{\mathbf{u}(s)} = \underbrace{\begin{pmatrix} -1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ \sin \gamma L & -\cos \gamma L & \sinh \gamma L & \cosh \gamma L \\ -\cos \gamma L & -\sin \gamma L & \cosh \gamma L & \sinh \gamma L \end{pmatrix}}_{\mathbf{K}(s)} \underbrace{\begin{pmatrix} C_1 \\ C_2 \\ C_3 \\ C_4 \end{pmatrix}}_{\mathbf{c}} \quad (9.273)$$

and

$$\underbrace{\begin{pmatrix} \frac{1}{\gamma} \theta_1(s) \\ \eta_2(s) \\ \frac{1}{\gamma} \theta_2(s) \\ \frac{1}{\gamma^3 EI} V_1(s) \end{pmatrix}}_{\mathbf{y}(s)} = \underbrace{\begin{pmatrix} 0 & 1 & 0 & 1 \\ \cos \gamma L & \sin \gamma L & \cosh \gamma L & \sinh \gamma L \\ -\sin \gamma L & \cos \gamma L & \sinh \gamma L & \cosh \gamma L \\ 0 & -1 & 0 & 1 \end{pmatrix}}_{\mathbf{G}(s)} \underbrace{\begin{pmatrix} C_1 \\ C_2 \\ C_3 \\ C_4 \end{pmatrix}}_{\mathbf{c}} \quad (9.274)$$

The matrix  $\mathbf{G}(s)$  is nonsingular, we can find the transfer function matrix  $\mathbf{H}(s)$  from the expression

$$\mathbf{y}(s) = \mathbf{H}(s)\mathbf{c}(s) = \mathbf{K}(s)\mathbf{G}(s)^{-1}\mathbf{u}(s) \quad (9.275)$$

where  $\mathbf{u}(s)$ ,  $\mathbf{y}(s)$ ,  $\mathbf{G}(s)$ , and  $\mathbf{K}(s)$  are defined in (9.259) and (9.260). Using a symbolic program like MATLAB or MAPLE, we find that the transfer function from the moment to the angle at  $x = 0$  is given by

$$\frac{\theta_1}{M_1}(s) = \frac{1}{EI} \frac{1 + \cos \gamma L \cosh \gamma L}{\gamma (\sin \gamma L \cosh \gamma L - \cos \gamma L \sinh \gamma L)} \quad (9.276)$$

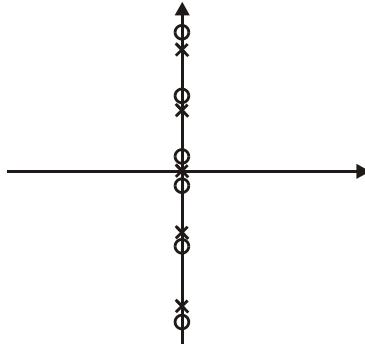
The zeros and the singularities are easily found for this transfer function as the numerator is equal to zero whenever the characteristic equation for the clamped-free beam is satisfied, whereas the denominator is zero when the characteristic equation of the pinned-free

Zeros		Singularities		(9.277)
$\gamma L$	$\frac{s}{c/L^2}$	$\gamma L$	$\frac{s}{c/L^2}$	
1.875104	$\pm j3.516$	0	$\pm 0$	
4.694091	$\pm j22.034$	3.926602	$\pm j15.418$	
7.854757	$\pm 61.697$	7.068583	$\pm j49.964$	
10.995541	$\pm j120.9019$	10.210176	$\pm j104.2477$	

Table 9.4: The first zeros and singularities for the transfer function from  $M_1(s)$  to  $\theta_1(s)$ .

beam is satisfied. These solutions are tabulated in standard textbooks like (Rao 1990). Then the zeros and singularities are given by Table 9.4.

The singularities and zeros are along the imaginary axis as shown in Figure 9.14. This agrees with the fact that the transfer function  $s\theta_1/M_1$  is positive real, which can be shown from energy arguments. The transfer function from the moment at  $x = 0$  to

Figure 9.14: Singularities and zeros for the transfer function from the motor torque to  $M_1$  the motor angle  $\theta_1$ . The singularities are marked with crosses, and the zeros are marked with circles.

the deflection at the other side of the beam is given by

$$\frac{\eta_2}{M_1}(s) = \frac{1}{EI} \frac{\sin \gamma L + \sinh \gamma L}{\gamma^2 (\sin \gamma L \cosh \gamma L - \cos \gamma L \sinh \gamma L)} \quad (9.278)$$

The singularities are the same as for the transfer function (9.276), while it is possible to verify that the numerator expression  $\sin z + \sinh z = 0$  has solutions of the type  $z = \rho(1 + j)$  where  $\rho$  is the solution of  $\tan \rho + \tanh \rho = 0$ . This gives the zeros and singularities shown in Table 9.5.

Note that this transfer function has zeros in the right half plane as shown in Figure 9.15.

The beam is connected to the motor by requiring

$$J_m s^2 \theta_1 = T - M_1 \quad (9.280)$$

which gives

$$\theta(s) = \frac{1 + \cos \gamma L \cosh \gamma L}{J_m s^2 (1 + \cos \gamma L \cosh \gamma L) + EI \gamma (\sin \gamma L \cosh \gamma L - \cos \gamma L \sinh \gamma L)} T(s) \quad (9.281)$$

Zeros		Singularities	
$\gamma L$	$\frac{s}{c/L^2}$	$\gamma L$	$\frac{s}{c/L^2}$
$2.365(1+j)$	$\pm 11.1865$	0	$\pm 0$
$5.498(1+j)$	$\pm 60.4560$	3.926602	$\pm j15.418$
$8.639(1+j)$	$\pm 149.2646$	7.068583	$\pm j49.964$

(9.279)

Table 9.5: The first zeros and singularities for the transfer function from the motor torque  $M_1(s)$  to tip position  $\eta_2(s)$ .

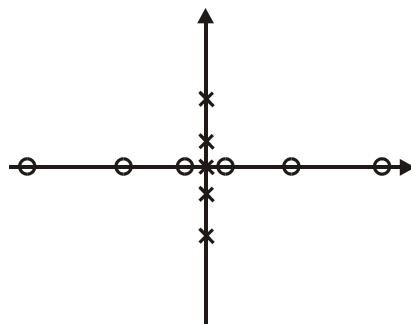


Figure 9.15: Singularities and zeros of the transfer function from the motor torque  $M_1$  to the end-point position  $\eta_2$ . The singularities are marked with crosses, and the zeros are marked with circles.

# **Part IV**

## **Balance equations**



# Chapter 10

## Kinematics of Flow

### 10.1 Introduction

Balance equations are differential equations that are derived from conservation laws for control volumes. The conservation laws include the conservation of mass, momentum, angular momentum, and energy. Fluid flow phenomena are important in the derivation of the balance laws of this chapter, and therefore the presentation starts with the kinematics of fluid flow. Then the transport theorem is presented, and it is shown how the transport theorem can be used to derive balance equations for typical control volumes. Balance equations for mass, momentum, angular momentum and energy are then developed using the mathematical tools presented in the first part of the chapter.

### 10.2 Kinematics

#### 10.2.1 The material derivative

Let  $x_1, x_2, x_3$  be the coordinates of a Cartesian frame with orthogonal unit vectors  $\vec{a}_1, \vec{a}_2, \vec{a}_3$ . A scalar function  $\phi = \phi(t, x_1, x_2, x_3)$  of time and position is called a scalar field. The time derivative of a scalar field  $\phi$  is, according to the usual definition of the derivative,

$$\frac{d\phi}{dt} = \frac{\partial\phi}{\partial t} + \frac{\partial\phi}{\partial x_1} \frac{dx_1}{dt} + \frac{\partial\phi}{\partial x_2} \frac{dx_2}{dt} + \frac{\partial\phi}{\partial x_3} \frac{dx_3}{dt}$$

We will also be dealing with *vector fields*  $\mathbf{u} = \mathbf{u}(t, x_1, x_2, x_3)$  where  $\mathbf{u} = (u_1, u_2, u_3)^T$ . The time derivative of a vector field  $\mathbf{u}$  is

$$\frac{d\mathbf{u}}{dt} = \frac{\partial\mathbf{u}}{\partial t} + \frac{\partial\mathbf{u}}{\partial x_1} \frac{dx_1}{dt} + \frac{\partial\mathbf{u}}{\partial x_2} \frac{dx_2}{dt} + \frac{\partial\mathbf{u}}{\partial x_3} \frac{dx_3}{dt}$$

The time derivative  $\frac{d\phi}{dt}$  of the scalar field  $\phi$  clearly depends on the time derivatives  $\dot{x}_i = \frac{dx_i}{dt}$  of the position coordinates  $x_i$ . This means that it must be specified for which time function  $\mathbf{x}(t)$  the derivative is taken. Two cases are common: One is the *spatial derivative* which is the derivative at a specific point  $\mathbf{x} = \mathbf{x}_0$  where  $\mathbf{x}_0$  is a constant vector. Then the position coordinates  $x_i$  are constants, and

$$\left. \frac{d\phi}{dt} \right|_{\mathbf{x}=\mathbf{x}_0} = \frac{\partial\phi}{\partial t}$$

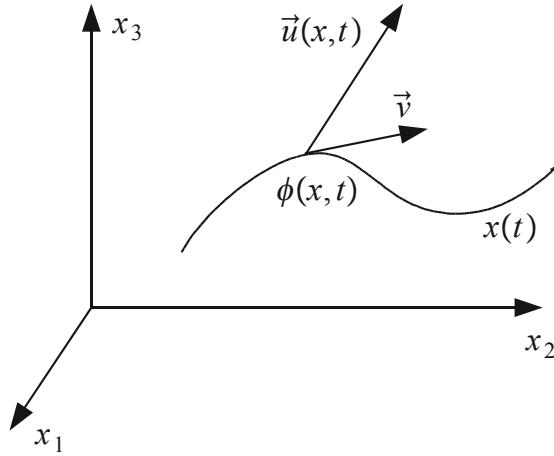


Figure 10.1: Coordinate frame with the motion  $\mathbf{x}(t)$  of a particle.

In the same way the spatial derivative of a vector field is

$$\frac{d\mathbf{u}}{dt} \Big|_{\mathbf{x}=\mathbf{x}_0} = \frac{\partial \mathbf{u}}{\partial t}$$

The other usual case is the *material derivative*, which is the time derivative when following a particular particle of the fluid. Then  $\dot{\mathbf{x}} = \mathbf{v}$ , where  $\mathbf{v} = (v_1, v_2, v_3)^T$  is the velocity of the fluid. The material derivative is widely used, and this motivates the introduction of the notation

$$\frac{D\phi}{Dt} := \frac{d\phi}{dt} \Big|_{\dot{\mathbf{x}}=\mathbf{v}}, \quad \frac{D\mathbf{u}}{Dt} := \frac{d\mathbf{u}}{dt} \Big|_{\dot{\mathbf{x}}=\mathbf{v}}$$

The material derivative of a scalar field  $\phi$  is defined by

$$\frac{D\phi}{Dt} := \frac{\partial \phi}{\partial t} + \frac{\partial \phi}{\partial x_1} v_1 + \frac{\partial \phi}{\partial x_2} v_2 + \frac{\partial \phi}{\partial x_3} v_3 \quad (10.1)$$

while the material derivative of a vector field  $\mathbf{u}$  is given by

$$\frac{D\mathbf{u}}{Dt} = \frac{\partial \mathbf{u}}{\partial t} + \frac{\partial \mathbf{u}}{\partial x_1} v_1 + \frac{\partial \mathbf{u}}{\partial x_2} v_2 + \frac{\partial \mathbf{u}}{\partial x_3} v_3$$

### 10.2.2 The nabla operator

The *nabla vector operator*  $\vec{\nabla}$  is defined in the Cartesian coordinate system by

$$\vec{\nabla} = \vec{a}_1 \frac{\partial}{\partial x_1} + \vec{a}_2 \frac{\partial}{\partial x_2} + \vec{a}_3 \frac{\partial}{\partial x_3} \quad (10.2)$$

When the vector operator  $\vec{\nabla}$  is applied to a scalar field  $\phi$ , we get the gradient vector

$$\vec{\nabla} \phi = \frac{\partial \phi}{\partial x_1} \vec{a}_1 + \frac{\partial \phi}{\partial x_2} \vec{a}_2 + \frac{\partial \phi}{\partial x_3} \vec{a}_3 \quad (10.3)$$

The nabla vector operator may be represented by a column vector as

$$\nabla := \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \frac{\partial}{\partial x_3} \end{pmatrix} \quad (10.4)$$

It follows that

$$\nabla \phi = \begin{pmatrix} \frac{\partial \phi}{\partial x_1} \\ \frac{\partial \phi}{\partial x_2} \\ \frac{\partial \phi}{\partial x_3} \end{pmatrix}, \quad \nabla \mathbf{u}^T = \left\{ \frac{\partial u_j}{\partial x_i} \right\} \quad (10.5)$$

It is then straightforward to show that

$$\vec{v} \cdot \vec{\nabla} \phi = \mathbf{v}^T \nabla \phi = \frac{\partial \phi}{\partial x_1} v_1 + \frac{\partial \phi}{\partial x_2} v_2 + \frac{\partial \phi}{\partial x_3} v_3 \quad (10.6)$$

With some care it is also found that

$$\mathbf{v}^T \nabla \mathbf{u} = (\mathbf{v}^T \nabla) \mathbf{u} = \frac{\partial \mathbf{u}}{\partial x_1} v_1 + \frac{\partial \mathbf{u}}{\partial x_2} v_2 + \frac{\partial \mathbf{u}}{\partial x_3} v_3 \quad (10.7)$$

In the vector notation with  $\vec{u} = u_1 \vec{a}_1 + u_2 \vec{a}_2 + u_3 \vec{a}_3$ , then

$$\vec{v} \cdot \vec{\nabla} \vec{u} = (\vec{v} \cdot \vec{\nabla}) \vec{u} = v_1 \frac{\partial \vec{u}}{\partial x_1} + v_2 \frac{\partial \vec{u}}{\partial x_2} + v_3 \frac{\partial \vec{u}}{\partial x_3} \quad (10.8)$$

The material derivative of a scalar field  $\phi$  and a vector field  $\mathbf{u}$  can be written

$$\frac{D\phi}{Dt} := \frac{\partial \phi}{\partial t} + \mathbf{v}^T \nabla \phi, \quad \frac{D\mathbf{u}}{Dt} := \frac{\partial \mathbf{u}}{\partial t} + \mathbf{v}^T \nabla \mathbf{u} \quad (10.9)$$

or, alternatively, in vector form as

$$\frac{D\phi}{Dt} = \frac{\partial \phi}{\partial t} + \vec{v} \cdot \vec{\nabla} \phi, \quad \frac{D\vec{u}}{Dt} = \frac{\partial \vec{u}}{\partial t} + \vec{v} \cdot \vec{\nabla} \vec{u} \quad (10.10)$$

### 10.2.3 Divergence

The divergence of a vector field  $\vec{u} = u_1 \vec{a}_1 + u_2 \vec{a}_2 + u_3 \vec{a}_3$  is the scalar

$$\vec{\nabla} \cdot \vec{u} = \nabla^T \mathbf{u} = \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3} \quad (10.11)$$

The divergence appears in many results. In particular, its usefulness is due to the divergence theorem:

**The Divergence Theorem:** Consider a volume  $V$  with a closed surface  $\partial V$  and an outwards pointing surface normal  $\vec{n}$ , where  $\vec{n}$  is a unit vector. Let  $dV$  be a volume element and  $dA$  a surface element. Then, for any vector field  $\vec{u} = \vec{u}(\mathbf{x})$  we have

$$\iint_{\partial V(t)} \vec{u} \cdot \vec{n} dA = \iiint_{V(t)} \vec{\nabla} \cdot \vec{u} dV \quad (10.12)$$

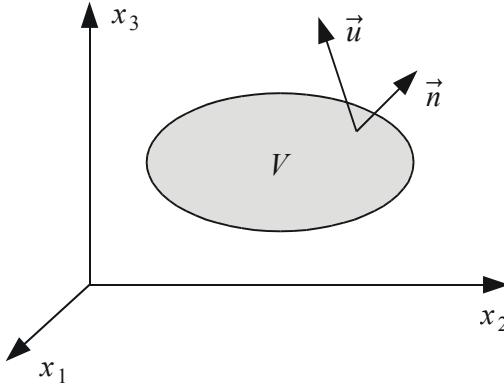


Figure 10.2: Volume  $V$  with outwards pointing surface normal  $\mathbf{n}$  and the vector  $\mathbf{u}$ .

**Example 151** A result related to the divergence theorem is

$$\iint_{\partial V(t)} \phi \vec{n} dA = \iiint_{V(t)} \vec{\nabla} \phi dV \quad (10.13)$$

The first component of this vector equation follows from the divergence theorem by letting  $\vec{u} = \phi \vec{a}_1$ , and the second and third element is found in a similar way.

**Example 152** The divergence of the vector  $\phi \vec{u}$  is

$$\begin{aligned} \vec{\nabla} \cdot (\phi \vec{u}) &= \frac{\partial \phi u_1}{\partial x_1} + \frac{\partial \phi u_2}{\partial x_2} + \frac{\partial \phi u_3}{\partial x_3} \\ &= \frac{\partial \phi}{\partial x_1} u_1 + \frac{\partial \phi}{\partial x_2} u_2 + \frac{\partial \phi}{\partial x_3} u_3 + \phi \left( \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3} \right) \end{aligned} \quad (10.14)$$

and we see that

$$\vec{\nabla} \cdot (\phi \vec{u}) = (\vec{\nabla} \phi) \cdot \vec{u} + \phi (\vec{\nabla} \cdot \vec{u}) \quad (10.15)$$

#### 10.2.4 Curl

The curl of a vector  $\vec{u} = \vec{u}(\mathbf{x})$  is the vector

$$\vec{\nabla} \times \vec{u} = \begin{vmatrix} \vec{a}_1 & \vec{a}_2 & \vec{a}_3 \\ \frac{\partial}{\partial x_1} & \frac{\partial}{\partial x_2} & \frac{\partial}{\partial x_3} \\ u_1 & u_2 & u_3 \end{vmatrix} = \sum_{i=1}^3 \sum_{j=1}^3 \varepsilon_{ijk} \vec{a}_i \frac{\partial u_k}{\partial x_j} \quad (10.16)$$

The coordinate form is seen to be

$$\nabla^\times \mathbf{u} = \begin{pmatrix} 0 & -\frac{\partial}{\partial x_3} & \frac{\partial}{\partial x_2} \\ \frac{\partial}{\partial x_3} & 0 & -\frac{\partial}{\partial x_1} \\ -\frac{\partial}{\partial x_2} & \frac{\partial}{\partial x_1} & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} \frac{\partial u_3}{\partial x_2} - \frac{\partial u_2}{\partial x_3} \\ \frac{\partial u_1}{\partial x_3} - \frac{\partial u_3}{\partial x_1} \\ \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2} \end{pmatrix} \quad (10.17)$$

where  $\nabla^\times$  is the skew symmetric form of  $\nabla$ . The curl of a vector is used in Stokes' Theorem:

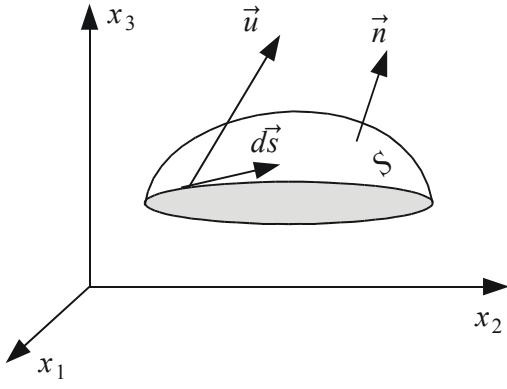


Figure 10.3: Surface  $S$  with surface normal  $\mathbf{n}$  and tangential differential vector  $d\mathbf{s}$  along the closed boundary  $\partial S$ .

**Stokes' Theorem:** Consider a surface  $S$  with a boundary  $\partial S$ , which is a closed curve. Let  $d\vec{s}$  be the differential position increment which is tangent to the curve  $\partial S$ . Let  $dA$  be an area element on the surface, and let  $\vec{n}$  be a surface normal so that the direction of  $d\vec{s}$  corresponds to a counter-clockwise rotation around  $\vec{n}$ . Then for any vector field  $\vec{u} = \vec{u}(\mathbf{x})$  we have

$$\oint_{\partial S} \vec{u} \cdot d\vec{s} = \iint_S (\vec{\nabla} \times \vec{u}) \cdot \vec{n} dA$$

We see from Stokes' Theorem that if the surface  $S$  is taken to be  $dA$ , then

$$\frac{1}{dA} \oint_{\partial S} \vec{u} \cdot d\vec{s} = (\vec{\nabla} \times \vec{u}) \cdot \vec{n} \quad (10.18)$$

The condition  $\vec{\nabla} \times \vec{u} = \vec{0}$  implies that

$$\oint_{\partial S} \vec{u} \cdot d\vec{s} = 0 \quad (10.19)$$

This is equivalent to the existence of a scalar function  $\psi(\mathbf{x})$  called the potential of  $\vec{u}(\mathbf{x})$ , so that  $\vec{u}(\mathbf{x}) = \vec{\nabla} \psi(\mathbf{x})$ . This is shown in basic textbooks on vector analysis. Here we just comment that this is a consequence of the result

$$\psi(\mathbf{x}_2) - \psi(\mathbf{x}_1) = \int_{\mathbf{x}_1}^{\mathbf{x}_2} \vec{\nabla} \psi(\mathbf{x}) \cdot d\vec{s} \quad (10.20)$$

**Example 153** The skew symmetric form of the column vector form  $\nabla^\times \mathbf{u}$  is found to be

$$(\nabla^\times \mathbf{u})^\times = \left\{ \frac{\partial u_i}{\partial x_j} - \frac{\partial u_j}{\partial x_i} \right\} \quad (10.21)$$

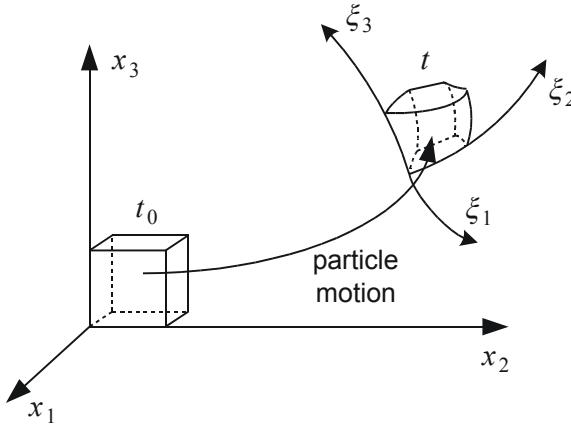


Figure 10.4: The spatial coordinate frame  $(x_1, x_2, x_3)$  and the material coordinate frame  $(\xi_1, \xi_2, \xi_3)$ . At time  $t_0$  the two frames coincide, then the material frame moves with the particles of the fluid so that each particle has the same coordinates  $\xi_1, \xi_2, \xi_3$  for all  $t$ .

**Example 154** The following results can be verified on the component level:

$$\vec{\nabla} \cdot (\vec{\nabla} \times \vec{u}) = 0 \quad (10.22)$$

$$\vec{\nabla} \times (\vec{\nabla} \phi) = 0 \quad (10.23)$$

$$\vec{\nabla} \times (\phi \vec{u}) = (\vec{\nabla} \phi) \times \vec{u} + \phi \vec{\nabla} \times \vec{u} \quad (10.24)$$

$$\frac{1}{2} (\vec{u} \cdot \vec{u}) = (\vec{u} \cdot \vec{\nabla}) \vec{u} + \vec{u} \times (\vec{\nabla} \times \vec{u}) \quad (10.25)$$

$$\vec{\nabla}^2 \vec{u} = \vec{\nabla} (\vec{\nabla} \cdot \vec{u}) - \vec{\nabla} \times (\vec{\nabla} \times \vec{u}) \quad (10.26)$$

### 10.2.5 Material coordinates

For fluids and deformable bodies the concept of *spatial coordinates* and *material coordinates* is useful. The spatial coordinates  $\mathbf{x} = (x_1, x_2, x_3)^T$  define a spatial grid which is constant. In contrast to this, the material coordinates  $\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_3)^T$  define a *material grid* where each particle of the fluid has a given position in the grid. Then, as the fluid deforms, the material grid deforms with the fluid so that each particle maintains its position in the grid. At initial time  $t_0$  the spatial coordinates and the material coordinates coincide, that is,  $\boldsymbol{\xi}(t_0) = \mathbf{x}(t_0)$ . For  $t \geq t_0$  the material coordinates will be a function of the spatial coordinates and *vice versa*, so that

$$\boldsymbol{\xi} = \boldsymbol{\xi} [\mathbf{x}(t), t] \quad \text{and} \quad \mathbf{x} = \mathbf{x} [\boldsymbol{\xi}(t), t] \quad (10.27)$$

### 10.2.6 The dilation

The *dilation* or *expansion* of a fluid (Aris 1989) is closely related to the *divergence* of the velocity, which will be shown in this section. A control volume  $V(t)$  is considered. The control volume is assumed to contain the same fluid particles as it moves with the flow as indicated in Figure 10.5. This means that the mass contained in the volume is

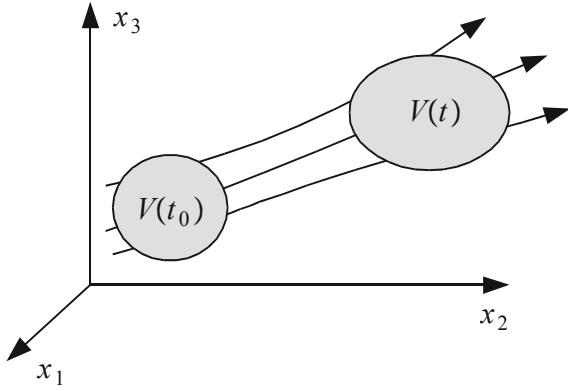


Figure 10.5: Material control volume containing the same set of particles for all  $t$ . The material volume  $V$  moves along with the particles of the fluid, and may be deformed and stretched.

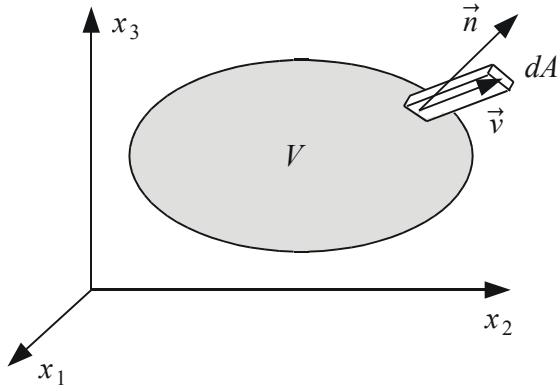


Figure 10.6: Material control volume  $V$  where the change in volume due to the velocity  $\mathbf{v}$  of a surface element  $dA$  is indicated.

constant, while the volume and the surface may change with the flow. Such a control volume will be called a *material control volume*. An area element  $dA(t)$  of the surface  $\partial V(t)$  of  $V(t)$  moves with the velocity  $\vec{v}$ . Therefore, the motion of the area element  $dA$  results in a change of volume with a rate  $\vec{v} \cdot \vec{n} dA$  where  $\vec{n}$  is the outwards surface normal at  $dA$ . Integrating over the whole surface  $\partial V(t)$  of the control volume results in

$$\frac{DV}{Dt} = \iint_{\partial V(t)} \vec{v} \cdot \vec{n} dA \quad (10.28)$$

Then the divergence theorem leads to

$$\frac{DV}{Dt} = \iiint_{V(t)} \vec{\nabla} \cdot \vec{v} dV \quad (10.29)$$

It is seen that if the divergence  $\vec{\nabla} \cdot \vec{v}$  is negative for all of  $V(t)$ , then the time derivative  $\frac{dV(t)}{dt}$  will be negative and the control volume will contract and become smaller in size.

Likewise, if the divergence is positive for all  $V(t)$ , the control volume will expand. An interesting result appears if the control volume is taken to be a *infinitesimal material volume element*  $dV(t)$  which is an infinitesimal volume element containing the same particles as the fluid flows. Then the divergence  $\vec{\nabla} \cdot \vec{v}$  can be taken to be a constant over the control volume, and the time derivative of  $dV$  is the material derivative. Then the divergence theorem gives the following result:

The time derivative of a material volume element  $dV$  is given by

$$\frac{D(dV)}{Dt} = (\vec{\nabla} \cdot \vec{v}) dV \quad (10.30)$$

This shows that the material volume element will diverge if the divergence of velocity is positive, and that it will contract if the divergence is negative.

**Example 155** Consider the specific volume

$$\hat{V} := \frac{dV}{dm} = \frac{1}{\rho} \quad (10.31)$$

which is widely used in thermodynamics. Here  $dV$  is the volume of the mass element  $dm$ , and  $\rho$  is the density. The material derivative of  $\hat{V}$  becomes

$$\frac{D\hat{V}}{Dt} = \frac{D}{Dt} \left( \frac{dV}{dm} \right) = \frac{1}{dm} \frac{D(dV)}{Dt} = \frac{(\vec{\nabla} \cdot \vec{v}) dV}{dm} = (\vec{\nabla} \cdot \vec{v}) \hat{V} \quad (10.32)$$

where  $\vec{v}$  is the velocity vector.

## 10.3 Orthogonal curvilinear coordinates

### 10.3.1 General results

So far we have been working with a Cartesian coordinate system  $(x_1, x_2, x_3)$  with orthogonal unit vectors  $\vec{i}_1, \vec{i}_2, \vec{i}_3$  along the coordinate axes. Other useful coordinate systems are cylindrical coordinates and spherical coordinates, which are examples of *orthogonal curvilinear coordinate systems*. The material in this section is based on (Milne-Thomson 1996, p. 62). The section can be skipped at a first reading. An orthogonal curvilinear coordinate system  $(y_1, y_2, y_3)$  is a coordinate system where the surfaces defined by  $y_1 = y_{1P}$ ,  $y_2 = y_{2P}$ ,  $y_3 = y_{3P}$  intersect orthogonally when  $y_{1P}$ ,  $y_{2P}$  and  $y_{3P}$  are constants. The point of intersection between the planes is denoted by  $P$ . If we draw the surfaces corresponding to  $y_1 = y_{1P}$ ,  $y_2 = y_{2P}$ ,  $y_3 = y_{3P}$  and  $y_1 = y_{1P} + \Delta y_1$ ,  $y_2 = y_{2P} + \Delta y_2$ ,  $y_3 = y_{3P} + \Delta y_3$  we get a figure which to the first order approximation is a parallel piped where one corner is the point  $P$ . The edges of the parallel piped are of length  $h_1 \Delta y_1$ ,  $h_2 \Delta y_2$  and  $h_3 \Delta y_3$  where  $h_1$ ,  $h_2$  and  $h_3$  are functions of the coordinates. To proceed we introduce a Cartesian coordinate system  $(z_1, z_2, z_3)$  with origin in the point  $P$  and with orthogonal unit vectors  $\vec{j}_1, \vec{j}_2, \vec{j}_3$  along the coordinate axes  $(y_1, y_2, y_3)$ , which coincide with the edges of the parallel piped. It is assumed that the unit vectors  $\vec{j}_1, \vec{j}_2, \vec{j}_3$  form a right handed system. In the  $(z_1, z_2, z_3)$  system the length of the edges are denoted  $\Delta z_1$ ,  $\Delta z_2$  and  $\Delta z_3$ . This means that

$$\Delta z_i = h_i \Delta y_i \quad (10.33)$$

and that the volume of the parallel piped is

$$\Delta V = \Delta z_1 \Delta z_2 \Delta z_3 = h_1 h_2 h_3 \Delta y_1 \Delta y_2 \Delta y_3 \quad (10.34)$$

The nabla operator at the point  $P$  described in the Cartesian coordinate system  $(z_1, z_2, z_3)$  is given by

$$\vec{\nabla} = \frac{\partial}{\partial z_1} \vec{j}_1 + \frac{\partial}{\partial z_2} \vec{j}_2 + \frac{\partial}{\partial z_3} \vec{j}_3 \quad (10.35)$$

and, if  $\phi$  is any scalar field, the gradient of  $\phi$  at the point  $P$  is

$$\vec{\nabla} \phi = \frac{\partial \phi}{\partial z_1} \vec{j}_1 + \frac{\partial \phi}{\partial z_2} \vec{j}_2 + \frac{\partial \phi}{\partial z_3} \vec{j}_3 \quad (10.36)$$

The gradient can be expressed in the orthogonal curvilinear coordinates  $(y_1, y_2, y_3)$  using

$$\begin{aligned} \vec{\nabla} \phi &= \frac{\partial \phi}{\partial y_1} \frac{\partial y_1}{\partial z_1} \vec{j}_1 + \frac{\partial \phi}{\partial y_2} \frac{\partial y_2}{\partial z_2} \vec{j}_2 + \frac{\partial \phi}{\partial y_3} \frac{\partial y_3}{\partial z_3} \vec{j}_3 \\ &= \frac{1}{h_1} \frac{\partial \phi}{\partial y_1} \vec{j}_1 + \frac{1}{h_2} \frac{\partial \phi}{\partial y_2} \vec{j}_2 + \frac{1}{h_3} \frac{\partial \phi}{\partial y_3} \vec{j}_3 \end{aligned} \quad (10.37)$$

and we may conclude that the nabla vector can be expressed in orthogonal curvilinear coordinates as

$$\vec{\nabla} = \frac{\vec{j}_1}{h_1} \frac{\partial}{\partial y_1} + \frac{\vec{j}_2}{h_2} \frac{\partial}{\partial y_2} + \frac{\vec{j}_3}{h_3} \frac{\partial}{\partial y_3} \quad (10.38)$$

The divergence at the point  $P$  of a vector

$$\vec{u} = u_1 \vec{j}_1 + u_2 \vec{j}_2 + u_3 \vec{j}_3 \quad (10.39)$$

is

$$\vec{\nabla} \cdot \vec{u} = \vec{\nabla} \cdot \sum_{i=1}^3 u_i \vec{j}_i = \sum_{i=1}^3 \left[ (\vec{\nabla} u_i) \cdot \vec{j}_i + u_i (\vec{\nabla} \cdot \vec{j}_i) \right] \quad (10.40)$$

while the curl is

$$\vec{\nabla} \times \vec{u} = \vec{\nabla} \times \sum_{i=1}^3 u_i \vec{j}_i = \sum_{i=1}^3 \left[ (\vec{\nabla} u_i) \times \vec{j}_i + u_i (\vec{\nabla} \times \vec{j}_i) \right] \quad (10.41)$$

Here we have used the identities

$$\vec{\nabla} \cdot (\phi \vec{a}) = (\vec{\nabla} \phi) \cdot \vec{a} + \phi (\vec{\nabla} \cdot \vec{a}) \quad (10.42)$$

$$\vec{\nabla} \times (\phi \vec{a}) = (\vec{\nabla} \phi) \times \vec{a} + \phi (\vec{\nabla} \times \vec{a}) \quad (10.43)$$

To compute the expressions for the divergence and the curl we need some intermediate results, namely expressions for the divergence and curl of the unit vectors  $\vec{j}_i$ . We do this by introducing a potential given by  $\phi = y_1$  with gradient

$$\vec{\nabla} y_1 = \frac{\vec{j}_1}{h_1} \quad (10.44)$$

Then from the identity  $\vec{\nabla} \times (\vec{\nabla} \phi) = (\vec{\nabla} \times \vec{\nabla}) \phi = \vec{0}$  we get

$$\vec{0} = \vec{\nabla} \times (\vec{\nabla} y_1) = \vec{\nabla} \times \left( \frac{\vec{j}_1}{h_1} \right) = \vec{\nabla} \left( \frac{1}{h_1} \right) \times \vec{j}_1 + \frac{1}{h_1} (\vec{\nabla} \times \vec{j}_1) \quad (10.45)$$

which implies that the curl of the unit vector  $\vec{j}_1$  is

$$\begin{aligned}\vec{\nabla} \times \vec{j}_1 &= h_1 \vec{\nabla} \left( \frac{1}{h_1} \right) \times \vec{j}_1 \\ &= h_1 \left( -\frac{1}{h_1^2} \right) \left( \frac{1}{h_1} \frac{\partial h_1}{\partial y_1} \vec{j}_1 + \frac{1}{h_2} \frac{\partial h_1}{\partial y_2} \vec{j}_2 + \frac{1}{h_3} \frac{\partial h_1}{\partial y_3} \vec{j}_3 \right) \times \vec{j}_1 \\ &= \frac{\vec{j}_2}{h_1 h_3} \frac{\partial h_1}{\partial y_3} - \frac{\vec{j}_3}{h_1 h_2} \frac{\partial h_1}{\partial y_2}\end{aligned}\quad (10.46)$$

The curl of  $\vec{j}_2$  and  $\vec{j}_3$  is found in the same way, and following expressions for the curl of the unit vectors result:

$$\vec{\nabla} \times \vec{j}_1 = \frac{\vec{j}_2}{h_1 h_3} \frac{\partial h_1}{\partial y_3} - \frac{\vec{j}_3}{h_1 h_2} \frac{\partial h_1}{\partial y_2} \quad (10.47)$$

$$\vec{\nabla} \times \vec{j}_2 = \frac{\vec{j}_3}{h_2 h_1} \frac{\partial h_2}{\partial y_1} - \frac{\vec{j}_1}{h_2 h_3} \frac{\partial h_2}{\partial y_3} \quad (10.48)$$

$$\vec{\nabla} \times \vec{j}_3 = \frac{\vec{j}_1}{h_3 h_2} \frac{\partial h_3}{\partial y_2} - \frac{\vec{j}_2}{h_3 h_1} \frac{\partial h_3}{\partial y_1} \quad (10.49)$$

The divergence of the unit vector  $\vec{j}_1$  is found from

$$\begin{aligned}\vec{\nabla} \cdot \vec{j}_1 &= \vec{\nabla} \cdot (\vec{j}_2 \times \vec{j}_3) \\ &= (\vec{\nabla} \times \vec{j}_2) \cdot \vec{j}_3 - \vec{j}_2 \cdot (\vec{\nabla} \times \vec{j}_3) \\ &= \frac{1}{h_2 h_1} \frac{\partial h_2}{\partial y_1} - \frac{1}{h_3 h_1} \frac{\partial h_3}{\partial y_1} \\ &= \frac{1}{h_1 h_2 h_3} \frac{\partial (h_2 h_3)}{\partial y_1}\end{aligned}\quad (10.50)$$

The divergence of  $\vec{j}_2$  and  $\vec{j}_3$  is found in the same way, and we can conclude that

$$\vec{\nabla} \cdot \vec{j}_1 = \frac{1}{h_1 h_2 h_3} \frac{\partial (h_2 h_3)}{\partial y_1} \quad (10.51)$$

$$\vec{\nabla} \cdot \vec{j}_2 = \frac{1}{h_1 h_2 h_3} \frac{\partial (h_3 h_1)}{\partial y_2} \quad (10.52)$$

$$\vec{\nabla} \cdot \vec{j}_3 = \frac{1}{h_1 h_2 h_3} \frac{\partial (h_1 h_2)}{\partial y_3} \quad (10.53)$$

With these results it is a straightforward, although it is a time-consuming exercise, to verify that the divergence of a vector in orthogonal curvilinear coordinates is

$$\vec{\nabla} \cdot \vec{u} = \frac{1}{h_1 h_2 h_3} \left[ \frac{\partial}{\partial y_1} (u_1 h_2 h_3) + \frac{\partial}{\partial y_2} (u_2 h_3 h_1) + \frac{\partial}{\partial y_3} (u_3 h_1 h_2) \right] \quad (10.54)$$

while the curl is

$$\vec{\nabla} \times \vec{u} = \frac{1}{h_1 h_2 h_3} \left| \begin{array}{ccc} h_1 \vec{j}_1 & h_2 \vec{j}_2 & h_3 \vec{j}_3 \\ \frac{\partial}{\partial y_1} & \frac{\partial}{\partial y_2} & \frac{\partial}{\partial y_3} \\ h_1 u_1 & h_2 u_2 & h_3 u_3 \end{array} \right| \quad (10.55)$$

We also present the result for the Laplacian  $\vec{\nabla}^2 \phi$  of a scalar field  $\phi$ , which is found by letting  $\vec{u} = \vec{\nabla} \phi$ . Then  $\vec{\nabla}^2 \phi = \vec{\nabla} \cdot \vec{u}$ , and it is seen that

$$\vec{\nabla}^2 \phi = \frac{1}{h_1 h_2 h_3} \left[ \frac{\partial}{\partial y_1} \left( \frac{\partial \phi}{\partial y_1} \frac{h_2 h_3}{h_1} \right) + \frac{\partial}{\partial y_2} \left( \frac{\partial \phi}{\partial y_2} \frac{h_3 h_1}{h_2} \right) + \frac{\partial}{\partial y_3} \left( \frac{\partial \phi}{\partial y_3} \frac{h_1 h_2}{h_3} \right) \right] \quad (10.56)$$

**Example 156** For spherical coordinates  $(r, \phi, \theta)$  the orthogonal unit vectors  $\vec{j}_r, \vec{j}_\phi, \vec{j}_\theta$  form a right handed system, and we have  $h_r = 1$ ,  $h_\phi = r$  and  $h_\theta = r \sin \phi$ . The vector  $\vec{u}$  is written

$$\vec{u} = u_r \vec{j}_r + u_\phi \vec{j}_\phi + u_\theta \vec{j}_\theta \quad (10.57)$$

while the gradient of a scalar field  $\psi$  is

$$\vec{\nabla} \psi = \frac{\partial \psi}{\partial r} \vec{j}_r + \frac{1}{r} \frac{\partial \psi}{\partial \phi} \vec{j}_\phi + \frac{\vec{j}_\theta}{r \sin \phi} \frac{\partial \psi}{\partial \theta} \quad (10.58)$$

We find that the divergence of  $\vec{u}$  is

$$\vec{\nabla} \cdot \vec{u} = \frac{1}{r^2} \frac{\partial (u_r r^2)}{\partial r} + \frac{1}{r \sin \phi} \frac{\partial (u_\phi \sin \phi)}{\partial \phi} + \frac{1}{r \sin \phi} \frac{\partial u_\theta}{\partial \theta} \quad (10.59)$$

while the curl is

$$\vec{\nabla} \times \vec{u} = \frac{1}{r^2 \sin \phi} \begin{vmatrix} \vec{j}_r & r \vec{j}_\phi & r \sin \phi \vec{j}_\theta \\ \frac{\partial}{\partial r} & \frac{\partial}{\partial \phi} & \frac{\partial}{\partial \theta} \\ u_r & r u_\phi & r \sin \phi u_\theta \end{vmatrix} \quad (10.60)$$

### 10.3.2 Cylindrical coordinates

For cylindrical coordinates  $(r, \theta, z)$  the orthogonal unit vectors  $\vec{j}_r, \vec{j}_\theta, \vec{j}_z$  form a right handed system, and we have  $h_r = 1$ ,  $h_\theta = r$  and  $h_z = 1$ . The vector  $\vec{u}$  is written

$$\vec{u} = u_r \vec{j}_r + u_\theta \vec{j}_\theta + u_z \vec{j}_z \quad (10.61)$$

the expression for  $\vec{\nabla}$  is

$$\vec{\nabla} = \vec{j}_r \frac{\partial}{\partial r} + \frac{\vec{j}_\theta}{r} \frac{\partial}{\partial \theta} + \vec{j}_z \frac{\partial}{\partial z} \quad (10.62)$$

The divergence of the vector  $\vec{u}$  is found to be

$$\vec{\nabla} \cdot \vec{u} = \frac{1}{r} \frac{\partial (u_r r)}{\partial r} + \frac{1}{r} \frac{\partial u_\theta}{\partial \theta} + \frac{\partial u_z}{\partial z} \quad (10.63)$$

while the curl is

$$\vec{\nabla} \times \vec{u} = \frac{1}{r} \begin{vmatrix} \vec{j}_r & r \vec{j}_\theta & \vec{j}_z \\ \frac{\partial}{\partial r} & \frac{\partial}{\partial \theta} & \frac{\partial}{\partial z} \\ u_r & r u_\theta & u_z \end{vmatrix} \quad (10.64)$$

The Laplacian of the scalar field  $\psi$  is

$$\vec{\nabla}^2 \psi = \frac{1}{r} \left[ \frac{\partial}{\partial r} \left( \frac{\partial \psi}{\partial r} r \right) + \frac{\partial}{\partial \theta} \left( \frac{\partial \psi}{\partial \theta} \frac{1}{r} \right) + \frac{\partial}{\partial z} \left( \frac{\partial \psi}{\partial z} r \right) \right] \quad (10.65)$$

$$= \frac{\partial^2 \psi}{\partial r^2} + \frac{1}{r^2} \frac{\partial^2 \psi}{\partial \theta^2} + \frac{\partial^2 \psi}{\partial z^2} + \frac{1}{r} \frac{\partial \psi}{\partial r} \quad (10.66)$$

The velocity  $\vec{v}$  may also be written

$$\vec{v} = v_1 \vec{i}_1 + v_2 \vec{i}_2 + v_3 \vec{i}_3 \quad (10.67)$$

$$= v_r \vec{j}_r + v_\theta \vec{j}_\theta + v_z \vec{j}_z \quad (10.68)$$

where the unit vectors  $\vec{i}_1, \vec{i}_2, \vec{i}_3$  of the Cartesian system  $(x_1, x_2, x_3)$  are constant vectors, while the unit vectors  $\vec{j}_r$  and  $\vec{j}_\theta$  of the cylindrical coordinate system changes as the particle moves. The material derivative is

$$\frac{D\vec{v}}{Dt} = \frac{Dv_1}{Dt} \vec{i}_1 + \frac{Dv_2}{Dt} \vec{i}_2 + \frac{Dv_3}{Dt} \vec{i}_3 \quad (10.69)$$

$$= \frac{Dv_r}{Dt} \vec{j}_r + \frac{Dv_\theta}{Dt} \vec{j}_\theta + \frac{Dv_z}{Dt} \vec{j}_z + v_r \frac{D\vec{j}_r}{Dt} + v_\theta \frac{D\vec{j}_\theta}{Dt} \quad (10.70)$$

where the time derivatives of the unit vectors are

$$\frac{D\vec{j}_r}{Dt} = \frac{\partial \vec{j}_r}{\partial t} + \frac{v_\theta}{r} \frac{\partial \vec{j}_r}{\partial \theta} = \frac{v_\theta}{r} \vec{j}_\theta \quad (10.71)$$

$$\frac{D\vec{j}_\theta}{Dt} = \frac{\partial \vec{j}_\theta}{\partial t} + \frac{v_\theta}{r} \frac{\partial \vec{j}_\theta}{\partial \theta} = -\frac{v_\theta}{r} \vec{j}_r \quad (10.72)$$

$$\frac{\partial \vec{j}_r}{\partial \theta} = \vec{j}_\theta, \quad \frac{\partial \vec{j}_\theta}{\partial \theta} = -\vec{j}_r \quad (10.73)$$

This gives the following expressions for the material derivative of the vector  $\vec{v}$ :

$$\frac{D\vec{v}}{Dt} = \left( \frac{Dv_r}{Dt} - \frac{v_\theta^2}{r} \right) \vec{j}_r + \left( \frac{Dv_\theta}{Dt} + \frac{v_r v_\theta}{r} \right) \vec{j}_\theta + \frac{Dv_z}{Dt} \vec{j}_z \quad (10.74)$$

and for the scalar field  $\psi$ :

$$\frac{D\psi}{Dt} = \frac{\partial \psi}{\partial t} + \vec{v} \cdot \vec{\nabla} \psi = \frac{\partial \psi}{\partial t} + v_r \frac{\partial \psi}{\partial r} + \frac{v_\theta}{r} \frac{\partial \psi}{\partial \theta} + v_z \frac{\partial \psi}{\partial z} \quad (10.75)$$

The Laplacian of the velocity is found from

$$\vec{\nabla}^2 \vec{v} = (\vec{\nabla}^2 v_1) \vec{i}_1 + (\vec{\nabla}^2 v_2) \vec{i}_2 + (\vec{\nabla}^2 v_3) \vec{i}_3 \quad (10.76)$$

$$= \vec{\nabla}^2 (v_r \vec{j}_r) + \vec{\nabla}^2 (v_\theta \vec{j}_\theta) + \vec{\nabla}^2 (v_z \vec{j}_z) \quad (10.77)$$

To proceed we need the Laplacian of the components in the cylindrical coordinate system, which are given by

$$\begin{aligned} \vec{\nabla}^2 (v_r \vec{j}_r) &= \frac{\partial^2 v_r}{\partial r^2} \vec{j}_r + \frac{1}{r^2} \left( \frac{\partial^2 v_r}{\partial \theta^2} \vec{j}_r + 2 \frac{\partial v_r}{\partial \theta} \frac{\partial \vec{j}_r}{\partial \theta} + v_r \frac{\partial^2 \vec{j}_r}{\partial \theta^2} \right) + \frac{\partial^2 v_r}{\partial z^2} \vec{j}_r + \frac{1}{r} \frac{\partial v_r}{\partial r} \vec{j}_r \\ &= \left( \frac{\partial^2 v_r}{\partial r^2} + \frac{1}{r^2} \frac{\partial^2 v_r}{\partial \theta^2} + \frac{\partial^2 v_r}{\partial z^2} + \frac{1}{r} \frac{\partial v_r}{\partial r} - \frac{v_r}{r^2} \right) \vec{j}_r + \frac{2}{r^2} \frac{\partial v_r}{\partial \theta} \vec{j}_\theta \end{aligned} \quad (10.78)$$

$$\begin{aligned} \vec{\nabla}^2 (v_\theta \vec{j}_\theta) &= \frac{\partial^2 v_\theta}{\partial r^2} \vec{j}_\theta + \frac{1}{r^2} \left( \frac{\partial^2 v_\theta}{\partial \theta^2} \vec{j}_\theta + 2 \frac{\partial v_\theta}{\partial \theta} \frac{\partial \vec{j}_\theta}{\partial \theta} + v_\theta \frac{\partial^2 \vec{j}_\theta}{\partial \theta^2} \right) \vec{j}_\theta \\ &= \left( \frac{\partial^2 v_\theta}{\partial r^2} + \frac{1}{r^2} \frac{\partial^2 v_\theta}{\partial \theta^2} + \frac{\partial^2 v_\theta}{\partial z^2} + \frac{1}{r} \frac{\partial v_\theta}{\partial r} - \frac{v_\theta}{r^2} \right) \vec{j}_\theta - \frac{2}{r^2} \frac{\partial v_\theta}{\partial \theta} \vec{j}_r \end{aligned} \quad (10.79)$$

$$\vec{\nabla}^2 (v_z \vec{j}_z) = \frac{\partial^2 v_z}{\partial r^2} \vec{j}_z + \frac{1}{r^2} \frac{\partial^2 v_z}{\partial \theta^2} \vec{j}_z + \frac{\partial^2 v_z}{\partial z^2} \vec{j}_z + \frac{1}{r} \frac{\partial v_z}{\partial r} \vec{j}_z$$

This leads to the following expression for the Laplacian of the velocity in cylindrical coordinates:

$$\begin{aligned}\vec{\nabla}^2 \vec{v} = & \left( \frac{\partial^2 v_r}{\partial r^2} + \frac{1}{r^2} \frac{\partial^2 v_r}{\partial \theta^2} + \frac{\partial^2 v_r}{\partial z^2} + \frac{1}{r} \frac{\partial v_r}{\partial r} - \frac{2}{r^2} \frac{\partial v_\theta}{\partial \theta} - \frac{v_r}{r^2} \right) \vec{j}_r \\ & + \left( \frac{\partial^2 v_\theta}{\partial r^2} + \frac{1}{r^2} \frac{\partial^2 v_\theta}{\partial \theta^2} + \frac{\partial^2 v_\theta}{\partial z^2} + \frac{1}{r} \frac{\partial v_\theta}{\partial r} + \frac{2}{r^2} \frac{\partial v_r}{\partial \theta} - \frac{v_\theta}{r^2} \right) \vec{j}_\theta \\ & + \left( \frac{\partial^2 v_z}{\partial r^2} + \frac{1}{r^2} \frac{\partial^2 v_z}{\partial \theta^2} + \frac{\partial^2 v_z}{\partial z^2} + \frac{1}{r} \frac{\partial v_z}{\partial r} \right) \vec{j}_z\end{aligned}\quad (10.80)$$

## 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

$$\iiint_{V_c(t)} \phi(\mathbf{x}, t) dV \quad (10.81)$$

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:

$$\frac{d}{dt} \int_{a(t)}^{b(t)} f(x, t) dx = \int_{a(t)}^{b(t)} \frac{\partial f(x, t)}{\partial t} dx + f(b, t) \frac{db}{dt} - f(a, t) \frac{da}{dt}. \quad (10.82)$$

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

$$\frac{d}{dt} \iiint_{V_c(t)} \phi(\mathbf{x}, t) dV = \iiint_{V_c} \frac{\partial \phi(\mathbf{x}, t)}{\partial t} dV + \iint_{\partial V_c} \phi \vec{v}_c \cdot \vec{n} dA \quad (10.83)$$

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

$$\frac{d}{dt} \iiint_{V_m(t)} \phi(\mathbf{x}, t) dV = \iiint_{V_m(t)} \frac{\partial \phi(\mathbf{x}, t)}{\partial t} dV + \iint_{\partial V_m(t)} \phi \vec{v} \cdot \vec{n} dA \quad (10.84)$$

To avoid the need to explain whether the volume is material or not, we define the notation

$$\frac{D}{Dt} \iiint_V \phi(\mathbf{x}, t) dV := \iiint_V \frac{\partial \phi(\mathbf{x}, t)}{\partial t} dV + \iint_{\partial V} \phi \vec{v} \cdot \vec{n} dA \quad (10.85)$$

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

$$\frac{D}{Dt} \iiint_V \phi(\mathbf{x}, t) dV = \iiint_V \left[ \frac{D\phi(\mathbf{x}, t)}{Dt} + \phi (\vec{\nabla} \cdot \vec{v}) \right] dV \quad (10.86)$$

and in divergence form as

$$\frac{D}{Dt} \iiint_V \phi(\mathbf{x}, t) dV = \iiint_V \left[ \frac{\partial \phi(\mathbf{x}, t)}{\partial t} + \vec{\nabla}(\phi \vec{v}) \right] dV \quad (10.87)$$

### 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

$$\frac{d}{dt} \iiint_{V_c} \phi(\mathbf{x}, t) dV = \frac{D}{Dt} \iiint_{V_c} \phi(\mathbf{x}, t) dV - \iint_{\partial V_c} \phi (\vec{v} - \vec{v}_c) \cdot \vec{n} dA \quad (10.88)$$

If the volume  $V_f$  is fixed in spatial coordinates, we get

$$\frac{d}{dt} \iiint_{V_f} \phi(\mathbf{x}, t) dV = \iiint_{V_f} \frac{\partial \phi(\mathbf{x}, t)}{\partial t} dV, \quad V_f \text{ is fixed.} \quad (10.89)$$

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

$$\iiint_{V_f} \frac{\partial \phi(\mathbf{x}, t)}{\partial t} dV = \frac{D}{Dt} \iiint_{V_f} \phi(\mathbf{x}, t) dV - \iint_{\partial V_f} \phi \vec{v} \cdot \vec{n} dA \quad (10.90)$$



# 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

$$m = \iiint_{V_m(t)} \rho dV \quad (11.1)$$

where  $\rho$  is the fluid density. This means that principle of mass conservation can be expressed in the form

$$\frac{D}{Dt} \iiint_V \rho dV = 0 \quad (11.2)$$

Then Reynolds' transport theorem with  $\phi = \rho$  leads to

$$\iiint_V \left[ \frac{D\rho}{Dt} + \rho(\vec{\nabla} \cdot \vec{v}) \right] dV = 0 \quad (11.3)$$

in material form and

$$\iiint_V \left[ \frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot (\rho \vec{v}) \right] dV = 0 \quad (11.4)$$

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

$$\underbrace{\frac{D\rho}{Dt}}_{\substack{\text{rate of change of} \\ \text{density in material} \\ \text{volume element}}} + \underbrace{\rho(\vec{\nabla} \cdot \vec{v})}_{\substack{\text{rate of change of density} \\ \text{due to divergence of} \\ \text{material volume element}}} = 0 \quad (11.5)$$

rate of change of density in material volume element      rate of change of density due to divergence of material volume element

and in divergence form,

$$\underbrace{\frac{\partial \rho}{\partial t}}_{\substack{\text{rate of change of} \\ \text{density in spatial} \\ \text{volume element}}} + \underbrace{\vec{\nabla} \cdot (\rho \vec{v})}_{\substack{\text{rate of change of density} \\ \text{due to convection out of} \\ \text{spatial volume element}}} = 0 \quad (11.6)$$

**Example 157** It is possible to derive the divergence form of the continuity equation from the material form using

$$\vec{\nabla} \cdot (\rho \vec{v}) = (\vec{\nabla} \rho) \cdot \vec{v} + \rho (\vec{\nabla} \cdot \vec{v}) \quad (11.7)$$

and the definition of the material derivative.

### 11.1.2 Integral form

For a fixed volume  $V_f$  we find from (10.90) that

$$\underbrace{\frac{d}{dt} \iiint_{V_f} \rho dV}_{\substack{\text{rate of change} \\ \text{of mass in} \\ \text{fixed volume}}} = - \underbrace{\iint_{\partial V_c} \rho \vec{v} \cdot \vec{n} dA}_{\substack{\text{net increase of} \\ \text{mass by} \\ \text{convection}}} \quad (11.8)$$

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  $\partial V_c$  of the volume:

$$\frac{d}{dt} \iiint_{V_c} \rho dV = \frac{D}{Dt} \iiint_{V_c} \rho dV - \iint_{\partial V_c} \rho (\vec{v} - \vec{v}_c) \cdot \vec{n} dA \quad (11.9)$$

Using the principle of mass conservation as expressed in (11.2) we arrive at the result

$$\frac{d}{dt} \iiint_{V_c} \rho dV = - \iint_{\partial V_c} \rho (\vec{v} - \vec{v}_c) \cdot \vec{n} dA \quad (11.10)$$

### 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  $\rho$  is the same all over the control volume. The fluid is assumed to be compressible with bulk modulus  $\beta$  so that

$$\frac{d\rho}{\rho} = \frac{dp}{\beta} \quad (11.11)$$

Then from (10.88) we have the mass balance in the form

$$\underbrace{\frac{d}{dt} \iiint_{V_c} \rho dV}_{\substack{\text{rate of change} \\ \text{of mass in} \\ \text{control volume}}} = \underbrace{\frac{D}{Dt} \iiint_{V_c} \rho dV}_{\substack{\text{This term equals} \\ \text{zero in view of} \\ (11.2)}} - \underbrace{\iint_{\partial V_c} \rho (\vec{v} - \vec{v}_c) \cdot \vec{n} dA}_{\substack{\text{net mass} \\ \text{flow into} \\ \text{control volume}}} \quad (11.12)$$

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

$$\frac{d}{dt}(\rho V_c) = w_1 - w_2 = \rho(q_1 - q_2) \quad (11.13)$$

Moreover, assume that the density  $\rho$  is the same all over the control volume. The fluid is assumed to be compressible with bulk modulus  $\beta$  so that

$$\frac{d\rho}{\rho} = \frac{dp}{\beta} \Rightarrow \dot{\rho} = \frac{\rho}{\beta} \dot{p} \quad (11.14)$$

Then the mass balance of a control volume  $V_c$  with compressible fluid with bulk modulus  $\beta$  can be written

$$\frac{V_c}{\beta} \dot{p} + \dot{V}_c = q_1 - q_2 \quad (11.15)$$

#### 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

$$w = - \iint_A \rho \vec{v} \cdot \vec{n} dA = - \iint_A \rho v \vec{i} \cdot (-\vec{i}) dA = \rho v A \quad (11.16)$$

If the velocity varies over the cross section of the pipe, then the mass flow is

$$w = - \iint_{\partial V_c} \rho \vec{v} \cdot \vec{n} dA = \rho \int v dA = \rho \bar{v} A \quad (11.17)$$

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  $\partial V_c$  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

$$\frac{d}{dt} m_c = \rho_1 \bar{v}_1 A - \rho_2 \bar{v}_2 A = w_1 - w_2 \quad (11.18)$$

where

$$m_c = \iiint_{V_c} \rho dV \quad (11.19)$$

is the mass inside the control volume.

**Example 158** We consider gas of density  $\rho$  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  $\rho$  is constant over the fixed volume  $V$ , while the density is  $\rho_1$  at the inlet. We assume that the velocity in the inlet pipe is in the  $x$  direction and of magnitude

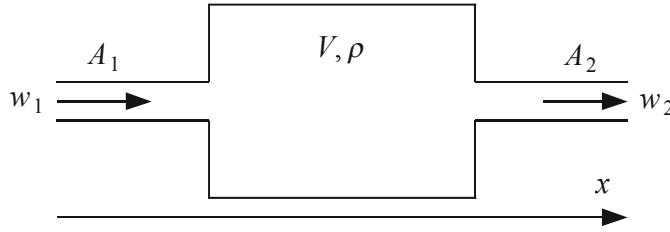


Figure 11.1: Volume  $V$  with mass flow  $w_1$  into the volume and  $w_2$  out of the volume.

$v_1$ . 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

$$V \frac{d\rho}{dt} = A_1 \rho v_1 - A_2 \rho v_2 \quad (11.20)$$

which may also be written

$$\frac{dm}{dt} = w_1 - w_2. \quad (11.21)$$

Here  $m = \rho V$  is the mass contained in the volume,  $w_1 = A_1 \rho v_1$  is the mass flow into the volume, and  $w_2 = A_2 \rho v_2$  is the mass flow out of the volume.

**Example 159** Water of constant density  $\rho$  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

$$\frac{d}{dt} (\rho A h) = w_1 - w_2 \quad (11.22)$$

which can be written

$$\dot{h} = \frac{1}{\rho A} (w_1 - w_2) \quad (11.23)$$

### 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

$$\frac{D}{Dt} \iiint_V \rho \phi dV = \iiint_V \left[ \frac{\partial(\rho\phi)}{\partial t} + \vec{\nabla} \cdot (\rho\phi \vec{v}) \right] dV \quad (11.24)$$

Then it is used that the material form (10.86) of Reynolds' transport theorem gives

$$\begin{aligned} \frac{D}{Dt} \iiint_V \rho \phi dV &= \iiint_V \left[ \frac{D(\rho\phi)}{Dt} + \rho\phi(\vec{\nabla} \cdot \vec{v}) \right] dV \\ &= \iiint_V \left\{ \rho \frac{D\phi}{Dt} + \phi \left[ \frac{D\rho}{Dt} + \rho(\vec{\nabla} \cdot \vec{v}) \right] \right\} dV \end{aligned} \quad (11.25)$$

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

$$\frac{D}{Dt} \iiint_V \rho \phi dV = \iiint_V \rho \frac{D\phi}{Dt} dV \quad (11.26)$$

By comparing this with (11.24) and accounting for the fact that the volume  $V$  is arbitrary, it is found that

$$\rho \frac{D\phi}{Dt} = \frac{\partial(\rho\phi)}{\partial t} + \vec{\nabla} \cdot (\rho\phi \vec{v}) \quad (11.27)$$

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

$$\iiint_V \rho \frac{D\phi}{Dt} dV = \iiint_V \frac{\partial(\rho\phi)}{\partial t} dV + \iint_{\partial V} \rho\phi (\vec{v} \cdot \vec{n}) dA \quad (11.28)$$

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  $\phi$  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

$$\frac{D}{Dt} \iiint_V \rho \vec{u} dV = \iiint_V \rho \frac{D\vec{u}}{Dt} dV \quad (11.29)$$

and

$$\rho \frac{D\vec{u}}{Dt} = \frac{\partial(\rho\vec{u})}{\partial t} + \vec{\nabla} \cdot (\rho\vec{v}\vec{u}) \quad (11.30)$$

The last term in (11.30) is verified in a Cartesian coordinate system with orthogonal unit vectors  $\vec{a}_i$  with the following computation:

$$\vec{\nabla} \cdot (\rho\vec{v}\vec{u}) = \sum_k \frac{\partial}{\partial x_k} \vec{a}_k \cdot \left( \rho \sum_j v_j \vec{a}_j \sum_i u_i \vec{a}_i \right) = \sum_i \frac{\partial}{\partial x_j} (\rho v_j u_i) \vec{a}_i \quad (11.31)$$

The integral form of (11.30) is found to be

$$\iiint_V \rho \frac{D\vec{u}}{Dt} dV = \iiint_V \frac{\partial(\rho\vec{u})}{\partial t} dV + \iint_{\partial V} \rho \vec{u} (\vec{v} \cdot \vec{n}) dA \quad (11.32)$$

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

$$\frac{\partial(\rho\phi)}{\partial t} + \vec{\nabla} \cdot (\rho\phi\vec{v}) = \rho \frac{D\phi}{Dt} = \rho \left( \frac{\partial\phi}{\partial t} + \vec{v} \cdot \vec{\nabla}\phi \right) \quad (11.33)$$

$$\frac{\partial(\rho\vec{u})}{\partial t} + \vec{\nabla} \cdot (\rho\vec{v}\vec{u}) = \rho \frac{D\vec{u}}{Dt} = \rho \left( \frac{\partial\vec{u}}{\partial t} + \vec{v} \cdot \vec{\nabla}\vec{u} \right) \quad (11.34)$$

### 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

$$\frac{d}{dt}m_k = \sum_{j=1}^r \iiint_V \nu_{kj} J_j dV \quad (11.35)$$

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

$$\iiint_V \left[ \frac{\partial \rho_k}{\partial t} + \vec{\nabla} \cdot (\rho_k \vec{v}_k) \right] dV = \sum_{j=i}^r \iiint_V \nu_{kj} J_j dV \quad (11.36)$$

where  $\rho_k$  is the density of component  $k$ , and  $\vec{v}_k$  is the velocity of component  $k$ . As the volume  $V$  is arbitrary, it follows that the *continuity equation of component  $k$*  is

$$\frac{\partial \rho_k}{\partial t} + \vec{\nabla} \cdot (\rho_k \vec{v}_k) = \sum_{j=i}^r \nu_{kj} J_j \quad (11.37)$$

Since mass is conserved in each of the chemical reactions it follows that

$$\sum_{k=i}^n \nu_{kj} J_j = 0 \quad (11.38)$$

Then, by adding the continuity equations of all components the continuity equation

$$\frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot (\rho \vec{v}) = 0 \quad (11.39)$$

where  $\rho$  is the *total density*

$$\rho = \sum_{k=1}^n \rho_k \quad (11.40)$$

and  $\vec{v}$  is the *barycentric velocity*, which is defined as the velocity of the center of mass

$$\vec{v} := \sum_{k=1}^n \frac{\rho_k \vec{v}_k}{\rho} \quad (11.41)$$

Define the *barycentric material derivative* by

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \vec{v} \cdot \vec{\nabla} \quad (11.42)$$

where  $\mathbf{v}$  is the barycentric velocity. Insertion into the continuity equation (11.37) for component  $k$  gives

$$\frac{D\rho_k}{Dt} - \vec{v} \cdot \vec{\nabla} \rho_k + \vec{\nabla} \cdot (\rho_k \vec{v}_k) = \sum_{j=i}^r \nu_{kj} J_j \quad (11.43)$$

The last term on the left side is expanded to give

$$\frac{D\rho_k}{Dt} - \vec{v} \cdot \vec{\nabla} \rho_k + \vec{\nabla} \cdot (\rho_k \vec{v}) + \vec{\nabla} \cdot [\rho_k (\vec{v}_k - \vec{v})] = \sum_{j=i}^r \nu_{kj} J_j \quad (11.44)$$

By defining the *diffusion flow* of component  $k$  as

$$\vec{j}_k = \rho_k (\vec{v}_k - \vec{v}) \quad (11.45)$$

and, accounting for (10.15), we find the following result:

The continuity equation for component  $k$  is found to be

$$\frac{D\rho_k}{Dt} = -\rho_k (\vec{\nabla} \cdot \vec{v}) - \vec{\nabla} \cdot \vec{j}_k + \sum_{j=i}^r \nu_{kj} J_j \quad (11.46)$$

while from (11.39) and (11.42) the continuity equation for the total density is

$$\frac{D\rho}{Dt} + \rho (\vec{\nabla} \cdot \vec{v}) = 0 \quad (11.47)$$

**Example 160** In terms of mass fractions

$$c_k = \frac{\rho_k}{\rho} \quad (11.48)$$

the continuity equation for component  $k$  becomes

$$\rho \frac{Dc_k}{Dt} = -\vec{\nabla} \cdot \vec{j}_k + \sum_{j=i}^r \nu_{kj} J_j \quad (11.49)$$

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  $\rho$  and velocity  $\vec{v}$ . The momentum of the volume element is  $\rho \vec{v} dV$ . Newton's law for this set of particles is

$$\frac{D}{Dt} (\rho \vec{v} dV) = d\vec{F}. \quad (11.50)$$

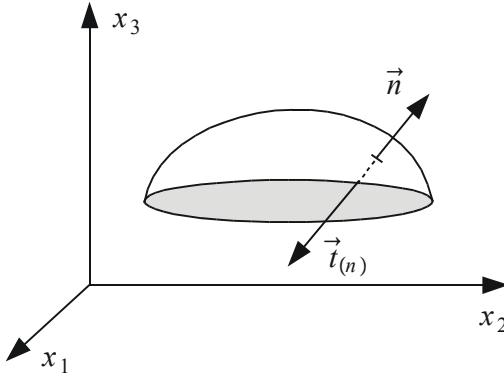


Figure 11.2: The stress vector in an inviscid fluid is parallel to the surface normal.

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  $\rho dV$  of the particles in the material volume element is constant, and it follows that

$$\frac{D}{Dt}(\rho \vec{v} dV) = \frac{D}{Dt}(\rho dV) \vec{v} + \rho \frac{D\vec{v}}{Dt} dV = \rho \frac{D\vec{v}}{Dt} dV \quad (11.51)$$

We may therefore write Newton's law in the form

$$\rho \frac{D\vec{v}}{Dt} dV = d\vec{F} \quad (11.52)$$

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

$$\iiint_V \rho \frac{D\vec{v}}{Dt} dV = \iiint_V d\vec{F} = \vec{F}^{(r)} \quad (11.53)$$

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  $\partial V$  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

$$\vec{F}^{(r)} = - \iint_{\partial V} p \vec{n} dA + \iiint_V \rho \vec{f} dV \quad (11.54)$$

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

$$\iiint_V \rho \frac{D\vec{v}}{Dt} dV = \iiint_V (-\vec{\nabla} p + \rho \vec{f}) dV \quad (11.55)$$

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

$$\rho \frac{D\vec{v}}{Dt} = -\vec{\nabla}p + \rho \vec{f} \quad (11.56)$$

Alternative formulations of Euler's equation found from (11.34) are the divergence form

$$\frac{\partial(\rho\vec{v})}{\partial t} + \vec{\nabla} \cdot (\rho\vec{v}\vec{v}) = -\vec{\nabla}p + \rho\vec{f} \quad (11.57)$$

and the formulation

$$\rho \frac{\partial \vec{v}}{\partial t} + \rho (\vec{v} \cdot \vec{\nabla}) \vec{v} = -\vec{\nabla}p + \rho \vec{f} \quad (11.58)$$

**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  $\rho$  is a constant, Euler's equation as given by (11.58) gives

$$\frac{\partial v}{\partial t} + v \frac{\partial v}{\partial x} = 0 \quad (11.59)$$

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

$$\frac{D}{Dt} \iiint_V \rho \vec{v} dV = \iiint_V \rho \frac{D\vec{v}}{Dt} dV \quad (11.60)$$

From (10.88) we have the following equation for a control volume  $V_c$

$$\frac{d}{dt} \iiint_{V_c} \rho \vec{v} dV = \frac{D}{Dt} \iiint_{V_c} \rho \vec{v} dV - \iint_{\partial V_c} \rho \vec{v} (\vec{v} - \vec{v}_c) \cdot \vec{n} dA \quad (11.61)$$

where  $\vec{v}$  is the velocity of the fluid and  $\vec{v}_c$  is the velocity of the surface  $\partial V_c$  of the control volume. Combining these two equations with (11.53) we get

$$\underbrace{\frac{d}{dt} \iiint_{V_c} \rho \vec{v} dV}_{\begin{array}{l} \text{rate of change} \\ \text{of momentum} \\ \text{in control} \\ \text{volume} \end{array}} = \underbrace{\vec{F}^{(r)}}_{\begin{array}{l} \text{resultant force} \\ \text{on fluid in} \\ \text{control} \\ \text{volume} \end{array}} - \underbrace{\iint_{\partial V_c} \rho \vec{v} (\vec{v} - \vec{v}_c) \cdot \vec{n} dA}_{\begin{array}{l} \text{net increase of} \\ \text{momentum} \\ \text{by convection} \end{array}} \quad (11.62)$$

**Example 162** For the system in Example 158 the momentum conservation in the  $x$  direction gives

$$\frac{d}{dt} \iiint_V v \rho dV = F + p_1 A_1 - p_2 A_2 + v_1 w_1 - v_2 w_2 \quad (11.63)$$

where  $F$  is the force in the  $x$  direction acting on the gas from the tank,  $p_1 A_1$  is the force due to pressure on the inlet, and  $p_2 A_2$  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

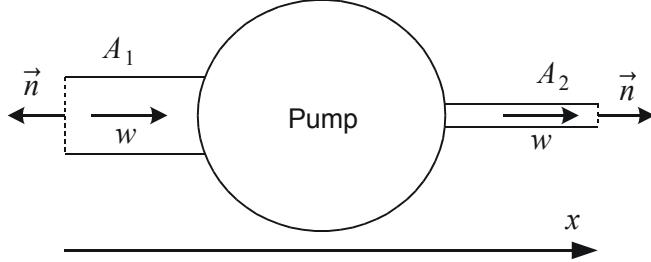


Figure 11.3: Schematic diagram of a waterjet.

We consider a waterjet (Figure 11.3) where water enters 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

$$A_1 \rho v_1 = A_2 \rho v_2 = w \quad (11.64)$$

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

$$F + v_1 A_1 \rho v_1 - v_2 \rho A_2 v_2 = 0. \quad (11.65)$$

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

$$T = - \left( 1 - \frac{A_2}{A_1} \right) w v_2 \approx -w v_2 \quad (11.66)$$

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  $T \approx -w v_2$  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  $\omega_m$ . The velocity of the conveyor belt is therefore  $v = \omega_m r$ .

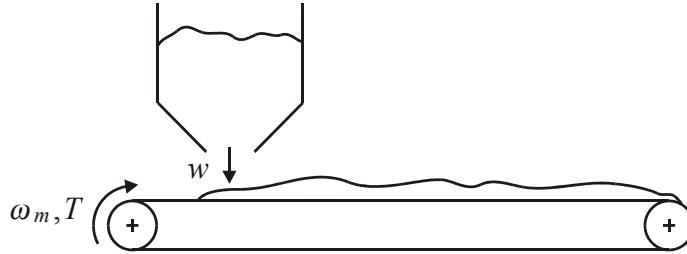


Figure 11.4: Sand of mass flow  $w$  falling down on a conveyor belt.

Here the mass  $m$  and the momentum  $p = mv$  of the sand are conserved quantities. The balance equation for the mass is

$$\frac{d}{dt}m = w - w_e, \quad (11.67)$$

while the balance equation for the momentum is

$$\frac{d}{dt}(mv) = -vw_e + F \quad (11.68)$$

Here  $F$  is the force from the conveyor belt on the sand. The equation of motion for the belt is

$$J\dot{\omega}_m = T - Fr \quad (11.69)$$

where  $J$  is the inertia experienced by the motor. The equation of motion can be expressed in terms of the velocity to give

$$\frac{J}{r^2}\dot{v} = \frac{1}{r}T - F. \quad (11.70)$$

The momentum equation gives

$$\dot{mv} + m\dot{v} = -vw_e + F \quad (11.71)$$

and insertion of the mass balance and the equation of motion gives

$$\left(m + \frac{J}{r^2}\right)\dot{v} = \frac{1}{r}T - vw \quad (11.72)$$

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

$$(\vec{v} \cdot \vec{\nabla})\vec{v} = \vec{\nabla} \left( \frac{1}{2}\vec{v}^2 \right) - \vec{v} \times (\vec{\nabla} \times \vec{v}) \quad (11.73)$$

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

$$\frac{\partial \vec{v}}{\partial t} + \vec{\nabla} \left( \frac{1}{2}\vec{v}^2 \right) - \vec{f} + \frac{1}{\rho}\vec{\nabla}p = \vec{0} \quad (11.74)$$

Suppose that the fluid is incompressible so that  $\rho$  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  $\phi$  so that  $\vec{v} = \vec{\nabla}\phi$ . Then Euler's equation can be written as the gradient equation

$$\vec{\nabla} \left[ \frac{\partial \phi}{\partial t} + \frac{1}{2}\vec{v}^2 + gz + \frac{p}{\rho} \right] = 0 \quad (11.75)$$

where it is used that  $\rho$  is a constant for incompressible fluids. This implies that

$$\frac{\partial \phi}{\partial t} + \frac{1}{2}\vec{v}^2 + \frac{p}{\rho} + gz = \text{constant} \quad (11.76)$$

which is the *irrotational Bernoulli equation*. In the stationary case we then have

$$\frac{1}{2}(\vec{v}_2^2 - \vec{v}_1^2) + \frac{(p_2 - p_1)}{\rho} + (z_2 - z_1)g = 0 \quad (11.77)$$

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*

$$\frac{\partial \phi}{\partial t} + \frac{1}{2}(\vec{\nabla}\phi) \cdot \vec{\nabla}\phi + \frac{p}{\rho} + gz = \text{constant} \quad (11.78)$$

### 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

$$\frac{\partial \vec{v}}{\partial t} + \vec{\nabla} \cdot \left( \frac{1}{2}\vec{v}^2 \right) - \vec{v} \times (\vec{\nabla} \times \vec{v}) - \vec{f} + \frac{1}{\rho}\vec{\nabla}p = \vec{0} \quad (11.79)$$

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):

$$\int \left[ \frac{\partial \vec{v}}{\partial t} + \vec{\nabla} \cdot \left( \frac{1}{2}\vec{v}^2 \right) - \vec{v} \times (\vec{\nabla} \times \vec{v}) - \vec{f} + \frac{1}{\rho}\vec{\nabla}p \right] \cdot d\vec{x} = 0 \quad (11.80)$$

where the differential  $d\vec{x}$  is parallel to the velocity and satisfies  $d\vec{x}/dt = \vec{v}$ . Then

$$\vec{v} \times (\vec{\nabla} \times \vec{v}) \cdot d\vec{x} = \mathbf{0} \quad (11.81)$$

and the integral expression becomes

$$\int \left[ \frac{\partial \vec{v}}{\partial t} + \vec{\nabla} \cdot \left( \frac{1}{2}\vec{v}^2 \right) - \vec{f} + \frac{1}{\rho}\vec{\nabla}p \right] \cdot d\vec{x} = 0 \quad (11.82)$$

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

$$\int_1^2 \frac{\partial |\vec{v}|}{\partial t} ds + \int_1^2 d \left( \frac{1}{2}\vec{v}^2 \right) + \int_1^2 gdz + \int_1^2 \frac{dp}{\rho} = 0 \quad (11.83)$$

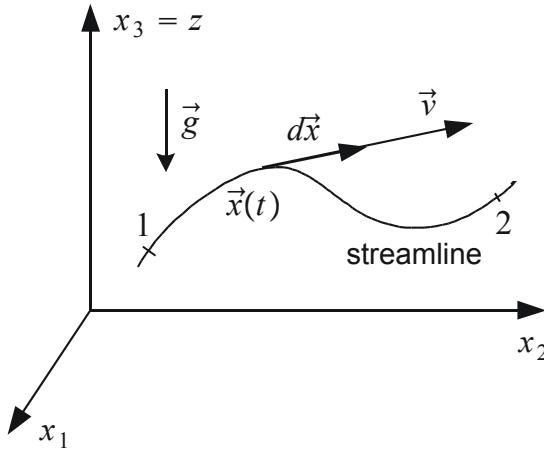


Figure 11.5: Streamline  $\mathbf{x}(t)$  with two points 1 and 2 on the streamline.

where 1 and 2 denotes two points on the same streamline. Two of the integrals are exact, and we find that

$$\int_1^2 \frac{\partial |\vec{v}|}{\partial t} ds + \frac{1}{2} (\vec{v}_2^2 - \vec{v}_1^2) + \int_1^2 \frac{dp}{\rho} + g(z_2 - z_1) = 0 \quad (11.84)$$

which is *Bernoulli's equation for frictionless flow along a streamline*. Under stationary conditions  $\partial |\vec{v}| / \partial t = 0$ , and

$$\frac{1}{2} (\vec{v}_2^2 - \vec{v}_1^2) + \int_1^2 \frac{dp}{\rho} + (z_2 - z_1) g = 0 \quad (11.85)$$

For incompressible flow  $\rho$  is a constant and

$$\frac{1}{2} (\vec{v}_2^2 - \vec{v}_1^2) + \frac{(p_2 - p_1)}{\rho} + (z_2 - z_1) g = 0, \quad 1 \text{ and } 2 \text{ on a streamline} \quad (11.86)$$

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

$$w(x, t) = \int_A \rho v dA = \rho \bar{v} A \quad (11.87)$$

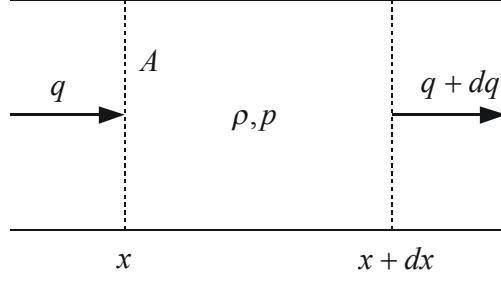


Figure 11.6: Volume element for hydraulic transmission line.

where  $\bar{v}$  is the average velocity. The mass balance is taken for the fixed differential control volume  $A dx$  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

$$A dx \frac{\partial \rho}{\partial t} = w - (w + dw) = -dw$$

Dividing by  $A dx$  we get

$$\frac{\partial \rho}{\partial t} = -\frac{1}{A} \frac{\partial w}{\partial x} \quad (11.88)$$

A change of variables from density  $\rho$  to pressure  $p$  is achieved in the mass balance using the constitutive equation  $dp = (\beta/\rho)d\rho$  where  $\beta$  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

$$\frac{\partial}{\partial t} (\rho \bar{v}) A dx = Ap - A(p + dp) + \int_A \rho v^2 dA - \int_A [\rho v^2 + d(\rho v^2)] dA - F dx \quad (11.89)$$

where  $F dx$  is the friction force. This gives

$$\frac{\partial w}{\partial t} = -A \frac{\partial p}{\partial x} - A \frac{\partial}{\partial x} \int_A \rho v^2 dA - F \quad (11.90)$$

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

$$\frac{\partial p}{\partial t} = -\frac{\beta}{\rho A} \frac{\partial w}{\partial x} \quad (11.91)$$

$$\frac{\partial w}{\partial t} = -A \frac{\partial p}{\partial x} - F \quad (11.92)$$

These equations are usually formulated in terms of the pressure  $p$  and the volumetric flow  $q$  by treating the density as a constant  $\rho_0$  so that  $w = \rho_0 q$ .

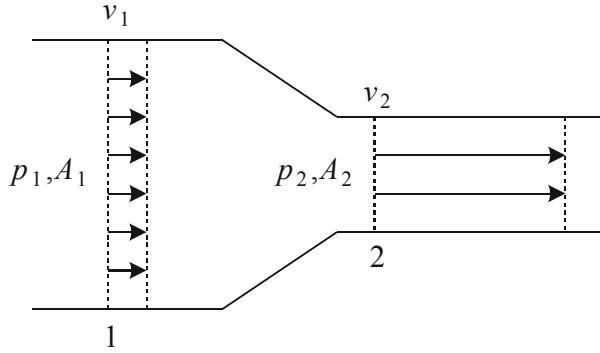


Figure 11.7: Incompressible fluid flowing through a pipe of cross section  $A_1$  with a restriction with cross section  $A_2$ .

The transmission line model linearized around  $q = 0$  and  $\rho = \rho_0$  is given by

$$\frac{\partial p}{\partial t} = -\frac{\beta}{A} \frac{\partial q}{\partial x} \quad (11.93)$$

$$\frac{\partial q}{\partial t} = -\frac{A}{\rho_0} \frac{\partial p}{\partial x} - \frac{F}{\rho_0} \quad (11.94)$$

### 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  $w_1 = \rho q_1$  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

$$q = v_1 A_1 = v_2 A_2 \quad (11.95)$$

Bernoulli's equation (11.86) gives

$$\frac{1}{2} v_1^2 + \frac{p_1}{\rho} = \frac{1}{2} v_2^2 + \frac{p_2}{\rho} \quad (11.96)$$

which gives

$$\begin{aligned} p_1 - p_2 &= \frac{\rho}{2} (v_2^2 - v_1^2) = \frac{\rho}{2} \left[ 1 - \left( \frac{A_2}{A_1} \right)^2 \right] v_2^2 \\ &= \left[ 1 - \left( \frac{A_2}{A_1} \right)^2 \right] \frac{\rho q_2^2}{2 A_2} \end{aligned} \quad (11.97)$$

This gives the following result:

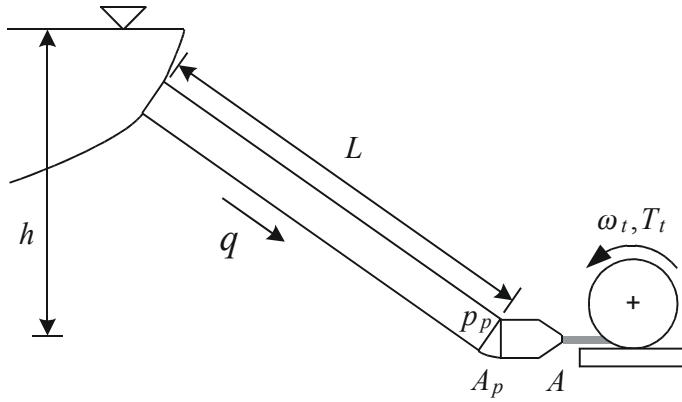


Figure 11.8:

Frictionless and incompressible flow through a restriction \$A\_2\$ in a pipe with cross section \$A\_1\$ is given by

$$q = A_2 \sqrt{\frac{2}{\rho} \frac{(p_1 - p_2)}{1 - \left(\frac{A_2}{A_1}\right)^2}} \quad (11.98)$$

If the flow is from a volume, then \$A\_1 \rightarrow \infty\$, and the expression becomes

$$q = A_2 \sqrt{\frac{2}{\rho} (p_1 - p_2)} \quad (11.99)$$

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

$$q = C_d A_2 \sqrt{\frac{2}{\rho} \frac{(p_1 - p_2)}{1 - \left(\frac{A_2}{A_1}\right)^2}} \quad (11.100)$$

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

$$J_t \dot{\omega}_t = T_t - T_L \quad (11.101)$$

where  $J_t$  is the moment of inertia of the turbine shaft,  $\omega_t$  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

$$H_{pq}(s) := \frac{-\Delta p_p}{\Delta q}(s) \quad (11.102)$$

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

$$p_p = \frac{\rho}{2} \left( \frac{q^2}{A^2} - \frac{q_0^2}{A_p^2} \right) \quad (11.103)$$

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  $\Delta q$  and rearranging we find that

$$\frac{\rho q_0^2}{A_0^3} \frac{\Delta A}{\Delta q}(s) = \frac{\rho \alpha q_0}{A_0^2} + \frac{-\Delta p_p}{\Delta q}(s) \quad (11.104)$$

The transfer function from the control input  $A$  to the flow  $q$  is then found to be given by

$$H_{qA}(s) := \frac{\Delta q}{\Delta A}(s) = \frac{q_0}{\alpha A_0} \frac{1}{1 + \frac{A_0^2}{\alpha \rho q_0} H_{pq}(s)} \quad (11.105)$$

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)

$$T_t = 2r_t \rho q(v - r_t \omega_t) = 2r_t \rho \left( \frac{q^2}{A} - qr_t \omega_t \right) \quad (11.106)$$

where  $r_t$  is the radius of the wheel. We will treat the shaft speed  $\omega_t$  as a constant in the linearization of the shaft torque. Linearization of the turbine torque equation (11.106) will then give

$$\Delta T = 2r_t \rho \frac{q_0^2}{A_0^2} \left( \beta \frac{A_0}{q_0} \Delta q - \Delta A \right) \quad (11.107)$$

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

$$\frac{\Delta T_t}{\Delta A}(s) = 2r_t \rho \frac{q_0^2}{A_0^2} \left( \beta \frac{A_0}{q_0} H_{qA}(s) - 1 \right) \quad (11.108)$$

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)

$$\frac{\Delta T_t}{\Delta A}(s) = \frac{2r_t \rho}{\gamma} \frac{q_0^2}{A_0^2} \frac{1 - \gamma \frac{A_0^2}{\alpha \rho q_0} H_{pq}}{1 + \frac{A_0^2}{\alpha \rho q_0} H_{pq}} \quad (11.109)$$

where the constant  $\gamma$  is given by

$$\gamma = \frac{1}{\frac{\beta}{\alpha} - 1} \approx \frac{1}{1 - \frac{r_t \omega_{t0} A_0}{q_0}}$$

The power on the turbine shaft is  $P = T_t \omega_t$ , 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  $\rho$ . 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

$$L\rho\dot{q} = mgh + A_p(p_0 - p_p) \quad (11.110)$$

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

$$H_{pq}(s) := \frac{-\Delta p_p}{\Delta q}(s) = \frac{\rho L s}{A_p} \quad (11.111)$$

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  $\mu$  by

$$T_r = 2 \frac{LA_0^2 q_{\max}}{\alpha q_0^2 A_p}, \quad \mu = \frac{q_0}{q_{\max}} \quad (11.112)$$

The transfer functions for the complete system is found from (11.109) to be

$$\frac{\Delta T_t}{\Delta A}(s) = \frac{2}{\gamma} \frac{r_t \rho q_0^2}{A_0^2} \frac{(1 - \gamma \mu \frac{T_r}{2}s)}{(1 + \mu \frac{T_r}{2}s)} \quad (11.113)$$

At full load with  $\mu = 1$  and  $\gamma = 2$ , the transfer function is

$$\frac{\Delta T_t}{\Delta A}(s) = \frac{r_t \rho q_0^2}{A_0^2} \frac{(1 - T_r s)}{(1 + \frac{T_r}{2}s)} \quad (11.114)$$

**Example 164** Francis or Kaplan type turbines are reaction turbines that are driven by power transfer from the water flow. The shaft torque is

$$T_{ft} = \frac{P}{\omega_t} \quad (11.115)$$

where

$$P = q\rho \left( \frac{1}{2} v^2 \right) = \frac{\rho}{2} \frac{q^3}{A^2} \quad (11.116)$$

is the power supplied to the turbine. Linearization of the power expression gives

$$\Delta P = \frac{\rho}{2} \frac{q_0^2}{A_0^2} \left( 3\Delta q - 2 \frac{q_0}{A_0} \Delta A \right) \quad (11.117)$$

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)

$$\frac{\Delta P}{\Delta A}(s) = \frac{\rho q_0^3}{A_0^3} \frac{(1 - \mu T_r s)}{(1 + \mu \frac{T_r}{2}s)} \quad (11.118)$$

where  $T_r$  and  $\mu$  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

$$H_{pq} = \frac{-\Delta p}{\Delta q}(s) = \frac{\rho c}{A_p} \tanh \frac{L}{c}s \quad (11.119)$$

Note that  $H_{pq}(s)$  tends to the incompressible solution  $\rho L s / A_p$  when  $c \rightarrow \infty$ , which corresponds to the incompressible case where  $\beta \rightarrow \infty$ . We find that when the compressibility effects of the water in the pipe is included the transfer functions to torque and power becomes

$$\frac{\Delta T_t}{\Delta A}(s) = \frac{2 r_t \rho q_0^2}{\gamma A_0^2} \frac{\left(1 - \gamma \frac{A_0^2}{\alpha q_0} \frac{c}{A_p} \tanh \frac{L}{c}s\right)}{\left(1 + \frac{A_0^2}{\alpha q_0} \frac{c}{A_p} \tanh \frac{L}{c}s\right)} \quad (11.120)$$

### 11.2.10 Example: Waterhammer

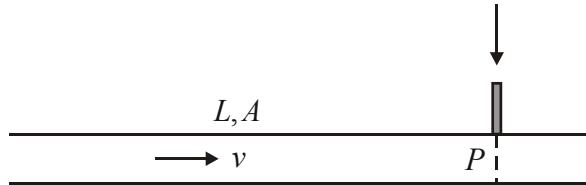


Figure 11.9: The waterhammer effect occurs when the pipe is suddenly closed at  $P$ .

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  $\rho$  flowing in a pipe of length  $L$  and cross section  $A$  will have a kinetic energy

$$K = \frac{1}{2} \rho V v^2 \quad (11.121)$$

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

$$dV = d\left(\frac{m}{\rho}\right) = -\frac{m}{\rho^2} d\rho = -\frac{m}{\rho^2} \frac{\rho}{\beta} dp = -\frac{V}{\beta} dp \quad (11.122)$$

Then it follows that if the fluid is instantaneously stopped the kinetic energy  $K$  will give an increase  $\Delta P$  due to compression, which is given by

$$\Delta P = - \int_1^2 p dV = \int_1^2 p \frac{V}{\beta} dp = \frac{1}{2} \frac{V}{\beta} (p_2^2 - p_1^2) \quad (11.123)$$

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

$$\sqrt{p_2^2 - p_1^2} = \rho c v \quad (11.124)$$

where  $c = \sqrt{\beta/\rho}$  is the sonic speed. In the case that the initial pressure  $p_1$  is small, this is approximated by

$$p_2 = \rho cv \quad (11.125)$$

**Example 165** For water  $c = 1500 \text{ m/s}$  and  $\rho = 10^3 \text{ kg/m}^3$ , 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  $5 \text{ m/s}$  gives a pressure rise of 75 bar. For hydraulic fluids  $c = 1250 \text{ m/s}$  and  $\rho = 800 \text{ kg/m}^3$  which gives  $p_2 = 10 \frac{\text{atm}}{\text{m/s}} \cdot v$  so that  $5 \text{ m/s}$  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

$$\rho \frac{D}{Dt} (\vec{r} \times \vec{v}) dV = \vec{r} \times d\vec{F} \quad (11.126)$$

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

$$\iiint_V \rho \frac{D}{Dt} (\vec{r} \times \vec{v}) dV = \frac{D}{Dt} \iiint_V \vec{r} \times \rho \vec{v} dV = \vec{N}_o \quad (11.127)$$

where

$$\vec{N}_o = \iiint_V \vec{r} \times d\vec{F} \quad (11.128)$$

is the moment about the point  $o$ .

The angular momentum equation is given by

$$\frac{D}{Dt} \iiint_V \vec{r} \times \rho \vec{v} dV = \vec{N}_o \quad (11.129)$$

For a general control volume  $V_c$  the angular momentum equation is written

$$\frac{d}{dt} \iiint_{V_c} \vec{r} \times \rho \vec{v} dV + \iint_{\partial V_c} (\vec{r} \times \rho \vec{v}) (\vec{v} - \vec{v}_c) \cdot \vec{n} dA = \vec{N}_o \quad (11.130)$$

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  $\omega$ . The

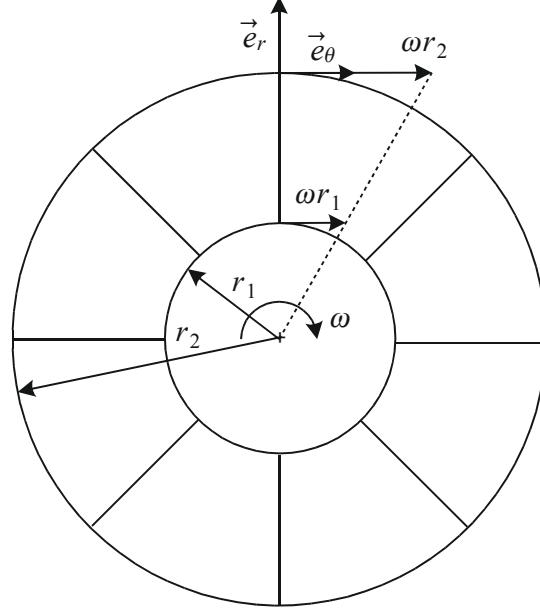


Figure 11.10: Centrifugal pump with radial blades.

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

$$w = 2\pi r_1 b v_{1r} = 2\pi r_2 b v_{2r} \quad (11.131)$$

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

$$\vec{v}_1 = \frac{w}{2\pi r_1 b} \vec{e}_r + r_1 \omega \vec{e}_\theta, \quad \vec{v}_2 = \frac{w}{2\pi r_2 b} \vec{e}_r + r_2 \omega \vec{e}_\theta \quad (11.132)$$

This gives

$$\vec{r}_1 \times \vec{v}_1 = r_1^2 \omega \vec{e}_z, \quad \vec{r}_2 \times \vec{v}_2 = r_2^2 \omega \vec{e}_z \quad (11.133)$$

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

$$\iiint_{V_c} \frac{\partial}{\partial t} (\rho r^2 \omega \vec{e}_z) dV + w (r_2^2 \omega - r_1^2 \omega) \vec{e}_z = \vec{T}_p \quad (11.134)$$

where  $\vec{T}_p = T_p \vec{e}_z$  is the load torque on the shaft. This gives

$$J_f \dot{\omega} + w\omega (r_2^2 - r_1^2) = T_p \quad (11.135)$$

where

$$J_f = \frac{\pi b \rho}{2} (r_2^4 - r_1^4) = \frac{m_f}{2} (r_1^2 + r_2^2) \quad (11.136)$$

is the moment of inertia due to the fluid, and  $m_f = \pi b \rho (r_2^2 - r_1^2)$  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

$$T_p = w\omega (r_2^2 - r_1^2) \quad (11.137)$$

The shaft power is

$$P_p = T_p \omega = w\omega^2 (r_2^2 - r_1^2) \quad (11.138)$$

### 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  $\beta_1$  and outlet angle  $\beta_2$ . The velocity  $\vec{v}_1$  at the inlet and the velocity  $\vec{v}_2$  at the blade outlet is written

$$\vec{v}_1 = v_{1r} \vec{e}_r + v_{1t} \vec{e}_\theta, \quad \vec{v}_2 = v_{2r} \vec{e}_r + v_{2t} \vec{e}_\theta \quad (11.139)$$

We get

$$\vec{r}_1 \times \vec{v}_1 = r_1 v_{1t} \vec{e}_z, \quad \vec{r}_2 \times \vec{v}_2 = r_2 v_{2t} \vec{e}_z \quad (11.140)$$

Then, proceeding as in the previous section, we get the shaft power

$$T_p = w (r_2 v_{2t} - r_1 v_{1t}) \quad (11.141)$$

The shaft power is found to be

$$P_p = T\omega = w\omega (r_2 v_{2t} - r_1 v_{1t}) \quad (11.142)$$

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  $\omega$  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

$$\tan \alpha_1 = \frac{v_{1t}}{v_{1r}}, \quad \tan \alpha_2 = \frac{v_{2t}}{v_{2r}} \quad (11.143)$$

We define  $\vec{W}_1$  and  $\vec{W}_2$  by

$$\vec{v}_1 = \vec{U}_1 + \vec{W}_1, \quad \vec{v}_2 = \vec{U}_2 + \vec{W}_2 \quad (11.144)$$

$$\vec{W}_1 = W_{1r} \vec{e}_r + W_{1t} \vec{e}_\theta, \quad \vec{W}_2 = W_{2r} \vec{e}_r + W_{2t} \vec{e}_\theta \quad (11.145)$$

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  $\beta_2$ . At design speed a design rule is to

select the inlet blade angle  $\beta_1$  so that the inlet flow will be along the blade at the inlet, so that  $\vec{W}_1$  will have direction given by  $\beta_1$ . Then

$$W_{1t} = -v_{1r}\cotan\beta_1, \quad W_{2t} = -v_{2r}\cotan\beta_2 \quad (11.146)$$

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

$$v_{1t} = 0 \Rightarrow \tan\beta_1 = \frac{v_{1r}}{U_1} \quad (11.147)$$

and the torque is found to be

$$\begin{aligned} T &= wr_2(U_2 - v_{2r}\cotan\beta_2) \\ &= wr_2\left(\omega r_2 - \frac{w}{2\pi r_2 b \rho} \cotan\beta_2\right) \end{aligned} \quad (11.148)$$

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

$$w = \rho A v \quad (11.149)$$

The equation of motion for the fluid is

$$\rho A L \dot{v} = F - F_{\text{out}} \quad (11.150)$$

where  $F_{\text{out}}$  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

$$F = \frac{\omega}{v} T = \rho A \frac{\omega}{w} T = \rho A \omega r_2 \left( \omega r_2 - \frac{Av}{2\pi r_2 b} \cotan\beta_2 \right) \quad (11.151)$$

and the equation of motion becomes

$$\rho A L \dot{v} = \rho A \omega^2 r_2^2 - v \frac{\rho A^2 \omega}{2\pi b} \cotan\beta_2 - F_{\text{out}} \quad (11.152)$$

The force consist of a term that is proportional to  $\omega^2$  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  $\beta_2 > 90^\circ$ , 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  $\beta_2 < 90^\circ$ , 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 a throttle with mass flow  $w_t(h) = C\sqrt{h}$ . The model for the system is

$$\dot{v} = -\frac{A}{2\pi b L} \omega \cotan\beta_2 v - \frac{g}{L} h + \frac{r_2^2 \omega^2}{L} \quad (11.153)$$

$$\dot{h} = \frac{A}{A_b} v - \frac{1}{A_b \rho} w_t(h) \quad (11.154)$$

where the pump velocity  $\omega^2$  is considered to be the control input. This can be achieved by velocity control of the motor driving the pump. Linearization gives

$$\dot{v} = a_{11}v + a_{12}h + b\omega^2 \quad (11.155)$$

$$\dot{h} = a_{21}v + a_{22}h \quad (11.156)$$

and the characteristic equation of the linearized system is found to be

$$\lambda^2 - (a_{11} + a_{22})\lambda - a_{12}a_{21} = 0 \quad (11.157)$$

Stability results whenever

$$a_{11} + a_{22} = -\frac{A}{2\pi bL}\omega \cot \beta_2 - \frac{1}{A_b \rho} \frac{dw_t}{dh} < 0 \quad (11.158)$$

which is the case if

$$\cot \beta_2 < -\frac{2\pi bL\omega}{\rho A A_b} \frac{dw_t}{dh} \quad (11.159)$$

This means that if the blade outlets are backswept so that  $\cot \beta_2 \geq 0$ , then the system will be stable. Forward swept blade outlets may cause instability depending on the system parameters.

**Example 167** Under stationary conditions it may be assumed that the mechanical power  $T\omega$  from the shaft is converted to power  $Fv$  supplied to the fluid, so that

$$F = \frac{\omega}{v} T \quad (11.160)$$

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 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

$$\dot{F} = \frac{1}{\alpha T_{\text{trans}}} \left( \frac{\omega}{v} T - F \right) \quad (11.161)$$

where  $T_{\text{trans}}$  can be taken to be the transport time of a fluid particle through the pump, and  $\alpha$  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

$$e = u + \frac{1}{2} \vec{v}^2 + \phi \quad (11.162)$$

is the specific energy,  $u$  is the specific internal energy,  $(1/2)\vec{v}^2$  is the specific kinetic energy, and  $\phi$  is the specific potential energy. We assume that the body forces are derived from the potential  $\phi$  in the sense that

$$\vec{\nabla} \phi = -\vec{f} \Rightarrow \frac{D\phi}{Dt} = (\vec{\nabla} \phi) \cdot \vec{v} = -\vec{f} \cdot \vec{v} \quad (11.163)$$

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

$$\frac{D}{Dt} \iiint_V \rho e dV = - \iint_{\partial V} p\vec{v} \cdot \vec{n} dA - \iint_{\partial V} \vec{j}_Q \cdot \vec{n} dA \quad (11.164)$$

The volume  $V$  is arbitrary, and it follows from the divergence theorem that

$$\underbrace{\rho \frac{D}{Dt} \left( u + \frac{1}{2} \vec{v}^2 + \phi \right)}_{\begin{array}{l} \text{rate of change} \\ \text{in internal, kinetic} \\ \text{and potential energy} \\ \text{for material} \\ \text{volume element} \end{array}} = - \underbrace{\vec{\nabla} \cdot (p\vec{v})}_{\begin{array}{l} \text{pressure work} \\ \text{on the surface of} \\ \text{the volume element} \end{array}} - \underbrace{\vec{\nabla} \cdot \vec{j}_Q}_{\text{heat conduction}} \quad (11.165)$$

The divergence form is found by changing the left hand side as follows:

$$\rho \frac{De}{Dt} = \frac{\partial}{\partial t} (\rho e) + \vec{\nabla} \cdot (\rho e \vec{v}) \quad (11.166)$$

If we leave out the potential energy, then (11.163) can be used to express the energy equation in the form

$$\underbrace{\rho \frac{D}{Dt} \left( \frac{1}{2} \vec{v}^2 + u \right)}_{\begin{array}{l} \text{rate of change} \\ \text{in internal and} \\ \text{kinetic energy} \\ \text{for material} \\ \text{volume element} \end{array}} = - \underbrace{\vec{\nabla} \cdot (p\vec{v})}_{\begin{array}{l} \text{pressure work} \\ \text{on the surface of} \\ \text{the volume element} \end{array}} - \underbrace{\vec{\nabla} \cdot \vec{j}_Q}_{\text{heat conduction}} + \underbrace{\rho \vec{v} \cdot \vec{f}}_{\begin{array}{l} \text{work of body} \\ \text{forces on volume} \\ \text{element}} \quad (11.167)$$

**Example 168** If the pressure is constant over the volume, then the pressure work can be written

$$\iint_{\partial V} p\vec{v} \cdot \vec{n} dA = p \iint_{\partial V} \vec{v} \cdot \vec{n} dA = p \frac{DV}{Dt} \quad (11.168)$$

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

$$\frac{D}{Dt} \iiint_V \rho e dV = \frac{d}{dt} \iiint_V \rho e dV + \iint_{\partial V} \rho e \vec{v} \cdot \vec{n} dA \quad (11.169)$$

Insertion of (11.164) gives the result

$$\frac{d}{dt} \iiint_V \rho e dV = - \iint_{\partial V} \rho \left( e + \frac{p}{\rho} \right) \vec{v} \cdot \vec{n} dA - \iint_{\partial V} \vec{j}_Q \cdot \vec{n} dA \quad (11.170)$$

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

$$h = u + \frac{p}{\rho} \quad (11.171)$$

Then the energy balance can be written

$$\underbrace{\frac{d}{dt} \iiint_V \rho \left( u + \frac{1}{2} \vec{v}^2 + \phi \right) dV}_{\begin{array}{l} \text{rate of change} \\ \text{of energy} \\ \text{in fixed volume} \end{array}} = - \underbrace{\iint_{\partial V} \rho \left( h + \frac{1}{2} \vec{v}^2 + \phi \right) \vec{v} \cdot \vec{n} dA}_{\begin{array}{l} \text{convected enthalpy,} \\ \text{kinetic energy and} \\ \text{potential energy}} - \underbrace{\iint_{\partial V} \vec{j}_Q \cdot \vec{n} dA}_{\begin{array}{l} \text{heat} \\ \text{conduction}} \quad (11.172)$$

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  $\rho$  and  $u$  are constants over the volume. Then the energy balance is

$$V \frac{d}{dt} (u\rho) = u_1 A_1 \rho_1 v_1 - u A_2 \rho v_2 + p_1 A_1 v_1 - p A_2 v_2 \quad (11.173)$$

which gives

$$\frac{d}{dt} E = h_1 w_1 - h w_2 \quad (11.174)$$

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

$$\dot{m}u + \dot{m}u = h_1 w_1 - h w_2. \quad (11.175)$$

Combining this with the mass balance

$$\dot{m} = w_1 - w_2 \quad (11.176)$$

we get a differential equation for the specific internal energy in the form

$$\dot{m}u = h_1 w_1 - h w_2 - u(w_1 - w_2) \quad (11.177)$$

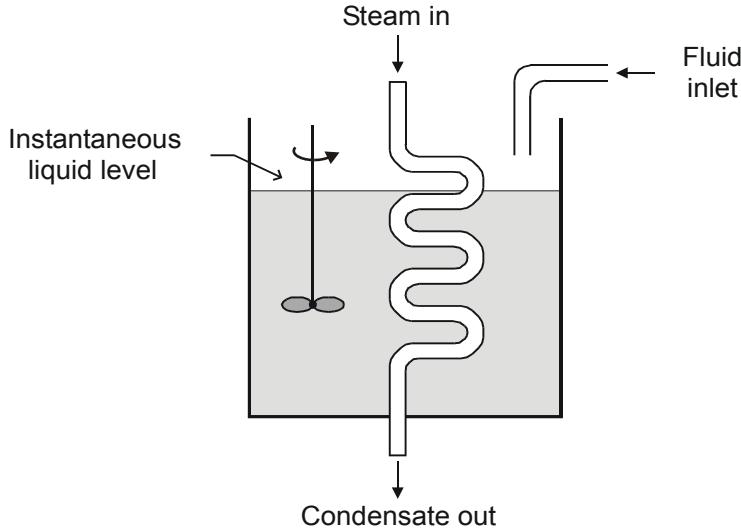


Figure 11.11: Water tank heated by a coil filled with steam.

which gives

$$\dot{u} = \frac{w_1}{m} (h_1 - u) - \frac{w_2 p}{m \rho} \quad (11.178)$$

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

$$\frac{d}{dt} (\rho V) = w \quad (11.179)$$

$$\frac{d}{dt} (\rho u V) = w u_1 + G z (T_s - T). \quad (11.180)$$

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

$$\left( \frac{d}{dt} \rho V \right) c_p T + \rho V c_p \frac{dT}{dt} = w c_p T_1 + G z (T_s - T). \quad (11.181)$$

Insertion of the mass balance in the energy balance gives

$$\rho V c_p \frac{dT}{dt} = w c_p T_1 - w c_p T + G z (T_s - T) \quad (11.182)$$

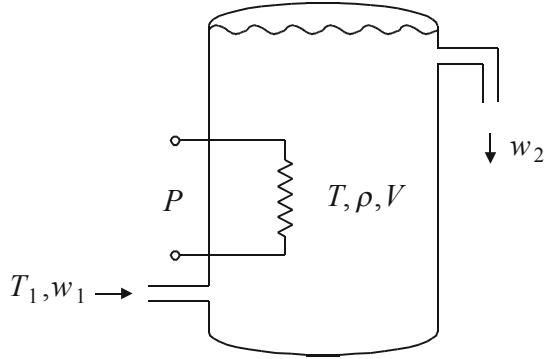


Figure 11.12: Heated tank.

and some straightforward manipulations lead to the model

$$\dot{z} = \frac{w}{\rho A} \quad (11.183)$$

$$\dot{T} = \frac{w}{\rho A z} (T_1 - T) + \frac{G}{\rho A c_p} (T_s - T). \quad (11.184)$$

**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  $\rho$  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  $mc_p T$  where  $m = \rho V$  is the mass of the liquid in the tank. The mass balance implies that

$$w_1 = w_2 = w \quad (11.185)$$

The energy balance is then

$$\frac{d}{dt} (mc_p T) = c_p w (T_1 - T) + P. \quad (11.186)$$

From the mass balance we have  $\dot{m} = 0$ , and we get

$$\dot{T} = \frac{w}{m} (T_1 - T) + \frac{P}{c_p m}. \quad (11.187)$$

### 11.4.3 General control volume

For a general control volume  $V_c$  Reynolds' transport theorem (10.88) gives

$$\frac{d}{dt} \iiint_{V_c} \rho e dV = \frac{D}{Dt} \iiint_{V_c} \rho e dV - \iint_{\partial V_c} \rho e (\vec{v} - \vec{v}_c) \cdot \vec{n} dA \quad (11.188)$$

From equation (11.164) we have the following expression for the material derivative of the energy:

$$\frac{D}{Dt} \iiint_{V_c} \rho e dV = - \iint_{\partial V_c} p \vec{v} \cdot \vec{n} dA - \iint_{\partial V_c} \vec{j}_Q \cdot \vec{n} dA \quad (11.189)$$

Combining the two equations we find that

$$\frac{d}{dt} \iiint_{V_c} \rho e dV = - \iint_{\partial V_c} \rho \left( e + \frac{p}{\rho} \right) (\vec{v} - \vec{v}_c) \cdot \vec{n} dA - \iint_{\partial V_c} p \vec{v}_c \cdot \vec{n} dA - \iint_{\partial V_c} \vec{j}_Q \cdot \vec{n} dA \quad (11.190)$$

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

$$\underbrace{\frac{d}{dt} \iiint_{V_c} \rho \left( u + \frac{1}{2} \vec{v}^2 + \phi \right) dV}_{\begin{array}{l} \text{rate of change} \\ \text{of energy} \\ \text{in control volume} \end{array}} = - \underbrace{\iint_{\partial V_c} \rho \left( h + \frac{1}{2} \vec{v}^2 + \phi \right) (\vec{v} - \vec{v}_c) \cdot \vec{n} dA}_{\begin{array}{l} \text{convected enthalpy,} \\ \text{kinetic energy and} \\ \text{potential energy}} - \underbrace{\iint_{\partial V_c} p \vec{v}_c \cdot \vec{n} dA}_{\begin{array}{l} \text{pressure work} \\ \text{due to change in} \\ \text{control volume} \end{array}} - \underbrace{\iint_{\partial V_c} \vec{j}_Q \cdot \vec{n} dA}_{\begin{array}{l} \text{heat} \\ \text{conduction} \end{array}} \quad (11.191)$$

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

$$\iint_{\partial V_c} p \vec{v}_c \cdot \vec{n} dA = p \iint_{\partial V_c} \vec{v}_c \cdot \vec{n} dA = p \dot{V}_c \quad (11.192)$$

#### 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  $u = c_p T$ , and energy flow is due to heat conduction according to the constitutive equation in the form of Fourier's law

$$\vec{j}_Q = -\alpha \vec{\nabla}(\rho c_p T) \quad (11.193)$$

where the  $\alpha$  is the thermal diffusivity in  $\text{m}^2/\text{s}$ . The energy balance is simply

$$\rho \frac{\partial u}{\partial t} = -\vec{\nabla} \cdot \vec{j}_Q \quad (11.194)$$

which in combination with Fourier's law with constant  $\alpha$  and  $\rho$  gives

$$\frac{\partial T}{\partial t} - \alpha \nabla^2 T = 0 \quad (11.195)$$

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

$$\frac{\partial T(x, t)}{\partial t} - \alpha \frac{\partial^2 T(x, t)}{\partial x^2} = 0 \quad (11.196)$$

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

$$\frac{\partial T(0, t)}{\partial x} = 0, \quad \frac{\partial T(L, t)}{\partial x} = u \quad (11.197)$$

Laplace transformation of (11.196) gives

$$\frac{\partial^2 T(x, s)}{\partial x^2} - \frac{s}{\alpha} T(x, s) = 0 \quad (11.198)$$

which has the solution

$$T(x, s) = A \cosh \left( \sqrt{\frac{s}{\alpha}} x \right) + B \sinh \left( \sqrt{\frac{s}{\alpha}} x \right) \quad (11.199)$$

with derivative

$$\frac{\partial T(x, s)}{\partial x} = A \sqrt{\frac{s}{\alpha}} \sinh \left( \sqrt{\frac{s}{\alpha}} x \right) + B \sqrt{\frac{s}{\alpha}} \cosh \left( \sqrt{\frac{s}{\alpha}} x \right) \quad (11.200)$$

The boundary condition at  $x = 0$  gives  $B = 0$ , and the boundary condition at  $x = L$  gives

$$A \sqrt{\frac{s}{\alpha}} \sinh \left( \sqrt{\frac{s}{\alpha}} L \right) = u \quad (11.201)$$

so that the temperature is given by

$$T(x, s) = \frac{\cosh \left( \sqrt{\frac{s}{\alpha}} x \right)}{\sqrt{\frac{s}{\alpha}} \sinh \left( \sqrt{\frac{s}{\alpha}} L \right)} u(s) \quad (11.202)$$

The transfer function from the heat flux to the temperature at  $x = L$  is found to be

$$\frac{T(L, s)}{u(s)} = \frac{\cosh \left( \sqrt{\frac{s}{\alpha}} L \right)}{\sqrt{\frac{s}{\alpha}} \sinh \left( \sqrt{\frac{s}{\alpha}} L \right)} \quad (11.203)$$

The zeros of the transfer function are found from

$$L \sqrt{\frac{s}{\alpha}} = j \left( k\pi + \frac{\pi}{2} \right) \Rightarrow \frac{s}{\alpha} = -\frac{1}{L^2} \left( k\pi + \frac{\pi}{2} \right)^2 \quad (11.204)$$

while the singularities are given by

$$L \sqrt{\frac{s}{\alpha}} = jk\pi \Rightarrow \frac{s}{\alpha} = -\frac{1}{L^2} (k\pi)^2 \quad (11.205)$$

Numerical values are given in Table 11.1.

Zeros		Singularities		
$L\sqrt{\frac{s}{\alpha}}$	$L^2\frac{s}{\alpha}$	$L\sqrt{\frac{s}{\alpha}}$	$L^2\frac{s}{\alpha}$	
1.5708	-2.4674	0	0	
4.7124	-22.207	3.1416	-9.8696	
7.8540	-61.685	6.2832	-39.478	
10.995541	-120.9019	9.4248	-88.826	

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  $\beta$ . Then the boundary conditions are changed to*

$$\frac{\partial T(0, t)}{\partial x} = 0, \quad \frac{\partial T(L, t)}{\partial x} = \beta[u - T(L, t)] \quad (11.207)$$

*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*

$$A = \frac{\beta}{\sqrt{\frac{s}{\alpha}} \sinh\left(\sqrt{\frac{s}{\alpha}}L\right) + \beta \cosh\left(\sqrt{\frac{s}{\alpha}}L\right)} u(s) \quad (11.208)$$

*and insertion in (11.199) gives the transfer function*

$$\frac{T(L, s)}{u(s)} = \frac{\cosh\left(\sqrt{\frac{s}{\alpha}}L\right)}{\frac{1}{\beta}\sqrt{\frac{s}{\alpha}} \sinh\left(\sqrt{\frac{s}{\alpha}}L\right) + \cosh\left(\sqrt{\frac{s}{\alpha}}L\right)} \quad (11.209)$$

## 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

for doing vector operations at the component level. We will see in the following that tensor notation is of particular use in connection with the computation of gradients and divergence of complicated vector expressions. All tensors in the following are Cartesian. Let  $a$  be a Cartesian frame with orthogonal unit vectors  $\vec{a}_1, \vec{a}_2, \vec{a}_3$  and let  $\vec{u}$  be a vector and  $\vec{D}$  be a dyadic given by

$$\vec{u} = \sum_{i=1}^3 u_i \vec{a}_i, \quad \vec{D} = \sum_{i=1}^3 \sum_{j=1}^3 d_{ij} \vec{a}_i \vec{a}_j \quad (11.210)$$

where  $u_i = \vec{u} \cdot \vec{a}_i$  is component  $i$  of  $\vec{u}$  and  $d_{ij} = \vec{a}_i \cdot \vec{D} \cdot \vec{a}_j$  is component  $i, j$  of  $\vec{D}$ . The vector  $\vec{u}$  and the dyadic  $\vec{D}$  are uniquely defined by their components. This means that we may represent  $\vec{u}$  by its generic component  $u_i$  and we may represent  $\vec{D}$  by its generic component  $d_{ij}$ .

The generic component  $u_i = \vec{u} \cdot \vec{a}_i$  of a vector  $\vec{u}$  is a *first order Cartesian tensor*, while the generic component  $d_{ij} = \vec{a}_i \cdot \vec{D} \cdot \vec{a}_j$  of a dyadic  $\vec{D}$  is a *second order Cartesian tensor*.

We introduce the *summation convention* where the summation symbol may be left out when it is summed over a repeated index in a product as in the scalar product

$$u_i v_i := \sum_{i=1}^3 u_i v_i = \mathbf{u}^T \mathbf{v} \quad (11.211)$$

The summation convention also applies to the vector expression

$$\vec{u} = \sum_{i=1}^3 u_i \vec{a}_i = u_i \vec{a}_i \quad (11.212)$$

and the dyadic expression

$$\vec{D} = \sum_{i=1}^3 \sum_{j=1}^3 d_{ij} \vec{a}_i \vec{a}_j = d_{ij} \vec{a}_i \vec{a}_j \quad (11.213)$$

In addition we will use a notation where subscript  $, i$  denotes partial differentiation with respect to  $x_i$ , so that

$$\phi_{,i} := \frac{\partial \phi}{\partial x_i} \quad \text{and} \quad v_{i,j} := \frac{\partial v_i}{\partial x_j} \quad (11.214)$$

In addition, subscript  $, ij$  denotes partial differentiation with respect to  $x_i$  and  $x_j$ , that is,

$$\phi_{,ij} := (\phi_{,i})_{,j} = \frac{\partial^2 \phi}{\partial x_i \partial x_j} \quad \text{and} \quad v_{i,jk} := (v_{i,j})_{,k} = \frac{\partial^2 v_i}{\partial x_j \partial x_k} \quad (11.215)$$

We let the summation convention apply to differentiation expressions so that

$$v_{j,ji} := \sum_{j=1}^3 v_{j,ji} = \frac{\partial}{\partial x_i} (\nabla^T \mathbf{v}) \quad \text{and} \quad v_{i,jj} := \sum_{j=1}^3 v_{i,jj} = \nabla^2 v_i \quad (11.216)$$

**Example 174** The tensor form of the material derivative of a scalar  $\phi$  is

$$\frac{D\phi}{Dt} = \frac{\partial\phi}{\partial t} + v_i \phi_{,i} \quad (11.217)$$

while the material derivative of a vector  $\mathbf{u}$  with elements  $u_i$  is

$$\frac{Du_i}{Dt} = \frac{\partial u_i}{\partial t} + v_j u_{i,j} \quad (11.218)$$

**Example 175** The divergence of velocity  $\vec{\nabla} \cdot \vec{v}$  can be written

$$\vec{\nabla} \cdot \vec{v} = \sum_{i=1}^3 v_{i,i} = v_{i,i} \quad (11.219)$$

while the scalar product between the  $\vec{v}$  and the gradient of a scalar  $\psi$  can be written

$$\vec{v} \cdot \vec{\nabla} \psi = \sum_{i=1}^3 v_i \psi_{,i} = v_i \psi_{,i} \quad (11.220)$$

**Theorem 1** The divergence theorem (10.12) can be written in tensor notation as

$$\iint_{\partial V(t)} u_i n_i dA = \iiint_{V(t)} u_{i,i} dV \quad (11.221)$$

Moreover, the related result (10.13) is written

$$\iint_{\partial V(t)} \phi n_i dA = \iiint_{V(t)} \vec{\nabla} \phi_{,i} dV \quad (11.222)$$

**Example 176** The Laplacian of a scalar  $\phi$  is

$$\nabla^2 \phi = \sum_{i=1}^3 \phi_{,ii} = \phi_{,ii} \quad (11.223)$$

where  $\nabla^2 = \vec{\nabla} \cdot \vec{\nabla} = \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_i}$  is the Laplacian operator.

**Example 177** The Hessian matrix of a scalar  $\phi$  is

$$\boldsymbol{\nabla} \boldsymbol{\nabla}^T \phi = \begin{pmatrix} \frac{\partial}{\partial x_1} \frac{\partial}{\partial x_1} & \frac{\partial}{\partial x_1} \frac{\partial}{\partial x_2} & \frac{\partial}{\partial x_1} \frac{\partial}{\partial x_3} \\ \frac{\partial}{\partial x_2} \frac{\partial}{\partial x_1} & \frac{\partial}{\partial x_2} \frac{\partial}{\partial x_2} & \frac{\partial}{\partial x_2} \frac{\partial}{\partial x_3} \\ \frac{\partial}{\partial x_3} \frac{\partial}{\partial x_1} & \frac{\partial}{\partial x_3} \frac{\partial}{\partial x_2} & \frac{\partial}{\partial x_3} \frac{\partial}{\partial x_3} \end{pmatrix} \phi = \begin{pmatrix} \phi_{,11} & \phi_{,12} & \phi_{,13} \\ \phi_{,21} & \phi_{,22} & \phi_{,23} \\ \phi_{,31} & \phi_{,32} & \phi_{,33} \end{pmatrix} \quad (11.224)$$

where  $\boldsymbol{\nabla} \boldsymbol{\nabla}^T = \left\{ \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_j} \right\}$  is the Hessian operator in matrix form.

**Example 178** The scalar  $\vec{\nabla} \cdot (\phi \vec{u})$ , which is the divergence of the vector  $\phi \vec{u}$  can be written in tensor notation as  $(\phi u_i)_{,i}$ . From the usual rule for the differentiation of products it follows that

$$(\phi u_i)_{,i} = \phi_{,i} u_i + \phi u_{i,i} \quad (11.225)$$

which is the tensor form of (10.15).

**Example 179** Matrix multiplication is conveniently expressed in tensor notation. Let  $\mathbf{x} = \{x_i\}$ ,  $\mathbf{y} = \{y_i\}$ ,  $\mathbf{A} = \{a_{ij}\}$ ,  $\mathbf{B} = \{b_{ij}\}$ ,  $\mathbf{C} = \{c_{ij}\}$ , and  $\mathbf{D} = \{d_{ij}\}$ . Then it is straightforward to verify that

$$\mathbf{y} = \mathbf{Ax} \Leftrightarrow y_i = a_{ij}x_j \quad (11.226)$$

$$\mathbf{C} = \mathbf{AB} \Leftrightarrow c_{ij} = a_{ik}b_{kj} \quad (11.227)$$

$$\mathbf{D} = \mathbf{A}^T \mathbf{B} \Leftrightarrow d_{ij} = a_{ki}b_{kj} \quad (11.228)$$

$$\mathbf{E} = \mathbf{AB}^T \Leftrightarrow e_{ij} = a_{ik}b_{jk} \quad (11.229)$$

### 11.5.3 The velocity gradient tensor

Viscous forces appear in fluids because of velocity gradients. To describe velocity gradients it is convenient to introduce the velocity gradient tensor defined by

$$v_{i,j} = \frac{\partial v_i}{\partial x_j} \quad (11.230)$$

This is the tensor form of the velocity gradient dyadic

$$\vec{\nabla} \vec{v} = v_{i,j} \vec{a}_i \vec{a}_j \quad (11.231)$$

while the corresponding matrix form is  $(\nabla \mathbf{v}^T)^T$ .

The velocity gradient tensor  $v_{i,j}$  is written

$$v_{i,j} = e_{ij} + \Omega_{ij} \quad (11.232)$$

where

$$e_{ij} := \frac{1}{2} (v_{i,j} + v_{j,i}) \quad (11.233)$$

is the symmetric *rate of strain tensor*, which is also called the *deformation tensor*, and

$$\Omega_{ij} := \frac{1}{2} (v_{i,j} - v_{j,i}) \quad (11.234)$$

is the skew-symmetric part of the velocity gradient tensor.

The matrix form of the rate of strain tensor is written

$$\mathbf{E} := \{e_{ij}\} = \left\{ \frac{1}{2} (v_{i,j} + v_{j,i}) \right\} \quad (11.235)$$

while

$$\boldsymbol{\Omega} := \{\Omega_{ij}\} = \left\{ \frac{1}{2} (v_{i,j} - v_{j,i}) \right\} \quad (11.236)$$

**Example 180** Define  $ds$  by

$$ds^2 = dx_i dx_i \quad (11.237)$$

which means that  $ds$  is the length of the differential vector  $d\mathbf{x} = (dx_1, dx_2, dx_3)^T$ . Then the material derivative of  $ds^2$  is found to be

$$\begin{aligned} \frac{1}{2} \frac{D}{Dt} (ds)^2 &= \frac{1}{2} \frac{D}{Dt} (dx_i dx_i) = dx_i \frac{D}{Dt} dx_i \\ &= dx_i \frac{D}{Dt} \frac{\partial x_i}{\partial \xi_j} d\xi_j = dx_i \frac{\partial v_i}{\partial \xi_j} d\xi_j \end{aligned} \quad (11.238)$$

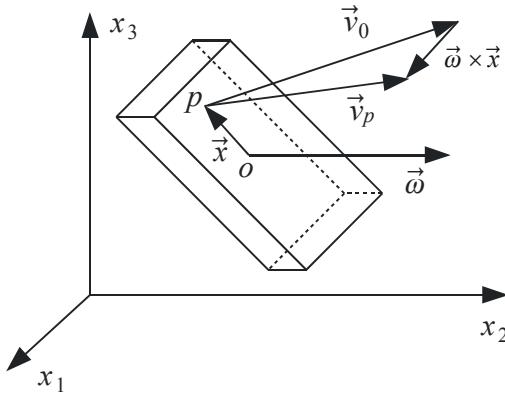


Figure 11.13: Rigid body with velocity  $\vec{v}_o$  of point  $o$  and angular velocity  $\vec{\omega}$ .

where  $\xi_j$  are the material coordinates. Then, because

$$\frac{\partial v_i}{\partial \xi_j} d\xi_j = \frac{\partial v_i}{\partial x_j} dx_j \quad (11.239)$$

it follows that

$$\frac{1}{2} \frac{D}{Dt} (ds)^2 = dx_i e_{ij} dx_j \quad (11.240)$$

We see that the deformation tensor  $e_{ij}$  is related to the stretching of the differential  $d\mathbf{x}$ .

#### 11.5.4 Example: The velocity gradient for a rigid body

Consider a rigid body with angular velocity  $\vec{\omega}(t)$  and velocity  $\vec{v}_o$  and position  $\vec{r}_o(t)$  of some specified point  $o$ , which is fixed in the rigid body. We consider a fixed point  $p$  in the rigid body with position  $\vec{r}(t, x_1, x_2, x_3)$ . The velocity of the point  $p$  is

$$\vec{v}(t, x_1, x_2, x_3) = \vec{v}_o(t) + \vec{\omega}(t) \times \vec{r}(t, x_1, x_2, x_3) \quad (11.241)$$

where  $\vec{r} = \vec{x} - \vec{r}_o$ . The velocity gradient  $\vec{\nabla} \vec{v}$  which describes the velocity variations over the rigid body is

$$\begin{aligned} \vec{\nabla} \vec{v} &= \vec{\nabla} \vec{v}_o + \vec{\nabla} (\vec{\omega} \times \vec{r}) = \vec{\omega} \times \vec{\nabla} \vec{r} = \vec{\omega} \times \vec{\nabla} \vec{x} = \vec{\omega}^\times \cdot \vec{\nabla} \vec{x} = \vec{\omega}^\times \cdot \vec{I} \\ &= \vec{\omega}^\times \end{aligned} \quad (11.242)$$

Here we have used the fact that  $\vec{v}_o$ ,  $\vec{r}_o$  and  $\vec{\omega}$  are functions of time only, and that  $\vec{\nabla} \vec{r} = \vec{\nabla} \vec{x} = \vec{I}$ . We see that for rigid-body motion the velocity gradient tensor is skew symmetric and given by

$$\vec{\nabla} \vec{v} = \vec{\omega}^\times \quad (11.243)$$

This means that for rigid body motion we have

$$\boldsymbol{\Omega} = \boldsymbol{\omega}^\times, \quad e_{ij} = 0 \quad (11.244)$$

The opposite is also true: If the rate of strain tensor  $e_{ij}$  is zero, then the motion is a rigid body motion. From this we conclude that a nonzero rate of strain tensor  $e_{ij}$  is a measure of how much the velocity field differs from rigid body motion, and in this sense  $e_{ij}$  is related to the rate of deformation of the fluid.

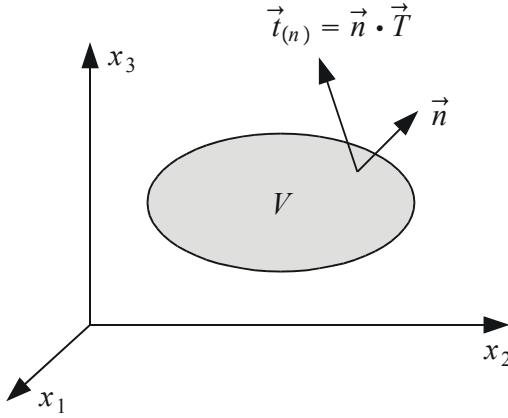


Figure 11.14: Volume  $V$  with surface normal  $\vec{n}$  and stress vector  $\vec{t}_{(n)}$ .

**Example 181** We recognize  $2\Omega$  in (11.236) as the skew symmetric form of the curl vector  $\nabla^\times \mathbf{v}$  as given in (10.21). This means that

$$(\nabla^\times \mathbf{v})^\times = 2\Omega \quad (11.245)$$

From this and (11.244) we conclude that for rigid body motion we have

$$\vec{\nabla} \times \vec{v} = 2\vec{\omega} \quad (11.246)$$

and

$$(\vec{\nabla} \times \vec{v})^\times = 2\vec{\nabla} \vec{v} \quad (11.247)$$

### 11.5.5 The stress tensor

The forces acting on a material volume  $V(t)$  are divided into *mass forces*  $\iiint_V \rho \vec{f} dV$  and *surface forces*  $\iint_{\partial V} \vec{t}_{(n)} dA$ . Here  $\vec{t}_{(n)}$  is the stress vector and  $\vec{t}_{(n)} dA$  is the contact force acting on the area element  $dA$ . The subscript  $(n)$  indicates that the stress vector  $\vec{t}_{(n)}$  is acting on a surface with outwards surface normal  $\vec{n}$  as shown in Figure 11.14. The stress vector  $\vec{t}_{(n)}$  can be expressed in the form

$$\vec{t}_{(n)} = \vec{n} \cdot \vec{T}$$

where  $\vec{T}$  is the dyadic form of the the *stress tensor*  $T_{ij}$ . We note that the corresponding matrix representation is  $\mathbf{T} = \{T_{ij}\}$ . By convention, the element  $T_{ij}$  denotes the contact force in the  $j$  direction exerted on a plane with surface normal in the  $i$  direction.

It is not straightforward to handle divergence terms involving the stress tensor in vector form. Because of this we use the more powerful tensor notation in component form, where the summation convention is used. Then, component  $i$  of the stress vector  $\vec{t}_{(n)}$  is written

$$t_{(n)i} = T_{ji} n_j \quad (11.248)$$

where the summation over  $j$  is implied by the summation convention. The divergence theorem (11.221) leads to the equation

$$\iint_{\partial V} T_{ji} n_j dA = \iiint_V T_{ji,j} dV \quad (11.249)$$

where the notation

$$T_{ji,j} = \frac{\partial T_{ji}}{\partial x_j} = \sum_{j=1}^3 \frac{\partial T_{ji}}{\partial x_j} \quad (11.250)$$

is used. In vector form this is written

$$\iint_{\partial V} \vec{t}_{(n)} dA = \iint_{\partial V} \vec{n} \cdot \vec{T} dA = \iiint_V (\vec{\nabla} \cdot \vec{T}) dV \quad (11.251)$$

The stress tensor can generally be written

$$\vec{T} = -p\vec{I} + \vec{\tau} \quad (11.252)$$

where  $p$  is the pressure and  $\vec{\tau}$  is the *viscous stress tensor*. We note that the equivalent tensor form is

$$T_{ij} = -p\delta_{ij} + \tau_{ij} \quad (11.253)$$

where  $\delta_{ij}$  is the Kronecker delta which is given by  $\delta_{ii} = 1$  and  $\delta_{ij} = 0$  for  $i \neq j$ . From

$$T_{ji,j} = -(p\delta_{ij})_{,j} + \tau_{ji,j} = -p_{,i} + \tau_{ji,j} \quad (11.254)$$

we find that

$$\vec{\nabla} \cdot \vec{T} = -\vec{\nabla} p + \vec{\nabla} \cdot \vec{\tau} \quad (11.255)$$

and we may use (11.251) to establish the following useful relation:

$$\iint_{\partial V} \vec{t}_{(n)} dA = \iiint_V (-\vec{\nabla} p + \vec{\nabla} \cdot \vec{\tau}) dV \quad (11.256)$$

which has the component form

$$\iint_{\partial V} t_{(n)i} dA = \iiint_V (-p_{,i} + \tau_{ji,j}) dV \quad (11.257)$$

**Example 182** We note that the matrix form of the main equations of this section is given by

$$\mathbf{T} = -p\mathbf{I} + \boldsymbol{\tau} \quad (11.258)$$

$$(\boldsymbol{\nabla}^T \mathbf{T})^T = -\boldsymbol{\nabla} p + (\boldsymbol{\nabla}^T \boldsymbol{\tau})^T \quad (11.259)$$

and

$$\iint_{\partial V} \mathbf{t}_{(n)} dA = \iiint_V \left[ -\boldsymbol{\nabla} p + (\boldsymbol{\nabla}^T \boldsymbol{\tau})^T \right] dV \quad (11.260)$$

**Remark 4** Some authors use  $\vec{T} = -p\vec{I} - \vec{\tau}$ , which means that they use the opposite sign for the viscous stress tensor.

### 11.5.6 Cauchy's equation of motion

The resultant force acting on a fluid of volume  $V$  is the sum of the surface forces and the body forces, which is written

$$\vec{F}^{(r)} = - \iint_{\partial V} \vec{t}_{(n)} dA + \iiint_V \rho \vec{f} dV \quad (11.261)$$

The divergence theorem gives

$$\vec{F}^{(r)} = \iiint_V (\vec{\nabla} \cdot \vec{T} + \rho \vec{f}) dV \quad (11.262)$$

and by combining this with the momentum balance (11.53) the result is

$$\iiint_V \rho \frac{D\vec{v}}{Dt} dV = \iiint_V (\vec{\nabla} \cdot \vec{T} + \rho \vec{f}) dV \quad (11.263)$$

As the volume  $V$  is arbitrary, this leads to *Cauchy's equation of motion*

$$\rho \frac{Dv_i}{Dt} = T_{ji,j} + \rho f_i \quad (11.264)$$

$$\rho \frac{D\vec{v}}{Dt} = \vec{\nabla} \cdot \vec{T} + \rho \vec{f} \quad (11.265)$$

which is stated both in component and vector form.

We express the stress tensor in terms of the pressure  $p$  and the viscous stress tensor  $\tau_{ij}$  as given in (11.253).

Cauchy's equation of motion for a fluid with viscosity is given in tensor form and vector form as

$$\rho \frac{Dv_i}{Dt} = -p_{,i} + \tau_{ji,j} + \rho f_i \quad (11.266)$$

$$\rho \frac{D\vec{v}}{Dt} = -\vec{\nabla} p + \vec{\nabla} \cdot \vec{\tau} + \rho \vec{f} \quad (11.267)$$

Cauchy's equation of motion can be written in divergence form by inserting (11.30). This gives

$$\frac{\partial(\rho \vec{v})}{\partial t} + \vec{\nabla} \cdot (\rho \vec{v} \vec{v}) = -\vec{\nabla} p + \vec{\nabla} \cdot \vec{\tau} + \rho \vec{f} \quad (11.268)$$

**Example 183** The component form of Cauchy's equation of motion (11.266) is

$$\rho \frac{Dv_1}{Dt} = -\frac{\partial p}{\partial x_1} + \frac{\partial \tau_{11}}{\partial x_1} + \frac{\partial \tau_{21}}{\partial x_2} + \frac{\partial \tau_{31}}{\partial x_3} + \rho f_1 \quad (11.269)$$

$$\rho \frac{Dv_2}{Dt} = -\frac{\partial p}{\partial x_2} + \frac{\partial \tau_{12}}{\partial x_1} + \frac{\partial \tau_{22}}{\partial x_2} + \frac{\partial \tau_{32}}{\partial x_3} + \rho f_2 \quad (11.270)$$

$$\rho \frac{Dv_3}{Dt} = -\frac{\partial p}{\partial x_3} + \frac{\partial \tau_{13}}{\partial x_1} + \frac{\partial \tau_{23}}{\partial x_2} + \frac{\partial \tau_{33}}{\partial x_3} + \rho f_3 \quad (11.271)$$

**Example 184** The momentum equation for a multi-component fluid is

$$\rho \frac{D\vec{v}}{Dt} = \vec{\nabla} \cdot \vec{T} + \sum_{k=1}^n \rho_k \vec{f}_k \quad (11.272)$$

We see that the only difference from a single-component fluid is in the term related to the mass forces  $\vec{f}_k$ .

### 11.5.7 Newtonian fluids

The results that have been derived in the previous sections for the stress tensor are kinematic and are valid for any fluid. The constitutive equations give expressions for the stress tensor of a particular fluid. Fluids can be arranged in different classes of fluids depending on which constitutive equations they satisfy. An example of this are the *Newtonian fluids*, which include water and oil. It is shown that the insertion of the viscous stress tensor of an incompressible Newtonian fluid in Cauchy's equation of motion leads to the Navier-Stokes equation.

Some notable fluids are not Newtonian. One example is blood, and another is the mud used in oil drilling. Non-Newtonian fluids will not be further discussed, but it is remarked that the equations of motion for such fluids can be obtained by inserting the viscous stress tensor of the specific fluid into Cauchy's equation of motion (11.266).

Newtonian fluids are fluids where the stress tensor is given by

$$\vec{\tau} = \lambda (\vec{\nabla} \cdot \vec{v}) \vec{I} + 2\mu \vec{E} \quad (11.273)$$

where  $\lambda$  and  $\mu$  are the *Lamé coefficients*, and  $\vec{E} = e_{ij} \vec{a}_i \vec{a}_j$  is the *rate of strain tensor* defined by (11.233).

The component form of the constitutive equation (11.273) is

$$\tau_{ij} = \lambda v_{k,k} \delta_{ij} + 2\mu e_{ij} \quad (11.274)$$

We note that for a Newtonian fluid the stress tensor is symmetric, so that  $\tau_{ji} = \tau_{ij}$ . To evaluate the term  $\tau_{ji,j} = \tau_{ij,j}$  in Cauchy's equation of motion (11.266) we first calculate

$$e_{ij,j} = \frac{1}{2} (v_{i,j} + v_{j,i})_{,j} = \frac{1}{2} (v_{i,jj} + v_{j,ij}) = \frac{1}{2} (v_{i,jj} + v_{j,ji}) \quad (11.275)$$

This gives

$$\begin{aligned} \tau_{ij,j} &= (\lambda v_{k,k} \delta_{ij})_{,j} + 2\mu e_{ij,j} \\ &= \lambda v_{k,ki} + \mu v_{j,ji} + \mu v_{i,jj} \end{aligned} \quad (11.276)$$

We may change indices in  $v_{j,ji}$  by noting that  $j$  is a dummy index so that  $v_{j,ji} = v_{k,ki}$ . This gives

$$\tau_{ij,j} = (\lambda + \mu) v_{k,ki} + \mu v_{i,jj} \quad (11.277)$$

In vector form this is written

$$\vec{\nabla} \cdot \vec{\tau} = (\lambda + \mu) \vec{\nabla} (\vec{\nabla} \cdot \vec{v}) + \mu \vec{\nabla}^2 \vec{v} \quad (11.278)$$

**Example 185** Let the velocity field be given by  $\mathbf{v} = (cx_2^2, 0, 0)^T$ , or equivalently, by

$$\vec{v} = cx_2^2 \vec{a}_1 \quad (11.279)$$

where  $\vec{a}_1, \vec{a}_2, \vec{a}_3$  are orthogonal unit vectors along the  $x_1, x_2, x_3$  axes of a Cartesian coordinate frame. The position is given by  $\mathbf{x} = (x_1, x_2, x_3)^T$ . Then the velocity gradient tensor is

$$\begin{aligned} \vec{\nabla} \vec{v} &= \left( \vec{a}_1 \frac{\partial}{\partial x_1} + \vec{a}_2 \frac{\partial}{\partial x_2} + \vec{a}_3 \frac{\partial}{\partial x_3} \right) cx_2^2 \vec{a}_1 \\ &= 2cx_2 \vec{a}_2 \vec{a}_1 \end{aligned} \quad (11.280)$$

The deformation tensor is the symmetric part of the velocity gradient tensor and is given by

$$\vec{E} = cx_2\vec{a}_2\vec{a}_1 + cx_2\vec{a}_1\vec{a}_2 \quad (11.281)$$

The divergence is found to be

$$\vec{\nabla} \cdot \vec{v} = 0 \quad (11.282)$$

while the Laplacian is

$$\vec{\nabla}^2 \vec{v} = \left( \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2} \right) cx_2^2\vec{a}_1 = 2c\vec{a}_1 \quad (11.283)$$

Then the viscous stress tensor is

$$\vec{\tau} = 2\mu cx_2\vec{a}_2\vec{a}_1 + 2\mu cx_2\vec{a}_1\vec{a}_2 \quad (11.284)$$

We see that the term  $\vec{\nabla} \cdot \vec{\tau}$  in the equation of motion is given by

$$\vec{\nabla} \cdot \vec{\tau} = \mu \vec{\nabla}^2 \vec{v} = 2\mu c\vec{a}_1 \quad (11.285)$$

Consider the surface with the surface normal  $\vec{n} = \vec{a}_2$ . Then the stress vector on that surface is

$$\vec{t}_{(n)} = \vec{n} \cdot \vec{\tau} = 2\mu cx_2\vec{a}_1 \quad (11.286)$$

The stress vector is seen to be in the direction of the flow.

**Example 186** It is common practice to use an assumption due to Stokes, which involves setting the relation between the Lamé coefficients so that

$$\lambda + \frac{2}{3}\mu = 0 \quad (11.287)$$

We will explain the motivation for this relation. Denote the mean of the diagonal terms of the stress tensor  $T_{ij}$  by  $-\bar{p}$ . By noting that the divergence is related to the diagonal terms of the rate of strain tensor according to

$$\vec{\nabla} \cdot \vec{v} = v_{k,k} = e_{kk} \quad (11.288)$$

we find that

$$-\bar{p} := \frac{1}{3}T_{ii} = -p + (\lambda + \frac{2}{3}\mu)e_{kk} \quad (11.289)$$

The reasoning that leads to (11.287) starts with the observation that for incompressible fluids  $e_{kk} = 0$ , which implies that  $\bar{p} = p$ . If it is assumed that (11.287) is valid, then  $\bar{p} = p$  also for compressible fluids. More details on this is found in (Aris 1989).

**Example 187** The constitutive equation for a Newtonian fluid originates from three assumptions (Aris 1989): First it is assumed that the stress tensor is linear in the velocity gradient tensor, which means that the viscous stress tensor can be written

$$\tau_{ij} = N_{ijkl}v_{k,l} \quad (11.290)$$

for some four-dimensional tensor  $N_{ijkl}$ . Second, it is assumed that  $N_{ijkl}$  is isentropic. This implies that  $N_{ijkl}$  can be expressed in terms of the Lamé parameters  $\lambda$ ,  $\mu$  and  $\kappa$  as

$$N_{ijkl} = \lambda\delta_{ij}\delta_{kl} + \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}) + \kappa(\delta_{ik}\delta_{jl} - \delta_{il}\delta_{jk}) \quad (11.291)$$

Third, symmetry is assumed in the sense that  $\tau_{ij} = \tau_{ji}$ , which implies that  $\kappa = 0$ . This gives the constitutive equation for a Newtonian fluid as

$$\lambda\delta_{ij}\delta_{kl} + \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})v_{k,l} = \lambda\delta_{ij}v_{k,k} + \mu(v_{i,j} + v_{j,i}) \quad (11.292)$$

### 11.5.8 The Navier-Stokes equation

If it is assumed that a fluid is *Newtonian*, then the momentum equation in the form of Cauchy's equation of motion is found by inserting (11.278) into (11.266)

Cauchy's equation of motion for a Newtonian fluid is given by

$$\rho \frac{D\vec{v}}{Dt} = -\vec{\nabla}p + (\lambda + \mu)\vec{\nabla}(\vec{\nabla} \cdot \vec{v}) + \mu(\vec{\nabla}^2 \vec{v}) + \rho\vec{f} \quad (11.293)$$

We note that the tensor form is

$$\rho \frac{Dv_i}{Dt} = -p_{,i} + (\lambda + \mu)v_{k,ki} + \mu v_{i,jj} + \rho f_i \quad (11.294)$$

We may write this out in each of the three orthogonal directions as

$$\rho \frac{Dv_1}{Dt} = -\frac{\partial p}{\partial x_1} + (\lambda + \mu)\frac{\partial}{\partial x_1}(\vec{\nabla} \cdot \vec{v}) + \mu(\vec{\nabla}^2 \vec{v}) + \rho f_1 \quad (11.295)$$

$$\rho \frac{Dv_2}{Dt} = -\frac{\partial p}{\partial x_2} + (\lambda + \mu)\frac{\partial}{\partial x_2}(\vec{\nabla} \cdot \vec{v}) + \mu(\vec{\nabla}^2 \vec{v}) + \rho f_2 \quad (11.296)$$

$$\rho \frac{Dv_3}{Dt} = -\frac{\partial p}{\partial x_3} + (\lambda + \mu)\frac{\partial}{\partial x_3}(\vec{\nabla} \cdot \vec{v}) + \mu(\vec{\nabla}^2 \vec{v}) + \rho f_3 \quad (11.297)$$

The divergence form of momentum equations for a Newtonian fluid is found to be

$$\frac{\partial(\rho\vec{v})}{\partial t} + \vec{\nabla} \cdot (\rho\vec{v}\vec{v}) = -\vec{\nabla}p + (\lambda + \mu)\vec{\nabla}(\vec{\nabla} \cdot \vec{v}) + \mu\vec{\nabla}^2\vec{v} + \rho\vec{f} \quad (11.298)$$

or, in tensor form,

$$\frac{\partial(\rho v_i)}{\partial t} + (\rho v_i v_j)_{,j} = -p_{,i} + (\lambda + \mu)v_{k,ki} + \mu v_{i,jj} + \rho f_i \quad (11.299)$$

In the incompressible case the divergence of the velocity is zero, and the equation of motion is found by inserting  $\vec{\nabla} \cdot \vec{v} = 0$  into (11.293). This leads to the Navier-Stokes equation.

Cauchy's equation of motion for an incompressible Newtonian fluid is given by

$$\rho \frac{D\vec{v}}{Dt} = -\vec{\nabla}p + \mu\vec{\nabla}^2\vec{v} + \rho\vec{f} \quad (11.300)$$

This is the classical Navier-Stokes equation

The tensor form of the classical Navier-Stokes equation is

$$\rho \frac{Dv_i}{Dt} = -p_{,i} + \mu v_{i,jj} + \rho f_i \quad (11.301)$$

The Euler equation for inviscid flow appears by setting the viscous forces to zero by inserting  $\mu = 0$ . Control of the Navier-Stokes equation is treated in great detail in (Aamo and Krstić 2003).

**Example 188** Insertion of the identity

$$\vec{\nabla}^2 \vec{v} = \vec{\nabla} (\vec{\nabla} \cdot \vec{v}) - \vec{\nabla} \times (\vec{\nabla} \times \vec{v}) \quad (11.302)$$

gives the alternative form

$$\rho \frac{D\vec{v}}{Dt} = -\vec{\nabla} p + (\lambda + 2\mu) \vec{\nabla} (\vec{\nabla} \cdot \vec{v}) - \mu \vec{\nabla} \times (\vec{\nabla} \times \vec{v}) + \rho \vec{f} \quad (11.303)$$

of the momentum equations for a Newtonian fluid. It is interesting to note how the divergence  $\vec{\nabla} \cdot \vec{v}$  and the curl  $\vec{\nabla} \times \vec{v}$  of the velocity appear in the two terms due to the stress tensor. If the fluid is incompressible, then the divergence term is zero, and if there is potential flow, then the curl term is zero.

**Example 189** In cylindrical coordinates the Navier-Stokes equation is found by applying the results of Section 10.3.2 to the coordinate-free form

$$\rho \frac{D\vec{v}}{Dt} = -\vec{\nabla} p + \mu \vec{\nabla}^2 \vec{v} + \rho \vec{f} \quad (11.304)$$

of (11.300) where

$$\vec{v} = v_r \vec{j}_r + v_\theta \vec{j}_\theta + v_z \vec{j}_z \quad (11.305)$$

and

$$\vec{\nabla} = \vec{j}_r \frac{\partial}{\partial r} + \frac{\vec{j}_\theta}{r} \frac{\partial}{\partial \theta} + \vec{j}_z \frac{\partial}{\partial z} \quad (11.306)$$

This results in

$$\begin{aligned} \frac{\partial v_r}{\partial t} + v_r \frac{\partial v_r}{\partial r} + \frac{v_\theta}{r} \frac{\partial v_r}{\partial \theta} + v_z \frac{\partial v_r}{\partial z} - \frac{v_\theta^2}{r} &= -\frac{1}{\rho} \frac{\partial p}{\partial r} \\ &+ \frac{\mu}{\rho} \left( \frac{\partial^2 v_r}{\partial r^2} + \frac{1}{r^2} \frac{\partial^2 v_r}{\partial \theta^2} + \frac{\partial^2 v_r}{\partial z^2} + \frac{1}{r} \frac{\partial v_r}{\partial r} - \frac{2}{r^2} \frac{\partial v_\theta}{\partial \theta} - \frac{v_r}{r^2} \right) \end{aligned} \quad (11.307)$$

$$\begin{aligned} \frac{\partial v_\theta}{\partial t} + v_r \frac{\partial v_\theta}{\partial r} + \frac{v_\theta}{r} \frac{\partial v_\theta}{\partial \theta} + v_z \frac{\partial v_\theta}{\partial z} + \frac{v_r v_\theta}{r} &= -\frac{1}{\rho r} \frac{\partial p}{\partial \theta} \\ &+ \frac{\mu}{\rho} \left( \frac{\partial^2 v_\theta}{\partial r^2} + \frac{1}{r^2} \frac{\partial^2 v_\theta}{\partial \theta^2} + \frac{\partial^2 v_\theta}{\partial z^2} + \frac{1}{r} \frac{\partial v_\theta}{\partial r} + \frac{2}{r^2} \frac{\partial v_r}{\partial \theta} - \frac{v_\theta}{r^2} \right) \end{aligned} \quad (11.308)$$

$$\begin{aligned} \frac{\partial v_z}{\partial t} + v_r \frac{\partial v_z}{\partial r} + \frac{v_\theta}{r} \frac{\partial v_z}{\partial \theta} + v_z \frac{\partial v_z}{\partial z} &= -\frac{1}{\rho} \frac{\partial p}{\partial z} \\ &+ \frac{\mu}{\rho} \left( \frac{\partial^2 v_z}{\partial r^2} + \frac{1}{r^2} \frac{\partial^2 v_z}{\partial \theta^2} + \frac{\partial^2 v_z}{\partial z^2} + \frac{1}{r} \frac{\partial v_z}{\partial r} \right) \end{aligned} \quad (11.309)$$

Further details on the systematic derivation of these equations are found in (Aris 1989) where also results for spherical coordinates are presented.

### 11.5.9 The Reynolds number

The Navier-Stokes equation can be made dimensionless by using the dimensionless variables

$$\mathbf{v}^* = \frac{\mathbf{v}}{U}, \quad \mathbf{x}^* = \frac{\mathbf{x}}{L}, \quad t^* = \frac{tU}{L}, \quad p^* = \frac{p}{\rho U^2}, \quad \nabla^* = \nabla L, \quad \mathbf{f}^* = \frac{\mathbf{f}L}{U^2} \quad (11.310)$$

where  $U$  is a characteristic velocity and  $L$  is a characteristic length. This gives

$$\rho \frac{D\mathbf{v}^*}{Dt^*} \frac{U^2}{L} = -\nabla^* p^* \frac{\rho U^2}{L} + \mu (\nabla^*)^2 \mathbf{v}^* \frac{U}{L^2} + \rho \mathbf{f}^* \frac{U}{L^2} \quad (11.311)$$

which simplifies to

$$\frac{D\mathbf{v}^*}{Dt^*} = -\nabla^* p^* + \mathbf{f}^* + \frac{1}{Re} (\nabla^*)^2 \mathbf{v}^* \quad (11.312)$$

where  $Re$  is the Reynolds number given by

$$Re = \frac{UL}{\nu} \quad (11.313)$$

and

$$\nu = \frac{\mu}{\rho} \quad (11.314)$$

is the kinematic viscosity. It is seen from (11.312) that the Reynolds number indicates the relative importance of viscosity compared to inertial forces.

### 11.5.10 The equation of kinetic energy

The kinetic energy  $\frac{1}{2}\rho\mathbf{v}^2$  is not a conserved quantity. However, an equation for the kinetic energy can be found from the momentum equation by observing that

$$\rho \frac{D}{Dt} \left( \frac{1}{2} \vec{v}^2 \right) = \vec{v} \cdot \left( \rho \frac{D\vec{v}}{Dt} \right) \quad (11.315)$$

This means that the equation for the kinetic energy can be found by premultiplying Cauchy's equation of motion (11.267) by  $\vec{v}$ , which gives

$$\rho \frac{D}{Dt} \left( \frac{1}{2} \vec{v}^2 \right) = \rho \vec{v} \cdot \vec{f} - \vec{v} \cdot \vec{\nabla} p + \vec{v} \cdot (\vec{\nabla} \cdot \vec{\tau}) \quad (11.316)$$

$$= \rho v_i f_i - v_i p_{,i} + \tau_{ij,i} v_j \quad (11.317)$$

In the last term on the right hand side we have interchanged the  $i$  and  $j$  index, which is possible as the term is a scalar, so that  $v_i \tau_{ji,j} = \tau_{jj,i} v_j$ .

From the product rule of differentiation of products it is straightforward to verify the expression

$$(\tau_{ij} v_j)_{,i} = \tau_{ij,i} v_j + \tau_{ij} v_{j,i} \quad (11.318)$$

The vector form of (11.318) is not quite as simple to derive, but we state the result, which is

$$\vec{\nabla} \cdot (\vec{\tau} \cdot \vec{v}) = \vec{v} \cdot (\vec{\nabla} \cdot \vec{\tau}) + \vec{\tau} : (\vec{\nabla} \vec{v}) \quad (11.319)$$

where  $:$  is the double dot product defined for four vectors  $\vec{u}, \vec{v}, \vec{w}, \vec{z}$  by

$$\vec{u} \vec{v} : \vec{w} \vec{z} = (\vec{u} \cdot \vec{w})(\vec{v} \cdot \vec{z}) \quad (11.320)$$

For two dyadics  $\vec{D} = d_{ij} \vec{a}_i \vec{a}_j$  and  $\vec{E} = e_{ij} \vec{a}_i \vec{a}_j$  this gives

$$\vec{D} : \vec{E} = d_{ij} e_{ij} \quad (11.321)$$

For matrix representations  $\mathbf{D} = \{d_{ij}\}$  and  $\mathbf{E} = \{e_{ij}\}$  the double dot product is defined in the same way as the scalar  $\mathbf{D} : \mathbf{E} = d_{ij} e_{ij}$ .

Equation (11.319) is verified in a Cartesian frame  $a$  by the dyadic computations

$$\vec{\nabla} \cdot (\vec{\tau} \cdot \vec{v}) = \frac{\partial}{\partial x_k} \vec{a}_k \cdot \tau_{ij} v_j \vec{a}_i = \frac{\partial}{\partial x_i} (\tau_{ij} v_j) = (\tau_{ij} v_j)_{,i} \quad (11.322)$$

$$\vec{v} \cdot (\vec{\nabla} \cdot \vec{\tau}) = v_l \vec{a}_l \cdot \left( \frac{\partial}{\partial x_k} \vec{a}_k \cdot \tau_{ij} \vec{a}_i \vec{a}_j \right) = \frac{\partial \tau_{ij}}{\partial x_i} v_j = \tau_{ij,i} v_j \quad (11.323)$$

$$\vec{\tau} : (\vec{\nabla} \vec{v}) = \tau d_{ij} \vec{a}_i \vec{a}_j : \frac{\partial v_l}{\partial x_k} \vec{a}_k \vec{a}_l = \tau_{ij} \frac{\partial v_j}{\partial x_i} = \tau_{ij} v_{j,i} \quad (11.324)$$

To complete the discussion we mention that the matrix formulation of the result can be shown on the component level to be

$$\nabla^T (\boldsymbol{\tau} \mathbf{v}) = (\nabla^T \boldsymbol{\tau}) \mathbf{v} + \boldsymbol{\tau} : (\nabla \mathbf{v}^T) \quad (11.325)$$

where at least the second term on the right side is not obvious in a derivation based on the use of the matrix notation.

Combination of (11.319) and (11.316) gives the following result:

The equation for the kinetic energy can be written

$$\underbrace{\rho \frac{D}{Dt} \left( \frac{1}{2} \vec{v}^2 \right)}_{\substack{\text{rate of change} \\ \text{in kinetic energy} \\ \text{for material} \\ \text{volume element}}} = \underbrace{\rho \vec{v} \cdot \vec{f}}_{\substack{\text{work of body} \\ \text{forces on volume} \\ \text{element}}} - \underbrace{\vec{\nabla} \cdot (p \vec{v})}_{\substack{\text{pressure work} \\ \text{on volume} \\ \text{element surface}}} + \underbrace{p \left( \vec{\nabla} \cdot \vec{v} \right)}_{\substack{\text{reversible} \\ \text{conversion} \\ \text{to internal} \\ \text{energy}}} + \underbrace{\vec{\nabla} \cdot (\vec{\tau} \cdot \vec{v})}_{\substack{\text{viscous work} \\ \text{on surface of} \\ \text{volume element}}} - \underbrace{\vec{\tau} : (\vec{\nabla} \vec{v})}_{\substack{\text{irreversible viscous} \\ \text{conversion to} \\ \text{internal energy}}} \quad (11.326)$$

Using the divergence theorem we find that the integral form of the equation for kinetic energy is

$$\frac{D}{Dt} \iiint_{V_c} \frac{\rho}{2} \vec{v}^2 dV = \iiint_{V_c} [\rho \vec{v} \cdot \vec{f} + p (\vec{\nabla} \cdot \vec{v}) - \vec{\tau} : (\vec{\nabla} \vec{v})] dV + \iint_{\partial V_c} \vec{v} \cdot (-p \vec{n} + \vec{n} \cdot \vec{\tau}) dA$$

**Example 190** Still another expression for the left hand side of (11.326) is found from (11.24):

$$\rho \frac{D}{Dt} \left( \frac{1}{2} \vec{v}^2 \right) = \rho \left[ \frac{\partial}{\partial t} \left( \frac{1}{2} \vec{v}^2 \right) + \vec{\nabla} \cdot \left( \frac{1}{2} \vec{v}^2 \right) \vec{v} \right] \quad (11.327)$$

**Example 191** For a Newtonian fluid the stress tensor is symmetric, and the term  $\boldsymbol{\tau} : (\nabla \mathbf{v}^T) = \boldsymbol{\tau} : \mathbf{E}$  is found to be positive from the calculation

$$\begin{aligned} \vec{\tau} : (\vec{\nabla} \vec{v}) &= \lambda (\vec{\nabla} \cdot \vec{v}) \vec{I} : \vec{E} + 2\mu \vec{E} : \vec{E} \\ &= \lambda (e_{ii})^2 + 2\mu e_{ij} e_{ij} \end{aligned} \quad (11.328)$$

which clearly shows that  $\vec{\tau} : (\vec{\nabla} \vec{v}) \geq 0$ .

**Example 192** Let the velocity field be given by  $\mathbf{v} = (cx_2^2, 0, 0)^T$ , or equivalently, by

$$\vec{v} = cx_2^2 \vec{a}_1 \quad (11.329)$$

where  $\vec{a}_1, \vec{a}_2, \vec{a}_3$  are orthogonal unit vectors along the  $x_1, x_2, x_3$  axes of a Cartesian coordinate frame. The position is given by  $\mathbf{x} = (x_1, x_2, x_3)^T$ . Then the deformation tensor is

$$\vec{E} = cx_2 \vec{a}_2 \vec{a}_1 + cx_2 \vec{a}_1 \vec{a}_2 \quad (11.330)$$

irreversible conversion to internal energy is found from (11.328) to be

$$\vec{\tau} : (\vec{\nabla} \vec{v}) = \lambda (e_{ii})^2 + 2\mu e_{ij} e_{ij} = 4\mu c^2 x_2^2 \quad (11.331)$$

which clearly is nonnegative. The viscous work on the surface corresponds to the divergence term

$$\begin{aligned} \vec{\nabla} \cdot (\vec{r} \cdot \vec{v}) &= \left( \frac{\partial}{\partial x_1} \vec{a}_1 + \frac{\partial}{\partial x_2} \vec{a}_2 + \frac{\partial}{\partial x_3} \vec{a}_3 \right) \cdot (2\mu c x_2 \vec{a}_2 \vec{a}_1 + 2\mu c x_2 \vec{a}_1 \vec{a}_2) \cdot cx_2^2 \vec{a}_1 \\ &= \left( \frac{\partial}{\partial x_1} \vec{a}_1 + \frac{\partial}{\partial x_2} \vec{a}_2 + \frac{\partial}{\partial x_3} \vec{a}_3 \right) \cdot 2\mu c^2 x_2^3 \vec{a}_2 \\ &= 6\mu c^2 x_2^2 \end{aligned} \quad (11.332)$$

**Example 193** If the stress tensor  $T_{ij}$  is symmetric, then

$$T_{ji} v_{i,j} = \frac{1}{2} (T_{ji} v_{i,j} + T_{ij} v_{j,i}) = T_{ij} e_{ij} \quad (11.333)$$

where  $e_{ij}$  is the rate of strain tensor. This gives

$$\vec{T} : (\vec{\nabla} \vec{v}) = \vec{T} : \vec{E}, \quad \vec{T} \text{ symmetric} \quad (11.334)$$

### 11.5.11 The energy balance for a viscous fluid

The material time derivative of the total energy in a volume  $V$  is equal to the heat flow into the volume due to the heat flux density  $\vec{j}_Q$  plus the power added from the contact stress

$$\vec{t}_{(n)} = -p\vec{n} + \vec{n} \cdot \vec{\tau} \quad (11.335)$$

acting on the surface. This is written

$$\frac{D}{Dt} \iiint_V \rho e dV = - \iint_{\partial V} p \vec{v} \cdot \vec{n} dA + \iint_{\partial V} \vec{v} \cdot (\vec{n} \cdot \vec{\tau}) dA - \iint_{\partial V} \vec{j}_Q \cdot \vec{n} dA \quad (11.336)$$

The volume  $V$  is arbitrary, and it follows from the divergence theorem that

$$\underbrace{\rho \frac{De}{Dt} \left( u + \frac{1}{2} \vec{v}^2 + \phi \right)}_{\substack{\text{rate of change} \\ \text{in internal, kinetic} \\ \text{and potential energy} \\ \text{for material} \\ \text{volume element}}} = - \underbrace{\vec{\nabla} \cdot (p \vec{v})}_{\substack{\text{pressure work} \\ \text{on the surface of} \\ \text{the volume element}}} + \underbrace{\vec{\nabla} \cdot (\vec{\tau} \cdot \vec{v})}_{\substack{\text{viscous work} \\ \text{on the surface of} \\ \text{the volume element}}} - \underbrace{\vec{\nabla} \cdot \vec{j}_Q}_{\substack{\text{heat} \\ \text{conduction}}}$$

**Example 194** The equation for the internal energy is found by subtracting (11.326) from this equation, which gives

$$\underbrace{\rho \frac{Du}{Dt}}_{\substack{\text{rate of change} \\ \text{in internal energy} \\ \text{for material} \\ \text{volume element}}} = - \underbrace{p (\vec{\nabla} \cdot \vec{v})}_{\substack{\text{reversible} \\ \text{conversion}}} + \underbrace{\vec{\tau} : (\vec{\nabla} \vec{v})}_{\substack{\text{irreversible viscous} \\ \text{conversion to} \\ \text{internal energy}}} - \underbrace{\vec{\nabla} \cdot \vec{j}_Q}_{\substack{\text{heat} \\ \text{conduction}}} \quad (11.337)$$

### 11.5.12 Fixed volume

If the volume  $V_f$  is fixed then Reynolds' transport theorem (10.90) gives

$$\frac{d}{dt} \iiint_{V_f} \rho e dV = \frac{D}{Dt} \iiint_{V_f} \rho e dV - \iint_{\partial V_f} \rho e \vec{v} \cdot \vec{n} dA \quad (11.338)$$

Insertion of (11.336) gives the result

$$\frac{d}{dt} \iiint_{V_f} \rho e dV = - \iint_{\partial V_f} \rho \left( e + \frac{p}{\rho} \right) \vec{v} \cdot \vec{n} dA + \iint_{\partial V_f} \vec{v} \cdot (\vec{n} \cdot \vec{\tau}) dA - \iint_{\partial V_f} \vec{j}_Q \cdot \vec{n} dA \quad (11.339)$$

where the first term on the right side is the convected energy plus the pressure work on the volume. Then the energy balance can be written

$$\underbrace{\frac{d}{dt} \iiint_{V_f} \rho \left( u + \frac{1}{2} \vec{v}^2 + \phi \right) dV}_{\substack{\text{rate of change} \\ \text{of energy} \\ \text{in fixed volume}}} = - \underbrace{\iint_{\partial V_f} \rho \left( h + \frac{1}{2} \vec{v}^2 + \phi \right) \vec{v} \cdot \vec{n} dA}_{\substack{\text{convected enthalpy,} \\ \text{kinetic energy and} \\ \text{potential energy}}} \\ + \underbrace{\iint_{\partial V_f} \vec{v} \cdot (\vec{\tau} \cdot \vec{n}) dA}_{\substack{\text{viscous work} \\ \text{on volume surface}}} - \underbrace{\iint_{\partial V_f} \vec{j}_Q \cdot \vec{n} dA}_{\substack{\text{heat} \\ \text{conduction}}} \quad (11.340)$$

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.

### 11.5.13 General control volume

For a general control volume  $V$  Reynolds' transport theorem (10.88) gives

$$\frac{d}{dt} \iiint_{V_c} \rho e dV = \frac{D}{Dt} \iiint_{V_c} \rho e dV - \iint_{\partial V_c} \rho e (\vec{v} - \vec{v}_c) \cdot \vec{n} dA \quad (11.341)$$

From equation (11.336) we then find that

$$\begin{aligned} \frac{d}{dt} \iiint_{V_c} \rho e dV &= - \iint_{\partial V_c} \rho \left( e + \frac{p}{\rho} \right) (\vec{v} - \vec{v}_c) \cdot \vec{n} dA - \iint_{\partial V_c} p \vec{v}_c \cdot \vec{n} dA \\ &\quad + \iint_{\partial V_c} \vec{v} \cdot (\vec{n} \cdot \vec{\tau}) dA - \iint_{\partial V_c} \vec{j}_Q \cdot \vec{n} dA \end{aligned} \quad (11.342)$$

where the first term on the right side is the convected energy plus the pressure work on the volume. Insertion of  $h = u + p/\rho$  gives

$$\underbrace{\frac{d}{dt} \iiint_{V_c} \rho(e) dV}_{\begin{array}{l} \text{rate of change} \\ \text{of energy} \\ \text{in control volume} \end{array}} = - \underbrace{\iint_{\partial V_c} \rho \left( h + \frac{1}{2} \vec{v}^2 + \phi \right) (\vec{v} - \vec{v}_c) \cdot \vec{n} dA}_{\begin{array}{l} \text{convected enthalpy,} \\ \text{kinetic energy and} \\ \text{potential energy} \end{array}} - \underbrace{\iint_{\partial V_c} p \vec{v}_c \cdot \vec{n} dA}_{\begin{array}{l} \text{pressure work} \\ \text{due to change in} \\ \text{control volume} \end{array}} + \underbrace{\iint_{\partial V_c} \vec{v} \cdot (\vec{\tau} \cdot \vec{n}) dA}_{\begin{array}{l} \text{viscous work} \\ \text{on volume surface} \end{array}} - \underbrace{\iint_{\partial V_c} \vec{j}_Q \cdot \vec{n} dA}_{\begin{array}{l} \text{heat} \\ \text{conduction} \end{array}} \quad (11.343)$$

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  $V_c$ .

# Chapter 12

## Gas dynamics

### 12.1 Introduction

The balance equations must be combined with thermodynamic results in the modeling of systems where gas flow plays an important part. This is the case in the modeling of engines, gas turbines and compressors, and for pipeline dynamics in the production and transport of oil and gas. Engines, turbines and compressors are designed so that losses due to viscosity and heat conduction are kept at a minimum. It will be shown in the following that such losses will be associated with the generation of entropy in the system. Because of this a well designed and optimized system will typically have gas dynamics that are close to isentropic. On background of this observation the concept of isentropic gas dynamics will be developed in the beginning of this chapter, and the results on isentropic gas dynamics will be used to formulate balance equations for important systems with gas dynamics.

### 12.2 Energy, enthalpy and entropy

#### 12.2.1 Energy

We consider a gas of mass  $m$  in a control volume  $V$ . The energy of the gas in the mass element  $dm$  is called the *specific energy*  $e$ , which is supposed to be given by

$$e = u + \frac{1}{2}\vec{v}^2 + \phi \quad (12.1)$$

where  $u$  is the *specific internal energy*,  $\frac{1}{2}\vec{v}^2$  is the *specific kinetic energy*, and  $\phi$  is the *specific potential energy*. The energy  $E$  of the total volume is given by

$$E = \iiint_V e \rho dV = \iiint_V u \rho dV + \iiint_V \frac{1}{2}\vec{v}^2 \rho dV + \iiint_V \phi \rho dV \quad (12.2)$$

$$= U + K + \Phi \quad (12.3)$$

where  $U$  is the internal energy,  $K$  is the kinetic energy and  $\Phi$  is the potential energy.

#### 12.2.2 Enthalpy

Enthalpy appears in convection terms in the energy balance to account for the convection of internal energy plus pressure work. This is seen in Example 169 and in equation

(11.343). The specific enthalpy is of great importance in gas flow problems where both internal energy and pressure work appear. This is the case in the modeling of compressors and turbines used in turbochargers, jet engines and power plants, and in the description of pipelines for gas transport. Note that enthalpy is not a conserved quantity.

The inverse of density  $\rho$  appears in many equations, and it is customary to define the *specific volume*  $\hat{V}$  as volume per mass unit, which is the inverse of density. In accordance with the notation used for other specific quantities, we would have liked to use  $v$  for specific volume, but we already used  $v$  for velocity. Therefore we use the notation

$$\hat{V} = \frac{1}{\rho} \quad (12.4)$$

The specific enthalpy is defined by

$$h = u + p\hat{V} \quad (12.5)$$

where  $\hat{V} = 1/\rho$  is the specific volume. Enthalpy is not a conserved quantity.

From the definition of specific enthalpy (12.5) we see that

$$dh = du + pd\hat{V} + \hat{V}dp \quad (12.6)$$

### 12.2.3 Specific heats

The internal energy for an ideal gas is a function of the temperature, and we define the specific heat  $c_v(T)$  by

$$du = c_v(T)dT \quad (12.7)$$

In many cases  $c_v(T)$  will be a constant, and in that case a change in the specific internal energy is given by

$$\Delta u = c_v \Delta T \quad (12.8)$$

For an ideal gas we have  $p\hat{V} = RT$ , where  $R$  is the universal gas constant, and it follows that

$$h = u + RT, \quad \text{ideal gas} \quad (12.9)$$

This implies that also the specific enthalpy will be a function of temperature for an ideal gas. We define the specific heat  $c_p(T)$  by

$$dh = c_p(T)dT \quad (12.10)$$

From (12.9) it follows that

$$dh = du + RdT, \quad \text{ideal gas} \quad (12.11)$$

and, consequently, for any gas that satisfy the ideal gas law we have the relation

$$c_p(T) = c_v(T) + R, \quad \text{ideal gas} \quad (12.12)$$

If  $c_v(T)$  is a constant, then a change in enthalpy is given by

$$\Delta h = c_p \Delta T \quad (12.13)$$

We define  $\kappa$  to be the ratio between  $c_p$  and  $c_v$ :

$$\kappa := \frac{c_p}{c_v} \quad (12.14)$$

The numerical value for air is  $\kappa = 1.4$ .

**Example 195** For an ideal gas we have the following useful expressions:

$$\frac{R}{c_v} = \frac{c_p - c_v}{c_v} = \frac{\frac{c_p}{c_v} - 1}{1} = \kappa - 1, \quad \text{ideal gas} \quad (12.15)$$

and

$$\frac{R}{c_p} = \frac{c_p - c_v}{c_p} = \frac{\frac{c_p}{c_v} - 1}{\frac{c_p}{c_v}} = \frac{\kappa - 1}{\kappa}, \quad \text{ideal gas} \quad (12.16)$$

### 12.2.4 Entropy

The concept of entropy provides us with a number of useful modeling tools for thermodynamic systems. The reason for this is that in engines, turbines and compressors, entropy can be seen as a book-keeping tool for certain energy-loss phenomena. In particular, if a component is designed so that no entropy is being generated in connection with gas flow, then this means that there is no energy loss due to viscosity or heat conduction. This will be highly desirable for most system components as this will typically imply that the thermal efficiency is high. If a process is designed so that no entropy is being generated, then the process is said to be isentropic. The design of systems that are isentropic or close to isentropic will therefore be the goal of the mechanical design. Then, as a successful design will be isentropic or close to isentropic, it can often be assumed in the development of the dynamic modeling that the gas flow is isentropic. This is very useful in the development of the balance equations. It is interesting to note that in this context, the important question for the control engineer is whether entropy is being generated or not, while the absolute value of the entropy is not an issue. Therefore, we will describe entropy in terms of its differential, and we will not be concerned about its absolute value.

The entropy of a mass element  $dm$  is called the *specific entropy* and is denoted  $s$ .

The specific entropy is defined in terms of its differential  $ds$  by

$$Tds = du + pd\hat{V} \quad (12.17)$$

Combination of this definition and (12.6) gives the following alternative expression:

$$Tds = dh - \hat{V}dp \quad (12.18)$$

### 12.2.5 The entropy equation

The entropy equation which will be derived in this section is quite interesting, as it gives a mathematical formulation of the second law of thermodynamics. This material is included to make it easier for the reader to understand the physical interpretation of entropy in the setting of the present chapter. The entropy equation is derived using the mass balance and the equation for internal energy. In addition, the derivation relies on the fact that certain forms of energy transfer can only run in one direction. In particular, heat flow is always in the direction of decreasing temperature, and internal work due to viscosity in combination with velocity gradients will always lead to a reduction of kinetic energy and an increase of internal energy.

From the definition (12.17) of entropy we have

$$T \frac{Ds}{Dt} = \frac{Du}{Dt} + p \frac{D\hat{V}}{Dt} \quad (12.19)$$

The continuity equation (11.5) gives

$$\rho \frac{D\hat{V}}{Dt} = \rho \frac{D}{Dt} \left( \frac{1}{\rho} \right) = -\frac{1}{\rho} \frac{D\rho}{Dt} = (\vec{\nabla} \cdot \vec{v}) \quad (12.20)$$

Insertion of this equation and (11.337) into (12.19) gives the following form of the entropy equation:

$$\underbrace{\rho \frac{Ds}{Dt}}_{\substack{\text{rate of change} \\ \text{in entropy in} \\ \text{material volume} \\ \text{element}}} = - \underbrace{\frac{1}{T} \vec{\nabla} \cdot \vec{j}_Q}_{\substack{\text{change in entropy} \\ \text{due to heat} \\ \text{conduction}}} + \underbrace{\frac{1}{T} \vec{\tau} : (\vec{\nabla} \vec{v})}_{\substack{\text{increase in entropy} \\ \text{from irreversible} \\ \text{viscous dissipation}}} \quad (12.21)$$

This can be further developed into a formulation which offers a more detailed interpretation by applying the product rule for differentiation to the first term on the right side.

This entropy equation is given by

$$\underbrace{\rho \frac{Ds}{Dt}}_{\substack{\text{rate of change} \\ \text{in entropy in} \\ \text{material volume} \\ \text{element}}} = - \underbrace{\vec{\nabla} \cdot \left( \frac{\vec{j}_Q}{T} \right)}_{\substack{\text{change in entropy} \\ \text{due to heat flow}}} - \underbrace{\frac{1}{T^2} \vec{j}_Q \cdot \vec{\nabla} T}_{\substack{\text{increase in entropy} \\ \text{due to heat} \\ \text{conduction within} \\ \text{volume element}}} + \underbrace{\frac{1}{T} \vec{\tau} : (\vec{\nabla} \vec{v})}_{\substack{\text{increase in entropy} \\ \text{from irreversible} \\ \text{viscous dissipation} \\ \text{within volume element}}} \quad (12.22)$$

The physical interpretation offered in the equation is supported by the integral form

$$\frac{D}{Dt} \iiint_V \rho s dV = - \iint_{\partial V} \left( \frac{\vec{j}_Q}{T} \right) \cdot \vec{n} dA + \iiint_V \left( -\frac{1}{T^2} \vec{j}_Q \cdot \vec{\nabla} T + \frac{1}{T} \vec{\tau} : (\vec{\nabla} \vec{v}) \right) dV \quad (12.23)$$

An illustration of the different phenomena is given in Figure 12.1. The heat generation  $\vec{\tau} : (\vec{\nabla} \vec{v})$  due to viscous dissipation of mechanical energy will always be positive. Moreover, it has been established that the heat flow will always have a negative component along the temperature gradient. These two results are summed up as

$$\vec{\tau} : (\vec{\nabla} \vec{v}) \geq 0, \quad \vec{j}_Q \cdot \vec{\nabla} T \leq 0 \quad (12.24)$$

Using the inequalities of (12.24) we then get the *Clausius-Duhem inequality* (Lin and Segel 1974) which we state both in differential and integral form:

$$\rho \frac{Ds}{Dt} \geq -\vec{\nabla} \cdot \left( \frac{\vec{j}_Q}{T} \right) \quad (12.25)$$

$$\frac{D}{Dt} \iiint_V \rho s dV \geq - \iint_{\partial V} \left( \frac{\vec{j}_Q}{T} \right) \cdot \vec{n} dA \quad (12.26)$$

The Clausius-Duhem inequality is often referred to as the second law of thermodynamics. Note, however, that a more precise formulation of the second law of thermodynamics is either of the equations (12.22) or (12.23) together with (12.24).

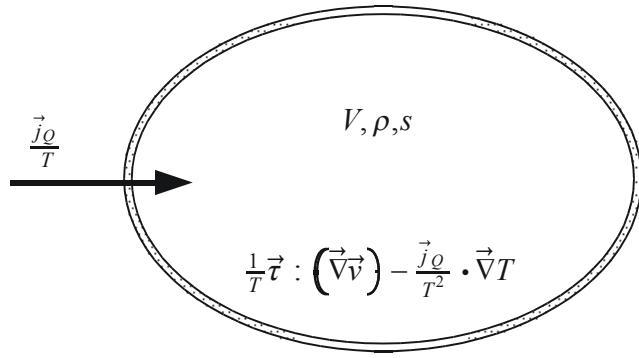


Figure 12.1: The rate of change of entropy in the material control volume  $V$  depends on the entropy flow  $\mathbf{j}_Q/T$  into the material volume due to heat flow over the boundary, and on the internal entropy production terms  $-\mathbf{j}_Q^T \nabla T/T^2$  and  $\boldsymbol{\tau} : \nabla \mathbf{v}^T/T$ , where the first term is related to temperature gradients and heat flow, and the second term is due to velocity gradients in combination with viscosity.

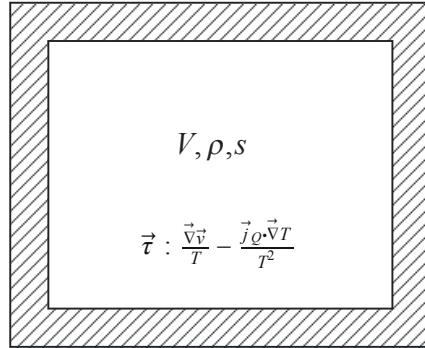


Figure 12.2: Isolated tank of volume  $V$ . The internal entropy production terms are indicated.

**Example 196** Consider a constant control volume  $V$  (Figure 12.2) filled with gas and with no exchange of gas with the outside. The walls of the volume are assumed to be isolated so that there is no heat conduction over the boundary  $\partial V$ . Then (12.26) give

$$\frac{D}{Dt} \iiint_V \rho s dV \geq 0 \quad (12.27)$$

which shows that the total entropy is constant or increasing. This result can be made more precise by using equation (12.23) which gives

$$\frac{D}{Dt} \iiint_V \rho s dV = - \iiint_V \frac{1}{T^2} \vec{j}_Q \cdot \vec{\nabla} T dV + \iiint_V \frac{1}{T} \vec{\tau} : (\vec{\nabla} \vec{v}) dV \quad (12.28)$$

We see from this equation and (12.24) that the entropy of the gas in an isolated tank with no gas exchange with the outside cannot decrease, and that the entropy will increase if there are either temperature gradients  $\vec{\nabla} T$ , or velocity gradients  $\vec{\nabla} \vec{v}$ .

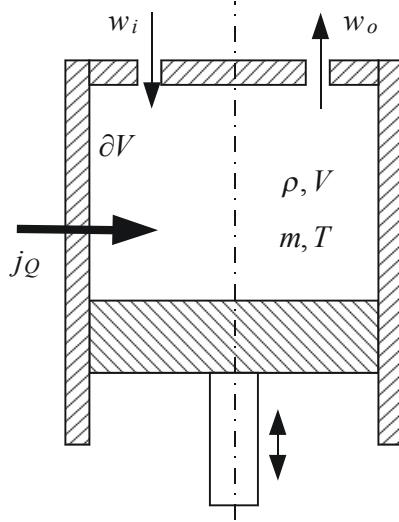


Figure 12.3: Cylinder with variable volume  $V$  due to a moving piston. The inlet mass flow in  $w_i$ , the outlet massflow is  $w_o$ , and the heat flow into the volume is  $j_Q$ .

**Example 197** In Fourier's law of heat conduction

$$\vec{j}_Q = -k\vec{\nabla}T \quad (12.29)$$

the heat flow is proportional to the negative temperature gradient. This gives

$$\vec{j}_Q \cdot \vec{\nabla}T = -k(\vec{\nabla}T)^2 \quad (12.30)$$

which clearly is negative.

### 12.2.6 Internal energy equation in terms of temperature

We consider a cylinder of variable volume  $V = Ax$  filled with an ideal gas. The temperature  $T$  and the pressure  $p$  are assumed to be constant over the volume. Gas with temperature  $T_i$  and pressure  $p_i$  is flowing into the volume with mass flow  $w_i = \rho_i q_i$ , where  $q_i$  is the volumetric flow and  $\rho_i$  is the density. Gas is flowing out of the volume with mass flow  $w_o = \rho q_o$  where  $q_o$  is the volumetric flow out of the volume. The mass balance is

$$\frac{d}{dt}m = w_i - w_o \quad (12.31)$$

We assume that the kinetic and potential energy is zero so that  $E = U = mu$ . Then the energy balance is

$$\frac{d}{dt}(mu) = h_i w_i - h w_o - p A \dot{x} + j_Q \quad (12.32)$$

where  $j_Q$  is the heat flow. We assume that  $c_v$  and  $c_p$  are constants, so that  $u = c_v T$  and  $h = c_p T$ . This gives

$$\frac{dm}{dt}c_v T + mc_v \dot{T} = c_p T_i w_i - c_p T w_o - p A \dot{x} + j_Q \quad (12.33)$$

Insertion of the mass balance (12.31) gives

$$c_v T (w_i - w_o) + mc_v \dot{T} = c_p T_i w_i - c_p T w_o - p A \dot{x} + j_Q \quad (12.34)$$

which leads to

$$mc_v \dot{T} = c_p w_i T_i - [c_v w_i + (c_p - c_v) w_o] T - p A \dot{x} + j_Q \quad (12.35)$$

Finally, it is noted that for an ideal gas,  $pV = mRT$ , and therefore  $pA = mRT/x$ . Insertion of this and the use of (12.15) leads to the energy balance in the form

$$\dot{T} = \frac{\kappa w_i}{m} T_i - \frac{w_i + (\kappa - 1) w_o}{m} T - (\kappa - 1) \frac{\dot{x}}{x} T + \frac{1}{mc_v} j_Q \quad (12.36)$$

### 12.2.7 Energy balance in terms of pressure

For an ideal gas with constant  $c_v$  the internal energy can be written

$$U = mc_v T = \frac{c_v}{R} p V = \frac{1}{\kappa - 1} p V. \quad (12.37)$$

The energy balance is assumed to be

$$\frac{dU}{dt} = h_i w_i - h w_o - p A \dot{x} + j_Q \quad (12.38)$$

and we get

$$\frac{1}{\kappa - 1} (pV + p\dot{V}) = h_i w_i - h w_o - p A \dot{x} + j_Q \quad (12.39)$$

which is simplified to

$$\dot{p}V = (\kappa - 1) (h_i w_i - h w_o + j_Q) - \kappa p A \dot{x}. \quad (12.40)$$

The specific enthalpy can be written

$$h = c_p T = \frac{c_p}{R} p \hat{V} = \frac{\kappa}{\kappa - 1} p \hat{V} \quad (12.41)$$

which gives

$$\dot{p}V = \kappa \left( p_i \hat{V}_i w_i - p \hat{V} w_o \right) + (\kappa - 1) j_Q - \kappa p A \dot{x} \quad (12.42)$$

We introduce the input volumetric flow  $q_i = \hat{V}_i w_i$  and the output volumetric flow  $q_o = \hat{V} w_o$  and get the expression

$$\dot{p} = -\kappa \frac{q_o}{V} p + \kappa \frac{q_i}{V} p_i - \kappa \frac{\dot{x}}{x} p + (\kappa - 1) \frac{j_Q}{V}. \quad (12.43)$$

**Example 198** An alternative way of writing this is

$$\frac{V}{\kappa p} \dot{p} = -q_o + q_i \frac{p_i}{p} - A \dot{x} + \frac{(\kappa - 1)}{\kappa p} j_Q \quad (12.44)$$

which is reminiscent of the mass balance for a hydraulic control volume.

### 12.2.8 Piston motion

The equation of motion for the piston is

$$m_p \ddot{x}_p = pA - F_L \quad (12.45)$$

where  $F_L$  is the load force. Suppose that the piston is connected to a crankshaft with a rod, and that the kinematic relation between the piston velocity  $\dot{x}$  and the crankshaft velocity  $\omega_{cs}$  is given by

$$\dot{x} = r(\theta_{cs})\omega_{cs} \quad (12.46)$$

where  $\theta_{cs}$  is the crankshaft angle. Then the equation of motion for the crankshaft is

$$J\dot{\omega}_{cs} = Fr(\theta) - T_L \quad (12.47)$$

where  $T_L$  is the load torque. The crankshaft is a mechanical two-port where one port has input  $F$  and output  $\omega_{cs}$ , and one port with input  $T_L$  and output  $\omega_{cs}$ .

Flexibility in the crankshaft can easily be included. This is done by inserting an elastic transmission and an inertia which is modelled as a mechanical two-port given by

$$J_1\dot{\omega}_1 = T_L - T_1 \quad (12.48)$$

$$T_L = D_1(\omega_m - \omega_1) + K_1(\theta_m - \theta_1). \quad (12.49)$$

where  $\omega_1$  is the shaft speed of the load shaft. This means that 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  $T_L$  and flow  $\omega_m$ , while the output port has effort  $T_1$  and flow  $\omega_1$ .

## 12.3 Isentropic conditions

### 12.3.1 Isentropic processes

*Isentropic processes* are important in a wide range of applications. Isentropic processes are processes where there is no entropy production in the sense that  $ds = 0$ . In view of (12.21) isentropic processes can occur if there is no heat conduction and no internal viscous work. From (12.17) and (12.18) we see that  $ds = 0$  implies

$$du = -pd\hat{V} \quad (12.50)$$

$$dh = \hat{V}dp \quad (12.51)$$

For gases that satisfy the ideal gas law  $p\hat{V} = RT$ , this can be written

$$c_v(T)dT = -\frac{RT}{\hat{V}}d\hat{V} \quad (12.52)$$

$$c_p(T)dT = \frac{RT}{p}dp \quad (12.53)$$

which leads to

$$\frac{dT}{T} = -\frac{R}{c_v(T)}\frac{d\hat{V}}{\hat{V}} \quad (12.54)$$

$$\frac{dT}{T} = \frac{R}{c_p(T)}\frac{dp}{p} \quad (12.55)$$

In view of (12.15) and (12.16) this gives the differential isentropic relations

$$\frac{dT}{T} = -[\kappa(T) - 1] \frac{d\hat{V}}{\hat{V}} \quad (12.56)$$

$$\frac{dT}{T} = \frac{\kappa(T) - 1}{\kappa(T)} \frac{dp}{p} \quad (12.57)$$

and, by combination of the two equations, the differential isentropic relation

$$\frac{dp}{p} = -\kappa(T) \frac{d\hat{V}}{\hat{V}} \quad (12.58)$$

Consider the differential equation

$$\frac{dx}{x} = a \frac{dy}{y} \Rightarrow \ln x = a \ln y + C_1 \Rightarrow e^{\ln x} = C (e^{\ln y})^a \quad (12.59)$$

$$\Rightarrow x = Cy^a \quad (12.60)$$

Thus, if  $x_1 = Cy_1^a$  is one solution, then the constant  $C$  is found from  $C = x_1/y_1^a$ , and any solution  $x_2 = Cy_2^a$  must satisfy

$$\frac{x_2}{x_1} = \left( \frac{y_2}{y_1} \right)^a \quad (12.61)$$

A reasonable assumption is that  $\kappa$  is a constant. Then we see from this derivation that (12.56, 12.57, 12.58) leads to the *isentropic relations*

$$\frac{T_1}{T_2} = \left( \frac{\hat{V}_2}{\hat{V}_1} \right)^{\kappa-1} \quad (12.62)$$

$$\frac{T_1}{T_2} = \left( \frac{p_1}{p_2} \right)^{\frac{\kappa-1}{\kappa}} \quad (12.63)$$

$$\frac{p_1}{p_2} = \left( \frac{\hat{V}_2}{\hat{V}_1} \right)^{\kappa} \quad (12.64)$$

**Example 199** The identity  $\rho\hat{V} = 1$  implies

$$\frac{d\rho}{\rho} = -\frac{d\hat{V}}{\hat{V}} \quad (12.65)$$

For isentropic processes (12.58) gives

$$dp = \kappa(T) \frac{p}{\rho} d\rho \quad (12.66)$$

For an ideal gas this leads to the result

$$dp = c^2 d\rho \quad (12.67)$$

where  $c = \sqrt{\kappa RT}$ , which we will see is the speed of sound.

**Example 200** The mass balance of a gas flowing through a fixed volume  $V$  with mass flow  $w_{in}$  into the volume and mass flow  $w_{out}$  out of the volume is

$$V\dot{\rho} = w_{in} - w_{out} \quad (12.68)$$

where  $\rho$  is the density of the gas. If the gas is ideal and the conditions are isentropic, then (12.67) applies, and the mass balance becomes

$$\frac{V}{c^2}\dot{p} = w_{in} - w_{out} \quad (12.69)$$

### 12.3.2 Stagnation state

In the modeling of flow the concept of a *stagnation state* is useful. The *stagnation state* is characterized by a *stagnation velocity*  $v_0 = 0$ , the *specific stagnation enthalpy*  $h_0$ , the *stagnation temperature*  $T_0$ , and the *stagnation pressure*  $p_0$ .

The specific stagnation enthalpy

$$h_0 = h + \frac{v^2}{2} \quad (12.70)$$

is defined as the sum of the specific enthalpy  $h$  and the specific kinetic energy  $v^2/2$ . The stagnation temperature  $T_0$  is the temperature corresponding to the specific stagnation enthalpy  $h_0$  in the sense that

$$T_0 - T = \frac{h_0 - h}{c_p} \quad (12.71)$$

when  $c_p$  is constant. The stagnation pressure  $p_0$  is defined by

$$\frac{p_0}{p} = \left(\frac{T_0}{T}\right)^{\frac{\kappa}{\kappa-1}} \quad (12.72)$$

From the definitions (12.71) and (12.72) it is seen that the stagnation temperature and the stagnation pressure can be found from

$$T_0 = T + \frac{v^2}{2c_p}, \quad p_0 = p \left(1 + \frac{v^2}{2c_p T}\right)^{\frac{\kappa}{\kappa-1}} \quad (12.73)$$

If a fluid with enthalpy  $h$ , temperature  $T$ , pressure  $p$  and velocity  $v$  is slowed down to zero velocity in an isentropic process, then the gas will be in the stagnation state with stagnation velocity  $v_0 = 0$ , stagnation enthalpy  $h_0$ , stagnation temperature  $T_0$ , and stagnation pressure  $p_0$ . To make a clear distinction between the pressure  $p$  and the stagnation pressure  $p_0$  it is customary to refer to  $p$  as the *static pressure*.

### 12.3.3 Energy balance for isentropic processes

In this section we investigate which condition that must be fulfilled for the energy balance to describe an isentropic process. We start with the relation

$$dT = -\frac{p}{c_v}d\hat{V} \quad (12.74)$$

which follows from  $Tds = c_vdT + pd\hat{V} = 0$ . Division with  $dt$  gives

$$\dot{T} = -\frac{p}{c_v}\frac{d}{dt}\left(\frac{V}{m}\right) = -\frac{p}{mc_v}\dot{V} + \frac{pV}{m^2c_v}\dot{m} \quad (12.75)$$

Then the ideal gas law  $pV = mRT$  gives

$$\dot{T} = (\kappa - 1) \left( \frac{\dot{m}}{m} - \frac{\dot{V}}{V} \right) T \quad (12.76)$$

The energy balance (12.36) satisfies this equation if the heat flux  $j_Q$  is zero, and the temperature of the mass flow into the volume has the same temperature as the temperature in the volume.

#### 12.3.4 The speed of sound

The derivation of the speed of sound is based on the following mass and momentum balances for a differential control volume:

$$\frac{\partial \rho}{\partial t} = -\vec{\nabla} \cdot (\rho \vec{v}) \quad (12.77)$$

$$\rho \frac{D \vec{v}}{D t} = -\vec{\nabla} p \quad (12.78)$$

The change of variable in the first equation from  $\rho$  to  $p$  requires an expression for  $dp$  as a function of  $d\rho$ . This is obtained by assuming isentropic conditions. For an ideal gas the ideal gas law  $p = \rho RT$  under isentropic conditions leads to

$$dp = \kappa R T d\rho \quad (12.79)$$

This gives

$$\frac{\partial p}{\partial t} = -\kappa R T \vec{\nabla} \cdot (\rho \vec{v}) \quad (12.80)$$

Linearization around  $(\rho_0, T_0, \vec{v}_0 = \vec{0})$  gives

$$\frac{\partial p}{\partial t} = -c^2 \rho_0 \vec{\nabla} \cdot \vec{v} \quad (12.81)$$

$$\rho_0 \frac{\partial \vec{v}}{\partial t} = -\vec{\nabla} p \quad (12.82)$$

where  $c^2 = \kappa R T_0$ . Then, time differentiation of the linearized pressure equation gives

$$\begin{aligned} \frac{\partial^2 p}{\partial t^2} &= -c^2 \rho_0 \frac{\partial}{\partial t} (\vec{\nabla} \cdot \vec{v}) \\ &= -c^2 \vec{\nabla} \cdot \rho_0 \frac{\partial \vec{v}}{\partial t} \end{aligned}$$

Insertion of the linearized velocity equation gives the wave equation

$$\frac{\partial^2 p}{\partial t^2} = c^2 \vec{\nabla}^2 p \quad (12.83)$$

where  $c = \sqrt{\kappa R T_0}$ .

d'Alembert's solution of the wave equation (12.83) (Kreyszig 1979) is in the form

$$p(x, t) = f(x + ct) + g(x - ct) \quad (12.84)$$

where  $f(x + ct) + g(x - ct)$  are required to satisfy initial conditions and boundary conditions. To verify that this is a solution to the wave equation we calculate

$$\frac{\partial^2}{\partial t^2} [f(x + ct) + g(x - ct)] = c^2 [f''(x + ct) + g''(x - ct)] \quad (12.85)$$

and

$$\frac{\partial^2}{\partial x^2} [f(x + ct) + g(x - ct)] = f''(x + ct) + g''(x - ct) \quad (12.86)$$

The function  $f(x + ct)$  can be interpreted as a wave that goes in the negative  $x$  direction with velocity  $-c$ , while  $g(x - ct)$  can be interpreted as a wave moving in the positive  $x$  direction with velocity  $c$ . From this we can conclude that the speed of sound is equal to  $c = \sqrt{\kappa RT}$ . For air at 15 degrees Centigrade this gives 340 m/s, which agrees with experimental data.

**Example 201** In the works of Sir Isaac Newton the speed of sound was derived under conditions that correspond to isothermal conditions. For ideal gases this gives  $dp = RTd\rho$  and the wave equation becomes

$$\frac{\partial^2 p}{\partial t^2} = \bar{c}^2 \vec{\nabla}^2 p \quad (12.87)$$

The speed of sound according to this result is  $\bar{c} = \sqrt{RT}$ , which for air at 15 degrees Centigrade gives 287 m/s. This was not in agreement with what was observed in experiments.

### 12.3.5 Helmholtz resonator

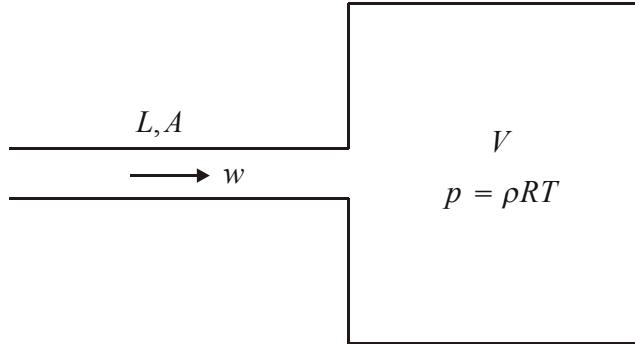


Figure 12.4: Helmholtz resonator

A Helmholtz resonator is a pipe of length  $L$  and cross section  $A$  which is open at one end, and which is connected to a volume  $V$ , called the plenum, at the other end. This is shown in Figure 12.4. It is assumed that the compressibility effects in the pipe are negligible, which means that it is assumed that the pressure  $p$  and the mass flow  $w$  do not vary along the pipe. Further, it is assumed that the velocity in the plenum can be approximated by zero. This means that the conditions in the plenum are isentropic. This implies  $dp = c^2 d\rho$  where  $c = \sqrt{\kappa RT}$  is the sonic speed. The dynamics of the Helmholtz

resonator is derived from the mass balance of the plenum and the momentum balance of the pipe. The model is found to be

$$\frac{V}{c^2} \frac{dp}{dt} = w \quad (12.88)$$

$$L \frac{dw}{dt} = -pA \quad (12.89)$$

The model can be further developed by differentiating the mass balance with respect to time, and by inserting the momentum balance. This gives

$$\frac{d^2p}{dt^2} + \omega_H^2 p = 0 \quad (12.90)$$

Here  $\omega_H$  is the Helmholtz frequency defined by

$$\omega_H = c \sqrt{\frac{A}{VL}} \quad (12.91)$$

**Example 202** Consider a Helmholtz resonator with a pressurizing device on the input that supplies air at a mass flow  $w$  with a pressure  $p_1 = p_1(w)$ . In addition, air is flowing out of the plenum with mass flow  $w_2 = w_2(p)$ . The model is then

$$\frac{dp}{dt} = \frac{c^2}{V} [w - w_2(p)] \quad (12.92)$$

$$\frac{dw}{dt} = \frac{A}{L} [p_1(w) - p] \quad (12.93)$$

Linearization gives

$$\frac{d\Delta p}{dt} = \frac{c^2}{V} \Delta w - \frac{c^2}{V} \frac{dw_2}{dp} \Delta p \quad (12.94)$$

$$\frac{d\Delta w}{dt} = -\frac{A}{L} \Delta p + \frac{A}{L} \frac{dp_1}{dw} \Delta w \quad (12.95)$$

For simplicity it is assumed that  $dw_2/dp \approx 0$ . Then the characteristic equation is

$$\lambda^2 - \frac{A}{L} \frac{dp_1}{dw} \lambda + \omega_H^2 = 0 \quad (12.96)$$

It is clear that the system will be stable if  $dp_1/dw \leq 0$ . Therefore, if  $p_1(w)$  has a positive slope, then the system becomes unstable (Figure 12.5).

## 12.4 Acoustic resonances in pipes

### 12.4.1 Dynamic model

Acoustic waves for an ideal gas are described by the wave equation

$$\frac{\partial^2 p}{\partial t^2} = c^2 \nabla^2 p \quad (12.97)$$

where  $c = \sqrt{\kappa RT}$  under the assumption of isentropic conditions. This is derived as for fluids. The wave variables defined in Section 4.5.7 will satisfy the transfer functions equations

$$a_2(s) = \exp(-Ts)a_1(s), \quad b_1(s) = \exp(-Ts)b_2(s) \quad (12.98)$$

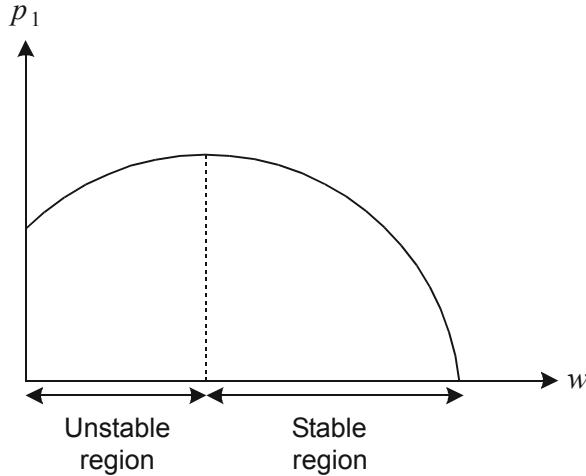


Figure 12.5: Pressure  $p_1$  as a function of massflow  $w$ . The unstable and stable regions are indicated.

and the boundary conditions

$$a_1(s) = g_1(s)b_1(s), \quad b_2(s) = g_2(s)a_2(s) \quad (12.99)$$

Combining these equations we get

$$[1 - \exp(-2Ts) g_1(s)g_2(s)] a_1(s) = 0 \quad (12.100)$$

### 12.4.2 Pipe closed at both ends

We consider acoustic waves in a pipe of length  $\ell$ , and describe the dynamics of the gas by the pressure  $p$  and the volumetric flow  $q$  through the pipe. The wave variables  $a$  and  $b$  are defined as for fluid flow through pipes. The boundary conditions are set to

$$q(0, s) = q(\ell, s) = 0 \quad (12.101)$$

which means that the pipe is closed at both ends. The impedances at the two ends are defined by  $p(0, s) = z_1(s)q(0, s)$  and  $p(\ell, s) = z_2(s)q(\ell, s)$ . The boundary conditions then correspond to impedances  $z_1 = \infty$  and  $z_2 = \infty$ . This leads to  $g_1 = g_2 = 1$  and  $b_1 = a_1$ ,  $a_2 = b_2$ , which gives

$$a_1(s) = \exp(-2Ts) a_1(s) \quad (12.102)$$

or

$$[1 - \exp(-2Ts)] a_1(s) = 0 \quad (12.103)$$

It is straightforward to check that the solution is the same if the pipe is open at both ends. This system has singularities whenever

$$\exp(-2Ts) = 1 \quad (12.104)$$

which means that the system has infinitely many singularities in

$$s = j \frac{k\pi}{T}, \quad k = 0, \pm 1, \pm 2, \dots \quad (12.105)$$

For each singularity there is an undamped resonance. The first resonance is at

$$s = j \frac{\pi}{T} \quad (12.106)$$

which corresponds to a frequency of

$$f = \frac{1}{2\pi} \frac{\pi}{T} = \frac{1}{2T} = \frac{c}{2\ell} \quad (12.107)$$

For air at 15 degrees Centigrade (288 K) the speed of sound is  $c = 340$  m/s. A pipe of length 10 m then gives a resonance frequency at  $f = 340/20 = 17$  Hz. To get a frequency of 440 Hz the pipe must be  $\ell = 340/(2 \cdot 440) = 0.386$  m. To get 256 Hz the pipe must be 0.66 m.

### 12.4.3 Pipe closed at one end

The boundary condition is changed to

$$p(0, s) = 0 \quad (12.108)$$

$$q(\ell, s) = 0 \quad (12.109)$$

which gives  $g_1 = -1$  and  $g_2 = 1$ . Then the system equation becomes

$$[1 + \exp(-2Ts)] a(s) = 0 \quad (12.110)$$

and the singularities are found at

$$s = j \frac{\pi + 2k\pi}{2T}, \quad k = 0, \pm 1, \pm 2, \dots \quad (12.111)$$

The first resonance is now at

$$f = \frac{1}{4T} = \frac{c}{4\ell} \quad (12.112)$$

To get 440 Hz at 288 K the pipe must then be 0.193 m, while 55 Hz results from a pipe of length 3.145 m.

Consider a pipe of length  $L$  and cross section  $A$  which is open at the inlet end, denoted by subscript 1, and which ends in a volume  $V$  at the outlet end denoted by subscript 2. It is assumed that the pressure is constant over the volume  $V$ , and that this pressure is equal to the pressure  $p_2 = p(\ell)$  at the outlet of the pipe. In this case the boundary conditions are

$$p(0, s) = 0 \quad (12.113)$$

$$p(\ell, s) = z_2(s)q(\ell, s) \quad (12.114)$$

The impedance  $z_2(s)$  at the outlet is found from the mass balance of the volume  $V$ . The mass balance is

$$\frac{d(\rho_2 V)}{dt} = \rho_2 q_2 \quad (12.115)$$

which gives

$$V \frac{dp_2}{dt} = \frac{\kappa p_2}{\rho_2} \rho q_2 \quad (12.116)$$

which is simplified to

$$\frac{dp_2}{dt} = \frac{\kappa p_2}{V} q_2 \quad (12.117)$$

and it is seen that for small  $q_2$  the impedance is

$$z_2(s) = \frac{\kappa p_2}{V} \frac{1}{s} \quad (12.118)$$

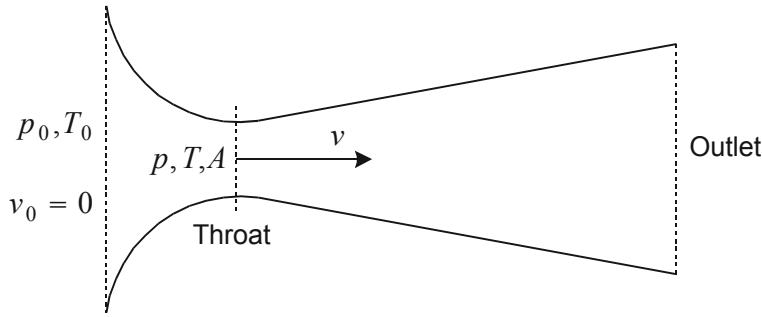


Figure 12.6: Isentropic gas flow from stagnation state  $p_0, T_0$  and velocity  $v_0 = 0$  through a restriction with throat cross section  $A$ . The throat states are denoted  $p$  and  $T$ , and the velocity at the throat is denoted  $v$ .

#### 12.4.4 Pressure measurement in diesel engine cylinder

Due to the high temperatures in a diesel cylinder it is desirable to measure the pressure in the cylinder by placing the pressure sensor in a bored pipe of small diameter. Suppose that the length of the bore is 3 cm, and that the pressure sensor is in the closed end of the pipe, while the open end is on the cylinder face. Then the resonance frequency of the pipe is

$$f = \frac{340}{4 \times 0.03} \sqrt{\frac{1000}{288}} = 5.28 \text{ kHz} \quad (12.119)$$

In a motor running at 6000 rpm, which corresponds to a shaft speed of 36000 degrees per second, the shaft will rotate  $36000/5280 = 6.8$  degrees in one resonance period. This means that accurate pressure measurements can not have a resolution better than about 6.8 degrees of crank-shaft angle under these conditions.

### 12.5 Gas flow

#### 12.5.1 Gas flow through a restriction

Gas flow through a valve or a restriction is usually modelled as *nozzle flow*, which is isentropic flow that is computed using the isentropic relations between the stagnation state in the inlet volume and the state in the throat of the restriction. It is convenient to use the Mach number

$$M = \frac{v}{c} \quad (12.120)$$

in the derivation of the mass flow. This is a dimensionless velocity, where the velocity  $v$  is scaled by the sonic velocity  $c = \sqrt{\kappa RT}$ .

The starting point for the derivation is the relation between the mass flow  $w$  and Mach number  $M = v/c$  which is found from the following calculation:

$$w = \rho A v = \rho A M c = \rho A M \sqrt{\kappa RT} = \frac{A p}{\sqrt{\kappa RT}} M \kappa \quad (12.121)$$

Here the ideal gas law  $p = \rho RT$  is used, and  $A$  is the cross section area of the nozzle.

This can then be written

$$\frac{w\sqrt{\kappa RT}}{Ap} = \kappa M \quad (12.122)$$

where the left hand side is a dimensionless form of the mass flow. To be able to reference the mass flow to the stagnation state 0, we use the expression

$$\frac{w\sqrt{\kappa RT_0}}{Ap_0} = \frac{w\sqrt{\kappa RT}}{Ap} \sqrt{\frac{T_0}{T}} \frac{p}{p_0} = \kappa M \sqrt{\frac{T_0}{T}} \frac{p}{p_0} \quad (12.123)$$

where  $p_0, T_0$  are stagnation states at the inlet. The isentropic relation

$$\frac{T_0}{T} = \left( \frac{p_0}{p} \right)^{(\kappa-1)/\kappa} \quad (12.124)$$

can then be used to find the following expression for isentropic flow of an ideal gas:

$$\frac{w\sqrt{\kappa RT_0}}{Ap_0} = \kappa M \left( \frac{T}{T_0} \right)^{(\kappa+1)/2(\kappa-1)} = \kappa M \left( \frac{p}{p_0} \right)^{(\kappa+1)/2\kappa} \quad (12.125)$$

Next, the steady-state energy equation for nozzle flow is written

$$c_p T_0 = c_p T + \frac{v^2}{2} = c_p T + \frac{M^2 c^2}{2} = c_p T + \frac{M^2 \kappa R T}{2} \quad (12.126)$$

where  $T_0$  is the stagnation temperature at the inlet,  $T$  is the temperature at the throat, while  $v$  is the velocity at the throat. Using  $R/c_p = (\kappa - 1)/\kappa$ , the energy equation can be written in the form

$$\frac{T_0}{T} = 1 + \frac{\kappa - 1}{2} M^2 \quad (12.127)$$

or, equivalently,

$$M = \left[ \frac{2}{\kappa - 1} \left( \frac{T_0}{T} - 1 \right) \right]^{1/2} = \left\{ \frac{2}{\kappa - 1} \left[ \left( \frac{p_0}{p} \right)^{(\kappa-1)/\kappa} - 1 \right] \right\}^{1/2} \quad (12.128)$$

Insertion of this expression for  $M$  in equation (12.125) for mass flow then leads to

$$\frac{w\sqrt{\kappa RT_0}}{Ap_0} = \kappa \left( \frac{p}{p_0} \right)^{(\kappa+1)/2\kappa} \left\{ \frac{2}{\kappa - 1} \left[ \left( \frac{p_0}{p} \right)^{(\kappa-1)/\kappa} - 1 \right] \right\}^{1/2}$$

By straightforward operations this can be rearranged into the well-known expressions for isentropic nozzle flow of an ideal gas.

Isentropic flow through a restriction will have mass flow given by

$$\frac{w\sqrt{\kappa RT_0}}{Ap_0} = \kappa \left( \frac{p}{p_0} \right)^{1/\kappa} \left\{ \frac{2}{\kappa - 1} \left[ 1 - \left( \frac{p}{p_0} \right)^{(\kappa-1)/\kappa} \right] \right\}^{1/2} \quad (12.129)$$

We note that this expressions are valid for  $M \leq 1$ . Sonic flow occurs when  $M = 1$ . It is seen from (12.127) that if  $\kappa = 1.4$  this corresponds to

$$\frac{p}{p_0} \Big|_{\text{sonic}} = \left( \frac{2}{\kappa + 1} \right)^{\frac{\kappa}{\kappa - 1}} = 0.52828 \quad (12.130)$$

The resulting mass flow at sonic conditions is seen from (12.125) to be

$$\frac{w\sqrt{\kappa RT_0}}{Ap_0} \Big|_{\text{sonic}} = \kappa \left( \frac{2}{\kappa + 1} \right)^{\frac{\kappa+1}{2(\kappa-1)}} \quad (12.131)$$

### 12.5.2 Example: Discharge of gas from tank

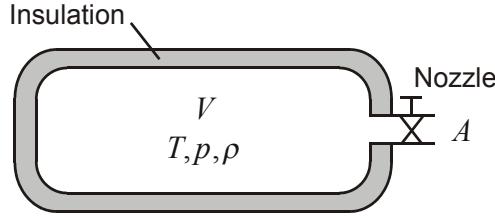


Figure 12.7: Insulated tank with gas discharge

The example of this section is adopted from (Bird et al. 1960, p. 480). A fluid is contained in an insulated tank of volume  $V$  as shown in Figure 12.7. The gas is discharged with mass flow  $w$  through a small nozzle with opening area  $A$ . The nozzle flow is assumed to be isentropic from the stagnation state  $T, p, \rho$  in the tank to the throat state  $T_a, p_a, \rho_a$  at the outlet of the nozzle. The throat state is set equal to the ambient state of the surrounding air, and the throat area is set to be the area  $A$  of the nozzle (Heywood 1988).

The mass balance is

$$\frac{dm}{dt} = -w \quad (12.132)$$

The energy of the gas is assumed to be equal to the internal energy  $U = \rho uV$  where  $u$  is the specific internal energy. The energy balance is then

$$\frac{d}{dt} (mu) = -wu - w\frac{p}{\rho} = -wh \quad (12.133)$$

where  $h = u + p/\rho$  is the specific enthalpy. Using the product rule for differentiation and mass balance we get

$$\dot{u} = -\frac{p}{\rho} \frac{w}{m} \quad (12.134)$$

Finally, we assume that  $u = c_v T$  and that the ideal gas law  $p = \rho RT$  is valid. Then the model can be written

$$\dot{m} = -w \quad (12.135)$$

$$\dot{T} = -(\kappa - 1) \frac{w}{m} T \quad (12.136)$$

where  $\kappa > 1$  and the mass flow is found from the nozzle flow equation to be

$$w = \frac{Ap}{\sqrt{\kappa RT}} \kappa \left( \frac{p_a}{p} \right)^{1/\kappa} \left\{ \frac{2}{\kappa - 1} \left[ 1 - \left( \frac{p_a}{p} \right)^{(\kappa-1)/\kappa} \right] \right\}^{\frac{1}{2}} \quad (12.137)$$

and the throat pressure  $p_a$  is the ambient pressure, and the stagnation pressure  $p$  is found from the ideal gas law to be

$$p = \frac{mRT}{V}. \quad (12.138)$$

From the model it is clear that the time derivatives of mass  $m$  and temperature  $T$  will be less than or equal to zero.

### 12.5.3 The Euler equation around sonic speed

We write the continuity equation (11.6) in the form

$$\frac{\partial \rho}{\partial t} = - (\vec{\nabla} \rho) \cdot \vec{v} - \rho (\vec{\nabla} \cdot \vec{v}) \quad (12.139)$$

and the Euler equation in the form (11.58) with zero mass force:

$$\rho \left[ \frac{\partial \vec{v}}{\partial t} + \vec{v} \cdot (\vec{\nabla} \vec{v}) \right] = -\vec{\nabla} p \quad (12.140)$$

We consider an ideal gas under isentropic conditions, which according to (12.67) implies that

$$\vec{\nabla} p = c^2 \vec{\nabla} \rho \quad (12.141)$$

The Euler equation then becomes

$$\rho \frac{\partial \vec{v}}{\partial t} = -\rho \vec{v} \cdot (\vec{\nabla} \vec{v}) - c^2 \vec{\nabla} \rho \quad (12.142)$$

We take the scalar product between the Euler equation and the velocity and get the following set of equations:

$$\frac{\partial \rho}{\partial t} = - (\vec{\nabla} \rho) \cdot \vec{v} - \rho (\vec{\nabla} \cdot \vec{v}) \quad (12.143)$$

$$\rho \frac{\partial \vec{v}}{\partial t} \cdot \vec{v} = -\rho \vec{v} \cdot (\vec{\nabla} \vec{v}) \cdot \vec{v} - c^2 (\vec{\nabla} \rho) \cdot \vec{v} \quad (12.144)$$

We consider a point, and align our coordinate frame so that the velocity  $\vec{v}$  is along the  $x_1$  axis, that is,

$$\vec{v} = v_1 \vec{a}_1 \quad (12.145)$$

Then, at this point our set of equations become

$$\frac{\partial \rho}{\partial t} = -\rho_{,1} v_1 - \rho v_{1,1} - \rho v_{2,2} - \rho v_{3,3} \quad (12.146)$$

$$\rho \left( \frac{\partial v_1}{\partial t} \right) v_1 = -\rho v_1^2 v_{1,1} - c^2 \rho_{,1} v_1 \quad (12.147)$$

By adding the equations we get

$$\frac{1}{\rho} \frac{\partial \rho}{\partial t} - \frac{1}{c^2} \left( \frac{\partial v_1}{\partial t} \right) v_1 = -(1 - M^2) v_{1,1} - (v_{2,2} + v_{3,3}) \quad (12.148)$$

Under stationary conditions we have

$$v_{2,2} + v_{3,3} = - (1 - M^2) v_{1,1} \quad (12.149)$$

This result has interesting implications if we consider speeds close to the sonic speed. Then the Mach number is close to zero, and in particular, at sonic speed  $M = 1$ , and  $1 - M^2 = 0$ . This gives

$$v_{2,2} + v_{3,3} = 0 \cdot v_{1,1} \quad (12.150)$$

This is satisfied in a flow field where  $v_2 = v_3 = 0$ , which is the case if all particles flow along the  $x_1$  axis. However, if  $M \rightarrow 1$  when  $v_{2,2} + v_{3,3} \neq 0$ , then  $v_{1,1} \rightarrow \infty$ . This means that the model consisting of the Euler equation and the continuity equation has a singularity at  $M = 1$ . This agrees with experimental evidence which shows that if the Mach number  $M$  approaches unity when there are disturbances in the flow patterns so that  $v_{2,2} + v_{3,3} \neq 0$ , then a large gradients in density and velocity occur, and this is called a *shock*. In a shock the assumption of isentropic conditions are no longer valid as there will be significant viscous energy dissipation, and (12.67) will no longer be valid.

# Chapter 13

## Compressor dynamics

### 13.1 Introduction

#### 13.1.1 Compressors

Compressors are used in a wide variety of applications. These includes turbojet engines used in aerospace propulsion, power generation using industrial gas turbines, turbocharging of internal combustion engines, pressurization of gas and fluids in the process industry, transport of fluids in pipelines and so on. Compressors can be divided into four general types: reciprocating, rotary, centrifugal and axial. Some authors use the term radial compressor when referring to a centrifugal compressor. Reciprocating and rotary compressors work by the principle of reducing the volume of the gas, and will not be considered further in this work. Centrifugal and axial compressors, also known as turbocompressors or continuous flow compressors, work by the principle of accelerating the fluid to a high velocity and then converting this kinetic energy into potential energy by decelerating the gas in diverging channels. In axial compressors the deceleration takes place in the stator blade passages, and in centrifugal compressors it takes place in the diffuser. The increase in potential energy of the fluid is manifested by a pressure rise. This conversion can be explained from Bernoulli's equation (11.86), which is written:

$$\frac{p_1}{\rho} + \frac{C_1^2}{2} + gz_1 = \frac{p_2}{\rho} + \frac{C_2^2}{2} + gz_2 \quad (13.1)$$

Here  $p$  is the pressure,  $\rho$  is the density of the fluid,  $C$  is the velocity of the fluid and  $gz$  the potential energy per unit mass. Subscripts 1 and 2 denotes properties before and after deceleration, respectively. The equation is a special case of the law of conservation of energy developed for flowing fluids. Bernoulli's equation states that the sum of kinetic energy  $\frac{C^2}{2}$ , potential energy  $gz$  and pressure head  $\frac{p}{\rho}$  at one set of conditions is equal to their sum at another set of conditions. Hence, decrease in kinetic energy implies an increase in potential energy and pressure. One obvious difference between the two types of turbocompressors is that, in axial compressors, the flow leaves the compressor in the axial direction, whereas in centrifugal compressors, the flows leaves the compressor in a direction perpendicular to the axis of the rotating shaft. Axial compressors can accept higher mass flow rates than centrifugal compressors for a given frontal area. This is one reason for axial compressors dominance in jet engines, where frontal area is of great

importance. Another reason for this is that for gas turbines, or jet engines, specific fuel consumption decreases with increasing pressure ratio. The axial compressor, using multiple stages, can achieve higher pressure ratio and efficiency than the centrifugal.

### 13.1.2 Surge and rotating stall

The useful range of operation of turbocompressors, both axial and centrifugal, is limited by choking at high mass flows when sonic velocity is reached in some component. The mass flow at which choking occurs depends on the rotational speed of the compressor, and the compressor is said to have reached the stone wall area in its operating domain when choking occurs. The mass flow at sonic velocity, that is  $M = v/c = 1$ , can be calculated by using (12.131).

At low mass flows the operation is limited by the onset of two instabilities known as surge and rotating stall. Surge is an axisymmetrical oscillation of the flow through the compressor, and is characterized by a limit cycle in the compressor characteristic. Surge oscillations are in most applications unwanted, and can in extreme cases even damage the compressor. Rotating stall is an instability where the circumferential flow pattern is disturbed. This is manifested through one or more stall cells of reduced, or stalled, flow that propagate around the compressor annulus at a fraction of the rotor speed. There are two different methods of dealing with the surge/stall-problem:

1. Traditionally, surge and rotating stall have been avoided by using control systems that prevent the operating point of the compression system to enter the unstable regime to the left of the *surge line*, that is the stability boundary.
2. A fundamentally different approach, known as active surge/stall control, is to use feedback to stabilize this unstable regime. This approach requires a model of the compression system in order to design a stabilizing controller. Active control can allow for both operation in the peak efficiency and pressure rise regions located in the neighborhood of the surge line, as well as an extension of the operating range of the compressor.

## 13.2 Centrifugal Compressors

### 13.2.1 Introduction

The centrifugal compressor consists essentially of a stationary casing containing a rotating impeller which imparts a high velocity to the fluid, and a number of fixed diverging passages in which the air is decelerated with a consequent rise in static pressure. The latter process is one of diffusion, and consequently, the part of the compressor containing the diverging passages is known as the diffuser. The impeller is mounted on a shaft which is either a direct extension of the drive shaft or a separate shaft supported by bearings and driven through a coupling. The shaft and impeller assembly, called the rotor, are seated in the casing.

The fluid flows into the inducer (also referred to as the impeller eye) and flow through the blade passage of the rotating impeller. Because of the rotation the tangential velocity of the fluid increases when the fluid flows outwards in the impeller, and the associated increase in the centrifugal force makes the static pressure increase through the impeller. A further pressure increase is obtained in the diffuser, where the pressure increase is due to the reduction of the velocity. It is common practice to design a compressor so that

about half the pressure rise occurs in the impeller and half in the diffuser. As no work is done on the fluid in the diffuser, all the energy is supplied to the fluid in the impeller, thus the energy transfer from the shaft power to the fluid energy will be determined by the conditions at the inlet and outlet of the impeller.

It is more difficult to obtain efficient deceleration of flow than it is to obtain efficient acceleration. If divergence in the diffuser is too rapid, the fluid will tend to break away from the walls of the diverging passage, reverse its direction and flow back in the direction of the pressure gradient. This may lead to the formation of eddies with consequent transfer of some kinetic energy into internal energy and a reduction of useful pressure rise. On the other hand, a small angle of divergence will lead to a long diffuser and high losses due to friction. In order to carry out the diffusion in as short a length as possible, the air leaving the impeller may be divided into a number of separate diverging passages separated by fixed diffuser vanes, resulting in a vaneless diffuser. However, in industrial applications where size may be of secondary importance a vaneless diffuser may have the economic advantage as it is much cheaper to manufacture than the vaneless diffuser. A vaneless diffuser is a simple annular channel, and is therefore also known as an annular diffuser, in which the radial velocity component is reduced by area increase and the tangential velocity component by the requirement of constant fluid angular momentum. If the disadvantage of the annular diffuser is its bulk, the advantage is its wide range of operation. A vaneless diffuser may have a higher peak efficiency than an annular diffuser, but its mass flow range is considerably less because of early stall of the diffuser vanes.

We consider a compressor driven by a drive torque  $\tau_d$ . The compressor is modeled as a momentum source in a duct of length  $L$  and cross section  $A$ . The duct is connected to a plenum, which is a volume  $V$ . The gas flows from the duct into the plenum, and then out of the plenum through a throttle. This is illustrated in Figure 13.4. The momentum delivered to the gas is the ideal momentum from the rotor minus momentum loss due to incidence loss and fluid friction loss. The usual *Greitzer surge model* (Greitzer 1976) for this system is derived from the mass balance of the plenum and the momentum equation of the duct.

### 13.2.2 Shaft dynamics

We assume that the compressor is driven by an electrical motor as illustrated in Figure 13.1. The shaft dynamics are given by

$$J\dot{\omega} = \tau_d - \tau_c \quad (13.2)$$

where  $J$  is the inertia of the compressor shaft and compressor wheel,  $\omega$  is the angular velocity of the shaft,  $\tau_d$  is the drive torque of the motor, and  $\tau_c$  is the compressor torque, which is the torque acting on the compressor shaft from the rotor blades. The compressor torque is equal to the rate of change of angular momentum

$$\tau_c = w(r_2 C_{\theta 2} - r_1 C_{\theta 1}) \quad (13.3)$$

where  $w$  is the mass flow,  $C_{\theta 1}$  is the tangential velocity of the fluid at the rotor inlet and  $C_{\theta 2}$  is the tangential fluid velocity at the rotor outlet. The radius at the inlet is  $r_1$  and the rotor radius at the outlet is denoted  $r_2$ , as shown in Figure 13.2. Assuming no pre-whirl, an assumption that is met in e.g. most turbochargers, we have that  $C_{\theta 1} = 0$ , and the torque becomes

$$\tau_c = wr_2 C_{\theta 2} \quad (13.4)$$

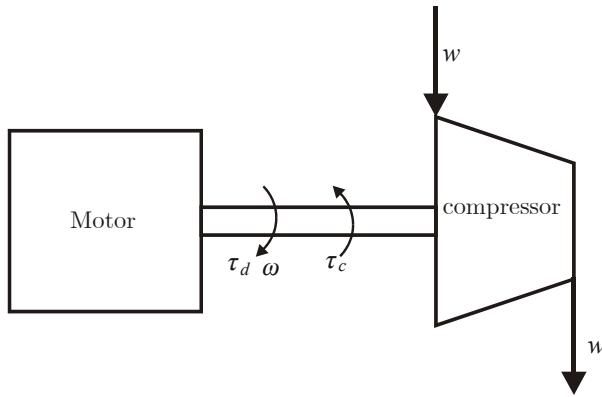


Figure 13.1: Centrifugal compressor with motor.

For backswept blades, we see from the velocity triangle in Figure 13.3 that

$$\begin{aligned} C_{\theta 2} &= U_2 - C_{r2} \cot \beta_{2b} \\ &= (1 - \phi \cot \beta_{2b}) U_2 \\ &= \mu(\phi) U_2 \end{aligned} \quad (13.5)$$

where  $C_{r2}$  is the radial flow velocity,

$$U_2 = r_2 \omega \quad (13.6)$$

is the tangential speed of the impeller tip,

$$\phi = \frac{C_{r2}}{U_2} = \frac{w}{\rho_1 A_1} \frac{r_2}{\omega} \quad (13.7)$$

is the flow coefficient, and  $\beta_{2b}$  is the blade angle at the impeller tip. The function

$$\mu(\phi) = 1 - \phi \cot \beta_{2b} \quad (13.8)$$

is the energy transfer coefficient. The torque is seen to be given by

$$\tau_c = w \mu(\phi) r_2^2 \omega \quad (13.9)$$

In the case of radial vanes  $\beta_{2b} = 90^\circ$  we have that  $\cot \beta_{2b} = 0$  and  $\mu = 1$ . Backwards swept blades have  $\beta_{2b} < 90^\circ$  in which case  $\mu$  decreases with increasing flow coefficient  $\phi$ . In practice the energy transfer coefficient is slightly less than the ideal value, and we may write

$$\mu(\phi) = \sigma (1 - \phi \cot \beta_{2b}) \quad (13.10)$$

where  $\sigma$  is the slip factor, which is slightly less than unity. A usual approximation is the Stanitz slip factor:

$$\sigma \approx 1 - \frac{2}{N} \quad (13.11)$$

where  $N$  is the number of blades.

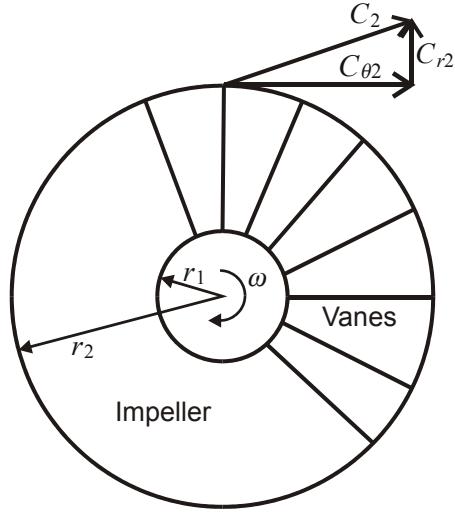


Figure 13.2: The rotating impeller viewed face-on. Note the radial vanes.

### 13.2.3 Compressor system

We will derive the dynamic equations for a compressor system that includes a compressor, a volume  $V_p$  called the plenum, and a duct of length  $L$  running from the compressor to the plenum. The model is derived from the mass balance of the plenum, the momentum equation for the duct and the pressure rise for the compressor. In the derivation the duct and the plenum is treated as in the derivation of the Helmholtz resonator. This means that the fluid in the duct is assumed to be incompressible and with mass flow

$$w = \rho A C \quad (13.12)$$

where  $\rho$  is the density in the duct,  $A$  is the constant cross section of the duct, and  $C$  is the velocity of the fluid in the duct, which is assumed to be constant along the duct.

### 13.2.4 Mass balance

The mass balance for the plenum is

$$V_p \dot{\rho}_p = w - w_t(p_p) \quad (13.13)$$

where  $\rho_p$  is the density in the plenum,  $w$  is the mass flow from the duct into the plenum, and  $w_t(p_p)$  throttle mass flow going out of the plenum. Suppose that the gas in the plenum is ideal and isentropic. Following the derivation leading to equation (12.67), we then have

$$dp_p = c_p^2 d\rho_p$$

where the sonic velocity in the plenum is given by

$$c_p = \sqrt{\kappa R T_p}$$

and the mass balance becomes

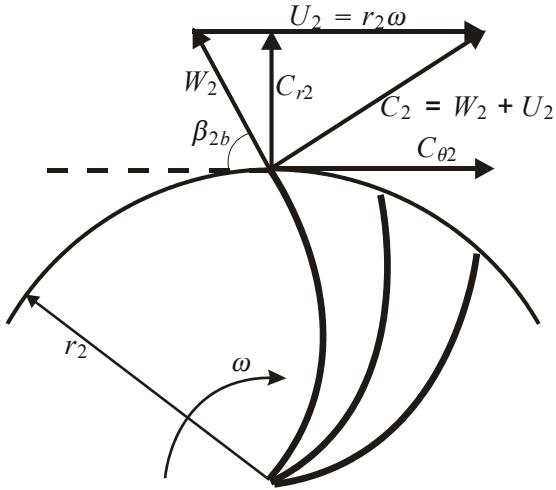


Figure 13.3: Velocity triangle at the impeller exit. Note the backswept vanes.

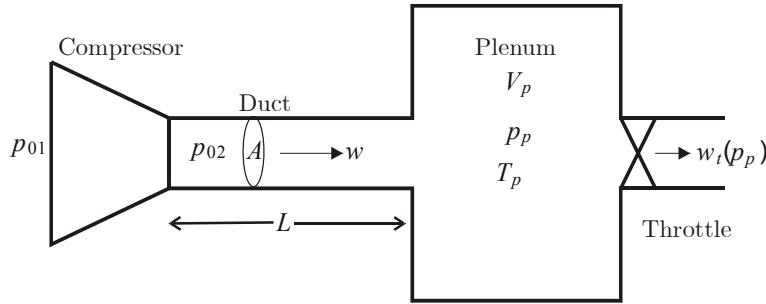


Figure 13.4: Compression system consisting of compressor, duct, plenum volume and throttle.

$$\dot{p}_p = \frac{c_p^2}{V_p} [w - w_t(p_p)] \quad (13.14)$$

In Greitzer's original derivation of this model, it is assumed that  $c_p \approx c_{01}$ , where  $c_{01}$  is the sonic velocity at ambient conditions. This is equivalent to assume  $T_p \approx T_{01}$ .

### 13.2.5 Momentum equation

We will now employ the momentum balance in order to find a differential equation for the duct mass flow. The mass flow in the duct is given by (13.12). The pressure at the inlet of the duct is the pressure at the outlet of the diffusor. The fluid enters the diffusor from the impeller with the stagnation pressure  $p_{02}(w, \omega)$ . In the diffuser the fluid is slowed down through an isentropic process where the stagnation enthalpy and hence the stagnation pressure is kept constant. Due to the reduction in the fluid velocity  $C$  the static pressure  $p = p_0 - \frac{1}{2}C^2$  is increased, and if the velocity is reduced to a value close

to zero, the static pressure at the outlet of the diffuser will be approximately equal to the stagnation pressure at the outlet of the impeller. Therefore, the pressure at the inlet of the duct is set to the stagnation pressure at the outlet of the impeller.

The momentum equation of the duct is then

$$\frac{d}{dt} (m_d C) = A p_{02}(w, \omega) - A p_p \quad (13.15)$$

where

$$m_d = L A \rho \quad (13.16)$$

is the mass in the duct,  $L$  is the duct length,  $p_{02}(w, \omega)$  is the stagnation pressure at the outlet of the compressor rotor which is assumed to be equal to the static pressure at the inlet of the duct, and  $p_p$  is the plenum pressure. The velocity  $C$  of the fluid can be written

$$C = \frac{q}{A} = \frac{w}{\rho A} \quad (13.17)$$

where  $q$  is the duct volume flow. By combining (13.15), (13.16) and (13.17), we get the momentum equation for the duct written as a differential equation in mass flow.

$$L \dot{w} = A [p_{02}(w, \omega) - p_p] \quad (13.18)$$

## 13.3 Compressor characteristic

### 13.3.1 Derivation

In (13.18) we need an expression for the pressure  $p_{02}(w, \omega)$  at the outlet of the compressor. The pressure rise in the compressor is from the stagnation pressure  $p_{01}$  at the inlet, to the stagnation pressure  $p_{02}$  at the rotor outlet. Generally, this rise in stagnation pressure can be obtained in an isentropic process  $1 \rightarrow 2s$  where  $p_{02s} = p_{02}$ . This isentropic process involves an increase in the stagnation enthalpy  $\Delta h_{02s} = h_{02s} - h_{01}$ . In practice, the pressure increase in the compressor is not isentropic, and there is an increase in entropy due to friction  $\Delta h_f$  and incidence losses  $\Delta h_i$ . The incidence losses arise from misalignment of the flow with respect to vane angles at off-design conditions, and the friction losses are due to friction between the fluid and the various solid surfaces in the compressor passages. To account for these losses, the compression from  $p_{01}$  to  $p_{02}$  is modelled as an isentropic process in series with an isobaric process, where the isobaric process involves an increase in entropy. We recall that for the isentropic process  $ds = 0$  and the enthalpy differential is  $dh = vdp$ , while for the isobaric process  $dp = 0$  and  $dh = Tds$ . This results in the following description: First, there is an isentropic process  $1 \rightarrow 2s$  which ends in a state with pressure  $p_{02s} = p_{02}$  and stagnation enthalpy  $h_{02s}$ . Then, there is an isobaric process  $2s \rightarrow 2$ , which ends in a state with pressure  $p_{02}$  and stagnation enthalpy  $h_{02}$ . The total process  $1 \rightarrow 2$  is illustrated in Figure 13.5. This gives

$$\begin{aligned} \Delta h_{02} &= \Delta h_{02s} + \int_{2s}^2 T ds \\ &= \Delta h_{02s} + \Delta h_i + \Delta h_f \end{aligned} \quad (13.19)$$

Here  $\Delta h_i$  is the incidence loss,  $\Delta h_f$  is the friction work, which both will be defined below, and

$$\Delta h_{02s} := \Delta h_{02} - \Delta h_i - \Delta h_f \quad (13.20)$$

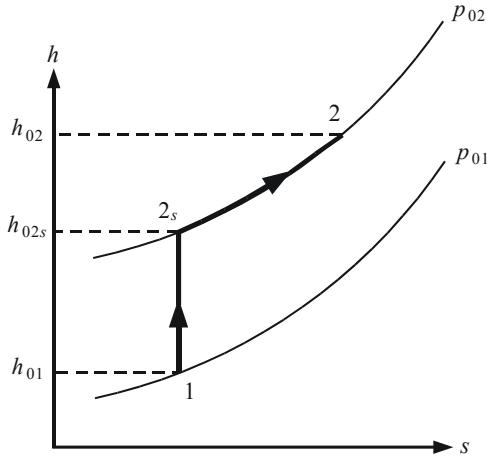


Figure 13.5: The isentropic and isobaric process in the compressor.

so that  $\Delta h_{02s}$  is the change in specific stagnation enthalpy that contributes to the acceleration of the gas in the duct.

The mechanical power transferred from the shaft to the gas is

$$P = \tau_c \omega$$

As there is no energy storage in the gas in the impeller, the power transferred from the rotor to the gas is equal to the rate of stagnation enthalpy increase

$$P = w \Delta h_{02}$$

of the fluid, where  $\Delta h_{02}$  is the increase in specific stagnation enthalpy. By using (13.3), this gives

$$w \Delta h_{02} = P = \tau_c \omega = w(U_2 C_{\theta 2} - U_1 C_{\theta 1}) \quad (13.21)$$

which is known as Euler's pump equation. Alternatively, we can use (13.9) and it follows that

$$w \Delta h_{02} = P = \tau_c \omega = w \mu(\phi) r_2^2 \omega^2 \quad (13.22)$$

and

$$\Delta h_{02}(\omega, C) = \frac{\tau_c \omega}{w} = \mu(\phi) r_2^2 \omega^2 \quad (13.23)$$

In off-design operation of the compressor there will be a mismatch between the fixed blade angle  $\beta_{1b}$  and the direction of the gas stream  $\beta_1$ . The loss associated with this is termed the incidence loss. The incidence loss  $\Delta h_i$ , expressed as a change of enthalpy, is determined according to the so-called NASA model of (Futral and Wasserbauer 1965). It is assumed that  $\Delta h_i$  is equal to a reduction in kinetic energy corresponding to an instantaneous change in fluid velocity at the blade inlet such that the axial velocity is unchanged, while the tangential velocity is changed to accommodate the fixed blade angle  $\beta_{1b}$ . The difference between the flow angle and the blade angle,

$$\beta_i = \beta_{1b} - \beta_1$$

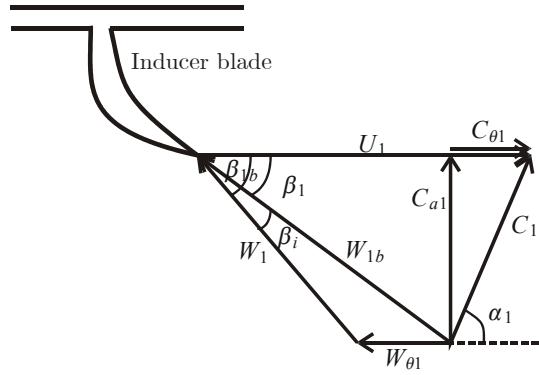


Figure 13.6: Velocity triangles at the inducer. Section through inducer at radius  $r_1$ .

is called the incidence angle. The velocity of the incoming gas relative to the inducer is termed  $W_1$ . Thus, with  $C_{\theta 1} = 0$ , we get from the triangles of Figure 13.6

$$\begin{aligned}\Delta h_i &= \frac{1}{2} W_{\theta 1}^2 \\ &= \frac{1}{2} (U_1 - \cot \beta_{1b} C_{a1})^2 \\ &= \frac{r_1^2}{2} \left( \omega - \cot \beta_{1b} \frac{A}{A_1 r_1} v \right)^2 \\ &= \frac{r_1^2}{2} (\omega - \alpha w)^2\end{aligned}\quad (13.24)$$

where  $U_1 = r_1 \omega$  is the rotor blade speed,  $C_{a1} = C(A/A_1)$  is the axial velocity at the blade inlet,  $\rho_1$  is the density at the blade inlet which is considered to be a constant, and

$$\alpha = \cot \beta_{1b} \frac{A}{A_1 r_1} \quad (13.25)$$

In order to calculate the friction loss, we treat the flow passages in the impeller as pipes with circular cross section. The friction loss is then modelled as

$$\Delta h_f = C_h \frac{l}{D} \frac{W_{1b}^2}{2}$$

where  $C_h$  is the surface friction loss coefficient,  $l$  is the mean channel length and  $D$  is the mean hydraulic diameter. The hydraulic diameter for a pipe of non-circular cross section can be calculated as

$$D = \frac{4A}{a}$$

where  $A$  is the cross sectional area and  $a$  is the length of the wetted perimeter. Using Figure 13.6, the friction loss can now be shown to be given by

$$\Delta h_f = k_f w^2 \quad (13.26)$$

where

$$k_f = \frac{C_h l}{2 D \rho_1^2 A_1^2 \sin^2 \beta_{1b}}$$

is a constant. Combining (13.23), (13.20), (13.24) and (13.26) results in the enthalpy transfer

$$\Delta h_{02s}(\omega, w) = \sigma r_2^2 \omega^2 - \frac{r_1^2}{2} (\omega - \alpha w)^2 - k_f w^2 \quad (13.27)$$

The stagnation pressure ratio is given by the stagnation temperature ratio of the isentropic process  $1 \rightarrow 2s$  according to

$$\frac{p_{02}}{p_{01}} = \frac{p_{02s}}{p_{01}} = \left( \frac{T_{02s}}{T_{01}} \right)^{\frac{\kappa}{\kappa-1}} \quad (13.28)$$

An expression for  $T_{02s}/T_{01}$  is found as

$$\frac{T_{02s}}{T_{01}} = \frac{h_{02s}}{h_{01}} = \frac{h_{01} + \Delta h_{02s}}{h_{01}} = 1 + \frac{\Delta h_{02s}}{h_{01}} = 1 + \frac{\Delta h_{02s}}{c_p T_{01}} \quad (13.29)$$

From (12.63) and (13.27) we find the expression for the pressure ratio

$$\Psi_c(w, \omega) = \frac{p_{02}}{p_{01}} = \left( 1 + \frac{\Delta h_{02s}}{c_p T_{01}} \right)^{\frac{\kappa}{\kappa-1}} \quad (13.30)$$

$$\Psi_c(w, \omega) = \left( 1 + \frac{\sigma r_2^2 \omega^2 - \frac{r_1^2}{2} (\omega - \alpha w)^2 - k_f w^2}{c_p T_{01}} \right)^{\frac{\kappa}{\kappa-1}} \quad (13.31)$$

such that

$$p_{02}(w, \omega) = \Psi_c(w, \omega) p_{01} \quad (13.32)$$

Inserting (13.32) into (13.18) gives the momentum balance

$$\dot{w} = \frac{A}{L} \left( \left( 1 + \frac{\Delta h_{02s}}{c_p T_{01}} \right)^{\frac{\kappa}{\kappa-1}} p_{01} - p_p \right) \quad (13.33)$$

or

$$\dot{w} = \frac{A}{L} (\Psi_c(w, \omega) p_{01} - p_p)$$

From (13.2), (13.14) and (13.33), the resulting compressor model is found to be

$$\dot{p}_p = \frac{c_p^2}{V_p} (w - w_t) \quad (13.34)$$

$$\frac{L}{A} \dot{w} = \left( 1 + \frac{\mu r_2^2 \omega^2 - \frac{r_1^2}{2} (\omega - \alpha w)^2 - k_f w^2}{c_p T_{01}} \right)^{\frac{\kappa}{\kappa-1}} p_{01} - p_p \quad (13.35)$$

$$J \dot{\omega} = \tau_d - w r_2^2 \mu \omega \quad (13.36)$$

### 13.3.2 The compressor characteristic at zero mass flow

Let  $C$  denote the fluid velocity and  $h$  the specific enthalpy. The stagnation enthalpy

$$h_0 = h + \frac{1}{2} C^2 \quad (13.37)$$

is increased over the impeller by the the energy

$$\tau_c \omega = (r_2 C_{\theta 2} - r_1 C_{\theta 1}) \omega w = (U_2 C_{\theta 2} - U_1 C_{\theta 1}) w \quad (13.38)$$

that is transferred from the compressor blades to the fluid. Thus, if we let 1 denote the impeller inlet and 2 denote the impeller outlet we have

$$(h_{20} - h_{10}) w = (U_2 C_{\theta 2} - U_1 C_{\theta 1}) w \quad (13.39)$$

When the mass flow tends to zero it seems reasonable to assume that the increase in static enthalpy is continuous at  $w = 0$ . Moreover, at zero mass flow through the compressor we may assume that  $C_{\theta 1} = U_1$  and  $C_{\theta 2} = U_2$ . This gives the following increase in static enthalpy for zero mass flow:

$$\Delta h_{0s} = h_{20} - h_{10} = U_2^2 - U_1^2 = \omega^2 (r_2^2 - r_1^2) \quad (13.40)$$

where  $r_2$  is the radius of the impeller outlet and  $r_1$  is the radius of the impeller inlet. The associated pressure rise for the compressor at zero mass flow is

$$\begin{aligned} \frac{p_{02}}{p_{01}} &= \left( 1 + \frac{h_{02} - h_{01}}{c_p T_{01}} \right)^{\frac{\kappa}{\kappa-1}} = \left( 1 + \frac{\omega^2 (r_2^2 - r_1^2)}{c_p T_{01}} \right)^{\frac{\kappa}{\kappa-1}} \\ &= \left( 1 + \frac{\rho_{01} \omega^2 (r_2^2 - r_1^2)}{p_{01} \frac{\kappa}{\kappa-1}} \right)^{\frac{\kappa}{\kappa-1}} \end{aligned} \quad (13.41)$$

where the ideal gas law has been used together with  $c_p = R\kappa / (\kappa - 1)$ .

**Example 203** By combining (13.37) and (13.39), the increase in enthalpy as opposed to stagnation enthalpy is found from

$$\left( h_1 + \frac{1}{2} C_1^2 - U_1 C_{\theta 1} \right) w = \left( h_2 + \frac{1}{2} C_2^2 - U_2 C_{\theta 2} \right) w. \quad (13.42)$$

We introduce the relative velocity  $W_i = C_i - U_i$ . In terms of the tangential components  $(\cdot)_{\theta i}$  and the radial components  $(\cdot)_{ri}$  we have

$$C_{\theta i} = U_i + W_{\theta i}, \quad C_{ri} = W_{ri} \quad (13.43)$$

and it follows that

$$C_i^2 = C_{\theta i}^2 + C_{ri}^2 = U_i^2 + 2U_i W_{\theta i} + W_{\theta i}^2 + W_{ri}^2 = U_i^2 + 2U_i W_{\theta i} + W_i^2 \quad (13.44)$$

We insert this expression in the enthalpy equation, and get

$$\left( h_1 + \frac{1}{2} W_1^2 - \frac{1}{2} U_1^2 \right) w = \left( h_2 + \frac{1}{2} W_2^2 - \frac{1}{2} U_2^2 \right) w \quad (13.45)$$

which in the terminology of (Cumpsty 1989) says that the rothalpy

$$I_i = h_i + \frac{1}{2} W_i^2 - \frac{1}{2} U_i^2 \quad (13.46)$$

is unchanged over the impeller. We may then compute the change in enthalpy at zero mass flow by assuming continuity when  $w \rightarrow 0$ . At zero mass flow we have  $C_i = U_i$  at the inlet and at the outlet, and therefore  $W_i = 0$ . It follows that the increase in enthalpy over the impeller at zero mass flow is

$$h_2 - h_1 = \frac{1}{2} (U_2^2 - U_1^2) = \frac{\omega^2}{2} (r_2^2 - r_1^2) \quad (13.47)$$

where  $r_2$  is the radius of the impeller outlet and  $r_1$  is the radius of the impeller inlet. As

$$h_{20} - h_{10} = h_2 - h_1 + \frac{1}{2} (C_2^2 - C_1^2) \quad (13.48)$$

this is consistent with the result for the stagnation enthalpy with  $C_1 = U_1$  and  $C_2 = U_2$ . Assuming isentropic pressure rise,

$$\Psi_c(w, \omega) = \frac{p_{02}}{p_{01}} = \left(1 + \frac{\Delta h}{c_p T_{01}}\right)^{\frac{\kappa}{\kappa-1}}, w > 0. \quad (13.49)$$

holds, for details see (Gravdahl, Egeland and Vatland 2001), and by combining (13.46) and (13.49), we get at zero mass flow

$$\Psi_c(0, N) = \Psi_o = \left(1 + \frac{\pi^2 N^2 (D_2^2 - D_1^2)}{2c_p T_{01}}\right)^{\frac{\kappa}{\kappa-1}}, \quad (13.50)$$

where  $N = 2\pi\omega$  is the rotational speed in revolutions per second.

**Remark 5** Close to zero mass flow there may be high incidence losses in the diffuser so that the fluid is not slowed down isentropically, and the kinetic energy is not recovered as a pressure rise. Therefore, the pressure at the inlet of the duct may be set to the static pressure at the outlet of the impeller, in which case the pressure rise is

$$\frac{p_2}{p_{01}} = \left(1 + \frac{h_{02} - h_{01}}{c_p T_{01}}\right)^{\frac{\kappa}{\kappa-1}} = \left(1 + \frac{\rho_{01}\omega^2 (r_2^2 - r_1^2)}{2p_{01}\frac{\kappa}{\kappa-1}}\right)^{\frac{\kappa}{\kappa-1}} \quad (13.51)$$

**Example 204** The centrifugal force on a material volume element is

$$\rho\omega^2 r dV \quad (13.52)$$

Integrating this gives the centrifugal loading at  $r = r_2$  for zero mass flow:

$$\int_{r_1}^{r_2} \rho\omega^2 r dr = \frac{1}{2}\rho_1\omega^2 (r_2^2 - r_1^2) \quad (13.53)$$

where the density has been approximated by  $\rho = \rho_1$ . Then the increase in static pressure at zero mass flow is

$$p_2 \approx p_1 + \frac{1}{2}\rho_1\omega^2 (r_2^2 - r_1^2) \quad (13.54)$$

and the pressure rise in terms of static pressure is

$$\frac{p_2}{p_1} = 1 + \frac{\rho_1\omega^2 (r_2^2 - r_1^2)}{2p_1} \quad (13.55)$$

This expression for the pressure rise is approximately equal to the expression (13.51) for a small pressure rise.

## 13.4 Compressor surge

### 13.4.1 The Greitzer surge model

Compressor surge is a serious problem in the control of compressors, and the main objective of compressor control systems is to avoid surge, as it reduces performance, and may lead to the destruction of the compressor. The compressor may surge if the pressure rise increases with increasing mass flow, which is the case in the unstable region of the compressor map. The mass flow and the plenum pressure will form an oscillating system, and even flow reversal may occur.

In the Greitzer model for the description of compressor surge dynamics the shaft dynamics are left out, and the compressor model (13.34–13.36) is written

$$\dot{p}_p = \frac{c_p^2}{V_p} [w - w_t(p_p)] \quad (13.56)$$

$$L\rho\dot{C} = (\Psi_c(w, \omega) - 1)p_{01} + p_{01} - p_p \quad (13.57)$$

The addition of  $\pm p_{01}$  in (13.57) has been done to facilitate the use of pressure differences below. This model is known as Greitzer's compressor model in dimensional form, and was first introduced in (Greitzer 1976). We introduce the nondimensional variables

$$\xi = \frac{t}{\omega_H} \quad (13.58)$$

$$\psi = \frac{p_p - p_{01}}{\frac{1}{2}\rho U^2}, \quad \psi_c(\phi) = \frac{(\Psi_c(w, \omega) - 1)p_{01}}{\frac{1}{2}\rho U^2} \quad (13.59)$$

$$\phi = \frac{C}{U} = \frac{w}{\rho AU}, \quad \phi_t(\psi) = \frac{w_t(p_p)}{\rho AU} \quad (13.60)$$

where  $\xi$  is the dimensionless time variable,  $\psi$  is the pressure coefficient,  $\psi_c(\phi)$  is the compressor characteristic,  $\phi$  is the flow coefficient,  $\phi_t(\psi)$  is the throttle flow coefficient, and

$$\omega_H = c_p \sqrt{\frac{A}{V_p L}} \quad (13.61)$$

is the Helmholtz frequency of the duct-plenum system.

**Example 205** If the temperature increase over the compressor is small, then the compressor characteristic can be approximated by

$$\begin{aligned} \psi_c(\phi) &= \frac{(\Psi_c(w, \omega) - 1)p_{01}}{\frac{1}{2}\rho_{01}U^2} = \left[ \left( 1 + \frac{\Delta h_{02s}(\omega, C)}{c_p T_{01}} \right)^{\frac{\kappa}{\kappa-1}} - 1 \right] \frac{p_{01}}{\frac{1}{2}\rho_{01}U^2} \\ &\approx \frac{\kappa}{\kappa-1} \frac{\Delta h_{02s}(\omega, C)}{c_p T_{01}} \frac{p_{01}}{\frac{1}{2}\rho_{01}U^2} \\ &= \frac{\Delta h_{02s}(\omega, C)}{\frac{1}{2}U^2} \frac{p_{01}}{RT_{01}\rho_{01}} \\ &= \frac{\Delta h_{02s}(\omega, C)}{\frac{1}{2}U^2} \\ &= 2\mu - \left( \frac{r_1}{r_2} \right)^2 (1 - r_2\alpha\phi) - 2k_f (\rho_{01}A)^2 \phi^2 \end{aligned} \quad (13.62)$$

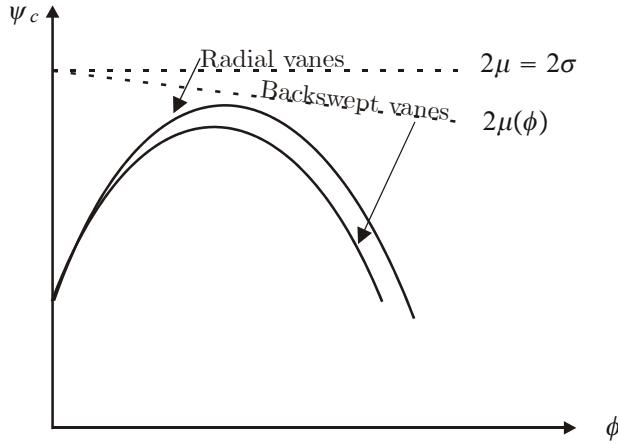


Figure 13.7: The compressor characteristic from (13.62).

where the expression for  $\Delta h_{02s}(\omega, w)$  from (13.27) has been used. Equation (13.62) clearly shows that the compressor characteristic is a function of the flow coefficient  $\phi$  when  $\Delta T_{02s}$  is small. The characteristic is plotted in Figure 13.7.

Using the normalized variables the model becomes

$$\frac{1}{2}\rho U^2 \omega_H \frac{d\psi}{d\xi} = \rho AU \frac{c_p^2}{V_p} [\phi - \phi_t(\psi)] \quad (13.63)$$

$$U \omega_H L \frac{d\phi}{d\xi} = \frac{1}{2} U^2 \psi_c(\phi) - \frac{1}{2} \rho U^2 \frac{1}{\rho} \psi \quad (13.64)$$

This gives the following result:

The Greitzer surge model is given by

$$\frac{d\psi}{d\xi} = \frac{1}{B} [\phi - \phi_t(\psi)] \quad (13.65)$$

$$\frac{d\phi}{d\xi} = B [\psi_c(\phi) - \psi] \quad (13.66)$$

where the nondimensional parameter

$$B = \frac{U}{2c_p} \sqrt{\frac{V_p}{AL}} = \frac{U}{2L\omega_H} \quad (13.67)$$

is Greitzer's B-parameter.

A large  $B$  corresponds to a small  $\omega_H$ . In axial compressors, the numerical value of  $B$  gives information about the type of instability the compressor will enter if operating beyond the surge line. A large  $B$  indicates that the compressor will enter surge, and a small  $B$  indicates that rotating stall will occur.

**Example 206 Remark 6** The pressure coefficient

$$\psi = \frac{\Delta p}{\frac{1}{2}\rho U^2} \quad (13.68)$$

has also been characterized as a temperature coefficient

$$\psi_{Cohen} = \frac{\Delta h_{0s}}{\frac{1}{2}U^2} \quad (13.69)$$

in (Cohen, Rogers and Saravanamuttoo 1996).

### 13.4.2 Linearization

Linearization of the Greitzer model (13.65)-(13.66) gives

$$\frac{d}{d\xi} \begin{pmatrix} \psi \\ \phi \end{pmatrix} = \begin{pmatrix} -\frac{1}{Bg_t} & \frac{1}{B} \\ -B & Bg_c \end{pmatrix} \begin{pmatrix} \psi \\ \phi \end{pmatrix} \quad (13.70)$$

where

$$g_c = \frac{\partial \psi_c}{\partial \phi}, \quad g_t = \left( \frac{\partial \phi_t}{\partial \psi} \right)^{-1} \quad (13.71)$$

are the slopes of the compressor characteristic and the throttle characteristic in a  $\phi, \psi$  diagram. The characteristic equation is

$$\lambda^2 + \left( \frac{1}{Bg_t} - Bg_c \right) \lambda + \left( 1 - \frac{g_c}{g_t} \right) = 0 \quad (13.72)$$

and the system is seen to be stable if

$$g_c < \frac{1}{B^2 g_t} \quad (13.73)$$

and

$$g_c < g_t \quad (13.74)$$

The case  $g_c > (B^2 g_t)^{-1}$  is usually referred to as a dynamic instability, while the case  $g_c < g_t$  is called a static instability. We consider a situation where the slope  $g_c$  of the compressor characteristic increases from some negative value. We see that if  $B$  becomes small, then a larger  $g_c$  is tolerated before the dynamic instability is encountered. If  $B$  becomes sufficiently small, that is if  $B < g_t^{-1}$ , then the static instability will occur before the dynamic instability.

**Example 207** At the top of the compressor characteristic we have  $g_c = 0$ , and the characteristic equation becomes

$$\lambda^2 + \frac{1}{Bg_t} \lambda + 1 = 0 \quad (13.75)$$

This corresponds to a normalized undamped natural frequency  $\omega_0^* = 1$ . We recall that the normalized time variable  $\xi = t\omega_H$  is scaled with the Helmholtz frequency, and the corresponding undamped natural frequency is found to be

$$\omega_0 = \omega_0^* \frac{\xi}{t} = \omega_H \quad (13.76)$$

We see that the undamped natural frequency for an equilibrium point at the top of the compressor characteristic is the Helmholtz frequency  $\omega_H$ .

**Example 208** We now assume that there is a valve at the compressor outlet. The valve characteristic is

$$\dot{w} = C_D A_2 \sqrt{\frac{2\rho(p_1 - p_2)}{1 - \left(\frac{A_2}{A_1}\right)^2}} \quad (13.77)$$

The plenum volume in this case is small, and we may assume that  $B < g_t^{-1}$ . Then the stability requirement is

$$g_c < g_t \quad (13.78)$$

that is, the static instability must be avoided.

**Example 209** If the compressor delivers gas to a pipeline the plenum volume is  $V_p \rightarrow \infty$  and it follows that  $B \rightarrow \infty$ . In this case  $\psi$  will be a constant, and the system is described by the first order model

$$\frac{d\phi}{d\xi} = B [\psi_c(\phi) - \psi] \quad (13.79)$$

This model is stable whenever

$$g_c < 0 \quad (13.80)$$

This means that a compressor connected to a pipeline is stable as long as the compressor characteristic has a negative slope.

### 13.4.3 Passivity of the Greitzer surge model

We will now investigate the passivity properties of the Greitzer model. The model is given as

$$\frac{d\psi}{d\xi} = -\frac{1}{B} [\phi - \phi_t(\psi)] \quad (13.81)$$

$$\frac{d\phi}{d\xi} = B [\psi_c(\phi) - \psi] \quad (13.82)$$

where (13.81) is the pressure dynamics, and (13.82) is the mass flow dynamics. The throttle flow  $\phi_t(\psi)$  is given as

$$\phi_t(\psi) = \gamma_t \sqrt{\psi} \quad (13.83)$$

Consider the positive function

$$V = V_1 + V_2 = \frac{1}{2} \left( B\psi^2 + \frac{1}{B}\phi^2 \right) \quad (13.84)$$

The time derivative of  $V_1$  along the solutions of (13.81) is

$$\frac{dV_1}{d\xi} = \psi\phi - \psi\phi_t(\psi) \quad (13.85)$$

Integrating (13.85) we see that

$$\begin{aligned} \int_0^T \psi\phi d\xi &= V_1(T) - V_1(0) + \int_0^T \psi\phi_t(\psi) d\xi \\ &> -V_1(0) + \int_0^T \psi\phi_t(\psi) d\xi \end{aligned}$$

which implies that the system with  $\phi$  as input and  $\psi$  as output has certain passivity properties depending on the value of the term  $\int_0^T \psi \phi_t(\psi) d\xi$ . The throttle in a compression system is a passive component, and from (13.83) it can be seen that a coefficient  $\kappa_1$  can always be chosen sufficiently small such that the throttle characteristic satisfies the condition

$$\psi \phi_t(\psi) \geq \kappa_1 \psi^2 \quad (13.86)$$

We now find that

$$\int_0^T \psi \phi d\xi > -V_1(0) + \kappa_1 \int_0^T \psi^2 d\xi \quad (13.87)$$

and it follows that the pressure dynamics are (strictly) passive.

The time derivative of  $V_2$  along the solutions of (13.82) is

$$\frac{dV_2}{d\xi} = \phi \psi_c(\phi) - \psi \phi \quad (13.88)$$

Following the same procedure as above, we find that

$$\begin{aligned} \int_0^T -\psi \phi d\xi &= V_2(T) - V_2(0) + \int_0^T \phi \psi_c(\phi) d\xi \\ &> -V_2(0) + \int_0^T \phi \psi_c(\phi) d\xi \end{aligned} \quad (13.89)$$

and the passivity properties of the system with input  $-\psi$  and output  $\phi$  depends on the value of the term  $\int_0^T \phi \psi_c(\phi) dt \xi$ . For operating points where the slope of the compressor characteristic is negative, the sector condition

$$\phi \psi_c(\phi) \geq \kappa_2 \phi^2$$

will hold, and

$$\int_0^T -\psi \phi d\xi > -V_2(0) + \kappa_2 \int_0^T \phi^2 d\xi \quad (13.90)$$

and it follows that the mass flow dynamics are (strictly) passive. This leads us to the known result that the feedback interconnection of the two systems, that is the Greitzer model, is a passive system when the operating point is located on a part of the compressor characteristic with negative slope.

In the case where the operating point is not located in an area of negative slope, that is the system is in an unstable (surge) condition, we still have that the pressure dynamics are passive according to (13.87). The flow dynamics are now not passive according to (13.90), but if it is possible to manipulate, through some external actuator, the compressor characteristic  $\psi_c(\phi)$ , (13.89) can be used for controller design. This approach was taken in (Gravdahl and Egeland 1998) using a close coupled valve for active surge control, and in (Gravdahl and Egeland 2002) using the drive itself.

#### 13.4.4 Curvefitting of compressor characteristic

When modelling a real compression system it might be difficult to obtain models of the compressor characteristic like (13.30) or (13.62). A compressor is usually equipped with a measured compressor map. This might be measured when the compressor is manufactured or when it is installed. It is then an alternative to use an approximation

of this measured characteristic in the model of the system. In this example it is shown how to use a polynomial approximation. The solid circles in Figure 13.8 are points read from an actual compressor map for a centrifugal compressor used in pipeline natural gas transport in the North Sea in Norway. It is to be noted that points near the stonewall area of the measured characteristic were not used. In that area choking is the dominant effect, and this is not taken into account in the modeling. Also, in order to ensure that the approximated constant speed lines do not cross each other in the negative flow area, one point for negative flow was chosen for each speed line. The zero flow pressure rise for each speed as calculated by (13.50), and the dotted lines are the third order polynomial approximations of the speed lines at five different speeds.

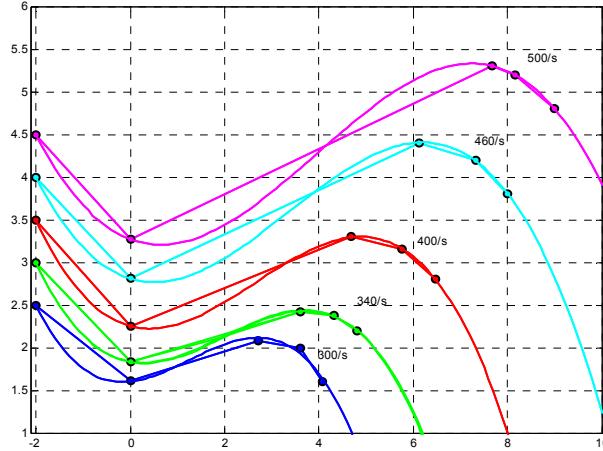


Figure 13.8: The measured speed lines (solid lines) and the polynomial approximations (dashed lines).

The approximations are calculated using the MATLAB function `polyfit`, and for the five chosen speed lines, results are:

$$\begin{aligned}\Psi_c(w, 300) &= 1.6024 - 0.0625w + 0.1668w^2 - 0.0441w^3, \\ \Psi_c(w, 340) &= 1.8291 - 0.0966w + 0.1825w^2 - 0.0304w^3, \\ \Psi_c(w, 400) &= 2.2511 - 0.1443w + 0.1908w^2 - 0.0240w^3, \\ \Psi_c(w, 460) &= 2.8092 - 0.1658w + 0.1783w^2 - 0.0177w^3, \\ \Psi_c(w, 500) &= 3.2699 - 0.2051w + 0.1744w^2 - 0.0147w^3.\end{aligned}$$

A compressor map is also continuous in the rotational speed, as can be seen from (13.30), so in order to simulate the system, there is a need for making the approximated map also continuous in rotational speed. For this reason, the coefficients of the third order polynomials are chosen to be functions of rotational speed. The polynomial approximation for each speed line can be written as

$$\Psi_c(w, N) = c_0(N) + c_1(N)w + c_2(N)w^2 + c_3(N)w^3,$$

where the functions

$$c_i(N) = c_{i0} + c_{i1}N + c_{i2}N^2 + c_{i3}N^3$$

are calculated by using polynomial approximation yet again. In Figure 13.9, the polynomial coefficients of the five polynomials are plotted as a function of rotational speed. As can be seen, a fairly good fit can be made with third order.

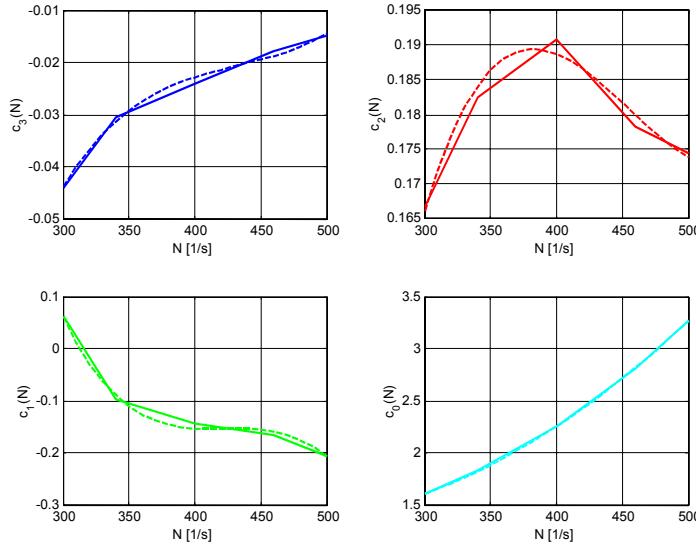


Figure 13.9: The coefficients  $c_i$  as functions of speed  $N$  (solid lines), and their polynomial approximations (dashed lines).

**Remark 7** For the presentation of compressor characteristics which describe the stationary performance of compressors another normalization is often used. This is based on using the Mach number to represent flow and blade velocity. We introduce as nondimensional variables the Mach numbers

$$M_C = \frac{C}{c_0}, \quad M_U = \frac{U}{c_0} \quad (13.91)$$

where

$$c_0 = \sqrt{\kappa R T_0} \quad (13.92)$$

and  $T_0 = T + \frac{c^2}{2}$  is the stagnation temperature. Then, using  $w = \rho A C$ , we find that

$$M_C = \frac{C}{c_0} = \frac{w}{\rho A \sqrt{\kappa R T_0}} = \frac{w R T_0}{A p_0 \sqrt{\kappa R T_0}} \frac{\rho_0}{\rho} = \frac{\rho_0}{\rho \sqrt{\kappa}} \frac{w \sqrt{R T_0}}{A p_0} \quad (13.93)$$

and that

$$M_U = \frac{U}{c_0} = \frac{\omega r}{\sqrt{\kappa R T_0}} = \frac{1}{\sqrt{\kappa}} \frac{\omega r}{\sqrt{R T_0}} \quad (13.94)$$

For plotting of compressor characteristics the following dimensionless variables are used: The corrected mass flow

$$\sqrt{\kappa} M_C = \frac{w \sqrt{R T}}{A p} \quad (13.95)$$

the corrected speed,

$$\sqrt{\kappa} M_U = \frac{\omega r}{\sqrt{RT}} \quad (13.96)$$

and the pressure ratio  $p_{02}/p_{01}$ .

### 13.4.5 Compression systems with recycle

As mentioned in Section 13.1.2, the surge problem in centrifugal compressor can be handled by using a recycle line around the compressor and thereby ensuring that the flow is maintained above a certain minimum value. Such a recycle concept is illustrated in Figure 13.10.

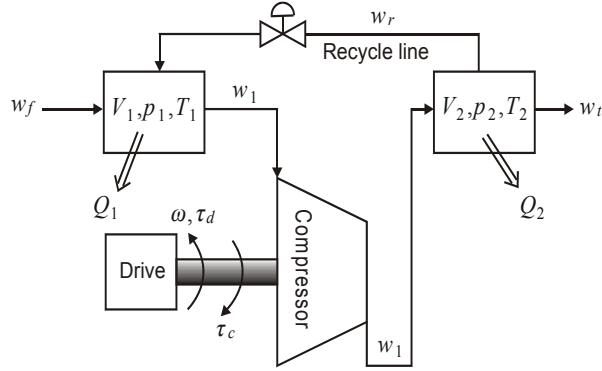


Figure 13.10: A centrifugal compressor with drive unit, upstream volume, downstream volume and recycle line.

A model of this system can be found by calculating the mass balances for the two volumes, calculating the momentum balance of the downstream compressor duct, and calculating the torque balance of the shaft:

$$\dot{p}_1 = \frac{c_{01}^2}{V_1} (w_f - w_1 + w_r) \quad (13.97a)$$

$$\dot{p}_2 = \frac{c_{01}^2}{V_2} (w_1 - w_r - w_t) \quad (13.97b)$$

$$\dot{w}_1 = \frac{A}{L_c} (\Psi_c(w_1, \omega)p_1 - p_2) \quad (13.97c)$$

$$\dot{\omega} = \frac{1}{J} (\tau_d - \tau_c), \quad (13.97d)$$

where,  $p_1$  is the pressure in the upstream volume  $V_1$ ,  $p_2$  is the pressure in the downstream volume  $V_2$ ,  $w_f$  is the feed mass flow,  $w_1$  is the mass flow through the compressor,  $w_r$  is the recycle mass flow,  $w_t$  is the throttle flow,  $\Psi_c(w_1, \omega)$  is the compressor characteristic,  $\omega$  is the rotational speed of the compressor,  $\tau_d$  is the drive torque,  $\tau_c$  is the compressor torque,  $J$  is the inertia of all rotating equipment,  $c_{01}$  is the sonic velocity at ambient conditions.

The mass flow through the recycle valve is given as

$$w_r = k_r \sqrt{\Delta p_r}, \quad (13.98)$$

where  $\Delta p_r$  is the pressure drop across the valve.

In order to ensure high efficiency, compression systems are often equipped with coolers. In order to take this into account, we need to study the energy flow in the system. By combining the mass balance

$$\dot{m} = w_{in} - w_{out}$$

for a volume  $V_j$  with the energy balance

$$\frac{d}{dt} U = \sum_i w_i h_i - Q,$$

and assuming ideal gas such that

$$pV = mRT,$$

one gets

$$\begin{aligned} \frac{d}{dt} (mu) &= \sum_i w_i c_p T_i - Q \\ (w_{in} - w_{out}) u + mc_v \dot{T} &= w_{in} c_p T_{in} - w_{ut} c_p T + Q \\ \dot{T} &= \frac{RT}{pVc_v} (w_{in} c_p T_{in} - (w_{ut} R + w_{in} c_v) T + Q) \end{aligned} \quad (13.99)$$

which is the energy balance for the volume in terms of temperature. Here  $U$  is internal energy,  $u = c_v T$  is specific internal energy,  $h = c_p T$  is specific enthalpy,  $c_p$  and  $c_v$  are the specific heats at constant pressure and volume,  $R = c_p - c_v$  is the specific gas constant and  $Q$  is heat flow. One energy balance (13.99) is used per volume.



# **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

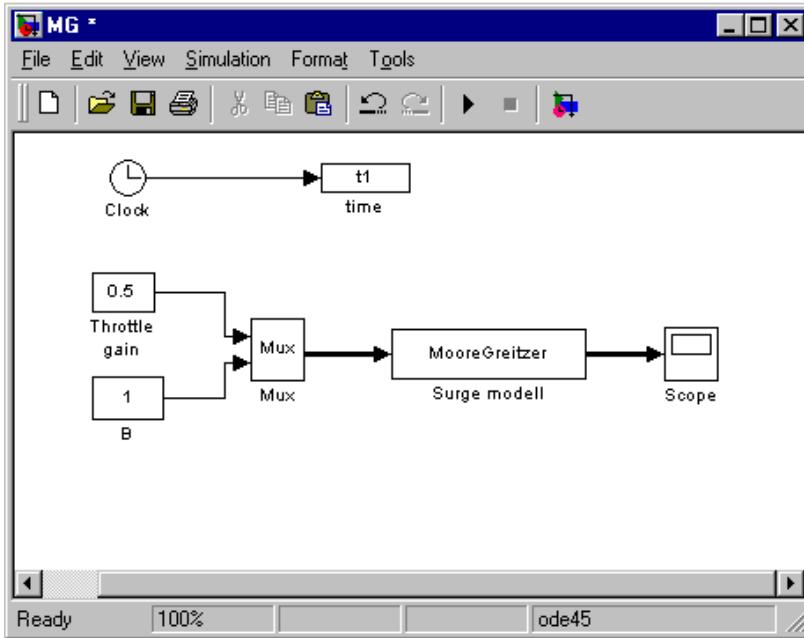


Figure 14.1: The model implemented in Simulink.

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

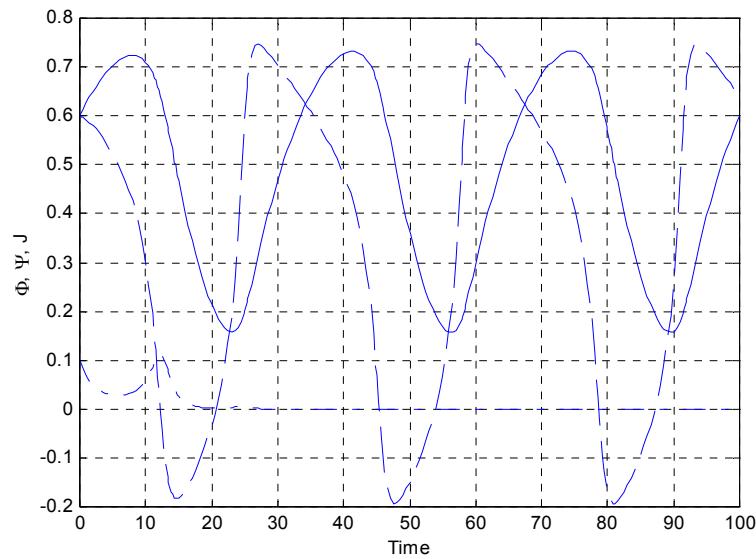


Figure 14.2: Simulation of rotating stall.  $\Phi$ ,  $\Psi$  and  $J$  are plotted with dashed solid and dash-dotted lines respectively.

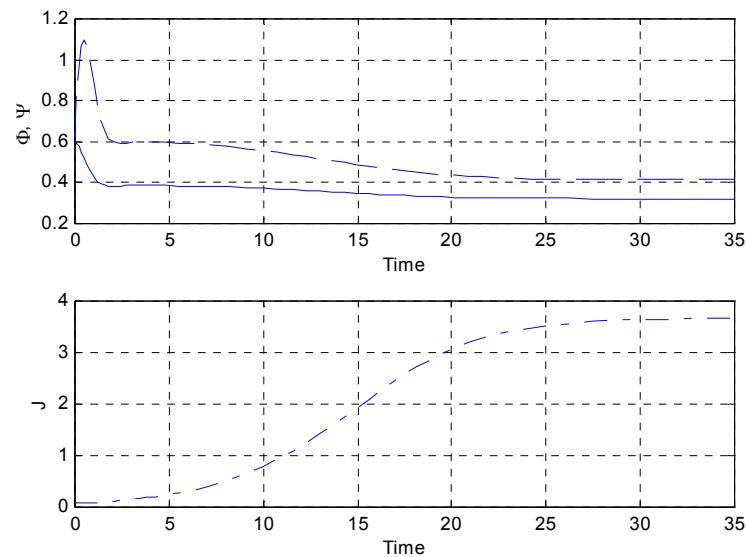


Figure 14.3: Simulation of surge.  $\Phi$ ,  $\Psi$  and  $J$  are plotted with dashed solid and dash-dotted lines respectively.

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

$$\dot{\Psi} = \frac{1}{4l_c B^2} (\Phi - \gamma_T \sqrt{\Psi}) \quad (14.1)$$

$$\dot{\Phi} = \frac{H}{l_c} \left( -\frac{\Psi - \psi_{c0}}{H} + 1 + \frac{3}{2} \left( \frac{\Phi}{W} - 1 \right) \left( 1 - \frac{J}{2} \right) - \frac{1}{2} \left( \frac{\Phi}{W} - 1 \right)^3 \right) \quad (14.2)$$

$$\dot{J} = J \left( 1 - \left( \frac{\Phi}{W} - 1 \right)^2 - \frac{J}{4} \right) \sigma \quad (14.3)$$

where  $\Psi$  is the nondimensional plenum pressure (pressure divided by density and the square of compressor rotational speed),  $\Phi$  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,  $\gamma_T$  is a parameter proportional to the throttle opening, and  $H$ ,  $W$ ,  $\psi_{c0}$  and  $\sigma$  are constants describing the compressor. Finally,

$$B = \frac{U}{2c} \sqrt{\frac{V_p}{A_c l_c}} \quad (14.4)$$

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

$$\dot{x} = v_x \quad (14.5)$$

$$\dot{y} = v_y \quad (14.6)$$

$$\dot{v}_x = 2v_y + x - \frac{m_1(x+m_2)}{r_1^3} - \frac{m_2(x-m_1)}{r_2^3} \quad (14.7)$$

$$\dot{v}_y = -2v_x + y - \frac{m_1 y}{r_1^3} - \frac{m_2 y}{r_2^3} \quad (14.8)$$

where

$$r_1 = \sqrt{(x+m_2)^2 + y^2} \quad (14.9)$$

$$r_2 = \sqrt{(x-m_1)^2 + y^2} \quad (14.10)$$

$$m_1 + m_2 = 1 \quad (14.11)$$

Orbit	Numerical values	
<b>1</b>	$x_0$	0.994
	$y_0$	0
	$v_{x0}$	0
	$v_{y0}$	-2.00158510637908252240537862224
	$T$	17.0652165601579625588917209
	$m_2$	0.012277471
<b>2</b>	$x_0$	0.994
	$y_0$	0
	$v_{x0}$	0
	$v_{y0}$	-2.0317326295573368357302057924
	$T$	11.124340337266085134999734047
	$m_2$	0.012277471
<b>3</b>	$x_0$	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

$$h = \frac{1}{2} (v_x^2 + v_y^2 - x^2 - y^2) - \frac{m_1}{r_1} - \frac{m_2}{r_2} \quad (14.12)$$

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=17.0652165601579625588917206249; %Periode
```

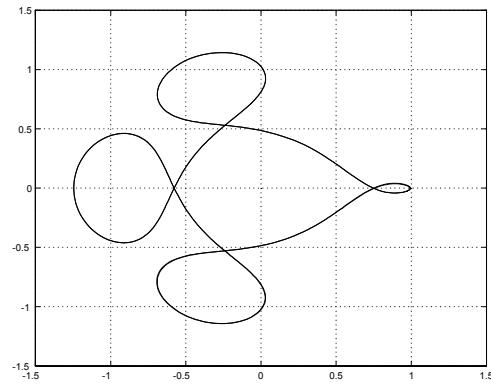


Figure 14.4: Periodic orbit 1 of the restricted three-body problem

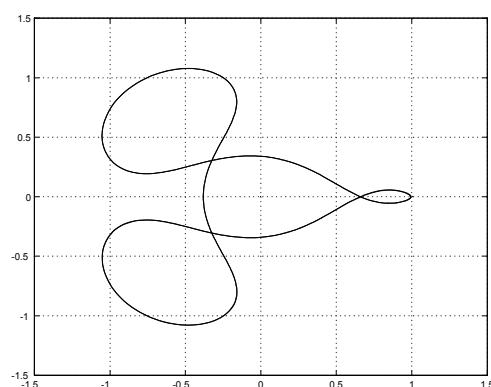


Figure 14.5: Periodic orbit 2 of the restricted three-body problem

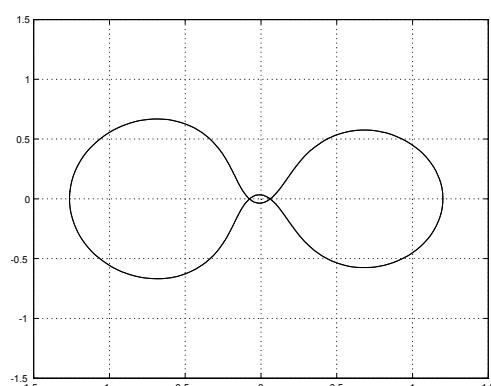


Figure 14.6: Periodic orbit 3 of the restricted three-body problem

```

tf2=11.124340337266085134999734047;
tf3=6.192169333131963970674;
x0=0.994; x03=1.2; y0=0; vx0=0.0; %Initial conditions
vy01=-2.00158510637908252240537862224;
vy02=-2.0317326295573368357302057924;
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)

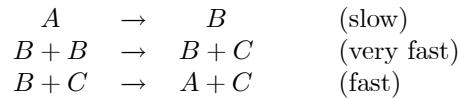
m2 = 0.012277471;
m1 = 1 - m2;
r13 = (((y(1) + m2)^2 + y(2)^2) ^ 1.5);
r23 = (((y(1) - m1)^2 + y(2)^2) ^ 1.5);
dydt = [ y(3)
          y(4)
          (2*y(4) + y(1) - m1*((y(1)+m2)/r13)...
          - m2*((y(1)-m1)/r23))
          (-2*y(3) + y(2) - m1*(y(2)/r13)...
          - m2*(y(2)/r23)) ];

```

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



in a closed tank has the mass balance equations

$$\begin{aligned}
 \dot{y}_1 &= -0.04y_1 + 10^4y_2y_3 & y_1(0) &= 1 \\
 \dot{y}_2 &= 0.04y_1 - 10^4y_2y_3 - 3 \cdot 10^7y_2^2 & y_2(0) &= 0 \\
 \dot{y}_3 &= 3 \cdot 10^7y_2^2 & y_3(0) &= 0
 \end{aligned} \tag{14.13}$$

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

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, t), \quad \mathbf{y}(t_0) = \mathbf{y}_0 \quad (14.14)$$

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  $(t_0, t_1, \dots, t_n, \dots, t_N)$  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,

$$\dot{\mathbf{y}}_L(t_n; t) = \mathbf{f}[\mathbf{y}_L(t_n; t)], \quad \mathbf{y}_L(t_n; t_n) = \mathbf{y}_n \quad (14.15)$$

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 difference of the computed solution  $\mathbf{y}_{n+1}$  from the local solution  $\mathbf{y}_L(t_n; t_{n+1})$  at time  $t_{n+1}$ :

$$\mathbf{e}_{n+1} = \mathbf{y}_{n+1} - \mathbf{y}_L(t_n; t_{n+1}) \quad (14.16)$$

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  $t_{n+1}$ :

$$\mathbf{E}_{n+1} = \mathbf{y}_{n+1} - \mathbf{y}(t_{n+1}) \quad (14.17)$$

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

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\phi(\mathbf{y}_n, t_n) \quad (14.18)$$

where  $\phi(\cdot)$  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

$$\mathbf{e}_{n+1} = O(h^{p+1}) \quad (14.19)$$

Here we have used the order notation  $O(\cdot)$  (Lin and Segel 1974). The function  $\phi(x)$  satisfies

$$\phi(x) = O[\psi(x)] \quad (14.20)$$

if there exists a constant  $C > 0$  so that

$$|\phi(x)| \leq C |\psi(x)| \quad (14.21)$$

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

$$\mathbf{y}_L(t_n; t_{n+1}) = \mathbf{y}_n + h\mathbf{f}(\mathbf{y}_n, t_n) + \dots + \frac{h^p}{p!} \frac{d^{p-1}\mathbf{f}(\mathbf{y}_n, t_n)}{dt^{p-1}} + \frac{h^{p+1}}{(p+1)!} \frac{d^p\mathbf{f}[\mathbf{y}_L(\tau), \tau]}{dt^p} \quad (14.22)$$

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

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(\mathbf{y}_n, t) + \dots + \frac{h^p}{p!} \frac{d^{p-1}\mathbf{f}(\mathbf{y}_n, t_n)}{dt^{p-1}} + O(h^{p+1}) \quad (14.23)$$

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

$$\dot{\mathbf{y}}^* = \mathbf{f}(\mathbf{y}^*, t), \quad \mathbf{y}^*(t_0) = \mathbf{y}_0^* \quad (14.24)$$

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

$$\dot{\mathbf{y}}^* + \Delta\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}^*, t) + \frac{\partial \mathbf{f}(\mathbf{y}, t)}{\partial \mathbf{y}} \Big|_{\mathbf{y}=\mathbf{y}^*} \Delta\mathbf{y} \quad (14.25)$$

We define the *Jacobian*  $\mathbf{J}$  of the system to be

$$\mathbf{J} = \frac{\partial \mathbf{f}(\mathbf{y}, t)}{\partial \mathbf{y}} \Big|_{\mathbf{y}=\mathbf{y}^*} = \left\{ \frac{\partial f_i(\mathbf{y}, t)}{\partial y_j} \Big|_{\mathbf{y}=\mathbf{y}^*} \right\} \quad (14.26)$$

and obtain the *linearization* of (14.14) which is

$$\Delta\dot{\mathbf{y}} = \mathbf{J}\Delta\mathbf{y} \quad (14.27)$$

The solution  $\Delta\mathbf{y}(t)$  of (14.27) can be expressed as a linear combination

$$\Delta\mathbf{y} = \sum_{i=1}^d q_i(t) \mathbf{m}_i \quad (14.28)$$

of solutions  $q_i(t)$  of the scalar differential equations

$$\dot{q}_i = \lambda_i q_i, \quad i = 1, \dots, d \quad (14.29)$$

where  $\mathbf{q} = (q_1, \dots, q_n)^T$ ,  $\lambda_i$  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

$$\Delta\mathbf{y}_{n+1} = \sum_{i=1}^d (q_i)_{n+1} \mathbf{m}_i \quad (14.30)$$

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

$$(q_i)_{n+1} = R(h\lambda_i)(q_i)_n \quad (14.31)$$

when applied to (14.29). Then

$$\Delta\mathbf{y}_n = \sum_{i=1}^d R^n(h\lambda_i)(q_i)_0 \mathbf{m}_i \quad (14.32)$$

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

$$|R(h\lambda_i)| \leq 1, \quad i = 1, \dots, d \quad (14.33)$$

where  $h$  is the step-length and  $\lambda_i$  is an eigenvalue of  $\mathbf{J}$ .

**Example 211** Consider the system

$$\dot{y}_1 = y_2 \quad (14.34)$$

$$\dot{y}_2 = -\gamma y_1^3 - cy_2 \quad (14.35)$$

The linearization around  $y_1 = 0, y_2 = 0$  is

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & -c \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \quad (14.36)$$

which has eigenvalues  $\lambda_1 = 0$  and  $\lambda_2 = -c$ . The linearization around a solution  $y_1^*(t), y_2^*(t)$  is

$$\begin{pmatrix} \Delta\dot{y}_1 \\ \Delta\dot{y}_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -3\gamma(y_1^*)^2 & -c \end{pmatrix} \begin{pmatrix} \Delta y_1 \\ \Delta y_2 \end{pmatrix} \quad (14.37)$$

The eigenvalues are given by

$$\lambda^2 + c\lambda + 3\gamma(y_1^*)^2 = 0 \quad (14.38)$$

which gives

$$\lambda = -\frac{c}{2} \pm \sqrt{\left(\frac{c}{2}\right)^2 - 3\gamma(y_1^*)^2} \quad (14.39)$$

which implies that  $\text{Re}[\lambda] \leq 0$ . We see that for large  $|y_1^*|$ , that is when  $3\gamma(y_1^*)^2 \gg \left(\frac{c}{2}\right)^2$ , 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*

$$\dot{y} = \lambda y \quad (14.40)$$

The numerical solution for this system is

$$y_{n+1} = R(h\lambda)y_n \quad (14.41)$$

where  $R(h\lambda)$  is the stability function for the method. Stability of the numerical scheme is ensured if the difference equation satisfies

$$|y_{n+1}| \leq |y_n| \quad (14.42)$$

and we see that this is ensured if

$$|R(h\lambda)| \leq 1 \quad (14.43)$$

This gives conditions on the time-step  $h$  and the location of the eigenvalue  $\lambda$  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

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(\mathbf{y}_n, t_n) \quad (14.44)$$

Comparison with (14.23) shows that the method is of order 1.

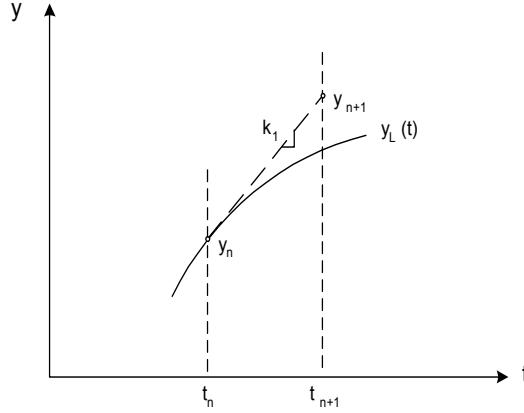


Figure 14.7: Euler's method

The linear stability of Euler's method can be investigated with the scalar test system

$$\dot{y} = \lambda y \quad (14.45)$$

Euler's method gives

$$y_{n+1} = y_n + h\lambda y_n = (1 + h\lambda)y_n \quad (14.46)$$

which shows that the stability function is

$$R(h\lambda) = 1 + h\lambda \quad (14.47)$$

Stability is ensured whenever

$$|R(h\lambda)| = |1 + h\lambda| \leq 1 \quad (14.48)$$

This is the case if  $h\lambda$  is inside the circle of radius one around  $-1$ . For real eigenvalues  $\lambda$  stability is ensured when

$$-\frac{2}{h} \leq \lambda \leq 0 \quad (14.49)$$

or, equivalently,

$$h \leq -\frac{2}{\lambda} \quad (14.50)$$

The region of stability is shown in Figure 14.14.

**Example 212** *The system*

$$\dot{y} = -y, \quad y(0) = 1 \quad (14.51)$$

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.

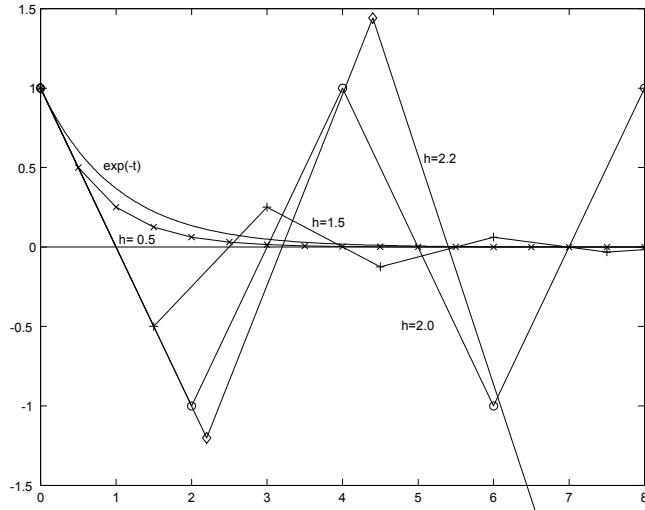


Figure 14.8: Calculated solutions for the system  $\dot{y} = -y$  with Euler's method for four different time steps  $h$ . The exact solution  $\exp(-t)$  is shown for comparison.

**Example 213** *The first order system*

$$\dot{y} = -\beta y^3 \quad (14.52)$$

in Euler's method gives the algorithm

$$y_{n+1} = y_n - h\beta y_n^3 \quad (14.53)$$

The linearization of the differential equation around zero gives

$$\dot{y} = 0 \quad (14.54)$$

that is, the test system  $\dot{y} = \lambda y$  with  $\lambda = 0$ . This is stable for all time steps  $h$ .

**Example 214** *The system*

$$\dot{y} = -\alpha y - \beta y^3 \quad (14.55)$$

in Euler's method gives

$$y_{n+1} = y_n - h(\alpha y_n + \beta y_n^3) \quad (14.56)$$

The linearization of the differential equation around zero gives

$$\dot{y} = -\alpha y \quad (14.57)$$

while the linearization around  $y^*$  gives

$$\Delta \dot{y} = - \left[ \alpha + 3\beta (y^*)^2 \right] \Delta y \quad (14.58)$$

A large  $|y^*|$  requires a small  $h$  for the stability condition to hold. Here, the eigenvalue is  $\lambda = - \left[ \alpha + 3\beta (y^*)^2 \right]$ , and the stability condition for the linearized system is

$$h \leq \frac{2}{\alpha + 3\beta (y^*)^2} \quad (14.59)$$

**Example 215** Consider the second order system

$$\ddot{x} = F(x, \dot{x}) \quad (14.60)$$

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

$$\dot{y}_1 = y_2 \quad (14.61)$$

$$\dot{y}_2 = F(y_1, y_2) \quad (14.62)$$

and Euler's method gives the integration algorithm

$$y_{1,n+1} = y_{1,n} + hy_{2,n} \quad (14.63)$$

$$y_{2,n+1} = y_{2,n} + hF(y_1, y_2) \quad (14.64)$$

### 14.3.2 The improved Euler method

The *improved Euler method* includes an evaluation  $\hat{\mathbf{y}}_{n+1} = \mathbf{y}_n + h\mathbf{f}(\mathbf{y}_n, t_n)$  according to Euler's method. Then an approximation of  $\mathbf{f}(\hat{\mathbf{y}}_{n+1}, t_{n+1})$  at the time  $t_{n+1}$  is computed using  $\hat{\mathbf{y}}_{n+1}$ . This value is used to improve the accuracy of the numerical solution  $\mathbf{y}_{n+1}$ . The method is given by

$$\mathbf{k}_1 = \mathbf{f}(\mathbf{y}_n, t_n) \quad (14.65)$$

$$\mathbf{k}_2 = \mathbf{f}(\mathbf{y}_n + h\mathbf{k}_1, t_n + h) \quad (14.66)$$

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{h}{2} (\mathbf{k}_1 + \mathbf{k}_2) \quad (14.67)$$

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

$$\mathbf{k}_2 = \mathbf{f}(\mathbf{y}_n, t_n) + h \frac{d\mathbf{f}}{dt}(\mathbf{y}_n, t_n) + \frac{h^2}{2} \frac{d^2\mathbf{f}}{dt^2}(\mathbf{y}_n, t_n) + O(h^3) \quad (14.68)$$

This gives the Taylor series

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(\mathbf{y}_n, t_n) + \frac{h^2}{2} \frac{d\mathbf{f}}{dt}(\mathbf{y}_n, t_n) + \frac{h^3}{4} \frac{d^2\mathbf{f}}{dt^2}(\mathbf{y}_n, t_n) + O(h^4) \quad (14.69)$$

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*

$$\dot{y} = -\alpha y^3 \quad (14.70)$$

in the improved Euler method gives the algorithm

$$k_1 = -\alpha y_n^3 \quad (14.71)$$

$$k_2 = -\alpha(y_n + hk_1)^3 \quad (14.72)$$

$$y_{n+1} = y_n + \frac{h}{2}(k_1 + k_2) \quad (14.73)$$

**Example 217** *Consider the second order system*

$$\ddot{x} + c\dot{x} + \gamma(x)x = 0 \quad (14.74)$$

which is set in standard form by using  $y_1 = x$  and  $y_2 = \dot{x}$ . Then the differential equation is written

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \end{pmatrix} = \begin{pmatrix} y_2 \\ -\gamma(y_1)y_1 - cy_2 \end{pmatrix} \quad (14.75)$$

The improved Euler's method gives the integration algorithm

$$\begin{pmatrix} k_{1,1} \\ k_{1,2} \end{pmatrix} = \begin{pmatrix} y_{2,n} \\ -\gamma(y_{1,n})y_{1,n} - cy_{2,n} \end{pmatrix} \quad (14.76)$$

$$\begin{pmatrix} k_{2,1} \\ k_{2,2} \end{pmatrix} = \begin{pmatrix} y_{2,n} + hk_{1,2} \\ -\gamma(y_{1,n} + hk_{1,1})(y_{1,n} + hk_{1,1}) - c(y_{2,n} + hk_{1,2}) \end{pmatrix} \quad (14.77)$$

$$y_{1,n+1} = y_{1,n} + \frac{h}{2}(k_{1,1} + k_{2,1}) \quad (14.78)$$

$$y_{2,n+1} = y_{2,n} + \frac{h}{2}(k_{1,2} + k_{2,2}) \quad (14.79)$$

### 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  $(\mathbf{y}(t + \frac{h}{2}), t + \frac{h}{2})$  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

$$\mathbf{k}_1 = \mathbf{f}(\mathbf{y}_n, t_n) \quad (14.80)$$

$$\mathbf{k}_2 = \mathbf{f}\left(\mathbf{y}_n + \frac{h}{2}\mathbf{k}_1, t_n + \frac{h}{2}\right) \quad (14.81)$$

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{k}_2 \quad (14.82)$$

A Taylor series expansion of  $\mathbf{k}_2$  gives

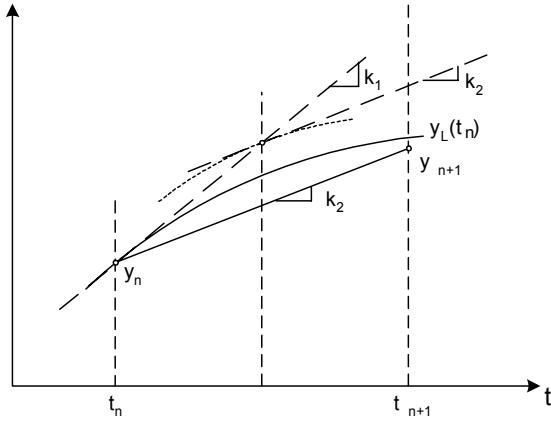


Figure 14.9: The modified Euler method

$$\mathbf{k}_2 = \mathbf{f}(\mathbf{y}_n, t_n) + \frac{h}{2} \frac{d\mathbf{f}}{dt}(\mathbf{y}_n, t_n) + \frac{\left(\frac{h}{2}\right)^2}{2} \frac{d^2\mathbf{f}}{dt^2}(\mathbf{y}_n, t_n) + O(h^3) \quad (14.83)$$

which gives

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(\mathbf{y}_n, t_n) + \frac{h^2}{2} \frac{d\mathbf{f}}{dt}(\mathbf{y}_n, t_n) + \frac{h^3}{8} \frac{d^2\mathbf{f}}{dt^2}(\mathbf{y}_n, t_n) + O(h^4) \quad (14.84)$$

The method is seen to be of order  $p = 2$ .

Application of the method to the test systems  $\dot{y} = \lambda y$  gives

$$\begin{aligned} k_2 &= \lambda \left(1 + \frac{h}{2}\lambda\right) y_n \\ y_{n+1} &= \left(1 + h\lambda + \frac{(h\lambda)^2}{2}\right) y_n \end{aligned}$$

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  $\mathbf{f}[\mathbf{y}(t_n + ch), t_n + ch]$  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  $\sigma$  stages if  $\sigma$  approximations, or stages, of the function derivative  $\mathbf{f}$  is used.

### 14.4.2 Numerical scheme

An explicit Runge-Kutta method with  $\sigma$  stages for the system

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, t) \quad (14.85)$$

is given by

$$\begin{aligned} \mathbf{k}_i &= \mathbf{f}\left(\mathbf{y}_n + h \sum_{j=1}^{i-1} a_{ij} \mathbf{k}_j, t_n + c_i h\right), \quad i = 1, \dots, \sigma \\ \mathbf{y}_{n+1} &= \mathbf{y}_n + h \sum_{j=1}^{\sigma} b_j \mathbf{k}_j \end{aligned}$$

The explicit Runge-Kutta method can be written out as

$$\mathbf{k}_1 = \mathbf{f}(\mathbf{y}_n, t_n) \quad (14.86)$$

$$\mathbf{k}_2 = \mathbf{f}(\mathbf{y}_n + h a_{21} \mathbf{k}_1, t_n + c_2 h) \quad (14.87)$$

$$\mathbf{k}_3 = \mathbf{f}(\mathbf{y}_n + h (a_{31} \mathbf{k}_1 + a_{32} \mathbf{k}_2), t_n + c_3 h) \quad (14.88)$$

$$\vdots \quad (14.89)$$

$$\mathbf{k}_\sigma = \mathbf{f}(\mathbf{y}_n + h (a_{\sigma 1} \mathbf{k}_1 + \dots + a_{\sigma, \sigma-1} \mathbf{k}_{\sigma-1}), t_n + c_\sigma h) \quad (14.90)$$

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h(b_1 \mathbf{k}_1 + \dots + b_\sigma \mathbf{k}_\sigma) \quad (14.91)$$

The equations for  $\mathbf{k}_1, \dots, \mathbf{k}_\sigma$  are called the stage computations. The interpolation parameters  $c_i, i \in \{2, \dots, \sigma\}$  are in the range  $0 \leq c_i \leq 1$  and form an increasing sequence, that is,  $0 \leq c_1 \leq \dots \leq c_\sigma \leq 1$ . The weighting parameters at stage  $i$  are denoted  $a_{ij}, i \in \{2, \dots, \sigma\}, j \in \{1, \dots, i-1\}$ , and satisfy the normalization condition

$$\sum_{j=1}^{i-1} a_{ij} = c_i \leq 1 \quad (14.92)$$

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

$$\begin{array}{c|ccccc} 0 & & & & & \\ c_2 & a_{21} & & & & \\ c_3 & a_{31} & a_{32} & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ c_\sigma & a_{\sigma 1} & a_{\sigma 2} & \dots & a_{\sigma, \sigma-1} & \\ \hline & b_1 & b_2 & \dots & b_{\sigma-1} & b_\sigma \end{array} \quad (14.93)$$

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  $\sigma$  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

$$\begin{array}{c|cc} 0 & & \\ \hline c_2 & a_{21} & \\ \hline & b_1 & b_2 \end{array}$$

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

$$\mathbf{y}_{n+1} = \mathbf{y}_n + (b_1 + b_2)h\mathbf{f}(\mathbf{y}_n, t_n) + b_2c_2h^2 \frac{d^2\mathbf{f}}{dt^2}(\mathbf{y}_n, t_n) + O(h^3) \quad (14.94)$$

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

$$b_1 + b_2 = 1 \quad (14.95)$$

$$b_2c_2 = \frac{1}{2} \quad (14.96)$$

**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). This 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

$$\begin{array}{c|cc} 0 & & \\ \hline 1 & 1 & \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

The modified Euler method has the array

$$\begin{array}{c|cc} 0 & & \\ \hline \frac{1}{2} & \frac{1}{2} & \\ \hline & 0 & 1 \end{array}$$

Heun's method has the following array

$$\begin{array}{c|ccc} 0 & & & \\ \hline \frac{1}{3} & & \frac{1}{3} & \\ \frac{2}{3} & 0 & \frac{2}{3} & \\ \hline & \frac{1}{4} & 0 & \frac{3}{4} \end{array}$$

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

$$\begin{array}{c|cccc} 0 & & & & \\ \hline \frac{1}{2} & & \frac{1}{2} & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{array}$$

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  $A = 0.01 \text{ m}^2$  and a vertical center axis pointing upwards with coordinate  $x$ . The cylinder

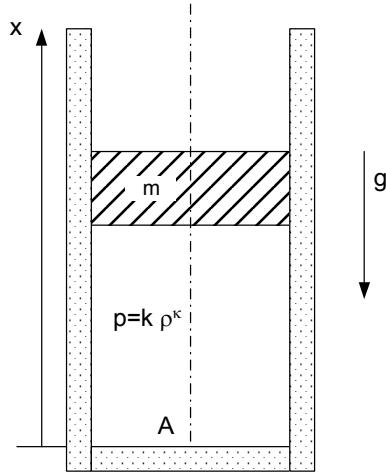


Figure 14.10: Pneumatic spring with gravity acting on the position.

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

$$\rho = \frac{m_a}{V_a} = \frac{m_a}{Ax} \quad (14.97)$$

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

$$p = p_0 \left( \frac{\rho}{\rho_0} \right)^\kappa = p_0 \left( \frac{x_0}{x} \right)^\kappa \quad (14.98)$$

where  $\kappa = 1.4$  and  $p_0 = 2 \cdot 10^5$  N/m<sup>2</sup> 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:

$$F = -mg + Ap = -mg + Ap_0 \left( \frac{x_0}{x} \right)^\kappa \quad (14.99)$$

where  $g = 10$  m/s<sup>2</sup>. 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

$$\ddot{x} + g [1 - x^{-\kappa}] = 0 \quad (14.100)$$

The standard form is  $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$  is obtained by setting

$$\mathbf{y} = \begin{pmatrix} x \\ v \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} v \\ -g[1 - x^{-\kappa}] \end{pmatrix} \quad (14.101)$$

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

$$\Delta \dot{\mathbf{y}} = \mathbf{J} \Delta \mathbf{y} \quad (14.102)$$

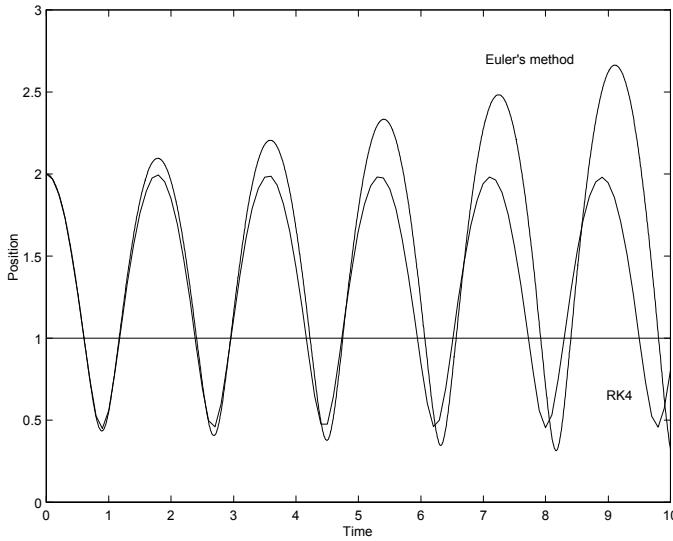


Figure 14.11: Position of piston integrated with Euler's method with  $h = 0.005$  s and with RK4 with  $h = 0.1$  s.

where

$$\Delta \mathbf{y} = \begin{pmatrix} x - 1 \\ v \end{pmatrix}, \quad \mathbf{J} = \begin{pmatrix} 0 & 1 \\ -g\kappa(x^*)^{-(\kappa+1)} & 0 \end{pmatrix} \quad (14.103)$$

The eigenvalues of the linearization are found to be

$$\lambda_{1,2} = \pm j\omega_0, \quad \omega_0 = \sqrt{g\kappa(x^*)^{-(\kappa+1)}} \quad (14.104)$$

Numerical values are

$x^*$	0.5	1	2
$\omega_0$	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  $\frac{1}{2}mv^2$ :

$$E = \frac{1}{\kappa - 1} p_0 A x^{-(\kappa-1)} + mgx + \frac{1}{2} mv^2 \quad (14.106)$$

The total energy has its minimum value at the equilibrium state where the energy is

$$E_{\min} = \frac{1}{\kappa - 1} p_0 A + mg = 7000 \text{ J} \quad (14.107)$$

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

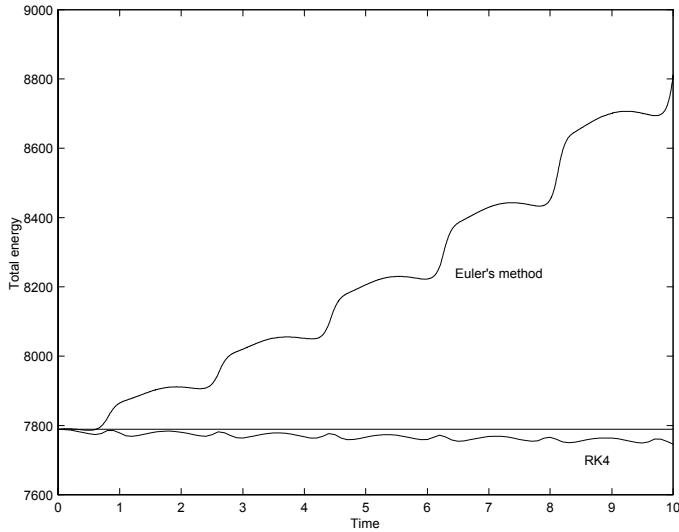


Figure 14.12: Total energy of solutions computed with Euler's method and RK4. It is seen that Euler's method increases the energy in the system, while RK4 gives a slight decrease in energy.

Euler's method gave a steady increase in energy which is not in agreement with the physics of the system as 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

$$\begin{aligned} k_1 &= \lambda y_n \\ &\vdots \\ k_\sigma &= \lambda [y_n + h(a_{\sigma 1}k_1 + \dots + a_{\sigma, \sigma-1}k_{\sigma-1})] \\ y_{n+1} &= y_n + h(b_1k_1 + \dots + b_\sigma k_\sigma) \end{aligned}$$

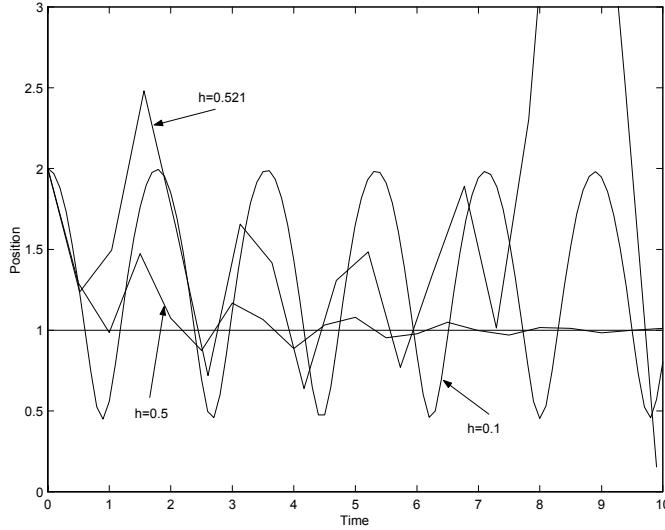


Figure 14.13: Simulation of pneumatic spring usning RK4 for three different step lengths.

In vector notation with  $\kappa = (k_1, k_2, \dots, k_\sigma)^T$  and  $\mathbf{1} = (1, 1, \dots, 1)^T$  this can be written

$$\kappa = \lambda(\mathbf{1}y_n + h\mathbf{A}\kappa) \quad (14.108)$$

$$y_{n+1} = y_n + h\mathbf{b}^T \kappa \quad (14.109)$$

Here  $\kappa$  can be solved from (14.108) and inserted into (14.109), which gives

$$R(h\lambda) = 1 + \lambda h \mathbf{b}^T (\mathbf{I} - h\lambda\mathbf{A})^{-1} \mathbf{1} \quad (14.110)$$

Alternatively, the system (14.108, 14.109) can be written

$$\begin{pmatrix} \mathbf{I} - h\lambda\mathbf{A} & \mathbf{0} \\ -h\mathbf{b}^T & 1 \end{pmatrix} \begin{pmatrix} \kappa \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} \lambda\mathbf{1} \\ 1 \end{pmatrix} y_n$$

From Cramer's rule it is seen that the stability function can be written

$$R(h\lambda) = \frac{\det [\mathbf{I} - \lambda h (\mathbf{A} - \mathbf{1}\mathbf{b}^T)]}{\det (\mathbf{I} - \lambda h\mathbf{A})} \quad (14.111)$$

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  $\det (\mathbf{I} - \lambda h\mathbf{A}) = 1$ . Using (14.111) we find:

For an explicit Runge-Kutta method the stability function can be written

$$R_E(h\lambda) = \det [\mathbf{I} - \lambda h (\mathbf{A} - \mathbf{1}\mathbf{b}^T)] \quad (14.112)$$

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  $\sigma$ .

**Example 219** Consider the improved Euler method where

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

Then

$$R_E(h\lambda) = \det \begin{pmatrix} 1 + \frac{\lambda h}{2} & \frac{\lambda h}{2} \\ -\frac{\lambda h}{2} & 1 + \frac{\lambda h}{2} \end{pmatrix} = 1 + \lambda h + \frac{(\lambda h)^2}{2} \quad (14.113)$$

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

$$y_L(t_n; t_{n+1}) = \left[ 1 + h\lambda + \frac{(h\lambda)^2}{2!} + \frac{(h\lambda)^3}{3!} + \dots \right] y_n \quad (14.114)$$

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

$$y_{n+1} = \left[ 1 + h\lambda + \frac{(h\lambda)^2}{2!} + \dots + \frac{(h\lambda)^p}{p!} + O(h^{p+1}) \right] y_n \quad (14.115)$$

It follows that the stability function for a explicit method of order  $p$  satisfies

$$R_E(\lambda h) = 1 + h\lambda + \frac{(h\lambda)^2}{2!} + \dots + \frac{(h\lambda)^p}{p!} + O(h^{p+1}) \quad (14.116)$$

The stability function of an explicit Runge-Kutta method with  $\sigma$  stages is a polynomial in  $\lambda h$  of degree less than or equal to the number of stages  $\sigma$ . If the method is of  $\sigma = p \leq 4$  stages the stability function must have exactly  $p$  terms, and this is only possible if

$$R_E(\lambda h) = 1 + h\lambda + \frac{(h\lambda)^2}{2!} + \dots + \frac{(h\lambda)^p}{p!} \quad \text{when } p = \sigma$$

**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.

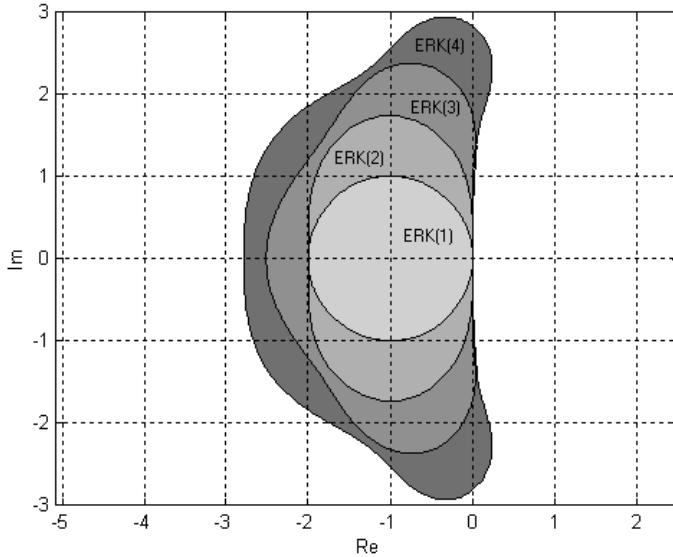


Figure 14.14: Regions of stability in  $s = h\lambda$  for the test system  $\dot{y} = \lambda y$  for the explicit Runge-Kutta methods. ERK(1): Euler's method, ERK(2): The modified and the improved Euler method, ERK(3): Heun's third order method, and ERK(4): The fourth order Rung-Kutta method RK4.

#### 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

$$\mathbf{k}_\sigma = \mathbf{f}(\mathbf{y}_{n+1}, t_{n+1}) \quad (14.117)$$

From the definition we see that in an FSAL method gives some savings in computations as

$$\mathbf{k}_\sigma^n = \mathbf{k}_1^{n+1} \quad (14.118)$$

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  $h |\lambda_{\max}|$  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  $\sigma$  stages for the system

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, t) \quad (14.119)$$

is given by

$$\mathbf{k}_1 = \mathbf{f}(\mathbf{y}_n + h(a_{11}\mathbf{k}_1 + \dots + a_{1\sigma}\mathbf{k}_\sigma), t_n + c_1 h) \quad (14.120)$$

$$\vdots \quad (14.121)$$

$$\mathbf{k}_\sigma = \mathbf{f}(\mathbf{y}_n + h(a_{\sigma 1}\mathbf{k}_1 + \dots + a_{\sigma\sigma}\mathbf{k}_\sigma), t_n + c_\sigma h) \quad (14.122)$$

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h(b_1\mathbf{k}_1 + \dots + b_\sigma\mathbf{k}_\sigma) \quad (14.123)$$

As for explicit Runge-Kutta methods, the interpolation parameters  $c_i$ ,  $i \in \{1, \dots, \sigma\}$  are in the range  $0 \leq c_i \leq 1$ . 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

$$y_{n+1} = y_n + \frac{h\lambda}{1 - h\lambda} y_n = \frac{1}{1 - h\lambda} y_n \quad (14.124)$$

The stability function is seen to be

$$R(h\lambda) = \frac{1}{1 - h\lambda} \quad (14.125)$$

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.

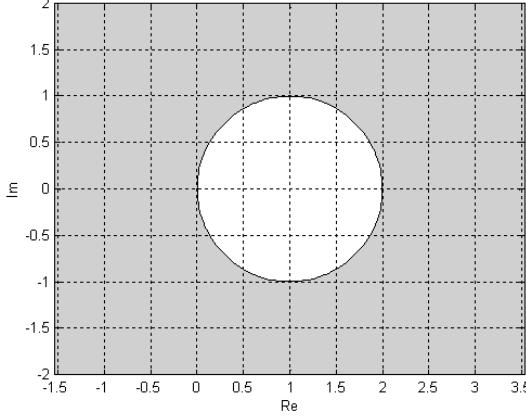


Figure 14.15: The shaded area shows where the implicit Euler method is stable as a function of the complex variable  $s = \lambda h$ .

#### 14.5.4 Trapezoidal rule

Consider the implicit Runge-Kutta method

$$\mathbf{k}_1 = \mathbf{f}(\mathbf{y}_n, t_n) \quad (14.126)$$

$$\mathbf{k}_2 = \mathbf{f}\left[\mathbf{y}_n + \frac{h}{2}(\mathbf{k}_1 + \mathbf{k}_2), t_n + h\right] \quad (14.127)$$

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{h}{2}(\mathbf{k}_1 + \mathbf{k}_2) \quad (14.128)$$

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

$$\mathbf{k}_2 = \mathbf{f}(\mathbf{y}_{n+1}, t_{n+1}) \quad (14.129)$$

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

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{h}{2} [\mathbf{f}(\mathbf{y}_n, t_n) + \mathbf{f}(\mathbf{y}_{n+1}, t_{n+1})] \quad (14.130)$$

which is known as the *trapezoidal rule*.

The stability function is found from (14.130) which for the test equation gives

$$y_{n+1} = y_n + \frac{h\lambda}{2} (y_n + y_{n+1}) \quad (14.131)$$

and it follows that

$$R(\lambda h) = \frac{1 + \frac{h\lambda}{2}}{1 - \frac{h\lambda}{2}} \quad (14.132)$$

We see that

$$|R(\lambda h)|^2 = \frac{\left(1 + \operatorname{Re}\left[\frac{h\lambda}{2}\right]\right)^2 + \left(\operatorname{Im}\left[\frac{h\lambda}{2}\right]\right)^2}{\left(1 - \operatorname{Re}\left[\frac{h\lambda}{2}\right]\right)^2 + \left(\operatorname{Im}\left[\frac{h\lambda}{2}\right]\right)^2} \quad (14.133)$$

and it follows that  $|R(\lambda h)| \leq 1$  and the method is stable for all  $\lambda$  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.

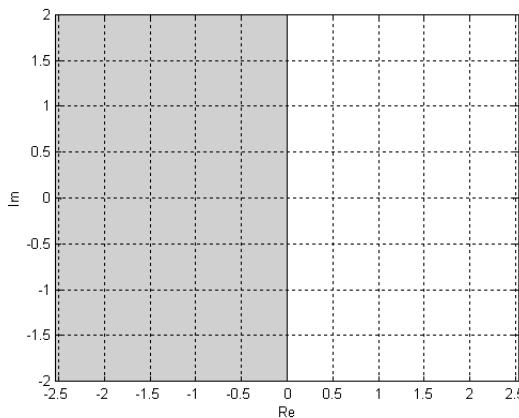


Figure 14.16: The shaded area shows where the trapezoidal rule is stable as a function of the complex variable  $s = \lambda h$ .

### 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

$$\mathbf{y}_{n+1} - \mathbf{y}_n = h\mathbf{f} \left[ \mathbf{y}_n + \frac{1}{2} (\mathbf{y}_{n+1} - \mathbf{y}_n), t_n + \frac{h}{2} \right]$$

which is simplifies to the following scheme

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f} \left( \frac{\mathbf{y}_n + \mathbf{y}_{n+1}}{2}, t_n + \frac{h}{2} \right) \quad (14.134)$$

which is called the implicit mid-point rule.

From (14.134) we find that the stability function for the implicit mid-point rule is

$$R(\lambda h) = \frac{1 + \frac{h\lambda}{2}}{1 - \frac{h\lambda}{2}} \quad (14.135)$$

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

$$\mathbf{k}_1 = \mathbf{f}(\mathbf{y}_n, t_n) \quad (14.136)$$

$$\mathbf{k}_2 = \mathbf{f}[\mathbf{y}_n + h[\theta\mathbf{k}_1 + (1-\theta)\mathbf{k}_2], t_n + h] \quad (14.137)$$

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h[\theta\mathbf{k}_1 + (1-\theta)\mathbf{k}_2] \quad (14.138)$$

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  $\mathbf{k}_2 = \mathbf{f}(\mathbf{y}_{n+1}, t_{n+1})$ . Then the expression for  $\mathbf{y}_{n+1}$  can be rewritten in the form

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h[\theta\mathbf{f}(\mathbf{y}_n, t_n) + (1-\theta)\mathbf{f}(\mathbf{y}_{n+1}, t_{n+1})]$$

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

$$R(h\lambda) = \frac{1 + h\lambda\theta}{1 - h\lambda(1-\theta)} \quad (14.139)$$

#### 14.5.7 Stability function

The application of an implicit Runge-Kutta method to a linear test system gives

$$\boldsymbol{\kappa} = \lambda(\mathbf{1}\mathbf{y}_n + h\mathbf{A}\boldsymbol{\kappa}) \quad (14.140)$$

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{b}^T \boldsymbol{\kappa} \quad (14.141)$$

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

$$R(h\lambda) = \left[ 1 + \lambda h \mathbf{b}^T (\mathbf{I} - h\lambda \mathbf{A})^{-1} \mathbf{1} \right] \quad (14.142)$$

$$R(h\lambda) = \frac{\det [\mathbf{I} - \lambda h (\mathbf{A} - \mathbf{1}\mathbf{b}^T)]}{\det (\mathbf{I} - \lambda h \mathbf{A})} \quad (14.143)$$

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.

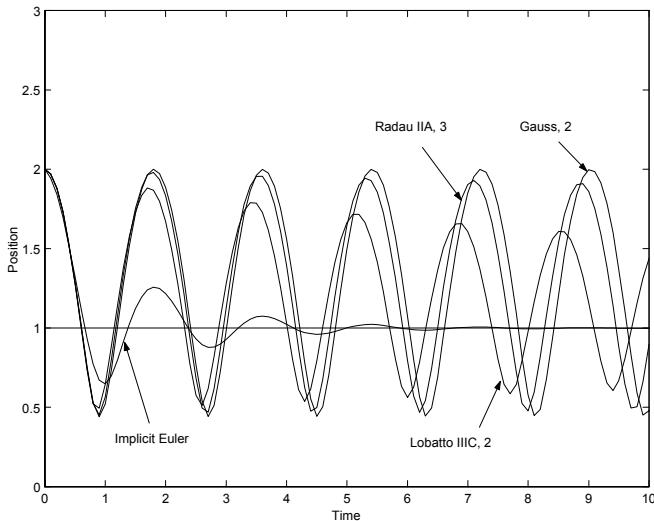


Figure 14.17: Position of piston computed with the implicit Runge-Kutta methods Gauss of order 2, Radau IIA of order 3, Lobatto IIIC of order 2 and the implicit Euler method.

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

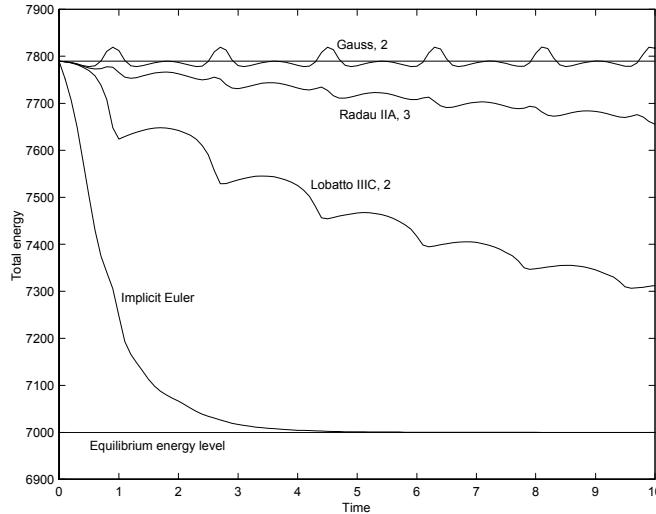


Figure 14.18: Total energy for the pneumatic system when the solution is computed with the implicit Runge-Kutta methods. The energy oscillates around the correct value for the Gauss of order 2, while the energy is numerically dissipated with the methods Radau IIA of order 3, Lobatto IIIC of order 2 and 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

$$\ddot{x}_1 + g \left( 1 - \frac{m}{m_1} x_1^{-\kappa} \right) + \frac{\omega_2^2}{2} (x_1 - x_2) = 0 \quad (14.144)$$

$$\ddot{x}_2 + g + \frac{\omega_2^2}{2} (x_2 - x_1) = 0 \quad (14.145)$$

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

$$\mathbf{y} = \begin{pmatrix} x_1 \\ v_1 \\ x_2 \\ v_2 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} v_1 \\ -g(1 - \frac{m}{m_1} x_1^{-\kappa}) - \frac{\omega_2^2}{2} (x_1 - x_2) \\ v_2 \\ -g - \frac{\omega_2^2}{2} (x_2 - x_1) \end{pmatrix} \quad (14.146)$$

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

$$\Delta \dot{\mathbf{y}} = \mathbf{J} \Delta \mathbf{y} \quad (14.147)$$

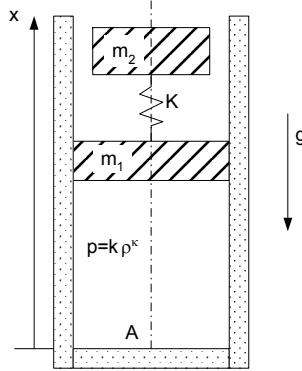


Figure 14.19: Pneumatic spring with mechanical resonance in load. The  $x$  coordinate of mass  $m_2$  is given an offset so that  $x_1 = x_2$  when the spring is unloaded.

where

$$\Delta \mathbf{y} = \begin{pmatrix} x_1 - x_1^* \\ v_1 \\ x_2 - x_2^* \\ v_2 \end{pmatrix}, \quad \mathbf{J} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -g \frac{m}{m_1} \kappa (x^*)^{-(\kappa-1)} - \frac{\omega_2^2}{2} & 0 & \frac{\omega_2^2}{2} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{\omega_2^2}{2} & 0 & -\frac{\omega_2^2}{2} & 0 \end{pmatrix} \quad (14.148)$$

The eigenvalues of the linearization around the equilibrium are found to be

$$\lambda_{1,2} = \pm j\omega_1, \quad \omega_1 = 3.7 \text{ rad/s} \quad (14.149)$$

$$\lambda_{3,4} = \pm j\omega_2, \quad \omega_2 = 1000 \text{ rad/s} \quad (14.150)$$

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.

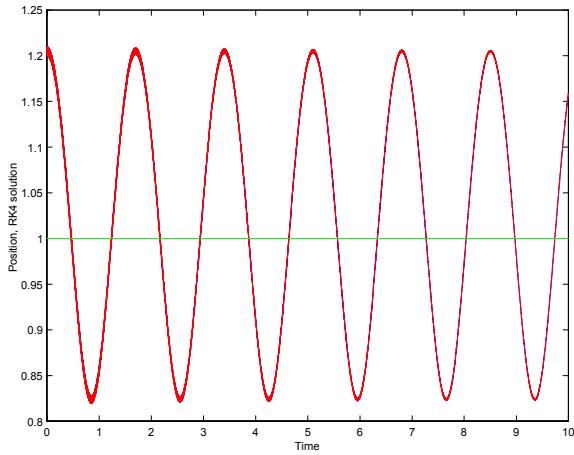


Figure 14.20: Position of the two masses computed with RK4 with time step  $h = 0.0005$  s.

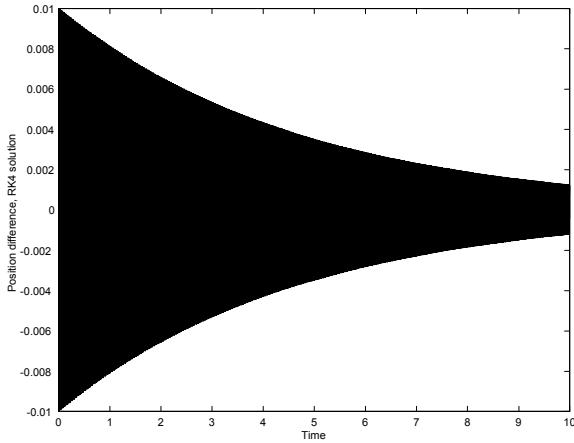


Figure 14.21: Offset equilibrium for the spring between masses one and two computed with RK4 with  $h = 0.0005$  s. The oscillation is seen to be lightly damped by the integration method.

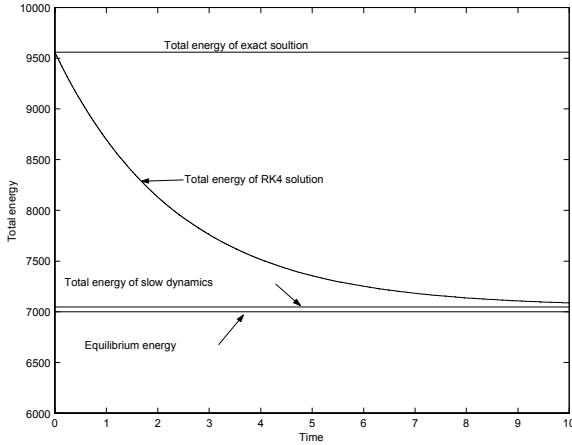


Figure 14.22: Total energy corresponding to the numerical solution computed with RK4 with  $h = 0.0005$  s. It is seen that the energy related to the fast dynamics is slowly damped out.

## 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

$$\lambda = \sigma + j\omega \quad (14.151)$$

where  $\sigma$  is the real part and  $j\omega$  is the imaginary part. It is assumed that  $\omega < \pi/h$ , that is,  $\omega$  is assumed to be less than the Nyquist frequency  $\pi/h$ . The local solution of the test system is

$$y_L(t_n; t_{n+1}) = e^{\lambda h} y_n$$

Consider a system  $\dot{y} = \mu y$ , which has the local solution

$$y_L(t_n; t_{n+1}) = e^{\mu h} y_n \quad (14.152)$$

The two systems will give the same local solutions at  $t_{n+1}$  whenever  $e^{\lambda h} = e^{\mu h}$  which is implied by

$$\mu = \lambda + j2k\frac{\pi}{h} = \sigma + j\left(\omega + 2k\frac{\pi}{h}\right), \quad k = 0, \pm 1, \pm 2, \dots \quad (14.153)$$

If

$$\mu = \lambda + 2kj\frac{\pi}{h}, \quad k = \pm 1, \pm 2, \dots \quad (14.154)$$

then the system  $\dot{y} = \mu y$  where  $\text{Im}(\lambda) > \pi/h$  will have the same solution as the system  $\dot{y} = \lambda y$  where  $\text{Im}(\lambda) < \pi/h$ . This phenomenon is called aliasing.

### 14.6.2 A-stability, L-stability

The test system  $\dot{y} = \lambda y$  is stable when  $\text{Re}\lambda \leq 0$ . 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

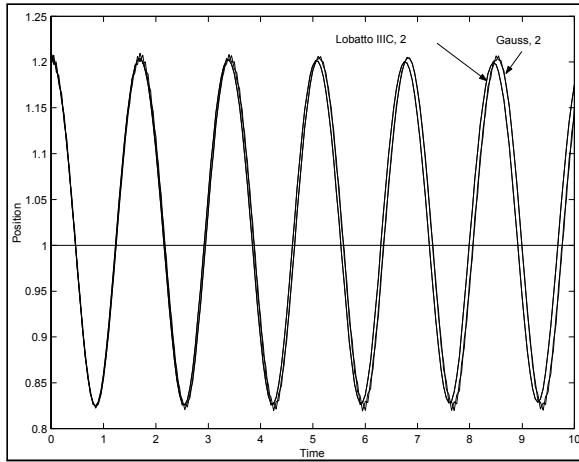


Figure 14.23: Position of the two masses computed with a Gauss method of order 2 and a Lobatto IIIC method of order 3. The time step was  $h = 0.05$  with both methods.

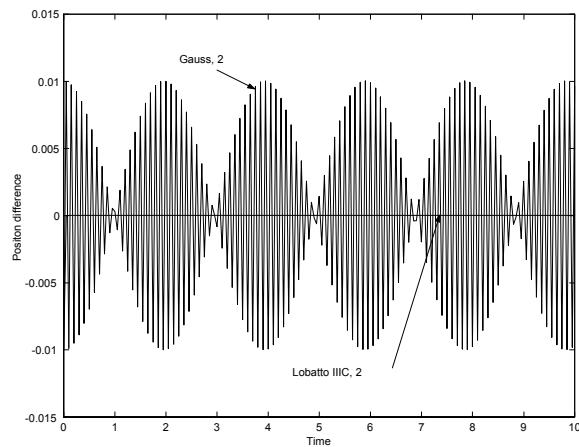


Figure 14.24: Offset in position between the two masses calculated with Gauss order 2 and Lobatto IIIC order 2 with  $h = 0.05$  s. The Gauss method gave no damping, but the energy of the fast dynamics was shifted to frequencies below the Nyquist frequency  $\omega_N = 62.8$  rad/s. The Lobatto method damped out the fast dynamics in one step.

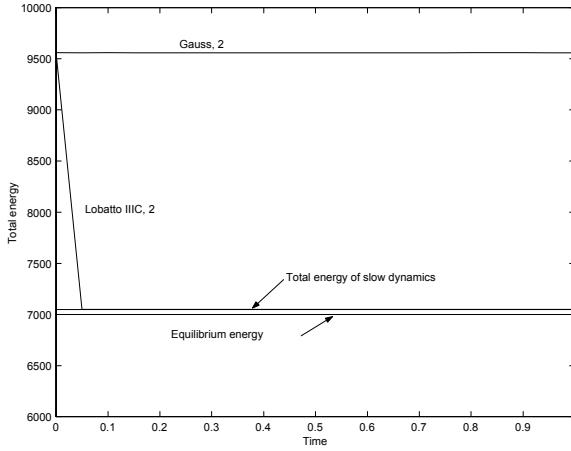


Figure 14.25: Total energy corresponding to the numerical solution of Gauss order 2 and Lobatto IIIC order 2. The Gauss method gave a constant total energy in agreement with the exact solution, while the Lobatto method damped out the energy associated with the fast dynamics.

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  $\text{Re } \lambda \leq 0$ .

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 than 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  $\pi/h$ . 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  $|R(j\omega h)| \rightarrow 0$  when  $\omega \rightarrow \infty$  for all systems  $\dot{y} = \lambda y$  where  $\lambda = j\omega$ .

**Example 221** We note that explicit Runge-Kutta methods have stability functions

$$R_E(\lambda h) = \det \left[ \mathbf{I} - \lambda h \left( \mathbf{A} - \mathbf{1}\mathbf{b}^T \right) \right] \quad (14.155)$$

It is clear that  $|R_E(\lambda h)| \rightarrow \infty$  whenever  $|\lambda| \rightarrow \infty$ , 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

$$R(h\lambda) = \left[ 1 + \lambda h \mathbf{b}^T (\mathbf{I} - h\lambda \mathbf{A})^{-1} \mathbf{1} \right]$$

(14.110) can be used also for implicit Runge-Kutta methods. Suppose that  $\mathbf{A}$  is nonsingular, and consider the case where  $\lambda h$  tends to infinity. Then if the limit  $R(\infty) := \lim_{s \rightarrow \infty} R(s)$  exists, it is given by

$$\begin{aligned} R(\infty) &= \lim_{s \rightarrow \infty} \left[ 1 + s \mathbf{b}^T (\mathbf{I} - s \mathbf{A})^{-1} \mathbf{1} \right] \\ &= \lim_{s \rightarrow \infty} \left[ 1 - s \mathbf{b}^T (s \mathbf{A})^{-1} \mathbf{1} \right] \\ &= 1 - \mathbf{b}^T \mathbf{A}^{-1} \mathbf{1} \end{aligned} \quad (14.156)$$

Moreover, it is noted that for an implicit Runge-Kutta method where

$$\mathbf{k}_\sigma = \mathbf{f}(\mathbf{y}_{n+1}, t_{n+1}) \quad (14.157)$$

then the weighting vector  $\mathbf{b}$  is equal to the last row in the stage matrix  $\mathbf{A}$ . This gives

$$\mathbf{b} = \mathbf{A}^T \mathbf{e}_\sigma \quad (14.158)$$

where  $\mathbf{e}_\sigma = (0, 0, \dots, 1)^T$  is a  $\sigma$ -dimensional unit vector. Insertion of (14.158) into (14.156) gives

$$R(\infty) = 1 - \lambda h \mathbf{e}_\sigma^T \mathbf{A} (\lambda h \mathbf{A})^{-1} \mathbf{1} = 1 - \mathbf{e}_\sigma^T \mathbf{1} = 0 \quad (14.159)$$

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  $\lambda_i$  that are large in the sense that  $|\lambda_i h|$  are much larger than 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  $\lambda_j$  so that  $|\lambda_j h| \rightarrow \infty$ , 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  $\omega$ .

**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

$$\begin{aligned} R(s) &= 1 + s\mathbf{b}^T(\mathbf{I} - s\mathbf{A})^{-1}\mathbf{1} \\ &= 1 + \frac{s}{2} \left[ \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \frac{\frac{s}{2}}{1-\frac{s}{2}} & \frac{1}{1-\frac{s}{2}} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right] \\ &= \frac{1 + \frac{s}{2}}{1 - \frac{s}{2}} \end{aligned}$$

which means that

$$|R(j\omega)| = 1$$

for all  $\omega$ .

#### 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

$$y_L(t_n; t_{n+1}) = e^{\lambda h} y_n \quad (14.160)$$

while the numerical solution is

$$y_{n+1} = R(\lambda h) y_n \quad (14.161)$$

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

$$e^s - R(s) \quad (14.162)$$

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  $\sigma$  stages approximates the exponential function  $e^s$  by the polynomial approximation

$$R(s) = 1 + \beta_1 s + \dots + \beta_\sigma s^\sigma$$

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

$$R(s) = 1 + s + \dots + \frac{s^p}{p!}$$

An implicit Runge-Kutta method of  $\sigma$  stages approximates  $e^s$  by the rational approximation

$$R(s) = \frac{1 + \beta_1 s + \dots + \beta_k s^k}{1 + \gamma_1 s + \dots + \gamma_m s^m}$$

		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	$\frac{1}{1-s+\frac{s^2}{2!}}$	$\frac{1 + \frac{1}{3}s}{1 - \frac{2}{3}s + \frac{1}{6}s^2}$	$\frac{1 + \frac{1}{2}s + \frac{1}{12}s^2}{1 - \frac{2}{3}s + \frac{1}{12}s^2}$	$\frac{1 + \frac{2}{5}s + \frac{3}{20}s^2 + \frac{1}{60}s^3}{1 - \frac{2}{5}s + \frac{1}{20}s^2}$
	3	$\frac{1}{1-s+\frac{s^2}{2!}-\frac{s^3}{3!}}$	$\frac{1 + \frac{1}{4}s}{1 - \frac{3}{4}s + \frac{1}{4}s^2 - \frac{1}{24}s^3}$	$\frac{1 + \frac{2}{3}s + \frac{1}{20}s^2}{1 - \frac{3}{5}s + \frac{3}{20}s^2 - \frac{1}{60}s^3}$	$\frac{1 + \frac{8}{2}s + \frac{s^2}{2} + \frac{s^3}{120}}{1 - \frac{s}{2} + \frac{s^2}{10} - \frac{s^3}{120}}$

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 Padé 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

$$P_m^k(s) = \frac{Q_{mk}(s)}{Q_{km}(-s)} \quad (14.163)$$

where

$$Q_{mk}(s) = 1 + \sum_{i=1}^k \frac{k! (m+k-i)!}{(k-i)! (m+k)!} \frac{s^i}{i!} \quad (14.164)$$

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:

$$P_m^k(s) \leq 1, \quad \text{when } \operatorname{Re}[s] \leq 0 \text{ for } k \leq m \leq k+2 \quad (14.165)$$

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

$$|P_m^m(j\omega)| = 1, \quad \text{for all } \omega \quad (14.166)$$

whereas for Padé approximations where the degree of the numerator polynomial is less than the degree of the denominator polynomials we have

$$|P_m^k(j\omega)| \rightarrow 0, \quad \text{when } \omega \rightarrow \infty \text{ for } m > k \quad (14.167)$$

Combining these results we arrive at the following result:

A one-step method with stability function

$$R(s) = P_m^m(s) \quad (14.168)$$

is A stable. A one-step method with stability function

$$R(s) = P_m^k(s) \quad \text{where } m = k+1 \text{ or } m = k+2 \quad (14.169)$$

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

$$R(s) = P_m^m(s) \quad (14.170)$$

which implies that the methods are A-stable.

**Example 227** Radau methods and Lobatto IIIC methods have stability functions

$$R(s) = P_m^k(s), \quad k < m \leq k+2 \quad (14.171)$$

and this implies that the methods are L-stable.

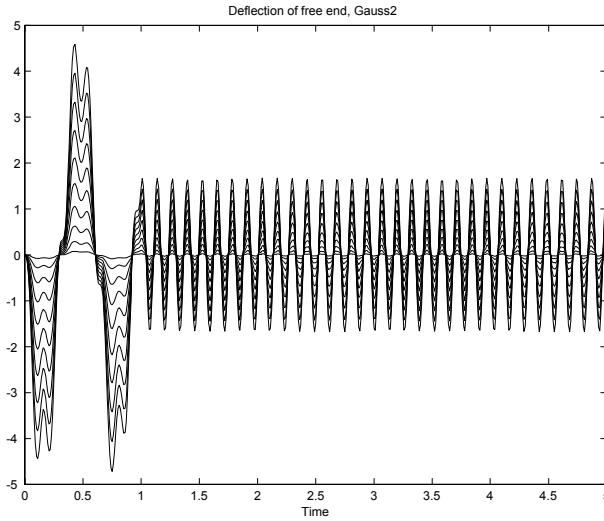


Figure 14.26: Simulation of vibrations in an Euler-Bernoulli beam with a Gauss method. There is an initial excitation that is switched off after 1 s. After this the vibrational energy of the system is a constant. The simulation reflects this.

### 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  $\omega_i$ ,  $i = 1, 2, \dots$ . 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

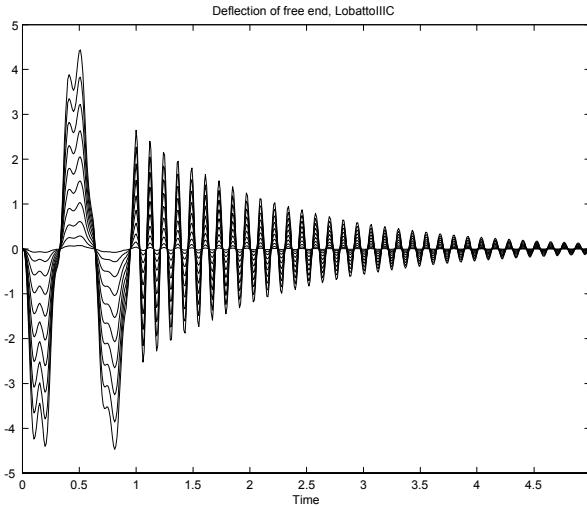


Figure 14.27: Simulation of vibrations in an Euler-Bernoulli beam with a Lobatto IIIC method. There is an initial excitation that is switched off after 1 s. After this the vibrational energy of the system is a constant. The simulation method is seen to introduce damping of the response, and the high 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

$$y_{n+1} = 0 \quad \text{when} \quad R(\lambda h) = 0 \quad (14.172)$$

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,  $|R(j\omega)| \rightarrow \infty$  when  $\omega \rightarrow \infty$ .

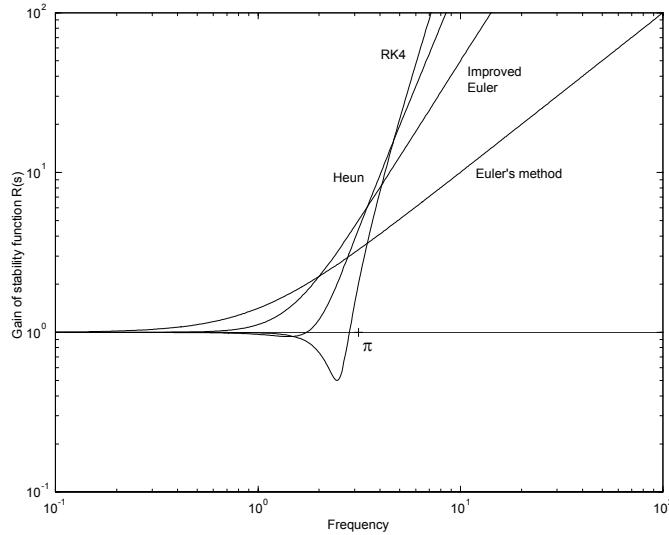


Figure 14.28: Stability function  $R(j\omega)$  of explicit Runge-Kutta methods evaluated for  $\lambda h = j\omega$ . The Nyquist frequency  $\omega_N$  is plotted at  $\omega_N h = \pi$ .

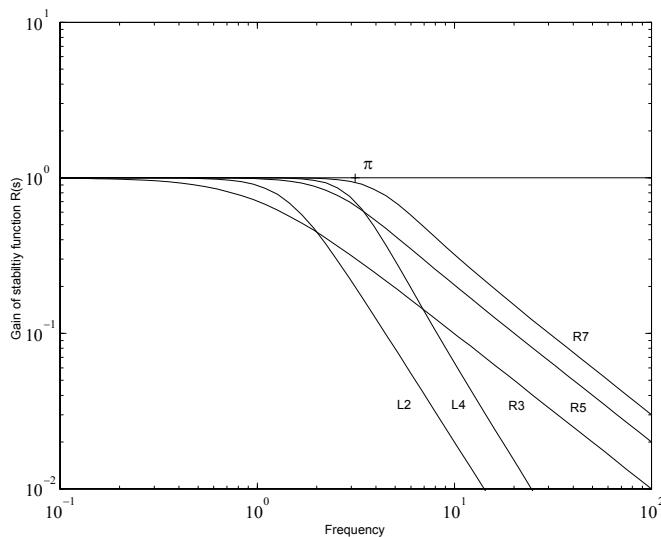


Figure 14.29: Stability function  $R(j\omega)$  of implicit Runge-Kutta methods evaluated for  $\lambda h = j\omega$ . The Nyquist frequency  $\omega_N$  is plotted at  $\omega_N h = \pi$ . The methods are seen to damp out frequency components over the Nyquist frequency. The Radau methods have a roll-off of -1, and the Lobatto IIIC methods have a roll-off of -2.

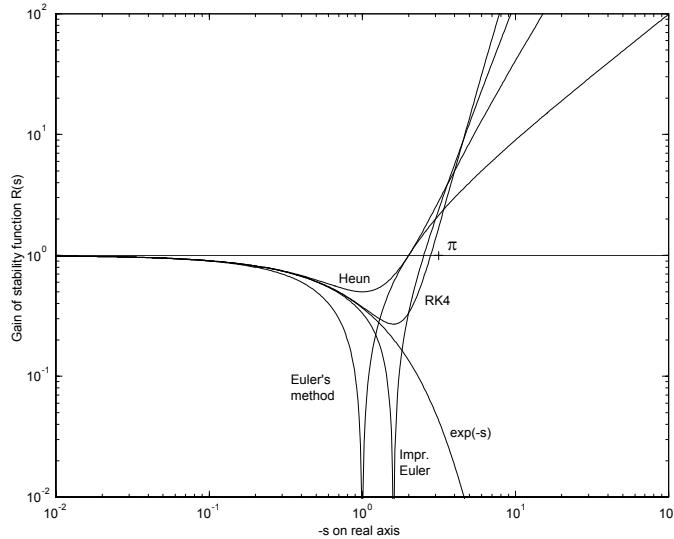


Figure 14.30: Absolute value of stability function  $|R(-s)|$  of explicit Runge-Kutta methods evaluated for  $\lambda h = -s$ . The exact value  $\exp(-s)$  is plotted for comparison.

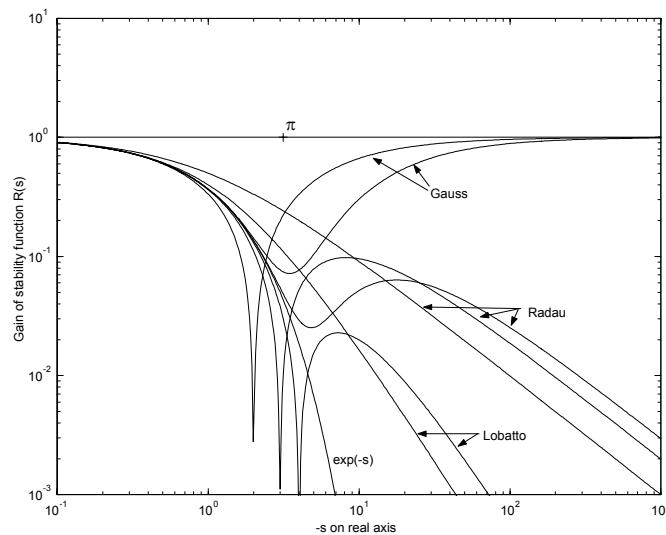


Figure 14.31: Absolute value of stability function  $|R(-s)|$  of implicit Runge-Kutta methods evaluated for  $\lambda h = -s$ . The exact value  $\exp(-s)$  is plotted for comparison. The Radau and Lobatto IIIC have a roll-off for  $-s$  large, while  $|R(-s)| \rightarrow 1$  when  $-s \rightarrow \infty$  for the Gauss methods.

A Gauss method with  $\sigma$  stages, which is an implicit Runge-Kutta method of order  $2\sigma$ , 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

$$\begin{aligned} s_1 &= -2 && \text{when } \sigma = 1 \\ s_{1,2} &= -3 \pm j1.7321 && \text{when } \sigma = 2 \\ s_{1,2} &= -3.6778 \pm j3.5088, \quad s_3 = -4.6444 && \text{when } \sigma = 3 \end{aligned}$$

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

$$R(s) = P_\sigma^{\sigma-1}(s) = \begin{cases} \frac{1}{1-s} & \text{when } \sigma = 1 \\ \frac{1+\frac{1}{3}s}{1-\frac{2}{3}s+\frac{1}{6}s^2} & \text{when } \sigma = 2 \\ \frac{1+\frac{2}{5}s+\frac{9}{20}s^2}{1-\frac{3}{5}s+\frac{3}{20}s^2-\frac{1}{60}s^3} & \text{when } \sigma = 3 \end{cases}$$

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  $\operatorname{Re}[\lambda(t)] \leq 0$  for all  $t \in [t_n, t_{n+1}]$ .

An implicit Runge-Kutta method for this system can be written

$$\boldsymbol{\kappa} = \mathbf{A}(\mathbf{1}y_n + h\mathbf{A}\boldsymbol{\kappa}) \quad (14.173)$$

$$y_{n+1} = y_n + h\mathbf{b}^T \boldsymbol{\kappa} \quad (14.174)$$

where  $\boldsymbol{\kappa} = (k_1 \dots k_\sigma)^T$  and  $\mathbf{1} = (1 \dots 1)^T$  and

$$\mathbf{A} = \operatorname{diag}(\lambda_1, \dots, \lambda_\sigma), \lambda_i = \lambda(t_n + c_i h) \quad (14.175)$$

Equation (14.173) gives  $\boldsymbol{\kappa} = (1 - h\mathbf{A}\mathbf{A})^{-1}\mathbf{A}\mathbf{1}y_n$ , and insertion into (14.174) gives

$$y_{n+1} = R_{AN}(h\mathbf{A})y_n$$

where we have defined the stability function

$$R_{AN}(h\mathbf{A}) = 1 + \mathbf{b}^T(\mathbf{I} - h\mathbf{A}\mathbf{A})^{-1}h\mathbf{A}\mathbf{1} \quad (14.176)$$

An implicit Runge-Kutta method is said to be AN-stable if  $\operatorname{Re}[\lambda_i] \leq 0$  implies that  $|R_{AN}(h\mathbf{A})| \leq 1$  and that  $(1 - h\mathbf{A}\mathbf{A})$  is nonsingular.

From this definition it is clear that

$$\boxed{\text{AN-stability}} \Rightarrow \boxed{\text{A-stability}} \quad (14.177)$$

**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

$$R_{AN}(h\mathbf{A}) = \frac{1 + \frac{h\lambda_1}{2}}{1 - \frac{h\lambda_2}{2}}$$

It is seen that if  $\lambda_2 = 0$  and  $\lambda_1$  is large, then  $|R_{AN}(h\mathbf{A})| > 1$ , and the method is not AN stable.

**Example 229** Cramer's rule can be used to find the function  $R_{AN}(h\mathbf{A})$  defined in (14.176) in the same way as the stability function  $R(h\lambda)$ . This gives

$$R_{AN}(h\mathbf{A}) = \frac{\det [\mathbf{I} - (\mathbf{A} - \mathbf{1}\mathbf{b}^T)h\mathbf{A}]}{\det (\mathbf{I} - \mathbf{A}h\mathbf{A})} \quad (14.178)$$

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  $R_{AN}(h\mathbf{A})$  is not a function of  $\lambda_\sigma$ , and the denominator of  $R_{AN}(h\mathbf{A})$  is not a function of  $\lambda_1$ . This means that if  $\lambda_2 = \dots = \lambda_\sigma = 0$ , then  $|R_{AN}(h\mathbf{A})|$  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  $R_{AN}(h\mathbf{A})$  is not a function of  $\lambda_1$ , and the denominator of  $R_{AN}(h\mathbf{A})$  is not a function of  $\lambda_\sigma$ . This means that if  $\lambda_1 = \dots = \lambda_{\sigma-1} = 0$ , then  $|R_{AN}(h\mathbf{A})|$  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

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, t) \quad (14.179)$$

and the scalar nonnegative function

$$V = \frac{1}{2} (\mathbf{y} - \tilde{\mathbf{y}})^T \mathbf{P} (\mathbf{y} - \tilde{\mathbf{y}}) \quad (14.180)$$

where  $\tilde{\mathbf{y}}$  is a solution of  $\frac{d}{dt}\tilde{\mathbf{y}} = \mathbf{f}(\tilde{\mathbf{y}}, t)$  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

$$\dot{V} = (\mathbf{y} - \tilde{\mathbf{y}})^T \mathbf{P} [\mathbf{f}(\mathbf{y}, t) - \mathbf{f}(\tilde{\mathbf{y}}, t)]$$

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

$$(\mathbf{y} - \tilde{\mathbf{y}})^T \mathbf{P} [\mathbf{f}(\mathbf{y}, t) - \mathbf{f}(\tilde{\mathbf{y}}, t)] \leq -\gamma (\mathbf{y} - \tilde{\mathbf{y}})^T \mathbf{P} (\mathbf{y} - \tilde{\mathbf{y}}), \quad \forall \mathbf{y}, \tilde{\mathbf{y}}$$

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

$$\|\mathbf{y}(t) - \tilde{\mathbf{y}}(t)\| \leq \left( \frac{\lambda_{\max}(\mathbf{P})}{\lambda_{\min}(\mathbf{P})} \right)^{\frac{1}{2}} \|\mathbf{y}(t_0) - \tilde{\mathbf{y}}(t_0)\| e^{-\gamma(t-t_0)}$$

This means that the two solutions  $\mathbf{y}(t)$  and  $\tilde{\mathbf{y}}(t)$  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

$$(\mathbf{y} - \tilde{\mathbf{y}})^T [\mathbf{f}(\mathbf{y}, t) - \mathbf{f}(\tilde{\mathbf{y}}, t)] \leq -\gamma (\mathbf{y} - \tilde{\mathbf{y}})^T (\mathbf{y} - \tilde{\mathbf{y}})$$

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  $\tilde{\mathbf{y}}_{n+1}$  starting from  $\tilde{\mathbf{y}}_n$  satisfies the condition

$$\|\mathbf{y}_{n+1} - \tilde{\mathbf{y}}_{n+1}\| \leq \|\mathbf{y}_n - \tilde{\mathbf{y}}_n\|$$

then the Runge-Kutta method is said to be B-stable.

Consider the linear time-varying test system

$$\dot{y} = \lambda(t)y \quad (14.181)$$

Then, for two solutions  $y(t)$  and  $\tilde{y}(t)$  we have

$$V = \frac{1}{2}(y - \tilde{y})^2 \Rightarrow \dot{V} = (y - \tilde{y})\lambda(t)(y - \tilde{y}) \quad (14.182)$$

It follows that if  $\operatorname{Re} \lambda(t) \leq 0$  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

B-stability	$\Rightarrow$	AN-stability	$\Rightarrow$	A-stability
-------------	---------------	--------------	---------------	-------------

(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  $i = 1, \dots, \sigma$  and

$$\mathbf{M} = \operatorname{diag}(\mathbf{b}) \mathbf{A} + \mathbf{A}^T \operatorname{diag}(\mathbf{b}) - \mathbf{b} \mathbf{b}^T \geq 0$$

We note that the elements of  $\mathbf{M}$  are given by

$$m_{ij} = b_i a_{ij} + b_j a_{ji} - b_i b_j \quad (14.184)$$

An algebraically stable Runge-Kutta method is B-stable, that is,

Algebraic stability	$\Rightarrow$	B-stability
---------------------	---------------	-------------

(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

$$\mathbf{Y}_i = \mathbf{y}_n + h \sum_{j=1}^{\sigma} a_{ij} \mathbf{f}(\mathbf{Y}_j, t_n + c_j h), \quad i = 1, \dots, \sigma \quad (14.186)$$

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h \sum_{i=1}^{\sigma} b_i \mathbf{f}(\mathbf{Y}_i, t_n + c_i h) \quad (14.187)$$

Then we denote the differences between the two solutions  $\mathbf{y}_n$  and  $\tilde{\mathbf{y}}_n$  by

$$\begin{aligned}\Delta \mathbf{y}_n &= \mathbf{y}_n - \tilde{\mathbf{y}}_n, \Delta \mathbf{y}_{n+1} = \mathbf{y}_{n+1} - \tilde{\mathbf{y}}_{n+1}, \Delta \mathbf{Y}_i = \mathbf{Y}_i - \tilde{\mathbf{Y}}_i \\ \Delta \mathbf{f}_i &= h \left[ \mathbf{f}(\mathbf{Y}_i, t_n + c_i h) - \mathbf{f}(\tilde{\mathbf{Y}}_i, t_n + c_i h) \right]\end{aligned}$$

where  $\mathbf{Y}_i$  is a vector corresponding to the vector  $\mathbf{y}_n$ , and  $\tilde{\mathbf{Y}}_i$  is a vector corresponding to the vector  $\tilde{\mathbf{y}}_n$ . Subtraction of the Runge-Kutta equations for the solution  $\tilde{\mathbf{y}}_n$  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

$$\begin{aligned}(\Delta \mathbf{y}_{n+1})^T \Delta \mathbf{y}_{n+1} &= (\Delta \mathbf{y}_n)^T \Delta \mathbf{y}_n + 2 \sum_{i=1}^{\sigma} b_i (\Delta \mathbf{f}_i)^T \Delta \mathbf{y}_n \\ &\quad + \sum_{i=1}^{\sigma} \sum_{j=1}^{\sigma} b_i b_j (\Delta \mathbf{f}_i)^T \Delta \mathbf{f}_j \\ &= (\Delta \mathbf{y}_n)^T \Delta \mathbf{y}_n + 2 \sum_{i=1}^{\sigma} b_i (\Delta \mathbf{f}_i)^T \left( \Delta \mathbf{Y}_i - \sum_{j=1}^{\sigma} a_{ij} \Delta \mathbf{f}_j \right) \\ &\quad + \sum_{i=1}^{\sigma} \sum_{j=1}^{\sigma} b_i b_j (\Delta \mathbf{f}_i)^T \Delta \mathbf{f}_j \\ &= (\Delta \mathbf{y}_n)^T \Delta \mathbf{y}_n + 2 \sum_{i=1}^{\sigma} b_i (\Delta \mathbf{f}_i)^T \Delta \mathbf{Y}_i \\ &\quad - \sum_{i=1}^{\sigma} \sum_{j=1}^{\sigma} m_{ij} (\Delta \mathbf{f}_i)^T \Delta \mathbf{f}_j \tag{14.188}\end{aligned}$$

where  $m_{ij}$  is element  $(i, j)$  of matrix  $\mathbf{M}$ . As

$$(\Delta \mathbf{f}_i)^T \Delta \mathbf{Y}_i = h \left[ \mathbf{f}(\mathbf{Y}_i, t_n + c_i h) - \mathbf{f}(\tilde{\mathbf{Y}}_i, t_n + c_i h) \right]^T (\mathbf{Y}_i - \tilde{\mathbf{Y}}_i) \leq 0$$

by assumption, and as  $\sum_{i=1}^{\sigma} \sum_{j=1}^{\sigma} m_{ij} (\Delta \mathbf{f}_i)^T \Delta \mathbf{f}_j \geq 0$  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

$$[\mathbf{f}(\mathbf{y}, t) - \mathbf{f}(\tilde{\mathbf{y}}, t)]^T \mathbf{P} (\mathbf{y} - \tilde{\mathbf{y}}) = 0 \tag{14.189}$$

then it follows from the derivation above that a Runge-Kutta method that satisfies

$$\mathbf{M} = \text{diag}(\mathbf{b}) \mathbf{A} + \mathbf{A}^T \text{diag}(\mathbf{b}) - \mathbf{b} \mathbf{b}^T = \mathbf{0}$$

will give

$$(\Delta \mathbf{y}_{n+1})^T \mathbf{P} \Delta \mathbf{y}_{n+1} = (\Delta \mathbf{y}_n)^T \mathbf{P} \Delta \mathbf{y}_n \tag{14.190}$$

Method	Order	Stability function	Linear stability	Nonlinear stability	Stiffly Accurate
Explicit, $p = \sigma$	$\sigma$	$P_0^\sigma$	$ h\lambda $ small	-	No
Gauss	$2\sigma$	$P_\sigma^\sigma$	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  $\hat{\mathbf{y}}_{n+1}$  of the same system is computed with a method

$$\begin{array}{c|c} \hat{\mathbf{c}} & \hat{\mathbf{A}} \\ \hline & \hat{\mathbf{b}}^T \end{array}$$

of order  $\hat{p} = p + 1$ . The computation of the value at  $t_{n+1}$  starts with  $\mathbf{y}_n = \hat{\mathbf{y}}_n$ . Then the

local solution  $\mathbf{y}_L(t_n; t_{n+1})$  satisfies

$$\mathbf{y}_L(t_n; t_{n+1}) = \mathbf{y}_{n+1} + \mathbf{e}_{n+1} = \hat{\mathbf{y}}_{n+1} + \hat{\mathbf{e}}_{n+1}$$

where  $\mathbf{e}_{n+1} = O(h^{p+1})$  is the local error in the computation of  $\mathbf{y}_{n+1}$ , and  $\hat{\mathbf{e}}_{n+1} = O(h^{p+2})$  is the local error in the computation of  $\hat{\mathbf{y}}_{n+1}$ . Because  $\hat{\mathbf{e}}_{n+1}$  is of higher order in  $h$  than  $\mathbf{e}_{n+1}$ , we can find an estimate of  $\mathbf{e}_{n+1}$  from

$$\hat{\mathbf{y}}_{n+1} - \mathbf{y}_{n+1} = \mathbf{e}_{n+1} - \hat{\mathbf{e}}_{n+1} \approx \mathbf{e}_{n+1}$$

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  $\hat{\mathbf{y}}_{n+1}$  is more accurate, so it makes more sense to use  $\hat{\mathbf{y}}_{n+1}$  as a starting point for the next time step. The use of  $\hat{\mathbf{y}}_{n+1}$  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  $\hat{\mathbf{y}}_{n+1}$  denotes the embedded low order solution.

To make the computations efficient, the two methods are usually designed so that  $\mathbf{c} = \hat{\mathbf{c}}$  and  $\mathbf{A} = \hat{\mathbf{A}}$ . Then the stage computations will be the same in both methods, and need only be done once. The solution  $\hat{\mathbf{y}}$  is said to be an *embedded solution* in this case. The algorithm is

$$\mathbf{k}_1 = \mathbf{f}(\mathbf{y}_n, t_n) \quad (14.191)$$

$$\vdots \quad (14.192)$$

$$\mathbf{k}_\sigma = \mathbf{f}(\mathbf{y}_n + h(a_{\sigma 1}\mathbf{k}_1 + \dots + a_{\sigma, \sigma-1}\mathbf{k}_{\sigma-1}), t_n + c_\sigma h) \quad (14.193)$$

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h(b_1\mathbf{k}_1 + \dots + b_\sigma\mathbf{k}_\sigma) \quad (14.194)$$

$$\hat{\mathbf{y}}_{n+1} = \mathbf{y}_n + h(\hat{b}_1\mathbf{k}_1 + \dots + \hat{b}_\sigma\mathbf{k}_\sigma) \quad (14.195)$$

$$\mathbf{e}_{n+1} = \hat{\mathbf{y}}_{n+1} - \mathbf{y}_{n+1} \quad (14.196)$$

The computation of  $\mathbf{y}_{n+1}$  and  $\hat{\mathbf{y}}_{n+1}$  is described by an array as follows:

$\mathbf{c}$	$\mathbf{A}$
$\mathbf{y}$	$\mathbf{b}^T$
$\hat{\mathbf{y}}$	$\hat{\mathbf{b}}^T$
$\mathbf{e}$	$\mathbf{E}^T$

where  $\mathbf{E} = \hat{\mathbf{b}} - \mathbf{b}$ .

Runge-Kutta-Fehlberg 4(5) is a method where  $\mathbf{y}$  is computed with order  $p = 4$  using five stages. The embedded solution  $\hat{\mathbf{y}}$  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	$\frac{1408}{2565}$	$\frac{2197}{4104}$	$-\frac{1}{5}$	
$\hat{y}$	$\frac{16}{135}$	0	$\frac{6656}{12825}$	$\frac{28561}{56430}$	$-\frac{9}{50}$	$\frac{2}{55}$
$\Delta e$	$\frac{1}{360}$	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  $\hat{\mathbf{y}}$  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	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$
$y$	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$
$\hat{y}$	$\frac{5179}{57600}$	0	$\frac{7571}{16695}$	$\frac{393}{640}$	$-\frac{92097}{339200}$	$\frac{187}{2100}$
$\Delta e$	$\frac{71}{57600}$	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	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{4}{9}$		
$y$	$\frac{21}{72}$	$\frac{1}{4}$	$\frac{3}{9}$	$\frac{1}{8}$		
$\hat{y}$	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{4}{9}$			
$\Delta e$	$-\frac{5}{72}$	$\frac{1}{12}$	$\frac{1}{9}$	$-\frac{1}{8}$		

### 14.7.2 Adjustment algorithm

Suppose that the specified accuracy is specified in terms of a tolerance  $e_{\text{tol}}$  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  $h_{\text{new}}$  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

$$h_{\text{new}} = h \left( \frac{e_{\text{tol}}}{\varepsilon_{n+1}} \right)^{\frac{1}{p+1}} \quad (14.197)$$

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:

$$\ln h_{\text{new}} = \ln h - \frac{1}{p+1} (\ln \varepsilon_{n+1} - \ln e_{\text{tol}}) \quad (14.198)$$

This can be compared with an incremental form of a PI controller

$$u_{n+1} = u_n - K_p (e_n - e_{n-1}) - K_p \frac{h}{T_i} e_n \quad (14.199)$$

One may compare the adjustment formula with an I controller. Proportional action is included using

$$\ln h_{\text{new}} = \ln h - K_p (\ln \varepsilon_{n+1} - \ln \varepsilon_n) - K_p \frac{h}{T_i} (\ln \varepsilon_{n+1} - \ln e_{\text{tol}}) \quad (14.200)$$

which gives the adjustment formula

$$h_{\text{new}} = h \left( \frac{e_{\text{tol}}}{\varepsilon_{n+1}} \right)^{K_p \frac{h}{T_i}} \left( \frac{\varepsilon_n}{\varepsilon_{n+1}} \right)^{K_p} \quad (14.201)$$

which is simplified to

$$h_{\text{new}} = h \left( \frac{e_{\text{tol}}}{\varepsilon_{n+1}} \right)^{K_p \left( 1 + \frac{h}{T_i} \right)} \left( \frac{\varepsilon_n}{e_{\text{tol}}} \right)^{K_p} \quad (14.202)$$

The following parameters have been suggested.

$$K_p = 0.4/(p+1), \quad T_i = 1.3h \quad (14.203)$$

## 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

$$\begin{aligned}\mathbf{z}_1 &= h [a_{11}\mathbf{f}(\mathbf{y}_n + \mathbf{z}_1, t_n + c_1 h) + \dots + a_{1\sigma}\mathbf{f}(\mathbf{y}_n + \mathbf{z}_\sigma, t_n + c_\sigma h)] \\ &\vdots \\ \mathbf{z}_\sigma &= h [a_{\sigma 1}\mathbf{f}(\mathbf{y}_n + \mathbf{z}_1, t_n + c_1 h) + \dots + a_{\sigma\sigma}\mathbf{f}(\mathbf{y}_n + \mathbf{z}_\sigma, t_n + c_\sigma h)]\end{aligned}$$

and to compute the solution  $\mathbf{y}_{n+1}$

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h [b_1\mathbf{f}(\mathbf{y}_n + \mathbf{z}_1, t_n + c_1 h) + \dots + b_\sigma\mathbf{f}(\mathbf{y}_n + \mathbf{z}_\sigma, t_n + c_\sigma h)]$$

**Example 232** If  $\mathbf{A}$  is nonsingular, the update can be found from

$$\mathbf{y}_{n+1} = \mathbf{y}_n + d_1\mathbf{z}_1 + \dots + d_\sigma\mathbf{z}_\sigma \quad (14.204)$$

where

$$(d_1, \dots, d_\sigma) = (b_1, \dots, b_\sigma) \mathbf{A}^{-1} \quad (14.205)$$

In particular, if  $a_{\sigma i} = b_i$  then

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \mathbf{z}_\sigma \quad (14.206)$$

To solve for  $\mathbf{z}_1, \dots, \mathbf{z}_\sigma$ , a Newton search method is used. The equation is written in vector form as

$$\mathbf{Z} = h (\mathbf{A} \otimes \mathbf{I}_\sigma) \mathbf{F}(\mathbf{Z})$$

where

$$\mathbf{Z} = \begin{pmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_\sigma \end{pmatrix}, \quad \mathbf{F}(\mathbf{Z}) = \begin{pmatrix} \mathbf{f}(\mathbf{y}_n + \mathbf{z}_1, t_n + c_1 h) \\ \vdots \\ \mathbf{f}(\mathbf{y}_n + \mathbf{z}_\sigma, t_n + c_\sigma h) \end{pmatrix}$$

are vectors of dimension  $d\sigma$ ,  $\mathbf{I}_\sigma$  is the  $\sigma \times \sigma$  identity matrix and

$$\mathbf{A} \otimes \mathbf{I}_\sigma = \begin{pmatrix} a_{11}\mathbf{I}_\sigma & \dots & a_{\sigma 1}\mathbf{I}_\sigma \\ \vdots & \ddots & \vdots \\ a_{1\sigma}\mathbf{I}_\sigma & \dots & a_{\sigma\sigma}\mathbf{I}_\sigma \end{pmatrix}$$

is the Kronecker tensor product of  $\mathbf{A}$  and  $\mathbf{I}_\sigma$ .

The solution is found by minimizing the function

$$L = [\mathbf{Z} - h (\mathbf{A} \otimes \mathbf{I}_\sigma) \mathbf{F}(\mathbf{Z})]^T [\mathbf{Z} - h (\mathbf{A} \otimes \mathbf{I}_\sigma) \mathbf{F}(\mathbf{Z})] \quad (14.207)$$

with respect to  $\mathbf{Z}$  using a Newton search, which is done by the iteration

$$\mathbf{H}(\mathbf{Z}^{i+1} - \mathbf{Z}^i) = -\mathbf{Z}^i + h (\mathbf{A} \otimes \mathbf{I}_\sigma) \mathbf{F}(\mathbf{Z}^i) \quad (14.208)$$

which is solved for  $\mathbf{Z}^{i+1}$ . Here  $\mathbf{Z}^i$  is iteration  $i$  of  $\mathbf{Z}$ , and

$$\mathbf{H} = \mathbf{I} - h (\mathbf{A} \otimes \mathbf{J}) = \begin{pmatrix} 1 - ha_{11}\mathbf{J} & \dots & -ha_{\sigma 1}\mathbf{J} \\ \vdots & \ddots & \vdots \\ -ha_{1\sigma}\mathbf{J} & \dots & 1 - ha_{\sigma\sigma}\mathbf{J} \end{pmatrix}$$

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  $\dots \mathbf{y}_{n-1}, \mathbf{y}_n, \mathbf{y}_{n+1} \dots$  at discrete time instants  $\dots t_{n-1}, t_n, t_{n+1} \dots$  for the system

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, t), \quad \mathbf{y}(t_0) = \mathbf{y}_0 \quad (14.209)$$

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

$$\mathbf{y}_n(\alpha), \quad \alpha \in [0, 1], \quad \mathbf{y}_n(0) = \mathbf{y}_n, \quad \mathbf{y}_n(1) = \mathbf{y}_{n+1} \quad (14.210)$$

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

$$\mathbf{k}_i = \mathbf{f}\left(\mathbf{y}_n + h \sum_{j=1}^{\sigma^*} a_{ij} \mathbf{k}_j, t_n + c_i h\right), \quad i = 1, \dots, \sigma^* \quad (14.211)$$

$$\mathbf{y}_n(\alpha) = \mathbf{y}_n + h \sum_{j=1}^{\sigma^*} b_j(\alpha) \mathbf{k}_j \quad (14.212)$$

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.

$$\begin{aligned} b_1(\alpha) &= \alpha^2(3 - 2\alpha)b_1 + \alpha(\alpha - 1)^2 \\ &\quad - \alpha^2(\alpha - 1)^2 \frac{5(2, 558, 722, 523 - 31, 403, 016\alpha)}{11, 282, 082, 432} \end{aligned} \quad (14.213)$$

$$b_2(\alpha) = 0 \quad (14.214)$$

$$b_3(\alpha) = \alpha^2(3 - 2\alpha)b_3 + \alpha^2(\alpha - 1)^2 \frac{100(882, 725, 551 - 15, 701, 508\alpha)}{32, 700, 410, 799} \quad (14.215)$$

$$b_4(\alpha) = \alpha^2(3 - 2\alpha)b_4 - \alpha^2(\alpha - 1)^2 \frac{25(443332067 - 31, 403, 016\alpha)}{1, 880, 347, 072} \quad (14.216)$$

$$b_5(\alpha) = \alpha^2(3 - 2\alpha)b_5 + \alpha^2(\alpha - 1)^2 \frac{32805(23, 143, 187 - 3, 489, 224\alpha)}{199, 316, 789, 632} \quad (14.217)$$

$$b_6(\alpha) = \alpha^2(3 - 2\alpha)b_6 - \alpha^2(\alpha - 1)^2 \frac{55(29, 972, 135 - 7, 076, 736\alpha)}{822, 651, 844} \quad (14.218)$$

$$b_7(\alpha) = \alpha^2(\alpha - 1) + \alpha^2(\alpha - 1)^2 \frac{10(7, 414, 447 - 829, 305\alpha)}{29, 380, 432} \quad (14.219)$$

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  $\alpha$ . 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

$$\mathbf{M}\ddot{\mathbf{u}} = \phi(\mathbf{u}) \quad (14.222)$$

where  $\mathbf{M}$  is a nonsingular matrix. Runge-Kutta methods can be implemented with the stage computations

$$\begin{aligned}\mathbf{k}_i &= \phi(\mathbf{u}_n + \sum_{j=1}^{i-1} a_{ij} \mathbf{k}_j, t_n + c_i h), \quad i = 1, \dots, \sigma \\ \mathbf{u}_{n+1} &= \mathbf{u}_n + \mathbf{M}^{-1} \left( h \sum_{j=1}^{\sigma} b_j \mathbf{k}_j \right)\end{aligned}$$

**Example 233** One example of this is in robotics where the equation of motion is of the form

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

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

$$\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}(\mathbf{q}) \end{pmatrix} \dot{\mathbf{u}} = \phi(\mathbf{u})$$

where

$$\mathbf{u} = \begin{pmatrix} \mathbf{q} \\ \dot{\mathbf{q}} \end{pmatrix}, \quad \phi(\mathbf{u}) = \begin{pmatrix} \dot{\mathbf{q}} \\ -\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}}) \dot{\mathbf{q}} - \mathbf{g}(\mathbf{q}) + \boldsymbol{\tau} \end{pmatrix}$$

The system could have been written in the form that has been used so far, that is,

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) = \begin{pmatrix} \dot{\mathbf{q}} \\ \mathbf{M}(\mathbf{q})^{-1} [-\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}}) \dot{\mathbf{q}} - \mathbf{g}(\mathbf{q}) + \boldsymbol{\tau}] \end{pmatrix}$$

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

$$L(\mathbf{y}) = \mathbf{w}^T \mathbf{y} \tag{14.223}$$

where  $\mathbf{w} = (w_1 \dots w_d)^T$  is a vector of constants, so that for all  $\mathbf{y}$  the time derivative along solutions of the system  $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, t)$  is zero, that is

$$\dot{L}(\mathbf{y}) := \mathbf{w}^T \dot{\mathbf{y}} = \mathbf{w}^T \mathbf{f}(\mathbf{y}, t) = 0 \quad \text{for all } \mathbf{y} \tag{14.224}$$

Such a function is called a *linear invariant*. Then it follows from

$$\mathbf{k}_i = \mathbf{f}(\mathbf{y}_n + \sum_{j=1}^{i-1} a_{ij} \mathbf{k}_j, t_n + c_i h) \quad (14.225)$$

that

$$\mathbf{w}^T \mathbf{k}_i = 0 \quad (14.226)$$

and

$$\mathbf{w}^T \mathbf{y}_{n+1} = \mathbf{w}^T \mathbf{y}_n + h \sum_{j=1}^{\sigma} b_j \mathbf{w}^T \mathbf{k}_j = \mathbf{w}^T \mathbf{y}_n \quad (14.227)$$

We see that linear invariants will be conserved when the solution is computed with any Runge-Kutta method.

**Example 234** Consider a chemical reaction  $A + B \rightarrow C$  in a closed tank. The total mass  $m$  of the chemical components will be constant due to the principle of conservation of mass. This is written

$$m = m_A + m_B + m_C = \text{const.} \quad (14.228)$$

where  $m_A$ ,  $m_B$  and  $m_C$  are the masses of each of the components  $A$ ,  $B$  and  $C$ . This means that the total mass  $m$  is a linear invariant of the system. The mass balance is assumed to be

$$\frac{d}{dt} m_A = -\nu_{CA}(m_A, m_B, m_C) \quad (14.229)$$

$$\frac{d}{dt} m_B = -\nu_{CB}(m_A, m_B, m_C) \quad (14.230)$$

$$\frac{d}{dt} m_C = \nu_{CA}(m_A, m_B, m_C) + \nu_{CB}(m_A, m_B, m_C) \quad (14.231)$$

where  $\nu_{CA}$  is the rate of mass transfer from  $A$  to  $C$ , and  $\nu_{CB}$  is the rate of mass transfer from  $B$  to  $C$ . Then, if the numerical solution to the mass balances (14.229–14.231) is computed with a Runge-Kutta method, the mass  $m$  will be conserved in the numerical solution.

### 14.9.3 Quadratic functions

Consider a system

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, t) \quad (14.232)$$

and a quadratic function

$$V(\mathbf{y}) = \frac{1}{2} \mathbf{y}^T \mathbf{P} \mathbf{y} \quad (14.233)$$

The time derivative of  $V$  along solutions of the system is given by

$$\dot{V}(\mathbf{y}) := \frac{\partial V(\mathbf{y})}{\partial \mathbf{y}} \dot{\mathbf{y}} = \mathbf{y}^T \mathbf{P} \mathbf{f}(\mathbf{y}, t) \quad (14.234)$$

We then have the following result (Hairer 1999):

If the system (14.232) is integrated with a Runge-Kutta method satisfying

$$m_{ij} = b_i a_{ij} + b_j a_{ji} - b_i b_j = 0 \quad (14.235)$$

then for the quadratic function  $V(\mathbf{y}) = (1/2)\mathbf{y}^T \mathbf{P} \mathbf{y}$  with time derivative  $\dot{V}(\mathbf{y})$  along the solutions of the system, the following result apply:

$$V(\mathbf{y}_{n+1}) = V(\mathbf{y}_n) + 2 \sum_{i=1}^{\sigma} b_i \dot{V}(\mathbf{Y}_i) \quad (14.236)$$

where  $\mathbf{Y}_i$  is the function value for  $\mathbf{y}$  corresponding to stage  $i$  as defined in (14.186). This means that if  $b_i \geq 0$ , then

$$\dot{V}(\mathbf{y}) > 0, \forall \mathbf{y} \implies V(\mathbf{y}_{n+1}) > V(\mathbf{y}_n) \quad (14.237)$$

$$\dot{V}(\mathbf{y}) = 0, \forall \mathbf{y} \implies V(\mathbf{y}_{n+1}) = V(\mathbf{y}_n) \quad (14.238)$$

$$\dot{V}(\mathbf{y}) < 0, \forall \mathbf{y} \implies V(\mathbf{y}_{n+1}) < V(\mathbf{y}_n) \quad (14.239)$$

This result follows from the calculation

$$\begin{aligned} \mathbf{y}_{n+1}^T \mathbf{P} \mathbf{y}_{n+1} &= \left( \mathbf{y}_n + h \sum_{i=1}^{\sigma} b_i \mathbf{f}(\mathbf{Y}_i) \right)^T \mathbf{P} \left( \mathbf{y}_n + h \sum_{j=1}^{\sigma} b_j \mathbf{f}(\mathbf{Y}_j) \right) \\ &= \mathbf{y}_n^T \mathbf{P} \mathbf{y}_n + 2h \sum_{i=1}^{\sigma} b_i \mathbf{Y}_i^T \mathbf{P} \mathbf{f}(\mathbf{Y}_i) \end{aligned} \quad (14.240)$$

$$-h^2 \sum_{i=1}^{\sigma} \sum_{j=1}^{\sigma} m_{ij} \mathbf{f}^T(\mathbf{Y}_i) \mathbf{P} \mathbf{f}(\mathbf{Y}_j) \quad (14.241)$$

where equations (14.186) and (14.187) are used, and the calculation is done along the lines of (14.188). The computed solution  $\mathbf{y}_{n+1}$  will then result in an increasing  $V(\mathbf{y}_{n+1})$  if  $V[\mathbf{y}(t)]$  is increasing for the exact solution  $\mathbf{y}(t)$ , it will result in an invariant  $V(\mathbf{y}_{n+1})$  if  $V[\mathbf{y}(t)]$  is an invariant, and it will give a decreasing  $V(\mathbf{y}_{n+1})$  if  $V[\mathbf{y}(t)]$  is a decreasing function for the exact solution.

#### 14.9.4 Quadratic invariants

In this section we will look closer at the case where the quadratic function  $V$  defined in (14.233) is invariant, which is the case if  $\dot{V}(\mathbf{y}) = 0$  for all  $\mathbf{y}$ . A numerical solution  $\mathbf{y}_n$  is found from a Runge-Kutta method that will be characterized by the matrix  $\mathbf{M} = \{m_{ij}\}$  where  $m_{ij}$  is defined in (14.235). The topic that is addressed in this section is to find conditions for the numerically computed invariant  $V(\mathbf{y}_n)$  to increase, decrease or stay invariant.

Suppose that the matrix  $\mathbf{P}$  is positive definite and symmetric. Then it follows that there is a matrix  $\mathbf{Q}$  so that  $\mathbf{P} = \mathbf{Q}^T \mathbf{Q}$ . Define  $\mathbf{g}_i = \mathbf{Q} \mathbf{f}_i(\mathbf{Y}_i)$ . Then, with  $\mathbf{M} = \{m_{ij}\}$  we have

$$\beta = \sum_{i=1}^{\sigma} \sum_{j=1}^{\sigma} m_{ij} \mathbf{g}_j^T \mathbf{g}_i = \sum_{k=1}^d \mathbf{v}_k^T \mathbf{M} \mathbf{v}_k \quad (14.242)$$

Method	eigenvalues of $\mathbf{M}$
Implicit Euler	1
Gauss order 2 (Implicit midpoint rule)	0
Gauss methods of any order	$0, \dots, 0$
Radau IA order 3	0, 0.1250
Radau IIA order 3	0, 0.5
Lobatto IIIA order 2 (Trapezoidal rule)	-0.25, 0.25
Lobatto IIIB order 2	-0.25, 0.25
Lobatto IIIC order 2	0, 0.5

Table 14.5: Eigenvalues of  $\mathbf{M}$  for some implicit Runge-Kutta methods

where the vector  $\mathbf{v}_k$  is defined by

$$\mathbf{v}_k := \begin{pmatrix} (\mathbf{g}_1)_k \\ \vdots \\ (\mathbf{g}_\sigma)_k \end{pmatrix} \quad (14.243)$$

and  $(\mathbf{g}_i)_k$  is element  $k$  of the vector  $\mathbf{g}_i$ . As it is assumed that  $\dot{V}(\mathbf{y}) = 0, \forall \mathbf{y}$ , it follows from (14.241) that

$$\mathbf{y}_{n+1}^T \mathbf{P} \mathbf{y}_{n+1} = \mathbf{y}_n^T \mathbf{P} \mathbf{y}_n - h^2 \sum_{k=1}^d \mathbf{v}_k^T \mathbf{M} \mathbf{v}_k$$

We recall that a quadratic form can be expressed in term of the eigenvalues of the matrix according to

$$\mathbf{v}_k^T \mathbf{M} \mathbf{v}_k = \sum_{i=1}^{\sigma} \lambda_i(\mathbf{M}) \alpha_k^2 \quad (14.244)$$

where the scalars  $\alpha_k^2$  depend on  $\mathbf{M}$  and  $\mathbf{v}_k$ . Therefore the quadratic form is positive if all the eigenvalues  $\lambda_i(\mathbf{M})$  of  $\mathbf{M}$  are positive, the quadratic form is zero if all the eigenvalues are zero, and the quadratic form is negative if all the eigenvalues are negative. We conclude that

Let  $V$  be the quadratic function  $V = (1/2)\mathbf{y}^T \mathbf{P} \mathbf{y}$  with symmetric and positive  $\mathbf{P}$ , and let  $\lambda_i(\mathbf{M})$  be an eigenvalue of  $\mathbf{M} = \{m_{ij}\}$ , where  $m_{ij}$  is defined in (14.235). Then

$$\dot{V}(\mathbf{y}) = 0, \forall \mathbf{y} \text{ and } \lambda_i(\mathbf{M}) < 0, \forall i \implies V(\mathbf{y}_{n+1}) \geq V(\mathbf{y}_n) \quad (14.245)$$

$$\dot{V}(\mathbf{y}) = 0, \forall \mathbf{y} \text{ and } \lambda_i(\mathbf{M}) = 0, \forall i \implies V(\mathbf{y}_{n+1}) = V(\mathbf{y}_n) \quad (14.246)$$

$$\dot{V}(\mathbf{y}) = 0, \forall \mathbf{y} \text{ and } \lambda_i(\mathbf{M}) > 0, \forall i \implies V(\mathbf{y}_{n+1}) \leq V(\mathbf{y}_n) \quad (14.247)$$

Table 14.5 shows the eigenvalues of  $\mathbf{M}$  for some implicit Runge-Kutta methods, while Table 14.6. shows the eigenvalues for some explicit methods. The eigenvalues were computed using the Symbolic Math Toolbox in MATLAB using a script like the one shown below which computes the eigenvalues for the Lobatto IIIC method.

```
syms A b M;
A=[1 -1; 1 1]/2;
b=[1 1]'/2;
```

Method	eigenvalues of $\mathbf{M}$
Euler's method	-1
Improved Euler	-0.5, 0
Modified Euler	-1.2, 0.2
Heun's method of order 3	-0.89, -0.06, 0.3
RK4	-0.27, -0.1, 0, 0.1

Table 14.6: Eigenvalues of  $\mathbf{M}$  for some explicit Runge-Kutta methods.

```
M=diag(b)*A+A'*diag(b)-b*b'
eig(M)
```

**Example 235** In the description of rigid body motion the rotation can be described using Euler parameters

$$\mathbf{y} = \begin{pmatrix} \eta \\ \epsilon \end{pmatrix} \quad (14.248)$$

where  $\eta = \cos \frac{\theta}{2}$ ,  $\epsilon = \mathbf{k} \sin \frac{\theta}{2}$ , and  $\mathbf{k}$  is a unit vector has the quadratic invariant

$$\mathbf{y}^T \mathbf{y} = 1 \quad (14.249)$$

These parameters are used e.g. in strap-down inertial navigation systems. The Euler parameters satisfy the differential equation

$$\begin{pmatrix} \dot{\eta} \\ \dot{\epsilon} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \omega^T \epsilon \\ \eta \omega + \mathbf{S}(\epsilon) \omega \end{pmatrix} \quad (14.250)$$

where  $\mathbf{S}(\epsilon)$  is the skew-symmetric form of  $\epsilon$  and  $\omega$  is the angular velocity vector of the rigid body. The invariant  $\mathbf{y}^T \mathbf{y} = 1$  will hold when the system is integrated with a Runge-Kutta method with  $m_{ij} = b_i a_{ij} + b_j a_{ji} - b_i b_j = 0$ . In applications it is usual to integrate with an explicit method and a projection

$$\mathbf{y} := \frac{\mathbf{y}}{|\mathbf{y}|} \quad (14.251)$$

#### 14.9.5 Symplectic Runge-Kutta methods

We consider a system with the Hamiltonian  $H = H(\mathbf{p}, \mathbf{q}, t)$  where

$$\mathbf{p} = \begin{pmatrix} p_1 \\ \vdots \\ p_d \end{pmatrix}, \quad \mathbf{q} = \begin{pmatrix} q_1 \\ \vdots \\ q_d \end{pmatrix} \quad (14.252)$$

are in the phase space  $\Omega \subset R^{2d}$ , that is  $(\mathbf{p}, \mathbf{q}) \in \Omega \subset R^{2d}$ . The Hamiltonian system of differential equations with Hamiltonian  $H$  is given by

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}, \quad \frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}, \quad i = 1, \dots, d$$

We refer to this system as  $\Sigma_H$ .

We note that  $H = H(\mathbf{p}, \mathbf{q})$ , which means that the Hamiltonian is not a function of time, then the Hamiltonian is an invariant of the Hamiltonian system, which is seen from

$$\dot{H} = \frac{\partial H}{\partial p_i} \frac{dp_i}{dt} + \frac{\partial H}{\partial q_i} \frac{dq_i}{dt} = 0 \quad (14.253)$$

We define

$$\mathbf{y} = \begin{pmatrix} \mathbf{p} \\ \mathbf{q} \end{pmatrix}$$

Then the Hamiltonian system can be written

$$\frac{d}{dt} \mathbf{y} = \mathbf{J}^{-1} \nabla H$$

where

$$\mathbf{J} = \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{0} \end{pmatrix}, \quad \mathbf{J}^{-1} = \mathbf{J}^T = \begin{pmatrix} \mathbf{0} & -\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix}$$

is a skew-symmetric matrix and

$$\nabla = \begin{pmatrix} \frac{\partial}{\partial \mathbf{p}} \\ \frac{\partial}{\partial \mathbf{q}} \end{pmatrix}$$

is the gradient operator.

We note that the Hamiltonian system has zero divergence in the sense that

$$\operatorname{div} \dot{\mathbf{y}} = \nabla^T \mathbf{J}^{-1} \nabla H = 0$$

which is a consequence of  $\mathbf{J}$  being skew-symmetric. Alternatively this is shown by

$$\operatorname{div} \dot{\mathbf{y}} = \sum_{i=1}^d \left( \frac{\partial}{\partial p_i} \frac{dp_i}{dt} + \frac{\partial}{\partial q_i} \frac{dq_i}{dt} \right) = \sum_{i=1}^d \left( -\frac{\partial^2 H}{\partial q_i \partial p_i} + \frac{\partial^2 H}{\partial p_i \partial q_i} \right) = 0$$

Let  $\mathbf{y} = \mathbf{y}(t)$  be the state of a Hamiltonian system at time  $t$ , and let  $\mathbf{y}^* = \mathbf{y}(t^*)$  be the state at time  $t^*$ . Then a fundamental property of Hamiltonian systems is that

$$\left( \frac{\partial \mathbf{y}^*}{\partial \mathbf{y}} \right)^T \mathbf{J} \left( \frac{\partial \mathbf{y}^*}{\partial \mathbf{y}} \right) = \mathbf{J} \quad (14.254)$$

where

$$\left( \frac{\partial \mathbf{y}^*}{\partial \mathbf{y}} \right) = \begin{pmatrix} \frac{\partial \mathbf{p}^*}{\partial \mathbf{p}} & \frac{\partial \mathbf{p}^*}{\partial \mathbf{q}} \\ \frac{\partial \mathbf{q}^*}{\partial \mathbf{p}} & \frac{\partial \mathbf{q}^*}{\partial \mathbf{q}} \end{pmatrix} \quad (14.255)$$

A system that satisfies (14.254) is said to be symplectic. A system is Hamiltonian if and only if it is symplectic.

**Example 236** Consider the system

$$\dot{q} = p, \quad \dot{p} = -\omega_0^2 q \quad (14.256)$$

which has the Hamiltonian

$$H = \frac{1}{2} p^2 + \frac{\omega_0^2}{2} q^2 \quad (14.257)$$

We see that

$$\dot{H} = p\dot{p} + \omega_0^2 q\dot{q} = -\omega_0^2 qp + \omega_0^2 qp = 0 \quad (14.258)$$

Then, as  $H$  is a quadratic invariant for this system, it follows that a Runge-Kutta method satisfying

$$\mathbf{M} = \text{diag}(\mathbf{b}) \mathbf{A} + \mathbf{A}^T \text{diag}(\mathbf{b}) - \mathbf{b}\mathbf{b}^T = \mathbf{0}$$

will result in

$$H_{n+1} = H_n \quad (14.259)$$

Consider the sensitivity function

$$\Psi(t) := \frac{\partial \mathbf{y}(t)}{\partial \mathbf{y}_0} \quad (14.260)$$

which gives the sensitivity of the true solution  $\mathbf{y}(t)$  with respect to the initial condition  $\mathbf{y}(0) = \mathbf{y}_0$ . We see that  $\Psi \in R^{2d \times 2d}$ . The time derivative of the sensitivity when the system evolves is found from the computation

$$\frac{d}{dt} \Psi(t) = \frac{d}{dt} \frac{\partial \mathbf{y}(t)}{\partial \mathbf{y}_0} = \frac{\partial}{\partial \mathbf{y}_0} \left( \frac{d\mathbf{y}(t)}{dt} \right) = \frac{\partial \mathbf{f}(t)}{\partial \mathbf{y}_0} = \frac{\partial \mathbf{f}(t)}{\partial \mathbf{y}(t)} \frac{\partial \mathbf{y}(t)}{\partial \mathbf{y}_0} \quad (14.261)$$

to be

$$\dot{\Psi} = \frac{\partial \mathbf{f}}{\partial \mathbf{y}} \Psi \quad (14.262)$$

If a Runge-Kutta method is applied to this equation, the computational scheme is given by

$$\mathbf{K}_1 = \frac{\partial \mathbf{f}}{\partial \mathbf{y}} (\Psi_n + h(a_{11}\mathbf{K}_1 + \dots + a_{1\sigma}\mathbf{K}_\sigma), t_n + c_1 h) \quad (14.263)$$

$$\vdots \quad (14.264)$$

$$\mathbf{K}_\sigma = \frac{\partial \mathbf{f}}{\partial \mathbf{y}_0} (\Psi_n + h(a_{\sigma 1}\mathbf{K}_1 + \dots + a_{\sigma\sigma}\mathbf{K}_\sigma), t_n + c_\sigma h) \quad (14.265)$$

$$\Psi_{n+1} = \Psi_n + h(b_1\mathbf{K}_1 + \dots + b_\sigma\mathbf{K}_\sigma) \quad (14.266)$$

where  $\mathbf{K}_i$ ,  $i = 1, \dots, \sigma$  are matrices of the same dimension as  $\Psi$ . Suppose that the same Runge-Kutta method is applied to the system  $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, t)$ , and that the solution at time  $t_{n+1}$  is computed to be  $\mathbf{y}_{n+1}$ . Then the sensitivity of the computed solution is found to be given by

$$\frac{\partial \mathbf{k}_1}{\partial \mathbf{y}_0} = \frac{\partial \mathbf{f}}{\partial \mathbf{y}} \left( \frac{\partial \mathbf{y}_n}{\partial \mathbf{y}_0} + h \left( a_{11} \frac{\partial \mathbf{k}_1}{\partial \mathbf{y}_0} + \dots + a_{1\sigma} \frac{\partial \mathbf{k}_\sigma}{\partial \mathbf{y}_0} \right), t_n + c_1 h \right) \quad (14.267)$$

$$\vdots \quad (14.268)$$

$$\frac{\partial \mathbf{k}_\sigma}{\partial \mathbf{y}_0} = \frac{\partial \mathbf{f}}{\partial \mathbf{y}} \left( \frac{\partial \mathbf{y}_n}{\partial \mathbf{y}_0} + h \left( a_{\sigma 1} \frac{\partial \mathbf{k}_1}{\partial \mathbf{y}_0} + \dots + a_{\sigma\sigma} \frac{\partial \mathbf{k}_\sigma}{\partial \mathbf{y}_0} \right), t_n + c_\sigma h \right) \quad (14.269)$$

$$\frac{\partial \mathbf{y}_{n+1}}{\partial \mathbf{y}_0} = \frac{\partial \mathbf{y}_n}{\partial \mathbf{y}_0} + h \left( b_1 \frac{\partial \mathbf{k}_1}{\partial \mathbf{y}_0} + \dots + b_\sigma \frac{\partial \mathbf{k}_\sigma}{\partial \mathbf{y}_0} \right) \quad (14.270)$$

Comparison with (14.263–14.266) shows that

$$\frac{\partial \mathbf{y}_{n+1}}{\partial \mathbf{y}_0} = \Psi_{n+1} \quad (14.271)$$

The condition for the solution to be symplectic is that

$$\left( \frac{\partial \mathbf{y}_{n+1}}{\partial \mathbf{y}_0} \right)^T \mathbf{J} \frac{\partial \mathbf{y}_{n+1}}{\partial \mathbf{y}_0} = \mathbf{J} \quad (14.272)$$

is therefore equivalent to

$$\Psi_{n+1}^T \mathbf{J} \Psi_{n+1} = \mathbf{J} \quad (14.273)$$

This expression will hold for all Runge-Kutta methods for which

$$m_{ij} = b_i a_{ij} + b_j a_{ji} - b_i b_j = 0 \quad (14.274)$$

as

$$V(\Psi) = \Psi^T \mathbf{J} \Psi = \mathbf{J} \quad (14.275)$$

holds for the exact solution, which shows that  $V(\Psi)$  is a quadratic invariant.

This means that for a Runge-Kutta method with  $m_{ij} = 0$ , which is the case for the Gauss methods, the numerically computed solution will be symplectic, and hence it must be the solution of a Hamiltonian system with a Hamiltonian that we denote  $H^*$ . For small time-steps and a smooth Hamiltonian  $H$ , the Hamiltonian system described by  $H^*$  will have solutions that are close to the solutions of the Hamiltonian system described by  $H$ .

From Tables 14.5 and 14.6 we may conclude that the quadratic invariant  $V(\Psi)$  will decrease for methods like Radau IA, Radau IIA, and Lobatto IIIC, while it will increase for methods like Euler's method, Modified Euler, and improved Euler. For methods like Lobatto IIIA, Lobatto IIIB, Heun's method of order 3, and RK4 the invariant  $V(\Psi)$  may decrease or increase.

## 14.10 Rosenbrock methods

A Rosenbrock method with  $\sigma$  stages for the system

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, t) \quad (14.276)$$

is given by (Hairer and Wanner 1996)

$$\begin{aligned} \mathbf{k}_i &= \mathbf{f}(\mathbf{y}_n + h \sum_{j=1}^{i-1} a_{ij} \mathbf{k}_j, t_n + c_i h) \\ &\quad + h \mathbf{J} \sum_{j=1}^i \rho_{ij} \mathbf{k}_j + \rho_i h \dot{\mathbf{f}}(\mathbf{y}_n, t_n), \quad i = 1, \dots, \sigma \\ \mathbf{y}_{n+1} &= \mathbf{y}_n + h \sum_{j=1}^{\sigma} b_j \mathbf{k}_j \end{aligned}$$

where  $\mathbf{J}$  is the Jacobian given by  $\mathbf{J} = \partial \mathbf{f}(\mathbf{y}_n, t_n) / \partial \mathbf{y}$ , the interpolation constants satisfy  $c_i = \sum_{j=1}^{i-1} a_{ij}$  as for the Runge-Kutta methods, and

$$\rho_i = \sum_{j=1}^i \rho_{ij}$$

The first term on the right side of the stage computations has the same form as the stage in an explicit Runge-Kutta method. A linearized term is added to the stage, which

makes the method implicit. However, while a Newton search is required at each time step to compute the stages in an implicit Runge-Kutta method, the stage computations in a Rosenbrock method can be done without iterations according to the formula

$$\begin{aligned}\mathbf{V}_i \mathbf{k}_i &= \mathbf{f}(\mathbf{y}_n + h \sum_{j=1}^{i-1} a_{ij} \mathbf{k}_j, t_n + c_i h) \\ &\quad + h \mathbf{J} \sum_{j=1}^{i-1} \rho_{ij} \mathbf{k}_j + \rho_i h \dot{\mathbf{f}}(\mathbf{y}_n, t_n)\end{aligned}\quad (14.277)$$

where

$$\mathbf{V}_i = \mathbf{I} - h \rho_{ii} \mathbf{J}$$

is a nonsingular matrix for a sufficiently small time step  $h$ .

For the test equation

$$\dot{y} = \lambda y$$

a Rosenbrock method gives

$$\begin{aligned}\boldsymbol{\kappa} &= \lambda (\mathbf{1} y_n + h \lambda (\mathbf{A} + \mathbf{R}) \boldsymbol{\kappa}) \\ y_{n+1} &= y_n + h \mathbf{b}^T \boldsymbol{\kappa}\end{aligned}$$

where  $\boldsymbol{\kappa} = (k_1 \dots k_\sigma)^T$ ,  $\mathbf{1} = (1 \dots 1)^T$  and  $\mathbf{R} = \{\rho_{ij}\}$ .

The stability function of a Rosenbrock method is given by

$$R(h\lambda) = \frac{\det [\mathbf{I} - \lambda h (\mathbf{A} + \mathbf{R} - \mathbf{1}\mathbf{b}^T)]}{\det [\mathbf{I} - \lambda h (\mathbf{A} + \mathbf{R})]} \quad (14.278)$$

It is seen that Rosenbrock methods can have the same type of stability function as an implicit Runge-Kutta method of the diagonally implicit type, which are implicit method with  $a_{ij} = 0$  for  $i > j$ . The main advantage with Rosenbrock methods is that they can be used for stiff systems without Newton iterations in the stage computations.

A second order method with L-stability developed by Wolfbrandt is given by

$$\begin{aligned}\mathbf{V} \mathbf{k}_1 &= \mathbf{f}(\mathbf{y}_n) \\ \mathbf{V} \mathbf{k}_2 &= \mathbf{f}(\mathbf{y}_n + \frac{2}{3} h \mathbf{k}_1) - \frac{4}{3(2 + \sqrt{2})} h \mathbf{J} \mathbf{k}_1 \\ \mathbf{y}_{n+1} &= \mathbf{y}_n + \frac{h}{4} (\mathbf{k}_1 + 3\mathbf{k}_2)\end{aligned}$$

where

$$\mathbf{V} = \mathbf{I} - \frac{1}{2 + \sqrt{2}} h \mathbf{J}$$

A modified second order Rosenbrock method with step size control is given by

$$\begin{aligned}\mathbf{V} \mathbf{k}_1 &= \mathbf{f}(\mathbf{y}_n, t_n) + h \rho \dot{\mathbf{f}}(\mathbf{y}_n, t_n) \\ \mathbf{V} \mathbf{k}_2 &= \mathbf{f}(\mathbf{y}_n, t_n) + \mathbf{f}\left(\mathbf{y}_n + \frac{h}{2} \mathbf{k}_1, t_n + \frac{h}{2}\right) - \mathbf{k}_1 + h \rho \dot{\mathbf{f}}(\mathbf{y}_n, t_n) \\ \mathbf{y}_{n+1} &= \mathbf{y}_n + h \mathbf{k}_2 \\ \mathbf{V} &= \mathbf{I} - h \rho \mathbf{J}, \quad \rho = \frac{1}{2 + \sqrt{2}}\end{aligned}$$

with step size control using a FSAL computation

$$\begin{aligned}\mathbf{V}\mathbf{k}_3 &= 2\mathbf{f}(\mathbf{y}_n, t_n) + \left(6 + \sqrt{2}\right)\mathbf{f}\left(\mathbf{y}_n + \frac{h}{2}\mathbf{k}_1, t_n + \frac{h}{2}\right) + \mathbf{f}(\mathbf{y}_{n+1}, t_{n+1}) \\ &\quad - 2\mathbf{k}_1 - \left(6 + \sqrt{2}\right)\mathbf{k}_2 + h\rho\dot{\mathbf{f}}(\mathbf{y}_n, t_n) \\ \text{error} &= \frac{h}{6}(\mathbf{k}_1 - 2\mathbf{k}_2 + \mathbf{k}_3)\end{aligned}$$

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  $h\mathbf{J}_{21}\mathbf{k}_1$ . 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-Basforth 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

$$\mathbf{P}(t_{n+1-i}) = \mathbf{f}(\mathbf{y}_{n+1-i}, t_{n+1-i}) =: \mathbf{f}_{n+1-i}, \quad i = 1, 2, \dots, q. \quad (14.279)$$

This is done with the polynomial

$$\mathbf{P}(t) = \sum_{i=1}^q \mathbf{f}_{n+1-i} L_i(t)$$

where  $L_i(t)$ ,  $i = 1, \dots, q$  are the fundamental Lagrange polynomials (Shampine et al. 1997)

$$L_i(t) = \prod_{j=1, j \neq i}^q \left( \frac{t - t_{n+1-j}}{t_{n+1-i} - t_{n+1-j}} \right), \quad i = 1, \dots, q$$

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  $\nabla$  by

$$\nabla \mathbf{y}_n = \mathbf{y}_n - \mathbf{y}_{n-1}$$

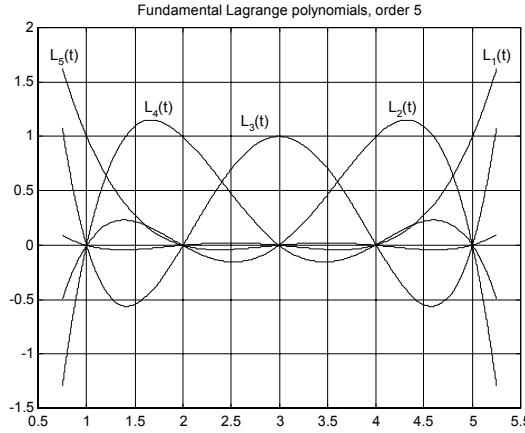


Figure 14.32: The Lagrange polynomials  $L_i(t_{6-j})$  for  $i = 1, \dots, 5$  when  $h = 1$ .

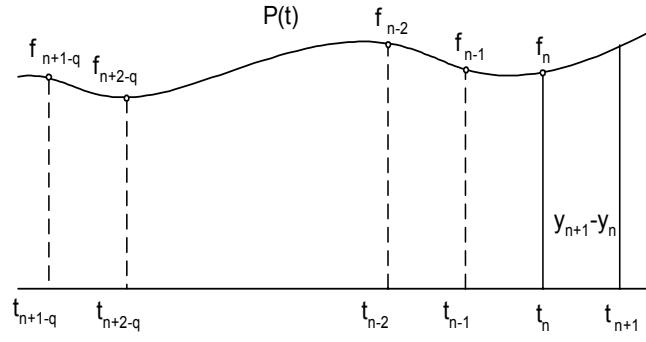


Figure 14.33: The explicit Adams method

Repeated use of the backward difference operator gives

$$\nabla^{m+1} \mathbf{y}_n = \nabla (\nabla^m \mathbf{y}_n) = \nabla^m \mathbf{y}_n - \nabla^m \mathbf{y}_{n-1}$$

for  $m = 0, 1, 2, \dots$  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

$$\mathbf{P}(t_n + \alpha h) = \sum_{m=0}^{q-1} \frac{\alpha(\alpha+1)\dots(\alpha+m-1)}{m!} \nabla^m \mathbf{f}_n$$

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

$$\gamma_m = \int_0^1 \frac{\alpha(\alpha+1)\dots(\alpha+m-1)}{m!} d\alpha$$

It can be shown that  $\gamma_m$  can be found recursively from the recurrence equation

$$\gamma_m + \frac{1}{2}\gamma_{m-1} + \dots + \frac{1}{m+1}\gamma_0 = 1$$

The numerical values for  $\gamma_m$  are found from the recurrence equation to be

$m$	0	1	2	3	4
$\gamma_m$	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

$$\mathbf{P}(t_{n+1-i}) = \mathbf{f}(\mathbf{y}_{n+1-i}, t_{n+1-i}), \quad i = 0, 1, \dots, q-1 \quad (14.280)$$

as shown in Figure 14.34.

This is achieved with

$$\mathbf{P}^*(t_n + \alpha h) = \sum_{m=0}^q \frac{(\alpha-1)\alpha(\alpha+1)\dots(\alpha+m-2)}{m!} \nabla^m \mathbf{f}_{n+1}$$

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

$$\gamma_m^* = \int_0^1 \frac{(\alpha-1)\alpha(\alpha+1)\dots(\alpha+m-2)}{m!} d\alpha$$

Numerical values are

$m$	0	1	2	3	4
$\gamma_m^*$	1	$-\frac{1}{2}$	$-\frac{1}{12}$	$-\frac{1}{24}$	$-\frac{19}{720}$

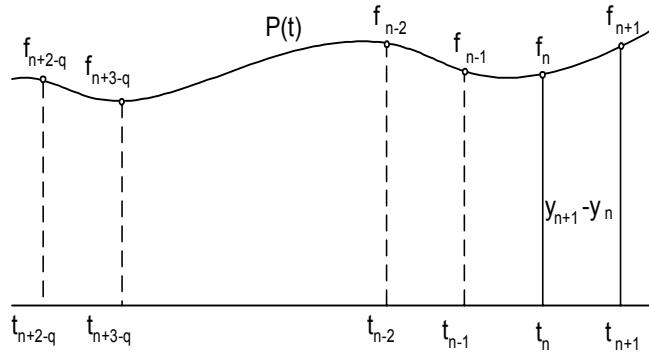


Figure 14.34: The implicit Adams method

which can also be found from the recurrence equation

$$\gamma_m^* + \frac{1}{2}\gamma_{m-1}^* + \dots + \frac{1}{m+1}\gamma_0^* = 0, \quad m \geq 0, \quad \gamma_0^* = 1$$

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

$$\hat{\mathbf{y}}_{n+1} = \mathbf{y}_n + h \sum_{m=0}^{q-1} \gamma_m \nabla^m \mathbf{f}(t_n, x_n)$$

with the explicit Adams method, and then use  $\hat{\mathbf{f}}_{n+1} := \mathbf{f}(t_{n+1}, \hat{\mathbf{y}}_{n+1})$  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

$$\mathbf{P}(t_{n-q+1}) = \mathbf{y}_{n-q+1}, \dots, \mathbf{P}(t_n) = \mathbf{y}_n, \mathbf{P}(t_{n+1}) = \mathbf{y}_{n+1} \quad (14.281)$$

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

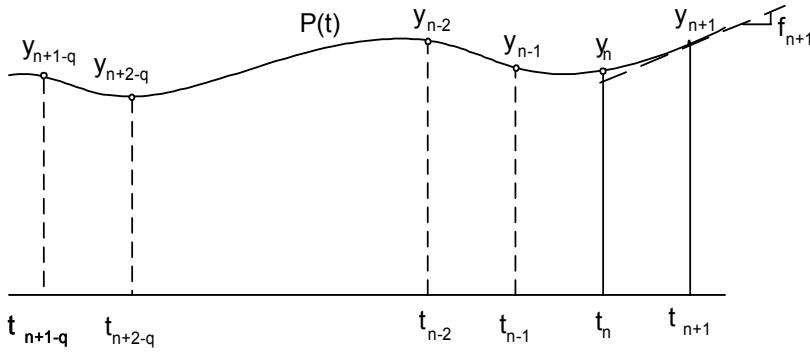


Figure 14.35: The BDF method

$$\mathbf{P}(t_n + \alpha h) = \left( 1 + \sum_{m=1}^q \frac{(\alpha - 1)\alpha(\alpha + 1) \dots (\alpha + m - 2)}{m!} \nabla^m \right) \mathbf{y}_{n+1}$$

From

$$\begin{aligned} \frac{d}{d\alpha} \mathbf{P}(t_n + \alpha h) \Big|_{\alpha=1} &= \sum_{m=1}^q \frac{d}{d\alpha} \frac{(\alpha - 1)\alpha(\alpha + 1) \dots (\alpha + m - 2)}{m!} \Big|_{\alpha=1} \nabla^m \mathbf{y}_{n+1} \\ &= \sum_{m=1}^q \frac{1}{m} \nabla^m \mathbf{y}_{n+1} \end{aligned}$$

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  $q = 1, \dots, 6$ :

$$\begin{aligned} \mathbf{y}_{n+1} - \mathbf{y}_n &= h\mathbf{f}_{n+1} \\ \frac{3}{2}\mathbf{y}_{n+1} - 2\mathbf{y}_n + \frac{1}{2}\mathbf{y}_{n-1} &= h\mathbf{f}_{n+1} \\ \frac{11}{6}\mathbf{y}_{n+1} - 3\mathbf{y}_n + \frac{3}{2}\mathbf{y}_{n-1} - \frac{1}{3}\mathbf{y}_{n-2} &= h\mathbf{f}_{n+1} \\ \frac{25}{12}\mathbf{y}_{n+1} - 4\mathbf{y}_n + 3\mathbf{y}_{n-1} - \frac{4}{3}\mathbf{y}_{n-2} + \frac{1}{4}\mathbf{y}_{n-3} &= h\mathbf{f}_{n+1} \\ \frac{137}{60}\mathbf{y}_{n+1} - 5\mathbf{y}_n + 5\mathbf{y}_{n-1} - \frac{10}{3}\mathbf{y}_{n-2} + \frac{5}{4}\mathbf{y}_{n-3} - \frac{1}{5}\mathbf{y}_{n-4} &= h\mathbf{f}_{n+1} \\ \frac{147}{60}\mathbf{y}_{n+1} - 6\mathbf{y}_n + \frac{15}{2}\mathbf{y}_{n-1} - \frac{20}{3}\mathbf{y}_{n-2} + \frac{15}{4}\mathbf{y}_{n-3} - \frac{6}{5}\mathbf{y}_{n-4} + \frac{1}{6}\mathbf{y}_{n-5} &= h\mathbf{f}_{n+1} \end{aligned}$$

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

$$\alpha_q \mathbf{y}_{n+1} + \alpha_{q-1} \mathbf{y}_n + \dots + \alpha_0 \mathbf{y}_{n+1-q} = h (\beta_q \mathbf{f}_{n+1} + \beta_{q-1} \mathbf{f}_n + \dots + \beta_0 \mathbf{f}_{n+1-q})$$

which is written

$$N(z) \mathbf{y}_{n+1} = h D(z) \mathbf{f}_{n+1} \quad (14.282)$$

where

$$\begin{aligned} N(z) &= \alpha_q z^q + \alpha_{q-1} z^{q-1} + \dots + \alpha_0 \\ D(z) &= \beta_q z^q + \beta_{q-1} z^{q-1} + \dots + \beta_0 \end{aligned}$$

and  $z$  is viewed as the time-shift operator defined by  $z^{-1} \mathbf{y}_{n+1} = \mathbf{y}_n$ .

Consider the linear test system

$$\dot{y} = \lambda y$$

Then the multistep method gives

$$N(z) \mathbf{y}_{n+1} = h \lambda D(z) \mathbf{y}_{n+1}$$

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  $\mathbf{y}_{n+1}$ . Equivalently, this is written

$$[N(z) - h \lambda D(z)] y(z) = 0 \quad (14.283)$$

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  $\lambda$  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  $\lambda$  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

$$h\lambda = \frac{N(e^{j\theta})}{D(e^{j\theta})}, \quad -\pi \leq \theta \leq \pi \quad (14.284)$$

#### 14.11.6 Stability of Adams methods

In terms of the  $z$  transformation the backwards differences operator  $\nabla$  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

$$zy(z) = y(z) + h\lambda \sum_{m=0}^q \gamma_m^* (1 - z^{-1})^m zy(z)$$

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  $\hat{y}_{n+1}$  of the explicit Adams method is inserted for  $y_{n+1}$  on the right hand side of the implicit method.

$$zy(z) = y(z) + h\lambda \left\{ \gamma_0^* z\hat{y}(z) + \gamma_1^* [z\hat{y}(z) - y(z)] + \gamma_2^* [z\hat{y}(z) - 2y(z) + z^{-1}y(z)] + \dots \right\}$$

After some calculation it can be established that  $h\lambda$  satisfies the second order equation

$$\begin{aligned} A(h\lambda)^2 + Bh\lambda + C &= 0 \\ A &= \left( \sum_{m=0}^q \gamma_m^* \right) \left[ \sum_{m=0}^{q-1} \gamma_m (1 - z^{-1})^m \right] \\ B &= (1 - z) \sum_{m=0}^q \gamma_m^* + z \sum_{m=0}^q \gamma_m^* (1 - z^{-1})^m \\ C &= 1 - z \end{aligned}$$

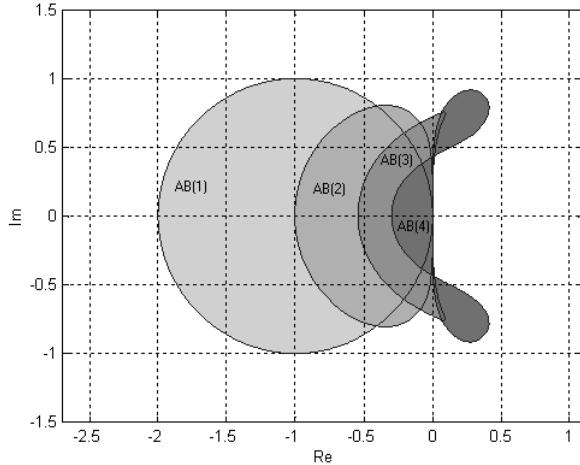


Figure 14.36: The stability regions of the explicit Adams (Adams-Basforth) methods of order 1 to 4. We recognize the stability area of AB(1) as that of Euler's method.

#### 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  $\text{Re}(\lambda) \leq 0$ .

#### 14.11.8 Frequency response

From

$$[N(z) - h\lambda D(z)] y(z) = 0 \quad (14.285)$$

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

$$N(z_p) - h\lambda D(z_p) = 0 \quad (14.286)$$

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  $y_{n+1} = y_n$ .

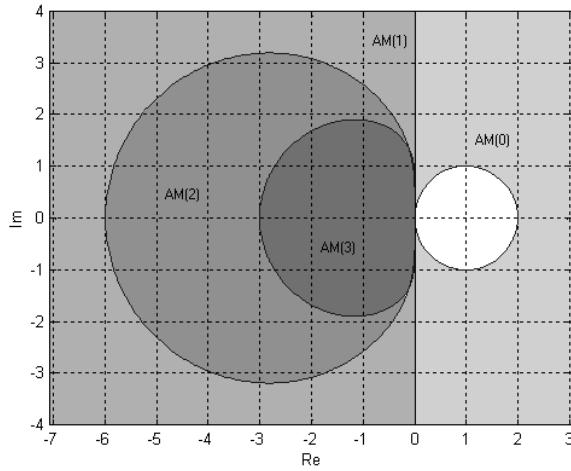


Figure 14.37: Stability areas of implicit Adams (Adams-Moulton) methods of order 1 to 4. The methods are denoted by  $\text{AM}(q)$  where  $q + 1$  is the order of the method. Note that  $\text{AM}(1)$  is the implicit Euler method, and  $\text{AM}(2)$  is the trapezoidal rule.

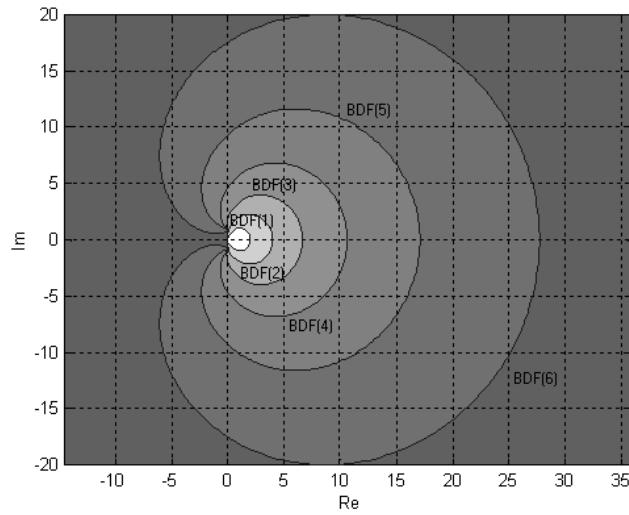


Figure 14.38: Stability areas for BDF methods of order 1 to 6. We note that  $\text{BDF}1$  is the implicit Euler method.

### 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  $|s| \rightarrow \infty$ , 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  $|s| \rightarrow \infty$  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

$$\alpha_q y_{n+1} + \alpha_{q-1} y_n + \dots + \alpha_0 = \lambda h y_{n+1} \quad (14.287)$$

leads to

$$(\alpha_q - \lambda h) y_{n+1} + \alpha_{q-1} y_n + \dots + \alpha_0 = 0 \quad (14.288)$$

We see that when  $\lambda h \rightarrow \infty$ , then  $y_{n+1} \rightarrow 0$ . In the  $z$  transform the result is found from

$$[(\alpha_q - \lambda h) z^q + \alpha_{q-1} z^{q-1} + \dots + \alpha_0] y(z) = 0 \quad (14.289)$$

where it is seen that when  $\lambda h \rightarrow \infty$  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

$$\mathbf{M}\dot{\mathbf{u}} = \phi(\mathbf{u})$$

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

$$\mathbf{M} = \begin{pmatrix} \mathbf{I}_{d_1} & \mathbf{0} \\ \mathbf{0} & \epsilon \mathbf{I}_{d_2} \end{pmatrix} \quad (14.290)$$

where  $d_1 + d_2 = d$  and  $\epsilon$  is a constant. It is seen that  $\mathbf{M}$  is singular whenever  $\epsilon = 0$ .

Let

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} = \mathbf{u}, \quad \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix} = \phi$$

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

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, \mathbf{z}) \quad (14.291)$$

$$\epsilon \dot{\mathbf{z}} = \mathbf{g}(\mathbf{y}, \mathbf{z}) \quad (14.292)$$

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*

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, \mathbf{z}) \quad (14.293)$$

$$\mathbf{0} = \mathbf{g}(\mathbf{y}, \mathbf{z}) \quad (14.294)$$

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

$$\dot{\mathbf{y}} = \mathbf{f}[\mathbf{y}, \mathbf{z}(\mathbf{y})] \quad (14.295)$$

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  $\epsilon$  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  $\mathbf{M}\dot{\mathbf{u}} = \phi(\mathbf{u})$  where  $\mathbf{M}$  is possibly nonsingular.

The system

$$\mathbf{M}\dot{\mathbf{u}} = \phi(\mathbf{u})$$

is said to be a differential algebraic equation of index 1 if it can be transformed into 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  $\epsilon$ . This is done by solving  $\mathbf{g}(\mathbf{Y}_j, \mathbf{Z}_j)$  from the equation for  $\epsilon \mathbf{Z}_i$ , which gives

$$h \mathbf{g}(\mathbf{Y}_j, \mathbf{Z}_j) = \epsilon \sum_{j=1}^{\sigma} \omega_{ij} (\mathbf{Z}_i - \mathbf{z}_n)$$

where  $\mathbf{A}^{-1} = \boldsymbol{\Omega} = \{\omega_{ij}\}$ . This expression is inserted into the equation for  $\epsilon \mathbf{z}_{n+1}$ , and the result is

$$\begin{aligned}\epsilon \mathbf{z}_{n+1} &= \epsilon \mathbf{z}_n + \epsilon \sum_{i=1}^{\sigma} b_i \sum_{j=1}^{\sigma} \omega_{ij} (\mathbf{Z}_i - \mathbf{z}_n) \\ &= \epsilon \left( 1 - \sum_{i=1}^{\sigma} \sum_{j=1}^{\sigma} b_i \omega_{ij} \right) \mathbf{z}_n + \epsilon \sum_{i=1}^{\sigma} \sum_{j=1}^{\sigma} b_i \omega_{ij} \mathbf{Z}_i\end{aligned}$$

We note that  $\epsilon$  may be cancelled from this equation, and recall from (14.156) that

$$R(\infty) = 1 - \mathbf{b}^T \mathbf{A}^{-1} \mathbf{1} = 1 - \sum_{i=1}^{\sigma} \sum_{j=1}^{\sigma} b_i \omega_{ij} \quad (14.296)$$

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  $\epsilon$  only appears in the equation for  $\epsilon \mathbf{Z}_i$ . If we let  $\epsilon$  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  $\epsilon$ .

For a general possibly singular  $\mathbf{M}$  the method is written

$$\begin{aligned}\mathbf{M}(\mathbf{U}_i - \mathbf{u}_n) &= h \sum_{j=1}^{\sigma} a_{ij} \phi(\mathbf{U}_j) \\ \mathbf{u}_{n+1} &= R(\infty)\mathbf{u}_n + \sum_{i=1}^{\sigma} \sum_{j=1}^{\sigma} b_i \omega_{ij} \mathbf{U}_i\end{aligned}$$

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

$$\mathbf{M}\dot{\mathbf{u}} = \phi(\mathbf{u})$$

the BDF and NDF method are given by

$$\sum_{m=1}^q \alpha_q (\mathbf{Mu})_{n+m-q} = h\phi(\mathbf{u}_{n+1})$$

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

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} = \mathbf{u}, \quad \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix} = \phi$$

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.



# Chapter 15

# Computational fluid dynamics

## 15.1 Introduction

As shown in Chapter 10, models from fluid mechanics and thermodynamics often involve partial differential equations. In particular this applies to the transport equations and the Navier-Stokes equations. These equations cannot be solved analytically except in special cases and there is a need for finding approximate solutions. Computational fluid dynamics (CFD) is the collection of tools and methods used for solving this through simulation. CFD covers both mathematical modelling of the application at hand, methods of discretization, numerical grid generation and methods of solving the sets of nonlinear algebraic equations arising from the discretization. CFD methods can be divided into at least three groups by the method of discretization. These three are finite difference, finite volume and finite element methods. Due to its simplicity and the fact that the approximation terms can be given a physical interpretation, we will in the following use *the finite volume method* to compute approximate numerical solutions to some important types of fluid dynamic problems.

The finite volume method uses the integral form of the balance equations as its starting point. The solution domain is divided into a finite number of control volumes (CV), and the conservation equations are applied to each CV. At the centroid of each CV lies a computational node at which the variables are to be calculated. Interpolation of the nodal values of neighboring nodes are used at the CV surfaces. This finally results in an algebraic equation for each node, in which a number of neighbor node values appear, and a system of equations can now be found. The size of this system depends on the size of the domain and the grid spacing, but a system of one million equations is not uncommon. The finite volume method always yields systems with equations with a large number of zero entries. The structure of the equations depend on the differencing scheme used, but in most cases a diagonal structure is ensured. This can be exploited by the solution technique, and iterative methods are often used. Further details on CFD are found in (Patankar 1980), (Anderson 1995), (Ferziger and Perić 1999) and (Versteeg and Malalasekera 1995)

## 15.2 Governing equations

The governing equations for fluid properties such as mass, velocity and energy have many similarities, and it will be advantageous to introduce a general variable  $\phi$  to denote a

conserved quantity like mass, momentum or energy. The balance equation for a conserved quantity  $\phi$  is written

$$\underbrace{\frac{\partial(\rho\phi)}{\partial t}}_{\text{rate of change of } \phi \text{ of spatial fluid element}} + \underbrace{\nabla^T(\rho\phi\mathbf{v})}_{\text{Net rate of flow of } \phi \text{ out of spatial fluid element}} = \underbrace{\nabla^T(\Gamma\nabla\phi)}_{\text{rate of change of } \phi \text{ due to diffusion}} + \underbrace{S_\phi}_{\text{rate of change of } \phi \text{ due to sources}} \quad (15.1)$$

in divergence form. Here  $\mathbf{v}$  is velocity,  $\rho$  is density and  $\Gamma$  is a diffusion coefficient. Equation (15.1) gives the continuity equation (11.6) by substituting  $\phi = 1$ , the momentum (Navier-Stokes) equations (11.299) by substituting  $\phi = v_i$ , and the energy equation (11.165) by setting  $\phi = e$ .

The key step in the finite volume method is the integration of equation (15.1) over a constant control volume  $V$  to give

$$\iiint_V \frac{\partial(\rho\phi)}{\partial t} dV + \iiint_V \nabla^T(\rho\phi\mathbf{v}) dV = \iint_V \nabla^T(\Gamma\nabla\phi) dV + \iiint_V S_\phi dV$$

By use of the divergence theorem as given by equation (10.12), we find

$$\underbrace{\frac{\partial}{\partial t} \left( \iiint_V \rho\phi dV \right)}_{\text{rate of change of } \phi} + \underbrace{\iint_{\partial V} \mathbf{n}^T(\rho\phi\mathbf{v}) dA}_{\text{Net rate of change of } \phi \text{ due to convection across the boundaries}} = \underbrace{\iint_{\partial V} \mathbf{n}^T(\Gamma\nabla\phi) dA}_{\text{Net rate of change of } \phi \text{ due to diffusion across the boundaries}} + \underbrace{\iiint_V S_\phi dV}_{\text{Net rate of creation of } \phi}$$

which describes the conservation of a fluid property for a finite size control volume  $V$ . In steady state problems, the rate of change term is zero, which leads to the integrated form of the steady state transport equation:

$$\iint_{\partial V} \mathbf{n}^T(\rho\phi\mathbf{v}) dA = \iint_{\partial V} \mathbf{n}^T(\Gamma\nabla\phi) dA + \iiint_V S_\phi dV$$

In transient time-dependent problems it is also necessary to integrate with respect to time  $t$ :

$$\begin{aligned} & \int \frac{\partial}{\partial t} \left( \iiint_V \rho\phi dV \right) dt + \int \iint_{\partial V} \mathbf{n}^T(\rho\phi\mathbf{v}) dA dt \\ &= \int \iint_{\partial V} \mathbf{n}^T(\Gamma\nabla\phi) dA dt + \int \iiint_V S_\phi dV dt \end{aligned}$$

### 15.3 Classification

Classification of partial differential equations is an important concept for solving such equations. Different methods can be developed for different types of equations such that

distinct properties of the equations can be utilized. The governing partial differential equations of fluid dynamics as derived in Chapter 10 are quasi linear. This means that the highest order derivatives occur linearly, they appear by themselves, multiplied with coefficients which are functions of the dependent variables. It is useful to examine the mathematical properties of such equations, as any numerical solution of the equations should exhibit the property of obeying the general mathematical properties of the governing equations. We will establish a classification of three types of differential equations: elliptic, hyperbolic and parabolic – all three of which are encountered in fluid dynamics.

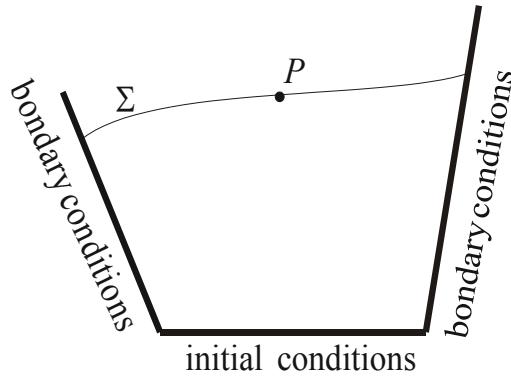


Figure 15.1: General propagation problem.

Consider the system of quasi-linear equations given below

$$a_1 \frac{\partial u}{\partial x} + b_1 \frac{\partial u}{\partial y} + c_1 \frac{\partial v}{\partial x} + d_1 \frac{\partial v}{\partial y} = f_1 \quad (15.2)$$

$$a_2 \frac{\partial u}{\partial x} + b_2 \frac{\partial u}{\partial y} + c_2 \frac{\partial v}{\partial x} + d_2 \frac{\partial v}{\partial y} = f_2 \quad (15.3)$$

where  $u$  and  $v$  are the dependent variables, and the coefficients  $a_i, b_i, c_i, d_i$  and  $f_i$  may be functions of  $x, y, u$  and  $v$ . Note that these are not the flow equations, but they are similar in some respects. In Figure 15.1, we have sketched a general problem where the solution for  $u$  and  $v$  is known below and on the curve  $\Sigma$ . The solution propagates from the known initial conditions and with known boundary conditions. In a point  $P$  on this curve we know the values of  $u$  and  $v$  and their directional derivatives in directions "downwards" from  $\Sigma$ . We are now interested in whether it is possible to decide if the solution above  $P$  is given by the information below and on the curve  $\Sigma$ . Or, equivalently: Are these data sufficient to decide the directional derivatives in  $P$  in directions "upwards" from  $\Sigma$ ? The directional derivatives can be found from the total differentials of  $u$  and  $v$ , which are given by

$$du = \frac{\partial u}{\partial x} dx + \frac{\partial u}{\partial y} dy \quad (15.4)$$

$$dv = \frac{\partial v}{\partial x} dx + \frac{\partial v}{\partial y} dy \quad (15.5)$$

Thus, the total differentials  $du$  and  $dv$  can be found if  $\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial v}{\partial x}$  and  $\frac{\partial v}{\partial y}$  are known, and we will now investigate under which conditions the partial derivatives in  $P$  can be

calculated from the values of  $u$  and  $v$  on  $\Sigma$ . Equations (15.2) to (15.5) can be written as a linear system in the four unknowns  $\frac{\partial u}{\partial x}$ ,  $\frac{\partial u}{\partial y}$ ,  $\frac{\partial v}{\partial x}$  and  $\frac{\partial v}{\partial y}$ :

$$\underbrace{\begin{pmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ dx & dy & 0 & 0 \\ 0 & 0 & dx & dy \end{pmatrix}}_{\mathbf{A}} \begin{pmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} \\ \frac{\partial v}{\partial y} \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ du \\ dv \end{pmatrix} \quad (15.6)$$

If  $u$  and  $v$  are known in  $P$ , then  $a_i, b_i, c_i, d_i$  and  $f_i$  are also known, and if  $\Sigma$  is known then  $dx$  and  $dy$  are known. Moreover, if  $u$  and  $v$  are known on  $\Sigma$ , then  $du$  and  $dv$  are known. A unique solution for the four partial derivatives will then exist if and only if  $\det A \neq 0$ , and the directional derivatives will have the same values above and below  $\Sigma$ . If on the other hand  $\det A = 0$ , (15.6) will have multiple solutions, and the partial derivative might be discontinuous over  $\Sigma$ . The characteristic equation of the system is found by setting  $\det A = 0$ :

$$\begin{vmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ dx & dy & 0 & 0 \\ 0 & 0 & dx & dy \end{vmatrix} = 0$$

↓

$$(a_1c_2 - a_2c_1)(dy)^2 - (a_1d_2 - a_2d_1 + b_1c_2 - b_2c_1)dxdy + (b_1d_2 - b_2d_1)(dx)^2 = 0$$

↓

$$(a_1c_2 - a_2c_1)\left(\frac{dy}{dx}\right)^2 - (a_1d_2 - a_2d_1 + b_1c_2 - b_2c_1)\frac{dy}{dx} + (b_1d_2 - b_2d_1) = 0 \quad (15.7)$$

The directions given by (15.7) are called characteristic directions, and a curve, plane or hyperplane of points where (15.7) is satisfied, is called a characteristic. The characteristic directions can be real and distinct, real and coinciding or imaginary dependent on the *discriminant*

$$D = (a_1d_2 - a_2d_1 + b_1c_2 - b_2c_1)^2 - 4(a_1c_2 - a_2c_1)(b_1d_2 - b_2d_1) = B^2 - 4AC$$

being positive, zero or negative. If  $D > 0$ , two distinct characteristic lines exist through each point in the  $xy$ -plane, and the system (15.2)-(15.3) is called *hyperbolic*, if  $D = 0$  one characteristic line exist through each point in the  $xy$ -plane, and the system (15.2)-(15.3) is called *parabolic*, and if  $D < 0$  the characteristic lines are imaginary, and the system is called *elliptic*.

Alternatively, consider the second order equation

$$a\frac{\partial^2 u}{\partial x^2} + b\frac{\partial^2 u}{\partial x \partial y} + c\frac{\partial^2 u}{\partial y^2} + e\frac{\partial u}{\partial x} + g\frac{\partial u}{\partial y} + hu = f \quad (15.8)$$

where  $a, b, c, d, e, f$  and  $g$  may be functions of  $x, y, u, \frac{\partial u}{\partial x}$  and  $\frac{\partial u}{\partial y}$ . As before, we will find conditions so that  $\frac{\partial^2 u}{\partial x^2}$ ,  $\frac{\partial^2 u}{\partial x \partial y}$  and  $\frac{\partial^2 u}{\partial y^2}$  are uniquely defined on  $\Sigma$  based on *known* values

of  $u$ ,  $\frac{\partial u}{\partial x}$  and  $\frac{\partial u}{\partial y}$  on  $\Sigma$ . Using the same argument leading to (15.6) we have

$$d\left(\frac{\partial u}{\partial x}\right) = \frac{\partial^2 u}{\partial x^2}dx + \frac{\partial^2 u}{\partial x \partial y}dy \quad (15.9)$$

$$d\left(\frac{\partial u}{\partial y}\right) = \frac{\partial^2 u}{\partial y \partial x}dx + \frac{\partial^2 u}{\partial y^2}dy \quad (15.10)$$

Equations (15.8), (15.9) and (15.10) can now be written

$$\begin{pmatrix} a & b & c \\ dx & dy & 0 \\ 0 & dx & dy \end{pmatrix} \begin{pmatrix} \frac{\partial^2 u}{\partial x^2} \\ \frac{\partial^2 u}{\partial x \partial y} \\ \frac{\partial^2 u}{\partial y^2} \end{pmatrix} = \begin{pmatrix} f - e\frac{\partial u}{\partial x} - g\frac{\partial u}{\partial y} - hu \\ d\left(\frac{\partial u}{\partial x}\right) \\ d\left(\frac{\partial u}{\partial y}\right) \end{pmatrix}$$

with characteristic equation

$$\begin{aligned} a(dy)^2 - b dxdy + c(dx)^2 &= 0 \\ \downarrow \\ a\left(\frac{dy}{dx}\right)^2 - b\frac{dy}{dx} + c &= 0 \end{aligned} \quad (15.11)$$

Based on (15.11), equation (15.8) will be hyperbolic if  $b^2 - 4ac > 0$ , parabolic if  $b^2 - 4ac = 0$  and elliptic if  $b^2 - 4ac < 0$ .

**Remark 8** Notice that the classification of the PDE (15.8), is dependent only on the coefficients of the second order derivatives.

We will now review some characteristic properties for each type of equation.

### 15.3.1 Hyperbolic equations

The characteristics of a hyperbolic equation in two variables are plotted in Figure 15.2. Information at point  $P$  influences only the region between the characteristics, that is the effect of a small disturbance at point  $P$  is felt **only** in the *region of influence*. Assume that boundary conditions have been specified on the  $y$ -axis. Then the solution can be found by "marching forward" along the  $x$ -axis, starting from the given boundary. The solution at  $P$  will depend only on the part of the boundary conditions that are given between points  $a$  and  $b$  on the  $y$ -axis. The region to the left of  $P$  is called the *domain of dependence*, that is properties at  $P$  depends **only** on what is happening in this region. The solution of hyperbolic equations can be set up as "marching" solutions, that is starting with the initial conditions, i.e. at the  $y$ -axis, and sequentially calculating the flow field step by step marching in the  $x$  direction. This technique is known as *space-marching*. Steady inviscid supersonic flow is an example of flow that is governed by a hyperbolic equation. Another example is unsteady, inviscid flow. But in this case the governing equation is hyperbolic with respect to time, as depicted in Figure 15.3. In such equations, the marching variable is always time, and the technique is known as *time-marching*.

**Example 237** The second order wave equation is given as

$$\frac{\partial^2 u}{\partial t^2} - c_s^2 \frac{\partial^2 u}{\partial x^2} = 0 \quad (15.12)$$

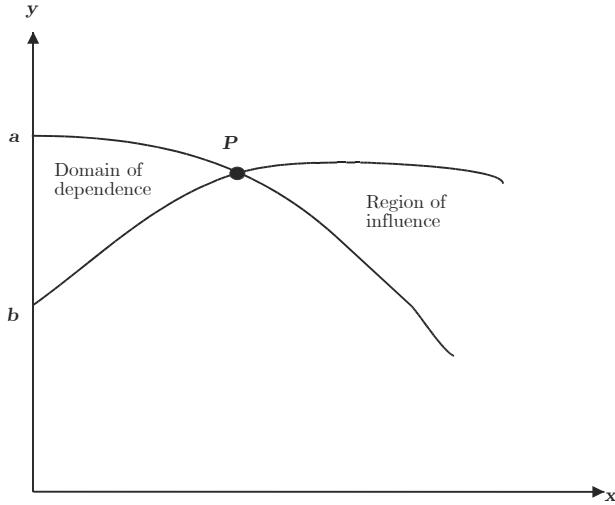


Figure 15.2: Characteristics for a hyperbolic equation in two dimensions.

From (15.8) we recognize  $a = 1$ ,  $b = 0$  and  $c = -c_s^2$ . The discriminant is given by

$$D = b^2 - 4ac = 4c_s^2 > 0$$

and the equation is hyperbolic. To find the characteristics, we now select  $dx/dt$  so that

$$a \left( \frac{dx}{dt} \right)^2 - b \frac{dx}{dt} + c = 0 \quad (15.13)$$

which is the characteristic equation, see (15.11). The characteristic directions are given by

$$\frac{dx}{dt} = \frac{0 \pm \sqrt{0 + 4 \cdot 1 \cdot c_s^2}}{2 \cdot 1} = \pm c_s$$

We conclude that the second order wave equation has two real characteristics.

### 15.3.2 Parabolic equations

The characteristic of a parabolic equation is shown as the dotted vertical line in Figure 15.4. Assume that initial conditions are given along the line  $ab$ , and boundary conditions are known along  $cd$  and  $ab$ . Information at  $P$  influences the region to the right of the characteristic, the region of influence. Parabolic equations are also solved by marching techniques. Steady boundary-layer flows and unsteady thermal conduction are examples of problems governed by parabolic equations. The latter being parabolic with respect to time.

**Example 238** The diffusion equation

$$\frac{\partial \phi}{\partial t} = \alpha \frac{\partial^2 \phi}{\partial x^2} \quad (15.14)$$

is a typical example of a parabolic equation. This can be established by recognizing from (15.8) that  $a = \alpha$ ,  $b = c = 0$ . We also have that  $g = 1$ , but this does not influence the

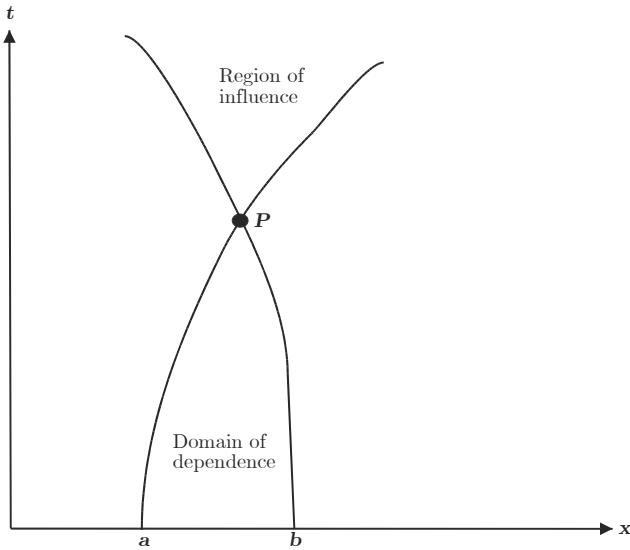


Figure 15.3: Characteristics for a equation hyperbolic with respect to time.

classification. The discriminant is

$$D = b^2 - 4ac = 0$$

and it follows that (15.14) is parabolic. The characteristic is found by

$$a \left( \frac{dt}{dx} \right)^2 - b \frac{dt}{dx} + c = 0 \quad (15.15)$$

$$\alpha \left( \frac{dt}{dx} \right)^2 = 0 \quad (15.16)$$

$$\frac{dt}{dx} = 0 \quad (15.17)$$

The parabolic equation has one characteristic given by  $\frac{dt}{dx} = 0$ .

### 15.3.3 Elliptic equations

For elliptic equations there are no limited regions of influence or domains of dependence. Information is propagated everywhere in all directions. Consider point  $P$  in Figure 15.5. A disturbance at  $P$  will be felt everywhere throughout the region, and the solution at  $P$  is influenced by the entire boundary  $abcd$ . Therefore the solution at  $P$  must be carried out simultaneously with the solution of all the other points in the domain and boundary conditions must be applied for the entire boundary. For this reason, marching solutions can not be used for elliptic equations. Boundary conditions can be a specification of the dependent variables  $u$  and  $v$  in which case the conditions are known as a *Dirichlet condition*, or they can be a specification of the derivatives such as  $\frac{\partial u}{\partial x}$  in which case they are called *Neumann conditions*. The boundary conditions can also be of mixed form, that is involving both Dirichlet and Neumann conditions. Examples of flow that are

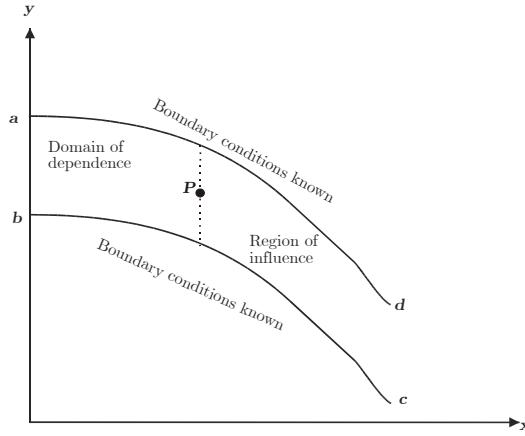


Figure 15.4: Characteristic of a parabolic equation

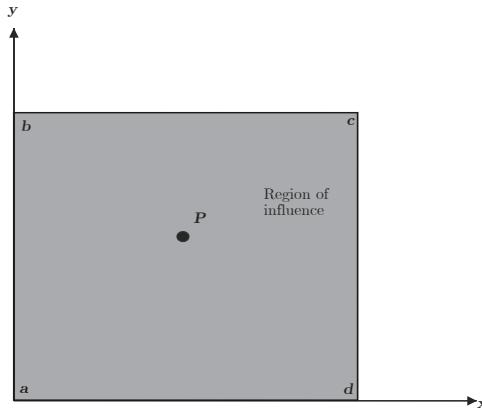


Figure 15.5: Domain and boundaries for an elliptic equation in two dimensions.

governed by elliptic equations include steady, subsonic, inviscid flow and incompressible inviscid flow.

**Example 239** *The Laplace equation which e.g. describes steady state conductive heat transfer is a typical example of an elliptic equation. In two dimensions we have*

$$\begin{aligned} \nabla^2\phi &= 0 \\ \frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} &= 0 \end{aligned} \tag{15.18}$$

By recognizing that  $a = 1$ ,  $b = 0$  and  $c = 1$  and calculating the discriminant

$$D = b^2 - 4ac = -4,$$

it can be concluded that (15.18) is elliptic. Further, the characteristic is found by

$$a \left( \frac{dx}{dy} \right)^2 - b \frac{dx}{dy} + c = 0 \quad (15.19)$$

$$\left( \frac{dx}{dy} \right)^2 = -1 \quad (15.20)$$

$$\frac{dx}{dy} = \pm i \quad (15.21)$$

The Laplace equation has no real characteristic, a fact that applies to all elliptic equations.

We will now present and apply the finite volume method of CFD to a number of well known problems of increasing complexity.

## 15.4 Diffusion

### 15.4.1 Introduction

We start by considering the simplest possible transport process: pure diffusion in steady state, which is governed by the equation

$$\underbrace{\nabla^T (\Gamma \nabla \phi)}_{\substack{\text{rate of change} \\ \text{of } \phi \text{ due to} \\ \text{diffusion}}} + \underbrace{S_\phi}_{\substack{\text{rate of change} \\ \text{of } \phi \text{ due to} \\ \text{sources}}} = 0 \quad (15.22)$$

and can be found by deleting the two terms on the left hand side of (15.1).

### 15.4.2 Finite volume method for stationary diffusion

#### 1D stationary diffusion

In one dimension, (15.22) is given by

$$\frac{d}{dx} \left( \Gamma \frac{d\phi}{dx} \right) + S = 0 \quad (15.23)$$

where  $\Gamma$  is called the diffusion coefficient, and  $S$  is the source term. Boundary conditions  $\phi_A = c_A$  and  $\phi_B = c_B$  where  $c_A$  and  $c_B$  are constants are provided. This process will be used to illustrate the three basic steps grid generation, discretization, solution of equations, in the finite volume method.

**Grid generation** In the finite volume method we want to divide the domain into discrete control volumes. A number of nodal points are placed between the boundaries  $A$  and  $B$ . Each node is placed in the center of its control volume, such that the control volume surfaces are positioned mid-way between the nodes. This arrangement is illustrated in Figure 15.6, and it is further detailed in Figure 15.7. The general node is named  $P$ ,

and its adjacent nodes  $W$  (west) and  $E$  (east), respectively. Lower case letters  $w$  and  $e$  are used for the boundaries. The distance between  $P$  and  $E$  is termed  $\delta x_{PE}$ , with other distances labeled in a similar manner, see Figure 15.7. Notice that the control volume width is given by  $\Delta x = \delta x_{we}$ . In two and three dimensions, the notation is extended with the points  $N$  (north),  $S$  (south),  $T$  (top) and  $B$  (bottom) and corresponding lower case letters for the boundaries.

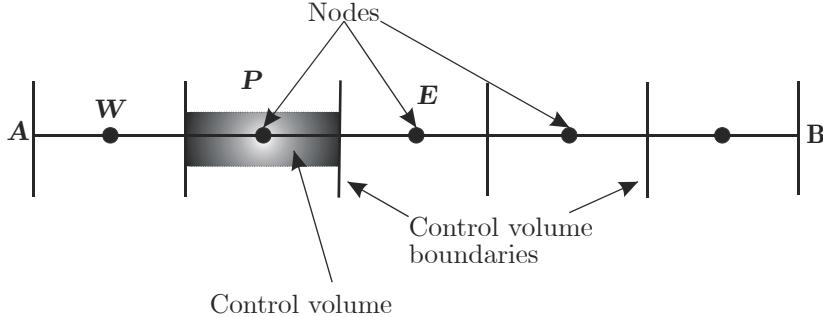


Figure 15.6: Control volume and boundaries.

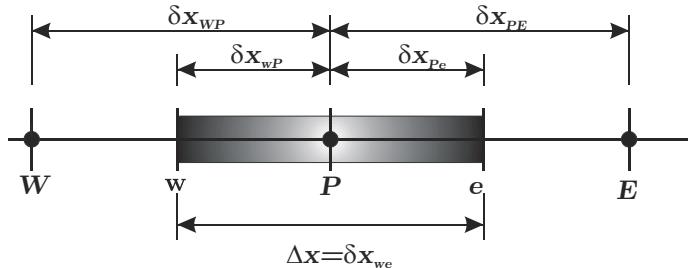


Figure 15.7: Labeling of nodes and distances in the finite volume method.

**Discretization** Integration of (15.23) over the control volume  $\Delta V$  gives

$$\int_{\Delta V} \frac{d}{dx} \left( \Gamma \frac{d\phi}{dx} \right) dV + \int_{\Delta V} S dV = 0$$

and by using the divergence theorem

$$\int_A \mathbf{n}^T \left( \Gamma \frac{d\phi}{dx} \right) dA + \int_{\Delta V} S dV = 0$$

and evaluate at the east and west surfaces of the control volume we get

$$\left( \Gamma A \frac{d\phi}{dx} \right)_e - \left( \Gamma A \frac{d\phi}{dx} \right)_w + \bar{S} \Delta V = 0 \quad (15.24)$$

where  $A$  is the cross-sectional area of the control volume face,  $\Delta V$  is the volume, and  $\bar{S}$  is average value of  $S$  over the control volume. Notice that this discretized equation has

a clear physical interpretation. It states that the diffusive flux of  $\phi$  leaving the  $e$  surface minus diffusive flux of  $\phi$  entering through the  $w$  surface is equal to the generation of  $\phi$  inside the control volume. This is in fact one of the most attractive features of this method. The resulting solution would imply that the integral conservation of quantities such as mass, temperature and energy is exactly satisfied over any group of control volumes. This holds for any number of grid points, and even coarse grid solutions exhibit exact integral balances. We will now use the central difference to evaluate the value of the diffusion coefficient  $\Gamma$  and the gradient  $\frac{d\phi}{dx}$  at the interfaces of the volume:

$$\Gamma_w = \frac{\Gamma_W + \Gamma_P}{2} \quad (15.25)$$

$$\Gamma_e = \frac{\Gamma_P + \Gamma_E}{2} \quad (15.26)$$

such that the diffusive terms in (15.24) can be evaluated as

$$\left( \Gamma A \frac{d\phi}{dx} \right)_w = \Gamma_w A_w \frac{\phi_P - \phi_W}{\delta x_{WP}} \quad (15.27)$$

$$\left( \Gamma A \frac{d\phi}{dx} \right)_e = \Gamma_e A_e \frac{\phi_E - \phi_P}{\delta x_{PE}} \quad (15.28)$$

The source term may be a function of  $\phi$ , and in such cases it may be linearized as

$$\bar{S}\Delta V = S_u + S_p\phi_P. \quad (15.29)$$

Substitution of (15.27), (15.28) and (15.29) into (15.24) gives

$$\underbrace{\left( \frac{\Gamma_e}{\delta x_{PE}} A_e + \frac{\Gamma_e}{\delta x_{WP}} A_w - S_p \right)}_{a_P} \phi_P = \underbrace{\left( \frac{\Gamma_w}{\delta x_{WP}} A_w \right)}_{a_W} \phi_W + \underbrace{\left( \frac{\Gamma_e}{\delta x_{PE}} A_e \right)}_{a_E} \phi_E + S_u$$

or

$$a_P\phi_P = a_W\phi_W + a_E\phi_E + S_u \quad (15.30)$$

which represent the discretized version of (15.22)

### Finite volume method for 2D steady state diffusion

In two dimensions, (15.22) is given by

$$\frac{\partial}{\partial x} \left( \Gamma \frac{\partial \phi}{\partial x} \right) + \frac{\partial}{\partial y} \left( \Gamma \frac{\partial \phi}{\partial y} \right) + S = 0 \quad (15.31)$$

Integration over the control volume gives

$$\int_{\Delta V} \frac{\partial}{\partial x} \left( \Gamma \frac{\partial \phi}{\partial x} \right) dx dy + \int_{\Delta V} \frac{\partial}{\partial y} \left( \Gamma \frac{\partial \phi}{\partial y} \right) dx dy + \int_{\Delta V} S dV = 0$$

and by following the same technique as in the 1D case with  $A_e = A_w = \Delta y$  and  $A_n = A_s = \Delta x$ , we get

$$\left[ \Gamma_e A_e \left( \frac{\partial \phi}{\partial x} \right)_e - \Gamma_w A_w \left( \frac{\partial \phi}{\partial x} \right)_w \right] + \left[ \Gamma_n A_n \left( \frac{\partial \phi}{\partial y} \right)_n - \Gamma_s A_s \left( \frac{\partial \phi}{\partial y} \right)_s \right] + \bar{S} \Delta V = 0 \quad (15.32)$$

By substituting

$$\begin{aligned} \Gamma_e A_e \left( \frac{\partial \phi}{\partial x} \right)_e &= \Gamma_e A_e \frac{\phi_E - \phi_P}{\delta x_{PE}} \\ \Gamma_w A_w \left( \frac{\partial \phi}{\partial x} \right)_w &= \Gamma_w A_w \frac{\phi_P - \phi_W}{\delta x_{WP}} \\ \Gamma_n A_n \left( \frac{\partial \phi}{\partial y} \right)_n &= \Gamma_n A_n \frac{\phi_N - \phi_P}{\delta y_{PN}} \\ \Gamma_s A_s \left( \frac{\partial \phi}{\partial y} \right)_s &= \Gamma_s A_s \frac{\phi_P - \phi_S}{\delta y_{SP}} \end{aligned}$$

into (15.32) and rearranging, we get

$$\underbrace{\left( \frac{\Gamma_e A_e}{\delta x_{PE}} + \frac{\Gamma_e A_w}{\delta x_{WP}} + \frac{\Gamma_s A_s}{\delta y_{SP}} + \frac{\Gamma_n A_n}{\delta y_{PN}} - S_p \right)}_{a_P} \phi_P = \underbrace{\left( \frac{\Gamma_w A_w}{\delta x_{WP}} \right)}_{a_W} \phi_W + \underbrace{\left( \frac{\Gamma_e A_e}{\delta x_{PE}} \right)}_{a_E} \phi_E + \underbrace{\left( \frac{\Gamma_s A_s}{\delta y_{SP}} \right)}_{a_S} \phi_S + \underbrace{\left( \frac{\Gamma_n A_n}{\delta y_{PN}} \right)}_{a_N} \phi_N + S_u$$

or

$$a_P \phi_P = a_W \phi_W + a_E \phi_E + a_S \phi_S + a_N \phi_N + S_u \quad (15.33)$$

which is the discretized version of (15.31).

#### Finite volume method for 3D steady state diffusion

In three dimensions, (15.22) is given by

$$\frac{\partial}{\partial x} \left( \Gamma \frac{\partial \phi}{\partial x} \right) + \frac{\partial}{\partial y} \left( \Gamma \frac{\partial \phi}{\partial y} \right) + \frac{\partial}{\partial z} \left( \Gamma \frac{\partial \phi}{\partial z} \right) + S = 0$$

which using the same technique as in the 2D case can be discretized as

$$a_P \phi_P = a_W \phi_W + a_E \phi_E + a_S \phi_S + a_N \phi_N + a_B \phi_B + a_T \phi_T + S_u$$

where

$a_W$	$a_E$	$a_S$	$a_N$	$a_B$	$a_T$	$a_P$
$\frac{\Gamma_w A_w}{\delta x_{WP}}$	$\frac{\Gamma_e A_e}{\delta x_{PE}}$	$\frac{\Gamma_s A_s}{\delta y_{SP}}$	$\frac{\Gamma_n A_n}{\delta y_{PN}}$	$\frac{\Gamma_b A_b}{\delta z_{BP}}$	$\frac{\Gamma_t A_t}{\delta z_{PT}}$	$a_W + a_E + a_S + a_N + a_B + a_T - S_p$

## 15.5 Solution of equations

In order to solve the problem, discretized equations like (15.30) have to be set up at each of the nodal points, and at the boundaries, the boundary conditions are incorporated. This will result in a system of ordinary linear algebraic equations which can be solved by a number of algorithms.

### 15.5.1 Worked example on stationary diffusion

This section is based on Example 4.2 in (Versteeg and Malalasekera 1995). Consider a large (so large that temperature gradients are only significant in the  $x$  direction) plate of thickness  $L = 2\text{cm}$  with thermal conductivity of  $k = 0.5 \text{ K}/(\text{Wm})$  and uniform heat generation of  $q = 1000 \text{ kW/m}^3$ . The boundary conditions are  $T_A = 100^\circ \text{ C}$  and  $T_B = 200^\circ \text{ C}$ . The temperature distribution is governed by

$$\frac{d}{dx} \left( k \frac{dT}{dx} \right) + q = 0$$

We divide the domain into five control volumes, that is  $\delta x = 0.004 \text{ m}$ , and consider a unit area in the  $yz$  plane. Integration over a control volume (using  $\bar{S}\Delta V = q\Delta V$ ) gives

$$\begin{aligned} \int_{\Delta V} \frac{d}{dx} \left( k \frac{dT}{dx} \right) dV + \int_{\Delta V} q dV &= 0 \\ \left( kA \frac{dT}{dx} \right)_e - \left( kA \frac{dT}{dx} \right)_w + q\Delta V &= 0 \\ k_e A \frac{T_E - T_P}{\delta x} - k_w A \frac{T_P - T_W}{\delta x} + qA\delta x &= 0 \end{aligned}$$

which is rearranged to

$$\left( \frac{k_e A}{\delta x} + \frac{k_w A}{\delta x} \right) T_P = \frac{k_e A}{\delta x} T_W + \frac{k_w A}{\delta x} T_E + qA\delta x \quad (15.34)$$

which, when comparing to (15.30), has the following coefficients:

$a_W$	$a_E$	$a_P$	$S_p$	$S_u$
$\frac{kA}{\delta x}$	$\frac{kA}{\delta x}$	$a_W + a_E - S_p$	0	$qA\delta x$

Equation (15.34) have to be somewhat modified at the boundaries in order to include the boundary conditions. A linear approximation is used for temperatures between the boundary point and the nodal point. So for node 1, we have

$$k_e A \frac{T_E - T_P}{\delta x} - k_A A \frac{T_P - T_A}{\delta x/2} + qA\delta x = 0$$

which have, using  $k_A = k$ , the following table of coefficients:

$a_W$	$a_E$	$a_P$	$S_p$	$S_u$
0	$\frac{kA}{\delta x}$	$a_W + a_E - S_p$	$-\frac{2kA}{\delta x}$	$qA\delta x + \frac{2kA}{\delta x} T_A$

Similar for node 5, we get

$a_W$	$a_E$	$a_P$	$S_p$	$S_u$
$\frac{kA}{\delta x}$	0	$a_W + a_E - S_p$	$-\frac{2kA}{\delta x}$	$qA\delta x + \frac{2kA}{\delta x} T_B$

Using numerical values, we get the following coefficients for all the nodes:

Node	$a_W$	$a_E$	$S_u$	$S_p$	$a_P$
1	0	125	$4000 + 250T_A$	-250	375
2	125	125	4000	0	250
3	125	125	4000	0	250
4	125	125	4000	0	250
5	125	0	$4000 + 250T_B$	-250	375

This set of equations can be written in matrix form:

$$\begin{pmatrix} 375 & -125 & 0 & 0 & 0 \\ -125 & 250 & -125 & 0 & 0 \\ 0 & -125 & 250 & -125 & 0 \\ 0 & 0 & -125 & 250 & -125 \\ 0 & 0 & 0 & -125 & 375 \end{pmatrix} \begin{pmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5 \end{pmatrix} = \begin{pmatrix} 29000 \\ 4000 \\ 4000 \\ 4000 \\ 54000 \end{pmatrix}$$

The matrix of coefficients is seen to be tridiagonal, which is generally the case in problems like this. The system can be found to have the solution

$$\begin{pmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5 \end{pmatrix} = \begin{pmatrix} 150 \\ 218 \\ 254 \\ 258 \\ 230 \end{pmatrix}$$

which can be found by using an appropriate technique for solving systems of ordinary algebraic equations. Particularly well suited is the TDMA (tridiagonal matrix algorithm) or Thomas algorithm, which can be applied iteratively.

## 15.6 Stability issues

Before we proceed to study more complicated flow phenomena, we need some concepts regarding stability of the numerical solutions. Convergence is the property of a numerical method to produce a solution which approaches the exact solution of the original problem as the grid spacing tends to zero. Consistent numerical schemes produce systems of algebraic equations which can be demonstrated to be equivalent to the original governing equation as the grid spacing tends to zero. Stability means that all solutions of the numerical scheme are uniformly bounded functions of the initial conditions for sufficiently small grid spacing. Convergence is usually very difficult to establish theoretically. However, for *linear* problems, convergence can be established using the following theorem.

**Theorem 2 (Lax's equivalence theorem)** *If a differential approximation to a initial and boundary value problem is consistent, stability is a necessary and sufficient condition for the scheme to be convergent.*

For CFD problems this is of limited value as the governing equations often are nonlinear, and in that case stability and consistency are necessary, but not sufficient conditions for convergence. (Patankar 1980) presented rules which yields robust finite volume calculations schemes. Robust schemes has three properties: conservativeness, boundedness

and transportiveness. As we shall see, these concepts are designed into finite volume schemes, and they are according to (Versteeg and Malalasekera 1995) commonly accepted as alternatives for the more mathematically rigorous concepts of convergence, consistency and stability. Stability will be studied in some more detail in section 15.10.

- **Conservativeness:** Any discretization scheme for diffusion convection problems must ensure that the flux through a common face is represented in a consistent manner in adjacent control volumes. Ideally, when calculating the overall flux balance, the fluxes between the control volumes should cancel out, leaving only the fluxes at the boundaries,
- **Boundedness:** The discretized equations at each node give rise to a set of algebraic equations that need to be solved. This set might be quite large, and iterative solution methods are often used. The Scarborough criterion (Scarborough 1966) states that a sufficient condition for a convergent iterative method can be expressed in terms of the values of the coefficients of the discretized equations. The criterion is given by

$$\frac{\sum |a_{nb}|}{|a'_P|} \begin{cases} \leq 1 & \text{at all nodes} \\ < 1 & \text{at one node at last} \end{cases} \quad (15.35)$$

where  $a'_P$  is the net coefficient for the central node  $P$ , that is

$$a'_P = a_P - S_P$$

and  $\sum |a_{nb}|$  is the net coefficient for all the neighboring nodes. Another criterion for boundedness is that *all coefficients of the discretized equation have the same sign*. Physically, this implies that the increase of  $\phi$  in one node should result in an increase of  $\phi$  in the neighboring nodes.

- **Transportiveness:** Define the *Peclet number* as

$$Pe = \frac{F}{D} = \frac{\rho u}{\Gamma/\delta x}$$

The Peclet number  $Pe$  is a measure of the relative strength of diffusion and convection, and the terms  $F$  and  $D$  will be further treated in Section 15.8. For pure diffusion we have  $Pe = 0$ , and for pure convection we have  $Pe = \infty$ .

## 15.7 Finite volume method for diffusion dynamics

We now include a rate of change term in the governing equation in order to study time dependent problems. Unsteady diffusion will be studied by looking at heat conduction, which in one dimension, is described by

$$\rho c \frac{\partial T}{\partial t} = \frac{\partial}{\partial x} \left( k \frac{\partial T}{\partial x} \right) + S \quad (15.36)$$

where  $\rho$  is the density (assumed constant),  $c$  is the specific heat of the material and  $k$  is the thermal conductivity. As seen in Section 15.3.2 this problem is parabolic in time, and we will solve it by marching in time from an initial distribution of temperature  $T$ .

Given the nodal values of  $T$  at time  $t$  (denoted  $T_P^0, T_W^0, T_E^0$  and so on), we will find the values of  $T$  at time  $t + \Delta t$  (denoted  $T_P^1, T_W^1, T_E^1$  and so on). Integrating over the control volume using the same techniques as before, we get

$$\int_{\Delta V} \rho c \frac{\partial T}{\partial t} dV = \int_{\Delta V} \frac{\partial}{\partial x} \left( k \frac{\partial T}{\partial x} \right) dV + \int_{\Delta V} S dV \quad (15.37)$$

$$\int_w^e \rho c \frac{\partial T}{\partial t} dV = \left( kA \frac{\partial T}{\partial x} \right)_e - \left( kA \frac{\partial T}{\partial x} \right)_w + \bar{S} \Delta V \quad (15.38)$$

Assuming that the temperature is uniform within each control volume, we have

$$\rho c \frac{\partial T}{\partial t} \Delta V = \left( kA \frac{\partial T}{\partial x} \right)_e - \left( kA \frac{\partial T}{\partial x} \right)_w + \bar{S} \Delta V$$

Using the theta method described in Section 14.5.6, we get.

$$\rho c \Delta V T_P^{n+1} = \rho c \Delta V T_P^n + \Delta t [ \theta f(T_P^n, t_n, x) + (1 - \theta) f(T_P^{n+1}, t_{n+1}, x) ] \quad (15.39)$$

where

$$f(T_P^n, t_n, x) = k_e \frac{T_E^n - T_P^n}{\delta x_{PE}} - k_w \frac{T_P^n - T_W^n}{\delta x_{WP}} + \bar{S} \Delta V$$

and

$$f(T_P^{n+1}, t_{n+1}, x) = k_e \frac{T_E^{n+1} - T_P^{n+1}}{\delta x_{PE}} - k_w \frac{T_P^{n+1} - T_W^{n+1}}{\delta x_{WP}} + \bar{S} \Delta V$$

As before, central differencing has been applied to the diffusion terms.

In (15.39),  $\theta = 1$  gives the Euler method,  $\theta = \frac{1}{2}$  gives the trapezoidal rule (usually referred to as the Crank-Nicolson method in the CFD literature) and  $\theta = 0$  gives the implicit Euler method. Using  $\Delta V = A \Delta x$ , equation (15.39) can also be written as

$$\begin{aligned} \left( \rho c \frac{\Delta x}{\Delta t} + (1 - \theta) \left( \frac{k_e}{\delta x_{PE}} + \frac{k_w}{\delta x_{WP}} \right) \right) T_P^{n+1} &= \frac{k_e}{\delta x_{PE}} ((1 - \theta) T_E^{n+1} + \theta T_E^n) \\ &\quad + \frac{k_w}{\delta x_{WP}} ((1 - \theta) T_W^{n+1} + \theta T_W^n) \\ &\quad + \left( \rho c \frac{\Delta x}{\Delta t} - \theta \left( \frac{k_e}{\delta x_{PE}} + \frac{k_w}{\delta x_{WP}} \right) \right) T_P^n \\ &\quad + \bar{S} \Delta x \end{aligned}$$

or

$$\begin{aligned} a_P T_P^{n+1} &= a_W ((1 - \theta) T_W^{n+1} + \theta T_W^n) + a_E ((1 - \theta) T_E^{n+1} + \theta T_E^n) \quad (15.40) \\ &\quad + (a_P^0 - \theta a_W - \theta a_E) T_P^n + \bar{S} \Delta x \end{aligned}$$

where

$$a_P = (1 - \theta) (a_W + a_E) + a_P^0$$

$$a_P^0 = \rho c \frac{\Delta x}{\Delta t}$$

$$a_W = \frac{k_w}{\delta x_{WP}}$$

$$a_E = \frac{k_e}{\delta x_{PE}}$$

We have now arrived at a discretized version of (15.36), and we will review three different schemes given by the theta method.

### Euler ( $\theta = 1$ )

In the Euler method, or explicit scheme,  $\theta = 1$ , and the source term is linearized as

$$\bar{S}\Delta x = S_u + S_P T_P^n$$

The discretized heat transfer equation is then given by

$$a_P T_P^{n+1} = a_W T_W^n + a_E T_E^n + (a_P^0 - a_W - a_E + S_P) T_P^n + S_u \quad (15.41)$$

where

$$\begin{aligned} a_P &= a_P^0 \\ a_P^0 &= \rho c \frac{\Delta x}{\Delta t} \\ a_W &= \frac{k_w}{\delta x_{WP}} \\ a_E &= \frac{k_e}{\delta x_{PE}} \end{aligned} \quad (15.42)$$

For this method to be stable, all coefficients in the discretized equation (15.41) need to be positive. This is satisfied if  $a_P^0 - a_W - a_E > 0$ . When assuming uniform grid spacing  $\delta x_{PE} = \delta x_{WP} = \Delta x$  and constant  $k = k_e = k_w$ , this can be written

$$\Delta t < \rho c \frac{(\Delta x)^2}{2k} \quad (15.43)$$

This sets a maximum limit on the time step size  $\Delta t$ , and reduction of  $\Delta x$  to improve spatial accuracy forces us to chose a much smaller time step in order to ensure stability.

### Crank-Nicolson ( $\theta = \frac{1}{2}$ )

In the Crank-Nicolson method,  $\theta = \frac{1}{2}$  and the source term is linearized as

$$\bar{S}\Delta x = S_u + S_P \frac{T_P^n + T_P^{n+1}}{2}$$

The discretized heat transfer equation is

$$a_P T_P^{n+1} = a_W \frac{T_W^n + T_W^{n+1}}{2} + a_E \frac{T_E^n + T_E^{n+1}}{2} + \left( a_P^0 - \frac{a_W}{2} - \frac{a_E}{2} + \frac{S_P}{2} \right) T_P^n + S_u$$

where

$$\begin{aligned} a_P &= \frac{1}{2}(a_W + a_E) + a_P^0 - \frac{S_P}{2} \\ a_P^0 &= \rho c \frac{\Delta x}{\Delta t} \\ a_W &= \frac{k_w}{\delta x_{WP}} \\ a_E &= \frac{k_e}{\delta x_{PE}} \end{aligned}$$

Again, all coefficients in the discretized equation must be positive, which is satisfied if

$$\Delta t < \rho c \frac{(\Delta x)^2}{k}$$

where  $\delta x_{PE} = \delta x_{WP} = \Delta x$  and constant  $k = k_e = k_w$ . Again, we notice the restriction on the time step  $\Delta t$ , which is only slightly less restrictive than for the Euler scheme.

### Implicit Euler ( $\theta = 0$ )

In the implicit Euler scheme,  $\theta = 0$  and the source term is linearized as

$$\bar{S}\Delta x = S_u + S_P T_P^{n+1}$$

and the discretized heat transfer equation is

$$a_P T_P^{n+1} = a_W T_W^{n+1} + a_E T_E^{n+1} + a_P^0 T_P^n + S_u$$

where

$$\begin{aligned} a_P &= a_W + a_E + a_P^0 - S_P \\ a_P^0 &= \rho c \frac{\Delta x}{\Delta t} \\ a_W &= \frac{k_w}{\delta x_{WP}} \\ a_E &= \frac{k_e}{\delta x_{PE}} \end{aligned}$$

As all coefficients in the discretized equation are positive, the method is unconditionally stable. CFD computations for two- and three-dimensional transient diffusion can be done using the same methods as in this section.

### Example 240 Simulation of temperature control using the Finite Volume CFD method

Consider an insulated rod equipped with a heating element at the east end as depicted in Figure 15.8. The one dimensional transient heat conduction equation is

$$\rho c \frac{\partial T}{\partial t} = \frac{\partial}{\partial x} \left( k \frac{\partial T}{\partial x} \right) + S$$

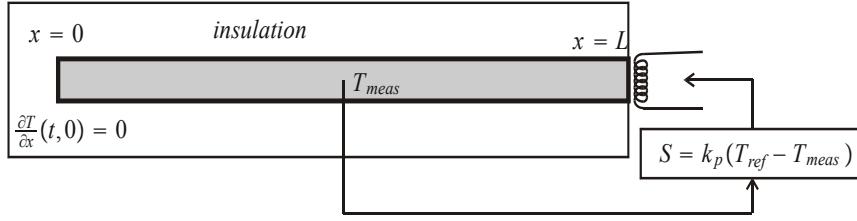


Figure 15.8: Temperature control of a rod. A heating element at the east end is controlled with feedback from the temperature at the middle of the rod. The rest of the rod is insulated.

where  $\rho c = 10^7 \text{ JK/m}^3$ ,  $k = 10 \text{ WK/m}$ , and length  $L = 0.02 \text{ m}$ . Initially  $T(x, 0) = 300 \text{ K}$ . The insulated west end implies the boundary condition  $\frac{\partial T}{\partial x}(t, 0) = 0$ . We will use a P controller to control the power  $S$  of the heating element which is located at the east end ( $x = L$ ). Temperature will be measured at the middle of the rod so that the source term will be given by

$$S = k_p(T_{ref} - T_{meas})$$

where  $k_p > 0$  is the controller gain  $T_{ref}$  is the temperature set point and  $T_{meas}$  is the temperature measurement. We use the Euler method for time discretization and divide the spatial domain into five control volumes such that  $\Delta x = 0.004 \text{ m}$ . The computational grid will be similar to the one used in Section 15.5.1. By using equation (15.40), we find for node 1

$$\rho c \left( \frac{T_P^{n+1} - T_P^n}{\Delta t} \right) \Delta x = \frac{k}{\Delta x} (T_E^n - T_P^n)$$

The discretized equations for node 2,3 and 4 are given by (15.41) and (15.42), and for node 5 we find

$$\rho c \left( \frac{T_P^{n+1} - T_P^n}{\Delta t} \right) \Delta x = k_p (T_{ref} - T_3^n) - \frac{k}{\Delta x} (T_P^n - T_W^n)$$

where  $T_{meas} = T_3$  has been used as the measurement is made at the middle. The time step  $\Delta t$  must satisfy (15.43), which leads to  $\Delta t < 8$ . We chose  $\Delta t = 2$ . Using numerical values, it can be shown that the equation for each node is given by

$$\begin{aligned} \text{Node 1: } & 200T_1^{n+1} = 25T_2^n + 175T_1^n \\ \text{Node 2-4: } & 200T_P^{n+1} = 25T_W^n + 25T_E^n + 150T_P^n \\ \text{Node 5: } & 200T_5^{n+1} = 25T_4^n + 175T_5^n + k'_p (T_{ref} - T_3) \end{aligned} \quad (15.44)$$

Notice that due to measurement and control not being collocated, the introduction of  $T_3$  in the equation for node 5 breaks the tridiagonal structure of the equations. The equation set (15.44) was solved numerically using MATLAB. The result is plotted in Figure 15.9. As can be seen, the desired temperature is reached throughout the rod, which is not surprising due to the insulation.

The equation set in the example was evaluated with the following MATLAB script.

```
kp=50; % Controller gain
T_ref=400; % Desired temperature
```

```

for i=1:5, % Initialization
    T(1,i)=300;
end
for t=2:150; % 150 timesteps
    T(t,1)=1/200*(25*T(t-1,2)+175*T(t-1,1)); %node 1
    T(t,2)=1/200*(25*T(t-1,1)+150*T(t-1,2)+25*T(t-1,3)); %node 2
    T(t,3)=1/200*(25*T(t-1,2)+150*T(t-1,3)+25*T(t-1,4)); %node 3
    T(t,4)=1/200*(25*T(t-1,3)+150*T(t-1,4)+25*T(t-1,5)); %node 4
    T(t,5)=1/200*(25*T(t-1,4)+175*T(t-1,5)+kp*(T_ref-T(t-1,3))); %node 5
end
figure(1)
surf(T')

```

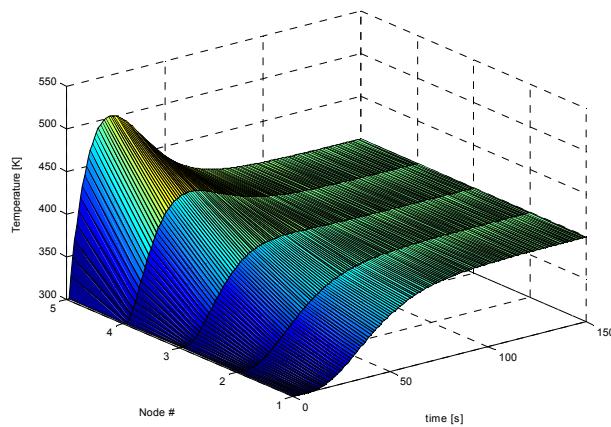


Figure 15.9: Simulation of the equation set (15.44).

## 15.8 Finite volumes for Convection-Diffusion

### 15.8.1 Introduction

Convection is caused by fluid flow. In this section we will study methods to obtain a solution for  $\phi$  for a given flow field. The transport equation for convection-diffusion of a general property  $\phi$  is given by (15.1), and repeated here for convenience:

$$\underbrace{\frac{\partial(\rho\phi)}{\partial t}}_{\text{rate of change of } \phi \text{ of fluid element}} + \underbrace{\nabla^T (\rho \mathbf{u} \phi)}_{\text{Net rate of flow of } \phi \text{ out of fluid element}} = \underbrace{\nabla^T (\Gamma \nabla \phi)}_{\text{rate of change of } \phi \text{ due to diffusion}} + \underbrace{S_\phi}_{\text{rate of change of } \phi \text{ due to sources}} \quad (15.45)$$

where  $\rho$  is the density,  $\mathbf{u}$  is the flow velocity vector and  $\Gamma$  is the diffusion coefficient.

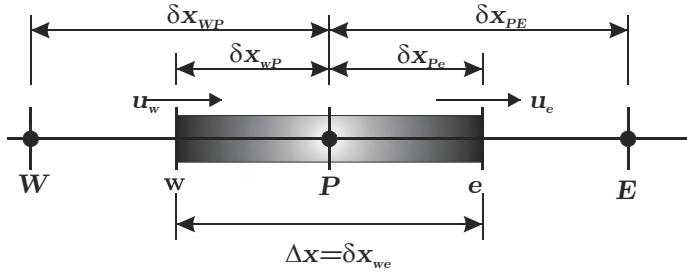


Figure 15.10: Control volume for 1D convection and diffusion.  $u_e$  is the flow velocity entering the CV, and  $u_w$  is the flow velocity leaving the CV.

### 15.8.2 Finite volume method for 1D diffusion and convection dynamics

In one dimension, (15.45) is given by

$$\frac{\partial}{\partial t} (\rho\phi) + \frac{\partial}{\partial x} (\rho u\phi) = \frac{\partial}{\partial x} \left( \Gamma \frac{\partial \phi}{\partial x} \right) + S, \quad (15.46)$$

where  $u$  is the  $x$  component of the flow velocity vector. Using the same methodology as in Section 15.7, integration of (15.46) over the control volume in Figure 15.10 gives

$$\int_{\Delta V} \frac{\partial}{\partial t} (\rho\phi) dV + (\rho Au\phi)_e - (\rho Au\phi)_w = \left( \Gamma A \frac{d\phi}{dx} \right)_e - \left( \Gamma A \frac{d\phi}{dx} \right)_w \quad (15.47)$$

$$\rho \frac{\partial \phi}{\partial t} \Delta V + (\rho Au\phi)_e - (\rho Au\phi)_w = \left( \Gamma A \frac{d\phi}{dx} \right)_e - \left( \Gamma A \frac{d\phi}{dx} \right)_w \quad (15.48)$$

We now define the convection and diffusion terms

$$\begin{aligned} F &= \rho u \\ D &= \frac{\Gamma}{\delta x} \end{aligned}$$

so that

$$\begin{aligned} F_w &= (\rho u)_w, F_e = (\rho u)_e \\ D_w &= \frac{\Gamma_w}{\delta x_{WP}}, D = \frac{\Gamma_e}{\delta x_{PE}} \end{aligned}$$

We choose here to use the implicit Euler scheme of Section 15.7 for integration, but other schemes might just as well be used. Linearizing the source term as

$$\bar{S}\Delta V = S_u + S_P\phi_P^{n+1}$$

the integrated version of (15.47) can be written

$$\begin{aligned} \rho\Delta V\phi_P^{n+1} &= \rho\Delta V\phi_P^n + \Delta t \left[ -F_e\phi_e^{n+1} + F_w\phi_w^{n+1} + D_e(\phi_E^{n+1} - \phi_P^{n+1}) \right. \\ &\quad \left. - D_w(\phi_P^{n+1} - \phi_W^{n+1}) + S_u + S_P\phi_P^{n+1} \right] \end{aligned} \quad (15.49)$$

In (15.49), the diffusion terms are approximated by using the central difference. Calculation of the convection terms at the  $e$  and  $w$  surfaces can be done by a number of different schemes, and some of them will now be reviewed.

### Central difference

One alternative is to use the central difference to approximate the convection terms. By substituting

$$\begin{aligned}\phi_e &= \frac{\phi_P + \phi_E}{2} \\ \phi_w &= \frac{\phi_W + \phi_P}{2}\end{aligned}$$

into (15.49) and rearranging, we get

$$\begin{aligned}&\left( \frac{\rho \Delta V}{\Delta t} + \left( D_w + \frac{F_w}{2} \right) + \left( D_e - \frac{F_e}{2} \right) + (F_e - F_w) - S_P \right) \phi_P^{n+1} \\ &= \frac{\rho \Delta V}{\Delta t} \phi_P^n + \left( D_w + \frac{F_w}{2} \right) \phi_W^{n+1} + \left( D_e - \frac{F_e}{2} \right) \phi_E^{n+1} + S_u\end{aligned}$$

or

$$a_P \phi_P^{n+1} = a_P^0 \phi_P^n + a_W \phi_W^{n+1} + a_E \phi_E^{n+1} + S_u$$

and we have the same type of discretized equation that we derived for the pure diffusion problems. Regarding conservativeness, it can easily be shown that fluxes of the type  $\Gamma \frac{(\phi_{i+1} - \phi_i)}{\Delta x}$  will cancel out in pairs, leaving only the fluxes at the boundaries, and thus the central differencing scheme is conservative. For a flow that is governed by (15.49) and simultaneously satisfies continuity, that is

$$F_e - F_w = 0$$

it follows that,  $a'_P = a_W + a_E + a_P^0$ . The Scarborough criterion can then be calculated as

$$\frac{\sum |a_{nb}|}{|a'_P|} = \frac{|a_W| + |a_E| + |a_P^0|}{|a_W + a_E + a_P^0|}$$

and it can be seen that the conditions of (15.35) are satisfied. As  $a_E = D_e - \frac{F_e}{2}$ , this coefficient can be negative if convection is sufficiently strong. Provided all other coefficients are positive, this will violate the requirement for boundedness. Moreover

$$a_E > 0 \implies D_e - \frac{F_e}{2} > 0 \implies \frac{F_e}{D_e} = Pe_e < 2$$

This means that the central differencing scheme in this case will produce bounded solutions only if the Peclet number satisfies  $Pe_e < 2$ . Regarding transportiveness, the scheme does not recognize the direction of the flow, and would not be useful for high  $Pe$ . The central difference scheme is accurate to second-order.

### The upwind scheme

To remedy this, the upwind scheme takes flow direction into account. If the flow direction is positive (from west in Figure 15.10), that is  $u_w > 0$  and  $u_e > 0$ , we chose

$$\phi_w = \phi_W \text{ and } \phi_e = \phi_P.$$

When inserted into (15.49) this gives the discretized equation

$$\begin{aligned} & \left( \frac{\rho \Delta V}{\Delta t} + (D_w + F_w) + D_e + (F_e - F_w) - S_P \right) \phi_P^{n+1} \\ &= \frac{\rho \Delta V}{\Delta t} \phi_P^n + (D_w - F_w) \phi_W^{n+1} + D_e \phi_E^{n+1} + S_u. \end{aligned} \quad (15.50a)$$

If the flow direction is negative (from east in Figure 15.10), that is  $u_w < 0$  and  $u_e < 0$ , we chose

$$\phi_w = \phi_P \text{ and } \phi_e = \phi_E.$$

and consequently

$$\begin{aligned} & \left( \frac{\rho \Delta V}{\Delta t} + D_w + (D_e - F_e) + (F_e - F_w) - S_P \right) \phi_P^{n+1} \\ &= \frac{\rho \Delta V}{\Delta t} \phi_P^n + D_w \phi_W^{n+1} + (D_e - F_e) \phi_E^{n+1} + S_u \end{aligned} \quad (15.51)$$

By combining (15.50a) and (15.51), the upwind scheme can be written

$$a_P \phi_P^{n+1} = a_P^0 \phi_P^n + a_W \phi_W^{n+1} + a_E \phi_E^{n+1}$$

where

$$\begin{aligned} a_P &= a_P^0 + a_W + a_E + (F_e - F_w) \\ a_P^0 &= \frac{\rho \Delta V}{\Delta t} \\ a_W &= D_w + \max(F_w, 0) \\ a_E &= D_e + \max(0, -F_e) \end{aligned}$$

The upwind scheme is conservative, the fact that  $a_P = a_W + a_E$  implies that the Scarborough criterion is met. Also, all the coefficient are positive, and transportiveness is built into the formulation. The upwind scheme is accurate to first order.

### The hybrid scheme

The hybrid differencing is based on a combination of central and upwind differencing schemes. For small Peclet numbers ( $Pe < 2$ ), central differencing is used while the upwind scheme is used for large Peclet numbers. The reason for this is to make use of the second order accuracy of the central differencing scheme. It can be shown that the hybrid scheme can be written

$$a_P \phi_P^{n+1} = a_P^0 \phi_P^n + a_W \phi_W^{n+1} + a_E \phi_E^{n+1}$$

with

$$\begin{aligned} a_P &= a_P^0 + a_W + a_E + (F_e - F_w) \\ a_P^0 &= \frac{\rho \Delta V}{\Delta t} \\ a_w &= \max \left( F_w, \left( D_w + \frac{F_w}{2} \right), 0 \right) \\ a_w &= \max \left( -F_e, \left( D_e - \frac{F_e}{2} \right), 0 \right) \end{aligned}$$

### Quadratic upwind scheme (QUICK)

The quadratic upstream interpolation for convective kinetics (QUICK) scheme (Leonard 1979) uses a three-point upstream-weighted quadratic interpolation for cell face values. The value of  $\phi$  at the cell face between two bracketing nodes  $i$  and  $i - 1$  and an upstream node  $n - 2$  is given by

$$\phi_{face} = \frac{6}{8}\phi_{i-1} + \frac{3}{8}\phi_i - \frac{1}{8}\phi_{i-2}$$

When  $u_w > 0$  and  $u_e > 0$ , a quadratic fit through  $WW$ ,  $W$ , and  $P$  is used to evaluate  $\phi_w$  and a quadratic fit through  $W$ ,  $P$ , and  $E$  is used to evaluate  $\phi_e$ . If  $u_w < 0$  and  $u_e < 0$ , values at  $W$ ,  $P$ , and  $E$  is used to evaluate  $\phi_w$  and  $P$ ,  $E$ , and  $EE$  is used to evaluate  $\phi_e$ . The diffusion terms are evaluated using central differencing. For one-dimensional convection diffusion, the QUICK-scheme can be summarized as

$$a_P\phi_P^{n+1} = a_P^0\phi_P^n + a_W\phi_W^{n+1} + a_E\phi_E^{n+1} + a_{WW}\phi_{WW}^{n+1} + a_{EE}\phi_{EE}^{n+1}$$

where

$$\begin{aligned} a_P &= a_P^0 + a_W + a_E + (F_e - F_w) + a_{WW} + a_{EE} \\ a_W &= D_w + \frac{6}{8}\alpha_w F_w + \frac{1}{8}\alpha_e F_e + \frac{3}{8}(1 - \alpha_w) F_w \\ a_{WW} &= -\frac{1}{8}\alpha_w F_w \\ a_E &= D_e - \frac{3}{8}\alpha_e F_e - \frac{6}{8}(1 - \alpha_e) F_e - \frac{1}{8}(1 - \alpha_w) F_w \\ a_{EE} &= \frac{1}{8}(1 - \alpha_e) F_e \end{aligned}$$

where  $\alpha_w = 1$  for  $F_w > 0$ ,  $\alpha_w = 0$  for  $F_w < 0$ ,  $\alpha_e = 1$  for  $F_e > 0$  and  $\alpha_e = 1$  for  $F_w < 0$ . The QUICK algorithm is accurate to third order. Notice that the QUICK algorithm will yield a penta-diagonal system of equations.

## 15.9 Pressure-velocity coupling

### 15.9.1 Introduction

Viscous incompressible flow is governed by the incompressible Navier-Stokes equations. These equations exhibit a mixed elliptic-parabolic behavioral, and therefore, as discussed in Section 15.3, the problem can not be solved by time marching techniques. The velocity components are governed by the momentum equations, which are particular cases of the general differential equation for  $\phi$ . The momentum equations, e.g. in 2D, are given by

$$\frac{\partial(\rho u)}{\partial t} + \frac{\partial}{\partial x}(\rho u^2) + \frac{\partial}{\partial y}(\rho uv) = \frac{\partial}{\partial x}\left(\mu \frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\mu \frac{\partial u}{\partial y}\right) - \frac{\partial p}{\partial x} + S_u \quad (15.52)$$

$$\frac{\partial(\rho v)}{\partial t} + \frac{\partial}{\partial x}(\rho uv) + \frac{\partial}{\partial y}(\rho v^2) = \frac{\partial}{\partial x}\left(\mu \frac{\partial v}{\partial x}\right) + \frac{\partial}{\partial y}\left(\mu \frac{\partial v}{\partial y}\right) - \frac{\partial p}{\partial y} + S_v \quad (15.53)$$

These equations can be shown to be parabolic. The velocity also satisfy the continuity equation

$$\frac{\partial}{\partial x}(\rho u) + \frac{\partial}{\partial y}(\rho v) = 0 \quad (15.54)$$

which is elliptic. The incompressible Navier-Stokes equations can be obtained from the compressible form simply by setting constant density  $\rho$ , which would lead us to believe that the equations can be solved numerically by using one of the previous mentioned techniques. This is unfortunately not the case. (Anderson 1995) cites a result where an approximate stability criterion for an explicit Navier-Stokes solution is shown to be

$$\Delta t \leq \frac{1}{|u|/\Delta x + |v|/\Delta y + a\sqrt{1/(\Delta x)^2 + 1/(\Delta y)^2}}$$

where  $\alpha$  is the sonic velocity. For a compressible flow  $a$  is finite, and it is possible to find a finite  $\Delta t$  to guarantee stability of the solution. However, for an incompressible flow, the sonic velocity is theoretically infinite, and we would get  $\Delta t = 0$ . Consequently, another method has to be found. In addition to the equations (15.52), (15.53) and (15.54) being nonlinear, another difficulty in solving this problem is the presence of the pressure gradient terms  $\frac{\partial p}{\partial x}$  and  $\frac{\partial p}{\partial y}$ . As can be seen we have no differential equation for  $\frac{\partial p}{\partial t}$ . We also recognize that a consequence of the system (15.52)-(15.54) being a parabolic-elliptic, we have no transient term in (15.54). It is interesting to compare this to a system of differential-algebraic equations (DAE).

**Remark 9** *If the flow is compressible, the continuity equation*

$$\frac{\partial}{\partial t}(\rho) + \frac{\partial}{\partial x}(\rho u) + \frac{\partial}{\partial y}(\rho v) = 0 \quad (15.55)$$

*may be used as a transport equation for density, and the energy equation is a transport equation for temperature. The pressure may then be obtained using the equation of state  $p = p(T, \rho)$ , and the the problem can be solved by using techniques described previously in this Chapter.*

For incompressible flow, there is by definition no connection between pressure and density. In this case, coupling between pressure and density introduces a constraint on the solution of the flow field: if the correct pressure field is applied in the momentum equations (15.52) and (15.53), the resulting velocity field should satisfy continuity (15.54). For these reasons, solution techniques for incompressible Navier-Stokes equations are different than for compressible Navier-Stokes equations. The pressure correction model, which will be presented next, are one such technique.

### 15.9.2 The staggered grid

When including the pressure gradient in the calculations, another problem arises. This is usually illustrated by an example.

**Example 241** *The, checkerboard pressure field of Figure 15.11 is highly irregular. Let us examine how this is represented in the discretized momentum equations. If the pressure drop across the CV is obtained by linear interpolation, we have in the  $x$  direction:*

$$p_w - p_e = \frac{p_W + p_P}{2} - \frac{p_P + p_E}{2} = \frac{p_W - p_E}{2}$$

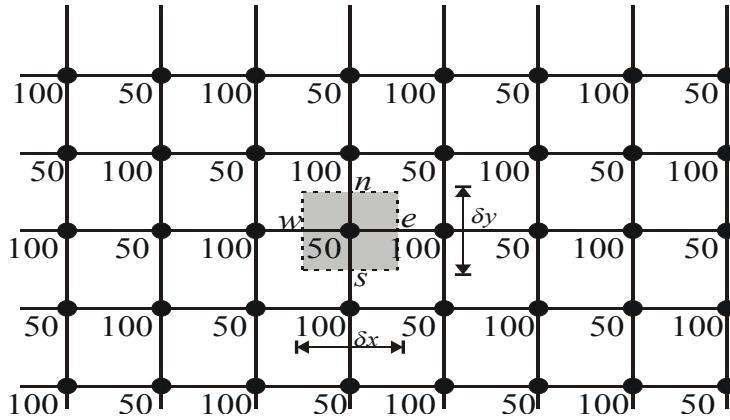


Figure 15.11: Discrete checkerboard pressure field

which implies

$$\frac{\partial p}{\partial x} = \frac{p_w - p_e}{\delta x} = \frac{p_w - p_e}{2\delta x} = \frac{100 - 100}{2\delta x} = 0 \quad (15.56)$$

and similarly in the  $y$  direction

$$\frac{\partial p}{\partial y} = \frac{p_N - p_S}{2\delta y} = \frac{100 - 100}{2\delta y} = 0 \quad (15.57)$$

We see that the value at the node  $P$  is cancelled out and not used in either of the calculations (15.56) or (15.57). This means that the pressure differences are calculated for alternate nodes, and not for adjacent ones. This might reduce the accuracy of the solution. However this is not the main problem: far more serious is the fact that a pressure field like the one in Figure 15.11 results in all the discretized gradients being zero at the nodal point. As a result, this pressure field would give the same momentum source as a uniform field.

**Remark 10** The problem described in Example 241 does not occur if the flow is compressible. The term  $\frac{\partial p}{\partial t}$  in the continuity equation (15.55) would damp out the checkerboard pattern.

It is clear from the Example 241 that if velocities are defined at the scalar grid nodes, the pressure gradients are not properly represented in the discretized momentum equations. A remedy for this is *the staggered grid*. The idea is to evaluate velocities on another grid than all the other variables. Pressure, temperature etc. are calculated at the ordinary nodal points, but velocities are calculated at the CV faces between the nodal points. A 2D staggered grid is shown in Figure 15.12. The  $x$  direction velocity  $u$  is calculated at the faces that are normal to the  $x$  direction, and the  $y$  direction velocity  $v$  is calculated at the faces that are normal to the  $y$  direction. The staggered grid arrangement solves the problem of the checkerboard pressure field. Alternatively, (Anderson 1995), suggest to use another approximation when calculating the pressure differences in (15.56) and (15.57). The upwind scheme might be one such scheme.

**Remark 11** Although example 241 illustrates that the staggered grid solves the problem of the checkerboard pressure field, it is rather unrealistic because it contains a nonphysical

pressure field. The main advantage of the staggered grid is in fact that it allows for the pressure difference between two adjacent grid points to be the driving force for the velocity component located between these grid points.

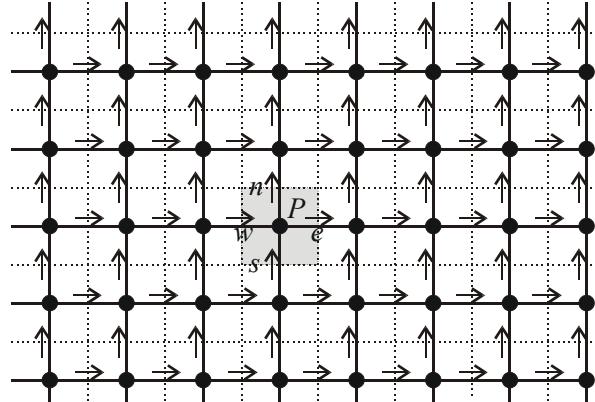


Figure 15.12: A staggered two dimensional grid. The  $u$ 's are stored at the  $\rightarrow$ , the  $v$ 's at the  $\uparrow$  and other variables  $\phi$  at the  $\bullet$ .

### 15.9.3 The momentum equations

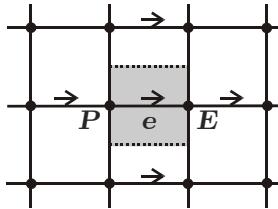


Figure 15.13: Control volume for  $u$ .

A staggered control volume for the momentum equation (15.52) is shown in Figure 15.13. As can be seen, the pressure difference  $p_P - p_E$  can be used to calculate the pressure force acting on the control volume for the velocity  $u$ . Using similar techniques as in the previous sections, the discretized transient momentum equation in the  $x$  direction on the staggered grid can be written in the form

$$a_e u_e^{n+1} = \sum a_{nb} u_{nb}^n + a_e^0 u_e^n + b + (p_P^n - p_E^n) A_e \quad (15.58)$$

where

$$A_u = \frac{\Delta V_u}{\Delta x}$$

and  $b$  is a collection of other source terms than the pressure gradient. In (15.58), integration with respect to time has been carried out by an explicit scheme such as the Euler method. The values of the coefficients  $a_e^0$ ,  $a_e$  and  $a_{nb}$  may be calculated by any of the

previous presented methods suitable for convection-diffusion problems (hybrid, upwind, QUICK).

Similarly, the discretized momentum equation in the  $y$  direction on the staggered grid can be written in the form

$$a_n v_n^{n+1} = \sum a_{nb} v_{nb}^n + a_n^0 v_n^n + b + (p_P^n - p_N^n) A_n \quad (15.59)$$

where

$$A_n = \frac{\Delta V_v}{\Delta y}$$

The extension to the  $z$  direction in a 3D problem is trivial.

**Remark 12** Notice that the term  $(p_P - p_E) A_u$  in (15.58) is the pressure force acting on the  $u$  control volume. This is physically correct and would not be possible without the staggered grid.

#### 15.9.4 The transient SIMPLE algorithm

SIMPLE (Patankar and Spalding 1972) stands for Semi-Implicit Method for Pressure-Linked Equations. The complete algorithm will be described below. Solving the momentum equations requires knowledge of the pressure field, and using an incorrect field results in a velocity field not satisfying continuity. We will make an initial guess  $p^*$  of the pressure field, and an imperfect velocity field based on this guessed pressure field will be denoted by  $u^*$ ,  $v^*$ ,  $w^*$ . Using (15.58), the  $u^*$  component will result from the solution of

$$a_e (u^*)_e^{n+1} = \sum a_{nb} (u^*)_{nb}^n + a_e^0 (u^*)_e^n + b + ((p^*)_P^n - (p^*)_E^n) A_e \quad (15.60)$$

where the notation  $(u^*)_e^{n+1}$  is used for expressing the velocity  $u^*$  at position  $e$  and time  $n+1$ . Similar equations are solved for  $v^*$  and  $w^*$ . Now we define the correction  $p'$  as

$$p = p^* + p'$$

and similarly for the velocities

$$u = u^* + u', v = v^* + v', w = w^* + w'$$

Subtracting (15.60) from (15.58) gives

$$a_e (u')_e^{n+1} = \sum a_{nb} (u')_{nb}^n + a_e^0 (u')_e^n + ((p')_P^n - (p')_E^n) A_e$$

At this time we set

$$\begin{aligned} \sum a_{nb} (u')_{nb}^n &= 0 \\ (u')_e^n &= 0 \end{aligned} \quad (15.61)$$

This is the main approximation of the SIMPLE algorithm and it results in the velocity-correction formula

$$\begin{aligned} (u')_e^{n+1} &= d_e ((p')_P^n - (p')_E^n) \\ u_e^{n+1} &= (u^*)_e^{n+1} + d_e ((p')_P^n - (p')_E^n) \end{aligned} \quad (15.62)$$

where

$$d_e = \frac{A_e}{a_e},$$

and  $a_e = f(\Delta t; \Delta x)$  depending on the numerical scheme used. The reason behind the approximation (15.61) is that we want to derive a equation for pressure correction  $p'$ . We construct a formula for  $p'$  that ensures convergence of the velocity field to a solution that satisfies continuity. As  $p'$  is a numerical artifice, and there is no reason to expect that the formula for predicting  $p'$  from one step to the next is physical correct. The approximation is analyzed in (Patankar 1980) and (Anderson 1995).

The above procedure can also be carried out in the  $y$  direction, which would lead to the velocity-correction formula

$$v_n^{n+1} = (v^*)_n^{n+1} + d_n ((p')_P^n - (p')_N^n) \quad (15.63)$$

We will now use the continuity equation (15.54) as a pressure correction equation. Integrating the continuity equation over the control volume centered at  $P$  and using the central difference for discretizing in  $x$  and  $y$  directions, we get

$$\rho \frac{u_e^{n+1} - u_w^{n+1}}{\Delta x} + \rho \frac{u_n^{n+1} - u_s^{n+1}}{\Delta y} = 0 \quad (15.64)$$

when evaluating the time invariant equation at time  $t + \Delta t$ . Inserting (15.62) and (15.63) into (15.64) it follows that

$$a_P (p')_P^n = a_E (p')_E^n + a_W (p')_W^n + a_N (p')_N^n + a_S (p')_S^n + b' \quad (15.65)$$

where the term

$$b' = \frac{(u^*)_e^{n+1} - (u^*)_w^{n+1}}{\Delta x} + \frac{(u^*)_n^{n+1} - (u^*)_s^{n+1}}{\Delta y}$$

is the continuity imbalance arising from the incorrect velocity field, and

$$a_P = \frac{d_w + d_e}{\Delta x} + \frac{d_n + d_s}{\Delta y}, \quad a_W = -\frac{d_w}{\Delta x}, \quad a_E = -\frac{d_e}{\Delta x}, \quad a_S = -\frac{d_s}{\Delta y}, \quad a_N = -\frac{d_n}{\Delta y}$$

The complete SIMPLE algorithm will now be stated for all three dimension.

1. Guess the pressure field  $p^*$
2. Solve the momentum equations for  $u^*, v^*, w^*$
3. Solve the pressure correction equation for  $p'$
4. Calculate  $p$  and  $u, v, w$  from  $p^*$  and  $u^*, v^*, w^*$
5. Solve the discretized equations for all other  $\phi$ . (If a  $\phi$  does not influence the flow field it is better to calculate it after the flow field calculations have converged)
6. If the solutions has converged: end, otherwise set  $p^* = p$ , and start over from step 2.

**Remark 13** The term *semi-implicit* in the name of the algorithm stems from the approximation (15.61). Without the approximation, the complete pressure correction field would have been coupled in one equation, giving a fully implicit equation. The approximation allows the pressure correction equation (15.65) to include terms only from the neighbor nodes, and it was termed only as *semi-implicit* by (Patankar and Spalding 1972).

Refinements to the SIMPLE algorithm in terms of computations effort and stability have produced algorithms such as SIMPLER, SIMPLEC and PISO. For an overview consult (Versteeg and Malalasekera 1995).

## 15.10 Von Neuman stability method

Although convergence might be difficult to establish, there exist a quite simple method to establish the stability of partial difference equations. Named after its inventor, the von Neuman method will be presented here. First the general idea will be presented, and then the detail will be made clear through a couple of examples.

Let  $D$  be the exact solution of the difference equation in question, and  $N$  be the corresponding numerical solution. The round off error, or hereafter the error, is then defined as

$$\varepsilon = N - D \Rightarrow N = D + \varepsilon \quad (15.66)$$

By showing that  $\varepsilon$  also satisfy the difference equation, it follows that the numerical solution  $N$  is stable if  $\varepsilon_i$  shrinks as the solution progresses from step  $n$  to step  $n + 1$ . That is, the solution will be stable if

$$\left| \frac{\varepsilon_P^{n+1}}{\varepsilon_P^n} \right| \leq 1 \quad (15.67)$$

and unstable otherwise. How to represent the error  $\varepsilon$  is shown in the example below.

**Example 242** Consider the one dimensional heat conduction equation

$$\frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2}$$

By representing  $\frac{\partial T}{\partial t}$  with a forward difference, and  $\frac{\partial^2 T}{\partial x^2}$  with a central difference we get

$$\frac{T_P^{n+1} - T_P^n}{\Delta t} = \frac{\alpha (T_W^n - 2T_P^n + T_E^n)}{(\Delta x)^2} \quad (15.68)$$

Let  $D$  be the exact solution of the difference equation (15.68), and  $N$  be the numerical solution. The error, is defined as

$$\varepsilon = N - D \Rightarrow N = D + \varepsilon.$$

The numerical solution  $N$  must satisfy the difference equation:

$$\frac{D_P^{n+1} + \varepsilon_P^{n+1} - D_P^n - \varepsilon_P^n}{\alpha \Delta t} = \frac{D_W^n + \varepsilon_W^n - 2D_P^n - 2\varepsilon_P^n + D_E^n + \varepsilon_E^n}{(\Delta x)^2}$$

and as  $D$  also satisfy (15.68), so must the error  $\varepsilon$ :

$$\frac{\varepsilon_P^{n+1} - \varepsilon_P^n}{\alpha \Delta t} = \frac{\varepsilon_W^n - 2\varepsilon_P^n + \varepsilon_E^n}{(\Delta x)^2} \quad (15.69)$$

Equation (15.69) is stable if

$$\left| \frac{\varepsilon_P^{n+1}}{\varepsilon_P^n} \right| \leq 1 \quad (15.70)$$

The error  $\varepsilon$  can be written as a Fourier series:

$$\varepsilon(x, t) = \sum_m A_m(t) e^{ik_m x} = \sum_m A_m(t) (\cos k_m x + i \sin k_m x) \quad (15.71)$$

where  $i = \sqrt{-1}$ ,

$$k_m = \frac{2\pi}{\lambda}$$

is the wave number and  $m$  is to be determined. It is assumed that the length of the domain on which the equation is solved is  $L$ . By using  $N + 1$  grid points, we have that

$$\Delta x = \frac{L}{N}$$

The smallest allowable wavelength in the Fourier series (15.71) is

$$\lambda_{\min} = \frac{2L}{N}$$

which is the wavelength of a sine (or cosine) function having all three zeros in adjacent gridpoints. Thus the highest wave number in the series is

$$k_{m,\max} = \frac{2\pi}{\lambda_{\min}} = \frac{2\pi}{L} \frac{N}{2}$$

which gives us the summation limits for (15.71):

$$\varepsilon(x, t) = \sum_{m=1}^{N/2} A_m(t) e^{ik_m x} = \sum_{m=1}^{N/2} A_m(t) (\cos k_m x + i \sin k_m x)$$

where

$$k_m = \left( \frac{2\pi}{L} \right) m$$

It is further assumed that the amplitude  $A_m$  varies with time as  $A_m(t) = e^{at}$ , where  $a$  is a constant. This implies

$$\varepsilon(x, t) = \sum_{m=1}^{N/2} e^{at} e^{ik_m x}$$

By substituting one term

$$\varepsilon_m(x, t) = e^{at} e^{ik_m x} \quad (15.72)$$

into (15.69), we get

$$\frac{e^{a(t+\Delta t)} e^{ik_m x} - e^{at} e^{ik_m x}}{\alpha \Delta t} = \frac{e^{at} e^{ik_m (x+\Delta x)} - 2e^{at} e^{ik_m x} + e^{at} e^{ik_m (x-\Delta x)}}{(\Delta x)^2}$$

which is simplified to

$$\begin{aligned} e^{a\Delta t} &= 1 + \frac{\alpha \Delta t}{(\Delta x)^2} (e^{ik_m \Delta x} + e^{-ik_m \Delta x} - 2) \\ &= 1 + \frac{2\alpha \Delta t}{(\Delta x)^2} (\cos(k_m \Delta x) - 1) \\ &= 1 - \frac{4\alpha \Delta t}{(\Delta x)^2} \sin^2 \left( \frac{k_m \Delta x}{2} \right) \end{aligned} \quad (15.73)$$

By combining (15.70), (15.72) and (15.73) we get

$$\begin{aligned} \left| \frac{\varepsilon_P^{n+1}}{\varepsilon_P^n} \right| &= \left| \frac{e^{a(t+\Delta t)} e^{ik_m x}}{e^{at} e^{ik_m x}} \right| = |e^{a\Delta t}| \\ &= \left| 1 - \frac{4\alpha\Delta t}{(\Delta x)^2} \sin^2 \left( \frac{k_m \Delta x}{2} \right) \right| \leq 1 \end{aligned}$$

which is the stability criterion. The factor

$$G \triangleq \left| 1 - \frac{4\alpha\Delta t}{(\Delta x)^2} \sin^2 \left( \frac{k_m \Delta x}{2} \right) \right|$$

is known as the amplification factor. The condition  $G \leq 1$  has two solutions, where the first is trivial, and the other leads to

$$1 - \frac{4\alpha\Delta t}{(\Delta x)^2} \sin^2 \left( \frac{k_m \Delta x}{2} \right) \geq -1$$

which is simplified to

$$\frac{\alpha\Delta t}{(\Delta x)^2} \leq \frac{1}{2}$$

which is the stability requirement for the difference equation (15.68) to be stable.

The exact form of the stability criterion  $G \leq 1$  depends on the form of the difference equation, that is both the original PDE and the discretization methods used. This is illustrated in the next example.

**Example 243** Given the partial differential equation for one dimensional convection (or the one dimensional wave equation)

$$\frac{\partial \phi}{\partial t} + c \frac{\partial \phi}{\partial x} = 0$$

Using a forward difference for  $\frac{\partial \phi}{\partial t}$  and Upwind difference for  $\frac{\partial \phi}{\partial x}$ :

$$\frac{\phi_P^{n+1} - \phi_P^n}{\Delta t} + c \frac{(\phi_P^n - \phi_E^n)}{\Delta x} = 0$$

which can be simplified to

$$\phi_P^{n+1} = (1 - C) \phi_P^n + C \phi_E^n \quad (15.74)$$

where

$$C = c \frac{\Delta t}{\Delta x}$$

is called the Courant number. The von Neuman stability analysis applied to (15.74), using

$$\varepsilon_m(x, t) = e^{at} e^{ik_m x}$$

gives

$$e^{a(t+\Delta t)} e^{ik_m x} = (1 - C) e^{at} e^{ik_m x} + C e^{at} e^{ik_m(x-\Delta x)}$$

and the amplification factor is given by

$$\begin{aligned} G &= \frac{\varepsilon_P^{n+1}}{\varepsilon_P^n} = e^{at} \\ &= (1 - C) + Ce^{at}e^{ik_m \Delta x} \\ &= 1 - C(1 - \cos k_m \Delta x) - iC \sin k_m \Delta x \end{aligned}$$

We now demand that

$$|G| = \sqrt{(1 - C(1 - \cos k_m \Delta x))^2 + (C \sin k_m \Delta x)^2} \leq 1$$

which after some elementary trigonometric manipulations gives

$$\sqrt{C^2 + (C - 1)^2} \leq 1 \Rightarrow C \leq 1$$

The above is an example of a more general stability result known as the Courant-Friedrichs-Lowy (CFL) condition. This condition, that is

$$C = c \frac{\Delta t}{\Delta x} \leq 1 \quad (15.75)$$

applies generally to explicit schemes for hyperbolic equations. Physically, the CFL condition indicates that for stability, a particle of fluid should not travel more than one spatial step-size  $\Delta x$  in one time step  $\Delta t$ .

The CFL condition is also the stability condition for the second order wave equation

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0 \quad (15.76)$$

studied in Example 237. There is a connection between the characteristic lines of a hyperbolic equation and the CFL condition. The following presentation is based on (Anderson 1995). The characteristic lines

$$x = \begin{cases} ct & (\text{right-running}) \\ -ct & (\text{left-running}) \end{cases}$$

for (15.76) are plotted in Figure 15.14. In both figures 15.14 a) and b) point  $b$  is the intersection of the right-running characteristic through grid-point  $W$  and the left-running characteristic through grid point  $E$ . This point also has a connection to the CFL condition. Let  $\Delta t_{C=1}$  denote the value of  $\Delta t$  given by (15.75) when  $C = 1$ , that is

$$\Delta t_{C=1} = \frac{\Delta x}{c}$$

In Figure 15.14 a) and b)  $\Delta t_{C=1}$  is exactly the distance between point  $P$  and point  $b$  given by the intersection of the characteristics. Now, study Figure 15.14 a) and assume  $C < 1$ . Then,  $\Delta t_{C<1} < \Delta t_{C=1}$ . Let point  $d$  correspond to the grid point directly above point  $P$  existing at time  $t + \Delta t_{C<1}$ . Since properties at point  $d$  are calculated numerically from the difference equation using information at grid points  $E$  and  $W$ , the *numerical domain* for point  $d$  is the triangle  $adc$ . The numerical domain is denoted  $\mathcal{D}_n$ . The *analytical domain* for  $d$  is the shaded area defined by the characteristics through  $d$ . The analytical domain

is denoted  $\mathcal{D}_a$ . Note that in this case, the numerical domain contains the analytical domain.

Now, consider Figure 15.14 b). Then  $C > 1$  and  $\Delta t_{C>1} > \Delta t_{C=1}$ . Let point  $d$  correspond to the grid point directly above point  $P$  existing at time  $t + \Delta t_{C>1}$ . As before, the numerical domain in this case is the triangle  $adc$ , and the analytical domain is defined by the characteristics through  $d$ . Note that in this case, the numerical domain does not contain all of the analytical domain. The case in 15.14 b) considered an unstable solution as  $C > 1$ , and we can now state the following interpretation of the CFL condition: For stability, the numerical domain must include all of the analytical domain, that is

$$\mathcal{D}_a \subset \mathcal{D}_n$$

Figure 15.14 a) can also illustrate *accuracy*. The analytical domain for point  $d$  is the shaded triangle. Note however, that the numerical grid points  $E$  and  $W$  are outside the domain of dependence of  $d$  and should therefore theoretically not influence the properties at  $d$ . On the other hand, the numerical calculation of the properties at  $d$  takes information from  $E$  and  $W$  into account. This leads to inaccurate results as there is a mismatch between the domain of dependence and the actual numerical data used. In view of accuracy it is therefore desirable to have the Courant number as close as possible to unity.

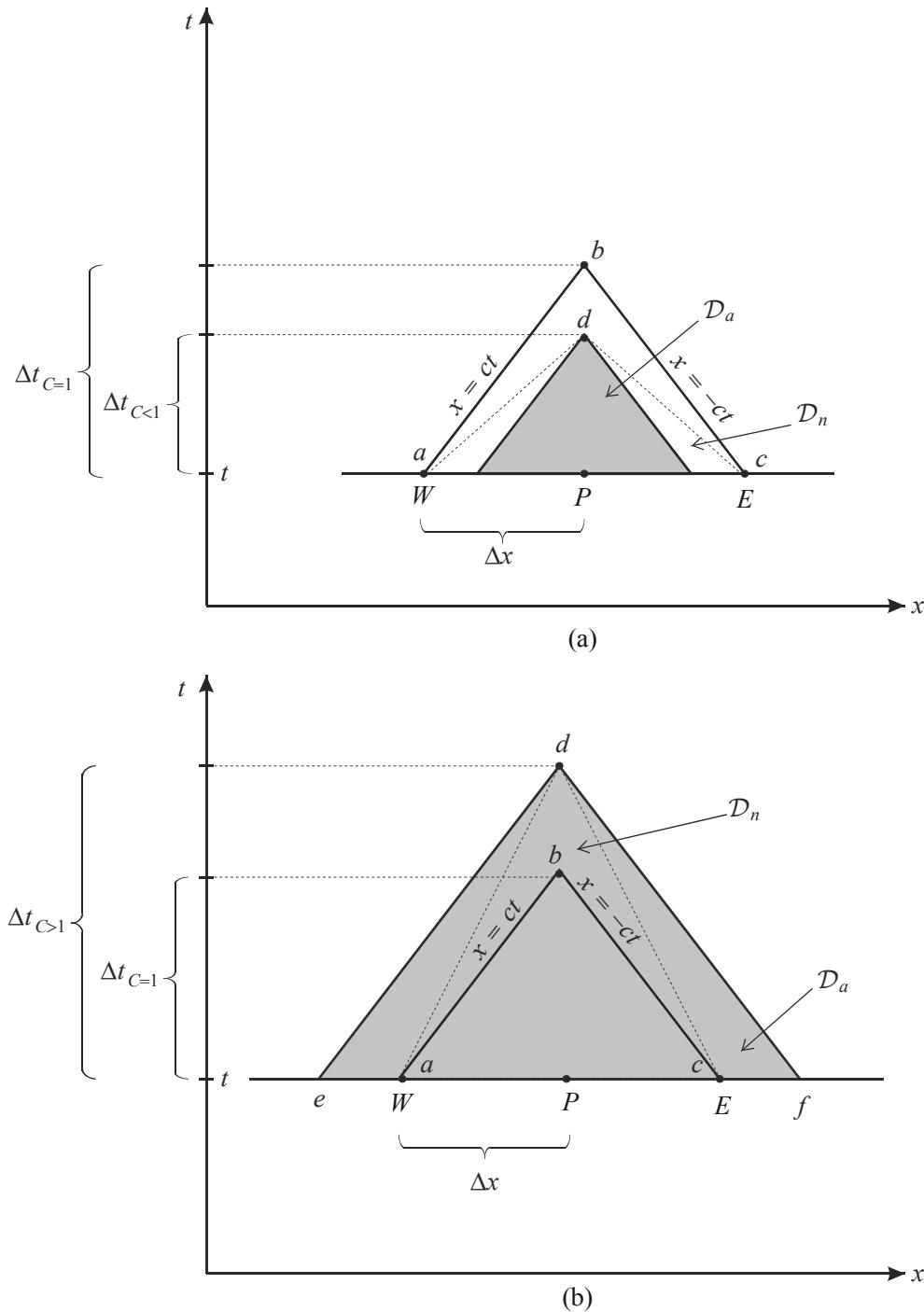


Figure 15.14: (a) Stable (b) Unstable



# Bibliography

- Aamo, O. and Fossen, T.: 2000, Finite element modelling of mooring lines, *Mathematics and Computers in Simulation* **53**, 415 – 422.
- Aamo, O. and Krstić, M.: 2003, *Flow Control by Feedback. Stabilization and Mixing*, Springer-Verlag, London.
- Anderson, B. and Moore, J.: 1989, *Optimal Control. Linear Quadratic Methods*, Prentice-Hall, Englewood Cliffs.
- Anderson, B. and Vongpanitlerd, S.: 1973, *Network Analysis and Synthesis*, Prentice-Hall, Englewood Cliffs, New Jersey.
- Anderson, J.-D.: 1995, *Computational fluid dynamics*, McGraw-Hill, New York.
- Angeles, J.: 1988, *Rational Kinematics*, Springer-Verlag, New York.
- Antsaklis, P. and Michel, A.: 1997, *Linear Systems*, McGraw-Hill, New York.
- Arimoto, S.: 1996, *Control Theory of Nonlinear Mechanical Systems: A Passivity-Based and Circuit-Theoretic Approach*, Oxford University Press, Oxford.
- Aris, R.: 1989, *Vectors, tensors, and the basic equations of fluid mechanics*, Dover Publications, New York.
- Armstrong-Hélouvry, B.: 1990, Stick-slip arising from stribeck friction, *Proceedings of the 1990 International conference on robotics and automation*, Cincinnati, OH, pp. 1377–1382.
- Armstrong-Hélouvry, B., Dupont, P. and Canudas de Wit, C.: 1994, A survey of models, analysis tools and compensation methods for the control of machines with friction, *Automatica* **30**(7), 1083–1138.
- Arnold, V.: 1989, *Mathematical Methods of Classical Mechanics*, 2nd edn, Springer-Verlag, New York.
- Aström, K. and Furuta, K.: 2000, Swinging up a pendulum by energy control, *Automatica* **36**(2), 287–295.
- Barabanov, N. and Ortega, R.: 2000, Necessary and sufficient conditions for passivity of the LuGre friction model, *IEEE Transactions on Automatic Control* **45**(4), 830–832.
- Bathe, K.-J.: 1996, *Finite Element Procedures*, Prentice-Hall, Englewood Cliffs, New Jersey.

- Bird, R., Stewart, W. and Lightfoot, E.: 1960, *Transport Phenomena*, John Wiley, New York.
- Blanke, M.: 1981, *Ship Propulsion Losses Related to Automatic Steering and Prime Mover Control*, PhD thesis, Danmarks tekniske højskole.
- Bremer, H.: 1988, über eine zentralgleichung in der dynamik, *Z. angew. Math. Mech.* **68**, 307–311.
- Canudas de Wit, C., Olsson, H., Åström, K. and Lischinsky, P.: 1995, A new model for control of systems with friction, *IEEE Transactions on Automatic Control* **40**(3), 419–424.
- Chen, C.: 1999, *Linear System Theory and Design*, Oxford University Press, Oxford.
- Cohen, H., Rogers, G. and Saravanamuttoo, H.: 1996, *Gas turbine theory*, 4th edn, Longman, Essex.
- Crandall, S., Karnopp, D., E.F. Kurtz, J. and Pridmore-Brown, D.: 1968, *Dynamics of Mechanical and Electromechanical Systems*, McGraw-Hill, New York.
- Cumpsty, N.: 1989, *Compressor Aerodynamics*, Longman.
- Dahl, P.: 1968, A solid friction model, *Technical Report TOR-0158(3107-18)-1*, The Aerospace corporation, El Segundo, Calif. 90245.
- Dahl, P.: 1976, Solid friction damping of mechanical vibrations, *AIAA Journal* **14**(12), 1675–1682.
- de Groot, S. and Mazur, P.: 1984, *Non-Equilibrium Thermodynamics*, Dover Publications, New York.
- der Waerden, B. V.: 1976, Hamilton's discovery of the quaternions, *Mathematics Magazine* **5**, 227 – 234.
- Dorf, R. and Bishop, R.: 2000, *Modern Control Systems*, 9th edn, Addison-Wesley, Reading, Massachusetts.
- Dormand, J. and Prince, P.: 1986, Runge-kutta triplets, *Comp. and Maths. with Applic.* **12A**, 1007–1017.
- Dupont, P., Hayward, V., Armstrong, B. and Altpeter, F.: 2002, Single state elastoplastic friction models, *IEEE Transactions Automatic Control* **47**(5), 787 – 792.
- Egeland, O. and Godhavn, J.-M.: 1994, Passivity-based attitude control of a rigid space-craft, *IEEE Transactions on Automatic Control* **39**, 842–846.
- Ellman, A. and Piché, R.: 1999, A two regime orifice flow formula for numerical simulation, *Journal of Dynamic Systems, Measurement, and Control* **121**(4), 721 – 724.
- Ervik, M.: 1971, *Regulatorer for Vannkraft Makiner*, Universitetsforlaget, Oslo.
- Evans, L.: 1998, *Partial Differential Equations*, Graduate Studies in Mathematics, American Mathematical Society, Providence, Rhode Island.
- Farell, J. and Barth, M.: 1999, *The Global Positioning System and Inertial Navigation*, McGraw-Hill, New York.

- Ferziger, J. and Perić, M.: 1999, *Computational Methods for Fluid Dynamics*, second edn, Springer Verlag, Heidelberg.
- Fitzgerald, A., Kingsley, C. and Umans, S.: 1983, *Electric Machinery*, 4th edn, McGraw-Hill, New York.
- Fossen, T.: 1994, *Guidance and Control of Ocean Vehicles*, John Wiley & Sons, Chichester.
- Fossen, T.: 2002, *Marine Control Systems: Guidance, Navigation and Control of Ships, Rigs and Underwater Vehicles*, Marine Cybernetics, Trondheim, Norway.
- Fuller, C., Elliott, S. and Nilsen, P.: 1996, *Active Control of Vibration*, Academic Press, London.
- Futral, S. and Wasserbauer, C.: 1965, Off design performance prediction with experimental verification for a radial-inflow turbine, *Technical Report TN D-2621*, NASA.
- Gavronski, W., Beech-Brandt, J., Ahlstrom, H. and Manieri, E.: 2000, Torque-bias profile for improved tracking of deep space network antennas, *IEEE Antennas and Propagation magazine* **42**, 35 – 45.
- Gevarter, W.: 1970, Basic relations for control of flexible vehicles, *AIAA Journal* **8**(4), 666 – 672.
- Goldstein, H.: 1980, *Classical Mechanics*, Addison Wesley, Reading, Massachusetts.
- Golub, G. and van Loan, C.: 1989, *Matrix computations*, 2nd edn, The John Hopkins University Press, Baltimore, Maryland.
- Goodson, R. and Leonard, R.: 1972, A survey of modeling techniques for fluid line transients, *Trans. of the ASME. Journal of Basic Engineering*. pp. 474 – 482.
- Gravdahl, J. and Egeland, O.: 1998, Two results on compressor surge control with disturbance rejection, *Proceedings of the 37th IEEE Conference on Decision and Control*. To appear.
- Gravdahl, J. and Egeland, O.: 1999, *Compressor surge and rotating stall: modeling and control*, Advances in Industrial Control, Springer-Verlag, London.
- Gravdahl, J. and Egeland, O.: 2002, Drive torque actuation in active surge control of centrifugal compressors, *Automatica* **38**(11), 1881–1893.
- Gravdahl, J., Egeland, O. and Vatland, S.: 2001, Active surge control of centrifugal compressors using drive torque, *Proceedings of 40th IEEE Conference on Decision and Control*, pp. 1286 –1291.
- Greitzer, E.: 1976, Surge and Rotating stall in axial flow compressors, Part I: Theoretical compression system model, *Journal of Engineering for Power* **98**, 190–198.
- Haessig, Jr., D. and Friedland, B.: 1991, On the modeling and simulation of friction, *ASME J. of Dynamic Systems, Measurement and Control* **113**, 354–362.
- Hairer, E.: 1999, Numerical geometric integration, Report found on <http://www.unige.ch/math/folks/hairer/polycop.html>.

- Hairer, E., Nørsett, S. and Wanner, G.: 1993, *Solving ordinary differential equations I: Nonstiff problems*, 2nd edn, Springer-Verlag Berlin.
- Hairer, E. and Wanner, G.: 1996, *Solving ordinary differential equations II: Stiff and differential-algebraic problems*, 2nd edn, Springer-Verlag Berlin.
- Hartman, P.: 1982, *Ordinary Differential Equations*, Elsevier, New York.
- Hess, D. and Soom, A.: 1990, Friction at a lubricated line contact operating at oscillating sliding velocities, *ASME Journal of Tribology* **112**(1), 147–152.
- Heywood, J.: 1988, *Internal combustion engine fundamentals*, McGraw-Hill.
- Holmboe, E. and Rouleau, W.: 1967, The effect of viscous shear on transients in liquid lines, *Journal of Basic Engineering, Trans. ASME, Series D* **89**, 174 – 180.
- Hughes, P.: 1974, Dynamics of flexible space vehicles with active attitude control, *Celestial Mechanics* **9**, 21–39.
- Hughes, P.: 1986, *Spacecraft Attitude Dynamics*, John Wiley, New York.
- Hutarew, G.: 1969, *Regelungstechnik : Kurze Einführung Am Beispiel der Drehzahlregelung Von Wasserturbinen*, Springer, Berlin.
- IEEE: 1987, IEEE standard on piezoelectricity, *Technical Report Standard 176-1987*, ANSI/IEEE.
- Isidori, A.: 1989, *Nonlinear Control Systems*, 2nd. edn, Springer-Verlag, Berlin.
- Joshi, S.: 1989, *Control of Large Flexible Space Structures*, Vol. 131 of *Lecture Notes in Control and Information Science*, Springer-Verlag, Berlin.
- Kane, T. and Levinson, D.: 1983, The use of Kane's dynamical equations in robotics, *Int. J. Robotics Research* **2**, 3–21.
- Kane, T. and Levinson, D.: 1985, *Dynamics: Theory and Applications*, McGraw-Hill, New York.
- Kane, T., Likins, P. and Levinson, D.: 1983, *Spacecraft Dynamics*, McGraw-Hill, New York.
- Karnopp, D.: 1985, Computer simulation of stick-slip friction in mechanical dynamic systems, *ASME J. of Dynamic Systems, Measurement and Control* **107**, 100–104.
- Karnopp, D., Margolis, D. and Rosenberg, R.: 2000, *System Dynamics. Modeling and Simulation of Mechatronic Systems*, 3rd. edn, Wiley-Interscience, New York.
- Kelkar, A. and Joshi, S.: 1996, *Control of Nonlinear Multibody Flexible Space Structures*, Springer-Verlag, London.
- Kelly, R., Carelli, R. and Ortega, R.: 1989, Adaptive motion control design of robot manipulators: An input-output approach, *International Journal of Control* **50**(6), 2563–2581.
- Khalil, H.: 1996, *Nonlinear systems*, 2nd edn, Prentice-Hall, Inc.

- Kiencke, U. and Nielsen, L.: 2000, *Automotive Control Systems for Engine, Driveline, and Vehicle*, Springer, Berlin.
- Kreyszig, E.: 1979, *Advanced Engineering Mathematics*, 4th edn, Wiley, New York.
- Kristiansen, E.: 2000, *Energy-based observer design for manipulators with flexible links*, Master's thesis, Norwegian University of Science and Technology.
- Krstić, M., Kanellakopoulos, I. and Kokotović, P.: 1995, *Nonlinear and Adaptive Control Design*, Wiley, New York.
- Kuo, B.: 1995, *Automatic Control Systems*, 7th edn, Prentice-Hall, Englewood Cliffs, New Jersey.
- Lamb, H.: 1945, *Hydrodynamics*, Dover Publications, New York.
- Lambert, J.: 1991, *Numerical methods for ordinary differential equations. The initial value problem*, John Wiley, Chichester.
- Lanczos, C.: 1986, *The Variational Principles of Mechanics*, Dover Publications, New York.
- Leonard, B.: 1979, A stable and accurate convective modelling procedure based on quadratic upstream interpolation, *Computer Methods in Applied Mechanics and Engineering* **19**(1), 59–98.
- Leonard, N.: 1997, Stability of bottom-heavy underwater vehicle, *Automatica* **33**(3), 331–346.
- Leonhard, W.: 1996, *Control of Electrical Drives*, 2nd edn, Springer, Berlin.
- Lewis, D. and Simo, J.: 1994, Conserving algorithms for the dynamics of hamiltonian systems on lie groups, *J. Nonlinear Science* **4**, 253 – 299.
- Lin, C. and Segel, L.: 1974, *Mathematics applied to deterministic problems in the natural sciences*, Macmillan, New York.
- Lohmiller, W. and Slotine, J.: 1998, On contraction analysis for nonlinear systems, *Automatica* **34**, pp. 683–696.
- Lovelock, D. and Rund, H.: 1989, *Tensors, Differential Forms, and Variational Principles*, Dover Publications, New York.
- Lozano, R., Brogliato, B., Egeland, O. and Maschke, B.: 2000, *Dissipative Systems Analysis and Control. Theory and Applications*, Springer-Verlag, London.
- Luh, J., Walker, M. and Paul, R.: 1980, On-line computational scheme for mechanical manipulators, *ASME J. Dynamic Syst., Meas., Contr.* pp. 69–76.
- Mäkinen, J., Piché, R. and Ellman, A.: 2000, Fluid transmission line modeling using a variational method, *J. Dynamic Systems, Measurement, and Control* **122**, 153 – 162.
- Marsden, J. and Ratiu, T.: 1994, *Introduction to Mechanics and Symmetry*, Springer-Verlag, New York.

- McCarthy, J.: 2000, *Geometric Design of Linkages*, Springer-Verlag, New York.
- Meirovitch, L.: 1967, *Analytical Methods in Vibrations*, Mamillan, New York.
- Meirovitch, L.: 1980, *Computational Methods in Structural Dynamics*, Sijthoff & Noordhoff, Alphen aan den Rijn, The Netherlands.
- Meisel, J.: 1966, *Principles of Electromechanical Energy Conversion*, McGraw-Hill, New York.
- Merritt, H.: 1967, *Hydraulic Control Systems*, John Wiley, New York.
- Milne-Thomson, L.: 1996, *Theoretical Hydrodynamics*, Dover Publications, New York.
- Mohan, N., Undeland, T. and Robbins, W.: 1989, *Power Electronics: Converters, Applications and Design*, John Wiley, New York.
- Moore, F. and Greitzer, E.: 1986, A theory of post-stall transients in a axial compressor systems: Part I—Development of equations, *Journal of Engineering for Gas Turbines and Power* **108**, 68–76.
- Murray, R., Li, Z. and Sastry, S.: 1994, *A Mathematical Introduction to Robotic Manipulation*, CRC Press, Boca Raton.
- Nicklasson, P.: 1996, *Passivity-based control of electric machines*, PhD thesis, Norwegian Institute of Technology, Dept of Engineering Cybernetics.
- Niemeyer, G. and Slotine, J.: 1991, Stable adaptive teleoperation, *IEEE J. Oceanic Engineering* **16**(1).
- Nijmeijer, H. and der Schaft, A. V.: 1990, *Nonlinear Dynamical Control Systems*, Springer-Verlag, Berlin.
- Nilsson, J.: 1983, *Electric Circuits*, Addison-Wesley, Reading, Massachusetts.
- Ortega, R., Loria, A., Nicklasson, P. and Sira-Ramirez, H.: 1998, *Passivity-Based Control of Euler-Lagrange Systems: Mechanical, Electrical and Electromechanical Applications*, Springer-Verlag.
- Patankar, S.: 1980, *Numerical Heat Transfer and Fluid Flow*, Hemisphere Publishing Corporation.
- Patankar, S. and Spalding, D.: 1972, A calculation procedure for heat, mass and momentum transfer in three-dimensional parabolic flows, *Int. J. Heat Mass Transfer* **15**, 1787–1806.
- Pazy, A.: 1983, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Vol. 44 of *Applied Mathematical Sciences*, Springer-Verlag, New York.
- Piché, R. and Ellman, A.: 1996, *A Fluid Transmission Line Model for Use with ODE Simulators*, Research Studies Press, Somerset, England, pp. 221 – 236.
- Rao, S.: 1990, *Mechanical Vibrations*, Addison-Wesley, Reading, Massachusetts.
- Robertson, R. and Schwertassek, R.: 1988, *Dynamics of Multibody Systems*, Springer-Verlag, Berlin.

- Rugh, W.: 1996, *Linear System Theory*, 2nd edn, Prentice-Hall, Englewood Cliffs.
- Sagatun, S. and Fossen, T. I.: 1991, Lagrangian formulation of underwater vehicles' dynamics, *Proc. IEEE Int. Conf. Systems, Man and Cybernetics*, pp. 1029–1034.
- Samson, C., Borgne, M. L. and Espiau, B.: 1991, *Robot Control: The Task Function Approach*, Vol. 22 of *Oxford Engineering Science Series*, Oxford University Press, Oxford.
- Sanz-Serna, J. and Calvo, M.: 1994, *Numerical Hamiltonian problems*, Chapman and Hall, London.
- Scarborough, J.: 1966, *Numerical Mathematical Analysis*, sixth edn, John Hopkins Press, Baltimore.
- Sciavicco, L. and Siciliano, B.: 2000, *Modeling and Control of Robot Manipulators*, Springer-Verlag, London Berlin Heidelberg.
- Sepulchre, R., Janković, M. and Kokotović, P.: 1997, *Constructive Nonlinear Control*, Springer-Verlag, London.
- Shahruz, S.: 1999, Boundary control of kirchhoff's non-linear string, *Int. Journal of Control* **72**(6), 560 – 563.
- Shampine, L.: 1994, *Numerical Solution of Ordinary Differential Equations*, Chapman and Hall, New York.
- Shampine, L., Allen, R. and Preuss, S.: 1997, *Fundamentals of numerical computing*, John Wiley, New York.
- Shampine, L. and Reichelt, M. W.: 1997, The matlab ode suite, *SIAM Journal on Scientific Computing* **18**(1). also in MATLAB Helpdesk, Online Manuals (in PDF).
- Shepperd, S.: 1978, Quaternion from rotation matrix, *J. Guidance and Control* **1**, 223–224.
- Skogestad, S. and Postlethwaite, I.: 1996, *Multivariable Feedback Control. Analysis and Design*, John Wiley, Chichester.
- Slotine, J.: 1991, *Applied Nonlinear Control*, Prentice-Hall, Englewood Cliffs, New Jersey.
- Slotine, J. and Li, W.: 1988, Adaptive manipulator control: A case study, *IEEE Transactions on Automatic Control* **33**, 995–1003.
- Spong, M. and Vidyasagar, M.: 1989, *Robot Dynamics and Control*, John Wiley, New York.
- Stecki, J. and Davis, D.: 1986a, Fluid transmission lines - distributed parameter models. part 1: A review of the state of the art, *Proc. Instn. Mech. Engrs.* **200**, 215 – 228.
- Stecki, J. and Davis, D.: 1986b, Fluid transmission lines - distributed parameter models. part 2: Comparison of models, *Proc. Instn. Mech. Eng.* **200**, 229 – 236.
- Strang, G.: 1988, *Linear Algebra and its Applications*, Saunders.
- Stribeck, R.: 1902, Die wesentlichen Eigenschaften der Gleit- und Rollenlager, *Zeitschrift des Vereines Deutcher Ingenieure* **46**(39), 1432–37.

- Szebehely, V.: 1967, *Theory of orbits*, Academic press, New York.
- Takegaki, M. and Arimoto, S.: 1981, A new feedback method for dynamic control of manipulators, *ASME J. Dyn. Syst. Meas. Control* **102**, 119–125.
- Titterton, D. and Weston, J.: 1997, *Strapdown Inertial Navigation Technology*, Peter Peregrinus on behalf of IEEE, London.
- Triantafyllou, M.: 1990, Cable mechanics with marine applications, *Lecture notes*, MIT.
- Vas, P.: 1990, *Vector Control of AC Machines*, Oxford University Press, Oxford.
- Versteeg, H. and Malalasekera, W.: 1995, *An Introduction to Computational Fluid Dynamics: The Finite Volume Method*, Longman, Essex.
- Walker, M. and Orin, D.: 1982, Efficient dynamic computer simulation of robotic mechanisms, *Journal of Dynamic Systems, Measurement and Control* **104**, 205 – 211.
- Watton, J.: 1989, *Fluid Power Systems. Modeling, Simulation, Analog and Microcomputer Control*, Prentice-Hall, New York.
- Weaver, W., Timoshenko, S. and Young, D.: 1990, *Vibration Problems in Engineering, 5th Ed.*, John Wiley & Sons, New York.
- Wen, J.-Y. and Kreutz-Delgado, K.: 1991, The attitude control problem, *IEEE Transactions on Automatic Control* **36**, 1148–1162.
- White, F.: 1999, *Fluid mechanics*, 4th edn, McGraw-Hill, New York.
- Wie, B. and Bryson, A.: 1987, Pole-zero modeling of flexible space structures, *Journal of Guidance, Control and Dynamics* **11**(6), 554 – 561.
- Woods, R.: 1983, A first-order square-root approximation for fluid transmission lines, in M. Franke and T. Drzewiecki (eds), *Fluid Transmission Line Dynamics 1983*, ASME, pp. 37–49.
- Yang, W. and Tobler, W.: 1991, Dissipative model approximation of fluid transmission lines using linear friction model, *ASME Journal of Dynamic Systems, Measurement and Control* **113**, 152–162.

# Index

- A-stability, 546, 550
- Adams methods
  - Explicit, 576
  - Implicit, 578
- Adams-Basforth methods, 576
- Adams-Moulton methods, 578
- Adjoint transformation
  - Rigid motion, 333
- Affine in the control, 4
- Algebraic stability, 558
- Aliasing, 544
- Amplification factor, 622
- AN-stability, 555
- Analytic, 18
- Angle-axis parameters, 227
- Angular velocity, 240
  
- B-stability, 557
- Barycentric material derivative, 423
- Barycentric velocity, 422
- BDF methods, 580
- Bernoulli's equation
  - Along a streamline, 429
  - Irrational, 428
- Bond graphs, 26
- Bound vector, 263
- Boundedness, 605
- Bulk modulus, 151
- Burger's equation, 425
- Butcher array, 527
  
- Cauchy's equation of motion, 455
- Causality, 41
- Cayley transformation, 239
- Cayley's formula, 239
- Center of gravity, 266
- Center of mass, 261
- Characteristic impedance, 34
- Christoffel symbols, 320
- Clausius-Duhem inequality, 468
- Closed loop transfer function, 14
  
- Coenergy, 103, 104
- Collocation, 361
- Complementary sensitivity function, 15
- Componets, 210
- Composite rotation, 220
- Compressor characteristic
  - Nondimesional form, 497
- Compressor dynamics, 485
- Computational causality, 26
- Configuration, 289
- Configuration space, 289
- Conservativness, 605
- Consistent mass matrix, 371
- Continuity equation, 417
  - Multi-component systems, 422
- Coordinate transformation matrix, 219
- Coordinate vector, 210
- Coordinate vectors
  - Differentiation, 242
- Coordinates, 210
- Corrected mass flow, 503
- Corrected speed, 504
- Coulomb friction, 192
- Couple, 265
- Courant number, 622
- Courant-Friedrichs-Lowy condition, 623
- Curvilinear coordinates, 408
  
- d'Alembert's principle, 291
- Dahl effect, 191
- deformation tensor, 451
- Degrees of freedom, 290
- Denavit-Hartenberg convention, 224
- Dense outputs, 565
- Differential-algebraic equations, 585, 586
  - Index 1, 586
  - Multistep methods, 589
  - Runge-Kutta method, 587
- Diffusion flow, 423
- Dilation, 406
- Direction cosines, 219

- Dirichlet condition, 597
- Discharge coefficient, 432
- Divergence, 406
- Divergence theorem, 403
- Dormand-Prince 5(4), 562
- Driving-point impedance, 26
- Dyadic, 215
  - Identity, 216
  - Inertia, 214
- Elasticity
  - Distributed parameter, 361
  - Lumped parameter, 361
- Electrical time constant, 88
- Elliptic equations, 594
- Embedded solution, 561
- Energy balance
  - Isentropic processes, 474
  - Pressure form, 471
  - Temperature form, 470
- Energy function, 349
- Enthalpy, 443
  - Stagnation, 474
- Entropy
  - Material derivative, 468
  - Specific, 467
- Entropy equation, 468
- Euler angles, 225
  - Classical, 226
  - Roll-Pitch-Yaw, 225
- Euler Bernoulli beam, 373
- Euler parameters, 231, 571
  - From rotation matrix, 236
- Euler rotation vector, 237
- Euler's equation of motion, 424
- Euler's method, 521, 528, 570, 574, 578
- Euler-angle singularity, 247
- Euler-Bernoulli beam, 390, 551
- Euler-Rodrigues parameters, 238
- Event detection, 566
- Explicit Adams methods, 576
- Explicit midpoint rule, 525
- Fehlberg 4(5), 561
- Field weakening, 119
- Field-oriented control, 130
- Flow coefficient, 497
- Flux linkage, 102
- Forces of constraints, 290
- Fourier's law, 446, 470
- FSAL method, 534, 562
- Gain margin, 44
- Gauss method, 537, 539, 547, 551, 560, 570
- Gear, 80
- Gear ratio, 80
- Generalized coordinates, 289, 314, 340
- Generalized force, 315
- Global error, 517
- Greitzer surge model, 487, 498
  - Linearization, 499
  - Normalized, 497
- Greitzer's B-parameter, 498
- Helmholtz frequency, 499
- Helmholtz resonator, 476
- Hessian matrix, 450
- Heun's method, 528
- Homogeneous transformation matrix, 223
- Hydraulic gear, 185
- Hydraulic motor
  - Pump controlled, 185
  - Valve controlled, 157
- Hydrodynamic motor, 141
- Hydrostatic motor, 141
- Hyperbolic equations, 594
- Hyrostatic gear, 185
- Ideal gas, 466
- Identity dyadic, 216
- Implicit Adams methods, 578
- Implicit Euler method, 535
- Implicit methods
  - Numerical solution, 563
- Implicit mid-point rule, 538, 539, 550
- Improved Euler method, 523, 528, 533
- Impulse response, 12
- Inertia dyadic, 214, 270
- Inertia matrix, 274
- Inertia tensor, 274
- Inertial frame, 259
- Infinite dimensional systems, 18
- Internal energy, 465
- Inverted pendulum, 281
- Inviscid fluid, 424
- Irrational transfer function, 17
- Isentropic process, 472
- Isentropic relations, 473
- Jacobi identity, 213

- Jacobian, 519, 565  
Kane's equation of motion, 297, 340  
Kinematic differential equations  
    Attitude deviation, 245  
    Euler angles, 247  
    Euler parameters, 248  
    Euler rotation, 250  
    Euler-Rodrigues parameters, 250  
    Rotation matrix, 240  
Kronecker tensor product, 564  
  
L-stability, 546, 547, 550  
Lame coefficients, 456  
Laplacian, 450  
Lie bracket  
    Rotations, 332  
Line of action, 263  
Linear invariant, 568  
Linear test system, 520  
Linearization, 5, 519  
Lobatta IIIC, 551  
Lobatto IIIA, 536, 539, 547, 550, 556, 560, 570, 574  
Lobatto IIIB, 540, 550, 556, 560, 570, 574  
Lobatto IIIC, 540, 550, 555, 560, 570, 574  
Local error, 517  
Local extrapolation, 561  
Local solution, 517  
Loop transfer function, 14  
  
Mach number, 480  
Magnetomotive force (mmf), 101, 122  
Material control volume, 407  
Material coordinates, 406  
Material derivative, 402  
Material volume, 414  
Matrix exponential, 230  
Mechanical time constant, 88  
mmf, 101, 122  
Modified Euler method, 525, 528  
Momentum balance, 423  
Momentum vector, 347  
Multiport, 21  
  
Nabla operator, 402  
Navier-Stokes equation, 459  
NDF methods, 581  
  
Neumann condition, 597  
Newton search, 564  
Newtonian fluid, 456, 458  
Newtonian frame, 259  
Noncollocation, 361  
Nonconsistent mass matrix, 371  
Normal plane, 254  
Nozzle flow, 480  
Numerical dissipation, 540  
  
O(.) notation, 518  
One-step method, 517  
Order, 518  
Osculating plane, 254  
  
Padé approximation, 18  
Padé approximations, 548  
Parabolic equations, 594  
Parallel axes theorem, 275  
Passive electrical one-port, 67  
Passivity, 361  
    Energy formulation, 63  
    PID controller, 65  
PD controller  
    Mechanical analog, 31  
PECE, 580  
Peclet number, 605  
Permutation symbol, 211  
Phase margin, 44  
Piezoelectric actuator, 114  
Piezoelectricity, 114  
Pole at infinity, 11  
Positive real, 56  
Pre-whirl, 440  
Predictor-Corrector method, 580  
Pressure  
    Stagnation, 474  
    Static, 474  
Pressure coefficient, 497  
Principle of virtual work, 290  
Proper transfer function, 11  
Pump controlled motor, 185  
  
Quaternion, 232  
Quaternions, 231  
    Identity quaternion, 234  
    Inverse quaternion, 234  
    Quaternion product, 232  
  
Radau IA, 539, 560, 570, 574

- Radau IIA, 539, 560, 570, 574
- Radau methods, 550
- Rate of production, 422
- Rate of strain tensor, 451, 456
- Rational transfer function, 11
- Rectifying plane, 254
- Reduction gear, 80
- Restriction
  - Gas flow, 480
  - Liquid flow, 431
  - Sonic gas flow, 482
- Reynolds number, 459
- Reynolds' transport theorem, 413
- Rigid body, 259
- RK4, 528, 535, 570, 574
- Rosenbrock method, 574
  - Stability function, 575
- Rotation dyadic, 229
- Rotation matrix, 219
  - Basic, 221
  - Classical Euler angles, 226
  - Composite rotations, 221
  - Euler angles, 225
  - Euler parameters, 231
  - Euler rotation vector, 237
  - Euler-Rodrigues parameters, 239
  - Simple rotation, 221
- Rothalpy, 495
- Runge-Kutta method
  - Continuous, 565
  - Differential-Algebraic, 587
  - Dormand-Prince 5(4), 562
  - Explicit, 526, 560
  - Fehlberg 4(5), 561
  - Implicit, 535, 563
  - Inertia matrix, 566
  - Stability function, 531
  - Step size selection, 560
  - Symplectic, 571
- Runge-Kutta methods
  - Property table, 560
- Sensitivity function, 14
- Serret-Frenet frame, 253
- Shock, 484
- Singularity of transfer function, 18
- Skew-symmetric form, 211
- $\text{SO}(3)$ , 220
- Sonic flow
  - Restriction, 482
- Space marching, 595
- Spatial coordinates, 406
- Specific energy, 465
- Specific entropy, 467
- Specific force, 258
- Specific internal energy, 465
- Specific kinetic energy, 465
- Specific volume, 408, 466
- Speed of sound, 475
- Stability
  - A-stability, 546
  - Algebraic stability, 558
  - AN-stability, 555
  - B-stability, 557
  - BDF methods, 583
  - Explicit Adams methods, 582
  - Implicit Adams methods, 582
  - L-stability, 546
  - Multistep methods, 581
  - Nonlinear analysis, 557
  - Pade approximations, 550
- Stability function, 518, 519, 531
- Stage computations, 526
- Stages, 526
- Staggered grid, 616
- Stagnation state, 474
- State space model, 4
- Static pressure, 474
- Step size selection, 560
- Stick-slip friction, 192
- Stiff systems, 535
- Stiffly accurate, 546, 588
- Stoke's assumption, 457
- Stokes' Theorem, 405
- Storage function, 63
- Stress tensor, 453
- Strubeck curve, 192
- Strubeck effect, 195
- Strictly proper transfer function, 11
- Summation convention, 449
- Telemanipulation, 70
- Theta method, 538
- Time marching, 595
- Torque, 265
- Torsion, 254
- Transfer function
  - Proper, 11
  - Strictly proper, 11
- Transmission line, 18

- Electric, 33
- Hydraulic, 73
- Lossless, 38
- Transportiveness, 605
- Trapezoidal rule, 537, 539, 547, 550, 556
- Twist vector, 246
- Undamped natural frequency, 42
- Unit quaternion, 233
- Unit step function, 12
- Valve controlled motor, 157
- Variable displacement pump, 185
- Variations, 325
- Vector, 209
- Vectors
  - Differentiation, 242
  - Virtual change, 325
  - Virtual displacement, 290
- Viscous stress tensor, 454
- Wave equation, 475
  - d'Alembert's solution, 475
- Wave variables, 37
- Zero at infinity, 11
- Zero crossing, 566