

SPRINGER OPTIMIZATION
AND ITS APPLICATIONS

33

Giuseppe Buttazzo · Aldo Frediani (Eds.)

Variational Analysis and Aerospace Engineering

 Springer

VARIATIONAL ANALYSIS AND AEROSPACE ENGINEERING

Springer Optimization and Its Applications

VOLUME 33

Managing Editor

Panos M. Pardalos (University of Florida)

Editor–Combinatorial Optimization

Ding-Zhu Du (University of Texas at Dallas)

Advisory Board

J. Birge (University of Chicago)

C.A. Floudas (Princeton University)

F. Giannessi (University of Pisa)

H.D. Sherali (Virginia Polytechnic and State University)

T. Terlaky (McMaster University)

Y. Ye (Stanford University)

Aims and Scope

Optimization has been expanding in all directions at an astonishing rate during the last few decades. New algorithmic and theoretical techniques have been developed, the diffusion into other disciplines has proceeded at a rapid pace, and our knowledge of all aspects of the field has grown even more profound. At the same time, one of the most striking trends in optimization is the constantly increasing emphasis on the interdisciplinary nature of the field. Optimization has been a basic tool in all areas of applied mathematics, engineering, medicine, economics and other sciences.

The series *Optimization and Its Applications* publishes undergraduate and graduate textbooks, monographs and state-of-the-art expository works that focus on algorithms for solving optimization problems and also study applications involving such problems. Some of the topics covered include nonlinear optimization (convex and nonconvex), network flow problems, stochastic optimization, optimal control, discrete optimization, multiobjective programming, description of software packages, approximation techniques and heuristic approaches.

VARIATIONAL ANALYSIS AND AEROSPACE ENGINEERING

Edited By

GIUSEPPE BUTTAZZO

University of Pisa, Italy

ALDO FREDIANI

University of Pisa, Italy

Editors

Giuseppe Buttazzo
Università di Pisa
Dipto. Matematica
Largo B. Pontecorvo, 5
56127 Pisa, Italy
buttazzo@dm.unipi.it

Aldo Frediani
Università di Pisa
Dipto. Ingegneria Aerospaziale
Via Diotisalvi, 2
56126 Pisa, Italy
a.frediani@ing.unipi.it

ISSN 1931-6828

ISBN 978-0-387-95856-9

e-ISBN 978-0-387-95857-6

DOI 10.1007/978-0-387-95857-6

Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009929316

AMS Subject Classifications (2000): 76-06, 79-06

© Springer Science+Business Media, LLC 2009

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

This book is dedicated to Angelo Miele on the occasion of his 85th birthday.

“This page left intentionally blank.”

Preface

In recent years, new mathematical methods and tools have been developed and applied extensively in the field of aerospace engineering, for example, finite element method, computational fluid dynamics, optimization, control, eigenvalues problems. The interaction between aerospace engineering and mathematics has been significant in the past for both engineers and mathematicians and will be even stronger in the future.

The School of Mathematics “Guido Stampacchia” of the “Ettore Majorana” Foundation and Centre of Scientific Culture is the most appropriate site for aerospace engineers and mathematicians to meet. The present volume collects the papers presented at the Erice Workshop held on September 8–16, 2007, which was organized in order to allow aerospace engineers and mathematicians from Universities, Research Centres, and Industry to debate advanced problems in aerospace engineering requiring extensive mathematical applications.

The editors are confident to capture the interest of people from both academia and industry, particularly, young researchers working on new frontiers of mathematical applications to engineering.

The workshop was dedicated to Angelo Miele, Professor at Rice University in Houston, on the occasion of his 85th birthday. Angelo Miele is both an eminent mathematician and a famous engineer, among other activities, able to conceive new scenarios for space exploration. He has been the advisor of many PhD students at Houston, who became well-known professors in universities worldwide and are speakers at this workshop.

Pisa,
July 2008

Giuseppe Buttazzo, Pisa (Italy)
Aldo Frediani, Pisa (Italy)

“This page left intentionally blank.”

Acknowledgments

This volume collects the contributions presented in the workshop on “Variational Analysis and Aerospace Engineering,” held in Erice on September 8–16, 2007. The workshop as well as the preparation of this volume have been possible thanks to the contributions of the following organisations and individuals:

- Università di Pisa, Pisa, Italy
- GNAMPA (Gruppo Nazionale per l’Analisi Matematica, la Probabilità e le loro Applicazioni), Italy
- Dipartimento di Ingegneria Aerospaziale di Pisa, Pisa, Italy
- Technical University of Technology, Delft, Holland
- Contessa Maria Fede Caproni, Roma, Italy
- IDS, Pisa, Italy
- AgustaWestland, Cascina Costa, Italy
- Fondazione Cassa di Risparmio di Pisa, Pisa, Italy
- Consorzio Etruria SCARL, Montelupo, Firenze, Italy

We gratefully acknowledge the E. Majorana Centre and Foundation for Scientific Culture and the precious help by Franco Giannessi and Emanuele Rizzo.

“This page left intentionally blank.”

Contents

1	Algorithm Issues and Challenges Associated with the Development of Robust CFD Codes	1
	Steven R. Allmaras, John E. Bussoletti, Craig L. Hilmes, Forrester T. Johnson, Robin G. Melvin, Edward N. Tinoco, Venkat Venkatakrishnan, Laurence B. Wigton and David P. Young	
1.1	Introduction	1
1.2	Algorithm Issues Related to the Solution of the Navier–Stokes Equations	2
1.2.1	Grid Adaption and Error Estimation	3
1.2.2	Discretization Issues	6
1.2.3	Higher Order Elements	12
1.2.4	Domain Decomposition and Linear Solver	16
1.3	Conclusions	19
	References	19
2	Flight Path Optimization at Constant Altitude	21
	Mark D. Ardema and Bryan C. Asuncion	
2.1	Introduction	21
2.2	Singular Optimal Control	23
2.3	The Cruise Problem	24
2.4	Fanjet Specific Fuel Consumption	26
2.5	An Example	28
2.6	Conclusions and Discussion	31
	References	32
3	A Survey on the Newton Problem of Optimal Profiles	33
	Giuseppe Buttazzo	
3.1	Introduction	33
3.2	Radially Symmetric Profiles	37
3.3	The Existence Result	40
	References	47

4	Innovative Rotor Blade Design Code	49
	Vittorio Caramaschi and Claudio Monteggia	
4.1	Introduction	50
4.2	Helicopter's Aeromechanics Outlines	51
4.3	Helicopter's Rotor Mathematical Model Features and Aeromechanics Codes Worldwide Status	56
4.4	AW Aeromechanics Code GYROX II	57
4.4.1	General Procedure	58
4.4.2	Rotor Hub Modelling Features	59
4.4.3	Pylon Modelling Features	61
4.4.4	Rotor Blade Structural Modelling Features	62
4.4.5	Rotor Aerodynamics	63
4.4.6	Solution Algorithms	66
4.4.7	Operational Main Features and Output Data	67
4.5	Applications	68
4.6	Conclusions	73
4.6.1	Short Term	73
4.6.2	Long Term	74
5	Fields of Extremals and Sufficient Conditions for the Simplest Problem of the Calculus of Variations in n-Variables	75
	Dean A. Carlson and George Leitmann	
5.1	Introduction	75
5.2	Notations and the Problem Definition	76
5.3	Leitmann's Direct Method	78
5.4	Fields of Extremals	80
5.5	Sufficient Conditions for Optimality	84
5.6	Conclusion	88
	References	88
6	A Framework for Aerodynamic Shape Optimization	91
	Giampiero Carpentieri and Michel J.L. van Tooren	
6.1	Introduction	91
6.2	Adjoint-Based Sensitivity Analysis	92
6.3	Optimization Framework	93
6.3.1	Flow Solver	94
6.3.2	Adjoint Solver	96
6.3.3	Shape Parameterization	98
6.3.4	Geometric Sensitivities	99
6.3.5	Optimization Algorithm	99
6.4	Optimization Test Cases	100
6.4.1	RAE2822 at $M_\infty = 0.73$ and $\alpha = 2^\circ$	100
6.4.2	NACA64A410 at $M_\infty = 0.75$ and $\alpha = 0^\circ$	101
6.4.3	NACA0012 at $M_\infty = 1.5$ and $\alpha = 2^\circ$	102
6.4.4	ONERA-M6 wing at $M_\infty = 0.84$ and $\alpha = 3.06^\circ$	103

6.5	Conclusions	105
	References	106
7	Optimal Motions of Multibody Systems in Resistive Media	107
	Felix L. Chernousko	
7.1	Introduction	107
7.2	Basic Equations	108
7.3	Linear Resistance	110
7.4	Relative Motions	110
7.5	Piecewise Linear Resistance	112
7.6	Quadratic Resistance	113
7.7	Dry Friction: Velocity-Control Motion	114
7.8	Dry Friction: Acceleration-Control Motion	120
7.9	Generalizations	124
7.10	Experiments	124
7.11	Conclusions	125
	References	125
8	Instationary Heat-Constrained Trajectory Optimization of a Hypersonic Space Vehicle by ODE–PDE-Constrained Optimal Control	127
	Kurt Chudej, Hans Josef Pesch, Markus Wächter, Gottfried Sachs and Florent Le Bras	
8.1	Introduction	128
8.2	Trajectory Optimization Problems with Active Cooling	130
8.3	Trajectory Optimization Problem with an Instationary Heat Constraint	134
8.4	Conclusions	140
	References	142
9	Variational Approaches to Fracture	145
	Gianpietro Del Piero	
9.1	Fracture as a Minimum Problem	145
9.2	The Numerical Solution	147
9.3	Energy Barriers and Local Minima	148
9.4	Barenblatt’s Regularization	151
9.5	Two Solution Strategies	153
9.6	The Dissipative Model	154
9.7	From Surface to Bulk Regularization	157
	References	161
10	On the Problem of Synchronization of Identical Dynamical Systems: The Huygens’s Clocks	163
	Rui Dilão	
10.1	Introduction	163
10.2	A Model for the Synchronization of the Two Pendulum Clocks ...	166

10.3	A Simple Clock Model	168
10.4	Synchronization of Two Pendulum Clocks with Equal Parameters .	169
10.5	Synchronization of Two Pendulum Clocks with Different Parameters: Robustness	178
10.6	Conclusions	179
	References	180
11	Best Wing System: An Exact Solution of the Prandtl's Problem	183
	Aldo Frediani and Guido Montanari	
11.1	Introduction	183
11.2	The Induced Drag for Lifting Multiwing Systems	184
11.3	The Problem of Minimum Induced Drag in a Box Wing	187
11.3.1	Case A: Elliptical Circulations on the Horizontal Wings and Zero on the Vertical Ones	191
11.3.2	Case B: Constant Circulations on the Horizontal Wings and Unknown on the Vertical Ones	192
11.3.3	Final Equations	194
11.4	The Optimum Lift Distribution Along the Vertical Wings	196
11.5	Results and Conclusions	197
	References	199
12	Numerical Simulation of the Dynamics of Boats by a Variational Inequality Approach	213
	Luca Formaggia, Edie Miglio, Andrea Mola and Anna Scotti	
12.1	Introduction	213
12.2	A Variational Approach to the Floating Body Problem	214
12.2.1	Characteristic Treatment of the Time Derivative	218
12.2.2	Enforcing the Constraint in the Hydrostatic Step	219
12.2.3	The Model for the Dynamics of a Rowing Scull	220
12.2.4	More Realistic Boundary Conditions	223
12.3	The Interaction Between the Boat and the Water	223
12.4	Numerical Results	224
12.4.1	Sinking and Pitching Motions	224
12.4.2	Reproducing Mean Motion Wave Pattern	225
12.4.3	An Example with the Full Dynamics	226
12.4.4	A Final Detail	226
	References	227
13	Concepts of Active Noise Reduction Employed in High Noise Level Aircraft Cockpits	229
	Hatem Foudhaili and Eduard Reithmeier	
13.1	Passive Versus Active Noise Reduction	230
13.2	Active Noise Cancellation	230
13.3	Active Structural/Acoustic Control (ASAC)	234
13.4	Active Aviation Headsets	237

13.5	An Aviation Communication Headset Prototype with Digital Adaptive Noise Reduction	238
13.6	Conclusions	240
	References	240
14	Lekhnitskii's Formalism for Stress Concentrations Around Irregularities in Anisotropic Plates: Solutions for Arbitrary Boundary Conditions	243
	Sotiris Koussios and Adriaan Beukers	
14.1	Introduction	243
14.2	Governing Equations	245
14.3	General Solution	246
14.4	Stress, Strain, and Displacements Formulation	247
14.5	Formulation of Boundary Conditions	248
	14.5.1 Forces	248
	14.5.2 Displacements	249
14.6	Solution Strategy	250
	14.6.1 Series Representation of the Boundary Conditions	250
	14.6.2 Transformation into a Single Variable	251
14.7	Boundary Conditions Evaluation	253
	14.7.1 Homogeneous Part	253
	14.7.2 Logarithmic Part	254
	14.7.3 Disturbance Field	256
14.8	Evaluation of Stresses and Displacements	259
14.9	Example	261
14.10	Conclusions	264
	References	265
15	Best Initial Conditions for the Rendezvous Maneuver	267
	Angelo Miele and Marco Ciarcià	
15.1	Introduction	268
15.2	Algorithm	269
15.3	System Description	271
	15.3.1 Multiple-Subarc Equations	272
	15.3.2 Inequality Constraint	273
	15.3.3 Particular Cases	274
	15.3.4 Boundary Conditions	274
	15.3.5 Performance Index	275
	15.3.6 Approaches	276
15.4	Minimum Fuel, Time Free	276
15.5	Results	277
15.6	Minimum Fuel, Time Given	280
	15.6.1 Results	282
15.7	Conclusions	287
	References	288

16	Commercial Aircraft Design for Reduced Noise and Environmental Impact	291
	S. Mistry, Howard Smith, and John P. Fielding	
16.1	Introduction	292
16.2	Simple Emission Trade-Off Study	292
16.2.1	Global Warming Costs	292
16.2.2	Noise Costs	293
16.2.3	Local Air Quality Cost (LAQ)	293
16.2.4	Annual Fuel Costs Fro Baseline Aircraft	294
16.2.5	Baseline Aircraft Environmental Costs	294
16.2.6	Summary of Trade-Offs	295
16.3	Aircraft Designs for Reduced Noise	295
16.3.1	Background	295
16.3.2	Baseline Aircraft Design and Noise Prediction	296
16.3.3	Low Airframe Noise Design Methodology	297
16.3.4	Low-Noise Aircraft Concept Brainstorming Process	297
16.3.5	Broad Delta Concepts	299
16.3.6	Airframe Approach Noise Prediction	302
16.3.7	Performance Comparison	303
16.4	The Cranfield A-6 Greenliner Project	304
16.4.1	Group Design Project Activities	304
16.4.2	Greenliner Description	305
16.4.3	Predicted Performance for the Greenliner	309
16.5	Conclusions	311
	References	312
17	Variational Approach to the Problem of the Minimum Induced Drag of Wings	313
	Maria Teresa Panaro, Aldo Frediani, Franco Giannessi and Emanuele Rizzo	
17.1	Introduction	314
17.2	Finite Span Wings	314
17.3	Problem of Minimum Induced Drag of a Straight Wing: An optimality condition	316
17.4	Duality: A New Approach to the Design of Wings	319
17.5	Direct Methods	325
17.5.1	Elliptic Distribution	325
17.5.2	Ritz Method	327
	References	342
18	Plastic Hinges in a Beam	343
	Danilo Percivale and Franco Tomarelli	
18.1	Elastic–Plastic Beam	343
18.2	Skew-Symmetric Load	347
	References	348

19 Problems of Minimal and Maximal Aerodynamic Resistance	349
Alexander Plakhov	
19.1 Introduction	349
19.2 Translational Motion	350
19.3 Translational Motion with Rotation: Two-Dimensional Case	355
19.3.1 Definition of Rough Body and Main Theorems	355
19.3.2 Problems of Minimal and Maximal Resistance for a Slowly Rotating Body	358
19.3.3 Mathematical Retroreflector	360
19.3.4 Effect of Magnus	361
References	365
20 Shock Optimization for Airfoil Design Problems	367
Olivier Pironneau	
20.1 Numerical Optimal Shape Design	367
20.1.1 An Academic Problem	367
20.1.2 Sensitivity Analysis	368
20.1.3 Conceptual Algorithm	369
20.2 Automatic Differentiation	370
20.2.1 Principle of Automatic Differentiation	370
20.2.2 Example of Application	371
20.3 Differentiability Issues	372
20.3.1 Extended Calculus of Variation	372
20.3.2 Sensitivity Analysis for Burgers' Equation	373
20.3.3 Application to Optimal Control	373
20.3.4 A Simple Example	374
20.3.5 Right and Wrong Schemes	374
20.4 Small Disturbances and Automatic Differentiations	376
References	377
21 Differential Games Treated by a Gradient-Restoration Approach	379
Mauro Pontani	
21.1 Introduction	379
21.2 Zero-Sum Differential Games	380
21.3 Numerical Solution of Two-Sided Optimization Problems	382
21.3.1 Transformation into Single-Objective Problem	382
21.3.2 Sequential Gradient-Restoration Algorithm	384
21.4 Homicidal Chauffeur Game	385
21.4.1 Formulation of the Problem	385
21.4.2 Method of Solution	386
21.4.3 Numerical Results	387
21.5 Orbital Pursuit-Evasion Game	388
21.5.1 Formulation of the Problem	389
21.5.2 Method of Solution	390
21.5.3 Numerical Results	392
21.6 Conclusions	395
References	395

22	Interval Methods for Optimal Control	397
	Andreas Rauh and Eberhard P. Hofer	
22.1	Introduction	398
22.2	Optimal and Robust Control of Dynamical Systems	399
22.2.1	Optimal Control of Discrete- and Continuous-Time Processes	400
22.2.2	Specification of Robustness in the Time Domain	401
22.2.3	Optimality Criteria for Systems with Uncertainties	402
22.3	Interval Arithmetic Optimization Algorithm	403
22.4	Parallelization of the Optimization Algorithm	405
22.5	Combination with Classical Controller Design	406
22.6	Validated Modeling and Simulation of Dynamical Systems with State-Dependent Switchings	407
22.7	Optimization Results	410
22.7.1	Interval Algorithm for Structure Optimization	410
22.7.2	Linear State Controller for Improvement of Robustness	413
22.7.3	Interval Algorithm for Parameter Optimization	415
22.8	Conclusions and Outlook on Future Work	416
	References	417
23	Application of Optimisation Algorithms to Aircraft Aerodynamics	419
	Emanuele Rizzo and Aldo Frediani	
23.1	Introduction	419
23.2	An Algorithm for the Search of Global Minima	424
23.3	Test Cases	428
23.3.1	Test Case 1 (Unconstrained): Ackley's Function	428
23.3.2	Test Case 2 (Unconstrained): Rastrigin's Function	431
23.3.3	Test Case 3 (Unconstrained): Rosenbrock's Function	431
23.3.4	Test Case 4 (Unconstrained): Schwefel's Function	432
23.4	The AEROSTATE Program: An Application to Aeronautics	434
23.4.1	Minimum Induced Drag of a Wing	435
23.4.2	Minimum Total Drag of a Wing	438
23.4.3	The Trimmed Aircraft	439
23.4.4	The PrandtlPlane	441
23.5	Conclusions	445
	References	445
24	Different levels of Optimisation in Aircraft Design	447
	Dieter Schmitt	
24.1	Air Transport System	448
24.2	Industrial Process of Aircraft Design	449
24.3	Different Levels of Aircraft Design vs. Development Phases	452
24.4	Tools Used in Different Phases	455
24.5	Conclusion	459
	References	459

25	Numerical and Analytical Methods for Global Optimization	461
	Paolo Teofilatto and Mauro Pontani	
25.1	Introduction	461
25.2	Green's Theorem Approach	463
25.3	Morse Theory Approach	469
25.4	Final Comments	474
	References	474
26	The Aeroservoelasticity Qualification Process in Alenia	477
	Vincenzo Vaccaro	
26.1	Introduction	477
26.2	Company Presentation	478
26.3	What Is Aeroelasticity	479
26.4	Aeroelastic Tradition in Alenia	480
26.5	Aeroservoelastic Certification Process	481
26.5.1	Analytical Models	482
26.5.2	Theoretical Background	484
26.5.3	Ground Test	486
26.5.4	Flight Test	486
26.5.5	Research and Future Developments	486
27	Further Steps Towards Quantitative Conceptual Aircraft Design	491
	Michel van Tooren, Gianfranco La Rocca and Teodor Chiciudean	
27.1	Introduction	491
27.2	The Systems Engineering Approach	496
27.3	Requirements on Computational Systems	496
27.4	The Design and Engineering Engine Concept	497
27.4.1	Describing Design Options	497
27.4.2	The Initiator	501
27.4.3	The Multi-model Generator	503
27.4.4	The Life-Cycle Analysis with Expert Tools	504
27.4.5	The Converger/Evaluator	504
27.4.6	The Agent-Based Framework	504
27.5	Results and Discussion	505
27.6	Conclusions	507
	References	508
28	Some Plebeian Variational Problems	509
	Piero Villaggio	
28.1	Introduction	509
28.2	Mechanical Plebeian Problems	510
28.3	Locomotion	513
28.4	Peeling and Cooking	515
28.5	Conclusions	518
	References	518

“This page left intentionally blank.”

Contributors

Steven R. Allmaras
The Boeing Company, Seattle, WA, USA
e-mail: steven.r.allmaras@boeing.com

Mark D. Ardema
Department of Mechanical Engineering, Santa Clara University, Real Santa Clara,
CA 95053, USA
e-mail: mardema@scu.edu

Bryan C. Asuncion
Department of Mechanical Engineering, Santa Clara University, Real Santa Clara,
CA 95053, USA
e-mail: basuncion@scu.edu

Adriaan Beukers
Delft University of Technology, Delft, The Netherlands
e-mail: a.beukers@tudelft.nl

John E. Bussioletti
The Boeing Company, Seattle, WA, USA
e-mail: john.e.bussioletti@boeing.com

Giuseppe Buttazzo
Dipartimento di Matematica, “L. Tonelli”, Università di Pisa, Pisa, Italy
e-mail: buttazzo@dm.unipi.it

Vittorio Caramaschi
AgustaWestland, Cascina Costa, VA, USA
e-mail: vittorio.caramaschi@agustawestland.com

Dean A. Carlson
Mathematical Reviews, American Mathematical Society, Ann Arbor, MI, USA
e-mail: dac@ams.org

Giampiero Carpentieri
Delft University of Technology, Delft, The Netherlands
e-mail: g.carpentieri@tudelft.nl

Felix L. Chernousko
Institute for Problems in Mechanics, Russian Academy of Sciences, Moscow,
Russia
e-mail: chern@ipmnet.ru

Teodor Chiciudean
Delft University of Technology, Delft, The Netherlands
e-mail: t.g.chiciudean@tudelft.nl

Kurt Chudej
Lehrstuhl für Ingenieurmathematik, Universität Bayreuth, Bayreuth, Germany
e-mail: kurt.chudej@uni-bayreuth.de

Marco Ciarcià
PhD Candidate, Aero-Astronautics Group, Rice University, Houston, TX, USA
e-mail: ciarcia@rice.edu

Rui Dilão
NonLinear Dynamics Group, Instituto Superior Técnico, Lisbon, Portugal
e-mail: rui@sd.ist.utl.pt

John P. Fielding
Department of Aerospace Engineering, Cranfield University, Cranfield,
Bedfordshire, UK
e-mail: j.p.fielding@cranfield.ac.uk

Luca Formaggia
MOX, Mathematics Department, Politecnico di Milano, Milan, Italy
e-mail: luca.formaggia@polimi.it

Hatem Foudhaili
Institute of Measurement and Automatic Control, Leibniz Universität, Hannover,
Germany
e-mail: hatem.foudhaili@imr.uni-hannover.de

Aldo Frediani
Dipartimento di Ingegneria Aerospaziale, “L. Lazzarino”, Università di Pisa, Pisa,
Italy
e-mail: a.frediani@ing.unipi.it

Franco Giannessi
Dipartimento di Matematica, “L. Tonelli”, Università di Pisa, Pisa, Italy
e-mail: gianness@dm.unipi.it

Craig L. Hilmes
The Boeing Company, Seattle, WA USA
e-mail: craig.l.hilmes@boeing.com

Eberhard P. Hofer
Institute of Measurement, Control, and Microtechnology, University of Ulm,
Ulm, Germany
e-mail: eberhard.hofer@uni-ulm.de

Forrester T. Johnson
The Boeing Company, Seattle, WA USA
e-mail: forrester.t.johnson@boeing.com

Sotiris Koussios
Delft University of Technology, Delft, The Netherlands
e-mail: s.koussios@tudelft.nl

Gianfranco La Rocca
Delft University of Technology, Delft, The Netherlands
e-mail: g.larocca@tudelft.nl

Floren Le Bras
Laboratoire de Recherches Balistiques et Aerodynamiques, Delegation Generale
pour l'Armement, Vernon, France
e-mail: florent.le-bras@polytechnique.org

George Leitmann
Department of Mechanical Engineering, University of California at Berkeley,
Berkeley, CA, USA
e-mail: gleit@berkeley.edu

Robin G. Melvin
The Boeing Company, Seattle, WA, USA
e-mail: robin.melvin@redwood.rt.cs.boeing.com

Angelo Miele
Aero-Astronautics Group, Rice University, Houston, TX, USA
e-mail: miele@rice.edu

Edie Miglio
MOX, Mathematics Department, Politecnico di Milano, Milan, Italy
e-mail: edie.miglio@polimi.it

S. Mistry
Department of Aerospace Engineering, Cranfield University, Cranfield,
Bedfordshire, UK
e-mail: s.mistry.2003@cranfield.ac.uk

Andrea Mola
MOX, Mathematics Department, Politecnico di Milano, Milan, Italy
e-mail: andrea.mola@polimi.it

Guido Montanari
Dipartimento di Matematica, "L. Tonelli", Università di Pisa, Pisa, Italy
e-mail: a.frediani@ing.unipi.it

Claudio Monteggia
AgustaWestland, Cascina Costa, VA, USA
e-mail: claudio.monteggia@agustawestland.com

Maria Teresa Panaro

Dipartimento di Matematica, “L. Tonelli”, Università di Pisa, Pisa, Italy

e-mail: mt.panaro@gmail.com

Danilo Percivale

Dipartimento di Ingegneria della Produzione, Università di Genova, Genova, Italy

e-mail: percival@dimet.unige.it

Hans Josef Pesch

German Institute of Science and Technology, Singapore

e-mail: hans-josef.pesch@uni-bayreuth.de

Olivier Pironneau

Laboratoire Jacques-Louis Lion, University of Paris VI &

IUF Paris, France

e-mail: pironneau@ann.jussieu.fr

Alexander Plakhov

University of Wales – Aberystwyth, Aberystwyth, UK on leave from Aveiro

University, Portugal

e-mail: axp@aber.ac.uk

Mauro Pontani

Scuola di Ingegneria Aerospaziale, University of Rome “La Sapienza,” Rome, Italy

e-mail: mauro.pontani@uniroma1.it

Andreas Rauh

Chair of Mechatronics, University of Rostock, D-18059 Rostock, Germany

e-mail: Andreas.Rauh@uni-rostock.de

Eduard Reithmeier

Institute of Measurement and Automatic Control, Leibniz Universitaet, Hannover, Germany

e-mail: eduard.reithmeier@imr.uni-hannover.de

Emanuele Rizzo

Dipartimento di Ingegneria Aerospaziale, “L. Lazzarino”, Università di Pisa, Pisa, Italy

e-mail: emanuele.rizzo@ing.unipi.it

Gottfried Sachs

Lehrstuhl für Flugmechanik und Flugregelung, Technische Universität München, München, Germany

e-mail: sachs@lfm.mw.tum.de

Dieter Schmitt

Airbus, Blagnac, France

e-mail: dieter.schmitt@airbus.com

Anna Scotti

MOX, Mathematics Department, Politecnico di Milano, Milan, Italy

e-mail: anna.scotti@mail.polimi.it

Howard Smith

Department of Aerospace Engineering, Cranfield University, Cranfield,
Bedfordshire, UK

e-mail: howard.smith@cranfield.ac.uk

Paolo Teofilatto

Scuola di Ingegneria Aerospaziale, University of Rome “La Sapienza”, Rome, Italy

e-mail: paolo.teofilatto@uniroma1.it

Edward N. Tinoco

The Boeing Company, Seattle, WA, USA

e-mail: edward.n.tinoco@boeing.com

Franco Tomarelli

Dipartimento di Matematica, Politecnico di Milano, Milano, Italy

e-mail: franco.tomarelli@polimi.it

Michel van Tooren

Delft University of Technology, Delft, The Netherlands

e-mail: m.j.l.vantooren@tudelft.nl

Venkat Venkatakrishnan

The Boeing Company, Seattle, WA, USA

e-mail: venkat.venkatakrishnan@boeing.com

Piero Villaggio

Dipartimento di Ingegneria Strutturale, Università di Pisa, Pisa, Italy

e-mail: dis@ing.unipi.it

Markus Wachter

German Institute of Science and Technology, Singapore

e-mail: markus.waechte@gist.edu.sg

Laurence B. Wigton

The Boeing Company, Seattle, WA, USA

e-mail: laurence.b.wigton@boeing.com

David P. Young

The Boeing Company, Seattle, WA, USA

e-mail: david.p.young@boeing.com

Chapter 1

Algorithm Issues and Challenges Associated with the Development of Robust CFD Codes

Steven R. Allmaras, John E. Bussioletti, Craig L. Hilmes, Forrester T. Johnson,
Robin G. Melvin, Edward N. Tinoco, Venkat Venkatakrishnan,
Laurence B. Wigton and David P. Young

Dedication We dedicate this chapter with thanks to Professor Angelo Miele whose student Gary Saaris made many important contributions to the development and application of CFD at Boeing and whose enthusiasm during several meetings at the Boeing Scientific Research Laboratories 44 years ago inspired Forrester Johnson to learn and apply optimization techniques to the design of Boeing vehicles.

Abstract Over the next 20 years, Boeing will likely develop, manufacture, sell, and support many thousands of vehicles that fly. During this period, Boeing project aerodynamicists need access to tools that accurately predict and confirm vehicle flight characteristics. Thirty years ago, these tools consisted almost entirely of analytic approximation methods, wind tunnel tests, and flight tests. With the development of increasingly powerful computers, numerical simulations of various approximations to the Navier–Stokes equations have begun supplementing these tools. Collectively, these numerical simulation methods have become known as computational fluid dynamics (CFD). This chapter describes the algorithm issues and challenges associated with the development of reliable Navier–Stokes codes that can be used by a wide variety of project engineers who do not necessarily have a deep background in numerical methods.

1.1 Introduction

This year we project that project engineers at Boeing Commercial Airplanes will run more than 50,000 computational fluid dynamics (CFD) cases on Boeing's PC clusters and Cray supercomputer. Most of these cases involve complex physics and complicated geometries. Clearly CFD has joined the wind tunnel and flight test

Steven R. Allmaras, John E. Bussioletti, Craig L. Hilmes, Forrester T. Johnson, Robin G. Melvin, Edward N. Tinoco, Venkat Venkatakrishnan, Laurence B. Wigton, David P. Young
The Boeing Company, Seattle, WA, USA

as primary flow analysis tools of the trade. CFD is now acknowledged to provide substantial value and has created a paradigm shift in the vehicle design, analysis, and support processes. This paradigm shift as well as the history of CFD development and use at Boeing was the main topic of [1]. In particular, Boeing's workhorse cruise configuration aerodynamic analysis and optimization code, TRANAIR, was described in detail. TRANAIR is a directly coupled full potential/boundary layer code which employs a solution-adapted rectangular grid. During the 1990s the limitation of full potential/boundary layer-coupled codes to the simulation of flows without significant flow separation led to the development and application of solutions to the Reynolds-averaged Navier–Stokes equations (RANS). At Boeing, various Navier–Stokes codes including CFL3D, Overflow, and CFD++ increasingly have been applied to the analysis of flows with mild to modest separation. The successful application of Navier–Stokes codes during the last 10 years has raised expectations among Boeing engineers that CFD can become a routine tool for the load analysis, stability and control analysis, and high-lift design processes. In fact, there is considerable thought that it is now feasible to populate databases involving tens of thousands of cases with results from Navier–Stokes CFD codes. However, before Navier–Stokes codes can routinely be used to populate databases, accuracy, reliability, efficiency, and usability issues need to be addressed. Gaps in data, inconsistent data, and long acquisition times seriously degrade the utility of a database. Even with current user aids, the application of Navier–Stokes codes to new configurations generally requires the services of an expert user.

1.2 Algorithm Issues Related to the Solution of the Navier–Stokes Equations

The generation of a “good grid” is still somewhat of an art and often quite labor intensive. Although everyone realizes that a good grid is necessary for accuracy and even convergence, there is no precise definition of what constitutes such a grid. In fact, the definition would probably vary from code to code and is certainly case dependent. Grid generation difficulties are reflected in the fact that although Navier–Stokes codes are now considered capable of generating more accurate results in a number of flow regimes, they are used far less frequently than TRANAIR at Boeing Commercial Airplanes. From our TRANAIR experience, it seems rather evident that solution-adaptive grids must be an essential feature for reliability and usability. This is especially true when computing flows at off-design conditions where our understanding of the flow physics is limited, making it difficult to generate suitable grids. However, these grids must now be anisotropic and, more than likely, quite irregular and unstructured. Simply generating and refining such grids is a rather complicated task, especially in three dimensions. Moreover, the routine use of adapted unstructured grids places a huge burden on improving discretization fidelity, as current discretization algorithms do not seem to do well with irregular spacings and cell shapes. TRANAIR suffered from this problem only in boundary cut cells and

here the scalar nature of the full potential equation allowed the design of a relatively robust discretization. On the plus side, the recent development of adjoint technology for predicting solution error offers the potential for a much more systematic and efficient grid refinement procedure than the heuristic adaptive refinement employed in TRANAIR.

Another issue for most Navier–Stokes solvers is nonlinear convergence. Guidelines requiring a certain number of digits of error reduction are not particularly reliable, especially when considering new configurations. In some cases we have seen drag and moment values change significantly at about seven to eight orders of magnitude reduction in residual error norms. Current solvers are relatively weak, which makes convergence slow and uncertain. When a CFD code does not seem to converge to a steady state, it is never clear whether the flow is inherently unsteady or whether the solver technology is just not powerful enough for the case under consideration. In many instances, the latter is the case. The use of Newton’s method combined with a powerful sparse solver preconditioner has allowed TRANAIR to achieve a convergence rate (nonlinear residuals driven to machine error) for project usage of over 97%. We feel the ability to drive residuals to machine error is critical for the Navier–Stokes equations as well.

Solutions involving complex flows often involve complex geometries as well, e.g., vortex generators, chines, spoilers, landing gear. Often 30–100 million grid points are now required to achieve reasonable accuracy. Problems of this size must be addressed using domain partitioning in conjunction with powerful global preconditioners. Higher order elements are certainly desirable for efficiency and for capturing latent features. However, stabilization and limiter technologies need to be advanced to handle such elements.

Boeing has for some time been involved in an effort to evaluate mainstream Navier–Stokes solution technology and explore alternatives and improvements, much the same as was done years ago in the case of potential flow. The project is called General Geometry Navier–Stokes Solver (GGNS) which is a joint effort between the CFD developers at Boeing and their colleagues at the Boeing Technical Research Center in Moscow. At the current time the project is considering steady-state RANS solutions. (This choice is a compromise between physical modeling and computational affordability in the production engineering environment. Detached Eddy simulation would be preferable from a physical modeling standpoint, particularly for cases of significant separation, but its computational requirements are likely factors of hundreds, if not thousands, that of steady RANS. We accept the fact that the limits of applicability for steady RANS will need to be determined through engineering practice in a manner similar to the learning process used with TRANAIR.) We now discuss in detail some of the technology directions and algorithm issues and challenges involved in the GGNS development effort.

1.2.1 Grid Adaption and Error Estimation

From our experience with TRANAIR grid adaption seems essential. The accuracy, nonlinear convergence, and grid generation benefits are too valuable to dismiss in

a production environment where thousands of simulations are carried out. With respect to grid generation benefits, the initial grid need only fill space and fit the vehicle geometry or at least the vehicle topology. The TRANAIR rectangular, cut cell, hierarchical grid approach is not suitable for Navier–Stokes flows involving arbitrarily oriented shear layers. However, through hierarchical grid refinement, it can resolve arbitrary geometrical features and can therefore be used to generate an initial grid if the cut cells can be divided into compatible convex sub-cells of simple shapes. Of course, this involves grid generation, but with fewer global topology issues. Alternatively, one can also use a number of advancing front grid generators, although we are not entirely satisfied with their robustness in the case of complex configurations. For solution purposes we prefer to use simplices as cells (triangles in 2-D and tetrahedrons in 3-D with possibly curved edges) due to the ease of implementation and stability analysis. However, the definition of the adapted grid can and should involve more complex cell shapes so long as they can be cut into compatible simplices when delivered to the solver.

The anisotropic hierarchical adaptive (AHA) grid technique, developed by Sergei Medvedev, Alex Martynov, Tamara Ivanova, and Sergei Degtyarev of the Keldysh Institute of Applied Mathematics, is an adaptive grid refinement scheme based on a concept of “macro cells” (defined by the initial grid). The refinement process proceeds by insertion of (hanging) nodes on edges of macro cells driven by edge-based error indicators. A subsequent tessellation process transforms each macro cell into a valid triangle (2-D) or tetrahedral (3-D) grid. The tessellation process involves edge/face swapping to achieve alignment with gradients in the error estimator and may also invoke isotropic refinement of the macro cell if there is no clear anisotropy in the error indicator gradients. A reference geometry is used to provide boundary definition at a higher resolution than offered by the initial grid and there are some “visibility conditions” that the initial grid must satisfy to allow the refinement process to successfully improve the boundary resolution, if needed. In 2-D the scheme has proven to be quite reliable and robust. Development of its extension to 3-D is still in progress.

We are also considering an alternative hierarchical refinement (HIREF) approach. Here all possible binary refinements of each cell are considered for their effect on the error indicator. Optimal cell refinements are then adjusted to be compatible with those of cell neighbors. Even using edge midpoints as new vertices, this type of refinement can ultimately orient cells along shear layers. HIREF is not as efficient as AHA, but the tessellation problem is simpler, especially in 3-D. Both methods allow derefinement, of course, as shocks can move and shear layers can change direction as the grid resolves flow features. An interesting variant periodically revisits the initial grid using the current fine-grid solution to hierarchically generate a new fine grid via function adaption.

Geometry representation for adaptive gridding must be “watertight” and adopt one of three discretizations:

1. Assume the initial grid provides sufficient resolution of the boundaries and forbid refinement of boundary definition.

2. Make use of a surrogate representation of the geometry, typically based on local polynomial approximations to allow a refined description, should it be needed in the solution adaptive gridding process.
3. Make direct use of some analytic representation of the geometry as typically provided by one or another CAD system.

The first approach is the simplest and is the one that has been used in TRANAIR. This does require the user to know in advance what is required in the way of surface resolution prior to obtaining a solution. The third approach would be ideal, but offers a number of difficulties with implementation, partly because of the variability in the analytic representations from different CAD systems and also due to a need to potentially re-tessellate a neighborhood of a cell, rather than just the cell itself. This is manageable in 2-D, based on our experiences with the adaptive schemes, but is an especially difficult issue in 3-D. We are currently pursuing the second option for use in 3-D. In that case, the use of curved cells adjacent to the boundary for the initial grid along with associated maps to straight cells in computational space allows arbitrary refinement without visibility issues.

After some testing we are convinced that the use of multiple adjoint solutions to assess grid convergence and to aid refinement/derefinement decisions must be an important feature of grid adaption. As with TRANAIR, our solver approach uses a powerful sparse solver preconditioner, so adjoints are relatively inexpensive to generate during post-processing. Some integral functional values of interest include lift, drag, and various moments. Estimates of changes in integral function values due to refinement of a given cell are obtained by integrating the adjoint against a change in residual over the cell. This change in residual is the difference between the solution residual and a residual evaluated with higher order reconstructed values on the refined grid. HIREF uses only an adjoint-based error indicator for anisotropic grid adaption whereas AHA uses a combination of adjoint and solution reconstruction-based error estimates.

Reconstruction is a difficult issue itself. Certainly, reconstruction in the presence of discontinuities such as shocks is a problem. However, for smooth flows even sophisticated least squares procedures can have great difficulties in the case of unstructured grids with high aspect ratio cells and wide disparities in local cell size. Near-singular neighbor node locations are not the only problem. Reconstruction in the neighborhood of curved viscous walls is a difficult issue for example. Straight-forward least squares reconstruction significantly underpredicts the normal gradient of tangential velocity. This is primarily due to the high aspect ratio cells and the fact that velocity is zero on the curved wall surface. A similar problem can also occur in curved shear layers out in the field.

In Fig. 1.1 we plot lift coefficient (c_l) versus number of adapted nodes for an airfoil at 2° angle-of-attack and freestream Mach number 0.2. The quantity c_{lq} is the adjoint estimated lift coefficient assuming uniform refinement. For reference, c_{l4x} is the actual lift coefficient for a uniformly refined grid. Clearly adjoint estimated integral quantities can be used to improve accuracy and assess grid convergence.

In Fig. 1.2a we plot an error indicator for the solution about the above airfoil on a handcrafted grid with 30,557 nodes. The error indicator represents the increment

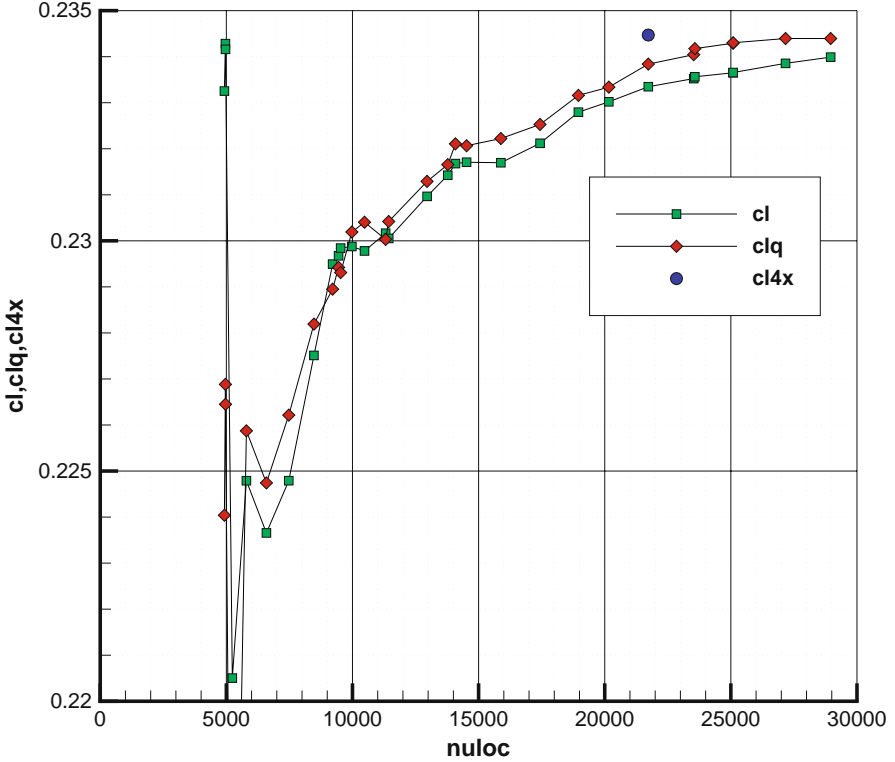


Fig. 1.1 Lift coefficient versus node count

in lift coefficient assuming optimal local cell refinement with HIREF. The computed lift coefficient on this handcrafted grid is 0.2441, which is significantly different from the lift coefficient of 0.2344 computed using an extremely fine grid. The handcrafted grid is especially designed to capture high gradients in the turbulent boundary layer, and it evidently does a good job of that. However, the grid is clearly not adequate in several smooth regions of the flow, where there is no direct indication of a problem. Adding 137 nodes to create an anisotropic adjoint-adapted grid in those regions (Fig. 1.2b) using HIREF yields a much more accurate lift coefficient of 0.2334, which can only be bettered in accuracy by a handcrafted grid containing 64,073 nodes.

1.2.2 Discretization Issues

We believe a discretization of the compressible Navier–Stokes equations augmented with a turbulence transport equation should have the following desirable characteristics:

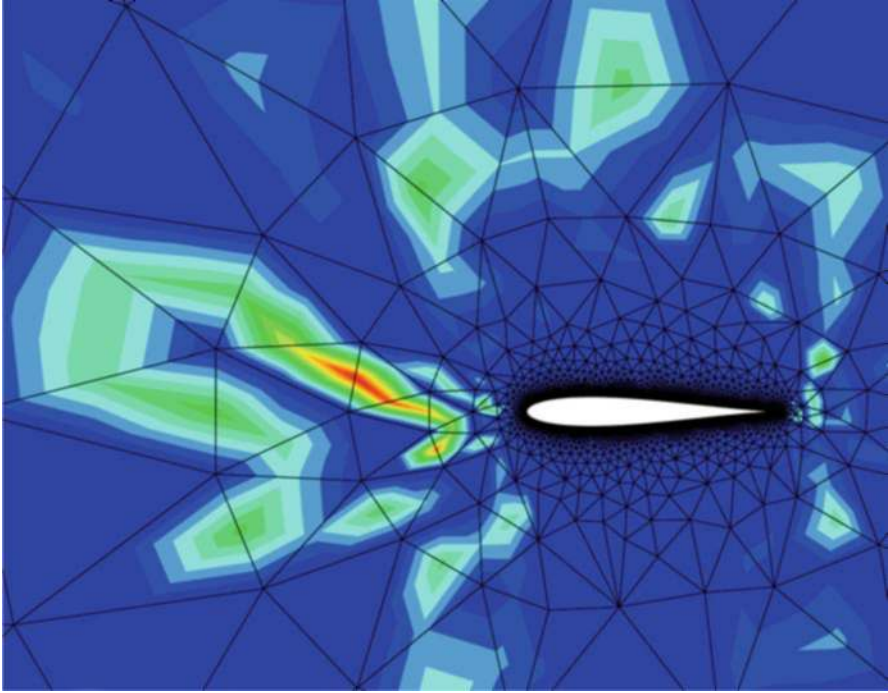


Fig. 1.2a Error indicator for turbulent flow around airfoil at 2° angle-of-attack (30,557-node grid); *red* indicates significant lift increment, and *blue* indicates no increment

1. Assuming an analytic solution exists, a physically reasonable discrete solution should exist on any grid and this solution should approach the analytic solution with grid refinement.
2. It should make minimal use of neighboring information for a given order of accuracy (compact stencil).
3. It should be easily extensible to higher order accuracy (at least for smooth flows).
4. The discrete formulation should satisfy any additional conservation laws or applicable maximum principles at a discrete level, e.g., no entropy production in inviscid flows, total enthalpy conservation in inviscid flows, no non-physical oscillations, etc.

Of these, item 4 is the most difficult to satisfy and is an active subject of research. Item 1 is critical for grid adaption methods, as the initial grids are likely to be quite coarse and it is difficult to proceed to a finer grid without a solution on a given grid. Item 1 is probably essential for fixed grid strategies as well, since to ensure that these grids are adequate everywhere on a new case would involve a prohibitively conservative gridding strategy. Singular or non-physical discretizations on underresolved grids probably account for a significant share of non-converged cases.

Finite volume (FV) schemes popularized in the 1970s and 1980s in CFD circles solve the integral form of the governing equations. These are generalizations

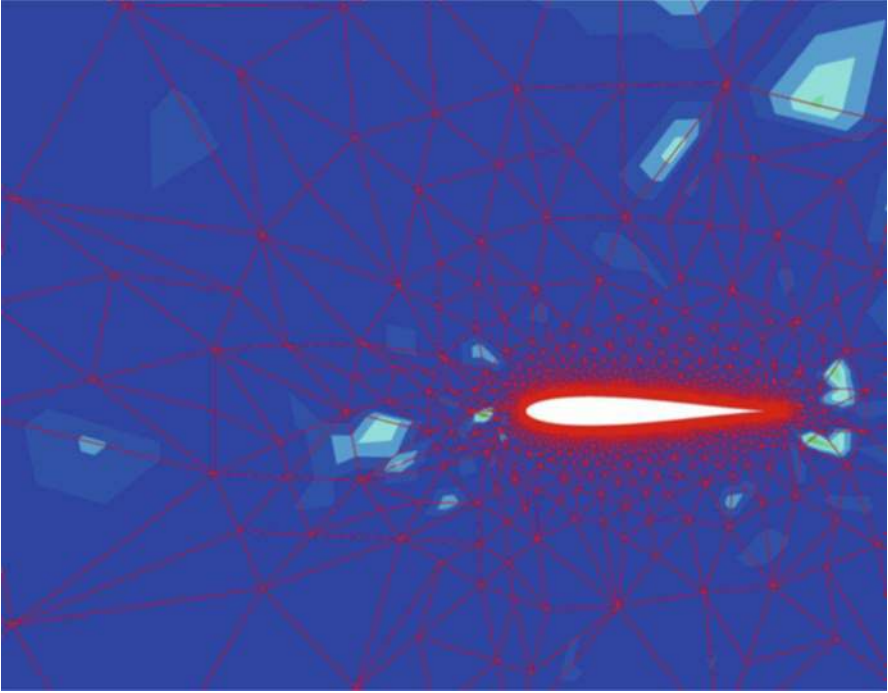


Fig. 1.2b Error indicator for turbulent flow around airfoil at 2° angle-of-attack (refined grid); *red* indicates significant lift increment, and *blue* indicates no increment; scale same as Fig. 1.2a

of finite-difference schemes, allowing weak solutions, such as shocks and contact discontinuities, to be captured with the correct jump conditions when the physical fluxes are augmented with dissipation terms. There is considerable literature on how to choose these terms. Additionally, to achieve better than first-order accuracy, the constant distribution within a cell is replaced by a locally piecewise linear distribution using neighboring data. This is termed “piecewise-linear reconstruction,” and the resulting scheme is second-order accurate for smooth flows. Limiters are usually employed to capture shocks and other discontinuities so that locally the schemes become first-order accurate. In the terminology of finite element methods FV schemes may be interpreted as using constant test functions and piecewise linear (reconstructed) basis functions. As for the characteristics above, FV methods satisfy item 1 if the grid is adequate. However, they use fairly large stencils (nearest and next-to-nearest) neighbors, and extensions to higher order are impractical because of increasingly larger stencils.

Finite element methods (FEM) have been popular in structural analysis for over 50 years, but have received much less attention in CFD until recently. These methods have in fact been in development in CFD for well over 20 years [2] and are the methods of choice for many production and research CFD codes. One advantage of these methods is that finite element theory provides a powerful framework for analyzing

accuracy and stability. A particularly accurate and robust family of schemes is the family of stabilized finite elements, e.g., streamwise upwind Petrov–Galerkin (SUPG) and discontinuous Galerkin (DG) methods. When using polynomial basis functions, these schemes are usually denoted as SUPG(k) and DG(k), where k refers to the degree of the polynomial basis. These schemes achieve $O(h^{k+1})$ accuracy for smooth flows. SUPG falls under the category of continuous finite element methods in that the basis functions are continuous across element boundaries. For example, if the unknowns are nodal values in a triangular mesh, the basis functions are piecewise linear over triangles and continuous across triangle edges. FEM make use of compact stencils. For example, with SUPG(1) one achieves second-order accuracy using linear basis functions, the stencil only involving nearest neighbors. In the case of DG, the lowest degree polynomial is of degree 0, resulting in a piecewise-constant basis. Thus, in the case of a triangular mesh, the basis function is constant within each cell and has jumps across edges. Having chosen the basis function, one then defines a numerical flux which introduces upwinding to achieve stability. SUPG introduces upwinding only in the streamwise direction. This provides adequate stability for smooth regions, but is inadequate at shocks and on coarse grids. With DG, the upwinding employed is similar to that used in FV methods. A DG(0) scheme that makes use of piecewise-constant basis functions and an approximate Riemann solver is identical to a first-order FV scheme and is a very stable method. SUPG must necessarily store the degrees of freedom associated with the linear basis function at the vertices with the higher order degrees of freedom stored at edges and cell centers. With DG, however, one usually stores the variables at the cell centers, although they can be stored at the vertices as well giving rise to a scheme termed nodal DG. The choice of the constant test function requires the definition of control volumes. In the case where variables are stored at cell centers, the control volumes are the triangles, whereas in the case of nodal DG, they are the dual to the triangles. The advantage of the nodal DG scheme is that it is possible to rely on continuous finite elements to discretize the viscous terms. With standard DG, the viscous terms have to be recast as a system of first-order partial differential equations (PDEs) with additional unknowns (that may be eliminated locally). See [2] for further discussion. Given a choice of basis functions and a numerical flux, the residual statement becomes the orthogonality between the PDEs and the test functions (taken to be the same as those defining the basis functions). Numerical quadrature then produces a system of nonlinear algebraic equations to be solved at each time step.

The extension to higher order with FEM is straightforward as far as the mechanics is concerned, but robustness remains a critical issue for nonlinear problems. One introduces additional degrees of freedom as needed to define the higher order basis functions. For example, SUPG(2) introduces additional degrees of freedom at edge midpoints so that a quadratic basis can be defined over a triangle and achieves third-order accuracy. DG(1) achieves second-order accuracy by storing the solution and its (scaled) derivative at grid points or at centers of cells. DG(2) achieves third-order accuracy by storing the solution, first and second derivatives.

FV methods are obviously conservative because they employ (numerical) flux balances over control volumes. In the case of DG(k), the residual statement with

the lowest order test function (constant) yields a similar result. With FEM, it is sometimes stated that these methods are not locally conservative. This is incorrect. Hughes et al. [3] have shown that FEM are indeed locally conservative. Venkatakrishnan et al. [4] have shown that it is actually possible to rewrite the local FEM residuals so they appear as flux balances over control volumes. This allows us to enhance the stability of SUPG(1) to deal with shocks and under-resolved features, by adding small amounts of DG(0) with a coefficient that is of $O(h)$ in smooth regions and $O(1)$ near shocks and unresolved features. The resulting method is competitive with FV methods. We typically see less false entropy generation using this scheme compared to a FV method on the same grid.

Forces and moments on a configuration or parts thereof are usually computed by numerical integration of physical fluxes, consisting of pressure and skin-friction contributions. These are more accurately computed using the numerical fluxes, because these are the fluxes in balance at convergence. Also, if the residuals are converged to machine-precision, forces may be computed by simple summation of momentum fluxes over the boundary of any subset of cells enclosing relevant portions of the configuration without loss of precision.

The computation of accurate moments can be as important to the design of a commercial transport as the computation of forces (e.g., lift and drag). Moment of momentum is an additional conservation law that holds for the continuous formulation. First, the use of numerical fluxes (as opposed to physical fluxes) generally results in more accurate estimates for moments. If the discrete scheme conserves moment of momentum, then it is possible to compute accurate moments by simple sums as with forces. It is easy to show that both Galerkin and SUPG obey this property as does DG(k), $k > 0$, but FV, including DG(0), does not. Since our scheme hybridizes SUPG and DG(0), we lose this property, although we believe certain variants of these schemes can recover it.

In order for a discrete scheme to yield a solution on any grid, the scheme must be robust. For a linear PDE, the discrete linear operator must be invertible. Unresolved features will manifest themselves as oscillations, which can be used in grid adaptation, and eventually these oscillations will disappear as the discrete solution approaches the continuous one. Alternate formulations exist for linear problems that eliminate non-physical oscillations altogether. These algorithms, typically nonlinear even for linear PDEs, often satisfy a discrete maximum principle or produce a linearized operator (Jacobian) that is always M -type. Unfortunately, we know of no extensions to these algorithms that provide similar properties for nonlinear systems of PDEs in multiple dimensions. Oscillations are much more of a problem for nonlinear PDEs. In the presence of oscillations, the Jacobian may become singular and/or the state may become unphysical (e.g., negative pressure). This is especially true on coarse grids and grids where there is a wide disparity in adjacent cell size and/or shape, as is often the case with adapted grids. Ideally we would like the Jacobian to always be M -type, but this is probably impossible for the RANS equations. Conceding this, we would hope that the Jacobian could always be made positive semi-definite or scalable to a positive semi-definite matrix, especially on an element by element basis. This appears to be feasible for the turbulent transport component

of the RANS equations, which is important as this component is often responsible for convergence problems. However, we have not managed to find a transformation allowing the full system to be positive semi-definite. At the very minimum we would like entropy stability, which is achievable. However, entropy stability does not guarantee total stability, as it only controls one component of the solution. During the transient stage of convergence we employ a pseudotime term which tends to make the Jacobian (with respect to entropy variables) more diagonally dominant. One would therefore like the steady-state Jacobian diagonal to not counter this time term. To this end, we have examined all sources of negative contributions to the diagonals. It is easy to see how upwinding strengthens the diagonal. One has to ensure, however, that the Galerkin terms do not counter this strengthening. A source term that has a Jacobian diagonal of the wrong sign must be rediscritized or else those terms with the correct sign enhanced. On occasion the mass equation must be scaled and added to the others. Through these means we have been able to ensure that the steady-state Jacobian diagonals are positive during all phases of solution. This seems to have considerably helped convergence.

Other discretization issues of importance involve the treatment of boundary conditions. We have developed a systematic means of strongly enforcing Dirichlet boundary conditions through the use of Lagrange multipliers. This formulation has the added benefit of producing adjoint fields which are continuous near boundaries in contrast to the typically discontinuous behavior observed for other methods of enforcing Dirichlet conditions. We have also found that weakly enforced boundary conditions can be problematic, particularly for slip walls in the presence of large aspect ratio cells. Discretizations of the compressible Euler and Navier–Stokes equations are known to have problems with stiffness and accuracy for flows approaching the incompressible limit. We note that even at modest freestream Mach numbers, portions of the flow field about a transport configuration can be nearly stagnant (e.g., in cove regions). This is also relevant for simulations of engine ground-test configurations, which are often conducted in quiescent flow. We are somewhat undecided as to how serious this issue might be in our cases. So far we have not found a low Mach number preconditioner that offers a significant advantage.

A good characterization of a discretization scheme in an adaptive context is the number of degrees of freedom required to achieve a specified level of accuracy. Another requirement is that the error remains bounded no matter how sub-optimally the grid is refined. The so-called torture test [4] compares various schemes on a linear problem, where grid refinement is done in the least optimal fashion. The results show that SUPG does much better than a FV scheme without limiting. Here the problem is linear and the solution is smooth, which should not require limiters.

Finally, we note an interesting consequence of enhancing convergence. Often failure to converge to a steady-state solution is attributed to inherent flow unsteadiness. In Fig. 1.3a we show lift coefficient versus angle-of-attack for a multi-element airfoil and in Fig. 1.3b lift coefficient versus drag coefficient. Flow features for 90° angle-of-attack case are shown in Figs. 1.4 and 1.5. An intermediate-adapted grid (suitable for viewing) is shown in Fig. 1.6. The nonlinear residuals at each angle-of-attack are converged to machine error. Our discretization is demonstrably second

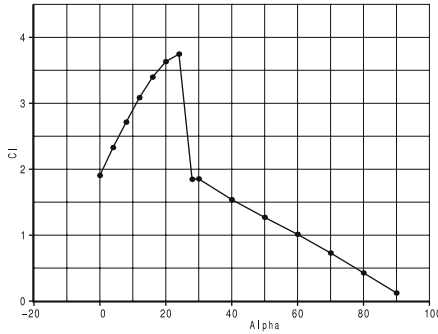


Fig. 1.3a Lift curve for multi-element airfoil

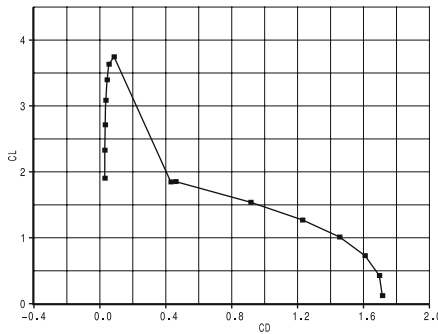


Fig. 1.3b Drag polar for multi-element airfoil

order and consistent with the Navier–Stokes equations supplemented by the Spalart–Allmaras turbulence model. From the adjoint prediction and other indications all solutions are grid converged. For the 90° angle-of-attack case 19 grids were used with the last grid totaling 255,672 grid points. Evidently the turbulence model allows a steady-state solution far beyond conditions where the flow would normally become unsteady. Thus, a certain amount of skepticism is in order before concluding that non-convergence necessarily implies unsteadiness.

1.2.3 Higher Order Elements

Higher order elements have the ability to dramatically reduce the degrees of freedom required to obtain a solution of given accuracy [4]. The difference in solution resources required between a second-order accurate scheme with 100 million nodes and a third-order accurate scheme with perhaps 3 million nodes would be enormous. In theory, viscous flow fields are smooth, so that the use of higher order elements should be very efficient. However, a grid around a transport aircraft would have to

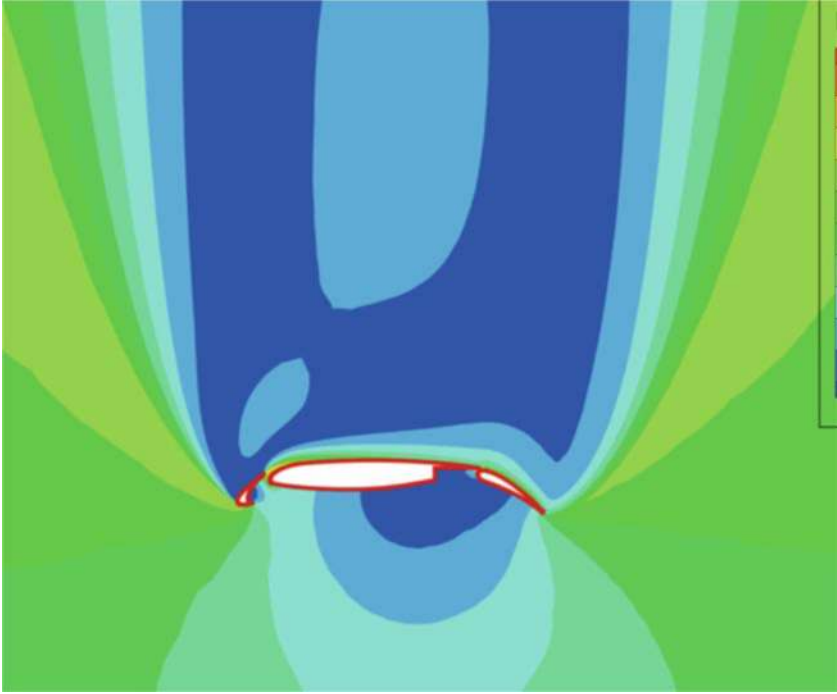


Fig. 1.4 Mach distribution about multi-element airfoil at 90° angle-of-attack

be prohibitively fine to treat certain flow features (e.g., shocks) as smooth. Even relatively smooth flow features can appear as discontinuities on coarse grids. To take advantage of higher order methods some mechanism is required to prevent non-physical oscillations, especially when the underlying fluxes are not defined for certain states. In order to simplify the code structure and not waste degrees of freedom, we favor an approach that gradually switches from p -refinement to h -refinement. If necessary, in certain parts of the flow field one can limit back to robust constant elements and rely on h -refinement. Such a procedure should be automatic using limiters. We note that many current limiters are not particularly robust and often as a result of the limiting algorithm higher order degrees of freedom become unused. In the following paragraph we note a reasonable way of trading degrees of freedom.

In Fig. 1.7 we show a uniformly refined triangle with numbered nodes and edge midpoints at which degrees of freedom u_i , $i = 1, 6$ are defined. For our purposes we assume the test functions for the finite element discretization are piecewise linear on the refined grid. Ordinarily a third-order method on the coarse triangle would employ upwinded fluxes using a quadratic element distribution interpolating the 6 degrees of freedom. To help achieve stability the inviscid fluxes can be modified by gradually limiting the interpolation function. In particular, we define the element distribution at an arbitrary point in any of the sub-triangles by

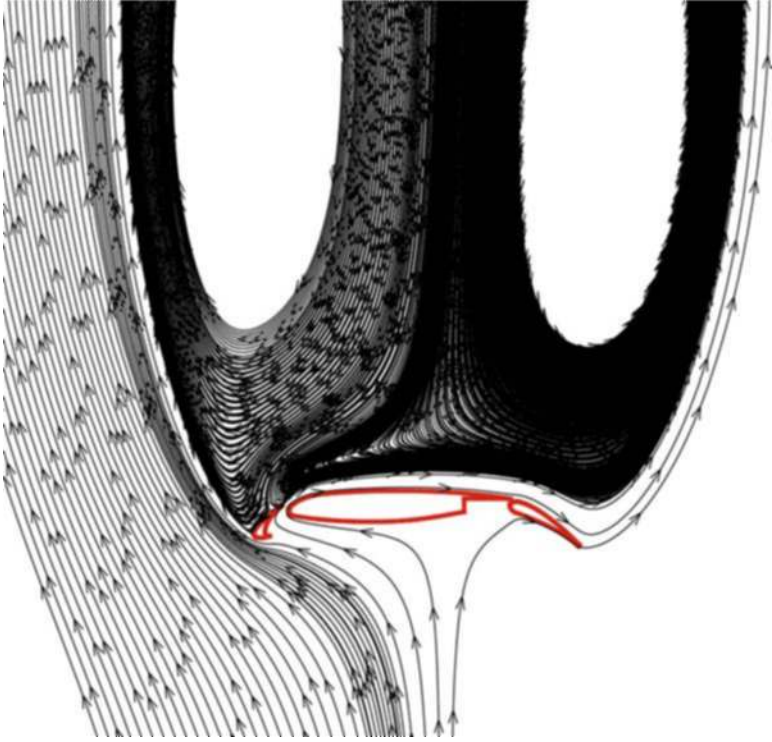


Fig. 1.5 Streamlines about multi-element airfoil at 90° angle-of-attack

$$u = (1 - \lambda)[(1 - \mu)u_Q + \mu u_L] + \lambda u_0 \quad (1.1)$$

Here u_Q is a quadratic distribution interpolating all six nodal values and u_L is a linear distribution interpolating the three nodal values of the sub-triangle containing the point. u_0 is the (constant) value of the degree of freedom associated with the dual sub-cell containing the point. Because of the discontinuous nature of the distribution, edge terms involving numerical fluxes must be added. One can also limit the fluxes instead of the unknowns u using the fact that continuous FEM are locally conservative with respect to dual cell boundaries as mentioned above. In Eq. (1.1), λ and μ are $O(h)$ in smooth regions and become $O(1)$ near discontinuities in solution and slope, respectively.

Once robustness issues regarding limiting are resolved, higher order (higher than linear) FEM will become practical. We believe that by just upgrading from linear to quadratic elements, the number of degrees of freedom required to achieve a given level of accuracy for our problem sizes will be reduced by over an order of magnitude. Because continuous quadratic elements in three dimensions require degrees of freedom at edge midpoints as well as nodes and there are approximately seven to eight times as many edge midpoints as nodes, the decision to use quadratic elements is not easily made despite the theoretical benefits. Here the use of p -multigrid can

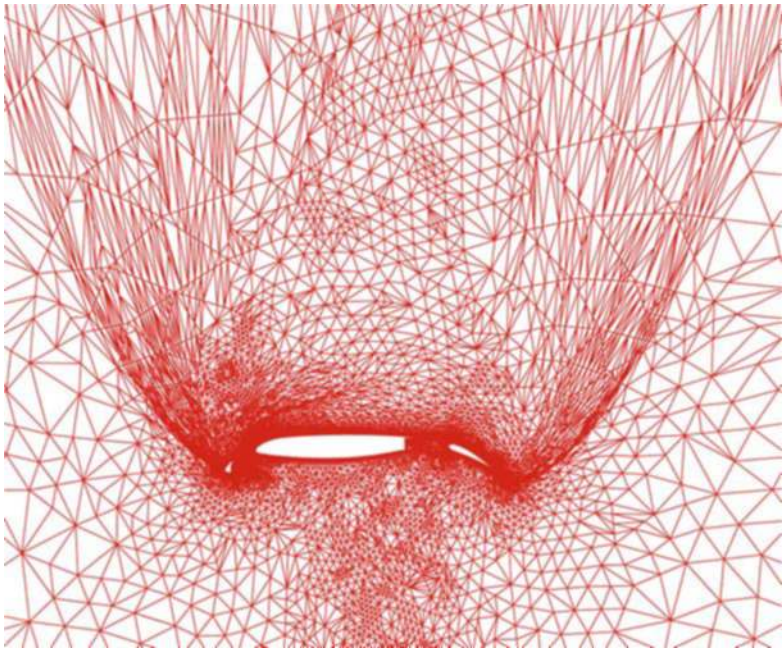


Fig. 1.6 Intermediate-adapted grid about multi-element airfoil at 90° angle-of-attack

help. Assuming a good reconstruction algorithm one can alternatively define edge midpoint values as a function of values at adjacent nodes and proceed directly with quadratic elements. Good limiters would be critical here. The size of the system of equations to be solved then reverts to that of linear elements. The net effect would

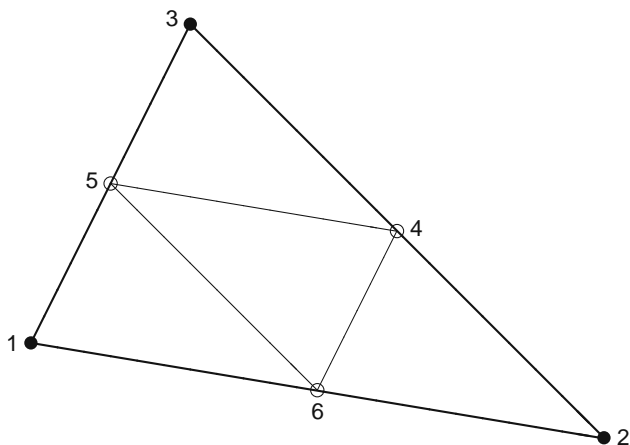


Fig. 1.7 Triangle degrees of freedom

be to enlarge the stencil, so that the very real accuracy benefit of a compact stencil would be lost. Nevertheless, the efficiency advantages of such an approach probably outweigh the drawbacks.

1.2.4 Domain Decomposition and Linear Solver

One of the most powerful and reliable methods for solving the nonlinear algebraic system of equations resulting from discretization is Newton's method. Newton's method proceeds by using a Taylor series expansion to define an update of the unknowns that will decrease the nonlinear equation residuals. A requirement for using Newton's method then is that the algebraic equations be differentiable, which is a somewhat difficult requirement when limiters are involved. The update is defined by the solution of a large sparse linear system. This update may be scaled at each step by a line search phase that computes an optimal step length. Such a step length can be chosen to achieve the maximum reduction in the nonlinear residual norm subject to various constraints on the solution state. Even though we are interested in a steady solution, we find it is advantageous to include the time derivative. The time step, which will be driven to infinity at steady state, can be used as a Levenberg–Marquart-type parameter to augment Newton's method enabling it to circumvent local extrema during the transient phase. In order to implement Newton's method, it is necessary to compute the Jacobian of the nonlinear residuals with respect to the variables being solved.

The Jacobian can be computed in many ways. The easiest way is to use finite differences. This requires repeated calls to the procedure evaluating residuals. The number of such calls can be minimized by coloring the sparse matrix graph or by proceeding on a cell-by-cell basis and accumulating contributions. The drawback of this approach is that it requires a valid step size and there may be no acceptable choice balancing round-off with truncation error. This can be a particular problem on adapted grids with high aspect ratio cells. Another option is to use complex variables in the function call. This approach does not require the subtraction of nearly equal quantities, but still requires a choice of step size and division by that parameter. Analytical Jacobian evaluation gets around these problems entirely, but can be tedious to code in large CFD codes. The operator overloading feature available in Fortran 90, C++, and other programming languages offers an easy and reasonably efficient way for computing the Jacobians, without requiring much modification of the source code. The Jacobians are obtained as a by-product as the residuals are being computed.

With Newton's method, a sparse linear system must be solved at each time step. For SUPG(1), recall that a stencil only involves nearest neighbors. In two dimensions, it is possible to invert this system of equations exactly using sparse direct methods which employ ordering algorithms such as quotient minimum degree or nested dissection to reduce the amount of fill-in that occurs during the factorization. In three-dimensions, however, direct solution is impractical on all but very small

problems. A typical problem size in three dimensions is about 10–20 million grid points with 6 or 7 unknowns per grid point. The fill-in associated with problems of this size leads to impossibly large memory requirements and unacceptable computation times. On the hardware side, distributed memory parallel computers offer the most cost-effective way to access large amounts of memory. Sparse direct methods are still unaffordable on such machines because the memory requirements are still too severe and also because these methods possess a limited amount of parallelism. Considering such issues, our method of choice is a preconditioned iterative method. The preconditioner is an approximate solve within each domain, together with a Schur complement or a coarse grid solver, all tied together with a GMRES driver.

For preconditioning within each domain, we use an incomplete factorization with drop tolerance (ILUT). The amount of fill-in is affected by the ordering of unknowns as well as the drop strategy. For Navier–Stokes equations with the one-equation Spalart–Allmaras turbulence model, dropping is done by treating the Jacobian as a block-matrix, with blocks of size 6×6 . Ordering of unknowns for complete LU factorization has been studied extensively, but is a relatively unexplored area of research when drop tolerance is used. Currently, we use standard reordering techniques, such as the quotient minimum degree (QMD) algorithm, that are based on the (symmetric) graph of the stencil, to determine the ordering within each sub-domain. As in the case of partitioning, reordering of unknowns based on QMD is done once during the preprocessing phase.

The grid and associated data structures should be distributed among the processors in the parallel computer so that the computational load is balanced. Note that the computational load varies during the two main phases of computation, i.e., the residual and Jacobian evaluation phase and the factorization and linear solution phase. During the first phase, computational load can be roughly taken to be proportional to the number of grid points on the processor, whereas during the second phase it cannot be estimated ahead of time, because it depends on the local stencil and the drop tolerance parameter. One way to mitigate this problem is to subdivide the problem into more sub-domains, assign multiple domains to each compute node, and periodically rearrange sub-domain assignment partially based on the computed fill-in.

Distributed memory computers are now based on compute nodes that have dual or even quad-core processors. On such machines, one can subdivide the problem on each compute node into two or four sub-domains and employ the standard preconditioned GMRES algorithm to couple these sub-domains as well. Alternatively, the domain can be treated as a single domain on the compute node and parallelism can be achieved at a loop level using OPENMP directives. The advantage of this approach is that the solver is implicit within the compute node and thus will require fewer GMRES outer iterations, while achieving almost perfect parallelism in the residual and Jacobian evaluation and forward/backsolve stages. The disadvantages with this approach are the higher memory requirements due to the larger domain size and the difficulty exploiting parallelism during the factorization phase.

The preconditioned GMRES algorithm described above is called an additive Schwarz method in domain decomposition literature. It is a block-Jacobi preconditioner where each block is solved using ILUT as the domain solver. The number of GMRES iterations grows as the number of sub-domains increases and also when the number of unknowns per sub-domain increases. This growth can be controlled by using a Schur complement method. The Schur complement results from block elimination of the interior of the sub-domains. This creates a smaller linear system involving only the unknowns at the inter-partition boundaries that is denser than the original system. It can be shown that employing block diagonal preconditioning for this system is the same as the original preconditioned GMRES method, but if one solves this linear subsystem, the number of outer GMRES iterations decreases dramatically. In the special case when the domain solver is exact, the outer problem converges in one GMRES iteration.

Another method for accelerating the convergence of linear problems is the multi-grid method. This method is particularly appropriate in the case of grid adaption, where suitable coarse grids are often available as a by-product of the adaption process (e.g., HIREF). In domain decomposition parlance, when a powerful sub-domain solver is used, this is called an additive Schwarz method with multiplicative coarse-grid correction. The discussion here will pertain to the latter interpretation. A coarse-grid linear system is solved that is driven by a restriction of fine-grid residuals, and the corrections obtained from the coarse grid are prolonged to the fine grid. The critical elements of this method are the transfer (prolongation and restriction) operators, and how one derives this coarse-grid system. The coarse-grid operator can be derived either from the fine grid and the transfer operators giving rise to a Galerkin coarse-grid operator or from the linearization of the discretization of the nonlinear problem on the coarse grid. The (linear) interpolation operators can be difficult to construct for complex curved geometries, especially on highly stretched grids. While we have had a moderate degree of success with the multigrid approach for special cases, e.g., the fine grid being a uniformly refined coarse grid, we prefer the Schur complement approach because of its algebraic approach and ease of implementation.

To summarize this section a number of issues pertaining to creating a production-level Navier–Stokes CFD tool have been addressed. We have also discussed our current approaches to addressing these, but many open problems remain. In the area of discretization, the main difficulty with regard to higher order methods is their lack of robustness. A fertile area of research here is how a baseline high-order scheme can be made to smoothly switch to lower order accuracy (possibly h -refined) when the solution is not smooth and also on under-resolved grids. In the area of solver technology, the quest for a scalable preconditioner that does not require unaffordable memory continues. Other topics of research are parallel dynamic load balancing in Newton’s method and solver methodology for higher order methods. In the area of grid generation and adaptation, robust initial grid generation and reliable refinement and coarsening of grids in three dimensions require further research.

1.3 Conclusions

At Boeing Commercial Airplanes, Seattle, CFD has evolved into a highly valued tool for the design, analysis, and support of cost-effective and high-performing commercial transports. The application of CFD today has revolutionized the process of aerodynamic design, and CFD has joined the wind tunnel and flight test as a critical tool of the trade. Experience to date has shown that CFD has had its greatest impact on the aerodynamic design of the cruise configuration of a transport aircraft. To a large extent this is due to the fact that a great deal of research and development have been devoted to algorithms and processes for allowing project users to reliably apply CFD to attached flow phenomena without the aid of CFD experts. There are numerous indications that CFD can be applied to flows covering more and more of the full flight envelope. Routine use here by project engineers will require further algorithm research, application studies, and process development. With regard to improved algorithms it appears that there are no magic bullets. A surprising number of issues must be addressed and there are still a number of open questions. Some of these have been considered in this chapter.

Acknowledgments The authors thank our colleague, Dmitrii Kamenetskii of the Keldysh Institute in Moscow, for valuable results regarding aspects of the SUPG method. We also thank our colleague, Andrey Wolkov of the Boeing Technical Research Center, Moscow, for similar results regarding the DG method. Finally, we thank our colleagues Victor Zhukov and Olga Feodoritova of the Keldysh Institute for valuable results regarding domain decomposition and multigrid methods.

References

1. Johnson, F.T., Tinoco, E.N. and Yu, N.J., Thirty Years of Development and Application of CFD at Boeing Commercial Airplanes, Seattle, AIAA-2003-3439, June, 2003; also in *Computers & Fluids*, 34 (2005), pp. 1115–1151.
2. Hughes, T. J. R. and Brooks, A., Multi-dimensional Upwind Scheme With No Crosswind Diffusion, *Finite Element Methods for Convection Dominated Flows*, New York, 1979, ASME.
3. Hughes, T. J. R., Engel, G., Mazzei, L., and Larson, M. G., The Continuous Galerkin Method is Locally Conservative, *J. Comp. Phys.*, 163 (2000), pp. 467–488.
4. Venkatakrishnan, V., Allmaras, S.R., Kamenetskii, D., and Johnson, F.T., Higher Order Schemes for the Compressible Navier-Stokes Equations, AIAA Paper 2003-3987, 16th AIAA Computational Fluid Dynamics Conference, June 23–26, 2003, Orlando, FL.

“This page left intentionally blank.”

Chapter 2

Flight Path Optimization at Constant Altitude

Mark D. Ardema and Bryan C. Asuncion

Abstract In this chapter we consider flight optimization at constant altitude for a variety of missions and propulsion systems and then focus on maximizing the range of a turbofan-powered aircraft. Most analyses of optimal transport aircraft flight begin with the assumption that the flight profile consists of three segments – climb, cruise, and descent. Indeed, this is the flight profile of all long-haul commercial flights today. The dominant stage of such flights, in terms of flight time, is the cruise segment. The air transportation industry is extremely competitive and even small changes in aircraft performance have significant impacts on the operation costs of airlines. Thus, there has been, and continues to be, great interest in optimizing the cruising flight of transport aircraft.

2.1 Introduction

In this chapter we consider flight optimization at constant altitude for a variety of missions and propulsion systems and then focus on maximizing the range of a turbofan-powered aircraft. Most analyses of optimal transport aircraft flight begin with the assumption that the flight profile consists of three segments – climb, cruise, and descent. Indeed, this is the flight profile of all long-haul commercial flights today. The dominant stage of such flights, in terms of flight time, is the cruise segment. The air transportation industry is extremely competitive and even small changes in aircraft performance have significant impacts on the operation costs of

Mark D. Ardema

Santa Clara University, 500 El Camino Real Santa Clara, CA 95053, USA,
e-mail: mardema@scu.edu

Bryan C. Asuncion

Santa Clara University, 500 El Camino Real Santa Clara, CA 95053, USA,
e-mail: basuncion@scu.edu

airlines. Thus, there has been, and continues to be, great interest in optimizing the cruising flight of transport aircraft.

The classical performance relation for cruising flight is the “Brequet range equation.” This is based on steady flight (constant speed and altitude), leaving only the range and mass as dynamic variables. Integrating the state equations associated with these two variables, assuming a constant lift-to-drag ratio, gives the Brequet equation:

$$R = B \ln \frac{m_0}{m_f} \quad (2.1)$$

where R is the range, m_0 is the initial mass, m_f is the final mass, and B is the Brequet factor, given by

$$B = \frac{\lambda V}{gC} \quad (2.2)$$

where λ is the lift-to-drag ratio, V the cruise speed, g the gravitational acceleration, and C the thrust-specific fuel consumption.

Thus, to optimize the flight path (in the sense of either maximizing range for a given mass ratio or maximizing the mass ratio for a given range), a search is conducted to find the point in the flight envelope (the portion of the (h, V) plane that does not violate any constraints) that maximizes B . Because the Brequet factor changes as fuel is burned off during the flight, the optimal (h, V) values change as well. Typically the optimum altitude increases during the flight, resulting in a steady “cruise climb.” Air traffic control requires that aircrafts hold specific altitudes; thus, the operational flight paths of long-haul transport aircraft are “step climbs” during which the altitude is increased at discrete times. Note that a cruise or step climb violates the assumption of flight at constant h but the rate of change of altitude is quite small.

Several authors have used more detailed math models to study aircraft cruise, models in which V and h are allowed to vary [1, 2]. These authors have investigated whether or not cyclic cruise is better than steady cruise. It was found that flight paths with large periodic changes in h , V , and throttle could be more fuel efficient for fixed-range missions. However, the improvement (reduction) in fuel consumption was very small, 1% at best. Furthermore, such flight paths would not be compatible with air traffic procedures as the altitude oscillations sometimes exceed 10,000 ft.

On the other hand, for endurance missions (maximum time) the non-steady flight paths gave great improvement relative to steady ones.

The work just discussed focuses on the interplay between h , V , and throttle setting. Because of the operational restrictions on cruise altitude, in the present chapter we look at aircraft cruise from a different point of view; in particular, we study the interplay between aircraft mass and speed at constant altitude. This problem was first considered by Miele [3] for rocket-powered aircraft. In [4] we extended these results to jet aircraft and made some preliminary calculations. This chapter extends these results further to other types of missions and propulsion systems. Because this is a singular optimal control problem, we begin with a review of that subject.

2.2 Singular Optimal Control

Consider a system whose math model is a set of state equations:

$$\dot{x} = f(x) + g(x)u \quad (2.3)$$

where $x \in \mathfrak{R}^n$ is the state vector and $u \in \mathfrak{R}$ is the scalar control variable bounded by $u_m \leq u \leq u_M$. The state may be free or fixed at $t = 0$ and $t = t_f$. It is desired to minimize

$$J = \int_0^{t_f} [f_0(x) + g_0(x)u] dt \quad (2.4)$$

This is a problem of singular optimal control [5, 6].

In [4] we found the singular arc in two different ways – by applying the maximum principle and using Green's theorem – and found that the Green's theorem approach was much easier and hence we use this method here. Application of Green's theorem to this type of problem was introduced independently by Miele [7] and Mancill [8]. These two works are quite different; Mancill approaches the problem as an identically non-regular problem of the calculus of variations, whereas Miele approaches it as a problem of optimal control. The method was developed into a powerful analytic tool by Miele [9]. The Green's theorem method only applies to the case of two state variables, say x and y , the case considered later in the chapter.

Now consider the optimization problem with x and y fixed at $t = 0$ and $t = t_f$:

$$\begin{aligned} \dot{x} &= f_x(x, y) + g_y(x, y)u \\ \dot{y} &= f_y(x, y) + g_x(x, y)u \\ J &= \int_0^{t_f} f_0(x, y) dt \end{aligned} \quad (2.5)$$

Eliminating the control between the state equations and substituting into the cost functional results in

$$J = \int_{(x_0, y_0)}^{(x_f, y_f)} (Adx + Bdy) \quad (2.6)$$

where

$$A = \frac{f_0 g_y}{f_x g_y - f_y g_x}, \quad B = \frac{f_0 g_x}{f_y g_x - f_x g_y} \quad (2.7)$$

Equation (2.6) is a line integral in the plane. Green's theorem relates line integrals around closed curves to area integrals. To use the theorem, consider the closed curve consisting of the curve to be optimized, C_1 , plus a fixed, but arbitrary, curve, C_2 , returning to the starting point; then Green's theorem is

$$\int_{C_1} (Adx + Bdy) + \int_{C_2} (Adx + Bdy) = \iint_A \left(\frac{\partial A}{\partial y} - \frac{\partial B}{\partial x} \right) dA \quad (2.8)$$

Since integral C_2 is fixed, minimizing integral C_1 is the same as minimizing the integral A . The critical curve associated with the latter integral is

$$\frac{\partial A}{\partial y} - \frac{\partial B}{\partial x} = 0 \quad (2.9)$$

and is optimizing. This equation is the singular arc.

2.3 The Cruise Problem

Consider an aircraft flying at constant altitude at a constant heading. The equations of motion are

$$\begin{cases} \dot{V} = \frac{T-D}{m} \dot{m} = -\beta \\ L = mg \end{cases} \quad (2.10)$$

where V is the speed; m is the mass; T , D , and L are the thrust, drag, and lift forces, respectively, and β is the fuel flow rate. It is assumed that

$$\begin{aligned} T(V) &= \Pi T_M(V) \\ \beta(V) &= C(V) T(V) \\ D &= AV^2 + \frac{BL^2}{V^2} \end{aligned} \quad (2.11)$$

where D represents a parabolic drag polar [13], $A = C_{D0} \frac{\rho s}{2}$, and $B = \frac{2K}{\rho s}$. The zero-lift drag coefficient, C_{D0} , the induced drag coefficient, K , the air density, ρ , and the reference area, s , are all taken as positive constants, a good assumption for flight at constant altitude of a subsonic aircraft. Π is the throttle setting where $\Pi_m \leq \Pi \leq 1$. In this chapter we consider three types of propulsion systems: rockets, jets, and internal combustion with propeller. Rockets are modeled by taking $C = \text{constant}$. Props are modeled by $C = (C_p/k)V$ where C_p and k are engine and propeller efficiencies, respectively. For low-speed flight ($M < 0.5$) C_p and k are nearly constant. For jet engines, C depends on various temperatures and pressures within the engine. In this chapter, we use EngineSim 19.11 to model the relationship between C and h , V , and Π .

It is desired to minimize

$$J = \int_0^{t_f} (E - V) dt \quad (2.12)$$

with t_f free. This is a problem with two states, V and m , and one control, Π . This cost functional is a weighted sum of minimum time and maximum range, with E being the weighting function. It has been found that this cost functional is closely related to direct operating cost. Special cases are maximum range ($E = 0$) and maximum endurance ($E = -8$). This is very similar to the problem first considered by Miele [3] and later appearing in Leitmann [5]. The differences are that instead of a rocket engine with constant exhaust exit velocity, we include air breathing engines with speed-dependent maximum thrust, T_M , and specific fuel consumption, C , and consider a more general cost functional.

To use Green's theorem, we begin by eliminating the control from Eq. (2.10) and substituting the result into Eq. (2.12); the result is

$$J = \int_0^{t_f} \left[(V-E) \frac{m}{D} DV + (V-E) \frac{1}{CD} dm \right] \quad (2.13)$$

Applying Green's theorem, this is equivalent to

$$J = \iint_A \left[\frac{\partial}{\partial m} \left((V-E) \frac{m}{D} \right) - \frac{\partial}{\partial V} \left((V-E) \frac{1}{CD} \right) \right] dV dm \quad (2.14)$$

Thus the singular arc is

$$\frac{\partial}{\partial m} \left[\left((V-E) \frac{m}{D} \right) \right] - \frac{\partial}{\partial V} \left[\left((V-E) \frac{1}{CD} \right) \right] = 0 \quad (2.15)$$

There are three special cases depending on propulsion system and mission:

Jet (Range):

$$m = \frac{V^2}{g} \sqrt{\frac{A(1+CV + \frac{V}{C}C_v)}{B(3+CV - \frac{V}{C}C_v)}}$$

Jet (Endurance):

$$m = \frac{V^2}{g} \sqrt{\frac{A(2+CV + \frac{V}{C}C_v)}{B(2+CV - \frac{V}{C}C_v)}}$$

Rocket (Range):

$$m = \frac{V^2}{g} \sqrt{\frac{A(1+CV)}{B(3+CV)}} \quad (2.16)$$

Rocket (Endurance):

$$m = \frac{V^2}{g} \sqrt{\frac{A}{B}}$$

Prop (Range):

$$m = \frac{V^2}{g} \sqrt{\frac{A(1+CV + \frac{C_p V}{kC})}{B(3+CV - \frac{C_p V}{kC})}}$$

Prop (Endurance):

$$m = \frac{V^2}{g} \sqrt{\frac{A(2+CV + \frac{C_p V}{kC})}{B(2+CV - \frac{C_p V}{kC})}}$$

There is a second-order necessary condition which optimal singular control must satisfy. The traditional approach uses the maximum principle to identify the switching function and then the switching function is differentiated until the control appears. This was done in [4] and great algebraic complexity was encountered.

However, [11] gives a Green's theorem approach to this condition and we use this here. The condition for the endurance mission is

$$\frac{\partial^2}{\partial V^2} \left(\frac{1}{CD} \right) \geq \frac{\partial^2}{\partial m \partial V} \left(\frac{m}{D} \right) \quad (2.17)$$

Carrying out the differentiation for a rocket engine gives

$$\begin{aligned} & 6A^2C^2V^2 - 18AC^2Em^2V^{-2} + 2C^2E^2m^4V^{-6} - 6Em^2V^{-2}C^2A \\ & + 4C_vCA^2V^3 - 4C_vCA^2V^3 - 4C_vCE^2m^4V^{-5} - C_{vv}CA^2V^4 \\ & - C_{vv}C2AEm^2 - C_{vv}CE^2m^4V^{-4} + 2C_v^2A^2V^4 + 4C_v^2AEm^2 \\ & + 2C_v^2E^2m^4V^{-4} \geq 12Em^2V^{-1}AC^3 - 4E^2m^4V^{-5}C^3 \\ & + 2E^2m^4V^{-5}C^3 \end{aligned} \quad (2.18)$$

where subscripts denote partial derivatives.

2.4 Fanjet Specific Fuel Consumption

Aircraft data show that specific fuel consumption is far from constant and can vary significantly over an aircraft's flight envelope. The equations for C are very complicated and highly specific to the engine. C is mainly a function of speed and air temperature. Altitude is a factor in C calculation, because air temperature varies with altitude.

For turbojets and turbofans, C depends on several temperatures in the engine, pressure ratios, bypass ratios, and the fuel to air mass ratio in the combustor. There are computer simulation capabilities that can provide the information we want, and here we use the NASA Glenn EngineSim [11].

The NASA Glenn EngineSim, setup for a CF6 turbofan sized for use on a 747-400, was sampled at various velocities to generate a C vs. V relationship for full throttle of the engine. Another condition was generated, where the engine thrust matches the drag of the aircraft with respect to speed, which is more accurate to what an engine mounted on an aircraft would provide. This was done by developing a table of speed and parabolic drag values, entering the speed into EngineSim, and then adjusting the throttle to match the drag.

Figure 2.1 shows these results for the general relation of C as a function of speed at constant altitude. The dashed line represents the full throttle condition, and as expected, the slower the aircraft is flying, the more efficient the turbofan becomes. The solid line represents the change to specific fuel consumption if in addition to the reduced speed, the throttle is reduced to match the drag. Notice that minimum throttle occurs at about 250 m/s.

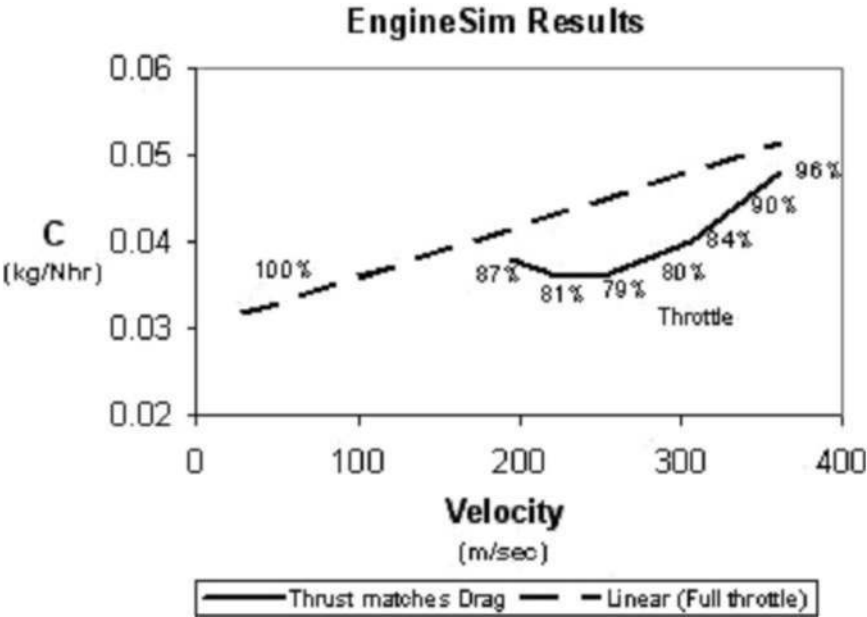


Fig. 2.1 C variation with respect to speed for a high-bypass turbofan

It is possible to simplify Eq. (2.16), max range with the jet, by considering values of the VC term. Subsonic aircraft speed does not get much above 300 m/s; however, C can vary significantly depending on the type of engine and application. As will be discussed later in this chapter, the CF6-80C2 engine example would have a C around 0.04 kg/N hr, which would be around 10^{-5} s/m. Therefore, its VC , a dimensionless number, would be on the order of 3×10^{-3} , which is small compared to 1 and 3. To give an order of scale to the VC value, the Concorde flying at Mach 2 only has a VC value of 0.018. The VC for an F-16 at full afterburner is around 0.03. For the VC term to be really large, the aircraft would need to have high speed and be using rocket engines. The Space Shuttle main engines have a VC of 1.67.¹ These numbers are summarized in Table 2.1. Thus the VC term may be neglected for our application.

Table 2.1 Magnitude of VC term with various aircrafts

Subsonic Jet	0.003
Concorde	0.02
F-16	0.03
Space Shuttle	1.7

¹ The Space Shuttle VC value is approximated by orbital speed divided by specific impulse and gravity. Specific impulse for the Space Shuttle Main Engines (SSME) is 428 s.

2.5 An Example

For example calculations, we choose a maximum range mission with a fanjet-powered aircraft for which the first of Eq. (2.16) applies. The aircraft selected is a Boeing 747-400 with General Electric CF6-80C2 high bypass, multistage, turbofan engines. Depending on the options installed on the engine, the CF6-80C2 is rated to produce max thrust of 282,500 N per engine (63,500 lbf) [12] at sea level.

The design range of a 747-400 is listed as 13,444 km (8,354 mi.), with a maximum takeoff mass of 362,875 kg (800,000 lb) [13, 14]. At takeoff, a full fuel load would be 120,205 kg (265,000 lb). This leaves 61,415 kg (135,400 lb) for passengers, crew, and cargo. It is easy to see that even small percentage changes to the amount of fuel required for each flight can provide significant potential for increased payload and thus for cost savings and increased revenue.

For the 747-400 aircraft and CF6-80C2 engine model, we input speed and throttle values to get C , thereby getting an order of magnitude of C_V to determine if it is significant or not. By using two mass-speed points, entering them into EngineSim, adjusting throttle, and looking at the change of C , we find C_V is on the order of 10^{-8} . With V/C being on the order of 10^8 , the term $(V/C)C_V$ is significant. Therefore, after eliminating VC , but leaving $(V/C)C_V$, the first of Eq. (2.16) becomes

$$m = \frac{V^2}{g} \sqrt{\frac{A \left(1 + \frac{V}{C} C_V\right)}{B \left(3 - \frac{V}{C} C_V\right)}} \quad (2.19)$$

Using EngineSim $(V/C)C_V$ is found to vary from 1.21 to 0.93 for the aircraft/engine combination considered here.

Equation (2.19) is plotted in Fig. 2.2 for various altitudes. The main feature is that the speed decreases as fuel is burned off.

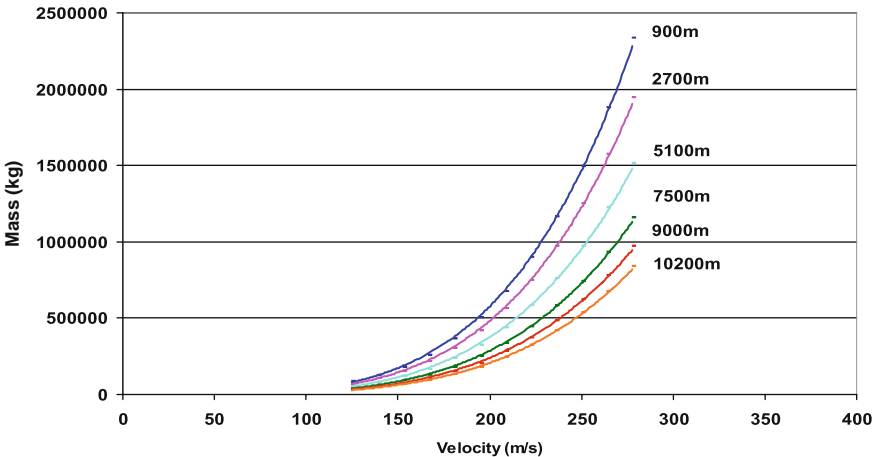


Fig. 2.2 Singular arcs and optimal path for 747-400 at different altitudes

Table 2.2 Comparative results for optimal path and standard cruise

Flight profile	Cruise speed (m/s)	Altitude (m)	Range (km)	Flight time (hrs)
Standard steady cruise	250	10,000	13,570	15.1
Optimal steady cruise	210	9,000	14,889	19.7
Optimal singular arc	219–206	9,800	14,904	19.5

We now conjecture how optimal paths look in the (m,V) plane. The aircraft climbs to the start of the cruise on the singular arc, follows the arc until cruise fuel is expended, and then descends. For our example, the entire cruise portion of the flight may follow the singular arc. A 747-400 with maximum fuel load flying the singular arc path for the entire cruise portion of the flight starts out burning about 10.0 kg/km and finishes burning 6.60 kg/km. In contrast, the standard steady speed path at the cruise speed of 250 m/s would start out burning 10.4 kg/km and finish burning 7.62 kg/km.

A computer simulation of the flight’s cruise portion shows that following the singular arc would result in an additional range. The steady cruise is performed at 250 m/s at 10,000 m altitude. The optimal singular arc is performed at 219–206 m/s at 9800 m altitude. With the results from Table 2.2, it can be observed that the optimal singular arc could have a maximum range of 14,904 km and a travel time of 19.50 hr.

Because many airlines’ costs are directly related to flight time, transport aircraft actually fly faster than the speed for maximum range with a given fuel load (equivalently minimum fuel for a given range). Thus, to make the comparison with singular flight fair, we maximized range in steady cruise. The results are included in Table 2.2. Optimal steady cruise is at 210 m/s speed and 9,000 m altitude. Optimal singular arc flight covers 115 more miles than optimal cruising flight, probably within the noise band of the numerical calculations.

Time history plots (Figs. 2.3, 2.4, 2.5 and 2.6) compare the optimal singular arc path with that for standard cruise in terms of speed, range, mass, and thrust,

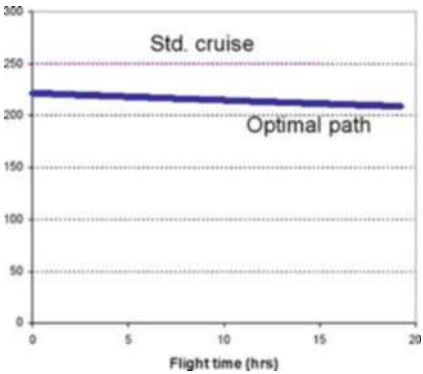


Fig. 2.3 Comparative time history plots of optimal path and standard cruise in terms of speed

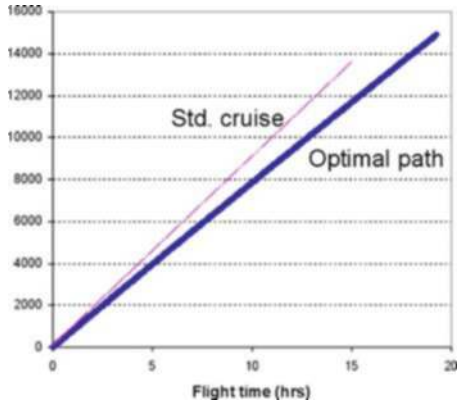


Fig. 2.4 Comparative time history plots of optimal path and standard cruise in terms of range

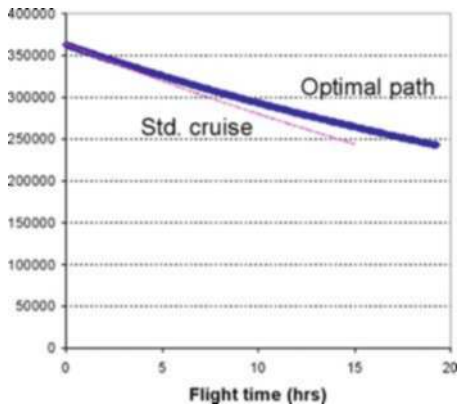


Fig. 2.5 Comparative time history plots of optimal path and standard cruise in terms of mass

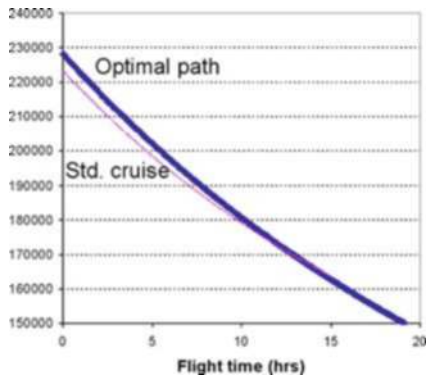


Fig. 2.6 Comparative time history plots of optimal path and standard cruise in terms of thrust

respectively. One interesting finding is that the optimal path results in higher thrust during the entire flight time. This would tend to disagree with the fuel savings finding, except that the aircraft is flying at a more efficient speed for the large turbofan engines. Because of this, the engines require less fuel to produce each unit of thrust and can therefore consume less fuel per unit distance flown even though the thrust is greater than for the standard cruise.

2.6 Conclusions and Discussion

We have analyzed aircraft cruise at constant altitude with a model with mass and speed as state variables. This is a singular optimal control problem and we have identified the singular arc for a weighted sum of time and range for a variety of propulsion systems. The singular arcs have been specified for six special cases; the combinations of endurance (maximum time) and maximum range for three propulsion systems: rockets, turbo and fanjets, and internal combustion/propeller. Example calculations using the Boeing 747-400 with General Electric CF6-80C2 engines for the maximum range mission show potential large savings in fuel. However, if steady cruise is optimized to maximize range, the difference in range between steady and singular arc flight is very small. In spite of this discouraging result, there are interesting avenues for future research.

First, the singular arcs of other propulsion systems need to be determined and evaluated numerically. Second, the endurance mission needs to be investigated. In view of the fact that non-steady flight paths have shown significant improvements relative to steady ones, there could be big improvements in flying singular arc paths. Third, the second-order necessary condition needs to be derived and numerically evaluated for all missions and propulsion systems.

Further study is required to determine the fuel consumption, range, and time over an entire flight. The boundary conditions on the cruise portion of the flight are different from the steady cruise case. For example, for the range mission, the singular arc cruise ends at a lower speed than it begins and thus the range gained in descent will be shorter. Thus the fuel consumption in climb and descent must be added to give a fair comparison.

There are obvious air traffic control issues with flying singular arc flight paths in controlled airspace. Mixing singular arc paths with steady cruise paths is clearly not acceptable. Even with all aircraft flying singular arc paths there is the issue of separating aircraft that is decelerating. In many parts of the world, however, airspace is not controlled and flight time is not critical. In such situations, singular arc flight may be employed immediately.

Perhaps the most important application of our results is to aircraft design. Key aircraft parameters (such as wing loading, aspect ratio, and wing sweep) could be chosen to move the singular arc to its optimum location in the mass-speed plane. The performance criteria should be a suitably weighted combination of fuel consumption and flight time so as to minimize direct operating cost.

Acknowledgments The authors thank Doug Pargett for the use of his computer program and for valuable discussions. We also thank P.K. Menon for suggesting the endurance mission and for pointing out the paper by Mancill.

References

1. Menon, P.K., Study of Aircraft Cruise, *Journal of Guidance, Control and Dynamics*, Vol. 12, No. 5, Sept–Oct 1989, pp. 631–639.
2. Sachs, G. and Christodoulou, T., Reducing Fuel Consumption of Subsonic Aircraft by Optimal Cyclic Cruise, *Journal of Aircraft*, Vol.24, No.9, 1987, pp. 616–622.
3. Miele, A., *The Calculus of Variations in Applied Aerodynamics and Flight Mechanics*, Optimization Techniques, edited by G. Leitmann, Academic Press, New York, 1962, pp. 100–171.
4. Pargett, Douglas and Ardema, Mark, Flight Path Optimization at Constant Altitude, Santa Clara University.
5. Leitmann, G., *The Calculus of Variations and Optimal Control*, Plenum Press, New York, 1981, pp. 225–237.
6. Bryson, A. and Ho, Y., *Applied Optimal Control*, Taylor and Francis, New York, 1975, pp. 246–270.
7. Miele, A., Problems of Minimum Time in Nonsteady Flight of Aircraft, *Atti della Accademia delle Scienze di Torino, Classe di Scienze Fisiche, Matematiche e Naturali*, Vol. 85, 1951, pp. 41–52.
8. Mancill, J. D., Identically Non-Regular Problems in the Calculus of Variations, *Mathematica Y Fisica Teorica*, Universidad Nacional del Tucuman, Republica Argentina, Vol.7, No.2, June 1950, pp. 131–139.
9. Miele, A., Extremization of Linear Integrals by Green's Theorem, Optimization Techniques, edited by G. Leitmann, Academic Press, New York, 1962, pp. 69–99.
10. Holt, Ashley, *Engineering Analysis of Flight Vehicles*, Addison-Wesley, New York, 1974, pp. 113–114.
11. NASA Glenn EngineSim, Ver. 1.6e [online application] <http://www.grc.nasa.gov/WWW/K-12/airplane/ngnsim.html> [cited 2 December 2004].
12. General Electric Aircraft Engines website: GE Transportation Aircraft Engines: CF6.: <http://www.geae.com/engines/commercial/cf6/cf6-80c2.html> [cited 2 December 2004].
13. Boeing Aircraft Company website: Boeing 747 Family <http://www.boeing.com/commercial/747family/technology.html> [cited 2 December 2004].
14. Jane's All the World's Aircraft 1997-98, 1988-89, Jane's Information Group Limited, Sentinel House, 1998.

Chapter 3

A Survey on the Newton Problem of Optimal Profiles

Giuseppe Buttazzo

Abstract This chapter aims to present a survey on some recent results about one of the first problems in the calculus of variations, namely Newton's problem of minimal resistance. Many variants of the problem can be studied, in relation to the various admissible classes of domains under consideration and to the different constraints that can be imposed. Here we limit ourselves essentially to the convex case. Other presentations in the workshop will deal with other kinds of domains.

3.1 Introduction

Finding the profile of a body that gives the minimal (aerodynamic or hydrodynamic) resistance to the motion is one of the first problems in the theory of the calculus of variations. In 1685 Sir Isaac Newton studied this problem presenting a very simple model to compute the resistance of a body moving through an inviscid and incompressible medium. In his words (from his *Principia Mathematica*),

If in a rare medium, consisting of equal particles freely disposed at equal distances from each other, a globe and a cylinder described on equal diameter move with equal velocities in the direction of the axis of the cylinder, (then) the resistance of the globe will be half as great as that of the cylinder. ... I reckon that this proposition will be not without application in the building of ships.

The history of this problem is well documented for instance in the book by Goldstine [12], and the problem can be roughly described as follows. Suppose a body moves with a given constant velocity through a fluid and suppose that the body covers a prescribed maximal cross section (orthogonal to the velocity vector) at its rear end: find the shape of the body which provides the minimal resistance.

Giuseppe Buttazzo

Università di Pisa, Dipartimento di Matematica, Largo B. Pontecorvo, 5, 56127 Pisa, Italy,
e-mail: buttazzo@dm.unipi.it

Of course, the solution depends on how we define the resistance of a body: the assumptions that Newton made in order to simplify the problem were the following:

- the fluid is composed by particles that are *mutually independent* and that move at a constant speed and velocity parallel to the stream direction;
- the resistance is only due to the shock interactions between the fluid particles and the surface of the body, and these shocks obey the usual laws governing *perfectly elastic* shocks;
- all other effects as *tangential friction*, *vorticity*, *turbulence* are neglected.

The assumptions above make the Newton's model a rather crude approximation to real physics; however, it appears to provide good results in the following situations: for a body in a rarefied gas with low speed, for bodies which move in an ideal gas with high Mach number, and for slender bodies.

The reader can find in Miele [21] a deep discussion about the conditions under which the assumptions above are fulfilled for realistic aerodynamical problems, as well as several variants of the Newton optimal profile problem (see also Hayes and Probstein [14] for applications to hypersonic aerodynamics).

Under the assumptions above it is easy to deduce the so-called *Newtonian sine-squared pressure law* which states that the pressure coefficient is proportional to $\sin^2 \vartheta$, ϑ being the inclination of the body contour with respect to the free-stream direction (see Fig. 3.1).

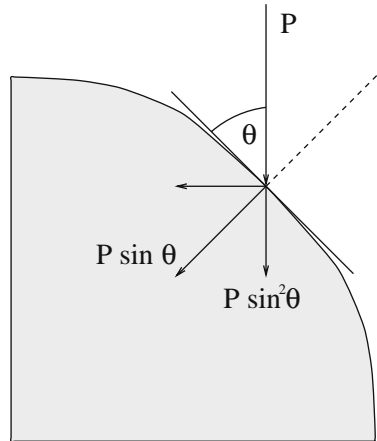


Fig. 3.1 The Newtonian sine-squared pressure law

If we denote by Ω the maximal cross section (orthogonal to the free-stream direction that we assume to be vertical downwards) and describe the front end of the body by a function $u(x)$, with $x \in \Omega$, we obtain that the shock occurring at the point $(x, u(x))$ provides a momentum, which slows the body down, proportional to $(1 + |\nabla u(x)|^2)^{-1}$. If we further assume that *each particle hits the body only once* after some easy calculations we obtain that the total resistance functional can be expressed as

$$\rho v^2 \int_{\Omega} \frac{1}{1 + |\nabla u|^2} dx$$

where ρ is the density of the fluid and v its velocity. Introducing the integral functional

$$F(u) = \int_{\Omega} \frac{1}{1 + |Du|^2} dx$$

we are then reduced to study the minimization problem

$$\min \left\{ F(u) : u \text{ admissible} \right\}. \quad (3.1)$$

Note that the integral functional F above is neither convex nor coercive. Therefore, obtaining an existence theorem for minimizers via the usual direct methods of the calculus of variations, based on weak lower semicontinuity and coercivity, may fail.

The determination of the admissible classes for problem (3.1) is a delicate issue. First of all we notice that without any constraint on the admissible functions u the problem above is meaningless: indeed, if we consider for every integer n the profiles given by the functions

$$u_n(x) = n \operatorname{dist}(x, \partial\Omega),$$

we easily deduce that

$$\lim_{n \rightarrow \infty} F(u_n) = 0$$

and so no minimizer may exist because this would imply $\inf F = 0$, while the functional F only assumes strictly positive values.

Therefore, a constraint on the maximal height, like $0 \leq u \leq M$, has to be added. However, without extra geometric assumptions, even this constraint does not provide the existence of an optimal profile. In fact, the sequence of functions

$$u_n(x) = M \sin^2(n|x|)$$

satisfies the constraint $0 \leq u_n \leq M$ but we still have

$$\lim_{n \rightarrow +\infty} F(u_n) = 0,$$

and by the same argument used before we may conclude that again the resistance functional F does not admit any minimizer in the considered class.

In order to fulfill the physical assumption that the fluid particles hit the body only once we restrict the analysis only to convex bodies, which turns out to consider as admissible the functions u which are bounded and concave on Ω . More precisely, we study the minimization problem

$$\min \left\{ F(u) : 0 \leq u \leq M, u \text{ concave on } \Omega \right\}. \quad (3.2)$$

We shall see in Sect. 3.3 that the concavity constraint on u is strong enough to provide an extra compactness which implies the existence of a minimizer. On the other hand, from the physical point of view, a motivation for this constraint is that,

thinking of the fluid as composed by many independent particles, each particle hits the body only once. If the body is not convex, it could happen that a particle hits the body more than once, but since $F(u)$ was constructed to measure only the resistance due to the first shock, it would no longer reflect the total resistance of the body.

Other kinds of constraints different from the bound on the maximal height $0 \leq u \leq M$ can be imposed on the class of nonnegative concave functions: for instance, we may consider a bound on the surface area of the body, like

$$\int_{\Omega} \sqrt{1 + |\nabla u|^2} dx + \int_{\partial\Omega} u dH^{n-1} \leq c,$$

or on its volume, like

$$\int_{\Omega} u dx \leq c.$$

We refer to some recent papers [1, 15, 28] for a detailed analysis on these other classes of convex bodies.

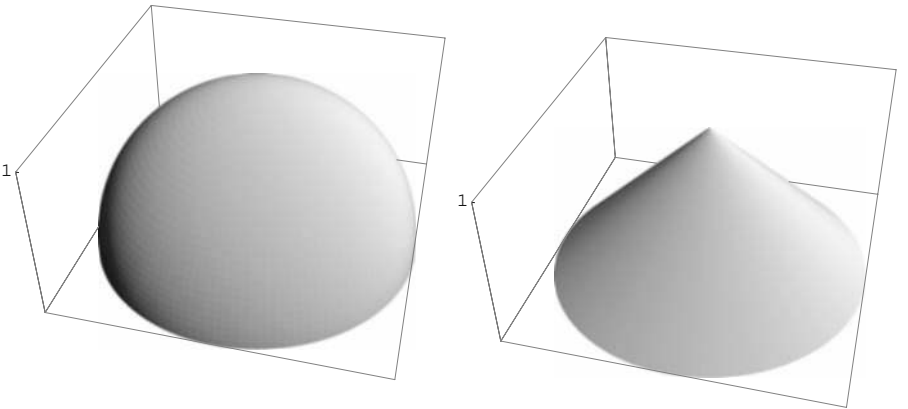


Fig. 3.2 (a) Half-sphere, (b) cone

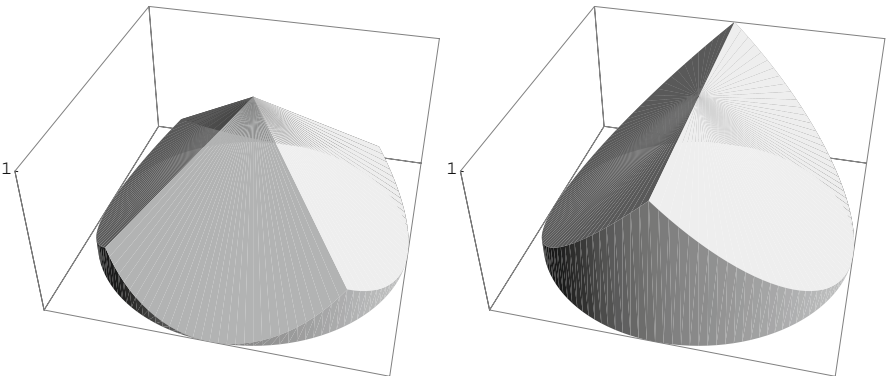


Fig. 3.3 (a) Pyramid 1, (b) pyramid 2

On the other hand, it is interesting to study the optimization of aerodynamical profiles also in a nonconvex framework; several classes of nonconvex bodies have been considered in the literature, with a different expression of the total resistance functional, and we refer to [4, 9, 10, 22] for all the relative details.

We may also define the *relative resistance* of a profile u , dividing the resistance $F(u)$ by the measure of the cross section Ω :

$$C_0(u) = \frac{F(u)}{|\Omega|}.$$

It is clear that we always have $0 \leq C_0(u) \leq 1$ and $C_0(u) = 1$ only if u is constant, that is, the profile is flat. In particular, if the body is a half-sphere of radius R we have $u(x) = \sqrt{R^2 - |x|^2}$ and an easy calculation gives the relative resistance

$$C_0(u) = \frac{F(u)}{\pi R^2} = 0.5$$

as predicted by Newton in 1685. Other bodies with the same value of C_0 are illustrated in Figs. 3.2 and 3.3.

In the next sections we analyse several issues about the optimization problem (3.2) together with a list of still open questions.

3.2 Radially Symmetric Profiles

The most studied case of the Newton problem of profile with minimal resistance is when the competing functions are supposed a priori with a radial symmetry, that is, the cross section Ω is a two-dimensional disk of radius R and the functions u which describe the profile only depend on the radial variable $r = |x|$. This is the case considered by Newton in 1685 and studied in many classical treatises in the calculus of variations (see for instance Funk [11], Kneser [16], Tonelli [26]). In this case, after integration in polar coordinates, the functional F can be written in the form

$$F(u) = 2\pi \int_0^R \frac{r}{1 + |u'(r)|^2} dr$$

so that the resistance minimization problem becomes

$$\min \left\{ \int_0^R \frac{r}{1 + |u'(r)|^2} dr : u \text{ concave}, 0 \leq u \leq M \right\}. \quad (3.3)$$

Several facts about the radial Newton problem can be shown; here we simply list them by referring to the several papers on the subject (see References) for all details.

- It is possible to show that the minimization problem (3.3) admits a solution u which satisfies the conditions $u(0) = M$ and $u(R) = 0$; moreover the optimal radial solution is unique.

- The minimum in (3.3) does not change if we minimize over the larger class of decreasing functions. Therefore, problem (3.3) can also be written in the form

$$\min \left\{ \int_0^R \frac{r}{1 + |u'(r)|^2} dr : u \text{ decreasing, } u(0) = M, u(R) = 0 \right\}. \quad (3.4)$$

Notice that, when the function u is not absolutely continuous, the symbol u' under the integral in (3.4) stands for the absolutely continuous part of u' .

- By using the functions $v(t) = u^{-1}(M - t)$, problem (3.4) can be rewritten in the more traditional form:

$$\min \left\{ \int_0^M \frac{vv'^3}{1 + v'^2} dr : v \text{ increasing, } v(0) = 0, v(M) = R \right\}. \quad (3.5)$$

Again, when v is a general increasing function, v' is a nonnegative measure, and (3.5) has to be intended in the sense of BV functions, as

$$\int_0^M \frac{vv'_a{}^3}{1 + v'_a{}^2} dt + \int_{[0,M]} vv'_s = \frac{R^2}{2} - \int_0^M \frac{vv'_a}{1 + v'_a{}^2} dt. \quad (3.6)$$

where v'_a and v'_s are, respectively, the absolutely continuous and singular parts of the measure v' with respect to Lebesgue measure. The equality in (3.6) has been obtained by replacing the product vv'_s by $vv' - vv'_a$.

- The minimization problem (3.4) admits the following Euler–Lagrange equation in its integrated form:

$$ru' = C(1 + u'^2)^2 \quad \text{on } \{u' \neq 0\} \quad (3.7)$$

for a suitable constant $C < 0$. From (3.7) the solution u can actually be explicitly computed in the parametric form, obtaining $u(r) = M$ on the interval $[0, r_0]$ and

$$\begin{cases} r(t) = \frac{r_0}{4t}(1 + t^2)^2 \\ u(t) = M - \frac{r_0}{4} \left(-\frac{7}{4} + \frac{3}{4}t^4 + t^2 - \ln t \right) \end{cases} \quad \forall t \in [1, T].$$

Here the quantities r_0 and T are defined through the strictly increasing function

$$f(t) = \frac{t}{(1 + t^2)^2} \left(-\frac{7}{4} + \frac{3}{4}t^4 + t^2 - \ln t \right) \quad \forall t \geq 1$$

by setting

$$T = f^{-1}(M/R), \quad r_0 = \frac{4RT}{(1 + T^2)^2}.$$

Notice that $|u'(r)| > 1$ for all $r > r_0$ and that $|u'(r_0^+)| = 1$; in particular, the derivative $|u'|$ never belongs to the interval $]0, 1[$.

- The optimal relative resistance C_0 of a radial body is then given by

$$C_0 = \frac{2}{R^2} \int_0^R \frac{r}{1+u'^2} dr$$

where u is the optimal solution above. We have $C_0 \in [0, 1]$ and it is easy to see that C_0 depends on M/R only. Some approximate calculations give

	$M/R = 1$	$M/R = 2$	$M/R = 3$	$M/R = 4$
r_0/R	0.35	0.12	0.048	0.023
C_0	0.37	0.16	0.082	0.049

- The following asymptotic estimates as $M/R \rightarrow +\infty$ hold:

$$\begin{aligned} r_0/R &\approx \frac{27}{16}(M/R)^{-3} & \text{as } M/R \rightarrow +\infty \\ C_0 &\approx \frac{27}{32}(M/R)^{-2} & \text{as } M/R \rightarrow +\infty. \end{aligned} \quad (3.8)$$

- Some optimal radial shapes for different values of the ratio M/R are shown in Figs. 3.4, 3.5 and 3.6.
- It is interesting to notice that the optimal frustum cone, that is, the frustum cone with height M , cross section radius R and minimal resistance, is only slightly less performant than the optimal radial body computed above. Indeed, its top radius \hat{r}_0 and its relative resistance \hat{C}_0 can be easily computed, and we find

$$\hat{C}_0 = \frac{\hat{r}_0}{R} = 1 - \frac{(M/R)^2}{2} (\sqrt{1 + 4(M/R)^{-2}} - 1),$$

with asymptotic behaviour

$$\hat{C}_0 \approx (M/R)^{-2} \quad \text{as } M/R \rightarrow +\infty.$$

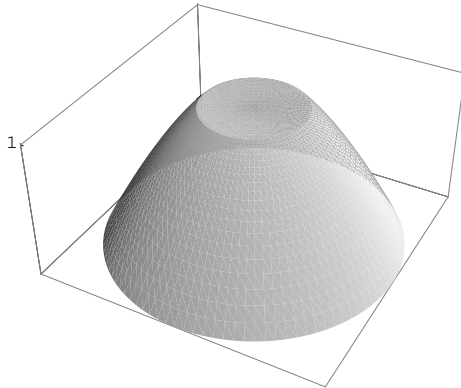


Fig. 3.4 The optimal radial shape for $M = R$

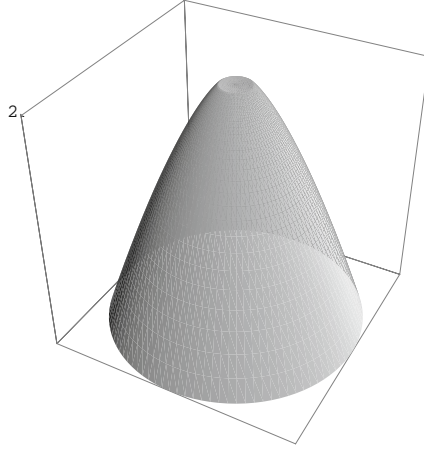


Fig. 3.5 The optimal radial shape for $M = 2R$

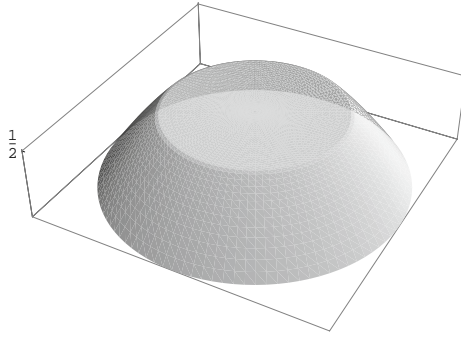


Fig. 3.6 The optimal radial shape for $M = R/2$

3.3 The Existence Result

We shall see here that in the case of a general cross section Ω it is still possible to show the existence of an optimal profile, even if little is known about its qualitative behaviour. We shall see that a necessary condition of optimality is that the optimal profile must be flat, in the sense that $\det D^2 u$ identically vanishes where u is of class C^2 . In particular, when Ω is a disk, this excludes the radial Newton solution and so the optimal solution cannot be radial. This also shows that the solution is not unique in general. Up to now it is not known if optimal solutions always have a *flat nose* and if they always assume the value zero at the boundary.

Denoting then by $C_M(\Omega)$ the class of concave functions on Ω that fulfill the inequalities $0 \leq u \leq M$ and by F the functional

$$F(u) = \int_{\Omega} \frac{1}{1 + |\nabla u|^2} dx,$$

we are concerned with the minimization problem

$$\min \left\{ F(u) : u \in C_M(\Omega) \right\}. \quad (3.9)$$

Note that, since every bounded concave function is locally Lipschitz continuous in Ω , the functional F in (3.9) is well defined on C_M . Moreover, as a consequence of Fatou's lemma, the functional F is lower semicontinuous with respect to the strong convergence of every Sobolev space $W^{1,p}(\Omega)$ or also $W_{loc}^{1,p}(\Omega)$.

The proof of the existence theorem for problem (3.9) relies on the following compactness result for the class $C_M(\Omega)$ (see [20]).

Lemma 3.1. *For every $M > 0$ and every $p < +\infty$ the class $C_M(\Omega)$ is compact with respect to the strong topology of $W_{loc}^{1,p}(\Omega)$.*

This allows us to apply successfully the direct methods of the calculus of variations and to obtain the following general result.

Theorem 3.1. *Let $f : \Omega \times \mathbb{R} \times \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$ be a function such that*

- (i) *f is nonnegative and measurable for the σ -algebra $\mathcal{L}_N \otimes \mathcal{B} \otimes \mathcal{B}_N$;*
- (ii) *for a.e. $x \in \Omega$ the function $f(x, \cdot, \cdot)$ is lower semicontinuous on $\mathbb{R} \times \mathbb{R}^N$. Then for every $M > 0$ the minimum problem*

$$\min \left\{ \int_{\Omega} f(x, u, \nabla u) dx : u \in C_M(\Omega) \right\} \quad (3.10)$$

admits at least a solution.

Remark 3.1. It is interesting to notice that in the existence theorem above no convexity assumptions with respect to ∇u on the integrand f are made. This is because the convexity is related to the lower semicontinuity of the cost functional F for the weak convergence of Sobolev spaces (see for instance Buttazzo [3]), while in our case, thanks to Lemma 3.1, we may work with the strong convergence. This approach can be used in different situations and a similar result can be obtained (see [4]), even if less justified physically, in the larger class of superharmonic functions:

$$E_M(\Omega) = \{u \in H_{loc}^1(\Omega) : 0 \leq u \leq M, \quad \Delta u \leq 0 \text{ in } \Omega\}.$$

Remark 3.2. As already mentioned, other constraints than prescribing the maximal height M of the body are possible, still keeping the convexity of the admissible bodies as a general requirement. For instance, if we prescribe a bound V on the volume of the body, we deal with the admissible class

$$C^V(\Omega) = \{u : \Omega \rightarrow \mathbb{R} : u \text{ concave}, u \geq 0, \int_{\Omega} u dx \leq V\}.$$

Alternatively, we can prescribe a bound S on the side surface of the body, so that the admissible class becomes

$$C(S, \Omega) = \{u : \Omega \rightarrow \mathbb{R} : u \text{ concave}, u \geq 0, \int_{\Omega} \sqrt{1 + |\nabla u|^2} dx \leq S\}.$$

In both cases we have a compactness result similar to the one of Lemma 3.1 and consequently an existence result similar to the one of Theorem 3.1. Indeed, if u is concave its sup-norm can be estimated in terms of its integral, as is easily seen by comparing the body itself with the cone of equal height:

$$V \geq \int_{\Omega} u dx \geq \frac{(\sup u) \text{meas}(\Omega)}{N+1}.$$

Then the volume class $C^V(\Omega)$ is included in the height class $C_M(\Omega)$ where $M = V(N+1)/\text{meas}(\Omega)$ and the corresponding compactness result follows from the one of Lemma 3.1.

The case of surface bound is similar: indeed, the sup-norm of a concave function can be estimated in terms of the surface of its graph, as is easily seen by comparing again the body itself with the cone of equal height:

$$S \geq \int_{\Omega} \sqrt{1 + |\nabla u|^2} dx \geq \frac{(\sup u) \mathcal{H}^{N-1}(\partial\Omega)}{N}.$$

Then the surface class $C(S, \Omega)$ is included in the height class $C_M(\Omega)$ where $M = SN/\mathcal{H}^{N-1}(\partial\Omega)$ and the corresponding compactness result again follows from the one of Lemma 3.1.

Once obtaining the existence result above we deal now with the question of deducing some necessary conditions of optimality. As in the radial case, it is possible to show that the slope $|\nabla u|$ of the solution is never in $]0, 1[$.

Theorem 3.2. *Let u be a solution of problem (3.9). Then for a.e. $x \in \Omega$ we have that $|\nabla u|(x) \notin]0, 1[$.*

The usual Euler–Lagrange equation gives the following first-order necessary condition of optimality, in the case of a general integrand $f(x, s, z)$.

Theorem 3.3. *Let u be a solution of problem (3.10); we assume that in an open set $\omega \subset \Omega$ the function u is smooth and belongs to the interior of the admissible class, that is*

(a) *u is of class $C^2(\omega)$;*

(b) *the maximal value M of u is not attained in ω ;*

(c) *u is strictly concave in the sense that its Hessian matrix is positive definite.*

We also assume that the integrand $f(x, s, z)$ appearing in (3.10) is sufficiently smooth. Then we have

$$-\text{div}(f_z(x, u, \nabla u)) + f_s(x, u, \nabla u) = 0 \quad \text{in } \omega.$$

In the case of Newton functional we have $f(x, s, z) = (1 + |z|^2)^{-1}$ and the equation above becomes

$$\operatorname{div} \left(\frac{\nabla u}{(1 + |\nabla u|^2)^2} \right) = 0 \quad \text{in } \omega.$$

Under the assumptions of Theorem 3.3 we can also perform the second variation; this gives for every test function ϕ

$$\int_{\omega} [f_{zz}(x, u, \nabla u) \nabla \phi \nabla \phi + 2f_{sz}(x, u, \nabla u) \phi \nabla \phi + f_{ss}(x, u, \nabla u) \phi^2] dx \geq 0.$$

In particular, for the Newton functional we obtain for every ϕ

$$\int_{\omega} \frac{2}{(1 + |\nabla u|^2)^3} \left(4(\nabla u \nabla \phi)^2 - (1 + |\nabla u|^2) |\nabla \phi|^2 \right) dx \geq 0. \quad (3.11)$$

Assume now for simplicity $N = 2$, the cross section Ω , a disk of radius R and let u be the optimal radial solution of the Newton problem computed in Sect. 3.2; we have seen that, outside a circle of radius r_0 where $u \equiv M$, the function u is smooth, strictly concave and does not attain the maximal value M . We are then in the conditions of Theorem 3.3 and then, using in (3.11) a test function ϕ of the form $\eta(r)\psi(\theta)$ with $\operatorname{spt} \eta \subset]r_0, R[$, we obtain

$$\int_{r_0}^R r dr \int_0^{2\pi} \left[\frac{4|u'(r)\eta'(r)\psi(\theta)|^2}{(1 + |u'(r)|^2)^3} - \frac{|\eta'(r)\psi(\theta)|^2 + |\eta(r)\psi'(\theta)|^2 r^{-2}}{(1 + |u'(r)|^2)^2} \right] d\theta \geq 0.$$

The same can be done using $\psi(k\theta)$ instead of $\psi(\theta)$, where k is an integer. In this case the previous inequality becomes

$$\int_{r_0}^R r dr \int_0^{2\pi} \left[\frac{4|u'(r)\eta'(r)\psi(k\theta)|^2}{(1 + |u'(r)|^2)^3} - \frac{|\eta'(r)\psi(k\theta)|^2 + k^2|\eta(r)\psi'(k\theta)|^2 r^{-2}}{(1 + |u'(r)|^2)^2} \right] d\theta \geq 0.$$

Letting $k \rightarrow +\infty$ gives then a contradiction and, by consequence, the following result.

Theorem 3.4. *Let Ω be a circle. Then an optimal solution of the Newton problem*

$$\min \left\{ \int_{\Omega} \frac{1}{1 + |\nabla u|^2} dx : u \in C_M \right\} \quad (3.12)$$

cannot be radial.

Remark 3.3. An immediate consequence of the nonradiality of the optimal Newton solutions is that problem (3.12) does not have a unique solution. In fact, rotating any nonradial solution u provides still a solution, as it is easy to verify, and therefore the number of solutions of problem (3.12) is infinite. It is not clear if a lack of symmetry in the domain Ω provides the uniqueness of the optimal solution u .

The fact that optimal profiles with circular cross section do not need to be radially symmetric can also be proved by exhibiting nonsymmetric profiles which are

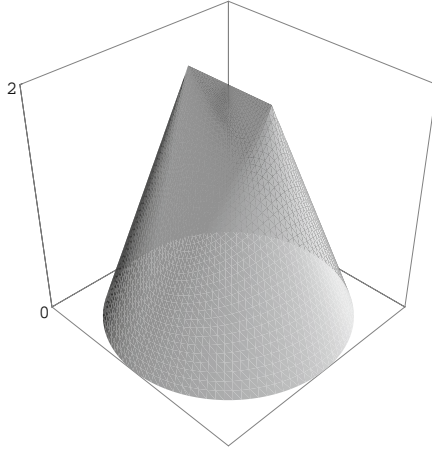


Fig. 3.7 A nonradial profile better than the optimal radial one

more performant than the optimal radial one (Fig. 3.7). This was first discovered by Guasoni in [13], who considered a body of the form obtained as the convex envelope of the set $(\Omega \times \{0\}) \cup (S \times \{M\})$ where S is a segment. Choosing in a suitable way the length of the segment S which represents the set $\{u = M\}$, we can compute the resistance of the profile and we have, taking into account the asymptotic estimates (3.8) seen in Sect. 3.2, that as $M/R \rightarrow +\infty$

$$F(u) \approx 0.77(M/R)^{-2} < \frac{27}{32}(M/R)^{-2} \approx F(u_{rad}).$$

Therefore, as M/R is large enough (larger than 2 in the Guasonio computation) the body above has a better performance than the optimal radial one, hence the optimal profile cannot be radially symmetric.

It remains to identify the optimal solutions. Surprisingly, we have that the optimal profiles have to be “flat” in the sense that the Hessian of optimal solutions u vanishes. More precisely, the following result holds.

Theorem 3.5. *Assume that u is an optimal solution for the Newton problem (3.9) which is of class C^2 in an open set $\omega \subset \Omega$ and that $u < M$ in ω . Then we have*

$$\det \nabla^2 u \equiv 0 \quad \text{in } \omega. \quad (3.13)$$

The proof of the result above can be easily obtained by contradiction. In fact, if the conclusion does not hold in a point $x_0 \in \omega$, since u is concave and of class C^2 we must have that the Hessian matrix $\nabla^2 u$ is negative definite in a neighbourhood U of x_0 . We may then perform the second variation argument, obtaining

$$\int_U \frac{2[4(\nabla u \cdot \nabla \phi)^2 - (1 + |\nabla u|^2)|\nabla \phi|^2]}{(1 + |\nabla u|^2)^3} dx \geq 0$$

for every test function ϕ with support in U . If a is a unitary vector orthogonal to $\nabla u(x_0)$, we choose a test function of the form

$$\phi(x) = \eta(x) \sin(ka \cdot x)$$

where k is an integer and η is a smooth function supported in a small neighbourhood of x_0 . Since

$$\nabla \phi(x) = \sin(ka \cdot x) \nabla \eta(x) + ka \cos(ka \cdot x) \eta(x)$$

passing to the limit as $k \rightarrow +\infty$ we obtain

$$\int_U \frac{2[4(a \cdot \nabla u)^2 - (1 + |\nabla u|^2)] \eta^2(x)}{(1 + |\nabla u|^2)^3} dx \geq 0$$

for every η . Letting now the support of η shrink to $\{x_0\}$ we find a contradiction, since $a \cdot \nabla u(x_0) = 0$.

Remark 3.4. The result of Theorem 3.5 gives once more the nonradiality of all optimal solutions. Indeed, the optimal radial functions u_{rad} do not satisfy the flatness condition (3.13).

The characterization of optimal Newton profiles (Figs. 3.8, 3.9, 3.10 and 3.11) is still an open question; the convexity constraint makes numerical computations rather difficult. In particular it is not clear if the upper region $\{u = M\}$ has dimension two or it reduces to a segment, and if the optimal solutions u are regular in the region $\{u < M\}$. The numerical computations below (taken from [17]) seem to disprove this last fact, but a rigorous proof is still missing.

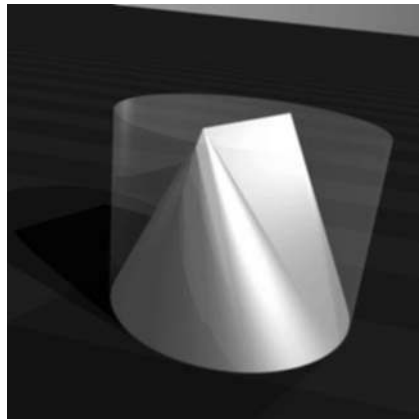


Fig. 3.8 A rather high optimal profile

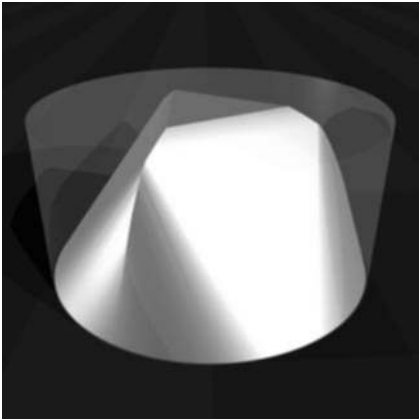


Fig. 3.9 A lower optimal profile

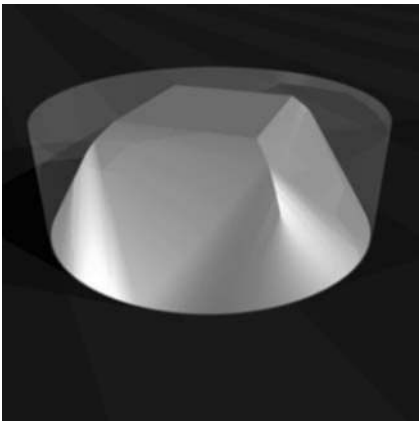


Fig. 3.10 A still lower optimal profile

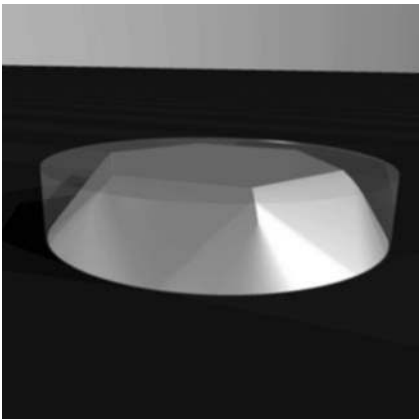


Fig. 3.11 A rather low optimal profile

References

1. M. Belloni, A. Wagner: *Newton's problem of minimal resistance in the class of bodies with prescribed volume*. J. Convex Anal., **10** (2) (2003), 491–500.
2. F. Brock, V. Ferone, B. Kawohl: *A symmetry problem in the calculus of variations*. Calc. Var., **4** (6) (1996), 593–599.
3. G. Buttazzo: *Semicontinuity, Relaxation and Integral Representation in the Calculus of Variations*. Pitman Res. Notes Math. Ser. **207**, Longman, Harlow (1989).
4. G. Buttazzo, V. Ferone, B. Kawohl: *Minimum problems over sets of concave functions and related questions*. Math. Nachr., **173** (1995), 71–89.
5. G. Buttazzo, M. Giaquinta, S. Hildebrandt: *One-dimensional Calculus of Variations: an Introduction*. Oxford University Press, Oxford (1998).
6. G. Buttazzo, P. Gusoni: *Shape optimization problems over classes of convex domains*. J. Convex Anal., **4** (1997), 343–351.
7. G. Buttazzo, B. Kawohl: *On Newton's problem of minimal resistance*. Math. Intell., **15** (1993), 7–12.
8. G. Carlier, T. Lachand-Robert: *Regularity of solutions for some variational problems subject to a convexity constraint*. Comm. Pure Appl. Math., **54** (5) (2001), 583–594.
9. M. Comte, T. Lachand-Robert: *Existence of minimizers for Newton's problem of the body of minimal resistance under a single impact assumption*. J. Anal. Math., **83** (2001), 313–335.
10. M. Comte, T. Lachand-Robert: *Newton's problem of the body of minimal resistance under a single-impact assumption*. Calc. Var. Partial Dif. Equations, **12** (2) (2001), 173–211.
11. P. Funk: *Variationsrechnung und ihre Anwendungen in Physik und Technik*. Grundlehren **94**, Springer-Verlag, Heidelberg (1962).
12. H. H. Goldstine: *A History of the Calculus of Variations from the 17th through the 19th Century*. Springer-Verlag, Heidelberg (1980).
13. P. Guasoni: *Problemi di ottimizzazione di forma su classi di insiemi convessi*. Tesi di Laurea, Università di Pisa, 1995–1996.
14. W. D. Hayes, R. F. Probstein: *Hypersonic Flow Theory*. Academic Press, New York (1966).
15. D. Horstmann, B. Kawohl, P. Villaggio: *Newton's aerodynamic problem in the presence of friction*. NoDEA Nonlinear Diff. Equations Appl., **9** (3) (2002), 295–307.
16. A. Kneser: *Ein Beitrag zur Frage nach der zweckmäßigsten Gestalt der Geschößspitzen*. Archiv der Mathematik und Physik, **2** (1902), 267–278.
17. T. Lachand-Robert, E. Oudet: *Minimizing within convex bodies using a convex hull method*. SIAM J. Optimiz., **16** (2) (2005), 368–379.
18. T. Lachand-Robert, M. A. Peletier: *An example of non-convex minimization and an application to Newton's problem of the body of least resistance*. Ann. Inst. H. Poincaré Anal. Non Linéaire, **18** (2) (2001), 179–198.
19. T. Lachand-Robert, M. A. Peletier: *Newton's problem of the body of minimal resistance in the class of convex developable functions*. Math. Nachr., **226** (2001), 153–176.
20. P. Marcellini: *Nonconvex integrals of the calculus of variations*. In “Methods of Nonconvex Analysis” (Varenna, 1989), Lecture Notes in Math. **1446**, Springer-Verlag, Berlin (1990), 16–57.
21. A. Miele: *Theory of Optimum Aerodynamic Shapes*. Academic Press, New York (1965).
22. A. Yu. Plakhov: *Newton's problem of the body of least resistance with a bounded number of collisions*. Russian Math. Surveys, **58** (1) (2003), 191–192.
23. A. Yu. Plakhov: *Newton's problem of minimal resistance for bodies containing a half-space*. J. Dynam. Control Systems, **10** (2) (2004), 247–251.
24. A. Yu. Plakhov: *Newton's problem of the body of minimal mean resistance*. Sb. Math., **195** (7–8) (2004), 1017–1037.
25. A. Yu. Plakhov, D. F. M. Torres: *Newton's aerodynamic problem in media of chaotically moving particles*. Sb. Math., **196** (5–6) (2005), 885–933.
26. L. Tonelli: *Fondamenti di Calcolo delle Variazioni*. Zanichelli, Bologna (1923).

27. N. Van Goethem: *Variational problems on classes of convex domains*. Commun. Appl. Anal., **8** (3) (2004), 353–371.
28. A. Wagner: *A remark on Newton's resistance formula*. ZAMM Z. Angew. Math. Mech., **79** (6) (1999), 423–427.

Chapter 4

Innovative Rotor Blade Design Code

Vittorio Caramaschi and Claudio Monteggia

Abstract The competitive advantage in helicopter world market is to develop a rotor design ‘tailored’ on specific, more demanding performances such as higher cruise speed and higher cruise altitude, but, at the same time, guaranteeing the maximum level of comfort for the crew and the passengers. To achieve this goal, it is normal practice to apply some design rules to the rotor aeromechanic behaviour but the residual hub loads transferred to the supporting pylon can still be so high that, in order to meet the desired threshold of the vibratory level, some vibration absorbers have to be installed as well. The reason for this has been up to now the poor to weak prediction capability of the vibratory rotor loads due to the incomplete knowledge in the rotor wake modelling and in the aerodynamics and structural interactions which are the sources of vibratory forces.

To overcome these difficulties AW has developed a new aeroelastic code, called GYROX II, FEM based, capable of representing any complex blade shape and hub/control system/pylon features. Details of the code, together with several results of the application of the code to twin-engine light-medium helicopters, are presented. Short- and medium-term upgrading of the code in order to become more attractive design tools in an integrated aeromechanics and flight mechanics environment is finally faced.

Vittorio Caramaschi
AgustaWestland, Cascina Costa (VA) , Italy,
e-mail: vittorio.caramaschi@agustawestland.com

Claudio Monteggia
AgustaWestland, Cascina Costa (VA), Italy,
e-mail: claudio.monteggia@agustawestland.com

4.1 Introduction

Today helicopter world market assets, following the recent significant expansion of the usage of such a vehicle, can be identified in the level of productivity and comfort of the offered class of products.

The word productivity is well expressed by the product of the payload and the maximum speed achievable or, alternatively, the range covered; the improvement of any of these parameters, leaving the other unchanged, is leading to a better productivity (Fig. 4.1).

An improvement of the productivity can be furthermore achieved by expanding the type of mission profile which can be flown: in this sense a tiltrotor (Fig. 4.2), doubling the maximum cruise speed and flying, like a turboprop, at higher cruise altitude (7500 m), but, like an helicopter, not requiring more than a small vertiport to takeoff and land, is going to offer a very attractive platform 'point to point' mission.

But the achievement of the target above cannot be claimed if the design of the vehicle is not improved as well in terms of vibroacoustic comfort and exterior noise produced.

The vibratory level, generated mostly by the main rotor, has to be minimized at any speed, both in the cockpit area, to reduce the required skill of the pilot and



Fig. 4.1 A109E helicopter



Fig. 4.2 BA609 tiltrotor

alleviate his workload, and in the main cabin, to provide the maximum comfort to the passengers. To fulfil the goal completely or, in other words, to maximize the benefit in terms of perception, the interior noise, this time generated mostly by the main gearbox, has to be reduced as well.

4.2 Helicopter's Aeromechanics Outlines

The main rotor, for a conventional helicopter configuration, is the primary source of vibratory forces; the genesis of these forces lies on the aerodynamic airloads of the blades and their aeroelastic interaction with the dynamics of the blades themselves (Fig. 4.3).

In terms of pure aerodynamic environment, these major features can be highlighted:

1. large fluctuations of local apparent wind, Mach number and, as a consequence, aerodynamic incidence, especially at very high forward speed due to the free stream anisotropy;
2. additional local apparent wind fluctuations produced by the interaction of the wake vorticity released by each blade with the other blades;
3. significant interaction, in terms of the wake spatial distribution, due to fuselage interference, varying with forward speed.

These three topics already clearly show why an helicopter rotor is no longer working in a steady-state aerodynamics environment, and so why the theories developed for the fixed wing aircrafts can just be taken as a reference. In order to provide a valuable prediction, computational tools have to include the unsteady aerodynamics effects up to, at very high speeds, the capability to represent a dynamic

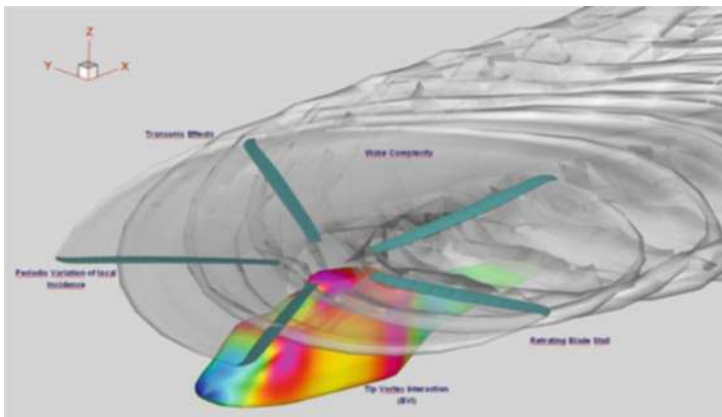


Fig. 4.3 Rotor aerodynamic environment

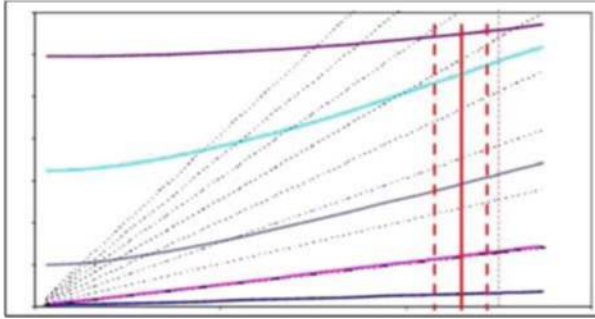


Fig. 4.4 Rotor fanplot

stall condition, significantly different from the static one experienced by airfoil/wing wind tunnel tests.

The airload produced by such a complex field is varying along the blade span but, also, along the azimuth, so making a Fourier analysis of any sectional force, selecting as fundamental harmonic the rotor rotational speed (1/rev), a discrete frequency spectrum with significant harmonics up to 8/rev to 10/rev can be found.

Now, since any rotor system is structurally characterized by some degree of flexibility/elasticity, the resultant vibratory forces are therefore dependent on the interaction (work done) of the aerodynamic airload with the rotor natural modes, in terms of frequency interaction and modal work (Fig. 4.4).

In turn, rotor mode frequencies and intermodal couplings are strongly influenced by the following design issues:

1. blade geometry;
2. blade structural properties;
3. hub features;
4. pylon and control system features.

Blade geometry

The basic parameters are the following:

1. airfoil distribution;
2. built-in twist and chord distributions;
3. sweep and camber distributions.

The required performances are normally largely driving the selection of the airfoils, the twist and the chord distributions along the blade (Fig. 4.5), but a refinement of the local twist and chord may be used, together with sweep and camber parameters to



Fig. 4.5 Rotor blade

- optimize the bound circulation and the consequent trail and skid vorticity generated;
- in particular for outboard blade region, to minimize the strength of the released vorticity and to address its trajectory in order to minimize the effects of the interaction with the other blades, affecting significantly the vibratory forces produced and the pressure peaks generating exterior noise.

Blade structural properties

According to the hub type, the elastic deformation of any blade can be affected, up to some extent, by the following structural properties distribution (Fig. 4.6):

- mass, centre of gravity and mass moment of inertia;
- beamwise and chordwise bending stiffnesses;
- shear centre location and torsional stiffness.

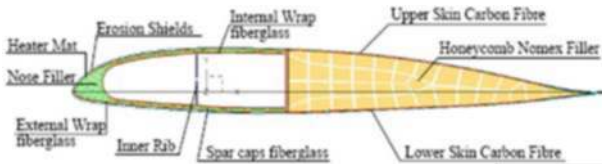


Fig. 4.6 Rotor blade structural section

Since many years the use of composite materials offers a wide design rule in terms of selection of the above parameters.

Basic mass and stiffness distributions normally confine bending modes frequencies, the level of elastic twist coupling being significantly controllable by means of centre of gravity, shear centre and torsional stiffness distributions.

These parameters are also essential to avoid the occurrence of any unstable coupling which could generate an aeroelastic instability (flutter, etc.).

Hub features

The choice of the hub configuration is normally done by an appropriate weighting function of the following attributes:

1. weight;
2. costs, both unit production cost and maintenance;
3. technological risk, in order to avoid excessive loads and/or unstable conditions.

The most common architectures worldwide considered are the following:

1. articulated;
2. hingeless/rigid or bearingless;
3. gimballed/teetering (two blades only).

Articulated hub

The blade is free to move in flap (out-of-plane), lead-lag (in-plane) and pitch motions about a single elastomeric spherical bearing. This hub type, provided that sufficient control stiffness is available, is very powerful in terms of overall dynamic/aeroelastic rotor tuning, and the technological risk is minimized; due to the lead-lag degree of freedom, it is well known that an unstable condition could occur on ground, named ground resonance. This phenomenon can be suppressed if a damper element (typically hydraulic, or elastomeric or fluidlastic) is introduced to damp the in-plane motion of the blades (Fig. 4.7).



Fig. 4.7 Articulated hub configuration

Hingeless/Rigid or Bearingless

From the point of view of weight and cost attributes, it is worldwide recognized to be an attractive choice, whilst, at the same time, from the technological risk viewpoint, is much more demanding in terms of overall effort to be produced (Fig. 4.8).



Fig. 4.8 Hingeless hub configuration

So, the final score in terms of costs and benefits is played largely by the details of the structural layout which is used to allow blade motions.

Normally the blade is allowed to move in flap and lead-lag deforming some elements of the hub system, soft-in-plane rotors still requiring a damper to prevent ground resonance.

As far as the torsional motion is concerned, this can be done using a bearing or, again, by deformation of an appropriate rotor head element.

Gimballed Hub or Teetering

Again with reference to an articulated rotor, an attempt to reduce the overall weight and costs may lead to this hub concept, where the cyclic flap motion of the blades is allowed by a single 3D spherical bearing (gimbal) or a single teetering hinge (for a two-blade rotor), whilst the rotor coning is produced by the deformation of suitable hub substructures (Fig. 4.9).

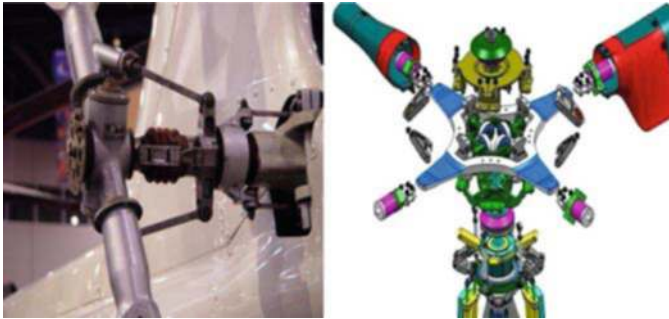


Fig. 4.9 (a) Teetering (b) Gimbal

Also in this case, depending on the in-plane stiffness, a damper may be required: it is nevertheless more common for this hub configuration to get a stiff-in-plane solution.

As far as the torsional motion is concerned, this could be accommodated by a bearing (normally elastomeric) or by an elastic strap element, reacting also the centrifugal force.

Pylon and Control System Features

If, from performances point of view, the pylon is not important and the control system is considered in terms of blade rigid impressed pitch only, from the aeromechanics side, it is well proven that the great sensitivity of vibratory rotor leads to the pylon ‘impedance’ and the control system stiffness (Fig. 4.10).

In other words, the rotor is normally generating larger forces than it would do in the hub fixed condition not only in the fundamental harmonics but also in the overall airframe vibrations producing additional non-harmonic rotor forces.

The word ‘pylon’ could include the main rotor supporting structure down to the cabin roof or even to the entire airframe.

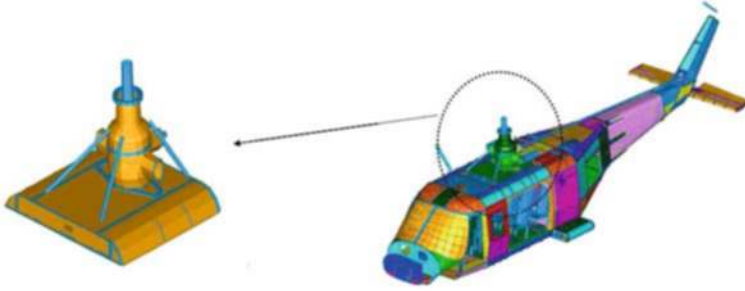


Fig. 4.10 Rotor pylon model

The control system normally includes a fixed and a rotating swashplate: by its nature the stiffness seen by any blade is continuously changing around the azimuth (Fig. 4.11). As already mentioned, the mechanical interaction between the rotating and fixed frames occurs through two different load paths: the rotor mast and the controls.

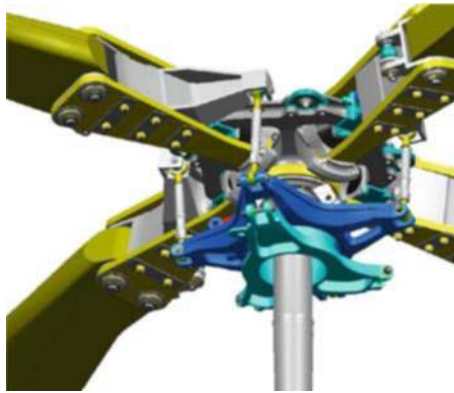


Fig. 4.11 Control chain

4.3 Helicopter's Rotor Mathematical Model Features and Aeromechanics Codes Worldwide Status

To face the vibratory loads issue, the mathematical model has

- to be so general to represent, as it is, any complex blade geometry related to the planform, twist or structural properties
- to be so general and detailed to fully describe rotor head and load paths, including the control system, as they are

- to be so general to allow the accurate description of the mechanical impedance of the pylon system
- to include an aerodynamics formulation consistent with the general finite element approach above and capable of capturing rotor wake features at all flight regimes and the interaction with the fuselage.

The status of the art in terms of vibratory load prediction has been worldwide recognized to be still at a poor to weak level because of the following reasons:

1. incomplete knowledge of the rotor wake, both isolated and including fuselage interaction (specific dedicated experimental tests were done in the past but they were referring mainly to performances rather than vibratory environment): further more complex testing is required to substantiate and validate any new advanced wake modelling;
2. incomplete knowledge of the aerodynamic-elastic mechanisms which are the root of the generated vibratory forces (again experimental campaigns are needed to reach the required level of confidence);
3. incomplete knowledge of the dynamics of the pylon and of the cabin (local vibration environment and modal damping), further complicated by its sensitivity to the manufacturing hardware scattering.

4.4 AW Aeromechanics Code GYROX II

In order to overcome the aforesaid difficulties and improve significantly the prediction of vibratory loads capability, AgustaWestland has developed internally a new aeroelastic methodology, named GYROX II (Fig. 4.12).

The primary task was dealing with rotor vibratory loads in steady-state conditions, but, before doing any forced response, it is straightforward to assess the stability of the equilibrium: this feature has therefore been included since the beginning of the code development.

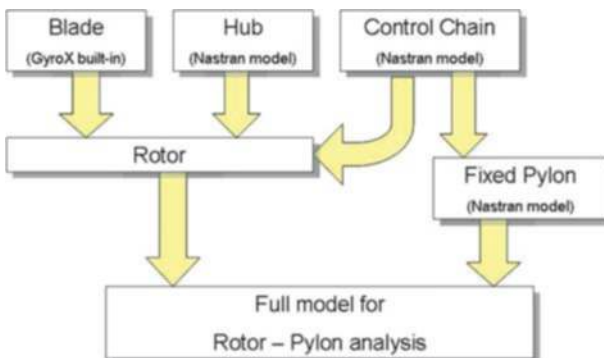


Fig. 4.12 GYROX II model architecture

Recently, in order to directly compare GYROX outputs with measured experimental time histories of any rotor/pylon signal, the capability to perform a transient response (to a gust or a step change of the impressed pitch input) has been added as well.

The use of the code, within any design process, might be seen at the very early stages, with the goal to design the most important rotor/pylon design variables, so, managing a synthetic set of data or in a subsequent step, when most of the data are available and the final aeroelastic requirement is requested.

The basic feature is that any rotor/pylon system can be thought as the assembly of four subsystems, which are the blade, the rotor head, the pylon and the controls, and which are interconnected at specific points, where the boundary conditions have to be preserved.

4.4.1 General Procedure

All the subsystems are FEM described using NASTRAN code, due to its general use in any technological area at AW; but, whilst the blade has a stand-alone input, directly managed by GYROX code, as far as the hub, the control chain and the pylon are concerned, geometrical description as well as their contributions in terms of mass, damping and stiffness matrices are imported from NASTRAN files.

Figures 4.13 and 4.14 show a sketch of the model architecture and GYROX II–NASTRAN interaction:

- whilst B is the subsystem ‘blade’ and S is the other subsystem (including hub, pylon and controls) the matrices can be partitioned as shown at the bottom of Fig. 4.13, in terms of Lagrangian physical degrees of freedom q , the corresponding matrices which, being part of the rotating system and part of the fixed one, include periodic coefficients as a function of the azimuth ψ .
- whilst, again, GYROX is generating, in a stand-alone way, all the linear and non-linear terms related to the rotating blade beam elements, to be added to the

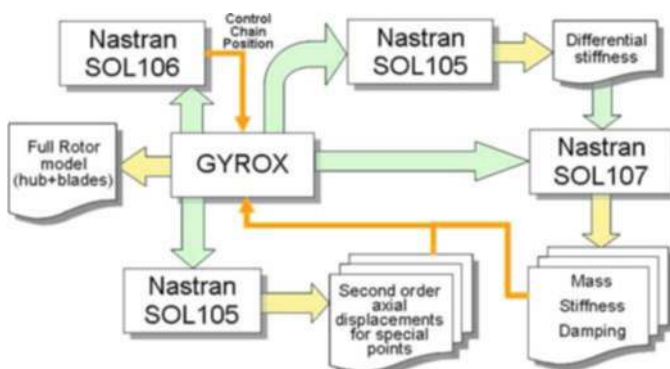


Fig. 4.13 GYROX II–NASTRAN interaction

basic mass and elastic stiffness matrices, the differential stiffnesses of any head element related to the recovery of the centrifugal loading and/or the upgrading of the control circuit geometry with the rotation itself are solved using NASTRAN 105, 106 and 107 solutions.

This approach is so general that each blade can be modelled separately from the others, leading to the possibility of processing an anisotropic rotor: this feature is extremely important when the assessment of manufacturing deviances has to be done, aiming to find acceptable standard criteria.

All the relevant matrices and the right-hand side independent aerodynamic loads and the non-linear aerodynamic and inertial terms are derived from the well-known Lagrangian equation, relative to the generic Lagrangian coordinate q_i

$$\frac{\partial T}{\partial \dot{q}_i} - \frac{\partial T}{\partial q_i} + \frac{\partial U}{\partial q_i} = Q \quad (4.1)$$

The q_i is representing a physical degree of freedom of any grid point lying in the rotor/pylon system, so, even if the rotor/pylon natural modes are computed internally to the code, a direct integration of the physical dofs is performed to solve such a non-linear problem, rather than using the modal superposition, much less suitable because of the uncertainties on which mode to include and their correctness.

4.4.2 Rotor Hub Modelling Features

The rotor hub is FEM described without any restriction in terms of mathematical model complexity, and thus it could range from a synthetic representation (as shown in Fig. 4.14) or a more detailed mesh (see Fig. 4.15).

The unique limitation is related only to the amount of available memory.

The general architecture could cover any layout; in particular, as already mentioned, articulated, hingeless, bearingless or gimbal/teetering hubs can be modelled.

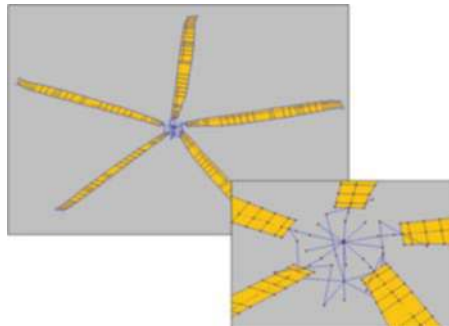


Fig. 4.14 GYROX II rotor model



Fig. 4.15 Shell model

The most important thing to be remembered when the mesh is generated is to give a prescribed identification to the grid points at the boundary with the blade, on one side, and with the rotor mast, on the other side, and, if the hub model includes the rotating part of the controls, it's interfacing grid points with the fixed part; doing that, it sorts out the most convenient breakdown of a rotor/pylon system bulk data, i.e.

1. the blades;
2. the hub and rotating controls;
3. the pylon and fixed controls;

with subsystems (b) and (c) that can be assembled from separated bulk data as well.

Figure 4.16 shows an example of a suitable FEM model of a tail rotor hub and control, which are rotating with the hub and the tail rotor mast, the interface within the hub subsystem with the fixed system being at the tail gearbox.

A damper might be modelled using either linear NASTRAN features (CDAMP elements) or, when non-linear characteristics have to be considered, directly the specific GYROX input. From the point of view of the geometrical layout, the damper might be connecting a blade to the hub or might be located between blades (interblade configuration), as shown in Fig. 4.17a and b.

A particular feature which can be modelled quite easily with FEM approach is the specific torque transfer system from the rotor mast to the hub which can be developed to obtain, on a gimballed hub, an homokinetic joint. Figure 4.18 shows schematically the joint: in a 'normal' joint it is known that ω_z , the angular velocity

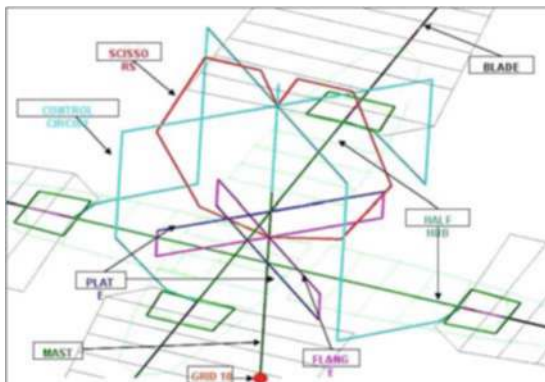


Fig. 4.16 Example of FEM hub and controls model



Fig. 4.17 (a) Blade to hub (b) Blade to blade

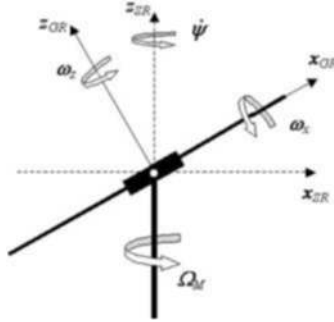


Fig. 4.18 Torque transfer scheme

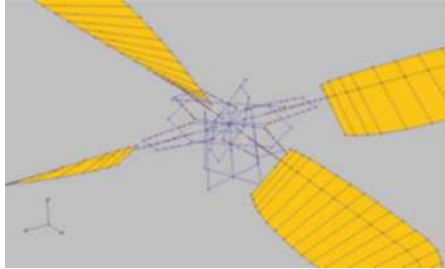


Fig. 4.19 Homokinetic torque transfer model

in the tilted hub plane, is no longer constant, but, if the mean value is equal to the mast angular velocity Ω_M , a 2/rev oscillating angular speed is superimposed to it.

A homokinetic joint is able to leave constant ω_z and this results in a strong effect on the gyroscopic forces acting on the rotor and the centrifugal restoring moments. This technology has been particularly implemented in tiltrotor proprotor hubs: Figure 4.19 shows an example of a tiltrotor homokinetic torque transfer system.

4.4.3 Pylon Modelling Features

As already mentioned, the work pylon may include a schematic condensed representation of what is appended below the rotor or much more complex models, as

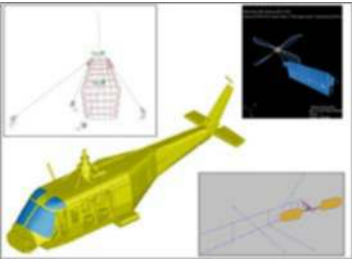


Fig. 4.20 Fixed pylon models

shown in Fig. 4.20, of the main gearbox supporting structure only, of the overall gearbox/airframe or, for a tiltrotor, of the nacelle and half wing or, for a tail rotor, of the tail boom/tail gearbox and tail rotor mast assembly.

The normal degrees of freedom are the physical ones unless dynamically reduced models of the airframe are used (NASTRAN superelements). If the fixed control system (fixed swashplate and actuators) is appended to such a pylon model, provided that the dynamic characteristics of the servo are known, an investigation of any aeroservoelastic problem can be addressed as well.

Finally, it is normal practice to disregard the drive system chain, being not so important in terms of vibratory force generation, and to study its torsional stability separately; it is nevertheless admissible and useful to include a schematic representation of this system too, in order to better identify any collective rotor lag mode.

4.4.4 Rotor Blade Structural Modelling Features

Any blade element (Fig. 4.21) is connecting two grid points, ‘A’ and ‘B’, normally lying on local 25% chord locus, so the overall blade is generally described by a curved line oriented in space; in particular a sweep angle η_A and a droop angle ν_A can be derived, from the grid point coordinates, to identify local beam

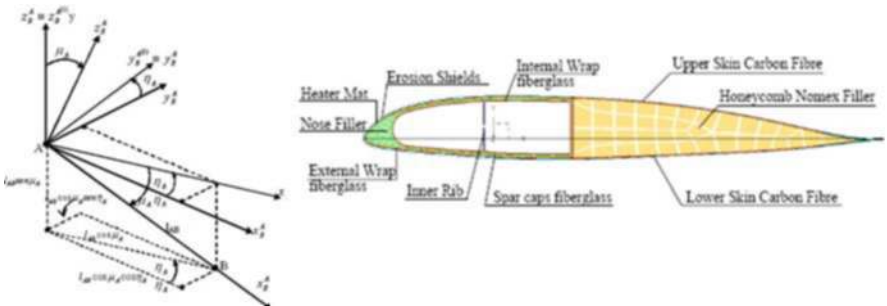


Fig. 4.21 Blade beam element and blade section scheme

reference system axis X_{ea} , the Y_{ea} axis being determined by local built-in twist θ_{twA} and Z_{ea} axis from the orthogonality condition.

The sectional properties may vary linearly along the element, as well as the twist. It is AW normal practice, dealing with composite blades, to derive such properties by means of a 2D FEM code named HANBA; the input to this code is in terms of blade section geometry and mesh of panel elements and stringers, the output being, respectively, axial, beamwise, chordwise and torsional stiffnesses, principal axes rotation, shear centre and neutral axis centre locations, mass and mass moment of inertia, centre of gravity location and inertial principal axes rotation.

Starting from this sectional properties, the elastic stiffness element is computed internally to GYROX by use of integration of the typical cubic functions or by user-defined polynomial shape functions (shear deformation is neglected, at the moment).

4.4.5 Rotor Aerodynamics

Regardless of the approach used to compute rotor inflow, a distribution of aerodynamic airfoil sections has to be given, where any cross section is pointing to the pertinent aerodynamic characteristics in terms of lift, drag and pitching moment coefficients (Fig. 4.22), for a set of geometric incidence from -180° to $+180^\circ$ and appropriate range of Mach numbers; these data normally refer to wind tunnel steady-state 2D characterization (see Fig. 4.23).

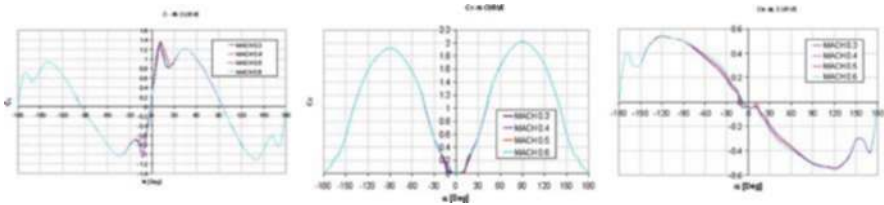


Fig. 4.22 Aerodynamic coefficients

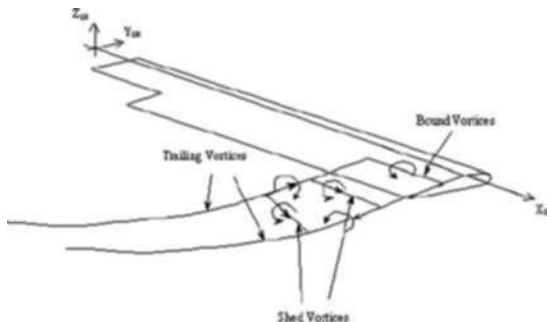


Fig. 4.23 Lifting line scheme

The basic features are modelled by these data:

1. the steady-state 2D stall behaviour, not fully representative of high-speed in-flight behaviour, where dynamic stall features should be more variable, but nevertheless important to build up the unsteady behaviour;
2. compressibility effects, well represented by the sensitivity with Mach numbers;
3. inverse flow state of the retreating blade, again in high-speed condition (the accuracy being not so reliable because of lack of experimental data availability).

The code has been built to include several options in terms of rotor inflow calculation, which can be grouped into two classes:

1. engineering strip theory based;
2. vortex wake based.

The first class of options is what can be easily found in any helicopter's textbook:

- uniform inflow is the most common option, the value of the inflow being determined by the equivalence of the rotor thrust computed by blade element theory and the Glauert momentum theory;
- other two options let the user to superimpose to the basic value above a distribution varying along the span and/or the azimuth: the meaning of these options is more in terms of process checkup, because it is well known that they are satisfactory for performance predictions, they are not at all for vibratory loads calculation.

If these options are used together with 2D sectional airfoil data, 3D effects relative to tip losses, yawed flow and radial drag must be accounted for using specific GYROX inputs as well as Theodorsen correction to deal with unsteady effects.

Second class of options is worldwide known to be the most powerful in terms of vibratory airloads generation, being the basic source of vibratory mechanism based on the intrinsic interaction of the wake-induced velocities with the effective blade incidence.

Again three options are actually included in the code:

- lifting line;
- lifting surface;
- free wake.

In the lifting line approach the blade is represented by a distribution of bound vortices at local 25% of the chord whilst the wake is including both trail vorticity, related to the rate of change of the bound vorticity along the span, and shed vorticity, balancing the rate of change along the azimuth and representing in this way the unsteady aerodynamic effects (Fig. 4.28).

Both trail and shed vorticity are developed in space just using the free stream velocity and the mean uniform inflow value, so that they can be considered as prescribed; the calculation process is iterative and makes use of 2D airfoil data so that high-incidence behaviour affected by steady stall is included, but 3D effects are nearly absent or they are added by means of semi-empirical formulations.

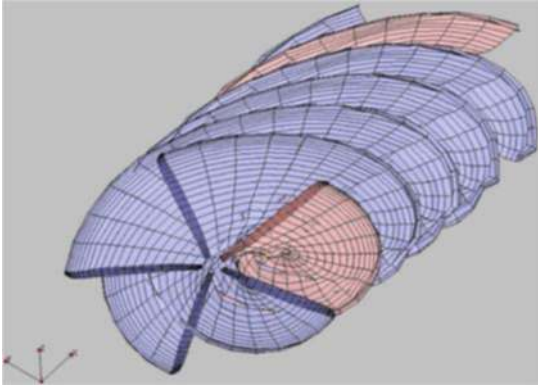


Fig. 4.24 Lifting surface scheme

A better approach to gather 3D effects is founded by the lifting surface methodology, where the blade is modelled by a mesh of panels along the span and the chord, and trail and shed vortices are departing from any bound vortex lying in each panel (Fig. 4.24). Here again the wake development in space and time is prescribed and because of the intrinsic panel modelling feature no stall behaviour can be addressed.

Free wake option is the most powerful approach (Fig. 4.25), where the evolution of the wake is not only based on the free stream velocity but on the mutual induction between vortices as well. Hereafter are listed the basic features of this option:

- wake vortex sheet generated by applying the Kutta condition at the trailing edge of the blades;
- time-stepping free wake vortex model, obtained by integrating the Biot–Savart law along the length of each filament, allowing to have refined roll-up models;

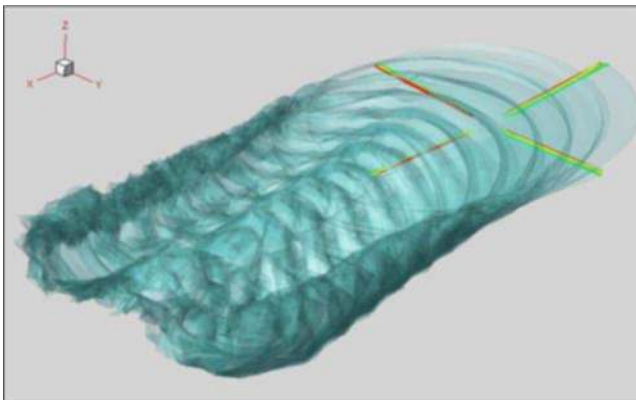


Fig. 4.25 Free wake vortex model

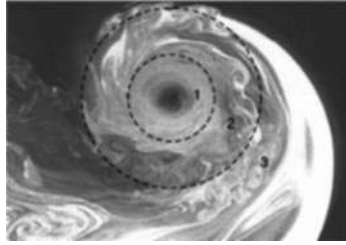


Fig. 4.26 Vortex structure

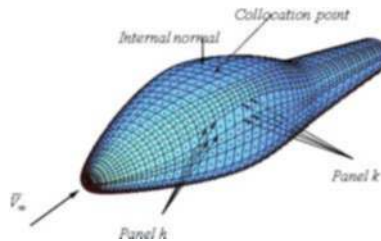


Fig. 4.27 Fuselage panel discretization

- Leishman–Ramasamy and Squire semi-empiric models of vortex growth in the rotor wake to take into account the dissipation of the vortices with the ‘age’ increasing (Fig. 4.26);
- recovery of viscous effects by means of C81 tables, going to look-up tables straightly with the potential lift coefficient value (taken totally into account 3D phenomena near blade root and tip);
- Reynolds number effect correction;
- a specific unsteady panel method in order to take into account the effect of fixed airframe on the rotors and for solving interactional aerodynamics (Fig. 4.27). A specific smart-fast algorithm has been developed for solving large linear systems (based on iterative/multiblock approach): once solving the system, local inviscid velocities are found by means of the potential 3D differentiation along the body surface.
 - 3D ‘constant strength singularity elements’ induced coefficients are evaluated by means of Bess and Smith formulas;
 - far-field simplifications are taken into account where necessary;
 - the doublet panel to vortex ring equivalence is used.

4.4.6 Solution Algorithms

The procedures solved by the code are the following:

1. steady-state forced response;
2. aeroelastic stability;
3. transient analysis.

The steady-state forced response is performed in the frequency domain. The Lagrangian unknowns being the harmonic coefficients of the physical degrees of freedom, are iteratively solved by means, alternatively, of the following algorithms:

1. harmonic balance;
2. Newton–Raphson technique.

As already mentioned, the input data, regarding the trim condition, are the collective and cyclic controls, the mast angle of attack and the free stream wind components. Alternatively, once a wind tunnel trim has to be assessed, a gradient method is used to find the controls, achieving the target static hub reactions.

The aeroelastic stability, because of the periodic coefficients, can only be assessed by means of Floquet technique or similar ones. Ripple method has been preferred here because it gives more physical insight: the physical dofs and the disk sectors where the unstable behaviour is generated are much more easily identified.

In summary

1. the disk is divided in a suitable number of sectors, where it is assumed that the system of equations is linear, so the overall matrices have constant (mean value in the sector) coefficients;
2. writing the compatibility equations in terms of motions components and its first time derivatives, a transition matrix is obtained, relating the initial condition ($az = 0^\circ$) to the final one ($az = 360^\circ$);
3. searching for the eigenvalues of such transition matrix, the system results to be stable if none of them has amplitude greater than unity;
4. sector by sector the overall rotor/pylon system modes are computed as well as its local damping (it might be useful to recall that a negative damping factor in one sector does not necessarily mean that the system is really unstable).

As a follow-up of such analysis, in order to compare the results directly with the nominal time histories of instrumented parameters, a direct integration in the time domain might be requested: Runge–Kutta or Adams–Bashforth algorithms are used as iterative processes on the complete response as a superimposition of the local sector-by-sector degrees of freedom.

4.4.7 Operational Main Features and Output Data

Because of its general usage, the architecture of the code is necessary modular and open: any feature can be included in the procedure, in terms of options used and/or algorithm to be activated.

From the output point of view, these are some of the main topics which can be exploited:

- rotor blade motions;
- overall blade loads, static, vibratory and/or Fourier analysis spectrum;
- detailed blade airloads, as above;
- hub loads, static and vibratory;

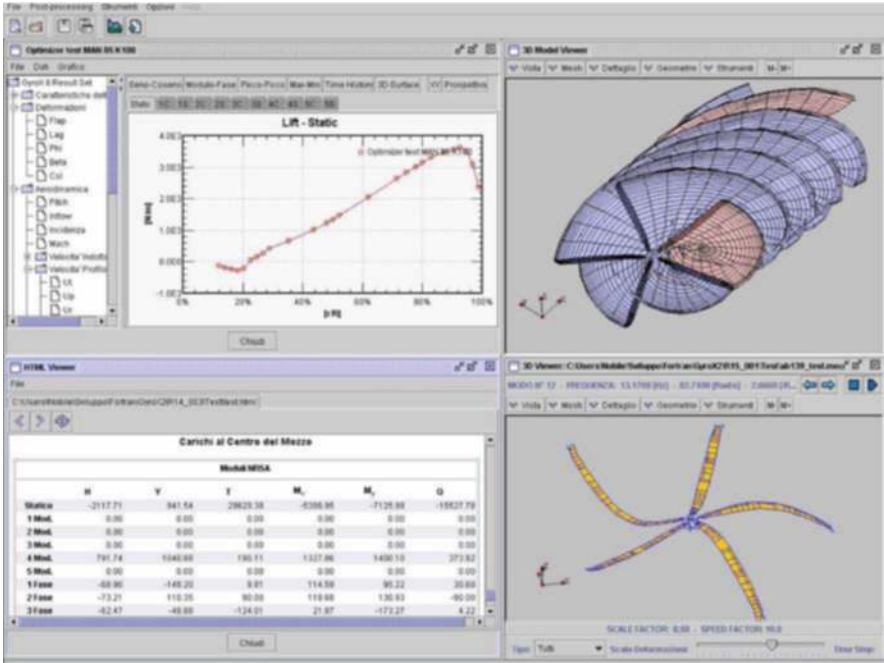


Fig. 4.28 GYROX II Graphical user interface

- damper load and stroke;
- control circuit loads;
- rotor/pylon in vacuo frequencies;
- overall integrated rotor/pylon eigenvalues (Ripple) and stability results.

These data can be sorted out by means of the graphical user interface (shown in Fig. 4.28), which is available to make pre- and post-processing job.

In addition to that, graphical tools are included to produce a fanplot of the rotor/pylon frequencies and/or, in terms of performances, a typical rotor polar (thrust vs. power).

4.5 Applications

The code has been validated through the correlation conducted on EC Research Program Wind Tunnel tests ('Helishape', tests on T7A blade or, more recently, 'Tiltaero', tests on ERICA tiltrotor half wing model) and, more extensively, comparing at medium to high speeds, the in-flight monitored parameters of twin engine light and medium helicopters.

In order to guarantee the most reliable cross-correlation, the following data have been processed:

- 1. trim hub loads;
- 2. blade bending moments (beamwise and chordwise);
- 3. damper stroke and loads;
- 4. pitch link loads;
- 5. vibratory hub loads;
- 6. cabin accelerometers.

Once achieving a satisfactory to good correlation, the code has also been used to predict the expected improvement related to a blade design change.

Figures 4.29, 4.30 and 4.31 show results in terms of trim and vibratory hub loads for a light twin engine helicopter in terms of baseline configuration and a revised one, where two antinodal masses have been added along blade span.

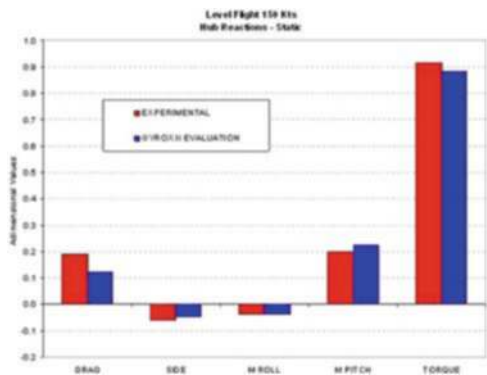


Fig. 4.29 Rotor static loads correlation

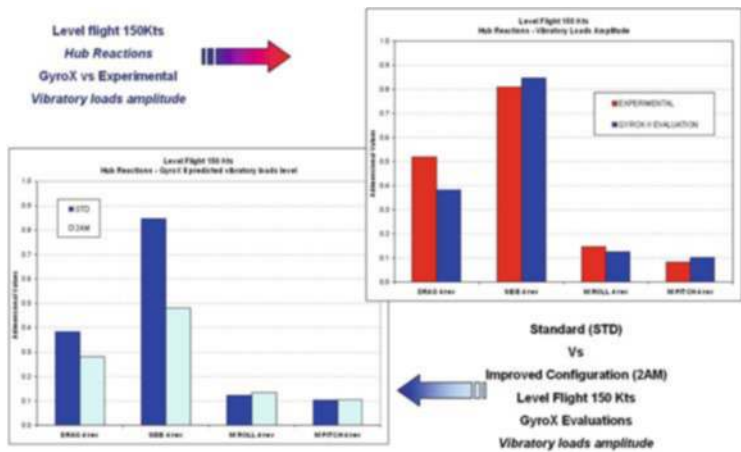


Fig. 4.30 Rotor vibratory loads correlation

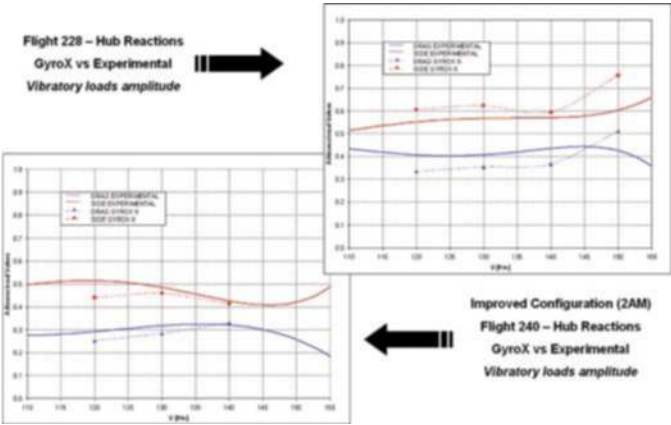


Fig. 4.31 Rotor In-plane hub reaction correlation

Figures 4.32, 4.33, 4.34 and 4.35 are pertinent vice versa to a medium twin-engine helicopter and they are showing, at 150 kts forward speed, the correlation of blade bending moments, damper load and vibratory hub loads.

From the aeroelastic stability point of view, the code has been used to assess the behaviour of a heavy helicopter tail rotor. Figures 4.36 and 4.37 show the correlation of the measured and expected rotor cyclic frequencies vs. rotational speed and the

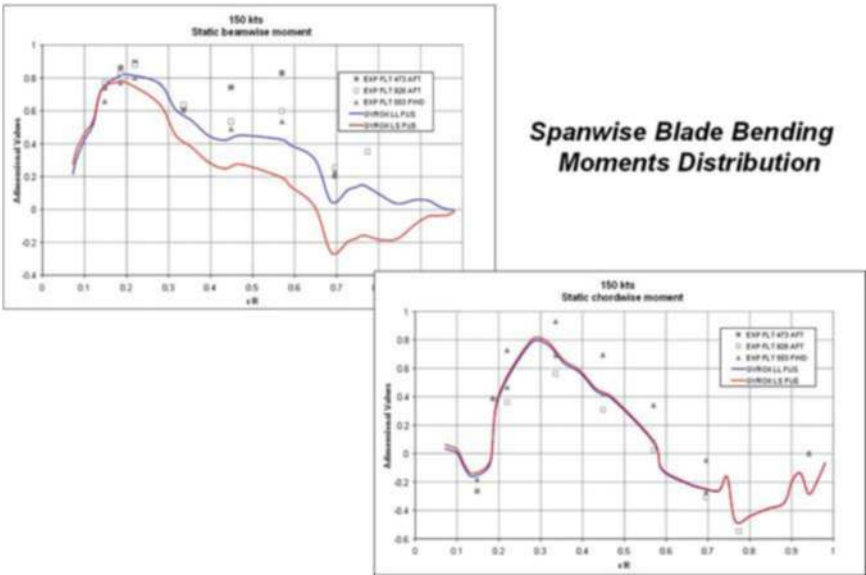


Fig. 4.32 Blade bending moment spanwise distribution correlation

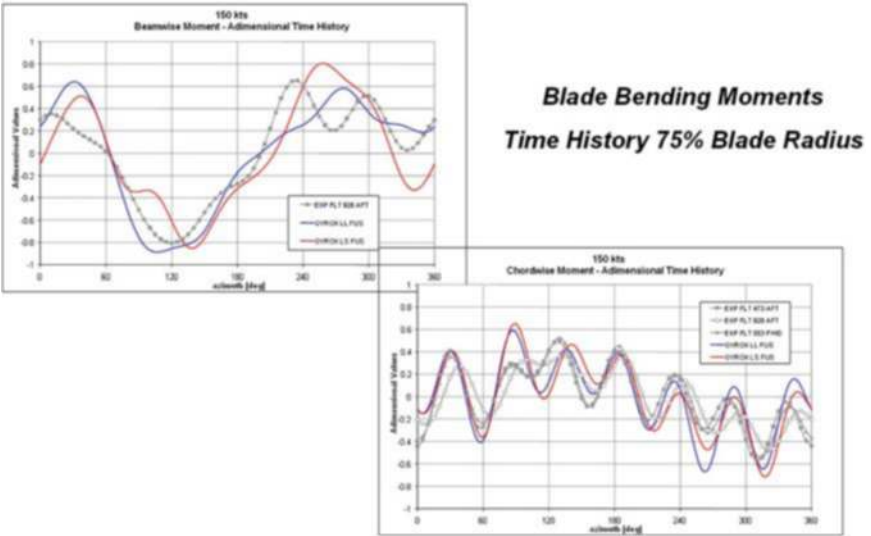


Fig. 4.33 Blade bending moment time history correlation

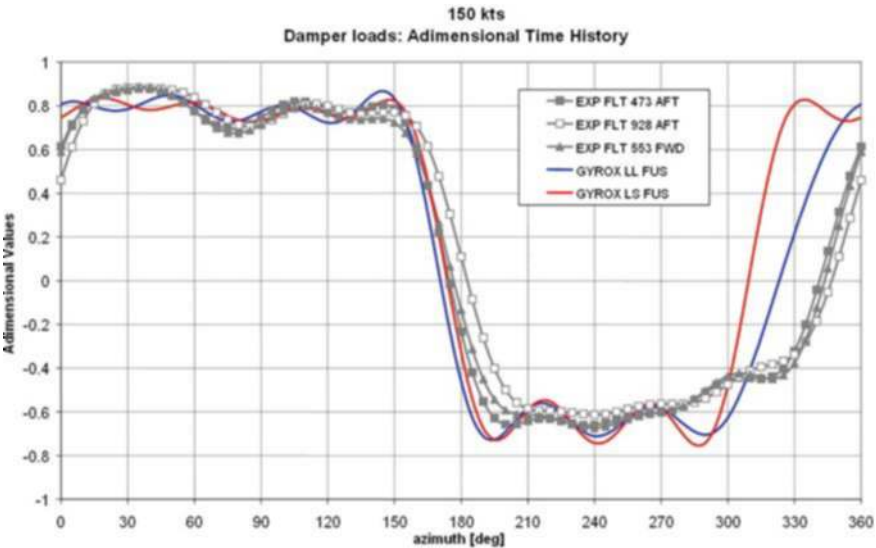


Fig. 4.34 Damper load correlation

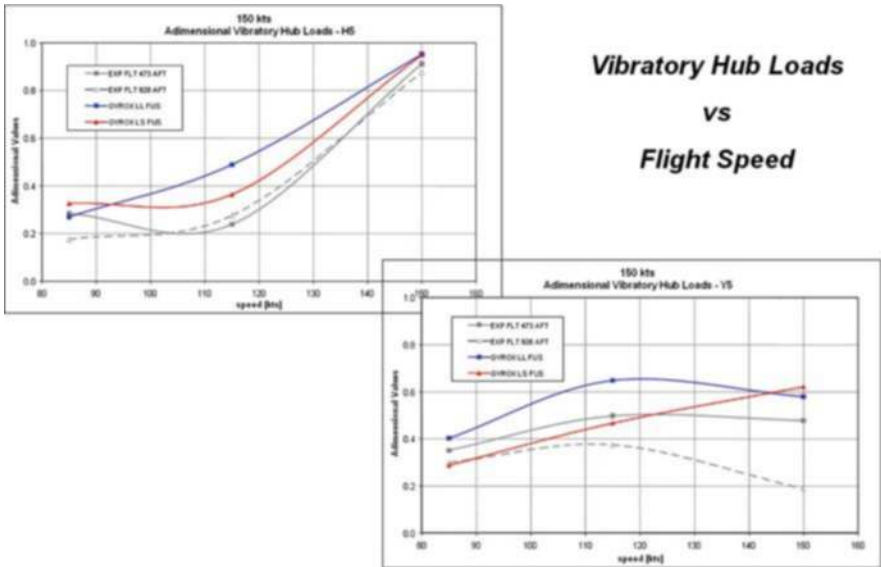


Fig. 4.35 Vibratory hub loads correlation vs. speed

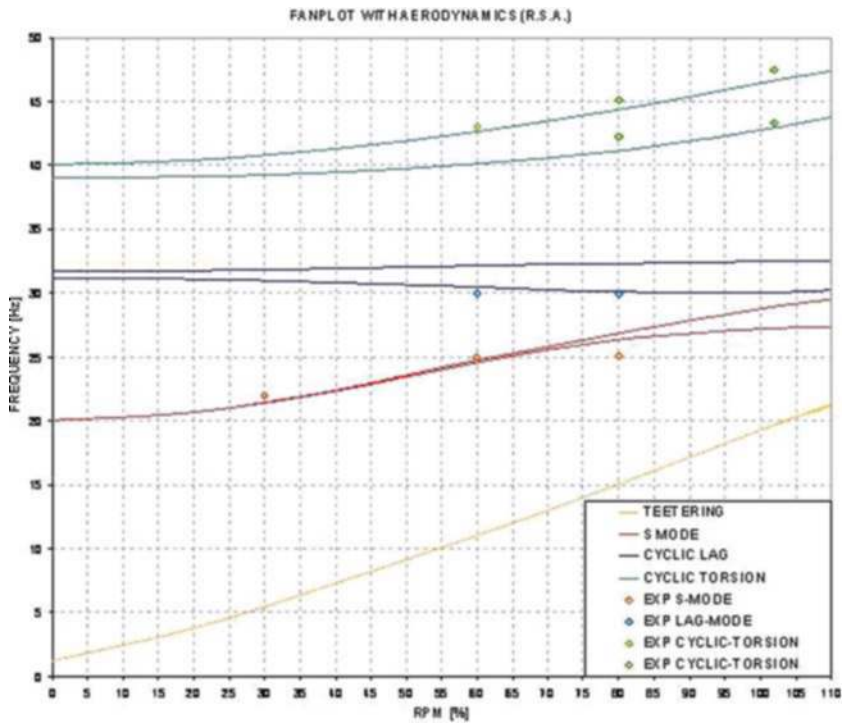


Fig. 4.36 Rotor cyclic frequencies vs. rotational speed

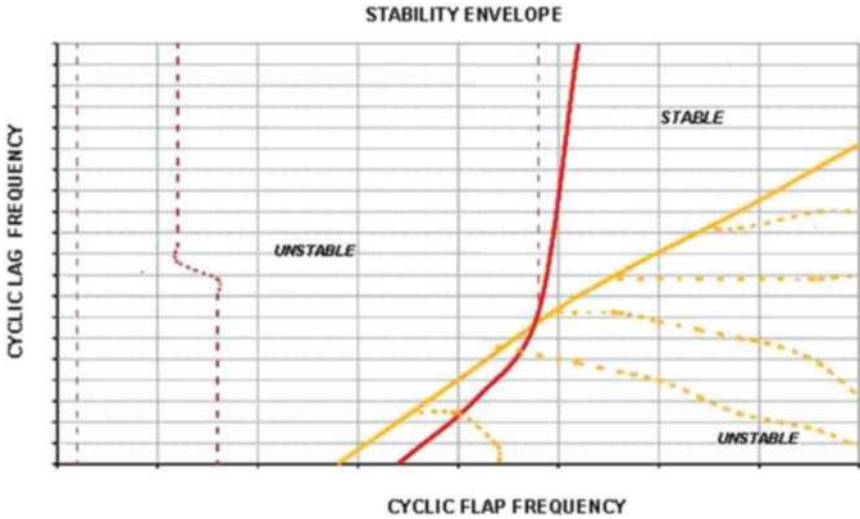


Fig. 4.37 Rotor stability carpet

carpet of the overall stability features, for a prescribed cyclic torsional frequency, as a function of rotor cyclic flap and lag modes.

4.6 Conclusions

The development of GYROX II dedicated tool to face the vibratory issue has demonstrated the validity of the approach. Nevertheless continuous improvement is required both on the modelling side (aerodynamics and structural dynamics) and the computational efficiency one.

4.6.1 Short Term

As far as the aerodynamics is concerned, two major features are presently envisaged to be included:

- an internal software re-shaping to allow easy link with any new or upgraded wake modelling routine;
- the implementation of a new 3D lifting line approach, which should provide the benefit of a lifting surface and should reduce the total running time, improving the use of GYROX as design tool.

In order to consolidate the aeroelastic mechanism, the actual and/or enhanced variant of the code will be used in the near future correlation with experimental dedicated wind tunnel campaign.

4.6.2 Long Term

A new finite element beam model with shape function more closed to the true elastic deformations, and allowing centrifugal hardening progressively increasing from tip to root region, is envisaged to be developed in medium to long term with two objectives:

- further improve the level of predictions;
- reduce the number of degrees of freedom in order to make more attractive the use of the code in an integrated aeromechanics and flight mechanics environment.

Chapter 5

Fields of Extremals and Sufficient Conditions for the Simplest Problem of the Calculus of Variations in n -Variables

Dean A. Carlson and George Leitmann

Abstract In a 1967 note, Leitmann observed that coordinate transformations may be used to deduce extrema (minimizers or maximizers) of integrals in the simplest problem of the calculus of variations. Subsequently, in a series of papers, starting in 2001, he revived this approach and extended it in a variety of ways. Shortly thereafter, Carlson presented an important generalization of this approach and connected it to Carathéodory's equivalent problem method. This in turn was followed by a number of joint papers addressing applications to dynamic games, multiple integrals, and other related topics.

For the simplest vector-valued variables problem of the calculus of variations, making use of the classical notion of fields of extremals, we employ Leitmann's direct method, as extended by Carlson, to present an elementary proof of Weierstrass' sufficiency theorem for strong local and global extrema.

5.1 Introduction

Coordinate transformations have long played a significant role in the analysis of many important problems in control theory as well as other disciplines. For dynamic optimization problems, perhaps the first important contribution in this arena was in the 1830s with Hamilton's introduction of coordinate transformations which was soon put on a firm theoretical foundation by Jacobi. This theory led to the Hamilton–Jacobi equation, an equation that is under extensive investigation to this day. The original goal of a coordinate transformation was to rewrite a system of differential equations, usually arising from a variational principle (i.e., the Euler–Lagrange

Dean A. Carlson

American Mathematical Society, Mathematical Reviews, 416 Fourth Street, Ann Arbor, MI 48103, USA, e-mail: dac@ams.org

George Leitmann

University of California, Berkeley, CA 94720, USA, e-mail: gleit@berkeley.edu

equations) in a form for which the solution was easier to obtain. While this was useful, from an optimization point of view this only assured a solution to the classical necessary condition and did not guarantee a solution of the optimization problem. Using such transformations to obtain sufficient conditions had to wait somewhat longer until in 1967 Leitmann [1] observed that such transformations could also be used to find the extrema (either a maximum or a minimum) for problems in the calculus of variations. This observation lay dormant until Leitmann revived it in [2] and subsequently some extensions to infinite horizon optimization problems, dynamic games, and other systems (see e.g., [3] and [4]). Shortly thereafter in Carlson [5] an important extension which connected Leitmann's method with Carathéodory's notion of equivalent variational problem was presented. Since then a number of joint papers by Carlson and Leitmann have extended these earlier works in a number of directions (see [6–8], and [9]). In these papers a coordinate transformation is used to obtain a new optimization problem in which there is a one-to-one correspondence between the extrema of the original problem and the new problem.

For the simplest free problem in the calculus of variations the classical Weierstrass sufficiency theorem requires the notion of a field of extremals. This is an n -parameter family of solutions to the Euler–Lagrange equations which cover a domain such that for each point of the domain there passes one and only one such extremal. In this chapter we show that if a field of extremals is viewed as a coordinate transformation, then the Leitmann's direct method provides an elementary proof of a version of the Weierstrass sufficiency theorem.

To demonstrate this result we begin in the next section by describing the problem of interest. In Sect. 5.3 we describe Leitmann's direct method. In Sect. 5.4 we recall the definition of a field of extremals and we present our main result in Sect. 5.5. We end with some concluding remarks in Sect. 5.6.

5.2 Notations and the Problem Definition

We consider the simplest free problem in the calculus of variation, namely that of minimizing a Lagrange-type functional

$$J(x(\cdot)) = \int_a^b L(t, x(t), \dot{x}(t)) dt, \quad (5.1)$$

over the class of functions $x(\cdot) : [a, b] \rightarrow \mathbb{R}^n$ which are differentiable with piecewise continuous derivatives (i.e., piecewise smooth) satisfying the fixed end point conditions

$$x(a) = x_a \quad \text{and} \quad x(b) = x_b. \quad (5.2)$$

In the above we assume that $L(\cdot, \cdot, \cdot) : [a, b] \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ has continuous partial derivatives up to order two. We will refer to this problem as problem (P).

Remark 5.1. Clearly, it is possible to restrict the domain of $L(\cdot, \cdot, \cdot)$ to a region $\mathcal{A} = [a, b] \times A_1 \times A_2 \subset [a, b] \times \mathbb{R}^n \times \mathbb{R}^n$, in which $A_1, A_2 \subset \mathbb{R}^n$ are open sets and fix ones attention to that region alone. In this way we could consider local sufficient conditions rather than global ones. Of course, now one restricts attention only to those trajectories $x(\cdot)$ for which $(t, x(t), \dot{x}(t)) \in \mathcal{A}$.

As is standard, in this chapter we consider three types of minimizers – global, strong, and weak – defined as follows.

Definition 5.1. A piecewise smooth trajectory $x^*(\cdot) : [a, b] \rightarrow \mathbb{R}^n$ satisfying the end conditions (5.2) is called

- a. a global minimum if and only if for each piecewise smooth trajectory $x(\cdot) : [a, b] \rightarrow \mathbb{R}^n$ one has

$$J(x^*(\cdot)) = \int_a^b L(t, x^*(t), \dot{x}^*(t)) dt \leq \int_a^b L(t, x(t), \dot{x}(t)) dt = J(x(\cdot)). \quad (5.3)$$

- b. a strong local minimum if and only if there exists an $\varepsilon > 0$ such that for any other piecewise smooth trajectory $x(\cdot) : [a, b] \rightarrow \mathbb{R}^n$ satisfying (5.2) and

$$\sup_{t \in [a, b]} \|x(t) - x^*(t)\| < \varepsilon, \quad (5.4)$$

one has that (5.3) holds.

- c. a weak local minimum if and only if there exists an $\varepsilon > 0$ such that for any other piecewise smooth trajectory $x(\cdot) : [a, b] \rightarrow \mathbb{R}^n$ satisfying (5.2), (5.4), and

$$\sup_{t \in [a, b]} \|\dot{x}(t) - \dot{x}^*(t)\| < \varepsilon, \quad (5.5)$$

one has that (5.3) holds.

Remark 5.2. Clearly it follows that a global minimizer is both a strong and a weak local minimizer and that a strong local minimizer is also a weak local minimizer.

The classical first-order necessary condition for a weak local minimizer (and hence a strong local or a global minimizer) is the Euler–Lagrange equation. That is, if $x^*(\cdot)$ is a weak local minimizer, then one has that it satisfies the nonlinear second-order system of differential equations¹

$$\left. \frac{d}{dt} \left(\frac{\partial L}{\partial p_j} \right) \right|_{(t, x^*(t), \dot{x}^*(t))} = \left. \frac{\partial L}{\partial x_j} \right|_{(t, x^*(t), \dot{x}^*(t))}, \quad j = 1, 2, \dots, n, \quad t \in [a, b]. \quad (5.6)$$

In closing this section we remark that these problems are the forerunners of optimal control problems of the form:

¹ Henceforth, for the sake of brevity, we omit the arguments of functions, here for $L(t, x, \beta)$.

$$\begin{aligned}
& \min \int_a^b f_0(t, x(t), u(t)) dt \\
& \text{subject to} \\
& \quad \dot{x}(t) = f(t, x(t), u(t)), \quad t \in [a, b], \\
& \quad x(a) = x_a \quad \text{and} \quad x(b) = x_b, \\
& \quad u(t) \in U(t, x(t)).
\end{aligned}$$

Remark 5.3. Under appropriate assumptions it is known that this control problem is equivalent to a free problem of the type described above in which the integrand $L(\cdot, \cdot, \cdot)$ is defined as

$$L(t, x, p) \doteq \inf \{ f_0(t, x, u) : p = f(t, x, u), u \in U(t, x) \},$$

where one assumes $L(t, x, p) = +\infty$ if the set $Q(t, x) = \{p : p = f(t, x, u), u \in U(t, x)\} = \emptyset$. Consequently, when such an $L(\cdot, \cdot, \cdot)$ has the requisite smoothness conditions, the results discussed below are applicable. One case where such a scheme can be implemented is when the equation

$$p = f(t, x, u)$$

can be solved for $u = \hat{u}(t, x, p) \in U(t, x)$. In this case $L(t, x, p) = f_0(t, x, \hat{u}(t, x, p))$ and $\hat{u}(t, x, p)$ is an “implicit” feedback control.

5.3 Leitmann’s Direct Method

In this section we present the main tool to be used to obtain our results. In 1967 Leitmann [1] introduced a direct sufficiency method that involved a transformation of coordinates which permitted him to deduce the solution of a calculus of variations problem by inspection. During the 1990s, he revived this method extending it to variational games and applying it to a number of problems arising in economic modeling. In Carlson [5], Leitmann’s method was compared and contrasted with Carathéodory’s notion of equivalent variational problem and finally combined into a new and improved version of Leitmann’s direct method. It is this improved version which we now describe. To do this we let $z(\cdot, \cdot) : [a, b] \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a continuously differentiable function such that the equation $x = z(t, \tilde{x})$ has a unique inverse $\tilde{x} = \tilde{z}(t, x)$ for all $t \in [a, b]$ and such that there is a one-to-one correspondence $x(t) \Leftrightarrow \tilde{x}(t)$ between the set of all piecewise smooth trajectories $x(\cdot)$ satisfying the boundary conditions (5.2) and all piecewise smooth functions $\tilde{x}(\cdot) : [a, b] \rightarrow \mathbb{R}^n$ satisfying

$$\tilde{x}(a) = \tilde{z}(a, x_a) \quad \text{and} \quad \tilde{x}(b) = \tilde{z}(b, x_b). \quad (5.7)$$

Remark 5.4. Observe that if the original integrand is defined on a restricted domain \mathcal{A} and $x(\cdot)$ is a piecewise smooth trajectory satisfying $(t, x(t), \dot{x}(t)) \in \mathcal{A}$ then we have that there exists a trajectory $\tilde{x}(\cdot)$ satisfying (5.7) and such that

$$(t, x(t), \dot{x}(t)) = (t, z(t, \tilde{x}(t)), \frac{\partial z}{\partial t}(t, \tilde{x}(t)) + \left(\frac{\partial z}{\partial \tilde{x}}(t, \tilde{x}(t)) \right)^T \dot{\tilde{x}}(t)) \quad \text{for } t \in [a, b].$$

This means that the trajectories $\tilde{x}(\cdot)$ are such that $(t, \tilde{x}(t), \dot{\tilde{x}}(t)) \in \tilde{\mathcal{A}}$ holds for all $t \in [a, b]$ in which

$$\tilde{\mathcal{A}} = \{(t, \tilde{x}, \tilde{p}) : (t, z(t, \tilde{x}), \frac{\partial z}{\partial t}(t, \tilde{x}) + \left(\frac{\partial z}{\partial \tilde{x}}(t, \tilde{x}) \right)^T \tilde{p}) \in \mathcal{A}\}. \quad (5.8)$$

In the above the partial derivative with respect to \tilde{x} denotes the gradient of $z(t, \cdot)$ and the notation a^T denotes the transpose of the vector $a \in \mathbb{R}^n$. We will adhere to this notation throughout the rest of this chapter.

Now let $\tilde{L}(\cdot, \cdot, \cdot) : [a, b] \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ be another integrand that enjoys the same properties as $L(\cdot, \cdot, \cdot)$ and consider the problem (\tilde{P}) of minimizing the integral functional

$$\tilde{J}(\tilde{x}(\cdot)) = \int_a^b \tilde{L}(t, \tilde{x}(t), \dot{\tilde{x}}(t)) dt \quad (5.9)$$

over all piecewise smooth trajectories $\tilde{x}(\cdot)$ satisfying the end conditions (5.7). With this notation we have the following definition.

Definition 5.2. We say the problems (P) and (\tilde{P}) are equivalent if and only if $x^*(\cdot)$ is a minimizer of (P) whenever $\tilde{x}^*(\cdot) = \tilde{z}(\cdot, x^*(\cdot))$ is a minimizer of (\tilde{P}) .

We now give Leitmann's direct sufficiency method.

Theorem 5.1. Let $z(\cdot, \cdot)$, $L(\cdot, \cdot, \cdot)$, and $\tilde{L}(\cdot, \cdot, \cdot)$ satisfy the general hypotheses described above. If there exists a C^1 function $G(\cdot, \cdot) : [a, b] \times \mathbb{R}^n \rightarrow \mathbb{R}$ such that the functional identity

$$L(t, x(t), \dot{x}(t)) - \tilde{L}(t, \tilde{x}(t), \dot{\tilde{x}}(t)) = \frac{d}{dt} G(t, \tilde{x}(t)) \quad (5.10)$$

holds for all $t \in [a, b]$ and all piecewise smooth trajectories $x(\cdot)$ satisfying the end conditions (5.2) with $\tilde{x}(t) = \tilde{z}(t, x(t))$, then the problems (P) and (\tilde{P}) are equivalent.

Proof. Leitmann [10]. □

Two immediate and useful corollaries are the following.

Corollary 5.1. The existence of $G(\cdot, \cdot)$ in (5.10) implies that the following identity holds for $(t, \tilde{x}, \tilde{p}) \in (a, b) \times \mathbb{R}^n \times \mathbb{R}^n$:

$$\begin{aligned} L(t, z(t, \tilde{x}), \frac{\partial z(t, \tilde{x})}{\partial t} + \left(\frac{\partial z(t, \tilde{x})}{\partial \tilde{x}} \right)^T \tilde{p}) - \tilde{L}(t, \tilde{x}, \tilde{p}) \\ \equiv \frac{\partial G(t, \tilde{x})}{\partial t} + \left(\frac{\partial G(t, \tilde{x})}{\partial \tilde{x}} \right)^T \tilde{p}. \end{aligned} \quad (5.11)$$

Corollary 5.2. The left-hand side of the identity (5.11) is linear in \tilde{p} , that is, it is of the form

$$\Theta(t, \tilde{x}) + \Psi(t, \tilde{x})^T \tilde{p}$$

and

$$\frac{\partial G(t, \tilde{x})}{\partial t} = \Theta(t, \tilde{x}) \quad \text{and} \quad \frac{\partial G(t, \tilde{x})}{\partial \tilde{x}} = \Psi(t, \tilde{x})$$

on $[a, b] \times \mathbb{R}$.

Remark 5.5. The utility of the above theorem rests on being able to choose not only the transformation $z(\cdot, \cdot)$ but also the integrand $\tilde{L}(\cdot, \cdot, \cdot)$ and the function $G(\cdot, \cdot)$. We further notice that the above results do not require the smoothness hypotheses that we have imposed here on the integrands $L(\cdot, \cdot, \cdot)$ and $\tilde{L}(\cdot, \cdot, \cdot)$.

Remark 5.6. The above theorem is an extension of Leitmann's original idea and includes it as a special case by taking $\tilde{L}(\cdot, \cdot, \cdot) = L(\cdot, \cdot, \cdot)$; see [2]. In addition, it includes the notion of equivalent variational problem found in Carathéodory's work (e.g., see Carathéodory's book [11]) by taking $z(\cdot, \cdot)$ to be the identity transformation (i.e., $z(t, \tilde{x}) \equiv \tilde{x}$). We further remark that these ideas are also closely related to the canonical transformations found in classical mechanics originally investigated by Hamilton and Jacobi in the 1830s.

5.4 Fields of Extremals

To apply the direct sufficiency method given in Theorem 5.1 requires three things. First one needs a transformation $z(\cdot, \cdot)$ satisfying the requisite hypotheses, second one requires the function $G(\cdot, \cdot)$, and finally one needs the new integrand $\tilde{L}(\cdot, \cdot, \cdot)$ to define the new variational problem (\tilde{P}). The last quantity, namely $\tilde{L}(\cdot, \cdot, \cdot)$, will be introduced in the next section. In this section we will see that the notion of a classical field of extremals allows us to determine an appropriate transformation $z(\cdot, \cdot)$ and to insure the existence of the function $G(\cdot, \cdot)$. We begin our discussion by developing the classical notion of a field of extremals. Our presentation follows that found in the book by N. I. Akhiezer [12] and begins with the following definition.

Definition 5.3. We say a region $D \subset [a, b] \times \mathbb{R}^n$ is a field for the functional (5.1) with slope function $\hat{p}(\cdot, \cdot) : D \rightarrow \mathbb{R}^n$ if and only if it satisfies the following two conditions:

1. The components of the vector-valued function $\hat{p}(\cdot, \cdot)$ have continuous first-order partial derivatives in D .
2. The "Hilbert invariant integral"

$$I_C = \int_C \left[L(t, x, \hat{p}(t, x)) - \left(\frac{\partial L}{\partial p} \bigg|_{(t, x, \hat{p}(t, x))} \right)^T \hat{p}(t, x) \right] dt \\ + \left(\frac{\partial L}{\partial p} \bigg|_{(t, x, \hat{p}(t, x))} \right)^T dx \quad (5.12)$$

depends only on the end points of the curve along which it is taken (i.e., it is path independent) for any curve C lying completely in D .

Associated with this field we define the **trajectories of the field** as the solutions of the system of first-order differential equations

$$\dot{x}(t) = \hat{p}(t, x(t)).$$

As a consequence of the existence theory for ordinary differential equation, through each point $(t, x) \in D$ there passes one and only one solution of this differential equation and therefore the collection of all such trajectories is an n -parameter family of trajectories which we denote by $x(\cdot) = \varphi(\cdot, \beta)$ in which $\varphi(\cdot, \cdot)$ is a twice continuously differentiable function defined on a region R in the space of variables $(t, \beta) \subset R \subset [a, b] \times \mathbb{R}^n$. If one follows the presentation in [12], as a consequence of the path independence of the integral (5.12) it follows easily that $t \rightarrow \varphi(t, \beta)$ satisfies the Euler–Lagrange equations (5.6) (see [1, Chapter 2]). That is, the functions $x(\cdot) = \varphi(\cdot, \beta) = \varphi(\cdot, \beta_1, \dots, \beta_n)$ form an n -parameter family of solutions of Euler–Lagrange equations (5.6) which is frequently called a **field of extremals**. A natural question to ask now is under what conditions does an n -parameter family of solutions, say $x(\cdot) = \varphi(\cdot, \beta)$, of (5.6) correspond to a field D for the functional (5.1) and what is the corresponding slope function $\hat{p}(\cdot, \cdot)$. We now direct our attention in this direction.

To begin we consider an n -parameter family of functions, $x(\cdot)$, given by

$$x(t) = \varphi(t, \beta) = \varphi(t, \beta_1, \beta_2, \dots, \beta_n).$$

For each choice of a parameter β the map $t \rightarrow \varphi(t, \beta)$ defines a curve in \mathbb{R}^n . We further suppose that this family of curves “simply covers” a region $D \subset [a, b] \times \mathbb{R}^n$ in the sense that through every point $(t, x) \in D$ exactly one and only one member of the family passes through this point (i.e., there exists only one choice of β such that $x = \varphi(t, \beta)$). In this way, this family of curves defines a point set R in the space of variables $(t, \beta) \in \mathbb{R} \times \mathbb{R}^n$, which we assume is simply connected. Further we assume that the functions $\varphi(\cdot, \cdot)$ are twice continuously differentiable in R , are such that the Jacobian

$$\det \left(\frac{\partial \varphi}{\partial \beta} \right) \neq 0, \quad (5.13)$$

everywhere in R , and are such that for each β the vector-valued function $t \rightarrow \varphi(t, \beta)$ satisfies the Euler–Lagrange equations (5.6). That is,

$$\left. \frac{d}{dt} \left(\frac{\partial L}{\partial p_j} \right) \right|_{(t, \varphi(t, \beta), \dot{\varphi}(t, \beta))} = \left. \frac{\partial L}{\partial x_j} \right|_{(t, \varphi(t, \beta), \dot{\varphi}(t, \beta))}, \quad j = 1, 2, \dots, n, \quad t \in [a, b], \quad (5.14)$$

in which $\dot{\varphi}(t, \beta)$ denotes the partial derivative of $\varphi(t, \beta)$ with respect to the t variable. A consequence of (5.13) is that we can uniquely solve the equation $x = \varphi(t, \beta)$ for β to obtain a twice continuously differentiable function which we denote by $\beta = \psi(t, x)$, for $(t, x) \in D$. Now, through each given point $(t, x) \in D$

there exists one and only one curve of the family which we denote by the function $s \rightarrow y(s) = \varphi(s, \psi(t, x))$, $a \leq s \leq b$ (i.e., we select the unique function from the family $\varphi(\cdot, \beta)$ corresponding to the parameter $\beta = \psi(t, x)$). The slope of the curve at any point $(s, y(s))$ on the curve is given by the formula

$$\dot{y}(s) = \dot{\varphi}(s, \beta) = \dot{\varphi}(s, \psi(t, x)).$$

In particular we notice that for $s = t$ this becomes

$$\dot{y}(t) = \dot{\varphi}(t, \psi(t, x)).$$

As $(t, x) \in D$ was chosen arbitrarily the mapping $(t, x) \rightarrow \dot{\varphi}(t, \psi(t, x))$ is well defined on D . With this observation we select as a candidate for the “slope function” $\hat{p}(\cdot, \cdot) : D \rightarrow \mathbb{R}^n$ defined by the formula

$$\hat{p}(t, x) = \dot{\varphi}(t, \psi(t, x)). \quad (5.15)$$

This function has continuous partial derivatives of the first order in D .

Our goal now is to find conditions which allow the region D to be a field with slope function $\hat{p}(\cdot, \cdot)$. Observe that, as a consequence of our construction, if D is a field with slope function $\hat{p}(\cdot, \cdot)$ the n -parameter family we started with will be the trajectories of the field. To find these additional conditions we rewrite the Hilbert invariant integral in terms of the variables (t, β) as

$$I_C = \int_{\mathcal{C}} L(t, \varphi(t, \beta), \dot{\varphi}(t, \beta)) dt + \left(\frac{\partial L}{\partial p} \Big|_{(t, \varphi(t, \beta), \dot{\varphi}(t, \beta))} \right)^T \frac{\partial \varphi}{\partial \beta} \Big|_{(t, \beta)} d\beta, \quad (5.16)$$

in which \mathcal{C} is the path in R corresponding to the path C in D , $\partial \varphi / \partial \beta$ denotes the Jacobian matrix of $\varphi(t, \cdot)$ (considered as a function of β alone for each fixed t), and $\dot{\varphi}(t, \beta)$ denotes the partial derivative of $\varphi(\cdot, \cdot)$ with respect to t . Now for this integral to be path independent it is necessary (and sufficient since R is simply connected) that the integrand be an exact differential. That is, there exists a function $G(\cdot, \cdot)$ such that

$$\frac{\partial G}{\partial t}(t, \beta) = L(t, \varphi(t, \beta), \dot{\varphi}(t, \beta)) \quad (5.17)$$

$$\frac{\partial G}{\partial \beta_j}(t, \beta) = \sum_{i=1}^n \frac{\partial \varphi_i}{\partial \beta_j}(t, \beta) \frac{\partial L}{\partial p_i}(t, \varphi(t, \beta), \dot{\varphi}(t, \beta)), \quad j = 1, 2, \dots, n.$$

The conditions for such a $G(\cdot, \cdot)$ to exist is that all of the second-order mixed partial derivatives of $G(\cdot, \cdot)$ be equal. This means that we first have the conditions

$$\frac{\partial}{\partial \beta_j} L(t, \varphi(t, \beta), \dot{\varphi}(t, \beta)) = \frac{\partial}{\partial t} \left(\sum_{i=1}^n \frac{\partial \varphi_i}{\partial \beta_j}(t, \beta) \frac{\partial L}{\partial p_i}(t, \varphi(t, \beta), \dot{\varphi}(t, \beta)) \right), \quad (5.18)$$

for $j = 1, 2, \dots, n$ and the conditions

$$\begin{aligned} \frac{\partial}{\partial \beta_r} \left(\sum_{i=1}^n \frac{\partial \varphi_i}{\partial \beta_s}(t, \beta) \right) \frac{\partial L}{\partial p_i}(t, \varphi(t, \beta), \dot{\varphi}(t, \beta)) \\ = \frac{\partial}{\partial \beta_s} \left(\sum_{i=1}^n \frac{\partial \varphi_i}{\partial \beta_r}(t, \beta) \frac{\partial L}{\partial p_i}(t, \varphi(t, \beta), \dot{\varphi}(t, \beta)) \right), \end{aligned} \quad (5.19)$$

for $r, s = 1, 2, \dots, n$, $r \neq s$. Carrying out the differentiation in (5.18) we have,

$$\sum_{i=1}^n \frac{\partial L}{\partial x_i} \frac{\partial \varphi}{\partial \beta_j} + \sum_{i=1}^n \frac{\partial L}{\partial p_i} \frac{\partial \dot{\varphi}}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial L}{\partial p_i} \frac{\partial^2 \varphi_i}{\partial t \partial \beta_j} + \sum_{i=1}^n \frac{\partial^2 L}{\partial t \partial p_i} \frac{\partial \varphi_i}{\partial \beta_j},$$

for $j = 1, 2, \dots, n$, where again we have suppressed the arguments of the functions for the sake of brevity. Using the fact that $\varphi(\cdot, \cdot)$ is twice continuously differentiable we have that

$$\frac{\partial \dot{\varphi}}{\partial \beta_j} = \frac{\partial^2 \varphi}{\partial \beta_j \partial t} = \frac{\partial^2 \varphi_i}{\partial t \partial \beta_j},$$

for $i, j = 1, 2, \dots, n$ so that the above expansion reduces to

$$\sum_{i=1}^n \left\{ \frac{\partial}{\partial x} \left(\frac{\partial L}{\partial p_i} \right) - \frac{\partial L}{\partial x_i} \right\} \frac{\partial \varphi_i}{\partial \beta_j} = 0, \quad j = 1, 2, \dots, n.$$

This last set of inequalities is automatically satisfied since we have assumed that the n -parameter family $\varphi(\cdot, \cdot)$ satisfies the Euler–Lagrange equations (5.6). Thus, we immediately have that (5.18) is satisfied. Expanding (5.19) we have

$$\begin{aligned} \sum_{i=1}^n \left\{ \frac{\partial^2 \varphi_i}{\partial \beta_r \partial \beta_s} \left(\frac{\partial L}{\partial p_i} \right) + \frac{\partial \varphi_i}{\partial \beta_s} \frac{\partial}{\partial \beta_r} \left(\frac{\partial L}{\partial p_i} \right) \right\} \\ = \sum_{i=1}^n \left\{ \frac{\partial^2 \varphi_i}{\partial \beta_s \partial \beta_r} \left(\frac{\partial L}{\partial p_i} \right) + \frac{\partial \varphi_i}{\partial \beta_r} \frac{\partial}{\partial \beta_s} \left(\frac{\partial L}{\partial p_i} \right) \right\}, \end{aligned}$$

for $r, s = 1, 2, \dots, n$, $r \neq s$. Again, since $\varphi(\cdot, \cdot)$ is twice continuously differentiable this last expression reduces to

$$\sum_{i=1}^n \left\{ \frac{\partial \varphi_i}{\partial \beta_s} \frac{\partial}{\partial \beta_r} \left(\frac{\partial L}{\partial p_i} \right) - \frac{\partial \varphi_i}{\partial \beta_r} \frac{\partial}{\partial \beta_s} \left(\frac{\partial L}{\partial p_i} \right) \right\} = 0, \quad (5.20)$$

for $r, s = 1, 2, \dots, n$ (observe that the summand corresponding to $r = s$ is automatically zero). The right-hand side of (5.20) is referred to as a Lagrange bracket. Thus, to summarize what we have done we state the following.

Theorem 5.2. *For an n -parameter family of functions $\varphi(\cdot, \cdot)$ satisfying the conditions outlined at the beginning of this section to be the trajectories of a field D for the functional (5.1) with slope function $\hat{p}(\cdot, \cdot)$ given by (5.15) it is necessary and*

sufficient that all of the Lagrange brackets (i.e., the right-hand side of (5.20) for all $r, s = 1, 2, \dots, n$) vanish identically on R .

Remark 5.7. In the case when $n = 1$, the conditions in (5.19) are vacuous. Therefore, any region $D \subset [a, b] \times \mathbb{R}$ that is simply covered by a 1-parameter family of solutions of the Euler–Lagrange equation (5.6) is a field for the functional (5.1).

We conclude our discussion of fields by making some connections to Leitmann’s direct sufficiency method. Let $D \subset [a, b] \times \mathbb{R}^n$ be a field for the objective functional (5.1) and let $\varphi(\cdot, \cdot)$ denote the n -parameter family of trajectories for the field. For each piecewise smooth trajectory $x(\cdot) : [a, b] \rightarrow \mathbb{R}^n$ whose graph lies in D (i.e., $\{(t, x(t)) : t \in [a, b]\} \subset D$) we can define the unique piecewise smooth function $\tilde{x}(\cdot) : [a, b] \rightarrow \mathbb{R}^n$ by means of the equation

$$x(t) = \varphi(t, \tilde{x}(t)), \quad t \in [a, b], \quad (5.21)$$

or equivalently by the inverse relation

$$\tilde{x}(t) = \psi(t, x(t)), \quad t \in [a, b]. \quad (5.22)$$

Moreover, we observe that the graph of $\tilde{x}(\cdot)$ is contained in R (i.e., $\{(t, \tilde{x}(t)) : t \in [a, b]\} \subset R$). Conversely, if $\tilde{x}(\cdot)$ is a piecewise smooth trajectory whose graph lies in R , then one can uniquely define a trajectory $x(\cdot) : [a, b] \rightarrow \mathbb{R}^n$ whose graph lies in D by means of the equation (5.21). In this way, we see that the n -parameter family of trajectories can be used to establish a one-to-one correspondence between the piecewise smooth trajectories with graphs in D and the piecewise smooth trajectories with graphs in R . Further, if we restrict $x(\cdot)$ to satisfy the fixed end conditions (5.2) then $\tilde{x}(\cdot) = \psi(\cdot, x(\cdot))$ satisfies the end conditions

$$\tilde{x}(a) = \tilde{x}_a = \psi(a, x_a) \quad \text{and} \quad \tilde{x}(b) = \tilde{x}_b = \psi(b, x_b). \quad (5.23)$$

From the above we see that we have a transformation of coordinates which we can exploit in Leitmann’s direct sufficiency method. More specifically, we can define $z(t, \tilde{x}) = \varphi(t, \tilde{x})$ for each $(t, \tilde{x}) \in R$. Furthermore, the inverse transformation $\tilde{z}(\cdot, \cdot) : D \rightarrow \mathbb{R}^n$ is defined by $\tilde{z}(t, x) = \psi(t, x)$ for each $(t, x) \in D$. In addition, we also have that because the Hilbert invariant integral (5.12) is path independent, there exists a function $G(\cdot, \cdot)$ for which (5.17) holds on R . Thus we see that two of the three components required to apply the direct sufficiency method can be obtained by the classical field theory.

5.5 Sufficient Conditions for Optimality

We now return to our original problem beginning with the following definition.

Definition 5.4. Let D be a field for the functional (5.1) with slope function $\hat{p}(\cdot)$ and let $x^*(\cdot) : [a, b] \times \mathbb{R}^n$ be a twice continuously differentiable solution of the

Euler–Lagrange equations (5.6) satisfying the fixed end conditions (5.2). We say $x^*(\cdot)$ is embedded in the field D for the functional (5.1) if and only if it corresponds to one of the trajectories of the field, that is, if there exists a (unique) parameter β^* such that $x^*(\cdot) = \varphi(\cdot, \beta^*)$.

Observe that if $x^*(\cdot)$ is embedded in a field D for the functional (5.1) then there exists a parameter β^* such that $x^*(t) = \varphi(t, \beta^*)$ for all $t \in [a, b]$. This means that we have $\beta^* = \psi(t, x^*(t))$ for all $t \in [a, b]$ and in particular we have $\beta^* = \psi(a, x_a) = \psi(b, x_b)$. Thus for every piecewise smooth trajectory $x(\cdot)$ that satisfies the fixed end conditions (5.2), its related trajectory $\tilde{x}(\cdot) = \psi(t, x(\cdot))$ satisfies the periodic end conditions

$$\tilde{x}(a) = \beta^* \quad \text{and} \quad \tilde{x}(b) = \beta^*.$$

Moreover, the related trajectory $\tilde{x}^*(\cdot)$ given by $\tilde{x}^*(t) = \psi(t, x^*(t)) \equiv \beta^*$ is a constant function.

At present we only have two-thirds of what is required to apply Leitmann's direct sufficiency method, namely the transformation $z(\cdot, \cdot)$ and the function $G(\cdot, \cdot)$. We still need to define the optimization problem (\tilde{P}). This requires us to select an integrand $\tilde{L}(\cdot, \cdot, \cdot) : [a, b] \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$. To do this we suppose, as above, that $x^*(\cdot)$ is embedded in a field D for the functional (5.1) with slope function $\hat{p}(\cdot, \cdot) = \phi(\cdot, \psi(\cdot, \cdot))$ and define the Weierstrass E -function, $E : [a, b] \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ by the formula

$$E(t, x, p, q) = L(t, x, q) - L(t, x, p) - \left(\frac{\partial L}{\partial p} \Big|_{(t, x, p)} \right)^T (q - p). \quad (5.24)$$

Observe that if we consider the function $p \rightarrow L(t, x, z)$, with (t, x) fixed the Weierstrass E -function represents the difference between $L(t, x, \cdot)$ evaluated at q and the linear part of its Taylor expansion about the point p . This means that for each fixed (t, x) there exists $\theta(t, x) \in (0, 1)$ such that

$$E(t, x, p, q) = \frac{1}{2}(q - p)^T \frac{\partial^2 L}{\partial p^2} \Big|_{(t, x, p + \theta(t, x)(q - p))} (q - p),$$

in which $\partial^2 L / \partial p^2$ denotes the Hessian matrix of the function $L(t, x, \cdot)$. In particular, when we have a field D for the functional (5.1) with slope $\hat{p}(\cdot, \cdot)$, we have the formula,

$$\begin{aligned} E(t, x, \hat{p}(t, x), q) &= \frac{1}{2}(q - \hat{p}(t, x))^T Q(t, x, q)(q - \hat{p}(t, x)) \\ &= \frac{1}{2}(q - \phi(t, \psi(t, x)))^T Q(t, x, q)(q - \phi(t, \psi(t, x))), \end{aligned}$$

where

$$Q(t, x, q) = \frac{\partial^2 L}{\partial p^2} \Big|_{(t, x, \hat{p}(t, x) + \theta(t, x)(q - \hat{p}(t, x)))} = \frac{\partial^2 L}{\partial p^2} \Big|_{(t, x, \phi(t, \psi(t, x)) + \theta(t, x)(q - \phi(t, \psi(t, x))))}.$$

With this notation we define our integrand $\tilde{L}(\cdot, \cdot, \cdot) : [a, b] \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ by the formula

$$\begin{aligned}\tilde{L}(t, \tilde{x}, q) &= E(t, \varphi(t, \tilde{x}), \dot{\varphi}(t, \tilde{x}), \dot{\varphi}(t, \tilde{x}) + \frac{\partial \varphi}{\partial \beta}(t, \tilde{x})q) \\ &= \frac{1}{2} q^T \left(\frac{\partial \varphi}{\partial \beta} \right)^T \bigg|_{(t, \tilde{x})} Q(t, \varphi(t, \tilde{x}), q) \frac{\partial \varphi}{\partial \beta} \bigg|_{(t, \tilde{x})} q,\end{aligned}\quad (5.25)$$

for $(t, \tilde{x}, q) \in [a, b] \times \mathbb{R}^n \times \mathbb{R}^n$.

Thus the optimization problem (\tilde{P}) is defined as minimizing the integral functional

$$\tilde{J}(\tilde{x}(\cdot)) = \int_a^b \tilde{L}(t, \tilde{x}(t), \dot{\tilde{x}}(t)) dt$$

over all piecewise smooth trajectories $\tilde{x}(\cdot) : [a, b] \rightarrow \mathbb{R}^n$ satisfying the fixed end conditions

$$\tilde{x}(a) = \beta^* \quad \text{and} \quad \tilde{x}(b) = \beta^*.$$

We now show that when we restrict admissible trajectories for (P) and (\tilde{P}) to have their graphs in D and R , respectively, then these two problems are equivalent.

Lemma 5.1. *Let $x^*(\cdot) : [a, b] \rightarrow \mathbb{R}^n$ be a twice continuously differentiable function that satisfies the Euler–Lagrange equation (5.6) and the fixed end conditions (5.2). Further suppose that $x^*(\cdot)$ is embedded in a field D for the functional (5.1). Then the problem (P) of minimizing (5.1) over all piecewise smooth trajectories $x(\cdot)$ satisfying the end conditions (5.2) and whose graphs lie in D is equivalent to the problem (\tilde{P}) of minimizing the functional (5.9), in which $\tilde{L}(\cdot, \cdot, \cdot)$ is as defined in (5.25), overall piecewise smooth trajectories $\tilde{x}(\cdot)$ satisfying the periodic boundary conditions $\tilde{x}(a) = \tilde{x}(b) = \beta^*$ and such that their graphs lie in R .*

Proof. As $x^*(\cdot)$ can be embedded in a field D , for each $(t, \beta) \in R$ and $q \in \mathbb{R}^n$ we can apply Taylor's theorem (in the last argument of $L(\cdot, \cdot, \cdot)$) with remainder to get

$$\begin{aligned}L(t, \varphi, \dot{\varphi} + \frac{\partial \varphi}{\partial \beta} q) &= L(t, \varphi, \dot{\varphi}) + \frac{\partial L}{\partial p} \bigg|_{(t, \varphi, \dot{\varphi})} \frac{\partial \varphi}{\partial \beta} q \\ &\quad + \frac{1}{2} q^T \left(\frac{\partial \varphi}{\partial \beta} \right)^T \frac{\partial^2 L}{\partial p^2} \bigg|_{(t, \varphi, \dot{\varphi} + \theta \frac{\partial \varphi}{\partial \beta} q)} \frac{\partial \varphi}{\partial \beta} q \\ &= L(t, \varphi, \dot{\varphi}) + \frac{\partial L}{\partial p} \bigg|_{(t, \varphi, \dot{\varphi})} \frac{\partial \varphi}{\partial \beta} q + \frac{1}{2} q^T \left(\frac{\partial \varphi}{\partial \beta} \right)^T Q(t, \varphi, q) \frac{\partial \varphi}{\partial \beta} q \\ &= L(t, \varphi, \dot{\varphi}) + \frac{\partial L}{\partial p} \bigg|_{(t, \varphi, \dot{\varphi})} \frac{\partial \varphi}{\partial \beta} q + \tilde{L}(t, \beta, q),\end{aligned}$$

in which φ , $\dot{\varphi}$, θ , and $\partial \varphi / \partial \beta$ are all evaluated at (t, β) and $\theta \in (0, 1)$. Thus, for all piecewise smooth trajectories $x(\cdot)$ whose graphs are in D and corresponding trajectory $\tilde{x}(\cdot) = \psi(\cdot, x(\cdot))$ whose graph is in R we have from the above that

$$\begin{aligned}
L(t, x(t), \dot{x}(t)) - \tilde{L}(t, \tilde{x}(t), \dot{\tilde{x}}(t)) &= L(t, \varphi(t, \tilde{x}(t)), \dot{\varphi}(t, \tilde{x}(t))) \\
&+ \left(\frac{\partial L}{\partial p} \Big|_{(t, \varphi(t, \tilde{x}(t)), \dot{\varphi}(t, \tilde{x}(t)))} \right)^T \frac{\partial \varphi}{\partial \beta} \Big|_{(t, \tilde{x}(t))} \dot{\tilde{x}}(t). \quad (5.26)
\end{aligned}$$

From the fact that D is a field for (5.1) we know there exists a function $G(\cdot, \cdot)$ defined on R such that the right-hand side of (5.26) may be represented as a total differential. That is, (5.26) becomes

$$\begin{aligned}
L(t, x(t), \dot{x}(t)) - \tilde{L}(t, \tilde{x}(t), \dot{\tilde{x}}(t)) &= \frac{\partial G}{\partial t}(t, \tilde{x}(t)) + \left(\frac{\partial G}{\partial \beta}(t, \tilde{x}(t)) \right)^T \dot{\tilde{x}}(t) \\
&= \frac{d}{dt} G(t, \tilde{x}(t)). \quad (5.27)
\end{aligned}$$

Thus we see that the fundamental identity (5.10) holds and the desired result follows. \square

We now are ready to prove the Weierstrass sufficiency theorem.

Theorem 5.3. *If $x^*(\cdot) : [a, b] \rightarrow \mathbb{R}^n$ is a twice continuously differentiable solution of the Euler–Lagrange equations (5.6) satisfying the end conditions (5.2) which can be embedded in a field D for the functional (5.1) and if for $(t, x) \in D$ and all $q \in \mathbb{R}^n$ the inequality*

$$E(t, x, \hat{p}(t, x), q) \geq 0, \quad (5.28)$$

where $\hat{p}(t, x) = \dot{\varphi}(t, \beta) = \dot{\varphi}(t, \psi(t, x))$ denotes the slope function of the field, then $x^*(\cdot)$ is a strong minimizer over the class of all piecewise smooth trajectories $x(\cdot)$ satisfying the end conditions (5.2) and whose graphs lie in D .

Remark 5.8. In the above theorem, if the field $D = [a, b] \times \mathbb{R}^n$, then the above result gives a global minimum. In general, however, the field $D \subset [a, b] \times \mathbb{R}^n$ and as such we obtain only a local strong minimum since D , a point set in $[a, b] \times \mathbb{R}^n$, places no restrictions on the derivatives of the trajectories.

Proof. To begin we observe that as a consequence of the inequality (5.28) we know that $\tilde{L}(t, \beta, q)$ given by (5.25) is non-negative. Further, since $x^*(\cdot)$ is embedded in the field D , it follows that there exists a unique parameter β^* such that $(t, \beta^*) \in R$ for all $t \in [a, b]$ and that its corresponding trajectory $\tilde{x}^*(t) \equiv \beta^*$ has its graph in R . From the above, the associated problem (\tilde{P}) is that of minimizing the non-negative functional

$$\begin{aligned}
\tilde{J}(\tilde{x}(\cdot)) &= \int_a^b \tilde{L}(t, \tilde{x}(t), \dot{\tilde{x}}(t)) dt \\
&= \int_a^b \frac{1}{2} \dot{\tilde{x}}(t)^T \left(\frac{\partial \varphi}{\partial \beta} \right)^T \Big|_{(t, \tilde{x}(t))} Q(t, \varphi(t, \tilde{x}(t)), \dot{\tilde{x}}(t)) \frac{\partial \varphi}{\partial \beta} \Big|_{(t, \tilde{x}(t))} \dot{\tilde{x}}(t) dt
\end{aligned}$$

over all piecewise smooth trajectories $\tilde{x}(\cdot) : [a, b] \rightarrow \mathbb{R}^n$ satisfying the end conditions

$$\tilde{x}(a) = \beta^* \quad \text{and} \quad \tilde{x}(b) = \beta^*,$$

and such that its graph $\{(t, \tilde{x}(t)) : t \in [a, b]\} \subset R$. We further notice that the functional $\tilde{J}(\cdot)$ is identically zero whenever $\dot{\tilde{x}}(t) \equiv 0$. Thus one solution of the associated problem is $\tilde{x}^*(t) \equiv \beta^*$. As a consequence of Lemma 5.1, we see that $x^*(\cdot)$ is indeed a minimizer of (5.1) over all piecewise trajectories whose graphs are in D and which satisfy the end conditions (5.2). \square

Remark 5.9. A version of the above presentation for the case $n = 1$ was presented earlier in Carlson and Leitmann [13]. The details presented here, in addition to considering $n \geq 1$, however, are more complete.

5.6 Conclusion

In this chapter we presented an elementary proof of the Weierstrass sufficiency theorem for a strong local minimum for a free problem in the calculus of variations. Our approach is to view the n -parameter family of field trajectories as a coordinate transformation which defines a one-to-one mapping between the piecewise smooth trajectories satisfying the requisite end conditions (5.2) and the piecewise smooth trajectories which satisfy a fixed set of periodic end conditions. In this way we could exploit Leitmann's direct sufficiency method to present our proof. From this we see that the direct sufficiency method can be viewed as a generalization of Weierstrass's classical result since in general the end points of the trajectories for problem (\tilde{P}) need not satisfy constant end point conditions in the direct sufficiency method. Although the results presented here are not new, our approach illustrates the potential of Leitmann's direct sufficiency method. Future directions of research include further applications in the theory of optimal control, differential games, and other areas of dynamic optimization.

References

1. Leitmann, G.: A note on absolute extrema of certain integrals. *International Journal of Non-Linear Mechanics* **2**, 55–59 (1967)
2. Leitmann, G.: On a class of direct optimization problems. *Journal of Optimization Theory and Applications* **108**(3), 467–481 (2001)
3. Dockner, E.J., Leitmann, G.: Coordinate transformation and derivation of open-loop Nash equilibrium. *Journal of Optimization Theory and Applications* **110**(1), 1–16 (2001)
4. Leitmann, G.: Some extensions of a direct optimization method. *Journal of Optimization Theory and Applications* **111**, 1–6 (2001)
5. Carlson, D.A.: An observation on two methods of obtaining solutions to variational problems. *Journal of Optimization Theory and Applications* **114**, 345–362 (2002)
6. Carlson, D.A., Leitmann, G.: An extension of the coordinate transformation method for open-loop Nash equilibria. *Journal of Optimization Theory and its Applications* **123**(1), 27–47 (2004)

7. Carlson, D.A., Leitmann, G.: A direct method for open-loop dynamic games for affine control systems. In: A. Haurie, G. Zaccour (eds.) *Dynamic Games: Theory and Applications*, pp. 37–55. Springer, New York, NY (2005)
8. Carlson, D.A., Leitmann, G.: The direct method for a class of infinite horizon dynamic games. In: C. Deissenberg, R.F. Hartl (eds.) *Optimal Control and Dynamic Games, Applications in Finance, Management Science and Economics*. Springer, New York, NY (2005)
9. Carlson, D.A., Leitmann, G.: A coordinate transformation method for the extremization of multiple integrals. *J. Optimization Theory and Applications* **127**, 523–533 (2005)
10. Leitmann, G.: A direct method of optimization and its applications to a class of differential games. *Dynamics of Continuous, Discrete and Impulsive Systems, Series A* **11**, 191–204 (2004)
11. Carathéodory, C.: *Calculus of Variations and Partial Differential Equations*. Chelsea, New York, NY (1982)
12. Akhiezer, N.I.: *The calculus of variations*. Translated from the Russian by Aline H. Frink. A Blaisdell Book in the pure and Applied Sciences. Blaisdell Publishing Co. (A Division of Random House, Inc.), New York-London (1962)
13. Carlson, D.A., Leitmann, G.: Fields of extremals and sufficient conditions for the simplest problem of the calculus of variations. *Journal of Global Optimization* **40**(1–3), 41–50 (2007).

“This page left intentionally blank.”

Chapter 6

A Framework for Aerodynamic Shape Optimization

Giampiero Carpentieri and Michel J.L. van Tooren

Abstract A framework for aerodynamic shape optimization is presented. It uses a shape parameterization method based on the Chebyshev polynomials, an unstructured finite-volume formulation for the solution of the Euler equations, and a discrete adjoint method for the computation of the sensitivity. The framework is demonstrated on 2D and 3D shape optimization problems for which the drag coefficient must be minimized, the lift coefficient must be kept constant, and several geometrical constraints must be satisfied.

6.1 Introduction

Shape optimization frameworks that use computational fluid dynamics (CFD) solvers may have several components, the number of which depends on how much they are sophisticated. The essential components of a CFD-based framework are the optimizer and the shape parameterization. The former drives the optimization process whereas the latter deforms the shape according to the values of the shape parameters. In addition, in order to avoid re-meshing, a mesh deformation algorithm may be present. More sophisticated frameworks also have the capability to efficiently compute the gradients by means of the adjoint method [1].

The level of complexity of the implementation varies for the different components. The mesh deformation algorithm, if based on the spring analogy, is straightforward to implement. The adjoint solver is probably the most complex component to implement [2–6]. In the case of a discrete adjoint solver the implementation of a code that

Giampiero Carpentieri
Delft University of Technology, Delft, The Netherlands,
e-mail: g.carpentieri@tudelft.nl

Michel J.L. van Tooren
Delft University of Technology, Delft, The Netherlands,
e-mail: m.j.l.vantooren@tudelft.nl

has memory requirements and execution times similar to that of the flow solver may be difficult. The main difficulty is that complex algorithms must be devised that perform matrix-vector products on-the-fly in order to avoid the storage of the matrices. In fact, matrix storage may not be affordable for large cases. Other difficulties are the differentiation of the flow solver, which may be very time consuming, and the solution of the adjoint equations. The parameterization of the shape also deserves particular attention because the effectiveness of the framework heavily depends on it. Failure to represent the design space completely may preclude the possibility of finding optimal designs. Orthogonal representations that feature completeness are available for 2D cases and are relatively easy to apply [7]. In contrast, applications to 3D cases require the parameterization to be extended in order to deal with complex geometries.

The present work describes an adjoint-based aerodynamic shape optimization framework for problems that are governed by the Euler equations. An overview of the framework is given first. Then, the different components of the framework are briefly described and the aforementioned issues are addressed. More detailed descriptions may be found in previous works published by the authors [8–10]. Finally, results are presented that demonstrate the effectiveness of the framework for 2D and 3D cases. Inviscid drag minimization problems with constraints on the lift and on the geometry are considered.

6.2 Adjoint-Based Sensitivity Analysis

The adjoint method is a very efficient way of performing sensitivity analysis in the context of aerodynamic design [4, 11, 12]. Sensitivity analysis is an additional solution phase. It is performed after the flow solution and evaluates the sensitivity of the flow functionals with respect to the shape parameters. In practice, sensitivity equations are derived and are solved by numerical procedures that are similar to that used for the flow equations. In the case of the adjoint method the equations are usually referred to as adjoint equations. The efficiency of the adjoint method is due to the fact that the gradients can be computed at a cost that is independent of the number of shape parameters and that is proportional to the number of functionals. The method is described in the following.

Consider the vector of functionals $\bar{J} = [J_1, J_2, \dots, J_M]$ and the vector of shape parameters $\bar{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_N]$. \bar{J} may contain the lift, the drag, and the pitching moment whereas $\bar{\alpha}$ may contain geometric parameters acting on the boundary, e.g., shape coefficients, airfoil thicknesses, twist angles. Note that the shape parameters are the design variables of the optimization problem. The functionals have a dependence $\bar{J} = \bar{J}(\mathbf{U}(\bar{\alpha}), \bar{\alpha})$ on the shape parameters, where $\mathbf{U}(\bar{\alpha})$ is the vector of conservative flow variables. Also, $\mathbf{R}(\mathbf{U}(\bar{\alpha}), \bar{\alpha})$ is the residual vector of the flow solution, which is known to satisfy the state equation $\mathbf{R} = \mathbf{0}$ in the case of a converged steady flow. In order to compute the gradient of the J_i functional with respect to

the parameters $\bar{\alpha}$ it is convenient to define the augmented functional L_i . The latter functional is augmented by the state equation and reads

$$L(\mathbf{U}, \bar{\alpha}, \mathbf{\Lambda}_i) = J_i(\mathbf{U}, \bar{\alpha}) - \mathbf{\Lambda}_i^T \mathbf{R}(\mathbf{U}, \bar{\alpha}). \quad (6.1)$$

At the stationary point $[\partial L / \partial \mathbf{U}, \partial L / \partial \mathbf{\Lambda}_i] = \mathbf{0}$, the sensitivity of the augmented functional with respect to the j th design variable α_j coincides with that of the original functional, i.e.,

$$\frac{dJ_i}{d\alpha_j} \equiv \frac{\partial L_i}{\partial \alpha_j} = \frac{\partial J_i}{\partial \alpha_j} - \mathbf{\Lambda}_i^T \frac{\partial \mathbf{R}}{\partial \alpha_j}. \quad (6.2)$$

Imposing the stationary condition on the augmented functional gives two equations. The one obtained by differentiating with respect to $\mathbf{\Lambda}_i$ is the state equation, which is satisfied by the steady flow solution. The one obtained by differentiating with respect to \mathbf{U} is the adjoint equation, which reads

$$\frac{\partial \mathbf{R}^T}{\partial \mathbf{U}} \mathbf{\Lambda}_i = \frac{\partial J_i^T}{\partial \mathbf{U}}. \quad (6.3)$$

The above equation must be solved in order to obtain the multipliers $\mathbf{\Lambda}_i$, which are used to evaluate the gradient according to Eq. (6.2). Usually the multipliers are referred to as adjoint variables.

6.3 Optimization Framework

Figure 6.1 shows the diagram of the framework implemented in the present work. The diagram illustrates the optimization process, which is repeated a certain number of design iterations in order to improve an existing design.

The optimizer drives the process by means of an optimization algorithm. At each design iteration it receives the functionals, eventually their gradients, and computes a new set of design variables $\bar{\alpha}$. The latter variables should give an improvement in the design, which means minimizing the objective function while satisfying the design constraints.

The design variables $\bar{\alpha}$ define the displacements $\Delta \mathbf{X}_B$ of the boundary surface. After the design variables are computed by the optimizer, the shape parameterization module takes them as input and generates the displacements. The module also generates the geometric functionals \bar{I} . For a 2D case the latter functionals may be the relative thickness, the nose radius, and the trailing edge angle. For a 3D case the same quantities may be evaluated at different sections along the span. Also the wing volume may be evaluated for the purpose of imposing a constraint on it.

The mesh deformation module imposes the displacements $\Delta \mathbf{X}_B$ on the wing surface and computes the deformed mesh coordinates \mathbf{X} . The deformation is realized

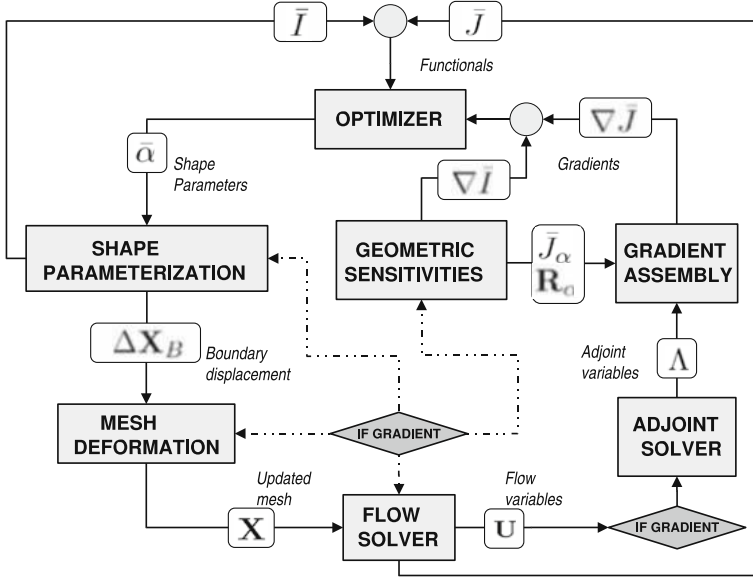


Fig. 6.1 Adjoint-based shape optimization framework

by the spring analogy method, which propagates the surface deformations into the volume mesh by means of Jacobi iterations.

The updated mesh is used by the flow solver to compute the flow field U and to evaluate the flow functionals \bar{J} needed by the optimizer. If the optimization algorithm requires the gradients, the adjoint solver computes the adjoint variables Λ . The latter variables are used, together with the geometric sensitivities, to assemble the gradients of the flow functionals $\nabla \bar{J}$.

The term geometric sensitivities is used to refer to the partial derivatives with respect to geometric quantities, e.g., the partial derivatives of the functional ($\bar{J}_\alpha = [\partial J_i / \partial \alpha_j]$, for $j = 1, N$ and $i = 1, M$) and those of the residuals vector ($R_\alpha = [\partial R / \partial \alpha_j]$, for $j = 1, N$). Compared to the flow sensitivity, the geometric sensitivities \bar{J}_α and R_α are inexpensive to compute. Only the sensitivity of the mesh deformation algorithm may be relatively expensive to compute because it involves an iterative solution procedure.

6.3.1 Flow Solver

The flow solver is based on an unstructured finite-volume formulation that discretize the Euler equations on the median-dual mesh [13]. The control volumes of the median-dual mesh are located on the nodes of the original mesh. An edge-based data structure is used, which makes no distinction between 2D and 3D or between different types of elements. A linear reconstruction scheme is employed, which reconstructs the primitive variables across the control volume interfaces, i.e.,

at the mid-point of each edge of the mesh. The reconstruction uses a least-squares or a Green–Gauss gradient and a multi-dimensional type of limiter. The Roe’s approximate Riemann solver is used to evaluate the discontinuous states across the interfaces. Weak boundary condition, i.e., zero normal fluxes, is enforced on the wall and on the symmetry type of boundaries. Flux-vector splitting is used for the boundaries at infinity. More details can be found in [3, 4].

The discretized flow equations are solved by means of an implicit pseudo-time stepping scheme, which is derived by applying the defect correction method to the semi-discrete form of the equations [14]. In practice, the scheme coincides with a backward Euler method that uses an approximate Jacobian [15]. It may be written as

$$\left(\mathbf{D}_t + \frac{\partial \tilde{\mathbf{R}}}{\partial \mathbf{U}} \right)^n (\mathbf{U}^{n+1} - \mathbf{U}^n) = -\mathbf{R}^n. \quad (6.4)$$

As already mentioned, \mathbf{R} and \mathbf{U} are the residuals and the conservative variables vector, respectively. \mathbf{D}_t is a diagonal matrix, which contains the control volumes divided by the local time steps, and $\partial \tilde{\mathbf{R}} / \partial \mathbf{U}$ is the first-order and approximate Jacobian of the residuals vector. The solution is initialized with the free-stream flow values. The above equation is solved and the solution is updated. The latter process is iterated until the flow is converged.

At each pseudo-time step n Eq. (6.4) is a sparse linear system of equations, which is solved iteratively by means of a symmetric-Gauss–Seydel procedure. If one defines $\Delta \mathbf{U} = \mathbf{U}^{n+1} - \mathbf{U}^n$, the linear iterative procedure may be written as

$$\Delta \mathbf{U}^{k+1} = \Delta \mathbf{U}^k - \mathbf{P}_M^{-1} \left[\mathbf{R}^n + \left(\mathbf{D}_t + \frac{\partial \tilde{\mathbf{R}}}{\partial \mathbf{U}} \right)^n \Delta \mathbf{U}^k \right]. \quad (6.5)$$

\mathbf{P}_M is the symmetric-Gauss–Seydel preconditioner, which may be expressed as

$$\mathbf{P}_M = (\mathbf{D}_M + \mathbf{L}_M) \mathbf{D}_M^{-1} (\mathbf{D}_M + \mathbf{U}_M), \quad (6.6)$$

where the matrices \mathbf{D}_M , \mathbf{L}_M , and \mathbf{U}_M are, respectively, the diagonal, the strictly lower, and the strictly upper part of the matrix at the left-hand side of Eq. (6.4). The structure of the preconditioner allows the product in Eq. (6.5) to be carried out with two sweeps on the nodes of the mesh. A forward sweep for the lower part followed by a backward sweep for the upper part. The linear iterations are stopped when the norm of the linear residuals vector is one order of magnitude smaller than the norm of the residuals vector \mathbf{R} . In practice, an average number of linear iterations $k \approx 5 - 8$ is usually satisfactory. Higher accuracies are not beneficial because of the approximate nature of the Jacobian.

An interesting feature of the solution scheme is the possibility of running matrix free, i.e., without storing the off-diagonal terms of the Jacobian matrix, \mathbf{L}_M and \mathbf{U}_M . The matrix-free option requires an amount of memory similar to that of an explicit scheme. However, if quantities are not stored, they must be re-computed at each iteration. Thus, the matrix-free option has higher requirements in terms of CPU-time.

6.3.2 Adjoint Solver

Although the adjoint equation in (6.3) is linear, the second-order Jacobian $\partial \mathbf{R} / \partial \mathbf{U}$ is poorly diagonally dominant. Consequently an iterative linear solver is not suitable for the solution. Instead, a widely used approach to the solution of the adjoint equations is that of using the flow solution scheme [4, 6, 16]. In practice, it means that the adjoint equations are treated as non-linear equations.

Application of the implicit scheme of Eq. (6.4) to the adjoint in Eq. (6.3) gives

$$\left(\mathbf{D}_t + \frac{\partial \tilde{\mathbf{R}}^T}{\partial \mathbf{U}} \right)^n (\mathbf{A}_i^{n+1} - \mathbf{A}_i^n) = - \left(\frac{\partial \mathbf{R}^T}{\partial \mathbf{U}} \mathbf{A}_i^n - \frac{\partial J_i^T}{\partial \mathbf{U}} \right). \quad (6.7)$$

The second-order Jacobian is now at the right-hand side of the equation and the solution process is driven by the diagonally dominant first-order Jacobian. The implicit scheme in the above equation, which is already available from the flow solver, has the advantage of being particularly robust.

The solution of Eq. (6.7) requires the assembly of matrix-vector products. At the left-hand side of the equation the matrix-vector products are assembled easily in one step, by looping over the edges of the mesh. At the right-hand side the assembly is more complicated because the Jacobian is second order. The presence of the second-order contribution, represented by the differentiation of the reconstruction operator, requires the product to be carried out at least in two steps [10]. Moreover, in order to keep the memory usage as low as possible, the assembly must be carried out on-the-fly. If \mathbf{G} is the vector of primitive variables gradients, the right-hand side product may be written as the sum of two contributions, i.e.,

$$\frac{\partial \mathbf{R}^T}{\partial \mathbf{U}} \mathbf{A}_i = \mathbf{P}_1 \mathbf{A}_i + \frac{\partial \mathbf{G}^T}{\partial \mathbf{U}} \mathbf{P}_2 \mathbf{A}_i. \quad (6.8)$$

\mathbf{P}_1 and \mathbf{P}_2 are sparse matrices. In the case of first-order accuracy, the second term on the right-hand side disappears. In practice, the products $\mathbf{P}_1 \mathbf{A}_i$ and $\mathbf{P}_2 \mathbf{A}_i$ are assembled first. Then, the transposed gradient operator is run with $\mathbf{P}_2 \mathbf{A}_i$ as input and the result is added to $\mathbf{P}_1 \mathbf{A}_i$.

The terms of the matrices \mathbf{P}_1 , \mathbf{P}_2 , and $\partial \mathbf{G} / \partial \mathbf{U}$ are obtained by differentiating the flow solver and by performing the transposition of the result. Differentiation may be very complicated because of the presence of non-linear functions. Performing the products between the transposed matrices and the vectors on-the-fly also increases the level of complexity of the implementation. In fact, the transposition implies that the operations within the differentiated code have to be reversed in a counter-intuitive way. Some authors proposed the use of automatic differentiation in order to speed up the derivation of the code [2, 4]. In the present work the adjoint solver is hand-coded. The hand-coding approach is demanding in terms of human work but gives the possibility to develop very efficient code in terms of memory and CPU-time requirements.

Equation (6.7) shows that the adjoint variables Λ_i must be calculated for each functional J_i , which means that the equations should be solved as many times as the number of functionals N . Instead, the equations may be solved simultaneously. In fact, as can be seen from Eq. (6.7), different equations have the same Jacobian in common. Since the Jacobian terms are expensive to compute, it saves time to perform the different matrix-vector products at the left- and right-hand side of Eq. (6.7) simultaneously. In practice, additional nested loops must be implemented, which allows the multiplication of the matrix terms with the different vectors.

Convergence histories of the flow and adjoint solvers for a 3D transonic case are shown in Fig. 6.3. The mesh used for the computation is shown in Fig. 6.2b. The convergence histories are plotted in terms of CPU time. In Fig. 6.3a the elements \mathbf{L}_M , \mathbf{U}_M , and \mathbf{D}_M have been stored for both the two solvers. It appears that the adjoint solver residual overlaps with the flow solver residual, i.e., one adjoint solution requires the same amount of time as one flow solution. In the case of multiple adjoint solutions, Fig. 6.3a shows that the simultaneous solution of two adjoints saves 25% of CPU time compared to two sequential solutions. In the case of three adjoint solutions, it shows that the CPU time saving rises to 33%.

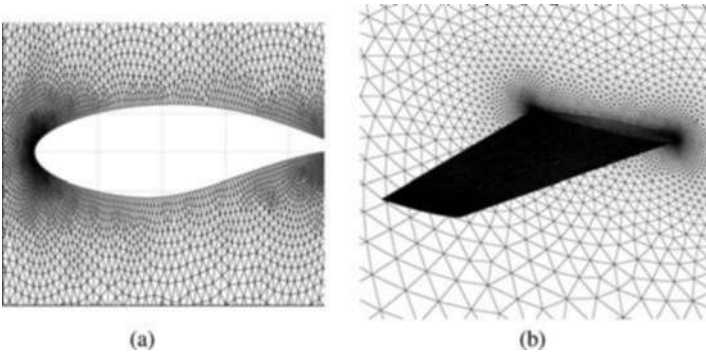


Fig. 6.2 Unstructured meshes: (a) REA2822 airfoil; (b) ONERA-M6 wing

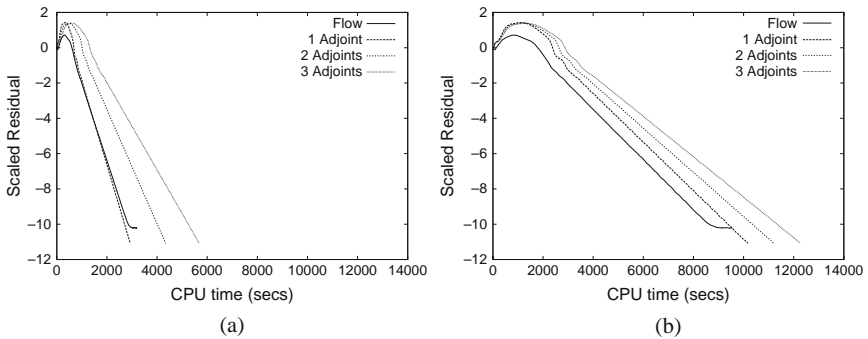


Fig. 6.3 ONERA-M6 wing at $M_\infty = 0.84$ and $\alpha = 3.06^\circ$. Flow and adjoint convergence histories: (a) storage of the preconditioner; (b) matrix-free preconditioning

The convergence histories of Fig. 6.3b have been produced using the matrix-free option, i.e., \mathbf{L}_M , \mathbf{U}_M are always computed on-the-fly. Clearly, there is a penalty in terms of time. As can be seen, the CPU time required to converge the flow is around three times more than in the storage case of Fig. 6.3a. However, the memory requirements are also reduced by a third. A single matrix-free adjoint solution required in this case around 10% more CPU time than the flow solution. Compared to the storage case, the time savings provided by the matrix-free option are amplified. As can be seen from Fig. 6.3b, two simultaneous adjoint solutions give around 45% time saving compared to sequential solutions. In the case of three adjoint solutions, the time saving rises to 60%.

6.3.3 Shape Parameterization

The shape parameterization uses Chebyshev polynomials. The basis functions T_k of the polynomials are not used directly. A linear combination D_k is used [7], which is defined as

$$D_k = T_k - T_{k+2}, \quad T_k(x^*) = \cos(k\gamma(x^*)), \quad \gamma(x^*) = \cos^{-1}(2\sqrt{x^*} - 1), \quad (6.9)$$

where $k \geq 0$ and $0 \leq x^* \leq 1$. The dimensionless coordinate is defined as $x^* = x/c$, where c is the chord section. The linear combination is necessary in order to ensure closure at the leading and trailing edges, i.e., for $x^* = 0$ and $x^* = 1$, respectively.

The displacements of an airfoil curve are written as

$$\Delta f(x) = c \sum_{k=0}^{N_x} \alpha_k D_k(x^*). \quad (6.10)$$

The coefficient α_k is the design variable in the optimization process. The displacements of a wing surface, in the direction normal to the wing planform, are written as

$$\Delta f(x, y) = c(y^*) \sum_{k=0}^{N_x} \left(\sum_{m=1}^{N_y} \alpha_{km} y^{*m-1} \right) D_k(x^*). \quad (6.11)$$

y^* is the scaled coordinate along the span-wise direction. Compared to the 2D parameterization, the above parameterization has variable coefficients along the span-wise direction. In fact, the term $\sum_{m=1}^{N_y} \alpha_{km} y^{*m-1}$ varies along the span linearly, if $N_y = 1$, or quadratically, if $N_y = 2$. The $N_x \times N_y$ coefficients α_{km} are the design parameters. If zero deformations must be specified at the tip or at the root of the wing, some of the coefficients are not considered as variables. Instead, they are kept fixed to certain values, which reflect the explicit imposition of a geometric constraint.

The shape parameterization may be differentiated in order to compute its first and second derivatives [9]. The derivatives can then be used to compute the trailing

edge angle and the nose radius of curvature. Constraints are enforced on the latter quantities during the optimization process.

6.3.4 Geometric Sensitivities

The computation of the gradients according to Eq. (6.3) requires the geometric sensitivities \bar{J}_α and \mathbf{R}_α . The gradient module computes matrix-vector products between the geometric sensitivities and the adjoint variables in order to assemble the gradient of the flow functionals $\nabla \bar{J}$. The geometric sensitivities may be computed by finite differences, which means that one additional evaluation of the functionals and of the residuals is needed for each design variable. Therefore, the geometric sensitivities module must be linked to the shape, mesh deformation, and flow solver modules. The links are represented by the dashed lines in Fig. 6.1. The geometric sensitivities module also produces the gradient of the geometric functionals $\nabla \bar{I}$.

Instead of finite differences, automatic differentiation (AD) may be used for the computation of the geometric sensitivities. In the present work the forward mode of the AD tool Tapenade [17] has been used to differentiate each routine of the code.

6.3.5 Optimization Algorithm

In the present work constrained optimization problems are considered. The drag coefficient is the objective function. An equality constraint is imposed on the lift coefficient, which must be kept constant. For 2D cases inequality constraints are imposed on the relative thickness, on the nose radius and on the trailing edge angle. For 3D cases, the latter quantities are considered at different sections along the span. Moreover, also an inequality constraint on the wing volume is used.

The aforementioned problems are solved by constrained optimization algorithms. A very simple Sequential Linear Programming algorithm has been employed. It is known as the method of centers [18]. It is certainly not the state of the art in optimization but is effective in solving the problems considered here. The algorithm tries to find the largest hypersphere that fits into the linearized design space, which is delimited by the gradients of the objective function and of the constraints. The optimum point is reached by computing the hypersphere and moving toward its center, repeating the process for a certain number of iterations until a satisfactory improvement is obtained. The lack of second-order information makes the algorithm inefficient in the neighborhood of the optimum. More details about the application of the algorithm can be found in [8]. Efficient Sequential Quadratic Programming algorithms have also been used successfully [9].

The use of constrained algorithms requires the computation of more than one adjoint. For instance, the cases presented here require the adjoint equations for the lift and for the drag to be solved. An alternative may be the use of unconstrained

algorithms, which include the constraints as penalty terms in the objective function and therefore require only one adjoint solution. However, the penalty approach may generate ill-conditioned problems [18]. Also, it may not be easy to satisfy the constraints accurately.

6.4 Optimization Test Cases

In the following four optimization test cases are presented. The first two are 2D transonic cases, the third one is a 2D supersonic case and the fourth one is a 3D transonic case. For all the cases only the shape parameters are used as design variables. The angle of attack is always kept constant. The mesh for the RAE2822 airfoil, which is shown in Fig. 6.2a, is deformed and used for all the other 2D cases presented below. The mesh for the ONERA-M6 optimization is that of Fig. 6.2b.

6.4.1 RAE2822 at $M_\infty = 0.73$ and $\alpha = 2^\circ$

Figure 6.4a, which shows the pressure contours around the airfoil, reveals that a shock is present on the upper surface. An optimization is carried out with the objective of reducing the drag. During the optimization the relative maximum thickness is not allowed to decrease; the nose curvature and the trailing edge angle cannot decrease more than 30 and 10% of their initial values, respectively; and the lift must be kept constant.

The optimization is completed after 15 design iterations. As can be seen in Fig. 6.4b the final airfoil is shock free. The objective function, which is the drag coefficient scaled by its initial value, is depicted in Fig. 6.5a. It shows a reduction of 40%. The inequality constraints are satisfied. The lift equality is very accurate, with percentage differences between the final and the initial lift coefficient of the

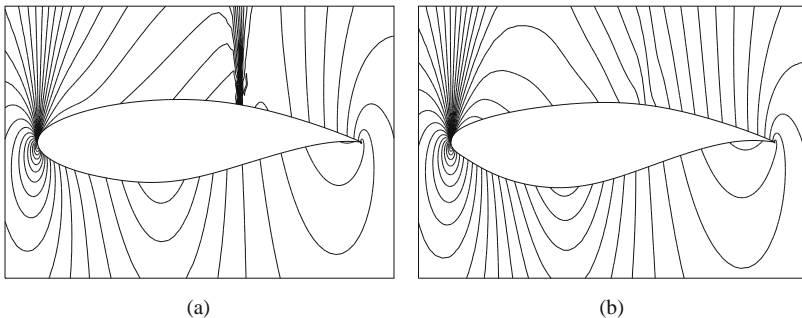


Fig. 6.4 RAE2822 optimization. Pressure contours: (a) RAE2822 airfoil; (b) optimized airfoil.

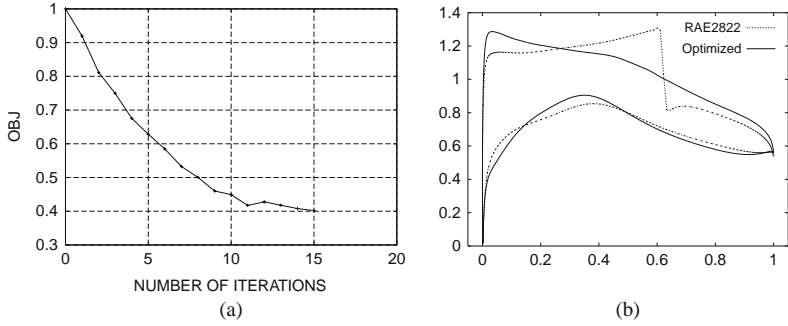


Fig. 6.5 RAE2822 optimization: (a) objective function; (b) comparison of initial and optimized Mach number distributions

order of 0.005%. A comparison between the initial and the optimized Mach number distributions is shown in Fig. 6.5b.

6.4.2 NACA64A410 at $M_\infty = 0.75$ and $\alpha = 0^\circ$

Figure 6.6a shows a strong shock on the upper side of the airfoil. The optimization settings are the same as that of the previous case, except for the nose radius and for the trailing edge angle. In this case the latter quantities cannot decrease more than 10% of their initial values.

The optimization is completed after 25 design iterations. The objective function is depicted in Fig. 6.7a. It shows a reduction of 85%. All the constraints are satisfied. Figure 6.6a shows that the final airfoil is shock-free. If compared to the initial NACA64A410 shape, the optimized shape appears to be rather complex and looks similar to that of a typical transonic airfoil. A comparison between the initial and the optimized Mach number distributions is shown in Fig. 6.7b.

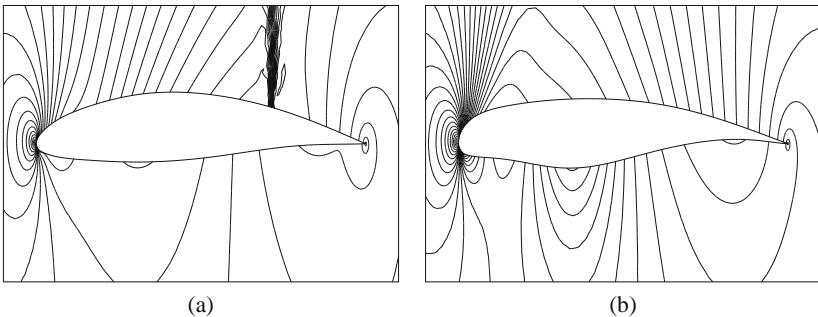


Fig. 6.6 NACA64A410 optimization. Pressure contours: (a) NACA64A410 airfoil; (b) optimized airfoil

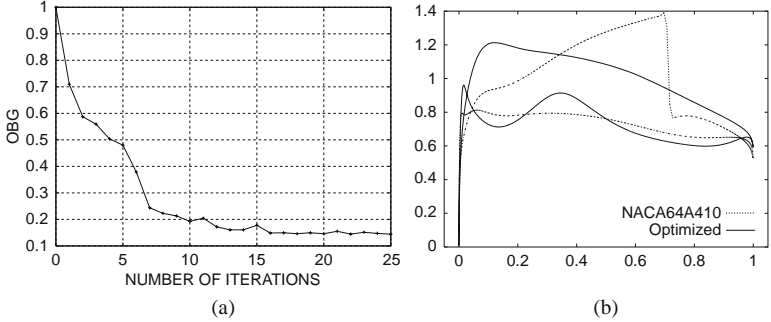


Fig. 6.7 NACA64A410 optimization: (a) objective function; (b) comparison of initial and optimized Mach number distributions

6.4.3 NACA0012 at $M_\infty = 1.5$ and $\alpha = 2^\circ$

A supersonic optimization is a good test case for the shape parameterization and for the geometric constraints. They both play a crucial role in the supersonic case. In fact, in order to reduce the drag, the nose is expected to become sharp so that the bow shock tends to be attached to it and oblique. The situation of a nose that tries to become sharp must be handled carefully. The danger exists that the upper and the lower curves can cross each other. In practice, airfoils do not have sharp noses but more likely noses with small radii of curvature. Thus, geometric constraints must be enforced properly that allow small radii to exist without generating unfeasible designs.

Figure 6.8a shows a strong detached bow shock. The NACA0012 airfoil has a 12% relative maximum thickness, which is unsuitable for supersonic flows. A more suitable value could be less than 6%. Therefore, the relative maximum thickness is allowed to halve. Moreover, the airfoil has a large nose radius. In order to allow the shock to move close to the nose, the nose radius is allowed to decrease up to 90% of its initial value. Also in this case, the Lift coefficient must be kept constant.

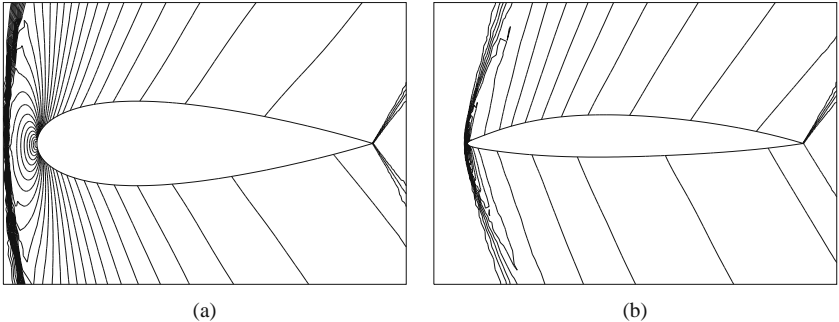


Fig. 6.8 NACA0012 optimization. Pressure contours: (a) NACA0012 airfoil; (b) optimized airfoil

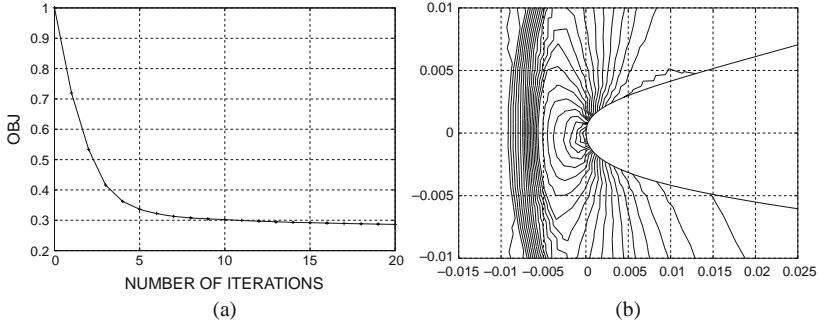


Fig. 6.9 NACA0012 optimization: (a) objective function; (b) magnification on the nose of the optimized airfoil

The objective function, see Fig. 6.9a, smoothly reduced 70% in 20 design iterations. The large reduction in drag is due to the repositioning of the shock, which is almost attached to the nose at the end of the optimization, see Fig. 6.8b. A magnification of the nose region is shown in Fig. 6.9b. It reveals that the shock is located at a distance from the nose that is less than 1% of the chord. As can be seen, the nose appears to be rounded and smooth. The curvature of the nose is now 10% of the initial curvature since both the upper and the lower curvature constraints are critical. The same is for the thickness constraint.

It is remarkable that also in this case, in spite of the challenging flow conditions and the large deformations involved, the whole framework proved to be effective and robust. Also, it is interesting to see that the final shape resembles that of lenticular airfoils, typically used in the supersonic regime.

6.4.4 ONERA-M6 wing at $M_\infty = 0.84$ and $\alpha = 3.06^\circ$

The ONERA-M6 wing is a swept wing with a flat planform. Each section of the wing has the same symmetric airfoil. For the flow condition considered here, the complex shock pattern on the upper surface of the wing is shown on the left part of Fig. 6.11. The optimization problem in this case is also a drag minimization at constant lift. Geometric constraints on the relative maximum thickness, on the nose radii and on the trailing edge angles, are imposed on the maximum values of these quantities along the span. An inequality constraint on the wing volume is also imposed, which avoids the volume to decrease below a certain value. The constraint may be useful in the case that a minimum volume is required for placing objects inside the wing, for instance a fuel tank. For the parameterization of the shape only the aforementioned shape functions are used. Twist of the wing sections is not implemented. Consequently the planform of the wing remains flat during the optimization process.

At the end of the optimization the drag is reduced by 30%. All the constraints are satisfied. Figure 6.10 shows the comparisons of the initial and the optimized

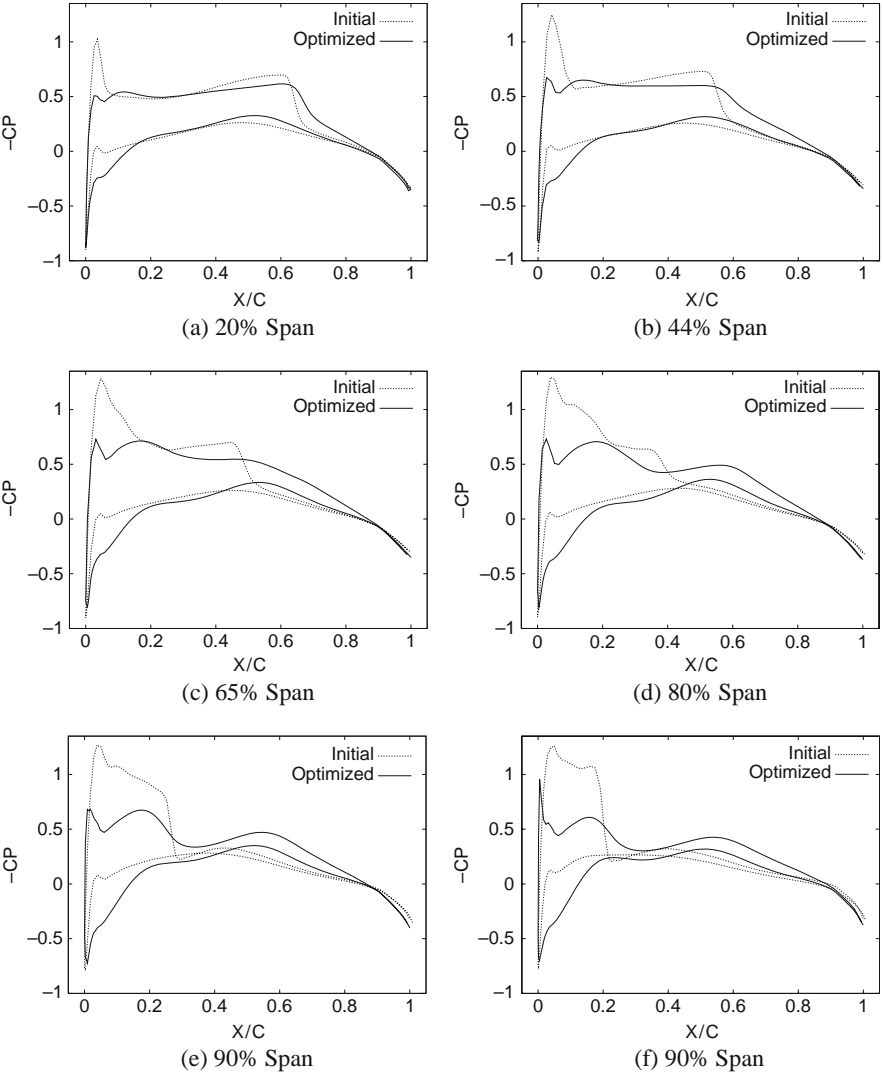


Fig. 6.10 ONERA-M6 optimization. Comparison of initial and optimized pressure distributions at different sections along the span

pressure distributions at different sections along the span. Figure 6.11 shows a comparison of the pressure on the upper surface of the wing. The strengths of the shocks appear to be reduced appreciably. It also appears that the velocity spikes along the leading edge of the original wing, which result in a strong shock, are reduced considerably in the optimized wing. The comparison of two wing section shapes is shown in Fig. 6.12. As can be seen, the optimized shape of the 95% section, which is close to the tip, does not appear to be very smooth in proximity of the leading

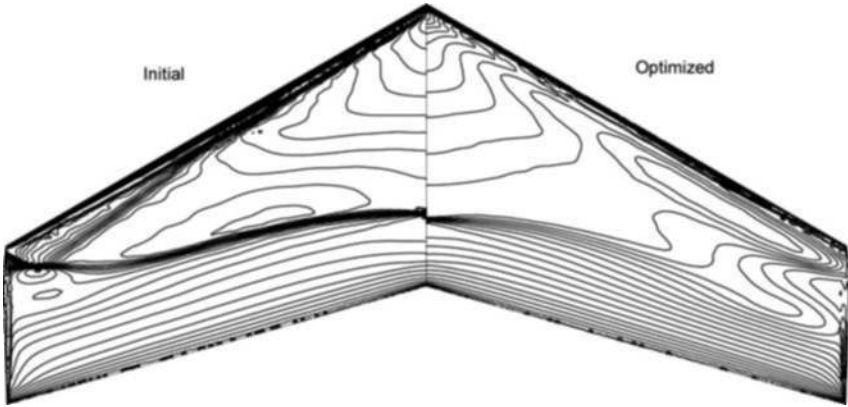


Fig. 6.11 ONERA-M6 optimization. Pressure contours: ONERA-M6 wing (*left*); Optimized wing (*right*)

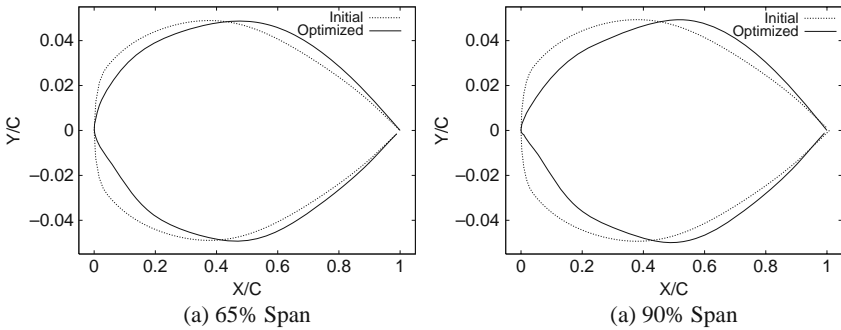


Fig. 6.12 ONERA-M6 optimization. Comparison of the initial and optimized airfoil shapes at two different wing sections

edge. Probably, smoother airfoils may be obtained if constraints on the sections are imposed in a different way. For instance, a possibility is to constrain the relative thickness at different point along the chord of the sections rather than imposing a constraint on its maximum value only. Also, it is possible that the implementation of twist deformations may have a positive effect on the smoothness. In fact, a fixed planform requires the optimizer to act only on the shape of the sections for the purpose of changing the pressure. The twist would give an additional degree of freedom to the optimizer.

6.5 Conclusions

An aerodynamic shape optimization framework has been presented. The main components of the framework have been briefly described and its effectiveness has been demonstrated on 2D and 3D optimization problems. Results are satisfactory for 2D

problems. Appreciable improvements in the design are observed. Also when large deformations are involved, the optimized shapes appear to be smooth. For 3D problems still some work is required. Although the improvement in the pressure distribution is appreciable, more appropriate geometric constraints must be defined in order to avoid non-smooth shapes. Also, deformations of the wing twist must be implemented and their effect on the optimization must be investigated.

Acknowledgments This research was supported by the Dutch Technology Foundation STW, applied science division of NWO and the technology program of the Dutch Ministry of Economic Affairs.

References

1. M.B. Giles, N.A. Pierce, An Introduction to the Adjoint Approach to Design, *Flow Turbul. Combust.* 65 (2000) 393–415.
2. B. Mohammadi, A New Optimal Shape Design Procedure for Inviscid and Viscous Turbulent Flows, *Int. J. Numer. Meth. Fluids* 25 (1997) 183–203.
3. N. Nemec, D.W. Zingg, Newton-Krylov Algorithm for Aerodynamic Design Using the Navier Stokes Equations, *AIAA J.* 40 (2002) 1146–1154.
4. M.B. Giles, M.C. Duta, J.-D. Müller, N.A. Pierce, Algorithm Developments for Discrete Adjoint Methods, *AIAA J.* 41 (2003) 198–205.
5. E.J. Nielsen, J. Lu, M.A. Park, D.L. Darmofal, An Implicit, Exact Dual Adjoint Solution Method for Turbulent Flows on Unstructured Grids, *Comput. Fluids* 33 (2004) 1131–1155.
6. O. Amoignon, M. Berggren, Adjoint of a Median-Dual Finite-Volume Scheme: Application to Transonic Aerodynamic Shape Optimization, Tech. Rep. 2006-13, Uppsala University, 2006.
7. A. Verhoff, D. Stooksberry, A.B. Cain, An Efficient Approach to Optimal Aerodynamic Design Part 1: Analytic Geometry and Aerodynamic Sensitivities, *AIAA Paper No. 93-0099*, 1993.
8. G. Carpentieri, M.J.L. van Tooren, B. Koren, Aerodynamic Shape Optimization by Means of Sequential Linear Programming Techniques, *ECCOMAS CFD*, 2006.
9. G. Carpentieri, M.J.L. van Tooren, B. Koren, Adjoint-Based Aerodynamic Shape Optimization on Unstructured Meshes, *J. Comput. Phys.* 224 (2007) 267–287.
10. G. Carpentieri, M.J.L. van Tooren, B. Koren, Development of the Discrete Adjoint for a 3D Unstructured Euler Solver, *J. Aircraft*, To appear.
11. J.C. Newman III, A.C. Taylor III, R.W. Barnwell, P.A. Newman, G.J.-W. Hou, Overview of Sensitivity Analysis and Shape Optimization for Complex Aerodynamic Configurations, *J. Aircraft* 36 (1999) 87–96.
12. B. Mohammadi, O. Pironneau, Shape Optimization in Fluid Mechanics, *Annu. Rev. Fluid Mech.* 36 (2004) 255–279.
13. T.J. Barth, Aspects of Unstructured Grids and Finite-Volume Solvers for the Euler and Navier-Stokes Equations, Lecture Series 1991-06, Von Karman Institute for Fluid Dynamics, 1991.
14. B. Koren, Defect Correction and Multigrid for an Efficient and Accurate Computation of Airfoil Flows, *J. Comput. Phys.* 77 (1988) 183–206.
15. D.J. Mavriplis, On Convergence Acceleration Techniques for Unstructured Meshes, *AIAA Paper No. 98-2966*, 1998.
16. D.J. Mavriplis, Formulation and Multigrid Solution of the Discrete Adjoint for Optimization Problems on Unstructured Meshes, *AIAA Paper No. 05-0319*, 2005.
17. <http://tapenade.inria.fr:8080/tapenade/index.jsp> Software Tapenade ©INRIA 2002, version 2.0
18. G.N. Vanderplaats, *Numerical Optimization Techniques for Engineering Design*, Vanderplaats Research & Development, Inc, 3rd ed., 2001.

Chapter 7

Optimal Motions of Multibody Systems in Resistive Media

Felix L. Chernousko

Abstract It is well known that a body containing internal masses can move in a resistive medium, if the internal masses perform oscillations relative to the body. In this chapter, progressive motions of a body carrying movable internal masses are considered for various resistance forces acting upon the body. The cases of linear and quadratic resistance as well as Coulomb's dry friction forces, both isotropic and anisotropic, are analyzed. Special classes of periodic motions of the internal masses are considered under constraints imposed on relative displacements, velocities, and accelerations of these masses. Optimal parameters of the relative internal motions are determined that correspond to the maximal average speed of the system as a whole. Results of the computer simulation and experimental data confirm the obtained theoretical results. The principle of motion analyzed in this chapter can be used for mobile robots, especially mini-robots, moving in tubes, in aggressive media, and in complex environment.

7.1 Introduction

A system of two or more bodies can move progressively in a resistive medium, if the bodies perform periodic motions relative to each other. One of these bodies (an inner one) can be contained within a certain closed cavity inside the other (outer) body, so that the system has no outward moving parts such as screws, wheels, legs, wings. This well-known principle of motion is utilized in various projects of mobile robots and underwater vehicles (see, e.g., [2, 12, and 13]).

In this chapter, simple models of this phenomenon are analyzed. The mechanical system under consideration consists of two rigid bodies of masses M and m . For brevity, these bodies will be called body M and mass m , respectively. Mass m moves

Felix L. Chernousko

Institute for Problems in Mechanics, Russian Academy of Sciences, Moscow, Russia
e-mail: chern@ipmnet.ru

periodically relative to the main body M which interacts with the outward medium and is subject to resistance forces.

Various kinds of resistance forces acting upon body M are considered, including linear and nonlinear resistance depending on the velocity of the body and also Coulomb's dry friction. The forces can be anisotropic, i.e., dependent on the direction of the velocity of body M .

The progressive motion of the system as a whole is controlled by the periodic motion of mass m relative to body M . Simple relative periodic motions are analyzed, and constraints are imposed on the relative displacements, velocities, and accelerations. Under the constraints imposed, optimal parameters of the periodic motions are determined that correspond to the maximal average speed of the system as a whole. The results obtained (see also [4–6]) enable one to evaluate the maximal possible speed of mobile mechanical systems that utilize the principle of motion based on relative oscillations of parts of the system moving in a resistive medium.

Experimental results confirm the practical implementability of this principle of motion.

7.2 Basic Equations

The system consists of two rigid bodies that can move along a straight line in a resistive medium (Fig. 7.1). Denote by x and v the absolute coordinate and velocity of the main body M , respectively, and by ξ , u , and w the displacement of the inner mass m relative to body M , its relative velocity, and acceleration, respectively.

The kinematic equations of motion of mass m relative to body M are

$$\dot{\xi} = u, \quad \dot{u} = w. \quad (7.1)$$

The dynamic equations for body M can be written as follows:

$$\dot{x} = v, \quad \dot{v} = -\mu w - r(v), \quad \mu = m/(M + m), \quad (7.2)$$

where $r(v)$ is the resistance force acting upon body M divided by the total mass of the system, $M + m$.

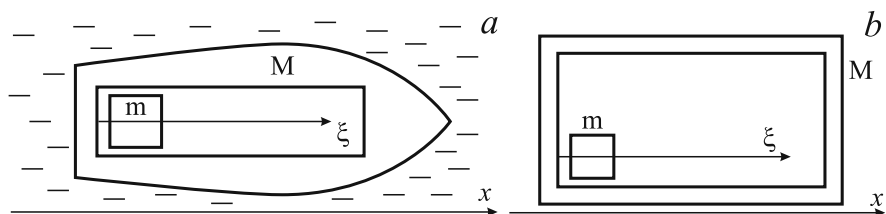


Fig. 7.1 Mechanical models

For the anisotropic linear resistance (Fig. 7.1a), the function $r(v)$ is given by

$$r(v) = k_+ v, \quad \text{if } v \geq 0; \quad r(v) = k_- v, \quad \text{if } v < 0. \quad (7.3)$$

Similarly, for the anisotropic quadratic resistance, this function has the form

$$r(v) = \kappa_+ |v|v, \quad \text{if } v \geq 0; \quad r(v) = \kappa_- |v|v, \quad \text{if } v < 0. \quad (7.4)$$

In Eqs. (7.3) and (7.4), k_+ , k_- , κ_+ , and κ_- are positive coefficients. For the isotropic case, $k_+ = k_-$ and $\kappa_+ = \kappa_-$.

For the case of anisotropic Coulomb's friction (Fig. 7.1b), the function $r(v)$ is given by

$$r(v) = f_+ g, \quad \text{if } v > 0; \quad r(v) = -f_- g, \quad \text{if } v < 0, \quad (7.5)$$

where g is the acceleration due to gravity, f_+ and f_- are coefficients of friction that can be different for onward and backward motions. If the inequalities

$$-f_+ g \leq \mu w \leq f_- g \quad (7.6)$$

hold and body M is at rest ($v = 0$), then it will stay at rest.

In what follows, the motion of mass m relative to body M is supposed to be periodic with a period T and bounded within a fixed interval:

$$0 \leq \xi(t) \leq L, \quad (7.7)$$

where $L > 0$ is given. Without loss of generality, it is assumed that at the beginning and at the end of the period mass m is at the left end of the interval, so that

$$\xi(0) = \xi(T) = 0, \quad u(0) = u(T) = 0. \quad (7.8)$$

The maximal admissible displacement $\xi(\theta) = L$ is reached at some instant $\theta \in (0, T)$.

The motion of the system is controlled by the relative motion of mass m , i.e., by functions $\xi(t)$, $u(t)$, and $w(t)$ subject to Eqs. (7.1) and conditions (7.7) and (7.8).

We will find relative motions of mass m such that the velocity $v(t)$ of body M is T -periodic, and the average velocity of the system $V = \Delta x / T$, where $\Delta x = x(T) - x(0)$, is maximal.

The periodicity of $v(t)$ implies that

$$v(0) = v(T) = v_0, \quad (7.9)$$

where v_0 is a constant. Below we will consider two cases: either v_0 is free and can be chosen or it is fixed and equal to zero ($v_0 = 0$).

7.3 Linear Resistance

Note that the anisotropic resistance (7.3) is, in fact, nonlinear, if $k_+ \neq k_-$. In the case of the linear resistance, we have $k_+ = k_- = k$. Let us substitute $r(v)$ from Eq. (7.3) and w from Eq. (7.1) into Eq. (7.2) and integrate the resulting equation to obtain

$$v(T) - v(0) = -\mu[u(T) - u(0)] - k[x(T) - x(0)].$$

Since $u(t)$ and $v(t)$ should be T -periodic, it follows from this equation that $x(T) = x(0)$ and, therefore, $V = 0$.

Hence, in the case of the isotropic linear resistance and for an arbitrary periodic relative motion of mass m , the system cannot move progressively and will only oscillate about some mean position.

7.4 Relative Motions

Let us confine ourselves to two simple classes of periodic relative motions of mass m ; these classes will be called *velocity-control* and *acceleration-control motions*. In *velocity-control*, or *two-phase motion*, the relative velocity $u(t)$ of mass m is regarded as a bounded piecewise constant control, and there are two intervals of constant velocity (Fig. 7.2). In *acceleration-control*, or *three-phase motion*, the relative acceleration $w(t)$ of mass m is regarded as a bounded piecewise constant control, and there are three intervals of constant acceleration (Fig. 7.3). These velocity-control and acceleration-control motions have, under the imposed periodicity conditions, the least possible number of intervals where the respective control (u or w) is constant.

Denote by τ_i the durations of intervals where the control is constant. For the velocity-control motion, we have

$$u(t) = u_1 \quad \text{for } t \in (0, \tau_1), \quad u(t) = u_2 \quad \text{for } t \in (\tau_1, T), \quad T = \tau_1 + \tau_2, \quad (7.10)$$

where u_1 and u_2 are positive constants.

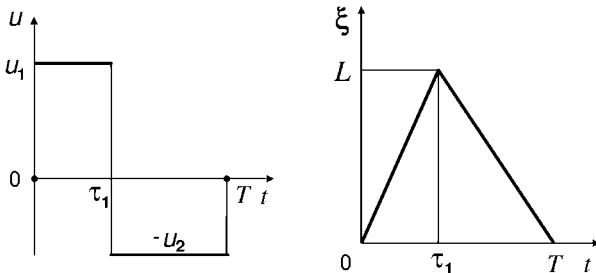


Fig. 7.2 Velocity-control motion

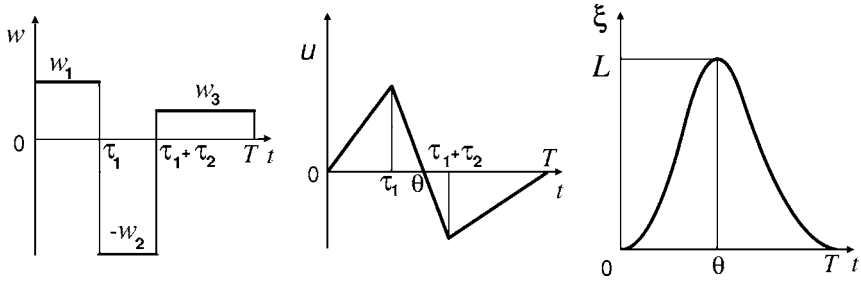


Fig. 7.3 Acceleration-control motion

Note that function $u(t)$ has jumps at $t = 0$ and $t = T$. For convenience and without loss of generality, we define $u(0) = u(T) = 0$ at these instants in accordance with Eq. (7.8).

The relative acceleration $w = \dot{u}$ of mass m for the velocity-control motion (7.10) is given by

$$w(t) = u_1 \delta(t) - (u_1 + u_2) \delta(t - \tau_1) + u_2 \delta(t - T), \quad (7.11)$$

where $\delta(t)$ is Dirac's delta function.

The velocity-control motion is determined by two parameters, u_1 and u_2 , and all other parameters are expressed in terms of u_1 and u_2 as follows:

$$\tau_1 = \theta = L/u_1, \quad \tau_2 = L/u_2, \quad T = L(u_1^{-1} + u_2^{-1}). \quad (7.12)$$

For the acceleration-control motion, we have

$$\begin{aligned} w(t) &= w_1 \quad \text{for } t \in (0, \tau_1), \quad w(t) = -w_2 \quad \text{for } t \in (\tau_1, \tau_1 + \tau_2), \\ w(t) &= w_3 \quad \text{for } t \in (\tau_1 + \tau_2, T), \quad T = \tau_1 + \tau_2 + \tau_3, \end{aligned} \quad (7.13)$$

where w_1, w_2 , and w_3 are positive constants. All other parameters are expressed in terms of w_1, w_2 , and w_3 as follows [5]:

$$\begin{aligned} \tau_1 &= \sqrt{\frac{2Lw_2}{w_1(w_1 + w_2)}}, \quad \tau_2 = \sqrt{\frac{2L}{w_2} \left[\left(\frac{w_1}{w_1 + w_2} \right)^{1/2} + \left(\frac{w_3}{w_2 + w_3} \right)^{1/2} \right]}, \\ \tau_3 &= \sqrt{\frac{2Lw_2}{w_3(w_2 + w_3)}}, \quad T = \sqrt{\frac{2L}{w_2} \left[\left(\frac{w_1 + w_2}{w_1} \right)^{1/2} + \left(\frac{w_2 + w_3}{w_3} \right)^{1/2} \right]}. \end{aligned} \quad (7.14)$$

The control parameters introduced above are subjected to constraints

$$0 < u_i \leq U, \quad i = 1, 2, \quad (7.15)$$

imposed on the relative velocity of the velocity-control motion, and constraints

$$0 < w_i \leq W, \quad i = 1, 2, 3, \quad (7.16)$$

imposed on the relative acceleration of the acceleration-control motion. Here, U and W are the maximal admissible velocity and acceleration of the relative motion, respectively.

The assumptions related to the velocity-control and acceleration-control motions correspond to different properties of the drives that can be used to ensure the relative movement of internal mass.

Let us now find optimal velocity-control and acceleration-control motions for resistance forces satisfying Eqs. (7.3), (7.4) and (7.5).

7.5 Piecewise Linear Resistance

First, we consider the case of the velocity-control motion for the piecewise linear resistance force described by Eq. (7.3). Let us substitute Eqs. (7.3) and (7.11) into Eq. (7.2) and integrate the resulting equation for $v(t)$ under the initial condition $v(0) = v_0$. We choose the parameter v_0 in Eq. (7.9) so that the obtained solution $v(t)$ is T -periodic. We have [6]

$$\begin{aligned} v(t) &= -\frac{\mu(u_1 + u_2)(1 - e_2)}{1 - e_1 e_2} \exp(-k_- t) \quad \text{for } t \in (0, \tau_1); \\ v(t) &= \frac{\mu(u_1 + u_2)e_2(1 - e_1)}{1 - e_1 e_2} \exp[-k_+(T - t)] \quad \text{for } t \in (\tau_1, T); \\ v_0 &= \frac{\mu[u_1 e_2(1 - e_1) - u_2(1 - e_2)]}{1 - e_1 e_2}, \quad e_1 = \exp(-k_- \tau_1), \quad e_2 = \exp(-k_+ \tau_2), \end{aligned} \quad (7.17)$$

where the parameters u_1, u_2, τ_1, τ_2 , and T satisfy Eq. (7.12). The function $v(t)$ from Eq. (7.17) is shown in Fig. 7.4.

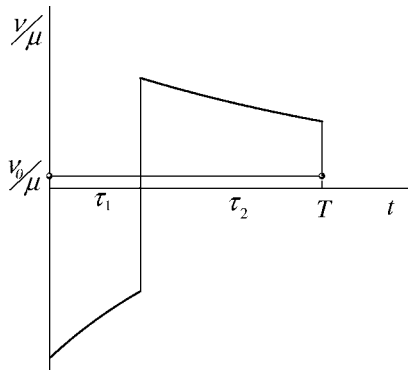


Fig. 7.4 Velocity $v(t)$ for piecewise linear resistance

To calculate the total displacement Δx , we integrate $v(t)$ from Eq. (7.17) over the period $[0, T]$. Divide the resulting expression by T and use Eq. (7.12) to obtain

$$V = \frac{\Delta x}{T} = \frac{\mu L(1 - e_1)(1 - e_2)(k_+^{-1} - k_-^{-1})}{(1 - e_1 e_2) \tau_1 \tau_2}. \quad (7.18)$$

Hence, $V > 0$ only if $k_+ < k_-$, which is physically quite natural.

For given μ , L , k_+ , and k_- , the average speed V from Eq. (7.18) depends on two parameters τ_1 and τ_2 or u_1 and u_2 . The maximization of V with respect to these parameters subject to constraint (7.15) provides

$$\begin{aligned} u_1 = u_2 = U, \quad \tau_1 = \tau_2 = L/U, \quad T = 2L/U, \\ V_{\max} = \frac{\mu U^2 L^{-1} (1 - e_1)(1 - e_2)(k_+^{-1} - k_-^{-1})}{1 - e_1 e_2}, \\ e_1 = \exp(-k_- L/U), \quad e_2 = \exp(-k_+ L/U). \end{aligned} \quad (7.19)$$

Thus, the optimal relative motion of mass m in this case is the motion with the maximal admissible speed U first from the point $\xi = 0$ to the point $\xi = L$, and then back, from $\xi = L$ to $\xi = 0$, with the same speed.

Formula (7.19) for V_{\max} is simplified, of $k_- L \ll U$. In this case we have $k_+ L \ll U$ and

$$V_{\max} = \mu U (k_- - k_+) / (k_- + k_+).$$

7.6 Quadratic Resistance

Consider now the velocity-control motion for the quadratic resistance described by Eq. (7.4). To simplify our equations, we restrict ourselves to the case of the isotropic quadratic resistance and assume that

$$\kappa_+ = \kappa_- = \kappa, \quad \mu \kappa L < 1, \quad v_0 = 0 \quad (7.20)$$

in Eqs. (7.4) and (7.9). Assumptions (7.20) imply the following relationship between parameters u_1 and u_2 :

$$u_2 = (1 - \mu \kappa L)(1 + \mu \kappa L)^{-1} u_1. \quad (7.21)$$

Under the assumptions made, the desired periodic velocity $v(t)$ of body M becomes

$$\begin{aligned} v(t) &= -\frac{\mu u_1}{1 + \mu u_1 \kappa t} \quad \text{for } t \in (0, \tau_1), \\ v(t) &= \frac{\mu u_1}{1 + \mu \kappa L + \mu u_1 \kappa (t - \tau_1)} \quad \text{for } t \in (\tau_1, T). \end{aligned} \quad (7.22)$$

By integrating velocity $v(t)$ from Eq. (7.22) over a period $[0, T]$, we find the total displacement Δx of body M over the period and the average speed as follows:

$$V = \Delta x / T = -(\kappa T)^{-1} \log(1 - \mu^2 \kappa^2 L^2) > 0. \quad (7.23)$$

Here, the period T can be expressed through parameters u_1 and u_2 by means of Eq. (7.12). Taking into account also relationship (7.21), we find that the maximal value of V in Eq. (7.23) over parameters u_1 and u_2 under the constraint (7.15) is attained, if

$$u_1 = U, \quad u_2 = (1 - \mu \kappa L)(1 + \mu \kappa L)^{-1}U,$$

and is equal to

$$V_{\max} = -U(2\kappa L)^{-1}(1 - \mu \kappa L) \log(1 - \mu^2 \kappa^2 L^2) > 0. \quad (7.24)$$

Note the substantial difference between piecewise linear and quadratic resistance. For the linear resistance the progressive motion is possible only in the anisotropic case, whereas for the quadratic resistance it is possible also in the isotropic case. For both types of resistance forces, we have $V_{\max} \rightarrow \infty$ as $U \rightarrow \infty$, see Eqs. (7.19) and (7.24).

7.7 Dry Friction: Velocity-Control Motion

Consider now the case of Coulomb's dry friction defined by Eq. (7.5). In this section, we will discuss the velocity-control motions described by Eqs. (7.10), (7.11) and (7.12) for two cases mentioned at the end of Sect. 7.2, namely, for the case of $v_0 = 0$ in Eq. (7.9) and the case where v_0 is to be chosen in order to maximize the average speed V .

Let us introduce the following notation:

$$\begin{aligned} u_0 &= (Lf_-g/\mu)^{1/2}, \quad u_i = u_0 x_i, \quad i = 1, 2; \quad U = u_0 X, \\ V &= \mu u_0 F, \quad c = f_+/f_-. \end{aligned} \quad (7.25)$$

Here, x_i , X , and F are non-dimensional parameters. Inequality (7.15) takes the form

$$0 < x_i \leq X, \quad i = 1, 2. \quad (7.26)$$

Consider first the case of $v_0 = 0$ that has been examined in [5]. It occurs that in this case two modes, a and b , shown in Fig. 7.5 are possible for body M . In mode a , body M is never in the state of rest ($v \neq 0$), whereas in mode b there is an interval of rest where $v = 0$.

Let us first assume that there is no upper bound on the velocity u of mass m , so that $U \rightarrow \infty$ in Eq. (7.15) and, therefore, $X \rightarrow \infty$ in Eq. (7.26). Then the maximal average speed V is attained in mode a , if $c \leq 1$, and mode b , if $c > 1$. The optimal motion is determined by the relationships [5]:

$$\begin{aligned} x_1 &= 1, \quad x_2 = c, \quad F = 1/2, \quad \text{if } c \leq 1; \\ x_1 &= x^*(c), \quad x_2 = c/x^*(c), \quad F = F^0(x^*(c), c), \quad \text{if } c > 1, \end{aligned} \quad (7.27)$$

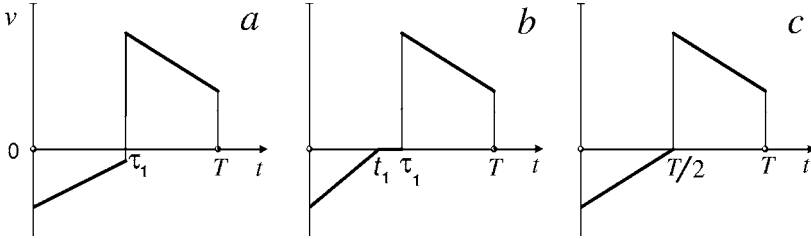


Fig. 7.5 Velocity-control modes (dry friction, $v_0 = 0$)

where the following notation is introduced:

$$x^*(c) = \left\{ (c/2)(c-1)^{-1} [1 - 3c + (9c^2 + 2c - 7)^{1/2}] \right\}^{1/2}, \quad (7.28)$$

$$F^0(x, c) = (x/2)(c + x^2)^{-1} [2c + x^2(1 - c)].$$

Note that here the average speed $V = \mu u_0 F$ is finite as $U \rightarrow \infty$, contrary to the cases of piecewise linear and quadratic resistance, see Eqs. (7.19) and (7.24).

If the friction is isotropic ($f_+ = f_- = f$, $c = 1$), we have, according to Eqs. (7.25), (7.27), and (7.12),

$$x_1 = x_2 = 1, \quad u_1 = u_2 = u_0 = (Lfg/\mu)^{1/2}, \quad F = 1/2,$$

$$\tau_1 = \tau_2 = (\mu L/fg)^{1/2}, \quad T = 2\tau_1, \quad V = (1/2)(\mu Lfg)^{1/2}.$$

This case is illustrated by Fig. 7.5c.

For the general case of finite U in Eq. (7.15), the velocity-control motions are implementable, if $X \geq \max(c^{1/2}, c)$. The optimal motion is described by the following relationships obtained in [5]:

$$\begin{aligned} (1) \quad & x_1 = X, \quad x_2 = c/X, \quad F = F^0(X, c), \quad \text{if } c^{1/2} \leq X \leq 1; \\ (2) \quad & x_1 = 1, \quad x_2 = c, \quad F = 1/2, \quad \text{if } c \leq 1 \text{ and } X \geq 1; \\ (3) \quad & x_1 = c/X, \quad x_2 = X, \quad F = F^0(c/X, c), \quad \text{if } 1 < c \leq X \leq c/x^*(c); \\ (4) \quad & x_1 = x^*(c), \quad x_2 = c/x^*(c), \quad F = F^0(x^*(c), c), \quad \text{if } c/x^*(c) \leq X. \end{aligned} \quad (7.29)$$

This optimal solution is illustrated by Fig. 7.6 where graphs of functions $X = c^{1/2}$ and $X = c/x^*(c)$ are indicated by the letters K and N , respectively. Together with the straight lines $c = 1$, $X = 1$, and $X = c$, they divide the region $X \geq \max(c^{1/2}, c)$, where the solution exists, into four regions indicated by the numbers 1, 2, 3, and 4. These numbers correspond to the four possibilities in Eq. (7.29). In Fig. 7.6, the motion of mode a from Fig. 7.5 occurs in region 2, and mode b occurs in the remaining regions 1, 3, and 4.

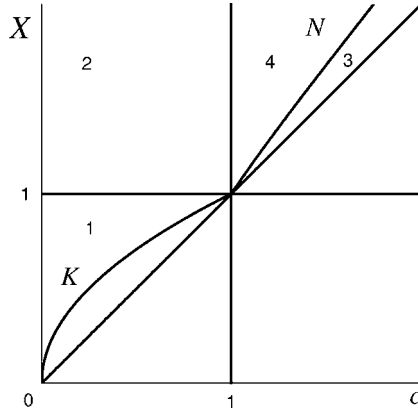


Fig. 7.6 Optimal velocity-control motion (dry friction, $v_0 = 0$)

The optimal solution for $v_0 = 0$ is completely defined by Eq. (7.27), (7.28) and (7.29) in the non-dimensional form. Using Eq. (7.25), one can return to the original dimensional variables.

Let us now consider the case where v_0 is a free parameter to be chosen.

According to Eqs. (7.2), (7.5), and (7.10), the velocity $v(t)$ of body M has two jumps at the ends of the period $[0, T]$ and one jump at the instant $t = \tau_1 \in (0, T)$ inside the period. Between these jumps, body M is subjected only to the constant friction force. The absolute value of its velocity here either decreases linearly in time or is equal to zero, if condition (7.6) is satisfied. Hence, the period $[0, T]$ can include not more than two intervals of rest where $v = 0$, one of these intervals can be placed before the instant $t = \tau_1$ and the other before $t = T$. Thus, the four modes shown in Fig. 7.7 are possible:

- A — no intervals of rest,
- B — one interval of rest (t_1, τ_1) ,
- C — one interval of rest (t_2, T) ,
- D — two intervals of rest (t_1, τ_1) and (t_2, T) .

Here, $0 \leq t_1 \leq \tau_1$ and $\tau_1 \leq t_2 \leq T$.

Let us first consider mode A. Using Eqs. (7.2), (7.5), and (7.10), and also the initial condition $v(0) = v_0$, we calculate successively

$$\begin{aligned}
 v(\tau_1 - 0) &= v_0 - \mu u_1 + f_- g \tau_1, \\
 v(t_1 + 0) &= v(\tau_1 - 0) + \mu(u_1 + u_2) = v_0 + \mu u_2 + f_- g \tau_1, \\
 v(T - 0) &= v(t_1 + 0) - f_+ g \tau_2 = v_0 + \mu u_2 + f_- g \tau_1 - f_+ g \tau_2, \\
 v(T) &= v(T - 0) - \mu u_2 = v_0 + f_- g \tau_1 - f_+ g \tau_2.
 \end{aligned} \tag{7.30}$$

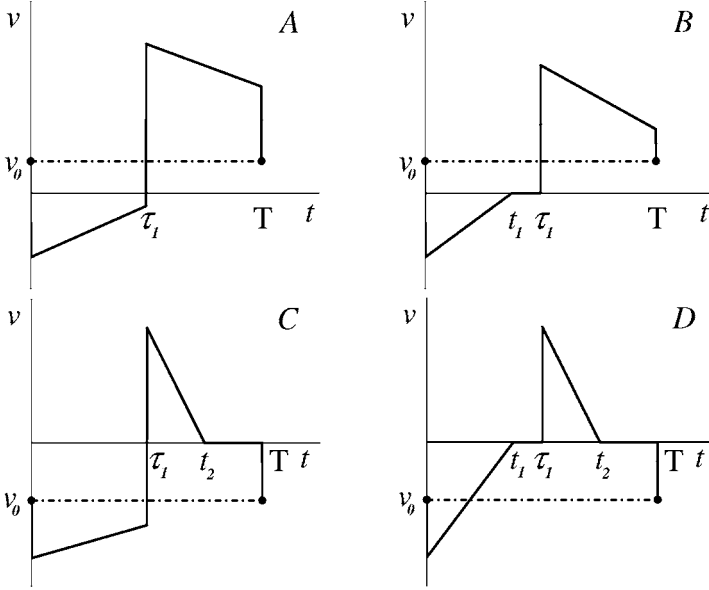


Fig. 7.7 Velocity-control modes (dry friction, $v_0 \neq 0$)

It follows from Eqs. (7.30) and the periodicity condition $v(T) = v_0$ that $f_- \tau_1 = f_+ \tau_2$. Taking into account relations (7.12) and (7.25), we obtain

$$cu_1 = u_2. \quad (7.31)$$

For mode A, inequalities $v(\tau_1 - 0) \leq 0$ and $v(T - 0) \geq 0$ must hold. These inequalities, together with Eqs. (7.30) and (7.31), imply

$$-\mu cu_1 \leq v_0 \leq \mu u_1 - f_- g \tau_1. \quad (7.32)$$

We will use the non-dimensional variables introduced in Eq. (7.25) and denote

$$v_0 = \mu u_0 x_0. \quad (7.33)$$

Let us express τ_1 by means of Eq. (7.12) and rewrite inequality (7.32) as follows:

$$-cx_1 \leq x_0 \leq x_1 - x_1^{-1}. \quad (7.34)$$

The left-hand side of inequality (7.34) should not exceed its right-hand side, hence

$$x_1 \geq (1 + c)^{-1/2}, \quad cx_1 = x_2. \quad (7.35)$$

The second equality (7.35) follows from Eqs. (7.25) and (7.31).

Thus, non-dimensional parameters x_0, x_1 , and x_2 of mode A must satisfy conditions (7.34) and (7.35). Integrate the function $v(t)$ from Fig. 7.7 over the interval

$[0, T]$ for mode A to evaluate the distance $\Delta x = x(T) - x(0)$ and the average speed $V = \Delta x/T$ of body M . After certain calculations using notation (7.25) and (7.33), we find

$$F = x_0 + (2x_1)^{-1}. \quad (7.36)$$

Modes B – D are analyzed analogously to mode A. For each of the modes, three conditions are obtained — one equality and two inequalities imposed on three parameters x_0, x_1 , and x_2 , and also the expression for the non-dimensional average speed F . The respective relations, similar to Eqs. (7.34), (7.35) and (7.36), have the form

$$\begin{aligned} B : x_0 &= x_1 - cx_2^{-1}, \quad cx_1 \leq x_2, \quad (x_1 + x_2)x_2 \geq c, \\ F &= x_1 - \frac{c(c+1)x_1}{2x_2(x_1 + x_2)}; \\ C : x_0 &= -x_2, \quad cx_1 \geq x_2, \quad (x_1 + x_2)x_1 \geq 1, \\ F &= \frac{(1+c)x_2}{2c(x_1 + x_2)x_1} - x_2; \\ D : x_0 &= -x_2, \quad (x_1 + x_2)x_1 \leq 1, \quad (x_1 + x_2)x_2 \leq c, \\ F &= \frac{(1-c)(x_1 + x_2)x_1x_2}{2c}. \end{aligned} \quad (7.37)$$

Modes A – D take place in the respective domains (7.37) in the plane of parameters $x_1 > 0, x_2 > 0$. These domains are shown in Fig. 7.8 for $c > 1$. According to Eq. (7.35), mode A occurs on the ray which is the boundary between domains B and C . The boundaries between domain D and domains B and C are the arcs of the respective hyperbolas $(x_1 + x_2)x_2 = c$, $x_1 \leq (1+c)^{-1/2}$ and $(x_1 + x_2)x_1 = 1$,

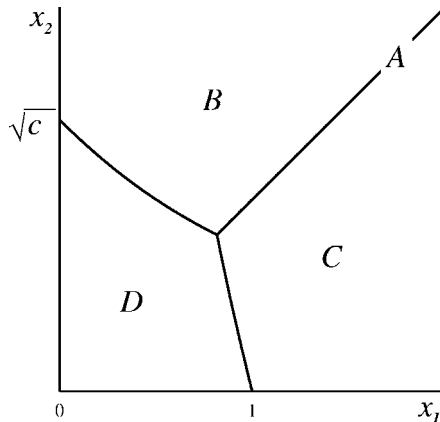


Fig. 7.8 Domains in x_1, x_2 -plane (dry friction, $v_0 \neq 0$)

$x_1 \geq (1+c)^{-1/2}$. These arcs and the ray $x_2 = cx_1$ meet at the point with the coordinates (Fig. 7.8)

$$x_1 = (1+c)^{-1/2}, \quad x_2 = c(1+c)^{-1/2}.$$

Thus, for each pair of non-dimensional parameters x_1, x_2 or dimensional ones u_1, u_2 or τ_1, τ_2 , see Eqs. (7.25) and (7.12), one can determine the respective mode of motion using Eqs. (7.35) and (7.37) or Fig. 7.8. Moreover, for modes B – D , one can also find the non-dimensional average speed F by means of Eqs. (7.37). As for mode A , one should first choose x_0 according to inequalities (7.34) and then evaluate F by means of Eq. (7.36). The values of the dimensional velocities u_1, u_2, v_0 , and V are determined by Eqs. (7.25) and (7.33).

Let us determine the optimal values of the parameters u_1, u_2 , and v_0 that correspond to the maximal possible average speed of body M under the constraints (7.15). In terms of the non-dimensional parameters (7.25) and (7.33), the optimization problem is stated as follows: find the parameters x_0, x_1 , and x_2 that correspond to the maximal value of F under constraints (7.26).

The function F is defined by equations (7.36) and (7.37) in the respective domains A – D .

Consider the behavior of the function F in domains A – D . Note that this function grows monotonically with x_0 in domain A , see Eq. (7.36). Hence, the optimal value of x_0 is given by the upper bound in Eq. (7.34) and, therefore,

$$x_0 = x_1 - x_1^{-1}, \quad F = x_1 - (2x_1)^{-1} \quad (7.38)$$

for mode A . The function F from Eq. (7.38) increases with x_1 . Therefore, the required maximum of F can be reached in domain A only at the maximal possible x_1 satisfying inequalities (7.26).

In domain B , the function F increases with x_2 . Hence, its maximum can be attained in B only at the maximal possible x_2 compatible with conditions (7.26). Furthermore, we have $\partial^2 F / \partial x_1^2 > 0$ for all $x_1 > 0, x_2 > 0$. Thus, F is a convex function of x_1 , and its maximum can be reached only at the ends of the permissible interval of parameter x_1 .

For mode C , the function F , according to Eq. (7.37), decreases monotonically with x_1 . Hence, its maximum is never reached within domain C and can be attained only on its boundaries with domains A and D , see Fig. 7.8.

In domain D , the function F increases with x_1 and x_2 , if $c < 1$. If $c > 1$, this function is negative and decreases with x_1 and x_2 in D .

Let us summarize our observations and find the required maximum of F over x_1 and x_2 subject to constraints (7.26).

If $c \leq 1$ and the point $x_1 = X, x_2 = X$ lies within domain D (it happens, if $X \leq (c/2)^{1/2}$), the required maximum of the function F is reached at this point. If this point is outside D , then it lies in domain B , and the maximum can be reached either at the same point or at the intersection of the line $x_2 = X$ with the boundary between domains B and D . A comparison of the respective values of F defined by Eq. (7.37) leads us to the conclusion that this maximum is always attained at $x_1 = x_2 = X$.

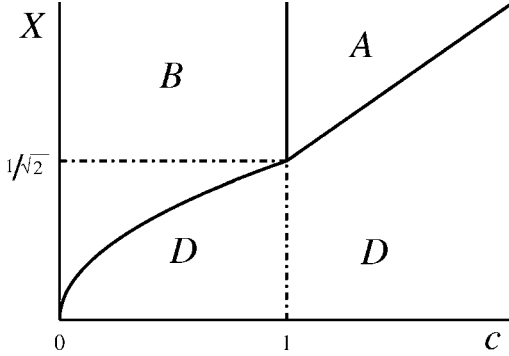


Fig. 7.9 Optimal velocity-control motion (dry friction, $v_0 \neq 0$)

If $c \geq 1$ and $X < 2^{-1/2}c$, then the function F is always negative. Its zero upper bound is approached, if $x_1 \rightarrow 0$ or $x_2 \rightarrow 0$. If $c \geq 1$ and $X \geq 2^{-1/2}c$, then the required maximum is attained in domain A at $x_1 = X/c$, $x_2 = X$.

The results obtained are presented in Fig. 7.9 and by the formulas

$$\begin{aligned}
 &c \leq 1, X \leq (c/2)^{1/2} : (x_1, x_2) \in D, \\
 &x_1 = x_2 = X, \quad x_0 = -X, \quad F = (1 - c)X^3/c; \\
 &c \leq 1, X \geq (c/2)^{1/2} : (x_1, x_2) \in B, \\
 &x_1 = x_2 = X, \quad x_0 = -X, \quad F = X - c(c + 1)(4X)^{-1}; \\
 &c \geq 1, X < c/2^{1/2} : (x_1, x_2) \in D, \\
 &x_1 \rightarrow 0 \quad \text{or} \quad x_2 \rightarrow 0, \quad x_0 \rightarrow 0, \quad F \rightarrow 0; \\
 &c \geq 1, X \geq c : (x_1, x_2) \in A, \\
 &x_1 = X/c, \quad x_2 = X, \quad x_0 = (X^2 - c^2)/(cX), \quad F = (2X^2 - c^2)/(2cX).
 \end{aligned} \tag{7.39}$$

We can return to the original dimensional variables in Eq. (7.39) using the notation (7.25) and (7.33).

We have considered two cases for the velocity-control motion in the presence of Coulomb's friction: the case where $v_0 = 0$ in Eq. (7.9) and the case where v_0 was chosen in the optimal way. Of course, the maximal value of the speed V is higher in the second case.

7.8 Dry Friction: Acceleration-Control Motion

The acceleration-control motions for the case of Coulomb's friction and for $v_0 = 0$ have been analyzed in [5]. These motions are described by eqs. (7.1), (7.2), and

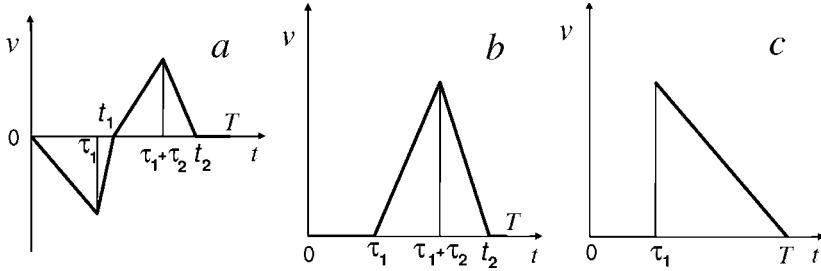


Fig. 7.10 Acceleration-control modes (dry friction, $v_0 = 0$)

(7.5). The acceleration of the internal mass m is defined by Eq. (7.13) and subjected to constraints (7.16).

The analysis shows that two modes of motion of body M , namely, modes a and b in Fig. 7.10, can occur here. Mode b contains an interval of rest of body M but does not include its backward motion, whereas mode a contains both intervals of rest and backward motion of body M . It has been shown [5] that mode b corresponds to a higher maximal average speed V of the system than mode a . Hence, we can restrict ourselves to mode b only.

To define the optimal acceleration of mass m , it is sufficient to determine three parameters w_1 , w_2 , and w_3 subject to constraints (7.16).

Let us introduce the non-dimensional variables according to the following notation:

$$\begin{aligned} w_i &= f - gy_i / \mu, \quad i = 1, 2, 3; \quad W = f - gY / \mu, \\ V &= (Lf - g\mu/2)\Phi. \end{aligned} \quad (7.40)$$

In the non-dimensional variables (7.40), constraints (7.16) take the form

$$0 < y_i \leq Y, \quad i = 1, 2, 3. \quad (7.41)$$

The optimization problem is reduced to the following one: find the values of parameters y_1, y_2 , and y_3 that satisfy constraints (7.41) and maximize the non-dimensional average velocity Φ specified by the relationship [5]

$$\Phi(y) = \frac{(y_1 y_3)^{1/2} (y_2 - c) [y_1^{1/2} (y_2 + y_3)^{1/2} + y_3^{1/2} (y_1 + y_2)^{1/2}]}{y_2^{1/2} (y_1 + y_2) (y_3 + c)}.$$

As a result of the lengthy analysis presented in [5], we come to the following results of optimization:

$$1) y_1 = y_3 = 1, \quad y_2 = Y, \quad \Phi = \frac{2(Y - c)}{(1 + c)Y^{1/2}(Y + 1)^{1/2}}$$

for $c \geq 1$, $Y > c$ and for $1/3 < c < 1$, $1 < Y < Y^*(c)$;

$$2) y_1 = y_2 = y_3 = Y, \quad \Phi = (2Y)^{1/2} \frac{Y-c}{Y+c}$$

for $c \leq 1/3$, $c < Y \leq Y^*(c)$ and for $c > 1/3$, $c < Y < 1$; (7.42)

$$3) y_1 = 1, \quad y_2 = Y, \quad y_3 = \frac{c^2(Y+1)}{Y-2c-c^2}, \quad \Phi = \frac{Y^{1/2}}{(Y+1)^{1/2}}$$

for $Y \geq 1$, $Y \geq Y^*(c)$;

$$4) y_1 = y_2 = Y, \quad y_3 = \frac{2c^2Y}{Y^2-2cY-c^2}, \quad \Phi = (Y/2)^{1/2}$$

for $3c \leq Y < 1$.

Here, the following denotation is used:

$$Y^*(c) = c[c+1+(c^2+2c+2)^{1/2}]. \quad (7.43)$$

The regions in the c, Y -plane corresponding to cases 1–4 in Eq. (7.42) are shown in Fig. 7.11 and marked by the respective numbers 1–4. Note that Figs. 7.11a and b correspond to different scales. Here, the straight lines $c = 1$, $Y = 1$, and $Y = c$, the segments of straight lines $c = 1/3$ and $Y = 3c$ for $Y \leq 1$, and the curve $Y = Y^*(c)$ defined by Eq. (7.43) are shown.

The solution does not exist below the line $Y = c$, i.e., for $Y < c$, and also in a very narrow domain S_1 defined by the inequalities

$$Y^*(c) < Y < 3c, \quad c \in (0, 1/3).$$

Domain S_1 can be seen in Fig. 7.11b.

Note that if constraint (7.16) is absent, i.e., if $W \rightarrow \infty$ in Eq. (7.16) and $Y \rightarrow \infty$ in Eq. (7.41), then case 3 from Eq. (7.42) occurs, and we have

$$y_1 = 1, \quad y_2 \rightarrow \infty, \quad y_3 = c^2, \quad \Phi = 1.$$

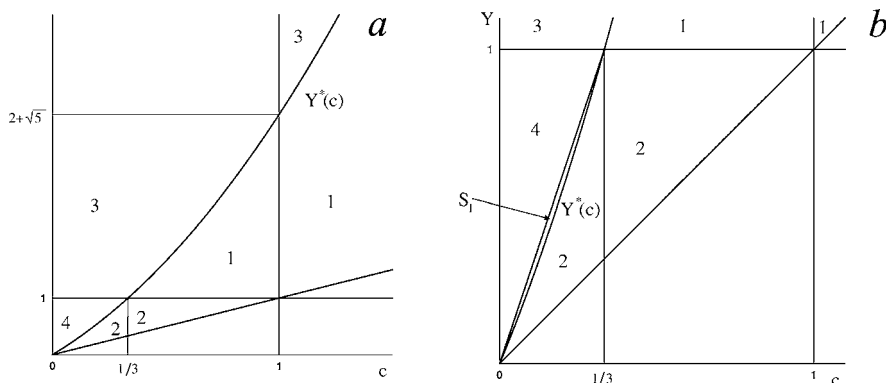


Fig. 7.11 Optimal acceleration-control motion (dry friction, $v_0 = 0$)

Here, the second interval in Eq. (7.13) degenerates into a jump of the velocity, and mode b shown in Fig. 7.10*b* is transformed into Fig. 7.10*c*.

Let us consider the case of the isotropic dry friction and set $c = 1$. Then relations (7.42) take the form

$$\begin{aligned} y_1 = y_3 = 1, \quad y_2 = Y, \quad \Phi &= \frac{Y-1}{Y^{1/2}(Y+1)^{1/2}} \\ \text{for } 1 < Y \leq 2 + \sqrt{5}; \\ y_1 = 1, \quad y_2 = Y, \quad y_3 &= \frac{Y+1}{Y-3}, \quad \Phi = \left(\frac{Y}{Y+1} \right)^{1/2} \\ \text{for } Y > 2 + \sqrt{5}. \end{aligned} \quad (7.44)$$

The transition from the non-dimensional variables to the original dimensional ones in Eqs. (7.42) and (7.44) can be accomplished using formulas (7.40). The time intervals, τ_i , $i = 1, 2, 3$, are defined by Eq. (7.14).

Let us note certain characteristic features of the optimal motions obtained in Sects. 7.7 and 7.8 for the velocity-control and acceleration-control motions with $v_0 = 0$ in the case of Coulomb's friction. If there are no velocity constraints (7.15), the maximal average velocity of the system for the velocity-control case is evaluated by Eqs. (7.25) and (7.29) as follows:

$$V \sim (\mu L f_{-g})^{1/2}/2.$$

The maximal average velocity for the acceleration-control case is determined by Eqs. (7.40) and (7.42):

$$V \sim (\mu L f_{-g}/2)^{1/2}.$$

Hence, the average velocities are of the same order and differ only in the coefficients: the velocity V is greater for the acceleration-control motion.

In both cases, these velocities are bounded, even if the upper limits for the relative velocity u or relative acceleration w of mass m tend to infinity ($U \rightarrow \infty$ or $W \rightarrow \infty$).

Furthermore, despite the significant differences in the assumptions made in these cases, there is a remarkable similarity between the corresponding optimal modes. Comparing Figs. 7.7*c* and 7.10*c*, we see that the optimal three-phase acceleration-control mode degenerates into the two-phase mode, in which velocity $v(t)$ of body M undergoes a jump between phases.

When constraints (7.15) are imposed on the velocity of mass m , the velocity-control motion is possible only if the non-dimensional parameter X is bounded from below by inequality $X \geq \max(c^{1/2}, c)$. Similarly, when constraints (7.16) are imposed on the acceleration of mass m , the acceleration-control motion is possible, only if the non-dimensional parameter Y is bounded from below ($Y > c$).

7.9 Generalizations

Other problems of optimal periodic motions of multibody systems in resistive media have been considered in [1, 3, 7, 8]. Optimal motions of a two-mass system have been analyzed in [3] in the case where both masses interact with the horizontal plane by means of Coulomb's friction forces.

The problem of optimal control for a rigid body containing a moving internal mass in the presence of isotropic Coulomb's friction between the body and the horizontal plane has been considered in [8]. The acceleration $w(t)$ of the internal mass has been subjected to the constraint $|w(t)| \leq W$, and Pontryagin's maximum principle has been applied. The obtained optimal acceleration occurs to be piecewise constant with three intervals of constancy but, by contrast to the three-phase motion considered above, the instants when $\xi(t) = 0$, $u(t) = 0$, and $v(t) = 0$ do not coincide.

The case of one or more internal masses moving in two directions (horizontally and vertically) in the vertical plane inside body M has been considered in [1, 7]. Due to the vertical motion of the internal mass, the normal reaction force exerted by body M upon the horizontal plane changes. Hence, Coulomb's friction force changes too. This force can be decreased during the forward motion of body M and increased during its backward motion and the state of rest. Thus, an additional increment of the average speed of the system can be attained. This phenomena has been analyzed and the corresponding optimal parameters of the motion have been determined [1].

7.10 Experiments

The principle of motion described above has been implemented in experimental models [9–11] shown in Fig. 7.12. Internal motions in these models are performed by an inverted pendulum [10], eccentric rotating wheels, or electromagnetic drives

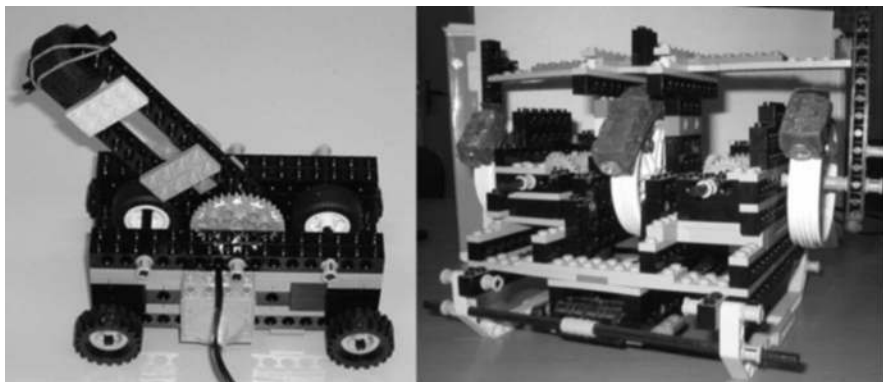


Fig. 7.12 Experimental models



Fig. 7.13 Mini-robot in a tube

[9]. The experiments have shown that the locomotion of mechanical systems controlled by internal moving masses is quite feasible.

Mini-robots that utilize the principle of motion under consideration and can move inside tubes have been designed and tested [9]. These robots move inside straight and curved tubes of diameter 0.5–3cm using friction force between the robot and the tube (Fig. 7.13).

7.11 Conclusions

Progressive motions of a rigid body controlled by internal periodic oscillations of internal masses relative to the body have been analyzed. For certain classes of relative motions, optimal controls have been found that correspond to the maximal average speed of the system as a whole in various resistive media. Experiments confirm the obtained theoretical results. The principle of motion considered in this chapter is of practical use for mobile robots, especially for mini-robots, that can move inside tubes, in corrosive media, and in complex environment.

Acknowledgments The research was supported by the Russian Foundation for Basic Research (Grants 07–01–92109 and 07–01–12015) and by the Program for the Support of Leading Russian Scientific Schools.

References

1. Bolotnik, N.N., Zeidis, I., Zimmermann, K., Yatsun, S.F.: Dynamics of controlled motions of vibration-driven systems. *Journal of Computer and Systems Sciences International* **45**, 831–840 (2006)

2. Breguet, J.-M., Clavel, R.: Stick and slip actuators: design, control, performances and applications. In: Proc. International Symposium on Micromechatronics and Human Science (MHS), pp. 89–95. IEEE, New York (1998)
3. Chernousko, F.L.: The optimum rectilinear motion of a two-mass system. *Journal of Applied Mathematics and Mechanics* **66**, 1–7 (2002)
4. Chernousko, F.L.: On the motion of a body containing a movable internal mass. *Doklady Physics* **50**, 593–597 (2005)
5. Chernousko, F.L.: Analysis and optimization of the motion of a body controlled by means of a movable internal mass. *Journal of Applied Mathematics and Mechanics* **70**, 915–941 (2006)
6. Chernousko, F.L.: Dynamics of a body controlled by internal motions. In: Hu, H.Y., Kreuzer, E. (eds.) Proc. IUTAM Symposium on Dynamics and Control of Nonlinear Systems with Uncertainty, pp. 227–236, Springer, Dordrecht (2007)
7. Chernousko, F.L., Zimmermann, K., Bolotnik, N.N., Yatsun, S.F., Zeidis, I.: Vibration-driven robots. In: Proc. of the Workshop on Adaptive and Intelligent Robots: Present and Future, pp. 26–31. Moscow (2005)
8. Figurina, T.Yu.: Optimal control of the motion of a two-body system along a straight line. *Journal of Computer and Systems Sciences International* **46**, 227–233 (2007)
9. Gradetsky, V., Solovtsov, V., Kniazkov, M., Rizzotto, G.G., Amato, P.: Modular design of electro-magnetic mechatronic microrobots. In: Proc. of the 6th International Conference on Climbing and Walking Robots CLAWAR, pp. 651–658. Catania (2003)
10. Li, H., Furuta, K., Chernousko, F.L.: A pendulum-driven cart via internal force and static friction. In: Proc. of the International Conference “Physics and Control”, pp. 15–17. St.-Petersburg (2005)
11. Li, H., Furuta, K., Chernousko, F.L.: Motion generation of the Capsbot using internal force and static friction. In: Proc. 45th Conference on Decision and Control, pp. 6575–6580. San Diego (2006)
12. Schmoeckel, F., Worn, H.: (2001). Remotely controllable mobile microrobots acting as nano positioners and intelligent tweezers in scanning electron microscopes (SEMs). In: Proc. International Conference on Robotics and Automation, pp. 3903–3913. New York (2001)
13. Vartholomeos, P., Papadopoulos, E.: Dynamics, design and simulation of a novel microrobotic platform employing vibration microactuators. *Trans. ASME. Journal of Dynamic Systems, Measurement, and Control* **128**, 122–133 (2006)

Chapter 8

Instationary Heat-Constrained Trajectory Optimization of a Hypersonic Space Vehicle by ODE–PDE-Constrained Optimal Control

Kurt Chudej, Hans Josef Pesch, Markus Wächter, Gottfried Sachs
and Florent Le Bras

Dedication This contribution is dedicated to Professor Angelo Miele on the occasion of his 85th birthday.

Abstract During ascent and reentry of a hypersonic space vehicle into the atmosphere of any heavenly body, the space vehicle is subjected, among others, to extreme aerothermic loads. Therefore, an efficient, sophisticated and lightweight thermal protection system is determinative for the success of the entire mission. For a deeper understanding of the conductive, convective and radiative heating effects through a thermal protection system, a mathematical model is investigated which is given by an optimal control problem subject to not only the usual dynamic equations of motion and suitable control and state variable inequality constraints but also an instationary quasi-linear heat equation with nonlinear boundary conditions. By this model, the temperature of the heat shield can be limited in certain critical regions. The resulting ODE–PDE-constrained optimal control problem is solved by

Kurt Chudej

Lehrstuhl für Ingenieurmathematik, Universität Bayreuth, Bayreuth, Germany,
e-mail: kurt.chudej@uni-bayreuth.de

Hans Josef Pesch

Lehrstuhl für Ingenieurmathematik, Universität Bayreuth, Bayreuth, Germany,
e-mail: hans-josef.pesch@uni-bayreuth.de

Markus Wächter

German Institute of Science and Technology, Singapore,
e-mail: markus.waechte@gist.edu.sg

Gottfried Sachs

Lehrstuhl für Flugmechanik und Flugregelung, Technische Universität München, München, Germany, e-mail: sachs@lfm.mw.tum.de

Florent Le Bras

Laboratoire de Recherches Balistiques et Aérodynamiques, Délégation Générale pour l'Armement, Vernon, formerly: École Polytechnique Paris, France,
e-mail: florent.le-bras@polytechnique.org

a second-order semi-discretization in space of the quasi-linear parabolic partial differential equation yielding a large-scale nonlinear ODE-constrained optimal control problem with additional state constraints for the heat load. Numerical results obtained by a direct collocation method are presented, which also include those for active cooling of the engine by the liquid hydrogen fuel. The aerothermic load and the fuel loss due to engine cooling can be considerably reduced by optimization.

8.1 Introduction

In hypersonic flight regimes, i.e. with Mach number greater than about 5, the aerothermic heating constitutes one of the major problems requiring new materials and new heat protection systems that can withstand temperatures up to 2000 K and simultaneously are lightweight and of low maintenance. Therefore, advanced mathematical models are required to compute and optimize the thermal load in hypersonic flight regimes. For that, one necessitates a sophisticated model which couples a usual flight path optimization problem with an instationary heating constraint. This will lead to an optimal control problem subject to a coupled system of ordinary differential equations (ODEs), which describes the equations of motion, and, in addition, a parabolic partial differential equation (PDE), which describes the heat conduction through the thermal protection system. One can refer the coupled nonlinear ODE–PDE system as a nonlinear partial differential algebraic equation system (PDAE) (with zero differential time index as well as MOL index and differential space index to be infinity [19]).

Among the different concepts for future space transportation systems, we concentrate in this chapter on the German Sanger II concept [17]. This concept is concerned with a two-stage winged aircraft, consisting of a rocket-propelled orbital stage and an airbreathing carrier being able to start and land horizontally.¹ Optimal flight paths for two-stage vehicle of Sanger type have been computed and analysed, e.g. by [2, 4, 5, 14, 30]. In this chapter, however, we concentrate on the lower stage only which may be considered for intercontinental hypersonic flights too.

Thermal protection systems (TPSs) are a necessity for hypersonic flights and consist of different insulated layers of suitable materials and thicknesses. For a realistic simulation of the heat transfer, a detailed mathematical model is required, which includes unsteady effects and also material properties; see, e.g. [8, 34]. Layered TPSs have been investigated in [13, 16, 36].

¹ Angelo Miele published some critical work on the NASP (National Aero-Space Plane), which was supposed to be a single-stage configuration powered by four power plants operating in sequence in different Mach number regimes (turbojet, ramjet, scramjet, rocket). His analysis was done under a grant from NASA-LRC. Miele’s results indicated that the NASP was not feasible as a single-stage configuration and this is why he urged NASA to look instead to at least a double-stage configuration. The associated publication can be found in the *Acta of the Academy of Sciences of Turin*, a dusty academy founded by Lagrange; see [22]. See also [23] and [24].

In this chapter, we concentrate on the reduction of the extreme heating loads by simultaneously optimizing a ODE–PDE system consisting of a sophisticated model for hypersonic flight regimes and a quasi-linear heat equation, both for heat conduction and transport, with nonlinear boundary conditions including radiation effects. Simultaneous flight path optimization and aerothermic heating have been investigated also in, e.g. [8–10, 13, 15, 20, 25, 28, 35, 36].

In addition, we investigate active cooling of the engine by the liquid hydrogen fuel. This is especially suited because of its large heat capacity and low temperatures; see also, e.g. [3, 11, 27, 29]. Moreover, it improves the combustion of the fuel by increasing its activation energy. Active cooling is a common design in liquid-fuel rocket propulsion.

This chapter is an expanded version of [6] and is based on the report [18] and the PhD thesis [34]. In contrast to [34], we treat the problem as an optimal control problem for a coupled ODE–PDE system. In terms of numerical methods for PDEs, in [34] a cell-oriented energy preserving method was used with lumped parameters for heat capacity and heat conductivity, called knot model approach; see [8]. This approach is equivalent to solving an initial-multipoint-boundary-value problem. In this chapter, we treat the coefficients as temperature dependent. Significant technical improvements may be achieved by both approaches with respect to efficiency, costs and weight of the TPS.

The chapter is organized as follows. In Sect. 8.1 a 2D flight path optimal control problem for minimum fuel consumption is, first of all, presented and numerically solved by a direct collocation method. Its optimal solution will constitute the reference trajectory. Then, three modified optimal control problems are investigated, two of which take into account that the cold hydrogen fuel can be used to cool down the structure and the turbo-ramjet engine to their individual demands and reused for powering the engine afterwards. In contrast, the third model variant simulates that the fuel is released after it has been flowed through the cooling devices.

Finally, a much more complex optimal control problem including an instationary quasi-linear heat equation in one, resp., two spatial dimensions with nonlinear boundary conditions and a state constraint is considered, in order to limit the temperature in critical regions more realistically, for example, at the stagnation point or around the engine. The scheme used for discretizing the partial differential equation involved is analysed and associated numerical results are discussed. This is the content of Sect. 8.3.

From a mathematical point of view, the resulting problem constitutes an optimal control problem subject to a coupled system of ODEs and a PDE with a multiplicity of control and state variable inequality constraints. The coupling is done not only through coefficients of the PDE depending on state variables of the ODEs but also through boundary conditions of the parabolic PDE. The controls occur in the ODE system only. Hence, this problem constitutes a new type of ODE–PDE-constrained optimization problem where the PDE is simultaneously controlled in a distributed and boundary-controlled manner. However, the controls exert its influence indirectly only via the state variables of the ODE subproblem. Besides the typical state and control constraints in flight dynamics, the PDE subproblem additionally exhibits a

state constraint which closes the dependency between ODE and PDE subproblem. This type of optimal control problem not discussed so far in the literature has given rise to a subsequent paper, in which a detailed analysis of an abstract prototype problem of similar kind is presented, cf. [26].

For an overview of the theory of optimal control subject to PDEs we refer to the new book of Tröltzsch [33]. State-constrained PDE optimal control problems and also quasi-linear problems are subject to actual research. Some first references on state-constrained PDE optimal control problems can be found in [33] too.

Because of the complexity of the final model we pursue here the so-called direct approach despite its known lack of reliability, cf. [12]. By semi-discretization in the spatial variables (i.e. using the vertical method of lines) the state- and control-constrained ODE–PDE optimal control problem is first transformed into a large-scale state- and control-constrained, but purely ODE-constrained, optimal control problem, which itself is solved by a direct collocation method.

The chapter is completed with some numerical results for the ODE–PDE-constrained optimization in Sect. 8.4.

8.2 Trajectory Optimization Problems with Active Cooling

The well-known optimal control problem for a minimum fuel, resp., maximum final mass range flight of a hypersonic space vehicle over a spherical rotational earth under the assumptions of a point mass model and, for the sake of simplicity, for a 2D flight over the equator is given by

$$\max_{\alpha(t), \delta_T(t)} m(t_f) \quad (8.1)$$

subject to

$$\begin{aligned} \dot{v} = & \frac{1}{m} \left(T(v, h; \alpha, \delta_T) \cos(\alpha + \sigma_T) - D(v, h; \alpha) \right) \\ & - g(h) \sin \gamma + \omega_E^2 r(h) \sin \gamma, \end{aligned} \quad (8.2)$$

$$\begin{aligned} \dot{\gamma} = & \frac{1}{mv} \left(T(v, h; \alpha, \delta_T) \sin(\alpha + \sigma_T) + L(v, h; \alpha) \right) \\ & + \cos \gamma \left(\frac{v}{r(h)} - \frac{g(h)}{v} + \frac{\omega_E^2 r(h)}{v} \right) + 2 \omega_E, \end{aligned} \quad (8.3)$$

$$\dot{h} = v \sin \gamma, \quad (8.4)$$

$$\dot{\zeta} = v \cos \gamma, \quad (8.5)$$

$$\dot{m} = - \begin{cases} \beta_T(v, h; \alpha, \delta_T) & \text{(no active cooling,} \\ & \text{resp., active cooling I),} \\ \max\{\beta_T(v, h; \alpha, \delta_T), \beta_C(v, h; \alpha, \delta_T)\} & \text{(active cooling II),} \\ \beta_T(v, h; \alpha, \delta_T) + \beta_C(v, h; \alpha, \delta_T) & \text{(active cooling III).} \end{cases} \quad (8.6)$$

A derivation of the equations of motion can be found, e.g. in Miele [21] and Vinh et al. [37].

The state variables are velocity v , flight path angle γ , altitude h , path length ζ and vehicle mass m . The control variables are angle of attack α and throttle setting δ_T . Note that due to the hypersonic flight also the angle of attack exerts an influence on the instantaneous fuel consumption (8.6). Finally, the flight time interval is $[0, t_f]$ with the final time t_f unspecified in case of no active cooling or active cooling I or II, resp., specified in case of active cooling III.

The approximations used here for the data fields of lift $L = C_L(v, h; \alpha) (\rho(h)/2) v^2 S$, drag $D = C_D(v, h; \alpha) (\rho(h)/2) v^2 S$ and thrust $T(v, h; \alpha, \delta_T)$, which are sufficiently often differentiable in the relevant altitude–Mach number flight envelope ($0 \text{ km} < h < 40 \text{ km}$, $0 \leq M \leq 7.2$), can be found in [34]; see also [7, 36]. The model includes the possibility of overfueled combustion. In [34], approximations are given for all relevant aerodynamic quantities as functions of air temperature Θ_{air} and air density ρ using the laws of thermodynamics. Air temperature Θ_{air} and air density ρ themselves are approximated as functions of the altitude h . The functions $g(h) = g_0 (r_E/r(h))^2$ and $r(h) = r_E + h$ in Eqs. (8.2), (8.3), (8.4) and (8.5), (8.6) denote the acceleration of gravity and the distance of the vehicle from the geocentre.

Furthermore, the optimal trajectory must obey certain control and state variable inequality constraints

$$-\frac{1.5}{180} \pi \leq \alpha \leq \frac{20}{180} \pi, \quad 0 \leq \delta_T \leq 1, \quad (8.7)$$

$$0 \leq n(v, h, m; \alpha) \leq 2, \quad 10 \text{ [kPa]} \leq \bar{q}(v, h) \leq 50 \text{ [kPa]} \quad (8.8)$$

limiting the angle of attack α , the throttle setting δ_T , the load factor $n(v, h, m; \alpha) = L(v, h; \alpha)/(m g_0)$ and the dynamic pressure $\bar{q}(v, h) = (\rho(h)/2) v^2$.

The optimal control problem is completed by the following initial and final conditions: $v(0) = 150 \text{ m/s}$, $\gamma(0) = 0 \text{ rad}$, $h(0) = 500 \text{ m}$, $\zeta(0) = 0 \text{ km}$, $m(0) = 244000 \text{ kg}$, $v(t_f) = 150 \text{ m/s}$, $\gamma(t_f) = 0 \text{ rad}$, $h(t_f) = 500 \text{ m}$ and $\zeta(t_f) = 9000 \text{ km}$.

In the present chapter, we will consider four different models for describing the fuel consumption. The first model (8.1), (8.2), (8.3), (8.4), (8.5) and (8.6, no active cooling) takes into account only the fuel consumption $\beta_T(v, h; \alpha, \delta_T)$ of the turbo-ramjet engine. It will be referred to as reference trajectory optimization problem. Its optimal solution is shown in Figs. 8.1 and 8.2 and yields a fuel consumption of 60614 kg and a final time of $t_f = 7023 \text{ s}$.

The other three models additionally include a complicated active cooling system of the engine modelled by an additional term $\beta_C(v, h; \alpha, \delta_T)$ where each major component of the engine is cooled down to its individual demand, such as walls and nozzle of the turbo combustion chamber, here to a temperature of $\Theta_E = 1600 \text{ K}$. Furthermore, in two of those three model variants the instantaneous amount of fuel for cooling must not exceed that for thrust, i.e.

$$\beta_C(v, h; \alpha, \delta_T) \leq \beta_T(v, h; \alpha, \delta_T). \quad (8.9)$$

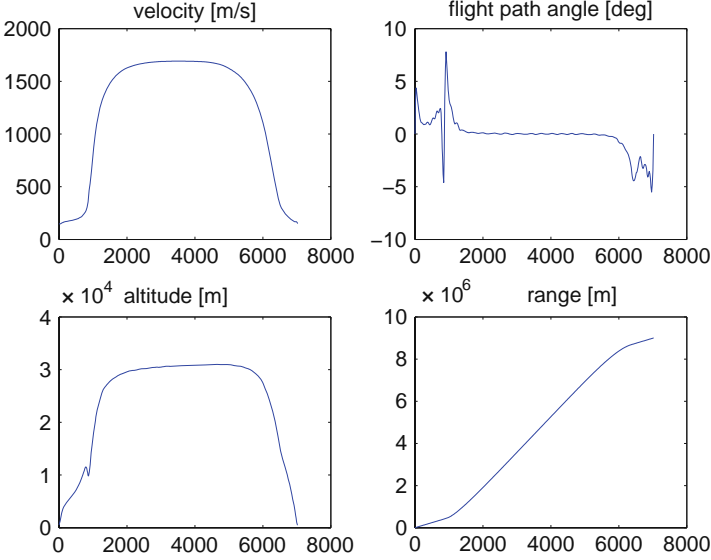


Fig. 8.1 Time histories of the state variables v , γ , h and ζ for the reference trajectory optimization problem (no active cooling), identical to active cooling I and II (backflow of fuel from cooling devices)

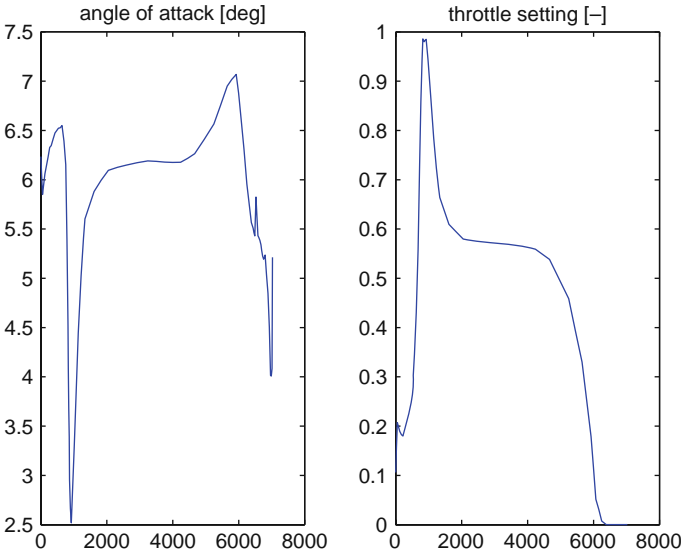


Fig. 8.2 Time histories of the control variables α and δ_T for the reference trajectory optimization problem (no active cooling), which is identical to active cooling I and II (backflow of fuel from cooling devices)

Concerning the first mass model (8.6, no active cooling), this constraint, when additionally taken into account, allows that the entire amount of fuel for the cooling circuit can be reused for the thrust afterwards, or, in other words, the flight path has to be controlled in such a way that the fuel needed for cooling must not exceed that for thrust. The variant (8.1), (8.2), (8.3), (8.4), (8.5) and (8.6, active cooling I) (8.7) is referred to as active cooling I. However, the optimal solution of this problem coincides with the reference trajectory shown in Figs. 8.1 and 8.2, since the state constraint (8.9) becomes nowhere active.

Therefore, the optimal trajectory with (8.6, active cooling I), (8.7) replaced by (8.6, active cooling II), coincides with the reference trajectory too. This variant would assess the fuel for cooling only as consumption, if it really exceeds that for thrust. It has been already investigated in [8, 34], where, however, the instantaneous fuel consumption β_C exceeds β_T at certain times and the surplus had therefore to be released.

In addition, Figs. 8.3 and 8.4 show the optimal trajectory for (8.1), (8.2), (8.3), (8.4), (8.5) and (8.6, active cooling III) when the fuel used for active cooling is not reused for the engines. When the terminal time is released to 7.500s, one obtains a fuel consumption of 84.597kg. This amount increases even to 91.504kg, if we screw down the terminal time to the optimal value of the reference trajectory.

All trajectories can be separated in three phases: ascent, cruise flight and descent of the hypersonic vehicle; see Fig. 8.1. Within about 1500s the cruising altitude is reached followed by a strong acceleration with maximum thrust; see Fig. 8.2. The cruise proceeds with almost constant flight path angle, while the two controls, angle of attack and throttle setting, as well as the lift coefficient are also almost constant; see Figs. 8.1 and 8.2 again. The latter stays in the vicinity of the lift coefficient for

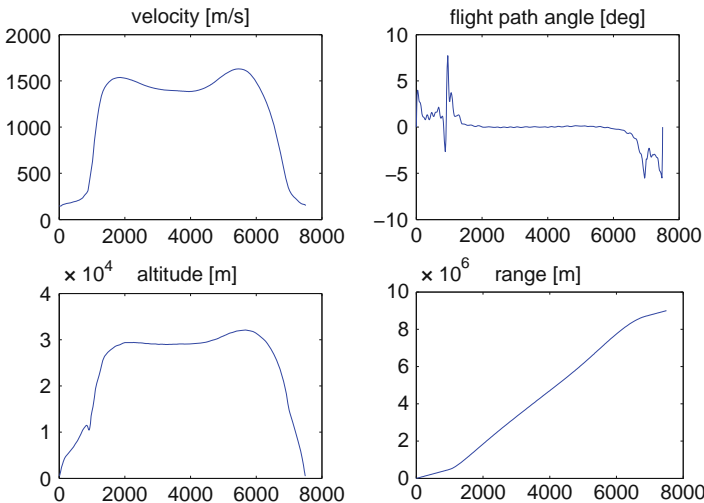


Fig. 8.3 Time histories of the state variables v , γ , h and ζ for the trajectory optimization problem with fixed final time $t_f = 7.500$ s, active cooling III (fuel from cooling devices is released)

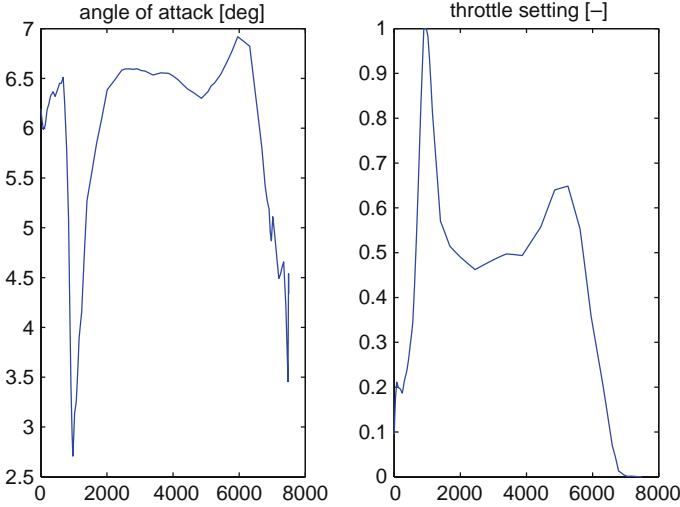


Fig. 8.4 Time histories of the control variables α and δ_T for the trajectory optimization problem with fixed final time $t_f = 7500$ s, active cooling III (fuel from cooling devices is released)

minimum drag, cf. [34]. A slight periodicity can be observed; see Fig. 8.2. This quasi-stationary flight phase covers about 60% of the entire flight time.

All numerical results have been obtained by the direct collocation optimal control software package DIRCOL of Stryk [31, 32].

So far we have not taken into account a goal-oriented reduction of the temperature of the TPS in critical regions in the optimization process. This will be the subject of the subsequent section.

8.3 Trajectory Optimization Problem with an Instationary Heat Constraint

In hypersonic flow the air stream impinges on the vehicle surface, and therefore its kinetic energy is transformed into thermic energy which leads to air temperatures of more than 2000 K. In order to reduce the costs for the necessary thermal protection system, the reference problem is expanded so that a limitation of the surface temperature can be obtained by optimal control. For this purpose, the system of ordinary differential equations (8.2), (8.3), (8.4), (8.5), and (8.6, no active cooling) for the dynamical behaviour of the vehicle is augmented by a parabolic partial differential equation for instationary heat flow and an additional state constraint for the temperature in critical regions of the surface.

For the sake of simplicity and computability, we restrict ourselves to the most critical regions, the stagnation point and the engine. When neglecting the heat flow

tangential to the layers of the TPS, it is sufficient to deal with a spatially 1D heat equation. This is, for example, suitable for an investigation of the heating at the stagnation point. However, for the investigation in the neighbourhood of the engine we have to deal with a spatially 2D problem. In both cases, the PDE subsystem for the heat load $\Theta(\mathbf{x}, t)$, $\mathbf{x} \in \Omega \subset \mathbf{R}^d$, $d = 1, 2$, $t \in (0, t_f)$, Ω open, bounded and lying in the vertical symmetry plane of the aircraft, is given by the following initial-boundary-value problem:

$$\begin{aligned} \rho(h) (c_p(\Theta) + c'_p(\Theta) \Theta) \frac{\partial \Theta}{\partial t} + \rho'(h) h c_p(\Theta) \Theta \\ - \nabla \cdot (\lambda(\Theta, p) \nabla \Theta) = 0 \quad \text{for all } (\mathbf{x}, t) \in \Omega \times (0, t_f), \end{aligned} \quad (8.10)$$

$$\Theta(\mathbf{x}, 0) = \Theta_0(\mathbf{x}) = 300 \text{ [K]} \quad \text{for all } \mathbf{x} \in \Omega, \quad (8.11)$$

$$\frac{\partial \Theta}{\partial \mathbf{n}}(\mathbf{x}, t) = q_{\text{conv}} - q_{\text{rad}} \quad \text{for all } \mathbf{x} \in \partial \Omega \quad \text{and all } t \in (0, t_f), \quad (8.12)$$

where the heat capacity c_p depends on the temperature Θ and the heat conductivity λ on both the temperature Θ and the pressure $p = \rho(h) R \Theta_{\text{air}}(h)$; R is the gas constant. Hence, the pressure itself is a function of the altitude h . Heat capacity and heat conductivity are given as follows:

$$\begin{aligned} c_p = R \frac{\kappa}{\kappa - 1} \quad \text{with} \quad \kappa = \frac{\sum_{i=0}^5 d_i \Theta^i}{\sum_{j=0}^1 e_j \Theta^j}, \\ \lambda(\Theta, p) = \begin{cases} \sum_{i=0}^3 g_i \Theta^i & \text{if } \Theta \leq 1400 \text{ [K]} \\ \xi_1 \sum_{i=0}^3 g_i \Theta^i + \xi_2 P_\lambda(\Theta, p) & \text{if } 1400 \text{ [K]} < \Theta < 1600 \text{ [K]} \\ P_\lambda(\Theta, p) & \text{if } 1600 \text{ [K]} \leq \Theta \end{cases}. \end{aligned}$$

Here, $\xi_1 = (1 - \tanh(100(\Theta - 1500)))/2$ and $\xi_2 = (1 + \tanh(100(\Theta - 1500)))/2$ denote transfer functions between the pressure-independent regime below 1400 K and the pressure-dependent regime above 1600 K. Furthermore, P_λ is a polynomial approximation of the heat conductivity based on suitable data fields; see [34] for details.

A derivation of Eq. (8.10) is given in the Appendix, Part A.

In the nonlinear boundary condition (8.12), the quantities q_{conv} and q_{rad} relate to the convective and radiative heat fluxes. If the outer normal \mathbf{n} points into the exterior environment of the vehicle, we use

$$\begin{aligned} q_{\text{conv}} &:= q_{\text{air}}(\nu, h, \Theta_{\text{air}}; \alpha; x_L, Q) \quad \text{with} \quad \Theta_{\text{air}} = \tau(\nu, h, \Theta_{\text{air}}; \alpha), \\ q_{\text{rad}} &:= \varepsilon \sigma (\Theta^4 - \Theta_{\text{air}}^4), \end{aligned}$$

where the air temperature Θ_{air} after the shock depends on the actual flight conditions. Furthermore, x_L denotes the length of the laminar-to-turbulent transition. It varies with the position Q on the vehicle surface. The constants ε and σ denote emissivity and Stefan–Boltzmann constant.

If the outer normal \mathbf{n} points into the interior of the vehicle, we use

$$\begin{aligned} q_{\text{conv}} &:= \alpha_q (\Theta - \Theta_{\text{int}}) , \\ q_{\text{rad}} &:= \varepsilon \sigma (\Theta^4 - \Theta_{\text{int}}^4) , \end{aligned}$$

with the heat transfer coefficient $\alpha_q = \alpha_q(\Theta_{\text{int}}, p)$ and the interior temperature Θ_{int} which relates to the inner surface of the TPS. This may be either a part of the vehicle structure or the inside air. Again all formulae can be found in [34].

For the sake of simplicity, we have not included a layered TPS which would require to include radiative fluxes at each boundary layer of the TPS, too, making Eqs. (8.10), (8.11) and (8.12) an initial-multipoint-boundary-value problem. By a so-called knot model approach this has been included in the optimization process in [34]. Actually this approach can be interpreted as a semi-discretization in space where the step size equals the thickness of the layers.

Finally, the model is completed by a state constraint on the temperature:

$$\Theta(\mathbf{x}, t) \leq \Theta_{\max}(\mathbf{x}) \quad \text{for all } (\mathbf{x}, t) \in \Omega \times (0, t_f). \quad (8.13)$$

In order to afford the numerical computations, we simplify the upper bound Θ_{\max} to a constant and confine the domain Ω in (8.13) either to a spatial interval (1D case), e.g. extending from the stagnation point inwards through the TPS by neglecting any fluxes tangentially to the TPS, or to a rectangle (2D case) in the vertical symmetry plane of the aircraft also pointing inwards through the TPS, e.g. for investigating heating effects near the engine. Indeed, it is sufficient to limit the temperature only in the most critical regions. In the 1D case, the parabolic partial differential equation is discretized with respect to the space variable x and with step size Δx using an implicit method. This is necessary because of the stiffness of the resulting ODE system. We have used the scheme

$$\begin{aligned} & \rho(h) (c_p(\Theta_1) + c'_p(\Theta_1) \Theta_1) \dot{\Theta}_1 \\ &:= \frac{1}{\Delta x} \left(\left(q_{\text{air}}(v, h, \Theta_{\text{air}}; \alpha; x_L, Q) - \varepsilon \sigma (\Theta_1^4 - \Theta_{\text{air}}^4) \right) \right. \\ & \quad \left. - \lambda \left(\frac{\Theta_1 + \Theta_2}{2} \right) \frac{\Theta_1 - \Theta_2}{\Delta x} \right) - \rho'(h) h c_p(\Theta_1) \Theta_1, \end{aligned} \quad (8.14)$$

$$\begin{aligned} & \rho(h) (c_p(\Theta_i) + c'_p(\Theta_i) \Theta_i) \dot{\Theta}_i \\ &:= \frac{1}{\Delta x} \left(\lambda \left(\frac{\Theta_{i-1} + \Theta_i}{2} \right) \frac{\Theta_{i-1} - \Theta_i}{\Delta x} - \lambda \left(\frac{\Theta_i + \Theta_{i+1}}{2} \right) \frac{\Theta_i - \Theta_{i+1}}{\Delta x} \right) \\ & \quad - \rho'(h) h c_p(\Theta_i) \Theta_i, \quad \text{for } i = 2, \dots, n-1, \end{aligned} \quad (8.15)$$

$$\begin{aligned}
& \rho(h) \left(c_p(\Theta_n) + c'_p(\Theta_n) \Theta_n \right) \dot{\Theta}_n \\
& := \frac{1}{\Delta x} \left(\lambda \left(\frac{\Theta_{n-1} + \Theta_n}{2} \right) \frac{\Theta_{n-1} - \Theta_n}{\Delta x} \right. \\
& \quad \left. - \left(\alpha_q (\Theta_n - \Theta_{\text{int}}) - \varepsilon \sigma (\Theta_n^4 - \Theta_{\text{int}}^4) \right) \right) \\
& \quad - \rho'(h) \dot{h} c_p(\Theta_n) \Theta_n,
\end{aligned} \tag{8.16}$$

where $\Theta_i = \Theta_i(t) := \Theta(x_i, t)$.

For the sake of brevity, we show, for the analogous discretization in 2D, only the stencil of the discretization; see Fig. 8.5. A detailed derivation is given in the Appendix, Part C.

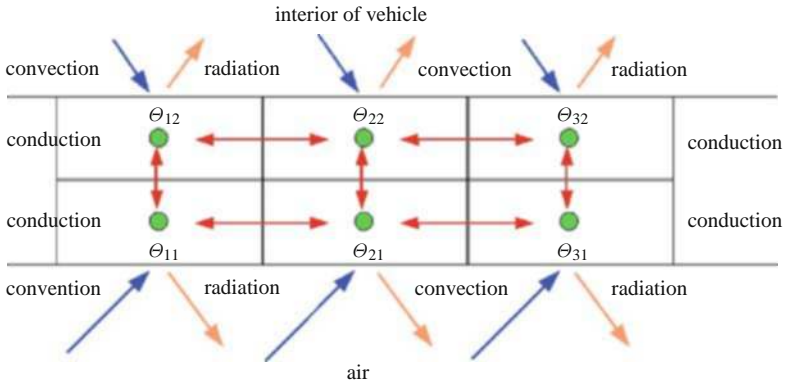


Fig. 8.5 Discretization stencil (here $n = 3, m = 2$) for the 2D case

Here, the notations are $\Theta_{ij} = \Theta_{ij}(t) := \Theta(\mathbf{x}_{ij}, t)$ with $\mathbf{x}_{ij} := (x_i, y_j)$, $i = 1, \dots, n$, $j = 1, \dots, m$. In case of dealing with a layered TPS the vertical conduction arrows in Fig. 8.5 are to be replaced by convection and radiation arrows likewise as at the initial or terminal lines.

These discretization schemes can be shown to be of order 2; see Appendix, Parts B and C.

The ODE part (8.2), (8.3), (8.4), (8.5) and (8.6, no active cooling) is now augmented by either the system (8.14), (8.15) and (8.16) or an equivalent system associated with the stencil of Fig. 8.5 including n , resp., nm additional discretized state variable inequality constraints

$$\Theta_i(t) \leq \Theta_{\max} \quad \text{for all } t \in (0, t_f) \quad \text{with } i = 1, \dots, n$$

or

$$\Theta_{i,j}(t) \leq \Theta_{\max} \quad \text{for all } t \in (0, t_f) \quad \text{with } i = 1, \dots, n, j = 1, \dots, m.$$

Note that due to the quasi-linearity of the PDE the resulting ODE system (8.14), (8.15) and (8.16) is nonlinear.

We now present the numerical results, first for the 1D case. Figure 8.6 shows the results for a limit temperature of $\Theta_{\max} = 1000$ K at the stagnation point. It exhibits a boundary arc on the first line which still can be seen on the subsequent lines, however, with decaying maximum temperatures.

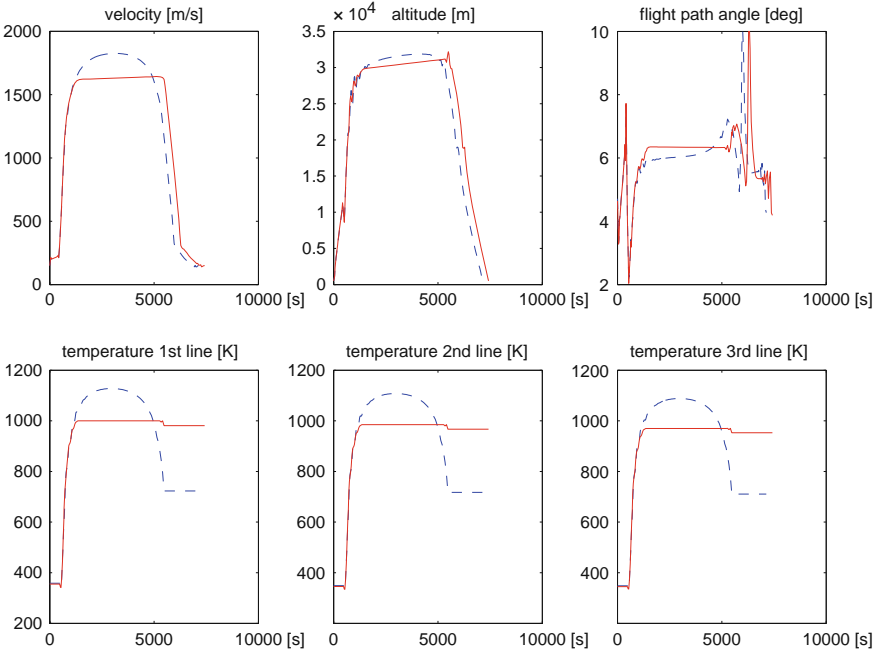


Fig. 8.6 Time histories of state variables v , h , and γ . Dashed lines: reference trajectory optimization problem. Solid lines: temperature-constrained optimization problem with $\Theta_{\max,i} = 1000$ K at the stagnation point

In order to reduce the temperature at the stagnation point, one obviously has to fly in lower altitudes at lower velocities. For a reduction of the stagnation point temperature compared to the reference optimization problem by 10%, one has to increase, thanks to optimization, the total fuel consumption by about 1% only; see Fig. 8.7.

It has to be noticed that we had to restrict the relevant time interval for the heat equation to the interval $[482 \text{ s}, 5478 \text{ s}]$. For larger intervals numerical results could not be obtained anymore. However, this does not play a role, since the heat load obviously is less than its maximum outside of this interval.

Finally, we present the numerical results for the 2D case close to the front part of the fuel tank; see Fig. 8.8. The temperature constraints do not become active here.

The temperatures also remain moderate at the lower surface; see Fig. 8.9.

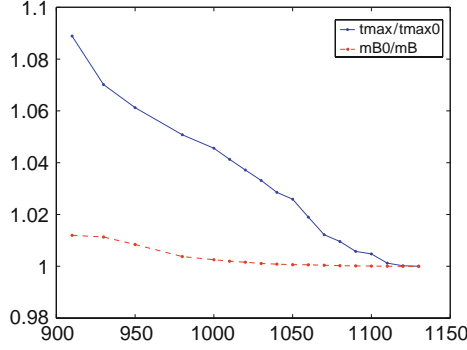


Fig. 8.7 Maximum temperature ratio $\Theta_{\max, \text{ref}}/\Theta_{\max}$ (upper curve) and ratio of the total fuel consumption $m_{\text{fuel, ref}}/m_{\text{fuel}}$ (lower curve) versus maximum temperature Θ_{\max} . The index ref refers to the reference optimization problem

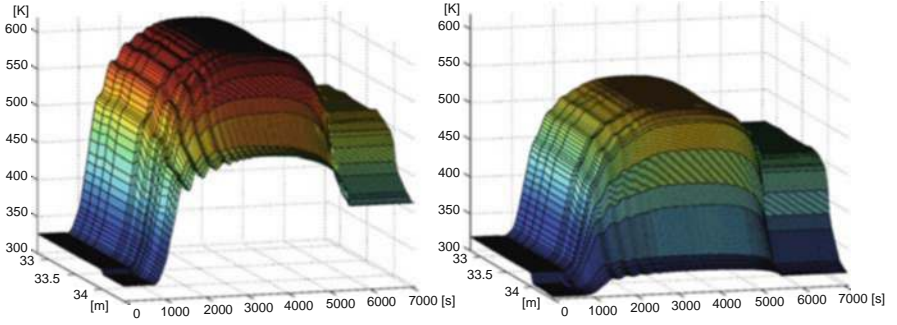


Fig. 8.8 Temperature profile $\Theta(x, \cdot, t)$ close to the front of the fuel tank on the surface (left: $\Theta(x, 0, t)$) and inside the structure (right: $\Theta(x, \Delta y, t)$) on time interval $[482 \text{ s}, 5478 \text{ s}]$

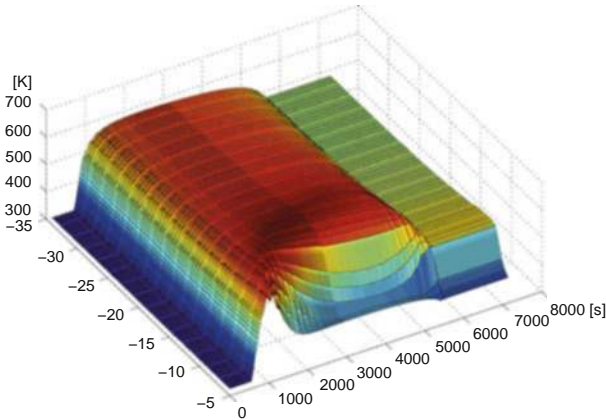


Fig. 8.9 Temperature profile $\Theta(x, 0, t)$ at the bottom side of the hull before the tank starts on time interval $[482 \text{ s}, 5478 \text{ s}]$

8.4 Conclusions

A complex mathematical model has been presented in order to control the heating of thermal protection systems of hypersonic aircraft. The model is a coupled system of nonlinear ordinary and a quasi-linear parabolic partial differential equation with nonlinear boundary conditions. Altogether the mathematical model presented here is an ODE–PDE control-and-state-constrained optimal control problem. By a semi-discretization in space using a finite volume method, this problem is transformed into a large-scale ODE control-and-state-constrained optimal control problem. Note that the resulting problem cannot be solved by standard software packages to higher accuracies. Direct ODE-constrained optimal control software as well as their incorporated SQP methods come to their limits. Nevertheless, despite the coarse discretization of the PDE, satisfactory results could be obtained, which show that detailed modelling and sophisticated numerical methods can help to determine the necessary dimensioning of thermal protection systems for hypersonic aircraft.

Moreover, the problem constitutes also a challenge for the growing field of PDE, resp., PDAE-constrained optimization in applied mathematics, since it contains features which have, so far, not been theoretically studied in the context of optimal control theory. For an abstract twin problem of an equivalent type, see [26].

Acknowledgments We are indebted to Prof. Dr. Oskar von Stryk for providing us with his direct optimal control software package DIRCOL.

Appendix

A. Proof of Eq. (8.10): Let $\Theta(\mathbf{x}, t)$ be a sufficiently smooth inhomogeneous temperature distribution in $\Omega \times \mathbf{R}_{\geq 0}$. According to Fourier's law, this gives rise to a conductive energy flux $q_{\text{cond}} = -\lambda(\Theta) \nabla \Theta$. Considering the energy density $\rho(h) c_p(\Theta) \Theta$ in an arbitrary subset $\omega \subset \Omega$, there holds, because of the integral form of the law of conservation of energy,

$$\frac{d}{dt} \int_{\omega} \rho(h(t)) c_p(\Theta(\mathbf{x}, t)) \Theta(\mathbf{x}, t) d\mathbf{x} = - \int_{\partial\omega} q_{\text{cond}}(\mathbf{x}, t) \cdot \mathbf{n} d\mathbf{s}.$$

Using Gauss' theorem, this implies the differential equation

$$\frac{d}{dt} (\rho(h(t)) c_p(\Theta(\mathbf{x}, t)) \Theta(\mathbf{x}, t)) = \nabla \cdot (\lambda(\Theta(\mathbf{x}, t)) \nabla \Theta(\mathbf{x}, t))$$

which yields Eq. (8.10). □

B. Proof of the consistency order of Eq. (8.15): Let $\Theta(x, t)$ be the sufficiently smooth exact solution of Eq. (8.10). Let $\mathcal{L}_{\Delta x}$ denote the difference operator of Eq. (8.15) due to the spatial step size Δx . For Eq. (8.15) being at least of second

consistency order, i.e. $\mathcal{L}_{\Delta x}(\Theta(x, t)) = \mathcal{O}(\Delta x^2)$, we only have to show that this scheme is symmetric, i.e. that $\mathcal{L}_{\Delta x} = \mathcal{L}_{-\Delta x}$. This, however, can be easily seen. By Taylor expansion it can be shown that the scheme (8.15) is of order 2 indeed. Moreover, it is also stable, hence convergent, since it is the 1D analogue of the finite volume scheme discussed below.

C. Derivation of the 2D difference scheme as finite volume method, resp., energy balance method: Starting point is the investigation of the energy balance in a finite volume $\omega_{i,j}$ of the discretized domain Ω :

$$\int_{\omega_{i,j}} \nabla \cdot (\lambda(\Theta(x_i, y_j)) \nabla \Theta(x_i, y_j)) \, d\omega = \int_{\partial\omega_{i,j}} \lambda(\Theta(x_i, y_j)) \nabla \Theta(x_i, y_j) \cdot \mathbf{n} \, ds,$$

where we suppress the dependence on t , since we are essentially interested in a spatial semi-discretization. First, we introduce the following notations:

$$\begin{aligned} x_{i+\frac{1}{2}} &= \frac{x_{i+1} - x_i}{2}, & y_{j+\frac{1}{2}} &= \frac{y_{j+1} + y_j}{2}, \\ \Delta x_i &= \frac{x_{i+1} - x_{i-1}}{2}, & \Delta y_j &= \frac{y_{j+1} - y_{j-1}}{2}, \\ \Delta x_{i+\frac{1}{2}} &= x_{i+1} - x_i, & \Delta y_{j+\frac{1}{2}} &= y_{j+1} - y_j, \\ \lambda_{i+\frac{1}{2}j} &= \lambda(\Theta(x_{i+\frac{1}{2}}, y_j)), & \lambda_{i+\frac{1}{2}j} &= \lambda(\Theta(x_i, y_{j+\frac{1}{2}})), \\ f_{i,j} &= [\rho(h) (c_p(\Theta) + c'_p(\Theta) \Theta) \dot{\Theta} + \rho'(h) \dot{h} c_p(\Theta) \Theta]_{x_i, y_j}. \end{aligned}$$

Using the support values $\Theta_{i,j}$ in the centre of each element $\omega_{i,j}$, one can then approximate the line integrals by

$$\begin{aligned} \int_{\partial\omega_{i,j} \text{ east}} \lambda \frac{\partial \Theta}{\partial x} \, dy &\approx \lambda_{i+\frac{1}{2}j} \frac{\Theta_{i+1j} - \Theta_{ij}}{\Delta x_{i+\frac{1}{2}}} \Delta y_j, \\ \int_{\partial\omega_{i,j} \text{ west}} \lambda \frac{\partial \Theta}{\partial x} \, dy &\approx \lambda_{i-\frac{1}{2}j} \frac{\Theta_{ij} - \Theta_{i-1j}}{\Delta x_{i-\frac{1}{2}}} \Delta y_j, \\ \int_{\partial\omega_{i,j} \text{ nord}} \lambda \frac{\partial \Theta}{\partial y} \, dx &\approx \lambda_{ij+\frac{1}{2}} \frac{\Theta_{ij+1} - \Theta_{ij}}{\Delta y_{i+\frac{1}{2}}} \Delta x_i, \\ \int_{\partial\omega_{i,j} \text{ south}} \lambda \frac{\partial \Theta}{\partial y} \, dx &\approx \lambda_{ij-\frac{1}{2}} \frac{\Theta_{ij} - \Theta_{ij-1}}{\Delta y_{j-\frac{1}{2}}} \Delta x_i, \end{aligned}$$

and

$$\int_{\omega_{i,j}} f \, d\omega \approx f_{i,j} \Delta x_i \Delta y_j$$

with f analogously defined as $f_{i,j}$.

Finally, the energy balance over the element $\omega_{i,j}$ yields

$$\begin{aligned} & \left[\lambda_{i+\frac{1}{2}j} \frac{\Theta_{i+1j} - \Theta_{ij}}{\Delta x_{i+\frac{1}{2}}} - \lambda_{i-\frac{1}{2}j} \frac{\Theta_{ij} - \Theta_{i-1j}}{\Delta x_{i-\frac{1}{2}}} \right] \Delta y_j \\ & + \left[\lambda_{ij+\frac{1}{2}} \frac{\Theta_{ij+1} - \Theta_{ij}}{\Delta y_{j+\frac{1}{2}}} - \lambda_{ij-\frac{1}{2}} \frac{\Theta_{ij} - \Theta_{ij-1}}{\Delta y_{j-\frac{1}{2}}} \right] \Delta x_i \\ & = f_{i,j} \Delta x_i \Delta y_j. \end{aligned}$$

The evaluation of λ at the midpoints $x_{i+\frac{1}{2}}$, resp. $y_{j+\frac{1}{2}}$, requires an interpolation, for example, $\Theta_{i+\frac{1}{2}j} = [\Theta(x_i, y_j) + \Theta(x_{i+1}, y_j)]/2 + \mathcal{O}(\Delta x)$. Herewith, we obtain the final 2D scheme analogue to (8.15):

$$\begin{aligned} & \left[\lambda \left(\frac{\Theta_{i-1j} + \Theta_{ij}}{2} \right) \frac{\Theta_{i-1j} - \Theta_{ij}}{\Delta x_{i-\frac{1}{2}}} - \lambda \left(\frac{\Theta_{ij} + \Theta_{i+1j}}{2} \right) \frac{\Theta_{ij} - \Theta_{i+1j}}{\Delta x_{i+\frac{1}{2}}} \right] \Delta y_j \\ & + \left[\lambda \left(\frac{\Theta_{ij-1} + \Theta_{ij}}{2} \right) \frac{\Theta_{ij-1} - \Theta_{ij}}{\Delta y_{j-\frac{1}{2}}} - \lambda \left(\frac{\Theta_{ij} + \Theta_{ij+1}}{2} \right) \frac{\Theta_{ij} - \Theta_{ij+1}}{\Delta y_{j+\frac{1}{2}}} \right] \Delta x_i \\ & = f_{i,j} \Delta x_i \Delta y_j. \end{aligned}$$

This yields the main loop of the 2D discretization scheme. In a similar way the boundary conditions are employed. In the case of Neumann conditions as given here, so-called ghost values outside the domain Ω need not be considered as for Dirichlet conditions. This scheme is also at least of order 2 because of its symmetry in Δx as well as Δy . It is known that finite volume schemes of this type are stable and thus convergent, see, e.g. [1]. Obviously, it is also a conservative scheme, since it preserves the energy not only locally but also globally if $\cup_{i=1, j=1}^{n,m} \omega_{i,j} = \Omega$. \square

References

1. T. Barth, M. Ohlberger: *Finite volume methods: foundation and analysis*. In: E. Stein, R. de Borst, T. J. R. Hughes (Eds.): *Encyclopedia of Computational Mechanics*, Volume 1, Fundamentals. John Wiley and Sons, Weinheim, Germany, 439–474, 2004.
2. R. Bayer, G. Sachs: *Optimal return-to-base cruise of hypersonic carrier vehicles*, Z. Flugwiss. Weltraumforsch. 19 (1995) 47–54.
3. M. Bouchez, S. Beyer, G. Cahuzac: *PTAH-SOCR fuel cooled composite material structure for dual mode ramjet and liquid rocket engines*, Proc. of the 40th AIAA/ASME/SAE/ASEE Joint Propulsion Conference and Exhibit, Fort Lauderdale, USA, 2004. AIAA 2004-3653.
4. W. Buhl, K. Ebert, H. Herbst: *Optimal ascent trajectories for advanced launch vehicles*, AIAA Fourth International Aerospace Planes Conf., Orlando, Florida, 1992. AIAA-92-5008.
5. R. Bulirsch, K. Chudej: *Combined optimization of trajectory and stage separation of a hypersonic two-stage space vehicle*, Z. Flugwiss. Weltraumforsch. 19 (1995) 55–60.
6. K. Chudej, M. Wächter, F. Le Bras: *Verringerung der thermischen Belastung eines Hyperschall-Flugsystems durch Trajektorienoptimierung*, Proc. Appl. Math. Mech. 5 (2005) 803–804.

7. J. Drexler: *Untersuchung optimaler Aufstiegsbahnen raketengetriebener Raumtransporter-Oberstufen*, PhD Thesis, Technische Universität München, Faculty of Mechanical Engineering, Munich, Germany, 1995.
8. M. Dinkelmann: *Reduzierung der thermischen Belastung eines Hyperschallflugzeugs durch optimale Bahnsteuerung*. PhD thesis, Technische Universität München, Faculty for Mechanical Engineering, Munich, Germany, 1997.
9. M. Dinkelmann, M. Wächter, G. Sachs: *Modelling and simulation of unsteady heat transfer effects on trajectory optimization of aerospace vehicles*, Math. Comput. in Simulat. 53 (2002) 389–394.
10. M. Dinkelmann, M. Wächter, G. Sachs: *Modelling of heat transfer and vehicle dynamics for thermal load reduction by hypersonic flight optimization*, Math. and Comp. Model. Dyn. Systems 8 (2002) 237–255.
11. D. Glass, A. Dilley, H. Kelley: *Numerical analysis of convection/transpiration cooling*, Proc. 9th International Space Planes and Hypersonic Systems and Technologies Conference, Norfolk, USA, 1999. AIAA 99-4911.
12. W. W. Hager: *Runge-Kutta methods in optimal control and the transformed adjoint system*, Numerische Mathematik 87 (2000) 247–282.
13. C. Jänsch, A. Markl: *Trajectory optimization and guidance for a Hermes-type reentry vehicle*, in Proc. of the AIAA Guidance, Navigation, and Control Conference, New Orleans, USA, 1991. AIAA-91-2659.
14. C. Jänsch, K. Schnepper, K. H. Well: *Trajectory optimization of a transatmospheric vehicle*, Proc. American Control Conf., Boston, (1991), 2232–2237.
15. H. Kreim, B. Kugelman, H. J. Pesch, M. Breitner: *Minimizing the maximum heating of a re-entering space shuttle: An optimal control problem with multiple control constraints*, Optim. Contr. Appl. Met. 17 (1996) 45–69.
16. C. Krishnaprakas: *Efficient solution of spacecraft thermal models using preconditioned conjugate gradient methods*, J. Spacecraft Rockets 35(1998) 760–764.
17. H. Kuczera, H. Hauck, P. Sacher: *The German hypersonics technology programme-status 1993 and perspectives*, Proc. of the 5th AIAA/DGLR International Aerospace Planes and Hypersonics Technologies Conference, Munich, Germany, 1993. AIAA-93-5159.
18. F. Le Bras: *Trajectoires optimales en vol hypersonique tenant compte de léchauffement instationnaire de l'avion*. Rapport de stage d'option scientifique, Ecole Polytechnique Paris, France, and Universität Bayreuth, Germany, 2004.
19. W. S. Martinson, P. I. Barton: *A differentiation index for partial differential-algebraic equations*, SIAM J. Sci. Comput. 21 (1999) 2295–2315.
20. E. Meese, H. Nørstrud: *Simulation of convective heat flux and heat penetration for a spacecraft at re-entry*, Aerosp. Sci. and Technol. 6 (2002) 185–194.
21. A. Miele: *Flight Mechanics I, Theory of Flight Paths*. Addison-Wesley, Reading, 1962.
22. A. Miele, W. Y. Lee, G. D. Wu: *Ascent Performance Feasibility of the National Aerospace Plane*, Atti della Accademia delle Scienze di Torino 131 (1997) 91–108.
23. A. Miele, S. Mancus: *Optimal Ascent Trajectories and Feasibility of Next-Generation Orbital Spacecraft*, J. Optim. Theory App. 95 (1997) 467–499.
24. A. Miele, S. Mancuso: *Design Feasibility via Ascent Optimality for Next-Generation Spacecraft*, Acta Astronaut. 45/11 (1999) 655–668.
25. H. J. Pesch: *Numerische Berechnung optimaler Steuerungen mit Hilfe der Mehrzielmethode dokumentiert am Problem der Rückführung eines Raumgleiters unter Berücksichtigung von Aufheizbegrenzungen*, Diploma Thesis, Universität Köln, Department of Mathematics, Cologne, Germany, 1973.
26. H. J. Pesch, A. Rund, W. von Wahl, S. Wendl: *On a Prototype Class of ODE-PDE State-Constrained Optimal Control Problems. Part 1: Analysis of the State-unconstrained Problems. Part 2: The State-constrained Problems*, submitted.
27. V. Rausch, C. McClinton: *NASA's Hyper-X program*, Proc. of the 51th International Astronautical Congress, IAF-00-V.4.01, Rio de Janeiro, Brasil, 2000.

28. J. Ring: *Flight trajectory control for thermal abatement of hypersonic vehicles*, Proc. of the 2nd AIAA International Aerospace Planes Conference, Orlando, USA, 1990.
29. G. Sachs, M. Dinkelmann: *Reduction of coolant fuel losses in hypersonic flight by optimal trajectory control*, J. Guid. Control Dynam. 19 (1996) 1278–1284.
30. G. Sachs, W. Schoder: *Optimal separation of lifting vehicles in hypersonic flight*, AIAA Guidance, Navigation, and Control Conference, New Orleans, LA, Aug. 12–14, 1991, Technical Papers, Vol. 1 (A91-49578 21-08). Washington, DC, American Institute of Aeronautics and Astronautics, 1991, p. 529–536.
31. O. v. Stryk: *Numerische Lösung optimaler Steuerungsprobleme: Diskretisierung, Parameteroptimierung und Berechnung der adjungierten Variablen*. Fortschritt-Berichte VDI, Reihe 8: Meß-, Steuerungs- und Regeltechnik, No. 441, VDI Verlag, Düsseldorf, Germany, 1995.
32. O. v. Stryk: *User's Guide for DIRCOL. A direct Collocation method for the numerical solution of optimal control problems*. Version 1.2. Lehrstuhl für Höhere Mathematik und Numerische Mathematik, Technische Universität München, Munich, Germany, 1995.
33. F. Tröltzsch: *Optimalsteuerung bei partiellen Differentialgleichungen*. Vieweg, Wiesbaden, Germany, 2005.
34. M. Wächter: *Optimalflugbahnen im Hyperschall unter Berücksichtigung der instationären Aufheizung*. PhD Thesis, Technische Universität München, Faculty of Mechanical Engineering, Munich, Germany, 2004.
35. M. Wächter, G. Sachs: *Unsteady heat load reduction for a hypersonic vehicle with a multi-point approach*. In: Optimal Control. Workshop at the University of Greifswald, 1.–3.10.2002, Report of the Collaborative Research Center 255 of the Deutsche Forschungsgemeinschaft (German Research Foundation), Technische Universität München, Munich, Germany, 2002, pp. 15–26.
36. R. Windhorst, M. D. Ardema, J. V. Bowles: *Minimum heating reentry trajectories for advanced hypersonic launch vehicles*, Proc. of the AIAA Guidance, Navigation, and Control Conference, New Orleans, USA, 1997. AIAA-97-3535.
37. N. X. Vinh, A. Busemann, R. D. Culp: *Hypersonic and Planetary Entry Flight Mechanics*. The University of Michigan Press, Ann Arbor, 1980.

Chapter 9

Variational Approaches to Fracture

Gianpietro Del Piero

Abstract Fracture mechanics is largely based on an assumption of A.A. Griffith, according to which the evolution of a crack in a fractured body is governed by the competition between the elastic strain energy and the energy spent in the crack growth. In the recent variational formulation of Francfort and Marigo, some analytical difficulties due to the presence of discontinuous displacements are avoided using the Γ -convergence theory. In it, the problem is approximated by a sequence of more regular problems, formulated in Sobolev spaces, and therefore solvable with standard finite elements techniques. A further regularization introduced by G.I. Barenblatt allows to solve the problem of determining the fracture onset in an initially unfractured body.

In this communication I discuss the role played by local energy minimizers in the post-fracture evolution. I show that a dissipation inequality leads to the formulation of an incremental problem, which determines a quasi-static evolution along a path made of local energy minimizers. The introduction of a dissipation inequality provides different responses at loading and unloading, in accordance with experimental evidence. Finally, I briefly discuss the possibility of bulk regularization, based on the assumption that the fracture energy be diffused in a three-dimensional region, instead of being concentrated on a singular surface.

9.1 Fracture as a Minimum Problem

The variational approach to fracture traces back to the paper of A.A. Griffith [21], in which the propagation of a crack is governed by the competition between elastic strain energy and fracture energy. Of the many basic questions left unsolved in the subsequent development of fracture mechanics, some are discussed in the paper by

Gianpietro Del Piero
Dipartimento di Ingegneria, Università di Ferrara, Ferrara, Italy.
e-mail: dlpgpt@unife.it

Francfort and Marigo [19]. They motivate a thorough revisiting of the formulation of the problem of fracture.

In Francfort and Marigo's re-formulation, the energy of a body Ω is assumed to be the sum of three contributions: the elastic strain energy, the energy of the loads, and the fracture energy:

$$E(u) = \int_{\Omega} w(\nabla u(x)) dx - \ell(u) + \gamma |\mathcal{S}(u)|. \quad (9.1)$$

Here w is the strain energy density, ℓ is the external load, and γ is the fracture energy per unit area (fracture toughness). Contrary to a standard assumption of continuum mechanics, the domain of E includes discontinuous displacement fields, in which all discontinuities are supposed to be of the jump type, see [1] or [24] for a precise definition. The set $\mathcal{S}(u)$ of all discontinuity points of u is called the jump set of u , and $|\mathcal{S}(u)|$ is the area of the jump set.

As formulated in [19], the problem of fracture consists in finding the global minimum of E under a given load ℓ . If the undeformed body Ω is not fractured, it is expected that the global minimizers for E be unfractured configurations for small ℓ and fractured configurations for sufficiently large ℓ . The fracture onset can be identified with the transition from the unfractured to the fractured regime, and the fractured minimizers describe the evolution of the fracture.

In fact, as we shall see, this picture is a bit too simple. Anyway, it reveals the evolutionary nature of the problem of fracture, since it requires the specification of a *loading program* $t \mapsto \ell_t$, with t a scalar *loading parameter*, to determine a corresponding family $t \mapsto u_t$ of minimizers for the functional (9.1). Even at this very basic level, a physical problem raises to complicate the picture. It is common experience that, in most materials, fractures do not heal: if during the evolution of the fracture the jump amplitude goes back to zero, the original unfractured configuration is not restored. This fact is accounted for by the supplementary condition

$$t > t^* \implies \mathcal{S}(u_{t^*}) \subseteq \mathcal{S}(u_t), \quad (9.2)$$

which is a mathematical description of the *irreversibility of fracture*. As a consequence of this assumption, the domain of E at t now depends on the solutions of the minimum problem at all $t^* < t$. Therefore, the family of the minimum problems, one for each t , is now transformed into a single evolutionary problem.

In what follows it is assumed that there is no load and that the body is subject to given displacements $\hat{u}(x)$ at a portion Γ of the boundary $\partial\Omega$. The displacements vary according to a given program $t \mapsto \hat{u}_t$, with t playing the role of the loading parameter.

There was a good reason for selecting this specific boundary condition [10, 19]. Assume that, for a given ℓ , the domain of E is unbounded in some direction u such that $\ell(u) > 0$. This occurs, for example, if there are surface tractions with a positive component in the direction of the outward normal to $\partial\Omega$. In this case u , and therefore $\ell(u)$, can take arbitrarily large values. Then there is no global minimizer, since the infimum of $E(u)$ is $-\infty$. As we will see later, the non-existence of global minimizers does not imply the non-existence of solutions for the fracture problem

if a *local minimum* strategy is adopted. But, before doing this, let us give a look to the numerical solution of the global minimum problem.

9.2 The Numerical Solution

In the absence of loads, one has to minimize the functional

$$E(u) = \int_{\Omega} w(\nabla u(x)) dx + \gamma |\mathcal{S}(u)|, \quad (9.3)$$

in the set of all displacement fields u which satisfy the boundary condition

$$u(x) = \hat{u}_t(x), \quad x \in \Gamma, \quad (9.4)$$

on a given portion Γ of the boundary and with the jump set restricted by the irreversibility condition

$$\mathcal{S}(u) \supseteq \mathcal{S}_t, \quad (9.5)$$

where \mathcal{S}_t depends on the previous history of the constraint law $t \mapsto \hat{u}_t$, through the solutions u_t of the corresponding minimum problems.

An appropriate function space for the domain of E is the set $SBV(\Omega)$ of all *special functions of bounded variation* [12], where *special* means that all discontinuities are of the jump type. In view of a numerical solution, the functional E can be approximated by a family of functionals defined in the Sobolev spaces $W^{1,p}(\Omega)$, using a result based on De Giorgi's Γ -convergence theory [5, 11].

A family $\varepsilon \mapsto E_\varepsilon$ of functionals is said to Γ -converge to E if the two following conditions are satisfied:

- (i) (semicontinuity) for every u in the domain of E and for every family $\varepsilon \mapsto u_\varepsilon$ converging to u ,

$$\liminf_{\varepsilon \rightarrow 0} E_\varepsilon(u_\varepsilon) \geq E(u),$$

- (ii) (*limsup inequality*) for every u in the domain of E there is at least one of such families such that

$$\limsup_{\varepsilon \rightarrow 0} E_\varepsilon(u_\varepsilon) \leq E(u).$$

An important property of Γ -convergence is that every converging sequence made of global minimizers of E_ε converges to a global minimizer of E . Thus, if a family of regularized functionals E_ε Γ -converges to E , a global minimizer for E is approximated by a global minimizer for E_ε with sufficiently small ε . In particular, if E_ε is defined on a Sobolev space, its global minimizers can be approximated by standard finite element procedures.

For the fracture problem, consider the functionals

$$E_\varepsilon(u, s) = \int_{\Omega} (s^2(x) + k_\varepsilon) w(\nabla u(x)) dx + \frac{\gamma}{2} \int_{\Omega'} (\varepsilon |\nabla s(x)|^2 + \frac{1}{\varepsilon} (1 - s(x))^2) dx, \quad (9.6)$$

with u and s in the Sobolev spaces $W^{1,p}(\Omega, \mathbb{R}^3)$ and $W^{1,p}(\Omega, [0, 1])$, respectively. The coefficient k_ε is a positive constant of order $o(\varepsilon^{p-1})$ which guarantees the coerciveness of the functional, and s is a scalar field which provides a regularized representation of fracture, in the sense that the extreme values $s = 1$ and $s = 0$ denote total absence of fracture and complete fracture, respectively.

The functionals (9.6) were introduced by Ambrosio and Tortorelli [2] to solve numerically the image segmentation problem of Mumford and Shah [23]. This problem can be viewed as a special case of the problem of fracture, with the following restrictions:

- (a) the body is two-dimensional,
- (b) the function w is quadratic, and
- (c) u is scalar valued.

In this restricted context, the Γ -convergence of the family (9.6) was proved in [2]. This result was used by Bourdin, Francfort, and Marigo [6] to obtain numerical solutions for the fracture problem in the case of antiplane shear. The extension to more general fracture problems was not trivial. In the same paper [6], a numerical solution was given for a problem of two-dimensional linear elasticity, with a vector-valued u . In doing this, they assumed the Γ -convergence of the family $\varepsilon \rightarrow E_\varepsilon$ in the vectorial case, a property which was proved only later [7].

In [17], Del Piero, Lancioni, and March considered problems in finite two-dimensional elasticity, removing the restrictions (b) and (c). They assumed the polyconvex energy density w of a compressible Blatz–Ko material [9]. In this case, too, the required Γ -convergence results were only partially known: while the semicontinuity property (i) for the polyconvex vectorial case had been proved in [20], the proof of the *limsup* condition (ii) is still an open problem.

A remarkable result in [6] is a prediction of the fracture onset in an initially unfractured body, and the main progress made in [17] was the elimination of the embarrassing symmetry of the responses in tension and in compression occurring in the linear elastic case. Anyway, as we shall see in the next section, neither the predictions given in [6] nor those in [17] are fully satisfactory.

9.3 Energy Barriers and Local Minima

To see why the predictions for crack initiation based on global energy minimization are unreliable, consider the simplest problem of a one-dimensional body made of a linear elastic material. In this problem w is quadratic, Ω reduces to a segment $(0, l)$, and the area $|\mathcal{S}(u)|$ is replaced by the number $\#u$ of the jumps:

$$E(u) = \frac{1}{2} \int_0^l k u'^2(x) dx + \gamma \#u. \quad (9.7)$$

Here k is a positive material constant (Young's modulus). The boundary condition (9.4) is replaced by the end conditions $u(0) = 0$ and $u(l) = \beta l$, where $\beta \geq 0$ is

the prescribed average elongation of the bar. This boundary condition can be more conveniently written in the form

$$\int_0^l u'(x) dx + \sum_{i=1}^{\#u} [u](x_i) = \beta l. \quad (9.8)$$

In this section we consider a monotonically increasing elongation, and we identify β with the loading parameter t . It is an elementary exercise in the calculus of variations to prove that, for every β and for any fixed value of the sum of the jump amplitudes, the minimum of E is attained at a configuration with constant u' . Then the boundary condition reduces to

$$u' + l^{-1} \sum_{i=1}^{\#u} [u](x_i) = \beta. \quad (9.9)$$

By substitution into the expression (9.7) of the energy, one gets the unconstrained minimum problem for the functional

$$E(u) = \frac{1}{2} kl \left(\beta - l^{-1} \sum_{i=1}^{\#u} [u](x_i) \right)^2 + \gamma \#u, \quad (9.10)$$

which now only depends on the number of the jumps and on the sum of their amplitudes. The minimum of E is easily determined for every fixed value of $\#u$. Indeed, if there are no jumps one simply has $E(u) = \frac{1}{2} kl \beta^2$, and for $\#u > 0$ the minimum is $E(u) = \gamma \#u$, with arbitrary positions and amplitudes of the jumps, but with the sum of the jump amplitudes equal to βl . These minima are plotted in Fig. 9.1a as functions of β , and the smallest of them is the global minimum for the given β .

From the figure we see that there are only two types of global minima, those with $\#u = 0$ and those with $\#u = 1$. The first occur for $\beta < \beta_o$ and the second for $\beta > \beta_o$; the transition value

$$\beta_o = \sqrt{\frac{2\gamma}{kl}} \quad (9.11)$$

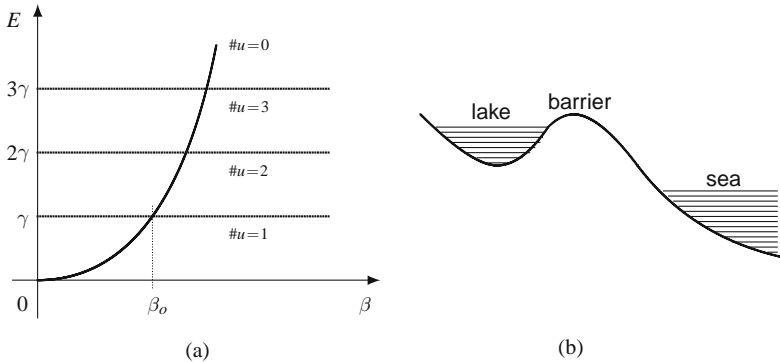


Fig. 9.1 Local minima for the uniaxial stretching of a bar (a) and an example of metastable equilibrium (b)

is obtained by equating the energy minima for $\#u = 0$ and $\#u = 1$. If fracture is determined by the global energy minimum, one concludes that β_o marks the onset of fracture. But a more accurate analysis shows that this conclusion may be uncorrect.

Let us identify the configuration space with the domain of the functional, and let us define a metric on it. In general, the choice of a metric is not only a matter of mathematical convenience. It also reflects specific assumptions on the structure of the continuum, see [14, Remark 5.4]. In particular, in a configurations space including configurations with jumps, the metric of the *total variation*

$$\|u\| = \int_0^l |u'(x)| dx + \sum_{i=1}^{\#u} |[u](x_i)|. \quad (9.12)$$

is inappropriate to the case in which the jumps represent material defects, for example, dislocations, which can move freely across the body, and is appropriate to the case in which the jumps represent the fractures in a fractured continuum, see [16, Section 6.3]. To clarify this point, consider the two following configurations:

- the unfractured configuration u_1 with $u'_1 = \beta_o$ and
- the fractured configuration u_2 with $u'_2 = 0$, $\#u_2 = 1$, and $[u_2] = l\beta_o$.

Their distance

$$\|u_2 - u_1\| = lu'_1 + [u_2] = 2l\beta_o \quad (9.13)$$

is finite, though the two configurations correspond to the same point in the energy-elongation plane of Fig. 9.1a. The transition from u_1 to u_2 at $\beta = \beta_o$ can take place only if there is a configuration path connecting u_1 and u_2 , continuous with respect to the metric $\|\cdot\|$ and with non-increasing energy. But this is not the case: in any continuous path from u_1 to u_2 , at a certain point a crack must open, and this requires an upward jump of the total energy of the order of γ . In other words, the transition from u_1 to u_2 is prevented by an *energy barrier* of height γ . A barrier of the same height prevents the transition from unfractured to fractured configuration at any $\beta > \beta_o$, though the energy of the fractured configuration is now smaller than the energy of the unfractured configuration with the same β .

Therefore, the bar follows the curve $\#u = 0$ forever. This conclusion can be extended to arbitrary loads and to higher-dimensional bodies of any shape [8]. It makes evident the failure of Griffith's theory in predicting the fracture onset [19], and explains why the predictions on the fracture onset made in [6, 17, 19] are unreliable. It also shows the irrelevance of the existence of global minimizers for the given loads, as long as there is an equilibrium path made of local minimum configurations.

The presence of energy barriers suggests that global energy minimization can be conveniently replaced by a *local minimum strategy*. In it, during its evolution $\beta \mapsto u(\beta)$ the body tends to follow a continuous curve in the configuration space, all points of which are local energy minimizers for the corresponding β . Within this strategy, a prediction of fracture onset is obtained by a proper regularization of Griffith's theory.

In a sense, that nature prefers a local minimum strategy is a fortune for everyday's life. In a lake located above the sea level, Fig. 9.1b, the water is in a situation of a

local minimum (*metastable equilibrium*), since it would like to join the underlying sea. This is prevented by the barrier constituted by the banks of the lake. In the very same way, most of everyday's life objects are in metastable equilibrium, because they admit a broken configuration with lower energy. Fortunately, they can reach it only if the energy barrier is somehow removed.

9.4 Barenblatt's Regularization

For the regularization of Griffith's energy, Barenblatt's idea [3] is to assume a continuous dependence of the fracture energy on the jump amplitude $[u]$. In fact, Griffith's energy is a discontinuous function of $[u]$, since it is equal to zero for $[u] = 0$ and to γ for $[u] \neq 0$. Griffith's and Barenblatt's energies are plotted in Fig. 9.2a and b. Figure 9.2c shows the special case of Dugdale's energy

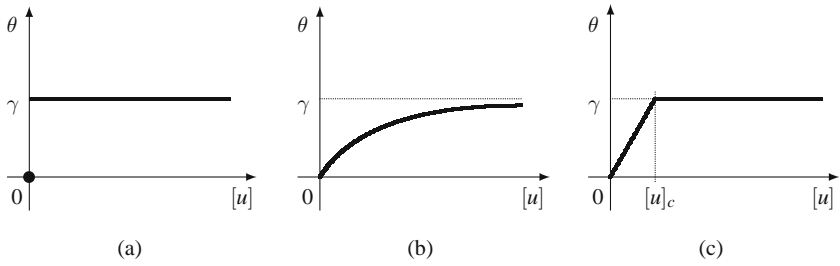


Fig. 9.2 The fracture energies of Griffith (a), Barenblatt (b), and Dugdale (c)

$$\theta([u]) = \begin{cases} \gamma \frac{[u]}{[u]_c} & \text{if } 0 \leq [u] < [u]_c \\ \gamma & \text{if } [u] \geq [u]_c \end{cases}, \quad (9.14)$$

with γ and $[u]_c$ positive constants. Due to its analytical simplicity, this is the expression of the energy I am going to use in the following. To avoid interpenetration, I assume that the jump amplitudes are non-negative:

$$[u] \geq 0. \quad (9.15)$$

For the moment, the question of the irreversibility of fracture is ignored. It will be discussed later. Consider a one-dimensional, linear elastic bar, in which the energy has the expression

$$E(u) = \frac{1}{2} \int_0^l k u'^2(x) dx + \sum_{i=1}^{\#u} \theta([u](x_i)), \quad (9.16)$$

with θ as in (9.14).

Just as in Griffith's case, it can be proved that u' is constant in each local or global minimizer. It can also be proved that if θ is concave, as it occurs in all energies shown in Fig. 9.2, then all equilibrium configurations u with $\#u > 1$ are unstable [18]. Then the only two cases of interest are $\#u = 0$ and $\#u = 1$. For them, the total energy reduces to

$$E(u) = \frac{1}{2} k l u'^2 + k \beta_c \min\{[u], [u]_c\}, \quad (9.17)$$

with

$$\beta_c = \frac{\gamma}{k[u]_c}, \quad (9.18)$$

and the boundary condition (9.9) reduces to

$$u' + l^{-1}[u] = \beta. \quad (9.19)$$

A configuration of the body is described by the two scalar variables u' and $[u]$. Due to the non-interpenetration condition (9.15), it coincides with the portion $[u] \geq 0$ of the plane $(u', [u])$.

The level curves for the energy E and for the load β are represented in Fig. 9.3. From (9.17) it follows that the curves of constant energy are parabolas with vertex on the horizontal axis in the strip $0 \leq [u] < [u]_c$ and horizontal lines in the half plane $[u] \geq [u]_c$. Moreover, by the boundary condition (9.19), the curves of constant load are parallel lines with slope $-l^{-1}$.

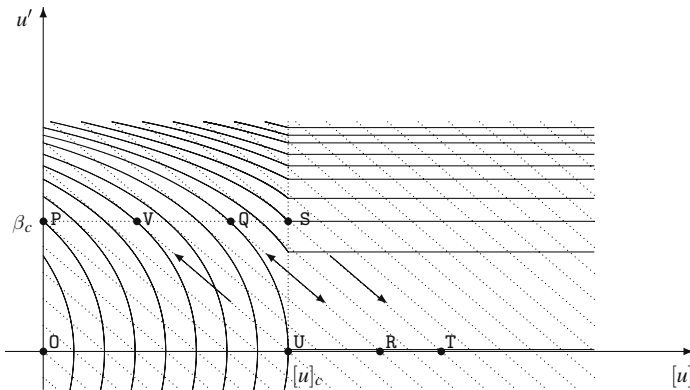


Fig. 9.3 Curves of constant energy (*solid lines*) and of constant load (*dotted lines*) in the configuration space. The *double arrow* shows the transition from partial to complete fracture according to the global minimum strategy, and the *two single arrows* show the same transition according to the local minimum strategy

In each load line, the (global or local) minima may be located either at points at which the load line is tangent to an energy line, or on the boundary $[u] = 0$ of the configuration space, or on the portion $\{u' = 0, [u] > [u]_c\}$ of the horizontal axis. The evolution of the system under a given load program is determined in the next section.

9.5 Two Solution Strategies

Consider a load program $t \mapsto \beta(t)$, where β is the average elongation of the bar and t is the loading parameter. We say that at a given t the load program is in a regime of *loading* if $\dot{\beta}(t) > 0$ and that it is in a regime of *unloading* if $\dot{\beta}(t) < 0$, where the dot denotes differentiation with respect to t . Let us determine first the evolution of the system at loading, starting from the origin 0. To do this, in Fig. 9.3, we follow the evolution of the minima on load lines corresponding to increasing β .

For small β the global minimum stays on the vertical axis, corresponding to the unfractured configurations $\{[u] = 0, u' = \beta\}$, until the loading line becomes tangent to one of the parabolas. This occurs at the point P at which $\beta = \beta_c$. For larger values of β the points on the vertical axis become maxima, while the global minima are located on the horizontal line $u' = \beta_c$. The system then leaves the vertical axis and follows this line.

On this line there is a configuration Q whose energy is the same as the energy on the horizontal half line $\{u' = 0, [u] \geq [u]_c\}$. Therefore, on the load line from Q there are two global minimizers, Q and R. For larger β , the points on the line $u' = \beta_c$ are local minimizers, and the global minimizers are located on the horizontal axis $u' = 0$.

According to the *global minimum* strategy followed in [17, 19], after reaching the point Q the system jumps to R and then follows the horizontal axis. But this prediction ignores the energy barrier between Q and R. It looks then more reasonable to follow the *local minimum* strategy, according to which the system follows the line $\beta = \beta_c$ of the local minima up to the point S = $([u]_c, \beta_c)$. At this point the load line has only one minimum at the point T of the horizontal axis. The system is then obliged to jump from S to T and then to follow the horizontal axis.

It must be said that this jump is a weak point of the local minimum strategy. Indeed, while in our one-dimensional example the load line from S has a single minimum and a single minimizing path, in higher dimension there may be a multiplicity of minimizing paths, and the model does not provide any criterion for selection. Moreover, if we identify t with the physical time we see that the transition from S to T is instantaneous, and it is hard to believe that an instantaneous transition does not involve any dynamic effect.

Consider now the unloading from the point T. In the global minimum strategy, the system traverses backward the path (OPQRT) followed at loading. In the local minimum strategy, the line $u' = 0$ of the local minima is followed up to the point U = $([u]_c, 0)$. Then the system jumps back to the global minimum point V which, depending on the slope of the load line,¹ may be located either on the line $u' = \beta_1$ or on the boundary line $[u] = 0$. After that, the system follows backward the loading path (OPV). Thus, in both cases the unloading process ends up at the initial unfractured configuration 0, which means that the fracture is not irreversible. Therefore, in both strategies a condition preventing the fracture healing is missing. Moreover,

¹ For a discussion of the *size effect* in the mechanics of fracture, determined here by the slope l^{-1} of the load lines, see, e.g., [4, 13].

in both strategies there is a transition in which the force jumps back from zero to $k\beta_c$. This upward jump of the force has never been observed in experiments.

In conclusion, the main progress brought by Barenblatt's model over Griffith's is a prediction for both fracture initiation and complete fracturing. Indeed in both strategies, global and local, the departure from the axis $[u] = 0$ at P characterizes the fracture onset, and the transition at which the axial force $\sigma = ku'$ instantaneously drops from $k\beta_c$ to zero is identified with complete fracture. However, Barenblatt's model is not fully satisfactory, since it predicts fracture healing and an upward jump of the force at unloading. For these reasons, a further refinement is needed.

9.6 The Dissipative Model

Calling *energies* the energetic terms of Griffith and Barenblatt suggests that they are recoverable. In fact, in the Barenblatt model the recovering is total in the global minimum strategy, in which the loading and unloading curves coincide, and is partial in the local minimum strategy, in which the hysteresis cycle (VSTUV) takes place. In the Griffith model, recovering is forbidden by the irreversibility condition (9.2) and not by a direct restriction on the recoverability of the energy.

A way for eliminating the drawbacks in the response of Barenblatt's model pointed out at the end of the previous section is to assume that the fracture energy θ is not recoverable [18]. To see the far-reaching consequences of this approach, consider again the problem of the homogeneous deformation of a bar subject to the boundary condition (9.19), with β varying according to a loading program $\beta = \beta(t)$. For it, define

- the elastic energy: $\Psi = \frac{1}{2}klu'^2$,
- the dissipation: $D = \theta([u])$, and
- the external power: $P = \sigma\dot{\beta}l$,

with θ as in (9.17). Any deformation process $t \mapsto (u'(t), [u](t))$ experienced by the bar must satisfy the power equation

$$P = \dot{\Psi} + \dot{D},$$

and the dissipation inequality

$$\dot{D} \geq 0.$$

These are the requirements imposed by the first and the second law of thermodynamics in this very simple context. By using the expressions of Ψ , D , and P and the boundary condition, the two conditions take the form

$$(\sigma - ku')lu' + (\sigma - \theta'([u]))[\dot{u}] = 0, \quad \theta'([u])[\dot{u}] \geq 0. \quad (9.20)$$

Moreover, because u' and $[u]$ can be varied independently, from the first equation one gets the force-elongation relation $\sigma = ku'$, and the above conditions reduce to

$$ku'[\dot{u}] = \theta'([u])[\dot{u}] \geq 0. \quad (9.21)$$

The evolution of u' and U is governed by a minimum principle, which can be stated as follows.² In a given load program $t \mapsto \beta(t)$, let $(u'(t), [u](t))$ be the configuration at t , let

$$E(t) = \Psi(u'(t)) + D([u](t)) \quad (9.22)$$

be the sum of the elastic energy and of the dissipation at t , and let $\tau > 0$ be a small increment of t . Then $(u'(t + \tau), [u](t + \tau))$ is the configuration which minimizes the functional E , subject to the dissipation inequality (9.20)₂ and to the boundary condition

$$u'(t + \tau) + l^{-1}[u](t + \tau) = \beta(t + \tau). \quad (9.23)$$

Notice that E coincides with the total energy E of the non-dissipative model and that the only difference in the two problems is the presence of the dissipation inequality in the dissipative model.

From the expressions of Ψ and D and from the expansions of $u'(t + \tau)$ and $[u](t + \tau)$ it follows that

$$\begin{aligned} E(t + \tau) = & \frac{1}{2}kl(u'^2 + 2\tau u'\dot{u}' + \tau^2(u'\ddot{u}' + \dot{u}'^2) + o(\tau^2)) \\ & + k\beta_c \min\{[u] + \tau[\dot{u}] + \frac{1}{2}\tau^2[\ddot{u}] + o(\tau^2), [u]_c\}, \end{aligned} \quad (9.24)$$

with u' , $[u]$, and their derivatives evaluated at t , and with

$$\dot{u}' + l^{-1}[\dot{u}] = \dot{\beta}, \quad \ddot{u}' + l^{-1}[\ddot{u}] = \ddot{\beta}. \quad (9.25)$$

Let us minimize E for configurations located at each of the three equilibrium lines of the non-dissipative model.³ On the line $\{[u] = 0, u' < \beta_c\}$ we have $u' = \beta$ and

$$E(t + \tau) = \frac{1}{2}kl\beta^2 + k\tau(\beta l \dot{u}' + \beta_c [\dot{u}]) + o(\tau). \quad (9.26)$$

Because β is fixed, it is sufficient to minimize

$$\beta l \dot{u}' + \beta_c [\dot{u}] = \beta l \dot{\beta} + (\beta_c - \beta)[\dot{u}]. \quad (9.27)$$

But $\dot{\beta}$ is fixed, β_c is greater than β by assumption, and $[\dot{u}] \geq 0$ because this is the only increment allowed at the boundary of the configuration space. Then the minimum is attained for $[\dot{u}] = 0$ and $\dot{u}' = \dot{\beta}$, which means continuation along the line $[u] = 0$.

On the horizontal segment $\{u' = \beta_c, 0 < [u] < [u]_c\}$, one has $\beta > \beta_c$ and

² This principle traces back to R. Hill's *principle of maximum plastic work* [22]. For the subsequent developments see [18] and the references cited therein.

³ These are not all possible equilibrium configurations for the dissipative model, because one of the effects of assuming θ dissipative is the creation of many new equilibrium configurations, see [18].

$$\begin{aligned}
& E(t + \tau) \\
&= \frac{1}{2} kl \beta_c^2 + k \beta_c [u] + k \tau \beta_c (l \dot{u}' + [\dot{u}]) + \frac{1}{2} k \tau^2 (\beta_c (l \ddot{u}' + [\ddot{u}]) + l \dot{u}'^2) + o(\tau^2) \quad (9.28) \\
&= E(t) + k \tau l \beta_c \dot{\beta} + \frac{1}{2} k l \tau^2 (\beta_c \ddot{\beta} + \dot{u}'^2) + o(\tau^2),
\end{aligned}$$

and because $E(t)$, $\dot{\beta}$, and $\ddot{\beta}$ are fixed, one is reduced to minimizing

$$l \dot{u}'^2 = (l \dot{\beta} - [\dot{u}])^2, \quad (9.29)$$

under the dissipation inequality, which here reduces to $[\dot{u}] \geq 0$ because $\theta'([u]) > 0$ for $0 < [u] < [u]_c$. The solution is

$$\begin{aligned}
\dot{u}' &= 0, \quad [\dot{u}] = \dot{\beta} l \quad \text{if } \dot{\beta} > 0, \\
\dot{u}' &= \dot{\beta}, \quad [\dot{u}] = 0 \quad \text{if } \dot{\beta} \leq 0.
\end{aligned} \quad (9.30)$$

Finally, on the half line $\{u' = 0, [u] > [u]_c\}$ one has

$$E(t + \tau) = \frac{1}{2} kl \tau^2 \dot{u}'^2 + o(\tau^2) + k \beta_c [u]_c, \quad (9.31)$$

and the problem is again reduced to minimizing the function (9.29). But now the dissipation inequality is identically satisfied as an equality, because $\theta'([u]) = 0$ for $[u] > [u]_c$. Therefore there is no constraint on $[\dot{u}]$, and the minimum is achieved for the increments

$$\dot{u}' = 0, \quad [\dot{u}] = \dot{\beta} l, \quad (9.32)$$

which correspond to continuation along the horizontal axis.

In Fig. 9.4a the directions of the continuations are shown in the configuration space, and the corresponding force-elongation incremental responses are represented in Fig. 9.4b. From the first picture we see that before fracture onset, segment $\{[u] = 0, u' < \beta_c\}$, the response is the same as in the non-dissipative model. Between fracture onset and complete fracture, segment $\{u' = \beta_c, 0 < [u] < [u]_c\}$, the response at loading is the same. At unloading, the non-dissipative model traces back the same segment, while the dissipative model follows the line $[\dot{u}] = 0$. The second

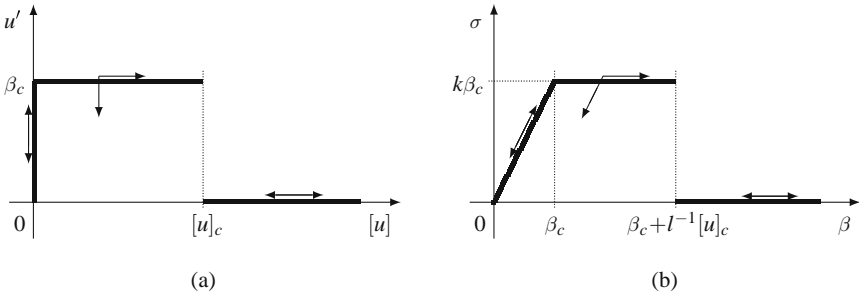


Fig. 9.4 Incremental loading-unloading response of the dissipative model, represented in the configuration space (a) and on the force-elongation curve (b)

picture shows that the response for $\beta_c < \beta < \beta_c + l^{-1}[u]_c$ coincides with the familiar *plastic loading – elastic unloading* response of perfect plasticity [15, 18]. Finally, after complete fracture, half line $\{u' = 0, [u] > [u]_c\}$, the response goes back and forth along the horizontal axis.

The equilibrium configurations in the dissipative model which are not equilibrium configurations in the non-dissipative model are those in the rectangle $\{0 < [u] < [u]_c, 0 < \beta < \beta_c\}$. They exhibit an elastic response, $\dot{u} = 0, \dot{u}' = \dot{\beta}$, at both loading and unloading. Therefore, in no case the dissipative model does exhibit the upward jump of the stress at unloading predicted by the non-dissipative model.

9.7 From Surface to Bulk Regularization

After removing the assumption that θ is recoverable, let us turn our attention to a second prejudice, whose removal opens the way to further new perspectives. This is the idea that the fracture energy, or dissipation as we may now call it, has a surfacic nature.

As discussed in Sect. 9.2, a three-dimensional fracture problem can be solved by bulk regularization, that is, by approximating the jumps by displacement fields defined all over the body. Here I examine the possibility of regularizing the model instead of its numerical approximation. This can be done by assuming that the fracture dissipation has a three-dimensional density and by interpreting the discontinuous fractured configurations as singular solutions due to strain localization.

Let me borrow from plasticity the idea that the deformation can be decomposed locally into an elastic and a plastic part. For simplicity I remain within the one-dimensional context, and I assume the additive decomposition

$$u' = e + p$$

of the deformation gradient u' , with the elastic part e and the plastic part p associated with an elastic energy density $w = w(e)$ and a dissipation density, respectively. The dissipation density need not be a function of the current configuration; what is essential is to assume that the *dissipation rate* φ be determined by the current values of the plastic strain and of the past history of the plastic strain rate:

$$\varphi(x, t) = \varphi((p(x, s), s \in [0, t]), \dot{p}(x, t)).$$

Specifically, I assume that w is quadratic:

$$w(e) = \frac{1}{2} k e^2, \quad (9.33)$$

and that the dissipation rate has the form

$$\varphi(p_m, \dot{p}) = \begin{cases} \sigma_o |\dot{p}| & \text{if } p_m < p_o \\ 0 & \text{if } p_m \geq p_o \end{cases}, \quad (9.34)$$

where p_m is a *state variable* representing the largest value taken by p in the previous deformation history:

$$p_m(x, t) = \max_{s \in [0, t]} p(x, s), \quad (9.35)$$

and σ_o and p_o are positive material constants, characterizing the first yielding and the complete fracture of the bar, respectively. Notice that the fact that φ is homogeneous of degree one with respect to \dot{p} characterizes the material as *rate independent*.

In a load-free bar subject to the boundary condition

$$\int_0^l (e(x, t) + p(x, t)) dx = l\beta(t), \quad (9.36)$$

the external power is again $\sigma l\dot{\beta}$, the power equation becomes

$$\sigma(t)l\dot{\beta}(t) = \int_0^l (ke(x, t)\dot{e}(x, t) + \varphi(p_m(x, t), \dot{p}(x, t))) dx, \quad (9.37)$$

and the dissipation inequality $\varphi \geq 0$ is automatically satisfied for φ as in (9.34). Combining the power equation with the boundary condition yields

$$\int_0^l ((\sigma(t) - ke(x, t))\dot{e}(x, t) + \sigma(t)\dot{p}(x, t) - \varphi(p_m(x, t), \dot{p}(x, t))) dx = 0, \quad (9.38)$$

and because \dot{e} is a free parameter we get $\sigma(t) = ke(x, t)$ for all x , and

$$\sigma(t)\dot{p}(x, t) = \varphi(p_m(x, t), \dot{p}(x, t)) \geq 0. \quad (9.39)$$

The evolution is still governed by the minimization of the (energy + dissipation) functional, which here takes the form

$$E(t + \tau) = \frac{1}{2}kl e^2(t + \tau) + \int_t^{t+\tau} \int_0^l \varphi(p_m(x, s), \dot{p}(x, s)) dx ds. \quad (9.40)$$

From the expansions of e and p it follows that

$$\begin{aligned} E(t + \tau) &= E(t) + \tau \int_0^l (\sigma \dot{e}(x) + \varphi(p_m(x), \dot{p}(x))) dx \\ &\quad + \frac{1}{2} \tau^2 \int_0^l (\sigma \ddot{e}(x) + k\dot{e}^2(x) + \varphi_{p_m}(x)\dot{p}(x) + \varphi_{\dot{p}}(x)\ddot{p}(x)) dx + o(\tau^2), \end{aligned} \quad (9.41)$$

with σ , e , $p(x)$, $\varphi(x)$, and their derivatives evaluated at t , and that

$$\int_0^l (\dot{e}(x) + \dot{p}(x)) dx = l\dot{\beta}, \quad \int_0^l (\ddot{e}(x) + \ddot{p}(x)) dx = l\ddot{\beta}. \quad (9.42)$$

I point out that for φ as in (9.34) the derivative φ_{p_m} vanishes everywhere except at the singular point $p_m = p_o$ and that $\varphi_{\dot{p}}$ is piecewise constant.

Let us look at the solution of the minimum problem in the two cases:

- (i) $p_m(x) < p_o$ almost everywhere in $(0, l)$ and
- (ii) $p_m(x) \geq p_o$ in a set \mathcal{J} with positive measure,

which correspond to the regime before and after complete fracture, respectively. In the first case we have $\varphi(p_m(x), \dot{p}(x)) = \sigma_o |\dot{p}(x)|$. Assume first that $|\sigma| < \sigma_o$. Then $\dot{p}(x) = 0$ at all x by (9.39). Moreover, by continuity, the inequality $|\sigma| < \sigma_o$ holds

for a finite time interval after t , and because $\dot{p}(x) = 0$ during this interval, one also has $\ddot{p}(x) = 0$ at t . Equation (9.41) then reduces to

$$\begin{aligned} E(t + \tau) &= E(t) + \tau \sigma \int_0^l \dot{e}(x) dx + \frac{1}{2} \tau^2 \int_0^l (\sigma \ddot{e}(x) + k \dot{e}^2(x)) dx + o(\tau^2) \\ &= E(t) + \tau \sigma l \dot{\beta} + \frac{1}{2} \tau^2 (\sigma l \ddot{\beta} + \int_0^l k \dot{e}^2(x) dx) + o(\tau^2), \end{aligned} \quad (9.43)$$

with the second inequality due to conditions (9.42) with $\dot{p}(x) = \ddot{p}(x) = 0$.

Then one has to minimize the integral of $\dot{e}^2(x)$, under the condition that the integral of $\dot{e}(x)$ be equal to $l\dot{\beta}$. The result is $\dot{e}(x) = \dot{\beta}$. Thus, when $|\sigma| < \sigma_o$ the system evolves along homogeneous deformations and without dissipation.

Now consider the case $\sigma = \sigma_o$. In it, (9.39) is satisfied by any $\dot{p}(x) \geq 0$, and (9.41) reduces to

$$\begin{aligned} E(t + \tau) &= E(t) + \tau \sigma_o \int_0^l (\dot{e}(x) + \dot{p}(x)) dx \\ &\quad + \frac{1}{2} \tau^2 \int_0^l (\sigma_o (\ddot{e}(x) + \ddot{p}(x)) + k \dot{e}^2(x)) dx + o(\tau^2). \end{aligned} \quad (9.44)$$

One has again to minimize the integral of $\dot{e}^2(x)$, but now under the condition

$$\int_0^l \dot{e}(x) dx = l\dot{\beta} - \int_0^l \dot{p}(x) dx \leq l\dot{\beta}. \quad (9.45)$$

The solution is

$$\begin{aligned} \dot{e}(x) &= 0, \quad \int_0^l \dot{p}(x) dx = \dot{\beta}l \quad \text{if } \dot{\beta} > 0, \\ \dot{e}(x) &= \dot{\beta}, \quad \dot{p}(x) = 0 \quad \text{if } \dot{\beta} \leq 0. \end{aligned} \quad (9.46)$$

It corresponds to dissipation at constant σ (*plastic loading*) if $\dot{\beta} < 0$, and to homogeneous evolution without dissipation (*elastic unloading*) if $\dot{\beta} > 0$. A similar result holds for $\sigma = -\sigma_o$.

Up to now, the well-known response of an elastic-perfectly plastic bar has been re-obtained. Now consider case (ii), $p_m(x) \geq p_o$ in \mathcal{J} , and assume first that $\sigma \neq 0$. In this case, from (9.39) we have

$$\dot{p}(x) = 0 \text{ in } \mathcal{J}, \quad |\sigma| = \sigma_o, \quad \sigma \dot{p}(x) \geq 0 \text{ in } (0, l) \setminus \mathcal{J}.$$

Then this case is possible only if $|\sigma| = \sigma_o$ all over the bar. For $\sigma = \sigma_o$, the functional (9.41) reduces to

$$E(t + \tau) = E(t) + \tau \left(\int_0^l \sigma_o \dot{e}(x) dx + \int_{(0, l) \setminus \mathcal{J}} \sigma_o \dot{p}(x) dx \right) + o(\tau). \quad (9.47)$$

Using the boundary condition (9.36), the term between parentheses transforms into

$$\int_0^l \sigma_o \dot{\beta} dx - \int_{\mathcal{J}} \sigma_o \dot{p}(x) dx,$$

and because $\dot{p}(x)$ can take any positive value, the infimum is $-\infty$. Therefore, there is no solution for $\sigma = \sigma_o$. A similar conclusion holds for $\sigma = -\sigma_o$.

For the remaining case $\sigma = 0$, from (9.39) we get $\varphi(p_m(x), p(x)) = 0$ almost everywhere in $(0, l)$, and

$$\dot{p}(x) \text{ arbitrary in } \mathcal{I}, \quad \dot{p}(x) = 0 \text{ in } (0, l) \setminus \mathcal{I}.$$

The functional (9.41) reduces to

$$E(t + \tau) = E(t) + \frac{1}{2} \tau^2 \left(\int_0^l k \dot{e}^2(x) dx + \int_{\mathcal{I}} \sigma_0 (\operatorname{sgn} \dot{p}(x)) \ddot{p}(x) dx \right) + o(\tau^2), \quad (9.48)$$

and one has to minimize the sum of the two integrals. The integrand function of the second integral is non-negative, because it is the second-order term of the expansion of $\varphi(p_m(x, \cdot), \dot{p}(x, \cdot))$ whose first-order term $\sigma_0 |\dot{p}(x)|$ is zero. Consequently, the function to be minimized is non-negative; the infimum, which is zero, is attained for

$$\dot{e}(x) = 0 \text{ in } (0, l), \quad \ddot{p}(x) = 0 \text{ in } \mathcal{I}.$$

Then the system evolves following the line $\sigma = 0$, that is, there is no energy recovering from any completely fractured configuration. The distribution of $\dot{p}(x)$ along the bar is undetermined, except for condition (9.42)₁ requiring that the integral of $\dot{p}(x)$ be equal to $l\dot{\beta}$.

By its very definition (9.35) the function $p_m(x, \cdot)$ is non-decreasing, and it increases only when $p(x) = p_m(x)$ and $\dot{p}(x) > 0$. From the foregoing analysis it emerges that this may occur only in the following two cases:

$$p_m < p_o \text{ and } \sigma = \sigma_o, \quad p_m \geq p_o \text{ and } \sigma = 0.$$

This determines the incremental response represented in Fig. 9.5.

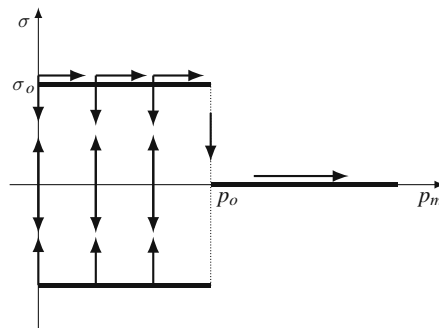


Fig. 9.5 Directions of the local incremental response for the model with bulk regularization

One sees that the only points accessible from the origin $\sigma = p_m = 0$ are those in the rectangle $\{0 \leq p_m < p_o, |\sigma| \leq \sigma_o\}$ and those on the half line $\{p_m \geq p_o, \sigma = 0\}$. The points in the rectangle correspond to elastic-plastic response and those on the half line correspond to complete fracture.

This picture describes the behavior of a single point and not that of the whole bar. Indeed, while σ , and therefore e , are constant over the bar, p need not be so, since in every dissipative process the distribution of the increments \dot{p} along the bar is undetermined. This means that the model is too elementary to completely determine the evolution of the bar. In particular, the local growth of p_m , which determines the total fracture, is totally out of control.

A standard strategy for determining punctually $\dot{p}(x)$ and $\dot{p}_m(x)$ is to add a higher-gradient dissipative term. Specifically, this term can be taken proportional to the square of the spatial gradient p' of p . With this choice, the functional (9.40) is replaced by

$$E(t + \tau) = \frac{1}{2} kl e^2(t + \tau) + \int_t^{t+\tau} \int_0^l \varphi(p_m(x, s), \dot{p}(x, s)) dx ds + \frac{1}{2} \alpha \int_0^l p'^2(x, t + \tau) dx, \quad (9.49)$$

with α a positive material constant. The study of this functional goes beyond the scope of this chapter. Anyway, a comparison with the functional (9.6) of Ambrosio and Tortorelli is very instructive. Indeed, in spite of their differences, due to the fact that one was tailored on the requirements of the Γ -convergence approximation and the other one reflects some physical modeling, they have a similar structure. Both consist of three terms: an elastic energy, a fracture energy now interpreted as dissipation, and a higher-gradient term.

This shows that numerical regularization and physical regularization go in the same direction and suggests the possibility of reducing the fracture problem to a problem of plasticity, with a supplementary element, in our example the state variable p_m , characterizing the occurrence of complete fracture.

Acknowledgments This work was supported by the Research Project *Mathematical Models for Materials Science* - PRIN 2005 of the Italian Ministry for University.

References

1. Ambrosio L., Fusco N., Pallara D.: Functions of Bounded Variation and Free Discontinuity Problems. Oxford Math. Monogr., Oxford University Press, New York (2000)
2. Ambrosio L., Tortorelli M.V.: On the approximation of free discontinuity problems. Boll. Un. Mat. Ital. **6-B**, 105–123 (1992)
3. Barenblatt G.I.: The mathematical theory of equilibrium cracks in brittle fracture. Adv. Appl. Mech. **7**, 55–129 (1962)
4. Bažant Z.P., Chen E.P.: Scaling of structural failure. Appl. Mech. Review **50**, 593–627 (1997)
5. Braides A.: Γ -Convergence for Beginners. Oxford University Press, Oxford (2002)
6. Bourdin B., Francfort G.A., Marigo J.-J.: Numerical experiments in revisited brittle fracture. J. Mech. Phys. Solids **48**, 797–826 (2000)
7. Chambolle A.: An approximation result for special functions with bounded deformation. J. Math. Pures Appl., IX Sér. **83**, 929–954 (2004)
8. Chambolle A., Giacomini A., Ponsiglione M.: Crack initiation in brittle materials. Arch. Rational Mech. Analysis **188**, 309–349 (2008)
9. Ciarlet P.G.: Mathematical Elasticity. Vol. I: Three-Dimensional Elasticity. North-Holland, Amsterdam (1987)
10. Dal Maso G., Toader R.: A model for the quasi-static growth of brittle fractures: existence and approximation results. Arch. Rational Mech. Anal. **162**, 101–135 (2002)
11. De Giorgi E., Franzoni T.: Su un tipo di convergenza variazionale. Atti Accad. Naz. Lincei, Rend. Cl. Sci. Fis. Mat. Natur. **58**, 842–850 (1975)
12. De Giorgi E., Ambrosio L.: Un nuovo funzionale del calcolo delle variazioni. Atti Accad. Naz. Lincei, Rend. Cl. Sci. Fis. Mat. Natur. **82**, 199–210 (1988)

13. Del Piero G.: One-dimensional ductile-brittle transition, yielding, and structured deformations. In: P. Argoul et al. (eds.) *Variations of Domains and Free-Boundary Problems in Solid Mechanics*, 203–210, Kluwer, Dordrecht (1999)
14. Del Piero G.: The energy of a one-dimensional structured deformation. *Math. Mech. Solids* **6**, 387–408 (2001)
15. Del Piero G.: Interface energies and structured deformations in plasticity. In: G. Dal Maso et al. (eds.) *Variational Methods for Discontinuous Structures*, 103–116, Birkhäuser, Basel (2002)
16. Del Piero G.: Foundations of the theory of structured deformations. In: Del Piero G., Owen D.R. (eds.) *Multiscale Modeling in Continuum Mechanics and Structured Deformations*. CISM Courses and Lectures n. 447 (2004)
17. Del Piero G., March R., Lancioni G.: A variational model for fracture mechanics: Numerical experiments. *J. Mech. Phys. Solids* **55**, 2513–2537 (2007)
18. Del Piero G., Truskinovsky L.: Elastic bars with cohesive energy. *Cont. Mech. Thermodynamics*, in press (DOI 10.1007/s00161-009-0101-9) (2009)
19. Francfort G.A., Marigo J.-J.: Revisiting brittle fracture as an energy minimization problem. *J. Mech. Phys. Solids* **46**, 1319–1342 (1998)
20. Fusco N., Leone C., March R., Verde A.: A lower semicontinuity result for polyconvex functionals in *SBV*, *Proc. R. Soc. Edinb. A*, **136**, 321–336 (2006)
21. Griffith A.A.: The phenomenon of rupture and flow in solids. *Phil. Trans. Roy. Soc. London A*, **221**, 163–198 (1920)
22. Hill R.: *The Mathematical Theory of Plasticity*. Oxford University Press (1950). Reprinted in: Oxford Classic Series, Clarendon Press, Oxford (1998)
23. Mumford D., Shah J.: Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. Pure Appl. Math.* **42** 577–685 (1989)
24. Vol’pert A.I., Hudjaev S.I.: *Analysis in Classes of Discontinuous Functions and Equations of Mathematical Physics*. Nijhoff, Dordrecht (1985)

Chapter 10

On the Problem of Synchronization of Identical Dynamical Systems: The Huygens's Clocks

Rui Dilão

Abstract In 1665, Christiaan Huygens reported the observation of the synchronization of two pendulum clocks hanged on the wall of his workshop. After synchronization, the clocks swung exactly in the same frequency and 180° out of phase—anti-phase synchronization. Here, we propose and analyze a new interaction mechanism between oscillators leading to exact anti-phase and in-phase synchronization of pendulum clocks, and we determine a sufficient condition for the existence of an exact anti-phase synchronous state. We show that exact anti-phase and in-phase synchronous states can coexist in phase space, and the periods of the synchronous states are different from the eigenperiods of the individual oscillators. We analyze the robustness of the system when the parameters of the individual pendulum clocks are varied, and we show numerically that exact anti-phase and in-phase synchronous states exist in systems of coupled oscillators with different parameters.

10.1 Introduction

In 26 February 1665, Christiaan Huygens, in a letter to his father [1], reported the observation of the synchronization of two pendulum clocks closely hanged on the wall of his workshop. After synchronization, the clocks swung exactly in the same frequency and 180° out of phase. For attachment distances less than 1 m, the clocks always synchronize with the 180° phase difference. For larger attachment distances (> 4.5 m), synchronization, if it occurs, takes longer times. Huygens also noted that if the two clocks were hanged in such a way that the planes of oscillation of the two pendulums were perpendicular, then synchronization did not occur. The Huygens

Rui Dilão

NonLinear Dynamics Group, Instituto Superior Técnico Av. Rovisco Pais, 1049-001 Lisbon, Portugal, e-mail: rui@sd.ist.utl.pt

observations were the first time that synchronization effects have been described scientifically.

Huygens justified the observed synchronization phenomena by the “sympathy that cannot be caused by anything other than the imperceptible stirring of the air due to the motion of the pendulum.”

In recent years, there has been a growing interest in the detailed analysis of synchronization phenomena, both from the theoretical and the experimental points of views. From the experimental point of view, Bennett et al. [2] built an experimental device consisting of two interacting pendulum clocks hanged on a heavy support, and this support was mounted on a low-friction wheeled cart. This device moves by the action of the tension forces originated by the swing of the two pendulums, and the interaction between the two clocks is due to the mobility of the heavy base of the clocks. If the difference between the natural or eigenfrequencies of the two clocks is less than 0.0009 Hz, the anti-phase synchrony between the two pendulum clocks is reached. If the difference between these frequencies is larger than 0.0045 Hz, the two clocks do not synchronize, running “uncoupled” or in a state of beating death [2]. For example, a difference of order of $\Delta\omega = 0.0009$ Hz for the two pendulum eigenfrequencies corresponds to a difference in the lengths of the pendulum rods of the order of $\Delta\ell = \sqrt{g}\ell^{3/2}\Delta\omega/4\pi$, which gives, for $\ell = 1$ m and $g = 9.8$ ms⁻², $\Delta\ell = 4$ mm, and for $\ell = 0.178$ m (the length of the pendulum rods used by Huygens, [1]), $\Delta\ell = 0.02$ mm, a precision that Huygens certainly could not achieve. According to Bennett et al. [2, p. 578], Huygens’s results depended on both talent and luck.

In the experiment of Bennett et al. [2], the in-phase synchronization is the natural way of synchronization of the two pendulum clocks. However, due to a detailed description of the Huygens findings, we can believe that the in-phase synchronization was never observed by Huygens.

Another experimental model mimicking the Huygens’s clocks system consists of two pendulums whose suspension rods are connected by a weak spring, and one of the two pendulums is driven by an external rotor [3] and [4]. In this system, the in-phase synchronization is approximately achieved with a small phase shift, and the experimental measurements and the model analysis both agree. The numerical results of Fradkov and Andrievsky for this device [4] show simultaneous and approximate in-phase and anti-phase synchronization, tuned by different initial conditions. However, if the pendulums have slightly different periods, the two oscillators may not synchronize [2, 4]. In another experimental device made of two rotors controlled by external torques [5, 6], Andrievsky et al. [5] reported the approximate anti-phase and in-phase synchronization of the two oscillators. In this experiment, the synchronization parameter is the stiffness of a spring connecting the two rotors.

For demonstration purposes, Pantaleone [7] built an experimental device made of two metronomes on a freely moving light wooden base. The base lies on two empty soda cans. In this metronomes experiment, the phase difference of the synchronous state of the two metronomes is close to 0°. Increasing the damping effect

on the freely moving base, the author reported approximate synchronization with a difference in phase close to 180° . From the theoretical point of view, the equations describing this experimental device lead to the Kuramoto synchronization model [8], where the synchronization mechanism is due to a non-linear effect associated with the phase difference between the oscillators.

In the Bennett et al. [2] and the Pantaleone [7] experimental systems, the interaction mechanism between oscillators is obtained by a moving base, an idea advanced by Kortweg [9]. In these systems, there is no clear evidence of what mechanisms are in the origin of the anti-phase synchronization, as described by Huygens. One common conclusion is that, if the pendulums have slightly different periods, the two oscillators may not synchronize [2, 4].

As the special type of collective rhythmicity just described occurs in several biological systems and several other natural phenomena [10], it is important to derive and to understand the interaction mechanisms leading to exact synchrony. Besides all the attention of the scientific community for this synchronization phenomenon, there is no clear evidence of a mechanism leading to anti-phase synchronization.

In this chapter, we propose and analyze a new interaction mechanism between oscillators leading to exact anti-phase and in-phase synchronization. The synchronization mechanism is obtained with a damped elastic spring, and the oscillators under analysis can be simple harmonic oscillators, pendulums, and any type of non-linear oscillators with a limit cycle in phase space, as is the case of pendulum clocks [11].

The main result of this chapter is to show that exact anti-phase synchronization can always be achieved for systems of coupled oscillators.

This chapter is organized as follows. In Sect. 10.2, we introduce the synchronization model for two generic oscillators, and we discuss its physical assumptions. Then, we make the approximation of small oscillations. In Sect. 10.3, we introduce a simplified pendulum clock model, and we prove that the ordinary differential equation describing the dynamics of the clock has a unique limit cycle in phase space. Based on the linear model derived in Sect. 10.2, in Sect. 10.4, we discuss the concept of anti-phase synchronization, and we find a sufficient condition for the existence of an exact anti-phase synchronized state for the two-pendulum clock system. This result is stated as Theorem 10.1, the main result of this chapter. After exploring numerically the phase space structure of the solutions of the model equations as a function of a control parameter, we show the existence of exact anti-phase and in-phase synchronization states for the Huygens' two-pendulum clock system. The tuning between the two types of synchronization regimes can be controlled through a damping constant associated with the (elastic) interaction mechanism and by the choice of initial conditions. These results justify some numerical results published previously [12].

In Sect. 10.5, we analyze numerically the persistence of in-phase and of anti-phase synchronization effects when we vary the parameters of the individual pendulum clocks. We show that the anti-phase and the in-phase synchronization regimes persist even if the two oscillators have different eigenperiods and parameters. Finally, in Sect. 10.6, we resume the conclusions of this chapter.

10.2 A Model for the Synchronization of the Two Pendulum Clocks

In the Huygens' two-pendulum clock system, the pendulums are hanged in a common support, and the only possible interaction between them is mediated by the tension forces generated by the oscillatory motion of the two pendulums. These tension forces propagate through the common support that we consider to be elastic. The role of the tension forces in the synchronization mechanism is corroborated by the Huygens's finding that when the planes of oscillation of the two hanged pendulums are mutually perpendicular, no synchronization is observed. In fact, the components of the tension forces generated by the motion of the pendulums are in the plane of motion of the pendulums and force the motion of the two attachment points. To be more specific, we consider the mechanical arrangement of Fig. 10.1, where the two pendulums have masses m_1 and m_2 , and lengths ℓ_1 and ℓ_2 , respectively.

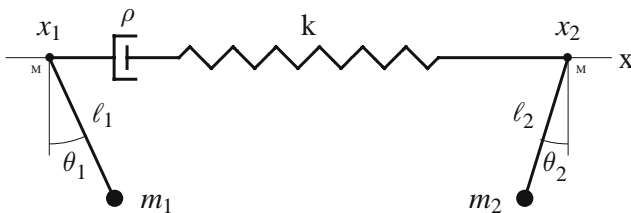


Fig. 10.1 Model to analyze the synchronization of the Huygens' two-pendulum clock system. The two pendulums are a representation of two non-linear oscillators. The interaction between the pendulums is mediated by the tension forces at the attachment points, and they propagate through an elastic and resistive media. Each attachment point is considered to have mass M

The pendulums are considered connected by a massless spring with stiffness constant k . The perturbations that propagate along the spring are damped, and the damping force is proportional to the velocity of the attachment points of the spring, with damping constant ρ . The damped spring simulates the interaction effects that propagate through the elastic and resistive media. We consider that the attachment points of the pendulums, located at the horizontal coordinates $x = x_1$ and $x = x_2$, have equal masses, and we denote this mass by M . As we shall see, the introduction of this mass is necessary to obtain explicitly the equations of motion.

The mechanical system of Fig. 10.1, considered without the damping forces, is described by the four degrees of freedom Lagrangian,

$$\begin{aligned}
 L = & \frac{1}{2}m_1(\ell_1^2\dot{\theta}_1^2 + \dot{x}_1^2 + 2\ell_1\dot{x}_1\dot{\theta}_1\cos\theta_1) + m_1g\ell_1\cos\theta_1 \\
 & + \frac{1}{2}m_2(\ell_2^2\dot{\theta}_2^2 + \dot{x}_2^2 + 2\ell_2\dot{x}_2\dot{\theta}_2\cos\theta_2) + m_2g\ell_2\cos\theta_2 \\
 & + \frac{1}{2}M(\dot{x}_1^2 + \dot{x}_2^2) - \frac{1}{2}k(x_2 - x_1)^2,
 \end{aligned} \tag{10.1}$$

where θ_1 and θ_2 are the angular coordinates of the two pendulums, g is the acceleration due to the gravity force, and the last two terms describe the interaction mechanism between the two pendulums. From (10.1), the Lagrange equations of motion of the two interacting pendulums are

$$\begin{aligned} m_1 \ell_1 \ddot{\theta}_1 + m_1 g \sin \theta_1 &= -m_1 \ddot{x}_1 \cos \theta_1 \\ m_2 \ell_2 \ddot{\theta}_2 + m_2 g \sin \theta_2 &= -m_2 \ddot{x}_2 \cos \theta_2 \\ (M + m_1) \ddot{x}_1 + m_1 \ell_1 \ddot{\theta}_1 \cos \theta_1 &= m_1 \ell_1 \dot{\theta}_1^2 \sin \theta_1 + k(x_2 - x_1) \\ (M + m_2) \ddot{x}_2 + m_2 \ell_2 \ddot{\theta}_2 \cos \theta_2 &= m_2 \ell_2 \dot{\theta}_2^2 \sin \theta_2 - k(x_2 - x_1). \end{aligned} \quad (10.2)$$

Introducing the damping effects and the escaping mechanism of the clocks into the system of equations (10.2), we obtain

$$\begin{aligned} m_1 \ell_1 \ddot{\theta}_1 + f_1(\theta_1, \dot{\theta}_1) + m_1 g \sin \theta_1 &= -m_1 \ddot{x}_1 \cos \theta_1 \\ m_2 \ell_2 \ddot{\theta}_2 + f_2(\theta_2, \dot{\theta}_2) + m_2 g \sin \theta_2 &= -m_2 \ddot{x}_2 \cos \theta_2 \\ (M + m_1) \ddot{x}_1 + 2\rho \dot{x}_1 + m_1 \ell_1 \ddot{\theta}_1 \cos \theta_1 &= m_1 \ell_1 \dot{\theta}_1^2 \sin \theta_1 + k(x_2 - x_1) \\ (M + m_2) \ddot{x}_2 + 2\rho \dot{x}_2 + m_2 \ell_2 \ddot{\theta}_2 \cos \theta_2 &= m_2 \ell_2 \dot{\theta}_2^2 \sin \theta_2 - k(x_2 - x_1), \end{aligned} \quad (10.3)$$

where ρ is the damping constant of the attachment points of the pendulums. The functions $f_1(\theta_1, \dot{\theta}_1)$ and $f_2(\theta_2, \dot{\theta}_2)$ describe the escaping mechanism of the pendulum clocks (Section 10.3).

The system of equations (10.3) implicitly defines a system of ordinary differential equations. If $M > 0$, $\ell_1 > 0$, $m_1 > 0$, $\ell_2 > 0$, and $m_2 > 0$, the system (10.3) can be solved algebraically in order to the higher derivatives, and we obtain

$$\begin{aligned} m_1 \ell_1 \ddot{\theta}_1 + f_1(\theta_1, \dot{\theta}_1) + m_1 g \sin \theta_1 &= -m_1 \cos \theta_1 F_1 \\ m_2 \ell_2 \ddot{\theta}_2 + f_2(\theta_2, \dot{\theta}_2) + m_2 g \sin \theta_2 &= -m_2 \cos \theta_2 F_2 \\ \ddot{x}_1 &= F_1 \\ \ddot{x}_2 &= F_2, \end{aligned} \quad (10.4)$$

where

$$\begin{aligned} F_1 &= \frac{f_1(\theta_1, \dot{\theta}_1) \cos \theta_1 + m_1 g \sin \theta_1 \cos \theta_1 - 2\rho \dot{x}_1 + m_1 \ell_1 \dot{\theta}_1^2 \sin \theta_1 + k(x_2 - x_1)}{M + m_1 \sin^2 \theta_1} \\ F_2 &= \frac{f_2(\theta_2, \dot{\theta}_2) \cos \theta_2 + m_2 g \sin \theta_2 \cos \theta_2 - 2\rho \dot{x}_2 + m_2 \ell_2 \dot{\theta}_2^2 \sin \theta_2 - k(x_2 - x_1)}{M + m_2 \sin^2 \theta_2}. \end{aligned} \quad (10.5)$$

The system of ordinary differential equations (10.8) provide a synchronization model for the Huygens' two-pendulum clock system [12]. Here, we will consider only the case $M > 0$. The case $M = 0$ will be analyzed elsewhere.

For small amplitude of oscillations in θ_1 and θ_2 , the system of equations (10.4) and (10.5) simplifies, and we obtain

$$\begin{aligned}
m_1 \ell_1 \ddot{\theta}_1 + f_1(\theta_1, \dot{\theta}_1) + m_1 g \theta_1 &= -\frac{m_1}{M} (f_1(\theta_1, \dot{\theta}_1) + m_1 g \theta_1 - 2\rho \dot{x}_1 + k(x_2 - x_1)) \\
m_2 \ell_2 \ddot{\theta}_2 + f_2(\theta_2, \dot{\theta}_2) + m_2 g \theta_2 &= -\frac{m_2}{M} (f_2(\theta_2, \dot{\theta}_2) + m_2 g \theta_2 - 2\rho \dot{x}_2 - k(x_2 - x_1)) \\
\ddot{x}_1 &= \frac{1}{M} (f_1(\theta_1, \dot{\theta}_1) + m_1 g \theta_1 - 2\rho \dot{x}_1 + k(x_2 - x_1)) \\
\ddot{x}_2 &= \frac{1}{M} (f_2(\theta_2, \dot{\theta}_2) + m_2 g \theta_2 - 2\rho \dot{x}_2 - k(x_2 - x_1)) .
\end{aligned} \tag{10.6}$$

In this chapter, our goal is to analyze the synchronization properties of the solutions of the system of ordinary differential equations (10.6). These equations are written in dimensional form and we use always the international system of units (SI units). In order to simplify the notation, in the following, parameter values will be written without the specifying the corresponding SI units.

To model the Huygens' two-pendulum clock experiment, we have to choose a specific form for the functions $f_1(\theta_1, \dot{\theta}_1)$ and $f_2(\theta_2, \dot{\theta}_2)$ describing the oscillatory behavior of pendulum clocks.

10.3 A Simple Clock Model

To restore the energy lost by a pendulum clock during one period, the sustained oscillations can be maintained by an impulsive force acting on the pendulum rod or by the force originated by the smooth unwinding of a circular spring attached to the pendulum balance wheel [11, pp. 169–200]. In any case, the dynamics of a pendulum clock is modeled by a two-dimensional dynamical system with a limit cycle in phase space. This same qualitative behavior can be obtained with an oscillator actuated by a non-linear damping force, piecewise proportional to the angular velocity of the pendulum. For small amplitude of oscillations, the proportionality constant is positive, and for large amplitudes of oscillations, the proportionality constant is negative.

To simplify our analysis, we take, as a qualitative model for a pendulum clock, the following second-order differential equation:

$$m\ell\ddot{\theta} + f(\theta; \lambda, \tilde{\theta})\dot{\theta} + mg\theta = 0, \tag{10.7}$$

where

$$f(\theta; \lambda, \tilde{\theta}) = \begin{cases} -2\lambda & \text{if } |\theta| < \tilde{\theta} \\ 2\lambda & \text{if } |\theta| \geq \tilde{\theta}. \end{cases} \tag{10.8}$$

The function $-f(\theta; \lambda, \tilde{\theta})\dot{\theta}$ is the damping force of the pendulum clock, and λ and $\tilde{\theta}$ are positive constants.

Proposition 10.1. *If $\lambda > 0$, $\tilde{\theta} > 0$, $m > 0$, $\ell > 0$, and $g > 0$, the second-order differential equation (10.7), with the damping function (10.8), has a unique and stable limit cycle in phase space.*

Proof. Assume that $\lambda > 0$, $\tilde{\theta} > 0$, $m > 0$, $\ell > 0$, and $g > 0$. Define the function

$$F(\theta; \eta, \tilde{\theta}) = \frac{1}{m\ell} \int_0^\theta f(s; \lambda, \tilde{\theta}) ds = \begin{cases} -2\eta\theta & \text{if } |\theta| < \tilde{\theta} \\ 2\eta\theta - 4\eta\tilde{\theta} & \text{if } |\theta| \geq \tilde{\theta}, \end{cases}$$

where, $\eta = \lambda/(m\ell)$ and $\omega^2 = g/\ell$. With the new coordinate $x = \dot{\theta} + F(\theta; \eta, \tilde{\theta})$, the differential equation (10.7), with damping function (10.8), has the Liénard form

$$\begin{cases} \dot{\theta} = x - F(\theta; \eta, \tilde{\theta}) \\ \dot{x} = -\omega^2\theta, \end{cases} \quad (10.9)$$

where $F(\theta; \eta, \tilde{\theta})$ is a continuous and odd function of θ . The existence, stability, and uniqueness of a limit cycle solution for eq. (10.9) follow from the Liénard theorem [13, pp. 179–181]. \square

In the conditions of Proposition 10.1, the differential equation (10.7), with damping function (10.8), has a unique limit cycle in phase space. In Fig. 10.2, we show the limit cycle in phase space of this non-linear oscillator for two values of the parameter $\omega^2 = g/\ell$.

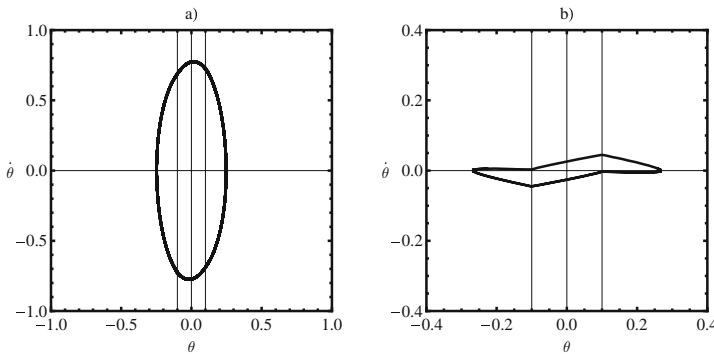


Fig. 10.2 Limit cycle solutions of the differential equation (10.7) with the damping function (10.8). In (a) we have chosen $\omega^2 = 9.8$, and in (b) $\omega^2 = 0.005$, where $\omega^2 = g/\ell$, $m = 1$, and $\ell = 1$. The parameter values of the damping function are $\lambda = 0.1$ and $\tilde{\theta} = 0.1$. The two vertical lines for $\theta = \pm\tilde{\theta}$ show the discontinuity of the tangential derivative along the limit cycles

The non-linear oscillator model defined by the differential equation (10.7), with damping function (10.8), has all the qualitative properties found in more accurate pendulum clock models [11] and, in the next sections, will be used to model Huygens's clocks.

10.4 Synchronization of Two Pendulum Clocks with Equal Parameters

From (10.6), the two interacting pendulum clocks with equal parameters are modeled by the system of differential equations,

$$\begin{aligned}
\ddot{\theta}_1 + \left(\frac{1}{m\ell} + \frac{1}{M\ell} \right) f(\theta_1; \lambda, \tilde{\theta}) \dot{\theta}_1 + \omega^2 \left(1 + \frac{m}{M} \right) \theta_1 - 2 \frac{\rho}{M\ell} \dot{x}_1 &= -\frac{k}{M\ell} (x_2 - x_1) \\
\ddot{\theta}_2 + \left(\frac{1}{m\ell} + \frac{1}{M\ell} \right) f(\theta_2; \lambda, \tilde{\theta}) \dot{\theta}_2 + \omega^2 \left(1 + \frac{m}{M} \right) \theta_2 - 2 \frac{\rho}{M\ell} \dot{x}_2 &= \frac{k}{M\ell} (x_2 - x_1) \\
\ddot{x}_1 - \frac{1}{M} f(\theta_1; \lambda, \tilde{\theta}) \dot{\theta}_1 - \frac{m}{M} g \theta_1 + 2 \frac{\rho}{M} \dot{x}_1 &= \frac{k}{M} (x_2 - x_1) \\
\ddot{x}_2 - \frac{1}{M} f(\theta_2; \lambda, \tilde{\theta}) \dot{\theta}_2 - \frac{m}{M} g \theta_2 + 2 \frac{\rho}{M} \dot{x}_2 &= -\frac{k}{M} (x_2 - x_1),
\end{aligned} \tag{10.10}$$

where $f(\theta; \lambda, \tilde{\theta})$ is given by (10.8) and $\omega^2 = g/\ell$.

Our goal here is to show that the solutions of the system of differential equations (10.10) synchronize in anti-phase. For that, we begin by analyzing the solutions of the system of equations (10.10) for the particular case where the initial conditions in θ_1 and θ_2 are bounded by $\tilde{\theta}$, that is, $|\theta_1(0)| < \tilde{\theta}$ and $|\theta_2(0)| < \tilde{\theta}$. As the system of equations (10.10), together with (10.8), is piecewise linear, and if $|\theta_1(t)| < \tilde{\theta}$ and $|\theta_2(t)| < \tilde{\theta}$, for every $t \in [0, t^*]$, where t^* is a positive constant, then the piecewise linear system of equations (10.10) can be written as the linear first-order system of differential equations:

$$\begin{pmatrix} \dot{\theta}_1 \\ \dot{v}_1 \\ \dot{\theta}_2 \\ \dot{v}_2 \\ \dot{x}_1 \\ \dot{w}_1 \\ \dot{x}_2 \\ \dot{w}_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & B & 0 & 0 & \frac{k}{M\ell} & \frac{2\rho}{M\ell} & -\frac{k}{M\ell} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & A & B & -\frac{k}{M\ell} & 0 & \frac{k}{M\ell} & \frac{2\rho}{M\ell} \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \frac{mg}{M} - \frac{2\lambda}{M} & 0 & 0 & 0 & -\frac{k}{M} & -\frac{2\rho}{M} & \frac{k}{M} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & \frac{mg}{M} - \frac{2\lambda}{M} & \frac{k}{M} & 0 & -\frac{k}{M} & -\frac{2\rho}{M} & 0 \end{pmatrix} \begin{pmatrix} \theta_1 \\ v_1 \\ \theta_2 \\ v_2 \\ x_1 \\ w_1 \\ x_2 \\ w_2 \end{pmatrix}, \tag{10.11}$$

where we have introduced the new variables $v_1 = \dot{\theta}_1$, $v_2 = \dot{\theta}_2$, $w_1 = \dot{x}_1$, $w_2 = \dot{x}_2$ and the new constants $A = -\omega^2(1 + m/M)$ and $B = 2\lambda(1/(M\ell) + 1/(m\ell))$. As, $|\theta_1(t)| < \tilde{\theta}$ and $|\theta_2(t)| < \tilde{\theta}$, for every $t \in [0, t^*]$, then, for the same initial conditions, the small amplitude solutions of (10.10) and (10.11) coincide in the interval $[0, t^*]$.

Denoting by Q the matrix in (10.11), we have $\text{Det } Q = 0$. A simple inspection of Q shows that Q has only one zero eigenvalue corresponding to the eigendirection defined by the equation $x_1 = x_2$ ¹. We investigate now the stability of the line of fixed points $x_1 = x_2$ of both equations (10.10) and (10.11).

Proposition 10.2. *If $\lambda > 0$, $M > 0$, $\ell > 0$, $m > 0$, and $\rho > 0$ are sufficiently small, then the systems differential equations (10.10) and (10.11) have a line of fixed points with coordinates $\theta_1 = 0$, $v_1 = 0$, $\theta_2 = 0$, $v_2 = 0$, $w_1 = 0$, $w_2 = 0$, $x_1 = x_2$, and this line of fixed points is Lyapunov unstable.*

Proof. We assume that $\lambda > 0$, $M > 0$, $\ell > 0$, and $m > 0$. The existence of the line of fixed points follows by solving the equation $Qy = 0$, where

¹ The vector $x^T = (0, 0, 0, 0, 1, 0, 1, 0)$ is such that $Mx = 0$.

$$y^T = (\theta_1, v_1, \theta_2, v_2, x_1, w_1, x_2, w_2).$$

For $\rho = 0$, the characteristic polynomial of the matrix Q in Eq. (10.11) is

$$q_{\rho=0}(x) = x^2(A + (B - x)x) (2gkm + M(A + (B - x)x) (Mx^2 + 2k)\ell - 4kx\lambda),$$

where $A = -\omega^2(1 + m/M)$ and $B = 2\lambda(1/(M\ell) + 1/(m\ell))$. As the polynomial $q_{\rho=0}(x)$ has the two roots,

$$\lambda = \frac{1}{2} \left(B \pm \sqrt{B^2 + 4A} \right),$$

with positive real parts $B > 0$, then by continuity of $Q(\rho)$ for sufficiently small $\rho > 0$, the characteristic polynomial $q_{\rho>0}(x)$ of the matrix Q has also eigenvalues with positive real parts. Therefore, for sufficiently small $\rho > 0$, $\lambda > 0$, $M > 0$, $\ell > 0$, $m > 0$, the line of fixed points of the linear system (10.11) is Lyapunov unstable. As both systems of equations (10.10) and (10.11) have the same phase space orbits near the common line of fixed points, the local properties of the flows are the same, and this is sufficient to prove the local instability of the flow defined by the system of equations (10.10). \square

Proposition 10.2 gives the conditions of nonconvergence to the quiescent state of the solutions of the system of linear equation (10.11). This quiescent state is the line of fixed points $x_1 = x_2$ in the eight-dimensional phase space. Under the conditions of the Proposition 10.2, the two pendulum clocks have sustained oscillations in the sense that $\lim_{t \rightarrow \infty} \theta_1(t) \neq 0$ and $\lim_{t \rightarrow \infty} \theta_2(t) \neq 0$.

To find a sufficient condition of existence of anti-phase synchronization of the two non-linear oscillators, we assume now that the asymptotic solutions of the system of equations (10.10) synchronize in anti-phase. As the two pendulum clocks are identical, with $t_n = t_0 + nh$, where $n = 0, 1, \dots$ and t_0 is the initial time, we must have

$$\begin{aligned} \lim_{t_n \rightarrow \infty} \theta_1(t_n) &= - \lim_{t_n \rightarrow \infty} \theta_2(t_n) \\ \lim_{t_n \rightarrow \infty} x_1(t_n) &= - \lim_{t_n \rightarrow \infty} x_2(t_n) \\ \lim_{t_n \rightarrow \infty} v_1(t_n) &= - \lim_{t_n \rightarrow \infty} v_2(t_n) \\ \lim_{t_n \rightarrow \infty} w_1(t_n) &= - \lim_{t_n \rightarrow \infty} w_2(t_n), \end{aligned} \tag{10.12}$$

for every $h > 0$.²

We now define the new variables $\theta = \theta_1 + \theta_2$, $v = v_1 + v_2$, $x = x_1 + x_2$, and $w = w_1 + w_2$. Then, adding the corresponding variables in the system of equations (10.11), we obtain

² In this context, we say that the two pendulum clocks synchronize in anti-phase if, for every $h > 0$, $\lim_{t_n \rightarrow \infty} \theta_1(t_n) = -\lim_{t_n \rightarrow \infty} \theta_2(t_n)$. As we shall see in Section 10.5, this definition can only be used in the case of identical pendulum clocks.

$$\begin{pmatrix} \dot{\theta} \\ \dot{v} \\ \dot{x} \\ \dot{w} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -\omega^2(1 + \frac{m}{M}) & 2\lambda(\frac{1}{M\ell} + \frac{1}{m\ell}) & 0 & \frac{2\rho}{M\ell} \\ 0 & 0 & 0 & 1 \\ \frac{mg}{M} & -\frac{2\lambda}{M} & 0 & -\frac{2\rho}{M} \end{pmatrix} \begin{pmatrix} \theta \\ v \\ x \\ w \end{pmatrix}. \quad (10.13)$$

In the following, we call the system of equations (10.13) the reduced system of equations associated with system (10.11). The reduced system of equations (10.13) has a line of fixed points with coordinates $\theta = 0$, $v = 0$, $w = 0$, and $x = \text{constant}$. If the line of fixed points of the reduced linear system (10.13) is asymptotically stable, then we have

$$\begin{aligned} \lim_{t \rightarrow \infty} \theta(t) &= \lim_{t \rightarrow \infty} \theta_1(t) + \lim_{t \rightarrow \infty} \theta_2(t) = 0 \\ \lim_{t \rightarrow \infty} v(t) &= \lim_{t \rightarrow \infty} v_1(t) + \lim_{t \rightarrow \infty} v_2(t) = 0 \\ \lim_{t \rightarrow \infty} w(t) &= \lim_{t \rightarrow \infty} w_1(t) + \lim_{t \rightarrow \infty} w_2(t) = 0. \end{aligned} \quad (10.14)$$

If, for every $h > 0$, the first condition in (10.12) is verified, the first condition in (10.14) is also verified, and synchronous solutions of Eq. (10.10) are also stable solutions of Eq. (10.13).

As, by Proposition 10.2, $\lim_{t_n \rightarrow \infty} \theta_1(t_n) \neq 0$ and $\lim_{t_n \rightarrow \infty} \theta_2(t_n) \neq 0$, for any initial condition away from the line of fixed points, any solution of the differential equation (10.10) that anti-phase synchronizes is also an asymptotically stable solution of the reduced system of equations (10.13).

Hence, if the line of fixed points of the reduced system (10.13) is asymptotically stable and the line of fixed points of the linear system of equations (10.11) is unstable (Proposition 10.2), then the solutions $\theta_1(t)$ and $\theta_2(t)$ of the system of equations (10.10) anti-phase synchronize. The asymptotic stability condition for the reduced linear system (10.13) together with the instability condition of Proposition 10.2 both give a sufficient condition for the existence of exact anti-phase synchronization of the two identical non-linear pendulum clocks.

To analyze the stability properties of the line of fixed points of the system of equations (10.13), we calculate the characteristic polynomial of the matrix P in (10.13):

$$\begin{aligned} p(y) &= y(mM\ell y^3 + y^2(2m\ell\rho - 2m\lambda - 2M\lambda) \\ &\quad + y(m^2\ell\omega^2 + mM\ell\omega^2 - 4\lambda\rho) + 2mg\rho) \\ &= yp_1(y). \end{aligned} \quad (10.15)$$

To the eigenvalue $x = 0$ corresponds the eigendirection e_x . This eigendirection is the line of fixed points of the linear system (10.13). If all the eigenvalues of the polynomial $p_1(y)$ in (10.15) have negative real parts, any initial condition away from the line of fixed points $x = \text{constant}$ evolves in time to this line of fixed points. Therefore, we have:

Theorem 10.1. *We consider the system of differential equations (10.10) with damping function (10.8). If $\lambda > 0$, $\hat{\theta} > 0$, $m > 0$, $\ell > 0$, $M > 0$, $k > 0$, $g > 0$, and $\rho > 0$ is sufficiently small, and if the reduced linear differential equation (10.13) has only non-positive eigenvalues, then the solutions of equation (10.10), with damping*

function (10.8), synchronize in anti-phase, in the sense that, for every $h > 0$,

$$\lim_{t_n \rightarrow \infty} \theta_1(t_n) = - \lim_{t_n \rightarrow \infty} \theta_2(t_n),$$

where $t_n = t_0 + nh$, $\lim_{t_n \rightarrow \infty} \theta_1(t_n) \neq 0$, and $\lim_{t_n \rightarrow \infty} \theta_2(t_n) \neq 0$. Moreover, if the polynomial

$$q(\rho) = 4m\ell\lambda\rho^2 - (4m\lambda^2 + 4M\lambda^2 + gm^3\ell)\rho + gm^3\lambda + 2gm^2M\lambda + gmM^2\lambda,$$

has two real roots ρ_1 and ρ_2 , and ρ obeys to the inequalities,

$$\begin{aligned} \rho_1 &< \rho < \rho_2 \\ \rho &> \rho_0 = \frac{\lambda}{\ell} \left(1 + \frac{M}{m}\right), \end{aligned}$$

then for any initial condition away from the line of fixed points of the system of equations (10.10), the solutions of Eq. (10.10), with damping function (10.8), anti-phase synchronize.

Proof. The sufficient condition for the existence of anti-phase synchronization of the two pendulum clocks has been derived before the statement of the theorem. The instability of the line of fixed points of the system of equations (10.10) has been proven in Proposition 10.2. To prove the condition of non-positivity of the eigenvalues of the matrix in the system of equations (10.13), we use the Routh–Hurwitz criterion [14]. By (10.15), as $p_1(y) = a_0y^3 + a_1y^2 + a_2y + a_3$, and as, by hypothesis, $a_0 > 0$ and $a_3 > 0$, by the Routh–Hurwitz criterion, if $a_1 > 0$ and $(a_1a_2 - a_0a_3) > 0$, then the polynomial $p_1(y)$ has only roots with negative real parts. As

$$(a_1a_2 - a_0a_3) = -q(\rho) = -4m\ell\lambda\rho^2 + (4m\lambda^2 + 4M\lambda^2 + gm^3\ell)\rho - gm^3\lambda - 2gm^2M\lambda - gmM^2\lambda,$$

the polynomial $q(\rho)$ has a global minimum for positive values of ρ and can have two real roots. This proves the first inequality of the theorem. The second inequality follows from the Routh–Hurwitz condition $a_1 = 2m\ell\rho - 2m\lambda - 2M\lambda > 0$. The existence of the asymptotic solution follows from the existence and uniqueness theorem for Eq. (10.13). \square

Theorem 10.1 gives a sufficient condition for the existence of exact anti-phase synchronization in the Huygens' two-pendulum clock system. To test numerically the results of Theorem 10.1, the parameters of the non-linear oscillator (10.7) and (10.8) have the values $g = 9.8$, $m = 1$, $\ell = 1$, $\lambda = 0.1$, and $\bar{\theta} = 0.1$. For these parameter values, the period of the solutions on the limit cycle of Eq. (10.7) is $T = 2.008$. The parameters describing the interaction between the pendulum clocks have been set to the values $k = 10$ and $M = 0.1$. In the following, ρ is a free parameter.

To test the conditions of Theorem 10.1, in Fig. 10.3, we have plotted the eigenvalues with the largest real part of the matrices Q and P of the linear systems (10.11) and (10.13) as a function of the damping parameter ρ . The zero eigenvalue has been excluded from the characteristic polynomials of the matrices Q and P . According

to Theorem 10.1, if $0.121 = \rho_1 < \rho < \rho_2 = 24.489$, $\rho > \rho_0 = 0.11$, and the matrix Q has eigenvalues with positive real parts, then the two pendulum clocks synchronize in anti-phase. Numerically, if $\rho < \rho_3 = 0.393$, the matrix Q of the system of equations (10.11) has positive eigenvalues. Therefore, the conditions of Theorem 10.1 imply that if $\rho \in [\rho_1, \rho_3]$, the two pendulum clocks synchronize in anti-phase (Fig. 10.3).

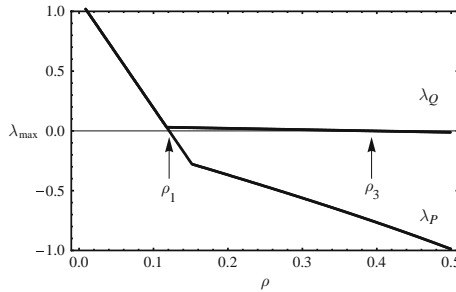


Fig. 10.3 Eigenvalues with the largest real part of the matrices Q and P of the linear systems (10.11) and (10.13), respectively, as a function of the damping parameter ρ . The other parameters have been fixed to the values: $g = 9.8$, $m = 1$, $\ell = 1$, $\lambda = 0.1$, $\tilde{\theta} = 0.1$, $k = 10$, and $M = 0.1$. If $\rho > \rho_1 = 0.121$ and $\rho < \rho_2 = 24.489$, all the eigenvalues of the matrix P have non-positive real parts, and the line of fixed points of system (10.13) is Lyapunov stable. If $\rho < \rho_3 = 0.393$, the matrix Q has eigenvalues with positive real parts and the line of fixed points of system (10.10) is Lyapunov unstable. By Theorem 10.1, if $\rho \in [\rho_1, \rho_3]$, the two pendulum clocks synchronize in anti-phase

By Theorem 10.1, the anti-phase synchronization between the two non-linear pendulum oscillators exists for $\rho \in [\rho_1 = 0.121, \rho_3 = 0.393]$. In Fig. 10.4, we show the time evolution of the angular coordinates and attachment points of the two pendulum clocks for $\rho = 0.37$ and calculated numerically from the system of equation (10.10). We have chosen for initial conditions the coordinate values, $\theta_1(0) = 0.2$, $\theta_2(0) = 0.3$, $x_1(0) = 0$, $x_2(0) = 0$, $\dot{\theta}_1(0) = 0$, $\dot{\theta}_2(0) = 0$, $\dot{x}_1(0) = 0$, and $\dot{x}_2(0) = 0$. The two non-linear pendulums and the attachment points synchronize in anti-phase. Numerically, the period of the synchronized state is $T = 2.457$, which differs from the eigenperiod of the uncoupled non-linear oscillators $T = 2.008$. The anti-phase synchronized state of the system of equations (10.10) corresponds to a stable limit cycle (isolated periodic orbit) in the eight-dimensional phase space.

In Fig. 10.5, we have decreased the parameter ρ to values below the transition values ρ_0 and ρ_1 of Theorem 10.1. In this case, the two non-linear oscillators also anti-phase synchronize. The numerical integration for several initial conditions shows that, for the damping parameter value $\rho = 0.1$, the system of equation (10.10) has a stable limit cycle in the eight-dimensional phase space. In this case, we are outside the conditions of Theorem 10.1.

For the parameter values of Figs. 10.4 and 10.5, but with the same initial conditions for the two pendulum clocks, asymptotically in time, the two pendulums synchronize with the same phase (in-phase). However, any small deviation of

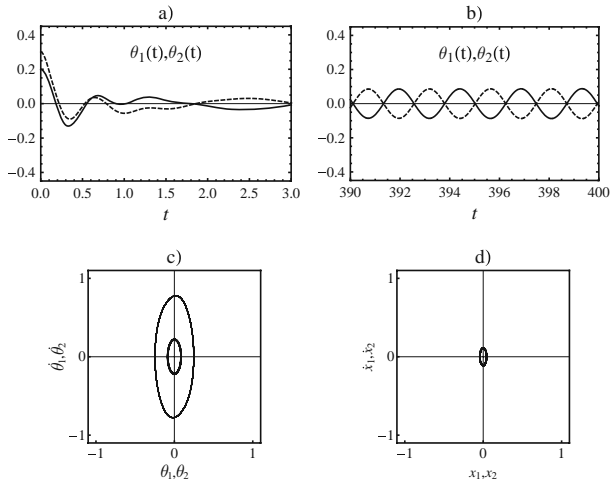


Fig. 10.4 Numerical solutions of the system of equations (10.10), with damping function (10.8), describing the coupling of two identical pendulum clocks. The parameter values of the simulation are $m = 1$, $\ell = 1$, $g = 9.8$, $k = 10$, $M = 0.1$, $\lambda = 0.1$, $\bar{\theta} = 0.1$, and $\rho = 0.37$. The initial conditions are $\theta_1(0) = 0.2$, $\theta_2(0) = 0.3$, $x_1(0) = 0$, $x_2(0) = 0$, $\dot{\theta}_1(0) = 0$, $\dot{\theta}_2(0) = 0$, $\dot{x}_1(0) = 0$, and $\dot{x}_2(0) = 0$. In (a) and (b), we show the time evolution of the angular coordinates of the two pendulum clocks, before and after anti-phase synchronization, respectively. In (c) and (d), we show the asymptotic solutions in the reduced phase space of the two pendulum clocks (c) and of the two attachment points (d). For comparison, in (c), we show the limit cycle solution (thin line curve) of the reference equation (10.7) with damping function (10.8). The period of the anti-phase oscillations is $T = 2.457$

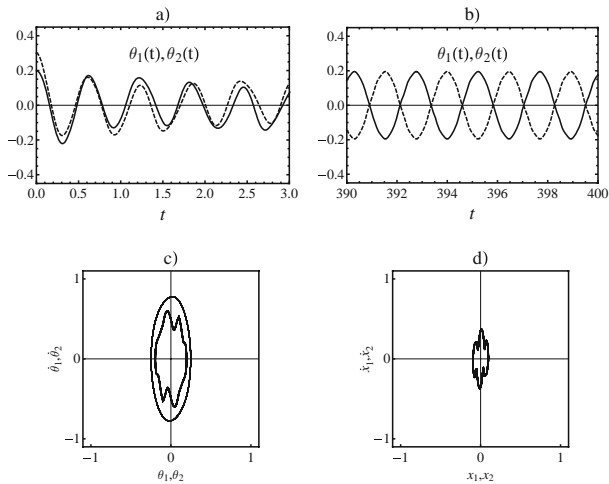


Fig. 10.5 Anti-phase synchronization of the two-pendulum clock system for the same parameter values of Fig. 10.4, except for the damping parameter ρ that, in this case, has the value $\rho = 0.1$. In this simulation, the period of the exact anti-phase oscillations is $T = 2.463$. In (c), the thin line curve is the limit cycle solution of the reference equation (10.7) with damping function (10.8)

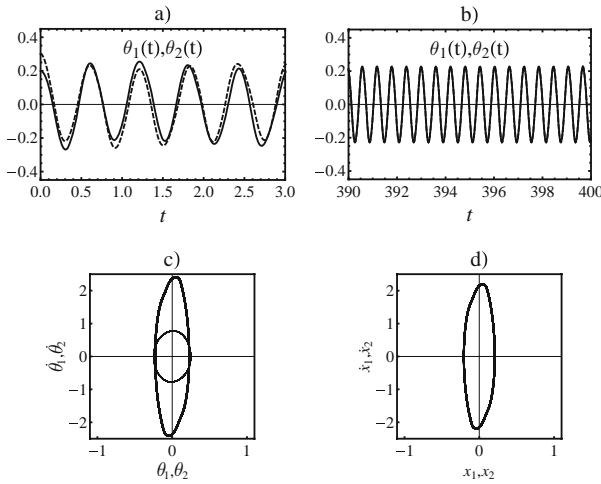


Fig. 10.6 In-phase synchronization of the two-pendulum clock system for the same parameter values of Fig. 10.4, except for the damping parameter ρ that, in this case, has the value $\rho = 0.01$. In this simulation, the initial conditions are $\theta_1(0) = 0.2$, $\theta_2(0) = 0.3$, $x_1(0) = 0$, $x_2(0) = 0$, $\dot{\theta}_1(0) = 0$, $\dot{\theta}_2(0) = 0$, $\dot{x}_1(0) = 0$, and $\dot{x}_2(0) = 0$. The period of the in-phase oscillations is $T = 0.606$. In (c), the thin line curve is the limit cycle solution of the reference equation (10.7) with damping function (10.8)

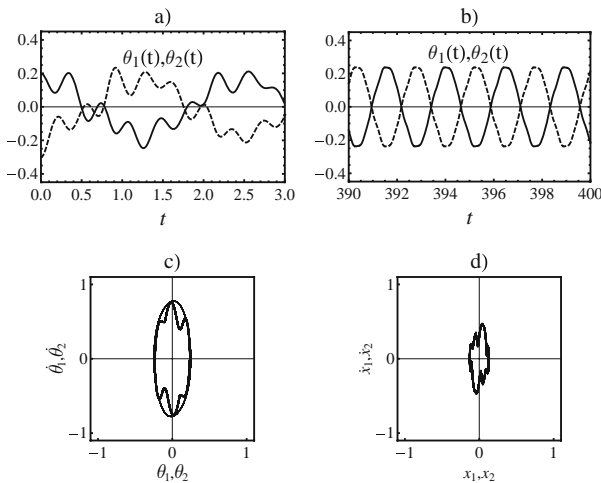


Fig. 10.7 Anti-phase synchronization of the two-pendulum clock system for the same parameter values of Fig. 10.6 and $\rho = 0.01$. In this simulation, the initial conditions are $\theta_1(0) = 0.2$, $\theta_2(0) = -0.3$, $x_1(0) = 0$, $x_2(0) = 0$, $\dot{\theta}_1(0) = 0$, $\dot{\theta}_2(0) = 0$, $\dot{x}_1(0) = 0$, and $\dot{x}_2(0) = 0$. Initially, the two pendulums are approximately in anti-phase. The period of the anti-phase oscillations is $T = 2.463$. In (c), the thin line curve is the limit cycle solution of the reference equation (10.7) with damping function (10.8)

one of the pendulums from these initial conditions, asymptotically in time, they anti-phase synchronize. This shows that, in this range of the control parameter ρ , there is an unstable limit cycle in phase space corresponding to an in-phase-synchronized regime.

Decreasing further the parameter ρ for $\rho < 0.066$ and for the same initial conditions as in Fig. 10.4, the two oscillators synchronize with the same phase (Fig. 10.6). However, changing the initial conditions for an approximate anti-phase initial state of the two pendulums, we obtain anti-phase synchronization, (Fig. 10.7). This simple fact shows that, for $\rho < 0.066$, there are at least two stable limit cycles in the eight-dimensional phase space, and these limit cycles have their own basins of attraction. These stable limit cycles exist for $0 \leq \rho < 0.066$.

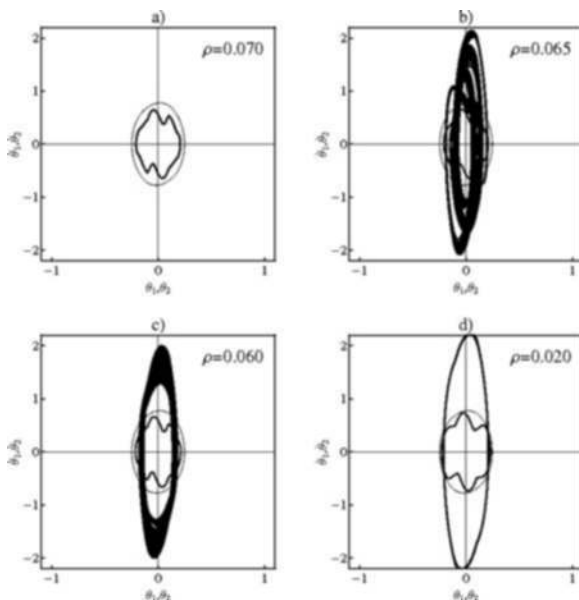


Fig. 10.8 Coexistence of anti-phase and in-phase synchronization and aperiodic regimes for the asymptotic solutions of the system of equations (10.10). We show the limit cycle solutions of the system equations (10.10) for the same parameter values of Fig. 10.4 and several values of ρ : (a) $\rho = 0.07$, (b) $\rho = 0.065$, (c) $\rho = 0.06$, and (d) $\rho = 0.02$. We have plotted the asymptotic solutions in phase space for two different initial conditions. In one case we have taken $\theta_1(0) = 0.2$ and $\theta_1(0) = 0.3$, with the other initial conditions equal to zero. In the other case, we have taken $\theta_1(0) = 0.2$ and $\theta_1(0) = -0.3$. For $\rho \geq 0.066$, different initial conditions lead always to anti-phase synchronization. For $\rho \leq 0.02$, we have three limit cycles in phase space, [12], two corresponding to anti-phase synchronization and the other to one in-phase synchronization. The thin line curve is the limit cycle solution of the reference equation (10.7), with damping function (10.8)

In Fig. 10.8, we show the transition from the anti-phase to the in-phase-synchronized asymptotic state as well as the bifurcation behavior of the attractors. We have superimposed in the same figure two stable limit cycles that can be

reached from different initial conditions. In Fig. 10.8(a), $\rho = 0.07$, there is only one asymptotically stable anti-phase-synchronized state. In Fig. 10.8(d), $\rho = 0.020$, there are three asymptotic stable synchronized states, one in-phase and the other in anti-phase. Fig. 10.8(b) and 10.8(c), one of the stable limit cycles for the asymptotic anti-phase synchronized states shows quasi-period behavior, which may correspond to a long transient. The basins of attraction of the anti-phase synchronized state that exists for $\rho < 0.066$ have been studied in [12].

Therefore, we have shown that the interaction mechanism proposed in Sect. 10.2 leads asymptotically in time to exact anti-phase synchronization of coupled identical oscillators. For a particular range of the control parameter ($0 \leq \rho \leq 0.06$), we can have anti-phase and in-phase synchronization, determined by different initial conditions. The tuning between the two types of synchronization regimes is controlled by the damping constant ρ .

10.5 Synchronization of Two Pendulum Clocks with Different Parameters: Robustness

In the previous analysis, we have considered that both oscillators are characterized by the same parameters. However, in real experiments this is not realistic and we must consider the persistence of synchronization when the parameters of the pendulums are different. Here, the persistence of the anti-phase and the in-phase synchronization states is analyzed numerically. We take, $m_1 = m$, $m_2 = m(1 + \varepsilon)$, $\ell_1 = \ell$, and $\ell_2 = \ell(1 + \delta)$, where ε and δ can have positive or negative values. In this case, the equations of motion (10.9) are rewritten as

$$\begin{aligned}
 \ddot{\theta}_1 + \left(\frac{1}{m\ell} + \frac{1}{M\ell} \right) f(\theta_1; \lambda, \tilde{\theta}) \dot{\theta}_1 + \omega^2 \left(1 + \frac{m}{M} \right) \theta_1 - 2 \frac{\rho}{M\ell} \dot{x}_1 &= - \frac{k}{M\ell} (x_2 - x_1) \\
 \ddot{\theta}_2 + \frac{1}{(1 + \varepsilon)(1 + \delta)} \left(\frac{1}{m\ell} + \frac{1}{M\ell} \right) f(\theta_2; \lambda, \tilde{\theta}) \dot{\theta}_2 + \frac{\omega^2}{(1 + \delta)} \left(1 + \frac{m}{M}(1 + \varepsilon) \right) \theta_2 \\
 - 2 \frac{\rho}{M\ell(1 + \delta)} \dot{x}_2 &= \frac{k}{M\ell(1 + \delta)} (x_2 - x_1) \\
 \ddot{x}_1 - \frac{1}{M} f(\theta_1; \lambda, \tilde{\theta}) \dot{\theta}_1 - \frac{m}{M} g \theta_1 + 2 \frac{\rho}{M} \dot{x}_1 &= \frac{k}{M} (x_2 - x_1) \\
 \ddot{x}_2 - \frac{1}{M} f(\theta_2; \lambda, \tilde{\theta}) \dot{\theta}_2 - \frac{m}{M} (1 + \varepsilon) g \theta_2 + 2 \frac{\rho}{M} \dot{x}_2 &= - \frac{k}{M} (x_2 - x_1),
 \end{aligned} \tag{10.16}$$

where $f(\theta; \lambda, \tilde{\theta})$ is given by (10.8).

We have integrated numerically the system of equations (10.16) for the same parameter values of Fig. 10.5 and the same initial conditions, but with $\varepsilon = 0.4$ and $\delta = 0.4$. The numerical results are presented in Fig. 10.9, and we conclude that the anti-phase-synchronized state still exists for large values of the parameters ε and δ .

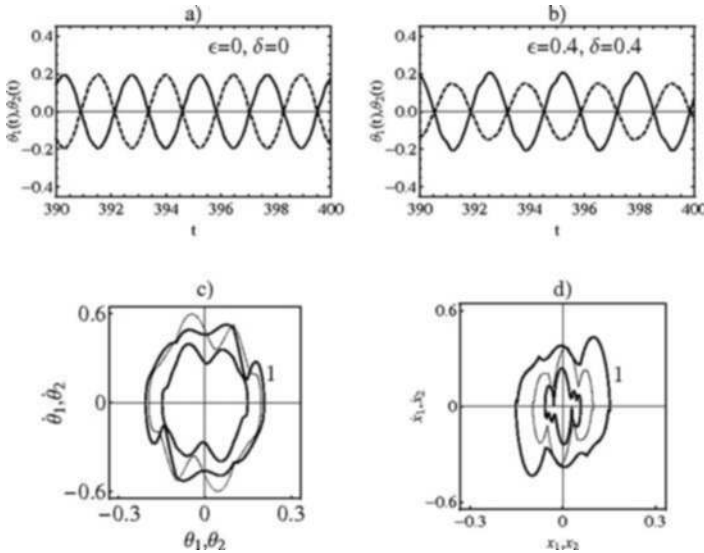


Fig. 10.9 Anti-phase synchronization of two pendulum clocks with different lengths and masses. In (a), we have the anti-phase synchronized state as in Fig. 10.5. In (b), we show the anti-phase-synchronized state for $\epsilon = 0.4$ and $\delta = 0.4$. In (c) and (d), we show the limit cycles in phase space (thick lines) of the angular coordinates and of the attachment points of the two pendulums. As the individual parameters of the two pendulum clocks are different, the two limit cycles have different shapes. The thin lines are the limit cycles for the cases $\epsilon = 0$ and $\delta = 0$. In (b), the period of oscillation is $T = 2.661$ and in (a) is $T = 2.463$. If the two non-linear pendulums have different parameters, the anti-phase synchronization condition, $\lim_{t_n \rightarrow \infty} \theta_1(t_n) = -\lim_{t_n \rightarrow \infty} \theta_2(t_n)$, introduced in Section 10.4 is no longer verified

Fixing ϵ to the value $\epsilon = 0.4$, the anti-phase-synchronized state still persists for $\delta \in [-0.1, 3.5]$. For $\delta = -0.2$, and the same initial conditions as in Fig. 10.9, the system in-phase synchronizes.

The comparison between Fig. 10.9(a) and (b) shows that the definition of anti-phase synchronization used in the previous section is specific to the case of non-linear oscillators with equal parameters. In Fig. 10.9(b), the two pendulums clearly synchronize in anti-phase, but, $\lim_{t_n \rightarrow \infty} \theta_1(t_n) \neq -\lim_{t_n \rightarrow \infty} \theta_2(t_n)$.

One important consequence of the approach presented here is that in order to synchronize oscillators characterized by different parameters, the equality between the eigenperiods of the oscillators is not required to obtain synchronization.

10.6 Conclusions

We have proposed a model of interaction between oscillators leading to exact anti-phase synchronization. This phenomenon has been observed for the first time, in 1665, by Christiaan Huygens.

The interaction parameters in our model are a damping constant ρ , a stiffness constant k of a linear spring, and a mass parameter M . This mass parameter is associated with the interaction, and not with the individual oscillators.

For moderated values of the (wet) damping constant ρ , $k > 0$, and $M > 0$, the asymptotic solutions of the model equations converge to a stable limit cycle in the eight-dimensional phase space of the model equations. Numerically, this limit cycle is the only stable attractor of the dynamics of the interacting oscillators. If the two pendulum clocks have the same initial conditions, the solutions synchronize with the same phase (in-phase), and this regime corresponds to an unstable attractor or limit cycle in the eight-dimensional phase space.

For smaller values of the damping constant ρ , three stable limit cycles in phase space coexist. Two of them correspond to anti-phase synchronized states of the two pendulum clocks, and the other corresponds to the in-phase synchronized state. The three limit cycles are reached by different initial conditions in the model equations.

The transition between the two different asymptotic regimes tuned by the damping parameter ρ , the first with one stable and one unstable limit cycle solutions, and the second one with three stable limit cycle solution, appears by a bifurcation from two limit cycles (one stable and the other unstable) to three stable limit cycles [12].

Changing the parameters of the individual pendulum clocks, we obtain the same synchronization properties as in the case of oscillators with identical parameters. This fact shows that the interaction mechanism purposed here is robust to changes in the parameters of the non-linear oscillators.

In all the cases analyzed numerically, the anti-phase and the in-phase synchrony occur with periods different from the eigenperiods of the individual oscillators. This shows that the equality between the eigenperiods of the individual oscillators is not required to obtain anti-phase or in-phase synchronization.

An important new issue introduced in the model is the possibility of existence of small movements of the attachment points of the pendulum clocks, a situation clearly avoided in the modern experimental devices, and diverging from the mechanism of synchrony proposed by Kortweg [9]. This explains why modern experimental setups have not been able to reproduce the original Huygens's results.

Acknowledgments I would like to thank the support of the Ettore Majorana Center for Scientific Culture and the hospitality of the organizers of the conference "Variational Analysis and Aerospace Engineering," dedicated to Prof. Angelo Miele on his 85th birthday. This work has been partially supported by a Fundação para a Ciência e a Tecnologia (FCT) pluriannual funding grant to the NonLinear Dynamics Group (GDNL).

References

1. A. Pikovsky, M. Rosenblum, J. Kurths, *Synchronization: A Universal Concept in Nonlinear Sciences*, Cambridge University Press, Cambridge, 2001.
2. M. Bennett, M.F. Schatz, H. Rockwood, K. Weisenfeld, Huygens's clocks, *Proc. R. Soc. Lond. A*, 458 (2002) 563–579.

3. M. Kumon, R. Washizaki, J. Sato, R.K.I. Mizumoto, Z. Iwai, Controlled synchronization of two 1-DOF coupled oscillators, in: Proc. of the 15th Triennial World Congress of IFAC, Barcelona, 2002.
4. A.L. Fradkov, B. Andrievsky, Synchronization and phase relations in the motion of two-pendulum system, *Int. J. Non-Linear Mechanics*, 42 (2007) 895–901.
5. B. Andrievsky, A. Fradkov, S. Gavrilov, V. Konoplev, Modeling and Synchronization of the Mechatronic Vibrational Stand, in Proc. 2nd Intern. Conf. "Physics and Control", IEEE, St. Petersburg, 2005, pp.165–168.
6. I.I. Blekhman, Yu.A. Bortsov, A.A. Burmistrov, A.L. Fradkov, S.V. Gavrilov, O.A. Kononov, B.P. Lavrov, V.M. Shestakov, P.V. Sokolov, O.P. Tomchina, Computer-controlled vibrational setup for education and research, in Proc. 14th IFAC World Congress, vol. M, 1999, pp. 193–197.
7. J. Pantaleone, Synchronization of metronomes, *Am. J. Phys.*, 70 (2002) 992–1000.
8. Y. Kuramoto, *Chemical Oscillations, Waves and Turbulence*, Springer-Verlag, Berlin, 1984.
9. D. J. Kortweg, *Les Horloges Sympathiques de Huygens*, Archives Neerlandaises, Series II, Tome XI, pp. 273–295, Martinus Nijhoff, The Hague, 1906.
10. S.H. Strogatz, I. Stewart, Coupled oscillators and biological synchronization, *Scient. Am.*, 269, n 6 (1993) 68–75.
11. A.A. Andronov, A.A. Witt, S.E. Khaikin, *Theory of Oscillators*, Pergamon, Oxford, 1966.
12. R. Dilão, Anti-phase and in-phase synchronization of nonlinear oscillators: The Huygens's clocks system, *Chaos* 19, 023118 (2009), DOI: 10.1063/1.3139117.
13. R. Hartman, *Ordinary Differential Equations*, Birkhäuser, Boston, 1982.
14. A. Hurwitz, On the Conditions under which an Equation has only Roots with Negative Real Parts, *Mathematische Annalen*, 46, (1895) 273–284. Reprinted in "Selected Papers on Mathematical Trends in Control Theory", R. Bellman, R. Kalaba (Ed.), Dover, New York, 1964.

“This page left intentionally blank.”

Chapter 11

Best Wing System: An Exact Solution of the Prandtl's Problem

Aldo Frediani and Guido Montanari

Abstract In 1924, Ludwig Prandtl published a fundamental paper in which he showed that the lifting system with minimum induced drag is a box wing. The results were obtained through an approximate procedure. In this chapter we obtain an exact solution of Prandtl's problem. In particular, we prove that the lift on the horizontal wings results from the superposition of a constant and an elliptical distribution and, on the vertical wings, it is butterfly shaped, as already shown by Prandtl. The discrepancies between the two solutions are discussed.

11.1 Introduction

The problem of minimum induced drag was tackled by Ludwig Prandtl before 1920. In 1924, he published a NACA paper [1], where he showed that a system with minimum induced drag exists and it is made of two horizontal wings connected at their tips with two vertical wings so as to obtain a box wing system. In order to introduce Prandtl's theory, we recall some classical results:

1. Among all the biplanes, there exist one with minimum induced drag, in which lift is equal on the two wings and the lift distribution is elliptical. The induced drag of the best biplane is lower than the induced drag of the best monoplane with the same span and total lift.
2. An optimum triplane exists, the induced drag of which is lower than that of a biplane (with the same total lift and wingspan); the same conclusion is valid for a configuration with $n + 1$ wings compared with n wings.

Aldo Frediani

Dipartimento di Ingegneria Aerospaziale, "L. Lazzarino", Università di Pisa, Pisa, Italy
e-mail: a.frediani@ing.unipi.it

Guido Montanari

Graduating Student in Mathematics,
Dipartimento di Matematica, "L. Tonelli", Università di Pisa, Pisa, Italy
e-mail: a.frediani@ing.unipi.it

According to Prandtl, the optimum drag arrangement of a multiplane is such that the upper and lower wings have the maximum lift and, in the internal wings, the lift decreases up to zero on the symmetry plane; when “ n ” (number of wings) is very high, any single wing can be approximated with a horseshoe tip vortex. The optimum system is a box one, where the two vertical wings are equivalent to tip vortex. This system was defined as “Best Wing System” by Prandtl. In the aforementioned Prandtl’s paper, the ratio between the induced drag of the best wing system and that of the best monoplane was calculated using an approximate procedure, though without a precise declaration of the underling assumptions.

In this chapter, the problem is tackled again with the aim of obtaining a closed form solution. The problem is that of assessing the circulation function along the vortex line which gives rise to the minimum induced drag in an asymptotic uniform stream. This condition of minimum is verified when the velocity induced by the free vortices is constant along the two horizontal wings and identically zero on the vertical wings. By applying this theorem, we can prove that, along the horizontal wings, the optimum lift distribution results from the superposition of a constant and an elliptical part and, on the vertical side wings, it is (according to Prandtl) butterfly shaped. The induced drag depends on the ratio h/b between the gap and the wingspan. The results show that, for values of h/b of possible practical applications (0.1–0.2), the induced drag predicted by Prandtl is nearly the same as that predicted by the present analysis. But, for higher values of h/b , Prandtl’s approximate solution is not accurate.

11.2 The Induced Drag for Lifting Multiwing Systems

We assume the lifting line theory, where $\Gamma(y)$ is the circulation function and $w(y)$ is the vertical component of the velocity induced by the trailing vortices (the only difference from zero). $\Gamma(y)$ satisfies the fundamental equation of the lifting line theory:

$$\Gamma(y) = \frac{1}{2} C_{l\alpha}(y) c(y) \left[V_{\infty} \alpha(y) - \frac{1}{4\pi} \int_{-\frac{b}{2}}^{\frac{b}{2}} \frac{d\Gamma}{d\eta} \frac{1}{\eta - y} d\eta \right]$$

where $\alpha(y)$, $C_{l\alpha}(y)$, and $c(y)$ are the local angle of attack, the derivative of the local lift coefficient, and the local wing chord, respectively; V_{∞} is the asymptotic flow velocity, b is the wingspan, and ρ is the air density.

The problem is, then, the following:

Given wing shape, ρ , V_{∞} , and the total lift $L = \rho V_{\infty} \int_{-\frac{b}{2}}^{\frac{b}{2}} \Gamma(y) dy$ to assess the local circulation $\Gamma(y)$ for which the induced drag is a minimum. In more precise terms the problem can be formulated as follows:

$$\begin{cases} \min D_i(y) \\ \text{subject to:} \\ \rho V_\infty \int_{-\frac{b}{2}}^{\frac{b}{2}} \Gamma(y) dy = L \\ \text{given geometry} \end{cases} \quad (11.1)$$

where

$$D_i = \rho \int_{-\frac{b}{2}}^{\frac{b}{2}} w(y) \Gamma(y) dy \quad (11.2)$$

is the induced drag.

In the case of a monoplane with minimum induced drag, $\Gamma(y)$ is elliptical, that is $\Gamma(y) = \Gamma_0 \sqrt{1 - (2y/b)^2}$; the induced velocity is constant along the span and $D_i = \frac{L^2}{q\pi b^2}$, where $q = \frac{1}{2}\rho V_\infty^2$ is the dynamic pressure. In the case of a biplane, we have

$$D_i = D_{11} + D_{22} + D_{12}^* + D_{21}^*, \quad (11.3)$$

where D_{hk}^* indicates the drag induced by the wing and trailing vortices of the wing h over the wing k and D_{kk} indicate the self-induced drag of the wing k .

Now, under classical simplifying hypotheses, we obtain

$$D_{11} = \frac{L_1^2}{q\pi b_1^2}, \quad D_{22} = \frac{L_2^2}{q\pi b_2^2} \quad (11.4)$$

where b_1 and b_2 are the upper and lower wing spans and

$$D_{12}^* + D_{21}^* = \rho \Gamma_1 \Gamma_2 \frac{4v_{12}}{\pi} \quad (11.5)$$

where $v_{12} = \frac{1}{4} \ln \left(\frac{\sqrt{l^2 + h^2}}{\sqrt{b^{*2} + h^2}} \right)$, $b^* = b_2 - b_1$ and $l = \frac{b_1 + b_2}{2}$ (Fig. 11.1); moreover, accepting the approximation that $\Gamma_1(y)$ and $\Gamma_2(y)$ are invariant along the span, we have the well-known result [2]:

$$D_i = D_{11} + D_{22} + D_{12}^* + D_{21}^* = \frac{1}{\pi q} \left(\frac{L_1^2}{b_1^2} + \frac{L_2^2}{b_2^2} + 2v_{12} \frac{L_1 L_2}{b_1 b_2} \right). \quad (11.6)$$

In the important case in which $b_1 = b_2 = b$, it is easy to prove that the induced drag is minimum when $L_1 = L_2 = L/2$ and the lift distribution is (very close to) elliptical on the two wings; the minimum induced drag is

$$D_i = \frac{L^2}{q\pi b^2} \left(\frac{1}{2} + \frac{1}{2} v_{12} \right) \quad (11.7)$$

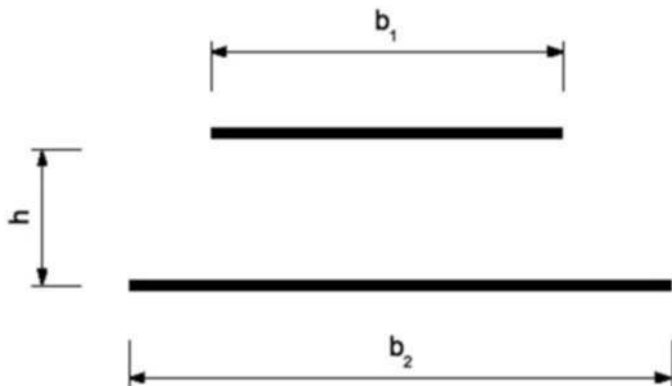


Fig. 11.1 Biplane geometric definitions

where $v_{12} = \frac{1}{4} \ln \left(\sqrt{\frac{b^2}{h^2} + 1} \right)$ (because $b^* = 0$ and $l = b$). By comparing a biplane with a monoplane with span b and elliptical lift distribution, we have

$$D_{i \text{ biplane}} \leq D_{i \text{ monoplane}}$$

when $\frac{b^2}{h^2} + 1 \leq e^8$ [2]. When $h \rightarrow \infty$, $v_{12} \rightarrow 0$; in this case $D_{i \text{ biplane}} = \frac{1}{2} D_{i \text{ monoplane}}$, whereas the condition $h \rightarrow 0$ (for which $v_{12} \rightarrow \infty$) is not considered here because the result depends on the hypothesis (not valid, in general) that the wake vortices are concentrated on the wing tips.

The condition of minimum induced drag for a given lift and a given front view of the system stems from the following two theorems due to Munk [3, 4].

Theorem 11.1. *The induced drag of a lifting system is invariant when the wake vortices are translated along the asymptotic speed direction.*

Owing to this theorem, all the vortices can be assumed to be on a plane normal to the asymptotic speed and the induced velocity w_n is due only to the trailing vortices (the contribution of the wing vortices is zero).

Theorem 11.2. *If two lifting elements are on the same plane normal to the asymptotic speed, the drag induced by the first on the second equals that induced by the second on the first.*

From these theorems it follows that the induced velocity in any point of an optimum lifting system is

$$w_n = w_0 \cos \varepsilon, \quad (11.8)$$

where w_0 is a constant and ε is the angle of the element ds with horizontal line [3, 4].

11.3 The Problem of Minimum Induced Drag in a Box Wing

We consider a box wing system composed of two horizontal wings connected by two vertical wings at the tips and, moreover, we suppose that the conditions of minimum induced drag are satisfied. Owing to the previous results, the induced velocity is identically zero on the side wings and is constant on the horizontal wings; the problem is to determine the optimum circulation function on the lifting system.

With reference to Fig. 11.2, let P_1 a point on the wing 1; the velocity induced in the point P_1 by the whole trailing vorticity can be calculated according to Biot-Savart's law, that is,

$$w(y_1^1) = \frac{1}{4\pi} \left[\int_{-b/2}^{b/2} \frac{d\Gamma_1}{dy_1} \frac{1}{y_1 - y_1^1} dy_1 + \int_{-b/2}^{b/2} \frac{d\Gamma_2}{dy_2} \frac{\cos^2 \alpha}{y_2 - y_1^1} dy_2 + \right. \\ \left. + \int_0^h \frac{d\Gamma_3}{dz_3} \frac{\cos \beta \sin \beta}{h - z_3} dz_3 + \int_0^h \frac{d\Gamma_4}{dz_4} \frac{\cos \delta \sin \delta}{z_4 - h} dz_4 \right]. \quad (11.9)$$

In the same way, the induced velocities on points P_2 , P_3 , and P_4 are

$$w(y_2^1) = \frac{1}{4\pi} \left[\int_{-b/2}^{b/2} \frac{d\Gamma_2}{dy_2} \frac{1}{y_2 - y_2^1} dy_2 + \int_{-b/2}^{b/2} \frac{d\Gamma_1}{dy_1} \frac{\cos^2 \alpha}{y_1 - y_2^1} dy_1 + \right. \\ \left. + \int_0^h \frac{d\Gamma_3}{dz_3} \frac{\cos \varepsilon \sin \varepsilon}{z_3} dz_3 - \int_0^h \frac{d\Gamma_4}{dz_4} \frac{\cos \theta \sin \theta}{z_4} dz_4 \right], \quad (11.10)$$

$$w(z_3^1) = \frac{1}{4\pi} \left[\int_{-b/2}^{b/2} \frac{d\Gamma_1}{dy_1} \frac{\cos^2 \beta}{z_3^1 - h} dy_1 + \int_{-b/2}^{b/2} \frac{d\Gamma_2}{dy_2} \frac{\sin^2 \varepsilon}{z_3^1} dy_2 + \right. \\ \left. - \int_0^h \frac{d\Gamma_4}{dz_4} \frac{\cos^2 \nu}{z_4 - z_3^1} dz_4 - \int_0^h \frac{d\Gamma_3}{dz_3} \frac{1}{z_3 - z_3^1} dz_3 \right], \quad (11.11)$$

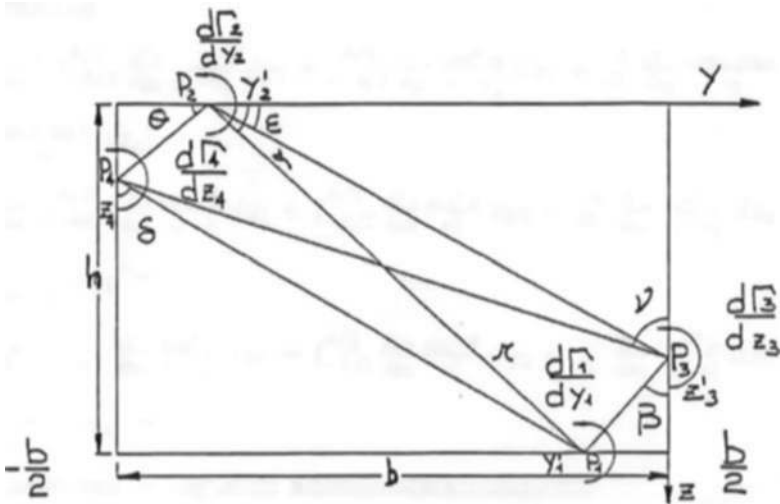


Fig. 11.2 Boxplane reference configuration

$$\begin{aligned}
 w(z_4^1) = \frac{1}{4\pi} & \left[\int_{-b/2}^{b/2} \frac{d\Gamma_1}{dy_1} \frac{\cos^2 \delta}{h - z_4^1} dy_1 - \int_{-b/2}^{b/2} \frac{d\Gamma_2}{dy_2} \frac{\sin^2 \theta}{z_4^1} dy_2 + \right. \\
 & \left. + \int_0^h \frac{d\Gamma_3}{dz_3} \frac{\cos^2 \nu}{z_3 - z_4^1} dz_3 + \int_0^h \frac{d\Gamma_4}{dz_4} \frac{1}{z_4 - z_4^1} dz_4 \right].
 \end{aligned} \quad (11.12)$$

Now, referring to Fig. 11.1 we introduce the following non-dimensional quantities:

$$\eta = \frac{y}{b/2}, \quad \mu = \frac{z}{h} \quad (11.13)$$

and

$$k = \frac{h}{b/2} \quad (11.14)$$

and apply the conditions of minimum induced drag; after some computation (recorded in Appendix 1), we obtain the following equations:

$$\begin{aligned}
 \int_{-1}^1 \frac{d\Gamma_1}{d\eta_1} \frac{1}{\eta_1 - \eta_1^1} d\eta_1 + \int_{-1}^1 \frac{d\Gamma_2}{d\eta_2} \frac{(\eta_2 - \eta_1^1)}{[(\eta_2 - \eta_1^1)^2 + k^2]} d\eta_2 + \\
 + \int_0^1 \frac{d\Gamma_3}{d\mu_3} \frac{(1 - \eta_1^1)}{[(1 - \eta_1^1)^2 + k^2(1 - \mu_3)^2]} d\mu_3 + \\
 - \int_0^1 \frac{d\Gamma_4}{d\mu_4} \frac{(1 + \eta_1^1)}{[(1 + \eta_1^1)^2 + k^2(1 - \mu_4)^2]} d\mu_4 = 2\pi b w_0,
 \end{aligned} \quad (11.15)$$

$$\begin{aligned}
 \int_{-1}^1 \frac{d\Gamma_1}{d\eta_1} \frac{(\eta_1 - \eta_2^1)}{[(\eta_1 - \eta_2^1)^2 + k^2]} d\eta_1 + \int_{-1}^1 \frac{d\Gamma_2}{d\eta_2} \frac{1}{\eta_2 - \eta_2^1} d\eta_2 + \\
 + \int_0^1 \frac{d\Gamma_3}{d\mu_3} \frac{(1 - \eta_2^1)}{[(1 - \eta_2^1)^2 + k^2\mu_3^2]} d\mu_3 + \\
 - \int_0^1 \frac{d\Gamma_4}{d\mu_4} \frac{(1 + \eta_2^1)}{[(1 + \eta_2^1)^2 + k^2\mu_4^2]} d\mu_4 = 2\pi b w_0,
 \end{aligned} \quad (11.16)$$

$$\begin{aligned}
 k^2 \int_{-1}^1 \frac{d\Gamma_1}{d\eta_1} \frac{(\mu_3^1 - 1)}{[(1 - \eta_1)^2 + k^2(\mu_3^1 - 1)^2]} d\eta_1 + \\
 + k^2 \int_{-1}^1 \frac{d\Gamma_2}{d\eta_2} \frac{\mu_3^1}{[(1 - \eta_2)^2 + k^2(\mu_3^1)^2]} d\eta_2 + \\
 + k^2 \int_0^1 \frac{d\Gamma_4}{d\mu_4} \frac{(\mu_3^1 - \mu_4)}{[4 + k^2(\mu_3^1 - \mu_4)^2]} d\mu_4 + \int_0^1 \frac{d\Gamma_3}{d\mu_3} \frac{1}{\mu_3^1 - \mu_3} d\mu_3 = 0,
 \end{aligned} \quad (11.17)$$

$$\begin{aligned}
& k^2 \int_{-1}^1 \frac{d\Gamma_1}{d\eta_1} \frac{(1 - \mu_3^1)}{\left[(1 + \eta_1)^2 + k^2 (1 - \mu_3^1)^2\right]} d\eta_1 + \\
& - k^2 \int_{-1}^1 \frac{d\Gamma_2}{d\eta_2} \frac{\mu_3^1}{\left[(1 + \eta_2)^2 + k^2 (\mu_3^1)^2\right]} d\eta_2 + \\
& k^2 \int_0^1 \frac{d\Gamma_3}{d\mu_3} \frac{(\mu_3 - \mu_3^1)}{\left[4 + k^2 (\mu_3 - \mu_3^1)^2\right]} d\mu_3 + \int_0^1 \frac{d\Gamma_4}{d\mu_4} \frac{1}{\mu_4 - \mu_4^1} d\mu_4 = 0.
\end{aligned} \tag{11.18}$$

Due to the symmetry, the two first equations become identical if we put $\mu_3 = 1 - \mu_3$ and $\mu_4 = 1 - \mu_4$, and the same occurs for the other equations if we put $\eta_1 = -\eta_1$ and $\eta_2 = -\eta_2$. Hence, putting $\Gamma_4 = -\Gamma_3$, the problem is reduced to the solution of the two integral equations:

$$\begin{aligned}
& \int_{-1}^1 \frac{d\Gamma_1}{d\eta_1} \frac{1}{\eta_1 - \eta_1^1} d\eta_1 + \int_{-1}^1 \frac{d\Gamma_2}{d\eta_2} \frac{(\eta_2 - \eta_1^1)}{\left[(\eta_2 - \eta_1^1)^2 + k^2\right]} d\eta_2 + \\
& + \int_0^1 \frac{d\Gamma_3}{d\mu_3} \frac{(1 - \eta_1^1)}{\left[(1 - \eta_1^1)^2 + k^2 (1 - \mu_3)^2\right]} d\mu_3 + \\
& \int_0^1 \frac{d\Gamma_3}{d\mu_3} \frac{(1 + \eta_1^1)}{\left[(1 + \eta_1^1)^2 + k^2 (1 - \mu_3)^2\right]} d\mu_3 = 2b\pi w_0,
\end{aligned} \tag{11.19}$$

$$\begin{aligned}
& k^2 \int_{-1}^1 \frac{d\Gamma_1}{d\eta_1} \frac{(\mu_3^1 - 1)}{\left[(1 + \eta_1)^2 + k^2 (1 - \mu_3^1)^2\right]} d\eta_1 + \\
& + k^2 \int_{-1}^1 \frac{d\Gamma_2}{d\eta_2} \frac{\mu_3^1}{\left[(1 + \eta_2)^2 + k^2 (\mu_3^1)^2\right]} d\eta_2 + \\
& - k^2 \int_0^1 \frac{d\Gamma_3}{d\mu_3} \frac{(\mu_3^1 - \mu_3)}{\left[4 + k^2 (\mu_3 - \mu_3^1)^2\right]} d\mu_3 + \int_0^1 \frac{d\Gamma_3}{d\mu_3} \frac{1}{\mu_3^1 - \mu_3} d\mu_3 = 0.
\end{aligned} \tag{11.20}$$

In (11.19) and (11.20), Γ_1 , Γ_2 , and Γ_3 are the unknown functions, defined on the horizontal and vertical wings, respectively. As for Γ_1 , it is easy to prove the following.

Proposition 11.1 *The optimum lift distribution on the horizontal wings is elliptical plus a constant.*

Proof. The proof is similar to the case of a biplane, taking into account that the induced drag produced by the vertical wings is zero because $w = 0$ (in the optimality conditions) [5]. It is worth noting that, contrary to a biplane, the lift distribution includes a constant, unknown, corresponding to $\Gamma\left(-\frac{b}{2}\right) = \Gamma\left(\frac{b}{2}\right) = \text{constant}$. Apart from this constant we obtain (11.21) and the minimum occurs when $a_n = c_n = 0$, $\forall n$; due to (11.7), we have the same solution as the best monoplane (elliptical lift).

We assume that Γ_1 and Γ_1' to be continuous and limited and, suppose to use the angular variable $\theta \in (0, \pi)$ with $\Gamma_1(0) = \Gamma_1(\pi) = c$, where, contrary to a conventional lifting segment, $c \neq 0$. Apart from the constant

$$\Gamma_1(\theta) = \sum_{n=1}^{\infty} a_n \sin(n\theta),$$

and

$$w_1(\theta) = -\frac{1}{2b} \sum_{n=1}^{\infty} n a_n \frac{\sin(n\theta)}{\sin \theta}.$$

Now we calculate the induced drag

$$D_i = D_{11} + D_{22} + D_{12}$$

where we put, $D_{12} = D_{12}^* + D_{21}^*$ and, for simplicity, assume again that the circulation distributions are constant on both the wings.

It is easy to show that D_{12} does not depend on the circulation on the side wings (where, $w = 0$ for hypothesis) and we obtain

$$D_{12} \simeq 2v_{12} \frac{L_1 L_2}{\pi q b^2}$$

$$D_{11} = \rho \int_{-\frac{b}{2}}^{\frac{b}{2}} \Gamma(y) w(y) dy = \frac{\rho}{8} \pi (a_1^2 + 2a_2^2 + \dots + na_n^2 + \dots)$$

$$L_1 = \rho V_{\infty} \int_{-\frac{b}{2}}^{\frac{b}{2}} \Gamma(y) dy = \rho V_{\infty} \frac{b}{4} \pi a_1 \Rightarrow 2 \frac{L_1^2}{\pi b^2 \rho V_{\infty}^2} = \frac{2}{\pi b^2 \rho V_{\infty}^2} \rho^2 V_{\infty}^2 \frac{b^2}{16} \pi^2 a_1^2 = \frac{\rho}{8} \pi a_1^2$$

and similar expressions are valid for D_{22} and L_2 so that, finally, we obtain

$$D_i = \frac{L^2}{q \pi b^2} \left(\frac{1}{2} + \frac{1}{2} v_{12} \right) + \frac{\rho}{8} \pi (2a_2^2 + 2c_2^2 + \dots + na_n^2 + nc_n^2 + \dots) \quad (11.21)$$

$$\text{where } v_{12} = \frac{1}{4} \ln \left(\sqrt{\frac{b^2}{h^2} + 1} \right).$$

The first term on the right-hand side gives the induced drag of the biplane with elliptical circulation on the two wings and the second term is the sum of positive terms; so, the induced drag is minimum when the second term is zero and, then, the circulation distribution on both the wings is elliptical, apart from a constant.

Hence, for clarity sake, we can reduce the general solution to the two cases: (A) the circulation is elliptical on the two horizontal wings and identically zero on the side wings (Fig. 11.3) and (B) the circulation is constant on the two horizontal wings and unknown on the side ones (Fig. 11.4).

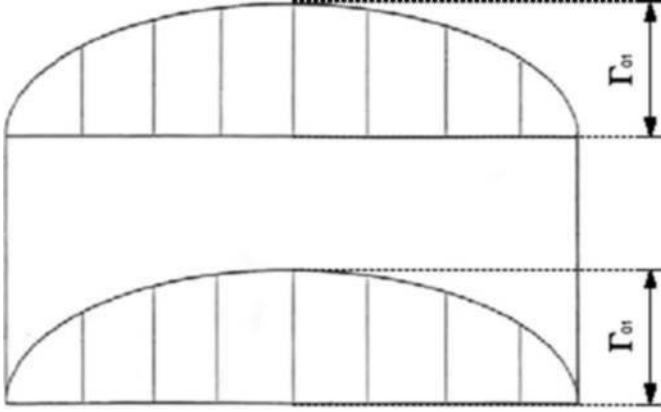


Fig. 11.3 Problem A: elliptical circulation along the horizontal wing vortex lines

11.3.1 Case A: Elliptical Circulations on the Horizontal Wings and Zero on the Vertical Ones

We start from

$$\Gamma_3 = 0, \quad (11.22)$$

$$\Gamma_1 = \Gamma_2 = \Gamma_{01} \sqrt{1 - \eta_1^2} \quad (11.23)$$

and using polar coordinates: $\eta_1 = \cos \varphi_1$ and $\eta_2 = \cos \varphi_2$, the induced velocity on the horizontal wings results from (11.19):

$$w_H^{(A)} = - \int_0^\pi \frac{\Gamma_{01} \cos \varphi_1}{\cos \varphi_1 - \cos \varphi_1^1} d\varphi_1 - \int_0^\pi \frac{\Gamma_{02} \cos \varphi_2 (\cos \varphi_2 - \cos \varphi_1^1)}{[(\cos \varphi_2 - \cos \varphi_1^1)^2 + k^2]} d\varphi_2. \quad (11.24)$$

Owing to the Glauert's integral, $\int_0^\pi \frac{\cos n\theta}{\cos \theta - \cos \phi} d\theta = \pi \frac{\sin n\phi}{\sin \phi}$ and $\Gamma_{02} = \Gamma_{01}$, the first integral becomes

$$- \int_0^\pi \frac{\Gamma_{01} \cos \varphi_1}{\cos \varphi_1 - \cos \varphi_1^1} d\varphi_1 = -\Gamma_{01} \pi$$

and the second integral is obtained in Appendix 2. As a final result we obtain

$$w_H^{(A)} = \pi \left[-\Gamma_{01} - \Gamma_{01} \frac{A}{(A^2 + k^2)} + \Gamma_{01} \frac{4}{(A^2 + k^2)} \left(\sqrt{\frac{2}{(A^2 + k^2)(\sqrt{V} - a)}} G + \right. \right. \\ \left. \left. - \frac{2k}{(A^2 + k^2)(\sqrt{V} - a)} F + \frac{R}{2(A^2 + k^2)} \right) \right]. \quad (11.25)$$

The induced velocity on the vertical wings is calculated from (11.20), under the conditions (11.23)

$$w_V^{(A)} = \frac{k^2}{4\pi} [(1 - \mu_3^1) \Gamma_{01} \int_0^\pi \frac{\cos \varphi_1}{[(1 - \cos \varphi_1)^2 + k^2 (\mu_3^1 - 1)^2]} d\varphi_1 + \\ - \mu_3^1 \Gamma_{02} \int_0^\pi \frac{\cos \varphi_2}{[(1 - \cos \varphi_2)^2 + k^2 (\mu_3^1)^2]} d\varphi_2] = 0$$

These integrals are calculated in Appendix 3 and the result is

$$w_V^{(A)} = \frac{k^2}{4} \left\{ (1 - \mu_3') \Gamma_{01} \sqrt{\frac{\sqrt{4 + k^2 (1 - \mu_3^1)^2} - k (1 - \mu_3')}{2k (1 - \mu_3') (4 + k^2 (1 - \mu_3')^2)}} (I_2 + \right. \\ \left. - H_2 \sqrt{\frac{k (1 - \mu_3') (\sqrt{4 + k^2 (1 - \mu_3^1)^2} - k (1 - \mu_3'))}{2 (4 + k^2 (1 - \mu_3')^2)}}) - \mu_3' \Gamma_{01} \right. \\ \left. \sqrt{\frac{\sqrt{4 + k^2 \mu_3^{12}} - k \mu_3'}{2 (4 + k^2 \mu_3'^2) k \mu_3'}} \left(I_1 - H_1 \sqrt{\frac{k \mu_3' (\sqrt{4 + k^2 \mu_3^{12}} - k \mu_3')}{2 (4 + k^2 \mu_3'^2)}} \right) \right\} \quad (11.26)$$

The symbols introduced in (11.25) and (11.26) are defined in Appendices 2 and 3.

11.3.2 Case B: Constant Circulations on the Horizontal Wings and Unknown on the Vertical Ones

We first remark that the circulation is continuous when passing from a horizontal wing to a vertical wing and vice versa; so, the circulation on top and bottom of the side wings is the same of that at the tips of the corresponding horizontal wings; we indicate this intensity as A_0 . As said before, A_0 is unknown together with the circulation function along the lateral wings (Fig. 11.4). The circulation along the lateral wings is assumed to be a cubic, this assumption is supposed to be adequate for applications of aircraft engineering:

$$\Gamma_3(\mu_3) = a\mu_3^3 + b\mu_3^2 + c\mu_3 + d \quad (11.27)$$

and we determine the coefficients a, b, c , and d by imposing that the boundary conditions are satisfied. The boundary conditions are

$$\Gamma_3(0) = 1, \quad \Gamma_3(1/2) = 0, \quad \Gamma_3(1) = -1. \quad (11.28)$$

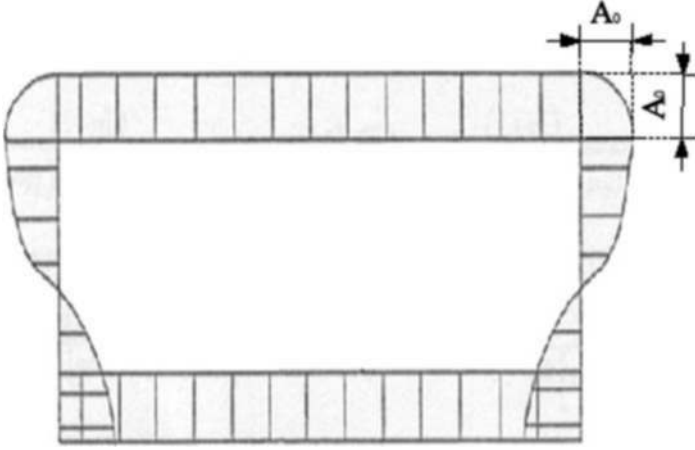


Fig. 11.4 Problem B: circulation distribution along the vortex line

By putting $e := \frac{\partial \Gamma_3}{\partial \mu_3} \Big|_{\mu_3=\frac{1}{2}}$, we obtain

$$a = -4(e+2), \quad b = 6e+12, \quad c = -2e-6, \quad d = 1, \quad (11.29)$$

and the circulation distribution is

$$\Gamma_3(\mu_3) = A_0 [-4(e+2)\mu_3^3 + 6(e+2)\mu_3^2 - 2(e+3)\mu_3 + 1]. \quad (11.30)$$

From (11.19), the induced velocity on the horizontal wing is

$$w_H^{(B)} = \int_0^1 \frac{d\Gamma_3}{d\mu_3} \frac{(1-\eta_1^1)}{[(1-\eta_1^1)^2 + k^2(1-\mu_3)^2]} d\mu_3 + \int_0^1 \frac{d\Gamma_3}{d\mu_3} \frac{(1+\eta_1^1)}{[(1+\eta_1^1)^2 + k^2(1-\mu_3)^2]} d\mu_3$$

or (as shown in Appendix 4)

$$\begin{aligned} w_H^{(B)} = A_0 & \left\{ \arctan \left(\frac{k}{1-\eta_1^1} \right) \left[\frac{12(e+2)(1-\eta_1^1)^2}{k^3} - \frac{2(e+3)}{k} \right] + \right. \\ & \arctan \left(\frac{k}{1+\eta_1^1} \right) \left[-\frac{2(e+3)}{k} \frac{12(e+2)(1+\eta_1^1)}{k^3} \right] + \\ & \ln \left[\frac{(1-\eta_1^1)^2}{k^2 + (1-\eta_1^1)^2} \right] \left(-\frac{6(e+2)(1-\eta_1^1)}{k^2} \right) - \frac{6(e+2)(1+\eta_1^1)}{k^2} \\ & \left. \ln \left[\frac{(1+\eta_1^1)^2}{k^2 + (1+\eta_1^1)^2} \right] - \frac{24(e+2)}{k^2} \right\} \end{aligned} \quad (11.31)$$

On the lateral wings we have, from (11.20)

$$w_V^{(B)} = \frac{1}{4\pi} \left\{ -k^2 \int_0^1 \frac{[-12(e+2)\mu_3^2 + 12(e+2)\mu_3 - 2(e+3)](\mu_3^1 - \mu_3)}{[4 + k^2(\mu_3^1 - \mu_3)^2]} d\mu_3 + \right. \\ \left. + \int_0^1 \frac{-12(e+2)\mu_3^2 + 12(e+2)\mu_3 - 2(e+3)}{\mu_3^1 - \mu_3} d\mu_3 \right\}$$

and, according to calculations reported in Appendix 4, we obtain

$$w_V^{(B)} = \frac{A_0}{4\pi} [24(e+2)\mu_3^1] + \ln\left(\frac{1-\mu_3^1}{\mu_3^1}\right)^2 (6(e+2)(\mu_3^1)^2 - 6(e+2)\mu_3^1 + \\ + e+3) + \ln\left(\frac{4+k^2(\mu_3^1-1)^2}{4+k^2(\mu_3^1)^2}\right) \left(\frac{6(e+2)(4+k^2(\mu_3^1)^2)}{k^2} - e+3\right) + \\ + 12(e+2)k^2(1+\mu_3^1) \left[-\frac{1}{k^2} - \frac{\mu_3^1}{2k^2} \ln\left(\frac{4+k^2(\mu_3^1-1)^2}{4+k^2(\mu_3^1)^2}\right) + \right. \\ \left. - \frac{2}{k^3} \left(\arctan\left(\frac{k}{2}(\mu_3^1-1)\right) - \arctan\left(\frac{k}{2}\mu_3^1\right)\right)\right] + \frac{k^2}{4} [(1-\mu_3^1) \\ \Gamma_{01} \sqrt{\frac{\sqrt{4+k^2(1-\mu_3^1)^2} - (1-\mu_3^1)}{2k(1-\mu_3^1)(4+k^2(1-\mu_3^1)^2)}} (I_2 + \\ - H_2 \sqrt{\frac{k(1-\mu_3^1)(\sqrt{4+k^2(1-\mu_3^1)^2} - (1-\mu_3^1))}{2(4+k^2(1-\mu_3^1)^2)}}) + \\ - \mu_3^1 \Gamma_{01} \sqrt{\frac{\sqrt{4+k^2\mu_3^{12}} - k\mu_3^1}{2(4+k^2\mu_3^{12})k\mu_3^{12}}} \left(I_1 - H_1 \sqrt{\frac{k\mu_3^1(\sqrt{4+k^2\mu_3^{12}} - k\mu_3^1)}{24+k^2\mu_3^{12}}}\right) \Big] \quad (11.32)$$

11.3.3 Final Equations

Collecting together the solutions of problems A and B, we have

$$w_H^{(A)} + W_H^{(B)} = 2\pi b w_0 \quad (11.33)$$

$$w_V^{(A)} + w_V^{(B)} = 0 \quad (11.34)$$

independent of the local coordinates η'_1 (on the horizontal wings) and μ'_3 (on the lateral wings)

$$\begin{aligned}
 A_0 & \left[\arctan\left(\frac{k}{1-\eta_1^1}\right) \left(\frac{12(e+2)(1-\eta_1^1)^2}{k^3} - \frac{2(e+3)}{k} \right) + \arctan\left(\frac{k}{1+\eta_1^1}\right) \right. \\
 & \left(\frac{12(e+2)(1+\eta_1^1)^2}{k^3} + \frac{-2(e+3)}{k} \right) - \frac{6(e+2)(1-\eta_1^1)}{k^2} \\
 & \ln\left(\frac{(1-\eta_1^1)^2}{k^2 + (1-\eta_1^1)^2}\right) - \frac{24(e+2)}{k^2} - \frac{6(e+2)(\eta_1^1+1)}{k^2} \\
 & \left. \ln\left(\frac{(1+\eta_1^1)^2}{k^2 + (1+\eta_1^1)^2}\right) \right] + \pi \left[-\Gamma_{01} - \Gamma_{01} \frac{A}{A^2+k^2} + \Gamma_{01} \frac{4}{A^2+k^2} \right. \\
 & \left. \left(\sqrt{\frac{2}{(A^2+k^2)(\sqrt{V}-a)}} G - \frac{2k}{(A^2+k^2)(\sqrt{V}-a)} F + \frac{1}{2(4+k^2)} R \right) \right] = 2\pi b w_0,
 \end{aligned} \tag{11.35}$$

$$\begin{aligned}
 \frac{A_0}{4\pi} & \left\{ 24(e+2)\mu_3^1 + \ln\left(\frac{1-\mu_3^1}{\mu_3^1}\right)^2 \left(6(e+2)(\mu_3^1)^2 - 6(e+2)\mu_3^1 + e+3 \right) + \right. \\
 & \ln\left(\frac{4+k^2(\mu_3^1-1)^2}{4+k^2(\mu_3^1)^2}\right) \left(\frac{6(e+2)(4+k^2(\mu_3^1)^2)}{k^2} - (e+3) \right) + 12(e+2)k^2 \\
 & (1+\mu_3^1) \left[-\frac{1}{k^2} - \frac{\mu_3^1}{2k^2} \ln\left(\frac{4+k^2(\mu_3^1-1)^2}{4+k^2(\mu_3^1)^2}\right) - \frac{2}{k^3} \left(\arctan\left(\frac{k}{2}(\mu_3^1-1)\right) + \right. \right. \\
 & \left. \left. - \arctan\left(\frac{k}{2}\mu_3^1\right) \right) \right] \right\} + \frac{k^2}{16} \left[(1-\mu_3^1)\Gamma_{01} \sqrt{\frac{\sqrt{4+k^2(1-\mu_3^1)^2} - (1-\mu_3^1)}{2k(1-\mu_3^1)(4+k^2(1-\mu_3^1)^2)}} \right. \\
 & \left(I_2 - H_2 \sqrt{\frac{k(1-\mu_3^1)(\sqrt{4+k^2(1-\mu_3^1)^2} - (1-\mu_3^1))}{2(4+k^2(1-\mu_3^1)^2)}} \right) + \\
 & \left. -\mu_3^1\Gamma_{01} \sqrt{\frac{\sqrt{4+k^2\mu_3^{12}} - k\mu_3^1}{2(4+k^2\mu_3^{12})k\mu_3^{12}}} \left(I_1 - H_1 \sqrt{\frac{k\mu_3^1(\sqrt{4+k^2\mu_3^{12}} - k\mu_3^1)}{24+k^2\mu_3^{12}}} \right) \right] = 0,
 \end{aligned} \tag{11.36}$$

where, as said before, the quantities present in (11.35) and (11.36) are defined in Appendices 2 and 3, and

$$I_2 = \frac{\sqrt{4 + k^2 (1 - \mu_3^1)^2}}{k (1 - \mu_3^1)},$$

$$H_2 = \frac{\sqrt{4 + k^2 \mu_3^{12}} \left(k (1 - \mu_3^1) + \sqrt{4 + k^2 (1 - \mu_3^1)^2} \right)}{2k (1 - \mu_3^1) \sqrt{k (1 - \mu_3^1) \left(\sqrt{4 + k^2 (1 - \mu_3^1)^2} - k (1 - \mu_3^1) \right)}}.$$

The problem of optimum is not influenced by the lift resultant; so, for simplicity sake, we assume

$$A_0 + \Gamma_{01} = 1. \quad (11.37)$$

The circulation distributions along the horizontal and lateral wings depend on the parameter e ; the problem of the minimum induced drag is reduced to the problem of the optimum value of e .

It is worth noting that the only non-dimensional geometric parameter is $k = \frac{h}{b/2}$.

11.4 The Optimum Lift Distribution Along the Vertical Wings

The optimum value of e cannot be obtained in a closed form, but it can be easily obtained numerically. Because e is present in both the equations, the optimization regards both of them. First, we choose to find the value of e which optimize the effects on the lateral wings and, therefore, we consider Eq. (11.36); after that, we verify that this optimum value of e minimizes the error in (11.35). Now, we have to define the concept of error; we assume as error the L^1 norm of the difference between the right- and the left-hand sides of (11.36) and, consequently, we obtain the following problem of minimum:

$$\min_e \int_0^1 |A_0 [(e+2)f(\mu_3^1) + g(\mu_3^1)] + h(\mu_3^1)| d\mu_3^1, \quad (11.38)$$

where the functions introduced have the following expressions:

$$\begin{aligned} f(\mu_3^1) = & \frac{1}{4\pi} \left\{ 24\mu_3^1 + \ln \left(\frac{1 - \mu_3^1}{\mu_3^1} \right)^2 \left(6(\mu_3^1)^2 - 6\mu_3^1 + 1 \right) \right. \\ & + \ln \left(\frac{4 + k^2(\mu_3^1 - 1)^2}{4 + k^2(\mu_3^1)^2} \right) \left(\frac{6(4 + k^2(\mu_3^1)^2)}{k^2} - 1 \right) + \\ & + 12k^2(1 + \mu_3^1) \left[-\frac{1}{k^2} - \frac{\mu_3^1}{2k^2} \ln \left(\frac{4 + k^2(\mu_3^1 - 1)^2}{(4 + k^2(\mu_3^1)^2)} \right) \right. \\ & \left. \left. - \frac{2}{k^3} \left[\arctan \left(\frac{k}{2}(\mu_3^1 - 1) \right) - \arctan \left(\frac{k}{2}\mu_3^1 \right) \right] \right] \right\}, \end{aligned}$$

$$g(\mu_3^1) = \frac{1}{4\pi} \left[\ln \left(\frac{1-\mu_3^1}{\mu_3^1} \right)^2 - \ln \left(\frac{4+k^2(\mu_3^1-1)^2}{4+k^2(\mu_3^1)^2} \right) \right],$$

$$h(\mu_3^1) = \frac{k^2}{4} \left[(1-\mu_3^1) \Gamma_{01} \sqrt{\frac{\sqrt{4+k^2(1-\mu_3^1)^2} - (1-\mu_3^1)}{2k(1-\mu_3^1)(4+k^2(1-\mu_3^1)^2)}} \right. \\ \left. (I_2 - H_2 \sqrt{\frac{k(1-\mu_3^1)(\sqrt{4+k^2(1-\mu_3^1)^2} - (1-\mu_3^1))}{2(4+k^2(1-\mu_3^1)^2)}}) + \right. \\ \left. -\mu_3^1 \Gamma_{01} \sqrt{\frac{\sqrt{4+k^2\mu_3^{12}} - k\mu_3^1}{2(4+k^2\mu_3^{12})k\mu_3^{12}}} \left(I_1 - H_1 \sqrt{\frac{k\mu_3^1(\sqrt{4+k^2\mu_3^{12}} - k\mu_3^1)}{24+k^2\mu_3^{12}}} \right) \right].$$

In the previous expression, the term $A_0 [(e+2)f(\mu_3^1) + g(\mu_3^1)]$ is the contribution of the lateral wings and $h(\mu_3^1)$ is the elliptical lift distribution on the horizontal wings.

In practice, we can fix a value of $A_0 \in (0, 1)$ and, by a numerical method examine the behavior of the function

$$m(\mu_3^1) := A_0 [(e+2)f(\mu_3^1) + g(\mu_3^1)] + h(\mu_3^1). \quad (11.39)$$

Numerical computation (Table 11.1) shows that the error is minimum when

$$e = -2; \quad (11.40)$$

from (11.29), the lift distribution along the vertical wings results to be linear and (with reference to the chosen reference frame), on the lateral wings, butterfly shaped; this result was predicted by Prandtl [1].

Table 11.1 Assessment of the optimum value of A_0

e	-4	-3	-2	-1
A_0	0.007	0.002	0.256	0.003
f	0.058	0.057	0.0137	0.060
<i>error</i>	5.8%	5.7%	1.37%	6.0%

11.5 Results and Conclusions

The circulation along the vortex line corresponding to the minimum of induced drag is made from the superposition of a constant and an elliptical distribution on the horizontal wings and butterfly shaped on the vertical ones. The induced drag of the box wing depends only on the non-dimensional parameter $k = 2h/b$; now, the problem is to determine, for any value of k , the rate $A_0/(1-A_0)$ between the

constant part and the maximum of the elliptical lift on the horizontal wings. This problem is solved by means of an algorithm based on the following steps:

- (i) numerical calculation of the L^1 norm of the difference between the right- and left-hand sides of one of the two equations of the system, taking A_0 as a parameter and
- (ii) assessment of the A_0 value for which the induced drag is minimum.

Preliminarily, for numerical reasons, we have to decide whether (11.35) (relevant to the horizontal wing) or (11.36) (vertical wing) has to be considered. Equation (11.35) is the most appropriate because the induced drag depends on the square of the lift and the whole lift is applied on the horizontal wings (on the vertical wings, the total induced drag is zero). Now, for any value of k , we search the optimum value of A_0 , using (11.35).

Figure 11.5 shows, for a given value of k , an example of error functions for the assessment of the optimum A_0 value, using both (11.35) and (11.36). It is evident that, using (11.35), the minimum is very sharp compared to the second one. Of course, the previous figure tells us that the optimum distribution made of a constant plus elliptical lifts on the horizontal wings has a strong influence on the same horizontal wings (where the whole induced drag takes place) and the opposite occurs on the side wings.

Finally, for any geometry, we can calculate the minimum induced drag and compare it with that of the best monoplane with the same total lift and wingspan. The results are shown in Fig. 11.6, together with the results obtained by Prandtl in [1]. The present and Prandtl's results are very close to each other for small values of k , but are somewhat different when tends to ∞ . In Prandtl's paper, we have

$$\frac{D_{bws}}{D_{bm}} = \frac{1 + 0.45h/b}{1.04 + 2.81h/b},$$

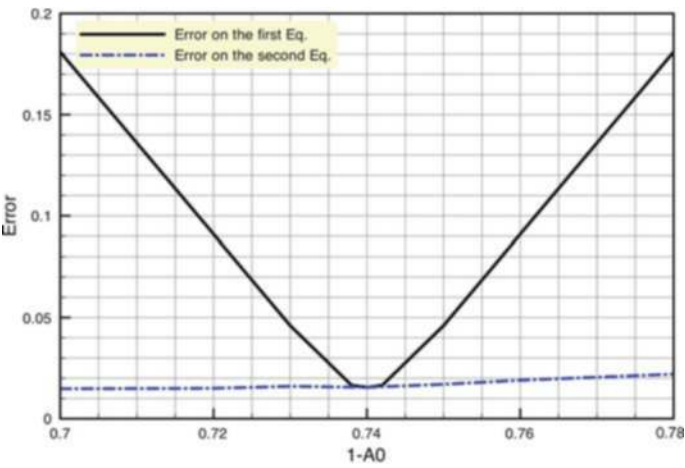


Fig. 11.5 Assessment of the optimum value of A_0

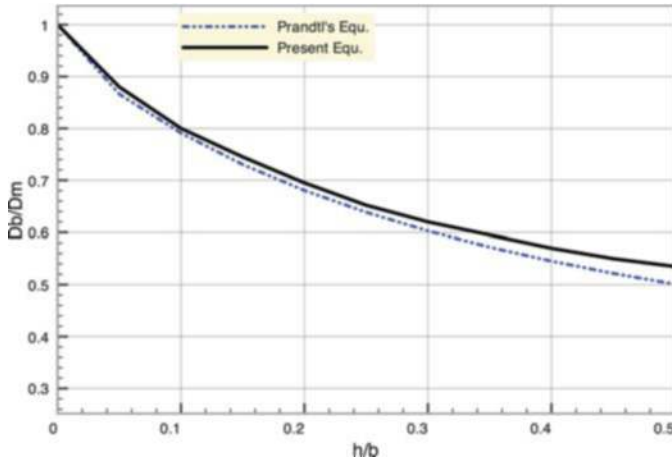


Fig. 11.6 Comparison between the present and the Prandtl's solution

where D_{bws} and D_{bm} are the induced drag on the best wing system and the best monoplane, respectively.

In Prandtl's results, the asymptotic behavior of the best wing system is more optimistic. The present results are limited by the hypothesis of a cubic distribution of circulation in the vertical wings.

Acknowledgments The authors gratefully acknowledge the contributions of Alberto Longhi in the formulation of the problem, Massimo Pappalardo of the Department of Applied Mathematics of Pisa, Piero Villaggio of the Department of Structural Engineering of Pisa, Luigi Polito and Emanuele Rizzo of the Department of Aerospace Engineering of Pisa.

References

1. Prandtl L.: Induced Drag of Multiplanes, NACA TN 182 (1924).
2. Pistolesi E.: Aerodinamica, ETS Editrice, Pisa.
3. Munk M.: Isoperimetrische Aufgaben aus der Theorie des Fluges, Inaugural Dissertation 1919, Gottinga (1919).
4. Munk M.: The minimum induced drag in airfoils, NACA 121(1924).
5. Montanari G.: Problemi di minimo della resistenza indotta in sistemi portanti, Graduating Thesis in Mathematics, Università di Pisa, 1998.

Appendix 1

Introducing the non-dimensional quantities

$$\eta = \frac{y}{b/2}, \quad \mu = \frac{z}{h}, \quad k = \frac{h}{b/2}$$

we have

$$\cos^2 \alpha = \left(\frac{y_1^1 - y_2}{\sqrt{(y_1^1 - y_2)^2 + h^2}} \right)^2 = \frac{b^2/4(\eta_1^1 - \eta_2)^2}{b^2/4(\eta_1^1 - \eta_2)^2 + h^2} = \frac{(\eta_1^1 - \eta_2)^2}{(\eta_1^1 - \eta_2)^2 + k^2}$$

$$\sin \beta = \frac{b/2 - y_1^1}{\sqrt{(b/2 - y_1^1)^2 + (h - z_3)^2}}$$

$$\cos \beta = \frac{h - z_3}{\sqrt{(b/2 - y_1^1)^2 + (h - z_3)^2}}$$

$$\cos \beta \sin \beta = \frac{(b/2 - y_1^1)(h - z_3)}{(b/2 - y_1^1)^2 + (h - z_3)^2} = \frac{hb/2(1 - \eta_1^1)(1 - \mu_3)}{b^2/4(1 - \eta_1^1)^2 + h^2(1 - \mu_3)^2} = k \frac{(1 - \eta_1^1)(1 - \mu_3)}{(1 - \eta_1^1)^2 + k^2(1 - \mu_3)^2}$$

$$\sin \delta = \frac{y_1^1 + b/2}{\sqrt{(y_1^1 + b/2)^2 + (h - z_4)^2}}$$

$$\cos \delta = \frac{h - z_4}{\sqrt{(y_1^1 + b/2)^2 + (h - z_4)^2}}$$

$$\cos \delta \sin \delta = \frac{(y_1^1 + b/2)(h - z_4)}{(y_1^1 + b/2)^2 + (h - z_4)^2} = \frac{b/2h(1 + \eta_1^1)(1 - \mu_4)}{b^2/4(1 + \eta_1^1)^2 + h^2(1 - \mu_4)^2} = k \frac{(1 + \eta_1^1)(1 - \mu_4)}{(1 + \eta_1^1)^2 + k^2(1 - \mu_4)^2}$$

and (11.9) becomes

$$\begin{aligned} & \frac{1}{2b\pi} \left[\int_{-1}^1 \frac{d\Gamma_1}{d\eta_1} \frac{1}{\eta_1 - \eta_1^1} d\eta_1 + \int_{-1}^1 \frac{d\Gamma_2}{d\eta_2} \frac{(\eta_2 - \eta_1^1)}{[(\eta_2 - \eta_1^1)^2 + k^2]} d\eta_2 \right. \\ & \quad + \int_0^1 \frac{d\Gamma_3}{d\mu_3} \frac{(1 - \eta_1^1)}{[(1 - \eta_1^1)^2 + k^2(1 - \mu_3)^2]} d\mu_3 + \\ & \quad \left. - \int_0^1 \frac{d\Gamma_4}{d\mu_4} \frac{(1 + \eta_1^1)}{[(1 + \eta_1^1)^2 + k^2(1 - \mu_4)^2]} d\mu_4 \right] = w_0 \end{aligned}$$

In the same way, it is easy to prove that the second equation becomes

$$\begin{aligned} & \frac{1}{2b\pi} \left[\int_{-1}^1 \frac{d\Gamma_1}{d\eta_1} \frac{(\eta_1 - \eta_2^1)}{[(\eta_1 - \eta_2^1)^2 + k^2]} d\eta_1 + \int_{-1}^1 \frac{d\Gamma_2}{d\eta_2} \frac{1}{\eta_2 - \eta_2^1} d\eta_2 \right. \\ & \quad \left. + \int_0^1 \frac{d\Gamma_3}{d\mu_3} \frac{(1 - \eta_2^1)}{[(1 - \eta_2^1)^2 + k^2\mu_3^2]} d\mu_3 + - \int_0^1 \frac{d\Gamma_4}{d\mu_4} \frac{(1 + \eta_2^1)}{[(1 + \eta_2^1)^2 + k^2\mu_4^2]} d\mu_4 \right] = w_0 \end{aligned}$$

In the third equation we have

$$\cos \beta = \frac{h - z_3^1}{\sqrt{(h - z_3^1)^2 + (b/2 - y_1)^2}}$$

$$\cos^2 \beta = \frac{h^2 (1 - \mu_3^1)^2}{h^2 (1 - \mu_3^1)^2 + b^2/4(1 - \eta_1)^2} = k^2 \frac{(1 - \mu_3^1)^2}{(1 - \eta_1)^2 + k^2 (1 - \mu_3^1)^2}$$

$$\sin \varepsilon = \frac{z_3^1}{\sqrt{(z_3^1)^2 + (b/2 - y_2)^2}}$$

$$\sin^2 \varepsilon = \frac{h^2 (\mu_3^1)^2}{h^2 (\mu_3^1)^2 + b^2/4(1 - \eta_2)^2} = k^2 \frac{(\mu_3^1)^2}{(1 - \eta_2)^2 + k^2 (\mu_3^1)^2}$$

$$\cos v = \frac{z_3^1 - z_4}{\sqrt{(z_3^1 - z_4)^2 + b^2}}$$

$$\cos^2 v = \frac{h^2 (\mu_3^1 - \mu_4)^2}{h^2 (\mu_3^1 - \mu_4)^2 + b^2} = k^2 \frac{(\mu_3^1 - \mu_4)^2}{4 + k^2 (\mu_3^1 - \mu_4)^2}$$

and we obtain

$$\begin{aligned} \frac{1}{4\pi} \left[k^2 \int_{-1}^1 \frac{d\Gamma_1}{d\eta_1} \frac{(\mu_3^1 - 1)}{[(1 - \eta_1)^2 + k^2 (\mu_3^1 - 1)^2]} d\eta_1 + k^2 \int_{-1}^1 \frac{d\Gamma_2}{d\eta_2} \frac{\mu_3^1}{[(1 - \eta_2)^2 + k^2 (\mu_3^1)^2]} d\eta_2 + \right. \\ \left. + k^2 \int_0^1 \frac{d\Gamma_4}{d\mu_4} \frac{(\mu_3^1 - \mu_4)}{[4 + k^2 (\mu_3^1 - \mu_4)^2]} d\mu_4 + \int_0^1 \frac{d\Gamma_3}{d\mu_3} \frac{1}{\mu_3^1 - \mu_3} d\mu_3 \right] = 0 \end{aligned}$$

Using a similar procedure for the fourth equation, we obtain

$$\begin{aligned} \frac{1}{4\pi} \left[k^2 \int_{-1}^1 \frac{d\Gamma_1}{d\eta_1} \frac{(1 - \mu_3^1)}{[(1 + \eta_1)^2 + k^2 (1 - \mu_3^1)^2]} d\eta_1 - k^2 \int_{-1}^1 \frac{d\Gamma_2}{d\eta_2} \frac{\mu_3^1}{[(1 + \eta_2)^2 + k^2 (\mu_3^1)^2]} d\eta_2 + \right. \\ \left. + k^2 \int_0^1 \frac{d\Gamma_3}{d\mu_3} \frac{(\mu_3 - \mu_3^1)}{[4 + k^2 (\mu_3 - \mu_3^1)^2]} d\mu_3 + \int_0^1 \frac{d\Gamma_4}{d\mu_4} \frac{1}{\mu_4 - \mu_4^1} d\mu_4 \right] = 0. \end{aligned}$$

Appendix 2

Putting, $\tan \frac{\varphi_2}{2} = t$, we have $\cos \varphi_2 = \frac{1-t^2}{1+t^2}$, $\sin \varphi_2 = \frac{2t}{1+t^2}$, $d\varphi_2 = \frac{2}{1+t^2} dt$;
besides, $\varphi_2 = 0 \rightarrow t = 0$ and $\varphi_2 = \pi \rightarrow t = \infty$.

Now, putting for simplicity sake, $\cos \varphi_1^1 = c$, it results

$$\begin{aligned} -\Gamma_{02} \int_0^\pi \frac{\cos \varphi_2 (\cos \varphi_2 - \cos \varphi_1^1)}{[(\cos \varphi_2 - \cos \varphi_1^1)^2 + k^2]} d\varphi_2 = -\Gamma_{02} \int_0^\infty \frac{\frac{1-t^2}{1+t^2} \left(\frac{1-t^2}{1+t^2} - c \right) \frac{2}{1+t^2}}{\left[\left(\frac{1-t^2}{1+t^2} - c \right)^2 + k^2 \right]} dt = \\ = -2\Gamma_{02} \int_0^\infty \frac{(c+1)t^4 - 2t^2 + 1 - c}{(1+t^2) [(c^2 + 1 + 2c + k^2)t^4 + 2(c^2 + k^2 - 1)t^2 + (c^2 + 1 - 2c + k^2)]} dt; \end{aligned}$$

We indicate as $A = c + 1$, $B = 1 - c$, $C = c^2 + k^2 - 1$

and

$$\begin{aligned}
 -\Gamma_{02} \int_0^\pi \frac{\cos \varphi_2 (\cos \varphi_2 - \cos \varphi_1^1)}{[(\cos \varphi_2 - \cos \varphi_1^1)^2 + k^2]} d\varphi_2 &= -2\Gamma_{02} \int_0^\infty \frac{At^4 - 2t^2 + B}{(1+t^2)[(A^2+k^2)t^4 + 2Ct^2 + (B^2+k^2)]} dt = \\
 &= -2\Gamma_{02} \int_0^\infty \frac{1}{1+t^2} \frac{A}{A^2+k^2} \frac{(A^2+k^2)t^4 - 2\frac{A^2+k^2}{A}t^2 + B\frac{A^2+k^2}{A}}{[(A^2+k^2)t^4 + 2Ct^2 + (B^2+k^2)]} dt = \\
 &= -2\Gamma_{02} \int_0^\infty \frac{1}{1+t^2} \frac{A}{A^2+k^2} \frac{[(A^2+k^2)t^4 + 2Ct^2 + (B^2+k^2)] - 2\left(C + \frac{A^2+k^2}{A}\right)t^2 + \left[\frac{B(A^2+k^2)}{A} - B^2 - k^2\right]}{[(A^2+k^2)t^4 + 2Ct^2 + (B^2+k^2)]} dt = \\
 &= -2\Gamma_{02} \frac{A}{A^2+k^2} \left[\arctan t \Big|_0^\infty - \int_0^\infty \frac{2(AC + A^2 + k^2)t^2 - (BA^2 + Bk^2 - B^2A - k^2A)}{A(t^2+1)[(A^2+k^2)t^4 + 2Ct^2 + (B^2+k^2)]} dt \right] = \\
 &= -2\Gamma_{02} \frac{1}{A^2+k^2} \left[\frac{\pi}{2}A - 2 \int_0^\infty \frac{(c^3 + 2c^2 + k^2c + 2k^2 + c)t^2 + (c^3 + k^2c - c)}{(t^2+1)[(A^2+k^2)t^4 + 2Ct^2 + (B^2+k^2)]} dt \right].
 \end{aligned}$$

The integration is obtained by decomposition, so we put

$$\begin{aligned}
 (t^2 + 1) [(A^2 + k^2)t^4 + 2(c^2 + k^2 - 1)t^2 + (B^2 + k^2)] &= \\
 = (t^2 + 1) [(c^2 + 1 + 2c + k^2)t^4 + 2(c^2 + k^2 - 1)t^2 + (c^2 + 1 - 2c + k^2)] &= 0.
 \end{aligned}$$

The roots of this equation are

$$t_1 = \frac{1}{\sqrt{A^2+k^2}} \left(-k\sqrt{\frac{2}{\sqrt{V}-a}} + i\sqrt{\frac{\sqrt{V}-a}{2}} \right)$$

$$t_2 = \frac{1}{\sqrt{A^2+k^2}} \left(k\sqrt{\frac{2}{\sqrt{V}-a}} - i\sqrt{\frac{\sqrt{V}-a}{2}} \right)$$

$$t_3 = \frac{1}{\sqrt{A^2+k^2}} \left(k\sqrt{\frac{2}{\sqrt{V}-a}} + i\sqrt{\frac{\sqrt{V}-a}{2}} \right)$$

$$t_4 = \frac{1}{\sqrt{A^2+k^2}} \left(-k\sqrt{\frac{2}{\sqrt{V}-a}} - i\sqrt{\frac{\sqrt{V}-a}{2}} \right)$$

$$t_5 = i$$

$$t_6 = -i$$

and the integral becomes

$$\begin{aligned}
 &\int_0^\infty \frac{(c^3 + 2c^2 + k^2c + 2k^2 + c)t^2 + (c^3 + k^2c - c)}{(t^2 + 1)[(c^2 + 1 + 2c + k^2)t^4 + 2(c^2 + k^2 - 1)t^2 + (c^2 + 1 - 2c + k^2)]} dt = \\
 &= \frac{1}{A^2+k^2} \left[\int_0^\infty \frac{Ft + G}{\left(t + k\sqrt{\frac{2}{(A^2+k^2)(\sqrt{V}-a)}} \right)^2 + \frac{\sqrt{V}-a}{2(A^2+k^2)}} dt \right. \\
 &\quad \left. + \int_0^\infty \frac{Ht + I}{\left(t - k\sqrt{\frac{2}{(A^2+k^2)(\sqrt{V}-a)}} \right)^2 + \frac{\sqrt{V}-a}{2(A^2+k^2)}} dt + \int_0^\infty \frac{St + R}{t^2 + 1} dt \right]
 \end{aligned}$$

where

$$S = 0, F = -H = \frac{A^2+k^2}{8k} \sqrt{\frac{(A^2+k^2)(\sqrt{V}-a)}{2}} \left(-c - c^2 - k^2 + \frac{-c - c^2 + c^3 + c^4 + k^2 + k^2 c + 2k^2 c^2 + k^4}{\sqrt{V}} \right),$$

$$R = -\frac{k^2+c^2+c}{2} (A^2+k^2), G = I = \frac{A^2+k^2}{4\sqrt{V}} (c^4 + c^3 - c^2 + 2k^2 c^2 - c + k^2 c + k^4 + k^2)$$

The three integrals can be solved explicitly.

As for the first one, we put $y = t + k \sqrt{\frac{2}{(A^2+k^2)(\sqrt{V}-a)}} \rightarrow$

so that

$$dy = dt,$$

$$t = 0 \rightarrow y = k \sqrt{\frac{2}{(A^2+k^2)(\sqrt{V}-a)}},$$

$$t = \infty \rightarrow y = \infty$$

and, after this substitution, we have

$$\begin{aligned} \int_0^\infty \frac{Ft+G}{\left(t+k\sqrt{\frac{2}{(A^2+k^2)(\sqrt{V}-a)}}\right)^2 + \frac{\sqrt{V}-a}{2(A^2+k^2)}} dt &= \\ = \int_{k\sqrt{\frac{2}{(A^2+k^2)(\sqrt{V}-a)}}}^\infty \frac{F\left(y-k\sqrt{\frac{2}{(A^2+k^2)(\sqrt{V}-a)}}\right)+G}{y^2 + \frac{\sqrt{V}-a}{2(A^2+k^2)}} dy &= \int_{k\sqrt{\frac{2}{(A^2+k^2)(\sqrt{V}-a)}}}^\infty \frac{F\left(y-k\sqrt{\frac{2}{(A^2+k^2)(\sqrt{V}-a)}}\right)+G}{\frac{\sqrt{V}-a}{2(A^2+k^2)}\left(1+2y^2\frac{A^2+k^2}{\sqrt{V}-a}\right)} dy. \end{aligned}$$

Now, with the position $z = y \sqrt{\frac{2(A^2+k^2)}{\sqrt{V}-a}}$

$$dz = \sqrt{\frac{2(A^2+k^2)}{\sqrt{V}-a}} dy, \text{ and because } y = k \sqrt{\frac{2}{(A^2+k^2)(\sqrt{V}-a)}},$$

$$z = \frac{2k}{\sqrt{V}-a},$$

$$y = \infty \rightarrow z = \infty;$$

and the previous integral becomes

$$\begin{aligned} &= \int_{\frac{2k}{\sqrt{V}-a}}^\infty \frac{G+F\left(\sqrt{\frac{\sqrt{V}-a}{2(A^2+k^2)}} z - k \sqrt{\frac{2}{(A^2+k^2)(\sqrt{V}-a)}}\right)}{1+z^2} \sqrt{\frac{2(A^2+k^2)}{\sqrt{V}-a}} dz = \\ &= \left(\sqrt{\frac{2(A^2+k^2)}{\sqrt{V}-a}} G - \frac{2k}{\sqrt{V}-a} F \right) [\arctan z]_{\frac{2k}{\sqrt{V}-a}}^\infty + \frac{F}{2} \int_{\frac{2k}{\sqrt{V}-a}}^\infty \frac{2z}{1+z^2} dz = \\ &= \left(\sqrt{\frac{2(A^2+k^2)}{\sqrt{V}-a}} G - \frac{2k}{\sqrt{V}-a} F \right) \left(\frac{\pi}{2} - \arctan \left(\frac{2k}{\sqrt{V}-a} \right) \right) + \frac{F}{2} [\ln(z^2+1)]_{\frac{2k}{\sqrt{V}-a}}^\infty \end{aligned}$$

As for the second integration, we put

$$y = t - k \sqrt{\frac{2}{(A^2+k^2)(\sqrt{V}-a)}} \text{ so that}$$

$$dy = dt;$$

$$t = 0 \rightarrow y = -k \sqrt{\frac{2}{(A^2+k^2)(\sqrt{V}-a)}},$$

$$t = \infty \rightarrow y = \infty$$

and we obtain

$$\begin{aligned} & \int_0^\infty \frac{Ht+I}{\left(t-k\sqrt{\frac{2}{(A^2+k^2)(\sqrt{V}-a)}}\right)^2 + \frac{\sqrt{V}-a}{2(A^2+k^2)}} dt = \\ & = \int_{-k}^\infty \sqrt{\frac{2}{(A^2+k^2)(\sqrt{V}-a)}} \frac{G-F\left(y+k\sqrt{\frac{2}{(A^2+k^2)(\sqrt{V}-a)}}\right)}{\frac{\sqrt{V}-a}{2(A^2+k^2)}\left(y^2\frac{2(A^2+k^2)}{\sqrt{V}-a}+1\right)} dy. \end{aligned}$$

With the position $z = y\sqrt{\frac{2(A^2+k^2)}{\sqrt{V}-a}}$

$$\begin{aligned} dz &= \sqrt{\frac{2(A^2+k^2)}{\sqrt{V}-a}} dy; \\ y &= -k\sqrt{\frac{2}{(A^2+k^2)(\sqrt{V}-a)}} \\ z &= -\frac{2k}{\sqrt{V}-a}, \\ y &= \infty \rightarrow z = \infty, \end{aligned}$$

the second integral becomes

$$\begin{aligned} & = \int_{-\frac{2k}{\sqrt{V}-a}}^\infty \frac{G-F\left(\sqrt{\frac{\sqrt{V}-a}{2(A^2+k^2)}}z+k\sqrt{\frac{2}{(A^2+k^2)(\sqrt{V}-a)}}\right)}{1+z^2} \sqrt{\frac{2(A^2+k^2)}{\sqrt{V}-a}} dz = \\ & = \left(\sqrt{\frac{2(A^2+k^2)}{\sqrt{V}-a}}G - \frac{2k}{\sqrt{V}-a}F\right) \left(\frac{\pi}{2} + \arctan\left(\frac{2k}{\sqrt{V}-a}\right)\right) - \frac{F}{2} [\ln(z^2+1)]_{-\frac{2k}{\sqrt{V}-a}}^\infty \int_0^\infty \frac{R}{t^2+1} dt = \\ & = R[\arctan t]_0^\infty = \frac{\pi}{2}R \end{aligned}$$

The solution of the third integral is simple and the result is

$$\begin{aligned} & \int_0^\infty \frac{(c^3+2c^2+k^2c+2k^2+c)t^2+(c^3+k^2c-c)}{(t^2+1)[(c^2+1+2c+k^2)t^4+2(c^2+k^2-1)t^2+(c^2+1-2c+k^2)]} dt = \\ & = \frac{1}{A^2+k^2} \left[\left(\sqrt{\frac{2(A^2+k^2)}{\sqrt{V}-a}}G - \frac{2k}{\sqrt{V}-a}F \right) \pi + \frac{\pi}{2}R \right] = \\ & = \pi \left(\sqrt{\frac{2}{(A^2+k^2)(\sqrt{V}-a)}}G - \frac{2k}{(A^2+k^2)(\sqrt{V}-a)}F + \frac{1}{2(A^2+k^2)}R \right). \end{aligned}$$

Finally

$$\begin{aligned} & \pi \left[-\Gamma_{01} - \Gamma_{02} \frac{A}{(A^2+k^2)} + \Gamma_{02} \frac{4}{(A^2+k^2)} \left(\sqrt{\frac{2}{(A^2+k^2)(\sqrt{V}-a)}}G \right. \right. \\ & \quad \left. \left. - \frac{2k}{(A^2+k^2)(\sqrt{V}-a)}F + \frac{R}{2(A^2+k^2)} \right) \right] \end{aligned}$$

and, because $\Gamma_{01} = \Gamma_{02}$, the induced velocity is given by (11.25).

Appendix 3

Putting $\tan\left(\frac{\varphi_1}{2}\right) = t = \tan\left(\frac{\varphi_2}{2}\right)$, we have

$$\int_0^\pi \frac{\cos \varphi_2}{\left[(1 - \cos \varphi_2)^2 + k^2 (\mu_3^1)^2\right]} d\varphi_2 = \int_0^\infty \frac{\frac{1-t^2}{1+t^2} \frac{2}{1+t^2}}{\left[\left(1 - \frac{1-t^2}{1+t^2}\right)^2 + k^2 (\mu_3^1)^2\right]} dt = \int_0^\infty \frac{2(1-t^2)}{4t^4 + k^2 (\mu_3^1)^2 (1-t^2)^2} dt.$$

The integration is obtained by decomposition; by putting $k^1 = k\mu_3^1$, we have

$$(4 + k_1^2)t^4 + 2k_1^2t^2 + k_1^2 = 0 \quad \text{and} \quad t^2 = \frac{-k_1^2 \pm 2ik_1}{4 + k_1^2}$$

The roots are expressed in the form: $t = a \pm ib$ with a and b real numbers; from the identity $t^2 = a^2 - b^2 + 2iab$, we obtain

$$a^2 - b^2 = -\frac{k_1^2}{4 + k_1^2}$$

$$2ab = \pm 2\frac{k_1}{4 + k_1^2}$$

and the solutions are

$$a = \pm \sqrt{\frac{-k_1^2 + k_1 \sqrt{4 + k_1^2}}{2(4 + k_1^2)}}$$

$$b = \pm \frac{\sqrt{2k_1}}{\sqrt{(4 + k_1^2)(-k_1^2 + \sqrt{4 + k_1^2})}}.$$

So, the roots of the equation $(4 + k_1^2)t^4 + 2k_1^2t^2 + k_1^2 = 0$ are

$$t_1 = \frac{1}{\sqrt{4 + k_1^2}} \left(\sqrt{\frac{k_1}{2} (\sqrt{4 + k_1^2} - k_1)} - \frac{\sqrt{2k_1}}{\sqrt{4 + k_1^2 - k_1}} i \right)$$

$$t_2 = \frac{1}{\sqrt{4 + k_1^2}} \left(-\sqrt{\frac{k_1}{2} (\sqrt{4 + k_1^2} - k_1)} + \frac{\sqrt{2k_1}}{\sqrt{4 + k_1^2 - k_1}} i \right)$$

$$t_3 = \frac{1}{\sqrt{4 + k_1^2}} \left(\sqrt{\frac{k_1}{2} (\sqrt{4 + k_1^2} - k_1)} + \frac{\sqrt{2k_1}}{\sqrt{4 + k_1^2 - k_1}} i \right) = \bar{t}_1$$

$$t_4 = \frac{1}{\sqrt{4 + k_1^2}} \left(-\sqrt{\frac{k_1}{2} (\sqrt{4 + k_1^2} - k_1)} - \frac{\sqrt{2k_1}}{\sqrt{4 + k_1^2 - k_1}} i \right) = \bar{t}_2$$

and the decomposition becomes

$$\begin{aligned} (4 + k_1^2)t^4 + 2k_1^2t^2 + k_1^2 &= \\ &= (4 + k_1^2) \left[\left(t - \sqrt{\frac{k_1(\sqrt{4 + k_1^2} - k_1)}{2(4 + k_1^2)}} \right)^2 + \frac{2k_1}{(4 + k_1^2)(\sqrt{4 + k_1^2} - k_1)} \right] \\ &\quad \left[\left(t + \sqrt{\frac{k_1(\sqrt{4 + k_1^2} - k_1)}{2(4 + k_1^2)}} \right)^2 + \frac{2k_1}{(4 + k_1^2)(\sqrt{4 + k_1^2} - k_1)} \right]. \end{aligned}$$

It follows that

$$\begin{aligned}
 & \int_0^\infty \frac{2(1-t^2)}{4t^2+k_1^2(1+t^2)^2} dt = \int_0^\infty \frac{2(1-t^2)}{(4+t^2)t^4+2k_1^2t^2+k_1^2} dt = \\
 & = \frac{1}{\sqrt{4+k_1^2}} \left[\int_0^\infty \frac{F_1t+G_1}{\left[\left(t - \sqrt{\frac{k_1(\sqrt{4+k_1^2}-k_1)}{2(4+k_1^2)}} \right)^2 + \frac{2k_1}{(4+k_1^2)(\sqrt{4+k_1^2}-k_1)} \right]} dt + \right. \\
 & \quad \left. + \int_0^\infty \frac{H_1t+I_1}{\left[\left(t + \sqrt{\frac{k_1(\sqrt{4+k_1^2}-k_1)}{2(4+k_1^2)}} \right)^2 + \frac{2k_1}{(4+k_1^2)(\sqrt{4+k_1^2}-k_1)} \right]} dt \right]
 \end{aligned}$$

where

$$G_1 = I_1 = \frac{\sqrt{4+k_1^2}}{k_1}$$

$$F_1 = -\frac{\sqrt{4+k_1^2}(k_1+\sqrt{4+k_1^2})}{k_1\sqrt{2k_1(\sqrt{4+k_1^2}-k_1)}}$$

$$H_1 = \frac{\sqrt{4+k_1^2}(k_1+\sqrt{4+k_1^2})}{k_1\sqrt{2k_1(\sqrt{4+k_1^2}-k_1)}}$$

Now, we solve separately the two integrals.

As for the first, we put $y = t - \sqrt{\frac{k_1(\sqrt{4+k_1^2}-k_1)}{2(4+k_1^2)}}$, so that

$$dt = dy,$$

$$t = 0 \rightarrow y = -\sqrt{\frac{k_1(\sqrt{4+k_1^2}-k_1)}{2(4+k_1^2)}},$$

$$t = \infty \rightarrow y = \infty,$$

and the result is

$$\int_0^\infty \frac{F_1t+G_1}{\left[\left(t - \sqrt{\frac{k_1(\sqrt{4+k_1^2}-k_1)}{2(4+k_1^2)}} \right)^2 + \frac{2k_1}{(4+k_1^2)(\sqrt{4+k_1^2}-k_1)} \right]} dt =$$

$$= \int_{-\sqrt{\frac{k_1(\sqrt{4+k_1^2}-k_1)}{2(4+k_1^2)}}}^{\infty} \frac{-F_1\left(y + \sqrt{\frac{k_1(\sqrt{4+k_1^2}-k_1)}{2(4+k_1^2)}}\right) + G_1}{\left(y^2 \frac{(4+k_1^2)(\sqrt{4+k_1^2}-k_1)}{2k_1}\right) \frac{2k_1}{(4+k_1^2)(\sqrt{4+k_1^2}-k_1)}} dy;$$

now, with the position $z = y \sqrt{\frac{k_1(\sqrt{4+k_1^2}-k_1)}{2(4+k_1^2)}}$,

$$dz = \sqrt{\frac{(4+k_1^2)(\sqrt{4+k_1^2}-k_1)}{2k_1}} dy,$$

$$y = -\sqrt{\frac{k_1(\sqrt{4+k_1^2}-k_1)}{2(4+k_1^2)}} \rightarrow$$

$$z = -\frac{\sqrt{4+k_1^2}-k_1}{2},$$

$$y = \infty \rightarrow z = \infty$$

we obtain

$$\begin{aligned} & \int_{-\sqrt{\frac{(4+k_1^2)(\sqrt{4+k_1^2}-k_1)}{2k_1}}}^{\infty} \frac{-F_1\left(\sqrt{\frac{2k_1}{(4+k_1^2)(\sqrt{4+k_1^2}-k_1)}} + \sqrt{\frac{(4+k_1^2)(\sqrt{4+k_1^2}-k_1)}{2k_1}}\right) + G_1}{z^2 + 1} \\ & \times \sqrt{\frac{(4+k_1^2)(\sqrt{4+k_1^2}-k_1)}{2k_1}} dz = \\ & = \sqrt{\frac{(4+k_1^2)(\sqrt{4+k_1^2}-k_1)}{2k_1}} \left\{ G_1 [\arctan z]_{-\frac{\sqrt{4+k_1^2}-k_1}{2}}^{\infty} \right. \\ & - \sqrt{\frac{(4+k_1^2)(\sqrt{4+k_1^2}-k_1)}{2(4+k_1^2)}} F_1 [\arctan z] + \\ & \left. - F_1 \sqrt{\frac{k_1}{2(4+k_1^2)(\sqrt{4+k_1^2}-k_1)}} [\ln(z^2 + 1)]_{-\frac{\sqrt{4+k_1^2}-k_1}{2}}^{\infty} \right\} \end{aligned}$$

As far as the second integral is concerned, we apply a procedure very similar to the previous one and we obtain

$$\int_0^{\infty} \frac{H_1 t + I_1}{\left[\left(t + \sqrt{\frac{k_1(\sqrt{4+k_1^2}-k_1)}{2(4+k_1^2)}} \right)^2 + \frac{2k_1}{(4+k_1^2)(\sqrt{4+k_1^2}-k_1)} \right]} dt =$$

$$\begin{aligned}
&= \sqrt{\frac{(4+k_1^2)(\sqrt{4+k_1^2}-k_1)}{2k_1}} \left\{ I_1 [\arctg z]_{\frac{\sqrt{4+k_1^2}-k_1}{2}}^{\infty} \right. \\
&\quad + H_1 \sqrt{\frac{k_1}{2(4+k_1^2)(\sqrt{4+k_1^2}-k_1)}} [\ln(z^2+1)]_{\frac{\sqrt{4+k_1^2}-k_1}{2}}^{\infty} + \\
&\quad \left. - H_1 \sqrt{\frac{k_1(\sqrt{4+k_1^2}-k_1)}{2(4+k_1^2)}} [\arctg z]_{\frac{\sqrt{4+k_1^2}-k_1}{2}}^{\infty} \right\}
\end{aligned}$$

Finally, by collecting together the results, the second equation of the system becomes

$$\int_0^{\infty} \frac{2(1-t^2)}{4t^2+k_1^2(1+t^2)} dt = \pi \sqrt{\frac{(\sqrt{4+k_1^2}-k_1)}{2k_1(4+k_1^2)}} \left[I_1 - \sqrt{\frac{k_1(\sqrt{4+k_1^2}-k_1)}{2(4+k_1^2)}} H_1 \right]$$

where

$$F_1 = -H_1 = -\frac{\sqrt{4+k_1^2}(k_1+\sqrt{4+k_1^2})}{k_1\sqrt{2k_1(\sqrt{4+k_1^2}-k_1)}}; \quad G_1 = I_1 = \frac{\sqrt{4+k_1^2}}{k_1}.$$

Appendix 4

The first equation of problem B becomes

$$\int_0^1 \frac{d\Gamma_3}{d\mu_3} \frac{(1-\eta_1^1)}{[(1-\eta_1^1)^2+k^2(1-\mu_3)^2]} d\mu_3 + \int_0^1 \frac{d\Gamma_3}{d\mu_3} \frac{(1+\eta_1^1)}{[(1+\eta_1^1)^2+k^2(1-\mu_3)^2]} d\mu_3$$

Putting, for simplicity sake $(1-\eta_1^1) = a$ and having assumed $\Gamma_3(\mu_3)$ as a cubic, the first integral becomes

$$\begin{aligned}
&A_0 \int_0^1 a \frac{-12(e+2)\mu_3^2+12(e+2)\mu_3-2(e+3)}{[a^2+k^2(1-\mu_3)^2]} d\mu_3 = \\
&= A_0 \int_0^1 a \frac{-2(e+3)}{a^2[1+\frac{k^2}{a^2}(1-\mu_3)^2]} d\mu_3 + \frac{12(e+2)A_0a}{2k^2} \int_0^1 \frac{2k^2\mu_3-2k^2+2k^2}{[a^2+k^2(1-\mu_3)^2]} d\mu_3 + \\
&- \frac{12A_0(e+2)a}{k^2} \int_0^1 \frac{k^2\mu_3^2+a^2+k^2-2k^2\mu_3}{[a^2+k^2-2k^2\mu_3+k^2\mu_3^2]} d\mu_3 + \frac{12A_0(e+2)a}{k^2} \int_0^1 \frac{a^2+k^2-2k^2\mu_3}{[a^2+k^2-2k^2\mu_3+k^2\mu_3^2]} d\mu_3 = \\
&(\text{putting } \frac{k}{a}(1-\mu_3) = y, \text{ so that } d\mu_3 = -\frac{a}{k}dy, \mu_3 = 0 \rightarrow y = \frac{k}{a}, \mu_3 = 1 \rightarrow y = 0), \\
&= -\frac{2A_0(e+3)}{k} \int_0^{\frac{k}{a}} \frac{1}{1+y^2} dy + \frac{6(e+2)A_0a}{k^2} \ln \left[a^2+k^2(1-\mu_3)^2 \right]_0^1 - \frac{12(e+2)A_0a}{k^2} +
\end{aligned}$$

$$\begin{aligned}
& + \frac{12A_0(e+2)a}{k^2} \int_0^1 \frac{a^2+k^2-2k^2\mu_3+k^2}{[a^2+k^2(1-\mu_3)^2]} d\mu_3 = \\
& = -\frac{2(e+3)}{k} A_0 \arctan\left(\frac{k}{a}\right) + \frac{6(e+2)A_0a}{k^2} \ln\left(\frac{a^2}{a^2+k^2}\right) - \frac{12(e+2)A_0a}{k^2} - \frac{12A_0(e+2)a}{k^2} \int_0^1 \frac{2k^2\mu_3-2k^2}{[a^2+k^2(1-\mu_3)^2]} d\mu_3 + \\
& + \frac{12A_0(e+2)a^2}{k^3} \int_0^{\frac{k}{a}} \frac{1}{[1+y^2]} dy = \\
& = A_0 \left\{ \arctan\left(\frac{k}{a}\right) \left[-\frac{2(e+3)}{k} + \frac{12(e+2)a^2}{k^3} \right] + \ln\left(\frac{a^2}{a^2+k^2}\right) \left[\frac{6(e+2)a}{k^2} - \frac{12(e+2)a}{k^2} \right] - \frac{12(e+2)a}{k^2} \right\}
\end{aligned}$$

The integral $\int_0^1 \frac{d\Gamma_3}{d\mu_3} \frac{(1+\eta_1^1)}{[(1+\eta_1^1)^2+k^2(1-\mu_3)^2]} d\mu_3$ is solved in the same way.

Hence, the contribution to the Problem B is

$$\begin{aligned}
A_0 \left\{ \arctan\left(\frac{k}{1-\eta_1^1}\right) \left[\frac{12(e+2)(1-\eta_1^1)^2}{k^3} - \frac{2(e+3)}{k} \right] + \arctan\left(\frac{k}{1+\eta_1^1}\right) \left[-\frac{2(e+3)}{k} + \right. \right. \\
\left. \left. + \frac{12(e+2)(1+\eta_1^1)}{k^3} \right] + \ln\left[\frac{(1-\eta_1^1)^2}{k^2+(1-\eta_1^1)^2}\right] \left(-\frac{6(e+2)(1-\eta_1^1)}{k^2} \right) + \right. \\
\left. - \frac{6(e+2)(1+\eta_1^1)}{k^2} \ln\left[\frac{(1+\eta_1^1)^2}{k^2+(1+\eta_1^1)^2}\right] - \frac{24(e+2)}{k^2} \right\}
\end{aligned}$$

The second contribution is

$$\begin{aligned}
& \frac{1}{4\pi} \left[-k^2 \int_0^1 \frac{d\Gamma_3}{d\mu_3} \frac{\mu_3^1-\mu_3}{[4+k^2(\mu_3^1-\mu_3)^2]} d\mu_3 + \int_0^1 \frac{d\Gamma_3}{d\mu_3} \frac{1}{\mu_3^1-\mu_3} d\mu_3 \right] = \\
& = \frac{A_0}{4\pi} \left[-k^2 \int_0^1 \frac{[-12(e+2)\mu_3^2+12(e+2)\mu_3-2(e+3)](\mu_3^1-\mu_3)}{[4+k^2(\mu_3^1-\mu_3)^2]} d\mu_3 + \int_0^1 \frac{-12(e+2)\mu_3^2+12(e+2)\mu_3-2(e+3)}{\mu_3^1-\mu_3} d\mu_3 \right].
\end{aligned}$$

The solution of the first integral is

$$\begin{aligned}
& -k^2 \int_0^1 \frac{[-12(e+2)\mu_3^2+12(e+2)\mu_3-2(e+3)](\mu_3^1-\mu_3)}{[4+k^2(\mu_3^1-\mu_3)^2]} d\mu_3 = 12(e+2) \int_0^1 \frac{(k^2\mu_3^2+4+k^2(\mu_3^1)^2-2k^2\mu_3^1\mu_3)(\mu_3^1-\mu_3)}{[4+k^2(\mu_3^1-\mu_3)^2]} d\mu_3 + \\
& -12(e+2) \int_0^1 \frac{(4+k^2(\mu_3^1)^2-2k^2\mu_3^1\mu_3)(\mu_3^1-\mu_3)}{[4+k^2(\mu_3^1-\mu_3)^2]} d\mu_3 + 12(e+2)k^2 \int_0^1 \frac{\mu_3(\mu_3^1-\mu_3)}{[4+k^2(\mu_3^1-\mu_3)^2]} d\mu_3 + \\
& -(e+3) \int_0^1 \frac{-2k^2(\mu_3^1-\mu_3)}{[4+k^2(\mu_3^1-\mu_3)^2]} d\mu_3 = \\
& = 12(e+2) \left(\mu_3^1 - \frac{1}{2} \right) - (e+3) \ln \left[\frac{4+k^2(\mu_3^1-1)^2}{4+k^2(\mu_3^1)^2} \right] + \frac{12(e+2)[4+k^2(\mu_3^1)^2]}{2k^2} \int_0^1 \frac{-2k^2(\mu_3^1-\mu_3)}{[4+k^2(\mu_3^1-\mu_3)^2]} d\mu_3 + \\
& + 12(e+2)k^2(1+\mu_3^1) \int_0^1 \frac{\mu_3(\mu_3^1-\mu_3)}{[4+k^2(\mu_3^1-\mu_3)^2]} d\mu_3 =
\end{aligned}$$

$$= 6(e+2)(2\mu_3^1 - 1) - (e+3) \ln \left[\frac{4+k^2(\mu_3^1-1)^2}{4+k^2(\mu_3^1)^2} \right] + \frac{6(e+2)}{k^2} (4+k^2(\mu_3^1)^2) \ln \left[\frac{4+k^2(\mu_3^1-1)^2}{4+k^2(\mu_3^1)^2} \right] +$$

$$12(e+2)k^2(1+\mu_3^1) \int_0^1 \frac{\mu_3(\mu_3^1-\mu_3)}{[4+k^2(\mu_3^1-\mu_3)^2]} d\mu_3;$$

the last term is easily integrated as follows:

$$\int_0^1 \frac{\mu_3(\mu_3^1-\mu_3)}{[4+k^2(\mu_3^1-\mu_3)^2]} d\mu_3 = - \int_0^1 \frac{\mu_3^2-\mu_3\mu_3^1}{[4+k^2(\mu_3^1-\mu_3)^2]} d\mu_3 = -\frac{1}{k^2} \int_0^1 \frac{k^2\mu_3^2-2k^2\mu_3\mu_3^1+4+k^2(\mu_3^1)^2}{[4+k^2(\mu_3^1-\mu_3)^2]} d\mu_3 +$$

$$\frac{1}{k^2} \int_0^1 \frac{4+k^2(\mu_3^1)^2-k^2\mu_3\mu_3^1}{[4+k^2(\mu_3^1-\mu_3)^2]} d\mu_3 =$$

$$= -\frac{1}{k^2} - \frac{\mu_3^1}{2k^2} \int_0^1 \frac{-2k^2(\mu_3^1-\mu_3)}{[4+k^2(\mu_3^1-\mu_3)^2]} d\mu_3 + \frac{4}{k^2} \int_0^1 \frac{1}{[4+k^2(\mu_3^1-\mu_3)^2]} d\mu_3 =$$

$$(\text{putting } \frac{k}{2}(\mu_3^1-\mu_3) = y) = -\frac{1}{k^2} - \frac{\mu_3^1}{2k^2} \ln \left[\frac{4+k^2(\mu_3^1-1)^2}{4+k^2(\mu_3^1)^2} \right] - \frac{2}{k^3} \int_{\frac{k}{2}\mu_3^1}^{\frac{k}{2}(\mu_3^1-1)} \frac{1}{1+y^2} dy =$$

$$= -\frac{1}{k^2} - \frac{\mu_3^1}{2k^2} \ln \left[\frac{4+k^2(\mu_3^1-1)^2}{4+k^2(\mu_3^1)^2} \right] - \frac{2}{k^3} [\arctan(\frac{k}{2}(\mu_3^1-1)) - \arctan(\frac{k}{2}\mu_3^1)]$$

As for the second integral, we have

$$\int_0^1 \frac{-12(e+2)\mu_3^2+12(e+2)\mu_3-2(e+3)}{\mu_3^1-\mu_3} d\mu_3 = -12(e+2) \int_0^1 \frac{\mu_3^2-(\mu_3^1)^2}{\mu_3^1-\mu_3} d\mu_3 - 12(e+2)(\mu_3^1) \int_0^1 \frac{1}{\mu_3^1-\mu_3} d\mu_3 +$$

$$+ 12(e+2) \int_0^1 \frac{\mu_3-\mu_3^1}{\mu_3^1-\mu_3} d\mu_3 + 12(e+2)\mu_3^1 \int_0^1 \frac{1}{\mu_3^1-\mu_3} d\mu_3 - 2(e+3) \int_0^1 \frac{1}{\mu_3^1-\mu_3} d\mu_3 =$$

$$= 12(e+2) \left[\frac{\mu_3^2}{2} + \mu_3\mu_3^1 \right]_0^1 + \ln \left(\frac{1-\mu_3^1}{\mu_3^1} \right) \left[12(e+2)(\mu_3^1)^2 - 12(e+2)\mu_3^1 + 2(e+3) \right] -$$

$$12(e+2) =$$

$$= 6(e+2)(2\mu_3^1+1) + \ln \left(\frac{1-\mu_3^1}{\mu_3^1} \right) \left[12(e+2)(\mu_3^1)^2 - 12(e+2)\mu_3^1 + 2(e+3) \right]$$

Finally, the contributions to the induced velocity in the Problem B become

$$\left\{ \begin{aligned}
 w_H^{(B)} &= A_0 \left[\arctan \left(\frac{k}{1-\eta_1^1} \right) \left(\frac{12(e+2)(1-\eta_1^1)^2}{k^3} - \frac{2(e+3)}{k} \right) + \arctan \left(\frac{k}{1+\eta_1^1} \right) \right. \\
 &\quad \left(\frac{-2(e+3)}{k} + \frac{12(e+2)(1+\eta_1^1)^2}{k^3} \right) - \frac{6(e+2)(1-\eta_1^1)}{k^2} \ln \left(\frac{(1-\eta_1^1)^2}{k^{2+(1-\eta_1^1)^2}} \right) + \\
 &\quad \left. - \frac{6(e+2)(\eta_1^1+1)}{k^2} \ln \left(\frac{(1+\eta_1^1)^2}{k^2+(1+\eta_1^1)^2} \right) - \frac{24(e+2)}{k^2} \right] \\
 w_V^{(B)} &= \frac{A_0}{4\pi} \left[24(e+2)\mu_3^1 + \ln \left(\frac{1-\mu_3^1}{\mu_3^1} \right)^2 \left(6(e+2)(\mu_3^1)^2 - 6(e+2)\mu_3^1 + (e+3) \right) + \right. \\
 &\quad + \ln \left(\frac{4+k^2(\mu_3^1-1)^2}{4+k^2(\mu_3^1)^2} \right) \left(\frac{6(e+2)(4+k^2(\mu_3^1)^2)}{k^2} - (e+3) \right) + \\
 &\quad + 12(e+2)k^2(1+\mu_3^1) \left(-\frac{1}{k^2} - \frac{\mu_3^1}{2k^2} \ln \left(\frac{4+k^2(\mu_3^1-1)^2}{4+k^2(\mu_3^1)^2} \right) - \right. \\
 &\quad \left. \frac{2}{k^3} \left(\arctan \left(\frac{k}{2}(\mu_3^1-1) \right) - \arctan \left(\frac{k}{2}\mu_3^1 \right) \right) \right) + \\
 &\quad + \frac{k^2}{4} \left[(1-\mu_3^1)\Gamma_{01} \sqrt{\frac{\sqrt{4+k^2(1-\mu_3^1)^2} - (1-\mu_3^1)}{2k(1-\mu_3^1)(4+k^2(1-\mu_3^1)^2)}} \right. \\
 &\quad \left(I_2 - H_2 \sqrt{\frac{k(1-\mu_3^1)(\sqrt{4+k^2(1-\mu_3^1)^2} - (1-\mu_3^1))}{2(4+k^2(1-\mu_3^1)^2)}} \right) + \\
 &\quad \left. - \mu_3^1\Gamma_{01} \sqrt{\frac{\sqrt{4+k^2\mu_3^{12}} - k\mu_3^1}{2(4+k^2\mu_3^{12})k\mu_3^{12}}} \left(I_1 - H_1 \sqrt{\frac{k\mu_3^1(\sqrt{4+k^2\mu_3^{12}} - k\mu_3^1)}{24+k^2\mu_3^{12}}} \right) \right]
 \end{aligned} \right\}$$

“This page left intentionally blank.”

Chapter 12

Numerical Simulation of the Dynamics of Boats by a Variational Inequality Approach

Luca Formaggia, Edie Miglio, Andrea Mola and Anna Scotti

Abstract In this chapter we present some recent numerical studies on fluid–structure interaction problems in the presence of free surface flow. We consider the dynamics of a rowing boat, simulated as a rigid body. We focus on an approach based on formulating the floating body problem as an inequality constraint on the water elevation. A splitting procedure is used to develop an efficient numerical scheme where the inequality constraint is imposed only on a wave-like equation representing an hydrostatic approximation of the hydrodynamic equations. Numerical tests demonstrate the effectiveness of the proposed procedure.

12.1 Introduction

The use of computational fluid dynamics (CFD) in boat design is traditionally based on potential flow theory, even if in the last years the use of Reynolds-averaged Navier–Stokes (RANS) codes has become increasingly more common. The role of CFD is of particular importance whenever performance optimisation is critical, such as in competition boats, where even a small advantage may be crucial.

Luca Formaggia

MOX, Mathematics Department, Politecnico di Milano, Milano, Italy,

e-mail: luca.formaggia@polimi.it

Edie Miglio

MOX, Mathematics Department, Politecnico di Milano, Milano, Italy,

e-mail: edie.miglio@polimi.it

Andrea Mola

MOX, Mathematics Department, Politecnico di Milano, Milano, Italy,

e-mail: andrea.mola@polimi.it

Anna Scotti

MOX, Mathematics Department, Politecnico di Milano, Milano, Italy,

e-mail: anna.scotti@mail.polimi.it

An overview on the numerical techniques for ship hydrodynamics may be found in [4, 5] and, more specifically, their relevance for high-performance sailing boat in [14].

In this field, most of the numerical investigations aim to assess the boat characteristics at a given fixed configuration. Furthermore, they usually compute a steady-state solution, even if sometimes this is reached through pseudotime stepping. Yet, simulating the full dynamics of a boat may be of great importance [1, 15]. We mention two cases: high-performance sailing boats and rowing sculls. In the former, the accurate simulation of the dynamics may allow for a better trimming of the boat [1, 16], better evaluating wave resistance [13] and in perspective the assessment of its performance during manoeuvring. For a competition rowing scull, accounting for the dynamics effects is even more important. Indeed, because of the periodic action at the oars and the movement of the oarsmen on the boat the motions of the scull are very complex and characterised by horizontal accelerations/decelerations, sinking and dipping. These secondary movements generate waves which dissipate part of rowers' energy, which could be better spent to move the boat forward.

In this chapter, we will give an account of some current research in this class of problems by focusing on a numerical model based on the solution of quasi-3D Navier–Stokes equations with free surface [11], where the presence of the boat is modelled through an inequality constraint. We show how the method is able to reproduce the general wave patterns of a moving scull.

12.2 A Variational Approach to the Floating Body Problem

We will consider the free-surface Navier–Stokes equations where part of the surface is subject to a constraint which is meant to represent the external surface of a boat. More precisely, we will consider for any $t \in (0, T)$, with $T > 0$, the domain

$$\Omega(t) = \{\mathbf{x} = (x, y, z) \in \mathbb{R}^3 : (x, y) \in \omega, z \in (-h, \eta(x, y, t))\}$$

sketched in Fig. 12.1 is occupied by a fluid, η being the description of the free surface of the fluid (measured with respect to the unperturbed water depth). The part of the boundary of Ω corresponding to the free surface is denoted by

$$\Gamma_s(t) = \{\mathbf{x} \in \mathbb{R}^3 : (x, y) \in \omega, z = \eta(x, y, t)\}.$$

Here, ω is an open-bounded connected subset of \mathbb{R}^2 and we are implicitly assuming that the free surface can be represented by a function of (x, y) , i.e. no wave breaking occurs during the motion. The bottom surface is

$$\Gamma_b = \{\mathbf{x} \in \mathbb{R}^3 : (x, y) \in \omega, z = -h\},$$

where h indicates the depth of the bottom surface (again measured with respect to the unperturbed water level) which, for the sake of simplicity, is assumed to be constant. The remaining portion of $\partial\Omega(t)$ is the far field

$$\Gamma_f(t) = \{\mathbf{x} \in \mathbb{R}^3 : (x, y) \in \partial\omega, z \in (-h, \eta(x, y, t))\}.$$

Clearly, $\partial\Omega(t) = \bar{\Gamma}_s(t) \cup \bar{\Gamma}_b \cup \bar{\Gamma}_f(t)$ at any time t .

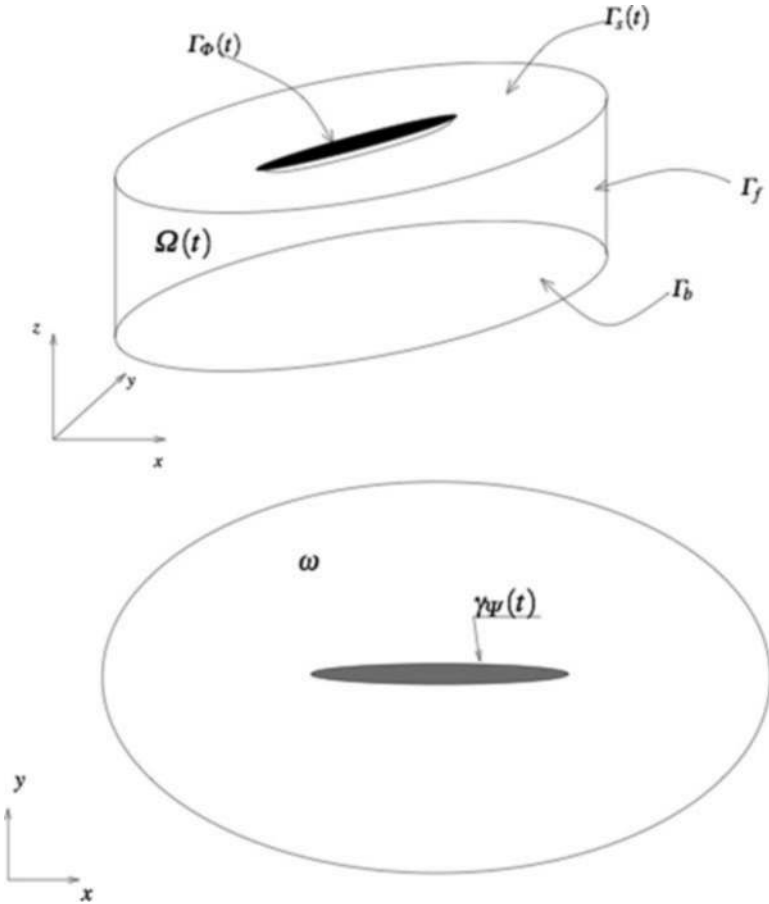


Fig. 12.1 The 3D computational domain (top) and its projection on the horizontal plane ω (bottom). The shadowed figure in the bottom picture represents the projection of the immersed part of the boat on the fluid surface, it is then part of ω

We now consider a continuous function

$$\Psi : \omega \times [0, T] \rightarrow \mathbb{R} \quad (12.1)$$

which is meant to represent the external surface of the hull of a boat, suitably extended to cover all ω (see Fig.12.2). We want to simulate the presence of a floating

boat by constraining the free-surface η to be at any time below Ψ . The evolution of Ψ will be normally given by the interaction of the fluid with the floating boat, yet in the following we assume that Ψ is given and that the (possibly empty) set

$$\gamma_\Psi(t) = \{(x, y) \in \omega : \Psi(x, y, t) = \eta(x, y, t)\} \quad (12.2)$$

is always strictly included in ω for all $t \in [0, T]$. This is clearly an “a priori” assumption, since η is one of our unknowns (and also Ψ when is obtained from the solution of a fluid–structure interaction problem). Yet, for many practical situations it is fulfilled whenever Ψ and ω are properly chosen.

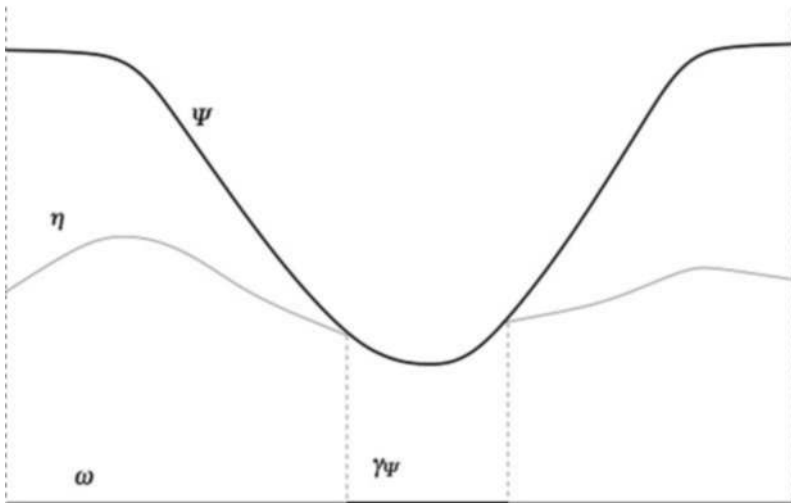


Fig. 12.2 A 2D view of the constrained problem. η is constrained to remain below Ψ at any time

In our case, $\gamma_\Psi(t)$ will denote the horizontal projection of the “submerged surface” of the boat:

$$\Gamma_\Psi(t) = \{(x, y) \in \gamma_\Psi(t), z = \Psi(x, y, t)\},$$

while \mathbf{U} and p denote the velocity and the pressure (scaled with the density), respectively. If $D = \{(\mathbf{x}, t) : t \in (0, T), \mathbf{x} \in \Omega(t)\}$, we have that $\mathbf{U} : D \rightarrow \mathbb{R}^3$ and $p : D \rightarrow \mathbb{R}$.

Furthermore, we put into evidence the x and y components of the velocity by writing

$$\mathbf{U} = (\mathbf{u}, w) = (u_x, u_y, w),$$

and indicating by ∇_{xy} and div_{xy} the gradient and the divergence operator in the (x, y) plane, respectively. The flow equations governing this problem may be conveniently written by introducing a Lagrange multiplier $\lambda : \omega \times (0, T) \rightarrow \mathbb{R}_+$ and solving for a.o. $t \in (0, T)$, the following system for the unknown \mathbf{U} , p , η and λ :

$$\begin{aligned}
\frac{D\mathbf{U}}{Dt} + \operatorname{div} \boldsymbol{\sigma}(\mathbf{U}) + \nabla p - \mathbf{g} &= \mathbf{0} & \text{in } \Omega(t), \\
\operatorname{div} \mathbf{U} &= 0 \\
\frac{\partial \eta}{\partial t} + u_x \frac{\partial \eta}{\partial x} + u_y \frac{\partial \eta}{\partial y} - w &= 0 & \text{in } \omega. \\
\lambda(\eta - \Psi) &= 0, \quad \lambda \geq 0, \quad \eta - \Psi \leq 0
\end{aligned} \tag{12.3}$$

with the additional dynamic condition

$$(\boldsymbol{\sigma}(\mathbf{U}) + p\mathbf{I}) \cdot \mathbf{n} - \lambda \mathbf{n} = 0, \quad \text{on } \Gamma_s(t), \tag{12.4}$$

being \mathbf{n} the outward normal of $\partial\Omega(t)$. This condition implies that the external pressure acting on the free-surface $\Gamma_s(t) \setminus \Gamma_\Psi(t)$ is constant and equal to zero.

Here, $\mathbf{g} = -g\mathbf{e}_z$ is the gravity acceleration, while $\boldsymbol{\sigma}$ denotes the viscous contribution to the internal stress, which in our case may be taken equal to $\boldsymbol{\sigma}(\mathbf{U}) = -\nu \nabla \mathbf{U}$, being ν the water kinematic viscosity, assumed constant. We have indicated by $\frac{D\mathbf{U}}{Dt} = \frac{\partial \mathbf{U}}{\partial t} + (\mathbf{U} \cdot \nabla) \mathbf{U}$ the material derivative. In deriving the first two equations in (12.3) we have assumed that the water density ρ is constant, and we have eliminated it from the equations by scaling. Let us note that the support of $\lambda(t)$ is always contained in $\gamma_\Psi(t)$.

System (12.3) has to be complemented with proper boundary conditions on $\Gamma_f(t)$ and Γ_b , which will be detailed later on, as well as initial conditions on \mathbf{U} and η .

We now exploit the special shape of the domain to operate on (12.3). First, we decompose the pressure as

$$p(\mathbf{x}, t) = g(\eta(x, y, t) - z) + q(\mathbf{x}, t) + \lambda(\mathbf{x}, t), \tag{12.5}$$

where q is the so-called “hydrodynamic correction” (see [11]), while $g(\eta - z)$ is the hydrostatic part.

Furthermore, we integrate the continuity equation along the z -direction by imposing $\mathbf{U} \cdot \mathbf{n} = 0$ on Γ_b and exploiting the kinematic interface condition to obtain

$$\begin{aligned}
\frac{D\mathbf{u}}{Dt} - \nu \Delta \mathbf{u} + g \nabla_{xy} \eta - \nu \frac{\partial \mathbf{u}}{\partial z} + \nabla_{xy} \lambda + \nabla_{xy} q &= \mathbf{0} & \text{in } \Omega(t), \\
\frac{Dw}{Dt} - \nu \Delta w + \frac{\partial q}{\partial z} &= 0 \\
\operatorname{div}_{xy} \mathbf{u} + \frac{\partial w}{\partial z} &= 0 \\
\frac{\partial \eta}{\partial t} + \operatorname{div}_{xy} \int_{-h}^{\eta} \mathbf{u} dz &= 0 & \text{in } \omega. \\
\lambda(\eta - \Psi) &= 0, \quad \lambda \geq 0, \quad \eta - \Psi \leq 0
\end{aligned} \tag{12.6}$$

The dynamic condition on $\Gamma_s(t)$ becomes $\boldsymbol{\sigma}(\mathbf{U}) \cdot \mathbf{n} = \mathbf{0}$, i.e. $\frac{\partial \mathbf{U}}{\partial n} = \mathbf{0}$. We have assumed, as usual in this type of derivations, that the dynamic pressure q is zero on $\Gamma_s(t)$. As a result, the Lagrange multiplier λ may be understood as the pressure field exerted on the water surface by the presence of the boat.

We wish now to solve this problem numerically. To this purpose we first reduce it to a simpler problem, more amenable to numerical analysis, by performing the integration in time.

12.2.1 Characteristic Treatment of the Time Derivative

We subdivide the time interval $[0, T]$ into N sub-intervals of width Δt and we denote with $t^n = n\Delta t$ the n th time step. The subscript n denotes the approximation at $t = t^n$ of the various time-dependent quantities. The method of characteristics consists in performing the following approximation:

$$\frac{DU}{Dt}(\mathbf{x}, t^{n+1}) \simeq \frac{U(\mathbf{x}, t^{n+1}) - U(\mathbf{X}((\mathbf{x}, t^{n+1}; t^n), t^n))}{\Delta t}, \quad (12.7)$$

where $\mathbf{X}((\mathbf{x}, t^{n+1}; t^n), t^n)$ is obtained by solving the following time backward differential problem for each $\mathbf{x} \in \Omega(t)$:

$$\begin{cases} \frac{d\mathbf{X}}{d\tau}(\mathbf{x}, t^{n+1}; t^{n+1} - \tau) = -\mathbf{U}(\mathbf{X}(\mathbf{x}, t^{n+1}; t^{n+1} - \tau), t^{n+1} - \tau), & \tau \in (0, \Delta t), \\ \mathbf{X}(\mathbf{x}, t^{n+1}; t^{n+1}) = \mathbf{x}. \end{cases}$$

More details on this technique may be found in [11] or in [3]. We will now use the short-hand notation \mathbf{X}^n to indicate $\mathbf{X}((\mathbf{x}, t^{n+1}; t^n), t^n)$ and replace (12.6) with the approximation

$$\begin{aligned} \frac{\mathbf{u}^{n+1} - \mathbf{u}^n(\mathbf{X}^n)}{\Delta t} - \nu \Delta \mathbf{u}^{n+1} + g \nabla_{xy} \eta^{n+1} + \nabla_{xy} \lambda^{n+1} + \nabla_{xy} q^{n+1} &= \mathbf{0} \\ \frac{w^{n+1} - w^n(\mathbf{X}^n)}{\Delta t} - \nu \Delta w^{n+1} + \frac{\partial q^{n+1}}{\partial z} &= 0 && \text{in } \Omega^{n+1}, \\ \operatorname{div}_{xy} \mathbf{u}^{n+1} + \frac{\partial w^{n+1}}{\partial z} &= 0 \\ \frac{\eta^{n+1} - \eta^n}{\Delta t} + \operatorname{div}_{xy} \int_{-h}^{\eta^{n+1}} \mathbf{u}^{n+1} dz &= 0 && \text{in } \omega, \\ \lambda^{n+1}(\eta^{n+1} - \Psi^{n+1}) &= 0, \quad \lambda \geq 0, \quad \eta^{n+1} - \Psi^{n+1} \leq 0 \end{aligned} \quad (12.8)$$

where the quantities at time t^n are assumed to be known.

This system of equations can be further modified by adopting an operator-splitting strategy analogous to the one employed in the well-known Chorin–Temam scheme [6] for incompressible fluid dynamics. More precisely, we first perform a *hydrostatic step*, which computes an intermediate velocity field $\tilde{\mathbf{u}}$, as well as λ^{n+1} and η^{n+1} by solving the system

$$\begin{aligned}
\frac{\tilde{\mathbf{u}} - \mathbf{u}^n(\mathbf{X}^n)}{\Delta t} - \nu \triangle \tilde{\mathbf{u}} - \nu \frac{\partial^2 \tilde{\mathbf{u}}}{\partial z^2} + g \nabla_{xy} \eta^{n+1} + \nabla_{xy} \lambda^{n+1} &= \mathbf{0} \quad \text{in } \Omega^{n+1} \\
\frac{\eta^{n+1} - \eta^n}{\Delta t} + \operatorname{div}_{xy} \int_{-h}^{\eta^{n+1}} \tilde{\mathbf{u}} dz &= 0 \\
\lambda^{n+1}(\eta^{n+1} - \Psi^{n+1}) &= 0, \quad \lambda \geq 0, \quad \eta^{n+1} - \Psi^{n+1} \leq 0
\end{aligned} \tag{12.9}$$

followed by a *correction step* for the actual computation of the solution

$$\begin{aligned}
\frac{\mathbf{u}^{n+1} - \tilde{\mathbf{u}}}{\Delta t} + \nabla_{xy} q^{n+1} &= \mathbf{0} \\
\operatorname{div}_{xy} \mathbf{u}^{n+1} + \frac{\partial w^{n+1}}{\partial z} &= 0 \\
\frac{w^{n+1} - w^n(\mathbf{X}^n)}{\Delta t} - \nu \triangle w^{n+1} + \frac{\partial q^{n+1}}{\partial z} &= 0,
\end{aligned} \tag{12.10}$$

in Ω^{n+1} .

It is possible to neglect the hydrodynamics pressure term q , using the so-called hydrostatic approximation. In that case, we set $q = 0$ everywhere, $\mathbf{u}^{n+1} = \tilde{\mathbf{u}}$ and we solve only the second equation in (12.10) to obtain the vertical component of the velocity. Another possible approximation is to neglect the term $-\nu \triangle \tilde{\mathbf{u}}$ in the first equation of (12.9); this has important consequence in the regularity of the solution (see [8] and [9]) and in the numerical scheme.

For what matters at the moment is to notice that with this splitting the unilateral constraint is imposed on a simpler set of equations. We now consider on how to apply the constraint in practise.

12.2.2 Enforcing the Constraint in the Hydrostatic Step

Let us consider (12.9) in more detail. We note that the problem is non-linear because the domain Ω^{n+1} is unknown, as it depends on η^{n+1} . In order to avoid a complex iterative procedure, we linearise the problem by computing a first approximation of Ω^{n+1} based on a full explicit treatment of the free-surface evolution. In practise, we first solve

$$\eta^* = \eta^n + \Delta t \operatorname{div}_{xy} \int_{-h}^{\eta^n} \mathbf{u}^n dz$$

and use it for the approximation of the domain at time t^{n+1} . The actual domain at time t^{n+1} will be calculated at the end of the step from the computed values of η^{n+1} . We now rewrite (12.9) where, for the sake of notation, we drop the superscript $(n+1)$ and the bar, and we set $\alpha = (\Delta t)^{-1}$.

The problem is to find \mathbf{u} , η and λ which satisfy

$$\begin{aligned}
\alpha \mathbf{u} - \nu \triangle \mathbf{u} - \nu \frac{\partial^2 \mathbf{u}}{\partial z^2} + g \nabla_{xy} \eta + \nabla_{xy} \lambda &= \mathbf{f}_u \quad \text{in } \Omega, \\
\alpha \eta + \operatorname{div}_{xy} \int_{-h}^{\eta^*} \mathbf{u} dz &= f_\eta \quad \text{in } \omega,
\end{aligned} \tag{12.11}$$

under the constraints

$$\lambda(\eta - \Psi) = 0, \quad \lambda \geq 0, \quad \eta - \Psi \leq 0, \quad \text{in } \omega, \quad (12.12)$$

being Ψ a given function. We have set $\mathbf{f}_u = \alpha \mathbf{u}''(\mathbf{X}^n)$ and $f_\eta = \alpha \eta''$. The boundary conditions are

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial n} &= 0 \quad \text{on } \Gamma_s \\ \mathbf{u} &= 0 \quad \text{on } \Gamma_b \cup \Gamma_f. \end{aligned} \quad (12.13)$$

It may be recognised that we are facing a classical saddle point problem which may be solved by duality techniques. For a given η and λ in $L^2(\omega)$, the first equation in (12.11) with boundary conditions (12.13) is well posed with $\mathbf{u} \in \mathbf{V} = \{\mathbf{v} \in [H^1(\Omega)]^2, \mathbf{v} = 0 \text{ on } \Gamma_b \cup \Gamma_f\}$.

More precisely, for a given $\phi \in \mathbf{V}'$ we indicate with $\mathbf{y} = \mathcal{F}(z)$ the element of \mathbf{V} which satisfies (in the sense of distribution) the equation

$$\alpha \mathbf{y} - \nu \Delta \mathbf{y} - \nu \frac{\partial^2 \mathbf{y}}{\partial z^2} = \mathbf{f}_u + \phi$$

in Ω , with boundary condition $\frac{\partial \mathbf{y}}{\partial n} = 0$ on Γ_s . The map \mathcal{F} is an isomorphism between \mathbf{V}' and \mathbf{V} . Therefore, (12.11) may be formally written in the equivalent form

$$\alpha \eta + \text{div}_{xy} \int_{-h}^{\eta^*} \mathcal{F}^{-1}(-g \nabla_{xy} \eta - \nabla_{xy} \lambda) dz = f_\eta, \quad \text{in } \omega, \quad (12.14)$$

which, for a given λ , provides an equation for η only. It may be verified that (12.14) is in fact akin to a wave equation for η and efficient numerical solution strategies may be devised for it [11]. We are now ready to state the Uzawa algorithm for our constrained problem.

For a given $\varepsilon > 0$, $\rho > 0$ and $\lambda^{(0)} \in L^2(\omega)$, with $\lambda^{(0)} \geq 0$, solve

$$\alpha \eta^{(k+1)} + \text{div}_{xy} \int_{-h}^{\eta^*} \mathcal{F}^{-1}(-g \nabla_{xy} \eta^{(k+1)} - \nabla_{xy} \lambda^{(k)}) dz = f_\eta, \quad \text{in } \omega,$$

and set

$$\lambda^{(k+1)} = \max(\lambda^{(k)} + \rho(\eta^{(k+1)} - \Psi), 0),$$

for $k = 0, 1, \dots$, until $\|\lambda^{(k+1)} - \lambda^{(k)}\|_{L^2(\omega)} \leq \varepsilon$.

The final iterate is used for the approximation of $\tilde{\mathbf{u}}$, λ^{n+1} and η^{n+1} . Finally, we either perform the full correction step (12.10) or, if we are making the hydrostatic assumption, we just compute the new w using the second equation in (12.10).

12.2.3 The Model for the Dynamics of a Rowing Scull

To compute the boat dynamics we need to couple between the fluid solver and an algorithm for the structural dynamics.

Here the boat is modelled as a rigid body and in the following [1, 2] we have considered two orthogonal Cartesian reference frames. The inertial reference system (O, x, y, z) and a body-fixed reference system (S, x^b, y^b, z^b) , whose origin is the boat centre of mass S , which translates and rotates with the boat. The xy -plane in the inertial reference system is parallel to the undisturbed water surface and the z -axis points upward. The body-fixed x^b -axis is directed from bow to stern and y^b is positive starboard.

The dynamics of the boat in the six degrees of freedom is described by the equations of linear and angular momentum, which in the inertial reference frame are given by

$$M\ddot{\mathbf{S}} = \mathbf{F} \quad (12.15)$$

and

$$\mathcal{R}\mathcal{I}\mathcal{R}^{-1}\dot{\Phi} + \Phi \times \mathcal{R}\mathcal{I}\mathcal{R}^{-1}\Phi = \mathbf{M}_G, \quad (12.16)$$

respectively. Here, M is the boat mass, $\ddot{\mathbf{S}}$ is the linear acceleration of the centre of mass, \mathbf{F} is the resultant of the external forces acting on the boat, $\dot{\Phi}$ and Φ are the angular acceleration and velocity, respectively. Finally, \mathbf{M}_G is the moment with respect to G acting on the boat, \mathcal{I} is the tensor of inertia of the boat about the body-fixed reference system axes and $\mathcal{R} = \mathcal{R}(\Phi)$ is the transformation matrix between the body-fixed and the inertial reference system (see [1] for details).

We here consider the application of the model to the dynamics of a rowing scull. A scull is a competition rowing boat where the oarsmen (also called scullers) hold both left and right oars and act on them synchronously, see Fig. 12.3. The problem is made difficult by the strong unsteadiness of the motion and the interaction with the free surface. Indeed, the varying forces at the oars and, even more importantly, the inertial forces due to the movement of the rowers (who slide over the boat during the rowing action) superimpose to the mean motion a complex system of secondary movements. The latter induce an additional drag, mainly because of the gravitational waves radiating from the boat. Their account can be useful during the design process of a new boat and to understand the effects of different rowing styles or crew composition.

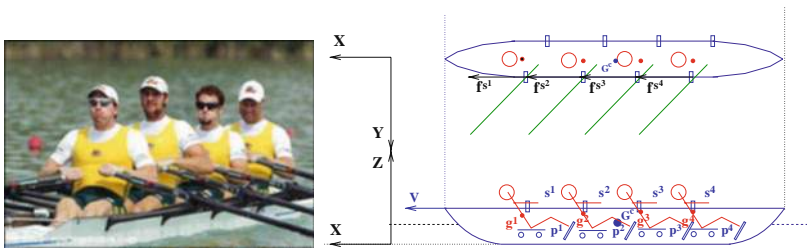


Fig. 12.3 An actual scull (coaxless quad) on the left and its model on the right

Because of the characteristics of a scull, we can assume as a first approximation that the motion takes place in the xy -plane. This implies a great simplification of (12.16), which reduces to a scalar equation, being $\Phi = \phi \mathbf{e}_y$.

The data we have usually available are the forces at the oarlocks \mathbf{F}_{o_j} , inferred from measurements taken on rowing machines and the movement of the rowers. Here, j runs over the number of rowers. The latter can be extracted from a kinematic model of the rower and measurements taken using video-imaging techniques [12], and is usually given as the position $\mathbf{g}_{ij} = \mathbf{g}_{ij}(t)$ of the centre of mass in the boat reference frame of portions of the body of the athlete (e.g. arm, forearm, legs), with corresponding mass m_{ij} (usually taken from anatomic tables as function of the sex, age and weight of the athlete). Here i runs over the number of parts into which the body has been subdivided. If we consider the system formed by the boat and the rowers, we need to provide the force exerted by the rowers on the oar as well, in the following indicated by \mathbf{F}_{h_j} . This force can be easily computed using a model of the oar action, therefore it is here assumed as given. We omit all details of the derivation of the model, which are rather standard and can be found in [7], and we provide only the final result. Let us indicate with

$$\mathcal{R} = \begin{bmatrix} \cos \phi & 0 & -\sin \phi \\ 0 & 1 & 0 \\ \sin \phi & 0 & \cos \phi \end{bmatrix}, \quad \mathcal{O} = \begin{bmatrix} -\sin \phi & 0 & -\cos \phi \\ 0 & 1 & 0 \\ \cos \phi & 0 & -\sin \phi \end{bmatrix}$$

the rotation matrix and its derivative w.r.t. ϕ and with M the mass of the boat. We have that

$$\begin{aligned} & (M + \sum_{i,j} m_{ij})\ddot{S} + \mathcal{O}(\sum_{i,j} m_{ij}\mathbf{g}_{ij})\ddot{\phi} + \mathcal{R}\sum_{j,j} m_{ij}\ddot{\mathbf{g}}_{ij} + 2\mathcal{O}(\sum_{i,j} m_{ij}\dot{\mathbf{g}}_{ij})\dot{\phi} \\ & - \mathcal{R}(\sum_{i,j} m_{ij}\mathbf{g}_{ij})\dot{\phi}^2 = \sum_{j=1}^n \mathbf{F}_{o_j} + \sum_{j=1}^n \mathbf{F}_{h_j} + (M + \sum_{i,j} m_{ij})\mathbf{g} + \mathbf{F}_{\text{Flow}} \end{aligned} \quad (12.17a)$$

and

$$\begin{aligned} & \mathcal{R}(\sum_{i,j} m_{ij}\mathbf{g}_{ij}) \times \ddot{S} + (I_{YY} + \sum_{i,j} m_{ij}|\mathbf{g}_{ij}|^2)\ddot{\phi} \\ & + 2(\sum_{i,j} m_{ij}\mathcal{R}\mathbf{g}_{ij} \times \mathcal{O}\dot{\mathbf{g}}_{ij})\dot{\phi} = -\mathcal{R}\sum_{i,j} m_{ij}\mathbf{g}_{ij} \times \mathcal{R}\ddot{\mathbf{g}}_{ij} \\ & + \mathcal{R}\sum_{j=1}^n \mathbf{g}_{s_j} \times \mathbf{F}_{s_j} + \mathcal{R}\sum_{j=1}^n \mathbf{g}_{m_j} \times \mathbf{F}_{m_j} \mathcal{R}\sum_{i,j} m_{ij}\mathbf{g}_{ij} \times \mathbf{g} + \mathbf{M}_{\text{Flow}}, \end{aligned} \quad (12.17b)$$

where the indexes i and j run from the number of body parts and the number of rowers, respectively. The dependence on t of the various terms is understood.

Equations (12.17) form a system of three non-linear second-order ordinary differential equations in the variables (S_x, S_z, ϕ) that must be complemented with a suitable fluid dynamic model in order to compute \mathbf{F}_{Flow} and \mathbf{M}_{Flow} and close the problem. For instance, the model proposed in the previous sections.

12.2.4 More Realistic Boundary Conditions

We need to make the boundary conditions on Γ_b and Γ_f more realistic. On the bottom, we normally prescribe a friction condition through a Chézy coefficient c_d . Being the bottom flat it corresponds on setting

$$w = 0 \quad \text{v} \frac{\partial \mathbf{u}}{\partial z} = c_d |\mathbf{u}| \mathbf{u}, \quad \text{on } \Gamma_b,$$

the non-linear term in the right-hand side being discretised in time in a semi-explicit fashion. On the far field, we employ a first-order linear radiation condition for the elevation, i.e. we impose

$$\frac{\partial \eta}{\partial t} + \sqrt{gh} \frac{\partial \eta}{\partial n} = 0, \quad \text{on } \Gamma_f(t),$$

which is approximated by using an extrapolation technique akin to the characteristic treatment of the time derivative already illustrated. The modifications to the numerical scheme are straightforward.

12.3 The Interaction Between the Boat and the Water

The scull dynamic model and the flow model have to interact. In particular the forces \mathbf{F}_{Flow} and the angular momentum \mathbf{M}_{Flow} acting on the boat depend on the flow solution. However, the dynamic condition (12.4) implies a zero tangential component of the normal stresses on the boat surface, while the normal component is simply given by λ . Therefore, the proposed fluid dynamics model is able to compute correctly pressure-induced forces, but neglects the viscous drag. Yet, for elongated geometries like a scull the viscous drag $\mathbf{F}_D(\mathbf{U}) = -R(\mathbf{U})\mathbf{e}_x^1$ can be estimated by standard empirical formula, which is quite accurate. Therefore, we may write that

$$\mathbf{F}_{\text{Flow}} = \int_{\Gamma_\Psi} \lambda \mathbf{n} d\gamma + \mathbf{F}_D(\mathbf{U}) = \int_{\omega} \lambda \left[\frac{\partial \Psi}{\partial x}, \frac{\partial \Psi}{\partial y}, 1 \right]^T dx dy + \mathbf{F}_D(\mathbf{U}).$$

We have here exploited the fact that the normal surface Γ_Ψ is given by $\mathbf{n} = (\sqrt{1 + |\nabla_{xy} \Psi|^2})^{-1} [\frac{\partial \Psi}{\partial x}, \frac{\partial \Psi}{\partial y}, 1]^T$ and, for the sake of completeness, we have given the general formula, while in the case of a scull the y component of \mathbf{F}_{Flow} is zero because of symmetry considerations. An analogous formula may be obtained for the computation of the couple \mathbf{M}_{Flow} induced by the action of the flow. We have also implicitly used the fact that $\lambda = 0$ outside the area where the boat is present.

The boat dynamical system describes the position of the boat and thus implicitly defines the function Ψ . Let $\mathcal{B}_0 = \{(x^b, y^b, z^b), (x^b, y^b) \in B \subset \mathbb{R}^2, z^b = \hat{r}_0^b(x^b, y^b)\}$

¹ In fact the drag is also a function of the submerged surface.

be the parametric description of the boat's external surface (the skin) in the boat reference frame, usually provided by means of analytic functions. We first extend \hat{r}^b with continuity to the whole \mathbb{R}^2 in a suitable way, and let \hat{r}^b indicate this extension. The extended boat geometry at time t can then be described as $\mathcal{B}(t) = \{(x, y, z), \quad (x, y) \in \mathbb{R}^2, z = r(x, y, t)\}$, where

$$r(x, y, t) = S_z(t) + \tan \phi(t)[x - S_x(t)] + \cos^{-1} \phi(t) \hat{r}^b(\mathcal{R}(\phi(t))[x - S(t)]).$$

Finally, Ψ can be taken as r restricted to ω .

12.4 Numerical Results

When considering the dynamics of the scull, the value of Ψ at each time step is given by solving equations (12.17), where the hydrodynamic forces are computed by integrating the surface stress provided by the Navier–Stokes model just presented.

For the space discretisation we have adopted a finite element scheme which employs Raviart–Thomas \mathbb{RT}_0 triangular elements in the (x, y) plane for \mathbf{u} and standard P^1 elements for w . The elevation η , as well as the multiplier λ , is approximated by a piecewise constant function (i.e. P_0 finite elements). Details are given in [11].

We have implemented a simple time integration procedure of the coupled problem. We evolve from time step t^n to t^{n+1} as follows:

- the body position is integrated explicitly using the fluid dynamic forces $\mathbf{F}_{\text{Flow}}^n$ and $\mathbf{M}_{\text{Flow}}^n$ computed from the flow solution at time step t^n ;
- once the approximation Ψ^{n+1} of the constraining surface is available, we solve the fluid dynamic problem using the Uzawa algorithm; and
- once η^{n+1} , λ^{n+1} and \mathbf{U}^{n+1} have been obtained we move to the next time step.

This explicit scheme is subject to an absolute stability condition. Yet, the time steps required to capture the rather fast dynamics of the generated waves have been found to be within the stability bounds, at least for the computations carried out so far.

The numerical results presented in the following section have been obtained using the hydrostatic assumption, i.e. the flow solution is computed neglecting the hydrodynamic pressure term q .

12.4.1 Sinking and Pitching Motions

For all the following dynamic simulation the scull has been approximated by a semi-ellipsoid. Its geometric, mass and inertia characteristics are summarised in Table 12.1.

The first simulation is a pure sinking motion: the hull is free to move in the z -direction subject to its weight. The initial position is at the centre of a square basin of

Table 12.1 Mass and inertia characteristics

Length [m]	Breadth [m]	Height [m]	Mass [kg]	\mathcal{I}_{yy} [kg m ²]
6	0.8	0.6	400	930

edge length 15 m. At time $t = 0$ the body is steady at $z = 0.6$ m over the free surface. The motion, represented in figure, is a sequence of damped oscillations and after a few seconds the vertical position levels off at $z = 0.30$ m. The asymptotic sinking is in good agreement with the theoretical equilibrium position of $z = 0.325$ m.

Pure pitching motion was also simulated: vertical position was fixed and a non-zero initial pitch angle $\theta_0 = 1.5^\circ$ was assigned. As in the sinking motion, oscillations damp out and pitch angle tends to its zero equilibrium value. Damping is considerably lower compared to the previous simulation, due to the smaller wave amplitude generated. This is also in accordance to experience. Figure 12.4 represents the wave pattern generated by sinking and pitching motion (beware: colour scales are different).

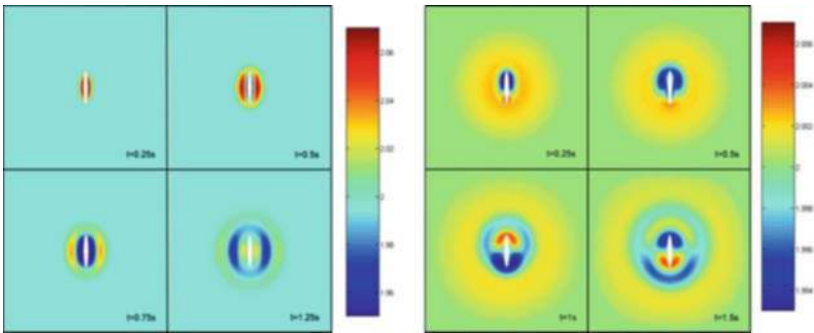


Fig. 12.4 Wave pattern for sinking (left) and pitching motion (right)

12.4.2 Reproducing Mean Motion Wave Pattern

A further test concerns the wave pattern generated by the advancing motion on free surface. In *shallow water* regime, i.e. for $H < 2v^2/g$, theory predicts for bow and stern waves a semi-angle $\beta = \arcsin\left(\frac{c}{v}\right)$, where the wave speed c is constant in the hydrostatic assumption and equal to \sqrt{gh} (see [10]). Setting $h = 3$ m this angle turns out to be 64.7° .

In Fig.12.5 the predicted angle is overlaid on the calculated wave pattern (colour scale is proportional to free-surface elevation). The agreement is satisfactory demonstrating the effectiveness of the procedure.

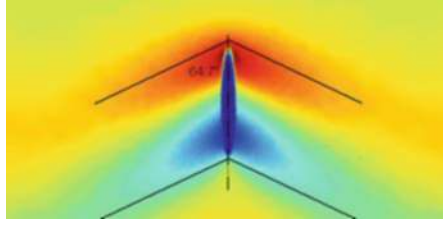


Fig. 12.5 The wave pattern generated and its expected angle

12.4.3 An Example with the Full Dynamics

We have here considered a coaxless quad scull. The first picture in Fig. 12.6 illustrates the wave pattern generated by the boat moving at the constant mean velocity, computed using the model given in Sect. 12.2. The second and third pictures illustrate that obtained at the instant of the catch and at the release, when the full dynamics of the boat is considered. We have assumed a stroke period of 1.5 s.

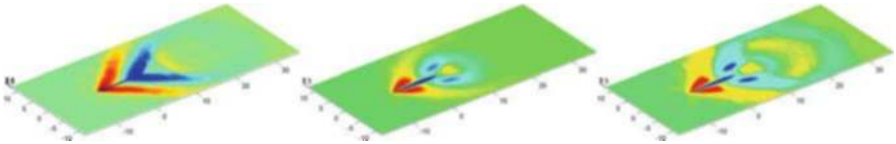


Fig. 12.6 The surface wave pattern for the mean motion (left) and at two different time instants obtained using the full boat dynamics

The alteration to the wave pattern caused by the secondary motions is evident. Comparison with experimental data is currently under way. So far, we have carried out only qualitative assessment comparing the wave pattern with that obtained from video recording, with good agreement.

12.4.4 A Final Detail

The numerical model described so far has a practical disadvantage. As the boat moves it will eventually reach the boundary of the computational domain. As a consequence, to simulate the boat during a race for reasonably long periods, we may need a rather large ω , with an increase in the computational costs. We have successfully overcome this problem by re-writing the flow equations in a non-inertial reference system with origin on the boat's centre of mass S , and axis directions kept fixed. What is needed is the addition of the inertial forces and some changes in the boundary conditions in the flow equations. In this way the variations in $\gamma\psi$ are only due to the sinking and pitching motions, while the boat centre remains fixed. For

the sake of brevity we have not reported here the modified equations even if the last computations here shown have been indeed computed this way.

Acknowledgments The authors wish to thank Filippi Lido s.r.l for the financial and technical support and in particular Ing. Alessandro Placido for having introduced them to the wonderful world of rowing. A thank also to Andrea Paradiso for making available some results from his master theses.

The authors want to remember the late Fausto Saleri, who has largely contributed to the development of some of the ideas here illustrated, before leaving us untimely.

References

1. R. Azcueta. Computation of turbulent free-surface flows around ships and floating bodies. *Ship Technol. Res.*, 49(2):46–69, 2002.
2. R. Azcueta. RANSE Simulations for Sailing Yachts Including Dynamic Sinkage & Trim and Unsteady Motions in Waves. In *High Performance Yacht Design Conference*, pages 13–20, Auckland, 2002.
3. K. Boukir, Y. Maday, B. Métivet, and E. Razafindrakoto. A high-order characteristics/finite element method for the incompressible Navier-Stokes equations. *Int. J. Num. Meth. Fluids*, 25(12):1421–1454, 1997.
4. U. P. Bulgarelli. The application of numerical methods for the solution of some problems in free-surface hydrodynamics. *J. Ship Res.*, 49(4):288–301, 2005.
5. U. P. Bulgarelli, C. Lugni, and M. Landrini. Numerical modelling of free-surface flows in ship hydrodynamics. *Int. J. Num. Meth. Fluids*, 43(5):465–481, 2003.
6. A. J. Chorin. Numerical solution of the Navier Stokes equations. *Math. Comp.*, 22:745–762, 1968.
7. L. Formaggia, E. Miglio, A. Mola, and N. Parolini. Fluid-structure interaction problems in free surface flows: application to boat dynamics. *Int. J. Num. Meth. Fluids*, 2007. DOI 10.1002/fld.1583 (in press).
8. J. L. Lions, R. Temam, and S. Wang. On the equations of the large-scale ocean. *Nonlinearity*, 5:1007–1053, 1992.
9. J. L. Lions, R. Temam, and S. Wang. On mathematical problems for the primitive equations of the ocean: the mesoscale midlatitude case. *Nonlinear Anal.*, 40:439–482, 2000.
10. C. C. Mei. *The applied dynamics of ocean surface waves*. World Scientific Publishing, Singapore, 1989. Second printing with corrections.
11. E. Miglio, A. Quarteroni, and F. Saleri. Finite element approximation of quasi-3D shallow water equations. *Comp. Meth. Appl. Mech. Engng.*, 174(3–4):355–369, 1999.
12. A. Mola, L Formaggia, and E. Miglio. Simulation of the dynamics of an olympic rowing boat. In *Proceedings of ECCOMAS CFD 2006, Egmond aan Zee, September 5–8, The Netherlands*. TU Delft, 2006. ISBN: 90-9020970-0.
13. H. Orihara and H. Miyata. Evaluation of added resistance in regular incident waves by computational fluid dynamics motion simulation using an overlapping grid system. *J. Mar. Sci. Technol.*, 8(2):47–60, 2003.
14. N. Parolini and A. Quarteroni. Mathematical models and numerical simulations for the America’s Cup. *Comp. Meth. Appl. Mech. Engng.*, 194(9–11):1001–1026, 2005.
15. N. Parolini and A. Quarteroni. Modelling and numerical simulation for yacht design. In *Proceedings of the 26th Symposium on Naval Hydrodynamics, Rome, Italy, 17–22 September 2006*, 2007. To appear.
16. C. Yang and R. Lohner. Calculation of ship sinkage and trim using a finite element method and unstructured grids. *Int. J. Comput. Fluid D.*, 16(3):217–227, 2002.

“This page left intentionally blank.”

Chapter 13

Concepts of Active Noise Reduction Employed in High Noise Level Aircraft Cockpits

Hatem Foudhaili and Eduard Reithmeier

Abstract During the past two decades, reducing exposure to high-level noise in aircraft cockpits by methods of active noise control (ANC) has aroused the interest of researchers. Also, some commercial applications were initiated by leading manufacturers. For this purpose, fundamentally different approaches were used. While active noise compensation reduces the noise level by generating an interfering antinoise, structural vibration control aims to limit sound emittance through active damping of the aircraft structure vibrations. These approaches are linked with very different financial and technical boundary conditions, which implied distinct degrees of success. The ANC approaches used in cockpit noise reduction will be summarised, and their success or failure reasons will be analysed. Thereafter, the focus will be set on the industrially more successful way of protecting pilots from high noise levels, which is the use of active headsets. The development and the current state of commercial products will be presented, and the requirements of future trends will be derived. These requirements consist in extending the band width of noise reduction and making the control adaptive to changing conditions. Finally, the development of a prototype of a new generation of ANC headsets is presented. The prototype combines standard feedback with adaptive feedforward control techniques and processes the control algorithms by an integrated DSP platform.

Hatem Foudhaili

Institute of Measurement and Automatic Control, Leibniz Universitaet, Hannover, Germany,
e-mail: hatem.foudhaili@imr.uni-hannover.de

Eduard Reithmeier

Institute of Measurement and Automatic Control, Leibniz Universitaet, Hannover, Germany,
e-mail: eduard.reithmeier@imr.uni-hannover.de

13.1 Passive Versus Active Noise Reduction

To distinct passive from active noise reduction, we advance a definition based on the information flow in a noise-reducing system. A system based on an open loop reaction mechanism to reduce sound energy is considered to be passive. Sound proofing and non-actuated resonator plates are examples of passive noise reduction devices. These are used for example to absorb sound energy in anechoic rooms to produce free-field conditions. We define an active noise reduction device as an autonomous decision support-based system with a closed loop reaction mechanism. Generally, these systems are reducing sound pressure level by means of actuators. In this chapter we will focus on the reduction of the exposure of pilots and passengers to high-level noise inside the interior space of aircrafts. This has been primarily achieved by means of passive sound proofing. But during the past two decades, active approaches of noise reduction increasingly emerged, mainly in research projects and at a lower degree in industrial applications. The motivation of active noise reduction is related to the fact that passive sound proofing requires the use of bulky materials to effectively reduce low-frequency noise. This comes obviously into conflict with the critical constraint of reducing the weight of aircrafts. Especially in the frequency range up to 100 Hz, active noise reduction could achieve considerably superior results at a lower weight load. We classify the concepts of active noise reduction employed in high noise level aircraft cockpits in three general categories:

- Active noise cancellation
- Active structural/acoustic control (also called structural vibration control)
- Active noise control in aviation headsets

Our classification is based on scientific and technical criteria and boundary conditions. Although there exists a great number of research publications related to active noise reduction, practical applications are still limited [9]. In order to procure a view about the applicability of active noise reduction in the aerospace industry, in this contribution a focus will be set on the reporting of industrial applications and application-oriented research activities.

13.2 Active Noise Cancellation

Active noise cancellation is the reduction of sound wave level through generation of a phase-delayed wave – generally called antinoise.

The superposition of the primary disturbing noise and the generated antinoise modifies the sound field characteristics and at some areas of the space, destructive interference leads to a cancellation of the disturbance. Figure 13.1 shows a realisation of active noise cancellation. This approach generally needs the feedback information of a so-called “error microphone” to generate a controlled antinoise. In some applications the generation of antinoise is additionally based on a reference signal of the disturbance source. This is not necessarily a microphone signal, it could be,

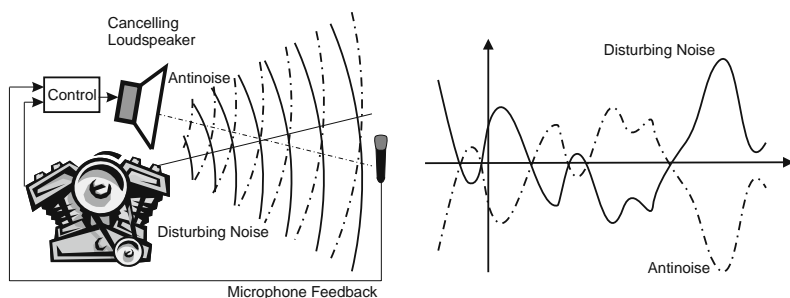


Fig. 13.1 Active noise cancellation

for example, a tachometer information of an engine. The actuators used in active noise cancellation are generally loudspeakers. Active noise cancellation is usually called active noise control, which is not exactly the same, since active noise control is a more general notion including, for example, controlled sound field design which may not pursue any aim of noise reduction.

Among the tasks of the advanced subsonic technology (AST) programme initiated by the National Aeronautics and Space Administration (NASA) between 1992 and 2000, active noise cancellation was incorporated as a potential promising technology to be used in the reduction of fan noise level.

Figure 13.2 shows the “active noise control fan” constructed by NASA, which is a low-speed fan specifically designed for active noise control testing. The system aimed to reduce fan noise level in both the inlet and the aft ducts via controlled loudspeakers.

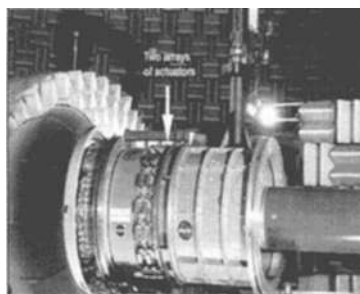


Fig. 13.2 NASA prototype for active noise cancellation of fan noise

A second example of a leading application-oriented research activity related to active noise cancellation is given by a cooperation work conducted by a consortium constituted by the German Aerospace Center (Deutsches Zentrum fuer Luft – und Raumfahrt – DLR), the European Aeronautic Defence and Space Company (EADS) and Germany’s leading aircraft engine manufacturer MTU aero engines. This work was incorporated as a sub-project of the research cluster Turbotech II from 1996 till 2000.

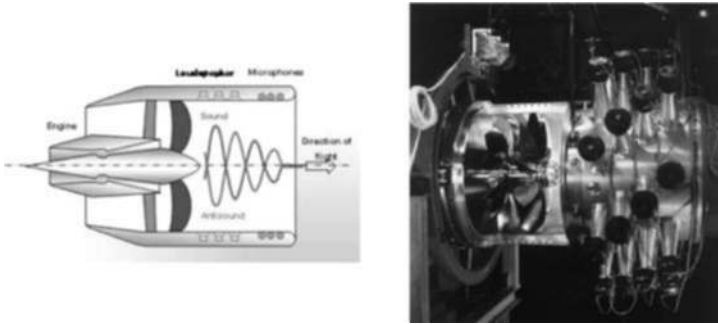


Fig. 13.3 DLR/EADS/MTU active noise cancellation of engine sound

The constructed prototype within this project is shown in Fig. 13.3. The active noise cancellation system incorporates 32 microphones and 32 loudspeakers. According to the German Aerospace Center in its final report [3], the project was completed with “great success”, as far as the prototype investigation is concerned. Currently, the German Aerospace Center is informing that works on active noise cancellation are being carried out in cooperation with the aircraft engine manufacturers MTU, SNECMA and Rolls-Royce to achieve an industrial realisation.

From the first-mentioned project of NASA, a widely deviant appreciation of the industrial applicability of active noise cancellation is reported. In its evaluation of the prospects of active noise cancellation technology [6], NASA mentioned that active noise cancellation was never successfully demonstrated in a relevant environment by the end of the AST programme. This programme element was dropped when the AST programme was terminated by the year 2000 and the work did not continue under the successor programme effort.

These widely deviant appreciations of the prospects of active noise cancellation in the given examples are symptomatic for the widely variegated presentiments among the researcher and manufacturer communities towards the technology during the last years. In the 1990s the majority of leading researchers in the field of active noise cancellation like P.A. Nelson and S.J. Elliott in 1993 [12], S.M. Kuo and D.R. Morgan in 1996 [1] and L.J. Eriksson in 1997 [4] prophesied a great success for the technology in this current decade. Only one leading researcher, C.H. Hansen, warned from too much “unfounded optimism in statements made in the media about the potential applications of the technology”, as he wrote in 1997 in [8]. Considering the current expansion of industrial applications, Hansen is right after all. In 2004 he reexamined the situation and stipulated that the unrestrained, unfounded and, as he accused, sometimes insincere optimism of the 1990s resulted in the current scepticism of manufacturers towards the active noise cancellation technology [9].

One of the more popular reasons for the narrowness of industrial applications was and still is the elevated costs related to the hardware requirements related to the technology. We confirm this reason but consider that it is lightheaded and counterproductive to restrict the discussion to this constraint. Actually, the optimism of the 1990s was widely based on the consideration that the main constraint to the

expansion of the technology is the high signal processing effort and that in some years the decrease of the costs of digital devices and the increase of their capacity will necessarily engender a breakthrough of the active noise cancellation technology. As could be noticed, during the last 15 years a tremendous change occurred in the costs and capacity of digital devices but this did not have a notable effect on the applicability of the active noise cancellation technology in real environments. During our research activity in the field of active noise cancellation we identified some other reasons which could explain the current situation:

- Complexity of the task of controlling a three-dimensional sound field.
- Signal processing and control engineering communities have a lack of knowledge of the physical limitations of noise control in real acoustical environments due to principles of room acoustics and special characteristics of acoustical sensors and actuators.
- Active noise cancellation is often realised by control engineers with methods of adaptive signal processing to solve a complex acoustical problem. Since only few researchers are well qualified in all these fields, there is an imperative need for a well-functioning multidisciplinary team with specialists from the involved fields of control engineering, signal processing and acoustics.
- Too much academic research without any ambition of application and an unbalanced ratio of fundamental research to application-oriented research.
- No possibility for volume production since each new environment requires a new custom-made solution.
- Existence of competing and promising technologies like active structural/acoustic control.

Thus, it becomes difficult to advance an expectation of the potentials of active noise cancellation for the future, particularly because there exist few examples of successfully functioning systems in real “common” environments. As far as aircraft cabins are concerned, in the past there was a unique example of successful industrial application of active noise cancellation. In 1994 the concern Ultra Electronics developed the system “UltraQuiet” as a retrofitting device for a Saab 2000 aircraft. Figure 13.4 shows the components of the noise cancellation system “UltraQuiet”.



Fig. 13.4 Components of the Ultra Electronics system of active noise cancellation “UltraQuiet”

The first introduced “UltraQuiet” system was a tonal active noise cancellation system for quieting cabins of turboprop and rear-engined jet aircrafts. In the following years this system was integrated to other turboprop aircrafts as a standard equipment like the Q-Series Dash 8 (since 1996) of Bombardier Aerospace, the Saab 240 (1996) and the Beech King Air 350. In 2004 Ultra Electronics reported that it has over 700 active noise cancellation systems in operation [7]. The amount of systems integrated as standard fit attests the success of this realisation. From the technical point of view the success could be explained with the reason that the noise reduction task was achieved by tonal noise cancellation, which means that the system generates only harmonic waves. It consists in the generation of a harmonic wave based on a reference signal (generally a tachometer signal) and the adaptation of only two parameters: the amplitude and the phase delay. This method is extensively simpler than broadband noise cancellation and provides very good results in a spatially confined sound field with a considerably dominant low-frequency harmonic, which is the case for the mentioned aircrafts. Our expectation for the future of active noise cancellation in aircrafts is that it has rather prospects of being used in niche markets, where the control task is simplified like the example stated above, than in general global noise reduction tasks. As an alternative noise reduction solution it could be primarily used as retrofitting in specific environments, where, for example, no transformation of the construction through passive or active vibration control devices is allowed. The use of loudspeakers which are already existing or which could be easily integrated in the interior space of an enclosure presents an advantage with respect to a potential use as a retrofitting solution. A second advantage of the active noise cancellation is the ability to design sound fields. Active noise control could focus on a certain point of the space in which the sound level is reduced or the spectrum is selectively changed with a minimum of effort.

13.3 Active Structural/Acoustic Control (ASAC)

Active structural/acoustic control (ASAC) aims to reduce sound level through vibration reduction of sound-emitting structures. This technology already exists as standard equipment in some automotive applications; it is now being developed by the same concerns which formerly did not succeed to realise reliable active noise cancellation systems. As far as industrial applications and application-oriented research activities are concerned, also in the aerospace industry we noticed in the last years a trend to give up research on active noise cancellation for the benefit of active structural/acoustic control. This could be deduced from the increasing active structural/acoustic control approaches investigated by application-oriented researchers and leading manufacturers in the last years. Within the same NASA AST programme mentioned in Sect. 13.2 diverse research activities related to active structural/acoustic control were carried on from 1992 to 2000. To reduce the sound emittance of engines, active rotor blades with embedded piezoelectric actuators to control the magnitude of blade vibrations were developed. In this regard, the HCC

(higher harmonic control) strategy aimed to realise an active blade pitch through excitation of the swashplate by dynamic actuators, while the IBC (individual blade control) technique fulfilled an active blade root pitch through replacement of the pitch links with high-frequency actuators. In 1997 within a research project of the Massachusetts Institute of Technology, individual blade control for the purpose of reducing rotor vibrations and noise was realised by an active flexible blade. Active fibre composites were used to induce shear stresses and hence a twisting moment along the blade. For the reduction of helicopter interior noise the concerns Daimler-Chrysler Aeroacoustics and Eurocopter presented in 1999 a new approach of active vibration isolation realised on a BK117 helicopter. They identified that the structure-born noise path via the gearbox struts is dominant and stipulated that it should be sufficient to control the structure-born noise by applying additional control forces to the strut. They constructed the prototype of smart gearbox struts shown in Fig. 13.5 where the control forces were induced through piezoceramic shells. The reached results were reported in [11].

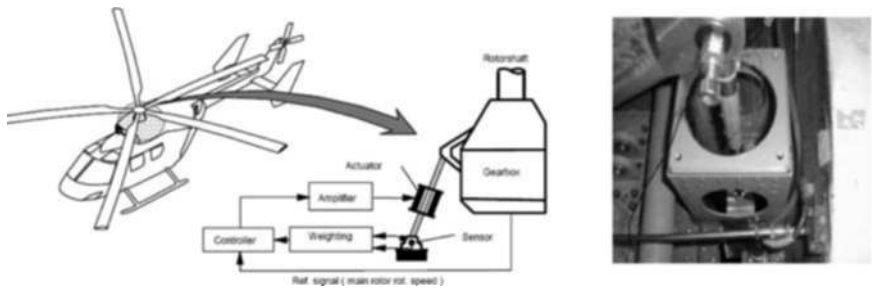


Fig. 13.5 DaimlerChrysler Aeroacoustics and Eurocopter smart gear struts

These were examples of realisations preventing vibrations to arise from the source. Other approaches of active structural/acoustic control aim to inhibit the transmission of the vibrations through the structure of the aircraft. To reduce the interior noise NASA in cooperation with Raytheon–Beech Aircraft used in the year 1999 in a prototype construction shown in Fig. 13.6, 21 inertial force actuators and

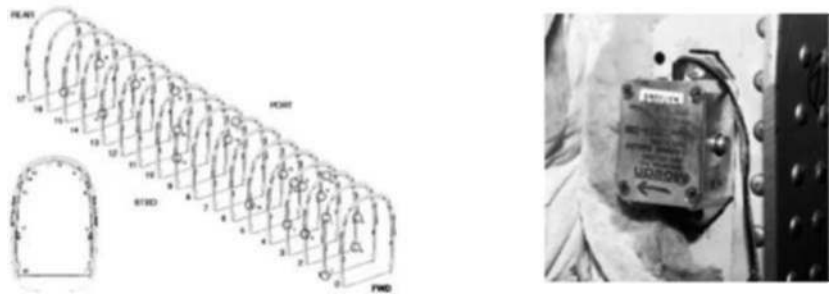


Fig. 13.6 NASA and Raytheon–Beech aircraft ASAC system with 21 inertial force actuators mounted to the aircraft frame

32 microphones mounted directly to the aircraft frame of a Raytheon–Beech 1900D. The reached results were reported in [2]. The actuators produce controlled inertial forces to counter excitation forces arising from the vibrating source. As an alternative to inertial force actuators, within the AST programme, NASA investigated the structure vibration reduction by means of piezoceramic actuators bonded to the outer surface of the trim panel [14]. In a similar approach the EADS Corporate Research Center France published in 2002 its results of using piezoelectric actuators to control vibrating plates and thus take influence on sound transmission in aircraft interior [13].

The trend we noticed of giving up the active noise cancellation technology for the benefit of the active structural/acoustic control technology is confirmed with the revealing example of the further development of the only existing commercial active noise cancellation system for the reduction of aircraft interior noise, “UltraQuiet”. Its developing concern, “Ultra Electronics”, introduced active tuned vibration attenuators (ATVAs) that were mounted to brackets fitted to the aircraft fuselage as actuators in replacement of loudspeakers. The active tuned vibration attenuators of the “UltraQuiet” system undertook the same task as the NASA inertial force actuators mentioned above of damping structure vibrations and hence reduce sound emittance. Microphones were further on used as sensors. Through controlled excitation of the ATVAs the vibrations of the aircraft structure and hence the emitted sound were reduced.

Ultra Electronics, in cooperation with Bombardier Aerospace, reported in 2002 about the realisation of the above-described system comprising 42 ATVAs and 84

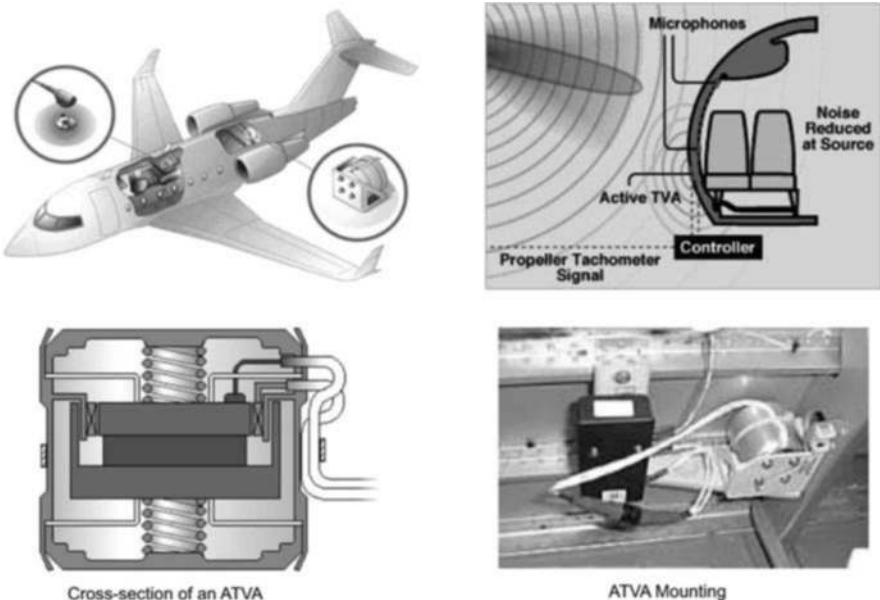


Fig. 13.7 Ultra Electronics ASAC system with active tuned vibration attenuators (ATVA)

sensors, of which 80 are microphones [10]. Ultra Electronics stated in this publication three reasons for its choice to use active structural/acoustic control instead of active noise cancellation. First, there are significantly more potential locations to install ATVAs than loudspeakers. This results in a “finer resolution” of potential actuator locations, which allows better spatial matching of the actuators relative to the sound field within the aircraft. Second, for a production system installing the ATVAs onto the fuselage is much simpler than installing loudspeakers through the trim. Third, unlike active noise cancellation, active structural/acoustic control allows both noise and vibration control. Similar reasons are given by publications of other manufacturers and research organisations like in [2, 11] and [13].

We extend the reasons for the use of active structural/acoustic control instead of active noise cancellation by the fact that active structural/acoustic control intervenes in the sound generation process at an early stage, effecting directly the primary sound source. The control of a vibrating plate is simpler than the control of a three-dimensional sound field, since complex effects of room acoustics like interferences and near-field/far-field characteristics are not to be taken into account. Also, the effect of plant time delay of acoustical transfer paths, which is from the point of view of controllability unfavourable, does not exist.

In the mentioned publication [10] from the year 2002, the manufacturers Ultra Electronics and Bombardier Aerospace announced that they had 53 Q400 aircrafts in service throughout the world, all with the active structural/acoustic control system installed. Currently, Bombardier Aerospace integrated this active structural/acoustic system as standard equipment of the Q-Series of Bombardier’s Dash 8 in replacement of the former active noise cancellation system.

13.4 Active Aviation Headsets

This approach is in fact an active noise cancellation solution, but due to very different technical and practical boundary conditions it will be treated apart. Since 10 years, active aviation headsets have been representing the unique widespread and successful commercial application of active noise cancellation. In fact it is a noise cancellation task in a very confined space where there is no need for global noise cancellation with multiple sensors and actuators. The following facts explain how the noise cancellation control task in headsets is simplified.

- Each ear cup of an active aviation headset is a single input/single output system, as shown in Fig. 13.8.
- The vicinity of the sensing microphone to the ear reduces the general three-dimensional sound field control task to a noise reduction problem at a unique point of space.
- The proximity of the actuator to the sensing microphone reduces the time delay of the control response.
- The almost unchanging conditions within the ear cup lead to a relatively unchanging plant compared to a general active noise cancellation task in a

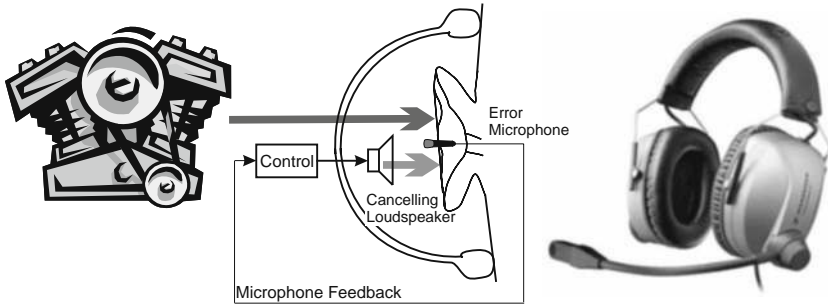


Fig. 13.8 Active aviation headset, theory and commercial application: ANC Headset HMEC450 of Sennheiser

room with changing reflection characteristics through potential geometrical rearrangements.

These favourable technical boundary conditions related to the ear cup enclosure made an industrial application possible at a very early stage of the research in the field of active noise cancellation. Additionally, from a cost-effectiveness point of view, an active aviation headset is designed for use in any environment and hence enables a volume production while an active noise cancellation system for a room is a custom-made solution for each different environment. The problem of custom-made production was noticed by C.H. Hansen who tried itself to implement active noise cancellation solutions for rooms [9]. He insisted that a breakout of the active noise cancellation in rooms could only be reached if the researchers try to develop less specific and more generic solutions which could be implemented and adapted by less-specialised staff. Already at the end of the 1980s, the first commercially successful active aviation headsets were manufactured independently by Bose and Sennheiser. From that time on, many manufacturers like David Clark, Peltor, Telex etc., are developing and successfully bringing to market active aviation headsets.

13.5 An Aviation Communication Headset Prototype with Digital Adaptive Noise Reduction

Commercial active aviation headsets have been based on non-adaptive, analogue and mainly feedback control techniques. However, during the last two decades the digital signal processing has been increasingly used by researchers in the domain of active noise control. The trend of ever-growing performance of processors simultaneously to the reduction of their size and costs has made possible the use of adaptive algorithms in practical applications. Particularly adaptive digital feedforward control techniques are considered to be realistic and promising approaches to be implemented in commercial active aviation headsets. Numerous works describe different active noise controller structures and optimisation algorithms by use of

either feedback or feedforward strategies. Especially in active headsets, the simultaneous use of both control strategies could be of great benefit [1].

In the following, a new prototype of an active noise cancellation headset is presented. This work was achieved in cooperation with the concern *Sennheiser electronic*. The noise cancellation strategy uses an adaptive feedforward control technique. The advantage of adaptive feedforward active noise control is the ability to control high frequencies and to focus on the reduction of the dominant frequency band of the disturbance. However, the implementation of digital adaptive algorithms is linked to high expenses. In a commercial application these expenses should be kept within a realistic limit. A promising issue was to confer a part of the cancelling task to a non-adaptive feedback controller, in order to save calculating and memory resources. This was achieved by a combination strategy of feedback and adaptive feedforward control, in which the adaptive feedforward component is intended to cancel high frequencies and to focus on specific dominant noise, while the feedback component is designed to cancel low-frequency noise. More detailed information related to the control strategy were provided by the authors in former publications [5, 15].

The realised prototype, which processes the control strategy by an integrated fixed-point DSP platform, was tested in the interior of a *Dornier DO228-212* turboprop aircraft, as shown in Fig. 13.9. During the flight, the sound pressure level averaged 105 dB SPL.



Fig. 13.9 Testing a new prototype of active aviation headset in a Dornier DO228-212

Figure 13.10 presents the results of the active noise reduction of the developed new headset prototype in comparison with a current commercial ANC headset. In the relevant frequency range of the disturbance of the turboprop aircraft (up to 1 kHz), the prototype outperforms the commercial headset in terms of noise reduction at an average of 15 dB. Especially within the dominant frequency band of the noise disturbance between 80 and 150 Hz, the new prototype was able to outper-

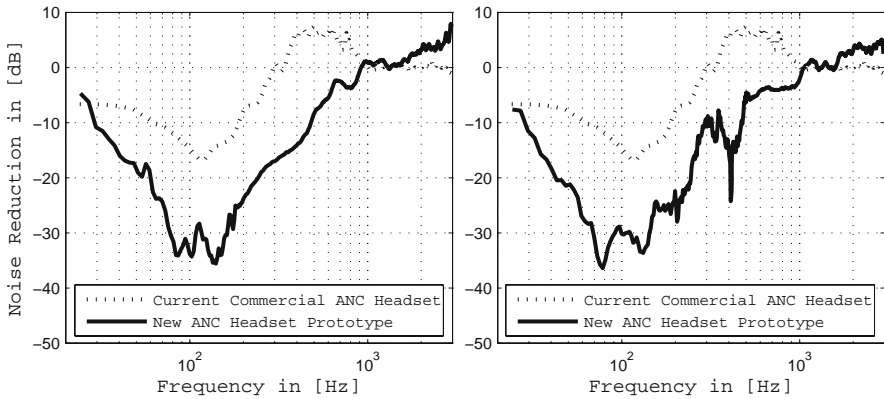


Fig. 13.10 Active noise reduction of a new prototype of ANC headset for the left and the right ear cup

form the commercial headset by 20 dB to reach a total active noise reduction of more than 30 dB. The adaptive digital control techniques offer a great improvement potential to the performance of active noise reduction aviation headsets. Nevertheless, the perspectives of commercial success in the future will depend on the ability to implement these techniques at a reasonable expense–benefit ratio. This includes the acceptance of a minor loss of performance for the benefit of a considerable save of realisation costs. Some issues could be given by the realisation of hybrid analogue/digital control schemes or the development of fast algorithms to be implemented on low-cost digital platforms.

13.6 Conclusions

This contribution provided an overview of active approaches used for the protection of passengers and pilots from high-level noise in aircraft cockpits. Three approaches were presented: active noise cancellation, active structural/acoustic control and active aviation headsets. These approaches are linked to different technical and financial boundary conditions, which implied distinct degrees of success. Related to each technique, some examples of applications of leading aerospace manufacturers and research organisations were given and their success or failure reasons were analysed. Finally, a prototype of a new generation of aviation headsets was presented.

References

1. Kuo, S.M., Morgan, D.R.: *Active Noise Control Systems, Algorithms and DSP Implementations*. Wiley-Interscience Publication, New York (1996)

2. Cabell, R., Palumbo, D., Viperman, J.: A Principal Component Feedforward Algorithm for Active Noise Control: Flight Test Results. *IEEE Transactions on Control Systems Technology*, **9**(1), 76–83 (2001)
3. Enghardt, L., Tapken, U., Neise, W., Schimming, P.: Experimentelle Untersuchungen zur aktiven Schallminderung. Abschlussbericht, Turbotech II, Teilprojekt 1.231 Foerderkennzeichen 0327040D (2000)
4. Eriksson, L.J.: A Primer on Active Sound and Vibration Control. *Sensors*, **14**(2), 18–31 (1997)
5. Foudhaili, H., Wolter, B., Reithmeier, E., Peissig, J.: Feedback-Feedforward aktive Laermkompensation fuer den Kopfhoeerer. *Fortschritte der Akustik*, 33. Jahrestagung fuer Akustik DAGA, Stuttgart, 705–6, (2007)
6. Golub, R.A., Rawls, J.W., Russell, J.W.: Evaluation of the Advanced Subsonic Technology Program Noise Reduction Benefits. NASA Center for AeroSpace Information, (2005)
7. Gorman, J., Hinchliffe, R., Stothers, I.: Active Sound Control on the Flight Deck of a C130 Hercules. *Proceedings of the 2004 International Symposium on Active Control of Sound and Vibration*, CD-ROM (2004)
8. Hansen, C.H.: Active Noise Control - from Laboratory to Industrial Implementation. *Proceedings of NOISE-CON97*, **1**, 3–38 (1997)
9. Hansen, C.H.: Current and Future Industrial Applications of Active Noise Control. *Proceedings of the 2004 International Symposium on Active Control of Sound and Vibration*, CD-ROM (2004)
10. Hinchliffe, R.A., Scott, I.A., Purver, M.J., Stothers, I.M.: Tonal Active Control in Production on a Large Turbo-prop Aircraft. *Proceedings of ACTIVE 02, The International Symposium on Active Control of Sound and Vibration*, CD-ROM (2002)
11. Maier, R., Pucher, M., Gemblar, W., Schweitzer, H.: Helicopter Interior Noise Reduction by Active Vibration Isolation with Smart Gearbox Struts. *Proceedings of ACTIVE 99, the International Symposium on Active Control of Sound and Vibration*, CD-ROM (1999)
12. Nelson, P.A., Elliott, S.J.: Active Noise Control. *IEEE Signal Processing Magazine*, **10**(4), 12–35 (1993)
13. Petitjean, B., Greffe, C.: Active Interior Noise Control: An Industrial Perspective. *Proceedings of the SPIE - The International Society for Optical Engineering*, **4698**, 133–42 (2002)
14. Stephens, D.G., Cazier, F.W.Jr.: NASA Noise Reduction Program for Advanced Subsonic Transports. *Noise Control Eng. J.*, **44**(3), 135–40 (1996)
15. Wolter, B., Foudhaili, H., Peissig, J., Reithmeier, E.: Combined Feedback and Adaptive Feedforward Active Noise Control in Headsets. *Proc. of Internoise, the 36th Int. Congress and Exhibition on Noise Control Engineering*, Istanbul, CD-ROM (2007)

“This page left intentionally blank.”

Chapter 14

Lekhnitskii's Formalism for Stress Concentrations Around Irregularities in Anisotropic Plates: Solutions for Arbitrary Boundary Conditions

Sotiris Koussios and Adriaan Beukers

Abstract Considering analytical methods in anisotropic elasticity, the complex potentials method (as extensively formulated by Lekhnitskii) may be regarded as a powerful tool. Among the various solutions generated by this approach, the analysis of thin anisotropic plates containing a geometrically simple irregularity is the most classical one as it reflects on an extensive collection of structures: from pin-loaded holes to cutouts in aircraft fuselages. In this chapter we outline the complete solution for this particular geometry where the boundary conditions on the edge of the irregularity (forces or displacements) are formulated in Fourier series. The analytical solutions provided here can directly be evaluated as a function of the external boundary loads and the coefficients in the Fourier series, which represent the boundary conditions at the edge of the irregularity. Therefore, the analytical solutions provided here are able to cover a large variety of structural problems. Although Lekhnitskii's formalism may be regarded as a well-established solution procedure, the availability of engineering-oriented, directly implementable solutions is rather limited. In this chapter we attempt to fill this gap.

14.1 Introduction

Classical engineering problems can usually be formulated as a (partial) differential equation (PDE) with appropriate boundary conditions ensuring existence and uniqueness for the derived solution.

Sotiris Koussios

Delft University of Technology, Kluyverweg 1, 2629 HS, Delft, The Netherlands,
e-mail: s.koussios@tudelft.nl

Adriann Beukers

Delft University of Technology, Kluyverweg 1, 2629 HS, Delft, The Netherlands,
e-mail: a.beukers@tudelft.nl

As this chapter focusses on linear elasticity where there is no energy dissipation, the approach of introducing a potential function to describe stresses and deformations seems to be suitable. This principle relies on the well-known Airy functions for elastostatic boundary value problems. The most classical reference work in this field is [7] where isotropic elasticity problems are well covered. For anisotropic materials, however, the Airy function-based theory needs a slight generalization; this is extensively researched by Muskhelishvili, Novozhilov, Sokolnikoff, and Lekhnitskii. In particular Lekhnitskii has written the classical reference work for such problems [5]. More recent works are published by Ting [8] and Rand and Rovenski [6]. Without claiming completeness for the presented literature list we mention here that, next to the Airy-like approach for elastostatic problems (usually referred to as Lekhnitskii formalism) the so-called Stroh formalism (comprehensively presented in [8]) is generally considered as the second major methodology to tackle such problems. Nevertheless, this chapter only focusses on the Lekhnitskii formalism, in particular on the analysis of stress concentrations around irregularities in thin anisotropic plates.

The main goal of this chapter is to outline the derivation of the above-said theory in a more practical context and to demonstrate its application on the generic case of a flat plate containing an irregularity with arbitrary force or displacement boundary conditions. These conditions are formulated as Fourier series. However, the generality of the presented methodology is strongly limited by the geometry of the irregularity; for “mathematically non simple” shapes the setup of an analytic solution will rapidly become impossible. In addition, since we assume here that the outer boundaries of the plate are “far away,” the applicability of the obtained results for small plates is obviously questionable. Nevertheless, the derived analytic solutions do certainly provide directions for optimization activities (e.g., minimization of the maximum equivalent stress, as formulated in a strength criterion). For aerospace applications, the solutions provided here are, for example, convenient for the estimation of stress fields around cutouts, e.g., window openings and pin-loaded holes (riveting) in aircraft fuselages.

Beginning with a short outline of the governing equations, Sect. 14.2, we derive the general solution and thereof generated stress and displacement fields (Sects. 14.3 and 14.4). The boundary conditions are formulated in Sect. 14.5, which is followed by the outline of the solution strategy, Sect. 14.6. In the latter we emphasize on series representations for the potential functions and the transformation of these series into single-variable-based complex polynomials. The actual evaluation of the boundary conditions is described in Sect. 14.7, where the obtained series for the potential (Airy-like) functions is categorized and analyzed. The quantification procedure for the resulting stress and displacement fields is outlined in Sect. 14.8. Next, in Sect. 14.9 we demonstrate the application of the Lekhnitskii formalism on an anisotropic plate with an unsymmetrically loaded hole boundary. Finally, some conclusions are given in Sect. 14.10.

As most related problems do actually involve mixed boundary conditions (a part of the boundary has described forces while another part is subjected to prescribed displacements), a direct translation of the complex power series representing the stress field into the displacement-associated series (complex polynomials as well) would be extremely beneficial; such a setup is part of ongoing research.

14.2 Governing Equations

We consider here a thin anisotropic plate in plane stress situation, Fig. 14.1. The square plate contains at the origin of the reference system an irregularity, for example, a hole. The irregularity is small as compared to the outer dimensions; therefore, we assume here that the plate is infinite and, on the outer edge, a set of constant loads is considered: p_x , p_y , p_{xy} . At the same time, the edge of the irregularity might be subjected to a prescribed set of load or displacement conditions, usually represented in a Fourier series [2]. In Fig. 14.1, the arrows N and T refer to, respectively, a radial and an axial load component (stress dimensions) applied on the edge of the hole.

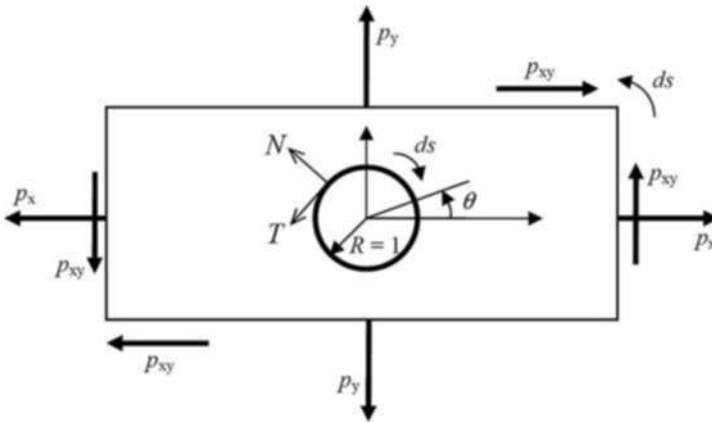


Fig. 14.1 Schematic representation of an in-plane-loaded thin rectangular laminate with a central hole

The idea is to set up the governing partial differential equation by considering the conditions for stress equilibrium and strain compatibility in plane state [7]. We assume the well-known stress equilibrium conditions in plane state [5, 7], the strain-displacements relationships, the constitutive equations

$$\begin{bmatrix} \varepsilon_x \\ \varepsilon_y \\ \gamma_{xy} \end{bmatrix} = \mathbf{C} \cdot \begin{bmatrix} \sigma_x \\ \sigma_y \\ \tau_{xy} \end{bmatrix} \quad (14.1)$$

where

$$\mathbf{C} = \begin{bmatrix} C_{11} & C_{12} & 0 \\ C_{12} & C_{22} & 0 \\ 0 & 0 & C_{66} \end{bmatrix} = \begin{bmatrix} \frac{1}{E_x} & -\frac{\nu_{xy}}{E_y} & 0 \\ -\frac{\nu_{xy}}{E_y} & \frac{1}{E_y} & 0 \\ 0 & 0 & \frac{1}{G_{xy}} \end{bmatrix}; \quad (14.2)$$

the (effective) engineering constants refer to the entire laminate and the presented quantities are deduced from a specific stacking of different anisotropic layers [1], see also Sect. 14.9.

We introduce a potential function $U(x, y)$, being such that

$$\begin{aligned}\sigma_x &= \frac{\partial^2 U(x, y)}{\partial y^2} \\ \sigma_y &= \frac{\partial^2 U(x, y)}{\partial x^2} \\ \tau_{xy} &= -\frac{\partial^2 U(x, y)}{\partial x \partial y}\end{aligned}\quad (14.3)$$

and we obtain

$$r^2 U_{xxxx} + 2a U_{xyxy} + U_{yyyy} = 0 \quad (14.4)$$

where

$$\begin{aligned}a &= \frac{C_{12} + \frac{C_{66}}{2}}{C_{11}} \\ r &= \sqrt{\frac{C_{22}}{C_{11}}}\end{aligned}\quad (14.5)$$

For isotropic materials they are both equal to 1. In this sense, a and r can be regarded as parameters indicating the anisotropy degree of the considered material. Their physical meaning can be explained by their original definition [2, 3, 5]:

$$\begin{aligned}a &= \frac{E_x}{2G_{xy}} - \nu_{xy} \\ r &= \sqrt{\frac{E_x}{E_y}}\end{aligned}\quad (14.6)$$

where E_x, E_y are the elasticity moduli in, respectively, the fiber direction and perpendicular to it, ν_{xy} the Poisson ratio (deformation in the y -direction due to a stress in the x -direction), and G_{xy} the shear modulus.

14.3 General Solution

The obtained PDE (14.4) can be decomposed in a product of linear differential operators:

$$\begin{aligned}(c^2 U_{xx} - U_{yy})(d^2 U_{xx} - U_{yy}) &= 0 \rightarrow \\ \rightarrow (cU_x + U_y)(cU_x - U_y)(dU_x + U_y)(dU_x - U_y) &= 0\end{aligned}\quad (14.7)$$

where c and d are (complex) constants. The product of these linear differential operators implies that the solution we look for must be based on a linear combination of the variables x and y . Therefore, we define

$$U(x, y) = F(z) \text{ where } z = x + sy \quad (14.8)$$

Substitution of this expression into (14.4) leads to the following characteristic equation:

$$s^4 + 2as^2 + r^2 = 0 \quad (14.9)$$

The roots are

$$\begin{aligned} s_1 &= \sqrt{\frac{r-a}{2}} + i\sqrt{\frac{r+a}{2}} \\ s_2 &= -\sqrt{\frac{r-a}{2}} + i\sqrt{\frac{r+a}{2}} \\ s_3 &= \sqrt{\frac{r-a}{2}} - i\sqrt{\frac{r+a}{2}} \\ s_4 &= -\sqrt{\frac{r-a}{2}} - i\sqrt{\frac{r+a}{2}} \end{aligned} \quad (14.10)$$

In these roots, the quantity $r + a$ is always positive [2, 3]. Depending on the laminate layup, the quantity $r - a$ can be both positive and negative; therefore, the following distinction is made:

$$\begin{aligned} r > a: \quad & \begin{aligned} s_1 &= \alpha + i\beta \\ s_2 &= \alpha - i\beta \end{aligned} \quad \text{where} \quad \begin{aligned} \alpha &= \sqrt{\frac{r-a}{2}} \\ \beta &= \sqrt{\frac{r+a}{2}} \end{aligned} \\ r < a: \quad & \begin{aligned} s_1 &= i(\gamma + \beta) \\ s_2 &= i(-\gamma + \beta) \end{aligned} \quad \text{where} \quad \gamma = \sqrt{\frac{a-r}{2}} \end{aligned} \quad (14.11)$$

The roots are apparently conjugates of each other:

$$\begin{aligned} r > a: \quad & s_1 = \bar{s}_3, \quad s_2 = \bar{s}_4 \\ r < a: \quad & s_1 = \bar{s}_4, \quad s_2 = \bar{s}_3 \end{aligned} \quad (14.12)$$

The general solution for (14.4) can now be constructed (for simplicity we neglect here the particular solution):

$$\begin{aligned} U(x, y) &= F_1(x + s_1y) + F_2(x + s_2y) + F_3(x + s_3y) + F_4(x + s_4y) \\ &= 2\Re(F_1(z_1) + F_2(z_2)) \end{aligned} \quad (14.13)$$

where $z_1 = x + s_1y$ and $z_2 = x + s_2y$. According to (14.12) we catch here the full set of solutions, regardless of the sign of $r - a$.

14.4 Stress, Strain, and Displacements Formulation

The general solution can now be substituted into the stress potentials, (14.3); the result is

$$\begin{aligned}
\sigma_x &= 2\Re(s_1^2 \Phi_1'(z_1) + s_2^2 \Phi_2'(z_2)) \\
\sigma_y &= 2\Re(\Phi_1'(z_1) + \Phi_2'(z_2)) \\
\tau_{xy} &= -2\Re(s_1 \Phi_1'(z_1) + s_2 \Phi_2'(z_2))
\end{aligned} \tag{14.14}$$

where

$$\Phi_k = \frac{dF_k}{dz_k} \quad \text{and} \quad \Phi_k' = \frac{d\Phi_k}{dz_k} \tag{14.15}$$

in which $k = 1, 2$

Substitution of these expressions into the constitutive relations (14.1) gives, after integration

$$\begin{aligned}
u &= 2\Re(u_1 \Phi_1(z_1) + u_2 \Phi_2(z_2)) + c_1 y + c_2 \\
v &= 2\Re(v_1 \Phi_1(z_1) + v_2 \Phi_2(z_2)) - c_1 x + c_3
\end{aligned} \tag{14.16}$$

in which

$$\begin{aligned}
u_k &= C_{11} s_k^2 + C_{12} \\
v_k &= \frac{C_{22}}{s_k} + C_{12} s_k
\end{aligned} \tag{14.17}$$

In addition, we define here for later use (Sect. 14.7.2)

$$w_k = \frac{C_{12}}{s_k} + C_{11} s_k \tag{14.18}$$

14.5 Formulation of Boundary Conditions

There are two kinds of boundary conditions for our problem: prescribed loads and prescribed displacements. In this section, the formulation of these conditions is briefly outlined.

14.5.1 Forces

With reference to the notations in Fig. 14.2 where the line AB represents a free edge (on the hole or the outer boundary of the plate) we have trivially

$$X ds = dy \sigma_x + dx \tau_{xy} \tag{14.19}$$

After replacing the stress components with their potentials (14.3) we obtain

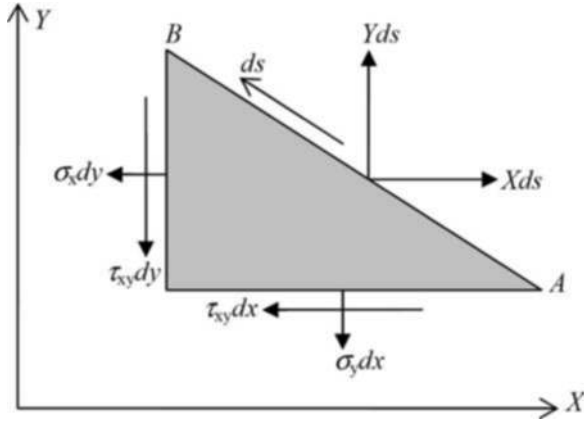


Fig. 14.2 Equilibrium of internal and external forces at a boundary

$$X = \frac{\partial U_y}{\partial y} \frac{dy}{ds} - \frac{\partial U_y}{\partial x} \frac{dx}{ds} = \frac{dU_y}{ds} \quad (14.20)$$

or, in integral form

$$U_y = 2\Re \left(\sum_{k=1}^2 s_k \Phi_k(z_k) \right) = c + \int X ds \quad (14.21)$$

where c is an arbitrary constant.

Similarly, we get

$$Y ds = dx \sigma_y + dy \tau_{xy} \quad (14.22)$$

and

$$U_x = 2\Re \left(\sum_{k=1}^2 \Phi_k(z_k) \right) = c - \int Y ds \quad (14.23)$$

With a given X, Y force distribution at the boundary (hole edge or outer edge) the conditions for further identifying the unknown Φ_k functions (14.14) are now formulated.

14.5.2 Displacements

The expressions relating the displacements at some point to the Φ_k functions (14.16) are given here in a slightly modified form where the displacements at the boundary are given as a function of the boundary contour length s (refer to Fig. 14.2 for positive ds)

$$\begin{aligned}
2\Re\left(\sum_{k=1}^2 u_k \Phi_k(z_k)\right) &= \underline{U}(s) \\
2\Re\left(\sum_{k=1}^2 v_k \Phi_k(z_k)\right) &= \underline{V}(s)
\end{aligned} \tag{14.24}$$

where $\underline{U}(s)$ and $\underline{V}(s)$ are the displacements in, respectively, the X and Y directions.

14.6 Solution Strategy

In this section, the solution procedure for the generic case of arbitrary force and displacement boundary conditions is outlined. As the unknown Φ_k functions are formulated in the z_1 and z_2 variables while the boundary conditions are given in either polar or cartesian coordinates, a procedure is here presented for converting the boundary conditions formulations and Φ_k into the same variable. The obtained representation consists of homogeneous (linear) and logarithmic terms, complemented by a power series. The latter represents the stress disturbance field due to the irregularity contained in the plate.

14.6.1 Series Representation of the Boundary Conditions

Since the complex functions as presented in (14.13) and (14.15) are analytic, it can be proven that they are expandable into Laurent series [4]; therefore, the left-hand side of the boundary conditions formulations (14.21), (14.23) and (14.24) can be represented as a power series in z_1 and z_2 :

$$\Phi_k(z_k) = \sum_{n=-\infty}^{n=+\infty} d_n^{(k)} z_k^n \tag{14.25}$$

In addition, we assume here that the right-hand side of the boundary conditions formulations is represented by a Fourier series which can be converted into a similar power series (14.27). Both sides of the boundary conditions are now expressed as power series. To satisfy the boundary conditions, one can employ the fundamental theorem of algebra:

$$\sum_{n=-\infty}^{n=+\infty} a_n x^n = \sum_{n=-\infty}^{n=+\infty} b_n x^n, \forall x \in \mathbb{C} \Leftrightarrow a_n \equiv b_n \tag{14.26}$$

where x denotes a variable (not to be confused with the x -coordinate). With this principle, the problem of determining the unknown Φ_k functions can be reduced into the identification of the (complex) constants in the corresponding series.

The series for representing the potential functions must generate limited stresses, everywhere and at the outer edge (infinity), the stress values should agree with p_x, p_y, p_{xy} , Fig. 14.1; hence, only non-positive power terms are allowed in the series $\Phi'_k(z_k)$

$$\Phi'_k(z_k) = \sum_{n=-\infty}^{n=0} g_n^{(k)} z_k^n$$

In regard to the boundary conditions at the edge of the hole, the following representation in Fourier series is here assumed for the introduced loads (see also Fig. 14.1). The series is truncated at m :

radial loads:

$$N_s(\Theta, m) = N_1 + \sum_{n=1}^m \left(\cos(n\Theta) N_{n+1}^{(c)} + \sin(n\Theta) N_{n+1}^{(s)} \right)$$

tangential loads:

$$T_s(\Theta, m) = T_1 + \sum_{n=1}^m \left(\cos(n\Theta) T_{n+1}^{(c)} + \sin(n\Theta) T_{n+1}^{(s)} \right) \quad (14.27)$$

where $N_n^{(c,s)}$ and $T_n^{(c,s)}$ are real constants and $0 \leq \Theta \leq 2\pi$. A similar formulation can be employed for the representation of prescribed displacements. However, since the plate is infinite, it is convenient to assume a constant load field on the outer edges as given in Fig. 14.1 (instead of displacements). Periodic load distributions are possible as well [9].

14.6.2 Transformation into a Single Variable

As previously mentioned, the coefficients of the power series are to be determined by the evaluation of the boundary conditions on the outer edge of the plate and the hole itself. For the outer edge such an evaluation is rather easy to perform since the plate is assumed to have straight edges at infinity. For the hole edge, however, we might face some problems:

- The hole or irregularity can have different shapes. For simplicity we assume here elliptical and quasi-rectangular shapes, which are typical in aerospace structures.
- The series for Φ_k contains different variables (z_1 and z_2); hence a transformation to a single one is highly desired.

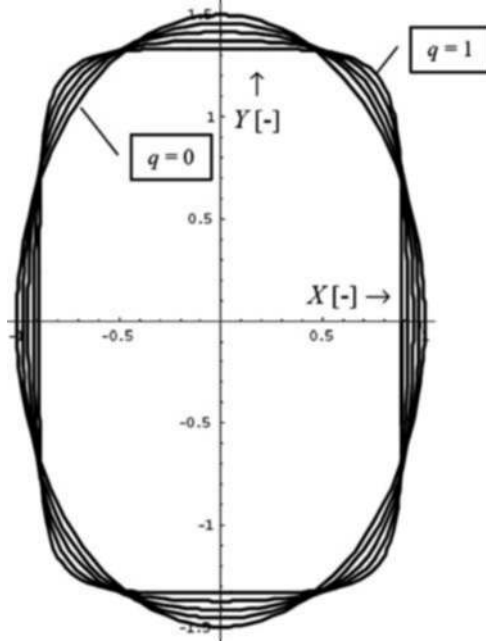


Fig. 14.3 Various hole shapes as generated by (14.28) ($\Xi = 1, H = 3/2$)

The following equation can represent the desired family of shapes:

$$\begin{aligned} x &= \Xi \left(\cos(\Theta) - \frac{1}{9}q \cos(3\Theta) \right) \\ y &= H \left(\sin(\Theta) + \frac{1}{9}q \sin(3\Theta) \right) \end{aligned} \quad (14.28)$$

For $q = 0$ we obtain an elliptical shape with a horizontal major axis Ξ and a vertical major axis H . For $q = 1$ we get the most rectangular hole still being convex everywhere (no re-entrant curvature), Fig. 14.3.

With the unity circle of complex numbers $z = \cos \Theta + i \sin \Theta$ in mind, the trigonometric terms can be written as

$$\begin{aligned} \cos(n\Theta) &= \frac{1}{2} (z^{-n} + z^n) \\ \sin(n\Theta) &= \frac{z^n - z^{-n}}{2i} \end{aligned} \quad (14.29)$$

With this, the variable z_k becomes

$$\begin{aligned} z_k &= \Xi \left(\cos \left(\frac{1}{2} \left(z + \frac{1}{z} \right) \right) - \frac{1}{9}q \cos \left(\frac{1}{2} \left(z^3 + \frac{1}{z^3} \right) \right) \right) \\ &\quad + s_k H \left(\frac{1}{9}q \sin \left(\frac{z^3 - \frac{1}{z^3}}{2i} \right) + \sin \left(\frac{z - \frac{1}{z}}{2i} \right) \right) \end{aligned} \quad (14.30)$$

The next step for the conversion of $\Phi_k(z_k)$ to $\Phi_k(z)$ is to express the unity circle z in terms of z_k . This implies the solution of a sixth-degree polynomial and will provide six roots. Since this solution can only be obtained by numerical techniques, we will limit the problem here to the class of ellipsoidal holes ($q = 0$). In this case, a second-degree equation must be solved, leading to

$$\zeta_k = \frac{z_k \pm \sqrt{-\Xi^2 - s_k^2 H^2 + z_k^2}}{\Xi - i s_k H} \quad (14.31)$$

For $x \rightarrow \cos \Theta$ and $y \rightarrow \sin \Theta$, the variables ζ_k reduce to $z = \cos \Theta + i \sin \Theta$. In regard to the right-hand side of the load boundary conditions (14.21) and (14.23), the Fourier series as given in (14.27) can directly be expressed in terms of z with (14.29).

14.7 Boundary Conditions Evaluation

From the series representation for Φ'_k (14.27) we obtain after integration

$$\Phi_k(z_k) = h_k z_k + A_k \ln(z_k) + \sum_{n=1}^{n=+\infty} g_n^{(k)} z_k^{-n} \quad (14.32)$$

This result contains three kinds of terms:

- h_k : these constants represent the homogeneous stress field (14.14) for the undisturbed plate
- A_k : these special terms represent the load resultants in X and Y directions (to be dequantified after integration of the load boundary conditions (14.21) and (14.23) around the complete contour of the hole)
- $g_n^{(k)}$: these higher order terms represent the disturbance field due to the hole presence

In the next subsections, the outlined cases are treated separately.

14.7.1 Homogeneous Part

Let the unknown h_k constants be represented by

$$h_k = (ib_k + d_k)(x + s_k y) \quad (14.33)$$

As z_k approaches infinity, only the linear terms of $\Phi_k(z_k)$ will provide a non-zero value for the stress field. In addition, this stress field should be constant throughout the plate (note that the stresses depend on the differentiated form of (14.32)

as given in (14.14). The outer edge loads are depicted in Fig. 14.1. To determine the unknowns d_k and b_k we need essentially four equations. These are provided by the force boundary conditions (14.21) and (14.23). However, this system provides only three equations since one of them is identical. Therefore, a fourth condition is needed. For an unambiguous determination of the searched constants, one should restrict rotation of the plate; this restriction leads to the fourth equation we need. The system that has to be solved becomes [2]

$$\begin{aligned} 2\Re \left(\sum_{k=1}^2 s_k h_k(z_k) \right) &= p_{xy}x - p_{xy}y \\ 2\Re \left(\sum_{k=1}^2 h_k(z_k) \right) &= p_{xy}y - p_{xy}x \\ \frac{\partial V(s)}{\partial x} - \frac{\partial U(s)}{\partial y} &= 0 \end{aligned} \quad (14.34)$$

With the aid of (14.11), the solution finally gets the form (irrespective of the sign for $r - a$)

$$h_k = \frac{p_x - p_y s_l^2}{2(s_k^2 - s_l^2)} - \frac{p_{xy}}{4s_k} \quad (14.35)$$

where $l = 3 - k$ and $k = 1, 2$.

14.7.2 Logarithmic Part

The calculation of the A_k constants relies on the evaluation of the boundary conditions (14.21) and (14.23) around the complete hole contour. During this operation, the other terms of (14.32) will give zero anyway (Cauchy Integral Theorem). Evaluation of the boundary conditions around the complete hole contour gives

$$\begin{aligned} \left(2\Re \left(\sum_{k=1}^2 s_k \left(h_k z_k + A_k \ln(z_k) + \sum_{n=1}^{n=+\infty} g_n^{(k)} z_k^{-n} \right) \right) \right)_0^{2\pi} &= \\ = c + \int_0^{2\pi} X ds = c - \int_0^{2\pi} X d\Theta = -R_x \\ \left(2\Re \left(\sum_{k=1}^2 \left(h_k z_k + A_k \ln(z_k) + \sum_{n=1}^{n=+\infty} g_n^{(k)} z_k^{-n} \right) \right) \right)_0^{2\pi} &= \\ = c - \int_0^{2\pi} Y ds = c + \int_0^{2\pi} Y d\Theta = R_y \end{aligned} \quad (14.36)$$

Note: $Rd\Theta = -ds$ with $R = 1$, Fig. 14.1. The variables z_k are now replaced by ζ_k (14.31). After expansion in series, the logarithmic part becomes

$$A_k \ln(z_k) = A_k \left(\ln(c) + \ln(\zeta_k) + \sum_{n=1}^{\infty} c_n \zeta_k^{-2n} \right) \quad (14.37)$$

On the edge of the hole contour, the variables ζ_k approach z . Within the limits $[0, 2\pi]$ this gives

$$\left(\ln(c) + \ln(z) + \sum_{n=1}^{\infty} c_n z^{-2n} \right)_0^{2\pi} = 2j\pi + i\Theta \text{ where } j \in \mathbb{N} \quad (14.38)$$

With this result we can now evaluate the summations in (14.36). However, as experienced for the calculation of the h_k terms (Sect. 14.7.1) these four equations contain an identical pair, hence a new condition is needed. The condition we are looking for relies on a physical explanation; as the edge of the hole deforms, the material should remain together. We assume that there are no cracks, discontinuities, or dislocations. In mathematical terms, the contour integrals of the displacements around the hole edge should provide zero values:

$$\begin{aligned} \int_0^{2\pi} \underline{U}(s) ds &= 0 \\ \int_0^{2\pi} \underline{V}(s) ds &= 0 \end{aligned} \quad (14.39)$$

With the aid of the displacement boundary conditions (14.24) and (14.36) and (14.39) we finally obtain

$$\begin{aligned} A_1 + A_2 - \bar{A}_1 - \bar{A}_2 &= \frac{R_y}{2\pi i} \\ A_1 s_1 + A_2 s_2 - \bar{s}_1 \bar{A}_1 - \bar{s}_2 \bar{A}_2 &= -\frac{R_x}{2\pi i} \\ A_1 u_1 + A_2 u_2 - \bar{u}_1 \bar{A}_1 - \bar{u}_2 \bar{A}_2 &= 0 \\ A_1 v_1 + A_2 v_2 - \bar{v}_1 \bar{A}_1 - \bar{v}_2 \bar{A}_2 &= 0 \end{aligned} \quad (14.40)$$

The solution becomes

$$A_k = \frac{i(R_y u_l + R_x w_k)}{(4\pi C_{11})(s_k^2 - s_l^2)}, \text{ where } l = 3 - k \quad (14.41)$$

in which R_x and R_y are the load resultants (14.36), C_{11} a compliance element (14.2) whereas w_k is defined in (14.18). The complex roots s_k, s_l are given in (14.10). Note that the constants A_k are multivalued (parameter j in (14.38)). However, the most important observation here is the fact that the logarithmic terms become only active in the case where the loads, applied on the edge of the hole, do have at least one non-zero resultant.

14.7.3 Disturbance Field

The departure points for the evaluation of the remaining terms $g_n^{(k)}$ are the load boundary conditions as formulated in (14.21) and (14.23). With the expressions for h_k (14.35), (14.36) and the relation $\log z = i\Theta$, we formulate

$$\begin{aligned} 2\Re \left(\sum_{k=1}^2 s_k \Phi_k^\circ(z_k) \right) &= \int_0^s X_n ds - p_x y + x p_{xy} + \frac{R_x \Theta}{2\pi} = f_x \\ 2\Re \left(\sum_{k=1}^2 \Phi_k^\circ(z_k) \right) &= - \int_0^s Y_n ds - p_y x + y p_{xy} - \frac{R_y \Theta}{2\pi} = f_y \end{aligned} \quad (14.42)$$

where $\Phi_k^\circ = \sum_{n=1}^{n=\infty} g_n^{(k)} z_k^{-n}$ stands for the part of the potential function Φ_k that corresponds to the power terms only (last term in (14.32)).

As the boundary conditions reflect on forces in, respectively, the X and Y directions, the Fourier series representing the radial and tangential loads (14.27) must accordingly be decomposed. The result is

$$\begin{aligned} X_s(\Theta, m) &= N_s(\Theta, m) \cos(\Theta) - T_s(\Theta, m) \sin(\Theta) \\ Y_s(\Theta, m) &= N_s(\Theta, m) \sin(\Theta) + T_s(\Theta, m) \cos(\Theta) \end{aligned} \quad (14.43)$$

After integration around the complete hole contour $[0, \pi/2]$, the resultant forces in, respectively, the X and Y directions become

$$\begin{aligned} R_x &= \pi \left(N_2^{(c)} - T_2^{(s)} \right) \\ R_y &= \pi \left(N_2^{(s)} + T_2^{(c)} \right) \end{aligned} \quad (14.44)$$

With the relations given in (14.43), the complete expression for f_x (as derived in (14.36)) can now be formulated as

$$\begin{aligned}
f_x(\Theta, m) = & -\sin(\Theta) \left(N_1 + p_x + \frac{1}{2} \left(N_3^{(c)} - T_3^{(s)} \right) \right) \\
& + \cos(\Theta) \left(p_{xy} - T_1 + \frac{1}{2} \left(T_3^{(c)} + N_3^{(s)} \right) \right) \\
& + \left(\sum_{n=2}^{m-1} \frac{\left(T_{n+2}^{(c)} - T_n^{(c)} + N_{n+2}^{(s)} + N_n^{(s)} \right) \cos(n\Theta)}{2n} \right) \\
& - \left(\sum_{n=2}^{m-1} \frac{\left(N_{n+2}^{(c)} + N_n^{(c)} - T_{n+2}^{(s)} + T_n^{(s)} \right) \sin(n\Theta)}{2n} \right) \\
& + \left(\sum_{n=m}^{m+1} \frac{\left(N_n^{(s)} - T_n^{(c)} \right) \cos(n\Theta)}{2n} \right) \\
& - \left(\sum_{n=m}^{m+1} \frac{\left(N_n^{(c)} + T_n^{(s)} \right) \sin(n\Theta)}{2n} \right)
\end{aligned} \tag{14.45}$$

Similarly, for f_y we obtain

$$\begin{aligned}
f_y(\Theta, m) = & +\sin(\Theta) \left(p_{xy} + T_1 + \frac{1}{2} \left(T_3^{(c)} + N_3^{(s)} \right) \right) \\
& + \cos(\Theta) \left(-N_1 - p_y + \frac{1}{2} \left(N_3^{(c)} - T_3^{(s)} \right) \right) \\
& - \left(\sum_{n=2}^{m-1} \frac{\left(-N_{n+2}^{(c)} + N_n^{(c)} + T_{n+2}^{(s)} + T_n^{(s)} \right) \cos(n\Theta)}{2n} \right) \\
& + \left(\sum_{n=2}^{m-1} \frac{\left(-T_{n+2}^{(c)} - T_n^{(c)} - N_{n+2}^{(s)} + N_n^{(s)} \right) \sin(n\Theta)}{2n} \right) \\
& - \left(\sum_{n=m}^{m+1} \frac{\left(N_n^{(c)} + T_n^{(s)} \right) \cos(n\Theta)}{2n} \right) \\
& + \left(\sum_{n=m}^{m+1} \frac{\left(N_n^{(s)} - T_n^{(c)} \right) \sin(n\Theta)}{2n} \right)
\end{aligned} \tag{14.46}$$

Note that the terms $\frac{R_y \Theta}{2\pi}$ and $\frac{R_y \Theta}{2\pi}$ automatically vanish as we evaluate the integrals in (14.42); the problem of the multivalued integral expression in (14.38) is hereby omitted. With (14.29), the expressions for f_x and f_y can be written in the following form:

$$\begin{aligned}
 f_x(\Theta, m) &= \sum_{n=1}^{m+1} P(n)z^{-n} + \sum_{n=1}^{m+1} \bar{P}(n)z^n \\
 f_y(\Theta, m) &= \sum_{n=1}^{m+1} Q(n)z^{-n} + \sum_{n=1}^{m+1} \bar{Q}(n)z^n
 \end{aligned} \tag{14.47}$$

where

$$\begin{aligned}
 n = 1 : \\
 P(n) &= \frac{\left(-iN_3^{(c)} + N_3^{(s)} + T_3^{(c)} + iT_3^{(s)} - 2iN_1 - 2ip_x + 2p_{xy} - 2T_1\right)}{4} \\
 2 \leq n \leq m-1 : \\
 P(n) &= \frac{\left(-T_n^{(c)} + T_{n+2}^{(c)} + N_n^{(s)} + N_{n+2}^{(s)}\right) - i\left(N_n^{(c)} + N_{n+2}^{(c)} + T_n^{(s)} - T_{n+2}^{(s)}\right)}{4n} \\
 m \leq n \leq m+1 : \\
 P(n) &= -\frac{iN_n^{(c)} - N_n^{(s)} + T_n^{(c)} + iT_n^{(s)}}{4n}
 \end{aligned} \tag{14.48}$$

and

$$\begin{aligned}
 n = 1 : \\
 Q(n) &= \frac{\left(N_3^{(c)} + iN_3^{(s)} + iT_3^{(c)} - T_3^{(s)} - 2N_1 + 2ip_{xy} - 2p_y + 2iT_1\right)}{4} \\
 2 \leq n \leq m-1 : \\
 Q(n) &= \frac{-i\left(-T_n^c - T_{n+2}^c + N_n^s - N_{n+2}^s\right) - \left(N_n^c - N_{n+2}^c + T_n^s + T_{n+2}^s\right)}{4n} \\
 m \leq n \leq m+1 : \\
 Q(n) &= -\frac{N_n^c + iN_n^s - iT_n^c + T_n^s}{4n}
 \end{aligned} \tag{14.49}$$

According to (14.42) (this time in series form) the following conditions must be satisfied:

$$\begin{aligned}
 f_x &= 2\Re\left(\sum_{k=1}^2 \sum_{n=1}^{m+1} s_k g_n^{(k)} z_k^{-n}\right) \\
 f_y &= 2\Re\left(\sum_{k=1}^2 \sum_{n=1}^{m+1} g_n^{(k)} z_k^{-n}\right)
 \end{aligned} \tag{14.50}$$

The f_x and f_y functions are expressed in z , hence we have to transform the related series into the same variable. With the aid of (14.31) a new series is here set up with $c_n^{(k)}$ as the unknown coefficients. For x, y approaching the hole edge we obtain $\zeta_k \rightarrow z$. Expansion of (14.50) into conjugate parts yields

$$\begin{aligned} 2\Re \left(\sum_{n=1}^{m+1} s_k g_n^{(k)} z_k^{-n} \right) &= 2\Re \left(\sum_{n=1}^{m+1} s_k c_n^{(k)} z^{-n} \right) = \sum_{n=1}^{m+1} s_k c_n^{(k)} z^{-n} + \sum_{n=1}^{m+1} \bar{s}_k \bar{c}_n^{(k)} z^{+n} \\ 2\Re \left(\sum_{n=1}^{m+1} g_n^{(k)} z_k^{-n} \right) &= 2\Re \left(\sum_{n=1}^{m+1} c_n^{(k)} z^{-n} \right) = \sum_{n=1}^{m+1} c_n^{(k)} z^{-n} + \sum_{n=1}^{m+1} \bar{c}_n^{(k)} z^{+n} \end{aligned} \quad (14.51)$$

Setting the obtained expressions equal to (14.47) gives, for every positive $n \in \mathbb{N}$, the following system of equations:

$$\begin{bmatrix} s_1 & s_2 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} c_n^1 \\ c_n^2 \end{bmatrix} = \begin{bmatrix} P(n) \\ Q(n) \end{bmatrix} \quad (14.52)$$

The solution for the unknown $c_n^{(k)}$ coefficients becomes

$$c_n^{(k)} = \frac{P(n) - Q(n)s_l}{s_k - s_l} \quad (14.53)$$

where $l = 3 - k$ and $k = 1, 2$ and $n = 1, 2, 3, \dots, m+1$.

14.8 Evaluation of Stresses and Displacements

With the expressions for h_k, A_k and $c_n^{(k)}$, the solution for $\Phi_k(z_k)$ attains the following form:

$$\Phi_k(z_k) = h_k z_k + A_k \ln(\zeta_k) + \sum_{n=1}^{m+1} c_n^{(k)} \zeta_k^{-n} \quad (14.54)$$

For the determination of the stresses (14.14) we have to differentiate $\Phi_k(z_k)$:

$$\Phi'_k(z_k) = h_k + A_k \frac{d \ln(\zeta_k)}{dz_k} + \sum_{n=1}^{m+1} c_n^{(k)} \frac{d \zeta_k^{-n}}{dz_k} \quad (14.55)$$

To simplify the differentiations we introduce

$$\begin{aligned} \Gamma_k &= \Xi + H i s_k \\ \Delta_k &= \sqrt{-\Xi^2 - s_k^2 H^2 + z_k^2} \end{aligned} \quad (14.56)$$

The variable ζ_k^{-n} can now be expressed as follows:

$$\zeta_k^{-n} = \left(\frac{z_k \pm \Delta_k}{\Gamma_k} \right)^n \quad (14.57)$$

The ζ_k related derivatives, as present in (14.55), become

$$\begin{aligned} \frac{d \ln(\zeta_k)}{dz_k} &= \frac{1}{\pm \Delta_k} \\ \frac{d \zeta_k^{-n}}{dz_k} &= \frac{n \left(\frac{z_k \pm \Delta_k}{\Gamma_k} \right)^n}{\pm \Delta_k} \end{aligned} \quad (14.58)$$

where the \pm originates from (14.31). The differentiated form of (14.54) gets finally the following form:

$$\Phi'_k = h_k + \frac{A_k}{\pm \Delta_k} + \sum_{n=1}^{m+1} \frac{c_n^{(k)} n \left(\frac{z_k \pm \Delta_k}{\Gamma_k} \right)^n}{\pm \Delta_k} \quad (14.59)$$

By substituting this expression into (14.14) we can evaluate the stresses over the entire plate domain:

$$\begin{aligned} \sigma_x &= 2\Re \left(\sum_{k=1}^2 s_k^2 \left(h_k + \frac{A_k}{\pm \Delta_k} + \sum_{n=1}^{m+1} \frac{c_n^{(k)} n \left(\frac{z_k \pm \Delta_k}{\Gamma_k} \right)^n}{\pm \Delta_k} \right) \right) \\ \sigma_y &= 2\Re \left(\sum_{k=1}^2 \left(h_k + \frac{A_k}{\pm \Delta_k} + \sum_{n=1}^{m+1} \frac{c_n^{(k)} n \left(\frac{z_k \pm \Delta_k}{\Gamma_k} \right)^n}{\pm \Delta_k} \right) \right) \\ \tau_{xy} &= -2\Re \left(\sum_{k=1}^2 s_k \left(h_k + \frac{A_k}{\pm \Delta_k} + \sum_{n=1}^{m+1} \frac{c_n^{(k)} n \left(\frac{z_k \pm \Delta_k}{\Gamma_k} \right)^n}{\pm \Delta_k} \right) \right) \end{aligned} \quad (14.60)$$

The stresses as formulated above refer to the x, y coordinate system as depicted in Fig. 14.1. To transform them in a polar system (to evaluate the radial and tangential stresses around, e.g., a circular hole) we apply the following transformation [1]:

$$\begin{bmatrix} \sigma_\rho \\ \sigma_\theta \\ \tau_{\rho\theta} \end{bmatrix} = \begin{bmatrix} \cos^2(\theta) & \sin^2(\theta) & 2\cos(\theta)\sin(\theta) \\ \sin^2(\theta) & \cos^2(\theta) & -2\cos(\theta)\sin(\theta) \\ -\cos(\theta)\sin(\theta) & \cos(\theta)\sin(\theta) & \cos^2(\theta) - \sin^2(\theta) \end{bmatrix} \cdot \begin{bmatrix} \sigma_x \\ \sigma_y \\ \tau_{xy} \end{bmatrix} \quad (14.61)$$

where θ is the angle $\sigma_\rho \angle \sigma_x$. Finally, according to (14.16), the displacements are

$$\begin{aligned} \underline{U} &= 2\Re \left(\sum_{k=1}^2 u_k \left(h_k z_k + A_k \ln(\zeta_k) + \sum_{n=1}^{m+1} c_n^{(k)} \left(\frac{z_k \pm \Delta_k}{\Gamma_k} \right)^n \right) \right) \\ \underline{V} &= 2\Re \left(\sum_{k=1}^2 v_k \left(h_k z_k + A_k \ln(\zeta_k) + \sum_{n=1}^{m+1} c_n^{(k)} \left(\frac{z_k \pm \Delta_k}{\Gamma_k} \right)^n \right) \right) \end{aligned} \quad (14.62)$$

An important aspect for the correct evaluation of the above expressions is the proper selection of the + or – sign. The corresponding computational procedure is explained in [2, 9, 10]. However, the proposed algorithms do not always lead to the correct selection, particularly in cases where $r - a > 0$. A publication outlining an improved sign selection algorithm is currently under preparation.

14.9 Example

We consider here a rectangular plate with a circular hole at the origin of the X, Y coordinate system, Fig. 14.1. The loads on the outer boundary are zero ($p_x = 0, p_y = 0, p_{xy} = 0$). The edge of the hole is loaded with the following radial and tangential components:

$$\begin{aligned} N(\Theta) &= N_a \sin^2(\Theta) & \text{for } 0 \leq \Theta \leq \pi \\ N(\Theta) &= 0 & \text{for } \pi < \Theta \leq 2\pi \\ T(\Theta) &= 0 & \text{for } 0 \leq \Theta \leq 2\pi \end{aligned} \quad (14.63)$$

where N_a is a real constant representing the peak load. The load situation is given in Fig. 14.4 where we depict the radial load according to (14.63) and its series approximation with $m = 10$, (14.27); the two graphs coincide (coincidence not visible).

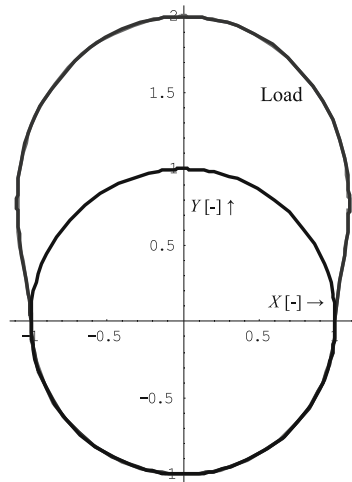


Fig. 14.4 The assumed radial loading on the contour of the hole (normalized)

In regard to the material properties, the plate consists here of a four-layer laminate (Table 14.1).

The engineering constants of a single layer are given in Table 14.2. The stiffness matrix of a single layer with respect to the so-called material coordinate system (the x -axis is aligned with the fiber direction) is given by [1]

Table 14.1 Laminate stacking

Layer number, p	Relative thickness, t	Fiber orientation, ϕ
1	0.4	$\pi/2$
2	0.2	0
3	0.2	$\pi/4$
4	0.2	$-\pi/4$

Table 14.2 Engineering constants of the basic layer

Name	Quantity	Dimension
E_x	145	GPa
E_y	7	GPa
G_{xy}	3.5	GPa
ν_{xy}	0.34	—

$$\mathbf{S} = \begin{pmatrix} \frac{E_x}{1 - \nu_{xy}\nu_{yx}} & \frac{\nu_{xy}E_y}{1 - \nu_{xy}\nu_{yx}} & 0 \\ \frac{\nu_{xy}E_y}{1 - \nu_{xy}\nu_{yx}} & \frac{E_y}{1 - \nu_{xy}\nu_{yx}} & 0 \\ 0 & 0 & G_{xy} \end{pmatrix} \quad (14.64)$$

According to [1], the compliance matrix of the entire laminate can be calculated as follows:

$$\mathbf{C}_{\text{laminate}} = \frac{1}{\sum_{p=1}^q t(p)} \left(\sum_{p=1}^q (\mathbf{M}(\phi(p)) \cdot \mathbf{S} \cdot (\mathbf{M}(\phi(p)))^T) \right)^{-1} \quad (14.65)$$

where q is the number of layers, $\phi(p)$ the fiber orientation angle of an individual layer p , and \mathbf{M} the stress transformation matrix as given in (14.61). The result is a 3×3 matrix, similar to (14.2).

With these compliance elements, the quantities a , r , s_k , u_k , ν_k , and w_k can now be determined. In addition, the given load distribution (14.63) is converted into a Fourier series [4]. With the external load vector (p_x, p_y, p_{xy}) and the Fourier coefficients N_n and T_n , the parameters $P(n)$ and $Q(n)$ can now be calculated. Next, for the determination of the $\Phi_k(z_k)$ series, the h_k , A_k , and $c_n^{(k)}$ coefficients have to be quantified. After this step, with (14.60), (14.61) and (14.62), we are able to provide the exact values for the stress and displacements field throughout the entire plate surface.

Some values for the stress field σ_y over the plate surface are given below in array format, where $-4 \leq x \leq 4$ (length) and $-4 \leq y \leq 4$ (height). The mesh size for these data is 0.5 and $N_a = 100$ [MPa]. The zeros reflect on the hole area, and the provided numbers are rounded.

$$\begin{pmatrix}
 -12 & -18 & -27 & -41 & -62 & -93 & -130 & -164 & -178 & -164 & -130 & -93 & -62 & -41 & -27 & -18 & -12 \\
 -7 & -12 & -20 & -33 & -56 & -93 & -142 & -191 & -212 & -191 & -142 & -93 & -56 & -33 & -20 & -12 & -7 \\
 -2 & -5 & -11 & -23 & -45 & -87 & -153 & -227 & -261 & -227 & -153 & -87 & -45 & -23 & -11 & -5 & -2 \\
 3 & 2 & -1 & -9 & -28 & -72 & -160 & -275 & -332 & -275 & -160 & -72 & -28 & -9 & -1 & 2 & 3 \\
 7 & 8 & 9 & 7 & -3 & -42 & -152 & -339 & -444 & -339 & -152 & -42 & -3 & 7 & 9 & 8 & 7 \\
 11 & 14 & 18 & 23 & 27 & 11 & -107 & -425 & -631 & -425 & -107 & 11 & 27 & 23 & 18 & 14 & 11 \\
 13 & 17 & 24 & 36 & 55 & 84 & 27 & -518 & -999 & -518 & 27 & 84 & 55 & 36 & 24 & 17 & 13 \\
 15 & 20 & 27 & 41 & 69 & 138 & 307 & 0 & 0 & 0 & 307 & 138 & 69 & 41 & 27 & 20 & 15 \\
 16 & 21 & 30 & 44 & 74 & 152 & 618 & 0 & 0 & 0 & 618 & 152 & 74 & 44 & 30 & 21 & 16 \\
 18 & 24 & 34 & 50 & 82 & 151 & 229 & 0 & 0 & 0 & 229 & 151 & 82 & 50 & 34 & 24 & 18 \\
 21 & 28 & 39 & 57 & 86 & 131 & 134 & 19 & -1 & 19 & 134 & 131 & 86 & 57 & 39 & 28 & 21 \\
 24 & 32 & 43 & 60 & 83 & 108 & 99 & 41 & 13 & 41 & 99 & 108 & 83 & 60 & 43 & 32 & 24 \\
 27 & 35 & 45 & 60 & 77 & 89 & 80 & 49 & 32 & 49 & 80 & 89 & 77 & 60 & 45 & 35 & 27 \\
 29 & 36 & 46 & 58 & 69 & 76 & 69 & 52 & 43 & 52 & 69 & 76 & 69 & 58 & 46 & 36 & 29 \\
 30 & 37 & 45 & 55 & 63 & 67 & 63 & 53 & 48 & 53 & 63 & 67 & 63 & 55 & 45 & 37 & 30 \\
 31 & 37 & 44 & 52 & 58 & 61 & 58 & 53 & 51 & 53 & 58 & 61 & 58 & 52 & 44 & 37 & 31 \\
 31 & 37 & 43 & 49 & 54 & 56 & 55 & 52 & 51 & 52 & 55 & 56 & 54 & 49 & 43 & 37 & 31
 \end{pmatrix} \quad (14.66)$$

The same σ_y stress field is depicted in Fig. 14.5 (mesh size = 0.1, range = 12×12). For the quantification of the stresses, one can refer to the above given array with stress values (14.66). The original and deformed hole shape is given in Fig. 14.6.

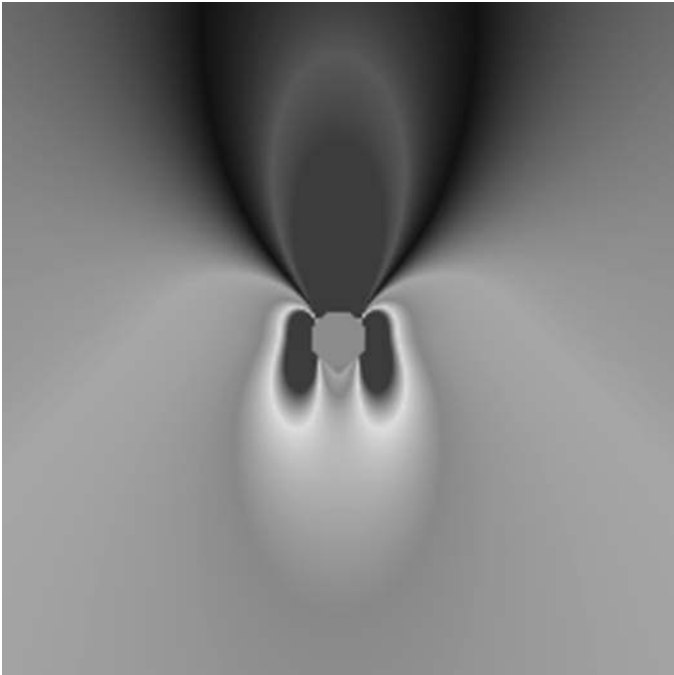


Fig. 14.5 The resulting $\sigma_y(x,y)$ stress field

To validate the results, we have calculated the radial stress at the edge of the hole $\sigma_\rho(\Theta, \rho)$ with $\rho = 1$ and $0 \leq \Theta \leq 2\pi$ according to (14.61). This distribution turned out to be equal to the original radial stress field, is given in (14.63) and depicted in Fig. 14.4; hence the load BC is identically satisfied.

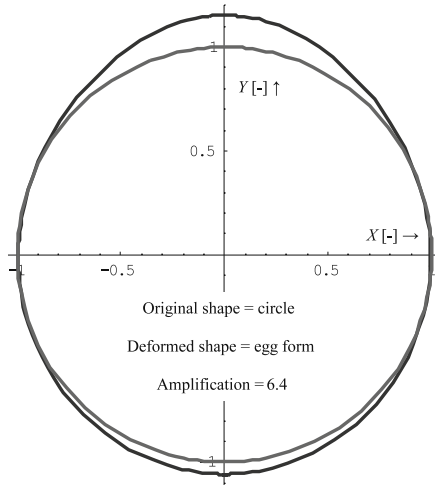


Fig. 14.6 The original and deformed hole shape, after application of the radial loading

14.10 Conclusions

In this chapter we have presented the derivation and application of the so-called Lekhnitskii formalism, for the case of thin anisotropic plates that contain geometrically simple irregularities. The main goal of this chapter is to provide a comprehensive outline of this theory in an engineering context and to show how this is applied.

After a short explanation of the governing equations, the general solutions have been derived with the generated stress and displacement field formulations. The boundary conditions are represented in series transformed into single-variable-based complex polynomials.

The generic formulation for the boundary conditions in Fourier series enables the solution of a wide range of problems. However, important conditions for obtaining analytical solutions are the geometrical simplicity of the contained irregularity, periodicity of the external loads (due to infinite plate dimensions), and the fact that the mechanical properties of the laminate should be constant throughout its surface. In addition, the load-deformation behavior is here assumed as linear (small deformation approach). Despite these limitations, however, the Lekhnitskii formalism does still provide a powerful tool for estimating stress and displacement fields (ideal for initial design procedures) and for predicting design modifications necessary to satisfy certain optimality conditions for the structure under consideration. In addition, the analytical formulation of the resulting stress and displacement fields in similar complex power series provides an opportunity for direct translation of these conditions into each other. This will certainly facilitate the solution of mixed boundary condition problems (combination of prescribed loads and displacements on, e.g., the hole boundary) and therefore is part of ongoing research.

References

1. Daniel, I.M., Ishai, O. *Engineering Mechanics of Composite Materials*. Second edition. Oxford University Press, New York (2006).
2. Jong, de, Th. *On the Calculation of Stresses in Pin-Loaded Anisotropic Plates*. PhD thesis. Faculty of Aerospace Engineering, Delft University of Technology, Delft (1987).
3. Koussios S. *Stress Concentrations around Holes*. Lecture Notes. Faculty of Aerospace Engineering, Delft University of Technology, Delft (2007).
4. Kreyszig, E. *Advanced Engineering Mathematics*. John Wiley & Sons Inc., New York (1999).
5. Lekhnitskii, S.G. *Anisotropic Plates*. Gordon and Bleach Science, New York (1968).
6. Rand, O., Rovenski, V. *Analytical Methods in Anisotropic Elasticity with Symbolic Computational Tools*. Birkhäuser, Boston (2005).
7. Timoshenko S.P., Goodier, J.N. *Theory of Elasticity*. McGraw-Hill Publishing Company, New York (1987).
8. Ting, T.C.T. *Anisotropic Elasticity: Theory and Applications*. Oxford University Press, New York (1996).
9. Tooren, van, M.J.L. *Sandwich Fuselage Design*. PhD thesis. Faculty of Aerospace Engineering, Delft University of Technology, Delft (1998).
10. Vuil, H.A. *On the Continuity of ζ_1 and ζ_2 in Lekhnitskii's Theory of Anisotropic Plates*. Internal Report. Faculty of Aerospace Engineering, Delft University of Technology, Delft (1981).

“This page left intentionally blank.”

Chapter 15

Best Initial Conditions for the Rendezvous Maneuver

Angelo Miele and Marco Ciarcià

Abstract The majority of the papers dedicated to the rendezvous problem have employed the assumption of given initial position and given initial velocity of the chaser spacecraft vis-à-vis the target spacecraft. In this research, the initial separation velocity components are assumed free and the initial separation coordinates are subject to only the requirement that the chaser-to-target distance is given. Within this frame, two problems are studied: time-to-rendezvous free and time-to-rendezvous given. It is assumed that the target spacecraft moves along a circular orbit, that the chaser spacecraft has variable mass, and that its trajectory is governed by three controls, one determining the thrust magnitude and two determining the thrust direction. Analyses performed with the multiple-subarc sequential gradient-restoration algorithm for optimal control problems show that the fuel-optimal trajectory is zero-bang; namely, it includes a long coasting zero-thrust subarc followed by a short powered max-thrust braking subarc. While the thrust direction of the powered subarc is continuously variable for the optimal trajectory, its replacement with a constant (yet optimized) thrust direction produces a very efficient guidance trajectory.

The optimization of the initial separation coordinates and velocities as well as the time lengths of all the subarcs is performed for several values of the initial distance in the range $5 \leq d_0 \leq 60$ km with particular reference to the rendezvous between the Space Shuttle (SS) and the International Space Station (ISS). This study is of interest because, for a preselected initial distance SS-to-ISS, it supplies not only the best initial conditions for the rendezvous maneuver, but also the corresponding final conditions of the ascent trajectory.

Angelo Miele

Research Professor and Foyt Professor Emeritus, Aero-Astronautics Group, Rice University, Houston, TX, USA e-mail: miele@rice.edu

Marco Ciarcià

PhD Candidate, Aero-Astronautics Group, Rice University, Houston, TX, USA
e-mail: ciarcia@rice.edu

15.1 Introduction

The rendezvous problem is as old as the space program. It has received renewed attention in the last 2 years for the following reason: NASA, DOD, and private industries have expressed interest in the development of technology enabling a chaser spacecraft to rendezvous and dock autonomously (without human intervention) with a target spacecraft.

The majority of the papers dedicated to the rendezvous problem have employed the assumption of given initial position and given initial velocity of the chaser spacecraft vis-à-vis the target spacecraft. In this research, the initial separation velocity components are assumed free and the initial separation coordinates are subject to the only requirement that the chaser-to-target distance is given.

Two problems are studied: time-to-rendezvous free and time-to-rendezvous given. It is assumed that the target spacecraft moves along a circular orbit around the Earth, that the chaser spacecraft has variable mass, and that its trajectory is governed by three controls, one determining the thrust magnitude and two determining the thrust direction. The Clohessy–Wiltshire equations [1] are used to describe the relative motion of the chaser vis-à-vis the target. There are two possible approaches to the above problems:

- The thrust controls are optimized together with the time lengths of all the subarcs and the initial values of the state variables.
- For each subarc, the thrust controls are kept constant; hence, they are replaced by thrust parameters which are optimized together with the time lengths of all the subarcs and the initial values of the state variables.

Systematic analyses performed by Miele, Weeks, and Ciarcià [2, 3] with the above approaches show that the performance index generated by the control approach 15.1 and the performance index generated by the parameter approach 15.1 are within less than 1% of one another. For this reason, we adopt the parameter optimization approach 15.1 instead of the control approach 15.1. Indeed, approach 15.1 is well suited to the generation of guidance trajectories.

The optimization of the initial separation coordinates and velocities as well as the time lengths of all the subarcs is performed for several values of the initial distance in the range $5 \leq d_0 \leq 60$ km with particular reference to the rendezvous between the Space Shuttle (SS) and the International Space Station (ISS). This study is of some importance because, for a preselected initial distance SS-to-ISS, it supplies not only the best initial conditions for the rendezvous maneuver, but also the corresponding final conditions of the ascent trajectory.

In the past, many articles have investigated the optimal rendezvous problem; see [4–20]. References [4–13] include papers which consider the impulsive thrust model and/or apply primer vector theory to determine instantaneous velocity changes. References [14–20] consider the finite-thrust model. In these papers, as well as in [4–13], the optimal rendezvous problem has been solved for the case of given initial relative position and relative velocity of the target vis-à-vis the chaser.

It appears that the problem of the absolute best initial conditions for the rendezvous problem has never been formulated, let alone solved.

15.2 Algorithm

In view of its flexibility, the sequential gradient-restoration algorithm (SGRA) has been employed for both the control approach 15.1 and the parameter approach 15.1.

In single-subarc form, SGRA was developed by Miele et al. during the period 1968–1986 [21–25]. It has proven to be a powerful tool for solving optimal trajectory problems of atmospheric and space flight. Applications and extensions of this algorithm have been reported in the United States, Japan, Germany, Spain, and other countries around the world; in particular, a version of this algorithm is used at NASA-JSC under the code name SEGRAM, developed by McDonnell Douglas Technical Service Company [26].

In multiple-subarc form, SGRA was developed by Miele and Wang [27, 28]. While the single-subarc SGRA deals with the optimization of a single system with initial and final boundary conditions, the multiple-subarc SGRA deals with the optimization of multiple systems with initial, final, and inner boundary conditions. In the multiple-subarc SGRA, a large and complicated overall system is decomposed into several subsystems along the time domain: each subsystem, having relatively simple properties, corresponds to a subarc; the connection between consecutive subsystems takes place via the inner boundary conditions.

In the application of the multiple-subarc SGRA, the actual time is denoted by ρ , the total number of subarcs is denoted by s , and the actual time length of each subarc is denoted by τ_i , $i = 1, 2, \dots, s$. Then, the end times ρ_i of all the subarcs are given by the recurrence relation

$$\rho_i = \rho_{i-1} + \tau_i, \quad i = 1, 2, \dots, s, \quad (15.1)$$

in which we set

$$\rho_0 = 0 \quad (15.2)$$

We denote by t the dimensionless virtual time, which is such that the time length of each subarc is normalized to 1. For the generic subarc i , the direct transformation from actual time to virtual time can be written as

$$t = (\rho - \rho_{i-1})/\tau_i, \quad i = 1, 2, \dots, s; \quad (15.3)$$

the corresponding inverse transformation from virtual time to actual time can be written as

$$\rho = \rho_{i-1} + \tau_i t, \quad i = 1, 2, \dots, s. \quad (15.4)$$

Then, with reference to the virtual time domain, the Bolza–Pontryagin optimization problem is formulated as follows: minimize the functional

$$I = \sum_{i=1}^s \int_0^1 f(x, u, \pi, t, i) dt + g(y, \pi), \quad (15.5)$$

with respect to the n -dimensional state vectors $x(t, i)$, the m -dimensional control vectors $u(t, i)$, and the p -dimensional parameter vector π which satisfy the n -dimensional differential constraints

$$x' = \Phi(x, u, \pi, t, i), \quad 0 \leq t \leq 1, \quad i = 1, 2, \dots, s, \quad (15.6)$$

plus a general q -dimensional boundary condition

$$\Psi(y, \pi) = 0, \quad q \leq 2ns, \quad (15.7)$$

incorporating the initial conditions, final conditions, and inner boundary conditions. In the above relations, s is the total number of subarcs, i is the index identifying a particular subarc, and the $2ns$ -vector y is given by

$$y = [x^T(0, i) \ x^T(1, i) \ \dots]^T, \quad i = 1, 2, \dots, s. \quad (15.8)$$

Clearly, the vector y includes the initial and final state vectors of all the subarcs. The time scaling factors τ_i , characterizing the map from the virtual time to the actual time, can be either fixed or free, depending on the optimization problem to be solved. In the latter case, the time scaling factors τ_i can be treated as elements of the unknown parameter vector π .

In the problem (15.5)-(15.4)-(15.5)-(15.6), the independent variable is the normalized time t , $0 \leq t \leq 1$; the prime denotes derivative with respect to the normalized time; f, g are scalar functions, Φ is an n -vector function, and Ψ is a q -vector function. The dependent variables are the state vectors $x(t, i)$, the control vectors $u(t, i)$, and the parameter vector π . Any trajectory satisfying the constraints (15.4), (15.5) and (15.6) is called a feasible trajectory; among the infinite number of feasible trajectories, we seek the special trajectory which minimizes the functional (15.5).

With the sequential gradient-restoration algorithm, a linear multipoint boundary-value problem is solved at each iteration using the method of particular solutions [21] to obtain a feasible solution and ultimately a locally optimal one. The SGRA algorithm involves a cyclical scheme whereby first the constraints (15.4), (15.5) and (15.6) are satisfied to a prescribed accuracy (restoration phase); then, using a first-order gradient method, a step is taken in the optimal direction to improve the performance index (gradient phase). Note that, in this implementation, (15.4), (15.5) and (15.6) are satisfied at the end of each restoration phase. The convergence tolerances for each phase are expressed as follows:

$$P \leq \zeta_1 \quad (15.9)$$

for the restoration phase and

$$Q \leq \zeta_2 \quad (15.10)$$

for the gradient phase, where P is the norm squared of the error in the constraints, Q is the norm squared of the error in the optimality conditions, and ζ_1, ζ_2 are small preselected constants.

The qualifier “sequential” is most appropriate for this algorithm due to the successive application of either a gradient iteration or a restorative iteration depending on the values of the scalar quantities P and Q . Specifically, three cases can occur:

- C1 – If $P > \zeta_1$, SGRA executes a restorative iteration leading to the decrease of the constraint error.
- C2 – If $P \leq \zeta_1$ but $Q > \zeta_2$, SGRA executes a gradient iteration leading to the decrease of the so-called augmented functional (original functional augmented by the constraints weighted via appropriate Lagrange multipliers).
- C3 – If $P \leq \zeta_1$ and $Q \leq \zeta_2$, convergence is declared and the algorithm stops.

In general, achieving specified target conditions is more important than obtaining an exact optimal solution; so, the constraint requirements (restoration phase) are stricter than the optimality condition requirements (gradient phase). This point of view is justified by the fact that the performance indexes of engineering problems tend to be quite flat, so that only small performance improvements occur after some point.

15.3 System Description

The relative motion of a chaser spacecraft P vis-à-vis a target spacecraft O in a circular orbit is expressed by the Clohessy–Wiltshire equations (CW) [1]. Two coordinate systems need to be defined: a fixed coordinate system $EXYZ$ and a moving coordinate system $Oxyz$ translating and rotating with angular velocity ω vis-à-vis $EXYZ$. The fixed coordinate system is centered at the Earth center E with axes X, Y, Z pointing to fixed directions in space. The moving coordinate system is of the *LVLH* type (local vertical, local horizontal) and is centered at the mass center O of the target spacecraft: the x -axis is aligned with the target velocity vector, positive backward; the y -axis is aligned with the orbital radius vector, positive upward; the z -axis is orthogonal to the xy -plane and completes the right-hand rule; hence, it is positive leftward with respect to the orbital plane of the target spacecraft.

let ρ denote the running time, $0 \leq \rho \leq \tau$, with τ the final time. Let M denote the instantaneous spacecraft mass; let M_0 denote the initial spacecraft mass; let $m = M/M_0$ denote the normalized mass, which is such that $m_0 = 1$.

Let \mathbf{T} denote the engine thrust vector with magnitude T and components T_x, T_y, T_z on the moving coordinate axes. Let $\boldsymbol{\sigma} = \mathbf{T}/M_0$ denote the thrust vector per unit initial mass with magnitude $\sigma = T/M_0$ and components $\sigma_x = T_x/M_0, \sigma_y = T_y/M_0, \sigma_z = T_z/M_0$ on the moving coordinate axes.

In second-order form, the CW equations are written as

$$\ddot{x} - 2\omega\dot{y} = \sigma_x/m, \quad (15.11)$$

$$\ddot{y} + 2\omega\dot{x} - 3\omega^2 y = \sigma_y/m, \quad (15.12)$$

$$\ddot{z} + \omega^2 z = \sigma_z/m, \quad (15.13)$$

$$\dot{m} = \sigma/V_e, \quad V_e = g_{ref} I_{sp}, \quad (15.14)$$

where the dot superscript denotes derivative with respect to the actual time ρ , with $0 \leq \rho \leq \tau$ and τ the final time. In Eqs. (15.11), (15.12) and (15.13), ω is the angular velocity of the moving reference frame vis-à-vis the fixed reference frame; terms linear in ω are due to the Coriolis acceleration; terms quadratic in ω are due in part to the transport acceleration and in part to the fact that the gravitational attractions on the chaser and target spacecraft are different in magnitude and direction. In Eq. (15.14), V_e is the jet exit velocity (relative to the engine frame), which in turn equals the product of the reference acceleration of gravity g_{ref} (acceleration of gravity at sea level on Earth) and the engine-specific impulse I_{sp} .

In first-order form, the CW equations are written as

$$\dot{x} = v_x, \quad (15.15)$$

$$\dot{y} = v_y, \quad (15.16)$$

$$\dot{z} = v_z, \quad (15.17)$$

$$\dot{v}_x = 2\omega v_y + \sigma_x/m, \quad (15.18)$$

$$\dot{v}_y = -2\omega v_x + 3\omega^2 y + \sigma_y/m, \quad (15.19)$$

$$\dot{v}_z = -\omega^2 z + \sigma_z/m, \quad (15.20)$$

$$\dot{m} = -\sigma/V_e, \quad V_e = g_{ref} I_{sp}, \quad (15.21)$$

where x, y, z are the separation coordinates in the downrange, radial, and transversal directions, and v_x, v_y, v_z are the separation velocities in the downrange, radial, and transversal directions. In Eqs. (15.15), (15.16), (15.17), (15.18), (15.19), (15.20) and (15.21), the dot superscript denotes derivative with respect to the actual time ρ , with $0 \leq \rho \leq \tau$ and τ the final time.

15.3.1 Multiple-Subarc Equations

The time transformation (15.3)–(15.4) allows us to rewrite the system (15.15), (15.16), (15.17), (15.18), (15.19), (15.20) and (15.21) in the following multiple-subarc form:

$$x' = \tau_i v_x, \quad (15.22)$$

$$y' = \tau_i v_y, \quad (15.23)$$

$$z' = \tau_i v_z, \quad (15.24)$$

$$v'_x = \tau_i (2\omega v_y + \sigma_x/m), \quad (15.25)$$

$$v'_y = \tau_i (-2\omega v_x + 3\omega^2 y + \sigma_y/m), \quad (15.26)$$

$$v'_z = \tau_i (-\omega^2 z + \sigma_z/m), \quad (15.27)$$

$$m' = \tau_i (-\sigma V_e), \quad V_e = g_{ref} I_{sp}, \quad (15.28)$$

in which the prime denotes derivative with respect to the normalized time t ,

$$t = (\rho - \rho_{i-1})/\tau_i, \quad 0 \leq t \leq 1, \quad i = 1, 2, \dots, s, \quad (15.29)$$

with

$$\rho_i = \rho_{i-1} + \tau_i, \quad 0 \leq t \leq 1, \quad i = 1, 2, \dots, s, \quad (15.30)$$

and $\rho_0 = 0$. In this formulation, the unknowns are the state vectors $[x(t, i) y(t, i) z(t, i) v_x(t, i) v_y(t, i) v_z(t, i) m(t, i)]$, the control vectors $[s_x(t, i) s_y(t, i) s_z(t, i)]$, and the parameters τ_i , the total time being

$$\tau = \sum \tau_i = \tau_1 + \tau_2 + \dots + \tau_s. \quad (15.31)$$

15.3.2 Inequality Constraint

The thrust T is bounded by the inequality

$$0 \leq T \leq T_{max}, \quad T = \sqrt{s_x^2 + s_y^2 + s_z^2}, \quad (15.32)$$

implying that the corresponding thrust acceleration σ is bounded by the inequality

$$0 \leq \sigma \leq \sigma_{max} \quad (15.33)$$

with

$$\sigma = T/M_0 = \sqrt{\sigma_x^2 + \sigma_y^2 + \sigma_z^2}. \quad (15.34)$$

Inequality (15.33) can be converted into equality via the nonsingular trigonometric transformation

$$\sigma = \frac{1}{2} \sigma_{max} (1 + \sin \alpha) \quad (15.35)$$

in which α denotes an auxiliary control.

Let θ denote the angle formed by the thrust vector and the xy -plane. Let Φ denote the angle which the projection of the thrust vector on the xy -plane forms with the x -axis. As a consequence, the original controls σ_x , σ_y , σ_z can be rewritten as

$$\sigma_x = \sigma \cos \theta \cos \Phi, \quad (15.36)$$

$$\sigma_y = \sigma \cos \theta \sin \Phi, \quad (15.37)$$

$$\sigma_z = \sigma \sin \theta. \quad (15.38)$$

The transformation (15.35), (15.36), (15.37) and (15.38) is nonsingular. It allows us to express the original controls σ_x , σ_y , σ_z in terms of the auxiliary controls α , θ , Φ . While the original controls σ_x , σ_y , σ_z are constrained by the inequality (15.33), the auxiliary controls α , θ , Φ must be considered as unconstrained: indeed, any choice of α , θ , Φ produces normalized thrust components [see Eqs (15.35), (15.36), (15.37) and (15.38)] which automatically satisfy the inequality (15.33).

15.3.3 Particular Cases

Previous works by Miele, Weeks, and Ciarcià [2, 3], neglecting the variability of the spacecraft mass, indicate that the solutions of the optimization problems include combinations of coasting zero-thrust subarcs and powered max-thrust subarcs. It is reasonable to expect that the same solution structure holds now that the variability of the mass is being considered. For a coasting zero-thrust subarc, $\alpha = -\pi/2$ and Eqs. (15.35), (15.36), (15.37) and (15.38) reduce to

$$\sigma = 0, \quad (15.39)$$

$$\sigma_x = 0, \quad (15.40)$$

$$\sigma_y = 0, \quad (15.41)$$

$$\sigma_z = 0, \quad (15.42)$$

For a powered max-thrust subarc, $\alpha = \pi/2$ and Eqs. (15.35), (15.36), (15.37) and (15.38) reduce to

$$\sigma = \sigma_{max}, \quad (15.43)$$

$$\sigma_x = \sigma_{max} \cos \theta \cos \Phi, \quad (15.44)$$

$$\sigma_y = \sigma_{max} \cos \theta \sin \Phi, \quad (15.45)$$

$$\sigma_z = \sigma_{max} \sin \theta, \quad (15.46)$$

15.3.4 Boundary Conditions

We consider the rendezvous between a target spacecraft and a chaser spacecraft under the following scenario. The orbit of the target spacecraft is circular and is

located at the Space Station altitude ($h = 390$ km, $r = 6768$ km); its eccentricity is 0.00 and its orbital inclination is 51.6° .

For the chaser spacecraft, the initial conditions are

$$x^2(0, 1) + y^2(0, 1) + z^2(0, 1) = d_0^2, \quad (15.47)$$

$$m(0, 1) = 1 \text{ (normalized mass)}, \quad (15.48)$$

where d_0 is the given initial separation distance. Also for the chaser spacecraft, the final conditions are

$$x(1, s) = 0 \text{ m}, \quad (15.49)$$

$$y(1, s) = 0 \text{ m}, \quad (15.50)$$

$$z(1, s) = 0 \text{ m}, \quad (15.51)$$

$$v_x(1, s) = 0 \text{ m/s}, \quad (15.52)$$

$$v_y(1, s) = 0 \text{ m/s}, \quad (15.53)$$

$$v_z(1, s) = 0 \text{ m/s}. \quad (15.54)$$

Finally, at the interface between any two contiguous subarcs, the following continuity conditions must be satisfied:

$$x(1, i) = x(0, i + 1), \quad i = 1, \dots, s - 1, \quad (15.55)$$

$$y(1, i) = y(0, i + 1), \quad i = 1, \dots, s - 1, \quad (15.56)$$

$$z(1, i) = z(0, i + 1), \quad i = 1, \dots, s - 1, \quad (15.57)$$

$$v_x(1, i) = v_x(0, i + 1), \quad i = 1, \dots, s - 1, \quad (15.58)$$

$$v_y(1, i) = v_y(0, i + 1), \quad i = 1, \dots, s - 1, \quad (15.59)$$

$$v_z(1, i) = v_z(0, i + 1), \quad i = 1, \dots, s - 1, \quad (15.60)$$

$$m(1, i) = m(0, i + 1), \quad i = 1, \dots, s - 1, \quad (15.61)$$

$$(15.62)$$

15.3.5 Performance Index

The performance index of interest in this work is the propellant mass ratio

$$I = m_p = 1 - m(1, s), \quad (15.63)$$

where

$$m_p = M_p/M_0, \quad m(1, s) = M(1, s)/M_0. \quad (15.64)$$

As a result, the Bolza–Pontryagin problem (15.5) is now reduced to a Mayer problem.

15.3.6 Approaches

Two possible approaches are considered:

- (I) For all the subarcs, the auxiliary controls $\alpha_i(t)$, $\theta_i(t)$, $\Phi_i(t)$ are optimized together with the time parameters τ_i and the initial values of the state variables $x(0, 1)$, $y(0, 1)$, $z(0, 1)$, $v_x(0, 1)$, $v_y(0, 1)$, $v_z(0, 1)$.
- (II) For all the subarcs, the auxiliary controls are replaced by the auxiliary parameters $\alpha_i(t)$, $\theta_i(t)$, $\Phi_i(t)$, which are then optimized together with the time parameters τ_i and the initial values of the state variables $x(0, 1)$, $y(0, 1)$, $z(0, 1)$, $v_x(0, 1)$, $v_y(0, 1)$, $v_z(0, 1)$.

Systematic analyses with the above approaches show that the performance index generated by the control approach 15.3.6 and the performance index generated by the parameter approach 15.3.6 are within less than 1% of one another. This is why, having in mind spacecraft guidance, we adopt the parameter optimization approach 15.3.6 instead of the control approach 15.3.6. Nevertheless, because of the great flexibility of the multiple-subarc sequential gradient-restoration algorithm, all the computations with either control and parameter approach 15.3.6 are done using the same algorithm, namely, the sequential gradient-restoration algorithm for optimal control problems.

15.4 Minimum Fuel, Time Free

For the minimum-fuel time-free problem, two possibilities were considered: a two-subarc zero-bang trajectory ($s = 2$) and a three-subarc bang-zero-bang trajectory ($s = 3$). Several tests were made and it was found that, from the point of view of the performance index, the two-subarc solution is better than the three-subarc solution: the addition of a third subarc, if feasible, causes an increase in the value of the performance index. In more detail, the best trajectory consists of a coasting zero-thrust subarc followed by a powered max-thrust braking subarc. Specifically, we have

$$a_1 = -p/2, s_1 = 0, \quad \text{subarc 1,} \quad (15.65)$$

$$a_2 = p/2, s_2 = s_{\max}, \quad \text{subarc 2.} \quad (15.66)$$

In light of (15.28), the performance index (15.63) reduces to

$$I = mp = (s_{\max}/V_e)t_2, \quad V_e = g_{\text{ref}}I_{sp}. \quad (15.67)$$

As a result, we deal with the following parameters:

$$t_1, \quad \text{subarc 1,} \quad (15.68)$$

$$t_2, q_2, f_2, \quad \text{subarc 2.} \quad (15.69)$$

Clearly, the parameter quadruplet

$$(t_1, t_2, q_2, f_2), \quad (15.70)$$

is unknown and must be found together with the sextuplet of initial values of the state variables

$$[x(0, 1), y(0, 1), z(0, 1), v_x(0, 1), v_y(0, 1), v_z(0, 1)]. \quad (15.71)$$

Clearly, the total number of unknowns is $n = 4 + 6 = 10$.

The essential constraints include the given initial distance condition (15.47) and the six final conditions (15.4) in which $s = 2$. Thus, the total number of essential constraints is $q = 1 + 6 = 7$. As a result, the optimal control problem degenerates into a mathematical programming problem in which the number of degrees of freedom is

$$NDF = n - q = 10 - 7 = 3. \quad (15.72)$$

This point of view is justified by the fact that the function of the differential constraints and the continuity conditions is merely that of connecting the end points of the two subarcs composing the extremal arc.

15.5 Results

The multiple-subarc sequential gradient-restoration algorithm was employed to solve the minimum-fuel time-free problem for several given values of the initial distance in the range $5 \leq d_0 \leq 60$ km. The max thrust-to-initial-mass ratio was set at the level $\sigma_{max} = 0.3 \text{ m/s}^2$. The assumed value for the specific impulse was $I_{sp} = 300$ s, corresponding to the gas exit velocity $V_e = 2.943 \text{ km/s}$.

For all values of the initial distance, it was found that the fuel-optimal trajectory is zero-bang; it includes a long coasting zero-thrust subarc followed by a short powered max-thrust subarc. Summary results can be found in Figs. 15.1 and 15.2 for all values of the initial distance. These figures show the nearly linear dependence of all the quantities considered from the initial distance: propellant mass ratio (Fig. 15.1), coasting subarc time (Fig. 15.2), powered subarc time (Fig. 15.3), initial separation coordinates (Fig. 15.4), and initial separation velocities (Fig. 15.5). Note that the total time $\tau = \tau_1 + \tau_2$ is nearly independent of the initial distance, $\tau = 4887$ s corresponding to the target angular travel $\chi = 317^\circ$; see (17.49). Because χ exceeds 180°

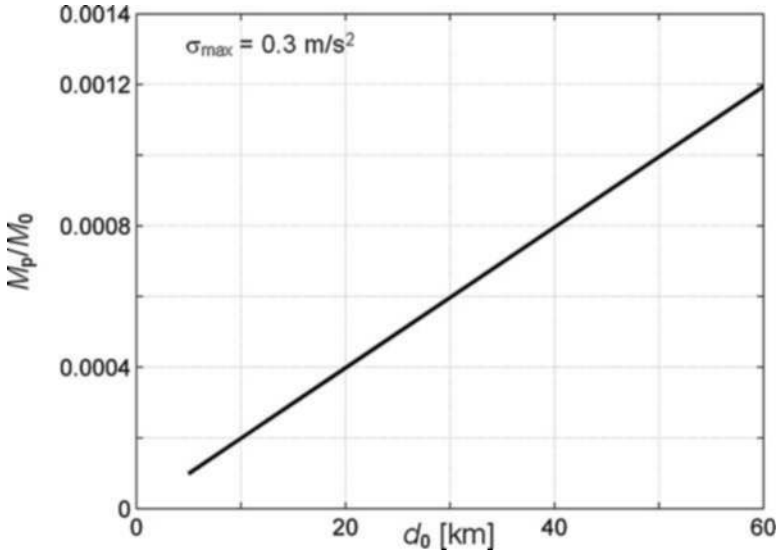


Fig. 15.1 Propellant mass ratio vs. initial distance, τ free (χ free)

for the time-free problem, this means that the target and chaser trajectories are flown partly in sunlight and partly in darkness, a circumstance rendering the rendezvous more difficult. Figure 15.1 shows the propellant mass ratio (ratio of propellant mass to the initial mass of the chaser). The order of magnitude of this ratio is 10^{-3} , meaning that the chaser mass is nearly constant during the rendezvous maneuver.

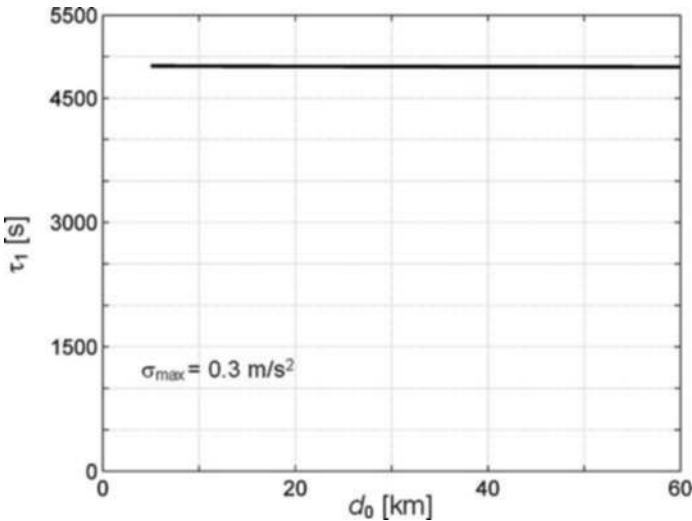


Fig. 15.2 Coasting subarc time vs. initial distance, τ free (χ free)

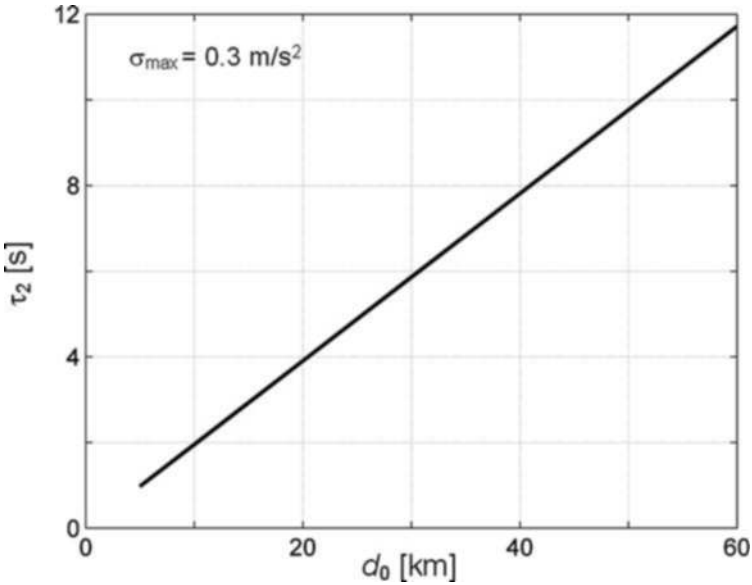


Fig. 15.3 Powered subarc time vs. initial distance, τ free (χ free)

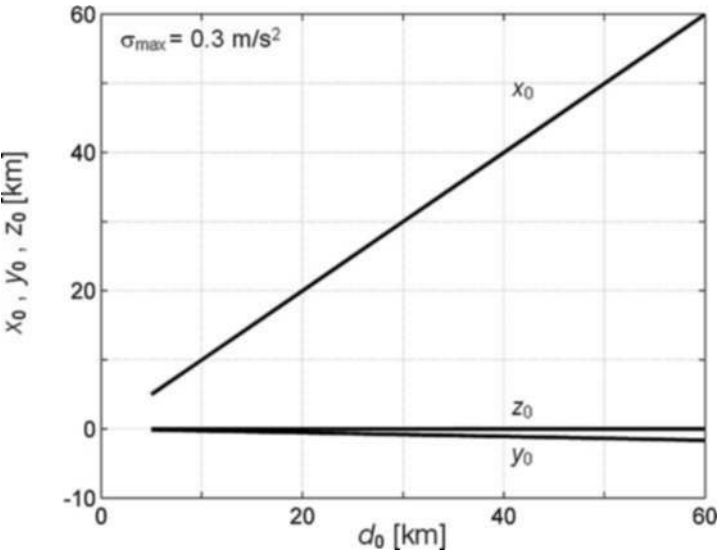


Fig. 15.4 Initial separation coordinates vs. initial distance, τ free (χ free)

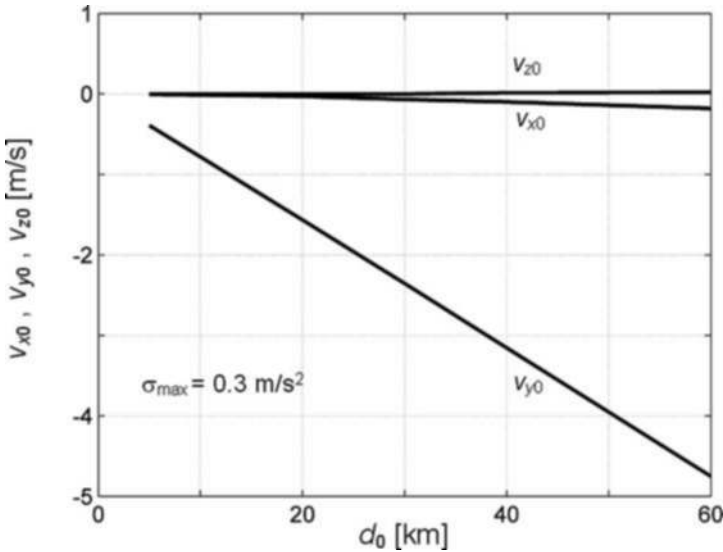


Fig. 15.5 Initial separation velocities vs. initial distance, τ free (χ free)

Figures 15.2 and 15.3 show the time lengths of the coasting subarc and the powered subarc; the coasting subarc duration is nearly constant, while the powered subarc duration increases linearly with the initial distance.

Figure 15.3 shows the initial separation coordinates and makes clear that the best initial position of the chaser spacecraft is behind and below the target spacecraft, albeit nearly in the same orbital plane of the target spacecraft.

Figure 15.4 shows the initial separation velocities and makes clear that the down-range velocity component is directed forward (this starts reducing the chaser-to-target distance), the radial velocity component is directed downward (this starts increasing the radial gap between chaser and target), while the transversal velocity component nearly vanishes (the chaser trajectory remains in the orbital plane of the target).

For the time-free case, more detailed information concerning the time history of the state variables and control variables can be found in [29].

15.6 Minimum Fuel, Time Given

The solution of the minimum-fuel time-free problem indicated that the rendezvous maneuver is flown partly in sunlight and partly in darkness, a complicating circumstance. To produce rendezvous trajectories flown completely in sunlight, we restudied the minimum-fuel problem with the added condition that the time-to-rendezvous is given. This is the same as constraining directly the angular travel c of the target spacecraft and indirectly the angular travel of the chaser spacecraft, with

$$\chi = \omega \tau. \quad (15.73)$$

Once more, we considered two possibilities: a two-subarc zero-bang trajectory ($s = 2$) and a three-subarc bang-zero-bang trajectory ($s = 3$). Once more, several tests were made and it was found that, from the point of view of the performance index, the two-subarc solution is better than the three-subarc solution: the addition of a third subarc, if feasible, causes an increase in the value of the performance index. In more detail, the best trajectory consists of a coasting zero-thrust subarc followed by a powered max-thrust braking subarc.

Specifically, we have

$$\alpha_1 = -\pi/2, \quad \sigma_1 = 0, \quad \text{subarc 1}, \quad (15.74)$$

$$\alpha_2 = \pi/2, \quad \sigma_2 = \sigma_{\max}, \quad \text{subarc 2}, \quad (15.75)$$

In light of (15.28), the performance index (15.63) reduces to

$$I = m_p = (s_{\max}/V_e) \tau_2, \quad V_e = g_{\text{ref}} I_{sp}, \quad (15.76)$$

and must be minimized subject to the constraints of the free-time problem plus the added constraint

$$\tau = \chi/\omega = \tau_1 + \tau_2 \quad (15.77)$$

with either τ or χ given.

Once more, we deal with the following parameters:

$$\tau_1, \quad \text{subarc 1}, \quad (15.78)$$

$$\tau_2, \theta_2, \Phi_2, \quad \text{subarc 2}. \quad (15.79)$$

Once more, we deal with the parameter quadruplet

$$(\tau_1, \tau_2, \theta_1, \Phi_2), \quad (15.80)$$

which is unknown and must be found together with the sextuplet of initial values of the state variables

$$[x(0, 1), y(0, 1), z(0, 1), v_x(0, 1), v_y(0, 1), v_z(0, 1)]. \quad (15.81)$$

Clearly, the total number of unknowns is still $n = 4 + 6 = 10$.

The essential constraints include the given initial distance condition (15.47), the six final conditions (15.49), (15.50), (15.51), (15.52), (15.53) and (15.54) in which $s = 2$, and the given time/angular travel condition (15.54). Thus, the total number of essential constraints is $q = 1 + 6 + 1 = 8$. As a result, the optimal control problem degenerates into a mathematical programming problem in which the number of degrees of freedom is

$$NDF = n - q = 10 - 8 = 2. \quad (15.82)$$

15.6.1 Results

The multiple-subarc sequential gradient-restoration algorithm was employed to solve the minimum-fuel time-given problem for several given values of the initial distance in the range $5 \leq d_0 \leq 60$ km. The max thrust-to-initial-mass ratio was set at the level $\sigma_{\max} = 0.3 \text{ m/s}^2$. The assumed value for the specific impulse was $I_{sp} = 300$ s, corresponding to the gas exit velocity, $V_e = 2.943 \text{ km/s}$. Three values were assumed for the given time,

$$\tau = 1847, 2309, 2771 \text{ s}, \quad (15.83)$$

corresponding to the target angular travel

$$\chi = 120, 150, 180^\circ. \quad (15.84)$$

For all values of the initial distance, it was found that the fuel-optimal trajectory is zero-bang; it includes a long coasting zero-thrust subarc followed by a short powered max-thrust subarc. Summary results can be found in Figs. 15.3, 15.4, 15.5 and 15.6 for all values of the initial distance. These figures show the nearly linear dependence of all the quantities considered from the initial distance: propellant mass ratio (Fig. 15.6, coasting subarc time (Fig. 15.7), powered subarc time (Fig. 15.8), initial separation coordinates (Figs. 15.9, 15.9, 15.11), and initial separation velocities (Figs. 15.10, 15.10, 15.12).

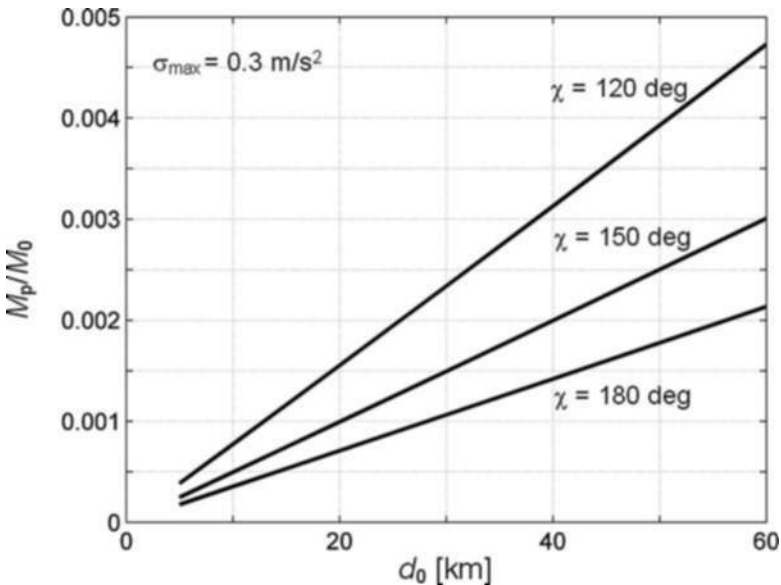


Fig. 15.6 Propellant mass ratio vs. initial distance, τ given (χ given)

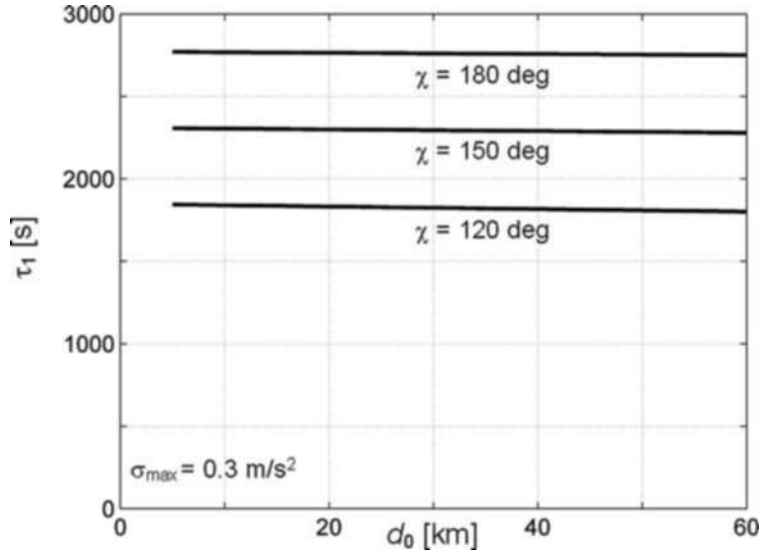


Fig. 15.7 Coasting subarc time vs. initial distance, τ given (χ given)

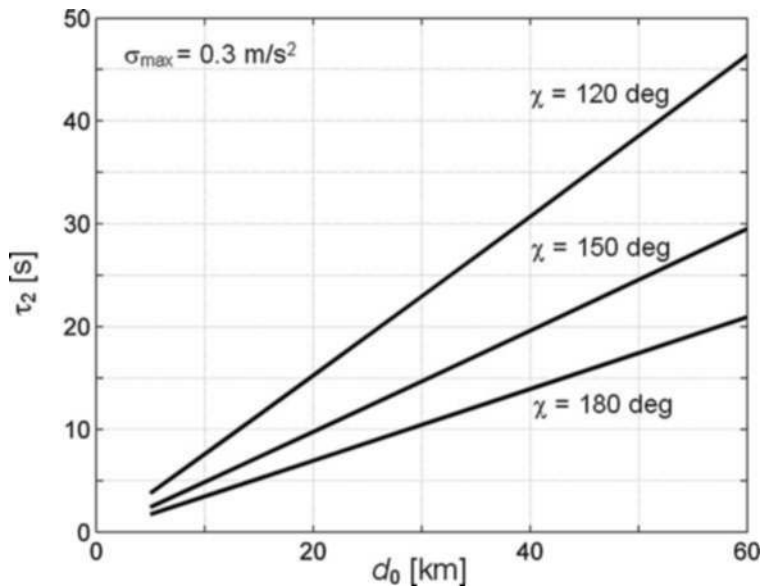


Fig. 15.8 Powered subarc time vs. initial distance, τ given (χ given)

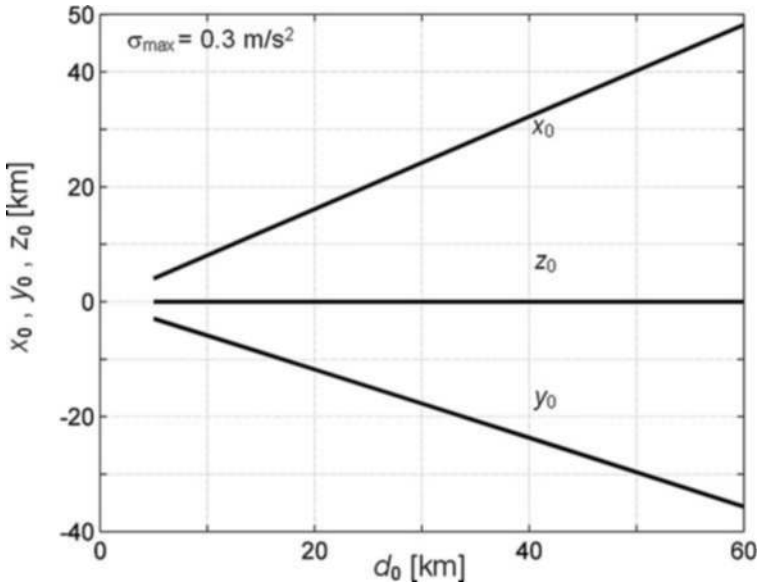


Fig. 15.9 Initial separation coordinates vs. initial distance, $\tau = 1847$ s ($\chi = 120^\circ$)

Propellant mass ratio. Comparison of Fig. 15.1 (time free, $\chi = 317^\circ$) with Fig. 15.6 (time given, hence target angular travel given) shows that, for any given initial distance, the propellant mass ratio increases by the multiplicative factor 1.8 for $\chi = 180^\circ$, by the multiplicative factor 2.5 for $\chi = 150^\circ$, and by the multiplicative factor 4.0 for $\chi = 120^\circ$. Therefore, there is a severe fuel penalty when performing the rendezvous maneuver entirely in sunlight (instead of partly in sunlight and partly in darkness), the penalty increasing as the target angular travel decreases.

Coasting subarc time. Comparison of Fig. 15.2 (time free, $\chi = 317^\circ$) with Fig. 15.7 (time given, hence target angular travel given) shows that, for any given initial distance, the coasting subarc time reduces according to the multiplicative factor 0.56 for $\chi = 180^\circ$, according to the multiplicative factor 0.47 for $\chi = 150^\circ$, and according to the multiplicative factor 0.37 for $\chi = 120^\circ$.

Powered subarc time. Comparison of Fig. 15.3 (time free, $\chi = 317^\circ$) with Fig. 15.8 (time given, hence target angular travel given) shows that, for any given initial distance, the powered subarc time increases by a multiplicative factor 1.8 for $\chi = 180^\circ$, by a multiplicative factor 2.5 for $\chi = 150^\circ$, and by a multiplicative factor 4.0 for $\chi = 120^\circ$. Note that the multiplicative factors for the powered subarc time are the same as the multiplicative factors of the propellant mass ratio.

Figures 15.9, 15.11, 15.13 show the initial separation coordinates and make clear that the best initial position of the chaser spacecraft is behind and below the target spacecraft, albeit nearly in the same orbital plane of the target spacecraft.

Figures 15.10, 15.12, 15.14 show the initial separation velocities and make clear that the initial downrange velocity component is directed forward (this starts reduc-

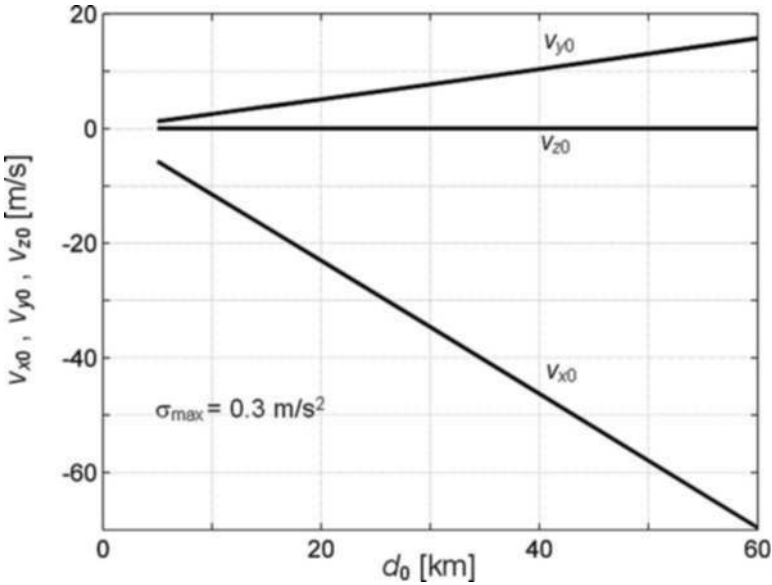


Fig. 15.10 Initial separation velocities vs. initial distance, $\tau = 1847 \text{ s}$ ($\chi = 120^\circ$)

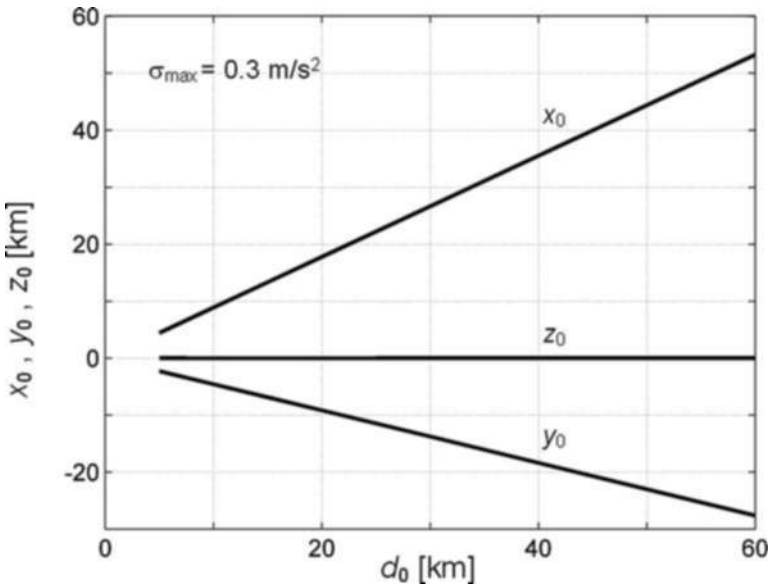


Fig. 15.11 Initial separation coordinates vs. initial distance, $\tau = 2309 \text{ s}$ ($\chi = 150^\circ$)

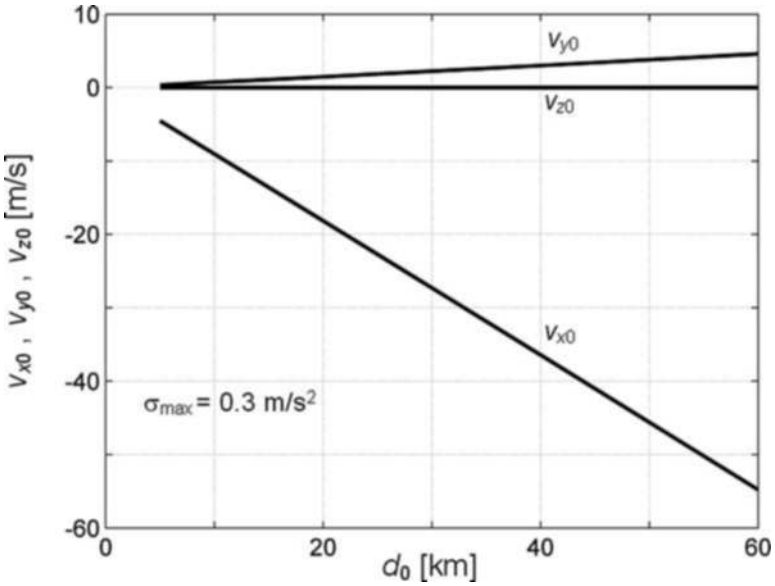


Fig. 15.12 Initial separation velocities vs. initial distance, $\tau = 2309 \text{ s}$ ($\chi = 150^\circ$)

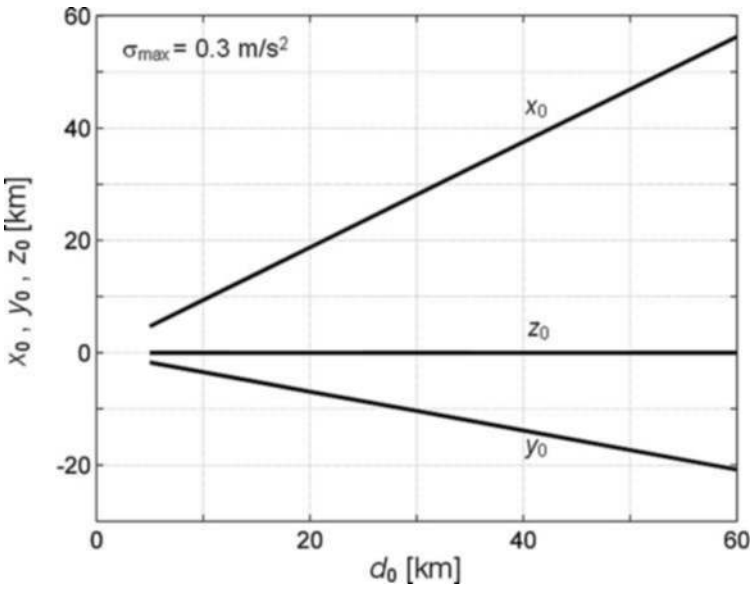


Fig. 15.13 Initial separation coordinates vs. initial distance, $\tau = 2771 \text{ s}$ ($\chi = 180^\circ$)

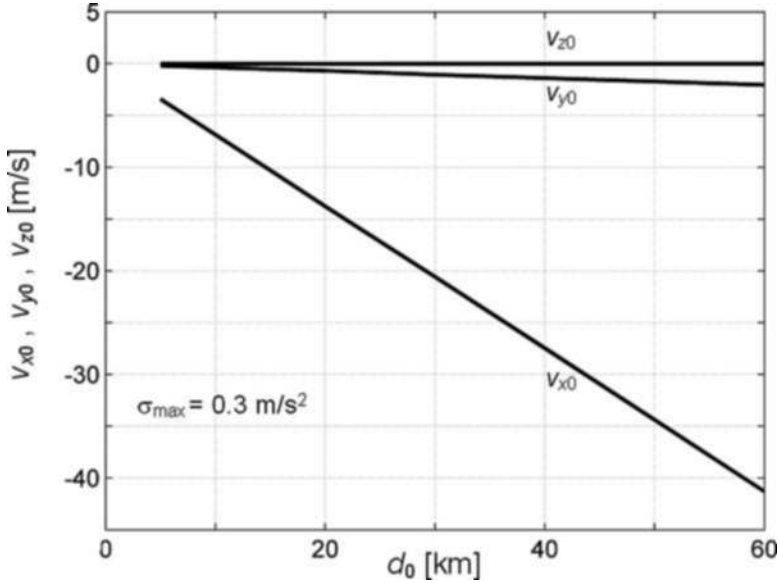


Fig. 15.14 Initial separation velocities vs. initial distance, $\tau = 2771$ s ($\chi = 180^\circ$)

ing the chaser-to-target distance), while the initial transversal velocity component nearly vanishes (the chaser trajectory remains in the orbital plane of the target).

For the time-given case, more detailed information concerning the time history of the state variables and control variables can be found in [29].

15.7 Conclusions

Most papers dealing with the rendezvous problem have assumed given initial conditions for the position and velocity of the chaser spacecraft vis-à-vis the target spacecraft. In this study, the initial separation velocity components are assumed free, while the initial separation coordinates are subject to the only requirement that the corresponding chaser-to-target distance is given. Two cases are considered: time-to-rendezvous free and time-to-rendezvous given, the latter being equivalent to constraining directly the angular travel of the target spacecraft and indirectly the angular travel of the chaser spacecraft. The optimization criterion is the mass of propellant consumed.

The analysis shows that, for all the values of the initial distance considered ($5 \leq d_0 \leq 60$ km), the resulting fuel-optimal trajectory is zero-bang; namely, it includes a long coasting zero-thrust subarc followed by a short braking max-thrust subarc. In all the examples, the chaser fuel-optimal trajectory is contained in the orbital plane of the target; in addition, the best initial position of the chaser is behind and below

the target. In both cases where the time-to-rendezvous is free and where the time-to-rendezvous is given, the propellant mass ratio (ratio of the propellant mass to the initial mass of the chaser) needed to execute the maneuver increases linearly with the initial distance.

If the time-to-rendezvous is free, the propellant mass ratio is of the order 10^{-3} , a very desirable value, but the resulting long trajectory is to be flown partly in sunlight and partly in darkness. If the time-to-rendezvous is given, namely if the angular travel χ of the target spacecraft is set at a value sufficiently smaller than the value characterizing the time-free rendezvous trajectory, the resulting optimal trajectory can be flown entirely in sunlight, albeit at the expense of an increase in propellant consumption according to a multiplicative factor 1.8 for $\chi = 180^\circ$, 2.5 for $\chi = 150^\circ$, and 4.0 for $\chi = 120^\circ$.

The above results were obtained having specifically in mind the rendezvous of the Space Shuttle (chaser) and the International Space Station (target). Once a given initial distance SS-to-ISS is preselected, the present analysis supplies not only the best initial conditions for the rendezvous trajectory but simultaneously the corresponding final conditions for the Space Shuttle ascent trajectory.

References

1. Clohessy, W.H., and Wiltshire, R.S., Terminal Guidance System for Satellite Rendezvous, *Journal of the Aerospace Sciences*, Vol. 27, No. 9, pp. 653–658, 1960.
2. Miele, A., Weeks, M.W., and Ciarcià, M., Optimal Trajectories for Spacecraft Rendezvous, *Journal of Optimization Theory and Applications*, Vol. 132, No. 3, pp. 353–376, 2007.
3. Miele, A., Ciarcià, M., and Weeks, M.W., Guidance Trajectories for Spacecraft Rendezvous, *Journal of Optimization Theory and Applications*, Vol. 132, No. 3, pp. 377–400, 2007.
4. Goldstein, A.A., Green, A.H., Johnson, A. T., Seidman, T.I., Fuel Optimization in Orbital Rendezvous, AIAA Paper 63–354, AIAA Guidance and Control Conference, Cambridge, Massachusetts, 1963.
5. Lion, P. M., and Handelsman, M., Primer Vector on Fixed-Time Impulsive Trajectories, *AIAA Journal*, Vol. 6, No. 1, pp. 127–132, 1968.
6. Jones, B. J., Optimal Rendezvous in the Neighborhood of a Circular Orbit, *Journal of the Astronautical Sciences*, Vol. 24, No. 1, pp. 55–90, 1976.
7. Jezewski, D. J., Primer Vector Theory Applied to the Linear Relative-Motion Equations, *Optimal Control Applications and Methods*, Vol. 1, No. 4, pp. 387–401, 1980.
8. Chiu, J. H., Optimal Multiple-Impulse Nonlinear Orbital Rendezvous, PhD Thesis, University of Illinois at Urbana-Champaign, 1984.
9. Prussing, J. E., and Chiu, J. H., Optimal Multiple-Impulse Time-Fixed Rendezvous between Circular Orbits, *Journal of Guidance, Control, and Dynamics*, Vol. 9, No. 1, pp. 17–22, 1986.
10. Carter, T. E., and Briant, J., Linearized Impulsive Rendezvous Problem, *Journal of Optimization Theory and Applications*, Vol. 86, No. 3, pp. 553–584, 1995.
11. Guzman, J., Mailhe, L., Schiff, C., and Hughes, S., Primer Vector Optimization: Survey of Theory and Some Applications, Paper IAC-02-A. 6.09, 53rd International Astronautical Congress, Houston, Texas, 2002.
12. Shen, H., and Tsiotras, P., Optimal Two-Impulse Rendezvous Using Multiple Revolution Lambert Solutions, *Journal of Guidance, Control, and Dynamics*, Vol. 26, No. 1, pp. 50–61, 2003.

13. Prussing, J.E., Optimal Two-Impulse and Three-Impulse Fixed-Time Rendezvous in the Vicinity of a Circular Orbit, *Journal of Spacecraft and Rockets*, Vol. 40, No. 6, pp. 952–959, 2003.
14. Paiewonsky, B., and Woodrow, P.J., Three-Dimensional Time-Optimal Rendezvous, *Journal of Spacecraft and Rockets*, Vol. 3, No. 11, pp. 1577–1584, 1966.
15. Carter, T.E., and Humi, M., Fuel-Optimal Rendezvous Near a Point in General Keplerian Orbit, *Journal of Guidance, Control, and Dynamics*, Vol. 10, No. 6, pp. 567–573, 1987.
16. Van Der Ha, J. C., Analytical Formulation for Finite-Thrust Rendezvous Trajectories, Paper IAF-88-308, 39th Congress of the International Astronautical Federation, Bangalore, India, 1988.
17. Carter, T.E., and Brient, J., Fuel-Optimal Rendezvous for Linearized Equations of Motion, *Journal of Guidance, Control, and Dynamics*, Vol. 15, No. 6, pp. 1411–1416, 1992.
18. Carter, T.E., Optimal Power-Limited Rendezvous of a Spacecraft with Bounded Thrust and General Linear Equation of Motion, *Journal of Optimization Theory and Applications*, Vol. 87, No. 3, pp. 487–515, 1995.
19. Carter, T.E., and Pardis, C. J., Optimal Power-Limited Rendezvous with Upper and Lower Bounds on Thrust, *Journal of Guidance, Control, and Dynamics*, Vol. 19, No. 5, pp. 1124–1133, 1996.
20. Park, C., Guibout, V., and Scheeres, D. J., Solving Optimal Continuous Thrust Rendezvous Problem with Generating Functions, *Journal of Guidance, Control, and Dynamics*, Vol. 29, No. 25, pp. 321–331, 2006.
21. Miele, A., Method of Particular Solutions for Linear Two-Point Boundary-Value Problems, *Journal of Optimization Theory and Applications*, Vol. 2, No. 4, pp. 260–273, 1968.
22. Miele, A., Pritchard, R.E., and Damoulakis, J.N., Sequential Gradient-Restoration Algorithm for Optimal Control Problems, *Journal of Optimization Theory and Applications*, Vol. 5, No. 4, pp. 235–282, 1970.
23. Miele, A., Tietze, J. L., and Levy, A. V., Summary and Comparison of Gradient-Restoration Algorithms for Optimal Control Problems, *Journal of Optimization Theory and Applications*, Vol. 10, No. 6, pp. 381–403, 1972.
24. Miele, A., Recent Advances in Gradient Algorithms for Optimal Control Problems, *Journal of Optimization Theory and Applications*, Vol. 17, Nos. 5–6, pp. 361–430, 1975.
25. Gonzalez, S., and Miele, A., Sequential Gradient-Restoration Algorithm for Optimal Control Problems with General Boundary Conditions, *Journal of Optimization Theory and Applications*, Vol. 26, No. 3, pp. 395–425, 1978.
26. Rishikof, B. H., McCormick, B. R., Pritchard, R. E., and Sponaugle, S. J., SEGRAM: A Practical and Versatile Tool for Spacecraft Trajectory Optimization, *Acta Astronautica*, Vol. 26, Nos. 8–10, pp. 599–609, 1992.
27. Miele, A., and Wang, T., Multiple-Subarc Sequential Gradient-Restoration Algorithm, Part 1: Algorithm Structure, *Journal of Optimization Theory and Applications*, Vol. 116, No. 1, pp. 1–17, 2003.
28. Miele, A., and Wang, T., Multiple-Subarc Sequential Gradient-Restoration Algorithm, Part 2: Application to a Multistage Launch Vehicle Design, *Journal of Optimization Theory and Applications*, Vol. 116, No. 1, pp. 19–39, 2003.
29. Miele, A., and Ciarcia, M., Optimal Starting Conditions for the Rendezvous Maneuver: Analytical and Computational Approach, Aero-Astronautics Report 361, Rice University, 2007.

“This page left intentionally blank.”

Chapter 16

Commercial Aircraft Design for Reduced Noise and Environmental Impact

S. Mistry, Howard Smith, and John P. Fielding

Abstract This chapter describes the noise and global warming effects produced by current commercial aircraft. It also describes noise and pollution sources and proposes technologies to mitigate them.

The chapter will then describe some estimates of the financial costs of aircraft pollution.

Significant environmental progress can only be made by more radical aircraft and engine configurations. A low-noise design methodology will be described, followed by results from using it to design a conventional baseline design, and a number of advanced low-noise designs. Particular attention will be given to the Broad Delta and blended-wing body designs. The chapter continues with a description of the 50,000 man-hours post-grad group design A-6 Greenliner project. This is a long-range 385 seat aircraft, which has been designed both to be quiet and to produce reduced atmospheric pollution. The A-6 is controversial, in that it is planned to fly at $M = 0.74$, thus trading speed for environmental effects and costs. Lessons learnt from these projects will be summarized and future plans discussed.

S. Mistry

Department of Aerospace Engineering, Cranfield University, Cranfield MK43 0AL, UK,
e-mail: s.mistry.2003@cranfield.ac.uk

Howard Smith

Department of Aerospace Engineering, Cranfield University, Cranfield MK43 0AL, UK,
e-mail: howard.smith@cranfield.ac.uk

John P. Fielding

Department of Aerospace Engineering, Cranfield University, Cranfield MK43 0AL, UK,
e-mail: j.p.fielding@cranfield.ac.uk

16.1 Introduction

The impact of aviation on the environment has been a major concern for at least two decades. Negative public opinion is growing rapidly and significant steps are required to alleviate aircraft emissions in a growing air transport market. There are worldwide concerns, and a significant initiative was taken in Europe in 2001, with the publication of quantified targets. These were summarized in [1] as

Europe...has it's Advisory Council for Aeronautical Research in Europe (ACARE), which, in 2001 committed the industry to developing technologies that would cut CO₂ by 50% per passenger kilometre, cut NO_x by 80% and halve perceived noise levels by 2020, relative to 2000 standards.

It is the authors' option that such large reductions will not only require new technologies and operating techniques but also need new aircraft configurations. Some earlier Cranfield University projects and technologies are described in [2], and more recent work is described below.

It is very difficult to assess the cost penalties of aviation pollution, but some understanding is required, so that trade-offs may be made with respect to other aircraft performance aspects. Such trade-offs will allow the production of a design process that may numerically allow for aircraft pollution effects. The first section of this chapter makes an initial attempt at this process.

The most important perceived local environmental issues are noise and emissions at airports. The second section of this chapter will discuss studies performed to mitigate noise, but will not address local air quality. The third section of the chapter describes an aircraft designed to mitigate global warming.

16.2 Simple Emission Trade-Off Study

16.2.1 Global Warming Costs

16.2.1.1 Quoted Figures

There are a number of conflicting reports which attempt to quantify the costs of global warming:

1. *The British Greener-by Design Report* [3]. This report was based on the British Government report [4, 5] and quoted the annual UK figure at £1,400 M. This equated to \$2,464 M at an exchange rate of £1 = \$1.76, which will be used for the rest of this section.
2. *The CLIMATECARE Web site* [6]. This web site allows the user to calculate the CO₂ burn for a particular flight per passenger, together with an off-set cost. The latter was calculated to be £7.5 (\$13.2) per tonne of CO₂. This equates to \$415 M per annum.

16.2.1.2 Investigation of the Above Figures

The authors were unable to obtain the basis of method 2 but method 1 seemed to be the most authoritative, and its basis has been investigated further. Discussion with the author of [3] led to its source document [4]. This was a report produced in 2003 by the UK's Department for Transport. It estimated the total UK aviation-related CO₂ emissions in 2000. This was related to the total fuel uplift for the relevant aircraft fleet. The estimated CO₂ was 27.33×10^6 tonnes. A factor of 3.67 was used to convert this to 7.446×10^6 tonnes of carbon. A radiative forcing index factor of 2.7 was used to allow for non-CO₂ global warming effects. This was based on simply scaling the 1992 IPCC report total global warming effects, relative to CO₂. Please note that this figure is subject to current vigorous debate. This gave a total figure of 20×10^6 tonnes of carbon. Reference [5] made estimates for the global warming effects of £35–140 per tonne of carbon and chose a value of £70. This gave the total annual cost as £1400 M, as quoted above. It is clear that this assumed value is crucial to the calculation. This figure was based on non-catastrophic changes to human health and agriculture caused by global warming. There is, however, considerable uncertainty in these figures. Another means of estimating the cost of carbon is shown in [5]. This shows a large number of estimates of prevention costs required to meet KYOTO targets and values of emission trading values. These vary widely, depending on the trading system used. The average values were some €75 per tonne of carbon or £50 per tonne or \$88. The latter figure was used here in subsequent calculations, but this value should be investigated in future studies.

16.2.2 Noise Costs

Reference [4] quoted noise costs of aviation in the United Kingdom in 2003 as

1. South-East England costs equate to 36–40 pence per passenger at Heathrow.
2. Other South-East England Airports were estimated at 5 pence per passenger.

These estimates were based on noise increments greater than 57 dBA Leq. It was stated that a sustained noise increase of 1 dBA was likely to reduce house prices by 0.5–1%. The fact that house prices are very expensive in London explains some of the differences. The authors assumed that the costs were for a complete landing, take-off cycle, and chose an average value of 22P, or 40c per passenger.

16.2.3 Local Air Quality Cost (LAQ)

Reference [4] states that most local aviation quality reduction is caused by NO₂ and PM₁₀. Cost estimates were made to quantify the costs of LAQ by examining Health Service cost increases associated with respiratory illnesses. Further studies in [4]

included a range of health and environmental impacts to estimate that LAQ costs are €1.5 (\$1.76). This value was assumed for this study.

16.2.4 Annual Fuel Costs Fro Baseline Aircraft

Reference [4] shows representative flights data for Boeing 747 and 737 aircraft:

1. B747. The aircraft flies 3724 n.miles and consumes 74.1 tonnes of fuel. Assuming an average speed of 500 nm per hour, this flightlength is approximately 7.5 hours. We thus have fuel burn/hour of approximately 10 tonnes. It is assumed that a B747 will have an annual utilization of 4500 hours, therefore fuel burn/year = 45,000 tonnes. There have been many fluctuations in fuel costs during recent months. Fuel costs at the time of writing were \$1.8/US gallon. Aviation fuel density is 3.02 KG/USG , therefore fuel cost/ $\text{KG} = \frac{1.8}{3.02} = \0.6
B747 annual fuel costs: $45,000 \cdot 1000 \cdot 0.6 = \$27M$
2. B737. Again from [4], a B737 burns 3.5 tonnes of fuel during a 600 n.mile flight. Assuming an average speed of 400 Kt and annual utilization of 3100 hours, we calculate that it will use 6860 tonnes of fuel.

At \$1.8/USG we calculate an annual fuel burn of

$$6860 \cdot 1000 \cdot 0.6 = \$4.116M$$

16.2.5 Baseline Aircraft Environmental Costs

16.2.5.1 Global Warming

Using the chosen value of \$88 per/tonne of carbon, we need to estimate annual costs based on the fuel burn of Sect. 16.2.4 above.

We know that 1 tonne of aviation fuel produces 3.15 tonnes of CO₂ [4]. This value is scaled by 2.7 to account for all radiative forcing.

Therefore 1 tonne of fuel = $3.15 \cdot 2.7 = 8.505$ tonnes CO₂

Conversion CO₂ to C we divide by 3.67 to give 2.32 tonnes.

We see that 1 tonne of fuel costs $\$2.32 \cdot 88 = \204

Annual global warming costs are thus the following:

$$\text{B747: } 45,000 \cdot 204 = \$9.18M$$

$$\text{B737: } 6860 \cdot 204 = \$1.4M$$

16.2.5.2 Noise Costs

B747. We will assume 400 passengers per aircraft, with an average flight length of 4.5 hours and 1000 flights per year. Using a cost of \$0.4 per passenger (Sect. 16.2.2).

Annual noise costs: $400 \cdot 1000 \cdot 0.4 = \$160,000$

B737. We will assume 150 passengers and an average flight length of 1.5 with an annual utilization of 3100 hours.

Annual noise cost: $3100 \cdot 150 \cdot 0.4 / 1.5 = \$124,000$

16.2.5.3 LAQ Costs

These are based on \$1.76 per passenger (Sect. 16.2.3). We therefore have

B747: $400 \cdot 1000 \cdot 1.76 = \$704,000$

B737: $150 \cdot 31,000 \cdot 1.76 / 1.5 = \$546,000$

16.2.6 Summary of Trade-Offs

Table in Fig. 16.1 shows a summary of the results from the above paragraphs. It must be stressed that it was difficult to obtain detailed cost estimates. A number of assumptions had to be made, and they have been stated. Using the assumptions made, it is clear that environmental costs can be up to 50% of fuel costs. Future studies will be needed to obtain more confidence in the results. It is quite clear, however, that there are huge potential environmental costs. It is also clear that there are many technologies that may be employed to mitigate them. A careful balance must be achieved.

Topic	Boeing 747	Boeing 737
Annual fuel burn (tonnes)	45,000	6,900
Annual global warming costs (USD)	9.18M	1.4M
Noise costs (USD)	160,000	124,000
LAQ costs (USD)	704,000	546,000
Annual environment costs (USD)	10.04M	2.07M
Annual fuel costs (USD)	27M	4.12M
% of av/fuel costs	37%	50.2%

Fig. 16.1 Summary of trade-offs: costs are based on individual aircraft years

16.3 Aircraft Designs for Reduced Noise

16.3.1 Background

Cambridge University and Massachusetts Institute of Technology recently concluded the “Silent Aircraft Initiative.” Cranfield University contributed to this

programme by investigating novel airframe and propulsion concepts [7]. This chapter will discuss some airframe aspects of this work.

The sources of airframe noise are mainly due to surface interference or obstructions to airflow (Fig. 16.2).

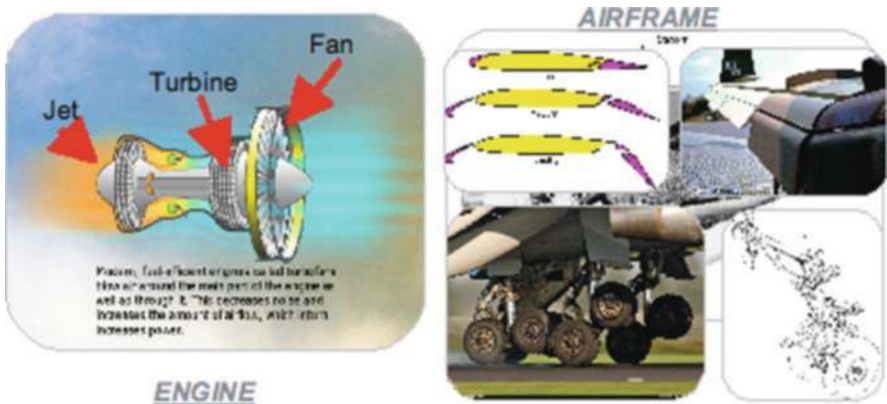


Fig. 16.2 Aircraft noise sources

The main contributors are undercarriage, leading edge (LE) devices, trailing edge (TE) flaps, and empennage. Although these provide the majority of noise, additional sources are present, such as the wing, wing–fuselage interface, wing–pylon, and pylon–engine nacelle joints.

Further noise sources include hatches, cavities, and surface vibrations during flight. Current trends in noise reduction have led to minor improvements which have little effect on overall noise. Some progress is promised by the use of landing gear fairings, but more radical aircraft configurations offer more significant savings, as described in this chapter. These will be compared with a baseline aircraft design, which is described below.

16.3.2 Baseline Aircraft Design and Noise Prediction

The initial aircraft specification was for a medium-range air transport with twin engines a 269 passenger payload, cruise speed of Mach 0.8, and a range of 4,020 nautical miles [7].

The method of [8] was used to perform the initial conceptual design of the aircraft. Structures were assumed to be primarily constructed from metallic materials but a 12% structure mass reduction was applied to allow for current levels of use of composite materials.

The configuration and mass predictions were very close to those of a current in-service aircraft with very similar performance characteristics. Figure 16.3 shows a computer-aided design model of the baseline aircraft.

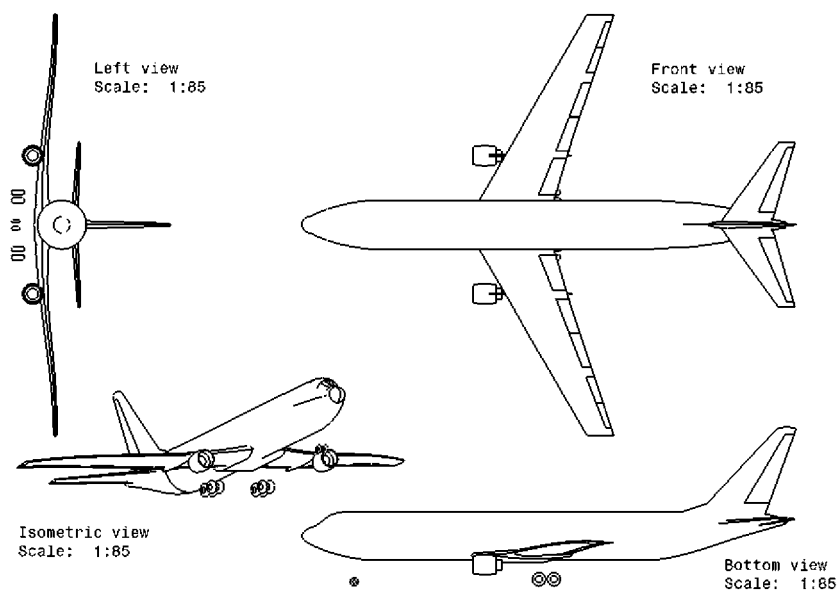


Fig. 16.3 Baseline aircraft CAD model

16.3.3 Low Airframe Noise Design Methodology

Figure 16.4 shows the main design stages. Semi-empirical design methods were used to produce the baseline aircraft design described above and the broad deltas described later. The resulting designs were compared with existing aircraft in terms of geometry, mass, and performance.

Several approach analyses were performed, as steep and slow approaches were found to be effective in reducing approach noise.

The final stage in the concept design analysis was to assess the noise produced from the airframes, which was achieved using simple methods.

The primary noise prediction tool was the ESDU semi-empirical low-fidelity computer model. This is applicable to most conventional aircraft configurations.

When combined with engine noise models, these results compared well with empirical data from similar aircraft.

16.3.4 Low-Noise Aircraft Concept Brainstorming Process

The development of novel designs was achieved through group brainstorming sessions to identify possibilities for multiple airframe configurations.

Reference [7] gives a full description of the structured down-selection process which led to the seven configurations shown in Fig. 16.5.

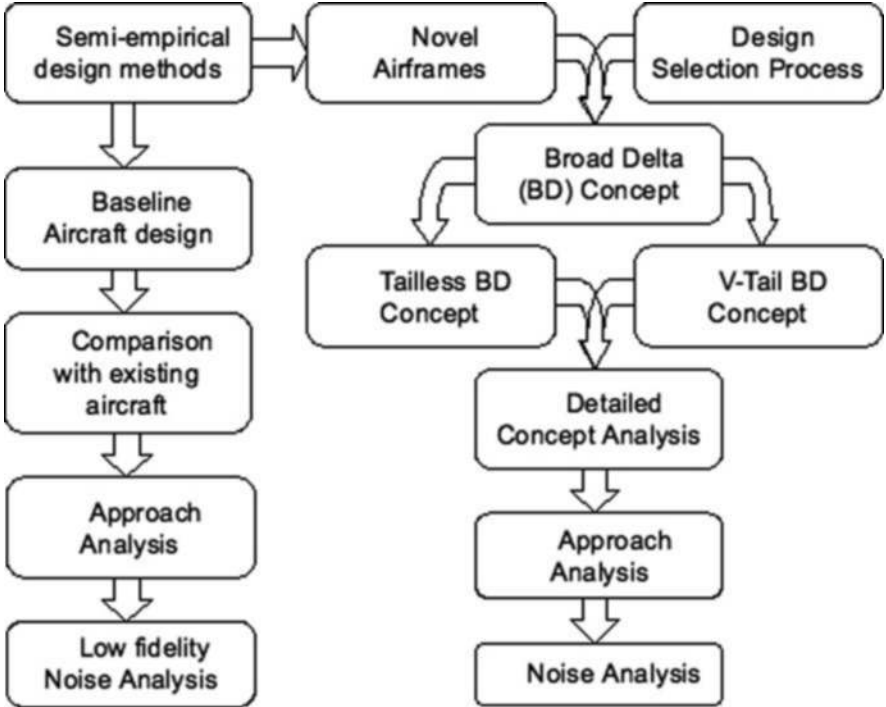


Fig. 16.4 Basic low airframe noise design methodology

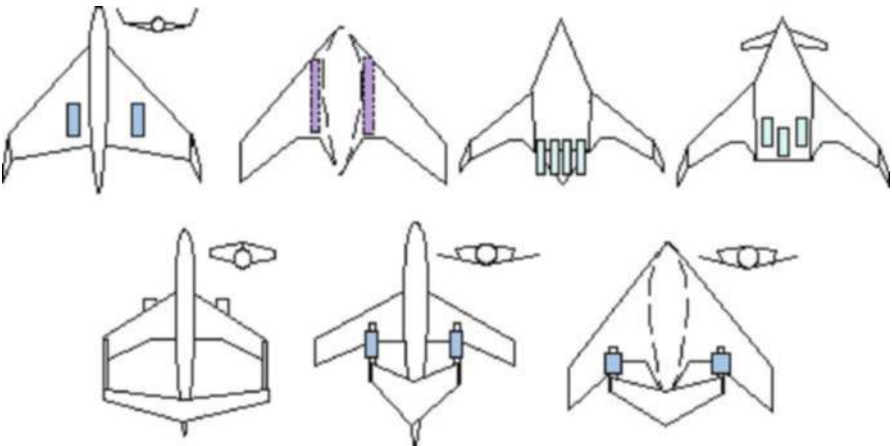


Fig. 16.5 Down-selected low-noise configurations

Some earlier work done at Cranfield University examined 650 and 250 seat blended-wing body aircraft [2].

The former was a long-range aircraft designed to requirements similar to those that led to the Airbus A380. Figure 16.6 shows this aircraft, while Fig. 16.7 shows the 250 seat BW-01 aircraft. The latter aircraft showed great promise, which is currently being assessed by the continued development of the Cranfield/BAE SYSTEMS KESTREL BWB uninhabited flying demonstration aircraft (Fig. 16.8).

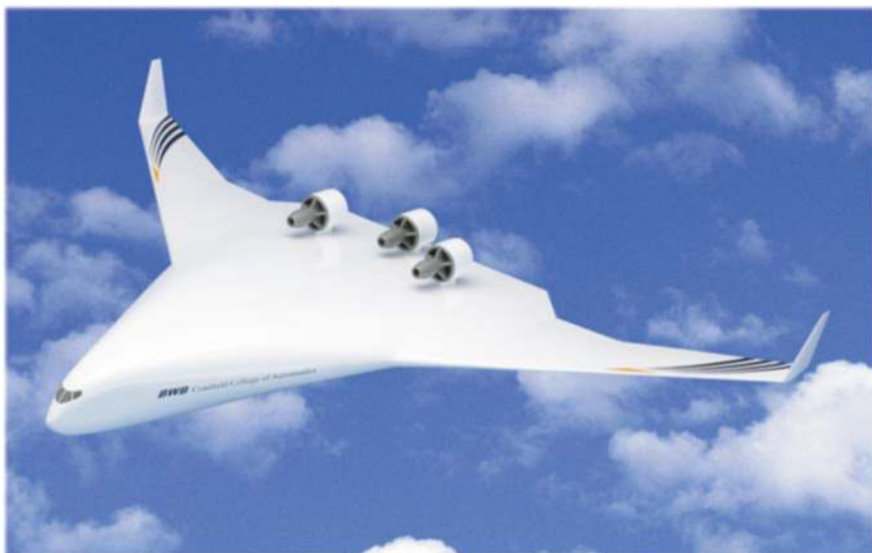


Fig. 16.6 Cranfield BW 98

MIT and Cambridge also performed design studies of blended-wing body aircraft [9].

It was felt that as these studies are examining BWB configurations, it would be useful to perform detailed examination of other down-selected low-noise configurations.

Current Cranfield work is studying joined wing aircraft, the broad deltas described below, and the more conventional A6 aft-engined aircraft is described at the end of this chapter.

16.3.5 Broad Delta Concepts

Reference [10] gives a detailed description of many airframe and engine concepts which were developed to meet low-noise goals.

The board delta (BD) was one of the most promising low-noise configurations. It has similarities to the baseline, as it has a conventional fuselage combined with

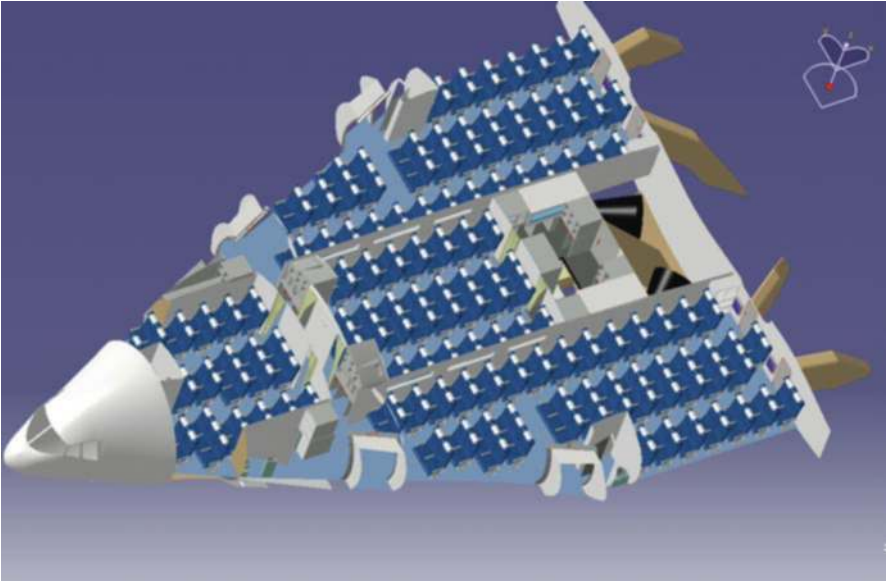


Fig. 16.7 Cranfield BW-01

a low aspect ratio delta wing. Two variants of this aircraft were designed in parallel to determine the effect that a stabilizing tail surface has both on noise and on performance (Fig. 16.9).

The BD appears to be similar to a conventional aircraft, but the large delta wing is closely integrated with the fuselage in a similar manner to the Avro Vulcan subsonic



Fig. 16.8 Kestrel

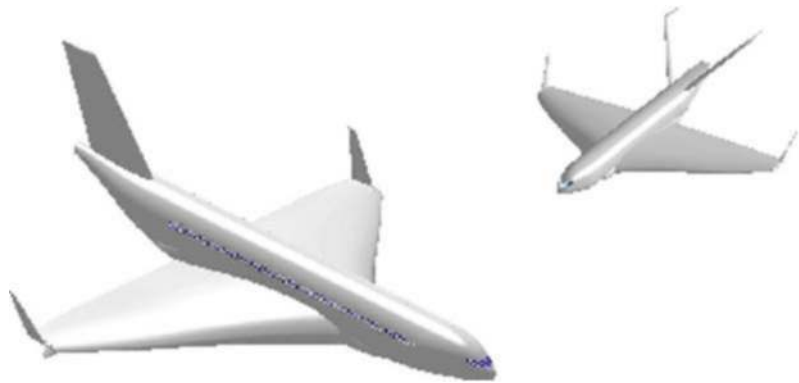


Fig. 16.9 BD concepts

strategic bomber. Such a configuration allows the partial or complete submerging of large-diameter power plants within the wing section. This has significant noise shielding and aerodynamic benefits.

Wing aspect ratio and taper ratios were optimized for minimum aircraft gross mass and incorporated the beneficial effects of winglets, which are significant at such low aspect ratios.

The fuselage is of a conventional cylindrical design, with a single-deck passenger layout. The SAI investigated leading edge (LE) carvings to provide additional lift at the nose of their BWB aircraft to reduce pitching moments. It was also decided to use this concept on the BD aircraft fuselage, as shown in Fig. 16.10.

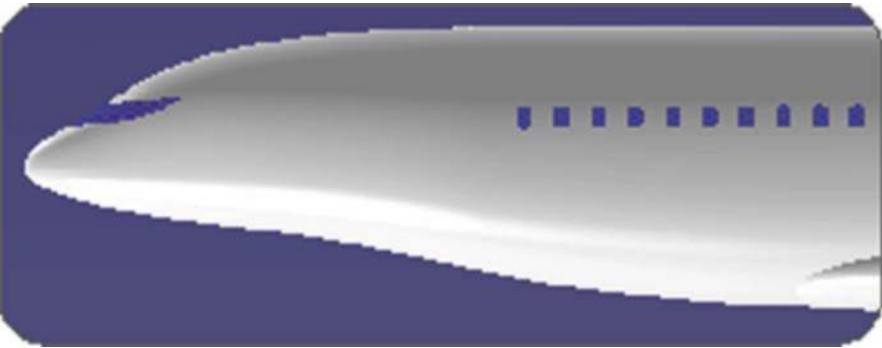


Fig. 16.10 “Carved” BD fuselage

Figure 16.11 shows an impression of the tailless aircraft, which incorporated semi-buried very high bypass ratio engines (16.20). Four engines were chosen rather than two, to keep engine diameters to acceptable levels.

A conventional landing gear fitted easily in the large wing root area. The V-tailed broad delta was more aerodynamically efficient, as the tail allowed the use of leading and trailing edge devices to increase lift coefficients for take-off



Fig. 16.11 Broad delta tailless

and landing. This allowed a significant increase in wing loading and cruise lift/drag ratio. This led to reduced fuel burn and mass. Figure 16.12 shows an early design of this aircraft, but it had not yet been fitted with power plants.

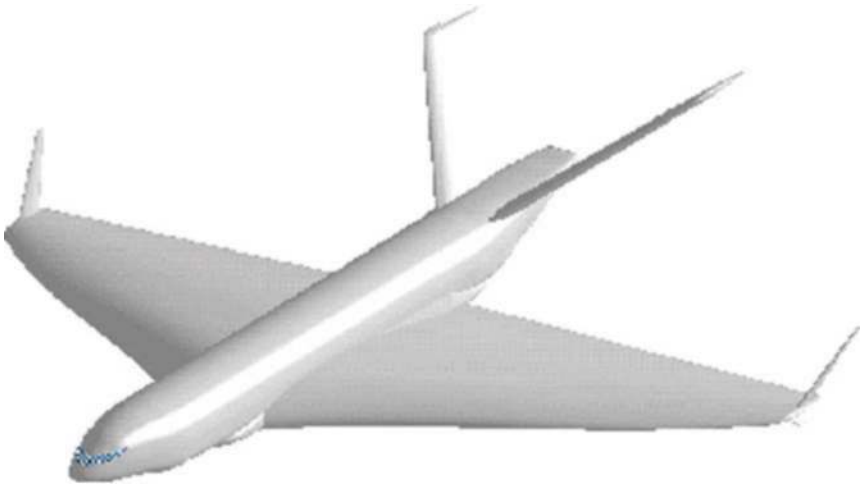


Fig. 16.12 V-Tail broad delta

16.3.6 Airframe Approach Noise Prediction

The selection of a steep flight path angle (FPA) was preferred because it allows a lower approach velocity thus reducing noise which is directly related by V_n .

The BD's reduced mass and wing loading, relative to the baseline, also contributed to approach velocity reduction.

Reducing approach velocity at high FPA, significantly reduces noise for the BD concepts compared with the baseline. Not only will there be a noise reduction due to velocity, but the BD tailless concept has no tail, and thus TE flaps or LE slats were not used. This removed three major noise sources of the airframe.

A V-tail BD had the noise disadvantage of a tail, but by using variable camber (slot-less) flaps, and drooped leading edge slats, approach noise will also be lower than that of the baseline airframe.

Noise reduction technologies may also be implemented to further reduce airframe component noise. Both BD concepts have two 4-wheel main bogies, and a twin nose wheel. Fairing the undercarriage components provides noise shielding of around 8–10 dB(A) [11]. Aerofoil noise may be reduced by using trailing edge brush technologies, which are claimed to reduce main wing noise by as much as 2 dB(A) [12]. These technologies may also be applied to conventional aircraft.

Increasing FPA from 3 to 6° for the tailless BD reduces airframe noise from 80.6 (A) to 73.5 dB(A) and the V-tail BD reduced airframe noise to 72 dB(A) using a 6° FPA.

The BD airframe noise for 6° approaches currently exceeds the SAI noise target of 60 dB(A) at the airport perimeter. Implementing undercarriage fairings and trailing edge brushes, however, could potentially reduce noise by a further 8–12 dB(A).

Even more noise reductions are achievable if a displaced threshold concept were used for landings. Baseline results suggest a reduction of 5 dB(A) by using a displaced threshold of 1 km.

It can thus be seen that the BD has the potential for meeting the extremely challenging 60 dBA target. Work is continuing on the prediction of take-off and sideline noise predictions of both airframe and engines – with promising initial results.

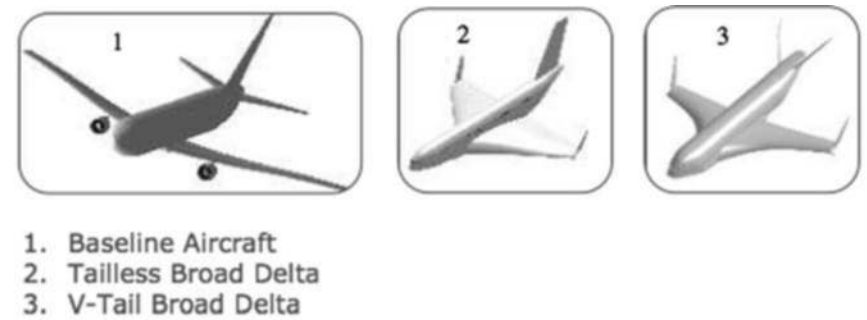
16.3.7 Performance Comparison

Consistent methods were used to compare the performance, mass, and noise characteristics for the baseline design and the two broad deltas. A significant exception, however, is that the power plant drag was not yet included for the broad deltas in Fig. 16.13.

BD noise performance was similar, but the tailed BD was chosen as the best BD configuration because of its better performance with respect to fuel burn and global warming.

It can be seen that the broad delta has significantly better noise characteristics than the baseline, with the potential for more improvements if fairings, brushes, and displaced thresholds are used. Such techniques, however, could also be used to improve conventional aircraft noise.

Another significant comparison is that for the same payload, range, and speed, the broad delta consumes only 80% of the baseline fuel with obvious global warming



	<i>Baseline</i>	<i>BD tailless</i>	<i>BD V-tail</i>
All up mass (AUM) [lb]	347,825	324,213	288,859
L/D	17.7	21.7	24.7
Mission fuel [lb]	96,139	76,096	61,490
Cruise CL	0.550	0.184	0.397
Approach velocity [m/s]	72.4	76,096	61,490
Noise at ICAO point [dB(A)]	93.8	73.5	72.0

Fig. 16.13 Low-noise performance comparisons. Note: the above figures do not include power plant drag for the broad deltas

benefits. This benefit will be reduced, when allowance is made for power plant drag – as is being currently investigated at Cranfield University.

16.4 The Cranfield A-6 Greenliner Project

16.4.1 Group Design Project Activities

Many Universities use group projects as powerful means of synthesizing aeronautical teaching and to give experience of design integration. Cranfield’s design course is unique in the magnitude of the student, staff, and equipment resources used in its annual projects. They allow significant in-depth design solutions at conceptual, preliminary, and detail design levels.

On the A-6 project, 50 graduate students were allocated responsibility for preliminary/detail designs of major parts of the aircraft such as the forward fuselage, a flying control surface, or an airframe system such as fuel, environmental control, propulsion, landing gear, avionics, or the control system. This allowed much more realistic estimates to be made of mass and performance and showed if the construction and system design methods were feasible.

This 8-month programme, using industry-standard design tools, was supervised by more than 10 faculty staff and was operated in what Cranfield term as a “virtual

industrial environment.” Some 50,000 man-hours of work are summarized in 50 project theses and in Stocking et al. [13].

16.4.2 Greenliner Description

The challenges of global warming and noise were the main motivating factors. The A-6’s design specifications and requirements were tailored to help the aviation industry to face up to the foreseen environmental challenges, both local and global, now and in the future. The aircraft conceptual design was performed by Smith [14] and formed the starting point for the group project. Four design variants were explored, namely composite and metallic airframe variants of aircraft with alternative V and U-tailed configurations. The chosen baseline aircraft was the V-tail composite variant (Fig. 16.14).



Fig. 16.14 Greenliner CAD model

The required performance is listed in Fig. 16.15. It can be seen to be a long-range, high-capacity aircraft in the class of the Boeing 777 or Airbus A330-600.

Great efforts were made to reduce fuel burn and aircraft noise, while ensuring low operating costs and high passenger comfort.

Specification	Values
<i>Airframe Life:</i>	70,000 hours (approximately 25 years)
<i>Design Mission:</i> Range Passengers	7500 nm 375 (Two Class)
<i>Accommodation & Capacity:</i>	
<i>Maximum Certified Capacity:</i>	440 (Single Class)
<i>Typical Configurations:</i>	400 Econ. (Single Class) 345 Econ. & 30 First Class (Two Class) 228 Econ., 54 Business & 24 First Class (Tri-Class)
<i>Cargo:</i>	32 LD-3 (Type A)
<i>Principal Geometry:</i> <i>Fuselage -</i> External Diameter Internal Diameter Overall length <i>Wing -</i> Span Aspect Ratio Gross Wing Area Leading edge sweepback Aerofoil Section	6.56 m 6.15 m 67 m 64 m 11.6 352.6 m ² 7.636° HSNLF(1)-0213 - Natural Laminar Flow (NLF)
<i>Mass:</i> Normal Take-off Max Landing Empty (OEM) Payload Fuel load	209,410 kg 168,500 kg 110,465 kg 35,625 kg 63,320 kg
<i>Power Plant:</i> Model	Rolls-Royce Trent 500 Derivative
Thrust (SL Static Rating)	266.2 kN
<i>Field Performance:</i> Max. Certificated Runway FAR Landing Distance	2500m (Max. All Up Mass takeoff @ ISA SL) 1752m (Max. Landing mass @ ISA SL)
<i>Cruise Speed:</i>	Mach 0.74
<i>Cabin Altitude:</i>	< 5500ft (average) to 7000ft (max)
<i>Design Requirements:</i>	EASA CS25 (Certification Specifications for Large Aeroplanes) AVD DES 0600/1 A-6 Project Specifications
<i>Specification Structural Limitations:</i>	+3.2g / -1.2g (SL, OEM, Gust Velocity 17.07 m/s)
<i>Design Objectives:</i>	To minimise the environmental impact of the operation of aircraft in the broadest sense including the reduction of both global and local impacts, better cabin comfort and considerations for sustainability issues.

Fig. 16.15 A-6 design specification

The wing was designed to use a natural laminar flow (NLF) aerofoil section with a very high aspect ratio to achieve drag reduction and thus, reduce the fuel consumption. This also achieves a large direct reduction of carbon emissions. This design decision led to a very low wing sweep which limited the cruise Mach number to 0.74. The aircraft’s engines were high mounted on the aft fuselage with a “butterfly” tail to provide noise shielding. Great efforts were also made by all team members to achieve weight reductions through trade-off studies, in order to attain better aircraft performance. Sustainability and weight reduction were the main factors considered during the material selection processes. Opportunities for better comfort and health of passengers during flight were also explored during the design of A-6. These took the form of significantly increased cabin size, improved seating, combined with a higher humidity and lower cabin altitude fuselage (Fig. 16.16). Figure 16.17 shows an example of the extensive finite element structural analysis that was used to aid design. This example is from the rear fuselage.

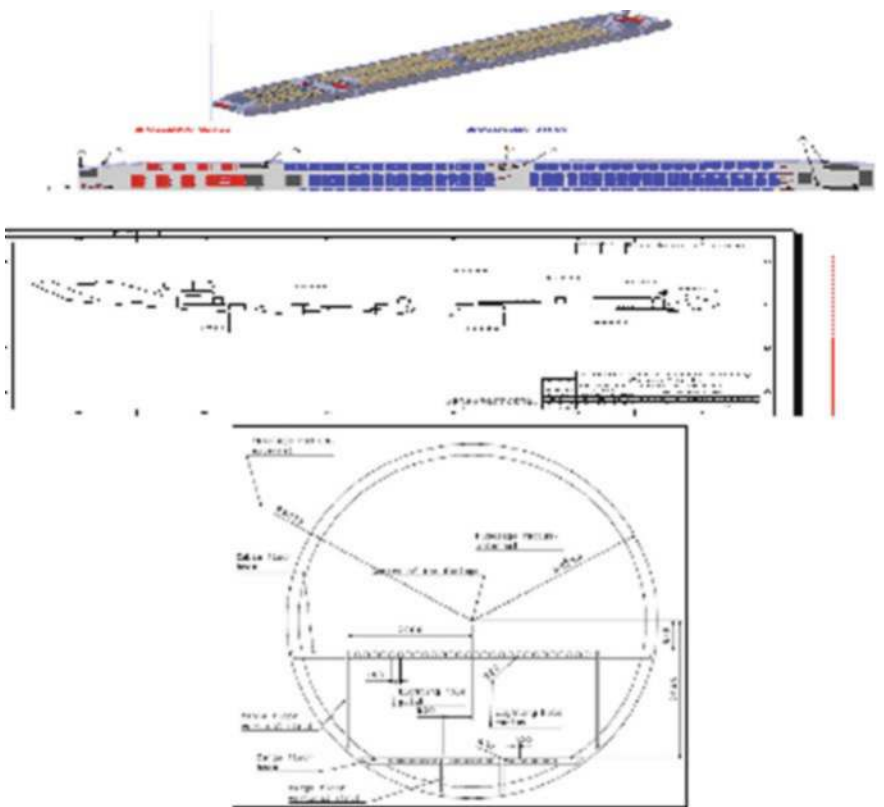


Fig. 16.16 A-6 Fuselage

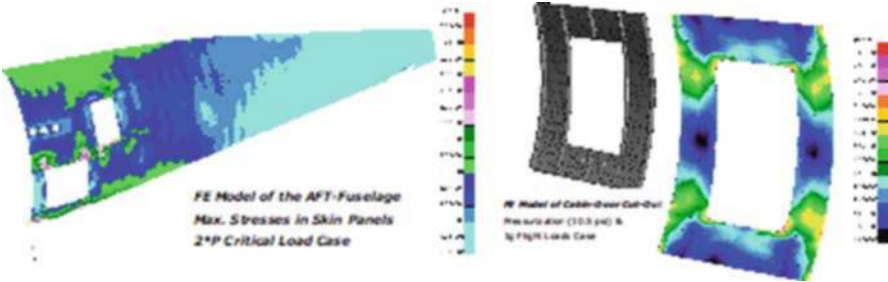


Fig. 16.17 Rear fuselage finite element model

A number of power plant options were examined and Fig. 16.18 shows the base-line Rolls-Royce Trent 500. The intake utilizes the negative scarf concept to reduce intake noise.

All airframe systems were designed, and Fig. 16.19 shows the main landing gear, while the schematic of the environmental control system is shown in Fig. 16.20. Extensive work was performed for the avionics systems, the displays of which are shown in Fig. 16.21.

Great attention was paid to aircraft certification requirements, safety, reliability, and maintainability. Figure 16.22 shows the CAD model used to assess wing system accessibility.

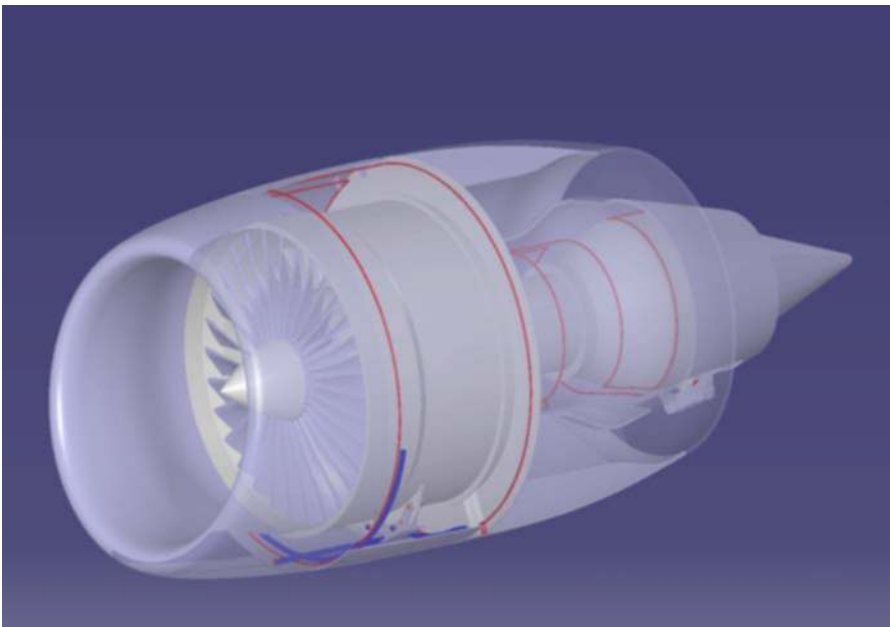


Fig. 16.18 Power plant

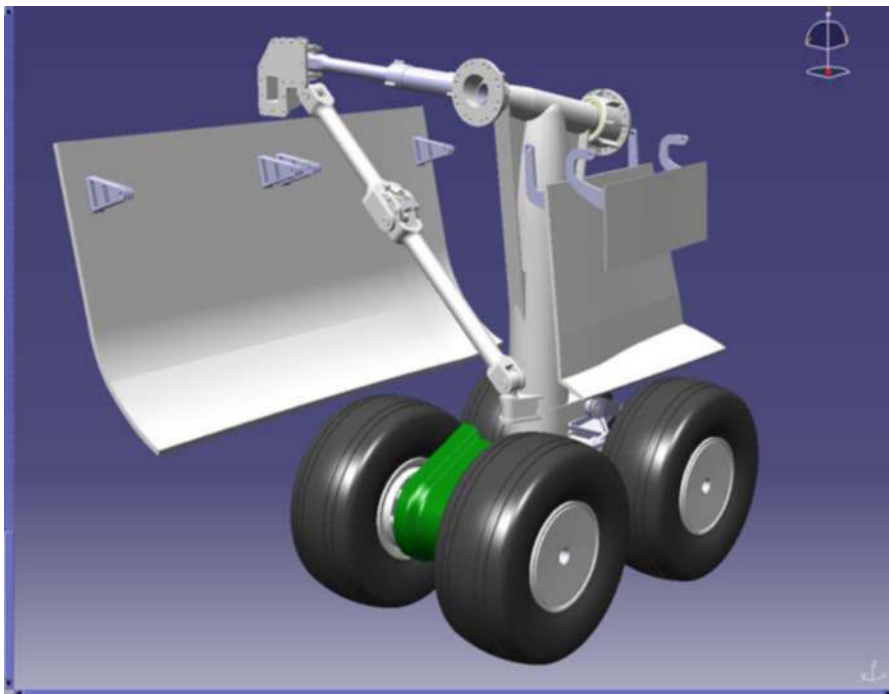


Fig. 16.19 Main landing gear

16.4.3 Predicted Performance for the Greenliner

Extensive calculations showed that the A-6 should meet or exceed its specified performance targets. The composite structure, V-tail derivative offered the most mass reduction and minimum fuel burn and pollution.

Work is continuing to verify mass estimates and to see how closely the A-6 can reach or exceed the ACARE targets mentioned above.

Acquisition and operating costs estimates have shown that the aircraft should be competitive with aircraft planned to be operated in the 2020 timeframe.

The main drawback of the aircraft is the fact that its cruise speed is some 90% of current aircraft values. This will lead to small increases in flight times at short and medium ranges, but becomes more significant at long ranges. The increased cabin dimensions, lower cabin altitude, and advanced avionics will make the aircraft more comfortable.

This should give some compensation for the increased journey times, as should the lower fares than those for other aircraft that would follow if carbon taxes were to be imposed.

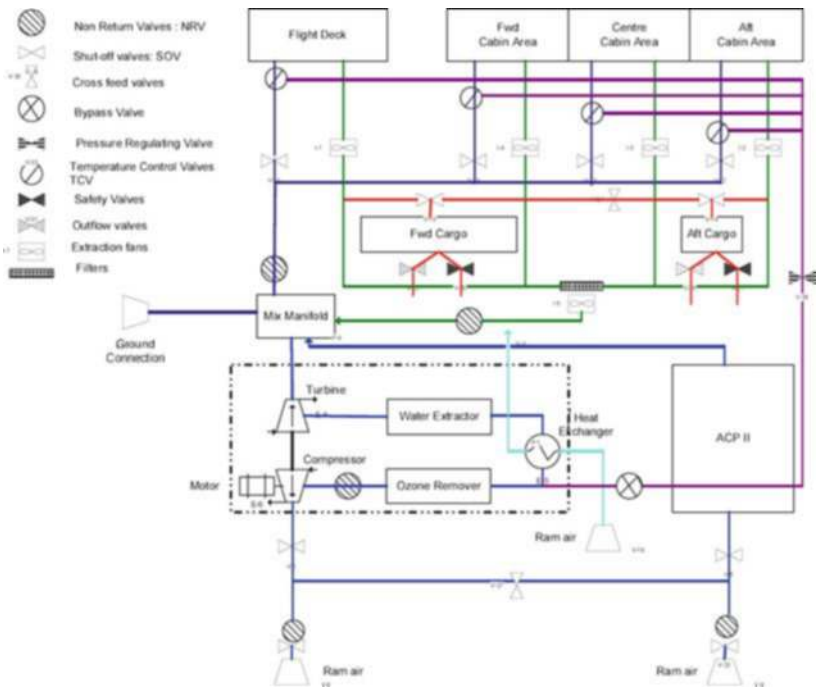


Fig. 16.20 Environmental control system

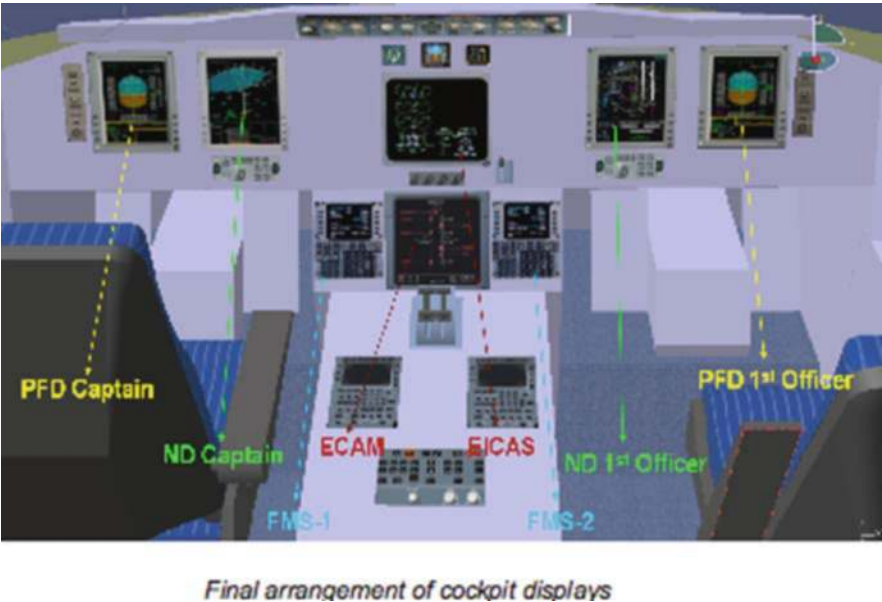


Fig. 16.21 Flight deck

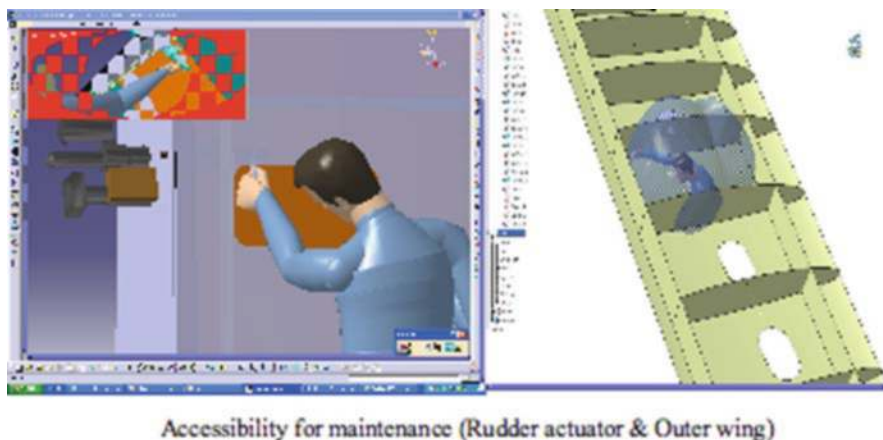


Fig. 16.22 Wing maintenance access

16.5 Conclusions

1. Aircraft-related noise, local air quality, and global warming are recognized as increasingly serious issues. They must be quickly addressed by new operational, economic, and technological means.
2. The current study has shown that realistic global warming costs are about one-third of fuel costs of long- and short-haul aircraft. Noise and local air quality costs are some 4 and 16% of fuel costs, respectively. Further work is required to substantiate these values.
3. The Silent Aviation Initiative noise targets can be met, if novel propulsion, airframe configurations, and advanced technologies are developed.
4. Current aircraft configurations and technologies can be developed incrementally to improve environmental impact, but current traffic growth is likely to increase pollution more than these savings. It is unlikely that the KYOTO and ACARE targets can be met by such means.
5. Intermediate aircraft configurations, such as the Cranfield A-6 Greenliner, should be able to significantly reduce costs, noise, and global warming. This is at modest risk and at lower cruise speeds.
6. The broad delta appears to be an alternative low-noise and emissions aircraft but at more project risk. The blended-wing body aircraft appears to offer even more performance advantages, but at more technical and economic risk.
7. Research should continue into new technologies and configurations, so that the claimed advantages may be substantiated. It is important to arrive at aircraft which simultaneously optimize cost, global warming and noise, and meet targets such as those proposed by ACARE.

References

1. A. Turner: Changed Climate – Special Report, in Flight International Magazine, 19–25th June 2007.
2. J. P. Fielding, H. Smith,: Development of Environmentally-Friendly Technologies and Configurations for Subsonic Jet Transports, International Congress of the Aeronautical Sciences (ICAS) 2002, Toronto, Canada.
3. J. E. Green, et al.: Air Travel-Greener by Design. Mitigating the Environmental Impact of Aviation: Opportunities and priorities. The Royal Aeronautical Society. July 2005
4. Aviation and the Environment using Economic Instruments: UK Department of Transport, ISBN 1 851126139. March 2003.
5. Aviation and Global Warming. UK Department of Transport. January 2004.
6. Climatecare website. www.climatecare.co.uk Accessed April 2006.
7. S. Mistry, G. Doulgeris, J. P. Fielding, P. Pilidis: Development of Silent Airframe Concepts and Innovative Cycle Propulsion Systems for Reduction of Aircraft Noise, International Congress of the Aeronautical Sciences (ICAS) 2006, Hamburg, Germany.
8. D. Howe, Aircraft Conceptual Design Synthesis, PE Publishing, UK, 2000.
9. J. I. Hileman, Z. S. Spakovszky, M. Drela, M. A. Sargeant: Aircraft Design for Silent Aircraft, 45th AIAA Sciences Meeting and Exhibit, Reno, Nevada, 8–11th January 2007.
10. G. Doulgeris, S. Mistry, J. P. Fielding, P. Pilidis: Development of a Broad Delta Airframe and Propulsion Concepts for Reducing Aircraft Noise around Airports, Paper 2007-01-3806, Society of Automotive Engineers (SAE) AeroTech Congress and Exhibition 2007, Los Angeles, CA, 17–20th September 2007.
11. A. Quayle, et. al.: Landing Gear for a Silent Aircraft, AIAA-2007-231, 45th AIAA Sciences Meeting and Exhibit, Reno, Nevada, 8–11th January 2007.
12. M. Herr, W. Dobrgynski: Experimental Investigations in Low Noise Trailing Edge Design, AIAA-2004-2804, 10th AIAA/CEAS Aeronautical Conference, Manchester 2004.
13. P. Stocking et al.: A-6 Greenliner Environmentally Benign Aircraft, Project Executive Summary, Cranfield University, October 2006.
14. H. Smith: A-6 Project Specification, Cranfield University, October 2006.

Chapter 17

Variational Approach to the Problem of the Minimum Induced Drag of Wings

Maria Teresa Panaro, Aldo Frediani, Franco Giannessi and Emanuele Rizzo

Abstract A closed form solution of the problem of the minimum induced drag of a finite span straight wing was given by Prandtl. In this chapter, a mathematical theory, based on a variational approach, is proposed in order to revise such a problem and provide one with a support for optimizing more complex wing configurations, which are becoming of interest for future aircraft. The first step of the theory consists in finding a class of functions (lift distributions) for which the induced drag functional is well defined. Then, in this class, the functional to be minimized is proved to be strictly convex; taking into account this result, it is proved that the global minimum solution exists and is unique. Subsequently, we introduce the *image space analysis* associated with a constrained extremum problem; this allows us to define the Lagrangian dual of the problem of the minimum induced drag and show how such a dual problem can supply a new approach to the design. After having obtained the Prandtl exact solution in the context of a variational formulation, a numerical algorithm, based on the Ritz method, is presented, and its convergence is proved.

Maria Teresa Panaro

Dipartimento di Matematica, “L. Tonelli,” Università di Pisa, Pisa, Italy,

e-mail: mt.panaro@gmail.com

Aldo Frediani

Dipartimento di Ingegneria Aerospaziale, “L. Lazzarino,” Università di Pisa, Pisa, Italy,

e-mail: a.frediani@ing.unipi.it

Franco Giannessi

Dipartimento di Matematica, “L. Tonelli,” Università di Pisa, Pisa, Italy,

e-mail: gianness@dm.unipi.it

Emanuele Rizzo

Dipartimento di Ingegneria Aerospaziale, “L. Lazzarino,” Università di Pisa, Pisa, Italy,

e-mail: emanuele.rizzo@ing.unipi.it

17.1 Introduction

The main parts of aerodynamic drag of an aircraft are the friction and the induced ones. Friction drag depends on the wetted surface; induced drag depends on the lift distribution along the lifting systems. According to the theorem of Kutta (1867–1914) and Jukowski (1847–1921), lift equals the product of asymptotic speed, density, and circulation (or vorticity). Ludwig Prandtl (1875–1953) and his collaborators at the University of Göttingen gave a significant contribution to aerodynamics. In the case of finite width straight wings, due to a theorem on the conservation of vorticity, the vorticity variation along the wing equals the vorticity released along the stream, which, in its turn, produces an induced velocity on the wing, in accordance with the Biot and Savart law. Prandtl gave a solution to the variational problem of assessing the lift distribution for which, given the total lift, the induced drag is a minimum. For this problem, the optimality condition implies a constant induced velocity along the wing span and the elliptical lifting distribution satisfy this condition. This result was fundamental in the history of aviation: all actual aircraft are designed in order to obtain an elliptical lift distribution as close as possible.

From these considerations, Prandtl's problem is considered again, with the aim of introducing an extensive mathematical analysis of the problem, taking into account recent results of the theory of constrained extremum problems, in particular, the image space analysis. This will lead, among other things, to formulate a Lagrangian dual problem of that of minimum induced drag. It is shown that such a dual problem is a new approach to the primary problem.

17.2 Finite Span Wings

In a steady, subsonic, and two-dimensional stream, the aerodynamic force acting on a solid body is given by

$$D = 0, \quad P = \rho \mathbf{V}_\infty \Gamma,$$

where D is the component along the asymptotic stream direction and P is the normal to D ; ρ and \mathbf{V}_∞ are the density and the asymptotic velocity, respectively. This result is known as the Kutta–Joukowski theorem and, accordingly, the drag on a profile is zero, independently of the lift. This result, even for inviscid fluids, is no longer valid when dealing with a finite span wing, where a drag induced by the lift (and, hence, named “induced drag”) is present due to three-dimensional effects.

In a finite span wing, the pressure difference between upper and lower sides produces tip horse shoe vortices which, in turn, on any wing section, induce a vertical downstream according to the well-known Biot–Savart law. Thus, in any section of the wing, the angle of incidence is locally modified by an angle α_i (of induced incidence), as shown in the following figure.

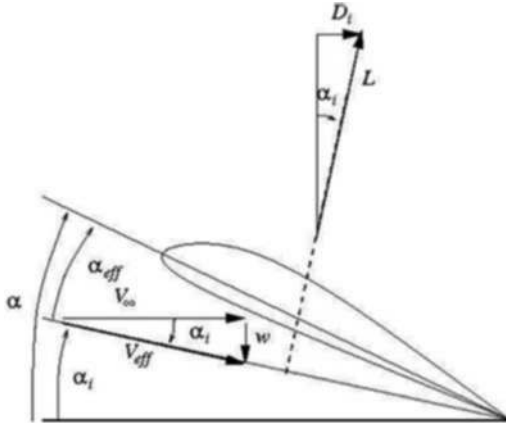


Fig. 17.1 α , geometric angle of incidence; α_i , angle of induced velocity; α_{eff} , actual angle of incidence.

With small α_i , the forces orthogonal to the stream (lift) and along the stream direction (induced drag) are

$$\begin{cases} L = F \cos \alpha_i = \int_{-b}^b \rho \Gamma V_\infty dy, \\ D_i = F \sin \alpha_i = \rho \int_{-b}^b w(y) \Gamma dy, \end{cases} \quad (17.1)$$

where $2b$ is the wingspan.

In the case of a large span/cord ratio, the wing can be assumed as a lifting line undergoing a circulation distribution along the span (Prandtl). According to the Biot–Savart law, the velocity induced in y_0 by an elementary free vortex $d\Gamma = \left(\frac{d\Gamma}{dy}\right) dy$ is given by the following:

$$dw(y_0) = \frac{1}{4\pi} \left(\frac{d\Gamma}{dy} \frac{dy}{y_0 - y} \right), \quad (17.2)$$

and velocity induced by the whole vorticity becomes

$$w(y_0) = \frac{1}{4\pi} \int_{-b}^b \frac{d\Gamma}{dy} \frac{dy}{y_0 - y}. \quad (17.3)$$

Combining the previous results, we have the induced drag

$$D_i = \frac{\rho}{4\pi} \int_{-1}^1 \int_{-1}^1 \frac{\Gamma'(x)\Gamma(y)}{y-x} dx dy.$$

Note that the double integral is defined by its principal value of Cauchy (see Sect. 17.6).



Fig. 17.2 Vortices on the wing wake: on the lifting line, the difference between vortices (or circulation) equals the free vortices detaching from trailing edge

17.3 Problem of Minimum Induced Drag of a Straight Wing: An optimality condition

Denote by \mathbb{R} and \mathbb{R}_+ the sets of reals and non-negative reals, respectively. We consider a straight wing, which is assumed to be a lifting segment of the real line; it is not restrictive to represent the segment by $T := [-1, 1] \subset \mathbb{R}$ (see figure 17.3).

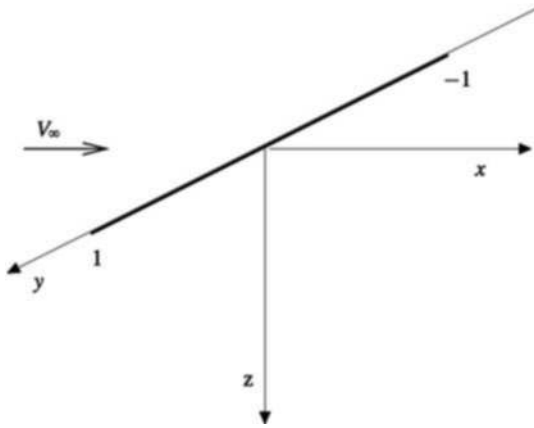


Fig. 17.3 Reference frame of the lifting line wing

Let Ω be an open set of \mathbb{R} , such that $\Omega \supset T$. We wish to determine a function $\Gamma : \Omega \rightarrow \mathbb{R}_+$, which minimizes the induced drag, denoted by $f(\Gamma)$, subject to a constraint on the total lift, denoted by $g(\Gamma)$. Let us now give this problem a mathematical formulation. To this end, let χ be a set of functions Γ , where a solution is sought. Thus, the problem can be formulated as

$$\min f(\Gamma) := \frac{\rho}{4\pi} \int_T \int_T \frac{\Gamma'(x)\Gamma(y)}{y-x} dx dy, \quad (17.4)$$

subject to

$$g(\Gamma) := \rho V_\infty \int_T \Gamma dx - c = 0, \quad (17.5)$$

$$\Gamma \in \chi, \quad (17.6)$$

where \min denotes the global minimum and c is a positive constant; “:=” denotes “equality by definition.”

First of all, the elements of χ must make f positive and satisfy given boundary conditions $\Gamma(-1) = \Gamma(1) = 0$. Furthermore, Γ must be such that the functionals f and g exist. This is guaranteed by the assumptions of Proposition 17.1 of Appendix 1. Such assumptions allow us to define f on the subset of $T \times T$ where $x = y$, giving it the value of its limit, so that the functional so extended – which, without any danger of confusion, will be denoted by the same f – turns out to be continuous.

Having restricted χ in order to guarantee the existence and the continuity of f and g , we now must assure the existence of the (constrained) minimum in (17.4), (17.5) and (17.6). This can be achieved in several ways. To this end, we introduce the following norm:

$$\|\Gamma\| := \max_{x \in T} \{ \|\Gamma(x)\|_2, \|\Gamma'(x)\|_2 \},$$

where $\|\cdot\|_2$ is the L^2 norm for the present problem.

A first way consists in restricting χ to be compact with respect to the above norm; a simple (but sufficient for the design of a wing) example is given by a set of bounded polynomials on T , with degree not greater than a fixed value. The assumptions of the Lebesgue Fundamental Theorem being fulfilled (Lebesgue measurability and uniform boundedness of each sequence), then the set of solutions to (17.5) is closed. Taking into account that a closed subset of a compact set is compact, the set of solutions to (17.5) and (17.6) is compact. This and the continuity of f give the existence of the minimum.

A second way consists in proving the strict convexity of the functional f (see Theorem 17.2 in Appendix 1) and show that, within the stated class χ , the first variation vanishes. Of course, this depends on the fact that the first variation vanishes on χ ; otherwise, nothing can be said, unless the convexity of χ is proved; but this is not an easy task.

Condition 17.1 *The minimum of problem (17.4), (17.5) and (17.6) is an increasing function of c , which tends to $+\infty$ as c tends to $+\infty$.*

The property expressed by the above condition, even if intuitively obvious from engineering viewpoint, is of no easy mathematical proof. Indeed, due to its importance in the analysis of the solution of (17.4), (17.5) and (17.6) when c is a parameter (and not a given number) as happens in Sect. 17.4, we should prove it; to avoid an excessive mathematical machinery, in the sequel we will assume it.

Having discussed the existence of the minimum of problem (17.4), (17.5) and (17.6), let us now consider an optimality condition. In Appendix 1 it is proved that

a class, where it is suitable to look for the circulation distribution Γ as solution of the minimum problem (17.4), (17.5) and (17.6), is the following:

$$\mathcal{X} = \{\Gamma \in AC[-1, 1], \Gamma' \in \mathcal{L}^{1+\varepsilon}(-1, 1), \text{ with } \varepsilon > 0, \Gamma(1) = 0, \\ \Gamma(-1) = 0, D_i(\Gamma) > 0\}.$$

Here, according to what was done by Munk [4], we can prove a necessary and sufficient optimality condition for Γ .

Theorem 17.1. *Let be $\Gamma \in \mathcal{X}$. Γ is solution of the isoperimetric problem (17.4), (17.5) and (17.6), if and only if $w(y) = \frac{1}{4\pi} \int_{-1}^1 \frac{\Gamma'(x)}{x-y} dx = \text{constant}$, $\forall y \in [-1, 1]$.*

Proof. Let $\Gamma, \Gamma_* \in \mathcal{X}$, and set $\delta\Gamma(x) := \Gamma - \Gamma_*$, with $\|\Gamma_*\| < \varepsilon$, $\varepsilon > 0$. Introduce the functions

$$\Gamma(z, \alpha) := \Gamma(z) + \alpha\delta\Gamma(z),$$

$$\Gamma'(z, \alpha) := \Gamma'(z) + \alpha\delta\Gamma'(z).$$

$\Gamma'(z, \alpha)$ is the derivative of $\Gamma(z, \alpha)$ with respect to z . Consider $J(\lambda)$ the functional

$$J(\lambda) := \int_T \int_T \left[\frac{\rho}{4\pi} \frac{\Gamma'(x)\Gamma(y)}{y-x} - \frac{\lambda\rho V_\infty}{2} \Gamma(y) \right] dx dy, \quad (17.7)$$

and let us calculate the variation

$$\begin{aligned} J(\lambda, \alpha) &:= \int_T \int_T \left[\frac{\rho}{4\pi} \frac{\Gamma'(x, \alpha)\Gamma(y, \alpha)}{y-x} - \frac{\lambda\rho V_\infty}{2} \Gamma(y, \alpha) \right] dx dy \\ &= \int_T \int_T \frac{\rho}{4\pi} \frac{(\Gamma'(x) + \alpha\delta\Gamma'(x))(\Gamma(y) + \alpha\delta\Gamma(y))}{y-x} dx dy + \\ &\quad - \int_T \int_T \frac{\lambda\rho V_\infty}{2} (\Gamma(y) + \alpha\delta\Gamma(y)) dx dy. \end{aligned}$$

When the derivative of J with respect to α is evaluated in zero, we find

$$\begin{aligned} J'(\lambda, 0) &= \int_T \int_T \left[\frac{\rho}{4\pi} \frac{\delta\Gamma'(x)\Gamma(y) + \delta\Gamma(y)\Gamma'(x)}{y-x} - \frac{\lambda\rho V_\infty}{2} \delta\Gamma(y) \right] dx dy \\ &= \int_T \int_T \left[\frac{\rho}{4\pi} \frac{\delta\Gamma(y)\Gamma'(x)}{y-x} - \frac{\lambda\rho V_\infty}{2} \delta\Gamma(y) \right] dx dy + \\ &\quad + \int_T \int_T \frac{\rho}{4\pi} \frac{\delta\Gamma'(x)\Gamma(y)}{y-x} dx dy. \end{aligned} \quad (17.8)$$

Integrating by parts the second term of Eq. (17.8) leads to

$$\begin{aligned} \frac{\rho}{4\pi} \int_T \int_T \frac{\delta \Gamma'(x) \Gamma(y)}{y-x} dx dy = & \left[\frac{\rho}{4\pi} \delta \Gamma(x) \int_T \frac{\Gamma(y)}{y-x} dy \right]_{-1}^1 \\ & - \frac{\rho}{4\pi} \int_T \delta \Gamma(x) \frac{d}{dx} \int_T \frac{\Gamma(y)}{y-x} dx dy. \end{aligned} \quad (17.9)$$

The first term in the right-hand side of Eq. (17.9) is null because $\delta \Gamma(-1) = \delta \Gamma(1) = 0$. Having put $t = y - x$, we have

$$\frac{d}{dx} \int_T \frac{\Gamma(y)}{y-x} dy = \frac{d}{dx} \int_T \frac{\Gamma(t+x)}{t} dt = \int_{-1-x}^{1-x} \frac{\Gamma'(t+x)}{t} dt.$$

Coming back to the previous variables, we find

$$\frac{d}{dx} \int_T \frac{\Gamma(y)}{y-x} dx = \int_T \frac{\Gamma'(y)}{y-x} dy. \quad (17.10)$$

By exchanging y with x in the right-hand side of Eq. (17.9) and remembering (17.10), the right-hand side of (17.8) becomes

$$J'(\lambda, 0) = \int_T \int_T \delta \Gamma(y) dy \left(\int_T \left(\frac{\rho}{2\pi} \frac{\Gamma'(x)}{y-x} - \frac{\lambda \rho V_\infty}{2} \right) dx \right). \quad (17.11)$$

Since $J'(\lambda, 0) = 0$, a sufficient condition is

$$w(y) = \frac{1}{4\pi} \int_T \frac{\Gamma'(x)}{x-y} dx = \text{constant}, \quad \forall y \in [-1, 1]. \quad (17.12)$$

This condition is necessary as well, after observing that the functional is convex; in fact, the second derivative of J with respect to α is

$$J''(\lambda, \alpha) = \frac{\rho}{4\pi} \int_T \int_T \frac{\delta \Gamma'(x) \delta \Gamma(y)}{y-x} dx dy, \quad (17.13)$$

where the quantity at the right-hand side is the elementary induced drag D_i , due to the lift and, therefore, is positive. \square

17.4 Duality: A New Approach to the Design of Wings

Now, we want to introduce the dual problem of (17.4), (17.5) and (17.6). To this end, let us say first of all something about the birth of duality.

A general feature of duality (it would be better to say dualism) consists in two entities, which express a sort of symmetry or complementary. In the field of optimization, such entities are a pair of constrained extremum problems. An early trace

of this – perhaps, the first – is due to Vecten, and, independently, to Fasbender (see [1, 2]) with reference to Fermat–Torricelli problem on a triangle (which consists in finding a point of a triangle – now called Torricelli point – which minimizes the sum of its distances from the vertices; the given problem is called *primal*): among all the equilateral triangles, which are circumscribed to a given triangle, to find one having maximum height; they showed that such a maximum height equals the minimum sum of the distances of Torricelli point from the vertices of the given triangle (see [1], page 235). Fermat–Torricelli and Vecten–Fasbender problems are a pair of constrained extremum problems, which enjoy the following properties:

- (i) they are defined by the same data;
- (ii) they search for opposite extrema;
- (iii) the values of their objective functions, corresponding to feasible solutions, form two sets of real numbers, which are separated;
- (iv) the two extrema are equal; the common value of the two extrema being, therefore, the separating element of two contiguous classes of real numbers.

The above problems enjoy further properties, which Vecten and Fasbender did not observe (and, perhaps, could not have noted at that time):

- (v) *relaxation*: the dual is equivalent to search, in the primal, for the best lower bound of the objective function obtained by *relaxing* the feasible region (of course, if the primal searches for the maximum – like in the problem of this chapter – then relaxation must be replaced by *contraction*);
- (vi) *reflexivity*: the dual of the dual problem is (equivalent to) the primal.

The result by Vecten and Fasbender has marked the birth of duality theory for constrained extremum problems. Subsequently, a few results appeared till when John von Neumann claimed the above (i)–(iv) for a linear programming problem. After von Neumann result, the theory of duality grew quickly; it achieved the present general form, when it was recognized to be a step of the image space analysis carried on through Hahn–Banach separation theory. Appendix 2 contains a short outline of image space analysis and on how it can lead to discover the theory of duality. Here, by a logic-intuitive way, we merely consider the essential steps to achieve the dual problem of (17.4), (17.5) and (17.6) which, for symmetry of language, is called *primal*.

Denote by $R := \{\Gamma \in \chi : \rho V \int_T \Gamma(x) dx - c = 0\}$ the feasible region of (17.4), (17.5) and (17.6). Let us start with the obvious remark that $\bar{\Gamma} \in R$ is a (global) minimum point of (17.4), (17.5) and (17.6), iff the system (in the unknown Γ ; the notation is the same as in Sect. 17.4)

$$u := f(\bar{\Gamma}) - f(\Gamma) > 0, \quad v := g(\Gamma) = 0, \quad \Gamma \in \chi \quad (17.14)$$

is impossible. u and v run in the images of χ through the functionals $f(\bar{\Gamma}) - f$ and g , respectively. Therefore, the space where Γ runs is paired with \mathbb{R}^2 where (u, v) runs; this \mathbb{R}^2 is called the *image space* associated with (17.4), (17.5) and (17.6); the set

$$\mathcal{H}_{\bar{\Gamma}} := \{(u, v) \in \mathbb{R}^2 : u = f(\bar{\Gamma}) - f(\Gamma), \quad v = g(\Gamma), \quad \Gamma \in \chi\}$$

is called the image set of (17.4), (17.5) and (17.6). By introducing the set

$$\mathcal{H} := \{(u, v) \in \mathbb{R}^2 : u > 0, v = 0\},$$

which mirrors the conditions in (17.14), we can say that $\bar{\Gamma} \in R$ is a (global) minimum point of (17.4), (17.5) and (17.6), iff

$$\mathcal{H} \cap \mathcal{K} = \emptyset. \quad (17.15)$$

It is trivial to note that (17.14) is impossible, iff (17.15) holds. While (17.14) has an algebraic appeal, (17.15) offers a geometrical approach. In fact, the disjunction (17.15) can be proved, by showing that there exists a line (of \mathbb{R}^2), say H^0 , such that \mathcal{H} and \mathcal{K} lie, respectively, in the halfplanes, say H^+ and H^- , the former open and the latter closed, defined by H^0 . Thus, taking into account that \mathcal{H} is a halfline (of \mathbb{R}^2) deprived of the vertex and \mathcal{K} can be replaced equivalently by a convex set (see Appendices 1, 2) and defining H^0, H^- and H^+ , respectively, by

$$\theta u + \lambda v = 0, \quad \theta u + \lambda v \leq 0, \quad \theta u + \lambda v > 0, \quad (\theta, \lambda) \in \mathbb{R}^2 \setminus \{0\}, \quad (17.16)$$

it is easy to note that (17.15) is equivalent to the existence of $\bar{\lambda} \in \mathbb{R}$, such that

$$\bar{\theta} u + \bar{\lambda} v \leq 0, \quad \forall (u, v) \in \mathcal{K}. \quad (17.17)$$

Due to the homogeneity of the inequalities (17.16) and (17.17), we might set $\theta = 1$; this is not done, because of the meaning that θ and λ will have in the subsequent application. In other words, \mathcal{H} being included by definition in H^+ , (17.15) holds if and (because of the convexity of both \mathcal{H} and \mathcal{K}) only if (17.17) holds. Now, by recalling the definition of u and v in (17.14), (17.17) turns out to be equivalent to the existence of $\bar{\theta} > 0$ and $\bar{\lambda}$, such that

$$\mathcal{L}(\bar{\Gamma}; \bar{\theta}, \bar{\lambda}) \leq \mathcal{L}(\Gamma; \bar{\theta}, \bar{\lambda}), \quad \forall \Gamma \in \chi, \quad (17.18)$$

where

$$\mathcal{L}(\Gamma; \theta, \lambda) := \int_T \int_T \left[\theta \frac{\rho}{4\pi} \frac{\Gamma'(x)\Gamma(y)}{y-x} - \lambda \rho V_\infty \frac{\Gamma(x)}{2} \right] dx dy \quad (17.19)$$

is the *Lagrangian function*.

In fact, $\bar{\Gamma} \in R$ implies $g(\bar{\Gamma}) = 0$, so that $\mathcal{L}(\Gamma; \bar{\theta}, \bar{\lambda}) = \theta f(\bar{\Gamma})$ and the inequality (17.18) is equivalent to (17.17).

We have thus shown that the fulfillment of inequality (17.18) is equivalent to prove the optimality of $\bar{\Gamma}$. However, to verify (17.18) is not an easy task. Therefore, in order to prove (17.18), we are led to evaluate, for each (θ, λ) , the minimum (in general, infimum) of $\mathcal{L}(\Gamma; \theta, \lambda)$ with respect to $\Gamma \in \chi$ (and not $\Gamma \in R$), and then the maximum (supremum, in general) of such minimum (which, obviously, depends

on θ, λ) with respect to θ, λ . In other words, we are led to introduce the following problem:

$$\max_{\theta > 0, \lambda \in \mathbb{R}} \min_{\Gamma \in \chi} \mathcal{L}(\Gamma; \theta, \lambda), \quad (17.20)$$

which is called *dual problem* of (17.4), (17.5) and (17.6). Due to the special properties of (17.4), (17.5) and (17.6), it is possible to prove that (17.20) enjoys the properties (i)–(iv); see Appendix 2.

The image space analysis allows one to achieve several other important properties and informations; see Appendix 2. In particular, we can draw that a solution $(\bar{\Gamma}; \bar{\theta}, \bar{\lambda})$ of the dual problem (17.19) enjoys this property: $\bar{\lambda}/\bar{\theta}$ is nothing more than the classic Lagrangian multiplier and *allows one to evaluate the change in the minimum induced drag consequent to a change in the value at which the total lift is constrained*.

Now, we are in the position to discuss a different approach to the design of the wing. As discussed in details in Appendix 2, to consider the induced drag as an objective (to be minimized) and the total lift as a constraint is absolutely subjective. An alternative approach, which does not oblige us to consider one of the two entities as constraint, is the following. *We consider both entities as objectives*, in the sense that *we aim to minimize the induced drag and to maximize the total lift* or, equivalently, to minimize the opposite of the total lift. To try to fulfill both objectives, we consider a combination of them

$$\theta f(\Gamma) - \lambda g(\Gamma), \quad \Gamma \in \chi, \quad (17.21)$$

where $\theta, \lambda > 0$. If we minimize (17.21) (depending on the weights θ and λ , the minimum may not exist, and the infimum may be $-\infty$), we certainly pursue both objectives, even if through a mixture of them. However, by itself, such a minimization does not give us any guarantee, until we identify (17.21) with (17.19) and we exploit the previous analysis. This way, we discover that the minimum of (17.21), which is obviously a function of the “weights” of the combination, is \leq of that of (17.4), (17.5) and (17.6). In other words, *by minimizing such a combination, we find a lower bound of the minimum of (17.4), (17.5) and (17.6)*. This result is rather intuitive: in setting up a combination of the objectives and, in addition, choosing arbitrary “weights,” *we have been “arbitrarily optimistic” as concerns the design of the wing*. At this point, it comes natural to search, among all such “optimistic designs,” for one which is the least optimistic; in other words, we look for the maximum, with respect to the “weights,” of the several above minima (found with respect to $\Gamma \in \chi$) of (17.21). But this is the dual problem of (17.4), (17.5) and (17.6), so that we obtain the same result as from (17.4), (17.5) and (17.6) (see Theorem 17.4 of Appendix 2 for details).

The dual problem of (17.4), (17.5) and (17.6), namely (17.20) (see also the 1st side of (17.19) of Appendix 2) depends on the constant c , even if it does not appear explicitly in (17.20). Now, replace c with the parameter ξ and denote the dual problem by $P^*(\xi)$; in other words, $P^*(\xi)$ is the dual of (17.4), (17.5) and (17.6), where c

is replaced by ξ . Let us now continue the interpretation of $P^*(\xi)$ and its exploitation for the design. By solving $P^*(\xi)$, we find the functions

$$\Gamma(\xi), \quad \theta(\xi), \quad \lambda(\xi), \quad f^*(\xi), \quad (17.22)$$

the last of which gives the maximum in (17.20) (or in the first side of (17.61) of Appendix 2). As said before for the case $\xi = c$, the ratio $\lambda(\xi)/\theta(\xi)$ gives a fundamental information for the design (see Appendix 2 for details). Thus, it is reasonable to assume that the designer can define a function, say $\varphi : [c_1, c_2] \rightarrow \mathbb{R}_+$ with c_1, c_2 given positive constants within which ξ must lie, which expresses a measure of the merit for the project consequent to the value of the ratio $\lambda(\xi)/\theta(\xi)$. It is reasonable to suppose also that φ be unimodal (so that it possesses maximum and unique maximum point). Then, the designer can now consider the problem

$$\max_{\xi \in [c_1, c_2]} \varphi \left(\frac{\lambda(\xi)}{\theta(\xi)} \right). \quad (17.23)$$

By solving it, he finds the unique maximum point, say $\bar{\xi}$. Consequently, with regard to the combination (17.21) of the two objectives, $\theta(\bar{\xi})$ and $\lambda(\bar{\xi})$ are “the best weights” with respect to the minimax criterion (expressed by the dual problem) and the criterion expressed by the merit function φ . This way, the designer avoids to perform an empirical choice of the weights. For such an approach, the image space analysis (see Appendix 2) has been instrumental: to see this, it is enough to note that the pair $(\theta(\xi), \lambda(\xi))$ is the gradient of a supporting line of the image set (or its conic extension) of (17.4), (17.5) and (17.6) (see Appendix 2, Definition 17.2); the functional form of this line is the core of the dual problem.

The above approach can be generalized in several ways. First of all, we can be faced with $l \geq 2$ objectives. In this case, by exploiting the reciprocity principle mentioned in (17.59) of Appendix 2, we can limit ourselves to consider $l - 1$ ratios of type λ/θ . In order to formulate (17.4), (17.5) and (17.6), we have chosen the constant c ; in as much as such a constant is replaced by a parameter, the determination of c becomes no longer essential. When the determination of the functions (17.22) may be computationally complex, the global analysis can be replaced by a local one.

To sum up the previous development, we can observe that the problem takes the remarkable role to free us from the irrational task to choose arbitrary (or with an empirical criterion) the weights of a combination of objectives. Of course, this fact can be generalized in various directions, in particular, when there are more than two entities/objectives; in such a case, the choice, among several entities, of one to be considered as objective may be much more difficult than in the present case of only two entities.

As concerns the computational aspects, let us observe that, due to Proposition 17.7, the dual problem of (17.4), (17.5) and (17.6) can be equivalently reduced to just one operation

$$\max_{\Gamma \in \mathcal{X}; \theta, \lambda > 0} \mathcal{L}(\Gamma; \theta, \lambda), \quad \text{subject to} \quad \mathcal{L}_\Gamma^l(\Gamma; \theta, \lambda) = 0, \quad (17.24)$$

where \mathcal{L}'_Γ denotes the first variation of \mathcal{L} with respect to Γ .

The second side of (17.61) of Appendix 2 offers a further interpretation in terms of designing a wing. For each design $\Gamma \in \mathcal{X}$, we consider again the combination (17.21), but, this time, we keep fixed Γ and we maximize it with respect to $\theta, \lambda > 0$ (depending on Γ , the maximum may not exist, and the supremum may be $+\infty$). This way, each design Γ is associated with a maximum of (17.21); this means to be “arbitrarily pessimistic.” Then, by minimizing such a result with respect to Γ , we look for the least pessimistic situation and free ourselves from the arbitrariness of the choice of the weights.

Let us now give a concise interpretation of the above development. The optimality of the circulation distribution $\bar{\Gamma}$ has been reduced to show separation, by means of a line, between two sets, the image set $\mathcal{H}_{\bar{\Gamma}}$ and \mathcal{H} . The separation line, namely H^0 (which in Appendix 2 is denoted, with a better notation, by $H^0(\bar{\theta}, \bar{\lambda}, 0)$), turned out to be a support line of the image set. The gradient of such a line, namely $(\bar{\theta}, \bar{\lambda})$, has shown to provide us with an extremely important information about the given problem: indeed, $\bar{\lambda}/\bar{\theta}$ is the so-called Lagrange multiplier and is (up to a constant) the (instantaneous) velocity with which the minimum induced drag changes with respect to the total lift. Hence, the dual problem of (17.4), (17.5) and (17.6) can be viewed as the search for “the best one” among the support lines of the image set, or H^0 . Thus, a spontaneous remark may arise: such a support line is not contained in the data which define (17.4), (17.5) and (17.6); being an adjunctive entity, which comes from the exterior of (17.4), (17.5) and (17.6), the line should have not an importance and be a mere catalyst; indeed, the support line is a tangent (Bouligand tangent, if at the supporting point the image set is not smooth), and this explains in a straightforward way its importance.

Another aspect of the above development has consisted in providing us with a way of proving the existence of the minimum of (17.4), (17.5) and (17.6), which is much easier than the classic one: in fact, a remarkable fact is that such a way requires to us to prove the existence of the extremum of a problem in a finite-dimensional space, namely the IS which in the present case is the Euclidean plane, notwithstanding the fact that the given problem is infinite dimensional (Γ runs in a Banach space), while the classic ones require to prove the existence of the extremum in an infinite-dimensional space (just that Banach space).

To sum up some of the wonderful aspect of the duality theory, we can say that the dual problem allows us

- (i) to achieve important theoretical, analytical results;
- (ii) to improve solving methods for the given problem;
- (iii) to obtain, with the dual variables, a knowledge on the given problem which, often, is more important than the solution itself of the given problem; for instance, if the given problem represents an engineering design, often its solution is not striking for the designer and merely refines what he already knows; on the contrary, almost always, the solution of the dual problem brings a precious and *unexpected* information or even leads to a new approach to the design, as shown in this subsection.

17.5 Direct Methods

In this section, we determine the solution, Γ , of the isoperimetric problem and propose a computation direct method to obtain a set of approximations to Γ , converging to Γ . In this method, the Γ circulation is obtained by means of the two following procedures:

- a classic Fourier expansion of Γ ;
- the Ritz method, with two minimizing sequences of the type

$$\Gamma_n(x) = \sum_{i=0}^n b_i W_i(x), \quad n \in \mathbb{N},$$

where $W_i = (1 - x^2)x^i$ in the former type and $W_i = (1 - x^2)^i$ in the latter one.

17.5.1 Elliptic Distribution

We put $y = \cos \theta$ and, hence $dy = -\sin \theta$, and consider the Fourier expansion of Γ , with the conditions $\Gamma(-1) = \Gamma(1) = 0$, or:

$$\Gamma = \sum_{n=1}^{\infty} a_n \sin(n\theta). \quad (17.25)$$

The expression of lift L becomes

$$\begin{aligned} L &= \rho V_{\infty} \int_{-1}^1 \Gamma(y) dy \\ &= \rho V_{\infty} \int_0^{\pi} \Gamma(\theta) \sin \theta d\theta \\ &= \rho V_{\infty} \sum_{n=1}^{\infty} a_n \int_0^{\pi} \sin \theta \sin(n\theta) d\theta \\ &= \rho V_{\infty} \left(a_1 \int_0^{\pi} \sin^2 \theta d\theta + \sum_{n=2}^{\infty} a_n \int_0^{\pi} \sin \theta \sin(n\theta) d\theta \right). \end{aligned}$$

Because $\int_0^{\pi} \sin(m\theta) \sin(n\theta) d\theta = 0$ if $n \neq m$, we have

$$L = \rho V_{\infty} a_1 \int_0^{\pi} \sin^2 \theta d\theta = \frac{\pi}{2} a_1 \rho V_{\infty}. \quad (17.26)$$

In $y_0 \in [-1, 1]$ the induced velocity is

$$w(y_0) = \frac{1}{4\pi} \int_{-1}^1 \frac{d\Gamma(y)}{dy} \frac{1}{y - y_0} dy, \quad (17.27)$$

or, equivalently

$$\begin{aligned} w(\theta_0) &= -\frac{1}{4\pi} \int_0^\pi \frac{d\Gamma(\theta)}{d\theta} \frac{1}{\cos \theta - \cos \theta_0} d\theta \\ &= -\frac{1}{4\pi} \sum_{n=1}^{\infty} na_n \int_0^\pi \frac{\cos(n\theta)}{\cos \theta - \cos \theta_0} d\theta. \end{aligned} \quad (17.28)$$

Due to the Glauert formula we obtain

$$\begin{aligned} w(\theta_0) &= -\frac{1}{4\pi} \sum_{n=1}^{\infty} na_n \left(\pi \frac{\sin(n\theta_0)}{\sin \theta_0} \right) \\ &= -\frac{1}{4} \sum_{n=1}^{\infty} na_n \frac{\sin(n\theta_0)}{\sin \theta_0}, \end{aligned} \quad (17.29)$$

and, finally,

$$\begin{aligned} D_i &= -\frac{\rho}{4} \int_0^\pi \left(\sum_{n=1}^{\infty} na_n \frac{\sin(n\theta)}{\sin \theta} \right) \left(\sum_{n=1}^{\infty} a_n \sin(n\theta) \right) - \sin \theta d\theta \\ &= \frac{\rho}{4} na_n^2 \int_0^\pi \sin^2(n\theta) d\theta = \frac{\rho\pi}{8} (a_1^2 + 2a_2^2 + \dots + na_n^2 + \dots) \end{aligned} \quad (17.30)$$

Because all the terms are positive and, in order to have a non-negative lift, we need $a_1 \neq 0$, the induced drag is minimum when $a_n^2 = 0$, $\forall n > 1$. Putting $a_1 = \Gamma_0$ the solution of the isoperimetric problem (17.4), (17.5) and (17.6) is

$$\Gamma(\theta) = \Gamma_0 \sin \theta, \quad (17.31)$$

or, in terms of y

$$\Gamma(y) = \Gamma_0 \sqrt{1 - y^2}, \quad (17.32)$$

and the induced drag D_i becomes

$$D_i = \frac{\rho\pi}{8} \Gamma_0^2. \quad (17.33)$$

Remarks 17.1 When: $\Gamma(\theta) = \Gamma_0 \sin \theta$, then

- the induced velocity w is constant, because

$$w(\theta) = -\frac{1}{4\pi} \int_0^\pi \frac{d\Gamma(\alpha)}{d\alpha} \frac{1}{\cos \alpha - \cos \theta} d\alpha = -\frac{\Gamma_0}{4} = \text{constant}; \quad (17.34)$$

- taking eqs. (17.26) and (17.8) into account, when Γ is elliptical, it is trivial to obtain the well-known result in aerodynamics

$$D_i = \frac{L^2}{2\pi\rho V_\infty}. \quad (17.35)$$

17.5.2 Ritz Method

In this subsection, we obtain an approximate solution of the isoperimetric problem (17.4), (17.5) and (17.6) by means of the Ritz method. The unknown circulation shape functions are of the following type:

$$\Gamma_n := \sum_{i=0}^n b_i W_i(x), \quad n \in \mathbb{N},$$

where, for example,

$$W_i(x) = b_i(1-x^2)^{i+1} \text{ and } W_i(x) = b_i(1-x^2)x^i,$$

in order to satisfy the kinematic boundary conditions $\Gamma_n(-1) = \Gamma_n(1) = 0$.

The two classes of polynomials are indicated as TIPO1 and TIPO2, respectively; both of them respect the boundary conditions $W_i(1) = W_i(-1) = 0$.

We remark that, even though polynomials TIPO1 are symmetric and TIPO2 are not, we do not need to assume any condition of symmetry from physics, because symmetry is intrinsic in the mathematical solution of the isoperimetric problem; in fact, for any Γ_n of TIPO2, the optimum solutions give $b_i = 0$, for all i odd. Now we describe a generic iteration for Γ_n .

17.5.2.1 Algorithm

- We write the induced drag D_i as a function of Γ_n , by solving the double integral according to the principal value of Cauchy; moreover, because $\Gamma_n = \sum_{i=0}^n b_i W_i(x)$, $n \in \mathbb{N}$, at any step we know $D_i(f_{n-2})$, relevant to the previous one. We obtain

$$\begin{aligned} D_i(\Gamma_n) = D_i(b_0, \dots, b_n) = D_i(f_{n-2}) &+ \frac{\rho}{4\pi} \sum_{i=n-1}^n \sum_{s=0}^n b_i b_s \int_{-1}^1 \int_{-1}^1 \frac{W_i(y) W'_s(x)}{y-x} dx dy + \\ &+ \frac{\rho}{4\pi} \sum_{i=0}^{n-2} \sum_{s=n-1}^n b_i b_s \int_{-1}^1 \int_{-1}^1 \frac{W_i(y) W'_s(x)}{y-x} dx dy. \end{aligned}$$

Because the functional D_i is quadratic with respect to Γ_n and Γ'_n , the result of the integration is a second-order homogeneous polynomial in b_i , $i = 0, \dots, n$.

- The lift is written as a function of Γ_n as well, and we have

$$L(\Gamma_n) = L(b_0, \dots, b_n) = \rho V_\infty \sum_{i=0}^n b_i \int_{-1}^1 W_i(y) dy.$$

The function L is linear in the unknowns b_i . The isoperimetric problem (17.4), (17.5) and (17.6) becomes:

$$(P) \quad \min D_i(b_0, \dots, b_n), \quad \text{s.t.} \quad L(b_0, \dots, b_n) = c, \quad (b_0, \dots, b_n) \in \mathbb{R}^{n+1}. \quad (17.36)$$

- The functional induced drag is a second-order homogeneous polynomial in b_i , and the condition that the derivatives of the Lagrangian with respect to b_i must be zero is equivalent to solve the following linear system:

$$Ax = b$$

where

- the matrix A of coefficients is the Hessian of the Lagrangian

$$J(b_0, \dots, b_n, \lambda) = D_i(b_0, \dots, b_n) - \lambda L(b_0, \dots, b_n);$$

- $b = [0, \dots, 0, -c]$;
- because of the convexity of the functional induced drag it results that $x = [\bar{b}_0, \dots, \bar{b}_n, \bar{\lambda}]$ is a point of minimum of the Lagrangian J .
- Once the $n+1$ -th $(\bar{b}_0, \dots, \bar{b}_n)$ we calculate $D_i(\bar{b}_0, \dots, \bar{b}_n)$ and, hence, D_i as a function of c

$$D_i(\bar{b}_0, \dots, \bar{b}_n) = \alpha_n c^2. \quad (17.37)$$

- Calculation of the Oswald coefficient “ e ,” that is

$$e := \frac{\bar{D}_i}{D_i(\bar{b}_0, \dots, \bar{b}_n)}, \quad (17.38)$$

where \bar{D}_i is the induced drag relevant to the elliptic circulation defined in (17.35).

The algorithm described before has been implemented by using the commercial code MapleV, with a symbolic computation.

As an example, we apply the iterative procedure with the following conditions:

- $\rho = 1$;
- $V_\infty = 1$;
- $c = 100$;
- elliptical circulation $\bar{\Gamma} = 63.6942675\sqrt{1-x^2}$;
- \bar{D}_i corresponding to $c = 100$ is worth: 1592.356688;
- let us indicate $\bar{\Gamma}_{max}$ the maximum value of $\bar{\Gamma}$ inside $[-1, 1]$.

17.5.2.2 Numerical Results

In this section, some numerical results are reported in order to show that the method is convergent when “ n ” becomes larger and larger.

Example 17.1

$$n = 8$$

Γ_8 is shown, for TIPO1 polynomials, in Fig. 17.4:

$$\begin{aligned} D_i(\Gamma_8) &= 1612.261147 \\ e &= 0.9876543207 \end{aligned}$$

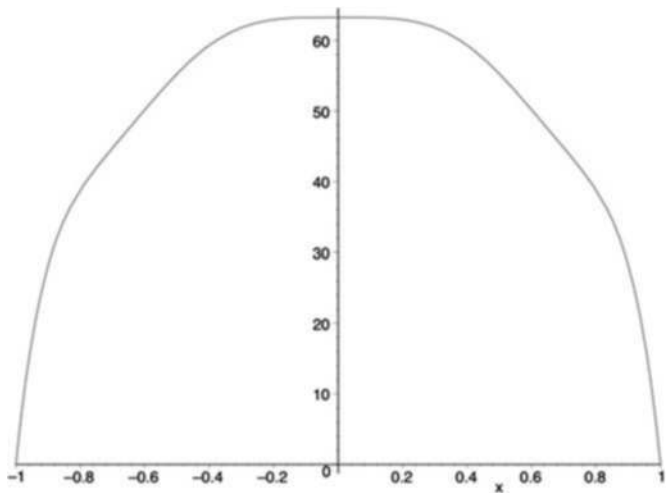


Fig. 17.4 Circulation distribution Γ_8

The error when $\overline{\Gamma}$ is approximated with $\Gamma_8(x)$ is shown in Fig. 17.5.

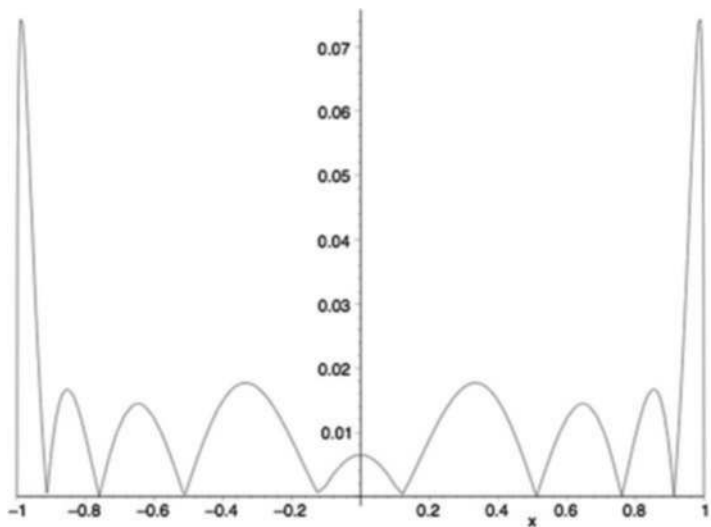


Fig. 17.5 Function $y(x) = \frac{|\Gamma_8(x) - \overline{\Gamma}(x)|}{\overline{\Gamma}_{max}}$

Example 17.2

$$n = 44$$

Γ_{44} is shown in Fig. 17.6:

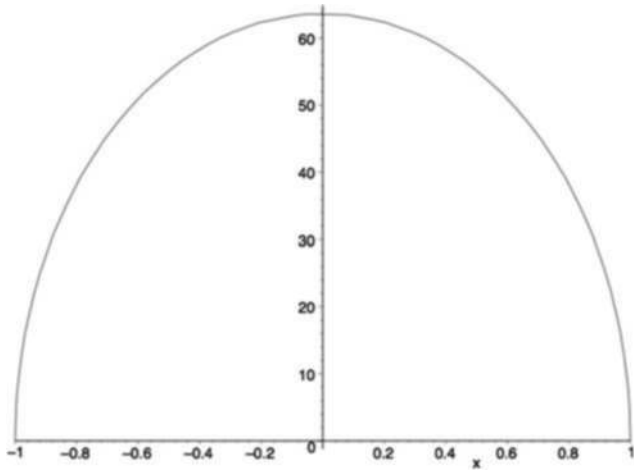


Fig. 17.6 Circulation distribution Γ_{44}

$$\begin{aligned} D_i(\Gamma_{44}) &= 1593.143425 \\ e &= 0.9995061731 \end{aligned}$$

Figure 17.7 shows, for any $x \in [-1, 1]$, the error when $\bar{\Gamma}$ is approximated with $\Gamma_{44}(x)$.

$$D_i(\Gamma_{44}) = 1594.252351, \quad e = 0.9988109392$$

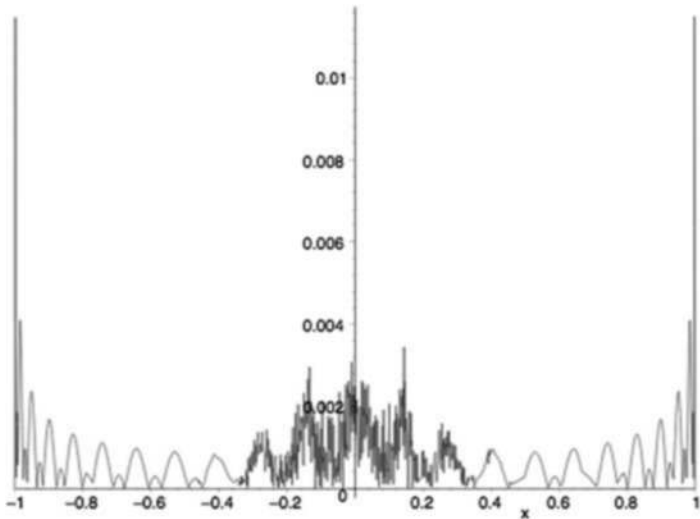


Fig. 17.7 Function $y(x) = \frac{|\Gamma_{44}(x) - \bar{\Gamma}(x)|}{\bar{\Gamma}_{max}}$

In the case of TIPO2 polynomials the results are very similar and they are not reported here for brevity sake. The results show that the iterative procedure converges to the exact solution and, also, that the solutions of TIPO1 and TIPO2 polynomials (with the same degree) give the same induced drag. As an example, Table 17.1 shows the TIPO1 main data relevant to some iterations.

Table 17.1 Numerical iterations relevant to TIPO1 polynomials

Degree	D_i	$\frac{D_i - \bar{D}_i}{\bar{D}_i}$	e
4	1658.704883	0.041666666646	0.9600000002
12	1601.835002	0.005952381192	0.9940828400
20	1595.975680	0.002272726976	0.9977324266
28	1594.252351	0.001190476364	0.9988109392
36	1593.520691	0.0007309938840	0.9992695401
44	1593.020170	0.0004166666960	0.9995835068
52	1592.923767	0.0003561256120	0.9996440012
64	1592.733666	0.0002367421840	0.9997633138

Appendix 1: Existence and Convexity of the Induced Drag Functional

Before proving the existence of the functional induced drag, we recall some useful definitions.

Definition 17.1 A function $f : [a, b] \rightarrow \mathbb{R}$ is absolutely continuous in $[a, b]$, and we write $f \in AC[a, b]$ iff, for any $\varepsilon > 0$ it exists $\delta > 0$ such that for any finite collections of disjoint intervals $]\alpha_i, \beta_i[$, $i = 1, \dots, k$, included in $[a, b]$ e with $\sum_{i=1}^k (\beta_i - \alpha_i) < \delta$, it results $\sum_{i=1}^k |f(\beta_i) - f(\alpha_i)| < \varepsilon$.

Definition 17.2 Let (Y, \mathcal{F}, μ) be a measure space and $1 \leq p < \infty$. We put $\mathcal{L}^p(Y) = \{f : Y \rightarrow \mathbb{R} : f \text{ is measurable and } \int_Y |f|^p d\mu < \infty\}$.

If q is conjugate exponent of p (i.e., $\frac{1}{p} + \frac{1}{q} = 1$, and, by stipulation, the conjugate exponent of 1 is ∞ and vice versa), we have $\|f\|_{\mathcal{L}^p(Y)} = \left[\int_Y |f|^p d\mu \right]^{\frac{1}{p}}$.

Hölder Inequality. If $f \in \mathcal{L}^p(Y)$ e $g \in \mathcal{L}^q(Y)$, then $fg \in \mathcal{L}^1(Y)$ and $\|fg\|_1 \leq \|f\|_p \|g\|_q$.

Proposition 17.1 Let $f \in AC] -1, 1[$ be such that

$$f(1) = f(-1) = 0, \quad f' \in \mathcal{L}^{1+\varepsilon}] -1, 1[, \text{ with } \varepsilon > 0.$$

Then

$$\int_{-1}^1 \int_{-1}^1 \frac{f'(x)f(y)}{y-x} dx dy$$

is convergent as a Cauchy improper integral.

Proof. Let us set

$$S_1(h) := \{(x, y) \in \mathbb{R}^2 : x+h < y < 1, -1 < x < 1-h\},$$

$$S_2(h) := \{(x, y) \in \mathbb{R}^2 : -1 < y < x-h, -1+h < x < 1\},$$

$$G_{S_i(h)}(f) := \int \int_{S_i(h)} \frac{f'(x)f(y)}{y-x} dx dy, \quad i = 1, 2.$$

Let us integrate by parts both $G_{S_1(h)}$ and $G_{S_2(h)}$

$$\begin{aligned} G_{S_1(h)}(f) &= \int_{-1}^{1-h} \int_{x+h}^1 \frac{f'(x)f(y)}{y-x} dx dy = \\ &= \int_{-1}^{1-h} \left[[\ln(y-x)f'(x)f(y)]_{x+h}^1 - \int_{x+h}^1 \ln(y-x)f'(x)f'(y) dy \right] dx = \\ &= - \int_{-1}^{1-h} \left[\ln(h)f'(x)f(x+h) - \int_{x+h}^1 \ln(y-x)f'(x)f'(y) dy \right] dx, \\ G_{S_2(h)}(f) &= \int_{-1+h}^1 \int_{-1}^{x-h} \frac{f'(x)f(y)}{y-x} dx dy = \\ &= \int_{-1+h}^1 \left[[\ln(x-y)f'(x)f(y)]_{-1}^{x-h} - \int_{-1}^{x-h} \ln(x-y)f'(x)f'(y) dy \right] dx = \\ &= \int_{-1+h}^1 \left[\ln(h)f'(x)f(x-h) - \int_{-1}^{x-h} \ln(x-y)f'(x)f'(y) dy \right] dx. \end{aligned}$$

It results that

$$\begin{aligned} &\int_{-1}^1 \int_{-1}^1 \frac{f'(x)f(y)}{y-x} dx dy \doteq \lim_{h \rightarrow 0} G_{S_1(h)}(f) + G_{S_2(h)}(f) = \\ &= \lim_{h \rightarrow 0} - \int_{-1+h}^1 \int_{-1}^{x-h} \ln(x-y)f'(x)f'(y) dy dx + \\ &\quad + \ln(h) \left(\int_{-1+h}^1 f'(x)f(x-h) dx - \int_{-1}^{1-h} f'(x)f(x+h) dx \right) + \\ &\quad - \int_{-1}^{1-h} \int_{x+h}^1 \ln(y-x)f'(x)f'(y) dy dx = - \int_{-1}^1 \int_{-1}^1 \ln|y-x|f'(x)f'(y) dx dy. \end{aligned}$$

The thesis is obtained by observing that

$$\left| \int_{-1}^1 \int_{-1}^1 \ln|y-x| f'(x) f'(y) dx dy \right| \leq \int_{-1}^1 |f'(x)| \int_{-1}^1 |\ln|y-x|| |f'(y)| dy dx, \quad (17.39)$$

and that from the Hölder disequality we have

$$\int_{-1}^1 |\ln|y-x|| |f'(y)| dy \leq \|f'\|_{\mathcal{L}^{1+\varepsilon}(-1,1)} \|\ln|y-x|\|_{\mathcal{L}^{\frac{1+\varepsilon}{\varepsilon}}(-1,1)}.$$

After having observed that, for suitable constants $\delta > 0$ and $C \in \mathbb{R}$, it results

$$\ln|y-x| \leq \frac{C}{|y-x|^\delta} \quad \forall |y-x| \in]0, 2],$$

we obtain

$$\|\ln|y-x|\|_{\mathcal{L}^{\frac{\varepsilon+1}{\varepsilon}}(-1,1)} < \eta, \eta \in \mathbb{R}.$$

From (17.39) we have, finally

$$\left| \int_{-1}^1 \int_{-1}^1 \ln|y-x| f'(x) f'(y) dx dy \right| \leq \eta \|f'\|_{\mathcal{L}^{1+\varepsilon}(-1,1)} \|f'\|_{\mathcal{L}^1(-1,1)} < \infty \quad (17.40)$$

as required. \square

Before proving the convexity of the functional, we recall the following:

Definition 17.3 Let K be a vector space. A function $f : K \rightarrow \mathbb{R}$ is called *convex*, if and only if

$$(1-\alpha)f(x) + \alpha f(y) \geq f((1-\alpha)x + \alpha y), \quad \forall x, y \in K, \quad \forall \alpha \in [0, 1]. \quad (17.41)$$

We say that function f is *strictly convex*, if and only if the inequality (17.41) holds strictly.

Equivalently,

Theorem 17.2. Let K be a vector space and $f : K \rightarrow \mathbb{R}$ be a function whatever. f is strictly convex on K , if and only if $\forall x, y \in K$ the quotient ratio

$$t \rightarrow R_y(t) = \frac{f(x+ty) - f(x)}{t}, \quad t \in \mathbb{R}_+ \setminus \{0\}$$

is an increasing function.

Proposition 17.2 Let be

$$\mathcal{X} = \{f \in AC([-1, 1]), f' \in \mathcal{L}^{1+\varepsilon}, f(-1) = f(1) = 0\}$$

and let us define the functional

$$J: \mathcal{X} \rightarrow \mathbb{R},$$

putting

$$J(f) = \int_{-1}^1 \int_{-1}^1 \frac{f'(x)f(y)}{y-x} dx dy,$$

where the double integral on the right-hand side exists in the Cauchy principal. So, we have

- (a) the functional J is not strictly convex on \mathcal{X} ;
- (b) the functional J is strictly convex on $\mathcal{X}^+ := \{f \in \mathcal{X} : J(f) > 0\}$.

Proof. (a) We calculate the difference quotient of J for $f, g \in \mathcal{X}$ whatever

$$\begin{aligned} R_g(t) &= \frac{J(f+tg) - J(f)}{t} \\ &= \frac{1}{t} \left(\int_{-1}^1 \int_{-1}^1 \frac{(f'(x) + tg'(x))(f(y) + tg(y))}{y-x} dx dy - \int_{-1}^1 \int_{-1}^1 \frac{f'(x)f(y)}{y-x} dx dy \right) \\ &= \int_{-1}^1 \int_{-1}^1 \frac{f'(x)g(y) + g'(x)f(y)}{y-x} dx dy + t \left(\int_{-1}^1 \int_{-1}^1 \frac{g'(x)g(y)}{y-x} dx dy \right). \end{aligned}$$

Now we calculate the derivative of the difference quotient:

$$R'_g(t) = \int_{-1}^1 \int_{-1}^1 \frac{g'(x)g(y)}{y-x} dx dy. \quad (17.42)$$

After Theorem 17.2 the functional J is not, in general, strictly convex; in fact there exist functions g for which the difference quotient $R_g(t)$ is decreasing, as for example, $g = -2 + \frac{3}{4}(1-x^2)^2$ because, $\forall f \in \mathcal{X}$, we get $R'_g(t) < 0$.

- (b) The strict convexity comes from Theorem 17.2, if we adjoin, as an hypothesis for the set \mathcal{X} , that the condition $J(f) > 0$ holds. \square

The consequence of Proposition 17.2 is that, if the minimum for f exists, then it is unique.

Appendix 2: Image Space Analysis

The study of the properties of the image of a real-valued function is an old one. However, in most cases the properties of the image have not been the purpose of the study and their investigation has occurred as an auxiliary step toward other achievements [2].

Traces of the idea of studying the images of functions involved in a constrained extremum problem go back to the work of C. Carathéodory. In the 1950s, R. Bellman, with his celebrated maximum principle, proposed – for the first time in the field of optimization – to replace the given unknown by a new one which runs in the image; however, also here the image is not the main purpose. Only in the late 1960s and 1970s some authors, independently from each other, brought explicitly such a study into the field of optimization (see Sect. 17.3.2 of [2]).

The approach consists in introducing the space, call it *image space* (for short, IS), where the images of functions of the given extremum problem run. Then a new problem is defined in the IS, which is equivalent to the given one. In a certain sense, such an approach has some analogies with what happens in the Theory of Measure when one goes from Mengoli–Cauchy–Riemann measure to the Lebesgue one.

The analysis in the IS must be viewed as a preliminary and auxiliary step – and not as a concurrent of the analysis in the given space – for studying a constrained extremum problem. When a statement has been achieved in the IS, then, of course, we have to write the corresponding (equivalent) statement in terms of the given space; the latter is, in general, difficult to be conceived without having at disposal the former. If this aspect is understood, then the IS analysis may be highly fruitful. In fact, in the IS we may have a sort of “regularization”: the conic extension (see Definition 17.5) of the image set (see Definition 17.4) of the given extremum problem may be convex or continuous or smooth when the given extremum problem does not enjoy the same property, so that convex or continuous or smooth analysis can be developed in the IS, but not in the given space. If the image set of an extremum problem is finite dimensional (as happens to (17.4), (17.5) and (17.6)), then it can be analysed, in the IS, by means of the some mathematical concepts which are used for the finite dimensional case, even if the domain of the given problem (χ in (17.4), (17.5) and (17.6)) is infinite dimensional. If the image set is infinite dimensional, by means of a suitable use of the selection theory of point-to-set maps, it is possible to postpone such an infinite dimensionality to the introduction of the IS, which, therefore, can be held finite dimensional. In this section, we understand that suitable assumptions have been made in order to let the extrema be achieved.

The IS approach arises naturally in as much as an optimality condition for an extremum problem is achieved through the impossibility of a system. By paraphrasing the very definition of global minimum point for (17.4), (17.5) and (17.6), we can say that $\bar{\Gamma} \in R := \{\Gamma \in \chi : \rho V_\infty \int_T \Gamma(x) dx - c = 0\}$ is a *global minimum point*, iff the system (in the unknown Γ)

$$f_{\bar{\Gamma}}(\Gamma) := f(\bar{\Gamma}) - f(\Gamma) > 0, \quad g(\Gamma) = 0, \quad \Gamma \in \chi \quad (17.43)$$

is impossible. This system leads immediately to introduce the image set of (17.4), (17.5) and (17.6).

Definition 17.4 *The set*

$$\mathcal{K}_{\bar{\Gamma}} := \{(u, v) \in \mathbb{R}^2 : u = f_{\bar{\Gamma}}(\Gamma), \quad v = g(\Gamma), \quad \Gamma \in \chi\}$$

is called the image of (17.4), (17.5) and (17.6).

By introducing the set

$$\mathcal{H} := \{(u, v) \in \mathbb{R}^2 : u > 0, \quad v = 0\},$$

which reflects the conditions of (17.43), it is trivial to state the following:

Proposition 17.3 $\bar{\Gamma} \in R$ is a global minimum point of (17.4), (17.5) and (17.6), if and only if

$$\mathcal{H} \cap \mathcal{K}_{\bar{\Gamma}} = \emptyset. \quad (17.44)$$

In passing, it is worth noting that minimization is the way of reading (in the sense of Galilei) the laws of nature (or human behavior), while the mathematical core of an extremum problem consists in proving the impossibility of a system or the disjunction of two sets, as (17.43) and (17.44) show.

As announced, we can now introduce the *image problem*:

$$\max(u), \quad \text{s.t. } (u, v) \in \mathcal{K}_{\bar{\Gamma}}, \quad v = 0, \quad (17.45)$$

and prove the following:

Proposition 17.4 Problems (17.4), (17.5) and (17.6) and (17.45) are equivalent, in the sense that (\hat{u}, \hat{v}) is a global maximum point of (17.45), if and only if it is the image, through the map $(f_{\bar{\Gamma}}(\Gamma), g(\Gamma))$, of a global minimum point, say $\hat{\Gamma}$, of (17.4), (17.5) and (17.6), and we have

$$f(\bar{\Gamma}) - \hat{u} = f(\hat{\Gamma}). \quad (17.46)$$

Proof. Only if. $(\hat{u}, \hat{v}) \in \mathcal{K}_{\bar{\Gamma}} \cap (\mathbb{R} \times \mathbb{O}) \Rightarrow \exists \hat{\Gamma} \in \mathcal{X}$, such that

$$u = f(\bar{\Gamma}) - f(\Gamma), v = g(\Gamma) = 0.$$

Taking into account these relations (the first of which proves the last claim), the assumption

$$\hat{u} \geq u, \quad \forall (u, v) \in \mathcal{K}_{\bar{\Gamma}} \cap (\mathbb{R} \times \mathbb{O}),$$

implies $f(\bar{\Gamma}) - f(\hat{\Gamma}) \geq f(\bar{\Gamma}) - f(\Gamma)$ or $f(\hat{\Gamma}) \leq f(\Gamma), \forall \Gamma \in \mathcal{X}$.

If. Set $\hat{u} := f(\bar{\Gamma}) - f(\hat{\Gamma}), \hat{v} := g(\hat{\Gamma})$, so that

$$(\hat{u}, \hat{v}) \in \mathcal{K}_{\bar{\Gamma}} \cap (\mathbb{R} \times \mathbb{O}).$$

From the assumption we draw $f(\hat{\Gamma}) \leq f(\Gamma), \forall \Gamma \in R$; by setting

$u := f(\bar{\Gamma}) - f(\Gamma)$ and $v := g(\Gamma)$, we have $f(\bar{\Gamma}) - f(\hat{\Gamma}) \geq f(\bar{\Gamma}) - f(\Gamma), \forall \Gamma \in R$, and hence $\hat{u} \geq u, \forall (u, v) \in \mathcal{K}_{\bar{\Gamma}} \cap (\mathbb{R} \times \mathbb{O})$. \square

Note that, while (17.4), (17.5) and (17.6) is infinite dimensional (its unknown runs in a Banach space), (17.45) is finite dimensional (its unknown runs in the Euclidean plane).

The theory of constrained extrema is full of proposals for changing the data of the given problem, without losing the extremum and extremum points, and with

the purpose of adding a desired property to the problem. Such proposals have been made essentially with reference to the given space. The IS approach suggests a new proposal, based on the following definition. cl denotes closure, and the difference is in the vector sense.

Definition 17.5 Let $\mathcal{Z} \subset \mathbb{R}^2$ denote a generic set of the IS associated with (17.4), (17.5) and (17.6). \mathcal{E} will denote the map which sends \mathcal{Z} into $\mathcal{Z} - cl\mathcal{H} \subset \mathbb{R}^2$; it is called conic extension of \mathcal{Z} .

Of course $\mathcal{Z} \subseteq \mathcal{E}(\mathcal{Z})$. In the sequel, we will consider only the conic extension of the image set, or $\mathcal{E}(\mathcal{K}_{\overline{\Gamma}})$. The above definition has been given for the particular problem (17.4), (17.5) and (17.6); obviously, it can be given for a general extremum problem (see [2], Def. 3.2).

Proposition 17.5 (17.44) holds, if and only if

$$\mathcal{H} \cap \mathcal{E}(\mathcal{K}_{\overline{\Gamma}}) = \emptyset \quad (17.47)$$

Proof. If. It is an obvious consequence of the inclusion $\mathcal{K}_{\overline{\Gamma}} \subseteq \mathcal{E}(\mathcal{K}_{\overline{\Gamma}})$.

Only if. Ab absurdo, suppose that $\exists z^1 \in \mathcal{K}_{\overline{\Gamma}}, \exists z^2 \in cl\mathcal{H}$ (so that $z^1 - z^2 \in \mathcal{E}(\mathcal{K}_{\overline{\Gamma}})$) and that $z^1 - z^2 \in \mathcal{H}$. Then, being \mathcal{H} the positive u -semi-axis (of the IS), we have

$$z^1 = (z^1 - z^2) + z^2 \in \mathcal{H} + cl\mathcal{H} = \mathcal{H},$$

and hence (17.44) is contradicted. \square

The above proposition shows that the optimality condition (17.44) still holds, if the image set (and therefore the data of (17.4), (17.5) and (17.6)) are modified according to Definition 17.5. This has an obvious consequence on the image problem (17.45), as shown by the following:

Proposition 17.6 Let Condition 17.1 hold.

(i) Problems (17.45) and

$$\max(u), \quad s.t. \quad (u, v) \in \mathcal{E}(\mathcal{K}_{\overline{\Gamma}}), \quad v = 0, \quad (17.48)$$

are equivalent in the sense of having the same maximum and maximum points.

(ii) Problem (17.48) has maximum.

Proof.

(i) Straightforward consequence of Propositions 17.3–17.5.

(ii) Because of Condition 17.1, $-f_{\overline{\Gamma}}(\Gamma)$ (see (17.43)) is coercive. Because of Proposition 17.2(b), $f_{\overline{\Gamma}}(\Gamma)$ is strictly concave. $g(\Gamma)$ is linear. Therefore, in the IS, the projection of $\mathcal{E}(\mathcal{K}_{\overline{\Gamma}})$ – as well as of $\mathcal{K}_{\overline{\Gamma}}$ – on the u -semi-axis is a closed (and bounded) segment. Hence, the assumptions of Theorem 3.2.3 of [2] are fulfilled. Such a theorem can thus be applied to achieve the thesis. \square

As announced in Sect. 17.4 (just before Condition 17.1), it is possible to prove the existence of the minimum in (17.4), (17.5) and (17.6) through IS: this is done by the above Proposition 17.6.

We have shown that a feasible $\bar{\Gamma} \in R$ is a (global) minimum point of (17.4), (17.5) and (17.6), iff (17.44) holds. In the general case (but also in the present one), to prove (17.44) is a difficult task. Therefore, a way of overcoming such a drawback consists in trying to show that \mathcal{H} and $\mathcal{K}_{\bar{\Gamma}}$ lie in two disjoint sets. The separation theory, whose “root” is the Hahn–Banach Linear Extension Theorem (but it was already present, even if in an implicit form, in Euclid!) is of great help.

Let us consider the function $w : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$, defined by

$$w(u, v; \theta, \lambda) := \theta u + \lambda v, \quad \theta, \lambda \in \mathbb{R}. \quad (17.49)$$

For each pair $(\theta, \lambda) \neq 0$, $w(u, v; \theta, \lambda) = 0$ identifies obviously a line, say H^0 , through the origin, of the IS (i.e., \mathbb{R}^2), where (u, v) runs; the IS is then split into two disjoint halfplanes:

$$H^-(\theta, \lambda, k) := \{(u, v) \in \mathbb{R}^2 : \theta u + \lambda v \leq k\}, \quad \theta, \lambda, k \in \mathbb{R},$$

$$H^+(\theta, \lambda, k) := \{(u, v) \in \mathbb{R}^2 : \theta u + \lambda v > k\}, \quad \theta, \lambda, k \in \mathbb{R}.$$

Of course, we have $\mathcal{H} \subset H^+(\theta, \lambda, 0)$, iff $\theta \neq 0$; thus, under this assumption, in order to prove (17.44) (and, hence, the optimality of $\bar{\Gamma}$) it is sufficient to show that $\exists \theta, \lambda \in \mathbb{R}$, with $\theta \neq 0$, such that

$$\mathcal{K}_{\bar{\Gamma}} \subseteq H^-(\theta, \lambda, 0), \quad (17.50)$$

or, equivalently (Proposition 17.5),

$$\mathcal{E}(\mathcal{K}_{\bar{\Gamma}}) \subseteq H^-(\theta, \lambda, 0). \quad (17.51)$$

In the general case, (17.50) – or (17.51) – is not necessary, as trivial examples show. However, it will be shown that, in the present case (17.4), (17.5) and (17.6), the inclusion (17.50) – or (17.51) – is also necessary.

Let ∂S and $\text{card} S$ denote the boundary and the cardinality of the set S , respectively; the difference between sets is denoted by “ \setminus ”; $H^0(\theta, \lambda, k) := \{(u, v) \in \mathbb{R}^2 : \theta u + \lambda v = k\}$ denotes a line (iff $(\theta, \lambda) \neq 0$) of the IS.

Proposition 17.7 *Let Condition 17.1 hold and $\bar{\Gamma} \in \chi$. $\mathcal{E}(\mathcal{K}_{\bar{\Gamma}})$ enjoys the following properties:*

- (i) $\mathcal{E}(\mathcal{K}_{\bar{\Gamma}})$ is strictly convex;
- (ii) $\partial \mathcal{E}(\mathcal{K}_{\bar{\Gamma}}) \subset \mathcal{K}_{\bar{\Gamma}}$;
- (iii) $\forall (u, v) \in \partial \mathcal{E}(\mathcal{K}_{\bar{\Gamma}})$, $\mathcal{E}(\mathcal{K}_{\bar{\Gamma}})$ admits a support, or $\exists (\theta, \lambda) \in \mathbb{R}^2$, with $\theta \neq 0$, and $\exists k \in \mathbb{R}$, such that

$$\mathcal{E}(\mathcal{K}_{\bar{\Gamma}}) \subset H^-(\theta, \lambda, k), \quad S := \mathcal{E}(\mathcal{K}_{\bar{\Gamma}}) \cap H^0(\theta, \lambda, k) \neq \emptyset, \quad \text{card} S = 1; \quad (17.52)$$

the same happens to $\mathcal{K}_{\bar{\Gamma}}$, or

$$\mathcal{K}_{\bar{\Gamma}} \subset H^-(\theta, \lambda, k), \quad \mathcal{K}_{\bar{\Gamma}} \cap H^0(\theta, \lambda, k) = S; \quad (17.53)$$

- (iv) $\mathcal{E}(\mathcal{K}_{\bar{\Gamma}})$ is regular, in the sense that, $\forall (u, v) \in \partial \mathcal{E}(\mathcal{K}_{\bar{\Gamma}})$, no supporting line (boundary of the Bouligand tangent cone to $\mathcal{E}(\mathcal{K}_{\bar{\Gamma}})$) at (u, v) to $\mathcal{E}(\mathcal{K}_{\bar{\Gamma}})$, is parallel to the u -axis of the IS.
 (v) at $v \geq 0$, a supporting line $H^0(\theta, \lambda, k)$ sub (iii) has $\theta > 0$ and $\lambda > 0$.

Proof. (i) Because of Proposition 17.2, $f(\Gamma)$ is strictly convex, so that $f_{\bar{\Gamma}}(\Gamma)$ (see (17.43)) is strictly concave. Hence, taking into account that $g(\Gamma)$ is linear in Γ and that $\mathcal{E}(\mathcal{K}_{\bar{\Gamma}})$ is the hypograph of $\mathcal{K}_{\bar{\Gamma}}$ (when $\mathcal{K}_{\bar{\Gamma}}$ is viewed as the point-to-set maps $v \rightrightarrows u$), $\mathcal{E}(\mathcal{K}_{\bar{\Gamma}})$ turns out to be strictly convex. (ii) It is a consequence of (i) and of the fact that $\mathcal{E}(\mathcal{K}_{\bar{\Gamma}})$ is the hypograph of $\mathcal{K}_{\bar{\Gamma}}$. (iii) The first two conditions of (17.52) come from the convexity of $\mathcal{E}(\mathcal{K}_{\bar{\Gamma}})$ and, as it concerns the existence of $\theta \neq 0$, from (iv); the last part of (17.52) is a consequence of the strict convexity of $\mathcal{E}(\mathcal{K}_{\bar{\Gamma}})$. Passing to (17.53), it is enough to observe that $\mathcal{E}(\mathcal{K}_{\bar{\Gamma}})$ is the hypograph (in the sense specified sub (i)) of $\mathcal{K}_{\bar{\Gamma}}$. (iv) Because of the Condition 17.1, $\forall v > 0, \exists (u, v) \in \mathcal{K}_{\bar{\Gamma}}$ (more general condition than the so-called Slater constraint qualification). Therefore, the existence of a supporting line $H^0(\theta, \lambda, k)$, parallel to the u -axis, account taken of (i), would require the boundedness of $\mathcal{K}_{\bar{\Gamma}}$ with respect to v and contradict the assumption. (v) It is a consequence of (iv) and of the Condition 17.1. \square

We are now ready to show that (17.50) and (17.51) are also necessary.

Proposition 17.8 *Let Condition 17.1 hold. $\bar{\Gamma} \in R$ is a (global) minimum point of (17.4), (17.5) and (17.6), if and only if $\exists(\bar{\theta}, \bar{\lambda}) > 0$, such that*

$$\mathcal{K}_{\bar{\Gamma}} \subset H^-(\bar{\theta}, \bar{\lambda}, 0) \quad \text{or equivalently} \quad \mathcal{E}(\mathcal{K}_{\bar{\Gamma}}) \subset H^-(\bar{\theta}, \bar{\lambda}, 0). \quad (17.54)$$

Proof. The sufficiency is an obvious consequence of what has been noted about (17.50). With regard to the necessity, the assumption that $\bar{\Gamma}$ be a (global) minimum point of (17.4), (17.5) and (17.6) implies (17.44) or (17.47). Because of Proposition 17.7(i), $\mathcal{E}(\mathcal{K}_{\bar{\Gamma}})$ and \mathcal{H} are separable; because of Proposition 17.7(iv), the separation line is of type $H^0(\bar{\theta}, \bar{\lambda}, 0)$ and does not contain the u -axis; because of Proposition 17.7(v), $\bar{\theta}$ and $\bar{\lambda}$ are positive.

Thus, the latter of (17.54) follows; the former is a consequence of the inclusion $\mathcal{K}_{\bar{\Gamma}} \subset \mathcal{E}(\mathcal{K}_{\bar{\Gamma}})$. \square

Since it is not easy to verify (17.54), it comes spontaneous to try to express the inclusion (17.54) through some extremum operators. In general, we cannot have equivalence; here it happens, due to Proposition 17.7.

Proposition 17.9 *Let Condition 17.1 hold and $\bar{\Gamma} \in R$.*

(i) *The equalities*

$$\min_{\theta, \lambda > 0} \max_{(u, v) \in \mathcal{K}_{\bar{\Gamma}}} (\theta u + \lambda v) = 0 \quad \text{and} \quad \min_{\theta, \lambda > 0} \max_{(u, v) \in \mathcal{E}(\mathcal{K}_{\bar{\Gamma}})} (\theta u + \lambda v) = 0 \quad (17.55)$$

are equivalent, respectively, to (17.54).

(ii) $\bar{\Gamma}$ is a global minimum point of (17.4), (17.5) and (17.6), if and only if (17.55) hold.

Proof. (i) Taking into account the definition of $\mathcal{E}(\mathcal{K}_{\bar{\Gamma}})$, it is enough to prove the equivalence between the second of (17.54) and the second of (17.55). Let the second of (17.54) hold. Because of Proposition 17.7(iii)–(v), $\exists \bar{\theta}, \bar{\lambda} > 0$ and $\exists(\bar{u}, \bar{v}) \in \mathbb{R}^2$, such that

$$\bar{\theta}u + \bar{\lambda}v \leq 0, \quad \forall (u, v) \in \mathcal{E}(\mathcal{K}_{\bar{\Gamma}}), \quad \bar{\theta}u + \bar{\lambda}v = 0 \Leftrightarrow (u, v) = (\bar{u}, \bar{v}).$$

Then, the maximum in the second of (17.55) is a non-negative function of (θ, λ) , which takes the value zero. Hence, the second of (17.55) follows. The reverse implication is obvious. (ii) Straightforward consequence of (i). \square

The left-hand side of the first of (17.55) is called *image dual problem*. In looking at the problems in (17.55), a question comes spontaneous: what kind of problem will we find, if the operators will be exchanged each other? Surprisingly, we do not find any problem, but just (17.45); this is expressed by the following:

Theorem 17.3 (Image Duality). *Let Condition 17.1 hold and $\bar{\Gamma} \in R$. We have*

$$\min_{\theta, \lambda > 0} \max_{(u, v) \in \mathcal{K}_{\bar{\Gamma}}} (\theta u + \lambda v) = \max_{(u, v) \in \mathcal{E}(\mathcal{K}_{\bar{\Gamma}})} \min_{\theta, \lambda > 0} (\theta u + \lambda v) = \max_{\substack{(u, v) \in \mathcal{K}_{\bar{\Gamma}} \\ v=0}} (u), \quad (17.56)$$

or

$$\min_{\theta, \lambda > 0} \max_{(u, v) \in \mathcal{E}(\mathcal{K}_{\bar{\Gamma}})} (\theta u + \lambda v) = \max_{(u, v) \in \mathcal{E}(\mathcal{K}_{\bar{\Gamma}})} \min_{\theta, \lambda > 0} (\theta u + \lambda v) = \max_{\substack{(u, v) \in \mathcal{E}(\mathcal{K}_{\bar{\Gamma}}) \\ v=0}} (u). \quad (17.57)$$

Proof. If $v \neq 0$, then the minimum in the second side of (17.56) may not exist and, in its place, the infimum – which is a function of (u, v) – is less than the value it takes at $v = 0$. Therefore, it is not restrictive to add the constraint $v = 0$ to the minimum in the second side of (17.56). Hence, taking into account that, due to the homogeneity of $\theta u + \lambda v$, it is not restrictive to assume $\theta = 1$, so that the minimization becomes obvious, the second equality of (17.56) follows. Between the first and second sides of (17.56) the inequality \geq holds as a special case of a well-known and classic inequality. The equality is a consequence of Proposition 17.7. A quite similar reasoning proves (17.57). \square

Once the IS analysis related to a given problem has been accomplished and some (image) statements have been proved in the IS, then such statements must be transferred to the given space, finding what we can call counterimage statements. To find image statements is, in general, much easier than to search for the counterimage statements directly in the given space; sometimes, in the given space it is difficult even to conceive a statement of this type. This is the main role of the IS analysis.

Now, let us write the counterimage statements of Proposition 17.8 and Theorem 17.3. To this end, consider the function

$$\mathcal{L}(\Gamma; \theta, \lambda) := \theta f(\Gamma) - \lambda g(\Gamma) = \int_T \int_T [\theta \frac{\rho}{4\pi} \frac{\Gamma'(x)\Gamma(y)}{y-x} - \lambda \rho V_\infty \frac{\Gamma(x)}{2}] dx dy, \quad (17.58)$$

which is called *Lagrangian function* associated to (17.4), (17.5) and (17.6). It expresses a (linear) combination of two entities, induced drag and the (difference between the) total lift (and a given constant, i.e., c). In the format (17.4), (17.5) and (17.6), the former is considered as an objective and the latter as a constraint. To adopt such a format is subjective. Therefore, why not to consider the *reciprocal problem*

$$\max[c + g(\Gamma)], \quad \text{subject to} \quad f(\Gamma) = d, \quad \Gamma \in \chi, \quad (17.59)$$

where d is a constant? In passing, we recall that the introduction of the reciprocal problem goes back to the ancient Greeks; under very general conditions (see [2], Sect. 17.5.5), it holds that, for suitable values of the constants c and d , a same Γ solves both (17.4), (17.5) and (17.6) and (17.59); this is known as *reciprocity principle*.

We will see that the theory of duality offers a way for overcoming the embarrassment of being obliged to choose between the formats (17.4), (17.5) and (17.6) and (17.59).

Proposition 17.10 *Let Condition 17.1 hold. $\bar{\Gamma} \in \chi$ is a (global) minimum point of (17.4), (17.5) and (17.6), if and only if $\exists(\bar{\theta}, \bar{\lambda}) > 0$, such that*

$$\mathcal{L}(\bar{\Gamma}; \theta, \lambda) = \mathcal{L}(\bar{\Gamma}; \bar{\theta}, \bar{\lambda}) \leq \mathcal{L}(\Gamma; \bar{\theta}, \bar{\lambda}), \quad \forall \Gamma \in \chi, \forall \theta, \lambda > 0. \quad (17.60)$$

Proof. The equality in (17.60) holds, iff $\bar{\Gamma} \in R$ or iff $\bar{\Gamma}$ is feasible for (17.4), (17.5) and (17.6). When such an equality holds, the inequality in (17.60) is equivalent to the optimality of $\bar{\Gamma}$, due to Proposition 17.8 (first inclusion). \square

Note that, unlike Proposition 17.8, in Proposition 17.10 $\bar{\Gamma}$ is assumed to merely belong to χ . A triplet $(\bar{\Gamma}, \bar{\theta}, \bar{\lambda})$ fulfilling (17.60) is called *saddle point* of \mathcal{L} ; the second side of (17.60) is the corresponding *saddle value*.

Theorem 17.4 (Duality). *Let Condition 17.1 hold and $\bar{\Gamma} \in R$. We have*

$$\max_{\theta, \lambda > 0} \min_{\Gamma \in \chi} \mathcal{L}(\Gamma; \theta, \lambda) = \min_{\Gamma \in \chi} \max_{\theta, \lambda > 0} \mathcal{L}(\Gamma; \theta, \lambda) = \min_{\Gamma \in R} f(\Gamma). \quad (17.61)$$

Proof. It is enough to use (17.43), Definition 17.4 and (17.58), and replace u and v in (17.56) with their expression in terms f, g, χ , and, finally apply Theorem 17.3. \square

The first side of (17.61) is called the *dual problem* of (17.4), (17.5) and (17.6), or of the third side of (17.61). The second side on (17.61) has been obtained from

the dual problem, by exchanging the order of the extremum operators; as announced above, it equals the primal problem. Note that the dual problem has nothing to share with the reciprocal problem, as it is easy to see, by comparing the first side of (17.61) with (17.59).

The IS analysis allows one to achieve several other important information. Proposition 17.7 – and, in particular, its (iii) – suggests the introduction of the following function:

$$u(\xi) := \max_{\substack{(u,v) \in \mathcal{K}_F \\ v=\xi}} (u) = \max_{\substack{(u,v) \in \mathcal{C}(\mathcal{K}_F) \\ v=\xi}} (u), \quad (17.62)$$

the second equality in (17.62) being due to Proposition 17.7(i). The function (17.62) is called *perturbation function* associated with (17.4), (17.5) and (17.6). When a problem does not enjoy a convexity property like (i) of Proposition 17.7, then the definition of the perturbation function is more general than (17.62). The perturbation function gives the value of the image problem (17.45) or (17.48), when the constraint $v = 0$ is replaced by $v = \xi$; since the image problem is related to (17.4), (17.5) and (17.6) by a relationship of type (17.46), then $u(\xi)$ allows one to know the change in the minimum in (17.4), (17.5) and (17.6) consequent to a change in the right-hand side of (17.5), where now zero is replaced by ξ . Furthermore, the (sub)derivative of $u(\xi)$ gives the (instantaneous) velocity of the minimum of (17.4), (17.5) and (17.6) with respect to the right-hand side of (17.5), namely ξ . In particular, at $\xi = 0$, the (instantaneous) velocity of the minimum of (17.4), (17.5) and (17.6) with respect to the right-hand side of (17.5) is given by $\bar{\lambda}/\bar{\theta}$. This number, which is nothing more than the classic Lagrangian multiplier, allows one to *evaluate the change in the minimum induced drag consequent to a change in the value at which the total lift is constrained*.

References

1. V. Boltyanski, H. Martini and V. Soltarr: *Geometric Methods and Optimization Problems*. Kluwer Academic Publishers, Dordrecht, 1999.
2. F. Giannessi: *Constrained optimization and Image Space Analysis*. Vol. 1: *Separation of Sets and Optimality Conditions*. Springer, New York, 2005.
3. F. Giannessi: *On the theory of Lagrangian duality*. Optimization Letters, Vol. 1, pp 9–20, Springer, New York, 2005.
4. Munk M.: The minimum induced drag in airfoils, NACA 121(1924).
5. Munk M.: Isoperimetrische Aufgaben aus der Theorie des Fluges, Inaugural Dissertation 1919, Gottinga (1919).
6. Prandtl L.: Induced Drag of Multiplanes, NACA TN 182 (1924).

Chapter 18

Plastic Hinges in a Beam

Danilo Percivale and Franco Tomarelli

Abstract This talk focuses minimization of one-dimensional free discontinuity problem with second-order energy dependent on jump integrals but not on the cardinality of the discontinuity set.

Related energies, describing loaded elastic–plastic beams, are not lower semi-continuous in BH (the space of displacements with second derivatives which are measures). Nevertheless we show that if a safe load condition is fulfilled then minimizers exist and they belong actually to SBH , say their second derivative has no cantor part. If in addition a stronger condition on load is fulfilled then minimizer is unique and belongs to the Sobolev space, H^2 . Moreover, we can always select one minimizer whose number of plastic hinges does not exceed two and is the limit of minimizers of penalized problems.

When the load stays in the gap between safe load and regularity condition then minimizers with hinges are allowed; if in addition the load is symmetric and strictly positive then there is uniqueness of minimizer, the hinges of such minimizer are exactly two and they are located at the endpoints.

If the beam is under the action of a skew-symmetric load then the safe load condition is less stringent than in the general case.

18.1 Elastic–Plastic Beam

Given f in $L^\infty(\mathbf{R})$, $E > 0$, $J > 0$, $\gamma > 0$, $L > 0$, we study the functional

Danilo Percivale

Dipartimento di Ingegneria della Produzione, Università di Genova, Piazzale Kennedy, Fiera del Mare Pad.D, Genova, Italy,

e-mail: percival@dimet.unige.it

Franco Tomarelli

Dipartimento di Matematica, Politecnico di Milano, 20133 Milano, Italy,

e-mail: franco.tomarelli@polimi.it

$$F(w) = \int_{\mathbf{R}} \left(\frac{EJ}{2} |\ddot{w}|^2 - fw \right) dx + \gamma \sum_{S_{\dot{w}}} |[\dot{w}]| \quad (18.1)$$

dependent on real-valued functions w with $\text{spt } w \subset [0, L]$ and w in $SBH(\mathbf{R})$ (e.g. w is an $L^1(\mathbf{R})$ function whose second derivative w'' is a Radon measure in \mathbf{R} without cantor part [1, 3, 4]). For any w in $SBH(\mathbf{R})$, \ddot{w} denotes the absolutely continuous part of w'' , $S_{\dot{w}}$ the singular set of $\dot{w} = w'$ and $[\dot{w}] = \dot{w}_+ - \dot{w}_-$.

Functional (18.1) describes the total energy associated to deformation of an elastic-plastic beam which is clamped at both endpoints and whose reference configuration is the horizontal interval $[0, L]$; w is the vertical displacement of the beam under the action of the vertical load f .

The crease points set $S_{\dot{w}}$ of a minimizer w may be interpreted as location of plastic hinges in the beam at equilibrium: functional (18.1) takes into account that the energy released in the deformation of a clamped elastic-plastic beam is the sum of elastic bending energy and of energy concentrated at plastic hinges. Jump points are not allowed (say $S_w = \emptyset$) for admissible displacements w which must be continuous.

The flexural rigidity EJ of the beam is given by the product of Young modulus E times the beam cross-section polar momentum of inertia J . The constant γ takes into account the resistance of the material to rotation at plastic hinges.

Unfortunately sequential w^*BH lower semicontinuity of functional (18.1) fails in SBH since absolutely continuous and jump part of w'' can merge in the limit.

We notice that (18.1) is convex, nevertheless compactness of minimizing sequences “a priori” may fail since the jump set $S_{\dot{w}}$ may be an infinite set even if $F(w) < \infty$.

We extend F to the whole BH with value $+\infty$ if $w \notin SBH$ or $\text{spt } w \not\subset [0, L]$ by defining $\mathcal{F}(w) : BH(\mathbf{R}) \rightarrow \mathbf{R} \cup \{+\infty\}$

$$\mathcal{F}(w) = \begin{cases} \int_{\mathbf{R}} \left(\frac{EJ}{2} |\ddot{w}|^2 - fw \right) dx + \gamma \sum_{S_{\dot{w}}} |[\dot{w}]|, & w \in SBH(\mathbf{R}), \text{spt } w \subset [0, L] \\ +\infty & \text{else.} \end{cases} \quad (18.2)$$

and we find a completely equivalent minimization problem (also \mathcal{F} fails to be lower semicontinuous with respect to the w^*BH topology).

The relaxed sequential w^*BH lower semi-continuous envelope of \mathcal{F} is difficult to handle, since it has an extra term containing the cantor part of second derivative and it takes into account the interplay between absolutely continuous and concentrated parts of energy.

The strategy to overcome this difficulty consists in three steps ([14], Sect. 18.2): first we introduce a sequence of penalized functionals \mathcal{F}^ε , which depend on parameter $\varepsilon > 0$ and are defined for every $w \in BH$:

$$\mathcal{F}^\varepsilon(w) = \begin{cases} \mathcal{F}(w) + \varepsilon \sharp(S_{\dot{w}}) & \text{if } w \in SBH(\mathbf{R}), \text{spt } w \subset [0, L] \\ +\infty & \text{else;} \end{cases} \quad (18.3)$$

then we study the nonconvex functionals \mathcal{F}^ε which are coercive and l.s.c. on BH but finite only in SBH ; eventually we jettison the parameter ε by showing that, provided the following safe load assumption on f holds true

$$\|f\|_{L^\infty} < 16 \frac{\gamma}{L^2}, \quad (18.4)$$

the minimizing sequences for \mathcal{F} are relatively compact in the w^*BH topology.

We prove that F , \mathcal{F} , its l.s.c envelope $sc^- \mathcal{F}$ and \mathcal{F}^* (the Γ limit of \mathcal{F}^ε) all achieve their minimum among w having support contained in $[0, L]$ and all their minima coincide [14]. Moreover, all minimizers w of F fulfil the following estimate for (absolutely continuous part of) bending moment $EJ\dot{w}$

$$\|\dot{w}\|_{L^\infty} \leq \gamma / (EJ),$$

and are balanced at creased points:

$$\dot{w}_\pm = \gamma \operatorname{sign}[\dot{w}] / (EJ).$$

The uniqueness of minimizer for F (or equivalently for \mathcal{F}) seems hard to tackle in general. Nevertheless we can always select minimizers of \mathcal{F} which have no more than two hinges. Moreover, we prove that strict sign of load (even without symmetry) entails that all minimizers exhibit no more than two hinges [12]. Under additional assumption (e.g. symmetry and strict sign of load) also uniqueness for minimizer of F holds true together with an explicit representation formula of minimizer [14].

The penalized functional (18.3) takes into account the total energy related to deformation of an elastic-plastic beam: the four terms correspond (in their order, referring to (18.2)) to the elastic bending energy, potential energy and concentrated plastic yielding together with a minimal threshold cost ε for the formation of any plastic hinge: functional (18.3) was deduced as a gamma limit by two-dimensional or three-dimensional thick approximation of the beam (see [8–11]) starting from classic models of damage [2, 7].

In a different framework (allowing for L^1 or even Radon measure load f) we showed a safe load condition ($\|f\|_{L^1} < 8\gamma/L$) and a regularity load condition ($\|f\|_{L^1} \leq 27\gamma/(4L)$) entailing, respectively, existence and H^2 regularity for minimizers of \mathcal{F}_ε : such gap between the safe and regularity load condition is very narrow and makes it difficult to check whether creased minimizers exist (actually they do exist [13]).

In [14] we deal with L^∞ load and in this framework we prove a sharp L^∞ safe load condition (i.e. $\|f\|_{L^\infty} < 16\gamma/L^2$) and also a sharp L^∞ regularity load condition (i.e. $\|f\|_{L^\infty} \leq 12\gamma/L^2$), entailing, respectively, existence and regularity for minimizers of both \mathcal{F}^ε and F : in this context we can prove that for any symmetric load f which stays in the gap and has a strict sign, then the minimizer is unique and has exactly two hinges located at the endpoints of the beam. The result is obtained by sharp estimates on the Green function and careful comparison between candidate minimizers.

Our analysis proves that the structures do not develop plastic hinges if the resistance γ fulfils

$$\gamma \geq \frac{L^2}{12} \|f\|_{L^\infty},$$

say a condition which entails that maximum bending moment of the purely elastic solution [16] does not exceed γ .

For generic data f in L^∞ , we show Euler–Lagrange equations and a Compliance Identity fulfilled by extremals of F (Theorems 3.1, 3.2 in [14]): they provide the essential tools in the comparison between competing functions with the aim of selecting minimizers with relevant qualitative properties, without quantitative knowledge about their derivative jumps.

We show an explicit formula for the gamma limit \mathcal{F}^* of \mathcal{F}^ε and show that the same L^∞ safe and regularity condition above (valid for $F, \mathcal{F}, \mathcal{F}^\varepsilon$) apply also to \mathcal{F}^* :

$$\begin{aligned} \mathcal{F}^*(w) &:= \Gamma(w^*BH) \lim_{\varepsilon \rightarrow 0} \mathcal{F}^\varepsilon(w) = \\ &= \begin{cases} \int_{\mathbf{R}} (\varphi^{**}(\ddot{w}) - fw) dx + \gamma \left(\sum_{S_{\ddot{w}}} |[w]| + |(w'')^c|_T \right) & \forall w \in BH : \text{spt } w \subset [0, L], \\ +\infty & \text{else.} \end{cases} \end{aligned} \quad (18.5)$$

where $(w'')^c$ is the cantor part of w'' , $|\cdot|_T$ denotes the total variation in \mathbf{R} and

$$\varphi^{**}(s) = \begin{cases} (EJ/2)s^2 & \text{if } |s| \leq \gamma/(EJ) \\ \gamma|s| - \gamma^2/(2EJ) & \text{if } |s| > \gamma/(EJ). \end{cases} \quad (18.6)$$

We emphasize that the estimate $|\ddot{w}| \leq \gamma/(EJ)$ a.e. in $(0, L)$ (proven for any minimizer w of \mathcal{F}) entails

$$\varphi^{**}(\ddot{w}) = \varphi(\ddot{w}) \quad \text{and} \quad \mathcal{F}^*(w) = \mathcal{F}(w) = F(w) \quad \forall w \in \text{argmin } \mathcal{F}. \quad (18.7)$$

All functionals $F, \mathcal{F}, \mathcal{F}^\varepsilon, \mathcal{F}^*$ refer to relaxed homogeneous Dirichlet boundary conditions (the beam is clamped at both endpoints); nevertheless, minimizers with hinges located at the boundary are not excluded: if this phenomenon takes place then also boundary creases add a positive cost in the energy.

The structure of minimizers under general symmetric load with a strict sign is completely described by main results ([14] Theorem 3.8). In the simple case of constant load $f \equiv -\lambda$, $\lambda > 0$, this analysis provides the following complete picture as long as λ increases:

- for $0 \leq \lambda \leq 12\gamma/L^2$, F has exactly one minimizer which turns out to be $C^3(\mathbf{R})$, say we are in the elastic regime;
- for $12\gamma/L^2 < \lambda < 16\gamma/L^2$, F still has exactly one minimizer but there is the development of two plastic hinges at the boundary;
- for $\lambda > 16\gamma/L^2$ there is collapse: the infimum of F is $-\infty$;
- in all the range $0 \leq \lambda < 16\gamma/L^2$ the minimizer is given by

$$z_\lambda(x) = -\lambda x^2(x-L)^2/(24EJ) - \frac{1}{(2EJ)}(\lambda L^2/12 - \gamma)^+ x(L-x),$$

the bending moment $EJ\ddot{z}_\lambda$ never exceeds γ , say $|\ddot{z}_\lambda(x)| \leq \gamma/(EJ)$, and

$$\min F = F(z_\lambda) = -\frac{EJ}{2} \int_0^L |\ddot{z}_\lambda|^2 = -\frac{1}{1440EJ} \lambda^2 L^5 - L \left((\lambda L^2/12 - \gamma)^+ \right)^2 / (2EJ);$$

- for $12\gamma/L^2 < \lambda < 16\gamma/L^2$ the (unique and creased) minimizer fulfils also

$$EJ\ddot{z}_\lambda(0) = \gamma \operatorname{sign}[\dot{z}_\lambda](0) = EJ\ddot{z}_\lambda(L) = \gamma \operatorname{sign}[\dot{z}_\lambda](L) = -\gamma$$

which express Euler–Lagrange identities for a minimizer with hinges at endpoints.

As far as it concerns penalized functionals, \mathcal{F}^ε , we remark that nonuniqueness phenomena may occur even for constant load: both smooth and creased minimizers may appear for suitable choice of constant load and parameter ε (Example 3.16 in [14]).

About the analysis of elastic–plastic plate we refer to [4–6, 10].

18.2 Skew-Symmetric Load

In case of skew-symmetric load, a less stringent safe load condition than (18.4) holds true (condition (18.4) is optimal for generic load and sharp for constant load [14]) as clarified by the following result ([15] Theorem 3.4).

Theorem 18.1. *Assume that $f \in L^\infty(\mathbf{R})$ fulfils*

$$\|f\|_{L^\infty(0,L)} < 8(3 + 2\sqrt{2}) \frac{\gamma}{L^2} \quad (\text{skew safe load condition}) \quad (18.8)$$

and

$$f(x) = f(L - x). \quad (18.9)$$

Then the functional (18.1) achieves a finite minimum and

$$\|\ddot{w}\|_{L^\infty} \leq \gamma \quad \forall w \in \operatorname{argmin} \mathcal{F}. \quad (18.10)$$

Moreover there is at least one minimizer w of \mathcal{F} such that $\sharp(S_w) \leq 2$ and

$$w(x) = -w(L - x).$$

Acknowledgments This research was partially supported by Italian M.U.R. (PRIN 2006, project Variational Problems with Multiple Scales)

References

1. Ambrosio L., Fusco N., Pallara D: Functions of Bounded Variation and Free Discontinuity Problems, Oxford Math. Mon., Oxford Univ. Press (2000)
2. Barenblatt G.I.: The formation of equilibrium cracks during brittle fracture, general ideas and hypotheses. Axially symmetric cracks. Appl. Math. Mech. (PMM) **23**, 622–636 (1959)
3. Carriero M., Leaci A. & Tomarelli F.: Plastic free discontinuities and special bounded hessian, C. R. Acad. Sci. Paris Sér. I Math. **314** no. 8, 595–600 (1992)
4. Carriero M., Leaci A. & Tomarelli F.: Special Bounded Hessian and elastic-plastic plate, Rend. Accad. Naz. Sci. XL, Mem. Mat. **5** no. 16, 223–258 (1992)
5. Carriero M., Leaci A. & Tomarelli F.: Strong solution for an Elastic Plastic Plate, Calc. Var. Partial Dif. Equ. **2** no. 2, 219–240 (1994)
6. Carriero M., Leaci A. & Tomarelli F.: Second Order Variational Problems with Free Discontinuity and Free Gradient Discontinuity. In: Calculus of Variations: Topics from the Mathematical Heritage of Ennio De Giorgi, Quad. Mat., 14, Depth. Math., Seconda Univ. Napoli, Caserta, 135–186 (2004)
7. Griffith A.A.: The phenomenon of rupture and flow in solids, Phyl.Trans. Roy.Soc. A, **221**, 163–198 (1920)
8. Percivale D.: Perfectly plastic plates, a variational definition, J.Reine Angew. Math. **411**, 39–50 (1990)
9. Percivale D. & Tomarelli F.: Scaled Korn-Poincaré inequality in BD and a model of elastic plastic cantilever, Asymptot. Anal. **23**, no. 3–4, 291–311 (2000)
10. Percivale D. & Tomarelli F.: From SBD to SBH: the elastic plastic plate, Interface Free Bound. **4**, no.2, 137–165 (2002)
11. Percivale D. & Tomarelli F.: From Special Bounded Deformation to Special Bounded Hessian: the elastic plastic beam, Math. Mod. Meth. Appl. S. **15**, no.7, 1009–1058 (2005)
12. Percivale D. & Tomarelli F.: Smooth and creased equilibria for elastic-plastic plates and beams. In: G. Dal Maso et al. (eds.) Variational Problems in Material Science, pp. 127–136, Progr. Nonlinear Differential Equations Appl., 68, Birkhäuser, Basel (2006)
13. Percivale D. & Tomarelli F.: Regular minimizers of free discontinuity problems, to appear.
14. Percivale D. & Tomarelli F.: A variational principle for plastic hinges in a beam, to appear.
15. Percivale D. & Tomarelli F.: How to prevent collapse of elastic-plastic beams under skew-symmetric load, to appear.
16. Villaggio P.: Mathematical Models for Elastic Structures, Cambridge Univ. Press (1997)

Chapter 19

Problems of Minimal and Maximal Aerodynamic Resistance

Alexander Plakhov

Abstract This is a review of results recently obtained by the author, related to problems of the body of minimal and maximal resistance. The cases of purely translational motion, as well as rotational and translational motions, are considered. The notions of rough body and law of scattering on a body are discussed. Connections with the Monge–Kantorovich problem of optimal mass transportation are revealed, and applications to the Magnus effect and retroreflectors are discussed.

19.1 Introduction

Consider a body moving through a homogeneous rarefied gas. It is assumed that

- (i) collisions of the body with the gas molecules are absolutely elastic;
- (ii) the gas is indeed very rare, so that mutual interaction of the molecules can be neglected; and
- (iii) there is no thermal motion in the gas.

Due to collisions of the body with the molecules, there is created a force: the gas resists the body's motion. The problem addressed in this chapter is: **Find the body having minimal or maximal resistance.**

The motivating examples are as follows:

- an *artificial satellite* moving around the Earth on low (100–200 km) altitudes (one needs to *minimize* the resistance)
- and
- *solar sail* (driving force \equiv resistance should be *maximized*).

Below we shall consider two different kinds of motion:

Alexander Plakhov

University of Wales – Aberystwyth, Aberystwyth SY23 3BZ, UK; on leave from Aveiro University, Portugal, e-mail: axp@aber.ac.uk

1. The body performs purely **translational** motion (Sect. 19.2);
2. It performs both **translational** and **rotational** motions (Sect. 19.3; we shall consider only the two-dimensional case).

To state the problem rigorously, one has to specify the classes of admissible bodies. This will be made in the forthcoming sections.

19.2 Translational Motion

Consider a reference system connected with the body, so that the body B does not move, and there is a homogeneous flow of particles moving vertically downward with the velocity $v = (0, 0, -1)$.

For an incident particle with the coordinates $(x, y, -t)$, denote by $v_B^+(x, y)$ the final velocity. That is, the particle initially moves with the velocity v , then makes several reflections from the body B , and finally gets away with the velocity $v_B^+(x, y)$. The momentum the particle transmits to the body is $\mu(v - v_B^+(x, y))$, where μ is the particle's mass. Summing up over all incident particles per unit time, one obtains that the resistance equals $-\rho R(B)$, where ρ is the flow density and

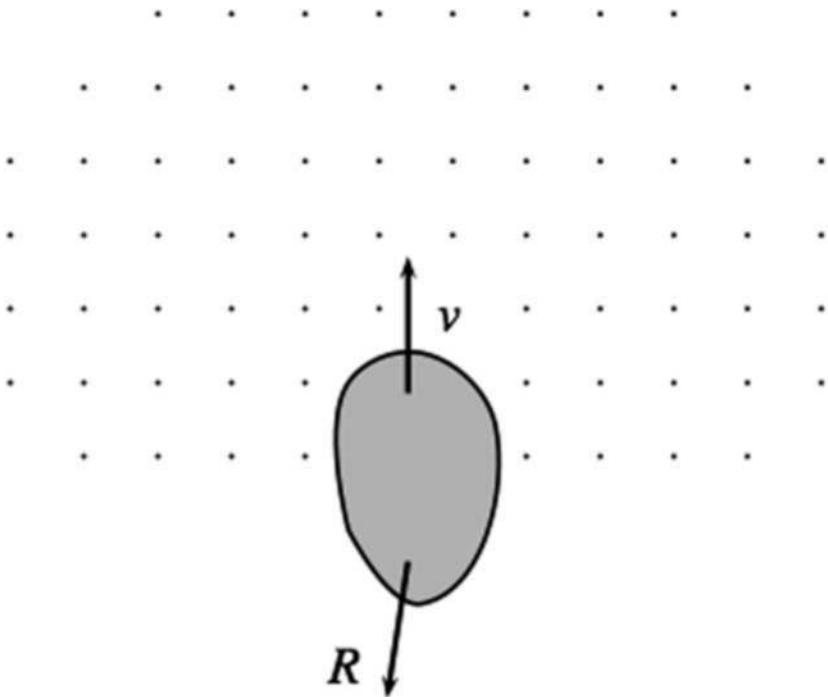


Fig. 19.1 Body moving in a rarefied medium

$$R(B) = \iint_{\mathbb{R}^2} (v_B^+(x, y) - v) dx dy. \quad (19.1)$$

Usually the problem reads as follows: minimize the modulus of the third component $|R_z(B)|$ of the resistance vector $R(B) = (R_x(B), R_y(B), R_z(B))$. Note, however, that in all known cases the body minimizing $|R_z|$ also minimizes $|R| = \sqrt{R_x^2 + R_y^2 + R_z^2}$ and vice versa; that is, these two minimization problems are equivalent.

Let us first briefly examine three classes of admissible bodies of fixed length and width:

- (I) convex and axially symmetric;
- (II) convex (generally nonsymmetric);
- (III) generally nonconvex and nonsymmetric.

By saying that a body has fixed length and width we mean that it is inscribed in a right circular cylinder with fixed length and radius, with the axis parallel to the direction of motion, and the projections of the body on the axis and on the plane orthogonal to the axis coincide with the corresponding projections of the cylinder.

Note that one could give another quite natural, but more restrictive, definition of a body of fixed length and width. Namely, the body must be contained in a fixed right circular cylinder and contain at least one orthogonal cross-section of it. In classes (I) and (II) this definition is equivalent to the former one. In class (III), the minimal resistance under this definition is equal to the minimal resistance under the former one (both are equal to zero), but the minimizing sequence of bodies should be constructed more carefully. See [1] for details.

(I) Newton in *Principia* [2] considered the problem of minimal resistance in the class of **convex** and **axially symmetric** bodies. The solution, for a special choice of the parameters (length)/(maximal width of the body) = 0.75, is shown in Fig. 19.2.

(II) In 1990s the problem was considered for **convex** (generally **nonsymmetric**) bodies (Buttazzo, Kawohl, Lachand-Robert et al.). The solution exists and does not coincide with Newton's one [3–5]. The problem is not completely solved till now, but there were found numerical solutions [6] and analytical solutions in the subclass of bodies with developable lateral surface [7].

Note that in the classes (I) and (II) the upper boundary of the body is the graph of a concave function $F : \Omega \rightarrow \mathbb{R}$, where $\Omega \subset \mathbb{R}^2$ is a circle, and the functional to be minimized (19.1) takes the simple analytic form

$$R(F) = 2 \iint_{\Omega} \frac{1}{1 + |\nabla F(x, y)|^2} dx dy.$$

In class (I) the function is radial, $F(x, y) = f(\sqrt{x^2 + y^2})$, and the functional has an even simpler form:

$$R(F) = 4\pi \int_0^L \frac{r}{1 + f'^2(r)} dr,$$

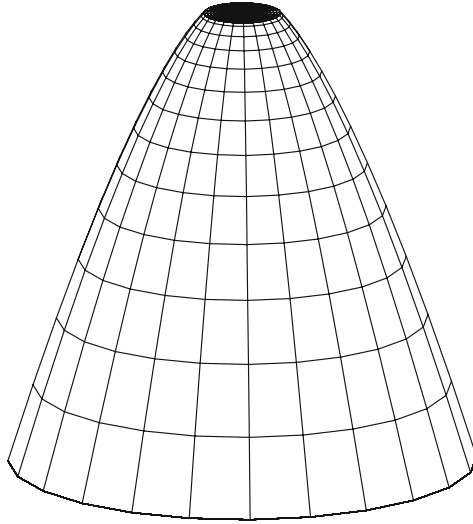


Fig. 19.2 Solution of Newton's problem

where L is the radius of the circle Ω . The problem is to minimize the functional $R(F)$ over all concave functions F such that $\sup F - \inf F \leq M$ (respectively, minimize the functional $R(f)$ over all concave monotone nonincreasing functions f such that $\sup f - \inf f \leq M$). Here L and M are the parameters of the problem.

(III) The solution in the class of **nonconvex** and **nonsymmetric** bodies is surprisingly simple. The answer is $\inf(\text{Resistance}) = 0$ [1]. Probably the infimum cannot be attained.

We will restrict ourselves to giving two ideas of solution, as illustrated in Figs. 19.3 and 19.4. These figures represent two-dimensional bodies equipped with several parabolic mirrors. The boundary of the mirrors and of the “central part” of the body is composed of parabolic arcs and straight line intervals. In Fig. 19.3, there are shown four pairs of parabolic arcs with common foci (shown by points) and vertical axes. The four small mirrors have the size ε . The motion of particles is shown by arrows; one can see that for most particles the equality $v^+ = v$ holds true, and only for a small portion of particles of order ε one has $v^+ \neq v$ (actually, $v^+ = -v$). Therefore, the resistance is of order ε .

In Fig. 19.4, there are also several pairs of confocal parabolas. Each pair consists of a parabolic arc with vertical axis and a small parabolic arc of size ε with horizontal axis. Here we have a similar behavior: all particles, except for a small portion of order ε , satisfy the relation $v^+ = v$, therefore the resistance is of order ε .

The three-dimensional bodies of small resistance are bodies of revolution obtained by rotating the corresponding figures with respect to the vertical axis. One should also add thin vertical “rods” connecting the mirrors with the main part of the body. The construction shown in Fig. 19.3 is applicable only if $(\text{length})/(\text{width of the body}) > 1/2$, while the construction in Fig. 19.4 is valid for arbitrary parameters of the problem.

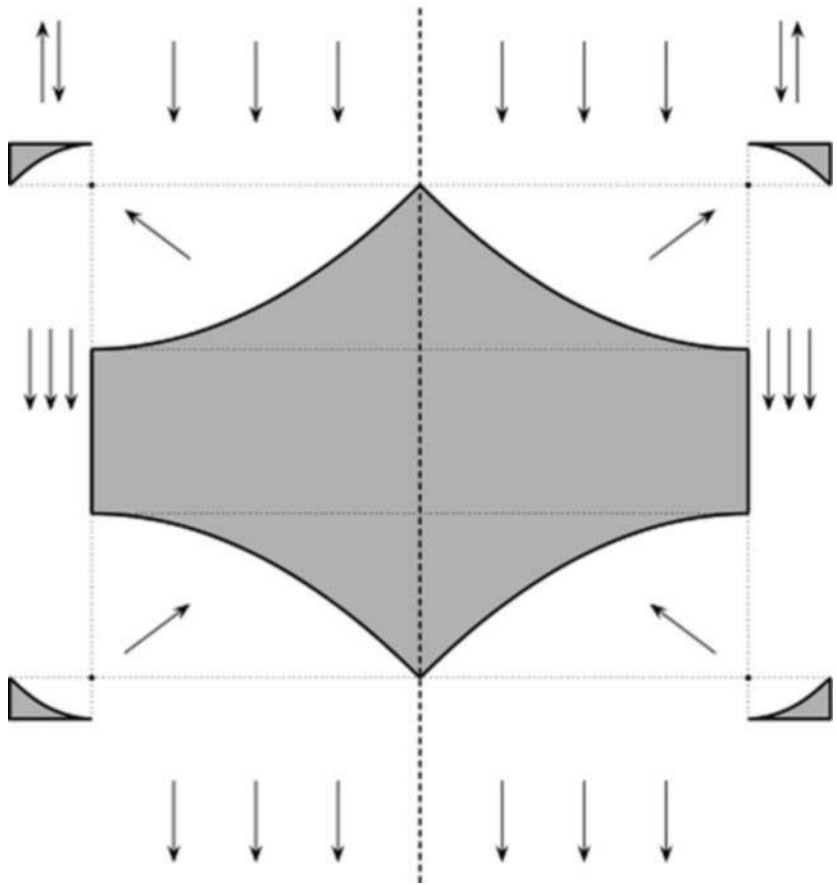


Fig. 19.3 Solution for the case $(\text{length})/(\text{width}) > 1/2$

Consider one more class of admissible bodies. Let B_1 and B_2 be two connected bounded sets such that $\mathcal{U}_\varepsilon(B_1) \subset B_2 \subset \mathbb{R}^3$, where $\mathcal{U}_\varepsilon(B_1)$ is the ε -neighborhood of B_1 .

(IV) The class of connected bodies B with piecewise smooth boundary such that $B_1 \subset B \subset B_2$.

The infimum of resistance in this class is also zero. Figure 19.5 gives an idea of the solution in the case where B_1 and B_2 are concentric cubes with the edges parallel to the coordinate axes. On this two-dimensional picture there is shown a square with several subsets (channels) removed. Each channel is the union of an input funnel, a tube, and an output funnel. The input and output funnels are trapezia obtained from an isosceles triangle with height ε and base $2\varepsilon^2$ by cutting off a smaller triangle with height ε^2 and base $2\varepsilon^3$. Each tube is a union of five “thin” rectangles and four quarters of circle of radius $2\varepsilon^3$. The “width” of rectangles is $2\varepsilon^3$, and the number

of channels is of order ε^{-2} . The parameter ε is chosen smaller than (size of B_2) – (size of B_1).

Each incident particle gets into one of the channels with the velocity $v = (0, 0, -1)$, then moves through the channel downwards, and finally gets out of the channel, the final velocity v^+ being almost equal to v : $v^+ - v = O(\varepsilon)$. This estimate is uniform over all incident particles.

The resulting body B_ε is a “sandwich”: each vertical cross-section of it parallel to the plane Oxz is similar either (a) to Fig. 19.5 or (b) to the square. The set of values y such that the cross-section is the square has measures less than ε . The resistance of B_ε is of order ε .

Indeed, formula (19.1) can be rewritten here as

$$R(B_\varepsilon) = \iint_{\Omega_\varepsilon^1} O(\varepsilon) dx dy + \iint_{\Omega_\varepsilon^2} 2 dx dy.$$

Here $\Omega_\varepsilon^1 \cup \Omega_\varepsilon^2$ is the projection of B_ε on the plane Oxy , that is, a square; the points of Ω_ε^1 belong to cross-sections of kind (a), and the points of Ω_ε^2 belong to cross-section of kind (b). The measure of Ω_ε^2 is of order ε ; therefore, $R(B_\varepsilon) = O(\varepsilon)$.

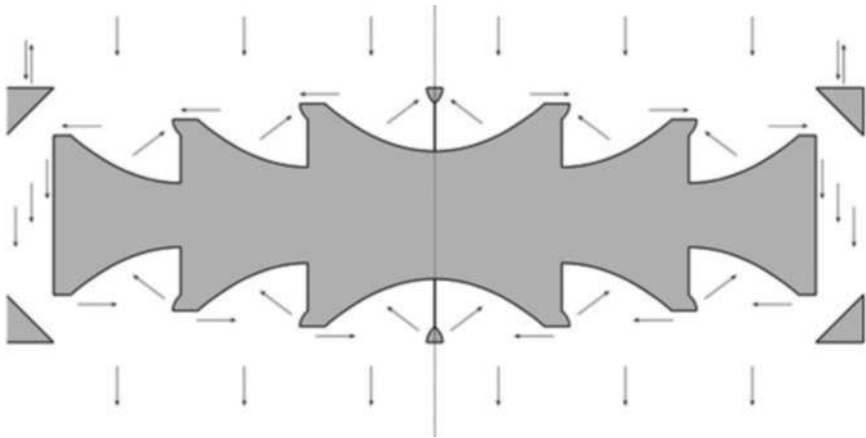


Fig. 19.4 Solution for arbitrary length and width

The result in class (IV) can be stated as follows: any body can be “damaged” in the ε -neighborhood of its boundary in such a way that the resulting body will have resistance less than ε . Damaging amounts to making a huge number of thin channels near the boundary: the incident flow gets into and passes through the channels and then gets out, the final velocity being almost equal to the velocity of incidence.

The mathematical solutions in the nonconvex cases (classes (III) and (IV)) do not seem very practical. They are not applicable in case of thermal motion of gas molecules or some rotational motion of the body.

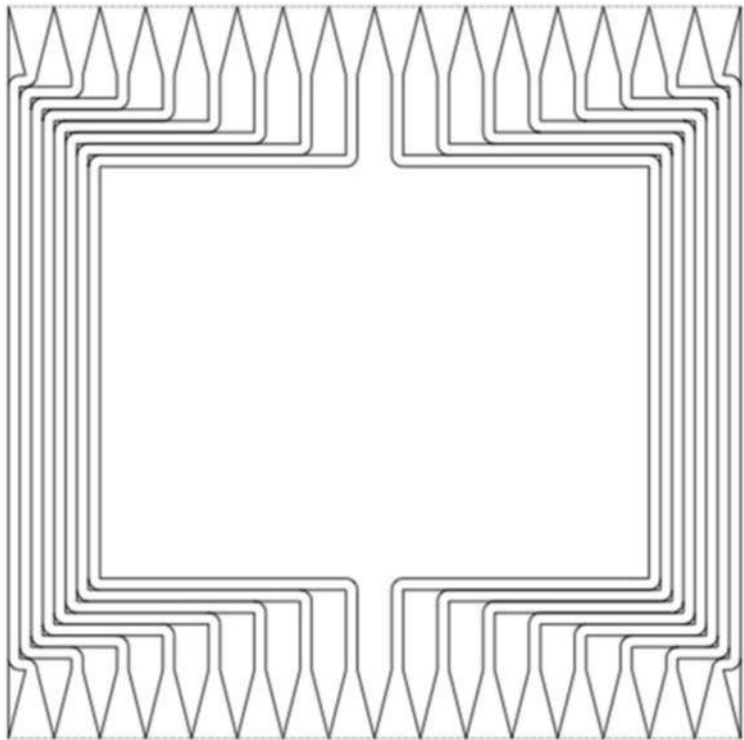


Fig. 19.5 Solution in the class (IV)

19.3 Translational Motion with Rotation: Two-Dimensional Case

19.3.1 Definition of Rough Body and Main Theorems

First introduce the notion of *rough body*. Consider three examples.

- (i) A circle.
- (ii) A “rough circle”. (There are “very small” triangular hollows on the boundary. Each hollow is a right isosceles triangle.)
- (iii) Another “rough circle” (with small rectangular hollows on the boundary).

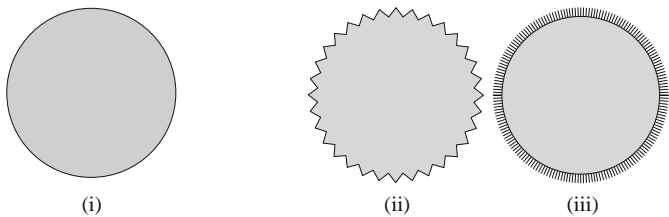


Fig. 19.6 Three bodies

The diagrams in Fig. 19.7 describe the laws of reflection from these bodies. Here φ is the angle of incidence and φ^+ is the angle of reflected particle, both angles vary between $-\pi/2$ and $\pi/2$ and are measured between the inner (outer) normal to the circumference and the velocity of incidence (of reflection).

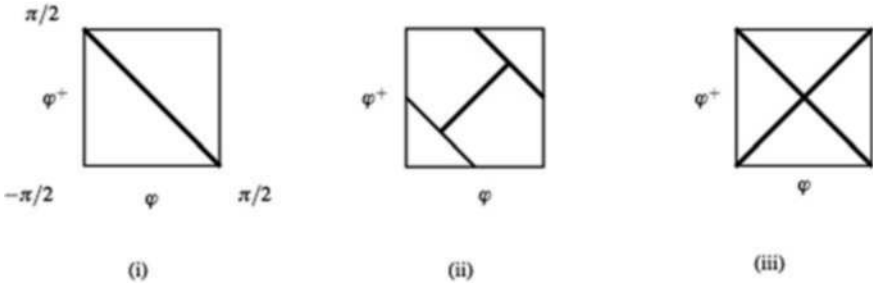


Fig. 19.7 Diagrams of reflection

In general, a *rough body* (= a body obtained by *roughening* a convex set B) is a set that coincides with B from the *macroscopical* point of view, but has *microscopical* irregularities on its boundary.

With any rough body one can associate a measure ν on the set $[-\pi/2, \pi/2] \times [-\pi/2, \pi/2] \times S^1$. Namely, randomly choose a particle incident on the body and detect (φ, φ^+, n) , where $\varphi \in [-\pi/2, \pi/2]$ is the angle of incidence, $\varphi^+ \in [-\pi/2, \pi/2]$ is the angle of reflected particle, and $n \in S^1$ is the outer unit normal to ∂B at the point where the particle hits B . The measure ν (**law of scattering**) is just the joint probability distribution of the triple (φ, φ^+, n) . We assume that φ and φ^+ are measured clockwise: φ is measured from $-n$ to the velocity of incidence, and φ^+ from n to the velocity of reflection.

Let us now give the rigorous definition of a body obtained by roughening B or just rough body.

Let n_x be the outer unit normal to ∂B at the point $x \in \partial B$. Consider a set with piecewise smooth boundary $\Omega \subset B$. Define the mapping $\varphi_{\Omega, B}^+ : \partial B \times [-\pi/2, \pi/2] \rightarrow [-\pi/2, \pi/2]$ as follows. For any $x \in \partial B$, $\varphi \in [-\pi/2, \pi/2]$, a particle starts moving at x , and the initial velocity makes the angle φ with $-n_x$. Then the particle makes several reflections from $\partial\Omega$, and finally, intersects ∂B again at a point x^+ and goes out of B , with the angle the velocity at that point makes with n_{x^+} being $\varphi_{\Omega, B}^+(x, \varphi)$. Introduce the shorthand notation $\square := [-\pi/2, \pi/2] \times [-\pi/2, \pi/2]$ and define the measure $\nu_{\Omega, B}$ on $\square \times S^1$ in the following way: for any Borel set $A \subset \square \times S^1$, $\nu_{\Omega, B}(A) := \mu \left(\{(x, \varphi) : (\nu, \varphi_{\Omega, B}^+(x, \varphi), n_x) \in A\} \right)$. Here $d\mu = \cos \varphi dx d\varphi$.

Definition 19.1. A sequence of sets with piecewise smooth boundary $\Omega_n \subset B$ represents a rough body \mathcal{B} , if

- (i) $\text{area}(B \setminus \Omega_n) \rightarrow 0$ as $n \rightarrow \infty$;
- (ii) there exists a $*$ -weak limit $\nu_{\mathcal{B}} := \lim_{n \rightarrow \infty} \nu_{\Omega_n, B}$.

Two sequences Ω_n and Ω'_n are equivalent, if the corresponding limiting measures coincide. By definition, a **rough body** \mathcal{B} is a class of equivalence of such sequences. It will also be called a **body obtained by roughening** B .

The measure $\nu_{\mathcal{B}}$ corresponding to a rough body \mathcal{B} is called the **law of scattering** on \mathcal{B} .

The following definition and theorem characterize the set of all possible scattering laws.

Let λ be the measure on $[-\pi/2, \pi/2]$ such that $d\lambda(\varphi) = \frac{1}{2} \cos \varphi d\varphi$. Let τ_B be the surface measure of B . Let $\pi_{\varphi, n}$, $\pi_{\varphi^+, n}$, and π_{φ, φ^+} be the projections on the corresponding subspaces; for example, $\pi_{\varphi, n} : (\varphi, \varphi^+, n) \mapsto (\varphi, n)$. Let $\pi_d : (\varphi, \varphi^+, n) \mapsto (\varphi^+, \varphi, n)$.

Definition 19.2. \mathcal{M}_B is the set of measures ν on $\square \times S^1$ such that

- (i) $\pi_{\varphi, n}^\# \nu = \lambda \otimes \tau_B = \pi_{\varphi^+, n}^\# \nu$;
- (ii) ν is invariant with respect to the exchange $\varphi \leftrightarrow \varphi^+$; that is, $\pi_d^\# \nu = \nu$.

Here $\pi^\# \nu$ means the push-forward of the measure ν by the mapping π .

Note in passing that $\nu(\square \times S^1) = |\partial B|$ for any $\nu \in \mathcal{M}_B$. Indeed, according to definition 19.2, $\nu(\square \times S^1) = \lambda \otimes \tau_B([-\pi/2, \pi/2] \times S^1) = \lambda([-\pi/2, \pi/2]) \cdot \tau_B(S^1) = 1 \cdot |\partial B|$.

Definition 19.3. \mathcal{M} is the set of measures η on \square such that

- (iii) $\pi_{\varphi}^\# \eta = \lambda = \pi_{\varphi^+}^\# \eta$, where $\pi_{\varphi}(\varphi, \varphi^+) = \varphi$ and $\pi_{\varphi^+}(\varphi, \varphi^+) = \varphi^+$;
- (iv) η is invariant with respect to the exchange $\varphi \leftrightarrow \varphi^+$.

Denote by $p\mathcal{M}$ the set of measures $p\eta$, $\eta \in \mathcal{M}$, where $p \in \mathbb{R}$. Consider the bodies obtained by roughening a convex body B .

Theorem 19.1. *The set of measures ν generated by these rough bodies coincides with \mathcal{M}_B .*

The measure $\eta_\Omega := \pi_{\varphi, \varphi^+}^\# \nu_{\Omega, \text{conv} \Omega}$ on \square contains important information about particle scattering on Ω . In particular, if Ω is convex, like in Fig. 19.6(i), then $\eta_\Omega = |\partial \Omega| \cdot \eta_0$, where $|\partial \Omega|$ is the length of $\partial \Omega$ and $d\eta_0 = \frac{1}{2} \cos \varphi \cdot \delta(\varphi + \varphi^+) d\varphi d\varphi^+$. If, like in Fig. 19.6(ii), the hollows on the boundary of a nonconvex body Ω are right isosceles triangles (a hollow is a connected component of $\text{conv} \Omega \setminus \Omega$), then $\eta_\Omega = |\partial(\text{conv} \Omega)| \cdot \eta_\nabla$, where $d\eta_\nabla = \cos \varphi [\chi_{[-\pi/2, -\pi/4]}(\varphi) \delta(\varphi + \varphi^+ + \frac{\pi}{2}) + \chi_{[-\pi/4, \pi/4]}(\varphi) \delta(\varphi - \varphi^+) + \chi_{[\pi/4, \pi/2]}(\varphi) \delta(\varphi + \varphi^+ - \frac{\pi}{2})] + |\sin \varphi| [\chi_{[-\pi/4, 0]}(\varphi) \delta(\varphi + \varphi^+ + \frac{\pi}{2}) - \chi_{[-\pi/4, \pi/4]}(\varphi) \delta(\varphi - \varphi^+) + \chi_{[0, \pi/4]}(\varphi) \delta(\varphi + \varphi^+ - \frac{\pi}{2})] d\varphi d\varphi^+$. One can verify that the measures η_0 and η_∇ belong to \mathcal{M} . Their supports are shown in Fig. 19.7 (i) and (ii).

Theorem 19.2. *The set of measures $\{\eta_\Omega; |\partial(\text{conv} \Omega)| = p\}$ is an everywhere dense subset of $p\mathcal{M}$ in the $*$ -weak topology.*

These theorems are useful tools for solving problems of minimal and maximal aerodynamic resistance. By using them, one can reduce various problems of optimal resistance to special problems of optimal mass transportation. The proof of theorem 19.1 can be found in [8], and theorem 19.2 is a direct consequence of theorem 2 from [9].

Let us consider some problems of resistance optimization.

19.3.2 Problems of Minimal and Maximal Resistance for a Slowly Rotating Body

A body (bounded connected set with piecewise smooth boundary) Ω moves through a rarefied gas and rotates slowly and uniformly; see Fig. 19.8. The force of the medium resistance acting on Ω is a vector-valued function of time, $R_t(\Omega)$. Let the period of rotation be T ; the function $R_t(\Omega)$ is periodic with the same period. We are interested in minimizing the modulus of the mean value of this function, $|\bar{R}(\Omega)|$, where $\bar{R}(\Omega) = \frac{1}{T} \int_0^T R_t(\Omega) dt$.

Let the body move vertically upwards with the velocity $(0, 1)$. In a reference system connected with the body, that is, moving upwards with the same velocity, one observes a flow of particles of velocity $v = (0, -1)$ falling down on the fixed body. The momentum transmitted by a particle to the body equals $\mu(v - v^+)$, where μ is the mass and v^+ is the final velocity of the particle. Let the particle first intersect $\partial(\text{conv } \Omega)$ at a point x and the vectors v and v^+ form angles φ and φ^+ , respectively, with the vector n_x . Here n_x is the outer normal to $\partial(\text{conv } \Omega)$ at x . Then the transmitted momentum can be written down as $\mu(\sin(\varphi - \varphi^+), 1 + \cos(\varphi - \varphi^+))$. Integrating this value over the measure η_Ω and properly choosing the flow density, one gets the mean resistance:

$$\bar{R}(\Omega) = -\frac{3}{4} \iint_{\square} (\sin(\varphi - \varphi^+), 1 + \cos(\varphi - \varphi^+)) d\eta_\Omega(\varphi, \varphi^+).$$

The factor $3/4$ is chosen for further convenience, as will be seen later.

Taking into account that η_Ω is symmetric and $\sin(\varphi - \varphi^+)$ is antisymmetric with respect to φ and φ^+ , one concludes that the first component of $\bar{R}(\Omega)$ is zero, therefore

$$|\bar{R}(\Omega)| = \frac{3}{4} \iint_{\square} (1 + \cos(\varphi - \varphi^+)) d\eta_\Omega(\varphi, \varphi^+).$$

Problem 19.1 Find

- (a) the body of minimal mean resistance;
- (b) the convex body of minimal mean resistance, provided that the body's area is fixed:

$$(a) \inf \{ |\bar{R}(\Omega)|; \text{Area}(\Omega) = a \text{ and } \Omega \text{ is convex} \};$$

$$(b) \inf_{\text{Area}(\Omega)=a} |\bar{R}(\Omega)|.$$

(a) If Ω is convex then

$$\begin{aligned} |\bar{R}(\Omega)| &= \frac{3}{4} \iint_{\square} (1 + \cos(\varphi - \varphi^+)) \cdot |\partial\Omega| d\eta_0(\varphi, \varphi^+) = \\ &= |\partial\Omega| \cdot \frac{3}{4} \int_{-\pi/2}^{\pi/2} (1 + \cos 2\varphi) \frac{1}{2} \cos \varphi d\varphi = |\partial\Omega|, \end{aligned}$$

thus one comes to the classical isoperimetric problem:

minimize $|\partial\Omega|$ provided that $\text{Area}(\Omega) = a$.

The solution is a circle of area a , and

$$\inf \{ |\bar{R}(\Omega)|; \text{Area}(\Omega) = a \text{ and } \Omega \text{ is convex} \} = 2\sqrt{\pi a}.$$

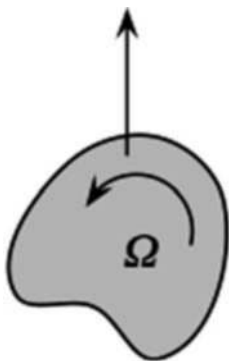


Fig. 19.8 Slowly rotating body

(b) Taking account of theorem 19.2, the problem is solved in two steps:

(i) find the measure η_* solving the problem

$$\inf_{\eta \in \mathcal{M}} \mathcal{F}(\eta), \text{ where } \mathcal{F}(\eta) = \iint_{\square} (1 + \cos(\varphi - \varphi^+)) d\eta(\varphi, \varphi^+) \quad (19.2)$$

and

(ii) find the sequence of sets Ω_n such that $|\partial(\text{conv} \Omega_n)| \rightarrow 2\sqrt{\pi a}$ and η_{Ω_n} weakly converges to η_* as $n \rightarrow \infty$.

The problem (19.2) is essentially the Monge–Kantorovich problem of optimal mass transportation. Its solution is given in [10]. The optimal measure η_* is supported on the union of five segments: (1) $\varphi + \varphi^+ = 0$, $-a_0 \leq \varphi \leq a_0$; (2) $3\varphi - \varphi^+ = \pi$, $\pi/6 \leq \varphi \leq a_0$; (3) $\varphi - 3\varphi^+ = \pi$, $\pi - 3a_0 \leq \varphi \leq \pi/2$; (4) $3\varphi - \varphi^+ = -\pi$, $-a_0 \leq \varphi \leq -\pi/6$; (5) $\varphi - 3\varphi^+ = -\pi$, $-\pi/2 \leq \varphi \leq -\pi + 3a_0$; and two

curves (6) $\sin \varphi - \sin \varphi^+ = \sin a_0 + \sin 3a_0$, $a_0 \leq \varphi \leq \pi - 3a_0$; (7) $\sin \varphi^+ - \sin \varphi = \sin a_0 + \sin 3a_0$, $-\pi + 3a_0 \leq \varphi \leq -a_0$, where $a_0 = 0.554$; in fig. 19.9, the support of η_* is shown in boldface. Note that the support of η_* is not the graph of a function; each value of φ that lies inside one of the two thin vertical strips bounded by dotted lines in the figure corresponds to two points in the diagram.

One has $\inf_{\eta \in \mathcal{M}} \mathcal{F}(\eta) = \mathcal{F}(\eta_*) = 0.9878\dots$, therefore the infimum in the non-convex minimization problem equals

$$\inf_{\text{Area}(\Omega)=a} |\bar{R}(\Omega)| = 0.9878\dots \cdot 2\sqrt{\pi a}.$$

Any minimizing sequence of sets approaches a circle of radius a , with the boundary of the sets becoming more and more complicated. In the spirit of the definition of rough body given above, one can associate the solution with a body \mathcal{B} obtained by roughening a circle of area a , so that any minimizing sequence represents this body. The law of scattering on this body is $v_{\mathcal{B}} = 2\sqrt{\pi a} \eta_* \otimes l$, where l is Lebesgue measure on S^1 .

In general, the mean resistance of a rough body \mathcal{B} is given by

$$\bar{R}(\mathcal{B}) = -\frac{3}{4} \iiint_{\square \times S^1} (\sin(\varphi - \varphi^+), 1 + \cos(\varphi - \varphi^+)) dv_{\mathcal{B}}(\varphi, \varphi^+, n).$$

Repeating the argument from the previous subsection, one comes to the formula

$$|\bar{R}(\mathcal{B})| = \frac{3}{4} \iiint_{\square \times S^1} (1 + \cos(\varphi - \varphi^+)) dv_{\mathcal{B}}(\varphi, \varphi^+, n). \quad (19.3)$$

19.3.3 Mathematical Retroreflector

Retroreflector is an optical device sending incident beams of light back to the origin. The most widely used retroreflectors are the so-called *cube corner* and *cat's eye*.

Denote the measure $\eta_* \in \mathcal{M}$ by $d\eta_* := \frac{1}{2} \cos \varphi \cdot \delta(\varphi - \varphi^+) d\varphi d\varphi^+$ and give the following.

Definition 19.4. Mathematical retroreflector is a rough body \mathcal{B}_* (body obtained by roughening a convex set B) such that $v_{\mathcal{B}_*} = \eta_* \otimes \tau_B$.

Note that the retroreflector \mathcal{B}_* is, at the same time, the body of **maximum resistance**. Indeed, using (19.3), for any body \mathcal{B} obtained by roughening B one has

$$|\bar{R}(\mathcal{B})| \leq \frac{3}{4} \iiint_{\square \times S^1} 2 \cdot dv_{\mathcal{B}}(\varphi, \varphi^+, n) = \frac{3}{2} v_{\mathcal{B}}(\square \times S^1) = \frac{3}{2} |\partial B|.$$

On the other hand,

$$|\bar{R}(\mathcal{B}_*)| = \frac{3}{4} \iiint_{\square \times S^1} (1 + \cos(\varphi - \varphi^+)) d\eta_*(\varphi, \varphi^+) d\tau_B(n) =$$

$$= \frac{3}{4} \int_{-\pi/2}^{\pi/2} 2 \cdot \frac{1}{2} \cos \varphi \, d\varphi \cdot \int_{S^1} d\tau_B(n) = \frac{3}{2} |\partial B|.$$

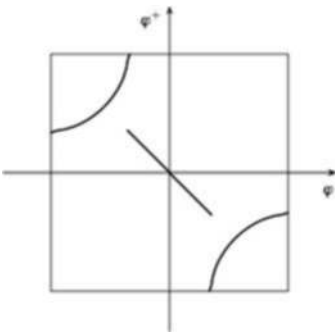


Fig. 19.9 Support of η_*

19.3.4 Effect of Magnus

This is a joint work with Tatiana Tchemisova, University of Aveiro [11].

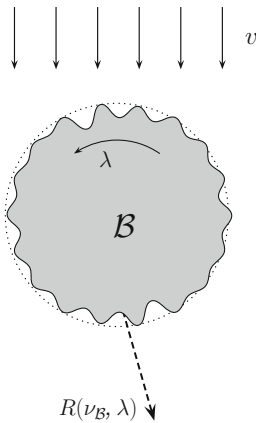


Fig. 19.10 Rapidly rotating rough circle

A rough unit circle \mathcal{B} rapidly rotates and at the same time moves through a rarefied gas (see Fig. 19.10). The velocity of translational motion is $-v$, $|v| = 1$. The force of gas resistance $R(\nu_{\mathcal{B}}, \lambda)$ acting on the body depends on the following parameters:

- $v_{\mathcal{B}}$, the law of scattering on \mathcal{B} ;
- λ , angular velocity of the body.

Problem 19.2 For each $\lambda \geq 0$ determine the set of all possible forces $\{R(v_{\mathcal{B}}, \lambda); \mathcal{B} \text{ is obtained by roughening the unit circle } B\}$.

For $\lambda = 0$, one takes by definition $R(v_{\mathcal{B}}, 0) := \lim_{\lambda \rightarrow 0^+} R(v_{\mathcal{B}}, \lambda)$.

Using theorem 19.1, one reformulates this problem in the following form:

Problem 19.3 For each $\lambda \geq 0$ determine the set of all possible values $\{R(v, \lambda); v \in \mathcal{M}_B\}$.

One has

$$R(v, \lambda) = - \int \int \int_{\square \times S^1} c_{\lambda}(x, y) dv(x, y, n),$$

where

(i) for $\lambda = 1$,

$$c_1(x, y) = 3 \sin^2 x \left(\frac{\cos(2x - y) + \cos x}{\sin(2x - y) + \sin x} \right) \cdot \chi_{(0, +\infty)}(x);$$

(ii) for $0 < \lambda < 1$,

$$c_{\lambda}(x, y) = \frac{3(\lambda \sin x + \sin \zeta(x))^3}{4 \sin \zeta(x)} \cos \frac{x - y}{2} \left(\frac{\cos \left(\zeta(x) + \frac{x - y}{2} \right)}{\sin \left(\zeta(x) + \frac{x - y}{2} \right)} \right)$$

where $\zeta(x) = \arccos(\lambda \cos x)$; and

(iii) for $\lambda > 1$,

$$c_{\lambda}(x, y) = \frac{3 \cos \frac{x - y}{2}}{2 \sin \zeta(x)} \left\{ (\lambda^3 \sin^3 x + 3 \lambda \sin x \sin^2 \zeta(x)) \cos \zeta(x) \left[\frac{\cos \frac{x - y}{2}}{\sin \frac{x - y}{2}} \right] + \right. \\ \left. + (3 \lambda^2 \sin^2 x \sin \zeta(x) + \sin^3 \zeta(x)) \sin \zeta(x) \left[\frac{-\sin \frac{x - y}{2}}{\cos \frac{x - y}{2}} \right] \right\} \cdot \chi_{\arccos(1/\lambda), +\infty}(x).$$

The detailed derivation of these formulas is given in the forthcoming paper [11].

Problem (19.3) is in fact a vector-valued problem of optimal mass transportation. It was solved numerically for several values of λ . The set $\{R(v_{\mathcal{B}}, \lambda)\}$, for $\lambda = 0, 0.1, 0.3$ and 1 , is shown in Figs. 19.11, 19.12, 19.13, and 19.14.

The points A, B, C , and D in Figs. 19.12, 19.13, and 19.14 correspond to the particular choices of $v = \eta \otimes l$, where $\eta \in \mathcal{M}$ and l is the Lebesgue measure on S^1 :

- A corresponds to η_{\star} (retroreflector measure);
- B corresponds to η_{∇} (measure generated by hollows having the shape of a right isosceles triangle);
- C corresponds to η_{mix} , where $d\eta_{\text{mix}} = \frac{1}{4} \cos x \cos y dx dy$;
- D corresponds to η_0 (convex body).

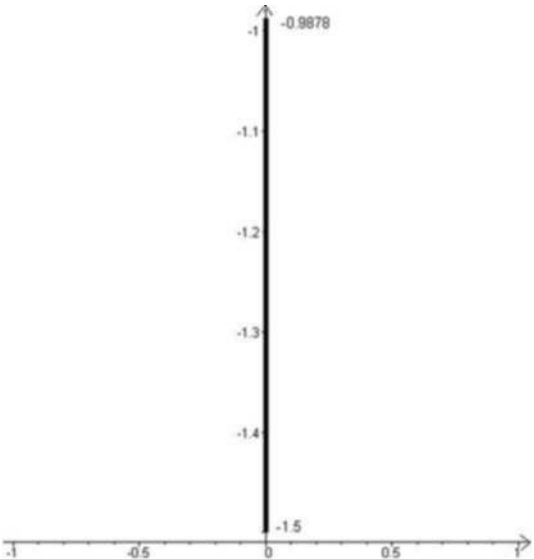


Fig. 19.11 The set of possible resistance forces $\{R(v_{\mathcal{B}}, \lambda)\}$ for $\lambda = 0$

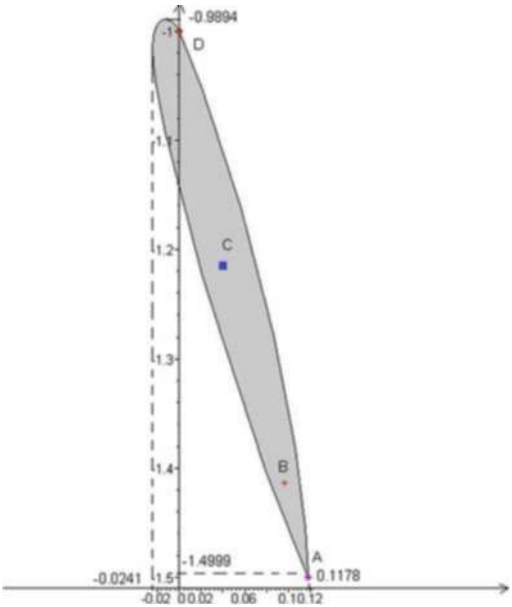


Fig. 19.12 The set of possible forces for $\lambda = 0.1$

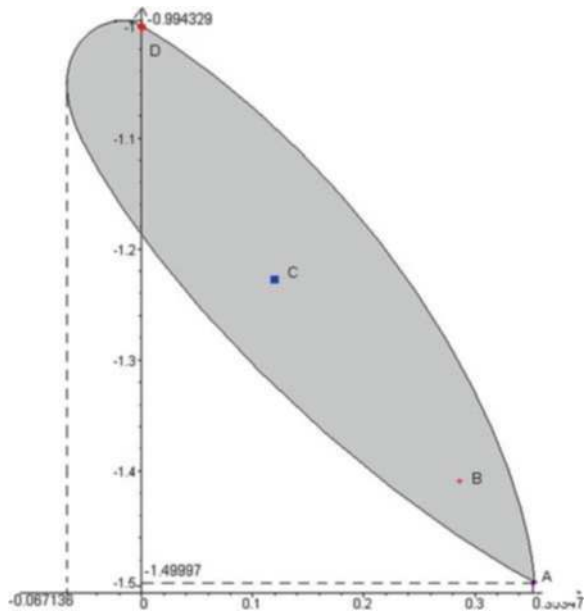


Fig. 19.13 The set of possible forces for $\lambda = 0.3$

The measures η_* , η_∇ , and η_0 are defined in Sect. 19.3.3 and 19.3.1.

If the horizontal component of resistance force is nonzero, $R_1(v_{\mathcal{B}}, \lambda) \neq 0$, then one says that the *Magnus effect* takes place. If, in particular, $R_1(v_{\mathcal{B}}, \lambda) > 0$, the effect is sometimes called the *reverse Magnus effect*. From Figs. 19.11, 19.12, 19.13,

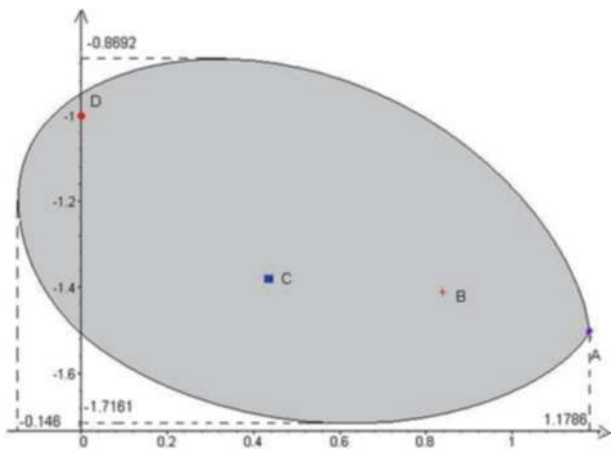


Fig. 19.14 The set of possible forces for $\lambda = 1$

and 19.14 one can see that the Magnus effect takes place for $\lambda > 0$ and observe that the reverse Magnus effect is much more typical than the “direct” one.

Note that, in our opinion, the inverse Magnus effect in rarefied gases is due to the two factors of different nature: (a) *nonelastic* interaction of gas molecules with the body surface; (b) *multiple reflections* caused by microscopical hollows on the body surface. The factor (a) was examined in [12, 14]. In these papers, the body surface was supposed to be convex, and thus, factor (b) was excluded from consideration.

On the contrary, here we concentrate on factor (b). In our model all reflections are perfectly elastic; therefore, factor (a) does not take place.

Acknowledgments This work was supported by *Centre for Research on Optimization and Control* (CEOC) from the “*Fundação para a Ciência e a Tecnologia*” (FCT), cofinanced by the European Community Fund FEDER/POCTI and by FCT (research project PTDC/MAT/72840/2006).

References

1. Plakhov, A.: Newton’s problem of the body of minimal resistance with a bounded number of collisions. *Russ. Math. Surv.* **58**, 191–192 (2003).
2. I. Newton, I.: *Philosophiae naturalis principia mathematica*. London: Streater (1687).
3. Buttazzo, G., Kawohl, B.: On Newton’s problem of minimal resistance. *Math. Intell.* **15**, No.4, 7–12 (1993).
4. Buttazzo, G., Ferone, V., Kawohl, B.: Minimum problems over sets of concave functions and related questions. *Math. Nachr.* **173**, 71–89 (1995).
5. Brock, F., Ferone, V., Kawohl, B.: A symmetry problem in the calculus of variations. *Calc. Var.* **4**, 593–599 (1996).
6. Lachand-Robert, T., Oudet, E.: Minimizing within convex bodies using a convex hull method. *SIAM J. Optimiz.* **16**, 368–379 (2006).
7. Lachand-Robert, T., Peletier, M. A.: Newton’s problem of the body of minimal resistance in the class of convex developable functions. *Math. Nachr.* **226**, 153–176 (2001).
8. Plakhov, A.: Billiard scattering on rough sets: Two-dimensional case. *SIAM J. Math. Anal.* **40**, 2155–2178 (2009).
9. Plakhov, A.: Billiards and two-dimensional problems of optimal resistance. *Arch. Rotational Mech. Anal.*, DOI:10.1007/s00205-008-0137-1(2008).
10. Plakhov, A.: Newton’s problem of the body of minimum mean resistance. *Sbornik: Math.* **195**, 1017–1037 (2004).
11. Plakhov, A., Tchemisova, T.: Force acting on a rough disk spinning in a flow of noninteracting particles. *Doklady Math.* **79**, 132–135 (2009).
12. Borg, K. I., Söderholm, L. H., Essén, H.: Force on a spinning sphere moving in a rarefied gas. *Phys. Fluids* **15**, 736–741 (2003).
13. Plakhov, A.: Newton’s problem of a body of minimal aerodynamic resistance. *Doklady Math.* **67**, 362–365 (2003).
14. Weidman, P. D., Herczynski, A.: On the inverse Magnus effect in free molecular flow. *Phys. Fluids* **16**, L9–L12 (2004).

“This page left intentionally blank.”

Chapter 20

Shock Optimization for Airfoil Design Problems

Olivier Pironneau

Abstract We begin by recalling the classical approach to solve shape design problems by gradient methods. Then we proceed to explain how automatic differentiation can simplify the analysis and illustrate this approach to a shape design problem where the shock is required to be at a certain place. The object of this chapter is to show that automatic differentiation seems to work even though the gradients and the calculus of variation have to be extended by Distribution theory.

20.1 Numerical Optimal Shape Design

20.1.1 An Academic Problem

Consider the problem of finding the shape of a wind tunnel Ω with uniform flow u_d in a region D (see Fig. 20.1). For simplicity let us assume two-dimensional-symmetry and irrotational inviscid flow computed by a stream function ψ . The problem then is

$$\min_{S \in \mathcal{S}_d} \{J(S) := \int_D |\psi - \psi_d|^2 : -\Delta \psi = 0, \text{ in } \Omega \quad \psi|_S = 0 \quad \psi|_{\Gamma-S} = \psi_d\} \quad (20.1)$$

where S is the control boundary, \mathcal{S}_d the set of admissible shapes, $\Gamma := \partial\Omega$, ψ_d contains the inflow and outflow conditions and is also such that $\nabla \times \psi_d|_D = u_d$.

Olivier Pironneau
LJLL, University of Paris VI & IUF, Paris, France,
e-mail: pironneau@ann.jussieu.fr

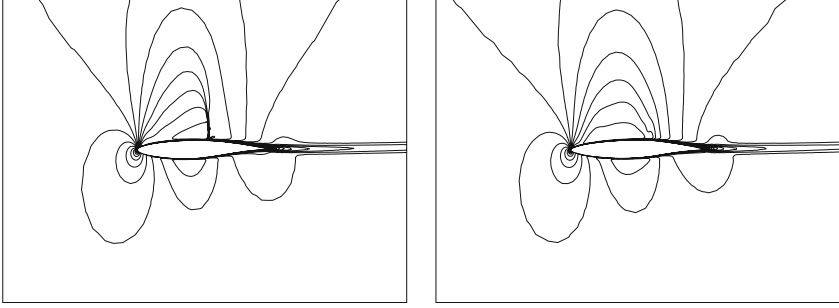


Fig. 20.1 Optimization of a wing profile at Mach 0.8 to minimize the drag. Since most of the drag is the pressure drag, the optimized profile is shock free (see [2])

20.1.2 Sensitivity Analysis

To study the effect on J of a change of shape S , it is convenient to consider normal variations of size proportional to $\alpha(\cdot)$ about S , or more generally about Γ :

$$\Gamma^\varepsilon := \{x + \varepsilon\alpha(x)n(x) : x \in \Gamma\}$$

with the convention that α is nonzero only on S and $n(x)$ is the normal to S at $x \in S$. Provided that S has no corner, all smooth variations of S are in that class and we can proceed to study the limit when $\varepsilon \rightarrow 0$ of $(J(S^\varepsilon) - J(S))/\varepsilon$. We begin with ψ . Let us translate the boundary conditions into the right-hand side f :

$$-\Delta \psi^\varepsilon = f \quad \text{in } \Omega^\varepsilon \quad \psi^\varepsilon = 0 \quad \text{on } \Gamma^\varepsilon := \{x + \varepsilon\alpha n : x \in \Gamma\} \quad (20.2)$$

If $\psi'_\alpha := \lim_{\varepsilon} \frac{1}{\varepsilon}(\psi^\varepsilon - \psi)$ exists then ψ is Gateau differentiable with respect to Γ in the direction α . If ψ'_α is linear in α then ψ is Frechet differentiable, and so on with higher derivatives. Assume that we can write

$$\psi^{\varepsilon\alpha} = \psi + \varepsilon\psi'_\alpha + \frac{\varepsilon^2}{2}\psi''_\alpha \quad (20.3)$$

Then, by linearity of the Laplace operator ψ' and ψ'' satisfy the same PDE but with $f = 0$.

Also by Taylor expansion, $x \in \Gamma$:

$$0 = \psi^{\varepsilon\alpha}(x + \varepsilon\alpha n) = \psi^{\varepsilon\alpha}(x) + \varepsilon\alpha \frac{\partial \psi^{\varepsilon\alpha}}{\partial n}(x) + \frac{\varepsilon^2\alpha^2}{2} \frac{\partial^2 \psi}{\partial n^2}(x) + \dots \quad (20.4)$$

Therefore,

$$-\Delta \psi'_\alpha = 0 \quad \psi'_\alpha|_\Gamma = -\alpha \frac{\partial \psi}{\partial n}, \quad -\Delta \psi''_\alpha = 0 \quad \psi''_\alpha|_\Gamma = -\alpha \frac{\partial \psi'_\alpha}{\partial n} - \frac{\alpha^2}{2} \frac{\partial^2 \psi}{\partial n^2} \quad (20.5)$$

Let us assume again that there is enough regularity to expand J in powers of ε :

$$J(S^\varepsilon) = \int_D |\psi^\varepsilon - \psi_d|^2 = \int_D |\psi - \psi_d|^2 + 2\varepsilon \int_D (\psi^\varepsilon - \psi_d) \psi'_\alpha + o(\varepsilon) \quad (20.6)$$

If J is Frechet differentiable there exists ξ such that $J'_\alpha = \int_S \xi \alpha$. To find ξ we must use the *adjoint trick* and introduce

$$-\Delta p = (\psi^\varepsilon - \psi_d) I_D, \quad p|_\Gamma = 0 \quad (20.7)$$

Then

$$2 \int_D (\psi^\varepsilon - \psi_d) \psi'_\alpha = -2 \int_\Omega \psi'_\alpha \Delta p = -2 \int_\Omega \Delta \psi'_\alpha p + \int_\Gamma \left(\frac{\partial p}{\partial n} \psi'_\alpha + \frac{\partial \psi'_\alpha}{\partial n} p \right) \quad (20.8)$$

So we have “proved” that J is Gateau differentiable, in fact:

$$J'_\alpha = 2 \int_S \frac{\partial p}{\partial n} \frac{\partial \psi}{\partial n} \alpha$$

and that $2 \frac{\partial p}{\partial n} \frac{\partial \psi}{\partial n}$ plays the role of the derivative of J with respect to S in the direction α .

20.1.3 Conceptual Algorithm

Let A be a smoother on Γ (for example, the Laplace–Beltrami operator). The following is a conceptual gradient method with fixed step size ρ :

- (1) Choose a shape S^0 , a small number $\rho > 0$ and set $m = 0$.
- (2) Compute ψ^m and p^m by solving

$$\begin{aligned} -\Delta \psi^m &= 0, \quad \psi^m|_{S^m} = 0, \quad \psi^m|_{\Gamma_d} = \psi_d \\ -\Delta p^m &= (\psi^m - \psi_d) I_D, \quad p|_{\Gamma^m} = 0 \end{aligned} \quad (20.9)$$

- (3) Set

$$\alpha = -\rho A \left(2 \frac{\partial p^m}{\partial n} \frac{\partial \psi^m}{\partial n} \right), \quad S^{m+1} = \{x + \alpha n : x \in S^m\} \quad (20.10)$$

- (4) Set $m \leftarrow m + 1$ and go to 1.

Every iteration is expected to decrease J because $\xi = 2 \frac{\partial p^m}{\partial n} \frac{\partial \psi^m}{\partial n}$ and

$$J(S^{m+1}) = J(S^m) + \int_{S^m} \xi \alpha = J(S^m) - \rho a \left(\frac{\partial p^m}{\partial n} \frac{\partial \psi^m}{\partial n}, \frac{\partial p^m}{\partial n} \frac{\partial \psi^m}{\partial n} \right) + o(\alpha)$$

and $a(z, z) > 0$ by hypothesis.

20.2 Automatic Differentiation

In practice for aerospace the state equations are very complex. Compressible turbulent flows may be modelled by (with standard notations)

$$\begin{aligned}
 \partial_t \rho + \nabla \cdot (\rho u) &= 0 \\
 \partial_t (\rho u) + \nabla \cdot (\rho u \otimes u) + \nabla \left(p + \frac{2}{3} \rho k \right) &= \nabla \cdot ((\mu + \mu_t) S) \\
 \partial_t (\rho E) + \nabla \cdot \left(\left(\rho E + p + \frac{5}{3} \rho k \right) u \right) &= \nabla \cdot ((\mu + \mu_t) S u) + \nabla \cdot ((\chi + \chi_t) \nabla T) \\
 \partial_t \rho k + \nabla \cdot (\rho u k) - \nabla \cdot ((\mu + \mu_t) \nabla k) &= S_k \\
 \partial_t \rho \varepsilon + \nabla \cdot (\rho u \varepsilon) - \nabla \cdot ((\mu + c_\varepsilon \mu_t) \nabla \varepsilon) &= S_\varepsilon.
 \end{aligned} \tag{20.11}$$

As it is painfully difficult to derive the sensitivities analytically for such a system we turn to automatic differentiation of computer programs and proceed to explain its principle.

20.2.1 Principle of Automatic Differentiation

Let $J(u) = |u - u_d|^2$, then its differential is

$$\delta J = 2(u - u_d)(\delta u - \delta u_d) \quad \frac{\partial J}{\partial u} = 2(u - u_d)(1.0 - 0.0) \tag{20.12}$$

Obviously the derivative of J with respect to u is obtained by putting $\delta u = 1$, $\delta u_d = 0$. Now suppose that J is programmed in C/C++ by

```
double J(double u, double u_d){
    double z = u-u_d;  z = z*(u-u_d);
    return z;
}
int main(){ double u=2,u_d = 0.1;
    cout << J(u,u_d) << endl;
}
```

Except for the embarrassing problem of returning both z , dz instead of z , a program which computes J and its differential can be obtained by writing above each line its differential:

```
double JandDJ(double u, double u_d, double du, double du_d,
               double *pdz)
{
    double dz = du - du_d, z = u-u_d;
    double dz = dz*(u-u_d) + z*(du - du_d);
    z = z*(u-u_d); *pdz = dz;
    return z;
}
int main()
{
    double u=2,u_d = 0.1;
    double dJ;
    cout << J(u,u_d,1,0,&dJ) << endl;
}
```


In C++ the program can be simplified further by encapsulating each variable and its derivative into one member of the class `ddouble` so defined:

```
class ddouble{
public: double val[2];
    ddouble(double a, double b=0){ v[0] = a; v[1]=b;}
    ddouble operator=(const ddouble& a)
    { val[1] = a.val[1]; val[0]=a.val[0]; return *this; }
};
```

Here `val[0]` holds the value of the variable and `val[1]` the value of the differential. Furthermore, the compiler itself can differentiate each line of the source code if we define the operators `+`, `-`, `*`, `/`, `sin`, `log` ... for members of the class `ddouble`. For instance

```
ddouble ddouble::operator-(const ddouble& a,const ddouble& b)
{
    ddouble c;
    c.v[1] = a.v[1] - b.v[1];    // (a-b)'=a'-b'
    c.v[0] = a.v[0] - b.v[0];    // c=a-b
    return c;
}
ddouble ddouble::operator*(const ddouble& a,const ddouble& b)
{
    ddouble c;
    c.v[1] = a.v[1]*b.v[0] + a.v[0]* b.v[1]; // (a*b)'=a'*b+a*b'
    c.v[0] = a.v[0] * b.v[0];    // c=a*b
    return c;
}
```

In the end, with a couple of pages of C++ code for the definition of the class `ddouble`, one can transform a C-program into one which computes all the sensitivities by simply changing with a text editor the keyword `double` into `ddouble`. There are limitations, however; the main one is that the adjoint trick is not used, so the method will not be computationally efficient if there are more than, say 50 parameters. When this number is exceeded one must use the *reverse mode*: see Griewank[1] or the web site of Tapenade:

<http://tapenade.inria.fr:8080>.

20.2.2 Example of Application

Using the reverse mode of Odyssee (now Tapenade), in 1998, B. Mohammadi [2] has minimized the drag of a wing profile at transonic speed for a flow governed by (20.11) plus wall laws. The criteria are composed of the drag plus penalization to keep the lift and area constant

$$J(u, p, \theta) = F \cdot u_\infty + \frac{1}{\varepsilon} |F \times u_\infty - C_l|^2 + \frac{1}{\beta} \left(\int_S dx - a \right)^2 \quad (20.13)$$

with $F = \int_S (p \mathbf{n} + (\mu \nabla u + \nabla u^T))$.

20.3 Differentiability Issues

Mohammed Hafez [3] raised an interesting question about the validity of calculus of variations in the presence of Shocks. The claim was that by considering only smooth variations δu of u one could not move the shock and so the variational set is not big enough.

To answer this question Bardos and the author [4] showed that calculus of variations can be extended to variations with Dirac masses and most of the rules are valid on the condition that whenever a product appears such as uv then its differential be written as

$$\delta(uv) = \bar{v}\delta u + \bar{u}\delta v \quad (20.14)$$

where $\bar{u} = \frac{u^+ + u^-}{2}$ if u is discontinuous.

20.3.1 Extended Calculus of Variation

Let S be the shock, function of time t and and a parameter a . Let C_S^1 be the space of $\{a, x\}$ - C^1 functions outside S . Then the a -derivative of $v \in C_S^1$ in the sense of distribution is

$$v'_a = \partial_a v - [v]x' \cdot n \delta_S \quad (20.15)$$

where $\partial_a v$ is the pointwise a -derivative and n the normal to the shock.

This result is easy to understand on a one-dimensional example. Let $H(x)$ be the Heavyside function ($H = 1$ if $x > 0$, 0 otherwise). Let v jump across $x(a)$ from v^- to v^+ . Then

$$\begin{aligned} v(x) &= v^-(x, a) + (v^+(x, a) - v^-(x, a))H(x - x(a)) \\ \Rightarrow v' &= v'^- + [v'_a]H(x - x(a)) - [v]x'(a)\delta(x - x(a)) \end{aligned} \quad (20.16)$$

Note that it can be shown that for all smooth w

$$\int_{\mathbb{R}} v_{a+\delta a} w = \int_{\mathbb{R}} (v_a + v'_a \delta a) w + o(|a|).$$

One can also show that (20.14) holds for functions of C_S^1 . Another important numerical observation from (20.15) is the following.

Proposition 20.1. *The shock displacement $d = x'_a \cdot n$ can be computed from the Dirac weight w of v'_a and the jump $[v]$ at the shock by*

$$d = -\frac{w}{[v]} \quad (20.17)$$

20.3.2 Sensitivity Analysis for Burgers' Equation

Consider

$$\partial_t u + \partial_x \frac{u^2}{2} = 0 \text{ in } Q := \mathbb{R} \times (0, +\infty), \quad u(x, 0) = u^0(x, a) \quad (20.18)$$

With compatible (entropy) u^0 , a discontinuity at $x = x(t, a)$, propagates at speed $\dot{x} := \partial_t x = \bar{u} := \frac{u^+ + u^-}{2}$. A sensitivity analysis with respect to a must differentiate both (20.18) and the above Rankine–Hugoniot condition. Thanks to this extended calculus of variation it is implicit in the formal calculus. To show this the basic idea is to notice that in the (x, t) space (20.18) is

$$\nabla \cdot \mathbf{v} = 0 \quad \text{with} \quad \mathbf{v} = \begin{pmatrix} u \\ \frac{u^2}{2} \end{pmatrix} \quad (20.19)$$

and to recall that $\nabla \cdot \mathbf{v} = \nabla_x \cdot \mathbf{v} + [\mathbf{v}]_S \cdot n \delta_S$, $\mathbf{v}'_a = \partial_a \mathbf{v} - [\mathbf{v}]_S x'_a \cdot n \delta_S$. This leads to the following.

Proposition 20.2. *The a -derivative of u satisfies the linearized Burgers equation*

$$\begin{aligned} \partial_t u'_a + \partial_x (\bar{u} u'_a) &= 0 \text{ in the sense of distribution i.e. for all smooth } w : \\ \int_Q w (\partial_t u'_a + \partial_x (u u'_a)) + \delta_S (-[u] \bar{u}'_a + d_t([u] x'_a)) dx dt &= 0, \quad u'_a(x, 0) = u^{0'}_a \end{aligned} \quad (20.20)$$

Numerical Consequences

A space–time approximation is most likely needed at the discrete level!

20.3.3 Application to Optimal Control

Consider the problem

$$\min_a \left\{ J := \frac{1}{2} \int_R |u(T) - u_d|^2 : \partial_t u + \partial_x \frac{u^2}{2} = 0 \quad u|_0 = u^0(a) \right\}$$

The extended calculus of variation is

$$\delta J = \int_R \overline{(u - u_d)}|_T \delta u(T) \quad \text{with} \quad \partial_t \delta u + \partial_x (\bar{u} \delta u) = 0, \quad \delta u(0) = \frac{du^0}{da} \delta a \quad (20.21)$$

Define the adjoint by

$$\partial_t p + \bar{u} \partial_x p = 0, \quad p(T) = \overline{(u - u_d)}|_T \quad (20.22)$$

and use an integration by parts:

$$0 = \int_{R \times (0, T)} u'_a (\partial_t p + \bar{u} \partial_x p) = \int_R p u'_a|_0^T \Rightarrow J'_a = \int_R p u'_a|_T = \int_R p|_0 u^{0'}_a \quad (20.23)$$

Notice that because of the bar in \bar{u} the adjoint contains implicitly

$$\frac{d}{dt} p(x(t), t) = 0, \quad p(x(T), T) = \overline{(u - u_d)}|_T \quad (20.24)$$

which defines it between the diverging characteristics issued from the shock at $x(T), T$. Notice also that p is continuous at the shock, but discontinuous on the characteristics from $x(T), T$.

20.3.4 A Simple Example

Consider

$$\partial_t u + \partial_x \left(\frac{u^2}{2} \right) = 0 \quad u(x, 0) = (1+a)(1-H(x)) \quad (20.25)$$

The analytical solution and its derivative with respect to a are

$$u = (1+a) \left(1 - H \left(x - \frac{1+a}{2} t \right) \right), \quad u'_a = 1 - H \left(x - \frac{1+a}{2} t \right) + \frac{1+a}{2} t \delta \left(x - \frac{1+a}{2} t \right)$$

Let

$$J(a) := \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} u(T)^2 = \frac{1}{4} (1+a)^2 ((1+a)T - 1) \Rightarrow J'(0) = \frac{3}{4} T - \frac{1}{2}$$

$$\text{By the extended theory } J' = \int_{-\frac{1}{2}}^{\frac{1}{2}} \bar{u} u'_a|_T = \int_{-\frac{1}{2}}^{\frac{1}{2}} \bar{p} u'_a|_0 = \int_{-\frac{1}{2}}^0 p(x, 0)$$

$$\partial_t p + \bar{u} \partial_x p = 0, \quad p|_T = \bar{u}|_T \Rightarrow p|_0 = 1 - \frac{1}{2} H \left(x + \frac{T}{2} \right) - \frac{1}{2} H \left(x - \frac{T}{2} \right) \quad (20.26)$$

20.3.5 Right and Wrong Schemes

Consider

$$J = \frac{1}{2} \int_{\mathbb{R}} |u|_T - 1|^2 dx : \partial_t u + \partial_x \frac{u^2}{2} = 0, \quad u|_0 = \text{atan}(10x + a)^- \quad (20.27)$$

and the following conservative finite difference scheme:

$$\frac{u_i^{m+1} - u_i^m}{\delta t} + u_i^+ \frac{u_i^m - u_{i-1}^m}{\delta x} - u_i^- \frac{u_{i+1}^m + u_i^m}{\delta x} = 0$$

$$\text{with } u_i^+ = \max \left\{ 0, \frac{u_i^m + u_{i-1}^m}{2} \right\}, \quad u_i^- = -\min \left\{ 0, \frac{u_i^m + u_{i-1}^m}{2} \right\} \quad (20.28)$$

Now consider the nonconservative characteristic finite difference scheme:

$$u_i^{m+1} = u_{j_i}^m(1 - \alpha) + u_{j_i+1}^m \alpha, \quad j_i = \text{int} \frac{x - u_i \delta t}{\delta x}, \quad \alpha = \frac{x - u_i \delta t}{\delta x} - j_i \quad (20.29)$$

and apply automatic differentiation (AD) to both of them. The results of Table 20.1 show that both schemes with AD compute correctly J' . However when u'_a is plotted we see that the conservative scheme gives the correct result while the characteristic scheme misses the Dirac masses (see Fig. 20.2).

Table 20.1 Results using schemes (20.28) and (20.29). Both schemes give the correct result J' . For comparison the derivative is also computed by finite difference on a

$J(0)$	$J(\delta a)$	$J'_a(0)$	$\frac{J(\delta a) - J(0)}{\delta a}$
0.0338517	0.033851	-0.00720163	-0.00720163
0.0338517	0.033851	-0.00719884	-0.00720173

Discrete Adjoint

The discrete adjoint for the conservative finite difference scheme is

$$\frac{p_i^{m-1} - p_i^m}{\delta t} + \frac{u_i^m}{\delta x} (p_i^m s_i - p_{i-1}^m s_{i-1}^- + p_{i+1}^m s_{i+1}^+) = 0 \quad (20.30)$$

with $s_i = \text{sign}(u_i^m + u_{i-1}^m)$

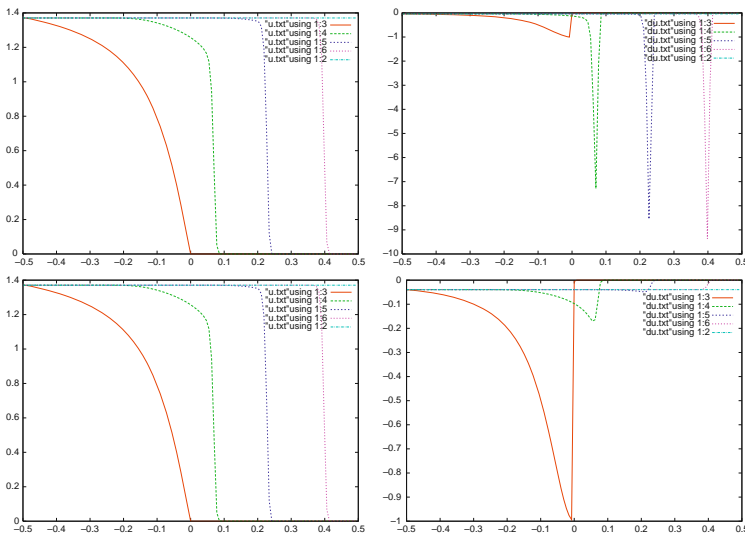


Fig. 20.2 Top displays u solution of Burgers' equation, while bottom displays u'_a at $T = 0.75$. The top two figures and the bottom-left one are correct but the bottom-right one is wrong

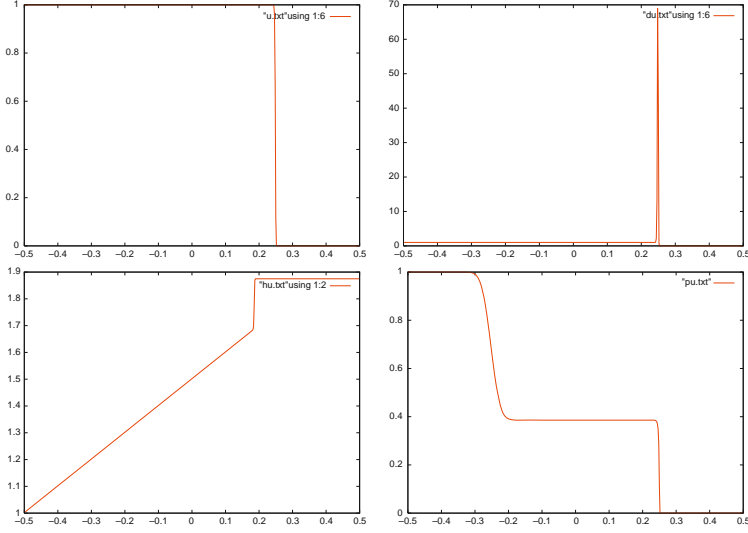


Fig. 20.3 Everything is correctly computed: $u, u'_a, \int^x u'_a dx, p$ but the weight of the Dirac mass in u'_a , which predicts the displacement of the shock position (see Proposition 20.1), can be apprehended only via the jump in $\int^x u'_a dx$

On the simple example again the adjoint p seems to be correctly calculated by the conservative scheme (see Fig. 20.3).

20.4 Small Disturbances and Automatic Differentiations

So far to our knowledge no one has been able to use the linearized fluid equations to dynamically compute the effect of small disturbances on a compressible flow at transonic speed when these induce a shock motion. This is because the linearized flow equations do not provide a mechanism for the motion of the shocks in a standard way.

Assuming that automatic differentiation gives the right results, it is then theoretically possible to compute the motion of shocks by Proposition 20.1. For instance consider a k-epsilon turbulent compressible flow around a NACA airfoil at Mach 0.8 at incidence 1° . Can we predict the change in the shock position when the Mach number is 0.85 by using the $|u|$ or rather the Mach m , its derivative m'_a and Proposition 20.1.

Figure 20.4 shows also m'_a and $\partial_x \mu$ where μ is solution of

$$0.01\mu - \frac{d^2\mu}{dx^2} = -\partial_x m'_a \quad \frac{\partial\mu}{\partial n}|_r = 0$$

This is a trick to integrate μ'_a in x because $\partial_x \mu = m'_a$; and so and by measuring the size of the jump of $\partial_x \mu$ at the shock we obtain the weight of the Dirac singularity

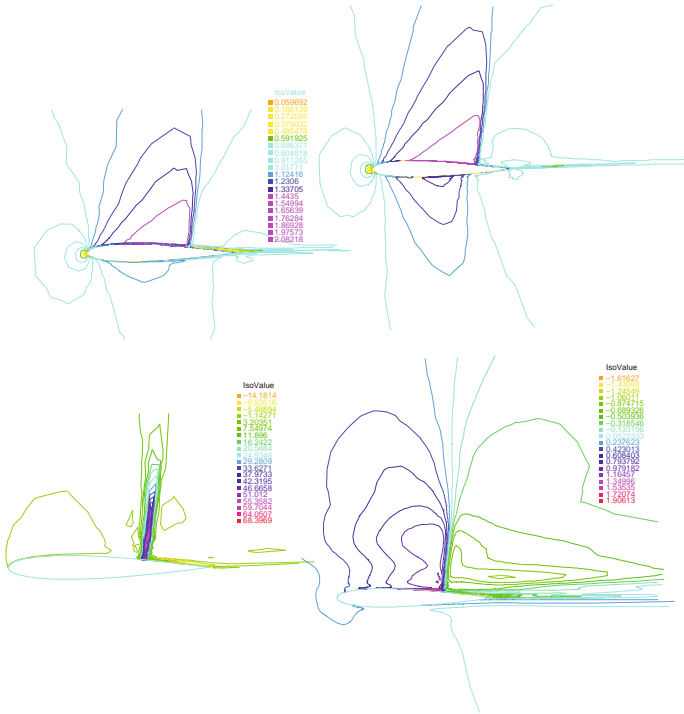


Fig. 20.4 Top: Mach lines around a NACA012 airfoil at $Mach_\infty = 0.8$ (left) and 0.85 (right) with one degree of incidence. Bottom: derivative of Mach number field m (not really useful) and $-\int^x m$ (right); the jump across the shock can be used to compute the displacement of the shock when $Mach_\infty$ changes from 0.8 to 0.85

of m'_a . We assume that the turbulence model has no real effect in the shock layer and that it is almost inviscid there. Figure 20.4 shows the Mach lines. Accordingly the size of the jump of $\partial_x \mu$ is roughly equal to 1.5 ± 0.2 , the jump of m is 0.43 ± 0.1 and so by Proposition 20.1 the shock displacement due to a change in Mach number at infinity of 0.05 will be $0.05 \times 1.5/0.43 = 0.17 \pm 0.06$ to be compared to the actual number which is more like 0.12 ; so the extended linear theory has overestimated the displacement but it seems to be correct within the range of numerical errors.

References

1. A. Griewank: *Evaluating Derivatives, Principles and Techniques of Algorithmic Differentiation*. Vol 19, Frontiers in Applied Mathematics. SIAM, Philadelphia. (2000).
2. B. Mohammadi and O. Pironneau: *Applied Numerical Optimal Shape Design*. Oxford U. Press. (2001).
3. A. Caughey and M. Hafez (ed.): *Pros and Cons of Airfoil Optimization*, Frontier of Computational Fluid Dynamics, World Scientific, Singapore. (1998).
4. C. Bardos and O. Pironneau : Derivatives and Control in the Presence of Shocks. Computational Fluid Dynamics Journal, vol 12, no.1 (April 2003).

“This page left intentionally blank.”

Chapter 21

Differential Games Treated by a Gradient-Restoration Approach

Mauro Pontani

Abstract When two competitive actors are involved in a flight path optimization, the problem can be modelled as a zero-sum game. This chapter describes a general procedure to convert the two-sided optimization problem into an optimal control problem. In addition, a numerical approach to the solution is proposed. The method is based on the joint application of an evolutionary algorithm (for providing a starting guess) and of the sequential gradient-restoration algorithm (to achieve the final solution). The homicidal chauffeur game – a classical example of zero-sum game – and an orbital pursuit-evasion game are considered to describe the method and test its effectiveness.

21.1 Introduction

In general, a flight path optimization problem can be treated as a single-objective problem (the classical optimal control problem) or a two-sided optimization problem. A single-objective optimization problem arises when a single actor is involved in the problem, which consists in finding the optimal control law that minimizes the objective function for the problem under consideration. However, in many situations, as the optimal interception of evasive missiles or the identification of the optimal maneuvers of two aircraft involved in an aerial combat, the two-sided approach seems considerably more appropriate. In this new context, the path optimization can be modelled as a zero-sum differential game, where two competitive “players” have conflicting objectives. The definition “zero-sum game” is related to the fact that the objectives of the two players are conflicting and, as a result, their sum is zero. Zero-sum games, first introduced by Isaacs [1], are also referred to as “pursuit-evasion games,” as they involve a pursuer trying to catch an evader.

Mauro Pontani

Scuola di Ingegneria Aerospaziale, University of Rome “La Sapienza”, via Eudossiana 16, 00184 Rome, Italy, e-mail: mauro.pontani@uniroma1.it

Many numerical algorithms have been proposed and employed to solve optimal control problems. However, they cannot be directly used to solve two-sided problems, since they are devoted to single-objective optimization problems. A possible approach to the numerical solution of differential games is based on the transformation of the two-sided optimization problem into a single-objective one. This goal can be achieved by including the adjoint variables associated to one player as additional components of an “extended” state, which is introduced in the re-formulation of the problem. The inclusion of the Lagrange multipliers in the extended state has the unfavorable consequence that a starting guess for these nonintuitive variables is needed. This difficulty can be faced by employing a systematic approach, based on the use of genetic algorithms (GAs) to provide first attempt values for the adjoint variables. Then, the sequential gradient-restoration algorithm, the well-known indirect method introduced by Angelo Miele since the 1960s [2–4], can be applied for solving the transformed problem.

In this research the classical homicidal chauffeur game [1] and an orbital pursuit-evasion game are considered. The latter problem consists in the identification of the optimal trajectories of two spacecraft the first, namely the pursuer, tries to reach the evader as quickly as possible, whereas the second, namely the evader, tries to delay the capture. The joint application of the evolutionary preprocessing technique and of the sequential gradient-restoration algorithm appears as a suitable approach to the numerical solution of the problem and several test cases are considered to prove the effectiveness of the method.

21.2 Zero-Sum Differential Games

When two competitive actors (usually referred to as “players”) are involved in a flight path optimization, the problem can be modelled as a zero-sum (or pursuit-evasion) differential game. Each player is associated to a state vector (\mathbf{x}_P or \mathbf{x}_E) and drives a dynamic system through its own set of control variables (denoted with \mathbf{u}_P for the pursuer P and with \mathbf{u}_E for the evader E). If “ $'$ ” denotes the derivative with respect to the actual time θ , the dynamic system is described by an uncoupled pair of state equations, each related to a single player:

$$\mathbf{x}'_P = \mathbf{f}_P(\mathbf{x}_P, \mathbf{u}_P, \theta) \quad (n_P \text{ differential equations}) \quad (21.1)$$

$$\mathbf{x}'_E = \mathbf{f}_E(\mathbf{x}_E, \mathbf{u}_E, \theta) \quad (n_E \text{ differential equations}) \quad (21.2)$$

with $\theta_0 \leq \theta \leq \theta_f$. In general, some components of the states at θ_0 can be unspecified. Usually, several boundary conditions hold for the problem at hand. In the most general formulation, these conditions can be collected in ψ , which is a vector function of the starting and final values of the states and of the initial and terminal times:

$$\psi(\mathbf{x}_{P0}, \mathbf{x}_{E0}, \mathbf{x}_{Pf}, \mathbf{x}_{Ef}, \theta_0, \theta_f) = \mathbf{0} \quad (q \text{ boundary conditions}) \quad (21.3)$$

Without any loss of generality, a problem of a Mayer type is considered:

$$J = \phi(\mathbf{x}_{P0}, \mathbf{x}_{E0}, \mathbf{x}_{Pf}, \mathbf{x}_{Ef}, \theta_0, \theta_f) \quad (21.4)$$

i.e., the objective cost is given as a function of the initial and final values of the states and of the initial and terminal times. In zero-sum games the pursuer tries to maximize J , whereas the evader tries to minimize J . Moreover, a pair of optimal strategies correspond to a saddle-point (SP) equilibrium solution and generate a saddle-point trajectory in the state space. By definition [5], an open-loop representation of an optimal feedback strategy is the strategy along the optimal trajectory as a function of the initial states only. Two basic properties relate open-loop strategies and feedback strategies:

- (i) if one of the two players deviates from his optimal open-loop strategy, his outcome worsens;
- (ii) if both players employ their own optimal open-loop strategies, the time histories of the optimal open-loop and of the optimal feedback strategies are identical.

First of all, a Hamiltonian and a function of terminal conditions are introduced as

$$H \doteq \lambda_P^T \mathbf{f}_P + \lambda_E^T \mathbf{f}_E \quad (21.5)$$

$$\Phi \doteq \phi + \mathbf{v}^T \boldsymbol{\psi} \quad (21.6)$$

where λ_P , λ_E , and \mathbf{v} are the adjoint variables conjugate to the state equations (21.1) and (21.2) and to the boundary conditions (21.3), respectively. The necessary conditions for the existence of an open-loop representation of a saddle-point solution can be viewed as an extension of the analogous conditions arising in optimal control theory. The following adjoint equations hold for the Lagrange multipliers λ_P and λ_E :

$$\lambda'_P = - \left[\frac{\partial H}{\partial \mathbf{x}_P} \right]^T = - \left[\frac{\partial \mathbf{f}_P}{\partial \mathbf{x}_P} \right]^T \lambda_P \quad (21.7)$$

$$\lambda'_E = - \left[\frac{\partial H}{\partial \mathbf{x}_E} \right]^T = - \left[\frac{\partial \mathbf{f}_E}{\partial \mathbf{x}_E} \right]^T \lambda_E \quad (21.8)$$

The boundary conditions for these adjoint variables are

$$\lambda_P(\theta_0) = - \left[\frac{\partial \Phi}{\partial \mathbf{x}_P(\theta_0)} \right]^T \quad \text{and} \quad \lambda_P(\theta_f) = \left[\frac{\partial \Phi}{\partial \mathbf{x}_P(\theta_f)} \right]^T \quad (21.9)$$

$$\lambda_E(\theta_0) = - \left[\frac{\partial \Phi}{\partial \mathbf{x}_E(\theta_0)} \right]^T \quad \text{and} \quad \lambda_E(\theta_f) = \left[\frac{\partial \Phi}{\partial \mathbf{x}_E(\theta_f)} \right]^T \quad (21.10)$$

As unbounded control variables are assumed, the following first-order conditions hold for \mathbf{u}_P and \mathbf{u}_E :

$$\left[\frac{\partial H}{\partial \mathbf{u}_P} \right]^T = \left[\frac{\partial \mathbf{f}_P}{\partial \mathbf{u}_P} \right]^T \lambda_P = \mathbf{0} \quad (21.11)$$

$$\left[\frac{\partial H}{\partial \mathbf{u}_E} \right]^T = \left[\frac{\partial \mathbf{f}_E}{\partial \mathbf{u}_E} \right]^T \lambda_E = \mathbf{0} \quad (21.12)$$

in conjunction with the second-order conditions:

$$H_{\mathbf{u}_P \mathbf{u}_P} \doteq \frac{\partial^2 H}{\partial \mathbf{u}_P^2} \leq 0 \quad (21.13)$$

$$H_{\mathbf{u}_E \mathbf{u}_E} \doteq \frac{\partial^2 H}{\partial \mathbf{u}_E^2} \geq 0 \quad (21.14)$$

i.e., the Hessian matrix $H_{\mathbf{u}_P \mathbf{u}_P}$ must be negative semidefinite, whereas the Hessian $H_{\mathbf{u}_E \mathbf{u}_E}$ must be positive semidefinite. The second-order conditions (21.13) and (21.14) represent the extension of the Clebsch–Legendre conditions [6] to zero-sum games. Equations (21.13) and (21.14) enforce the respective first-order conditions, which properly constitute necessary conditions for stationarity of the solution with respect to the control variables \mathbf{u}_P and \mathbf{u}_E . Finally, the transversality conditions are related to the initial and terminal times (θ_0 and θ_f) and are written as follows [5, 6]:

$$\frac{\partial \Phi}{\partial \theta_0} - H_0 = 0 \quad \text{if } \theta_0 \text{ is unspecified} \quad (21.15)$$

$$\frac{\partial \Phi}{\partial \theta_f} + H_f = 0 \quad \text{if } \theta_f \text{ is unspecified} \quad (21.16)$$

If either θ_0 or θ_f is unspecified, the respective relationship becomes unnecessary.

Equations (21.1), (21.2) and (21.3) and (21.7), (21.8), (21.9), (21.10), (21.11), (21.12), (21.13), (21.14), (21.15) and (21.16) form a two-point boundary value problem (TPBVP) where the unknowns are the n_P -dimensional vectors $\mathbf{x}_P(\theta)$ and $\lambda_P(\theta)$, the n_E -dimensional vectors $\mathbf{x}_E(\theta)$ and $\lambda_E(\theta)$, the m_P -dimensional vector $\mathbf{u}_P(\theta)$, the m_E -dimensional vector $\mathbf{u}_E(\theta)$, the q -dimensional time-independent multiplier ν , and (possibly) the times θ_0 and θ_f .

21.3 Numerical Solution of Two-Sided Optimization Problems

The numerical solution of two-sided optimization problems is based on the formal conversion of the problem into a single-objective one, followed by the application of the sequential gradient-restoration algorithm (SGRA). This issue is addressed in the next section, whereas an outline of the main features of the SGRA is reported in Sect. 21.3.2.

21.3.1 Transformation into Single-Objective Problem

Algorithms devoted to optimal control problems are unable to solve zero-sum differential games, since they are designed to find the minimum value of a single-objective

function related to a given dynamic system. However, recently an innovative approach has been proposed [7, 8]; it allows the formal transformation of a zero-sum game into an optimal control problem. This method is based on the following points:

- the control for one player, for instance the pursuer P , is found from the optimality conditions (21.11 and 21.13) and can be expressed as

$$\mathbf{u}_P = \mathbf{u}_P(\mathbf{x}_P, \lambda_P, \theta) \quad (21.17)$$

- the control for the other player, i.e., the evader E , is found numerically
- an “extended state” is defined as follows:

$$\tilde{\mathbf{x}}(\theta) = [\mathbf{x}_P^T(\theta) \quad \mathbf{x}_E^T(\theta) \quad \lambda_P^T(\theta)]^T \quad (\tilde{n}\text{-dimensional vector}) \quad (21.18)$$

- a new control variable, which includes \mathbf{u}_E only, is introduced:

$$\tilde{\mathbf{u}}(\theta) = \mathbf{u}_E(\theta) \quad (\tilde{m}\text{-dimensional vector}) \quad (21.19)$$

Hence, the extended dynamic equations related to $\tilde{\mathbf{x}}$ can be written by taking into account the state equations (21.1)–(21.2) and the adjoint equation (21.7):

$$\tilde{\mathbf{x}}'(\theta) = \left[\mathbf{f}_P^T \quad \mathbf{f}_E^T \quad -\lambda_P^T \left[\frac{\partial \mathbf{f}_P}{\partial \mathbf{x}_P} \right] \right]^T \doteq \tilde{\mathbf{f}} \quad (21.20)$$

where $\mathbf{f}_P = \mathbf{f}_P(\mathbf{x}_P, \mathbf{u}_P(\mathbf{x}_P, \lambda_P, \theta), \theta) = \mathbf{f}_P(\mathbf{x}_P, \lambda_P, \theta)$. The extended boundary conditions are

$$\tilde{\boldsymbol{\psi}} = [\boldsymbol{\psi}^T \quad \boldsymbol{\psi}_{EXT}^T]^T = \mathbf{0} \quad (\tilde{q} \text{ boundary conditions}) \quad (21.21)$$

and now include the additional term $\boldsymbol{\psi}_{EXT}$, which consists of the boundary conditions related to the multipliers λ_P , after eliminating the components of $\boldsymbol{\psi}$ from the relationships (21.9)–(21.10). In Sect. 21.4 and 21.5 the steps leading to the definition of $\boldsymbol{\psi}_{EXT}$ will be detailed. Definitely, the zero-sum game has been transformed into the following optimal control problem:

$$\min_{\tilde{\mathbf{u}}(\theta)} J \quad \text{subject to the constraints (21.20) and (21.21)} \quad (21.22)$$

The inclusion of the adjoint equations for P has the unfavorable consequence that the starting guesses for P ’s adjoint variables are needed. These guesses affect the sequential gradient-restoration algorithm convergence. In addition, adjoint variables usually have a nonintuitive significance. As a consequence, the use of a trial-and-error approach is likely to be unsuccessful and an alternative approach is desirable. Evolutionary methods represent a systematic approach to the solution of this difficulty because they do not require any a priori information about the solution. As a matter of fact, they have already been successfully employed as preprocessing techniques in differential game contexts [7, 8]. The complete set of the unknown parameters forms an individual; each generation is composed of a large number of

individuals. After a specified (large) number of generations the method is expected to produce the best individual, which represents the optimal approximate solution to the problem. Genetic algorithms are characterized by a poor numerical accuracy. Yet, this property is not a limitation when they are used as a preprocessing technique, i.e., just to provide a reasonable guess for the successive use of the SGRA.

21.3.2 Sequential Gradient-Restoration Algorithm

The sequential gradient-restoration algorithm (SGRA) is a first-order indirect method introduced by Angelo Miele [2–4], and devoted to optimal control problems, consisting in the minimization of a performance index J . The dynamic system is assumed to be subject to the following constraints:

$$\tilde{\mathbf{x}}'(\theta) = \tilde{\mathbf{f}}(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}, \boldsymbol{\pi}, \theta) \quad (21.23)$$

$$\tilde{\boldsymbol{\psi}}(\tilde{\mathbf{y}}_{0u}, \tilde{\mathbf{y}}, \boldsymbol{\pi}) = \mathbf{0} \quad (21.24)$$

where

- $\tilde{\mathbf{x}}$ is the \tilde{n} -dimensional state vector
- $\tilde{\mathbf{u}}$ is the \tilde{m} -dimensional control vector
- $\boldsymbol{\pi}$ is the p -dimensional vector including all the time-independent parameters to be optimized
- $\tilde{\mathbf{y}}_{0u}$ collects all the \tilde{b} unspecified initial values of $\tilde{\mathbf{x}}$
- $\tilde{\mathbf{y}}$ includes all the final values of $\tilde{\mathbf{x}}$

Without any loss of generality, the optimal control problem can be written as a Mayer problem:

$$\min_{\tilde{\mathbf{u}}(\theta), \boldsymbol{\pi}} J \quad \text{where} \quad J = \phi(\tilde{\mathbf{y}}_{0u}, \tilde{\mathbf{y}}, \boldsymbol{\pi}) \quad (21.25)$$

The SGRA is characterized by a high numerical accuracy and is based on the cyclic execution of two phases:

- the **gradient phase**, aimed at decreasing the value of the objective function while the constraints are satisfied to first order;
- the **restoration phase**, aimed at decreasing the constraint error while avoiding excessive changes in the control and parameter vectors.

The SGRA has been also formulated in multiple subarc form [9], which formulation has two additional features:

- (i) enhanced robustness in solving complex problems with different timescales;
- (ii) ability to solve problems with discontinuities in the state and control variables.

The sequential gradient-restoration algorithm stops when both the constraints and the (first order) optimality conditions are satisfied to the respective (prefixed) accuracies.

21.4 Homicidal Chauffeur Game

This section is focused on the solution of the classical *homicidal chauffeur game*, which is one of most well-studied differential game problems. It consists in the identification of the saddle-point trajectories that lead to the capture of a low-speed, highly maneuverable evader (the “pedestrian”) by a high-speed less-maneuverable pursuer, the “chauffeur”. Termination of the game occurs when the distance between the two players becomes the capture radius, R_{cap} . The objective function is related to the interception time, which is to be minimized by the pursuer and maximized by the evader. Each player is assumed to possess complete and instantaneous information on the state of the opponent player. The method of solution is based on the application of the SGRA, after the transformation of the dual-sided optimization problem into an optimal control problem.

21.4.1 Formulation of the Problem

The pursuer and the evader are assumed to move with their respective constant velocities v_P and v_E in a two-dimensional plane (with coordinates x, y). The pursuer velocity is greater than that of the evader, whereas this latter player has superior maneuverability, i.e., he is able to change instantaneously his turning angle u_E , which is assumed as the control variable. Conversely, the pursuer has a minimum turning radius equal to 1. This circumstance implies that his turning rate ω is constrained: $-1 \leq \omega \leq 1$. Hence, in the inertial frame (x, y) , the equations of motion are

$$\begin{cases} \dot{x}_P' = v_P \cos \varphi \\ \dot{y}_P' = v_P \sin \varphi \\ \dot{\varphi}' = \omega \end{cases} \quad \begin{cases} \dot{x}_E' = v_E \cos u_E \\ \dot{y}_E' = v_E \sin u_E \end{cases} \quad (21.26)$$

where φ and u_E are the turning angles of the two players, measured in counterclockwise sense from the x -axis. An unconstrained control variable can be introduced for the pursuer by letting $\omega = \cos u_P$. As a result, the state vectors related to the two players are given by

$$\mathbf{x}_P = [x_P \quad y_P \quad \varphi]^T \quad \mathbf{x}_E = [x_E \quad y_E]^T \quad (21.27)$$

and the control vectors \mathbf{u}_P and \mathbf{u}_E coincide with the scalar variables u_P and u_E , respectively. For all the cases which will be considered, the initial conditions of the pursuer (at $\theta_0 = 0$) are the following:

$$x_P(0) = 0, \quad y_P(0) = 0, \quad \varphi(0) = \frac{\pi}{2} \quad (21.28)$$

and different initial positions of the evader correspond to distinct illustrative cases, whose numerical solution is reported in Sect. 21.4.3.

Termination occurs when the distance between the two players becomes the capture radius R_{cap} :

$$[x_P(\theta_f) - x_E(\theta_f)]^2 + [y_P(\theta_f) - y_E(\theta_f)]^2 - R_{cap}^2 = 0 \quad (21.29)$$

The left-hand side of (21.29) is included in ψ , which collects all the boundary conditions (according to (21.3)). Finally, the objective function is

$$J = -\theta_f^2 \quad (= \phi) \quad (21.30)$$

and must be maximized by the pursuer and minimized by the evader.

21.4.2 Method of Solution

The necessary conditions described in Sect. 21.2 allow the definition of the two-point boundary-value problem related to the game at hand. Letting $\lambda_P = [\lambda_1 \ \lambda_2 \ \lambda_3]^T$ and $\lambda_E = [\lambda_4 \ \lambda_5]^T$, the relationships (21.7)–(21.8) yield

$$\lambda_1' = \lambda_2' = \lambda_4' = \lambda_5' = 0 \quad (21.31)$$

$$\lambda_3' = v_P (\lambda_1 \sin \varphi - \lambda_2 \cos \varphi) \quad (21.32)$$

Due to (21.31), the Lagrange multipliers λ_1 , λ_2 , λ_4 , and λ_5 turn out to be constant. The respective boundary conditions (21.9) and (21.10) are

$$\lambda_3(\theta_f) = 0 \quad (21.33)$$

$$\lambda_1 = -\lambda_4 = 2v[x_1(\theta_f) - x_4(\theta_f)] \quad (21.34)$$

$$\lambda_2 = -\lambda_5 = 2v[x_2(\theta_f) - x_5(\theta_f)] \quad (21.35)$$

where v is the adjoint variable conjugate to the boundary condition (21.29). After combining (21.34) and (21.35) one obtains

$$\lambda_1 [x_2(\theta_f) - x_5(\theta_f)] - \lambda_2 [x_1(\theta_f) - x_4(\theta_f)] = 0 \quad (21.36)$$

The numerical solution of the dynamic game at hand requires the expression of the control u_P as a function of λ_P and \mathbf{x}_P through the necessary conditions (21.11) and (21.13), which yield

$$\lambda_3 \sin u_P = 0 \quad \text{and} \quad \lambda_3 \cos u_P \geq 0 \quad (21.37)$$

These conditions, in conjunction with (21.32), allow writing u_P as follows:

$$u_P = \begin{cases} 0 & \text{if } \lambda_3 > 0 \\ \pi/2 & \text{if } \lambda_3 = 0 \\ \pi & \text{if } \lambda_3 < 0 \end{cases} \quad \text{and} \quad \lambda_1 \sin \varphi - \lambda_2 \cos \varphi = 0 \quad (21.38)$$

The control for the evader, u_E , is found numerically by the SGRA ($\tilde{\mathbf{u}} = u_E$). The conversion of the dual-sided optimization problem into an optimal control problem is based on the inclusion of the time-varying components of λ_P in $\tilde{\mathbf{x}}$:

$$\tilde{\mathbf{x}} = [x_P \ y_P \ \varphi \ x_E \ y_E \ \lambda_3]^T \quad (21.39)$$

whereas the constant components λ_1 and λ_2 are included in the parameter vector π , as well as the unspecified terminal time θ_f :

$$\pi = [\lambda_1 \ \lambda_2 \ \theta_f]^T \quad (21.40)$$

The auxiliary vector $\tilde{\mathbf{y}}_{0u}$ is composed of the unknown components of $\tilde{\mathbf{x}}(0)$: $\tilde{\mathbf{y}}_{0u} = \lambda_3(0)$. Finally, the constraints (21.29), (21.33), and (21.36) are included in $\tilde{\psi}$.

The transformation of the game into a single-objective problem is thus completed. For each case, starting from a reasonable guess “solution” (not necessarily provided by a genetic algorithm in this simple application), the SGRA is employed to find the refined numerical solution to the game.

21.4.3 Numerical Results

The following six cases, corresponding to different initial position of the evader, have been considered:

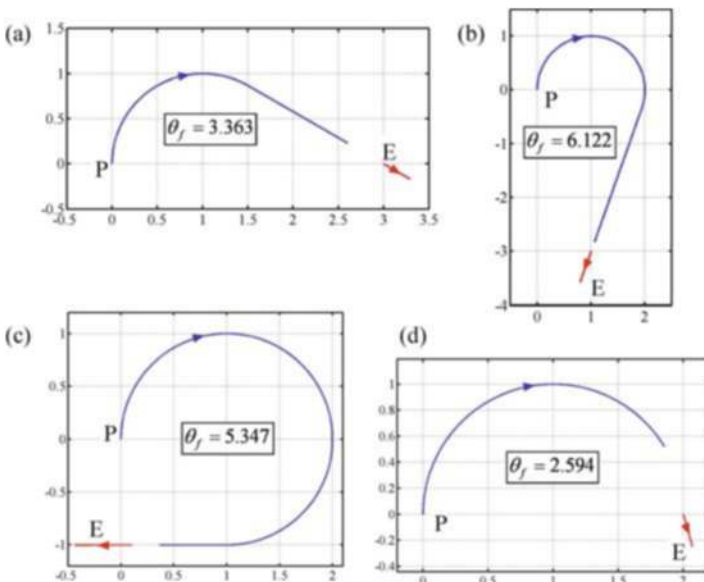


Fig. 21.1 Saddle-point trajectories for cases 1–4

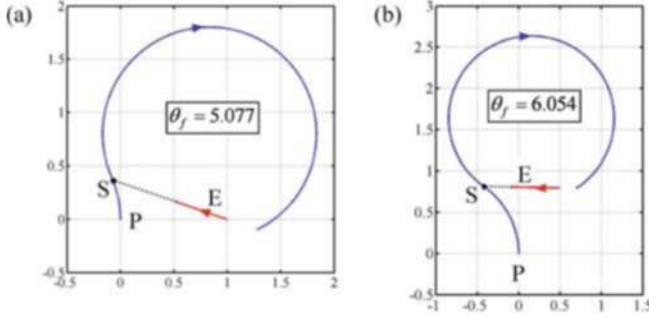


Fig. 21.2 Saddle-point trajectories for cases 5 and 6

1. $\begin{cases} x_E(0) = 3 \\ y_E(0) = 0 \end{cases}$
2. $\begin{cases} x_E(0) = 1 \\ y_E(0) = -3 \end{cases}$
3. $\begin{cases} x_E(0) = 0.1 \\ y_E(0) = -1 \end{cases}$
4. $\begin{cases} x_E(0) = 2 \\ y_E(0) = 0 \end{cases}$
5. $\begin{cases} x_E(0) = 1 \\ y_E(0) = 0 \end{cases}$
6. $\begin{cases} x_E(0) = 0.5 \\ y_E(0) = 0.8 \end{cases}$

With reference to these cases, the optimal saddle-point trajectories are illustrated in Figs. 21.1 and 21.2, which include the capture times in the insets. The velocities v_P and v_E are set to 1 and 0.1, respectively, whereas the capture radius, R_{cap} , is set to 0.8. In all cases, the optimal saddle-point trajectory of the evader is represented by a rectilinear path. For cases 5 and 6, portrayed in Fig. 21.2, the optimal trajectory of the pursuer is bang–bang, and the optimal path of the evader is correctly directed toward the switching point S , where the control u_P switches from 0 to π . All the numerical results, albeit obtained only for six special cases, are consistent with those reported in the literature [1, 5].

21.5 Orbital Pursuit-Evasion Game

This section is concerned with the solution of an *orbital pursuit-evasion game*, consisting in the identification of the saddle-point trajectories that lead to the capture of an evasive spacecraft by a pursuing spacecraft. Termination occurs when P gets the instantaneous position of E and a sufficient condition will be stated to ensure that actually capture ends the game. The objective function is related to the interception time, which is to be minimized by the pursuer and maximized by the evader. Each player is assumed to possess complete and instantaneous information on the state of the opponent player. The method of solution is based on the joint application of the genetic algorithm NSGA-II developed by Deb [10] as preprocessing technique and of the sequential gradient-restoration algorithm.

21.5.1 Formulation of the Problem

This research employs a point-mass model to describe the motion of the two spacecraft, in the context of a two-degree-of-freedom problem of optimal interception. This means that orbital motion of both spacecraft is confined to the same orbital plane. For the sake of simplicity, atmospheric forces are not included in the model. In addition, a constant thrust-to-mass ratio is assumed for both spacecraft, so the thrust pointing angle is the only control for both players.

A sufficient condition ensuring that interception concludes the game is

$$\frac{T_P}{m_P} > \frac{T_E}{m_E} \quad (21.41)$$

i.e., the thrust-to-mass ratio of the pursuer is assumed greater than that of the evader in all the cases which will be investigated. The condition (21.41) states that the pursuer has superior capabilities with respect to the evader. As unbounded controls are assumed for both spacecraft, this condition implies that interception can occur in a finite time.

With reference to Fig. 21.3, the orbital motion of the two spacecraft can be described through the following state variables:

$$\mathbf{x}_P = [v_{rP} \ v_{\theta P} \ r_P \ \xi_P]^T \quad \mathbf{x}_E = [v_{rE} \ v_{\theta E} \ r_E \ \xi_E]^T \quad (21.42)$$

where v_r and v_θ are the radial and the horizontal components of the velocities (\mathbf{v}_P and \mathbf{v}_E), r denotes the radius (i.e., the distance from the center of the attracting body) and ξ represents the angular displacement of the position vector \mathbf{r} from the inertial

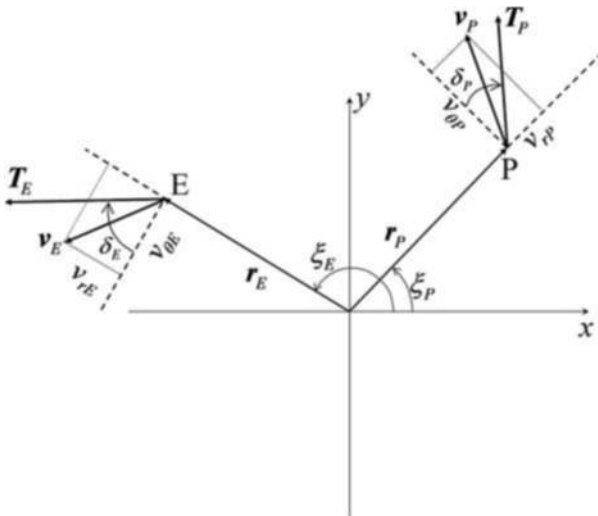


Fig. 21.3 Rotating frames of the two spacecraft

axis x . If δ_P and δ_E are the thrust pointing angles of the two players, the control variables are simply given by $\mathbf{u}_P = \delta_P$ and $\mathbf{u}_E = \delta_E$.

The equations of motion may be written in the rotating frames portrayed in Fig. 21.3:

$$v'_{ri} = \frac{T_i}{m_i} \sin \delta_i - \frac{\mu - v_{\theta i}^2 r_i}{r_i^2} \quad (21.43)$$

$$v'_{\theta i} = \frac{T_i}{m_i} \cos \delta_i - \frac{v_{ri} v_{\theta i}}{r_i} \quad (21.44)$$

$$r'_i = v_{ri} \quad (21.45)$$

$$\xi'_i = \frac{v_{\theta i}}{r_i} \quad (21.46)$$

where $i = P$ or $i = E$ and μ represents the planetary constant of the attracting body.

The two spacecraft are assumed to be placed in two circular (distinct) orbits at θ_0 , which is set to 0. In addition, the x -axis is such that $\xi_P(0) = 0$. Hence, the starting conditions are expressed as follows:

$$v_{rP}(0) = 0, \quad v_{\theta P}(0) = \sqrt{\frac{\mu}{r_{P0}}}, \quad r_P(0) = r_{P0} (= \text{given}), \quad \xi_P(0) = 0 \quad (21.47)$$

$$v_{rE}(0) = 0, \quad v_{\theta E}(0) = \sqrt{\frac{\mu}{r_{E0}}}, \quad r_E(0) = r_{E0} (= \text{given}), \quad \xi_E(0) = \xi_{E0} \quad (21.48)$$

The angular displacement between the two spacecraft is ξ_{E0} and affects the outcome of the game, i.e., the interception time.

The boundary conditions for the problem at hand are the following:

$$r_P(\theta_f) - r_E(\theta_f) = 0 \quad \text{and} \quad \xi_P(\theta_f) - \xi_E(\theta_f) = 0 \quad (21.49)$$

whereas the objective function is given by

$$J = -\theta_f^2 \quad (21.50)$$

J must be maximized by the pursuer and minimized by the evader.

21.5.2 Method of Solution

The necessary conditions described in Sect. 21.2 allow the definition of the two-point boundary value problem related to the game at hand. First of all, the adjoint variables conjugate to the state equations of P and E are written as follows:

$$\lambda_P = [\lambda_1 \quad \lambda_2 \quad \lambda_3 \quad \lambda_4]^T \quad \text{and} \quad \lambda_E = [\lambda_5 \quad \lambda_6 \quad \lambda_7 \quad \lambda_8]^T \quad (21.51)$$

whereas the Lagrange multiplier related to (21.49) is

$$\mathbf{v} = [v_1 \quad v_2]^T \quad (21.52)$$

Omitting some details for the sake of brevity, (21.11), (21.12), (21.13) and (21.14) yield the following expressions for the control variables:

$$\sin \delta_P = \frac{\lambda_1}{\sqrt{\lambda_1^2 + \lambda_2^2}} \quad \text{and} \quad \cos \delta_P = \frac{\lambda_2}{\sqrt{\lambda_1^2 + \lambda_2^2}} \quad (21.53)$$

$$\sin \delta_E = -\frac{\lambda_5}{\sqrt{\lambda_5^2 + \lambda_6^2}} \quad \text{and} \quad \cos \delta_E = -\frac{\lambda_6}{\sqrt{\lambda_5^2 + \lambda_6^2}} \quad (21.54)$$

In addition, as $\lambda'_4 = 0$ and $\lambda'_8 = 0$, λ_4 and λ_8 are constant. The boundary conditions for the adjoint variables λ_P and λ_E turn out to be

$$\lambda_1(\theta_f) = \lambda_2(\theta_f) = 0 \quad (21.55)$$

$$\lambda_3(\theta_f) = v_1 \quad (21.56)$$

$$\lambda_4(\theta_f) = \lambda_4 = v_2 \quad (21.57)$$

$$\lambda_5(\theta_f) = \lambda_6(\theta_f) = 0 \quad (21.58)$$

$$\lambda_7(\theta_f) = -v_1 \quad (21.59)$$

$$\lambda_8(\theta_f) = \lambda_8 = -v_2 \quad (21.60)$$

After eliminating v_1 and v_2 , (21.56), (21.57), (21.59) and (21.60) can be replaced with the following equations:

$$\lambda_3(\theta_f) + \lambda_7(\theta_f) = 0 \quad (21.61)$$

$$\lambda_4 + \lambda_8 = 0 \quad (21.62)$$

Finally, the transversality condition (21.16) (related to the unknown terminal time) yields

$$2\theta_f - \lambda_3(\theta_f) [v_{rP}(\theta_f) - v_{rE}(\theta_f)] - \lambda_4 \left[\frac{v_{\theta P}(\theta_f)}{r_P(\theta_f)} - \frac{v_{\theta E}(\theta_f)}{r_E(\theta_f)} \right] = 0 \quad (21.63)$$

Equations (21.55), (21.58), and (21.61), (21.62) and (21.63), in conjunction with the boundary conditions (21.49) represent a set of nine constraints to be satisfied.

In the GA preprocessing, each individual corresponds to a set of parameters representing the unknown values of the state and costate variables at $\theta_0 (= 0)$. For the problem at hand, this set consists of nine unknown parameters and also includes the time θ_f :

$$\{\lambda_1(0), \quad \lambda_2(0), \quad \lambda_3(0), \quad \lambda_4, \quad \lambda_5(0), \quad \lambda_6(0), \quad \lambda_7(0), \quad \lambda_8, \quad \theta_f\} \quad (21.64)$$

For each individual, the GA numerically integrates the state and the adjoint equations (which are not reported for the sake of brevity) and evaluates the constraint violation. Then the GA selects the best individual, i.e., the set of parameters corresponding to the minimum constraint violation.

To get a refined solution through the SGRA, the zero-sum game must be reformulated as an optimal control problem. First of all, P 's adjoint variables are included into the extended state $\tilde{\mathbf{x}}$:

$$\tilde{\mathbf{x}} = [v_{rP} \quad v_{\theta P} \quad r_P \quad \xi_P \quad v_{rE} \quad v_{\theta E} \quad r_E \quad \xi_E \quad \lambda_1 \quad \lambda_2 \quad \lambda_3]^T \quad (21.65)$$

with the exception of the constant component λ_4 , which is inserted in the parameter vector π (also including the time θ_f):

$$\pi = [\theta_f \quad \lambda_4]^T \quad (21.66)$$

The auxiliary vector $\tilde{\mathbf{y}}_{0u}$ is composed of all the unknown components of $\tilde{\mathbf{x}}(0)$:

$$\tilde{\mathbf{y}}_{0u} = [\lambda_1(0) \quad \lambda_2(0) \quad \lambda_3(0)]^T \quad (21.67)$$

The control variable $\tilde{\mathbf{u}}$ is simply $\tilde{\mathbf{u}} = \delta_E$. Finally, the constraints (21.49), (21.55), (21.58), and (21.61), (21.62) and (21.63) are considered. Only those constraints that are independent of λ_E are taken into account in the SGRA. Hence, $\tilde{\psi}$ is composed of five components ($\tilde{q} = 5$), associated to the constraints (21.49), (21.55), and (21.63). The transformation of the game into a single-objective problem is thus completed.

Starting from the guess "solution" provided by the GA, the SGRA is employed to find the numerical solution to such transformed problem.

21.5.3 Numerical Results

For the numerical solution of the interception problem, canonical units for distances and times (denoted with DU and TU, respectively) were used. The planetary constant of the attracting body was set to $1 \text{ DU}^3/\text{TU}^2$.

A first test case was run to check the performance of the method, using the following data:

$$r_{P0} = 1 \text{ DU} \quad r_{E0} = 1.05 \text{ DU} \quad \xi_{E0} = 20 \text{ deg} \quad (21.68)$$

$$\frac{T_P}{m_P} = 0.05 \frac{\text{DU}}{\text{TU}^2} \quad \frac{T_E}{m_E} = 0.0025 \frac{\text{DU}}{\text{TU}^2} \quad (21.69)$$

The related optimal control angles and saddle-point trajectories are illustrated in Figs. 21.4 and 21.5. In particular, Fig. 21.4 shows that the GA preprocessing generates a reasonable guess for the subsequent application of the SGRA. However, the time history of the thrust pointing angle of the evader, δ_E , is substantially altered by the SGRA with respect to the result produced by the GA preprocessing.

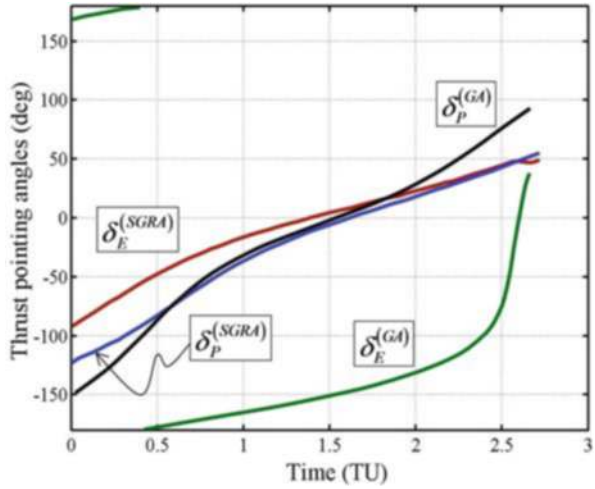


Fig. 21.4 Preprocessed and optimal control angles for the first test case

Other test cases were performed using the values $r_{P0} = 1\text{DU}$, $r_{E0} = 1.05\text{DU}$, and $T_E/m_E = 0.0025\text{DU}/\text{TU}^2$ for all of them. The results for different values of (T_P/m_P) and ξ_{E0} are summarized in Table 21.1, which reports the interception times θ_f . As expected, θ_f increases as ξ_{E0} increases and decreases as (T_P/m_P) increases. Several saddle-point trajectories are portrayed in Figs. 21.6 and 21.7, for different values of (T_P/m_P) and ξ_{E0} (shown in the insets).

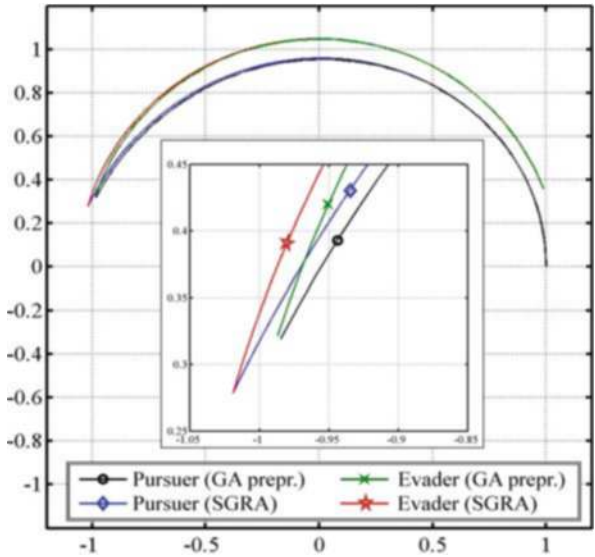


Fig. 21.5 Preprocessed and saddle-point trajectories for the first test case

Table 21.1 Interception time θ_f (TU) for different values of (T_p/m_P) and ξ_{E0}

T_p/m_P (DU/TU ²)		0.02	0.04	0.06	0.08	0.10
ξ_{E0} (deg)	10	2.4758	1.9890	1.7379	1.5711	1.4483
	20	3.6315	2.9189	2.5542	2.3129	2.1349
	30	4.3276	3.4756	3.0490	2.7695	2.5642
	40	4.8344	3.8619	3.3897	3.0836	2.8601
	50	5.2667	4.1476	3.6458	3.3176	3.0795
	60	5.6923	4.4057	3.8505	3.5015	3.2504
	70	6.1814	4.6250	4.0223	3.6520	3.3882
	80	6.8698	4.8338	4.1733	3.7797	3.5028
	90	8.2336	5.0480	4.3126	3.8922	3.6009

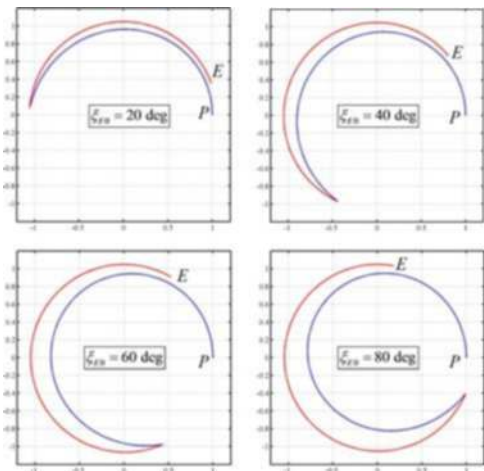


Fig. 21.6 Saddle-point trajectories for the case $(T_p/m_P) = 0.04\text{DU}/\text{TU}^2$

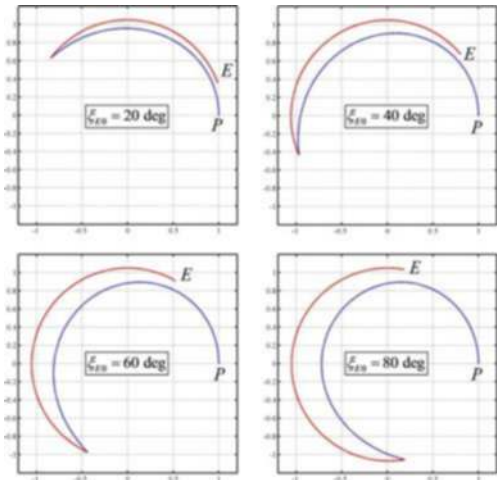


Fig. 21.7 Saddle-point trajectories for the case $(T_p/m_P) = 0.08\text{DU}/\text{TU}^2$

21.6 Conclusions

In the field of optimal aerospace trajectories, all the problems that involve two competitive actors can be appropriately modelled as zero-sum games. If compared to the vast number of studies regarding optimal control problems, only a limited number of researches have been concerned with the numerical solution of differential games [7, 11–16]. In some cases [11, 12], a simplified dynamics has been assumed in order to achieve analytical results for problems of possible practical interest. In the present research, a method is proposed for the numerical solution of zero-sum differential games. The method described in this chapter is based on the joint application of a genetic algorithm and of a local (indirect) algorithm (the sequential gradient-restoration algorithm), after the reformulation of the game as a single-objective problem. This approach has been successfully applied to the solution of two games: the well-known homicidal chauffeur game and an orbital pursuit-evasion game. Definitely, this work proves the concept of formulating (and solving) the problem of the optimal interception of an evasive target as a dynamic game.

For the orbital pursuit-evasion game, in all the cases the GA has produced a single approximate solution, corresponding to the minimum constraint violation. However, with minor revisions, the GA preprocessing should be able – at least in principle – to find more than a single solution that satisfies all the constraints with a prefixed accuracy. Hence, theoretically speaking, the method should be capable of investigating the existence of possible multiple saddle-point solutions.

In conclusion, with reference to the theoretical foundations of zero-sum games [1, 5], additional investigations are needed for the numerical characterization of singular surfaces, which often arise in dynamic game contexts. This final remark is also coherent with the considerations made by Breitner et al. [15] about the barrier, which is a type of singular surface.

References

1. Isaacs, R.: *Differential Games*. Wiley, New York (1964)
2. Miele, A., Pritchard, R.E.: Gradient methods in control theory, part 2: sequential gradient-restoration algorithm. Aero-Astronautics Report No. 62, Rice University, Houston (1969)
3. Gonzalez, S., Miele, A.: Sequential gradient-restoration algorithm for optimal control problems with general boundary conditions. *Journal of Optimization Theory and Applications*, **26**, No. 3, 395–425 (1978)
4. Miele, A., Wang, T., Basapur, V.K.: Primal and dual formulations of sequential gradient-restoration algorithms for trajectory optimization problems. *Acta Astronautica*, **13**, No. 8, 491–505 (1986)
5. Basar, T., Olsder, J.C.: *Dynamic Noncooperative Game Theory*. SIAM, Philadelphia (1999)
6. Bryson, A.E., Ho, Y.C.: *Applied Optimal Control*. Hemisphere, New York (1965)
7. Horie, K., Conway, B.A.: Optimal fighter pursuit-evasion maneuvers found via two-sided optimization. *Journal of Guidance, Control, and Dynamics*, **29**, No. 1, 105–112 (2006)
8. Horie, K.: Collocation with nonlinear programming for two-sided flight path optimization. Ph.D. Thesis, University of Illinois at Urbana-Champaign, Urbana (2002)

9. Miele, A., Wang, T.: Multiple-Subarc Gradient-Restoration Algorithm, part 1: algorithm structure. *Journal of Optimization Theory and Applications*, **116**, No. 1, 1–17 (2003)
10. Deb, K.: *Multi-Objective Optimization using Evolutionary Algorithms*. Wiley, Chichester (2001)
11. Breakwell, J.V., Merz, A.W.: Minimum required capture radius in a coplanar model of the aerial combat problem. *AIAA Journal*, **15**, No. 8, 1089–1094 (1977)
12. Guelman, M., Shinar, J., Green, A.: Qualitative study of a planar pursuit evasion game in the atmosphere. *Journal of Guidance, Control, and Dynamics*, **13**, No. 6, 1136–1142 (1990)
13. Roberts, D.A., Montgomery, R.C.: Development and application of a gradient method for solving differential games. Langley Research Center, NASA TN D-6502, Washington D.C. (1971)
14. Jarmark, B., Merz, A.W., Breakwell, J.V.: The variable speed tail-chase aerial combat problem. *Journal of Guidance, Control, and Dynamics*, **4**, No. 3, 323–328 (1981)
15. Breitner, M.H., Pesch, H.J., Grimm, W.: Complex differential games of pursuit-evasion type with state constraints, part I: necessary conditions for open-loop strategies. *Journal of Optimization Theory and Applications*, **78**, No. 3, 419–441 (1993)
16. Raivio, T., Ehtamo, H.: Visual aircraft identification as a pursuit-evasion game. *Journal of Guidance, Control, and Dynamics*, **23**, No. 4, 701–708 (2000)

Chapter 22

Interval Methods for Optimal Control

Andreas Rauh and Eberhard P. Hofer

Abstract Bellman's discrete dynamic programming is one of the most general approaches to solve optimal control problems. For discrete-time dynamical systems, it is, at least theoretically, capable to determine globally optimal control laws. In most practical cases, both state and control variables are subject to constraints. Due to the necessity for gridding of the range of both state and control variables in numerical implementations of dynamic programming, the computational effort grows exponentially with increasing system dimensions. This fact is well known as the *curse of dimensionality*. Furthermore, gridding of intervals representing uncertain system parameters is inevitable, if dynamic programming is used for the design of optimal controllers for systems with uncertainties. In this contribution, an interval arithmetic procedure for the design of optimal and robust controllers is presented. This procedure relies on the basic concepts of dynamic programming. Sophisticated techniques for the exclusion of non-optimal control strategies significantly reduce the computational burden. Since interval techniques can be applied to both continuous-time and discrete-time dynamical systems, the interval arithmetic optimization approach presented in this chapter is applicable to both cases. In addition, the inclusion of effects of uncertain parameters in the underlying optimality criteria is demonstrated. For that purpose, interval arithmetic routines for analysis and design of optimal and robust controllers have been developed. Details about computationally efficient implementations of interval arithmetic optimization procedures and numerical results for a mechanical positioning system with state-dependent switchings between different dynamical models for viscous and Coulomb friction are summarized.

Andreas Rauh

Chair of Mechatronics, University of Rostock, D-18059 Rostock, Germany,
e-mail: Andreas.Rauh@uni-rostock.de

Eberhard P. Hofer

Institute of Measurement, Control, and Microtechnology, University of Ulm, D-89069 Ulm, Germany, e-mail: Eberhard.Hofer@uni-ulm.de

This work was performed while Andreas Rauh was with the Institute of Measurement, Control, and Microtechnology, University of Ulm, Germany.

22.1 Introduction

In recent years, different optimization techniques for dynamical systems have been developed for both discrete-time and continuous-time systems. In the following, the state equations for discrete-time and continuous-time systems are assumed to be given in state-space representation by difference equations and sets of ordinary differential equations (ODEs).

The most important optimization procedures for *continuous-time* systems described by ODEs are based on *Pontryagin's maximum principle* [19] and the *Hamilton–Jacobi–Bellman equation* [5, 8]. The maximum principle leads to a boundary value problem for a set of ODEs while the Hamilton–Jacobi–Bellman equation is a nonlinear partial differential equation. In both cases, it is usually necessary to apply numerical techniques to determine optimal solutions for nonlinear real-world models.

For *discrete-time* systems, *Bellman's dynamic programming* is the most universal approach to calculate globally optimal control strategies [2, 3]. Since most implementations of dynamic programming rely on gridding of the admissible range of both state and control variables, this approach suffers from the so-called *curse of dimensionality*, i.e., the computational effort grows exponentially for increasing dimensions of the state and control vectors.

Numerical routines implementing dynamic programming techniques for nonlinear continuous-time processes make use of time discretization of the underlying state equations. If the resulting discretization errors are neglected, often considerable deviations from the original continuous-time systems arise. Therefore, such phenomena should be taken into account during optimization.

Interval arithmetic methods [6, 14] have already been applied successfully to enclose time discretization errors by guaranteed interval bounds in validated simulations of dynamical systems. These simulations are usually applied to determine guaranteed enclosures of all reachable states. Due to their capability to compute guaranteed state enclosures, interval methods are one possible approach to consider uncertain parameters in the analysis of the robustness of open-loop and closed-loop control systems as well as in the design of control strategies which are robust against parameter variations. One further application of interval methods is global optimization of static mathematical functions. Until now, only a few approaches toward the use of interval techniques in optimal controller design for dynamical systems exist. Recently, an algorithm determining piecewise constant control sequences which are restricted to a given small number of switchings has been published by Y. Lin and M. A. Stadtherr for low-dimensional problems, see e.g., [10, 11]. In contrast to the approach which will be presented in the following, effects of uncertain parameters have not been taken into account explicitly by Y. Lin and M. A. Stadtherr.

In the literature related to optimization of dynamical systems with uncertainties, state-dependent switchings between different dynamical models describing the plant to be controlled have not been considered. It is often necessary to consider state-dependent switchings in mathematical system models to obtain adequate descriptions for nonlinear phenomena such as friction in mechanical systems. A friction

characteristic with uncertain parameters is taken into account in the application scenario which is studied in this chapter.

The prerequisite for the use of interval methods in a framework for the design of optimal and robust controllers is not only the extension of optimality criteria to handle the influence of parameter uncertainties but also the computationally efficient implementation of the algorithms. The implemented optimization strategies have to tackle problems caused by the curse of dimensionality by an intelligent search for the global optimum of the performance index. These approaches include disregarding all control sequences which are either not admissible due to the violation of state constraints or not optimal with respect to the performance index. Non-admissible as well as non-optimal control sequences are eliminated already in early stages of the optimization process to avoid unnecessary computations. Although this technique helps to reduce the computational burden significantly, it cannot eliminate the curse of dimensionality completely. In contrast to other implementations of dynamic programming, gridding can be avoided when using interval methods. Instead, the suggested approach relies on an adaptive refinement of control variable intervals near the optimum of the performance index. Thus, the computed results are not subject to errors introduced by rounding toward nearest grid points.

In parallelized implementations, intermediate solutions which are constant for several subsequent time steps can be included as a further means to reduce the computational effort. These intermediate solutions provide upper bounds for the global optimum of the cost function to be minimized. This property can be used as an additional criterion to exclude non-optimal control strategies in early stages of the optimization. The presented optimization approach is suitable for generalization to the case of multiple control variables.

In Sect. 22.2, a detailed formulation of the considered optimization problems is given for discrete-time and continuous-time systems. The recently developed interval arithmetic optimization routine—which can be applied to both structure and parameter optimization—is introduced in Sect. 22.3. Section 22.4 summarizes a parallelized implementation of the suggested optimization routines. In Sect. 22.5, possibilities for combination of the interval arithmetic optimization technique with classical controller design are discussed to reduce the controlled systems' sensitivity with respect to parameter variations. For demonstration purposes, optimal control of a simplified continuous-time mechanical positioning system with state-dependent switchings between different dynamical models is discussed in the Sects. 22.6 and 22.7. These models describe the system's motion under consideration of viscous friction together with Coulomb friction subject to parameter uncertainties [23, 25]. Finally, in Sect. 22.8, an outlook on future research is given.

22.2 Optimal and Robust Control of Dynamical Systems

In this section, the problem of optimal and robust control of both discrete-time and continuous-time processes is formulated. In both cases, the goal is to transfer an

initial state vector into a desired final state vector such that a predefined performance index is minimized by choosing an admissible control strategy. In general, the two different problems of *parameter optimization* on the one hand and *structure optimization* on the other hand can be distinguished.

In the *parameter optimization* problem, parameters of a controller with a fixed structure, e.g., a P-, PI-, PID-, or linear state controller, have to be determined. In contrast, the result of *structure optimization* is an optimal open-loop or closed-loop control strategy which is determined without making any a priori assumptions about the structure of the controller.

22.2.1 Optimal Control of Discrete- and Continuous-Time Processes

For discrete-time dynamical systems

$$x_{k+1} = g_k(x_k, p_k, u_k, k) \quad , \quad (22.1)$$

with the state vector $x_k \in \mathbb{R}^{n_x}$ and the vector $p_k \in \mathbb{R}^{n_p}$ of system parameters, an initial state x_0 is transferred into a desired final state $x_f = x_{k_{max}}$, such that the performance index

$$J = g_{J, k_{max}}(x_{k_{max}}, p_{k_{max}}, k_{max}) + \sum_{k=0}^{k_{max}-1} g_{J, k}(x_k, p_k, u_k, k) \quad (22.2)$$

is minimized by calculation of an admissible control sequence $u_k \in \mathbb{R}^{n_u}$.

Analogously, continuous-time processes described by the set of ODEs

$$\dot{x}(t) = f(x(t), p(t), u(t), t) \quad , \quad (22.3)$$

with the state vector $x(t) \in \mathbb{R}^{n_x}$ and the vector $p(t) \in \mathbb{R}^{n_p}$ of system parameters can be considered. Again, an initial state x_0 is transferred into a desired final state $x_f = x(t_f)$ by calculation of a control law $u(t) \in \mathbb{R}^{n_u}$ such that the performance index

$$J = f_{t_f}(x(t_f), p(t_f), t_f) + \int_0^{t_f} f_0(x(t), p(t), u(t), t) dt \quad (22.4)$$

is minimized.

For both optimization problems, a finite time horizon is considered in the following. The time horizon is denoted by $k \in [0; k_{max}]$ in the discrete-time case and by $t \in [0; t_f]$ in the continuous-time case. In addition to exactly known initial and final states, also free boundary conditions or states from a certain predefined region of initial or final states can be investigated.

In many practical situations, uncertainties of system parameters have to be taken into account during the optimization process. If guaranteed bounds for these values are known, the uncertain system parameters can be described by the intervals $p_k \in$

$\left[\underline{p}_k; \overline{p}_k\right]$ and $p(t) \in \left[\underline{p}(t); \overline{p}(t)\right]$, resp. These intervals represent the maximum possible tolerances of the system parameters. Parameter variations are taken into account by additional discrete-time state equations

$$p_{k+1} = p_k + \Delta p_k \quad \text{with} \quad \Delta p_k \in \left[\underline{\Delta p}_k; \overline{\Delta p}_k\right] \quad (22.5)$$

or, in the continuous-time case, by additional ODEs

$$\dot{p}(t) = \Delta p(t) \quad \text{with} \quad \Delta p(t) \in \left[\underline{\Delta p}(t); \overline{\Delta p}(t)\right] \quad (22.6)$$

to describe the effects of the bounded variation rates Δp_k and $\Delta p(t)$, respectively. During the optimization procedure, limitations of all control variables u_k and $u(t)$, respectively, have to be considered. In this chapter, it is assumed that the control variables are bounded by the intervals $u_k \in [\underline{u}_k; \overline{u}_k]$ and $u(t) \in [\underline{u}(t); \overline{u}(t)]$, respectively, which do not need to be constant over the given time horizon.

22.2.2 Specification of Robustness in the Time Domain

As already pointed out, one of the main properties of the presented optimization procedure is its capability to directly deal with interval uncertainties of the system parameters. Hence, robustness of the controlled system with respect to parameter variations as well as optimality of a control sequence under the influence of parameter uncertainties have to be defined.

Robustness specifications as shown in Fig. 22.1 are assumed to be given by worst-case bounds of all system states which must not be violated during the transfer of the initial state into the desired final state using a control sequence which is completely inside its bounds for all points of time [22]. Furthermore, only the definition of either free final states or bounded regions of admissible final states makes sense in the case of uncertain parameters, since, in general, it is not possible to eliminate the influence of parameter uncertainties completely using one common control sequence for all possible parameter values.

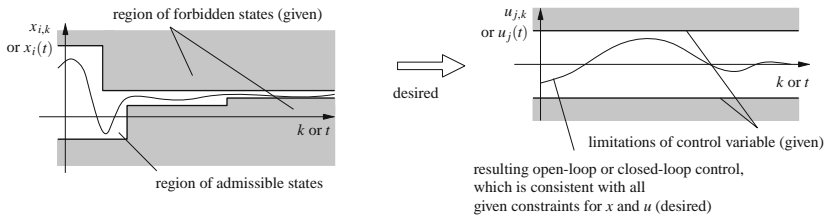


Fig. 22.1 Specification of robustness in the time domain for the state variables x_i , $i = 1, \dots, n_x$, and the control variables u_j , $j = 1, \dots, n_u$

22.2.3 Optimality Criteria for Systems with Uncertainties

In the case of simultaneous consideration of the above-mentioned bounds of the state variables and an optimization criterion as defined in (22.2) and (22.4), a control sequence is said to be optimal if it does not violate any of the specified bounds; at the same time the optimal control sequence must lead to the *smallest possible upper bound* of the performance index if the maximum influence of the uncertainties of initial states and parameters is investigated.

The interval arithmetic optimization algorithm presented in the following aims at calculating piecewise constant control sequences (with N time intervals in which the control is piecewise constant) within the prescribed bounds of the control variable intervals. The resulting trajectories of the state variables have to be included completely in the regions of admissible states for each possible value of the uncertain parameters p .

For discrete-time systems (22.1), the switching points for the control input are given by

$$\{k^s\} := \{k_1^s = 0, k_2^s, \dots, k_N^s, k_{N+1}^s = k_{\max}\} \quad . \quad (22.7)$$

Then, the control variable is constant for all $k \in \{k_i^s, k_i^s + 1, \dots, k_{i+1}^s - 1\}$, $i = 1, \dots, N$, i.e.,

$$u_{k_i^s}^s = u_{k_i^s+1}^s = \dots = u_{k_{i+1}^s-1}^s \quad . \quad (22.8)$$

After evaluating the state equations for subintervals of the maximum admissible range of u_k , guaranteed bounds for the performance index are obtained by

$$[J] := g_{J, k_{\max}} \left([x_{k_{\max}}], [p_{k_{\max}}], k_{\max} \right) + \sum_{k=0}^{k_{\max}-1} g_{J, k} \left([x_k], [p_k], [u_k], k \right) \quad . \quad (22.9)$$

The expression (22.9) is evaluated by replacement of all arithmetic operations by their corresponding interval counterparts.

In the continuous-time case, the switching points

$$\{t_k^s\} := \{t_1^s = t_0, t_2^s, \dots, t_N^s, t_{N+1}^s = t_f\} \quad (22.10)$$

are specified to define the piecewise constant control

$$u_k := u(t) = \text{const} \quad \text{for all } t \in [t_k^s; t_{k+1}^s] \quad (22.11)$$

and $k = 1, \dots, N$. To determine guaranteed enclosures $[J]$ of the performance index, the state equations are extended by the integrand of (22.4) according to

$$\begin{bmatrix} \dot{x}(t) \\ J(t) \end{bmatrix} = \begin{bmatrix} f(x(t), p(t), u(t), t) \\ f_0(x(t), p(t), u(t), t) \end{bmatrix} \quad \text{with} \quad \begin{bmatrix} x(0) \\ J(0) \end{bmatrix} = \begin{bmatrix} x_0 \\ 0 \end{bmatrix} \quad (22.12)$$

and

$$[J] := f_{t_f}([x(t_f)], [p(t_f)], t_f) + [J(t_f)] \quad . \quad (22.13)$$

To calculate guaranteed state enclosures for *discrete-time* systems, recursive interval evaluation of the state equations, global optimization techniques, as well as consistency tests and state–space transformations aiming at the reduction of overestimation are used. For *continuous-time* systems, arbitrary validated ODE solvers such as VNODE (using Taylor series expansions of the solution of an initial value problem IVP based on higher-order time derivatives of the ODE) [15–17], COSY VI (Taylor model-based solver with additional series expansion in the initial states) [4, 12], VSPODE [9], or VALENCIA-IVP¹ [1, 20] can be applied.

22.3 Interval Arithmetic Optimization Algorithm

The interval arithmetic optimization algorithm presented in this section is an extension of a procedure presented by the authors in [21, 25]. In addition to the original version, the new version can not only deal with discrete-time systems with nominal parameters; both discrete-time and continuous-time dynamical models including parameter uncertainties as well as constraints for the state variables representing time-domain robustness specifications can be handled. For continuous-time processes, a piecewise constant control law with a predefined sampling time, which is independent of the step sizes used by the underlying validated ODE solvers, is computed.

Step OPT 1 Based on a backward evaluation of the state equations from the final to the initial point of time, i.e., from $k = k_{max}$ to $k = 0$, and from $t = t_f$ to $t = 0$, respectively, guaranteed enclosures of all controllable states are determined. For discrete-time systems, the state equation (22.1) is solved for the state vector x_k under the assumption that an interval enclosure for x_{k+1} is known. If an analytical solution

$$x_k = \tilde{g}_k(x_{k+1}, p_k, u_k, k) \quad (22.14)$$

does not exist, interval Newton methods are used instead. Analogously, for continuous-time systems backward integration of the state equation

$$\dot{x}(t) = f(x(t), p(t), u(t), t) \quad (22.15)$$

is performed for given regions of final states $x(t_f)$ until the point of time $t = 0$ is reached. In both cases, the computation is performed for the known interval bounds of the uncertain parameters and the bounded range of the control variable intervals.

Simultaneously, interval enclosures of the performance indices (22.2) and (22.4) are determined according to

$$\begin{aligned} J_k &= J_{k+1} + g_{J,k}(x_k, p_k, u_k, k) \\ &= J_{k+1} + g_{J,k}(\tilde{g}_k(x_{k+1}, p_k, u_k, k), p_k, u_k, k) \end{aligned} \quad (22.16)$$

¹ VALENCIA-IVP is a validated ODE solver developed by A. Rauh and E. Auer. It is capable to compute guaranteed state enclosures for IVPs for both ODEs and differential algebraic systems, see also <http://www.valencia-ivp.com>.

and

$$J_t = f_{t_f}(x(t_f), p(t_f), t_f) + \int_t^{t_f} f_0(x(\tau), p(\tau), u(\tau), \tau) d\tau, \quad (22.17)$$

where the terminal cost functions are denoted by $g_{J, k_{\max}}$ (and f_{t_f} , respectively). The costs for transfer from the time step k to k_{\max} (or from t to t_f) are denoted by J_k (and J_t). The general idea of the presented optimization algorithm is the minimization of the performance index J_0 by repeated splitting of the control variable intervals. This procedure leads to an approximation of the globally optimal control sequences $\{u_k^*\}$ (and $u^*(t)$) as well as the corresponding optimal trajectories $\{x_k^*\}$ (and $x^*(t)$) for all $k \in [0; k_{\max}]$ (and $t \in [0; t_f]$).

Using interval techniques for backward evaluation of the state equations, validated enclosures of the regions of attraction are calculated which can be transferred into the desired final state under consideration of all possible values of the uncertain system parameters. The intervals $[J_k]$ and $[J_t]$ represent worst-case bounds of the range of the performance index for all control variables from the admissible range $[\underline{u}; \bar{u}]$. Thus, the influence of interval uncertainties is directly expressed in terms of the maximum variations of the system states and the performance index. Backward evaluation of the state equations is omitted if the final states $x_{k_{\max}}$ (and $x(t_f)$, respectively) are unbounded.

Step OPT 2 In this Step, the state equations are evaluated from the initial to the final point of time. Together with the results of **Step OPT 1**, candidates for control sequences are eliminated which do not allow to transfer a given initial state x_0 or interval box $[x_0]$ of initial states into the desired final state or region of final states. Additionally, non-optimal control sequences are eliminated which are detected by comparison of the performance index intervals of several candidates.

Step OPT 3 The interval widths for $\{[x_k]\}$ (and $[x(t)]$) as well as $\{[u_k]\}$ (and $[u(t)]$) of candidates for optimal control strategies are reduced by repeated forward and backward evaluations in the **Steps OPT 1** and **OPT 2** as long as a further improvement is possible.

Step OPT 4 During the global optimization procedure two strategies have proven successful. First, control sequences $\{[u_k]\}_{sup}$ (and $\{[u(t)]\}_{sup}$) are selected such that the corresponding supremum of the performance index is smaller than the supremum of all other candidates. This leads to a fast reduction of the upper bound of the necessary costs. For both discrete-time and continuous-time systems, the control variable interval is split at the point of time \tilde{k}^s according to

$$\tilde{k}^s := \arg \max_{j=0, \dots, N-1} \left\{ \text{diam} \left(\left[\frac{\partial J}{\partial u_j} \right] \right) \cdot \text{diam}([u_j]) \right\} \quad (22.18)$$

with $\text{diam}([u_j]) = \bar{u}_j - \underline{u}_j$. For splitting of the control interval at the point of time $k = \tilde{k}^s$ (or $t = t_{\tilde{k}^s}$) the largest reduction of the diameter of the performance index interval is expected. For continuous-time systems with piecewise constant control strategies

$u(t_k)$, the Eq. (22.18) is evaluated after discretization of the state equations with the same step size h_k that is used for validated integration. Second, the control sequences $\{[u_k]\}_{inf}$ (and $\{[u(t)]\}_{inf}$) with the smallest infimum of the performance index are selected to improve the lower bound of the necessary costs. Here, \tilde{k}^s is chosen again as in (22.18). *Suboptimal* control laws are obtained by performing the optimization under the assumption of control strategies which are *constant* for several subsequent time steps, i.e., for $N < k_{max}$ in the discrete-time case and $t_{k+1}^s - t_k^s = M \cdot h_k$, $M \in \mathbb{N} \setminus \{1\}$, for continuous-time systems.

The optimization is stopped if the diameter of the performance index interval of the best-known approximation of the optimal control sequence is smaller than a user-defined value and, at the same time, if the distance between the performance index of this candidate and the estimate for the global infimum of the necessary costs becomes smaller than another user-defined value.

Step OPT 5 Output of the best-known approximation for the optimal control sequences $\{u_k^*\}$ (and $u^*(t)$).

This approximation of the optimal control sequence is not ensured to be globally optimal. However, the results of the previous steps provide *guaranteed lower and upper bounds of the costs* which are necessary to perform the control task. By choosing appropriate stopping criteria for the optimization process in **Step OPT 4**, the user can make sure that the deviation between the costs for the approximation of the optimal control strategy and the global infimum of the performance index (assuming piecewise constant control strategies in both cases) is smaller than a prescribed value. The robustness of a control strategy is expressed directly in terms of the computed interval bounds for the states and the performance index which take into account the uncertainties of both initial conditions and system parameters.

22.4 Parallelization of the Optimization Algorithm

To reduce the computational effort by exclusion of non-optimal control sequences, the optimization algorithm has been parallelized according to Fig. 22.2. For that purpose, the MATLAB DISTRIBUTED COMPUTING TOOLBOX [29] is used, if the underlying routines for validated evaluation of the state equations are available as MATLAB code as in the application scenario in Sects. 22.6 and 22.7.

In the parallelized optimization routine, m independent tasks are automatically distributed to available standard PCs in a network. By simultaneous splitting of L_j , $j = 1, \dots, m$, different control sequences in all m tasks which are executed in parallel, a repeated selection of a single candidate is avoided. After a predefined number of iterations, non-optimal control sequences are eliminated by synchronization of the results of all tasks. Then, either the output of the best approximation for the globally optimal control strategy or a restart of the parallelized optimization routine is possible.

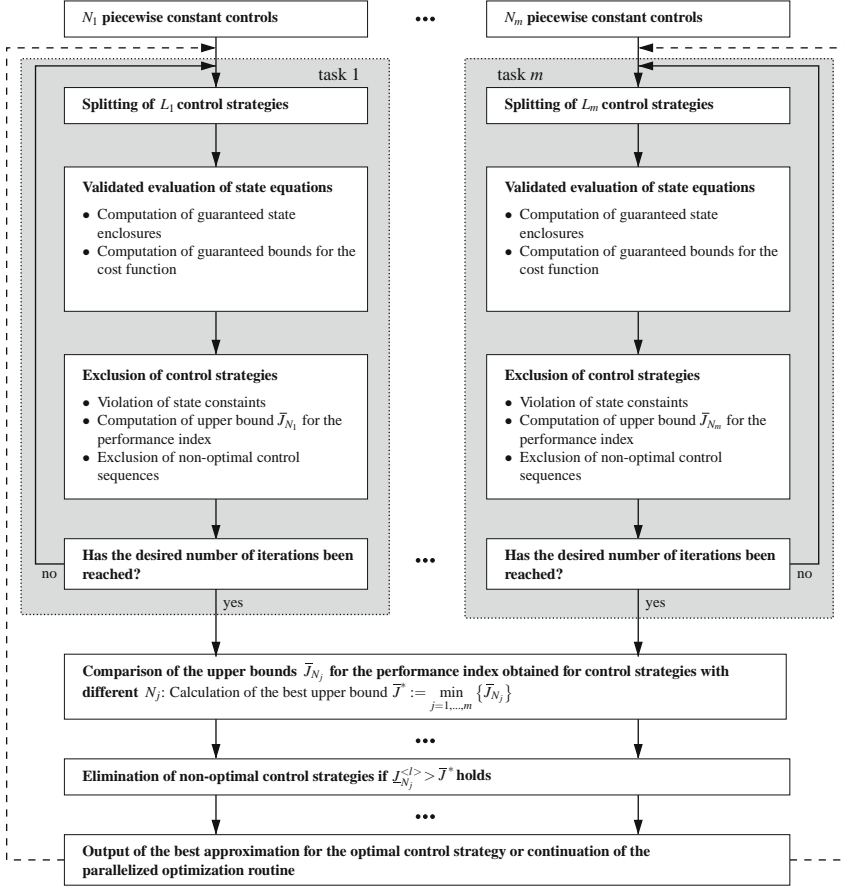


Fig. 22.2 Parallelized implementation of the interval arithmetic procedure for the calculation of optimal control strategies

22.5 Combination with Classical Controller Design

As it will be shown in the application scenario in Sect. 22.7, combinations of control laws resulting from structure optimization and suitable closed-loop controllers can be applied successfully to reduce the influence of parameter variations on the performance of a controlled dynamical system.

In general, the interval arithmetic optimization procedure described in Sect. 22.3 can be used for both structure and parameter optimizations. For $N = 1$, optimal constant control variables are obtained which correspond to the solution of a parameter optimization problem. Hence, also both constant and time-varying parameters of controllers with a given structure can be determined with the help of the presented optimization algorithm if a suitable performance index is given and if the controller parameters are interpreted as control inputs in the **Steps OPT 1–5**.

22.6 Validated Modeling and Simulation of Dynamical Systems with State-Dependent Switchings

In this section, modeling and optimization of a dynamical system with state-dependent switchings between different dynamical models are studied. This system represents a simplified model of a mechanical positioning unit with a sliding mass m as well as the given initial position $x_1(0) = 0$ and the given initial velocity $x_2(0) = 0$. The state-dependent switching characteristic reflects the different dynamical behavior of the system in the static and sliding friction modes. During the optimization, an optimal accelerating force $u(t) = F_a(t)$ has to be determined such that the region of admissible final states $[x_1(t_f)] = [0.9; 1.1]$, $[x_2(t_f)] = [-0.1; 0.1]$ is reached for $t_f = 5$ for all uncertain parameters of the friction characteristic $F_f(x_2)$.

According to [23], where a general simulation procedure for continuous-time dynamical systems with state-dependent switchings has been introduced in more detail, the positioning system is described by

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ \frac{1}{m} (F_a(t) - F_f(x_2)) \end{bmatrix} \quad (22.19)$$

with the state vector $x = [x_1 \ x_2]^T$. The friction characteristic $F_f(x_2)$ is described by three discrete model states $\mathcal{S} := \{S_1, S_2, S_3\}$, which are

- S_1 : sliding friction for motion in “negative” (backward) direction,
- S_2 : static friction, and
- S_3 : sliding friction for motion in “positive” (forward) direction.

In Fig. 22.3, the influence of interval parameters for the static friction coefficient $[F_s] := [\underline{F}_s; \overline{F}_s]$ as well as the sliding friction coefficient $[\mu] := [\underline{\mu}; \overline{\mu}]$ is depicted. The resulting sliding friction force is given by

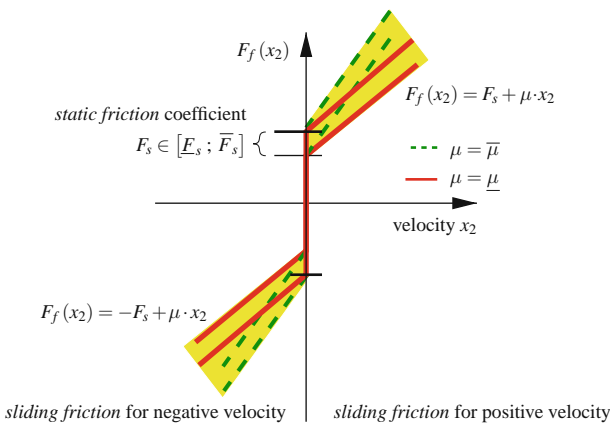


Fig. 22.3 Friction characteristic with uncertain sliding friction coefficient $[\mu]$ and uncertain static friction coefficient $[F_s]$

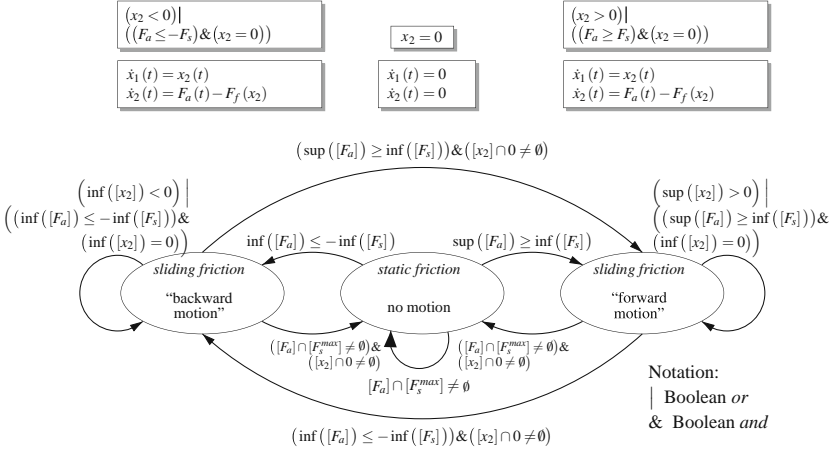


Fig. 22.4 State transition diagram for the friction characteristic with uncertainties

$$F_f(x_2) = \begin{cases} -[F_s] + [\mu] \cdot x_2 & \text{for } S_1 = \text{true} \\ +[F_s] + [\mu] \cdot x_2 & \text{for } S_3 = \text{true} \end{cases} \quad (22.20)$$

The static friction force is given by

$$F_f(x_2) \in [F_s^{max}] := [-\bar{F}_s; \bar{F}_s] \quad \text{for } S_2 = \text{true} \quad (22.21)$$

The state transition diagram in Fig. 22.4 illustrates the three model states S_1 , S_2 , and S_3 together with the conditions for all possible transitions between these states. In the case of parameter uncertainties, several states \mathcal{S} can be active simultaneously. The transition conditions T_i^j from state S_i to state S_j are expressed in terms of the state variables, the system parameters, and the control variable. In Fig. 22.4, the intersection operator for n -dimensional interval vectors $[v]$ and $[w]$ is defined as

$$[v] \cap [w] = \begin{cases} [\max\{\underline{v}, \underline{w}\}; \min\{\bar{v}, \bar{w}\}] & \text{if } \max\{\underline{v}_i, \underline{w}_i\} \leq \min\{\bar{v}_i, \bar{w}_i\} \quad \forall i = 1, \dots, n \\ \emptyset & \text{otherwise} \end{cases} \quad (22.22)$$

In the following, an extension of a Taylor series-based integration algorithm is described which allows for computation of validated enclosures of all reachable states in the case of state-dependent switchings by detecting all points of time at which transition conditions are activated or at which one of the discrete model states S_i is deactivated. The following procedure can also be applied to any other validated ODE solver if appropriate adjustments are made. For further information about interval methods for the computation of guaranteed state enclosures for dynamical systems with state-dependent switchings, the reader is referred to [18, 26] and the references therein.

Step SIM 1 A bounding box $[B_{a,k}]$ of all reachable states in the time interval $[t_k ; t_{k+1}]$ is calculated by a Picard iteration for the state equation $f_a(x(t), p, u(t), t)$ which is the union of *all* active models at $t = t_k$, i.e.,

$$\mathbb{F}_{f_a} \supseteq \bigcup_{i \in \mathcal{J}_a} \mathbb{F}_{f_{S_i}} , \quad (22.23)$$

where \mathbb{F}_{f_a} and $\mathbb{F}_{f_{S_i}}$ represent the *exact* value sets

$$\mathbb{F}_{f_a} := \{y \mid y = f_a(x, p, u, t_k) \ \forall x \in [x(t_k)], p \in [p(t_k)], u \in [u(t_k)]\} \quad (22.24)$$

and

$$\mathbb{F}_{f_{S_i}} := \{y \mid y = f_{S_i}(x, p, u, t_k) \ \forall x \in [x(t_k)], p \in [p(t_k)], u \in [u(t_k)]\} \quad (22.25)$$

of f_a and f_{S_i} under consideration of all interval arguments with

$$\mathcal{J}_a := \{i \mid S_i = \text{true}\} . \quad (22.26)$$

The subscript a indicates that the state equation f_a consists of the union of all *active* models.

Step SIM 2 If an additional transition condition is activated for one of the active models S_i , $i \in \mathcal{J}_a$, during the time interval $[t_k ; t_{k+1}]$, the bounding box $[B_{a,k}]$ in **Step SIM 1** has to be re-computed. I.e., for $\tilde{\mathcal{J}}_a \neq \mathcal{J}_a$ all additionally activated models have to be taken into account, where $\tilde{\mathcal{J}}_a$ is defined by

$$\tilde{\mathcal{J}}_a := \mathcal{J}_a \cup \left\{ j \mid \left(T_i^j([B_{a,k}], u([t_{k,k+1}])) = \text{true} \right) \cap (i \in \mathcal{J}_a) \right\} . \quad (22.27)$$

Then, the modified state equation $\tilde{f}_a(x(t), p, u(t), t)$ is determined such that

$$\mathbb{F}_{\tilde{f}_a} \supseteq \bigcup_{i \in \tilde{\mathcal{J}}_a} \mathbb{F}_{f_{S_i}} \quad (22.28)$$

with the exact value sets

$$\mathbb{F}_{\tilde{f}_a} := \{y \mid y = \tilde{f}_a(x, p, u, [t_{k,k+1}]) \ \forall x \in [B_{a,k}], p \in [p_k], u \in u([t_{k,k+1}])\} ,$$

$$\mathbb{F}_{f_{S_i}} := \{y \mid y = f_{S_i}(x, p, u, [t_{k,k+1}]) \ \forall x \in [B_{a,k}], p \in [p_k], u \in u([t_{k,k+1}])\} ,$$

and

$$[t_{k,k+1}] := [t_k ; t_{k+1}]$$

holds. In the case $\tilde{\mathcal{J}}_a = \mathcal{J}_a$, the evaluation is continued with **Step SIM 3**. Otherwise, the modification of f_a is continued with $\tilde{f}_a := \tilde{f}_a$ and $\mathcal{J}_a := \tilde{\mathcal{J}}_a$.

Step SIM 3 Interval enclosures of the state vector $[x_{k+1}]$ at t_{k+1} are computed by validated integration of $f_a(x(t), p, u(t), t)$ as defined in **Step SIM 2**. In Taylor series-based methods, the analytical expression for $f_a(x(t), p, u(t), t)$ is also used to compute the required Taylor coefficients, see [23].

Step SIM 4 All model states have to be deactivated which can no longer be active at t_{k+1} . Afterward, the simulation is continued with **Step SIM 1** for the next time interval $[t_{k+1,k+2}] := [t_{k+1} ; t_{k+2}]$.

22.7 Optimization Results

After an algorithm for validated integration of ODEs with state-dependent switchings has been presented in the previous section, results for optimal control of the dynamical system (22.19) with the control input $u(t) = F_a(t) \in [-1 ; 1]$ and the mass $m = 1.0$ as well as the interval parameters $F_s \in [0.015 ; 0.050]$ and $\mu \in [0.001 ; 0.010]$ are summarized in the following.

22.7.1 Interval Algorithm for Structure Optimization

22.7.1.1 Application of the Optimization Routine Without Parallelization

For structure optimization, a piecewise constant control strategy

$$u(t) =: u_{k-1} = \text{const} \quad \text{for } t \in \Delta T \cdot [(k-1) ; k] \quad (22.29)$$

is assumed, where $t_f = 5.0$, $\Delta T = \frac{t_f}{N}$, and $k = 1, \dots, N$ are given. The performance index is defined as

$$J = \underbrace{\int_0^{t_f} \left((x_1(t) - 1)^2 + x_2(t)^2 + u(t)^2 \right) dt}_{=: J_A} + 100 \cdot \underbrace{\frac{t_f}{50} \cdot \sum_{k=1}^{k_{\max}} (u_k - u_{k-1})^2}_{=: J_B} \quad (22.30)$$

with an integral term J_A weighting states and control variables over the complete time horizon and an additive term J_B to avoid unnecessarily large variations of the control variable between two subsequent points of time t_k and t_{k+1} . The optimization routine presented in this chapter is applied to three different problems which are depicted in Fig. 22.5.

First, the optimization is performed for the nominal system parameters \underline{F}_s and $\underline{\mu}$. The resulting time response for $x_1(t)$ and $x_2(t)$ is denoted by *case (A)*. Here, two different approximate solutions of the optimization problem are distinguished, namely a control sequence with $N = N_1 = 5$ and a control sequence with $N = N_2 = 50$ piecewise constant values of the input variable. The resulting interval enclosures of the best approximations of the optimal final states are

$$[x^{N_1}(t_f)] = \begin{bmatrix} [1.0104 ; 1.0137] \\ [0.0981 ; 0.0998] \end{bmatrix} \quad (22.31)$$

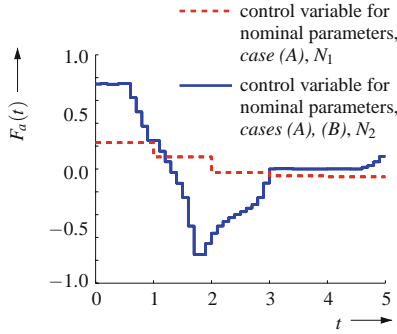
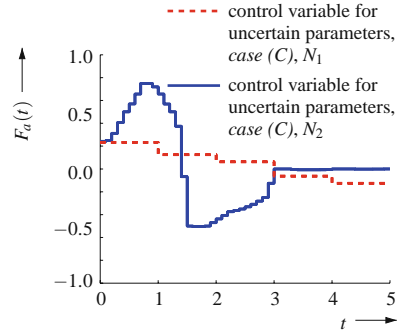
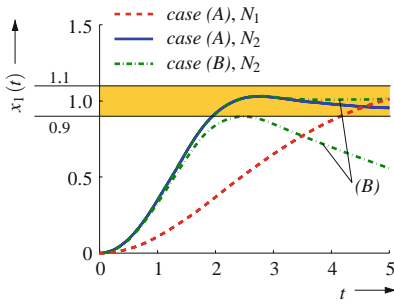
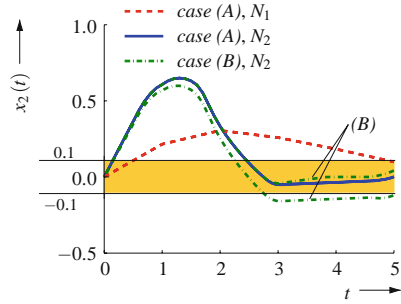
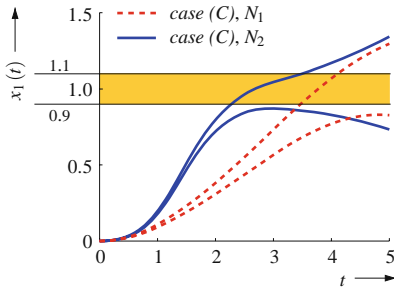
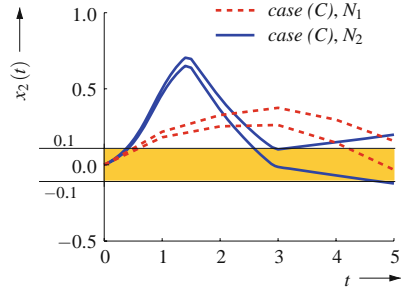
(a) Optimized accelerating force F_a .(b) Optimized accelerating force F_a .(c) Position x_1 , cases (A), (B).(d) Velocity x_2 , cases (A), (B).(e) Position x_1 , case (C).(f) Velocity x_2 , case (C).

Fig. 22.5 Control variable $F_a(t)$ of the positioning system with friction as well as interval enclosures of the state variables $x_1(t)$ and $x_2(t)$ in the cases (A), (B), and (C)

as well as

$$[x^{N_2}(t_f)] = \begin{bmatrix} [0.9531; 0.9544] \\ [0.0009; 0.0015] \end{bmatrix} \quad (22.32)$$

with the corresponding upper bounds

$$\bar{J}_{N_1} = 2.5567 \quad \text{and} \quad \bar{J}_{N_2} = 5.3705 \quad (22.33)$$

for the necessary costs.

Second, the resulting approximation for the optimal control sequence is applied to the uncertain system model leading to the time responses denoted by *case (B)*. The corresponding upper bound of the performance index is given by $\bar{J}_{N_2} \leq 5.5241$ if the control strategy for $N_2 = 50$ is considered.

Third, the optimization is performed directly for the uncertain parameters (*case (C)*). In this case, the resulting upper bounds of the performance index with $N_1 = 5$ and $N_2 = 50$ are

$$\bar{J}_{N_1} = 2.9696 \quad \text{and} \quad \bar{J}_{N_2} = 7.6130, \quad \text{respectively} \quad (22.34)$$

The resulting enclosures for the final state are

$$[x^{N_1}(t_f)] = \begin{bmatrix} [0.8265; 1.2957] \\ [-0.0319; 0.1572] \end{bmatrix} \quad (22.35)$$

and

$$[x^{N_2}(t_f)] = \begin{bmatrix} [0.7326; 1.3424] \\ [-0.1223; 0.1987] \end{bmatrix}, \quad \text{respectively} \quad (22.36)$$

Due to the direct consideration of the uncertainties in the optimization procedure, time responses for both state variables are obtained which are overlapping completely with the desired region of admissible final states given by $[x_1(t_f)] = [0.9; 1.1]$ and $[x_2(t_f)] = [-0.1; 0.1]$. However, in both *cases (B) and (C)*, the uncertainties of the final states are larger than the specified tolerances. In the following Subsections, linear state feedback controllers are introduced to reduce this influence of the uncertain system parameters.

In the *cases (A) and (C)*, the optimization has been performed with up to 20,000 evaluations of the set of state equations for the complete time horizon using a prototypical MATLAB implementation relying on the interval arithmetic toolbox INTLAB [27, 28]. To reduce the computing time and to prevent repeated interval splitting of a single control strategy, the state equations have been vectorized such that they are evaluated for up to $L = 20$ different control sequences simultaneously.

22.7.1.2 Application of the Optimization Routine with Parallelization

In order to demonstrate the use of parallelization techniques in the solution of optimal control problems using the suggested interval approach, two different tasks (see also Fig. 22.2) have been parameterized with $N_1 = 5$ and $N_2 = 50$ as well as $L_1 = L_2 = 20$. After 10 evaluations of each task, their results have been synchronized. In total, again 20,000 splittings of the range of the control interval (the same number of splittings as in the non-parallelized case) have been performed for each of the tasks. The corresponding result is

$$[x(t_f)] = \begin{bmatrix} [1.0285; 1.0288] \\ [0.0997; 0.0998] \end{bmatrix} \quad \text{with} \quad \bar{J} = 2.5533 . \quad (22.37)$$

After comparison with the result of the non-parallelized evaluation in *case (A)*, it can be seen that the parallelization improves the performance of the optimization algorithm by avoiding to get stuck in local minima. This problem, for example, occurs in the case $N_2 = 50$ in the non-parallelized search for the global optimum of the performance index.

Using the interval arithmetic optimization routine, a good approximation for the optimal solution of the control problem for nominal system parameters—which fulfills all given constraints for the state variables at the final point of time t_f —is obtained after 20,000 splittings of control intervals.

To determine an estimate for the computational effort that is necessary to compute solutions for the same optimization problem using an approach which is not based on the presented interval techniques, a control sequence with $N_2 = 50$ pre-defined switching points is considered. This task corresponds to a 50-dimensional global optimization problem. If a pure gridding of the range of the control variable is performed (similar to discrete dynamic programming) using M grid points for each of the control values to be determined, M^{50} evaluations of the state equations become necessary instead of the 20,000 evaluations in the interval arithmetic case. Therefore, even for $M = 2$ the resulting number of evaluations of the state equations is of the order 10^{15} . In practice, even much larger numbers of evaluations of the state equations will be necessary, since suitable approximations for the optimal control sequence can, in general, only be determined if also values from the interior of its admissible range are considered. The case $M = 2$ corresponds to a bang–bang control consisting of the two bounds of the range of u . A further increase of the computational effort is caused by parameter uncertainties which can only be taken into account in a non-interval-based extension of Bellman’s discrete dynamic programming if their range is also represented by grid points. Using the interval approach, the enclosures of the state variables and the corresponding cost function are always determined for the complete range of the considered control and parameter intervals.

22.7.2 Linear State Controller for Improvement of Robustness

Since it has not been possible in the previously discussed *cases (B) and (C)* to find one common control strategy for all parameter values such that the intervals of the final position and final velocity are completely included in $[x_1(t_f)] = [0.9; 1.1]$ and $[x_2(t_f)] = [-0.1; 0.1]$, an extension of the system by a closed-loop linear state controller is investigated. The feedback gain K of the linear state controller in Fig. 22.6 is determined by pole placement with $\lambda_1 = \lambda_2 = -1$ using Ackermann’s formula for the nominal system parameters \underline{F}_s and $\underline{\mu}$. For these parameters, $K = [1 \ 1.999]$ is obtained. The resulting accelerating force in Fig. 22.7 is defined by

$$u(t) = K \cdot (x_d(t) - x(t)) + u_{opt}(t) , \quad (22.38)$$

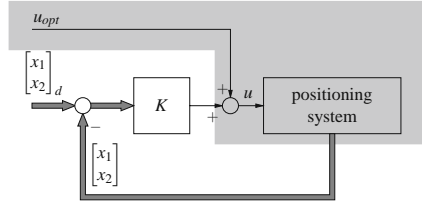


Fig. 22.6 Block diagram for extension of the dynamical system by a linear state feedback controller (feedback of the complete state vector x)

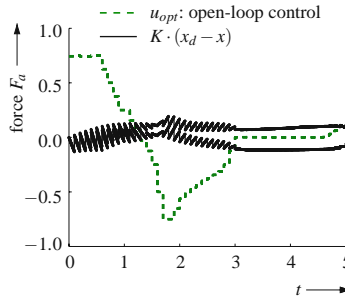


Fig. 22.7 Bounds for the accelerating force resulting from the linear state controller

where $u_{opt}(t)$ is the result of the structure optimization for nominal system parameters with $N_2 = 50$. The reference trajectory $x_d(t)$ is the corresponding time response of the nominal system. As it is shown in the Figs. 22.8(a) and (b), the modified control law now leads to trajectories which are completely included in the region of admissible final states. This has been proven by the guaranteed simulation approach summarized in Sect. 22.6.

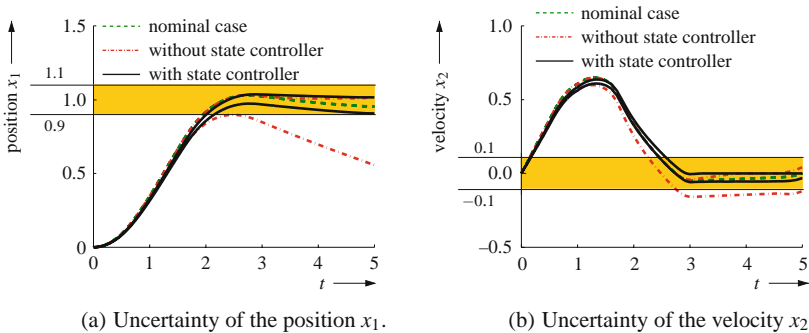


Fig. 22.8 Interval enclosures of the state variables x_1 and x_2 after extension of the system by a linear state controller

22.7.3 Interval Algorithm for Parameter Optimization

In the previous subsections, an interval arithmetic algorithm for the calculation of guaranteed enclosures of the states of dynamical systems with uncertainties and state-dependent switchings between various dynamical models has been applied for two different purposes. First, it has been applied in an interval-based routine to compute approximate solutions for the so-called structure optimization problem. In this case, optimal open-loop control sequences for both the nominal and the uncertain system models have been calculated. Second, the underlying validated integration routine has been used for robustness analysis by computation of guaranteed enclosures of the states of the uncertain mechanical positioning system which has been extended by a linear state controller. This controller has been parameterized using classical techniques for pole placement in order to improve the controlled system's robustness with respect to uncertain parameters.

In general, application of pole placement to nominal system models as well as linearization of nonlinear models is not always sufficient in order to meet specific requirements for the dynamics of closed-loop control systems. Therefore, the following example aims at illustrating how to apply the interval arithmetic optimization routines to the optimal parameterization of a state controller with a given structure. For that purpose, the gain vector $K = [k_1 \ k_2]$ of the linear state controller in Fig. 22.6 and Eq. (22.38) is determined by parameter optimization after interpretation of k_1 and k_2 as *time invariant*, i.e., *constant* control inputs. This corresponds to the setting $N = 1$ in the interval arithmetic optimization routine. Now, optimal constant values for the parameters k_1 and k_2 are determined such that the performance index (22.30) is minimized, where $u_{opt}(t)$ and $x_d(t)$ are given as in Sect. 22.7.2.

In Fig. 22.9, the resulting state enclosures are shown for the approximation

$$K \in \begin{bmatrix} [0.7031 ; 0.7444] \\ [0.2734 ; 0.3126] \end{bmatrix}^T \quad (22.39)$$

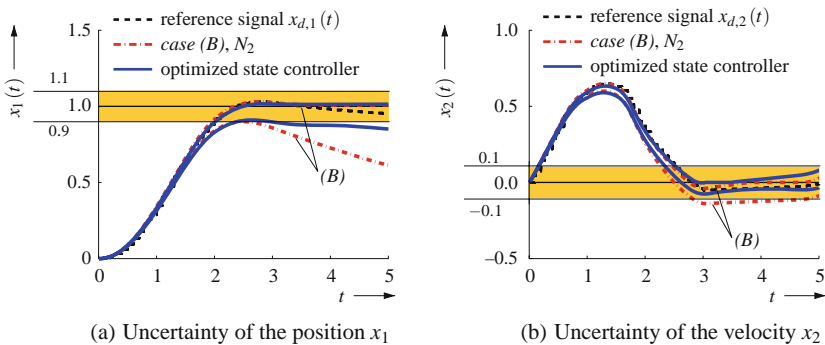


Fig. 22.9 Interval enclosures for the state variables x_1 and x_2 after parameterization of the linear state controller using parameter optimization techniques

of the optimal gain factors with the corresponding supremum

$$\bar{J} = 5.3755 \quad (22.40)$$

for the necessary costs. In contrast to the open-loop operation which corresponds to $K = [0 \ 0]$, the upper bound of the necessary costs is reduced from $\bar{J} = 5.5241$ in the previously discussed *case (B)* (and from $\bar{J} = 5.4930$ in the case of the linear state controller which has been parameterized using pole assignment for the nominal system model in the sliding friction mode in Sect. 22.7.2) by approximately 2.7% (and 2.1%, respectively). The enclosure

$$[x(t_f)] = \begin{bmatrix} [0.8500 ; 1.0139] \\ [-0.0344 ; 0.0808] \end{bmatrix} \quad (22.41)$$

of the corresponding final state shows a significant reduction of the uncertainty resulting from the parameter intervals $[F_s]$ and $[\mu]$. In this subsection, the previously specified tolerances for $x_1(t_f)$ and $x_2(t_f)$ have not been considered during calculation of the optimal values for K .

22.8 Conclusions and Outlook on Future Work

In this chapter, an interval arithmetic optimization algorithm which is applicable to both structure and parameter optimizations for discrete-time as well as continuous-time dynamical systems has been presented. It has been applied to a simplified model of a mechanical positioning system with state-dependent switching characteristics. For validated simulation of such systems, an extension of a Taylor series-based validated integration algorithm is used to detect the points of time at which transition conditions between the discrete model states are activated or at which one of these model states becomes invalid.

In future research, the optimization routine will be extended by solutions of continuous-time optimal control problems determined by application of Pontryagin's maximum principle to suitable approximations of the considered nonlinear dynamical systems [13, 24]. The approximation techniques which will be applied are based on Carleman linearization which approximates nonlinear systems by higher-dimensional models [7, 13, 24]. The goal of this extension is to speed up the optimization process and to reduce the computational burden in the case of high-dimensional real-world processes by improved guaranteed initial upper bounds for the maximum necessary costs. Furthermore, the optimization routine will be extended to the calculation of guaranteed bounds for continuous control laws with given bounded variation rates.

These extensions are a first step toward the development of a general interval arithmetic framework for the design and analysis of robust and optimal controllers for linear and nonlinear dynamical systems with uncertainties. On the one hand, improvements of approximate solutions can be achieved by suitable classical design

approaches. On the other hand, interval techniques allow for mathematically rigorous verification of the robustness of the controlled systems as well as for adaptation of controller structures and parameters if robustness specifications cannot be fulfilled using classical approaches.

References

1. E. Auer, A. Rauh, E. P. Hofer, and W. Luther. Validated Modeling of Mechanical Systems with SmartMOBILE: Improvement of Performance by ValEncIA-IVP. In *Proc. of Dagstuhl Seminar 06021: Reliable Implementation of Real Number Algorithms: Theory and Practice, Lecture Notes in Computer Science 5045*, Springer-Verlag, pages 1–27, 2008.
2. R. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.
3. R. Bellman, editor. *Mathematical Optimization Techniques*. University of California Press, Berkeley, California, 1963.
4. M. Berz and K. Makino. Verified Integration of ODEs and Flows Using Differential Algebraic Methods on High-order Taylor Models. *Reliable Computing*, 4:361–369, 1998.
5. A. A. Feldbaum. *Optimal Control Systems*. Academic Press, New York, 1965.
6. L. Jaulin, M. Kieffer, O. Didrit, and É. Walter. *Applied Interval Analysis*. Springer-Verlag, London, 2001.
7. K. Kowalski and W.-H. Steeb. *Nonlinear Dynamical Systems and Carleman Linearization*. World Scientific, Singapore, 1991.
8. G. Leitmann. *An Introduction to Optimal Control*. McGraw-Hill, New York, 1966.
9. Y. Lin and M. A. Stadtherr. Validated Solution of Initial Value Problems for ODEs with Interval Parameters. In *Proc. of the 2nd NSF Workshop on Reliable Engineering Computing*, Savannah GA, 2006.
10. Y. Lin and M. A. Stadtherr. Deterministic Global Optimization for Dynamic Systems Using Interval Analysis. In *CD-Proc. of the 12th GAMM-IMACS Intl. Symposium on Scientific Computing, Computer Arithmetic, and Validated Numerics SCAN 2006*, Duisburg, Germany, 2007. IEEE Computer Society.
11. Y. Lin and M. A. Stadtherr. Deterministic Global Optimization of Nonlinear Dynamic Systems. *AIChE Journal*, 53(4):866–875, 2007.
12. K. Makino. *Rigorous Analysis of Nonlinear Motion in Particle Accelerators*. PhD thesis, Michigan State University, 1998.
13. J. Minisini, A. Rauh, and E. P. Hofer. Carleman Linearization for Approximate Solutions of Nonlinear Control Problems: Part 1 – Theory. In *Proc. of the 14th Intl. Workshop on Dynamics and Control*, Moscow-Zvenigorod, Russia, 2007, *Advances in Mechanics: Dynamics and Control*, F. L. Chernousko, G. V. Kostin, V. V. Saurin (Eds), pages 215–222, Nauka, Moscow, 2008.
14. R. Moore. *Methods and Applications of Interval Analysis*. SIAM, Philadelphia, 1979.
15. N. S. Nedialkov. *Computing Rigorous Bounds on the Solution of an Initial Value Problem for an Ordinary Differential Equation*. PhD thesis, Graduate Department of Computer Science, University of Toronto, 1999.
16. N. S. Nedialkov. *The Design and Implementation of an Object-Oriented Validated ODE Solver*. Kluwer Academic Publishers, Dordrecht, 2002.
17. N. S. Nedialkov. Interval Tools for ODEs and DAEs. In *CD-Proc. of the 12th GAMM-IMACS Intl. Symposium on Scientific Computing, Computer Arithmetic, and Validated Numerics SCAN 2006*, Duisburg, Germany, 2007. IEEE Computer Society.
18. N. S. Nedialkov and M. v. Mohrenschildt. Rigorous Simulation of Hybrid Dynamic Systems with Symbolic and Interval Methods. In *Proc. of the American Control Conference ACC*, pages 140–147, Anchorage, USA, 2002.

19. L. S. Pontryagin, V. G. Boltjanskij, R. V. Gamkrelidze, and E. F. Misčenko. *The Mathematical Theory of Optimal Processes*. Interscience Publishers, New York, 1962.
20. A. Rauh, E. Auer, and E. P. Hofer. VALENCIA-IVP: A Comparison with Other Initial Value Problem Solvers. In *CD-Proc. of the 12th GAMM-IMACS Intl. Symposium on Scientific Computing, Computer Arithmetic, and Validated Numerics SCAN 2006*, Duisburg, Germany, 2007. IEEE Computer Society.
21. A. Rauh and E. P. Hofer. Interval Arithmetic Optimization Techniques for Uncertain Discrete-Time Systems. In E. P. Hofer and E. Reithmeier, editors, *Proc. of the 13th Intl. Workshop on Dynamics and Control, Modeling and Control of Autonomous Decision Support Based Systems*, pages 141–148, Wiesensteig, Germany, 2005. Shaker Verlag, Aachen.
22. A. Rauh, M. Kletting, H. Aschemann, and E. P. Hofer. Robust Controller Design for Bounded State and Control Variables and Uncertain Parameters Using Interval Methods. In *Proc. of the Intl. Conference on Control and Automation ICCA'05*, pages 777–782, Budapest, Hungary, 2005.
23. A. Rauh, M. Kletting, H. Aschemann, and E. P. Hofer. Interval Methods for Simulation of Dynamical Systems with State-Dependent Switching Characteristics. In *Proc. of the IEEE Intl. Conference on Control Applications CCA 2006*, pages 355–360, Munich, Germany, 2006.
24. A. Rauh, J. Minisini, and E. P. Hofer. Carleman Linearization for Approximate Solutions of Nonlinear Control Problems: Part 2 – Applications. In *Proc. of the 14th Intl. Workshop on Dynamics and Control, Moscow-Zvenigorod, Russia, 2007, Advances in Mechanics: Dynamics and Control*, F. L. Chernousko, G. V. Kostin, V. V. Saurin (Eds), pages 266–273, Nauka, Moscow, 2008.
25. A. Rauh, J. Minisini, and E. P. Hofer. Interval Techniques for Design of Optimal and Robust Control Strategies. In *CD-Proc. of the 12th GAMM-IMACS Intl. Symposium on Scientific Computing, Computer Arithmetic, and Validated Numerics SCAN 2006*, Duisburg, Germany, 2007. IEEE Computer Society.
26. R. Rihm. Über Einschließungsverfahren für gewöhnliche Anfangswertprobleme und ihre Anwendung auf Differentialgleichungen mit unstetiger rechter Seite (in German). PhD thesis, University of Karlsruhe, Germany, 1993.
27. S. M. Rump. IntLab (Version 5.4). <http://www.ti3.tu-harburg.de/~rump/intlab/>.
28. S. M. Rump. INTLAB — INTerval LABoratory. In T. Csendes, editor, *Developments in Reliable Computing*, pages 77–104. Kluwer Academic Publishers, 1999.
29. The MathWorks, Inc. MATLAB Distributed Computing Toolbox — User's Guide. <http://www.mathworks.com>.

Chapter 23

Application of Optimisation Algorithms to Aircraft Aerodynamics

Emanuele Rizzo and Aldo Frediani

Abstract Reduced operating costs and low environmental impact are required for modern commercial aircraft. The today airplanes are so optimized that a significant increase in performances can hardly be achieved and new configurations are now of interest, supposing to allow a jump forward in drag reduction. In the present work, optimization algorithms are applied to search new aircraft configurations able to satisfy different constraints. The algorithms are first applied to benchmarking problems in order to test their performances and evaluate the computational time. The same algorithms are applied to classical problems of aerodynamics like the optimum wings. At the end, the problems of trim and stability of flight are tackled; we look for wing planforms which satisfy a set of constraints defining the feasible region both in cruise condition and in low-speed condition (i.e. when high lift devices are deployed). Finally, a new ultralight optimised aircraft is presented as an example of application.

23.1 Introduction

After the introduction into service of turbofan engine and pressurised cabin, only small changes in aircraft configurations for civil transport can be observed up to-day (an example is emphasized in Fig. 23.1, where, after an appropriate scaling, the planforms of two commercial aircraft separated by 40 years of evolution are superimposed).

Emanuele Rizzo

Dipartimento di Ingegneria Aerospaziale, Università di Pisa, via Caruso, 56122 Pisa, Italy,
e-mail: emanuele.rizzo@ing.unipi.it

Aldo Frediani

Dipartimento di Ingegneria Aerospaziale, Università di Pisa, via Caruso, 56122 Pisa, Italy,
e-mail: a.frediani@ing.unipi.it

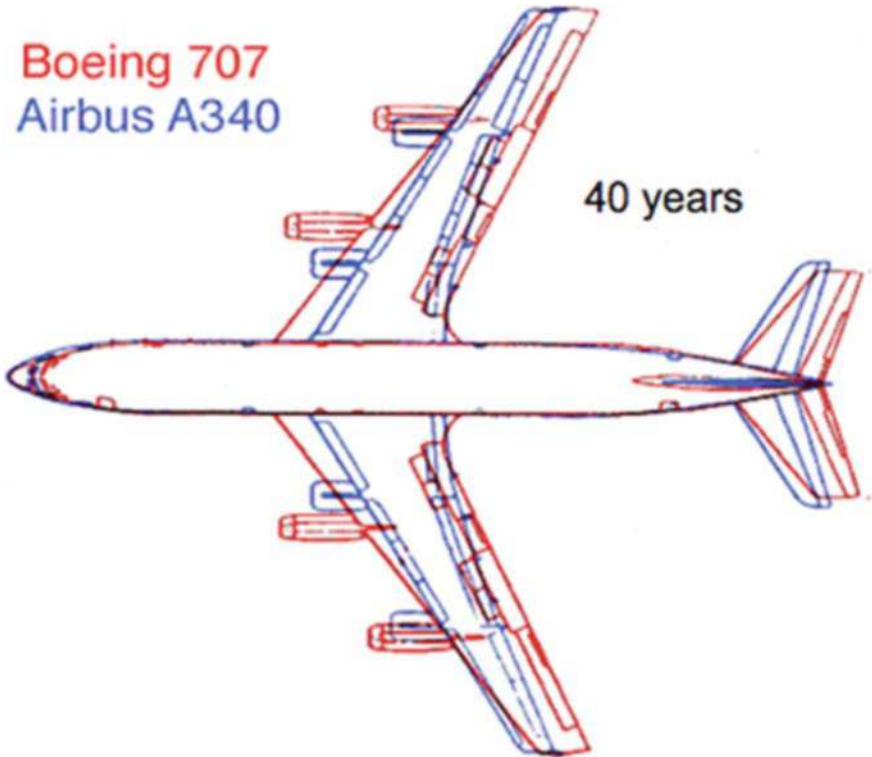


Fig. 23.1 Forty years of evolution in the aircraft geometry: from Boeing 707 to Airbus A340 (from [11])

The Airbus A380 with its 79.8 m of wingspan and 73 m of length is the largest commercial aircraft in the world (Fig. 23.2), but it also represents the upper technological limit of the conventional aircraft configuration (fuselage, wing, tail).

At the beginning of this century, the sustainable growth and the greening of air transport have become the two main targets that European Union has set to be reached within the 2020 [2]. In addition to that, airline companies operate in a high-competitive market, where low costs together with high safety standards decree their commercial success. Within this frame, the aeronautic world is facing a new scenario, where new solutions in transport aviation must be found.

The most important proposal for future aircraft is the Blended Wing Body (BWB) (Fig. 23.3, [3]), the C-Wing (Fig. 23.4, [4]) from USA, and the PrandtlPlane from Italy (Fig. 23.5, [5]).

Improvements in aerodynamics and/or structures are theoretically possible with all the new proposals. BWB and C-Wing have open wings, similar to the conventional ones; in the PrandtlPlane proposal, the wing system is a closed box-type. This last represents the engineering application of the Best Wing System concept introduced by Prandtl in 1924 [6], [23], [24], that is the system that minimises the induced drag ([5, 7, 8]). The PrandtlPlane configuration is the most critical as far



Fig. 23.2 Airbus A380: the largest commercial airplane in the world



Fig. 23.3 The Blended Wing-Body (BW) concept

as the aerodynamic interference between the wings is concerned, and, moreover, there is a strict connection between aerodynamic efficiency and static stability of flight. The control surfaces of a PrandtlPlane are unconventional; e.g. pitch control is obtained by two elevators at the front and rear wing roots, moving in opposition of phase so that they generate a pure pitch moment. The optimization of a PrandtlPlane is the most general problem, because any other configuration can be obtained as a particular case from it.

In the present work, we model a PrandtlPlane configuration with a certain number of parameters and, by means of an optimisation procedure, we seek the optimum combination minimizing the drag and satisfying the constraints on both trim and stability of flight.

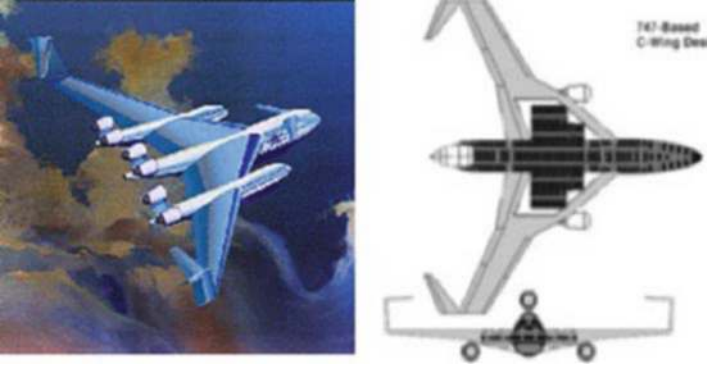


Fig. 23.4 The C-Wing concept



Fig. 23.5 The PrandtlPlane concept

For the reader's convenience, we recall some results in the theory of optimization useful to better identify which tools are used in the present work.

In this chapter, the *objective function* $f(x)$ is a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$; discrete parameters as the number of engines or the number of passengers are considered as fixed. The objective represents the drag of the aircraft and it is always positive.

The *feasible set* Ω is the set where the vector of the variables x runs; in this chapter it is defined by means of a certain number of equalities and inequalities (say, m and p) such that

$$\Omega = \{x \in \mathbb{R}^n : g_i(x) \leq 0, \quad i = 1, \dots, p, \quad h_j(x) = 0, \quad j = 1, \dots, m\}.$$

The model of optimization problem used in this chapter to find the wing shape which minimizes the drag is

$$\begin{cases} \min f(x) \\ g(x) \leq 0 \\ h(x) = 0 \\ x \in \mathbb{R}^n \end{cases} \quad (23.1)$$

In the most general case, functions f , g and h are non-linear, giving rise to a Non Linear Programming (NLP) problem; moreover, we require that they are at least continuously differentiable in \mathbb{R}^n .

When $\Omega = \mathbb{R}$, (23.1) becomes an *unconstrained problem*. For this class of problems the following lemmas state necessary and sufficient conditions for the minimum.

Lemma 23.1 (Necessary Optimality Condition (NOC)). *For unconstrained problem: x^* is a stationary point if $\nabla f(x^*) = 0$.*

Lemma 23.2 (Second-Order Sufficient Condition (SOSC)). *For unconstrained problem: x^* is a minimum point if the Hessian matrix at x^* , $\nabla^2 f(x^*)$, is positive definite.*

When $\Omega \subset \mathbb{R}$, the problem is a *constrained problem* and Theorem 23.1 states necessary optimality conditions (NOC). Second-order sufficient conditions (SOSC) may be also defined (see [9, 10] and [11] for more details).

Theorem 23.1 (Karush–Khun–Tucker (KKT)). *Let the Lagrange function be*

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^p \lambda_i g_i(x) + \sum_{j=1}^m \mu_j h_j(x),$$

and the constraint qualification¹ be verified in x^ , then the point x^* is a solution of problem 23.1 if exist μ^* and $\lambda^* \geq 0$ satisfying the following system:*

$$\begin{cases} \nabla_x L(x^*, \lambda^*, \mu^*) = 0 \\ \sum_{i=1}^p \lambda_i g_i(x^*) = 0 \\ h_j(x^*) = 0 \end{cases} \quad j = 1, \dots, m \quad (23.2)$$

There are different algorithms solving problem (23.1) on the basis of the NOC and, often, on SOSC conditions but, in general, they seek only local solutions. The search for global solutions is an important task for engineering purposes mainly for the following reasons:

¹ The constraint qualification is a regularity property of constraints at x^* ; see [9, 11] for more details.

- (i) The objective function and the constraints are “black boxes”.
- (ii) Numerical issues associated to the evaluation of objective and constraints introduce numerical noise.
- (iii) When applied to the study of new projects, a good exploration property of algorithms is mandatory.

Point (i) means that explicit expressions for objective and constraints are not available; the value of these functions at point x is given by a numerical code (the so called *oracle* or *black box*) and, therefore, general properties of the functions are not known a priori.

Point (ii) is related to the previous point 23.1, where the numerical evaluations of function values are often approximated and numerical errors are superimposed and the function may be affected by high-frequency oscillations with local peaks and valleys (local minima).

Point (iii) deals with the effectiveness of the algorithm. New projects and innovative designs are often defined in an almost unknown domain, and the starting point may be very far from the minimum.

In the present chapter, an algorithm for the search of global minima ([12] and [13]) has been implemented. This algorithm uses local solvers; here two types have been tested: the well-known Sequential Quadratic Programming (SQP) algorithm and the Mesh Adaptive Direct Search (MADS) algorithm ([14] and [15]). Both solvers are applied to benchmarking problems, based on the results obtained, and one has been applied to the design of a PrandtlPlane ULM airplane. For this configuration, the lift, induced drag and pitching moment are calculated by a numerical solver based on the Vortex Lattice Method [16], whilst the viscous drag is simply estimated by means the flat plate analogy. The final result of the optimization algorithm is the final aircraft configuration. A flying scaled model is under construction.

23.2 An Algorithm for the Search of Global Minima

Problem (23.1) is a NLP problem, that is a minimization problem with at least one non linear function. It is well known that algorithms for solving constrained problems derive directly from algorithms for unconstrained problems. In this framework, two main categories for optimization algorithms are identified:

- gradient-based methods;
- direct search methods.

Methods belonging to the first category are very popular and they have been applied in many different fields. Among these Newton, quasi-Newton and conjugate gradient methods are the most widely used. Theoretical background for these methods is well defined and further details may be found in the literature [9–11].

Methods belonging to the second category are relatively new and they are conceived to avoid the calculation of derivatives. The most ancient and famous direct search method is the method of simplex, even though mathematical theory is

developing towards the family of Global Pattern Search (GPS) algorithms [16]. The Mesh Adaptive Direct Search (MADS) Algorithm [18] belongs to this family and it is a probabilistic algorithm.

Both classes seek only local minima of unconstrained problems. Constrained problems may be faced by means the use of *penalty functions* or by a Lagrangian duality formulation like in the Sequential Quadratic Programming (SQP) method. The first approach is more suitable for direct methods [15], whilst the second approach is still a grad-based method.

An example of penalty function is the *Augmented Lagrangian penalty function*. In this method, both the minimum point x^* and an estimate of the Lagrange multipliers (μ^*, λ^*) are sought. An augmented Lagrange function is defined as follows:

$$L_a[x, \lambda, \mu, \varepsilon] = f(x) + \lambda \max \left\{ g(x), -\frac{\varepsilon}{2} \lambda \right\} + \mu h(x) + \frac{1}{\varepsilon} \left\| \max \left\{ g(x), -\frac{\varepsilon}{2} \lambda \right\} \right\|^2 + \frac{1}{\varepsilon} \|h(x)\|^2 \quad (23.3)$$

The algorithm consists in the following steps:

STEP1 Initialization: given $\varepsilon_0, \mu_0, \lambda_0, x^0 \in \mathbb{R}^n$ and set $k = 0$;

STEP2 Set $x^s = x^0$ and find an unconstrained local minimum x^k starting from x^s of

$$\min_{x \in \mathbb{R}^n} L_a(x, \mu_k, \lambda_k, \varepsilon_k)$$

STEP3 If (x^k, μ_k, λ_k) is a KKT point, STOP
otherwise:

STEP4 Update the penalty parameter $\varepsilon_{k+1} \in (0, \varepsilon_k]$

STEP5 Update the multiplier estimates to μ_{k+1} and λ_{k+1}

STEP6 Set $x^s = x^k, k = k + 1$ and go to STEP2

Rules on how to update the penalty parameter and the Lagrange multiplier estimates may be found in [9, 11] and [14] when there are bound box constraints. In [15] the method is specialized to the GPS algorithm, this technique has been investigated in the present work in a constrained version of MADS algorithm.

In order to extend local minimum methods for finding global minima it is possible to proceed as follows. This algorithm, taken from [12] and [13], can be viewed as a multistart algorithm; its main idea is to take information from the local minima in order to build a function giving a direction of the basin of attraction. The algorithm is based on the following assumptions and hypotheses:

- (i) The objective function has a structure so that it can be viewed as a superimposition of “noise” to the underlying function (*funnel structure*, see Fig. 23.6).
- (ii) The function $L(x)$, “minima found by a local optimizer starting from different points”, is defined; it is a step function (Fig. 23.7).
- (iii) Given the current point x^k , K points are uniformly sampled inside a ball with centre in x^k and radius r , producing K observations of the function $L(x)$ (Fig. 23.8). A rule how to choose r is presented later on.

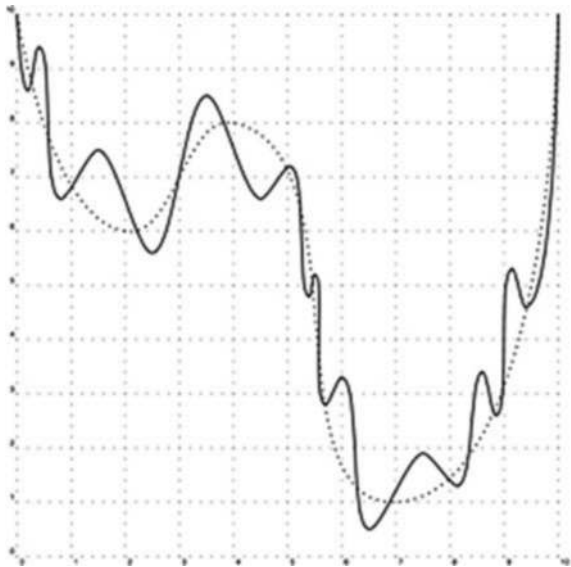


Fig. 23.6 Function having a funnel structure [12]

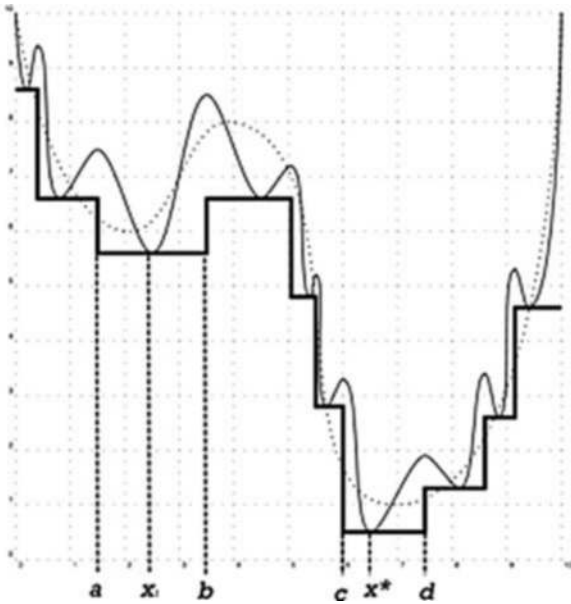


Fig. 23.7 The step function $L(x)$ [12]

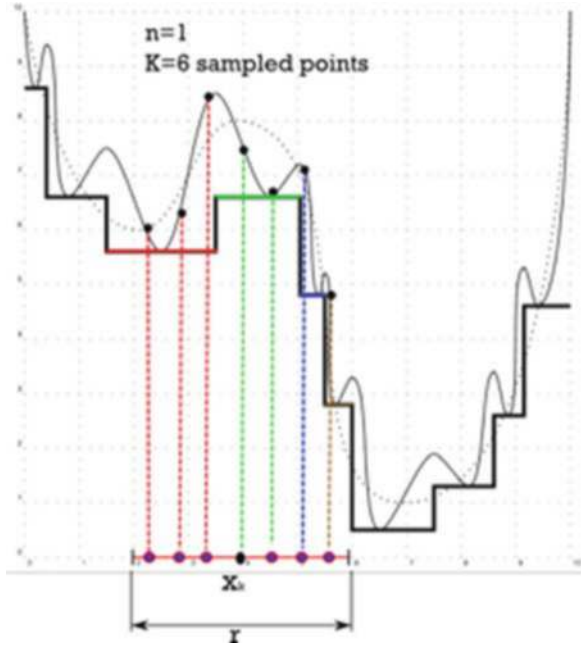


Fig. 23.8 Sampling points around the current point x^k [12]

(iv) The smoothed function $\widehat{L}_g^B(x)$ is built as

$$\widehat{L}_g^B(x) = \frac{\sum_{i=1}^K L(y_i) g(\|y_i - x\|)}{\sum_{i=1}^K g(\|y_i - x\|)}$$

where $g(z) = e^{-z^2/(2\sigma)^2}$ is the Gaussian operator and $\sigma = \frac{r}{\sqrt[3]{K}}$ is the standard deviation.

- (v) The smoothed function is minimised and its minimum gives directions on the basin of attraction (Fig. 23.9).
- (vi) The true function is evaluated at this point; if its value is lower than before, x^k moves at this point and go to 3, otherwise there is no improvement in the objective and the algorithm terminates.

In order to improve the exploration capabilities of the algorithm, instead of terminate the search as soon as no improvement is observed, a new sampling may be repeated; the “number of no improvements” is a user-defined parameter.

The radius of the ball r is another user-defined parameter. It defines how large is the sampling area of starting points of local solvers. In the present chapter it has been chosen as $r = \frac{ub-lb}{2}$, where ub and lb are the upper and low boundaries, respectively.

This algorithm showed great performances in finding global solutions as it will be shown in the following tests cases.

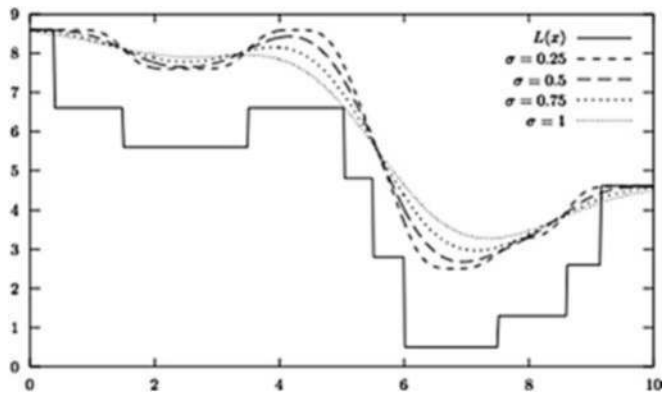


Fig. 23.9 Example of smoothing of the step function [12]

23.3 Test Cases

In order to check the performances of the algorithms, they have been applied to some benchmarking problems taken from the literature. The first class of problems includes very noisy objective functions. The second class is constituted by NLP constrained problems.

23.3.1 Test Case 1 (Unconstrained): Ackley’s Function

A classical benchmarking function to test optimization algorithms is the Ackley’s function [19]. A 3D plot of the Ackley’s function is shown in Fig. 23.11. The function presents a deep basin of attraction but it is very noisy in a neighbor of the minimum and high-frequency waves are superimposed everywhere.

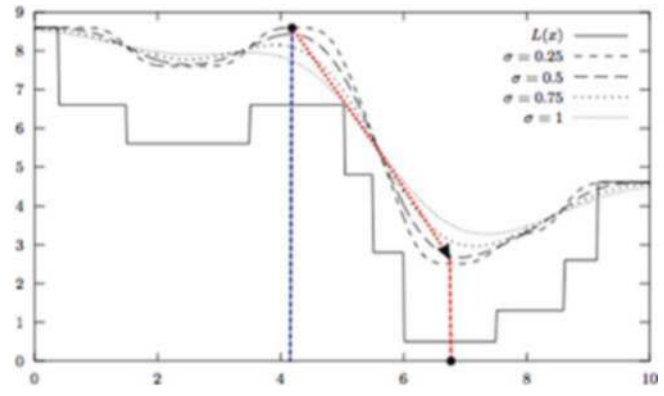


Fig. 23.10 Minimization of the smoothed function [12]

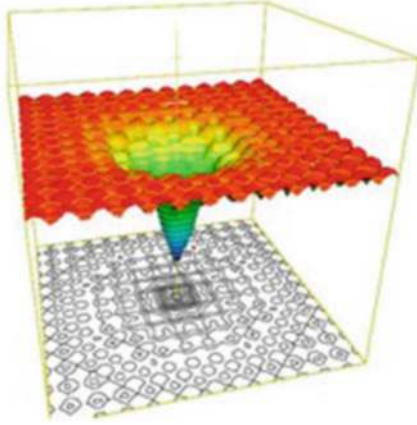


Fig. 23.11 3D plot of the Ackley's function

$$f(x) = -20e^{-0.2\sqrt{\frac{1}{n}\sum_{i=1}^n x_i^2}} - e^{\frac{1}{n}\sum_{i=1}^n \cos 2\pi x_i} + 20 + e \quad (23.4)$$

Domain: $-32.768 \leq x \leq 32.768$.

Point of global minimum: $x^* = (0 \ 0 \ \dots \ 0)$.

Global minimum: $f(x^*) = 0$.

In Tables 23.1 and 23.2, the minima found with the two algorithms are summarised. It can be observed that the global solver LOCSMOOTH with local solver SQP solver gives results far from the exact solution ($f(x^*) = 0$), whereas the solution is closer to the optimum when MADS is used as local solver. This behavior could be due to the highly noised nature of the function.

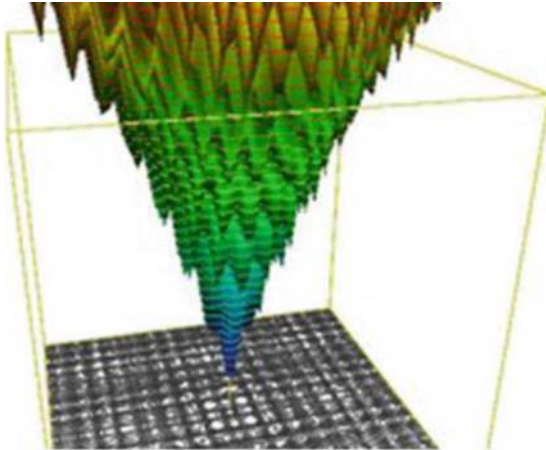


Fig. 23.12 A zoom near the origin of the Ackley's function

Table 23.1 Numerical solutions of the Ackley’s function: SQP+LocSmooth Solver

n	$f(x^*)$	NFval
2	10.1203	824
5	5.9783	7422
10	15.0208	15,392
15	13.8393	26,243
20	19.8439	10,395
25	19.9737	16,066
30	20.2286	8974

Table 23.2 Numerical solutions of the Ackley’s function: MADS+LocSmooth Solver

n	$f(x^*)$	NFval
2	3.03×10^{-6}	10,837
5	9.1×10^{-6}	22,013
10	0.1395×10^{-4}	95,319
15	0.1577×10^{-4}	190,992
20	0.1559×10^{-4}	554,557
25	0.1479×10^{-4}	390,157
30	0.1536×10^{-4}	1,000,328

Worth of notice is the high number of function evaluations (*NFval*) generated by the MADS solver; this is the main drawback of this method, as it is shown in the Fig. 23.13, where the angle formed by the search direction and the $-\nabla f$ direction is plotted versus the number of variables. The solutions found by both the methods deteriorate as the space dimensions grow.

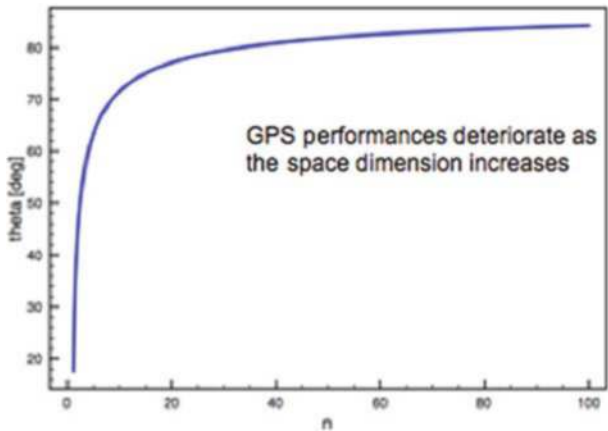


Fig. 23.13 Performances of GPS methods as the number of variables increases

23.3.2 Test Case 2 (Unconstrained): Rastrigin's Function

This is a typical example of non-linear multimodal function. It was first proposed by Rastrigin [20] as a 2D function and has been generalized by Mühlenbein et al. in [21]. This function presents a large search space and a large number of local minima. The typical funnel structure is present (see Fig. 23.14), and in this case, the LOCSMOOTH technique is expected to be effective. According to the conclusions from test case 1, the gradient-based solver (SQP) gives bad results due to the presence of high-frequency noise in the objective (see Table 23.2), whereas better results are obtained when the MADS solver is used as local solver. Even in the present case the quality of the solution deteriorates as the space dimensions grow.

$$f(x) = 10n + \sum_{i=1}^n x_i^2 - 10 \cos 2\pi x_i \quad (23.5)$$

Domain: $-5.12 \leq x \leq 5.12$.

Point of global minimum: $x^* = (0 \ 0 \ \dots \ 0)$.

Global minimum: $f(x^*) = 0$.

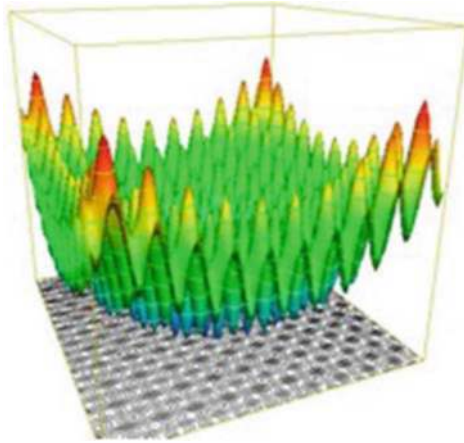


Fig. 23.14 3D plot of the Rastrigin's function

23.3.3 Test Case 3 (Unconstrained): Rosenbrock's Function

Rosenbrock's valley is a classic optimization problem, also known as "banana function". The global optimum is inside a long, narrow, parabolic-shaped flat valley. To seek the basin of attraction is trivial; more difficult is the convergence to the global optimum and, hence, this problem has been repeatedly used to assess the performance of optimization algorithms. A 3D plot of the Rosenbrock's function is shown in Fig. 23.15.

Table 23.3 Numerical solutions of the Rastrigin’s function: SQP+LocSmooth Solver

n	$f(x^*)$	NFval
2	0.995	437
5	4.9798	1690
10	9.9496	7038
15	21.889	6553
20	20.8940	10,363
25	0.995	14,181
30	18.9042	18,614

$$f(x) = \sum_{i=1}^{n-1} 100(x_{i+1} - x_i)^2 + (1 - x_i)^2$$

(23.6)

Domain: $-2.48 \leq x \leq 2.48$.
Point of global minimum: $x^* = (1 \ 1 \ \dots \ 1)$.
Global minimum: $f(x^*) = 0$.

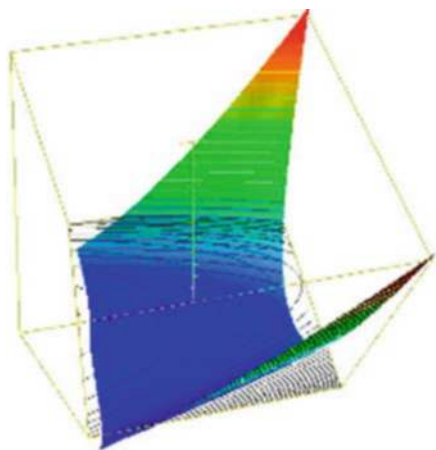


Fig. 23.15 3D plot of the Rosenbrock’s function

In this problem, the SQP local solver gives better results than MADS. Again, MADS generated a huge number of function evaluations compared to SQP (about 10,000 times greater).

23.3.4 Test Case 4 (Unconstrained): Schwefel’s Function

Schwefel’s function [22] is deceptive in that the global minimum is geometrically distant from the next best local minima. Therefore, the search algorithms are potentially prone to convergence in the wrong direction.

Table 23.4 Numerical solutions of the Rastrigin's function: MADS+LocSmooth Solver (best results over 10 runs)

n	$f(x^*)$	NFval
2	2.11×10^{-10}	3137
5	6.35×10^{-9}	24,986
10	0.244×10^{-7}	37,806
15	0.889×10^{-7}	89,374
20	0.685×10^{-7}	257,906
25	0.833×10^{-7}	432,970
30	0.114×10^{-6}	820,628

Table 23.5 Numerical solutions of the Rosenbrock's function: SQP+LocSmooth Solver

n	$f(x^*)$	NFval
2	6.14×10^{-11}	3968
5	1.7×10^{-10}	9416
10	0.996×10^{-9}	22,419
15	0.6965×10^{-9}	39,776
20	0.159×10^{-9}	45,309
25	0.6145×10^{-8}	34,251
30	0.2937×10^{-8}	119,417

Table 23.6 Numerical solutions of the Rosenbrock's function: MADS+LocSmooth Solver (best results over 10 runs)

n	$f(x^*)$	NFval
2	8.97×10^{-10}	58,916
5	0.0866×10^{-3}	1,243,050
10	0.1059×10^{-3}	3,345,535
15	0.0933×10^{-3}	980,867
20	0.1644×10^{-3}	6,758,729
25	0.1955×10^{-3}	1,1527,938
30	0.0202×10^{-3}	13,462,135

$$f(x) = \sum_{i=1}^n -x_i \sin \sqrt{\|x_i\|} \quad (23.7)$$

Domain: $-500 \leq x \leq 500$.

Point of global minimum: $x^* = 420.9687 (1 \ 1 \ \dots \ 1)$.

Global minimum: $f(x^*) = 418.9829 \cdot n$ (see Table 23.7).

Numerical results of the optimization are summarized in Tables 23.8 and 23.9, in the case of SQP and MADS local solvers, respectively, and with three sampled points for every ball. Worth of notice is that both techniques fail in finding the

Table 23.7 Optimum values for Schwefel’s function

n	2	5	10	15	20	25	30
$f(x^*)$	-837.9658	-2094.914	-4189.83	-6284.744	-8379.66	-10474.57	-12569.5

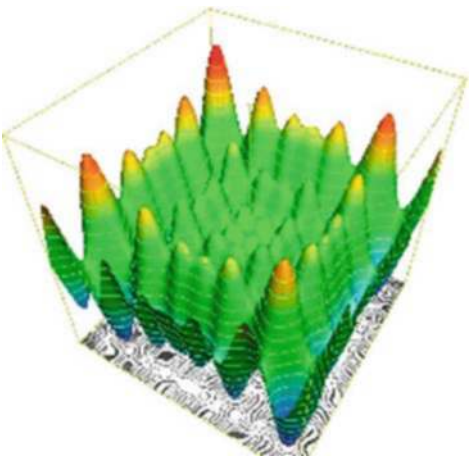


Fig. 23.16 3D plot of the Schwefel’s function

solution for $n > 2$. Moreover, they show similar results but MADs accounts once again a larger number of function evaluations.

23.4 The AEROSTATE Program: An Application to Aeronautics

The previous algorithms, in particular the global search algorithm LocSmooth, are now applied to seek the minimum of several problem relevant to aerodynamics. The first problem is the minimization of the induced drag of a wing; the problem is then

Table 23.8 Numerical results of the Schwefel’s function, number of sampled points $K = 3$. SQP local solver

n	$f(x^*)$	NFval
2	-837.9658	405
5	-1.3432×10^3	1723
10	-3.2378×10^3	2814
15	-4.6843×10^3	12,304
20	-5.7102×10^3	11,983
25	-6.9138×10^3	33,268
30	-8.5547×10^3	51,044

Table 23.9 Numerical results of the Schwefel's function, number of sampled points $K = 3$. MADs local solver (best results over 10 runs)

n	$f(x^*)$	NFval
2	-837.9658	5042
5	-1.6193×10^3	38,081
10	-3.1178×10^3	38,320
15	-4.8787×10^3	139,909
20	-7.194×10^3	531,061
25	-10.276×10^3	111,862
30	-8.939×10^3	338,409

generalised by adding the viscous drag. Another class of aerodynamic problems concerns non-planar wing systems, in particular biplane, wing-box and whole aircraft configurations (wing plus tail), with constraints on geometry, static stability of flight (minimum and maximum allowable static margin of stability) and trim. This latter application led to the development of the **AEROSTATE** (AERodynamic Optimization with STatic stability and Trim Evaluator) program; a program capable of finding the optimum wing planforms of whatever configuration under geometrical and aerodynamic constraints both in cruise and in TO/LAN conditions.

23.4.1 Minimum Induced Drag of a Wing

A classical problem of aerodynamics is to find the lift distribution over an isolated wing giving the minimum induced drag. As it is well known, the optimum lift distribution is an ellipse. Here the same problem is simulated and solved by AEROSTATE.

In order to set the optimisation problem, the wing span is divided into a certain number of trapezoidal trunks (bays) and, for each one, 12 geometric variables are defined (Figs. 23.17 and 23.18), namely

- coordinates of the point at leading edge of the root chord (x_0, y_0, z_0) ;
- root and tip chord lengths;
- span;
- dihedral angle;
- swept angle;
- root and tip twist angles;
- root and tip airfoils.

When flaps and slats are present in the bay, other six variables are added:

- flap chord to bay chord ratio c_f/c ;
- flap angle of rotation δ_f ;
- flap axis of rotation;

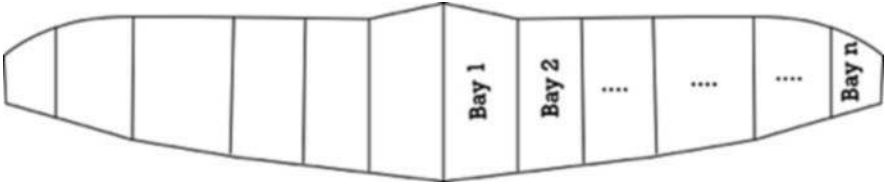


Fig. 23.17 Geometric approximation of a wing

- slat chord to bay chord ratio c_s/c ;
- slat angle of rotation δ_s ;
- slat axis of rotation.

and the total number of variables for each bay becomes 18.

The problem can be formulated as

$$\begin{cases} \min D_{ind}(x) \\ L(x) = W \\ lb \leq x \leq ub \end{cases} \quad (23.8)$$

where D_{ind} is the induced drag, $L(x)$ is the lift, W is the weight, lb is the vector of lower boundaries and ub the vector of upper boundaries. The variable x is the vector of the design variable. For this problem, the variables are the chord lengths at six equally spaced sections over the wing and the angle of attack α , for a total of seven

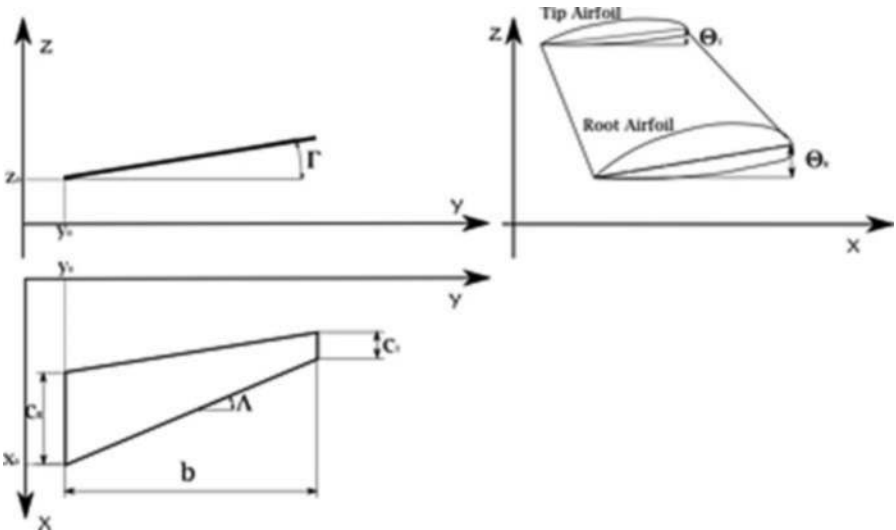


Fig. 23.18 Geometric approximation of a wing

Table 23.10 Starting data for the isolated wing

α	C_L	C_{Di}	e	S
1 deg	0.2718	0.0021	0.9319	12 m ²

variables. The weight is fixed in 5,000 N and the chords may vary between 0 and 3 m, whereas the angle of attack varies between -5° and 5° .

The starting geometry together with starting Trefftz plane and starting aerodynamics data are presented in Table 23.10 and Figs. 23.19, and 23.20.

Results of the optimization are reported in Table 23.11, Figs. 23.21 and 23.22. Worth of notice is that the optimized lift distribution is close to an ellipse as well as the wing planform.

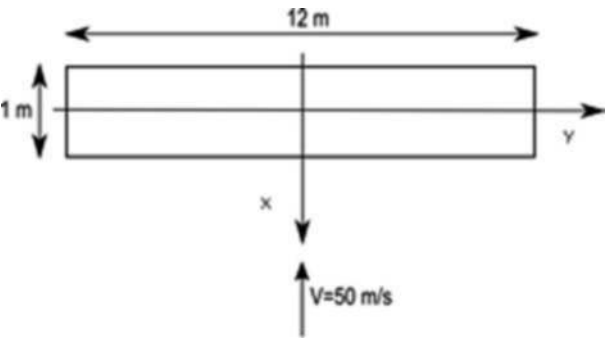


Fig. 23.19 Starting geometry of a wing

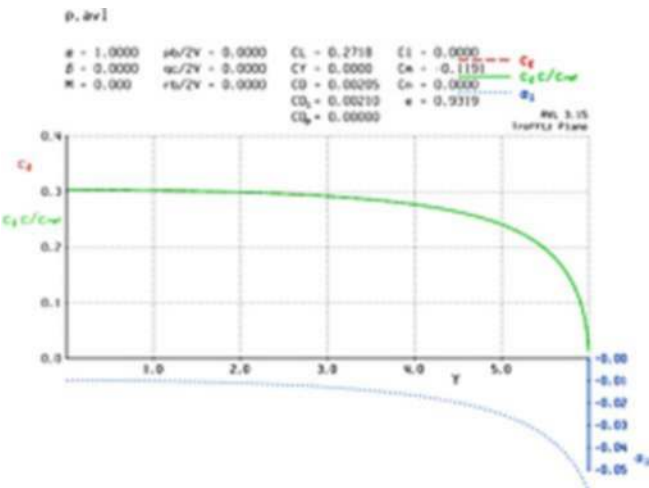


Fig. 23.20 Lift distribution calculated in the Trefftz plane for the isolated wing

Table 23.12 Optimized data for the wing total drag problem

	<i>SQP</i> LocSmooth	<i>MADS</i> LocSmooth
$S \text{ m}^2$	6.67	6.57
AR	21.6	21.9
C_L	0.49332	0.4966
C_{Di}	0.00363	0.0036
C_{Diteo}	0.00358	0.00358
C_{Dvisc}	0.0063858	0.0063871
$\alpha \text{ deg}$	5	5
e	0.9854	0.9952
$D = D_i + D_v$	101 N	100 N
$\Delta D_v \%$	-43.7%	-44.35%
$\Delta D \%$	-34.6%	-35.2%

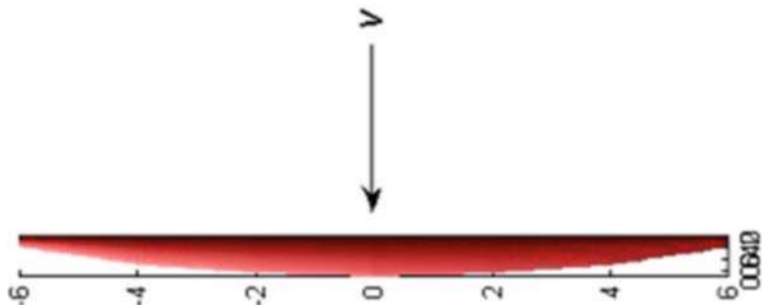


Fig. 23.23 Optimized geometry for the total drag problem (MADS+LocSmooth result)

objective function is highly non-linear because the friction drag coefficient depends on some power of the Reynolds number, which, in turn, is linearly dependent from the chord length.

Table 23.12 summarizes the main results relevant to the geometry of Fig. 23.23; Fig. 23.24 shows the lift distribution calculated on the Trefftz Plane.

Worth of notice is that the shape of the wing in Fig. 23.24 is very close to the today most efficient gliders.²

23.4.3 The Trimmed Aircraft

When an entire configuration is analyzed, an adjunctive equation concerning the equilibrium to rotations about the center of gravity has to be considered.³ The equilibrium to rotations together with the vertical equilibrium equation define the *trim*

² In the design of a wing other disciplines than aerodynamics are involved, like structures (neglected here).

³ Traditionally this equation is neglected in the aerodynamic optimisation. A tail is added at the end to accomplish equilibrium to rotations.

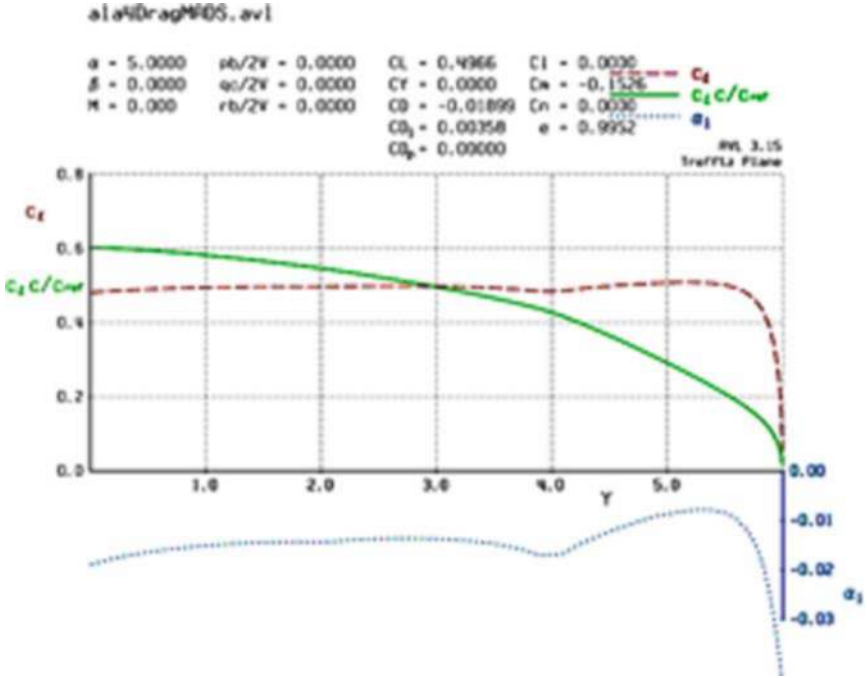


Fig. 23.24 Trefftz plane lift distribution for the total drag problem (MADS+LocSmooth result)

condition. Moreover, a condition on the *static stability* of flight has to be added to the previous equations. All these conditions are summarized in the following equations:

$$\sum_i^N L_i = W \quad (23.9)$$

$$\sum_i^N M_i|_{CG} = 0 \quad (23.10)$$

$$\frac{\partial M_{CG}}{\partial \alpha} < 0 \quad (23.11)$$

where N is the total number of lifting surfaces, the subscript CG indicates the center of gravity as pole of moments, W is the weight of the airplane and α is the reference angle of attack. Equation (23.9) states the vertical equilibrium and it is valid for rectilinear, levelled flight; Eq. (23.10) states the equilibrium to rotations about the center of gravity and Eq. (23.11) states the static equilibrium of flight. By defining the *static margin of stability* MoS

$$MoS = \frac{X_{NP} - X_{CG}}{mac} \quad (23.12)$$

where mac is the reference mean aerodynamic chord.

Equations (23.10) and (23.11) in the previous system may be translated in terms of engineering parameters, that is

$$|X_{PC} - X_{CG}| = 0 \quad (23.13)$$

$$MoS_{min} \leq \frac{X_{NP} - X_{CG}}{mac} \leq MoS_{max} \quad (23.14)$$

where X_{PC} , X_{CG} and X_{NP} indicate the position on the longitudinal axis of the center of pressures, the center of gravity and the neutral point, respectively. The two limits MoS_{min} and MoS_{max} are linked to the airplane maneuverability and their values come from experience.

23.4.4 The PrandtlPlane

As said before, the PrandtlPlane is an innovative aircraft configuration aiming at improving aerodynamic efficiency together with a weight save and therefore to reduce the Direct Operating Costs (DOC). It was conceived for commercial airplanes and for Very Large Aircraft (VLA), but the concept may be applied even to small airplane as ULM. The optimum condition, stating an equal repartition of lift between the two wings, is apparently in contrast with the requirement of static stability of flight; a great care must be taken in designing the two wings, the right combination of geometric parameter (swept angle, twist and chord distribution) must be set in order to satisfy both aerodynamic requirements on the Best Wing System and flight static stability conditions. Within this frame, the application of optimization methods seems to be the natural approach to solve the problem. As for the isolated wing, the aerodynamics of the lifting surfaces are modeled by a Vortex Lattice Method (VLM), and the friction drag is evaluated by means the flat plate analogy. As far as geometry is concerned, every wing is divided into two bays (three sections define a wing) and the design variables are taken as the twist angles and the chords at every section, the swept angle for every bay and the deflection angles of elevator and flaps, for a total of 18 geometric variables. Moreover, the angles of attack during cruise and landing are added. Finally, the total number of variables for the optimization is 20. The airfoils are fixed.

In terms of optimization, the problem can be stated as

$$\left\{ \begin{array}{l} \min D(x) = D_{ind} + D_{visc} \\ W_{min}^*|_{cr,Lan} \leq L_{cr,Lan} \leq W_{max}^*|_{cr,Lan} \\ |X_{CG} - X_{PC}|_{cr,Lan} \leq \delta \\ MoS_{min} \leq MoS \leq MoS_{max} \\ C_L|_{Lan} \leq C_L|_{Stall} \\ lb \leq x \leq ub \end{array} \right. \quad (23.15)$$

where

$$x = (c_i \ \theta_i \ \Lambda_j \ \alpha_{AV} \ \alpha_{BV} \ \delta_e \ \delta_f)$$

$$i = 1, \dots, 6$$

$$j = 1, \dots, 5$$

$\delta \in \mathbb{R}_+$, MoS is the static margin of stability, L is the lift, W is the weight, X_{CG} the longitudinal position of the centre of gravity, X_{PC} the longitudinal position of the pressure centre, C_L is the lift coefficient, the subscripts cr and Lan refer to the cruise and landing conditions, respectively, and the *Stall* subscript refer to stall condition. The following boundaries have been fixed:

$$\begin{aligned} 0.5 \text{ m} &\leq c_i \leq 1.5 \text{ m} & i = 1, \dots, 6 \\ -10 \text{ deg} &\leq \theta_i \leq 10 \text{ deg} & i = 1, \dots, 6 \\ 0 \text{ deg} &\leq \Lambda_i \leq 35 \text{ deg} & i = 1, 2, \text{ forward wing} \\ -35 \text{ deg} &\leq \Lambda_i \leq 0 \text{ deg} & i = 3, 4, \text{ rearward wing} \\ -3 \text{ deg} &\leq \alpha_{AV} \leq 3 \text{ deg} \\ 0 \text{ deg} &\leq \alpha_{BV} \leq 16 \text{ deg} \\ -20 \text{ deg} &\leq \delta_e \leq 20 \text{ deg} \\ 0 \text{ deg} &\leq \delta_f \leq 30 \text{ deg} \\ 4950 \text{ N} &\leq W \leq 5050 \text{ N} \\ 5 \% &\leq MoS_{cr, Lan} \leq 20 \% \end{aligned} \quad (23.16)$$

The starting geometry is shown in Fig. 23.25a, whereas different optimized geometries are shown in Fig. 23.25b, 23.26a and 23.26b. The OPT1 geometry (Fig. 23.25b) is the solution given by the LocSmooth global solver with the SQP local solver; for technological and manufacturing reasons, the geometric variations of this solution are not allowed, therefore an adjunctive constraint on the maximum relative swept angle between two bays has been added. The solution of the problem with the new constraint is the OPT2 geometry (Fig. 23.26a). This last solution still presents some small manufacturing issues, and the final optimization is performed “by hand” giving the solution shown in Fig. 23.25a. In Fig. 23.27 are reported the variations of the wing surface and drag for the different solutions. From Fig. 23.28 it can be inferred that the drag reduction is due both to a reduction of the wetted



Fig. 23.25a PrandtlPlane starting geometry

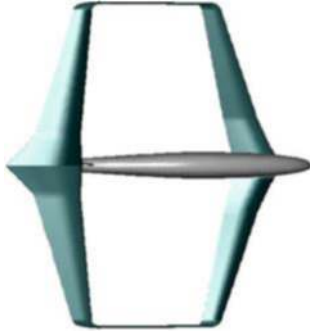


Fig. 23.25b OPT1 geometry

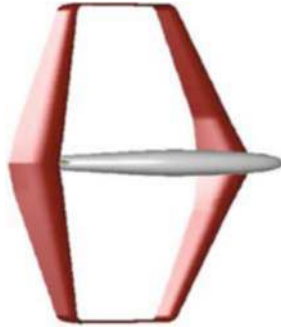


Fig. 23.26a OPT2 geometry

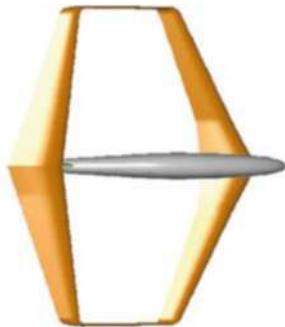


Fig. 23.26b Final (OPT3) geometry

surface and to a best distribution of lift; for this class of airplanes, with low wing loading, this last component weights in the measure of 3–6% of the global drag. Worth of notice is that when constraints are added, the minimum is higher than before (compare OPT2 results with OPT1).

A flying 1:3 scaled model of the aircraft as shown in the final geometry in Fig. 23.26b is under construction (Fig. 23.29).

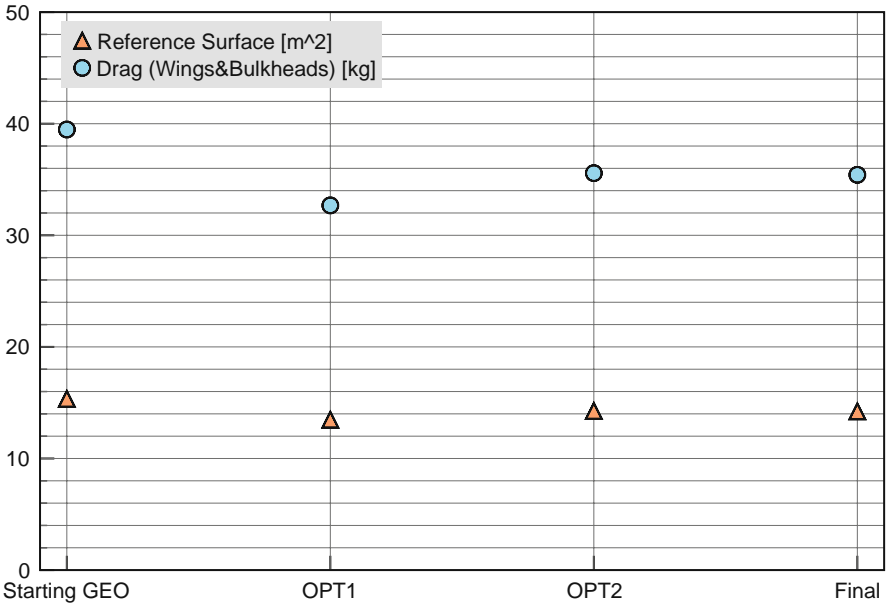


Fig. 23.27 Surface and drag variations during optimization

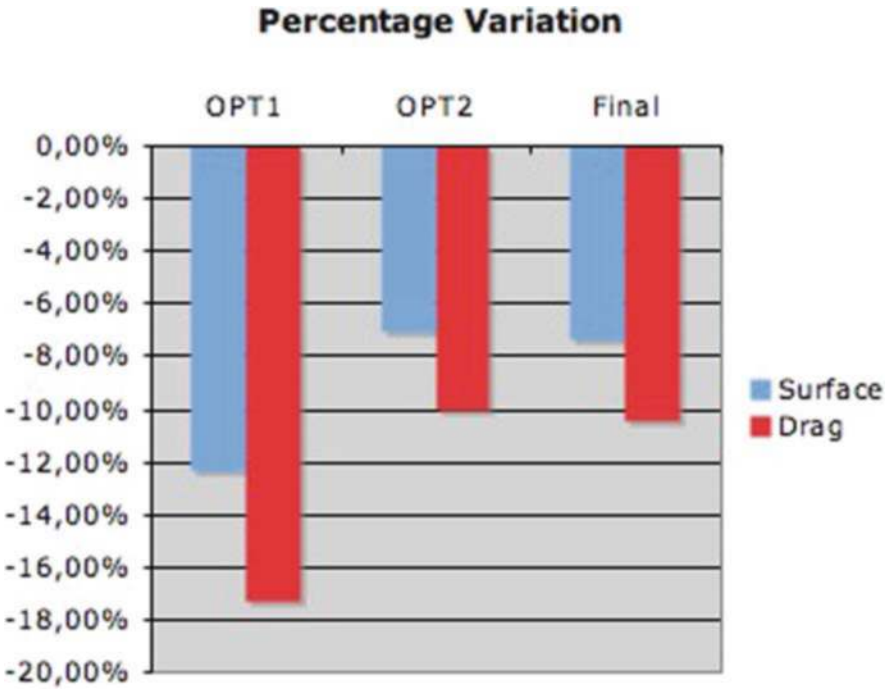


Fig. 23.28 Percentage variations of surface and drag



Fig. 23.29 Flying 1:3 scaled model under construction

23.5 Conclusions

Some optimisation algorithms have been presented and tested on benchmarking problems. In particular, an algorithm for the search of global minima demonstrated its effectiveness, finding global solution both in the case of closed form function and when a *black box* predicts objective and constraints. The algorithm has been applied to the preliminary design of an innovative, non-conventional small airplane, giving a final solution satisfying the imposed constraints and improving the aerodynamic efficiency. A scaled flying model of the final geometry so obtained is under construction.

References

1. I. Kroo, Aircraft Design: Synthesis and Analysis. <http://adg.stanford.edu/aa241/AircraftDesign.html>, (2007)
2. Busquin, P., Evans, R., Lagardere, J.-L.: The Future of European Aeronautics: a shared VISION for 2020, Air and Space Europe, Aeronautics Days 2001, Elsevier (2001).
3. Liebeck, R.H.: Design of the blended-wing-body subsonic transport. In: Lecture Series on Innovative configurations and advanced concepts for future civil aircraft, Von Karman Institute, ISBN 2-930389-62-1, VKI 2005–06, (2005).
4. McMasters, J.H.: A U.S. perspective on future commercial airliner design. In: Lecture Series on Innovative configurations and advanced concepts for future civil aircraft, Von Karman Institute, ISBN 2-930389-62-1, VKI 2005–06, (2005).

5. Frediani, A.: The Prandtlwing. In: Lecture Series on Innovative configurations and advanced concepts for future civil aircraft, Von Karman Institute, ISBN 2-930389-62-1, VKI 2005-06, (2005).
6. Prandtl L.: Induced Drag of Multiplanes, NACA TN 182 (1924).
7. Frediani A., Montanari G., Pappalardo M.: Sul problema di Prandtl della Minima Resistenza Indotta di un Sistema Portante (in Italian), Proceedings of the 15th national Italian Conference AIDAA (1999).
8. A. Frediani, E. Rizzo, C. Bottoni, J. Scanu, G. Iezzi: A 250 Passenger PrandtlPlane Transport Aircraft Preliminary Design, *Aerotecnica Missili e Spazio* Vol. 84 4/2005 (2005).
9. Di Pillo G., Palagi L.: Nonlinear Programming: Introduction, Unconstrained and Constrained Optimization, Tech. Rep. 25–01, Dipartimento di Informatica e Sistemistica “A. Ruberti”, Universit di Roma “La Sapienza”. <http://ftp.dis.uniroma1.it/PUB/OR/palagi/papers01/tr25-01.pdf> (2001)
10. Mangasarian O. L.: Non Linear Programming, SIAM (1994).
11. Fletcher R.: Practical Methods of Optimization, Vol. 1 & 2, Wiley (1981).
12. Addis, B., Locatelli, M., Schoen, F.: Local Optima Smoothing for Global Optimization, *Optimization Methods and Software*, Taylor & Francis, Vol. 20, No. 4–5, August–October 2005 417–437 (2005).
13. Addis, B., Leyffer, S.: A Trust-Region Algorithm for Global Optimization, *Computational Optimization and Applications*, Springer Netherlands, Vol. 35, No. 3 (2006).
14. Conn, A., Gould, N. I. M., Toint, P. L.: A Globally Convergent Augmented Lagrangian Algorithm for Optimization with General Constraints and Simple Bounds *SIAM J. NUMER. ANAL.*, Vol. 28, No. 2, 545–572 (1991).
15. Lewis R. M., Torczon V.: A Globally Convergent Augmented Lagrangian Pattern Search Algorithm for Optimization with General Constraints and Simple Bounds *SIAM J. OPTIM.*, Vol. 12, No. 4, 1075–1089 (1998).
16. Drele M.: AVL (Athena Vortex Lattice). <http://web.mit.edu/drele/Public/web/avl/> (2007)
17. Torczon V.: On the convergence of Pattern Search Algorithms *SIAM J. OPTIM.*, Vol. 7, No. 1, 1–25 (1997).
18. Audet, C., Dennis, J.E.: Mesh Adaptive Direct Search Algorithm for Constrained Optimization *SIAM J. OPTIM.*, Vol. 17, No.1, 188–217 (2006).
19. Ackley, D. H.: “A connectionist machine for genetic hill climbing”. Boston: Kluwer Academic Publishers, 1987.
20. Törn, A. and Zilinskas, A.: “Global Optimization”. Lecture Notes in Computer Science, N° 350, Springer-Verlag, Berlin, 1989.
21. Mühlenbein H., Schomisch D. and Born, J.: “The Parallel Genetic Algorithm as Function Optimizer”. *PARALLEL COMPUT.* 17, 619–632 (1991).
22. Schwefel, H.-P.: Numerical optimization of computer models. Chichester: Wiley & Sons, 1981.
23. Munk M.: Isoperimetrische Aufgaben aus der Theorie des Fluges, Inaugural Dissertation 1919, Gottinga (1919).
24. Munk M.: The minimum induced drag in airfoils, NACA 121(1924).

Chapter 24

Different levels of Optimisation in Aircraft Design

Dieter Schmitt

Abstract Air transport is still one of the continuously growing industry sectors worldwide. Like all industrial sectors there is the constant request to reduce cost, improve quality and enhance security and safety. The aircraft as the central mean for transportation is under similar threats. But if we want to improve the aircraft further we have to understand the air transport system itself, identify the different actors and their role, identify each partners strengths and weaknesses and then identify the areas where further aircraft improvements will bring best value to the system.

The presentation will start to quickly describe the air transport system with its main elements and partners. ACARE, the European consortium for air transport, has developed their Vision 2020 and defined two strategic research agendas to achieve the defined goals that can be used as a good basis for the next challenges.

The aircraft design process can be described in four different levels. The first level is the air transport system which defines the environment and constraints in which the aircraft can be operated. In this first level, the market requirements for the aircraft have to be derived. Once the market requirements are identified, then the industrial process starts:

- Which aircraft in terms of size and range would best fit?
- What is the competitive situation?
- What type of aircraft, a derivative or a new design?
- What level of technology and risks should be taken?
- What propulsion system and how to secure exclusivity?
- Who and how many risk sharing partners/subcontractors?

Here a compromise between the different aspects of marketing, engineering, finance and production has to be developed.

Dieter Schmitt

Airbus, F-31707 Blagnac Cedex, France, e-mail: dieter.schmitt@airbus.com

At the third level then is the purely engineering task, handled by the chief engineer. He has to define together with his engineering teams from aerodynamics, structures, aeroelasticity, propulsion, cabin, etc., the suitable aircraft configuration which fulfils all these requirements as a compromise between the different disciplines. At this level, the aircraft performance or the DOC (direct operating cost) can be the yardstick to measure and identify the improvements compared to previous and competitors design.

Today there is even a fourth level of aircraft optimisation. This is at the level of aircraft subsystems design, where another optimisation of functionalities is needed.

- What is the best way to control the aircraft?
- What is the future architecture to ensure communication within the aircraft (cable, wireless, mixed), between aircraft and ground and also for entertainment with all the new features like onboard TV, video on demand, use of mobile phones, etc.
- What is the optimum way of onboard power generation, distribution and economic consumption?

Optimisation is needed in all of these described levels. But it is very often fairly difficult to understand and define the system boundaries, the related optimisation parameters and target functions.

This chapter tries to give a global overview, but will not aim to provide all answers!

24.1 Air Transport System

The air transport system is today a global business and has several stakeholders who have to work closely together to provide an efficient transportation system. Figure 24.1 highlights the main actors and their interrelationship.

The transport system is traditionally an essential activity of national interest for each state/government and this is especially true for air transport, even and because this happens mainly internationally.

Transport recognises a privilege compared to other domains of the economy:

- Economic reasons:
 - Export-oriented industries need proper means of transport for goods.
 - Production areas need fast, cheap and reliable ways of transport.
- National reasons:
 - Demonstration of power/sovereignty, one of the reasons to have an air force to demonstrate air supremacy.
 - National air fleet for reserve for transport needs during war.
 - Prestige for national “flag carrier” (Air France, Iberia, Air India, Alitalia, British Airways, ...).

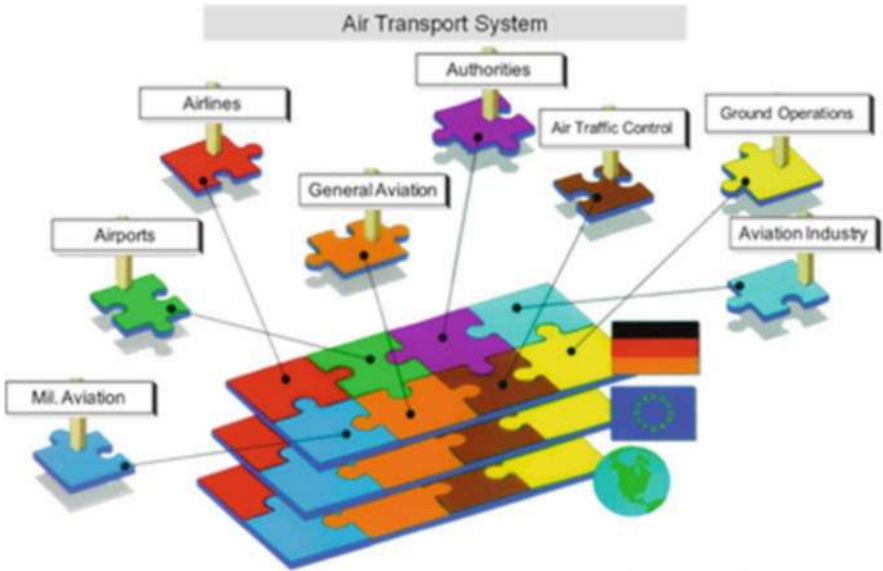


Fig. 24.1 Air transport system

Due to the internationality of air transport on one side and the air supremacy of the states on the other side a lot of contact points and common interests exist between state/government and air transport!

The strong link between economic growth and air transport can be seen by the surprising correlation between domestic growth product GDP and the transportation performance, expressed in revenue passenger miles (RPM).

As shown in Fig. 24.2 there is a strong similarity in percentage growth per year for GDP and RPM (revenue passenger miles), which means in simple terms sold air tickets per year. All world crises like 11th of September or beginning of Iraq war can be seen directly in the GDP curve but are also directly reflected by the sold air tickets. The RPM is, however, by a factor of 2–3 higher, which means air travel request is a very sensitive indicator for world economy and its growth potential.

24.2 Industrial Process of Aircraft Design

The aircraft industry is characterised by a very long cycle. It takes about 15–20 years to define a new product (aircraft type); this aircraft type will then be adapted to changing market conditions and will be produced in different versions for the next 20–50 years (B747 is in production since 1974 and a new version -800 has just being defined!) and even the last-produced aircraft will still stay in service for another 20 years. Figure 24.3 shows this typical life cycle of a civil aircraft programme [1–3].

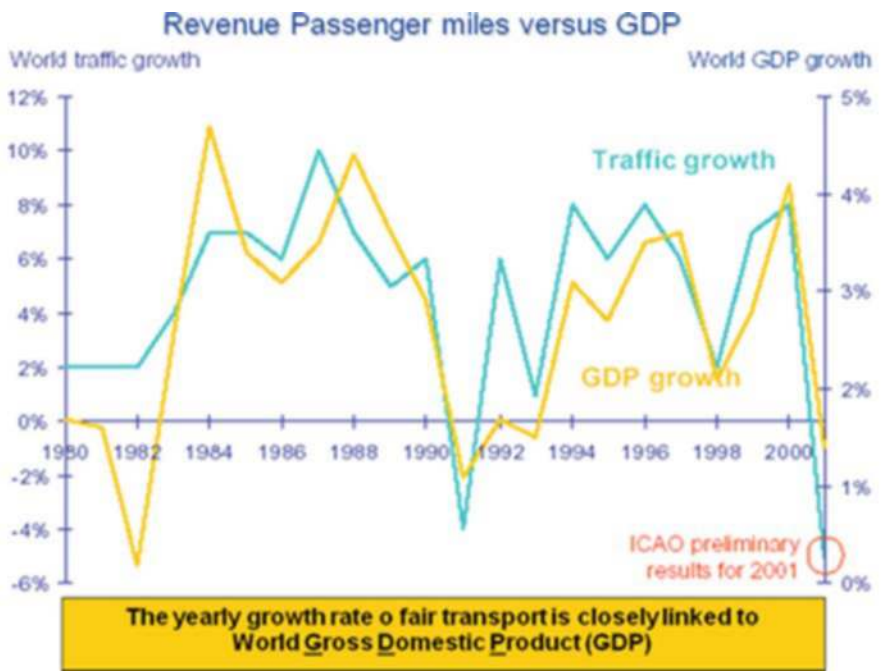


Fig. 24.2 Revenue passenger miles versus GDP

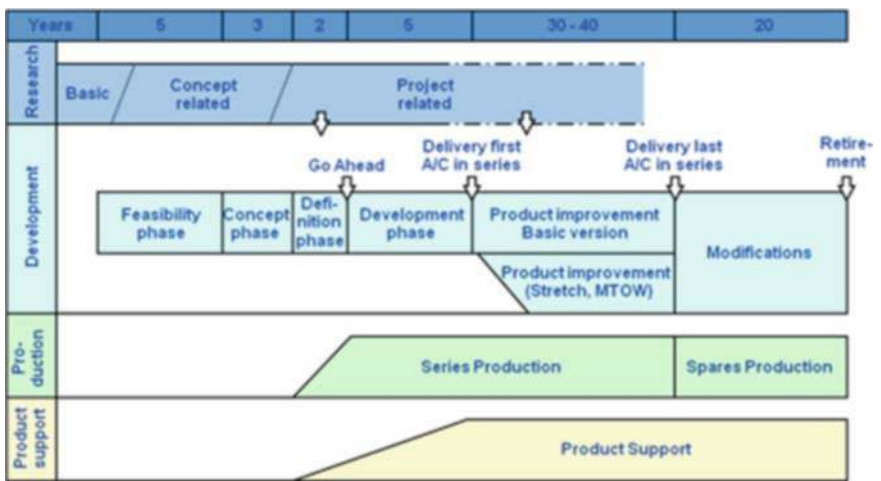


Fig. 24.3 Typical life cycle of a civil programme

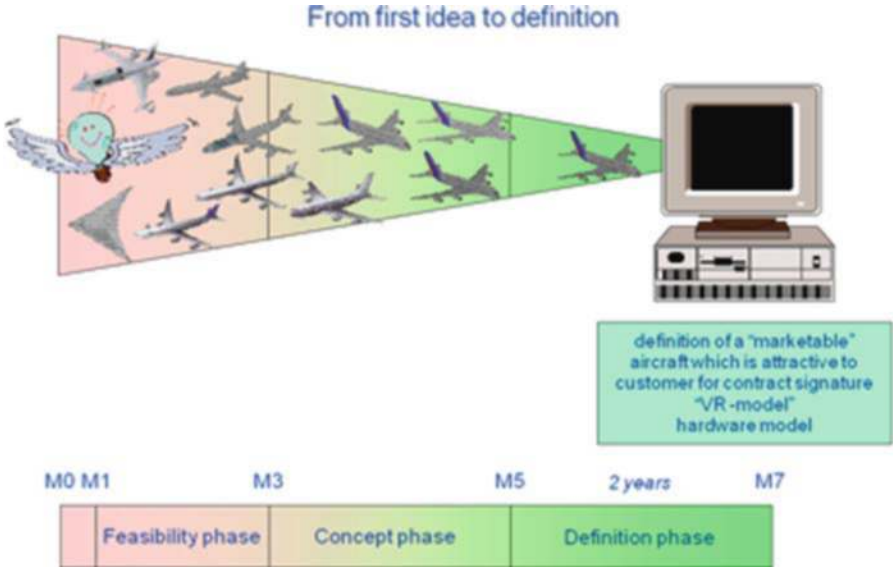


Fig. 24.4 From first idea to definition

But in this chapter, we want to concentrate on the aircraft design aspects, which starts with the first idea and ends with the decision and commitment of the aircraft manufacturer, to build the aircraft and give performance and delivery guarantees. This is normally characterised by the milestone “Go Ahead” or also often in literature defined as milestone M7 [4, 5].

Figure 24.4 defines the three phases from the first idea of a new aircraft concept to the final definition of a “marketable aircraft” which is attractive for new customers, which has convinced the launching customers to sign a contract and which has shown to the management of the aircraft manufacturer that there is a good potential in this programme to be profitable over the whole life cycle. At milestone M7, the new aircraft is fully defined as virtual product. Once the “Go Ahead” decision is taken, then the real development process bringing the “virtual product” definition into a “real product” will start, with all the high financial investments needed to set up the production line, have the first hardware parts produced (classically called “first metal cut” but today – in the world of carbon fibre materials – better called “first part production”), begin the final assembly line, start with the first flight of the first test aircraft, leading to the type certification of the aircraft and finally the delivery of the first aircraft to the customer and the “entry into service”. Figure 24.5 shows this development phase, which runs in parallel to the production process.

Figure 24.5 defines all 15 milestones, from M0 “product idea established” till M14 “end of development phase”.

The crucial point in this development cycle is milestone M7, the “Go Ahead” decision. This milestone separates the “definition phase” from the “development phase”. In the development phase, there is a virtual product definition, which is

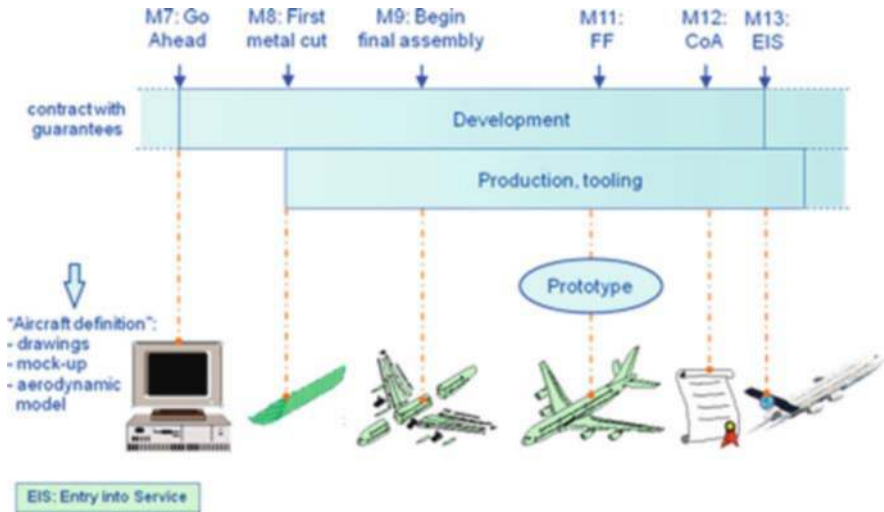


Fig. 24.5 From Go Ahead to EIS

defined in all details of external shapes, internal structural design, aircraft control and systems architecture definition, possible cabin interior layouts, A/C system installation and space allocation, etc. A very simplistic view could be to state the following: The aircraft is well defined in all critical aspects in order to deliver the necessary performance and it is now only necessary to transfer the virtual design into reality, i.e. a production aircraft. A lot of tools, methods, processes in the engineering world are known and have to be further improved to ensure that the development phase will be mastered in the most efficient way in terms of time, cost and quality.

24.3 Different Levels of Aircraft Design vs. Development Phases

The A/C product definition phase is more difficult to plan efficiently and identify a unique optimum process. Why? During the long preparation phases for a new aircraft type, very contradictory aspects from marketing, financial, business, engineering and production have to be agreed upon to arrive at a "marketable" aircraft definition. It is nearly impossible to structure and define a good process from "first idea" to "virtual product defined". To better understand the complexity of this definition phase, the easiest way is to start from the milestone M7. To get a decision from the board for launching the aircraft, the launching conditions, which have to be defined upfront, must have been fulfilled. Launch conditions are, for example, the number of aircraft ordered, number and quality of airlines and quality of contracts. In order to achieve the launch conditions, the customers (launch airlines) want to have a well-defined aircraft specification, performance and delivery guarantees, credibilities about the programme planning, the engine commitments, etc. This

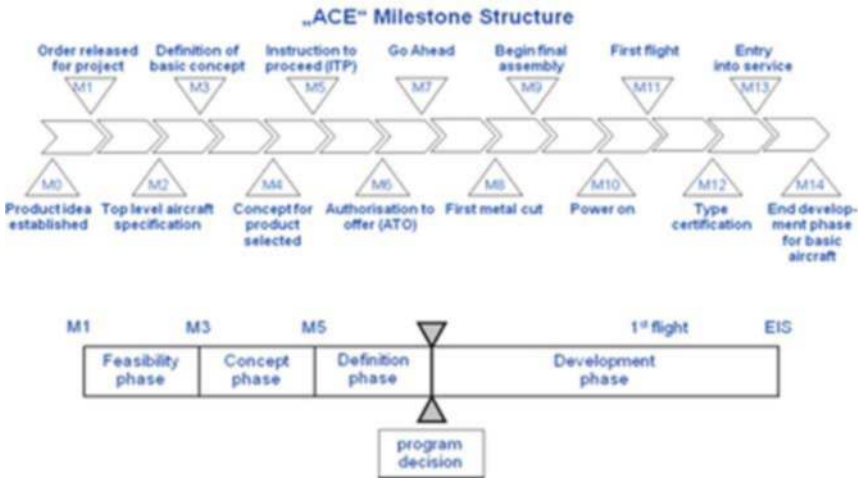


Fig. 24.6 “ACE” milestone structure

FIELDS OF ACTIVITIES

Dependencies by GO-AHEAD

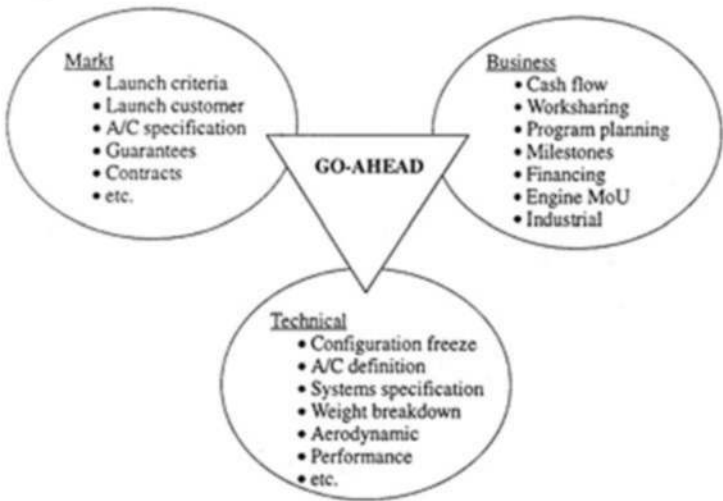


Fig. 24.7 A/C development plan for Go Ahead

means that the technical specification has to be defined even around milestone M5, when the aircraft has been authorised to be offered to the market.

Figure 24.7 shows these interdependencies in a simplified way. With the three major blocks market, business/finance and technical this complex process has now to be restructured, starting from M7 backwards. So all technical aspects, especially a good aircraft definition, weight breakdown and performance calculations, have

to be ready around M5 in order to go to the market and offer the aircraft to the launching customers. The process for “aircraft definition” is so very different from the development process of the aircraft. Main reasons are the following.

There is no clearly fixed engineering target. The marketing department will never be able to nor can really issue a clear market specification, which could be used to define the aircraft in engineering terms. The best market specification says: *define an aircraft configuration which is “marketable”* and this means for the engineering team:

- the payload–range capability is about fixed,
- the technology level should be high but cost efficient for the user,
- the competition will not wait for your final “product definition,”
- your “product proposal” has to show a “significant” market benefit relative to existing products,
- the schedule to achieve “Go Ahead” is defined but will depend on market situation,
- the management normally is reluctant to spend the necessary money in advance to let you develop a mature technology base,
- etc.

For a successful new product launch, the classical three factors (time, cost, quality) have to be matched simultaneously. In a simplified way again, the a.m. process elements from market, business and technical matters can be linked to the factors time–market; cost–business; technical–quality (see Fig. 24.8 and [4]).



Fig. 24.8 Time–cost–quality



Fig. 24.9 Aviation industry

Some 10 years ago, the new product definition was mainly a task from engineering to define the best possible aircraft. This has changed today and a close cooperation between marketing, finance and engineering is mandatory to define and prepare a new product for the market.

In reality the aeronautical industry has today an even more complex structure (see Fig. 24.9), where there are up to 10 different directorates in the industry, who all want to be involved in the next product definition.

If we define the air transport system as level 1 (Fig. 24.1) and the aeronautical industry with its different directorates as level 2 (Fig. 24.9), there is the next level below (level 3), the engineering level (see Fig. 24.10) with all its technical disciplines, which has to be successfully integrated for an optimum technical definition. In addition, there is a decision to be taken about a “Make or Buy” policy.

24.4 Tools Used in Different Phases

Which tools are used today to define an aircraft in a multidisciplinary way? Each major aeronautical discipline has developed its own sophisticated tools to do the necessary analysis and trade-off studies for the best of the new aircraft definition. Figure 24.11 is characterising a classical aircraft design approach. In a first preliminary sizing effort, the aircraft basic dimensions and characteristics are defined. Then, each technical domain is further optimising with its tools the basic aircraft concept, using all sorts of tools, from statistical to analytical and highly multidisciplinary ones, depending on the level of detail required. Figure 24.12 shows the

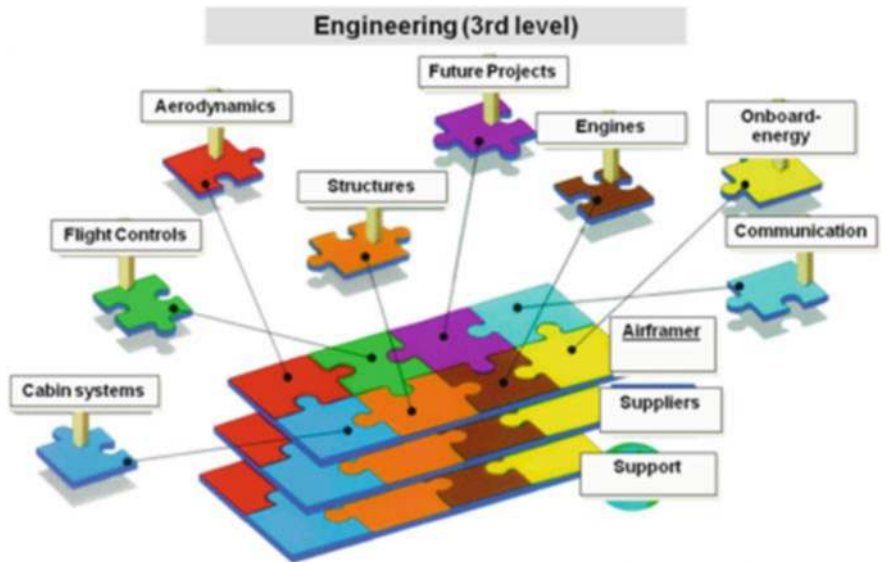


Fig. 24.10 Engineering level

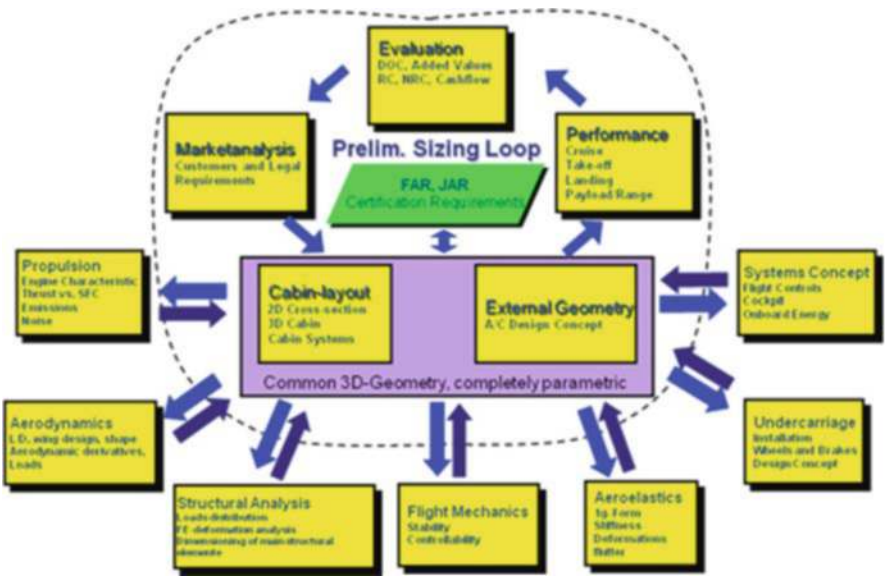


Fig. 24.11 Aircraft design approach

same need for an integrated approach, but is better illustrating the time aspect from the first market idea to a detailed aircraft concept, validated by integrated tools for “authorisation to offer” to the market.

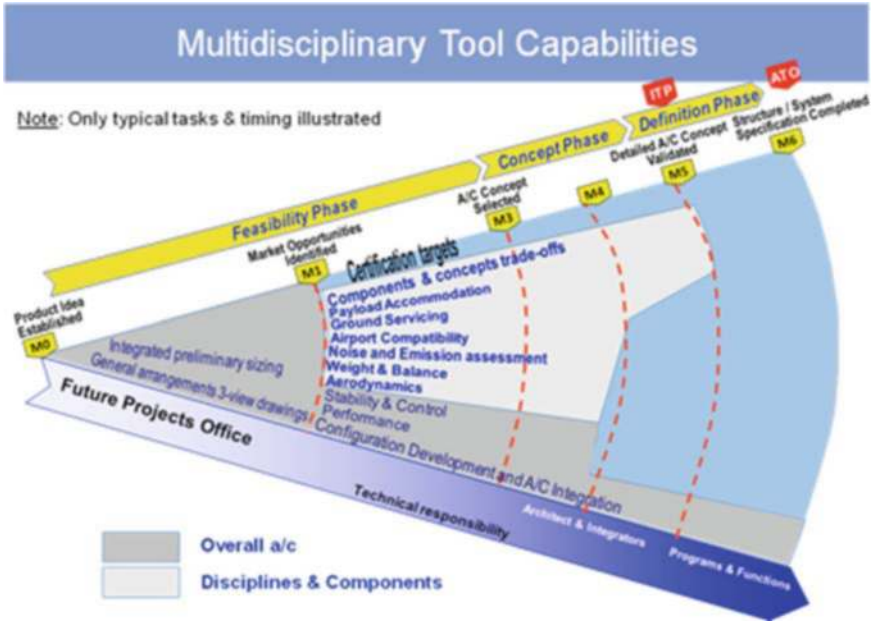


Fig. 24.12 Multidisciplinary tool capabilities

Another way to show the new approach is demonstrated in Fig. 24.13. In layer 3, all aeronautical domain centres have their well-specified, highly complex specialist tools, which have been validated during the last aircraft projects and which are maintained and further upgraded on a continuous basis. The big challenge is the connection of these specialist tools in layer 2, where there is a harmonised interface which allows an easy exchange of data with the other domains, either for providing input to specific request or for validating and verifying results from the other domains. A common geometric CAD basis is mandatory: CATIA V5 and further developments are standard today. But more is needed than just a common geometrical database. The ultimate goal is a layer 1 toolset, which can control the concept sizing and optimisation during the whole aircraft definition process by respecting and integrating the highly sophisticated tools from the specific domains.

During the feasibility phase (see Figs. 24.4 and 24.6), the following tools are used today:

- Statistically based tools
- a huge database of own and competitors A/C
- detailed database for geometry, aerodynamics, weight estimation, systems
- closed loop tools

This means there is a lot of expert knowledge collected, which is integrated in the tools, and all tools are properly tested and validated.

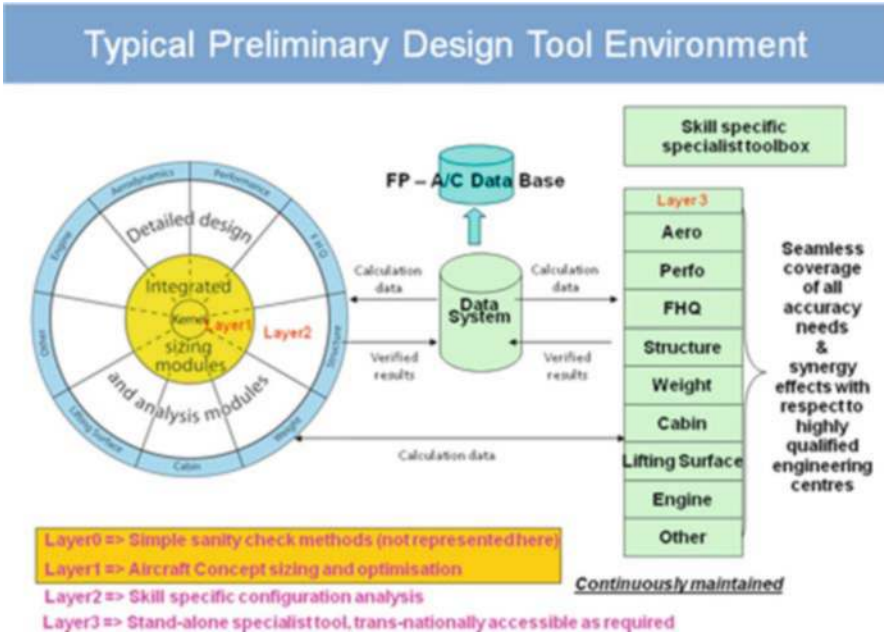


Fig. 24.13 Typical preliminary design tool environment

There is little use of optimisation and optimisation tools in industry! The understanding of a solution, the transparency of the solution is of prime importance to achieve credibility. The future: open loop tools will be used, which will easily adapt and cooperate with other domain tools.

During the concept and definition phases, the following approach is used in engineering today:

- Common geometry basis
- Integrated sizing tools
- All specialists have access to all data and info (more dream than reality!)
- Each engineering domain maintains and updates its tools and parts

The future: open the engineering toolbox and integrate market, finance and customers inputs.

The major progress is requested to change from level 3 (Fig. 24.10) to level 2 and use the consequence of a programme management organisation. This means that the programme management is also master of the toolbox used by each/all directorates. The key words are

- programme management via distributed enterprise;
- internal: use of digital mock-up (DMU), configuration management (Windchill), etc., with support and integration from all business units;
- management of supply chain; and
- use of customer forums to integrate customer needs.

The future will lead to

- Extended enterprise.

And finally it will end for the aircraft integrator in

- specify components and wait for delivery.

24.5 Conclusion

The following conclusions can be drawn:

- The industry is not really using optimisation tools and will not do so in future!
- Optimisation tools are limited to aeroelastic, aerodynamic and structural optimisations!
- A main problem is to convince the “non-technical management” for the need of a next level of tool integration on a programme management basis.
- To convince the “non-technical management” for a further integration of tools and a common toolset, a lot of simplified argumentations are required, mainly based on cost argumentations and issues.
- All aircraft design work in future will be done by integrated tools with transparent data input/output and a lot of sensitivity studies to define robust solution, not the best solution.

References

1. Jenkinson, L.R., Simpkin, P., Rhodes, D., Civil Jet Aircraft Design Arnold Publishers, London 1999.
2. John P. Fielding, Introduction to Aircraft Design Cambridge University Press, Cambridge 1999.
3. John E. Steiner, How Decision are Made – Major Considerations for Aircraft Programs, AIAA, 1982.
4. Dieter Schmitt, Concurrent Engineering in Aeronautics, Lecture held for ECATA Course, Technische Universität München, LLT; Jan. 2000.
5. McMasters, J.H and Kroo, Advanced configurations for very large transport airplanes, Aircraft Design Vol. 1, S. 217–242.
6. Walter Dolezal, Success factors for Digital Mock-ups (DMU) in Complex Aerospace Product Development, Dissertation , TU München, Munich, LLT

“This page left intentionally blank.”

Chapter 25

Numerical and Analytical Methods for Global Optimization

Paolo Teofilatto and Mauro Pontani

Abstract Global optimization is an important issue in the field of optimal aerospace trajectories. The joint use of global (approximate) numerical methods and of accurate local methods is one of the current approaches to global optimization. Nevertheless, the existence of global analysis techniques for investigating the possible multiplicity of results in optimization problems is of great interest. One of these techniques was the analysis of optimal trajectories through the Green's theorem. This approach, formerly introduced by Miele, has been applied to find singular optimal solutions related to missile, spacecraft, and aircraft trajectories. Another approach is based on the Morse theory, which relates the number of singular points of the objective function to the topology of the space where this function is defined. In particular, the so-called Morse inequalities provide a lower bound on the number of the local minima of the objective function. In this paper, Miele's and Morse's approaches will be recalled and applied to some problems in flight mechanics.

25.1 Introduction

Global optimization is an important issue in the field of optimal aerospace trajectories. In general, optimal control problems involve a dynamic system governed by a set of differential equations (the state equations), subject to some boundary conditions, and require the minimization of an objective cost (also termed "cost function"). An optimal control problem can be translated into a two-point boundary-value problem (TPBVP) by introducing a new set of variables (the Lagrangian

Paolo Teofilatto

Scuola di Ingegneria Aerospaziale, University of Rome "La Sapienza," via Eudossiana 16, 00184 Rome, Italy, e-mail: Paolo.Teofilatto@uniroma1.it

Mauro Pontani

Scuola di Ingegneria Aerospaziale, University of Rome "La Sapienza," via Eudossiana 16, 00184 Rome, Italy, e-mail: mauro.pontani@uniroma1.it

multipliers) and the Euler–Lagrange necessary conditions for optimality. The TP-BVP usually is not amenable to an analytic solution so numerical methods are to be employed. Shooting methods represent a class of numerical methods for solving such problems. Starting guess values are given to the unknown variables (generally the initial values of the Lagrangian multipliers, λ_0), then the boundary conditions are checked. If they are not satisfied to the prefixed accuracy, the initial guess is improved. A well-known drawback characterizes shooting methods, e.g., their poor robustness. As a matter of fact, a suitable guess must be provided to achieve the convergence to the final (optimal) result. In addition, Lagrangian multipliers usually do not have a straightforward physical meaning and their dependence on the initial values of the state variables, \mathbf{x}_0 , can be strongly nonlinear (as shown for instance in [1]). However, even providing a reasonable guess, an additional inconvenience cannot be avoided, e.g., the locality of the final result. In fact, the solution found by the numerical method corresponds to a single critical point of the cost function, more specifically the critical point “closest” to the starting guess. A possible approach to find other (locally) optimal solutions is based on the use of smart numerical algorithms to perform a global search on the space of solutions. Evolutionary methods (also referred to as Genetic Algorithms, GA) are able to generate a multiplicity of approximate locally optimal solutions. They employ an effective search technique that discretizes the space of the initial conditions λ_0 , thus reducing the number of trials while guessing the values of the unknown λ_0 . The joint use of genetic algorithms (for providing the starting guess) and local methods (for refinement) has been proved rather effective to find global optimal solutions in the context of aerospace optimization problems.

However, examples show that the number of critical points found by a GA may depend on the discretization in the λ_0 ’s space. For instance, let us consider the Euler equations of an axisymmetric rigid body controlled along the axis of symmetry:

$$\dot{x} = zy \quad \dot{y} = -zx \quad \dot{z} = u$$

The control u is chosen to maneuver the body from an initial state (x_0, y_0, z_0) to a final state (x_T, y_T, z_T) in a specified time T while minimizing the cost function $J = \int_0^T u^2(t) \mathcal{D}t$. It is relatively straightforward to prove that for the cost function J an infinite number of minima exists, where the GA could stop if a suitable region in the search space is not selected. Also the stepsize is relevant: in the above example (with $(x_0, y_0, z_0) = (1, 0, 0)$, $(x_T, y_T, z_T) = (0, 1, 0)$, $T = 3$) it is easy to show that minima for J can be missed if a stepsize greater than $8\pi/9$ on the $\lambda_y(0)$ axis or greater than $4\pi/3$ on the $\lambda_z(0)$ axis is chosen.

In general, the number of local solutions (i.e., of critical points) that a GA can find depends on the ranges where the values of the unknown variables are searched, see Fig. 25.1. Hence, analytical methods able to provide some information about the global optimal solution (or at least able to estimate the number of critical points) seems undoubtedly useful.

In this paper we recall two analytical methods, the first based on Green’s theorem and the second related to Morse theory. Green’s theorem approach was introduced by Angelo Miele in the 1950s (see the original papers [2–7]). This method is able to

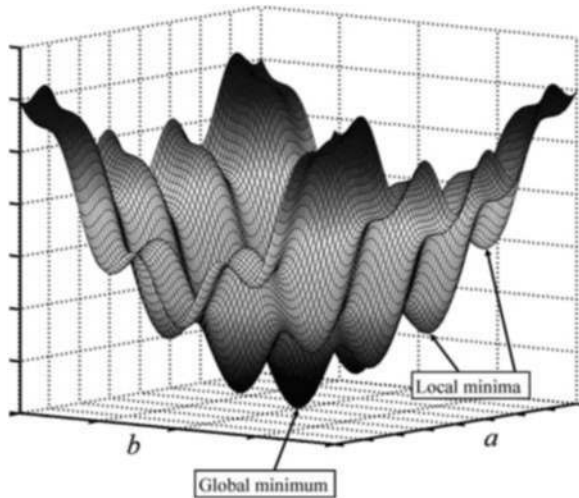


Fig. 25.1 Schematic representation of the minima of a cost function depending on two parameters

find the global optimal solution to problems possibly admitting an infinite number of critical points. In addition, the method under consideration has been successfully applied to identify some singular optimal solutions related to aircraft, spacecraft, and missile trajectories ([8–10]). As Green’s theorem approach is very effective as well as easy to be applied, it has been employed in many different fields: for instance in Biochemical Engineering ([11–13]), Management Science, and Biomechanics ([14–17]).

Another approach is based on Morse theory [18], which relates the number of critical points of the cost function J to the topology of the space where J is defined. In particular, the so-called Morse inequalities provide a lower bound on the number of local minima of the cost function.

In this chapter Miele’s and Morse approaches will be recalled and applied to some problems, in order to prove their effectiveness and their capability to drive numerical algorithms toward the global optimum.

25.2 Green’s Theorem Approach

An analytical method aimed at distinguishing minima among all the critical points, and eventually able to find the global optimum, has been proposed by Angelo Miele since 1950. The method is based on the application of Green’s theorem to optimal control problems defined on a planar region Ω of the state variables, with the cost function J given as the line integral of a linear differential form $\sigma (= \sigma_x \mathcal{D}x + \sigma_y \mathcal{D}y)$. Let the initial and final conditions $\mathbf{x}_0, \mathbf{x}_f$ be given on Ω . The choice of an admissible control function u_1 defines a curve γ_1 connecting \mathbf{x}_0 with \mathbf{x}_f , and the cost function related to such a control is

$$J(u_1) = \int_{\gamma_1} \sigma$$

If another control function u_2 is chosen, a different curve γ_2 is determined, and the difference between the two cost functions is equal to the line integral defined along the closed loop $\gamma_1 - \gamma_2$. Due to Green's theorem this integral is equal to the surface integral defined on the region Σ interior to the closed loop Σ

$$J(u_1) - J(u_2) = \int_{\gamma_1} \sigma - \int_{\gamma_2} \sigma = \int_{\gamma_1 - \gamma_2} \sigma = \int_{\Sigma} \mathcal{D}\sigma$$

where the two-form $\mathcal{D}\sigma$ is written as follows:

$$\omega = \mathcal{D}\sigma = \left(\frac{\partial \sigma_y}{\partial x} - \frac{\partial \sigma_x}{\partial y} \right) \mathcal{D}x \mathcal{D}y$$

The difference between the two cost functions depends on the sign of the two-form ω : for instance ω positive on Σ implies $J(u_1) > J(u_2)$. In general, if ω has a definite sign on Ω , the optimal solution is achieved at the boundary of the region Ω , $\partial\Omega$.

Example 25.1. A spacecraft is spin stabilized with an angular velocity $p(t) \geq p_0 \neq 0$. If a given variation of the roll angle ϕ_f is to be achieved in minimum time (for instance to point a sensor), the (simplified) dynamics can be represented by the differential system:

$$\begin{cases} \dot{\phi} = p \\ \dot{p} = u \end{cases} \quad (\|u\| \leq 1) \quad (25.1)$$

with initial and final conditions

$$\phi(0) = \phi_0, \quad p(0) = p_0, \quad \phi(t_f) = \phi_f, \quad p(t_f) = p_0$$

and cost function:

$$J = t_f = \int_0^{t_f} \mathcal{D}t = \int_{\mathbf{x}_0}^{\mathbf{x}_f} \frac{\mathcal{D}\phi}{p} \quad (25.2)$$

The region Ω attainable from the initial conditions and satisfying the constraint $p(t) \geq p_0$ is portrayed in Fig. 25.2. The left upper boundary of Ω is the parabola through \mathbf{x}_0 :

$$p^2 = 2\phi + C_0^+, \quad C_0^+ = p_0^2 - 2\phi_0$$

corresponding to the choice $u = 1$ for the control function, which belongs to the boundary of the space of the admissible control functions. The right upper boundary of Ω is the parabola through \mathbf{x}_f :

$$p^2 = -2\phi + C_0^-, \quad C_0^- = p_f^2 + 2\phi_f$$

Such trajectory corresponds to the choice $u = -1$. The lower boundary is associated to the condition $p(t) = p_0$. With reference to Fig. 25.3, let the curve I be a solution of (25.1) corresponding to a specified control function u_1 and II the curve corresponding to u_2 . The difference between the two cost functions is

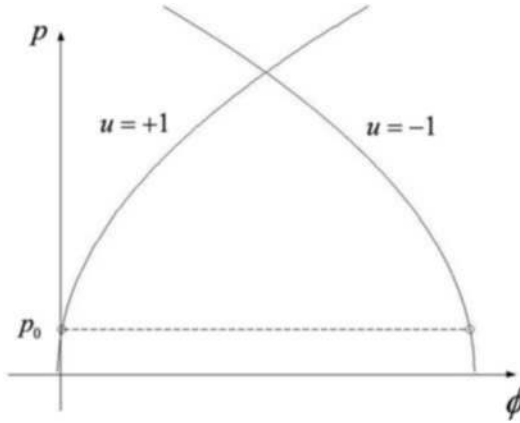


Fig. 25.2 The attainable region for the problem in Example 25.1

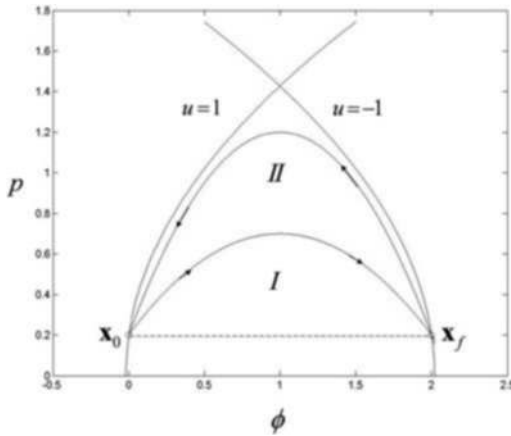


Fig. 25.3 The closed curve where the cost function difference is evaluated

$$J(u_1) - J(u_2) = (I) \int_{x_0}^{x_f} \sigma - (II) \int_{x_0}^{x_f} \sigma = (I) \int_{x_0}^{x_f} \sigma + (II) \int_{x_f}^{x_0} \sigma = \oint_{\Gamma} \sigma$$

where Γ is the closed curve from x_0 to x_0 followed in counterclockwise sense.

The application of Green's theorem leads to

$$\oint_{\Gamma} \sigma = \int_{\Sigma} \omega$$

where Σ is the interior of the closed loop Γ , and the two-form ω is

$$\omega = \mathcal{D}\sigma = \frac{1}{p^2} \mathcal{D}\phi \mathcal{D}p$$

The two-form ω is positive definite on Ω , so $J(u_1) > J(u_2)$ and the minimum time strategy corresponds to the boundary of the region Ω . Hence, the optimal solution is bang–bang, that is

$$u^* = \begin{cases} +1 & \text{if } 0 \leq t < t_1 \\ -1 & \text{if } t_1 < t \leq t_f^* \end{cases} \quad (25.3)$$

where the minimum time is $t_f^* = 2t_1$, and t_1 is the switching time

$$t_1 = \sqrt{p_0^2 + (\phi_f - \phi_0) - p_0}$$

Example 25.2. The determination of the minimum consumption impulsive transfer between coplanar Keplerian orbits has been a classical problem in orbital mechanics [19]. Many important contributions were given by several researchers, but a definitive answer for transfers between coplanar and coaxial orbits was given in [20] using Green’s theorem. The region Ω for such a problem is portrayed in Fig. 25.4 where the perigee radius x_p is associated to the horizontal axis, and the vertical axis X represents the inverse of the apogee radius x_a ($X = 1/x_a$). The region is bounded by the horizontal axis $X = 0$, representing parabolic trajectories of different perigee radii, and by the hyperbola $X x_p = 1$, representing circular orbits. In Fig. 25.4 dimensionless coordinates are used with respect to the Earth radius, so that $X = x_p = 1$ is the Earth surface and Ω is defined to the right of the straight line $x_p = 1$. The application of the Euler–Lagrange conditions to impulsive transfers between Keplerian trajectories [21] leads to the following property: for elliptic orbits (locally) optimal impulses must be applied tangentially at apogee or at perigee. In Fig. 25.4 these

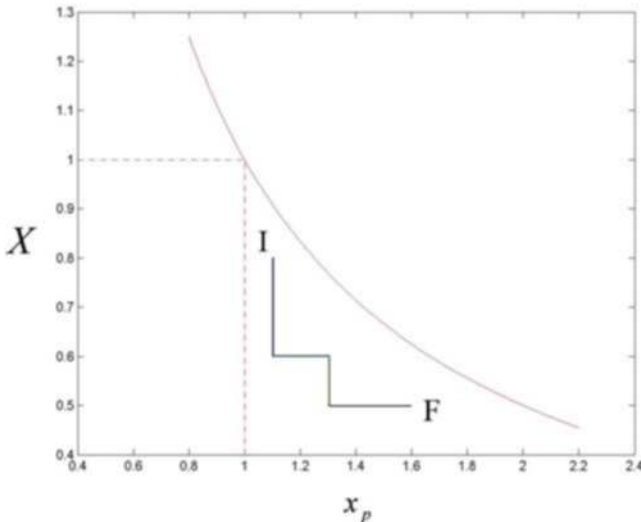


Fig. 25.4 Attainable region for transfers between elliptic orbits

locally optimal transfers are represented as horizontal or vertical segments in the (x_p, X) plane. This figure shows an example of a four-impulse locally optimal transfer performed through a perigee–apogee–perigee–apogee sequence of impulses. If γ is any piecewise straight line representing a maneuver in the (x_p, X) plane, the related variation of velocity ΔV needed to perform the transfer can be computed through the integral [20]

$$\Delta V = \int_{\gamma} \sigma = \int_{\gamma} \sigma_p \mathcal{D}x_p + \sigma_X \mathcal{D}X$$

where

$$\sigma_p = \pm \sqrt{\frac{\mu}{2}} \frac{X}{\sqrt{x_p} (1 + Xx_p)^{\frac{3}{2}}} \quad (+) \text{ if } x_p \text{ increases, } (-) \text{ if } x_p \text{ decreases}$$

$$\sigma_X = \pm \sqrt{\frac{\mu}{2} x_p} \frac{1}{(1 + Xx_p)^{\frac{3}{2}}} \quad (+) \text{ if } X \text{ increases, } (-) \text{ if } X \text{ decreases}$$

For instance, let us consider a transfer from a low circular orbit to a higher orbit, as in Fig. 25.5. The two-impulse Hohmann transfer Ia_2F is represented and the related ΔV is the integral

$$\Delta V_1 = \int_{Ia_2F} \sigma$$

Let us compare the Hohmann transfer with any other transfer satisfying the constraint $X \geq X_F$, that is the apogee of any transit orbit is at lower altitude than that of the final orbit. Locally optimal transfers are piecewise straight lines in Fig. 25.5.

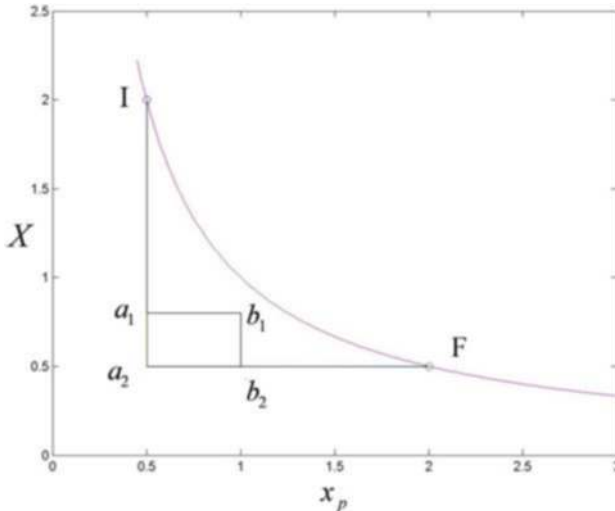


Fig. 25.5 Comparison between the Hohmann transfer and a four-impulse transfer

Let us consider for instance the four-impulse transfer $Ia_1b_1b_2F$, with

$$\Delta V_2 = \int_{Ia_1b_1b_2F} \sigma$$

The difference in the cost function between the two strategies is

$$\begin{aligned} \Delta V &= \Delta V_1 - \Delta V_2 = \int_{Ia_2F} \sigma - \int_{Ia_1b_1b_2F} \sigma = \\ &= \left(\int_{Ia_1} \sigma + \int_{a_1a_2} \sigma + \int_{a_2b_2} \sigma + \int_{b_2F} \sigma \right) - \left(\int_{Ia_1} \sigma + \int_{a_1b_1} \sigma + \int_{b_1b_2} \sigma + \int_{b_2F} \sigma \right) = \\ &= \oint_{a_1a_2b_2b_1a_1} \sigma \end{aligned}$$

i.e., the difference is equal to the line integral along the boundary of the rectangle $a_1a_2b_2b_1a_1$ (followed in counterclockwise sense). Through Green's theorem one obtains

$$\Delta V = \int_{\Sigma} \omega = \int_{\Sigma} \left(\frac{\partial \sigma_X}{\partial x_p} - \frac{\partial \sigma_p}{\partial X} \right) \mathcal{D}x_p \mathcal{D}X$$

where Σ is the region inside the rectangle; since $\omega < 0$ in such a region, the two-impulse Hohmann transfer is less expensive. The same analysis can be performed to show that the two-impulse strategy is optimal with respect to any other number of impulses. Therefore Green's theorem identifies the Hohmann transfer as the global optimal solution among the infinitely many locally optimal solutions, if the constraint $X \geq X_F$ holds. If such a constraint is removed and in the case of hyperbolic trajectories, see [20, 22].

In the above examples the two-form ω has a definite sign on the region Ω , and, as a consequence, the optimal solution is at the boundary $\partial\Omega$. Green's method allows the determination of the global optimal solution also when ω changes its sign [23].

Example 25.3. Consider the following optimal control problem [23]:

$$\begin{cases} \dot{x}_1 = x_1 + u x_2 & x_1(0) = 1, x_2(0) = 0 \\ \dot{x}_2 = x_2 + u x_1 & x_1(t_f) = 3, x_2(t_f) = 0 \end{cases} \quad (25.4)$$

with the (scalar) control variable constrained to the interval $[-1, 1]$ (i.e. $\|u\| \leq 1$), and the following cost function:

$$J = \int_0^{t_f} \left(x_1 + \frac{x_2}{2} + \log(x_1 + x_2) \right) \mathcal{D}t \quad (25.5)$$

The objective function (25.5) can be rewritten as

$$J = \int_{x_0}^{x_f} \left[x_1 + \frac{x_2}{2} + \log(x_1 + x_2) \right] \left(\frac{x_1 \mathcal{D}x_1 - x_2 \mathcal{D}x_2}{x_1^2 - x_2^2} \right) \quad (25.6)$$

Then the two-form ω is given by

$$\omega = \frac{x_1 + 2x_2 + 2}{x_2^2 - x_1^2}$$

The attainable region Ω is partitioned into two sub-regions, where $\omega < 0$ and $\omega > 0$, as shown in Fig. 25.6. With reference to Fig. 25.6, the application of Green's theorem leads to the identification of the optimal path from I to F. Due to the sign of ω , the path IDB is bettered by the path IAB, and the path ACF is bettered by the path ABF. Hence, the optimal path from I to F is unequivocally composed of the following three segments:

- IA, associated to the choice $u = 1$ for the control variable;
- AB, associated to an intermediate value for u ($-1 < u < 1$);
- BF, associated to the choice $u = 1$.

Green's theorem method has been also generalized to some multidimensional control problems in [24].

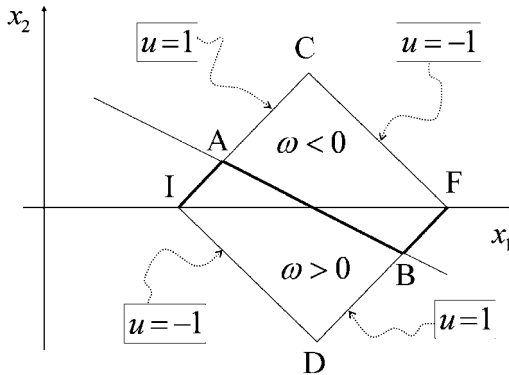


Fig. 25.6 Optimal path in the state space for Example 25.3

25.3 Morse Theory Approach

Let us return to the elementary optimal control problem in Example 25.1:

$$\begin{cases} \dot{\phi} = p \\ \dot{p} = u \end{cases} \quad (25.7)$$

A minimum energy maneuver in a specified time interval $[0, T]$ is such that the quadratic cost functional

$$J = \frac{1}{2} \int_0^T u^2(t) \mathcal{D}t$$

is minimized. The control u , which is assumed to be square integrable ($\|u\|_{L_2} < \infty$), drives the system from a specified initial state to a given terminal state:

$$\phi(0) = 0, \quad p(0) = 0, \quad \phi(T) = \phi_f, \quad p(T) = 0$$

The Hamiltonian function is introduced as

$$H = \lambda_\phi p + \lambda_p u - \frac{1}{2}u^2$$

so the equations for the multipliers are

$$\dot{\lambda}_\phi = -\frac{\partial H}{\partial \phi} = 0 \quad \dot{\lambda}_p = -\frac{\partial H}{\partial p} = -\lambda_\phi$$

Hence

$$\lambda_\phi = c_1, \quad \lambda_p = -c_1 t + c_2$$

The Euler–Lagrange condition $\partial H / \partial u = 0$ yields the optimal control law:

$$u^* = \lambda_p = -c_1 t + c_2 \quad (25.8)$$

After inserting (25.8) in (25.5) one gets the solution for the state variables:

$$\begin{aligned} \phi(t) &= -c_1 \frac{t^3}{6} + c_2 \frac{t^2}{2} \\ p(t) &= -c_1 \frac{t^2}{2} + c_2 t \end{aligned}$$

The final conditions determine univocally the constants c_1 and c_2 :

$$\begin{cases} -c_1 \frac{T^3}{6} + c_2 \frac{T^2}{2} = \phi_f \\ -c_1 \frac{T^2}{2} + c_2 T = 0 \end{cases} \quad (25.9)$$

so the optimal control is *unique*.

Suppose now that the terminal condition is not a single state but any point on the unit circle S^1 , parameterized with the angle θ :

$$\phi_f = \cos \theta \quad p_f = \sin \theta$$

The optimal control law is again (25.8) but the final conditions yield

$$\begin{cases} -c_1 \frac{T^3}{6} + c_2 \frac{T^2}{2} = \cos \theta \\ -c_1 \frac{T^2}{2} + c_2 T = \sin \theta \end{cases} \quad (25.10)$$

Hence, the cost function can be expressed as a function of θ

$$J = J(\theta) = \frac{2}{T^3} [(T^2 - 3) \sin(2\theta) - 3T \cos(2\theta)]$$

and the stationarity condition $\partial J / \partial \theta = 0$ produces the four solutions:

$$\theta_k = \frac{1}{2} \arctan \frac{3T}{T^2 - 3} + k \frac{\pi}{2} \quad k = 0, 1, 2, 3 \quad (25.11)$$

The four extremal solutions reach the unit circle in four points split at 90° , the value of θ_0 depends on the final time T as shown in (25.8). It is apparent that the cost function J has two equivalent minima (for $\theta = \theta_0, \theta_2$) and two equivalent maxima (for $\theta = \theta_1, \theta_3$). Figure 25.6 shows the resulting four different trajectories for $T = 2$. Symmetry with respect to the origin is due to the fact that if u is an admissible control also $-u$ is admissible with the same value of the cost function. Such a symmetry shows that the problem can split into two problems considering the target region N as the disjoint union of the left and the right semi-circles: $N = S_- \cup S_+$. The problem with target region $N_- = S_-$ is equivalent to the problem with target region $N_+ = S_+$.

The multiplicity of solutions to the above problem (unlike the target point (ϕ_f, p_f) problem) depends on the topology of the final target region. Namely N is composed of two sets that are not simply connected. In contrast, in the former example – the target point problem – a unique solution exists [25]. This is due to the fact that the target region is convex. One can conclude that there exists a relationship between the topology of the target region N and the number of the critical points of J .

Such a relationship is very well understood in the framework of functions $f : M \rightarrow \mathfrak{R}$ defined on a finite dimensional manifold M . Morse theory [18] states that there exist coordinates around any non-degenerate critical point \mathbf{x}_0 such that the function f is locally

$$f(\mathbf{x}) = f(\mathbf{x}_0) - \underbrace{x_1^2 - x_2^2 - \dots - x_k^2}_{\text{negative}} + \underbrace{x_{k+1}^2 + \dots x_n^2}_{\text{positive}}$$

The number k is referred to as the index of f at the critical point \mathbf{x}_0 (and it is independent of the coordinates, which can be arbitrary). Let us take, for instance, $n = 2$. If $k = 0$ the critical point \mathbf{x}_0 is a minimum, if $k = 1$ \mathbf{x}_0 is a saddle point, and if $k = 2$ \mathbf{x}_0 is a maximum.

The so-called Morse inequalities relate the number of critical points of index k of the function f to the topology of the space where f is defined. Suppose that f has only non-degenerate critical points on M and let $c_k(f)$ be the number of critical points of index k , then

$$c_k(f) \geq \dim(H_k(M)) \quad \text{Morse inequalities} \quad (25.12)$$

In (25.12) $H_k(M)$ is the k -homology group of M in the sense of singular homology. For a precise definition of these groups, see [26].

In particular, the dimension of the 0-homology group of M , $H_0(M)$, is equal to the number of connected components of M , and the dimension of the 1-homology group

$H_1(M)$ is equal to the number of closed paths on M which cannot be contracted into a point. For instance the target space $N = S_- \cup S_+$ of the example at the beginning of this section has the following homology groups (the double of the homology groups of a circle): $\dim(H_0(N)) = 2$, $\dim(H_1(N)) = 2$, $\dim(H_k(N)) = 0$ for $k \geq 2$.

If the Morse inequalities (25.12) are applied with $f = J$ and $M = N$ it turns out that two minima and two saddle points (at least) must exist for the cost function J . However, the straightforward application of (25.12) to the optimal control problem at hand leads to dealing with two interesting topics:

- (1) the cost function J is not defined on N , but it is defined on the (infinite dimensional) space of control functions \mathbf{M} that generate a flow connecting the initial condition \mathbf{x}_0 to the target space N .
- (2) two minima and two maxima for J (and not two saddle points) can be found by applying the Euler–Lagrange conditions

With reference to the issue (25.1), previous researches [27–29] have proved that, under proper hypotheses, the Morse inequalities hold for the cost function defined over \mathbf{M} :

$$c_k(J) \geq \dim(H_k(\mathbf{M}))$$

This means that, under appropriate assumptions, the space of the control functions \mathbf{M} preserves the same topological properties of the target space N . In fact \mathbf{M} has the structure of a fibred space: at any point \mathbf{y} of N there is an attached space $\pi^{-1}(\mathbf{y})$ consisting of those control functions that generate flows connecting \mathbf{x}_0 and \mathbf{y} in the fixed time T . Then the space \mathbf{M} is obtained considering all these fibers $\pi^{-1}(\mathbf{y})$ together in a bundle (see Fig. 25.7). If all the fibers $\pi^{-1}(\mathbf{y})$ are contractible, i.e., homotopically equivalent to a point, then the topology of the bundle \mathbf{M} coincides with the topology of the base space N [30]. For the optimal control problem (25.5), as well as for any linear control problem with target space N , it is trivial to prove that

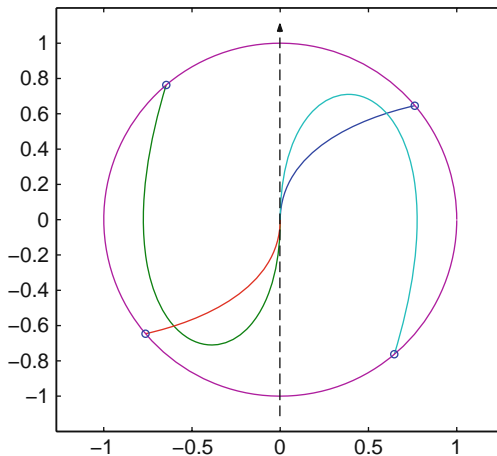


Fig. 25.7 The four trajectories corresponding to the four critical control functions

the control function space $\pi^{-1}(\mathbf{y})$ is convex for any $\mathbf{y} \in N$, therefore contractible. This circumstance implies that

$$H_k(\mathbf{M}) = H_k(N)$$

In conclusion one has

$$c_k(J) \geq H_k(N)$$

and this occurs for the example (25.9).

With reference to the issue (25.2), J exhibits two maxima and two minima when the control u belongs to the class of linear control functions (as dictated by the Euler–Lagrange conditions). Regarding J as a function of the entire control space of the square integrable functions, the two maxima correspond to saddle points, whereas the two minima are again minima. These properties are evident just considering a quadratic form for the control u :

$$u = at^2 + bt + c$$

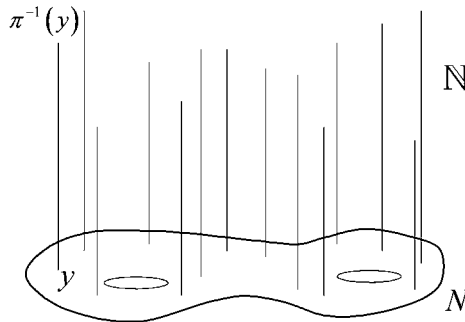


Fig. 25.8 The fiber bundle of admissible controls

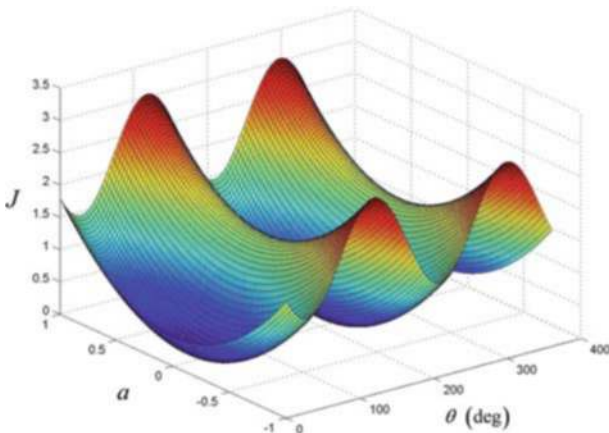


Fig. 25.9 The cost function J as a function of θ and a

The terminal conditions lead to the determination of the parameters b and c as functions of a . Hence the two-parameter function $J = J(\theta, a)$, portrayed in Fig. 25.8, exhibits two minima and two saddle points, as expected. The two saddle points correspond to maximum points over the curve associated to $a = 0$, which identifies the case of linear control functions.

Therefore the number of critical points of J are $c_0(J) = 2$, $c_1(J) = 2$, and $c_k(J) = 0$ for $k \geq 2$, and the number of critical points of J of degree k is equal to the dimension of the k -homology group of the target space N .

25.4 Final Comments

To the authors' knowledge, the Green's theorem approach proposed by Miele is the only versatile, analytical method able to find global optimal solutions in optimal control problems. The method has been successfully applied in many different fields of research.

Morse theory approach may provide lower bounds on the number of critical points of the cost function. The method can be used to drive a global numerical search, as that performed by employing Genetic Algorithms. At present rather elementary applications of Morse theory to optimal control are known. This is due to the difficult computations of the homology groups. On the other hand, the dimension of these homology groups is invariant with respect to the cost function J , provided that J satisfies certain conditions. Then the number of critical points of an optimal control problem with a specified cost function J can be derived by solving the same problem with a different cost function, thus allowing an easier solution. This could disclose the application of Morse theory to more complex optimal control problems.

References

1. Teofilatto, P., De Pasquale, E.: A non linear adaptive guidance for last stage control. *Journal of Aerospace Engineering, Part G* **213**, 45–55 (1999)
2. Miele, A.: Minimum time in nonsteady flight of aircraft (in Italian). *Atti della Accademia delle Scienze di Torino*, **85**, 41–52 (1950)
3. Miele, A.: General optimal solutions for aircraft in nonsteady flight (in Italian). *L'Aerotecnica*, **32**, 135–142 (1952)
4. Miele, A.: Optimal flight trajectories of turbojet aircraft (in Italian). *L'Aerotecnica*, **32**, 206–219 (1952)
5. Miele, A.: On non steady climb of turbojet aircraft. *Journal of the Astronautical Sciences*, **21**, 781–783 (1954)
6. Miele, A.: Minimum time flight trajectories (in Italian). *Atti della Accademia delle Scienze di Torino*, **21**, 80–87 (1954)
7. Cicala, P., Miele, A.: Brachistocronic maneuvers of constant mass, *Journal of the Astronautical Sciences*, **2**, 286–288 (1955)
8. Miele, A.: General variational theory of flight paths of rocket powered aircraft, missiles and satellite carriers, *Astronautica Acta*, **4**, 11–21 (1958)

9. Miele, A.: Flight mechanics and variational problems of linear type, *Journal of the Astronautical Sciences*, **25**, 286–288 (1958)
10. Miele, A.: Extremization of linear integrals by Green's theory. In: Leitmann, G. (ed) *Optimization Techniques*. Academic Press (1962)
11. YamarĖ, T. et al.: Start up of chemostat: application of fed-batch culture, *Biotechnology and Bioengineering*, **21**, 111–129 (1978)
12. Weigard, W.: Maximum cell productivity by repeated fed-batch culture, *Biotechnology and Bioengineering*, **23**, 249–266 (1980)
13. Constantinides, A.: Application of optimization methods to the control of fermentation processes, *Annals of the New York Academy of Sciences*, **326**, 193–221 (1979)
14. Sethi, S.: Optimal institutional advertising: minimum time problems, *Journal of Optimization theory and applications*, **14**, 213–231 (1974)
15. Sethi, S.: Optimal advertising policy with the contagion model, *Journal of Optimization theory and applications*, **29**, 615–627 (1979)
16. Sethi, S.: Optimal quarantine programs for controlling epidemic spread, *Journal of Optimization theory and applications*, **29**, 202–212 (1978)
17. Stabble, P., Maronsky, R.: Minimum time running and swimming, *Journal of Biomechanics*, **29**, 245–249 (1996)
18. Milnor, J.: *Morse theory*. Princeton University Press, Princeton (1969)
19. Edelbaum, P.: How many impulses ? , *Astronautics and Aeronautics*, **5**, 245–249 (1997)
20. Hazelrigg, G.: Globally optimal impulsive transfers via Green's theorem, *Journal of Guidance, Control, and Dynamics*, **7**, 462–470 (1984)
21. Lawden, D.: *Optimal trajectories for space navigation*. Butterworths, London (1963)
22. Pontani, M.: Simple Method to Determine Globally Optimal Orbital Transfers. *Journal of Guidance, Control and Dynamics*, **32**, No. 3, 899–914 (2009)
23. Hermes, H., Heynes, G.: On nonlinear control problem with control appearing linearly, *SIAM Journal of Control*, **1**, 85–108 (1963)
24. Heynes, G.: The optimality of a totally singular vector control: an extension of the Green's theorem to higher dimensions, *SIAM Journal of Control*, **4**, 662–677 (1966)
25. Neustadt, L.: Minimum effort control systems, *SIAM Journal of Control*, **1**, 16–31 (1962)
26. Spainer, E.: *Algebraic topology*, Mc Graw-Hill, New York (1966)
27. Palis, R., Smale, S.: A generalized Morse theory, *Bulletin of the American Mathematical Society*, **70**, 165–172 (1964)
28. Agrachev, A.A., Vakhramnev, S.A.: Morse theory and optimal control problems. In *Progress in System Control Theory*, Birkhauser, Boston 1–11 (1991)
29. Vakhramnev, S.A.: Hilbert manifolds with corners of finite codimension and the theory of optimal control, *Journal of Mathematical Sciences*, **53**, 176–223 (1991)
30. Vakhramnev, S.A.: Morse theory and the Lyusternik-Shnirelman theory in geometric control theory, *Journal of Mathematical Sciences*, **71**, 2434–2485 (1994)

“This page left intentionally blank.”

Chapter 26

The Aeroservoelasticity Qualification Process in Alenia

Vincenzo Vaccaro

Abstract Since the beginning of aviation history, flutter was recognized as a catastrophic event for an aircraft, consequence of a dynamic instability when speed is too high, due to the interaction between the aeroelastic forces acting on wing and tail and their dynamic characteristics.

With the progress of aircraft design, aircraft speed increased and the consolidation of the monoplane configuration led to a lower torsional stiffness of wings than biplanes. The consequence was that flutter became a serious problem, as demonstrated by the series of accidents usually accompanied by the destruction of the machine. It was evident the necessity was to develop theories and methodologies able to predict the phenomenon and to validate the design for the maximum speed to be achieved. This development is still continuing and has provided the modern aeronautical engineer with very powerful tools but still not sufficient to exclude specific flight trials aimed to confirm the prediction of aircraft speed margins before flutter.

This chapter has been intended to explain flutter to an audience with limited knowledge on aeronautics and to present the design and verification processes followed by engineers to produce aircraft that, accordingly to aeronautical regulations, are free from flutter within their flight envelope.

26.1 Introduction

The design of aircraft has progressed together with the capability to simulate the aerodynamic forces that allow an object heavier than air to fly. Very complex mathematic formulations have been developed to achieve this goal, and the progress in numerical analysis and computational power has led to an optimization of shapes and structure for minimum drag and weight objectives.

Vincenzo Vaccaro
Alenia Aeronautica Corso Marche 41, 10146 Turin, Italy,
e-mail: vvaccaro@aeronautica.alenia.it

These formulations and computational tools are even more sophisticated when applied to aeroelasticity, where the mutual influence between the airframe elasticity and the aerodynamic forces has to be simulated.

Aim of this chapter is to give a very general overview on the process of aeroelastic qualification of an aircraft, from the design to ground and flight testing. This will be done presenting the experience matured in this field by Alenia Aeronautica.

26.2 Company Presentation

With over 90 years of history and successes, today Alenia Aeronautica is the Italian leader in the aeronautics sector and a major international player in the aerospace industry, with full system development and integration capabilities in the most advanced fields, including high-performance combat aircraft, military and commercial transport aircraft, unmanned aerial vehicles, mission systems aircraft, advanced aerostructures for airliners, overhaul, maintenance and modification of military and commercial aircraft. With high investments in research and development and strong and established relationships with the other key aerospace players both in Europe and in North America, Alenia Aeronautica is a renowned player in world leading programs, both military and commercial, including C-27J, Eurofighter Typhoon, ATR regional aircraft, Boeing 787, Lockheed Martin Joint Strike Fighter, Airbus A380, and UAV (Unmanned Aerial Vehicle) technological demonstrator programs such as Sky-X, Molynx and Neuron. Its global positioning is further reinforced by the continuous support of its parent company Finmeccanica, a key global aerospace and defense player operating today through its products in more than 60 countries worldwide. At the same time Alenia Aeronautica offers its customers a comprehensive and attentive support service. Thanks to its strategy of selective competition, Alenia Aeronautica is the effective, efficient and flexible risk-sharing partner that creates value for its shareholders while strengthening its position in crucial business areas, where it can apply its core competencies such as advanced systems and aerostructures. Today, Alenia Aeronautica can rely on a first-class international network of subsidiaries, including Alenia Aermacchi, Alenia Aeronavali, Alenia SIA, Quadrics, Alenia North America, Alenia Hellas and joint ventures like Eurofighter (with BAE Systems and EADS), ATR (with EADS), GMAS (with L3 Communications) and Global Aeronautica (with Vought Industries), and long-established relationships with Boeing and Lockheed Martin. In 2005 Alenia Aeronautica recorded consolidated revenues in excess of 2.046 billion Euro with a workforce of 11,200 skilled technicians distributed among several advanced Centres of Excellence in Italy and abroad, working every day to secure a continuing success.

Technology is a key element in the Alenia Aeronautica goal of providing reliable and cost-effective solutions to market requirements and complex challenges. An experienced integrator of complex systems, Alenia Aeronautica uses innovative product and process technologies and invests considerable resources in personnel training and technology research. In 2005 the company invested 25% of turnover

in R&D and self-funded over half the cost of its about 50 current research projects, including activities related to the Sky-X UAV technological demonstrator. Alenia Aeronautica research focuses on technologies most relevant to business priorities and cooperative opportunities. The main R&D areas are related to aircraft system design and integration with particular attention to UAV aspects and technologies for aerostructures. System integration R&D covers the entire life cycle, from concept to design, production and certification, and aerostructures research concentrates on materials, structural architectures, non-destructive tests (NDT), and industrial processes. Moreover Alenia benefits from its cross-knowledge opportunity to apply dual technologies (civil and military) to all its business lines. Alenia Aeronautica is at the leading edge of design. Product configuration data and life-cycle traceability allow effective fleet support. The company has extensive competencies in complex flight system simulation, including pilot in the loop. Together, these unique skills lead to shorter development time, faster access to flight test and early achievement of operational readiness. The company has a wide range of testing systems and facilities to support research and technology development. These include laboratories for aeromechanical, structural and system tests, equipment acceptance, EMC tests facilities, integration rigs, the Sky-light simulator to support visual human-machine interface of advanced cockpits and flight simulators to support entire aircraft life cycle. The flight simulators are an integral part of company's synthetic environment, an extended network of applications, models, simulation, equipment that form a virtual representation of the real world for both manned and unmanned aircraft. To remain at the forefront of technology, Alenia Aeronautica constantly monitors and improves its product design and development processes competencies and trains its staff in new technologies and processes in cooperation with leading Italian universities.

At Turin-Caselle, Alenia Aeronautica operates a full-fledged flight test centre, capable of performing experimental and production flight tests. The Alenia Aeronautica Eurofighter, C-27J and AMX simulators attest to the company's ability to model complex situations effectively. This affords a better understanding of project specifications and the ability to study the product's operational modes. Alenia has an established, well-structured collaboration research activity network with major national and international aerospace centres, universities and aerospace companies.

26.3 What Is Aeroelasticity

Aeroelasticity is the discipline devoted to the study of the mutual influence between the elasticity of an aircraft structural component and the aerodynamic forces that are applied to it. Aerodynamic forces change because of the deformation of the structure that, on the other hand, depends on the loads that are consequent. As a result of this coupling there are some characteristics of the aircraft that have to be assessed and verified, since potentially catastrophic owing to an unstable response of the structure under the action of the changing aerodynamic forces. In particular,

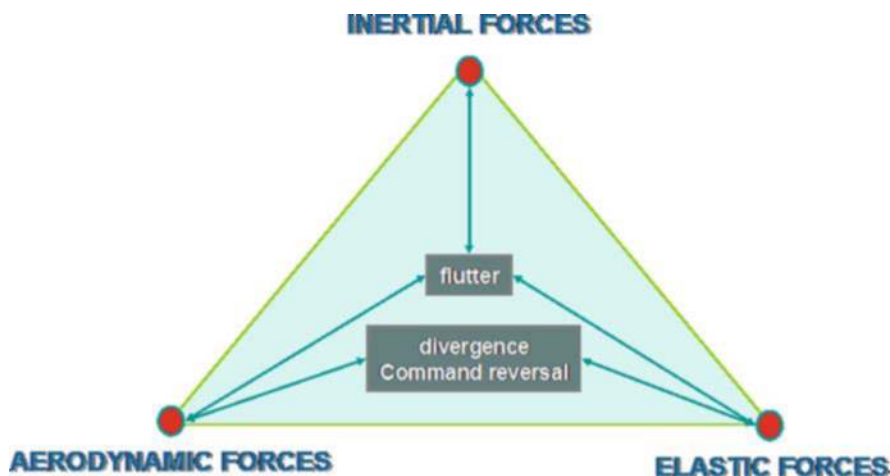


Fig. 26.1 The aeroelasticity triangle

the effect of the aerodynamic forces increases with speed and there is a limit that, if overcome, leads to an unstable equilibrium among the aerodynamic, the elastic and inertia forces that determine the response. Figure 26.1 illustrates these concepts.

A further complication has been added by the introduction of the fly-by-wire concept in modern aircraft, which implies a coupling also with the FCS, that picks up the dynamic response of the airframe through its sensors and reacts exciting the structure by the consequent deflection of the control surfaces. Aeroservoelasticity is the new discipline that treats these aspects.

Therefore, in the design of an aircraft it is essential to guarantee that the speed at which aeroelastic phenomena can occur has adequate margins with respect to the maximum speed achievable by the aircraft.

The theoretical prediction of phenomena like wing divergence, control reversal and flutter is rather complex and based on structural and aerodynamic models that have become more and more sophisticated and realistic, requiring significant computational resources.

Notwithstanding the high level of fidelity achieved by modern models a validation through ground testing activities and a final verification in flight are still necessary to demonstrate that a new aircraft is free from aeroelastic instabilities. However, thanks to the high level of reliability achieved by the aeroelastic analysis, the amount of testing has been reduced as well as the risk and the cost of all the certification process.

26.4 Aeroelastic Tradition in Alenia

Aeroelasticity has always been a discipline to which the company has devoted a lot of efforts, since the beginning of its history. As a result, a strong competence in this



Fig. 26.2 Alenia responsibilities for aeroelastic aspects in the most important projects

field has been achieved, thanks to the transmission of a state-of-the-art know-how to young engineers, covering analysis, ground resonance test and flutter trials.

Without going too back in time, the progresses made in the last 30 years and the experience and competence matured on several military and civil programs have put the Alenia aeroelastic group in the condition to always play a role of high responsibility in the aeroelastic qualification of an aircraft. A summary of the most significant contributions has been made and illustrated in Fig. 26.2. They cover transport, military aircraft and UAVs programmes that have seen Alenia as owner of the project or prime partner in the last years.

26.5 Aeroservoelastic Certification Process

The certification process consists in the demonstration by analysis, ground and flight testing that the system is compliant with specific requirements of civil or military regulations. In the case of aeroelasticity this means that the aircraft shall be free from flutter or any other aeroelastic instability within its flight envelope, extended with some margins. The structural coupling with the FCS has to be attenuated to levels that are acceptable for the stability and control of the aircraft.

This process is rather long and starts during the first loop of design of the aircraft, when a preliminary structural layout is available. In the past the approach was to design the airframe with additional margins in order to reduce the risk of aeroelastic instabilities. The consequence was a heavier aircraft and the aeroelastic certification process merely a validation of the design. The modern design is instead based on a multidisciplinary optimized design that allows to consider aeroelastic constraints

since the beginning. The advantage is to have a less conservative design and, as a consequence, a weight saving. Of course, the MDO approach requires a significant know-how and the availability of codes, computational tools and models at the state of the art. The process is also much more complex owing to the necessity of integrating several specialists (structure, aerodynamics, aeroelasticity, FCS, weights, etc.) in a single team, overlapping their competences.

Models and analysis are not sufficient for the certification. Even considering that probability of occurrence of phenomena like flutter or divergence is quite low in the modern aircraft design, their impact on structural integrity would be catastrophic and a confirmation of theoretical predictions by ground and flight testing is necessary to exclude this risk.

All civil (FAR, EASA) and military regulations (MIL) define the process that has to be followed to certify an aircraft for aeroelastic aspects. This process is quite similar for both and is based on analytical predictions of flutter speed for several aircraft configurations (fuel, payload, external stores) and flight conditions (Mach number and altitude), GRT and flutter trials for the most significant cases.

26.5.1 Analytical Models

A basilar capability in the qualification process is that of simulating aeroelastic phenomena. This means that a specific model has to be developed in order to simulate, at a proper level of fidelity, the dynamic characteristics of the airframe and the aerodynamic forces that act on the airframe. It is much simpler to talk about aeroelasticity referring to the most significant component of an aircraft: the wing. Nowadays, a typical aeroelastic model is obtained coupling a finite element model (FEM) to an aerodynamic mesh. Coupling these two elements means that the modelling allows to integrate aerodynamic pressures on each panel of the mesh and to transfer the resulting forces and moments to the grids of the structural elements. On the other direction, the same modelling permits to transfer the displacements of the structural grids to the aerodynamic mesh. Once the mutual influence between structural deformations and aerodynamic forces is implemented it is possible to simulate typical response analyses, both in time and in frequency domain, and eigenvalue calculations to assess intrinsic stability of the system. The level of fidelity and complexity of the models used depends on the characteristics of the aircraft. If the flight envelope is limited in Mach number well below the transonic range ($Mach \leq 0.7$) a linear DLM model is sufficient to simulate the unsteady aerodynamics. The simplicity of the mesh (no profile effects using plane boxes and slender bodies) and the fast generation of aero data using small CPU capabilities are the most attractive characteristics of this sort of model. But when the Mach number gets higher, and this is the typical case of modern liners, it is necessary to use much more complex models based on CFD analysis. The aerodynamic mesh will be very fine, simulating the actual external surface of the wing and giving the possibility to take into account compressibility and viscous effect that, if neglected, would lead

to non-conservative aeroelastic predictions. The amount of time for the preparation and validation of the model and the CPU power required make this analysis very complex, long and expensive. From the structural point of view, the complexity of the model depends on the layout of the main components. A slender wing can be easily simulated by a beam of proper elastic characteristics and mass distribution. For delta or low aspect ratio wings, typical of combat aircraft, the model cannot be so simple and the main elements of the structure (spars, ribs, skin) are to be represented. Depending on the complexity of the aero and structural models, the coupling between the two meshes can be a very difficult challenge. This is because the two meshes are designed following very different criteria and, as a consequence, they have no grids in common and very complex tools of geometric interpolation must be developed. Whilst for the FEM there are very valid commercial products (NASTRAN, ELFINI) diffused in the aeronautical industry, the same cannot be said for CFD tools. Most of CFD programs used today in the main industries have been developed and validated in home together with the geometric interpolation tools and the response simulation program. Different models can be used during the development of the project, depending on the objective of the analysis. When a lot of parametric analyses are to be performed the current standard is to use simple DLM models, referring to CFD only when the design is well consolidated and a refined analysis is necessary. Figure 26.3 gives an idea of the typology of models that are used to build up an aeroservoelastic model, highlighting the interaction among different types of mathematical means and the strong multidisciplinary characteristics of this discipline.

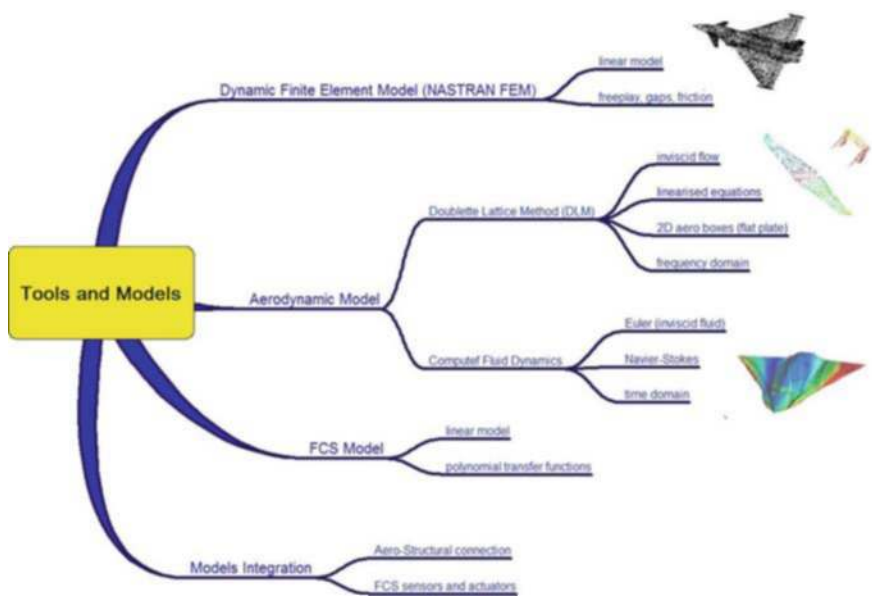


Fig. 26.3 Overview of models used in aeroservoelasticity

26.5.2 Theoretical Background

Aeroelastic equations of motion, in the frequency domain, are obtained by means of the Laplace transform of the equations of motion in generalized coordinates of a dynamic system excited by aerodynamic forces. The following system of equations is the extreme synthesis of basis for aeroelastic analysis:

$$([M_{hh}]s^2 + [B_{hh}]s + [K_{hh}] + q_\infty[Q_{hh}(s)])\{\xi(s)\} = 0 \quad (26.1)$$

where

- $[M_{hh}]$, $[B_{hh}]$, $[K_{hh}]$ are the mass, damping and stiffness generalised matrices;
- $[Q_{hh}(s)]$ generalised aerodynamic matrix;
- $q_\infty = 1/2\rho V^2$ dynamic pressure (ρ is the air density and V the airspeed);
- $s = \sigma + j\omega$ Laplace variable;
- $\xi(s)$ modal displacements.

Aerodynamic matrices are approximated by rational expressions after the introduction of the nondimensional Laplace variable p and of nondimensional reduced frequency k .

$$p = sL/V = g + ik, \quad k = \omega L/V \quad (26.2)$$

where

- V airspeed;
- ω frequency (rad/s);
- L reference length.

$$[Q_{hh}(s)] = [A_0] + L/V[A_1]s + L_2/V_2[A_2]s^2 + [D]([I]s - V/L[R]) - 1[E]s \quad (26.3)$$

Using the approximation for the aerodynamic matrices the equations of motion can be transformed in a state-space form, where it is possible to easily include control terms.

$$\begin{cases} \{dX/dt\} = [A]\{X\} + [B]\{u\} \\ \{Y\} = [C]\{X\} + [D]\{u\} \end{cases} \quad (26.4)$$

The first set of equations representing the equation of motion of the aeroservoelastic system, the second ones a set of output parameters expressed in terms of the plant states and control variables.

State-Space formulation is the most common and consolidated approach to aeroservoelastic analysis when linear aerodynamic assumptions are applicable. State-Space formulation is the most common and consolidated approach to aeroservoelastic analysis when linear aerodynamic assumptions are applicable. The generation of matrices $[A_0]$, $[A_1]$, $[A_2]$, etc., needed for the approximation of aerodynamic contributions, is carried out using tabular values of $[Q_{hh}]$ available for various reduced frequency values and then proceeding by a least squares approach.

When aerodynamic linear assumptions are not applicable (transonic range, turbulent flow, wide deformations) the CFD approach is usually followed. Models are, in this case, much more complex, time domain simulations are usually performed coupling the aerodynamic mesh to the structural model and computational costs are very high, requiring high-performance machines and long CPU time. The application of CFD is therefore limited to a selection of significant cases and results quite often used to correct linear models.

CFD tools are based on the discretisation of fluid dynamic equations (Euler and Navier–Stokes equations) on the nodes of a support grid. The main role of CFD is to reduce wind tunnel and flight test activities, and relevant costs, through the use of a very sophisticated and reliable aerodynamic theoretical tool. This has a consequence on global product development costs, with significant reductions, and a shorter time to market of the product. The role of CFD is, therefore, to complement with wind tunnel and flight testing that anyhow are fundamental and necessary for the final validation of the tool. Figure 26.4 shows an example of CFD application, with a grid mesh of the aerodynamic surface, a plot of the pressure distribution on the surface and some comparison, at specific wing stations, of CFD data versus experimental ones.

For aeroelastic studies, using CFD as aerodynamic data generator, it is necessary to couple the aerodynamic mesh to the structural mesh, in order to transfer displacements and forces between the two and then simulate the aeroelastic phenomena. Moreover, the main step to perform aeroelastic simulations using CFD is the integration of aerodynamics and structural dynamic solvers.

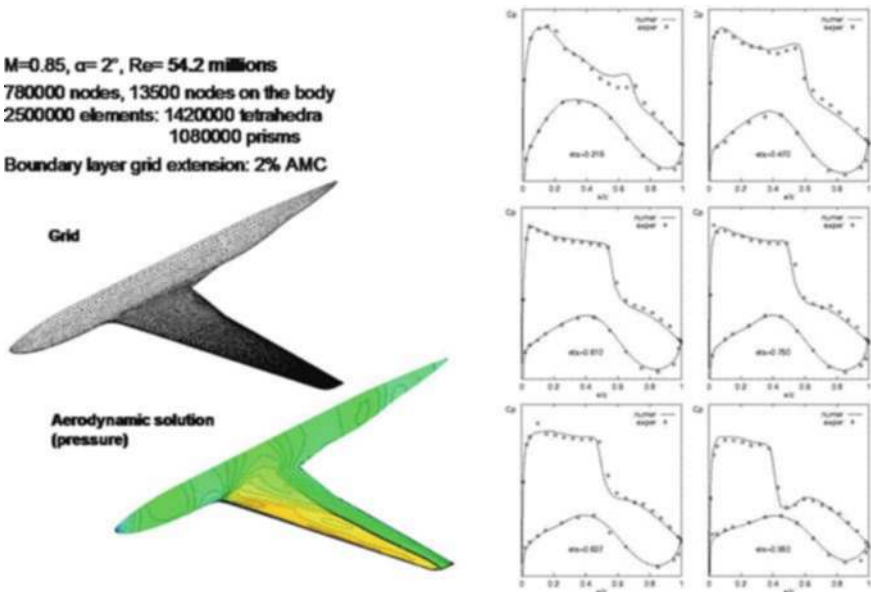


Fig. 26.4 Example of CFD data compared with wind tunnel test results

26.5.3 Ground Test

Notwithstanding the high level of fidelity achieved by modern models there is always the necessity to validate the most significant results by test results. The ground test activities are the first step of validation before the first flight.

To solve all the uncertainties inherent in the dynamic model a GRT on the complete aircraft is usually performed. Modal frequencies and shapes are measured and compared with theoretical predictions, in order to upgrade the model for an adequate representation of the real aircraft.

In some cases, especially for novel aircraft configurations that present significant uncertainties from the aeroelastic point of view, a flutter wind tunnel model is a solution to reduce the risk for the program. The model is scaled in terms of not only geometry but also dynamic characteristics. This means that the model has to be flexible and with a proper mass distribution, in order to simulate in the wind tunnel the aeroelastic behaviour of the full-scale aircraft. The results of the wind tunnel test will be used to assess whether the aerodynamics used for the flutter predictions are correct or need to be corrected to get a simulation closer to test results.

The combination of GRT and wind tunnel test results should solve most of the model uncertainties and allows the release of safe first flight clearances.

The GRT is a qualification step required by all aeronautical regulations for new aircraft or when changes that can have impact on flutter are introduced in the aircraft layout.

26.5.4 Flight Test

Owing to the high risk that flutter, being potentially catastrophic, represents for an aeronautical programme a prudent and controlled expansion of the speed envelope has to be performed, besides all the analytical and ground test activities already carried out. For this scope it is required to instrument the aircraft tasked for the flutter trials with a set of accelerometers and a monitoring system, in order to check, in real time during the speed envelope expansion, the dynamic response of the wing to sudden (abrupt pilot command) or tuned excitations (control surface oscillations). Similarly to the GRT, also flutter trials are required by regulations and represent the final demonstration that the aircraft is free from any aeroelastic instability inside its flight envelope. Figure 26.5 shows a general overview of the links among the disciplines and activities mentioned before.

26.5.5 Research and Future Developments

There are several research themes dedicated to improve the reliability of aeroelastic predictions and the design of aircraft to reduce weight. In fact, the most common solution adopted to attenuate aeroelastic effects is to increase the wing stiffness, with consequent weight increment.

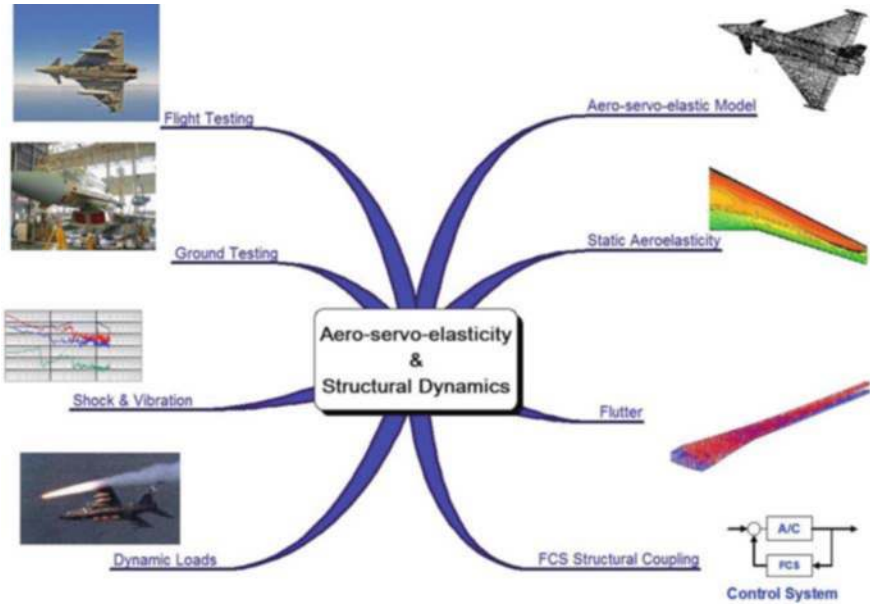


Fig. 26.5 Overview of aeroservoelastic Links

A more reliable flutter prediction is seen as a benefit in terms of risk and cost reduction, since it would be possible to reduce the amount of test activities and explore complex cases difficult to test using the validated model (failure cases, flight conditions difficult for testing, etc.).

The weight saving improves the aircraft efficiency, reduces the fuel consumption with saving on the operational cost and impact on environment and represents a very ambitious target for future projects. The availability of sophisticated models is essential to carry out an aeroelastic optimization using weight as a cost function. The main field of research to improve models is the application of CFD to aeroelasticity and the validation of these tools using wind tunnel test results. The inclusion of structural non-linearity is another topic that attracts a lot of interest, above all for the solution of limit cycle oscillation phenomena that occur in the transonic regime. Parallel to the development of these models there are plenty of studies on new mathematical methods and numerical tools that make the calculation feasible and more efficient. Quite often, in fact, the computational power required is very high and a reduction of its cost, together with the computer requirements, is the only way to render this approach affordable and fast enough for large-scale industrial applications.

An important tool to validate CFD for aeroelastic applications is to use a wind tunnel model that, besides to replicate the aerodynamic shape of the body, replicates also the structural dynamics and is instrumented in order to measure both dynamic response (accelerometers) and unsteady aerodynamic pressure (kulite). Figures 26.6 and 26.7 show a sketch of a wind tunnel model designed for flutter investigation in transonic (left) and the model, built, assembled and ready for wind tunnel testing.

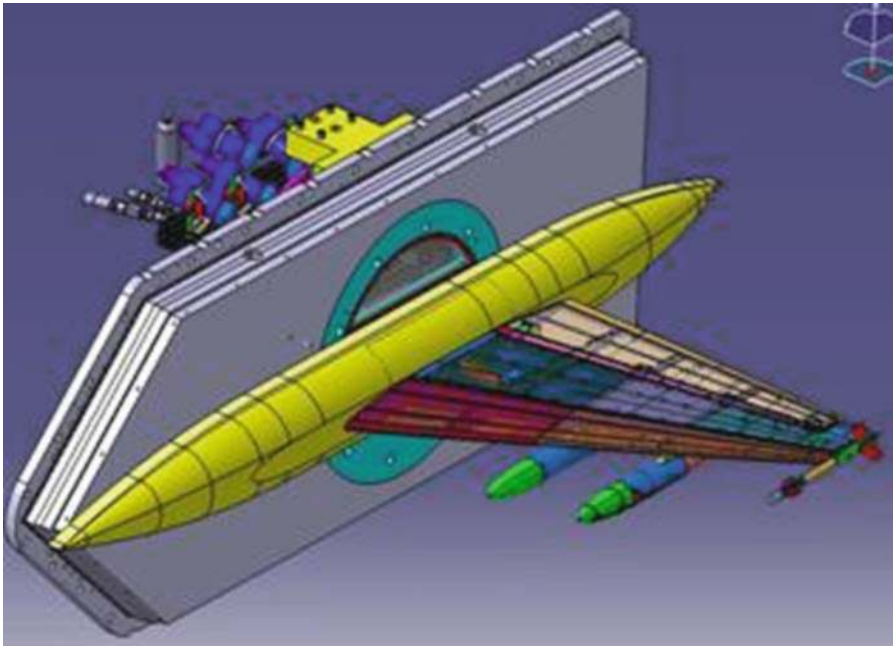


Fig. 26.6 Wind tunnel models for flutter investigation



Fig. 26.7 Wind tunnel models for flutter investigation

In the past, the way to consider aeroelasticity at the beginning of a project was that of applying to design loads by a factor between 1.2 and 1.5, design the airframe structure using these loads and then by verifying the aeroelastic characteristics at the end of the design. This process does not lead to a design that is efficient for the weight of the aircraft. Another field of research, that aims to consider aeroelasticity since the preliminary phase of the design in a more efficient way, is the multidisciplinary optimization or MDO. The new philosophy is based on the availability of reliable and efficient models of the aircraft that allow to design the structure applying loads and considering aeroelastic characteristics at the same time. The target of the optimization process is to produce a structure that can sustain the design loads, satisfies aeroelastic constraints (no instabilities or excessive deformations under manoeuvre loads that can compromise the aerodynamic or control efficiency) and has the minimum weight. Also in this case, the quite advanced state of the art achieved in this field has been possible thanks to the parallel development of mathematical and numerical methods that with their application have made this approach feasible and with promising results.

“This page left intentionally blank.”

Chapter 27

Further Steps Towards Quantitative Conceptual Aircraft Design

Michel van Tooren, Gianfranco La Rocca and Teodor Chiciudean

Abstract To cope with the large growth of air transport and the tightening requirements on noise and emissions, it is expected that radical new aircraft concepts will be needed. The design of such new concepts will require high-fidelity computational systems to support the designer in his search through the design space. The design and engineering engine (DEE) is a concept for such computational systems that offers generative distributed product modelling based on knowledge-based engineering, design domain search based on multi-disciplinary design optimization principles and initial design vector determination based on the feasibility principle and agent technology to couple the components of the DEE in a flexible and distributed fashion. The different components have been successfully tested in several small-scale projects.

27.1 Introduction

Many societal changes will be required in the coming decades to sustain the Earth. Science and technology will play an important role to provide new insights in the status and the future availability of natural resources and will have to supply new solutions to human needs for food, transport, energy and leisure. Also aviation will be challenged heavily. The expected global growth in transport of passengers

Michel van Tooren

Delft University of Technology, Kluyverweg 1, 2629 HS Delft, The Netherlands,
e-mail: m.j.l.vantooren@tudelft.nl

Gianfranco La Rocca

Delft University of Technology, Kluyverweg 1, 2629 HS Delft, The Netherlands,
e-mail: g.larocca@tudelft.nl

Teodor Chiciudean

Delft University of Technology, Kluyverweg 1, 2629 HS Delft, The Netherlands,
e-mail: t.g.chiciudean@tudelft.nl

and freight is such that the air transport system will need severe changes in all its elements. Vision 2020 states that for the age of sustainable growth the required improvements in quality, performance, safety, security and economy combined with the need for reduced environmental impact and reduced impact of accidents will need a future aircraft system that differs as much from the current aircraft system as the current system differs from that of 1930. This can be interpreted as a need for breakthrough technology and radical new concepts.

Since the publication of Vision 2020 several new aircraft programmes have been launched. The presented generations of aircraft already show some new technologies (see Table 27.1). The question is, are current steps sufficient to meet the future demands or do we need much more radical changes?

Table 27.1 Some features of the next generation of civil transport aircraft

Subsystems/aspects	New PAX aircraft	VLJs
Architecture	Conventional: long almost cylindrical fuselage, low wing, VTP, low HTP, wing-mounted podded engines, tricycle landing gear	Conventional: (almost) cylindrical fuselage, low wing, T-tail, rear fuselage mounted podded engines, tricycle landing gear
Airframe	Hybrid: composite shells, metal (Al and Ti) ribs, frames, brackets, joints and mounts. Low drag paint. Lean approach towards production	Different solutions from mainly composite to friction stir welded metal. Production aiming at large numbers, lean approach
Powerplant	Bleed-air-free, low-specific fuel consumption, low emission, low noise. More electric aircraft	Conventional turbofan
Flight control system	Flight-by-wire, vertical gust suppression	Mechanical controls
Environmental control system	Bleed-air-free. Cabin pressure altitude 6000 ft	Bleed-air-based. Sea level cabin upto 21500 ft

If we focus on the architecture, there are many alternatives under study. To name a few:

- Prandtl planes, aiming for minimized induced drag, Fig. 27.1
- Blended wing bodies, aiming for minimized wetted surface, Fig. 27.2
- Oblique flying wings, aiming for low induced drag, low parasite/wave drag, Fig. 27.3

Even more radical designs are possible. An example is the Aerodina Lenticulara, Fig. 27.4 and of Henri Coanda, Fig. 27.5. He had his own vision 2020 already 40 years ago. In 1967, at a Symposium organized by the Romanian Academy he said,

These airplanes we have today are no more than a perfection of a toy made of paper children use to play with. My opinion is we should search for a completely different flying machine,



Fig. 27.1 The Pisa University PrandtlPlane concept

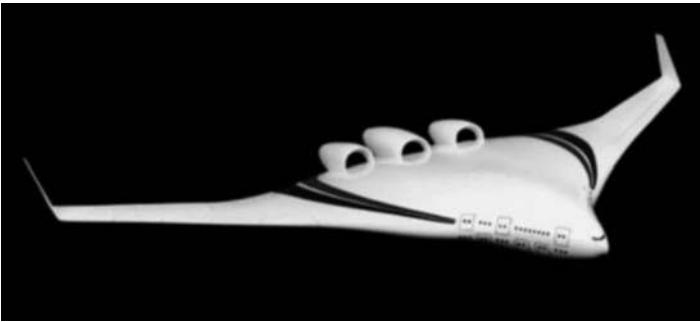


Fig. 27.2 The Boeing Blended Wing Body aircraft concept



Fig. 27.3 United States Defence Advanced Research Projects Agency (DARPA) oblique wing concept



Fig. 27.4 The Aerodina Lenticulara: VTOL using the Coanda effect, normal cruise



Fig. 27.5 Henri Coanda, the Romanian aircraft design pioneer

based on other flying principles. I consider the aircraft of the future, that which will take off vertically, fly as usual and land vertically. This flying machine should have no parts in movement.

Coanda's vision is very hard to embody. Especially the vertical take-off and landing and no parts in movement are far from trivial. Only rocket propulsion combined with ramjets and scramjets could give us the required propulsion system, not an easy

acceptable and feasible design option. Doing this for passenger aircraft will require radical different configurations.

These alternative concepts and many other design questions like

- Why bring your landing gear along if you can make high precision landings?
- Why use podded engines?
- Why still fly with a pilot?
- Is sitting the best position for a passenger?

are all very interesting but apparently not attractive enough to result in an intensive search for their implementation. There is no (civil) application of the alternative configurations/solutions mentioned above so far. What we need is a good way to find out what all the proposed paradigm shifts could bring us. Since trying out all the different concepts full scale is an unaffordable option, we need to test these ideas in a virtual world. Although the current optimization algorithms are very powerful we are lacking a proper way to feed these algorithms with sound parametric multi-disciplinary product descriptions in an automated way. If we solve this issue we can take a next step towards higher fidelity virtual product development. The cost of design and engineering will significantly reduce and by the improved search of the design space we enhance the probability to find real improvements with respect to noise, drag, weight, power plant efficiency and required airport space, compared to current dominant designs.

The proposed solution is an efficient computational system that supports the systems engineering approach applied in aerospace industry. It provides a means to

- explore design domains
- increase yield of current analysis tools potential
- free human minds for creative activities
- reduce cost of engineering

It will be a tool to challenge the problems ahead and to investigate new design options, which are hard to try at first in real aircraft. This virtual search in the design domain (generalized multi-disciplinary design optimization) is a way to reduce schedule and budget risks.

Designing is moving around, with a high level of uncertainty, in an object-based space. In that sense it is like variational analysis. The initial system, believed to comply with the design requirements, is synthesized by human creativity. The best state of this system is found using optimization techniques, so the design problem can be seen as an object-based variational analysis. The determination of the following two sensitivities is essential to find a proper solution:

$$\frac{\partial \text{preferred_design_concept}}{\partial \text{requirement}_i}$$

$$\frac{\partial \text{preferred_design_concept}}{\partial \text{available_design_option}_i}$$

27.2 The Systems Engineering Approach

The aircraft industry uses a more or less standardized approach to structure aircraft development programs. The approach is called systems engineering and is using the following assumptions:

- Each complex product can be seen as a system.
- Each system has a life cycle consisting of different phases:
 - it is conceived (starting with a need (market) or a seed (invention))
 - it grows (design and engineering)
 - it is born and raised (production)
 - it has a professional life (operation)
 - it needs care (support and maintenance)
 - it dies (phase out + re-use/re-cycling)
- Each life cycle phase generates requirements that have to be taken into account during the design and engineering phase of the complex product.
- Each of the life cycle phases can require the development of related systems.

This process forces a design team to express proper requirements for each life cycle phase, to have design options for each of these requirements and to be able to assess the behaviour of a system synthesized from these options in order to compare the system's behaviour with the requirements and value each result of the search (Design for X).

27.3 Requirements on Computational Systems

The last century has brought large-scale automation in industrial production to cope with increasing labour cost. Similar automation will have to take place in design and engineering. The introduction of computers in the last decades of the last century was related to almost every domain of human activity and was often sold under the flag of automation but it turned out to be more of a strong stimulus for new activities and generation of additional information, often without a link to the initiating problem. The last decade, the competition in human costs has entered also the aircraft design and engineering practice. Therefore we need to have a second look and split the activities in engineering in repetitive and creative elements. In this way we can differentiate in non-recurring and recurring engineering costs and see if we can reduce the recurring costs by automation.

A computerized SE-support system will have to control the information burst, automate the current repetitive human actions and automate the human interaction with tools as found in current SE practice. It therefore needs to

- allow quantitative specification of requirements from the different life cycle phases

- allow distributed parametric multi-disciplinary descriptions of the design options (concepts) for all the system elements
- estimate initial values for design parameters and variables (named feasilization)
- provide automatic search for optimal values of design parameter and variables
- provide automatic analysis model input
- provide linkage to distributed analysis tools to derive all the properties of the system related to the requirements from the different life cycle phases
- provide automatic interpretation of analysis output results
- offer flawless connections between all its elements

The development of these features needs a very thorough understanding of how people currently act and connect to perform their role in the systems engineering process. Only in that way the resulting computational system will successfully support the exploitation of the multi-disciplinary design optimization (MDO) methodology in a distributed design environment.

27.4 The Design and Engineering Engine Concept

The computational system, baptized design and engineering engine (DEE), is constructed such that it resembles the working of a human team in a design process. Its formal definition is A DEE is an advanced design system to support and accelerate the design process of complex products, through the automation of non-creative and repetitive design activities [1]. A DEE consists of a multi-disciplinary collection of design and analysis tools, able to automatically interface and exchange data and information, Fig 27.6.

The proposed DEE's main components are

- (i). Reqs/design options specifier
- (ii). Initiator
- (iii). multi-model generator (MMG)
- (iv). Expert tools covering all the analyses required to derive the behaviour of the system in the different life cycle phases
- (v). Converger and evaluator (the global optimizer)
- (vi). An (agent-based) framework

Each of these components is discussed in the following subsections.

27.4.1 Describing Design Options

Each system is supposed to be a synthesis of design options selected to meet a set of system and subsystem requirements. The user of the DEE can specify which design domain he wants to be searched, using the so-called high-level primitives (HLPs) as basic building blocks [4]. A primitive is a consistent set of parameters/

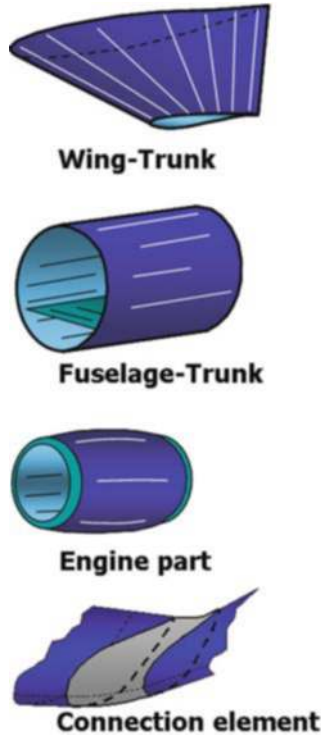


Fig. 27.7 Some high-level primitives used in a DEE for conceptual aircraft design

27.4.1.1 Describing Requirements

Requirements are translated into an objective function, constraint functions and bounds on design parameters/variables. Each of these functions and bounds is based on the parameters and variables in the HLPs or on behaviour computed using an instantiation of the HLPs. Examples can be a minimization of noise, expressed in awakenings, annoyance or money. For each potential design solution under consideration a description is made using HLPs. The HLPs should be such that the parameters and variables involved are sufficient to allow calculations of the behaviour related to the objective function. The same is valid for constraint functions, e.g. deliver enough lift, limit cost to, etc. The third way of expressing requirements is to specify bounds on values of the parameters/variables (e.g. size of wing span) in the HLPs. The “degrees of freedom” of the HLPs and/or combinations thereof are the variables/parameters within the DEE process.

To be able to use the product model defined by a set of HLPs, for example, to visualize the result or to transfer knowledge about the product to an analysis tool, so-called capability modules (CMs) are defined [4]. The capability modules are formalized versions of engineering operations normally done by humans on HLPs,

like generating 3D views or meshing surfaces to prepare for FE analysis, etc. They will be discussed in more detail in the section about the multi-model generator.

How to define and implement the primitives? The current available primitives are implemented using knowledge-based engineering (KBE). This is a technology based on the use of dedicated software tools (i.e. KBE systems) that are able to capture and re-use product and process engineering knowledge [8]. Instead of “drawing” the engineer “describes” his ideas in a collection of objects. The following five important “lowest common denominator” features are intrinsic in any generative KBE system:

- (i). Functional coding style: Programs return values, rather than modifying things in memory or in the model.
- (ii). Declarative coding style: There is no “begin” or “end” to a KBE model – only a description of the items to be modelled.
- (iii). Runtime value caching and dependency tracking: The system computes and memorizes those things which are required – and only those things which are required (no more, no less).
- (v). Dynamic data types: Slot values and object types do not have to be specified ahead of time. They are inferred automatically at runtime from the instantiated data. Their data types can also change at runtime. In fact, the entire structure and topology of a model tree can change, depending on the inputs.
- (vi). Automatic memory management: When an object or a piece of data is no longer accessible to the system, the runtime environment automatically reclaims it.

These conditions are not met by (most) parametric CAD systems today. They still are geometry focussed and more fit to record a finalized design than to build parametric models to start a design. Using a KBE system the designer describes his idea to the computer as a set of objects. Very important for successful usage of KBE systems is that the right knowledge about product and processes is elicited and captured before the actual coding is done. Therefore formalized methods, so-called knowledge acquisition techniques, are employed [5].

In practice the HLPs are implemented as classes as defined by the sample pseudo-code below:

```
(define-object-class classname (Mixin-list)
:input-slots (variablenames1)
:computed-slots (variablenames2 (expressions leading to a value for variable3))
:objects (childnames :type)
:functions (functionnames (operations)))
```

Class-name: For each HLP a descriptive name is chosen. The current system has wing trunks, connection elements, fuselage trunks and many others.

Mixin-list: This is a list of superclasses or other classes from which the class here specified will inherit all the characteristics (attributes and components). The classes

specified here can either be formal superclasses (of which the class specified in the define-object is an actual specialization), or just other classes with which this class is to share attributes and parts.

Input-slots: This is a list of parameters to be assigned in order to generate an instantiation of the given class. This set of parameters represents the so-called “class protocol.”

Computed-slots: These slots are generally expressions which return a value when computed. These expressions can either be production rules or any other mathematical, logic or engineering rule. To compute these expressions it is possible to use and combine values of other slots, such as the input-slots, or the slots inherited by the classes specified in the mixin-list or slots of the children defined in the given define-object-class.

Objects: This is the list of objects that may be contained in an instance of the define-object. They are also called “children” of the define-object-class instance. For each object, the following must be specified:

- the object name
- the name of the relative class to be instantiated (by using the keyword type)
- values for the input-slots required for the instantiation. This parameter list must sufficiently match the protocol of the class to be instantiated (i.e. at least its required input-slots).

In many cases the class defines how to make a 3D view of the object when instantiated.

Functions: These are similar to computed slots, but, like functions in raw Lisp, they can accept arguments, and their computed return values are not cached as is done by default with computed slots.

The capability modules are implemented in two ways. The preferred option is as Classes coupled with the HLPs through the mix-in list. In some cases it is more practical to define a CM in the function part of a HLP directly.

If we combine a proper set of HLPs and CMs we have a formal definition of a product family including the engineering processes that can be performed on this family. So instead of a drawing we can use a class diagram as defined by UML to record and communicate our thoughts. An example of such a diagram including the link to instantiations (different aircraft types) of the model is shown in Fig. 27.8.

27.4.2 The Initiator

The HLPs define an object as a consistent set of parameters and variables. To be able to start a search for the best solution within the design domain defined by the HLPs, we need a start value for each of the parameters and variables. The initiator component of the DEE is responsible for generating such start values. The process is called feasilization [2] and follows the normal behaviour of engineers, namely obtaining an approximated feasible solution to the design problem by assuming

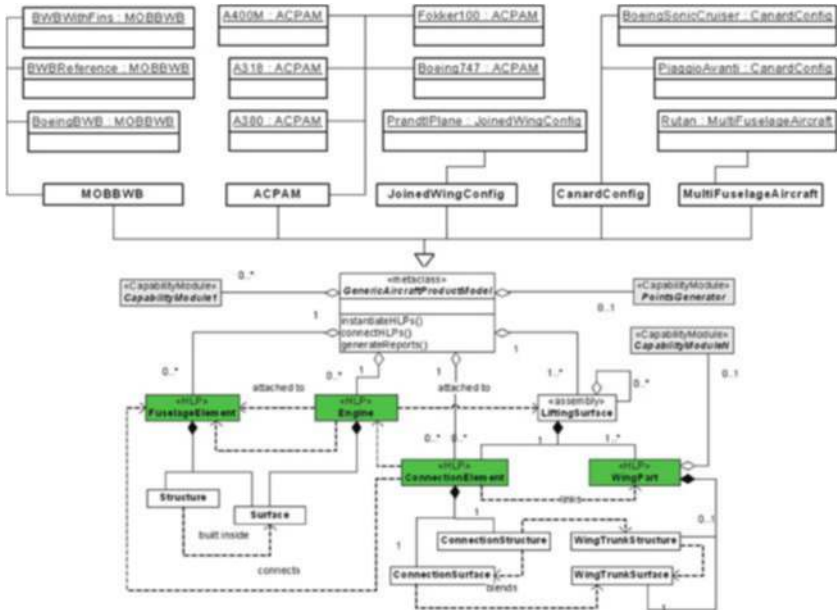


Fig. 27.8 The design and engineering engine (DEE) paradigm

- a reduced set of requirements
- an iteratively decomposed (and independent) set of sub-problems
- simplified design solutions (design options)
- simplified behavioural models (schematic models)

Considering the fact that we want to use the DEE for novel designs it is important to use first principle-based methods to estimate product behaviour and not use statistical estimates.

The feasilization process for wing-like structures can be found in [2]. During the development of this process it became clear that the feasilization requires a DEE by itself with all its components. It also became clear that it is beneficial to describe the combination of the design problem and its verified solution as a so-called generalized engineering object, Fig. 27.9. The structure of the object and the selection of its elements are based on the philosophy that for every well-engineered product it is possible to define the following relationship between requirements, design options and their fitness for purpose:

criteria (behaviour (properties (design_option), testCases)) = designValue

To obtain the DesignValue (the value of the multi-objective function describing the design problem at hand), we need criteria (requirements) with which we can compare the actual behaviour of our design. This behaviour requires the properties of the product and a set of appropriate test cases with which the behaviour can be estimated.

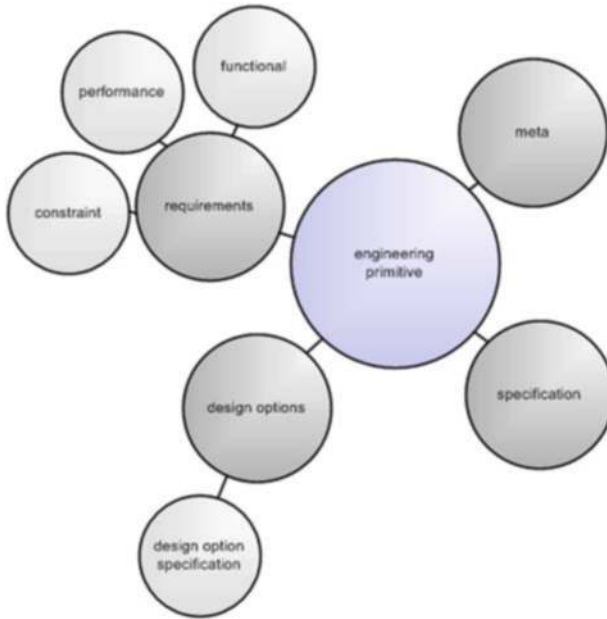


Fig. 27.9 The engineering object

The feasilization process uses this approach to give an initial value to each variable and parameter of each design option. Currently a feasilization process is being implemented for the outer surface definition of wing trunks; an adjoint formulation of an Euler code will be used for airfoil/trunk local optimization [3].

27.4.3 The Multi-model Generator

The multi-model generator is a knowledge-based engineering (KBE) application (MMG) [4] providing two functions

- (i). Support of the definition of product models, based on high-level primitives
- (ii). Support of the creation of multiple views on the product model by means of capability modules which generate input files for various analysis tools.

Especially the definition and implementation of the capability modules is a complex issue matter. Getting robust modules to prepare input for FE calculations or for cost calculations requires a thorough understanding of these processes as performed by human engineers. The specialists spent a substantial percentage of their time on inventing work-arounds to deal with program peculiarities and bugs. In general, however, a KBE system will be able to mimic this behaviour when proper knowledge acquisition has been done.

27.4.4 The Life-Cycle Analysis with Expert Tools

The determination of the system behaviour is done with a collection of analysis tools. Depending on the problem at hand the proper selection of tools is made. Since many commercial analysis tools expect a human user to communicate with the tool through an interactive interface, it is not a simple task to have tools flexibly incorporated in the DEE. As far as possible the battle of the engineers with these tools is mimicked with the capability modules. Sometimes it is necessary to add additional code between MMG and analysis tools to create a robust and sufficiently product independent linkage between MMG and tool.

27.4.5 The Converger/Evaluator

The converger checks if results from the expert tools are valid (converged). Sometimes this functionality is delivered by the expert tool itself but in many cases the results from the analysis tool still have to be judged separately. This can be done based on test cases.

The evaluator is an optimization routine that controls the search in the solution domain. It evaluates each design option analysed with the expert tools for its fit with the requirements. A wide range of methods can be applied: sequential linear programming, Sequential quadratic programming, genetic algorithms, heuristic methods, etc. where necessary, a limited amount of behavioural data from the expert tools is approximated with surrogate models (e.g. response surfaces) to make the optimization affordable. Multi-level optimization methods like Bliss and Target Cascading have not been tried yet in the DEE context.

27.4.6 The Agent-Based Framework

To make the components of the DEE work as a single service, they have to be flexibly connected, taking into account the fact that they can be distributed over different locations and can be running in different environments. The solution has been found in the SE practice where design team members solve these communication and geographical issues.

A DEE can be seen as an integrated product team (IPT) or design built team (DBT), where more actors, with different roles and functionalities, co-operate in the design process, Fig. 27.10. To make the DEE components act as virtual team members, they are wrapped in agents that communicate with each other and, if necessary, with the human team members [9]. The most senior agent performs the master functions, thus acting as project manager. When the master agent is unavailable, an automatic fall back system transfers the management to the next most senior agent.

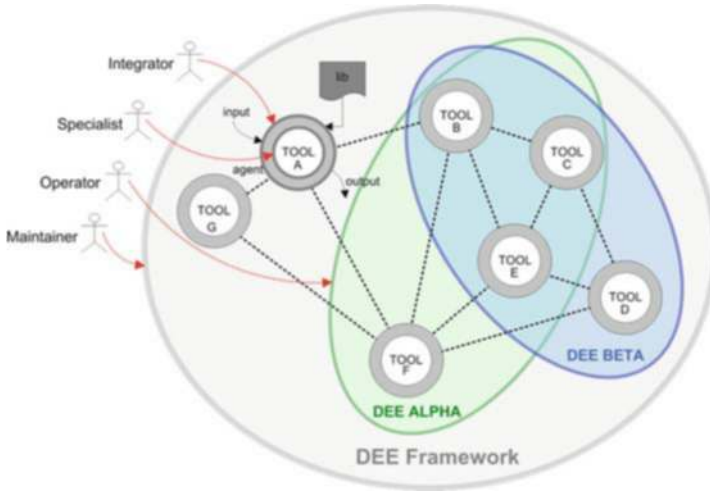


Fig. 27.10 The DEE agent-based framework

Four main functions have been identified for the agents:

- (i). Resource management: which resources connected and available to the network
- (ii). Resource interfacing: communication between elements and actors
- (iii). Process execution support: transformation process management
- (iv). Information flow control: external and internal data and data request management

The language of the agents is a major issue. Currently XML is used to import and export information about products and processes. Also for the engineering object, XML is used as a basis. Work is ongoing to define domain-specific languages that ease the communication between the components and between human and components. The language will most likely include UML-like diagrams to communicate with the user and use ontologies to do grammar and semantics checking on communication between humans and the DEE.

27.5 Results and Discussion

The different components of the DEE concept have been developed and tested in various projects. The MMG approach, using HLPs and CMs has been tried successfully in the 5th framework project Multi-Disciplinary Optimization of A Blended Wing Body Aircraft (MOB) [6], Fig. 27.11.

Since the completion of that project, the work continued [1], Fig. 27.12, and currently the ICAD-based MMG of MOB is being turned into a Genwork International

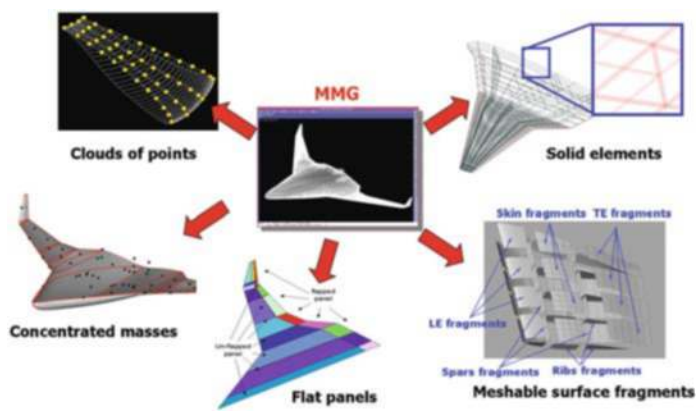


Fig. 27.11 The MMG of the EU 5th framework project MOB

GDL-based KBE application. In addition the approach is being applied in other domains like wind energy [7], Fig. 27.13.

The initiator approach has been tested in several structural design projects and proved to be highly successful. Extension to other product domains and other disciplinary fields is on its way. The linkage between MMG and different expert tools has been achieved for various tools. The later development of the agents-based framework has helped to make the links more flexible. The expert tools used are mostly commercial codes. For the aerodynamics, however, an adjoint solver-based CFD tool has been made in-house to analyse and optimize the outer shape of airfoils and 3D objects. The future coupling to the MMG will make this a very powerful tool. Different optimization algorithms have been used in various projects. The

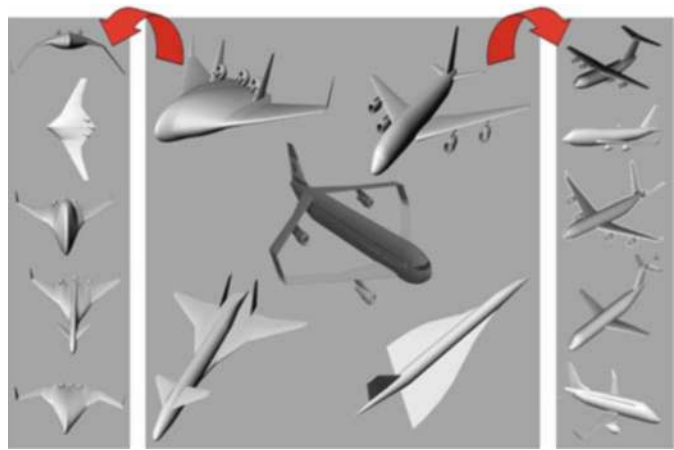


Fig. 27.12 The DEE current potential of the ICAD-based MMG for aircraft



Fig. 27.13 The MMG under development for wind turbines

implementations used have mostly been commercial tools. Only the sequential linear programming has been implemented in-house.

Although no full MDO case has been solved by the DEE approach yet, the spin-off of the development has resulted in successful applications. Work will therefore continue, aiming for a full implementation of the proposed methodology.

27.6 Conclusions

The systems engineering process can be supported by a computational system structured according to the design and engineering engine structure. It allows an object-based variational approach to design optimization.

The components of a DEE need a multiple of technologies to make them effective:

- Specifying the design problem: MDO, XML
- Specifying the design space: KBE, XML
- Initiating search process: Heuristics, MDO
- Linking to expert tools: KBE, agents
- Searching the design space: MDO (incl SLP, SQP, GA, DoE)
- Recording product/process results: XML

Prototypes of each component have been tested in various programs. A full integration of the tools to perform a real high-fidelity multi-disciplinary design optimization has not yet been achieved but will remain the ultimate proof of the approach.

Acknowledgments The authors want to thank Airbus UK, Airbus Hamburg, Stork AESP and Genworks International for their continuous support.

References

1. La Rocca G, van Tooren M (2007) Enabling Distributed Multi-Disciplinary Design of Complex Products: a Knowledge Based Engineering Approach. *J Design Research* 5: 3:333–352
2. Schut E, van Tooren M (2007) Design “Feasibilization” Using Knowledge-Based Engineering and Optimization Techniques. *J Aircraft* 44:6
3. Carpentieri G, Koren B, van Tooren M, (2007) Adjoint-Based Aerodynamic Shape Optimization on Unstructured Meshes. *J Comp Physics* 224:1:267–287
4. LaRocca G (2007), Knowledge Based Engineering Techniques to Support Aircraft Design And Multidisciplinary Analysis and Optimisation. Technical University of Delft, The Netherlands
5. Milton N, La Rocca G (2008), Knowledge Technologies. Polimetrica Monza/Milano
6. Morris A, Arendsen P, La Rocca, et al. (2004) MOB – a European project on multidisciplinary design optimisation. 24th ICAS Congress, Yokohama, Japan
7. Chiciudean T, LaRocca G, van Tooren (2008) A knowledge Based Engineering Approach to Support Automatic Design of Wind Turbine Blades. CIRP Design Conference, Twente, The Netherlands
8. Cooper D, LaRocca G (2007) Knowledge-based Techniques for Developing Engineering Applications in the 21st Century. 7th AIAA ATIO Conference, Belfast, Northern Ireland
9. Berends J, van Tooren M (2007) Design of a Multi-Agent Task Environment Framework to Support Multidisciplinary Design and Optimisation. 45th AIAA ASME Conference, Reno, USA

Chapter 28

Some Plebeian Variational Problems

Piero Villaggio

Abstract Several variational problems worthy to be treated are suggested by the operations we perform everyday after millennia of accumulated experiences.

28.1 Introduction

The Calculus of Variations is one of the most fascinating branches of mathematics. Its birthdate is the year 1744, when L. Euler proposed a method for obtaining a differential equation for solving variational problems. Problems of this kind were known since the remote antiquity, suggested by geometry, like that of the geodesic, of the minimal surface of revolution, of the curve of given perimeter enclosing the maximal area.

At the end of 17th century, new variational problems were proposed by mechanics, like that of brachistochrone (solved by Johann Bernoulli) or that of the projectile of least resistance (solved by Newton). Both kinds of problems were tackled by clever but not rigorous procedures, whereas Euler's method constituted a tool for treating them systematically.

In the following two centuries, the mathematics of Calculus of Variations was rendered more precise, in particular by the work of Weierstrass and Tonelli, and also its sectors of application were unexpectedly widened. Variational methods proved themselves to be essential for describing the motion of planets or of system of particles, judging the conditions of stability or collapse of a structure, controlling the trajectory of a rocket, and many other questions again. We may call these the “noble” problems of the Calculus of Variations.

But an attentive observer may object that the classical problems quoted above do not exhaust the list of the possible, important, applications of the theory. Many

Piero Villaggio

Dipartimento di Ingegneria Strutturale, Università di Pisa, via Diotisalvi, 2 – 56126 Pisa, Italy,
e-mail: dis@ing.unipi.it

natural phenomena, like the motion of winds, the onset of earthquakes, the shape of mountains and glaciers, the growing of trees and bones, have a variational formulation. And we must acknowledge that we are observing an increasing interest in these topics in the last two decades. However, there are still other problems, ever more important, which have been ignored until now. They concern the mathematical setting of the operations we perform in everyday life as, for example, hammering a nail into a domestic wall, cutting a piece of soap, peeling a protective membrane from a piece of cheese, walking, running, climbing, and cooking foods. We call “plebeian” these problems and – to the satisfaction of Karl Marx – we hope that they will be considered by scientist.

28.2 Mechanical Plebeian Problems

Hammering constituted for millennia one of the essential devices for the survival of the “*Homo sapiens*,” whose life was rendered possible by his capacity for shaking nuts, skulls, and connecting tables of wood for constructing huts or boats (Fig. 28.1).

Hammering is not uniform. If we observe a carpenter driving a nail, we notice that his distribution of effort during this operation is not uniform. At the beginning, his blows are light but frequent, at the end violent and rare. This strategy is not arbitrary but is the result of a precise criterion of economy. As the nail progressively enters inside the fibers of wood, the force required in order to permit the penetration becomes larger and larger. The total work that must be expended is prescribed (it is the work done by friction between wood and nail), but the distribution of this work during the time is free, and the carpenter instinctively adopts the law physiologically less tiring, namely keeping the mechanical power constant. This method may have industrial applications too, as in shaking a hot block of steel with a mallet or driving a pile in the soil. But are we sure that it is followed?

Another vital operation necessary for the survival of human beings is cutting, as, for example, splitting a piece of wood along its fibers with a blade (Fig. 28.2) for making a fire, or removing the skin of a, just killed, animal.

In contrast to hammering, cutting is not violent because the advance of the blade is obtained by exerting a continuous force and not by subsequent blows. We may predict that the blade reaches the bottom smoothly, but, against the expectations, we



Fig. 28.1 Hammering is not uniform

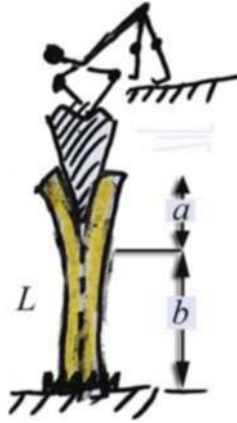


Fig. 28.2 The operation of cutting. There is a critical length a after which splitting is catastrophic

experience that for certain materials like wood (not skin) the separation under the edge of the blade is slow until a certain distance a from the top and then it suddenly jumps to the bottom along the remaining distance b .

Carpenters have known this strange behavior of wood since the beginning of the neolithic age and prevented the damages produced by the abrupt fall of the blade on the base by interposing a soft layer L between the bottom of the stump and the ground. Now the phenomenon is well understood in the light of fracture theory (cf. Bilek and Burns [1], Burrig and Keller [4]). When the split is shorter than a the energy is purely elastic and stored in the two halves separated by the advancing blade. As soon as the tip of the blade has reached the lower end of the interval a , then the elastic energy is suddenly converted into detachment energy localized along the prolongation b of the line of the fracture a . Then the blade falls violently downward and its remaining kinetic energy (if any) is eventually absorbed by the layer L . Thus the equation governing the process is

The Total Energy stored in a (strain energy of the deformed halves minus the exterior work exerted by the blade descended along a distance a) = Detachment energy developed along the line b plus residual kinetic energy of the blade at the bottom (absorbed by the layer L).

This equation may appear an abstract paralogism, but it is not so, because knowing some physical properties of the specimen of wood, like its elastic modulus and the specific detachment energy between its fibers, we can describe the operation of splitting in its smallest details. For example, we can predict instant by instant the time position of the blade after its first contact with the top of the specimen.

There are other important applications of the theory of brittle splitting. One is the onset of earthquakes as a consequence of the sudden detachment and gliding of a fault which, due to a local cooling, tends to contract and then, abruptly, slips along a lower layer. A second problem solvable with the theory of fracture is that of the shape of mountains. Pieces of rock tumble systematically down from the summits of mountains as a consequence of the phenomenon called “permafrost.”

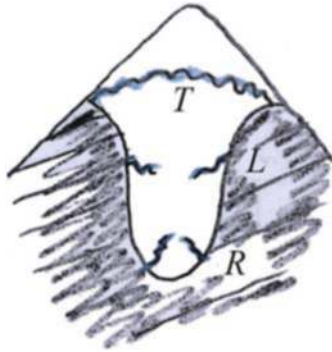


Fig. 28.3 Crevasses are distributed according to an impressive regularity

There is a clever but rudimentary theory for understanding *a posteriori* the changing of shape of the faces of mountains by a geometrical balance of the volume of debris that constantly falls down. This is the so-called “expoliation theory of slopes” (cf. Scheidegger [6] chap. 3) but a mechanical formulation of the problem is still lacking. Another appealing problem is the explanation of the shape and extent of a crevasses in glaciers. A glacier is a huge tongue of ice and snow flowing from the top of a mountain with an average velocity of 100 m/year in the Alps and 1000 m/year in the Himalayas. The tongue presents a bergschrund *T*, lateral cracks *L*, and radial cracks *R* at the bottom (Fig. 28.3). These lines of fracture are surprisingly regular (cf. Sturm–Zintl [7]) and can be explained by an energy balance accounting for the fracture energy unleashed by the crevasses.

A branch of Calculus of Variations rich of a long tradition, from Galilei to Maxwell and Michell, studies the optimum design of structures. Its purpose is that of shaping a structure exerting a given mechanical function with the minimum amount of material. The challenge of mechanical and civil engineers is just that of finding structures with least weight. But the same criterion was followed by nature during

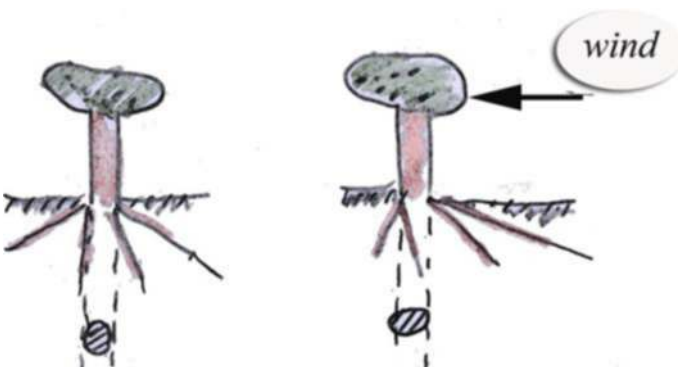


Fig. 28.4 (a) no wind, (b) wind

billions of years for shaping the forms of cells, cell aggregates, and skeleton. Any one who has, even superficially, looked over the great work of D'Arcy Thompson [3] knows how nature is teleologically clever in proposing its solutions. An example may illustrate the devices of nature (cf. Mattek [5]). Consider a tree planted on a sandy terrain as happens close to the coasts of an ocean. If the region is not windy, the cross-section of the trunk is circular and the roots are distributed symmetrically (Fig. 28.4(a)). But, if the wind tends to blow along a particular direction the cross-section grows oval in the direction of the wind and roots spread more vigorous on the windward side where the soil is subject to tension (Fig. 28.4(b)). This is a typical case of adaptive growth.

28.3 Locomotion

According to D'Arcy Thompson ([3], Chap. 2) the brothers Weber proposed the first "pendular" theory of human locomotion in 1836. Notwithstanding some later criticisms with proper qualifications, the theory is still accepted today. The act of walking is regarded as the motion of an inverted pendulum where each leg, alternatively, carries the weight of the body from a position G to a subsequent position G' , describing an arc of amplitude α (Fig. 28.5).

The linear distance $\overline{GG'}$ represents the length of each step, and the mechanical work is given by the rise h of the weight to the highest point of the arc $\widehat{GG'}$. If l is the length of a leg, the extent of each stride is proportional to l , but the time of swing varies inversely as \sqrt{l} . Therefore the velocity, measured as length/time, will also vary as \sqrt{l} . Therefore a larger man, or animal, goes faster than the smaller, but only in the ratio of the square root of his linear dimensions.

This is the dry, though iniquitous, conclusion of the pendular theory, but it ignores the fact that a man is able to lengthen his steps and increase their rhythm. The walker must cover a prescribed distance, say L , with the least fatigue, and asks himself which velocity v and which total time T are the most convenient in order to render the energetic expense a minimum. The choice is not obvious because too

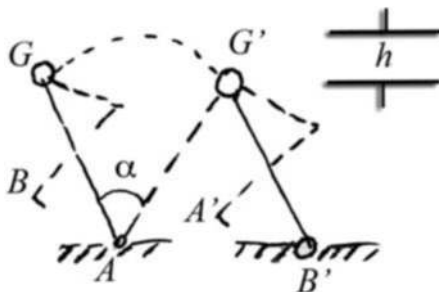


Fig. 28.5 The pendular theory

great a velocity increases the so-called athletic part of the energy, and too slow a motion increases the metabolic part of the energy steadily consumed even at rest. In mathematical terms, the problem can be formulated as follows. Assume that the dependence of the energy \mathcal{E} on v and T has the form

$$\mathcal{E} = \kappa_1 v^2 L + \kappa_2 T, \quad (28.1)$$

where κ_1, κ_2 are two positive constants. Physiologists agree that \mathcal{E} depends quadratically on v and linearly on T . On the other hand, T and v are related by the condition $vT = L$, and hence, after substitution into (28.1), we find the two (unique) values v_0, T_0 minimizing \mathcal{E} , namely

$$v_0 = \sqrt[3]{\frac{\kappa_2}{2\kappa_1}}, \quad T_0 = \frac{L}{v_0}. \quad (28.2)$$

We may object that this solution is not very realistic because it ignores the fact that the ground is not uniform and hence the walker is compelled to proceed slower at some points and faster at others. In this case we may still apply an expression of dissipated energy analogous to (28.1) but in the differential form:

$$d\mathcal{E} = \kappa_1(x)v(x)^2 dx + \kappa_2(x)dt, \quad (28.3)$$

where $\kappa_1(x), \kappa_2(x)$ are two strictly positive functions (not necessarily continuous) defined in the interval $0 \leq x \leq L$. Of course $dt = \frac{dx}{v(x)}$ in (28.3), and therefore the best velocity $v(x)$ minimizes the functional

$$\int_0^L d\mathcal{E} = \int_0^L \left(\kappa_1(x)v^2 + \frac{\kappa_2(x)}{v} \right) dx, \quad (28.4)$$

and the result is

$$v(x) = \sqrt[3]{\frac{\kappa_2(x)}{2\kappa_1(x)}}. \quad (28.5)$$

Another form of locomotion, by far older than walking, is climbing, practiced by men for 8 millions of years (according to Diamond [4]), first for survival and recently for an esthetic passion (the alpinism). The simplest but still non-trivial case of climbing is the ascent vertical-ladder (Fig. 28.6). The progression, step by step, is obtained starting from a position, in which legs are contracted and arms extended, to a superior place where the placement of limbs is inverted. In doing this the centre of mass of the body describes an arc $\widehat{GG'}$ of an unspecified curve, which is not a circle but is periodically repeated at each step. Then we conclude that the old pendular theory is valid even in climbing!

But, unlike horizontal walking, climbing requires the alternative movement of extension of legs and contraction of arms. Arms are, however, weaker than legs and therefore the effort spent in the position (a) of Fig. 28.6, where legs are bent and arm straightened, is much higher than that required in the position (b). Thus the local

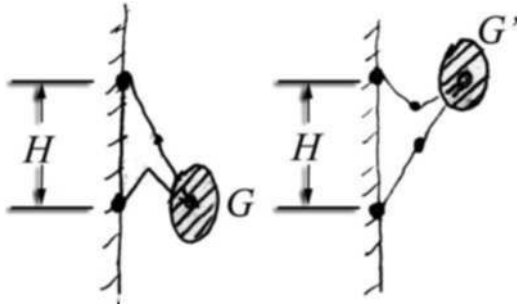


Fig. 28.6 (a) Initial position. (b) Final position

amount of energy dissipated during a single step of height H can be represented as

$$d\mathcal{E} = \kappa_1(x)v^2(x)dx + \kappa_2(x)dt + Wdx, \quad (28.6)$$

where x ($0 \leq x \leq H$) is the instantaneous level of G and W the weight of the climber. In addition, $\kappa_1(x)$, $\kappa_2(x)$ are two strictly positive and strictly monotone decreasing functions in the interval $0 \leq x \leq H$. The calculation of the best distribution of $v(x)$ proceeds as before.

There are other forms of animal locomotion, like running, jumping, swimming, and flying. These problems have been studied for three centuries, since the times of Galilei and Borelli, but their variational formulations are relatively recent.

28.4 Peeling and Cooking

The discovering and regular use of fire (about 13,000 years ago, according to Diamond [4]) represented a radical revolution in preparing food. Men realized that roasted meat is more edible, tasty, and, above all, more easily conservable, especially during warm seasons. Cooking was later extended to vegetables and fruits but with more sophisticated techniques like boiling or exposure to vapor. This heritage of knowledge is still exploited today.

Cooking foods requires a more advanced degree of intelligence, possessed only by man, for it is obtained by combining mechanics with heat propagation. Primitive men found clever procedures for cooking foods, although rudimentary, scientifically correct. But the use of heat for cooking is preceded by the purely mechanical operation of skinning since humans soon learnt the advantage of removing the outer layer of the bodies of killed animals before exposing them to the flames. The same procedure was performed with the husk of fruits before eating them. In the absence of cutting tools, the removal of the surface layer from a corpse or a fruit is obtained by peeling. In its most schematic formulation, peeling consists in detaching a semi-infinite tape glued to a substrate by exerting a tensile force P at the end O (Fig. 28.7). The instantaneous shape of the tape is a sort of truncated angle whose



Fig. 28.7 Peeling of a tape

vertex V slowly travels from left to right. The man exerts his pull along a direction making an angle α with respect to the horizontal. Let us assume that the tape is inextensible and the support is rigid, while the only soft element is a thin layer interposed between the first ones. And suppose that this layer is elastic-perfectly plastic. If P is small the tape will remain bonded to its support, but, as soon as P exceeds a given value depending upon angle α and the yield stress of the layer, the tape will start to detach. We now ask which is the best value of the angle α allowing peeling with the minimum P . Despite its apparent simplicity the problem is very difficult for it requires the calculation of the stress state in the layer. However, an approximate solution is the following (cf. Burrige and Keller [2]). Assume that the force P acts exclusively on the vertex V , with its two components $N = P \sin \alpha$, $T = P \cos \alpha$, and that breaking occurs as soon as N , T are such that (an integrated von Mises criterion)

$$N^2 + 4T^2 = P^2(\sin^2 \alpha + 4 \cos^2 \alpha) = 4K_0^2, \quad (28.7)$$

where K_0 is a given constant. Then P attains a maximum when $\alpha = 0$, namely when P is exerted parallel to the plane. We instinctively follow this rule when we try to detach a thin membrane, glued to a table, with the fingers!

The situation is completely different whenever the tape possesses a flexural stiffness, say B , and behaves like a beam (Fig. 28.8). Now the progressive detachment of the beam is due to the fact that the bending moment at V has reached a critical value M_0 , that is,

$$P \sin \alpha \cdot \overline{O'V} = M_0. \quad (28.8)$$

Here the minimum of P is achieved for $\alpha = \frac{\pi}{2}$, namely when P is perpendicular to the plane. Observe that, unlike the previous situation, the magnitude of P decreases little by little as V goes away from O' . We realize this when we open a can of sardines pulling the upper ring with a finger!

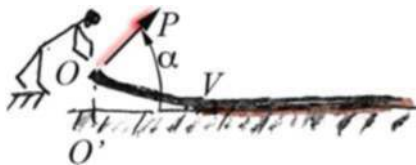


Fig. 28.8 Peeling of a beam

The fundamental question in cooking is that of distributing the temperature in the interior of a solid (meat, bread, a cake) by prescribing the temperature on its surface. Unfortunately, heat propagation is a slow process and hence, if we wish to roast a steak uniformly, we must wait for too long. Therefore millennia of experiences have codified the methods for cooking foods. These prescriptions are now printed on the label of aliments we buy everyday. They play on the compromise of tolerating a partially edible crust provided that the interior is sufficiently cooked. The problem can be formulated as follows. Consider a rod of length, say π , initially at temperature zero but whose end points $x = 0$, $x = \pi$ are taken at temperature T_0 . Let the rod occupy the interval $0 \leq x \leq \pi$ and put κ , the conductivity, equal to 1. The distribution of temperature at any point of the rod at each time is explicitly representable with the method of separation of variables. The result is

$$T(x, t) = T_0 \left(1 - \sum_{n=1,3,5} e^{-n^2 t} \sin nx \right). \quad (28.9)$$

This formula allows us to determine the time of maintenance of an rod in an oven having temperature T_0 in order that the mid point $x = \frac{\pi}{2}$ reaches a given temperature $T_1 < T_0$. Truncating the series in (28.9) at its first term we can write

$$T\left(\frac{\pi}{2}, t\right) = T_1 \simeq T_0 (1 - e^{-t}), \quad (28.10)$$

whence we obtain

$$t = \ln \left(\frac{1}{1 - T_1/T_0} \right). \quad (28.11)$$

Provided that T_1 is fixed, a device for shortening the time of cooking is that of increasing T_0 . But this may overheat the food close to the ends. Good cooks play between these conflicting needs!

Another appealing example of thermal regulation is offered by the resistance of warm-blooded animals to a cold climate. Already more than 150 years ago the physiologist Carl Bergman arrived at the conclusion that a smaller animal must produce more heat (per unit of mass) than a large one in order to keep balance with surface-loss (cf. D'Arcy Thompson [3], p- 25–26). For this reason the smaller animal needs more food. But the one-dimensional model of a rod may explain the phenomenon more precisely. Consider the bar of Fig. 28.9 and assume, this time, that the ends $x = 0$, $x = \pi$ are kept at temperature zero while the initial temperature in the interior $T(x, 0)$ is a constant, say T_0 . The distribution of temperature at subsequent times is



Fig. 28.9 Heating of a rod

$$T(x, t) = T_0 \sum_{1, 3, \dots} e^{-n^2 t} \sin nx. \quad (28.12)$$

Suppose now that the length of the bar is not π but $N\pi$ where N is an integer number. Now the law of propagation of temperature is

$$T(x, t) = T_0 \sum_{1, 3, \dots} e^{-\left(\frac{n}{N}\right)^2 t} \sin \frac{n}{N} x. \quad (28.13)$$

At the mid points $x = \frac{\pi}{2}$ and $x = N\frac{\pi}{2}$ we have the following approximate values of the temperature

$$T\left(\frac{\pi}{2}, t\right) \simeq T_0 e^{-t}, \quad T_N\left(N\frac{\pi}{2}, t\right) \simeq T_0 e^{-\left(\frac{1}{N}\right)^2 t}. \quad (28.14)$$

This implies that, at the same time t_1 , the temperature is a small animal is catastrophically lower than in a large one and hence the former hardly survives in arctic regions.

28.5 Conclusions

Plebeian problems are important for at least three reasons: they offer a rational re-visitation of several technical achievements of the past; suggest new devices for improving our actual life; propose exciting questions to mathematicians. They are now becoming popular.

References

1. Bilek, Z. J. and Burns, S.J.: Crack propagation in wedged double cantilevered beam specimens. *J. Mech. Phys. Solids*, 22, 85–95 (1974).
2. Burridge, A. and Keller, J. B.: “Peeling, Slipping and Cracking – Some One-Dimensional Free-Boundary Problems in Mechanics.” *SIAM Rev.*, 20, 1, 31–61 (1978).
3. D’Arcy Thompson, W.: *On Growth and Form*. Cambridge: Cambridge University Press (1961).
4. Diamond, J.: *The Third Chimpanzee*. New York: Harper (2006).
5. Mattek, C.: *Design in Nature*. Berlin/Heidelberg: Springer (1998).
6. Scheidegger, A. E.: *Theoretical Geomorphology*. Berlin/Heidelberg: Springer (1991).
7. Sturm, G. and Zintl, F.: *Sicheres Klettern in Fels und Eis*. München: BLV (1969).