Lecture Notes in Applied and Com

Vincent Acary
Bernard Brogliato

Numerica
for Nonsn
Dynamica

Applications in Mechanics

Lecture Notes in Applied and Computational Mechanics

Volume 35

Series Editors

Prof. Dr.-Ing. Friedrich Pfeiffer
Prof. Dr.-Ing. Peter Wriggers

Lecture Notes in Applied and Computational Mechanics

Edited by F. Pfeiffer and P. Wriggers

Further volumes of this series found on our homepage: springer.com

Vol. 35: Acary, V.; Brogliato, B.
Numerical Methods for Nonsmooth Dynamical Systems:
Applications in Mechanics and Electronics
540 p. 2008 [978-3-540-75391-9]

Vol. 34: Flores, P.; Ambrósio, J.; Pimenta Claro, J.C.;
Lankarani Hamid M.
Kinematics and Dynamics of Multibody Systems
with Imperfect Joints: Models and Case Studies
186 p. 2008 [978-3-540-74359-0]

Vol. 33: Niesłony, A.; Macha, E.
Spectral Method in Multiaxial Random Fatigue
146 p. 2007 [978-3-540-73822-0]

Vol. 32: Bardzokas, D.I.; Filshtinsky, M.L.; Filshtinsky, L.A. (Eds.)
Mathematical Methods
in Electro-Magneto-Elasticity
530 p. 2007 [978-3-540-71030-1]

Vol. 31: Lehmann, L. (Ed.)
Wave Propagation in Infinite Domains
186 p. 2007 [978-3-540-71108-7]

Vol. 30: Stupkiewicz, S. (Ed.)
Micromechanics of Contact and Interphase Layers
206 p. 2006 [978-3-540-49716-5]

Vol. 29: Schanz, M.; Steinbach, O. (Eds.)
Boundary Element Analysis
571 p. 2006 [978-3-540-47465-4]

Vol. 28: Helmig, R.; Mielke, A.; Wohlmuth, B.I. (Eds.)
Multifield Problems in Solid and Fluid Mechanics
571 p. 2006 [978-3-540-34959-4]

Vol. 27: Wriggers P., Nackenhorst U. (Eds.)
Analysis and Simulation of Contact Problems
395 p. 2006 [978-3-540-31760-9]

Vol. 26: Nowacki, J.P.
Static and Dynamic Coupled Fields in Bodies
with Piezoeffects or Polarization Gradient
209 p. 2006 [978-3-540-31668-8]

Vol. 25: Chen C.-N.
Discrete Element Analysis Methods
of Generic Differential Quadratures
282 p. 2006 [978-3-540-28947-0]

Vol. 24: Schenk, C.A., Schuëller. G
Uncertainty Assessment of Large
Finite Element Systems
165 p. 2006 [978-3-540-25343-3]

Vol. 23: Frémond M., Maceri F. (Eds.)
Mechanical Modelling and Computational Issues
in Civil Engineering
400 p. 2005 [978-3-540-25567-3]

Vol. 22: Chang C.H.
Mechanics of Elastic Structures with Inclined Members:
Analysis of Vibration, Buckling and Bending of X-Braced
Frames and Conical Shells
190 p. 2004 [978-3-540-24384-7]

Vol. 21: Hinkelmann R.
Efficient Numerical Methods and Information-Processing
Techniques for Modeling Hydro- and Environmental
Systems
305 p. 2005 [978-3-540-24146-1]

Vol. 20: Zohdi T.I., Wriggers P.
Introduction to Computational Micromechanics
196 p. 2005 [978-3-540-22820-2]

Vol. 19: McCallen R., Browand F., Ross J. (Eds.)
The Aerodynamics of Heavy Vehicles:
Trucks, Buses, and Trains
567 p. 2004 [978-3-540-22088-6]

Vol. 18: Leine, R.I., Nijmeijer, H.
Dynamics and Bifurcations
of Non-Smooth Mechanical Systems
236 p. 2004 [978-3-540-21987-3]

Vol. 17: Hurtado, J.E.
Structural Reliability: Statistical Learning Perspectives
257 p. 2004 [978-3-540-21963-7]

Vol. 16: Kienzler R., Altenbach H., Ott I. (Eds.)
Theories of Plates and Shells:
Critical Review and New Applications
238 p. 2004 [978-3-540-20997-3]

Vol. 15: Dyszlewicz, J.
Micropolar Theory of Elasticity
356 p. 2004 [978-3-540-41835-1]

Vol. 14: Frémond M., Maceri F. (Eds.)
Novel Approaches in Civil Engineering
400 p. 2003 [978-3-540-41836-8]

Vol. 13: Kolymbas D. (Eds.)
Advanced Mathematical and Computational
Geomechanics
315 p. 2003 [978-3-540-40547-4]

Vol. 12: Wendland W., Efendiev M. (Eds.)
Analysis and Simulation of Multifield Problems
381 p. 2003 [978-3-540-00696-1]

Vol. 11: Hutter K., Kirchner N. (Eds.)
Dynamic Response of Granular and Porous Materials
under Large and Catastrophic Deformations
426 p. 2003 [978-3-540-00849-1]

Vol. 10: Hutter K., Baaser H. (Eds.)
Deformation and Failure in Metallic Materials
409 p. 2003 [978-3-540-00848-4]

Numerical Methods for Nonsmooth Dynamical Systems

Applications in Mechanics and Electronics

Vincent Acary · Bernard Brogliato

With 81 Figures and 4 Tables

Springer

Vincent Acary and Bernard Brogliato
INRIA, team-project BIPOP
Inovallée
655 avenue de l'Europe
Montbonnot, 38334 Saint-Ismier
France
vincent.acary@inrialpes.fr
bernard.brogliato@inrialpes.fr

ISBN: 978-3-540-75391-9 e-ISBN: 978-3-540-75392-6

Lecture Notes in Applied and Computational Mechanics ISSN 1613-7736

Library of Congress Control Number: 2007939833

Cover design: WMXDesign GmbH, Heidelberg

Printed on acid-free paper

9 8 7 8 6 5 4 3 2 1 0

springer.com

À Céline et Martin
À Laurence et Bastien

Preface

This book concerns the numerical simulation of dynamical systems whose trajectories may not be differentiable everywhere. They are named *nonsmooth* dynamical systems. They make an important class of systems, first because of the many applications in which nonsmooth models are useful, secondly because they give rise to new problems in various fields of science. Usually nonsmooth dynamical systems are represented as differential inclusions, complementarity systems, evolution variational inequalities, each of these classes itself being split into several subclasses. The book is divided into four parts, the first three parts being sketched in Fig. 0.1. The aim of the first part is to present the main tools from mechanics and applied mathematics which are necessary to understand how nonsmooth dynamical systems may be numerically simulated in a reliable way. Many examples illustrate the theoretical results, and an emphasis is put on mechanical systems, as well as on electrical circuits (the so-called Filippov's systems are also examined in some detail, due to their importance in control applications). The second and third parts are dedicated to a detailed presentation of the numerical schemes. A fourth part is devoted to the presentation of the software platform SICONOS. This book is not a textbook on numerical analysis of nonsmooth systems, in the sense that despite the main results of numerical analysis (convergence, order of consistency, etc.) being presented, their proofs are not provided. Our main concern is rather to present in detail how the algorithms are constructed and what kind of advantages and drawbacks they possess.

Nonsmooth mechanics (resp. nonsmooth electrical circuits) is a topic that has been pioneered and developed in parallel with convex analysis in the 1960s and the 1970s in western Europe by J.J. Moreau, M. Schatzman, and P.D. Panagiotopoulos (resp. by the Dutch school of van Bockhoven and Leenaerts), then followed by several groups of researchers in Montpellier, Munich, Eindhoven, Marseille, Stockholm, Lausanne, Lisbon, Grenoble, Zurich, etc. More recently nonsmooth dynamical systems (especially complementarity systems) emerged in the USA, a country in which, paradoxically, complementarity theory and convex analysis (which are central tools for the study of nonsmooth mechanical and electrical systems) have been developed since a long time. Though nonsmooth mechanics and more generally nonsmooth dynamical systems have long been studied by mechanical engineers (impact mechanics

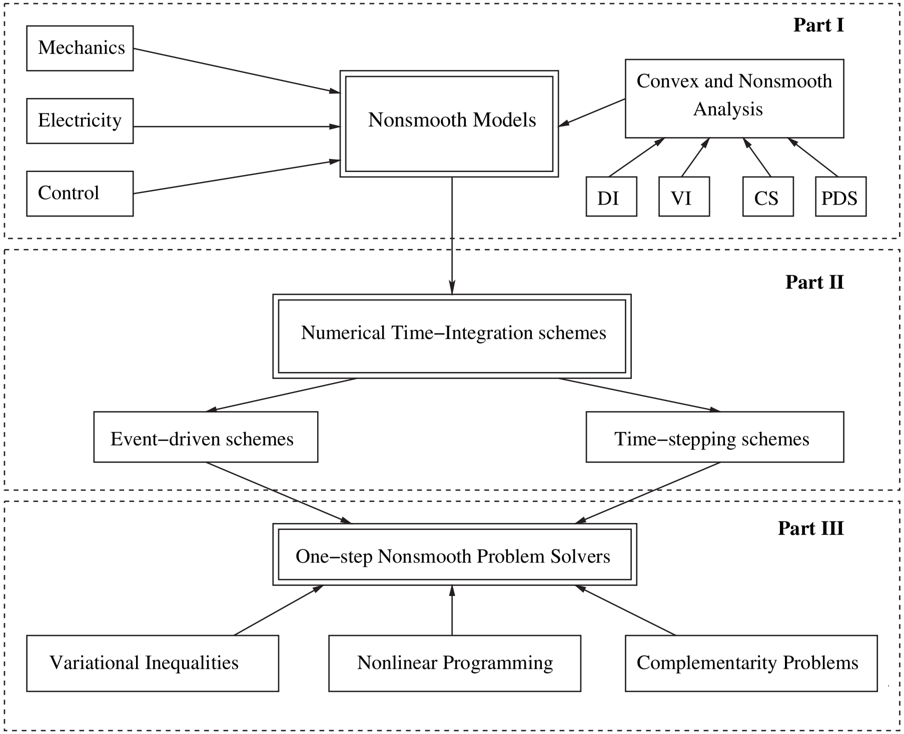


Fig. 0.1. Book Synopsis

can be traced back to ancient Greeks!) and applied mathematicians, their study has more recently attracted researchers of other scientific communities like systems and control, robotics, physics of granular media, civil engineering, virtual reality, haptic systems, image synthesis. We hope that this book will increase its dissemination.

We warmly thank Claude Lemaréchal (INRIA Bipop) for his many comments and discussions on Chap. 12 and Mathieu Renouf (LAMCOS-CNRS, Lyon) whose joint work with the first author contributed to Chap. 13. We also thank Professor F. Pfeiffer (Munich), an ardent promoter of nonsmooth mechanical systems, for his encouragements to us for writing this monograph, and Dr. Ditzinger (Springer Verlag). This work originated from a set of draft notes for a CEA-EdF-INRIA spring school that occurred in Rocquencourt from May 29 to June 02, 2006. The authors thank M. Jean (LMA-CNRS, Marseille, France) for his collaboration to this school and part of the preliminary draft. We would finally like to mention that part of this work was made in the framework of the European project SICONOS IST 2001-37172, from which the software platform SICONOS emerged. In particular the works of F. Pérignon and P. Denoyelle, expert engineers in the INRIA team-project Bipop, are here acknowledged.

Montbonnot,
August 2007

Vincent Acary
Bernard Brogliato

Contents

1 Nonsmooth Dynamical Systems: Motivating Examples and Basic Concepts 1

1.1 Electrical Circuits with Ideal Diodes 1

1.1.1 Mathematical Modeling Issues 2

1.1.2 Four Nonsmooth Electrical Circuits 5

1.1.3 Continuous System (Ordinary Differential Equation) 7

1.1.4 Hints on the Numerical Simulation of Circuits (a) and (b) . 9

1.1.5 Unilateral Differential Inclusion 12

1.1.6 Hints on the Numerical Simulation of Circuits (c) and (d) . 14

1.1.7 Calculation of the Equilibrium Points 19

1.2 Electrical Circuits with Ideal Zener Diodes 21

1.2.1 The Zener Diode 21

1.2.2 The Dynamics of a Simple Circuit 23

1.2.3 Numerical Simulation by Means of Time-Stepping Schemes 28

1.2.4 Numerical Simulation by Means of Event-Driven Schemes 38

1.2.5 Conclusions 40

1.3 Mechanical Systems with Coulomb Friction 40

1.4 Mechanical Systems with Impacts: The Bouncing Ball Paradigm . . . 41

1.4.1 The Dynamics 41

1.4.2 A Measure Differential Inclusion 44

1.4.3 Hints on the Numerical Simulation of the Bouncing Ball . . 45

1.5 Stiff ODEs, Explicit and Implicit Methods, and the Sweeping Process 50

1.5.1 Discretization of the Penalized System 51

1.5.2 The Switching Conditions 52

1.5.3 Discretization of the Relative Degree Two Complementarity System 53

1.6 Summary of the Main Ideas 53

Part I Formulations of Nonsmooth Dynamical Systems

2 Nonsmooth Dynamical Systems: A Short Zoology 57
2.1 Differential Inclusions 57
2.1.1 Lipschitzian Differential Inclusions 58
2.1.2 Upper Semi-continuous DIs and Discontinuous Differential Equations 61
2.1.3 The One-Sided Lipschitz Condition 68
2.1.4 Recapitulation of the Main Properties of DIs 71
2.1.5 Some Hints About Uniqueness of Solutions 73
2.2 Moreau's Sweeping Process and Unilateral DIs 74
2.2.1 Moreau's Sweeping Process 74
2.2.2 Unilateral DIs and Maximal Monotone Operators 77
2.2.3 Equivalence Between UDIs and other Formalisms 78
2.3 Evolution Variational Inequalities 80
2.4 Differential Variational Inequalities 82
2.5 Projected Dynamical Systems 84
2.6 Dynamical Complementarity Systems 85
2.6.1 Generalities 85
2.6.2 Nonlinear Complementarity Systems 88
2.7 Second-Order Moreau's Sweeping Process 88
2.8 ODE with Discontinuities 92
2.8.1 Order of Discontinuity 92
2.8.2 Transversality Conditions 93
2.8.3 Piecewise Affine and Piecewise Continuous Systems 94
2.9 Switched Systems 98
2.10 Impulsive Differential Equations 100
2.10.1 Generalities and Well-Posedness 100
2.10.2 An Aside to Time-Discretization and Approximation 104
2.11 Summary 104

3 Mechanical Systems with Unilateral Constraints and Friction 107
3.1 Multibody Dynamics: The Lagrangian Formalism 107
3.1.1 Perfect Bilateral Constraints 109
3.1.2 Perfect Unilateral Constraints 110
3.1.3 Smooth Dynamics as an Inclusion 112
3.2 The Newton–Euler Formalism 112
3.2.1 Kinematics 112
3.2.2 Kinetics 115
3.2.3 Dynamics 117
3.3 Local Kinematics at the Contact Points 123
3.3.1 Local Variables at Contact Points 123
3.3.2 Back to Newton–Euler's Equations 126
3.3.3 Collision Detection and the Gap Function Calculation 128

3.4 The Smooth Dynamics of Continuum Media 131
3.4.1 The Smooth Equations of Motion 131
3.4.2 Summary of the Equations of Motion 135
3.5 Nonsmooth Dynamics and Schatzman's Formulation 135
3.6 Nonsmooth Dynamics and Moreau's Sweeping Process 137
3.6.1 Measure Differential Inclusions 137
3.6.2 Decomposition of the Nonsmooth Dynamics 137
3.6.3 The Impact Equations and the Smooth Dynamics 138
3.6.4 Moreau's Sweeping Process 139
3.6.5 Finitely Represented $\mathscr{C}$ and the Complementarity Formulation 141
3.7 Well-Posedness Results 143
3.8 Lagrangian Systems with Perfect Unilateral Constraints: Summary . 143
3.9 Contact Models 144
3.9.1 Coulomb's Friction 145
3.9.2 De Saxcé's Bipotential Function 148
3.9.3 Impact with Friction 151
3.9.4 Enhanced Contact Models 153
3.10 Lagrangian Systems with Frictional Unilateral Constraints and Newton's Impact Laws: Summary 161
3.11 A Mechanical Filippov's System 162

4 Complementarity Systems 165
4.1 Definitions 165
4.2 Existence and Uniqueness of Solutions 167
4.2.1 Passive LCS 168
4.2.2 Examples of LCS 169
4.2.3 Complementarity Systems and the Sweeping Process 170
4.2.4 Nonlinear Complementarity Systems 172
4.3 Relative Degree and the Completeness of the Formulation 173
4.3.1 The Single Input/Single Output (SISO) Case 174
4.3.2 The Multiple Input/Multiple Output (MIMO) Case 175
4.3.3 The Solutions and the Relative Degree 175

5 Higher Order Constrained Dynamical Systems 177
5.1 Motivations 177
5.2 A Canonical State Space Representation 178
5.3 The Space of Solutions 180
5.4 The Distribution DI and Its Properties 180
5.4.1 Introduction 180
5.4.2 The Inclusions for the Measures ν_i 182
5.4.3 Two Formalisms for the HOSP 183
5.4.4 Some Qualitative Properties 186
5.5 Well-Posedness of the HOSP 187
5.6 Summary of the Main Ideas of Chapters 4 and 5 188

6 Specific Features of Nonsmooth Dynamical Systems 189
6.1 Discontinuity with Respect to Initial Conditions 189
6.1.1 Impact in a Corner 189
6.1.2 A Theoretical Result 190
6.1.3 A Physical Example 191
6.2 Frictional Paroxysms (the Painlevé Paradoxes) 192
6.3 Infinity of Events in a Finite Time 193
6.3.1 Accumulations of Impacts 193
6.3.2 Infinitely Many Switchings in Filippov's Inclusions 194
6.3.3 Limit of the Saw-Tooth Function in Filippov's Systems . . . 194

Part II Time Integration of Nonsmooth Dynamical Systems

7 Event-Driven Schemes for Inclusions with AC Solutions 203
7.1 Filippov's Inclusions 203
7.1.1 Introduction 203
7.1.2 Stewart's Method 205
7.1.3 Why Is Stewart's Method Superior to Trivial Event-Driven Schemes? 213
7.2 ODEs with Discontinuities with a Transversality Condition 215
7.2.1 Position of the Problem 215
7.2.2 Event-Driven Schemes 215

8 Event-Driven Schemes for Lagrangian Systems 219
8.1 Introduction 219
8.2 The Smooth Dynamics and the Impact Equations 221
8.3 Reformulations of the Unilateral Constraints at Different Kinematics Levels 222
8.3.1 At the Position Level 222
8.3.2 At the Velocity Level 222
8.3.3 At the Acceleration Level 223
8.3.4 The Smooth Dynamics 224
8.4 The Case of a Single Contact 225
8.4.1 Comments 227
8.5 The Multi-contact Case and the Index Sets 229
8.5.1 Index Sets 229
8.6 Comments and Extensions 230
8.6.1 Event-Driven Algorithms and Switching Diagrams 230
8.6.2 Coulomb's Friction and Enhanced Set-Valued Force Laws . 231
8.6.3 Bilateral or Unilateral Dynamics? 232
8.6.4 Event-Driven Schemes: Lötstedt's Algorithm 232
8.6.5 Consistency and Order of Event-Driven Algorithms 236
8.7 Linear Complementarity Systems 240
8.8 Some Results 241

9 Time-Stepping Schemes for Systems with AC Solutions 243
9.1 ODEs with Discontinuities 243
9.1.1 Numerical Illustrations of Expected Troubles 243
9.1.2 Consistent Time-Stepping Methods 247
9.2 DIs with Absolutely Continuous Solutions 251
9.2.1 Explicit Euler Algorithm 252
9.2.2 Implicit θ-Method 256
9.2.3 Multistep and Runge–Kutta Algorithms 258
9.2.4 Computational Results and Comments 263
9.3 The Special Case of the Filippov's Inclusions 266
9.3.1 Smoothing Methods 266
9.3.2 Switched Model and Explicit Schemes 267
9.3.3 Implicit Schemes and Complementarity Formulation 269
9.3.4 Comments 271
9.4 Moreau's Catching-Up Algorithm for the First-Order Sweeping Process 271
9.4.1 Mathematical Properties 272
9.4.2 Practical Implementation of the Catching-up Algorithm . . . 273
9.4.3 Time-Independent Convex Set K 274
9.5 Linear Complementarity Systems with $r \leqslant 1$ 275
9.6 Differential Variational Inequalities 279
9.6.1 The Initial Value Problem (IVP) 280
9.6.2 The Boundary Value Problem 281
9.7 Summary of the Main Ideas 283

10 Time-Stepping Schemes for Mechanical Systems 285
10.1 The Nonsmooth Contact Dynamics (NSCD) Method 285
10.1.1 The Linear Time-Invariant Nonsmooth Lagrangian Dynamics 286
10.1.2 The Nonlinear Nonsmooth Lagrangian Dynamics 289
10.1.3 Discretization of Moreau's Inclusion 293
10.1.4 Sweeping Process with Friction 295
10.1.5 The One-Step Time-Discretized Nonsmooth Problem 296
10.1.6 Convergence Properties 303
10.1.7 Bilateral and Unilateral Constraints 305
10.2 Some Numerical Illustrations of the NSCD Method 307
10.2.1 Granular Material 307
10.2.2 Deep Drawing 309
10.2.3 Tensegrity Structures 309
10.2.4 Masonry Structures 309
10.2.5 Real-Time and Virtual Reality Simulations 311
10.2.6 More Applications 314
10.2.7 Moreau's Time-Stepping Method and Painlevé Paradoxes . 315
10.3 Variants and Other Time-Stepping Schemes 315
10.3.1 The Paoli–Schatzman Scheme 315
10.3.2 The Stewart–Trinkle–Anitescu–Potra Scheme 317

11 Time-Stepping Scheme for the HOSP 319
11.1 Insufficiency of the Backward Euler Method 319
11.2 Time-Discretization of the HOSP 321
11.2.1 Principle of the Discretization 321
11.2.2 Properties of the Discrete-Time Extended Sweeping Process 322
11.2.3 Numerical Examples 324
11.3 Synoptic Outline of the Algorithms 325

Part III Numerical Methods for the One-Step Nonsmooth Problems

12 Basics on Mathematical Programming Theory 331
12.1 Introduction 331
12.2 The Quadratic Program (QP) 331
12.2.1 Definition and Basic Properties 331
12.2.2 Equality-Constrained QP 335
12.2.3 Inequality-Constrained QP 338
12.2.4 Comments on Numerical Methods for QP 344
12.3 Constrained Nonlinear Programming (NLP) 345
12.3.1 Definition and Basic Properties 345
12.3.2 Main Methods to Solve NLPs 347
12.4 The Linear Complementarity Problem (LCP) 351
12.4.1 Definition of the Standard Form 351
12.4.2 Some Mathematical Properties 352
12.4.3 Variants of the LCP 355
12.4.4 Relation Between the Variants of the LCPs 357
12.4.5 Links Between the LCP and the QP 359
12.4.6 Splitting-Based Methods 363
12.4.7 Pivoting-Based Methods 367
12.4.8 Interior Point Methods 374
12.4.9 How to chose a LCP solver? 379
12.5 The Nonlinear Complementarity Problem (NCP) 379
12.5.1 Definition and Basic Properties 379
12.5.2 The Mixed Complementarity Problem (MCP) 383
12.5.3 Newton–Josephy's and Linearization Methods 384
12.5.4 Generalized or Semismooth Newton's Methods 385
12.5.5 Interior Point Methods 388
12.5.6 Effective Implementations and Comparison of the Numerical Methods for NCPs 388
12.6 Variational and Quasi-Variational Inequalities 389
12.6.1 Definition and Basic Properties 389
12.6.2 Links with the Complementarity Problems 390
12.6.3 Links with the Constrained Minimization Problem 391
12.6.4 Merit and Gap Functions for VI 392

12.6.5 Nonsmooth and Generalized equations 396
12.6.6 Main Types of Algorithms for the VI and QVI 398
12.6.7 Projection-Type and Splitting Methods 398
12.6.8 Minimization of Merit Functions 400
12.6.9 Generalized Newton Methods 401
12.6.10 Interest from a Computational Point of View 401
12.7 Summary of the Main Ideas 401

13 Numerical Methods for the Frictional Contact Problem 403
13.1 Introduction 403
13.2 Summary of the Time-Discretized Equations 403
13.2.1 The Index Set of Forecast Active Constraints 403
13.2.2 Summary of the OSNSPs 405
13.3 Formulations and Resolutions in LCP Forms 407
13.3.1 The Frictionless Case with Newton's Impact Law 407
13.3.2 The Frictionless Case with Newton's Impact and Linear Perfect Bilateral Constraints 408
13.3.3 Two-Dimensional Frictional Case as an LCP 409
13.3.4 Outer Faceting of the Coulomb's Cone 410
13.3.5 Inner Faceting of the Coulomb's Cone 414
13.3.6 Comments 417
13.3.7 Weakness of the Faceting Process 418
13.4 Formulation and Resolution in a Standard NCP Form 419
13.4.1 The Frictionless Case 419
13.4.2 A Direct MCP for the 3D Frictional Contact 419
13.4.3 A Clever Formulation of the 3D Frictional Contact as an NCP 420
13.5 Formulation and Resolution in QP and NLP Forms 422
13.5.1 The Frictionless Case 422
13.5.2 Minimization Principles and Coulomb's Friction 423
13.6 Formulations and Resolution as Nonsmooth Equations 424
13.6.1 Alart and Curnier's Formulation and Generalized Newton's Method 424
13.6.2 Variants and Line-Search Procedure 429
13.6.3 Other Direct Equation-Based Reformulations 430
13.7 Formulation and Resolution as VI/CP 432
13.7.1 VI/CP Reformulation 432
13.7.2 Projection-type Methods 433
13.7.3 Fixed-Point Iterations on the Friction Threshold and Ad Hoc Projection Methods 434
13.7.4 A Clever Block Splitting: the Nonsmooth Gauss–Seidel (NSGS) Approach 437
13.7.5 Newton's Method for VI 440

Part IV The SICONOS Software: Implementation and Examples

14 The SICONOS Platform 443
- 14.1 Introduction 443
- 14.2 An Insight into SICONOS 443
  - 14.2.1 Step 1. Building a Nonsmooth Dynamical System 444
  - 14.2.2 Step 2. Simulation Strategy Definition 447
- 14.3 SICONOS Software 448
  - 14.3.1 General Principles of Modeling and Simulation 448
  - 14.3.2 NSDS-Related Components 451
  - 14.3.3 Simulation-Related Components 456
  - 14.3.4 SICONOS Software Design 457
- 14.4 Examples 460
  - 14.4.1 The Bouncing Ball(s) 460
  - 14.4.2 The Woodpecker Toy 463
  - 14.4.3 MOS Transistors and Inverters 464
  - 14.4.4 Control of Lagrangian systems 466

A Convex, Nonsmooth, and Set-Valued Analysis 475
- A.1 Set-Valued Analysis 475
- A.2 Subdifferentiation 475
- A.3 Some Useful Equivalences 476

B Some Results of Complementarity Theory 479

C Some Facts in Real Analysis 481
- C.1 Functions of Bounded Variations in Time 481
- C.2 Multifunctions of Bounded Variation in Time 482
- C.3 Distributions Generated by RCLSBV Functions 483
- C.4 Differential Measures 486
- C.5 Bohl's Distributions 487
- C.6 Some Useful Results 487

References 489

Index 519

List of Acronyms

Absolutely Continuous (AC)
Affine Variational Inequality (AVI)

Complementarity Problem (CP)
Constraint Qualification (CQ)

Differential Algebraic Equation (DAE)
Dynamical (or Differential) Complementarity System (DCS)
Differential Inclusion (DI)

Evolution Variational Inequality (EVI)

Karush-Kuhn-Tucker (KKT)

Linear Complementarity Problem (LCP)
Linear Complementarity System (LCS)
Linar Independence Constraint Qualification (LICQ)

Mixed Complementarity Problem (MCP)
Measure Differential Inclusion (MDI)
Mixed Linear Complementarity Problem (MLCP)
Mixed Linear Complementarity System (MLCS)

Nonlinear Complementarity Problem (NCP)
NonLinear Programming (NLP)
Nonsmooth Gauss–Seidel (NSGS)

Ordinary Differential Equation (ODE)
Onestep NonSmooth Problem (OSNSP)

Positive Definite (PD)
Positive Semi–Definite (PSD)

Quadratic Program (QP)

Successive Linear Complementarity Problem (SLCP)
Successive Quadratic Program (SQP)

Unilateral Differential Inclusion (UDI)

Variational Inequality (VI)

List of Algorithms

1 Stewart's event-driven method . . . 212
2 Stewart's switching point location method . . . 213
3 Stewart's active-set updating procedure . . . 213
4 Event -driven procedure on a single time-step with one contact . . . 228
5 Event -driven procedure on a single time-step with several contacts . . 231
6 Leine's switch model . . . 268
7 Time-stepping for passive LCS . . . 277
8 NSCD method. Linear case. . . . 299
9 NSCD method. Nonlinear case with Newton's method . . . 304
10 Sketch of the active-set method for convex QP . . . 340
11 Sketch of the general splitting scheme for the LCP . . . 363
12 Sketch of the general splitting scheme with line search for the LCP(M,q) with M symmetric . . . 366
13 Murty's least index pivoting method . . . 369
14 Lemke's method . . . 372
15 General scheme of the primal–dual interior point methods . . . 377
16 Hyperplane projection method (Konnov, 1993) . . . 400
17 Fixed-point algorithm on the friction threshold . . . 436
18 Analytical resolution of the two-dimensional frictional contact subproblem . . . 438
19 Approximate solution of the three-dimensional frictional contact subproblem . . . 440

1

Nonsmooth Dynamical Systems: Motivating Examples and Basic Concepts

The aim of this introductory material is to show how one may write the dynamical equations of several physical systems like simple electrical circuits with nonsmooth elements, and simple mechanical systems with unilateral constraints on the position and impacts, Coulomb friction. We start with circuits with ideal diodes, then circuits with ideal Zener diodes. Then a mechanical system with Coulomb friction is analyzed, and the bouncing ball system is presented. These physical examples illustrate gradually how one may construct various mathematical equations, some of which are equivalent (i.e., the same "initial" data produce the same solutions). In each case we also derive the time-discretization of the continuous-time dynamics, and gradually highlight the discrepancy from one system to the next. All the presented tools and algorithms that are briefly presented in this chapter will be more deeply studied further in the book.

1.1 Electrical Circuits with Ideal Diodes

Though this book is mainly concerned with mechanical systems, electrical circuits will also be considered. The reasons are that on one hand electrical circuits with nonsmooth elements are an important class of physical systems, on the other hand their dynamics can nicely be recast in the family of evolution problems like differential inclusions, variational inequalities, complementarity systems, and some piecewise smooth systems. There is therefore a strong analogy between nonsmooth circuits and nonsmooth mechanical systems. This similarity will naturally exist also at the level of numerical simulation, which is the main object of this book.

The objective of this section is to show that electrical circuits containing so-called ideal diodes possess a dynamics which can be interpreted in various ways. It can be written as a complementarity system, a differential inclusion, an evolution variational inequality, or a variable structure system. What these several formalisms really mean will be made clear with simple examples.

1.1.1 Mathematical Modeling Issues

Let us consider the four electrical circuits depicted in Fig. 1.3. The diodes are supposed to be ideal, i.e., the characteristic between the current $i(t)$ and the voltage $v(t)$ (see Fig. 1.1a for the notation) satisfies the *complementarity* conditions:

$$0 \leqslant i(t) \perp v(t) \geqslant 0\,. \tag{1.1}$$

This set of conditions merely means that both the variables current $i(t)$ and voltage $v(t)$ have to remain nonnegative at all times t and that they have to be orthogonal one to each other. So $i(t)$ can be positive only if $v(t) = 0$, and vice versa. The complementarity condition (1.1) between the current across the diode and its voltage certainly represents the most natural way to define the diode characteristic. It is quite similar

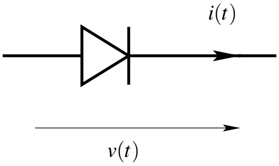


Fig. 1.1a. The diode component

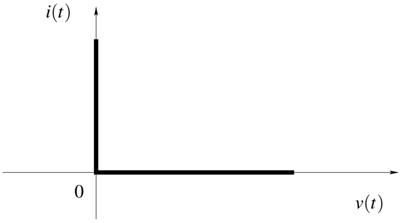


Fig. 1.1b. Characteristics of an ideal diode. A complementarity condition

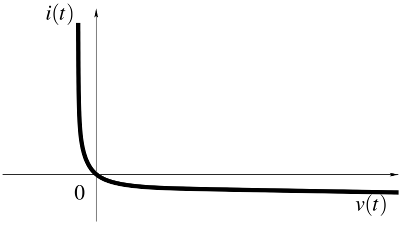


Fig. 1.1c. The graph of the Shockley's law

to the relations between the contact force and the distance between the system and an obstacle, in unilateral mechanics,[1] see Sect. 1.4.

Naturally, other models can be considered for the diode component. The well-known Shockley's law, which is one of the numerous models that can be found in standard simulation software, can be defined as

$$i = i_{\mathrm{s}} \exp(-\frac{v}{\alpha} - 1) , \tag{1.2}$$

where the constant α depends mainly on the temperature. This law is depicted in Fig. 1.1c. This model may be considered to be more physical than the ideal one, because the residual saturation current, i_{s} is taken into account as a function of the voltage across the diode. The same remark applies in mechanics for a compliant contact model with respect to unilateral rigid contact model. Nevertheless, in the numerical practice, the ideal model reveals to be better from the qualitative point of view and also from the quantitative point of view. One of the reasons is that exchanging the highly stiff nonlinear model as in (1.2) by a nonsmooth multivalued model (1.1) leads to more robust numerical schemes. Moreover it is easy to introduce a residual current in the complementarity formalism as follows:

$$0 \leqslant i(t) + \varepsilon_1 \perp v(t) + \varepsilon_2 \geqslant 0 \tag{1.3}$$

for some $\varepsilon_1 \geqslant 0$, $\varepsilon_2 \geqslant 0$. This results in a shift of the characteristic of Fig. 1.1b.

The relation in (1.1) will necessarily enter the dynamics of a circuit containing ideal diodes. It is consequently crucial to clearly understand its meaning. Let us notice that the relation in (1.1) defines the *graph* of a *multivalued function* (or multifunction, or set-valued function), as it is clear that it is satisfied for any $i(t) \geqslant 0$ if $v(t) = 0$. This graph is depicted in Fig. 1.1b.

Using basic convex analysis (which in particular will allow us to accurately define what is meant by the gradient of a function that is not differentiable in the usual way), a nice interpretation of the relation in (1.1) and of its graph in Fig. 1.1b can be obtained with *indicator functions of convex sets*. The indicator of a set K is defined as

$$\psi_K(x) = \begin{cases} 0 & \text{if } x \in K \\ +\infty & \text{if } x \notin K \end{cases} . \tag{1.4}$$

This function is highly nonsmooth on the boundary ∂K of K, since it even possesses an infinite jump at such points! It is therefore nondifferentiable at $x \in \partial K$. Nevertheless, if K is a convex set then $\psi_K(\cdot)$ is a convex function, and it is *subdifferentiable* in the sense of convex analysis. Roughly speaking, one will consider *subgradients* instead of the usual gradient of a differentiable function. The subgradients of a convex function are vectors γ defining the directions "under" the graph of the function. More precisely, γ is a subgradient of a convex function $f(\cdot)$ at x if and only if it satisfies

[1] At this stage the similarity between both remains at a pure formal level. Indeed a more physical analogy would lead us to consider that it is rather a relation between a velocity and a force that corresponds to (1.1).

$$f(y) - f(x) \geqslant \gamma^{\mathrm{T}}(y - x) \tag{1.5}$$

for all y. Normally the subdifferential is denoted as $\partial f(\cdot)$, and $\partial f(x)$ can be a set (containing the subgradients γ).

Let us now consider the particular case of the indicator function of $K = I\!R^+ = \{x \in \mathrm{I\!R} \mid x \geqslant 0\}$. Though this might be at first sight surprising, this function is subdifferentiable at $x = 0$. Its subdifferential is given by

$$\partial \psi_{I\!R^+}(x) = \begin{cases} \{0\} & \text{if } x > 0 \\ (-\infty, 0] & \text{if } x = 0 \end{cases} . \tag{1.6}$$

Indeed one checks that when $x \geqslant 0$, then $\psi_{I\!R^+}(y) \geqslant \gamma(y - x)$ for all $y \in I\!R$ can be satisfied if and only if $\gamma = 0$. Now if $x = 0$, $\psi_{I\!R^+}(y) \geqslant \gamma y$ is satisfied for all $y \in I\!R$ if and only if $\gamma \leqslant 0$. One sees that at $x = 0$ the subdifferential is a set, since it is a complete half space. In fact the set $\partial \psi_{I\!R^+}(x)$ is equal to the so-called normal cone to $I\!R^+$ at the point x (Fig. 1.2). This can be generalized to convex sets $K \subset I\!R^n$, so that the subdifferential $\psi_K(x)$ is the normal cone to the set K, computed at the point $x \in K$ and denoted by $N_K(x)$. If the boundary of K is differentiable, this is simply a half-line normal to the tangent plane to K at x, and in the direction outward K.

It becomes apparent that the graph of the subdifferential of the indicator of $I\!R^+$ resembles a lot the corner law depicted in Fig. 1.1b. Actually, one can now deduce from (1.6) and (1.1) that

$$i(t) \in -\partial \psi_{I\!R^+}(v(t)) \quad \Longleftrightarrow \quad v(t) \in -\partial \psi_{I\!R^+}(i(t)) . \tag{1.7}$$

The symmetry between these two inclusions is clear from Fig. 1.1b: if one inverts the multifunction (exchange $i(t)$ and $v(t)$ in Fig. 1.1b), then one obtains exactly the same graph. Actually this is a very particular case of duality between two variables. In a more general setting the graph inversion procedure does not yield the graph of the same multifunction, but the graph of its conjugate. And inverting once again allows

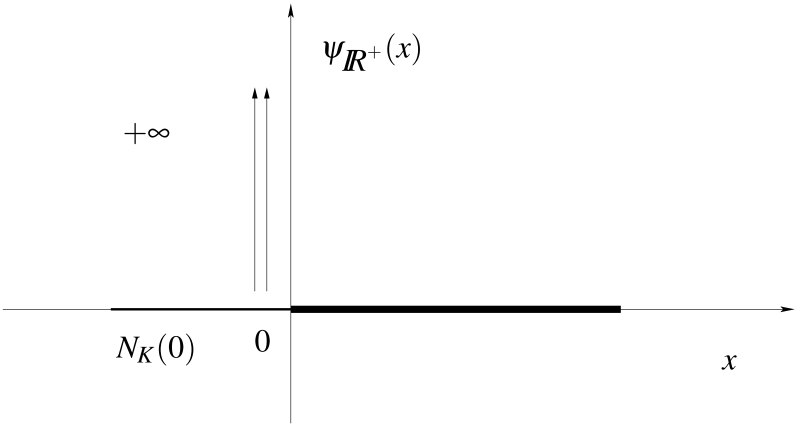


Fig. 1.2. The indicator function of $I\!R^+$ and the normal cone at $x = 0$, $N_K(0) = \partial \psi_{I\!R^+}(0) = I\!R^-$

one to recover the original graph under some convexity and properness assumption: this is the very basic principle of duality (Luenberger, 1992).

Let us focus now on the *inclusions* in (1.7). As a matter of fact, one may check that the first one is equivalent to: for any $v(t) \geqslant 0$,

$$\langle i(t), u - v(t) \rangle \geqslant 0, \ \forall \ u \geqslant 0 \tag{1.8}$$

and to: for any $i(t) \geqslant 0$,

$$\langle v(t), u - i(t) \rangle \geqslant 0, \ \forall \ u \geqslant 0 \ . \tag{1.9}$$

The objects in (1.8) and (1.9) are called a *Variational Inequality (VI)*.

We therefore have three different ways of looking at the ideal diode characteristic: the complementarity relations in (1.1), the inclusion in (1.7), and the variational inequality in (1.8). Our objective now is to show that when introduced into the dynamics of an electrical circuit, these formalisms give rise to various types of dynamical systems as enumerated at the beginning of this section.

Remark 1.1. Another variational inequality can also be written: for all $i(t) \geqslant 0$, $v(t) \geqslant 0$,

$$\langle j - i(t), u - v(t) \rangle \geqslant 0 \ , \ \forall \ j, u \geqslant 0 \ . \tag{1.10}$$

Having attained this point, the reader might legitimately wonder what is the usefulness of doing such an operation, and what has been gained by rewriting (1.1) as in (1.7) or as in (1.8). Let us answer a bit vaguely: several formalisms are likely to be useful for different tasks which occur in the course of the study of a dynamical system (mathematical analysis, time-discretization and numerical simulation, analysis for control, feedback control design, and so on). In this introductory chapter, we just ask the reader to trust us: all these formalisms are useful and are used. We will see in the sequel that there exists a lot of other ways to write the complementary condition such as zeroes of special functions or extremal points of a functional. All these formulations will lead to specific ways of studying and solving the system.

1.1.2 Four Nonsmooth Electrical Circuits

In order to derive the dynamics of an electrical circuit we need to consider Kirchoff's laws as well as the constitutive relations of devices like resistors, inductors, and capacitors (Chua et al., 1991). The constitutive relation of the ideal diode is the complementarity relation (1.1) while in the case of resistors, inductors, and capacitors we have the classical linear relations between variables like voltages, currents, and charges. Thus, taking into account those constitutive relations and using Kirchoff's laws it follows that the dynamical equations of the four circuits depicted in Fig. 1.3 are given by

$$(\mathbf{a}) \quad \begin{cases} \dot{x}_1(t) = x_2(t) - \dfrac{1}{RC}x_1(t) - \dfrac{\lambda(t)}{R} \\ \dot{x}_2(t) = -\dfrac{1}{LC}x_1(t) - \dfrac{\lambda(t)}{L} \\ 0 \leqslant \lambda(t) \perp w(t) = \dfrac{\lambda(t)}{R} + \dfrac{1}{RC}x_1(t) - x_2(t) \geqslant 0 \end{cases} \tag{1.11}$$

$$(\mathbf{b}) \quad \begin{cases} \dot{x}_1(t) = -x_2(t) + \lambda(t) \\ \dot{x}_2(t) = \dfrac{1}{LC}x_1(t) \\ 0 \leqslant \lambda(t) \perp w(t) = \dfrac{1}{C}x_1(t) + R\lambda(t) \geqslant 0 \end{cases} \tag{1.12}$$

$$(\mathbf{c}) \quad \begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) = -\dfrac{R}{L}x_2(t) - \dfrac{1}{LC}x_1(t) - \dfrac{\lambda(t)}{L} \\ 0 \leqslant \lambda(t) \perp w(t) = -x_2(t) \geqslant 0 \end{cases} \tag{1.13}$$

$$(\mathbf{d}) \quad \begin{cases} \dot{x}_1(t) = x_2(t) - \dfrac{1}{RC}x_1(t) \\ \dot{x}_2(t) = -\dfrac{1}{LC}x_1(t) + \dfrac{\lambda(t)}{L} \\ 0 \leqslant \lambda(t) \perp w(t) = x_2(t) \geqslant 0 \end{cases} \tag{1.14}$$

where we considered the current through the inductors for the variable $x_2(t)$, and for the variable $x_1(t)$ the charge on the capacitors as state variables.

Let us now make use of the above equivalent formalisms to express the dynamics in (1.11)–(1.14) in various ways. We will generically call the dynamics in (1.11)–(1.14) a Linear Complementarity System (LCS), a terminology introduced in van der Schaft & Schumacher (1996). An LCS therefore consists of a linear differential equation with state (x_1, x_2), an external signal $\lambda(\cdot)$ entering the differential equation, and a set of complementarity conditions which relate a variable $w(\cdot)$ and $\lambda(\cdot)$. Since $w(\cdot)$ is itself a function of the state and possibly of $\lambda(\cdot)$, the complementarity conditions play a crucial role in the dynamics. The variable λ may be interpreted as a Lagrange multiplier.

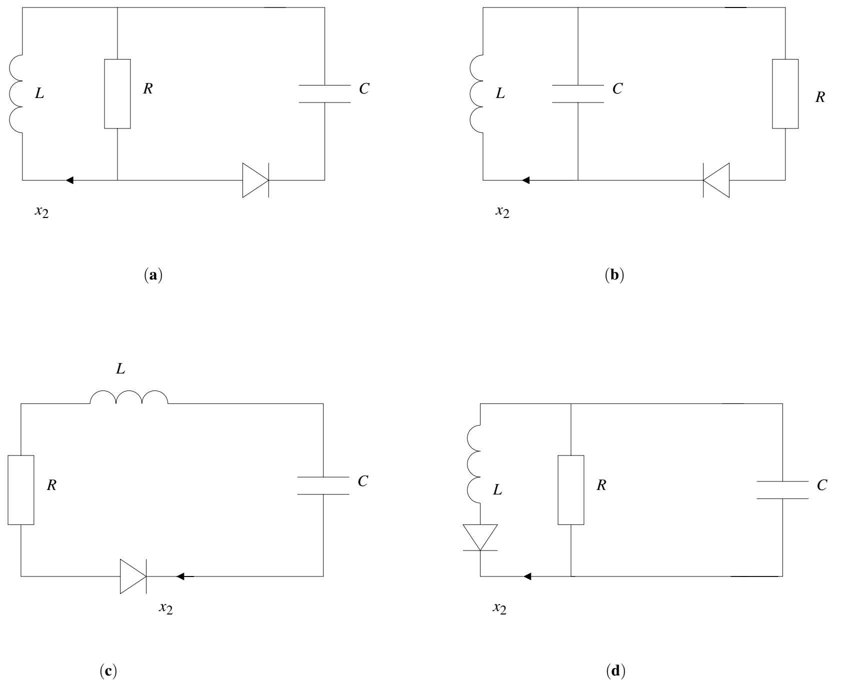


Fig. 1.3. RLC circuits with an ideal diode

1.1.3 Continuous System (Ordinary Differential Equation)

Let us consider for instance the circuit (**a**) whose dynamics is in (1.11). Its complementarity conditions are given by

$$\begin{cases} w(t) = \dfrac{\lambda(t)}{R} + \dfrac{1}{RC}x_1(t) - x_2(t) \\ \\ 0 \leqslant \lambda(t) \perp w(t) \geqslant 0 \end{cases} . \tag{1.15}$$

If we consider $\lambda(t)$ as the unknown of this problem, then the question we have to answer to is: does it possess a solution, and if yes is this solution unique? Here we must introduce a basic tool that is ubiquitous in complementarity systems: the Linear Complementarity Problem (LCP). An LCP is a problem which consists of solving a set of complementarity relations as

$$\begin{cases} w = M\lambda + q \\ \\ 0 \leqslant \lambda \perp w \geqslant 0 \end{cases} , \tag{1.16}$$

where M is a constant matrix and q a constant vector, both of appropriate dimensions. The inequalities have to be understood component-wise and the relation $w \perp \lambda$ means $w^{\mathrm{T}}\lambda = 0$. A fundamental result on LCP (see Sect. 12.4) guarantees that there is a unique λ that solves the LCP in (1.16) for any q if and only if M is a so-called P-matrix (i.e., all its principal minors are positive). In particular, positive definite matrices are P-matrices.

Taking this into account, it is an easy task to see that there is a unique solution $\lambda(t)$ to the LCP in (1.15) given by

$$\lambda(t) = 0 \qquad \text{if } \frac{1}{RC}x_1(t) - x_2(t) \geqslant 0\,, \tag{1.17}$$

$$\lambda(t) = -\frac{1}{C}x_1(t) + Rx_2(t) > 0 \qquad \text{if } \frac{1}{RC}x_1(t) - x_2(t) < 0\,. \tag{1.18}$$

Evidently we could have solved this LCP without resorting to any general result on existence and uniqueness of solutions. However, we will often encounter LCPs with several tenth or even hundreds of variables (i.e., the dimension of M in (1.16) can be very large in many applications). In such cases solving the LCP "with the hands" rapidly becomes intractable. So $\lambda(t)$ in (1.11) considered as the solution at time t of the LCP in (1.15) can take two values, and only two, for all $t \geqslant 0$.

Another way to arrive at the same result for circuit (**a**) is to use once again the equivalence between (1.1) and (1.7). It is straightforward then to see that (1.15) is equivalent to

$$\lambda(t) + \frac{1}{C}x_1(t) - Rx_2(t) \in -\partial\psi_{\mathbb{R}^+}(\lambda(t)) \tag{1.19}$$

(we have multiplied the left-hand side by R and since $\partial\psi_{\mathbb{R}^+}(\lambda(t))$ is a cone $R\partial\psi_{\mathbb{R}^+}(\lambda(t)) = \partial\psi_{\mathbb{R}^+}(\lambda(t))$). It is well known in convex analysis (see Appendix A) that (1.19) is equivalent to

$$\lambda(t) = \mathrm{Proj}_{\mathbb{R}^+}\left[-\frac{1}{C}x_1(t) + Rx_2(t)\right]\,, \tag{1.20}$$

where $\mathrm{Proj}_{\mathbb{R}^+}$ is the projection on $\mathbb{R}^+$. Since $\mathbb{R}^+$ is convex (1.20) possesses a unique solution. Once again we arrive at the same conclusion. The surface that splits the phase space (x_1, x_2) in two parts corresponding to the "switching" of the LCP is the line $-\frac{1}{C}x_1(t) + Rx_2(t) = 0$. On one side of this line $\lambda(t) = 0$, and on the other side $\lambda(t) = -\frac{1}{C}x_1(t) + Rx_2(t) > 0$. We may write (1.11) as

$$\begin{cases} \begin{bmatrix} \dot{x}_1(t) = x_2(t) - \dfrac{1}{RC} x_1(t) \\ \dot{x}_2(t) = -\dfrac{1}{LC} x_1(t) \end{bmatrix} & \text{if } -\dfrac{1}{C} x_1(t) + R x_2(t) < 0\,, \\ \begin{bmatrix} \dot{x}_1(t) = 0 \\ \dot{x}_2(t) = -\dfrac{R}{L} x_2(t) \end{bmatrix} & \text{if } -\dfrac{1}{C} x_1(t) + R x_2(t) \geqslant 0\,, \end{cases} \tag{1.21}$$

that is a piecewise linear system, or as

$$\dot{x}(t) - Ax(t) = B \,\mathrm{Proj}_{\mathbb{R}^+} \left[-\frac{1}{C} x_1(t) + R x_2(t) \right]\,, \tag{1.22}$$

where the matrices A and B can be easily identified.

The fact that the projection operator in (1.20) is a Lipschitz-continuous single-valued function (Goeleven et al., 2003a) shows that the equation (1.22) is an Ordinary Differential Equation (ODE) with a Lipschitz-continuous vector field.[2] We therefore conclude that this complementarity system possesses a global unique and differentiable solution, as a standard result on ODEs (Coddington & Levinson, 1955).

Exactly the same analysis can be done for the circuit (**b**) which is also an ODE.

1.1.4 Hints on the Numerical Simulation of Circuits (a) and (b)

The circuit (**a**) can be simulated with any standard one-step and multistep methods like explicit or implicit (backward) Euler, mid-point, or trapezoidal rules (Hairer et al., 1993, Chap. II.7), which apply to ordinary differential equations with a Lipschitz right-hand side. Nevertheless, all these methods behave globally as a method of order one as the right-hand side is not differentiable everywhere (Hairer et al., 1993; Calvo et al., 2003).

As an illustration, a simple trajectory of the circuit (**a**) is computed with an explicit Euler scheme and a standard Runge–Kutta of order 4 scheme. The results are depicted in Fig. 1.4. With the initial conditions, $x_1(0) = 1$, $x_2(0) = -1$, we observe only one event or switch from one mode to the other. Before the switch, the dynamics is a linear oscillator in x_1 and after the switch, it corresponds to a exponential decay in x_1.

We present in Fig. 1.5 a slightly more rich dynamics with the circuit (**b**), which corresponds to a half-wave rectifier. When the diode blocks the current, $\lambda = 0, w > 0$, the dynamics of the circuit is a pure linear LC oscillator in x_2. When the constraint is active $\lambda > 0, w = 0$ and the diode lets the positive current pass: the dynamics is a damped linear oscillator in x_1. The interest of the circuit (**b**) with respect to the circuit (**a**) is that if R is small other switches are possible in circuit (**b**).

[2] It is also known that the solutions of LCPs as in (1.16) with M a P-matrix are Lipschitz-continuous functions of q (Cottle et al., 1992, Sect. 7.2). So we could have deduced this result from (1.15) and the complementarity formalism of the circuit.

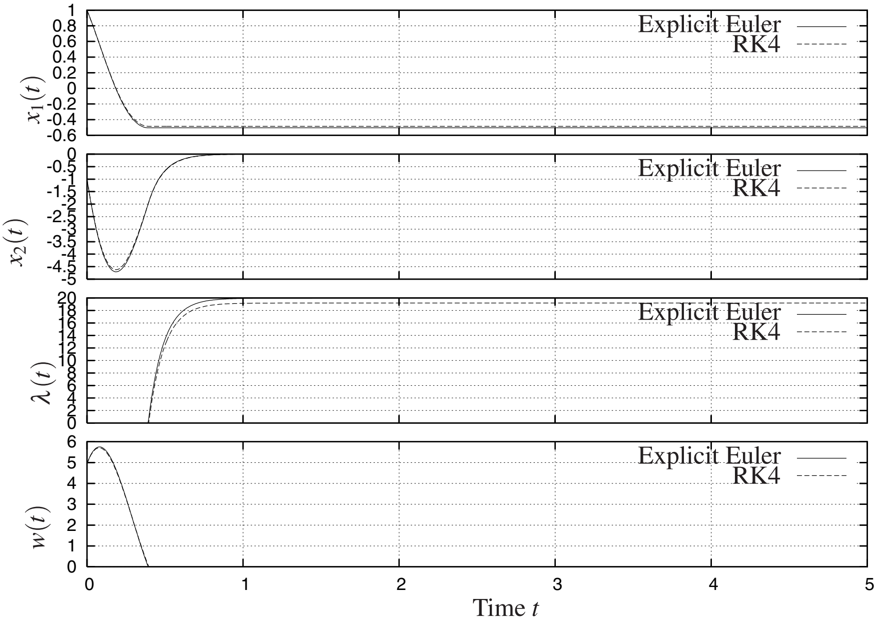


Fig. 1.4. Simulation of the circuit (**a**) with the initial conditions $x_1(0) = 1,\ x_2(0) = -1$ and $R = 10,\ L = 1,\ C = \dfrac{1}{(2\pi)^2}$. Time step $h = 5 \times 10^{-3}$

The Question of the Order

It is noteworthy that even in this simple case, where the "degree" of nonsmoothness is rather low (said otherwise, the system is a gentle nonsmooth system), applying higher order "time-stepping" methods which preserve the order $p \geqslant 2$ is not straightforward. By *time-stepping method*, we mean here a time-discretization method which does not consider explicitly the possible times at which the solution is not differentiable in the process of integration.

Let us now quote some ideas from Grüne & Kloeden (2006) which accurately explain the problem of applying standard higher order schemes: *In principle known numerical schemes for ordinary differential equations such as Runge–Kutta schemes can be applied to switching systems, changing the vector field after each switch has occurred. However, in order to maintain the usual consistency order of these schemes, the integration time steps need to be adjusted to the switching times in such a way that switching always occurs at the end of an integration interval. This is impractical in the case of fast switching, because in this case an adjustment of the scheme's integration step size to the switching times would lead to very small time steps causing an undesirably high computational load.* Such a method for the time integration of nonsmooth systems, which consists in locating and adjusting the time step to the events will be called an *event-driven method*. If the location of the events

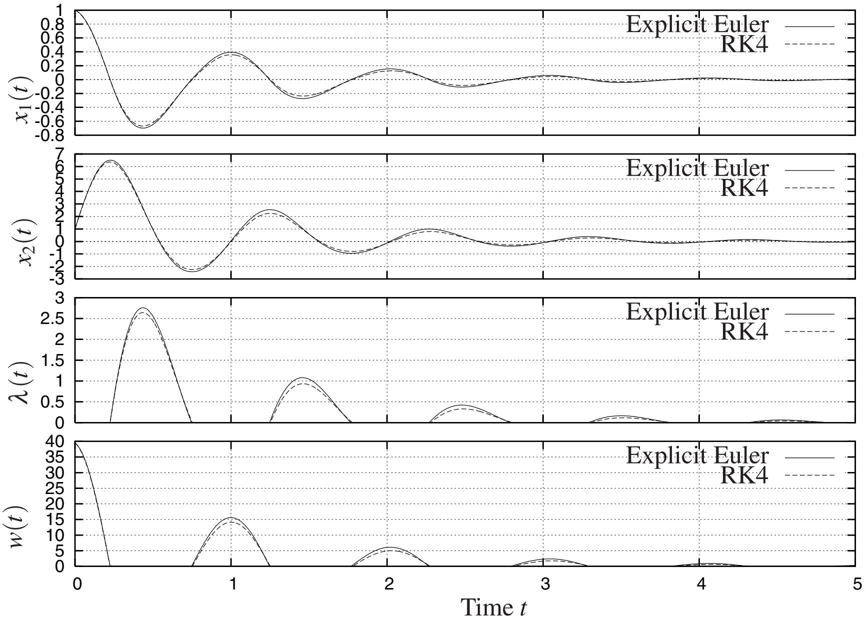


Fig. 1.5. Simulation of the circuit (**b**) with the initial conditions $x_1(0) = 1$, $x_2(0) = 1$ and $R = 10$, $L = 1$, $C = \frac{1}{(2\pi)^2}$. Time step $h = 5 \times 10^{-3}$

is sufficiently accurate, the global order of the integration method can be retrieved. If one is not interested in maintaining the order of the scheme larger than one, however, one may apply Runge–Kutta methods directly to an ODE as (1.22).

There are three main conclusions to be retained from this:

1. When the instants of nondifferentiability are not known in advance, or when there are too many such times, then applying an "event-driven" method with order larger than one may not be tractable.
2. We may add another drawback of event-driven methods that may not be present in the system we have just studied, but will frequently occur in the systems studied in this book. Suppose that the events (or times of nondifferentiability, or switching times) possess a finite accumulation point. Then an event-driven scheme will not be able to go further than the accumulation, except at the price of continuing the integration with some ad hoc, physically and mathematically unjustified trick.
3. Finally, there exist higher order standard numerical schemes which continue to perform well for some classes of nonsmooth systems, but at the price of decreasing the global order to one (see Sect. 9.2). However, this global low-order behavior can be compensated by an adaptive time-step strategy which takes benefits from the high accuracy of the time-integration scheme on smooth phases.

It is noteworthy that the events that will be encountered in the systems examined throughout the book usually are not exogenous events but state dependent, hence not known in advance. Therefore, the choice between the event-driven methods or the time-stepping methods depends strongly on the type of systems under study. We will come back later on the difference between time-stepping and event-driven numerical schemes and their respective ranges of applications (especially for mechanical systems).

The Question of the Stability of Explicit Schemes

As we said earlier, the nonsmoothness of the right-hand side destroys the order of convergence of the standard time-stepping integration scheme. Another aspect is the stability, especially for explicit schemes. Most of the results on the stability of numerical integration schemes are based on the assumption of sufficient regularity of the right-hand side.

The question of the simulation of ODEs with discontinuities will be discussed in Sects. 7.2 and 9.1. Some numerical illustrations of troubles in terms of the order of convergence and the stability of the methods are given in Sect. 9.1 where the dynamics of the circuits (**a**) and (**b**) are simulated.

1.1.5 Unilateral Differential Inclusion

Let us now turn our attention to circuit (**c**). This time the complementarity relations are given by

$$0 \leqslant \lambda(t) \perp w(t) = -x_2(t) \geqslant 0\,. \tag{1.23}$$

Contrary to (1.15), it is not possible to calculate $\lambda(t)$ directly from this set of relations. At first sight there is no LCP that can be constructed (indeed now we have a zero matrix M).

Let us, however, imagine that there is a time interval $[\tau,\tau+\varepsilon)$, $\varepsilon > 0$, on which the solution $x_2(t) = 0$ for all $t \in [\tau,\tau+\varepsilon)$. Then on $[\tau,\tau+\varepsilon)$ one has necessarily $-\dot{x}_2(t) \geqslant 0$, otherwise the unilateral constraint $-x_2(t) \geqslant 0$ would be violated. Actually all the derivatives of $x_2(\cdot)$ are identically 0 on $[\tau,\tau+\varepsilon)$. The interesting question is: what happens on the right of $t = \tau+\varepsilon$? Is there one derivative of $x_2(\cdot)$ that becomes positive, so that the system starts to detach from the constraint $x_2 = 0$ at $t = \tau+\varepsilon$? Such a question is important, think for instance of numerical simulation: one will need to implement a correct test to determine whether or not the system keeps evolving on the constraint, or quits it. In fact the test consists of considering the further complementarity condition

$$0 \leqslant \lambda(t^+) \perp -\dot{x}_2(t) = \frac{R}{L}x_2(t^+) + \frac{1}{LC}x_1(t) + \frac{\lambda(t^+)}{L} \geqslant 0 \tag{1.24}$$

which is an LCP to be solved only when $x_2(t) = 0$. The fact that this LCP possesses a solution $\lambda(t) - \dot{x}_2(t) > 0$ is a sufficient condition for the system to change its *mode* of evolution. We can solve for $\lambda(t)$ in (1.24) exactly as we did for (1.15). Both are

LCPs with a unique solution. However, this time the resulting dynamical system is not quite the same, since we have been obliged to follow a different path to get the LCP in (1.24).

In order to better realize this big discrepancy, let us use once again the equivalence between (1.1) and (1.7). We obtain that $\lambda(t) \in -\partial\psi_{\mathbb{R}^+}(-x_2(t))$. Inserting this inclusion in the dynamics (1.13) yields

$$(\mathbf{c}) \quad \begin{cases} \dot{x}_1(t) - x_2(t) = 0 \\ \\ \dot{x}_2(t) + \dfrac{R}{L}x_2(t) + \dfrac{1}{LC}x_1(t) \in \dfrac{1}{L}\partial\psi_{\mathbb{R}^+}(-x_2(t)) \end{cases} \tag{1.25}$$

where it is implicitly assumed that $x_2(0) \leqslant 0$ so that the inequality constraint $x_2(t) \leqslant 0$ will be satisfied for all $t \geqslant 0$.

Passing from the LCP (1.23) to the LCP (1.24) and then from (1.13) to (1.25) can be viewed similarly as the index-reduction operation in a Differential Algebraic Equation (DAE). Indeed, the LCP on x_2 in (1.23) is replaced by the LCP on $\dot{x}_2$ in (1.24).

Unilateral Differential Inclusion

More compactly, (1.25) can be rewritten as

$$-\dot{x}(t) + Ax(t) \in B\partial\psi_{\mathbb{R}^+}(w(t)) \tag{1.26}$$

which we can call a Unilateral Differential Inclusion (UDI) where the matrices A and B can be easily identified. The reason why we employ the word *unilateral* should be obvious. It is noteworthy that the right-hand side of (1.26) is generally a set that is not reduced to a single element, see (1.6). It is also noteworthy that the complementarity conditions are included in the UDI in (1.26). Obviously, the dynamics in (1.26) is not a variable structure or discontinuous vector field system. It is something else.

Evolution Variational Inequality

Using a suitable change of coordinate $z = Rx$, $R = R^{\mathrm{T}} > 0$, it is possible to show (Goeleven & Brogliato, 2004; Brogliato, 2004) that (1.26) can also be seen as an Evolution Variational Inequality (EVI). This time we make use of the equivalence between (1.7) and (1.8) and of a property of electrical circuits composed of resistors, capacitors, and inductances (they are dissipative). Then (1.26) is equivalent to the EVI

$$\begin{cases} \left\langle \dfrac{\mathrm{d}z}{\mathrm{d}t}(t) - RAR^{-1}z(t), v - z(t) \right\rangle \geqslant 0, \forall\, v \in K, \text{ a.e. } t \geqslant 0 \\ \\ z(t) \in K, t \geqslant 0\ , \end{cases} \tag{1.27}$$

where $K = \{(z_1, z_2) | -(0\ 1)\, R^{-1} z \geqslant 0\}$ and a.e. means almost everywhere (the solution not being a priori differentiable everywhere). As a consequence of how the set K is constructed, having $z(t) \in K$ is equivalent to having $x_2(t) \leqslant 0$. In fact it can be shown that the EVI in (1.27) possesses unique continuous solutions which are right differentiable (Goeleven & Brogliato, 2004). It is remarkable at this stage to notice that both (1.22) and (1.26) possess unique continuous solutions, however, the solutions of the inclusion (1.26) are less regular.

1.1.6 Hints on the Numerical Simulation of Circuits (c) and (d)

Let us now see how the differential inclusion (1.26) and the LCS in (1.13) may be time-discretized for numerical simulation purpose. Let us start with the LCS in (1.13).

A Direct Backward Euler Scheme

Mimicking the backward Euler discretization for ODEs, a time-discretization of (1.13) is

$$\begin{cases} x_{1,k+1} - x_{1,k} = h x_{2,k+1} \\ x_{2,k+1} - x_{2,k} = -h\frac{R}{L} x_{2,k+1} - \frac{h}{LC} x_{1,k+1} - \frac{h}{L}\lambda_{k+1} \\ 0 \leqslant \lambda_{k+1} \perp -x_{2,k+1} \geqslant 0 \end{cases}, \tag{1.28}$$

where x_k is the value, at time t_k of a grid $t_0 < t_1 < \cdots < t_N = T$, $N < +\infty$, $h = \frac{T - t_0}{N} = t_k - t_{k-1}$, of a step function $x^N(\cdot)$ that approximates the analytical solution $x(\cdot)$.

Let us denote $a(h) = 1 + h\frac{R}{L} + h^2 \frac{1}{LC}$. Then we can rewrite (1.28) as

$$\begin{cases} x_{1,k+1} - x_{1,k} = h x_{2,k+1} \\ x_{2,k+1} = (a(h))^{-1} \left\{ x_{2,k} - \frac{h}{LC} x_{1,k} - \frac{h}{L}\lambda_{k+1} \right\} \\ 0 \leqslant \lambda_{k+1} \perp -(a(h))^{-1} \left\{ x_{2,k} - \frac{h}{LC} x_{1,k} \right\} + (a(h))^{-1} \frac{h}{L}\lambda_{k+1} \geqslant 0 \end{cases}. \tag{1.29}$$

Remark 1.2. This time-stepping scheme is made of a discretization of the continuous dynamics (the first two lines of (1.29)) and of a LCP whose unknown is λ_{k+1}. We shall call later on the LCP resolution a one-step algorithm. Here the LCP is scalar and can easily be solved by inspection. In higher dimensions specific solvers will be necessary. This is the object of Part III of this book.

Remark 1.3. The LCP matrix M (here a scalar) is equal to $(a(h))^{-1}\frac{h}{L} > 0$ for all $h > 0$, which tends to 0 as $h \to 0$. This is not very good in practice when very small steps are chosen. To cope with this issue, let us choose as the unknown the variable $\bar{\lambda}_{k+1} = h\lambda_{k+1}$. We then solve the LCP

$$0 \leqslant \bar{\lambda}_{k+1} \perp -(a(h))^{-1}\left\{x_{2,k} - \frac{h}{LC}x_{1,k}\right\} + (a(h))^{-1}\frac{1}{L}\bar{\lambda}_{k+1} \geqslant 0\,. \tag{1.30}$$

It is noteworthy that this does not change the result of the algorithm, because the set of nonnegative reals is a cone. This LCP is easily solved:

$$\text{If } x_{2,k} - \frac{h}{LC}x_{1,k} < 0, \text{ then } \bar{\lambda}_{k+1} = 0\,. \tag{1.31}$$

$$\text{If } x_{2,k} - \frac{h}{LC}x_{1,k} \geqslant 0 \text{ then } \bar{\lambda}_{k+1} = L\left\{x_{2,k} - \frac{h}{LC}x_{1,k}\right\} \geqslant 0\,. \tag{1.32}$$

Inserting these values into (1.29) we get:

$$x_{2,k+1} = \begin{cases} (a(h))^{-1}\left\{x_{2,k} - \frac{h}{LC}x_{1,k}\right\} & \text{if } x_{2,k} - \frac{h}{LC}x_{1,k} < 0 \\ \\ 0 & \text{if } x_{2,k} - \frac{h}{LC}x_{1,k} \geqslant 0 \end{cases}\,. \tag{1.33}$$

A Discretization of the Differential Inclusion (1.26)

Let us now propose an implicit time-discretization of the differential inclusion in (1.26), as follows:

$$\begin{cases} x_{1,k+1} - x_{1,k} = hx_{2,k+1} \\ \\ x_{2,k+1} - x_{2,k} + \frac{hR}{L}x_{2,k+1} + \frac{h}{LC}x_{1,k+1} \in \frac{1}{L}\partial\psi_{\mathbb{R}^+}(-x_{2,k+1}) \end{cases} \tag{1.34}$$

Notice that we can rewrite the second line of (1.34) as

$$x_{2,k+1} - (a(h))^{-1}\left\{x_{2,k} - \frac{h}{LC}x_{1,k}\right\} \in \partial\psi_{\mathbb{R}^+}(-x_{2,k+1}) \tag{1.35}$$

where we have dropped the factor $\frac{1}{L}$ because $\partial\psi_{\mathbb{R}^+}(-x_{2,k+1})$ is a cone.

Let us now use two properties from convex analysis. Let $K \subset \mathbb{R}^n$ be a convex set, and let x and y be vectors of $\mathbb{R}^n$. Then

$$x - y \in -\partial\psi_K(x) \Longleftrightarrow x = \text{prox}[K;y]\,, \tag{1.36}$$

where "prox" means the closest element of K to y in the Euclidean metric, i.e., $x = \text{argmin}_{z\in K}\frac{1}{2} \parallel z - y \parallel^2$ (see (A.8) for a generalization in a metric M). Moreover using the chain rule of Proposition A.3 one has

$$\partial \psi_{I\!R^+}(-x) = -\partial \psi_{I\!R^-}(x) \, . \tag{1.37}$$

Using (1.36) and (1.37) one deduces from (1.35) that

$$x_{2,k+1} - (a(h))^{-1} \left\{ x_{2,k} - \frac{h}{LC} x_{1,k} \right\} \in -\partial \psi_{I\!R^-}(x_{2,k+1}) \, , \tag{1.38}$$

so that the algorithm becomes

$$\begin{cases} x_{1,k+1} - x_{1,k} = h x_{2,k+1} \\ \\ x_{2,k+1} = \mathrm{prox}\left[I\!R^-; (a(h))^{-1} \left\{ x_{2,k} - \dfrac{h}{LC} x_{1,k} \right\} \right] \end{cases} . \tag{1.39}$$

We therefore have proved the following:

Proposition 1.4. *The algorithm (1.28) is equivalent to the algorithm (1.34). They both allow one to advance from step k to step $k+1$, solving the proximation in (1.39).*

In Figs. 1.6 and 1.7, simulation results of the presented algorithm are given.

1.1.6.1 Approximating the Measure of an Interval

It is worthy to come back on the trick presented in Remark 1.3 that has been used to calculate the solution of the LCP in (1.29), i.e., to calculate $\bar{\lambda}_{k+1} = h\lambda_{k+1}$ rather than λ_{k+1}.

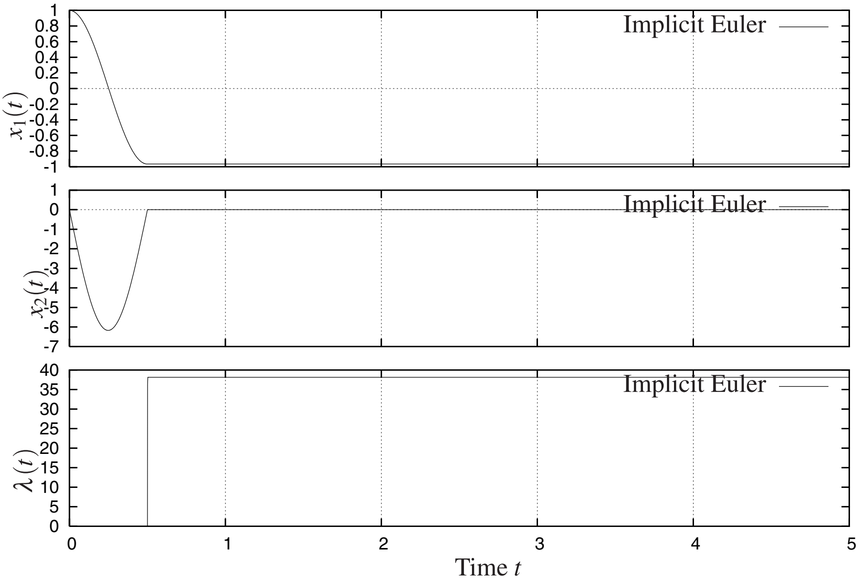


Fig. 1.6. Simulation of the circuit (**c**) with the initial conditions $x_1(0) = 1, \; x_2(0) = 0$ and $R = 0.1, \; L = 1, \; C = \frac{1}{(2\pi)^2}$. Time step $h = 1 \times 10^{-3}$

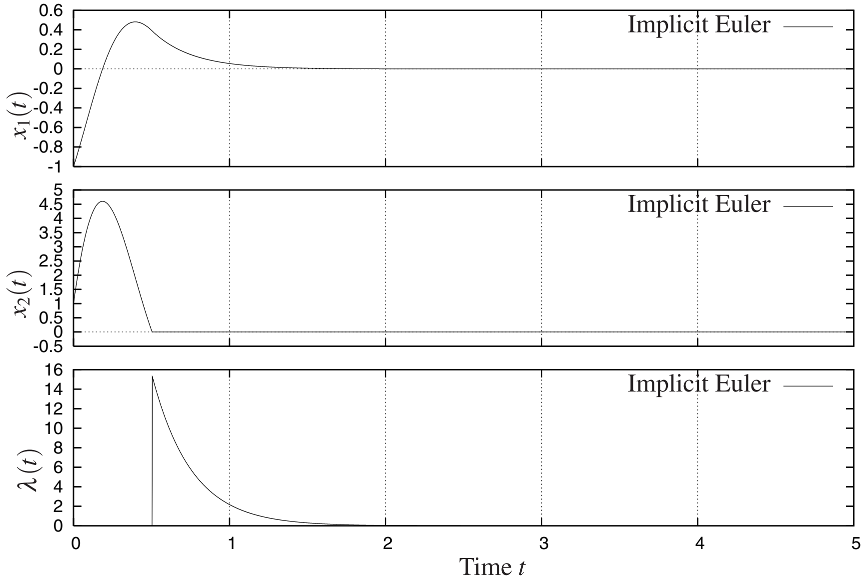


Fig. 1.7. Simulation of the circuit (**d**) with the initial conditions $x_1(0) = 1,\ x_2(0) = -1$ and $R = 10,\ L = 1,\ C = \frac{1}{(2\pi)^2}$. Time step $h = 1 \times 10^{-3}$

First of all, it follows from (1.34) and (1.28) that the element of the set $\partial \psi_{\mathbb{R}^+}(-x_{2,k+1})$ is not λ_{k+1}, but $\bar{\lambda}_{k+1}$. Retrospectively, our "trick" therefore appears not to be a trick, but a natural thing to do. Second, this means that the primary variables which are used in the integration are not $(x_{1,k}, x_{2,k}, \lambda_{k+1})$, but $(x_{1,k}, x_{2,k}, \bar{\lambda}_{k+1})$. Suppose that the initial value for the variable $x_2(\cdot)$ is negative. Then its right limit (supposed at this stage of the study to exist) has to satisfy $x_2(0^+) \geqslant 0$. Thus a jump occurs initially in $x_2(\cdot)$, so that the multiplier λ is at $t = 0$ a Dirac measure:[3]

$$\lambda = -L(x_2(0^+) - x_2(0^-))\delta_0 \tag{1.40}$$

The numerical scheme has to be able to approximate this measure! It is not possible numerically to achieve such a task, because this would mean approximating some kind of infinitely large value over one integration interval. However, what is quite possible is to calculate the value of

$$\mathrm{d}\lambda([t_k, t_{k+1}]) = \int_{[t_k, t_{k+1}]} \mathrm{d}\lambda\ , \tag{1.41}$$

i.e., the measure of the interval $[t_k, t_{k+1}]$.

[3] Throughout the book, right and left limits of a function F$(\cdot)$ will be denoted as $F(t^+)$ or $F^+(t)$, and $F(t^-)$ or $F^-(t)$, respectively.

Outside atoms of λ this is easy as λ is simply the Lebesgue measure. At atoms of λ this is again a bounded value. In fact, $\bar{\lambda}_{k+1} = h\lambda_{k+1}$ is an approximation of the measure of the interval by $\mathrm{d}\lambda$ i.e.,

$$\bar{\lambda}_{k+1} = h\lambda_{k+1} \approx \int_{[t_k,t_{k+1}]} \mathrm{d}\lambda \tag{1.42}$$

for each time-step interval.

Such an algorithm is therefore guaranteed to compute only *bounded* values, even if state jumps occur. Such a situation is common when we consider mechanical systems (see Sect. 1.4), dynamical complementarity systems (see Chap. 4), or higher relative degree systems (see Chap. 5).

Remark 1.5. A noticeable discrepancy between the equations (1.11) of the circuit (**a**) and the equations (1.13) of the circuit (**c**) is as follows. The complementarity relations in (1.11) are such that for any initial value of $x_1(\cdot)$ and $x_2(\cdot)$, there always exist a bounded value of the multiplier λ (which is a function of time and of the states) such that the integration proceeds. Such is not the case for (1.13), as pointed out just above. The *relative degree* r between w and λ plays a significant role in the dynamics (the relative degree is the number of times one needs to differentiate w in order to make λ appear explicitly: in (1.11) one has $r = 0$, but in (1.13) one has $r = 1$). A comprehensible presentation of the notion of relative degree is given in Chap. 4.

1.1.6.2 The Necessity of an Implicit Discretization

Another reason why considering the discretization of the inclusion in (1.25) is important is the following. Suppose one writes an explicit right-hand side $\partial\psi_{\mathbb{R}^+}(-x_{2,k})$ in (1.34) instead of the implicit form $\partial\psi_{\mathbb{R}^+}(-x_{2,k+1})$. Then after few manipulations and using (1.36) one obtains

$$\begin{array}{c} a(h)x_{2,k+1} + \dfrac{h}{LC}x_{1,k} - x_{2,k} \in \partial\psi_{\mathbb{R}^+}(-x_{2,k}) \\ \Updownarrow \\ x_{2,k} = \mathrm{prox}\left[\mathbb{R}^+; -a(h)x_{2,k+1} - \dfrac{h}{LC}x_{1,k}\right] \end{array} \tag{1.43}$$

which is absurd.

The implicit way of discretizing the inclusion is thus the only way that leads to a sound algorithm. This will still be the case with more general inclusions with right-hand sides of the form $\partial\psi_K(x)$ for some domain $K \subset \mathbb{R}^n$.

Let us now start from the complementarity formalism (1.28), with an explicit form

$$0 \leqslant \lambda_{k+1} \perp -x_{2,k} \geqslant 0\,. \tag{1.44}$$

Then we get the complementarity problem

$$0 \leqslant \lambda_{k+1} \perp -\left(1+\frac{h}{R}L\right)x_{2,k+1} - \frac{h}{LC}x_{1,k+1} - \frac{h}{L}\lambda_{k+1} \geqslant 0\,. \tag{1.45}$$

Clearly this complementarity problem cannot be used to advance the algorithm from step k to step $k+1$. This intrinsic implicit form of the discretization of the Differential Inclusion (DI) we work with here is not present in other types of inclusions, where explicit discretizations are possible, see Chap. 9.

1.1.7 Calculation of the Equilibrium Points

It is expected that studying the equilibrium points of complementarity systems as in (1.13) and (1.11) will lead either to a Complementarity Problem (CP) (like LCPs), or inclusions (see (1.7)), or variational inequalities (see (1.8)). Let us point out briefly the usefulness of the tools that have been introduced above, for the characterization of the equilibria of the class of nonsmooth systems we are dealing with.

In general one cannot expect that even simple complementarity systems possess a unique equilibrium. Consider for instance circuit (**c**) in (1.13). It is not difficult to see that the set of equilibria is given by $\{(x_1^*, x_2^*)|\, x_1^* \leqslant 0, x_2^* = 0\}$.

Let us consider now (1.26) and its equivalent (1.27). The fixed points z^* of the EVI in (1.27) have to satisfy

$$\langle -RAR^{-1}z^*, v - z^*\rangle \geqslant 0, \forall\, v\, \in K\,. \tag{1.46}$$

This is a variational inequality, and the studies concerning existence and uniqueness of solutions of a Variational Inequality (VI)are numerous. We may for instance use results in Yao (1994) which relate the set of solutions of (1.46) to the monotonicity of the operator $x \mapsto -RAR^{-1}x$. In this case, monotonicity is equivalent to semi-positive definiteness of $-RAR^{-1}$ and strong monotonicity is equivalent to positive definiteness of $-RAR^{-1}$ (Facchinei & Pang, 2003, p. 155). If the matrix $-RAR^{-1}$ is semi-positive definite, then Yao (1994, theorem 3.3) guarantees that the set of equilibria is nonempty, compact, and convex. If $-RAR^{-1}$ is positive definite, then from Yao (1994, theorem 3.5) there is a unique solution to (1.46), consequently a unique equilibrium for the system (1.26).

The monotonicity is of course a sufficient condition only. In order to see this, let us consider a linear complementarity system

$$\begin{cases} \dot{x}(t) = Ax(t) + B\lambda(t) \\ 0 \leqslant Cx(t) + D \perp \lambda(t) \geqslant 0 \end{cases}. \tag{1.47}$$

The fixed points of this LCS are the solutions of the problem

$$\begin{cases} 0 = Ax^* + B\lambda \\ 0 \leqslant Cx^* + D \perp \lambda \geqslant 0 \end{cases}. \tag{1.48}$$

If we assume that A is invertible, then we can construct the following LCP

$$0 \leqslant -CA^{-1}B\lambda + D \perp \lambda \geqslant 0 \tag{1.49}$$

which is not to be confused with the LCP in (1.24). If the matrix $-CA^{-1}B$ is a P-matrix then this LCP has a unique solution λ^* and we conclude that there is a unique equilibrium state $x^* = -A^{-1}B\lambda^*$. Clearly there is no monotonicity argument in this reasoning as the set of P-matrices contains that of positive definite matrices (i.e., a P-matrix is not necessarily positive definite).

As an illustration we may consider once again the circuits and (**c**) and (**d**). In the case of (1.13) we have

$$A = \begin{pmatrix} 0 & 1 \\ -\frac{1}{LC} & -\frac{R}{L} \end{pmatrix}, -CA^{-1}B = 0, \text{ and } D = 0\,. \tag{1.50}$$

There is an infinity of solutions for the LCP in (1.49), as pointed out above. In the case of (1.14) we have

$$A = \begin{pmatrix} -\frac{1}{RC} & 1 \\ -\frac{1}{LC} & 0 \end{pmatrix}, -CA^{-1}B = \frac{1}{R} > 0, \text{ and } D = 0\,. \tag{1.51}$$

There is a unique solution. We leave it to the reader to calculate explicitly the solutions (or the set of solutions). It is easily checked that no one of the two matrices $-A$ is semi-positive definite and they therefore do not define monotone operators. The sufficient criterion alluded to above is therefore not applicable.

In the case of circuit (**a**) with dynamics in (1.11), the fixed points are given as the solutions of a complementarity problem of the form

$$\begin{cases} 0 = Ax^* + B\lambda \\ 0 \leqslant Cx^* + D\lambda \perp \lambda \geqslant 0 \end{cases}, \tag{1.52}$$

where

$$A = \begin{pmatrix} -\frac{1}{RC} & 1 \\ -\frac{1}{LC} & 0 \end{pmatrix}, B = -\begin{pmatrix} \frac{1}{R} \\ \frac{1}{L} \end{pmatrix}, C = \begin{pmatrix} \frac{1}{RC} & -1 \end{pmatrix}, \text{ and } D = \frac{1}{R} \tag{1.53}$$

Since A is invertible, with inverse

$$A^{-1} = \begin{pmatrix} 0 & -1 \\ \frac{1}{LC} & -\frac{1}{RC} \end{pmatrix}$$

one can express x^* as $x^* = -A^{-1}B\lambda$. Therefore $Cx^* + D\lambda = (D - CA^{-1}B)\lambda$, and the LCP is: $0 \leqslant \lambda \perp (D - CA^{-1}B)\lambda \geqslant 0$. The solution is $\lambda = 0$ independently of the sign of the scalar $D - CA^{-1}B$. This can also be seen from the inclusion $\lambda \in -\partial\psi_{\mathbb{R}^+}((D - CA^{-1}B)\lambda)$, taking into account (1.6).

It is noteworthy that computing the fixed points of our circuits may be done by solving LCPs. In dimension 1 or 2, this may be done by checking the two or four possible cases, respectively. In higher dimensions, such enumerative procedures become impossible, and specific algorithms for solving LCPs (or other kinds of CPs) have to be used. Such algorithms will be described later in Part III.

1.2 Electrical Circuits with Ideal Zener Diodes

1.2.1 The Zener Diode

Let us consider, now, a further electrical device: the ideal Zener diode whose schematic symbol is depicted in Fig. 1.8a. A Zener diode is a type of diode that permits current to flow in the forward direction like a normal diode, but also in the reverse direction if the voltage is larger than the rated breakdown voltage known as "Zener knee voltage" or "Zener voltage" denoted by $V_z > 0$. The ideal characteristic between the current $i(t)$ and the voltage $v(t)$ can be seen in Fig. 1.8b.

Let us seek an analytical representation of the current–voltage characteristic of the ideal Zener diode. For this we are going to use some convex analysis tools and make some manipulations: subdifferentiate, conjugate, invert. Let us see how this works, with Fig. 1.9 as a guide.

The inversion consists of expressing $v(t)$ as a function of $-i(t)$: this is done in Fig. 1.9b. Computing the subderivative of the function $f(\cdot)$ of Fig. 1.9c, one gets the multivalued mapping of Fig. 1.9b. Indeed we have

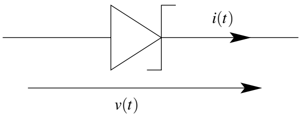


Fig. 1.8a. The Zener diode schematic symbol

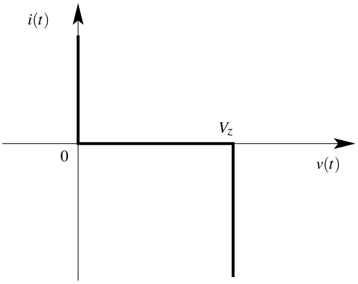


Fig. 1.8b. The ideal characteristic of a Zener diode

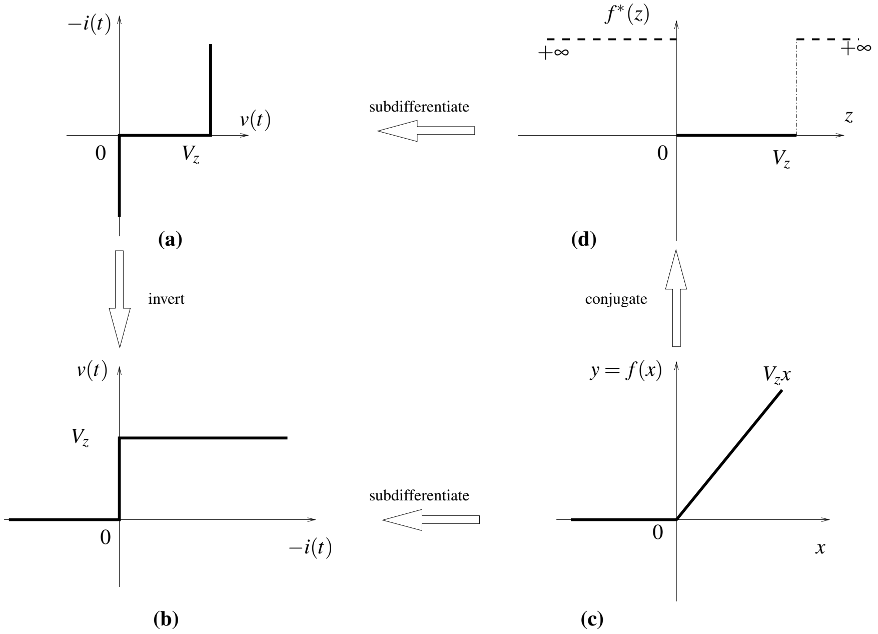


Fig. 1.9. The Zener diode characteristic

$$f(x) = \begin{cases} V_z x \text{ if } x \geqslant 0 \\ \\ 0 \quad \text{if } x < 0 \end{cases} \tag{1.54}$$

from which it follows that the subdifferential of $f(\cdot)$ is

$$\partial f(x) = \begin{cases} V_z & \text{if } x > 0 \\ \\ [0, V_z] & \text{if } x = 0 \\ \\ 0 & \text{if } x < 0 \end{cases} . \tag{1.55}$$

Notice that the function $f(\cdot)$ is convex, proper, continuous, and that the graphs of the multivalued mappings of Fig. 1.9a and b are maximal monotone. *Monotonicity* means that if you pick any two points $-i_1$ and $-i_2$ on the abscissa of Fig. 1.9b, and the corresponding v_1 and v_2, then it is always true that

$$\langle -i_1 - (-i_2), v_1 - v_2 \rangle \geqslant 0 \tag{1.56}$$

Similarly for Fig. 1.9a *maximality* means that it is not possible to add any new branch to the graphs of these mappings, without destroying the monotonicity. This is indeed the case for the graphs of Fig. 1.9a and b.

Let us now introduce the notion of the conjugate of a convex function $f(\cdot)$ that is defined as

$$f^*(z) = \sup_{x\in\mathbb{R}}(\langle x,z\rangle - f(x)) \,. \tag{1.57}$$

Let us calculate the conjugate of the function $f(\cdot)$ above:

$$f^*(z) = \sup\nolimits_{x\in\mathbb{R}} \begin{cases} xz - V_z x \text{ if } x \geqslant 0 \\ \\ xz \qquad\quad \text{if } x < 0 \end{cases} = \sup\nolimits_{x\in\mathbb{R}} \begin{cases} x(z - V_z) \text{ if } x \geqslant 0 \\ \\ xz \qquad\quad \text{if } x < 0 \end{cases}$$

$$\begin{aligned} &= \begin{cases} \begin{cases} +\infty \text{ if } z > V_z \\ 0 \quad \text{if } z \leqslant V_z \end{cases} \\ \\ \begin{cases} 0 \quad \text{if } z \geqslant 0 \\ +\infty \text{ if } z < 0 \end{cases} \end{cases} = \begin{cases} +\infty \text{ if } z < 0 \text{ and } z > V_z \\ \\ 0 \quad \text{if } 0 \leqslant z \leqslant V_z \end{cases} \\ \\ &= \psi_{[0,V_z]}(z) \,, \end{aligned} \tag{1.58}$$

where we retrieve the indicator function that was already met when we considered the ideal diode, see Sect. 1.1.1.

We therefore deduce from Fig. 1.9 that

$$-i(t) \in \partial\psi_{[0,V_z]}(v(t)), \text{ whereas } v(t) \in \partial f(-i(t)) \,. \tag{1.59}$$

The function $f(\cdot) = \psi^*_{[0,V_z]}(\cdot)$ is called in convex analysis the *support* function of the set $[0,V_z]$. It is known that the support function and the indicator function of a convex set are conjugate to one another.

We saw earlier that the subderivative of the indicator function of a convex set is also the normal cone to this convex set. Here we obtain that $\partial\psi_{[0,V_z]}(v(t))$ is the normal cone $N_{[0,V_z]}(v(t))$, that is $\mathbb{R}^-$ when $v(t) = 0$ and $\mathbb{R}^+$ when $v(t) = V_z$. It is the singleton $\{0\}$ when $0 < v(t) < V_z$.

1.2.2 The Dynamics of a Simple Circuit

Differential Inclusions and Filippov's Systems

Now that these calculations have been led, let us consider the dynamics of the circuit in Fig. 1.3c, where we replace the ideal diode by an ideal Zener diode. Choosing the same state variables (x_1 is the capacitor charge, x_2 is the current through the circuit), we obtain:

$$\begin{cases} \dot{x}_1(t) = x_2(t) \\ \\ \dot{x}_2(t) + \dfrac{R}{L}x_2(t) + \dfrac{1}{LC}x_1(t) = \dfrac{1}{L}v(t) \end{cases} , \tag{1.60}$$

where $v(\cdot)$ is the voltage of the Zener diode. We saw that $v(t) \in \partial f(-i(t))$, thus we get

$$\begin{cases} \dot{x}_1(t) - x_2(t) = 0 \\ \\ \dot{x}_2(t) + \frac{R}{L}x_2(t) + \frac{1}{LC}x_1(t) \in \frac{1}{L}\partial f(-x_2(t)) \end{cases}, \tag{1.61}$$

which is a differential inclusion.

Compare the inclusions in (1.25) and in (1.61). They look quite similar, however, the sets in their right-hand sides are quite different. Indeed the set in the right-hand side of (1.25) is unbounded, whereas the set in the right-hand side of (1.61) is bounded, as it is included in $[0,V_z]$. More precisely, the set-valued mapping $\partial f(\cdot)$ is nonempty, compact, convex, upper semi-continuous, and satisfies a linear growth condition: for all $v \in \partial f(x)$ there exists constants k and a such that $\| v \| \leqslant k \| x \| + a$.

The differential inclusion (1.61) possesses an absolutely continuous solution, and we may even assert here that this solution is unique for each initial condition, because in addition the considered set-valued mapping is maximal monotone, see Lemma 2.13, Theorem 2.41. This is also sometimes called a Filippov's system or a Filippov's DI, associated with the switching surface $\Sigma = \{x \in I\!R^2 \mid x_2 = 0\}$. See Sect. 2.1 for a precise definition of Filippov's systems. Simple calculations yield that the vector field in the neighborhood of Σ is as depicted in Fig. 1.10. The surface Σ is crossed transversally by the trajectories when $x_1(t) < 0$ and $x_1(t) > CV_z$. It is an attracting surface when $x_1(t) \in [0,V_z]$ (where t means the time when the trajectory attains Σ). According to Filippov's definition of the solution, Σ is a sliding surface in the latter case, which means that $x_2(t) = 0$ after the trajectory has reached this portion of Σ. Notice that we may rewrite the second line in (1.61) as

$$\dot{x}_2(t) + \frac{R}{L}x_2(t) + \frac{1}{LC}x_1(t) = \lambda(t), \;\; \lambda(t) \in \frac{1}{L}\partial f(-x_2(t)) \, . \tag{1.62}$$

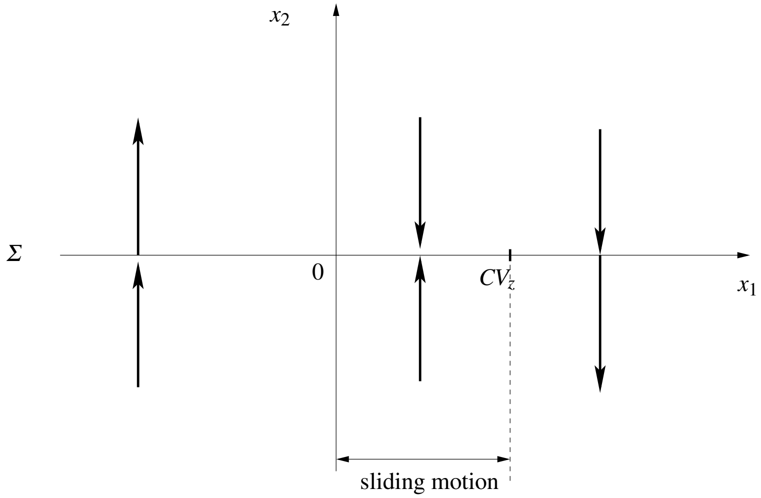


Fig. 1.10. The vector field on the switching surface Σ

Despite passing from (1.61) to (1.62) looks like wasted effort, it means that the inclusion in (1.61) is equivalent to integrate its left-hand side by looking for an element of the set in its right-hand side, at each time instant. This is in fact the case for *all* the differential inclusions that we shall deal with in this book. In other words the integration proceeds along Σ with an element $\lambda \in \partial f(0)$ such that $\lambda(t) = \frac{x_1(t)}{C}$, where t is the "entry" time of the trajectory in Σ (notice that as long as $x_2 = 0$ then x_1 remains constant).

Remark 1.6. The fact that the switching surface Σ is attracting in $x_1(t) \in [0, V_z]$, is intimately linked with the maximal monotonicity of the set-valued mapping $\partial f(\cdot)$. This mapping is sometimes called a *relay* function in the systems and control community (Fig. 1.11).

A First Complementarity System Formulation

Let us now seek a complementarity formulation of the multivalued mapping $\partial f(\cdot) = \partial \psi^*_{[0,V_z]}(\cdot)$ whose graph is in Fig. 1.9a. Let us introduce two slack variables (or multipliers) λ_1 and λ_2, and the set of conditions:

$$\begin{cases} 0 \leqslant \lambda_1(t) \perp -i(t) + |i(t)| \geqslant 0 \\ 0 \leqslant \lambda_2(t) \perp i(t) + |i(t)| \geqslant 0 \\ \lambda_1(t) + \lambda_2(t) = V_z \\ v(t) = \lambda_2(t) \end{cases} . \tag{1.63}$$

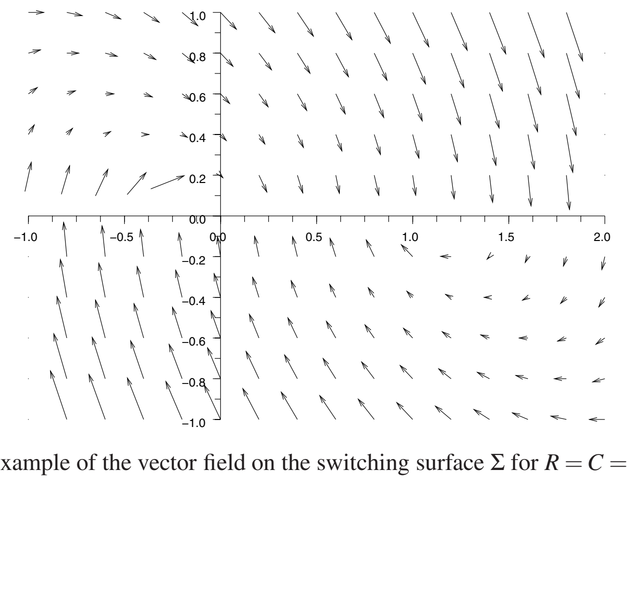


Fig. 1.11. Example of the vector field on the switching surface Σ for $R = C = L = V_z = 1$

Let us check by inspection that indeed (1.63) represents the mapping of Fig. 1.9a. If $-i(t) > 0$, then $-i(t) + |i(t)| > 0$, so $\lambda_1(t) = 0$ and $\lambda_2(t) = V_z = v(t)$ (and $i(t) + |i(t)| = 0$). If $-i(t) < 0$ then $i(t) + |i(t)| > 0$, so $\lambda_2(t) = 0$, and $\lambda_1(t) = V_z$ (and $-i(t) + |i(t)| = 0$) and $v(t) = \lambda_2(t) = 0$. Now if $i(t) = 0$, then one easily calculates that $0 \leqslant \lambda_1(t) \leqslant V_z$, $0 \leqslant \lambda_2(t) \leqslant V_z$. Thus $0 \leqslant v(t) \leqslant V_z$.

Thanks to the complementary formulation (1.63), the inclusion (1.61) can be formulated as a Dynamical (or Differential) Complementarity System (DCS)

$$\begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) + \dfrac{R}{L} x_2(t) + \dfrac{1}{LC} x_1(t) = \dfrac{1}{L} v(t) \\ 0 \leqslant \lambda_1(t) \perp -x_2(t) + |x_2(t)| \geqslant 0 \\ 0 \leqslant \lambda_2(t) \perp \quad x_2(t) + |x_2(t)| \geqslant 0 \\ \lambda_1(t) + \lambda_2(t) = V_z \\ v(t) = \lambda_2(t) \end{cases} . \tag{1.64}$$

This DCS is not an LCS due to the presence of the absolute value function in the complementarity condition and the two last algebraic equations. We notice that the variables $\lambda_1(t)$ and $\lambda_2(t)$ can be eliminated from (1.64) using the last two equalities, leading to another formulation of the DCS:

$$\begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) + \dfrac{R}{L} x_2(t) + \dfrac{1}{LC} x_1(t) = \dfrac{1}{L} v(t) \\ 0 \leqslant V_z - v(t) \perp -x_2(t) + |x_2(t)| \geqslant 0 \\ 0 \leqslant v(t) \perp x_2(t) + |x_2(t)| \geqslant 0 \end{cases} \tag{1.65}$$

which is neither an LCS.

A Mixed Linear Complementarity Formulation

It is possible from (1.63) to obtain a so-called Mixed Linear Complementarity System (MLCS) which is a generalization of an LCS with an additional system of linear equations. The goal is to obtain after discretization a so-called Mixed Linear Complementarity Problem (MLCP) which is a generalization of an LCP with an additional system of linear equations, such that

$$\begin{cases} Au + Cw + a = 0 \\ 0 \leqslant w \perp Du + Bw + d \geqslant 0 \end{cases} . \tag{1.66}$$

To obtain an MLCS formulation, let us introduce the positive part and the negative part of the current $i(t)$ as

$$i^+(t) = \frac{1}{2}(i(t) + |i(t)|) = \max(0, i(t)) \geqslant 0\,, \tag{1.67}$$

$$i^-(t) = \frac{1}{2}(i(t) - |i(t)|) = \min(0, i(t)) \leqslant 0\,. \tag{1.68}$$

The system (1.63) can be rewritten equivalently as

$$\begin{cases} 0 \leqslant \lambda_1(t) \perp i^+(t) - i(t) \geqslant 0 \\ 0 \leqslant \lambda_2(t) \perp i^+(t) \geqslant 0 \\ i(t) = i^-(t) + i^+(t) \\ \lambda_1(t) + \lambda_2(t) = V_z \\ v(t) = \lambda_2(t) \end{cases}, \tag{1.69}$$

where the absolute value has disappeared, but a linear equation has been added. Substitution of two of the last three equations into the complementarity conditions leads to an intermediate complementarity formulation of the relay function as

$$\begin{cases} 0 \leqslant \lambda_1(t) \perp i^+(t) - i(t) \geqslant 0 \\ 0 \leqslant v(t) \perp i^+(t) \geqslant 0 \\ \lambda_1(t) + v(t) = V_z \end{cases} \tag{1.70}$$

or as

$$\begin{cases} 0 \leqslant V_z - v(t) \perp i^+(t) - i(t) \geqslant 0 \\ 0 \leqslant v(t) \perp i^+(t) \geqslant 0 \end{cases}. \tag{1.71}$$

The linear dynamical system (1.60) together with one of the reformulations (1.69), (1.70), or (1.71) leads to an MLCS formulation. Nevertheless, the complete substitution of the equation into the complementarity condition yields a DCS

$$\begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) + \frac{R}{L}x_2(t) + \frac{1}{LC}x_1(t) = \frac{1}{L}v(t) \\ 0 \leqslant V_z - v(t) \perp x_2^+(t) - x_2(t) \geqslant 0 \\ 0 \leqslant v(t) \perp x_2^+(t) \geqslant 0 \end{cases}, \tag{1.72}$$

which is neither an LCS nor an MLCS.

A Linear Complementarity Formulation

Due to the simplicity of the equations involved in the MLCS formulation (1.71), it is possible to find an LCS formulation of the dynamics. Indeed, the following system

$$\begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) + \dfrac{R}{L} x_2(t) + \dfrac{1}{LC} x_1(t) = \dfrac{1}{L}\lambda_2(t) \\ x_2^+(t) = x_2(t) - x_2^-(t) \\ \lambda_1(t) = V_{\mathrm{z}} - \lambda_2(t) \\ 0 \leqslant \begin{pmatrix} x_2^+(t) \\ \lambda_1(t) \end{pmatrix} \perp \begin{pmatrix} \lambda_2(t) \\ -x_2^-(t) \end{pmatrix} \geqslant 0 \end{cases} \tag{1.73}$$

can be recast into the following LCS form

$$\begin{cases} \dot{x}(t) = Ax(t) + B\tilde{\lambda}(t) \\ w(t) = Cx(t) + D\tilde{\lambda}(t) + g \\ 0 \leqslant w(t) \perp \tilde{\lambda}(t) \geqslant 0 \end{cases} \tag{1.74}$$

with

$$A = \begin{bmatrix} 0 & 1 \\ -\dfrac{1}{LC} & -\dfrac{R}{L} \end{bmatrix}, B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, C = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, D = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, g = \begin{bmatrix} 0 \\ V_{\mathrm{z}} \end{bmatrix}. \tag{1.75}$$

The reformulation appears to be a special case for more general reformulations of relay systems or two-dimensional friction problems into LCS. For more details, we refer to Pfeiffer & Glocker (1996) and to Sect. 9.3.3. In the more general framework of ODE with discontinuous right-hand side, an LCS reformulation can be found in Chap. 7.

1.2.3 Numerical Simulation by Means of Time-Stepping Schemes

In view of this preliminary material, we may consider now the time-discretization of our system. Clearly our objective here is still to introduce the topic, and the reader should not expect an exhaustive description of the numerical simulation of the system.

1.2.3.1 Explicit Time-Stepping Schemes Based on ODE with Discontinuities Formulations

A forward Euler scheme may be applied on an ODE with discontinuities of the form, $\dot{x} = f(x,t)$, where the right-hand side may possess discontinuities (see Sect. 2.8). For the right-hand side of the circuit with the Zener diode, a switched model may be given by

$$f(x,t) = \begin{cases} \begin{bmatrix} x_2 \\ -\frac{R}{L}x_2 - \frac{1}{LC}x_1 \end{bmatrix} & \text{for } -x_2 < 0 \qquad (1.76a) \\ \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \text{for } -x_2 = 0 \qquad (1.76b) \\ \begin{bmatrix} x_2 \\ -\frac{R}{L}x_2 - \frac{1}{LC}x_1 - V_z \end{bmatrix} & \text{for } -x_2 > 0. \qquad (1.76c) \end{cases}$$

The simulation for this choice of the right-hand side is illustrated in Fig. 1.12. We can observe that some "chattering" effects due to the fact that the sliding mode given by (1.76b) cannot be reached due to the numerical approximation on x_2. This artifact results in spurious oscillations of the diode voltage $v(t) = \lambda(t)$ and the diode current $x_2(t) = \omega(t)$ as we can observe on the zoom in Fig. 1.13.

One way to circumvent the spurious oscillations is to introduce a "sliding band", i.e., an interval where the variable x_2 is small in order to approximate the sliding mode. This interval can be for instance chosen as $|x_2| \leqslant \eta$ such that the new right-hand side is given by

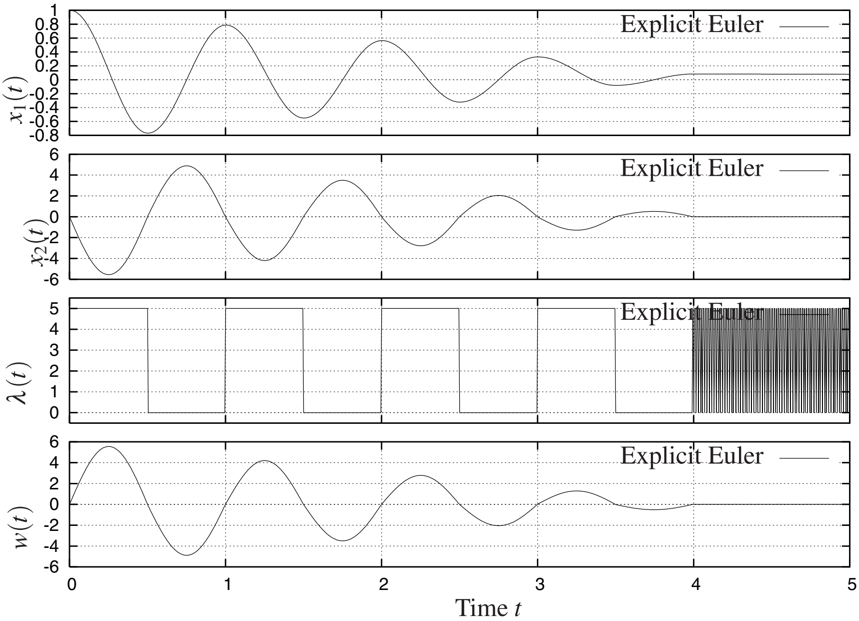


Fig. 1.12. Simulation of the RLC circuit with a Zener diode with the initial conditions $x_1(0) = 1, x_2(0) = 1$ and $R = 0.1, L = 1, C = \frac{1}{(2\pi)^2}$. Explicit Euler scheme with the right-hand side defined by (1.76). Time step $h = 5 \times 10^{-3}$

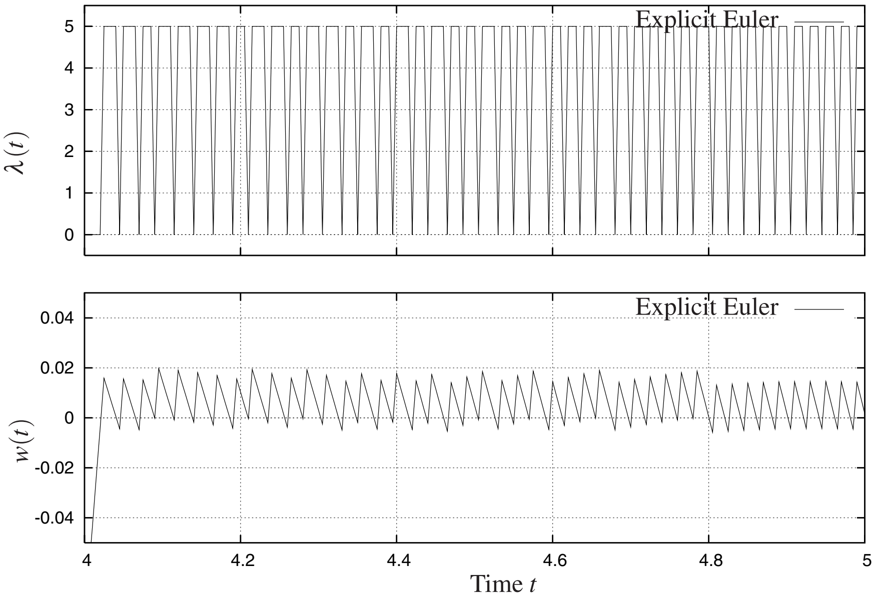


Fig. 1.13. Zoom on the "chattering" behavior simulation of the RLC circuit with a Zener diode with the initial conditions $x_1(0) = 1, x_2(0) = 1$ and $R = 0.1, L = 1, C = \frac{1}{(2\pi)^2}$. Explicit Euler scheme with the right-hand side defined by (1.76). Time step $h = 5 \times 10^{-3}$

$$
f(x,t) = \begin{cases}
\begin{bmatrix} x_2 \\ -\frac{R}{L}x_2 - \frac{1}{LC}x_1 \end{bmatrix} & \text{for } -x_2 < -\eta \qquad (1.77a) \\
\begin{bmatrix} x_2 \\ -\frac{R}{L}x_2 \end{bmatrix} & \text{for } |x_2| \leqslant \eta \qquad (1.77b) \\
\begin{bmatrix} x_2 \\ -\frac{R}{L}x_2 - \frac{1}{LC}x_1 - V_{\mathrm{z}} \end{bmatrix} & \text{for } -x_2 > \eta \qquad (1.77c)
\end{cases}
$$

Simulation results depicted in the Figs. 1.14 and 1.15 show that the spurious oscillations have been cancelled.

The switched models (1.76) and (1.77) are incomplete models. In more general situations they may fail due the lack of conditions for the transition from the sliding mode to the other modes. Clearly, the value of the dual variable $\lambda(t) = v(t)$ has to be checked to know if the system stays in the sliding mode. We will see in Sect. 9.3.3 that all these conditional statements can be in numerous cases replaced by an LCP formulation.

It is noteworthy that the previous numerical trick is not an universal solution for the problem of chattering. Indeed, the switched model given by the right-hand side (1.77) allows the solution to stay near the boundary of the sliding band. The new

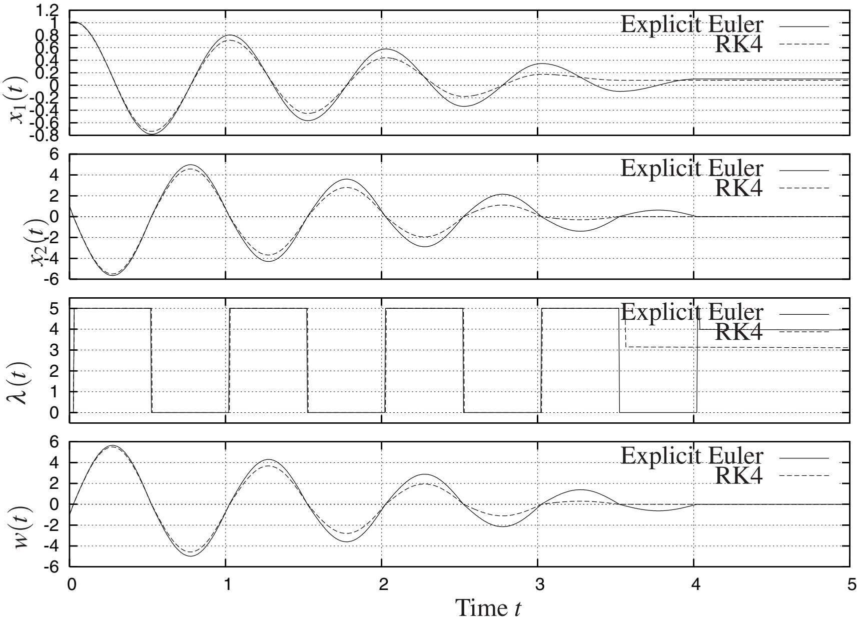


Fig. 1.14. Simulation of the RLC circuit with a Zener diode with the initial conditions $x_1(0) = 1, x_2(0) = 1$ and $R = 0.1, L = 1, C = \frac{1}{(2\pi)^2}$. Euler and four order Runge–Kutta explicit scheme with the right-hand side defined by (1.77). Time step $h = 5 \times 10^{-3}$

model is still a discontinuous system and therefore some numerical instabilities of the ODE solver can appear. More smart approaches for the choice of the right-hand side in the sliding band can be found in Karnopp (1985), Leine et al. (1998), Leine & Nijmeijer (2004) and will be described in Sect. 9.3.2.

The fact that we are able to express the Filippov's DI as an equivalent model of ODE with a switched right-hand side allows one to use any other explicit schemes such as explicit Runge–Kutta methods. In Figs. 1.15 and 1.16, the results of the simulation with the right-hand side (1.76) and (1.77) are depicted. The conclusions are the same as above. One notices also that two different methods provide different results (see Figs. 1.14 and 1.15). We will discuss in Sect. 9.2 the question of the order and the stability of such a higher order method for Filippov's DIs.

1.2.3.2 Explicit Discretization of the Differential Inclusion and the Complementarity Systems

Explicit Discretization of the Differential Inclusion (1.61)

Consider the forward Euler method

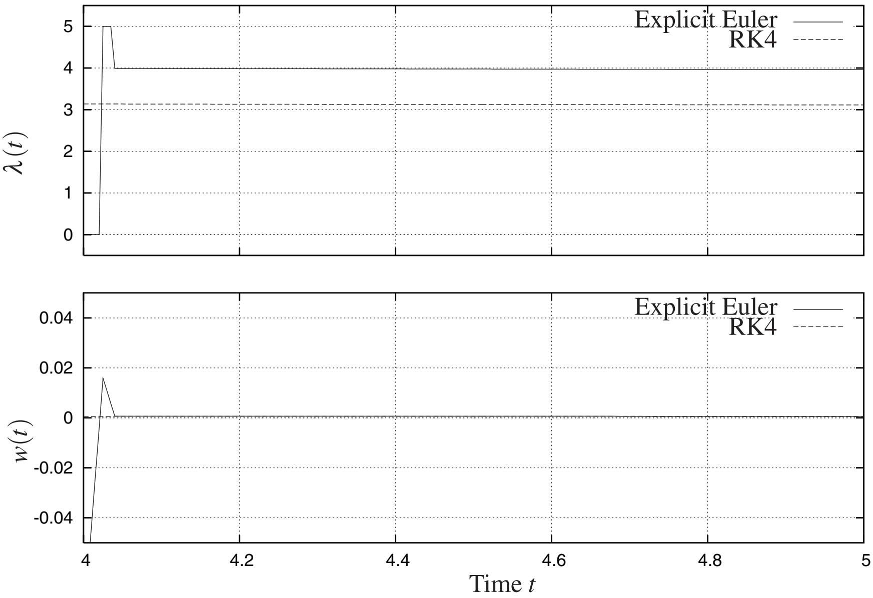


Fig. 1.15. Simulation of the RLC circuit with a Zener diode with the initial conditions $x_1(0) = 1, x_2(0) = 1$ and $R = 0.1, L = 1, C = \frac{1}{(2\pi)^2}$. Euler and four order Runge–Kutta explicit scheme with the right-hand side defined by (1.77). Time step $h = 5 \times 10^{-3}$

$$\begin{cases} x_{1,k+1} - x_{1,k} = hx_{2,k} \\ x_{2,k+1} - x_{2,k} + \dfrac{hR}{L}x_{2,k} + \dfrac{h}{LC}x_{1,k} \in \dfrac{h}{L}\partial f(-x_{2,k}) \,, \end{cases} \tag{1.78}$$

where x_k is the value, at time t_k of a grid $t_0 < t_1 < \cdots < t_N = T$, $N < +\infty$, $h = \dfrac{T - t_0}{N} = t_k - t_{k-1}$, of a step function $x^N(\cdot)$ that approximates the analytical solution $x(\cdot)$.

Compare with the time-discretization of the inclusion (1.25) that is proposed in Sect. 1.1.5. This time considering an implicit scheme is not mandatory (this may improve the overall quality of the numerical integration especially from the stability point of view, but is not a consequence of the dynamics contrary to what happens with (1.25)). One of the major discrepancies with the circuit (1.25) is that the values of x_2 are no longer constrained to stay in a set by the inclusion (1.78).

Explicit Discretization of the Complementarity Systems (1.65)

Let us investigate how the complementarity system (1.65) may be discretized. We get

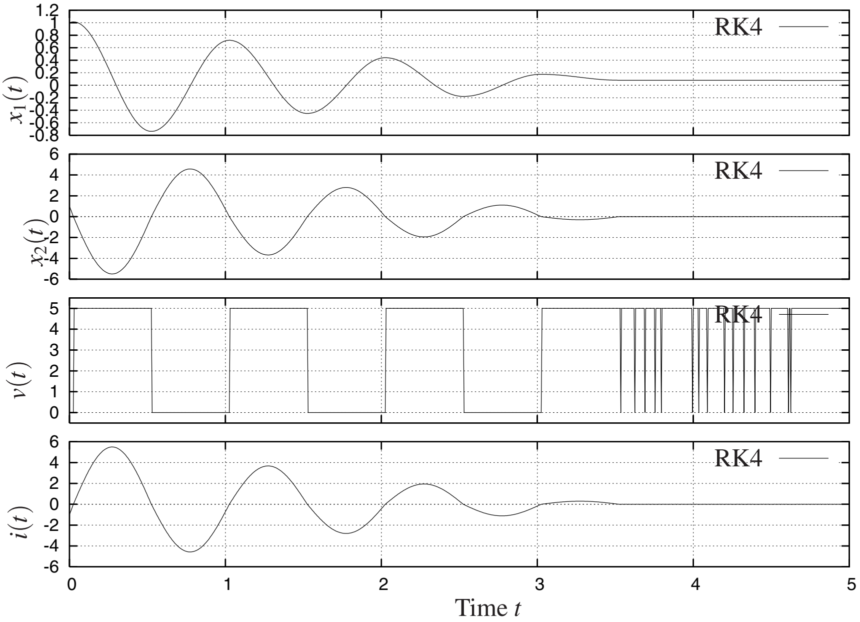


Fig. 1.16. Simulation of the RLC circuit with a Zener diode with the initial conditions $x_1(0) = 1, x_2(0) = 1$ and $R = 0.1, L = 1, C = \frac{1}{(2\pi)^2}$. Four order Runge–Kutta explicit scheme with the right-hand side defined by (1.76). Time step $h = 5 \times 10^{-3}$

$$\begin{cases} x_{1,k+1} - x_{1,k} = hx_{2,k} \\ x_{2,k+1} - x_{2,k} + \dfrac{hR}{L}x_{2,k} + \dfrac{h}{LC}x_{1,k} = \dfrac{h}{L}\lambda_{2,k} \\ 0 \leqslant V_\mathrm{z} - \lambda_{2,k} \perp -x_{2,k} + |x_{2,k}| \geqslant 0 \\ 0 \leqslant \lambda_{2,k} \perp \quad x_{2,k} + |x_{2,k}| \geqslant 0 \end{cases} . \tag{1.79}$$

One computes that if $x_{2,k} > 0$ then $\lambda_{2,k} = 0$, while $x_{2,k} < 0$ implies $\lambda_{2,k} = V_\mathrm{z}$. Moreover $x_{2,k} = 0$ implies that $\lambda_{2,k} \in [0, V_\mathrm{z}]$. We conclude that the two schemes in (1.78) and (1.79) are the same.

However, the complementarity formalism does not bring any advantage over the inclusion formalism, as it does not yield neither an LCP nor an MLCP, even with the reformulation proposed in the preceding section. The main reason for that is not the presence of absolute values in the complementarity formalism which can be avoided by adding an equality, but the fact that $\lambda_{2,k}$ has to be complementary to the positive part of $x_{2,k}$ which is not an unknown at the beginning of the step.

For instance, if we choose the MLCS formulation given by the dynamical system (1.60) and the formulation (1.70), we get the following complementarity problem

$$\begin{cases} x_{1,k+1} - x_{1,k} = hx_{2,k} \\ x_{2,k+1} - x_{2,k} + \frac{hR}{L}x_{2,k} + \frac{h}{LC}x_{1,k} = \frac{h}{L}\lambda_{2,k} \\ 0 \leqslant \lambda_{1,k} \perp x_{2,k}^{+} - x_{2,k} \geqslant 0 \\ 0 \leqslant \lambda_{2,k} \perp x_{2,k}^{+} \geqslant 0 \\ \lambda_{1,k} + \lambda_{2,k} = V_{\mathrm{z}} \end{cases} . \tag{1.80}$$

In such a "fake" complementarity problem, one has to perform the procedure described in the Remark 1.7, which implies to choose a threshold on the value of $x_{2,k}$.

To conclude this part, whatever the mathematical formalism which is used to formulate the dynamics, explicit discretizations lead to algorithms without any sense.

Remark 1.7. One has to choose a value for $\lambda_{2,k}$ in the interval $[0, V_{\mathrm{z}}]$ when $x_{2,k} = 0$. More concretely when implementing the algorithm on a computer, one has to choose a threshold $\eta > 0$ such that $x_{2,k}$ is considered to be null when $|x_{2,k}| \leqslant \eta$. One possibility is to choose the Filippov's solution that makes the trajectory slide on the surface $\Sigma = \{x \in \mathrm{I\!R}^2 \mid x_2 = 0\}$. If $x_{1,k} \notin [0, CV_{\mathrm{z}}]$ we have seen that the trajectories cross transversally Σ. Thus the chosen value of $\lambda_{2,k}$ is not important. If $x_{1,k} \in [0, CV_z]$ one may simply choose $\lambda_{2,k} = \frac{x_{1,k}}{C}$ or $\lambda_{2,k} = -\frac{L}{h}x_{2,k} + Rx_{2,k} + \frac{x_{1,k}}{C}$ to keep $x_{2,k+1}$ in the required neighborhood of Σ. With the solution, we have also to check the value of the dual variable $v(t) = \lambda_2(t)$ to know when the application of this rule has to be stopped.

1.2.3.3 An Implicit Time-Stepping Scheme

Implicit Discretization of the Differential Inclusion (1.61)

Let us try the following implicit scheme[4]:

$$\begin{cases} x_{1,k+1} - x_{1,k} = hx_{2,k+1} \\ x_{2,k+1} - x_{2,k} + \frac{hR}{L}x_{2,k+1} + \frac{h}{LC}x_{1,k+1} \in \frac{h}{L}\partial f(-x_{2,k+1}) . \end{cases} \tag{1.81}$$

After some manipulations this may be rewritten as

$$\begin{cases} x_{1,k+1} - x_{1,k} = hx_{2,k+1} \\ x_{2,k+1} + a(h)\left[\frac{h}{LC}x_{1,k} - x_{2,k}\right] \in a(h)\frac{h}{L}\partial f(-x_{2,k+1}) , \end{cases} \tag{1.82}$$

[4] The scheme chosen here is fully implicit for the sake of simplicity.

where $a(h) = \left(1 + \frac{hR}{L} + \frac{h^2}{LC}\right)^{-1}$.

Denoting

$$b = a(h)\left[\frac{h}{LC}x_{1,k} - x_{2,k}\right]$$

the second line of (1.82) may be rewritten as

$$x_{2,k+1} + b \in a(h)\frac{h}{L}\partial f(-x_{2,k+1}) \ . \tag{1.83}$$

It is this inclusion that we are going to examine now. This will allow us to illustrate graphically why the monotonicity is a crucial property. In Fig. 1.17 the graph of the linear function

$$\mathscr{D}_b = \left\{(\lambda_{2,k+1}, x_{2,k+1}) \in I\!R^2 \mid \quad \lambda_{2,k+1} = x_{2,k+1} + b\right\}$$

is depicted for three values of b, together with the graph of the set-valued function,

$$\mathscr{G} = \left\{(\lambda_{2,k+1}, x_{2,k+1}) \in I\!R^2 \mid \quad \lambda_{2,k+1} \in a(h)\frac{h}{L}\partial f(-x_{2,k+1})\right\} \ .$$

It is apparent that for any value of b, there is always a single intersection between the two graphs. One concludes that the generalized equation (1.83) with unknown $x_{2,k+1}$ has a unique solution, which allows one to advance the algorithm from k to $k+1$.

If there is an exogenous input $u(t)$ that acts on the system so that the dynamics is changed to

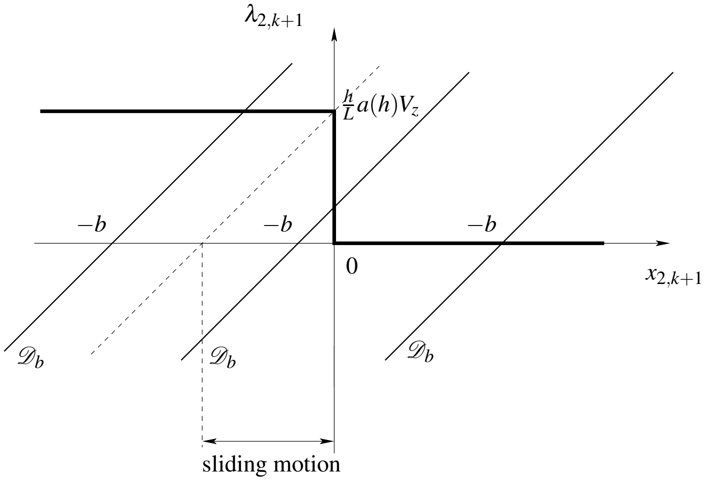


Fig. 1.17. Implicit scheme: the maximal monotone case

$$\dot{x}_2(t)+\frac{R}{L}x_2(t)+\frac{1}{LC}x_1(t)+\frac{u(t)}{L}\in\frac{1}{L}\partial f(-x_2(t)) \tag{1.84}$$

then the variable b is changed to $b+a(h)\frac{u_k}{L}$. Varying u_{k+1} corresponds to a horizontal translations of the straight lines in Fig. 1.17.

Remark 1.8 (A nonmonotone example). Suppose now that the dynamics is

$$x_{2,k+1}+b\in -a(h)\frac{h}{L}\partial f(-x_{2,k+1})\;. \tag{1.85}$$

We know this is not possible with the circuit we are studying. For the sake of the reasoning we are leading let us imagine this is the case. Then we get the situation depicted in Fig. 1.18. There exist values of b for which the generalized equation has two or three solutions. Uniqueness is lost.

Remark 1.9 (Comparison with the procedure in Remark 1.7). Coming back to Fig. 1.17, one sees that the values of b that yield a sliding motion along the surface Σ, correspond to all the values such that the graph of the linear function intersects the vertical segment of the graph of the multifunction. Contrarily to what happens with the explicit scheme where a threshold has to be introduced, “detecting” the sliding motion is now the result of a resolution of the intersection problem. No artificial threshold is needed due to the fact that we have to verify the inclusion of a value into a set of nonempty interior.

Implicit Discretization of the Complementarity Systems

Let us choose one of the LCS formulations described in the previous section given by the dynamics (1.73). An implicit time-discretization is given by

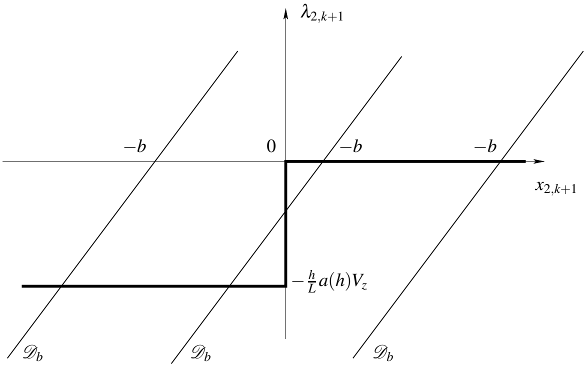


Fig. 1.18. Implicit scheme: the nonmonotone case

$$\begin{cases} x_{1,k+1} - x_{1,k} = h x_{2,k+1} \\ x_{2,k+1} - x_{2,k} + \dfrac{hR}{L} x_{2,k+1} + \dfrac{h}{LC} x_{1,k+1} = \dfrac{h}{L} \lambda_{2,k+1} \\ x^{+}_{2,k+1} = x_{2,k+1} - x^{-}_{2,k+1} \\ \lambda_{1,k+1} = V_{\mathrm{z}} - \lambda_{2,k+1} \\ 0 \leqslant \begin{pmatrix} x^{+}_{2,k+1} \\ \lambda_{1,k+1} \end{pmatrix} \perp \begin{pmatrix} \lambda_{2,k+1} \\ -x^{-}_{2,k+1} \end{pmatrix} \geqslant 0 \end{cases} . \tag{1.86}$$

Using the previous notations for $a(h)$ and b, we get the following system

$$\begin{cases} x_{1,k+1} - x_{1,k} = h x_{2,k+1} \\ x_{2,k+1} + b = a(h) \dfrac{h}{L} \lambda_{2,k+1} \\ x^{+}_{2,k+1} = x_{2,k+1} - x^{-}_{2,k+1} \\ \lambda_{1,k+1} = V_{\mathrm{z}} - \lambda_{2,k+1} \\ 0 \leqslant \begin{pmatrix} x^{+}_{2,k+1} \\ \lambda_{1,k+1} \end{pmatrix} \perp \begin{pmatrix} \lambda_{2,k+1} \\ -x^{-}_{2,k+1} \end{pmatrix} \geqslant 0 \end{cases} . \tag{1.87}$$

The value of $\lambda_{2,k+1}$ is obtained at each time step by the following LCP

$$\begin{cases} w = \begin{bmatrix} \dfrac{a(h)h}{L} & 1 \\ -1 & 0 \end{bmatrix} z + \begin{bmatrix} -b \\ V_{\mathrm{z}} \end{bmatrix} \\ 0 \leqslant w \perp z \geqslant 0 \end{cases} \tag{1.88}$$

with $w = [x^{+}_{2,k+1}\ , \lambda_{1,k+1}]^{\mathrm{T}}$ and $z = [\lambda_{2,k+1}\ , x^{-}_{2,k+1}]^{\mathrm{T}}$. We see in this case that the interest of the LCS formulation is to open the door to LCP solvers instead of having to check the modes.

Simulation Results

The simulation results are presented in Fig. 1.19. We can notice that the spurious oscillations in Figs. 1.12, 1.13 and 1.16 have disappeared due to the fact that the sliding is correctly modeled with the implicit approach.

1.2.3.4 Convergence Properties

Consider the explicit Euler scheme in (1.78). Then there exists a subsequence of the sequence $\{x^n(\cdot)\}_n$ that converges uniformly as $n \to +\infty$ to some (the) solution of the

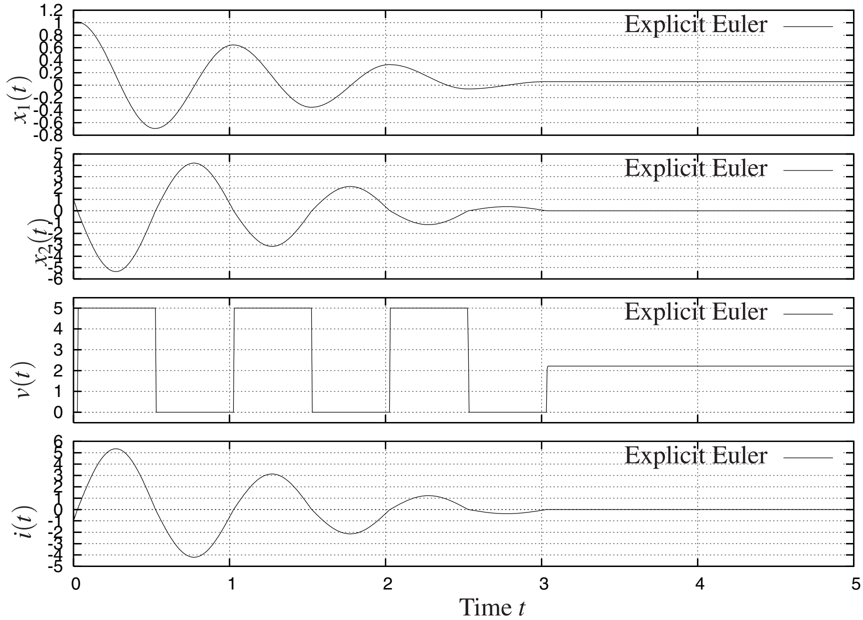


Fig. 1.19. Simulation of the RLC circuit with a Zener diode with the initial conditions $x_1(0) = 1, x_2(0) = 1$ and $R = 0.1, L = 1, C = \dfrac{1}{(2\pi)^2}$. Implicit Euler scheme. Time step $h = 5 \times 10^{-3}$

inclusion in (1.78). This is a consequence of Theorem 9.5. A similar result applies to the implicit scheme in (1.81), considered as a particular case of a linear multistep algorithm.

More details will be given in Chap. 9 on one-step and multistep time-stepping methods for differential inclusion with absolutely continuous solutions such as Filippov's DI. When uniqueness of solutions holds, more can be said on the convergence of the scheme, see Theorems 9.8, 9.9 and 9.11.

1.2.4 Numerical Simulation by Means of Event-Driven Schemes

The Filippov's DI (1.61) may also be simulated by means of *event-driven schemes*. We recall that the event-driven approach is based on a time integration of an ODE or a DAE between two nonsmooth *events*. At events, if the evolution of the system is nonsmooth, then a reinitialization is applied. From the numerical point of view, the time integration on smooth phases is performed by any standard one-step or multistep ODE or DAE solvers. This approach needs an accurate location of the events in time which is based on some root-finding procedure.

In order to illustrate a little bit more what can be an event-driven approach for a Filippov's differential inclusion with an exogenous signal $u(t)$, we introduce the notion of *modes*, where the system evolves smoothly. Three modes can be defined as follows,

$$\begin{array}{lll}
\text{mode} - & : \begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) + \dfrac{R}{L}x_2(t) + \dfrac{1}{LC}x_1(t) + \dfrac{u(t)}{L} = 0 \end{cases} & \text{if } x_2 \in \mathscr{I}_- \,, \\[2ex]
\text{mode } 0 & : \begin{cases} \dot{x}_1(t) = 0 \\ \dot{x}_2(t) = 0 \end{cases} & \text{if } x_2 \in \mathscr{I}_0 \,, \\[2ex]
\text{mode} + & : \begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) + \dfrac{R}{L}x_2(t) + \dfrac{1}{LC}x_1(t) + \dfrac{u(t)}{L} = V_\mathrm{z} \end{cases} & \text{if } x_2 \in \mathscr{I}_+ \,,
\end{array} \tag{1.89}$$

respectively associated with the three sets,

$$\begin{array}{l}
\mathscr{I}_-(t) = \{i \in \mathbb{R} \mid i < 0\} \\[1ex]
\mathscr{I}_0(t) \ = \{i \in \mathbb{R} \mid i = 0\} \\[1ex]
\mathscr{I}_+(t) = \{i \in \mathbb{R} \mid i > 0\}
\end{array} . \tag{1.90}$$

In each mode, the dynamical system is represented by an ODE that can be integrated by any ODE solver. The transition between two modes is activated when the sign of a guard function changes, i.e., when an event is detected.

For the modes, "$-$" and "+", it suffices to check that the sign of i is changing to detect an event. A naive approach is to check when the variable x_2 is crossing a threshold $\varepsilon > 0$ sufficiently small. This naive approach may lead to numerical troubles, such as chattering due to the possible drift from the constraint $x_2 = 0$ in the mode when we integrate $\dot{x}_2(t) = 0$. To avoid this artifact, it is better to check the guard functions $v(t)$ and $V_\mathrm{z} - v$, which are dual to the current x_2^+ and x_2^- in the complementarity formalism, see (1.71) with $x_2 = i$. We will see in Chap. 7 that considering a complementarity formulation, or more generally, a formulation that exhibits a duality leads to powerful event-driven schemes.

Once the event is detected, a mode transition has to be performed to provide the time integrator with the new next mode. The operation is made by inspecting the sign of $\dot{x}_2(t)$ at the event by solving for instance the inclusion. We will see also in Chap. 7 that a good manner to perform this task is to relay the mode transition onto a CP resolution.

1.2.5 Conclusions

The message of Sects. 1.2.3 and 1.2.4 is the following: explicit schemes, when applied to Filippov's systems like (1.60), yield poor results. One should prefer implicit schemes. More details on the properties of various methods are provided in Chap. 9. The picture is similar for event-driven algorithms, where one has to be careful with the choice of the variable to check mode transitions. Mode transitions should preferably be steered by the multiplier λ rather than by the state $x(\cdot)$. In mechanics with Coulomb friction, this is equivalent to decide between sticking and sliding, watching whether or not the contact force lies strictly inside the friction cone or on its boundary. For Filippov's inclusion Stewart's method is described in Sect. 7.1.2.

1.3 Mechanical Systems with Coulomb Friction

In this section we treat the case of a one-degree-of-freedom mechanical system subject to Coulomb friction with a bilateral constraint and a constant normal force, as depicted in Fig. 1.20. Its dynamics is given by

$$m\ddot{q}(t) + f(t) \in -mg\mu \ \mathrm{sgn}(\dot{q}(t)) \,, \tag{1.91}$$

where $q(\cdot)$ is the position of the mass, $f(\cdot)$ is some force acting on the mass, g is the gravity, $\mu > 0$ is the friction coefficient. The sign multifunction is defined as

$$\mathrm{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ [-1,1] & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases} \,. \tag{1.92}$$

In view of the foregoing developments one deduces that

$$\mathrm{sgn}(x) = \partial |x| \,, \tag{1.93}$$

i.e., the subdifferential of the absolute value function. It is easy to see that this system is quite similar to the circuit with an ideal Zener diode in (1.61). It can also be expressed using a complementarity formalism as follows:

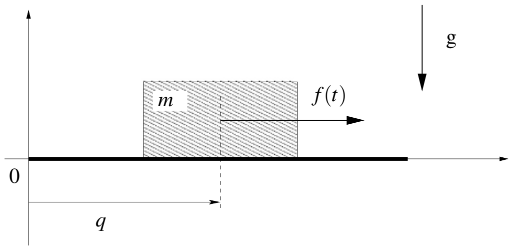


Fig. 1.20. A one-degree-of-freedom mechanical system with Coulomb friction

$$\begin{cases} 0 \leqslant \lambda_1 \perp -x + |x| \geqslant 0 \\ 0 \leqslant \lambda_2 \perp x + |x| \geqslant 0 \\ \lambda_1 + \lambda_2 = 2 \\ \mathrm{sgn}(x) = \dfrac{\lambda_1 - \lambda_2}{2} \end{cases} \tag{1.94}$$

which is quite similar to the set of relations in (1.63). Consequently what has been done for the Zener diode can be redone for such a simple system with Coulomb friction, which is a Filippov's DI.

Similarly to the Zener circuit, the one-degree-of-freedom mechanical system with Coulomb friction can be formulated as an LCS, introducing the positive and the negative parts of the velocity:

$$\begin{cases} \dot{q}(t) = v(t) \\ m\dot{v}(t) + f(t) = -\lambda(t) = \dfrac{1}{2}(\lambda_2(t) + \lambda_1(t)) \\ v^+(t) = v(t) - v^-(t) \\ \lambda_1(t) = 2mg\mu - \lambda_2(t) \\ 0 \leqslant \begin{pmatrix} \lambda_1(t) \\ v^+(t) \end{pmatrix} \perp \begin{pmatrix} -v^-(t) \\ \lambda_2(t) \end{pmatrix} \geqslant 0 \end{cases} . \tag{1.95}$$

1.4 Mechanical Systems with Impacts: The Bouncing Ball Paradigm

In this section some new notions are used, which are all defined later in the book.

1.4.1 The Dynamics

Let us write down the dynamics of a ball with mass *m*, subjected to gravity and to a unilateral constraint on its position, depicted in Fig. 1.21:

$$\begin{cases} m\ddot{q}(t) + f(t) = -mg + \lambda \\ 0 \leqslant q(t) \perp \lambda \geqslant 0 \\ \dot{q}(t^+) = -e\dot{q}(t^-) \text{ if } q(t) = 0 \text{ and } \dot{q}(t^-) \leqslant 0 \\ q(0) = q_0 \geqslant 0, \dot{q}(0^-) = \dot{q}_0 \end{cases} . \tag{1.96}$$

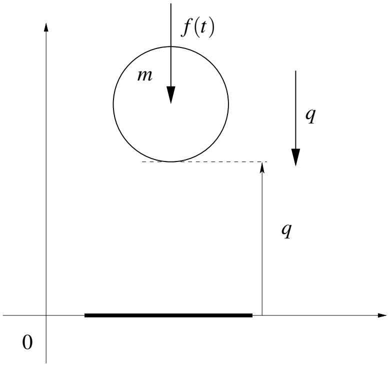


Fig. 1.21. The one-dimensional bouncing ball

The variable λ is a Lagrange multiplier that represents the contact force: it has to remain nonnegative. The complementarity condition between $q(t)$ and λ implies that when $q(t) > 0$ then $\lambda = 0$, while $\lambda > 0$ is possible only if $q(t) = 0$. This is a particular contact model which excludes effects like magnetism (nonzero contact force with $q(t) > 0$) or gluing (negative contact force). This relationship between q and λ is a set-valued function whose graph is as in Fig. 1.1b. The third ingredient in (1.96) is an impact law, which reinitializes the velocity when the trajectory tends to violate the inequality constraint.

Let us analyze the dynamics (1.96) on phases of smooth motion, i.e., either $q(t) > 0$ or $q(t) = 0$ for all $t \in [a,b]$, for some $0 \leqslant a < b$. As seen above the complementarity condition implies that $\lambda(t) = 0$ in the first case. In the second case it allows for $\lambda(t) \geqslant 0$. Let us investigate how the multiplier may be calculated, employing a reasoning similar to the one in Sect. 1.1.5 to get the LCP in (1.24). On $[a,b)$ one has $q(t) = 0$ and $\dot{q}(t) = 0$. So a necessary condition for the inequality constraint not to be violated in a right neighborhood of b is that $\ddot{q}(t^+) \geqslant 0$ on [a, b].

Actually as shown by Glocker (2001, Chap. 7) it is possible to reformulate the contact force law in (1.96), i.e.,

$$-\lambda(t) \in \partial\psi_{\mathbb{R}^+}(q(t)) \Leftrightarrow 0 \leqslant \lambda(t) \perp q(t) \geqslant 0 \tag{1.97}$$

(compare with (1.1), (1.6), (1.7)) at the acceleration level as follows:

$$-\lambda(t^+) \in \begin{cases} 0 & \text{if } q(t) > 0 \\ 0 & \text{if } q(t) = 0 \text{ and } \dot{q}(t^+) > 0 \\ 0 & \text{if } q(t) = 0 \text{ if } \dot{q}(t^+) = 0 \text{ and } \ddot{q}(t^+) > 0 \\ [-\infty, 0] & \text{if } q(t) = 0 \text{ if } \dot{q}(t^+) = 0 \text{ and } \ddot{q}(t^+) = 0 \end{cases} . \tag{1.98}$$

All the functions are expressed as their right limits, since given the state of the system at some instant of time, one is interested to know what happens in the very near future of this time.

Let us now focus on the calculation of $\lambda(t^+)$ in the latter case. Using the dynamics one has

$$m\ddot{q}(t^+) + f(t^+) = -mg + \lambda(t^+)\,. \tag{1.99}$$

From the third and fourth lines of (1.98) we deduce that

$$\begin{cases} 0 \leqslant \ddot{q}(t^+) \perp \lambda(t^+) \geqslant 0 \text{ if } q(t) = 0 \text{ and } \dot{q}(t^+) = 0 \\ \lambda(t^+) = 0 \qquad\qquad\qquad \text{if } (q(t), \dot{q}(t^+)) > 0 \end{cases} . \tag{1.100}$$

where the lexicographical inequality means that the first non zero element has to be positive. Inserting (1.99) into the first line of (1.100) yields

$$0 \leqslant -\frac{1}{m} f(t^+) - g + \frac{1}{m}\lambda(t^+) \perp \lambda(t^+) \geqslant 0 \tag{1.101}$$

which is an LCP with unknown $\lambda(t^+)$. We therefore have derived an LCP allowing us to compute the multiplier. However, this time two differentiations have been needed, when only one differentiation was sufficient to get (1.24).

Remark 1.10. One can rewrite (1.100) as

$$\begin{cases} -\lambda(t^+) \in \partial\psi_{I\!R^+}(\ddot{q}(t^+)) \text{ if } q(t) = 0 \text{ and } \dot{q}(t^+) = 0 \\ \lambda(t^+) = 0 \qquad\qquad\qquad \text{if } (q(t), \dot{q}(t^+)) > 0 \end{cases} . \tag{1.102}$$

Similarly a contact force law at the velocity level can be written as

$$\begin{cases} -\lambda(t^+) \in \partial\psi_{I\!R^+}(\dot{q}(t^+)) \text{ if } q(t) = 0 \\ \lambda(t^+) = 0 \qquad\qquad\qquad \text{if } q(t) > 0 \end{cases} . \tag{1.103}$$

Such various formulations of the contact law strongly rely on Glocker's Proposition C.8 in Appendix C. Notice that inserting (1.97) into (1.96) allows us to express the first and second lines of (1.96) as an inclusion in the cone $\partial\psi_{I\!R^+}(q(t))$

To complete this remark, the whole system (1.103) can be rewritten as a single inclusion as

$$-\lambda(t^+) \in \partial\psi_{T_{I\!R^+}(q(t))}\left(\dot{q}(t^+)\right)\,, \tag{1.104}$$

where $T_{I\!R^+}(q(t))$ is the tangent cone to $I\!R^+$ at $q(t)$: it is equal to $I\!R$ if $q(t) > 0$, and equal to $I\!R^+$ if $q(t) \leqslant 0$. In the same way the whole system (1.102) can be rewritten as

$$-\lambda(t^+) \in \partial\psi_{T_{T_{I\!R^+}(q(t))}(\dot{q}(t^+))}\left(\ddot{q}(t^+)\right)\,, \tag{1.105}$$

where $T_{T_{I\!R^+}(q(t))}(\dot{q}(t^+))$ is the tangent cone at $\dot{q}(t^+)$ to the tangent cone at $q(t)$ to $\mathrm{I\!R}^+$. We will see again such cones in Sect. 5.4.2, for higher order systems.

1.4.2 A Measure Differential Inclusion

Suppose that the velocity is a function of local bounded variation (LBV). This implies that the discontinuity instants are countable, and that for any $t \geqslant 0$ there exists an $\varepsilon > 0$ such that on $(t, t+\varepsilon)$ the velocity is smooth. This also implies that at jump instants the acceleration is a Dirac measure. In fact, the acceleration is the *Stieltjes measure*, or the *differential measure* of the velocity (see Definition C.4).

If we assume that the position $q(\cdot)$ is an Absolutely Continuous (AC) function, we may say that the velocity is equal to some Lebesgue integrable and LBV function $v(\cdot)$ such that

$$q(t) = q(0) + \int_0^t v(s)\mathrm{d}s \, . \tag{1.106}$$

We denote the acceleration as the differential measure dv associated with $v(\cdot)$.

With this material in mind, let us rewrite the system (1.96) as the following DI involving measures:

$$-m\,\mathrm{d}v - f(t)\mathrm{d}t - mg\,\mathrm{d}t \in \partial\psi_{T_{\mathbb{R}^+}(q(t))}\left(\frac{v(t^+) + ev(t^-)}{1+e}\right) . \tag{1.107}$$

We recall that $T_{\mathbb{R}^+}(q(t))$ is the tangent cone to $\mathbb{R}^+$ at $q(t)$. Therefore the right-hand side of the inclusion in (1.107) is the normal cone to the tangent cone $T_{\mathbb{R}^+}(q(t))$, calculated at the "averaged" velocity $\dfrac{v(t^+) + ev(t^-)}{1+e}$, where $v(t^+)$ is the right limit of $v(\cdot)$ at t, and $v(t^-)$ is the left limit.

Let us check that (1.96) and (1.107) represent the same dynamics. On an interval $(t, t+\varepsilon)$ on which the solution is smooth (infinitely differentiable) then

$$v(t) = \dot{q}(t), \quad \mathrm{d}v = \ddot{q}(t)\mathrm{d}t, \quad \frac{v(t^+) + ev(t^-)}{1+e} = \dot{q}(t) \, . \tag{1.108}$$

Thus we obtain

$$-m\ddot{q}(t) - f(t) - mg \in \partial\psi_{T_{\mathbb{R}^+}(q(t))}(\dot{q}(t)) \, . \tag{1.109}$$

We considered intervals of time on which no impact occur, i.e., either $q(t) > 0$ (free motion) or $q(t) = 0$ (constrained motion). In the first case $T_{\mathbb{R}^+}(q(t)) = \mathbb{R}$ so that $\partial\psi_{T_{\mathbb{R}^+}(q(t))}(\dot{q}(t)) = \{0\}$. In the second case $T_{\mathbb{R}^+}(q(t)) = \mathbb{R}^+$. The right-hand side is therefore equal to the normal cone $\partial\psi_{\mathbb{R}^+}(\dot{q}(t))$. So if $\dot{q}(t) = 0$ we get $\partial\psi_{\mathbb{R}^+}(0) = \mathbb{R}^-$. If $\dot{q}(t) > 0$ we get $\partial\psi_{\mathbb{R}^+}(\dot{q}(t)) = \{0\}$. In other words either the velocity is tangential to the constraint (in this simple case zero) and we get the inclusion $-m\ddot{q}(t) - f(t) - mg \in \mathbb{R}^-$, or the velocity points inside the admissible domain and $-m\ddot{q}(t) - f(t) - mg = 0$. One may see the cone in the right-hand side of (1.107) as a way to represent in one shot the contact force law both at the position and the velocity levels.

Let us now consider an impact time t. Then d$v = (v(t^+) - v(t^-))\delta_t$. Since the Lebesgue measure has no atoms, the terms $-f(t)\mathrm{d}t - mg\,\mathrm{d}t$ disappear and we get

$$-m(v(t^+) - v(t^-)) \in \partial\psi_{\mathbb{R}^+}\left(\frac{v(t^+) + ev(t^-)}{1+e}\right) . \tag{1.110}$$

The fact that the inclusion of the measure $m\,\mathrm{d}v$ into a cone can be written as in (1.110) is proved rigorously in Monteiro Marques (1993) and Acary et al. (in press). Since the right-hand side is a cone we can simplify the m and we finally obtain

$$-\frac{v(t^+)+ev(t^-)}{1+e}+v(t^-)\in\partial\psi_{\mathbb{R}^+}\left(\frac{v(t^+)+ev(t^-)}{1+e}\right). \tag{1.111}$$

Now using (1.36) and the fact that $v(t^-)\leqslant 0$ it follows that $v(t^+)+ev(t^-)=0$, which is the impact rule in (1.96).

The measure differential inclusion in (1.107) therefore encompasses all the phases of motion in one compact formulation. It is a particular case of the so-called *Moreau's sweeping process*.

1.4.3 Hints on the Numerical Simulation of the Bouncing Ball

Let us provide now some insights on the consequences of the dynamics in (1.96) and in (1.107) in terms of numerical algorithms.

1.4.3.1 Event-Driven Schemes

One notices that (1.96) contains in its intrinsic formulation some kind of conditional statements ("if...then" test procedure). Such a formalism is close to event-driven schemes. Therefore, we may name it an event-driven-like formalism. Two smooth dynamical modes can be defined from the dynamics in (1.96):

$$\left[\begin{array}{lll}
\text{Mode 1 “free flight”:} & \begin{cases} m\ddot{q}(t^+)+f(t)=-mg \\ \lambda=0 \end{cases} & \text{if } (q(t),\dot{q}(t^+))>0 \\
\text{Mode 2 “contact”:} & \begin{cases} m\ddot{q}(t^+)+f(t)=-mg+\lambda \\ 0\leqslant\ddot{q}(t^+)\perp\lambda\geqslant 0 \end{cases} & \text{if } q(t)=0,\dot{q}(t^+)=0
\end{array}\right. .$$

The sketch of the time integration is as follows:

0. Given the initial data, q_0 $\dot{q}_0$, apply the impact rule if necessary ($q_0=0$ and $\dot{q}_0<0$).
1. Determine the next smooth dynamical mode.
2. Integrate the mode with a suitable ODE or a DAE solver until the constraint is violated.
3. Make an accurate detection/localization of the impact so that the order is preserved.
4. Apply the impact rule if necessary and go back to the step 1.

In the implementation of this algorithm, three issues have to be solved:

- *The time integration of the smooth dynamical modes*. In our simple example, the mode "free flight" is a simple ODE which can be solved by any ODE solver. The mode "contact" needs the computation of the Lagrange multiplier. This can be done by solving λ assuming $\ddot{q}(t) = 0$ and then integrating an ODE or integrating the free flight under the constraints $\ddot{q}(t) = 0$ with a DAE solver.
- *The localization of the event*. The event detection in the mode "free flight" is given by inspecting the sign of $q(\cdot)$. In the mode "contact", this can be done efficiently by inspecting the sign of the Lagrange multiplier λ. All these event detection procedures are implemented with root-finding procedures.
- *The mode transition procedure*. After an event has been detected, the next smooth dynamical mode has to be selected. For that, the sign of the right limit of the acceleration and the Lagrange multiplier λ has to be inspected.

The problem one will face when implementing such an event-driven scheme is that the algorithm stops if there is an accumulation of events (here the impacts). This is the case for the bouncing ball in (1.96) when $f(\cdot) = 0$ and $0 \leqslant e < 1$. How to go "through" the accumulation point? One needs to know what happens after the accumulation, an information which usually is unavailable.

It may be concluded that event-driven algorithms are suitable if there are not too many impacts, and that in such a case an accurate detection/localization of the events may assure an order $p \geqslant 2$ and a good precision during the smooth phases of motion. We had already reached such conclusions in Sect. 1.1.4.

1.4.3.2 Moreau's Time-Stepping Scheme

Let us now turn our attention to the sweeping process in (1.107):

$$\begin{cases} -m\,\mathrm{d}v - f(t)\mathrm{d}t - mg\,\mathrm{d}t = \mathrm{d}\lambda \\ \\ \mathrm{d}\lambda \in \partial\psi_{T_{\mathbb{R}^+}(q(t))}\left(\dfrac{v(t^+) + ev(t^-)}{1+e}\right) . \end{cases} \tag{1.112}$$

The time integration on a time interval $(t_k, t_{k+1}]$ of the first line of this dynamics can be written as

$$\int_{(t_k,t_{k+1}]} m\,\mathrm{d}v + \int_{t_k}^{t_{k+1}} f(t) + mg\,\mathrm{d}t = -\mathrm{d}\lambda\left((t_k,t_{k+1}]\right) . \tag{1.113}$$

Using the definition of a differential measure, we get

$$m(v(t_{k+1}^+) - v(t_k^+)) + \int_{t_k}^{t_{k+1}} f(t) + mg\,\mathrm{d}t = -\mathrm{d}\lambda\left((t_k,t_{k+1}]\right) . \tag{1.114}$$

Let us adopt the convention that

$$v_{k+1} \approx v(t_{k+1}^+) \tag{1.115}$$

and

$$\mu_{k+1} \approx \mathrm{d}\lambda\left((t_k,t_{k+1}]\right), \tag{1.116}$$

that is, the right limit of the velocity $v(t_{k+1}^+)$ is approximated by v_{k+1}, and the measure of the interval $(t_k,t_{k+1}]$ by $\mathrm{d}\lambda$ is approximated by μ_{k+1}. Let us propose the following implicit scheme, which we may call the discrete-time Moreau's second-order sweeping process:

$$\begin{cases} q_{k+1}-q_k = hv_{k+1} \\ m(v_{k+1}-v_k)+h(f_{k+1}+mg) = -\mu_{k+1} \\ \mu_{k+1} \in \partial\psi_{T_{\mathbb{R}^+}(q_k)}\left(\dfrac{v_{k+1}+ev_k}{1+e}\right) \end{cases}. \tag{1.117}$$

After some manipulations (1.117) is rewritten as

$$\begin{cases} q_{k+1}-q_k = hv_{k+1} \\ v_{k+1} = -ev_k+(1+e)\mathrm{prox}[T_{\mathbb{R}^+}(q_k);-b_k] \\ b_k = -v_k+\dfrac{h}{m(1+e)}f_{k+1}+\dfrac{hg}{1+e} \end{cases}. \tag{1.118}$$

Though it looks like that, such a scheme is *not* an implicit Euler scheme. The reasons why have already been detailed in the context of the electrical circuit (**c**) in Sect. 1.1.6 and are recalled here:

- First of all notice that the time step $h > 0$ does not appear in the right-hand side of (1.117). Indeed the set

$$\partial\psi_{T_{\mathbb{R}^+}(q_k)}\left(\frac{v_{k+1}+ev_k}{1+e}\right)$$

  is a cone, whose value does not change when pre-multiplied by a positive constant.
- Secondly, notice that the terms $hf_{k+1}+hmg$ do not represent forces, but forces times one integration interval h, i.e., an impulse. This is the copy of (1.107) in the discrete-time setting. As alluded to above, the dynamics (1.107) is an inclusion of *measures*. In other words, mg is a force, and it may be interpreted as the density of the measure $mg\,\mathrm{d}t$. The integral of $mg\,\mathrm{d}t$ over some time interval is in turn an impulse. As a consequence, the element μ_{k+1} inside the normal cone in the right-hand side of (1.117) is the approximation of the impulse calculated over an interval $(t_k,t_{k+1}]$, as the equation (1.116) confirmed. It is always a *bounded* quantity, even at an impact time.

From a numerical point of view, two major lessons can be learned from this work. First, the various terms manipulated by the numerical algorithm are of finite values. The use of differential measures of the time interval $(t_k,t_{k+1}]$, i.e., $\mathrm{d}v((t_k,t_{k+1}]) =$

$v(t_{k+1}^{+}) - v(t_k^{+})$ and $\mu_{k+1} = \mathrm{d}\lambda\left((t_k, t_{k+1}]\right)$, is fundamental and allows a rigorous treatment of the nonsmooth evolutions. When the time step $h > 0$ converges to zero, it enables one to deal with finite jumps. When the evolution is smooth, the scheme is equivalent to a backward Euler scheme. We can remark that nowhere an approximation of the acceleration is used. Secondly, the inclusion in terms of velocity allows us to treat the displacement as a secondary variable. A viability lemma ensures that the constraints on $q(\cdot)$ will be respected at convergence. We will see further that this formulation gives more stability to the scheme.

These remarks might be viewed only as some numerical tricks. In fact, the mathematical study of the second-order MDI by Moreau provides a sound mathematical ground to this numerical scheme.

1.4.3.3 Simulation of the Bouncing Ball

Let us now provide some numerical results when the time-stepping scheme is applied. They will illustrate some of its properties. In Fig. 1.22, the position, the velocity, and the impulse are depicted. We can observe that the accumulation of impact is approximated without difficulties. The crucial fact that there is no detection of the impact times allows one to pass over the accumulation time. The resulting impulse after the accumulation corresponds to the time integration over a time step of the weight of the ball.

In Fig. 1.23, the energy balance is drawn. We can observe that the total energy is only dissipated at impact. This property is due to the fact that the external forces are constant and therefore, the integration of the free flight is exact. We will see later in the book that these property is retrieved in most general cases by the use of energy-conserving schemes based on θ-methods.

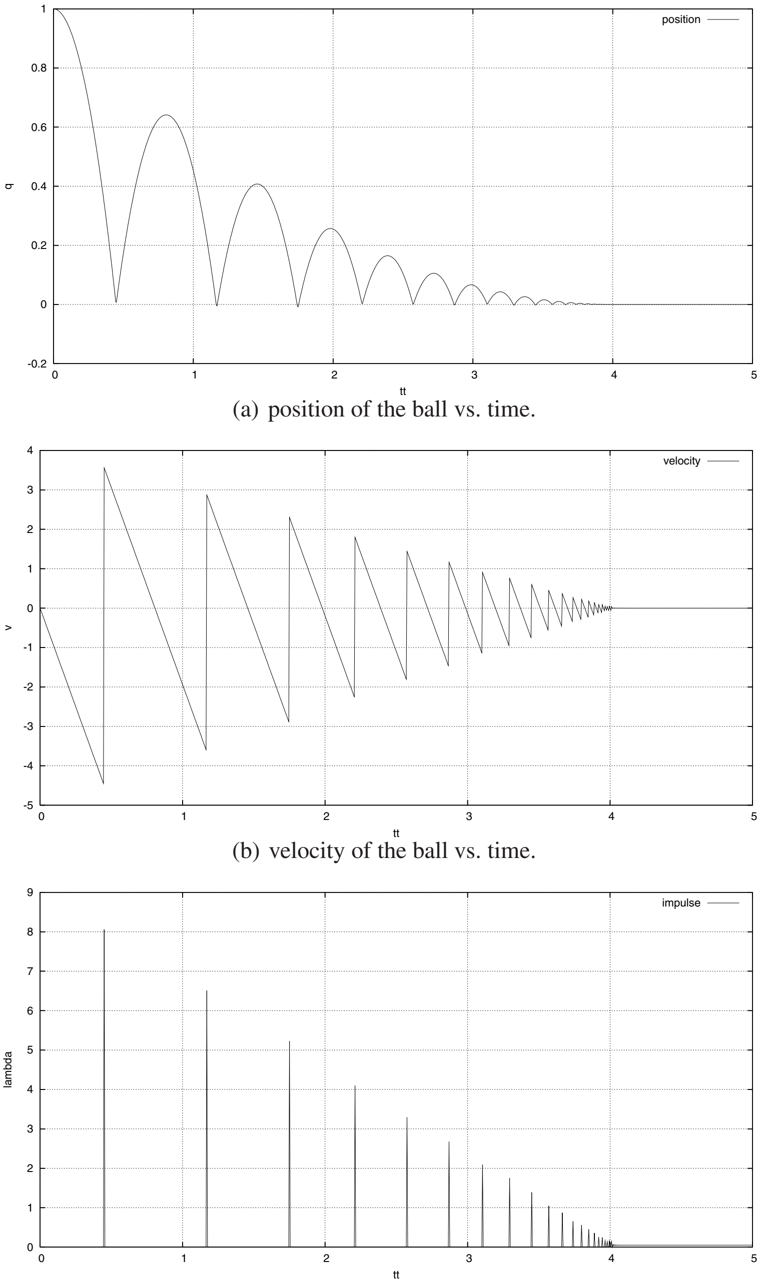
1.4.3.4 Convergence Properties of Moreau's Time-Stepping Algorithm

The convergence of Moreau's time-stepping scheme has been shown in Monteiro Marques (1993), Mabrouk (1998), Stewart (1998), and Dzonou & Monteiro Marques (2007) under various assumptions. Various other ways to discretize such measure differential inclusions with time-stepping algorithms exist together with convergence results. They will be described later in the book.

1.4.3.5 Analogy with the Electrical Circuit

Let us consider again the electrical circuit discrete-time dynamics in (1.34), where we change the notation as $x_{1,k} = q_k$ and $x_{2,k} = v_k$:

$$\begin{cases} q_{k+1} - q_k = h v_{k+1} \\ v_{k+1} - v_k + \dfrac{hR}{L} v_{k+1} + \dfrac{h}{LC} q_{k+1} \in -\partial \psi_{\mathbb{R}^-}(v_{k+1}) \end{cases} . \tag{1.119}$$



(a) position of the ball vs. time.

(b) velocity of the ball vs. time.

(c) impulse vs. time.

Fig. 1.22. Simulation of the bouncing Ball. Moreau's time-stepping scheme. Time step $h = 5 \times 10^{-3}$

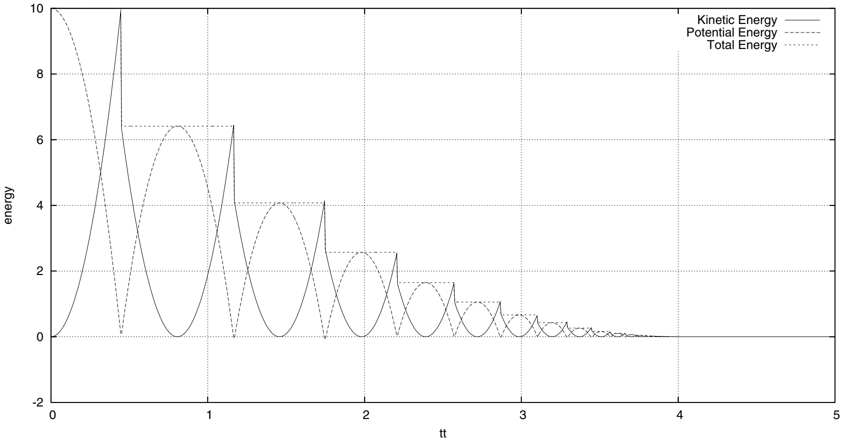


Fig. 1.23. Simulation of the bouncing ball. Moreau's time-stepping scheme. Time step $h = 5 \times 10^{-3}$. Energy vs. time

Let us now consider that the term $f(t) = a_1 v(t) + a_2 q(t)$ for some positive constants a_1 and a_2, and let us take $e = 0$. Then the discretization in (1.117) becomes

$$\begin{cases} q_{k+1} - q_k = h v_{k+1} \\ v_{k+1} - v_k + \dfrac{h a_1}{m} v_{k+1} + \dfrac{h a_2}{m} q_{k+1} + hg \in -\partial \psi_{T_{\mathbb{R}^+}(q_k)}(v_{k+1}) \end{cases} . \qquad (1.120)$$

One concludes that the only difference between both discretizations (1.119) and (1.120) is that the tangent cone $T_{\mathbb{R}^+}(q_k)$ in mechanics is changed to the set $\mathbb{R}^-$ in electricity. This is a simplification, as the tangent cone "switches" between $\mathbb{R}$ and $\mathbb{R}^+$.

With this in mind we may rely on several results to prove the convergence properties of the schemes in (1.119) and (1.120). Convergence results for dissipative electrical circuits may be found in Sect. 9.5.

1.5 Stiff ODEs, Explicit and Implicit Methods, and the Sweeping Process

The bouncing ball dynamics in (1.96) may be considered as the limit when the stiffness $k \longrightarrow +\infty$ of a compliant problem in which the unilateral constraint is replaced by a spring (a penalization) with $k > 0$. It is known that the discretization of a penalized system may lead to stiff systems when k is too large, see e.g. Sect. VII.7 in Hairer et al. (1993). Explicit schemes fail and implicit schemes have to be applied to stiff problems, however, their efficiency may decrease significantly when the required tolerance is small because of possible oscillations with high frequency leading to small step sizes (Hairer et al., 1993, p. 541). Clearly, the rigid body modeling that yields a complementarity formalism and a discretization of the sweeping process via Moreau's time-stepping algorithm may then be of great help.

Let us illustrate this on an even simpler example. A mass $m = 1$ colliding a massless spring-dashpot, whose dynamics is

$$\ddot{q}(t) = u(t) + \begin{cases} -kq(t) - d\dot{q}(t) & \text{if } q(t) \geqslant 0 \\ 0 & \text{if } q(t) \leqslant 0 \end{cases} \tag{1.121}$$

The limit as $k \to +\infty$ is the relative degree two complementarity system

$$\begin{cases} \ddot{q}(t) = u(t) + \lambda \\ 0 \leqslant \lambda \perp q(t) \geqslant 0 \\ \dot{q}(t^+) = -e\dot{q}(t^-) \text{ if } q(t) = 0 \text{ and } \dot{q}(t^-) < 0 \end{cases} \tag{1.122}$$

1.5.1 Discretization of the Penalized System

An explicit discretization of (1.121) yields during the contact phases of motion[5]:

$$\begin{cases} \dfrac{\dot{q}_{i+1} - \dot{q}_i}{h} = -kq_i - d\dot{q}_i + u_{i+1} \\ \dfrac{q_{i+1} - q_i}{h} = \dot{q}_i \end{cases} \Leftrightarrow \begin{pmatrix} q_{i+1} \\ \dot{q}_{i+1} \end{pmatrix} = \begin{pmatrix} 1 & h \\ -hk & 1 - hd \end{pmatrix} \begin{pmatrix} q_i \\ \dot{q}_i \end{pmatrix} + \begin{pmatrix} 0 \\ h \end{pmatrix} u_{i+1} \tag{1.123}$$

The eigenvalues γ_1 and γ_2 of $\begin{pmatrix} 1 & h \\ -hk & 1 - hd \end{pmatrix}$ have a modulus equal to $\frac{1}{2}\sqrt{(2 - hd)^2 + h^2(4k - d^2)}$. The condition for the modulus to be < 1 is $h < \frac{d}{k}$. Therefore, if k is too large then the explicit Euler method is unstable, the system is stiff. Let us now try a fully implicit Euler method. In order to simplify the calculations, we consider $d = 0$, i.e. the system is conservative. One obtains

$$\begin{cases} \dfrac{\dot{q}_{i+1} - \dot{q}_i}{h} = -kq_{i+1} + u_{i+1} \\ \dfrac{q_{i+1} - q_i}{h} = \dot{q}_{i+1} \end{cases} \Leftrightarrow \begin{pmatrix} q_{i+1} \\ \dot{q}_{i+1} \end{pmatrix}$$
$$= a(h,k) \begin{pmatrix} 1 & h \\ -hk & 1 \end{pmatrix} \begin{pmatrix} q_i \\ \dot{q}_i \end{pmatrix} + ha(h,k) \begin{pmatrix} h \\ 1 \end{pmatrix} u_{i+1} \tag{1.124}$$

with $a(h,k) = (1 + h^2k)^{-1}$. This problem is no longer stiff since the modulus of the eigenvalues in this time is equal to 1 (in case $d > 0$ we would obtain a modulus smaller than 1 for any $h > 0$). However, the ratio of the imaginary and the real part of the eignevalues is $h\sqrt{k}$, indicating indeed possible high-frequency oscillations.

[5] The discretization is written with i instead of k to avoid confusion between the stiffness and the number of steps.

1.5.2 The Switching Conditions

We have not discussed yet about the switching condition between the free and the contact motions. Let us rewrite the system (1.121) with $d = 0$ as

$$\ddot{q}(t) = u(t) - \max(kq(t), 0) \tag{1.125}$$

This is easily shown to be equivalent to the relative degree zero complementarity system

$$\begin{cases} \ddot{q}(t) = u(t) - \lambda(t) \\ 0 \leqslant \lambda(t) \perp \lambda(t) - kq(t) \geqslant 0 \end{cases} \tag{1.126}$$

whose implicit Euler discretization is

$$\begin{cases} \dot{q}_{i+1} - \dot{q}_i = hu_{i+1} - h\lambda_{i+1} \\ q_{i+1} - q_i = h\dot{q}_{i+1} \\ 0 \leqslant \lambda_{i+1} \perp \lambda_{i+1} - kq_{i+1} \geqslant 0 \end{cases} \tag{1.127}$$

which after few manipulations becomes the LCP

$$0 \leqslant \lambda_{i+1} \perp (1 + h^2k)\lambda_{i+1} - kh\dot{q}_i - kh^2u_{i+1} - kq_i \geqslant 0 \tag{1.128}$$

that is easily solved for λ_{i+1} and permits to advance the method from step i to step $i+1$. With the switching condition $q_{i+1} \geqslant 0$ or $q_{i+1} \leqslant 0$, one retrieves the implicit method (1.124). If the complementarity relation is taken as $0 \leqslant \lambda_{i+1} \perp \lambda_{i+1} - kq_i \geqslant 0$ and $q_{i+1} - q_i = h\dot{q}_i$, one recovers the explicit method with a switching condition $q_i \geqslant 0$ or $q_i \leqslant 0$. We conclude that the complementarity formulation of (1.121) allows us to clarify the choice of the switching variable and of the manner to compute the new state *via* an LCP, but does not bring any novelty concerning the stiff/nonstiff issue. One also notes that the explicit method for (1.125) yields again (1.123). Therefore, applying an explicit Euler method to (1.121), (1.125), or (1.126) is equivalent. The implicit discretization of (1.125), i.e. $\dot{q}_{i+1} = \dot{q}_i + hu_{i+1} - h\max(kq_i + kh\dot{q}_{i+1}, 0)$, is obviously also equivalent to (1.127). But its direct solving without resorting to the LCP in (1.128) is not quite clear. One may say that the CP formalism is a way to implicitly discretize the projection.

All these comments apply to the circuits **(a)** and **(b)** in (1.11) (1.12), and the various formulations in (1.15) through (1.22).

Remark 1.11. Without the complementarity interpretation in (1.126) that yields the LCP (1.128), one may encounter difficulties in implementing the switching with q_{i+1} and $q_{i+1} - q_i = h\dot{q}_{i+1}$, because the system is a piecewise linear system with an implicit switching condition. Consequently, one often chooses an implicit method with an explicit switching variable $q_{i+1} - q_i = h\dot{q}_i$. This boils down to a semi explicit/implicit method which also yields a stiff system.

1.5.3 Discretization of the Relative Degree Two Complementarity System

Moreau's time stepping method for (1.122) is

$$\begin{cases} \frac{\dot{q}_{i+1}+e\dot{q}_i}{1+e} = \mathrm{prox}[T_{\mathbb{R}^+}(q_{i+1}); \dot{q}_i + \frac{h}{1+e}u(t_{i+1})] \\ \\ q_{i+1} = q_i + h\dot{q}_i \end{cases} \tag{1.129}$$

which is nothing else but solving a simple LCP (or a QP) at each step. It is noteworthy that we could have written a fully implicit scheme with $q_{i+1} = q_i + h\dot{q}_{i+1}$ without modifying the conclusion: Moreau's time stepping method is not stiff.

1.6 Summary of the Main Ideas

- Simple physical systems yield different types of dynamics:
  - ODEs with Lipschitz-continuous vector field
  - Differential inclusions with compact, convex right-hand sides (like Filippov's inclusions)
  - Differential inclusions in normal cones (like Moreau's sweeping process)
  - Measure differential inclusions
  - Evolution variational inequalities
  - Linear complementarity systems

  Some of these formalisms may be shown to be equivalent, see (Brogliato et al. (2006)).
- The nonsmooth formalisms may be useful to avoid stiff problems. All these systems possess solutions which are not differentiable everywhere, and may even jump (absolutely continuous, locally bounded variation solutions).
- There exist two types of numerical schemes for the integration of these nonsmooth systems:
  - *The event-driven (or event-tracking) schemes.* One supposes that between events (instants of nondifferentiability), the solutions are differentiable enough, so that any standard high-order scheme (Runge–Kutta methods, extrapolation methods, multistep methods, ...) may be used until an event is detected. The event detection/localization has to be accurate enough so that the order is preserved. Once the event has been treated, continue the integration with your favorite scheme. This procedure may fail when there are too many events (like for instance an accumulation).
  - *The time-stepping (or event-capturing) schemes.* The whole dynamics (differential and algebraic parts) is discretized in one shot. Habitually low-order (Euler-like) schemes are used (other, higher order methods may in some cases be applied, however, the nonsmoothness brings back the order to one). Advancing the scheme from step k to step $k+1$ requires to solve a complementarity problem, or a quadratic problem, or a projection algorithm. Convergence results have been proved.

- Though the time-stepping schemes look like Euler schemes, they are not. The primary variables are chosen so that even in the presence of Dirac measures, all the calculated quantities are bounded for all times. These schemes do not try to approximate the Dirac measures at an impact. They approximate the measures of the integration intervals, which indeed are always bounded. From a mathematical point of view, this may be explained from the fact that the right-hand sides are cones (hence pre-multiplication by the time step $h > 0$ is equivalent to pre-multiplication by 1).
- There are strong analogies between nonsmooth electrical circuits and nonsmooth mechanical systems. More may be found in Möller & Glocker (2007). The solutions of nonsmooth electrical circuits may jump, so that they are rigorously represented by *measure differential inclusions*. The fact that switching networks may contain Dirac measures has been noticed since a long time in the circuits literature (Bedrosian & Vlach, 1992). Proper simulation tools for nonsmooth systems are necessary, because the integrators based on stiff, so-called "physical" models may provide poor, unreliable results (Bedrosian & Vlach, 1992).

Part I

Formulations of Nonsmooth Dynamical Systems

2

Nonsmooth Dynamical Systems: A Short Zoology

The goal of this chapter is to present several examples of well-identified nonsmooth dynamical systems (NSDS). The purpose is not to be exhaustive nor to provide all the mathematical details, but rather to fix the notation and give an overview of the variety of NSDS, so that the reader gets an idea of what the realm of NSDS looks like. Equivalences exist between the formalisms that are presented and will be pointed out.

2.1 Differential Inclusions

In this section we review some "classical" notions and formalisms of differential inclusions, with an emphasis on the so-called Filippov's systems. In Sect. 2.2 another type of inclusions will be examined, which usually do not satisfy the same set of assumptions.

Definition 2.1 (Differential inclusion). *A differential inclusion (DI) may be defined by*

$$\dot{x}(t) \in F(t,x(t)),\ t \in [0,T],\ \ x(0) = x_0\ , \tag{2.1}$$

where $x : \mathbb{R} \rightarrow \mathbb{R}^n$ *is a function of time* t, $\dot{x} : \mathbb{R} \rightarrow \mathbb{R}^n$ *is its time derivative,* $F : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ *is a set-valued map which associates to any point* $x \in \mathbb{R}^n$ *and time* $t \in \mathbb{R}$ *a set* $F(t,x) \subset \mathbb{R}^n$, *and* $T > 0$.

In general the inclusion will be satisfied almost everywhere on $[0,T]$, because $x(\cdot)$ may not be differentiable for all $t \in [0,T]$. If $x(\cdot)$ is absolutely continuous then $\dot{x}(\cdot)$ is defined up to a set of Lebesgue measure zero on $[0,T]$. In fact it happens that there are several very different types of differential inclusions, depending on what the sets $F(x)$ look like. For instance when $n = 1$, some sets $F(x)$ may contain bounded vertical segments (the sign multivalued function) or unbounded vertical lines (see Fig. 1.1b), some others may be such that their graph has positive measure in the plane ($F(x) = [-1,1]$ for all x), some others may have all of these features, etc. One may expect that not only the modeling motivations that yield such DIs are not the same, but that the subsequent mathematical and numerical analysis will differ a lot as well. A significant number of dynamical systems can be described by a differential inclusion. First of them are the standard ODEs

$$\dot{x}(t) = f(x(t),t) \tag{2.2}$$

taking for the set-valued map the singleton $F(t,x) = \{f(x,t)\}$. An implicit differential equation

$$f(\dot{x}(t),x(t)) = 0 \tag{2.3}$$

can also be cast into a DI by defining the set-valued map as $F(x) = \{v \mid f(v,x) = 0\}$.

The definition, existence, and uniqueness of solutions for such types of dynamical systems is not a trivial task and strongly depends on the boundary conditions which are prescribed (initial or Cauchy conditions or two-point boundary conditions for instance) and naturally on the regularity of the function $x(\cdot)$ together with the set-valued map $F(\cdot,\cdot)$.

In standard books on DI (Aubin & Cellina, 1984; Deimling, 1992; Smirnov, 2002), the trajectory $x(\cdot)$ is usually assumed to be absolutely continuous. This is the case for instance with the Lipschitzian and upper semi-continuous set-valued right-hand sides (see below). Most of such systems can be regarded as nonsmooth dynamical systems due to two main reasons. The first one is the extensive use of nonsmooth analysis and set-valued analysis to study their properties and the second one is related with the possible nonsmoothness of the time derivative $\dot{x}(\cdot)$ due to the constraints imposed by the inclusion.

We will give in the sequel some illustrative examples of such differential inclusions where the nonsmoothness plays an important role. The cases of Lipschitzian and upper semi-continuous right-hand sides are the most well known in the literature. We will recall their definition quickly. For our purpose, we are more interested in the unilateral differential inclusions which are more representative of our applications. This type of DI is characterized by unbounded sets as set-valued right-hand sides, while more usual DIs are based on compact sets.

2.1.1 Lipschitzian Differential Inclusions

Let us deal with the autonomous case, i.e., $\dot{x}(t) \in F(x(t))$.

Definition 2.2 (Lipschitzian DI). *A DI is said to be Lipschitzian if the set-valued map $F : \mathbb{R} \rightarrow \mathbb{R}^n$ satisfies the following conditions:*

1. *the sets $F(x)$ are closed and convex for all $x \in \mathbb{R}^n$;*
2. *the set-valued map $F(\cdot)$ is Lipschitzian with a bounded constant l, i.e.,*

$$\exists\, l \geqslant 0, \quad F(x_1) \subset F(x_2) + l\|x_1 - x_2\|B_n \tag{2.4}$$

for all $x_1 \in \mathbb{R}^n$, $x_2 \in \mathbb{R}^n$, where B_n is the unit ball of $\mathbb{R}^n$, i.e., $B_n = \{y \in \mathbb{R}^n \mid \|y\| \leqslant 1\}$.

The Lipschitz continuity is sometimes stated with a function $l : \mathbb{R}^+ \rightarrow \mathbb{R}^+$, $l(\cdot)$ Lebesgue integrable. Recall that given two sets A and $B \subset \mathbb{R}^n$, one has $A+B = \{a+b \mid a \in A, b \in B\}$. Thus for instance $[-1,1]+[-1,1] = [-2,2]$, and $[-2,-1]+[1,2] = [-1,1]$.

Example 2.3. Let $F : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto \begin{cases} [bx, ax] \text{ if } x \geqslant 0 \\ \\ [ax, bx] \text{ if } x \leqslant 0 \end{cases}$, with $a > b > 0$. Let $x_1 < 0$ and $x_2 > 0$. Then $F(x_1) = [ax_1, bx_1]$, $F(x_2) = [bx_2, ax_2]$, and $a|x_1 - x_2|B_1 = [-a(x_2 - x_1), a(x_2 - x_1)]$. Therefore $F(x_2) + a|x_1 - x_2|B_1 = [(b-a)x_2 + ax_1, 2ax_2 - ax_1]$, and one may check that $F(x_1) \subset [(b-a)x_2 + ax_1, 2ax_2 - ax_1]$. A similar reasoning may be done when $x_1 > 0$ and $x_2 < 0$. We conclude that $F(\cdot)$ is Lipschitz continuous with $l = a$ in (2.4).

The most common examples of Lipschitzian DI are issued from the control and systems theory. Let us consider an ODE depending on a control parameter $u \in U \subset \mathbb{R}^m$:

$$\dot{x}(t) = f(x(t), u) \; , \tag{2.5}$$

where $f : \mathbb{R}^n \times U \rightarrow \mathbb{R}^n$ is assumed to be a continuous function satisfying a Lipschitz condition in x and such that the set $f(x, U)$ is closed and convex for all $x \in \mathbb{R}^n$. If $u(t)$ is an admissible control (for instance all bounded measurable functions satisfying $u(t) \in U$ a.e.), the Cauchy problem

$$\dot{x}(t) = f(x(t), u(t)),\ t \in [0, T],\ x(0) = x_0 \tag{2.6}$$

has a solution $x(\cdot)$. The connection between differential inclusions and such a control system is given by the following DI:

$$\dot{x}(t) \in \cup_{u \in U} f(x(t), u) \; . \tag{2.7}$$

The solution of the Cauchy problem (2.6) is a solution of the DI (2.7) and thanks to a result of Filippov, the converse statement is also true in the sense that there exists a solution $v(t)$ of the inclusion (2.7) which is also a solution of (2.6). More formally we have the following result (Li, 2007; Nieuwenhuis, 1981).

Theorem 2.4. *Let the set of admissible control inputs U be compact, and let $f : \mathbb{R}^n \times U \rightarrow \mathbb{R}^n$ be continuous with $F(x) = \{f(x,u) \mid u \in U\}$ convex for each $x \in \mathbb{R}^n$. The open-loop system (2.5) is equivalent to the DI $\dot{x}(t) \in F(x(t))$.*

Under the stated assumptions, the set-valued mapping $F(\cdot)$ is continuous with compact convex images. Equivalence means that any absolutely continuous function $x(\cdot)$ that satisfies (2.5) with an admissible $u(\cdot)$ satisfies the DI, and there always exists a solution of the DI that is also a solution of the ODE (2.5). As the next example shows both formalisms are in fact not really equivalent to each other.

Example 2.5. Consider the scalar controlled system

$$\dot{x}(t) = x(t)u(t),\ \ x(0) = x_0\ , \tag{2.8}$$

where $x(t) \in \mathbb{R}$, $u(\cdot)$ is a measurable function that takes its values in $[-1,1]$, that is $u(t) \in [-1,1]$ for all $t \in \mathbb{R}$. Given a control input $u(\cdot)$ one finds that the solution of the ODE (2.8) is $x(t) = x_0 \exp\left(\int_0^t u(s)\mathrm{d}s\right)$. In particular if $x_0 = 0$ then $x(t) = 0$ for all $t \geqslant 0$. We now associate to the ODE (2.8) the DI

$$\dot{x}(t) \in [-x(t), x(t)],\ \ x(0) = x_0\ . \tag{2.9}$$

Let $x_0 = 0$. Then $x(t) = 0$ for all $t \geqslant 0$ is a solution, but $x(t) = 0$ for $0 \leqslant t \leqslant \sqrt{2}$ and $x(t) = t^2$ for $t \geqslant \sqrt{2}$ are also solutions. Thus embedding the system into a DI formalism adds more solutions. In a sense the DI is more loose than the ODE.

The next lemma is a tool for testing the Lipschitz continuity of a set-valued map.

Lemma 2.6. *Let the graph of $F(\cdot)$, i.e., the set $\{(x,y) \in \mathbb{R}^n \times \mathbb{R}^n \mid y \in F(x)\}$, be closed and convex. If $F(x_0) \subset mB_n$, $m > 0$, x_0 ∈Int(dom(F)), then $F(\cdot)$ is Lipschitzian in a neighborhood of x.*

Example 2.7. $F(x) = [-1,1]$, $F(x) = [-x, 2x]$ are Lipschitz continuous. The relay multifunction, or sign multifunction, does not satisfy the assumptions of the lemma. The values of $F(x)$ are convex, but not its graph. Actually it is not Lipschitz continuous.

The following holds (Smirnov, 2002):

Lemma 2.8. *Let the set $F(x)$ satisfy the conditions of Definition 2.2. Then for any $x_0 \in \mathbb{R}^n$ there exists a solution to the DI (2.1) on $\mathbb{R}^+$ with $x(0) = x_0$.*

Let us end this section on Lipschitz-ontinuous DIs with a theorem due to Filippov. It is a result that is useful for the study of discrete approximations, like the Euler method presented in Sect. 9.2. In fact it enables one to prove that every approximated solution of the discretized DI contains in its neighborhood a solution of the DI, when the time step is small enough.

Theorem 2.9. *Let $F(\cdot,\cdot)$ be Hausdorff continuous and Lipschitz continuous with constant $L(\cdot)$ in the region $\{(t,x) \mid t \geqslant 0, ||x - y(t)|| \leqslant b\}$. Let $y : \mathbb{R}^+ \rightarrow \mathbb{R}^n$ be absolutely continuous and dist$(\dot{y}(t)), F(t, y(t)) \leqslant g(t)$ almost everywhere in $\mathbb{R}^+$, where $g(\cdot)$ is integrable. Let x_0 be such that $||x_0 - y(0)|| < b$, and*

$$m(t) = \int_0^t L(s)ds,$$

$$v(t) = \exp(m(t))\left(||x_0 - y(0)|| + \int_0^t \exp(-m(s))g(s)ds\right).$$

Then there exists a solution of the DI: $\dot{x}(t) \in F(t, x(t))$ on the interval $\Delta = \{t \in \mathbb{R}^+ \mid v(t) \leqslant b\}$ that satisfies

$$||x(t) - y(t)|| \leqslant v(t) \ \textit{for all}\ t \in \Delta,\ x(0) = x_0,$$

$$||\dot{x}(t) - \dot{y}(t)|| \leqslant L(t)v(t) + g(t) \ \textit{almost everywhere on}\ \Delta.$$

The distance between a point and a set is as in appendix A. Theorem 2.9 is a technical result that is used in Theorems 9.1 and 9.2, which concern the properties of discretized inclusions. In these theorems upper bounds on the distance between the approximated solution and one solution of the continuous-time DI are obtained.

2.1.2 Upper Semi-continuous DIs and Discontinuous Differential Equations

Another class of DIs is the class of the upper semi-continuous DIs which plays a fundamental role in optimal control theory and in the study of ODE with discontinuous right-hand side.

Definition 2.10 (Upper semi-continuous DI). *A DI is said to be upper semi-continuous if the set-valued map $F : \mathrm{I\!R} \to \mathrm{I\!R}^n$ satisfies the following conditions:*

1. *the sets $F(x)$ are closed and convex for all $x \in \mathrm{I\!R}^n$;*
2. *the set-valued map $F(\cdot)$ is upper semi-continuous, i.e., for every open set M containing $F(x), x \in \mathrm{I\!R}$, there exists a neighborhood Ω of x such that $F(\Omega) \subset M$.*

Upper semi-continuity for set-valued mappings is sometimes called *outer semi-continuity* (Rockafellar & Wets, 1998).[1] An equivalent formulation is that $\limsup_{x \to x_0} F(x) \subset F(x_0)$: then $F(\cdot)$ is outer semi-continuous at x_0. It may be more appropriate to use outer rather than upper semi-continuous. Indeed when specialized to single-valued functions, Definition 2.10 implies the continuity. However, upper semi-continuous single-valued functions may be discontinuous (when upper semi-continuity in the sense of single-valued functions is considered). One concludes that the upper semi-continuity for set-valued functions (Definition 2.10) and upper semi-continuity for single-valued functions are two different notions with the same name. A useful result taken from Smirnov (2002) is as follows (other characterizations of upper semi-continuity can be found in Chap. 1, Sect. 1 of Deimling, 1992):

Proposition 2.11. *Assume that the graph of the set-valued mapping $F(\cdot)$, i.e., $\{(x,y) \in \mathrm{I\!R}^n \times \mathrm{I\!R}^n \mid y \in F(x)\}$, is closed and the closure of the set $N_\delta = \{F(x) \mid ||x - x_0|| < \delta\}$, $\delta > 0$, is compact. Then $F(\cdot)$ is upper semi-continuous at x_0. On the other hand, if $F(\cdot)$ is upper semi-continuous, then its graph is closed.*

Using this proposition one may immediately conclude that the sign set-valued function (also called the relay function in systems and control) is upper semi-continuous. Indeed all the sets N_δ are equal either to $\{1\}$ or to $\{-1\}$ or to $[-1,1]$. Another useful result is that all maximal monotone mappings are upper semi-continuous (exercise 12.8 in Rockafellar & Wets, 1998): indeed maximal monotonicity implies that the graph is closed. Obviously not all upper semi-continuous set-valued mappings are monotone.

[1] After a suggestion made by J.B. Hiriart-Urruty.

Example 2.12. Consider the set-valued mapping $x \mapsto \partial\Psi_K(x)$ with $K = \mathbb{R}^+$ and $\psi_K(\cdot)$ is the indicator function of K, see (1.4). The graph of this mapping is similar to the graph depicted in Fig. 1.1b, with the vertical branch directed to the bottom. This set-valued mapping is upper semi-continuous. Indeed $F(0) = \mathbb{R}^-$. The open sets containing $F(0)$ are of the form $(-\infty, a)$ with $a > 0$. In any neighborhood of 0 one has $F(x) = \emptyset$ or $F(x) = 0$, so $F(x) \subset (-\infty, a)$.

Let us state an existence result for upper semi-continuous DIs, taken from Deimling (1992).

Lemma 2.13. *Let $F(x)$ satisfy the conditions of Definition 2.10, and in addition $||F(x)|| \leqslant c(1 + ||x||)$ for some $c > 0$ and all $x \in \mathbb{R}^n$. Then there is an absolutely continuous solution to the DI (2.1) on $\mathbb{R}^+$, for every $x_0 \in \mathbb{R}^n$.*

This result extends to time-varying inclusions $F(t,x)$ (theorem 5.1 in Deimling, 1992). Examples in Sect. 5.2 of Deimling (1992) show that upper semi-continuity alone is not sufficient to guarantee the existence. The convexification of the sets $F(x)$ is needed. Notice that Lemma 2.13 does not apply to the right-hand side in Example 2.12 which obviously does not satisfy the growth condition.

Example 2.14. Let $F : \mathbb{R} \to \mathbb{R}, x \mapsto [-1, 1]$. This set-valued function satisfies the conditions of Lemma 2.13. Let $x(0) = 0$. Then $x(t) = 0$ for all $t \geqslant 0$ is a solution of the DI, as well as $x(t) = at$ for any $a \in [-1, 1]$. There is an infinity of solutions.

Example 2.15. The scalar DI

$$\dot{x}(t) \in \begin{cases} [-8x(x+1)^2, 2] \text{ if } -1 \leqslant x \leqslant 0 \\ \\ -8x(x+1)^2 \qquad \text{otherwise} \end{cases} \tag{2.10}$$

is upper semi-continuous, and the sets $F(x)$ are closed and convex for all $x \in \mathbb{R}^n$.

Let us now consider the case of an ODE with a discontinuous right-hand side:

$$\dot{x}(t) = f(x(t)),\ t \in [0, T],\ x(0) = x_0\ , \tag{2.11}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ is a bounded function. If $f(\cdot)$ is not continuous, then the Cauchy problem associated with this ODE may have no solution. A standard example is given by the following right-hand side:

$$f(x,t) = \begin{cases} 1 & \text{if } x < 0 \\ -1 & \text{if } x \geqslant 0 \end{cases} \tag{2.12}$$

with the initial condition $x(0) = x_0$. It is clear that the problem has no solution in the usual sense. Indeed, if $x(t) < 0$, then the solution is of the form $x(t) = t + x_0$. On the contrary, if $x(t) \geqslant 0$, then the solution is of the form $x(t) = -t + x_0$. Each solution reaches the point $x = 0$ and cannot leave it. Unfortunately, the function $x(t) \equiv 0$ does not satisfy the equation, since $\dot{x} = 0 \neq f(0) = -1$.

2.1.2.1 Filippov's Inclusions

A way to circumvent the problem encountered with (2.12) is to define a new type of solutions or more precisely to change slightly the model. This is the goal of the Filippov solutions. The minimal requirement is to ensure the existence of solutions in simple cases for any initial conditions and to coincide with the standard solution of ODE with continuous right-hand side. Filippov (1988) introduced a method to fulfill these requirements by considering the set-valued map $F(\cdot)$ with $F(x)$ defined by

$$\dot{x}(t) \in F(x(t)) = \bigcap_{\varepsilon>0} \bigcap_{\mu(N)=0} \overline{\mathrm{conv}} f((x(t)+\varepsilon B_n) \setminus N) \,, \tag{2.13}$$

where B_n is the unit ball of $\mathbb{R}^n$, and the sets N are all sets of zero Lebesgue measure. $\overline{\mathrm{conv}}\,(\cdot)$ denotes the closure of the convex hall, and the set $\{x+\varepsilon B_n\} = \{y \in \mathbb{R}^n / y \in \{x\}+\varepsilon B_n\}$. The notation $F((x+\varepsilon B_n)\setminus N)$ means all the vector fields $F(y)$ with y in the set $\{x+\varepsilon B_n\}$ minus sets N. Since the function $f(\cdot)$ is assumed to be bounded and the graph of $F(\cdot)$ is closed, the set-valued map $F(\cdot)$ is upper semi-continuous (see Proposition 2.11). It is noteworthy that $F(x) = \{f(x)\}$ whenever $f(\cdot)$ is continuous (see Theorem 2.22 for more properties of the right-hand side in (2.13)). Thanks to Lemma 2.13, the Cauchy problem $\dot{x}(t) \in F(x(t)), x(0) = x_0$, with $F(\cdot)$ in (2.13) always has an absolutely continuous solution. Moreover, some mathematical properties like compactness, connectedness, and boundary property are retrieved as in the case of standard ODE. Usually, an absolutely continuous function is said to be a Filippov solution or a solution in the sense of Filippov of the Cauchy problem (2.11) if it is a solution of the DI (2.13).

Remark 2.16. Results similar to Proposition 2.11 to characterize the right-hand side of Filippov's systems exist in the literature, see for instance theorem 2.3 in Kastner-Maresch (1992) or proposition 1 in Aubin & Cellina (1984).

A geometrical interpretation of the definition (2.13) and of the Filippov solution can be given for a function $f : \mathbb{R}^n \to \mathbb{R}^n$ which is discontinuous on a smooth surface S. Let us assume that the surface S separates $\mathbb{R}^n$ into two open domains Ω^+ and Ω^-. The set $F(x), x \in S$, is the segment comprised between the two values:

$$f^-(x) = \lim_{x' \to x, x' \in \Omega^-} f(x'), \quad f^+(x) = \lim_{x' \to x, x' \in \Omega^+} f(x') \,. \tag{2.14}$$

Two behaviors for the dynamical system can be envisaged:

- If the set $F(x)$ lies on one side of the tangent plane $T(x,S)$ of the surface S at x, then the solution passes from one side to the other. This is called a *crossing* solution.
- If the set $F(x)$ intersects the tangent plane, $T(x,S)$, the intersection point $f_S(x) = F(x) \cap T(x,S)$ defines the velocity along the surface. In this case, such a solution is called a *sliding* solution. This is the case when the right-hand side is directed towards the surface S on both sides (then S is said to be *attractive*).
- When the vector fields on the surface S are directed to the interior of the domains Ω^+ and Ω^- for $x \in \Omega^+$ and $x \in \Omega^-$, respectively, S is said to be *repulsive* and there may exist the so-called *spontaneous* switches.

In general there may be portions of S which are attractive, portions which are crossing, and others which are repulsive. It is noteworthy that the value of $f(\cdot)$ on S is irrelevant in Filippov's framework, as zero measure sets are excluded from the definition of the right-hand side $F(x)$.

A similar idea has been developed in Utkin (1977). A typical case is depicted in Fig. 2.1. It is noteworthy that in general the system may not be defined by a single switching surface. There may exist several surfaces, with nonvoid intersections. See Sect. 7.1 for more examples. Filippov's inclusions were originally motivated by control applications. But they also find applications in the modeling of genetic networks (Grognard et al., 2007).

2.1.2.2 Examples and Comments

Let us now provide some more examples and comments, which highlight the relationships that may exist between Filippov's definition of a solution, and other definitions. See also Sect. 7.1.

Example 2.17. Let us consider

$$\dot{x}(t) = \begin{cases} 1 \text{ if } x(t) \neq 0 \\ \\ 0 \text{ if } x(t) = 0 \end{cases} \tag{2.15}$$

with $x(t) \in \mathbb{R}$. According to (2.13) it follows that Filippov's inclusion for this discontinuous ODE is

$$\dot{x}(t) = 1 \;, \tag{2.16}$$

which is not an inclusion but a (very) simple ODE.

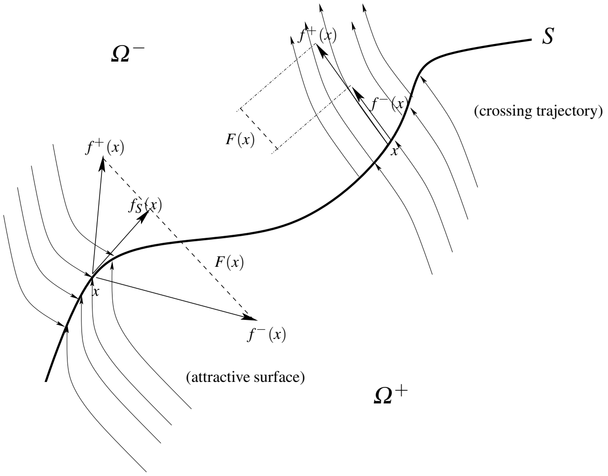


Fig. 2.1. The switching surface S

Example 2.18. Let us consider

$$\dot{x}(t) = \begin{cases} 1 \text{ if } x \in \mathbb{Q} \\ 0 \text{ if } x \notin \mathbb{Q} \end{cases} \tag{2.17}$$

with $x(t) \in \mathbb{R}$. Now from (2.13) and since the set of rational numbers $\mathbb{Q}$ is of zero Lebesgue measure in $\mathbb{R}$, one finds that the Filippov's inclusion is

$$\dot{x}(t) = 0\,, \tag{2.18}$$

which once again is a simple ODE.

Example 2.19. This example is taken from Stewart (1990). Let us consider

$$\dot{x}(t) = g(t) - \mathrm{sgn}(x(t)) \tag{2.19}$$

with $x(t) \in \mathbb{R}$, $|g(t)| \leqslant 1$ for all t, and we suppose first that $\mathrm{sgn}(\cdot)$ is single valued. Let $x(0) = 0$, $t_0 = 0$. Suppose that $x(t_1) > 0$ for some $t_1 > 0$. Since we are looking for an absolutely continuous solution, there must exist some $t_2 \in (0, t_1)$ such that $x(t_2) > 0$ and thus $\dot{x}(t_2) = g(t_2) - \mathrm{sgn}(x(t_2)) > 0$. But since $|g(t_2)| \leqslant 1$ and $x(t_2) > 0$, one has $\dot{x}(t_2) \leqslant 0$, a contradiction. Therefore $x(t_1) \leqslant 0$ for all $t_1 > 0$. A similar reasoning may be done starting with $x(t_1) < 0$, and one then concludes that $x(t_1) \geqslant 0$ for all $t_1 > 0$. Consequently $x(t) = 0$ for all $t \geqslant 0$. But this implies that $\dot{x}(t) = g(t) - \mathrm{sgn}(0) = 0$ almost everywhere on $\mathbb{R}$. Except if $g(t)$ corresponds to the value $\mathrm{sgn}(0)$ this is impossible. Now let us transform (2.19) into a Filippov's inclusion. We obtain

$$\dot{x}(t) - g(t) \in \mathrm{sgn}(x(t)) = \begin{cases} 1 & \text{if } x(t) > 0 \\ [-1,1] & \text{if } x(t) = 0 \\ -1 & \text{if } x(t) < 0\,. \end{cases} \tag{2.20}$$

This is a time-varying inclusion. Measurability of $g(\cdot)$ assures the existence of at least one absolutely continuous solution, for any initial data (from Proposition 2.11 it follows that $F(t,\cdot)$ is upper semi-continuous, and theorem 5.2 in Deimling, 1992 then applies).

Example 2.20 (Carathéodory's solutions vs. Filippov's solutions). Consider once again Example 2.17. If we initialize the system at $x(0) = x_0 < 0$, then the solution will be $x(t) = t$ until it attains 0 at time $-x_0$. Then $\dot{x}(-x_0) = 0$ and the solution stays at 0. If $x_0 = 0$ the solution is $x(t) = 0$ for all $t \geqslant 0$. If $x_0 > 0$ then $x(t) = t$ for all $t \geqslant 0$. It is apparent that such solutions drastically differ from Filippov's solutions. They are called Carathéodory's solutions. As a further illustration consider the following:

$$\dot{x}(t) = \begin{cases} -1 & \text{if } x(t) < 0 \\ \{-2,0,2\} & \text{if } x(t) = 0 \\ 1 & \text{if } x(t) > 0\,. \end{cases} \tag{2.21}$$

Filippov's inclusion is

$$\dot{x}(t) = \begin{cases} -1 & \text{if } x(t) < 0 \\ [-1,1] & \text{if } x(t) = 0 \\ 1 & \text{if } x(t) > 0 \end{cases} \tag{2.22}$$

because the sets of zero measure are eliminated from the definition of the right-hand side in (2.13), so the values at $x = 0$ are not relevant. In other words, Filippov's inclusion ignores what may happen on the switching surface S, here equal to $x = 0$. One may consider the Carathéodory solutions of (2.21). When $x_0 = 0$ then one may have $\dot{x}(0) = 2$ (or $\dot{x}(0) = -2$). Then $x(t)$ is positive (or negative) in a right neighborhood of $t = 0$, and $\dot{x}(t)$ jumps to 1 (or -1) so that $x(t) = t$ (or $x(t) = -t$) for $t > 0$. Carathéodory and Filippov's solutions are the same, except that their derivatives may differ on a set of zero measure (in $\mathrm{I\!R}$). This is not important as absolutely continuous functions have a derivative that is defined up to a set of zero measure. Consider now the following definition of the right-hand side of the inclusion, which is sometimes proposed (Smirnov, 2002):

$$\dot{x}(t) \in F(x(t)) = \bigcap_{\varepsilon>0} \overline{\mathrm{conv}} f(x(t) + \varepsilon B_n)\,, \tag{2.23}$$

where B_n is the unit ball of $\mathrm{I\!R}^n$. This definition does not exclude sets of measure zero. At $x = 0$ we get $F(\varepsilon B_n) = \{-2,-1,0,1,2\}$ for all $\varepsilon > 0$, and $\overline{\mathrm{conv}}\, F(\varepsilon B_n) = [-2,2]$. The system in (2.21) then becomes the inclusion

$$\dot{x}(t) = \begin{cases} -1 & \text{if } x(t) < 0 \\ [-2,2] & \text{if } x(t) = 0 \\ 1 & \text{if } x(t) > 0\,. \end{cases} \tag{2.24}$$

Similar comments may be done concerning the solutions of (2.24) and the Filippov's solutions. In mechanics, the choice between (2.22) and (2.24) corresponds either to neglecting the static coefficient of dry friction (that is usually larger than the dynamic coefficient of friction) or to modeling it. Existence of solutions is assured for similar reasons as in (2.20). However, physical motivations may oblige one to add extra information in the system. For instance in mechanics, a sliding system with a nonzero initial velocity may have the velocity that converges to zero in finite time. The coefficient of friction when the contact force enters the Coulomb's friction cone is equal to the dynamic coefficient (here it would be 1). This means that the trajectory "ignores" the portion $(1,2]$ of the graph of the set-valued function in (2.24). When the trajectory comes from velocity zero (sticking mode) and goes to a sliding mode, the coefficient of friction reaches the value 2. This sort of effect may be properly modeled with a hysteresis behavior, adding a memory to the dynamics.

Example 2.21. As a further illustration of the different notions of a solution that may exist for a system, consider

$$x^{(3)}(t) = -\mathrm{sgn}(x(t))\ . \tag{2.25}$$

Then as shown in Pogromsky et al. (2003) there are infinitely many solutions to the Filippov's inclusion that start from $x(0) = \dot{x}(0) = \ddot{x}(0) = 0$. In fact all these solutions emerge from the origin of the state space with an accumulation of switches of the sign function (i.e., of crossing points of $x = 0$), that is a right-accumulation at $t = 0$.[2] On the contrary there is a unique locally analytic solution (sometimes called a forward solution in the computer science literature) on some interval of the form $[0, \varepsilon)$.

The conclusion of this section is that the way one embeds discontinuous ODEs into differential inclusions is not unique and may yield various results in terms of the obtained solutions. Sometimes the choice may be guided by physical considerations. We have insisted here more on Filippov's inclusions as this concept seems to be the more used, especially in control theory and applications with the so-called *variable structure systems*. The question that comes next is: does there correspond a specific way to numerically approximate each of these different notions? Specific algorithms for simulating Filippov's systems are described in Sect. 7.1 and in Sect. 9.3. Simulating (2.24) may lead one to resort to other types of dynamics like hysteresis effects, which will not be tackled in this book (see Hui & Zhu, 1995 for an event-driven algorithm). Consider Example 2.17. There is little chance that Carathéodory solutions may be computed correctly, to say nothing of Example 2.18. In both cases, however, Filippov's solutions are easily obtained.

2.1.2.3 Links with Subdifferentiation and Some Properties

One issue one faces when seeing (2.13) is: how may this set be computed in practice? In simple cases, several of the above examples have shown how this may be done, in a rather automatic way. The results presented next show a close connection between the right-hand side in (2.13) and generalized gradients of functions which are not differentiable, but only subdifferentiable (in the sense of Clarke or of convex analysis). For ease of exposition let us denote the right-hand side of (2.13) as

$$\mathscr{F}[f](x),$$

where the $\mathscr{F}$ is for Filippov. The next theorem is taken from Paden & Sastry (1987) and concerns the calculus for $\mathscr{F}[f](x)$. Let $2^{\mathrm{I\!R}^n}$ denote the set of subsets of $\mathrm{I\!R}^n$. Then $\mathscr{F}$ maps $\{f \text{ Lebesgue integrable} \mid f : \mathrm{I\!R}^n \to \mathrm{I\!R}^m\}$ into $\{g \mid \mathrm{I\!R}^n \to 2^{\mathrm{I\!R}^m}\}$.[3]

[2] This is a right-accumulation because the accumulation is situated on the right of $t = 0$.

[3] Usually $m = n$.

Theorem 2.22. *The map $\mathscr{F}$ has the following properties:*

(i) Assume that $f : \mathrm{IR}^n \to \mathrm{IR}^m$ is locally bounded. Then there exists a set $N_f \subset \mathrm{IR}^n$ of zero Lebesgue measure, such that for all sets $N \subset \mathrm{IR}^n$ of zero Lebesgue measure, one has

$$\mathscr{F}[f](x) = \mathrm{conv}\{\lim f(x_i) \mid x_i \to x,\, x_i \notin N_f \cup N\}\ . \tag{2.26}$$

(ii) Assume that $f, g : \mathrm{IR}^n \to \mathrm{IR}^m$ are locally bounded. Then

$$\mathscr{F}[f+g](x) \subset \mathscr{F}[f](x) + \mathscr{F}[g](x)\ . \tag{2.27}$$

(iii) Assume that $f_j : \mathrm{IR}^n \to \mathrm{IR}^{m_j}$, $j \in \{1,2,..,n\}$ are locally bounded. Then

$$\mathscr{F}[\times_{j=1}^{n} f_j](x) \subset \times_{j=1}^{n} \mathscr{F}[f_j](x)\ , \tag{2.28}$$

where $\times$ denotes the Cartesian product.

(iv) Let $g : \mathrm{IR}^n \to \mathrm{IR}^m$ be continuously differentiable (i.e., $g(\cdot) \in C^1(\mathrm{IR}^n;\mathrm{IR}^m)$), rank$\frac{\partial g}{\partial x}(x) = n$, and $f : \mathrm{IR}^m \to \mathrm{IR}^p$ be locally bounded. Then

$$\mathscr{F}[f \circ g](x) = \mathscr{F}[f](g(x))\ . \tag{2.29}$$

(v) Let the matrix-valued function $g : \mathrm{IR}^n \to \mathrm{IR}^{p\times m}$ be continuous, and $f : \mathrm{IR}^n \to \mathrm{IR}^m$ be locally bounded. Then

$$\mathscr{F}[f \circ gf](x) = g(x)\mathscr{F}[f](x)\ , \tag{2.30}$$

where $gf(x) = g(x)f(x) \in \mathrm{IR}^p$.

(vi) Let the function $V : \mathrm{IR}^n \to \mathrm{IR}$ be locally Lipschitz continuous. Then

$$\mathscr{F}[\nabla V](x) = \partial V(x)\ , \tag{2.31}$$

where $\partial V(x)$ denotes the Clarke's generalized gradient of $V(\cdot)$ at x (see Definition A.2).

(vii) Let $f : \mathrm{IR}^n \to \mathrm{IR}^m$ be continuous. Then

$$\mathscr{F}[f](x) = \{f(x)\}\ . \tag{2.32}$$

In Paden & Sastry (1987) the closed-loop dynamics of a mechanical system controlled by a nonsmooth state feedback is computed thanks to the properties of Theorem 2.22. The expression in (2.26) is also useful to compute Filippov's solution of piecewise continuous systems. See Sects. 2.8.3 and 2.9.

2.1.3 The One-Sided Lipschitz Condition

This property that is useful to show uniqueness of solutions of DIs was introduced for stiff ODEs by Dekker & Verwer (1984) and Butcher (1987) and for DIs in Kastner-Maresch (1990–91) and Dontchev & Lempio (1992). It was already used by Filippov (1964) to prove the uniqueness of solutions for ODEs with discontinuous right-hand side. Let us provide a definition that may be found in Dontchev & Farkhi (1998).

Definition 2.23. *The set-valued map $F : \mathbb{R}^n \to 2^{\mathbb{R}^n} \setminus \emptyset$ where $F(t,x)$ is compact for all $x \in \mathbb{R}^n$ and all $t \geqslant 0$ is called one-sided Lipschitz continuous (OSLC) if there is an integrable function $L : \mathbb{R}^+ \to \mathbb{R}$ such that for every $x_1, x_2 \in \mathbb{R}^n$, for every $y_1 \in F(t,x_1)$, there exists $y_2 \in F(t,x_2)$ such that*

$$\langle x_1 - x_2, y_1 - y_2 \rangle \leqslant L(t)||x_1 - x_2||^2.$$

Let A be a nonempty compact set. If we introduce the support function $\sigma(x,A) = \max_{a \in A} \langle x, a \rangle$, an equivalent definition is

$$\sigma(x - y, F(t,x)) - \sigma(x - y, F(t,y)) \leqslant L(t)||x - y||^2.$$

It is noteworthy that $L(\cdot)$ may be constant, time-varying, positive, negative, or zero. We recall that here $\langle \cdot, \cdot \rangle$ simply means the inner product in $\mathbb{R}^n$, but the OSLC condition may also be formulated for other inner products. The next examples are taken from Dontchev & Farkhi (1998) and Lempio (1992).

Example 2.24. The set-valued map $F : x \mapsto \begin{cases} [0,1] & \text{if } x < 0 \\ \\ [-1,1] & \text{if } x \geqslant 0 \end{cases}$ is OSLC with $L \geqslant 0$. Indeed if $x_1 > 0$ and $x_2 > 0$ or if $x_1 < 0$ and $x_2 < 0$, one can always choose, whatever $y_1 \in F(x_1)$, a y_2 that is equal to y_1. Thus $(x_1 - x_2, y_1 - y_2) = 0$. Let $x_1 \geqslant 0$ and $x_2 < 0$. If $y_1 \geqslant 0$ it is always possible to choose $y_2 = y_1$ so that $(x_1 - x_2, y_1 - y_2) = 0$. If $y_1 < 0$, for instance $y_1 = -1$, we get $(x_1 - x_2, y_1 - y_2) = (x_1 - x_2)(-1 - y_2) \leqslant 0$. In case $x_1 < 0$ and $x_2 \geqslant 0$, one has $y_1 \in [0,1]$ so it is always possible to choose $y_2 = y_1$. However, this multifunction is not Lipschitz continuous. Indeed take $x_1 = -\varepsilon < 0$, $x_2 = \varepsilon$, and let $\varepsilon \to 0$. Since $F(x_1) = [0,1]$ and $F(x_2) = [-1,1]$, finding a bounded constant $l \geqslant 0$ such that (2.4) holds is not possible as $\varepsilon \to 0$.

Example 2.25. Let $F(t,x) = L(h(t) - x) + \dot{h}(t) + L - L\,\mathrm{sgn}(x)$, where $L > 0$, $x \in \mathbb{R}$, $t \in [0,2]$, and $h(t) = -\frac{4}{\pi}\arctan(t-1)$. Then $F(\cdot,\cdot)$ is OSLC with constant $-L$.

In Kastner-Maresch (1990–91) and Dontchev & Lempio (1992) the OSLC is defined by replacing "there exists $y_2 \in F(t,x_2)$" with "for all $y_2 \in F(t,x_2)$". Both definitions are not equivalent.[4] Let us call the OSLC where "there exists $y_2 \in F(t,x_2)$" is replaced by "for all $y_2 \in F(t,x_2)$" the uniform OSLC condition (UOSLC). Clearly UOSLC implies OSLC. But the converse is not true, so that the OSLC condition of Definition 2.23 extends the notions of UOSLC, Lipschitz continuity, dissipativity (monotonicity). Consider once again Example 2.24. Let $x_1 = \varepsilon > 0$ and $x_2 = -\varepsilon < 0$. Thus $(x_1 - x_2)(y_1 - y_2) = 2\varepsilon(y_1 - y_2)$. Let $y_1 = -1$, then $(x_1 - x_2)(y_1 - y_2) = -2\varepsilon(1 + y_2)$ and since $y_2 \geqslant 0$ we get $(x_1 - x_2)(y_1 - y_2) \leqslant 0$. Thus any constant $L \geqslant 0$ is suitable. Now take $y_1 = 1$ so that $(x_1 - x_2)(y_1 - y_2) = 2\varepsilon(1 - y_2)$. In the first definition we can choose $y_2 = 1$ so that $(x_1 - x_2)(y_1 - y_2) = 0$. Now in the second definition

[4] Note also that in Definition 2.23, the sets $F(t,x)$ are required to be compact, which is not the case for the second definition hereafter named UOSLC.

this must hold for all $y_2 \in [0,1]$. Let us try $y_2 = 0$. We get $(x_1 - x_2)(y_1 - y_2) = 2\varepsilon > 0$. Any negative L is impossible. Take $L \geqslant 0$. We get $2\varepsilon \leqslant 4L\varepsilon^2$. Thus $L \geqslant \frac{1}{2\varepsilon}$ that diverges as ε approaches 0. Thus this $F(\cdot)$ is not UOSLC, though it is OSLC.

Still another way to define OSLC is proposed in Dontchev (2002), for multivalued maps with $F(x)$ convex and compact for each x, and uses the one-sided directional derivative of the function $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto ||x_1 - x_2||$, with respect to $y_1 - y_2$, as

$$\lim_{h \to 0, h>0} \frac{1}{h} \left(||(x_1 - x_2) + h(y_1 - y_2)|| - ||x_1 - x_2||\right) \leqslant L||x_1 - x_2|| \,. \tag{2.33}$$

The left-hand side is equal to $\sup_{z \in \partial ||x_1 - x_2||} \langle z, y_1 - y_2 \rangle$ (proposition 1.3.2 in Goeleven et al., 2003a). When $x_1 \neq x_2$ we get that $||y_1 - y_2|| \leqslant L||x_1 - x_2||$. When $x_1 = x_2$ the subdifferential $\partial ||x_1 - x_2||$ is the unit ball B_n. Therefore the supremum is attained at $\frac{y_1 - y_2}{||y_1 - y_2||}$ and we get $||y_1 - y_2|| \leqslant 0$, i.e., $y_1 = y_2$. Using this inequality or using the above inequality does not make much difference.

Remark 2.26. Here we chose to name *uniform* OSLC those set-valued mappings that satisfy the OSLC inequality for every $y_1 \in F(t, x_1)$ and for every $y_2 \in F(t, x_2)$. In the literature on the subject, one usually calls OSLC what we call UOSLC *and* OSLC (hence a possible confusion), while UOSLC is reserved for time-varying OSLC maps where the OSLC property holds uniformly in t (Kastner-Maresch, 1990–91).

Example 2.27. All set-valued mappings that may be written as $F(t,x) = f(t,x) - \varphi(x)$, where $\varphi \colon \mathbb{R}^n \to \mathbb{R}^n$ is a multivalued monotone mapping and $f(t,x)$ is Lipschitz continuous, are UOSLC. The OSLC constant L is equal to $\max(0, \lambda)$, where λ is the Lipschitz constant of the function $f(\cdot, \cdot)$.

Example 2.28. Consider $F(x) = \text{sgn}(x)$, the set-valued sign function. For all x_1, x_2, and $y_1 \in F(x_1)$, $y_2 \in F(x_2)$, one has $\langle x_1 - x_2, y_1 - y_2 \rangle \geqslant 0$. Therefore the multifunction $-F(\cdot)$ satisfies $\langle x_1 - x_2, -y_1 + y_2 \rangle \leqslant 0$ and is UOSLC with constant $L = 0$ (this is consistent with Example 2.27 with $\varphi(x) = \partial |x|$). However, $F(\cdot)$ is not OSLC, hence not UOSLC. Indeed take $x_1 > 0$, $x_2 < 0$, so that $y_1 = 1$, $y_2 = -1$. We get $(x_1 - x_2)(y_1 - y_2) = 2(x_1 - x_2) > 0$. OSLC implies that $2(x_1 - x_2) \leqslant L(x_1 - x_2)^2$ for some L. A negative L is impossible, and a nonnegative L yields $L \geqslant \frac{2}{x_1 - x_2}$. As $x_1 - x_2$ approaches 0, L diverges to infinity.

Example 2.29. Let $F: \mathbb{R}^n \to \mathbb{R}^n, x \mapsto Ax + K$, where A is a constant matrix and K is a closed convex cone. Such a multifunction is called a *convex process*, as its graph is a convex cone that contains the origin (Rockafellar, 1970). It is even a strict closed convex process since $\text{dom}(F) = \mathbb{R}^n$. From theorem 2.12 in Smirnov (2002) it follows that $F(\cdot)$ is Lipschitzian, with Lipschitz constant equal to $\sup_{x \in B_n} \inf_{v \in F(x)} |v|$. Hence it is UOSLC.

Example 2.30. The mapping $x \mapsto -x^{\frac{1}{3}} + [-1, 1]$ is OSLC, but it is not UOSLC (Dontchev & Farkhi, 1998).

We therefore have: Lipschitz continuity ⇒ UOSLC ⇒ OSLC. The next lemma demonstrates the usefulness of the UOSLC condition (Kastner-Maresch, 1990–91).

Lemma 2.31. *Let $F(\cdot,\cdot)$ be UOSLC with constant L, and let $x_1 : [t_0,+\infty) \to \mathrm{IR}^n$, $x_2 : [t_0,+\infty) \to \mathrm{IR}^n$ be two absolutely continuous solutions of the DI: $\dot{x}(t) \in F(t,x(t))$, i.e., $\dot{x}_1(t) \in F(t,x_1(t))$ and $\dot{x}_2(t) \in F(t,x_2(t))$ almost everywhere on $[t_0,+\infty)$. Then*

$$||x_1(t) - x_2(t)|| \leqslant \exp(L(t-t_0))\,||x_1(t_0) - x_2(t_0)|| \tag{2.34}$$

for all $t \geqslant t_0$. In particular, the DI: $\dot{x}(t) \in F(t,x(t))$ enjoys the uniqueness of solutions property.

As one may expect, uniqueness is also important from the numerical point of view. If uniqueness fails, then one can only expect that the sequence of approximated solutions (or a subsequence of it) converges towards *some* solution of the DI. Results similar to Lemma 2.31 can be obtained with the OSLC condition of Definition 2.23, see theorem 3.2 and corollary 3.3 in Dontchev & Farkhi (1998). From the numerical point of view, error estimates for the Euler scheme for inclusions with OSLC right-hand sides can be calculated. Moreover the set-valued Euler iterations approximate invariant attracting sets when $L < 0$. Variants of Theorem 2.9 are proposed in Dontchev & Farkhi (1998) for right-hand sides which satisfy an OSLC condition.

2.1.4 Recapitulation of the Main Properties of DIs

The results that have been presented in the foregoing sections are a rapid overview of all the available results for differential inclusions with absolutely continuous solutions (sometimes called ordinary DIs). Let us recapitulate some of the properties of the set $F(t,x)$ that are the most encountered in the mathematical literature on existence and uniqueness of solutions:

- Convexity: see, e.g., Filippov's convexification and Lemmas 2.13 and 2.8. See also examples in Sect. 5.2 of Deimling (1992).
- Compactness.
- Upper (or outer) semi-continuity: see Lemma 2.13.
- Lower semi-continuity: see Chap. 6 in Deimling (1992) for existence of solutions.
- Boundedness: there exists $a > 0$ such that $F(x) \subset aB_n$ for all $x \in \mathrm{IR}^n$; see for instance Sect. 4.3 in Smirnov (2002).
- Lipschitz continuous, see (2.4); another, equivalent formulation using the Hausdorff distance between sets (see Definition A.1) is as follows. The set-valued map $F : \mathrm{IR}^n \to 2^{\mathrm{IR}^n} \setminus \emptyset$ is Lipschitz continuous if for all x_1 and x_2 there exists a constant $l \geqslant 0$ such that $d_{\mathrm{H}}(F(x_1),F(x_2)) \leqslant l||x_1 - x_2||$. This is used in Lemma 2.8.
- One-sided Lipschitz continuity: see Lemma 2.31.
- Maximal monotone: a multivalued set $F(\cdot)$ is said to be monotone if for all $x_1, x_2 \in \mathrm{dom}(F)$ and all $y_1 \in F(x_1)$, $y_2 \in F(x_2)$, one has

$$\langle x_1 - x_2, y_1 - y_2 \rangle \geqslant 0.$$

One sees that if $F(\cdot)$ is one-sided Lipschitz continuous as in Definition 2.23, then $-F(\cdot)+L(t)$ is monotone for each t. On the other hand, if $F(\cdot)$ is monotone, then $-F(\cdot)$ is OSLC with constant $L=0$. This is the case of the sign multifunction: it is monotone and $-\mathrm{sgn}(\cdot)$ is OSLC. But notice that sign$(\cdot)$ is not OSLC, see Example 2.28.

Maximality means that the graph of the set-valued mapping contains the graphs of all the possible monotone mappings obtained from $F(\cdot)$ by cutting some pieces of its graph. Said differently, it is impossible to add new pieces to the graph of $F(\cdot)$ without destroying the monotonicity. For all $x,y \in \mathrm{I\!R}^n$ that satisfy $\langle y - F(z), x-z\rangle \geqslant 0$ for all $z \in \mathrm{dom}(F)$, then $y \in F(x)$. Maximality therefore refers to the inclusion of graphs. All the mappings defined as the subdifferential of some convex function $\varphi(\cdot)$ are maximal monotone (including functions whose domain is not the whole of $\mathrm{I\!R}^n$ like the indicator function $\psi_K(\cdot)$ of a convex set $K \subset \mathrm{I\!R}^n$). See Sect. 2.2 and Theorem 2.41.

- Hypomonotone: the map satisfies

$$\langle x_1 - x_2, y_1 - y_2\rangle \geqslant -l||x_1 - x_2||^2$$

for some $l \geqslant 0$ and all $x_1, x_2 \in \mathrm{dom}(F)$. A one-sided Lipschitz-continuous map with $L(t)$ constant and nonnegative is such that $-F(\cdot)$ is hypomonotone. Clearly hypomonotonicity with $l=0$ is monotonicity.

- Closed: see for instance Filippov's convexification.
- The linear growth condition: there exists an integrable function $\lambda : \mathrm{I\!R}^+ \to \mathrm{I\!R}^+$ such that $||F(t,x)|| \leqslant \lambda(t)(1+||x||)$ for all $x \in \mathrm{I\!R}^n$ and almost all $t \geqslant 0$. This is used in many results to guarantee the global existence of solutions (on $\mathrm{I\!R}^+$).
- Nonemptiness: a basic requirement that avoids situations such as $\dot{x}(t) \in \emptyset$ for all $x(0)$.
- Continuity: the mapping $F(\cdot)$ is called continuous if it is continuous with respect to the Hausdorff distance. There are mappings with closed values which are continuous but not upper semi-continuous and mappings with closed values which are lower and upper semi-continuous but not continuous (Deimling, 1992).

There are many more properties one may encounter in the literature on DIs (see for instance the index of Deimling, 1992): α-condensing, almost continuous, almost lower semi-continuous, almost upper semi-continuous, hyperaccretive, nonexpansive, homogeneous, one-sided contractive, strengthened expansive, relaxed accretive, and so on. Finding one's way through this dense forest is not always an easy task for the nonspecialist. [5]

Let us notice an important point that is sometimes (most often) not clearly stated. It is not because the uniform one-sided Lipschitz-continuous (UOSLC) condition is by itself more general than the maximal monotone (MM) condition (because all UOSLC maps $F(\cdot)$ are such that $-F(\cdot)+L\cdot$ is monotone) that all the results based

[5] By nonspecialist, one usually means someone who has not spent at least 3 years of PhD plus 1 year of post-doc, working 40 h per week on differential inclusions. This makes a lot of people, indeed.

on the OSLC contain the results based on the MM. For instance theorem 3.2 in Dontchev & Farkhi (1998) assumes compactness of $F(x)$ for all x and the linear growth condition. Thus inclusions in normal cones are excluded.

Example 2.32 (The relay function). This is the multifunction $\mathrm{sgn}(x)$ as in (2.20). It has convex, compact, closed, bounded values $F(x)$, it is upper semi-continuous, maximal monotone, satisfies a linear growth condition, but it is not Lipschitz continuous.

Example 2.33 (A normal cone). Consider $F(x) = -N_C(f(x))$ where $C \subset \mathbb{R}^n$ is a convex nonempty set, and $N_C(\cdot)$ is the normal cone of convex analysis. Take $C = \mathbb{R}^+$ and $f(x) = -|x| - 1$. Then clearly $N_C(f(x)) = \emptyset$. Now let $f(x) = x$. Then $F(x)$ is not compact for $x = 0$, does not satisfy a linear growth condition, is not Lipschitz continuous, but it is maximal monotone (the graph of $-N_C(\cdot)$ is in Fig. 1.1b).

Remark 2.34. One should not confuse the convexity of the graph of the set-valued function and the convexity of the values $F(x)$ for each x. These are totally decoupled notions. The graph of the set-valued sign function is not convex, but each set $\mathrm{sgn}(x)$ is convex (either a singleton or the interval $[-1,1]$). The graph of $F(x) = [-1,1]$ for all x is convex, as well as that of $F(x) = [-x,x]$ for $x \geqslant 0$ and $F(x) = [x,-x]$ for $x \leqslant 0$.

2.1.5 Some Hints About Uniqueness of Solutions

It has already been seen that the one-sided Lipschitz condition assures the uniqueness of solutions. Suppose that the DI in (2.21) has two solutions $x_1(\cdot)$ and $x_2(\cdot)$ defined on $\mathbb{R}^+$, with initial data $x_1(0)$ and $x_2(0)$. Suppose that $\langle x_1 - x_2, y_1 - y_2\rangle \leqslant 0$ for all x_1, $x_2 \in \mathbb{R}^n$, and all $y_1 \in F(t,x_1)$, $y_2 \in F(t,x_2)$: this means that the set-valued mapping $-F(t,\cdot)$ is monotone for each t. Then we obtain

$$\begin{aligned}\int_0^t \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}||x_1(s) - x_2(s)||^2\right)\mathrm{d}s &= \tfrac{1}{2}||x_1(t) - x_2(t)||^2 - \tfrac{1}{2}||x_1(0) - x_2(0)||^2 \\ &= \textstyle\int_0^t \langle x_1(s) - x_2(s), \dot{x}_1(s) - \dot{x}_2(s)\rangle \mathrm{d}s \qquad (2.35) \\ &= \textstyle\int_0^t \langle x_1(s) - x_2(s), y_1(s) - y_2(s)\rangle \mathrm{d}s\end{aligned}$$

with $y_1(s) \in F(s,x_1(s))$ and $y_2(s) \in F(s,x_2(s))$. Thus we get

$$\frac{1}{2}||x_1(t) - x_2(t)||^2 - \frac{1}{2}||x_1(0) - x_2(0)||^2 \leqslant 0\,, \tag{2.36}$$

which implies that $||x_1(t) - x_2(t)|| \leqslant ||x_1(0) - x_2(0)||$ for all $t \geqslant 0$. Suppose now that the mapping $-F(t,\cdot)$ is hypomonotone for each t and some constant $l \geqslant 0$. Then the mapping $x \mapsto -F(t,x) + lx$ is monotone for each t. We obtain

$$
\begin{aligned}
\tfrac{1}{2}||x_1(t)-x_2(t)||^2-\tfrac{1}{2}||x_1(0)-x_2(0)||^2 &= \textstyle\int_0^t \langle x_1(s)-x_2(s), y_1(s)-y_2(s)\rangle \mathrm{d}s \\
&= \textstyle\int_0^t (\langle x_1(s)-x_2(s), y_1(s)-lx_1(s) \\
&\quad -y_2(s)+lx_2(s)\rangle + l||x_1(s)-x_2(s)||^2)\,\mathrm{d}s \\
&\leqslant l\ \textstyle\int_0^t ||x_1(s)-x_2(s)||^2\mathrm{d}s\,.
\end{aligned}
\tag{2.37}
$$

Let us denote $f(t)=\frac{1}{2}||x_1(t)-x_2(t)||^2$. We can rewrite (2.37) as

$$
f(t)-f(0)\leqslant 2l\int_0^t f(s)\mathrm{d}s\,. \tag{2.38}
$$

Using Grownwall's Lemma C.7, it follows that

$$
f(t)\leqslant f(0)\exp(2lt)\,. \tag{2.39}
$$

Therefore from (2.39) we deduce that in case $x_1(0)=x_2(0)$, then $x_1(\cdot)=x_2(\cdot)$. The proof of Lemma 2.31 is similar (Kastner-Maresch, 1990–91).

Remark 2.35. It happens that some properties of the right-hand side are necessary to prove the *existence* of solutions (Lipschitz continuity, convexity of $F(x)$, upper or lower semi-continuity, maximality, etc.), while others are sufficient to guarantee the *uniqueness* of solutions (monotonicity, hypomonotonicity, OSLC).

2.2 Moreau's Sweeping Process and Unilateral DIs

For our purpose on unilateral dynamics and the targeted applications, the following particular classes of DI are of great interest: Moreau's sweeping process and unilateral DIs. Let us consider for instance a closed nonempty convex set $K\subset \mathbb{R}^n$. The normal cone to K may be defined as follows:

$$
N_K(x)\overset{\Delta}{=}\{s\in \mathbb{R}^n : \langle s, y-x\rangle \leqslant 0,\ \text{for all } y\in K\}
$$

while the tangent cone is the polar of the normal cone, which means

$$
T_K(x)\overset{\Delta}{=}[N_K(x)]^\circ \overset{\Delta}{=}\{d\in \mathbb{R}^n : \langle s, d\rangle \leqslant 0,\ \text{for all } s\in N_K(x)\}.
$$

If $x\notin K$, we set $N_K(x)=T_K(x)=\emptyset$. It is also possible to start from the definition of tangent cones and then define the normal cone (Hiriart-Urruty & Lemaréchal, 2001).

2.2.1 Moreau's Sweeping Process

Let us start with a DI which is of particular interest for our applications, the so-called first-order Moreau's sweeping process (Fig. 2.2 Moreau, 1971, 1972, 1977).

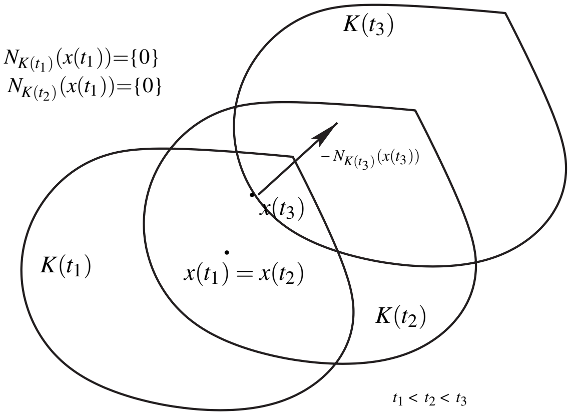


Fig. 2.2. The first-order sweeping process

Definition 2.36 (First-order Moreau's sweeping process). *Moreau's sweeping process (of first order) is defined by the following DI:*

$$\begin{cases} -\dot{x}(t) \in N_{K(t)}(x(t)) \;\; t \in [0,T] \\ \\ x(0) = x_0 \in K(0) \end{cases}, \tag{2.40}$$

where $K(t)$ is a moving closed and nonempty convex set, $T > 0$.

This terminology is explained by the fact that $x(t)$ can be viewed as a point which is swept by a moving convex set. As shown in Moreau (1977), under some mild conditions on the mapping $t \mapsto K(t)$, a solution to (2.40) possesses right and left limits, and $x(t) = \text{proj}[K(t);x(t^-)]$, $x(t^+) = \text{proj}[K(t^+);x(t)]$, for all $t \in (0,T)$.

A solution $x(\cdot)$ for such a DI is assumed to be differentiable almost everywhere satisfying (2.40) and the inclusion $x(t) \in K(t), t \in [0,T]$. In simple cases, the set-valued application $t \mapsto K(t)$ is supposed to be Lipschitz continuous, i.e.,

$$\exists\, l \geqslant 0, \quad d_{\mathrm{H}}(K(t),K(s)) \leqslant l|t-s|\,, \tag{2.41}$$

where d_{H} is the Hausdorff distance between two closed sets (see Definition A.1 and recall that (2.41) is equivalent to (2.4)). For instance, let $g : [0,T] \to \mathbb{R}^n$ be a Lipschitz-continuous function, the set-valued map $K(t) = C + g(t)$ given by a translation of a constant convex set C satisfies (2.41). It is noteworthy that the convex set can also change in shape. When the set $K(t)$ satisfies (2.41), it is possible to prove the existence of a solution which is Lipschitz continuous with the same constant l and the uniqueness in the class of absolutely continuous functions. The following theorem states the existence, the uniqueness, the dependence on initial data, and the dependence on the moving set of solutions.

Theorem 2.37. *Suppose that the mapping $t \mapsto K(t)$ is Lipschitz continuous in the Hausdorff distance with constant l and $K(t)$ is nonempty, closed, and convex for every $t \in [0,T]$. Let $x_0 \in K(0)$. Then there exists a solution $x : [0,T] \to \mathrm{I\!R}^n$ of the DI in (2.40) which is Lipschitz continuous with constant l. In particular $||\dot{x}(t)|| \leqslant l$ for almost every $t \in (0,T)$. Moreover the solution is unique in the class of absolutely continuous functions. Next, if $x_1(\cdot)$ and $x_2(\cdot)$ are two solutions with $x_1(0) = x_{10}$ and $x_2(0) = x_{20}$, one has $||x_1(t) - x_2(t)|| \leqslant ||x_{10} - x_{20}||$. Finally, let $t \mapsto K(t)$ and $t \mapsto C(t)$ be two moving nonempty, closed convex sets with Lipschitz constants l and c, respectively. Let $x(\cdot)$ denote the solution of the sweeping process with $K(t)$, and $z(\cdot)$ the solution with $C(t)$. Then*

$$||x(t) - z(t)||^2 \leqslant ||x_0 - z_0||^2 + 2(c+l) \int_0^t d_{\mathrm{H}}(C(s), K(s))\mathrm{d}s \,. \tag{2.42}$$

Remark 2.38. The Lipschitz sweeping process is the sweeping process with a Lipschitz-continuous set-valued mapping $t \mapsto K(t)$. This is not to be confused with the DIs with Lipschitz-continuous right-hand sides in Sect. 2.1.1. Indeed let t be fixed. Then Lipschitz continuity in x means that there exists a bounded constant l such that for all $x_1, x_2 \in \mathrm{I\!R}^n$ one has $N_{K(t)}(x_1) \subset N_{K(t)}(x_2) + l||x_1 - x_2||B_n$. Consider $n = 1$ and $K(t) = K = \mathrm{I\!R}^+$. When $x_1 = 0$ one has $N_K(x_1) = \mathrm{I\!R}^-$, and when $x_2 \neq 0$ one has $N_K(x_2) = \{0\}$. Letting x_2 approach 0 one sees that the inclusion of sets cannot be satisfied with a bounded l. The Lipschitz sweeping process is not a Lipschitzian DI in the sense of Sect. 2.1.1.

Many extensions have been studied. To cite a few of them, we refer to Kunze & Monteiro Marqués (2000), Castaing et al. (1993), Benabdellah et al. (1996), Castaing & Monteiro-Marques (1996) and Thibault (2003). For instance, the state-dependent sweeping process

$$\begin{cases} -\dot{x}(t) \in N_{K(t,x(t))}(x(t)) \;\; t \in [0,T] \\ \\ x(0) = x_0 \in K(0) \end{cases} \tag{2.43}$$

has been studied in Kunze & Monteiro Marques (1998). One of these extensions which is of utmost importance is the RCBV (right-continuous and of bounded variation) sweeping process:

$$\begin{cases} -\mathrm{d}x \in N_{K(t)}(x(t)) \;\; (t \geqslant 0) \\ \\ \;\; x(0) = x_0 \end{cases} , \tag{2.44}$$

where the solution $x(\cdot)$ is searched as a function of bounded variations (BV). The measure $\mathrm{d}x$ associated with the BV function x is called a differential measure or a Stieltjes measure, see Definition C.4. We will come back later on what is the meaning of the inclusion of a measure into a cone.

The Lipschitz assumption (2.41) is no longer made on the convex set $K(t)$ but is replaced by

$$d_{\mathrm{H}}(K(t),K(s)) \leqslant r(t)-r(s) \tag{2.45}$$

for some right-continuous nondecreasing function $r:[0,T] \to \mathrm{I\!R}$. A moving set $K(t)$ that satisfies such a condition is called RCBV (see Sect. C.2).

Theorem 2.39. *Let $t \mapsto K(t)$ be RCBV, such that every $K(t) \subset \mathrm{I\!R}^n$ is nonempty, closed, and convex. Let $x_0 \in K(0)$. Then (2.44) has a unique RCBV solution.*

Both Theorems 2.37 and 2.39 have been proved in Moreau (1977) and Monteiro Marques (1993) (theorem 1.5 in this book). Intuitively, when the moving set $K(t)$ is BV, then it may have discontinuities and may jump from one position $K(t^-)$ to another position $K(t^+)$ such that $K(t^+)$ does not contain $x(t^-)$. Then the state $x(\cdot)$ also has to jump so that $x(t^+) \in K(t^+)$. We conclude that the set of instants at which $x(\cdot)$ jumps is contained in (possibly equal to) the set of instants at which $K(\cdot)$ jumps. Said differently, the atoms of the differential measure dx must be times where $K(\cdot)$ is discontinuous. Obviously if $K(\cdot)$ jumps at t but $x(t)$ stays inside $K(t^+)$ then $x(\cdot)$ is continuous at t.

An important class of sweeping processes are the so-called *perturbed sweeping processes*. They are inclusions of the form

$$\begin{cases} -\mathrm{d}x + f(x(t),t)\mathrm{d}t \in N_{K(t)}(x(t)) \quad (t \geqslant 0) \\ \\ x(0) = x_0 \end{cases} \tag{2.46}$$

for some vector field $f(\cdot,\cdot)$. The inclusion is written in (2.46) in terms of measures, to allow the possibility for the solution to possess jumps. The well-posedness of such inclusions has been studied, with various assumptions on $f(\cdot,\cdot)$ and $K(t)$, in Benabdellah et al. (1997), Edmond & Thibault (2005, 2006) and Brogliato & Thibault (2006), to cite a few. In particular, some nonsmooth electrical circuits can be written as perturbed sweeping processes Brogliato & Thibault, 2006, so that this class of DIs has a particular importance in applications. We almost showed it on a particular case in Sect. 1.1.5 (in the treated example of circuit **(c)**, the set K is constant, but it could be easily rendered time-varying by adding a current source in the circuit).

2.2.2 Unilateral DIs and Maximal Monotone Operators

Definition 2.40 (Unilateral differential inclusion). *A unilateral differential inclusion (UDI) is defined as*

$$-(\dot{x}(t) + f(x(t)) + g(t)) \in N_K(x(t)), \; x(0) = x_0 \in K \,, \tag{2.47}$$

where the closed convex nonempty set $K \subset \mathrm{I\!R}^n$ is the feasible set and $g:\mathrm{I\!R}^+ \to \mathrm{I\!R}^n$ and $f:\mathrm{I\!R}^n \to \mathrm{I\!R}^n$.

A solution of such a DI is understood as an absolutely continuous function $t \mapsto x(t)$. It is noteworthy that any solution of such a DI must by construction lie in the convex set K. The link with maximal monotone operators is strong. Let us consider the following inclusion:

$$\dot{x}(t) + A(x(t)) + g(t) \ni 0, \;\; x(0) = x_0 \in \mathrm{dom}(A)\,, \tag{2.48}$$

where $A(\cdot)$ is a maximal monotone operator and $g(\cdot)$ an absolutely continuous function of time. If we consider an operator $f(\cdot)$ which is monotone and Lipschitz continuous, then the operator

$$A(x(t)) = f(x(t)) + N_K(x(t)) \tag{2.49}$$

is maximal monotone. Results for such differential inclusions abound in the literature, see for instance Brezis (1973) and Goeleven et al. (2003a). Let us state the following that is a generalized version of the Hille–Yosida theorem.

Theorem 2.41. *Consider the differential inclusion (2.48), where $A : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ is maximal monotone and $g : \mathbb{R}^+ \to \mathbb{R}^n$ is absolutely continuous. Then there exists a unique Lipschitz-continuous solution $x(\cdot)$ satisfying the inclusion almost everywhere.*

There exist so many variations of Theorem 2.41 that writing all of them would bring us much too far from the topic of this book, is numerical simulation. The interested reader may have a look at Brezis (1973), Goeleven et al. (2003a,b), Aubin & Cellina (1984), Adly & Goeleven (2004) and references therein. Recall that maximal monotonicity implies upper semi-continuity, as we stated in Sect. 2.1.2. This does not mean at all that Theorem 2.41 can be deduced from Lemma 2.13. Indeed many maximal monotone operators do not satisfy the conditions of Lemma 2.13.

2.2.3 Equivalence Between UDIs and other Formalisms

In Brogliato et al. (2006), a result of equivalence in terms of solutions has been given between the implicit differential inclusion

$$-(\dot{x}(t) + f(x(t)) + g(t)) \in N_{\mathrm{T}_K(x(t))}(\dot{x}(t)) \tag{2.50}$$

and (2.47), provided that the UDI (2.47) has the so-called slow solution, that is $\dot{x}(t)$ is of minimal norm in $N_K(x(t)) + f(x,t) + g(t)$. We will see that an inclusion into a tangent cone taken at $\dot{x}(\cdot)$ will also be introduced in second-order dynamics.

Special Case when K is Finitely Represented

Let the nonempty closed convex set K be finitely represented such that

$$K = \{x \in \mathbb{R}^n \mid h(x) \leqslant 0\} \tag{2.51}$$

where the function $h(x) = (h_1(x), \ldots, h_m(x))^{\mathrm{T}}$ and the functions $h_i(x) : \mathbb{R}^n \to \mathbb{R}$ are assumed to be continuously differentiable with gradients denoted by $\nabla h_i(x)$. In this case, two other definitions of the normal and the tangent cones can be written. For $x \in K$, we denote by

$$I(x) = \{i \in \{i, \ldots, m\} \mid h_i(x) = 0\} \tag{2.52}$$

the set of active constraints at x. The tangent cone can be defined by

$$T^h(x) = \{s \in \mathbb{R}^n \mid \langle \nabla h_i(x), s\rangle \leqslant 0, i \in I(x)\} \tag{2.53}$$

and the normal cone as its polar cone, i.e.,

$$N^h(x) := [T^h(x)]^\circ = \Big\{ \sum_{i \in I(x)} \lambda_i \nabla h_i(x), \lambda_i \geqslant 0, i \in I(x) \Big\} . \tag{2.54}$$

It is always true that $N_K(x) \supset N^h(x)$ and $T_K(x) \subset T^h(x)$. A key assumption to guarantee that $N_K = N^h$ and equivalently $T_K = T^h$ is commonly called a constraints qualification condition in convex analysis and optimization theory. For instance, the so-called *Mangasarian-Fromowitz* assumption

$$\left.\begin{array}{l} \sum_{i \in I(x)} \lambda_i \nabla h_i(x) = 0 \\ \text{with } \lambda_i \geqslant 0 ,\ i \in I(x) \end{array}\right\} \quad \Longrightarrow \quad \lambda_i = 0 ,\ i \in I(x) \tag{QC.1}$$

is a kind of constraints qualification which ensures $N_K = N^h$ and $T_K = T^h$. Such conditions hold in particular if the gradients of the active constraints at x are linearly independent. When the $h_i(\cdot)$, $1 \leqslant i \leqslant m$, are convex, it can be seen that (QC.1) is equivalent to the so-called *Slater* assumption:

$$\exists\, \bar{x} \in \mathbb{R}^n : h_i(\bar{x}) < 0 , i = 1, \ldots, m. \tag{QC.2}$$

Let us now introduce the following DI :

$$-(\dot{x}(t) + f(x(t)) + g(t)) \in N^h(x(t)) . \tag{2.55}$$

Obviously, this last UDI is equivalent to the DI (2.47) if and only if $N_K = N^h$. It is also equivalent to the following dynamical system which is a particular type of dynamical complementarity systems, see Sect. 2.6,

$$\begin{cases} -\dot{x}(t) = f(x(t)) + g(t) + \nabla h(x(t)) \lambda(t) \\ 0 \leqslant -h(x(t)) \perp \lambda(t) \geqslant 0 , \end{cases} \tag{2.56}$$

where $\lambda(\cdot)$ is assumed to be measurable.

2.3 Evolution Variational Inequalities

Let us first recall what a variational inequality (VI) is. Given a nonempty closed convex set $K \subset \mathbb{R}^n$, a VI is defined as: find $x \in K$ such that

$$\langle F(x), y - x \rangle \geqslant 0, \ \forall y \in K , \tag{2.57}$$

where $F : \mathbb{R}^n \to \mathbb{R}^n$. From the definition of a normal cone, this definition can be rewritten as the inclusion

$$-F(x) \in N_K(x) . \tag{2.58}$$

More details on this problem can be found in Sect. 12.6. The link between (2.58) and subdifferentiation of convex analysis is clear, as (2.58) means that $-F(x)$ is a subgradient of the indicator function of the set K. This is generalized to any proper convex function $\phi(\cdot)$, see Appendix A.

Following the definition of a VI, the evolution variational inequalities are defined as follows.

Definition 2.42 (Evolution variational inequalities). *Given a nonempty closed convex set $K \subset \mathbb{R}^n$, an evolution variational inequality (EVI) is defined as: find $x(\cdot) \in K$ such that*

$$\langle \dot{x}(t) + f(x(t)), y - x(t) \rangle \geqslant 0, \ \forall \ y \in K, \ x(0) = x_0 \in K , \tag{2.59}$$

which is equivalent to the following unilateral DI:

$$-(\dot{x}(t) + f(x(t))) \in N_K(x(t)), \ x(0) = x_0 \in K . \tag{2.60}$$

Again the equivalence holds because this is just a rewriting of the same object, the normal cone to K. Seminal studies of such VI and EVI have been carried in infinite-dimensional spaces, such as Hilbert spaces, see the work of Lions & Stampacchia (1967) and Kinderlehrer & Stampacchia (1980). The book of Goeleven et al. (2003a) gives a good overview of the various mathematical approaches for VIs and EVIs. In Harker & Pang (1990) and Facchinei & Pang (2003), the presentation is focused on finite-dimensional VI and the associated algorithms to solve such problems.

Some variants for the definition of the EVI can be found in the literature and the terminology is still not very well fixed. For instance, an EVI sometimes referred to as a parabolic VI is defined as follows: find $x : [0,T] \to \mathbb{R}^n$ such that

$$\langle \dot{x}(t), y - x(t) \rangle + a(x(t), y - x(t)) + \phi(y) - \phi(x(t)) \geqslant \langle f(t), y - x(t) \rangle \tag{2.61}$$

for all $y \in K$, where $a : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous, bilinear, and elliptic (i.e., $a(x,x) \geqslant \alpha \|x\|^2, \alpha > 0$), $f : [0,T] \to \mathbb{R}^n$ belongs to $\mathscr{L}^2$, and $\phi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is convex, proper, and lower semi-continuous. The above EVI can be further generalized by replacing the bilinear form $a(x,y)$ by $\langle F(x), y \rangle$ where $F(x)$ is a maximal monotone operator. In order to obtain the previous definition of EVI given by (2.59) it suffices to choose $\phi(\cdot)$ as the indicatrix $\psi_K(\cdot)$ of the set K, $f = 0$, and $a(x, y - x) = \langle f(x), y \rangle$.

Example 2.43. Let us consider the dynamics of a one degree-of-freedom system subject to Coulomb and viscous friction. It may be written as the inclusion

$$m\ddot{q}(t) + c\dot{q}(t) + kq(t) \in -\partial\varphi(\dot{q}(t)) \,, \tag{2.62}$$

where $c > 0$ is the damping coefficient, $k > 0$ is the stiffness of a spring acting on the mass, and $\varphi(\dot{q}) = \mu|\dot{q}|$, $\mu > 0$ is the dry friction coefficient. As we have already seen in the foregoing sections and chapter, one has $\partial\varphi(\dot{q}) = \mu\ \mathrm{sgn}(\dot{q})$, where $\mathrm{sgn}(\cdot)$ is the sign multifunction. The inclusion in (2.62) may be rewritten as the evolution variational inequality (see Appendix A.3)

$$\langle m\ddot{q}(t) + c\dot{q}(t) + kq(t), v - \dot{q}(t)\rangle + \varphi(v) - \varphi(\dot{q}(t)) \geqslant 0 \tag{2.63}$$

for all $v \in \mathrm{I\!R}$ and almost everywhere in $\mathrm{I\!R}^+$. Let us rewrite (2.62) as a first-order system

$$\dot{x}(t) + Ax(t) \in -\partial\Phi(x(t)) \tag{2.64}$$

with $x^{\mathrm{T}} = (x_1 n x_2) = (q, \dot{q})$ and A is easily calculable, the function $\Phi(x) = \frac{\mu}{\sqrt{m}}|x_2|$. Then (2.64) is equivalent to the variational inequality (see Appendix A)

$$\langle \dot{x}(t) + Ax(t), v - x(t)\rangle + \Phi(v) - \Phi(x(t)) \geqslant 0 \tag{2.65}$$

for all $v \in \mathrm{I\!R}^2$ and almost everywhere in $\mathrm{I\!R}^+$.

The link between EVIs, unilateral DIs, and maximal monotone operators is also strong. For instance, the existence and uniqueness theorem for maximal monotone operators holds for

$$\langle \dot{x}(t) + f(x(t)) + g(t), v - x\rangle \geqslant 0 \text{ for all } v \in K \,, \tag{2.66}$$

where K is nonempty closed convex. In Brogliato et al. (2006), an existence result is given for this last EVI under the assumption that $f(\cdot)$ is continuous and hypomonotone.

For $g \equiv 0$ and $f(x) = Ax$, the EVI is called a linear evolution variational inequality (LEVI) (Goeleven & Brogliato, 2004). To complete this description, let us define what is the so-called quasi-variational inequality (QVI) and the associated evolution quasi-variational inequality (EQVI). The QVI may be defined by finding the solution $x \in K(x)$ such that

$$\langle F(x), y - x\rangle \geqslant 0, \text{ for all } y \in K(x) \,. \tag{2.67}$$

The major discrepancy with the standard VI is the dependence of K on the variable x which leads to strong mathematical difficulties. In the same way, the EQVI may be defined by finding $x \in K(x)$ such that

$$\langle \dot{x}(t) + f(x(t)), y - x(t)\rangle \geqslant 0, \text{ for all } y \in K(x(t)) \,. \tag{2.68}$$

We will see in the sequel that the Coulomb's friction model with unilateral contact and the second-order Moreau's sweeping process enter into such a class of problems. There also exist so-called *hemi variational inequalities* that were introduced by P.G. Panagiotopoulos. Roughly speaking, VIs are related to the subdifferentiation of convex functions. Hemi VIs are related to subdifferentiation of locally Lipschitz continuous functions and Clarke's normal cone (see Definition A.2). We do not tackle hemi VIs in this book, see Goeleven et al. (2003a).

To end this part on VIs, let us state a well-posedness result taken from Goeleven & Brogliato (2004) that is inspired by corollary 2.2 of Goeleven et al. (2003b), which is itself an extension of Kato's theorem (Kato, 1970).

Theorem 2.44. *Let K be a nonempty closed convex subset of $\mathbb{R}^n$ and let $A \in \mathbb{R}^{n\times n}$ be a real matrix of order n. Suppose that $F : \mathbb{R}^n \to \mathbb{R}^n$ can be written as*

$$F(\cdot) = F_1(\cdot) + \Phi'(\cdot),$$

where $F_1(\cdot)$ is Lipschitz continuous and $\Phi(\cdot) \in C^1(\mathbb{R}^n;\mathbb{R})$ is convex. Let $t_0 \in \mathbb{R}$ and $x_0 \in K$ be given. Then there exists a unique $x \in C^0([t_0,+\infty);\mathbb{R}^n)$ such that

$$\frac{dx}{dt}(\cdot) \in \mathscr{L}^{\infty}_{\mathrm{loc}}([t_0,+\infty);\mathbb{R}^n)\ , \tag{2.69}$$

$$x(\cdot) \textit{ is right-differentiable on } [t_0,+\infty)\ , \tag{2.70}$$

$$x(t_0) = x_0\ , \tag{2.71}$$

$$x(t) \in K,\ t \geqslant t_0\ , \tag{2.72}$$

$$\langle \frac{dx}{dt}(t) + Ax(t) + F(x(t)), v - x(t)\rangle \geqslant 0,\ \forall v \in K,\ a.e.\ t \geqslant t_0\ . \tag{2.73}$$

The well-posedness results of Theorem 2.44 continue to hold for a controlled LEVI(A,B,K) defined as $\langle \frac{\mathrm{d}x}{\mathrm{d}t}(t) + Ax(t) + Bu(t) + F(x(t)), v - x(t)\rangle \geqslant 0,\ \forall v \in K$, with $B \in \mathbb{R}^{m\times n}$, $u \in C^0([t_0,+\infty);\mathbb{R}^m)$ and $\frac{\mathrm{d}u}{\mathrm{d}t} \in \mathscr{L}^1_{\mathrm{loc}}([t_0,+\infty);\mathbb{R}^m)$, see Goeleven et al. (2003a).

2.4 Differential Variational Inequalities

In Pang & Stewart (in press), a framework that is more general than EVIs is settled and the so-called differential variational inequalities are introduced.

Definition 2.45 (Differential variational inequalities). *A differential variational inequality (DVI) can be defined as follows:*

$$\begin{aligned} \dot{x}(t) &= f(t,x(t),\lambda(t)), && (2.74)\\ \lambda(t) &= \mathrm{SOL}(K,F(t,x(t),\cdot)), && (2.75)\\ 0 &= \Gamma(x(0),x(T))\ , && (2.76) \end{aligned}$$

where

- *$x : [0,T] \to \mathbb{R}^n$ is the differential trajectory (state variable),*

- $\lambda : [0,T] \to \mathbb{R}^m$ *is the algebraic trajectory (or Lagrange multiplier),*
- $f : [0,T] \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ *is the ODE right-hand side,*
- $F : [0,T] \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ *is the VI function,*
- K *is nonempty closed convex subset of* $\mathbb{R}^m$*,*
- $\Gamma : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ *is the boundary conditions function. For an initial value problem (IVP), the function is equal to* $\Gamma(x,y) = x - x_0$ *and for a linear boundary value problem (BVP) equal to* $\Gamma(x,y) = Mx + Ny - b$ *for some matrices* $M \in \mathbb{R}^{n\times n}$ *and* $N \in \mathbb{R}^{n\times n}$ *and a vector* $b \in \mathbb{R}^n$*.*

The notation $\lambda(t) = \mathrm{SOL}(K,\Phi)$ *means that* $\lambda(t) \in K$ *is the solution of the following VI:*

$$(v-\lambda)^{\mathrm{T}}\Phi(\lambda) \geqslant 0, \quad \forall v \in K\,. \tag{2.77}$$

A pair of functions (x,λ) is the solution in the sense of Carathéodory of the DVI if $x(\cdot)$ is absolutely continuous and $\lambda(\cdot)$ is (Lebesgue) integrable satisfying

$$x(t) = x(s) + \int_s^t f(\tau,x(\tau),\lambda(\tau))\,\mathrm{d}\tau, \quad \forall s \leqslant t \tag{2.78}$$

and for any continuous $\tilde{\lambda} : [0,T] \to K$ such that

$$\int_0^T (\tilde{\lambda}(t) - \lambda(t))^{\mathrm{T}} F(t,x(t),\lambda(t))\,\mathrm{d}t \geqslant 0\,. \tag{2.79}$$

The latter condition implies that $\lambda(t) \overset{\text{a.e.}}{=} \mathrm{SOL}(K, F(t,x(t),\cdot))$.

Special Cases of DVI

The DVI framework includes the following:

- Differential algebraic equations

$$\begin{cases} \dot{x}(t) = f(t,x(t),\lambda(t)) \\ \lambda(t) = F(t,x(t),\lambda(t))\,. \end{cases} \tag{2.80}$$

- Differential complementarity systems

$$\begin{cases} \dot{x}(t) = f(t,x(t),\lambda(t)) \\ K \ni \lambda(t) \perp F(t,x(t),\lambda(t)) \in K^*\,, \end{cases} \tag{2.81}$$

  where K and K^* are a pair of dual closed convex cones ($K^* = -K^\circ$, where K° is the polar cone). The LCSs are also a special case of DVI (see Sect. 2.6).
- Evolution variational inequalities

$$-(\dot{x}(t) + f(x(t))) \in N_K(x(t))\,. \tag{2.82}$$

- When K is a cone, the preceding EVI is equivalent to a differential CS of the type

$$\begin{cases} \dot{x}(t) + f(x(t)) = \lambda(t) \\ K \ni x(t) \perp \lambda(t) \in K^* \,. \end{cases} \tag{2.83}$$

- When K is finitely represented, i.e., $K = \{x \in \mathbb{R}^n \mid h(x) \leqslant 0\}$, then under some appropriate constraints qualifications, we obtain another dynamical CS which is often called a gradient complementarity system (see Sect. 4.1):

$$\begin{cases} \dot{x}(t) + f(x(t)) = -\nabla h(x(t))\lambda(t) \\ 0 \leqslant -h(x(t)) \perp \lambda(t) \geqslant 0 \,. \end{cases} \tag{2.84}$$

- Finally, if K is a closed convex and nonempty set then the EVI is equivalent to the following DVI:

$$\begin{cases} \dot{x}(t) + f(x(t)) = w(t) \\ 0 = x(t) - y(t) \\ y(t) \in K,\, (v - y(t))^{\mathrm{T}} w(t) \geqslant 0, \forall v \in K \,. \end{cases} \tag{2.85}$$

Example 2.46. The circuit (**c**) in (1.13) is a gradient CS with $g(x) = x_2$ (one may replace $\lambda(t)$ by $\lambda'(t) = \frac{1}{L}\lambda(t)$ in the complementarity condition without changing the system). Other examples are given in Pang & Stewart (in press).

Remark 2.47. Consider (2.84). Suppose the system evolves in the subspace $\{x \mid h(x) = 0\}$ on some nonzero closed time interval I. Then as we saw in Sect. 1.4 the complementarity may be rewritten as $0 \leqslant -\nabla h^{\mathrm{T}}(x(t))\dot{x}(t) \perp u(t) \geqslant 0$ for $t \in$Int(I). Replacing $\dot{x}(t)$ by its value one then gets $0 \leqslant -\nabla h^{\mathrm{T}}(x(t))(-f(x(t)) - \nabla h(x(t))u(t)) \perp u(t) \geqslant 0$, which is an LCP with matrix $\nabla h^{\mathrm{T}}(x(t))\nabla h(x(t)) \geqslant 0$.

2.5 Projected Dynamical Systems

An other type of NSDS are the so-called projected dynamical systems (PDS) which have been introduced in Dupuis & Nagurney (1993) and Nagurney & Zhang (1996), see also Henry (1973). Various definitions for PDS exist and are fortunately equivalent in most situations. To start with one of them, let us consider the following definition:

Definition 2.48 (Projected dynamical systems). *Let us consider a nonempty closed and convex subset K of $\mathbb{R}^n$. A projected dynamical system (PDS) is defined as*

$$\dot{x}(t) = \Pi_K\left(x(t); -(f(x(t)) + g(t))\right) \,, \tag{2.86}$$

where $\Pi_K : K \times \mathbb{R}^n \to \mathbb{R}^n$ is the operator

$$\Pi_K(x;v) = \lim_{\delta\downarrow 0} \frac{\mathrm{proj}_K(x+\delta v)-x}{\delta} \; . \tag{2.87}$$

Let us recall that the projection operator $\mathrm{proj}_K : \mathbb{R}^n \to K$ is defined as

$$\|\mathrm{proj}_K(z)-z\| = \inf_{y\in K} \|y-z\| \; . \tag{2.88}$$

The definition of the operator $\Pi_K(\cdot,\cdot)$ corresponds to the one-sided Gâteaux derivative of the projection operator for $x \in K$, i.e., when $\mathrm{proj}_K(x) = x$. A classical result of convex analysis (see for instance Hiriart-Urruty & Lemaréchal, 2001) states that

$$\Pi_K(x;v) = \mathrm{proj}_{T_K(x)}(v) \; , \tag{2.89}$$

where $T_K(x)$ is the tangent cone to K at x. Therefore, the PDS can be equivalently rewritten as

$$\dot{x}(t) = \mathrm{proj}_{T_K(x(t))}\left(-(f(x(t))+g(t))\right) \; . \tag{2.90}$$

In Brogliato et al. (2006), the PDS (2.90) is proved to be equivalent to the implicit UDI(2.50) and therefore to be equivalent to the UDI (2.47) if the slow solution is selected (similar equivalences were shown by Henry, 1973; Cornet, 1983). For results and definitions in infinite-dimensional spaces (Hilbert spaces), we refer to the works in Cojocaru (2002) and Cojocaru & Jonker (2003).

2.6 Dynamical Complementarity Systems

2.6.1 Generalities

In van der Schaft & Schumacher (2000), the class of dynamical complementarity systems is defined as

$$\begin{cases} f(\dot{x}(t),x(t),t,y(t),\lambda(t)) \\ \\ K^* \ni w(t) \perp \lambda(t) \in K \; , \end{cases} \tag{2.91}$$

where $K \subset \mathbb{R}^m$ is a closed nonempty convex cone and K^* its dual cone (identify $f(\cdot)$, $w(\cdot)$, C, and K^* in the dynamical CS in (1.64) and (1.65)). In this formulation, the dynamics is given in an implicit way as in a DAE. It is possible to provide a semi-explicit formulation in assigning to y the role of an output and to λ the role of an input of the dynamical system, i.e.,

$$\begin{cases} \dot{x}(t) = f(x(t),t,\lambda(t)) = 0 \\ \\ w(t) = h(x(t),\lambda(t)) \\ \\ K^* \ni w(t) \perp \lambda(t) \in K \end{cases} \; . \tag{2.92}$$

In order to be able to give well-posedness results, one has to focus on subclasses of (2.92), that is a much too large class of complementarity systems. We will see this in Chap. 4. In Heemels & Brogliato (2003), some other extensions can be found which include explicitly a control input $u(\cdot)$.

Example 2.49. Consider the following scalar system, with $u(\cdot)$ a measurable function:

$$\begin{cases} \dot{x}(t) = u(t) + \lambda(t) \\ 0 \leqslant w(t) = x(t) \perp \lambda(t) \geqslant 0 \\ x(0) \geqslant 0\,. \end{cases} \tag{2.93}$$

It is seen that when $x(t) > 0$ then $\lambda(t) = 0$ and $\dot{x}(t) = u(t)$. When $x(t) = 0$ things have to be looked at more carefully. We do not state a well-posedness result here, the next reasoning merely aims at showing “how this works”. Suppose that $x(t) = 0$ on $(t_1, t_1 + \varepsilon)$, $\varepsilon > 0$. Then from Glocker's Proposition C.8 it follows that the complementarity can be expressed as $0 \leqslant \dot{x}(t) = u(t) + \lambda(t) \perp \lambda(t) \geqslant 0$, that is an LCP with a unique solution whatever $u(t)$. If $u(t) > 0$ then $\lambda(t) = 0$ and $\dot{x}(t) = u(t) > 0$ (the state leaves the constraint). If $u(t) \leqslant 0$ then $\lambda(t) = -u(t) \geqslant 0$ and $\dot{x}(t) = 0$. Thus in all cases there exists a $\lambda(t)$ which allows one to integrate the system.

Example 2.50. Consider the following scalar system, with $u(\cdot)$ a measurable function:

$$\begin{cases} \dot{x}(t) = -x(t) + \lambda(t) \\ 0 \leqslant w(t) = x(t) + \lambda(t) + u(t) \perp \lambda(t) \geqslant 0 \\ x(0) = x_0 \in \mathbb{R}\,. \end{cases} \tag{2.94}$$

The complementarity relations define an LCP with unknown $\lambda(t)$ and a unique solution. If $x(t) + u(t) > 0$ then $\lambda(t) = 0$ and $\dot{x}(t) = -x(t)$. When $x(t) + u(t) \leqslant 0$ then $\lambda(t) = -x(t) - u(t) \geqslant 0$ and $\dot{x}(t) = -2x(t) - u(t)$.

Rather than listing an exhaustive set of dynamical CS, we will focus our attention on two specifications of them. The first one is the class of linear complementarity systems (LCS):

$$\begin{cases} \dot{x}(t) = Ax(t) + B\lambda(t) \\ w(t) = Cx(t) + D\lambda(t) \\ 0 \leqslant w(t) \perp \lambda(t) \geqslant 0 \end{cases} \tag{2.95}$$

with the matrices $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}, C \in \mathbb{R}^{m\times n}, D \in \mathbb{R}^{m\times m}$. In this case, the general cone K is given by the nonnegative orthant of $\mathbb{R}^m$, i.e., $K = \mathbb{R}^m_+$. The well-posedness, and even the very meaning of the dynamics in (2.95), is not trivial. An extensive study of such systems can be found in Heemels et al. (2000) and

Heemels (1999). In Acary et al. (in press) the LCS (2.95) is embedded into a differential inclusion that is an extension of Moreau's sweeping process. Solutions are distributions, whose degree is related to the relative degree between y and λ. Consider for instance the two electrical circuits of Chap. 1, in (1.13) and (1.11). It is easily checked that the relative degree for (1.13) is $r = 1$ whereas for (1.11) it is $r = 0$. Both possess continuous solutions. This is the same for Examples 2.49 ($r = 1$) and 2.50 ($r = 0$). Consider now the bouncing ball in its complementarity formalism in (1.96). This time $r = 2$ and the solutions may be discontinuous, so that the acceleration is a measure. One realizes on these simple examples that there is a strong correlation between the relative degree between the two complementary variables y and λ and the smoothness of the solutions. More will be said on this in Chap. 5.

Remark 2.51. The term *linear* CS does not mean at all that the dynamical systems in (2.95) are linear. In fact they are strongly nonlinear and nonsmooth.

Example 2.52. Let us consider the electrical circuit with ideal diodes in Fig. 2.3, with $R_1, R_2, R_3 \geqslant 0, L_2, L_3 > 0, C_4 > 0$. One has $0 \leqslant -u_{\mathrm{D}_4} \perp x_2 \geqslant 0$ and $0 \leqslant -u_{\mathrm{D}_1} \perp -x_3 + x_2 \geqslant 0$, where u_{D_4} and u_{D_1} are the voltages of the diodes. The dynamical equations of this circuit are the following ones:

$$\begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) = -\left(\frac{R_1+R_3}{L_3}\right) x_2(t) + \frac{R_1}{L_3} x_3(t) - \frac{1}{L_3 C_4} x_1(t) + \frac{1}{L_3}\zeta_1(t) + \frac{1}{L_3}\zeta_2(t) + \frac{u(t)}{L_3} \\ \dot{x}_{3(t)} = -\left(\frac{R_1+R_2}{L_2}\right) x_3(t) + \frac{R_1}{L_2} x_2(t) - \frac{1}{L_2}\zeta_1(t) + \frac{u(t)}{L_2} \\ 0 \leqslant \begin{pmatrix} \zeta_1(t) \\ \zeta_2(t) \end{pmatrix} \perp \begin{pmatrix} -x_3(t) + x_2(t) \\ x_2(t) \end{pmatrix} \geqslant 0 \,, \end{cases} \tag{2.96}$$

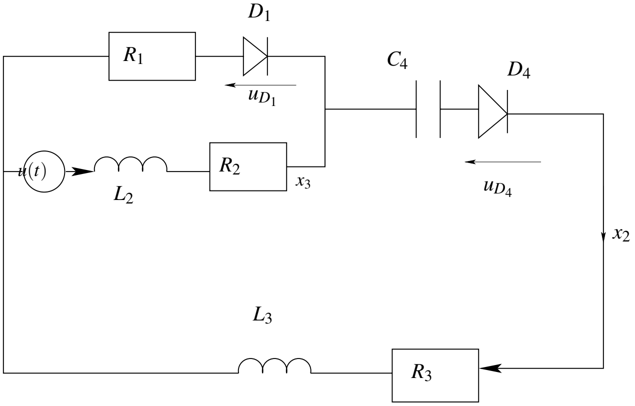


Fig. 2.3. A circuit with ideal diodes

where $x_1(\cdot)$ is the time integral of the current across the capacitor, $x_2(\cdot)$ is the current across the capacitor, and $x_3(\cdot)$ is the current across the inductor L_2 and resistor R_2, $-\zeta_1$ is the voltage of the diode D_1 and $-\zeta_2$ is the voltage of the diode D_4. The system in (2.96) can be written compactly as $\dot{x}(t) = Ax(t) + B\zeta(t) + Eu(t)$, $0 \leqslant \zeta(t) \perp y(t) = Cx(t) \geqslant 0$, with

$$A = \begin{pmatrix} 0 & 1 & 0 \\ -\frac{1}{L_3C_4} & -\frac{R_1+R_3}{L_3} & \frac{R_1}{L_3} \\ 0 & \frac{R_1}{L_2} & -\frac{R_1+R_2}{L_2} \end{pmatrix},$$

$$B = \begin{pmatrix} 0 & 0 \\ \frac{1}{L_3} & \frac{1}{L_3} \\ -\frac{1}{L_2} & 0 \end{pmatrix},\ C = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 1 & 0 \end{pmatrix}, E = \begin{pmatrix} 0 \\ \frac{1}{L_3} \\ \frac{1}{L_2} \end{pmatrix}.$$

2.6.2 Nonlinear Complementarity Systems

Finally, let us introduce a nonlinear complementarity system where the input λ enters into the dynamical system in an affine way:

$$\begin{cases} \dot{x}(t) = f(x(t),t) + g(x(t))\lambda \\ \\ w(t) = h(x(t),\lambda) \\ \\ K^* \ni w(t) \perp \lambda \in K\,. \end{cases} \tag{2.97}$$

This nonlinear special case is interesting because we can recognize the structure given by a Lagrange multiplier λ associated with a constraint, if the function $g(x)$ is equal to the gradients of the constraints, i.e., $g(x) = \nabla h(x)$. We will see further the analogy with the Lagrangian systems. The well-posedness of such systems have received attention in van der Schaft & Schumacher (1998) (in which it is a priori assumed that solutions are right-continuous) and in Brogliato & Thibault (2006) when the triple $(f(x), g(x), h(x))$ satisfies a dissipation equality (in which case the NLCS can be transformed into a perturbed Moreau's sweeping process), see Sect. 4.2.4.

2.7 Second-Order Moreau's Sweeping Process

This formalism has been developed in Moreau (1983), and aims at providing a dynamical framework for Lagrangian systems with unilateral constraints and/or Coulomb friction. Such systems will be examined in detail in Chap. 3. They are *measure differential inclusions* (MDI), i.e., DIs that involve measures (this term was coined by J.J. Moreau). An introduction through the bouncing ball can be found in Chap. 1, see (1.107) and (1.109). We already saw MDIs in Sect. 2.2, see the BV first-order sweeping process in (2.44). The discrepancy between the first-order sweeping process with a BV set $K(\cdot)$ and the second-order sweeping process is that in the former the state jumps are due to the variation of $K(t)$. If $K(t)$ varies smoothly, so will the state. The source of state jumps in the second-order sweeping process is rather due to the relative degree issue that has been briefly discussed at the end of Sect. 2.6 for the LCS in (2.95). Let us now introduce the second-order sweeping process. This is an MDI into a normal cone:

$$-M(q(t))\mathrm{d}v - f(q(t),v(t),t)\mathrm{d}t \in N_{T_{\Phi}(q(t))}(v(t^+)),\ \ q(0) \in \Phi\ , \tag{2.98}$$

where $\Phi = \{q \in \mathbb{R}^n \mid h(q) \leqslant 0\}$ for some $h : \mathbb{R}^n \to \mathbb{R}^m$, Φ is the admissible domain, $q(\cdot)$ is a vector of generalized coordinates, and it is supposed that the time-function $q : \mathbb{R}^+ \to \mathbb{R}^n$ is absolutely continuous, $v(\cdot)$ is the BV generalized velocity, i.e., $q(t) = q(0) + \int_0^t v(s)\mathrm{d}s$, $M(q) = M^{\mathrm{T}}(q) > 0$ is the $n \times n$ inertia matrix, $f(\cdot,\cdot,\cdot)$ contains all inertial forces (Coriolis, centrifugal forces) and external actions like inputs or disturbances and forces that derive from a potential (gravity, elastic forces, etc.). The measure $\mathrm{d}v$ is the acceleration and the Stieltjes measure associated with the BV velocity $v(\cdot)$. The cone $T_{\Phi}(q(t))$ is the tangent cone to the domain Φ, calculated at the position $q(t)$. Thus the right-hand side of (2.98) is the normal cone to $T_{\Phi}(q(t))$, calculated at the velocity $v(t^+)$. The differential inclusion (2.98) looks weird. Three questions arise:

- How does it work?
- Is it well-posed?
- What is it useful for?

The bouncing ball example in Chap. 1 already brought a partial answer to the first question, but we will come back on this later. The second question received positive answers in the works of Monteiro Marques (1993), Mabrouk (1998) and Dzonou & Monteiro Marques (2007), which prove that given $q(0) \in \Phi$ there exists a trajectory with absolutely continuous position and RCLBV velocity (see these references for the assumptions on the data, in particular on Φ). Uniqueness also holds under some mild conditions (Ballard, 2000): the functions $h_i(\cdot)$ and all other data (like external forces) have to be piecewise analytic. However, it is well-known that solutions may not depend continuously on the initial conditions when the boundary of Φ is not smooth, see Chap. 6. The answer to the third point is contained in Chap. 11, where the numerical simulation of (2.98) is described.

Remark 2.53. The discrepancy between the perturbed sweeping process in (2.46) and (2.98) is that the moving set $T_{\Phi}(q)$ is state dependent whereas $K(t)$ in (2.46) is a function of time. Also $T_{\Phi}(q)$ is always a polyhedral convex cone (provided some regularity on Φ is imposed), whereas $K(t)$ is just a convex set. In the same way (2.98) is not equivalent to (2.50). Indeed from Theorem 2.41, the DI in (2.50), which is equivalent to the DI in (2.47) when slow solutions are selected, possesses a Lipschitz-continuous solution $x(\cdot)$. But the solutions of (2.98) with $\Phi \subset \mathbb{R}^n$, a closed domain, are such that $v(\cdot)$ is BV. The source of the discrepancy is that the whole state vector $x(\cdot)$ is constrained to evolve in $K(\cdot)$ in (2.50) or (2.47), whereas only $q(\cdot)$ is constrained in Φ in (2.98).

Let us now bring new material that answers the first question above. Suppose first that $q(t) \in \mathrm{Int}(\Phi)$. Then the tangent cone $T_\Phi(q) = \mathbb{R}^n$, and consequently the normal cone to the tangent cone is the singleton $\{0\}$. This holds whatever $v(t^+)$. In this case (2.98) reduces to the usual Lagrange equations. Suppose now that $q(t) \in \partial\Phi$, the boundary of Φ (since Φ is assumed to be a closed domain it has a boundary). Then $T_\Phi(q)$ is no longer the whole ambient space, but is a cone. Some cases are depicted in Fig. 2.4. Now the value of the right-hand side of (2.98) depends on the right limit of the velocity $v(t^+)$ (remember this is an *implicit* formulation of the dynamics). Clearly $v(t^+)$ must belong to $T_\Phi(q)$, otherwise the right-hand side of (2.98) is the empty set. If $v(t^+)$ points inside $T_\Phi(q)$, the normal cone is equal to the singleton $\{0\}$. Once again we are back to the usual Lagrange dynamics. If $v(t^+)$ is on the boundary $\partial\Phi$ then the right-hand side of (2.98) is a (nonempty) convex cone. This means that it is allowed that a nonzero multiplier λ belongs to $N_{T_\Phi(q(t))}(v(t^+))$.

Let us analyze this from another point of view, which makes the implicit formulation become explicit (and therefore certainly more telling). At the atoms of the differential measure dv (these atoms correspond to the impact times), the measure DI in (2.98) is equivalent to

$$M(q(t))[v(t^+) - v(t^-)] \in -N_{T_\Phi(q(t))}(v(t^+)) \; . \tag{2.99}$$

Now using the equivalences in (A.8), it follows that (2.99) is equivalent to

$$v(t^+) = \underset{z \in T_\Phi(q(t))}{\mathrm{argmin}} \; \frac{1}{2}(z - v(t^-))^{\mathrm{T}} M(q(t))(z - v(t^-)) \tag{2.100}$$

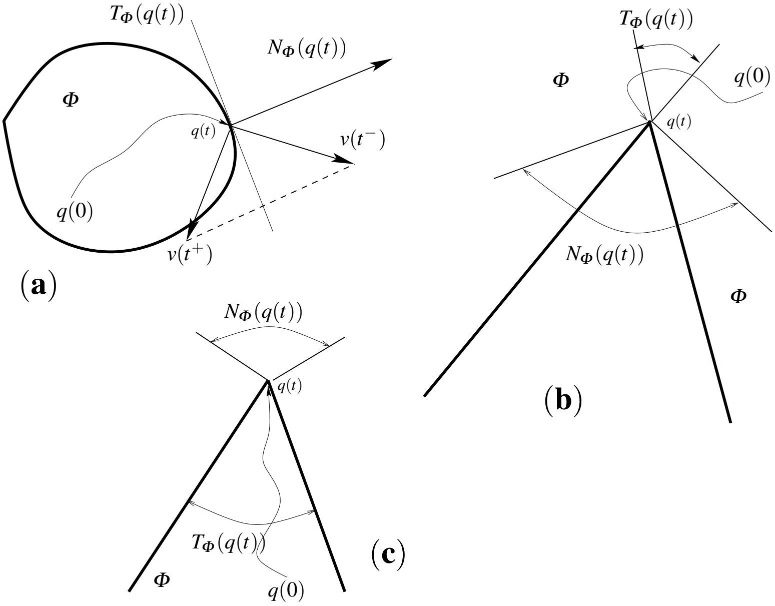


Fig. 2.4. Tangent and normal cones

that is also often written by Moreau as follows:

$$v(t^+) = \mathrm{prox}_{M(q(t))}[T_\Phi(q(t)); v(t^-)] \,. \tag{2.101}$$

One concludes that at atoms of dv, the measure DI in (2.98) imposes a "plastic" impact on the velocity: the post-impact velocity lies on the boundary of $T_\Phi(q(t))$.

Remark 2.54. The above framework easily admits other sorts of impacts. One can introduce a coefficient of restitution $e \in [0,1]$ as follows: the right-hand side of (2.98) is replaced by

$$N_{T_\Phi(q(t))}\left(\frac{v(t^+)+ev(t^-)}{1+e}\right) \,. \tag{2.102}$$

At an impact we obtain

$$M(q(t))[v(t^+)-v(t^-)] \in -N_{T_\Phi(q(t))}\left(\frac{v(t^+)+ev(t^-)}{1+e}\right) \,, \tag{2.103}$$

which we can rewrite equivalently as

$$\frac{v(t^+)+ev(t^-)}{1+e} - v(t^-) \in -(M(q(t)))^{-1}N_{T_\Phi(q(t))}\left(\frac{v(t^+)+ev(t^-)}{1+e}\right) \tag{2.104}$$

from which we deduce (see (A.8))

$$\frac{v(t^+)+ev(t^-)}{1+e} = \mathrm{prox}_{M(q(t))}[T_\Phi(q(t)); v(t^-)] \,, \tag{2.105}$$

i.e.,

$$v(t^+) = -ev(t^-) + (1+e)\mathrm{prox}_{M(q(t))}[T_\Phi(q(t)); v(t^-)] \,. \tag{2.106}$$

Other expressions may be found in Mabrouk (1998). The impact rule in (2.106) assures that the kinetic energy decreases at impact times, provided $e \in [0,1]$. In practice the proximation has to be solved with an optimization algorithm. Multiple impacts are complex phenomena and are still an on-going subject of research (Acary & Brogliato, 2005; Glocker, 2004). Obviously Moreau's rule in (2.106) often produces a post-impact velocity that does not fit with observed experimental results. However, it has to be considered as a fundamental step towards a geometrical description of multiple impacts processes. This is the starting point for the framework developed in Glocker (2004).

Remark 2.55. We have introduced normal and tangent cones to convex sets. Moreau defines the tangent cone as

$$T_\Phi(q) = \{v \in \mathrm{I\!R}^n \mid \text{for all } i \in I(q),\ v^{\mathrm{T}}\nabla h_i(q) \leqslant 0\},$$

where $I(q)$ is the set of active constraints, i.e., $I(q) = \{i \in \{1,m\} \mid h_i(q) \geqslant 0\}$. Then the normal cone is defined as the polar cone to the tangent cone. The active set embeds cases $h_i(q) > 0$ so that it can be calculated numerically when the constraints are slightly violated. Both cones hence defined are closed polyhedral sets. Notice that Φ need not be convex.

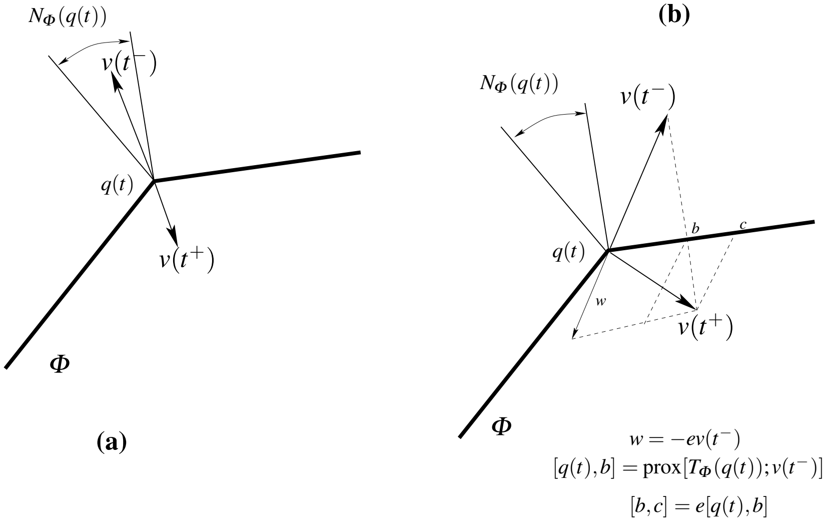


Fig. 2.5. The impact process

The impact in Fig. 2.5 a has $v(t^-) \in N_\Phi(q(t))$, and its projection on $T_\Phi(q(t))$ is zero. From (2.106) one has $v(t^+) = -ev(t^-)$. In Fig. 2.5b one has $v(t^-) \notin N_\Phi(q(t))$ and $v(t^-) \notin \Phi$. The projection of the pre-impact velocity on the tangent cone is the segment $[q(t),b]$. The vector $w = -ev(t^-)$. The segment $[b,c]$ is equal to e $\mathrm{prox}[T_\Phi(q(t));v(t^-)$. From (2.106) one obtains the post-impact velocity. One notes that the impact rule keeps the tangential component of the velocity: only the normal component is changed. Also for the ease of exposition and drawing it is supposed that the inertia matrix $M(q(t))$ is the identity. Notice that the angle between the two faces of the domain Φ in Fig. 2.5 is larger than $\frac{\pi}{2}$. If it is smaller than $\frac{\pi}{2}$ then the pre-impact velocity is always inside $N_\Phi(q(t))$, and the scenario of Fig. 2.5a is the generic one. In general all the quantities (angles, projections) have to be calculated in the kinetic metric defined as $\langle x,y\rangle_q = x^{\mathrm{T}}M(q)y$ for two vectors $x,y \in \mathbb{R}^n$.

2.8 ODE with Discontinuities

2.8.1 Order of Discontinuity

Let us introduce a definition of the order of discontinuity of a function.

Definition 2.56 (Order of discontinuity of a function). *We will say that a discontinuity of a function has order $q \geqslant 0$ if the function has a finite jump in at least one of the partial derivative of order q and has continuous derivatives of order $q-1$, $q-2$,...,0.*

In this section, we will discuss some properties of dynamical systems with low order of discontinuity, i.e., ODEs of the form

$$\dot{x}(t) = f(x(t),t)\ , \tag{2.107}$$

where the right-hand side function $f(\cdot,\cdot)$ has a low order of discontinuity $q \geqslant 0$. The notion of "low order of discontinuity" has no precise definition. It has to be understood regarding the numerical methods that we want to use and the mathematical analysis that we want to perform. Indeed, if the order $q \geqslant 1$ then the standard theory of ODEs can be applied. But numerical solving methods and stability analysis have to take care about the possible nonsmoothness of the gradients of $f(\cdot,\cdot)$. We will see in Sects. 7.2 and 9.1 that higher order time-integration schemes for ODE have some difficulties to deal with such dynamical systems.

2.8.2 Transversality Conditions

Let us consider dynamical systems of the form

$$\dot{x}(t) = \begin{cases} f^-(t,x(t)) \text{ if } & g(t,x(t)) \leqslant 0 \\ \\ f^+(t,x(t)) \text{ if } & g(t,x(t)) > 0 \end{cases}\ , \tag{2.108}$$

where $f^-(\cdot)$ and $f^+(\cdot)$ are locally Lipschitz in x and $g : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}$ is smooth (infinitely differentiable). The vector field $f(\cdot,\cdot)$ made of $f^-(\cdot,\cdot)$ and $f^+(\cdot,\cdot)$ may jump at the switching surface or it may be continuous but with a discontinuous derivative of order $q \geqslant 1$.

According to Definition 2.56, one may speak of a system with an order of discontinuity q. When $q = 0$ the jump is in $f(\cdot,\cdot)$ itself and one may resort to Filippov's formalism to study the system. When $q = 1$ then $f(\cdot,\cdot)$ is continuous when crossing the switching surface; however, its derivative is discontinuous. In other words, the right derivative of $f^-(\cdot,\cdot)$ and the left derivative of $f^+(\cdot,\cdot)$ at the switching surface are not equal. A fundamental assumption is made:

Assumption 1 (Transversality). *There exists $\delta > 0$ such that for all $t \in \mathbb{R}^+$ and all $x \in \mathbb{R}^n$*

$$\begin{cases} \frac{\partial g}{\partial t} + \frac{\partial g}{\partial x} f^-(t,x) \geqslant \delta \\ \\ \frac{\partial g}{\partial t} + \frac{\partial g}{\partial x} f^+(t,x) \geqslant \delta \end{cases}. \tag{2.109}$$

Thanks to Assumption 1, all the trajectories that attain the switching surface cross it so that $g(t,x)$ changes its sign. There is no sliding trajectory that remains on the switching surface and no spontaneous jumps. Thus for any initial data there is a unique AC solution with a finite number of switching states on any finite interval. At each switching state, the derivative $x^{(q+1)}(\cdot)$ has a finite jump. In the rest of this section it will be supposed that Assumption 1 holds true.

Example 2.57. The system

$$\begin{cases} \dot{x}(t) = -\mathrm{sgn}(t)|1-|t||x^2(t) \\ \\ x(-2) = \frac{2}{3},\ t \in [-2,2] \end{cases} \tag{2.110}$$

has two discontinuities of order $q = 1$ at $t = 1$ and $t = -1$ and a discontinuity of order $q = 0$ at $t = 0$ (Hairer et al., 1993).

2.8.3 Piecewise Affine and Piecewise Continuous Systems

These two sorts of dynamical systems are popular in the systems and control community.

Definition 2.58 (Piecewise affine systems). *A piecewise affine (PWA) system can be defined as follows:*

$$\dot{x}(t) = A_i x(t) + a_i, \quad x(t) \in X_i\ , \tag{2.111}$$

where

- *the set $\{X_i\}_{i\in I}$ with $X_i \subset \mathbb{R}^n$, $i \in I$, is a partition of the state space into a number of closed (possibly unbounded) polyhedral cells with disjoint interior; the index set of the cells is denoted by $I \subset \mathbb{N}$;*
- *the matrix $A_i \in \mathbb{R}^{n\times n}$ and the vector $a_i \in \mathbb{R}^n$ define an affine system on each cell.*

The sets X_i are defined by a finite representation of the type

$$X_i = \{y(t)\ \mid\ C_i y(t) \geqslant D_i\}\ , \tag{2.112}$$

where the inequality has to be understood component-wise, the matrix $C_i \in \mathbb{R}^{m\times n}$ and the vector $D_i \in \mathbb{R}^m$ define the polyhedral cell i. By a partition of the state space it is meant that $\cup_{i\in I} X_i = \mathbb{R}^n$. The solution of PWA systems may be defined as a continuous piecewise $\mathscr{C}^1$ function $x(\cdot) \in \cup_{i\in I} X_i$ on the time interval $[0,T]$. The function $x(\cdot)$ is a solution of the system (2.111) if for every $t \in [0,T]$ such that the derivative $\dot{x}(\cdot)$ is defined, the equation $\dot{x}(t) = A_i x(t) + a_i$ holds for all i with $x(t) \in X_i$. The PWA possesses such piecewise $\mathscr{C}^1$ solutions if at each switching point one has continuity of the vector field (conditions like $A_i x(t) + a_i = A_j x(t) + a_j$ have to be satisfied as the state goes from X_i to X_j). Other sufficient conditions may be found in Imura & van der Schaft (2000) and Imura (2003) yielding to well-posed switch-driven piecewise affine systems. Another way to avoid the DI formalism is to postulate an assumption of transversality (see Assumption 1).

Remark 2.59. Definition 2.58 is relatively rough, but can suffice to understand what types of solutions are sought. Indeed, if some discontinuity of the right-hand side is allowed, the canonical problem with the sign function can be cast into such a formalism. We know that the existence of solutions is not guaranteed for such a discontinuous ODE (see (2.12)) which has to be recast into a well-posed framework like

Filippov's inclusions. Johansson & Rantzer (1998) circumvent this problem excluding arbitrarily such cases.

A proper definition of the solution can be given with the DI:

$$\dot{x}(t) = \mathrm{conv}_{j\in J}\{A_i x(t) + a_i\} \text{ with } J = \{j \mid x(t) \in X_j\}\ . \tag{2.113}$$

It is seen that a discontinuity in the vector field may occur only on the boundaries ∂X_i. Let us denote $N = \cup_{i\in I}\partial X_i$, a set which has zero measure in $\mathbb{R}^n$. Then from (2.26) Filippov's right-hand side may be calculated with the formula

$$\mathscr{F}[f](x) = \mathrm{conv}\{\lim f(x_n) \mid x_n \to x,\ x_n \notin N\}$$

so that (2.113) is recovered if the state x belongs to the cells X_j, $j \in J$.

Remark 2.60. Consider the LCS in (2.95). Suppose that D is an $m\times m$ P-matrix. Then from Theorem B.3 it follows that $\lambda(\cdot)$ is a piecewise linear function of x, and the LCS is consequently an ODE with a piecewise linear right-hand side. Using the same notation as in Theorem B.3 the LCS can be rewritten as (Camlibel et al., 2006)

$$\dot{x}(t) = [A - B_{\bullet\alpha}(D_{\alpha\alpha})^{-1}C_{\alpha\bullet}]x(t) \text{ if } \begin{pmatrix} -(D_{\alpha\alpha})^{-1} & 0 \\ -D_{\bar{\alpha}\alpha}(D_{\alpha\alpha})^{-1} & I_{\bar{\alpha}\bar{\alpha}} \end{pmatrix}\begin{pmatrix} C_{\alpha\bullet} \\ C_{\bar{\alpha}\bullet} \end{pmatrix} x(t) \geqslant 0\ , \tag{2.114}$$

where the notation $B_{\bullet\alpha}$ is the matrix constructed from B by taking all the rows, and the columns indexed by integers in $\alpha \subset \{1,...,m\}$, and $C_{\alpha\bullet}$ is constructed from C by taking rows indexed in α and all columns. We have already seen a particular case of an LCS that can be rewritten as a piecewise linear system in Chap. 1, see (1.11) and (1.21).

The next two examples are inspired from Heemels & Brogliato (2003).

Example 2.61. Consider the system

$$\dot{x}(t) = \begin{cases} f_{11}(x(t)) \text{ if } h_1(x(t)) > 0 \text{ and } h_2(x(t)) > 0 \\ f_{10}(x(t)) \text{ if } h_1(x(t)) > 0 \text{ and } h_2(x(t)) < 0 \\ f_{01}(x(t)) \text{ if } h_1(x(t)) < 0 \text{ and } h_2(x(t)) > 0 \\ f_{00}(x(t)) \text{ if } h_1(x(t)) < 0 \text{ and } h_2(x(t)) < 0 \end{cases} \tag{2.115}$$

with $x(t) \in \mathbb{R}^n$. It is assumed that the smooth functions $h_1(\cdot)$ and $h_2(\cdot) : \mathbb{R}^n \to \mathbb{R}$ are such that the ambient space $\mathbb{R}^n$ is divided into four parts, each part corresponds to the activation of a vector field. The condition $f_{11}(x) = f_{10}(x) = f_{00}(x) = f_{01}(x)$ for x such that $h_1(x) = h_2(x) = 0$ guarantees that the vector field is continuous and therefore (2.115) is an ODE. From property (vii) of Theorem 2.22, Filippov's convexification always yields $\mathscr{F}[f](x) = \{f(x)\}$ in such a case. Let us investigate what happens when the vector field jumps occur on the codimension one surfaces $\Sigma_i = \{x \in \mathbb{R}^2 \mid$

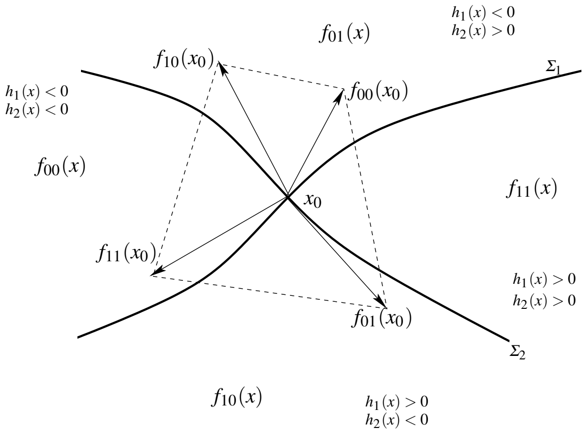


Fig. 2.6. Filippov's convexification on Σ_{12}

$h_i(x) = 0\}$ and on the codimension two subspace $\Sigma_{12} = \{x \in \mathbb{R}^2 \mid h_1(x) = h_2(x) = 0\} = \Sigma_1 \cap \Sigma_2$. We may use (2.26) for the calculation of $\mathscr{F}[f](x)$. For instance on Σ_{12} we get

$$\mathscr{F}[f](x) = \text{conv}\{\lim f(x_i) \mid x_i \to x,\ x_i \notin \Sigma_1 \cap \Sigma_2\} \tag{2.116}$$

that is the convex hull of the four vector fields at x, i.e., $f_{11}(x)$, $f_{10}(x)$, $f_{00}(x)$, and $f_{01}(x)$. This is depicted in Fig. 2.6, where we have denoted x_0 the value of the state on Σ_{12}. The convex hull is in dashed line. Let us make two comments. First the vector fields are not defined on Σ_1 or Σ_2 in (2.115). Therefore the dynamics in (2.115) is not complete. It is completed either by imposing some continuity on the switching surfaces or by embedding the system into Filippov's framework. Second, and most importantly, Filippov's convexification allows us to construct a new model, which we know is well-posed in the sense that there exists an absolutely continuous solution for any $x(0) = x_0 \in \mathbb{R}^n$. However, uniqueness is not guaranteed (additional conditions have to be imposed as we saw in Sects. 2.1.3 and 2.1.5), and even if it is, one should know more to integrate the system through $\Sigma_1 \cap \Sigma_2$. Is this an attractive, repulsive, crossing subspace? The reader is referred to Sects. 7.1.1 and 7.1.2 where a powerful method invented by D. Stewart allows one to determine the future solutions when the trajectory attains a switching surface. Notice that one may also rewrite (2.115) quite similarly as (7.4) or (7.5), which once again shows the close link between the piecewise smooth system (2.115) and differential inclusions, via the sign set-valued function.

Example 2.62. We consider the LCS

$$\begin{cases} \dot{x}(t) = Ax(t) + b\lambda(t) + eu(t) \\ \\ 0 \leqslant w(t) = c^{\mathrm{T}}x(t) + d\lambda(t) \perp \lambda(t) \geqslant 0 \\ \\ x(0) = x_0 \in \mathbb{R}^n \end{cases}, \tag{2.117}$$

where $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^{n\times 1}$, $c \in \mathbb{R}^{n\times 1}$, $d \in \mathbb{R}$, and $e \in \mathbb{R}^{n\times 1}$. Suppose $d > 0$. Then the complementarity relation is a LCP with unknown $\lambda(t)$ and a unique solution whatever $c^{\mathrm{T}}x(t)$, that is Lipschitz continuous in x (as the solution of an LCP with positive definite matrix, see Theorems B.2 and B.3). We may rewrite the system as

$$\dot{x}(t) = \begin{cases} Ax(t) + eu(t) & \text{if } c^{\mathrm{T}}x(t) \geqslant 0 \\ \\ (A - bd^{-1}c^{\mathrm{T}})x(t) + eu(t) & \text{if } c^{\mathrm{T}}x(t) \leqslant 0 \end{cases} . \tag{2.118}$$

This shows that some LCS (with relative degree $r = 0$) can be interpreted as piecewise linear systems (take u as a constant to recover the class of piecewise affine systems of this section). We have already seen this in Chap. 1 with the electrical circuit (**b**). Let us now suppose that $d = 0$ and that $c^{\mathrm{T}}b > 0$ (and $c^{\mathrm{T}}x_0 \geqslant 0$). Then we can rewrite the system as

$$\dot{x}(t) = \begin{cases} Ax(t) + eu(t) & \text{if } (c^{\mathrm{T}}x(t), c^{\mathrm{T}}Ax(t) + c^{\mathrm{T}}eu(t)) \succeq 0 \\ \\ P(Ax(t) + eu(t)) & \text{if } c^{\mathrm{T}}x(t) = 0 \text{ and } c^{\mathrm{T}}Ax(t) + c^{\mathrm{T}}eu(t) \leqslant 0 \end{cases} \tag{2.119}$$

with $P = I_n - b(c^{\mathrm{T}}b)^{-1}c^{\mathrm{T}}$, and $\succeq$ is the lexicographical inequality (the first nonzero element has to be nonnegative). The way to go from (2.117) to (2.119) is similar to what has been done with the electrical circuit (c) in Chap. 1. The condition $c^{\mathrm{T}}b > 0$ is a fundamental one, which means that the principal Markov parameter of the system is positive. It guarantees that when the system evolves on the surface $\{x \in \mathbb{R}^n \mid c^{\mathrm{T}}x = 0\}$, then the LCP: $0 \leqslant \lambda(t) \perp c^{\mathrm{T}}(Ax(t) + b\lambda(t) + eu(t)) \geqslant 0$ possesses a unique solution $\lambda(t)$ that is a Lipschitz-continuous function of $x(t)$ and $u(t)$. The positivity of the leading Markov parameter is an essential condition for the well-posedness that will be encountered in more general cases of LCS, see Chap. 5. Therefore in the second case also we have been able to go from an LCS to a kind of piecewise linear system. As we said in Chap. 1 concerning circuit (c), this is not really what is usually called in the literature a PWL system, as the switching conditions show.[6] This is a differential inclusion in a normal cone as the complementarity relation $0 \leqslant c^{\mathrm{T}}x(t) \perp \lambda(t) \geqslant 0$ is equivalent to the inclusion $-\lambda(t) \in N_K(x(t))$, where $K = \mathbb{R}^+$ (see (A.9)). This inclusion defines a set-valued mapping.

Remark 2.63. Recall that an LCP: $0 \leqslant \lambda \perp \lambda + b \geqslant 0$ does not define a set-valued, but a single-valued mapping (see Sect. A.3). This is why well-posed LCS with $r = 0$ are not inclusions but ODEs.

Let us now turn our attention to another class of piecewise systems often used in the systems and control scientific community.

Definition 2.64 (Piecewise continuous systems). *A piecewise continuous (PWC) system can be defined by*

[6] To convince oneself of this, try to draw the cells defined by the lexicographical inequality, when $n = 2$. They do not fit with Definition 2.58.

$$\dot{x}(t) = f_i(x(t),t), \quad x(t) \in X_i,\, x(0) = x_0 \,, \tag{2.120}$$

where the continuous functions $f_i : \mathrm{I\!R}^n \times [0,T] \to \mathrm{I\!R}^n$ *define a continuous system on each cell* X_i*.*

In a general way, it is difficult to understand PWA and PWC systems without referring to one of the following formalisms:

- ODE with Lipschitz right-hand side,
- Filippov differential inclusions,
- Higher relative degree systems.

PWA or PWC or PWL (linear) or PWS (smooth) systems have to be recast in one of these classes, to be given a meaning in terms of existence and uniqueness of solutions (in other words, what is a solution for a general PWA or PWC system?). This is what is done properly in Orlov (2005) (see Definition 2.1 in this chapter).

2.9 Switched Systems

Switched systems can be defined as

$$\dot{x}(t) = f_{\sigma(t)}(x(t)),\, t \in [0,T] \,, \tag{2.121}$$

where $\sigma : [0,T] \to \mathrm{I\!R}$ is called a switching signal usually taking integer values i, and the vector fields $f_i(\cdot)$ are locally Lipschitz continuous. To assure the existence of solutions to the time-varying system (2.121) one may simply resort to Carathéodory Theorem. In particular it is sufficient that $f_{\sigma(t)}(x)$ be Lebesgue measurable in t for each x. For instance Boscain (2002) considers controlled systems of the form $\dot{x}(t) = u(t)Ax(t) + (1 - u(t))Bx(t)$ where $u(\cdot) : \mathrm{I\!R}^+ \to [0,1]$ is measurable. By a switching signal, one may also mean a signal that exhibits only a finite number of discontinuities in any finite time interval and is right-continuous. Accumulations of switches are consequently excluded from such a framework. A related definition is the following (Hespanha, 2004; Vu & Liberzon, 2005):

$$\dot{x}(t) = f_{\sigma}(x(t)),\, t \in [0,T],\, \sigma(\cdot) \in \mathscr{S} \,, \tag{2.122}$$

where $\mathscr{S}$ is a set of piecewise constant signals, called the switching signals. The difference between (2.121) and (2.122) is that solutions of (2.122) are parameterized by $x(0)$ and σ, whereas in (2.121) solutions are parameterized only by $x(0)$. For further stability study, the solutions of such switching systems may be supposed to be continuous piecewise differentiable, i.e., they are continuous functions of time whose derivative may exhibit a finite number of jumps in any finite interval of time.

One may also define switched systems starting from a PWC of a PWA system. However, in such a case the switching function $\sigma(\cdot)$ is no longer exogenous, but state dependent, because the times at which the vector field changes depend on whether or not the state $x(\cdot)$ has reached some set boundary. As for PWA of PWC systems, one has in such a case to be careful with the definition of switched systems, since it may easily happen that Carathéodory solutions do not exist, and one has to resort to Filippov's modeling to give a meaning to (2.121). Again one way to guarantee the well-posedness with continuous piecewise differentiable solutions and without changing the model is to impose that $f_{\sigma(t_k^+)}(x) = f_{\sigma(t_k^-)}(x)$, where t_k is a switching time (a time at which the function $\sigma(\cdot)$ jumps).

Example 2.65. Consider the following system (inspired from Hespanha, 2004). Let

$$\dot{x}(t) = -\sigma(x(t))x(t), \ \ \sigma \in \mathscr{S} \ , \tag{2.123}$$

where $\mathscr{S}$ contains all the pairs (x,σ) with $x(\cdot) : \mathrm{I\!R}^+ \rightarrow \mathrm{I\!R}$ piecewise differentiable, and $\sigma(\cdot) : \mathrm{I\!R}^+ \rightarrow \mathrm{I\!R}^+$ is piecewise constant, with

$$\sigma(x(t)) = \begin{cases} 0 \ \ \text{if}\, x(t) = 0 \\ \\ 2^n \ \text{if}\ |x(t)| \in [2^{-n-1}, 2^{-n}),\ n \in \mathbb{Z} \ . \end{cases} \tag{2.124}$$

Consider $x = \frac{1}{2^n}$, $n < +\infty$. Let us call the vector field in (2.123) and (2.124) $f(x)$, which is depicted in Fig. 2.7. Then $f(x^+) = -\frac{1}{2}$ and $f(x^-) = -1$. Similarly $x = -\frac{1}{2^n}$, with a jump between $\frac{1}{2}$ and 1. Therefore the vector field jumps at each $|x| = \frac{1}{2^n}$, and there is an accumulation of jumps as x approaches zero. At $x = 0$ one has $f(0) = 0$ since $[2^{-n-1}, 2^{-n}) \rightarrow \{0\}$ as $n \rightarrow +\infty$, and $f(0^+) = (-1, -\frac{1}{2}]$, $f(0^-) = [\frac{1}{2}, 1)$. Filippov's convexification for this system replaces the jumps by segments $[-1, -\frac{1}{2}]$ for $x > 0$ and segments $[\frac{1}{2}, 1]$ for $x < 0$. At $x = 0$ it imposes $f(0) = [-1, 1]$. One may use (2.26) to compute these sets. Here we may highlight the importance of defining the system with closed bracket $[2^{-n-1}, 2^{-n})$ in (2.124), if one considers Carathéodory solutions instead of Filippov's solutions. Filippov's solutions ignore the value of the discontinuous vector field at points $x = \pm\frac{1}{2^n}$. Carathéodory solutions do not ignore it. If one had defined $\sigma(\cdot)$ with open sets $(2^{-n-1}, 2^{-n})$, assigning the value 0 at $x = \pm\frac{1}{2^n}$, then Carathéodory solutions would simply stop at $x = \pm\frac{1}{2^n}$ for some n. Filippov's solutions would not.

The system is designed in such a way that all the switching surfaces $\Sigma_n = \{x \in \mathrm{I\!R} \mid x = \frac{1}{2^n}\}$ and $\Sigma_{-n} = \{x \in \mathrm{I\!R} \mid x = -\frac{1}{2^n}\}$ are crossed transversally. Therefore the spontaneous jumps (see Sect. 7.1.2) which are typical in Filippov's systems with repulsive switching surfaces are ruled out, and the uniqueness of solutions holds. In order to study the stability of the origin $x = 0$, one may choose a Lyapunov function $V(x) = x^2$. However, one has to resort to the specific tools of convex analysis to study the variations of $V(\cdot)$ along the (absolutely continuous) solutions. The set-valued right-hand side hence constructed is not Lipschitz continuous nor uniformly one-sided Lipschitz continuous nor monotone.

Systems with switching set-valued mappings are analyzed in Mancilla-Aguilar, Garcia, Sontag, & Wang (2005) (i.e., the locally Lipschitz vector fields $f_i(\cdot)$ are replaced by locally Lipschitz set-valued mappings $F_i(\cdot)$).

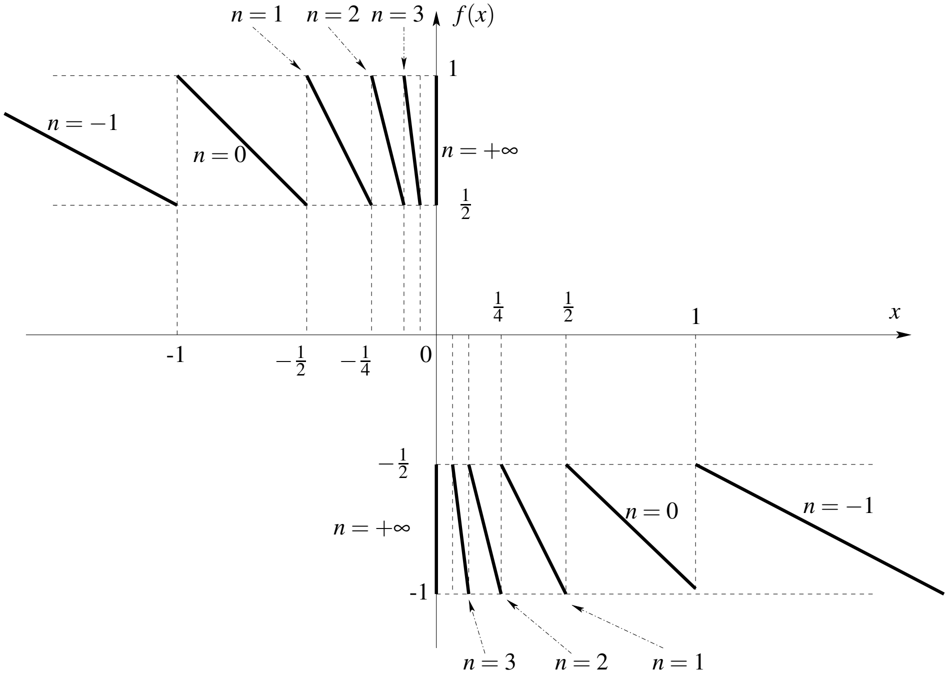


Fig. 2.7. A switched vector field

2.10 Impulsive Differential Equations

2.10.1 Generalities and Well-Posedness

Roughly speaking, impulsive ODEs are ODEs with inputs that may be Dirac measures. Hence the state may jump. One way to write down the dynamics of impulsive ODEs is

$$\begin{cases} \dot{x}(t) = f(x(t),t) \text{ for all } t \neq t_k \\ \\ x(t_k^+) - x(t_k^-) = I(x(t_k^-)) \text{ for all } t = t_k(x) \\ \\ x(0) = x_0, t \geqslant 0 \end{cases}, \tag{2.125}$$

where the sequence of times $\{t_k\}_{k\geqslant 0}$ may be purely exogenous or state dependent, i.e., $t_k = t_k(x)$. The dynamics in (2.125) may represent physical systems like a predator–prey system in which some quantity of one of the species (pikes, trouts) is added or subtracted. It is implicitly supposed in (2.125) that the solutions possess right and left limits everywhere. This imposes some restrictions on the set

$\{t_k\}_{k\geqslant 0}$ as for instance one cannot have $\{t_k\}_{k\geqslant 0} = \mathbb{Q}$. Usually one imposes that $0 < t_1 < t_2 \cdots < t_k < \cdots$, and even more: $t_{k+1} - t_k > \gamma$ for some $\gamma > 0$ and all $k \geqslant 0$. Therefore solutions of (2.125), if any, are piecewise differentiable.

The dynamics of impulsive ODEs is sometimes written as

$$\dot{x}(t) = f(x(t),t) + g(x(t),t)\dot{u},\; x(0) = x_0,\; t \geqslant 0\,, \tag{2.126}$$

where $u(\cdot)$ is a function of bounded variation (see Sect. C.1) or a Lebesgue measurable function. As $g(\cdot,\cdot)$ may be state dependent and $\dot{u}$ is a measure (that should preferably be written as the Stieltjes measure du), the well-posedness issues associated with (2.126) require much care. Actually (2.126) should rather be seen as an equality of distributions, under the condition that the right-hand side is itself well defined as a distribution. Well-posedness results have been published (Bressan & Rampazzo, 1993; Orlov, 1985, 2000) that concerns impulsive ODEs as

$$\dot{x}(t) = f(x(t)) + g(x(t))\dot{u},\; x(0) = x_0,\; t \geqslant 0\,, \tag{2.127}$$

where $f(\cdot)$, $g(\cdot)$ are $C^1[\mathbb{R}^n;\mathbb{R}^n]$, the vector field $g(\cdot)$ is such that for every x the map $t \mapsto \exp(tg)(x)$ is defined for all $t \in \mathbb{R}$, and $u(\cdot)$ is a bounded, measurable function of time. Let us now summarize the results of Bressan & Rampazzo (1993), which are very close in spirit to Orlov (1985). In what follows the notation $(\exp(th)(x)$ means the value at time t of the solution of the ODE $\dot{y}(t) = h(y(t))$, $y(0) = x$

Definition 2.66. *A function $x : [0,T] \to \mathbb{R}^n$ is a* generalized solution *of (2.127) if $x(t) = \exp(u(t)g)(y(t))$, where $y(\cdot)$ is a Carathéodory solution of the ODE:*

$$\dot{y}(t) = f^u(t,y(t)),\;\; y(0) = \exp(-u(0)g)(x_0)\,, \tag{2.128}$$

where $f^u(t,y) = (\exp(-u(t)h))_* f(\exp(u(t)h)(y))$, and

$$(\exp th)_* f(x) = \lim_{\varepsilon\to 0} \frac{(\exp(th)(x+\varepsilon f(x)) - (\exp(th)(x)}{\varepsilon}\;.$$

When $u(\cdot)$ is a smooth function, this definition is the classical definition of a solution for an ODE.

Theorem 2.67. *Under the stated assumptions, there exists $T > 0$ such that the Cauchy problem (2.127) has a unique generalized solution on $[0,T]$. If in addition the linear growth conditions $||f(x)|| \leqslant c(1+||x||)$, $\|\frac{\partial g}{\partial x}(x)\| \leqslant c$ hold for some c and all $x \in \mathbb{R}^n$, then the solution exists globally on $[0,T]$ and is uniquely defined.*

Example 2.68. As an example consider the scalar impulsive ODE

$$\dot{x}(t) = x(t)\dot{u},\;\; x(0) = x_0 \tag{2.129}$$

so that $f(\cdot) = 0$, $g(x) = x$, and we choose $u(t) = 0$ if $t < 0$, $u(t) = 1$ if $t \geqslant 0$. This $u(\cdot)$ is right continuous, BV, and $\dot{u} = \delta_0$, the Dirac measure at $t = 0$, and

$\dot{u} = 0$ almost everywhere. One has $f^u(\cdot) = 0$, so that (2.128) is the ODE $\dot{y}(t) = 0$, $y(0) = \exp(-g)(x_0)$ is the solution at $t = -1$ of the ODE $\dot{z}(t) = z(t)$, $z(0) = x_0$, that is $y(0) = \frac{x_0}{e} = y(t)$.[7] The generalized solution is then given by $x(t) = \exp(u(t)x)(y(t)) = \exp(u(t)x)(y(0))$, that is the solution at time $u(t)$ of the ODE $\dot{z}(t) = z(t)$, $z(0) = y(0)$. For $t < 0$ we have $u(t) = 0$ so $x(t) = y(0) = \frac{x_0}{e}$. For $t \geqslant 0$ we have $u(t) = 1$ so $x(t) = y(0)e = x_0$. Since we are interested by solutions on $\mathbb{R}^+$ we conclude that the generalized solution of (2.129) is the constant x_0.

Let us now choose $u(t) = 0$ if $t < 1$, $u(t) = 1$ if $t \geqslant 1$. The ODE (2.128) is $\dot{y}(t) = 0$, $y(0) = \exp(0)(x_0)$, i.e., $y(0) = x_0$, so that $y(t) = x_0$. Thus the generalized solution is $x(t) = \exp(u(t)g)(y(t))$, that is the value at time $u(t)$ of the solution of the ODE $\dot{z}(\tau) = z(\tau)$, $z(0) = y(t) = x_0$, that is $z(t) = x_0 \exp(\tau)$ for all $\tau \geqslant 0$. For $t < 1$ we have $u(t) = 0$ so $x(t) = x_0$, for $t \geqslant 1$ we have $u(t) = 1$ so that $x(t) = x_0 e$. We therefore obtain a solution of (2.129) that jumps from x_0 to ex_0 at time $t = 1$.

We realize on this simple example that the generalized solution contains the intuitive jump that should occur in the solution when the input is a Dirac. This is, however, not so obvious because at the instant of jump $t = 1$, the right-hand side of (2.129) multiplies a function (discontinuous at $t = 1$) and a Dirac measure. According to the theory of distributions this is a mathematical object that does not exist! The notion of generalized solution proposed in Bressan & Rampazzo (1993) and Orlov (1985) overcomes this obstacle.

On such a simple example the dynamics can be integrated directly. Indeed we can write (2.129) as

$$\frac{\dot{x}(t)}{x(t)} = \dot{u},\ \ x(0) = x_0\ . \tag{2.130}$$

provided $x(t) \neq 0$. Integrating both sides (supposing $\dot{u}$ is integrable) we get

$$\ln(x(t)) = \ln(x_0) + u(t) - u(0)\ \Rightarrow\ \ x(t) = x_0 \exp(u(t) - u(0)) \tag{2.131}$$

The above results with the two values for $u(\cdot)$ are recovered with the direct integration method. In a sense, the notion of generalized solution proposed in Bressan & Rampazzo (1993) means that we are looking for the solution of a system that is not exactly the system in (2.127), but another ODE whose solutions satisfy (2.127) almost everywhere. In other words, one may say that the right and sound definition of the system is in (2.131), and that writing it as in (2.129) requires much care when $u(\cdot)$ is not differentiable. The procedure that we employed to transform and integrate the impulsive ODE (2.129) is generalized to a broader class of systems in Orlov (1985) when some commutativity conditions are fulfilled (see also Brogliato (1999), Sect. 1.4).

Another class of impulsive ODEs that is encountered in the literature is as follows:

[7] e is the Neperian logarithm constant, i.e., $\ln(e) = 1$ and $\exp(1) = e$.

$$\begin{cases} \dot{x}(t) = f(x(t),t) & \text{if } x(t) \notin \mathscr{X} \\ \\ x(t^+) - x(t^-) = f_d(x(t),t) & \text{if } x(t) \in \mathscr{X} \end{cases} \tag{2.132}$$

where $\mathscr{X}$ is a so-called resetting set, and it is assumed that

(i) If $x \in \mathscr{X}$ then $x + f_d(x,t) \notin \mathscr{X}$,.
(ii) If at time t one has $x(t) \in \bar{\mathscr{X}} \setminus \mathscr{X}$, then there exists an $\varepsilon > 0$ such that for all $0 < \delta < \varepsilon$, one has $x(t+\delta) \notin \mathscr{X}$.
(iii) The vector field $f(x,t)$ is such that between state jumps, the system is well-posed.

The first assumption means that a trajectory cannot enter the resetting set through a point that belongs to its closure but not to $\mathscr{X}$ itself. The second assumption implies that any trajectory that enters $\mathscr{X}$ is instantaneously directed outside $\mathscr{X}$. Thus no trajectory can reach the interior of the resetting set, and the state discontinuities are separated, i.e., they satisfy $0 < t_1 < t_2 \cdots < t_k < \cdots$. However, the framework admits accumulations of resetting times.

Remark 2.69. The assumptions (i) and (ii) therefore introduce a notion of unilaterality in (2.132), since one may define a set in which trajectories cannot penetrate. However, one must not confuse these systems with complementarity systems, and in particular mechanical systems with unilateral constraints, from which they differ a lot. Indeed complementarity systems may live on lower dimensional subspaces, which is not the case of (2.132) (and of none of the other presented impulsive ODEs). Moreover complementarity systems possess very specific features due to the complementarity conditions, which are absent in (2.132). Let us illustrate this on a simple example.

$$\dot{x}(t) = \sin\left(x(t) + \frac{5\pi}{4}\right) + \cos\left(x(t) + \frac{3\pi}{4}\right)\dot{u}(t),\ \ x(0^-) = x_0,\ \ x(t) \in \mathbb{R}\,, \tag{2.133}$$

where $u(\cdot)$ is of bounded variation. Applying Theorem 2.67, this impulsive ODE has a unique global generalized solution. Consider now a complementarity system that looks like (2.133):

$$\begin{cases} \dot{x}(t) = \sin\left(x(t) + \frac{5\pi}{4}\right) + \cos\left(x(t) + \frac{3\pi}{4}\right)\lambda(t),\ \ x(0^-) = x_0,\ \ x(t) \in \mathbb{R} \\ \\ 0 \leqslant x(t) \perp \lambda(t) \geqslant 0\,. \end{cases} \tag{2.134}$$

Suppose that $x_0 = 0$. Then if $\lambda(0) = 0$ one gets $\dot{x}(0) = \sin(\frac{5\pi}{4}) < 0$. It is necessary that there exists a $\lambda(0) > 0$ such that $\dot{x}(0) \geqslant 0$. However, since $\cos(\frac{3\pi}{4}) < 0$, this is not possible and necessarily $\dot{x}(0) < 0$. If $x_0 < 0$, then an initial jump must occur and $x(0^+) \geqslant 0$. If $x(0^+) = 0$ the previous analysis applies. One sees that defining generalized solutions as in Theorem 2.67 is not sufficient. Therefore the complementarity system in (2.134) is not well-posed, despite its resemblance with the impulsive ODE in (2.133). Example 2.7 of the bouncing ball in Haddad et al. (2006), which is inept from the mechanical point of view, clearly shows that the formalism (2.132) is not a suitable framework for complementarity dynamical systems, as it cannot even describe what happens on the constraint surface during persistent contact (it simply does not allow for persistent contact).

Further information on impulsive ODEs may be found in Bainov & Simeonov (1989). There also exist models which somewhat mix unilaterality and impulsive terms, see for instance neural networks models (Tonnelier & Gerstner, 2003). Let us again insist on the fact that the impulsive ODEs presented in this section and measure DIs or complementarity systems, are quite distinct formalisms.

2.10.2 An Aside to Time-Discretization and Approximation

Normally this chapter is not dedicated to numerical schemes. The time-discretization of most of the NSDS presented in this chapter will be studied in Part II of the book. However, we will not see again impulsive ODEs in Chaps. 7–11, so their time-discretization is briefly presented now. Let us consider (2.127). It is assumed that the same assumptions as in Theorem 2.67 hold. The impulsive ODE is studied on a time interval $[0,T]$, $T > 0$.

Theorem 2.70. *Let the control input $u(\cdot) \in \mathscr{L}^1([0,T];\mathrm{I\!R}) \times \mathscr{L}^\infty([0,T];\mathrm{I\!R})$ and be pointwise defined at a given $\tau \in [0,T]$ and a $t = 0$. Let $\{u_n(\cdot)\}_{n\in\mathrm{I\!N}}$ be a sequence in $W^{1,2}([0,T];\mathrm{I\!R})$, such that the variation of $u_n(\cdot)$ on $[0,T]$ satisfies $var_{[0,T]}(u_n) \leqslant L$ for some constant L, and $u_n(\cdot) \to u(\cdot)$ in $\mathscr{L}^1([0,T];\mathrm{I\!R})$. Moreover suppose that $u_n(0) \to u(0)$ and $u_n(\tau) \to u(\tau)$. Then the Carathéodory solutions of the ODE (2.127) with the control input $u_n(\cdot)$ satisfy $x_n(\cdot) \to x(\cdot)$ in $\mathscr{L}^1([0,T];\mathrm{I\!R}^n)$ and $x_n(\tau) \to x(\tau)$, where $x(\cdot)$ is the generalized solution of (2.127) with the control input $u(\cdot)$ pointwise defined at $t = 0$ and $t = \tau$.*

We recall that $u_n(\cdot) \to u(\cdot)$ in $\mathscr{L}^1([0,T];\mathrm{I\!R})$ means that $\int_0^T ||u_n(t) - u(t)||\mathrm{d}t \to 0$ as $n \to +\infty$. Also $W^{1,2}([0,T];\mathrm{I\!R})$ is a Sobolev space, i.e., functions which are in $\mathscr{L}^2([0,T];\mathrm{I\!R})$ and whose first-order derivative is also in $\mathscr{L}^2([0,T];\mathrm{I\!R})$. This theorem signifies that the generalized solutions in Definition 2.66 can be approximated when the control input is replaced by some approximation. This is not strictly speaking a time-discretization of the impulsive ODE. One may think of Theorem 2.70 as a first step towards the definition of a nonimpulsive ODE that approximates (2.127), and this nonimpulsive ODE can be discretized with a usual scheme (Euler, Runge–Kutta).

2.11 Summary

In this chapter we have presented a number of systems that may be classified under the general umbrella “nonsmooth dynamical systems”. The nonsmoothness comes from the fact that their solutions are not differentiable everywhere or, worse, they may be discontinuous. Differential inclusions occupy a large place in the NSDS class. There are various types of differential inclusions, with very different properties of their right-hand sides: Lipschitz continuous, upper semi-continuous, maximal monotone, one-sided Lipschitz continuous, normal cones to convex sets, etc. Variational inequalities with dynamics (also named evolution variational inequalities), projected dynamical systems, complementarity systems are other kinds of NSDS which are sometimes closely linked with some kinds of differential inclusions, even equivalent. As a consequence the mathematical tools which are used to study NSDS come from convex analysis, nonsmooth analysis, complementarity theory, the theory of variational inequalities (there are many sorts of these as well). Some more NSDS exist, like various sorts of impulsive ordinary differential equations, piecewise *schtroumpf* systems (where *schtroumpf* may mean linear, affine, continuous, smooth). There are more than those presented in this chapter. Our aim is not at being exhaustive, but only at presenting the most encountered ones in the literature, with their properties and features and with examples that illustrate the theory. This is a necessary step before tackling time-discretization.

The reader may have the feeling that the large number of NSDS that exist and are presented in this chapter is an obstacle to get a clear picture of their main properties. However, this is unavoidable. When one wants to obtain accurate results then one necessarily has to narrow the class of systems under study. Introducing larger classes which embed several other subclasses (as for instance the DVIs of Sect. 2.4) generally has a limited usefulness. Of much greater interest is the study of the relationships between the existing formalisms, like possible equivalences (Brogliato et al., 2006).

3

Mechanical Systems with Unilateral Constraints and Friction

This chapter aims at providing a rapid overview on modeling aspects of nonsmooth mechanical systems. Some of the material has already been presented on particular examples in Chaps. 1 and 2. We start with the Lagrange and the Newton–Euler formalisms when there are no nonsmooth effects. The local kinematics between two bodies that make contact are detailed. Then nonsmoothness is introduced and Moreau's sweeping process is derived. The chapter ends with a short presentation of the various contact models for impact and friction one may encounter.

3.1 Multibody Dynamics: The Lagrangian Formalism

Let us consider a system of n_{b} rigid bodies parameterized by a set of generalized coordinates $q(t) \in \mathbb{R}^n$, whose motion is defined on a time interval $[0,T]$, $T > 0$. The generalized velocities $v(t) \in \mathbb{R}^n$ are usually defined as the derivative with respect to time of these generalized coordinates: $v(t) = \frac{\mathrm{d}q}{\mathrm{d}t}(t)$. In the classical Lagrangian setting, the equations of motion are derived from the Lagrange's equations as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L(q(t),v(t))}{\partial v_i}\right) - \frac{\partial L(q(t),v(t))}{\partial q_i} = Q_i(q(t),t), \quad i \in \{1\dots n\}\,, \tag{3.1}$$

where the Lagrangian of the system

$$L(q,v) = T(q,v) - V(q)$$

is composed of the kinetic energy

$$T(q,v) = \frac{1}{2}v^{\mathrm{T}}M(q)v$$

and the potential energy of the system, $V(q)$. The vector $Q(q,t) \in \mathbb{R}^n$ denotes the set of generalized forces corresponding to the parameterization q and is determined using the principle of virtual work (see the example of the pendulum in a section below). The matrix $M(q)$, called the mass matrix contains all the masses and the moments of inertia. In most applications one has $M(q) = M^{\mathrm{T}}(q) > 0$, however, this is not always the case, see Remark 3.8.

With some standard algebraic manipulations, the Lagrange equations (3.1) can be put in a more usual way

$$M(q(t))\frac{\mathrm{d}v}{\mathrm{d}t}(t) + N(q(t),v(t)) = Q(q(t),t) - \nabla V(q(t)) \;, \tag{3.2}$$

where the vector

$$N(q,v) = \left[\frac{1}{2}\sum_{k,l}\frac{\partial M_{ik}}{\partial q_l} + \frac{\partial M_{il}}{\partial q_k} - \frac{\partial M_{kl}}{\partial q_i}, i = 1\dots n\right] \tag{3.3}$$

collects the nonlinear inertial terms, i.e., the gyroscopic accelerations.

If we allow one to introduce nonlinear interactions between bodies of the systems and external applied forces which do not derive from a potential, we will use the following more general form for the equation of motion:

$$M(q(t))\frac{\mathrm{d}v}{\mathrm{d}t}(t) + N(q(t),v(t)) + F_{\mathrm{int}}(t,q(t),v(t)) = F_{\mathrm{ext}}(t) \;, \tag{3.4}$$

where

- $F_{\mathrm{int}} : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$ collects the nonlinear interactions between bodies, called also the internal forces which are not necessarily derived from a potential.
- $F_{\mathrm{ext}} : \mathbb{R} \rightarrow \mathbb{R}^n$ collects all the external applied loads.

It is noteworthy that the dynamics of deformable continuum media, discretized in space, for instance by a finite element method, can be cast into such a formulation, see Sect. 3.4. In the sequel, the nonlinear inertial terms will be integrated to F_{int} to lighten the notation.

A Particular Case: Linear Time-Invariant Systems

In this case, the operators defined above are linear time invariant (LTI):

- $M(q) = M \in \mathbb{R}^{n\times n}$ is the mass matrix.
- $F_{\mathrm{int}}(t,q,v) = Cv + Kq$ where $C \in \mathbb{R}^{n\times n}$ is the viscosity matrix and $K \in \mathbb{R}^{n\times n}$ is the stiffness matrix.

When the mass matrix is a constant, then the nonlinear inertial torques are zero, see the expression of $N(q,v)$ above. The geometrical meaning is that the configuration manifold of the system has curvature zero. For instance a double pendulum in the plane has its configuration manifold that is a torus. Necessarily its mass matrix is configuration dependent. In a kinematic chain where all the joints are prismatic, the mass matrix is constant: the configuration space is $\mathbb{R}^n$.

Time Boundary Conditions

The boundary conditions are given for an initial value problem (IVP) as

$$t_0 \in I\!R, \quad q(t_0) = q_0 \in I\!R^n, \quad v(t_0) = v_0 \in I\!R^n \;, \tag{3.5}$$

and for a boundary value problem (BVP):

$$(t_0, T) \in I\!R \times I\!R, \quad \Gamma(q(t_0), v(t_0), q(T), v(T)) = 0 \;. \tag{3.6}$$

3.1.1 Perfect Bilateral Constraints

When a multibody system is considered, some relationships (like mechanical constraints) between the variables are usually imposed between bodies which constrain the dynamics of the system. These relationships can be of various types: boundary conditions or joints are one of them. There are two ways of considering such relationships. The first one is to take them into account through the parameterization, reducing in this way the number of degrees of freedom (this is a procedure often used for the design of feedback control of mechanical systems). This is particularly well suited for boundary conditions of linear constraints. The other way is to consider bilateral constraints and an associated set of Lagrange multipliers.

Let us consider a set of m bilateral constraints on the generalized coordinates:

$$h_j(q,t) = 0, \quad j \in \{1 \ldots m\} \;, \tag{3.7}$$

where the functions $h_j(\cdot)$ are sufficiently smooth with regular gradients, $\nabla_q h_j(\cdot,\cdot)$. The function $h : [0,T] \times \mathrm{I\!R}^n \to \mathrm{I\!R}^m$ is defined as the vector collecting the functions $h_j(\cdot)$,

$$h(q,t) = [h_1(q,t), \ldots, h_m(q,t)]^{\mathrm{T}} \;. \tag{3.8}$$

The bilateral constraints define the configuration manifold $\mathscr{M}(t)$, in which the system must evolve:

$$\mathscr{M}(t) = \left\{ q(t) \in I\!R^n \;\mid\; h_j(q,t) = 0, \quad j \in \{1 \ldots m\} \right\} \;. \tag{3.9}$$

These bilateral constraints are usually enforced by a set of Lagrange multipliers, $\mu \in I\!R^m$. Therefore, the equations of motion are given by

$$M(q(t))\frac{\mathrm{d}v}{\mathrm{d}t}(t) + N(q(t), v(t)) + F_{\mathrm{int}}(t, q(t), v(t)) = F_{\mathrm{ext}}(t) + \nabla_q h(q(t), t)\,\mu \;, \tag{3.10}$$

where the terms $\nabla_q h(q,t)\,\mu$ represent the generalized forces or generalized reactions due to the constraints.

This description of holonomic bilateral constraints can be a little generalized by introducing the tangent space to the manifold $\mathscr{M}$ at q

$$T_{\mathscr{M}}(q) = \{\xi \in \mathrm{I\!R}^n \;\mid\; \nabla_q h^{\mathrm{T}}(q,t)\xi = 0\} \tag{3.11}$$

and the normal space as the orthogonal to the tangent space[1]

$$N_{\mathscr{M}}(q) = \{\eta \in \mathbb{R}^n \mid \eta^{\mathrm{T}}\xi = 0, \forall \xi \in T_{\mathscr{M}}\} . \tag{3.12}$$

It is noteworthy that the linearly independent rows of the gradient $\nabla_q h(q,t)$ form a basis of $N_{\mathscr{M}}(q)$. The bilateral holonomic constraints are said to be perfect if the multipliers μ satisfy the following inclusion:

$$r = \nabla_q h(q,t)\,\mu \in N_{\mathscr{M}}(q) . \tag{3.13}$$

We will see in the sequel that this formulation in terms of an inclusion is very useful in practice. We will also omit the term $\nabla_q h(q,t)\,\mu$ that corresponds to the bilateral constraints for the sake of simplicity and because the main concern of this book is about unilateral constraints.

Remark 3.1. One usually writes $r = -\nabla_q h(q,t)\,\mu$ so that all the gradients that enter the dynamics have the same sign. In the bilateral case since the multiplier μ is not signed this is not important.

3.1.2 Perfect Unilateral Constraints

In the Lagrangian setting, the unilateral constraints are usually described by a set of ν inequalities

$$g^{\alpha}(q,t) \geqslant 0, \quad \alpha \in \{1 \dots \nu\} , \tag{3.14}$$

where the functions $g^{\alpha}(\cdot)$ are assumed to be sufficiently smooth with regular gradients. The function $g : \mathbb{R}^n \times [0,T] \to \mathbb{R}^{\nu}$ is defined as the vector collecting the functions $g^{\alpha}(\cdot)$,

$$g(q,t) = \left[g^1(q,t), \dots, g^{\nu}(q,t)\right]^{\mathrm{T}} . \tag{3.15}$$

These unilateral constraints define the subset $\mathscr{C}(t)$ of the configuration space where the system is constrained to evolve

$$\mathscr{C}(t) = \{q \in \mathscr{M}(t) \mid g^{\alpha}(q,t) \geqslant 0, \alpha \in \{1 \dots \nu\}\} \tag{3.16}$$

As for the bilateral constraints, the unilateral constraints are enforced in the equations of motion by a set of Lagrange multipliers $\lambda \in I\!R^{\nu}$ such that the equation of motion is given by

$$M(q(t))\frac{\mathrm{d}v}{\mathrm{d}t}(t) + N(q(t),v(t)) + F_{\mathrm{int}}(t,q(t),v(t)) = F_{\mathrm{ext}}(t) + \nabla_q g(q(t),t)\,\lambda \tag{3.17}$$

where the vector $\lambda \in \mathbb{R}^{\nu}$ collects the components λ_{α},

$$\lambda = [\lambda_1, \lambda_2, ..., \lambda_{\nu}]^{\mathrm{T}} . \tag{3.18}$$

[1] A metric based on the mass matrix is also habitually used.

The vector $n_\alpha(q,t) = \nabla_q g^\alpha(q,t)$ is a normal vector (not necessarily unit) to the surface $\partial\mathscr{C}(t)$ directed toward the admissible region $\mathscr{C}(t)$.

In a perfect unilateral constraint setting, it is assumed that the reaction force lies along the normal vectors. Finally, when the function $g^\alpha(\cdot,\cdot)$, is positive, the corresponding reaction force must be zero, which leads to the following complementarity condition (the so-called Signorini condition):

$$g^\alpha(q,t) \geqslant 0, \quad \lambda_\alpha \geqslant 0, \quad \lambda_\alpha g^\alpha(q,t) = 0, \quad \alpha \in \{1\dots\nu\} \tag{3.19}$$

which will be denoted in the sequel as

$$0 \leqslant g(q,t) \perp \lambda \geqslant 0\,. \tag{3.20}$$

The vector inequalities in (3.20) have to be understood component-wise.

In a more general way, the outward normal cone to the set $\mathscr{C}(t)$ is defined as

$$N_{\mathscr{C}(t)}(q(t)) = \{y \in \mathrm{IR}^n \mid y = -\textstyle\sum_\alpha \lambda_\alpha \nabla g^\alpha(q,t),\ \lambda_\alpha \geqslant 0, \text{ for all } \alpha \text{ such that}$$
$$g^\alpha(q,t) = 0\}\ . \tag{3.21}$$

Defining the generalized force $r \in \mathrm{IR}^n$ corresponding to the unilateral constraints as

$$r = \sum_\alpha \nabla_q g^\alpha(q,t)\,\lambda_\alpha \tag{3.22}$$

or more compactly as

$$r = \nabla_q g(q,t)\,\lambda \tag{3.23}$$

the complementarity condition can be formulated as an inclusion into the normal cone:

$$-r \in N_{\mathscr{C}(t)}(q(t)) \tag{3.24}$$

Remark 3.2. Under the constraint qualification: for all $x \in \mathscr{C}(t)$, there exists $d \in I\!R^n$ such that $\nabla g^{\alpha,\mathrm{T}}(q,t)d > 0$ for all α such that $g^\alpha(q,t) = 0$, then the normal cone in (3.21) and the normal cone of convex analysis $N_{\mathscr{C}(t)}(q(t)) = \{s \in I\!R^n \mid s^\mathrm{T}(y - q(t)) \leqslant 0 \text{ for all } y \in \mathscr{C}(t)\}$ are equal.

Remark 3.3. It is sometimes discussed whether or not it is preferable to eliminate the constraints by reducing the number of coordinates, or to keep the coordinates but add Lagrange multipliers (Baraff, 1996). This is a sound question when bilateral constraints are considered. When unilateral constraints come into play, however, one has no choice. The Lagrange multipliers are mandatory, otherwise one would spend one's time switching the coordinates. Indeed it is noteworthy that complementarity systems may live on lower-dimensional subspaces.

Notice that we have not yet introduced any nonsmoothness in the dynamics. This will be done later.

3.1.3 Smooth Dynamics as an Inclusion

Combining the equation of motion (3.17) and the inclusion (3.24), the smooth dynamics of a unilaterally constrained Lagrangian dynamical system can be written as the following differential inclusion:

$$-M(q(t))\frac{\mathrm{d}v}{\mathrm{d}t}(t)+N(q(t),v(t))+F_{\mathrm{int}}(t,q(t),v(t))-F_{\mathrm{ext}}(t)\in N_{\mathscr{C}(t)}(q(t))\,. \quad (3.25)$$

As we said in Chap. 2, a huge amount of work has been published in the literature on DIs, but this kind of inclusion is very particular for two main reasons:

- The right-hand side is neither bounded and then nor compact. This yields a UDI.
- The inclusion and the constraints concern the second-order time derivative of q, i.e., the acceleration. This fact leads to strong difficulties, and consequently tools for UDI based on monotone set-valued operator and first-order sweeping process cannot be used.

Such kind of inclusions yields in most of the cases a nonsmooth evolution where the velocity may have jumps, and therefore the acceleration cannot be defined in the usual sense. In Sects. 3.5 and 3.6, we will describe briefly some works that tackle the nonsmooth problem in its integrality, i.e., a UDI on the second-order derivative with a nonsmooth evolution as it has been developed in Schatzman (1978), Moreau (1988b, 1983), Monteiro Marques (1993), and Kunze & Monteiro Marqués (2000).

3.2 The Newton–Euler Formalism

This section is dedicated to present another way to derive the dynamical equations of a mechanical system, sometimes called the vectorial dynamics. The screw formalism is chosen. We recall basic definitions and results from kinematics, kinetics, and dynamics of a rigid body, or a system of rigid bodies (including particles). The presentation is a little bit sketchy, but a complete exposition of the Newton–Euler mechanics is outside the scope of this book. More details may be found in McCarthy (1990), Glocker (2001), and Arnold (1989). In this section only the smooth dynamics is considered. Contact problems are tackled in Sect. 3.9.

3.2.1 Kinematics

We denote as (ξ) a 3-dimensional Euclidean space of points ($(\xi)\equiv I\!R^3$), and (E) the associated linear space of vectors ($(E)\equiv I\!R^3$). Let O be a point of (ξ) and $(\mathbf{i},\mathbf{j},\mathbf{k})$ a basis of (E). Then $\mathscr{R}=(O,\mathbf{i},\mathbf{j},\mathbf{k})$ is a coordinate system (in short, a c.s.). To each point $M\in(\xi)$ we associate three reals a,b,c such that $OM=a\mathbf{i}+b\mathbf{j}+c\mathbf{k}$. In short $OM=(a,b,c)^{\mathrm{T}}$. Usually $\mathscr{R}$ is an orthonormal and direct [2] c.s., i.e., $||\mathbf{i}||=1$, $||\mathbf{j}||=1$, $||\mathbf{k}||=1$, $\mathbf{i}^{\mathrm{T}}\mathbf{j}=\mathbf{i}^{\mathrm{T}}\mathbf{k}=\mathbf{j}^{\mathrm{T}}\mathbf{k}=0$, $\mathbf{i}\times\mathbf{j}=\mathbf{k}$, $\mathbf{k}\times\mathbf{i}=\mathbf{j}$, $\mathbf{j}\times\mathbf{k}=\mathbf{i}$. For a rigid body evolving

[2] i.e., right handed.

in a 3-dimensional space, one usually considers two c.s.: $\mathscr{R}_0$ a Galilean c.s. and $\mathscr{R}$ a c.s. attached to the body, fixed to the body, and moving in $\mathscr{R}_0$. But, this is not an obligation.

Position, Velocity, Acceleration

The position of a point $M \in (\xi)$ in $\mathscr{R}$, parameterized with time, is the vector $OM(t)$. The trajectory of M in a c.s. $\mathscr{R}$ is the set of points fixed in $\mathscr{R}$, with which M coincides along its motion in $\mathscr{R}$. The velocity of M in $\mathscr{R}$ at t is

$$V(M/\mathscr{R},t) = \left[\frac{\mathrm{d}}{\mathrm{d}t}OM(t)\right]_{\mathscr{R}} . \tag{3.26}$$

The vector $V(M/\mathscr{R},t)$ is a vector tangent to the trajectory of M in $\mathscr{R}$. The acceleration of M in $\mathscr{R}$ is

$$\Gamma(M/\mathscr{R},t) = \left[\frac{\mathrm{d}^2}{\mathrm{d}t^2}OM(t)\right]_{\mathscr{R}} = \left[\frac{\mathrm{d}}{\mathrm{d}t}V(M/\mathscr{R},t)\right]_{\mathscr{R}} . \tag{3.27}$$

Angular Velocity

Let $\mathscr{R}_0 = (O_0, \mathbf{i}_0, \mathbf{j}_0, \mathbf{k}_0)$ and $\mathscr{R}_1 = (O_1, \mathbf{i}_1, \mathbf{j}_1, \mathbf{k}_1)$ be two c.s. Then

$$\begin{cases} \left(\dfrac{\mathrm{d}\mathbf{i}_1}{\mathrm{d}t}\right)_{\mathscr{R}_0} = \Omega(\mathscr{R}_1/\mathscr{R}_0) \times \mathbf{i}_1 \\ \left(\dfrac{\mathrm{d}\mathbf{j}_1}{\mathrm{d}t}\right)_{\mathscr{R}_0} = \Omega(\mathscr{R}_1/\mathscr{R}_0) \times \mathbf{j}_1 \\ \left(\dfrac{\mathrm{d}\mathbf{k}_1}{\mathrm{d}t}\right)_{\mathscr{R}_0} = \Omega(\mathscr{R}_1/\mathscr{R}_0) \times \mathbf{k}_1 \end{cases} \tag{3.28}$$

and $\Omega(\mathscr{R}_1/\mathscr{R}_0) = p\mathbf{i}_1 + q\mathbf{j}_1 + r\mathbf{k}_1 = \begin{pmatrix} p \\ q \\ r \end{pmatrix}_{\mathscr{R}_1}$ is the rotational vector (or angular velocity) of $\mathscr{R}_1$ with respect to $\mathscr{R}_0$. It is unique.

Let a vector $u = \alpha(t)\mathbf{i}_1 + \beta(t)\mathbf{j}_1 + \gamma(t)\mathbf{k}_1$. Then

$$\left(\frac{\mathrm{d}u}{\mathrm{d}t}\right)_{\mathscr{R}_0} = \left(\frac{\mathrm{d}u}{\mathrm{d}t}\right)_{\mathscr{R}_1} + \Omega(\mathscr{R}_1/\mathscr{R}_0) \times u \tag{3.29}$$

where $\left(\frac{\mathrm{d}u}{\mathrm{d}t}\right)_{\mathscr{R}_1} = \dot{\alpha}(t)\mathbf{i}_1 + \dot{\beta}(t)\mathbf{j}_1 + \dot{\gamma}(t)\mathbf{k}_1$, $\left(\frac{\mathrm{d}u}{\mathrm{d}t}\right)_{\mathscr{R}_0} = \dot{\alpha}_0(t)\mathbf{i}_1 + \dot{\beta}_0(t)\mathbf{j}_1 + \dot{\gamma}_0(t)\mathbf{k}_1$ for some $\alpha_0(\cdot)$, $\beta_0(\cdot)$, $\gamma_0(\cdot)$, and $\Omega(\mathscr{R}_1/\mathscr{R}_0) \times u = \alpha(t)\left(\frac{\mathrm{d}\mathbf{i}_1}{\mathrm{d}t}\right)_{\mathscr{R}_0} + \beta(t)\left(\frac{\mathrm{d}\mathbf{j}_1}{\mathrm{d}t}\right)_{\mathscr{R}_0} + \gamma(t)\left(\frac{\mathrm{d}\mathbf{k}_1}{\mathrm{d}t}\right)_{\mathscr{R}_0}$. The term $\left(\frac{\mathrm{d}u}{\mathrm{d}t}\right)_{\mathscr{R}_0}$ represents the time derivative as observed from $\mathscr{R}_0$, while the term $\left(\frac{du}{dt}\right)_{\mathscr{R}_1}$ represents the time derivative as observed from $\mathscr{R}_1$.

Example 3.4. Let $\mathbf{k}_1 = \mathbf{k}_2 = \mathbf{k}$, $O_0 = O_1 = O$, and let $\mathscr{R}_1$ have a rotational velocity $\dot{\theta}(t)\mathbf{k}$ w.r.t. $\mathscr{R}_0$ that is fixed. Let a point M be fixed in $\mathscr{R}_1$, i.e., $u = OM = \alpha\mathbf{i}_1 + \beta\mathbf{j}_1$. Obviously $\left(\frac{\mathrm{d}u}{\mathrm{d}t}\right)_{\mathscr{R}_1} = 0$. Then $\left(\frac{\mathrm{d}u}{\mathrm{d}t}\right)_{\mathscr{R}_0} = \dot{\theta}(t)[\mathbf{k} \times (\alpha\mathbf{i}_1 + \beta\mathbf{j}_1) = \dot{\theta}(t)[\alpha\mathbf{j}_1 - \beta\mathbf{i}_1]$.

Let $\mathscr{R}_0$, $\mathscr{R}_1$, and $\mathscr{R}_2$ be three c.s. Then $\Omega(\mathscr{R}_2/\mathscr{R}_0) = \Omega(\mathscr{R}_2/\mathscr{R}_1) + \Omega(\mathscr{R}_1/\mathscr{R}_0)$. More generally the relation of Chasles states that $\Omega(\mathscr{R}_n/\mathscr{R}_0) = \sum_{i=1}^{n} \Omega(\mathscr{R}_i/\mathscr{R}_{i-1})$.

Composition of Velocities

Let M have a motion in $\mathscr{R}_1$ and let $\mathscr{R}_1$ be moving with respect to $\mathscr{R}_0$. Then

$$V(M/\mathscr{R}_0) = V(M/\mathscr{R}_1) + V(O_1/\mathscr{R}_0) + \Omega(\mathscr{R}_1/\mathscr{R}_0) \times O_1M\,. \tag{3.30}$$

This is known as the absolute velocity equal to the relative velocity plus the transferred velocity.

Composition of Accelerations

Let M have a motion in $\mathscr{R}_1$ and let $\mathscr{R}_1$ be moving with respect to $\mathscr{R}_0$. Then

$$\begin{aligned}\Gamma(M/\mathscr{R}_0) = \Gamma(M/\mathscr{R}_1) + \Big\{&\Gamma(O_1/\mathscr{R}_0) + \left[\tfrac{\mathrm{d}}{\mathrm{d}t}\Omega(\mathscr{R}_1/\mathscr{R}_0)\right]_{\mathscr{R}_0} \times O_1M \\ &+\Omega(\mathscr{R}_1/\mathscr{R}_0) \times [\Omega(\mathscr{R}_1/\mathscr{R}_0) \times O_1M]\Big\} \\ &+ 2\Omega(\mathscr{R}_1/\mathscr{R}_0) \times V(M/\mathscr{R}_1).\end{aligned} \tag{3.31}$$

This is known as the absolute acceleration equal to the relative acceleration plus the transferred acceleration plus the Coriolis acceleration. This may be obtained by applying twice (3.29).

Remark 3.5. From the form of the Coriolis acceleration, one deduces that the rotational motion of anticyclones is clockwise in the Northern Hemisphere, and counter clockwise in the Southern Hemisphere. This is due to the fact that the angular velocity of the earth in the Copernic c.s. points outside the ground in the north, and inside the ground in the south. A falling stone always accelerates towards the ground. Thus the angle between the two vectors of the Coriolis acceleration is $< \frac{\pi}{2}$ in the south, and $> \frac{\pi}{2}$ in the north. So the vector product changes its sign from one hemisphere to the other.

Velocity Composition in a Solid

Let (S) be a rigid body (a solid), and let M and N be two points of (S). Let $\mathscr{R}_0$ be a c.s., and $\mathscr{R}$ be a c.s. associated to (S). Thus $V(M,S/\mathscr{R}_0) \overset{\Delta}{=} V(M,\mathscr{R}/\mathscr{R}_0)$. One has

$$V(M,S/\mathscr{R}_0) = V(N,S/\mathscr{R}_0) + MN \times \Omega(S/\mathscr{R}_0)\,. \tag{3.32}$$

The formula (3.32) is called *Varignon's formula*, after the French mathematician Pierre Varignon (1654–1722).

The Kinematic Screw (or Twist)

Varignon's formula leads us to the introduction of the following screw:

$$\mathscr{V}_{M,S/\mathscr{R}_0} = \begin{bmatrix} \Omega(S/\mathscr{R}_0) \\ \\ V(M,S/\mathscr{R}_0) \end{bmatrix}_{M,\mathscr{R}} \tag{3.33}$$

that is usually called the *twist* of the body (S) or the *kinematic screw*. A screw is computed at a point M, and in a coordinate system $\mathscr{R}$. Varignon's formula (3.32) allows one to compute the twist at another point than M. The angular velocity is the resultant of the twist, and it does not vary with the point M at which the twist is calculated. The linear velocity is the moment of the twist that varies along Varignon's formula (3.32).

Rotational Parameterization and Euler Angles

Introducing the twist of a solid leads us to introduce the coordinates of a solid in (ξ). The Euler angles are three angles which allow one to determine the angular position of a rigid body in the 3-dimensional space (ξ). We denote them as ψ, θ, and φ. They correspond to successive rotations of the body (S) (or of a c.s. attached to (S)) around three axis of three successive intermediate c.s.. Let (S) have a rotational motion w.r.t. the c.s. $\mathscr{R}_0$. Then

$$(\mathbf{i}_0,\mathbf{j}_0,\mathbf{k}_0) \overset{(\psi,\mathbf{k}_0)}{\longrightarrow} (\mathbf{u},\mathbf{v},\mathbf{k}_0) \overset{(\theta,\mathbf{u})}{\longrightarrow} (\mathbf{u},\mathbf{w},\mathbf{z}) \overset{(\varphi,\mathbf{z})}{\longrightarrow} (\mathbf{x},\mathbf{y},\mathbf{z}) \, .$$

There are other sets of angles that correspond to other rotations. Of much interest to us is the Olinde–Rodrigues formula:

$$\Omega(S/\mathscr{R}_0) = \begin{pmatrix} p \\ q \\ r \end{pmatrix} = \begin{pmatrix} \sin\varphi\sin\theta & \cos\varphi & 0 \\ \sin\theta\cos\varphi & -\sin\varphi & 0 \\ \cos\theta & 0 & 1 \end{pmatrix} \begin{pmatrix} \dot{\psi} \\ \dot{\theta} \\ \dot{\varphi} \end{pmatrix} \tag{3.34}$$

which relates the derivative of the Euler angles to the angular velocity (also called the instantaneous velocity vector) expressed in the c.s. $\mathscr{R}$ attached to (S). Notice that $\det(\mathscr{M}(\varphi,\theta)) = -\sin\theta$ so that the Olinde–Rodrigues formula is singular at $\theta = k\pi$, $k \in \mathbb{N}$. Moreover $\mathscr{M}(\varphi,\theta)$ is not a Jacobian matrix (because it is not symmetric, see Theorem 12.61), which means that the Olinde–Rodrigues formula is not integrable. This is important when we consider the Lagrange equations of a rigid body. One cannot choose the angular velocity as the derivative of the generalized coordinates.

3.2.2 Kinetics

Kinetics is concerned with inertia operators. The gravity center (or center of mass) of a mechanical system (E) (i.e., a set of particles, rigid bodies moving in (ξ)) with a mass $m_E > 0$ is the unique point $G \in (\xi)$ such that

$$m_E AG = \int_{(E)} AP \, \mathrm{d}m \tag{3.35}$$

for an arbitrary point A. Let (S) be a rigid body. If A is fixed in (S) and $\mathscr{R}$ is fixed w.r.t. (S), then

$$m_S AG = \int_{(S)} AP \, \mathrm{d}m \tag{3.36}$$

is a fixed vector in $\mathscr{R}$. The sums are understood along all points $P \in (S)$ and infinitesimal mass elements dm.

Inertia Matrix of a Mechanical System (E)

Let $\mathscr{R} = (O, \mathbf{i}, \mathbf{j}, \mathbf{k})$ be an orthonormal c.s., the coordinates of points $M \in (E)$ in $\mathscr{R}$ being $(x, y, z)^{\mathrm{T}}$. Then the inertia matrix of (E) in $\mathscr{R}$ is

$$\mathbf{I}(O,E) = \begin{pmatrix} \int_{(E)} (y^2+z^2)\mathrm{d}m & -\int_{(E)} xy \, \mathrm{d}m & -\int_{(E)} xz \, \mathrm{d}m \\ -\int_{(E)} yx \, \mathrm{d}m & \int_{(E)} (x^2+z^2)\mathrm{d}m & -\int_{(E)} yz \, \mathrm{d}m \\ -\int_{(E)} xz \, \mathrm{d}m & -\int_{(E)} yz \, \mathrm{d}m & \int_{(E)} (y^2+x^2)\mathrm{d}m \end{pmatrix} . \tag{3.37}$$

The diagonal terms are the moments of inertia w.r.t. the axis $(O, \mathbf{i})$, $(O, \mathbf{j})$, $(O, \mathbf{k})$, respectively. Clearly if (E) is a solid and the c.s. $\mathscr{R}$ is attached to this body (fixed w.r.t. it), then $\mathbf{I}(O,S)$ is a constant matrix. Let G be the center of mass of the system (E). Then

$$\mathbf{I}(O,E)u = \mathbf{I}(G,E)u + m_E OG \times (u \times OG) \tag{3.38}$$

for all vectors $u \in \mathbb{R}^3$. The moment of inertia of (E) w.r.t. a straight line Δ crossing O (that we may name the axis (O, δ) where $\delta \in \mathbb{R}^3$ is a unit vector of $\varDelta$) is

$$\mathbf{I}_\Delta = \delta^{\mathrm{T}} \mathbf{I}(O,E)\delta \;\; (= \int_{(E)} ||\delta \times OM||^2 \mathrm{d}m) \tag{3.39}$$

Let Δ_a pass through G, and Δ be another line parallel to Δ_a passing through O. The distance between Δ_a and Δ is $d \geqslant 0$. Then

$$\mathbf{I}_\Delta(E) = \mathbf{I}_{\Delta_a}(E) + Md^2 \, , \tag{3.40}$$

which is Huygens' theorem. The moment of inertia w.r.t. a point O is

$$\mathbf{I}_O(E) = I_G(E) + m||OG||^2 \;\; (= \int_{(E)} ||OM||^2 \mathrm{d}m) \, . \tag{3.41}$$

The Kinetic Screw

This is the screw of linear and angular momenta:

$$\mathscr{K}_{A,\mathscr{R}} = \begin{bmatrix} m_E V(G/\mathscr{R}) = \int_{(E)} V(M/\mathscr{R})\mathrm{d}m \\ \sigma(A, E/\mathscr{R}) = \int_{(E)} AM \times V(M/\mathscr{R})\mathrm{d}m \end{bmatrix}_{A,\mathscr{R}} . \tag{3.42}$$

If the system is a rigid body in 3 dimensions, the integrals have to be computed over the body, so that $\int_{(E)}$ is understood as $\int\int\int_{(E)}$. For a rigid body (S), let $O \in (S)$ and be fixed in the c.s. $\mathscr{R}$. Then

$$\sigma(O,S/\mathscr{R}) = \mathbf{I}(O,S)\Omega(S/\mathscr{R})\,. \tag{3.43}$$

By Varignon's formula for screws, we obtain

$$\sigma(A,S/\mathscr{R}) = \mathbf{I}(O,S)\Omega(S/\mathscr{R}) + m_S AO \times \Omega(S/\mathscr{R}) \tag{3.44}$$

for any point $A \in (S)$. From a general point of view, the angular momentum at the gravity center G is given by

$$\sigma(G,S/\mathscr{R}) = \mathbf{I}(G,S)\Omega(S/\mathscr{R})\,. \tag{3.45}$$

The Kinetic Energy of a Solid (S)

Let $O \in (S)$ be fixed in the c.s. $\mathscr{R}$. The kinetic energy of (S) in $\mathscr{R}$ is:

$$\begin{aligned} T(S/\mathscr{R}) &= \tfrac{1}{2}\textstyle\int_{(S)} \|V(M/\mathscr{R})\|^2 \mathrm{d}m \\ &= \frac{1}{2}\Omega^{\mathrm{T}}(S/\mathscr{R})\mathbf{I}(O,S)\Omega(S/\mathscr{R})\,. \end{aligned} \tag{3.46}$$

In general we have

$$T(S/\mathscr{R}) = \frac{1}{2}m_S\|V(G/\mathscr{R})\|^2 + \frac{1}{2}\Omega^{\mathrm{T}}(S/\mathscr{R})\mathbf{I}(G,S)\Omega(S/\mathscr{R})\,, \tag{3.47}$$

where all the quantities have to be expressed in the same c.s. One sees that $T(S/\mathscr{R}) = \frac{1}{2}\mathscr{K}^{\mathrm{T}}_{G,\mathscr{R}}\mathscr{V}_{G,S/\mathscr{R}}$.

3.2.3 Dynamics

Since dynamics concerns the relationship between acceleration and forces, let us first introduce the screw of external actions exerted on a system (E), called the *wrench*:

$$\mathscr{W}_{A,\mathscr{R}} = \begin{bmatrix} F \\ T(A) \end{bmatrix}_{A,\mathscr{R}}\,, \tag{3.48}$$

where $F \in \mathbb{R}^3$ is the vector of external forces acting on (E), $T(A)$ is the external torque at $A \in (E)$ acting on (E). By Varignon's formula we get

$$T(B) = T(A) + BA \times F \tag{3.49}$$

as F is the resultant of the screw and is not changed if the point at which the wrench is calculated changes. If the system (E) is at equilibrium, the fundamental principle of statics says that $\mathscr{W}_{A,\mathscr{R}} = 0$. The wrench includes all types of external actions on (E), like forces created by contact and Coulomb friction, or gravity, or any other effect. In case of contact between two bodies, special care has to be taken concerning the way the actions are written. We shall come back on this later.

The Dynamic Screw

The screw of acceleration, or *dynamic screw* of a system (E), is defined as

$$\mathscr{D}_{A,\mathscr{R}} = \begin{bmatrix} m_E\Gamma(G/\mathscr{R}) \\ \\ \delta(A,E/\mathscr{R}) = \int_{(E)} AM \times \Gamma(M/\mathscr{R})\mathrm{d}m \end{bmatrix}_{A,\mathscr{R}} . \tag{3.50}$$

Let A be an arbitrary point in (E). Then the dynamic moment and the angular momentum are related as:

$$\delta(A,E/\mathscr{R}) = \frac{\mathrm{d}}{\mathrm{d}t}[\sigma(A,E/\mathscr{R})]_{\mathscr{R}} + m_E V(A/\mathscr{R}) \times V(G/\mathscr{R}) \,, \tag{3.51}$$

where the second term of the right-hand side vanishes if A is fixed in $\mathscr{R}$ or if $A = G$. The dynamic moment of a solid (S) is computed at the gravity center G and using (3.45) as:

$$\delta(G,S/\mathscr{R}) = \frac{\mathrm{d}}{\mathrm{d}t}[\mathbf{I}(G,S)\Omega(S/\mathscr{R})]_{\mathscr{R}} \,. \tag{3.52}$$

One deduces that when $\mathscr{R}$ is attached to (S) so that $\mathbf{I}(G,S)$ is a constant 3×3 matrix, then

$$\delta(G,S/\mathscr{R}) = \mathbf{I}(G,S)\frac{\mathrm{d}}{\mathrm{d}t}[\Omega(S/\mathscr{R})]_{\mathscr{R}} \,. \tag{3.53}$$

Change of Coordinates in Screws

Suppose that the system under investigation is studied in a base Galilean c.s. $\mathscr{R}_0$, and that a c.s. $\mathscr{R}$ is attached to the system. Let A be the 3×3 rotation matrix and $d \in \mathbb{R}^3$ the translation vector, which define the transformation of coordinates from $\mathscr{R}$ to $\mathscr{R}_0$: $X = Ax + d$, where X denote the coordinates in $\mathscr{R}_0$ and x the coordinates in $\mathscr{R}$. Such a transformation is sometimes called a *spatial displacement* or a *motion*. Let $\mathscr{S} = \begin{pmatrix} r \\ m \end{pmatrix} \in \mathbb{R}^6$ be a screw expressed at a point and in $\mathscr{R}$. Then when expressed in $\mathscr{R}_0$ the screw $\mathscr{S}$ becomes the screw $\mathscr{S}' = \begin{pmatrix} R \\ M \end{pmatrix} = \begin{pmatrix} A & 0_{3\times3} \\ DA & A \end{pmatrix} \begin{pmatrix} r \\ m \end{pmatrix}$, with $D = \begin{pmatrix} 0 & -d_3 & d_2 \\ d_3 & 0 & -d_1 \\ -d_2 & d_1 & 0 \end{pmatrix}$, the skew-symmetric matrix of the vector product. So we get $R = Ar$ and $M = d \times Ar + Am$.

The motion of a body is represented by trajectories of $\mathscr{R}$ in $\mathscr{R}_0$. When the body moves, equivalently $\mathscr{R}$ moves. Suppose that a spatial displacement (A,d) is applied to both $\mathscr{R}$ and $\mathscr{R}_0$, which are transformed into $\mathscr{R}'$ and $\mathscr{R}'_0$. The screws associated to the body under study are transformed as indicated above. If the spatial displacement is just a rotation ($d = 0$), then the new components of the screws generically are $R = Ar$ and $M = Am$. If it is just a translation ($A = 0$) then $R = r$ and $M = d \times r + m$. We recover here Varignon's formula which is at the base of coordinate change for screws, and indicates that the resultant does not depend on the point at which the screw is calculated, whereas the moment does depend on it.

The Fundamental Principle of Dynamics

$$\mathscr{D}_{\mathrm{A},\mathscr{R}} = \mathscr{W}_{\mathrm{A},\mathscr{R}} \,. \tag{3.54}$$

It is clear that this dynamical equilibrium may be written at any point A and in any c.s. $\mathscr{R}$. Let the system under study be a rigid body (S), and let the c.s. $\mathscr{R}$ be fixed with respect to the body and $A = G$ (the center of mass). One obtains the so-called Newton–Euler equations for the motion in a Galilean c.s. $\mathscr{R}_0$. Newton's dynamics is in $\mathscr{R}_0$:

$$m_S\Gamma(G/\mathscr{R}_0) = F \,. \tag{3.55}$$

The dynamic equilibrium when applied in a moving c.s. $\mathscr{R}$ with the same origin O_0 as the one of $\mathscr{R}_0$ and for a constant rotational velocity $\Omega(\mathscr{R}/\mathscr{R}_0)$ has to incorporate the inertial forces which take into account the fact that $\mathscr{R}$ is not a Galilean c.s. One obtains instead of (3.55):

$$m_S\Gamma(G/\mathscr{R}) = F - 2m_S\Omega(\mathscr{R}/\mathscr{R}_0) \times \left(\frac{\mathrm{d}u}{\mathrm{d}t}\right)_{\mathscr{R}} - m_S\Omega(\mathscr{R}/\mathscr{R}_0) \times (\Omega(\mathscr{R}/\mathscr{R}_0) \times u) \tag{3.56}$$

with $u = O_0G$. The two inertial forces that appear in the right-hand side of (3.56) are the Coriolis and centrifugal forces. These are the inertial forces that someone moving on the body would experience.

Example 3.6. Let us consider a particle M with mass $m > 0$, mounted at the edge of a massless rod of length $l > 0$ that rotates in the plane $(O_0, \mathbf{i}_0, \mathbf{j}_0)$. The c.s. $(O_0, \mathbf{i}_1, \mathbf{j}_1, \mathbf{k})$ is attached to the particle–rod system, and the angle of rotation is $\theta(\cdot)$. Then $\Gamma(M/\mathscr{R}_1) = 0$, $\Omega(\mathscr{R}_1/\mathscr{R}_0) = \dot{\theta}(t)\mathbf{k}$, $u = l\mathbf{i}_1$. Thus one gets from (3.56): $0 = ml\dot{\theta}^2 + T$, where T is the tension in the rod, along $(O_0, \mathbf{i}_1)$. We retrieve the well-known fact that the centrifugal force is balanced by the tension in the rod.

From (3.53) and (3.29) we obtain Euler's equations

$$\mathbf{I}(G,S)\frac{\mathrm{d}}{\mathrm{d}t}[\Omega(S/\mathscr{R}_0)] + \Omega(S/\mathscr{R}_0) \times \mathbf{I}(G,S)\Omega(S/\mathscr{R}_0) = T(G) \tag{3.57}$$

where we have dropped the subscript $\mathscr{R}$ but it is understood that all the vectors are expressed and computed in $\mathscr{R}$. The vector product reflects the rotation of $\mathscr{R}$ w.r.t. $\mathscr{R}_0$.

Remark 3.7. Despite the inertia matrix in (3.57) is a constant, nonlinear inertial torques act on the system, represented in $\Omega(S/\mathscr{R}_0) \times \mathbf{I}(G,S)\Omega(S/\mathscr{R}_0)$. This is in contrast with the Lagrange equations (3.2), in which a constant mass matrix implies no nonlinear inertial forces (see (3.3)). However, Euler equations (3.57) are not Lagrange equations, because the instantaneous rotation $\Omega(S/\mathscr{R}_0)$ usually is not the derivative of the body's coordinates with the Euler angles (see (3.34)).

Relationships Between Newton–Euler Mechanics and Lagrange Equations

Let us consider a solid (S) moving in (ξ). One choice for its generalized coordinates is a 6-dimensional vector $(x, y, z, \psi, \theta, \varphi)^{\mathrm{T}}$ where (x, y, z) are the coordinates of the mass center in some Galilean c.s. $\mathscr{R}_0$, and (ψ,θ,φ) are the Euler angles. This is a minimal parameterization. Let us denote $\chi=(\psi,\theta,\varphi)^{\mathrm{T}}$. Using (3.34), i.e., $\Omega(S/\mathscr{R}_0)=\mathscr{M}(\phi,\theta)\dot{\chi}$, and (3.57) one obtains

$$\mathscr{M}^{\mathrm{T}}\mathbf{I}(G,S)\mathscr{M}\ddot{\chi}+\mathscr{M}^{\mathrm{T}}\mathbf{I}(G,S)\dot{\mathscr{M}}\dot{\chi}+\mathscr{M}^{\mathrm{T}}\mathscr{M}\dot{\chi}\times\mathscr{M}^{\mathrm{T}}\mathbf{I}(G,S)\mathscr{M}\dot{\chi}=\mathscr{M}^{\mathrm{T}}T(G)\,, \tag{3.58}$$

where all the arguments are dropped for convenience. The Euler dynamics is now under a Lagrangian formalism. One sees that the mass matrix is no longer constant, as expected. However, it has singularities, since the Olinde–Rodrigues matrix is singular. Concatenating (3.58) with (3.55) one gets the Lagrange dynamics for the body moving in the Galilean c.s. $\mathscr{R}_0$. It is noteworthy that one may perform any other diffeomorphic (generalized) coordinate transformation $z=Z(x,y,z,\chi)$ to rewrite the obtained Lagrangian dynamics, mixing translational and rotational coordinates if needed. Doing so one realizes that the Lagrange equations and the Newton–Euler equations really pertain to different worlds.

Some Comments on Newton, Euler and Lagrange Dynamics

It is a common thought that one may derive the dynamics of a system using any of these three approaches. As shown in Antman (1998) things are more subtle. Following Antman (1998) let us study a simple pendulum (Fig. 3.1), which will also provide us with the opportunity to illustrate some of the above developments of classical mechanics.

We suppose that the coordinates systems $\mathscr{R}_0=(O;\mathbf{i},\mathbf{j},\mathbf{k})$ and $\mathscr{R}_1=(O;\mathbf{e}_1,\mathbf{e}_2,\mathbf{k})$ are right-handed. We have $\mathbf{e}_1=\cos\theta\mathbf{i}+\sin\theta\mathbf{j}$, $\mathbf{e}_2=-\sin\theta\mathbf{i}+\cos\theta\mathbf{j}$, from which $\mathbf{i}=\cos\theta\mathbf{e}_1-\sin\theta\mathbf{e}_2$ and $\mathbf{j}=\sin\theta\mathbf{e}_1+\cos\theta\mathbf{e}_2$. Applying (3.29) to $\left(\frac{\mathrm{d}\mathbf{e}_1}{\mathrm{d}t}\right)_{\mathscr{R}_0}$ we obtain

$$\begin{aligned}\left(\frac{\mathrm{d}\mathbf{e}_1}{\mathrm{d}t}\right)_{\mathscr{R}_0}&=\left(\frac{\mathrm{d}\mathbf{e}_1}{\mathrm{d}t}\right)_{\mathscr{R}_1}+\begin{pmatrix}0\\0\\\dot{\theta}\end{pmatrix}\times\begin{pmatrix}\cos\theta\\\sin\theta\\0\end{pmatrix}\\&=0-\dot{\theta}\sin\theta\mathbf{i}+\dot{\theta}\cos\theta\mathbf{j}\,.\end{aligned}\tag{3.59}$$

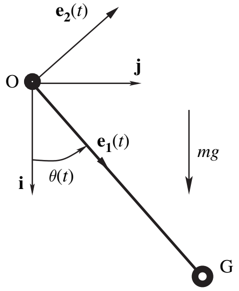


Fig. 3.1. A simple pendulum

Indeed the derivative of $\mathbf{e}_1$ as observed from $\mathscr{R}_1$, i.e., the trivial equality $\mathbf{e}_1 = \mathbf{e}_1$, is zero. Reusing (3.29) for the second-order derivative and noting that $\frac{\mathrm{d}\mathbf{e}_1}{\mathrm{d}t} = \dot{\theta}\mathbf{e}_2$ we obtain

$$\begin{aligned} \left(\frac{\mathrm{d}^2\mathbf{e}_1}{\mathrm{d}t^2}\right)_{\mathscr{R}_0} &= \left(\frac{\mathrm{d}^2\mathbf{e}_1}{\mathrm{d}t^2}\right)_{\mathscr{R}_1} + \begin{pmatrix} 0 \\ 0 \\ \dot{\theta} \end{pmatrix} \times \begin{pmatrix} -\dot{\theta}\sin\theta \\ \dot{\theta}\cos\theta \\ 0 \end{pmatrix} \qquad (3.60) \\ &= \ddot{\theta}\mathbf{e}_2(t) + \dot{\theta}^2\mathbf{e}_1(t)\,. \end{aligned}$$

Therefore the acceleration of the pendulum tip with respect to $\mathscr{R}_0$ is $l\ddot{\theta}\mathbf{e}_2 + l\dot{\theta}^2\mathbf{e}_1$, which can also be expressed in $\mathscr{R}_0$ as $(-l\ddot{\theta}\sin\theta + l\dot{\theta}^2\cos\theta)\mathbf{i} + (l\ddot{\theta}\cos\theta + l\dot{\theta}^2\sin\theta)\mathbf{j}$. Let us assume that a reaction force $h(t) = h_1(t)\mathbf{e}_1(t) + h_2(t)\mathbf{e}_2(t)$ acts at the joint O, and another force $f(t) = f_1(t)\mathbf{e}_1(t) + f_2(t)\mathbf{e}_2(t)$ acts at the gravity center G. The wrench at O is therefore equal to

$$\mathscr{W}_{O,\mathscr{R}_1} = \begin{bmatrix} (h_1(t)+f_1(t)+mg\cos\theta(t))\mathbf{e}_1(t) + (h_2(t)+f_2(t)-mg\sin\theta(t))\mathbf{e}_2(t) \\ T(O) \end{bmatrix}, \qquad (3.61)$$

where the torque $T(O) = h(t) \times OO + f(t) \times GO + mg\mathbf{i} \times GO = (lf_2(t) - mgl\sin\theta(t))\mathbf{k}$. From the fundamental principle of the dynamics we obtain

$$\begin{bmatrix} m\Gamma(G) = ml\ddot{\theta}(t)\mathbf{e}_2(t) + l\dot{\theta}^2(t)\mathbf{e}_1(t) \\ ml^2\ddot{\theta}(t) \end{bmatrix} = \mathscr{W}_{O,\mathscr{R}_1}\,. \qquad (3.62)$$

So we obtain (since $I_G = 0$)

$$\begin{cases} ml\dot{\theta}^2(t) = h_1(t) + f_1(t) + mg\cos\theta(t) \\ ml\ddot{\theta}(t) = h_2(t) + f_2(t) - mg\sin\theta(t) \\ ml^2\ddot{\theta}(t) = lf_2(t) - mgl\sin\theta(t). \end{cases} \qquad (3.63)$$

The first two equations of (3.63) are the Newton's dynamics of the pendulum. The third equation is the Euler's dynamics. It is sometimes said that Euler dynamics is a consequence of Newton dynamics. Clearly this holds only if $h_2(t) = 0$. The question raised by Antman (1998) is: why should this be so? It seems that there exist no fundamental reason that sustains this claim. By further studying a compound pendulum subject to couples at G and O, it is concluded in Antman (1998) that the Euler dynamics is the right way to write the dynamics of such a pendulum, because Newton's dynamics yields mechanically unexplainable equations.

Let us now turn our attention to Lagrange's equations. Let the coordinates of G be x and y in $\mathscr{R}_0$, and those of O be x_o and y_o. From (3.47) the kinetic energy of the free pendulum is $T(\dot{\theta},\dot{x},\dot{y}) = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2)$, as indeed the moment of inertia at G is zero, all the mass being concentrated at G. The potential energy is $U(q) = mgx$. The

Lagrangian function is $L(\theta,\dot{\theta},x,y,\dot{x},\dot{y}) = T(\dot{\theta},\dot{x},\dot{y}) - U(x)$. Since O is fixed, there are two bilateral constraints $x_O = x - l\,\cos\theta = 0$, $y_O = y - l\,\sin\theta = 0$, with which two Lagrange multipliers λ_1 and λ_2 are associated, respectively. Let us choose the generalized coordinates of the free system as $(x,y,\theta)^{\mathrm{T}}$. From (3.10) we can write the Lagrange equations for the pendulum evolving in the plane and subject to the bilateral constraints as

$$\begin{cases} m\ddot{x}(t) = \lambda_1(t) + mg + f_1'(t) \\ m\ddot{y}(t) = \lambda_2(t) + f_2'(t) \\ 0.\ddot{\theta}(t) = l\,\sin\theta\,\lambda_1 - l\,\cos\theta\,\lambda_2 \\ x = l\,\cos\theta,\ \ y = l\,\sin\theta, \end{cases} \tag{3.64}$$

where $f' = (f_1', f_2')^{\mathrm{T}}$ is the force applied at G, expressed in $\mathscr{R}_0$. More generally the vector $Q(\cdot)$ in (3.1) is determined from the principle of virtual work. The work of the external force $f(\cdot)$ acting at G, when G varies its position of δx and δy, is $\delta W = f_1'\delta x + f_2'\delta y$. Thus the vector $Q(\cdot)$ has to satisfy $\delta W = Q^{\mathrm{T}}\delta q$ for an arbitrary variation δq of the generalized coordinates. Equaling both expressions of δW yields (since this is true for arbitrary variations) that $Q = (f_1', f_2', 0)^{\mathrm{T}}$. Differentiating the constraints twice, we obtain $\ddot{x} = -\ddot{\theta}l\,\sin\theta - \dot{\theta}^2 l\,\cos\theta$ and $\ddot{y} = l\ddot{\theta}\,\cos\theta - l\dot{\theta}^2\sin\theta$. One may then calculate the Lagrange multipliers from the first two equations of (3.64) as (dropping the time argument)

$$\begin{cases} \lambda_1 = -ml\ddot{\theta}\,\sin\theta - ml\dot{\theta}^2\cos\theta - mg - f_1' \\ \lambda_2 = ml\ddot{\theta}\,\cos\theta - ml\dot{\theta}^2\sin\theta - f_2'\,. \end{cases} \tag{3.65}$$

Inserting these values into the third equality of (3.64) and using $f_1' = f_1\cos\theta - f_2\sin\theta$ and $f_2' = f_1\sin\theta + f_2\cos\theta$ gives

$$-ml^2\ddot{\theta}(t) - mgl\,\sin\theta(t) + lf_2(t) = 0 \tag{3.66}$$

which is nothing else but the Euler's dynamics in (3.63). It is noteworthy that the only physical assumption that has been done here is that the generalized force corresponding to the constraints is orthogonal to the constrained manifold $\{q \in \mathrm{IR}^3 \mid x = l\,\cos\theta,\ \text{and}\ y = l\,\sin\theta\}$. We conclude that Lagrange's dynamics and Newton's dynamics are not quite equivalent one an other, since one cannot derive the first ones from the second ones without further assumption.

Remark 3.8. Notice that due to the choice of generalized coordinates we made and the fact that the mass is concentrated at G, the obtained mass matrix in (3.64) is singular. The singularity does not come from the constraints. Is this an artifact? There are five unknown functions ($x(\cdot)$, $y(\cdot)$, $\theta(\cdot)$, $\lambda_1(\cdot)$, $\lambda_2(\cdot)$). There are five equations (three in (3.64) plus the two constraints) if the θ-dynamics is in (3.64), only four equations if it is not. So it seems that choosing θ as one of the generalized coordinates, is mandatory if one wants to integrate the system.

3.3 Local Kinematics at the Contact Points

Let us provide some explanations on how the unilateral constraints, and more generally nonsmooth contact laws, may be introduced into the Newton–Euler dynamics and the Lagrange dynamics, starting from what happens locally at the contact points between the bodies that compose the system.

3.3.1 Local Variables at Contact Points

The main issue is to get an expression of the distance between a rigid body (S) and an obstacle, that may compactly be written as $g(x,y,z,\psi,\theta,\varphi) \geqslant 0$ for some function $g(\cdot)$ or between two bodies. To study this point, let us examine the so-called local kinematics at a contacting point (a contacting point is a point of the system that is likely to become a contact point in the near future). By local variables it is meant any variables introduced to describe the physical behavior of the system. In usual mechanical situations these variables go by pairs (velocities, forces). Let O and O' be two contacting bodies, and P, P' the proximal points belonging to the bodies O and O', respectively (Fig. 3.2).

The relative velocity is defined as usual, though being extended to noncontacting bodies. Consider the particles M and M', lying respectively on the bodies O and O', coinciding with the geometrical points P and P'. Let $V(M)$ and $V(M')$ be the velocity vectors of the particles M and M', belonging to the bodies O and O'. The relative velocity of O with respect to O' is defined as,

$$U = V(M) - V(M') \ .$$

The indices $_{\text{N}}$ and $_{\text{T}}$, are used to define normal and tangential components, the index $_{\text{N}}$ standing for components on $\mathbf{n}$, the index $_{\text{T}}$ standing for the pair of components on $\mathbf{t}$, $\mathbf{s}$, in the 3 dimensional case, on $\mathbf{t}$ in the 2 dimensional case.

- $\mathbf{n}$ is the normal vector directed along $P'P$, where P and P' are the proximal points.
- $U = \begin{pmatrix} U_{\text{T}} \\ U_{\text{N}} \end{pmatrix} \in \mathrm{I\!R}^3$ are the components of the relative velocity, U^- and U^+ are the left and right velocities, respectively.
- $R = \begin{pmatrix} R_{\text{T}} \\ R_{\text{N}} \end{pmatrix} \in \mathrm{I\!R}^3$ are the components of the reaction from O' onto O.
- $g(\cdot)$ is the gap function, i.e., the signed distance $\overline{P'P}$.

In order to obtain the gap function as a function of the bodies' coordinates (center of gravity coordinates and Euler angles), one needs to perform an analysis of the bodies' geometries. Usually the bodies' boundaries have to be parameterized with the coordinates, so that the signed distance $\overline{P'P}$ can be written as a function of these coordinates, i.e., $\overline{P'P} = g(x,y,z,x',y',z',\chi,\chi')$. In general this is not an easy task. See Sect. 3.3.3 for insights on the practical calculation in a software package. Some calculations for the nontrivial example of a wheelset on a track are presented in Soellner & Führer (1998, example 5.5.1). The complementarity conditions at the contacting points P and P' are, in the frictionless case where $R_{\text{T}} = 0$,

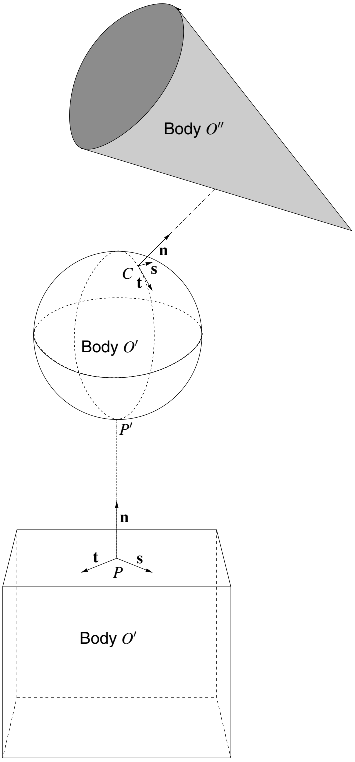


Fig. 3.2. Definition of the local frame

$$0 \leqslant \overline{P'P} \perp R \geqslant 0 \tag{3.67}$$

or equivalently $0 \leqslant \overline{P'P}^{\mathrm{T}} \mathbf{n} \perp R_{\mathrm{N}} \geqslant 0$. Note that U_{T} is the sliding velocity vector which will be used in frictional laws. Usual kinematics yields relations connecting local variables to generalized variables. For instance consider the example of two disks O and O', with radius r and r', velocities at the center of O given by $\dot{q}_1$, $\dot{q}_2$, rotation $\dot{q}_3$, velocities at the center of O' given by $\dot{q}'_1$, $\dot{q}'_2$, rotation $\dot{q}'_3$. The relative velocity is

$$\begin{cases} U_{\mathrm{T}} = (\dot{q}_1 - \dot{q}'_1)t_1 + (\dot{q}_2 - \dot{q}'_2)t_2 + r\dot{q}_3 + r'\,\dot{q}'_3\,, \\ U_{\mathrm{N}} = (\dot{q}_1 - \dot{q}'_1)n_1 + (\dot{q}_2 - \dot{q}'_2)n_2\,. \end{cases} \tag{3.68}$$

Similar relations between the local reaction R and the total momentum exerted on the two bodies may be written. The relations (3.68) appear as linear relations connecting the components U of the relative velocities to the generalized variables $\dot{q}$. In general situations, dealing with a collection of contactors, and a collection of contacts[3] labeled by superscripts $^{\alpha}$, there exists a linear relation relating the relative velocity at

[3] By contact it is meant, either a pair of bodies, the “contactors”, a pair of proximal points, or the geometrical point where proximal points coincide when the gap is null. Contacts

the contact α and the generalized variable,

$$U^{\alpha} = H^{\alpha,\mathrm{T}}(q)\,v\,,\ (v(\cdot) = \dot{q}(\cdot) \text{ almost everywhere})\,. \tag{3.69}$$

There exists also a dual relation relating the representative r^{α} of the local reaction R^{α} for the parameterization q:

$$r^{\alpha} = H^{\alpha}(q)\,R^{\alpha}\,. \tag{3.70}$$

The matrices $H^{\alpha}(q)$ and $H^{\alpha\,\mathrm{T}}(q)$ are transposed linear mappings (in the sense that $v \longrightarrow H^{\alpha\,\mathrm{T}}(q)\,v$, $R \longrightarrow H^{\alpha}(q)\,R^{\alpha}$ are linear, but $q \longrightarrow H^{\alpha,\mathrm{T}}(q)\,v$, $q \longrightarrow H^{\alpha}(q)\,R^{\alpha}$ are not necessarily linear). A last relation, whose writing is often omitted or misunderstood,[4] is

The derivative with respect to time of the gap function $t \longrightarrow g^{\alpha}(t)$ is the normal relative velocity U_{N},

$$\dot{g}^{\alpha}(\cdot) = U_{\mathrm{N}}^{\alpha}(\cdot) = \nabla g^{\alpha,\mathrm{T}}(q)v(\cdot) \tag{3.71}$$

and the second-order derivative of $g^{\alpha}(\cdot)$ is the normal relative acceleration

$$\ddot{g}^{\alpha}(\cdot) = \dot{U}_{\mathrm{N}}^{\alpha}(\cdot) \tag{3.72}$$

that is equal to $\nabla g^{\alpha,\mathrm{T}}(q)\dot{v} + \frac{\mathrm{d}}{\mathrm{d}t}(\nabla g^{\alpha,\mathrm{T}}(q))v$.

This last relation means that the relative acceleration is not equal to $\nabla g^{\alpha,\mathrm{T}}(q)\ddot{q}$ as it is sometimes wrongly written.

Finally, the vectors collecting the components for each contact are defined by

$$U = \begin{bmatrix} U^1 \\ \vdots \\ U^{\nu} \end{bmatrix}, \quad U_{\mathrm{N}} = \left[U_{\mathrm{N}}^1, \ldots, U_{\mathrm{N}}^{\nu}\right]^{\mathrm{T}}, \quad U_{\mathrm{T}} = \begin{bmatrix} U_{\mathrm{T}}^1 \\ \vdots \\ U_{\mathrm{T}}^{\nu} \end{bmatrix}, \tag{3.73}$$

$$R = \begin{bmatrix} R^1 \\ \vdots \\ R^{\nu} \end{bmatrix}, \quad R_{\mathrm{N}} = \left[R_{\mathrm{N}}^1, \ldots, R_{\mathrm{N}}^{\nu}\right]^{\mathrm{T}}, \quad R_{\mathrm{T}} = \begin{bmatrix} R_{\mathrm{T}}^1 \\ \vdots \\ R_{\mathrm{T}}^{\nu} \end{bmatrix}, \tag{3.74}$$

$$g = [g^1, \ldots, g^{\nu}]^{\mathrm{T}}, \quad H(q) = [H^1(q), \ldots, H^{\nu}(q)], \tag{3.75}$$

$$r = \sum_{\alpha} r^{\alpha} = \sum_{\alpha} H^{\alpha}(q)\,R^{\alpha} = H(q)\,R. \tag{3.76}$$

are labeled with Greek letters and an index is a pair of indices labeling the two contacting bodies. This is clear from the picture where bodies are convex and there exists only a pair of contacting points. When "flat" bodies are near each other, there may be a multiplicity of proximal points: the system of labeling must then be more sophisticated. For instance, one may choose some parts of the bodies playing the role of bodies in the system of labeling. Discussing these questions, one encounters into numerical computation and sorting algorithms techniques.

[4] or whose definition derives from considerations on "relative displacements", themselves badly defined, in some incremental approach of kinematics.

The vector $r \in \mathrm{I\!R}^n$ is the vector of reaction forces that enters the Lagrange equations written in the generalized coordinates q. Each $H^{\alpha}(q)$ has dimensions $n \times 3$, so that $H(q)$ has dimensions $n \times 3\nu$. If the system under study is the body O one obtains:

$$M(q(t))\frac{\mathrm{d}v}{\mathrm{d}t}(t) + N(q(t), v(t)) = Q(q(t), t) - \nabla V(q(t)) + H(q(t))R\,, \qquad (3.77)$$

where q is a 6-dimensional vector of generalized coordinates for the body, being comprised of the gravity center coordinates and the Euler angles. If there is only one contact point P at which the body O touches another body, then $R \in \mathrm{I\!R}^3$. In general there may be several such contact points so that $R \in \mathrm{I\!R}^{3\nu}$. The dynamics (3.77) is composed of Newton's equation (3.55) and Euler's equation (3.58), so that

$$M(q) = \begin{pmatrix} m_O I_3 & 0_3 \\ & \\ 0_3 & \mathscr{M}^{\mathrm{T}}\mathbf{I}(G,O)\mathscr{M} \end{pmatrix}.$$

Starting from (3.70) it is easy to write the following expression

$$r^{\alpha} = H_{\mathrm{T}}^{\alpha}(q)R_{\mathrm{T}}^{\alpha} + H_{\mathrm{N}}^{\alpha}(q)R_{\mathrm{N}}^{\alpha} \qquad (3.78)$$

for some $H_{\mathrm{T}}^{\alpha}(q)$ and $H_{\mathrm{N}}^{\alpha}(q)$ with appropriate dimensions (as $R_{\mathrm{T}}^{\alpha} \in \mathrm{I\!R}^2$ and $R_{\mathrm{N}}^{\alpha} \in \mathrm{I\!R}$). Collecting all contacts this allows one to deduce that

$$r = H_{\mathrm{T}}(q)R_{\mathrm{T}} + H_{\mathrm{N}}(q)R_{\mathrm{N}} \qquad (3.79)$$

for some matrices $H_{\mathrm{T}}(q)$ and $H_{\mathrm{N}}(q)$. Inserting this into the Lagrange equations (3.77) one obtains a formulation extensively used in Pfeiffer & Glocker (1996) and Glocker (2001) and popularized by these authors. We have developed the equations for one body ($q(t) \in \mathrm{I\!R}^6$), but the Lagrange dynamics (3.77) may be written for n_b bodies so that $q(t) \in \mathrm{I\!R}^{6n_b}$. Then when there are ν contacting points between the bodies, one forms a matrix $H(q)$ that is $6n_b \times 3\nu$. The Lagrange dynamics (3.77) may also be written for a kinematic chain (like a robot manipulator) which undergoes some unilateral contacts. In this case it is supposed that the vector $q(\cdot)$ is a suitable generalized coordinates vector obtained after a possible elimination of redundant variables due to bilateral constraints (joints).

3.3.2 Back to Newton–Euler's Equations

Consider the dynamic equilibrium of the body O, on which the body O' applies the contact force R at the point P. From (3.55) and (3.58) (the Lagrange equation for the body O), we can choose $q = (x, y, z, \chi^{\mathrm{T}})^{\mathrm{T}}$ as the vector of generalized coordinates. We obtain

$$\begin{cases} m_0\Gamma(G/\mathscr{R}_0) = F_{\mathscr{R}_0} + [I_3\ O_3]H(q)R \\ \\ \mathbf{I}(G,0)\frac{\mathrm{d}}{\mathrm{d}t}[\Omega(0/\mathscr{R}_0)] + \Omega(0/\mathscr{R}_0) \times \mathbf{I}(G,0)\Omega(0/\mathscr{R}_0) = T(G) + \\ \\ \qquad\qquad\qquad\qquad\qquad\qquad + \mathscr{M}^{-\mathrm{T}}[0_3\ I_3]H(q)R. \end{cases} \qquad (3.80)$$

Now if the local reaction R at the contact point P is expressed in the c.s. $\mathscr{R}$ fixed in the body 0 as the vector R_0, we obtain from the principle of dynamics (see also (3.56))

$$\begin{cases} m_0\Gamma(G/\mathscr{R}) = F_{\mathscr{R}} + R_0 \\ \\ \mathbf{I}(G,0)\frac{\mathrm{d}}{\mathrm{d}t}[\Omega(0/\mathscr{R}_0)] + \Omega(0/\mathscr{R}_0) \times \mathbf{I}(G,0)\Omega(0/\mathscr{R}_0) = T(G) + GP \times R_0 \end{cases} \tag{3.81}$$

which can also be expressed at the contact point P using the same c.s. $\mathscr{R}$ translated with $d = GP$. We may also start from (3.77) and use (3.69) to get for each contact α:

$$\begin{aligned} \dot{U}^{\alpha}(t) &= H^{\alpha,\mathrm{T}}(q)M^{-1}(q)H(q)R(t) + \\ &\quad + \dot{H}^{\alpha,\mathrm{T}}(q)v + H^{\alpha,\mathrm{T}}(q)M^{-1}(q)[-N(q,v) + Q(q,t) - \nabla V(q)] \,, \end{aligned} \tag{3.82}$$

where $q = q(t)$ and $v = v(t)$. Now applying (3.82) to the body O in $\mathrm{I\!R}^3$ and supposing that there is a single contact point P (labeled α) gives in short

$$\dot{U}^{\alpha}(t) = W^{\alpha}(q)R^{\alpha}(t) + F^{\alpha}(q,v,t) \,, \tag{3.83}$$

where $W^{\alpha}(q) = H^{\alpha,\mathrm{T}}(q)M^{-1}(q)H^{\alpha}(q)$ is a Delassus' operator. By Varignon's formula and (3.34) we have

$$H^{\alpha,T}(q) = [I_3 \;\; \mathscr{M}_{PG}\mathscr{M}^{-1}(\varphi,\theta)] \tag{3.84}$$

with $\mathscr{M}_{PG}u = PG \times u$ for any $u \in \mathrm{I\!R}^3$. Note that

$$M^{-1}(q) = \begin{pmatrix} \frac{1}{m_S} I_3 & 0_3 \\ \\ 0_3 & [\mathscr{M}^{\mathrm{T}}(\varphi,\theta)\mathbf{I}(G,S)\mathscr{M}(\varphi,\theta)]^{-1} \end{pmatrix} .$$

so that $W^{\alpha}(q)$ can be computed from φ and θ, the Euler angles.

Suppose now that there are $\nu \geqslant 2$ contacts, n_b bodies, and consider (3.82). Collecting all the equations (3.83) for the contacts $1,\ldots,\nu$, one obtains a compact formulation

$$\dot{U}(t) = W(q)R(t) + F(q,v,t) \,, \tag{3.85}$$

where $W(q) = H^{\mathrm{T}}(q)M^{-1}(q)H(q)$ is a Delassus' operator of dimension $3\nu \times 3\nu$. It will be seen in Chaps. 8 and 10 that depending on the type of algorithm that is implemented (event-driven or time-stepping), the Delassus' operator may be modified for numerical integration purposes.

Remark 3.9 (Solvability at impacts). At an impact time t, one gets ν equations $U_{\mathrm{N}}^{\alpha}(t^+) = -e_{\alpha}U_{\mathrm{N}}^{\alpha}(t^-)$, 3ν equations $U(t^+) - U(t^-) = W(q(t))P$, and 4ν unknowns: the 3ν-dimensional post-impact velocity $U(t^+)$ and the ν components P_{N}^{α} of the percussion vector. It is noteworthy that the reactions are supposed to be normal to the α tangent planes at each contact point, but the tangential velocities are not necessarily continuous at impacts! Inertial couplings may induce tangential velocity jumps despite there no friction.

Notice that we have obtained (3.85) starting with the arrangement for U in (3.73). If instead one forms U and R as $U^{\mathrm{T}} = (U_{\mathrm{T}}^{\mathrm{T}}, U_{\mathrm{N}}^{\mathrm{T}})^{\mathrm{T}}$, and $R^{\mathrm{T}} = (R_{\mathrm{T}}^{\mathrm{T}}, R_{\mathrm{N}}^{\mathrm{T}})^{\mathrm{T}}$, then a similar formulation as (3.85) is obtained. However, we may then decompose the matrix $W(q)$ as

$$W(q) = \begin{pmatrix} W_{\mathrm{TT}}(q) & W_{\mathrm{TN}}(q) \\ \\ W_{\mathrm{NT}}(q) & W_{\mathrm{NN}}(q) \end{pmatrix} . \tag{3.86}$$

Therefore one can write $\dot{U}_{\mathrm{N}}(t) = W_{\mathrm{NN}}(q)R_{\mathrm{N}}(t) + F_{\mathrm{N}}(q,v,t)$ as long as $R_{\mathrm{T}} = 0$ (the frictionless case). In order to construct a LCP whose solution is the normal contact force R_{N}, one needs to know which contacts are activated, i.e., which indices $\alpha \in \{1,...,\nu\}$ satisfy $g^{\alpha}(q) \leqslant 0$. Numerically one will have to forecast the contacts and construct a set of forecasted indexes.

3.3.3 Collision Detection and the Gap Function Calculation

This section is a short summary of an important module of any software package: the management of contacts status. Approximation of the shapes and approximate calculation of impact times are generally CPU-time-intensive tasks (Eberhard, 1999). Many works have been dedicated to collision detection, e.g., von Herzen et al. (1990), Mirtich (1997), Hubbard (1996), and Ponamgi et al. (1995) to cite a few. Roughly this module requires to calculate, explicitly or implicitly, the expressions for $g^{\alpha}(q)$ and solve $g^{\alpha}(q) = 0$ (the signed distance $\overline{P'P}$), i.e., determinate the points that are going to touch, which are not necessarily the ones which are the closest at the instant of the computation, so several pairs of points have to be watched simultaneously. Even in very simple cases such as one degree-of-freedom systems, various numerical methods may be used to calculate the times t_k such that $g^{\alpha}(q(t_k)) = 0$, see Sect. 8.6.5. Their influence on the algorithm properties (consistency, order) may be significant.

The main problem is that an exact analytical description of the objects shapes, even when this is possible, is quite time consuming. Secondly one has to calculate with a suitable numerical routine the times t_k. In case of accumulation of impacts and for multiple contacts, the problem is harder because the influence of deciding the end of the series t_k, $k > 0$, according to the machine accuracy, is not easy to quantify. Micro-collisions phenomenon (Hurmuzlu, 1998) prove that it is possible in some cases that there is a large quantity of rebounds, but finite number of collisions, and an escape out of the constraint surface after a finite time. Things even complicate for multiple impacts. What is the influence on the long-run motion if one decides instead that one constraint becomes active?

Another issue is that it may not be possible to define all the constraints $g^{\alpha}(q) \geqslant 0$: in many practical situations there would be too many! Hence one usually employs procedures that eliminate useless constraints, i.e., those bodies which are too far to one another to be likely to collide in the next future steps of integration. Consequently one implements rough tests that select the bodies which may collide, and fine tests to compute the collision times (Eberhard, 1999). Rough tests usually consist of surrounding the bodies by simple volumes (spheres, boxes) and watching whether they overlap or not. Concerning the finest tests, the main approaches are (see Hubbard, 1996, for a review):

- Classification of typical contacts and geometries (Wang et al., 1997, 1999; Conti et al., 1992; Goyal et al., 1994; Han et al., 1993. In other words, process the real surface of the objects and the type of contact (circle–circle, circle–line, angle–line, etc.). These methods are essentially studied in the mechanical engineering literature. They are restricted to certain types of geometries contained in the available library developed for the software. If the body surfaces are simple enough to be described by analytical curves, one gets an explicit function $g^{\alpha}(q(t))$. The next step is to solve numerically $g^{\alpha}(q(t)) = 0$—which can be done with a Newton–Raphson method or a polynomial root-finding routine, since in case of several roots Newton–Raphson may compute the wrong zero and there is penetration before the algorithm decides that contact has occurred. Other authors (Wu et al. 1986; Wang et al., 1999) use a time step halving process until $g^{\alpha}(q(t-k)) = 0$ is satisfied within a specified tolerance. For instance, for two bodies with parametric surfaces one faces a nonlinear 5-dimensional root-finding problem (von Herzen et al., 1990). These methods are, however, less fast and more complex to implement than the 2-dimensional ones (Baraff, 1993).
- For 2-dimensional systems, one can approximate the bodies B_i, $i \in \{1,...,N\}$, by polygons made of edges and nodes N_i. Two main methods are used (Eberhard, 1999): the node-in-polygon test (NIPT) and the ray-crossing approach (RCA), see Fig. 3.3a and b, respectively. Let n_i be the number of nodes N_i in polygon i, and α_{ii} the angle (N_jN_i, N_jN_{i11}). Then if $\sum_{\text{nodes of} Bi} \alpha_{ii} = 0$, the node $N_j \notin B_i$; if $\sum_{\text{nodes of} B_i} \alpha_{ii} = 2\pi$, then $N_j \in B_i$: the bodies intersect. The RCA consists of looking at the number $\tilde{n}$ of intersections of a straight half-line (a ray) emanating from N_j, with the polygon ∂B_i. If $\tilde{n}$ is odd then $N_j \in B_i$, if $\tilde{n}$ is even then $N_j \notin B_i$. The RCA is more robust than the NIPT. Both methods are $O(n_in_j)$ for two bodies B_i and B_j. However, their generalization to 3-dimensional systems is not easy (Eberhard, 1999).
- Approximation of the object surfaces and of the impact times by bounding boxes methods (Hubbard, 1996) are more efficient for 3-dimensional systems. These methods are essentially studied in the computer science literature. If the bodies are convex and subject to gravity (or more generally to any vector field that is integrable), it is possible to approximate the distance $g^{\alpha}(q)$ and to calculate a lower bound on the impact time (Mirtich, 1997). The approximation can be refined as much as the constraints (desired accuracy, speed of computation) permit to do it. In von Herzen et al. (1990), it is pointed out that just watching positions to determine collision times cannot work since contact may occur between two sampling instants t_i and t_{i+1} while $g^{\alpha}(q(t_i)) > 0$ and $g^{\alpha}(q(t_{i+1})) > 0$. So including the velocity information in the algorithm is mandatory. Adaptive subdivision of the bodies into simple volumes (polygons or polyhedra; Eberhard, 1999), spheres (Mirtich, 1997; Hubbard, 1996), rectangular prisms (von Herzen et al., 1990), and incorporation of a Lipschtiz boundedness condition on $g^{\alpha}(\cdot)$ allows one to

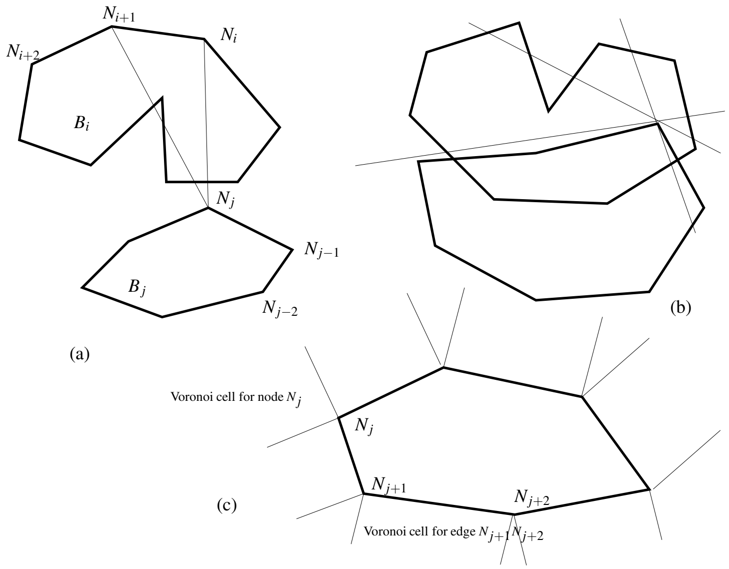


Fig. 3.3. Collision detection methods

approximate the collision times (von Herzen et al., 1990; Filip et al., 1986). This method is called bounding box schemes: each object is surrounded by bounding boxes. When these boxes overlap, the objects must be close to one another. Then a more accurate collision test is made once more. Bounding box schemes allow one to avoid testing all possible contacts ($= O(N^2)$ for N bodies), but to focus on objects in close proximity only. Roughly speaking, the Lipschitz bounds permit to approximate the next step motion of each simple volume (or surface) and to determine if a collision has occurred. A refinement of the mesh can be used to increase the accuracy of the collision time computation, in an adaptive way. These methods apply well to convex bodies. Nonconvex bodies can be decomposed into convex parts to be treated. Voronoi regions for polytopes (Ponamgi et al., 1995; Lin & Canny, 1991) are used to maintain a list of closest distances during the simulation.[5] The change in Voronoi cells from one step to the next one is usually small, facilitating the calculations. An implementation of the Lin–Canny algorithm with a running time linear in N can be found in Cohen et al. (1995). Baraff (1990) proposes a coherence-based bounding box test that is $O(N)$.

The interested reader may also have a look at the survey (Agarwal et al., 2002) and at Muth et al. (2007), where the ray-crossing and the fast multipole methods are compared. We shall not come back on these issues in this book.

[5] A Voronoi cell associated to an object consists of the set of points whose distance to this object is the smallest. The object can be a node, an edge, a face.

3.4 The Smooth Dynamics of Continuum Media

This section intends to briefly recall how one may derive the Lagrange equations for continuum media. It happens that once this step is achieved, the resulting dynamics is equivalent to the dynamics of any other mechanical system made of rigid bodies. Therefore the subsequent numerical analysis and simulation is the same as the ones for a rigid body. This is why there is no specific chapter on deformable bodies in this book.

3.4.1 The Smooth Equations of Motion

In this section, the smooth equations of motion of a collection of N continuum media are introduced in a quite usual setting. The continuum medium is identified at time $t \in [0,T]$ by its volume in $\mathrm{I\!R}^d$. The integer $d = 1,2,3$ denotes the space dimension, of interior

$$\Omega^{\alpha}(t) \subset \mathrm{I\!R}^d, i \in \{1,\ldots,N\}$$

and boundary $\partial\Omega^{\alpha}(t)$. If Ω^{α} is a deformable continuum medium, the equations of motion are introduced through the principle of virtual powers in a finite strain Lagrangian setting permitting a space-discretization based on a conventional finite element method. If Ω^{α} is assumed to be a rigid body, the equations of motion will be described by a finite set of coordinates. In both cases, possibly after a space-discretization, the equations of motion will be formulated and treated in a single finite-dimensional framework.

A material particle is described by its position X in a reference frame at $t = 0$ and by its current position $x = \varphi(X,t)$ at time t. For a Lagrangian description, we also assume we know at least formally the function $X = \psi(x,t)$. The displacement is defined by $u(x,t) = x - X = x - \psi(x,t)$ and the velocity and the acceleration are denoted by $\dot{u}$ and $\ddot{u}$. Most of the Lagrangian variables expressed in terms of X are denoted by capital letters, for instance, $U(X,t)$ for the displacement, and denoted by lower case for the associated Eulerian variables, in this case $u(x,t)$. This convention can be summarized by $u(x,t) = u(\varphi(X,t),t) = U(X,t)$.

Principle of Virtual Powers in Continuum Mechanics

Starting from the equation of motion in Eulerian coordinates,

$$\operatorname{div}\sigma(x,t) + \rho(x,t)b(x,t) = \rho(x,t)\ddot{u}(x,t), \;\; \forall x \in \Omega^{\alpha}(t)\,, \tag{3.87}$$

where $\sigma(x,t)$ is the Cauchy stress tensor and $b(x,t)$ is the density of body forces, the principle of virtual power states that

$$\int_{\Omega^{\alpha}(t)} (\ddot{u}(x,t) - b(x,t))\hat{v}(x,t)\,\mathrm{d}m(x,t) = \int_{\Omega^{\alpha}(t)} \operatorname{div}\sigma(x,t)\hat{v}(x,t)\,\mathrm{d}\omega(x,t) \tag{3.88}$$

for all virtual velocities denoted by $\hat{v}(x,t)$. The measure $\mathrm{d}\omega(x,t)$ denotes the Lebesgue measure in $\mathrm{I\!R}^d$ at x and the measure $\mathrm{d}m(x,t) = \rho(x,t)\,\mathrm{d}\omega(x,t)$ is the mass measure. With the help of the Green formulas, the principle of virtual power is usually reformulated as

$$\begin{aligned}\int_{\Omega^\alpha(t)} (\ddot{u}(x,t) - b(x,t))\hat{v}(x,t)\,\mathrm{d}m(x,t) = &-\int_{\Omega^\alpha(t)} \sigma(x,t):\nabla\hat{v}(x,t)\,\mathrm{d}\omega(x,t) \\ &+\int_{\partial\Omega^\alpha_F(t)} t(x,t)\hat{v}(x,t)\,\mathrm{d}s(x,t) + \int_{\Gamma^i_c(t)} r(x,t)\hat{v}(x,t)\,\mathrm{d}s(x,t)\end{aligned} \tag{3.89}$$

where $A:B = A_{ij}B^{ij}$ is the double contracted tensor product, and $t(x,t) = \sigma(x,t).n(x,t)$ is the applied forces on the boundary of outward normal n and $r(x,t)$ the reaction forces due to the unilateral contact and friction. The measure $\mathrm{d}s(x,t)$ is the Lebesgue measure at $x \in \partial\Omega^\alpha$. The symmetry of the Cauchy stress tensor in absence of density of momentum allows one to introduce the symmetric deformation rate tensor,

$$D(x,t) = \frac{1}{2}(\nabla^{\mathrm{T}}v(x,t) + \nabla v(x,t)) \tag{3.90}$$

leading to the standard expression of the virtual power of the internal forces of cohesion

$$\hat{\mathscr{P}}_{\mathrm{int}} = -\int_{\Omega^\alpha(t)} \sigma(x,t):\nabla\hat{v}(x,t)\,\mathrm{d}\omega(x,t) = -\int_{\Omega^\alpha(t)} \sigma(x,t):\hat{D}(x,t)\,\mathrm{d}\omega(x,t)\,. \tag{3.91}$$

In order to formulate the principle of virtual power in a total Lagrangian framework, the second Piola–Kirchhoff tensor,

$$S(X,t) = F^{-1}\det(F)\sigma^{\mathrm{T}}F^{-\mathrm{T}}$$

is introduced, where $F = \dfrac{\partial x}{\partial X} = \dfrac{\partial\varphi(X,t)}{\partial X}$ is the deformation gradient. The virtual power of the internal forces is then rewritten as

$$\hat{\mathscr{P}}_{\mathrm{int}} = -\int_{\Omega^\alpha(t)} \sigma(x,t):\nabla\hat{v}(x,t)\,\mathrm{d}\omega(x,t) = -\int_{\Omega^\alpha(0)} S(X,t):\hat{L}(X,t)\,\mathrm{d}\Omega(X,0) \tag{3.92}$$

where $L = \dot{F}$, and the notation $\hat{\cdot}$ denotes the virtual quantities. Finally, the principle of virtual power in a total finite strain Lagrangian framework in terms of the convected Lagrangian variable X, is

$$\begin{aligned}\int_{\Omega^\alpha(0)} (\ddot{U}(X,t) - B(X,t))\hat{V}(X,t)\,\mathrm{d}M(X,0) = &-\int_{\Omega^\alpha(0)} S(X,t):\hat{L}(X,t)\,\mathrm{d}\Omega(X,0) \\ &+\int_{\partial\Omega^\alpha(0)} T(X,t)\hat{V}(X,t)\,\mathrm{d}S(X,t) + \int_{\Gamma^i_c(0)} R(X,t)\hat{V}(X,t)\,\mathrm{d}S(X,t)\end{aligned} \tag{3.93}$$

where the applied forces laws on the boundary satisfy

$$T(X,t) = S(X,t)F^{\mathrm{T}}(X,t)N(X,t)$$

and

$$R(X,t) = S(X,t)F^{\mathrm{T}}(X,t)N(X,t).$$

It is noteworthy that the interactions between bodies that are taken into account in this model are only given by the forces through the interface with unilateral contact and friction.

For the constitutive material laws, a large panel of models can be taken into account in this framework and in the numerical applications. The formulation of the constitutive laws is based on the standard thermodynamics of irreversible processes (Germain et al., 1983) or based on a variational formulation of incremental stress–strain relation deriving from a pseudo-elastic potential (Ortiz & Stainier, 1999). If the bulk response of the material is supposed to be linear elastic, that is,

$$S(X,t) = K(X,T) : E(X,t) \; , \tag{3.94}$$

where E is the Green–Lagrange strain tensor,

$$E(X,t) = \frac{1}{2}(F^{\mathrm{T}}(X,t)F(X,t) - I(X,t)) \; , \tag{3.95}$$

where the tensor I is the identity tensor and $K(X,T)$ is the fourth-order tensor of elastic properties.

The finite element discretization is conventional and is based on this principle of virtual power in this total Lagrangian framework. Choosing some isoparametric element leads to the following approximation

$$\begin{aligned} &U(X,t) = \textstyle\sum_h N^h(X,t)U_h(t), \quad \dot{u}(X,t) = \textstyle\sum_h N^h(X,t)\dot{U}_h(t) \\ &\ddot{U}(X,t) = \textstyle\sum_h N^h(X,t)\ddot{U}_h(t) \; , \end{aligned} \tag{3.96}$$

where N^h are the shape functions and U_h the finite set of displacement at nodes. Substituting this approximation into the principle of virtual power and simplifying with respect to the virtual field yields a space-discretized equation of motion of the form

$$M(U_h)\ddot{U}_h + F(t, U_h, \dot{U}_h) = R \; , \tag{3.97}$$

where $M(U_h)$ is the consistent or lumped mass matrix, the vector $F(t, U_h, \dot{U}_h)$ collects the internal and external discretized forces, and R are the discretized forces, due to the contact model.

Principle of Virtual Powers in Rigid Body Mechanics

In rigid body mechanics, it is assumed that the power of the cohesion internal forces vanishes for a rigid motion given by the following set of virtual velocity field,

$$\mathscr{V} = \{\hat{v}(x,t) = \hat{v}_O(t) + \hat{\omega}(t) \times (x - x_O), \;\; \forall x \in \Omega^{\alpha}(t)\} \; , \tag{3.98}$$

where O is a geometrical point fixed with respect to the body, x_O is the position of this point $v_O(t)$ is its velocity, and $\omega(t)$ the angular velocity of the body at O. This assumption yields

$$\int_{\Omega^{\alpha}(t)} (\ddot{u}(x,t) - b(x,t))\hat{v}(x,t)\,\mathrm{d}m(x,t) = \int_{\partial\Omega^{\alpha}(t)} t(x,t)\hat{v}(x,t)\,\mathrm{d}s(x,t) \tag{3.99}$$

for all $\hat{v}(x,t) \in \mathscr{V}$. The equation of motion can be derived choosing a particular virtual velocity as:

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}\displaystyle\int_{\Omega^{\alpha}(t)} \dot{u}(x,t)\,\mathrm{d}m(x,t) = \int_{\Omega^{\alpha}(t)} b(x,t)\,\mathrm{d}m(x,t) + \int_{\partial\Omega^{\alpha}(t)} t(x,t)\,\mathrm{d}s(x,t) \\ \dfrac{\mathrm{d}}{\mathrm{d}t}\displaystyle\int_{\Omega^{\alpha}(t)} (x - x_O) \times \dot{u}(x,t)\,\mathrm{d}m(x,t) = \int_{\Omega^{\alpha}(t)} (x - x_O) \times g(x,t)\,\mathrm{d}m(x,t) \\ \qquad\qquad + \displaystyle\int_{\partial\Omega^{\alpha}(t)} (x - x_O) \times t(x,t)\,\mathrm{d}s(x,t)\ . \end{cases} \tag{3.100}$$

Various descriptions of the equations of motion of a rigid body can be deduced from the principle of virtual power choosing particular kinematics. Without going into further details, the Newton–Euler formulation can be chosen to write the kinematics with respect to the center of mass G_i of the body Ω^{α} in Eulerian coordinates:

$$\begin{cases} \dot{u}(x,t) = v_{G_i}(t) + \omega_i(t) \times (x - x_{G_i}) \\ \ddot{u}(x,t) = \dot{v}_{G_i}(t) + \dot{\omega}_i(t) \times (x - x_{G_i}) + \omega_i(t) \times (\omega_i(t) \times (x - x_{G_i}))\ . \end{cases} \tag{3.101}$$

Substituting (3.101) into the equations of motion (3.100) yields the Newton–Euler equations,

$$\begin{bmatrix} M & 0 \\ 0 & I \end{bmatrix} \frac{\mathrm{d}}{\mathrm{d}t} \begin{bmatrix} v_{G_i}(t) \\ \omega_i(t) \end{bmatrix} + \begin{bmatrix} 0 \\ \omega_i(t) \times I\omega_i(t) \end{bmatrix} = \begin{bmatrix} f_{\mathrm{ext}}(t) \\ m_{\mathrm{ext}}(t) \end{bmatrix}\ , \tag{3.102}$$

where

$$\begin{aligned} M &= \int_{\Omega^{\alpha}(t)} \mathrm{d}m(x,t) = \int_{\Omega^{\alpha}(0)} \mathrm{d}M(X,0) \\ I &= \int_{\Omega^{\alpha}(t)} (x - x_{G_i})^{\mathrm{T}}(x - x_{G_i})\,\mathrm{d}m(x,t) \\ &= \int_{\Omega^{\alpha}(0)} (X - X_{G_i})^{\mathrm{T}}(X - X_{G_i})\,\mathrm{d}M(X,0) \\ f_{\mathrm{ext}}(t) &= \int_{\Omega^{\alpha}(t)} b(x,t)\,\mathrm{d}m(x,t) + \int_{\partial\Omega^{\alpha}(t)} t(x,t)\,\mathrm{d}s(x,t) \\ m_{\mathrm{ext}}(t) &= \int_{\Omega^{\alpha}(t)} (x - x_{G_i}) \times b(x,t)\,\mathrm{d}m(x,t) + \int_{\partial\Omega^{\alpha}(t)} (x - x_{G_i}) \times t(x,t)\,\mathrm{d}s(x,t). \end{aligned} \tag{3.103}$$

Usually, a second-order form of the dynamics is obtained with the help of the following parameterization of the vector ω_i:

$$\omega_i(t) = D_i(\Psi,t)\dot{\Psi}_i(t)\ , \tag{3.104}$$

where $D(\Psi,t)$ is supposed to be a diffeomorphism. For a collection of N rigid bodies, a usual way is to introduce a set a generalized coordinates z such that

$$z = \left[[x_{G_i}, \Psi_i]_{i\in\{1\ldots N\}} \right]^{\mathrm{T}} \tag{3.105}$$

assuming that the positions and the orientations of the bodies are uniquely determined by z. With this variable, after some algebraic manipulations, the equations of motion can be written as:

$$M(z(t))\ddot{z}(t) + F(t, z(t), \dot{z}(t)) = r(t)\ . \tag{3.106}$$

It is noteworthy that this formulation allows us to add some internal forces between bodies expressed in terms of the generalized coordinates $z(\cdot)$.

3.4.2 Summary of the Equations of Motion

In the sequel, the equations of motion of space-discretized continuum media and rigid bodies will be treated in the same setting. In order to summarize the equations (3.97), (3.106), and (3.4), we introduce the finite vector of variables q which can represent the discretized displacement U_h or any generalized coordinates of the rigid motion z. Hence, the equations of motion will be written as

$$M(q(t))\ddot{q}(t) + F(t, q(t), \dot{q}(t)) = r(t)\ , \tag{3.107}$$

where q collects the variables U_h and z.

3.5 Nonsmooth Dynamics and Schatzman's Formulation

In Schatzman (1978), a mathematical formulation of the nonsmooth dynamics in the scalar case $q(t) \in \mathbb{R}$ is proposed. Let us consider a nonempty closed convex set K, not reduced to a singleton, i.e., $K = [a,b]$ for some reals a and b, possibly infinite, $a < b$. The second-order nonsmooth dynamics is written as follows for a continuous function $q(\cdot)$, from $[0,T]$ to $\mathbb{R}$, which takes its value in K, i.e., $q \in \mathscr{C}^0([0,T],K)$, and

$$\begin{cases} \ddot{q}(t) = f(\cdot, q(t), \dot{q}(t)) + \mu, \\ \\ \langle \mu, v - q(t) \rangle \leqslant 0,\ \forall v \in \mathscr{C}^0([0,T],K)\ , \end{cases} \tag{3.108}$$

where

- The function $\dot{q}(\cdot)$ is chosen to be a BV function, which must have discontinuities at the boundary ∂K of K.
- The first equation has to be understood in the sense of distributions. The term μ is a real measure on $[0,T]$.

- The function $f\colon [0,T]\times\mathrm{I\!R}\times\mathrm{I\!R}\to\mathrm{I\!R}$ is assumed to be a Lipschitz-continuous function with respect to its last two arguments.

Let us give some insights on this formulation, that is a variational inequality formalism. Indeed the second line of (3.108) may be written as $\langle\ddot{q}(t)-f(t,q(t),\dot{q}(t)), v-q(t)\rangle\leqslant 0$, for all $v\in\mathscr{C}^0([0,T],K)$. If $q(t_0)\in\min K$ with a strictly negative left velocity $\dot{q}^-(t_0)<0$, the velocity after a jump must be nonnegative, i.e., $\dot{q}^+(t_0)\geqslant 0$ and the second-order derivative must have a Dirac mass at t_0. The second condition in (3.108) is equivalent to

$$\begin{aligned}&\mathrm{supp}(\mu)\in\{\mathrm{t}\mid \mathrm{q(t)}\in\partial\mathrm{K}\},\quad \mu\geqslant 0 \text{ on } \{\mathrm{t}\mid\mathrm{q(t)}=\min\mathrm{K}\}\\ &\mu\leqslant 0 \text{ on } \{t\mid q(t)=\max K\}\,.\end{aligned}\tag{3.109}$$

In the language of convex analysis and differential inclusions, the system (3.108) is equivalent to

$$\ddot{q}(t)+\partial\psi_K(q(t))\in f(t,q(t),\dot{q}(t))\,,\tag{3.110}$$

where $\psi_K(\cdot)$ is the indicator function of K. For an initial value problem, Schatzman (1978) also defines consistent initial condition for the Cauchy problem. Let us recall what is the tangent cone to K in this simple case

$$T_K(x)=\begin{cases}\mathrm{I\!R} & \text{if } x\in\mathrm{Int}(K)\\ \mathrm{I\!R}^- & \text{if } x=\max K\\ \mathrm{I\!R}^+ & \text{if } x=\min K\,.\end{cases}\tag{3.111}$$

The set of consistent Cauchy data is defined by $\{(q,\dot{q})\in\mathrm{I\!R}^2\mid q\in K,\dot{q}\in T_K(q)\}$.

To complete the model, a constitutive law has to be given for the impact rule. In Schatzman (1978), a purely elastic impact law is chosen, i.e., $\dot{q}^+(t)=-\dot{q}^-(t)$. With this model, a nonuniqueness example is given even with $\mathscr{C}^\infty$ data. In Paoli & Schatzman (1993), a Newton impact law is chosen

$$\dot{q}^+(t)=-e\dot{q}^-(t),\quad e\in[0,1]\,.\tag{3.112}$$

Finally, in Paoli & Schatzman (1999), the finite-freedom dynamics is introduced in the form

$$M(q(t))\ddot{q}(t)+\mu=f(\cdot,q(t),p(t))\,,\tag{3.113}$$

where $p=M(q)\dot{q}$ is the generalized momentum of Hamiltonian mechanics. The major interest of the formulation is that it is proved to be equal to the limit of a smooth penalized model in the rigid limit. This formulation gave rise to a time-stepping scheme presented in Chap. 10.

Remark 3.10. This Schatzman–Paoli formulation is very similar to the following Moreau's sweeping process. The major difference is that the Moreau's sweeping process proposes a unified framework for the dynamics in terms of velocity and direct inclusions of measures into a cone which depends on the velocity. This has important consequences on the numerical implementation, because the cones which appear in Moreau's formulation (the second-order sweeping process) are polyhedral cones.

3.6 Nonsmooth Dynamics and Moreau's Sweeping Process

3.6.1 Measure Differential Inclusions

With the presence of the unilateral constraints, the evolution of the systems is usually no longer smooth. Especially, the velocity $v(\cdot) = \dot{q}(\cdot)$ may encounter jumps and must be considered as a function of bounded variations (BV) in time. With this assumption, the equation of motion is rewritten in terms of right-continuous BV (RCBV) function, denoted as $v^+(\cdot) = \dot{q}^+(\cdot)$.[6]

The generalized coordinates, assumed to be absolutely continuous, are deduced from the velocity by the standard integration of a function of bounded variations:

$$q(t) = q(t_0) + \int_{t_0}^{t} v^+(t)\,\mathrm{d}t\ , \tag{3.114}$$

where dt is the Lebesgue measure.

If the velocity is a BV function, the acceleration is no longer defined everywhere as the derivative in the classical sense of the velocity. The notion of differential measure, or a special Stieltjes measure provides the right substitute to this notion as a derivative of the velocity in the sense of the distributions. In the same way, the generalized force r is to be considered as a real measure, denoted dr.

The equation of motion (3.17) is formulated in terms of a measure differential equation:

$$\begin{cases} M(q(t))\mathrm{d}v + N(q(t), v^+(t))\mathrm{d}t + F_{\mathrm{int}}(t, q(t), v^+(t))\mathrm{d}t = F_{\mathrm{ext}}(t)\mathrm{d}t + \mathrm{d}r \\ \\ v^+(t) = \dot{q}^+(t) \end{cases} \tag{3.115}$$

on $[0,T]$, and with admissible initial data.

Remark 3.11. Notice that the dynamics is written in terms of the RCBV function $v^+(\cdot)$. It may also be possible to write the dynamics in terms of left-continuous BV function $v^-(\cdot)$, as

$$\begin{cases} M(q(t))\mathrm{d}v + N(q(t), v^-(t))\mathrm{d}t + F_{\mathrm{int}}(t, q(t), v^-(t))\mathrm{d}t = F_{\mathrm{ext}}(t)\mathrm{d}t + \mathrm{d}r \\ \\ v^-(t) = \dot{q}^-(t) \end{cases} \tag{3.116}$$

on $[0,T]$. Since we are interested only in forward integration of the dynamics, we keep only the form (3.115).

3.6.2 Decomposition of the Nonsmooth Dynamics

Thanks to the Lebesgue decomposition theorem and its variants, the differential measure dv is decomposed as

[6] Functions of bounded variations always possess right and left limits.

$$\mathrm{d}v = \gamma\,\mathrm{d}t + (v^+ - v^-)\mathrm{d}\nu + \mathrm{d}v_s \tag{3.117}$$

where

- $\gamma(\cdot) = \ddot{q}(\cdot)$ is the acceleration defined in the usual sense.
- $v^+ - v^-$ is the difference between the right-continuous and the left-continuous functions associated with the BV function $v(\cdot) = \dot{q}(\cdot)$, and $\mathrm{d}\nu$ is a purely atomic measure with atoms at the time t_i of discontinuities of $v(\cdot)$, i.e.,

$$\mathrm{d}\nu = \sum_i \delta_{t_i}\,. \tag{3.118}$$

- dv_s is a singular measure with respect to $\mathrm{d}t + \mathrm{d}\nu$ which we will neglect for practical reasons.

In the same way, the measure $\mathrm{d}r$ can be decomposed as follows:

$$\mathrm{d}r = f\ \ \mathrm{d}t + p\ \ \mathrm{d}\nu + \mathrm{d}r_s\,, \tag{3.119}$$

where:

- $f(\cdot)$ is the Lebesgue measurable force.
- p is the purely atomic impact impulsion such that

$$p\,\mathrm{d}\nu = \sum_i p_i \delta_{t_i}\,. \tag{3.120}$$

- $\mathrm{d}r_s$ is a singular force measure with respect to $\mathrm{d}t + \mathrm{d}\nu$ which we will also neglect.

3.6.3 The Impact Equations and the Smooth Dynamics

Inserting (3.117) and (3.119) in (3.115), the dynamics is written as an equality of measures

$$\begin{aligned} &M(q(t))\gamma(t)\mathrm{d}t + M(q(t))(v^+(t) - v^-(t))\mathrm{d}\nu + N(q(t), v^+(t))\mathrm{d}t + \\ &\qquad + F_{\mathrm{int}}(t, q(t), v(t))\mathrm{d}t = F_{\mathrm{ext}}(t)\mathrm{d}t + f(t)\mathrm{d}t + p\,\mathrm{d}\nu \end{aligned} \tag{3.121}$$

and can be split into the atomic part and the Lebesgue part in terms of $v^+(\cdot)$:

$$\begin{cases} M(q(t))(v^+(t) - v^-(t))\mathrm{d}\nu = p\,\mathrm{d}\nu \\ M(q(t))\gamma(t)\mathrm{d}t + N(q(t), v^+(t))\mathrm{d}t + F_{\mathrm{int}}(t, q(t), v(t))\mathrm{d}t = F_{\mathrm{ext}}(t)\mathrm{d}t + f(t)\mathrm{d}t\,. \end{cases} \tag{3.122}$$

It is supposed that the unilateral constraints are $g(q) \geqslant 0$, see (3.14) and (3.15). Due to the definition (3.118) of the measure $\mathrm{d}\nu$, the impact equations can be written at the time t_i of discontinuities:

$$M(q(t_i))(v^+(t_i) - v^-(t_i)) = p_i\,. \tag{3.123}$$

This is an algebraic equation. The smooth dynamics which is valid almost everywhere for the Lebesgue measure dt (dt a.e.) is governed by the following equation:

$$M(q(t))\gamma^{+}(t)+N(q(t),v^{+}(t))+F_{\text{int}}(t,q(t),v^{+}(t))=F_{\text{ext}}(t)+f^{+}(t) \qquad (3.124)$$

$dt-$a.e., where we assume that $f^{+}(\cdot)=f^{-}(\cdot)=f(\cdot)\,(dt-a.e.)$. Obviously the same type of separation between smooth and nonsmooth motions can be performed with the Newton–Euler's equations. The impact dynamics then links the jump in the center of mass velocity and the impulsive contact force, and the instantaneous angular velocity jump with the impulsive contact reaction moment.

3.6.4 Moreau's Sweeping Process

Moreau's sweeping process is a mathematical setting which combines a dynamics described in terms of measure as in (3.115) together with a description of the unilateral constraint including an impact law. We already described quickly the sweeping process in Sects. 1.4 and 2.7. A key stone of this formulation is the inclusion in terms of velocity. Indeed, the inclusion (3.24) is "replaced" by[7]

$$-\mathrm{d}r \in N_{T_{\mathscr{C}}(q(t))}(v^{+}(t)) \ , \qquad (3.125)$$

where $\mathscr{C}$ is the admissible domain of the configuration space. We do not make any assumption on $\mathscr{C}$ here, but one should keep in mind that the right-hand side of (3.125) may be meaningless for some too general sets $\mathscr{C}$. In most of the cases with practical interest, $\mathscr{C}$ is finitely represented, i.e., it is represented as in (3.16). In such a case one just has to take care that $\mathrm{Int}(T_{\mathscr{C}}(q)) \neq \emptyset$, which is equivalent to the existence of a hyperplane in $\mathbb{R}^{n}$, not containing the origin, which intersects all the half-lines generated by the gradients $\nabla g^{\alpha}(q)$ of the active constraints (Moreau, 1985b). This inclusion will be called the inclusion in terms of velocity. Two features of (3.125) have to be mentioned:

- *The inclusion concerns measures.* Therefore, it is necessary to define what is the inclusion of a measure into a cone.
- *The inclusion is written in terms of velocity* $v^{+}(\cdot)$ rather than of the coordinates $q(\cdot)$.

As we can define an inequality constraint on a measure, it is possible to define a relevant meaning for the inclusion (3.125). Roughly speaking, when the measure possesses a density with respect to the Lebesgue measure,

$$\mathrm{d}r = r'\mathrm{d}t = f(t)\mathrm{d}t \ . \qquad (3.126)$$

Then the inclusion is equivalent to the inclusion of $f(\cdot)$ which is a real function of time, into the cone at time t. When the measure possesses an atom

$$\mathrm{d}r = p\delta \ , \qquad (3.127)$$

[7] Actually it is proved by J.J. Moreau, from convex analysis, that the inclusion (3.125) is satisfied.

where δ is the Dirac measure and p the amplitude of the atom usually called the percussion, the inclusion is equivalent to say that p is included into the cone. Naturally, the same illustration can be made for inequality constraints on measures. For more details, we refer to Monteiro Marques (1993), Kunze & Monteiro Marqués (2000), Stewart (2001), and Acary et al. (in press).

A viability lemma due to Moreau (1999) ensures that the inclusion in terms of velocity (3.125) together with admissible initial conditions on the position implies that the constraints on the coordinates are always satisfied. In fact, we always have (see, e.g., Brogliato, 2004 for a proof)

$$N_{T_C(q)}(v^+) \subset N_{\mathscr{C}}(q).$$

The reverse is not true. A key assumption has to be added which is related to the notion of impact laws. Indeed, if the constraint is active, i.e., d$r > 0$, then the post-impact velocity $v^+(\cdot)$ is equal to zero. For instance if an impact occurs, the post-impact velocity vanishes. The model is an inelastic (plastic) impact rule.

As done in Moreau (1988b) and Mabrouk (1998), the impact rule can be enhanced with normal and tangential coefficients as follows

$$-\mathrm{d}r \in N_{T_{\mathscr{C}}(q(t))}\left(\frac{v^+(t)+ev^-(t)}{1+e}\right) . \tag{3.128}$$

Inserting (3.128) in (3.123) and using (A.8) one obtains

$$v^+(t) = -ev^-(t) + (1+e)\mathrm{prox}_{M(q(t))}[T_{\mathscr{C}}(q(t));v^-(t)] \tag{3.129}$$

with $\mathrm{prox}_{M(q(t))}[T_{\mathscr{C}}(q(t));v^-(t)] = \mathrm{argmin}_{z\in T_{\mathscr{C}}(q(t))}\,\frac{1}{2}(z-v^-(t))^{\mathrm{T}}M(q(t))(z-v^-(t))$ that is numerically tractable since $T_{\mathscr{C}}(q)$ is a polyhedral set. Let $\nu = 1$, i.e., there is only one constraint. This may also be written after some calculations as (q stands for $q(t)$)

$$v^+(t) = v^-(t) - (1+e)M^{-1}(q)\nabla g(q)[\nabla g^{\mathrm{T}}(q)M^{-1}(q)\nabla g(q)]^{-1}\nabla g^{\mathrm{T}}(q)v^-(t)\,, \tag{3.130}$$

where the multiplier is given by

$$\lambda = -(1+e)[\nabla g^{\mathrm{T}}(q)M^{-1}(q)\nabla g(q)]^{-1}\nabla g^{\mathrm{T}}(q)v^-(t)$$

and $p_i = \nabla g(q)\lambda$ in (3.123). One may also obtain (3.130) directly from (3.129) from the expression of the projection on the tangent cone. If the local relative velocity satisfies $U_{\mathrm{N}}^+(t) = -eU_{\mathrm{N}}^-(t)$ and $U_{\mathrm{T}}^+(t) = U_{\mathrm{T}}^-(t)$ (the case of a frictionless surface), and if $U_{\mathrm{N}}(\cdot) = \dot{g}(\cdot) = \nabla g^{\mathrm{T}}(q)v(\cdot)$, then (3.130) is a consequence of the impact dynamics. Moreau's rule is equivalent to Newton's impact rule, however, it is formulated in generalized coordinates and supplies the whole velocity in one shot. When $\nu \geqslant 2$, multiple impacts may occur when the trajectory hits several constraint boundaries at the same time. Moreau's rule also provides a result for the post-impact velocity in this case (notice that (3.129) is written without assuming that $\nu = 1$). Whether or not the obtained solution is physically sound is another problem. The modeling of multiple impacts is a topic still under investigation at the time of writing of this book. We just mention the fact that Moreau's sweeping process furnishes a geometrical framework that may be used for further research in the field of multiple impacts, and refer to Glocker (2004) and Acary & Brogliato (2003) for more information. Moreau's rule in (3.129) generalizes Newton's law. In Pfeiffer & Glocker (1996) it is proposed to extend Poisson's model, sometimes called the kinetic model. This is done by solving two LCPs, one corresponding to the compression phase, the other one to the expansion phase (despite in rigid body theory there are no deformations, so this is to be understood as some kind of approximation of the compliant case).

3.6.5 Finitely Represented $\mathscr{C}$ and the Complementarity Formulation

Let $\mathscr{C}$ be finitely represented, i.e.,

$$\mathscr{C} = \{q \in \mathscr{M}(t) \mid g^\alpha(q) \geqslant 0, \alpha \in \{1 \dots \nu\}\} . \tag{3.131}$$

In this case the tangent cone is a convex polyhedral set defined by Moreau as

$$T_{\mathscr{C}}(q) = \{z \in \mathrm{I\!R}^n \mid z^{\mathrm{T}} \nabla g^\alpha(q) \geqslant 0, \text{ for all } \alpha \in I(q)\} , \tag{3.132}$$

where $I(q)$ is the set of indices of the active constraints, i.e., $I(q) = \{\alpha \in \{1,..,\nu\} \mid g^\alpha(q) \leqslant 0\}$.[8] Then we can decompose the measure dr and the velocity $V^+(\cdot) = \nabla g^{\mathrm{T}}(q) v^+(\cdot)$ as follows:

$$\mathrm{d}r = \sum_\alpha \nabla g^\alpha(q)\, \mathrm{d}\lambda_\alpha \tag{3.133}$$

$$U^+ = \left[U^{\alpha,+} = \nabla g^{\alpha,\mathrm{T}}(q)\, v^+, \alpha \in \{1 \dots \nu\}\right] . \tag{3.134}$$

If some constraints qualification condition holds, then the inclusion (3.125) can be written equivalently as

$$-\mathrm{d}\lambda_\alpha \in N_{T_{\mathrm{I\!R}_+}(g^\alpha(q))}(U^{\alpha,+}) , \tag{3.135}$$

or

$$\begin{cases} \text{If } g^\alpha(q) \leqslant 0, \text{ then } 0 \leqslant U^{\alpha,+} \perp \mathrm{d}\lambda^\alpha \geqslant 0 \\ \text{If } g^\alpha(q) > 0, \text{ then } \mathrm{d}\lambda^\alpha = 0 . \end{cases} \tag{3.136}$$

This corresponds to a plastic impact ($e = 0$). Replacing V_α^+ by $V_\alpha^+ + eV_\alpha^-$ in (3.135) and (3.136) allows one to take into account other restitutions with $e \in [0,1]$. From Claim 6.1 in Brogliato (1999) this is equivalent to formulate the impact at the generalized velocity level as in (3.130).

Remark 3.12. We have not written U^α for the velocity because the term $\nabla g^{\alpha,\mathrm{T}}(q)\, v^+$ does not necessarily represent the local kinematics variable as in Sect. 3.3. It does for a particular choice of the functions $g^\alpha(\cdot)$ as the gap functions.

[8] This definition permits to compute the tangent cone even when the constraints are violated, which is needed numerically.

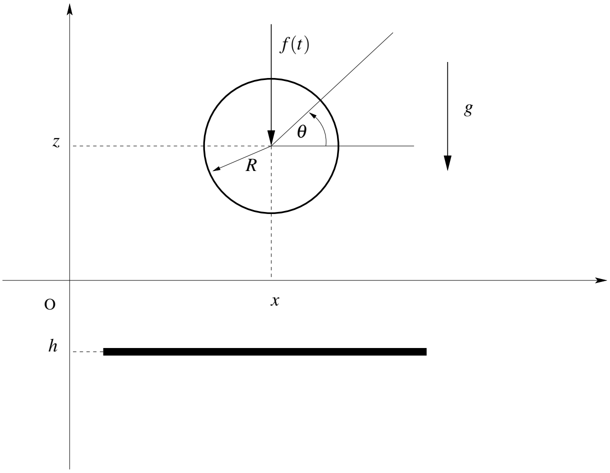


Fig. 3.4. Two-dimensional bouncing ball on a rigid plane

Example 3.13 (The set $\mathscr{C}$ equal to $\mathbb{R}^+$). To illustrate the last point on a very simple example, let us take for example an admissible set $\mathscr{C}$ equal to $\mathbb{R}^+$. The complementarity relation

$$-\mathrm{d}r \in N_{\mathscr{C}}(q) \Leftrightarrow 0 \leqslant q \perp \mathrm{d}r \geqslant 0 \tag{3.137}$$

is replaced by

$$-\mathrm{d}r \in N_{T_{\mathscr{C}}(q)}(v^+) \Leftrightarrow \begin{cases} \text{if } q \leqslant 0, & \text{then } 0 \leqslant v^+ \perp \mathrm{d}r \geqslant 0 \\ \\ \text{if } q > 0, & \text{then } \mathrm{d}r = 0\,. \end{cases} \tag{3.138}$$

Example 3.14 (The example of the bouncing ball). Let us consider a ball of mass m and radius R, described by three generalized coordinates $q = [z,x,\theta]^{\mathrm{T}}$. The ball is subjected to the gravity g and a vertical external force $f(t)$. The system is also constituted by a rigid plane, defined by its position h with respect to the axis Oz. We assume that the position of the plane is fixed. The physical problem is depicted in Fig. 3.4. The ball bounces on the rigid plane, introducing a constraint on its position. We consider also that the behavior of the system at impacts is governed by a Newton impact law with a coefficient of restitution $e \in [0,1]$.

Lagrangian Dynamics: We construct all the terms which define a Lagrangian NSDS as in (3.115). In our special case, the model is completely linear:

$$q = \begin{bmatrix} z \\ x \\ \theta \end{bmatrix}, \; M(q) = \begin{bmatrix} m & 0 & 0 \\ 0 & m & 0 \\ 0 & 0 & I \end{bmatrix} \text{ where } I = \frac{3}{5}mR^2$$

$$N(q,\dot{q}) = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \; F_{\text{int}}(q,\dot{q},t) = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \; F_{\text{ext}}(t) = \begin{bmatrix} -mg \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} f(t) \\ 0 \\ 0 \end{bmatrix}. \tag{3.139}$$

Kinematics Relations: The unilateral constraint requires that

$$\mathscr{C} = \{q \mid g^1(q) = z - R - h \geqslant 0\}, \tag{3.140}$$

so we identify the terms of the equation (3.133)

$$-\mathrm{d}r = [1,0,0]^{\mathrm{T}}\mathrm{d}\lambda_1 \tag{3.141}$$

$$\tag{3.142}$$

$$U_1^+ = [1,0,0] \begin{bmatrix} \dot{z} \\ \dot{x} \\ \dot{\theta} \end{bmatrix} = \dot{z}. \tag{3.143}$$

Nonsmooth Laws: The following contact laws can be written:

$$\begin{cases} \text{if } g^1(q) \leqslant 0, & \text{then } 0 \leqslant U_1^+ + eU_1^- \perp \mathrm{d}\lambda_1 \geqslant 0 \\ \text{if } g^1(q) \geqslant 0, & \text{then } \mathrm{d}\lambda_1 = 0\,. \end{cases} \tag{3.144}$$

3.7 Well-Posedness Results

There have been numerous studies concerning the existence, uniqueness, and continuous dependence of solutions. For this last point see Sect. 6.1. We may cite Dzonou & Monteiro Marques (2007) for the sweeping process, and Ballard (2000). Under some mild assumptions like piecewise analycity of the data and functional independence of the constraint functions $g^\alpha(\cdot)$, then the position $q(\cdot)$ is absolutely continuous, and the velocity $v(\cdot)$ is RCLBV. The interest of many well-posedness proofs (like the one in Dzonou & Monteiro Marques, 2007) is that they are led with time-discretizations of the dynamics (the catching-up algorithm) whose convergence properties are studied, therefore proving the consistency of the algorithm.

3.8 Lagrangian Systems with Perfect Unilateral Constraints: Summary

In the Lagrangian setting, the equation of motion with the perfect unilateral constraints are given by

$$\begin{cases} M(q(t))\mathrm{d}v + N(q(t), v^+(t))\mathrm{d}t + F_{\mathrm{int}}(t, q(t), v^+(t))\mathrm{d}t = F_{\mathrm{ext}}(t)\mathrm{d}t + \mathrm{d}r \\[2ex] q(t) = q(t_0) + \displaystyle\int_{t_0}^{t} v^+(t)\,\mathrm{d}t \\[2ex] -\mathrm{d}r \in N_{T_{\mathscr{C}}(q(t))}(v^+(t))\,. \end{cases} \tag{3.145}$$

If $\mathscr{C} = \{q \in \mathscr{M}(t) \mid g^{\alpha}(q) \geqslant 0, \alpha \in \{1 \dots \nu\}\}$ is finitely represented, we have

$$\begin{cases} M(q(t))\mathrm{d}v + N(q(t), v^+(t))\mathrm{d}t + F_{\mathrm{int}}(t, q(t), v^+(t))\mathrm{d}t = F_{\mathrm{ext}}(t)\mathrm{d}t + \mathrm{d}r \\[2ex] q(t) = q(t_0) + \displaystyle\int_{t_0}^{t} v^+(t)\,\mathrm{d}t \\[2ex] \mathrm{d}r = \displaystyle\sum_{\alpha} \nabla g^{\alpha}(q)\,\mathrm{d}\lambda_{\alpha} \\[2ex] U^{\alpha,+} = \left[U^{\alpha,+}, \alpha \in \{1 \dots \nu\}\right] \text{ with } U^{\alpha,+} = \nabla g^{\alpha,\mathrm{T}}(q)\, v^+ \\[2ex] -\mathrm{d}\lambda^{\alpha} \in N_{T_{\mathbb{R}_+}(g_{\alpha})}(U^{\alpha,+}), \text{ or } 0 \leqslant U^{\alpha,+} \perp \mathrm{d}\lambda^{\alpha} \geqslant 0 \text{ if } g^{\alpha}(q) \leqslant 0\,. \end{cases} \tag{3.146}$$

These systems can be enhanced with friction and impacts rules, see Sect. 3.9. Recall that if $g^{\alpha}(q)$ represents the signed distance between two bodies of a system, then $V_{\alpha}^{+}(\cdot)$ is the normal relative velocity $U_{\mathrm{N}}^{\alpha}(\cdot)$ between the two bodies (called the contactors).

3.9 Contact Models

There are two basic, fundamental contact models: Newton's law for frictionless impacts that states that $U_{\mathrm{N}}^{+}(t) = -eU_{\mathrm{N}}^{-}(t)$ and $U_{\mathrm{T}}^{+}(t) = U_{\mathrm{T}}^{-}(t)$, and Coulomb model of friction. In this section the basic models are presented, and some extensions which allow one to take into account more mechanical effects are also examined. Within the framework of multibody systems simulation that is the topic of this book, any contact/impact models should satisfy the following properties:

(i) Dissipativity or, more generally, thermodynamical consistency (see Sect. 3.9.4.1)
(ii) Multivalued property at zero tangential relative velocity
(iii) Keep the number of parameters as low as possible
(iv) Parameters with mechanical meaning and identifiable from experiments in a reliable way
(v) Numerical tractability

3.9.1 Coulomb's Friction

We already presented some models of friction in Chap. 1, in the 2-dimensional case. We now focus on the 3-dimensional Coulomb friction. Only the fundamental basic model is presented here, see Chap. 13 for more details linked to the numerical implementation.

3.9.1.1 Three-Dimensional Coulomb's Friction

The notations are those of Sect. 3.3. We assume that the gap is closed, i.e., $P = P'$ and the two bodies touch at P, with a tangent plane of contact spanned by $\mathbf{t}$ and $\mathbf{s}$. Coulomb's model links the reaction force $R \in \mathbb{R}^3$ to the tangential relative velocity[9] $U_{\mathrm{T}}(\cdot) \in \mathbb{R}^2$, through the friction cone $\mathbf{C}$. The cone $\mathbf{C}$ is a second-order convex cone with its apex at the contact point P, whose sections by planes parallel to the tangent plane are discs, and the angle between the normal $\mathbf{n}$ and any vector PM with M on the boundary of $\mathbf{C}$ is equal to $\arctan\mu$. The coefficient $\mu \geqslant 0$ is the friction coefficient. The Coulomb friction cone is depicted in Fig. 3.5.

Coulomb's friction says the following. If $g(q) = 0$ then

$$\begin{cases} \text{If } U_{\mathrm{T}}(t) = 0 \text{ then } R \in \ \mathbf{C} \\ \\ \text{If } U_{\mathrm{T}}(t) \neq 0 \text{ then } ||R_{\mathrm{T}}(t)|| = \mu |R_{\mathrm{N}}| \text{ and there exists a scalar } a \geqslant 0 \\ \\ \qquad\qquad \text{such that } R_{\mathrm{T}}(t) = -aU_{\mathrm{T}}(t) \end{cases} \tag{3.147}$$

where we recall that in the above notation $R_{\mathrm{T}} \in \mathbb{R}^2$, $R_{\mathrm{N}} \in \mathbb{R}$. Thus Coulomb's model says that if the sliding velocity is not zero, then the reaction R lies on the boundary of $\mathbf{C}$, and its projection on the tangent plane has the same direction as but opposite sense to the sliding velocity. When the sliding velocity is zero, R is in $\mathbf{C}$, possibly on its boundary. The fact that $U_{\mathrm{T}} = 0$ and R be on $\partial\mathbf{C}$ is therefore a necessary condition to have a transition from sticking to sliding. It is noteworthy that the dynamics of a system together with (3.147) defines an implicit dynamics (this is more visible with (3.150) below). A sliding case is depicted in Fig. 3.6.

The Coulomb model is also often written as follows:

$$||R_{\mathrm{T}}(t)|| \leqslant \mu |R_{\mathrm{N}}| \text{ and } \begin{cases} ||R_{\mathrm{T}}(t)|| < \mu |R_{\mathrm{N}}| \Rightarrow U_{\mathrm{T}}(t) = 0 \\ \\ ||R_{\mathrm{T}}(t)|| = \mu |R_{\mathrm{N}}| \Rightarrow \text{ and there exists a scalar } b \geqslant 0 \\ \\ \qquad\qquad \text{such that } U_{\mathrm{T}}(t) = -bR_{\mathrm{T}}(t)\ . \end{cases} \tag{3.148}$$

[9] That one may also call the sliding velocity.

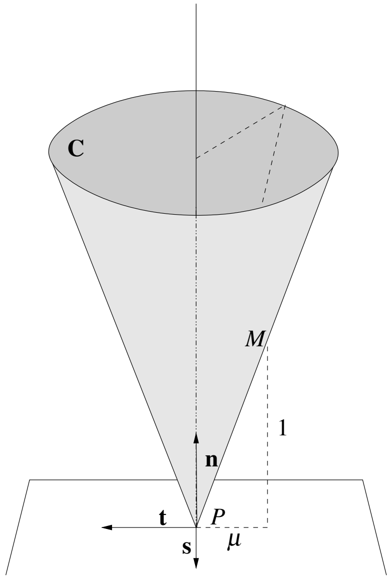


Fig. 3.5. Three-dimensional Coulomb's friction cone

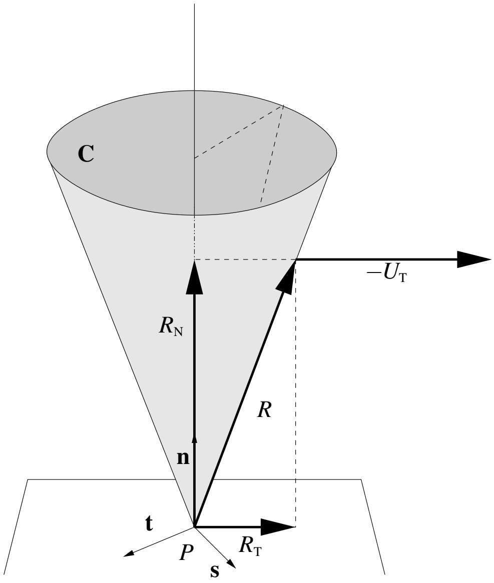


Fig. 3.6. Coulomb's friction. The sliding case

Coulomb's Friction as an Inclusion

Let $\mathbf{D}_1$ be a given closed convex subset of the common tangent plane between the two contacting bodies. Let $\mathbf{D} = \mu \mathbf{D}_1$. Let us introduce the following inclusion (Moreau, 1988b) using the indicator function $\psi_{\mathbf{D}}(\cdot)$:

$$-U_{\mathrm{T}} \in \partial \psi_{\mathbf{D}}(R_{\mathrm{T}}) \,. \tag{3.149}$$

The meaning of (3.149) is as follows. If $R_{\mathrm{T}} \in \mathrm{Int}(\mathbf{D})$, then the normal cone to $\mathbf{D}$ is the singleton $\{0\}$, so the tangential relative velocity is null. If $R_{\mathrm{T}} \in \partial \mathbf{D}$, the boundary of $\mathbf{D}$, then $-U_{\mathrm{T}}$ is in the normal cone to $\mathbf{D}$ computed at R_{T}. Let $\mathbf{D}_1$ be a disc with radius $|R_{\mathrm{N}}|$, so that when R_{T} is on the boundary of $\mathbf{D}$ one has $||R_t|| = \mu |R_n|$. Then the normal cone to $\mathbf{D}$ at R_{T} is nothing else but the ray passing through the center of the disc (the apex of the cone $\mathbf{C}$) and whose direction is that of R_{T}. If we denote $\mathbf{d} = \frac{R_{\mathrm{T}}}{||R_{\mathrm{T}}||}$ the sliding direction, then $-U_{\mathrm{T}} = b\mathbf{d}$ for some real $b \geqslant 0$. One sees that in this case (3.149) does represent the model in (3.148). When $\mathbf{D}_1$ is not a disc the model may incorporate anisotropic effects (the friction coefficient may vary with the direction of sliding).

Since $\mathbf{D}$ is nonempty closed convex, one may use convex analysis to rewrite (3.149) in its dual form that can be inserted in the dynamics:

$$R_{\mathrm{T}} \in \partial \psi^*_{\mathbf{D}}(-U_{\mathrm{T}}) \,. \tag{3.150}$$

More rigorously one should write (3.150) with the right velocity U_{T}^+, since there may exist velocity jumps. When the friction is isotropic, one has $\psi^*_{\mathbf{D}}(\cdot) = \mu \ ||\cdot||$. This function is called a dissipation function. Starting from (3.150) one may recover (3.147). One deduces that both ways of writing the Coulomb model as in (3.147) or (3.148) are equivalent.

It is also possible to formulate the friction model at the acceleration level (Glocker, 2001). In the 3-dimensional case, formulating the Coulomb friction with complementarity relations is less easy than in the 2-dimensional case. See Chap. 13 where various solutions for the numerical implementation are described.

Coulomb's Friction as a VI

Let us end this section with some other formulations of Coulomb's friction, using some equivalences in Sect. A.3. Then (3.149) appears to be equivalent to

$$\begin{cases} R_{\mathrm{T}} \in \mathbf{D} \\ \langle U_{\mathrm{T}}, z - R_{\mathrm{T}} \rangle \geqslant 0 \text{ for all } z \in \mathbf{D} \end{cases} \tag{3.151}$$

and to

$$R_{\mathrm{T}} = \mathrm{proj}_{\mathbf{D}}[R_{\mathrm{T}} - \rho U_{\mathrm{T}}], \text{ for all } \rho > 0 \,. \tag{3.152}$$

Remark 3.15. The friction model may also be written in the configuration space of the generalized coordinates (Erdmann, 1994; Moreau, 1988b; Génot &

Brogliato, 1998 and Sect. 6.6 in Brogliato, 1999). Though this has little usefulness for the numerical implementation, it may be used to explain apparently paradoxical behaviors of mechanical systems with Coulomb friction and unilateral constraints, which are then interpreted as the generalized friction cone dipping in the constraints.

Remark 3.16. Another way to represent the 3-dimensional Coulomb friction has been proposed in Klarbring (1986a), Klarbring & Björkman (1988), and Stewart (2000). It will be described in detail in Chap. 13.

3.9.1.2 The Maximum Dissipation Principle

Let us consider (3.151). The inequality may be rewritten as

$$\langle U_{\mathrm{T}}, -R_{\mathrm{T}}\rangle \geqslant \langle U_{\mathrm{T}}, z\rangle \tag{3.153}$$

for all $z \in \mathbf{D}$, and where $R_{\mathrm{T}} \in \mathbf{D}$. This means that the power dissipated by the tangential component of the reaction is maximal w.r.t. all the powers that may be dissipated by other forces within the friction disk $\mathbf{D}$. Moreau named this the principle of maximal dissipation. It is crucial to remind that the maximization is done w.r.t. all forces in the disk $\mathbf{D}$, not w.r.t. all forces in the friction cone. Another interpretation is that the maximum dissipation principle is only valid when the normal reaction R_{N} is assumed to be given.

3.9.2 De Saxcé's Bipotential Function

Moreau's Superpotential

Let us make a brief summary of the different behavior laws one encounters in mechanics. Basically, there are primal variables (deformations, displacements, velocities, etc.), dual variables (stress, forces), and a scalar product (work, power, etc.). A fundamental function is the potential function. A potential function may be a differentiable function. For instance the potential of elasticity takes the general form

$$V(q) = -\frac{1}{2}q^{\mathrm{T}}Kq = -\langle F, q\rangle = -F^{\mathrm{T}}q$$

so that

$$F = -\frac{\partial V}{\partial q} = Kq,$$

with $K = K^{\mathrm{T}} > 0$ and q a displacement vector.

The potential function may also not be differentiable, but only subdifferentiable in the sense of convex analysis. For instance let

$$V(q) = \psi_K(q),$$

then

$$-F \in \partial \psi_K(q)$$

expresses, when K is a convex set, the complementarity between F and q: if q is in the interior of K, then $F = 0$. If q is on the boundary of K, then $-F$ belongs to the normal cone to K at q.

Such a potential function has been named a superpotential by Moreau (1974). More generally a superpotential is a lower semi-continuous proper convex function $\phi(\cdot)$ such that the inclusion $x \in \partial\phi(y)$ holds between two dual variables x and y and expresses some physical law.

By elementary convex analysis one has also $y \in \partial\phi^*(x)$, where $\phi^*(\cdot)$ is the conjugate function of $\phi(\cdot)$ (Hiriart-Urruty & Lemaréchal, 2001). The second inclusion expresses the so-called inverse law. It holds that

$$y \in \partial\phi^*(x) \;\Leftrightarrow\; x \in \partial\phi(y) \;\Leftrightarrow\; \phi(y) + \phi^*(x) = x^{\mathrm{T}} y\,. \tag{3.154}$$

The last equality is Fenchel's equality. It states that the couple (x,y) is extremal for the superpotential $\phi(\cdot)$. In the first example of the elasticity, the Fenchel's equality reads with $V(q) = -\frac{1}{2}q^{\mathrm{T}}Kq$ and $V^*(-F) = -\frac{1}{2}F^{\mathrm{T}}K^{-1}F$:

$$V(q) + V^*(-F) = (-F)^{\mathrm{T}}q,$$

where the two dual variables are q and $-F$. The second example gives

$$\psi_K(q) + \psi_K^*(-F) = -F^{\mathrm{T}}\, q.$$

De Saxcé's Bipotential

Let us consider now the Coulomb model in (3.147). Does there exist a superpotential for such a law? The answer is negative. The inclusion in (3.149) might let one think it is, but recall that $\mathbf{D}$ in (3.149) depends on R_{N}. One cannot find a proper convex function $\phi(\cdot)$ such that $-R \in \partial\phi(U)$ and $U \in \partial\phi^*(-R)$, and such that (3.154) is satisfied. But, an extension exists where one replaces the superpotential by a *bipotential*.

The bipotential was introduced in De Saxcé (1992). Let us make a rough introduction to bipotentials.

Definition 3.17. *A bipotential is a function* $b\colon \mathrm{I\!R}^n \times \mathrm{I\!R}^n \to \mathrm{I\!R} \cup \{+\infty, -\infty\}$, $(x,y) \mapsto b(x,y)$, *such that*

- $b(\cdot,\cdot)$ *is convex w.r.t. x for fixed y, and convex w.r.t. y when x is fixed.*
- *For all couples* $(x,y) \in \mathrm{I\!R}^n \times \mathrm{I\!R}^n$ *one has* $b(x,y) \geqslant x^T y$.

A pair of dual variables (x,y) *is said extremal if the equality holds, i.e.,* $b(x,y) = x^T y$.

By definition, any extremal pair satisfies

$$\begin{cases} \text{For all } x' : \;\; b(x',y) - b(x,y) \geqslant y^{\mathrm{T}}(x'-x) \\[2ex] \text{For all } y' : \;\; b(x,y') - b(x,y) \geqslant x^{\mathrm{T}}(y'-y). \end{cases} \tag{3.155}$$

One recognizes that (3.155) can be formulated in terms of subdifferentials (see (1.5)) as:

$$\begin{cases} x \in \partial_y b(x,y) \\ \\ y \in \partial_x b(x,y) \end{cases} \tag{3.156}$$

The bipotential has been introduced in De Saxcé (1992, 1995) and De Saxcé & Feng (1991) for $x = v$ a velocity, and $y = r$ a force or for nonassociated behavior law of materials with $x = \dot{\varepsilon}$ a strain rate tensor and $y = \sigma$ the Cauchy stress tensor.

Bipotential Formulation of Coulomb's Friction

Let us now show that the Coulomb friction may be expressed with a bipotential function. The notation in the following theorem is as in Sect. 3.3, where the contact between two bodies O and O' is considered.

Theorem 3.18. *(De Saxcé, 1992) The function*

$$b(-U,R) = \psi_{\mathbb{R}^-}(-U_N) + \psi_{\mathbf{C}}(R) + \mu R_N ||U_T|| \tag{3.157}$$

is a bipotential. Moreover the extremal pairs $(-U,R)$ of this bipotential satisfy the Coulomb friction model equations, i.e.,

$$\begin{cases} -U \in \partial_R b(-U,R) \\ \\ R \in \partial_{-U} b(-U,R)\ . \end{cases} \tag{3.158}$$

It may be checked that the bipotential function in (3.157) can also be written as:

$$b(-U,R) = \begin{cases} \mu\ R_{\mathrm{N}}||U_{\mathrm{T}}|| \text{ if } R \in \mathbf{C} \text{ and } U_{\mathrm{N}} \geqslant 0 \\ \\ +\infty \qquad\quad \text{otherwise}\ . \end{cases} \tag{3.159}$$

The proof of Theorem 3.18 uses the fact that an equivalent expression of Coulomb's friction is

$$-(U_{\mathrm{N}} + \mu\ ||U_{\mathrm{T}}||, U_{\mathrm{T}})^{\mathrm{T}} \in \partial\psi_{\mathbf{C}}(R)\ . \tag{3.160}$$

Indeed let R be inside $\mathbf{C}$. Then the right-hand side of (3.160) is reduced to $\{0\}$ so that $U_{\mathrm{T}} = 0$ and $U_{\mathrm{N}} = 0$. When $R \in \partial\mathbf{C}$ and $R \neq 0$, then there exists $\gamma \geqslant 0$ such that $U_{\mathrm{N}} + \mu||U_{\mathrm{T}}|| = \gamma\mu$ and $-U_{\mathrm{T}} = \gamma\frac{R_{\mathrm{T}}}{||R_{\mathrm{T}}||}$. The second equality comes from the last two components of the inclusion (3.160) that is an inclusion quite the same as (3.149). Thus $||U_{\mathrm{T}}|| = \gamma$ so that from the first equality $U_{\mathrm{N}} = 0$: there is sliding between both bodies at the contact point $P = P'$.

Coulomb's Friction as a Second-Order Cone Complementarity Problem

Introducing the polar cone of **C**, denoted by $\mathbf{C}^\circ$

$$\mathbf{C}^\circ = \{v \in \mathbb{R}^n \mid r^T v \leqslant 0, \forall r \in \mathbf{C}\} \,, \tag{3.161}$$

the last inclusion (3.160) can be written as a second-order cone complementarity problem ,

$$\mathbf{C}^\circ \ni -[U_{\mathrm{N}} + \mu \, ||U_{\mathrm{T}}||, U_{\mathrm{T}}]^T \perp R \in \mathbf{C} \,. \tag{3.162}$$

Finally, the second-order cone complementarity problem will be described with the help of the dual cone (opposite of the polar cone),

$$\mathbf{C}^* = \{v \in \mathbb{R}^n \mid r^T v \geqslant 0, \forall r \in \mathbf{C}\} \,, \tag{3.163}$$

as

$$\mathbf{C}^* \ni [U_{\mathrm{N}} + \mu \, ||U_{\mathrm{T}}||, U_{\mathrm{T}}]^T \perp R \in \mathbf{C} \,. \tag{3.164}$$

In Chaps. 10 and 13, the following notation will be introduced for the modified velocity $\widehat{U}$ defined by

$$\widehat{U} = [U_{\mathrm{N}} + \mu \, ||U_{\mathrm{T}}||, U_{\mathrm{T}}]^T \,. \tag{3.165}$$

This notation provides us with a synthetic form of the Coulomb friction as

$$-\widehat{U} \in \partial \psi_{\mathbf{C}}(R) \,, \tag{3.166}$$

or

$$\mathbf{C}^* \ni \widehat{U} \perp R \in \mathbf{C} \,. \tag{3.167}$$

These relations are depicted in Fig. 3.7 in the sliding case 3.7.

De Saxcé's bipotential function yields specific numerical algorithms that will be described in Sect. 13.7.

3.9.3 Impact with Friction

The case of impact with friction is of great interest and the mixing of both models is not obvious. Such models have been proposed in Moreau (1988b, 1994b), Pfeiffer & Glocker (1996), and Payr & Glocker (2005). It is for instance proposed in Moreau (1988b) to extend (3.149) and (3.150) to densities, i.e., to impulses.

In Moreau (1988b, 1994b), tangential coefficients of restitution are introduced. We denote P_{N} the normal impulse of the force R_{N}, and P_{T} the tangential impulse of the force R_{T}, at the contact point P. Let $\mathbf{D} = \{-P_{\mathrm{T}} \mid ||P_{\mathrm{T}}|| \leqslant \mu \, P_{\mathrm{N}}\}$.[10] Then at time t of the impact

[10] The absolute value on the normal component is not useful here as this is always nonnegative, see the first inclusion in (3.170).

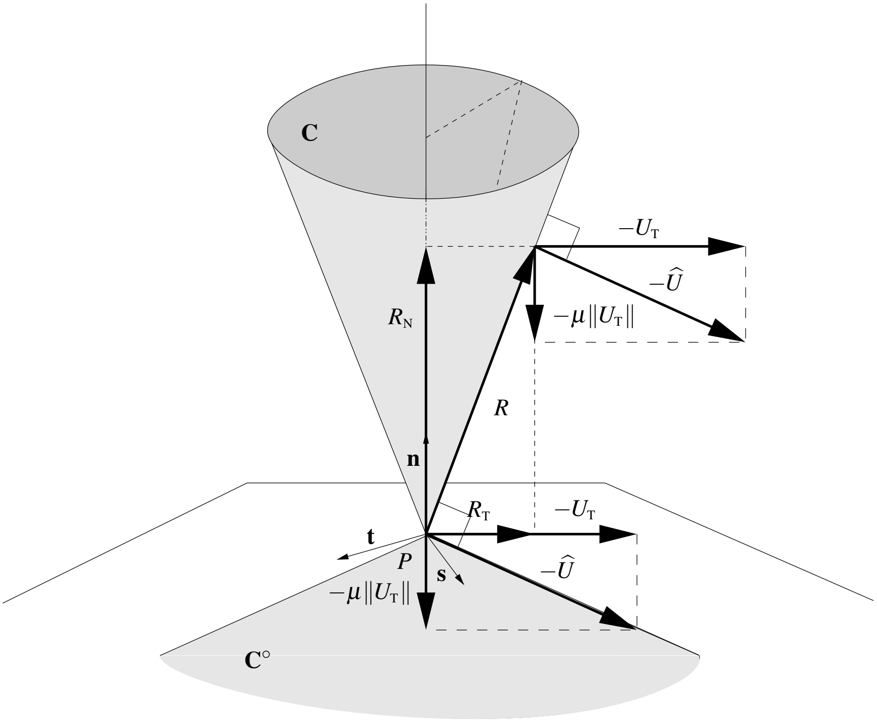


Fig. 3.7. Coulomb's friction and the modified velocity $\widehat{U}$. The sliding case.

$$\begin{cases} -P_N \in \partial\psi^*_{\mathbb{R}^-}\left(\dfrac{1}{1+\rho}U_N^+(t) + \dfrac{\rho}{1+\rho}U_N^-(t)\right) \\ -P_T \in \partial\psi^*_{\mathbf{D}}\left(\dfrac{1}{1+\tau}U_T^+(t) + \dfrac{\tau}{1+\tau}U_T^-(t)\right)\ . \end{cases} \tag{3.168}$$

Here ρ and τ are constants with values in the interval $[0,1]$. These laws can be reformulated equivalently in a more common form as

$$\begin{cases} -P_N \in \partial\psi^*_{\mathbb{R}^-}(U_N^+(t) + e_N U_N^-(t)) \\ -P_T \in \partial\psi^*_{\mathbf{D}}(U_T^+(t) + e_T U_T^-(t))\ , \end{cases} \tag{3.169}$$

where $e_N \in [0,1)$ and $e_T \in (-1,1)$ are the normal and tangential restitution coefficients.

When the motion is continuous similar inclusions are proposed replacing the impulses by the forces. Notice that we can equivalently write (3.168) in its conjugate form as

$$\begin{cases} U_N^+(t) + e_N U_N^-(t) \in \partial\psi_{\mathbb{R}^-}(-P_N) \\ U_T^+(t) + e_T U_T^-(t) \in \partial\psi_{\mathbf{D}}(-P_T) \end{cases} \tag{3.170}$$

thanks to the convexity of the sets $\mathbb{R}^-$ and $\mathbf{D}$. It follows immediately from the first inclusion in (3.170) that $P_{\mathrm{N}} > 0 \Rightarrow U_{\mathrm{N}}^+(t) + e_{\mathrm{N}} U_{\mathrm{N}}^-(t) = 0$. In fact, the first inclusion is equivalent to

$$0 \leqslant U_{\mathrm{N}}^+(t) + e_{\mathrm{N}} U_{\mathrm{N}}^-(t) \perp P_{\mathrm{N}} \geqslant 0 \,. \tag{3.171}$$

Let $\mu = 0$, then $\mathbf{D} = \{0\}$ so that $P_{\mathrm{T}} = 0$ as well. The inclusions (3.168) can be inserted in the right-hand side of (3.81) to deduce the jump of the velocity of the center of mass G and in the instantaneous angular velocity Ω.

This model involves three parameters per contact point: e_{N}, e_{T}, μ. The tangential restitution e_{T} is introduced as an additional parameter that permits to handle special problems such as the super ball rebound, which otherwise cannot be simulated. Obviously e_{T} and μ are linked in the sense that $\mu = 0 \Rightarrow e_{\mathrm{T}} = -1$, so that $U_{\mathrm{T}}^+(t) = U_{\mathrm{T}}^-(t)$. Consequently the triple $(e_{\mathrm{N}}, -1, 0)$ yields a frictionless impact with Newton's kinematic law.

Energy Balance Considerations

It is well known that Newton's impact law and Coulomb's friction yields the violation of the dissipation principle. Various authors (Brach, 1990; Wang & Mason, 1992; Stronge, 2000) stress that the system energy may increase in collisions with friction. Indeed, if a single coefficient $\rho = \tau$ is used, Moreau (1994b) derived the energy balance of a collision for a multi-contact system in the form

$$\mathscr{E}^- - \mathscr{E}^+ = \frac{1}{2}(v^+ - v^-)^{\mathrm{T}} M(q,t)(v^+ - v^-)^{\mathrm{T}} \delta - \sum_{\alpha} U^{\alpha,\mathrm{T}} P^{\alpha} \tag{3.172}$$

with $\delta = (1-\rho)/(1+\rho) \in [0,1]$ called the dissipation index. If the mass matrix $M(q,t)$ is Positive Semi–Definite (PSD), the quadratic term is positive. The term $-\sum\limits_{\alpha} U^{\alpha,T} P^{\alpha}$ is always positive respecting the principle of dissipation. Chosing $\delta \in$ [0,1], ensures that the energy decreases.

This is why one has to impose some conditions to guarantee that (3.168) defines a dissipative mapping, see, e.g., Sect. 4.2 in Brogliato (1999) for an extended overview of impact models that mix normal laws with Coulomb friction at the impulse level and more details on 3-parameter impact laws. Let us say that (3.168) provides a framework for impacts with friction that extends the sweeping process rule. In Leine & van de Wouw (2007) conditions are given that assure the dissipativity of (3.168).

3.9.4 Enhanced Contact Models

Let us now provide some few examples of contact/impact models that fit within the framework of nonsmooth systems. The challenge is to derive contact laws that lend themselves to a reliable numerical treatment, i.e., which can be solved with CP or QP solvers.

3.9.4.1 A Thermo-mechanical Framework

As for the behavior law of materials, a very useful and rigorous framework for writing enhanced contact laws is provided by the thermodynamics or precisely, the thermodynamics of continuum media (see standards references in Germain et al., 1983). This framework can be extended to surface behavior laws by postulating the existence of free energy and (pseudo- or super-) potentials of dissipation of surfaces. Without entering into deep details, we give here the basic recipes to write consistent contact laws. A very complete exposition of a coherent thermodynamical approach for the modeling of contact problems may be found in Frémond (2002).

Let us consider that the surface is described by the following variables:

- A set of state (external) variables g, g_{T} and its derivatives $U_{\mathrm{N}}, U_{\mathrm{T}}$. The variable g corresponds usually to the gap function, i.e., $g = g(q)$ and $\dot{g} = U_{\mathrm{N}}$. It can be enhanced to take into account an initial thickness of the interface. Note that the definition of the so-called tangential displacement g_{T} is not a straightforward task. The definition of a reference point $g_{\mathrm{T}}(t_0)$ to define the g_{T} from $\dot{g}_{\mathrm{T}} = U_{\mathrm{T}}$ needs usually the contact to be initially closed.
- Together with these external variables, a set of internal variables $\beta \in \mathrm{IR}^{\eta}$ is added to describe internal physical phenomena at the surface (wear, damage, etc.).

The dual variables to the external and internal variables are, respectively, the contact forces $R_{\mathrm{N}}, R_{\mathrm{T}}$ and the thermodynamical forces associated to β denoted by X_{β}. Usually, the forces are decomposed in reversible parts denoted by the superscript R and irreversible parts denoted by the superscript IR such that

$$\begin{cases} R_{\mathrm{N}} = R_{\mathrm{N}}^{\mathrm{R}} + R_{\mathrm{N}}^{\mathrm{IR}} \\ R_{\mathrm{T}} = R_{\mathrm{T}}^{\mathrm{R}} + R_{\mathrm{T}}^{\mathrm{IR}} \,. \end{cases} \tag{3.173}$$

The state laws can be written by postulating the existence of a surface free energy density $W_S(g, g_{\mathrm{T}}, \beta)$ and the derivation rule

$$\begin{cases} -R_{\mathrm{N}}^{\mathrm{R}} \in \partial_g W_S(g, g_{\mathrm{T}}, \beta) \\ -R_{\mathrm{T}}^{\mathrm{R}} \in \partial_{g_{\mathrm{T}}} W_S(g, g_{\mathrm{T}}, \beta) \\ -X_{\beta} \in \partial_{\beta} W_S(g, g_{\mathrm{T}}, \beta) \,. \end{cases} \tag{3.174}$$

From the second principle of thermodynamics, the local Clausius–Duhem inequality postulates that the intrinsic mechanical dissipation (ignoring for a moment the thermal dissipation) should be positive, that is

$$-\dot{W}_S + R^{\mathrm{T}} U \geqslant 0 \,. \tag{3.175}$$

We assume that the total derivative with respect to time of W_S denoted by $\dot{W}_S$ can be written as

$$\begin{aligned}\dot{W}_S &= \dot{g}^{\mathrm{T}}\partial_g W_S + \dot{g}_{\mathrm{T}}^{\mathrm{T}}\partial_{g_{\mathrm{T}}} W_S + \dot{\beta}^{\mathrm{T}}\partial_\beta W_S \\ &= U_{\mathrm{N}}^{\mathrm{T}}\partial_g W_S + U_{\mathrm{T}}^{\mathrm{T}}\partial_{g_{\mathrm{T}}} W_S + \dot{\beta}^{\mathrm{T}}\partial_\beta W_S\,.\end{aligned} \tag{3.176}$$

Due to the choice of the state laws (3.174), the time derivative of W_S can be written

$$\dot{W}_S = -(R_{\mathrm{N}}^{\mathsf{R}}U_{\mathrm{N}} + R_{\mathrm{T}}^{\mathsf{R},\mathrm{T}}U_{\mathrm{T}} + X_\beta^{\mathrm{T}}\dot{\beta})\,. \tag{3.177}$$

The local form of Clausius–Duhem inequality is therefore

$$-(R_{\mathrm{N}}^{\mathsf{IR}}U_{\mathrm{N}} + R_{\mathrm{T}}^{\mathsf{IR},\mathrm{T}}U_{\mathrm{T}}) + X_\beta^{\mathrm{T}}\dot{\beta} \geqslant 0\,. \tag{3.178}$$

A more simple way to ensure that the mechanical dissipation is positive and satisfy (3.178) is to postulate the existence of a surface pseudo-potential of dissipation $\Phi_S(g, g_{\mathrm{T}}, \beta, U_{\mathrm{N}}, U_{\mathrm{T}}, \dot{\beta})$ which is a lower semi-continuous proper convex function. The constitutive laws are written as follows:

$$\begin{cases} -R_{\mathrm{N}}^{\mathsf{IR}} \in \partial_{U_{\mathrm{N}}}\Phi_S(g, g_{\mathrm{T}}, \beta, U_{\mathrm{N}}, U_{\mathrm{T}}, \dot{\beta}) \\ -R_{\mathrm{T}}^{\mathsf{IR}} \in \partial_{U_{\mathrm{T}}}\Phi_S(g, g_{\mathrm{T}}, \beta, U_{\mathrm{N}}, U_{\mathrm{T}}, \dot{\beta}) \\ -X_\beta \in \partial_{\dot{\beta}}\Phi_S(g, g_{\mathrm{T}}, \beta, U_{\mathrm{N}}, U_{\mathrm{T}}, \dot{\beta})\,. \end{cases} \tag{3.179}$$

Remark 3.19. The standard frictional and unilateral contact law correspond to the choice

$$\begin{aligned} W_S(g) &= \psi_{\mathrm{IR}_+}(g) \\ \Phi_S(U_{\mathrm{T}}) &= \mu R_{\mathrm{N}}\|U_{\mathrm{T}}\|. \end{aligned} \tag{3.180}$$

As we said before, the formulation in terms of pseudo-potentials of the Coulomb's law is not adequate because the normal reaction appears in the right-hand side of the second equation.

3.9.4.2 Enhanced Coulomb's friction with Elasticity and Damping

Clearly Coulomb friction, despite being a complex contact law, is sometimes not rich enough to represent some phenomena. Let us illustrate briefly how one may enrich it without losing its nice property of being tractable with complementarity tools.

In the simplest case of Coulomb friction, one has

$$R_{\mathrm{T}}^{\mathsf{R}} = 0 \quad \text{and} \quad R_{\mathrm{T}}^{\mathsf{IR}} \in \partial\psi_{\mathbf{D}}^*(-U_{\mathrm{T}})\,. \tag{3.181}$$

We say that $R_{\mathrm{T}}^{\mathsf{IR}}$ derives from the nonsmooth potential function (the superpotential) $\psi_{\mathbf{D}}^*(-U_{\mathrm{T}})$.

Let us consider now $R_{\mathrm{T}}^{\mathsf{R}} = -kg_{\mathrm{T}}$, where $\dot{g}_{\mathrm{T}} = U_{\mathrm{T}}$ and q_{T} is a "tangential" displacement. Then the reversible force derives from the smooth potential $W_S(g_{\mathrm{T}}) = -\frac{1}{2}g_{\mathrm{T}}^{\mathrm{T}}Kg_{\mathrm{T}}$. The enhanced Coulomb's law with elasticity may be written as follows:

$$\begin{cases} R_{\mathrm{T}}(t) \in -\mu |R_{\mathrm{N}}(t)|\, \partial |U_{\mathrm{T}}(t)| - K g_{\mathrm{T}}(t) \\ \dot{g}_{\mathrm{T}}(t) = U_{\mathrm{T}}(t)\,. \end{cases} \tag{3.182}$$

Doing so, one has added to the velocity/force multivalued characteristic a displacement/force characteristic, which may physically model some micro-displacements which occur *during the sticking phase*. Let us now consider that $R_{\mathrm{T}}^{\mathrm{IR}}$ derives from the superpotential

$$\Psi_S(U_t) = \psi_{\mathbf{D}}^*(-U_{\mathrm{T}}) - \frac{1}{2} U_{\mathrm{T}}^{\mathrm{T}} C U_t,$$

where C is a matrix of viscous friction (damping). One obtains:

$$\begin{cases} R_{\mathrm{T}}(t) \in -\mu |R_{\mathrm{N}}(t)|\, \partial |U_{\mathrm{T}}(t)| - C U_{\mathrm{T}}(t) - K g_{\mathrm{T}}(t) \\ \dot{g}_{\mathrm{T}}(t) = U_{\mathrm{T}}(t)\,. \end{cases} \tag{3.183}$$

Comments

It is noteworthy that the multivalued property of the characteristic $(U_{\mathrm{T}}, R_{\mathrm{T}}^{\mathrm{IR}})$ is kept, but the force varies linearly with the velocity outside zero. Obviously one may add other smooth potentials to $\psi_{\mathbf{D}}^*(-U_{\mathrm{T}})$ in order to take into account other force/velocity behaviors during the sliding phases.

One sees that neither (3.182) nor (3.183) are regularizations of the multivalued sector of the Coulomb's law: the fundamental multivalued property is kept and it assures that the sticking modes exist. In particular there is no spurious drift behavior such as sliding from zero initial velocity and with very small external actions, or contact force oscillations while in a sticking mode that may occur with other types of models, see, e.g., some comments in Dupont et al. (2002).

It is possible to choose a more complex potential for $R_{\mathrm{T}}^{\mathrm{IR}}$ in order to model Stribeck effects in sliding modes. Obviously one may also approximate the curves for the sliding modes with piecewise linear characteristic, and introduce supplementary Lagrange multipliers so that the whole model is treated in a complementarity framework.

Other types of friction models mixing dry friction and linear springs, known as the Persoz' gephyroidal models, are analyzed in Bastien & Lamarque (2007). They fit within the nonsmooth framework and are numerically tractable. Interestingly enough, quite similar models exist in electronics (Addi et al., 2007).

3.9.4.3 Notes and Comments on Friction Models

There are many different ways to model friction. Depending on the task and on the objective, the most appropriate model may vary a lot (see, e.g., the survey of Bona & Indri, 2005, in the field of robotics and Bliman & Sorine, 1995, in systems and control).

In multibody mechanics, it is often preferable to have a model with as less parameters as possible, and to keep the multivalued feature of Coulomb's model that is a fundamental physical feature. There are many reasons for this, some of which have been already pointed out in the book (problems of identification of the parameters, property of the model to lend itself to a reliable numerical treatment as a complementarity problem, avoiding stiff equations and regularization at zero velocity, reaching the sticking mode in finite time, avoiding contact force oscillations during the sticking phases, etc.). This is why macroscopic models such as Dahl, LuGre, Bliman-Sorine, Leuven etc. (see Bona & Indri, 2005) are rarely used in multibody mechanics, because they would not bring much to the field as they fail to respect all or some of the items (i)–(v) above. It is noteworthy that the efficiency of such tribological models has rarely (if never) been shown on systems with several contact points. Even for control and feedback stabilization issues, the Coulomb model may prove to be quite efficient and may yield robust stabilization solutions (Doris, 2007).

An interesting contribution is in Kikuuwe et al. (2005) where the implicit discretization of Coulomb friction is rediscovered and extended to more sophisticated multivalued nonsmooth models (equations (14a)–(14c) in Kikuuwe et al., 2005 exactly correspond to the procedure described in Fig. 1.17). The so-called Masing model (that consists of an interconnection of springs, dampers, and dry friction elements) and a modified Dahl's model are examined in Bastien et al. (2007). The viscous Masing model consists of a damper $c > 0$, a spring k_0, and a spring k in series with a dry friction element with coefficient μ, mounted in parallel. Its dynamics is given by an inclusion of the form

$$\begin{cases} -\dot{w}(t) + U_{\mathrm{T}}(t) \in N_{[-1,1]}\left(\frac{kw(t)}{\mu}\right) \\ R_{\mathrm{T}}(t) = kw(t) + k_0 g_{\mathrm{T}}(t) + cU_{\mathrm{T}} - k_0 l_0 \\ \dot{g}_{\mathrm{T}}(t) = U_{\mathrm{T}}(t)\,, \end{cases} \tag{3.184}$$

where l_0 is a spring-free length, $g_{\mathrm{T}} = g_{\mathrm{T},e} + g_{\mathrm{T},f}$, $g_{\mathrm{T},f}$ is the displacement of the dry friction element, $g_{\mathrm{T},e}$ is the displacement of the spring k, and $w = z_{\mathrm{T},e} - l$, l is a spring-free length.

As shown in Bastien et al. (2007), the inclusion (3.184) is of the type (2.48) with $g(t,x)$ being Lipschitz in x and with $\mathscr{L}^2$-bounded time derivatives. So it is wellposed. Analysis, simulations, and experiments are carried out on a belt tensioner system. The chosen numerical scheme is an implicit Euler method for the Masing model that is nonsmooth, and a multistep solver for the Dahl's model. It is concluded that the Masing model is better to reproduce stick–slip transitions. The modified Dahl's model is efficient for modeling the intermediate stick–slip state. Other models made of springs and dry friction elements exist, like the Prandtl's model, which also gives rise to a maximal monotone inclusion as (2.48).

A good account on rheological models (i.e., models made of springs, dampers, and Coulomb friction elements mounted in series and parallel interconnections) is given in Bastien et al. (2000). A friction model that extends Coulomb friction is presented in Leine & Glocker (2003), within the framework of multivalued mappings.

3.9.4.4 Cohesion Laws

We present in this section an example of complex contact laws written in the framework of the thermodynamics of continuum media. This is an adhesive contact model coupling closely unilateral contact, Coulomb's friction, and surface damage. It has been proposed by Cangémi (1997) (see also Cangémi et al., 1996; Raous et al., 1999) as an extension of Frémond's model of adherence with unilateral contact (Frémond, 1982, 1987, 1988) in the context of fiber/matrix interface in composite materials. This model has been further developed and studied by Monerie (2000) (see also Monerie & Raous, 1999; Chaboche et al., 2001) which demonstrates the usefulness and the ability of the model to predict complex fracture process and the very correlation with experimental data. In Acary (2001), Monerie & Acary (2001), Jean et al. (2001), and Acary & Monerie (2006), the cohesive zone model has been formulated in a nonsmooth dynamics context, and a numerical solving method is proposed.

We propose here to summarize the main equations of the model. Besides the standard state variable of the contact surface g and g_{T}, an adhesion variable $\beta \in [0,1]$ is introduced to model the rate of adherence of the surface. This variable plays the same role as the standard damage variable in continuum media. If $\beta = 1$, the interface is adhesive and sound. If $\beta = 0$, the interface is broken and there is no more adhesion. Furthermore, a nonlocal formulation can be proposed introducing the gradient of the adhesion variable $\nabla\beta$. In this section the time variable is dropped for convenience.

The state laws are stated as follows,

$$\begin{cases} -R_{\mathrm{N}}^{\mathrm{R}} \in \partial_g W_S(g, g_{\mathrm{T}}, \beta, \nabla\beta), \\ -R_{\mathrm{T}}^{\mathrm{R}} \in \partial_{g_{\mathrm{T}}} W_S(g, g_{\mathrm{T}}, \beta, \nabla\beta), \\ -X_\beta \in \partial_\beta W_S(g, g_{\mathrm{T}}, \beta, \nabla\beta), \\ -X_{\nabla\beta} \in \partial_{\nabla\beta} W_S(g, g_{\mathrm{T}}, \beta, \nabla\beta)\,. \end{cases} \tag{3.185}$$

with the following free energy density

$$W_S(g, g_{\mathrm{T}}, \beta, \nabla\beta) = \frac{1}{2}\beta^2 c_{\mathrm{N}} g^2 + \frac{1}{2}\beta^2 c_{\mathrm{T}} \|g_{\mathrm{T}}\|^2 - w h(\beta) + \frac{1}{2}k\|\nabla\beta\|^2 + \Psi_{\mathbb{R}^+}(g) + \Psi_{[0,1]}(\beta)\,. \tag{3.186}$$

The parameters of the free energy are

1. The initial stiffnesses c_{N} and c_{T}, homogeneous to elasticity modulus per unit of length
2. An energy w dissipated by the decohesion without the energy dissipated by the viscosity
3. A smooth function h which allows one to model the dissipated energy with respect to β
4. The energy associated with the gradient of adhesion, $\nabla\beta$ chosen as a quadratic function with parameter k

The constitutive laws are written as

$$\left\{\begin{array}{ll} R_{\mathrm{N}}^{\mathsf{IR}} & = 0 \\[2ex] -R_{\mathrm{T}}^{\mathsf{IR}} & \in \partial_{U_{\mathrm{T}}} \Phi_S(U_{\mathrm{T}}, \dot{\beta}) \\[2ex] (X_\beta - X_{\nabla\beta}) & \in \partial_{\dot{\beta}} \Phi_S(U_{\mathrm{T}}, \dot{\beta})\,, \end{array}\right. \tag{3.187}$$

where the pseudo-potential of dissipation is given by

$$\Phi_S(U_{\mathrm{T}}, \dot{\beta}) = \mu(\beta)(R_{\mathrm{N}}^{\mathsf{R}} + \beta^2 c_{\mathrm{N}} g)\|U_{\mathrm{T}}\| + \frac{b}{p+1}\|\dot{\beta}\|^{p+1} + \Phi_{\mathbb{R}^-}(\dot{\beta})\,, \tag{3.188}$$

where b is a viscosity parameter, $p \in \mathbb{N}$, and $\mu(\beta)$ is coefficient of friction which may depend on β. It allows especially to introduce the friction only when the interface is damaged.

Without entering into deeper details, the cohesive zone model leads to the following sets of equations:

1. Unilateral contact with elasticity and adhesion

$$(R_{\mathrm{N}}^{\mathsf{R}} + \beta^2 c_{\mathrm{N}} g) \geqslant 0, \qquad g \geqslant 0, \qquad (R_{\mathrm{N}}^{\mathsf{R}} + \beta^2 c_{\mathrm{N}} g)\, g = 0\,. \tag{3.189}$$

2. Coulomb's friction with elasticity and adhesion

$$\left\{\begin{array}{l} R_{\mathrm{T}}^{\mathsf{R}} = -\beta^2 c_{\mathrm{T}} g_{\mathrm{T}}, \qquad R_{\mathrm{N}}^{\mathsf{R}} = R_{\mathrm{N}}, \\ \bullet \quad \|U_{\mathrm{T}}\| > 0, \quad \|R_{\mathrm{T}}^{\mathsf{IR}}\| = \mu(\beta)(R_{\mathrm{N}}^{\mathsf{R}} + \beta^2 c_{\mathrm{N}} g), \quad \dfrac{R_{\mathrm{T}}^{\mathsf{IR}}}{\|R_{\mathrm{T}}^{\mathsf{IR}}\|} = -\dfrac{U_{\mathrm{T}}}{\|U_{\mathrm{T}}\|}, \\ \bullet \quad \|U_{\mathrm{T}}\| = 0, \quad \|R_{\mathrm{T}}^{\mathsf{IR}}\| < \mu(\beta)(R_{\mathrm{N}}^{\mathsf{R}} + \beta^2 c_{\mathrm{N}} g)\,. \end{array}\right. \tag{3.190}$$

3. Dynamics of the adhesive behavior

$$\left\{\begin{array}{ll} b\dot{\beta} = -\left[\left(w h'(\beta) - \beta(c_{\mathrm{N}} g^2 + c_{\mathrm{T}}\|g_{\mathrm{T}}\|^2) - k\nabla\beta\right)^-\right]^{(1/p)}, & \text{si } \beta \in [0,1] \\ b\dot{\beta} \leqslant -\left[\left(w h'(\beta) - \beta(c_{\mathrm{N}} g^2 + c_{\mathrm{T}}\|g_{\mathrm{T}}\|^2) - k\nabla\beta\right)^-\right]^{(1/p)}, & \text{si } \beta = 1\,. \end{array}\right. \tag{3.191}$$

Introducing the following change of variables

$$\widetilde{R}_{\mathrm{N}} = R_{\mathrm{N}} + \beta^2 c_{\mathrm{N}} g, \quad \widetilde{R}_{\mathrm{T}} = R_{\mathrm{T}} + \beta^2 c_{\mathrm{T}} g_{\mathrm{T}}\,, \tag{3.192}$$

the cohesive zone model can be recast into the synthetic form of the unilateral contact with Coulomb's friction,

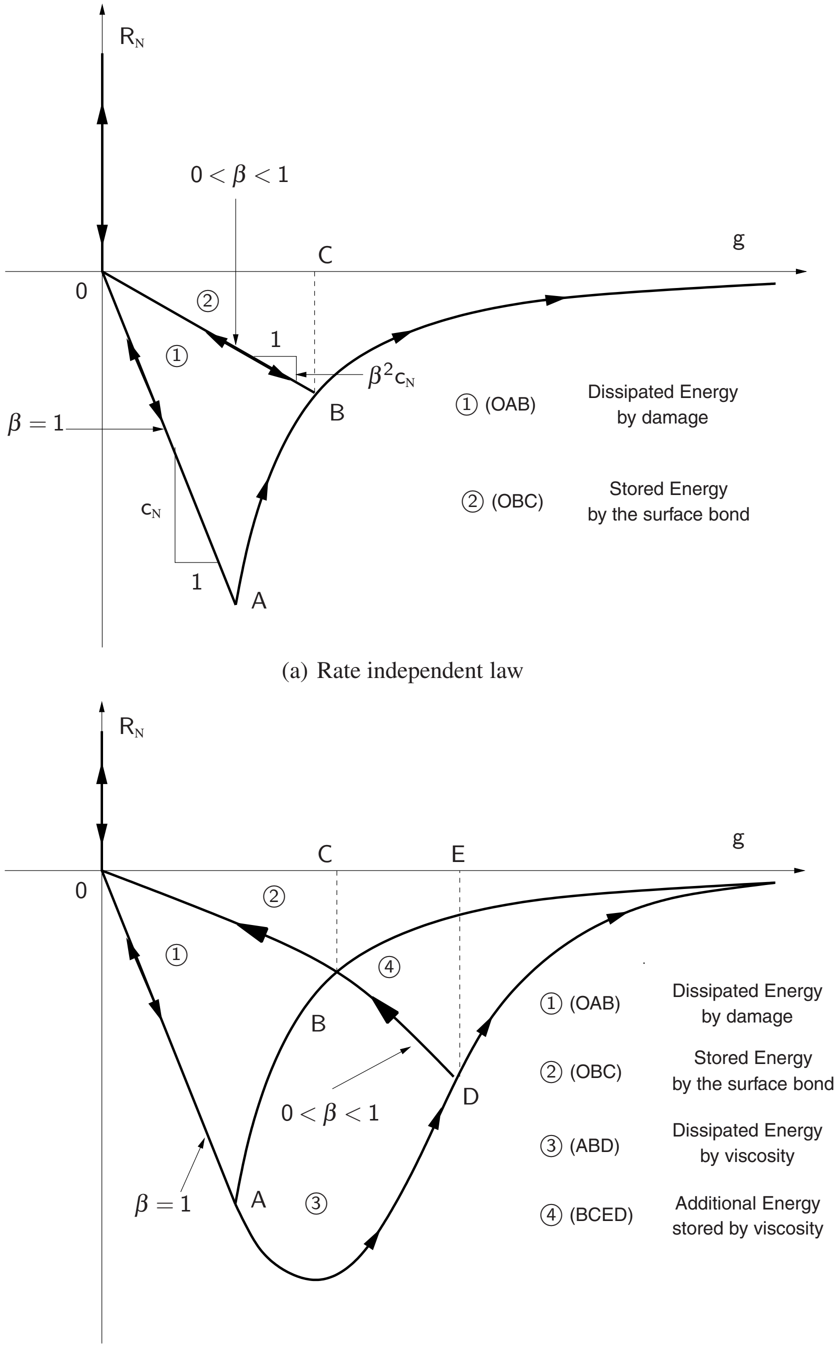
$$\left\{\begin{array}{l} -g \in \partial\psi_{\mathbb{R}^+}(\widetilde{R}_{\mathrm{N}}) \\ -U_{\mathrm{T}} \in \partial\psi_{\mathbf{D}(\mu(\beta)\widetilde{R}_{\mathrm{N}})}(\widetilde{R}_{\mathrm{T}}) \end{array}\right. \tag{3.193}$$

with the evolution of the adhesion variable

$$\dot{\beta} = f(\beta, g, g_{\mathrm{T}}). \tag{3.194}$$

This reformulation is a keystone of the numerical method for solving such a model.

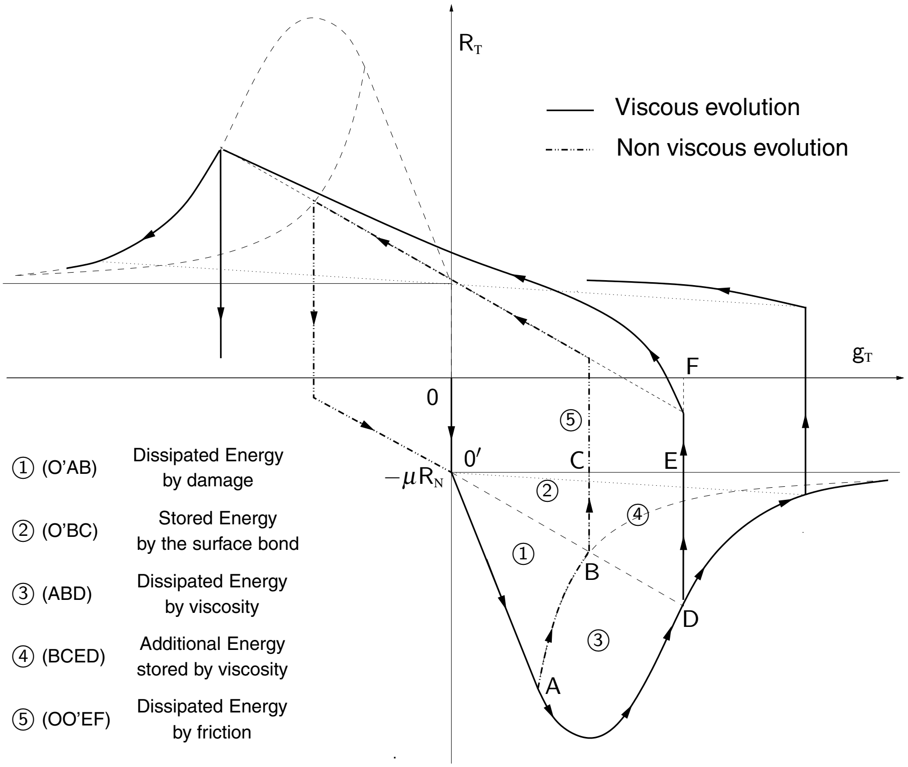
In Figs. 3.8 and 3.9, some responses of the model to periodic loading are depicted in order to give a flavor of the modeled behavior.


R_N
$0<\beta<1$
C
g
0
②
①
1
$\beta^2 c_N$
B
$\beta=1$
① (OAB)
Dissipated Energy by damage
② (OBC)
Stored Energy by the surface bond
c_N
1
A
R_N
C
E
g
0
②
①
④
B
D
$0<\beta<1$
$\beta=1$
A
③
① (OAB)
Dissipated Energy by damage
② (OBC)
Stored Energy by the surface bond
③ (ABD)
Dissipated Energy by viscosity
④ (BCED)
Additional Energy stored by viscosity

(a) Rate independent law

(b) Rate dependent law (viscosity)

Fig. 3.8. Uniaxial traction/compression test in the normal direction



(a) Viscous Law

Fig. 3.9. Periodic shear test

3.10 Lagrangian Systems with Frictional Unilateral Constraints and Newton's Impact Laws: Summary

To summarize, we give the equation of motion in the standard framework of the unilateral constraints with Newton's impact law and Coulomb's friction:

$$\begin{cases} M(q(t))dv + N(q(t),v^{+}(t))dt + F_{\mathrm{int}}(t,q(t),v^{+}(t))dt = F_{\mathrm{ext}}(t)dt + dr \\ q(t) = q(t_0) + \displaystyle\int_{t_0}^{t} v^{+}(t)dt \\ dr = \displaystyle\sum_{\alpha} dr^{\alpha} = \sum_{\alpha} H^{\alpha}(q)dR^{\alpha} = H(q)dR \\ U^{+} = [U^{\alpha,+}, \alpha \in \{1\dots\nu\}] \text{ with } U^{\alpha,+} = H^{\alpha,T}(q)v^{+} \\ \mathbf{C}^{\alpha,*} \ni [U_N^{\alpha,+} + e^{\alpha}U_N^{\alpha,-} + \mu^{\alpha}||\ U_T^{\alpha,+}||,\ U_T^{\alpha,+}]^T \perp dR^{\alpha} \in \mathbf{C}^{\alpha}, \quad \alpha \in \{1\dots\nu\} \\ \mathbf{C}^{\alpha} = \left\{X \in \mathbb{R}^3, ||X_T|| \leqslant \mu^{\alpha}|X_N|\right\} \quad \alpha \in \{1\dots\nu\} \end{cases} \tag{3.195}$$

This model can be enhanced with the nonsmooth contact models, especially with Poisson and Moreau's impact laws, that have been presented all along the Sect. 3.9. Note that we have assumed that Coulomb's friction is still valid for the impulses in case of impacts.

3.11 A Mechanical Filippov's System

Let us consider the system in Fig. 3.10, where the two friction cones are represented with dashed lines. Its dynamics is

$$\begin{cases} \ddot{q}_1(t) + k(q_1(t) - q_2(t)) \in -\mu \ \mathrm{sgn}(\dot{q}_1(t)) \\ \\ \ddot{q}_2(t) + k(q_2(t) - q_1(t)) \in -\mu \ \mathrm{sgn}(\dot{q}_2(t)) \\ \\ q_1(0) = q_{10}, q_2(0) = q_{20}, \dot{q}_1(0) = \dot{q}_{10}, \dot{q}_2(0) = \dot{q}_{20} \ , \end{cases} \tag{3.196}$$

where $\mu \geqslant 0$ is the coefficient of friction. It is assumed that both masses are $m = 1$. It can be verified that this inclusion satisfies the conditions of Theorem 2.41, so that there exists a unique global Lipschitz solution for any initial data. Let us perform the variable change $q = q_1 - q_2$, $z = q_1 + q_2$. One obtains:

$$\begin{cases} \ddot{z}(t) \in -\mu \ (\mathrm{sgn}(\dot{q}(t) + \dot{z}(t)) + \mathrm{sgn}(-\dot{q}(t) + \dot{z}(t)) \\ \\ \ddot{q}(t) + 2kq(t) \in -\mu \ (\mathrm{sgn}(\dot{q}(t) + \dot{z}(t)) - \mathrm{sgn}(-\dot{q}(t) + \dot{z}(t)) \ . \end{cases} \tag{3.197}$$

As long as $\mu > 0$ this system is dissipative. In fact considering the Lyapunov function candidate $V(q_1, q_2, \dot{q}_1, \dot{q}_2) = \frac{1}{2}\dot{q}_1^2 + \frac{1}{2}\dot{q}_2^2 + \frac{1}{2}k(q_1 - q_2)^2$, one finds along the trajectories of (3.196)

$$\dot{V}(t) \in -\mu \ |\dot{q}_1(t)| - \mu \ |\dot{q}_2(t)| \ . \tag{3.198}$$

Then from theorems 3.1 and 3.2 in Shevitz & Paden (1994), the fixed point $\dot{q}_1^* = 0$, $\dot{q}_2^* = 0$, $q_1^* = q_2^*$ is uniformly stable, and all trajectories converge asymptotically in the set $M = \{q_1, q_2, \dot{q}_1, \dot{q}_2 \ | \ \dot{q}_1 = \dot{q}_2 = 0, \ q_1 - q_2 \in [-\frac{\mu}{k}, \frac{\mu}{k}]\}$.

The two switching surfaces are $\Sigma_1 = \{(\dot{q}, \dot{z}) \ | \ \dot{q} = \dot{z}\}$ and $\Sigma_2 = \{(\dot{q}, \dot{z}) \ | \ \dot{q} = -\dot{z}\}$, as depicted in Fig. 3.11. The vector fields and the Filippov's convex set are also depicted in Fig. 3.11 for the value $k = 0$. When $k > 0$ the Filippov's inclusion is

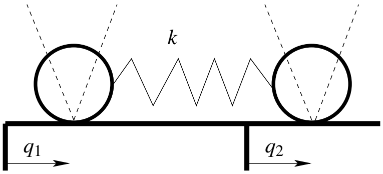


Fig. 3.10. Mechanical system with Coulomb friction

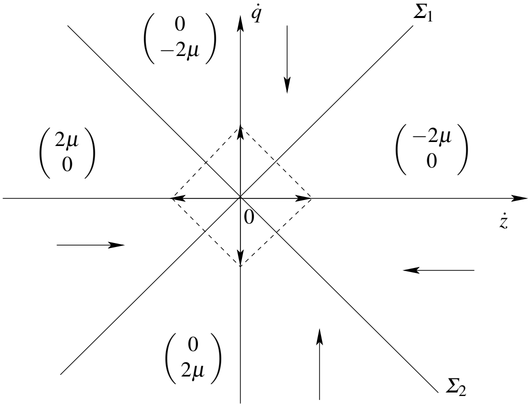

$\dot{q}$
Σ_1
$\begin{pmatrix} 0 \\ -2\mu \end{pmatrix}$
$\begin{pmatrix} 2\mu \\ 0 \end{pmatrix}$
$\begin{pmatrix} -2\mu \\ 0 \end{pmatrix}$
0
$\dot{z}$
$\begin{pmatrix} 0 \\ 2\mu \end{pmatrix}$
Σ_2

Fig. 3.11. Vector fields and inclusion at the origin ($k = 0$)

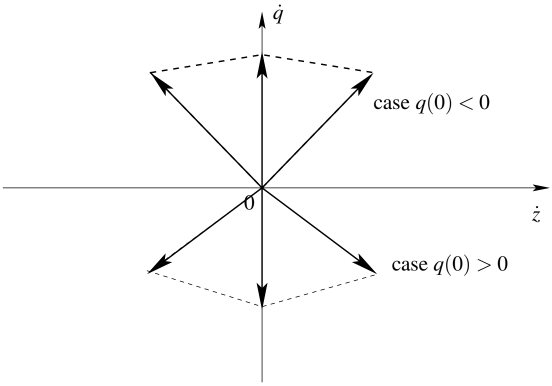


Fig. 3.12. Inclusion at the origin ($k > 0$)

depicted for different initial displacement conditions in Fig. 3.12. The origin in the figures correspond to a surface of codimension 2 in the state space $\mathbb{R}^4$. When $q(0) \neq 0$ the vertical component of the vector field has to be corrected by a value $-2kq(0)$, and the horizontal component remains unchanged. Notice that since uniqueness of solutions holds, repulsive switching surfaces with spontaneous switches cannot exist. Surfaces are crossed transversally, or are attractive.

This system has been analyzed in detail in Pratt et al. (2007), when one of the masses is acted upon by an external force. It is shown analytically that the trajectory undergoes an infinity of events (sticking–sliding transitions) for some critical values of the external load and particular initial data. This system is therefore presented as a possible benchmark to test the validity of numerical schemes.

4

Complementarity Systems

Complementarity systems have been introduced in Sect. 2.6, where some examples have been given. In this chapter they are examined in more details.

4.1 Definitions

Let us provide several formal definitions of complementarity systems. In view of the preceding chapters (Chap. 3 on Lagrangian systems, but see also Chap. 2), it should be clear that the dynamics which are written below may not be complete, in the sense that starting from $x(0)$, the ingredients that are proposed may not be sufficient to integrate the trajectories: it may even happen that one cannot start the system, i.e., the right limit of the state $x(0^+)$ cannot be calculated![1] Anyway, instead of adding what is missing (a state jump rule), we first give a partial description of the dynamics. Then the state jump rules will be examined in particular cases, in this chapter and in Chap. 5. In what follows $x(t) \in \mathrm{I\!R}^n$ and $w(t) \in \mathrm{I\!R}^m$. In all the definitions, an initial data is given as $x(0) = x_0$, without further constraint on it.

Definition 4.1 (Generalized dynamical complementarity systems). *A generalized dynamical complementarity system in a semi-explicit form is defined by*

$$\begin{cases} \dot{x}(t) = f(x(t), t, \lambda(t)) \\ \\ w(t) = h(x(t), \lambda(t)) \\ \\ K^* \ni w(t) \perp \lambda(t) \in K\,, \end{cases} \tag{4.1}$$

where the nonempty cone $K \subset \mathrm{I\!R}^n$ and $K^* = \{x \in \mathrm{I\!R}^n \mid x^{\mathrm{T}} y \geqslant 0$ for all $y \in K\}$ is its dual cone.

[1] Consider the bouncing ball with a constraint $q \geqslant 0$, with initial data $q(0) = 0$ and $\dot{q}(0^-) < 0$: without the velocity jump, time integration cannot proceed.

The term generalized is used when the set K is a general cone of $\mathbb{R}^n$. If this cone is a nonnegative orthant like $\mathbb{R}^N_+$, we speak of DCS or shortly CS.

Definition 4.2 (Dynamical complementarity systems). *A dynamical complementarity system (DCS) in an explicit form is defined by*

$$\begin{cases} \dot{x}(t) = f(x(t), t, \lambda(t)) \\ \\ w(t) = h(x(t), \lambda(t)) \\ \\ 0 \leqslant w(t) \perp \lambda(t) \geqslant 0\ . \end{cases} \tag{4.2}$$

If the smooth dynamics and the input/output function are linear, we speak of linear complementarity systems.

Definition 4.3 (Linear complementarity systems). *A linear complementarity system (LCS) is defined by*

$$\begin{cases} \dot{x}(t) = Ax(t) + B\lambda(t) \\ \\ w(t) = Cx(t) + D\lambda(t) \\ \\ 0 \leqslant w(t) \perp \lambda(t) \geqslant 0\ . \end{cases} \tag{4.3}$$

One may be led to work with more complex forms of linear complementarity systems such as

Definition 4.4 (Mixed linear complementarity systems). *A mixed linear complementarity system (MLCS) is defined by*

$$\begin{cases} E\dot{x}(t) = Ax(t) + B\lambda(t) \\ \\ Mw(t) = Cx(t) + D\lambda(t) \\ \\ 0 \leqslant w(t) \perp \lambda(t) \geqslant 0\ . \end{cases} \tag{4.4}$$

It is noteworthy that such formalisms naturally arise in the modeling of electrical circuits and are therefore not artificial. If both the matrices E and M are square full rank, we are back to an LCS as in Definition 4.3. See for instance example 7 in Brogliato (2003) for a system that fits within MLCS.

Finally, specific nonlinear complementarity systems are usually defined as follows:

Definition 4.5 (Nonlinear complementarity systems). *A nonlinear complementarity system (NLCS) is defined by*

$$\begin{cases} \dot{x}(t) = f(x(t),t) + g(x(t))\lambda(t) \\ \\ w(t) = h(x(t),\lambda(t)) \\ \\ 0 \leqslant w(t) \perp \lambda(t) \geqslant 0\,. \end{cases} \tag{4.5}$$

If $g(x) = -\nabla h(x)$, one obtains the so-called gradient-type complementarity systems which are defined as follows:

Definition 4.6 (Gradient complementarity system). *A gradient complementarity system (GCS) is defined by*

$$\begin{cases} \dot{x}(t) + f(x(t)) = \nabla g(x(t))\lambda(t) \\ \\ w(t) = g(x(t)) \\ \\ 0 \leqslant w(t) \perp \lambda(t) \geqslant 0\,. \end{cases} \tag{4.6}$$

More details on the definitions and the mathematical properties of CS can be found in Heemels (1999), Camlibel (2001), Heemels et al. (2000), van der Schaft & Schumacher (2000), Heemels & Brogliato (2003), and Brogliato (2003). The CSs presented in this section are autonomous. Obviously one may define non autonomous CS, with exogenous inputs. For instance the nonautonomous LCS dynamics is

$$\begin{cases} \dot{x}(t) = Ax(t) + B\lambda(t) + Eu(t) \\ \\ w(t) = Cx(t) + D\lambda(t) + Fu(t) \\ \\ 0 \leqslant w(t) \perp \lambda(t) \geqslant 0\,, \end{cases} \tag{4.7}$$

where the exogenous term $u(\cdot)$ should satisfy some regularity conditions, so that the well-posedness can be proved. Such LCS with inputs frequently occur in electrical circuits with current or voltage sources, see, e.g., Example 2.52.

4.2 Existence and Uniqueness of Solutions

The class of complementarity systems is too vast to allow one to state a general well-posedness result for *all* systems of Definitions 4.1–4.6 in one shot. Therefore one has to focus on subclasses of the above classes (notice that frictionless Lagrangian complementarity systems are one subclass of NLCS). The results given in this section concern only a subclass of LCS (passive LCS). Few words on passive NLCS are also written. This may appear quite narrow. The reader should, however, keep in mind that CSs are complex dynamical systems, and that the quantity of results one may obtain is inversely proportional to the degree of generality of the considered class of systems. Imposing linearity and passivity seriously narrows down the class, but permits to obtain precise results.

4.2.1 Passive LCS

We consider here the systems as in (4.3), where the quadruple (A,B,C,D) satisfies a passivity constraint, i.e., the LMI

$$\begin{pmatrix} A^{\mathrm{T}}P+PA & PB-C^{\mathrm{T}} \\ B^{\mathrm{T}}P-C & -D-D^{\mathrm{T}} \end{pmatrix} \leqslant 0 \tag{4.8}$$

has a solution $P=P^{\mathrm{T}} \geqslant 0$. The next theorem is taken from Heemels et al. (2002) and Camlibel et al. (2002b). The solutions are understood as Bohl distributions, see Appendix C.5.

Theorem 4.7. *Assume that (A,B,C,D) is passive, the pair (A,B) is controllable, the pair (C,A) is observable, and the matrix $\begin{pmatrix} B \\ D+D^T \end{pmatrix}$ has full column rank. Then the LCS (4.3) has a unique solution on $\mathrm{I\!R}^+$ such that λ is a Bohl distribution of degree 2 (i.e., a measure), and $x(\cdot)$ is a Bohl function that is in $\mathscr{L}^2(\mathrm{I\!R}^+,\mathrm{I\!R}^n)$. Moreover let $\lambda_{imp}=\lambda_0\delta_0$. Then the state initial jump is given by $x(0^+)=x(0^-)+B\lambda_0$.*

Similar well-posedness results hold in the nonautonomous case (4.7). However, in this case, there may exist state jumps at times $t>0$. This depends a lot on the regularity of the signal $u(\cdot)$. Roughly speaking the state may jump when the signal $u(\cdot)$ has a discontinuity. Let us provide some details on the state reinitializing mapping. A general jump rule is provided in Camlibel et al. (2002b), i.e., a rule that works for any matrix $D \geqslant 0$. For this we need to define the set Q_D of solutions of the LCP$(0,D)$, i.e., the LCP: $0 \leqslant \lambda \perp D\lambda \geqslant 0$. Its dual cone is $Q_D^* = \{x \in \mathrm{I\!R}^m \mid x^{\mathrm{T}}y \geqslant 0, \text{ for all } y \in Q_D\}$. For the sake of generality, the jump rules are presented for the nonautonomous LCS in (4.7). They concern possible state jumps at $t=0$, but can easily be generalized to any time $t>0$ at which a jump is needed:

(i) The jump multiplier λ_0 is uniquely determined as the solution of the complementarity problem

$$Q_D \ni \lambda_0 \perp Cx(0^-)+Fu(0)+CB\lambda_0 \in Q_D^* \,. \tag{4.9}$$

(ii) The post-initialization state $x(0^+)$ is the unique minimizer of

$$\begin{aligned} &\min\ \tfrac{1}{2}(z-x(0^-))^{\mathrm{T}}P(z-x(0^-)) \\ &\text{subject to: } Cz+Fu(0) \in Q_D^* \,, \end{aligned} \tag{4.10}$$

where P is any positive definite solution of the Kalman–Yakubovic–Popov Lemma LMI in (4.8).

(iii) The jump multiplier is the unique minimizer of

$$\begin{aligned} &\min\ \tfrac{1}{2}(x(0^-)+Bz)^{\mathrm{T}}P(x(0^-)+Bz)+z^{\mathrm{T}}Fu(0) \\ &\text{subject to: } z\in Q_D\ , \end{aligned} \tag{4.11}$$

where P is any positive definite solution of the Kalman–Yakubovic–Popov Lemma LMI in (4.8).

Remark 4.8. The authors (Heemels et al., 2000, 2002; Camlibel et al., 2002b) were originally motivated by the jump rule of Moreau's second-order sweeping process, which has been adapted to passive LCS. It is therefore not surprising that the above jump rule and the sweeping process restitution law with a restitution coefficient equal to zero (see (2.99), (2.100), or (2.101)) are quite similar. Such rules can therefore be used in an event-driven scheme, when an event is detected. They are well approximated by backward Euler methods, see Sect. 9.5.

Theorem 4.7 extends to nonautonomous LCS as in (4.7) provided the input $u(\cdot)$ is a piecewise Bohl function, i.e., it is right-continuous and there exists a partition of $\mathbb{R}^+$ into intervals $[t_k,t_{k+1})$ (i.e., $\mathbb{R}^+=\cup_{k\geqslant 0}[t_k,t_{k+1})$) such that $u(\cdot)$ is a Bohl function on each interval $[t_k,t_{k+1})$.

Remark 4.9. When $B=C^{\mathrm{T}}$ it is possible to use some results for evolution variational inequalities to study the well-posedness and the stability of LCSs (Goeleven & Brogliato, 2004). Indeed, in such a case the LCS can be recast into EVIs, and more precisely into linear EVIs:

$$\langle \dot{x}(t)+f(x(t)),y-x(t)\rangle \geqslant 0,\ \forall y\in K, \quad \text{with} \begin{cases} K=\{x\in\mathbb{R}^n, Cx\geqslant 0\} \\ f(x)=-Ax\ . \end{cases} \tag{4.12}$$

The existence and uniqueness of solutions of (4.12) in the relevant class of function is given in Theorem 2.44.

4.2.2 Examples of LCS

Several examples (physical or nonphysical) of LCS have been presented in Chap. 1 and Sect. 2.6. To complete this section, an example of nonexistence and nonuniqueness of solutions is provided for a LCS of relative degree 0. This example is taken from Heemels & Brogliato (2003).

Example 4.10. Let us consider the following LCS:

$$\begin{cases} \dot{x}(t)=-x(t)+\lambda(t) \\ w(t)=x(t)-\lambda(t) \\ 0\leqslant w(t)\perp\lambda(t)\geqslant 0 \\ x(0)=x_0\ . \end{cases} \tag{4.13}$$

The LCP that allows one to calculate $\lambda(t)$ in (4.13) is: $0 \leqslant \lambda(t) \perp x(t) - \lambda(t) \geqslant 0$ which has no solution at t for $x(t) < 0$. If $x(t) \geqslant 0$ then one can take either $\lambda(t) = 0$ or $\lambda(t) = x(t)$. This system is therefore equivalent to

$$\dot{x}(t) = \begin{cases} -x(t) & \text{if } x(t) \geqslant 0 \\ 0 & \text{if } x(t) \geqslant 0 \\ \emptyset & \text{if } x(t) < 0\,, \end{cases} \tag{4.14}$$

which leads to nonexistence of solutions for $x(0) < 0$ and to nonuniqueness for $x(0) > 0$.

As one may guess, it is not too difficult to invent other examples of LCS which do not possess solutions of any kind or which do not enjoy the uniqueness of solutions property. In general strong conditions have to be imposed so that CSs are well-posed.

4.2.3 Complementarity Systems and the Sweeping Process

A link exists between LCS and Moreau's perturbed sweeping process in (2.46), when the triple (A,B,C) is positive real and $D = 0$. Indeed in such a case a suitable variable change allows one to rewrite the LCS as an inclusion into a normal cone (Brogliato & Thibault, 2006). Positive realness of (A,B,C) implies that a linear matrix inequality $A^{\mathrm{T}}P + PA \leqslant 0, PB = C^{\mathrm{T}}$ is satisfied for some P, with $P = P^{\mathrm{T}} > 0$ when the pair (C,A) is observable and the pair (A,B) is controllable or stabilizable (see theorem 3.29 in Brogliato et al., 2007). The variable change is defined as $z = Rx$ with $RR = P$, i.e., R is a symmetric square root of P. This is the case of the circuit in Fig. 2.3 for which the above linear matrix inequality has the solution $P = \begin{pmatrix} \frac{1}{C_4} & 0 & 0 \\ 0 & L_3 & 0 \\ 0 & 0 & L_2 \end{pmatrix}$ Brogliato & Goeleven, (2005). The time-variation of the set $K(t)$ is then due to the input $u(t)$. Depending on the smoothness of $u(\cdot)$ the solutions may be absolutely continuous or of bounded variation. Let us show this on simple examples.

Example 4.11. Consider first the autonomous case, as treated in Brogliato (2004):

$$\begin{cases} \dot{x}(t) = Ax(t) + B\lambda(t) \\ 0 \leqslant Cx(t) \perp \lambda(t) \geqslant 0 \\ x(0) \geqslant 0\,, \end{cases} \tag{4.15}$$

where $x(t) \in \mathrm{IR}^n$, $C \in \mathrm{IR}^{m \times n}$, (A,B,C) is positive real, B has full column rank, and (A,B,C) is a minimal realization. Using the equivalence (A.9) in the appendix, the complementarity relation is rewritten as

$$-\lambda(t) \in \partial\psi_K(Cx(t)) \tag{4.16}$$

with $K = (\mathbb{R}^+)^m$. Thus we obtain that (4.15) is equivalent to

$$\dot{x}(t) \in Ax(t) - B\partial\psi_K(Cx(t)) \,. \tag{4.17}$$

Using that $PB = C^{\mathrm{T}}$ and $RR = P$ we get

$$R\dot{x}(t) \in RAR^{-1}Rx(t) - R^{-1}C^{\mathrm{T}}\partial\psi_K(CR^{-1}Rx(t)) \tag{4.18}$$

and with $z = Rx$

$$\dot{z}(t) \in RAR^{-1}z(t) - R^{-1}C^{\mathrm{T}}\partial\psi_K(CR^{-1}z(t)) \,. \tag{4.19}$$

Let us define $f(\cdot) = CR^{-1} \cdot \circ\psi_K(\cdot)$, so that $\partial f(\cdot) = (CR^{-1})^{\mathrm{T}}\partial\psi_K(\cdot)$, from the chain rule for compositions of convex lower semi-continuous functions and linear continuous mappings (see Proposition A.3). Therefore (4.15) is finally transformed into the DI

$$\dot{z}(t) \in RAR^{-1}z(t) - \partial f(z(t)), \;\; z(0) = Rx(0) \,. \tag{4.20}$$

The function $f(\cdot)$ is convex and lower semi-continuous, since K is convex. Its subderivative may also be written as $N_{\bar{K}}(z)$, where $\bar{K} = \{z \in \mathbb{R}^n \mid CR^{-1}z \in K\}$. Thus we finally get

$$\dot{z}(t) \in RAR^{-1}z(t) - N_{\bar{K}}(z(t)), \;\; z(0) = Rx(0) \,. \tag{4.21}$$

The inclusion in (4.21) can easily be cast into the perturbed sweeping process in (2.46). Actually as shown in Brogliato (2004) its well-posedness can also be analyzed with Theorem 2.41, because it is an autonomous system.

In Brogliato & Thibault (2006) the nonautonomous case is analyzed, i.e., systems of the form

$$\begin{cases} \dot{x}(t) = Ax(t) + B\lambda(t) + Eu(t) \\[2ex] 0 \leqslant Cx(t) + Fu(t) \perp \lambda(t) \geqslant 0 \\[2ex] x(0) \geqslant 0 \,, \end{cases} \tag{4.22}$$

where $u(\cdot)$ is supposed to be either absolutely continuous or BV. The presence of the function of time $u(\cdot)$ in the complementarity relations implies that the set K in (4.21) becomes a time-varying set $K(t)$. Thus nonautonomous LCSs with positive real (A,B,C) naturally lend themselves to an interpretation through Moreau's sweeping process. In case $u(\cdot)$ is BV, then the state $x(\cdot)$ may jump and is itself BV. Thus the inclusion is a measure differential inclusion as (2.46). In such a case, one may say that the right formulation of these systems is the measure DI (2.46), and that in case $u(\cdot)$ is absolutely continuous and $K(t)$ can be described as in (4.22), the measure DI becomes a DI which in turn is equivalent to a LCS.

Remark 4.12. From (4.19) at an atom of dz (when $z(\cdot)$ is BV) one obtains

$$z(t^+) - z(t^-) \in -R^{-1}C^{\mathrm{T}}\,\partial\psi_K(CR^{-1}z(t^+)) \;\Leftrightarrow\; z(t^+) = \mathrm{prox}[\bar{K}; z(t^-)] \tag{4.23}$$

with $\bar{K}$ defined above. In the autonomous case (4.19) dz may have an atom at $t = 0$ only. In the nonautonomous case atoms may exist for any $t \geqslant 0$, depending on the exogenous signal regularity.

4.2.4 Nonlinear Complementarity Systems

The case of nonlinear CS is more tricky, as expected. Works in this area may be found in van der Schaft & Schumacher (1998a) and Brogliato & Thibault (2006). Let us briefly describe the results of Brogliato & Thibault (2006), which are based on previous well-posedness results for the perturbed sweeping process obtained in Edmond & Thibault (2005, 2006) and on the variable change described in Sect. 4.2.3. The dissipativity properties of an uncontrolled, unconstrained system extracted from the controlled nonsmooth system are central in the well-posedness proof.

Let us focus on the following class of complementarity systems:

$$\begin{cases} \dot{x}(t) = f(x(t)) + B\lambda(t) + e(x(t), u(t)) \\ 0 \leqslant \lambda(t) \perp w(t) = c(x(t)) + g(u(t)) \geqslant 0 \,, \end{cases} \tag{4.24}$$

where $x(t) \in \mathrm{I\!R}^n$, $u(t) \in \mathrm{I\!R}^p$, $w(t) \in \mathrm{I\!R}^m$.

Assumption 2. *The mappings $f(\cdot)$ and $e(\cdot)$ are continuous, $g(\cdot)$ is supposed to be locally Lipschitz continuous, $a(0) = 0$, $g(0) = 0$, $e(\cdot, 0) = 0$.*

Assumption 3. *The uncontrolled system*

$$\begin{cases} \dot{x}(t) = a(x(t)) + B\lambda(t) \\ w(t) = c(x(t)) \end{cases} \tag{4.25}$$

is dissipative with respect to the supply rate $w = w^{\mathrm{T}}\lambda$, and there exists a positive function $V(\cdot)$ such that $V(0) = 0$ and

$$c^{\mathrm{T}}(x) = \frac{\partial V}{\partial x}^{\mathrm{T}}(x)B \,. \tag{4.26}$$

Assumption 4. *The function $V(\cdot)$ is of class $C^3(\mathrm{I\!R}^n; \mathrm{I\!R}^+)$ and the Hessian $\frac{\partial^2 V}{\partial x^2}(x)$ is positive definite and symmetric for all $x \in \mathrm{I\!R}^n$. Moreover $\left(\frac{\partial^2 V}{\partial x^2}(x)\right)^{\frac{1}{2}}$ is integrable as a function of x, and*

$$\frac{\partial h}{\partial x}(x) = \left(\frac{\partial^2 V}{\partial x^2}(x)\right)^{\frac{1}{2}} \overset{\Delta}{=} \Lambda(x).$$

for some diffeomorphism $h : \mathrm{I\!R}^n \to \mathrm{I\!R}^n$. Finally there exists a constant $\rho > 0$ such that for all $x \in \mathrm{I\!R}^n$

$$\rho B_m \subset B^T \Lambda(x)(B_n) + (\mathrm{I\!R}^+)^m \,. \tag{4.27}$$

The next lemma and the next theorem are proved in Brogliato & Thibault (2006). It is noteworthy that in the right-hand side of (4.28) below the set $S(t)$ may not be convex. Consequently the normal cone is not the normal cone of convex analysis, but a generalization of it, known as the limiting normal cone of $S(t)$ (Rockafellar & Wets, 1998). This generalization is needed in the framework of nonlinear systems, for which it is too restrictive to assume convexity of all the ingredients.

Lemma 4.13. *Suppose that Assumptions 2, 3, and 4 hold. Then the state space transformation $z = h(x)$ allows one to transform the NLCS in (4.24) into the perturbed sweeping process*

$$-\dot{z}(t) + \frac{\partial h}{\partial x}^{\mathrm{T}}(x) f(h^{-1}(z(t))) + \frac{\partial h}{\partial x}^{\mathrm{T}}(x) e(h^{-1}(z(t)), u(t)) \in \partial \psi_{S(t)}(z(t)) \,, \quad (4.28)$$

where $S(t) \overset{\Delta}{=} \{z \mid \; c(h^{-1}(z)) + g(u(t)) \geqslant 0\}$.

The proof of Lemma 4.13 roughly follows the same steps as that done in Sect. 4.2.3. The existence and uniqueness of solutions comes now.

Theorem 4.14. *Consider the system in (4.24) and suppose that assumptions 2, 3, and 4 hold. Let $u(\cdot)$ be locally absolutely continuous, and $z_0 \in S(0)$. Then there exists some $T > 0$ such that the perturbed differential inclusion (4.28) with z_0 as initial condition has at least one locally absolutely continuous solution on $[0, T)$ and the solution is unique whenever $\frac{\partial^2 V}{\partial x^2}(\cdot)$ is bounded on the convex hull $\mathrm{co}(\mathrm{Rge}\, S)$ of $\mathrm{Rge}(S)$.*

If, in addition, $f(\cdot)$ and $e(\cdot, u)$ are locally Lipschitz continuous and the mapping $(t, z) \mapsto \frac{\partial h}{\partial x}^{\mathrm{T}}(x) f(h^{-1}(z(t))) + \frac{\partial h}{\partial x}^{\mathrm{T}}(x) e(h^{-1}(z(t)), u(t))$ in (4.28) satisfies a linear growth condition, then T may be taken equal to $+\infty$.

Theorem 4.14 can be extended to $u(\cdot)$ RCBV, so that the inclusion in (4.28) becomes a measure differential inclusion. However, in this case uniqueness of solutions is not proved. The interest of such a result, in, this book, is mainly to show that nonlinear complementarity systems may be well-posed, at the price of imposing stringent conditions.

Remark 4.15. As long as the well-posedness is concerned, the positiveness of $V(\cdot)$ is not necessary. The essential tool is relation (4.26).

4.3 Relative Degree and the Completeness of the Formulation

A fundamental notion to study such systems is the notion of *relative degree*. The name is taken from the usual terminology in systems and control and can be defined independently of the complementarity relation. Let us consider a linear system in state representation given by the quadruple (A, B, C, D):

$$\begin{cases} \dot{x}(t) = Ax(t) + B\lambda(t) \\ \\ w(t) = Cx(t) + D\lambda(t) \end{cases} \tag{4.29}$$

with $x(t) \in \mathbb{R}^n$, $w(t) \in \mathbb{R}^m$, and $\lambda(t) \in \mathbb{R}^m$, and $x(0) = x_0$.

4.3.1 The Single Input/Single Output (SISO) Case

In the SISO case ($m = 1$), the relative degree is defined by the first nonzero Markov parameter. The sequence of Markov parameters is defined by the sequence of scalars

$$D, CB, CAB, CA^2B, \ldots, CA^{r-1}B, \ldots \; . \tag{4.30}$$

In fact, these parameters arise naturally when we derive the output $w(\cdot)$ with respect to time:

$$\begin{aligned} w(t) &= Cx(t) + D\lambda(t) \\ \dot{w}(t) &= CAx(t) + CB\lambda(t), \text{ if } D = 0 \\ \ddot{w}(t) &= CA^2x(t) + CAB\lambda(t), \text{ if } D = 0 \text{ and } CB = 0 \\ &\vdots \\ w^{(r)}(t) &= CA^r x(t) + CA^{r-1}B\lambda(t), \text{ if } D = 0 \\ &\qquad \text{and } CA^jB = 0, \text{ for all } j = 0, \ldots, r-2. \\ &\vdots \end{aligned} \tag{4.31}$$

The first nonzero Markov parameter allows us to define the "output" $w(\cdot)$ as an explicit function of the "input" $\lambda(\cdot)$. The existence of a finite relative degree is guaranteed by the existence of a nonzero transfer function or a nonzero input/output operator $u(\cdot) \mapsto w(\cdot)$. The relation with the transfer function is as follows. The transfer function $H : \mathbb{C} \to \mathbb{C}$ of the system (4.29) is given by

$$H(s) = D + C(sI_n - A)^{-1}B \; . \tag{4.32}$$

We may write

$$H(s) = \frac{N(s)}{D(s)} \; ,$$

where $D(s)$ is a polynomial of degree n and $N(s)$ is a polynomial of degree $l \leqslant n$. The relative degree r of the system (A, B, C, D) is defined equivalently as the difference between the degrees of the denominator and numerator polynomials of $H(s)$, i.e., $r = n - l$. Note that $0 \leqslant r \leqslant n$.

4.3.2 The Multiple Input/Multiple Output (MIMO) Case

When $m > 1$, a notion of uniform vector relative degree r can be defined as follows (Sannuti, 1983). The parameters in (4.30) are $m \times m$ matrices. If D is nonsingular, the relative degree is equal to 0. Otherwise, it is assumed to be the first positive integer r such that

$$CA^iB = 0, \quad i = 0, \ldots, q-2 \tag{4.33}$$

while

$$CA^{r-1}B \text{ is nonsingular} . \tag{4.34}$$

We may then denote $\bar{r} = (r, r, ..., r)^{\mathrm{T}} \in \mathrm{I\!R}^m$, the vector relative degree, or simply r. It is also possible to define nonuniform relative degree vector, for some classes of nonlinear systems, see for instance Isidori (1995). In systems and control, the relative degree is a very useful notion to derive various sorts of canonical state space realizations, for control analysis and design.

Example 4.16. Consider a gradient CS as in (4.6), but with linear dynamics:

$$\begin{cases} \dot{x}(t) + Ax(t) = G^{\mathrm{T}}\lambda(t) \\ \\ w(t) = Gx(t) \\ \\ 0 \leqslant w(t) \perp \lambda(t) \geqslant 0 \,. \end{cases} \tag{4.35}$$

Then $\dot{w}(t) = -GAx(t) + GG^{\mathrm{T}}\lambda(t)$. Provided the $m \times m$ matrix GG^{T} is square full rank (equivalently G has rank m), the vector relative degree is equal to $\bar{r} = (1, ..., 1)^{\mathrm{T}} \in \mathrm{I\!R}^m$. One says that the system has a uniform relative degree $r = 1$.

4.3.3 The Solutions and the Relative Degree

A general discussion is done in Acary et al. (in press), whose work is summarized in Chap. 5. We will consider here a very simple example to illustrate our purpose.

Example 4.17. Let us consider the following LCS:

$$\begin{cases} \dddot{x}(t) = \lambda, \; x(0) = x_0 \geqslant 0 \\ \\ w(t) = x(t) \\ \\ 0 \leqslant w(t) \perp \lambda \geqslant 0 \,. \end{cases} \tag{4.36}$$

Obviously it has $r = 3$. When the constraint $w = x \geqslant 0$ becomes active at t, i.e., $x(t) = 0$, the sign of the derivatives of $w(t)$ (i.e., of $x(t)$) will govern the future behavior of the system. If $\dot{x}(t^-) > 0$, one can keep this value and work with a continuous velocity. The system will instantaneously leave the constraints. If $\dot{x}(t^-) < 0$, the velocity needs to jump so that the trajectory respects the unilateral constraint in a right neighborhood of t. Therefore it is necessary in such a case that the velocity be nonsmooth and jumps to some positive value $\dot{x}(t^+) > 0$. The velocity has to be considered as a function of bounded variations and it remains to define how it jumps. The second derivative $\ddot{x}(\cdot)$ is a measure with an atom at t, and the third derivative is a derivative (in the sense of distributions) of a measure (something like $\dot{\delta}_t$). Consequently, the term λ is expected to be also a derivative of measure, i.e., a distribution of degree 3.

Consequently, a constraint of the type $\lambda \geqslant 0$ has no mathematical meaning for (4.36). The LCS formalism is not adapted when $r \geqslant 3$ and has to be replaced by a more general formalism that makes sense with distributional solutions. We will come back on the problem of higher relative degree in Chap. 5.

5

Higher Order Constrained Dynamical Systems

The material presented in this chapter is taken from Acary et al. (in press) and Acary & Brogliato (2006, 2005). It follows from the arguments in Sect. 4.3 that the framework of linear complementarity systems is well suited for systems with a relative degree between the complementarity variables $\lambda(\cdot)$ and $w(\cdot)$ less or equal to one. When the relative degree $r \geqslant 2$, such a formalism is no longer appropriate. In other words, suppose one is looking for a sound mathematical formalism, such that a unilaterally constrained system of the general form

$$\begin{cases} \dot{x}(t) = Ax(t) + B\lambda \\ \\ w(t) = Cx(t) \geqslant 0, \;\; x(0) = x_0 \end{cases} \tag{5.1}$$

with $A \in \mathrm{IR}^{n\times n}$, $B \in \mathrm{IR}^{n\times m}$, $C \in \mathrm{IR}^{m\times n}$, possesses unique solutions on IR^+, for all $x_0 \in \mathrm{IR}^n$. From Sect. 4.3.3, the solutions of such a system are likely to be distributions, so that the mere writing of the differential equation in (5.1) may not be suitable. The following sections aim at briefly introducing the so-called *higher order sweeping process* (HOSP), which is a differential inclusion whose solutions are distributions. The HOSP is an extension of the second-order sweeping process that is a measure DI. The HOSP may be named a *distribution DI*.

5.1 Motivations

Roughly speaking, the primary interest for studying such a DI is to provide one with a time-stepping method allowing to integrate a system like (4.36) with $r \geqslant 3$. Indeed the backward Euler method presented in Sect. 9.5 does not work in such a case. This is due to the fact that an algorithm as (9.72) is able to approximate measures (at the price of slightly modifying it), but it cannot approximate derivatives of Dirac measures. Examples are treated in Sect. 11.1 which demonstrate this.

Apart from the pleasure of being able to integrate systems with $r \geqslant 3$, there exist other motivations. The necessary first-order optimality conditions that stem from

Pontryagin's principle with inequality state constraints have the form of a LCS with $r \geqslant 2$ (van der Schaft & Schumacher, 2000). The viability problem may also involve systems with distributions (Kinzebulatov, 2007). The viability problem may be thought of as finding distributional controls such that the domain $\{x \in \mathbb{R}^n \mid Cx \geqslant 0\}$ remains invariant. The problem that is solved by the HOSP is to find the right relationships between the multiplier and the state so that the same goal is attained.

5.2 A Canonical State Space Representation

Let us introduce a state space representation for (5.1) which will later on allow us to design the HOSP. This representation is based on the relative degree r between $w(\cdot)$ and $\lambda(\cdot)$, considered for the time being as smooth functions. Let us first deal with the SISO case of Sect. 4.3.1, i.e., $m = 1$. We therefore suppose that the transfer function $H(s) = C(sI_n - A)^{-1}B \neq 0$, so that $1 \leqslant r \leqslant n$. Then there exists a full-rank matrix $W \in \mathbb{R}^{n\times n}$ such that (see, e.g., Sannuti (1983)):

$$WB = \begin{pmatrix} 0^{r-1} \\ CA^{r-1}B \\ 0^{n-r} \end{pmatrix}$$

$$CW^{-1} = \begin{pmatrix} 1 & 0_{n-1} \end{pmatrix}$$

and

$$WAW^{-1} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0_{n-r} \\ 0 & 0 & 1 & \ddots & \vdots & \vdots \\ \vdots & & \ddots & \ddots & 0 & \vdots \\ 0 & \dots & \dots & 0 & 1 & 0_{n-r} \\ d_1 & d_2 & d_3 & \dots & d_r & d_\xi^{\mathrm{T}} \\ B_\xi & 0^{n-r} & 0^{n-r} & \dots & 0^{n-r} & A_\xi \end{pmatrix}, \tag{5.2}$$

where $A_\xi \in \mathbb{R}^{(n-r)\times(n-r)}, B_\xi \in \mathbb{R}^{(n-r)\times 1}$, and $(d^{\mathrm{T}}, d_\xi^{\mathrm{T}}) = (CA^rW^{-1})^{\mathrm{T}}$ with $d^{\mathrm{T}} = (d_1, \dots, d_r)$. The notation $0^n = (0,0,\dots,0) \in \mathbb{R}^{1\times n}$, and $0_n = (0^n)^{\mathrm{T}} \in \mathbb{R}^n$.

Actually, the framework that is presented next is essentially linked to systems with $r \geqslant 1$. The existence of a relative degree allows one to perform a state space transformation with new state vector $z = Wx$,

$$z^{\mathrm{T}} = (z_1, z_2, \dots, z_r, \xi^{\mathrm{T}}) = (\bar{z}^{\mathrm{T}}, \xi^{\mathrm{T}}),\ \xi \in \mathbb{R}^{n-r} \tag{5.3}$$

such that the new state space representation is (see Sannuti (1983))

$$\begin{cases} \dot{z}(t) = WAW^{-1}z(t) + WB\lambda(t) \\ z(0) = Wx_0 \\ w(t) = CW^{-1}z(t) \geqslant 0 \end{cases}, \tag{5.4}$$

that is,

$$\begin{cases} \dot{z}_1(t) = z_2(t) \\ \dot{z}_2(t) = z_3(t) \\ \dot{z}_3(t) = z_4(t) \\ \\ \vdots \\ \\ \dot{z}_{r-1}(t) = z_r(t) \\ \dot{z}_r(t) = CA^r W^{-1} z(t) + CA^{r-1} B \lambda(t) \\ \\ \dot{\xi}(t) = A_\xi \xi(t) + B_\xi z_1(t) \\ \\ w(t) = z_1(t) \geqslant 0 \\ \\ z(0) = z_0 \end{cases} . \tag{5.5}$$

Moreover

$$CA^r W^{-1} z = d^{\mathrm{T}} \bar{z} + d_\xi^{\mathrm{T}} \xi . \tag{5.6}$$

In systems and control theory, the dynamics $\dot{\xi} = A_\xi \xi + B_\xi z_1$ is called the *zero dynamics*, so we shall denote the state space form in (5.5) the ZD representation.

Example 5.1. Let us consider the state representation in (5.5) with

$$A = \begin{pmatrix} 2 & 7 & 3-2\alpha & -2\beta+2 \\ -1 & -3 & -1+\alpha & \beta-1 \\ 0 & 0 & 0 & 1 \\ 1 & 2 & 1 & 0 \end{pmatrix}, \; B = \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \; C = \begin{pmatrix} 1 \; 2 \; 0 \; 0 \end{pmatrix}$$

where $\alpha, \beta \in \mathbb{R}$. The transfer function of this system is given by

$$H(s) = \frac{s^2 - 1}{s^4 + s^3 - (1+\alpha)s - 1 - \beta} .$$

Then the transformation matrix

$$W = \begin{pmatrix} 1 \; 2 \; 0 \; 0 \\ 0 \; 1 \; 1 \; 0 \\ 0 \; 0 \; 1 \; 0 \\ 0 \; 0 \; 0 \; 1 \end{pmatrix}$$

allows one to transform the system (A, B, C) into the ZD canonical state space representation (WAW^{-1}, WB, CW^{-1}) where

$$WAW^{-1} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & -1 & \alpha & \beta \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}, \; WB = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \; CW^{-1} = \begin{pmatrix} 1 \; 0 \; 0 \; 0 \end{pmatrix}$$

so that the dynamics in (5.5) is given by

$$\begin{cases} \dot{z}_1(t) = z_2(t) \\ \dot{z}_2(t) = \lambda(t) - z_1(t) - z_2(t) + \alpha\xi_1(t) + \beta\xi_2(t) \\ \dot{\xi}_1(t) = \xi_2(t) \\ \dot{\xi}_2(t) = \xi_1(t) + z_1(t) \\ w(t) = z_1(t) \end{cases} . \tag{5.7}$$

5.3 The Space of Solutions

The reader is referred to B Section C.3, for details on how the solutions are constructed and the employed notation. Roughly speaking, the space of solutions for the HOSP generalizes that of complementarity Lagrangian systems with absolutely continuous positions and BV velocities, so that the acceleration is a differential measure. However, due to the relative degree $\geqslant 1$, the number of state variables that may have jumps is not restricted to the velocity. The higher derivatives may also be discontinuous, up to the order r (see the canonical form in (5.5)), while the zero dynamics vector $\xi(\cdot)$ is a continuous function of time in the HOSP framework. This is reflected in (C.6).

5.4 The Distribution DI and Its Properties

The HOSP formalism has some fundamental properties. In particular it is of great interest to understand how the inequality $\lambda \geqslant 0$, which is meaningless when λ is a distribution of degree > 2 (the derivatives of the Dirac measure), may be formulated in a rigorous way.

5.4.1 Introduction

As the example of Sect. 4.3.3 shows, in general the possible solutions of (5.5) cannot be defined in a class of smooth functions. Consider for instance the initial data $z_{0,i} \leqslant -\delta$ for some $\delta > 0$ and all $1 \leqslant i \leqslant r$. Then, since the unilateral constraint $z_1 \geqslant 0$ must be satisfied, it is necessary that $z_1(0^+) \geqslant 0$, i.e., z_1 needs to "jump" to some nonnegative value. It results that z_1 cannot be continuous and the derivatives in (5.5) must be considered in the sense of distributions. At this stage we can just say that a jump mapping is needed. Its form will depend on the type of system one handles (in mechanics, this is the realm of impact mechanics (Brogliato, 1999)). If one considers (5.5) as an equality of distributions of class $\mathscr{T}_\infty(I)$ (see Definition C.2), then we can rewrite it as

$$\begin{cases} Dz_1 = z_2 \\ Dz_2 = z_3 \\ Dz_3 = z_4 \\ \\ \vdots \\ \\ Dz_{r-1} = z_r \\ \\ Dz_r = CA^r W^{-1} z + CA^{r-1} B\lambda \\ \\ D\xi = A_\xi \xi + B_\xi z_1 . \end{cases} , \tag{5.8}$$

where $D.$ stands for the derivation in the sense of distributions. Consider the above initial conditions on $\{z_i\}$ $(1 \leqslant i \leqslant r)$. Then Dz_1 is a distribution of degree 2 and we get $Dz_1 = \{\dot{z}_1\} + \sigma_{z_1}(0)\delta_0 = z_2$. Consequently Dz_2 is a distribution of degree 3 and $Dz_2 = D^2 z_1 = D\{\dot{z}_1\} + \sigma_{z_1}(0) D\delta_0 = \{\dot{z}_2\} + \sigma_{\{\dot{z}_1\}}(0)\delta_0 + \sigma_{z_1}(0) D\delta_0 = z_3$, and $\{\dot{z}_1\} = \{z_2\}$. Then Dz_3 is a distribution of degree 4, and we get $Dz_3 = D\{\dot{z}_2\} + \sigma_{\{dotz_1\}}(0) D\delta_0 + \sigma_{z_1}(0) D^2\delta_0 = \{\dot{z}_3\} + \sigma_{\{\dot{z}_2\}}(0)\delta_0 + \sigma_{\{\dot{z}_1\}}(0) D\delta_0 + \sigma_{z_1}(0) D^2 \delta_0 = z_4$, and $\{\dot{z}_2\} = \{z_3\}$, and so on. Thus $\sigma_{\{\dot{z}_1\}}(0) = \{z_2\}(0^+) - \{z_2\}(0^-)$, $\sigma_{\{\dot{z}_2\}}(0) = \{z_3\}(0^+) - \{z_3\}(0^-)$, and so on. Until now we have decomposed only the left-hand side of the dynamics as distributions of some degrees. Now let us get back to the distributional dynamics in (5.8). Starting from $Dz_1 = z_2$, one deduces that the right-hand side has to be of the same degree has the left-hand side. This means that the right-hand side is equal to $\{z_2\} + \nu_1$, where ν_1 is a distribution of degree 2, i.e., a measure. Similarly from $Dz_2 = z_3$ one deduces that $z_3 = \{z_3\} + \tilde{\nu}_2$, where $\tilde{\nu}_2$ has degree 3 and can therefore further be decomposed as $\nu_2 + \tilde{\nu}_1$, with $\deg(\nu_2) = 2$ and $\deg(\tilde{\nu}_1) = 3$. It is not difficult to see that $\tilde{\nu}_1 = D\nu_1$. Therefore $Dz_2 = \{z_3\} + \nu_2 + D\nu_1$. The variables ν_1 and ν_2 are slack variables (or Lagrange multipliers), and are measures of the form $\nu_i = \int_I \mathrm{d}\nu_i$, where $\mathrm{d}\nu_i$ is a Stieltjes measure generated by a $\mathscr{F}_\infty(I;\mathbb{R})$-function. Continuing the reasoning until Dz_r, we obtain $Dz_r = CA^r W^{-1}\{z\} + CA^{r-1}B\lambda$ where $\deg(\lambda)$ =$\deg(Dz_r) = r+1$. Consequently from (5.8) one gets

$$\begin{cases} Dz_1 = \{z_2\} + \nu_1 \\ Dz_2 = \{z_3\} + D\nu_1 + \nu_2 \\ Dz_3 = \{z_4\} + D^2\nu_1 + D\nu_2 + \nu_3 \\ \\ \vdots \\ Dz_i = \{z_{i+1}\} + D^{(i-1)}\nu_1 + D^{(i-2)}\nu_2 + \ldots + D\nu_{i-1} + \nu_i \\ \vdots \\ Dz_{r-1} = \{z_r\} + D^{(r-2)}\nu_1 + \ldots + D\nu_{r-2} + \nu_{r-1} \\ \\ Dz_r = CA^r W^{-1}\{z\} + CA^{r-1}B\lambda . \end{cases} \tag{5.9}$$

We keep the notation λ for the multiplier which appears in the last line. One sees that λ in (5.9) can be given a meaning as

$$\lambda = (CA^{r-1}B)^{-1}[D^{(r-1)}\nu_1 + \ldots + D\nu_{r-1}] + \nu_r \tag{5.10}$$

provided $CA^{r-1}B \neq 0$ (invertible in the multivariable case $m \geqslant 2$ with relative degree $\bar{r} = (r, r, ..., r)^{\mathrm{T}} \in \mathrm{IR}^m$). Then λ is uniquely defined as in (5.10). It is important at this stage to realize that λ is the unique source of higher degree distributions in the system which will allow the state to jump. Therefore the measures ν_i have themselves to be considered as sub-multipliers. The expression in (5.10) is important, as it will enable us to generalize the positivity of λ when λ is a measure, to the case when it is not a measure.

Remark 5.2. In view of the nature of the solutions as explained in Sects. 5.3 and C.3, we may write $\mathrm{d}\nu_i$ as

$$\mathrm{d}\nu_i = \chi_i(t)\mathrm{d}t + \mathrm{d}\mathscr{J}_i, \tag{5.11}$$

where $\chi_i \in \mathscr{F}_\infty(I; \mathrm{IR})$ and $\mathrm{d}\mathscr{J}_i$ is an atomic measure with countable set of atoms generated by a right-continuous jump function $\mathscr{J}_i$. Let $1 \leqslant i \leqslant r-1$ be given. We know that $Dz_i = z_{i+1}$ and thus $\{Dz_i\} = \{z_{i+1}\}$. Thus $\nu_i = \ll Dz_i - Dz_i \gg$. It results that $\mathrm{d}\nu_i$ is an atomic measure and thus

$$\chi_i(t) = 0, \text{ a.e. } t \in I, \ (1 \leqslant i \leqslant r-1). \tag{5.12}$$

This means that except for ν_r, the other measures ν_i are purely atomic (they act only at the state jump instants). The nonatomic part of ν_r allows the state to move on the constraint boundary $\{z_1 = 0\}$. This is the equivalent of the contact force multiplier in complementarity Lagrangian systems.

Remark 5.3. The fundamental difference between (2.125) and (5.9) is pointed out in remark 1.9 in Brogliato (1999).

5.4.2 The Inclusions for the Measures ν_i

Let K be a nonempty closed convex subset of IR. We denote by $T_K(x)$ the tangent cone of K at $x \in \mathrm{IR}$ defined by

$$T_K(x) = \overline{\mathrm{cone}}(K - \{x\}), \tag{5.13}$$

where $\mathrm{cone}(K - \{x\})$ denotes the cone generated by $K - \{x\}$ and $\overline{\mathrm{cone}}(K - \{x\})$ denotes the closure of $\mathrm{cone}(K - \{x\})$, i.e., $\overline{\mathrm{cone}}(K - \{x\}) = \overline{\mathrm{cone}(\mathrm{K} - \{\mathrm{x}\})}$. The definition in (5.13) allows us to take into account constraints violations. Note that

$$T_{\mathrm{IR}^+}(x) = \begin{cases} \mathrm{IR} & \text{if } x > 0 \\ \mathrm{IR}^+ & \text{if } x \leqslant 0 \end{cases}$$

and

$$T_{\mathrm{IR}}(x) = \mathrm{IR}$$

Let us now set

$$\Phi \stackrel{\Delta}{=} \mathrm{IR}^+. \tag{5.14}$$

For $z \in \mathrm{IR}^r$, we set

$$Z_i = (z_1, z_2, ..., z_i), \quad (1 \leqslant i \leqslant r). \tag{5.15}$$

By convention, we set $Z_0 = 0$ and

$$T^0_\Phi(Z_0) = \Phi$$

and we define

$$\begin{cases} T^1_\Phi(Z_1) = T_\Phi(z_1), \\ T^2_\Phi(Z_2) = T_{T^1_\Phi(Z_1)}(z_2), \\ \quad\vdots \\ T^r_\Phi(Z_r) = T_{T^1_\Phi(Z_{r-1})}(z_r), \end{cases}$$

that is,

$$T^i_\Phi(Z_i) = T_{T^{i-1}_\Phi(Z_{i-1})}(z_i),\ 1 \leqslant i \leqslant r.$$

Remark 5.4. In the multivariable case $m \geqslant 2$ with vector relative degree $\bar{r}$, we have $\Phi = (\mathbb{R}^+)^m$, $Z^l_0 = 0, Z^l_i = (z^l_1, z^l_2, ..., z^l_i)$, $1 \leqslant i \leqslant r, 1 \leqslant l \leqslant m$, and

$$T^i_\Phi(Z_i) = \times^m_{l=1} T^i_\Phi(Z^l_i),\ 1 \leqslant i \leqslant r.$$

Starting from (5.8), (5.9) the HOSP is written as follows:

$$\mathrm{d}\nu_i \in -\partial\psi_{T^{i-1}_\Phi(\{Z_{i-1}\}(t^-))}(\{z_i\}(t^+)) \quad (1 \leqslant i \leqslant r) \tag{5.16}$$

with ν_i in (5.9). Here $\{z_i\}(0^-)$ $(1 \leqslant i \leqslant r)$ will be given (by convention) so as to define some initial conditions for the process. The sets

$$\partial\psi_{T^{i-1}_\Phi(\{Z_{i-1}\}(t^-))}(\{z_i\}(t^+))\ (1 \leqslant i \leqslant r)$$

are nonempty closed convex cones. The positivity of λ is now understood as the positivity of ν_r as

$$\mathrm{d}\nu_r \in -\,\partial\psi_{T^{r-1}_\Phi(\{Z_{r-1}\}(t^-))}(\{z_r\}(t^+)).$$

5.4.3 Two Formalisms for the HOSP

We are now going to introduce two ways to formalize the HOSP. The first one, called the *distributional formalism*, has solutions in the space $\mathscr{T}_\infty(I)$ of distributions (see Definitions C.1 and C.2). The second one, called the *measure differential formalism*, has solutions in the space of functions $\mathscr{F}_\infty(I;\mathbb{R})$ (see (C.6)). Both are linked in a one-to-one way. The measure differential formalism is important because the time-stepping scheme presented in Chap. 11 is built to approximate its solutions (and not the solutions of the distributional formalism). It happens that the numerical way to approximate distributions of degree $\geqslant 3$ has not yet been discovered. Hence the passage through the measure differential formalism is mandatory for numerical purposes.

Let $T > 0$, $T \in \mathbb{R} \cup \{+\infty\}$ be given and set $I = [0, T)$. Let

$$z_0^T = (\bar{z}_0^T, \xi_0^T)$$

be given in $\mathbb{R}^n$ with $\bar{z}_0 \in \mathbb{R}^r$ and $\xi_0 \in \mathbb{R}^{n-r}$.

Distributional Formalism: Problem SP$(z_0; I)$ Find $z_1, \ldots, z_r \in \mathscr{T}_\infty(I)$ and $\xi_i \in \mathscr{T}_\infty(I)$ $(1 \leqslant i \leqslant n-r)$ satisfying the distributional equations:

$$\begin{cases} Dz_1 - z_2 = 0 \\ Dz_2 - z_3 = 0 \\ Dz_3 - z_4 = 0 \\ \\ \vdots \\ \\ Dz_{r-1} - z_r = 0 \\ Dz_r - CA^r W^{-1}\{z\} = CA^{r-1}B\lambda \\ \\ D\xi = A_\xi \xi + B_\xi z_1 \end{cases}, \tag{5.17}$$

$$\lambda = (CA^{r-1}B)^{-1}\left[\sum_{i=1}^{r-1} D^{(r-i)} \ll Dz_i - \{z_{i+1}\} \gg \right] + \ll Dz_r - CA^r W^{-1}\{z\} \gg \tag{5.18}$$

the measure differential inclusions on $(0,T)$:

$$\begin{cases} d\{z_1\} - \{z_2\}(t)\mathrm{d}t \ \in -\partial\psi_\Phi(\{z_1\}(t^+)) \\ \\ d\{z_2\} - \{z_3\}(t)\mathrm{d}t \in -\partial\psi_{T^1_\Phi(\{Z_1\}(t^-))}(\{z_2\}(t^+)) \\ \vdots \\ d\{z_i\} - \{z_{i+1}\}(t)\mathrm{d}t \in -\partial\psi_{T^{i-1}_\Phi(\{Z_{i-1}\}(t^-))}(\{z_i\}(t^+)) \\ \vdots \\ d\{z_{r-1}\} - \{z_r\}(t)\mathrm{d}t \in -\partial\psi_{T^{r-2}_\Phi(\{Z_{r-2}\}(t^-))}(\{z_{r-1}\}(t^+)) \\ \\ (CA^{r-1}B)^{-1}[d\{z_r\} - CA^r W^{-1}\{z\}(t)\mathrm{d}t] \in -\partial\psi_{T^{r-1}_\Phi(\{Z_{r-1}\}(t^-))}(\{z_r\}(t^+)) \end{cases} \tag{5.19}$$

and the initial conditions:

$$\begin{cases} \{z_1\}(0^+) - z_{0,1} \ \in -\partial\psi_\Phi(\{z_1\}(0^+)) \\ \\ \{z_2\}(0^+) - z_{0,2} \in -\partial\psi_{T^1_\Phi(Z_{0,1})}(\{z_2\}(0^+) \\ \vdots \\ \{z_i\}(0^+) - z_{0,i} \in -\partial\psi_{T^{i-1}_\Phi(Z_{0,i-1})}(\{z_i\}(0^+)) \\ \vdots \\ \{z_{r-1}\}(0^+) - z_{0,r-1} \in -\partial\psi_{T^{r-2}_\Phi(Z_{r-2})}(\{z_{r-1}\}(0^+)) \\ \\ (CA^{r-1}B)^{-1}[\{z_r\}(0^+) - z_{0,r}] \in -\partial\psi_{T^{r-1}_\Phi(Z_{0,r-1})}(\{z_r\}(0^+)) \end{cases} \tag{5.20}$$

and

$$\{\xi\}(0^+) = \xi_0. \tag{5.21}$$

Measure Differential Formalism: Problem MP(z_0; I) Find $z_i \in \mathscr{F}_\infty(I;\mathbb{R})$ $(1 \leqslant i \leqslant r)$ and $\xi_i \in \mathscr{F}_\infty(I;\mathbb{R})$ $(1 \leqslant i \leqslant n-r)$ such that

$$dz_i - z_{i+1}(t)\mathrm{d}t \in -\partial\psi_{T_\Phi^{i-1}(Z_{i-1}(t^-))}(z_i(t^+)) \quad \text{on} \quad I \ \ (1 \leqslant i \leqslant r-1), \tag{5.22}$$

$$(CA^{r-1}B)^{-1}[dz_r - CA^rW^{-1}z(t)\mathrm{d}t] \in -\partial\psi_{T_\Phi^{r-1}(Z_{r-1}(t^-))}(z_r(t^+)) \quad \text{on} \quad I, \tag{5.23}$$

and

$$\mathrm{d}\xi - (A_\xi\xi(t) + B_\xi z_1(t))\mathrm{d}t = 0 \quad \text{on} \ I. \tag{5.24}$$

The system in (5.22) and (5.23) has to be interpreted in the following sense: Find nonnegative real-valued Radon measure $\mathrm{d}\mu_i$ relative to which the Lebesgue measure $\mathrm{d}t$ and the Stieltjes measure $\mathrm{d}z_i$ possess densities $\frac{\mathrm{d}t}{\mathrm{d}\mu_i}$ and $\frac{\mathrm{d}z_i}{\mathrm{d}\mu_i}$, respectively, such that

$$\frac{\mathrm{d}z_i}{\mathrm{d}\mu_i}(t) - z_{i+1}(t)\frac{\mathrm{d}t}{\mathrm{d}\mu_i}(t) \in -\partial\psi_{T_\Phi^{i-1}(Z_{i-1}(t^-))}(z_i(t^+)), \ \mathrm{d}\mu_i - \text{a.e. } t \in I \tag{5.25}$$

with $(1 \leqslant i \leqslant r-1)$, and

$$(CA^{r-1}B)^{-1}\left[\frac{\mathrm{d}z_r}{\mathrm{d}\mu_r}(t) - CA^rW^{-1}z(t)\frac{\mathrm{d}t}{\mathrm{d}\mu_r}(t)\right] \in -\partial\psi_{T_\Phi^{r-1}(Z_{r-1}(t^-))}(z_r(t^+)),$$

$$\mathrm{d}\mu_r - \text{a.e. } t \in I. \tag{5.26}$$

The two formalisms are related as explained in the next proposition.

Proposition 5.5. *(i) Let $(z_1,\ldots,z_r,\xi) \in (\mathscr{T}_\infty(I))^n$ be a solution of problem SP(z_0; I). Then*

$$deg(z_i) \leqslant i \ \ (1 \leqslant i \leqslant r),$$

$$z_1 = \{z_1\} \in \mathscr{F}_\infty(I;\mathbb{R}), \ \ \xi = \{\xi\} \in (\mathscr{F}_\infty(I;\mathbb{R}))^{n-r} \cap (C^0(I;\mathbb{R}))^{n-r}$$

and $(\{z_1\},\ldots,\{z_r\},\xi)$ is a solution of problem MP(z_0; I).

(ii) Let $(w_1,\ldots,w_r,\xi) \in (\mathscr{F}_\infty(I;\mathbb{R}))^n$ be a solution of problem MP(z_0; I) such that for each $1 \leqslant i \leqslant r-1$, the measure $dw_i - w_{i+1}dt$ is atomic. Let $z_1,\ldots,z_r$ be defined by

$$z_1 \stackrel{\Delta}{=} w_1$$

and

$$z_i \stackrel{\Delta}{=} w_i + \sum_{j=1}^{i-1}\Big(\sum_{t_k \in E_0(w_j)} (w_j(t_k^+) - w_j(t_k^-))\delta_{t_k}^{(i-j-1)}\Big) \ \ (2 \leqslant i \leqslant r).$$

Then $(z_1,\ldots,z_r,\xi) \in (\mathscr{T}_\infty(I))^n$ and is a solution of problem SP(z_0; I).

Proposition 5.5 is crucial since it implies that once the solutions of the problem MP have been calculated, it is a simple matter to deduce the solutions of the problem SP. This relies a lot on the properties of the distributions generated by RCSLBV functions, like the fact that the set of jumps of the state is countable, and that all the state variables possess right and left limits everywhere.

Let us rewrite the inclusion (5.19) in a more compact form. Let us define the matrices $G \in \mathbb{R}^{r\times r}, \bar{G} \in \mathbb{R}^{n\times r}$ and $H \in \mathbb{R}^{r\times n}$ as follows:

$$\bar{G} = \begin{pmatrix} G \\ 0_{(n-r)\times r} \end{pmatrix} \tag{5.27}$$

with

$$G = \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \ldots & 0 & 1 & 0 \\ & & & & \\ 0 & \ldots & \ldots & 0 & CA^{r-1}B \end{pmatrix} \tag{5.28}$$

and

$$H = (I_r \;\; 0_{r\times(n-r)}). \tag{5.29}$$

Let us now set

$$\mathrm{d}\nu \stackrel{\Delta}{=} (\mathrm{d}\nu_1, ..., \mathrm{d}\nu_r, 0_{1\times(n-r)})^{\mathrm{T}}.$$

We have

$$\mathrm{d}\{z\} = WAW^{-1}\{z\}\mathrm{d}t + \bar{\mathrm{G}}\,\mathrm{d}\nu = WAW^{-1}\{z\}^{+}\mathrm{d}t + \bar{\mathrm{G}}\,\mathrm{d}\nu. \tag{5.30}$$

The upper script $+$ in $\{z\}^+$ means that since the functions $\{z\}(\cdot)$ is right-continuous, taking the function or its right limit is equivalent for the integration.

5.4.4 Some Qualitative Properties

The HOSP being an extension of the sweeping process, one may expect that the jump rules for the variables $\{z_i\}(\cdot)$ will look like the velocity jumps of the second-order sweeping process. This is confirmed by the next proposition.

Proposition 5.6. *Let $m = 1$, and z be a solution of problem $SP(z_0;\ I)$. Then, for each $t \in I$ and for all $1 \leqslant i \leqslant r-1$, we have*

$$\{z_i\}(t^+) - \{z_i\}(t^-) \in -\partial\psi_{T_{\Phi}^{i-1}(\{Z_{i-1}\}(t^-))}(\{z_i\}(t^+))$$

if and only if

$$\{z_i\}(t^+) = prox\left[T_{\Phi}^{i-1}(\{Z_{i-1}\}(t^-)); \{z_i\}(t^-)\right].$$

If $CA^{r-1}B > 0$ then

$$\{z_r\}(t^+) - \{z_r\}(t^-) \in -CA^{r-1}B\,\partial\psi_{T_{\Phi}^{r-1}(\{Z_{r-1}\}(t^-))}(\{z_r\}(t^+))$$

if and only if

$$\{z_r\}(t^+) = prox\left[T_\Phi^{r-1}(\{Z_{r-1}\}(t^-)); \{z_r\}(t^-)\right]$$

These results continue to hold in the MIMO case $m \geqslant 2$, provided the Markov parameter $CA^{r-1}B = (CA^{r-1}B)^{\mathrm{T}} > 0$. One ingredient of LCS in (4.3) is the complementarity relation between the multiplier λ and a linear function of the state $w(\cdot)$. This is generalized as follows in the HOSP.

Theorem 5.7. *Let z be a solution of problem $SP(z_0; I)$. Then, for each $t \in I$, we have*

$$0 \leqslant z_1(t^+) \perp d\nu_r(\{t\}) \geqslant 0 \tag{5.31}$$

and

$$0 \leqslant z_1(t^+) \perp \chi_r(t) \geqslant 0, \text{ a.e. } t \in I. \tag{5.32}$$

From the complementarity conditions (5.32) and from (5.12), and from the transformed dynamics (5.5), one easily deduces the LCP that $\chi_r(t)$ satisfies on time intervals on which the solution is smooth and evolves on the boundary $\{z_1 = 0\}$. Due to (5.31) and (5.32) one may consider that the HOSP is also an extension of the LCSs, with relative degree $\geqslant 1$.

5.5 Well-Posedness of the HOSP

Let us first recall that for $z \in \mathbb{R}^n$, we use the notation $z^{\mathrm{T}} = (\bar{z}^{\mathrm{T}}, \xi^{\mathrm{T}})$ as in (5.3) with $\bar{z} \in \mathbb{R}^r$ and $\xi \in \mathbb{R}^{n-r}$. Let $z_0^{\mathrm{T}} = (\bar{z}_0^{\mathrm{T}}, \xi_0^{\mathrm{T}})$ be given.

Definition 5.8. *Let $0 \leqslant a < b \leqslant T \leqslant +\infty$ be given. We say that a solution $z \in (\mathscr{T}_\infty([0,T)))^n$ of problem $SP(z_0;\ [0,T))$ is regular on $[a,b)$ if for each $t \in [a,b)$, there exists a right neighborhood $[t, t+\sigma)$ $(\sigma > 0)$ such that the restriction of $\{z\}$ to $[t, t+\sigma)$ is analytic.*

This definition does not preclude the existence of accumulations of state jumps, as the size of the right neighborhoods is not lower bounded. If t^* is such an accumulation, then a $\sigma > 0$ exists for all $t < t^*$, for t^*, and for all $t > t^*$.

Theorem 5.9. (Global existence and uniqueness) *Let $m = 1$, and $\Lambda = \sup_{x \in \mathbb{R}^n, x \neq 0} \frac{\|WAW^{-1}x\|}{\|x\|}$ (see (5.2)). Suppose that $CA^{r-1}B > 0$. For each $z_0 \in \mathbb{R}^n$, problem $SP(z_0;\ [0,+\infty))$ has at least one regular solution z such that:*

- **(i)** $z_1 \equiv \{z_1\} \geqslant 0$ *on* $\mathbb{R}^+$*;*
- **(ii)** $\|\{z\}(t)\| \leqslant e^{\Lambda t}\|z_0\|, \ \forall\, t \in \mathbb{R}^+$.
- **(iii)** *(Uniqueness in the class of regular solutions) If z^* denotes a regular solution of problem* $\mathbf{SP}(z_0;\ [0,T^*))$ $(0 < T^* \leqslant +\infty)$ *then* $\langle z^*, \varphi\rangle = \langle z, \varphi\rangle, \ \forall \varphi \in C_0^\infty([0,T);\mathbb{R}^n)$.

It is stressed that uniqueness holds in the class of regular solutions only (as we saw in Example 2.21, various solutions may exist at the same time for nonsmooth systems). In the MIMO case, Theorem 5.9 continues to hold, provided the matrix $CA^{r-1}B$ is a Stieltjes matrix, i.e., a nonsingular symmetric M-matrix. This assumption secures that the matrix $CA^{r-1}B$ is positive definite and $(CA^{r-1}B)^{-1}$ is nonnegative in the sense that $(CA^{r-1}B)^{-1}_{ij} \geqslant 0$ for all $i, j \in \{1, ..., n\}$.

5.6 Summary of the Main Ideas of Chapters 4 and 5

The class of dynamical systems subject to state inequality constraints may be embedded into complementarity dynamical systems. In turn, complementarity systems are much too general to be studied without restricting oneself to subclasses, like linear CS with dissipative properties, gradient CS, etc. The relative degree between the two complementary variables is a fundamental parameter. CSs form an important class of dynamical systems due to a wide range of applications. The relationships between CSs and other nonsmooth dynamical systems are numerous, and one may take advantage of them to study their well-posedness and their simulation.

6

Specific Features of Nonsmooth Dynamical Systems

The material in the foregoing chapters shows that some nonsmooth dynamical systems are rather "gentle": for instance linear complementarity systems with a direct feedthrough matrix D that is a P-matrix are ordinary differential equations with a Lipschitz-continuous right-hand side. The class of "gentle" NSDS is, however, a small subclass. Most of the NSDS may have either jumping solutions or nonunique solutions or solutions that do not depend continuously on the initial data. This is the case of complementarity Lagrangian systems: when multiple impacts occur, solutions may be discontinuous with respect to the initial conditions. When Coulomb friction is present, the contact force may diverge to infinity and produce unexpected subsequent motion: this is the well-known Painlevé paradoxes, better called *frictional paroxysms*.[1] We give a brief account of these two phenomena in this chapter.

6.1 Discontinuity with Respect to Initial Conditions

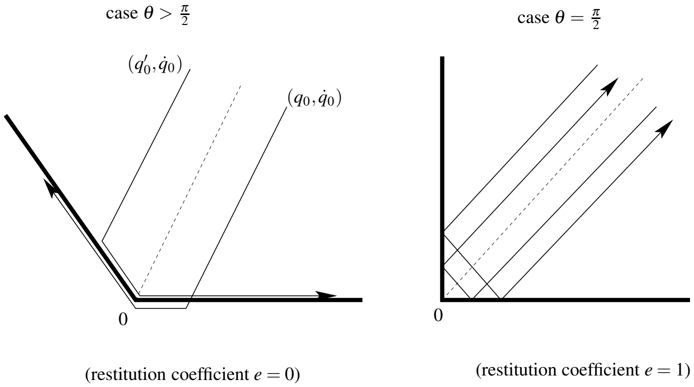
Let us consider here a complementarity Lagrangian system, with several unilateral constraints (i.e., $\nu \geqslant 2$ in (3.14)). This phenomenon is closely linked to the kinetic angle between the codimension one surfaces $\Sigma_\alpha = \{q \in \mathbb{R}^n \mid g^\alpha(q) = 0\}$, $1 \leqslant \alpha \leqslant \nu$. The kinetic angle between Σ_α and Σ_β is defined as

$$\cos\theta_{\alpha\beta} = \frac{\nabla g^{\alpha,\mathrm{T}}(q) M^{-1}(q) \nabla g^\beta(q)}{(\nabla g^{\alpha,\mathrm{T}}(q) M^{-1}(q) \nabla g^\alpha(q))^{\frac{1}{2}} (\nabla g^{\beta,\mathrm{T}}(q) M^{-1}(q) \nabla g^\beta(q))^{\frac{1}{2}}}. \tag{6.1}$$

6.1.1 Impact in a Corner

The simplest case of a multiple impact is depicted in Fig. 6.1, where we consider a particle with mass $m > 0$. The left figure depicts the case of a kinetic angle (equal here to the Euclidean angle since we deal with a particle) larger than $\frac{\pi}{2}$. Moreau's

[1] After a suggestion of Jean Jacques Moreau.


case $\theta > \frac{\pi}{2}$

case $\theta = \frac{\pi}{2}$

(restitution coefficient $e = 0$)

(restitution coefficient $e = 1$)

Fig. 6.1. Impact at a corner

rule in (2.101) is chosen for the impact. The dashed line is the bisector of the angle. A trajectory initialized on the bisector with its velocity oriented along the bisector, will hit at the corner O and stick there. Trajectories initialized with the same velocity, but on each side of the bisector, will remain close to each other until they hit the boundary. Then due to the impacts and the constraint boundaries orientations, they will diverge from each other and from the corner O. Solutions are discontinuous with respect to the initial data.

The figure on the right is for an angle equal to $\frac{\pi}{2}$. The impact being elastic, there is no kinetic energy loss, and trajectories remain in a neighborhood of the bisector line. Solutions are continuous with respect to the initial data.

These two cases graphically illustrate that depending on the geometry of the singularity, solutions may be discontinuous with respect to the initial data. In the next two sections a more general result due to Paoli (2005a) and a detailed mechanical example due to Heemels et al. (2000) are presented. We also refer to Kozlov & Treshchev (1991) for more details on the conditions that guarantee the continuity property. Notice that from the numerical point of view, the discontinuity with respect to the initial data means that there is a certain degree of uncertainty (or randomness) in the result. Indeed depending on the numerical uncertainty, the time-step value, and various thresholds to be implemented, the solution will "choose" different paths. This simply reflects the reality of mechanics: systems that possess such a property will behave in a random way or will be very sensitive to initializations.

6.1.2 A Theoretical Result

The next result is taken from Paoli (2005a). We consider a Lagrangian system as in (3.4), with perfect time-invariant unilateral constraints (3.14) and Moreau's impact law (2.103)–(2.106). The following assumptions are supposed to hold:

(a) $M(q) = M^{\mathrm{T}}(q) > 0$ and is continuously differentiable.
(b) All data expressing generalized forces are continuous functions.

(c) The constraints functions $g^{\alpha}(\cdot)$ are continuously differentiable, with locally Lipschitz gradient that does not vanish in a neighborhood of $g^{\alpha}(q) = 0$.
(d) The active constraints are functionally independent, i.e., their gradients are independent vectors of $\mathbb{R}^n$.
(e) The initial data satisfy $g^{\alpha}(q_0) \geqslant 0$ for all $\alpha \in \{1, ..., \nu\}$, and $\nabla g^{\alpha,\mathrm{T}}(q_0)\dot{q}_0 \geqslant 0$ when $g^{\alpha}(q_0) = 0$.

These are constraints qualification conditions and admissibility of the initial data. Let $F(q, \dot{q}, t)$ generically denote the generalized forces in (3.4). In the next theorem, we shall consider a sequence of data $\{q_{0,k}, \dot{q}_{0,k}, M_k, F_k, g^{\alpha,k}, 1 \leqslant \alpha \leqslant \nu\}_{k\geqslant 0}$. The set of admissible positions is $\Phi_k = \{z \in \mathbb{R}^n \mid g^{\alpha}(z) \geqslant 0, \; \forall\, 1 \leqslant \alpha \leqslant \nu\}$, and for all $z \in \mathbb{R}^n$ we define the set of active constraints indices $I_k(z) = \{i \in \{1, .., \nu\} \mid g^{\alpha,k}(z) \leqslant 0\}$.

Theorem 6.1. *Let the following hold:*

- *The sequence $\{q_{0,k}, \dot{q}_{0,k}\}_{k\geqslant 0}$ converges to a limit $\{q_0, \dot{q}_0\}$.*
- *The sequences $\{M_k(\cdot)\}_{k\geqslant 0}$, $\{dM_k(\cdot)\}_{k\geqslant 0}$, $\{g^{\alpha,k}(\cdot)\}_{k\geqslant 0}$, and $\{\nabla g^{\alpha,k}(\cdot)\}_{k\geqslant 0}$, $1 \leqslant \alpha \leqslant \nu$, converge uniformly on the compact subsets of $\mathbb{R}^n$ to limits $M(\cdot)$, $dM(\cdot)$, $g^{\alpha}(\cdot)$, and $\nabla g^{\alpha}(\cdot)$, respectively.*
- *The sequence $\{F_k(\cdot)\}_{k\geqslant 0}$ converges uniformly on the compact subsets of $[0, T] \times \mathbb{R}^n \times \mathbb{R}^n$ to a limit $F(\cdot, \cdot, \cdot)$.*
- *The set of data $\mathscr{D} = (q_0, \dot{q}_0, M(\cdot), F(\cdot), g^{\alpha}(\cdot))$, $1 \leqslant \alpha \leqslant \nu$, satisfies the above assumptions (a)–(e).*
- *There exists $\delta \in [0, T]$ such that for all $k \geqslant 0$, the Cauchy problem associated with the initial data $\mathscr{D}_k$ admits a solution $q_k(\cdot)$ defined on $[0, T]$.*

Then there exists $\delta' \in (0, \delta]$ and a subsequence of $\{q_k(\cdot)\}_{k\geqslant 0}$ which converges uniformly on $[0, \delta']$ to a limit $q(\cdot) \in C^0([0, \delta']; \Phi)$ such that $\frac{d}{dt}q(\cdot) = \dot{q}(\cdot)$ is a BV function. There exist also measures λ such that the dynamical equations (3.17) and (3.20) are satisfied by the limits. Moreover, let the kinetic angle conditions for the limit functions

$$\langle \nabla g^{\alpha}(q), M^{-1}(q) \nabla g^{\beta}(q) \rangle \leqslant 0 \;\; \textit{if}\, e = 0, \tag{6.2}$$

$$\langle \nabla g^{\alpha}(q), M^{-1}(q) \nabla g^{\beta}(q) \rangle = 0 \;\; \textit{if}\, e \in (0, 1] \tag{6.3}$$

hold for all indices $\alpha, \beta \in I(q)$, $\alpha \neq \beta$, and for all $t \in (0, \delta)$. Then the limit velocity $\dot{q}(\cdot)$ also satisfies Moreau's impact rule in (2.103)–(2.106).

The solutions are therefore continuous with respect to the initial data $(q_0, \dot{q}_0)$. Indeed if we denote the solution starting at $(q_1, \dot{q}_1)$ at time $t = 0$ as $\varphi(t; q_1, \dot{q}_1)$, we have that $\lim_{k\to+\infty} \varphi(t; q_{0,k}, \dot{q}_{0,k}) = \varphi(t; q_0, \dot{q}_0)$ for any sequence $\{q_{0,k}, \dot{q}_{0,k}\}_{k\geqslant 0}$, and $\varphi(t; q_0, \dot{q}_0)$ is a solution of the complementarity Lagrangian system.

6.1.3 A Physical Example

The discontinuity with respect to initial data may be seen as a sensitivity of the solutions to the order in which the constraints are activated. This is visible in the

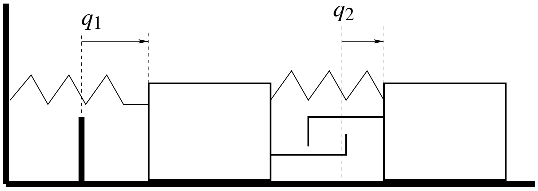


Fig. 6.2. A two-cart system with a hook

above corner example. In Heemels et al. (2000) calculations have been made for the system of Fig. 6.2, using Moreau's impact rule in (2.100).

With an obvious definition of the state variables and $q_1 = x_1$, $q_2 = x_2$, the dynamics of the two-cart system is:

$$\begin{cases} \dot{x}_1(t) = x_3(t) \\ \dot{x}_2(t) = x_4(t) \\ \dot{x}_3(t) = -2x_1(t) = x_2(t) + \lambda_1 + \lambda_2 \\ \dot{x}_4(t) = x_1(t) - x_2(t) - \lambda_2 \\ \\ 0 \leqslant x_1(t) \perp \lambda_1(t) \geqslant 0 \\ 0 \leqslant x_1(t) - x_2(t) \perp \lambda_2(t) \geqslant 0 \end{cases} . \tag{6.4}$$

Consider the initial data $x_0(\varepsilon) = (\varepsilon, \varepsilon, -2, 1)^{\mathrm{T}}$, with $\varepsilon \geqslant 0$. For $\varepsilon = 0$ the solution initially jumps to the origin $(0,0,0,0)^{\mathrm{T}}$. This is a plastic impact. After this collision the system stays at rest at its equilibrium position. For $\varepsilon > 0$, the constraint of the hook $x_1(t) - x_2(t) \geqslant 0$ becomes active, and there is a jump to $(\varepsilon, \varepsilon, -\frac{1}{2}, -\frac{1}{2})^{\mathrm{T}}$. After this event the system evolves on the hook constraint, until it hits the stop $x_1(t) \geqslant 0$. The pre-impact state is $(0, 0, -\frac{1}{2} + g(\varepsilon), -\frac{1}{2} + g(\varepsilon))^{\mathrm{T}}$ where $g(\cdot)$ is a continuous function with $g(0) = 0$. The post-impact state is $(0, 0, 0, -\frac{1}{2} + g(\varepsilon))^{\mathrm{T}}$, which converges to $(0, 0, 0, -\frac{1}{2})^{\mathrm{T}}$ when $\varepsilon \to 0$. Further calculations show that the initial state $(0, -\varepsilon, -2, 1)^{\mathrm{T}}$, $\varepsilon \geqslant 0$, yields after two jumps to the state $(0, 0, \frac{1}{2}, \frac{1}{2})^{\mathrm{T}}$. This is a trajectory that first hits the stop and then hits the hook constraint. So the solution starting at $(0, 0, -2, 1)^{\mathrm{T}}$ is discontinuous with respect to the initial conditions. Looking at Theorem 6.1, we deduce that the kinetic angle between the two constraints boundaries is larger than $\frac{\pi}{2}$.

6.2 Frictional Paroxysms (the Painlevé Paradoxes)

It has been known since a long time that the introduction of Coulomb friction in mechanical systems with bilateral or unilateral constraints may yield to apparently strange and paradoxical situations. Roughly speaking, there may exist states for which the system possesses several (possibly an infinity of) solutions (indeterminate states) or no solutions (inconsistent states), and the contact force may diverge to infinity. Such problems exist not only in dynamical situations, but also in equilibrium statics positions. In the statics case, the typical examples of a rod or a disc inserted between convergent walls are treated for instance in Moreau (2006) and Sect. 5.1 in Brogliato (1999). In the dynamical case, the problem is that the coupling between the unilaterality and Coulomb friction complicates the system's behavior a lot. There may exist subsets of the state space in which no solution exists or several solutions may exist at each instant. There may also exist configurations at which the contact force takes unbounded values. The classical planar problem of a rigid rod sliding on a rough plane is analyzed in detail in Génot & Brogliato (1999). It is shown that while in sliding motion, the dynamics reduces to a scalar singular ODE of the form $\dot{x}(t) = \frac{f(x(t))}{g(x(t))}$, where $f(\cdot)$ and $g(\cdot)$ may be simultaneously singular. When reaching the neighborhood of such singular states, the trajectories either detach from the constraint or go through the singularity. What happens is that though the contact force diverges, the impulse (i.e., the integral of the force) remains bounded. Such a phenomenon has therefore no link with an impact. A second important behavior occurs when a trajectory tends to enter a subset of the state space, in which the contact force LCP has no solution at all (a subset of inconsistent states). Then a principle of maximal dissipation (i.e., a supplementary modeling assumption) says that the velocity jumps to some value at which the LCP is solvable. This is sometimes called a *tangential impact*.

We do not investigate this topic further as this would lead us too far away from our main purpose. We simply mention that Moreau's catching-up algorithm is able to reproduce frictional paroxysms. See Liu et al. (2007) for applications in robotics and Zhao et al. (2007) for experimental validations.

6.3 Infinity of Events in a Finite Time

The occurrence of an infinity of events (impacts, stick-slip transitions with Coulomb friction) within a finite time interval is an important and common phenomenon in many nonsmooth systems. From a numerical point of view this may create big problems. Let us briefly recall two typical cases.

6.3.1 Accumulations of Impacts

It is well known that the bouncing ball in (1.96) with $e \in (0,1)$ and $q(0) > 0$ has a trajectory that stabilizes in finite time on $q = 0$ after an infinity of rebounds, see for instance Brogliato (1999, Sect. 7.1.4). More complex cases have been analyzed which share the same property, see Cabot & Paoli (2007) and Wang (1993). Except in particular cases, it is expected that most complementarity mechanical systems possess trajectories with accumulations of events. An important property that is secured by the well-posedness results that may be found in Ballard (2000), Dzonou & Monteiro Marques (2007), Paoli & Schatzman (2002a,b), Mabrouk (1998) and Monteiro Marques (1993) is that since the velocity is LBV, then the set of impact times is countable. Consequently solutions are smooth in the right neighborhood of any $t \geqslant 0$. This is a property that is also encountered in higher order systems of Chap. 5, see Theorem 5.9.

6.3.2 Infinitely Many Switchings in Filippov's Inclusions

Let us consider the planar example:

$$\left\{\begin{array}{l} \dot{x}_1(t) \in -\mathrm{sgn}(x_1(t)) + 2\mathrm{sgn}(x_2(t)) \\ \dot{x}_1(t) \in -2\mathrm{sgn}(x_1(t)) - \mathrm{sgn}(x_2(t)) \end{array}\right. , \tag{6.5}$$

where $\mathrm{sgn}(\cdot)$ is the set-valued sign function. The trajectories initialized outside the origin reach the origin in finite time and with an infinite number of crossings of the switching surfaces $x_1 = 0$ and $x_2 = 0$. The finite time convergence is easy to establish as the time intervals between two switches satisfy a geometric series and consequently have a finite sum. For instance starting at $(0,1)$ yields a convergence time equal to $\sum_{n=1}^{+\infty} \frac{1}{3^n} = \frac{1}{2}$.

Let us now reverse the time in the system (6.5):

$$\left\{\begin{array}{l} \dot{x}_1(t) \in \mathrm{sgn}(x_1(t)) - 2\mathrm{sgn}(x_2(t)) \\ \dot{x}_1(t) \in 2\mathrm{sgn}(x_1(t)) + \mathrm{sgn}(x_2(t)) \end{array}\right. . \tag{6.6}$$

There is an infinity of trajectories which start with the initial data $(0,0)$, and except for the trivial solution that stays at the origin, they all cross the switching surfaces an infinity of times. This system has an infinity of spontaneous switches from the origin.

6.3.3 Limit of the Saw-Tooth Function in Filippov's Systems

Suppose one integrates the differential inclusion $\dot{x}(t) \in -\mathrm{sgn}(x(t))$, $x(0) = x_0 > 0$, with some delay in the switch between $+1$ and -1 because some kind of hysteresis function is implemented around the switching surface $x = 0$. Such a procedure is often used in systems and control, in order to avoid too many switches in practice when the system attains the sliding surface. The solution $x_\varepsilon(\cdot)$ is then a saw-toothed, or zig-zag function, i.e., a function that oscillates around $x = 0$, with peaks at $-\varepsilon < 0$ and $+\varepsilon > 0$ occurring at times t_k with $t_{k+1} - t_k = 2\varepsilon$. On intervals (t_k, t_{k+1}) the solution is linear with slope $+1$ or -1, alternatively. The derivative $\dot{x}_\varepsilon(\cdot)$ thus exists almost everywhere and is equal either to $+1$ or to -1 on the intervals (t_k, t_{k+1}). Therefore the second-order derivative is equal almost everywhere to 0 and is a Dirac measure at times t_k. Let the hysteresis size go to zero, i.e., let $\varepsilon \to 0$. Then $x_\varepsilon(\cdot)$ converges uniformly towards the zero function. Clearly the number of "events" (the instants t_k) goes to infinity on any interval of time with positive measure. Though the derivative seems to converge to the zero function, it does not because $|\dot{x}_\varepsilon(t)| = 1$ almost everywhere. The limit of the second-order derivative deserves attention. Indeed $\ddot{x}(\cdot)$ is no longer a function but consists of an accumulation of Dirac measures at each time t! One mathematical interpretation is that the infinitesimal zig-zag curve can be assigned the slopes $+1$ and -1 with probability $\frac{1}{2}$ at each t.

The crucial conclusion to be drawn from this example is that the type of “infinity of events” that occurs is quite different from the above ones. A commonly used terminology coming from computer science is to name *Zeno* all phenomena involving some way or another an infinity of events within a finite time interval. As these examples show, this is a very vague notion lacking serious mathematical foundations. One should better speak of solutions as being piecewise *something* (meaning that they are *something* on intervals of the form $[t_k, t_{k+1})$ with $t_{k+1} - t_k > \delta > 0$ for all k) or BV or absolutely continuous or continuous with a piecewise *something* derivative, etc.

Part II

Time Integration of Nonsmooth Dynamical Systems

Introduction

The following notation is used throughout this part. We denote by $0 = t_0 < t_1 < \cdots < t_k < \cdots < t_N = T$ a finite partition (or a subdivision) of the time interval $[0,T]$ $(T > 0)$. The integer N stands for the number of time intervals in the subdivision. The length of a time step is denoted by $h_k = t_{k+1} - t_k$. For simplicity sake, we consider only in the sequel a constant time length $h = h_k$ $(0 \leqslant k \leqslant N-1)$. Then $N = \frac{T}{h}$. The approximation of $f(t_k)$, the value of a real function $f(\cdot)$ at the time t_k, is denoted by f_k .

In this part we shall concentrate on two types of algorithms for nonsmooth systems: event-driven and time-stepping. Let us first provide a brief description of such schemes.

Event-driven algorithms: The principle of the event-driven schemes is based on the time decomposition of the dynamics in modes, time intervals in which the dynamics is smooth, and discrete events, i.e., times where the dynamics is nonsmooth. It is based on the following assumptions guaranteeing the existence and the consistency of such a decomposition:

- The definition and the localization of the discrete events. The events may be defined as the instants when the dynamics is nonsmooth or not sufficiently regular, and we suppose that the set of such events is negligible with respect to the Lebesgue measure.
- The definition of time intervals of nonzero length based on the fact that the events are of finite number and "well separated" in time. Clearly, for a dynamics based on BV functions, this assumption is not satisfied. If we assume that finite accumulations of impacts or Zeno-state will not occur, the decomposition can be obtained. A way to avoid such situations is to change the model slightly.

From the numerical point of view, the event-driven schemes use the decomposition in time of the dynamics in order to solve the following steps:

- detect and solve the nonsmooth dynamics at events with a reinitialization rule of the state,
- integrate the smooth dynamics between two events with any ODE or DAE solvers with root findings.

The event-driven methods may also be called *nonsmooth event-tracking methods*, borrowing from the partial differential equations terminology (LeVeque, 1990). Consider complementarity systems. Then the calculation of the reinitializations and the detection of the events of the "constraint deactivation" type are monitored by complementarity problems. This is a fundamental point that implies event-driven methods for LCS are not enumerative methods. See below for more details.

Time-stepping algorithms: The principle of time-stepping schemes is to write down a time-discretization of the whole dynamical system (the smooth dynamics, the complementarity conditions) and to form a nonsmooth one-step problem which, once solved, allows the scheme to advance from step k to step $k+1$. Contrary to event-driven schemes, the detection of the events is considered on the same footing as the rest of the integration process, i.e., there is no accurate event detection algorithms and reinitialization at the event. These schemes may be called *nonsmooth event-capturing methods* (LeVeque, 1990).

These two methods have been examined on particular cases in Chap. 1 (see Sects. 1.1.6, 1.2.3, and 1.2.4). Event-driven strategies have the following drawbacks. First, if the number of events is too large, the algorithm cannot efficiently advance in time. Secondly, the method is very sensitive to the numerical tolerances used for the detection of the events (*the choice of the espilons*). Thirdly, it needs a reformulation of the generalized equations at different kinematic levels. They are well suited for systems with well-separated events. The advantage is that during periods with no events, the integration is accurate. The main advantages of time-stepping methods are that they accommodate with large numbers of events (even accumulations) and are able to work without accurate detections of the events. A drawback is their low order.

Remark 6.2. To shed more light on the last point raised about event-driven schemes, let us consider once again the case of an ideal diode. Instead of relying on the tools described in Sect. 1.1.1, one may simply consider the diode as a pure logical component thanks to conditional "if" and "then" statements. The curve of Fig. 1.1b can be parameterized by a parameter s, and the following script may be defined (Elmqvist et al., 2001; Mattsson et al., 1999):

$$\begin{aligned} \mathsf{off} &= \mathsf{s} < 0 \\ \lambda &= \textbf{if} \quad \mathsf{off} \quad \textbf{then} \quad -\mathsf{s} \quad \textbf{else} \quad 0 \\ y &= \textbf{if} \quad \mathsf{off} \quad \textbf{then} \quad 0 \quad \textbf{else} \quad \mathsf{s} \end{aligned}$$

Similar representations can be performed with ideal switches, piecewise linear model of MOS transistors. The main difficulty to view systems with ideal components this way is that for each new boolean variable like off, two modes of the hybrid dynamical system are possible. If we introduce n-boolean variables, in the worst case, 2^n modes have to be checked. Therefore the problem complexity is exponential and the problem quickly becomes intractable in practice. The mere writing of the dynamics becomes so cumbersome that it is not possible. This is directly related to the issue discussed in Sect. 8.6.1 and switching diagrams: it seems difficult to draw such a diagram with 2^n nodes when n is larger than say 6. It is then no surprise that the examples which are usually presented with such an approach are of (very) low dimensions. In practice n can be very large (about 100 for relatively simple electrical circuits like buck converters, to several thousands for mechanical systems with frictional unilateral constraints), rendering the use of CP solvers mandatory.

On the contrary, in the nonsmooth approach the discretized problem at each step can be reformulated as an LCP of the form

$$\begin{cases} w = Mz + q \\ 0 \leqslant w \perp z \geqslant 0. \end{cases} \tag{6.7}$$

Under some assumptions on the matrix M and on the vector q, numerical algorithms can be used with polynomial complexity, avoiding an exhaustive enumerative verification of each mode at exponential time.

To conclude this comment, from the mathematical point of view, the nonsmooth framework yields precise definitions of solutions together with uniqueness and existence results under appropriate assumptions. It is also quite useful for stability and control analysis (Goeleven & Brogliato, 2004, 2005; Brogliato, 2004; Camlibel et al., 2002b; Brogliato & Goeleven, 2005; Brogliato et al., 2007). The point is that it allows one to study some properties by looking at the properties of the CP, and not by analyzing conditional statements. From the numerical point of view, the use of specific algorithms (time-stepping schemes, LCP solvers with polynomial complexity) leads to an efficient simulation environment which takes advantage of the research works led in mathematical programming, see Chaps. 12 and 13. This is the object of this part, Part III, and Part IV.

7

Event-Driven Schemes for Inclusions with AC Solutions

In this chapter we deal with differential inclusions which possess absolutely continuous (AC) solutions. Consequently only the derivative of the solutions may possibly jump. This class of inclusions contains some of the examples examined in the foregoing chapters.

7.1 Filippov's Inclusions

7.1.1 Introduction

As we saw in Sect. 2.1, Filippov's systems are a specific type of differential inclusion in which the right-hand side is made of (smooth) vector fields that switch when the trajectory attains some surface $S \subset \mathrm{I\!R}^n$. The codimension of the switching surface and the existence of so-called *sliding* motions play an important role in the way the system may be simulated. There are different ways to define a Filippov's inclusion. One may suppose that $S = \{x \in \mathrm{I\!R}^n \mid c(x) = 0\}$ for some smooth function c: $\mathrm{I\!R}^n \to \mathrm{I\!R}$. The codimension of S is then 1, and the surface S divides the ambient space $\mathrm{I\!R}^n$ into two parts. Let us define m smooth functions $c_i : \mathrm{I\!R}^n \to \mathrm{I\!R}$. A more general case is when $\mathrm{I\!R}^n$ is divided into several subsets, whose boundaries are surfaces $S_A = \{x \in \mathrm{I\!R}^n \mid c_{a_1}(x) = c_{a_2}(x) = ... = c_{a_j}(x) = 0\}$, where $A = \{a_1, a_2, .., a_j)$ is a subset of $(1, 2, ..., m)$. Then the codimension of S_A is equal to j, the cardinal of A. In most applications one starts with codimension one switching surfaces, and the intersection of these surfaces define codimension $\geqslant 2$ switching submanifolds. One may also start directly from the definition of disjoint open sets R_i, $i = 1, ..., m$, which cover the ambient space with piecewise smooth boundary ∂R_i, and such that $f(x) = f_i(x)$ whenever $x \in R_i$. Then one may suppose that each set $R_i = \{x \mid g_i(x) < 0\}$ for some function $g_i : \mathrm{I\!R}^n \to \mathrm{I\!R}$. The relationship between both descriptions will be clarified in Sect. 7.1.2.

When the switching surface $S = \{x \in \mathrm{I\!R}^n \mid c(x) = 0\}$ is of codimension one, Filippov's notion of a solution says that

$$\dot{x}(t) \in \alpha f^{+}(x(t)) + (1 - \alpha) f^{-}(x(t)) \tag{7.1}$$

with $\alpha \in [0,1]$, and where the two vector fields are as in (2.14). On one side of S one has $\alpha = 1$, and on the other side $\alpha = 0$. On an attractive surface S the vector field is a convex combination of both vector fields, tangent to S. Notice that we can rewrite (7.1) equivalently as

$$\begin{aligned}\dot{x}(t) &\in \frac{1+\mathrm{sgn}(c(x(t))}{2} f^{+}(x(t)) + \frac{1-\mathrm{sgn}(c(x(t))}{2} f^{-}(x(t)) \\ &= \frac{1}{2} f^{+}(x(t)) + \frac{1}{2} f^{-}(x(t)) + \frac{1}{2}\mathrm{sgn}(c(x(t))\{f^{+}(x(t)) - f^{-}(x(t))\}\end{aligned} \tag{7.2}$$

where we recall that the sign set-valued function is defined as $\mathrm{sgn}(x) = 1$ if $x > 0$, $\mathrm{sgn}(x) = -1$ if $x < 0$, and $\mathrm{sgn}(x) \in [-1,1]$ if $x = 0$. Suppose now we are given two smooth functions $c_1 : \mathrm{I\!R}^n \to \mathrm{I\!R}$ and $c_2 : \mathrm{I\!R}^n \to \mathrm{I\!R}$ that separate the ambient space into four disjoint open subsets $R_1 = \{x \mid c_1(x) > 0 \text{ and } c_2(x) < 0\}$, $R_2 = \{x \mid c_1(x) < 0 \text{ and } c_2(x) < 0\}$, $R_3 = \{x \mid c_1(x) < 0 \text{ and } c_2(x) > 0\}$, and $R_4 = \{x \mid c_1(x) > 0 \text{ and } c_2(x) > 0\}$, with $\cup_{i=1}^{4} R_i \cup_{i=1}^{4} \partial R_i = \mathrm{I\!R}^n$. We denote $S = \{x \in \mathrm{I\!R}^n \mid c_1(x) = c_2(x) = 0\}$. Filippov's solution satisfies

$$\dot{x}(t) \in \sum_{i=1}^{4} \alpha_i f_i(x(t)) \tag{7.3}$$

with $\sum_{i=1}^{4} \alpha_i = 1$ and $\alpha_i \geqslant 0$, $i = 1,..,4$. In the interior of R_i one has $\alpha_i = 1$ and $\alpha_j = 0$, $j \neq i$. Notice that when $n = 2$ then S has dimension zero, so that the sliding motion reduces to $\dot{x}(t) = 0$ in S. One can rewrite (7.3) as

$$\begin{aligned}\dot{x}(t) \in\ & \frac{1+\mathrm{sgn}(c_1(x(t))}{2} \frac{1-\mathrm{sgn}(c_2(x(t))}{2} f_1(x(t)) \\ &+ \frac{1-\mathrm{sgn}(c_1(x(t))}{2} \frac{1-\mathrm{sgn}(c_2(x(t))}{2} f_2(x(t)) \\ &+ \frac{1-\mathrm{sgn}(c_1(x(t))}{2} \frac{1+\mathrm{sgn}(c_2(x(t))}{2} f_3(x(t)) \\ &+ \frac{1+\mathrm{sgn}(c_1(x(t))}{2} \frac{1+\mathrm{sgn}(c_2(x(t))}{2} f_4(x(t))\end{aligned} \tag{7.4}$$

that is

$$\begin{aligned}\dot{x}(t) \in\ & \frac{1}{4}\{f_1(x(t)) + f_2(x(t)) + f_3(x(t)) + f_4(x(t))\} \\ &+ \frac{1}{4}\mathrm{sgn}(c_1(x(t))\{f_1(x(t)) - f_2(x(t)) - f_3(x(t)) + f_4(x(t))\} \\ &+ \frac{1}{4}\mathrm{sgn}(c_2(x(t))\{-f_1(x(t)) - f_2(x(t)) + f_3(x(t)) + f_4(x(t))\} \\ &+ \frac{1}{4}\mathrm{sgn}(c_1(x(t)c_2(x(t))\{-f_1(x(t)) + f_2(x(t)) - f_3(x(t)) + f_4(x(t))\}.\end{aligned} \tag{7.5}$$

The passage from (7.4) to (7.5) is done by assuming that the meaning of $\mathrm{sgn}(c_i(x))$ when $c_i(x) = 0$ is that $\mathrm{sgn}(c_i(x)) = \lambda_i$ for some $\lambda_i \in [-1,1]$. This permits the factorizations in (7.5). One also checks that the sum of the four coefficients in (7.4) is always equal to 1, and that each coefficient lives in $[0,1]$. Thus the right-hand side of (7.4) is the convex hull of the set of vectors $\{f_1(x), f_2(x), f_3(x), f_4(x)\}$. Notice that the extension of (7.4) toward more complex cases with more functions $c_i(\cdot)$ and more subsets R_i may not be straightforward.

7.1.2 Stewart's Method

In this section Stewart's event-driven algorithm for solving discontinuous ODEs with high accuracy (Stewart, 1990) is presented. This algorithm relies on a specific way to describe the switching conditions, and on the construction of an LCP whose solution(s) allows one to compute the set of active switching surfaces. Doing so, the switching conditions look like the conditions encountered in complementarity systems. Most importantly, when several solutions exist (if for instance the switching surface is not attractive like in the inclusion $\dot{x}(t) \in \mathrm{sgn}(x(t))$ and $x(0) = 0$), then the LCP gives all the possible solutions. No "guess procedure" is needed to advance the integration.

7.1.2.1 Basic Assumptions

The starting point of Stewart's method is that there exists a family of disjoint open sets R_i, $i = 1, \ldots, m$, with $\mathbb{R}^n = \overline{\cup_{i=1}^m R_i}$, and

$$f(x) = f_i(x) \text{ whenever } x \in R_i$$

where $f_i(\cdot)$, $i = 1, \ldots, m$, is a family of smooth functions. It is assumed that the boundaries ∂R_i are piecewise smooth. The initial value problem (2.11) is therefore a discontinuous ODE that can be embedded into Filippov's inclusions. Let us define the *active set* as

$$I(x) = \{i \mid x \in \partial R_i\} \subset \{1, \ldots, m\}. \tag{7.6}$$

The inclusion is therefore

$$\dot{x}(t) \in \mathrm{conv}\{f_i(x(t)) \mid i \in I(x(t))\}. \tag{7.7}$$

Several basic assumptions are in order.

Assumption 5. *The sets R_i are given in terms of* discriminant *functions $h_i(\cdot)$, $i = 1, \ldots, m$, by*

$$R_i = \{x \in \mathbb{R}^n \mid h_i(x) < h_j(x) \textit{ for all } j \neq i\}. \tag{7.8}$$

Assumption 6. *The active-set function $t \mapsto I(x(t))$ changes value for only finitely many $t \in [t_0, t_f]$, where $[t_0, t_f]$ is the interval of integration.*

Assumption 7. *The functions $f_i(\cdot)$, $h_i(\cdot)$, $\nabla h_i(\cdot)$ are Lipschitz continuous for all $i = 1, .., m$.*

Assumption 6 means that trajectories which undergo an infinity of switchings (as for instance when the trajectories spiral down to asymptotically stable equilibrium point) are not permitted. One may also understand it as integrating such motions on a finite time interval only. The interval $[t_0,t_f]$ can therefore be divided into a finite number of subintervals (t_r,t_{r+1}) on which the active set $I(x(t))$ is constant, denoted as I_r. A solution $x(\cdot)$ that evolves along such a finite partition is called *piecewise active*. Let us provide some insight on Assumption 5, that is a very specific way to describe the sets R_i. Let us consider the two examples in Sect. 7.1.1. In the codimension one case (7.1) and (7.2) there are two sets R_1 and R_2 which can be described as

$$\begin{array}{l} R_1 = \{x \in \mathrm{I\!R}^n \mid -c(x) < c(x)\} \\ \\ R_2 = \{x \in \mathrm{I\!R}^n \mid c(x) < -c(x)\} \end{array} . \tag{7.9}$$

One may check that $R_1 = \{x \in \mathrm{I\!R}^n \mid c(x) > 0\}$ and $R_2 = \{x \in \mathrm{I\!R}^n \mid c(x) < 0\}$, and that $h_1(x) = -c(x)$ and $h_2(x) = c(x)$. These two sets correspond to what is denoted Ω^+ and Ω^- in Sect. 2.1.2. In the case of the system in (7.3) and (7.4), the following description is possible:

$$\begin{cases} h_1(x) = -c_1(x) - c_2(x) \\ h_2(x) = c_1(x) - c_2(x) \\ h_3(x) = c_1(x) + c_2(x) \\ h_4(x) = -c_1(x) + c_2(x). \end{cases} \tag{7.10}$$

Then one has

$$\begin{aligned} R_1 &= \{x \mid h_1(x) < h_j(x), j = 2,3,4\} = \{x \mid c_1(x) > 0 \text{ and } c_2(x) < 0\} \\ R_2 &= \{x \mid h_2(x) < h_j(x), j = 1,3,4\} = \{x \mid c_1(x) < 0 \text{ and } c_2(x) < 0\} \\ R_3 &= \{x \mid h_3(x) < h_j(x), j = 1,2,4\} = \{x \mid c_1(x) < 0 \text{ and } c_2(x) > 0\} \\ R_4 &= \{x \mid h_4(x) < h_j(x), j = 1,2,3\} = \{x \mid c_1(x) > 0 \text{ and } c_2(x) > 0\}. \end{aligned}$$

Suppose there are three functions $c_1(\cdot)$, $c_2(\cdot)$, and $c_3(\cdot)$ that divide the ambient space $\mathrm{I\!R}^n$ into three subsets $R_1 = \{x \mid c_1(x) > 0 \text{ and } c_3(x) < 0\}$, $R_2 = \{x \mid c_1(x) < 0 \text{ and } c_2(x) > 0\}$, and $R_3 = \{x \mid c_2(x) < 0 \text{ and } c_3(x) > 0\}$. Then one may add five other subsets with nonswitching conditions at their boundaries, and define eight functions $h_i(\cdot)$ and eight subsets R_i in a way similar to (7.10).

Example 7.1. As an example let us consider

$$\dot{x}(t) - g(t) \in \mathrm{sgn}(x(t)),\; x(0) = 0, t_0 = 0 \tag{7.11}$$

with $|g(t)| \leqslant 1$ for all t. Then $R_1 = \{x \in \mathrm{I\!R} \mid x > 0\}$, $f_1(x,t) = [g(t) - 1, 1]^{\mathrm{T}}$, $h_1(x,t) = -x$, and $R_2 = \{x \mid x < 0\}$, $f_2(x,t) = [g(t) + 1, 1]^{\mathrm{T}}$, $h_2(x,t) = x$. This system is not autonomous, however, considering a new state variable $\dot{y}(t) = 1$, $y(0) = 0$ brings it into the class of autonomous systems.

The next result shows that under some hypotheses on the functions $h_i(\cdot)$, the active set $I(x)$ can be computed from equalities, and furthermore some properties of the sets R_i are in order.

Assumption 8. *The functions $h_i(\cdot)$, $i = 1,...,m$, are such that for each $x \in \mathrm{I\!R}^n$ the set $\{\nabla h_i(x) \mid i \in I(x)\}$ is geometrically independent.*

A set of vectors $V_k = \{v_1,....,v_k\}$ is said to be *geometrically independent* if the affine plane generated by V_k (that is the set of all linear combinations $\sum_{i=1}^k a_i v_i$, with $\sum_{i=1}^k a_i = 0$) is not generated by any strict subset of V_k.

Lemma 7.2. *Suppose Assumptions 5–8 hold. Let $J \subset \{1,...,m\}$. The set $R_J = \{x \mid I(x) = J\}$ is a manifold of dimension $n+1-card(J)$. If J ranges over all subsets of $\{1,...,m\}$ with $card(J) \geqslant 2$, then $\cup_J R_J$ has measure zero in $\mathrm{I\!R}^n$.*

It will be seen later that Lemma 7.2 imposes in fact some qualification constraints on the sets R_i, so that a certain LCP will be well-posed.

7.1.2.2 Calculation of the Solution with a Constant Active Set

Let us assume that on the interval (t_r, t_{r+1}) the active set $I_r(x)$ is known and constant. The solution $x(\cdot)$ is absolutely continuous and satisfies

$$\dot{x}(t) \in \mathrm{conv}\{f_i(x(t)) \mid i \in I_r\} \tag{7.12}$$

almost everywhere on (t_r, t_{r+1}). As we saw above, the existence of $\dot{x}(t)$ implies that one can find $\mathrm{z}_l(t)$ such that

$$\dot{x}(t) = \sum_{l \in I_r} \mathrm{z}_l(t) f_i(x(t)), \;\; \sum_{l \in I_r} \mathrm{z}_l(t) = 1, \;\; \mathrm{z}_l(t) \geqslant 0 \text{ for all } l \in I_r. \tag{7.13}$$

The $\mathrm{z}_l(t)$ correspond to the α_i in (7.3), where the time dependency indicates that they refer to the derivative at time t. From Assumption 5 one has $h_i(x(t)) = h_j(x(t))$ for all $i, j \in I_r$, and all $t \in (t_r, t_{r+1})$.

Lemma 7.3. *Let $\mu(t) = \sum_{l \in I_r} \nabla h_i(x(t)) f_l(x(t)) \mathrm{z}_l(t)$ for all $i \in I_r$. Let $m_{ij}(x) = \nabla h_i(x) f_j(x)$, $M_{I_r}(x) = [m_{ij}(x) \mid i,j \in I_r]$, $\mathrm{z}_{I_r} = [\mathrm{z}_j(t) \mid j \in I_r]$. Then the system*

$$\begin{cases} M_{I_r}(x(t))\mathrm{z}_{I_r}(t) = \mu(t) e \\ \\ \mathrm{z}_{I_r}(t) \geqslant 0, \;\; e^{\mathrm{T}} \mathrm{z}_{I_r}(t) = 1 \end{cases} \tag{7.14}$$

where $e = (1\;1\;1....1)^{\mathrm{T}}$, either has no solution or its solution is unique for each t.

The proof uses (7.13) and the fact that $h_i(x(t)) = h_j(x(t))$ for all $i, j \in I_r$. The next problem is how to guarantee that the system (7.14) has a unique solution, and how to calculate it.

Example 7.4. Let us continue with Example 7.1. For $x = 0$ and $I_r = \{1,2\}$ one has

$$M_{I_r} = \begin{pmatrix} 1-g(t) & -1-g(t) \\ -1-g(t) & 1+g(t) \end{pmatrix}. \tag{7.15}$$

Let us add to Assumption 8 the following:

Assumption 9. *The set $\{f_i(x) \mid i \in I_r\}$ is geometrically independent. Let V be the vector space parallel to the affine plane generated by $\{f_i(x) \mid i \in I_r\}$, and W be the vector space parallel to the affine plane generated by $\{\nabla h_i(x) \mid i \in I_r\}$. Then $V \cap W^{\perp} = \{0\}$.*

Lemma 7.5. *Let Assumptions 8 and 9 hold, and $M_{I_r}(x)$ be of dimension $k \times k$. Then the matrix $\tilde{M}_{I_r}$ of dimension $(k-1) \times (k-1)$ whose entries are $\tilde{m}_{ij} = m_{ij} - m_{ik} - m_{kj} + m_{kk}$ for $i, j < k$, is nonsingular.*

Example 7.6. In Example 7.4 we get that $\tilde{M}_{I_r} = (1-g(t)) - (-1-g(t)) - (-1+g(t)) + (1+g(t)) = 4$.

We are now stating a result which shows how to calculate the terms $\mathrm{z}_l(t)$ for $l \in I_r$, when the matrix $\tilde{M}_{I_r}$ is nonsingular. The notation $M_{I_r,\alpha} = M_{I_r} + \alpha e e^{\mathrm{T}} = [m_{ij} + \alpha \mid i, j \in I_r]$ will be used. Also the inequality $M_{I_r,\alpha} > 0$ means that $(M_{I_r,\alpha})_{ij} > 0$ for all $i, j \in I_r$.

Theorem 7.7. *Let Assumptions 5–9 hold, and consider a solution $x(\cdot)$ of the Filippov's inclusion, with $I(x(t)) = I_r$ for all $t \in (t_r, t_{r+1})$. Then there are unique functions $\mathrm{z}_l(\cdot)$ for $l \in I_r$, such that for $t \in (t_r, t_{r+1})$:*

$$\begin{cases} \dot{x}(t) = \sum_{l \in I_r} \mathrm{z}_l(t) f_l(x(t)) \\ \sum_{l \in I_r} \mathrm{z}_l(t) = 1, \;\; \mathrm{z}_l(t) \geqslant 0, \;\; \textit{for all } l \in I_r \end{cases}. \tag{7.16}$$

The functions $\mathrm{z}_{I_r}(\cdot)$ can be computed as

$$\mathrm{z}_{I_r}(t) = \frac{\hat{z}_{I_r}(t)}{e^{\mathrm{T}} \hat{z}_{I_r}(t)}, \;\; \textit{where } \hat{z}_{I_r}(t) = (M_{I_r,\alpha}(x(t)))^{-1} e \tag{7.17}$$

and α is chosen so that $M_{I_r,\alpha}(x(t)) > 0$.

We notice that under the stated assumptions (7.16) and (7.17) is a smooth ODE.

Example 7.8. Let us continue Example 7.6. For the calculation of z_{I_r} we may take $\alpha = 3$, so that

$$M_{I_r,\alpha}(x,t) = M_{I_r}(x,t) + 3ee^{\mathrm{T}} = \begin{pmatrix} 4-g(t) & 2-g(t) \\ 2+g(t) & 4+g(t) \end{pmatrix}$$

and

$$\hat{z}_{I_r} = (M_{I_r,\alpha}(x,t))^{-1} e = \frac{1}{6} \begin{pmatrix} 1+g(t) \\ 1-g(t) \end{pmatrix}$$

so that

$$\mathrm{z}_{I_r}(x,t) = \frac{1}{2} \begin{pmatrix} 1+g(t) \\ 1-g(t) \end{pmatrix}.$$

One may then compute that

$$\dot{x}(t) = \mathrm{z}_1(x,t)(f_1(x,t))_1 + \mathrm{z}_2(x,t)(f_2(x,t))_1 = -\frac{1}{2}(1-g^2(t)) + \frac{1}{2}(1-g^2(t)) = 0.$$

Consequently $\dot{x}(t) = x(t) = 0$ for $t \geqslant 0$. This shows that the system is in a sliding mode on the "surface" $x = 0$.

7.1.2.3 Calculation of the Active Set

In the previous section, it has been shown that given $x(t_r)$ and I_r, one can expect to be able to calculate the solution $x(\cdot)$ on an interval (t_r, t_{r+1}). The next problem is to find I_{r+1}, given $x(t_{r+1})$. If this can be done, then the solution can be constructed on the whole of $[t_0, t_f]$ by concatenation, since the solution is assumed to be piecewise active. The basic assumption now is that both t_{r+1} and $x(t_{r+1})$ are known. The active set I_{r+1} will be computed thanks to a suitable LCP.

Theorem 7.9. *Let $x(\cdot)$ be a solution of the Filippov's inclusion on a time interval $[t', t'']$ with $I(x(t')) = I_0$ and $I(x(t)) = I$ for all $t \in (t', t'')$. Assume that Assumptions 8 and 9 hold for all $t \in (t', t'')$, and that α is chosen so that $M_{I_0,\alpha}(x(t')) > 0$. Then the LCP*

$$0 \leqslant w = M_{I_0,\alpha}(x(t'))z - e \perp z \geqslant 0 \tag{7.18}$$

has a solution $(\hat{z}, \hat{w})$ such that

$$\{i \mid \hat{z}_i > 0\} \subseteq I \subseteq \{i \mid \hat{w}_i = 0\}. \tag{7.19}$$

Now, let an initial condition $x_0 \in \mathbb{R}^n$ and t' be given, and set $I_0 = I(x_0)$. Choose α so that $M_{I_0,\alpha}(x_0) > 0$. Then if $(\hat{z}, \hat{w})$ is a solution of the LCP in (7.18) such that

$$\{i \mid \hat{z}_i > 0\} = I = \{i \mid \hat{w}_i = 0\} \tag{7.20}$$

and M_{I,x_0} satisfies the conditions of Lemma 7.5, there is a $t'' > t'$ and a solution of the Filippov's inclusion on $[t', t'']$ such that $x(t') = x_0$ and $I(x(t)) = I$ for all $t \in (t', t'')$.

Therefore the central tool of Theorem 7.9 is the LCP in (7.18) that is constructed from the data of the system at time t': $x(t')$ and the active set at t'. Theorem 7.9 says that provided some conditions on $M_{I_0,\alpha}$ are satisfied, there is a solution to the LCP:

the reason is that the imposed condition implies that the matrix is strictly copositive, which is a sufficient condition for a LCP to possess a solution (see Theorem B.2). No uniqueness result is shown, however. The lack of uniqueness reflects that there may not be a unique evolution for the inclusion, as $\dot{x}(t) \in \mathrm{sgn}(x(t))$, $x(0) = 0$, shows. The second part of Theorem 7.9 proves that the active set in the right neighborhood of t' can be deduced from an LCP, and if the LCP has several solutions, there may exist several future active sets. The solution that is assumed to be piecewise active, can consequently be continued, perhaps in a nonunique way, in the right neighborhood of t'. The first part of the theorem shows that the future active set can always be upper and lower bounded using the active set of a particular solution of the LCP.

Example 7.10. Consider $\dot{x}(t) = \mathrm{sgn}(x(t))$, and suppose that $x(t') = 0$. Thus $I_0 = I(x(t')) = \{1,2\}$ (the initial state is on the boundaries of both $R_1 = \{x \mid x > 0\}$ and $R_2 = \{x \mid x < 0\}$) and

$$M_{I_0}(x(t')) = \begin{pmatrix} -1 & 1 \\ & \\ 1 & -1 \end{pmatrix}. \tag{7.21}$$

Let us set $\alpha = 3$. The LCP $0 \leqslant M_{I_0}(x(t'))z - e \perp z \geqslant 0$ has three solutions:

- (i) $\hat{z} = \frac{1}{6}[1,1]^{\mathrm{T}}$, $\hat{w} = [0,0]^{\mathrm{T}}$, $I = \{1,2\}$,
- (ii) $\hat{z} = \frac{1}{2}[1,0]^{\mathrm{T}}$, $\hat{w} = [0,1]^{T}$, $I = \{1\}$,
- (iii) $\hat{z} = \frac{1}{2}[0,1]^{\mathrm{T}}$, $\hat{w} = [1,0]^{\mathrm{T}}$, $I = \{2\}$.

These three solutions show that the system may evolve along three different trajectories: stay on the switching surface, enter R_1, or enter R_2.

7.1.2.4 Determination of the Switching Points and Times

The solution of the inclusion is supposed to be piecewise active, and on an interval (t_r, t_{r+1}) with constant active set I_r, the inclusion is in fact a smooth ODE

$$\dot{x}(t) = \sum_{i \in I_r} \mathrm{z}_i(t) f_i(x(t)), \;\; \mathrm{z}_{I_r}(x) = \frac{\hat{z}_{I_r}(x)}{e^{\mathrm{T}}\hat{z}_{I_r}(x)}, \;\; \hat{z}_{I_r}(x) = (M_{I_r,\alpha}(x))^{-1}e. \tag{7.22}$$

As an example we may choose once again the inclusion $\dot{x}(t) \in -\mathrm{sgn}(x(t))$, $x(0) = 3$. On $[0,3)$ one has $x(t) = -t$ and on $[3,+\infty)$ one has $x(t) = 0$. There are events that indicate a switching point:

- One of the functions $\mathrm{z}_l(t)$, $l \in I$, crosses 0.
- The function $h_j(x) - \min_{i \in I} h_i(x)$ crosses 0 for some $j \neq I$.
- Spontaneous switching points t^* like in Example 7.10.

The spontaneous switchings occur only when the LCP$(M_{I,\alpha}(x(t')), -e)$ has multiple solutions. A switching function is introduced:

$$\psi(I,x,t) = \min[\min_{i\in I} \mathrm{z}_i(t), \min_{j\notin I} h_j(x) - \min_{i\in I} h_i(x), t^* - t]. \tag{7.23}$$

When $\psi(I,x,t) = 0$, one has either one $\mathrm{z}_i(t)$ which crosses 0, or $\min_{j\notin I} h_j(x) - \min_{i\in I} h_i(x)$ which means that the boundary ∂R_j of a set R_j has been attained (and this ∂R_j was not in the active set I), or there is a spontaneous switching t^* at t. Finally since it is not possible to determine exactly 0 on a computer, one introduces an ε-active set as

$$I_\varepsilon(x) = \{i \mid h_i(x) < \min_j h_j(x) + \varepsilon\}. \tag{7.24}$$

Defining an ε-active set corresponds to the usual boundary layer one is obliged to define around the switching surface to make the algorithm implementable on a computer. When I_ε is used in the algorithm, it is possible that the switching function $\psi(I_\varepsilon, x(t), t)$ takes negative values $\geqslant -\varepsilon$. Thus the algorithm may detect a switching point immediately because of this "uncertainty" that is introduced in the switching determination. A possible way to deal with this issue is to detect a switching point only if in addition $\psi(I_\varepsilon, x(t), t)$ is decreasing at the zero crossing. This is similar to detecting an impact in mechanics, not only when the position has attained a certain value, but also when the normal velocity is negative.

Remark 7.11. It is noteworthy that $I_\varepsilon(x)$ is used only to detect the event, and not to integrate the system along the switching surface. This is therefore completely different from the S_ε band of the smoothing method of Sect. 9.3.2. See Sect. 7.1.3 for more explanations.

7.1.2.5 The Algorithm

Roughly speaking the event-driven algorithm uses any ODE solver to integrate during the smooth parts of the motion, and any LCP solver when time comes to compute the solutions of an LCP. Stewart's method comes into play when a switching point has to be determined.

Remark 7.12. The difference with some other integration methods is that Stewart's algorithm steers the integration process with the active set $I(x)$, and the active set is reinitialized using the solution(s) of a suitable LCP. This is clearly where its superiority is. It does not regularize the system, it does not try to maintain the solution on the switching surface by some trick like forcing the discrete solution to lie on S solving a Newton algorithm at each step.

7.1.2.6 Convergence Results

Let us make two assumptions that will be used in the next theorem.

Assumption 10. *The matrix $M_I(x) = [\nabla h_i(x) f_j(x) \mid i,j \in I]$ satisfies the conditions of Lemma 7.5 for all $x \in \mathrm{I\!R}^n$ and $I \subseteq I(x)$ where $I \neq \emptyset$.*

Assumption 11. *The matrix $M_{I(x)}(x) = [\nabla h_i(x) f_j(x) \mid i, j \in I(x)]$ of dimension $k \times k$ is such that the vector $e = (1, 1, .., 1)^T$ cannot be expressed as a linear combination of $k-1$ columns of the matrix $[I - M]$, where I is here the identity matrix of appropriate size.*

The order of accuracy of the smooth ODE $\dot{x}(t) = g(t, x(t))$ solver is denoted as $\omega(h)$. If $x_h(\cdot)$ is the numerical solution and $x(\cdot)$ the analytical solution of the ODE then

$$\begin{cases} ||x_h(t) - x(t)||_\infty \leqslant K_1(t, g)\omega(h) \\ \\ ||\dot{x}_h(t) - \dot{x}(t)||_\infty \leqslant K_2(t, g)\omega(h) \end{cases} \tag{7.25}$$

for all t if $h > 0$ is small enough. The next theorem characterizes the convergence and accuracy of Stewart's event-driven scheme.

Theorem 7.13. *Let Assumptions 10 and 11 hold, and the solution of the Filippov's inclusion be piecewise active. Suppose also that $\varepsilon = \varepsilon(h)$, $\eta = \eta(h)$, and $\varepsilon(h), \eta(h) \to 0$ as $h \to 0$. Let Stewart's algorithm generate a sequence of approximations $x_h(\cdot)$ on $[t_0, t_f]$. Then a limit exists as $h \to 0$ and all the limits are solutions of the Filippov's inclusion. For $h > 0$ sufficiently small, there exists suitable choices of I and t^* in step 5 of the algorithm, independent of h, such that Stewart's algorithm generates numerical approximations $x_h(\cdot)$ where*

$$||x_h(t) - x(t)||_\infty = O(\omega(h)) \tag{7.26}$$

Algorithm 1 Stewart's event-driven method

```
Require: $t_0, T$ interval of integration
Require: $x_0$ initial data
Require: $h > 0$ time–step
  ======== Initialization phase ========
  $k \leftarrow 0$ index of the time step
  $t^\star \leftarrow \infty$
  ======== loop in time ========
  while $t_k < T$ do

    $s_{k+1} \leftarrow t_{k+1} + h$
    $y_{k+1} \leftarrow$ odesolver($t_k, s_{k+1}, x_k, ,$ rhs) with rhs given by ODE (7.16) (7.17)
    $\psi_{k+1} \leftarrow \psi(I, y_{k+1}, s_{k+1})$

    if $\psi_{k+1} < 0$ and $\psi_{k+1} < \psi_k$ then
      Switching event. Call the switching point location Algorithm 2
    else
      Accepted step. Update the state.
      $x_{k+1} \leftarrow y_{k+1}$
      $t_{k+1} \leftarrow s_{k+1}$
      $k \leftarrow k + 1$
    end if
  end while
```

Algorithm 2 Stewart's switching point location method

```
Require: $\eta > 0$ tolerance for locating zeros
Require: $\varepsilon > 0$ tolerance for determining $I_\varepsilon$
Ensure: $t'$ switching time
Ensure: New state $t_{k_1}, x_{k+1}, I, \psi_{k+1}$
  Root finding of $\psi$. Locate an interval
      $[a,b] \subset (t_k, s_{k+1}), |b-a| \leqslant \eta$
  such that
      $\exists \tau \in [a,b] \mid \psi(I, x(\tau), \tau) = 0$
  $t' \leftarrow b$
  $t_{k+1} \leftarrow b$
  $x_{k+1} \leftarrow x(t_{k+1}) = x(t')$
  $I_0 \leftarrow I_\varepsilon(x_{k+1})$
  Update the index set I thanks to Algorithm 3
  $\psi_{k+1} \leftarrow \psi(I, y_{k+1}, s_{k+1})$
```

Algorithm 3 Stewart's active-set updating procedure

```
Require: $I_0$
Ensure: $I$ new index set
Ensure: $t^\star$
  Compute $M_{I_0,\alpha}(x(t'))$
  Compute all the solutions $(\hat{z}_p, \hat{w}_p)$ of $LCP(M_{I_0,\alpha}(x(t')), -e)$ in (7.18)
  $I_p \leftarrow \{i, \mid \hat{z}_p, i > 0\}$ for all $p = 1, 2. \dots$
  if $p == 1$ then
    $I \leftarrow I_p$
  else
    $I \leftarrow I_p$ for a chosen $p$
  end if
  if $LCP(M_{I,\alpha}(x(t')), -e)$ does not have a unique solution then
    Choose $t^\star > t'$
  end if
```

on $[t_0, t_f]$ *provided* $\omega(h) = o(\varepsilon(h))$ *and* $\eta(h) = O(\omega(h))$.

Recall that for two functions $f(\cdot)$ and $g(\cdot)$, $f = O(g)$ at x_0 means that there exists $\beta > 0$ such that $|f(x)| \leqslant \beta |g(x)|$ in a neighborhood of a point x_0. One says that $f = O(g)$ at x_0 if for all $\delta > 0$ there is a neighborhood of x_0 such that $|f(x)| \leqslant \delta |g(x)|$.

7.1.3 Why Is Stewart's Method Superior to Trivial Event-Driven Schemes?

Stewart's method may not be the most intuitive way to simulate Filippov's systems. In order to better understand it, let us work on the simple classical example $\dot{x}(t) \in -\mathrm{sgn}(x(t))$, where $\mathrm{sgn}(\cdot)$ is the set-valued sign function. Suppose $x(0) = 1$. A "naive" event-driven scheme works as follows:

(1) Choose your favorite method, and integrate until the first step k where $x_k < 0$.
(2) Stop the integration, and refine the event detection (i.e., go backwards, recalculate with a shorter time step, and stop this process when the accuracy is considered acceptable by yourself).
(3) Take the new value $\bar{x}_k$ as a starting value and continue to integrate with your favorite method.
(4) Redo step (2) when needed.
With such a method, usually the solution of the discrete inclusion will oscillate around the sliding surface, because the state will go from one side to the other side of the switching surface (here $x = 0$). It is possible to design a wiser procedure at step (2) to minimize these oscillations. But the method intrinsically contains the "oscillating process" around the sliding attractive surface. Moreover the treatment of switching surfaces with high codimension is not easy.

Let us now take a drastically different approach. Instead of driving the switches with the values of $x(\cdot)$, we are going to drive them with the values of a multiplier. As long as the algorithm detects that the switching surface is an attractive sliding surface, then the multiplier is kept to a certain value that corresponds to a given mode (the sliding mode). In such a mode the system is an ODE with an equality constraint, that is a differential algebraic equation (DAE). The reduced-order system is integrated, either by integrating the reduced dynamics explicitly or with the addition of a multiplier associated with the constraint. Even if there is some drift from the switching surface, the system remains in the constrained mode. If the surface becomes a crossing surface at some point, then the multiplier switches to a new value and the trajectory leaves the neighborhood of the switching surface.

A beginning of answer has also been given in Sect. 1.2.3. Consider a mechanical system subject to Coulomb friction. The right way to monitor the switches between the modes (stick and slip motions) is to look at the contact force orientation: is it inside the friction cone, or on the boundary of the friction cone? Remind that the contact force is a Lagrange multiplier. This is quite different from switching the vector field each time a function of the velocity passes through zero. See also Sect. 9.7 for further arguments.

Numerical results are provided in Stewart (1990). Two time-stepping schemes (a multistep algorithm and a θ-method which we will describe later in Chap. 9, with time step $h = 0.005$) are compared to the event-driven scheme. Simple second-order examples with sign functions are tested. The solver used on periods of smooth motion (between the events) is a variable order/variable stepsize multistep method, with added interpolation between previous and current time, and a zero location algorithm. The LCP solver is based on a method described in Al-Khayyal (1987). The main conclusions reported in Stewart (1990) are: a much better accuracy (about 10^4 to 10^5 smaller errors, i.e., about 4 or 5 orders better), about half to third the number of functions evaluations. Moreover decreasing the error tolerance slightly affects the number of functions evaluations, indicating that higher accuracy may be obtained with little additional effort.

7.2 ODEs with Discontinuities with a Transversality Condition

7.2.1 Position of the Problem

Let us consider dynamical systems given by the following ODE with one discontinuity on the hyper-surface $\{x \in \mathbb{R}^n \mid g(x,t) = 0\}$,

$$\dot{x}(t) = f(t,x(t)) = \begin{cases} f^-(t,x(t)) \text{ if } & g(t,x(t)) \leqslant 0 \\ \\ f^+(t,x(t)) \text{ if } & g(t,x(t)) > 0 \end{cases} \tag{7.27}$$

where $f^-(\cdot)$ and $f^+(\cdot)$ are locally Lipschitz in x, and $g: \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}$ is smooth (infinitely differentiable). We assume that the right-hand side f has a discontinuity of order q (see Definition 2.56). If $q = 0$, the vector field $f(\cdot,\cdot)$ made of $f^-(\cdot,\cdot)$ and $f^+(\cdot,\cdot)$ jump at the switching surface and the transversality Assumption 1 has to be satisfied. If $q \geqslant 1$, it may be continuous but with a discontinuous derivative of order $q \geqslant 1$.

When $q \geqslant 1$, the standard ODE theory for existence and uniqueness of solutions applies. Nevertheless, higher order schemes which assume sufficient regularity in the right-hand side f can have some hard troubles. For instance, the order of accuracy of the method is not reached or the condition of stability of the method is called into question. In the case of a discontinuity in f, i.e., an order of discontinuity $q = 0$, standard time-integration scheme can be applied only if the transversality assumption is made. Naturally, the problems that have been evoked for smoother systems are more topical.

Two strategies are implemented to remedy these problems. In Chap. 9 we will review *time-stepping schemes* that apply to such discontinuous systems even with order $q = 0$ and without the transversality condition. Here, we will give some insight on the event-driven that have developed to integrate ODEs with discontinuities.

7.2.2 Event-Driven Schemes

Outline of the Strategy

Without entering into deep details, event-driven strategies for ODEs with discontinuities are based on the standard three-stage method:

1. Perform a time integration of the smooth vector field up to the next nonsmooth event with any standard ODE solver.
2. Locate with a prescribed accuracy the time of the next nonsmooth event.
3. Reinitialize the system at the time of the event if necessary.

Note that as for general event-driven schemes, the nonsmooth events are supposed to be well-separated and finitely countable.

Roughly speaking, the only additional difficulty with respect to the time integration of smooth ODEs is the detection and location of nonsmooth events, if the discontinuities are described by some $g(t,x(t))$ functions as in (7.27). The detection of a nonsmooth event amounts to checking changes of signs of the function g and the accurate location of the event amounts to finding zeroes of the function $g(x(t),t)$. Finding roots of such a function implies that the trajectory is known with a good accuracy. Otherwise a very accurate root detection is a loss of computational time. The nonsmooth event is usually detected by a change of sign of g within a time step. An event is detected if

$$g(x_{k+1},t_{k+1})g(x_k,t_k) \leqslant 0. \tag{7.28}$$

A second test is performed to be sure that a short-lived change of sign has not been omitted by checking if

$$g'(x_{k+1},t_{k+1})g'(x_k,t_k) \leqslant 0. \tag{7.29}$$

If the second test (7.29) is true, the time step is arbitrarily reduced. We will review in the next paragraph methods to find nonsmooth events.

Accurate Locations of Nonsmooth Events

The following family of methods to accurately locate the nonsmooth events can be enumerated:

1. *Brute-force bisection and Newton's method.* Pritsker & Hunt (1973) used a bisection method to accurately locate the event. The integration is repeated with the step size halved. Cellier & Rufer (1975) avoided the bisection by using Newton's method. When this is either not applicable or failed to converge they use a secant method. These two approaches are reliable. Although the latter method is quicker, brute-forces approaches are expensive and many costly evaluations are still required. In Mannshardt (1978), a simplified Newton procedure is used where the Jacobian matrix of g is only evaluated on time. Other techniques such as linear interpolation, "regula falsi", and "Illinois" method are described in Moler (1997). As always, the numerical techniques from the discrete/continuous combined simulation are well suited only for small systems.
2. *Inverse interpolation.* The idea of the inverse interpolation is to use the precomputed values of the approximate solution x_k at time t_k and to interpolate them by a polynomial expression of given order in time in function of x. This method is quite efficient when the root of this polynomial is found inside the interval of study and the order of the interpolation is consistent with the order of consistency of the time-integration scheme. In Ellison (1981), a third-order inverse Hermite interpolation is used in conjunction with a third-order Runge–Kutta method. Usually, the use of external interpolation of the results has to be avoided in favor of local interpolants (see Item 4.).
3. *Augmented ODE systems.* Carver (1978) proposed to augment the initial ODE system (7.27) with the following additional ODE:

$$\begin{cases} z'(t) = (g(x(t),t))' = \nabla_x^{\mathrm{T}} g(x(t),t) f(x(t),t) + \dfrac{\partial g}{\partial t}(x(t),t), \\ z(t_0) = g(x(t_0),t_0). \end{cases} \tag{7.30}$$

The integration of the whole system is performed with any ODE solver with automatic time-step adjustment. The accurate location of the nonsmooth event is performed by using the inherent polynomial interpolation given by the backward differentiation formula (BDF) time-integration method. Originally, Carver (1978) used the standard Hindmarsh–Gear ODE solver (Gear, 1970; Hindmarsh, 1974) with the Nordsieck step size control (Nordsieck, 1962). The polynomial expansion obtained from the approximate gradient of the solution is solved for finding roots.

4. *Local interpolants and error-based detecting.* In Gear & Østerby (1984) the location is based on monitoring the local error of integration in a multi step method (implicit Adams with predictor–corrector (PECE)). If the error is very large within a step, a nonsmooth event is suspected. Assuming a priori that the order of discontinuity is 0 or 1, the event is located in a smaller interval by reducing the step size up to a prescribed tolerance. In Enright et al. (1988), local interpolants of Runge–Kutta methods (Enright et al., 1986) are used to locate the event up to a prescribed tolerance. Indeed, the local interpolants and their associated defects are sampled, and by means of bisection, the event is located. Note that this approach does not require the knowledge of g. In Shampine et al. (1991), the idea of Carver (1978) to add an "event dynamics"as in (7.30) when g is known is used with Runge–Kutta interpolants as in Enright et al. (1988).

Remark 7.14. The first aim of the methods developed in Mannshardt (1978) (simplified Newton's method), in Gear & Østerby (1984) (error monitoring), and in Enright et al. (1988) (defect and local interpolants) is not really to locate accurately the time events. They can be used for this purpose but their first goal is to locate some events inside a time step of a prescribed length. Assuming that the error made by the time integration is at least of order 0 inside this time step and the number of events is finite, the authors chose a time-step size, and therefore the accuracy of event location with respect to the order of consistency of the method. This strategy allows them to keep the global order of accuracy of the methods. For all these reasons, we have preferred to detail these methods in Sect. 9.1 that is devoted to time-stepping methods for ODE with discontinuities.

8

Event-Driven Schemes for Lagrangian Systems

The dynamics of nonsmooth Lagrangian systems has been presented in Chap. 3, to which the reader is referred for details (see also Sect. 2.7). In this chapter, we assume that the well-posedness assumptions of Sect. 3.7 hold whenever the corresponding systems are examined. Obviously the specific features described in Chap. 6 may occur. In this chapter, we will present an event-driven scheme for Lagrangian dynamical systems in a very simplified way. Indeed, in full generality, the equations that govern such systems are quite complicated. Especially, we will assume that the considered Lagrangian system is only subjected to perfect unilateral constraints with the Newton impact rules. More details on the frictional case with Poisson impact law can be found in Pfeiffer & Glocker (1996), Glocker (2001), Abadie (1998, 2000).

8.1 Introduction

Let us briefly recall a fundamental feature of mechanical systems subjected to complementarity conditions that make them clearly depart from the switched and impulsive systems of Sects. 2.9 and 2.10. We suppose that there is a single unilateral constraint $g(q) \geqslant 0$, and we disregard the impacts. The system is

$$\begin{cases} M(q(t))\ddot{q}(t) + F(q(t),\dot{q}(t),t) = \nabla g(q)\lambda \\ \\ 0 \leqslant \lambda \perp g(q) \geqslant 0\,. \end{cases} \tag{8.1}$$

One may consider that there are two modes: either the system is not in contact ($g(q) > 0$) or it is in contact ($g(q) = 0$). Denoting the time intervals for each mode as $\mathscr{I}_i^{\mathrm{nc}}$ and $\mathscr{I}_i^{\mathrm{c}}$, respectively, and assuming that the system is well-posed with $\mathbb{R}^+ = \cup_{i \geqslant 0}(\mathscr{I}_i^{\mathrm{nc}} \cup \mathscr{I}_i^{\mathrm{c}})$, one obtains (we drop the time argument)

$$M(q)\ddot{q} + F(q,\dot{q},t) + \nabla g(q)(\nabla g^{\mathrm{T}}(q)M^{-1}(q)\nabla g(q))^{-1}G(q,\dot{q},t) = 0 \text{ for all } t \in \mathscr{I}_i^{\mathrm{c}} \tag{8.2}$$

and

$$M(q)\ddot{q} + F(q,\dot{q},t) = 0 \text{ for all } t \in \mathscr{I}_i^{\mathrm{nc}}\,. \tag{8.3}$$

The dynamics in (8.2) comes from solving the LCP: $0 \leqslant \lambda \perp \nabla g^{\mathrm{T}}(q)M^{-1}(q)\nabla g(q)\lambda + G(q,\dot{q},t) \geqslant 0$, where $G(q,\dot{q},t) \geqslant 0$ collects nonlinear terms that come from differentiating twice $g(q(t))$. From (8.2) and (8.3) one may be tempted to consider that the Lagrangian system (8.1) is a piecewise smooth or a switched system as those of Sect. 2.9. Notice, however, that (8.2) and (8.3) have been written under the assumption that the solutions possess a certain form: this is an a priori assumption on the dynamics, which has to be proved to hold true. Secondly, the problem of determining the conditions that make the dynamics switch from (8.2) to (8.3) and vice versa no longer appears in the formalism (8.2) and (8.3). Something is missing. Obviously the intervals $\mathscr{I}_i^{\mathrm{nc}}$ and $\mathscr{I}_i^{\mathrm{c}}$ are not exogenous but state dependent. This "something" is nothing else but the complementarity relations, which monitor the detachment and persistent contact conditions. One cannot get rid of the multiplier λ when treating mechanical systems, even if the velocities are assumed to be continuous. Finally the ODE (8.2) hides the fact that on the intervals $\mathscr{I}_i^{\mathrm{c}}$, the constraint is active, i.e., $g(q) = 0$. The dynamics is therefore that of a DAE, and it can be rewritten after a suitable change of coordinate as a reduced-order ODE with additional algebraic conditions linking the multiplier to q, $\dot{q}$, and $\ddot{q}$, see for instance McClamroch & Wang (1988). Complementarity systems live on lower-dimensional subspaces, that is not the case of switched, piecewise something, or impulsive systems. Once again ignoring the complementarity conditions is impossible. This is true for mechanical systems and for LCS. See more information on this point in Heemels & Brogliato (2003) and Brogliato (2003) and in Sect. 8.6.1.

Let us end this introduction by embedding (8.1) into an inclusion as in (3.25). One may rewrite (8.2) and (8.3) as

$$\begin{aligned} & M(q)\ddot{q} + F(q,\dot{q},t) + \nabla g(q)(\nabla g^{\mathrm{T}}(q)M^{-1}(q)\nabla g(q))^{-1}G(q,\dot{q},t) = 0 \\ & \text{if } g(q) = 0 \quad \text{and} \quad \nabla g^{\mathrm{T}}(q)\dot{q} \leqslant 0 \\ & \text{or if } g(q) = \nabla g^{\mathrm{T}}(q)\dot{q} = 0 \quad \text{and} \quad \nabla g^{\mathrm{T}}(q)\ddot{q} + \tfrac{\mathrm{d}}{\mathrm{d}t}(\nabla g^{\mathrm{T}}(q))\dot{q} \leqslant 0 \end{aligned} \tag{8.4}$$

and

$$M(q)\ddot{q} + F(q,\dot{q},t) = 0 \text{ if } (g(q), \nabla g^{\mathrm{T}}(q)\dot{q}, \nabla g^{\mathrm{T}}(q)\ddot{q} + \frac{\mathrm{d}}{\mathrm{d}t}(\nabla g^{\mathrm{T}}(q))\dot{q}) \succ 0, \tag{8.5}$$

where $\frac{\mathrm{d}^2}{\mathrm{d}t^2}(g(q(t))) = \nabla g^{\mathrm{T}}(q)\ddot{q} + \frac{\mathrm{d}}{\mathrm{d}t}(\nabla g^{\mathrm{T}}(q))\dot{q}$. Rigorously the quantities in (8.4) and (8.5) are estimated at their right limits. Doing this assumption, it is possible to use the results in van der Schaft & Schumacher (1998) to assert that the trajectories of (3.25) and of (8.4) and (8.5) are the same on intervals with no velocity jumps. The system then looks like a switched system, though the domain appearing in (8.4) and defined with a lexicographical inequality, is neither closed nor open. The disadvantage of (8.4) and (8.5) is that: (1) there is no contact force; (2) if the number of constraints m is large, the formalism becomes quite cumbersome to write down since there are 2^m modes. As a consequence further studies may be rendered extremely difficult because the compactness of the inclusion or of the complementarity formalisms is lost. In particular, we will see in this chapter that solving LCPs is a very convenient way to integrate the motion with event-driven algorithms.

In conclusion, complementarity conditions cannot be dispensed with, even in event-driven methods.

8.2 The Smooth Dynamics and the Impact Equations

The Impact Equations

The impact equations (3.123) can be written at the time t_i of velocity discontinuities in the following algebraic way:

$$M(q(t_i))(v^+(t_i) - v^-(t_i)) = p_i \; . \tag{8.6}$$

This equation will be solved at the time of impact together with an impact law. For a Newton impact law one obtains

$$\begin{cases} M(q(t_i))(v^+(t_i) - v^-(t_i)) = p_i \\ U_{\mathrm{N}}^+(t_i) = \nabla g^{\mathrm{T}}(q(t_i))v^+(t_i) \\ U_{\mathrm{N}}^-(t_i) = \nabla g^{\mathrm{T}}(q(t_i))v^-(t_i) \\ p_i = \nabla g(q(t_i))P_{\mathrm{N},i} \\ 0 \leqslant U_{\mathrm{N}}^+(t_i) + eU_{\mathrm{N}}^-(t_i) \perp P_{\mathrm{N},i} \geqslant 0 \; , \end{cases} \tag{8.7}$$

where U_{N} is defined in Sect. 3.3 and $g(\cdot)$ is the gap function. This problem can be reduced to the local unknowns $U_{\mathrm{N}}^+(t_i)$ and $P_{\mathrm{N},i}$ if the matrix $M(q(t_i))$ is assumed to be invertible. One obtains the following LCP at time t_i of discontinuities of $v(\cdot)$:

$$\begin{cases} U_{\mathrm{N}}^+(t_i) = \nabla g^{\mathrm{T}}(q(t_i))(M(q(t_i)))^{-1}\nabla g(q(t_i))P_{\mathrm{N},i} + U_{\mathrm{N}}^-(t_i) \\ 0 \leqslant U_{\mathrm{N}}^+(t_i) + eU_{\mathrm{N}}^-(t_i) \perp P_{\mathrm{N},i} \geqslant 0 \end{cases} \tag{8.8}$$

from which $P_{\mathrm{N},i}$ may be computed.

The Smooth Dynamics

The smooth dynamics which is valid almost everywhere for the Lebesgue measure dt (dt – a.e.) is governed by (3.124):

$$M(q(t))\gamma^+ + F_{\text{int}}(t,q(t),v^+(t)) = F_{\text{ext}}(t) + f^+(t) \quad (dt - \text{a.e.})\,, \tag{8.9}$$

where we assume that $f^+(\cdot) = f^-(\cdot) = f(\cdot)$ $(dt - \text{a.e.})$. The following smooth system is then to be solved $(dt - \text{a.e.})$:

$$\begin{cases} M(q(t))\gamma^+(t) + F_{\text{int}}(t,q(t),v^+(t)) = F_{\text{ext}}(t) + f^+(t) \\ \\ f^+(t) = \nabla g(q(t))F^+(t) \\ \\ 0 \leqslant g(q(t)) \perp F^+(t) \geqslant 0\,. \end{cases} \tag{8.10}$$

In order to solve this system at each time t, i.e., to know the configuration after each event and to integrate it numerically, it is useful to express the complementarity laws at different kinematics levels. This is done in the following section.

8.3 Reformulations of the Unilateral Constraints at Different Kinematics Levels

8.3.1 At the Position Level

Let us consider the complementarity conditions $0 \leqslant g(q(t)) \perp F^+(t) \geqslant 0$. From (A.9) this is equivalently rewritten as the inclusion (dropping the time argument)

$$-F^+ \in N_K(g(q)) \tag{8.11}$$

with $K = \mathbb{R}^+$.

8.3.2 At the Velocity Level

The gap function $t \mapsto g(q(t))$ can be differentiated with respect to time as follows in the Lagrangian setting:

$$\begin{cases} \dot{g}(q(t^+)) = U_{\text{N}}^+(t) = \nabla g^{\text{T}}(q(t))v^+(t) \\ \\ \ddot{g}(q(t^+)) = \dot{U}_{\text{N}}^+(t) = \Gamma_{\text{N}}(t^+) = \nabla g^{\text{T}}(q(t))\gamma^+(t) + \frac{\text{d}}{\text{d}t}(\nabla g^{\text{T}}(q(t)))v^+(t)\,. \end{cases} \tag{8.12}$$

The complementarity condition $0 \leqslant g(q(t)) \perp F^+(t) \geqslant 0$ must be written now at different kinematic levels, i.e., in terms of the right velocity $U_{\text{N}}^+(\cdot)$ and in terms of the right accelerations $\Gamma_{\text{N}}^+(\cdot)$.

Assuming that $U_{\text{N}}^+(\cdot)$ is right-continuous by definition of the right limit of a BV function, the complementarity condition implies, in terms of the velocity, the following relation:

$$-F^+ \in \begin{cases} 0 & \text{if } g(q) > 0 \\ 0 & \text{if } g(q) = 0,\ U_{\mathrm{N}}^+ > 0 \\ (-\infty, 0] & \text{if } g(q) = 0,\ U_{\mathrm{N}}^+ = 0\,. \end{cases} \tag{8.13}$$

A rigorous proof of this assertion can be found in Glocker (2001). This relation is the representation of the complementarity condition at the velocity level, which can be written more compactly as

$$-F^+ \in \begin{cases} 0 & \text{if } g(q) > 0 \\ N_{\mathbb{R}^+}(U_{\mathrm{N}}^+) & \text{if } g(q) = 0 \end{cases} \tag{8.14}$$

and using the notation of the tangent cone, we can write (8.14) as

$$-F^+ \in N_{T_{\mathbb{R}^+}(g(q))}(U_{\mathrm{N}}^+)\,. \tag{8.15}$$

In a complementarity formalism, this relation can be written as

$$\begin{cases} \text{If } g(q) = 0 \text{ then } 0 \leqslant U_{\mathrm{N}}^+ \perp F^+ \geqslant 0 \\ \text{If } g(q) > 0 \text{ then } F^+ = 0\,. \end{cases} \tag{8.16}$$

8.3.3 At the Acceleration Level

In the same way, the complementarity condition can be written at the acceleration level as follows:

$$-F^+ \in \begin{cases} 0 & \text{if } g(q) > 0 \\ 0 & \text{if } g(q) = 0, U_{\mathrm{N}}^+ > 0 \\ 0 & \text{if } g(q) = 0, U_{\mathrm{N}}^+ = 0, \Gamma_{\mathrm{N}}^+ > 0 \\ (-\infty, 0] & \text{if } g(q) = 0, U_{\mathrm{N}}^+ = 0, \Gamma_{\mathrm{N}}^+ = 0\,. \end{cases} \tag{8.17}$$

A rigorous proof of this assertion can be found in Glocker (2001). As before, the equation can be written more compactly as

$$-F^+ \in \begin{cases} 0 & \text{if } g(q) > 0 \\ 0 & \text{if } g(q) = 0, U_{\mathrm{N}}^+ > 0 \\ N_{\mathbb{R}^+}(\Gamma_{\mathrm{N}}^+) & \text{if } g(q) = 0, U_{\mathrm{N}}^+ = 0\,, \end{cases} \tag{8.18}$$

which is to be compared to (8.14), or with the tangent cone notation, we obtain Glocker's inclusion, where time dependency is dropped:

$$-F^+ \in N_{T_{T_{\mathrm{I\!R}^+}(g(q))}(U_{\mathrm{N}}^+)}(\Gamma_n^+)\,, \tag{8.19}$$

which is to be compared to (8.15). Finally, in the complementarity formalism we can write

$$\begin{cases} 0 \leqslant \Gamma_{\mathrm{N}}^+ \perp F^+ \geqslant 0 & \text{if } g(q)=0 \text{ and } U_{\mathrm{N}}^+ = 0 \\ \\ F^+ = 0 & \text{otherwise}\,. \end{cases} \tag{8.20}$$

We can see that the right measurable force F^+ is nonnull if the contact is active ($g(q)=0$) and the right velocity vanishes ($U_{\mathrm{N}}^+ = 0$). This result seems to be very reasonable from the mechanical point of view.

Remark 8.1. It is apparent from the foregoing developments that ($K = \mathrm{I\!R}^+$)

$$N_K(g(q)) \supset N_{T_{\mathrm{I\!R}^+}(g(q))}(U_{\mathrm{N}}^+) \supset N_{T_{T_{\mathrm{I\!R}^+}(g(q))}(U_{\mathrm{N}}^+)}(\Gamma_n^+)\,. \tag{8.21}$$

Also one may define further normal cones by continuing the differentiation. This is what has been done for the higher order sweeping process in Chap. 5. In fact, for a system with relative degree r, it suffices to consider the cones up to the derivative of order $r-1$. For Lagrangian systems, $r=2$ and Moreau's sweeping process considers the velocity level. This is sufficient to get a complete formulation of the dynamics, which allows one to integrate the system. This is also sufficient to design a sound time-stepping scheme.

8.3.4 The Smooth Dynamics

The system (3.124) that we have to solve for the smooth dynamics can be written at the acceleration level as follows:

$$\begin{cases} M(q(t))\gamma^+(t) + F_{\mathrm{int}}(t,q,v^+) = F_{\mathrm{ext}}(t) + f^+(t) \\ \\ \Gamma_{\mathrm{N}}(t^+) = \nabla g^{\mathrm{T}}(q(t))\gamma^+(t) + \frac{\mathrm{d}}{\mathrm{d}t}(\nabla g^{\mathrm{T}}(q(t)))v^+(t) \\ \\ f^+(t) = \nabla g(q(t))F^+(t) \\ \\ -F^+(t) \in N_{T_{T_{\mathrm{I\!R}^+}(g(q(t)))}(U_{\mathrm{N}}^+(t))}(\Gamma_n(t^+))\,. \end{cases} \tag{8.22}$$

When the condition, $g(q)=0$ and $U_{\mathrm{N}}^+ = 0$ is satisfied, we obtain the following LCP:

$$\begin{cases} M(q(t))\gamma^+(t) + F_{\mathrm{int}}(t,q,v^+) = F_{\mathrm{ext}}(t) + \nabla g(q(t))F^+(t) \\ \\ \Gamma_{\mathrm{N}}^+(t) = \nabla g^{\mathrm{T}}(q)\gamma^+(t) + \frac{\mathrm{d}}{\mathrm{d}t}(\nabla g^{\mathrm{T}}(q))v^+(t) \\ \\ 0 \leqslant \Gamma_{\mathrm{N}}^+(t) \perp F^+(t) \geqslant 0\,, \end{cases} \tag{8.23}$$

which can be reduced on variable Γ_{N}^{+} and F^{+}, if $M(q(t))$ is invertible, as

$$\begin{aligned}\Gamma_{\mathrm{N}}^{+} = \nabla g^{\mathrm{T}}(q)M^{-1}(q(t))(-F_{\mathrm{int}}(t,q,v^{+})+F_{\mathrm{ext}}(t)) + \tfrac{\mathrm{d}}{\mathrm{d}t}(\nabla g^{\mathrm{T}}(q))v^{+} \\ +\nabla g^{\mathrm{T}}(q)M^{-1}(q)\nabla g(q(t))F^{+}(t) \\ 0 \leqslant \Gamma_{\mathrm{N}}^{+}(t) \perp F^{+}(t) \geqslant 0\,.\end{aligned} \tag{8.24}$$

The LCP in (8.24) has $F^{+}(t)$ as its unknown. When the system evolves in a mode without any modification of the set of active constraints, then the dynamics possesses the same degree of smoothness as the external forces (if those forces are smooth, the solutions are smooth as well). The LCP in (8.24) is then used to determine the right limit of the contact force. If this right limit is positive, contact is kept (i.e., the active set does not change on the right of t). If it is zero, the active set may change. Detachment is checked with the sign of the right acceleration whose expression is in (8.24). Obviously it is supposed in (8.24) that the constraints $g(q)$ are active, i.e., $g(q)=0$, and in addition that $U_{\mathrm{N}}^{+}=0$.

8.4 The Case of a Single Contact

Let us suppose that an approximation of the right state and of the associated local variables are known at t_k, i.e.,

$$(q_k, v_k, \gamma_k) \approx (q(t_k), v^{+}(t_k), \gamma^{+}(t_k)) \tag{8.25}$$

$$(g_k, U_{\mathrm{N},k}, \Gamma_{\mathrm{N},k}) \approx (g(t_k), U_{\mathrm{N}}^{+}(t_k), \Gamma_{\mathrm{N}}^{+}(t_k)) \tag{8.26}$$

Two smooth dynamics may be integrated between t_k and t_{k+1}:

1. *The constraint is not active.* If the constraint is not active, the following system is integrated with $F^{+}=0$:

$$M(q(t))\gamma^{+}(t) + F_{\mathrm{int}}(t,q,v) = F_{\mathrm{ext}}(t)\,. \tag{8.27}$$

   In this case, we associate to this step an integer, $\mathrm{status}_k = 0$.
2. *The constraint is active.* The smooth system (8.23) is numerically integrated with the bilateral constraint $\Gamma_{\mathrm{N}}^{+}=0$, i.e.,

$$\begin{bmatrix} M(q(t)) & -\nabla g(q(t)) \\ \nabla g(q(t)) & 0 \end{bmatrix} \begin{bmatrix} \gamma^{+}(t) \\ F^{+}(t) \end{bmatrix} = \begin{bmatrix} -F_{\mathrm{int}}(t,q(t),v(t^{+})) + F_{\mathrm{ext}}(t) \\ \frac{\mathrm{d}}{\mathrm{d}t}(\nabla g^{\mathrm{T}}(q(t)))v^{+}(t) \end{bmatrix}. \tag{8.28}$$

   In this case, we associate to this step an integer, $\mathrm{status}_k = 1$.

At the end of the time step, the following procedure with the decision tree may be applied:

- **Case 1**: $\mathrm{status}_k = 0$. Integrate the system (8.27) on the time interval $[t_k, t_{k+1}]$:
  - **Case 1.1**: $g_{k+1} > 0$
    The constraint is still not active. We set $\mathrm{status}_{k+1} = 0$.

- **Case 1.2**: $g_{k+1} = 0, U_{\mathrm{N},k+1} < 0$
  In this case an impact occurs. The value $U_{\mathrm{N},k+1} < 0$ is considered as the pre-impact velocity U_{N}^- and the impact equation (8.8) is solved. After, we set $U_{\mathrm{N},k+1} = U_{\mathrm{N}}^+$. Two cases are then possible:
  - **Case 1.2.1**: $U_{\mathrm{N}}^+ > 0$
    Just after the impact, the relative velocity is positive. The constraint ceases to be active and we set $\text{status}_{k+1} = 0$.
  - **Case 1.2.2**: $U_{\mathrm{N}}^+ = 0$
    The relative post-impact velocity vanishes. In this case, in order to determine the new status, we solve the LCP (8.24). Three cases are then possible:
    - **Case 1.2.2.1**: $\Gamma_{\mathrm{N},k+1} > 0, F_{k+1} = 0$
      The constraint is still not active. We set $\text{status}_{k+1} = 0$.
    - **Case 1.2.2.2**: $\Gamma_{\mathrm{N},k+1} = 0, F_{k+1} > 0$
      The constraint has to be activated. We set $\text{status}_{k+1} = 1$.
    - **Case 1.2.2.3**: $\Gamma_{\mathrm{N},k+1} = 0, F_{k+1} = 0$
      This case is undetermined. We need to know the value of $\dot{\Gamma}_{\mathrm{N}}^+ = \lim_{\varepsilon \to 0, \varepsilon > 0} \frac{\Gamma_{\mathrm{N}}^+(t+\varepsilon) - \Gamma_{\mathrm{N}}^+(t)}{\varepsilon}$.
- **Case 1.3**: $g_{k+1} = 0, U_{\mathrm{N},k+1} = 0$
  In this case, we have a grazing constraint. To know what the status should be for the future time, we compute the value of $\Gamma_{\mathrm{N},k+1}, F_{k+1}$ thanks to the LCP (8.24) assuming that $U_{\mathrm{N}}^+ = U_{\mathrm{N}}^- = U_{\mathrm{N},k+1}$. Three cases are then possible:
  - **Case 1.3.1**: $\Gamma_{\mathrm{N},k+1} > 0, F_{k+1} = 0$
    The constraint is still not active. We set $\text{status}_{k+1} = 0$.
  - **Case 1.3.2**: $\Gamma_{\mathrm{N},k+1} = 0, F_{k+1} > 0$
    The constraint has to be activated. We set $\text{status}_{k+1} = 1$.
  - **Case 1.3.3**: $\Gamma_{\mathrm{N},k+1} = 0, F_{k+1} = 0$
    This case is undetermined. We need to know the value of $\dot{\Gamma}_{\mathrm{N}}^+$ in solving an LCP of higher order.
- **Case 1.4**: $g_{k+1} = 0, U_{\mathrm{N},k+1} > 0$
  The activation of the constraint has not been detected. We seek for the first time t_* such that $g = 0$. We set $t_{k+1} = t_*$. Then we perform all of this procedure keeping $\text{status}_k = 0$.
- **Case 1.5**: $g_{k+1} < 0$
  The activation of the constraint has not been detected. We seek for the first time t_* such that $g = 0$. We set $t_{k+1} = t_*$. Then we perform all of this procedure keeping $\text{status}_k = 0$.

- **Case 2**: $\text{status}_k = 1$ Integrate the system (8.28) on the time interval $[t_k, t_{k+1}]$
  - **Case 2.1**: $g_{k+1} \neq 0$ or $U_{\mathrm{N},k+1} = 0$
    Something is wrong in the time integration or the drift from the constraints is too large.
  - **Case 2.2**: $g_{k+1} = 0, U_{\mathrm{N},k+1} = 0$
    In this case, we assume that $U_{\mathrm{N}}^+ = U_{\mathrm{N}}^- = U_{\mathrm{N},k+1}$ and we compute $\Gamma_{\mathrm{N},k+1}, F_{k+1}$ thanks to the LCP (8.24). Three cases are then possible:

- **Case 2.2.1**: $\Gamma_{\mathrm{N},k+1} = 0, F_{k+1} > 0$
  The constraint is still active. We set $\text{status}_{k+1} = 1$.
- **Case 2.2.2**: $\Gamma_{\mathrm{N},k+1} > 0, F_{k+1} = 0$
  The bilateral constraint is no longer valid. We seek for the time t_* such that $F^+ = 0$. We set $t_{k+1} = t_*$ and we perform the integration up to this instant. We perform all of this procedure at this new time t_{k+1}.
- **Case 2.2.3**: $\Gamma_{\mathrm{N},k+1} = 0, F_{k+1} = 0$
  This case is undetermined. We need to know the value of $\dot{\Gamma}_{\mathrm{N}}^+$.

The complementarity conditions of **Case 1.3.3** are constructed as follows:

$$-\dot{F}^+ \in \begin{cases} 0 & \text{if } g(q) > 0 \\ 0 & \text{if } g(q) = 0, U_{\mathrm{N}}^+ > 0 \\ 0 & \text{if } g(q) = 0, U_{\mathrm{N}}^+ = 0, \Gamma_{\mathrm{N}}^+ > 0 \\ 0 & \text{if } g(q) = 0, U_{\mathrm{N}}^+ = 0, \Gamma_{\mathrm{N}}^+ = 0, \dot{\Gamma}_{\mathrm{N}}^+ > 0 \\ (-\infty, 0] & \text{if } g(q) = 0, U_{\mathrm{N}}^+ = 0, \Gamma_{\mathrm{N}}^+ = 0, \dot{\Gamma}_{\mathrm{N}}^+ = 0\,, \end{cases} \tag{8.29}$$

which may be written more compactly as the inclusion

$$-\dot{F}^+ \in N_{T_{T_{T_{\mathrm{I\!R}^+}(g(q))}(U_{\mathrm{N}}^+)}(\Gamma_{\mathrm{N}}^+)}(\dot{\Gamma}_{\mathrm{N}}^+).$$

One may then differentiate the dynamics so as to get an equality similar to (8.24), for $\dot{\Gamma}_{\mathrm{N}}^+$. The LCP is obtained from $0 \leqslant \dot{\Gamma}_{\mathrm{N}}^+ \perp \dot{F}^+ \geqslant 0$. It is clear that this relies on the assumption that the data are sufficiently regular, so that the differentiation can be performed. Both inclusions are written under the assumption of **Case 1.3.3** that says $F^+ = 0$ at the considered instant. They are consequently closely related to the lexicographical complementarity conditions as introduced in (8.5).

8.4.1 Comments

In practical situations, all the tests are made up to an accuracy threshold. All statements of the type $g(q) = 0$ are replaced by $|g(q)| < \varepsilon$. The role played by these epsilons can be very important and they are quite difficult to size.

If the ODE solver is able to perform the root finding of the function $g(q) = 0$ for $\text{status}_k = 0$ and $F^+ = 0$ for $\text{status}_k = 1$ within the integration algorithm, the cases 1.4, 1.5 and 2.2.2 can be suppressed in the decision tree. If the drift from the constraints is controlled into the ODE solver by an error computation, the case 2.1 can also be suppressed. With these assumptions, it is possible to rearrange the decision tree in Algorithm 4 below.

Algorithm 4 Event -driven procedure on a single time-step with one contact

```
Require: $t_k, g_k, U_{N,k}, status_k$, values at the beginning of the time step
Ensure: $t_{k+1}, g_{k+1}, U_{N,k+1}, status_{k+1}$, values at the end of the time step

  //======= Computation of the provisional values of $(g_{k+1}, U_{N,k+1})$
  if $status_k = 0$ then
     $(g_{k+1}, U_{N,k+1}) \leftarrow$ time–integration of (8.27) up to an event
  end if
  if $status_k = 1$ then
     $(g_{k+1}, U_{N,k+1}) \leftarrow$ time–integration of (8.28) up to an event
  end if

  //======= Reinitialization and update of the index sets
  //    The constraint is still not active. (case 1.1)
  if $g_{k+1} > 0$ then
     $status_{k+1} \leftarrow 0$
  end if
  //    The constraint is active $g_{k+1} = 0$ and an impact occur $U_{N,k+1} < 0$ (case 1.2)
  if $g_{k+1} = 0, U_{N,k+1} < 0$ then
     $U_N^- \leftarrow U_{N,k+1}$
     Solve the LCP (8.8)
     $U_{N,k+1} \leftarrow U_N^+$
     if $U_{N,k+1} > 0$ then
        $status_{k+1} \leftarrow 0$
     end if
  end if
  //    The constraint is active $g_{k+1} = 0$ without impact (case 1.2.2, case 1.3, case 2.2)
  if $g_{k+1} = 0, U_{N,k+1} = 0$ then
     solve the LCP (8.24)
     if $\Gamma_{N,k+1} = 0, F_{k+1} > 0$ then
        $status_{k+1} \leftarrow 1$
     else if $\Gamma_{N,k+1} > 0, F_{k+1} = 0$ then
        $status_{k+1} \leftarrow 0$
     else if $\Gamma_{N,k+1} = 0, F_{k+1} = 0$ then
        //Undetermined case.
     end if
  end if
```

Remark 8.2. When $g_{k+1} = 0$ and $U_{N,k+1} = 0$, the LCP at the acceleration level (8.24) is solved. In the one-contact case, it is possible through a naive approach to choose to deactivate the constraint just by analyzing the result of the integration of the smooth system with a bilateral constraint after the event. If $F_{k+1} < 0$, the constraint may be suppressed and the contact ceases to be active. The procedure, which is simpler than the LCP solving, is no longer valid in the multi-contact case. Indeed, some famous examples such as the Delassus example (Pérès, 1953) or the two blocks example

(Pfeiffer & Glocker, 1996) show that the contacts that have to be suppressed are not necessarily the contacts for which the force is negative after a bilateral integration. We will come later on another interest of the LCP when the constraints are not linearly dependent and when we want to simplify the algorithm 4.

8.5 The Multi-contact Case and the Index Sets

8.5.1 Index Sets

Index sets have already been introduced in Sect. 7.1.2 when we dealt with Stewart's event-driven method for Filippov's systems. Here the goal is the same: determine which constraints are active in order to construct the LCP in (8.24). In the multi-contact case, i.e., the case with $\nu > 1$ constraints, Algorithm 4 is extended by the introduction of the following index sets. The index set I is the set of all unilateral constraints in the system:

$$I = \{1, \ldots, \nu\} \subset \mathbb{N} \,. \tag{8.30}$$

The index set I_{c} is the set of all active constraints of the system,

$$I_{\mathrm{c}} = \{\alpha \in I \mid g^{\alpha} = 0\} \subseteq I \,, \tag{8.31}$$

and the index set I_{s} is the set of all active constraints of the system with a relative velocity equal to zero,

$$I_{\mathrm{s}} = \{\alpha \in I_{\mathrm{c}} \mid U_{\mathrm{N}}^{\alpha} = 0\} \subseteq I_{\mathrm{c}} \,. \tag{8.32}$$

We can notice that thanks to these index sets the conditional statements in Algorithm 4 can be generalized to the multi-contact case. In the same manner we can make the impact equation (8.8) and the smooth dynamics (8.23) precise by

$$\begin{cases} M(q(t_i))(v^+(t_i) - v^-(t_i)) = p_i \\[2ex] U_{\mathrm{N}}^+(t_i) = \nabla g^{\mathrm{T}}(q(t_i))v^+(t_i) \\[2ex] U_{\mathrm{N}}^-(t_i) = \nabla g^{\mathrm{T}}(q(t_i))v^-(t_i) \\[2ex] p_i = \nabla g(q(t_i))P_{\mathrm{N},i} \\[2ex] P_{\mathrm{N},i}^{\alpha} = 0; U_{\mathrm{N}}^{\alpha,+}(t_i) = U_{\mathrm{N}}^{\alpha,-}(t_i), \quad \forall \alpha \in I \setminus I_{\mathrm{c}} \\[2ex] 0 \leqslant U_{\mathrm{N}}^{+,\alpha}(t_i) + eU_{\mathrm{N}}^{-,\alpha}(t_i) \perp P_{\mathrm{N},i}^{\alpha} \geqslant 0, \quad \forall \alpha \in I_{\mathrm{c}} \,. \end{cases} \tag{8.33}$$

Using the fact that $P_{\mathrm{N},i}^{\alpha} = 0$ for all $\alpha \in I \setminus I_{\mathrm{c}}$, this problem can be reduced on the local unknowns $U_{\mathrm{N}}^+(t_i), P_{\mathrm{N},i}$ for all $\alpha \in I_{\mathrm{c}}$. For the smooth dynamics (8.23), we suppose that $I_{\mathrm{c}} \setminus I_{\mathrm{s}} = \emptyset$. We obtain

$$\begin{cases} M(q(t))\gamma^{+}(t)+F_{\mathrm{int}}(t,q(t),v(t^{+}))=F_{\mathrm{ext}}(t)+\nabla g(q(t))F^{+}(t) \\ \Gamma_{\mathrm{N}}^{+}(t)=\nabla g^{\mathrm{T}}(q(t))\gamma^{+}(t)+\frac{\mathrm{d}}{\mathrm{d}t}(\nabla g^{T}(q(t)))v^{+}(t) \\ F^{+,\alpha}(t)=0, \quad \forall\alpha\in I\setminus I_{\mathrm{s}} \\ 0\leqslant \Gamma_{\mathrm{N}}^{+,\alpha}(t)\perp F^{+,\alpha}(t)\geqslant 0, \quad \forall\alpha\in I_{\mathrm{s}}\,. \end{cases} \tag{8.34}$$

Finally, we rewrite the bilateral smooth dynamics as

$$\begin{cases} M(q(t))\gamma^{+}(t)+F_{\mathrm{int}}(t,q(t),v(t^{+}))=F_{\mathrm{ext}}(t)+\nabla g(q(t))F^{+}(t) \\ \Gamma_{\mathrm{N}}^{+}(t)=\nabla g^{\mathrm{T}}(q(t))\gamma^{+}(t)+\frac{\mathrm{d}}{\mathrm{d}t}(\nabla g^{\mathrm{T}}(q(t)))v^{+}(t) \\ F^{+,\alpha}(t)=0, \quad \forall\alpha\in I\setminus I_{\mathrm{s}} \\ \Gamma_{\mathrm{N}}^{+,\alpha}(t)=0, \quad \forall\alpha\in I_{\mathrm{s}}\,. \end{cases} \tag{8.35}$$

This bilateral dynamics is integrated up to an event given by the root finding of the following function:

$$\begin{cases} g^{\alpha}(t)=0, \quad \forall\alpha\in I\setminus I_{\mathrm{s}} \\ F^{+,\alpha}(t)=0, \quad \forall\alpha\in I_{\mathrm{s}}\,. \end{cases} \tag{8.36}$$

Algorithm 5 is then an extension of the single-contact case.

8.6 Comments and Extensions

8.6.1 Event-Driven Algorithms and Switching Diagrams

A switching diagram consists of nodes that represent various modes of the dynamics and arrows linking the nodes that represent the conditions of switching between the nodes. Strictly speaking it is possible to rewrite Algorithms 4 and 5 with such diagrams. The main issue is that the number of nodes may quickly become quite large: a mechanical system with 30 frictionless unilateral contacts has 2^{30} modes. It is impossible to draw a diagram with such a number of nodes! However, the LCP that corresponds to this mechanical system has 30 unknowns and is therefore of size 30: it is easy to solve numerically. In other words, the switching diagram is the enumerative way of solving the LCP. Usually, switching diagrams are used when the trajectory that is to be simulated is a priori known, so that the number of nodes and arrows becomes low (see, e.g., the simulation of the woodpecker toy in Soellner &

Algorithm 5 Event -driven procedure on a single time-step with several contacts

```
Require: $t_k, g_k, U_{N,k}, I_{c,k}, I_{s,k}$ values at the beginning of the time step
Ensure: $t_{k+1}, g_{k+1}, U_{N,k+1}, I_{c,k+1}, I_{s,k+1}$ values at the end of the time step

  //======= Computation of the provisional values of
  //                    $(g_{k+1}, U_{N,k+1})$ and $I_{c,k+1}, I_{s,k+1}$
  $(g_{k+1}, U_{N,k+1}) \leftarrow$ time–integration of (8.35) according to $I_{c,k}$ and $I_{s,k}$ up to an event
  Update the index sets $I_{c,k+1}$ and $I_{s,k+1}$

  //======= Reinitialization and update of the index sets
  //   Impacts occur.
  if $I_{c,k+1} \setminus I_{s,k+1} \neq \emptyset$ then
    Solve the LCP (8.33).
    Update the index-set $I_{c,k+1}$ and provisional $I_{s,k+1}$
    Check that $I_{c,k+1} \setminus I_{s,k+1} = \emptyset$
  end if
  if $I_{s,k+1} \neq \emptyset$ then
    Solve the LCP (8.34)
    for $\alpha \in I_{s,k+1}$ do
      if $\Gamma_{N,\alpha,k+1} > 0, F_{\alpha,k+1} = 0$ then
        remove $\alpha$ from $I_{s,k+1}$ and $I_{c,k+1}$
      else if $\Gamma_{N,\alpha,k+1} = 0, F_{\alpha,k+1} = 0$ then
        Undetermined case.
      end if
    end for
  end if
  // Go to the next time step
```

Führer 1998, Sect. 6.3.5). Clearly such an assumption is not reasonable in most situations and one may say that switching diagrams are of little use for the simulation of nonsmooth multibody systems, and more generally for complementarity dynamical systems.

8.6.2 Coulomb's Friction and Enhanced Set-Valued Force Laws

The case of the Coulomb's friction can theoretically be treated in the same way. Two index sets I_r and I_t are added. The set I_r is the set of sticking or rolling contacts:

$$I_\mathrm{r} = \{\alpha \in I_\mathrm{s} \mid U_\mathrm{N}^\alpha = 0, \|U_\mathrm{T}\| = 0\} \subseteq I_\mathrm{s}\,, \tag{8.37}$$

and

$$I_\mathrm{t} = \{\alpha \in I_\mathrm{s} \mid U_\mathrm{N}^\alpha = 0, \|U_\mathrm{T}\| > 0\} \subseteq I_\mathrm{s} \tag{8.38}$$

is the set of slipping or sliding contact. Together with these new index sets, new events have to be checked corresponding to transitions from sticking to slipping and vice versa. Checking the transitions is not an easy task especially in the three-dimensional case. For more enhanced set-valued forces laws, we refer to Glocker (2001).

8.6.3 Bilateral or Unilateral Dynamics?

Between two events, the smooth dynamics with bilateral constraints (8.28) is solved. The main reason for this choice is that we assumed that the problem with bilateral constraints is easier and cheaper to integrate. This is not always the case. Another way is to continue to integrate the unilateral dynamics (8.34) assuming that this dynamics is smooth between two events and smooth if the event at the end of the interval is a constraint deactivation.

There are two main reasons to do like this:

1. The first reason is the detection of the events of the type: $F^{+,\alpha} = 0$ or "$F^{+,\alpha}$ vanishing". This type of events necessitates to stop the integration, solve the LCP (8.34) to see if a detachment will occur. If the integration is done directly with this LCP, an event of the type: $\Gamma_{\mathrm{N}}^{+,\alpha} > 0$ guarantees that the detachment occurs. If we assume that the dynamics is smooth after a detachment, an event of the type: $g^{\alpha}(q(t+\varepsilon)) > 0$, $\varepsilon > 0$ can be triggered. In the other case, we can continue to integrate the system without any modifications in the index sets.
2. The second reason is that the forces $F^{+} = 0$ are in usual cases not uniquely defined. On the contrary, the accelerations are defined in a unique way (see the theorem due to J.J. Moreau in Brogliato, 1999, theorem 5.4). The main reason is that the constraints are not linearly independent in most practical situations. If we integrate the dynamics with bilateral constraints, we need to use pseudo-inverses to obtain the acceleration or to compress the bilateral constraints. These operations are expensive from the numerical point of view. The LCP solvers that are used to solve (8.34) are in most cases able to find the acceleration in a unique way even if the constraints are not linearly independent. This point will be detailed in the following chapter.
   Finally, in the bilateral case, it is possible to detect $F^{+,\alpha} = 0$ while there is no detachment due the indeterminacy of the contact forces. This fact is related to the Delassus example. If the unilateral dynamics is solved, we seek for an event $\Gamma_{\mathrm{N}}^{+,\alpha} > 0$ or $g^{\alpha} > 0$. This problem is suppressed.

8.6.4 Event-Driven Schemes: Lötstedt's Algorithm

Event-driven algorithms have been proposed in Tzitzouris & Pang (2002), Fetecau et al. (2003), and Lötstedt (1984). Though they are sometimes presented as time-stepping schemes, they may in fact be classified as event-driven schemes, because they involve some impact time and state detection procedures. As a consequence the order may be strictly larger than 1 (for instance an order 2 is reported in Fetecau et al., 2003), which is impossible in time-stepping methods where no specific algorithm is implemented to accurately detect the impact times and states. Increasing the order when detection procedures are implemented is not surprising in view of the results of Sect. 8.6.5. One may also find interesting applications in Pfeiffer et al. (2006) like roller coasters and drop tower hydraulics, where comparisons between experimental results and numerical results are shown.

Let us describe the scheme of Lötstedt (1984), which may historically be the first attempt to construct a dedicated event-driven scheme for complementarity Lagrangian systems with and without friction. Though Lötstdedt's algorithm may be of little interest now due to the many works and improvements that have been performed since its publication, it keeps an historical value and deserves some place in such a monograph.

The Frictionless Case

Let us consider first the frictionless case. The following numerical scheme is proposed to compute the state at step k:

$$\begin{cases} q_k = \frac{1}{\alpha_0^1} b_k^1 \\ v_k = \frac{1}{\alpha_0^2}[h\beta_0^2 M^{-1}(Q_k + \nabla g_k \lambda_k) + b_k^2] \\ 0 \leqslant \nabla g_k^{\mathrm{T}} v_k \perp \lambda_k \geqslant 0 \, , \end{cases} \tag{8.39}$$

where

$$\begin{cases} b_k^1 = h\sum_{i=0}^r \beta_i^1 v_{k-i} - \sum_{i=1}^r \alpha_i^1 q_{k-i} \\ b_k^2 = h\sum_{i=1}^r \beta_i^2 M^{-1}(Q_{k-i} + \nabla g_{k-i}\lambda_{k-i}) - \sum_{i=1}^r \alpha_i^2 v_{k-i} \, . \end{cases} \tag{8.40}$$

The complementarity relations in (8.39) correspond to the active constraints at step k. They encompass the persistent contact as well as plastic impacts phases. The formulas in (8.39) and (8.40) correspond to two linear r-step methods. The notation f_k stands for $f(q_k)$. The coefficients α_i^1 and β_i^1 are determined from an Adams–Bashforth family of explicit formulas, see, e.g., page 250 in Garcia de Jalon & Bayo (1994), denoted as AB-r. The coefficients $\beta_0^2 = 1$ and $\beta_i^2 = 0$ for all $i = 1, 2, ..., r$. The second equation in (8.39) is a backward difference formula, denoted BDF-r. Notice that the mass matrix M is assumed to be constant (hence the Coriolis and centrifugal torques are zero), which restricts the application to simple mechanical systems with Euclidean configuration space (like collections of particles). It is, however, argued that this is just a matter of convenience to allow for an easy factorization of M, and that the extension towards $M(q)$ is possible. The torque $Q_k = Q(q_k, v_k, t_k)$ therefore contains gravity, viscous friction, and external actions (like control inputs). The integration step is chosen constant, equal to $h > 0$. When $Q = Q(t, q)$, it is shown in Lötstedt (1984) that an LCP whose unknown is λ_k can be formulated from (8.39). This LCP can be rephrased as a quadratic program:

$$\min_{\lambda_k \geqslant 0} \frac{1}{2} \lambda_k^{\mathrm{T}} \nabla g_k^{\mathrm{T}} M^{-1} \nabla g_k \lambda_k + h^{-1} \nabla g_k^{\mathrm{T}} [b_k^2 + hM^{-1} Q_k] \, . \tag{8.41}$$

Consequently, the set of equations in (8.39) allows one to advance the solution in time from $k-1$ to k. The methods AB-1 (forward Euler)-BDF-1 and AB-2-BDF-2 are chosen, where it is recalled that it is useless to use methods of order > 3 (linear multistep A-stable methods have an accuracy of order < 2, i.e., at most $O(h^2)$, see

Garcia de Jalon & Bayo, 1994, pp. 250–251). After discontinuities in v_k or $\dot{v}_k$ (which are detected from the value of the impulse on one step with a threshold under which it is considered to be zero), the AB-1-BDF-1 algorithm is used during two steps to restart the simulation (it is known that multistep methods are not self-starting and require the help of a single-step algorithm initially).

When $Q = Q(q,v,t)$, then the LCP formulation is lost. However, Lötstedt proves that provided the matrix

$$A(v) = M - \frac{h}{\alpha_0^2}\frac{\partial Q}{\partial v}(q,v,t) \tag{8.42}$$

is full rank and $\nabla g_k^{\mathrm{T}} A^{-1} \nabla g_k$ is positive definite,[1] then (8.39) still possesses a unique solution so that the algorithm can be used to safely advance the solution in time. However, this time λ_k is generally the solution of a NCP (a quick look at the second equation in (8.39) allows one to realize this). The condition in (8.42) can be used with the implicit function theorem to express $v_k = f_k(\lambda_k)$ for some function $f_k(\cdot)$. The second condition is used to prove the existence of a solution to the NCP. A way to solve the NCP is proposed, based on functional iteration. In summary, Lötstedt algorithm is given as follows:

- Compute q_k using AB-1 or AB-2, with h such that the local error in q_k is smaller than $h\varepsilon$ for a prescribed tolerance ε (ways to estimate such a h are provided).
- Calculate ∇g_k to a prescribed accuracy and calculate $\dot{v}_k = M^{-1}(Q_k + \nabla g_k)$ and v_k by BDF-1 or BDF-2.
- Test whether velocities and accelerations are discontinuous between t_{k-1} and t_k, due to either an impact (detected from a nonzero value of the impulse) or the activation of a new constraint ($g^\alpha(t_{k-1}) > 0$ and $g^\alpha(t_k) \leqslant 0$ for some α) or the deactivation of a constraint. The time of such jumps is calculated by inverse linear interpolation. After a shock a new velocity v_{k+1} is computed by a collision rule. Then restart the algorithm at the first step with AB-1 and the new set of active constraints.
- Test the detachment conditions by checking whether one entry of the vector λ_k passes through zero, and whether the corresponding entry in the normal velocity $\nabla g_k^{\mathrm{T}} v_k$ is positive. Then refresh the set of inactive constraints if needed.
- End.

Remark 8.3. Lötstedt also shows that the LCP(λ_k) can be reformulated as the minimization problem

$$\min \ ||\sum_{i=0}^{r} \alpha_i^2 v_{k-i} - hM^{-1}Q_k||_M, \ \ \nabla g_k^{\mathrm{T}} v_k \geqslant 0\,. \tag{8.43}$$

Dissipativity of (8.39)–(8.41) plus the impact rule and convergence of the algorithm are not proved.

[1] This is a kind of iteration matrix.

Constraints with 2D Friction

The same algorithms AB-1-BDF-1 or AB-2-BDF-2 are used as in the frictionless case. When 2D friction is incorporated in the algorithm, one has to add the tangential contribution of the contact force in the right-hand side of the second equation in (8.39). The contact force is split into two parts as: $[G(q)+H(q,v)]\lambda$. Roughly $G(q)\lambda$ contains the normal generalized force and the contribution of the sticking contacts, whereas $H(q,v)\lambda$ accounts for the sliding contacts. The vector λ contains the normal multipliers $\lambda_{n,j}$ and the tangential ones $\lambda_{t,j}$ for the jth contact. There are two features in the algorithm. The first one is the approximation of λ_k, the second one is the calculation of the impulses at the shock instants. Let us denote the jth component of λ by λ^j and its kth iteration by λ_k^j. Then the approximated value is $\bar{\lambda}_k^j = \lambda_{k-1}^j + h_k \frac{\lambda_{k-1}^j - \lambda_{k-2}^j}{h_{k-1}}$, for a variable step of integration h_k. A QP is constructed that allows the computation of the term $G(q)\lambda$. It possesses the advantage of assuring that the tangential force is opposite the tangential acceleration. But it has the strong drawback that sliding generally implies the QP matrix to be nonsymmetric, rendering the problem harder to solve. It is clear that the introduction of $\bar{\lambda}_k^j$ in the dynamical equations modifies the subsequent calculations in a nonphysical manner right after the first step and should be avoided. Special procedures are also used after a shock and a discontinuity in the acceleration. The error introduced in v_k by the use of $\bar{\lambda}_k^j$ in a permanent contact phase are shown to be $\mathrm{O}(h^3)$ when $h_k = h > 0$, a constant. They are $\mathrm{O}(h)$ after a reinitialization of the velocity or of the acceleration.

The second point (calculation of the impulse at a shock instant) is formulated as follows. Taking frictional effects at impacts into account, let us denote the right-hand side of the impact algebraic equation as $P_j = G(q(t_j)\Lambda_j$, where Λ_j is a vector of normal and tangential percussions and t_j is an impact time. Then Lötstedt proposes to calculate the impulse from the QP:

$$\left.\begin{array}{l} \min\ \frac{1}{2}[v(t_j^+) - v(t_j^-)]^{\mathrm{T}} M [v(t_j^+) - v(t_j^-)] \\ \\ W^{\mathrm{T}} v = G^{\mathrm{T}} v(t_j^+),\ v \geqslant 0,\ v^{\mathrm{T}} W \Lambda_j = \Lambda_j^{\mathrm{T}} G^{\mathrm{T}} v(t_j^+) = 0 \\ \\ W = \begin{pmatrix} I & 0 \\ f_{\mathrm{imp}} I & -I \\ f_{\mathrm{imp}} I & I \end{pmatrix} \end{array}\right. , \tag{8.44}$$

where I is the identity matrix with dimension equal to the number of active constraints and f_{imp} can be considered as an impulse ratio (Brach, 1990). The main problem with the calculation in (8.44) is that although it looks like a plastic impact rule, it is not: there may be rebounds. In addition, if there is a tangential velocity reversal during the shock (i.e., the post and pre-impact tangential velocities have opposite signs), then there may be a kinetic energy gain at the shock instant.

Remark 8.4. The algorithm in Tzitzouris & Pang (2002) is close in spirit to Lötstedt scheme (time-stepping with accurate detection of contacting times). It uses a trapezoidal discretization of the continuous frictionless dynamics (implicit one-step scheme, solved by a Newton method with an initial guess from a Euler discretization) and an adaptive step size procedure. Several simple examples show that h may decrease to very small values as 10^{-12} s during the simulation. Lemke algorithm is used to solve the contact force LCP and the impact percussion LCP.

8.6.5 Consistency and Order of Event-Driven Algorithms

8.6.5.1 Some Classical Definitions and Results

Let us first recall several basic definitions that apply to a numerical method of the form

$$x_{k+1} = x_k + h_k \phi(t_k, x_k, h_k) \text{ for all } k \geqslant 0,\ t_0 = 0,\ t_{k+1} = t_k + h_k\,. \tag{8.45}$$

Definition 8.5. *The numerical method in (8.45) is said to be consistent for the ODE $\dot{x}(t) = f(x(t),t)$, $x(0) = x_0$, if for any solution of this ODE the consistency error*

$$\sum_{k=0}^{N-1} ||x(t_{k+1}) - x(t_k) - h_k \phi(t_k, x(t_k), h_k)|| \tag{8.46}$$

tends to 0 when $h = \max_{0 \leqslant k \leqslant N} h_k$ tends to 0.

Definition 8.6. *The numerical method in (8.45) is said to be stable if there exists a constant M, not depending on h_k, such that for all sequences $\{x_k\}_{0 \leqslant k \leqslant N}$, $\{z_k\}_{0 \leqslant k \leqslant N}$, and $\{\varepsilon_k\}_{0 \leqslant k \leqslant N}$ verifying*

$$\begin{cases} x_{k+1} = x_k + h_k \phi(t_k, x_k, h_k) \\ \\ z_{k+1} = z_k + h_k \phi(t_k, z_k, h_k) + \varepsilon_k\,, \end{cases} \tag{8.47}$$

one has

$$\max_{0 \leqslant k \leqslant N} ||z_k - x_k|| \leqslant M \left(||x_0 - z_0|| + \sum_{k<N} ||\varepsilon_k|| \right) . \tag{8.48}$$

Definition 8.7. *The numerical method in (8.45) is said to be convergent if*

$$\lim_{h \to 0} \max_{0 \leqslant k \leqslant N} ||x(t_k) - x_k|| = 0\,. \tag{8.49}$$

Theorem 8.8. *If the numerical method in (8.45) is consistent and stable, then it is convergent.*

The next result concerns one-step methods.

Definition 8.9. *The numerical method in (8.45) is said to be of order $p > 0$ if there exists a constant K, depending only on $x(\cdot)$ and $\phi(\cdot)$, such that*

$$\sum_{k=0}^{N-1} ||x(t_{k+1}) - x(t_k) - h_k \phi(t_k, x(t_k), h_k)|| \leqslant K\, h^p \tag{8.50}$$

for any solution $x(\cdot)$ of the ODE $\dot{x}(t) = f(x(t),t)$, $x(0) = x_0$ such that $x(\cdot) \in C^{p+1}([0,T], \mathrm{I\!R}^n)$.

Then the following result holds:

Theorem 8.10. *If the numerical method in (8.45) is stable and of order p, and of* $f(\cdot,\cdot) \in C^p([0,T] \times \mathrm{IR}, \mathrm{IR}^n)$*, one has*

$$||x(t_k) - x_k|| \leqslant MK\, h^p, \ \ \forall\, k \leqslant N\,. \tag{8.51}$$

8.6.5.2 Application to the Event-Driven Methods

The next results are taken from Janin & Lamarque (2001). They concern a one degree-of-freedom mechanical system of the form

$$\begin{cases} \ddot{x}(t) + 2a\dot{x}(t) + \omega_1^2 x(t) = f(t) + \lambda \\ 0 \leqslant \lambda \perp -x(t) + x_{\max} \geqslant 0 \\ \dot{x}(t^+) = -e\dot{x}(t^-) \text{ if } x(t) = x_{\max} \text{ and } \dot{x}(t^-) < 0 \end{cases} \tag{8.52}$$

with $f(\cdot) \in \mathscr{L}^1_{\mathrm{loc}}(\mathrm{IR}^+)$, $e \in [0,1]$. Notice that Definition 8.9 and Theorem 8.10 apply to systems with a certain degree of smoothness. However, as shown in Janin & Lamarque (2001) they can be adapted to the case of piecewise $C^{p+1}([0,T], \mathrm{IR}^n)$ solutions and piecewise $C^p([0,T] \times \mathrm{IR}, \mathrm{IR}^n)$ vector fields. Before stating the consistency and order results of the event-driven methods, it is necessary to say few words on the way the impact times are approximated, because this is going to have a strong influence on the overall scheme order.

Definition 8.11. *Let an impact be detected at step* $k+1$*, i.e.,* $x_k < x_{\max}$ *and* $x_{k+1} > x_{\max}$*. Let us apply a linear interpolation of the solution that maps* t_k *to* x_k *and* t_{k+1} *to* x_{k+1} *and infer which time is mapped to* $x_{\max}$*. This is called the (IM1) approximation.*

Assume that the scheme provides an estimate v_k of the derivative (the velocity) at step k.

Definition 8.12. *Let us apply a second-order interpolation of the solution that maps* t_k *to* x_k *and* t_{k+1} *to* x_{k+1} *and whose derivative at* t_k *is* v_k *and infer which time is mapped to* $x_{\max}$*. This is called the (IM2) approximation.*

Definition 8.13. *Let us apply a dichotomy procedure between* t_k *and* t_{k+1} *to the approximation* $x_s(t)$ *of the displacement provided by the scheme applied on* $[t_k, t]$*, until reaching a time* $t^* \in (t_k, t_{k+1})$ *such that* $x_s(t^*)$ *is close to* $x_{\max}$*. The iterative procedure is stopped when the difference between two consecutive times falls under a chosen precision, that is set to* h^4 *for the Runge–Kutta RK24 method. This is called the (IM3) approximation.*

One drawback of the (IM3) method is that the computer precision may rapidly be reached when h is decreased. We then have the next results that couple the order of the method between events and the precision of the approximation of the impact times, for the system (8.52).

Once an approximation t^* of the impact time has been calculated, one proceeds as follows:

- Integrate the trajectory on $[t_k, t^*)$ with the chosen method,[2] to get an approximation of the pre-impact velocity v_*^-.
- Calculate the post-impact velocity $v_*^+ = -ev_*^-$.
- Integrate the trajectory from $(x_{\max}, v_*^+)$ on $(t^*, t_{k+1}]$ with the chosen method, to obtain (x_{k+1}, v_{k+1}).

It is possible that another impact has occurred on $(t^*, t_{k+1}]$. Then one may choose to apply another detection procedure on this interval. We shall then make a difference between event-driven methods with at most one impact detection per time step and those with multiple impact detection per time step.

In such event-driven schemes it is mandatory to also implement a sticking mode detection (which we called a bilateral constraint in the foregoing sections). One tests whether $x_{k+1} > x_{\max}$. If it does, then it is considered that the system is in a sticking mode at t_{k+1} and one sets x_{k+1} to $x_{\max}$ and v_{k+1} to 0. If sticking is occurring at t_k, then one only checks the sign of the acceleration to determine the subsequent motion. If the acceleration is negative, it is apparent from (8.52) that sticking persists. If the acceleration is positive, then two cases have to be considered. If there is one impact detection per step, then the above procedure on impact detection may be applied to compute the value (x_{k+1}, v_{k+1}). If one uses a multiple impact detection procedure per step, then the algorithm approximates impact times between t_k and t_{k+1} in order to approximate the time $\bar{t}$ when sticking ends. This time satisfies $f(\bar{t}) = \omega_1^2 x_{\max}$. Then integrate on $[\bar{t}, t_{k+1}]$ with the chosen method.

Remark 8.14. It becomes clear that an event-driven method is not simple to implement. Several different procedures have to be implemented together. The sticking mode detection that is described here seems to work because the considered system is simple. In more complex cases this may become impossible. Other event detection procedures exist, see, e.g., Turner (2001) for a noniterative procedure based on a time-scale of the continuous dynamics, where the new time-variable is the inequality constraint entering the model. The woodpecker problem is tested with a Runge–Kutta fixed-step method. It is shown that $N = 2000$ integration steps are needed to detect all the events on a period of 1 s (i.e., when $N > 2000$, the number of detected events no longer increases).

The Case of Finite Number of Impacts

Let us consider that there is at most one impact detection at each time step. The next theorem concerns trajectories with a finite number of impacts on the interval of integration $[0, T]$.

[2] Adapting the time step.

Theorem 8.15. *(Event-driven with finite number of impacts)*

(i) [Order 2] Let us consider a one-step convergent numerical method, at least of order 2 for any sufficiently smooth ODE, to which we associate the approximation method (IM1), in order to obtain an approximation of the solution of the oscillator in (8.52), with $f(\cdot)$ differentiable and with a bounded derivative on $\mathbb{R}^+$. Then the resulting event-driven scheme is consistent and of order 2.

(ii) [Order 3] Let us consider a one-step convergent numerical method, at least of order 3 for any sufficiently smooth ODE, to which we associate the approximation method (IM2), in order to obtain an approximation of the solution of the oscillator in (8.52), with $f(\cdot)$ twice differentiable and with a bounded second derivative on $\mathbb{R}^+$. Then the resulting event-driven scheme is consistent and of order 3.

(iii) [Order 4] Let us consider a one-step convergent numerical method, at least of order 4 for any sufficiently smooth ODE, to which we associate the approximation method (IM3), in order to obtain an approximation of the solution of the oscillator in (8.52), with $f(\cdot)$ three-times differentiable and with a bounded third derivative on $\mathbb{R}^+$. Then the resulting event-driven scheme is consistent and of order 4.

Since it is a common feature of mechanical systems to possess accumulations of impacts (and possibly also of stick-slip transitions when Coulomb friction is present), Theorem 8.15 is restricted to so-called vibro-impact systems, whose solutions are piecewise continuous with separated velocity jumps. In Janin & Lamarque (2001) an extension is proposed that is presented next.

The Case with One Accumulation of Impacts

Now we consider systems which may have an accumulation of impacts on $[0,T]$, followed by a sticking mode. Still only methods with at most one impact detection at each time step are considered.

Proposition 8.16. *Let us consider a convergent one-step method of order $p \geqslant 2$ for any smooth ODE, to which we associate (IM1). Suppose that $f(\cdot)$ is differentiable with locally Lipschitz derivative. Let us also assume that the trajectory of the system in (8.52) has one accumulation of impacts on $[0,T]$. Let $h = \frac{T}{N}$, (x_k^N, v_k^N) be the approximation of the solution $(x(\cdot), \dot{x}(\cdot))$ given by the numerical method at t_k, and $\varepsilon_k = (\varepsilon_k^1, \varepsilon_k^2)$ the consistency error between t_k and t_{k+1} defined by $\varepsilon_k^1 = |x(t_k) - x_k|$ and $\varepsilon_k^2 = |\dot{x}(t_k) - v_k|$. Then we have*

(i) If there is no impact on (t_k, t_{k+1}), then $\varepsilon_k^1 \leqslant c_1 h^3$ and $\varepsilon_k^2 \leqslant c_2 h^3$,
(ii) if there is at least one impact on (t_k, t_{k+1}), then $\varepsilon_k^1 \leqslant c_3 h^2$ and $\varepsilon_k^2 \leqslant c_4 h$,
(iii) if t_k lies in a sticking phase, then $\varepsilon_k^1 \leqslant c_5 h^3$ and $\varepsilon_k^2 \leqslant c_6 h^2$,
where c_i, $1 \leqslant i \leqslant 6$ are constants.

Remark 8.17. The foregoing order and consistency results can be extended to event-driven schemes with a multiple impact localization per step. This, however, does not increase the orders.

Numerical Experiments

The theoretical results have been confirmed by numerical tests in Janin & Lamarque (2001). Three methods are used in the tests: a Newmark, a Runge–Kutta RK24, and a Dormand–Price Runge–Kutta DOPRI5. The system is excited by a periodic function $f(\cdot)$ and two sets of parameters are chosen so that one system has periodic trajectories with separated impacts and the other one has trajectories with one accumulation of impacts.

- When the Newmark method is used, the order of the event-driven scheme is insensitive to the localization method (IM1), (IM2), or (IM3). For relatively high values of h, the velocity is no longer accurately approximated. For trajectories with finite number of impacts, the order is 2. For systems with one accumulation of impacts, the order is 1 for the velocity and 2 for the displacement.
- For the RK24 method, the estimated order of the event-driven scheme is 1 with (IM1), 2 with (IM2), and 3 with (IM3), as predicted by the theoretical results. For trajectories with one impact accumulation, the impact approximation method does not affect the order of the event-driven scheme. An accurate impact time approximation is not useful as the order is always 2 for the displacement and 1 for the velocity. It is also noticed that the accurate detection of times when sticking ends is interesting only when RK24 is used with (IM2) or (IM3).
- Applying a multiple impact localization procedure with Newmark's method does not improve the performance of the event-driven scheme. It is thus useless to try to detect as many impacts as possible in this case.
- On the contrary, multiple impact localizations improve the event-driven schemes with the RK24 method. Applying iteratively (IM2) yields an order 3 for velocity and displacement. Applying iteratively (IM3) yields an order 4 for velocity and displacement.
- The DOPRI5 method has order 5 when applied to smooth systems. When coupled with an iterative (IM3) multiple impact localization procedure at each step, it provides an event-driven scheme of order 4 in displacement and velocity.

The computational times are reported in Janin & Lamarque (2001). It follows that the RK24 method with (IM2), iterative (IM2), and (IM3), and the DOPRI5 method with the iterative (IM3) are the fastest numerical schemes for the considered system.

8.7 Linear Complementarity Systems

Though this chapter is dedicated to mechanical systems, let us make an aside to LCSs. Let us consider an LCS as in (2.95) or (4.3). It follows from the material of Chap. 5 that without any restriction on the relative degree between λ and y, the solutions may be distributions. Nevertheless, let one assume that

- The solutions are regular in the sense of Definition 5.8, with $\sigma \geqslant \sigma_{\min} > 0$ for some $\sigma_{\min}$. In other words the solutions are piecewise analytic.

- If needed, a jump rule has been defined so that the system is well-posed (the HOSP formalism of Chap. 5 provides such a rule, see also (4.9)–(4.11) and Heemels et al., 2000 for other jump rules).

Then an event-driven algorithm can be implemented in a way similar to the above event-driven schemes for mechanical systems. In particular it is necessary to monitor the index sets for active and inactive constraints.

8.8 Some Results

Event-driven strategies have been applied successfully in various application cases. They usually concern systems with few degrees of freedom, but several unilateral contacts with friction. The simulation of circuit breakers for virtual prototyping is described in Abadie (2000). Many applications may be found in Pfeiffer & Glocker (1996) and Leine et al. (2003): hammering in gears, gear rattling in gearboxes, ship-turning gear, turbine blade damper, friction clutch vibrations, the woodpecker toy, drilling machines, landing gear dynamics, assembly processes, a tumbling toy. The influence of Painlevé paradoxes on the dynamics and the bifurcations is analyzed in Leine et al. (2002).

9

Time-Stepping Schemes for Systems with AC Solutions

9.1 ODEs with Discontinuities

In this section, some specific features of the numerical time integration of ODEs with discontinuities introduced in Sect. 2.8 are reviewed. If the order of discontinuity q is equal to 0, i.e., the right-hand side of the ODE f possesses a jump, then the transversality Assumption 1 has to be made to guarantee the existence of standard solution to the ODE. If $q \geqslant 1$, standard assumptions such as Lipschitz continuity of the right-hand side ensures the application of the standard theory of ODEs.

As we saw in Chap. 1 and in the discussion in Sect. 7.2, any standard one-step and multistep methods can be applied to approximate such ODE systems with discontinuities. Nevertheless, some care has to be taken about the efficiency, the local order of consistency, the global order of accuracy, and stability results. In this section, we will first illustrate these points on some numerical experiments on the nonsmooth circuits introduced in Chap. 1. In a second stage, we will give some of the remedies that can be applied to retrieve the order of accuracy of higher order Runge–Kutta method.

For a thorough description of one-step and multi-step ODE solvers which are termed by their acronyms (RK32, DOPR154, ...), we refer to the following monographs (Hairer et al., 1993; Hairer & Wanner, 1996). Precise definitions of standard notions for ODE solvers such as convergence, order of convergence, order of consistency can also be found in the above cited references. They will not be recalled.

9.1.1 Numerical Illustrations of Expected Troubles

Order of Convergence

In Calvo et al. (2003), the authors observe *that when the differential equation is sufficiently smooth the local error estimate for a Runge–Kutta method behaves as a certain positive power* $\mathrm{O}(h^p)$ *while in the presence of a discontinuity of order* $q < p$, *the local error estimates behave as* $\mathrm{O}(h^{(q+1)})$. Similar remarks can be found in Hairer et al. (1993). In Gear & Østerby (1984), the authors remark that the control of the step size usually based on the a priori knowledge of the order of consistency of the method can fail and lead to a large amount of unsuccessful steps.

To the best of the authors' knowledge, there is neither a proof of such results nor a proof of the order of a Runge–Kutta method of order p for a differential equation with discontinuities of order $q < p$. A proof of convergence of one-step and multistep methods exists for a class of differential inclusions with absolutely continuous solutions whose derivatives may jump (see Sect. 9.2). They can be applied to this case for a discontinuity of order 0. The order result of Theorem 9.16 is also given and states that the methods are of global order 1, but a crucial point is that the set of instants when the derivative of the absolutely continuous solution has jumps has to be a finite set.

Mannshardt (1978) proposed the first serious work on the global order of accuracy of one-step methods for ODEs with discontinuous right-hand side and transversality conditions. He notes that "Almost every Runge–Kutta method remains convergent after a transition but only with order 1 (the order of consistency decreases to 0 during a transition)". We will present in the next section how he succeeded to overcome this difficulty.

To complete this paragraph, we present some numerical illustrations of the defects in the order of convergence when some discontinuities are present in the right-hand side. Let us consider the circuits (**a**) and (**b**) depicted in Fig. 1.3. The ODE systems satisfied by the circuit equations are, respectively,

$$(\mathbf{a}) \quad \begin{cases} \dot{x}_1(t) = x_2(t) - \dfrac{1}{RC}x_1(t) - \dfrac{\lambda(t)}{R} \\[2ex] \dot{x}_2(t) = -\dfrac{1}{LC}x_1(t) - \dfrac{\lambda(t)}{L} \\[2ex] 0 \leqslant \lambda(t) \perp w(t) = \dfrac{\lambda(t)}{R} + \dfrac{1}{RC}x_1(t) - x_2(t) \geqslant 0, \end{cases} \tag{9.1}$$

$$(\mathbf{b}) \quad \begin{cases} \dot{x}_1(t) = -x_2(t) + \lambda(t) \\[2ex] \dot{x}_2(t) = \dfrac{1}{LC}x_1(t) \\[2ex] 0 \leqslant \lambda(t) \perp w(t) = \dfrac{1}{C}x_1(t) + R\lambda(t) \geqslant 0. \end{cases} \tag{9.2}$$

Let us recall that the equations of the circuit (**a**) and (**b**) yield ODEs with a continuous right-hand side, i.e., the order of discontinuity is at least $q \geqslant 1$.

In Fig. 9.1, the global error is plotted with respect to the time step. The simulation which is performed is the time integration on $t \in [0,5]$ of the circuit (**a**) with the initial conditions $x_1(0) = 1, x_2(0) = -1$, and the data $R = 10, L = 1, C = \frac{1}{(2\pi)^2}$. In this case, the diode is assumed to be always on. The results are given for four methods: explicit Euler, embedded pairs of Runge–Kutta RK32, RKF45, and DOPRI54 (see Hairer et al., 1993, for details). We can observe that the order of each method is easily

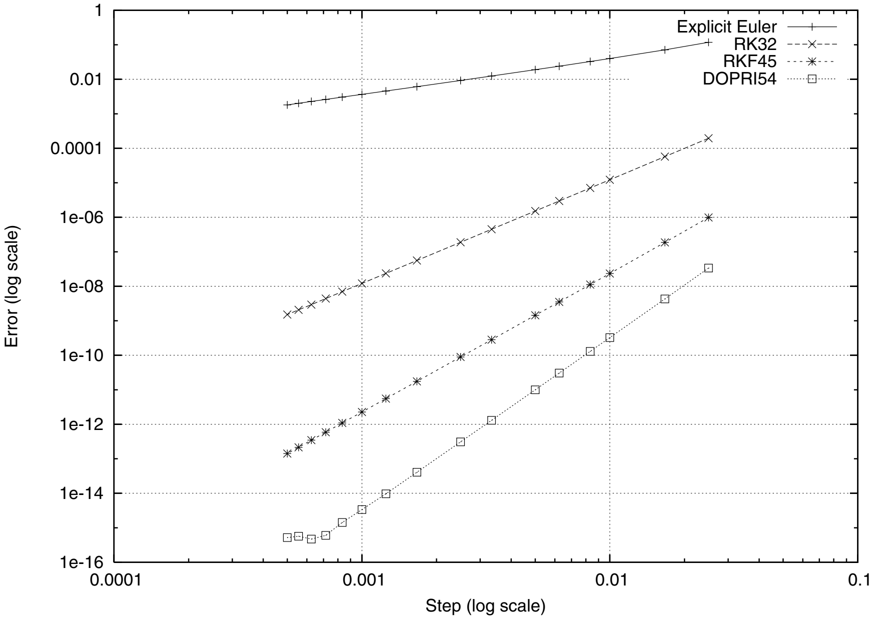


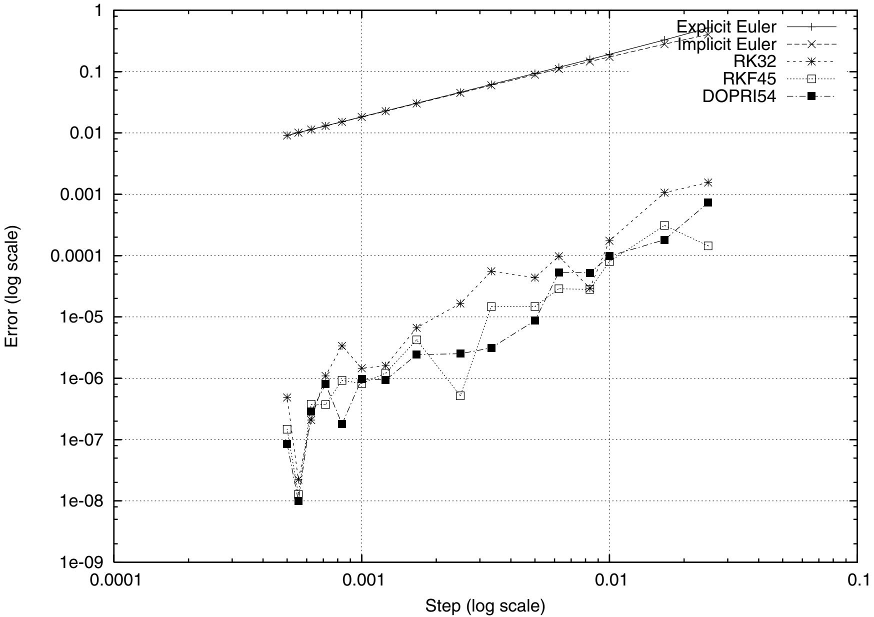
Fig. 9.1. Order of convergence without discontinuities of standard ODE solvers

identified. Thanks to the accuracy of the DOPRI54 method, the machine precision attained is double for a time step equal to 0.0005. In Fig. 9.2, the same numerical experiments are carried out but with the ideal diode model. We can easily see that the orders of convergence of the methods are destroyed.

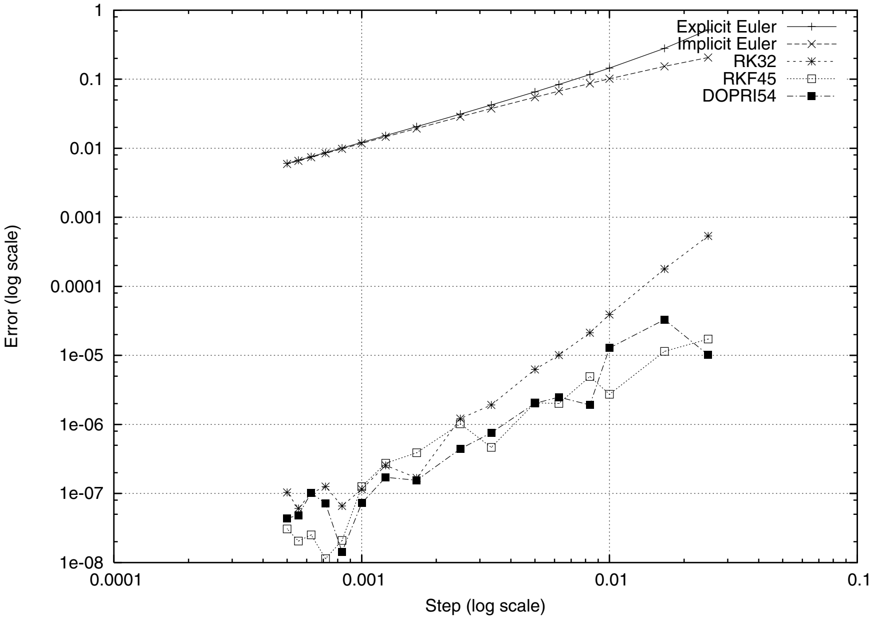
The main conclusion that we can draw is that it is not suitable for higher order integration schemes for ODE with discontinuities.

Stability Results

From the stability point of view, the standard results on the stability domain for the explicit Runge–Kutta schemes are also put into question when some discontinuities are present in the right-hand side. We propose a numerical illustration of this trouble on the simulation of the circuit $(\mathbf{a})$ with $R = 10,000$. The exact solution with $x_1(0) = 1, x_2(0) = -1$, and $L = 1, C = \frac{1}{(2\pi)^2}$ is depicted in Fig. 9.3. In Fig. 9.4, some simulation results with a time step $h = 1 \times 10^{-4}$ are plotted for three schemes: explicit Euler, RK32, DOPRI54. Clearly in Fig. 9.4a and b, some instabilities appear just after the switch and never disappear. In Fig. 9.4c, the instability disappears quickly for this time step $h = 1 \times 10^{-4}$. In Fig. 9.5, the integration with the Runge–Kutta scheme DOPRI54 and $h = 5 \times 10^{-3}$ yields serious instability troubles. Note that each mode can be simulated with coarser time-steps inside the stability domain. The switch and the nonsmoothness of the solution destroys the stability property of the schemes.



(a) Circuit **(a)**



(b) Circuit **(b)**

Fig. 9.2. Order of convergence with discontinuities of standard ODE solvers

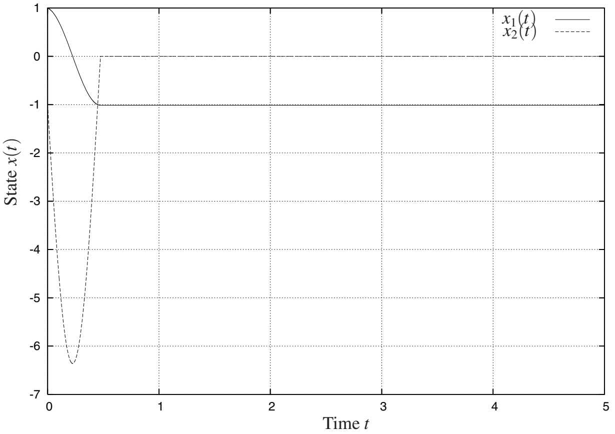


Fig. 9.3. Exact solution for the circuit (**a**) with the initial conditions $x_1(0) = 1, x_2(0) = -1$, and $R = 10,00, L = 1, C = \frac{1}{(2\pi)^2}$. Time step $h = 5 \times 10^{-3}$

9.1.2 Consistent Time-Stepping Methods

In general, discontinuities of order q have the effect of decreasing the order of the method, even when higher order methods are used (see Sect. 9.1.1, see also Sect. 9.2.3 where time-stepping multistep and Runge–Kutta schemes are presented).

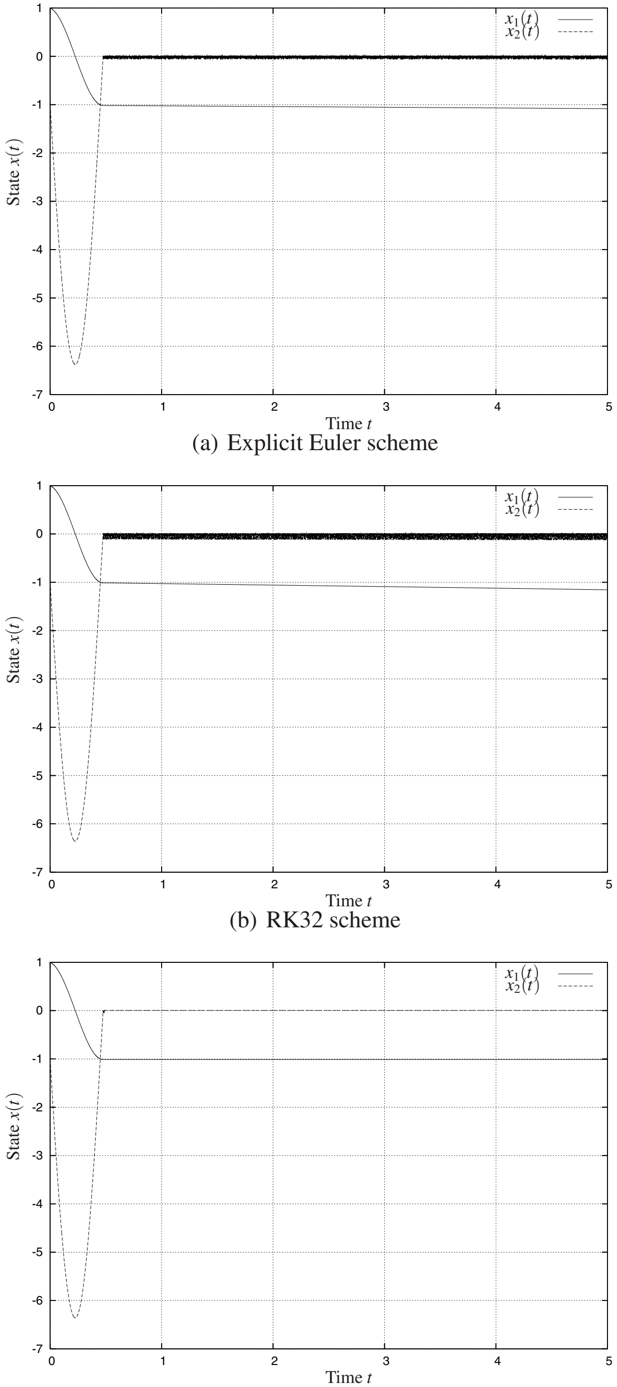
It is consequently of interest to study higher order time-stepping schemes. This may be realized with specific time-stepping strategies in which a discontinuity detection process with low computational cost is used to control the time step and therefore the error of convergence. As we said at the end of Chap. 7, the use of a detection procedure has not to be misunderstood. The goal of the procedure is not to accurately locate the event but to ensure that the integration error on a time step where events are located is sufficiently small. This is the reason why we choose to present these methods as time-stepping strategies. One of the other discrepancies with standard event-driven approaches presented in Sect. 7.2 is as follows: the time-stepping methods presented in this section are proved to be convergent and the order of convergence is theoretically shown.

This type of time-stepping methods originated with the work of Mannshardt (1978) in the context of one-step methods and has been extended and improved in Gear & Østerby (1984) and Enright et al. (1988) and finally in Calvo et al. (2003). Let us start with the pioneering work of Mannshardt (1978).

Sketch of Mannshardt's Method

Let us start with the pioneering work of Mannshardt (1978). As in Sect. 7.2, let us consider dynamical systems given by the following ODE with one discontinuity on the hyper-surface $\{x, t \mid g(x,t) = 0\}$,

$$\dot{x}(t) = f(x,t) = \begin{cases} f^-(t, x(t)) & \text{if } g(t, x(t)) \leqslant 0 & (9.3a) \\ f^+(t, x(t)) & \text{if } g(t, x(t)) > 0, & (9.3b) \end{cases}$$



(a) Explicit Euler scheme

(b) RK32 scheme

(c) DOPRI54 scheme

Fig. 9.4. Stability and instability of explicit Runge–Kutta schemes. Simulation of the circuit (**a**) time step $h = 1 \times 10^{-4}$

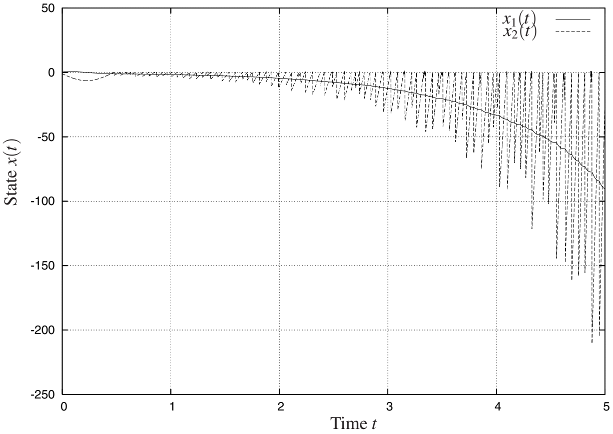


Fig. 9.5. Instability of explicit Runge–Kutta scheme DOPRI54. Simulation of the circuit (**a**) time step $h = 5 \times 10^{-3}$

where $f^-(\cdot)$ and $f^+(\cdot)$ are locally Lipschitz in x and $g : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}$ is smooth (infinitely differentiable). We assume that the right-hand side f has a discontinuity of order q (see Definition 2.56) and that the transversality Assumption 1 holds.

Let us recall some definitions and notations for the one-step time-integration methods already introduced in Sect. 8.6.5.1. A one-step method with an increment function $\phi(\cdot)$ can be defined as follows on the time-step $[t_k, t_{k+1}]$:

$$\begin{cases} x_{k+1} = x_k + h_k \phi(t_k, x_k, h_k) \\ t_{k+1} = t_k + h_k \end{cases} \tag{9.4}$$

or more generally

$$\begin{cases} \chi(t) = x_k + (t - t_k)\phi(t_k, x_k, t - t_k) \\ t_{k+1} = t_k + h_k, \quad x_{k+1} = \chi(t_{k+1}). \end{cases} \tag{9.5}$$

The principle of the method developed in Mannshardt (1978) is as follows:

1. Let ϕ^- be the increment function which integrates (9.3a) with an order p (of consistency and convergence). The provisional value x^-_{k+1} at $t_{k+1} = t_k + h_k$ is given by
$$x^-_{k+1} = \chi^-(t_{k+1}) \tag{9.6}$$
with
$$\chi^-(t) = x_k + (t - t_k)\phi^-(t_k, x_k, t - t_k). \tag{9.7}$$
2. If an event is located in $[t_k, t_{k+1}]$, compute an approximation $\tilde{t}$ of the time of event $t^\star$ such that
$$\tilde{t} = t^\star + \mathrm{O}(h^{p+1}). \tag{9.8}$$
This step is performed by a simplified Newton method.
3. Use the following approximation of the part of the time step:
$$\tilde{x} = \chi^-(\tilde{t}). \tag{9.9}$$

4. Let ϕ^+ be the increment function (which may be identical with ϕ^-) which integrates (9.3b) with an order p (of consistency and convergence). The value $x(t_{k+1})$ can be given at the end of the time step by

$$x_{k+1} = \psi^+(t_{k+1}) \tag{9.10}$$

with

$$\psi^+(t) = \tilde{x} + (t - \tilde{t})\phi^+(\tilde{t}, \tilde{x}, t - \tilde{t}). \tag{9.11}$$

The total increment function for the time step reads as

$$\phi(t_k, x_k, h_k) = \sigma\phi^-(t_k, x_k, \sigma h_k) + (1 - \sigma)\phi^+(\tilde{t}, \tilde{x}, (1 - \sigma)h_k) \tag{9.12}$$

with $\sigma = \dfrac{\tilde{t} - t_k}{h_k}$. The key result is that the order of consistency of this total increment function is p if the increment functions ϕ^- and ϕ^+ are of order p and the condition 9.8 holds. Furthermore, asymptotic developments of the local truncation error are given which allow one to control the time-step size.

Some remarks can be made on this method:

- It is implicitly assumed that there is only one event in the time step. Indeed with the transversality condition (Assumption 1), the number of events is assumed to be finite in a finite time interval. So we can choose a sufficiently small time step such that there is only one event per step.
- We argue that the result on the order of convergence should continue to hold if there is an arbitrarily large number of events inside a time step provided that they are contained in an interval $[\tilde{t}_1, \tilde{t}_1]$ such that $\tilde{t}_2 - \tilde{t}_1 = \mathcal{O}(h^{p+1})$. This is one of the reasons why we class such a method as a time-stepping method. Furthermore, the method could be improved by using a first-order method on this time step.

Improvements and Extensions

Gear & Østerby (1984) proposed a similar approach but without the knowledge of the switching function $g(\cdot)$ in the context of multistep Adams methods with predictor–corrector steps (PECE). In a first stage, the discontinuity is detected by examining the behavior of the step size control procedure. This detection is based on the following ad hoc test:

- The new step h'_k that is predicted by the code after a rejected step h_k implies a reduction factor less than $\frac{1}{2}$, i.e., $h'_k \leqslant \frac{h_k}{2}$,
- or in the last three steps the code had at least two rejections.

Once the discontinuity is detected, it is located and the order of discontinuity is determined by sampling (bisection) the right-hand side and performing time integration on halved time steps. This reduction of the time step is made up to obtain a sufficiently small time step to pass through the discontinuity with the method with the right order of convergence and an acceptable local error estimate.

Similar to the Mannshardt method, the event is not necessarily located accurately and we can expect that multiple events can be taken into account in a single time step.

In Enright et al. (1988), the thorough study of the interpolants for the Runge–Kutta formulas (Enright et al., 1986) is used to locate the discontinuity and to evaluate its order. As in Gear & Østerby (1984), the knowledge of the switching function is not required. The interpolants can be used through a switching function to find events. They are directly studied together with the defects of interpolation to determine the amplitude, the order of the discontinuity. The time-step size and the order of consistency of the Runge–Kutta method is then adapted to pass in a suitable manner the discontinuity.

In Calvo et al. (2003), the aim is to construct a low-cost technique that uses the function evaluations computed on $[t_{k-1},t_k]$ and $[t_{k-2},t_{k-1}]$, which allows one to either confirm or disregard the suspected discontinuity. The presence of a discontinuity is detected thanks to a simple test. A suitable function $q(h_k)$ is constructed such that either $q(h_k)$ is small if the switching surface has not been crossed or $q(h_k)$ is very large if the switching surface has been crossed. The function $q(h_k)$ is the quotient of two linear functions of the Runge–Kutta scheme DOPRI54 evaluations f_1, f_2,..., f_7 (computed on the last successful step), and f_8, f_9 computed at the current failed step. An accurate description on how these linear functions are computed would bring us too far. Let us insist on the fact that the procedure automatically detects the presence of a discontinuity within the integration interval, using only quantities computed by the Runge–Kutta code.

The vector field $f(t,x)$ is evaluated at a particular point (t,u) with $t \in (t_k,t_k+ch_k)$ and $u \approx x(t;t_{k-1},x_{k-1})$. It is assumed that a discontinuity has been detected within (t_k,t_k+ch_k). The algorithm is constructed with a pair of linear forms that use only the vector field at the stages $f(t_{k-1}+c_ih_k,Y_i)$, $i=1,...,s$ of the accepted step and the evaluation $f(t,u)$. The quotient $v(h_k)$ of these forms behaves differently depending on the discontinuity being to the left or to the right of t. It is $1+\mathrm{O}(h_k)$ on the left and $\mathrm{O}(h_k^{-j})$, $j \geqslant 1$, on the right. The discontinuity test consists of checking $v(h_k) \leqslant K$ for some K (to be suitably determined). If yes, then the discontinuity is considered to be to the right of t. If the user wants a precise location of the discontinuity, a bisection technique (advancing the step to $t_k+c\frac{h_k}{2}$) is applied that consists of taking n iterations such that $\frac{ch_k}{2^{n+1}}$ is smaller than the tolerated error. Similar to the previous process, this process uses only quantities that are computed by the Runge–Kutta adaptive scheme. It is not an interpolation procedure as the ones in Definitions 8.11–8.13.

The final step of this scheme is a procedure for crossing the discontinuity and restarting the algorithm. Numerical experiments show that the new adaptive Runge–Kutta scheme supersedes the classical one, with much smaller errors (factors 10^{-1} to 10^{-5} are obtained) and less calculations.

9.2 DIs with Absolutely Continuous Solutions

In this section, time-stepping algorithms for the differential inclusions of Sect. 2.1 are presented. These DIs all possess solutions (possibly nonunique) which are absolutely continuous. Therefore their derivatives may possess discontinuities. The explicit Euler scheme, the θ-method, multistep, and Runge–Kutta algorithms are examined. Roughly speaking, the results which are reported for DIs which possess several solutions state that the discretization process provides one with the approximation of some solution of the continuous-time DI. On the other hand any solution of the DI may be approximated by a solution of the discrete-time inclusion (but it may not be obvious to find it in practice).

9.2.1 Explicit Euler Algorithm

The first approach which is reviewed here is the explicit Euler algorithm that takes the form

$$x_{k+1} - x_k \in h\,F(t_k, x_k) \tag{9.13}$$

in which $h = \frac{T}{N} > 0$ is the time step, $0 = t_0 < t_1 < t_2 < \cdots < t_N = T$ is the interval of integration, $N \in \mathbb{N}$, and the solution is approximated by a piecewise linear function $x^N : [0,T] \to \mathbb{R}^n$, $x^N(t) = x_k + \frac{1}{h}(t - t_k)(x_{k+1} - x_k)$, for $t_k \leqslant t < t_{k+1}$, $k = 0, 1, ..., N-1$. Notice that the inclusion in (9.13) may be rewritten as

$$\begin{cases} x_{k+1} - x_k = h\,\zeta_k \\ \zeta_k \in F(t_k, x_k). \end{cases} \tag{9.14}$$

One has to choose at each step an element ζ_k inside the set $F(t_k, x_k)$. This is called a *selection* procedure. The type of result that one may obtain is closely linked to the properties of the right-hand side, from which upperbounds may be calculated. Notice that if the set $F(t_k, x_k)$ is a cone, then the Euler scheme is simply $x_{k+1} - x_k \in F(t_k, x_k)$: the value of the time step is irrelevant in such a case.

9.2.1.1 Lipschitzian Right-Hand Sides

The first result concerns Lipschitzian DIs with bounded values and is taken from Smirnov (1991).

Theorem 9.1. *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a set-valued map with closed convex values satisfying $F(x) \subset bB_n$, where $b > 0$. Suppose that $F(\cdot)$ is Lipschitz continuous with a constant $l > 0$. Then*

(i) For any solution $x(\cdot)$ of the DI: $\dot{x}(t) \in F(x(t))$, $x(0) = x_0$, there exists solutions of (9.13), with the initial data $x_0^N = x(0)$, such that the functions $w^N(t) = \frac{1}{h}(x^N(t_{k+1}) - x^N(t_k))$, $t \in [kh, (k+1)h)$, $0 \leqslant k \leqslant N-1$, converge to $\dot{x}(\cdot)$ in the $\mathscr{L}^2([0,T];\mathbb{R}^n)$ norm.

(ii) For any solution $x^N(\cdot)$ of the discrete-time inclusion (9.13) with the initial data $x_0^N = x(0)$, there exists a solution $x(\cdot)$ of the DI: $\dot{x}(t) \in F(x(t))$, $x(0) = x_0$, that satisfies $||x(t) - y(t)|| \leqslant hb\,\exp(lT)$ and $||\dot{x}(t) - w(t)|| \leqslant hlb\,(\exp(lT) + 1)$, $t \in [0,T]$, where $w(t) = \frac{1}{h}(x_{k+1} - x_k)$, $t \in [kh, (k+1)h)$, $0 \leqslant k \leqslant N-1$, and $y(t) = x_0 + \int_0^t w(s)ds$.

One sees that $y(\cdot) = x^N(\cdot)$ in (ii). There is no uniqueness of solutions in Lemma 2.8. It is therefore expected that the approximated solutions do not converge or are not close to *the* solution of the continuous-time DI, but only to *one* existing (absolutely continuous) solution. The first part of Theorem 9.1 shows that whatever the solution of the DI may be, the discrete-time inclusion may always produce an approximation of this solution. The second part shows that given an approximated solution, one can always find a corresponding function of time that is close to it, and that is a solution of the DI. Result (ii) shows that the scheme has order 1. The proof of Theorem 9.1 (ii) strongly relies on Theorem 2.9. A similar result, also based on Theorem 2.9, is as follows:

Theorem 9.2. *Let $F(\cdot,\cdot)$ be continuous, Lipschitz continuous in x on bounded sets of $\mathbb{R}^n$, with compact and convex values $F(t,x)$ for all $x \in \mathbb{R}^n$ and each $t \in [0,T]$. Let also a linear growth condition $||\zeta|| \leqslant c(1+||x||)$ for all $\zeta \in F(t,x)$, $x \in \mathbb{R}^n$, $t \in [0,T]$. Then for every $\varepsilon > 0$, there exists N^* such that for every $N > N^*$ and for every solution $x^N(\cdot)$ of the discrete inclusion with initial data $x^N(0) = x_0$, there exists a solution $x(\cdot)$ of the DI: $\dot{x}(t) \in F(t,x(t))$, $x(0) = x_0$, such that*

$$\max_{t\in[0,T]} ||x^N(t) - x(t)|| \leqslant \varepsilon. \tag{9.15}$$

If in addition $F(\cdot)$ has $\mathbb{R}^n$ as its domain of definition and is Lipschitz continuous in the set $\{z \in \mathbb{R}^n \mid ||z - x(t)|| \leqslant \varepsilon$, for some $t \in [0,T]\}$, then there exists N^* such that for every $N > N^*$

$$\max_{0\leqslant k\leqslant N} ||x^N(t_k) - x(t_k)|| \leqslant ch \tag{9.16}$$

for some constant c.

This was proved in Pshenichny (1980).

Example 9.3. Consider $\dot{x}(t) \in [-2+\sin(x(t)), 2+\sin(x(t))]$. The set-valued function $F(\cdot)$ is Lipschitz continuous with constant $l = 1$. Indeed $[-2+\sin(x), 2+\sin(x)] \subset [-2+\sin(y) - |x-y|, 2+\sin(y)+|x-y|]$, since $\sin(x) - \sin(y) > -|x-y|$ and $\sin(x) - \sin(y) < |x-y|$. It is also bounded as $F(x) \subset [-4,4]$ for all $x \in \mathbb{R}$. The sets $F(x)$ are closed and convex (but the graph of the multifunction is not convex). Therefore one may apply Theorem 9.1 to the discrete inclusion $x_{k+1} \in x_k + h[-2+\sin(x_k), 2+\sin(x_k)]$.

Let us now state a result that concerns specific DIs of the class (2.5):

$$\dot{x}(t) \in A(t)x(t) + B(t)U, \;\; x(0) = x_0 \in X_0, \tag{9.17}$$

where $A(\cdot)$ and $B(\cdot)$ are $n \times n$ and $n \times p$ matrices, differentiable functions with Lipschitz-continuous derivatives, and both $U \in \mathbb{R}^p$ and $X_0 \in \mathbb{R}^n$ are convex compact sets. Such DIs may arise in some optimal control problems. The proposed time-discretization of (9.17) is

$$\begin{cases} x_{k+1} \in A_{h,k}x_k + B_{h,k}U, \;\; X_{0,h} = X_0 \\ A_{h,k} = I_n + \frac{h}{2}(A(t_k) + A(t_{k+1})) + \frac{h^2}{2}A^2(t_{k+1}) \\ B_{h,k} = \frac{h}{2}(B(t_k) + B(t_{k+1})) + \frac{h^2}{2}A(t_{k+1})B(t_{k+1}) \end{cases} \tag{9.18}$$

for $0 \leqslant k \leqslant N-1$, and a piecewise linear function $x^N(\cdot)$ is constructed from this discretization. Let us define the reachable set of the DI in (9.17), starting from the set X_0, as

$$\begin{aligned} R(X_0;T) = \{x(T) \mid x : [0,T] \to \mathrm{I\!R}^n, &\text{ absolutely continuous, } x(\cdot) \text{ satisfies} \\ &(9.17) \text{ almost everywhere on } [0,T],\, x(0) \in X_0\}. \end{aligned} \tag{9.19}$$

The next result is taken from Veliov (1992).

Theorem 9.4. *Under the stated assumptions, the approximated solution* $x^N(\cdot)$ *of the inclusion in (9.18) satisfies*

$$d_H(x^N(t_N), R(X_0;T)) \leqslant c\, h^2 \tag{9.20}$$

for some constant c.

There is therefore an order 2 of convergence; however, the convergence does not concern the solutions of the DI, but the reachable set.

These results that apply to DIs with a Lipschitz right-hand side, say that if the time-step $h >$ is taken sufficiently small, then every solution of the discrete inclusion has in its neighborhood a solution of the continuous-time DI. One may conclude that Euler schemes approximate such DIs correctly. This, however, does not mean that one will always be content with such an algorithm.

9.2.1.2 Upper Semi-continuous Right-Hand Sides

The next theorem may be found in many works under various forms and can be traced back to Filippov (1988). Let X_h denote the set of solutions of the discrete inclusion (9.13), with $x^N(0) = x(0) = x_0$.

Theorem 9.5. *Let the set-valued map* $F(\cdot,\cdot)$ *take nonempty, convex, compact values and satisfy a linear growth condition* $||\zeta|| \leqslant c(1+||x||)$ *for all* $\zeta \in F(t,x)$ *and some constant c,* $x \in \mathrm{I\!R}^n$, $t \in [0,T]$. *Then every sequence* $\{x^N(\cdot)\}_{N \in \mathrm{I\!N}}$ *with* $x^N(\cdot) \in X_h$ *for* $N \in \mathrm{I\!N}$ *has a subsequence which converges uniformly in* $[0,T]$ *to some solution of the DI:* $\dot{x}(t) \in F(t,x(t))$, $x(0) = x_0$, *as* $N \to +\infty$.

Let X denote the set of solutions of the continuous-time DI. The convergence of Theorem 9.5 means that the set $\{f(\cdot) \in C^0([0,T];\mathrm{I\!R}^n) \mid \liminf_{h\to 0} d_{\mathrm{H}}(f, X_h) = 0\}$ is a subset of X. Obviously if the DI has a unique solution, the subsequence converges to it. Recall from Proposition 2.11 that the conditions of Theorem 9.5 guarantee the outer semi-continuity. In practice Theorem 9.5 may not provide sufficient results; see, e.g., Sect. 9.7 for comments on explicit schemes.

9.2.1.3 One-Sided Lipschitz-Continuous Right-Hand Sides

As we have seen in Sect. 2.1.3, the OSLC of the right-hand side guarantees nice properties of the solutions. For instance the UOSLC condition guarantees the uniqueness of solutions (Lemma 2.1.3). The next result, taken from Lempio (1992), is about the order of the Euler method. Let us consider the perturbed Euler method

$$\begin{cases} x_{k+1} - x_k = h\,\zeta_k + h\varepsilon_k \\ \zeta_k \in F(t_k, x_k) \\ x^N(0) = x(0) + \varepsilon_0, \end{cases} \tag{9.21}$$

where ε_0 is the error on the initial approximation and ε_k reflects all the errors made when selecting an element ζ_k in $F(t_k, x_k)$.

Theorem 9.6. *Let $F : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a set-valued mapping with closed graph, satisfying a uniform one-sided Lipschitz condition (UOSLC), and with domain $\mathbb{R}$. Let the solution of the DI: $\dot{x}(t) \in F(t, x(t))$, $x(0) = x_0$, exist and be piecewise Lipschitz continuously differentiable on $\mathbb{R}$. Let the initial error $|\varepsilon_0|$ and the mean error $\frac{1}{h}\sum_{i=1}^{N} |\varepsilon_i|$ be of order 1 as functions of the step size $h > 0$. Then the order of convergence of the Euler method in (9.21) is equal to 1.*

This theorem is interesting as it is similar to the results of Janin & Lamarque (2001) that apply to event-driven schemes of Lagrangian systems. Provided the "errors" respect some order condition, the whole scheme possesses a minimum order. In Lagrangian systems, the errors come from the velocity jumps. In Theorem 9.6 they are a consequence of the numerical uncertainty.

Let now the interval of integration be $[0,1]$, i.e., we take $T = 1$. Let $A \subset \mathbb{R}^n$ be compact and $F : [0,1] \times A \to \mathbb{R}^n$ with $F(t,x)$ convex and compact, uniformly bounded, measurable in t, continuous in x. We define the averaged $\mathscr{L}^1$−moduli as

$$\Xi(F,A,h) = \int_0^1 \Xi(F,A,t,h)\mathrm{d}t, \;\; \tau(F,A,h) = \int_0^1 \tau(F,A,t,h)\mathrm{d}t,$$

where $\Xi(F,A,t,h) = \sup\{d_{\mathrm{H}}(F(t,x),F(t,y)) \mid ||x-y|| \leqslant h,\ x \in A, y \in A\}$, and $\tau(F,A,t,h) = \sup\{\sup(d_{\mathrm{H}}(F(s,x),F(r,x)) \mid s,r \in [t-\frac{h}{2}, t+\frac{h}{2}] \cap \mathbb{R}) \mid x \in A\}$. $\Xi(F,A,t,h)$ and $\tau(F,A,t,h)$ are called the local moduli of continuity of $F(\cdot,\cdot)$ with respect to each argument t and x. The next lemma is taken from Dontchev & Farkhi (1998).

Lemma 9.7. *Suppose that there exists an integrable function $\lambda : [0,1] \to \mathbb{R}^+$ such that $||F(t,x)|| \leqslant \lambda(t)(1+||x||)$ for all $x \in \mathbb{R}^n$ and almost all $t \in [0,1]$. Then every trajectory of the discrete inclusion (9.14) with initial data $x^N(0) \in K_0$ is bounded by*

$$\max_{t\in[0,1]} ||x^N(t)|| \leqslant \exp(\Lambda)(|K_0| + \Lambda), \;\; \Lambda = \sup_{N\in\mathbb{N}} \frac{1}{N}\sum_{i=0}^{N} \lambda(t_i), \tag{9.22}$$

where $|K_0| = \sup\{||x|| \mid x \in K_0\}$*. Suppose in addition that* $F(\cdot,\cdot)$ *defined on* $[0,1] \times \mathrm{IR}^n$ *has nonempty, compact, and convex values and is OSLC with an integrable function* $L(\cdot)$*. Suppose also that the linear growth function* $\lambda(\cdot)$ *is a constant* λ*, so that all trajectories of the continuous time and of the discrete inclusions are contained in a bounded set A and all the derivatives in a bounded set B. Then for every solution* $x(\cdot)$ *of the DI:* $\dot{x}(t) \in F(t,x(t))$ *for almost all* $t \in [0,1]$*,* $x(0) \in K_0$*, there exists a trajectory* $x^N(\cdot)$ *of the discrete inclusion (9.14) with initial data* $x^N(0) \in K_0$ *such that*

$$\max_{t\in[0,1]} ||x(t) - x^N(t)|| \leqslant c[\tau(F,A,h) + \Xi(F,A,h)], \tag{9.23}$$

where $c = \exp(m(1))\max(2,|B|)$*,* $m(t) = \int_0^t L(r)dr$*,* $m(t) = \int_0^t L(s)ds$*.*

We recall the notation $||F(t,x)|| = \sup\{||z|| \mid z \in F(t,x)\}$. This lemma states results of the same nature as Theorems 9.1 and 9.2: an explicit Euler algorithm yields a "good" approximation of the continuous-time DI, in the sense that to every solution of the DI one may associate a trajectory of the discrete inclusion. Suppose for instance that the OSLC function $L(\cdot)$ is a constant $L < 0$ with $|L|$ large enough. Then $m(t) = Lt$, and $c = \exp(L)\max(2,|B|)$. The set B is the set within which the derivative of the trajectories lies. If the right-hand side is gentle enough with a linear growth constant $\lambda > 0$ that is very small, then $|B|$ is small also and c is small. The upper bound in (9.23) therefore reflects the "agitation" of the DI.

9.2.2 Implicit θ-Method

The θ-method for the discretization of the DI: $\dot{x}(t) \in F(t,x(t))$, $x(0) = x_0$, on the interval $[0,T]$, yields the discrete inclusion:

$$\begin{cases} x_{k+1} \in x_k + hF(t_k,x_{k+\theta}) \\ \\ x_{k+\theta} = \theta x_{k+1} + (1-\theta)x_k, \;\; \theta \in [0,1]. \end{cases} \tag{9.24}$$

The explicit Euler method is for $\theta = 0$, the mid-point rule is for $\theta = \frac{1}{2}$, the fully implicit Euler method is for $\theta = 1$. In the next theorem it is assumed that $F(t,x) = f(t,x) - A(x)$, where $f : [0,T] \times \mathrm{IR}^n \to \mathrm{IR}^n$ is Lipschitz continuous in both arguments, i.e., $||f(t,x) - f(s,y)|| \leqslant k(|t-s| + ||x-y||)$ for all $t,s \in [0,T]$ and all $x,y \in \mathrm{IR}^n$ and some constant k. The set-valued mapping $A(\cdot)$ is maximal monotone, with closed domain. The intervals on which the solution stays in the subspaces $\Sigma_{\mathscr{J}} = \{x \in \mathrm{IR}^n \mid x_j = 0,\; j \in \mathscr{J}\}$, where $\mathscr{J}$ is a subset of $\{1,2,...,n\}$, are called the *stiction* intervals. We call the *junction* times the times such that $x(t) \in \Sigma_{\mathscr{J}}$, and $x_j(t+\varepsilon) \neq 0$, $x_j(t-\varepsilon) \neq 0$ for $\varepsilon > 0$ arbitrarily small, $j \in \mathscr{J}$. The next theorem is taken from Elliot (1985).

Theorem 9.8. *Let* $x(\cdot)$ *be the (unique) solution of the DI:* $\dot{x}(t) \in f(t,x(t)) - A(x(t))$ *on* $[0,T]$*,* $x(0) = x_0$*. Let the discrete inclusion (9.24) be initialized with* $x^N(0) = x_0$*. Assume that there are at most a finite number of stiction intervals, and there are at most a finite number of junction times in* $[0,T]$*. Then for* $h > 0$ *sufficiently small one has*

$$||x^N(t_k) - x(t_k)|| \leqslant ch \tag{9.25}$$

for some constant c independent of h and N.

One reminds here that the switching instants must be of finite number in the integration interval. This is also the case of Stewart's event-driven method in Sect. 7.1.2. Theorem 9.8 guarantees convergence as $h \to 0$ for all $\theta \in [0,1]$. However, this does not mean that the algorithms will all behave in the same way when θ varies from 0 to 1. See Sects. 9.2.4.2 and 9.7 for details on this aspect. Other results for similar maximal monotone inclusions (UDIs) have been obtained in Bastien & Schatzman (2002) in the fully implicit case, following the work in Lippold (1990). An order $\frac{1}{2}$ is proved without relying on any sort of assumption on the solution. The order is shown to be 1 when the maximal monotone multifunction is $\partial\psi_K(\cdot)$, the indicator of a nonempty closed convex set K.

Let us now describe a result that holds for Filippov's DIs constructed from a switching codimension one surface $S \subset \mathbb{R}^n$ and two vector fields as in (2.14). The next theorem is taken from Kastner-Maresch (1992) and concerns an implicit midpoint algorithm.

Theorem 9.9. *Let us consider $\theta = \frac{1}{2}$ in the θ-method (9.24). Let the set-valued mapping $F(\cdot,\cdot)$ be a Filippov's right-hand side, satisfying a linear growth condition, and a uniform one-sided Lipschitz-continuous (UOSLC) condition with constant L. Assume further that*

- *The solutions of the DI: $\dot{x}(t) \in F(t,x(t))$, $x(0) = x_0$, are piecewise twice continuously one-sided differentiable.*
- *The initial approximation satisfies $||x_0 - x^N(0)|| = O(N^{-1})$.*
- *$h \leqslant \frac{K}{N}$ for some $K > 0$.*

Then there exists constants c and N^* (possibly depending on L) such that for all $N > N^*$ and all solutions of the DI,

$$||x_k - x(t_k)|| \leqslant \frac{c}{N} \tag{9.26}$$

for all $0 \leqslant k \leqslant N$.

Recall that Filippov's convexification procedure implies that the set-valued map is upper semi-continuous. Also the assumptions assure that the solutions of the DI are unique (see Lemma 2.31). In its original formulation the theorem holds with a varying step h_i and with perturbations included in the analysis. Once again the result holds provided the trajectories do not cross the switching surface too often. The notation $f = \mathrm{O}(g)$ for two functions $f(\cdot)$ and $g(\cdot)$ means that there exists a real $\beta \geqslant 0$ such that $||f(x)|| \leqslant \beta||g(x)||$. A function is one-sided differentiable at x if the quantities $\lim_{h\to 0, h>0} \frac{f(x+h)-f(x)}{h}$ or $\lim_{h\to 0, h<0} \frac{f(x+h)-f(x)}{h}$ exist and are right and left continuous, respectively.

Remark 9.10. Possibly these time-stepping methods accommodate for accumulations of switching instants, in the sense that the algorithm will not produce diverging solutions. But the order of the scheme then cannot be guaranteed (or at least it cannot be proved analytically). Such accumulations often occur in stabilization problems with discontinuous feedback controllers.

9.2.3 Multistep and Runge–Kutta Algorithms

The Euler explicit method is not the only numerical scheme one may use to discretize Lipschitz or upper semi-continuous DIs. Multistep and Runge–Kutta methods possess the advantage over Euler methods that they are more accurate on intervals where the trajectories are differentiable. This does not mean that their order is larger than 1.

9.2.3.1 The Linear Multistep Algorithm

Let reals a_i, b_i, $0 \leqslant i \leqslant r$,[1] be given, with $a_r \neq 0$, $|a_0| + |b_0| > 0$. The starting values are x_k, $k = 0, 1, ..., r-1$, and the corresponding starting selections are $\zeta_k \in F(t_k, x_k)$, $k = 0, 1, ..., r-1$. As for the Euler method, a procedure has to be chosen to select the ζ_k, see Sect. 9.2.4.1. Moreover the first $r-1$ steps must be initialized by another method (a multistep method with $\bar{r} < r$ steps or a one-step method like the Euler scheme). The algorithm is advanced from step $k-1$ to step k as follows:

For $k = 1, 2, ...N$, compute x_k from

$$\begin{cases} \sum_{i=0}^{r} a_i x_{k-r+i} = h \sum_{i=0}^{r} b_i \zeta_{k-r+i} \\ \\ \zeta_k \in F(t_k, x_k). \end{cases} \tag{9.27}$$

The case $r = 1$, $b_1 = 0$, $b_0 = a_0 = 1$, $a_1 = -1$, is the explicit Euler scheme. When $b_r \neq 0$ the multistep method is implicit. The next theorem is due to Taubert (1981).

Theorem 9.11. *We consider the DI: $\dot{x}(t) \in F(t, x(t))$, $x(0) = x_0$. Let $F : [0, T] \times \mathrm{I\!R}^n \rightarrow \mathrm{I\!R}^n$ be a set-valued mapping with nonempty, closed, convex values $F(t, x)$ that is bounded (there exists a constant c such that for all $(t, x) \in \mathrm{I\!R} \times \mathrm{I\!R}^n$ and any $z \in F(t, x)$, one has $||z|| \leqslant c$) and upper semi-continuous. Assume further that*

- *(Strong stability) The roots of the characteristic polynomial $\sum_{i=0}^{r} a_i \lambda^i$ are, except the simple root $\lambda = 1$, of absolute value $|\lambda| < 1$.*
- *(Consistency) $\sum_{i=0}^{r} a_i = 0$, $\sum_{i=0}^{r} i a_i = \sum_{i=0}^{r} b_i$, $b_i \geqslant 0$, $0 \leqslant i \leqslant r$.*
- *(Starting field) There exists a constant M, independent of $h > 0$, such that $||x_i^N - x_0|| = o(h)$ for $0 \leqslant i \leqslant r-1$.*

[1] This r has nothing to do with the relative degree introduced elsewhere in the book, obviously.

Assume finally that the DI enjoys the uniqueness of solutions property. Then on any compact interval of $[0,T] \ni t$, $T > 0$, *the solutions of the discretized inclusion (9.27) are such that*

$$\lim_{h \to 0, t_n \to t} ||x^N(t_n) - x(t)|| = 0, \tag{9.28}$$

where $t_n = nh$.

Recall that for two functions $f(\cdot)$ and $g(\cdot)$, $f = o(g)$ means that for all $\varepsilon > 0$ one has $||f(x)|| \leqslant \varepsilon ||g(x)||$. The starting field condition thus means that $||x_i^N - x_0|| \leqslant hM$ for some constant M depending on T and $F(\cdot,\cdot)$. A variant of Theorem 9.11 is as follows (Dontchev & Lempio, 1992):

Theorem 9.12. *Let all the assumptions of Theorem 9.11 hold, with the starting field condition* $||x_{k+1}^N - x_k^N|| \leqslant hM$, $k = 0, 1, ..., r-2$, *for all* $N \in \mathbb{N}$, *and a constant* M *independent of* $h > 0$. *Let also the approximations of the initial value* $x_0 = x(0)$ *satisfy* $\lim_{h \to 0} x^N(0) = x_0$. *Then the sequence of continuous piecewise linear functions* $\{x^N(\cdot)\}_{N \in \mathbb{N}}$ *contains a subsequence which converges uniformly to a solution of the DI.*

In the implicit case $b_r \neq 0$, the existence of x_k in (9.27) is assured if the generalized equation

$$0 \in \sum_{i=0}^{r-1} a_i x_{k-r+i} - h \sum_{i=0}^{r-1} b_i \zeta_{k-r+i} - h b_r F(x_k) + a_r x_k$$

has a solution. This may be proved with a fixed-point theorem, see Taubert (1981).

9.2.3.2 The Runge–Kutta Method

Runge–Kutta algorithms of order p are[2] meant to approximate the solution and its first p derivatives, in the sense that the Taylor series of the exact solution $x(h)$ and the Taylor series of the approximated solution $x^N(t_1)$ coincide up to the term h^{p-1}, including it. The solutions of the DIs we are studying usually are absolutely continuous, and therefore have a derivative that may jump (consequently higher order derivatives are distributions). The first question one may ask to one's self is: how will a Runge–Kutta algorithm of order $p > 1$ behave with such nonsmooth solutions? A reassuring point is that the multistep methods described above behave correctly in the sense that they lose their order, but are still convergent. Similarly the implicit mid-point rule has order 1.

Assumption 12. *It is assumed that the set-valued mapping is upper semi-continuous, satisfies a growth condition, and is UOSLC. Moreover its domain is* $\mathbb{R}^n$, *i.e., for each* t, *the set* $\{x \in \mathbb{R}^n \mid F(t,x) \neq \emptyset\} = \mathbb{R}^n$.

[2] The order in Hairer et al. (1993) is our definition of the order minus 1, see definition 1.2, Chap. II.1 in that book.

From the results of Chap. 2, the DI: $\dot{x}(t) \in F(t,x(t))$, $x(0) = x_0$, has a unique absolutely continuous solution on $\mathrm{I\!R}^+$ for each x_0.

To simplify the presentation it is assumed that a constant time step $h > 0$ is chosen. However, varying steps h_k are implementable. The algorithm is advanced from step k to step $k+1$ as follows:

For $k = 0,2,...,N-1$, compute x_k from

$$\begin{cases} X_j = x_k + h\ \sum_{i=1}^{r} a_{ji}\zeta_{k+1,j} \\ \zeta_{k+1,j} \in F(t_k + c_i h, X_i),\ \ i,j \in (1,2,...,r)\ , \\ x_{k+1} = x_k + h\ \sum_{i=1}^{r} b_i \zeta_{k+1,i} \end{cases} \tag{9.29}$$

where the coefficients c_i, b_i, $1 \leqslant i \leqslant r$, and a_{ji}, $1 \leqslant j \leqslant r$, $1 \leqslant i \leqslant r$, are computed so as to satisfy some constraints. It is supposed that $0 \leqslant c_i \leqslant 1$ for all $1 \leqslant i \leqslant r$. A perturbed version of the algorithm (9.29) is as follows, where δ_i, $1 \leqslant i \leqslant r = 1$, represents the disturbance due to nonperfect initialization.

For $k = 0,2,...,N-1$, compute x_k from

$$\begin{cases} X_j = x_j + h\ \sum_{i=1}^{r} a_{ji}\zeta_{k+1,j} + \delta_j \\ \zeta_{k+1,j} \in F(t_k + c_i h, X_i),\ \ i,j \in (1,2,...,r)\ , \\ x_{k+1,p} = x_{k,p} + h\ \sum_{i=1}^{r} b_i \zeta_{k+1,i} + \delta_{r+1} \end{cases} \tag{9.30}$$

where the subscript p is for perturbed. Since such schemes may be implicit, it is important to test their *practicability*, i.e., are the discrete inclusions solvable? If yes the step $k \rightarrow k+1$ is said to be (uniquely) practicable. The next two notions will be used for convergence of the scheme.

Definition 9.13. *(Convergence stability) Let x_{k+1} and $x_{k+1,p}$ be two arbitrary uniquely practicable parallel steps of the schemes (9.29) and (9.30), respectively. The Runge–Kutta method is said* C-stable *if there exists constants $c_0 \geqslant 0$ and $h_0 > 0$ with*

$$||x_{k+1} - x_{k+1,p}|| \leqslant (1 + c_0 h)\ ||x_k - x_{k,p}|| \tag{9.31}$$

for all $h \in (0, h_0]$.

Definition 9.14. *(B-stage stability) Let x_{k+1} and $x_{k+1,p}$ be two uniquely solvable steps of the schemes (9.29) and (9.30), respectively. The Runge–Kutta method is said* BS-stable *if there exists constants $c_1 > 0$, $h_1 > 0$, such that*

$$||x_{k+1,p} - x_{k+1}|| \leqslant c_1\ \max(||\delta_1||, ..., ||\delta_{r+1}||) \tag{9.32}$$

for $h \in (0, h_1]$, and all $(t_k, x_k) \in [0,T] \times \mathrm{I\!R}^n$, all δ_1, δ_2,..., δ_{r+1}.

Finally the *order of consistency* of the scheme is equal to $p \geqslant 0$ if there exist constants $d \geqslant 0$, $h_1 > 0$, such that

$$||x_{k+1}^N - x(t_{k+1})|| \leqslant d\, h^{p+1} \tag{9.33}$$

for all $h \in (0, h_1]$, $t_{k+1} \in [0, T]$, and for all solutions to the continuous-time DI.[3] In order to state the next lemma, we need some more definitions. A function $f(\cdot)$ is right-sided continuously differentiable at t if the quantity $\dot{f}^+(t) = \lim_{\varepsilon \to 0, \varepsilon > 0} \frac{f(t+\varepsilon)-f(t)}{\varepsilon}$ exists and is right-continuous. Then we define

$$M^+ = \max_{t \in [0,T]} ||\dot{x}^+(t)||. \tag{9.34}$$

Let us finally define the following perturbations:

$$\begin{cases} \delta_j \overset{\Delta}{=} x(t + c_j h) - x(t_k) - h\, \sum_{i=1}^r a_{ji} \dot{x}^+(t + c_j h), \;\; j = 1, ..., r \\ \delta_{r+1} \overset{\Delta}{=} x(t_{k+1}) - x(t_k) - h\, \sum_{i=1}^r b_i \dot{x}^+(t + c_j h). \end{cases} \tag{9.35}$$

We are now ready to state the first result.

Lemma 9.15. *Let the set-valued mapping $F(\cdot,\cdot)$ satisfy Assumption 12, and the solution to the DI be right-sided continuously differentiable, with $M^+ < +\infty$. Then the step $k \to k+1$ is practicable with the perturbations in (9.35). Furthermore there exists a constant $\hat{d} > 0$ such that*

$$||\delta_j|| \leqslant \hat{d} h, \;\; j = 1, ..., r+1. \tag{9.36}$$

The constant $\hat{d}$ depends only on M^+ and on the parameters of the method.

In addition to the order of consistency, one defines the order of convergence. It is assumed that the three steps:

1. the unperturbed step starting at (t_k, x_k): $x_k \mapsto x_{k+1}$,
2. the unperturbed step starting at $(t_k, x(t_k))$: $x(t_k) \mapsto x_{k+1}$,
3. the perturbed step starting at $(t_k, x(t_k))$ and ending on the solution $x(t_{k+1})$: $x(t_k) \mapsto x(t_{k+1})$,

are uniquely practicable for $N \geqslant N_0$. The Runge–Kutta method has *order of convergence* $p > 0$ if there exist constants $c > 0$ and $\bar{N} > 0$ such that

$$||x^N(t_N) - x(t_N)|| \leqslant c \left(\frac{1}{N}\right)^p \tag{9.37}$$

for all $N \geqslant \bar{N} \geqslant N_0$ and all solutions of the DI. The constants c and $\bar{N}$ may depend on the OSLC constant L, on M^+ and on the parameters of the method. The main result from Kastner-Maresch (1990–91) is as follows:

[3] We recall that the approximated piecewise continuous solution is x^N: $[0, T] \to \mathbb{R}^n$, with $x^N(t) = x_k + \frac{1}{h}(t - t_k)(x_{k+1} - x_k)$, for $t \in [t_k, t_{k+1})$, and that $x(t_k)$ denotes the value at time t_k of the solution of the DI.

Theorem 9.16. *Let for a fixed $p \in \mathbb{N}$ the following assumptions hold:*

- *Assumption 12 on the set-valued mapping holds.*
- *There exists a partition of $[0,T]$ into finitely many subintervals such that the solution of the DI is right-sided continuously differentiable on $[0,T]$, and $(p+1)$-times continuously differentiable on each subinterval.*
- *The simplifying assumptions*

$$\begin{cases} k\ \sum_{i=1}^{r} b_i c_i^{k-1} = 1, \ \ (k=1,...,p) \\ \\ k\ \sum_{i=1}^{r} a_{ji} ci^{k-1} = c_j^k \ \ (k=1,...,p\, and\ j=1,...,r). \end{cases} \tag{9.38}$$

- *The Runge–Kutta method is BS-stable and C-stable.*
- *The initial approximation $x_0^N = x^N(t_0)$ of $x_0 = x(0)$ satisfies*

$$||x_0^N - x_0|| = \mathrm{O}\left(\frac{1}{N}\right). \tag{9.39}$$

Then the order of convergence is at least equal to 1.

The theorem does not say that the order of convergence is equal to 1, it may possibly be larger. However, it seems that if the derivative of the solution has a (finite) number of jumps, as allowed by the assumptions, then the order cannot be larger than 1.

Remark 9.17. We have seen that the class of set-valued maps for which Theorem 9.16 applies contains maps of the form $f(t,x) - \beta(x)$ where $f(\cdot,\cdot)$ is Lipschitz continuous and $\beta(\cdot)$ is a maximal monotone mapping. However, this class does not include *normal cones*, i.e., right-hand sides of the form $\partial\psi_K(\cdot)$, whose domain is not the whole of $\mathbb{R}^n$, but K. Let us examine the case of such right-hand sides, with $r = 1$. The Runge–Kutta method reads

$$\begin{cases} x_{k+1} x_k \in h b_1\ F(x_k + h a_{11} \zeta_{k1}) \\ \\ \zeta_{k1} \in F(x_k + h a_{11} \zeta_{k1}), \end{cases} \tag{9.40}$$

which is the mid-point scheme (the θ-method with $\theta = \frac{1}{2}$). Suppose that $F(x) = -N_K(x)$ for some closed nonempty convex set $K \subset \mathbb{R}^n$. Then we obtain

$$\begin{cases} x_{k+1} x_k \in -h b_1\ N_K(x_k + h a_{11} \zeta_{k1}) \\ \\ \zeta_{k1} \in -N_K(x_k + h a_{11} \zeta_{k1}). \end{cases} \tag{9.41}$$

We may rewrite (9.41) as

$$x_k + h a_{11} \zeta_{k1} + \zeta_{k1} - x_k - h a_{11} \zeta_{k1} \in -N_K(x_k + h a_{11} \zeta_{k1}) \tag{9.42}$$

that is equivalent to (see Sect. A.3)

$$x_k + ha_{11}\zeta_{k1} = \text{prox}[K; (ha_{11} - 1)\zeta_{k1} + x_k]. \tag{9.43}$$

It is not clear how to use (9.43) to advance from step k to step $k+1$ with the method in (9.41). Let K be a nonempty closed convex cone. Notice that we may rewrite (9.41) as

$$x_k + ha_{11}\zeta_{k1} \in \partial\psi_K^*(-\zeta_{k1}) = N_{K^o}(-\zeta_{k1}), \tag{9.44}$$

where $\psi_K^*(\cdot)$ is the conjugate function of the indicator of K, and K^o is the polar cone to K (which is another nonempty convex cone). We deduce that

$$-\zeta_{k1} = \text{prox}\left[K^o; \frac{1}{ha_{11}} x_k\right] \tag{9.45}$$

so that

$$x_{k+1} - x_k \in -N_K\left(x_k + ha_{11}\,\text{prox}\left[K^o; -\frac{1}{ha_{11}} x_k\right]\right). \tag{9.46}$$

Once again the meaning of such an inclusion is not clear. It seems that the Runge–Kutta method with normal cones to convex sets is not practicable in the above sense.

We end this section on Runge–Kutta methods with the following result:

Lemma 9.18. *Let all the assumptions and conditions of Theorem 9.16 be satisfied. Suppose that the solution of the DI is $(p+1)$-times continuously differentiable on $[0,T]$, and that the initial approximation satisfies $||x_0^N - x_0|| = O(h^p)$. Then the order of convergence of the scheme is equal to p.*

One may wonder what Lemma 9.18 is useful to, since the solutions we are dealing with usually are absolutely continuous. In fact it shows that if there are portions of the trajectory on which the trajectory is smooth enough, then the order is preserved. Therefore the accuracy is good for systems that have solutions with separated times of nondifferentiability. This may also be useful in the context of event-driven methods, provided an accurate enough event detection is coupled to the Runge–Kutta algorithm, see Sect. 7.2.

9.2.4 Computational Results and Comments

Let us consider the two DIs: $\dot{x}(t) \in [-1,1]$, $t \geqslant 0$, and $\dot{x}(t) \in -\text{sgn}(x(t))$, $t \geqslant 0$, where $\text{sgn}(\cdot)$ is the multivalued sign function. Obviously they are quite different. The first DI satisfies the conditions of Lemma 2.8 and has an infinity of solutions for any $x(0) \in \mathbb{R}$. The second DI satisfies the conditions of Theorem 2.41 and has a unique solution for all $x(0) \in \mathbb{R}$. What is the best algorithm to simulate these DIs? It seems that the first DI is not well suited for event-driven schemes (there are no events!). The above results (Theorems 9.1, 9.2, 9.4) seem appropriate to characterize its approximated solutions. For the second DI, should one prefer an event-driven scheme using for instance Stewart's method or some kind of time-stepping scheme (Euler or θ-method as in Theorem 9.8 or a multistep method as in Theorem 9.11)?

9.2.4.1 The Selection Procedure

Consider $\dot{x}(t) \in [-1,1]$, $t \geqslant 0$, which clearly has an infinite number of solutions. Its discretization is $x_{k+1} = x_k + h\zeta_k$, $\zeta_k \in [-1,1]$. According to Theorem 9.1, one may choose at each step any ζ_k inside $[-1,1]$, and there always exists a solution of the DI that is in a neighborhood of the obtained discrete trajectory (the piecewise linear function $x^N(\cdot)$). In practice, however, one may want to approximate a particular solution inside the bundle of solutions of the DI. For instance the minimum norm selection may be a choice. For the multistep method this boils down to the algorithm:

$$\begin{array}{c} \min \frac{1}{2}\zeta_k^{\mathrm{T}}\zeta_k \\ \text{subject to } \frac{1}{h}\sum_{i=0}^{r} a_i x_{j-r+i} = \sum_{i=0}^{r} b_i \zeta_{j-r+i} \\ \zeta_k \in F(t_k, x_k) \end{array} \tag{9.47}$$

at each step k. The explicit Euler method corresponds to the choice $r = 1$, $b_0 = 1$, $b_1 = 0$, $a_0 = 1$, and $a_1 = -1$. Thus the algorithm in (9.47) becomes

$$\begin{array}{c} \min \frac{1}{2}\zeta_k^{\mathrm{T}}\zeta_k \\ \text{subject to } x_{k+1} - x_k = h\zeta_k \\ \zeta_k \in F(t_k, x_k) \end{array}, \tag{9.48}$$

which we may rewrite with $y_k = \frac{x_{k+1}-x_k}{h}$ as

$$\begin{array}{c} \min \frac{h}{2} y_k^{\mathrm{T}} y_k \\ \text{subject to } y_k \in F(t_k, x_k). \end{array} \tag{9.49}$$

If $F(t_k,x_k)$ contains $\{0\}$ then $y_k = 0$, so $x_{k+1} = x_k$. We see that y_k is nothing else but the projection of the vector $x = 0$ onto the set $F(t_k,x_k)$. We therefore arrive at the discrete inclusion $\frac{x_{k+1}-x_k}{h} = \mathrm{proj}_{F(t_k,x_k)}(0)$. Thus looking for the minimal norm solution is equivalent to solving $\dot{x}(t) = \mathrm{proj}_{F(t,x(t))}(0)$. If $F(t,x)$ is convex for each t and x, then the right-hand side is Lipschitz continuous, so that the DI boils down to an ODE with Lipschitz right-hand side (this is the case for Filippov's inclusions, which by construction always have a convex right-hand side). For instance the DI: $\dot{x}(t) \in [-1,1]$ with its minimal norm solution is simply the scalar ODE: $\dot{x}(t) = 0$. Consider now the DI: $\dot{x}(t) \in -\mathrm{sgn}(x(t))$. Then algorithm (9.47) boils down to finding $y_k \in -\mathrm{sgn}(x_k)$ at each step. If $x_k \neq 0$ then $y_k = 1$ or -1. If $x_k = 0$ then $y_k = 0$.[4] The minimum norm solution seems to be *the* solution of the DI, which has a unique solution for each initial condition. Such solutions are sometimes called the *slow solutions* (Brezis, 1973).

Remark 9.19. Starting from $\dot{x}(t) \in [-1,1]$ we arrive at $\dot{x}(t) = 0$ simply by adding a constraint on the norm of the solutions' variation. This may pose a philosophical issue concerning differential inclusions which possess infinitely many solutions (and thus are far from enjoying the uniqueness of solutions property). An apparently tiny

[4] In practice ε-layers have to be implemented to test the value zero.

constraint brings such a loose system into the simplest and most structured ODE one may imagine. This questions the usefulness of a DI like $\dot{x}(t) \in [-1,1]$. It is possible that in most of the practical situations, one may work a little more and find physical arguments that eliminate a lot (here an infinity, as $+\infty - 1 = +\infty$) of solutions. We may classify DIs into two main classes: loose DIs with (too) many solutions and structured DIs with unique solutions.

The next theorem continues Theorem 9.12 for the multistep method (9.27) and is taken from Dontchev & Lempio (1992) and Filippov (1967).

Theorem 9.20. *Let all the assumptions of Theorem 9.12 hold, and in addition let $b_r = 0$ and $F(\cdot,\cdot)$ be Hausdorff continuous. Then the sequence of continuous piecewise linear functions $\{x^N(\cdot)\}_{N\in \mathbb{N}}$, corresponding to minimal norm selections, contains a subsequence which converges uniformly to a continuously differentiable solution of the DI.*

It is interesting to see that selecting the minimal norm leads to selecting a particular solution that is continuously differentiable. The above example $\dot{x}(t) \in [-1,1]$ is an illustration. If one chooses any ζ_k at each step, then the solution of the discrete inclusion approximates some solution of the DI. When selecting the minimum norm ζ_k at each step, one finds $\dot{x}(t) = 0$ whose solutions are smooth. Other selections exist. The minimal variation selection is

$$\begin{array}{c} \min ||v_k - v_{k-1}|| \\ \text{subject to } \frac{1}{h}\sum_{i=0}^{r} a_i x_{j-r+i} = v_k \\ v_k \in \sum_{i=0}^{r} b_i F(t_{j-r+i}, x_{j-r+i}) \end{array}. \tag{9.50}$$

Then the following holds (Dontchev & Lempio, 1992) and still concerns the multistep method (9.27).

Theorem 9.21. *Let the set-valued mapping $F(\cdot,\cdot)$ be bounded, upper semicontinuous, with nonempty, closed, convex values $F(t,x)$ for all t and x, and be Lipschitz-continuous on $[0,T] \times \mathbb{R}^n$. Let $b_r = 0$, and the starting selections satisfy*

$$||v_{k+1} - v_k|| \leqslant hM, \ \ 0 \leqslant k \leqslant r-2$$

for some constant M independent of h. Then the sequence of piecewise linear continuous functions $(x^N(\cdot), v^N(\cdot))_{N\in\mathbb{N}}$ contains a subsequence which converges uniformly to a pair $(y(\cdot), v(\cdot))$, where $y(\cdot)$ is a solution of the DI, with Lipschitz-continuous derivative $v(\cdot)$.

9.2.4.2 Numerical Results

The above results are of a theoretical and general nature. This is not sufficient in practice. For instance it is well known that the explicit ($\theta = 0$) and implicit ($\theta = 1$) Euler methods may yield numerical results that differ significantly (Elliot, 1985; Stewart &

Anitescu, 2006). Explicit Euler methods, when applied to Filippov's systems, may yield very poor results with strong oscillations around the switching surface, see, e.g., Lempio (1992) and Figs. 1.12 and 1.13. Such numerically induced oscillations may be reduced with Runge–Kutta methods, see example 3.3 in Dontchev & Lempio (1992); however, they still are present. Finding a good selection procedure may improve a lot, which is not surprising in view of the above comments. Let us investigate on the classical scalar example $\dot{x}(t) \in -\mathrm{sgn}(x(t))$ discretized with a θ-method as $x_{k+1} - x_k \in h\,\mathrm{sgn}(x_k + \theta(x_{k+1} - x_k))$ of how oscillations appear. When $|x_k| \leqslant h$, the scheme yields $x_k + \theta(x_{k+1} - x_k) = 0$. Thus $x_{k+1} = \frac{1-\theta}{\theta} x_k$ is the approximation of the system in the neighborhood of $x = 0$. If $0 < \theta < \frac{1}{2}$, an oscillation occurs since $\frac{1-\theta}{\theta} > 1$: the discrete trajectory will quit the layer $|x_k| \leqslant h$, come back into it, and so on. For $\frac{1}{2} < \theta < 1$ one gets $\frac{1-\theta}{\theta} < 1$, so there will still be oscillations; however, they exponentially decay inside the boundary layer. Now the case $\theta = 1$ is the implicit method (or backward Euler) and no oscillations occur (see Fig. 1.19 for an illustration). We refer the reader to Sect. 9.7 for a study of the implicit case and some comments. Such conclusions about the discrepancy between the implicit and explicit cases were found for instance in (Elliot, 1985) and also in Stewart (1990). The simulation results in Stewart (1990) indicate that the multistep method may also produce oscillations around the switching surface.

Stewart's event-driven method or an implicit time-stepping method seems to be mandatory if one desires high accuracy and acceptable approximations of the derivative of the state trajectory seen as a Filippov's solution, even in cases where nonsmooth events are rare. In fact the oscillations around the switching surface correspond to having a multiplier that switches very rapidly between two discrete values (see Fig. 1.12), despite in theory the multiplier should keep a constant value. In the applications where one needs the value of the multiplier (e.g., for feedback design) this may be extremely problematic.

9.3 The Special Case of the Filippov's Inclusions

Though most of the foregoing results apply to a larger class of differential inclusions than Filippov's inclusions, the particular case of Filippov's inclusions deserves attention because on one hand they find many applications and on the other hand their specific structure allows one to derive specific schemes.

9.3.1 Smoothing Methods

In this section, let us consider a Filippov system with a switching surface $S = \{x \in \mathbb{R}^n \mid c(x) = 0\}$ of codimension one. The notation is the same as introduced in Sect. 7.1.1. The Filippov's notion of a solution says that

$$\dot{x}(t) \in \alpha f^+(x(t)) + (1-\alpha) f^-(x(t)) \tag{9.51}$$

with $\alpha \in [0,1]$ and where the two vector fields are as in (2.14). On one side of S one has $\alpha = 1$, and on the other side $\alpha = 0$. On an attractive surface S the vector field is a convex combination of both vector fields, tangent to S. As we saw in Sect. 7.1.1, such a system can be expressed with sign multifunctions, see (7.2). The idea behind smoothing is to replace the sign multifunction by an approximation, i.e., a single-valued function like a sigmoid. The advantage is that one obtains an ODE with a continuous right-hand side. The drawback is that this ODE may not be easily tractable numerically, because a good approximation around the switching surface implies a function with a large slope: the ODE may be stiff, and the gain is not clear. On the other hand avoiding the stiffness drawback is possible, however, at the price of a bad approximation. We may say that such smoothing methods should be avoided.

9.3.2 Switched Model and Explicit Schemes

Consider the case when S is of codimension one, i.e., (9.51). The switching surface S is thickened to a band S_ε with a thickness $\varepsilon > 0$. In practice the thickness parameter should be chosen small enough, to preserve the accuracy of the scheme. Let us denote $n(x) = \nabla c(x)$ the normal to S at x. We may then define the following subspaces:

$$\begin{aligned}
\mathscr{U} &= \{x \in \mathrm{IR}^n \mid n^{\mathrm{T}}(x) f^+(x) < 0 \text{ and } n^{\mathrm{T}}(x) f^-(x) > 0\}\\
\mathscr{Q} &= \{x \in \mathrm{IR}^n \mid n^{\mathrm{T}}(x) f^+(x) > 0 \text{ and } n^{\mathrm{T}}(x) f^-(x) < 0\}\\
\mathscr{T}_+ &= \{x \in \mathrm{IR}^n \mid n^{\mathrm{T}}(x) f^+(x) > 0 \text{ and } n^{\mathrm{T}}(x) f^-(x) > 0\}\\
\mathscr{T}_- &= \{x \in \mathrm{IR}^n \mid n^{\mathrm{T}}(x) f^+(x) < 0 \text{ and } n^{\mathrm{T}}(x) f^-(x) < 0\}.
\end{aligned}$$

One has $S_\varepsilon = \mathscr{U} \cup \mathscr{Q} \cup \mathscr{T}_+ \cup \mathscr{T}_-$. The subspace $\mathscr{U}$ is the attractive subspace, $\mathscr{Q}$ is the repulsive subspace, and $\mathscr{T}_+$ and $\mathscr{T}_-$ correspond to the part of the switching surface where trajectories cross the surface (transversal intersection between the trajectories and S). Following Leine & Nijmeijer (2004) we may choose α in (7.1) as

$$\alpha = \frac{n^{\mathrm{T}}(x) f^-(x) + \tau^{-1} c(x)}{n^{\mathrm{T}}(x) f^-(x) + f^+(x)} \tag{9.52}$$

for some $\tau > 0$, so that

$$\dot{c}(x(t)) = -\tau^{1} c(x(t)). \tag{9.53}$$

It is noteworthy that this choice is equivalent to smoothing the sign function at 0. To see this consider the case when $c(x) = x$. Then (9.53) becomes $\dot{x}(t) = -\tau^{-1} x(t)$ when $x(t)$ lies in $\mathscr{U} \subset S_\varepsilon = [-\varepsilon, \varepsilon]$. Thus the multivalued sign function is replaced by a single-valued linear function with slope τ^{-1} and $\tau = \varepsilon$ (hence the notation τ^{-1} in (9.52) and (9.53)). The algorithm, called the *switch model* in Leine & Nijmeijer (2004), is described in Algorithm 6.

Certainly the strongest drawback of this method is that when the number of switching surfaces that define the Filippov's inclusion increases, the complexity of the procedure becomes an obstacle for the implementation. We may say that this

Algorithm 6 Leine's switch model

```
Require: $\mathsf{t}$ time instant
Require: $\mathsf{x}$ state at time $\mathsf{t}$
Require: $f^{+}(\cdot,\cdot), f^{-}(\cdot,\cdot), c(\cdot), n = \nabla c(\cdot)$
Require: $\varepsilon$ thickness parameter
Ensure: $\mathsf{y}$ value of the ode r.h.s. i.e. $\mathsf{y} = \mathsf{rhs}(\mathsf{x},\mathsf{t})$

  $\mathsf{h} \leftarrow \mathsf{c}(\mathsf{x})$
  $\mathsf{n} \leftarrow \nabla \mathsf{c}(\mathsf{x})$
  $\mathsf{f}^{+} \leftarrow \mathsf{f}^{+}(\mathsf{t},\mathsf{x})$
  $\mathsf{f}^{-} \leftarrow \mathsf{f}^{-}(\mathsf{t},\mathsf{x})$
  if $|\mathsf{c}| \geqslant \eta$ then
    // Smooth motion
    if $\mathsf{h} > \varepsilon$ then
      $\mathsf{y} \leftarrow \mathsf{f}^{+}$
    else
      $\mathsf{y} \leftarrow \mathsf{f}^{-}$
    end if
  else
    if $\mathsf{n}^{\mathsf{T}}\mathsf{f}^{+} > 0$ and $\mathsf{n}^{\mathsf{T}}\mathsf{f}^{-} > 0$ then
      // transition
      $\mathsf{y} \leftarrow \mathsf{f}^{+}$
      $\mathsf{x} \in \mathscr{T}_{+}$
    end if
    if $\mathsf{n}^{\mathsf{T}}\mathsf{f}^{+} < 0$ and $\mathsf{n}^{\mathsf{T}}\mathsf{f}^{-} < 0$ then
      // transition
      $\mathsf{y} \leftarrow \mathsf{f}^{-}$
      $\mathsf{x} \in \mathscr{T}_{-}$
    end if
    if $\mathsf{n}^{\mathsf{T}}\mathsf{f}^{+} < 0$ and $\mathsf{n}^{\mathsf{T}}\mathsf{f}^{-} > 0$ then
      // attractive sliding mode
      $\alpha \leftarrow \dfrac{\mathsf{n}^{\mathsf{T}}\mathsf{f}^{-} + \tau^{-1}\mathsf{h}}{\mathsf{n}^{\mathsf{T}}(\mathsf{f}^{-} + \mathsf{f}^{+}}$
      $\mathsf{y} \leftarrow \alpha\mathsf{f}^{+} + (1-\alpha)\mathsf{f}^{-}$
      $\mathsf{x} \in \mathscr{U}$
    end if
    if $\mathsf{n}^{\mathsf{T}}\mathsf{f}^{+} > 0$ and $\mathsf{n}^{\mathsf{T}}\mathsf{f}^{-} < 0$ then
      // repulsive sliding mode
      $\mathsf{y} \leftarrow \mathsf{f}^{+}$ or $\mathsf{y} = \mathsf{f}^{-}$
      $\mathsf{x} \in \mathscr{Q}$
    end if
  end if
```

method works when there are no more than two switching surfaces. Otherwise, Stewart's event-driven method in Sect. 7.1.2 or some time-stepping method of Sect. 9.2 should be chosen. We will see in the next section that in the case of mechanical systems with Coulomb friction, a method that solves the selection problem of the multiplier and of the mode with a suitable LCP can also be advantageously chosen.

9.3.3 Implicit Schemes and Complementarity Formulation

Let us consider a particular case of Filippov's inclusions, known in the systems and control literature as *linear relay systems* (Camlibel, 2001). Their dynamics takes the form

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ y(t) = Cx(t) + Du(t) \\ u_i(t) \in \mathrm{sgn}(-y_i(t)), \;\; 1 \leqslant i \leqslant m \end{cases} \tag{9.54}$$

with $u(t) \in \mathrm{I\!R}^m$, $x(t) \in \mathrm{I\!R}^n$, $y(t) \in \mathrm{I\!R}^m$. It should be clear from the material and the manipulations done for instance in Sect. 1.2 that (9.54) can be rewritten as a complementarity system. The most immediate way to discretize (9.54) is with a backward Euler scheme, i.e.,

$$\begin{cases} x_{k+1} = x_k + hAx_{k+1} + hBu_{k+1} \\ y_{k+1} = Cx_{k+1} + Du_{k+1} \\ u_{i,k+1} \in \mathrm{sgn}(-y_{i,k+1}), \;\; 1 \leqslant i \leqslant m \end{cases} . \tag{9.55}$$

Therefore the advance from step k to step $k+1$ requires to solve

$$\begin{cases} u_{k+1} = [C(I_n - hA)^{-1}hB + D]^{-1}[y_{k+1} - C(I_n - hA)^{-1}x_k] \\ u_{i,k+1} \in \mathrm{sgn}(-y_{i,k+1}), \;\; 1 \leqslant i \leqslant m \end{cases} , \tag{9.56}$$

that is to find the intersection(s) between the graph of a single-valued and of a multivalued function. One convenient way to do this is to rewrite the sign multifunction as a complementarity problem, then inject this into the dynamics, and end up with an LCP to be solved numerically. Actually, the inclusions $u_i(t) \in \mathrm{sgn}(-y_i(t))$ for $1 \leqslant i \leqslant m$ are equivalent to the complementarity problem:

$$\begin{cases} \lambda^1 = \frac{1}{2}(e - u) \\ y^2 = \frac{1}{2}(e + u) \\ -y = y^1 - \lambda^2 \\ \\ 0 \leqslant \lambda^1 \perp y^1 \geqslant 0 \\ \\ 0 \leqslant \lambda^2 \perp y^2 \geqslant 0 \end{cases} , \tag{9.57}$$

where $e = (1,1,...,1)^{\mathrm{T}}$ is of appropriate dimension. One has $u = e - 2\lambda^1$ and $y^2 = e - \lambda^1$. Inserting (9.57) into (9.54) and after some manipulations we obtain the following:

$$\begin{cases} \dot{x}(t) = Ax(t) + Be - 2B\lambda^1(t) \\ \begin{pmatrix} y^1 \\ y^2 \end{pmatrix} = \begin{pmatrix} -Cx(t) - De \\ e \end{pmatrix} + \begin{pmatrix} 2D & I_m \\ -I_m & 0_m \end{pmatrix} \begin{pmatrix} \lambda^1 \\ \lambda^2 \end{pmatrix} , \\ 0 \leqslant \lambda^1 \perp y^1 \geqslant 0 \\ 0 \leqslant \lambda^2 \perp y^2 \geqslant 0 \end{cases} \tag{9.58}$$

which may be named an *affine complementarity system* (Brogliato, 2003). After some manipulations one obtains the LCP

$$\begin{cases} y^1_{k+1} = -(D + C(I_n - hA)^{-1}hB)e - C(I_n - hA)^{-1}x_k \\ \qquad +2(D + C(I_n - ha)^{-1}hB)\lambda^1_{k+1} + \lambda^2_{k+1} \\ y^2_{k+1} = e - \lambda^1_{k+1} \qquad\qquad\qquad\qquad . \\ 0 \leqslant \begin{pmatrix} y^1_{k+1} \\ y^2_{k+1} \end{pmatrix} \perp \begin{pmatrix} \lambda^1_{k+1} \\ \lambda^2_{k+1} \end{pmatrix} \geqslant 0 \end{cases} \tag{9.59}$$

Provided this LCP can be solved for some $(\lambda^1_{k+1}, \lambda^2_{k+1})$, the algorithm may be advanced as $x_{k+1} = (I_n - hA)^{-1}x_k + (I_n - hA)^{-1}hBe - 2h(I_n - hA)^{-1}B\lambda^1_{k+1}$. It should be obvious that a necessary and sufficient condition for solvability of this LCP for any x_k is that its matrix be a P-matrix. A convergence result is given (Camlibel, 2001), showing the consistency of the method.

Theorem 9.22. *Consider the system in (9.54) and suppose that the rational transfer function $C(sI_n - A)^{-1}B + D$ is a P-matrix for all sufficiently large $s \in \mathbb{R}$, and that $D \geqslant 0$. Let the sequence of time steps $\{h_k\}_{k \geqslant 0}$ converge to zero. Consider the piecewise functions $(x^N(\cdot), \lambda^N(\cdot), y^N(\cdot))$ that approximate $(x(\cdot), \lambda(\cdot), y(\cdot))$ on the interval of integration $[0, T]$. The following holds for any sequence $\{h_k\}_{k \geqslant 0}$ that converges to zero:*

(a) There exists a subsequence $\{h_{k_i}\}_{k_i \geqslant 0}$ of $\{h_k\}_{k \geqslant 0}$ such that $\{\lambda^N_{k_i}(\cdot), y^N_{k_i}(\cdot)\}$ converges weakly in $\mathscr{L}_2$ to some limit functions $(\lambda(\cdot), y(\cdot))$ and $\{x^N_{k_i}(\cdot)\}$ converges in $\mathscr{L}_2$ to some limit function $x(\cdot)$.
(b) The limit triple $(x(\cdot), \lambda(\cdot), y(\cdot))$ is a solution of the dynamical system (9.54).
(c) In case of uniqueness of the solutions of (9.54), the results of item (a) hold for the complete sequence $\{h_k\}_{k \geqslant 0}$.

Uniqueness of solutions depends on the quadruple (A, B, C, D).

Remark 9.23. the very big difference between such a backward Euler scheme, and an explicit discretization as in (9.14) is to be noticed. The implicit way may be named a *dual way* to discretize, as it is equivalent to look for Lagrange multipliers λ_{k+1} at each step. The mode switches are monitored by the LCP, i.e., by the multiplier value. They are not monitored by the state x_k. The convergence results are of the same nature; however, the practical implementations and results differ a lot. Theorem 9.8 and the θ-method apply to such relay systems when $D = 0$. However, one notes that the order of convergence result of Theorem 9.8 is obtained under the assumption that the trajectories are piecewise differentiable. This is not needed in Theorem 9.22.

9.3.4 Comments

Consider the Filippov system in (7.5). Using the complementarity formulation of the multivalued sign function, one may rewrite this system as a complementarity system. One may then proceed to discretizing this CS as done in the foregoing subsection for the relay systems. Similar to Stewart's event-driven method, all relies on a suitable way to rewrite the inclusion.

9.4 Moreau's Catching-Up Algorithm for the First-Order Sweeping Process

In this section, we give some details about the seminal work of Jean Jacques Moreau on the numerical time integration of the sweeping process. Let us consider the first-order sweeping process with a BV solution:

$$\begin{cases} -\mathrm{d}x \in N_{K(t)}(x(t^+)) \;\; (t \geqslant 0) \\ x(0) = x_0 \end{cases} . \tag{9.60}$$

When the solution is absolutely continuous, then $\mathrm{d}x = \dot{x}(t)\mathrm{d}t$, and since the right-hand side is a cone, the left-hand side may be simplified to $-\dot{x}(t)$. Under suitable hypothesis on the multivalued function $t \mapsto K(t)$, numerous convergence and consistency results (Monteiro Marques, 1993, Kunze & Monteiro Marqués, 2000) have been given together with well-posedness results, using the so-called "catching-up algorithm" defined in Moreau (1977):

$$-(x_{k+1} - x_k) \in \partial\psi_{K(t_{k+1})}(x_{k+1}), \tag{9.61}$$

where x_k stands for the approximation of the right limit of $x(\cdot)$. It is noteworthy that the case with a Lipschitz-continuous moving set is also discretized in the same way.

By elementary convex analysis (see Sect. A.3), the inclusion (9.61) is equivalent to

$$x_{k+1} = \mathrm{prox}[K(t_{k+1}); x_k]. \tag{9.62}$$

Contrary to the standard backward Euler scheme with which it might be confused, the catching-up algorithm is based on the evaluation of the measure dx on the interval $(t_k, t_{k+1}]$, i.e., $\mathrm{d}x((t_k, t_{k+1}]) = x^+(t_{k+1}) - x^+(t_k)$. Indeed, the backward Euler scheme is based on the approximation of $\dot{x}(t)$ which is not defined in a classical sense for our case. When the time step vanishes, the approximation of the measure dx tends to a finite value corresponding to the jump of $x(\cdot)$. This remark is crucial for the consistency of the scheme. Particularly, this fact ensures that we handle only finite values.

Figure 9.6 depicts the evolution of the discretized sweeping process. The name *catching-up* is clear from the figure: the algorithm makes x_k catch up with the moving set $K(t_k)$, so that it stays inside the moving set.

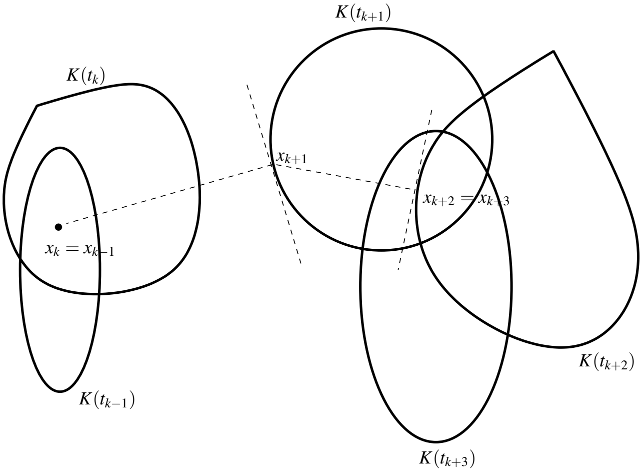


Fig. 9.6. The catching-up algorithm

9.4.1 Mathematical Properties

It is noteworthy that the catching-up algorithm is a central tool to prove Theorems 2.37 and 2.39. As a consequence, some properties of the catching-up algorithm, that is a discretized version of the first-order sweeping process, can be deduced directly from these well-posedness proofs. We give below a brief account of the properties of the discretized sweeping process. More may be found in Monteiro Marques (1993) and Kunze & Monteiro Marqués (2000). Let us first deal with the Lipschitz sweeping process.

Theorem 9.24. *Suppose that the conditions of Theorem 2.37 are satisfied. Consider the algorithm in (9.61), with a fixed time step $h = \frac{T}{N} > 0$. Let $m \in \mathbb{N}$ be such that $mT < N$. Then*

(a) $var_{[0,T]}(x^N) \leqslant ||x^N(0)|| + lT$, for all $t \in [t_k, t_{k+1}]$ and all $N \in \mathbb{N}$,

(b) $||x^N(t) - x^N(s)|| \leqslant l\left(|t-s| + \frac{2}{m}\right)$, for all $t, s \in [t_k, t_{k+1}]$,

(c) from which it follows that $||x(t) - x(s)|| \leqslant l|t-s|$ for all $t, s \in [0, T]$, where $(x(t) - x(s))$ is the limit in the weak sense of $\{x^N(t) - x^N(s)\}_{N \in \mathbb{N}}$,

(d) $||\dot{x}^N(t)|| \leqslant l$ for all $t \neq t_k$, where $\dot{x}^N(t) = \frac{1}{h}(x_{k+1} - x_k)$ for $t \in [t_k, t_{k+1})$,

(e) the "velocity" $\dot{x}^N(\cdot)$ converges weakly to $\dot{x}^*(\cdot)$, i.e., for all $\varphi(\cdot) \in \mathscr{L}^1([0,T]; \mathbb{R}^n)$ one has

$$\int_0^T \langle \dot{x}^N(t), \varphi(t) \rangle dt \to \int_0^T \langle \dot{x}^*(t), \varphi(t) \rangle dt,$$

(f) $x^N(\cdot) \to x(\cdot)$ uniformly and $\dot{x}(\cdot) = \dot{x}^*(\cdot)$ almost everywhere in $[0, T]$,

(g) the limit satisfies $\dot{x}(t) \in N_{K(t)}(x(t))$ almost everywhere in $[0, T]$.

Theorem 2.37 is proved thanks to the discrete inclusion analysis. This explains why the upper bounds in Theorem 9.24 are rather rough and do not allow to conclude on the order of the method. These upper bounds are sufficient for step (g) to be achieved. One may say that the works using the discrete-time sweeping process are oriented toward mathematical analysis rather than numerical analysis. In the RCBV case, the catching-up algorithm may be used also to prove Theorem 2.39, with similar steps as in Theorem 9.24.

Remark 9.25. Using higher order numerical schemes is at best useless, more often it is dangerous when solutions possess discontinuities (hence a measure differential inclusion). Basically, a general way to obtain a finite difference-type scheme of order n is to write a Taylor expansion of order n or higher. Such a scheme is meant to approximate the nth derivative of the discretized function. If the solution we are dealing with is obviously not differentiable, what is the meaning of using a scheme with order $n \geqslant 2$? Such a scheme will try to approximate derivatives which do not exist. At the times of nondifferentiability, it may introduce in the solution artificial unbounded terms creating oscillations, etc., see Vola et al. (1998). In summary, higher order numerical schemes are inadequate for time-stepping discretization of NSDS with state jumps, i.e., in particular measure differential inclusions as (9.60). To say nothing of DIs whose solutions are distributions of higher degree, see Chaps. 5 and 11.

Remark 9.26. Many of the results which are stated in this chapter, say that one may extract a subsequence from the sequence $\{x^N(\cdot)\}_{N\in\mathbb{N}}$, that converges towards a limit that is a solution of the continuous time DI. Roughly speaking, the proofs are made of three main steps: (i) derive various upper bounds on the solution of the discrete inclusion and its derivative and their variations (steps (a)–(d) in Theorem 9.24), (ii) use various forms of the Arzela–Ascoli, Banach–Alaoglu theorems, which allow one to extract convergent subsequences (steps (e) and (f) in Theorem 9.24), (iii) prove that the limit of the subsequence is a solution of the continuous-time DI (step (g) in Theorem 9.24). Step (iii) is not the most easy one.

9.4.2 Practical Implementation of the Catching-up Algorithm

The catching-up algorithm as stated in (9.62) cannot be directly implemented. In order to achieve this, one needs to perform further steps. Suppose that $K(t) = \{x \in \mathbb{R}^n \mid Cx + D(t) \geqslant 0\}$ for some constant matrix $C \in \mathbb{R}^{m\times n}$ and a time-varying vector $D(t) \in \mathbb{R}^m$. Using Proposition A.3 we may rewrite the sweeping process as the LCS

$$\begin{cases} -\dot{x}(t) = C^T\lambda(t) \\ \\ 0 \leqslant \lambda(t) \perp Cx(t) + D(t) \geqslant 0 \end{cases} \tag{9.63}$$

where the multiplier $\lambda(t)$ is the equivalent of a selection as in ordinary differential inclusions. The time-discretization of this gradient LCS with a backward Euler scheme gives

$$\begin{cases} x_{k+1} - x_k = hC^T\lambda_{k+1} \\ \\ 0 \leqslant \lambda_{k+1} \perp Cx_{k+1} + D_{k+1} \geqslant 0 \end{cases} \tag{9.64}$$

After few manipulations one obtains the following LCP

$$0 \leqslant \lambda_{k+1} \perp hCC^T \lambda_{k+1} + Cx_k + D_{k+1} \geqslant 0 \tag{9.65}$$

whose unknown is λ_{k+1} and whose matrix is the symmetric semi-positive definite matrix hCC^T. It is noteworthy that usually one may have m much larger than n, so that CC^T is not full rank. Some care has to be taken in order to solve (9.65). One may have a look at Sect. 12.4 for an overview of complementarity problems and their solvers. In particular QP solvers may be quite useful in this symmetric PSD context, see Sect. 12.4.5.

Remark 9.27. Recall from Sect. 4.2.3 that in case a positive real constraint is imposed, close relationships exist between the sweeping process and LCS. In such a case the LCS may be recast after a suitable state vector change, into gradient complementarity systems, which in turn may be rewritten as an inclusion into a normal cone. We therefore conclude that the first-order sweeping process discretization, and the material that follows in Sect. 9.5, are very similar. The discrepancies are that the catching-up algorithm does not consider nonzero vector fields; however, it does not need Assumptions 13 and 14, and the complementarity function $w(\cdot)$ may depend explicitly on time.

9.4.3 Time-Independent Convex Set *K*

In the case of a time-independent convex set K, the sweeping process (9.60) is of poor interest. The solution can only have a jump at the initial time if the initial condition does not satisfy the inclusion into K, i.e., $x_0 \notin K$. In this case, the problem can be viewed as a sweeping with a convex set which moves at the initial time from a convex set containing x_0, to K.

The case of a time-independent convex set is more interesting if nontrivial terms are added. Let us recall now the UDI (2.47), i.e.,

$$-(\dot{x}(t) + f(x(t)) + g(t)) \in N_K(x(t)), \quad x(0) = x_0. \tag{9.66}$$

Mimicking (9.61), the inclusion can be discretized as

$$-(x_{k+1} - x_k) + h(f(x_{k+1}) + g(t_{k+1})) = \mu_{k+1} \in N_K(x_{k+1}). \tag{9.67}$$

In this discretization, we have chosen to evaluate the measure dx by the approximated value μ_{k+1}. If the initial condition does not satisfy the inclusion at the initial time, the jump in the state can be treated in a consistent way.

If the constant set is equal to $K = \mathbb{R}^n_+$, the previous problem can be written as a nonlinear complementarity problem:

$$\begin{cases} (x_{k+1} - x_k) - h(f(x_{k+1}) + g(t_{k+1})) = \mu_{k+1} \\ \\ 0 \leqslant x_{k+1} \perp \mu_{k+1} \geqslant 0 \end{cases} \tag{9.68}$$

and if the term $f(x)$ is linear, i.e., $f(x) = Ax$, we obtain the following LCP(q,M):

$$\begin{cases} (I_n - hA)x_{k+1} - (x_k + hg(t_{k+1})) = \mu_{k+1} \\ \\ 0 \leqslant x_{k+1} \perp \mu_{k+1} \geqslant 0 \end{cases} \tag{9.69}$$

with $M = (I_n - hA)$ and $q = -(x_k + hg(t_{k+1}))$. It is noteworthy that the value μ_{k+1} approximates the measure $\mathrm{d}\lambda$ on the time interval $[t_k, t_{k+1})$ rather than directly the value of λ, where λ is the multiplier of the LCS, see Sect. 2.6.

Remark 9.28. We will see later in Sect. 9.5 that the discretization proposed in Camlibel et al. (2002a) for LCS is very similar to this discretization, in particular, if the set K is polyhedral described by

$$K = \{x \in \mathrm{IR}^n \mid Cx \geqslant 0\}. \tag{9.70}$$

If a constraint qualification holds, the DI (9.66) in the linear case $f(x) = -Ax$ is equivalent to the following LCS:

$$\begin{cases} \dot{x}(t) = Ax(t) + C^{\mathrm{T}}\lambda(t) \\ \\ w(t) = Cx(t) \\ \\ 0 \leqslant w(t) \perp \lambda(t) \geqslant 0. \end{cases} \tag{9.71}$$

In this case, the catching-up algorithm yields

$$\begin{cases} x_{k+1} - x_k = hAx_{k+1} + C^{\mathrm{T}}\mu^{k+1} \\ \\ w_{k+1} = Cx_{k+1} \\ \\ 0 \leqslant w_{k+1} \perp \mu_{k+1} \geqslant 0. \end{cases} \tag{9.72}$$

9.5 Linear Complementarity Systems with $r \leqslant 1$

What is presented in this section is very close to what has been presented for relay systems in Sect. 9.3.3. In Camlibel et al. (2002a), Heemels (1999), and Camlibel (2001), a backward Euler time-stepping method is designed for LCS of the form

$$\begin{cases} \dot{x}(t) = Ax(t) + B\lambda(t) \\ \\ w(t) = Cx(t) + D\lambda(t) \\ \\ 0 \leqslant w(t) \perp \lambda(t) \geqslant 0 \\ \\ x(0) = x_0 \end{cases} \tag{9.73}$$

with $x(t) \in \mathrm{I\!R}^n$, $w(t) \in \mathrm{I\!R}^m$. The following fundamental assumption is made: the quadruple (A,B,C,D) is passive (see, e.g., Brogliato et al., 2007), i.e., there exists $P = P^{\mathrm{T}} \geqslant 0$ such that

$$\begin{pmatrix} A^{\mathrm{T}}P + PA & PB - C^{\mathrm{T}} \\ B^{\mathrm{T}}P - C & -D - D^{\mathrm{T}} \end{pmatrix} \leqslant 0. \tag{9.74}$$

In case the pair (C,A) is observable and the pair (A,B) is controllable, then there exists $P > 0$ that solves (9.74). It can be deduced from this linear matrix inequality that $D \geqslant 0$ and $A^{\mathrm{T}}P + PA \leqslant 0$. If $D = 0$ then it follows that $PB = C^{\mathrm{T}}$, so that $CB = B^{\mathrm{T}}PB \geqslant 0$. The condition that B has full column rank implies that $CB > 0$. In other words, the leading Markov parameter of the system (A,B,C,D) is under a dissipativity assumption either D of CB. Let $m = 1$: the relative degree r between y and λ is equal to 0 (if $D > 0$) or 1 ($D = 0$ and $CB > 0$). When $m > 1$ things become more complex; however, under some additional assumptions the case $m = 1$ can be generalized with a so-called vector relative degree $r \in \mathrm{I\!R}^m$ (see Sect. 4.3). The fact that the relative degree satisfies $r = 0$ or $r = 1$ is extremely important to understand the dynamics of the LCS (9.73). This means that apart possibly at the initial time, solutions will be continuous, see Chaps. 4 and 5 for details.

A backward Euler scheme is applied to evaluate the time derivative $\dot{x}(\cdot)$ leading to the following scheme:

$$\begin{cases} \dfrac{x_{k+1} - x_k}{h} = Ax_{k+1} + B\lambda_{k+1} \\ \\ w_{k+1} = Cx_{k+1} + D\lambda_{k+1} \\ \\ 0 \leqslant \lambda_{k+1} \perp w_{k+1} \geqslant 0 \end{cases}, \tag{9.75}$$

which can be reduced to an LCP by a straightforward substitution:

$$0 \leqslant \lambda_{k+1} \perp C(I_n - hA)^{-1}x_k + (hC(I_n - hA)^{-1}B + D)\lambda_{k+1} \geqslant 0. \tag{9.76}$$

In the sequel, such an LCP will be denoted as (w_{k+1}, λ_{k+1}) =LCP(M, b_{k+1}) where

$$M = hC(I_n - hA)^{-1}B + D, \tag{9.77}$$

$$b_{k+1} = C(I_n - hA)^{-1}x_k. \tag{9.78}$$

The scheme to compute the approximating piecewise continuous solutions is developed in Algorithm 7.

The next result is taken from Camlibel et al. (2002a). The solutions of the continuous-time LCS are as in Chap. 5, see Theorem 4.7. Therefore they are smooth everywhere on $\mathrm{I\!R}^+$, except possibly at the initial time (when $D > 0$ there is no initial jump). If there is an initial jump, then the multiplier λ is a Dirac measure at time $t = 0$. Let us make the following assumptions:

Algorithm 7 Time-stepping for passive LCS

```
Require: $t_0, T$
Require: $x_0$ initial data
Require: h time–step
Ensure: $x_k, w_k, \lambda_k$ solution of (9.75)
  $k \leftarrow 0$
  $W \leftarrow (I_n - hA)^{-1}$
  $M \leftarrow hCWB + D$
  while $t_k < T$ do
    $b_{k+1} \leftarrow CWx_k$
    $w_{k+1}, \lambda_{k+1} \leftarrow$ solution of $LCP(M, b_{k+1})$
    $x_{k+1} \leftarrow W(x_k + hB\lambda_{k+1})$
    $t_{k+1} \leftarrow t_k + h$
    $k \leftarrow k+1$
  end while
```

Assumption 13. *There exists $h^* > 0$ such that for all $h \in (0, h^*)$ the LCP(M, b_{k+1}) has a unique solution for all b_{k+1}.*

Assumption 14. *The system (A,B,C,D) is minimal (the pair (A,B) is controllable, the pair (C,A) is observable) and B is of full column rank.*

In particular the approximation of the Dirac measure at $t = 0$ is given by $h\lambda_0\delta_0$. Assumption 13 secures that Algorithm 7 generates a unique output at each step, for $h > 0$ small enough. Let us examine the algorithm in a particular case.

Example 9.29. Let us choose a scalar system, with $A = B = C = 1$, $D = 0$, and the initial data $x(-1) = -1$. Then we get $x_{k+1} = -(1+h)^{-1}[h\lambda_{k+1} + x_k]$, and the LCP$(M, b_{k+1})$ is $0 \leqslant \lambda_{k+1} \perp (1+h)^{-1}x_k + h(1+h)^{-1}\lambda_{k+1} \geqslant 0$. The algorithm is initialized at step $k = -1$ with $x^N(-1) = x(0) = -1$. Simple calculations yield

- $k = -1$: $\lambda_0 = \frac{1}{h}$ and $x_0 = 0$,
- $k = 0$: $\lambda_1 = 0$ and $x_1 = 0$,
- $k \geqslant 1$: $\lambda_k = 0$ and $x_k = 0$.

Therefore the algorithm produces an initial jump from $x^N(-1) = -1$ to $x^N(0) = 0$, and after this the solution remains stuck at $x_k = 0$. We notice that $\lambda_0 \rightarrow +\infty$ as $h \rightarrow 0$; however, the value $h\lambda_0 = 1$ is bounded. The obtained solution is quite consistent with the jump rule in (4.9), (4.10), or (4.11). We have $Q_D = Q_0 = \mathbb{R}^+$ and $Q_D^* = \mathbb{R}^+$. Thus (4.9) is the LCP: $0 \leqslant \lambda_0 \perp x(0^-) + \lambda_0 \geqslant 0$, whose solution is $\lambda_0 = 1$ with the above choice of the initial data. From Theorem 4.7 the initial jump is such that $x(0^+) = x(0^-) + \lambda_0 = 0$. The numerical method produces the right initial jump.

Let us now state a convergence result (Camlibel et al., 2002a). The interval of integration is $[0, T]$, $T > 0$. The convergence is understood as $\lim_{h\rightarrow 0}\langle x^N(t) - x(t), \varphi(t)\rangle = 0$ for all $\varphi \in \mathscr{L}^2([0,T]; \mathbb{R}^n)$ and all $t \in [0, T]$, which is the weak convergence in $\mathscr{L}^2([0,T]; \mathbb{R}^n)$.

Theorem 9.30. *Consider the LCS in (9.73) with $D \geqslant 0$ and let Assumption 13 hold. Let $(\lambda_k^N, x_k^N, w_k^N)$ be the output of Algorithm 7, with the initial impulsive term being approximated by $(h\lambda_0, hx_0, hw_0)$. Assume that there exists a constant $\alpha > 0$ such that for $h > 0$ small enough,one has $||h\lambda_0|| \leqslant \alpha$ and $||\lambda_k^N|| \leqslant \alpha$ for all $k \geqslant 0$. Then for any sequence $\{h_k\}_{k\geqslant 0}$ that converges to zero, one has the following:*

(i) There exists a subsequence $\{h_{k_l}\} \subseteq \{h_k\}_{k\geqslant 0}$ such that $(\{\lambda^N\}_{k_l}, \{w^N\}_{k_l})$ converges weakly to some (λ, w) and $\{x^N\}_{k_l}$ converges to some $x(\cdot)$.

(ii) The triple $(\lambda, x(\cdot), w)$ is a solution of the LCS in (9.73) on $[0,T]$ with initial data $x(0) = x_0$.

(iii) If the LCS has a unique solution for $x(0) = x_0$, the whole sequence $(\{\lambda^N\}_k, \{w^N\}_k)$ converges weakly to (λ, w) and the whole sequence $\{x^N\}_k$ converges to $x(\cdot)$.

If the quadruple (A,B,C,D) is such that Assumption 14 holds and is passive, then (iii) holds.

We emphasize the notation $x(\cdot)$ since the solutions are functions of time, whereas the notation λ and w means that these have to be considered as measures. To be quite rigorous we should have chosen such a notation in (9.73).

Remark 9.31. As seen on the simple example, the initial value of the multiplier λ may be unbounded. If λ has a unique atom at $t = 0$, this is not very bothering for the implementation. However, if the system undergoes repeated state jumps (when for instance an external excitation is present as in (4.7)), this becomes untractable in practice. Then, as we already said in Chap. 1, a better way to implement the algorithm is to calculate $h\lambda_k$, and not λ_k. This is what is done in the time-discretization of the first-order sweeping process.

In the case $D > 0$ (relative degree 0), the LCS is equivalent to a standard system of ODEs with a Lipschitz-continuous vector field (see Goeleven & Brogliato, 2004, remark 10).[5] The result of convergence is then the standard result of convergence for the Euler backward scheme. In the case $D \geqslant 0$ (if $m = 1$ this is a relative degree equal to 1), the initial condition must satisfy the unilateral constraints $w_0 = Cx_0 \geqslant 0$. Otherwise, the approximation $\frac{x_{k+1} - x_k}{h}$ has no chance to converge if the state possesses a jump.

Remark 9.32. Following Remark 9.28, we can note some similarities with the catching-up algorithm. Two main differences have, however, to be noted:

- The first one is that the sweeping process can be equivalent to an LCS under the condition $C = B^{\mathrm{T}}$. In this way, the previous time-stepping scheme extends the catching-up algorithm to more general systems.
- The second major discrepancy is as follows. The catching-up algorithm does not approximate the time derivative $\dot{x}(\cdot)$ as

$$\dot{x}(t) \approx \frac{x(t+h) - x(t)}{h}, \tag{9.79}$$

[5] As a simple consequence of (A.8) or of Theorem B.3.

but it approximates the measure of the time interval by

$$\mathrm{d}x((t,t+h]) = x^+(t+h) - x^+(t). \tag{9.80}$$

This difference leads to a consistent time-stepping scheme if the state possesses an initial jump. A direct consequence is that the primary variable μ_{k+1} in the catching-up algorithm is homogeneous to a measure of the time interval.

We will see in Chap. 11 some examples of systems with relative degree $r \geqslant 3$ where the scheme (9.75)–(9.76) does not work at all due to the fact that the solutions are strongly nonsmooth (they are distributions of degree $\geqslant 2$, see Chap. 5).

Remark 9.33. In the case of a relative degree 0, the following scheme based on a θ-method ($\theta \in [0,1]$) should also work:

$$\begin{cases} \dfrac{x_{k+1}-x_k}{h} = A(\theta x_{k+1} + (1-\theta)x_k) + B(\theta\lambda_{k+1} + (1-\theta)\lambda_k) \\ w_{k+1} = Cx_{k+1} + D\lambda_{k+1} \\ 0 \leqslant \lambda_{k+1} \perp w_{k+1} \geqslant 0 \end{cases} \tag{9.81}$$

because a continuously differentiable trajectory is expected. It has been successfully tested on nonsmooth electrical circuits of relative degree 0, in the semi-implicit case $\theta \in [1/2, 1]$ (Denoyelle & Acary, 2006). An interesting feature of such θ-methods is the energy-conserving property that they exhibit for $\theta = 1/2$. We will see in the following section that the scheme can be viewed as a special case of the time-stepping scheme proposed in Pang & Stewart (in press).

9.6 Differential Variational Inequalities

In Pang & Stewart (in press), several time-stepping schemes are designed for DVI which are separable in λ:

$$\dot{x}(t) = f(t,x(t)) + B(x(t),t)\lambda(t), \tag{9.82}$$

$$\lambda(t) = \mathrm{SOL}(K, G(t,x(t)) + F(\cdot)). \tag{9.83}$$

We recall that the second equation means that $\lambda(t) \in K$ is the solution of the following VI:

$$(v - \lambda(t))^{\mathrm{T}}(G(t,x(t)) + F(\lambda(t))) \geqslant 0, \forall v \in K. \tag{9.84}$$

Two cases are treated with a time-stepping scheme: the initial value problem (IVP) and the boundary value problem (BVP).

9.6.1 The Initial Value Problem (IVP)

Let us start with the initial value problem:

$$\begin{cases} \dot{x}(t) = f(t,x(t)) + B(x(t),t)\lambda(t) \\ \lambda(t) = \mathrm{SOL}(K, G(t,x(t)) + F(\cdot)) \\ x(0) = x_0 \end{cases} . \tag{9.85}$$

The proposed time-stepping method is given as follows:

$$x_{k+1} - x_k = h\,[f(t_k, \theta x_{k+1} + (1-\theta)x_k) + B(x_k,t_k)\lambda_{k+1}], \tag{9.86}$$

$$\lambda_{k+1} = \mathrm{SOL}(K, G(t_{k+1},x_{k+1}) + F(\cdot)). \tag{9.87}$$

If $\theta = 0$, an explicit discretization of $\dot{x}(\cdot)$ is realized leading to the one-step nonsmooth problem

$$x_{k+1} = x_k + h\,[f(t_k,x_k) + B(x_k,t_k)\lambda_{k+1}], \tag{9.88}$$

where λ_{k+1} solves the $\mathrm{VI}(K, F_{k+1})$ with

$$F_{k+1}(\lambda) = G(t_{k+1}, h\,[f(t_k,x_k) + B(x_k,t_k)\lambda]) + F(\lambda). \tag{9.89}$$

In the last VI, the value λ_{k+1} can be evaluated in an explicit way with respect to x_{k+1}. It is noteworthy that even in the explicit case, the VI is always solved in an implicit way, i.e., for x_{k+1} and λ_{k+1}.

If $\theta \in (0,1]$, we obtain a semi-implicit method where the pair (u_{k+1}, x_{k+1}) solves the $\mathrm{VI}(\mathbb{R}^n \times K, F_{k+1})$ with

$$F_{k+1}(x,\lambda) = \begin{bmatrix} x - x_k - h\,[f(t_k, \theta x + (1-\theta)x_k) + B(x_k,t_k)\lambda] \\ G(t_{k+1},x) + F(\lambda) \end{bmatrix}. \tag{9.90}$$

In Pang & Stewart (in press), the convergence of the semi-implicit case is proved. For this, a continuous piecewise linear function $x^N(\cdot)$ is built by interpolation of the approximate values x_k,

$$x^N(t) = x_k + \frac{t-t_k}{h}(x_{k+1} - x_k), \forall t \in (t_k, t_{k+1}], \tag{9.91}$$

and a piecewise constant function λ^N is built such that

$$\lambda^N(t) = \lambda_{k+1}, \forall t \in (t_k, t_{k+1}]. \tag{9.92}$$

It is noteworthy that the approximation $x^N(\cdot)$ is constructed as a continuous function but that $\lambda^N(\cdot)$ may be discontinuous.

The existence of a subsequence of λ^N, x^N denoted by $\lambda^{N_\nu}, x^{N_\nu}$ such that

- x^{N_ν} converges uniformly to $\hat{x}$ on $[0,T]$,
- λ^{N_ν} converges weakly to $\hat{\lambda}$ in $\mathscr{L}^2((0,T);\mathbb{R}^n)$

is proved under the following assumptions:

1. $f(\cdot)$ and $G(\cdot)$ are Lipschitz continuous on $\Omega = [0,T] \times \mathbb{R}^n$.
2. $B(\cdot)$ is a continuous bounded matrix-valued function on Ω.
3. K is closed and convex (not necessarily bounded).
4. $F(\cdot)$ is continuous.
5. $SOL(K, q+F) \neq \emptyset$ and convex such that $\forall q \in G(\Omega)$, the following growth condition holds:

$$\exists \rho > 0, \sup\{\|\lambda\| \mid \lambda \in \mathrm{SOL}(K, q+F)\} \leqslant \rho(1+\|q\|). \tag{9.93}$$

This assumption is used to prove that a pair (λ_{k+1}, x_{k+1}) exists for the VI (9.90). This assumption of the type "growth condition" is quite usual to prove the existence of solutions of VIs with a fixed-point theorem (see Facchinei & Pang, 2003).

Furthermore, under either one of the following two conditions:

- $F(\lambda) = D\lambda$ (i.e., linear VI) for some positive semi-definite matrix, D,
- $F(\lambda) = \Psi(E\lambda)$, where Ψ is Lipschitz continuous, and there exists $c > 0$ such that

$$\|E\lambda_{k+1} - E\lambda_k\| \leqslant ch, \tag{9.94}$$

all limits $(\hat{x}, \hat{\lambda})$ are weak solutions of the initial value DVI.

This proof of convergence provides us with an existence result for such DVIs separable in λ.

Remark 9.34. The linear growth condition which is a strong assumption in most of the practical cases can be dropped. In this case, some monotonicity assumption has to be made on $F(\cdot)$ and a strong monotonicity assumption on the map $\lambda \mapsto G(t,x) \circ (r + B(t,x)\lambda)$ for all $t \in [0,T], x \in \mathbb{R}^n, r \in \mathbb{R}^n$. We refer to Pang & Stewart (in press) for more details. If $G(x,t) = Cx$, the last assumption means that CB is positive definite.

9.6.2 The Boundary Value Problem

Let us consider now the boundary value problem with linear boundary function

$$\begin{cases} \dot{x}(t) = f(t, x(t)) + B(x(t), t)\lambda(t) \\ \lambda(t) = \mathrm{SOL}(K, G(t, x(t)) + F(\cdot)) \\ b = Mx(0) + Nx(T) \end{cases} .$$

The time-stepping proposed by Pang & Stewart (in press) is as follows:

$$\begin{cases} x_{k+1} - x_k = h\,[f(t_k, \theta x_{k+1} + (1-\theta)x_k) + B(x_k, t_k)\lambda_{k+1}], \\ \quad k \in \{0, \dots, N-1\} \\ \lambda_{k+1} = \mathrm{SOL}(K, G(t_{k+1}, x_{k+1}) + F(\cdot)), \quad k \in \{0, \dots, N-1\} \end{cases} \tag{9.95}$$

plus the boundary condition

$$b = Mx_0 + Nx_N. \tag{9.96}$$

The system is a coupled and large size VI for which the numerical solution is not trivial. The existence of the discrete-time trajectory is ensured under the following assumption:

1. $F(\cdot)$ monotone and VI solutions have linear growth.
2. The map $\lambda \mapsto G(t,x) \circ (r + B(t,x)\lambda)$ is strongly monotone.
3. $M+N$ is nonsingular and satisfies

$$\exp(T\psi_x) < 1 + \frac{1}{\|(M+N)^{-1}N\|},$$

where $\psi_x > 0$ is a constant derived from problem data.

The convergence of the discrete-time trajectory is proved if $F(\cdot)$ is linear.

9.6.2.1 Generalized LCS with $D = 0$ and $B = C^{\mathrm{T}}$

In the case of a generalized LCS with the condition $D = 0$ and $B = C^{\mathrm{T}}$, one obtains the following discretized LEVI:

$$\begin{cases} x_{k+1} - x_k = h\left[p + A\theta x_{k+1} + A(1-\theta)x_k + C^T\lambda_{k+1}\right], \\ \quad k \in \{0, \dots, N-1\} \\ K \ni \lambda_{k+1} \perp q + Cx_{k+1} \in K^* \\ b = Mx_0 + Nx_N \end{cases} . \tag{9.97}$$

The convergence theorem is obtained under the following assumptions:

- $q \in C.\mathrm{I\!R}^n + K^\circ$.
- $M+N$ is nonsingular.
- $C(M+N)^{-1}(b + N.\mathrm{I\!R}^n) \subset K^\circ$.
- $\exp(T\psi_x) < 1 + \dfrac{1}{\|(M+N)^{-1}N\|}$, $\psi_x > 0$,

It is noteworthy that these results are among the first on the BVP for a special case of nonsmooth dynamical systems.

9.7 Summary of the Main Ideas

Both in this chapter and in Chap. 7, we have reviewed most of the methods which allow one to simulate nonsmooth dynamical systems with AC solutions. Let us try to provide a general picture of these methods, with the simplest example of a Filippov's inclusion, i.e.,

$$\dot{x}(t) \in -\mathrm{sgn}(x(t)). \tag{9.98}$$

As we already pointed out, this is an interesting example since it belongs to several subclasses of differential inclusions (Filippov's, maximal monotone, convex and compact sets $F(x)$, etc.). We briefly examined the θ-method for this inclusion in Sect. 9.2.4.2. Let us re-examine two ways to discretize (9.98):

$$x_{k+1} - x_k = -h\,\mathrm{sgn}(x_{k+1}) \ \ \text{(the implicit method)} \tag{9.99}$$

and

$$x_{k+1} - x_k = -h\,\mathrm{sgn}(x_k) \ \ \text{(the explicit method).} \tag{9.100}$$

The implicit method in (9.99) consists in determining the intersection of two graphs: the graph of the single-valued mapping $z(x_{k+1}) = x_{k+1} - x_k$ that is a straight line and the graph of the multivalued mapping $x_{k+1} \mapsto -h\,\mathrm{sgn}(x_{k+1})$. One easily obtains

$$\begin{cases} |x_k| \leqslant h \implies x_{k+1} = x_k \\ x_k > h \quad\implies x_{k+1} = x_k - h \\ x_k < -h \implies x_{k+1} = x_k + h \end{cases}. \tag{9.101}$$

Actually (see Sect. 9.3.3, see also Sect. 1.2) discretizing this way is equivalent to working in a complementarity formalism, i.e., to look for a multiplier λ_{k+1} at each step. Convergence results have been proved. We may therefore consider the implicit method as a "dual" method, in which mode switches are triggered from the value of a suitable Lagrange multiplier. It is noteworthy here that this general feature is also shared by Stewart's event-driven method of Sect. 7.1.2. As an example we may consider $x_0 = 2h > 0$. Then $x_k = h$ for all $k \geqslant 1$. If $x_0 = -2h$ then $x_k = h$ for all $k \geqslant 1$. Filippov's solution is well approximated. The implicit way of discretizing Coulomb model has been rediscovered in Kikuuwe et al. (2005, (14)), where several comparisons with other models prove its superiority (sticking in finite time, no oscillations of the contact force). It is often pointed out that extension towards multiple frictional contacts is hard, because the events detection becomes cumbersome. However, the use of the complementarity formalism and of LCPs permits to overcome such difficulties.

Let us now deal with (9.100). In a sense, one may see the explicit method as a method that uses the value of a multiplier at the foregoing step. Consider $x_0 = 2h$. Then $x_1 = h$ and $x_2 = h - h = 0$. At this stage we encounter a problem that was absent in the implicit method: 0 does not exist on a computer, so that x_2 has to be approximated by some ε. One solution is to define a layer around 0 within which some kind of stabilization is performed, see Leine's method in Sect. 9.3. This, however, does not extend easily to higher codimension switching surfaces. Basic time-stepping schemes work as follow: at step k test whether $|x_k| \leqslant \varepsilon$. If yes, then pick any $\xi_k \in [-h,h]$ and compute $x_{k+1} = x_k + \xi_k$. If no, advance $x_{k+1} = x_k \pm 1$. Then redo the test. According to the theoretical results (see Theorem 9.5), this defines a convergent scheme. However, in practice and for $h > 0$ this results in oscillations around zero and untimely switches of the multiplier, see Fig. 1.12 (see also Fig. 1 in Galias & Yu (2007), where an explicit Euler time-stepping scheme is applied to a sliding-mode controlled feedback system: the discretized trajectories oscillate around the switching surface during the sliding motion). Event-driven strategies, with accurate zero detections, will suffer the same drawback. We conclude that mode switch driven explicitly by the state does not represent a good solution. This is certainly the most important fact to be retained from Chaps. 7 and 9.

Higher order methods (Runge–Kutta, multistep) may improve the quality of the results when the instants of nondifferentiability are rare. Then indeed there are long enough periods during which the algorithm behaves as for a smooth system. The same type of behavior is observed with event-driven schemes, which should then be preferred if high accuracy is needed between the nonsmooth events.

10

Time-Stepping Schemes for Mechanical Systems

This chapter is dedicated to the presentation of time-stepping schemes for mechanical systems with unilateral constraints and/or friction. Roughly speaking, these methods consist in a time-discretization of the dynamics which can be advanced from step k to step $k+1$ by solving specific one-step nonsmooth problems (complementarity problems). The one-step nonsmooth problems may be thought of as the equivalent of any Newton method one needs to update an implicit Euler method for ODEs. Since solving one-step nonsmooth problems is a major issue, a whole part of the book is dedicated to it: Part III. In this chapter we first present the discretized version of Moreau's sweeping process (of second order), without and with friction. Then some other time-stepping algorithms are examined. Summaries of some numerical experiments that have been published in the related literature are presented in order to provide the reader with a rough idea on the capabilities of such time-stepping schemes.

10.1 The Nonsmooth Contact Dynamics (NSCD) Method

In this section, a time-discretization method of the Lagrange dynamical equation (3.115), consistent with the nonsmooth character of the solution, is presented. It is assumed in this section, as in the other sections, that $v^+(\cdot) = \dot{q}^+(\cdot)$ is a locally bounded variation function. The equation of motion reads as,

$$\begin{cases} M(q(t))\mathrm{d}v + N(q(t), v^+(t))\mathrm{d}t + F_{\mathrm{int}}(t, q(t), v^+(t))\,\mathrm{d}t = F_{\mathrm{ext}}(t)\,\mathrm{d}t + \mathrm{d}r \\ \\ v^+(t) = \dot{q}^+(t) \\ \\ q(0) = q_0 \in \mathscr{C}(0),\ \dot{q}(0^-) = \dot{q}_0. \end{cases} \tag{10.1}$$

where $\mathscr{C}(t)$ is defined in (3.16)

We also assume that $F_{\mathrm{int}}(\cdot)$ and $F_{\mathrm{ext}}(\cdot)$ are continuous with respect to time. This assumption is made for the sake of simplicity to avoid the notation $F^+_{\mathrm{int}}(\cdot)$ and $F^+_{\mathrm{ext}}(\cdot)$.

Finally, we will condense the nonlinear inertia terms and the internal forces to lighten the notation. We obtain

$$\begin{cases} M(q(t))\mathrm{d}v + F(t,q(t),v^+(t))\,\mathrm{d}t = F_{\mathrm{ext}}(t)\,\mathrm{d}t + \mathrm{d}r \\ \\ v^+(t) = \dot{q}^+(t) \\ \\ q(0) = q_0 \in \mathscr{C}(0),\ \dot{q}(0^-) = \dot{q}_0. \end{cases} \tag{10.2}$$

The NSCD method, also known as the contact dynamics (CD), is due to the seminal works of J.J. Moreau (1983, 1985a, 1988b, 1994b, 1999) and M. Jean (1988, 1999) (see also Jean & Pratt 1985; Jean & Moreau 1991, 1992). A lot of improvements and variants have been proposed over the years. In this section, we take liberties with these original works, but we choose to present a version of the NSCD method which preserves the essential of the original work. Some extra developments and interpretations are added which are only under our responsibility. To come back to the source of the NSCD method, we encourage to read the above references.

10.1.1 The Linear Time-Invariant Nonsmooth Lagrangian Dynamics

For the sake of simplicity of the presentation, the linear time-invariant case is considered first. The nonlinear case will be examined later in this chapter:

$$\begin{cases} M\,\mathrm{d}v + (Kq(t) + Cv^+(t))\,\mathrm{d}t = F_{\mathrm{ext}}(t)\,\mathrm{d}t + \mathrm{d}r \\ \\ v^+(t) = \dot{q}^+(t). \end{cases} \tag{10.3}$$

10.1.1.1 Time-Discretization of the Dynamics

Integrating both sides of this equation over a time step $(t_k, t_{k+1}]$ of length $h > 0$, one obtains

$$\begin{cases} \displaystyle\int_{(t_k,t_{k+1}]} M\,\mathrm{d}v + \int_{t_k}^{t_{k+1}} (Cv^+(t) + Kq(t))\,\mathrm{d}t = \int_{t_k}^{t_{k+1}} F_{\mathrm{ext}}\,\mathrm{d}t + \int_{(t_k,t_{k+1}]} \mathrm{d}r, \\ \\ \displaystyle q(t_{k+1}) = q(t_k) + \int_{t_k}^{t_{k+1}} v^+(t)\,\mathrm{d}t. \end{cases} \tag{10.4}$$

By definition of the differential measure dv, we obtain

$$\int_{(t_k,t_{k+1}]} M\,\mathrm{d}v = M\int_{(t_k,t_{k+1}]} \mathrm{d}v = M\,(v^+(t_{k+1}) - v^+(t_k)). \tag{10.5}$$

Note that the right velocities are involved in this formulation. The impulse $\displaystyle\int_{(t_k,t_{k+1}]} \mathrm{d}r$ of the reaction on the time interval $(t_k, t_{k+1}]$ emerges as a natural unknown. The equation of the nonsmooth motion can be written under an integral form as:

$$\begin{cases} M\,(v(t_{k+1}) - v(t_k)) = \displaystyle\int_{t_k}^{t_{k+1}} (-Cv^+(t) - Kq(t) + F_{\mathrm{ext}}(t))\,\mathrm{d}t + \int_{(t_k,t_{k+1}]} \mathrm{d}r\,, \\ \\ q(t_{k+1}) = q(t_k) + \displaystyle\int_{t_k}^{t_{k+1}} v^+(t)\,\mathrm{d}t. \end{cases} \tag{10.6}$$

Choosing a numerical method boils down to choosing a method of approximation for the remaining integral terms. Since discontinuities of the derivative $v(\cdot)$ are to be expected if some shocks are occurring, i.e., dr has some atoms within the interval $(t_k,t_{k+1}]$, it is not relevant to use high-order approximations integration schemes for dr (this was pointed out in Remark 9.25). It may be shown on some examples that, on the contrary, such high-order schemes may generate artifact numerical oscillations (see Vola et al., 1998).

The following notation will be used:

- q_k is an approximation of $q(t_k)$ and q_{k+1} is an approximation of $q(t_{k+1})$.
- v_k is an approximation of $v^+(t_k)$ and v_{k+1} is an approximation of $v^+(t_{k+1})$.
- p_{k+1} is an approximation of $\displaystyle\int_{(t_k,t_{k+1}]} \mathrm{d}r$.

A popular first-order numerical scheme, the so-called θ-method, is used for the term supposed to be sufficiently smooth:

$$\int_{t_k}^{t_{k+1}} Cv + Kq\,\mathrm{d}t \approx h\,[\theta(Cv_{k+1} + Kq_{k+1}) + (1-\theta)(Cv_k + Kq_k)]$$
$$\int_{t_k}^{t_{k+1}} F_{\mathrm{ext}}(t)\,\mathrm{d}t \approx h\,[\theta(F_{\mathrm{ext}})_{k+1} + (1-\theta)(F_{\mathrm{ext}})_k]\,.$$

The displacement, assumed to be absolutely continuous, is approximated by

$$q_{k+1} = q_k + h\,[\theta v_{k+1} + (1-\theta)v_k]\,.$$

Taking into account all these discretizations, the following time-discretized equation of motion is obtained:

$$\begin{cases} M(v_{k+1} - v_k) + h\,[\theta(Cv_{k+1} + Kq_{k+1}) + (1-\theta)(Cv_k + Kq_k)] = \\ \qquad\qquad = h\,[\theta(F_{\mathrm{ext}})_{k+1} + (1-\theta)(F_{\mathrm{ext}})_k] + p_{k+1} \\ q_{k+1} = q_k + h\,[\theta v_{k+1} + (1-\theta)v_k]\,. \end{cases} \tag{10.7}$$

Finally, introducing the expression of q_{k+1} in the first equation of (10.7), one obtains:

$$\begin{aligned} \left[M + h\theta C + h^2\theta^2 K\right](v_{k+1} - v_k) &= -hCv_k - hKq_k - h^2\theta K v_k \\ &\quad + h\,[\theta(F_{\mathrm{ext}})_{k+1}) + (1-\theta)(F_{\mathrm{ext}})_k] + p_{k+1}\,, \end{aligned} \tag{10.8}$$

which can be written as:

$$v_{k+1} = v_{\text{free}} + \widehat{M}^{-1} p_{k+1} \,, \tag{10.9}$$

where

- the matrix
$$\widehat{M} = \left[M + h\theta C + h^2\theta^2 K\right] \tag{10.10}$$
is usually called the *iteration matrix*.
- The vector
$$\begin{aligned} v_{\text{free}} &= v_k + \widehat{M}^{-1}\left[-hCv_k - hKq_k - h^2\theta K v_k \right.\\ &\left. + h\left[\theta (F_{\text{ext}})_{k+1}) + (1-\theta)(F_{\text{ext}})_k\right]\right] \end{aligned} \tag{10.11}$$
is the so-called “free” velocity, i.e., the velocity of the system when reaction forces are null.

10.1.1.2 Comments

Let us make some comments on the above developments:

- The iteration matrix $\widehat{M} = \left[M + h\theta C + h^2\theta^2 K\right]$ is supposed to be invertible, since the mass matrix M is usually positive definite and h is supposed to be small enough. The matrices C and K are usually semi-definite positive since rigid motions are allowed to bodies.
- When $\theta = 0$, the θ-scheme is the explicit Euler scheme. When $\theta = 1$, the θ-scheme is the fully implicit Euler scheme. When dealing with a plain Ordinary Differential Equation (ODE)
$$M\ddot{q}(t) + C\dot{q}(t) + Kq(t) = F(t) \tag{10.12}$$
the θ-scheme is unconditionally stable for $0.5 < \theta \leqslant 1$. It is conditionally stable otherwise.
- Equation (10.9) is a linear form of the dynamical equation. It appears as an affine relation between the two unknowns, v_{k+1} that is an approximation of the right derivative of the Lagrange variable at time t_{k+1} and the impulse p_{k+1}. Notice that this scheme is fully implicit. Nonsmooth laws have to be treated by implicit methods.
- From a numerical point of view, two major features appear. First, the different terms in the numerical algorithm will keep finite values. When the time step h vanishes, the scheme copes with finite jumps. Secondly, the use of differential measures of the time interval $(t_k, t_{k+1}]$, i.e., $\mathrm{d}v((t_k, t_{k+1}]) = v^+(t_{k+1}) - v^+(t_k)$ and $\mathrm{d}r((t_k, t_{k+1}])$, offers a rigorous treatment of the nonsmooth evolutions. It is to be noticed that approximations of the acceleration are ignored.

These remarks on the contact dynamics method might be viewed only as some numerical tricks. In fact, the mathematical study of the second-order MDI by Moreau provides a sound mathematical ground to this numerical scheme. It is noteworthy that convergence results have been proved for such time-stepping schemes in Monteiro Marques (1993), Stewart (1998), Mabrouk (1998), and Dzonou & Monteiro Marques (2007), see below.

10.1.2 The Nonlinear Nonsmooth Lagrangian Dynamics

10.1.2.1 Time-Discretization of the Dynamics

Starting from the nonlinear dynamics (10.2), the integration of both sides of this equation over a time step $(t_k, t_{k+1}]$ of length $h > 0$ yields

$$\begin{cases} \displaystyle\int_{(t_k,t_{k+1}]} M(q)\mathrm{d}v + \int_{t_k}^{t_{k+1}} F(t,q(t),v^+(t))\,\mathrm{d}t = \int_{t_k}^{t_{k+1}} F_{\mathrm{ext}}(t)\,\mathrm{d}t + \int_{(t_k,t_{k+1}]} \mathrm{d}r, \\ \displaystyle q(t_{k+1}) = q(t_k) + \int_{t_k}^{t_{k+1}} v^+(t)\,\mathrm{d}t. \end{cases} \tag{10.13}$$

The first term is generally approximated by

$$\int_{(t_k,t_{k+1}]} M(q)\,\mathrm{d}v \approx M(q_{k+\gamma})\,(v_{k+1} - v_k), \tag{10.14}$$

where $q_{k+\gamma}$ generalizes the standard notation for $\gamma \in [0,1]$ such that

$$q_{k+\gamma} = (1-\gamma)q_k + \gamma q_{k+1}. \tag{10.15}$$

The a priori smooth terms are evaluated with a θ-method, chosen in this context for its energy conservation ability,

$$\int_{t_k}^{t_{k+1}} F(t,q,v)\,\mathrm{d}t \approx h\tilde{F}_{k+\theta}, \tag{10.16}$$

where $\tilde{F}_{k+\theta}$ is an approximation with the following dependencies

$$\tilde{F}(t_k, q_k, v_k, t_{k+1}, q_{k+1}, v_{k+1}, t_{k+\theta}, q_{k+\theta}, v_{k+\theta}).$$

The mid-values $t_{k+\theta}, q_{k+\theta}, v_{k+\theta}$ are defined by

$$\begin{cases} t_{k+\theta} = \theta t_{k+1} + (1-\theta)t_k \\ q_{k+\theta} = \theta q_{k+1} + (1-\theta)q_k \ , \quad \theta \in [0,1]. \\ v_{k+\theta} = \theta v_{k+1} + (1-\theta)v_k \end{cases} \tag{10.17}$$

Remark 10.1. The choice of the approximated function $\tilde{F}(\cdot)$ strongly depends on the nature of the internal forces that are modeled. For the linear elastic behavior of homogeneous continuum media, this approximation can be made by

$$\tilde{F}_{k+\theta} = \frac{1}{2}K\colon [E(q_k) + E(q_{k+1})] : F(q_{k+1/2}), \tag{10.18}$$

where $E(\cdot)$ is the Green–Lagrange strain tensor, which leads to an energy conserving algorithm as in Simo & Tarnow (1992). For nonlinear elastic other smooth nonlinear behaviors, we refer to the work of Gonzalez (2000), Laursen & Meng (2001) and references therein for the choice of the discretization and the value of θ.

The displacement, assumed to be absolutely continuous, is approximated by

$$q_{k+1} = q_k + h v_{k+\theta}.$$

The following nonlinear time-discretized equation of motion is obtained:

$$\begin{cases} M(q_{k+\gamma})(v_{k+1} - v_k) + h\tilde{F}_{k+\theta} = p_{k+1} \\ \\ q_{k+1} = q_k + hv_{k+\theta} \end{cases} . \tag{10.19}$$

In its full generality and at least formally, substituting the expression of $q_{k+\gamma}, q_{k+1}$, and $q_{k+\theta}$, the first line of the problem can be written under the form of a residue $\mathscr{R}$ depending only on v_{k+1} such that

$$\mathscr{R}(v_{k+1}) = p_{k+1}. \tag{10.20}$$

In the last expression, we have omitted the dependence to the known values at the beginning of the time step, i.e., q_k and v_k.

10.1.2.2 Linearizing the Dynamics

The system of equations (10.20) for v_{k+1} and p_{k+1} can be linearized yielding a Newton's procedure for solving it. This linearization needs the knowledge of the Jacobian matrix $\nabla\mathscr{R}(\cdot)$ with respect to its argument to construct the tangent linear model.

Let us consider that we have to solve the following equations:

$$\mathscr{R}(u) = 0 \tag{10.21}$$

by a Newton's method where

$$\mathscr{R}(u) = M(q_{k+\gamma})(u - v_k) + h\tilde{F}_{k+\theta}. \tag{10.22}$$

The solution of this system of nonlinear equations is sought as a limit of the sequence $\{u^{\tau}_{k+1}\}_{\tau\in\mathbb{N}}$ such that

$$\begin{cases} u^0_{k+1} = v_k \\ \\ \mathscr{R}_L(u^{\tau+1}_{k+1}) = \mathscr{R}(u^{\tau}_{k+1}) + \nabla\mathscr{R}(u^{\tau}_{k+1})(u^{\tau+1}_{k+1} - u^{\tau}_{k+1}) = 0. \end{cases} \tag{10.23}$$

In practice, all the nonlinearities are not treated in the same manner and the Jacobian matrices for the nonlinear terms involved in the Newton's algorithm are only computed in their natural variables. In the following, we consider some of the most widely used approaches.

The Nonlinear Mass Matrix

The derivation of the Jacobian of the first term of $\mathscr{R}(\cdot)$ implies to compute

$$\nabla_u \left(M(q_{k+\gamma}(u))(u - v_k)\right) \text{ with } q_{k+\gamma}(u) = q_k + \gamma h[(1-\theta)v_k + \theta u]. \tag{10.24}$$

One gets

$$\begin{aligned}\nabla_u \left(M(q_{k+\gamma}(u))(u-v_k)\right) &= M(q_{k+\gamma}(u)) + \left[\nabla_u M(q_{k+\gamma}(u))\right](u-v_k) \\ &= M(q_{k+\gamma}(u)) + \left[h\gamma\theta\nabla_q M(q_{k+\gamma}(u))\right](u-v_k).\end{aligned} \tag{10.25}$$

Remark 10.2. The notation $\nabla_u M(q_{k+\gamma}(u))(u-v_k)$ is to be understood as follows:

$$\nabla_u M(q_{k+\gamma}(u))(u-v_k) = \frac{\partial}{\partial u}[M(q_{k+\gamma}(u))(u-v_k)]$$

which is denoted as $\frac{\partial M_{ij}}{\partial q^l}(q_{k+\gamma}(u))(u^l - v_k^l)$ in tensorial notation. □

A very common approximation consists in considering that the mass matrix evolves slowly with the configuration in a single time step, that is, the term $\nabla_q M(q_{k+\gamma})$ is neglected and one gets,

$$\nabla_u(M(q_{k+\gamma}(u))(u-v_k)) \approx M(q_{k+\gamma}(u)). \tag{10.26}$$

The Jacobian matrix $\nabla\mathscr{R}(\cdot)$ is evaluated in u_{k+1}^τ which yields for the equation (10.26)

$$\nabla_u(M(q_{k+\gamma})(u_{k+1}^\tau - v_k)) \approx M(q_k + \gamma h[(1-\theta)v_k + \theta u_{k+1}^\tau]). \tag{10.27}$$

The prediction of the position which plays an important role will be denoted by

$$\tilde{q}_{k+1}^\tau = q_k + \gamma h[(1-\theta)v_k + \theta u_{k+1}^\tau]. \tag{10.28}$$

Very often, the matrix $M(q_{k+\gamma})$ is only evaluated at the first Newton's iteration with $u_{k+1}^0 = v_k$ leading the approximation for the whole step:

$$M(q_k + \gamma h[(1-\theta)v_k + \theta u_{k+1}^\tau]) \approx M(q_k + h\gamma v_k). \tag{10.29}$$

Another way to interpret the approximation (10.29) is to remark that this evaluation is just an explicit evaluation of the predictive position (10.28) given by $\theta = 0$:

$$\tilde{q}_{k+1} = q_k + h\gamma v_k. \tag{10.30}$$

Using this prediction, the problem (10.19) is written as follows:

$$\begin{cases} M(\tilde{q}_{k+1})(v_{k+1} - v_k) + h\tilde{F}_{k+\theta} = p_{k+1} \\ \\ q_{k+1} = q_k + hv_{k+\theta} \\ \\ \tilde{q}_{k+1} = q_k + h\gamma v_k. \end{cases} \tag{10.31}$$

The Nonlinear Term $F(t,q,v)$

The remaining nonlinear term is linearized providing the Jacobian matrices of $F(t,q,v)$ with respect to q and v. This expression depends strongly on the choice of the approximation $\tilde{F}_{k+\theta}$. Let us consider a pedagogical example, which is not necessarily the best as the Remark 10.1 suggests but which is one of the simplest,

$$\tilde{F}_{k+\theta} = (1-\theta)F(t_k,q_k,v_k) + \theta F(t_{k+1},q_{k+1},v_{k+1}). \tag{10.32}$$

The computation of the Jacobian of $\tilde{F}_{k+\theta}(t,q(u),u)$ for

$$q(u) = q_k + h[(1-\theta)v_k + \theta u]$$

is given for this example by

$$\begin{aligned}\nabla_u \tilde{F}_{k+\theta}(t,q,u) &= \theta \nabla_u F(t,q(u),u) \\ &= \theta \nabla_q F(t_{k+1},q(u),u) \nabla_u q(u) + \theta \nabla_u F(t,q(u),u) \\ &= h\theta^2 \nabla_q F(t,q(u),u) + \theta \nabla_u F(t,q(u),u).\end{aligned} \tag{10.33}$$

The standard tangent stiffness and damping matrices K_t and C_t are defined by

$$\begin{aligned}K_t(t,q,u) &= \nabla_q F(t,q,u) \\ C_t(t,q,u) &= \nabla_u F(t,q,u).\end{aligned} \tag{10.34}$$

In this case, the Jacobian of $\tilde{F}_{k+\theta}(t,q(u),u)$ may be written as

$$\nabla_u \tilde{F}_{k+\theta}(t,q,u) = h\theta^2 K_t(t,q,u) + \theta C_t(t,q,u). \tag{10.35}$$

The complete Newton's iteration can then be written as

$$\widehat{M}_{k+1}^{\tau+1}(u_{k+1}^{\tau+1} - u_{k+1}^{\tau}) = \mathscr{R}(u_{k+1}^{\tau}) + p_{k+1}^{\tau+1}, \tag{10.36}$$

where the iteration matrix is evaluated as

$$\widehat{M}_{k+1}^{\tau+1} = (M(\tilde{q}_{k+1}^{\tau}) + h^2\theta^2 K_t(t_{k+1},q_{k+1}^{\tau},u_{k+1}^{\tau}) + \theta h C_t(t,q_{k+1}^{\tau},u_{k+1}^{\tau})) \tag{10.37}$$

(compare with (10.10)).

Remark 10.3. The choice of $\theta = 0$ leads to an explicit evaluation of the position and the nonlinear forces terms. This choice can be interesting if the time step has to be chosen relatively small due to the presence a very rapid dynamical process. This can be the case in crashes applications or in fracture dynamics (Acary & Monerie (2006)). In this case, the iteration matrix reduces to $\widehat{M}_{k+1}^{\tau+1} = M(\tilde{q}_{k+1}^{\tau})$ avoiding the expensive evaluation of the tangent operator at each time step.

This choice must not be misunderstood. The treatment of the nonsmooth dynamics continues to be implicit.

10.1.3 Discretization of Moreau's Inclusion

In Generalized Coordinates

Let us propose the following:

$$\begin{cases} p_{k+1} \in -N_{T_{\mathscr{C}}(\tilde{q}_{k+1})}\left(\dfrac{v_{k+1}+ev_k}{1+e}\right) \\ \\ \tilde{q}_{k+1} = q_k + h\gamma v_k, \end{cases} \tag{10.38}$$

where the second equality is just a rewriting of (10.9).

The first equality is a discretization of (3.128). Using (A.8) and the linear dynamics (10.9), it follows that

$$\begin{cases} v_{k+1} = -ev_k + (1+e)\mathrm{prox}_{\widehat{M}}[T_{\mathscr{C}}(\tilde{q}_{k+1}); v_{\mathrm{free}}] \\ \\ \tilde{q}_{k+1} = q_k + hv_k. \end{cases} \tag{10.39}$$

The choice for $\tilde{q}_k$ is not unique. For instance Moreau takes $\tilde{q}_{k+1} = q_k + \frac{h}{2}v_k$, i.e., $\gamma = 1/2$ (Moreau, 1999).

When $\mathscr{C}$ is finitely represented, the inclusion in a normal cone in the first line of (10.38) can be rewritten as a complementarity problem. Under some qualification constraints (like nonemptiness of the interior of $T_{\mathscr{C}}(q)$), the tangent cone is a polyhedral cone, since it is either the intersection of half-spaces or the whole ambient space. Thus the normal cone in the right-hand side of (10.38) is also a convex polyhedral cone as in (3.132)—in general different from the normal cone to the admissible domain defined in (3.21). When one or several constraints are activated, i.e., $g^{\alpha}(\tilde{q}_{k+1}) < 0$ for some values of $\alpha \in \{1,...,\nu\}$, one may use the representation in (3.146) to derive the complementarity conditions of this problem (written for $e_{\mathrm{N}} = 0$ in (3.146)).

In Local Coordinates

Recall from Sect. 3.3 that one may also work with the local coordinates that describe the kinematics at the contacting points between bodies. When the contact reactions $R^{\alpha} \in \mathrm{I\!R}^3$ are included in Lagrange equations, one obtains (3.77). Then the contact law is written in terms of the local reactions R^{α} and the local velocities $U^{\alpha} \in \mathrm{I\!R}^3$. In such a case, one directly works with complementarity relations. Let us consider the material of Sect. 3.6.5. The notation used here is the same as in Sect. 3.3. The following notation is used[1] for the local variables setting, where $\approx$ means "is an approximation of":

$$U_{k+1} \approx U^+(t_{k+1}), U_k \approx U^+(t_k),$$

[1] In this paragraph, for simplicity sake, the upper indices $^{\alpha,\beta}$ labeling the contacts are sometimes omitted.

$$g_{k+1} \approx g(t_{k+1}), g_k \approx g(t_k), P_{k+1} \approx \int_{(t_k,t_{k+1}]} \mathrm{d}R.$$

Following the implicit way of discretizing, the discretization of the kinematic laws is proposed as follows:

$$U^{\alpha}_{k+1} = H^{\alpha,\mathrm{T}}(q_{k+1})\, v_{k+1}\,, \tag{10.40}$$

$$p^{\alpha}_{k+1} = H^{\alpha}(q_{k+1})\, P^{\alpha}_{k+1}\,, \quad p_{k+1} = \sum_{\alpha} p^{\alpha}_{k+1}$$

and by analogy with the following formula,

$$q_{k+1} = q_k + h\,[\theta v_{k+1} + (1-\theta) v_k] \tag{10.41}$$

it follows that one has to discretize the gap function as

$$g^{\alpha}_{k+1} = g^{\alpha}_{k} + h\left[\theta U^{\alpha}_{\mathrm{N}k+1} + (1-\theta) U^{\alpha}_{\mathrm{N}k}\right]$$

The local Newton's law of impact is time-discretized in a fully implicit way through the complementarity conditions

$$\begin{cases} \text{If } g^{\alpha}(\tilde{q}_{k+1}) \leqslant 0 \text{ then } 0 \leqslant P^{\alpha}_{\mathrm{N},k+1} \perp U^{\alpha}_{\mathrm{N},k+1} + e^{\alpha} U^{\alpha}_{\mathrm{N},k} \geqslant 0 \\ \text{If } g^{\alpha}(\tilde{q}_{k+1}) > 0 \text{ then } P^{\alpha}_{\mathrm{N},k+1} = 0 \end{cases} \tag{10.42}$$

where we indicated e^{α} since there is no reason that all restitution coefficients be the same.

Notice that the evaluation of the gap function at $\tilde{q}_{k+1}$ exactly corresponds to the evaluation of the tangent cone at $\tilde{q}_{k+1}$.

Remark 10.4. The choice of implicit discretizations is not made randomly. As demonstrated in Sect. 1.1.6.2, implicit discretizations are the only sound way to discretize such nonsmooth systems.

Remark 10.5. [Coping with penetration] All the treatments of the unilateral constraints are written at the velocity level. We cannot expect the original constraints at the position to be satisfied exactly. This is the price to pay to be able to integrate with a time-stepping scheme an evolution on which some impacts are expected. This question is closely related to the notion of index or relative degree in DAE (Brenan et al., 1989). As in the DAE theory, several patches can be applied to circumvent the problem. The Baumgarte stabilization of constraints (Baumgarte, 1972) can be extended to unilateral constraints. We can also add fictitious multipliers on the position level to project the local violation of constraints on the constraints. Jean (1999) proposed a clever way to discretize the gap function in order to satisfy the constraints at the position and the velocity levels at the end of the time step. The *consistency of the gap approximation with unilateral condition* means that the discretized gap function $\bar{g}_k \overset{\Delta}{=} \bar{g}(\tilde{q}_{k+1})$ should satisfy the implication[2]

[2] The superscript α is dropped.

$$\bar{g}_{k+1} = 0 \text{ and } \bar{g}_k = 0 \Rightarrow U_{\mathrm{N},k+1} = 0. \tag{10.43}$$

Possible choices are $\bar{g}_{k+1} = g(\tilde{q}_{k+1}) + (1-\theta)hU_{\mathrm{N},k+1}$ and $\bar{g}_k = g(\tilde{q}_k) + (1-\theta)hU_{\mathrm{N},k}$. Then $\bar{g}_{k+1}$ is approximately the gap corresponding to the configuration $q_{k+1} + (1-\theta)$ hv_{k+1}, whereas $\bar{g}_k$ is approximately the gap corresponding to the configuration $q_k + (1-\theta)hv_k$. By discretizing the "true" gap as above as $g_{k+1} = g_k + \theta hU_{\mathrm{N},k+1} +$ $(1-\theta)hU_{\mathrm{N},k}$ it follows that the consistency property is satisfied for $\bar{g}$. Clearly the consistency cannot hold for g_{k+1} and g_k.

Obviously, another solution consists of choosing $h > 0$ small enough while still working at the velocity level, so that the penetration is negligible. One may also project the trajectory on the constraint, from time to time.

10.1.4 Sweeping Process with Friction

The natural framework when friction acts at the contact points is that of local kinematics, presented in Sect. 10.1.3 in (10.40)–(10.42). The time-discretization of Coulomb's model in (3.147) is done as follows

$$\begin{cases} \text{If } U^{\alpha}_{\mathrm{T},k+1} = 0 \text{ then } P^{\alpha}_{k+1} \in \mathbf{C}^{\alpha} \\ \text{If } U^{\alpha}_{\mathrm{T},k+1} \neq 0 \text{ then } \|P^{\alpha}_{\mathrm{T},k+1}\| = \mu^{\alpha}|P^{\alpha}_{\mathrm{N},k+1}| \text{ and there exists a scalar } a \geqslant 0 \\ \qquad\qquad \text{such that } P^{\alpha}_{\mathrm{T},k+1} = -aU^{\alpha}_{\mathrm{T},k+1} \end{cases} \tag{10.44}$$

where $\mathbf{C}^{\alpha} = \{P \mid \|P_{\mathrm{T}}\| \leqslant \mu^{\alpha} P_{\mathrm{N}}\}$. We note that the time-discretized friction may equivalently be written as

$$-U^{\alpha}_{\mathrm{T},k+1} \in N_{\mathbf{D}^{\alpha}_{k+1}}(P^{\alpha}_{k+1}) \tag{10.45}$$

with $\mathbf{D}^{\alpha}_{k+1} = \{z \in \mathbb{R}^2 \mid \|z\| \leqslant \mu^{\alpha} P^{\alpha}_{\mathrm{N},k+1}\}$. The other formulation of the Coulomb's model such as (3.152) follows straightforwardly,

$$P^{\alpha}_{\mathrm{T},k+1} = \mathrm{proj}_{\mathbf{D}^{\alpha}_{k+1}}[P^{\alpha}_{\mathrm{T},k+1} - \rho U^{\alpha}_{\mathrm{T},k+1}], \ \ \rho > 0. \tag{10.46}$$

A Second-Order Cone Complementarity Problem

The formulation based on the De Saxcé's bipotential (3.160) takes into account the unilateral contact and the friction model. The discretization follows the implicit rule in which we include the restitution law

$$-(U^{\alpha}_{\mathrm{N},k+1} + e^{\alpha}U^{\alpha}_{\mathrm{N},k} + \mu^{\alpha}\, \|U^{\alpha}_{\mathrm{T},k+1}\|, U^{\alpha}_{\mathrm{T},k+1})^{\mathrm{T}} \in \partial\psi_{\mathbf{C}}(P^{\alpha}_{k+1}). \tag{10.47}$$

The second-order cone complementarity problem (see (3.164))

$$\mathbf{C}^{\alpha,*} \ni \left[U^{\alpha}_{\mathrm{N},k+1} + eU^{\alpha}_{\mathrm{N},k} + \mu^{\alpha}\, \|U^{\alpha}_{\mathrm{T},k+1}\|, U^{\alpha}_{\mathrm{T},k+1}\right]^{\mathrm{T}} \perp P^{\alpha}_{k+1} \in \mathbf{C}^{\alpha} \tag{10.48}$$

summarizes the time-discretized unilateral contact with Coulomb's friction model. Introducing the modified local velocity as

$$\widehat{U}^{\alpha}_{k+1} = \left[U^{\alpha}_{\mathrm{N},k+1} + eU^{\alpha}_{\mathrm{N},k} + \mu^{\alpha}\, ||U^{\alpha}_{\mathrm{T},k+1}||, U^{\alpha}_{\mathrm{T},k+1}\right]^{\mathrm{T}} \tag{10.49}$$

the second-order cone complementarity problem may be written as

$$\mathbf{C}^{\alpha,*} \ni \widehat{U}^{\alpha}_{k+1} \perp P^{\alpha}_{k+1} \in \mathbf{C}^{\alpha}. \tag{10.50}$$

10.1.5 The One-Step Time-Discretized Nonsmooth Problem

Once the nonsmooth Lagrangian dynamics (Sects. 10.1.1 and 10.1.2) and the Moreau's inclusion (Sect. 10.1.3) have been time-discretized, a time-discretized Onestep Nonsmooth Problem (OSNSP) can be written under various forms depending mainly on the kind of nonlinearities in the problem and the choices in keeping local and/or generalized variables.

Two main types of nonlinearities have to be addressed:

- *The nonlinearities in the dynamics in the term $F(t,q,v)$ and the mass matrix $M(q)$*. We will call these nonlinearities, *global nonlinearities*. They usually are treated differently. The nonlinear inertia terms $N(q,v)$ and the mass matrix $M(q)$ are often discretized in an explicit way assuming that the configuration does not evolve too much in a time step. On the contrary, the nonlinear internal forces $F_{\mathrm{int}}(t,q,v)$ are fully implicitly discretized and subsequently treated by a Newton-like method.
- *The nonlinearities in the constraints*, $g(q)$. We will call the nonlinearities the *local nonlinearities*. These nonlinearities only depend on the configuration. They are often treated explicitly assuming once again that the configuration evolves slowly in a time step.

Three choices in the variables to state the one-step time-discretized can be listed:

- *Reduction to the local variables*. Using the kinematics law and if the time-discretized dynamics is either linear or linearized, it is possible to reduce all the problems in terms of local variables.
- *Reduction to the generalized variables*. If the Moreau's differential inclusion or more generally the frictional contact law is written in terms of generalized coordinates, it is possible to formulate the problems only in terms of generalized variables.
- *The mixed problem with local and generalized variables*. In the nonlinear case, the problem appears naturally as a mixed problem.

We will try in this section to summarize the main kinds of one-step time-discretized problems mainly depending on the choice of the numerical treatments of the global and local nonlinearities and the possible reduction to local coordinates.

10.1.5.1 The Linear Time-Invariant Case

In this case, the nonsmooth Lagrangian dynamics and the constraints are assumed to be linear. We have seen that the linear time-invariant dynamics can be discretized to obtain

$$\widehat{M}(v_{k+1} - v_{\text{free}}) = p_{k+1}, \tag{10.51}$$

where v_{free} and $\widehat{M}$ are defined in (10.10) and (10.11).

In local coordinates, the discretization of the kinematics law is given by (10.40) which yields in the linear time-invariant framework

$$U^{\alpha}_{k+1} = H^{\alpha,\mathrm{T}}\, v_{k+1}\,, \tag{10.52}$$

$$p^{\alpha}_{k+1} = H^{\alpha}\, P^{\alpha}_{k+1}\,, \quad p_{k+1} = \sum_{\alpha} p^{\alpha}_{k+1} \ .$$

Adding the time-discretized contact law with Coulomb's friction (10.50), the time-discretized mixed linear Onestep NonSmooth Problem (OSNSP), denoted as $(\mathscr{P}_{\text{ML}})$ is obtained

$$\boxed{(\mathscr{P}_{\text{ML}}) \begin{cases} \widehat{M}(v_{k+1} - v_{\text{free}}) = p_{k+1} = \sum\limits_{\alpha} p^{\alpha}_{k+1} \\[2ex] U^{\alpha}_{k+1} = H^{\alpha,T}\, v_{k+1}; \qquad p^{\alpha}_{k+1} = H^{\alpha}\, P^{\alpha}_{k+1} \\[2ex] \text{If } \ g^{\alpha}(\tilde{q}_{k+1}) \leqslant 0 \ \text{ then} \\ \qquad \widehat{U}^{\alpha}_{k+1} = \left[U^{\alpha}_{\text{N},k+1} + eU^{\alpha}_{\text{N},k} + \mu^{\alpha}\, ||U^{\alpha}_{\text{T},k+1}||, U^{\alpha}_{\text{T},k+1}\right]^{\mathrm{T}} \\ \qquad \mathbf{C}^{\alpha,*} \ni \widehat{U}^{\alpha}_{k+1} \perp P^{\alpha}_{k+1} \in \mathbf{C}^{\alpha} \\[2ex] \text{If } \ g^{\alpha}(\tilde{q}_{k+1}) > 0 \ \text{ then } \ P^{\alpha}_{k+1} = 0 \end{cases}}$$

Reduction to Local Coordinates

Rewriting the time-discretized dynamics in local coordinates for each $\alpha \in \{1 \dots \nu\}$ leads to

$$U^{\alpha}_{k+1} = H^{\alpha,\mathrm{T}} \widehat{M}^{-1} \sum_{\beta} H^{\beta}\, P^{\beta}_{k+1} + H^{\alpha,\mathrm{T}} v_{\text{free}}. \tag{10.53}$$

More compactly, these equations can be written as

$$U^{\alpha}_{k+1} = \widehat{W}^{\alpha\alpha} P_{k+1} + U^{\alpha}_{\text{locfree}}, \quad \alpha \in \{1 \dots \nu\}, \tag{10.54}$$

where the so-called Delassus' operator for the constraints α is equal to

$$\widehat{W}^{\alpha\alpha} = H^{\alpha,\mathrm{T}} \widehat{M}^{-1} H^{\alpha}. \tag{10.55}$$

The local free velocity $U^{\alpha}_{\text{locfree}}$ represents the velocity that the system would have at step $k+1$ if $P^{\alpha}_{k+1} = 0$, i.e.,

$$\begin{aligned} U^{\alpha}_{\mathrm{locfree}} &= H^{\alpha,\mathrm{T}} v_{\mathrm{free}} + \sum_{\beta\neq\alpha} H^{\alpha,\mathrm{T}} \widehat{M}^{-1} H^{\beta} P^{\beta}_{k+1} \\ &= H^{\alpha,\mathrm{T}} v_{\mathrm{free}} + \sum_{\beta\neq\alpha} \widehat{W}^{\alpha\beta} P^{\beta}_{k+1}. \end{aligned} \tag{10.56}$$

The previous equations (10.54) are gathered in the following form for all constraints $\alpha \in \{1\ldots\nu\}$ thanks to (3.73):

$$U_{k+1} = \widehat{W} P_{k+1} + U_{\mathrm{free}}. \tag{10.57}$$

With this notation, the complete Delassus' operator can be written as

$$\widehat{W} = H^{\mathrm{T}} \widehat{M}^{-1} H. \tag{10.58}$$

Adding the time-discretized contact law with Coulomb's friction (10.50), the time-discretized linear OSNSP, denoted by $(\mathscr{P}_{\mathrm{L}})$ is obtained:

$$\boxed{(\mathscr{P}_{\mathrm{L}}) \left\{ \begin{array}{l} U_{k+1} = \widehat{W} P_{k+1} + U_{\mathrm{free}} \\ \\ \text{If } g^{\alpha}(\tilde{q}_{k+1}) \leqslant 0 \text{ then} \\ \qquad \widehat{U}^{\alpha}_{k+1} = \left[U^{\alpha}_{\mathrm{N},k+1} + e U^{\alpha}_{\mathrm{N},k} + \mu^{\alpha} \, ||U^{\alpha}_{\mathrm{T},k+1}||, U^{\alpha}_{\mathrm{T},k+1} \right]^{\mathrm{T}} \\ \qquad \mathbf{C}^{\alpha,*} \ni \widehat{U}^{\alpha}_{k+1} \perp P^{\alpha}_{k+1} \in \mathbf{C}^{\alpha} \\ \\ \text{If } g^{\alpha}(\tilde{q}_{k+1}) > 0 \text{ then } P^{\alpha}_{k+1} = 0 \end{array} \right.}$$

The NSCD method for the linear case is summarized in Algorithm 8.

The Frictionless Case

In the frictionless case, i.e., $P^{\alpha}_{\mathrm{T},k+1} = [0,0]^{\mathrm{T}}$, the problem $(\mathscr{P}_{\mathrm{L}_1})$ can be further reduced. To do this, we decompose $\widehat{W}^{\alpha}$ as:

$$\widehat{W}^{\alpha} = \begin{pmatrix} \widehat{W}^{\alpha}_{\mathrm{TT}} & \widehat{W}^{\alpha}_{\mathrm{TN}} \\ \\ \widehat{W}^{\alpha}_{\mathrm{NT}} & \widehat{W}^{\alpha}_{\mathrm{NN}} \end{pmatrix}. \tag{10.59}$$

Let us now construct the Linear Complementarity Problem (LCP) with unknown $P^{\alpha}_{\mathrm{N},k+1}$. We assume first that there is a single contact α. We denote $e_3 = (0\;0\;1)^{\mathrm{T}}$, so that $U^{\alpha}_{\mathrm{N},k+1} = e_3^{\mathrm{T}} U^{\alpha}_{k+1}$, whereas $P^{\alpha}_{k+1} = P^{\alpha}_{\mathrm{N},k+1} e_3$ since the reaction is normal to the common tangent plane between the bodies.

Consequently we may rewrite (10.54) as

$$U^{\alpha}_{\mathrm{N},k+1} = e_3^T \widehat{W}^{\alpha} e_3 P^{\alpha}_{\mathrm{N},k+1} + e_3^T U^{\alpha}_{\mathrm{free}} = \widehat{W}^{\alpha}_{\mathrm{NN}} P^{\alpha}_{\mathrm{N},k+1} + U^{\alpha}_{\mathrm{N,free}}. \tag{10.60}$$

Algorithm 8 NSCD method. Linear case.

```
Require: $\mathsf{M}, \mathsf{K}, \mathsf{C}, \mathsf{F}_{\mathsf{ext}}$ linear dynamics (10.3)
Require: $\mathsf{H}^{\alpha}, \mathsf{g}^{\alpha}(\cdot)$, for all $\alpha \in \mathsf{I} = \{1 \dots \nu\} \subset \mathbb{N}$, kinematic relations (10.52)
Require: $\mathsf{e}, \mu$ frictional contact law.
Require: $\mathsf{t}_0, \mathsf{T}$ time–integration interval
Require: $\mathsf{q}_0, \mathsf{v}_0$ initial data
Require: $\mathsf{h}, \theta, \gamma$ time-step and integration parameters
Ensure: $(\{\mathsf{q}_\mathsf{k}\}, \{\mathsf{v}_\mathsf{k}\}, \{\mathsf{p}_\mathsf{k}\}, \{\mathsf{U}_\mathsf{k}\}\{\mathsf{P}_\mathsf{k}\}), \mathsf{k} \in \{1, 2, \dots\}$

  $\mathsf{k} \leftarrow 0$
  $\mathsf{U}_0 \leftarrow \mathsf{H}^\mathsf{T} \mathsf{v}_0$
  // Computation of time independent matrices
  $\widehat{\mathsf{M}} \leftarrow \left[\mathsf{M} + \mathsf{h}\theta\mathsf{C} + \mathsf{h}^2\theta^2\mathsf{K}\right]$ // iteration matrix (10.10)
  $\widehat{\mathsf{W}}^{\alpha\beta} \leftarrow \mathsf{H}^{\alpha,\mathsf{T}}\widehat{\mathsf{M}}^{-1}\mathsf{H}^{\beta}, (\alpha,\beta) \in \{1 \dots \nu\}^2$ // Delassus operator (10.55)

  // Time integration
  while $\mathsf{t}_\mathsf{k} < \mathsf{T}$ do
    $\mathsf{v}_{\mathrm{free}} \leftarrow \mathsf{v}_\mathsf{k} + \widehat{\mathsf{M}}^{-1}\left[-\mathsf{hCv}_\mathsf{k} - \mathsf{hKq}_\mathsf{k} - \mathsf{h}^2\theta\mathsf{Kv}_\mathsf{k}\ + \mathsf{h}\left[\theta(\mathsf{F}_{\mathrm{ext}})_{\mathsf{k}+1}) + (1-\theta)(\mathsf{F}_{\mathrm{ext}})_\mathsf{k}\right]\right]$
    // Update of the index set of forecast active constraints
    $\tilde{\mathsf{q}}_{\mathsf{k}+1} \leftarrow \mathsf{q}_\mathsf{k} + \mathsf{h}\gamma\mathsf{v}_\mathsf{k}$
    $\mathsf{I}_\mathsf{a}(\tilde{\mathsf{q}}_{\mathsf{k}+1}) \leftarrow \{\alpha \in \mathsf{I} \mid \mathsf{g}^{\alpha}(\tilde{\mathsf{q}}_{\mathsf{k}+1}) \leqslant 0\} \subseteq \mathsf{I}$
    //One-step nonsmooth problem update
    for $\alpha \in \mathsf{I}_\mathsf{a}$ do
      $\mathsf{U}_{\mathrm{free}} \leftarrow \mathsf{H}^\mathsf{T}\mathsf{v}_{\mathsf{free}}$
      Assemble (if necessary) $\widehat{\mathsf{W}}$ with $\widehat{\mathsf{W}}^{\alpha,\beta}, (\alpha,\beta) \in \mathsf{I}_\mathsf{a}^2$
    end for
    //Resolution of the one-step nonsmooth problem
    if $\mathsf{I}_\mathsf{a} \neq \emptyset$ then
      $[\mathsf{U}_{\mathsf{k}+1}, \mathsf{P}_{\mathsf{k}+1}] \leftarrow$ solution of OSNSP $(\mathscr{P}_L)$ (see Chap. 13)
    end if
    // State update
    $\mathsf{p}_{\mathsf{k}+1} \leftarrow \sum_{\alpha \in \mathsf{I}_\mathsf{a}} \mathsf{H}^{\alpha}\mathsf{P}^{\alpha}_{\mathsf{k}+1}$
    $\mathsf{v}_{\mathsf{k}+1} \leftarrow \mathsf{v}_{\mathsf{free}} + \widehat{\mathsf{M}}^{-1}\mathsf{p}_{\mathsf{k}+1}$
    $\mathsf{q}_{\mathsf{k}+1} \leftarrow \mathsf{q}_\mathsf{k} + \mathsf{h}\left[\theta\mathsf{v}_{\mathsf{k}+1} + (1-\theta)\mathsf{v}_\mathsf{k}\right]$
    $\mathsf{t}_\mathsf{k} \leftarrow \mathsf{t}_{\mathsf{k}+1}$
    $\mathsf{k} \leftarrow \mathsf{k}+1$
  end while
```

Inserting this into the first line of (10.42) yields the scalar LCP with unknown $P^{\alpha}_{\mathrm{N},k+1}$:

$$0 \leqslant P^{\alpha}_{\mathrm{N},k+1} \perp \widehat{W}^{\alpha}_{\mathrm{NN}} P^{\alpha}_{\mathrm{N},k+1} + U^{\alpha}_{\mathrm{N,free}} + eU^{\alpha}_{\mathrm{N},k} \geqslant 0 \tag{10.61}$$

assuming that $g^{\alpha}(\tilde{q}_{k+1}) \leqslant 0$. Let us now deal with the general case of a system with $\nu \geqslant 2$ contacts. The decomposition of $\widehat{W}$ may be written

$$\widehat{W} = \begin{pmatrix} \widehat{W}_{\text{TT}} & \widehat{W}_{\text{TN}} \\ \\ \widehat{W}_{\text{NT}} & \widehat{W}_{\text{NN}} \end{pmatrix}. \tag{10.62}$$

When dealing with frictionless contacts, $P_{\text{T},k+1} = 0$ so that inserting (10.62) into (10.57) we get:

$$U_{\text{N},k+1} = \widehat{W}_{\text{NN}} P_{\text{N},k+1} + U_{\text{N,free}} \tag{10.63}$$

leading to the time-discretized linear OSNSP without friction, denoted by $(\mathscr{P}_{\text{LWF}})$:

$$\boxed{(\mathscr{P}_{\text{LWF}}) \begin{cases} U_{\text{N},k+1} = \widehat{W}_{\text{NN}} P_{\text{N},k+1} + U_{\text{N,free}} \\ \\ \text{If } g^{\alpha}(\tilde{q}_{k+1}) \leqslant 0 \text{ then } 0 \leqslant P^{\alpha}_{\text{N},k+1} \perp U^{\alpha}_{\text{N},k+1} + e^{\alpha} U^{\alpha}_{\text{N},k} \geqslant 0 \\ \\ \text{If } g^{\alpha}(\tilde{q}_{k+1}) > 0 \text{ then } P^{\text{N},\alpha}_{k+1} = 0 \end{cases}}$$

10.1.5.2 The Fully Nonlinear Case

In the fully nonlinear case, the time-discretized dynamics yields:

$$\mathscr{R}(v_{k+1}) = p_{k+1}, \tag{10.64}$$

where $\mathscr{R}$ is the nonlinear residue

$$\mathscr{R}(u) = M(q_{k+\gamma})(u - v_k) + h\tilde{F}_{k+\theta}. \tag{10.65}$$

Adding the time-discretized contact law with Coulomb's friction (10.50), the time-discretized mixed nonlinear OSNSP denoted by $(\mathscr{P}_{\text{MNL}})$, is obtained:

$$\boxed{(\mathscr{P}_{\text{MNL}}) \begin{cases} \mathscr{R}(v_{k+1}) = p_{k+1} = \sum\limits_{\alpha} p^{\alpha}_{k+1} \\ \\ U^{\alpha}_{k+1} = H^{\alpha,\text{T}}(q_k + 1)\, v_{k+1}; \qquad p^{\alpha}_{k+1} = H^{\alpha}(q_k + 1)\, P^{\alpha}_{k+1} \\ \\ \text{If } g^{\alpha}(\tilde{q}_{k+1}) \leqslant 0 \text{ then} \\ \qquad \widehat{U}^{\alpha}_{k+1} = \left[U^{\alpha}_{\text{N},k+1} + eU^{\alpha}_{\text{N},k} + \mu^{\alpha}\, ||U^{\alpha}_{\text{T},k+1}||, U^{\alpha}_{\text{T},k+1} \right]^{\text{T}} \\ \qquad \mathbf{C}^{\alpha,*} \ni \widehat{U}^{\alpha}_{k+1} \perp P^{\alpha}_{k+1} \in \mathbf{C}^{\alpha} \\ \\ \text{If } g^{\alpha}(\tilde{q}_{k+1}) > 0 \text{ then } P^{\alpha,}_{k+1} = 0 \end{cases}}$$

10.1.5.3 Linearizing the Dynamics and the Constraints

The linearization of the dynamics by a Newton's procedure yields the following time-discretized linearized dynamics in the form

$$\widehat{M}_{k+1}^{\tau+1}(u_{k+1}^{\tau+1}-u_{k+1}^{\tau})=\mathscr{R}(u_{k+1}^{\tau})+p_{k+1}^{\tau+1}. \tag{10.66}$$

The nonlinearity in the kinematics relations is processed in the same manner as in the mass matrix. Let us consider the function

$$U(v)=H^{\mathrm{T}}(q_{k+\gamma}(v))v \text{ with } q_{k+\gamma}(v)=q_k+\gamma h[(1-\theta)v_k+\theta v]. \tag{10.67}$$

The linearization necessitates to compute the Jacobian, $\nabla U(v)$, that is[3]

$$\begin{aligned}\nabla_v U(v)&=\nabla_v(H^{\mathrm{T}}(q_{k+\gamma}(v))v)\\&=\nabla_v(H^{\mathrm{T}}(q_{k+\gamma}(v)))v+H^{\mathrm{T}}(q_{k+\gamma}(v))\\&=h\theta\gamma\nabla_q(H^{\mathrm{T}}(q))v+H^{\mathrm{T}}(q_{k+\gamma}(v)).\end{aligned} \tag{10.68}$$

As for the mass matrix, the second-order term $h\theta\gamma\nabla_q(H^{\mathrm{T}}(q_{k+\gamma}(v)))vv$ is often neglected. One gets for the approximation of the Jacobian

$$\nabla_v U(v)=H^{\mathrm{T}}(q_{k+\gamma}(v)). \tag{10.69}$$

The tangent linear model around the point u_{k+1}^{τ} is given by

$$\begin{aligned}U_{\mathrm{L}}(u_{k+1}^{\tau+1})&=H^{\mathrm{T}}(q_{k+\gamma}(u_{k+1}^{\tau}))u_{k+1}^{\tau}+\nabla_v U(v_{k+1}^{\tau})(u_{k+1}^{\tau+1}-u_{k+1}^{\tau})\\&=H^{\mathrm{T}}(q_{k+\gamma}(u_{k+1}^{\tau}))u_{k+1}^{\tau}+\left[H^{\mathrm{T}}(q_{k+\gamma}(u_{k+1}^{\tau}))\right](u_{k+1}^{\tau+1}-u_{k+1}^{\tau})\\&=\left[H^{\mathrm{T}}(q_{k+\gamma}(u_{k+1}^{\tau}))\right]u_{k+1}^{\tau+1}.\end{aligned} \tag{10.70}$$

As for the mass matrix, the prediction of the position is given by

$$\tilde{q}_{k+1}^{\tau}=q_k+h\gamma[(1-\theta)v_k+\theta u_{k+1}^{\tau}] \tag{10.71}$$

which is often evaluated explicitly as the first iteration of the Newton's loop for $u_{k+1}^{0}=v_k$. One gets

$$\tilde{q}_{k+1}=q_k+h\gamma v_k. \tag{10.72}$$

The adjoint relation on p_{k+1} is similarly treated.

To summarize, one gets the following linearized kinematics relations for each contact α as

$$\begin{aligned}U_{k+1}^{\alpha,\tau+1}&=\left[H^{\alpha,\mathrm{T}}(\tilde{q}_{k+1}^{\tau})\right]u_{k+1}^{\tau+1}\\p_{k+1}^{\alpha,\tau+1}&=\left[H^{\alpha}(\tilde{q}_{k+1}^{\tau})\right]P_{k+1}^{\alpha,\tau+1}.\end{aligned} \tag{10.73}$$

[3] We adopt the same convention for the differentiation as in Remark 10.2.

Adding the time-discretized contact law with Coulomb's friction (10.50), the time-discretized mixed linearized OSNSP denoted by $(\mathscr{P}_{\mathrm{ML}\tau})$, is obtained

$$\boxed{(\mathscr{P}_{\mathrm{ML}\tau})\begin{cases} \widehat{M}_{k+1}^{\tau+1}(u_{k+1}^{\tau+1}-u_{k+1}^{\tau}) = \mathscr{R}(u_{k+1}^{\tau}) + p_{k+1}^{\tau+1} = \mathscr{R}(u_{k+1}^{\tau}) + \sum\limits_{\alpha} p_{k+1}^{\alpha,\tau+1} \\[2ex] U_{k+1}^{\alpha,\tau+1} = \left[H^{\alpha,\mathrm{T}}(\tilde{q}_{k+1}^{\tau})\right] u_{k+1}^{\tau+1} \\[2ex] p_{k+1}^{\alpha,\tau+1} = \left[H^{\alpha}(\tilde{q}_{k+1}^{\tau})\right] P_{k+1}^{\alpha,\tau+1} \\[2ex] \text{If } g^{\alpha}(\tilde{q}_{k+1}^{\tau}) \leqslant 0 \text{ then} \\ \qquad \widehat{U}_{k+1}^{\alpha,\tau+1} = \left[U_{\mathrm{N},k+1}^{\alpha,\tau+1} + eU_{\mathrm{N},k}^{\alpha} + \mu^{\alpha}\, ||U_{\mathrm{T},k+1}^{\alpha,\tau+1}||, U_{\mathrm{T},k+1}^{\alpha,\tau+1}\right]^{\mathrm{T}} \\ \qquad \mathbf{C}^{\alpha,*} \ni \widehat{U}_{k+1}^{\alpha,\tau+1} \perp P_{k+1}^{\alpha,\tau+1} \in \mathbf{C}^{\alpha} \\[2ex] \text{If } g^{\alpha}(\tilde{q}_{k+1}^{\tau}) > 0 \text{ then } P_{k+1}^{\alpha,\tau+1} = 0 \end{cases}}$$

Reduction to the Local Variables

The same reduction as in the linear case can be performed. Without entering into deeper details, we define the so-called Delassus' operator for the constraints α as

$$\widehat{W}_{k+1}^{\alpha\alpha,\tau+1} = H^{\alpha,\mathrm{T}}(\tilde{q}_{k+1}^{\tau})\widehat{M}_{k+1}^{\tau+1,-1} H^{\alpha}(\tilde{q}_{k+1}^{\tau})$$

and the free velocity as

$$U_{k+1,\mathrm{free}}^{\tau+1} = H^{\alpha,\mathrm{T}}(\tilde{q}_{k+1}^{\tau})v_{k+1,\mathrm{free}}^{\tau+1} \tag{10.74}$$

with

$$v_{k+1,\mathrm{free}}^{\tau+1} = \widehat{M}_{k+1}^{\tau+1,-1}\left[\mathscr{R}(u_{k+1}^{\tau})\right] + u_{k+1}^{\tau}. \tag{10.75}$$

Adding the time-discretized contact law with Coulomb's friction (10.50), the time-discretized linearized OSNSP denoted by $(\mathscr{P}_{\mathrm{L}\tau})$, is obtained

$$\boxed{(\mathscr{P}_{L\tau})\begin{cases} U_{k+1}^{\tau+1} = \widehat{W}_{k+1}^{\tau+1} P_{k+1}^{\tau+1} + U_{k+1,\mathrm{free}}^{\tau+1} \\[2ex] \text{If } g^{\alpha}(\tilde{q}_{k+1}^{\tau}) \leqslant 0 \text{ then} \\ \qquad \widehat{U}_{k+1}^{\alpha,\tau+1} = \left[U_{\mathrm{N},k+1}^{\alpha,\tau+1} + eU_{\mathrm{N},k}^{\alpha} + \mu^{\alpha}\, ||U_{\mathrm{T},k+1}^{\alpha,\tau+1}||, U_{\mathrm{T},k+1}^{\alpha,\tau+1}\right]^{\mathrm{T}} \\ \qquad \mathbf{C}^{\alpha,*} \ni \widehat{U}_{k+1}^{\alpha,\tau+1} \perp P_{k+1}^{\alpha,\tau+1} \in \mathbf{C}^{\alpha} \\[2ex] \text{If } g^{\alpha}(\tilde{q}_{k+1}) > 0 \text{ then } P_{k+1}^{\alpha,\tau+1} = 0 \end{cases}}$$

The NSCD method for the linear case is summarized in Algorithm 9.

Remark 10.6. The Delassus' operator that is used in event-driven schemes is precisely the matrix $W(q)$, not the matrix $\widehat{W}(q)$, see (8.8). The major discrepancy between event-driven and time-stepping methods is that the nonlinear terms are not considered at the impact times in an event-driven scheme. Therefore the above issues on the two Delassus' operators are irrelevant. These problems are specific to time-stepping algorithms.

10.1.6 Convergence Properties

The convergence properties of time-discretizations of Moreau's sweeping process have been pioneered in Monteiro Marques (1985, 1993). Let us now describe the results in Dzonou et al. (2006), which extend previous convergence studies in Mabrouk (1998) and Monteiro Marques (1993). It is assumed that the forces $F_{\text{int}}(t,q,v)$ and $F_{\text{ext}}(\cdot)$ define continuous mappings, and that they are locally Lipschitz continuous with respect to q and v. It is also assumed that $M(q) = M^{\mathrm{T}}(q) > 0$ and that the unilateral constraint $g(\cdot)$ is a $C^{1,\frac{1}{2}}$ function with nonzero gradient in the neighborhood of its zero level set. Let the interval of integration be $[0,T]$, $T > 0$, and the time step be $h = \frac{T}{N}$, $1 \leqslant N \in \mathbb{N}$, so that $t_k = kh$. The two sequences $\{q_{N,k}\}_{0 \leqslant k \leqslant N}$ and $\{v_{N,k}\}_{0 \leqslant k \leqslant N}$ are defined as

$$\begin{cases} q_{N,0} = q_0 \\ \\ v_{N,0} = -e\dot{q}_0 + (1+e)\text{proj}_{q_0}[T_{\mathscr{C}}(q_0);\dot{q}_0] \end{cases} \tag{10.76}$$

and for all $0 \leqslant k \leqslant N-1$

$$\begin{cases} q_{N,k+1} = q_{N,k} + hv_{N,k} \\ \\ v_{N,k+1} = -ev_{N,k} + \\ \\ \qquad\qquad + (1+e)\text{proj}_{q_{N,k+1}}[T_{\mathscr{C}}(q_{N,k+1}); v_{N,k} + \frac{h}{1+e}M^{-1}(q_{N,k+1})F_{N,k+1}) \end{cases} \tag{10.77}$$

where the initial data are those of (10.1), and $F_{N,k+1}$ is an approximate value of $F_{\text{int}}(t,q,v) - F_{\text{ext}}(t)$ that may be chosen as an implicit function $F_{N,k+1} = F(t_{k+1}, q_{N,k+1}, M(q_{N,k+1})v_{N,k+1})$. The projection operator is defined as

$$\text{proj}_q[T_{\mathscr{C}}(q);x] = \begin{cases} x - (x^{\mathrm{T}}\nabla g(q))^+ \frac{M^{-1}(q)\nabla g(q)}{\nabla g^{\mathrm{T}}(q)M^{-1}(q)\nabla g(q)} & \text{if } g(q) \leqslant 0 \\ \\ x & \text{otherwise} \end{cases} \tag{10.78}$$

with $a^+ = \max(0,a)$ for all reals a. One recognizes the projection in the kinetic metric that was already used in (3.130). It is noteworthy that the presence of the matrix $M(q)\dot{q}$ in the right-hand side of (10.77) is odd from a mechanics point of

Algorithm 9 NSCD method. Nonlinear case with Newton's method

```
Require: $\mathsf{M}(\cdot), \mathsf{F}(\cdot), \mathsf{F}_{\mathsf{ext}}(\cdot)$ nonlinear Dynamics (10.2)
Require: $\mathsf{K}_{\mathsf{t}}(\mathsf{t},\mathsf{q},\mathsf{u}) = \nabla_{\mathsf{q}}\mathsf{F}(\mathsf{t},\mathsf{q},\mathsf{u}), \mathsf{C}_{\mathsf{t}}(\mathsf{t},\mathsf{q},\mathsf{u}) = \nabla_{\mathsf{u}}\mathsf{F}(\mathsf{t},\mathsf{q},\mathsf{u})$ tangent operators
Require: $\mathsf{H}^{\alpha}(\cdot), \mathsf{g}^{\alpha}(\cdot),$ for all $\alpha \in \mathsf{I} = \{1\dots\nu\} \subset \mathbb{N}$, kinematic relations (10.67)
Require: $\mathsf{e}, \mu$ frictional contact law.
Require: $\mathsf{t}_0, \mathsf{T}$ time–integration interval
Require: $\mathsf{q}_0, \mathsf{v}_0$ initial data
Require: $\mathsf{h}, \theta, \gamma, \varepsilon,$ time–step and integration and Newton parameters
Ensure: $(\{\mathsf{q}_{\mathsf{k}}\}, \{\mathsf{v}_{\mathsf{k}}\}, \{\mathsf{p}_{\mathsf{k}}\}, \{\mathsf{U}_{\mathsf{k}}\}\{\mathsf{P}_{\mathsf{k}}\}), \mathsf{k} \in \{1,2,\dots\}$

  $k \leftarrow 0$
  $U_0 \leftarrow H^T v_0$
  // Time integration
  while $t_k < T$ do
    // Newton loop
    $\tau \leftarrow 0, \quad \text{error} \leftarrow \infty$
    $q^{\tau}_{k+1} \leftarrow q_k, \quad u^{\tau}_{k+1} \leftarrow v_k$
    $\mathscr{R}(u^{\tau}_{k+1}) \leftarrow h\tilde{F}_{k+\theta}$
    while error $> \varepsilon$ do
      // Evaluate tangent operators
      $\widehat{M}^{\tau+1}_{k+1} \leftarrow (M(\tilde{q}^{\tau}_{k+1}) + h^2\theta^2 K_t(t_{k+1}, q^{\tau}_{k+1}, u^{\tau}_{k+1}) + \theta h C_t(t, q^{\tau}_{k+1}, u^{\tau}_{k+1}))$ (10.37)
      $u^{\tau+1}_{\mathrm{free}} \leftarrow u^{\tau}_{k+1} + \left[\widehat{M}^{\tau+1}_{k+1}\right]^{-1} \mathscr{R}(u^{\tau}_{k+1})$ (10.75)
      // Update of the index set of forecast active constraints
      $\tilde{q}^{\tau}_{k+1} \leftarrow q_k + h\gamma[(1-\theta)v_k + \theta u^{\tau}_{k+1}]$
      $I_a(\tilde{q}^{\tau}_{k+1}) \leftarrow \{\alpha \in I \mid g^{\alpha}(\tilde{q}^{\tau}_{k+1}) \leqslant 0\} \subseteq I$
      //One-step non smooth problem update
      for $\alpha \in I_a$ do
        $U^{\tau+1}_{\mathrm{free}} \leftarrow H^T(\tilde{q}^{\tau}_{k+1}) u^{\tau+1}_{\mathrm{free}}$
        $\widehat{W}^{\alpha\beta,\tau+1}_{k+1} \leftarrow H^{\alpha,T}(\tilde{q}^{\tau}_{k+1}) \left[\widehat{M}^{\tau+1}_{k+1}\right]^{-1} H^{\beta}(\tilde{q}^{\tau}_{k+1})$ for all $(\alpha,\beta) \in I_a^2$
        Assemble (if necessary) $\widehat{W}^{\tau+1}_{k+1}$ with $\widehat{W}^{\alpha\beta,\tau+1}_{k+1}, (\alpha,\beta) \in I_a^2$
      end for
      //Resolution of the one-step non smooth problem
      if $I_a \neq \emptyset$ then
        $[U^{\tau+1}_{k+1}, P^{\tau+1}_{k+1}] \leftarrow$ solution of OSNSP $(\mathscr{P}_{L\tau})$ (see Chap. 13)
      end if
      $p^{\tau+1}_{k+1} \leftarrow \sum_{\alpha \in I_a} H^{\alpha} P^{\alpha,\tau+1}_{k+1}$
      $q^{\tau+1}_{k+1} \leftarrow q_k + h\left[\theta u^{\tau+1}_{k+1} + (1-\theta)v_k\right]$
      $\mathscr{R}(u^{\tau+1}_{k+1}) \leftarrow M(\tilde{q}^{\tau}_{k+1})(u^{\tau+1}_{k+1} - v_k) + h\tilde{F}_{k+\theta}$
      error $\leftarrow \|R(u^{\tau}_{k+1}) + p^{\tau+1}_{k+1}\|$
      $\tau \leftarrow \tau + 1$
    end while
    // State update
    $p_{k+1} \leftarrow p^{\tau+1}_{k+1}, \quad q_{k+1} \leftarrow q^{\tau+1}_{k+1}$
    $v_{k+1} \leftarrow u^{\tau+1}_{\mathrm{free}} + \left[\widehat{M}^{\tau+1}_{k+1}\right]^{-1} p^{\tau+1}_{k+1}$
    $t_k \leftarrow t_{k+1}, \quad k \leftarrow k+1$
  end while
```

view, as this term is nothing else but the generalized momentum that is a variable usually associated with Hamiltonian mechanics.

One finally defines the approximate solutions as the piecewise linear functions $q^N(t) = q_0 + \int_0^t v^N(s)\mathrm{d}s$ and

$$v^N(t) = \begin{cases} v_{N,k} \text{ if } t \in [t_k, t_{k+1}),\ 0 \leqslant k \leqslant N-1 \\ \\ v_{N,N,} \text{ if } t = T. \end{cases}$$

Theorem 10.7. *Let $R > \dot{q}_0^T M(q_0)\dot{q}_0$. Then there exists $T(R) > 0$ such that for any solution $(q(\cdot), v(\cdot))$ of the sweeping process (3.115) and (3.128) defined on $[0, \tilde{T}]$ ($0 < \tilde{T} \leqslant T$) the following estimates hold for all $t \in [0, \min(\tilde{T}, T(R))]$:*

$$||q(t) - q_0|| \leqslant R,\ \ v^{\mathrm{T}}(t)M(q(t))v(t) \leqslant R^2$$

Moreover there exists a subsequence of $\{q^N(\cdot), v^N(\cdot)\}_{N \geqslant 1}$, still denoted as $\{q^N(\cdot), v^N(\cdot)\}_{N \geqslant 1}$, such that $q^N(\cdot) \to q(\cdot)$ in $C^0([0, \min(T, T(R))], \mathrm{I\!R}^n)$, $v^N(\cdot) \to v(\cdot)$ pointwise in $[0, \min(T, T(R))]$, and $(q(\cdot), v(\cdot))$ is a solution of the sweeping process (3.115) (3.128) on $[0, \min(T, T(R))]$.

The same type of proof is derived in Dzonou & Monteiro Marques (2007), however, only a local result with $e = 0$ is treated.

Remark 10.8. The way the discretized sweeping process is written in (10.76) and (10.77) has a pure mathematical interest. Indeed in the numerical practice, implementing a projection on a convex set is far from trivial. One has to perform subsequent steps to put the inclusion under a suitable form so that numerical simulation can be performed (in other words, the only things one is able to solve efficiently are CPs or NonLinear Programming (NLP)s). Thanks to the polyhedrality of $T_{\mathscr{C}}(q)$, those steps can be done.

10.1.7 Bilateral and Unilateral Constraints

Bilateral constraints may be treated either by

- (i) reducing the system's dimension by eliminating some coordinates
- or (ii) introducing Lagrange multipliers.

If solution (i) is chosen, nothing has to be added to the above developments once the new generalized coordinates have been obtained. Let solution (ii) be chosen, where the bilateral constraints are given by $f(q) = 0$ with $f(q) \in \mathrm{I\!R}^m$. The unilateral constraints are still $g^\alpha(q) \geqslant 0$, $1 \leqslant \alpha \leqslant \nu$.

We may generically write the Lagrange equations in (10.1) as

$$\begin{cases} M(q)\mathrm{d}v + F(q, v^+, t)\mathrm{d}t = r_\mu \mathrm{d}t + \mathrm{d}r_\lambda \\ \\ r_\mu = \nabla F(q)\mu,\ \ \mathrm{d}r_\lambda = \nabla g(q)\mathrm{d}\lambda \\ \\ F(q) = 0,\ \ 0 \leqslant g(q) \perp \lambda \geqslant 0 \end{cases} \tag{10.79}$$

where $\mu(\cdot) = [\mu_1, \dots, \mu_m]^{\mathrm{T}}$ is a function of time, whereas $\mathrm{d}\lambda = [\mathrm{d}\lambda^1, \dots, \mathrm{d}\lambda^{\nu}]^{\mathrm{T}}$ is a measure.

The Mixed Linear OSNSP with Bilateral Constraints

Let us consider the problem $(\mathscr{P}_{\mathrm{ML}})$ in which we add m perfect linear bilateral constraints

$$F(q) = Gq + b = 0, \tag{10.80}$$

where the Jacobian matrix of these constraints $\nabla F^{\mathrm{T}}(q)$ is given by $G^{\mathrm{T}}(q) \in \mathbb{R}^{n\times m}$. We will denote this OSNSP by $(\mathscr{P}_{\mathrm{ML}_b})$

$$(\mathscr{P}_{\mathrm{ML}_b}) \begin{cases} \widehat{M}(v_{k+1} - v_{\mathrm{free}}) = p_{k+1} + GP_{\mu,k+1} \\ \\ G^{\mathrm{T}} v_{k+1} = 0 \\ \\ U^{\alpha}_{k+1} = H^{\alpha,\mathrm{T}}\, v_{k+1} \\ \\ p^{\alpha}_{k+1} = H^{\alpha}\, P^{\alpha}_{k+1} \\ \\ \text{If } \; g^{\alpha}(\tilde{q}_{k+1}) \leqslant 0 \; \text{ then} \\ \mathbf{C}^{\alpha,*} \ni \left[U^{\alpha}_{\mathrm{N},k+1} + eU^{\alpha}_{\mathrm{N},k} + \mu^{\alpha}\, ||U^{\alpha}_{\mathrm{T},k+1}||, U^{\alpha}_{\mathrm{T},k+1}\right]^{\mathrm{T}} \perp P^{\alpha}_{k+1} \in \mathbf{C}^{\alpha} \\ \\ \text{If } \; g^{\alpha}(\tilde{q}_{k+1}) > 0 \; \text{ then } \; P^{\alpha}_{k+1} = 0 \end{cases}$$

The treatment of the bilateral constraints is made in a natural way at the velocity level. Clearly, the bilateral constraints at the position level are satisfied at the end of the time step if they are satisfied at the initial time. Indeed, we have

$$\begin{aligned} F(q_{k+1}) &= G\,q_{k+1} + b \\ &= G\,(q_k + h[(1-\theta)v_k + \theta v_{k+1}]) + b \\ &= G\,(q_k) + b \\ &= F(q_k). \end{aligned} \tag{10.81}$$

This property is no longer true with nonlinear bilateral constraints. This problem is related to the index of the underlying Differential Algebraic Equation (DAE).

Nonlinear Bilateral Constraints

Let us for the sake of simplicity consider the linear time-invariant case, and let us perform the same steps as in Sect. 10.1.1. We obtain

$$v_{k+1} = v_{\text{free}} + \widehat{M}(p_{\mu,k+1} + p_{\lambda,k+1}) \tag{10.82}$$

where p_μ and p_λ are the impulses corresponding to $r_\mu(\cdot)$ and $\mathrm{d}r_\lambda$, respectively. Notice that $\nabla F^{\mathrm{T}}(q(t))v^+(t) = 0$. Introducing Moreau's inclusion and the expression for q_{k+1} in (10.7) the discretized problem may be written as

$$\begin{cases} p_{\lambda,k+1} \in -N_{T_{\mathscr{C}}(\tilde{q}_{k+1})}\left(\frac{v_{k+1}+ev_k}{1+e}\right) \\ \nabla F^T(q_{k+1})v_{k+1} = 0 \\ v_{k+1} = v_{\text{free}} + \widehat{M}(p_{\mu,k+1} + p_{\lambda,k+1}). \end{cases} \tag{10.83}$$

It will be seen in Sect. 13.3.2 how the set of equations in (10.83) may be solved.

10.2 Some Numerical Illustrations of the NSCD Method

This section is devoted to summarize the numerous results that have been obtained with the Moreau–Jean's NSCD method, in various application domains.

10.2.1 Granular Material

Let us report some simulation results which concern granular material. The NSCD method has been tested on various granular matter systems and has been shown to enable one to reproduce well-known important macroscopic phenomena.

A vertically shaken cylindrical vessel with 3999 beads with diameter 0.2 cm, and one bead with diameter 0.5 cm is simulated in Moreau (1994a). The vessel has a diameter 3.5 cm. A 3-parameter contact law as in (3.168) is chosen. The friction coefficient is $\mu = 0.8$, the normal restitution is $e_{\mathrm{N}} = 0.95$, the tangential restitution is $e_{\mathrm{T}} = 0.4$. These values hold at all contacts, i.e., between the beads and between the beads and the vessel boundaries. The vessel is shaked vertically with frequency 25 Hz, and peak-to-peak amplitude 0.2 cm, a motion with maximal acceleration equal to $2.51g$. This represents a large acceleration that makes the pack of 4000 beads lose contact with the vessel bottom for a part of each period. The propagation of collisions when the lowest beads hit the vessel bottom therefore induces a strong agitation in the whole pack. The set of numerical simulations presented in Moreau (1994a) demonstrates that bulk segregation, or size segregation, that is the tendency of large objects to migrate upward with respect to the surrounding smaller ones, is reproduced by the simulation. The importance of boundary effects between the beads and the vessel is pointed out: peripheral beads experience large downward forces. This size segregation, known as the Brazil nuts effect, is shown to occur also with different parameters: zero friction at the vessel boundary, and $e_{\mathrm{N}} = 0.9$. Between the beads one takes $e_{\mathrm{T}} = 0.4$, $e_{\mathrm{N}} = 0.9$, and $\mu = 0.5$. The vessel with diameter 3 cm is filled with 2000 beads with diameter between 0.2 and 0.1 cm, and 200 beads with diameter 0.02 cm. The vertical motion of the vessel is sinusoidal, frequency 20 Hz, peak-to-peak amplitude 0.25 cm. The 200 smaller beads are initially placed on top. They undergo large ballistic flights and are progressively captured in the bigger beads layer. After about 100 shakes, all the smaller beads are trapped at bottom. Other numerical results may be found in Radjai & Wolf (1998) and Moreau (1994b).

Simulation results for a Couette granular flow are presented in Jean (1999) and Zervos et al. (2000). Samples of 1200, 2400, 4000 and 16,000 polydisperse disks or rolls are kept within two drums. The outer drum is a membrane subject to a constant pressure 75 kPa. The inner drum rotates with a constant speed 0.1 tr/min. The friction coefficient between the grains is $\mu = 0.5$ and it is $\mu = 0.75$ between the grains and the drums. It is observed that the sample behaves quasi-rigidly, except close to the boundary of the inner drum. An interface layer forms near the internal rotating disk, with thickness five times the mean diameter of the grains. Inside this layer, tangential displacements localize and present a steep gradient that fades out almost exponentially with the radial position. The numerical results are compared to experimental results (Daudon et al., 1997; Lerat et al., 1995) and are shown to provide a good qualitative prediction of the process outcome.

Several sets of numerical experiments are reported in Renouf & Alart (2004a): a depositing process of particles in a 1 m × 1 m-size box under gravity, with 1000–33,000 disks, and 7200 disks with mean diameter 3 mm in a rotating drum with diameter 450 mm, with angular velocity 3 rpm. The friction parameter is $\mu = 0.4$ and $e_{\text{N}} = 0.92$. It is noteworthy that the number of unilateral contacts with friction that is treated in such examples is of several tenth of thousands and corresponds to the size of the complementarity problems to be solved by the one-step nonsmooth problem solver. Various test configurations are reported in Renouf & Alart (2004a): biaxial test (a constant pressure is applied on the right boundary and a constant velocity is applied to the upper side of the sample), shear tests (an angular velocity is imposed to the lateral sides of the square and a constant pressure is maintained on the upper side), free surface compaction (the sample of disks is compacted by moving one of the lateral walls of the box with a nonmonotone speed). The objective of Renouf & Alart (2004a) is to compare different one-step nonsmooth problem solvers, using the NSCD method. This specific comparison is outside the scope of this chapter and is examined in Chap. 13. Further results may be found in Renouf et al. (2005b,c).

Let us finish this short overview on numerical experiments for granular media with an application concerning ballast modeling (Saussine et al., 2004a,b 2006).The grains are supposed to be pentagonal (Saussine et al., 2006) in the 2-dimensional case or convex polyhedrons (Saussine et al., 2004b) in the 3-dimensional case. An algorithm dedicated to the determination of geometrical intersections between convex polyhedrons has been developed (Saussine, 2004) and is used in these numerical experiments. Ballast settlement is the result of several millions of loading cycles. Each cycle corresponds to the passing of an axle on the rails. The numerical experiments in Saussine et al. (2006) are made of 361 grains with diameter 1 cm, 242 grains with diameter 1.5 cm and 121 grains with diameter 2 cm. The contact parameters are $e_{\text{N}} = 0$ and $\mu = 0.5$. The grains are let to fall under gravity on a sublayer with a stiffness of 411 N/m and a viscosity of 80 N/m s^{-2}. The loadings are sinusoidal forces $F(t) = -1000 - 500\cos(40\pi t + \pi)$, and several thousands of loading cycles are applied to the sample (up to 20,000 cycles). Quite interestingly the numerical results are compared to experimental results, and show good agreement. The 3-dimensional test in Saussine et al. (2004b) are made with samples composed of 25,000 grains prepared by deposition under gravity.

Remark 10.9. Such numerical experiments that involve a very large number of contacts usually are very long. For instance it is reported in Saussine et al. (2006) a computation time of 3 weeks on a Pentium 4 (2.5 GHz) Unix station. This is what motivates works on parallelization (Renouf & Alart, 2004b; Renouf et al., 2004; Alart et al., 2003). This does not mean that such time-stepping schemes are impractical for real-time applications, see below.

It is noteworthy that the NSCD method has had a significant impact in the Physics community for the study of granular media (see, e.g., Radjai et al., 1996, 1997, 1998, 1999; Bratberg et al., 2002; Nouguier et al., 2000; Radjai, 1999; Radjai & Roux, 2002).

10.2.2 Deep Drawing

Applications concerning deep drawing are also presented in Jean (1999) and Jourdan et al. (1998a,b). The NSCD method is tested in SIMEM3, a deep drawing simulation software of the car company Renault. The Numisheet'93 congress benchmark example (a U bending example) is used. Spring-back effects are taken into account. Results are better than those obtained with other software packages (implicit methods with quasi-static model and Newton–Raphson iteration, for which convergence problems appear due to ill-conditioned stiffness matrix, or explicit methods with dynamic model for which very small time steps are needed). Comparisons with experiments show good agreement with the numerics. Convergence results are proved in Jourdan et al. (1998a) with a nonlinear block Gauss–Seidel one-step nonsmooth problem solver.

10.2.3 Tensegrity Structures

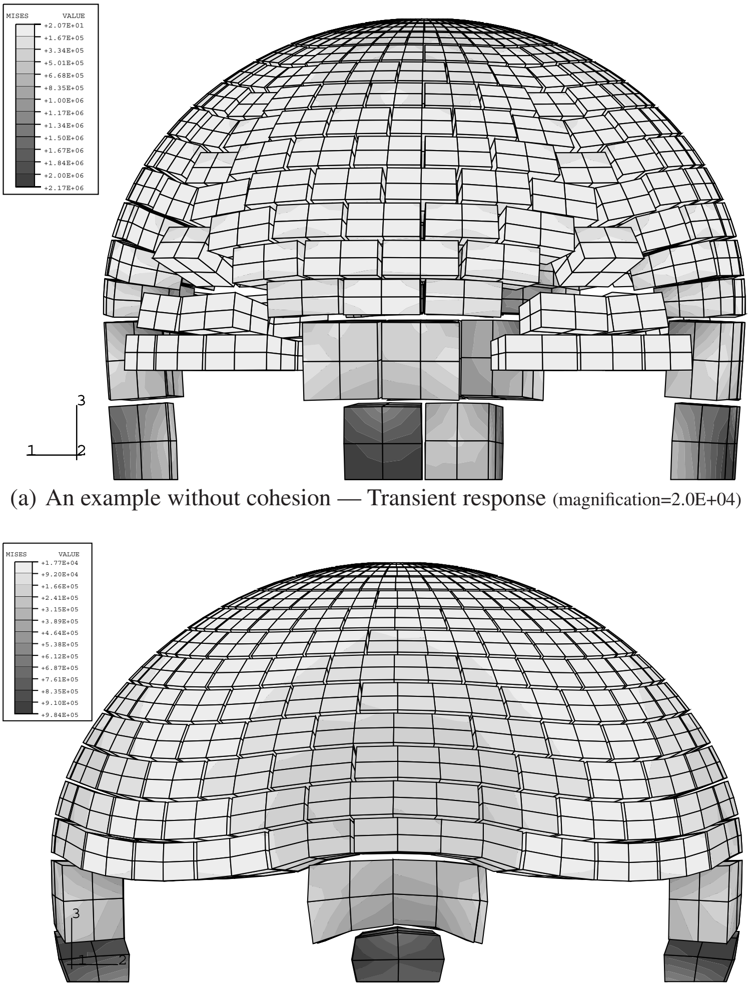
Tensegrity structures are made of cables and beams, with "unilateral" cables: the cables are either infinitely rigid in the compression sense and flexible in the extension sense or the contrary. In Motro (2006) the next definition is given: *systems in a stable selfstress state including a discontinuous set of compressed components inside a continuum of tensioned components*. Depending on the cables being elastic extendable or not, some complementarity relations between the total stress in the cable and the deformation may be written. This gives rise to mechanical systems with complementarity conditions that lend themselves to a treatment with the NSCD method. Results may be found in Nineb et al. (2005, 2006).

10.2.4 Masonry Structures

The NSCD method is applied to the simulation of masonry structures in Acary & Jean (1998, 2000), Acary et al. (1999), Acary (2001), Jean et al. (2001).

A numerical test dome with diameter 10 m, 300 granite blocks set on four pillars, subjected to gravity is depicted in Fig. 10.1. The blocks are composed of eight H8 finite elements, and the structure is composed of 2400 H8 elements, 8100 nodes, and has twenty four 300 degrees of freedom. Each block face has 16 candidate points to contact, and the total number of candidates is 9176. Enhanced contact laws are used (cohesive frictional laws).

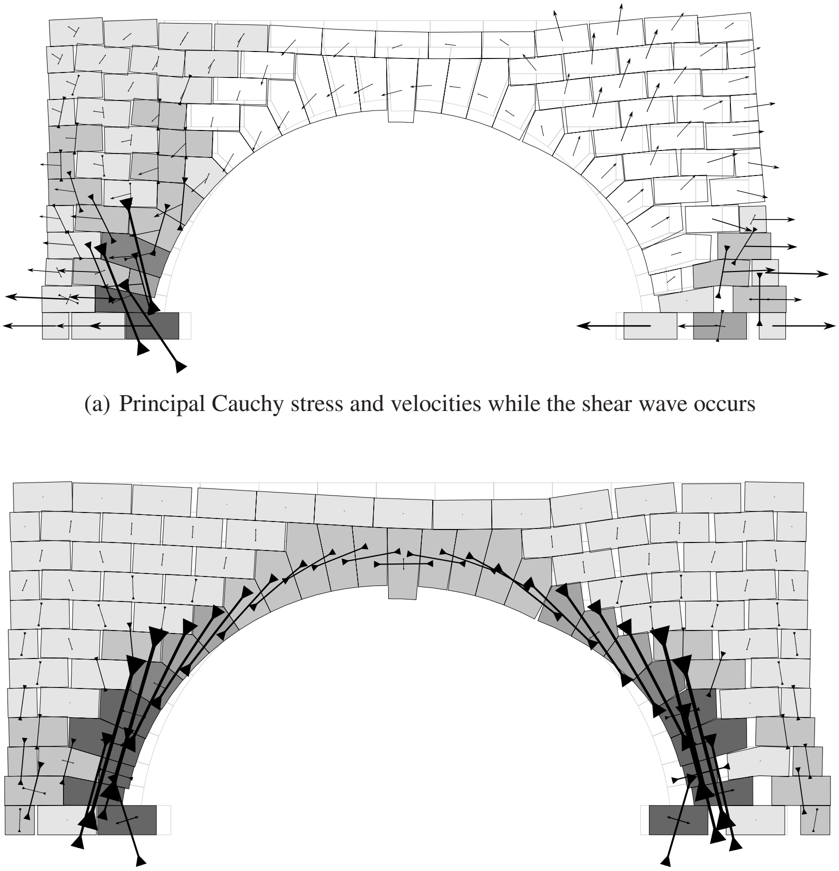
Other numerical examples illustrate the ability of the NSCD method to simulate fracture processes in divided materials. See Figs. 10.2 and 10.3.



(a) An example without cohesion — Transient response (magnification=2.0E+04)

(b) With cohesion, equilibrium state (magnification=3E+06)

Fig. 10.1. Dome on four pillars under gravity load



(a) Principal Cauchy stress and velocities while the shear wave occurs

(b) Principal Cauchy stress after the shear wave solicitation

Fig. 10.2. Earthquake simulations. An arch-bridge subjected to a shear wave (×200)

10.2.5 Real-Time and Virtual Reality Simulations

The NSCD method may also adapt itself to real-time simulations (Renouf et al., 2005a) and virtual reality (Kaufman et al., 2005). A simulation of 2016 chess pieces falling through a hopper and stacking is presented in Kaufman et al. (2005). The number of colliding bodies and the number of detected contacts are presented, as well as the wall-time as a function of the total number of contacts in a step. The implementation is linear in the total number of bodies being simulated and in the total number of contact points detected at each step (about 10^5). The frictional model is altered to accelerate the simulation process.

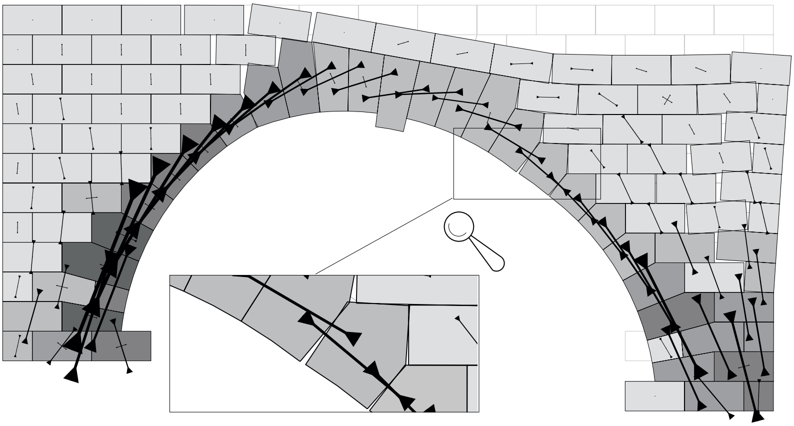


Fig. 10.3. Stress after a ground settlement $-4\,\mathrm{cm}(\times 10)$

The real-time capabilities of the NSCD method are demonstrated in Renouf et al. (2005a) on two samples of dense assemblies: sphere settling and a ball hitting a masonry structure, as depicted in Fig. 10.4.

Spheres Settling

The following numerical parameters are chosen: the time step value h, and $g_{\max}$ as the maximal interpenetration between contactors in the sample. The physical parameters are $e_{\mathrm{N}} = 0.4$, $\mu = 0.4$, and $e_{\mathrm{T}} = 0$. The rule is to find h such in order to preserve a value of $g_{\max}$ as small as possible, to ensure the quality of the simulation and the real-time constraint. The ratio between the simulated time and the elapsed CPU time is denoted as Sp. The real-time constraint will be preserved if $Sp \geqslant 1$. We performed settling in a box with a frictional contact interaction law and using different numbers of spheres: 80, 160, and 320. Simulation results are shown in the Table 10.1. The

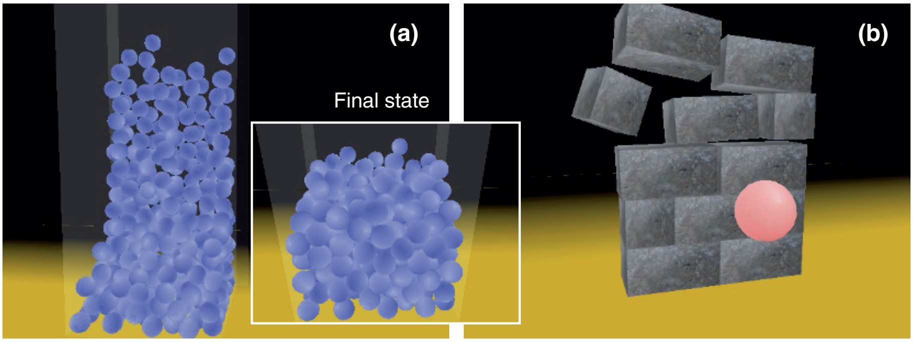


Fig. 10.4. Sphere settling and ball hitting a brick wall

Table 10.1. Results of simulation of spheres settlings

n_b	n_c	h	Sp	$g_{\max}(\%)$
80	271	0.02	1.24	0.3
80	267	0.04	2.28	1.2
160	587	0.02	0.95	0.5
160	584	0.04	1.50	1.8
320	1218	0.02	0.50	0.6
320	1275	0.04	0.75	2.2

calculations have been led on an Opteron 242 with a 2 GHz processor. n_b is the number of spheres and n_c is the number of contacts.

The simulations of samples composed of 80 and 160 spheres respect the real-time constraint and keep a good simulation quality (less than 2% of constraint violation). For the bigger samples (320 spheres), it is difficult to preserve both the time constraint and the quality of the solution. Nevertheless the value of the speed-up (0.75) is not so small and some numerical optimization should allow one to obtain the respect of the time constraint. Note that for the frictionless packing, the real-time constraint is reached for a larger sample of 800 spheres: the time needed by the solver is smaller due to the smaller number of unknowns.

The difficulty in this kind of simulation is the large variation in the number of contacts. It increases quickly to reach a stabilized value (e.g., Fig. 10.5b)). The large number of status modifications does not help the frictional contact solver to reach a solution. During the settling, the number of iterations N_{it} reaches the maximal value ($t \in [0,15]$) and during a stabilization phase ($t \in [15,35]$) N_{it} has erratic variations from the minimal to the maximal value to keep a stabilized evolution below. The fact that iterative methods can benefit from the solution of the previous time step to initialize the algorithm is one of the reasons of the quasi-smooth evolution. Moreover with an iterative method a good approximation of the solution is obtained quickly

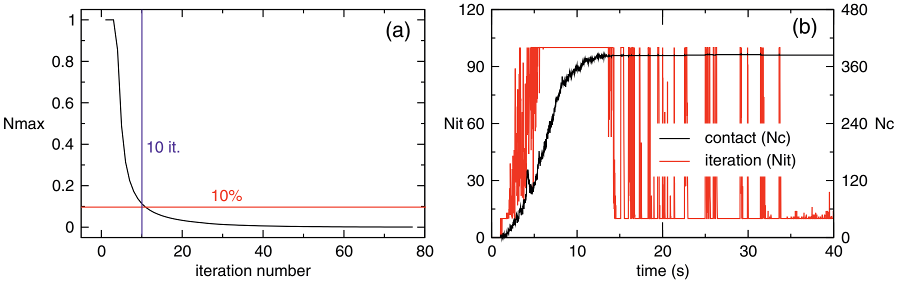


Fig. 10.5. (a) Evolution of the convergence criterion during the iterative process. **(b)** Parallel between the evolution of iterations and the number of contacts during a whole simulation process

as shown in Fig. 10.5(a). For a sample with 600 contacts, an approximation of the solution with an error of 10% is obtained in 10 iterations only.

Virtual Masonry

Some difficulties are related to the numerical simulation of masonry structures. The first one is linked to the location of the contact points and their number. When we consider two blocks in a face/face contact, four dependent contacts at least are considered. This strategy, which preserves the time taken by the detection algorithm, handicaps direct methods as Lemke's algorithm. The second difficulty concerns the introduction of friction. When the problem is formulated as an LCP, the matrix of the LCP is no longer symmetric. This appears as a problem for Lemke as well as the PATH solvers. It may be related to the observation of Klarbring on the class of matrices unsolvable by LCP solvers (Klarbring, 1986a).

10.2.6 More Applications

In Le Saux et al. (2005) Moreau's time-stepping scheme is implemented to simulate the dynamics of a rolling disk on a flat support, with various kinds of friction models (resistance against sliding, pivoting, and rolling). Comparisons of the energy decay calculated analytically and the numerical energy profiles show good agreement. Comparisons with available experimental results are also done, which allow the authors to determine the dominant mechanism of dissipation in this process: the contour friction, which models resistance against the movement of the contact point along the contour of the disk. In Transeth et al. (2006b) a snake robot with 11 links is studied when dropped on the floor, and in Transeth et al. (2006a) when it undulates between obstacles. The dynamics of fracture has been studied numerically in Dubois (2005) and Acary & Monerie (2006). Mechanical systems with Coulomb friction and time-delayed terms are considered in Lamarque et al. (2003). The NSCD method is tested in Pratt et al. (2007) on systems with bilateral constraints and Coulomb friction as in Sect. 3.11. Thorough numerical tests and comparisons with analytical results show that the time-stepping NSCD scheme may approximate correctly trajectories with a large number of sticking–sliding transitions. However, when the number of transitions is too large, the time-stepping scheme no longer provides correct results in terms of the time t_{rest} when the system comes to rest. This is explained in Pratt et al. (2007) by the fact that the number of events is growing exponentially with the choice of the parameters and initial data that are made therein, but the NSCD time-stepping scheme is unable to calculate such an exponential increase. In such a case, since the dynamics between the events is easily calculable analytically, it is possible that an event-driven method would supersede the time-stepping scheme. This conclusion seems to be in contrast with the usual statement that time-stepping schemes behave better than event-driven ones when the number of events is too large. Notice, however, that in general this may still be the case because implementing an event-driven scheme when the trajectories between the events are not analytically calculable necessitates an accurate numerical detection of the events which may rapidly prevent the use of an event-driven procedure. The results of Pratt et al. (2007) rather point out one possible deficiency of the NSCD time-stepping method.[4] Results on the woodpecker toy may be found in Glocker & Studer (2005).

10.2.7 Moreau's Time-Stepping Method and Painlevé Paradoxes

The coupling of unilaterality and Coulomb friction may result in so-called frictional paroxysms or Painlevé paradoxes, see Sect. 6.2. As demonstrated in the seminal paper (Moreau, 1988b) on an example, the discretized sweeping process is able to handle such behaviors since it is a true discretization of the maximum dissipation principle. This strong property of impulse–velocity time-stepping schemes is also pointed out in Anistescu (2006).

10.3 Variants and Other Time-Stepping Schemes

10.3.1 The Paoli–Schatzman Scheme

This time-stepping method originates from the dynamics presented in Sect. 3.5, especially (3.113). It is supposed that the admissible domain $\mathscr{C}$ of the configuration space is either of class C^3 (i.e., one can find a C^3 function $g(\cdot)$ such that $\mathscr{C} = \{q \in \mathrm{I\!R}^n \mid g(q) \geqslant 0\}$) or finitely represented as in (3.16) with C^1 time-invariant functions $g^{\alpha}(\cdot)$ (in this latter case one has $e_{\mathrm{N}} = 0$). The initial data in (3.113) are $q(0) = q_0$ and $p(0) = M(q_0)\dot{q}(0) = p_0$. The initial conditions for the discretized algorithm is q_0 at step 0, and $q_1 = q_0 + hM^{-1}(q_0)p_0 + hz(h)$, where $z(h) \to 0$ as $h \to 0$. Given q_{k-1} and q_k, q_{k+1} is given by

$$\begin{cases} q_{k+1} = -eq_k + (1+e)\mathrm{proj}\left[\mathscr{C}; \dfrac{2q_k - (1-e)q_{k-1} + h^2 F_k}{1+e}\right] \\[2ex] F_k = F\left(t_k, q_k, q_{k-1}, \dfrac{q_{k+1} - q_{k-1}}{2h}, h\right). \end{cases} \tag{10.84}$$

F_k is the approximate of $F(t,q,q,v,0) = M^{-1}(q)f(t,q(t),p)$ with $p = M(q)v$ in (3.113). Defining the discrete velocity as $v_k = \frac{q_{k+1}-q_k}{h}$ one may rewrite it as

$$v_k - v_{k-1} - hF_k = \frac{(1+e)(z_k - w_k)}{h} \tag{10.85}$$

with $w_k = \frac{2q_k-(1-e)q_{k-1}+h^2F_k}{1+e}$, $z_k = \frac{q_{k+1}+eq_{k-1}}{1+e} = \mathrm{proj}[\mathscr{C}; w_k]$. As we already pointed out above, such formulations are quite impractical when implementation is to be

[4] Surprisingly enough, no mention of the values of the time step $h > 0$ is made in Pratt et al. (2007). It would have been extremely interesting to present the variation of t_{rest} as a function of h: since the solutions converge, t_{rest} necessarily decreases as $h \to 0$.

envisaged. This is why none of the above experimental numerical results are implemented this way. Indeed calculating a projection onto a convex set is not an easy task. It is therefore of some interest to rewrite the algorithm in (10.84) under a more tractable form. Using (A.8) one obtains

$$M(q_k)[q_{k+1}-2q_k+q_{k-1}]-h^2M(q_k)F_k \overset{\Delta}{=} p_{k+1} \in -N_{\mathscr{C}}\left(\frac{q_{k+1}+eq_{k-1}}{1+e}\right). \quad (10.86)$$

The interest for performing this step is that provided $\mathscr{C}$ is finitely represented, the inclusion into the normal cone can be, under some constraint qualification, recast into a nonlinear complementarity problem:

$$\begin{cases} w_{k+1} = g\left(\dfrac{q_{k+1}+eq_{k-1}}{1+e}\right) \\ \\ p_{k+1} = \nabla g\left(\dfrac{q_{k+1}+eq_{k-1}}{1+e}\right)\mu_{k+1} \\ \\ 0 \leqslant w_{k+1} \perp \mu_{k+1} \geqslant 0. \end{cases} \quad (10.87)$$

If $g(q)=Aq+B$, i.e., the admissible domain is a polyhedral set, then one obtains

$$\begin{cases} w_{k+1} = A\left(\dfrac{q_{k+1}+eq_{k-1}}{1+e}\right)+B \\ \\ p_{k+1} = A^{\mathrm{T}}\mu_{k+1} \\ \\ 0 \leqslant w_{k+1} \perp \mu_{k+1} \geqslant 0. \end{cases} \quad (10.88)$$

This can be in turn rewritten as

$$0 \leqslant \mu_{k+1} \perp AM^{-1}(q_k)A^{\mathrm{T}}\mu_{k+1}+2q_k-q_{k-1}+h^2M^{-1}(q_k)F_k \geqslant 0 \quad (10.89)$$

that is an LCP with unknown μ_{k+1}.

Remark 10.10. The multiplier that belongs to the normal cone in Moreau's time-stepping scheme in (10.38) has a natural and physical interpretation as the approximation of an impulse, because the argument in the normal cone is the velocity. Such is not the case for the Paoli–Schatzman scheme in which the multiplier μ_{k+1} has no clear mechanical meaning, the argument in the normal cone in (10.86) being some combination of positions. Another advantage of Moreau's sweeping process (of order 2) is that the sets $T_{\mathscr{C}}(q)$ and $N_{T_{\mathscr{C}}(q)}(v)$ are polyhedral sets, independently of $\mathscr{C}$. This is not the case in the Schatzman–Paoli formulation, where the projection is done directly on $\mathscr{C}$, not on $T_{\mathscr{C}}(q)$.

When a contact is detected with the boundary of $\mathscr{C}$, the scheme in (10.84) reverses the normal velocity in two steps. This may be an issue, so modifications have been proposed so that the normal velocity is reversed in one step only Nqi, 1997.

Convergence results have been shown in Paoli & Schatzman (2002a,b) and Paoli (2005b).

10.3.2 The Stewart–Trinkle–Anitescu–Potra Scheme

These authors have presented several variants and extensions of the fundamental velocity–impulse algorithm of the sweeping process (Stewart & Trinkle, 1996; Stewart, 2000; Anitescu et al., 1999; Anitescu & Potra, 1997; Potra et al., 2006). The main contributions are the introduction (following Klarbring, 1986b; Klarbring & Björkman, 1988) of a special way to represent the 3-dimensional Coulomb friction (see Chap. 13) and proofs of existence and convergence. We may classify these results in two main parts:

- Convergence of the solutions of the time-stepping algorithm with facetized Coulomb friction, and one contact (Stewart 1998). This seems to be the only proof of convergence for time-stepping schemes when friction is present, and is an important extension of the results in Monteiro Marques (1993). It includes the Painlevé issues.
- Existence of solutions to LCPs or NCPs for the one-step nonsmooth problem, see e.g., Anitescu & Potra (1997), and reformulation of the one-step nonsmooth problem (Anistescu, 2006; Anitescu & Hart, 2004). This will be tackled in Chap. 13.

Simulation results are presented in Anistescu (2006); Anistescu & Hart, 2004; Anitescu & Hart, 2004). They essentially aim at illustrating some properties of the presented methods on low-dimensional systems. Size segregation is demonstrated in Anitescu & Hart (2004) and Anistescu (2006) with 210 disks in a 2-dimensional vessel with $\mu = 0.5$ and $e_{\mathrm{N}} = 0.5$. Other numerical results can be found in Stewart & Trinkle (1996), Trinkle et al. (2001) and Son et al. (2004). A falling rod and a chain of four balls are presented in Stewart & Trinkle (1996), a robotic application may be found in Son et al. (2004), and a 3-dimensional sphere on a rough plan or on a spherical surface is presented in Trinkle et al. (2001).

Remark 10.11. It is noteworthy that most of the numerical experiments that are presented in this chapter mainly focus on the choice of the one-step nonsmooth problem solver.

Remark 10.12. In Stewart & Trinkle (1996), the treatment is similar to the NSCD method of Moreau (1998) and Jean & Moreau (1992) except that the numerical formulation ensures that there is no interpenetration of rigid bodies. Indeed, as discussed in Jean (1999), it is possible to directly impose a constraint on the position and to associate the impulse as complementary variable when we use a purely backward Euler scheme and an inelastic impact law. This is mainly due to the fact that the gap is positively homogeneous to the relative velocity. Unfortunately, this trick is no longer possible when one wants an elastic impact law and when we use a θ-method. Other tricks can be found in Jean (1999) to attempt to satisfy both the impact law and the constraint on the position with a θ-method.

11

Time-Stepping Scheme for the HOSP

This chapter is dedicated to present a time-stepping method for the higher order Moreau's sweeping process (the HOSP) described in Chap. 5. We start by presenting some simple examples which prove that the backward Euler method for LCS (see (9.75)) does not work for the HOSP systems (5.1) where the relative degree $r \geqslant 2$. The material of this chapter is taken from Acary et al. (in press). The time-stepping scheme that is presented in the following sections is constructed to approximate the solutions of the measure differential formalism in (5.22)–(5.24). Using Proposition 5.5 one can recover the solutions of the distributional formalism in (5.17)–(5.20). This is because the time-stepping schemes we are using are able to approximate the measure of an interval. But they are not able to approximate distributions of any degree.

11.1 Insufficiency of the Backward Euler Method

In Sect. 9.5, it has been seen that a simple backward Euler method may provide good results (convergence) when it is applied to some linear complementarity systems with relative degree $r = 0$ or $r = 1$. In particular, passive LCS lend themselves well to such discretization, see Theorem 9.30. When the relative degree is larger, the scheme in (9.75) no longer works. This is illustrated by few examples.

Example 11.1. Let us consider an LCS with the following matrix definition:

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \; B = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \; C = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}, \; D = 0\,. \tag{11.1}$$

The relative degree r of this LCS is equal to 2 ($D = 0, CB = 0, CAB \neq 0$). If we consider the initial data $x_0 = (0, -1, 0)^{\mathrm{T}}$, we obtain by a straightforward application of the scheme (9.75) the following solution:

$$x_k = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \; \forall k \geqslant 1, \tag{11.2}$$

$$\lambda_1 = \frac{1}{h}, \quad \lambda_k = 0, \; \forall k \geqslant 2. \tag{11.3}$$

We can remark that the multiplier λ_1 which is the solution of the LCP at the first step, tends toward $+\infty$ when h vanishes. In this example, the state $x(\cdot)$ seems to be well approximated but both the LCP matrix and the multiplier tend to inconsistent values when h vanishes. This inconsistency is just the result of an attempt to approximate the point value of a distribution, which is nonsense.

If we consider now the initial data $x_0 = (-1,-1,0)^{\mathrm{T}}$, we obtain the following numerical solution from (9.75) :

$$x_k = \begin{pmatrix} k \\ \\ \dfrac{1}{h} \\ \\ 0 \end{pmatrix}, \forall k \geqslant 1, \tag{11.4}$$

$$\lambda_1 = \frac{1}{h^2}, \quad \lambda_k = 0, \forall k \geqslant 2. \tag{11.5}$$

With such an initial data, the exact solution should be $x_k = 0, \forall k \geqslant 1$. We can see that there is an inconsistency in the result because the first component of the approximate state does not depend on the time step. We cannot expect that this approximation converges to the exact solution.

Example 11.2. Let us consider another simple example:

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, B = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, C = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}, D = 0. \tag{11.6}$$

In this case, the relative degree r is equal to 3. The direct discretization of the system leads to the same problem as in the previous example even in the case where the initial data satisfies the constraints. Let us consider $x_0 = (0,-1,0)^{\mathrm{T}}$. From (9.75), we obtain the following numerical solution:

$$x_k = \begin{pmatrix} \dfrac{k(k+1)}{2h} \\ \\ k \\ \\ \dfrac{1}{h} \end{pmatrix}, \forall k \geqslant 1, \tag{11.7}$$

$$\lambda_1 = \frac{1}{h^2}, \quad \lambda_k = 0, \forall k \geqslant 2. \tag{11.8}$$

This solution cannot converge to an analytical solution.

11.2 Time-Discretization of the HOSP

11.2.1 Principle of the Discretization

Let us start with a generic equation of the measure differential formalism for the extended sweeping process (5.22) for $1 \leqslant i \leqslant r-1$,

$$\begin{cases} \mathrm{d}z_i - z_{i+1}(t)\mathrm{d}t = \mathrm{d}\nu_i, \\ \mathrm{d}\nu_i \in -\partial\psi_{T_\Phi^{i-1}(Z_{i-1}(t^-))}(z_i(t^+)) \end{cases}. \tag{11.9}$$

It results that an evaluation of this MDI on the time interval $(t_k, t_{k+1}]$ yields

$$\begin{cases} \mathrm{d}z_i((t_k,t_{k+1}]) - \int_{(t_k,t_{k+1}]} z_{i+1}(\tau)\mathrm{d}\tau = \mathrm{d}\nu_i((t_k,t_{k+1}]) \\ \mathrm{d}\nu_i((t_k,t_{k+1}]) \in \overline{\mathrm{conv}}\big(\cup_{\tau\in(t_k,t_{k+1}]} -\partial\psi_{T_\Phi^{i-1}(Z_{i-1}(\tau^-))}(z_i(\tau^+))\big). \end{cases} \tag{11.10}$$

The values of the measures $\mathrm{d}z_i((t_k,t_{k+1}])$ and $\mu_{i,k+1} \overset{\Delta}{=} \mathrm{d}\nu_i((t_k,t_{k+1}])$ are kept as primary variables and this fact is crucial for the consistency of the method for the nonsmooth evolutions, as we already saw in Chaps. 1 and 10. The integral term is approximated thanks to

$$\int_{(t_k,t_{k+1}]} z_{i+1}(\tau)\mathrm{d}\tau \approx hz_{i+1}(t_{k+1}^+) = hz_{i+1,k+1} \tag{11.11}$$

and then we obtain

$$z_{i,k+1} - z_{i,k} - hz_{i+1,k+1} = \mu_{i,k+1}. \tag{11.12}$$

For the approximation of the inclusion, the union of convex cones is approximated in the following way:

$$\overline{\mathrm{conv}}\big(\cup_{\tau\in(t_k,t_{k+1}]} -\partial\psi_{T_\Phi^{i-1}(Z_{i-1}(\tau^-))}(z_i(\tau^+))\big) \approx -\partial\psi_{T_\Phi^{i-1}(Z_{i-1}(t_k^-))}(z_i(t_{k+1}^+)). \tag{11.13}$$

Assuming, as in (11.11), that the approximation of z_i is constant on each interval $(t_k, t_{k+1}]$, we get

$$\mu_{i,k+1} \in -\partial\psi_{T_\Phi^{i-1}(Z_{i-1,k})}(z_{i,k+1}). \tag{11.14}$$

Finally, the time integration of a generic equation of the MDI in (5.22) is given by

$$z_{i,k+1} - z_{i,k} - hz_{i+1,k+1} = \mu_{i,k+1} \in -\partial\psi_{T_\Phi^{i-1}(Z_{i-1,k})}(z_{i,k+1}) \;\; (1 \leqslant i \leqslant r-1). \tag{11.15}$$

The last equation (5.23) is discretized in the same way as

$$\begin{cases} z_{r,k+1} - z_{r,k} - hCA^rW^{-1}z_{k+1} = CA^{r-1}B\,\mu_{r,k+1}, \\ \mu_{r,k+1} \in -\partial\psi_{T_\Phi^{r-1}(Z_{r-1,k})}(z_{r,k+1}). \end{cases} \tag{11.16}$$

For the zero dynamics defined in (5.24), we use for the sake of simplicity[1] an Euler backward scheme,

$$\xi_{k+1} - \xi_k - hA_\xi \xi_{k+1} - hB_\xi z_{1,k+1} = 0. \tag{11.17}$$

The following notation is used for the discretized variables. Let us denote the discretized state vector by

$$z_{k+1} = [z_{1,k+1}, \dots, z_{r,k+1}, \xi_{k+1}^{\mathrm{T}}]^{\mathrm{T}} = [\bar{z}_{k+1}^{\mathrm{T}}, \xi_{k+1}^{\mathrm{T}}]^{\mathrm{T}},$$

the vector of discretized multipliers by μ_{k+1}, i.e.,

$$\mu_{k+1} = [\mu_{1,k+1}, \dots, \mu_{r,k+1}]^{\mathrm{T}}.$$

Then the discrete-time system in (11.15), (11.16), and (11.17) can be rewritten compactly as (see (5.27) and (5.28))

$$z_{k+1} - z_k = hWAW^{-1} z_{k+1} + \bar{G}\mu_{k+1} \tag{11.18}$$

which is the discrete-time counterpart to (5.30).

Definition 11.3 (Extended Moreau's time-stepping scheme). *The inclusions in* (11.15)*,* (11.16)*, and* (11.17) *define a numerical time integration of the higher order sweeping process* $SP(z_0;[0,T])$ *that we call the extended Moreau's time-stepping (EMTS) scheme.*

11.2.2 Properties of the Discrete-Time Extended Sweeping Process

Let us make the following:

Assumption 15. *The triple* $(WAW^{-1}, \bar{G}, H)$ *is observable, controllable, and positive real.*

where the matrices are defined in (5.27)–(5.29).

11.2.2.1 Dissipativity

Let us consider (11.15), (11.16), and (11.17), and the matrix

$$J = \begin{pmatrix} G^{-1} & 0_{r\times(n-r)} \\ 0_{(n-r)\times r} & J_\xi \end{pmatrix}$$

with J_ξ symmetric and positive definite $(n-r)\times(n-r)$ real matrix. We have the following Proposition.

Proposition 11.4. *Suppose that Assumption 15 holds. Then:*

$$\frac{1}{2} z_{k+1}^T J z_{k+1} - \frac{1}{2} z_k^T J z_k \leqslant -\frac{1}{2}(z_{k+1}-z_k)^{\mathrm{T}} J (z_{k+1}-z_k) + h z_{k+1}^T JWAW^{-1} z_{k+1} \tag{11.19}$$

for all $0 \leqslant k \leqslant N-1$.

[1] Depending on the regularity of z_1, a higher order scheme may be used for the time-integration of the zero dynamics.

11.2.2.2 Boundedness

Proposition 11.5. *Suppose that Assumption 15 holds. There exists a constant* $\alpha > 0$ *such that for all* $h > 0$ *and all* $0 \leqslant k \leqslant N-1$, $||z_k|| \leqslant \alpha$. *Moreover, for any given* $h^* > 0$, *there exists a constant* $M \equiv M(h^*) > 0$ *such that* $||\bar{G}\mu_k|| \leqslant M, \forall h \in (0, h^*)$.

11.2.2.3 Local Bounded Variation

What follows is strongly inspired from Monteiro Marques' work in Monteiro Marques (1993, lemma 2.5). We first notice that since all the cones $T_{\Phi}^i(\cdot)$ in Sect. 5.4.2 are either $\mathbb{R}$ or $\mathbb{R}^+$, it follows that the closed ball $\bar{B}(a,R) = \{z \in \mathbb{R} \mid ||z-a|| \leqslant R\} \subset T_{\Phi}^i(\cdot)$ for any $a > 0$ and $R < \frac{a}{2}$. We define $z_i^N : [0,T) \to \mathbb{R}; t \mapsto z_i^N(t)$ as the step function given by $z_i^N(t) = z_{i,k}$ for all $t \in [t_k, t_{k+1})$, $0 \leqslant k \leqslant N-1$ and $z_i^N(t_N) = z_{i,N}$, $1 \leqslant i \leqslant r$.

Proposition 11.6. *Suppose that Assumption 15 holds. The total variation of* z_i^N, $1 \leqslant n$, *in* $[0,T]$ *is bounded above according to:*

$$\begin{aligned}
&var(z_i^N, [0,T]) \leqslant \tfrac{1}{2R}|z_{i,0} - a|^2 + \tfrac{\alpha^2}{2R}T^2 + \alpha T(1 + \tfrac{1}{R}|z_{i,0} - a|) \qquad (1 \leqslant i \leqslant r-1)\\
&var(z_r^N, [0,T]) \leqslant \tfrac{1}{2R}|z_{r,0} - a|^2 + \tfrac{\beta^2\alpha^2}{2R}T^2 + \beta\alpha T(1 + \tfrac{1}{R}|z_{1,0} - a|)\\
&var(\xi^N, [0,T]) \leqslant (\gamma + \delta)\alpha T
\end{aligned} \tag{11.20}$$

where $|||CA^rW^{-1}||| \leqslant \beta$, $|||A_\xi||| \leqslant \gamma$ *and* $|||B_\xi||| \leqslant \delta$, *whereas* α *is as in Proposition 11.5. Moreover there exists a constant* $K > 0$ *such that for all* $N \in \mathbb{N}, N \geqslant 1$*:*

$$var(z^N, [0,T]) \leqslant K. \tag{11.21}$$

Consider the step function $\mu^N : [0,T] \to \mathbb{R}^r; t \mapsto \mu^N(t)$ such that $\mu^N(t) = \mu_{k+1}$ for all $t \in [t_k, t_{k+1})$ $(0 \leqslant k \leqslant N-1)$ and $\mu^N(t_N) = \mu_N$.

Proposition 11.7. *Suppose that Assumption 15 holds. For any given* $h^* > 0$, *there exists a constant* $K' \equiv K'(h^*) > 0$ *such that*

$$var(\mu^N, [0,T]) \leqslant K', \forall h \in (0, h^*). \tag{11.22}$$

11.2.2.4 Convergence

We now denote $\{z^N\}$ the sequence of functions constructed from the functions $z^N(\cdot)$, and similarly for μ^N.

Proposition 11.8. *Suppose that Assumption 15 holds. There exists a subsequence* $\{z^{N_k}\}$ *of* $\{z^N\}$ *which converges pointwise to some function* $z : [0,T] \to \mathbb{R}^n$, *such that* $var(z, [0,T]) \leqslant K$, *and a subsequence* $\{\mu^{N_k}\}$ *of* $\{\mu^N\}$ *which converges pointwise to some function* $\mu(\cdot) : [0,T] \to \mathbb{R}^r$ *such that* $var(\mu, [0,T]) \leqslant K'$.

Remark 11.9. The convergence of μ^N towards a LBV *function* reflects the fact that the primary variables are $\mu_{i,k+1} \overset{\Delta}{=} \mathrm{d}\nu_i((t_k,t_{k+1}])$. Hence the Dirac measures do not appear in the limit $\mu(\cdot)$ which is by construction a (bounded) function.

Proposition 11.10. *Suppose that Assumption 15 holds. If $z(\cdot)$ is right-continuous, then for every continuous function of bounded variation $\varphi : [0,T] \to \mathrm{I\!R}$ we have*

$$\int_{(s,t]} \varphi \, dz_i^{N_k} \to \int_{(s,t]} \varphi \, dz_i \;\; (s<t) \;\; as \; N_k \to +\infty \; (1 \leqslant i \leqslant r). \tag{11.23}$$

Remark 11.11. It has not yet been proved that the limits are solutions of the continuous-time HOSP. However, the examples treated in the next section suggest that this is indeed the case.

11.2.3 Numerical Examples

Let us consider again Examples 11.1 and 11.2, to which we apply the EMTS of Definition 11.3.

Example 11.12. (Example 11.1 continued) If we apply the EMTS scheme, we obtain the following solution:

$$x_k = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \forall k \geqslant 1, \tag{11.24}$$

$$\mu_{1,1} = 1, \quad \mu_{2,1} = 1, \tag{11.25}$$

$$\mu_{1,k} = 0, \, \mu_{2,k} = 0, \, \forall k \geqslant 2, \tag{11.26}$$

which converges to the time-continuous solution of the higher order Moreau's sweeping process, i.e., $x(0) = x_0, x(t) = (0,0,0)^{\mathrm{T}}, \forall t > 0$.

Example 11.13. (Example 11.2 continued) The solution given by the EMTS scheme is

$$x_k = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \forall k \geqslant 1, \tag{11.27}$$

$$\mu_{1,1} = 1, \quad \mu_{2,1} = 1, \quad \mu_{3,1} = 0, \tag{11.28}$$

$$\mu_{i,k} = 0, \, \forall k \geqslant 2, i = 1, \ldots, 3, \tag{11.29}$$

which is the time-continuous solution of the higher order Moreau's sweeping process, i.e., $x(0) = x_0, x(t) = (0,0,0)^{\mathrm{T}}, \forall t > 0$.

Obviously these examples are not a general proof that the EMTS solutions converge to limits which are solutions of the HOSP. We expect, however, that this is the case. In Acary et al. (in press) it is further shown that the zero dynamics may have a strong influence on the dynamics of the system.

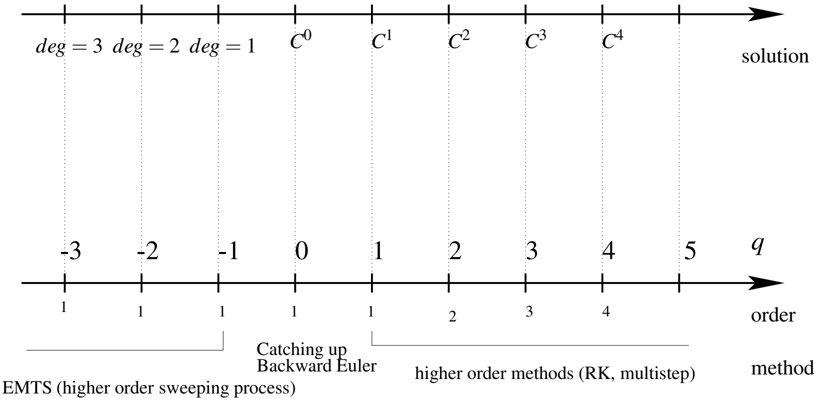


Fig. 11.1. Orders and degrees of discontinuities

11.3 Synoptic Outline of the Algorithms

Figure 11.1 provides a rapid overview of the orders of the various schemes that are presented in Part II, in correspondence with the vector field degree of discontinuity q (Definition 2.56) and the solution regularity. The degree of a function is in (C.10). Also the degree of discontinuity is taken as $-i+1$ if the degree of the distribution in the right-hand side is i (recall that from a general perspective, the systems we are dealing with are to be seen as equalities of distributions, see for instance (5.8), (5.9), and (5.10)). If the right-hand side contains a Dirac measure (degree 2), then $q = -1$ and the solution is a function with a jump, i.e., with a degree 1. If the right-hand side is a discontinuous function (degree 1), then $q = 0$ and the solution is a continuous function C^0. The cases $q \leqslant -1$ concern the higher order sweeping process, whose time-stepping discretization is named the EMTS.

Part III

Numerical Methods for the One-Step Nonsmooth Problems

Introduction

When computing the solution of a smooth ODE with an implicit Euler scheme, one needs to solve at each step a nonlinear equation using some root-finding algorithm (like a Newton's algorithm). Then the scheme can be advanced from step k to step $k+1$. This is the same for the algorithms that have been described in Part II of this book. The schemes can be advanced if at each step some nonsmooth problem is solved (like a linear complementarity problem). We call these problems the *one-step discretized problems*. In this chapter various such one-step problems and their numerical solvers are reviewed. Properly choosing a one-step problem solver is an important feature of the obtained algorithm (event-driven or time-stepping).

12

Basics on Mathematical Programming Theory

12.1 Introduction

As we have seen along the previous chapters, the time-discretization of nonsmooth dynamical systems leads to systems of equalities and inequalities that have to be solved at each time step. Most of these systems are often well-known problems in the mathematical programming theory and a lot of theoretical analysis and solving methods have been proposed over the years.

The goal of this chapter is to give some basic elements on these problems issued by the mathematical programming community without claiming to substitute the reference textbooks. The aim is at helping the reader, who is not familiar with the optimization theory, (a) to identify what the kind of problem yielded from the time-discretization is, (b) to know what the main solving algorithms and their principles are and finally (c) to help the reader to choose a solving method for a specific class of problems. We do not claim that the presentation is rigorous or exhaustive, but we hope that it will help the reader to find a path in the jungle of the optimization methods.

12.2 The Quadratic Program (QP)

12.2.1 Definition and Basic Properties

Let us start with a well-known problem in the mathematical programming theory: the Quadratic Program, which can be defined as follows.

Definition 12.1 (Quadratic Program (QP)). *Let $Q \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Given the matrices $A \in \mathbb{R}^{m_i \times n}$, $C \in \mathbb{R}^{m_e \times n}$, and the vectors $p \in \mathbb{R}^n$, $b \in \mathbb{R}^{m_i}$, $d \in \mathbb{R}^{m_e}$, the Quadratic Program (QP) denoted by $\mathrm{QP}(Q,p,A,b,C,d)$ is to find a vector $z \in \mathbb{R}^n$ such that*

$$\begin{aligned} &\textit{minimize} \;\; q(z) = \frac{1}{2} z^{\mathrm{T}} Q z + p^{\mathrm{T}} z \\ &\textit{subject to}\; Az - b \geqslant 0 \\ &\qquad\qquad\;\; Cz - d = 0 \end{aligned} \tag{12.1}$$

The set $\mathscr{D} = \{z \in \mathbb{R}^n \mid \quad Az - b \geqslant 0, Cz - d = 0\}$ *is called the feasible set of the Quadratic Program (QP).*

Associated Lagrangian Function

The following Lagrangian function is usually associated with this constrained optimization problem:

$$\mathscr{L}(z, \lambda, \mu) = \frac{1}{2} z^{\mathrm{T}} Q z + p^{\mathrm{T}} z - \lambda^{\mathrm{T}}(Az - b) - \mu^{\mathrm{T}}(Cz - d)\,, \tag{12.2}$$

where $(\lambda, \mu) \in \mathbb{R}^{m_i} \times \mathbb{R}^{m_e}$ are the Lagrange multipliers.

First-Order Optimality Conditions

The following theorem defines the so-called first-order optimality conditions which are necessary conditions for a point to be an optimal point of the Quadratic Program (QP) (12.1).

Theorem 12.2 (First-order necessary optimality conditions or Karush–Kuhn–Tucker (KKT) conditions). *The first-order optimality conditions or Karush–Kuhn–Tucker (KKT) conditions of the QP can be stated as follows: suppose that $\bar{z}$ is a local optimum of the QP* (12.1). *Then there exist two vectors of Lagrange multipliers $(\lambda, \mu) \in \mathbb{R}^{m_i} \times \mathbb{R}^{m_e}$ such that*

$$\begin{cases} \nabla_z \mathscr{L}(\bar{z}, \lambda, \mu) = Q\bar{z} + p - A^{\mathrm{T}}\lambda - C^{\mathrm{T}}\mu = 0 \\ \\ C\bar{z} - d = 0 \\ \\ 0 \leqslant \lambda \perp A\bar{z} - b \geqslant 0\,. \end{cases} \tag{12.3}$$

Any solution of the first-order optimality conditions is called a stationary point of the QP.

The Karush–Kuhn–Tucker (KKT) conditions are necessary conditions, i.e. any solution $\bar{z}$ of the KKT condition is a solution of the QP in (12.1). We will see later that this set of equations gives rise to variants of LCP, especially the MLCP and the horizontal LCP according to the type of constraints. If there is no inequality constraint, the KKT system is a linear system.

Remark 12.3. For the more general NonLinear Programming (see Sect. 12.3), a Constraint Qualification (CQ) condition is added to guarantee that KKT conditions are necessary conditions to the optimization problem. The existence and uniqueness of the Lagrange multipliers is not ensured for a particular solution $\bar{z}$ for instance, if the constraints are not linearly independent.

Second-Order Optimality Conditions

The second-order optimality conditions can be separated into necessary and sufficient conditions for optimality. We refer the reader to Bonnans et al. (2003) for a complete description in the framework of the nonlinear programming.

The Dual Problem and Lagrangian Relaxation

The problem (12.1) is usually referred to as the primal problem. A dual problem can be introduced considering the Lagrangian function (12.2). Due to the particular form of the Lagrangian function, the QP problem is equivalent to solving

$$\min_z \max_{\lambda \geqslant 0, \mu} \mathscr{L}(z, \lambda, \mu) \, . \tag{12.4}$$

The idea of Lagrangian relaxation is to invert the min and max, introducing the dual function

$$\theta(\lambda, \mu) = \min_z \mathscr{L}(z, \lambda, \mu) \tag{12.5}$$

and the dual problem

$$\max_{\lambda \geqslant 0, \mu} \theta(\lambda, \mu) \, . \tag{12.6}$$

In the particular case of a QP where the matrix Q is Positive Definite (PD), the dual function is equal to

$$\begin{aligned} \theta(\lambda, \mu) &= \min_z \mathscr{L}(z, \lambda, \mu) = \mathscr{L}(Q^{-1}(A^{\mathrm{T}}\lambda + C^{\mathrm{T}}\mu - p), \lambda, \mu) \\ &= -\tfrac{1}{2}(A^{\mathrm{T}}\lambda + C^{\mathrm{T}}\mu - p)^{\mathrm{T}} Q^{-1} (A^{\mathrm{T}}\lambda + C^{\mathrm{T}}\mu - p) \\ &\quad + b^{\mathrm{T}}\lambda + d^{\mathrm{T}}\mu \end{aligned} \tag{12.7}$$

and we obtain the following dual problem:

$$\max_{\lambda \geqslant 0, \mu} -\frac{1}{2}(A^{\mathrm{T}}\lambda + C^{\mathrm{T}}\mu - p)^{\mathrm{T}} Q^{-1} (A^{\mathrm{T}}\lambda + C^{\mathrm{T}}\mu - p) + b^{\mathrm{T}}\lambda + d^{\mathrm{T}}\mu \, , \tag{12.8}$$

which is a QP with only inequality constraints of positivity. The strong duality theorem asserts that if the matrices Q and $AQ^{-1}A^{\mathrm{T}}$ are symmetric semi-definite positive, if the primal problem (12.1) has an optimal solution, then the dual has also an optimal solution. We will see later the interest of the dual formulation (12.8).

Basic Properties

Quadratic problems can always be solved or shown to be infeasible in a finite number of iterations. However, the problem can be more or less hard to solve depending on the properties of the matrix Q and on the number of inequality constraints. For instance, the QP with an indefinite matrix is a NP-hard problem

(Murty & Kabadi, 1987)[1] and the QP is said to be nonconvex. The nonconvex QP can have several stationary points and local minima.

If Q is PSD, the QP is said to be convex. In this case, polynomial time and robust algorithms can be found. If the matrix Q is a Positive Definite (PD) matrix, we say that the QP is strictly convex. With the positive definiteness assumption, the existence and uniqueness of an optimal point $\bar{z}$ are ensured (unless the feasible domain $\mathscr{D}$ is empty). The existence is due to the fact that the function is coercive and the feasible set $\mathscr{D}$ is closed and convex (possibly unbounded).

If the matrix is indefinite, the existence of a solution is not ensured. As an example consider the QP:

$$\begin{aligned} &\min_{(x,y)} \quad x+y^2 \\ &\text{subject to } x \leqslant 0\,. \end{aligned} \tag{12.9}$$

The "minimum" occurs at $x=-\infty$ for all y and is $-\infty$. We will see further that more precise conditions can be given on the existence and uniqueness of solutions.

Before discussing the numerical methods for convex programs, the following generalized QP can be introduced:

$$\begin{aligned} &\text{minimize } q(z) = \frac{1}{2} z^{\mathrm{T}} Q z + p^{\mathrm{T}} z \\ &\text{subject to } z \in \mathscr{D} \end{aligned} \tag{12.10}$$

for discussing the existence and uniqueness of solutions. If the matrix Q is a PD matrix, the strict convexity ensures the existence of an optimum for all $\mathscr{D} \neq \emptyset$ and the uniqueness of the optimum if $\mathscr{D}$ is convex. The question of the existence of the multipliers and the dual problem is more subtle. If $\mathscr{D}$ is defined by a finite number of inequalities, say

$$\mathscr{D} = \{z \mid g_i(z) \leqslant 0, i=1,\dots,m\}\,, \tag{12.11}$$

the existence of the multipliers is subjected to some qualification constraints. One of the consequences of convexity is that the KKT multipliers coincide with the solutions of the dual problem (12.6).

Degenerate QP

A QP is said to be degenerate when at least one of the following situations is met:

(a) the columns of the matrices C and A corresponding to the active constraints at the solution $\bar{z}$ are not linearly independent,
(b) the strict complementarity condition in (12.3) fails to hold at the solution $\bar{z}$, that is, the optimal Lagrange multiplier λ has a vanishing component for an inequality constraint which is active, $\lambda_i = (A\bar{z}-b)_i = 0$.

The degeneracy of the QP can cause many numerical troubles. We will try to express them while exposing the numerical methods for solving QP. The nonconvex QP with an indefinite matrix Q is also sometimes called a degenerate QP. In the sequel we prefer to call it a nonconvex QP.

[1] The proof is based on a copositive matrix Q which is not PSD.

12.2.2 Equality-Constrained QP

The motivation for this section is that a large class of numerical methods are based on the iterative solutions of equality-constrained QPs.

12.2.2.1 Existence and Uniqueness

In this section, we will give some insights on QPs with only equality constraints:

$$\begin{aligned} &\text{minimize } q(z) = \frac{1}{2} z^{\mathrm{T}} Q z + p^{\mathrm{T}} z \\ &\text{subject to } Cz - d = 0\,. \end{aligned} \tag{12.12}$$

The KKT conditions (12.3) can be written in the form of a linear system:

$$\begin{bmatrix} Q & -C^{\mathrm{T}} \\ C & 0 \end{bmatrix} \begin{bmatrix} \bar{z} \\ \mu \end{bmatrix} = \begin{bmatrix} -p \\ d \end{bmatrix}. \tag{12.13}$$

From now on, the matrix C is assumed to be full row rank, i.e., the constraints are linearly independent.

For numerical purposes, the KKT system (12.13) is usually rewritten as

$$\begin{bmatrix} Q & C^{\mathrm{T}} \\ C & 0 \end{bmatrix} \begin{bmatrix} -r \\ \mu \end{bmatrix} = \begin{bmatrix} g \\ h \end{bmatrix}, \tag{12.14}$$

where $\bar{z} = z_0 + r$, z_0 is any estimate of the solution, r is the desired step to the solution, and

$$h = Cz_0 - d, \quad g = p + Qz_0, \quad r = \bar{z} - z_0\,. \tag{12.15}$$

In order to give some results on this linear system, a matrix $Z \in \mathbb{R}^{n\times(m_e)}$ is introduced whose columns form a basis of the null space of the matrix C^{T}, i.e., $\mathrm{Ker}\,C^{\mathrm{T}}$. The matrix Z has full rank, i.e., it has rank m_e and

$$C^{\mathrm{T}} Z = 0\,. \tag{12.16}$$

The following result is well known in the mathematical programming theory.

Lemma 12.4. *Let C^T have full row rank, and assume that the reduced Hessian matrix, $Z^T QZ$, is PD. Then the KKT matrix*

$$K = \begin{bmatrix} Q & C^T \\ C & 0 \end{bmatrix} \tag{12.17}$$

is nonsingular, and there exists a unique vector pair $(\bar{z}, \mu)$ satisfying (12.13).

We refer to Fletcher (1987) for a proof of this lemma. In fact, using second-order optimality conditions yields the following result:

Lemma 12.5. *Suppose that the assumptions of Lemma 12.4 are satisfied. Then (12.12) has a unique (global) solution, which is the unique solution of (12.13).*

12.2.2.2 Numerical Methods for Solving the KKT System of Equality-Constrained QPs

Although the KKT system (12.14) is a linear system, the fact that the KKT matrix is always indefinite for $m_e \geqslant 1$ leads to some numerical difficulties to solve it. Three major classes of numerical methods can be listed for solving the KKT system of equality-constrained QP: (i) direct methods, on the whole KKT system, (ii) the range-space methods, and (iii) the null-space methods.

Direct Methods of the Whole KKT Matrix

The first way to solve the whole KKT system (12.14) is to perform a factorization of the matrix K. The fact that the KKT matrix is indefinite prevents the use of the standard Cholesky method for the triangular factorization of symmetric matrices. Although the Gaussian elimination with partial pivoting can be used to perform a LU factorization, the most effective solution is to use a dedicated symmetric indefinite factorization for taking into account symmetry (Bunch & Parlett, 1971; Bunch & Kaufman, 1977; Bunch et al., 1976; Higham, 1997). For a general symmetric matrix K, the form of the factorization is as follows:

$$P^{\mathrm{T}} KP = LDL^{\mathrm{T}} , \tag{12.18}$$

where P is permutation matrix, D is block-diagonal matrix containing only 1×1 or 2×2 blocks. The permutation matrix is only introduced to maintain numerical stability and sparsity. This approach can be quite effective on some problems if the heuristics in choosing the permutation matrix for the numerical stability do not destroy the sparsity. Another way to solve the whole KKT system is to apply an iterative method. The indefiniteness precludes the use of the conjugate gradient method, but the QMR methods (Freund & Nachtigal, 1991) and least-squares approaches such as LSQR method (Paige & Saunders, 1982) can be efficient.

Range-Space Method

The range-space method is based on the assumption that the QP matrix Q is positive definite. In this case, an elimination of the block equation on $-r$ in (12.14) can be performed by multiplying by CQ^{-1} and subtracting the second equation. One obtains a symmetric PD linear system on the Lagrange multiplier μ:

$$(CQ^{-1}C^{\mathrm{T}})\mu = CQ^{-1}g - h\,. \tag{12.19}$$

Once we have solved the system (12.19) for μ by standard methods (Cholesky, etc.), the vector r and the optimal point z are retrieved by solving the first block equation:

$$Qr = C^{\mathrm{T}}\mu - g\,. \tag{12.20}$$

Note that this approach just amounts to solving the dual problem (12.7–12.8), which is written here as

$$\max_{\mu} -\frac{1}{2}(C^{\mathrm{T}}\mu - p)^{\mathrm{T}}Q^{-1}(C^{\mathrm{T}}\mu - p) + d^{\mathrm{T}}\mu \, . \tag{12.21}$$

The range-space method is useful when the cost of obtaining the inverse of Q or the factorization of $CQ^{-1}C^{\mathrm{T}}$ is reasonable. This is the case when Q is diagonal or block diagonal or when the number of equality constraints is small with respect to the number of unknowns.

Remark 12.6. A case of a number of equality constraints small with respect to the number of unknowns is for instance the large finite element applications with perfect unilateral contacts when the number of nodes involved in the contact area is small with respect to the global number of nodes in the mesh.

Null-Space Method

On the contrary, the null-space method does not require the positive definiteness of Q. Hereafter we only assume that the conditions for existence and uniqueness of Lemma 12.4 are satisfied (C^{T} is full row rank and $Z^{\mathrm{T}}QZ$ is PD). The null space method requires the computation of the matrix Z whose columns span the null-space of C. This matrix can be computed with orthogonal factorizations or, in the case of sparse problems, by factorization of a submatrix of C. Given a feasible vector z_0 (which is just a particular solution of the system $Cz = d$), any feasible vector can be expressed as

$$z = z_0 + Zw, \quad w \in \mathbb{R}^{m_e} \, . \tag{12.22}$$

The equality-constrained QP (12.12) is then equivalent to the following unconstrained QP:

$$\text{minimize } \frac{1}{2}w^{\mathrm{T}}Z^{\mathrm{T}}QZw + (Qz_0 + p)^{\mathrm{T}}Zw \, . \tag{12.23}$$

If the reduced Hessian $Z^{\mathrm{T}}QZ$ is PD, then the unique solution $\bar{w}$ is given by the solution of the following linear system:

$$Z^{\mathrm{T}}QZw = -Z^{\mathrm{T}}(Qz_0 + p) \, . \tag{12.24}$$

The optimal solution of the original problem (12.12) is then retrieved by using (12.22). The Lagrange multipliers are computed using the first-order optimality conditions,

$$Q\bar{z} + p + C^{\mathrm{T}}\mu = 0 \, , \tag{12.25}$$

which are uniquely solvable if C has full rank by

$$\mu = -(CC^{\mathrm{T}})^{-1}C(Q\bar{z} + p) \, . \tag{12.26}$$

Details on the numerical computation of the null-space matrix Z and the computation of the particular solution together with an efficient evaluation of (12.22) can be found in Fletcher & Johnson (1995). In contrast to the range-space method, the null-space method is efficient if the number $n - m_e$ of degrees of freedom is small. The main drawback is the computation of the matrix Z which can be a hard task in large-scale applications. The choice of the matrix Z, which is not uniquely defined, can also have an important impact on the conditioning of the system. For small and medium applications, an orthonormal Z is chosen in order to keep the conditioning of Q.

The symmetric PD linear system (12.24) can be solved by any standard solvers. Factorization methods can be used also by iterative methods such as the conjugate gradient method. As always, the main difficulty is to find good pre-conditioners.

Comments

It is difficult to give some a priori rules on the use and the effectiveness of the direct, range-space, and null-space methods. These rules depend strongly on the structure of the QP. However, the range-space methods are recommended if the matrix Q is PD and if $CQ^{-1}C^{\mathrm{T}}$ can be computed cheaply. All of these methods imply to solve linear systems. According to their size and their properties, direct or iterative solvers have to be chosen. The efficiency of the QP solvers depends strongly on the efficiency of the underlying linear solver.

The case of the degenerate QP can cause numerical troubles. Indeed, the linear dependence of the columns of the matrix C can cause difficulty in the computation of the null-space matrix Z in null-space methods, and in the range-space methods the matrix $CQ^{-1}C^{\mathrm{T}}$ can become singular. Some regularization can be nevertheless introduced to overcome this problem.

12.2.3 Inequality-Constrained QP

In order to simplify the notation for the presentation of the active-set methods for QP, we introduce the following notation for the inequality-constrained QP (12.1):

$$\begin{aligned} &\text{minimize } q(z) = \frac{1}{2} z^{\mathrm{T}} Q z + p^{\mathrm{T}} z \\ &\text{subject to } h_i^{\mathrm{T}} z - g_i \geqslant 0, \quad i \in \mathscr{I} \\ &\qquad\qquad\quad h_i^{\mathrm{T}} z - g_i = 0, \quad i \in \mathscr{E}\,, \end{aligned} \tag{12.27}$$

where $\mathscr{E}$ and $\mathscr{I}$ are finite sets of indices. The relation with (12.1) is obvious in the sense that the matrix $H = [h_i]^{\mathrm{T}}$ and the vector g are given by

$$H = \begin{bmatrix} A \\ \\ C \end{bmatrix}, \quad g = \begin{bmatrix} b \\ \\ d \end{bmatrix}. \tag{12.28}$$

This formulation allows one to define easily what the active set, $\mathscr{A}(\bar{z})$, is at an optimal point $\bar{z}$ in the following way:

$$\mathscr{A}(\bar{z}) = \{ i \in \mathscr{E} \cup \mathscr{I} \mid h_i^{\mathrm{T}} z - g_i = 0 \}\,. \tag{12.29}$$

The active set contains the index sets of equality constraints $\mathscr{E}$ and the subset of the inequality constraints $\mathscr{I}$, such that equality is satisfied.

Using the active-set notation, the first-order optimality conditions can be simplified to

$$\begin{cases} \nabla_z \mathscr{L}(\bar{z},\lambda,\mu) = Q\bar{z} + p - H^{\mathrm{T}}\lambda = 0 \\ h_i^{\mathrm{T}}\bar{z} - g_i = 0, \quad \forall i \in \mathscr{A}(\bar{z}) \\ h_i^{\mathrm{T}}\bar{z} - g_i \geqslant 0, \quad \forall i \in \mathscr{I} \setminus \mathscr{A}(\bar{z}) \\ \lambda \geqslant 0, \quad \forall i \in \mathscr{I} \cap \mathscr{A}(\bar{z}) \end{cases} . \tag{12.30}$$

12.2.3.1 Active-Set Methods

An active-set method starts by using a guess of the active set of constraints $\mathscr{A}(\bar{z})$ and solves the corresponding equality-constrained QP

$$\begin{aligned} &\text{minimize } q(z) = \frac{1}{2}z^{\mathrm{T}}Qz + p^{\mathrm{T}}z \\ &\text{subject to } h_i^{\mathrm{T}}z - g_i = 0, \quad i \in \mathscr{A}(\bar{z}) \end{aligned} \tag{12.31}$$

by one of the methods previously exposed in Sect. 12.2.2. The active set is then updated using the information of the nonactive constraints and the Lagrange multipliers. One index is added or dropped up until convergence is obtained.

In practice, the notion of working set $\mathscr{W}_k$ is introduced for each iteration. It is a subset of the active set $\mathscr{A}$ and it consists of all the equality constraints, $i \in \mathscr{E}$, plus some active inequality constraints. Not necessarily all active inequality constraints are included in the working set. Especially, an important requirement is to impose that the active constraints are linearly independent.

Given z_k and $\mathscr{W}_k$ at the iteration k, we compute the step $r_k = z - z_k$ from the following QP:

$$\begin{aligned} &\text{minimize } \frac{1}{2}r^{\mathrm{T}}Qr + s_k^{\mathrm{T}}r \\ &\text{subject to } h_i^{\mathrm{T}}r = 0, \quad i \in \mathscr{W}_k \end{aligned} \tag{12.32}$$

with $s_k = Qz_k + p$. The solution of (12.32) can be computed by any method presented in Sect. 12.2.2.

If $r_k \neq 0$, we have to choose a step length α_k as large as possible to maintain the feasibility with respect to all the constraints. It is noteworthy that all the equalities in the working set $\mathscr{W}_k$ continue to hold in the direction r_k. Indeed, we have the following property:

$$h_i^{\mathrm{T}}(z_k + \alpha_k r_k) = h_i^{\mathrm{T}}z_k = g_i \tag{12.33}$$

so the constraint value $h_i^{\mathrm{T}}z$ is constant along the direction r_k. An explicit formula can be derived for α_k (see Fletcher, 1987):

$$\alpha_k = \min\left\{1, \min_{i \in \mathscr{W}_k, h_i^{\mathrm{T}} r_k < 0} \left\{ \frac{g_i - h_i^{\mathrm{T}} z_k}{h_i^{\mathrm{T}} r_k} \right\}\right\} . \tag{12.34}$$

The new iterate z_{k+1} is set to $z_k + \alpha_k r_k$. If $\alpha_k < 1$ the constraints for which the minimum in (12.34) is achieved are called the blocking constraints. The new working set $\mathscr{W}_{k+1}$ is updated by adding one of the blocking constraints to $\mathscr{W}_k$. We continue in this way until an optimal point $\hat{z} = z_k$ is obtained over the working set $\hat{\mathscr{W}} = \mathscr{W}_k$. In this situation, which corresponds to $\alpha_{k-1} = 1$, the step $\hat{r} = r_k$ is going to be equal to zero.

If the step $r_k = 0$, the Lagrange multiplier $\hat{\lambda} = \lambda_k$ from (12.31) satisfies

$$\sum_{i \in \hat{\mathscr{W}}} h_i \hat{\lambda}_i = g = Q\hat{z} + p . \tag{12.35}$$

If all of the multipliers $\hat{\lambda}_i$ for $i \in \hat{\mathscr{W}} \cap \mathscr{I}$ are nonnegative, we have found a stationary point which respects the first-order optimality conditions. Otherwise the objective function q may be decreased by dropping the inequality constraint corresponding to some negative $\hat{\lambda}_i$.

The algorithm of the active-set method for convex QP is described in Algorithm 10.

Algorithm 10 Sketch of the active-set method for convex QP

```
Require: $Q, p, H, g$
Ensure: $\bar{z}, \lambda$

  Compute a feasible initial point $z_0$.
  Compute the working set $\mathscr{W}_0$ at $z_0$.
  IsTheSolutionNotFound ← true
  while IsTheSolutionNotFound do
    Solve the equality constrained QP (12.32) for $r_k$ and $\hat{\lambda} = \lambda_k$ satisfying (12.35).
    if $r_k = 0$ then
      if $\hat{\lambda}_i \geqslant 0, \forall i \in \mathscr{W}_k \cap \mathscr{I}$ then
        $\bar{z} \leftarrow z_k$
        IsTheSolutionNotFound ← false
      else
        $j \leftarrow \mathrm{argmin}_{j \in W_k \cap \mathscr{I}} \{\hat{\lambda}_j\}$
        $z_{k+1} \leftarrow z_k$
        $\mathscr{W}_{k+1} \leftarrow \mathscr{W}_k \setminus \{j\}$
      end if
    else
      Compute $\alpha_k$ according to (12.34).
      $z_{k+1} \leftarrow z_k + \alpha_k r_k$.
      Update $\mathscr{W}_{k+1}$ by adding one of the blocking constraints if any.
    end if
  end while
```

Comments and References

The above presentation of the active-set method is not complete and is only conceptual. For more details on how to compute the feasible initial point z_0, how to construct a working set with only linear independent constraints, and how to efficiently update the factorization of the KKT system for each resolution of the subproblem, we refer to the following textbooks: Gill et al. (1981), Fletcher (1987), and Nocedal & Wright (1999).

If the QP is strictly convex, the convergence, i.e., the finite termination of the algorithm, can be shown under the assumption that the step length α_k does not vanish whenever the step r_k is nonzero. This assumption prevents the phenomenon of cycling. Cycling occurs when a constraint is added and dropped iteratively in the working set without any change in the iteration z_k.

The Nonconvex Case

Most of the serious implementations of QP solvers have suitable heuristics to adapt the search direction and the step length in the case that r_k does not point toward a minimizer of (12.32). In this case, in the context of null-space method, the inertia controlling method is used in which the reduced Hessian $Z^{\mathrm{T}}QZ$ is not permitted to have more than one negative eigenvalue. Pseudo-constraints are maintained in the working set to keep the positive definiteness of the reduced Hessian $Z^{\mathrm{T}}QZ$, and the negative curvature is detected and used as search directions by means of Cholesky. More details can be found in Fletcher (1971) and Gill et al. (1991).

12.2.3.2 Gradient Method with Projections

The major drawback of the active-set methods is the slow evolution of the identification of the constraints that are active at the solution. Indeed, at most one constraint can be added or dropped in the working set at each iteration. For large-scale problems, the lower bound on the number of iterations can be a severe drawback. Among the methods that try to identify quickly the active set of constraints, we can cite the gradient method with projection. If the standard gradient method has the drawback of a slow convergence rate, it has the advantage to allow large changes in the active set and then helps to quickly identify a suitable active set.

The gradient projection method is interesting if the computational cost of the projection $P_{\mathscr{D}}(\cdot)$ onto the feasible set $\mathscr{D}$ is cheap. This is especially the case for a bound constrained QP such as

$$\begin{aligned} &\text{minimize } q(z) = \frac{1}{2} z^{\mathrm{T}} Q z + p^{\mathrm{T}} z \\ &\text{subject to } l \leqslant z \leqslant u \,, \end{aligned} \tag{12.36}$$

where $l \in \mathbb{R}^m$ and $u \in \mathbb{R}^m$ stand for the lower and upper bounds.

The GPCG method proposed in Moré & Toraldo (1991) is based on a two-step procedure at each iteration: the gradient projection stage and the conjugate gradient projection stage. The following sections will give the basic principles of these stages.

Gradient Projection Stage

Given z_k at iteration k, a first step consists in generating a sequence of iterates $y_0 = z_k, y_1, \ldots$, by a gradient projection algorithm such that

$$y_{j+1} = P_{\mathscr{D}}(y_j - \alpha_j \nabla q(y_j)) \, , \tag{12.37}$$

where $\alpha_j > 0$ is chosen by a projected search so that the objective function is decreased, i.e., $q(y_j + 1) < q(y_j)$. We will discuss later how a projected search can be implemented. The gradient projection stage is used to select a new face corresponding to a new active set. The gradient projection stage is stopped if it fails to make reasonable progress or when a suitable active set is found, that is when the index j_k is the first to satisfy the two following tests for j:

$$\begin{aligned} &\mathscr{A}(y_j) = \mathscr{A}(y_{j-1}) \\ &q(y_{j-1}) - q(y_j) \leqslant \nu_1 \max\{q(y_{l-1}) - q(y_l), 1 \leqslant l < j\}, \quad \nu_1 > 0 \, . \end{aligned} \tag{12.38}$$

At the end of this step, we set $z^\star = y_{j_k}$.

Conjugate Gradient Stage

The second step is to compute an approximation of the solution of the QP (12.36) on the face defined by the active given by the first step. This means to solve the following equality-constrained QP:

$$\begin{aligned} &\text{minimize } q(z^\star + d) = \frac{1}{2} d^{\mathrm{T}} Q d + p^{\mathrm{T}} (z^{\star\mathrm{T}} Q + d) \\ &\text{subject to } d_i = 0, \quad \forall i \in \mathscr{A}(z^\star) \, . \end{aligned} \tag{12.39}$$

The equality-constrained QP (12.39) can be easily solved recognizing that the null-space matrix Z is easily computable. Null-space methods lead to an unconstrained QP in the variable w:

$$\text{minimize } \tilde{q}(w) = \frac{1}{2} w^{\mathrm{T}} A w + r^{\mathrm{T}} w \, , \tag{12.40}$$

where $A = Z^{\mathrm{T}} Q Z$ is the reduced Hessian on the “free” variables with respect to $\mathscr{A}(z^\star)$. The QP (12.40) can be solved by conjugate gradient solvers which generate a sequence of w_i. The solver is stopped when the following test is satisfied for j_k:

$$\tilde{q}(w_{j-1}) - \tilde{q}(w_j) \leqslant \nu_2 \max\{\tilde{q}(w_{l-1}) - \tilde{q}(w_l), 1 \leqslant l < j\}, \quad \nu_2 > 0 \, . \tag{12.41}$$

We set $d_k = Z w_{j_k}$ as the new search direction for a new projected search defining the next iterate as

$$z_{k+1} = P_{\mathscr{D}}(z^\star + \alpha_k d_k) \, . \tag{12.42}$$

In case the solution is in the face defined by $\mathscr{A}(z^\star)$, the conjugate gradient method must be continued until it reaches an accurate solution. For detecting such a situation, a necessary condition is verified based on the notion of the so-called binding set:

$$\mathscr{B}(z) = \{i \mid z_i = l_i \text{ and } (\nabla q(x))_i \geqslant 0, \quad \text{or} \quad z_i = u_i \text{ and } (\nabla q(x))_i \leqslant 0\} \ . \tag{12.43}$$

If $\mathscr{A}(z_{k+1}) = \mathscr{B}(z_{k+1})$, the conjugate gradient method is continued.

Projection Searches

As we said before, the evaluation of α_j in (12.37) and α_k in (12.42) is based on a projected search. It consists in finding a value for $\alpha > 0$ such that the function

$$\phi_k(\alpha) = q(P_{\mathscr{D}}(z_k + \alpha d_k)) \tag{12.44}$$

is sufficiently decreased. The sufficient decrease condition requires that $\alpha > 0$ satisfies

$$\phi_k(\alpha) \leqslant \phi_k(0) + \mu \nabla^{\mathrm{T}} q(z_k)(P_{\mathscr{D}}(z_k + \alpha d_k) - z_k), \quad \mu \in \left(0, \frac{1}{2}\right) \ . \tag{12.45}$$

This is done by testing the values of α_l generated by the decreasing sequence

$$\alpha_0 > 0, \quad \alpha_{l+1} \in [\gamma_1 \alpha_l, \gamma_2 \alpha_l], \quad 0 < \gamma_1 < \gamma_2 < 1 \ . \tag{12.46}$$

Note that $P_{\mathscr{D}}(z_k + \alpha d_k)$ is a piecewise linear function. Its breakpoints are given by changes in the active set. In the case of a simple feasible set given by the bound constraints, these breakpoints can be computed explicitly. The function $\phi_k(\alpha)$ is a piecewise quadratic function.

Comments and Variants

In Conn et al. (1988), Wright (1989) and Nocedal & Wright (1999), some variants of the previous algorithms can be found. Especially, the first step can be defined by only one iterate of the gradient projection method but with a exact computation of a local minimizer of the piecewise quadratic function. This can be done by testing each linear interval of the piecewise linear path $P_{\mathscr{D}}(z_k + \alpha d_k)$. In Friedlander & Leyffer (2006), the idea of the gradient projection method is also used to identify quickly the active set, but together with an augmented Lagrangian approach (see Sect. 12.3) and filter methods. In Bertsekas (1982), a second-order acceleration mechanism is inserted in the spirit of the active-set approach.

12.2.3.3 Interior Point Methods

Recent trends seem to favor the so-called interior point methods for solving large QPs. As the name indicates, the key idea is to approach the solution of the problem by a sequence of iterates which stay, as long as possible, far away from constraints, i.e., in the strict interior of the feasible domain. The primal methods also known as barrier methods use the logarithmic function as barrier functions to enforce the iterates to stay far away from the constraints. Such methods have been extensively studied as the book of Fiacco & McCormick (1968) witnesses. In the context of nonlinear programming, these methods have been forgotten due to the poor numerical efficiency and the lack of robust methods to drive the penalty parameter (see Sect. 12.3).

The interior point methods have known a new interest in the optimization community for their ability to provide tools for the complexity theory (see the book of Nesterov & Nemirovskii, 1993, for a general theory). Indeed, from the seminal work of Karmarkar (1984), the efficiency of interior point methods for linear programming has been extensively studied. Such studies have been improved and extended many times to other problems such as QP and LCP (see the survey papers of Freund & Mizuno 1996; Potra & Wright, 2000).

From the numerical point of view, the numerical efficiency is more debatable. Clearly, the best algorithms from the complexity point of view are not necessarily the best from the practical efficiency point of view. However, some interior point methods seem to enjoy a good practical efficiency, even if their complexity properties are worse. Mainly the primal–dual interior point methods seem to provide the basic framework to efficient algorithms. They are based on solving the first-order optimality conditions for both primal and dual variables. As we said earlier in the beginning of Sect. 12.2, the first optimality conditions for a QP are a MLCP. Therefore, without going into further details in this section, we will present some aspects of the interior point methods for LCPs in Sect. 12.4.8.1.

For more details on interior point methods, we refer to the monographs: den Hertog (1994), Wright (1996b), Ye (1997), and Bonnans et al. (2003)

12.2.4 Comments on Numerical Methods for QP

12.2.4.1 How to Choose the Right Method?

It is quite difficult to give hard and simple rules for choosing a method with respect to the others. Nevertheless, we will give some advices which are relatively widely admitted in the mathematical programming community:

1. Active-set methods are the best suited
   - for small to medium system sizes ($n < 5000$),
   - when a good initial point is known especially for the active-set identification point of view, for instance, in sequential quadratic programming or at each step of a dynamical process,
   - when an exact solution is searched. Active-set methods can be used as “purification” techniques of interior point methods.

   Recall that several methods are available to solve the equality-constrained subproblem depending on the structure of the original QP.
2. Gradient projection methods are well suited for large QP with simple constraints (simple inequality, bound constrained, etc.).

3. Interior point methods are well suited
    - for large systems without the knowledge of a good starting point,
    - when the problem has a special structure that can be exploited directly in solving the Newton iteration.

12.2.4.2 Special Interest of the QP Method for the One-Step Discretized Problem

Anticipating on the discussion in Sect. 12.4, it can be convenient to recast some complementarity problems into the QP formulation if it is possible. One of the reasons is that there is now a huge collection of very robust and efficient QP solvers (freely available or commercially distributed) which are able to deal with difficult problems such as

- convex QPs with PSD matrices,
- redundant and linearly dependent constraints,
- nonconvex QPs.

For all of these methods, the fact that the problem is an optimization problem and not only a direct system to solve such as the KKT conditions helps us to stabilize the problem and to ensure by globalization the convergence of the algorithm to a local optimum. This fact is crucial from the practical point of view. For a presentation of such methods which are more or less variants of the methods presented in this section, we refer to the work of Gould & Toint (2002).

We will see in Chap. 13 that such types of problems can arise in solving some frictional contact problems, where some constraints are often redundant leading to hyperstaticity and where the friction and its approximation induce nonconvexity.

12.3 Constrained Nonlinear Programming (NLP)

12.3.1 Definition and Basic Properties

The NonLinear Programming (NLP) problem is somehow a generalization of the QP problem. It consists in finding the minimum of a nonlinear function under nonlinear equality and inequality constraints. This very general problem in optimization can be defined as

Definition 12.7 (Nonlinear programming (NLP) problem). *Given a differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$ and two differentiable mappings $g\colon \mathbb{R}^n \to \mathbb{R}^{m_i}$, $h\colon \mathbb{R}^n \to \mathbb{R}^{m_e}$, the nonlinear programming problem is to find a vector $z \in \mathbb{R}^n$ such that*

$$\begin{aligned} &\textit{minimize} \;\; f(z) \\ &\textit{subject to}\; g(z) \geqslant 0 \\ &\qquad\qquad\;\; h(z) = 0\,. \end{aligned} \tag{12.47}$$

As usual, the Lagrangian of this NLP problem is associated as follows:

$$\mathscr{L}(z,\lambda,\mu) = f(z) - \lambda^{\mathrm{T}} g(z) - \mu^{\mathrm{T}} h(z) , \tag{12.48}$$

where $(\lambda,\mu) \in \mathrm{I\!R}^{m_i} \times \mathrm{I\!R}^{m_e}$ are the Lagrange multipliers.

First-Order Optimality Conditions

In contrast with the QP case, existence of Lagrange multipliers (μ,λ) is not straightforward. Some Constraint Qualification (CQ) conditions must hold at least at the optimal point. Numerous types of (CQ) exist. Let us give the definition of the most popular (CQ), the so-called Linar Independence Constraint Qualification (LICQ). For that we introduce for convenience purposes for the equivalent NLP

$$\begin{array}{l} \text{minimize } f(z) \\ \\ \text{subject to } c_i(z) = 0, i \in \mathscr{E} \\ \qquad\qquad\quad c_i(z) \geqslant 0, i \in \mathscr{I} \end{array} , \tag{12.49}$$

where $\mathscr{E}$ and $\mathscr{I}$ are finite sets of indices. The relation between (12.47) and (12.49) is obvious. The active-set $\mathscr{A}(\bar{z})$ at a point $\bar{z}$ is defined by

$$\mathscr{A}(\bar{z}) = \{ i \in \mathscr{E} \cup \mathscr{I} \mid c_i(\bar{z}) = 0 \} . \tag{12.50}$$

Definition 12.8 (Linar Independence Constraint Qualification (LICQ)). *Given the point $\bar{z}$ and the active set $(\bar{z})$ we say that the Linar Independence Constraint Qualification (LICQ) holds if the set of active constraint gradients $\{\nabla c_i(\bar{z}) \mid i \in \mathscr{A}(\bar{z})\}$ is linearly independent.*

For more general (CQ) conditions, we refer to Mangasarian (1969), Fletcher, (1987), Hiriart-Urruty & Lemaréchal (1993), and Bonnans et al. (2003). With the previous definition of the Linar Independence Constraint Qualification (LICQ), the first-order optimality conditions can be stated as follows:

Theorem 12.9 (First-order necessary optimality conditions or KKT conditions). *Suppose that $\bar{z}$ is a local optimum of the NLP problem* (12.47) *and the LICQ holds at $\bar{z}$, then there exists two vectors of Lagrange multiplier $(\lambda,\mu) \in \mathrm{I\!R}^{m_i} \times \mathrm{I\!R}^{m_e}$ such that*

$$\begin{cases} \nabla_z \mathscr{L}(\bar{z},\lambda,\mu) = \nabla f(\bar{z}) - \nabla g^{\mathrm{T}}(\bar{z})\lambda - \nabla h^{\mathrm{T}}(\bar{z})\mu = 0 \\ \\ h(\bar{z}) = 0 \\ \\ 0 \leqslant \lambda \perp g(\bar{z}) \geqslant 0 . \end{cases} \tag{12.51}$$

Any solution of the first-order optimality conditions is called a stationary point of the NLP.

Second-Order Optimality Conditions

As for the QP, the second-order optimality conditions can be separated into necessary and sufficient conditions for optimality. We refer the reader to Bonnans et al. (2003). One basic ingredient of the sufficient conditions is the definite positiveness of the projected Hessian of the Lagrangian.

The Dual Problem

As for the QP problem, an analog dual problem to (12.6) can be defined based on the dual function (12.48) by introducing the dual function

$$\theta(\lambda,\mu) = \min_{z} \mathscr{L}(z,\lambda,\mu) \tag{12.52}$$

and the dual problem

$$\max_{\lambda \geqslant 0,\mu} \theta(\lambda,\mu)\,. \tag{12.53}$$

Basic Properties

Generally, the existence of a minimizer is not ensured. Several assumptions can lead to the existence of (possibly several) minimizers. Boundedness from below or coercivity of f can be invoked for the existence of minimizers. Uniqueness of a global minimizer is ensured in the convex case, that is when the function f is strictly convex and the set feasible set $\mathscr{D}$ is also convex. For more details, we refer to the famous textbook: Hiriart-Urruty & Lemaréchal (1993).

12.3.2 Main Methods to Solve NLPs

We will not enter into the details of the implementation of the numerical methods for solving NLPs. One of the reason is that the subject is very wide and the algorithms are generally quite complex. As we did at the beginning of this chapter, we will just list the main methods and their advantages, giving some pointers to useful references.

12.3.2.1 Penalty, Barrier, and Augmented Lagrangian Approaches

Exterior Penalty Approach

The most natural idea to solve the NLP (12.47) is to transform the original problem into an unconstrained NLP with the help of a penalty function. The most well-known penalty function is quadratic penalty function. In this case, we end up with the following unconstrained NLP:

$$\text{minimize } f(z) + \frac{1}{2\varepsilon}\|h(z)\|^2 + \frac{1}{2\varepsilon}\|\max(0,-g(x))\|^2, \tag{12.54}$$

where ε is the penalty parameter. Most of the algorithms based on the penalty approach consist in solving a sequence of problems (12.54) for a sequence of penalty parameters $\{\varepsilon_k\}$ which tends to 0. At any stage of the algorithm, the constraints are not fulfilled exactly. They are exactly satisfied only in the limit $\varepsilon \to 0$. By vanishing the penalty parameter, we penalize the constraints more severely, and the optimal point is found in the exterior of the feasible set. We call sometimes this method the *exterior penalty method*. From the computational point of view, these methods behave poorly. As the penalty parameter tends to 0, the problem becomes stiff and ill-conditioned. Nevertheless, its very easy implementation can attract users for small simple problems.

Barrier Methods

In contrast to the exterior penalty approach, the barrier method penalizes the sequence of iterates in the interior of the feasible set. For an inequality-constrained NLP, the log barrier function is the most well-known function and yields the following unconstrained NLP,

$$\text{minimize } f(z) - \varepsilon \textstyle\sum_{i=1}^{m_i} \log \; g_i(x) \; . \tag{12.55}$$

As with the penalty approach, the algorithm attempts to generate a sequence of solutions of the problem (12.55) for a sequence of parameters $\{\varepsilon_k\}$ which tends to 0. The monitoring of the parameter ε and the difficulty to solve the problem for small parameters may have dramatic consequences on the practical behavior of the method. This method which was extensively studied in Fiacco & McCormick (1968) has been abandoned in this primal form but gives rise to the primal–dual interior point methods for NLP. As for the interior point methods for LCP which can be applied to the optimality conditions of a QP, the analog can be implemented for the couple NonLinear Complementarity Problem (NCP)/NLP. More details can be found in the references cited in Sect. 12.2.3.3.

Augmented Lagrangian Approach

We complete this section with the augmented Lagrangian approach, which is sometimes called the *exact penalty approach*. The key idea is to introduce new terms in the Lagrangian function ("to augment the Lagrangian") to penalize the constraints in an exact way, that is, when the constraints are satisfied these added terms vanish in the objective function. One example of augmented Lagrangian function can be given for an inequality-constrained NLP,

$$\mathscr{L}_\sigma(z,\lambda,\mu) = \mathscr{L}(z,\lambda,\mu) + \lambda^{\mathrm{T}} \max\left(\frac{-\lambda}{\sigma}, g(x)\right) + \frac{\sigma}{2}\left\|\max\left(\frac{-\lambda}{\sigma}, g(x)\right)\right\|^2 , \tag{12.56}$$

where the max function has to be taken component-wise. Many other augmented Lagrangian functions have been defined with various theoretical properties and computational efficiencies. For a review and a discussion, we refer to the papers of Rockafellar (1973, 1974, 1976a, 1979, 1993) and the book Bonnans et al. (2003). The principle of the algorithms based on the augmented Lagrangian is to generate a sequence of iterates, z^k, which minimizes the augmented Lagrangian for a sequence of Lagrange multipliers (λ^k, μ^k). The efficiency of the augmented Lagrangian approach has been shown on very large class of problems. For a practical implementation, we refer to the code LANCELOT (Conn et al., 1992).

12.3.2.2 Successive Quadratic Program (SQP)

For the sake of simplicity, we expose only the principle of the Successive Quadratic Program (SQP) method for the following equality-constrained NLP,

$$\begin{array}{l} \text{minimize } f(z) \\ \text{subject to } h(z) = 0\,. \end{array} \tag{12.57}$$

We assume that the LICQ for all z holds, that is the matrix $\nabla h(z)$ has full row rank. From the KKT conditions, for any solution, $\bar{z}$ of (12.57) there exists a unique Lagrange multiplier $\mu \in \mathbb{R}^{m_e}$ such that

$$\begin{cases} \nabla f(\bar{z}) + \nabla h(\bar{z})^{\mathrm{T}} \mu = 0 \\ h(\bar{z}) = 0\,. \end{cases} \tag{12.58}$$

The Successive Quadratic Program (SQP) method for solving (12.57) is a Newton-like method on the system (12.58). Starting at the current point (z_k, μ_k), the step $(\Delta z_k, \Delta \mu_k)$ such that

$$\begin{bmatrix} z_{k+1} \\ \mu_{k+1} \end{bmatrix} = \begin{bmatrix} z_k \\ \mu_k \end{bmatrix} + \begin{bmatrix} \Delta z_k \\ \Delta \mu_k \end{bmatrix} \tag{12.59}$$

solves the following linear system:

$$\begin{bmatrix} W_k & -A_k^{\mathrm{T}} \\ A_k^{\mathrm{T}} & 0 \end{bmatrix} \begin{bmatrix} \Delta z_k \\ \Delta \mu_k \end{bmatrix} = \begin{bmatrix} -\nabla f(z_k) + A_k^{\mathrm{T}} \mu_k \\ -h(z_k) \end{bmatrix}, \tag{12.60}$$

where $A_k^{\mathrm{T}} = \nabla^{\mathrm{T}} h(z_k)$ and the matrix W_k is either the Hessian of $\mathscr{L}$ at z_k, i.e., $W_k = \nabla^2_{zz} \mathscr{L}(z, \mu)$ in Newton's method or an approximation of the Hessian in the quasi-Newton method such as BFGS, which is updated at each iteration.

An alternative view of the system (12.60) is to define the following QP:

$$\begin{array}{l} \text{minimize } q(z) = \dfrac{1}{2} \Delta z W_k \Delta z + \nabla f(z_k)^{\mathrm{T}} \Delta z \\[2ex] \text{subject to } A_k \Delta z - h(z_k) = 0\,, \end{array} \tag{12.61}$$

which is well defined if W_k is PD on the null space of the constraints and the matrix A_k has full row rank. In this case, the unique solution of the QP is the solution of the linear system (12.60). The two alternative ways to view the SQP method are very interesting. From the theoretical point of view, the analysis is based on Newton's approach. From the practical point of view, the approach uses the QP solvers tools exposed in Sect. 12.2.3. The SQP method is easily extended to the general NLP problem with inequality-constrained NLP by introducing the following QP:

$$\begin{aligned} &\text{minimize } q(z) = \frac{1}{2}\Delta z W_k \Delta z + \nabla f(z_k)^{\mathrm{T}} \Delta z \\ &\text{subject to } A_k \Delta z - h(z_k) = 0, \\ &\qquad\qquad\; B_k \Delta z - g(z_k) \geqslant 0\,, \end{aligned} \tag{12.62}$$

where $B_k = \nabla g(z_k)$.

The practical implementation of SQP relies on the choice and the monitoring of the underlying QP solvers. Two choices are available: (a) to solve only equality constrained QPs with an active set updated in the outer algorithm or (b) to solve the complete QP (12.62) directly with an update of the active inside the QP solver. The second choice is preferable if a good QP solver is at hand. The SQP are always globalized by line-search or trust-region methods as for the standard Newton's and quasi-Newton's methods for the unconstrained NLP. We will not enter in more details of the SQP which would necessitate a whole chapter. We refer to Nocedal & Wright (1999, Chap. 18) and Bonnans et al. (2003, Part III).

12.3.2.3 Gradient Projection Methods

As explained in Sect. 12.2.3.2, the gradient projection method is interesting for large-scale problems if the computational cost of the projection $P_{\mathscr{D}}$ onto the feasible set $\mathscr{D}$ is cheap. This is especially the case for bound constraints, sphere constraints, Cartesian products of simple constraints.

The Goldstein–Levitin–Polyak gradient projection method (Goldstein, 1964; Levitin & Polyak, 1966) consists of the iteration

$$y_{j+1} = P_{\mathscr{D}}(y_j - \alpha_j \nabla f(y_j))\,, \tag{12.63}$$

where $\alpha_j \geqslant 0$ denotes the step size. Levitin & Polyak (1966) proved the convergence of the method under the assumption that f is Lipschitz with constant L and the feasible set is convex. The proof is given for step sizes that satisfy

$$0 < \varepsilon \leqslant \alpha_j \leqslant \frac{2(1-\varepsilon)}{L}, \quad \text{for all } j\,, \tag{12.64}$$

where ε is any scalar with $0 < \varepsilon \leqslant 2/(2+L)$.

In Bertsekas (1976), a generalized Armijo line-search procedure is given. This rule extends the previous convergence results and provides us with an efficient computational procedure for choosing the step. Furthermore, the approach allows one to combine the gradient direction with some modified Newton directions improving by the way the rate of convergence. As we said earlier in Sect. 12.2.3.2, Bertsekas (1982) completes his analysis of his second-order acceleration mechanism.

Remark 12.10. The gradient-projection methods must not be confused with the projected gradient method and the Rosen gradient projection method (Rosen, 1960, 1961). It is well known that the projected gradient method which consists of the iterate

$$y_{j+1} = y_j + \alpha_j P_{\mathscr{D}}(-\nabla f(y_j)) \tag{12.65}$$

does not converge. Rosen's gradient projection method (Rosen, 1960, 1961) is based on projecting the search direction into the subspace tangent to the active constraints. Unfortunately, the computation of the Rosen projected gradient is based on the update of the active set of constraints that does not allow large changes at each step. Acceleration procedures using active-set strategies have also been proposed for the Rosen method to obtain super-linear convergence (Gill & Murray, 1975). However, active-set strategies preclude the application of such methods onto large-scale problems.

12.4 The Linear Complementarity Problem (LCP)

12.4.1 Definition of the Standard Form

The LCP is a widespread problem in mathematical programming theory. A usual definition of this problem can be formulated as follows:

Definition 12.11 (Linear complementarity problem, LCP). *Given $M \in \mathbb{R}^{n\times n}$ and $q \in \mathbb{R}^n$, the linear complementarity problem is to find a vector $z \in \mathbb{R}^n$, denoted by* $\mathrm{LCP}(M,q)$ *such that*

$$\begin{cases} w = Mz + q \\ \\ 0 \leqslant z \perp w \geqslant 0. \end{cases} \tag{12.66}$$

The solution set of $\mathrm{LCP}(M,q)$ *is denoted by* $\mathrm{SOL}(M,q)$.

The inequalities have to be understood component-wise and the relation $x \perp y$ means $x^{\mathrm{T}}y = 0$. A vector z such that the inequalities $z \geqslant 0$ and $Mz+q \geqslant 0$ are satisfied is said to be feasible. A LCP is said to be feasible if a feasible vector exists. The following standard index sets are defined, for any vector z:

$$\begin{aligned} \alpha(z) &= \{i \mid z_i > 0 = w_i = (Mz+q)_i\} \\ \beta(z) &= \{i \mid z_i = 0 = w_i = (Mz+q)_i\} \\ \gamma(z) &= \{i \mid z_i = 0 < w_i = (Mz+q)_i\}. \end{aligned} \tag{12.67}$$

A solution $\bar{z}$ of $\mathrm{LCP}(M,q)$ is said to be degenerate if the index set $\beta(\bar{z})$ is a nonempty set.

12.4.2 Some Mathematical Properties

For an exhaustive presentation of the LCP and its mathematical properties, we refer to Cottle et al. (1992) and Murty (1988). We recall in this section only the main mathematical properties. The proofs can be found in the standard monographs cited above.

The P-Matrix Property

Mathematical results concerning the LCP are associated with a large number of matrix classes, see Cottle et al. (1992, Chap. 3) for an almost exhaustive presentation). Among these classes, one is fundamental since it yields existence and uniqueness of the solution: the class of P-matrices defined below.

Definition 12.12 (P-matrix). *A matrix, $M \in \mathbb{R}^{n\times n}$, is said to be a P-matrix if all its principal minors are positive.*

Recall that for a matrix $A \in \mathbb{R}^{n\times n}$ and an index set $\alpha \subset \{1,\dots,n\}$, the submatrix $A_{\alpha\alpha}$, which is the matrix whose entries lie in the rows and columns of A indexed by α, is called a principal submatrix of A. The determinant $\det(A_{\alpha\alpha})$ is called a principal minor of A.

The following theorem gives a first characterization of a P-matrix:

Theorem 12.13. *Let $M \in \mathbb{R}^{n\times n}$. The following statements are equivalent:*

(a) M is a P-matrix
(b) M reverses the sign of no nonzero vector,[2] *i.e.,*

$$x \circ Mx \leqslant 0 \quad \Longrightarrow \quad x = 0. \tag{12.68}$$

This property can be written equivalently,

$$\forall x \neq 0, \exists i \text{ such that } x_i(Mx)_i > 0. \tag{12.69}$$

(c) All real eigenvalues of M and its principal submatrices are positive.

The Existence and Uniqueness Theorem

The existence and uniqueness of solutions to $\mathrm{LCP}(M,q)$ can be characterized by the following theorem:

Theorem 12.14. *A matrix $M \in \mathbb{R}^{n\times n}$ is a P-matrix if and only if* $\mathrm{LCP}(M,q)$ *has a unique solution for all vectors $q \in \mathbb{R}^n$.*

[2] A matrix $A \in \mathbb{R}^{n\times n}$ reverses the sign of a vector $x \in \mathbb{R}^n$ if $x_i(Ax)_i \leqslant 0, \forall i \in \{1,\dots,n\}$. The Hadamard product $x \circ y$ is the vector with coordinates $x_i y_i$.

The Case of the PD Matrix

It is noteworthy that if M is a symmetric P-matrix, then M is PD due to the statement (c) of Theorem 12.13 or directly Definition 12.12. It is clear also that (not necessarily symmetric) PD matrices belong to the class of the P-matrices. Therefore, Theorem 12.14 holds directly for a PD matrix.

Remark 12.15. In most monographs, the notion of PD matrix usually implies the notion of symmetry of the matrix. This is mainly due to the fact that positiveness is related to the positiveness of the bilinear mapping $x^{\mathrm{T}}Mx$ which is always equal to $\frac{1}{2}x^{\mathrm{T}}(M^{\mathrm{T}}+M)x$. This also due to the extension for Hermitian matrices in $\mathbb{C}^{n\times n}$. Indeed, $\frac{1}{2}x^*(M^*+M)x$ is in $\mathbb{R}$ but not x^*Mx. In contrast to the QP context, the symmetry of a PD matrix is not assumed in the LCP framework.

In practice, this “P-matrix” assumption is difficult to check via numerical computation. Especially it is not possible in polynomial time. But a PD matrix (not necessarily symmetric), which is a P-matrix, is often encountered in applications.

The P_0-Matrix Property

Let us start with the definition of a P_0-matrix.

Definition 12.16 (P_0-matrix). *A matrix $M \in \mathbb{R}^{n\times n}$ is said to be a P_0-matrix if all its principal minors are nonnegative.*

The following theorem gives a first characterization of a P_0-matrix.

Theorem 12.17. *Let $M \in \mathbb{R}^{n\times n}$. The following statements are equivalent:*

(a) M is a P_0-matrix.
(b) For any $x \neq 0$, there exists i such that $x_i \neq 0$ and $x_i(Mx)_i \geqslant 0$.
(c) All the real eigenvalues of M and of its principal submatrices are nonnegative.
(d) For each $\varepsilon > 0$, $M+\varepsilon I$ is a P-matrix.

It is noteworthy that if M is a symmetric P_0-matrix, then M is PSD. Conversely, the class of PSD matrices (not necessarily symmetric) belongs to the class of P_0-matrices.

As for the linear system with a PSD matrix, the existence and uniqueness of solutions of the LCP are not guaranteed with a P_0-matrix. The following theorem gives some results on the uniqueness of w if the matrix is a so-called column adequate matrix.

Theorem 12.18. *Let $M \in \mathbb{R}^{n\times n}$. The following statements are equivalent:*

(a) For all $q \in K(M) = \{q \mid SOL(M,q) \neq \emptyset\}$, if z and $\bar{z}$ are any two solutions of LCP(M,q) *then*

$$w = Mz+q = M\bar{z}+q = \bar{w}, \tag{12.70}$$

that is w is uniquely defined.

(b) Every vector whose sign is reversed by M belongs to the null space of M,

$$x \circ Mx \leqslant 0 \Longrightarrow Mx = 0. \tag{12.71}$$

(c) M is a P_0-matrix and is column adequate, that is, for each index set $\alpha \subset \{1,\ldots,n\}$, one has

$$\det\ M_{\alpha\alpha} = 0 \Longrightarrow M_{\bullet\alpha} \textit{ has linearly dependent columns,} \tag{12.72}$$

where $M_{\bullet\alpha}$ is the submatrix of M composed of the columns indexed by α.

In order to make precise the notion of existence and uniqueness of solutions, new matrix classes have been introduced. We will not enter the details, refering to the book of Cottle et al. (1992); but we list some of the properties that can be useful when discussing numerical algorithms.

Definition 12.19 (*Q*-and *Q* —

Definition 12.19 (Q-and Q_0-matrices). *The class of matrices M for which* LCP(M,q) *has a solution for all q is denoted by Q and its elements are called Q-matrices. The class of matrices M for which* LCP(M,q) *has a solution whenever it is feasible is denoted by Q_0 and its elements are called Q_0-matrices.*

Definition 12.20 (Sufficient matrix). *A matrix $M \in \mathbb{R}^{n\times n}$ is called a column sufficient matrix if it satisfies*

$$x \circ Mx \leqslant 0 \Longrightarrow x_i(Mx_i) = 0, \forall i \in \{1,\ldots,n\}. \tag{12.73}$$

The matrix M is called row sufficient if M^{T} is column sufficient. If M is both column and row sufficient, M is a sufficient matrix.

For a recent work on necessary and sufficient conditions on the solvability of feasible LCPs, we refer to Kostreva & Yang (2004).

Other Mathematical Properties and Matrix Classes

The $P_\star$ property was introduced by Kojima et al. (1991) in the context of the interior point method for LCP.

Definition 12.21. *A matrix $M \in \mathbb{R}^{n\times n}$ is said to be a $P_\star(\kappa)$-matrix for $\kappa \geqslant 0$ if for all $x \in \mathbb{R}^n$*

$$(1+4\kappa)\sum_{i\in\mathscr{I}_+(x)} x_i(Mx)_i + \sum_{i\in\mathscr{I}_-(x)} x_i(Mx)_i \geqslant 0, \tag{12.74}$$

where the index sets are defined by

$$\mathscr{I}_+(x) = \{i \mid x_i(Mx)_i > 0\}, \quad \mathscr{I}_-(x) = \{i \mid x_i(Mx)_i < 0\}. \tag{12.75}$$

A matrix $M \in \mathbb{R}^{n\times n}$ is said to a $P_\star$-matrix if it is a $P_\star(\kappa)$-matrix for some $\kappa \geqslant 0$, i.e.,

$$P_\star = \cup_{\kappa\geqslant 0} P_\star(\kappa). \tag{12.76}$$

A surprising result of Väliaho (1996) shows that the $P_\star$-matrices are the sufficient matrices. The $P_\star$-property (12.74) can be reformulated as

$$z = Mx \Longrightarrow x^{\mathrm{T}}z \geqslant -4\kappa \sum_{i\in\mathscr{I}_+(x)} x_i z_i. \tag{12.77}$$

Numerical Checking of the Matrix Property

Unfortunately, the P property cannot be checked in polynomial time. This remark is true for most of the matrix classes reviewed here. However, the classes of PD and PSD matrices can be checked in polynomial time. This is also the case for a $P_\star(\kappa)$-matrix for a given fixed κ, which is so far the largest class of matrices relevant for numerics and which can be checked in polynomial time.

12.4.3 Variants of the LCP

We present here some variants of the standard form of the LCP which are convenient for our applications.

12.4.3.1 The Mixed Linear Complementarity Problem (MLCP)

Definition 12.22. *Given the matrices $A \in \mathrm{IR}^{n\times n}$, $B \in \mathrm{IR}^{m\times m}$, $C \in \mathrm{IR}^{n\times m}$, $D \in \mathrm{IR}^{m\times n}$, and the vectors $a \in \mathrm{IR}^n, b \in \mathrm{IR}^m$, the Mixed Linear Complementarity Problem (MLCP) denoted by* $\mathrm{MLCP}(A,B,C,D,a,b)$ *consists in finding two vectors $u \in \mathrm{IR}^n$ and $v \in \mathrm{IR}^m$ such that*

$$\begin{cases} Au + Cv + a = 0 \\ \\ 0 \leqslant v \perp Du + Bv + b \geqslant 0 \end{cases} . \tag{12.78}$$

The Mixed Linear Complementarity Problem (MLCP) can be defined equivalently in the following form denoted by $\mathrm{MLCP}(M,q,\mathscr{E},\mathscr{I})$*:*

$$\begin{cases} w = Mz + q \\ w_i = 0, \forall i \in \mathscr{E} \\ 0 \leqslant z_i \perp w_i \geqslant 0, \forall i \in \mathscr{I}, \end{cases} \tag{12.79}$$

where $\mathscr{E}$ and $\mathscr{I}$ are finite sets of indices such that $\mathrm{card}(\mathscr{E} \cup \mathscr{I}) = n$ *and* $\mathscr{E} \cap \mathscr{I} = \emptyset$.

The Mixed Linear Complementarity Problem (MLCP) is a mixture between an LCP and a system of linear equations.

12.4.3.2 The Horizontal and the Vertical LCP

The horizontal LCP has emerged as an important variant of the standard LCP, especially in the framework of interior point methods.

Definition 12.23. *Given the matrices $Q \in \mathrm{IR}^{n\times n}$ and $R \in \mathrm{IR}^{n\times n}$ and a vector $q \in \mathrm{IR}^n$, the horizontal LCP denoted by* $\mathrm{hLCP}(Q,R,q)$ *consists in finding two vectors $x \in \mathrm{IR}^n$ and $s \in \mathrm{IR}^n$ such that*

$$\begin{cases} Qx + Rs = q \\ \\ 0 \leqslant x \perp s \geqslant 0 \end{cases} . \tag{12.80}$$

The vertical LCP appears naturally in optimization theory and control theory.

Definition 12.24. *Given the matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ and two vectors $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^n$, the vertical Linear Complementarity Problem (LCP) denoted by* vLCP(A,B,a,b) *consists in finding a vector $x \in \mathbb{R}^n$ such that*

$$0 \leqslant Ax + a \perp Bx + b \geqslant 0. \tag{12.81}$$

12.4.3.3 The Geometric LCP

Another way to look at LCPs is to define the so-called geometric LCP.

Definition 12.25. *Let us consider an affine subspace $\mathscr{M} \subset \mathbb{R}^{2n}$. The geometric LCP, denoted by* gLCP$(\bar{s}, \Phi)$*, consists in finding two vectors $w \in \mathbb{R}^n$ and $z \in \mathbb{R}^n$ such that*

$$\begin{cases} s = (w,z) \in \mathscr{M} \\ \\ 0 \leqslant w \perp z \geqslant 0. \end{cases} \tag{12.82}$$

An affine subspace is usually specified by a vector $\bar{s}$ and a vector subspace Φ such that

$$\mathscr{M} = \bar{s} + \Phi = \{s \in \mathbb{R}^{2n} \mid s - \bar{s} \in \Phi\}. \tag{12.83}$$

The geometric LCP has been first studied by Güler (1995), who called it a generalized LCP.

12.4.3.4 The Extended and the Generalized LCP

In Mangasarian & Pang (1995), a notion of extended LCP is introduced to generalize all of the preceding definitions. We will see that some good properties can be factorized on the extended form.

Definition 12.26. *Given the matrices $M \in \mathbb{R}^{m\times n}$ and $N \in \mathbb{R}^{m\times n}$ and a nonempty polyhedral set $\mathscr{K}$, the extended LCP denoted by* xLCP$(M,N,\mathscr{K})$ *consists in finding two vectors $w \in \mathbb{R}^n$ and $z \in \mathbb{R}^n$ such that*

$$\begin{cases} Mz + Nw \in \mathscr{K} \\ 0 \leqslant w \perp z \geqslant 0. \end{cases} \tag{12.84}$$

The above definition of the extended LCP is equivalent to the generalized LCP introduced by Ye (1993) in the context of interior point methods.

Definition 12.27. *Given the matrices $M \in \mathbb{R}^{m\times n}$, $N \in \mathbb{R}^{m\times n}$, $Q \in \mathbb{R}^{m\times k}$ and a vector $q \in \mathbb{R}^m$, the generalized LCP consists in finding three vectors $w \in \mathbb{R}^n$, $z \in \mathbb{R}^n$, and $y \in \mathbb{R}^k$ such that*

$$\begin{cases} Mz + Nw + Qy = q \\ 0 \leqslant w \perp z \geqslant 0 \\ y \geqslant 0 \end{cases} . \tag{12.85}$$

12.4.4 Relation Between the Variants of LCPs

The Generalization Way

The horizontal LCP, the vertical LCP, the Mixed Linear Complementarity Problem (MLCP), and the standard LCP can be written explicitly as an extended LCP. This relation can be written formally as

$$\text{hLCP} \subset \text{xLCP}, \quad \text{vLCP} \subset \text{xLCP}, \quad \text{MLCP} \subset \text{xLCP}, \quad \text{LCP} \subset \text{xLCP}, \qquad (12.86)$$

and the equivalence between extended LCP and generalized LCP can also be written formally as xLCP = gLCP. In the same way, the following relations are easily obtained:

$$\text{LCP} \subset \text{hLCP}, \quad \text{LCP} \subset \text{vLCP}, \quad \text{LCP} \subset \text{MLCP}. \qquad (12.87)$$

The Specialization Way

In order to obtain the equivalence between nontrivial forms of LCPs we need to add some assumptions.

Clearly, if the matrix A is nonsingular in the MLCP (12.78), we may solve the embedded linear system to obtain u and then reduce the MLCP to a LCP with $q = b - DA^{-1}a, M = b - DA^{-1}C$. If $n = m$ and the matrix C is nonsingular, the MLCP reduces to a vertical LCP. If $n = m$ and the matrix D is nonsingular, the MLCP reduces to a horizontal LCP. In the same way, if R (or Q) is nonsingular in the horizontal LCP (12.80), it can be formulated as a standard LCP.

Under weaker assumptions on the monotonicity of matrices, equivalence can also be proved. Let us define a monotonicity property of a couple of matrices.

Definition 12.28. *The couple of matrices (Q,R) satisfies the monotonicity property if*

$$Qz + Rw = 0 \Longrightarrow z^T w \geqslant 0. \qquad (12.88)$$

The horizontal LCP is said to be monotone if the couple of matrices (R,S) satisfies the monotonicity property.

Theorem 12.29. *Any monotone horizontal LCP (12.80) can be reformulated as an LCP in the standard form (12.66),*

$$monotone\ \text{hLCP} \subset \text{LCP}. \qquad (12.89)$$

A proof can be found in Güler (1995) based on maximal monotone operator properties. A simpler proof can be found in Bonnans & Gonzaga (1996) and Bonnans et al. (2003) which is based on a QR decomposition of the matrix R.

Theorem 12.30. *Let us consider a MLCP in the form (12.79). If the matrix M is PSD, then the MLCP can be reformulated as an LCP in the standard form (12.66),*

$$monotone\ \text{MLCP} \subset \text{LCP}. \qquad (12.90)$$

A proof of this theorem can be found in Wright (1996a).

The work of Anitescu et al. (1995) generalizes the equivalence results for the more general class of matrices that satisfy the $P_\star$ property. This property can be extended to matrix pairs for the horizontal LCP.

Definition 12.31. *A matrix pair $(M,N) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$ is said to be a $P_\star(\kappa)$-matrix pair for $\kappa \geqslant 0$ if*

$$Mz + Nx = 0 \Longrightarrow x^{\mathrm{T}} z \geqslant -4\kappa \sum_{i \in \mathscr{I}_+(x)} x_i z_i. \tag{12.91}$$

In the same way, the $P_\star$ property can be extended to the geometric LCP by

$$\dim \Phi = n \text{ and } x^{\mathrm{T}} z \geqslant -4\kappa \sum_{i \in \mathscr{I}_+(x)} x_i z_i, \forall (x,z) \in \Phi. \tag{12.92}$$

Theorem 12.32. *Let (Q,R) be a $P_\star(\kappa)$-matrix pair. The horizontal LCP (12.80),* hLCP(Q,R,q)*, can be reformulated as an LCP in the standard form (12.66). Let M be a $P_\star(\kappa)$-matrix. The MLCP (12.78),* MLCP$(M,q,\mathscr{E},\mathscr{I})$*, can be reformulated as an LCP in the standard form (12.66). Let Φ be a vector subspace satisfying (12.92). The geometric LCP (12.82),* gLCP$(\bar{s},\Phi)$*, can be reformulated as an LCP in the standard form (12.66).*

12.4.4.1 Mathematical Properties of the Variants

Other references on the variants of the LCP and their mathematical properties can be found in Gowda & Sznajder (1994), Sznajder & Gowda (1995), Gowda (1996), Güler (1995), and Anitescu et al. (1995). Especially, the P property, the P_0 property, the Q property, and the $P_\star$ property have been extended with the corresponding theorems.

12.4.4.2 Own Interest of the Variants from a Numerical Point of View

Even in the case where the equivalence has been shown, the own interest of the variants of the standard LCPs lies in the following points:

(a) Most problems are formulated naturally in terms of these variants. A reformulation, when it is possible, possibly destroys the physical meaning of the unknowns, leading to some difficulties in the interpretation of the results.
(b) Most importantly, the transformations from a variant to another can destroy the special structure of the problem (sparsity, conditioning) leading to numerical troubles.

Following these remarks, we see that we have interest to implement and design numerical solvers able to take advantage of the special structure of the variants of the LCP. One way should be to design an LCP solver for the most general class of LCPs. However, the approach could lead to inefficient algorithms due to the imposed generalization. We will see in Sect. 12.4.8.2 that interior point methods seem to have succeeded in combining generality with efficiency in the solvers.

12.4.5 Links Between the LCP and the QP

12.4.5.1 The KKT Conditions of a QP as a Variant of LCPs

The first-order optimality conditions of QPs (12.3) lead to various forms of LCPs depending on the type of constraints. We reformulate in this section the KKT conditions as LCPs and we attempt, when possible, to quickly give some equivalence results.

In fact, solutions of a QP satisfy the KKT conditions which can be formulated as an LCP. Conversely, it cannot be said that the solutions of an LCP are solutions of a QP; this requires additional assumptions on the matrix of the LCP.

Standard Form of the LCP with a Symmetric Matrix

Let us start with the simplest case.

Theorem 12.33. *Let $Q = Q^T$, a PSD matrix. Consider the following QP:*

$$\begin{aligned} &\textit{minimize} \;\; q(z) = \frac{1}{2} z^T Q z + p^T z \\ &\textit{subject to } z \geqslant 0. \end{aligned} \tag{12.93}$$

Its solutions (if any) are the solutions of the KKT system, which is the LCP

$$\begin{cases} \nabla \mathscr{L}(z,w) = Qz + p - w = 0 \\ \\ 0 \leqslant z \perp w \geqslant 0. \end{cases} \tag{12.94}$$

MLCP

The following more complicated example is taken from Wright (1996a). Let us consider the following QP:

$$\begin{aligned} \text{minimize } & q(z) = \frac{1}{2} z^{\mathrm{T}} Q z + p^{\mathrm{T}} z \\ \text{subject to } & Az - b \geqslant 0 \\ & Cz - d = 0 \\ & z_i \geqslant l_i, i \in \mathscr{L} \subset \{1, \ldots, n\} \\ & z_i \leqslant u_i, i \in \mathscr{U} \subset \{1, \ldots, n\} \end{aligned} \quad . \tag{12.95}$$

By defining

$$\begin{aligned} E_{\mathscr{L}} &= [e_i^{\mathrm{T}}]_{i \in \mathscr{L}}, \quad l = [l_i]_{i \in \mathscr{L}}, \\ E_{\mathscr{U}} &= [e_i^{\mathrm{T}}]_{i \in \mathscr{U}}, \quad u = [u_i]_{i \in \mathscr{U}}, \end{aligned} \tag{12.96}$$

where e_i is the ith unit vector from the standard basis, the first-order optimality conditions can be stated in the form of MCLP$(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{a}, \tilde{b})$ with

$$\tilde{A} = \begin{bmatrix} Q & -C^{\mathrm{T}} \\ C & 0 \end{bmatrix}, \quad \tilde{B} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \tilde{C} = \begin{bmatrix} -E_{\mathscr{L}}^{\mathrm{T}} & E_{\mathscr{U}}^{\mathrm{T}} & -A^{\mathrm{T}} \\ 0 & 0 & 0 \end{bmatrix}, \quad \tilde{D} = \begin{bmatrix} E_{\mathscr{L}} & 0 \\ -E_{\mathscr{U}} & 0 \\ A & 0 \end{bmatrix} \tag{12.97}$$

and

$$\tilde{a} = \begin{bmatrix} p \\ -d \end{bmatrix}, \quad \tilde{b} = \begin{bmatrix} -l \\ u \\ -b \end{bmatrix}. \tag{12.98}$$

If the QP is convex, the matrix Q is PSD. The matrix $\tilde{A}$ is PSD as the sum of a PSD matrix containing Q and a skew-symmetric matrix.

Monotone Horizontal LCP

The first-order optimality conditions for the following QP

$$\begin{aligned} &\text{minimize } q(z) = \frac{1}{2} z^{\mathrm{T}} Q z + p^{\mathrm{T}} z \\ &\text{subject to } z \geqslant 0 \\ &\qquad\qquad Cz - d = 0 \end{aligned} \tag{12.99}$$

give rise to a horizontal LCP (12.80) of the form:

$$\begin{cases} Rz + Sw = q \\ \\ 0 \leqslant w \perp z \geqslant 0. \end{cases} \tag{12.100}$$

Indeed, the KKT conditions (12.3) are here

$$\begin{cases} Q\bar{z} + p - \lambda - C^{\mathrm{T}} \mu = 0 \\ \\ C\bar{z} - d = 0 \\ \\ 0 \leqslant \lambda \perp \bar{z} \geqslant 0, \end{cases} \tag{12.101}$$

which can be directly put into an MLCP form. A horizontal LCP can be obtained if the matrix C has full row rank via an elimination of the multiplier μ. Introduce a matrix Z, whose columns span the null space of C, then (12.101) can be rewritten as

$$\begin{cases} Z^{\mathrm{T}}(Q\bar{z} + p - \lambda) = 0 \\ \\ C\bar{z} - d = 0 \\ \\ 0 \leqslant \lambda \perp \bar{z} \geqslant 0, \end{cases} \tag{12.102}$$

which can be put into the form (12.100) with

$$R = \begin{bmatrix} Z^{\mathrm{T}}Q \\ C \end{bmatrix}, \quad S = \begin{bmatrix} -Z^{\mathrm{T}} \\ 0 \end{bmatrix}, \quad q = \begin{bmatrix} -Z^{\mathrm{T}}p \\ d \end{bmatrix}. \tag{12.103}$$

If the QP (12.99) is convex, the KKT conditions are a particular case of a monotone horizontal LCP. Indeed, the relation $Rz + S\lambda = 0$ is here

$$Z^{\mathrm{T}}(Qz - \lambda) = 0, \quad C\bar{z} = 0, \tag{12.104}$$

which implies that

$$z^{\mathrm{T}}(Qz - \lambda) = 0. \tag{12.105}$$

The monotonicity property follows from the fact that Q is PSD matrix.

12.4.5.2 QP Reformulations of LCPs

Standard Form of the LCP with an Asymmetric Matrix

If the LCP matrix M in (12.66) is asymmetric, we can consider the following QP:

$$\begin{array}{l} \text{minimize } q(z) = z^{\mathrm{T}}(Mz + p) = \frac{1}{2}z^{\mathrm{T}}(M + M^{\mathrm{T}})z + z^{\mathrm{T}}p \\ \text{subject to } Mz + p \geqslant 0 \\ \qquad\qquad\quad z \geqslant 0 \end{array} . \tag{12.106}$$

The KKT conditions for this QP can be written as

$$\begin{cases} (M + M^{\mathrm{T}})\bar{z} + p - M^{\mathrm{T}}\lambda_1 - \lambda_2 = 0 \\ 0 \leqslant M\bar{z} + p \perp \lambda_1 \geqslant 0 \\ 0 \leqslant \bar{z} \perp \lambda_2 \geqslant 0 \end{cases} . \tag{12.107}$$

Clearly, if (12.106) has an optimal value z^* such that $q(z^*) = 0$, then z^* solves the LCP(M, p) in (12.66). The following lemma is valid for any matrix $M \in \mathbb{R}^{n \times n}$. The proof can be found in Cottle et al. (1992).

Lemma 12.34. *If the* LCP(M, p) *in (12.66) is feasible, then*

(a) the QP (12.106) has an optimal solution,
(b) there exist $\bar{z}$, λ_1, λ_2 such that the KKT (12.107) conditions are satisfied,
(c) the vectors $\bar{z}$ and λ_1 satisfy

$$(\bar{z} - \lambda_1)_i (M^T(\bar{z} - \lambda_1))_i \leqslant 0, \quad \forall i \in \{1, \ldots, n\}. \tag{12.108}$$

If some additional assumption is made on the matrix M, then $(\bar{z},\lambda_2)$ solves (12.66). This holds for example when M is PSD (but not necessarily symmetric) or a P-matrix or a row-sufficient matrix. The key property is to have $\lambda_1 = \bar{z}$ in (b) of Lemma 12.34 and this property is usually obtained via (c).

Unfortunately, the QP (12.106) is nonconvex in general. Therefore robust algorithms, essentially for convex QP, cannot be applied. Nevertheless, as we noted in Sect. 12.2, some methods are able to behave correctly with nonconvex QPs. Although there is no clear numerical comparisons, as far as we know, dedicated solvers for LCPs might perform better – and be more favored by users.

Example 12.35. Let us consider the LCP(M,q) with the following data:

$$M = \begin{bmatrix} 1 & -3 \\ 0 & 1 \end{bmatrix}, \quad q = \begin{bmatrix} -1 \\ -1 \end{bmatrix}. \tag{12.109}$$

Clearly, the matrix M is a P-matrix, but not a PD matrix. The associated QP (12.106) is given by

$$\begin{array}{l} \text{minimize } q(z) = \frac{1}{2} z^{\mathrm{T}} \begin{bmatrix} 2 & -3 \\ -3 & 2 \end{bmatrix} z + z^{\mathrm{T}} \begin{bmatrix} -1 \\ -1 \end{bmatrix} \\[2ex] \text{subject to } \begin{bmatrix} 1 & -3 \\ 0 & 1 \end{bmatrix} z + \begin{bmatrix} -1 \\ -1 \end{bmatrix} \geqslant 0 \\[2ex] \qquad\qquad z \geqslant 0 \end{array} , \tag{12.110}$$

which is a nonconvex QP.

Standard Form of the LCP with a Symmetric Matrix

Theorem 12.36. *Let $Q = Q^T$ a PD matrix. Consider the following three QPs:*

$$\begin{cases} \textit{minimize } \; q(z) = \frac{1}{2} z^T Q z + p^T z \\[2ex] \textit{subject to } z \geqslant 0 \end{cases} , \tag{12.111}$$

$$\begin{cases} \textit{minimize } \; q(z) = \frac{1}{2} z^T Q z + p^T z \\[2ex] \textit{subject to } Az - b \geqslant 0 \end{cases} , \tag{12.112}$$

and

$$\begin{cases} \textit{minimize } \; q(z) = \frac{1}{2} z^T Q z - b^T w \\[2ex] \textit{subject to } A^T w - Qz = p \\ \qquad\qquad\quad w \geqslant 0 \end{cases} . \tag{12.113}$$

Then the next statements hold:

(a) The KKT conditions formulated by the LCP $0 \leqslant z \perp p + Qz \geqslant 0$ *are necessary and sufficient for the vector* z *to be a globally optimal solution of the QP in (12.111).*
(b) (Dorn's duality) If $\bar{z}$ *solves the QP in (12.112) then there exists* $\bar{w}$ *such that* $(\bar{x}, \bar{w})$ *solves the QP in (12.113). Moreover the two extrema are equal.*
(c) (Dorn's converse duality) If $(\bar{x}, \bar{w})$ *solves the QP in (12.113), then there exists* $\hat{z}$ *with* $Q\hat{z} = Q\bar{z}$ *such that* $\hat{z}$ *solves the QP in (12.112).*

We devote the rest of this section to the numerical solvers.

12.4.6 Splitting-Based Methods

The principle of the splitting-based methods for solving LCPs is to decompose the matrix M as the sum of two matrices B and C,

$$M = B + C, \tag{12.114}$$

which define the splitting. Then LCP(M, q) is solved via a fixed-point iteration.

For an arbitrary vector z^{ν} we consider LCP(B,q^{ν}) with

$$q^{\nu} = q + Cz^{\nu} \tag{12.115}$$

A vector $z = z^{\nu}$ solves LCP(M, q) if and only if z^{ν} is itself a solution of LCP(B,q^{ν}).

All of the variants of splitting methods are based on various choices for the splitting (B, C). The subproblem, LCP$(B, q + Cz^{\nu})$, needs to have at least one solution, i.e., B has to be a Q-matrix. From a practical point of view, the splitting must also lead to a subproblem which is relatively easy to solve. A clear analogy can be drawn with iterative splitting techniques for linear systems. A general scheme for the splitting method is presented in Algorithm 11.

Algorithm 11 Sketch of the general splitting scheme for the LCP

```
Require: $M, q, tol$
Require: $(B, C)$ a splitting of $M$
Ensure: $z, w$ solution of LCP$(M, q)$.
  Compute a feasible initial point $z_0 \geqslant 0$.
  $\nu \leftarrow 0$
  while $error > tol$ do
    Solve the LCP$(B, q + Cz^{\nu})$.
    Set $z^{\nu+1}$ as an arbitrary solution.
    Evaluate $error$.
  end while
```

Projected Jacobi Method

The most trivial choice for the matrix B is to choose the identity matrix or any positive diagonal matrix D. The subproblem LCP$(B, q + Cz^{\nu})$ is then reduced to a component-wise maximum, i.e.,

$$z^{\nu+1} = \max\{0, z^{\nu} - D^{-1}(q + Mz^{\nu})\}. \tag{12.116}$$

In particular, if the matrix D is chosen as the diagonal part of the matrix M, i.e., $D = \mathrm{diag}(m_{ii})$, we obtain the projected Jacobi method. The word *Projected* refers to the projection onto the nonnegative orthant related to the unilateral constraint.

Projected Gauss–Seidel and Projected Successive Overrelaxation (PSOR) Methods

Based on the Gauss–Seidel method for linear systems, the following splitting of M can be used:

$$M = B + C, \text{ with } B = L + \omega^{-1}D, \quad C = U, \tag{12.117}$$

where the matrices L and U are, respectively, the strictly lower part and upper part of the matrix M and $\omega \in (0,2)$ is an arbitrary relaxation parameter. In this case we obtain a projected successive overrelaxation (PSOR) scheme where the iterate z^{k+1} is given by

$$z_i^{k+1} = \max\left(0, z_i^k - \omega M_{ii}^{-1}\left(q_i + \sum_{j<i} M_{ij} z_i^{k+1} + \sum_{j\geqslant i} M_{ij} z_i^k\right)\right), \quad i = 1, \ldots, n. \tag{12.118}$$

When $\omega = 1$ the PSOR method is called the projected Gauss–Seidel (PGS) algorithm.

Other types of splitting can also be used. Especially, we can also take advantage of the possible block structure of the matrix M. If a similar decomposition is made with respect to the block-diagonal matrix and strictly upper and lower block matrices, each smaller subproblem can be solved by any standard method for LCPs, even possibly by an iterative splitting. We will see in Chap. 13 that this approach can improve the rate of convergence of the method.

Convergence Results

Chapter 5 in Cottle et al. (1992) and Chap. 9 in Murty (1988) give many results on the convergence of the splitting method. The basic ingredients are on one hand the symmetry and the positive definiteness of the matrix and on the other hand the contraction of the mapping defined in the fixed-point algorithm. For suitable choices of the splitting, the results provide us with the convergence to the solution of the LCP(M, q) for a symmetric PSD matrix M. Nevertheless, some strong conditions have to be satisfied for the splitting and then for the matrix M. In particular, the cases of an asymmetric PSD matrix and of a P-matrix are not covered by the assumptions.

To the best of our knowledge, few results concern the rate of convergence. Practical experience shows that the convergence is rather slow. In fact, the set of active constraints is quickly identified but the convergence to an accurate solution is thereafter slow. In Murty (1988, p. 380), a standard example of an LCP with a symmetric PD matrix for which the convergence is extremely slow is given. Another example of this type will given in Chap. 13.

The advantage of such methods is their easy implementation, the low cost in terms of memory, and the fact that they preserve special structures of the matrix M, especially sparsity. Another advantage of iterative splitting schemes is their ability to be parallelized.

Regularized PSOR Methods

In some applications, the diagonal matrix of M, $D = \operatorname{diag}(m_{ii})$, is not invertible. This can be the case for some PSD matrix, in which some diagonal entries vanish. The example of the diode bridge rectifier in Chap. 14 exhibits for instance an LCP matrix which is PSD, see (14.5) and (14.6). It is then possible to adapt the previous PSOR scheme by considering the following successive LCPs:

$$\begin{cases} w = Mz + q + \rho(z - \tilde{z}) = (M + \rho I)z + q - \rho\tilde{z} \\ \\ 0 \leqslant w \perp z \geqslant 0 \end{cases} . \tag{12.119}$$

If (12.119) is solved by $\tilde{z} = z$ then $\tilde{z}$ solves LCP(M,q). The key idea is to adapt the PSOR algorithm to include the condition $\tilde{z} = z$ in the fixed-point iterates. The regularized projected successive overrelaxation (RPSOR) scheme is given by

$$z_i^{k+1} = \max(0, z_i^k - \omega(M_{ii} + \rho)^{-1} \left(q_i + \sum_{j<i} M_{ij} z_i^{k+1} + \sum_{j \geqslant i} M_{ij} z_i^k - \rho z_i^k \right) \tag{12.120}$$

for $i = 1, \ldots, n$. When $\omega = 1$ the RPSOR method is called the regularized projected Gauss–Seidel (RPGS) algorithm.

In a more general setting, the regularization process is interesting because of the following reason. Even when they converge, the iterative methods are more stable with PD matrices. One goal of the regularization is to transform a problem with a PSD matrix into a sequence of subproblems with a PD matrix.

Line Search in the Symmetric Case

In order to globalize the convergence of the splitting algorithms a line-search procedure can enhance the basic scheme presented in Algorithm 11. Algorithm 12 describes the main steps of the algorithm. If $z^\star$ is a solution of LCP$(B, q + Cz^\nu)$, the line search determines the step size in the direction d^ν defined by $d^\nu = z^\star - z^\nu$. If ${d^\nu}^{\mathrm{T}} M d^\nu \leqslant 0$, then $\alpha^\nu = 1$; otherwise α^ν must be a nonnegative number satisfying

$$f(z^\nu + \alpha^\nu d^\nu) = \min\{f(z^\nu + \alpha^\nu d^\nu), z^\nu + \alpha^\nu d^\nu \geqslant 0, \alpha \geqslant 0\}, \tag{12.121}$$

Algorithm 12 Sketch of the general splitting scheme with line search for the LCP(M,q) with M symmetric

```
Require: $M, q, tol$ and $(B, C)$ a splitting of a symmetric matrix $M$
Ensure: $z, w$ solution of LCP$(M, q)$.
  Compute a feasible initial point $z_0 \geqslant 0$.
  $\nu \leftarrow 0$
  while $error > tol$ do
    Solve the LCP$(B, q + Cz^{\nu})$.
    Set $z^{\star}$ as an arbitrary solution.
    Set $d^{\nu} \leftarrow z^{\star} - z^{\nu}$ as the search direction.
    Determine the step size $\alpha^{\nu}$ by a line-search procedure.
    Set $z^{\nu+1} \leftarrow z^{\nu} + \alpha^{\nu} d^{\nu}$.
    Evaluate $error$.
  end while
```

where $f(\cdot)$ is the objective function defined by

$$f(z) = z^{\mathrm{T}}(Mz + q). \tag{12.122}$$

In the symmetric case, a clear analogy can be drawn with the following QP:

$$\begin{aligned} &\text{minimize } q(z) = \frac{1}{2} z^{\mathrm{T}} M z + q^{\mathrm{T}} z \\ &\text{subject to } z \geqslant 0 \end{aligned} \tag{12.123}$$

and d^{ν} is a descent direction of the objective function. Indeed, we have

$$f(z^{\nu}) - f(z^{\nu+1}) = -\alpha^{\nu} d^{\nu\mathrm{T}}(q + Mz^{\nu}) - \frac{\alpha^{\nu 2}}{2} d^{\nu\mathrm{T}} M d^{\nu}. \tag{12.124}$$

If $d^{\nu\mathrm{T}} M d^{\nu} \leqslant 0$, the choice of $\alpha^{\nu} = 1$ ensures a decrease of the function $f(\cdot)$,

$$f(z^{\nu}) - f(z^{\nu+1}) \geqslant -d^{\nu\mathrm{T}}(q + Mz^{\nu}), \tag{12.125}$$

and a new iterate $z^{\nu+1}$ is nonnegative by construction. If $d^{\nu\mathrm{T}} M d^{\nu} > 0$, the function $f(z^{\nu} + \alpha d^{\nu})$ defines a strictly convex QP in α whose optimal solution is attained for

$$\bar{\alpha} = -\frac{\alpha^{\nu} d^{\nu\mathrm{T}}(q + Mz^{\nu})}{d^{\nu\mathrm{T}} M d^{\nu}} > 0. \tag{12.126}$$

If $z^{\nu} + \bar{\alpha} d^{\nu} \geqslant 0$ then we can choose $\alpha^{\nu} = \bar{\alpha}$, otherwise we can find an α_{nu} $[1, \bar{\alpha}]$ such that the new iterate $z^{\nu+1}$ is nonnegative.

Convergence for this algorithm is given for a symmetric matrix M and a PD matrix B.

Link with the Gradient Projection Method of QP

The gradient projection method (see Sect. 12.2.3.2) for the QP (12.123) is defined by

$$z_{k+1} = P_{\mathscr{D}}(z_k - \alpha_k(Mz^k + q)). \tag{12.127}$$

In other terms,

$$z_{k+1} = \max(0, (I - \alpha_k M)z^k - q). \tag{12.128}$$

In this case, the gradient projection method with line search is the simplest PSOR method with a splitting based on the matrix $B = I$. We can then consider that all the presented methods for symmetric LCP are better than the pure gradient projection methods. On the other hand, the specific gradient projection for QP presented in Sect. 12.2.3.2 may be a good alternative to obviate the slow convergence of QP solvers. This approach has been developed in Kocvara & Zowe (1994) for LCPs with a symmetric PD matrix. Other well-known techniques to improve greatly the rate of convergence of gradient projection methods are the multi-grid techniques. We refer to Hackbusch & Mittelmann (1983), Mandel (1984), Brandt & Cryer (1983), and Oosterlee (2003) for the study and the development of such methods.

Line-Search in the Asymmetric Case

The line-search can also be used to extend the results of convergence of the standard PSOR method in the particular case of asymmetric PSD matrices and P-matrices. In this case, the analogy with a QP (12.106) is used. The splitting is made on the symmetric matrix $M + M^{\mathrm{T}}$ and B is chosen as a PD matrix. In this case, the direction produced by LCP$(B, q + Cz^{\nu})$ or equivalently QP$(B, q + Cz^{\nu})$ provides a descent direction for $f(z) = z^{\mathrm{T}}(Mz + q)$. The same line-search procedure can be used as before.

The convergence result for this algorithm is given for a row-sufficient matrix M and a PD matrix B. The choice of B as a PD matrix is not so restrictive. Choosing for instance the lower triangular part of a P-matrix is sufficient. We have then a general iterative algorithm with provable convergence for a P-matrix.

Another possibility to solve the LCP with an asymmetric row-sufficient matrix would be to directly apply the gradient projection method of Sect. 12.2.3.2. Unfortunately, the cost of the projection on the feasible set $\mathscr{D} = \{z \mid z \geqslant 0, Mz + q \geqslant 0\}$ may be prohibitive.

12.4.7 Pivoting-Based Methods

Principle

Let us consider a LCP(M, q). If $q \geqslant 0$, then $z = 0$ solves the problem. If there exists an index r such that

$$q_r < 0 \qquad \text{and} \qquad m_{rj} \leqslant 0, \quad \forall j \in \{1, \ldots, n\} \tag{12.129}$$

then there is no vector $z \geqslant 0$ such that $q_r + \sum_j m_{rj} z_i \geqslant 0$. Therefore the LCP is infeasible, thus unsolvable. The LCP rarely possesses these properties in its standard form. The goal of pivoting methods is to derive, by performing pivots, an equivalent system that has one of the previous properties.

Pivotal Algebra

The pivotal algebra is at the heart of pivoting methods. This procedure for LCPs is a simple extension of the pivoting operations in Gauss elimination schemes for linear systems. Let us consider the following linear system:

$$w = q + Mz, \tag{12.130}$$

which is represented in a tableau as

	1	z_1	$\dots$	z_n
w_1	q_1	m_{11}	$\dots$	m_{1n}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
w_n	q_1	m_{n1}	$\dots$	m_{nn}

In the above tableau, the dependent variables w_i are called the basic variables and the independent variables z_i are called the nonbasic variables. A simple pivot consists in exchanging a basic variable w_r with a nonbasic variable z_s. This operation is possible if and only if $m_{rs} \neq 0$ and yields a new definition of the tableau with (w', z', q', M') such that

$$\left.\begin{array}{lll} w'_r = z_s, & w'_i = w_i, & i \neq r \\ z'_s = w_r, & z'_j = z_j, & j \neq s \\ & & \\ q'_r = -q_r/m_{rs}, & q'_i = q_i - (m_{is}/m_{rs})q_r, & i \neq r \\ & & \\ m'_{rs} = 1/m_{rs}, & m'_{is} = m_{is}/m_{rs}, & i \neq r \\ m'_{rj} = -m_{rj}/m_{rs}, \quad j \neq s, & m'_{ij} = m_{ij} - (m_{is}/m_{rs})m_{rj}, & i \neq r, j \neq s \end{array}\right. . \tag{12.131}$$

This pivot operation will be denoted by

$$(w', z', M', q') = \Pi_{rs}(w, z, M, q). \tag{12.132}$$

The representation as a tableau is interesting because the operations on the vector q are identical to those on one column of the matrix M, except for the pivot column.

The simple pivot operation can be extended to a block pivot operation exchanging the same number of basic and nonbasic variables, provided that the pivot block is nonsingular. We will only detail here a principal block pivoting. A block pivoting is said to be a principal block pivoting if the pivot block is a principal submatrix of M. In this case, let us consider an index set α and its complement $\bar{\alpha}$. If the principal submatrix $M_{\alpha\alpha}$ is nonsingular, the principal block pivoting operation consists in

$$\begin{array}{llll} w'_\alpha &= z_\alpha, & w'_{\bar\alpha} &= w_{\bar\alpha} \\ z'_\alpha &= w_\alpha, & z'_{\bar\alpha} &= z_{\bar\alpha} \\ & & & \\ q'_\alpha &= -M_{\alpha\alpha}^{-1} q_\alpha, & q'_{\bar\alpha} &= q_{\bar\alpha} - M_{\bar\alpha\alpha} M_{\alpha\alpha}^{-1} q_\alpha \\ & & & \\ M'_{\alpha\alpha} &= M_{\alpha\alpha}^{-1}, & M'_{\alpha\bar\alpha} &= -M_{\alpha\alpha}^{-1} M_{\alpha\bar\alpha} \\ M'_{\bar\alpha\alpha} &= M_{\bar\alpha\alpha} M_{\alpha\alpha}^{-1}, & M'_{\bar\alpha\bar\alpha} &= M_{\bar\alpha\bar\alpha} - M_{\bar\alpha\alpha} M_{\alpha\alpha}^{-1} M_{\alpha\bar\alpha} \end{array} \quad . \tag{12.133}$$

This principal block pivoting operation will be denoted by

$$(w', z', M', q') = \Pi_\alpha(w, z, M, q). \tag{12.134}$$

The above matrix $M'_{\bar\alpha\bar\alpha}$ is just an instance of the Schur complement with respect to the square matrix $M_{\alpha\alpha}$.

The success of the pivoting method is based on the conservation of the fundamental properties of the matrix M under principal pivoting and principal rearrangement. Among them, the PD and PSD matrices are invariant under pivoting operations. Furthermore, the P-matrix property and the sufficiency are also conserved. The case of the P_0 property is more tricky and needs some additional care, see Cottle et al. (1992, Sect. 4.1).

Murty's Least Index Method

Among the simple principal pivoting methods, also called "Bard-type" algorithms, we have chosen to present Murty's least index method (Murty, 1974) because it has the interest to be one of the simplest pivoting methods to solve an LCP with a P-matrix. It can be described by Algorithm 13.

Algorithm 13 Murty's least index pivoting method

```
Require: $M, q$
Ensure: $z, w$ solution of $LCP(M, q)$ with M a P–matrix.
  $\nu \leftarrow 0$
  $q^\nu \leftarrow q, \quad M^\nu \leftarrow M$
  while $q^\nu \not\geq 0$ do
    Choose the pivot row of index r such that
        $r = \min\{i, q_i^\nu < 0\}$                                  (12.135)
    Pivoting $w_r^\nu$ and $z_r^\nu$.
        $(w^{\nu+1}, z^{\nu+1}, M^{\nu+1}, q^{\nu+1}) \leftarrow \Pi_{rr}(w^\nu, z^\nu, M^\nu, q^\nu)$    (12.136)
    $\nu \leftarrow \nu + 1$
  end while
  $(z^\nu = 0, w^\nu = q^\nu)$ solves $LCP(M^\nu, q^\nu)$.
  Recover the solution of $LCP(M, q)$.
```

Originally, the standard Bard-type methods were designed for LCP(M,q) such that $q = Pb$ and $M = PP^{\mathrm{T}}$, and quickly extended to the case $q = Pb$ and $M = PAP^{\mathrm{T}}$ with a PD matrix A.[3] The only difference with respect to the original Bard-Type scheme is the method for choosing the pivot. In Bard-Type methods, the selection method is

$$r = \arg_i \min\{q_i^\nu\} \tag{12.137}$$

but in Murty's least index method the selection is made thanks to

$$r = \min\{i, q_i^\nu < 0\}. \tag{12.138}$$

It has been observed that the selection rule (12.137) leads to cycling in the degenerate cases. Cycling occurs when the same set of basic variables is found after a certain number of pivoting operations. The method for choosing the pivot in (12.138) which gives the name to the Murty's least index method is crucial for the finite termination of the algorithm. We will see at the end of this section that there exist other methods to prevent cycling, such as the lexicographic degeneracy resolution. For a theoretical analysis of Murty's least index method, we refer to Murty (1988, pp. 258–259).

Minimum Ratio Test

Before presenting a complementary pivot algorithm, we expose the so-called *minimum ratio test*, well known in the linear programming theory. Let us consider a system of linear equations and inequalities of the form

$$\begin{cases} y_i + a_{is}x_s = b_i \\ b_i \geqslant 0 \end{cases} ,i = 1,\dots,m. \tag{12.139}$$

The nonbasic variable x_s is called a driving variable, as it "drives" the values of the basic variables y_i. The largest value of x_s for which $y_i \geqslant 0$ for all $i = 1,\dots,m$ is defined by

$$\hat{x}_s = \sup\{x_s \quad | \quad b_i - a_{is}x_s \geqslant 0, i = 1,\dots,m.\}. \tag{12.140}$$

We set $\hat{x}_s = +\infty$ when $a_{is} \leqslant 0$ for all $i = 1,\dots.m$. If $a_{is} > 0$ for at least one i, then

$$\hat{x}_s = \min\left\{\frac{b_i}{a_{is}} \quad | \quad a_{is} > 0\right\}. \tag{12.141}$$

The minimum ratio test consists in finding an index r such that $a_{rs} > 0$ and $\hat{x}_s = b_r/a_{rs}$ i.e.,

$$r = \arg_i \min\left\{\frac{b_i}{a_{is}} \quad | \quad a_{is} > 0\right\}. \tag{12.142}$$

The variable y_r is called the blocking variable as it blocks the increase of the variable x_s under the nonnegativity constraints $y_i \geqslant 0$ for all $i = 1,\dots,m$. If $\hat{x}_s = \infty$, the variable x_s is said to be unblocked.

The interest of the minimum ratio test is that exchanging the (basic) blocking y_r with the (nonbasic) driving variable, x_s, preserves the nonnegativity of the constant in the right-hand side of (12.139).

[3] Although this form seems to be very restrictive, the case of mechanical systems with perfect unilateral constraints can be cast into this LCP form. See Chap. 13 for more details.

Lemke's Algorithm

Lemke's algorithm (Lemke & Howson, 1964; Lemke, 1965) belongs to the larger class of the complementary pivot algorithms. The name of this class of algorithms is derived from the selection rule of the entering variable in each step, which is always the complementary variable of the dropping variable in the previous step.

For the sake of simplicity, the theoretical justifications and the class of matrices for which the algorithm works will not be treated here in detail. Nevertheless, we attempt to give the basic lines of Lemke's method, in order to justify its choice for specific applications.

A particularity of Lemke's method, and more generally the complementary pivot algorithm, is to consider the following augmented LCP:

$$\begin{cases} w = Mz + q + dz_0 \\ w_0 = q_0 - d^{\mathrm{T}} z \\ 0 \leqslant w \perp z \geqslant 0 \\ 0 \leqslant w_0 \perp z_0 \geqslant 0 \end{cases} \tag{12.143}$$

for a sufficiently large scalar $q_0 \geqslant 0$ and vector $d > 0$. This augmented LCP will be denoted by LCP$(\tilde{M}, \tilde{q})$ with

$$\tilde{z} = \begin{bmatrix} z_0 \; z^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \quad \tilde{w} = \begin{bmatrix} w_0 \; w^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \quad \tilde{q} = \begin{bmatrix} q_0 \; q^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}, \quad \tilde{M} = \begin{bmatrix} 0 & -d^{\mathrm{T}} \\ d & M \end{bmatrix}. \tag{12.144}$$

In contrast to w and z, we assume that the index starts at zero for the components of $\tilde{w}$ and $\tilde{z}$ such that

$$\tilde{z}_0 = z_0,\ \tilde{z}_i = z_i, i = 0, \dots, n, \quad \tilde{w}_0 = w_0,\ \tilde{w}_i = w_i, i = 0, \dots, n. \tag{12.145}$$

The LCP$(\tilde{M}, \tilde{q})$ is known to always possess a solution. Its solution solves LCP(M, q) if $z_0 = 0$. Lemke's method seeks for such a solution. The vector d, which is user-supplied, is often called the covering vector. The augmented LCP allows one to obtain a first feasible basic solution. Indeed, there exists a value $\bar{z}_0 \geqslant 0$ such that

$$w = q + dz_0 \geqslant 0, \quad \forall\, z_0 \geqslant \bar{z}_0. \tag{12.146}$$

This value is given by

$$\bar{z}_0 = \max_i \frac{-q_i}{d_i}. \tag{12.147}$$

If for some i, $q_i < 0$, then $\bar{z}_0 > 0$. The way to obtain a first feasible basic solution is to pivot z_0 with a basic variable. The first pivot row index α is chosen by the minimum ratio such that

$$\alpha = \arg_i \min \left\{ -\frac{q_i}{d_i} \quad | \quad q_i < 0 \right\}. \tag{12.148}$$

It is unique under some nondegeneracy assumptions. This pivot index is chosen such that the basic variable component $w = w_\alpha$ equals zero for $z_0 = \bar{z}_0$. Lemke's method starts by pivoting z_0 and w_α. The remaining part of Lemke's method is described in Algorithm 14.

Algorithm 14 Lemke's method

```
Require: $\mathsf{M},\mathsf{q}$ LCP data and $\mathsf{c}$ the covering vector
Ensure: $\mathsf{z},\mathsf{w}$ solution of $\mathrm{LCP}(\mathsf{M},\mathsf{q})$
  if $\mathsf{q} \geqslant 0$ then $\mathsf{z} = 0, \mathsf{w} = \mathsf{q}$ solves the $\mathrm{LCP}(\mathsf{M},\mathsf{q})$ end if.
  $\nu \leftarrow 0, \quad \tilde{\mathsf{z}}^\nu \leftarrow \left[\mathsf{z}_0\ \mathsf{z}\right]^\mathsf{T}, \quad \tilde{\mathsf{w}}^\nu \leftarrow \left[\mathsf{w}_0\ \mathsf{w}\right]^\mathsf{T}, \quad \tilde{\mathsf{q}}^\nu \leftarrow \left[\mathsf{q}_0\ \mathsf{q}\right]^\mathsf{T}, \quad \tilde{\mathsf{M}}^\nu \leftarrow \begin{bmatrix} 0 & -\mathsf{c}^\mathsf{T} \\ \mathsf{c} & \mathsf{M} \end{bmatrix}$
  Find an index $\alpha > 1$ by using the minimum ratio test,

      $\alpha \leftarrow \arg_\mathsf{i} \min \left\{ -\frac{\mathsf{q}_\mathsf{i}}{\mathsf{c}_\mathsf{i}} \quad | \quad \mathsf{q}_\mathsf{i} < 0 \right\}$        (12.149)

  Pivot $\tilde{\mathsf{z}}_0^\nu = \mathsf{z}_0$ and $\tilde{\mathsf{w}}_\alpha^\nu = \mathsf{w}_\alpha$.

      $(\tilde{\mathsf{w}}^{\nu+1}, \tilde{\mathsf{z}}^{\nu+1}, \tilde{\mathsf{M}}^{\nu+1}, \tilde{\mathsf{q}}^{\nu+1}) \leftarrow \Pi_{\alpha,0}(\tilde{\mathsf{w}}^\nu, \tilde{\mathsf{z}}^\nu, \tilde{\mathsf{M}}^\nu, \tilde{\mathsf{q}}^\nu)$        (12.150)

  Set the index of the driving variable $\mathsf{d} \leftarrow \alpha$. The driving variable is $\mathsf{z}_\alpha^\nu$.
  $\mathsf{IsFound} \leftarrow \mathsf{false}, \quad \mathsf{IsNotFound} \leftarrow \mathsf{false}$

  while $\mathsf{IsFound} = \mathsf{false}$ and $\mathsf{IsNotFound} = \mathsf{false}$ do
    Step 1. Determination of the blocking variable $\tilde{\mathsf{w}}_\mathsf{b}^\nu$
    if $\exists \mathsf{i}, \mathsf{m}_\mathsf{id}^\nu < 0$ then
      Use the minimum ratio test,

          $\mathsf{b} \leftarrow \arg_\mathsf{i} \min \left\{ -\frac{\mathsf{q}_\mathsf{i}^\nu}{\mathsf{m}_\mathsf{id}^\nu} \quad | \quad \mathsf{m}_\mathsf{id}^\nu < 0 \right\}$        (12.151)

    else
      The blocking variable is $\mathsf{w}_0$. $IsNotFound = true$
    end if
    Step 2. Pivoting. The driving variable is blocked.
    if $\mathsf{b} = \alpha$ then
      The blocking variable is $\mathsf{z}_0$. Pivoting $\tilde{\mathsf{w}}_\mathsf{b}^\nu = \mathsf{z}_0$ and $z_d^\nu$.

          $(\tilde{\mathsf{w}}^{\nu+1}, \tilde{\mathsf{z}}^{\nu+1}, \tilde{\mathsf{M}}^{\nu+1}, \tilde{\mathsf{q}}^{\nu+1}) \leftarrow \Pi_{\mathsf{b}+1\mathsf{d}}(\tilde{\mathsf{w}}^\nu, \tilde{\mathsf{z}}^\nu, \tilde{\mathsf{M}}^\nu, \tilde{\mathsf{q}}^\nu)$        (12.152)

      The solution is found. $\tilde{\mathsf{z}}^{\nu+1}$ solves $\mathrm{LCP}(\mathsf{M}^{\nu+1}, \mathsf{q}^{\nu+1})$ with $\mathsf{z}_0 = 0$.
      $\mathsf{IsFound} \leftarrow \mathsf{true}$
    else
      Pivoting the blocking variable $\tilde{\mathsf{w}}_\mathsf{b}^\nu$ and the driving variable $\mathsf{z}_\mathsf{d}^\nu = \mathsf{z}_0$.

          $(\tilde{\mathsf{w}}^{\nu+1}, \tilde{\mathsf{z}}^{\nu+1}, \tilde{\mathsf{M}}^{\nu+1}, \tilde{\mathsf{q}}^{\nu+1}) \leftarrow \Pi_\mathsf{bd}(\tilde{\mathsf{w}}^\nu, \tilde{\mathsf{z}}^\nu, \tilde{\mathsf{M}}^\nu, \tilde{\mathsf{q}}^\nu)$        (12.153)

    end if
    Set the index of the driving variable $\mathsf{d} \leftarrow \mathsf{b}$.
    $\nu \leftarrow \nu + 1$
  end while
  if $\mathsf{IsNotFound} = \mathsf{true}$ then
    Interpret the output in terms of infeasibility or unsolvability.
  end if
  if $\mathsf{IsFound} = \mathsf{true}$ then Recover the solution of $\mathrm{LCP}(\mathsf{M},\mathsf{q})$ end if.
```

Lexicographic Degeneracy Resolution

Some adaptation in pivoting algorithms such as Lemke's method avoid cycling during the pivoting process and are of crucial importance in practical situations. The solution is to base the choice of the pivot and the minimum ratio test on a lexicographic ordering (Cottle et al., 1992; Murty, 1988). This is called the *lexicographic degeneracy resolution*. The notion of lexicographic ordering of vectors has been introduced to obtain a specific ordering relation between vectors defined as follows.

Definition 12.37. *A vector* $z = (z_1, ..., z_n)^T \in \mathbb{R}^n$ *is said to be lexicographically positive (resp. negative) and denoted* $z \succ 0$ *(resp.* $z \prec 0$ *), if* $z \neq 0$ *and its first nonzero component (i.e., lowest indexed) is strictly positive (resp. negative).*

Using this definition, a vector of $\mathbb{R}^n$ is either lexicographically positive, negative, or zero. This classification cannot be obtained by the usual ordering relations $\leqslant$ and $\geqslant$ of $\mathbb{R}$.

Definition 12.38. *Let two arbitrary vectors* z_1 *and* $z_2 \in \mathbb{R}^n$*. Then* z_1 *is lexicographically greater than (resp. less than)* z_2 *if and only if* $z^1 - z^2 \succ 0$ *(resp.* $z^1 - z^2 \prec 0$*). In this case we note* $z^1 \succ z^2$ *(resp.* $z^1 \prec z^2$*).*

Thus for a set of vectors $\{z^1, ..., z^k\}$, z^m is a lexico minimum (resp. maximum) if for each $i = 1, ..., k$, $z^m \prec z^i$ (resp. $z^m \succ z^i$).

Proposition 12.39. *Every nonempty finite subset of* $\mathbb{R}^n$ *has a unique lexicographic minimum and a unique lexicographic maximum.*

The standard LCP(M,q) tableau is introduced to start the algorithm with a feasible basis, where the variable z_0 is associated with the covering vector d and Q is chosen to be a nonsingular matrix with lexicographic positive rows. The identity matrix is often chosen. The following notation, which is used as standard matrix formulation, is also introduced:

$$\mathcal{Q} = [\, q \mid Q \,] \quad \text{and} \quad \mathcal{M} = [\, d \mid M \,].$$

The pivoting operations are generalized to Table 12.1 and are also applied to Q, providing $Q^{\nu+1}$ from Q^{ν}.

Using lexicographic ordering and Proposition 12.39, the choice of the pivot variable is unique and allows one to obtain a solution when the problem is degenerate.

Table 12.1. Classical augmented LCP tableau

1	w	z_0	z
q	Q	d	M

$w \geqslant 0, \quad z \geqslant 0, \quad z_0 \geqslant 0$

The first minimum ratio test (12.149) for computing b in Algorithm 14 is replaced by the analog with the lexicographic ordering:

$$b = \arg_i \operatorname*{lexicomin}_{Q^{\nu}_{i0}<0} \left\{ -\frac{1}{d_i} Q^{\nu}_{i\bullet} \right\} \text{ for } \nu = 0. \tag{12.154}$$

The second minimum ratio test (12.151) is also replaced by its analog with the lexicographic ordering:

$$b = \arg_i \operatorname*{lexicomin}_{m^{\nu}_{id}<0} \left\{ -\frac{1}{m^{\nu}_{id}} Q^{\nu}_{i\bullet} \right\}. \tag{12.155}$$

The success of the lexicographic degeneracy resolution is based on the fact that the pivot selection rule implies a strict decrease of a vector-valued function at each pivoting. This strict decrease ensures that every basic solution is used only one time.

What Are the LCPs for Which the Pivoting Method Works?

This question is rather complicated and is difficult to answer in few lines. We prefer to refer to the monographs of Cottle et al. (1992) and Murty (1988) for a rigorous treatment of this subject. A result ensures that for a feasible LCP with a matrix M which is copositive plus,[4] the complementarity pivot algorithm terminates at a solution of the LCP. If it does not, the LCP is feasible. Other results can be found for other classes of matrices (semi-monotone, P_0-matrices, etc.) in the above cited books.

12.4.8 Interior Point Methods

As said at the end of Sect. 12.2.3.3, the literature on this subject and the number of published algorithms is large. We restrict this section to practical considerations rather than theoretical complexity properties.

12.4.8.1 The Horizontal Monotone LCP

Let us start with the horizontal monotone LCP defined by

$$\begin{cases} Qx + Rs = q \\ 0 \leqslant x \perp s \geqslant 0 \end{cases} \tag{12.156}$$

together with the monotonicity property

$$Qx + Rs = 0 \Longrightarrow s^{\mathrm{T}} x \geqslant 0. \tag{12.157}$$

We assume the monotonicity property because most of the interior point methods have been designed and have been proved to converge under these assumptions. One fundamental property of the horizontal monotone LCP is that the set of solutions (x,s) is convex. We will see in Sect. 12.4.8.2 that this property can be extended to larger class of LCPs.

[4] This means: $x^{\mathrm{T}}Mx \geqslant 0$ for all $x \geqslant 0$ and $[x^{\mathrm{T}}Mx = 0, x \geqslant 0] \Rightarrow (M + M^{\mathrm{T}})x = 0$.

Basic Principle of the Primal–Dual Interior Point Methods

Recalling that the horizontal monotone LCP (12.156) can be viewed as the KKT conditions of a convex QP, the designation "primal–dual method" is derived from the fact that we solve the problem for both x and s which are primal and dual variables of a QP. One fundamental notion in primal–dual interior point methods is the notion of central path defined below.

Definition 12.40 (Central path). *The central path for the horizontal monotone LCP (12.156) is the set of points (x,s) defined by*

$$\begin{cases} x \circ s = \mu \mathbb{1} \\ Qx + Rs = q \\ x \geqslant 0, \quad s \geqslant 0 \end{cases} \tag{12.158}$$

for μ describing the half-line, $\mathbb{R}_+$. Here, $\mathbb{1}$ is the vector whose components are all equal to 1.

Obviously, for $\mu = 0$, the central path equation (12.158) is equivalent to the horizontal monotone LCP (12.156). We can remark that for $\mu > 0$, any point of the central path lies in the strictly primal–dual feasible domain defined by

$$\mathscr{F}^\circ = \{x, s \in \mathbb{R}^n \mid Qx + Rs = q, x > 0, s > 0\}. \tag{12.159}$$

The central path can also be interpreted as the locus of minima in $\mathscr{F}^\circ$ of a particular mixed potential

$$\Phi(x,s) = x^{\mathrm{T}} s - \mu \sum_{i=1}^{n} \log x_i s_i = x^{\mathrm{T}} s + \mu \phi(x) + \mu \phi(s), \tag{12.160}$$

where the logarithmic potential

$$\phi(x) = -\sum_{i=1}^{n} \log x_i \tag{12.161}$$

is strictly convex. More precisely, every point (x^μ, s^μ, μ) of the central path is the unique solution of the following optimization problem:

$$\begin{array}{ll} \min_{x,s} & \Phi(x,s) \\ \text{subject to} & Qx + Rs = q \\ & x > 0 \\ & s > 0 \end{array} . \tag{12.162}$$

Recalling that solving the horizontal monotone LCP (12.156) amounts to solving the quadratic problem

$$\begin{array}{ll} \min_{x,s} & x^{\mathrm{T}} s \\ \text{subject to} & Qx + Rs = q \\ & x \geqslant 0 \\ & s \geqslant 0, \end{array} \tag{12.163}$$

the potential $\Phi(x,s)$ can be interpreted as the logarithmic penalty associated with (12.163). As with primal interior points and barrier methods, we see the link between the primal–dual interior point methods and the logarithmic penalty.

Driving the iterates toward a solution of the horizontal LCP can be seen from two points of view. The first one is an approximate minimization of the mixed potential (12.160) (or some other variants) for a sequence of μ that converges to 0. Then one speaks of potential reduction methods. Alternatively, one may say that the central path is approximated for a sequence of μ that converges to 0.

In any case, the direction between two iterates is the Newton direction associated with

$$\begin{cases} x \circ s = \sigma\mu \mathbb{1} \\ Qx + Rs = q. \end{cases} \tag{12.164}$$

The strict feasibility assumption is made, i.e., $(x,s) \in \mathscr{F}^{\circ}$ and $\sigma \in [0,1]$ is the reduction parameter of μ. Linearizing the problem (12.164) around the current point (x,s) results in the following linear system for the direction (u,v):

$$\begin{cases} s \circ u + x \circ v = \sigma\mu \mathbb{1} - x \circ s \\ Qu + Rv = 0 \end{cases} . \tag{12.165}$$

We introduce a matrix notation of the previous system:

$$\begin{bmatrix} S & X \\ Q & R \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sigma\mu \mathbb{1} - x \circ s \\ 0 \end{bmatrix}, \tag{12.166}$$

where the matrix $S \in \mathbb{R}^{n\times n}$ and $X \in \mathbb{R}^{n\times n}$ are defined by $S = \mathrm{diag}(s)$ and $X = \mathrm{diag}(x)$. Two extreme choices of σ are often encountered in practice:

(a) The value $\sigma = 1$ defines the so–called centering direction (or a centralization displacement). Indeed a Newton step points toward (x^{μ}, s^{μ}) on the central path: $x_i^{\mu} s_i^{\mu} = \mu, i = 1,\dots,n$. A displacement along a centering direction makes little progress, if any, toward reducing the value of μ.

(b) The value $\sigma = 0$ defines the so-called affine-scaling direction, which the standard Newton step for the system

$$\begin{cases} x \circ s = 0 \\ Qx + Rs = q. \end{cases} \tag{12.167}$$

This step therefore should ensure a decrease of μ.

Most of the algorithms choose an intermediate value for σ to have a good trade-off between reducing μ and improving centrality. Finally, once the direction is chosen through a value of σ, a step length α has to be chosen in the direction (u,v) to respect the strict feasibility. Algorithm 15 describes a general scheme for primal–dual interior point methods.

The literature on this subject is vast and a huge amount of algorithms have been proved to converge for various selection strategies of α_k and σ_k. All of these algorithms may differ slightly in their formulation and it is always difficult to compare

Algorithm 15 General scheme of the primal–dual interior point methods

```
Require: $Q, R, q, tol$
Require: $(x_0, s_0) \in \mathscr{F}^\circ$
Ensure: $x, s$ solution of $\mathrm{hLCP}(Q, R, q)$
  $\mu_0 \leftarrow \frac{x_0^T s_0}{n}$
  $k \leftarrow 0$
  while $\mu_k > tol$ do
    Solve
        $$\begin{bmatrix} S^k & X^k \\ Q & R \end{bmatrix} \begin{bmatrix} u^k \\ v^k \end{bmatrix} = \begin{bmatrix} \sigma^k \mu^k \mathbb{1} - x^k \circ s^k \\ 0 \end{bmatrix} \qquad (12.168)$$
    for some $\sigma_k \in (0, 1)$.
    Choose $\alpha_k$ such that
        $$(x^{k+1}, s^{k+1}) \leftarrow (x^k, s^k) + \alpha_k (u^k, v^k) \qquad (12.169)$$
    is strictly feasible i.e., $x^{k+1} > 0, s^{k+1} > 0$
    $\mu_k \leftarrow \frac{x_k^T s_k}{n}$
  end while
```

their practical efficiencies. In what follows we present formally only standard algorithms which are representative of a class of interior point methods.

Path-Following or Central Path Following Methods

The path-following primal–dual interior point methods generate a sequence of strictly feasible points satisfying *approximately* the central path equation (12.158) for a sequence of μ that converges to 0. The approximation may be measured by the centrality measure

$$\delta(s, x, \mu) = \left\| \frac{x \circ s}{\mu} - \mathbb{1} \right\|_2 . \qquad (12.170)$$

In most methods, the sequence of points is constrained to lie in one of the following two neighborhoods of the central path: the small neighborhood parametrized by θ

$$\mathscr{N}_2(\theta) = \{(x, s) \in \mathscr{F}^\circ \mid \ \|x \circ s - \mu \mathbb{1}\|_2 \geqslant \mu\theta\} \text{ for some } \theta \in (0, 1) \qquad (12.171)$$

and the large neighborhood parametrized by ε

$$\mathscr{N}_{-\infty}(\varepsilon) = \{(x, s) \in \mathscr{F}^\circ \mid x_i s_i \geqslant \mu\varepsilon\} \text{ for some } \varepsilon \in (0, 1). \qquad (12.172)$$

Path-following methods follow the general scheme Algorithm 15 and differ in the selection of the neighborhood type. Once a neighborhood type and size has been selected, the method chooses a relation between the parameters σ_k, α_k, and the size of the neighborhood, θ or ε. Most well-known instances of path-following methods are (i) the short-step path-following method, (ii) the long-step path-following method, and (iii) the standard predictor–corrector path-following algorithm. We will only sketch these algorithms; indeed they have only a theoretical interest. Only the long-step path-following methods have proved to be efficient in several situations.

The short-step path-following algorithm starts with a feasible point in a relatively small neighborhood $\mathscr{N}_2(\theta)$ of the central path and generates a sequence of point with stays in $\mathscr{N}_2(\theta)$. This choice induces a relation between θ and the value σ_k, typically, $\theta = 0.4$ and $\sigma_k = 1 - 0.4/\sqrt{n}$. The step length is chosen equal to 1. If the short-step methods are interesting from the complexity point of view, the value of σ_k implies very low convergence rate, especially for large systems. Such interior point methods have therefore almost only a theoretical interest.

The standard predictor–corrector as it has been published by Mizuno et al. (1993) is based on a two-step procedure and a pair of two neighborhoods, $\mathscr{N}_2(\theta_1)$ and $\mathscr{N}_2(\theta_2)$ with $\theta_1 < \theta_2$. The predictor step, which is an affine-scaling step with $\sigma_k = 0$, reduces the value μ starting from $\mathscr{N}_2(\theta_1)$ and choosing the step length such that the iterate ends in $\mathscr{N}_2(\theta_2)$. On the contrary, the corrector step, which is a centralization step with $\sigma_k = 1$, improves the centralization starting from $\mathscr{N}_2(\theta_2)$ and choosing the step-length equal to 1 such that the iterate ends in $\mathscr{N}_2(\theta_1)$. The standard predictor–corrector scheme improves the short-step path-following scheme from the practical point of view; and both schemes have the same theoretical complexity. Nevertheless, the use of a small neighborhood such as $\mathscr{N}_2$ restricts fast convergence during the early iterations of the algorithm.

The long-step path-following method uses a large neighborhood of the central path, $\mathscr{N}_{-\infty}(\varepsilon)$ with small value of ε, let us say, 10^{-3}. The centering parameter σ is chosen between the values $\sigma_{\min}$ and $\sigma_{\max}$. The step length is chosen such that the iterates stay in the large neighborhood, $\mathscr{N}_{-\infty}(\varepsilon)$. In practice, this algorithm seems to be better than the two previous algorithms.

Practical Implementations

Most of the practical implementations of interior point methods are based on Mehrotra's predictor–corrector algorithm (Mehrotra, 1992) which is an infeasible interior point method. The term "infeasible" refers to the fact that the equality constraint $Qx + Rx = q$ is relaxed into $Qx + Rx = q + r$ and the algorithm tries to minimize both μ and r. The infeasible strategy allows one to start with primal–dual initial points which are just strictly positive. Furthermore, the Mehrotra predictor–corrector scheme uses some tricks to correct the Newton directions and to evaluate adaptively the centering parameter. Fore more details, we refer to Wright (1996b) and Bonnans et al. (2003).

Finally, a crucial point in the efficiency of interior point methods is the linear solvers to compute Newton's directions. More details can be found on this important aspect in Wright (1996b, Chap. 11) and Andersen et al. (1996, Sect. 4). Another question is the purification stage which can be useful if we want to know accurately what are the active constraints; this is also discussed in Bonnans et al. (2003).

12.4.8.2 The General Case

A lot of theoretical results have been extended to more general LCPs, particularly, the $P_\star(\kappa)$ class of matrices. We refer to the following works for precise statements of the algorithm: Anitescu et al. (1997), Potra & Liu (2005), Potra & Sheng (1997), and Illés et al. (2007) and references therein. The question of practical efficiency with respect to convergence and complexity results is largely open.

12.4.9 How to Choose an LCP Solver?

As with the QP, it is also difficult to give firm and easy rules to choose the right LCP solver. Nevertheless, we can attempt to give the following advices:

1. The splitting methods are well suited
   - for very large and well-conditioned LCP. Typically, the LCPs with symmetric PD matrix are solved very easily by a splitting method,
   - when a good initial solution is known in advance.
2. The pivoting techniques are well suited
   - for small to medium system sizes ($n < 5000$),
   - for "difficult problems" when the LCP has only a P-matrix, sufficient matrix, or copositive plus matrix,
   - when one wants to test the solvability of the system.
3. Finally, interior point methods can be used
   - for large-scale problems without the knowledge of a good starting point,
   - when the problem has a special structure that can be exploited directly in solving the Newton direction with an adequate linear solver.

We also mention the existence of "generalized or semi-smooth Newton's methods" which we postpone to Sect. 12.5.4. The algorithms, their implementation, and the convergence results are similar for Nonlinear Complementarity Problem (NCP)s and LCPs. Only in the linear case can some part of the algorithm be optimized using the linearity.

12.5 The Nonlinear Complementarity Problem (NCP)

12.5.1 Definition and Basic Properties

The NCP is somehow a nonlinear version of a LCP defined as follows :

Definition 12.41 (Nonlinear Complementarity Problem (NCP)). *Given a mapping* $F\colon \mathbb{R}^n \to \mathbb{R}^n$, *the Nonlinear Complementarity Problem (NCP) denoted by* $\mathrm{NCP}(F)$ *is to find a vector* $z \in \mathbb{R}^n$ *such that*

$$0 \leqslant z \perp F(z) \geqslant 0. \tag{12.173}$$

A vector z is called feasible (respectively, strictly feasible) for the $\mathrm{NCP}(F)$ *if* $z \geqslant 0$ *and* $F(z) \geqslant 0$ *(respectively,* $z > 0$ *and* $F(z) > 0$*).*

The following standard index sets are defined, for any vector z:

$$\begin{aligned} \alpha(z) &= \{i \mid z_i > 0 = F_i(z)\} \\ \beta(z) &= \{i \mid z_i = 0 = F_i(z)\}\,. \\ \gamma(z) &= \{i \mid z_i = 0 < F_i(z)\} \end{aligned} \tag{12.174}$$

A solution $\bar{z}$ of NCP(F) is said to be degenerate if the index set $\beta(\bar{z})$ is a nonempty set.

Basic Existence and Uniqueness Properties

The analog property of (semi)-positive definiteness of M in LCP(M,q) is the monotonicity property of the function $F(\cdot)$, for which we recall some definitions.

Definition 12.42. *A given mapping $F\colon X \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is said to be*

(a) monotone on X if

$$(x-y)^T(F(x)-F(y)) \geqslant 0, \textit{ for all } x,y \in X; \tag{12.175}$$

(b) strictly monotone on X if

$$(x-y)^T(F(x)-F(y)) > 0, \textit{ for all } x,y \in X, x \neq y; \tag{12.176}$$

(c) strongly monotone on X if there exists $\mu > 0$ such that

$$(x-y)^T(F(x)-F(y)) \geqslant \mu\|x-y\|^2, \textit{ for all } x,y \in X; \tag{12.177}$$

(d) pseudo-monotone on X if

$$(x-y)^T F(y) \geqslant 0 \Rightarrow (x-y)^T F(x) \geqslant 0, \textit{ for all } x,y \in X. \tag{12.178}$$

For an affine mapping $F(x) = Mx+q$, (a) (respectively (b)) means that M is PSD (respectively PD). Furthermore, if $F(\cdot)$ is continuously differentiable on a open convex set $\mathscr{D}$, the following results hold:

Theorem 12.43. *Given a continuously differentiable mapping $F\colon \mathscr{D} \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ on the open convex set $\mathscr{D}$, the following statements are valid:*

(a) $F(\cdot)$ is monotone on $\mathscr{D}$ if and only if $\nabla^T F(x)$ is PSD for all $x \in \mathscr{D}$.
(b) $F(\cdot)$ is strictly monotone on $\mathscr{D}$ if $\nabla^T F(x)$ is PD for all $x \in \mathscr{D}$.
(c) $F(\cdot)$ is strongly monotone on $\mathscr{D}$ if and only if $\nabla^T F(x)$ is uniformly PD for all $x \in \mathscr{D}$, i.e.,

$$\exists \mu > 0, \qquad z^T \nabla^T F(x) z^T \geqslant \mu\|z\|^2, \quad \forall x \in \mathscr{D}. \tag{12.179}$$

Theorem 12.44. *Given a continuous mapping $F\colon X \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$, the following statements hold:*

(a) If $F(\cdot)$ is monotone on $X = \mathbb{R}^n_+$, the NCP(F) *has a convex (possibly empty) solution set. Furthermore, if there exists a strictly feasible point, the* NCP(F) *has a nonempty and compact solution set.*
(b) If $F(\cdot)$ is strictly monotone on $X = \mathbb{R}^n_+$, the NCP(F) *has at most one solution.*
(c) If $F(\cdot)$ is strongly monotone on $X = \mathbb{R}^n_+$, the NCP(F) *has a unique solution.*

The proof of this result can be found in Moré & Rheinbolt (1973). The above results are standard results related to the equivalent reformulation of NCPs as VIs. In fact, numerous results can be stated directly in the context of VIs and then can be applied to NCPs as a specification of VIs. Some of these results that can be found in Harker & Pang (1990) and Facchinei & Pang (2003) will be given in Sect. 12.6. Some results which are derived from the study of LCPs concern only certain classes of VIs which contain the NCP. This is the case for a VI on a box, i.e., a Cartesian product of n-dimensional closed intervals, for which the notion of P-function generalizes that of P-matrix.

Definition 12.45. *A given mapping $F : X \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is said to be*

(a) a P-function on X if

$$\max_{i=1,\dots,n} (x_i - y_i)(F_i(x) - F_i(y)) > 0, \qquad \forall x, y \in X, x \neq y; \tag{12.180}$$

(b) a uniform P-function if

$$\exists\, \mu > 0, \quad \max_{i=1,\dots,n} (x_i - y_i)(F_i(x) - F_i(y)) \geqslant \mu \|x-y\|^2, \qquad \forall x, y \in X, x \neq y. \tag{12.181}$$

Obviously, if $F(\cdot)$ is strictly monotone on X, then it is a P-function on X. If $F(\cdot)$ is strongly monotone on X, then it is a uniform P-function on X. With the above definitions we have the following result which is valid for VIs over boxes. We state here this result in the particular case of NCP.

Theorem 12.46. *Given a continuous mapping $F\colon X \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$, the following statements hold:*

(a) If $F(\cdot)$ is a P-function on X, then the NCP(F) *has at most one solution.*
(b) If $F(\cdot)$ is a uniform P-function on X, then the NCP(F) *has a unique solution.*

The proof can be found in Moré (1974).

Reformulations of the NCP in Terms of NLP

An NCP can be reformulated as several forms of well-known problems of mathematical programming (see Sect. 12.6 for VI and inclusion into a normal cone.). We focus our interest in this section on the reformulation in terms of NLP. This reformulation is the analog of the reformulation of an LCP as a QP. The following result generalizes the linear/quadratic case.

Theorem 12.47. *Given a mapping* $F : X \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$. *A vector* $\bar{z}$ *solves the* NCP(F) *if and only if* $\bar{z}$ *solves the following NLP:*

$$\begin{aligned} &\textit{minimize} \;\; z^T F(z) \\ &\textit{subject to} \;\; F(z) \geqslant 0 \\ &\qquad\qquad\quad z \geqslant 0 \end{aligned} \tag{12.182}$$

and the value of the objective function is equal to zero, i.e.,

$$\bar{z}^T F(\bar{z}) = 0. \tag{12.183}$$

Note that in the previous theorem, nothing has been said about feasibility. As with LCPs, the complete equivalence with a NLP can also be considered. This discussion will be pursued in the more general setting of VIs in Sect. 12.6.

General minimization problems can also be written, based on the notion of merit function.

Definition 12.48. *A function* $\Psi \colon \mathbb{R}^n \longrightarrow \mathbb{R}_+$ *is called a merit function for the* NCP(F) *if it has both the properties*

(a) $\Psi(z) \geqslant 0$ *for all* $x \in \mathbb{R}^n$,
(b) $\Psi(z) = 0$ *if and only if* z *solves* NCP(F).

If $\Psi(\cdot)$ is a merit function, the following unconstrained minimization problem is relevant:

$$\min_x \Psi(x). \tag{12.184}$$

There are many merit functions in the literature. To cite a few of them, the following implicit Lagrangian suggested in Mangasarian & Solodov (1993)

$$\begin{aligned} \Psi(x) = \sum_{i=1}^{n} \Bigg[& x_i F_i(x) + \frac{1}{2\alpha}(\max{}^2(0, x_i - \alpha F_i(x)) - x_i^2 \\ & + \max{}^2(0, F_i(x) - \alpha x_i) - F_i^2(x)) \Bigg] \end{aligned} \tag{12.185}$$

is an example of merit function. The Fischer–Burmeister merit function (Facchinei & Soares, 1997) can also be mentioned:

$$\Psi(x) = \frac{1}{2} \sum_{i=1}^{n} \left(\sqrt{x_i^2 + F_i^2(x)} - x_i - F_i(x) \right)^2. \tag{12.186}$$

We will see that for most of the equation-based formulations (see Sect. 12.5.4) of an NCP it is possible to define a merit function. One could think to solve directly the minimization problem (12.184) with one of these merit functions. The difficulty lies in the fact that to have efficient minimization solvers, the merit must be at least twice differentiable. This is not the case for most merit functions. These functions are preferably used to monitor the global convergence of the numerical methods by line-search procedures.

Other Reformulations

Many other reformulations of NCPs are also widely used for numerical purposes, such as VI reformulation and equation-based reformulation. They are closely related to numerical methods, therefore we will present them in the subsequent sections, which are inspired by the excellent review paper of Ferris & Kanzow (2002).

12.5.2 The Mixed Complementarity Problem (MCP)

Similar to the LCP with respect to the MLCP, the Mixed Complementarity Problem (MCP) is a special case of the NCP where the system is defined by a set of nonlinear equations, while the complementarity is only applied to some variables and functions. This leads to the following problem definition:

Problem 12.49 (Mixed Complementarity Problem (MCP)). Given two mappings $G\colon \mathrm{I\!R}^{n_1} \times \mathrm{I\!R}^{n_2}_{+} \mapsto \mathrm{I\!R}^{n_1}$ and $H\colon \mathrm{I\!R}^{n_1} \times \mathrm{I\!R}^{n_2}_{+} \to \mathrm{I\!R}^{n_2}$, the Mixed Complementarity Problem (MCP) mixed complementarity problem (MCP) denoted by $\mathrm{MCP}(G,H)$ is to find a pair of a vectors $u, v \in \mathrm{I\!R}^{n_1} \times \mathrm{I\!R}^{n_2}$ such that

$$\begin{cases} G(u,v) = 0 \\ \\ 0 \leqslant v \perp H(u,v) \geqslant 0. \end{cases} \tag{12.187}$$

The following problem is equivalent to Problem 12.49:

Problem 12.50. Given two index sets $\mathscr{C}$ (for constrained) and $\mathscr{F}$ (for free) forming a partition of the set $\{1,2,\ldots,n\}$ and two mappings $F_{\mathscr{C}}\colon \mathrm{I\!R}^n \to \mathrm{I\!R}^c$, $F_{\mathscr{F}}\colon \mathrm{I\!R}^n \to \mathrm{I\!R}^f$, such that $f + c = n$, find a vector $z \in \mathrm{I\!R}^n$ such that

$$\begin{cases} F_{\mathscr{F}}(z) = 0 \\ \\ 0 \leqslant z_{\mathscr{C}} \perp F_{\mathscr{C}}(z) \geqslant 0. \end{cases} \tag{12.188}$$

When only box constraints are encountered, we obtain the so-called box-constrained Mixed Complementarity Problem (MCP) given by

Definition 12.51 (Box-constrained MCP). *Given a mapping $F : \mathrm{I\!R}^n \to \mathrm{I\!R}^n$ and bounds $b_l, b_u \in R^n \cup \{-\infty, +\infty\}$, the box-constrained MCP is to find a vector $z \in \mathrm{I\!R}^n$ and vectors $v, w \in \mathrm{I\!R}^n$ such that*

$$\begin{cases} F(z) = w - v \\ \\ 0 \leqslant (z - b_l) \perp w \geqslant 0 \\ \\ 0 \leqslant v \perp (b_u - z) \geqslant 0. \end{cases} \tag{12.189}$$

The relation with the NLP problem (12.47) is given by the KKT necessary conditions which can be written as

$$\begin{cases} \mathscr{L}(z,u,v) = \nabla f(z) + u\nabla g(z) + v\nabla h(z) = 0 \\ h(z) = 0 \\ 0 \leqslant g(z) \perp u \geqslant 0 \end{cases} , \tag{12.190}$$

which is a MCP.

12.5.3 Newton–Josephy's and Linearization Methods

Newton–Josephy's Method

Assuming that $F(\cdot)$ is continuously differentiable, a natural idea for solving a NCP is to linearize $F(\cdot)$ at a current iterate z_k in order to retrieve a LCP at each step. If the standard Newton method to linearize $F(\cdot)$ is used, the following LCP

$$0 \leqslant z \perp F(z_k) + \nabla F(z_k)(z - z_k) \geqslant 0 \tag{12.191}$$

has to be solved to obtain z_{k+1}. This method is known as the Newton–Josephy's method or Successive Linear Complementarity Problem (SLCP). Convergence, which has been studied in Josephy (1979), shows that the iterates are locally well defined and fastly convergent under the fundamental property of *strong regularity*. This property introduced by Robinson (1980) in the context of generalized equations is defined below.

Definition 12.52. *Let $\bar{z}$ be a solution of* NCP(F) *and the associated standard index sets $\alpha = \alpha(\bar{z})$ and $\beta = \beta(\bar{z})$ defined in (12.174). The vector $\bar{z}$ is called a strongly regular solution if the submatrix*

$$[\nabla F(\bar{z})]_{\alpha\alpha} \tag{12.192}$$

is nonsingular and the Schur complement

$$[\nabla F(\bar{z})]_{\beta\beta} - [\nabla F(\bar{z})]_{\beta\alpha}[\nabla F(\bar{z})]^{-1}_{\alpha\alpha}[\nabla F(\bar{z})]_{\alpha\beta} \tag{12.193}$$

is a P-matrix.

Just as the standard Newton method, the Newton–Josephy method works extremely well with a good starting point, i.e., locally. Many difficulties arise to monitor the global convergence. As with the SQP method 12.3, the two major problems are: an unsolvable one-step LCP far from the solution and a choice of a merit function that does not destroy the fast local convergence. The theoretical study of this method is analog to the study of successive QP (SQP) approach. A short treatment can be found in Cottle et al. (1992, Sect. 7.4).

The Newton–Robinson Method and the PATH Solver

For theoretical considerations, Robinson (1988,1992) proposed to use a linearization of the so-called normal map (see Sect. 12.6 for a general definition for VIs)

$$F^{\mathrm{nor}}(y) = F(y^+) + (y - y^+), \tag{12.194}$$

where $y^+ = \max(0, y)$ stands for the positive part of y. Equivalence with the NCP(F) is as follows: y is a zero of the normal map if and only if y^+ solves NCP(F). The Newton–Robinson method uses a piecewise linear approximation of the normal map, namely

$$L_k(y) = F(y_k^+) + \nabla F(y_k^+)(y^+ - y_k^+) + y - y^+. \tag{12.195}$$

The Newton iterate y_{k+1} is a zero of $L_k(\cdot)$. The same y_{k+1} would be obtained by Newton–Josephy's method if z^k were set to y_k^+ in (12.191).

The great advantage of Newton–Robinson's method from the numerical point of view is its ability to be efficiently globalized. Indeed, Ralph (1994) extended the damped Newton method for smooth equations to the normal map via a path $p^k(t)$ from y^k to y^{k+1} defined by

$$L_k(p_k(t)) = (1 - t)F^+(y_k). \tag{12.196}$$

A "path-search" (as opposed to line search) is then performed using the merit function $\|F_+(y)\|$. Standard theory of damped Newton's method can be extended to prove standard local and global convergence results (Ralph, 1994; Dirkse & Ferris, 1995). The construction of the piecewise linear path p_k is based on the use of pivoting methods. Each pivot corresponds to a kink in the path. In Dirkse & Ferris (1995), a modification of Lemke's algorithm is proposed to construct the path. Once the path is constructed, a path-search procedure is performed, that is to find a point on the path that satisfies some descent properties. Several path-search strategies are also proposed.

The PATH solver (Dirkse & Ferris, 1995) is an efficient implementation of Newton–Robinson's method together with the path-search scheme. It has proved to be very efficient on a large variety of problems (Billups et al., 1997). We will not discuss in detail the implementation of the PATH solver. Many improvements and evolutions have been proposed over the years, we refer especially to Ferris & Munson (1999) and Munson (2000) where a merit function based on the Fischer–Burmeister merit function (12.186) is used. In practice, the implementation of the PATH solver addresses the MCP stated in Sect. 12.5.2.

12.5.4 Generalized or Semismooth Newton's Methods

The principle of the generalized or semismooth Newton's method for LCPs is based on a reformulation in terms of possibly nonsmooth equations using the so-called C-function also called NCP-function.

Definition 12.53. *A function $\phi : \mathbb{R}^2 \to \mathbb{R}$ is called a C-function (for complementarity) if*

$$0 \leqslant w \perp z \geqslant 0 \Longleftrightarrow \phi(w,z) = 0. \tag{12.197}$$

Well-known examples of C-function are

$$\begin{aligned}
\phi(w,z) &= \min(w,z), && (12.198a)\\
\phi(w,z) &= \max(0,w-\rho z)-w,\ \ \rho > 0, && (12.198b)\\
\phi(w,z) &= \max(0,z-\rho w)-z,\ \ \rho > 0, && (12.198c)\\
\phi(w,z) &= \sqrt{w^2+z^2}-z-w, && (12.198d)\\
\phi(w,z) &= \lambda(\sqrt{w^2+z^2}-z-w)-(1-\lambda)w_+z_+),\ \lambda \in (0,1), && (12.198e)\\
\phi(w,z) &= -wz+\frac{1}{2}\min^2(0,w+z). && (12.198f)
\end{aligned}$$

The function (12.198d) is called the Fischer–Burmeister function (Fischer, 1992) and (12.198e) the penalized Fischer–Burmeister function (Chen et al., 2000). The particularity of the function (12.198f) is that it is differentiable on the whole space $\mathbb{R}^2$ Evtushenko & Purtov, 1984). Many other C-functions can be found in the literature (Mangasarian, 1976; Sun & Qi, 1999; Qi & Yang, 2002, to mention a few).

Then defining the following function associated with NCP(F):

$$\Phi(z) = \begin{bmatrix} \phi(F_1(z),z_1) \\ \vdots \\ \phi(F_i(z),z_i) \\ \vdots \\ \phi(F_n(z),z_n) \end{bmatrix}, \tag{12.199}$$

we obtain as an immediate consequence of the definitions of $\varphi(\cdot)$ and $\Phi(\cdot)$ the following equivalence.

Proposition 12.54. *Let $\phi(\cdot)$ be a C-function and the corresponding operator $\Phi(\cdot)$ defined by (12.199). A vector $\bar{z}$ is a solution of* NCP(F) *if and only if $\bar{z}$ solves the nonlinear system of equations $\Phi(z) = 0$.*

The generalized or semismooth Newton's method consists in applying a Newton-type algorithm for searching a zero of $\Phi(\cdot)$. If the operator $\Phi(\cdot)$ is chosen to be at least locally Lipschitzian, it is therefore almost everywhere differentiable (Rademacher's theorem). Then the generalized Clarke Jacobian, $\partial\Phi(z)$, can be defined using the limiting Jacobian by

$$\partial\Phi(z) = \operatorname{conv}\{H \in \mathbb{R}^{n\times n} \mid H = \lim_{k\to+\infty} \nabla\Phi(z_k),\ \text{for } z = \lim_{k\to+\infty} z_k, z_k \notin \mathcal{N}_\Phi\} \tag{12.200}$$

where $\mathcal{N}_\Phi$ is the set (of zero Lebesgue measure) on which the function $\Phi(\cdot)$ is not differentiable.

The standard Newton method is generalized to the nonsmooth case by the following scheme:

$$z_{k+1} = z_k - {H_k}^{-1}\Phi(z_k), \quad H_k \in \partial\Phi(z_k). \tag{12.201}$$

Because the set $\partial\Phi(z_k)$ may not be a singleton (if z_k is a point of discontinuity of $\Phi(\cdot)$), we have to select an arbitrary element for H_k.

Remark 12.55. If the C-function (12.198f) and $F(\cdot)$ are continuously differentiable on the whole space, the mapping $\Phi(\cdot)$ also satisfies this property. In this case, the standard Newton method can be applied directly. Unfortunately, an interesting result of Kanzow & Kleinmichel (1995) shows that the Jacobian of $\Phi(\cdot)$ is singular at any degenerate solution of (12.173), thus preventing fast local convergence.

Choices of the C-Function and the Semismoothness Property

First of all, the choice of the C-function relies on the four major properties:

(a) Suitability of the generalized Jacobian, $\partial\Phi$. Intuitively, it is natural to prefer $\Phi(\cdot)$ with "small" generalized Jacobians. In fact, if H_k in (12.201) is allowed to take very different values, the resulting sequence z_k will likely behave unwieldily. Incidentally, we mention that computing a limiting Jacobian is not a difficult operation: in practice, one just formal differentiation pretending that $\partial\Phi(z_k)$ is a singleton.

(b) Invertibility of the elements of $\partial\Phi$. Clearly from (12.201), H_k must be invertible for each k, and in fact convergence is unlikely if H_k tends to a degenerate matrix. From this point of view, we recall the conclusion of negative Remark 12.55.

(c) Semismoothness of $\Phi(\cdot)$. It is a key property for the convergence of the generalized Newton's method. This is the main reason why the method is often termed in the literature a semismooth Newton's method. Primary results on convergence of semismooth methods can be found in Qi & Sun (1993).

(d) Existence of a differentiable merit function associated with $\Phi(\cdot)$. Eventually, in order to monitor the global convergence of Newton's method, some line search has to be performed along the Newton direction

$$d_k = -{H_k}^{-1}\Phi(x_k) \tag{12.202}$$

to get a sufficient decrease of the associated merit function

$$\Psi(z) = \frac{1}{2}\Phi(z)^{\mathrm{T}}\Phi(z). \tag{12.203}$$

Differentiability of $\Psi(\cdot)$ increases the ability of this procedure to drive z_k toward a zero of $\Phi(\cdot)$.

The Fischer–Burmeister (12.198d) and the penalized Fischer–Burmeister (12.198e) functions enjoy all of these properties if the solution of the NCP is strongly regular (Chen et al., 2000). As far as we know, the choice of the Fischer–Burmeister functions is the best compromise for the numerical efficiency, even on nonmonotone complementarity problems.

Remark 12.56. Pang (1990, 1991) and co-workers (Gabriel & Pang, 1992; Pang & Gabriel, 1993) proposed a generalized Newton method based on the min function in (12.198a). Following the comments in Ferris & Kanzow (2002), a Newton method based on this function is more difficult to globalize.

We will discuss in Sect. 12.5.6 the relative efficiency of various forms of generalized Newton's methods. In particular, most of the variants of the standard Newton's method such as inexact Newton's method, Levenberg–Marquardt method have also been implemented in the semismooth framework (De Luca et al., 2000).

Remark 12.57. The standard LCP and its variants can also be treated by generalized Newton's methods. The computation of the gradients is in this case even simpler.

12.5.5 Interior Point Methods

Interior point methods have also been extended from NLP to NCP. The monotone case is treated in Potra & Ye (1996) by a potential reduction method. The principle is the same as for the LCP. The only difference lies in the fact that, at each iteration, a nonlinear problem has to be solved up to a prescribed tolerance. The way how the tolerance is monitored together with the various intrinsic parameter of the standard interior point method generates a long list of algorithms, in which it is rather difficult to find a way. As we said, the interior point methods have also been extended to nonlinear programs.

12.5.6 Effective Implementations and Comparison of the Numerical Methods for NCPs

In contrast to the interior point methods, it is not difficult to find comparisons of numerical methods based on Newton's method for solving NCPs. In the context of MCP, we refer to the paper of Billups et al. (1997) for an impressive comparison of the following implementation of solvers:

- MILES (Rutherford, 1993) which is an implementation of the classical Newton–Josephy method (see Sect. 12.5.3),
- PATH which has been described at the end of Sect. 12.5.3,
- NE/SQP (Gabriel & Pang, 1992; Pang & Gabriel, 1993) which is a generalized Newton's method based on the minimum function (12.198a); the search direction is computed by solving a convex QP at each iteration,
- QPCOMP (Billups & Ferris, 1995) which is an enhancement of the NE/SQP algorithm to allow iterates to escape from local minima,
- SMOOTH (Chen & Mangasarian, 1996) which is based on solving a sequence of smooth approximations of the NCP,
- PROXI (Billups, 1995) which is a variant of the QPCOMP algorithm using a nonsmooth Newton solver rather than a QP solver,

- SEMISMOOTH (DeLuca et al., 1996) which is an implementation of a semismooth Newton method using the Fischer–Burmeister function,
- SEMICOMP (Billups, 1995) which is an enhancement of SEMISMOOTH based on the same strategy as QPCOMP.

All of these comparisons, which have been made in the framework of the MCP (12.189), show that the PROXI, PATH, and SMOOTH are superior on a large sample of test problems.

For a comparison of the variants of the SEMISMOOTH algorithm, we refer to De Luca et al. (2000).

12.6 Variational and Quasi-Variational Inequalities

12.6.1 Definition and Basic Properties

The VI problem may be defined as follows:

Definition 12.58 (Variational inequality (VI) problem). *Let X be a nonempty subset of $\mathbb{R}^n$ and let F be a mapping from $\mathbb{R}^n$ into itself. The variational inequality problem, denoted by* $\mathrm{VI}(X,F)$*, is to find a vector $z \in \mathbb{R}^n$ such that*

$$F^T(z)(y-z) \geqslant 0, \forall\, y \in X. \tag{12.204}$$

We denote the solution set of (12.204) by Ω.

Usually, the set X is assumed to be closed and convex. The function F is also assumed to be continuous; nevertheless some generalized VI are also defined for set-valued mappings (Harker & Pang, 1990). If X is a closed set and F continuous, the solution set of $\mathrm{VI}(X,F)$ denoted by $\mathrm{SOL}(X,F)$ is always a closed set.

A geometrical interpretation of the $\mathrm{VI}(X,F)$ leads to the equivalent formulation in terms of inclusion into a normal cone of X, i.e.,

$$-F(x) \in N_X(x) \tag{12.205}$$

or equivalently

$$0 \in F(x) + N_X(x). \tag{12.206}$$

This may be deduced directly from the variational definition of a normal cone. It is noteworthy that the $\mathrm{VI}(X,F)$ extends the problem of solving nonlinear equations of the form $F(x) = 0$, taking $X = \mathbb{R}^n$ in (12.206). If $F(z) = Mz + q$ is affine, the $\mathrm{VI}(X,F)$ is called Affine Variational Inequality (AVI) and is denoted by $\mathrm{AVI}(X,q,M)$.

If X is polyhedral, we say that the $\mathrm{VI}(X,F)$ is linearly constrained or is a linearly constrained VI. An important case is the box-constrained VI where the set X is a closed box (possibly unbounded) of $\mathbb{R}^n$, i.e.,

$$K = \{x \in \mathbb{R}^n \mid -\infty \leqslant a_i \leqslant x \leqslant b_i \leqslant +\infty\}. \tag{12.207}$$

Basic Existence and Uniqueness Properties

The basic ingredients for the existence and possibly the uniqueness of solutions of VIs are (a) the degree theory and the fixed-point approaches and (b) the monotonicity property. The degree theory and the fixed-point approaches for VI are well presented in Goeleven et al. (2003a). The monotonicity property is used just as in Theorem 12.44 for NCP. The P-property and its variants cannot be applied to general VI but some notions of F-uniqueness can be introduced. The notion of copositivity can also be a good substitute in the more general framework of VI (see Facchinei & Pang, 2003, for more details).

Quasi-variational Inequalities

We end this section with the definition of a quasi-variational inequality.

Definition 12.59 (Quasi-variational inequality (QVI) problem). *Let X be a multivalued mapping $\mathbb{R}^n \rightsquigarrow \mathbb{R}^n$ and let F be a mapping from $\mathbb{R}^n$ into itself. The quasi-variational inequality problem, denoted by* $\mathrm{QVI}(X,F)$*, is to find a vector $z \in \mathbb{R}^n$ such that*

$$F^T(z)(y-z) \geqslant 0, \forall y \in X(z). \tag{12.208}$$

This problem is a very hard problem from the existence and uniqueness point of view. Unfortunately, the frictional contact problem studied in the following chapter belongs to this class of problems.

For the reader interested in the theory of VI we refer to Goeleven et al. (2003a) and Facchinei & Pang (2003) and the survey paper, Harker & Pang (1990). In these works, some extensions of VIs can also be found where F is multivalued.

12.6.2 Links with the Complementarity Problems

The following complementarity problem over cones can be defined:

Definition 12.60 (Complementarity Problem (CP)). *Given a closed convex cone $K \subset \mathbb{R}^n$ and a mapping $F\colon \mathbb{R}^n \to \mathbb{R}^n$, the complementarity problem, denoted by* $\mathrm{CP}(K,F)$*, is to find a vector $z \in \mathbb{R}^n$ such that*

$$K \ni z \perp F(z) \in K^*, \tag{12.209}$$

where K^* is the dual (negative polar) cone of K defined by

$$K^* = \{d \in \mathbb{R}^n \mid v^T d \geqslant 0, \forall v \in K\}. \tag{12.210}$$

We say that a vector x is feasible to the $\mathrm{CP}(K,F)$ if

$$z \in K \text{ and } F(z) \in K^\star. \tag{12.211}$$

When K is the nonnegative orthant $\mathbb{R}^n_+$, the CP is a NCP. Furthermore, if $F(z) = Mx+q$ is affine, $\mathrm{CP}(\mathbb{R}^n_+,F)$ is $\mathrm{LCP}(M,q)$.

Let $X = K \subset \mathbb{R}^n$ be a cone. A vector x solves the VI(X,F) if and only if x solves the CP(K,F). If K is equal to the nonnegative orthant of $\mathbb{R}^n_+$, a vector x solves the VI(X,F) if and only if x solves the NCP(F).

A box-constrained VI is equivalent to a NCP or a MCP choosing the bounds a_i and b_i in the right way. If, in CP(K,F), K is polyhedral and $F(\cdot)$ is affine, we get an LCP.

An interesting nonpolyhedral example is when

$$K = \{z \in \mathbb{R}^{n+1} \mid z_0 \geqslant \|(z_1,\ldots,z_n)\|\}, \tag{12.212}$$

which is the so-called *second-order cone* or *ice-cream cone*.

12.6.3 Links with the Constrained Minimization Problem

Let us consider for instance the following NLP:

$$\begin{array}{l} \text{minimize } G(z) \\ \text{subject to } z \in K, \end{array} \tag{12.213}$$

where $G(\cdot)$ is supposed to be continuously differentiable. If the set K is convex, any local minimizer $\bar{z}$ of (12.213) must satisfy the following first-order optimality conditions:

$$(y-\bar{z})^{\mathrm{T}}\nabla G(\bar{z}) \geqslant 0, \forall y \in K. \tag{12.214}$$

The latter problem defines clearly VI$(K,\nabla G)$. If the function $G(\cdot)$ is convex, the stationary point which solves VI$(K,\nabla G)$ does solve (12.213) (and is unique if $G(\cdot)$ is strictly convex). Therefore, under the convexity assumptions on $G(\cdot)$ and K, the NLP (12.213) is equivalent to VI$(K,\nabla G)$.

The converse relation between a general VI(K,F) and the NLP can be obtained if the mapping $F(\cdot)$ is a gradient map, that is if a mapping G exists such that $F = \nabla G$. The question relies on the integrability of $F(\cdot)$ which is closely related to the symmetry of the Jacobian matrix ∇F^{T} as the following theorem shows.

Theorem 12.61. *Let $F\colon \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$ be a continuously differentiable mapping on a convex set Ω; then the following statements are equivalent:*

(a) There exists a real-valued function $G(\cdot)$ such that $F(x) = \nabla G^T(x)$ on Ω.
(b) The Jacobian matrix, $\nabla F^T(x)$, is symmetric on Ω.
(c) The integral of $F(\cdot)$ along any closed curve in Ω is zero.

Thanks to this condition, we see how the VI, and therefore its specializations, CP, NCP, LCP, may be related to an optimization problem. We see also in the preceding sections that if the problem is more structured, as can be the case with the specializations of the VIs, it is possible to have deeper equivalences.

12.6.4 Merit and Gap Functions for VI

Merit Function

In the spirit of merit functions for NCP presented in Sect. 12.5.6, the following definition can be stated.

Definition 12.62. *A function $\Psi\colon X \to \mathbb{R}_+$ is called a merit function for the* $\mathrm{VI}(X,F)$ *if it has both the properties:*

(a) $\Psi(z) \geqslant 0$ *for all* $z \in X$.
(b) $\Psi(z) = 0$ *and* $z \in X$ *if and only if* z *solves* $\mathrm{VI}(X,F)$.

If $\Psi(\cdot)$ is a merit function, the following constrained minimization problem can be considered:

$$\begin{array}{l}\text{minimize } \Psi(z)\\ \text{subject to } z \in X.\end{array} \tag{12.215}$$

The set solutions of the $\mathrm{VI}(X,F)$ coincides with the global solution of (12.215) and the optimal value of this problem is zero.

Gap Function

Definition 12.63. *The (primal) gap function, G: $\mathbb{R}^n \to \mathbb{R}_+ \cup \{+\infty\}$, for the* $\mathrm{VI}(X,F)$ *is given by*

$$G(x) = \sup_{y\in X} \quad F^T(z)(x-y). \tag{12.216}$$

This is a nonnegative extended value function which can be possibly infinite. Note that (12.216) is a convex program. We can observe that $\bar{z}$ is a solution of the $\mathrm{VI}(X,F)$ if and only if $\bar{z}$ is the global solution of the so-called gap-constrained minimization problem

$$\begin{array}{l}\text{minimize } G(z)\\ \text{subject to } z \in X\end{array} \tag{12.217}$$

and $G(\bar{z}) = 0$.

The following theorem summarizes the properties of the gap function.

Theorem 12.64. *For any $x \in X$, let $Y(x)$ denote the (possible empty) set of optimal solutions to (12.216). The function $G(\cdot)$ in (12.216) satisfies the following properties:*

1. *$G(\cdot)$ is a merit function for the* $\mathrm{VI}(X,F)$.
2. *$G(\cdot)$ is lower semi-continuous.*
3. *If X is bounded and $F \in \mathscr{C}^1(X)$, then $G(\cdot)$ is Lipschitz continuous on X.*
4. *If $F \in \mathscr{C}^1(X)$, then $G(\cdot)$ is differentiable at $z \in X$ if $Y(z) = y(z)$ ($Y(z)$ is a singleton). We then have*

$$\nabla G(z) = F(z) + \nabla F(z)^T(z - y(z)). \tag{12.218}$$

5. *If $F \in \mathscr{C}^1(X)$ and monotone then if $z \notin \Omega$ and $Y(z) = y(z)$, the direction $d = y(z) - z$ is a feasible direction of descent with respect to $G(\cdot)$ at x. The directional derivative satisfies*
$$G'(z;d) = \nabla G(z)^T d \leqslant -G(z). \tag{12.219}$$
6. *$G(\cdot)$ is convex on X if $F(z)^T z$ is convex and each component of $F(\cdot)$ is concave on X.*
7. *Any solution z of* $\mathrm{VI}(X,F)$ *satisfies the fixed-point problem*
$$z \in \Omega \leftrightarrow z = Y(z). \tag{12.220}$$
8. *Any solution z of* $\mathrm{VI}(X,F)$ *satisfies the stationary point condition under the conditions on $F(\cdot)$ in item 5*
$$z \in \Omega \leftrightarrow G'(z;y-z) \geqslant 0, \quad \textit{for all } y \in X. \tag{12.221}$$

References for the proofs of these properties can be found in Larsson & Patriksson (1994). The properties in items 5 and 7 are the keystones of the descent methods and the projection methods for monotone VIs (see Sect. 12.6.6).

Let $X = K \subset \mathrm{I\!R}^n$ be a cone. The gap function is

$$G(z) = \begin{cases} F(z)^{\mathrm{T}} z & \text{if } F(z) \in K^* \\ +\infty & \text{otherwise.} \end{cases} \tag{12.222}$$

In this case, the gap-constrained minimization problem attains a very simple form summarized in the following theorem:

Theorem 12.65. *Let $X = K \subset \mathrm{I\!R}^n$ be a cone. A vector $\bar{z}$ solves the* $\mathrm{VI}(X,F)$ *or equivalently the* $\mathrm{CP}(K,F)$ *if and only if $\bar{z}$ solves the following constrained minimization problem*

$$\begin{array}{ll} \textit{minimize} & F(z)^T z \\ \textit{subject to} & z \in K \\ & F(z) \in K^* \end{array} \tag{12.223}$$

and the value of the objective function is equal to zero, i.e.,

$$F(\bar{z})^T \bar{z} = 0. \tag{12.224}$$

Note that for $K = \mathrm{I\!R}^n_+$, the problem (12.223) is exactly (12.182).

Remark 12.66. As noted by Facchinei & Pang (2003, p. 89), merit functions can be used in the design of numerical algorithms based on the minimization of the merit function with the hope to obtain one of its global minimum. However, merit functions are typically not convex. Therefore we cannot guarantee to attain their global minima. The stationary point obtained by the minimization of (12.223) is not necessarily a solution of the $\mathrm{VI}(X,F)$. In contrast to the LCP/QP reformulation where sufficient matrices ensure the equivalence (see Sect. 12.4.5), conditions for equivalences are an open issue for VI/CP.

The Generalized and Regularized Gap Function

Auchmuty (1989) develops a class of merit functions for VIs based on the following function, $\widetilde{L}(z,y)$: $X \times X \to \mathbb{R}$, with

$$\widetilde{L}(z,y) = f(z) - f(y) + [F(z) - \nabla f(y)]^{\mathrm{T}}(x-y), \tag{12.225}$$

where f: $\mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is convex, lower semi-continuous, and in $\mathscr{C}^1$ on X. This function is related to the $\mathrm{VI}(X,F)$ thanks to the following saddle-point problem.

Theorem 12.67. *If* $(\bar{z},\bar{y}) \in X \times X$ *satisfies the saddle-point problem*

$$\widetilde{L}(\bar{z},y) \leqslant \widetilde{L}(\bar{z},\bar{y}) \leqslant \widetilde{L}(z,\bar{y}), \quad \text{for all } (z,y) \in X \times X, \tag{12.226}$$

then $\bar{z}$ *solves* $\mathrm{VI}(X,F)$.

This theorem states that saddle points of $\widetilde{L}$ are in the set of solution of $\mathrm{VI}(X,F)$. The saddle–point formulation for the $\mathrm{VI}(X,F)$ allows one to consider two types of optimization problems. The first one is based on the definition of the primal generalized gap function

$$\widetilde{G}(z) = \sup_{y \in X} \quad \widetilde{L}(z,y), \qquad z \in X, \tag{12.227}$$

and the corresponding optimization formulation is

$$\inf_{z \in X} \widetilde{G}(z). \tag{12.228}$$

Note that (12.227) is a convex program. Solving $\mathrm{VI}(X,F)$ is equivalent to minimizing $\widetilde{G}(\cdot)$ over X, so that $\widetilde{G}(\cdot)$ is a merit function. The general properties of the primal generalized gap function (12.227) can be found in Larsson & Patriksson (1994). The second optimization is the dual optimization problem that we will not develop here.

We will focus in the remaining part of this section on a special case of the generalized gap function, the so-called regularized gap function due to Fukushima (1992). The function $f(\cdot)$ is specified as follows:

$$f(z) = \frac{1}{2} z^{\mathrm{T}} Q z, \tag{12.229}$$

where Q is a given symmetric and PD matrix in $\mathbb{R}^{n \times n}$. In this case, the problem (12.227) reduces to find the unique point $y(z)$ such that

$$\begin{array}{ll} \text{minimize} & \|y - (z - Q^{-1}F(z))\|_Q \\ \text{subject to} & y \in X \end{array}, \tag{12.230}$$

that is

$$y(x) = \mathrm{prox}_Q[X; z - Q^{-1}F(z)]. \tag{12.231}$$

The regularized gap function will be denoted by

$$\begin{aligned} G_1(z) &= \sup_{y\in X} F(z)^{\mathrm{T}}(z-y) - \frac{1}{2}(z-y)^{\mathrm{T}}Q(z-y) \\ &= F(z)^{\mathrm{T}}(z-y(z)) - \frac{1}{2}(z-y(z))^{\mathrm{T}}Q(z-y(z)) \end{aligned} \tag{12.232}$$

or more generally by

$$\begin{aligned} G_\alpha(z) &= \sup_{y\in X} F(z)^{\mathrm{T}}(z-y) - \frac{\alpha}{2}(z-y)^{\mathrm{T}}Q(z-y), \quad \alpha > 0 \\ &= F(z)^{\mathrm{T}}(z-y_\alpha(z)) - \frac{\alpha}{2}(z-y_\alpha(z))^{\mathrm{T}}Q(z-y_\alpha(z)), \end{aligned} \tag{12.233}$$

where $y_\alpha(z) = \mathrm{prox}_Q[X; z - \frac{1}{\alpha}Q^{-1}F(z)]$.

The following theorem summarizes several properties of the regularized gap function G_1.

Theorem 12.68. *For any $x \in X$, let $y(x)$ denote the optimal solution to (12.230). The function $G_1(\cdot)$ satisfies the following properties:*

1. *$G_1(\cdot)$ is a merit function for the $\mathrm{VI}(X,F)$.*
2. *If $F(\cdot)$ is continuous, then $G_1(\cdot)$ is also continuous.*
3. *If $F(\cdot)$ is $\mathscr{C}^1$, then $G_1(\cdot)$ is also $\mathscr{C}^1$; moreover,*

$$\nabla G_1(z) = F(z) + (\nabla F(z) - Q)^T(z - y(z)). \tag{12.234}$$

4. *Any solution z of $\mathrm{VI}(X,F)$ satisfies the fixed-point problem*

$$z \in \Omega \Longleftrightarrow z = y(z). \tag{12.235}$$

5. *If $F \in \mathscr{C}^1$ with a positive definite Jacobian ∇F, then*

$$\nabla G_1(z)^T(y-z) \geqslant 0, \quad \textit{for all } y \in X \Rightarrow z \in \Omega. \tag{12.236}$$

6. *If $F \in \mathscr{C}^1$ with a positive definite Jacobian ∇F, then*

$$z \notin \Omega \Rightarrow \nabla G_1(z)^T(y(z)-z) < 0. \tag{12.237}$$

7. *If $F \in \mathscr{C}^1$ with a positive definite Jacobian ∇F and also strongly monotone on X, then*

$$\nabla G_1(z)^T(y(z)-z) \leqslant -m_F\|y(x)-x\|^2, \quad m_F > 0. \tag{12.238}$$

Remark 12.69. We reproduce here a very interesting interpretation due to Larsson & Patriksson (1994) on a possible choice of the function $f(\cdot)$. Since (12.227) is a convex program, evaluating $\widetilde{G}(z)$ is equivalent to finding a solution $y(z)$ to the following VI:

$$\nabla_y \widetilde{L}(z, y(z))^{\mathrm{T}}(y - y(z)) \leqslant 0, \quad \text{for all } y \in X. \tag{12.239}$$

If we assume that $F(\cdot)$ is approximated by the gradient of a convex function $f(\cdot)$, the error made is $F - \nabla f$. For some fixed z, this error may be taken into account by adding to ∇f the error term given by $F(z) - \nabla f(z)$. We obtain then the following approximate symmetric VI:

$$[\nabla f(y(z)) + F(z) - \nabla f(z)]^{\mathrm{T}} (y - y(z)) \leqslant 0, \quad \text{for all } y \in X. \tag{12.240}$$

But from (12.225), it follows that (12.239) and (12.240) are equivalent.

Let us assume for a moment that ∇F is symmetric. One may choose $f(\cdot)$ such that $F = \nabla f$ and (12.240) reduces to VI(X, F). This possibility to approximate $F(\cdot)$ exactly for symmetric VI suggests an analog symmetrization strategy for the asymmetric case. We then define

$$f(z) = \int_{\Gamma} F(s)^{\mathrm{T}} s \ \mathrm{d}s = \int_0^1 \Theta(t)\, \mathrm{d}t, \tag{12.241}$$

where Γ is a curve from 0 to z, and $\Theta(t) = F(tz)^{\mathrm{T}} z$. This approximation is exact if $F(\cdot)$ is the gradient of $f(\cdot)$ and could be a good approximation for modest asymmetric case.

With this interpretation, it seems that the symmetric approximation ∇f of $F(\cdot)$ is a more natural choice than a constant matrix Q as in (12.232).

Finally, we introduce a last version of regularized gap function, the so-called D-gap function. In the spirit of Mangasarian & Solodov (1993), who introduced the notion of implicit Lagrangian for NCPs, Peng (1997) and Yamashita et al. (1997) proposed the D-gap function defined by

$$G_{\alpha,\beta}(z) = G_{\alpha}(z) - G_{\beta}(z), \quad \text{with } \alpha > \beta > 0. \tag{12.242}$$

The main advantage of the D-gap function lies in the fact that it is an unconstrained merit function in the sense that

1. $G_{\alpha,\beta}(z) \geqslant 0$ for all $z \in \mathbb{R}^n$,
2. $G_{\alpha,\beta}(z) = 0$ if and only if z solves VI(X, F),

leading to the following unconstrained minimization problem reformulation of the VI(X, F):

$$\text{minimize} \qquad G_{\alpha,\beta}. \tag{12.243}$$

It was shown that any stationary point $\bar{z}$ of (12.243) with $\nabla F(\bar{z})$ positive definite is a solution of the VI(X, F). If X is a box, then $\nabla F(\bar{z})$ needs only to be a P-matrix. Based on these properties, we will see in Sect. 12.6.6 that unconstrained descent methods were developed to solve the VI(X, F) and convergence to a solution was shown when is $F(\cdot)$ strongly monotone or, if X is a box, when is $F(\cdot)$ a uniform P-function.

12.6.5 Nonsmooth and Generalized equations

In this section, we define nonsmooth equations or generalized equations. Let us start with the definition of a generalized equation given in the pioneering work of Robinson (1979).

Problem 12.70 (Generalized equation (GE) problem). Let $\Omega \subset \mathrm{I\!R}^n$ be an open set. Given a continuously Fréchet differentiable mapping $F\colon \Omega \subset \mathrm{I\!R}^n \to \mathrm{I\!R}^n$ and a maximal monotone operator $T\colon \mathrm{I\!R}^n \rightsquigarrow \mathrm{I\!R}^n$, find a vector $z \in \mathrm{I\!R}^n$ such that

$$0 \in F(z) + T(z). \tag{12.244}$$

We recall that a set-valued operator $T(\cdot)$ is monotone if for each couple $(z,y),(z^\star,y^\star)$ in the graph of $T(\cdot)$, one has

$$(z-z^\star)^{\mathrm{T}}(y-y^\star) \geqslant 0 \tag{12.245}$$

and maximal monotone if its graph is not strictly contained in the graph of any other monotone operator.

The GE problem is closely related to CP problems and to the NLP. For instance, the NCP (12.173) can represented into a GE by

$$0 \in F(z) + N_{\mathrm{I\!R}^n_+}(z) \tag{12.246}$$

and the MCP (12.51), which provides the KKT conditions for the NLP, can be recast into a GE of the form

$$0 \in F(z) + N_K(z),\ z \in \mathrm{I\!R}^{n+m_e+m_i} \tag{12.247}$$

with

$$\begin{cases} F(z) = \begin{bmatrix} \mathscr{L}(z,u,v) \\ -g(z) \\ -h(z) \end{bmatrix} \\ \\ K = \mathrm{I\!R}^n \times \mathrm{I\!R}^{m_i}_+ \times \mathrm{I\!R}^{m_e} \end{cases} . \tag{12.248}$$

General Reformulation of a VI with the Normal and the Natural Map

The following two results extend the reformulation of MCP and NCP as generalized equations.

Proposition 12.71. *Let $X \subset \mathrm{I\!R}^n$ be a closed convex set and a mapping $F\colon X \to \mathrm{I\!R}^n$ and let P_X denote the projection operator onto X. The following statement holds:*

$$x \textit{ solves } \mathrm{VI}(X,F) \Longleftrightarrow \mathbf{F}^{nat}_X(x) = 0, \tag{12.249}$$

where F^{nat}_X is the so-called natural map, defined by

$$\mathbf{F}^{nat}_X(y) = y - P_X(y - F(y)). \tag{12.250}$$

Proposition 12.72. *Let $X \subset \mathrm{I\!R}^n$ be a closed convex set and a mapping $F\colon X \to \mathrm{I\!R}^n$. The following statement holds:*

$$x \textit{ solves } \mathrm{VI}(X,F) \Longleftrightarrow x = P_X(z) \textit{ for some } z \textit{ such that } \mathbf{F}^{nor}(z) = 0, \tag{12.251}$$

where $F^{nor}_X(\cdot)$ is the so-called normal map, defined by

$$\mathbf{F}^{nor}_X(y) = F(P_X(y)) + y - P_X(y). \tag{12.252}$$

The equations $\mathbf{F}_X^{\text{nat}}(x) = 0$ and $\mathbf{F}_X^{\text{nor}}(z) = 0$ allow one to state projection-type algorithm and generalized Newton's method in a very general framework. This is one of the reasons why a lot of effort has been made to characterize these two operators. Furthermore, such formulations lead themselves to a sensitivity analysis to data. For more details, we refer to the recent treatment of Facchinei & Pang (2003).

12.6.6 Main Types of Algorithms for the VI and QVI

As said before, the equation-based reformulations of VIs and the merit function equivalences pave the way to three main types of algorithms:

(a) projection-type and splitting methods,
(b) minimization of merit functions,
(b) generalized Newton Methods,
(c) interior and smoothing methods.

Other classes of methods can also be cited. For instance, the proximal point algorithm (Martinet, 1970; Rockafellar, 1976b) for solving an inclusion of the form

$$T(x) \ni 0, \tag{12.253}$$

where $T(\cdot)$ is a maximal monotone operator, can also be invoked to solve monotone VIs. Interior and smoothing methods have also been developed for solving VIs (Facchinei & Pang, 2003, Chap. 10). When the set X is polyhedral, some simplex-like methods have been implemented for VIs. We will give some insights on these methods in the next paragraphs. We refer to Harker & Pang (1990) and Ferris & Kanzow (2002) for a survey of such methods.

12.6.7 Projection-Type and Splitting Methods

The projection-type methods provide one with a family of methods which are easy to implement and robust if the projection onto the set X is cheap to compute. Such methods do not need the knowledge of the Jacobian of F. Nevertheless, convergence requires a monotonicity-like assumption and the rate of convergence one can expect is linear in most cases

Basic Fixed-Point Scheme

The basic projection-type method performs a fixed-point iteration based on the natural map, namely

$$z_{k+1} = P_X(z_k - F(z_k)), \tag{12.254}$$

where $z_0 \in X$ is a given starting point. In order to improve the convergence of the fixed-point iteration (Brezis & Sibony, 1967/1968; Sibony, 1970) a (small) parameter $\gamma > 0$ is usually introduced as follows:

$$z_{k+1} = P_X(z_k - \gamma F(z_k)). \tag{12.255}$$

Using the Banach fixed-point theorem, this iterative scheme can be proved to be globally convergent provided γ is sufficiently small, and with a linear convergence rate for a strongly monotone and Lipschitz-continuous function F (Auslender, 1976; Bertsekas & Tsitsiklis, 1989). This method has been extended by means of a variable step size in Marcotte & Wu (1995) as follows:

$$z_{k+1} = P_X(z_k - \gamma_k F(z_k)). \tag{12.256}$$

The convergence is then improved to co-coercive functions on X, i.e.,

$$(F(z) - F(y))^{\mathrm{T}}(z-y) \geqslant c\|F(x) - F(y)\|^2, c > 0, \quad \text{for all } (z,y) \in X \times X, \tag{12.257}$$

together with the Lipschitz assumption.

The Extragradient Method

The extragradient method (Korpelevich, 1976) is also a well-known method for VI which improves the previous projection method. It has been extensively studied in Khobotov (1987), Marcotte (1991), and Nagurney (1993) and references therein. It can formulated as

$$z_{k+1} = P_X(z_k - \gamma F(P_X(z_k - \gamma F(z_k)))). \tag{12.258}$$

The convergence of this method requires that the function F is Lipschitz-continuous and pseudo-monotone and that a solution exists. The convergence rate one can expect is also linear. This method has been further extended in Solodov & Tseng (1996) by

$$z_{k+1} = z_k - \tau P^{-1}\left[T_\gamma(z_k) - T_\gamma(P_X(z_k - \gamma F(z_k)))\right], \quad \tau > 0, \tag{12.259}$$

where $P \in \mathbb{R}^{n\times n}$ is a PD matrix and either $T_\gamma = I - \gamma F$, or if F is affine $T_\gamma = I - \gamma M^T$. The parameter γ is chosen such that T_γ is strongly monotone. The convergence result for such a method is equivalent to the extragradient method but it needs only one projection per iteration.

The Hyperplane Projection Method

Besides the monotonicity-like assumption, the drawback of the three previous methods is the requirement of the Lipschitz constant of F to ensure convergence. In Golshtein & Tretyakov (1996), subgradient-like methods have been studied to provide convergence with diminishing but nonsummable sequences of steps. To face this problem, the hyperplane projection method has been introduced by Konnov (1993). The convergence has been proved under the assumptions that F is a continuous pseudo-monotone mapping. The method is described in Algorithm 12.6.7.

Splitting Methods

In the case of the AVI(X,q,M), most of the previous projection methods have been extended by splitting the matrix M as for the LCP case in Tseng (1990, 1995), Marcotte & Wu (1995), and Eckstein & Ferris (1998). The convergence of the schemes has been proved under monotonicity-like assumptions.

Algorithm 16 Hyperplane projection method (Konnov, 1993)

```
Require: $F, X$
Require: $z_0 \in X, \tau > 0, \sigma \in (0,1)$
Ensure: $z$ solution of $\mathrm{VI}(X,F)$ with $F$ a continuous pseudo-monotone mapping
  $k \leftarrow 0$
  while error > tol do
    $y_k \leftarrow P_X(z_k - \tau F(z_k))$
    (Armijo line–search) Find the smallest integer, $i \in \mathbb{N}$ such that

        $F(2^{-i}y_k + (1-2^{-i})z_k)^T (z_k - y_k) \geqslant \frac{\sigma}{\tau}\|z_k - y_k\|^2$        (12.260)

    $i_k \leftarrow i$
    $x^k \leftarrow 2^{-i_k} y_k + (1-2^{-i}) z_k$
    $H_k \leftarrow \{z \in nbR^n \mid F(x_k)^T (z - x_k) = 0\}$
    $w_k \leftarrow P_{H_k}(z_k) = z_k - \frac{F(x_k)^T (z_k - x_k)}{\|F(x_k)\|^2} F(x_k)$
    $z_{k+1} \leftarrow\leftarrow P_X(w_k)$
    $k \leftarrow k+1$
    Evaluate error.
  end while
```

12.6.8 Minimization of Merit Functions

The key idea in this section is to substitute the VI problem by a minimization problem. As we saw in Sect. 12.6.3, a direct reformulation into a NLP is not possible if the Jacobian of F is asymmetric. Nevertheless, based on the notion of gap and merit functions presented in Sect. 12.6.4, minimization reformulations can be proposed.

Using the gap function (12.216), it is possible to design iterative descent methods for solving the problem (12.217). At each iteration, a linear program has to be solved for z^k over the constraint $z^k \in X$. This class of methods has been proved to converge for a compact set X under the assumption that F is monotone (Marcotte, 1985; Marcotte & Dussault, 1987).

There have been many descent methods based on the regularized gap function (12.233) proposed in the literature (Fukushima, 1992; Taji et al., 1993; Zhu & Marcotte, 1993, 1994). All of these methods are proved to be convergent under monotonicity assumption. Their study has been extended to a more general framework in Facchinei & Pang (2003, Sect. 10.4.4). At the iteration k, the descent direction is given by

$$d_k = y_\alpha(z^k) - z^k = \mathrm{prox}_Q\Big[X; z^k - \frac{1}{\alpha}Q^{-1}F(z^k)\Big] - z^k. \tag{12.261}$$

At each step, either a line search is performed along this direction such that

$$z^{k+1} = z^k + \tau^k d^k \tag{12.262}$$

or the regularization parameter α is reduced. Due to the form of the descent direction, these methods can provide a framework for the study of the projection-type methods described previously. The main drawback of this method is that the iterates must remain feasible in solving the constrained minimization problem (12.228). For a complex set X, this minimization can be expensive. The descent methods based on the D-gap function extend the methods based on the regularized gap function providing us with an unconstrained minimization process.

12.6.9 Generalized Newton Methods

Newton's methods for VIs are based on the normal map. As with the NCP, linearizing of the normal map provides a general framework to develop Newton's methods. We will not enter into the details of Newton's method for VI; for more details, we refer to Facchinei & Pang (2003, Chaps. 7 and 8) and to Robinson (1982).

12.6.10 Interest from a Computational Point of View

The VI is a very general framework which encompasses all of the complementarity problems. Therefore, it offers a very interesting basis for theoretical mathematical developments. Unfortunately, the price of this generality is that a lot of the specific structure of the problems is lost when we generalize to VI. From the numerical point of view, the consequence is that the specific structure of problems cannot be exploited and leads to poor qualitative properties of the algorithms or strong assumptions like the monotonicity for convergence. Furthermore, most of the algorithms that are designed for very general VIs with any assumption on the structure of the set X are only conceptual. Indeed, without any structure of the set X, it is very hard to compute efficiently the projection on this set and even more difficult to obtain linear approximation of the normal map. Therefore, in most practical cases, the set X is finitely represented by inequalities. As much as possible, it is more interesting to recast the problem into a complementarity framework.

12.7 Summary of the Main Ideas

- Most of the problems expressed in terms of equalities and inequalities such as LCP, MLCP, MCP, or NCP can be cast into an optimization problem involving a minimization process under various assumptions (symmetry, monotonicity, etc.) through the first-order optimality conditions. Most of the formulations of the time-discretized problems naturally give birth to systems of equalities and inequalities, but it can be interesting from a computational point of view to use the analog optimization problem. Indeed, the minimization of an objective function provides us with some stabilization and globalization results which are very useful in practice. When this is possible we advocate to model the problem into the natural optimization form which involves for instance energetic balances. Furthermore, robust and efficient algorithms for minimization are more prevalent than those for CP solvers.

- Concerning the question of the choice of solvers for a specific problem, some remarks have already been made at the end of several sections. A fundamental specificity of the problems we address is that they are embedded in a dynamic evolution via a time-discretization. Therefore, a good starting guess can be provided to the solver using the information of the preceding time step. Naturally, the nonsmooth character of the evolution may go against this assumption; but even when a nonsmooth event occurs, some information can be retrieved. In this context, solvers that can take advantage of a good starting guess have to be preferred. The remark is closely related to the notion of *warm start* of an algorithm.
- Finally, the efficiency of most algorithms presented in this chapter relies for an important part on the efficiency of the underlying linear algebra solvers. Algorithms able to deeply exploit the structure of the problems have to be preferred.

13

Numerical Methods for the Frictional Contact Problem

13.1 Introduction

The aim of this chapter is to provide some details on how the various Onestep Nonsmooth Problem (OSNSP) which have to be solved at each step of the time-stepping schemes may be solved with the tools presented in Chap. 12. It is noteworthy that even for event-driven algorithms, one has to solve similar problems.

13.2 Summary of the Time-Discretized Equations

In this section we recall the main OSNSPs that have been obtained in Chap. 10. For the sake of readability, we present these problems only in the case of the unilateral contact with Coulomb's friction. The cases of the bilateral constraints and the enhanced nonsmooth laws are omitted. Most of the algorithms in this section can be adapted to the latter cases.

13.2.1 The Index Set of Forecast Active Constraints

The index set I of all unilateral constraints in the system is denoted as in Chap. 8 by

$$I = \{1 \dots \nu\} \subset \mathbb{N}. \tag{13.1}$$

The index set I_a is the set of all *forecast* active constraints of the system and it is denoted by

$$I_a(\tilde{q}_{k+1}) = \{\alpha \in I \mid g^{\alpha}(\tilde{q}_{k+1}) \leqslant 0\} \subseteq I\,, \tag{13.2}$$

where $\tilde{q}_{k+1}$ is as in (10.39). For each index $\alpha \in I_a(\tilde{q}_{k+1})$, we specify the notation introduced in (3.73). The reduced matrices H^a and $\widehat{W}^a$ corresponding to the local unknowns U^a_{k+1} and P^a_{k+1} are such that

$$
\begin{aligned}
U^a_{k+1} &= \left[U^\alpha_{k+1}\right]_{\alpha\in I_a(\tilde{q}_{k+1})}\\
U^a_{\mathrm{N},k+1} &= \left[U^\alpha_{\mathrm{N},k+1}\right]_{\alpha\in I_a(\tilde{q}_{k+1})}\\
U^a_{\mathrm{T},k+1} &= \left[U^\alpha_{\mathrm{T},k+1}\right]_{\alpha\in I_a(\tilde{q}_{k+1})}\\
\\
P^a_{k+1} &= \left[P^\alpha_{k+1}\right]_{\alpha\in I_a(\tilde{q}_{k+1})}\\
P^a_{\mathrm{N},k+1} &= \left[P^\alpha_{\mathrm{N},k+1}\right]_{\alpha\in I_a(\tilde{q}_{k+1})}\\
P^a_{\mathrm{T},k+1} &= \left[P^\alpha_{\mathrm{T},k+1}\right]_{\alpha\in I_a(\tilde{q}_{k+1})}\\
\\
H^a(\tilde{q}_{k+1}) &= [H^\alpha(\tilde{q}_{k+1})]_{\alpha\in I_a(\tilde{q}_{k+1})}\\
\\
H^a_{\mathrm{N}}(\tilde{q}_{k+1}) &= [H^\alpha_{\mathrm{N}}(\tilde{q}_{k+1})]_{\alpha\in I_a(\tilde{q}_{k+1})}\\
\\
H^a_{\mathrm{T}}(\tilde{q}_{k+1}) &= [H^\alpha_{\mathrm{T}}(\tilde{q}_{k+1})]_{\alpha\in I_a(\tilde{q}_{k+1})}\\
\\
p = \textstyle\sum_\alpha p^\alpha &= \textstyle\sum_{\alpha\in I_a(\tilde{q}_{k+1})} H^\alpha(\tilde{q}_{k+1})\,P^\alpha \; = H^a(\tilde{q}_{k+1})\,P\,.
\end{aligned}
\tag{13.3}
$$

It is noteworthy that the last equation implies that $P^\alpha = 0$ for all $\alpha \in I \setminus I_a(\tilde{q}_{k+1})$. The time-discretized contact law with Coulomb's friction, that is,

$$
\begin{cases}
\text{If } \; g^\alpha(\tilde{q}_{k+1}) \leqslant 0 \; \text{ then}\\
\qquad \widehat{U}^\alpha_{k+1} = \left[U^\alpha_{\mathrm{N},k+1} + eU^\alpha_{\mathrm{N},k} + \mu^\alpha\, ||U^\alpha_{\mathrm{T},k+1}||, U^\alpha_{\mathrm{T},k+1}\right]^{\mathrm{T}}\\
\qquad \mathbf{C}^{\alpha,*} \ni \widehat{U}^\alpha_{k+1} \perp P^\alpha_{k+1} \in \mathbf{C}^\alpha\\
\\
\text{If } \; g^\alpha(\tilde{q}_{k+1}) > 0 \; \text{ then } \; P^{\alpha,}_{k+1} = 0
\end{cases}
\tag{13.4}
$$

can therefore be written as

$$
\mathbf{C}^{\alpha,*} \ni \widehat{U}^\alpha \perp P^\alpha_{k+1} \in \mathbf{C}^\alpha, \;\; \forall\, \alpha \in I_a(\tilde{q}_{k+1}) \tag{13.5}
$$

assuming implicitly that $P^\alpha = 0$ for all $\alpha \in I \setminus I_a(\tilde{q}_{k+1})$ and introducing the modified local velocity

$$
\widehat{U}^\alpha_{k+1} = \left[U^\alpha_{\mathrm{N},k+1} + e^\alpha U^\alpha_{\mathrm{N},k} + \mu^\alpha\, ||U^\alpha_{\mathrm{T},k+1}||, U^\alpha_{\mathrm{T},k+1}\right]^{\mathrm{T}}\,. \tag{13.6}
$$

The complementarity problems (13.5) can be gathered for all $\alpha \in I_a(\tilde{q}_{k+1})$ with the following notation

$$
\prod_{\alpha\in I_a(\tilde{q}_{k+1})} \mathbf{C}^{\alpha,*} \ni \widehat{U}^a_{k+1} \perp P^a_{k+1} \in \prod_{\alpha\in I_a(\tilde{q}_{k+1})} \mathbf{C}^\alpha \tag{13.7}
$$

where the variable $\widehat{U}^a_{k+1}$ gathers the value of $\widehat{U}^\alpha_{k+1}$ for all $\alpha \in I_a(\tilde{q}_{k+1})$.

In the same manner, when it is possible to construct a Delassus' operator (in the linear and linearized case), we reduce it on the set of forecast active constraints such that for the linear case

$$\left\{\begin{array}{ll}\widehat{W}^a & = H^{a,\mathrm{T}}\widehat{M}^{-1}H^a \\ \\ \widehat{W}^a_{\mathrm{NT}} & = H^{a,\mathrm{T}}_{\mathrm{N}}\widehat{M}^{-1}H^a_{\mathrm{T}} \\ \\ \widehat{W}^a_{\mathrm{NN}} & = H^{a,\mathrm{T}}_{\mathrm{N}}\widehat{M}^{-1}H^a_{\mathrm{N}} \\ \\ \widehat{W}^a_{\mathrm{TT}} & = H^{a,\mathrm{T}}_{\mathrm{T}}\widehat{M}^{-1}H^a_{\mathrm{T}}\end{array}\right. . \tag{13.8}$$

Remark 13.1. From the computational point of view, the local Delassus's operators

$$\left\{\begin{array}{ll}\widehat{W}^{\alpha\alpha} & = H^{\alpha,\mathrm{T}}\widehat{M}^{-1}H^\alpha \\ \\ \widehat{W}^{\alpha\beta} & = H^{\alpha,\mathrm{T}}\widehat{M}^{-1}H^\beta\end{array}\right. \tag{13.9}$$

are computed in different ways:

- *Linear time-invariant case*. In the linear case, the local Delassus' operators are computed in the initialization phase of the algorithm and before the time-integration loop.
- *Linearized time-invariant case.* The local Delassus' operators $\alpha \in I$ are updated in each Newton's loop or at the beginning of the time step depending on the choice of the prediction $\tilde{q}_{k+1}$ (see (10.39) and below).

The inverse of $\widehat{M}$ is never computed explicitly and therefore never stored except in very special cases (diagonal matrix, small matrix, etc.). In general, we perform a Gaussian elimination with partial pivoting (LU, Cholesky, etc.). The local Delassus' operators are computed by means of a dedicated back substitution. If the solver exploits the block structure of the matrix, the global Delassus' matrix is never assembled. One stores only the list of the local Delassus' operators.

Remark 13.2. The superscript a will be omitted in the sequel to lighten the notation. It will also be assumed that $P^\alpha = 0$ for all $\alpha \in I \setminus I_a(\tilde{q}_{k+1})$.

13.2.2 Summary of the OSNSPs

The Time-Discretized Linear OSNSP $(\mathscr{P}_{\mathrm{L}})$

The time-discretized linear OSNSP, denoted by $(\mathscr{P}_{\mathrm{L}})$ is given by

$$\boxed{(\mathscr{P}_{\mathrm{L}})\left\{\begin{array}{l}U_{k+1} = \widehat{W}P_{k+1} + U_{\mathrm{free}} \\ \\ \left.\begin{array}{l}\widehat{U}^\alpha_{k+1} = \left[U^\alpha_{\mathrm{N},k+1} + e^\alpha U^\alpha_{\mathrm{N},k} + \mu^\alpha\,||U^\alpha_{\mathrm{T},k+1}||, U^\alpha_{\mathrm{T},k+1}\right]^{\mathrm{T}} \\ \\ \mathbf{C}^{\alpha,*} \ni \widehat{U}^\alpha_{k+1} \perp P^\alpha_{k+1} \in \mathbf{C}^\alpha\end{array}\right\}\forall\alpha \in I_a(\tilde{q}_{k+1})\end{array}\right.}$$

Some further comments can be made from a numerical point of view:

- The Delassus' matrix $\widehat{W}$ is assumed to be at least a PSD matrix. This assumption is reasonable if we assume that the iteration matrix $\widehat{M}$ is *PD*. The definiteness can be lost if the matrix H has not full rank, which is common in practical large-scale applications.
- The problem $(\mathscr{P}_L)$ can take into account the bilateral constraints if these constraints are condensed when reducing the problem to local coordinates.
- The time-discretized linearized OSNSP, denoted by $(\mathscr{P}_{\mathrm{L}\tau})$ in Chap. 10, has exactly the same structure than $(\mathscr{P}_{\mathrm{L}})$. Therefore, all solvers for $(\mathscr{P}_{\mathrm{L}})$ process also the $(\mathscr{P}_{\mathrm{L}\tau})$. We will not make a particular presentation for the linearized case.

The Time-Discretized Mixed Linear OSNSP $(\mathscr{P}_{\mathrm{ML}})$

The time-discretized mixed linear OSNSP $(\mathscr{P}_{\mathrm{ML}})$ is given by

$$\boxed{(\mathscr{P}_{\mathrm{ML}})\begin{cases} \widehat{M}(v_{k+1}-v_{\mathrm{free}}) = p_{k+1} = \displaystyle\sum_{\alpha\in I_a(\tilde{q}_{k+1})} p^{\alpha}_{k+1} \\[2ex] U^{\alpha}_{k+1} = H^{\alpha,T}\, v_{k+1} \\[2ex] p^{\alpha}_{k+1} = H^{\alpha}\, P^{\alpha}_{k+1} \\[2ex] \left.\begin{array}{l} \widehat{U}^{\alpha}_{k+1} = \left[U^{\alpha}_{\mathrm{N},k+1} + e^{\alpha}U^{\alpha}_{\mathrm{N},k} + \mu^{\alpha}\,||U^{\alpha}_{\mathrm{T},k+1}||, U^{\alpha}_{\mathrm{T},k+1}\right]^{\mathrm{T}} \\ \mathbf{C}^{\alpha,*} \ni \widehat{U}^{\alpha}_{k+1} \perp P^{\alpha}_{k+1} \in \mathbf{C}^{\alpha} \end{array}\right\} \forall\alpha\in I_a(\tilde{q}_{k+1}) \end{cases}}$$

The Time-Discretized Mixed Nonlinear OSNSP $(\mathscr{P}_{\mathrm{MNL}})$

The time-discretized mixed nonlinear OSNSP denoted by $(\mathscr{P}_{\mathrm{MNL}})$ is obtained

$$\boxed{(\mathscr{P}_{\mathrm{MNL}})\begin{cases} \mathscr{R}(v_{k+1}) = p_{k+1} = \displaystyle\sum_{\alpha\in I_a(\tilde{q}_{k+1})} p^{\alpha}_{k+1} \\[2ex] U^{\alpha}_{k+1} = H^{\alpha,T}(q_k+1)\, v_{k+1} \\[2ex] p^{\alpha}_{k+1} = H^{\alpha}(q_k+1)\, P^{\alpha}_{k+1} \\[2ex] \left.\begin{array}{l} \widehat{U}^{\alpha}_{k+1} = \left[U^{\alpha}_{\mathrm{N},k+1} + e^{\alpha}U^{\alpha}_{\mathrm{N},k} + \mu^{\alpha}\,||U^{\alpha}_{\mathrm{T},k+1}||, U^{\alpha}_{\mathrm{T},k+1}\right]^{\mathrm{T}} \\[2ex] \mathbf{C}^{\alpha,*} \ni \widehat{U}^{\alpha}_{k+1} \perp P^{\alpha}_{k+1} \in \mathbf{C}^{\alpha} \end{array}\right\} \forall\alpha\in I_a(\tilde{q}_{k+1}) \end{cases}}$$

13.3 Formulations and Resolutions in LCP Forms

13.3.1 The Frictionless Case with Newton's Impact Law

LCP Formulation

Remind that the frictionless contact problem in the form $(\mathscr{P}_{\mathrm{L}})$ can be written as $(\mathscr{P}_{\mathrm{LWF}})$:

$$\boxed{(\mathscr{P}_{\mathrm{LWF}}) \begin{cases} U_{\mathrm{N},k+1} = \widehat{W}_{\mathrm{NN}} P_{\mathrm{N},k+1} + U_{\mathrm{N,free}} \\ \left.\begin{aligned} &\widehat{U}^{\alpha}_{\mathrm{N},k+1} = U^{\alpha}_{\mathrm{N},k+1} + e^{\alpha} U^{\alpha}_{\mathrm{N},k} \\ &0 \leqslant \widehat{U}^{\alpha}_{\mathrm{N},k+1} \perp P^{\alpha}_{\mathrm{N},k+1} \geqslant 0 \end{aligned}\right\} \forall \alpha \in I_a(\tilde{q}_{k+1}) \end{cases}}$$

The formulation in terms of LCP (12.66) is straightforward,

$$\begin{cases} U_{\mathrm{N},k+1} = \widehat{W}_{\mathrm{NN}} P_{\mathrm{N},k+1} + U_{\mathrm{N,free}} \\ \widehat{U}_{\mathrm{N},k+1} = U_{\mathrm{N},k+1} + e \circ U_{\mathrm{N},k} \\ 0 \leqslant P_{\mathrm{N},k+1} \perp \widehat{U}_{\mathrm{N},k+1} \geqslant 0 \end{cases} \tag{13.10}$$

where the vector e collects the coefficients of restitution for $\alpha \in I_a(\tilde{q}_{k+1})$, and $x \circ y$ is the Hadamard product of the vectors x and y.

To obtain a proper LCP formulation it suffices to write

$$\begin{cases} \widehat{U}_{\mathrm{N},k+1} = \widehat{W}_{\mathrm{NN}} P_{\mathrm{N},k+1} + U_{\mathrm{N,free}} - e \circ U_{\mathrm{N},k} \\ 0 \leqslant \widehat{U}_{\mathrm{N},k+1} \perp P_{\mathrm{N},k+1} \geqslant 0 \end{cases} \tag{13.11}$$

and we can conclude that $(\widehat{U}_{\mathrm{N},k+1}, P_{\mathrm{N},k+1})$ solves the following LCP

$$\mathrm{LCP}(\widehat{W}_{\mathrm{NN}}, U_{\mathrm{N,free}} - e \circ U_{\mathrm{N},k}) \,. \tag{13.12}$$

LCP Resolution

Almost all the methods exposed in the Sect. 12.4 can be applied to solve (13.12). The main reason is that the matrix $\widehat{W}_{\mathrm{NN}}$ is a symmetric PSD matrix, provided that $\widehat{M}$ is PD. The fact that $\widehat{W}_{\mathrm{NN}}$ is PSD and not necessarily PD can cause troubles in numerical applications. This is due to the rank deficiency of H and can be interpreted in terms of redundant constraints. In practice, it may happen that the splitting-based algorithms have difficulties to converge.

Fortunately, the local velocities are uniquely defined. Only the multiplier $P_{\mathrm{N},k+1}$ is not unique if the matrix H is rank deficient. This point can easily be proved by considering the equivalent QP formulation (see Theorem 12.33). One way to circumvent this problem is to use numerical algorithms that look at priority for the local velocities, $U_{\mathrm{N},k+1}$ or the robust QP solvers presented in Sect. 12.2.

The question of the existence and uniqueness of solutions for the frictionless problem and the ability to compute a solution by a numerical treatment are discussed in Lötstedt (1982) and Baraff (1993). The context of these works is quite different in the sense that Lötstedt (1982) only studied smooth motion with unilateral constraints and Baraff (1993) studied the problem with an event-driven strategy. Nevertheless, most of the conclusions on the existence and uniqueness of solutions may be applied to the other frictionless problems.

13.3.2 The Frictionless Case with Newton's Impact and Linear Perfect Bilateral Constraints

Remind that the mixed linear OSNSP with linear perfect bilateral constraints $G^{\mathrm{T}}q+b=0$ is given by

$$\boxed{(\mathscr{P}_{\mathrm{ML}_b})\begin{cases}\widehat{M}(v_{k+1}-v_{\mathrm{free}})=p_{k+1}+GP_{\mu,k+1}\\[2ex] G^{\mathrm{T}}v_{k+1}=0\\[2ex] U^{\alpha}_{\mathrm{N},k+1}=H^{\alpha,\mathrm{T}}_{\mathrm{N}}\,v_{k+1}\\[2ex] p^{\alpha}_{k+1}=H^{\alpha}_{\mathrm{N}}\,P^{\alpha}_{\mathrm{N},k+1}\\[2ex] \left.\begin{array}{l}\widehat{U}^{\alpha}_{\mathrm{N},k+1}=U^{\alpha}_{\mathrm{N},k+1}+e^{\alpha}U^{\alpha}_{\mathrm{N},k}\\[2ex] 0\leqslant\widehat{U}^{\alpha}_{\mathrm{N},k+1}\perp P^{\alpha}_{\mathrm{N},k+1}\geqslant 0\end{array}\right\}\forall\alpha\in I_a(\tilde{q}_{k+1})\end{cases}}$$

This problem $(\mathscr{P}_{\mathrm{ML}_b})$ can be cast directly into the form of an MLCP in (12.78), but we prefer to substitute the generalized velocities $\dot{q}_{k+1}=v_{k+1}$ thanks to

$$v_{k+1}=v_{\mathrm{free}}+\widehat{M}^{-1}\left(GP_{\mu,k+1}+H_{\mathrm{N}}P_{\mathrm{N},k+1}\right)=0 \tag{13.13}$$

in order to obtain the following MLCP in the form (12.79),

$$\begin{cases} G^{\mathrm{T}} v_{\mathrm{free}} + G^{\mathrm{T}} \widehat{M}^{-1} G P_{\mu,k+1} + G^{\mathrm{T}} \widehat{M}^{-1} H_{\mathrm{N}} P_{k+1} = 0 \\ \widehat{U}^{\alpha}_{\mathrm{N},k+1} = \left[H_{\mathrm{N}}^{\mathrm{T}} v_{\mathrm{free}} + e \circ U_{\mathrm{N},k} + G^{\mathrm{T}} \widehat{M}^{-1} G P_{\mu,k+1} + G^{\mathrm{T}} \widehat{M}^{-1} H_{\mathrm{N}} P_{k+1} \right] \\ 0 \leqslant \widehat{U}^{\alpha}_{\mathrm{N},k+1} \perp P^{\alpha}_{\mathrm{N},k+1} \geqslant 0 \,. \end{cases} \tag{13.14}$$

MLCP Resolution

As we said in Sect. 12.4, the multiplier μ may be statically solved provided that the Schur complement matrix $G^{\mathrm{T}} \widehat{M}^{-1} G$ is nonsingular. In this case, we obtain an LCP in a standard form. This operation is expensive from a computational point of view and is usually not performed, except in the special case of “simple” bilateral constraints such as bound constraints corresponding to the boundary conditions. Indeed, the structure of the matrix G^{T} for bound constraints is very simple and it is then possible to work directly on the matrix $\widehat{M}$ before the Gaussian elimination to take into account such constraints. Bound constraints arise naturally when we deal with prescribed boundary conditions on a multibody system or continuum media discretized by a finite element method.

The MLCP (13.14) can also solved by any of the MLCP solvers detailed in Sect. 12.4. The class of iterative algorithms (projection/splitting, interior point methods) have to be favored because it is easier to exploit the block structure of the MLCP. It should be possible to show that the MLCP is monotone under the assumption that $\widehat{M}$ is a PD matrix, which is a usual assumption for the iteration matrix.

13.3.3 Two-Dimensional Frictional Case as an LCP

As Chap. 1 mentioned, it is possible to write the two-dimensional frictional unilateral contact problem as an LCP in standard form. We refer to the works of Pfeiffer & Glocker (1996) and Glocker (1999) for a detailed presentation of the method.

Unfortunately, the LCP matrix is no longer a symmetric PD matrix, not even a P-matrix. Nevertheless, the pivoting methods such as the Lemke’s method can process the LCP. This is a byproduct of the study of Stewart & Trinkle (1996) and Anitescu & Potra (1997) in the more general case of the faceting of the second-order Coulomb’s cone (see Sects. 13.3.4 and 13.3.5). The drawback of the pivoting methods are that they are very expensive on large-scale applications. In “gentle” cases of large applications, we can try to solve this LCP with an iterative scheme normally well suited for symmetric PSD matrix. The projection/splitting methods exposed at the end of Sect. 12.4.6 for asymmetric row-sufficient matrices would be an issue.

The question of existence and uniqueness of solutions to the resulting LCP was studied by Lötstedt (1981) and Baraff (1993) in the context of an event-driven approach or an analytical study at events. In this context, the existence of solutions is not guaranteed (see for instance the Painlevé example in Sect. 6.2). Therefore, the numerical solvers such as Lemke’s algorithm can undergo a lot of difficulties. The question of finite termination of Lemke’s algorithm is addressed in Baraff (1993).

One of the other important discrepancy between the event-driven algorithms and the time-stepping methods lies in the existence of solutions of the OSNSP. This fact is of utmost importance for the numerical robustness of the simulations.

More generally, the two-dimensional Coulomb's law can be viewed as a scalar piecewise linear multivalued function as the relay characteristics. With this insight, we can use results on the equivalence between special kinds of piecewise linear multi-valued functions and LCP (see Facchinei & Pang, 2003, p. 213). These works would allow to express more complicated piecewise linear model of friction in terms of LCPs.

13.3.4 Outer Faceting of the Coulomb's Cone

In this section, we drop for a moment the subscripts k and $k+1$ and the superscript α to lighten the notation. We recall that α denotes the indexes of the contacts in the set $I_a(\tilde{q}_{k+1})$ in (13.2).

Contrary to the two-dimensional frictional contact problem, the three-dimensional case cannot be directly cast into an LCP standard form. This is mainly due to the second-order cone $\mathbf{C}$ which cannot be written as a polyhedral cone. The nonlinear nature of the section of the friction cone, i.e., the disk $\mathbf{D}(\mu R_{\mathrm{N}})$ defined by

$$\mathbf{D}(\mu R_{\mathrm{N}}) = \{R_{\mathrm{T}} \mid \sigma(R_{\mathrm{T}}) = \mu R_{\mathrm{N}} - \|R_{\mathrm{T}}\| \geqslant 0\} \tag{13.15}$$

adds new difficulties from the formulation point of view.

To overcome this difficulty, some approximations have been proposed which consist in faceting $\mathbf{C}$. The following presentation is partly inspired from Glocker (1999) where a very clear and concise presentation can be found.

Following the work in Klarbring (1986a) and Klarbring & Björkman (1988), the friction disk $\mathbf{D}$ can be approximated by an outer polygon:

$$\mathbf{D}_{outer}(\mu R_{\mathrm{N}}) = \bigcap_{i=1}^{\omega} \mathbf{D}_i(\mu R_{\mathrm{N}}) \quad \text{with } \mathbf{D}_i(\mu R_{\mathrm{N}}) = \left\{R_{\mathrm{T}}, \sigma_i(R_{\mathrm{T}}) = \mu R_{\mathrm{N}} - c_i^{\mathrm{T}} R_{\mathrm{T}} \geqslant 0\right\} . \tag{13.16}$$

Where $\omega \in \mathbb{N}$ is the number of Facets. The functions $\sigma_i(R_{\mathrm{T}})$ are the friction saturation with respect to the cone $\mathbf{D}_i(\mu R_{\mathrm{N}})$ generated by an outward unit vector $c_i \in \mathbb{R}^2$ (e.g., Fig. 13.1a)). We now assume that the contact law (3.149) is of the form

$$-U_{\mathrm{T}} \in N_{\mathbf{D}_{outer}(\mu R_{\mathrm{N}})}(R_{\mathrm{T}}) . \tag{13.17}$$

From Rockafellar (1970), the normal cone to $\mathbf{D}_{outer}(\mu R_{\mathrm{N}})$ is given by

$$N_{\mathbf{D}_{outer}(\mu R_{\mathrm{N}})}(R_{\mathrm{T}}) = \Sigma_{i=1}^{\omega} N_{\mathbf{D}_i(\mu R_{\mathrm{N}})}(R_{\mathrm{T}}) \tag{13.18}$$

and the inclusion can be stated as:

$$-U_{\mathrm{T}} = \Sigma_{i=1}^{\omega} \kappa_i \partial \sigma_i(R_{\mathrm{T}}), \quad 0 \leqslant \sigma_i(R_{\mathrm{T}}) \perp \kappa_i \geqslant 0 . \tag{13.19}$$

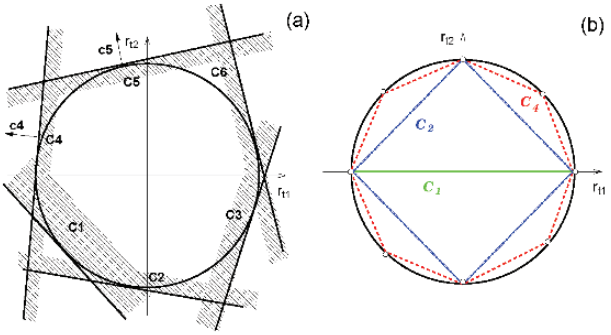


Fig. 13.1. Approximation of the base of the Coulomb cone by an outer approximation (**a**) and by an interior 2ω-gon (**b**)

Since $\sigma_i(R_{\mathrm{T}})$ is linear with respect to R_{T}, we obtain the following MLCP:

$$-U_{\mathrm{T}} = \Sigma_{i=1}^{\omega} \kappa_i c_i, \quad 0 \leqslant \sigma_i(R_{\mathrm{T}}) \perp \kappa_i \geqslant 0\,. \tag{13.20}$$

Assuming for the sake of simplicity that the vectors c_i are chosen equal for all contacts α, the time-discretized linear OSNSP, $(\mathscr{P}_{\mathrm{L}})$, can be written as

$$(\mathscr{P}_{\mathrm{L}}) \begin{cases} U_{k+1} = \widehat{W} P_{k+1} + U_{\mathrm{free}} \\ \left.\begin{array}{l} -U^{\alpha}_{\mathrm{T},k+1} = \Sigma_{i=1}^{\omega} \kappa_i^{\alpha} c_i \\ \sigma_i(P^{\alpha}_{\mathrm{T},k+1}) = \mu P^{\alpha}_{\mathrm{N},k+1} - c_i^{\mathrm{T}} P^{\alpha}_{\mathrm{T},k+1} \\ 0 \leqslant U^{\alpha}_{\mathrm{N},k+1} + e^{\alpha} U^{\alpha}_{\mathrm{N},k} \perp P^{\alpha}_{\mathrm{N},k+1} \geqslant 0 \\ 0 \leqslant \sigma_i^{\alpha}(P^{\alpha}_{\mathrm{T},k+1}) \perp \kappa_i^{\alpha} \geqslant 0 \end{array}\right\} \forall \alpha \in I_a(\tilde{q}_{k+1})\,. \end{cases} \tag{13.21}$$

The fact that the friction saturation functions $\sigma_i(P_{\mathrm{T},k+1})$ are linear shows that the previous problem (13.21) is an MLCP.

Remark 13.3. The number of linear constraints should be at least chosen such that $\omega > 1$. Indeed for $\omega \leqslant 1$, the pyramidal friction cone is not pointed and some nonzero values of R_{T} are allowed for $R_{\mathrm{N}} < 0$.

The LCP in a Single-Contact Case

Generally, the MLCP (13.21) can be reduced into an LCP in standard form assuming that at least one pair of vectors c_i is linearly independent. This process is analogous to the process described in Chap. 1 for the Zener diode, where we have succeeded in transforming the MLCS (1.71) in the LCS (1.73).

As we said earlier, the most simple way to transform an MLCP into an LCP is to compute a Schur complement of the MLCP matrix, which necessitates to invert a submatrix. To be able to invert a submatrix of the MLCP (13.21), we assume that a pair of c_i^α vectors is linearly independent for $i \in \mathscr{P}^\alpha \subset \{1\dots\omega^\alpha\}$, where it is recalled that ω^α is the number of facets of the approximation of the cone at the contact α. Following Glocker (1999), we introduce the following notation,

$$\begin{aligned}
&\mathscr{R} = \{1\dots\omega\} \setminus \mathscr{P}^\alpha \\
&I_{\mathscr{P}^\alpha} = \left[\, c_i^\alpha \,\right]_{\mathscr{P}^\alpha} \\
&I_{\mathscr{R}^\alpha} = \left[\, c_i^\alpha \,\right]_{\mathscr{R}^\alpha} \ .
\end{aligned} \tag{13.22}$$

Thanks to this notation, we may write

$$\sigma_i^\alpha(\lambda_{\mathrm{T}}^\alpha) = \mu^\alpha R_{\mathrm{N}}^\alpha - c_i^{\alpha,\mathrm{T}}\,\lambda_{\mathrm{T}}^\alpha, \ \ \forall i \in \{1\dots\omega\} \tag{13.23}$$

as

$$\begin{aligned}
\sigma_{\mathscr{P}^\alpha}(\lambda_{\mathrm{T}}^\alpha) &= \mu_{\mathscr{P}^\alpha} R_{\mathrm{N}}^\alpha - I_{\mathscr{P}^\alpha}^{\mathrm{T}}\,\lambda_{\mathrm{T}}^\alpha \\
\sigma_{\mathscr{R}^\alpha}(\lambda_{\mathrm{T}}^\alpha) &= \mu_{\mathscr{R}^\alpha} R_{\mathrm{N}}^\alpha - I_{\mathscr{R}^\alpha}^{\mathrm{T}}\,\lambda_{\mathrm{T}}^\alpha
\end{aligned} \tag{13.24}$$

where the vector $\mu_{\mathscr{P}^\alpha}$ and $\mu_{\mathscr{R}^\alpha}$ are defined by

$$\mu_{\mathscr{P}^\alpha} = \begin{bmatrix} \mu^\alpha \\ \mu^\alpha \end{bmatrix} \in \mathrm{I\!R}^2, \quad \mu_{\mathscr{R}^\alpha} = \begin{bmatrix} \mu^\alpha \\ \vdots \\ \mu^\alpha \end{bmatrix} \in \mathrm{I\!R}^{\omega^\alpha - 2} \ . \tag{13.25}$$

Since $I_{\mathscr{P}^\alpha}$ is assumed to be invertible, one obtains

$$\lambda_{\mathrm{T}}^\alpha = I_{\mathscr{P}^\alpha}^{-\mathrm{T}} \mu_{\mathscr{P}^\alpha} R_{\mathrm{N}} - I_{\mathscr{P}^\alpha}^{-\mathrm{T}} \sigma_{\mathscr{P}^\alpha} \tag{13.26}$$

and then by substitution,

$$\sigma_{\mathscr{R}^\alpha}(R_{\mathrm{T}}) = \mu_{\mathscr{R}^\alpha} R_{\mathrm{N}}^\alpha - I_{\mathscr{R}^\alpha}^{\mathrm{T}} I_{\mathscr{P}^\alpha}^{-\mathrm{T}} \mu_{\mathscr{P}^\alpha} r_{\mathrm{N}}^\alpha + I_{\mathscr{R}^\alpha}^{\mathrm{T}} I_{\mathscr{P}^\alpha}^{-\mathrm{T}} \sigma_{\mathscr{P}^\alpha} \ . \tag{13.27}$$

In the same manner, the equation

$$-u_{\mathrm{T}}^\alpha = \Sigma_{i=1}^{\omega^\alpha} \kappa_i^\alpha c_i^\alpha = I_{\mathscr{P}^\alpha} \kappa_{\mathscr{P}^\alpha} + I_{\mathscr{R}^\alpha} \kappa_{\mathscr{R}^\alpha} \tag{13.28}$$

can be written as

$$\kappa_{\mathscr{P}^\alpha} = -I_{\mathscr{P}^\alpha}^{-1} U_{\mathrm{T}} - I_{\mathscr{P}^\alpha}^{-1} I_{\mathscr{R}^\alpha} \kappa_{\mathscr{R}^\alpha} \ . \tag{13.29}$$

We drop the superscript α to lighten the notation. Substituting the value of $P_{\mathrm{T},k+1}$ given by the discrete analog to (13.26) into the first equation of (13.21) and substituting the velocity $U_{\mathrm{T},k+1}$ into the discrete analog to (13.29) one obtains the following LCP in standard form:

$$\left\{\begin{array}{l} \begin{bmatrix} U_{\mathrm{N},k+1}+eU_{\mathrm{N},k} \\ \kappa_{\mathscr{P}} \\ \sigma_{\mathscr{R}} \end{bmatrix} = M \begin{bmatrix} P_{\mathrm{N},k+1} \\ \sigma_{\mathscr{P}} \\ \kappa_{\mathscr{R}} \end{bmatrix} + q \\ \\ 0 \leqslant \begin{bmatrix} U_{\mathrm{N},k+1}+eU_{\mathrm{N},k} \\ \kappa_{\mathscr{P}} \\ \sigma_{\mathscr{R}} \end{bmatrix} \perp \begin{bmatrix} P_{\mathrm{N},k+1} \\ \sigma_{\mathscr{P}} \\ \kappa_{\mathscr{R}} \end{bmatrix} \geqslant 0 \end{array}\right. \tag{13.30}$$

where

$$M = \begin{bmatrix} \widehat{W}_{\mathrm{NN}} + \widehat{W}_{\mathrm{NT}} I_{\mathscr{P}}^{-\mathrm{T}} \mu_{\mathscr{P}} & -\widehat{W}_{\mathrm{NT}} I_{\mathscr{P}}^{-\mathrm{T}} & 0 \\ -I_{\mathscr{P}}^{-1}[\widehat{W}_{\mathrm{TN}} + \widehat{W}_{\mathrm{TT}} I_{\mathscr{P}}^{-\mathrm{T}} \mu_{\mathscr{P}}] & I_{\mathscr{P}}^{-1} \widehat{W}_{\mathrm{TT}} I_{\mathscr{P}}^{-\mathrm{T}} & -I_{\mathscr{P}}^{-1} I_{\mathscr{R}} \\ \mu_{\mathscr{R}} - I_{\mathscr{R}}^{-\mathrm{T}} I_{\mathscr{P}}^{-\mathrm{T}} \mu_{\mathscr{P}} & I_{\mathscr{R}}^{-\mathrm{T}} I_{\mathscr{P}}^{-\mathrm{T}} & 0 \end{bmatrix} \tag{13.31}$$

and

$$q = \begin{bmatrix} U_{\mathrm{N,free}} + eU_{\mathrm{N,free}} \\ -I_{\mathscr{P}}^{-1} U_{\mathrm{T,free}} \\ 0 \end{bmatrix} . \tag{13.32}$$

Example 13.4 ($\mathbf{D}(\mu R_N)$ is approximated by a square). The example when $\mathbf{D}(\mu R_{\mathrm{N}})$ is approximated by a square with edges of length $2\mu_0$ is given by Glocker (1999). The directions c_i can be taken as

$$c_1 = [1 \quad 0]^{\mathrm{T}}, \quad c_2 = [0 \quad 1]^{\mathrm{T}}, \quad c_3 = [-1 \quad 0]^{\mathrm{T}}, \quad c_4 = [0 \quad -1]^{\mathrm{T}} . \tag{13.33}$$

Choosing $\mathscr{P} = \{1,2\}$ and then $\mathscr{R} = \{3,4\}$, one gets

$$I_{\mathscr{P}} = -I_{\mathscr{R}} = I_2 , \tag{13.34}$$

where I_2 is the identity matrix of $\mathrm{I\!R}^{2\times 2}$. The LCP matrix in (13.31) becomes

$$M = \begin{bmatrix} \widehat{W}_{\mathrm{NN}} + \widehat{W}_{\mathrm{NT}} \mu_{\mathscr{P}} & -\widehat{W}_{\mathrm{NT}} & 0 \\ -\widehat{W}_{\mathrm{TN}} - \widehat{W}_{\mathrm{TT}} \mu_{\mathscr{P}} & \widehat{W}_{\mathrm{TT}} & I_2 \\ 2\mu_{\mathscr{P}} & -I_2 & 0 \end{bmatrix} . \tag{13.35}$$

The LCP in the Multi-contact Case

In the multi-contact case, the matrix notation must be enlarged to extend the formulation (13.30) and (13.31). Let us first introduce the index sets

$$\mathscr{P} = \{\mathscr{P}^{\alpha} \mid \alpha \in I_a(\tilde{q}_{k+1})\}, \quad \mathscr{R} = \{\mathscr{R}^{\alpha} \mid \alpha \in I_a(\tilde{q}_{k+1})\} . \tag{13.36}$$

In order to perform this extension, we introduce the following notation:

$$\mu_{\mathscr{P}} = \begin{bmatrix} \mu_{\mathscr{P}^1} & & & & \\ & \ddots & & (0) & \\ & & \mu_{\mathscr{P}^\alpha} & & \\ & (0) & & \ddots & \\ & & & & \mu_{\mathscr{P}^\nu} \end{bmatrix}, \quad \mu_{\mathscr{R}} = \begin{bmatrix} \mu_{\mathscr{R}^1} & & & & \\ & \ddots & & (0) & \\ & & \mu_{\mathscr{R}^\alpha} & & \\ & (0) & & \ddots & \\ & & & & \mu_{\mathscr{R}^\nu} \end{bmatrix} \tag{13.37}$$

for $\mu_{\mathscr{P}} \in \mathrm{I\!R}^{2a\times a}$ and $\mu_{\mathscr{R}} \in \mathrm{I\!R}^{(\Sigma_\alpha(\omega^\alpha-2)a)\times 2a}$ where $a \leqslant \nu$ is the cardinal of $I_a(\tilde{q}_{k+1})$. Finally, we define

$$\mathrm{I}_{\mathscr{P}} = \begin{bmatrix} I_{\mathscr{P}^1} & & & & \\ & \ddots & & (0) & \\ & & I_{\mathscr{P}^\alpha} & & \\ & (0) & & \ddots & \\ & & & & I_{\mathscr{P}^\nu} \end{bmatrix}, \quad I_{\mathscr{R}} = \begin{bmatrix} I_{\mathscr{R}^1} & & & & \\ & \ddots & & (0) & \\ & & I_{\mathscr{R}^\alpha} & & \\ & (0) & & \ddots & \\ & & & & I_{\mathscr{R}^\nu} \end{bmatrix} \tag{13.38}$$

for $I_{\mathscr{P}} \in \mathrm{I\!R}^{2a\times 2a}$ and $I_{\mathscr{R}} \in \mathrm{I\!R}^{(\Sigma_\alpha(\omega^\alpha-2)a)\times 2a}$.

With the notation (13.36), (13.37), and (13.38), the LCP given by (13.30), (13.31), and (13.32) is valid for the multicontact case. As we said earlier, if the solver exploits the block structure of the matrix, the global Delassus' matrix is never assembled. We store only the list of the local Delassus' operators.

Example 13.5 ($\mathbf{D}(\mu R_{\mathrm{N}})$ is approximated by a square (continued)). The example when $\mathbf{D}(\mu R_{\mathrm{N}})$ is approximated by a square with edges of length $2\mu_0$ can be extended to the multi-contact case. Choosing $\mathscr{P}^\alpha = \{1,2\}$ and then $\mathscr{R}^\alpha = \{3,4\}$ for each contact α, one gets

$$I_{\mathscr{P}} = -I_{\mathscr{R}} = I_{2a}\,, \tag{13.39}$$

where I_{2a} is the identity matrix of $\mathrm{I\!R}^{2a\times 2a}$. The LCP matrix in (13.31) becomes

$$M = \begin{bmatrix} \widehat{W}_{\mathrm{NN}} + \widehat{W}_{\mathrm{NT}}\mu_{\mathscr{P}} & -\widehat{W}_{\mathrm{NT}} & 0 \\ -\widehat{W}_{\mathrm{TN}} - \widehat{W}_{\mathrm{TT}}\mu_{\mathscr{P}} & \widehat{W}_{\mathrm{TT}} & I_{2a} \\ 2\mu_{\mathscr{P}} & -I_{2a} & 0 \end{bmatrix}. \tag{13.40}$$

13.3.5 Inner Faceting of the Coulomb's Cone

In this section, we drop for a moment the subscripts k and $k+1$ and the superscript α to lighten the notation.

Another approach is based on an inner approximation as exposed in Al-Fahed et al. (1991) and Stewart & Trinkle (1996). The idea is to approach the friction disk by an interior polygon with ω edges (e.g., Fig.13.1b)

$$\mathbf{D}_{inner}(\mu R_{\mathrm{N}}) = \left\{ R_{\mathrm{T}} = D\beta \mid \beta \geqslant 0, \mu R_{\mathrm{N}} \geqslant \mathbb{1}^{\mathrm{T}} \beta \right\} \tag{13.41}$$

where $\beta \in \mathbb{R}^2$, $\mathbb{1} = [1, \ldots, 1]^{\mathrm{T}} \in \mathbb{R}^{\omega}$, the columns of the matrix $D \in \mathbb{R}^{2\times\omega}$ are the direction vectors d_j which are the coordinates of the vertices of the polygon. For the sake of simplicity, we assumed that for every i there is j such that $d_i = -d_j$.

Following the same process as in the previous case and rearranging the equation, we obtain the following MLCP:

$$\begin{cases} R_{\mathrm{T}} = D\beta \\ \\ 0 \leqslant \beta \perp \lambda\,\mathbb{1} + D^{\mathrm{T}} U_{\mathrm{T}} \geqslant 0 \\ 0 \leqslant \lambda \perp \mu R_{\mathrm{N}} - \mathbb{1}^{\mathrm{T}} \beta \geqslant 0 \end{cases} \tag{13.42}$$

where $\lambda \in \mathbb{R}$.

The time-discretized linear OSNSP, $(\mathscr{P}_{\mathrm{L}})$, can be written as the following MLCP:

$$\begin{cases} U_{k+1} = \widehat{W} P_{k+1} + U_{\mathrm{free}} \\ \\ \left.\begin{array}{l} P^{\alpha}_{T,k+1} = D^{\alpha} \beta^{\alpha}_{k+1} \\ \\ 0 \leqslant U^{\alpha}_{\mathrm{N},k+1} + e^{\alpha} U^{\alpha}_{\mathrm{N},k} \perp P^{\alpha}_{\mathrm{N},k+1} \geqslant 0 \\ 0 \leqslant \beta^{\alpha}_{k+1} \perp \lambda^{\alpha}_{k+1} \mathbb{1}^{\alpha} + D^{\alpha,\mathrm{T}} U^{\alpha}_{\mathrm{T},k+1} \geqslant 0 \\ 0 \leqslant \lambda \perp \mu P^{\alpha}_{\mathrm{N},k+1} - \mathbb{1}^{\mathrm{T},\alpha} \beta^{\alpha}_{k+1} \geqslant 0 \end{array}\right\} \forall \alpha \in I_a(\tilde{q}_{k+1})\,. \end{cases} \tag{13.43}$$

The LCP in a Single-Contact Case

We drop the superscript α to lighten the notation. Substituting the value of $P_{\mathrm{T},k+1}$ given by the second equation in (13.43) in the first equation in (13.43), one gets the LCP in standard form:

$$\begin{cases} \begin{bmatrix} U_{\mathrm{N},k+1} + eU_{\mathrm{N},k} \\ \kappa_{k+1} \\ \sigma_{k+1} \end{bmatrix} = M \begin{bmatrix} P_{\mathrm{N},k+1} \\ \beta_{k+1} \\ \lambda_{k+1} \end{bmatrix} + q \\ \\ 0 \leqslant \begin{bmatrix} U_{\mathrm{N},k+1} + eU^{\alpha}_{\mathrm{N},k} \\ \kappa_{k+1} \\ \sigma_{k+1} \end{bmatrix} \perp \begin{bmatrix} P_{\mathrm{N},k+1} \\ \beta_{k+1} \\ \lambda_{k+1} \end{bmatrix} \geqslant 0 \end{cases} \tag{13.44}$$

where

$$M = \begin{bmatrix} \widehat{W}_{\mathrm{NN}} & \widehat{W}_{\mathrm{NT}} D & 0 \\ \\ D^{\mathrm{T}} \widehat{W}_{\mathrm{TN}} & D^{\mathrm{T}} \widehat{W}_{\mathrm{TT}} D & \mathbb{1} \\ \\ \mu & -\mathbb{1}^{\mathrm{T}} & 0 \end{bmatrix} \tag{13.45}$$

and

$$q = \begin{bmatrix} U_{\mathrm{N,free}} + eU_{\mathrm{N,free}} \\ \\ D^{\mathrm{T}}(U_{\mathrm{T,free}}) \\ \\ 0 \end{bmatrix} . \tag{13.46}$$

The variables $\kappa_{k+1} \in \mathbb{R}^{\omega}$ and $\sigma_{k+1} \in \mathbb{R}$ are given by the following equations:

$$\kappa_{k+1} = \lambda_{k+1} + D^{\mathrm{T}} U_{\mathrm{T},k+1}, \qquad \sigma_{k+1} = \mu P_{\mathrm{N},k+1} - 1\!\!1^{\mathrm{T}} \beta_{k+1} . \tag{13.47}$$

The LCP in the Multicontact Case

In the multicontact case, the matrix notation must be enlarged to extend the formulation (13.45) and (13.46). In order to perform this extension, we introduce the following notation for all $\alpha \in I_a(\tilde{q}_{k+1})$:

$$\begin{cases} \mu = \mathrm{diag}\,(\mu^{\alpha}) \in \mathbb{R}^{a \times a} \\ \\ e = \mathrm{diag}\,(\mu^{\alpha}) \in \mathbb{R}^{a \times a} \\ \\ 1\!\!1 = \mathrm{diag}\,(1\!\!1^{\alpha}) \in \mathbb{R}^{(\Sigma_{\alpha}\, \omega^{\alpha}) \times a} . \end{cases} \tag{13.48}$$

With the help of this notation, the LCP (13.44) is still valid is the multi-contact case.

Let us note a result on the existence of solutions and their numerical computations (Stewart & Trinkle, 1996).

Proposition 13.6. *Let $\widehat{W}$ be a PD matrix. The LCP defined by (13.44), (13.45), and (13.46) possesses solutions, which can be computed by Lemke's algorithm (Algorithm 14) provided precautions are taken against cycling due to degeneracy.*

Proof. Let us prove first that the matrix M in (13.45) is copositive. For that, we choose a vector $z = [P_{\mathrm{N},k+1}\, \beta_{k+1}\, \lambda_{k+1}]^{\mathrm{T}} \geqslant 0$, and we compute

$$z^{\mathrm{T}} M z = \begin{bmatrix} P_{\mathrm{N},k+1} \\ D\beta_{k+1} \end{bmatrix}^{\mathrm{T}} \widehat{W} \begin{bmatrix} P_{\mathrm{N},k+1} \\ D\beta_{k+1} \end{bmatrix} + \mu P_{\mathrm{N},k+1} \lambda_{k+1} . \tag{13.49}$$

Since $\widehat{W}$ is assumed to be a PD matrix and $\mu \geqslant 0$, one obtains

$$z^{\mathrm{T}} M z \geqslant 0 :, \text{ for all } z \geqslant 0 , \tag{13.50}$$

i.e., M is copositive on the nonnegative orthant. Note, however, that M is not copositive plus, since $z^{\mathrm{T}} M z = 0$ implies that $P_{\mathrm{N},k+1} = 0$ and $D\beta_{k+1} = 0$, but *not* $\lambda = 0$, and $(M + M^{\mathrm{T}})z \neq 0$ if $\lambda > 0$.

Nevertheless, theorem 3.8.6 in Cottle et al. (1992, p. 179) asserts that if M is copositive and the implication

$$[z \geqslant 0, \quad Mz \geqslant 0, \quad z^{\mathrm{T}} M z = 0] \Longrightarrow z^{\mathrm{T}} q \geqslant 0 \tag{13.51}$$

is valid, then $LCP(M,q)$ has a solution. In our case, the right-hand side in (13.51) implies that

$$z = \begin{bmatrix} 0 \\ 0 \\ \lambda_{k+1} \end{bmatrix} \quad \text{with } \lambda_{k+1} \geqslant 0\,. \tag{13.52}$$

Using the special form of q given by (13.46), we conclude easily that the above cited theorem applies. We have then proved that the $LCP(M,q)$ has a solution.

Theorems 4.4.12 and 4.4.13 in Cottle et al. (1992, p. 277) assert that Lemke's algorithm 14 will compute a solution provided precautions are taken against cycling due to degeneracy. □

13.3.6 Comments

The presentation of the outer cone faceting is adapted to the time-stepping method where the constraint is treated at the velocity level. In the original works of Klarbring (1986a) and Klarbring & Björkman (1988), the method was presented in the context of the quasi-static modeling of elastic continuum media with unilateral contact and friction. In the context of nonsmooth dynamics, the book of Pfeiffer & Glocker (1996) presents how the two-dimensional contact friction with the Poisson impact law can be formulated as an LCP within an event-driven strategy. This latter approach can be straightforwardly extended to the 3-dimensional case.

To the best our knowledge, the first inner approximation of the Coulomb cone is due to Al-Fahed et al. (1991). In this latter work, fingered robot grippers are studied under the quasi-static assumption. The solvability of the resulting LCP is proved with the help the VI reformulation (see Sect. 12.6). These results are a special case of the theoretical study of Fichera (1972) on boundary value problems in elasticity with unilateral constraints.

The above cited result of Stewart & Trinkle (1996) has been originally formulated with a constraint on the position (see Remark 10.12). It has been easily extended to the case with bilateral constraints together with a Poisson impact law in Anitescu & Potra (1997). This result is based on the formulation of the Poisson law as a two-stage LCP as in Pfeiffer & Glocker (1996).

Other existence and uniqueness results can be found in Pang & Trinkle (1996) and Trinkle et al. (1997) for the approximated model of friction with an outer pyramid. These results are stated in the framework of finite-force models, i.e., without tangential collisions and no discontinuities in the tangential velocities. Furthermore, the solutions are known to exist only for small enough friction coefficients.

In Pang & Stewart (1999), general results of existence of solutions are given in a more unified framework; in particular, most of the time-discretized or time-incremental problems (quasi-static rigid and elastic body, dynamic of rigid bodies with collisions, etc.) with unilateral contact and friction are considered. Friction models are also unified comprising the standard Coulomb model, the faceted model, and more general friction models.

From the computational point of view, the faceted friction model is mainly solved by Lemke's algorithm 14 provided precautions are taken against cycling due to degeneracy. In Trinkle et al. (1997), a feasible interior point method (see Sect. 15) is compared to Lemke's algorithm. It is quite difficult to conclude about the relative efficiency of this algorithm. The first reason is that Lemke's algorithm fails to compute a solution on all the problem data. The choice of the covering vector seems to be problematic. There is also no mention of the use of a degeneracy resolution strategy. The conditions of the convergence of the feasible interior point method also seems to impose some constraints on the maximal value of the friction coefficient μ.

13.3.7 Weakness of the Faceting Process

Let us illustrate some weaknesses of the method that consists of faceting the Coulomb cone. Let us consider a ball of mass m lying on a horizontal plane under gravity g. A cycling external force defined by

$$F(t) = \begin{cases} \mu m g(\cos\frac{\pi}{3}\,\mathbf{i} + \sin\frac{\pi}{3}\,\mathbf{j}), & t \in [15k, 5+15k) \\ -\mu m g\,\mathbf{i}, & t \in [5+15k, 10+15k) \\ \mu m g(\cos\frac{\pi}{3}\mathbf{i} + -\sin\frac{\pi}{3}\mathbf{j}), & t \in [10+15k, 15(k+1)) \end{cases}, \ k \in \mathbb{N} \quad (13.53)$$

is applied to the ball. The norm of $F(t)$ is chosen such that the ball slides and the trajectory of the ball must match with an equilateral triangle. In particular the initial and the final points of a cycle must coincide.

Figure 13.2 a depicts the trajectories without approximation of the Coulomb cone, labeled by (w.a.) and with the faceting approximation (13.16) for $\omega = 2$ labeled by C2. Figure 13.2b depicts the trajectories obtained respectively with

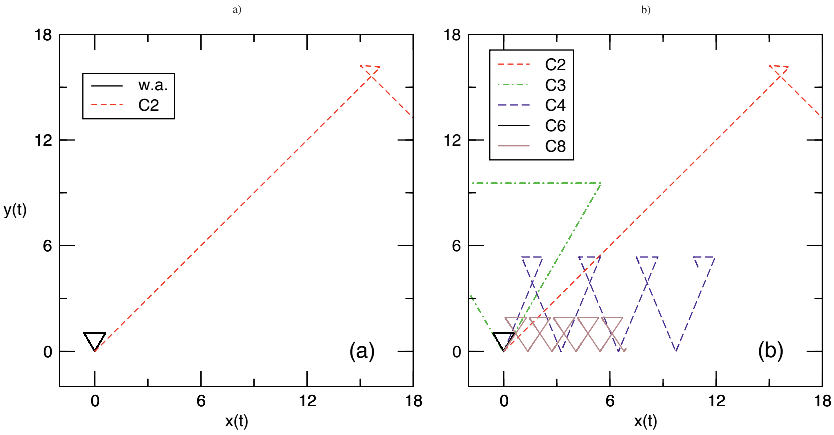


Fig. 13.2. Ball trajectory under cycling loading

$\omega \in \{2,3,4,6,8\}$ and labeled respectively by *C2,C3,C4,C6,C8*. Note that the time period has been simulated in each test.

Computations using faceting of the cone lead to unrealistic behaviors. The main reason is that the computed friction force does generally not oppose the direction of sliding. Indeed, the maximum dissipation principle together with a faceted cone implies that the contact force always lies along one of the edges of the polyhedral cone. To better approach the solution we must use a higher order approximation introducing a higher number of unknowns.

13.4 Formulation and Resolution in a Standard NCP Form

13.4.1 The Frictionless Case

Let us start with the mixed nonlinear OSNSP ($\mathscr{P}_{\mathrm{MNL}}$) in the frictionless case

$$\begin{cases} \mathscr{R}(v_{k+1}) = \displaystyle\sum_{\alpha \in I_a(\tilde{q}_{k+1})} H_{\mathrm{N}}^{\alpha}(q_k+1)\, P_{\mathrm{N},k+1}^{\alpha} \\ U_{\mathrm{N},k+1}^{\alpha} = H_{\mathrm{N}}^{\alpha,\mathrm{T}}(q_k+1)\, v_{k+1} \\ \left.\begin{array}{l} \widehat{U}_{\mathrm{N},k+1}^{\alpha} = U_{\mathrm{N},k+1}^{\alpha} + e^{\alpha} U_{\mathrm{N},k}^{\alpha} \\ 0 \leqslant \widehat{U}_{\mathrm{N},k+1}^{\alpha} \perp P_{\mathrm{N},k+1}^{\alpha} \geqslant 0 \end{array}\right\} \forall \alpha \in I_a(\tilde{q}_{k+1})\,. \end{cases} \tag{13.54}$$

This problem can be stated in the following NCP for v_{k+1} and $P_{\mathrm{N},k+1}^{\alpha}$, $\alpha \in I_a(\tilde{q}_{k+1})$

$$\begin{cases} \mathscr{R}(v_{k+1}) = \displaystyle\sum_{\alpha \in I_a(\tilde{q}_{k+1})} H_{\mathrm{N}}^{\alpha}(q_k+1)\, P_{\mathrm{N},k+1}^{\alpha} \\ 0 \leqslant H_{\mathrm{N}}^{\alpha,\mathrm{T}}(q_k+1)\, v_{k+1} + e^{\alpha} U_{\mathrm{N},k}^{\alpha} \perp P_{\mathrm{N},k+1}^{\alpha} \geqslant 0, \qquad \forall \alpha \in I_a(\tilde{q}_{k+1})\,. \end{cases} \tag{13.55}$$

13.4.2 A Direct MCP for the 3D Frictional Contact

The Coulomb friction model (3.147) can be easily reformulated into the following CP:

$$\begin{cases} R_{\mathrm{T}}\, \|U_{\mathrm{T}}\| + \|R_{\mathrm{T}}\| U_{\mathrm{T}} = 0 \\ 0 \leqslant \|U_{\mathrm{T}}\| \perp \mu R_{\mathrm{N}} - \|R_{\mathrm{T}}\| \geqslant 0\,. \end{cases} \tag{13.56}$$

Let us denote $\kappa = \|U_{\mathrm{T}}\|$, the norm of U_{T}. The following CP can be stated

$$\begin{cases} \kappa = \|U_{\mathrm{T}}\| \\ \\ R_{\mathrm{T}}\,\kappa + \|R_{\mathrm{T}}\| U_{\mathrm{T}} = 0 \\ \\ 0 \leqslant \kappa \perp \mu R_{\mathrm{N}} - \|R_{\mathrm{T}}\| \geqslant 0\ . \end{cases} \tag{13.57}$$

To obtain a complete MCP formulation we need to add the complementarity relation between U_{N} and R_{N} and the dynamics. For the sake of simplicity, let us consider the linear OSNSP $(\mathscr{P}_{\mathrm{L}})$. The following MCP can be then written as

$$\begin{cases} U_{k+1} = \widehat{W} P_{k+1} + U_{\mathrm{free}} \\ \\ \left.\begin{array}{l} \kappa^{\alpha}_{k+1} = \|U^{\alpha}_{\mathrm{T},k+1}\| \\ \tilde{U}^{\alpha}_{\mathrm{N},k+1} = U^{\alpha}_{\mathrm{N},k+1} + e^{\alpha} U^{\alpha}_{\mathrm{N},k} \\ \\ P^{\alpha}_{\mathrm{T},k+1}\,\kappa^{\alpha}_{k+1} + \|P^{\alpha}_{\mathrm{N},k+1}\| U^{\alpha}_{\mathrm{T},k+1} = 0 \\ \\ 0 \leqslant \kappa^{\alpha}_{k+1} \perp \mu^{\alpha} P^{\alpha}_{\mathrm{N},k+1} - \|P^{\alpha}_{\mathrm{T},k+1}\| \geqslant 0 \\ 0 \leqslant \tilde{U}^{\alpha}_{\mathrm{N},k+1} \perp P^{\alpha}_{\mathrm{N},k+1} \geqslant 0 \end{array}\right\} \forall \alpha \in I_a(\tilde{q}_{k+1}) \end{cases} \tag{13.58}$$

which can be casted into the MCP form (12.187) with $u = [U^{\alpha}_{\mathrm{T},k+1}, P^{\alpha}_{\mathrm{N},k+1}, P^{\alpha}_{\mathrm{T},k+1}]^{\mathrm{T}}$ and $v = [\kappa^{\alpha}_{k+1}, U_{\mathrm{N}}]^{\mathrm{T}}$.

Besides the difficulty to directly deal with a MCP in general form, the main drawback of this formulation is the lack of differentiability of the mappings involved in the complementarity preventing the use of most of the nonsmooth Newton solvers (see Sect. 12.5.4).

13.4.3 A Clever Formulation of the 3D Frictional Contact as an NCP

In Glocker (1999), the nondifferentiability of the previous formulation has been overcome in a very elegant way. We will give here a short overview of such a formulation. Starting from the inclusion (3.149) with the standard definition (13.15) of the friction disk $\mathbf{D}(\mu r_{\mathrm{N}})$, Glocker (1999) adds three inequalities

$$\sigma_i(R_{\mathrm{T}}) = \mu R_{\mathrm{N}} - e_i^{\mathrm{T}} R_{\mathrm{T}} \geqslant 0,\ \ i = 1,2,3 \tag{13.59}$$

where e_1, e_2, e_3 are three unit outward vector defined by

$$e_i = [\cos\alpha_i, \sin\alpha_i], \quad \alpha_i = \frac{(4i-3)\pi}{6}\ . \tag{13.60}$$

We can remark that

$$\mathbf{D}(\mu R_{\mathrm{N}}) = \cap_{i=1}^{3} \mathbf{D}_i \cap \mathbf{D}(\mu r_{\mathrm{N}})\ , \tag{13.61}$$

thus the Coulomb's frictional law remains as in (3.149). This normal cone condition leads to

$$-U_{\mathrm{T}} \in \Sigma_{i=1}^{3} e_i \kappa_i + \partial\sigma_D(R_{\mathrm{T}})\,\kappa_D \tag{13.62}$$

where $\sigma_D(R_{\mathrm{T}}) = \mu^2 R_{\mathrm{N}}^2 - \|R_{\mathrm{T}}\|^2$ is a nonlinear friction saturation associated with the second-order cone.

The trick introduced by Glocker lies in the reformulation of this inclusion into an equation of the form

$$-U_{\mathrm{T}} = \Sigma_{i=1}^{3} e_i \kappa_i + 2R_{\mathrm{T}}\,\kappa_D, \qquad 0 \leqslant \kappa_j \perp \sigma_j \geqslant 0,\ \ j = 1,2,3,D\,. \tag{13.63}$$

MCP Formulation

For the sake of simplicity, let us consider the linear OSNSP ($\mathscr{P}_{\mathrm{L}}$). A more general MCP can be easily written starting from ($\mathscr{P}_{\mathrm{MNL}}$). The previous CP formulation yields

$$\begin{cases} U_{k+1} = \widehat{W} P_{k+1} + U_{\mathrm{free}} \\ \left.\begin{array}{l} -U^{\alpha}_{\mathrm{T},k+1} = \displaystyle\sum_{i=1}^{3} e_i \kappa^{\alpha}_{i,k+1} + 2P^{\alpha}_{\mathrm{T},k+1}\,\kappa^{\alpha}_{C,k+1} \\ \sigma^{\alpha}_i(P^{\alpha}_{\mathrm{T},k+1}) = \mu^{\alpha} P^{\alpha}_{\mathrm{N},k+1} - e_i^{\mathrm{T}} P^{\alpha}_{\mathrm{T},k+1}, i = 1,2,3 \\ \sigma^{\alpha}_D(P^{\alpha}_{\mathrm{T},k+1}) = (\mu^{\alpha} P^{\alpha}_{\mathrm{N},k+1})^2 - \|P^{\alpha}_{\mathrm{T},k+1}\|^2 \\ \\ 0 \leqslant U^{\alpha}_{\mathrm{N},k+1} + e^{\alpha} U^{\alpha}_{\mathrm{N},k} \perp P^{\alpha}_{\mathrm{N},k+1} \geqslant 0 \\ 0 \leqslant \kappa^{\alpha}_{j,k+1} \perp \sigma^{\alpha}_{j,k+1} \geqslant 0, j = 1,2,3,D \end{array}\right\} \forall \alpha \in I_a(\tilde{q}_{k+1})\,. \end{cases} \tag{13.64}$$

From this formulation, the derivation of an NCP can be made following the path of Sect. 13.3.4. We specify the notation of Sect. 13.3.4 by denoting

$$\begin{aligned} \mathscr{P}^{\alpha} &= \{1,2\} \\ I_{\mathscr{P}^{\alpha}} &= \begin{bmatrix} e_1 & e_2 \end{bmatrix}\,. \end{aligned} \tag{13.65}$$

One gets the following NCP of dimension 4

$$\begin{cases} \begin{bmatrix} U_{\mathrm{N},k+1} + eU_{\mathrm{N},k} \\ \kappa_{\mathscr{P}} \\ \sigma_3 \\ \sigma_D \end{bmatrix} = M \begin{bmatrix} P_{\mathrm{N},k+1} \\ \sigma_{\mathscr{P}} \\ \kappa_3 \\ \kappa_D \end{bmatrix} + g\left(\begin{bmatrix} P_{\mathrm{N},k+1} \\ \sigma_{\mathscr{P}} \\ \kappa_3 \\ \kappa_D \end{bmatrix}\right) + q \\ 0 \leqslant \begin{bmatrix} U_{\mathrm{N},k+1} + eU_{\mathrm{N},k} \\ \kappa_{\mathscr{P}} \\ \sigma_3 \\ \sigma_D \end{bmatrix} \perp \begin{bmatrix} P_{\mathrm{N},k+1} \\ \sigma_{\mathscr{P}} \\ \kappa_3 \\ \sigma_D \end{bmatrix} \geqslant 0 \end{cases} \tag{13.66}$$

where

$$M = \begin{bmatrix} \widehat{W}_{\mathrm{NN}} + \widehat{W}_{\mathrm{NT}} I_{\mathscr{P}}^{-\mathrm{T}} \mu_{\mathscr{P}} & -\widehat{W}_{\mathrm{NT}} I_{\mathscr{P}}^{-\mathrm{T}} & 0 & 0 \\ -I_{\mathscr{P}}^{-1} [\widehat{W}_{\mathrm{TN}} + \widehat{W}_{\mathrm{TT}} I_{\mathscr{P}}^{-\mathrm{T}} \mu_{\mathscr{P}}] & I_{\mathscr{P}}^{-1} \widehat{W}_{\mathrm{TT}} I_{\mathscr{P}}^{-\mathrm{T}} & -I_{\mathscr{P}}^{-1} e_3 & 0 \\ \mu - e_3^{\mathrm{T}} I_{\mathscr{P}}^{-\mathrm{T}} \mu_{\mathscr{P}} & e_3^{\mathrm{T}} I_{\mathscr{P}}^{-\mathrm{T}} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{13.67}$$

$$g\left(\begin{bmatrix} P_{\mathrm{N},k+1} \\ \sigma_{\mathscr{P}} \\ \kappa_3 \\ \kappa_D \end{bmatrix} \right) = \begin{bmatrix} 0 \\ -2I(\mu_{\mathscr{P}} P_{\mathrm{N},k+1} - \sigma_{\mathscr{P}})\kappa_D \\ 0 \\ (\mu P_{\mathrm{N},k+1})^2 - \|\mu_{\mathscr{P}} P_{\mathrm{N},k+1} - \sigma_{\mathscr{P}}\|_I^2 \end{bmatrix}, \text{ with } I = I_{\mathscr{P}}^{-1} I_{\mathscr{P}}^{-\mathrm{T}} \tag{13.68}$$

and

$$q = \begin{bmatrix} U_{\mathrm{N,free}} + e U_{\mathrm{N,free}} \\ -I_{\mathscr{P}}^{-1} U_{\mathrm{T,free}} \\ 0 \\ 0 \end{bmatrix}. \tag{13.69}$$

The form of the function $F(z) = Mz + g(z) + q$ with $z = \left[P_{\mathrm{N},k+1}, \sigma_{\mathscr{P}}, \kappa_3, \kappa_D,\right]^{\mathrm{T}}$ shows that the computation of the Jacobian $\nabla F^{\mathrm{T}}(z) = M + \nabla g^{\mathrm{T}}(z)$ only requires the computation of $\nabla g^{\mathrm{T}}(z)$. This Jacobian is given by

$$\nabla g^{\mathrm{T}}(z) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -2I\mu_{\mathscr{P}}\kappa_D & 2I\kappa_D & 0 & -2IP_{\mathscr{P},k+1} \\ 0 & 0 & 0 & 0 \\ 2\mu^2 P_{\mathrm{N},k+1} - 2P_{\mathscr{P},k+1}^{\mathrm{T}} I^{\mathrm{T}} \mu_{\mathscr{P}} & 2P_{\mathscr{P},k+1}^{\mathrm{T}} I^{\mathrm{T}} & 0 & 0 \end{bmatrix} \tag{13.70}$$

with $P_{\mathscr{P},k+1} = \mu_{\mathscr{P}} P_{\mathrm{N},k+1} - \sigma_{\mathscr{P}}$.

Note that the function $F(z)$ is continuously differentiable and the structure of the matrix M in (13.67) is analogous to those in (13.31).

13.5 Formulation and Resolution in QP and NLP Forms

13.5.1 The Frictionless Case

QP and NLP Reformulations

Let us start with the linear frictionless contact problem in the form $(\mathscr{P}_{\mathrm{LWF}})$ which can be reformulated as the following LCP (see Sect. 13.3.1)

$$\begin{cases} \widehat{U}_{\mathrm{N},k+1} = \widehat{W}_{\mathrm{NN}} P_{\mathrm{N},k+1} + U_{\mathrm{N,free}} - e \circ U_{\mathrm{N},k} \\ 0 \leqslant \widehat{U}_{\mathrm{N},k+1} \perp P_{\mathrm{N},k+1} \geqslant 0. \end{cases} \tag{13.71}$$

The question of the reformulation of (13.71) into a QP amounts to checking the assumptions for the equivalence between a QP and an LCP. This question has already been developed in Sect. 12.4.5. Under the standard assumption that $\widehat{W}_{\mathrm{NN}}$ is symmetric PSD, the following equivalent QP can be solved:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2} P^{\mathrm{T}}_{\mathrm{N},k+1} \widehat{W}_{\mathrm{NN}} P_{\mathrm{N},k+1} + P^{\mathrm{T}}_{\mathrm{N},k+1} (U_{\mathrm{N,free}} - e \circ U_{\mathrm{N},k}) \\ &\text{subject to} \quad P_{\mathrm{N},k+1} \geqslant 0. \end{aligned} \tag{13.72}$$

A NLP reformulation can be considered as in Sect. 12.5.1 for the NCP given in (13.55). This reformulation is not straightforward due to the particular form of the NCP (13.55). Nevertheless, it should be possible to state a NLP provided that the function involved in the NCP is a gradient mapping.

Numerical Methods

The choice of the numerical methods to solve the previous minimization problems depends strongly on the size and the structure of the problem and follows the recommendations at the end of Sect. 12.2.4. Generally speaking, gradient projection and interior point methods are well suited for large-scale systems. Active-set methods are interesting for small scale, possibly ill-conditioned systems. Generally speaking, the minimization formulation improves the robustness of the algorithm.

The idea of using splitting and block splitting to solve the frictionless contact problem dates back to Glowinski et al. (1976) and Mittelmann (1978). In Mittelmann (1980, 1981a,b), an acceleration method is proposed in the spirit of the work of Moré & Toraldo (1991) presented in Sect. 12.2.3.2. In a first phase a splitting method is used. In a second phase, a preconditioned conjugate gradient with projection is used to improve the rate of convergence.

13.5.2 Minimization Principles and Coulomb's Friction

When the Coulomb's friction is involved, there is no direct associated minimization formulation. This is mainly due to the fact that Coulomb's friction is a nonassociated friction law De Saxcé (1991), where the relative velocity is not in the normal cone of the Coulomb's cone. Nevertheless, it is possible to state generalized optimization as saddle-point problems or minimization problems plus a condition on the value of the objective function at the optimal point as in Theorem 12.65. The framework of VIs and CPs is better suited to write these reformulations. This is the reason why we postpone these developments in Sect. 13.7.

13.6 Formulations and Resolution as Nonsmooth Equations

In the pioneering works of Curnier & Alart (1988), Alart & Curnier (1991) and Alart (1993), a generalized Newton method is proposed for the resolution of the three-dimensional frictional contact problem. It is based on a generalized equations reformulation. The principle of this Newton method is analogous to those presented in Sect. 12.5.4. The only discrepancy lies in the reformulation of the frictional contact problem as a system of generalized equations, which does not rely on a NCP formulation. The term "dedicated" refers to the fact that no intermediate NCP is written; the equation-based reformulation is derived directly from the three-dimensional frictional contact problem. The section ends with some alternative equation-based formulations and a line-search procedure.

13.6.1 Alart and Curnier's Formulation and Generalized Newton's Method

13.6.1.1 At the Level of the Local Variables

For the sake of simplicity, let us start with the linear OSNSP $(\mathscr{P}_{\mathrm{L}})$ defined by

$$\begin{cases} U_{k+1} = \widehat{W} P_{k+1} + U_{\mathrm{free}} \\ \left.\begin{array}{l} \widehat{U}^{\alpha}_{k+1} = \left[U^{\alpha}_{\mathrm{N},k+1} + e^{\alpha} U^{\alpha}_{\mathrm{N},k} + \mu^{\alpha}\, ||U^{\alpha}_{\mathrm{T},k+1}||, U^{\alpha}_{\mathrm{T},k+1}\right]^{\mathrm{T}} \\ \mathbf{C}^{\alpha,*} \ni \widehat{U}^{\alpha}_{k+1} \perp P^{\alpha}_{k+1} \in C^{\alpha} \end{array}\right\} \forall \alpha \in I_a(\tilde{q}_{k+1}). \end{cases} \tag{13.73}$$

Using the equivalent formulation of the unilateral contact and friction in terms of projection operators $\mathrm{proj}(\cdot)$ (see (A.12), (A.13), and (3.152)), the previous linear OSNSP (13.73) can be written as

$$\begin{cases} U_{k+1} = \widehat{W} P_{k+1} + U_{\mathrm{free}} \\ \left.\begin{array}{l} P^{\alpha}_{\mathrm{N},k+1} = \mathrm{proj}_{\mathbb{R}_+}(P^{\alpha}_{\mathrm{N},k+1} - \rho^{\alpha}_{\mathrm{N}}\,(U^{\alpha}_{\mathrm{N},k+1} + e^{\alpha} U^{\alpha}_{\mathrm{N},k})) \\ P^{\alpha}_{\mathrm{T},k+1} = \mathrm{proj}_{\widehat{\mathbf{D}}^{\alpha}(P^{\alpha}_{\mathrm{N},k+1}, U^{\alpha}_{\mathrm{N},k+1})}(P^{\alpha}_{\mathrm{T},k+1} - \rho^{\alpha}_{\mathrm{T}} \circ U^{\alpha}_{\mathrm{T},k+1}) \end{array}\right\} \forall \alpha \in I_a(\tilde{q}_{k+1}) \end{cases} \tag{13.74}$$

where $\rho^{\alpha}_{\mathrm{N}} > 0$, $\rho^{\alpha}_{\mathrm{T}} \in \mathbb{R}^2_+ \setminus \{0\}$ for all $\alpha \in I_a(\tilde{q}_{k+1})$ and the modified friction disk is

$$\widehat{\mathbf{D}}^{\alpha}(P^{\alpha}_{\mathrm{N},k+1}, U^{\alpha}_{\mathrm{N},k+1}) = \mathbf{D}(\mu(\mathrm{proj}_{\mathbb{R}_+}(P^{\alpha}_{\mathrm{N},k+1} - \rho^{\alpha}_{\mathrm{N}}\,(U^{\alpha}_{\mathrm{N},k+1} + e^{\alpha} U^{\alpha}_{\mathrm{N},k})))) \tag{13.75}$$

for all $\alpha \in I_a(\tilde{q}_{k+1})$. We recall that $\cdot \circ \cdot$ is the Hadamard product of vectors.

As we saw earlier in Sects. 12.5.4 and 12.6.5, the use of the projection operators $\mathrm{proj}(\cdot)$, or more generally the natural and normal map, allows one to restate a CP or a VI into a system of nonlinear nonsmooth equations,

$$\Phi(U_{k+1},P_{k+1}) = \begin{bmatrix} -U_{k+1}+\widehat{W}P_{k+1}+U_{\mathrm{free}} \\ \\ P_{\mathrm{N},k+1}-\mathrm{proj}_{\mathbb{R}^a_+}(P_{\mathrm{N},k+1}-\rho_{\mathrm{N}}\circ(U_{\mathrm{N},k+1}+e\circ U_{\mathrm{N},k})) \\ \\ P_{\mathrm{T},k+1}-\mathrm{proj}_{\widehat{\mathbf{D}}(P_{\mathrm{N},k+1},U_{\mathrm{N},k+1})}(P_{\mathrm{T},k+1}-\rho_{\mathrm{T}}\circ U_{\mathrm{T},k+1}) \end{bmatrix} = 0, \tag{13.76}$$

where the following notation has been introduced:

$$\begin{aligned} \rho_{\mathrm{N}} &= [\rho_{\mathrm{N}}^{\alpha}]^{\mathrm{T}},\ \text{for all } \alpha \in I_a(\tilde{q}_{k+1}) \\ \rho_{\mathrm{T}} &= [\rho_{\mathrm{T}}^{\alpha}]^{\mathrm{T}},\ \text{for all } \alpha \in I_a(\tilde{q}_{k+1}) \\ \widehat{\mathbf{D}}(P_{\mathrm{N},k+1},U_{\mathrm{N},k+1}) &= \prod_{\alpha\in I_a(\tilde{q}_{k+1})} \widehat{\mathbf{D}}^{\alpha}(P^{\alpha}_{\mathrm{N},k+1},U^{\alpha}_{\mathrm{N},k+1}). \end{aligned} \tag{13.77}$$

We recall that a is the cardinal of $I_a(\tilde{q}_{k+1})$ and the symbol $\prod$ represents the Cartesian product of sets.

The solution procedure is based on a nonsmooth Newton method as presented in Sect. 12.5.4 on the system (13.76). One of the basic ingredients of the method is the computation of an element of the Clarke generalized Jacobian $\partial\Phi(U,P)$. We propose to derive here a possible solution. Let us denote the one element of the generalized Jacobian by $H(U,P) \in \partial\Phi(U,P)$ which has the structure

$$H(U,P) = \begin{bmatrix} -I & 0 & \widehat{W}_{\mathrm{NN}} & \widehat{W}_{\mathrm{NT}} \\ \\ 0 & -I & \widehat{W}_{\mathrm{TN}} & \widehat{W}_{\mathrm{TT}} \\ \\ \partial_{U_{\mathrm{N}}}\Phi_2(U,P) & 0 & \partial_{P_{\mathrm{N}}}\Phi_2(U,P) & 0 \\ \\ 0 & \partial_{U_{\mathrm{T}}}\Phi_3(U,P) & \partial_{P_{\mathrm{N}}}\Phi_3(U,P) & \partial_{P_{\mathrm{T}}}\Phi_3(U,P) \end{bmatrix}, \tag{13.78}$$

where the components of Φ are defined by

$$\begin{aligned} \Phi_1(U,P) &= -U_{k+1}+\widehat{W}P_{k+1}+U_{\mathrm{free}} \\ \Phi_2(U,P) &= P_{\mathrm{N}}-\mathrm{proj}_{\mathbb{R}^a_+}(P_{\mathrm{N}}-\rho_{\mathrm{N}}\circ(U_{\mathrm{N}}+e\circ U_{\mathrm{N},k})) \\ \Phi_3(U,P) &= P_{\mathrm{T}}-\mathrm{proj}_{\widehat{\mathbf{D}}(P_{\mathrm{N}},U_{\mathrm{N}})}(P_{\mathrm{T},k+1}-\rho_{\mathrm{T}}\circ U_{\mathrm{T}}). \end{aligned} \tag{13.79}$$

To explicitly compute the value of the Clarke generalized Jacobian of $\Phi_2(U,P)$ and $\Phi_3(U,P)$, we restrict ourselves to the single-contact case. Remind that

$$\partial\,\mathrm{proj}_{\mathbb{R}_+}(x) = \begin{cases} 1 \text{ if } x>0 \\ [0,1] \text{ if } x=0 \\ 0 \text{ if } x<0 \end{cases}, \tag{13.80}$$

$$\text{proj}_{\mathbf{D}(1)}(x) = \begin{cases} x & \text{if } x \in \mathbf{D}(1) \\ \dfrac{x}{\|x\|} & \text{if } x \notin \mathbf{D}(1) \end{cases}, \tag{13.81}$$

$$\partial\,\text{proj}_{\mathbf{D}(1)}(x) = \begin{cases} I_2 & \text{if } x \in \mathbf{D}(1) \setminus \partial\mathbf{D}(1) \\ \dfrac{I_2}{\|x\|} - \dfrac{xx^{\mathrm{T}}}{\|x\|^3} & \text{if } x \notin \mathbf{D}(1) \\ I_2 + (s-1)xx^{\mathrm{T}}, s \in [0,1] & \text{if } x \in \partial\mathbf{D}(1). \end{cases} \tag{13.82}$$

An element of the generalized Jacobian can be chosen as

$$\begin{aligned}
\partial_{U_{\mathrm{N}}}\Phi_2(U,P) &= \begin{cases} \rho_{\mathrm{N}} & \text{if } P_{\mathrm{N}} - \rho_{\mathrm{N}}U_{\mathrm{N}} > 0 \\ 0 & \text{otherwise} \end{cases} \\
\partial_{P_{\mathrm{N}}}\Phi_2(U,P) &= \begin{cases} 0 & \text{if } P_{\mathrm{N}} - \rho_{\mathrm{N}}U_{\mathrm{N}} > 0 \\ 1 & \text{otherwise} \end{cases} \\
\partial_{U_{\mathrm{T}}}\Phi_3(U,P) &= \begin{cases} \rho_{\mathrm{T}} I_{2\times 2} & \text{if } P_{\mathrm{T}} - \rho_{\mathrm{T}}U_{\mathrm{T}} \in \widehat{\mathbf{D}}(P_{\mathrm{N}}, U_{\mathrm{N}}) \\ \rho_{\mathrm{T}}\mu P_{\mathrm{N}}\,\Gamma(P_{\mathrm{T}} - \rho_{\mathrm{T}}U_{\mathrm{T}}) & \text{otherwise} \end{cases} \\
\partial_{R_{\mathrm{N}}}\Phi_3(U,P) &= \begin{cases} 0 & \text{if } P_{\mathrm{T}} - \rho_{\mathrm{T}}U_{\mathrm{T}} \in \widehat{\mathbf{D}}(P_{\mathrm{N}}, U_{\mathrm{N}}) \\ -\mu P_{\mathrm{N}} \dfrac{P_{\mathrm{T}} - \rho_{\mathrm{T}}U_{\mathrm{T}}}{\|P_{\mathrm{T}} - \rho_{\mathrm{T}}U_{\mathrm{T}}\|} & \text{otherwise} \end{cases} \\
\partial_{R_{\mathrm{T}}}\Phi_3(U,P) &= \begin{cases} 0 & \text{if } P_{\mathrm{T}} - \rho_{\mathrm{T}}U_{\mathrm{T}} \in \widehat{\mathbf{D}}(P_{\mathrm{N}}, U_{\mathrm{N}}) \\ I_2 - \rho_{\mathrm{T}}\mu P_{\mathrm{N}}\,\Gamma(P_{\mathrm{T}} - \rho_{\mathrm{T}}U_{\mathrm{T}}) & \text{otherwise} \end{cases}
\end{aligned}, \tag{13.83}$$

where the function $\Gamma(\cdot)$ is defined by

$$\Gamma(x) = \frac{I_{2\times 2}}{\|x\|} - \frac{xx^{\mathrm{T}}}{\|x\|^3}. \tag{13.84}$$

The multi-contact case is treated as well collecting the generalized Jacobians for each contact.

We can take benefits from the structure of the element of the generalized Jacobian in (13.78) which can be written as

$$H(U,P) = \begin{bmatrix} -I & \widehat{W} \\ A & B \end{bmatrix}, \tag{13.85}$$

where the matrices A and B can be easily identified. The Newton iteration amounts to solving for U^{i+1} and P^{i+1} the linear system at the iteration i

$$H(U^i,P^i)\begin{bmatrix} U^{i+1}-U^i \\ P^{i+1}-P^i \end{bmatrix} = -\Phi(U^i,P^i). \tag{13.86}$$

Due to the special structure of H in (13.85), it is interesting to substitute the value of $U^{i+1}-U^i$ in the second equation and to obtain the reduced linear system

$$(A\widehat{W}+B)(P^{i+1}-P^i) = -A\Phi_1(U^i,P^i) - \begin{bmatrix} \Phi_2(U^i,P^i) \\ \Phi_3(U^i,P^i) \end{bmatrix}. \tag{13.87}$$

Once we have obtained the new value of P^{i+1}, we compute U^{i+1} with the first block of equation in (13.85).

Remark 13.7. In Curnier & Alart (1988), Alart & Curnier (1991) and Alart (1993), the reformulation of (13.73) in (13.74) is motivated by the introduction of an augmented Lagrangian function. As said in Sect. 12.3, there are a lot of augmented Lagrangian functions. The quadratic augmentation of the constraints is the most usual one. It has been introduced in its basic form for equality constraints by Hestenes (1969) and Powell (1969) and has been extended to inequality constraints by Rockafellar (1973). The work of Rockafellar (1973, 1974, 1976a, 1979, 1993) is very interesting in the sense that it bridges the gap between the augmented Lagrangian method and the notion of the proximal operator and the Moreau–Yosida regularization (Moreau, 1965). The proximal point algorithm, which derives from this analysis, provides us with a numerical tool for solving a minimization problem based on the augmented Lagrangian function.

We will not enter here into deeper details because the augmented Lagrangian is more theoretical tool rather than a numerical method. It allows one to write some "augmented" KKT conditions such as those obtained with the projection operator in (13.74); but it does not provide us with a solution procedure. Indeed, the notion of augmented Lagrangian functions have also been used in Simo & Laursen (1992) and Laursen & Simo (1993a,b) and in De Saxcé & Feng (1991) but with completely different solution procedures.

13.6.1.2 At the Level of the Generalized Variables

The use of a generalized Newton method for solving the contact friction problems advocates to include the treatment of the *global nonlinearities* (see the discussion at the beginning of Sect. 10.1.5) directly into the Newton method. In this section, we will expose some ingredients of the formulation of the so-called global generalized Newton method.

For the sake of simplicity, let us consider first the mixed linear OSNSP ($\mathscr{P}_{\mathrm{ML}}$) defined by

$$\begin{cases} \widehat{M}(v_{k+1}-v_{\text{free}}) = p_{k+1} = \displaystyle\sum_{\alpha \in I_a(\tilde{q}_{k+1})} p^{\alpha}_{k+1} \\[2ex] U^{\alpha}_{k+1} = H^{\alpha,\mathrm{T}}\, v_{k+1} \\[2ex] p^{\alpha}_{k+1} = H^{\alpha}\, P^{\alpha}_{k+1} \\[2ex] \left.\begin{array}{l} \widehat{U}^{\alpha}_{k+1} = \left[U^{\alpha}_{\mathrm{N},k+1} + e^{\alpha} U^{\alpha}_{\mathrm{N},k} + \mu^{\alpha}\, \|U^{\alpha}_{\mathrm{T},k+1}\|, U^{\alpha}_{\mathrm{T},k+1}\right]^{\mathrm{T}} \\[2ex] \mathbf{C}^{\alpha,*} \ni \widehat{U}^{\alpha}_{k+1} \perp P^{\alpha}_{k+1} \in \mathbf{C}^{\alpha} \end{array}\right\} \forall \alpha \in I_a(\tilde{q}_{k+1}). \end{cases} \tag{13.88}$$

Using the same development than in the previous section, the mixed linear OSNSP (13.88) can be written as a set of nonlinear nonsmooth equations as

$$\Phi(v_{k+1}, P_{k+1}) = \begin{bmatrix} \widehat{M}(v_{k+1}-v_{\text{free}}) \\ \qquad -H_{\mathrm{N}}\, \mathrm{proj}_{\mathbb{R}^a_+}(P_{\mathrm{N},k+1} - \rho_{\mathrm{N}} \circ (H_{\mathrm{N}} v_{k+1} + e \circ H_{\mathrm{N}} v_k)) \\ \qquad -H_{\mathrm{T}}\, \mathrm{proj}_{\widehat{\mathbf{D}}(P_{\mathrm{N},k+1}, v_{k+1})}(P_{\mathrm{T},k+1} - \rho_{\mathrm{T}} \circ H_{\mathrm{T}} v_{k+1}) \\[2ex] H_{\mathrm{N}}\left[P_{\mathrm{N},k+1} - \mathrm{proj}_{\mathbb{R}^a_+}(P_{\mathrm{N},k+1} - \rho_{\mathrm{N}} \circ (H_{\mathrm{N}} v_{k+1} + e \circ H_{\mathrm{N}} v_k))\right] \\[2ex] H_{\mathrm{T}}\left[P_{\mathrm{T},k+1} - \mathrm{proj}_{\widehat{\mathbf{D}}(P_{\mathrm{N},k+1}, v_{k+1})}(P_{\mathrm{T},k+1} - \rho_{\mathrm{T}} \circ H_{\mathrm{T}} v_{k+1})\right] \end{bmatrix} = 0, \tag{13.89}$$

where the modified friction disk is expressed in terms of the generalized velocity v_{k+1} as

$$\widehat{\mathbf{D}}(P_{\mathrm{N},k+1}, v_{k+1}) = \prod_{\alpha \in I_a(\tilde{q}_{k+1})} \mathbf{D}(\mu^{\alpha}(\mathrm{proj}_{\mathbb{R}_+}(P^{\alpha}_{\mathrm{N},k+1} - \rho^{\alpha}_{\mathrm{N}}(H^{\alpha}_{\mathrm{N}} v_{k+1} + e^{\alpha} H^{\alpha}_{\mathrm{N}} v_k))). \tag{13.90}$$

In the same way, the mixed nonlinear OSNSP ($\mathscr{P}_{\mathrm{MNL}}$) may be considered

$$\begin{cases} \mathscr{R}(v_{k+1}) = p_{k+1} = \displaystyle\sum_{\alpha \in I_a(\tilde{q}_{k+1})} p^{\alpha}_{k+1} \\[2ex] U^{\alpha}_{k+1} = H^{\alpha,\mathrm{T}}(q_k+1)\, v_{k+1} \\[2ex] p^{\alpha}_{k+1} = H^{\alpha}(q_k+1)\, P^{\alpha}_{k+1} \\[2ex] \left.\begin{array}{l} \widehat{U}^{\alpha}_{k+1} = \left[U^{\alpha}_{\mathrm{N},k+1} + e^{\alpha} U^{\alpha}_{\mathrm{N},k} + \mu^{\alpha}\, \|U^{\alpha}_{\mathrm{T},k+1}\|, U^{\alpha}_{\mathrm{T},k+1}\right]^{\mathrm{T}} \\[2ex] \mathbf{C}^{\alpha,*} \ni \widehat{U}^{\alpha}_{k+1} \perp P^{\alpha}_{k+1} \in \mathbf{C}^{\alpha} \end{array}\right\} \forall \alpha \in I_a(\tilde{q}_{k+1}). \end{cases} \tag{13.91}$$

This OSNSP can give rise to the following equation-based reformulation,

$$\Phi(v_{k+1},P_{k+1}) = \begin{bmatrix} \mathscr{R}(v_{k+1}) - H_{\mathrm{N}}\,\mathrm{proj}_{\mathrm{IR}^a_+}(P_{\mathrm{N},k+1} - \rho_{\mathrm{N}} \circ (H_{\mathrm{N}}v_{k+1} + e \circ H_{\mathrm{N}}v_k)) \\ -H_{\mathrm{T}}\,\mathrm{proj}_{\widehat{\mathbf{D}}(P_{\mathrm{N},k+1},v_{k+1})}(P_{\mathrm{T},k+1} - \rho_{\mathrm{T}} \circ H_{\mathrm{T}}v_{k+1}) \\ \\ H_{\mathrm{N}}\left[P_{\mathrm{N},k+1} - \mathrm{proj}_{\mathrm{IR}^a_+}(P_{\mathrm{N},k+1} - \rho_{\mathrm{N}} \circ (H_{\mathrm{N}}v_{k+1} + e \circ H_{\mathrm{N}}v_k))\right] \\ \\ H_{\mathrm{T}}\left[P_{\mathrm{T},k+1} - \mathrm{proj}_{\widehat{\mathbf{D}}(P_{\mathrm{N},k+1},v_{k+1})}(P_{\mathrm{T},k+1} - \rho_{\mathrm{T}} \circ H_{\mathrm{T}}v_{k+1})\right] \end{bmatrix} = 0. \tag{13.92}$$

Although very technical, the computation of an element of the Clarke generalized gradient is similar to the linear case exposed at the beginning of Sect. 13.6.1. We will not detail the algebraic manipulations. At each iteration of the Newton method, a linear system of the form (13.87) has to be solved. The use of iterative solvers for conjugate gradient solvers can be relevant if some good pre-conditioners are used. This work is done and detailed in Alart & Lebon (1995) where some incomplete LU (ILU) pre-conditioners are used.

13.6.2 Variants and Line-Search Procedure

In Christensen et al. (1998), Christensen & Pang (1998), and Christensen (2000), the authors developed a very similar method to Alart–Curnier's method using the following simplified equation-based reformulation:

$$\Phi(U_{k+1},P_{k+1}) = \begin{bmatrix} -U_{k+1} + \widehat{W}P_{k+1} + U_{\mathrm{free}} \\ \\ P_{\mathrm{N},k+1} - \mathrm{proj}_{\mathrm{IR}^a_+}(P_{\mathrm{N},k+1} - \rho_{\mathrm{N}} \circ (U_{\mathrm{N},k+1} + e \circ U_{\mathrm{N},k})) \\ \\ P_{\mathrm{T},k+1} - \mathrm{proj}_{\mathbf{D}(P_{\mathrm{N},k+1})}(P_{\mathrm{T},k+1} - \rho_{\mathrm{T}} \circ U_{\mathrm{T},k+1}) \end{bmatrix} = 0. \tag{13.93}$$

The other discrepancies with Alart–Curnier's method lie in (a) the use of the semismoothness and the B-differentiability for the justification of the method; and (b) in the introduction of a line-search procedure. The concept of B-differentiability seems to have a poor interest in this context. Indeed, B-differentiable Newton's method (Pang, 1990) is more a conceptual method rather than an efficient numerical method, essentially because the Newton's direction is the solution of a nonlinear system, which is very difficult to solve. Nevertheless, in Christensen & Pang (1998), the semismoothness of the operator Φ is shown. We recall that the semismoothness is a key property for convergence of generalized Newton's method (see Sect. 12.5.4).

One of the original contributions of these works is the introduction of a line-search procedure. As explained in Sect. 12.5.4, nonsmooth Newton's method can be globalized using a line-search procedure based on a merit function such as

$$\Psi(U,P) = \frac{1}{2}\Phi(U,P)^{\mathrm{T}}\Phi(U,P). \tag{13.94}$$

We denote by z^i the current iterate

$$z^i = \begin{bmatrix} U^i \\ P^i \end{bmatrix} \tag{13.95}$$

and by $d^i = z^{i+1} - z^i$ the Newton's direction computed by solving the linear system (13.86). In Christensen et al. (1998), the proposed line search at the iteration k can be described as follows. Let $\alpha^i = \rho^{m^i}$, where $\rho \in (0,1)$ and m^i is the smallest nonnegative integer m for which the following decrease criterion holds:

$$\Psi(z^i + \rho^m d^i) \geqslant (1 - 2\sigma\rho^m)\Psi(z^i), \tag{13.96}$$

where $\sigma \in (0, \frac{1}{2})$ is a given parameter. Once this criterion is satisfied, the next iterate is set to $z^{i+1} = z^k + \alpha^i d^i$.

Unfortunately, there is no clear analysis of the influence of the previous line-search procedure on the global and local convergence of the methods. According to Ferris & Kanzow (2002), a Newton method based on the min function is more difficult to globalize than semismooth Newton method based on other C-function (see Remark 12.56). It would be interesting to confirm this point of view by a thorough analysis.

13.6.3 Other Direct Equation-Based Reformulations

In Park & Kwak (1994) and Leung et al. (1998), another equation-based reformulation of the three-dimensional frictional contact problems is proposed. The key idea is to write a two-dimensional problem in the plane defined by the sliding direction and the normal vector of the local frame at contact. Let us just recall how the two-dimensional frictional problem given for $U_{\mathrm{T}} \in \mathbb{R}$ and $R_{\mathrm{T}} \in \mathbb{R}$

$$\begin{cases} \text{If } U_{\mathrm{T}} = 0 \text{ then } |R_{\mathrm{T}}| \leqslant \mu R_{\mathrm{N}} \\ \text{If } U_{\mathrm{T}} < 0 \text{ then } R_{\mathrm{T}} = \mu R_{\mathrm{N}} \\ \text{If } U_{\mathrm{T}} > 0 \text{ then } R_{\mathrm{T}} = -\mu R_{\mathrm{N}} \end{cases} \tag{13.97}$$

can be reformulated in terms of equations. Using the fact that the unilateral contact[1]

$$\text{If } g(q) \leqslant 0 \text{ then } 0 \leqslant U_{\mathrm{N}} \perp R_{\mathrm{N}} \geqslant 0 \tag{13.98}$$

can be written equivalently as

$$\text{If } g(q) \leqslant 0, \quad \Psi_1(U_{\mathrm{N}}, R_{\mathrm{N}}) = \min(U_{\mathrm{N}}, R_{\mathrm{N}}) = 0, \tag{13.99}$$

Leung et al. (1998) proposed to write the two-dimensional frictional contact problem together with (13.99) as

[1] For the sake of simplicity, we assume here that the motion is smooth. The time-discretized equation below will take care of this fact.

$$\begin{aligned}\Psi_2(U_{\mathrm{T}},R_{\mathrm{T}}) &= U_{\mathrm{T}}+\min(0,\mu \max(0,R_{\mathrm{N}}-U_{\mathrm{N}})+R_{\mathrm{T}}-U_{\mathrm{T}}) \\ &\quad +\max(0,-\mu \max(0,R_{\mathrm{N}}-U_{\mathrm{N}})+R_{\mathrm{T}}-U_{\mathrm{T}})=0\,. \qquad (13.100)\end{aligned}$$

In order to extend the equation-based reformulation of the two-dimensional frictional contact problem (13.99) and (13.100) to the three-dimensional case, the sliding angle θ is introduced as a variable. In the local contact frame, the tangential reaction R_{T} is written as

$$R_{\mathrm{T}}=\begin{bmatrix}R_{\mathrm{T}1}\\ R_{\mathrm{T}2}\end{bmatrix}=R_{\mathrm{S}}\begin{bmatrix}\cos\theta\\ \sin\theta\end{bmatrix} \qquad (13.101)$$

and the tangential velocity as

$$U_{\mathrm{T}}=\begin{bmatrix}U_{\mathrm{T}1}\\ U_{\mathrm{T}2}\end{bmatrix}=U_{\mathrm{S}}\begin{bmatrix}\cos\theta\\ \sin\theta\end{bmatrix}. \qquad (13.102)$$

With this new variable, the two-dimensional model (13.100) can be used directly, that is

$$\Psi_2(U_{\mathrm{S}},R_{\mathrm{S}})=0. \qquad (13.103)$$

The solving procedure is based on a smoothing procedure of the min and the max functions. This approach is very similar to the approach of Chen & Mangasarian (1996) commented in Sect. 12.5.4.

Remark 13.8. Leung et al. (1998) used directly the equation (13.103) in their solving procedure. However, if the sliding angle is defined on the tangential velocity, the variable U_{S} has to be equal to $\|U_{\mathrm{T}}\|$. Therefore the two-dimensional friction model should be reduced to

$$\begin{cases}\text{If } U_{\mathrm{S}}=0 \quad \text{then } |R_{\mathrm{S}}|\leqslant \mu R_{\mathrm{N}}\\ \\ \text{If } U_{\mathrm{S}}>0 \quad \text{then } R_{\mathrm{S}}=-\mu R_{\mathrm{N}}\end{cases} \qquad (13.104)$$

This simplification does not seem to be taken into account in Leung et al. (1998).

In Xuewen et al. (2000), another direct equation-based reformulation is presented. The reformulation is written down as

$$\begin{cases}\Psi_1(U,R)=\min(U_{\mathrm{N}},R_{\mathrm{N}})=0\\ \Psi_2(U,R)=\min(\|U_{\mathrm{T}}\|,\mu R_{\mathrm{N}}-\|R_{\mathrm{T}}\|)\\ \Psi_3(U,R)=|U_{\mathrm{T}1}R_{\mathrm{T}2}-U_{\mathrm{T}2}R_{\mathrm{T}1}|+\max(0,U_{\mathrm{T}1}R_{\mathrm{T}1})\end{cases}\quad . \qquad (13.105)$$

This system of nonsmooth equations is solved by a generalized Newton method with a line-search procedure similar to those presented in Sect. 13.6.2. Some comparisons have been performed with the smoothing method of Leung et al. (1998). Unfortunately, it is quite difficult to figure out some conclusions mainly due to the fact that the conditions of the numerical experiments are not the same.

13.7 Formulation and Resolution as VI/CP

13.7.1 VI/CP Reformulation

Reformulation of the Linear OSNSP $(\mathscr{P}_{\mathrm{L}})$

Let us start with the linear OSNSP $(\mathscr{P}_{\mathrm{L}})$ defined by

$$\begin{cases} U_{k+1} = \widehat{W} P_{k+1} + U_{\mathrm{free}} \\ \left.\begin{array}{l} \widehat{U}^{\alpha}_{k+1} = \left[U^{\alpha}_{\mathrm{N},k+1} + e^{\alpha} U^{\alpha}_{\mathrm{N},k} + \mu^{\alpha}\, ||U^{\alpha}_{\mathrm{T},k+1}||, U^{\alpha}_{\mathrm{T},k+1}\right]^{\mathrm{T}} \\ \mathbf{C}^{\alpha,*} \ni \widehat{U}^{\alpha}_{k+1} \perp P^{\alpha}_{k+1} \in \mathbf{C}^{\alpha} \end{array}\right\} \forall \alpha \in I_a(\tilde{q}_{k+1}). \end{cases} \tag{13.106}$$

With the following definitions for the Cartesian product of Coulomb cones,

$$\mathbf{C} = \prod_{\alpha \in I_a(\tilde{q}_{k+1})} \mathbf{C}^{\alpha}, \qquad \mathbf{C}^* = \prod_{\alpha \in I_a(\tilde{q}_{k+1})} \mathbf{C}^{\alpha,*}, \tag{13.107}$$

the following CP over cones can be written

$$\begin{cases} \widehat{U}_{k+1} = \widehat{W} P_{k+1} + U_{\mathrm{free}} - G(P_{k+1}) \\ \mathbf{C}^* \ni \widehat{U}_{k+1} \perp P_{k+1} \in \mathbf{C}. \end{cases} \tag{13.108}$$

We assume that the vectors U_{k+1} and P_{k+1} are ordered in a suitable manner to satisfy the cone inclusion. The function $G : \mathbb{R}^{3a} \to \mathbb{R}^{3a}$ defined by

$$G(P) = \left[\left[\mu^{\alpha} \| [\widehat{W} P + U_{\mathrm{free}}]^{\alpha}_{\mathrm{T}} \| + e^{\alpha} U^{\alpha}_{\mathrm{N},k}, 0\right], \alpha \in I_a(\tilde{q}_{k+1}) \right]^{\mathrm{T}} \tag{13.109}$$

is a nonlinear and nonsmooth function of P.

The formulation in terms of VI is straightforward due to the equivalence between VIs and CPs. The linear OSNSP $(\mathscr{P}_{\mathrm{L}})$ is equivalent to the following VI

$$(\widehat{W} P_{k+1} + U_{\mathrm{free}} - G(P_{k+1}))^{\mathrm{T}} (P^* - P_{k+1}) \geqslant 0, \quad \text{for all } P^* \in \mathbf{C}. \tag{13.110}$$

Reformulation of the Mixed Nonlinear OSNSP $(\mathscr{P}_{\mathrm{MNL}})$

Let us consider now the mixed nonlinear OSNSP $(\mathscr{P}_{\mathrm{MNL}})$ given by

$$\begin{cases} \mathscr{R}(v_{k+1}) = p_{k+1} = \displaystyle\sum_{\alpha \in I_a(\tilde{q}_{k+1})} p^{\alpha}_{k+1} \\[2ex] U^{\alpha}_{k+1} = H^{\alpha,\mathrm{T}}(q_k+1)\, v_{k+1} \\[2ex] p^{\alpha}_{k+1} = H^{\alpha}(q_k+1)\, P^{\alpha}_{k+1} \\[2ex] \left.\begin{array}{l} \widehat{U}^{\alpha}_{k+1} = \left[U^{\alpha}_{\mathrm{N},k+1} + e^{\alpha} U^{\alpha}_{\mathrm{N},k} + \mu^{\alpha}\, \|U^{\alpha}_{\mathrm{T},k+1}\|, U^{\alpha}_{\mathrm{T},k+1}\right]^{\mathrm{T}} \\[2ex] \mathbf{C}^{\alpha,*} \ni \widehat{U}^{\alpha}_{k+1} \perp P^{\alpha}_{k+1} \in \mathbf{C}^{\alpha} \end{array}\right\} \forall \alpha \in I_a(\tilde{q}_{k+1}). \end{cases} \tag{13.111}$$

Introducing the Cartesian products of the Coulomb cones, we obtain the following CP

$$\begin{cases} \mathscr{R}(v_{k+1}) = \displaystyle\sum_{\alpha \in I_a(\tilde{q}_{k+1})} H^{\alpha}(q_k+1)\, P^{\alpha}_{k+1} \\[2ex] \mathbf{C}^* \ni g(v_{k+1}) \perp P_{k+1} \in \mathbf{C}. \end{cases} \tag{13.112}$$

The function g: $\mathbb{R}^n \rightarrow \mathbb{R}^n$ defined by

$$\begin{aligned} g(v_{k+1}) = \Bigg[& [H^{\alpha,\mathrm{T}}_{\mathrm{N}}(q_k+1)\, v_{k+1} + e^{\alpha} U^{\alpha}_{\mathrm{N},k} + \mu^{\alpha}\, \|H^{\alpha,\mathrm{T}}_{\mathrm{T}}(q_k+1)\, v_{k+1}\|, \\ & H^{\alpha,\mathrm{T}}_{\mathrm{T}}(q_k+1)\, v_{k+1}], \alpha \in I_a(\tilde{q}_{k+1}) \Bigg]^{\mathrm{T}}. \end{aligned} \tag{13.113}$$

The CP (13.112) is kind of mixed CP over cones and its reformulation in terms of VI is not straightforward. For the sake of simplicity, this question is left open focusing on the more linear standard case.

13.7.2 Projection-type Methods

The projection-type methods for VIs presented in Sect. 12.6.6 can be applied to the VI (13.110). We recall that the most basic projection-type method for the VI

$$F(P)^{\mathrm{T}}(P^* - P) \geqslant 0, \quad \text{for all } P^* \in \mathbf{C} \tag{13.114}$$

is a fixed-point method such that

$$P^{i+1} = \mathrm{proj}_{\mathbf{C}}(P^i - \rho F(P^i)), \quad \rho > 0. \tag{13.115}$$

For the three-dimensional frictional contact, i.e., $F(P) = \widehat{W}P + U_{\mathrm{free}} - G(P)$, it yields

$$P^{i+1} = \mathrm{proj}_{\mathbf{C}}\left[(I - \rho\widehat{W})P^i - \rho G(P^i) + \rho U_{\mathrm{free}}\right], \quad \rho > 0\,. \tag{13.116}$$

This method has been initiated by De Saxcé & Feng (1991) and extensively tested in Feng (1991, 1995) and De Saxcé & Feng (1998). The authors term this method an Uzawa's method for solving the variational inequality. Indeed, as we have seen in Sect. 12.6.4, a VI can be viewed as a saddle-point problem. The standard Uzawa's method consists in alternative iterative in the primal and the dual formulations to find the saddle-point of a Lagrangian function.

The extra-gradient method can also be directly applied, yielding the following iterative scheme:

$$P^{i+1} = \mathrm{proj}_{\mathbf{C}}\left[P^i - \rho\, F\left(\mathrm{proj}_{\mathbf{C}}(P^i - \rho F(P^i))\right)\right] . \tag{13.117}$$

More generally, it is also possible to apply the hyperplane projection method and the splitting plus projection methods for VIs. All these methods show a good behavior on practical large-scale problems in terms of robustness and computational cost. Indeed, the cost of the projection onto the Cartesian product of second-order cone is cheap. Nevertheless, the convergence rate is quite slow.

Unfortunately, most of the known convergence proofs for these algorithms are based on monotonicity-like assumptions. A less restrictive assumption is the pseudo-monotonicity for the hyperplane projection method. Therefore, standard convergence proofs cannot be applied to the special case of the three-dimensional frictional contact. It should be, however, possible to prove the convergence by fixed-point arguments under suitable assumptions on the values of μ.

13.7.3 Fixed-Point Iterations on the Friction Threshold and Ad Hoc Projection Methods

Numerous projection-type methods has already been developed in the literature. Most of them are based on a fixed-point iteration method procedure on the friction threshold and a projection onto the Tresca friction cylinder.

Tresca's Friction

To be more precise, the so-called Tresca friction model can be invoked

$$\begin{cases} \text{If } U_{\mathrm{T}} = 0 \text{ then } ||R_{\mathrm{T}}|| \leqslant \theta \\ \\ \text{If } U_{\mathrm{T}} \neq 0 \text{ then } ||R_{\mathrm{T}}(t)|| = \theta, \quad \text{and } \exists a \geqslant 0 \text{ such that } R_{\mathrm{T}}(t) = -aU_{\mathrm{T}}(t) \end{cases} \tag{13.118}$$

where θ is the friction threshold. The Tresca model is equivalent to the following inclusions

$$-U_{\mathrm{T}} \in \partial \psi_{\mathbf{D}_\theta}(R_{\mathrm{T}}), \quad R_{\mathrm{T}} \in \partial \psi^*_{\mathbf{D}_\theta}(-U_{\mathrm{T}}) \tag{13.119}$$

where $\mathbf{D}_\theta$ is a disk with radius θ. The fact that the radius of the friction disk does not depend on R_{N} allows one to state the whole three-dimensional contact friction problem for the Tresca model as

$$-U \in \partial \psi_{\mathbf{T}_\theta}(R), \quad \text{or } R \in \partial \psi^*_{\mathbf{T}_\theta}(-U) \tag{13.120}$$

where $\mathbf{T}_\theta = \mathrm{I\!R}_+ \times \mathbf{D}_\theta$ is the friction cylinder or the Tresca cylinder. Thus it is an associated law of friction.

With the Tresca friction model and the Newton impact law, the linear OSNSP $(\mathscr{P}_\mathrm{L})$ can be written as

$$\begin{cases} U_{k+1} = \widehat{W} P_{k+1} + U_\mathrm{free} \\ -\left(U^\alpha_{k+1} + \left[e^\alpha \circ U^\alpha_{\mathrm{N},k}, 0\right]^\mathrm{T}\right) \in \partial \psi_{\mathbf{T}^\alpha_\theta}(P^\alpha_{k+1}), \quad \forall \alpha \in I_a(\tilde{q}_{k+1}). \end{cases} \tag{13.121}$$

Introducing the Cartesian product of the Tresca cylinders such that

$$\mathbf{T} = \prod_{\alpha \in I_a(\tilde{q}_{k+1})} \mathbf{T}^\alpha_\theta, \qquad \mathbf{T}^* = \prod_{\alpha \in I_a(\tilde{q}_{k+1})} \mathbf{T}^{\alpha,*}_\theta, \tag{13.122}$$

the following inclusion can be written for the linear OSNSP $(\mathscr{P}_\mathrm{L})$ with Tresca's friction

$$-\left(\widehat{W} P_{k+1} + U^e_\mathrm{free}\right) \in \partial \psi_\mathbf{T}(P_{k+1}) \tag{13.123}$$

where $U^e_\mathrm{free} = U_\mathrm{free} + \left[\left[e^\alpha \circ U^\alpha_{\mathrm{N},k}, 0\right], \quad \alpha \in I_a(\tilde{q}_{k+1})\right]^\mathrm{T}$. The relative velocity can be deduced by

$$U_{k+1} = \widehat{W} P_{k+1} + U_\mathrm{free}. \tag{13.124}$$

The inclusion (13.123) can be reformulated as the following VI

$$\left(\widehat{W} P_{k+1} + U^e_\mathrm{free}\right)^\mathrm{T} (P^* - P_{k+1}), \quad \text{for all } P^* \in \mathbf{T}. \tag{13.125}$$

The interest of such a formulation lies in the fact that the function $\widehat{W} P_{k+1} + U^e_\mathrm{free}$ is a gradient mapping if the matrix $\widehat{W}$ is symmetric. Under the assumption that $\widehat{W}$ is a symmetric PSD matrix, we may consider the following minimization problem

$$\begin{array}{ll} \text{minimize} & \dfrac{1}{2} P^\mathrm{T}_{k+1} \widehat{W} P_{k+1} + P^\mathrm{T}_{k+1} U^e_\mathrm{free} \\ \text{subject to} & P_{k+1} \in \mathbf{T} \end{array} \tag{13.126}$$

Indeed, the VI (13.125) or the inclusion (13.123) are the first-order optimality conditions of the constrained minimization problem (13.126). Under the convexity assumption of the objective function and the feasible set, i.e., in our case the PSD property of $\widehat{W}$, these KKT conditions are equivalent to the minimization problem.

Fixed-Point Methods on the Friction Threshold

A key idea of numerous projection-type methods dedicated to the Coulomb's friction problem is to perform a fixed-point iteration on the friction threshold of a Tresca's friction problem. The general algorithm can be stated as in Algorithm 17.

The methods presented in the literature differ in the choice of the numerical method for solving the Tresca minimization problem (13.126) and how this numerical method is driven. We will list in the next paragraphs some of the most widespread choices.

Gradient Projection Methods

In Jean & Touzot (1988), Horkay et al. (1989) and Mehrez (1991), standard gradient projection methods and Rosen's method (see Sect. 12.2.3.2) are used to solve (13.126). Several strategies are developed to perform this step into the fixed-point procedure. Either the problem (13.126) is solved up to a given accuracy or very roughly with one or a few iterations of the gradient projection solver.

Splitting Methods

As we said earlier in Sect. 13.5.1, the idea of using (block) splitting, i.e., PSOR methods, to solve the frictionless contact problem date back to the pioneering works of Glowinski et al. (1976). In Lebon & Raous (1992), this method is used to solve the Tresca two-dimensional frictional contact problem similar to (13.126). We will see further in Sect. 13.7.4 that another splitting method has been developed for the three-dimensional frictional contact problem where the coupling between the fixed-point algorithm and a PSOR method is deeper.

Algorithm 17 Fixed-point algorithm on the friction threshold

```
Require: $\mathsf{W}, \mathsf{U}^{\mathsf{e}}_{\mathrm{free}}, \mathsf{U}_{\mathrm{free}}, \mu$
Require: $\mathsf{P}^{0}_{\mathsf{k}+1}$
Ensure: $\mathsf{U}_{\mathsf{k}+1}, \mathsf{P}_{\mathsf{k}+1}$ solution of the OSNSP $(\mathscr{P}_L)$
  $\mathsf{i} \leftarrow 0$
  while error $>$ tol do
    $\theta^{\mathsf{i}} \leftarrow \mu \circ \mathsf{P}^{\mathsf{i}}_{\mathrm{N},\mathsf{k}+1}$
    Solve (possibly inexactly) the minimization problem (13.126) that is
        $$\mathsf{P}^{\mathsf{i}+1}_{\mathsf{k}+1} \leftarrow \operatorname*{argmin}_{\mathsf{P} \in \mathbf{T}_{\theta^{\mathsf{i}}}} \frac{1}{2} \mathsf{P}^{\mathsf{T}} \widehat{\mathsf{W}} \mathsf{P} + \mathsf{P}^{\mathsf{T}} \mathsf{U}^{\mathsf{e}}_{\mathrm{free}}$$
    $\mathsf{i} \leftarrow \mathsf{i} + 1$
    Evaluate error.
  end while
  $\mathsf{U}_{\mathsf{k}+1} \leftarrow \widehat{\mathsf{W}} \mathsf{P}_{\mathsf{k}+1} + \mathsf{U}_{\mathrm{free}}$
```

Comments

The acceleration methods (Mittelmann, 1981a,b; Calamai & More, 1987; Moré & Toraldo, 1991; Wright, 1990) used to improve the rate of convergence of the gradient projection method should be very interesting to implement in the context of the two-dimensional friction, in which the constraints are merely bound constraints. In the three-dimensional context, the advantage of these methods is less obvious.

The methods are often justified using the term of gradient projection method of gradient with splitting. As we said before, there is no straightforward minimization principle associated with the three-dimensional frictional contact problem. The term gradient-like methods is then ambiguous, because it refers to the minimization of an objective function. In any case, the function $F(x) = \frac{1}{2}P^{\mathrm{T}}\widehat{W}P + P^{\mathrm{T}}U_{\mathrm{free}}$ is a function that we have to minimize to find the right contact impulses P.

13.7.4 A Clever Block Splitting: the Nonsmooth Gauss–Seidel (NSGS) Approach

In this section, we outline a specification of the general splitting method for Affine Variational Inequality (AVI) or CP to the frictional contact problem. The so-called nonsmooth Gauss–Seidel initiated by Jean & Moreau (1991, 1992) and further studied in Jourdan et al. (1998a) is based on the following two remarks:

1. The Delassus operator W is usually sparse block structured in multibody dynamics (see the formulation (13.8) and (13.9)). The splitting is chosen to take advantage of this structure.
2. Each subproblem of frictionless/frictional contact for a single contact α can be either analytically solved or easily approximated.

The splitting algorithm can be stated using (13.9)

$$\begin{cases} U_{k+1}^{\alpha,i+1} - \widehat{W}^{\alpha\alpha} P_{k+1}^{\alpha,i+1} = U_{\mathrm{free}}^{\alpha} + \sum\limits_{\beta<\alpha} \widehat{W}^{\alpha\beta} P_{k+1}^{\beta,i+1} + \sum\limits_{\beta>\alpha} \widehat{W}^{\alpha\beta} P_{k+1}^{\beta,i} \\[2ex] \widehat{U}_{k+1}^{\alpha,i+1} = \left[U_{\mathrm{N}}^{\alpha,i+1} + e^{\alpha} U_{\mathrm{N},k}^{\alpha} + \mu^{\alpha}\, ||U_{\mathrm{T},k+1}^{\alpha,i+1}||, U_{\mathrm{T},k+1}^{\alpha,i+1}\right]^{\mathrm{T}} \\[2ex] \mathbf{C}^{\alpha,*} \ni \widehat{U}_{k+1}^{\alpha,i+1} \perp P_{k+1}^{\alpha,i+1} \in \mathbf{C}^{\alpha} \end{cases} \tag{13.127}$$

for all $\alpha, \beta \in I_a(\tilde{q}_{k+1})$. The index i corresponds to the iteration in the Gauss–Seidel method. A parameter of relaxation ω can be introduced leading to the Non Smooth Gauss–Seidel (NSGS) method with overrelaxation.

Let us now give some details on the resolution of the local problem which can be defined by

$$\begin{cases} U^{\alpha,i+1}_{k+1} = \widehat{W}^{\alpha\alpha} P^{\alpha,i+1}_{k+1} + q^{\alpha,i+1} \\ \widehat{U}^{\alpha,i+1}_{k+1} = \left[U^{\alpha,i+1}_{\mathrm{N}} + e^{\alpha} U^{\alpha}_{\mathrm{N},k} + \mu^{\alpha} \,||U^{\alpha,i+1}_{\mathrm{T},k+1}||, U^{\alpha,i+1}_{\mathrm{T},k+1} \right]^{\mathrm{T}} \\ \mathbf{C}^{\alpha,*} \ni \widehat{U}^{\alpha,i+1}_{k+1} \perp P^{\alpha,i+1}_{k+1} \in \mathbf{C}^{\alpha} \end{cases} \tag{13.128}$$

where

$$q^{\alpha,i+1} = U^{\alpha}_{\mathrm{free}} + \sum_{\beta<\alpha} \widehat{W}_{\alpha\beta} P^{\beta,i+1}_{k+1} + \sum_{\beta>\alpha} \widehat{W}_{\alpha\beta} P^{\beta,i}_{k+1}$$

is a known value at the step α of the iteration i.

Analytical Solutions for the Frictionless and the Two-Dimensional Case

In the frictionless case and in the 2-dimensional case, the subproblem (13.128) can be solved analytically. Indeed, for the frictionless case, the solution is given by

$$P^{\alpha,i+1}_{\mathrm{N},k+1} = \max\left(0, -\frac{q^{\alpha,i+1}_{\mathrm{N}} + e^{\alpha} U^{\alpha}_{\mathrm{N},k}}{\widehat{W}^{\alpha\alpha}_{\mathrm{NN}}}\right), \qquad P^{\alpha,i+1}_{\mathrm{T},k+1} = 0 \tag{13.129}$$

Algorithm 18 Analytical resolution of the two-dimensional frictional contact subproblem

```
Require: $\widehat{W}^{\alpha\alpha}, q^{\alpha,i+1}, \mu^{\alpha}$
Ensure: $U^{\alpha,i+1}_{k+1}, P^{\alpha,i+1}_{k+1}$ solution of (13.128) in 2D

  if $q^{\alpha,i+1}_{N} + e^{\alpha} U^{\alpha}_{N,k} > 0$ then
    $P^{\alpha,i+1}_{k+1} \leftarrow 0$
  else
    $P^{\alpha,i+1}_{k+1} \leftarrow -\widehat{W}^{\alpha\alpha,-1} q^{\alpha,i+1}$
    if $P^{\alpha,i+1}_{T,k+1} + \mu^{\alpha} P^{\alpha,i+1}_{N,k+1} > 0$ then
      $P^{\alpha,i+1}_{N,k+1} \leftarrow -\frac{q^{\alpha,i+1}_{N} + e^{\alpha} U^{\alpha}_{N,k}}{\widehat{W}^{\alpha\alpha}_{NN} - \mu^{\alpha} \widehat{W}^{\alpha\alpha}_{NT}}$
      $P^{\alpha,i+1}_{T,k+1} \leftarrow -\mu^{\alpha} P^{\alpha,i+1}_{N,k+1}$
    end if
    if $P^{\alpha,i+1}_{T,k+1} - \mu^{\alpha} P^{\alpha,i+1}_{N,k+1} < 0$ then
      $P^{\alpha,i+1}_{N,k+1} \leftarrow -\frac{q^{\alpha,i+1}_{N} + e^{\alpha} U^{\alpha}_{N,k}}{\widehat{W}^{\alpha\alpha}_{NN} + \mu^{\alpha} \widehat{W}^{\alpha\alpha}_{NT}}$
      $P^{\alpha,i+1}_{T,k+1} \leftarrow +\mu^{\alpha} P^{\alpha,i+1}_{N,k+1}$
    end if
  end if
```

and the relative velocity is found as

$$U_{k+1}^{\alpha,i+1} = \widehat{W}^{\alpha\alpha} P_{k+1}^{\alpha,i+1} + q^{\alpha,i+1}. \tag{13.130}$$

In (13.129), the value $q_{\rm N}^{\alpha,i+1}$ denotes as usual $q^{\alpha,i+1^{\rm T}} e_{\rm N}$ with $e_{\rm N} = [1\ 0\ 0]^{\rm T}$.

In the 2-dimensional case, an analysis can be performed assuming that there exists a unique solution to the problem given by the intersection of graphs as in Sect. 1.2. The solution is given by Algorithm 18. In Mitsopoulou & Doudoumis (1987, 1988), a similar resolution is used to solve the 2-dimensional frictional case. It is assumed that the solution exists and is unique.

Approximate 3D Resolution Based on a Projection onto the Friction Disk

In the 3-dimensional case, the analytical solution of the local subproblem is not known. Various strategies can be implemented. We start with several strategies based on the projection onto the friction disk.

This approximate solution is found by mimicking Algorithm 18 for the 3-dimensional case. The procedure is given in Algorithm 19. The solution of this algorithm is not an exact solution of the problem, at least when the projection step is performed. Nevertheless, the overall algorithm is able to converge with this approximate solution.

Approximate 3D Resolution Based on a Projection onto the Friction Cone

A direct projection onto the friction cone can be implemented. It amounts to use the projection-type methods for VI described in Sect. 13.7.2 that is

$$P_{k+1}^{\alpha,i+1,j+1} = \mathrm{proj}_{\mathbf{C}} \left[(I - \rho \widehat{W}) P_{k+1}^{\alpha,i+1,j} - \rho G^{\alpha,i+1}(P_{k+1}^{\alpha,i+1,j}) + q^{\alpha,i+1} \right], \quad \rho > 0 \tag{13.131}$$

where j stands for the iteration index of the projection-type methods. The function $G^{\alpha,i+1}(\cdot)$ is the restriction of the function $G(\cdot)$ defined in (13.109) for a single contact α assuming the other contact values are known, that is:

$$G^{\alpha,i+1}(P) = \left[\mu^{\alpha} \| [\widehat{W}^{\alpha\alpha} P + q^{\alpha,i+1}]_{\rm T} \| + e^{\alpha} U_{{\rm N},k}^{\alpha}, 0 \right]^{T}. \tag{13.132}$$

Two strategies can be proposed:

1. Perform just one iteration of (13.131). One gets a dedicated splitting method for the VI (13.110) improving therefore the method (13.116).
2. Perform iterations of (13.131) up to a given accuracy to obtain an exact solution of the subproblem.

Approximate 3D Resolution Based on General Methods

Finally, we can invoke to solve the contact subproblem any general solver presented in this chapter. We think especially of the nonsmooth Newton method of Alart & Curnier (1991) and its variant. This strategy is used in Jourdan et al. (1998a) and Jean (1999).

Algorithm 19 Approximate solution of the three-dimensional frictional contact subproblem

```
Require: $\widehat{W}^{\alpha\alpha}, q^{\alpha,i+1}, \mu^\alpha$
Ensure: $U^{\alpha,i+1}_{k+1}, P^{\alpha,i+1}_{k+1}$ approximate solution of (13.128) in 3D

  if $q_{N}^{\alpha,i+1} + e^\alpha U^\alpha_{N,k} > 0$ then
    $P^{\alpha,i+1}_{k+1} \leftarrow 0$
  else
    $P^{\alpha,i+1}_{k+1} \leftarrow -\widehat{W}^{\alpha\alpha,-1} q^{\alpha,i+1}$
    if $\|P^{\alpha,i+1}_{T,k+1}\| > \mu |P^{\alpha,i+1}_{N,k+1}|$ then
      $P^{\alpha,i+1}_{T,k+1} \leftarrow \mathrm{proj}_{\mathbf{D}(\mu^\alpha P^{\alpha,i+1}_{N,k+1})}(P^{\alpha,i+1}_{T,k+1})$
    end if
  end if
```

Comments

The NSGS solver has been proved to be very robust and efficient on a large collection of heterogeneous problems (see Acary & Jean, 2000; Acary, 2001; Jean, 1999; Moreau, 1994b, 1999; Renouf et al., 2004; Saussine et al., 2004a, 2006). Although it suffers from a slow convergence rate (usually linear), the algorithm has shown to be robust and parsimonious with the memory. The solver has also been used in an event-driven framework by Abadie (2000).

13.7.5 Newton's Method for VI

We end up this section with the description of a nonsmooth Newton method for the VI (13.110). As explained in Sect. 12.6.6, it is possible to design Newton's method for VI by reformulating the VI or CP as a set of nonsmooth equations by means of the natural map. Let us start with such a formulation. A vector P solves (13.110) if and only if the following nonsmooth equations holds

$$\mathbf{F}^{nat}_C(P) = P - \mathrm{proj}_{\mathbf{C}}((I - \widehat{W})P + U_{\text{free}} - G(P)) = 0 \tag{13.133}$$

or equivalently

$$\Phi(P) = P - \mathrm{proj}_{\mathbf{C}}((I - \rho\widehat{W})P + \rho U_{\text{free}} - \rho G(P)) = 0, \quad \rho > 0. \tag{13.134}$$

The key idea of nonsmooth Newton method is to use a linear approximation of the nonsmooth map $\Phi(P)$. This can be done by using the Clarke generalized subgradient (see Sect. 13.6) to generate iterates such that

$$P^{i+1} = P^i - H^{i-1}\Phi(P^i), \text{ with } H^{i-1} \in \partial\Phi(P^i). \tag{13.135}$$

Part IV

The Siconos Software: Implementation and Examples

14

The Siconos Platform

14.1 Introduction

The Siconos Platform is a scientific computing software dedicated to the modeling, simulation, control, and analysis of nonsmooth dynamical systems (NSDS), mainly developed in the Bipop team-project at INRIA[1] in Grenoble, France, and distributed under GPL GNU license.

Siconos aims at providing a general and common tool for nonsmooth problems in various scientific fields like applied mathematics, mechanics, robotics, electrical circuits, and so on. However, the platform is not supposed to re-implement the existing dedicated tools already used for the modeling of specific systems, but to possibly integrate them. For instance, strong collaborations exist with HuMAns (humanoid motion modeling and control[2]) or LMGC90 (multibody contact mechanics[3]) software packages.

14.2 An Insight into Siconos

The present part is dedicated to a short presentation of the general writing process for a problem treated with Siconos, through a simple example. The point is to introduce the main functionalities, the main steps required to model and simulate the systems behavior, before going more into details in Sect. 14.3, where the NSDS will be described.

The chosen example is a four-diode bridge wave rectifier as shown in Fig. 14.1.

An LC oscillator, initialized with a given voltage across the capacitor and a null current through the inductor, provides the energy to a load resistance through a full-wave rectifier consisting of a four ideal diodes bridge. Both waves of the oscillating

[1] The French National Institute for Research in Computer Science and Control (`http://bipop.inrialpes.fr`).

[2] `http://bipop.inrialpes.fr/software/humans/index.html`.

[3] `http://www.lmgc.univ-montp2.fr~/\~dubois/LMGC90/`.

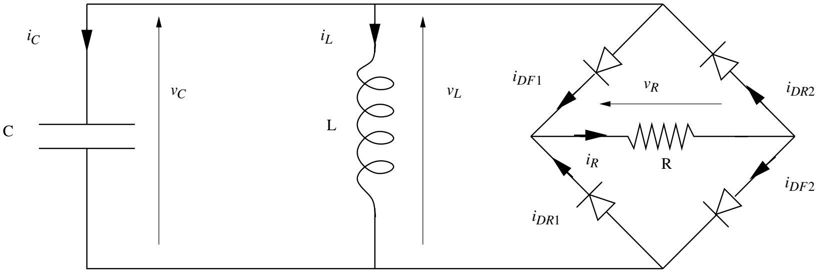


Fig. 14.1. A four-diode bridge wave rectifier

voltage across the LC are provided to the resistor with current flowing always in the same direction. The energy is dissipated into the resistor and results in a damped oscillation.

One of the ways to define a problem with SICONOS consists in writing a C++ file. In the following, for the diode bridge example, only snippets of the C++ commands will be given, just to enlighten the main steps. It is noteworthy that one can also use an XML description as shown in the bouncing ball example in Sect. 14.4.1 or the Python interface.

14.2.1 Step 1. Building a Nonsmooth Dynamical System

In the present case, the oscillator is a time-invariant linear dynamical system, and using the Kirchhoff current and voltage laws and branch constitutive equations, its dynamics is written as (see Fig. 14.1 for the notation)

$$\begin{bmatrix} \dot{v}_L \\ \dot{i}_L \end{bmatrix} = \begin{bmatrix} 0 & -\dfrac{1}{C} \\ \dfrac{1}{L} & 0 \end{bmatrix} \cdot \begin{bmatrix} v_L \\ i_L \end{bmatrix} + \begin{bmatrix} 0\ 0 & -\dfrac{1}{C} & \dfrac{1}{C} \\ 0\ 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} -v_{\mathrm{DR1}} \\ -v_{\mathrm{DF2}} \\ i_{\mathrm{DF1}} \\ i_{\mathrm{DR2}} \end{bmatrix}. \tag{14.1}$$

If we denote

$$x = \begin{bmatrix} \dot{v}_L \\ \dot{i}_L \end{bmatrix}, \quad \lambda = \begin{bmatrix} -v_{\mathrm{DR1}} \\ -v_{\mathrm{DF2}} \\ i_{\mathrm{DF1}} \\ i_{\mathrm{DR2}} \end{bmatrix}, \quad A = \begin{bmatrix} 0 & \dfrac{-1}{C} \\ \dfrac{1}{L} & 0 \end{bmatrix}, \quad r = \begin{bmatrix} 0\ 0 & \dfrac{-1}{C} & \dfrac{1}{C} \\ 0\ 0 & 0 & 0 \end{bmatrix} .\lambda \tag{14.2}$$

the dynamical system (14.1) results in

$$\dot{x} = Ax + r. \tag{14.3}$$

The first step of any SICONOS problem is to define and build some `DynamicalSystemobjects` objects. The corresponding command lines to build a `FirstOrderLinearTIDS` object are:

```
// User-defined parameters
unsigned int ndof = 2;  // number of degrees of freedom of
your system
double Lvalue = 1e-2;  // inductance
double Cvalue = 1e-6;  // capacitance
double Rvalue = 1e3;   // resistance
double Vinit = 10.0;   // initial voltage
// DynamicalSystem(s)
SimpleMatrix A(ndof,ndof);  // All components of A are
automatically  set to 0.
A(0,1) = -1.0/Cvalue;
A(1,0) = 1.0/Lvalue;
// initial conditions vector
SimpleVector x0(ndof);
x0(0) = Vinit;
// Build a First Order Linear and Time Invariant Dynamical
System
//           using A matrix and x0 as initial state.
FirstOrderLinearTIDS * oscillator = new FirstOrderLinearTIDS
(1,x0,A);
```

The suffix DS to the name of a class such as the FirstOrderLinearTIDS object means that this class inherits from the general class of DynamicalSystem.

Thereafter, it is necessary to define the way the previously defined dynamical systems will interact together. This is the role of the Interaction object composed of a Relation object, a set of algebraic equations, and of a NonSmoothLaw object.

The linear relations between voltage and current inside the circuit are given by

$$\begin{bmatrix} i_{\mathrm{DR1}} \\ i_{\mathrm{DF2}} \\ -v_{\mathrm{DF1}} \\ -v_{\mathrm{DR2}} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ -1 & 0 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} v_L \\ i_L \end{bmatrix} + \begin{bmatrix} \frac{1}{R} & \frac{1}{R} & -1 & 0 \\ \frac{1}{R} & \frac{1}{R} & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} -v_{\mathrm{DR1}} \\ -v_{\mathrm{DF2}} \\ i_{\mathrm{DF1}} \\ i_{\mathrm{DR2}} \end{bmatrix}, \tag{14.4}$$

which can be stated by the linear equation

$$y = Cx + D\lambda \tag{14.5}$$

with

$$y = \begin{bmatrix} i_{\mathrm{DR1}} \\ i_{\mathrm{DF2}} \\ -v_{\mathrm{DF1}} \\ -v_{\mathrm{DR2}} \end{bmatrix}, \quad D = \begin{bmatrix} \frac{1}{R} & \frac{1}{R} & -1 & 0 \\ \frac{1}{R} & \frac{1}{R} & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad \lambda = \begin{bmatrix} -v_{\mathrm{DR1}} \\ -v_{\mathrm{DF2}} \\ i_{\mathrm{DF1}} \\ i_{\mathrm{DR2}} \end{bmatrix}. \tag{14.6}$$

Completed with the relation between r and λ (see (14.2)) it results in a linear equation as

$$r = B\lambda. \tag{14.7}$$

This corresponds to a SICONOS FirstOrderLinearTIR object, i.e., a linear and time-invariant coefficients relation. The corresponding code is as follows:

```
// -- Interaction --
// - Relations -
unsigned int ninter = 4; // dimension of your Interaction
= size of y and lambda vectors
SimpleMatrix B(ndof,ninter);
B(0,2) =-1.0/Cvalue ;
B(0,3) = 1.0/Cvalue;
SimpleMatrix C(ninter,ndof);
C(2,0) = -1.0;
C(3,0) = 1.0;
// the Relation:
FirstOrderLinearTIR * myRelation = new FirstOrderLinearTIR
 (C,B);
SimpleMatrix D(ninter,ninter); D(0,0) = 1.0/Rvalue;
D(0,1) = 1.0/Rvalue; D(0,2) = -1.0; D(1,0) = 1.0/Rvalue;
D(1,1) = 1.0/Rvalue; D(1,3) = -1.0; (2,0) = 1.0; D(3,1)=1.0;
myRelation->setD(D);
```

To complete the `Interaction` object, a nonsmooth law is needed to define what the behavior will be when a nonsmooth event occurs.

Thus the behavior of each diode of the bridge, supposed to be ideal, can be described with a complementarity condition between current and reverse voltage (variables (y,λ)). Depending on the diode position in the bridge, y stands for the reverse voltage across the diode or for the diode current. Then, the complementarity conditions, results of the ideal diodes characteristics, are given by

$$\begin{array}{l} 0 \leqslant -v_{\mathrm{DR1}} \perp i_{\mathrm{DR1}} \geqslant 0 \\ 0 \leqslant -v_{\mathrm{DF2}} \perp i_{\mathrm{DF2}} \geqslant 0 \\ 0 \leqslant i_{\mathrm{DF1}} \perp -v_{\mathrm{DF1}} \geqslant 0 \\ 0 \leqslant i_{\mathrm{DR2}} \perp -v_{\mathrm{DR2}} \geqslant 0 \end{array} \Longleftrightarrow 0 \leqslant y \perp \lambda \geqslant 0, \tag{14.8}$$

which corresponds to a `ComplementarityConditionNSL` object which is an inherited class form of the `NonSmoothLaw` class. The SICONOS code is as follows:

```
   // NonSmoothLaw definition
   unsigned int nslawSize = 4;
   NonSmoothLaw * myNslaw = new ComplementarityConditionNSL
   (nslawSize) ;
```

The `Interaction` is built using the concerned `DynamicalSystem`, the `Relation`, and the `NonSmoothLaw` defined above:

```
   // A name and a id-number for the Interaction
   string nameInter = "InterDiodeBridge";
   unsigned int numInter = 1;
   unsigned int ninter = 4; // ninter is the size of y
   Interaction* myInteraction = new Interaction(nameInter,
   allDS, numInter,ninter, myNslaw, myRelation);
```

When the `DynamicalSystem` and `Interaction` have been clearly defined, they are gathered into a NSDS:

```
// NonSmoothDynamicalSystem construction
NonSmoothDynamicalSystem* myNSDS = new NonSmoothDynamical
System (oscillator,myInteraction);
```

Finally the NSDS is inserted into a `Model`, an object that will link the NSDS to the strategy of simulation. It also defines the time boundaries of the simulation:

```
// Model construction
double t0 = 0; // Initial time
double T = 10; // Total simulation time
Model * DiodeBridge = new Model(t0,T);
// The pre-built NSDS is linked to the DiodeBridge Model.
DiodeBridge->setNonSmoothDynamicalSystemPtr(myNSDS);
```

From this point, the diodes bridge system is completely defined by the `NonSmoothDynamicalSystem` object named `myNSDS` and handled by the `Model` object `DiodeBridge`. In the next section, a strategy of simulation will be defined and applied to this model.

14.2.2 Step 2. Simulation Strategy Definition

It is now necessary to define the way the dynamical behavior of the `NonSmoothDynamicalSystem` will be computed. This is the role of `Simulation` class. In Siconos, two different strategies of simulation are available: the time-stepping schemes or the event-driven algorithms. To be complete, a `Simulation` object requires

- a discretization of the considered time interval of study,
- a time-integration method for the dynamics,
- a way to formalize and solve the possibly nonsmooth problems.

For the diode bridge example, the Moreau's time-stepping scheme is used (Sect. 9.4), where the integration of the equations over the time steps is based on a θ-method. The nonsmooth problem is written as an LCP and solved thanks to a projected Gauss–Seidel algorithm (Sect. 12.4.6). The resulting code in Siconos is

```
double h =  1.0e-6;  // Time step
// The time discretisation, linked to the Model.
TimeDiscretisation * td = new TimeDiscretisation(h,
DiodeBridge);
Simulation * s = new TimeStepping(td);
// Moreau Integrator for the dynamics:
double theta = 0.5;
Moreau* myIntegrator = new Moreau(oscillator,theta,s);
// One Step nonsmooth problem:
string solverName = "PGS"; // nonsmooth problem solver
name.
OneStepNSProblem* myLCP = new LCP(s, "LCP", solverName,
101, 0.0001, "max", 0.6 );
```

Note that the Simulation is connected to the Model thanks to the TimeDiscretisation.

The last step is the simulation process with first the initialization and then the time-loop:

```
    // Simulation process
    s->initialize();
    s->run()
```

14.3 SICONOS Software

14.3.1 General Principles of Modeling and Simulation

The SICONOS software is mostly written in C++ and thus integrally relies on the object-oriented paradigm. In this first section we will not get into details on how to build these objects,[4] but rather on what they are and what they are used for.

As explained in Sect. 14.2, the central object is the Model. The model is the overall object composed of a nonsmooth dynamical system and a simulation object. The nonsmooth dynamical system object contains all of the informations to describe the system and the simulation object contains all of the informations to simulate it. The compulsory process to handle a problem with SICONOS is first to build a nonsmooth dynamical system and then to describe a simulation strategy, see Sect. 14.3.1.2. Additionally, a control of the Model object can possibly be defined, see Sect. 14.3.1.3.

The way the software is written relies also on this "cutting-out" with clearly separated modeling and simulation components as explained in Sect. 14.3.4.

14.3.1.1 NSDS Modeling in SICONOS Software

An NSDS can be viewed as a set of dynamical systems that may interact in a nonsmooth way through interactions. The modeling approach in the SICONOS platform consists in considering the NSDS as a graph with dynamical systems as nodes and nonsmooth interactions as branches. Thus, to describe each element of this graph in SICONOS, one needs to define a NonSmoothDynamicalSystem object composed of a set of DynamicalSystem objects and a set of Interaction objects.

A DynamicalSystem object is just a set of equations to describe the behavior of a single dynamical system, with some specific operators, initial conditions, and so on. A complete review of the dynamical systems available in SICONOS is given in Sect. 14.3.2.1.

[4] This is the role of the tutorial, users, guide or others manuals that may be found at http://siconos.gforge.inria.fr/

An Interaction object describes the way one or more dynamical systems are linked or may interact. For instance, if one considers a set of rigid bodies, the Interaction objects define and describe what happens at contact. The Interaction object is characterized by some "local" variables, y (also called output), and λ (input) and is composed of

- a NonSmoothLaw object that describes the mapping between y and λ,
- a Relation object that describes the equations between the local variables (y, λ) and the global ones (those of the DynamicalSystem object).

One can find a review of the various possibilities for the Relation and the NonSmoothLaw objects in Sects. 14.3.2.2 and 14.3.2.3. As summarized in Fig. 14.2, building a problem in SICONOS relies on the proper identification and construction of some DynamicalSystems and of all the potential interactions.

14.3.1.2 Simulation Strategies for the NSDS Behavior

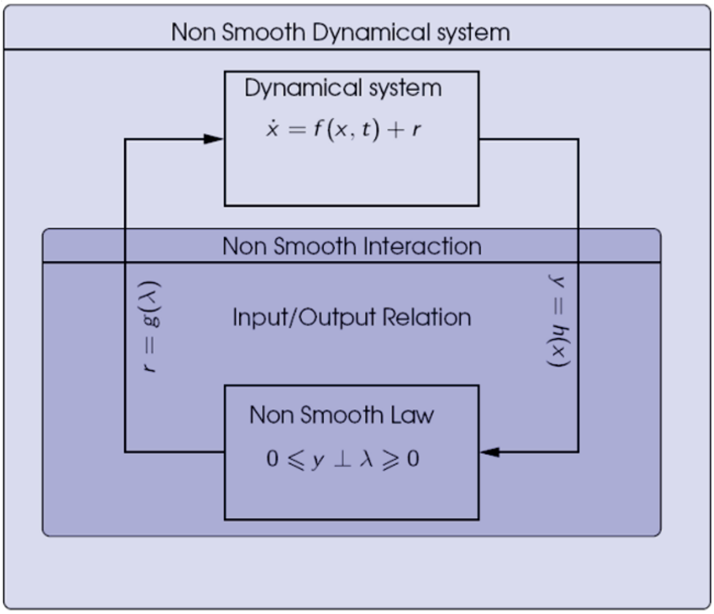
Once an NSDS has been fully designed and described thanks to the objects detailed above, it is necessary to build a Simulation object, namely to define the way the nonsmooth response of the NSDS will be computed.

First of all, let us introduce the Event object, which is characterized by a type and a time of occurrence. Each event has also a process method which defines a list of actions that are executed when this event occurs. These actions depend on the object type. For the objects related to nonsmooth time events, namely NonSmoothEvent, an action is performed only if an event-driven strategy is chosen. For the SensorsEvents and ActuatorEvent related to control tools (see Sect. 14.3.1.3), an action is performed for both time-stepping and event-driven strategy at the times defined by the control law. Finally, thanks to a registration mechanism, user-defined events can be added.

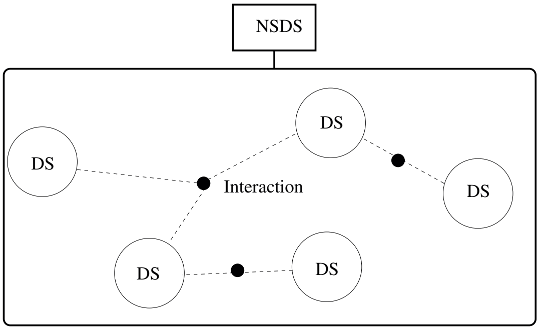
To build the Simulation object, we first define a discretization, using a TimeDiscretisation object, to set the number of time steps and their respective size. Note that the initial and final time values are part of the Model. The time instants of this discretization define TimeDiscretisationEvent objects used to initialize an EventsManager object, which contains the list of Event objects and their related methods. The EventsManager object belongs to the simulation and will lead the simulation process: the system integration is always done between a "current" and a "next" event. Then, during simulation, events of different types may be added or removed, for example when the user creates a sensor or when an impact is detected.

Thereafter, to complete the Simulation object, we need

- some instructions on how to integrate the smooth dynamics over a time step, which is the role of the OneStepIntegrator objects,
- some details on how to formalize and solve the nonsmooth problems when they occur, this is done with the OneStepNSProblem objects.



(a) A simple `NonSmoothDynamicalSystem` with one `DynamicalSystem` object and one `Interaction`



(b) The graph structure of a complex NSDS with `DynamicalSystem` objects as nodes and nonsmooth `Interaction` object as branches

Fig. 14.2. SICONOS nonsmooth dynamical system modeling principle

To summarize, a `Simulation` object is composed of a `TimeDiscretisation`, a set of `OneStepIntegrator` plus a set of `OneStepNSProblem` and belongs to a `Model` object. The whole simulation process is led by the chosen type of strategy, either time-stepping or event-driven. To proceed, one needs to instantiate one of the classes that inherits from `Simulation` object: `TimeStepping` or `EventDriven`.

14.3.1.3 Control Tools

In SICONOS, some control can be applied on a NSDS. The principle is to get information from the systems thanks to some `Sensor` objects, used by some `Actuator` objects to act on the NSDS components. Each `Sensor` or `Actuator` object has its own `TimeDiscretisation` object, a list of time instants where data are to be captured for sensors or where action occurs for actuators. Those instants are scheduled as events into the simulation's `EventsManager` object and thus processed when necessary.

The whole control process is handled thanks to a `ControlManager` object, which is composed of a set of `Sensor` objects and another set of `Actuator` objects. The `ControlManager` object "knows" the `Model` object and thus all its components.

Each `DynamicalSystem` object has a specific variable, named z, which is a vector of discrete parameters (see Sect. 14.3.2.1). To control the systems with a sampled control law, the `Actuator` object sets the values of z components according to the user instructions.

14.3.2 NSDS-Related Components

In the following paragraphs, we turn our attention to the specific types of systems, relations, and laws available in the platform.

14.3.2.1 Dynamical Systems

The most general way to write dynamical systems in SICONOS is

$$g(\dot{x}, x, t, z) = 0,$$

which is a n-dimensional set of equations where

- t is the time,
- $x \in \mathbb{R}^n$ is the state,[5]
- the vector of algebraic variables $z \in \mathbb{R}^s$ is a set of discrete states, which evolves only at user-specified events. The vector z may be used to set some perturbation parameters or to stabilize the system with a sampled control law.

[5] The typical dimension of the state vector can range between a few degrees of freedom and more than several hundred thousands, for example for mechanical or electrical systems. The implementation of the software has been done to deal either with small- or large-scale problems.

Under some specific conditions, we can rewrite this as

$$\dot{x} = \text{rhs}(x,t,z),$$

where "rhs" means right-hand side. Note that in that case $\nabla_{\dot{x}} g(\cdot,\cdot,\cdot,\cdot)$ must be invertible. From this generic interface, some specific dynamical systems are derived, to fit with different application fields. They are separated into two categories: first- and second-order (Lagrangian) systems, and then specialized according to the type of their operators (linear or not, time invariant, etc.).

The following list reviews the dynamical system implemented in SICONOS:

- `FirstOrderNonLinearDS` class, which describes the nonlinear dynamical systems of first order in the form

$$\begin{cases} M\dot{x}(t) = f(t,x(t),z) + r \\ \\ x(t_0) = x_0 \end{cases} \tag{14.9}$$

  with M a $n \times n$ matrix, $f(x,t,z)$ the vector field, and r the input due to the nonsmooth behavior.
- `FirstOrderLinearDS` class, which describes the linear dynamical systems of first order in the form (coefficients may be time invariant or not)

$$\begin{cases} \dot{x}(t) = A(t,z)x(t) + b(t,z) + r \\ \\ x(t_0) = x_0. \end{cases} \tag{14.10}$$

  Simple Electrical circuits for instance fit into this formalism, as shown in the diode bridge example in Sect. 14.2.
- `LagrangianDS` class, which describes the Lagrangian nonlinear dynamical systems in the form

$$\begin{cases} M(q,z)\ddot{q} + \text{NNL}(\dot{q},q,z) + F_{\text{int}}(t,\dot{q},q,z) = F_{\text{ext}}(t,z) + p \\ \\ q(t_0) = q_0, \quad \dot{q}(t_0) = v_0 \end{cases}, \tag{14.11}$$

  where q denotes the generalized coordinates, NNL the nonlinear inertia operator, F_{int} the internal, nonlinear, forces and F_{ext} the external forces, depending only on time. This formalism corresponds to mechanics and can be written in a simpler manner as

$$\begin{cases} M(q,z)\ddot{q} = f_L(t,\dot{q},q,z) + p \\ \\ q(t_0) = q_0, \quad \dot{q}(t_0) = v_0 \end{cases}. \tag{14.12}$$

  The full-form (14.11) with several operators has been designed to fit different users habits, depending on the application field (multibody mechanics, robotics, solid and structures mechanics through Finite Element Method (FEM)).

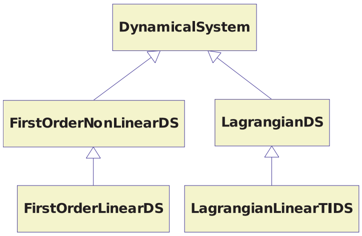


Fig. 14.3. DynamicalSystem-type classes

- `LagrangianLinearTIDS` class, which describes the Lagrangian linear and time-invariant coefficients systems:

$$\begin{cases} M\ddot{q} + C\dot{q} + Kq = F_{\text{ext}}(t,z) + p \\ q(t_0) = q_0, \quad \dot{q}(t_0) = v_0 \end{cases}, \tag{14.13}$$

where C and K are, respectively, the classical damping and stiffness matrices.

As illustrated in Fig. 14.3, all the classes inherit from the `DynamicalSystem` class.

14.3.2.2 Relations

As explained above, some relations between local, (y,λ), and global variables (x,r), have to be set to describe the interactions between systems. The general form of these algebraic equations is

$$\begin{cases} y = \text{output}(x,t,z,\dots) \\ r = \text{input}(\lambda,t,z,\dots) \end{cases} \tag{14.14}$$

and is contained in the abstract `Relation` class. Any other `Relation` objects are derived from this one.

As for `DynamicalSystems` they are separated in first- and second-order relations and specified according to the type and number of variables, the linearity of the operators, etc. The possible cases are as follows:

- `FirstOrderR` class, which describes the nonlinear relations of first order as

$$\begin{cases} y = h(X,t,Z) \\ R = g(\lambda,t,Z). \end{cases} \tag{14.15}$$

Note that we use upper case for all variables related to `DynamicalSystem` objects. Remember that a `Relation` object applies through the `Interaction` object to a set of dynamical systems, and thus, X, Z,… are concatenation of x, z,… of the `DynamicalSystem` objects involved in the relation.

- FirstOrderLinearTIR class, which describes the first-order linear and time-invariant relations:
$$\begin{cases} y = CX + FZ + D\lambda + e \\ \\ R = B\lambda. \end{cases} \tag{14.16}$$
Once again, see for instance the diode bridge example in Sect. 14.2.
- `LagrangianScleronoumousR` class: the scleronomic constraints case, where the relation depends only on the global coordinates of the dynamical systems,
$$\begin{cases} y = h(Q,Z) \\ \\ \dot{y} = G_0(Q,Z)\dot{Q} \\ \\ P = \nabla h^{\mathrm{T}}(Q,Z)\lambda = G_0(Q,Z)^{\mathrm{T}}\lambda \end{cases} \tag{14.17}$$
with
$$G_0(Q,Z) = \nabla_Q h(Q,Z). \tag{14.18}$$
- `LagrangianRheonomousR`: in that case, the relation depends also on time.
$$\begin{cases} y = h(Q,t,Z) \\ \\ \dot{y} = G_0(Q,t,Z)\dot{Q} + \dfrac{\partial h}{\partial t}(Q,t,Z) \\ \\ P = G_0(Q,t,Z)^{\mathrm{T}}\lambda \end{cases} \tag{14.19}$$
with
$$G_0(Q,t,Z) = \nabla_Q h(Q,t,Z). \tag{14.20}$$
- `LagrangianCompliantR` class: there, the relation depends on λ. For instance in the mechanical case, this may correspond to a spring, since it links a force to a displacement.
$$\begin{cases} y = h(Q,\lambda_0,Z) \\ \\ \dot{y} = G_0(Q,\lambda_0,Z)\dot{Q} + G_1(Q,\lambda_0,Z)\lambda_1 \\ \\ P = G_0(Q,\lambda_0,Z)^{\mathrm{T}}\lambda_0 \end{cases} \tag{14.21}$$
with
$$\begin{cases} G_0(Q,\lambda_0,Z) = \nabla_Q h(Q,\lambda_0,Z) \\ \\ G_1(Q,\lambda_0,Z) = \nabla_{\lambda_0} h(Q,\lambda_0,Z) \end{cases} \tag{14.22}$$
and λ_0 the multiplier corresponding to y, while λ_1 corresponds to $\dot{y}$.

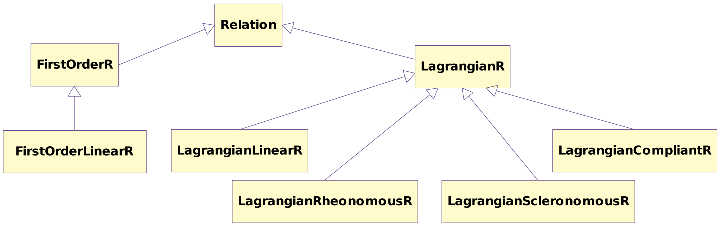


Fig. 14.4. Relation-type classes

- `LagrangianLinearR` class: the simplest one, with linear and time-invariant relations between local and global variables.

$$\begin{cases} y = HQ + D\lambda + FZ + b \\ P = H^{\mathrm{T}}\lambda. \end{cases} \tag{14.23}$$

As shown in Fig. 14.4, all the classes inherit from the `Relation` class.

14.3.2.3 Nonsmooth Laws

The `NonSmoothLaw` object is the last required object to complete the `Interaction` object. We present here a list of the existing laws in SICONOS:

- `ComplementarityConditionNSL` class which models a complementarity condition as

$$0 \leqslant y \perp \lambda \geqslant 0. \tag{14.24}$$

- `NewtonImpactNSL` class which models the unilateral contact with the Newton's impact law, known also as the Moreau's impacting rule:

$$\text{if } y(t) = 0, \quad 0 \leqslant \dot{y}(t^+) + e\dot{y}(t^-) \perp \lambda \geqslant 0. \tag{14.25}$$

- `RelayNSL` class which models the simple relay mapping as

$$\begin{cases} \dot{y} = 0\colon\ |\lambda| \leqslant 1 \\ \dot{y} \neq 0\colon\ \lambda = \operatorname{sign}(y). \end{cases} \tag{14.26}$$

- `NewtonImpactFrictionNSL` class which models the unilateral contact with Coulomb's friction in 2D and 3D as: $y = [y_{\mathrm{N}}, y_{\mathrm{T}}]^{\mathrm{T}}$, $\lambda = [\lambda_{\mathrm{N}}, \lambda_{\mathrm{T}}]^{\mathrm{T}}$,

$$\text{if} \quad y_{\mathrm{N}} = 0, \begin{cases} 0 \leqslant \dot{y}_{\mathrm{N}} \perp \lambda_{\mathrm{N}} \geqslant 0 \\ \dot{y}_{\mathrm{T}} = 0, \|\lambda_{\mathrm{T}}\| \leqslant \mu\lambda_{\mathrm{N}} \\ \dot{y}_{\mathrm{T}} \neq 0, \lambda_{\mathrm{T}} = -\mu\lambda_{\mathrm{N}} \operatorname{sign}(\dot{y}_{\mathrm{T}}). \end{cases} \tag{14.27}$$

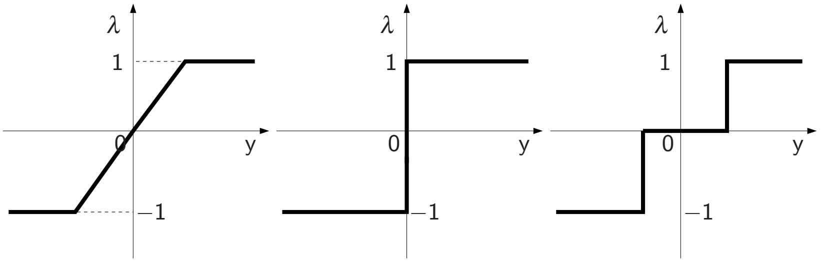


Fig. 14.5. Some multivalued piecewise linear laws: saturation, relay, relay with dead zone

- `PiecewiseLinearNSL` class which models 1D piecewise linear set-valued mapping with fill-in graphs as depicted in Fig. 14.5 (see also Fig. 14.6).

14.3.3 Simulation-Related Components

14.3.3.1 Integration of the Dynamics

To integrate the dynamics over a time step or between two events, `OneStepIntegrator` objects have to be defined. Two types of integrators are available at the time in the platform, listed below and represented in Fig. 14.7a:

- `Moreau` class for Moreau's time-stepping scheme, based on a θ-method,
- `Lsodar` class for the event-driven strategy; this class is an interface for LSODAR, odepack integrator (see `http://www.netlib.org/alliant/ode/doc`).

14.3.3.2 Formalization and Solving of the Nonsmooth Problems

Depending on the encountered situation, various formalizations for the nonsmooth problem are available:

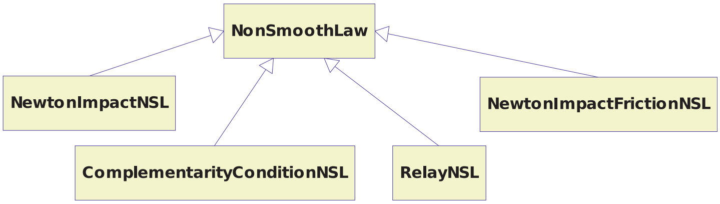
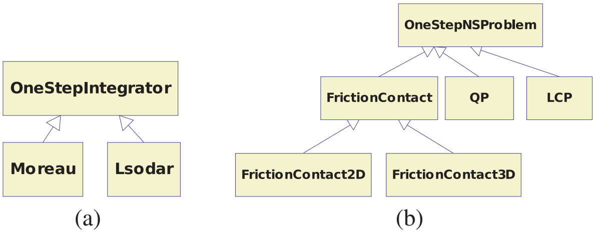


Fig. 14.6. NonSmoothLaw-type classes



(a)

(b)

Fig. 14.7. (**a**) One-step integrators classes, (b) one-step nonsmooth problem classes

- `LCP` class which describes the linear complementarity problem (12.66)

$$\begin{cases} w = Mz + q \\ \\ 0 \leqslant w \perp z \geqslant 0 \end{cases},$$

- `FrictionContact2D(3D)` class, for two(three)-dimensional contact and friction problems, described in Sect. 3.9.1
- `QP` class for the quadratic programming problem (1.1)
- `Relay` class for the relay problem.

From a practical point of view, the solving of nonsmooth problems relies on low-level algorithms (from the SICONOS/Numerics package).

14.3.4 SICONOS Software Design

14.3.4.1 Overview

SICONOS is composed of three main parts: Numerics, Kernel and Front-End, as represented in Fig. 14.8 below.

The SICONOS**/Kernel** is the core of the software, providing high-level description of the studied systems and numerical solving strategies. It is fully written in C++, using extensively the STL utilities. A complete description of the Kernel is given in Sect. 14.3.4.2.

The SICONOS**/Numerics** part holds all low-level algorithms, to compute basic well-identified problems (ordinary differential equations, LCP, QP, etc).

The last component, SICONOS**/Front-End**, provides interfaces with some specific command-languages such as Python or Scilab. This to supply more pleasant and easy-access tools for users, during pre/post-treatment. Front-End is only an optional pack, while the Kernel cannot work without Numerics.

14.3.4.2 The SICONOS/Numerics library

The SICONOS/Numerics library which is a stand-alone library, contains a collection of low-level numerical routines in C and F77 to solve linear algebra problems and OSNSP. It is based on well−known netlib libraries such as BLAS/LAPACK, ATLAS, Templates. Numerical integration of ODE is also provided thanks to ODEPACK. (LSODE solver.) At the time, the following OSNSP solvers are implemented:

- LCP solvers:

– Splitting based methods of Sect. 12.4.6 (PSOR, PGS, RPSOR, RPGS)
– Lemke's algorithm of Sect. 12.4.7
– Newton's method of Sect. 12.5.4

- MLCP solvers:

– Splitting based methods of Sect. 12.4.6 (PSOR, PGS, RPSOR, RPGS)

- NCP solvers.

– Newton's method based on the Fischer–Burmeister function
– Interface to the PATH solver described in Sect. 13.5.3

- QP solver based on QLD due to Prof. K.Schittkowski of the University of Bayreuth, Germany (modification of routines due to Prof. MJD Powell at the University of Cambridge).

- Frictional contact solvers:

– Projection-type methods of Sect. 13.7.2
– NSGS splitting based method of Sect. 13.7.4
– Alart–Curnier's method of Sect. 13.6.1
– NCP reformulation method of Sect. 13.4.3

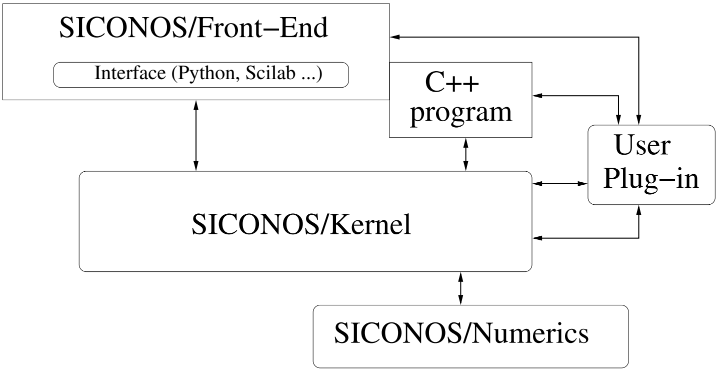


Fig. 14.8. General design of SICONOS software

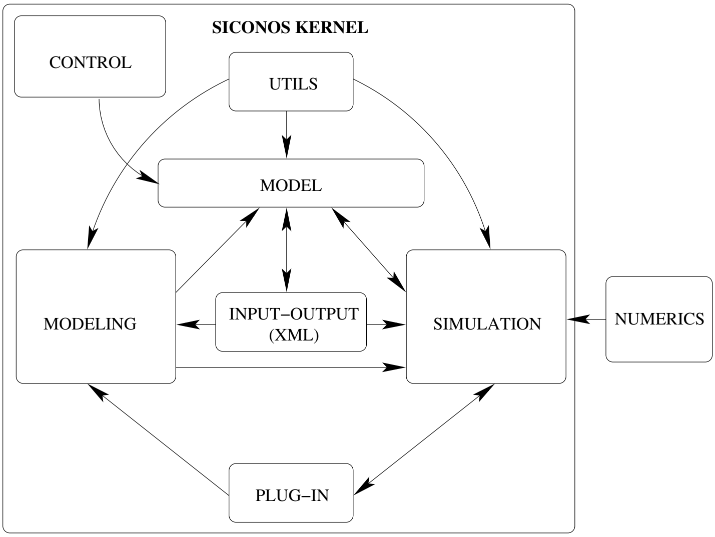


Fig. 14.9. Kernel components dependencies

14.3.4.3 Siconos Kernel Components

As previously said, Kernel is the central and main part of the software. The whole dependencies among Kernel parts are fully depicted in Fig. 14.9.

All the Kernel implementation is based on the principle we gave in Sect. 14.3.1. It is mainly composed of two rather distinct parts, modeling and simulation, that handle all the objects used, respectively, in the NSDS modeling (see Sect. 14.3.1.1) and the Simulation description (see Sect. 14.3.1.2).

The Utils module contains tools, mainly to handle classical objects such as matrices or vectors and is based on the Boost library,[6] especially, uBLAS,[7] a C++ library that provides BLAS functionalities for vectors, dense and sparse matrices.

The Input–Output module concerns objects for data management in XML format, thanks to the libxml2 see footnote[8] library. More precisely, all the description of the Model, NSDS and Simulation, can be done thanks to an XML input file. An example of such a file is given in Sect. 14.4.1.

Control package provides objects like `Sensor` and `Actuator`, to add control of the dynamical systems through the `Model` object, as explained in Sect. 14.3.1.3.

A plug-in system is available, mainly to allow the user to provide one's own computation methods for some specific functions (vector field of a dynamical system,

[6] `http://www.boost.org`.

[7] `http://www.boost.org/libs/numeric/ublas/doc/index.htm`.

[8] http://xmlsoft.org/

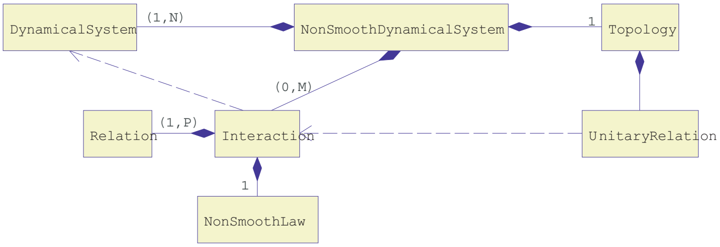


Fig. 14.10. Simplified class diagram for Kernel modeling part

mass, etc.), this without having to recompile the whole platform. Moreover, the platform is designed in a way that allows user to add dedicated modules through object registration and object factories mechanisms (for example to add a specific nonsmooth law, a user-defined sensor, etc.).

To conclude, class diagrams for modeling and simulation components are given in Figs. 14.10 and 14.11, which make clearer the various links between all the objects presented before.

14.4 Examples

14.4.1 The Bouncing Ball(s)

We consider a ball of mass m and radius R, described by three generalized coordinates $q = (z,x,\theta)^{\mathrm{T}}$. The ball is subjected to the gravity g. The system is also constituted by a rigid plane, defined by its position h with respect to the axis Oz. We assume that the position of the plane is fixed.

The equation of motion of the ball is given by

$$M\ddot{q}(t) = F_{\mathrm{ext}}(t) + P \tag{14.28}$$

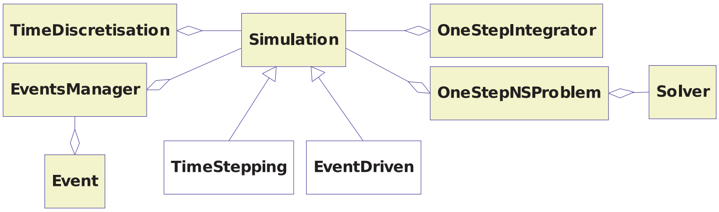


Fig. 14.11. Simplified class diagram for Kernel simulation part

with M the inertia matrix, P the force due to the nonsmooth law, i.e., the reaction at the impact times, and $F_{\text{ext}}(t): \mathscr{R} \mapsto \mathscr{R}^n$ the given external force:

$$M = \begin{bmatrix} m & 0 & 0 \\ 0 & m & 0 \\ 0 & 0 & I \end{bmatrix}, \quad I = 3/5mR^2, \quad F_{\text{ext}} = \begin{bmatrix} -mg \\ 0 \\ 0 \end{bmatrix}. \tag{14.29}$$

The ball bounces on the rigid plane, introducing a constraint on its vertical position, given by

$$z - R - h \geqslant 0. \tag{14.30}$$

We introduce y as the distance between the ball and the floor and λ as the multiplier that corresponds to the reaction at contact. Then from (14.30), we get

$$y = Hq + b = [1\ 0\ 0]q - R - h \tag{14.31}$$

completed by

$$P = H^{\mathrm{T}}\lambda. \tag{14.32}$$

Finally we need to introduce a nonsmooth law to define the behavior of the ball at impact. The unilateral constraint is such that

$$0 \leqslant y \perp \lambda \geqslant 0 \tag{14.33}$$

completed with a Newton impact law, for which we set the restitution coefficient e to 0.9:

$$\text{if } y(t) = 0 \text{ and } \dot{y}(t^-) \leqslant 0, \text{ then } \dot{y}(t^+) = -e\dot{y}(t^-), \tag{14.34}$$

t^+ and t^- being post and pre-impact times. Then, (14.28) fits with `LagrangianLinearTIDS` (14.13) formalism, (14.31) and (14.32) with `LagrangianLinearR` (14.23), and (14.33) and (14.34) with `NewtonImpactNSL` (14.25). If we use XML for the Model description, the part corresponding to the NSDS will look like

```
<NSDS bvp=’false’>
   <DS_Definition>
      <LagrangianLinearTIDS number=’1’>
         <Id> Ball </Id>
         <q0 vectorSize=’3’>1.0 0.0 0.0</q0>
         <Velocity0 vectorSize=’3’>0.0 0.0 0.0</Velocity0>
         <FExt vectorPlugin="BallPlugin:ballFExt"/>
         <Mass matrixRowSize=’3’ matrixColSize=’3’>
            <row>1.0 0.0 0.0</row>
            <row>0.0 1.0 0.0</row>
            <row>0.0 0.0 1.0</row>
            </Mass>
      </LagrangianLinearTIDS>
   </DS_Definition>
   <Interaction_Definition>
      <Interaction number=’1’ Id=’Ball-Ground’>
         <size> 1 </size>
         <DS_Concerned all=’true’></DS_Concerned>
         <Interaction_Content>
            <LagrangianLinearRelation>
               <H matrixRowSize=’1’ matrixColSize=’3’>
                  <row> 1.0  0.0  0.0</row>
               </H>
            </LagrangianLinearRelation>
            <NewtonImpactLaw>
               <e>0.9</e>
            </NewtonImpactLaw>
         </Interaction_Content>
      </Interaction>
   </Interaction_Definition>
</NSDS>
```

For the simulation, we use a Moreau’s time-stepping scheme and an LCP formalization with a Lemke solver:

```
<Simulation type=’TimeStepping’>
   <TimeDiscretisation>
      <h>0.005</h>
   </TimeDiscretisation>
   <OneStepIntegrator_Definition>
      <Moreau>
         <DS_Concerned vectorSize=’1’>1</DS_Concerned>
         <Theta all="0.5"></Theta>
      </Moreau>
   </OneStepIntegrator_Definition>
   <OneStepNSProblem>
      <LCP>
         <Solver type="Lemke" maxIter="101" />
      </LCP>
   </OneStepNSProblem>
</Simulation>
```

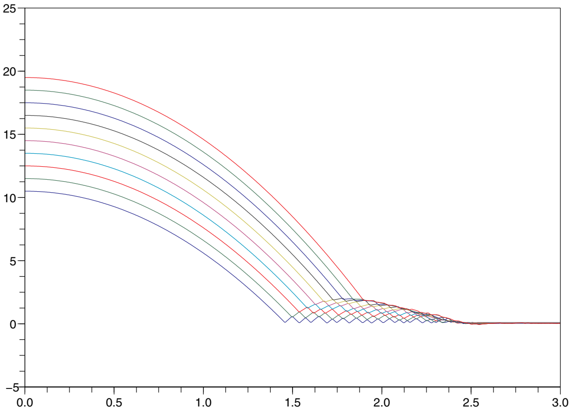


Fig. 14.12. Vertical displacement of the 10 lowest beads according to time

And then, in the C++ input file we have

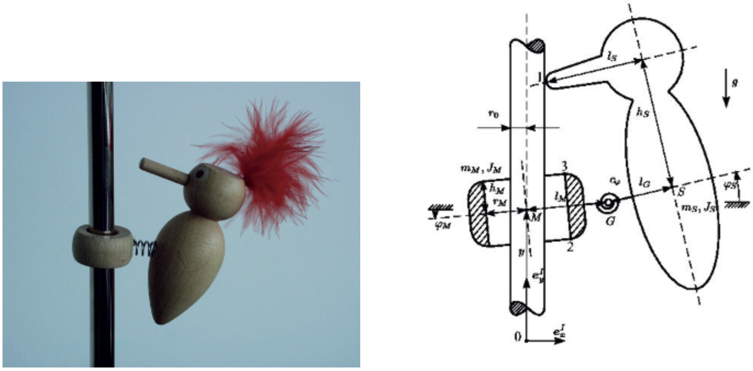
```
// Loading of the XML input file that describes the model:
Model * bouncingBall = new Model("./BallTS.xml");
// Get the simulation object
Simulation* s = bouncingBall->getSimulationPtr();
// ...
// Initialize and run ...
s->initialize();
s->run();
```

We consider now a column of 1000 spherical beads, in contact or not, falling down to the ground. The modeling is quite the same as for the single ball, one just has to define one dynamical system for each bead and one interaction for each potential contact between two beads. The interest of this example lies in the important number of degrees of freedom (i.e., the size of the vector q) and of relations (the size of y and λ) that is equal to 1000. Figure 14.12 displays vertical displacements of the 10 lowest beads according to time.

14.4.2 The Woodpecker Toy

The woodpecker toy is presented in Fig. 14.13a and b and consists of a sleeve, a spring, and the woodpecker. The hole in the sleeve is slightly larger than the diameter of the pole, thus allowing a kind of pitching motion interrupted by impacts with friction. Its dynamical behavior shows both impact and friction phenomena.

The woodpecker toy is a system which can only operate in the presence of friction as it relies on combined impacts and jamming. Among other things, an animation of the toy can be found at: `http://www.zfm.ethz.ch/~leine/toys.htm`.



(a) (b)

Fig. 14.13. The woodpecker toy. Courtesy of Christoph Glocker (1995) and Remco Leine, ETH Zürich

Some results obtained with SICONOS are presented in Fig. 14.14.[9]

14.4.3 MOS Transistors and Inverters

14.4.3.1 Piecewise Linear Model of a MOS Transistor

One can benefit from a simplification of devices models (e.g., MOS models) in the form of a piecewise linear representation instead of the complicated formula implemented in SPICE simulators. For instance, in Leenaerts & Van Bokhoven (1998), the authors considered the Sah model of the NMOS static characteristic:

$$I_{DS} = \frac{K}{2} \cdot (f(V_{\mathrm{G}} - V_{\mathrm{S}} - V_{\mathrm{T}}) - f(V_{\mathrm{G}} - V_{\mathrm{D}} - V_{\mathrm{T}}))$$

with

$K = \dfrac{\mu C_{\mathrm{OX}} W}{L}$

μ mobility of majority carriers

(sample values of 750 $\mathrm{cm}^2\,\mathrm{V}^{-1}\,\mathrm{s}^{-1}$ for an NMOS, 250 $\mathrm{cm}^2\,\mathrm{V}^{-1}\,\mathrm{s}^{-1}$ for a PMOS)

$C_{OX} = \dfrac{\varepsilon_{\mathrm{SiO}_2}}{t_{\mathrm{OX}}}$

[9] This system has been implemented in SICONOS by M. Moeller from the Mechanical Engineering Department of ETH Zurich, following the examples proposed in Leine et al. (2003).

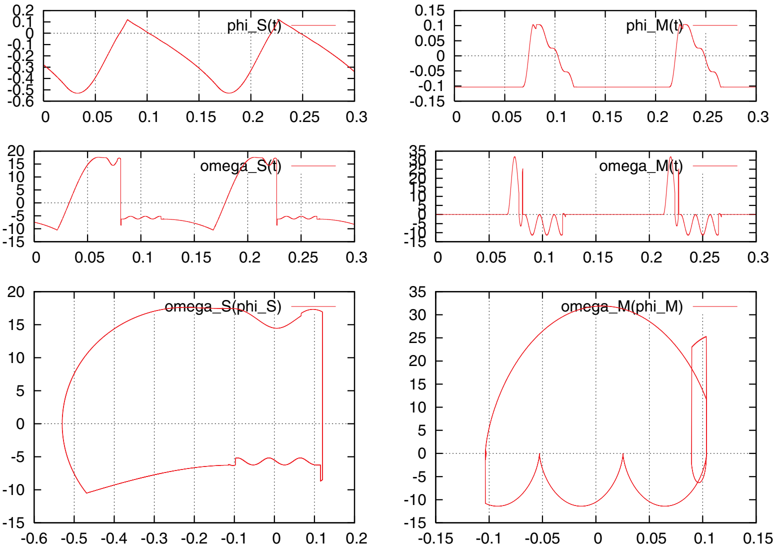


Fig. 14.14. Simulation results for the woodpecker toy using the SICONOS platform

$\varepsilon_{\mathrm{SiO}_2} = \varepsilon_{r\ \mathrm{SiO}_2} \cdot \varepsilon_0$ $(\varepsilon_{r\ SiO_2} \approx 3.9)$

t_{OX} oxide thickness $\approx 4\,\mathrm{nm}$ in a recent $180\,\mathrm{nm}$ technology

W channel width

L channel length $\approx 130\,\mathrm{nm}$ in a recent $180\,\mathrm{nm}$ technology

V_{T} threshold voltage depending on technology, V_{BS} , temperature ≈ 0.25–$1\,\mathrm{V}$

The function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is defined as

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ x^2 & \text{if } x \geqslant 0. \end{cases}$$

The piecewise and quadratic nature of this function is approximated by the following six segments piecewise linear function in Leenaerts & Van Bokhoven (1998) (see Fig. 14.15):

$$f_{\mathrm{PWL}}(x) = \begin{cases} 0 & \text{if } x < 0 \\ 0.09x & \text{if } 0 \leqslant x < 0.1 \\ 0.314055x - 0.0224055 & \text{if } 0.1 \leqslant x < 0.2487 \\ 0.780422x - 0.138391 & \text{if } 0.2487 \leqslant x < 0.6185 \\ 1.94107x - 0.856254 & \text{if } 0.6185 \leqslant x < 1.5383 \\ 4.82766x - 5.29668 & \text{if } 1.5383 \leqslant x\,. \end{cases}$$

The relative error between $f(\cdot)$ and $f_{\mathrm{PWL}}(\cdot)$ is kept below 0.1 for $0.1 \leqslant x < 3.82$. The absolute error is less than 2×10^{-3} for $0 \leqslant x < 0.1$ and 0 for negative x. In practice, the values of $V_{\mathrm{G}}, V_{\mathrm{S}}, V_{\mathrm{D}}, V_{\mathrm{T}}$ in logic integrated circuits allow a good approximation of f by f_{PWL}.

Figure 14.16 displays the static characteristic $I_{\mathrm{DS}}(V_{\mathrm{GS}}, V_{\mathrm{DS}})$ of an NMOS obtained with the SPICE level 1 model and the piecewise linear approximation of the Sah model. The following parameter values were used: $\varepsilon_{r\ \mathrm{SiO_2}} = 3.9$, $t_{\mathrm{OX}} = 20\,\mathrm{nm}$, $\mu = 750\,\mathrm{cm}^2\,\mathrm{V}^{-1}\,\mathrm{s}^{-1}$, $W = 1\ \mu\mathrm{m}$, $L = 1\ \mu\mathrm{m}$, $V_{\mathrm{T}} = 1\ \mathrm{V}$. Bottom figures include results of both models with two different viewpoints to display the regions where differences appear.

14.4.3.2 Inverter Chain

This simple model of an NMOS transistor was adapted to the PMOS transistor and both models were used to simulate an inverter chain (see Fig. 14.17). The output of each inverter is loaded by the intrinsic capacitances of transistors (with values of a few fF) and a load capacitor of 50 fF representing the wiring between successive inverters.

In these early simulations, the dynamical behavior of the MOS transistor was simplified by keeping the intrinsic capacitances C_{GS} and C_{GD} independent from voltages. Of course, this differs from the Meyer nonlinear capacitances implemented in the SPICE level 1 model. Comparisons between simulation results with SPICE and SICONOS for a selection of inverters output voltage and MOS currents can be found in Denoyelle & Acary (2006).

14.4.4 Control of Lagrangian systems

14.4.4.1 Control principle in SICONOS/Control

Two strategies are available to implement a control law in the SICONOS platform:

Nonlinear Continuous Control with Switches

The control can be implemented in external functions F_{int} and F_{ext}. For an accurate simulation of the control law with Newton's method, the Jacobian of the control with respect to q and $\dot{q}$ must be provided (finite–difference approximation can be also used). For an explicit evaluation of the control law, the second strategy is preferable.

Sampled Discrete Control with Delay

Thanks to `Actuator` and `Sensor` objects, it is possible to schedule events of control type in the stack of the `EventsManager` object. The `Sensor` object is able to store any data of the model whenever an event is reached. The `Actuator` object is able to compute the control law with the stored values in the sensors. This strategy allows one to implement "real" sampled control laws with delay and switches independently of the time-step chosen for time-integration. It may be convenient for studying robustness of the control in sampled cases with delay. For the `SensorsEvents`

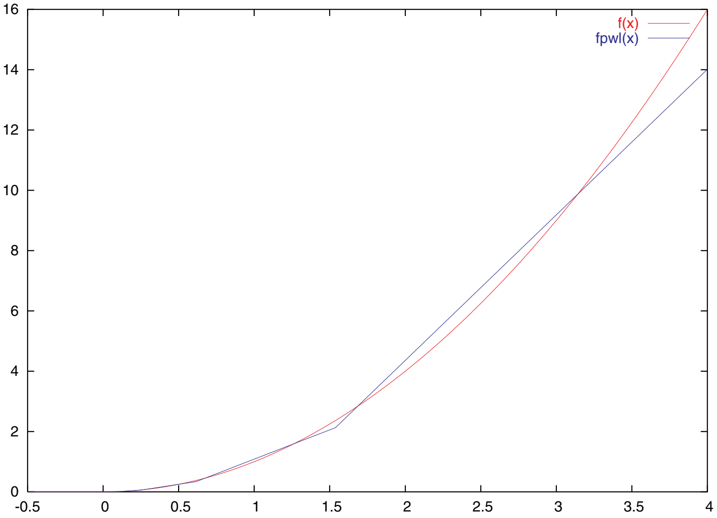


Fig. 14.15. Piecewise linear approximation of $f(\cdot)$

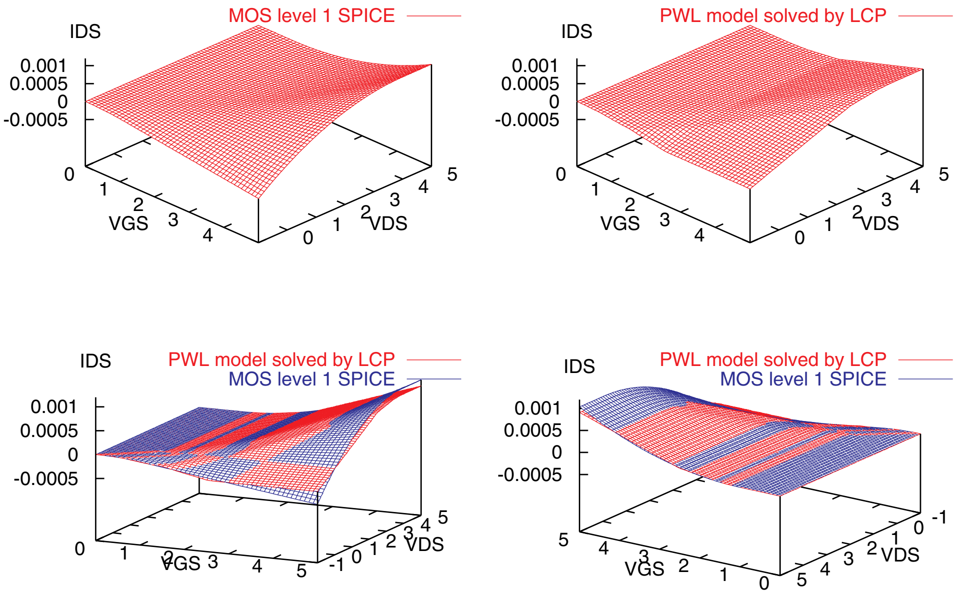


Fig. 14.16. Static characteristic of an NMOS transistor with a simple PWL model and SPICE level 1 model

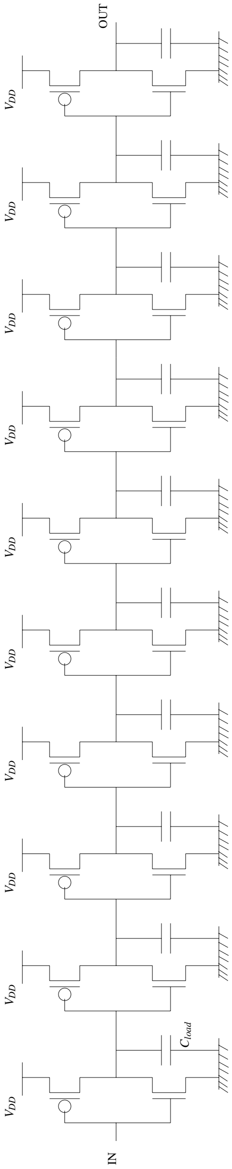


Fig. 14.17. Inverter chain in CMOS

and `ActuatorEvent` related to control tools, an action is performed for both time-stepping and event-driven strategies at the times defined by the control law.

14.4.4.2 The Switching Nonlinear Control of a two-link Manipulator

In the sequel, let us introduce a model that allows us to test the Moreau's time-stepping algorithm of the SICONOS platform presented in the previous sections. Precisely, we consider a simple planar two-link manipulator whose end effector must track a desired circular trajectory that leaves the admissible domain. In order to accomplish its task the manipulator has to follow the constraint from the point where the circle leaves the admissible domain to the point where the circle re-enters in it.

The time domain representation of the manipulator task can be described as (see Brogliato et al. (1997)):

$$\mathbb{R}^+ = \Omega_0 \cup I_0 \cup \Omega_1 \cup \Omega_2 \cup I_1 \cup \ldots \cup \Omega_{2k} \cup I_k \cup \Omega_{2k+1} \cup \ldots \tag{14.35}$$

where Ω_{2k} corresponds to free-motion phases, Ω_{2k+1} corresponds to constrained-motion phases and I_k represents the transient between free and constrained phases. It is worth to point out that during the phases I_k some impacts occur. The constraints defining the admissible domain are supposed frictionless and unilateral.

Controller Design

In order to overcome some difficulties that can appear in the controller definition, the dynamical system (14.11) will be expressed in the generalized coordinates introduced in McClamroch & Wang (1988). The coordinates are $q \in \mathbb{R}^2$, with $q = \begin{bmatrix} q_1 \\ q_2 \end{bmatrix}$, such that the admissible domain $\Phi = \{q_1(t) \geqslant 0\}$ and then the set of complementary relations can be rewritten as $0 \leqslant \lambda \perp Dq \geqslant 0$ with $D = (1,0) \in \mathbb{R}^2$. The controller used here consists of different low-level control laws for each phase of the system. More precisely, the controller can be expressed as

$$T(q)U = \begin{cases} U_{nc} & \text{for } t \in \Omega_{2k} \\ U_t & \text{for } t \in I_k \\ U_c & \text{for } t \in \Omega_{2k+1} \end{cases} \tag{14.36}$$

where $T(q) = \begin{pmatrix} T_1(q) \\ T_2(q) \end{pmatrix} \in \mathbb{R}^{2\times 2}$.

Roughly speaking, we deal with a passivity-based control law (see for instance Brogliato et al. (2007)) but some of the nonlinear terms are compensated during the constrained phases Ω_{2k+1}. We note also that the transition between constrained and free phases is monitored via a LCP. The closed-loop stability analysis can be found in (Bourgeot & Brogliato, 2005). Some of the events(impacts, detachment from the constraint) are state-dependent. Some others (switch between U_{nc} and U_t) are exogenous. The SICONOS/Control toolbox is able to simulate all these events, and to record them.

Dynamics Equation based on Lagrangian Formulation

We consider the following notations (see Fig. 14.18): θ_i represents the joint angle of the joint i, m_i is the mass of link i, I_i denotes the moment of inertia of link i about the axis that passes through the center of mass and is parallel to the Z axis, l_i is the length of link i, and g denotes the gravitational acceleration.

Let us consider that the constraint is given by the ground (i.e. $y = 0$), thus the associated admissible domain is $\Phi = \{(x, y) \mid y \geqslant 0\}$. One introduces the generalized coordinates $q = \begin{bmatrix} y \\ x \end{bmatrix}, y \geqslant 0$ where (x, y) are the Cartesian coordinates of the end effector. However, we suppose that only a half of the circle is in the admissible domain. Concluding the system has to track a half circle and then to follow the ground from the point where the circle leaves the admissible domain to the point where the circle re-enters in it. Using the Lagrangian formulation we derive the dynamical equations of the system. Precisely, the inertia matrix is given by:

$$\begin{aligned} M_{11} &= \frac{m_1 l_1^2}{4} + m_2 \left(l_1^2 + \frac{l_2^2}{4} l_1 l_2 \cos\theta_2 \right) + I_1 + I_2 \\ M_{12} &= M_{21} = \frac{m_2 l_2^2}{4} + \frac{m_2 l_1 l_2}{2} \cos\theta_2 + I_2 \\ M_{22} &= \frac{m_2 l_2^2}{4} + I_2 \end{aligned}$$

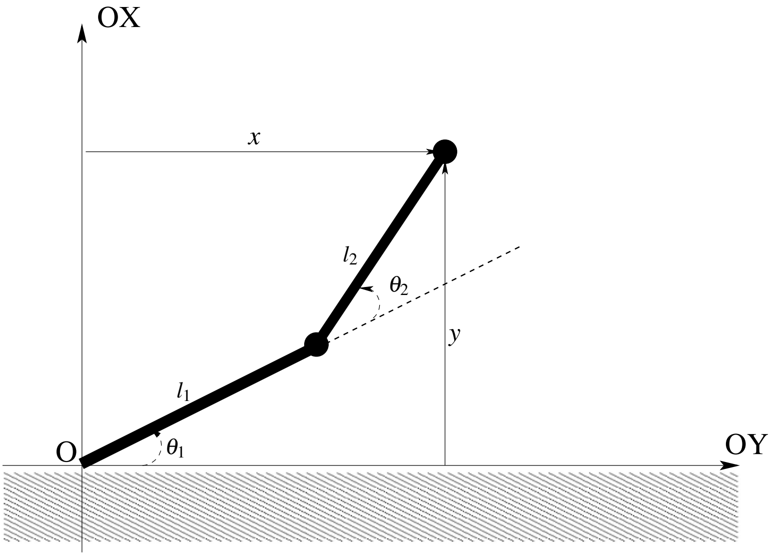


Fig. 14.18. Two-link planar manipulator

The nonlinear term containing Coriolis and centripetal forces is:

$$\begin{aligned} C_{11} &= -m_2 l_1 l_2 \dot{\theta}_2 \sin\theta_2,\ C_{12} = -\frac{m_2 l_1 l_2}{2}\dot{\theta}_2 \sin\theta_2 \\ C_{21} &= \frac{m_2 l_1 l_2}{2}\dot{\theta}_1 \sin\theta_2, \quad C_{22} = 0 \end{aligned}$$

and the term containing conservative forces is:

$$\begin{aligned} G_1 &= \frac{g}{2}[l_1(2m_1 + m_2)\cos\theta_1 + m_2 l_2 \cos(\theta_1 + \theta_2)] \\ G_2 &= \frac{m_2 g l_2}{2}\cos(\theta_1 + \theta_2) \end{aligned}$$

Obviously the generalized coordinates are obtained using the following transformation:

$$\begin{aligned} y &= l_1 \sin\theta_1 + l_2 \sin(\theta_1 + \theta_2) \\ x &= l_1 \cos\theta_1 + l_2 \cos(\theta_1 + \theta_2) \end{aligned}$$

Implementation Details

The simulations were done using a nonlinear continuous control strategy. Precisely the term $NNL(\dot{q}, q, z)$ of equation (14.11) has been identified as $C(\dot{q}, q)\dot{q} + G(q)$ and the switching control has been introduced using the function F_{int}. Therefore, the Jacobian of N and F_{int} with respect to q and $\dot{q}$ have been explicitly computed and inserted in the algorithm. The SICONOS/Control toolbox also allows one to introduce a time-delay in the feedback loop.

14.4.4.3 Numerical Results

The stability analysis of the model and figures illustrating the behavior of the system during each phase of the motion (particularly during transition phases where the corresponding Lyapunov function is almost decreasing) can be found in (Morărescu & Brogliato, 2008). In the sequel, we discuss only some numerical aspects related to the time-stepping simulation strategy chosen in this work. The choice of a time-stepping algorithm was mainly dictated by the presence of accumulations of impacts which render the use of event-driven methods difficult[10]. The numerical values used for the dynamical model are $l_1 = l_2 = 0.5m$, $I_1 = I_2 = 1kg.m^2$, $m_1 = m_2 = 1kg$. It is noteworthy that the simulation results do not depend essentially on the chosen time-step for the scheme but, a smaller time-step allows to capture more precisely the behavior of the system. As it can be seen in Fig. 14.19 the real trajectory and the lengths of each transition phase are almost unchanged starting with a sufficiently small time-step ($h = 10^{-3}$).

We do not insist too much on the simulation results during the free-motion phases since the smoothness of the system is guaranteed on these phases and the behavior

[10] An event-driven algorithm is also available in SICONOS. Its use in case of accumulations needs some ad hoc numerical tricks to pass through the accumulation.

of the system is clear. The most interesting phases from the numerical point of view are the transition (accumulation of impacts) phases. It is worth to clarify that the

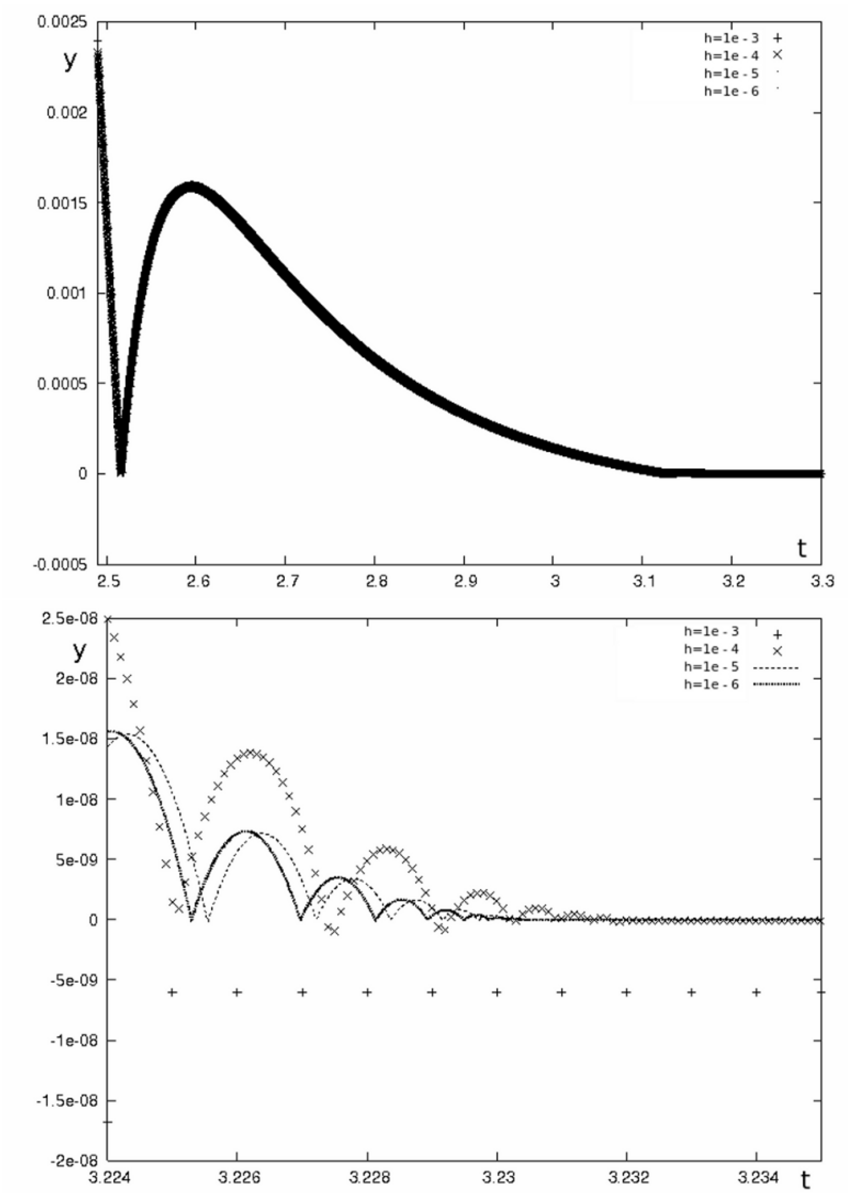


Fig. 14.19. Top: The variation of q_1 during transition phase for $h \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$; **Bottom**: Zoom at the end of transition phase

Table 14.1. Length of the transition phase I_1 and CPU time with respect to the time-step h

h	$10^{-3}s$	$10^{-4}s$	$10^{-5}s$	$10^{-6}s$
$\lambda(I_1)$	0.945	0.9536	0.9525	0.9523
CPU time	$1.5s$	$11.2s$	$111.3s$	$1072.2s$

number of impacts during the transition phases is not so important and the major issue is the finiteness of these phases. To be more clear we present in Table 14.1 and 14.2 some numerical values. In Table 14.1 one can see that the length of the transition phase I_1 with respect to the time-step h does not vary significantly when the time-step decreases. Let us also denote by *CPU* the computing time necessary for the simulation (using an Intel(R) Core(TM)2 CPU 6300 1.86GHz) of one cycle (5 seconds).

The evolution of the number of impacts n_i with respect to the restitution coefficient e_{N} and the time-step h is quite different (see Table 14.2). As expected, n_i becomes larger when the restitution coefficient increases. Also, one can see that the accumulation of impacts can be captured with a higher precision when the time-step becomes smaller. However, a higher number of captured impacts does not change the global behavior of the system and the transition phase ends almost in the same moment when h varies, see $\lambda(I_1)$ in Table 14.1. Other details on the dependence of the trajectories on the control parameters can be found in (Acary et al., 2008).

14.5 Notes

The software can be downloaded at `http://siconos.gforge.inria.fr/`, where one can also find an installation guide, a tutorial, the full doxygen documentation of the code, support, mailing lists and all that sort of utilities.

Note that the above presentation is only an overall view which is moreover likely to change. The implementation of the SICONOS/Kernel and SICONOS/Numerics libraries is still in progress. Users are invited to check on `http://siconos.gforge.inria.fr` for the contents of future releases. The technical report (Acary & Pérignon, 2007) will be updated with the new functionalities and the new examples.

Table 14.2. Number of impacts detected n_i with respect to the time-step h and the coefficient of restitution e_{N}

$e_{\mathrm{N}} \backslash h$	$10^{-3}s$	$10^{-4}s$	$10^{-5}s$	$10^{-6}s$
0.2	$n_i = 3$	$n_i = 5$	$n_i = 6$	$n_i = 8$
0.5	$n_i = 6$	$n_i = 9$	$n_i = 12$	$n_i = 16$
0.7	$n_i = 9$	$n_i = 16$	$n_i = 23$	$n_i = 29$
0.9	$n_i = 23$	$n_i = 40$	$n_i = 64$	$n_i = 81$
0.95	$n_i = 32$	$n_i = 67$	$n_i = 108$	$n_i = 161$

A

Convex, Nonsmooth, and Set-Valued Analysis

A.1 Set-Valued Analysis

Definition A.1 (Hausdorff distance). *Let A and B be two nonempty sets of* IR^n. *We define the distance between a point x and a set A as*

$$\rho(x,A) = \inf_{a \in A} ||x-a||$$

and

$$d_H(A,B) = \max\left\{\sup_{x\in A} \rho(x,B), \sup_{x\in B} \rho(x,A)\right\} \tag{A.1}$$

which is the Hausdorff distance between A and B.

The Hausdorff continuity of a function means that the function is continuous with the Hausdorff distance.

A.2 Subdifferentiation

Definition A.2 (Clarke's generalized derivative). *Let a function* $f \colon \mathrm{IR}^n \to \mathrm{IR}$ *be locally Lipschitz continuous. The* generalized gradient *of* $f(\cdot)$ *at x is defined as the set*

$$\partial f(x) = \mathrm{conv}\{\lim \nabla f(x_i) \mid x_i \to x, x_i \notin \Omega_f \cup N\} \tag{A.2}$$

where Ω_f *is the set of Lebesgue measure zero where* ∇f *does not exist, and N is an arbitrary set of zero Lebesgue measure.*

The next proposition is a generalization of the chain rule (Goeleven et al., 2003a), when a convex function is composed with a linear mapping.

Proposition A.3. *Let $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex lower semi-continuous function and $A\colon \mathbb{R}^m \to \mathbb{R}^n$ be a linear continuous operator. Assume that a point $y_0 = Ax_0$ exists at which $f(\cdot)$ is finite and continuous. The subdifferential in the sense of convex analysis of the composite functional $f \circ A\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is given by*

$$\partial(f \circ A)(x) = A^{\mathrm{T}} \partial f(Ax), \ \ \forall\, x \in \mathbb{R}^n. \tag{A.3}$$

We remind that what is called the subdifferential in the sense of convex analysis, is the set of subgradients as in (1.5).

A.3 Some Useful Equivalences

Let $\phi(\cdot)$ be a proper, convex lower semi-continuous function $\mathbb{R}^n \to \mathbb{R}$. Then for each $y \in \mathbb{R}^n$ there exists a unique $x \overset{\Delta}{=} P_\phi(y) \in \mathbb{R}^n$ such that

$$\langle x - y, v - x\rangle + \phi(v) - \phi(x) \geqslant 0, \ \text{ for all } v \in \mathbb{R}^n. \tag{A.4}$$

The mapping $P_\phi\colon \mathbb{R}^n \to \mathbb{R}^n$ is called the *proximation operator*. It is single-valued, nonexpansive and continuous. The next equivalences hold:

$$\begin{gathered} u \in \mathbb{R}^n\colon \langle Mu + q, v - u\rangle + \phi(v) - \phi(u) \geqslant 0, \ \text{ for all } v \in \mathbb{R}^n \\ \Updownarrow \\ u \in \mathbb{R}^n\colon u = P_\phi(u - (Mu + q)) \\ \Updownarrow \\ u \in \mathbb{R}^n\colon Mu + q \in -\partial\phi(u). \end{gathered} \tag{A.5}$$

The first formulation in (A.5) is called a VI of the second kind. Setting $\phi(\cdot) = \psi_K(\cdot)$ with $K \subset \mathbb{R}^n$ nonempty closed convex, one obtains a VI of the first kind:

$$u \in K\colon \langle Mu + q, v - u\rangle \geqslant 0, \ \text{ for all } v \in K. \tag{A.6}$$

Setting $\phi(\cdot) = \varphi(\cdot) + \psi_K(\cdot)$ with $\varphi(\cdot)$ a proper, convex lower semi-continuous function $\mathbb{R}^n \to \mathbb{R}$, one gets a mixed VI:

$$u \in K\colon \langle Mu + q, v - u\rangle + \varphi(v) - \varphi(u) \geqslant 0, \ \text{ for all } v \in K. \tag{A.7}$$

Let $M = M^{\mathrm{T}} > 0$ be a $n \times n$ matrix and $K \subset \mathbb{R}^n$ be a closed convex nonempty set. Then

$$\begin{array}{c} M(x-y) \in -N_K(x) \\ \Updownarrow \\ x = \mathrm{prox}_M[K;y] \\ \Updownarrow \\ x = \mathrm{argmin}_{z\in K}\, \frac{1}{2}(z-y)^{\mathrm{T}} M(z-y) \\ \Updownarrow \\ x = \mathrm{proj}_M(K;y), \end{array} \tag{A.8}$$

where proj_M indicates that the projection is done in the metric defined by M. In some other parts of the book, when $M = I_n$ we simply denote the projection as $\mathrm{proj}_K(\cdot)$ or as $P_K(\cdot)$.

If $K = (\mathrm{I\!R}^m)^+$ and $M = I_n$, we get

$$y \in -N_K(x) \iff 0 \leqslant y \perp x \geqslant 0 \tag{A.9}$$

for any $x, y \in \mathrm{I\!R}^m$. Let $K \subset \mathrm{I\!R}^n$ be a closed convex nonempty cone and K^o is its polar cone, i.e.,

$$K^o = \{x \in \mathrm{I\!R}^n \mid x^{\mathrm{T}} y \leqslant 0 \text{ for all } y \in K\}.$$

One has $(K^o)^o = K$. Moreover the conjugate function of the indicator of K is the indicator function of K^o, i.e.,

$$\psi_K^*(\cdot) = \psi_{K^o}(\cdot).$$

Thus $\partial\psi_K^*(\cdot) = N_{K^o}(\cdot)$, the normal cone to K^o. Finally

$$y \in \partial\psi_K(x) \Leftrightarrow x \in N_{K^o}(y).$$

This is equivalent to $x \in K$, $y \in K^o$, and $x^{\mathrm{T}} y = 0$.

More generally when K is a nonempty closed convex cone, then (A.9) is extended to

$$y \in -N_K(x) \iff -K^0 \ni y \perp x \in K \tag{A.10}$$

which in turn is equivalent to the VI in (A.6) with $-y = Mu + q$ and $x = u$.

Remark A.4. The normal cone is usually defined as the polar cone to the tangent cone, in the Euclidean metric. It is also possible to define it as the polar cone in the kinetic metric, as follows:

$$N_K^*(q) = \{w \in \mathrm{I\!R}^n \mid w^{\mathrm{T}} M(q) v \leqslant 0,\ \forall\, v \in T_K(q)\} = M^{-1}(q) N_K(q). \tag{A.11}$$

This simply reflects the fact that the normal to a surface $\{x \in \mathbb{R}^n \mid c(x) = 0\} \subset \mathbb{R}^n$, c: $\mathbb{R}^n \to \mathbb{R}$ a smooth function, equipped with a metric $M = M^{\mathrm{T}} > 0$ is equal to $M^{-1}\mathbf{n}$, where $\mathbf{n}$ is the (usual) Euclidean gradient, i.e., $(\frac{\partial c}{\partial x_1}, \frac{\partial c}{\partial x_2}, \ldots, \frac{\partial c}{\partial x_n})^{\mathrm{T}} \in \mathbb{R}^n$. It is possible to express Moreau's impact rule with the kinetic normal cone.

The usefulness of such an operation for numerical purpose is, however, doubtful. Continuing with some equivalences which are used in the formulation of Coulomb's friction, we have the following. Let K be a nonempty convex set:

$$\begin{aligned} y \in \partial \psi_K(x) &\iff x \in K,\ \langle y, z - x\rangle \leqslant 0 \text{ for all } z \in K \\ &\iff x = \mathrm{proj}_K(x + \rho y), \text{ for all } \rho > 0. \end{aligned} \tag{A.12}$$

The last equivalence can be shown as follows. Since the right-hand side is a cone, we have for any $\rho > 0$: $y \in \rho^{-1}\partial\psi_K(x)$, equivalently $\rho y \in \partial\psi_K(x)$. Thus equivalently $-\rho y \in -\partial\psi_K(x) \iff x - (x + \rho y) \in -\partial\psi_K(x)$, from which we deduce using (A.8) that $x = \mathrm{prox}[K; x + \rho y]$. When $K = \mathbb{R}^+$ we get using (A.9) and (A.12):

$$0 \leqslant x \perp y \geqslant 0 \iff y = \mathrm{proj}_K(y - \rho x) \iff x = \mathrm{proj}_K(x - \rho y) \tag{A.13}$$

for all $\rho > 0$.

Remark A.5 (Projection). Throughout the book the projection operator is used and is denoted differently depending on the context, or for the sake of briefness of the expressions. The projection on a set K may be denoted as $P_K(\cdot)$, or as $\mathrm{proj}_K(\cdot)$, or as $\mathrm{proj}[K;\cdot]$. When the projection is made with the metric defined by the kinetic energy, we may write it as $\mathrm{proj}_M[K;\cdot]$ or as $\mathrm{proj}_q[K;\cdot]$.

B

Some Results of Complementarity Theory

Many results on complementarity problems and systems are provided throughout the book. Here we recall some results that concern copositive matrices and the solution of an LCP.

Definition B.1. *A matrix $A \in \mathbb{R}^{n\times n}$ is said to be* strictly copositive *if*

$$x \geqslant 0 \text{ and } x \neq 0 \ \Rightarrow \ x^T A x > 0.$$

Theorem B.2. *If $A \in \mathbb{R}^{n\times n}$ is strictly copositive then the LCP(A,q)*

$$0 \leqslant Ax + q \perp x \geqslant 0$$

has a solution for every $q \in \mathbb{R}^n$. The LCP(A,q) has a unique solution for every $q \in \mathbb{R}^n$ if and only if A is a P-matrix.

Copositive matrices are also useful for studying the Lyapunov stability of some dynamical systems' fixed points like evolution variational inequalities, see Goeleven & Brogliato (2004), and a class of complementarity systems (Camlibel et al., 2006).

Let α and β be subsets of $\{1,2,...,n\}$. In the next theorem $A_{\alpha\beta}$ denotes the submatrix constructed from A by taking rows indexed in α and columns indexed by β.

Theorem B.3. *Let us consider the LCP(A,q) and let A be a P-matrix. Then for each $q \in \mathbb{R}^n$ there exists an index set $\alpha \subset \{1,2,...,n\}$ with complement $\bar{\alpha}$ such that*

- *(i) $-(A_{\alpha\alpha})^{-1}q_\alpha \geqslant 0$ and $q_{\bar{\alpha}} - A_{\bar{\alpha}\alpha}(A_{\alpha\alpha})^{-1}q_\alpha \geqslant 0$,*
- *the unique solution x of the LCP(A,q) is given by $x_\alpha = -(A_{\alpha\alpha})^{-1}q_\alpha$ and $x_{\bar{\alpha}} = 0$.*

In particular the solution mapping $q \mapsto x$ is a piecewise linear function on $\mathbb{R}^n$.

C

Some Facts in Real Analysis

C.1 Functions of Bounded Variations in Time

Let I be an interval, and define a subdivision S_n of I as $x_0 < x_1 < \cdots < x_n$. The variation of a function $f\colon \mathrm{I\!R} \rightarrow \mathrm{I\!R}^n$ on I with respect to the subdivision S_n is defined as

$$\mathrm{var}_{I,S_n}(f) = \sum_{i=0}^{n} ||f(x_{i+1}) - f(x_i)||.$$

The function $f(\cdot)$ is said to have a bounded variation on I if

$$\sup_{S_n} \mathrm{var}_{I,S_n}(f) \leqslant C$$

for some bounded constant C. Then $\mathrm{var}_I(f)$ is called the total variation of $f(\cdot)$ on I. A function that has a bounded variation on any compact subinterval of I is said to be of *local bounded variation* (LBV). If it is right-continuous and LBV it will be denoted RCLBV.

BV functions have the following fundamental properties:

- Let E_f be the set of points x where $f(\cdot)$ has discontinuities. Then E_f is countable.
- If $f(\cdot)$ is BV, then it is Riemann integrable.
- BV functions have left and right limits at all points (of their domain of definition).[1]
- The derivative of a BV function can be decomposed into three parts: a Lebesgue integrable part, a purely atomic measure, and a measure that is singular with respect to the Lebesgue measure and is nonatomic (see below).
- Functions of special bounded variation (SBV) possess a derivative that is the sum of a Lebesgue integrable function, and a purely atomic measure. The third part vanishes for SBV functions.

[1] Throughout the book the right (left) limits at t are denoted either as $f^+(t)$ ($f^-(t)$) or as $f(t^+)$ ($f(t^-)$).

In most engineering applications, it may be reasonably assumed that the derivative of a BV function is just the sum of an integrable function and a purely atomic measure of the form $\sum_i \delta_i$ for some set of i:

- We denote by $\mathrm{LBV}(I;\mathbb{R}^n)$ the space of functions of locally bounded variation, i.e., of bounded variation on every compact subinterval of I.
- We denote by $\mathrm{RCLBV}(I;\mathbb{R}^n)$ the space of right-continuous functions of locally bounded variation. It is known that if $x \in \mathrm{RCLBV}(I;\mathbb{R}^n)$ and $[a,b]$ denotes a compact subinterval of I, then x can be represented in the form (see, e.g., Shilov & Gurevich (1966))

  $$x(t) = \mathscr{J}_x(t) + [x](t) + \zeta_x(t), \forall t \in [a,b],$$

  where $\mathscr{J}_x$ is a jump function, $[x]$ is an absolutely continuous function, and ζ_x is a singular function. Here $\mathscr{J}_x$ is a jump function in the sense that $\mathscr{J}_x$ is right continuous and given any $\varepsilon > 0$, there exist finitely many points of discontinuity $t_1,...,t_N$ of $\mathscr{J}_x$ such that $\sum_{i=1}^N \| \mathscr{J}_x(t_i) - \mathscr{J}_x(t_i^-)\| + \varepsilon > \ \mathrm{var}(\mathscr{J}_x,[a,b])$, $[x]$ is an absolutely continuous function in the sense that for every $\varepsilon > 0$, there exists $\delta > 0$ such that $\sum_{i=1}^N \|[x](\beta_i) - [x](\alpha_i)\| < \varepsilon$, for any collection of disjoint subintervals $]\alpha_i,\beta_i] \subset [a,b] (1 \leqslant i \leqslant N)$ such that $\sum_{i=1}^N (\beta_i - \alpha_i) < \delta$, and ζ_x is a singular function in the sense that ζ_x is a continuous and of bounded variation function on $[a,b]$ such that $\dot{\zeta}_x = 0$ almost everywhere on $[a,b]$.
- By $u \in \mathrm{RCSLBV}(I;\mathbb{R}^n)$ it is meant that x is a right-continuous function of special locally bounded variation, i.e., x is of bounded variation and can be written as the sum of a jump function and an absolutely continuous function on every compact subinterval of I. So, if $x \in \mathrm{RCSLBV}(I;\mathbb{R}^n)$ then

  $$x = [x] + \mathscr{J}_x, \tag{C.1}$$

  where $[x]$ is a locally absolutely continuous function called the absolutely continuous component of x and $\mathscr{J}_x$ is uniquely defined up to a constant by

  $$\mathscr{J}_x(t) = \sum_{t \geqslant t_n} x(t_n^+) - x(t_n^-) = \sum_{t \geqslant t_n} x(t_n) - x(t_n^-), \tag{C.2}$$

  where $t_1, t_2, ..., t_n, ...$ denote the countably many points of discontinuity of x in I.

C.2 Multifunctions of Bounded Variation in Time

A moving set $t \mapsto K(t)$ is said to be right-continuous bounded variation in time on $[0,T]$, if there exists a right-continuous nondecreasing function r: $[0,T] \to \mathbb{R}$ such that

$$d_{\mathrm{H}}(K(t),K(s)) \leqslant r(t) - r(s), \text{ for all } 0 \leqslant s \leqslant t \leqslant T.$$

Let $r(0) = 0$. For any partition $0 = t_0 < t_1 < \cdots < t_N = T$ of $[0,T]$, this yields

$$\sum_{i=0}^{N-1} d_{\mathrm{H}}(K(t_{i+1}),K(t_i)) \leqslant \sum_{i=0}^{N-1}[r(t_{i+1})-r(t_i)] = r(T).$$

Therefore the first inequality can be interpreted as requiring that $t \mapsto K(t)$ is of bounded variation. We conclude that the above definition of the variation of a function can be extended to set-valued functions where the Euclidean distance is replaced by the Hausdorff distance.

C.3 Distributions Generated by RCLSBV Functions

The material in this section is entirely taken from Sect. 2 of Acary et al. (in press). Let I be the real interval given by

$$I = [\alpha, \beta),$$

where $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R} \cup \{+\infty\}$. The support supp$\{\varphi\}$ of a function $\varphi\colon I \to \mathbb{R}$ is defined by $\mathrm{supp}\{\varphi\} := \overline{\{t \in I \mid \varphi(t) \neq 0\}}$. We denote by $C_0^\infty(I)$ the space of real-valued $C^\infty(I)$-mappings with compact support contained in the open interval (α,β), and $\mathscr{D}'(I)$ is the space of Schwartz distributions on I, i.e., the space of linear continuous forms on $C_0^\infty(I)$. Recall that for $T \in \mathscr{D}'(I)$, the (generalized) derivative of T is defined by

$$\langle DT, \varphi \rangle = -\langle T, \dot{\varphi} \rangle, \forall\ \varphi \in C_0^\infty(I).$$

The (generalized) derivative of order n is then given by

$$D^n T = D(D^{n-1}T)\ (n \geqslant 2)$$

that is

$$\langle D^n T, \varphi \rangle = (-1)^n \langle T, \varphi^{(n)} \rangle, \forall\ \varphi \in C_0^\infty(I).$$

For $a \in I$, we denote by δ_a the Dirac distribution at a, defined by

$$\langle \delta_a, \varphi \rangle = \varphi(a), \forall\ \varphi \in C_0^\infty(I).$$

Note that $\delta_a = D\mathscr{H}(.-a)$ where $\mathscr{H}$ is the Heaviside function:

$$\mathscr{H}(t) = \begin{cases} 1 & \text{if } t \geqslant 0 \\ 0 & \text{if } t < 0. \end{cases} \tag{C.3}$$

The support supp$\{T\}$ of a distribution $T \in \mathscr{D}'(I)$ is defined by $\mathrm{supp}\{T\} \overset{\delta}{=} I \backslash \mathscr{O}$ where $\mathscr{O} \subset I$ denotes the largest open set in I on which T vanishes in the sense that $\langle T, \varphi \rangle = 0, \forall\ \varphi \in C_0^\infty(I)$ with support contained in $\mathscr{O}$:

- Let $h \in \mathrm{RCSLBV}(I;\mathbb{R})$ be given. We will denote by $E_0(h)$ the countable set of points of discontinuity $t_1, t_2, ..., t_k, ...$ of h. As seen above, h can be written as the sum of a locally absolutely continuous function $[h]$ and the locally jump function $\mathscr{J}_h$ given by

$$\mathscr{J}_h(t) = \sum_{t \geqslant t_k} \sigma_h(t_k),$$

where for $t \in I$,

$$\sigma_h(t) \overset{\Delta}{=} h(t^+) - h(t^-) = h(t) - h(t^-)$$

denotes the jump of h at t. It is clear that if $t \in I \backslash E_0(h)$ then $\sigma_h(t) = 0$.

- We will denote by $\hat{h}^{(1)}(t)$ the right derivative (if it exists) of the absolutely continuous part $[h]$ of $h \in \mathrm{RCSLBV}(I;\mathrm{I\!R})$ at t, i.e.,

$$\hat{h}^{(1)}(t) \overset{\Delta}{=} \frac{\mathrm{d}^+[h]}{\mathrm{d}t}(t) = \lim_{\sigma \to 0^+} \frac{[h](t+\sigma) - [h](t)}{\sigma}.$$

We have thus

$$h = [h] + \mathscr{J}_h \tag{C.4}$$

and

$$\mathrm{d}h = \hat{h}^{(1)}\mathrm{d}t + \mathrm{d}\mathscr{J}_h \tag{C.5}$$

The measure $\mathrm{d}\mathscr{J}_h$ is atomic as a measure concentrated on the set $E_0(h)$ of countably many points of discontinuity of h in I, i.e., $\mathrm{d}\mathscr{J}_h(A) = 0, \forall A \in \mathscr{B}(\mathrm{I\!R}), A \subset I \backslash E_0(h)$.

- Let us now set

$$\begin{cases} \mathscr{F}_0(I;\mathrm{I\!R}) = \mathrm{RCSLBV}(I;I\!\!R) \\ \\ \mathscr{F}_1(I;\mathrm{I\!R}) = \{h \in \mathscr{F}_0(I;\mathrm{I\!R}) : \hat{h}^{(1)} \in \mathrm{RCSLBV}(I;\mathrm{I\!R})\} \\ \\ \mathscr{F}_2(I;\mathrm{I\!R}) = \{h \in \mathscr{F}_1(I;\mathrm{I\!R}) : \hat{h}^{(2)} \overset{\Delta}{=} \frac{\mathrm{d}^+}{\mathrm{d}t}[\hat{h}^{(1)}] \in \mathrm{RCSLBV}(I;\mathrm{I\!R})\} \\ \\ \vdots \\ \\ \mathscr{F}_k(I;\mathrm{I\!R}) = \{h \in \mathscr{F}_{k-1}(I;\mathrm{I\!R}) : \hat{h}^{(k)} \overset{\Delta}{=} \frac{\mathrm{d}^+}{\mathrm{d}t}[\hat{h}^{(k-1)}] \in \mathrm{RCSLBV}(I;\mathrm{I\!R})\} \end{cases} \tag{C.6}$$

and

$$\mathscr{F}_\infty(I;\mathrm{I\!R}) = \cap_{k \in \mathrm{I\!N}} \mathscr{F}_k(I;\mathrm{I\!R}).$$

We standardize the notation by setting $\hat{h}^{(0)} \overset{\Delta}{=} h$. Note that $\hat{h}^{(\alpha)} \in \mathrm{RCSLBV}(I;\mathrm{I\!R})$ $(\alpha \geqslant 1)$ means that the absolutely continuous function $[\hat{h}^{(\alpha-1)}]$ admits a right derivative $\hat{h}^{(\alpha)}(t) = \frac{\mathrm{d}^+}{\mathrm{d}t}[\hat{h}^{(\alpha-1)}](t)$ at each $t \in I$ and $\hat{h}^{(\alpha)}$ is of special local bounded variation over I.

Let $n \in \mathrm{I\!N}$ be given.

Definition C.1. *We say that a Schwartz distribution $T \in \mathscr{D}'(I)$ is of class $\mathscr{T}_n$ on I provided that there exists a function $F \in \mathscr{F}_\infty(I;\mathrm{I\!R})$ such that $T = D^n F$.*

Let us now denote by $\mathscr{T}_n(I)$ the set of all distributions of class $\mathscr{T}_n$ on I, i.e.,

$$\mathscr{T}_n(I) = \{T \in \mathscr{D}'(I) : \exists F \in \mathscr{F}_\infty(I;\mathrm{I\!R}) \text{ such that } T = D^n F\}.$$

It is clear that

$$\mathscr{T}_0(I) = \mathscr{F}_\infty(I;I\!\!R).$$

If $T \in \mathscr{T}_1(I)$ then there exists $F \in \mathscr{F}_\infty(I;\mathrm{I\!R})$ such that

$$\langle T,\varphi\rangle = \int_I \varphi\, \mathrm{d}\hat{F}^{(0)} = \int_I \varphi\, \mathrm{d}F, \forall\varphi \in C_0^\infty(I).$$

More generally, if $T \in \mathscr{T}_n(I)$ for some $n \geqslant 2$, then there exists $F \in \mathscr{F}_\infty(I;\mathrm{I\!R})$ such that, for all $\varphi \in C_0^\infty(I)$,

$$\begin{aligned}\langle T,\varphi\rangle =& \int_I \varphi\, \mathrm{d}\hat{F}^{(n-1)} + \sum_{i=2}^{n}\Big(\sum_{t_k\in E_0(\hat{F}^{(n-i)})\cap\mathrm{supp}\{\varphi\}} (\hat{F}^{(n-i)}(t_k^+) - \hat{F}^{(n-i)}(t_k^-))\langle\delta_{t_k}^{(i-1)},\varphi\rangle\Big)\\ =& \int_I \varphi\, \hat{F}^{(n)}\mathrm{d}t + \sum_{i=1}^{n}\Big(\sum_{t_k\in E_0(\hat{F}^{(n-i)})\cap\mathrm{supp}\{\varphi\}} \sigma_{\hat{F}^{(n-i)}}(t_k)\langle\delta_{t_k}^{(i-1)},\varphi\rangle\Big)\,. \end{aligned} \tag{C.7}$$

For a distribution $T \in \mathscr{T}_n(I)$, as expressed in (C.7), we may clearly identify the "function part" $\{T\}$ and the "measure part" $\ll T \gg$ respectively by

$$\{T\} = \hat{F}^{(n)} \tag{C.8}$$

and (if $n \geqslant 1$)

$$\langle \ll T \gg, \varphi\rangle = \int_I \varphi\, \mathrm{d}\hat{F}^{(n-1)},\ \forall\ \varphi \in C_0^\infty(I)\,. \tag{C.9}$$

We will also use the notation d $\ll$ T $\gg$ to denote the Stieltjes measure d$\hat{F}^{(n-1)}$ generated by $\hat{F}^{(n-1)} \in \mathrm{RCSLBV}(I;I\!\!R)$. Here $\{T\}$ is a RCSLBV function and d $\ll$ T $\gg$ is a Stieltjes measure. For pedagogical reasons, we use the two different notation d $\ll$ T $\gg$ and $\ll$ T $\gg$ to denote, respectively, the Radon measure defined on the Borel sets and the corresponding distribution, i.e.,

$$\langle \ll T \gg, \varphi\rangle = \int_I \varphi\ \ \mathrm{d} \ll T \gg,\ \forall\ \varphi \in C_0^\infty(I)\,.$$

It will be also convenient to use the notation $\{T^{(k)}\}$ to denote the "function part" of D^kT, i.e.,

$$\{T^{(k)}\} = \{D^kT\} = \hat{F}^{(n+k)}\,.$$

Definition C.2. *We say that a Schwartz distribution $T \in \mathscr{D}'(I)$ is of class $\mathscr{T}_\infty$ on I provided that there exist $n \in \mathrm{I\!N}$ and a function $F \in \mathscr{F}_\infty(I;\mathrm{I\!R})$ such that $T = D^nF$.*

Defining $\mathscr{T}_\infty$ therefore allows one to encompass all distributions of class $\mathscr{T}_n$, $n \in \mathbb{N}$. The set of Schwartz distributions of class $\mathscr{T}_\infty$ on I will be denoted by $\mathscr{T}_\infty(I)$. It is clear that

$$\mathscr{T}_\infty(I) = \cup_{n\in\mathbb{N}} \mathscr{T}_n(I).$$

For $T \in \mathscr{T}_\infty(I)$, we define the degree "deg(T)" of T in the following way: Let n be the smallest integer such that $T \in \mathscr{T}_n(I)$, we set

$$\deg(T) = \begin{cases} n+1 \text{ if } n \geqslant 1 \\ 1 \text{ if } n=0 \text{ and } E_0(\{T\}) \neq \emptyset \\ 0 \text{ if } n=0 \text{ and } E_0(\{T\}) = \emptyset \end{cases}. \tag{C.10}$$

Remark C.3. The distributions of degree 0 are the continuous $\mathscr{F}_\infty$-functions while the distributions of degree 1 are the discontinuous $\mathscr{F}_\infty$-functions. The right-continuous Heaviside function is of degree 1, the Dirac distribution δ_a $(a \in I)$ is of degree 2, the distribution $D^{(n)}\delta_a$ $(a \in I)$ is of degree $n+1$.

C.4 Differential Measures

Details on differential measures may be found in (Schwartz, 1993; Monteiro Marques, 1993; Moreau, 1988a).

Definition C.4. *Let $x : I \to \mathbb{R}^n$ be a BV function, $I \neq \emptyset$, $I \subseteq \mathbb{R}$. Let $\varphi(\cdot)$ be a continuous real function on I, with compact support. Let $\mathscr{P}$ denote the set of finite partitions of I, each partition P_N with nodes $t_0 < t_1 < \ldots < t_N$. Let $\theta_k \in [t_{k-1}, t_k]$ for all intervals of the partition P_N. The Riemann–Stieltjes sums $S(\varphi, P_N, \theta; x) = \sum_{k=1}^{N} \varphi(\theta_k)(x(t_k) - x(t_{k-1}))$ converge as $N \to +\infty$ to a limit independent of the θ_k. This limit is denoted as*

$$\int \varphi \, dx \tag{C.11}$$

where dx is the differential measure associated to $x(\cdot)$. The map $x \mapsto dx$ is linear.

If $x(\cdot)$ is constant, $\mathrm{d}x = 0$. If $\mathrm{d}x = 0$ and $x(\cdot)$ is right-continuous in the interior of I, then $x(\cdot)$ is constant. If $x(\cdot)$ is a step function, then $\mathrm{d}x$ is the sum of a finite collection of Dirac measures with atoms at the discontinuity points of $x(\cdot)$. For $a \leqslant b$, $a, b \in I$,

$$\begin{aligned}
\mathrm{d}x([a,b]) &= x(b^+) - x(a^-), \\
\mathrm{d}x([a,b)) &= x(b^-) - x(a^-), \\
\mathrm{d}x((a,b]) &= x(b^+) - x(a^+), \\
\mathrm{d}x((a,b)) &= x(b^-) - x(a^+).
\end{aligned}$$

In particular, we have

$$\mathrm{d}x(\{a\}) = x(a^+) - x(a^-)$$

For $x \in \mathrm{LBV}(I;\mathbb{R}^n)$, x^+ and x^- denote the functions defined by

$$x^+(t) = x(t^+) = \lim_{s\to t, s>t} x(s), \;\; \forall t \in I, t < \sup\{I\}$$

and

$$x^-(t) = x(t^-) = \lim_{s\to t, s<t} x(s), \;\; \forall t \in I, t > \inf\{I\}$$

(where $\sup\{I\}$ (resp. $\inf\{I\}$) denotes the supremum (resp. infinum) of the set I). If $x, y \in \mathrm{LBV}(I;\mathbb{R}^n)$ then $x^{\mathrm{T}}y \in \mathrm{LBV}(I;\mathbb{R})$ and

$$\mathrm{d}(x^{\mathrm{T}}y) = (y^-)^{\mathrm{T}}\mathrm{d}x + (x^+)^{\mathrm{T}}\mathrm{d}y = (y^+)^{\mathrm{T}}\mathrm{d}x + (x^-)^{\mathrm{T}}\mathrm{d}y. \tag{C.12}$$

Let us also recall that

$$2(x^-)^{\mathrm{T}}\mathrm{d}x \leqslant \mathrm{d}(x^{\mathrm{T}}x) = (x^+ + x^-)^{\mathrm{T}}\mathrm{d}x \leqslant 2(x^+)^{\mathrm{T}}\mathrm{d}x. \tag{C.13}$$

C.5 Bohl's Distributions

Definition C.5. *A Bohl function $f(\cdot)$ is a continuous function having rational Laplace transforms, defined on $\mathbb{R}^+$ into $\mathbb{R}^n$. One has $f(t) = A\exp(Bt)C$ for some matrices A, B, C of suitable dimensions. The one-sided Laplace transform of a Bohl function is given by $\hat{f}(s) = A(sI_n - B)^{-1}C$. It is rational and strictly proper (i.e., its denominator has a degree strictly larger than the degree of the numerator). The inverse Laplace transform of a rational, strictly proper function, is a Bohl function.*

Bohl functions appear as the solutions of linear differential equations with constant coefficients. They are constants, exponentials, sines, and cosines.

Definition C.6. *A Schwartz' distribution T is a Bohl distribution, if it can be decomposed as $T = T_{imp} + T_{reg}$, where T_{reg} is a Bohl function, and $T_{imp} = \sum_{i=0}^{l} T^i \delta_0^{(i)}$, where $\delta_t^{(i)}$ is the ith derivative of the Dirac measure with atom at time t, and some real numbers T^i. The integer $l+2$ may be called the degree of the distribution.*

Hence the Dirac measure has a degree 2. Notice that the Bohl's distributions are a particular case of the distributions generated by RCSLBV functions described in Sect. C.3.

C.6 Some Useful Results

Lemma C.7 (Gronwall's Lemma). *Suppose $f\colon \mathbb{R}^+ \to \mathbb{R}^+$ is a continuous function, and constants $b \geqslant 0$ and $c \geqslant 0$ are given. Then if*

$$f(t) \leqslant b + \int_0^t f(s)ds, \ \textit{for all } t \geqslant 0 \tag{C.14}$$

it follows that

$$f(t) \leqslant b \ \exp(ct) \tag{C.15}$$

The next result is proposition 7.1.1 in Glocker (2001).

Proposition C.8. *Let $f\colon t \mapsto f(t)$ be a right-continuous function, and let $F(t) = F(0) + \int_0^t f(s)ds$. If $f(0) > 0$, then there exists a time interval $(0,t^*)$, $t^* > 0$, such that $F(t) > F(0)$ for all $t \in (0,t^*)$.*

This proposition allows one to deduce some information on the evolution of the position of a mechanical system, when the velocity is right-continuous and positive. For instance, if a constraint $h(q)$ is active on some time interval I (i.e., $h(q(t)) = 0$ for all $t \in I$), then $\frac{\mathrm{d}^+}{\mathrm{d}t}(h(q(t)) = \nabla h^{\mathrm{T}}(q(t))\dot{q}(t^+) > 0$ implies that $h(q(\cdot)) > 0$ in a right neighborhood of t. The same reasoning may be applied when both $h(q(t)) = 0$ and $\nabla h^{\mathrm{T}}(q(t))\dot{q}(t^+) = 0$ on I, using this time the acceleration. And so on. Let us state a similar result with more regularity assumptions.

Let $f\colon \mathbb{R} \to \mathbb{R}$ be n times right differentiable, i.e., for all $t \in \mathbb{R}$ there exists $\eta > 0$ such that $\dot{f}^+(\tau) = \lim_{h \to 0, h>0} \frac{f(\tau+h)-f(\tau)}{h}$ exists on $[t, t+\eta)$, $\ddot{f}^+(\tau) = \lim_{h\to 0,h>0} \frac{\dot{f}^+(\tau+h)-\dot{f}^+(\tau)}{h}$ exists on $[t,t+\eta)$, and so on. For the ease of writing let us drop the $+$ upperscript. Denote $Df_j(t) = (f^{(j)}(t), f^{(j+1)}(t), ..., f^{(n)}(t))$, $1 \leqslant j \leqslant n$. The lexicographical inequality $x \succ 0$, $x \in \mathbb{R}^{1\times n}$ is a row, means that the first nonzero element of x is positive, and $x \neq 0$.

Proposition C.9. *Under the stated assumptions, suppose that for some $t \in \mathbb{R}$ one has $Df_j(t) \succ 0$ for some $1 \leqslant j \leqslant n$. Then $f^{(j-1)}(s) > f^{(j-1)}(\tau)$ for all $s \in (t,t^*)$ and some $t^* > t$.*

Proof: There exists an i with $j \leqslant i \leqslant n$ such that $f^{(i)}(t) > 0$. We may then apply iteratively Proposition C.8. □

References

M. Abadie. *Simulation Dynamique de Mécanismes, Prise en Compte du Contact Frottant*. PhD thesis, Université des sciences et techniques du Languedoc, Montpellier II, 1998.

M. Abadie. Dynamic simulation of rigid bodies: modelling of frictional contact. B. Brogliato, editor, *Impacts in Mechanical Systems: Analysis and Modelling*, volume 551 of *Lecture Notes in Physics (LNP)*, pp. 61–144. Springer, 2000.

V. Acary. *Contribution to the Numerical and Mechanical Modelling of Masonry Buildings*. PhD thesis, Université d'Aix-Marseille II, 2001. In French.

V. Acary & B. Brogliato. Concurrent multiple impacts modelling – case-study of a 3-ball chain. K.J. Bathe, editor, *Second MIT Conference on Computational Fluid and Solid Mechanics*, pp. 1842–1847. Elsevier, June 2003.

V. Acary & B. Brogliato. Numerical time integration of higher order *dynamical systems with* state constraints. *Euromech Conference ENOC-2005, Eindhoven, August 7–12*, 2005.

V. Acary & B. Brogliato. Higher order Moreau's sweeping process. P. Alart, O. Maisonneuve & R. T. Rockafellar, editors, *Chapter 22 of Nonsmooth Mechanics and Analysis. Theoretical and Numerical Advances*, pp. 261–278. Springer, 2006.

V. Acary & M. Jean. Numerical simulation of monuments by the contacts dynamics method. DGEMN-LNEC-JRC, editor, *Monument-98, Workshop on Seismic Perfomance of Monuments*, pp. 69–78. Laboratório Nacional de engenharia Civil (LNEC), November 12–14 1998.

V. Acary & M. Jean. Numerical modeling of three dimensional divided structures by the nonsmooth contact dynamics method: application to masonry structure. B.H.V. Topping, editor, *The Fifth International Conference on Computational Structures Technology 2000*, pp. 211–222. Civil-Comp Press, September 6–8 2000.

V. Acary & Y. Monerie. Nonsmooth fracture dynamics using a cohesive zone approach. INRIA Research Report 6032, http://hal.inria.fr/docs/00/11/66/87/PDF/RR-6032.pdf, 2006.

V. Acary & F. Pérignon. An introduction to Siconos. Technical Report TR-0340, INRIA, http://hal.inria.fr/inria-00162911/en/, 2007.

V. Acary, B. Brogliato & D. Goeleven. Higher order Moreau's sweeping process: mathematical formulation and numerical simulation. *Mathematical Programming Series A*, in press.

V. Acary, I.-C Morărescu, F. Pérignon & B. Brogliato. Numerical simulation of nonsmooth systems and switching control with the SICONOS/control toolbox. *6th Euromech Conference ENOC*, Saint Petersburg, Russia, June, 30- July, 4, 2008.

V. Acary, J.Y. Blaise, P. Drap, M. Florenzano, S. Garrec, M. Jean & D. Merad. NSCD method applied to mechanical simulation of masonry in historical buildings using MOMA. *XVII CIPA International Symposium*. CIPA (International Committee for Architectural Photogrammetry), Olinda, Brazil, October, 3–6 1999.

K. Addi, S. Adly, B. Brogliato & D. Goeleven. A method using the approach of Moreau and Panagiotopoulos for the mathematical formulation of non-regular circuits in electronics. *Nonlinear Analysis: Hybrid Systems*, 1, pp. 30–43, 2007.

S. Adly & D. Goeleven. A stability theory for second-order nonsmooth dynamical systems with application to friction problems. *Journal de Mathématiques Pures et Appliquées*, 83, pp. 17–51, 2004.

P.K. Agarwal, L.J. Guibas, H. Edelsbrunner, J. Erickson, M. Isard, S. Har-Peled, J. Hershberger, C. Jensen, L. Kavraki, P. Koehl, M. Lin, D. Manocha, D. Metaxas, B. Mirtich & D. Mount. Algorithmic issues in modeling motion. *ACM Computing Surveys*, 34, pp. 550–572, 2002.

P. Alart. Critère d'injectivité et de surjectivité pour certaines applications de $\mathbb{R}^n$ dans lui-même: application à la mécanique du contact. *RAIRO, Modélisation Mathématique et Analyse Numérique*, 2(27), pp. 203–222, 1993.

P. Alart & A. Curnier. A mixed formulation for frictional contact problems prone to Newton like solution method. *Computer Methods in Applied Mechanics and Engineering*, 92(3), pp. 353–375, 1991.

P. Alart & F. Lebon. Solution of frictional contact problems using ILU and coarse/fine preconditioners. *Computational Mechanics*, 16(2), pp. 98–105, 1995.

P. Alart, M. Barboteu & M. Renouf. Parallel computational strategies for multicontact problems: applications to cellular and granular media. *International Journal for Multiscale Computational Engineering*, 1(4), pp. 419–430, 2003.

A.M. Al-Fahed, G.E. Stavroulakis & P.D. Panagiotopulos. Hard and soft fingered robot grippers. The linear complementarity approach. *Zeitschrift für Angewandte Mathematik und Mechanik*, 71, pp. 257–265, 1991.

F. Al-Khayyal. An implicit enumeration procedure for the general linear complementarity problem. *Mathematical ProgrammingStudy*, 31, pp. 1–20, 1987.

E.D. Andersen, J. Gondzio, C. Meszaros & X. Xu. Implementation of interior point methods for large scale linear programming. Papers 96.3, Ecole des Hautes Etudes Commerciales, Universite de Geneve, 1996. Available at http://ideas.repec.org/p/fth/ehecge/96.3.html.

M. Anistescu. Optimization-based simulation of nonsmooth rigid multibody dynamics. *Mathematical Programming Series A*, 105, pp. 113–143, 2006.

M. Anistescu & G.D. Hart. A constraint-stabilized time-stepping approach for rigid multibody dynamics with joints, contact and friction. *International Journal for Numerical Methods in Engineering*, 60, pp. 2335–2371, 2004.

M. Anitescu & G.D. Hart. A fixed-point iteration approach for multibody dynamics with contact and small friction. *Mathematical Programming Series B*, 101, pp. 3–32, 2004.

M. Anitescu & F.A. Potra. Formulating dynamic multi-rigid-body contact problems with friction as solvable linear complementarity problems. *Nonlinear Dynamics, Transactions of A.S.M.E.*, 14, pp. 231–247, 1997.

M. Anitescu, G. Lesaja & F. Potra. Equivalence between different formulations of the linear complementarity problem. Technical Report 71, University of Iowa Technical Reports in Computational Mathematics, Iowa City, IA 52242, USA, 1995.

M. Anitescu, G. Lesaja & F. Potra. An infeasible–interior–point predictor–corrector algorithm for the P_*–geometric LCP. *Applied Mathematics and Optimization*, 36, pp. 203–228, 1997.

M. Anitescu, F.A. Potra & D.E. Stewart. Time-stepping for the three dimensional rigid body dynamics. *Computer Methods in Applied Mechanics and Engineering*, 177, pp. 183–4197, 1999.

S.S. Antman. The simple pendulum is not so simple. *SIAM Review*, 40, pp. 927–930, 1998.

V.I. Arnold. *Mathematical Methods of Classical Mechanics.* Graduate Texts in Mathematics. Springer, 1989.

J.P. Aubin & A. Cellina. *Differential Inclusions: Set-Valued Maps and Viability Theory.* Springer, Berlin, 1984.

G. Auchmuty. Variational principles for variational inequalities. *Numerical Functional Analysis and Optimization*, 10, pp. 863–874, 1989.

A. Auslender. *Optimisation. Méthodes Numériques.* Masson, Paris, 1976.

D.D. Bainov & P.S. Simeonov. *Systems with Impulse Effect Stability, Theory and Applications. Ellis Horwood Series in Mathematics and Its Applications*. Ellis Horwood Limited, Chichester, 1989.

P. Ballard. The dynamics of discrete mechanical systems with perfect unilateral constraints. *Archives for Rational Mechanics and Analysis*, 154, pp. 199–274, 2000.

D. Baraff. Curved surfaces and coherence for non-penetrating rigid body simulation. *Computer Graphics (Proceedings of SIGGRAPH)*, 24, pp. 19–28, 1990.

D. Baraff. Issues in computing contact forces for nonpenetrating rigid bodies. *Algorithmica*, 10, pp. 292–352, 1993.

D. Baraff. Linear-time dynamics using Lagrange multipliers. *Proceedings of SIGGRAPH 1996, CDROM Version*, pp. 137–146, 1996.

J. Bastien & C.H. Lamarque. Persoz' gephyroidal model described by a maximal monotone differential inclusion. *Archives of Applied Mechanics*, 2007. in press.

J. Bastien & M. Schatzman. Numerical precision for differential inclusions with uniqueness. *ESAIM M2AN: Mathematical Modelling and Numerical Analysis*, 36(3), pp. 427–460, 2002.

J. Bastien, M. Schatzman & C.H. Lamarque. Study of some rheological models with a finite number of degrees of freedom. *European Journal of Mechanics A/Solids*, 19, pp. 277–307, 2000.

J. Bastien, G. Michon, L. Manin & R. Dufour. An analysis of the modified Dahl and Masing models: application to a belt tensioner. *Journal of Sound and Vibrations*, 302, pp. 841–864, 2007.

J. Baumgarte. Stabilization of constraints and integral of motion for nonpenetrating rigid bodies. *Computer Methods in Applied Mechanics and Engineering*, 1, pp. 1–16, 1972.

D. Bedrosian & J. Vlach. Time-domain analysis of networks with internally controlled switches. *IEEE Transactions on Circuits and Systems– I: Fundamental Theory and Applications*, 39, pp. 199–212, 1992.

H. Benabdellah, C. Castaing, A. Salvadori & A. Syam. Nonconvex sweeping process. *Journal of Applied Analysis*, 2(2), pp. 217–240, 1996.

H. Benabdellah, C. Castaing & A. Salvadori. Compactness and discretization methods for differential inclusions and evolution problems. *Atti del Seminario Mathematico e Fisico dell Universita di Modena*, XLV, pp. 9–51, 1997.

D. Bertsekas. On the Goldstein–Levitin–Polyak gradient projection method. *IEEE Transactions on Automatic Control*, 21(2), pp. 174–184, 1976.

D.P. Bertsekas. Projected Newton methods for optimization problems with simple constraints. *SIAM Journal on Control and Optimization*, 20, pp. 221–246, 1982.

D.P. Bertsekas & J.N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Prentice–Hall Inc., Englewood Cliffs, NJ, 1989.

S.C. Billups. *Algorithms for Complementarity Problems and Generalized Equations*. PhD thesis, University of Wisconsin, Madison, 1995. Available as Technical Report MP-TR-1995-14.

S.C. Billups & M.C. Ferris. QPCOMP: a quadratic programming based solver for mixed complementarity problems. Technical Report MP-TR-1995-09, University of Wisconsin, Madison, 1995.

S.C. Billups, S.P. Dirkse & M.C. Ferris. A comparison of large scale mixed complementarity problem solvers. *Computational Optimization and Applications*, 7, pp. 3–25, 1997.

P.A. Bliman & M. Sorine. Easy-to-use realistic dry friction models for automatic control. *3rd European Control Conference*, pp. 3788–3794, Roma, Italy, September 1995.

B. Bona & M. Indri. Friction compensation in robotics: an overview. *44th IEEE Conference on Decision and Control, and the European Control Conference 2005*, pp. 4360–4367, Seville, Spain, December 2005.

J.F. Bonnans & C.C. Gonzaga. Convergence of interior-point algorithms for the monotone linear complementarity problem. *Mathematics of Operations Research*, 21, pp. 1–25, 1996.

J.F. Bonnans, J.C. Gilbert, C. Lemarèchal & C.A. Sagastizábal. *Numerical Optimization: Theoretical and Practical Aspects*. Springer, 2003.

U. Boscain. Stability of planar switched systems: the linear single input case. *SIAM Journal of Control and Optimization*, 41(1), pp. 89–112, 2002.

J.-M. Bourgeot & B. Brogliato. Tracking control of complementary lagrangian systems. *International Journal of Bifurcation and Chaos*, 15(6), pp. 1839–1866, 2005.

R.M. Brach. *Mechanical Impact Dynamics*. Wiley, New York, 1990.

A. Brandt & C.W. Cryer. Multigrid algorithms for the solution of linear complementarity problems arising from free boundary problems. *SIAM Journal of Scientific and Statistical Computing*, 4, pp. 655–684, 1983.

I. Bratberg, F. Radjai & A. Hansen. Dynamic rearrangements and packing regimes in randomly deposited two-dimensional granular beds. *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics*, 66(3), page (3 Pt 1):031303, 2002.

K.E. Brenan, S. Campbell & L.R. Petzold. *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*. North-Holland, Amsterdam, 1989.

A. Bressan & F. Rampazzo. On differential systems with quadratic impulses and their applications to Lagrangian mechanics. *SIAM Journal of Control and Optimization*, 31(5), pp. 1205–1220, 1993.

H. Brezis. *Opérateurs Maximaux Monotones et Semi-groupe de Contraction dans les Espaces de Hilbert*. North Holland, Amsterdam, 1973.

H. Brezis & M. Sibony. Méthodes d'approximation et d'itération pour les opérateurs monotones. *Archives for Rational Mechanics and Analysis*, 28, pp. 59–82, 1967/1968.

B. Brogliato. *Nonsmooth Mechanics: Models, Dynamics and Control, 2nd Edition*. Springer, London, 1999.

B. Brogliato. Some perspectives on the analysis and control of complementarity systems. *IEEE Transactions on Automatic Control*, 48(6), pp. 918–935, 2003.

B. Brogliato. Absolute stability and the Lagrange–Dirichlet theorem with monotone multivalued mappings. *Systems and Control Letters*, 51, pp. 343–353, 2004.

B. Brogliato & D. Goeleven. The Krakovskii–LaSalle invariance principle for a class of unilateral dynamical systems. *Mathematics of Control, Signals, and Systems*, 17, pp. 57–76, 2005.

B. Brogliato & L. Thibault. Well-posedness results for non-autonomous dissipative complementarity systems. *INRIA research Report RR-5931, http://hal.inria.fr/docs/00/07/95/71/PDF/RR-5931.pdf*, 2006.

B. Brogliato, S.-I. Niculescu & P. Orhant. On the control of finite-dimensional mechanical systems with unilateral constraints. *IEEE Trans. Autom. Contr.*, 42(2), pp. 200–215, 1997.

B. Brogliato, A. Daniilidis, C. Lemaréchal & V. Acary. On the equivalence between complementarity systems, projected systems and differential inclusions. *Systems and Control Letters*, 55, pp. 45–51, 2006.

B. Brogliato, R. Lozano, B. Maschke & O. Egeland. *Dissipative Systems Analysis and Control, 2nd Edition*. Springer, 2007.

J.R. Bunch & L. Kaufman. Some stable methods for calculating inertia and solving symmetric linear systems. *Mathematics of Computation*, 31(137), pp. 163–179, 1977.

J.R. Bunch & B.N. Parlett. Direct methods for solving symmetric indefinite syetms of linear equations. *SIAM Journal on Numerical Analysis*, 8(639–655), 1971.

J.R. Bunch, L. Kaufman & B.N. Parlett. Decomposition of a symmetric system. *Handbook Series Linear Algebra. Numerische Mathematik*, 27, pp. 95–109, 1976.

J.C. Butcher. *The Numerical Analysis of Ordinary Differential Equations–Runge-Kutta and General Linear Methods*. Wiley, New York, 1987.

A. Cabot & L. Paoli. Asymptotics for some vibro-impact problems with a linear dissipation term. *Journal de Mathématiques Pures et Appliquées*, 87, pp. 291–323, 2007.

P.H. Calamai & J.J. More. Projected gradient methods for linearly constrained problems. *Mathematical Programming*, 39(1), pp. 93–116, 1987.

M. Calvo, J.L. Montijano & L. Rández. On the solution of discontinuous ivp by adaptive Runge–Kutta codes. *Numerical Algorithms*, 33, pp. 163–182, 2003.

M.K. Camlibel. *Complementarity Methods in the Analysis of Piecewise Linear Dynamical Systems*. PhD thesis, Katholieke Universiteit Brabant, 2001. ISBN: 90 5668 073X.

M.K. Camlibel, W.P.M.H. Heemels & J.M. Schumacher. Consistency of a time-stepping method for a class of piecewise-linear networks. *IEEE Transactions on Circuits and Systems I*, 49, pp. 349–357, 2002a.

M.K. Camlibel, W.P.M.H. Heemels & J.M. Schumacher. On linear passive complementarity systems. *European Journal of Control*, 8, pp. 220–237, 2002b.

M.K. Camlibel, J.S. Pang & J. Shen. Conewise linear systems: non-zeroness and observability. *SIAM Journal of Control and Optimization*, 45(5), pp. 1769–1800, 2006.

L. Cangémi. *Frottement et adhèrence : modèle, traitement numérique et application à l'interface fibre/matrice*. PhD thesis, Université d'Aix-Marseille II, 1997.

L. Cangémi, M. Cocu & M. Raous. Adhesion and friction model for the fibre/matrix interface of a composite. *Third Biennal Joint Comference on Engineering System Design and Analysis*, pp. 157–163, Montpellier, 1996. A.S.M.E.

M.B. Carver. Efficient integration over discontinuities in ordinary differential equation simulations. *Mathematics and Computers in Simulation*, 20, pp. 190–196, 1978.

C. Castaing & M.D.P. Monteiro-Marques. Evolution problems associated with non-convex closed moving sets with bounded variation. *Portugaliae Mathematica*, 1, pp. 73–87, 1996.

C. Castaing, T.X. Duc Ha & M. Valadier. Evolution equations governed by the sweeping process. *Set-Valued Analysis*, 1, pp. 109–139, 1993.

F.E. Cellier & D.F. Rufer. Algorithm suited for the solution of initial value problems in engineering applications. M.H. Hamza, editor, *Proceedings of International Symposium and Course SIMULATION*, pp. 160–165, Zürich, 1975. Acta Press.

J.-L. Chaboche, F. Feyel & Y. Monerie. Interface debonding models: a viscous regularization with a limited rate dependency. *International Journal of Solids and Structures*, 38(18), pp. 3127–3160, 2001.

C. Chen & O.L. Mangasarian. A class of smoothing functions for nonlinear and mixed complementarity problems. *Computational Optimization and Applications*, 5, pp. 97–138, 1996.

B. Chen, X. Chen & C. Kanzow. A penalized Fischer–Burmeister NCP-function. *Mathematical Programming*, 88, pp. 211–216, 2000. In extended versiion as preprint 126 of the Institute of Mathematics, University of Hamburg, 1997.

P. Christensen, A. Klarbring, J. Pang & N. Stromberg. Formulation and comparison of algorithms for frictional contact problems. *International Journal for Numerical Methods in Engineering*, 42, pp. 145–172, 1998.

P.W. Christensen. *Computational Nonsmooth Mechanics. Contact, Friction and Plasticity*. Dissertations No. 657, Department of Mechanical Engineering, Linkoping University, 2000.

P.W. Christensen & J.S. Pang. Frictional contact algorithms based on semismooth newton methods. M. Fukushima & L. Qi, editors, *Reformulation – Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods*, pp. 81–116. Kluwer Academic, Dordrecht, 1998.

L.O. Chua, C.A. Desoer & E.S. Kuh. *Linear and Nonlinear Circuits*. McGraw-Hill, 1991.

E.A. Coddington & N. Levinson. *Theory of Ordinary Differential Equations*. McGraw-Hill, New York, 1955.

J. Cohen, M. Lin, D. Manocah & K. Ponamgi. I-collide–an interactive and exact collision detection system for large scale environments. *ACM Interactive 3D Graphics Conference, Monterey*, pp. 112–120, 1995.

M.G. Cojocaru. *Projected Dynamical Systems on Hilbert Spaces*. PhD thesis, Department of Mathematics and Statistics, Quenn's University, Kingston, Ontario, Canada, 2002.

M.G. Cojocaru & L.B. Jonker. Existence of solutions to projected differential equations on Hilbert spaces. *Proceedings of the AMS*, 132(1), pp. 183–193, 2003.

A.R. Conn, N.I.M. Gould & P.L. Toint. Testing a class of methods for solving minimization problems with simple bounds on the variables. *Mathematics of Computation*, 50, pp. 399–430, 1988.

A.R. Conn, N.I.M. Gould & P.L. Toint. *LANCELOT: a FORTRAN package for large-scale nonlinear optimization (Release A),* Number 17 in *Springer Series in Computational Mathematics*. Springer Verlag, New York, 1992.

C. Conti, P. Corron & P. Michotte. A computer aided kinematic analysis system for mechanism design and computer simulation. *Mechanisms and Machine Theory*, 27, pp. 563–574, 1992.

B. Cornet. Existence of slow solutions for a class of differential inclusions. *Journal of Mathematical Analysis and Application*, 96, pp. 130–147, 1983.

R.W. Cottle, J. Pang & R.E. Stone. *The Linear Complementarity Problem*. Academic Press, Boston, MA, 1992.

A. Curnier & P. Alart. A generalized Newton method for contact problems with friction. *Journal de Mécanique Théorique et Appliquée*, suppl. 1–7, pp. 67–82, 1988.

D. Daudon, J. Lanier & I. Vardoulakis. A micromechanical comparison between experimental results and numerical simulation of a bi-axial test on 2D granular material. R.P. Behringer & J.T. Jenkins, editors, *Powder and Grains 97*, pp. 219–222. Balkema, Rotterdam, 1997.

T. De Luca, F. Facchinei & C. Kanzow. A theoritical and numerical comparison of some semismooth algorithms for complementarity problems. *Computational Optimization and Applications*, 16, pp. 173–205, 2000.

G. De Saxcé. Une généralisation de l'inégalité de Fenchel et ses applications aux lois constitutives. *Comptes Rendus de l'Académie des Sciences*, t 314, série II, pp. 125–129, 1992.

G. De Saxcé. The bipotential method, a new variational and numerical treatment of the dissipative laws of materials. *10th International Conference on Mathematical and Computer Modelling and Scientific Computing*, Boston, USA, July 1995.

G. De Saxcé & Z.-Q. Feng. New inequality and functional for contact with friction: the implicit standard material approach. *Mechanics of Structures and Machines*, 19, pp. 301–325, 1991.

G. De Saxcé & Z.-Q. Feng. The bipotential method: a constructive approach to design the complete contact law with friction and improved numerical algorithms. *Mathemetical and Computer Modelling*, 28(4), pp. 225–245, 1998.

K. Deimling. *Multivalued Differential Equations*. Walter de Gruyter, 1992.

K. Dekker & J.G. Verwer. *Stability of Runge–Kutta Methods for Stiff Nonlinear Differential Equations,* vol. 2. CWI Monographs. North-Holland, Amsterdam, 1984.

T. DeLuca, F. Facchinei & C. Kanzow. A semismooth equation approach to the solution of nonlinear complementarity problems. *Mathematical Programming*, 75, pp. 407–439, 1996.

D. den Hertog. *Interior-Point Approach to Linear, Quadratic and Convex Programming*. Kluwer Academic, Boston, MA, 1994.

P. Denoyelle & V. Acary. The non-smooth approach applied to simulating integrated circuits and power electronics. Evolution of electronic circuit simulators towards fast-SPICE performance. *INRIA Research Report 0321, http://hal.inria.fr/docs/00/08/09/20/PDF/RT-0321.pdf*, 2006.

S.P. Dirkse & M.C. Ferris. The path solver: a non-monotone stabilization scheme for mixed complementarity problems. *Optimization Methods and Software*, 5, pp. 123–156, 1995.

A. Dontchev. Properties of one-sided Lipschitz multivalued maps. *Nonlinear Analysis TMA*, 49, pp. 13–20, 2002.

T. Dontchev & E. Farkhi. Stability and Euler approximation of one-sided Lipschitz differential inclusions. *SIAM Journal of Control and Optimization*, 36(2), pp. 780–796, 1998.

A.L. Dontchev & F. Lempio. Difference methods for differential inclusions: a survey. *SIAM Reviews*, 34(2), pp. 263–294, 1992.

A. Doris. *Output–Feedback Design for Non-smooth Mechanical Systems: Control Synthesis and Experiments*. PhD thesis, Technological University of Eindhoven, Mechanical Engineering Department, September 2007.

F. Dubois. Fracturation as a nonsmooth contact dynamic problem. E. Onate & D.R.J. Owen, editors, *VIII International Conference on Computational Plasticity, COMPLAS VIII*, Barcelona, 2005. CIMNE, Interantional Center for Numerical Methods in Engineering.

P. Dupont, V. Hayward, B. Armstrong & F. Altpeter. Single state elastoplastic friction models. *IEEE Transactions on Automatic Control*, 47(5), pp. 787–792, May 2002.

P. Dupuis & A. Nagurney. Dynamical systems and variational inequalities. *Annals of Operations Research,* 44(1–4), pp. 9–42, 1993.

R. Dzonou & M.D.P. Monteiro Marques. A sweeping process approach to inelastic contact problems with general inertia operators. *European Journal of Mechanics A/Solids*, 26(3), pp. 474–490, 2007.

R. Dzonou, M.D.P. Monteiro Marques & L. Paoli. Sweeping process for impact problems with a general inertia operators. C.A. Mota Soares et al., editors, *III European Conference on Computational Mechanics, Solids, Structures and Coupled Problems in Engineering*, Lisbon, Portugal, June 2006.

P. Eberhard. Collision detection and contact approaches for a hybrid system/finite element simulation. F. Pfeiffer & Ch. Glocker, editors, *Proceedings of IUTAM Symposium on Unilateral Multibody Contacts*, pp. 193–202. Kluwer, 1999.

J. Eckstein & M.C. Ferris. Operator-splitting methods for monotone affine variational inequalities, with a parallel application to optimal control. *INFORMS Journal on Computing*, 10(2), pp. 218–235, 1998.

J.F. Edmond & L. Thibault. Relaxation of an optimal control problem involving a perturbed sweeping process. *Mathematical Programming Series B*, 104, pp. 347–373, 2005.

J.F. Edmond & L. Thibault. BV solutions of nonconvex sweeping process differential inclusion with perturbation. *Journal of Differential Equations*, 226, pp. 135–179, 2006.

C.M. Elliot. On the convergence of a one-step method for the numerical solution of an ordinary differential inclusion. *IMA Journal of Numerical Analysis*, 5, pp. 3–21, 1985.

D. Ellison. Efficient automatic integration of ordinary differential equations with discontinuities. *Mathematics and Computers in Simulation*, 23, pp. 12–20, 1981.

H. Elmqvist, S.E. Mattsson & M. Otter. Object-oriented and hybrid modeling in modelica. *Journal Européen des systèmes automatisés*, 35(4), pp. 395–404, 2001.

W.H. Enright, K.R. Jackson, S.P. Norsett & P.G. Thomsen. Interpolants for Runge–Kutta formulas. *ACM Transactions on Mathematical Software*, 12(3), pp. 193–218, 1986.

W.H. Enright, K.R. Jackson, S.P. Norsett & P.G. Thomsen. Effective solution of discontinuous IVPs using a Runge–Kutta formula pair with interpolants. *Applied Mathematics and Computation*, 27, pp. 313–335, 1988.

M. Erdmann. On a representation of friction in configuration space. *International Journal of Robotics Research*, 13, pp. 240–271, 1994.

Y.G. Evtushenko & V.A. Purtov. Sufficient conditions for a minimum for nonlinear programming problems. *Soviet Mathematics Doklady*, 30, pp. 313–316, 1984.

F. Facchinei & J.S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*, volume I & II of *Springer Series in Operations Research.* Springer, New York, 2003.

F. Facchinei & J. Soares. A new merit function for nonlinear complementarity problems and a related algorithm. *SIAM Journal on Optimization*, 7, pp. 225–247, 1997.

Z-Q. Feng. *Contribution à la modélisation des problèmes non linéaires: contact, plasticité et endommagement*. PhD thesis, Université Technologique de Compiègne, 1991.

Z.-Q. Feng. 2D and 3D frictional contact algorithms and applications in a large deformation context. *Communications in Numerical Methods in Engineering*, 11, pp. 409–416, 1995.

M.C. Ferris & C. Kanzow. *Complementarity and Related Problems*, chapter Complementarity and Related Problems. Oxford University Press, 2002.

M. C. Ferris & T. S. Munson. Interfaces to path 3.0: design, implementation and usage. *Computational Optimizations and Applications*, 12(1-3), pp. 207–227, 1999.

R.C. Fetecau, J.E. Mardsen, M. Ortiz & M. West. Nonsmooth Lagrangian mechanics and variational collision integrators. *SIAM Journal of Applied Dynamical Systems*, 2(3), pp. 381–416, 2003.

A.V. Fiacco & G.P. McCormick. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. Wiley, New York, 1968. (Reprint by SIAM, Philadelphia, 1990).

G. Fichera. *Existence Theorems in Elasticity. Boundary Value Problems in Elasticity with Unilateral Constraints*, volume VIa/2 Mechanics of Solids II. Truesdell, C. ed., chapter in Encyclopedia of Physics. Flügge, S. ed., pp. 347–424. Springer Verlag, Berlin, 1972.

D. Filip, R. Magedson & R. Marhot. Surface algorithm using bound derivatives. *Computer Aided Geometric Design*, 3, pp. 255–311, 1986.

A.F. Filippov. Differential equations with discontinuous right-hand-side. *AMS Translations*, 42, pp. 199–231, 1964.

A.F. Filippov. Classical solutions of differential equations with multivalued right-hand-side. *SIAM Journal of Control and Optimization*, 5, pp. 609–621, 1967.

A.F. Filippov. *Differential Equations with Discontinuous Right Hand Sides*. Kluwer, Dordrecht, 1988.

A. Fischer. A special Newton-type optimization method. *Optimization*, 2, pp. 269–284, 1992.

R. Fletcher. A general quadratic programming algorithm. *Journal of the Institute and Its Applications*, 7, pp. 76–91, 1971.

R. Fletcher. *Practical Methods of Optimization*. Wiley, Chichester, 1987.

R. Fletcher & T. Johnson. On the stability of null-space methods fo KKT systems. Numerical Analysis Report NA/167, Department of Mathematics, University of Dundee, 1995.

M. Frémond. Equilibre des structures qui adhèrent à leur support. *Comptes Rendus de l'Académie des Sciences*, 295, Série II, pp. 913–916, 1982.

M. Frémond. Adhérence des solides. *Journal de Mécanique Théorique et Appliquée*, 6(3), pp. 383–407, 1987.

M. Frémond. Contact with adhesion. In Moreau & Panagiotopoulos (1988), pp. 177–222.

M. Frémond. *Non-smooth Thermo-mechanics*. Springer, Berlin-Heidelberg, 2002.

R.M. Freund & S. Mizuno. Interior point methods: current status and future directions. *Optima*, 51, pp. 1–9, 1996.

R.W. Freund & N.M. Nachtigal. QMR: a quasi-minimal residual method for non-Hermitian linear systems. *Numerische Mathematik*, 60(1), pp. 315–339, 1991.

M.P. Friedlander & S. Leyffer. Gradient projection for general quadratic programs. Technical Report ANl/MCS-P1370-0906, Argonne National Laboratory, IL, 2006.

M. Fukushima. Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems. *Mathematical Programming*, 53, pp. 99–110, 1992.

S.A. Gabriel & J.S. Pang. An inexact NE/SQPmethod for solving the nonlinear complementarity problem. *Computational Optimization and Applications*, 1, pp. 67–91, 1992.

Z. Galias & X. Yu. Euler's discretization of single input sliding-mode control systems. *IEEE Transactions on Automatic Control*, 52(9), pp. 1726–1730, September 2007.

J. Garcia de Jalon & E. Bayo. *Kinematic and Dynamic Simulation of Multibody Systrems: The Real-Time Challenge*. *Mechanical Engineering Series*. Springer, 1994.

C.W. Gear. The automatic integration of ordinary differential equations. *Communications of the ACM*, 14(3), pp. 176–190, 1970.

C.W. Gear & O. Østerby. Solving ordinary differential equations with discontinuities. *ACM Transactions on Mathematical Software*, 10(1), pp. 23–44, 1984.

F. Génot & B. Brogliato. New results on Painlevé paradoxes. *INRIA Research Report 3366, http://hal.inria.fr/docs/00/07/33/23/PDF/RR-3366.pdf*, 1998.

F. Génot & B. Brogliato. New results on Painlevé paradoxes. *European Journal of Mechanics A/Solids*, 18, pp. 653–677, 1999.

P. Germain, Q.S. Nguyen & P. Suquet. Continuum thermodynamics. *Journal of Applied Mechanics, Transactions of ASME*, 50, 50th Anniversary Issue, pp. 1010–1020, 1983.

P.E. Gill & W. Murray. *Numerical Methods for Constrained Optimzation*. Academic Press, New York, 1975.

P.E. Gill, W. Murray & M.H. Wright. *Practical Optimization*. Academic Press, New York, 1981.

P.E. Gill, W. Murray, M.A. Saunders & M.H. Wright. Inertia-controlling methods for general quadratic programming. *SIAM Review*, 33(1), pp. 1–36, 1991.

C. Glocker. Dynamik von Starrkörpersystemen mit Reibung und Stöss en Dissertation. *Technischen Universität München*, 1995.

C. Glocker. Formulation of spatial contact situations in rigid multibody systems. *Computer Methods in Applied Mechanics and Engineering*, 177, pp. 199–214, 1999.

C. Glocker. *Set-Valued Force Laws: Dynamics of Non-smooth Systems*, volume 1 of *Lecture Notes in Applied Mechanics*. Springer Verlag, 2001.

C. Glocker. Concepts for modeling impacts without friction. *Acta Mechanica*, 168 (1–2), pp. 1–19, 2004.

C. Glocker & C. Studer. Formulation and preparation for numerical evaluation of linear complementarity systems in dynamics. *Multibody Systems Dynamics*, 13, pp. 447–463, 2005.

R. Glowinski, J.L. Lions & R. Trémolières. *Approximations des Inéquations Variationnelles*. Dunod, Paris, 1976.

D. Goeleven & B. Brogliato. Stability and instability matrices for linear evolution variational inequalities. *IEEE Transactions on Automatic Control*, 49(4), pp. 521–534, 2004.

D. Goeleven & B. Brogliato. Necessary conditions of asymptotic stability for unilateral dynamical systems. *Nonlinear Analysis: Theory, Methods and Applications*, 61, pp. 961–1004, 2005.

D. Goeleven, D. Motreanu, Y. Dumont & M. Rochdi. *Variational and Hemivariational Inequalities: Theory, Methods and Applications; Volume I: Unilateral Analysis and Unilateral Mechanics*. Nonconvex Optimization and its Applications. Kluwer Academic Publishers, 2003a.

D. Goeleven, D. Motreanu & V.V. Motreanu. On the stability of stationary solutions of parabolic variational inequalities. *Advances in Nonlinear Variational Inequalities*, 6, pp. 1–30, 2003b.

A.A. Goldstein. Convex programming in Hilbert space. *Bulletin of American Mathematical Society*, 70(5), pp. 709–710, 1964.

E.G. Golshtein & N.V. Tretyakov. *Modified Lagrangians and Monotone Maps in Optimization*. Wiley, New York, 1996.

O. Gonzalez. Exact enrgy and momentum conserving algorithms for general models in nonlinear elasticity. *Computer Methods in Applied Mechanics and Engineering*, 190, pp. 1762–1783, 2000.

N.I.M. Gould & P.L. Toint. Numerical methods for large-scale non-convex quadratic programming. A.H. Siddiqi & M. Kocvara, editors, *Trends in Industrial and Applied Mathematics*. Kluwer Academic, Dordrecht, 2002.

M.S. Gowda. On the extended linear complementarity problem. *Mathematical Programming*, 72(1), pp. 33–50, 1996.

M.S. Gowda & R. Sznajder. The generalized order linear complementarity problem. *SIAM Journal on Matrix Analysis and Applications*, 15, pp. 779–795, 1994.

S. Goyal, E.N. Pinson & F.W. Sinden. Simulation of dynamics of interacting rigid bodies including friction. I: general problem and contact model. II: software system design and implementation. *Engineering with Computers*, 10, pp. 162–195, 1994.

F. Grognard, H. de Jong & J.L. Gouzé. *Bio and Control Theory: Current Challenges (Queinnec et al, Eds)*, chapter Piecewise-linear models of genetic regulatory networks: theory and example, pp. 137–159. Number 357 in *Lecture Notes in Control and Information Sciences*. Springer, Berlin Heidelberg, 2007.

L. Grüne & P.A. Kloeden. Higher order numerical approximation of switching systems. *Systems and Control Letters*, 55, pp. 746–754, 2006.

O. Güler. Generalized linear complementarity problems. *Mathematics of Operations Research*, 20, pp. 441–448, 1995.

W. Hackbusch & H.D. Mittelmann. On multi-grid methods for variational inequalities. *Numerische Mathematik*, 42, pp. 65–76, 1983.

W.M. Haddad, V. Chellaboina & S.G. Nersesov. *Impulsive and Hybrid Dynamical Systems. Stability, Dissipativity and Control.* Princeton Series in Applied Mathematics, 2006.

E. Hairer & G. Wanner. *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems.* Springer, 1996.

E. Hairer, S.P. Norsett & G. Wanner. *Solving Ordinary Differential Equations I. Nonstiff Problems.* Springer, 1993.

I. Han, B.J. Gilmore & M.M. Ogot. The incorporation of arc boundaries and stick/slip friction in a rule-based simulation algorithm for dynamic mechanical systems with changing topologies. *ASME Journal of Mechanical Design*, 115, pp. 423–434, 1993.

P.T. Harker & J.-S. Pang. Finite-dimensional variational inequality and complementarity problems: a survey of theory, algorithms and applications. *Mathematical Programming*, 48, pp. 160–220, 1990.

W.P.M.H. Heemels. *Linear Complementarity Systems. A Study in Hybrid Dynamics.* PhD thesis, Technical University of Eindhoven, 1999. ISBN 90-386-1690-2.

W.P.M.H. Heemels & B. Brogliato. The complementarity class of hybrid dynamical systems. *European Journal of Control*, 9, pp. 311–349, 2003.

W.P.M.H. Heemels, J.M. Schumacher & S. Weiland. Linear complementarity systems. *SIAM Journal on Applied Mathematics*, 60, pp. 1234–1269, 2000.

W.P.M.H. Heemels, M.K. Camlibel & J.M. Schumacher. On the dynamical analysis of piecewise-linear networks. *IEEE Transactions on Circuits and Systems I*, 49, pp. 315–327, 2002.

C. Henry. An existence theorem for a class of differential equations with multivalued right-hand-side. *Journal of Mathematical Analysis and Applications*, 41, pp. 179–186, 1973.

J.P. Hespanha. Uniform stability of switched linear systems: extensions of LaSalle's invariance principle. *IEEE Transactions on Automatic Control*, 49(4), pp. 470–482, 2004.

M.R. Hestenes. Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, 4, pp. 303–320, 1969.

N.J. Higham. Stability of the diagonal pivoting method with partial pivoting. *SIAM Journal of Matrix Analysis and Applications*, 18(1), pp. 52–65, 1997.

A.C. Hindmarsh. GEAR: ordinary differential equations solver. Technical Report UCID-30001, Rev 3, Lawrence Livermore Laboratory, December 1974.

J.B. Hiriart-Urruty & C. Lemaréchal. *Convex Analysis and Minimization Algorithms*, volumes I and II. Springer, Berlin, 1993.

J.B. Hiriart-Urruty & C. Lemaréchal. *Fundamentals of Convex Analysis.* Springer, 2001.

F. Horkay, M. Jean & F. Mehrez. Unilateral contact nd dry friction in numerical simulation of deep drawing. *Proceedings of the Numiform 89*, Fort Collins, Colorado, June 26–30 1989.

P.M. Hubbard. Approximating polyhedra with spheres fot time-critical collision detection. *ACM Transactions on Graphics*, 15(3), pp. 179–210, 1996.

S.Y.R. Hui & J. Zhu. Numerical modelling and simulation of hysteresis effects in magnetic cores using transmission-line modelling and the Preisach theory. *IEE Proceedings of Electric Power Applications*, 142(1), pp. 57–62, 1995.

Y. Hurmuzlu. An energy-based coefficient of restitution for planar impacts of slender bars with massive external surfaces. *Journal of Applied Mechanics, Transactions of ASME*, 65, pp. 952–962, 1998.

T. Illés, M. Nagy & T. Terlaky. Polynomial interior point methods for general LCPs. Technical Report ORR 2007-3, Eötvös Loránd University of Sciences, Department of Operations Research, Budapest, Hungary, 2007. Available at `http://www.optimization-online.org/DB_HTML/2007/04/1635.html`.

J. Imura. Well-posedness analysis of switch-driven piecewise affine systems. *IEEE Transactions on Automatic Control*, 48(11), pp. 1926–1935, 2003.

J. Imura & A.J. van der Schaft. Characterization of well-posedness of piecewise-linear systems. *IEEE Transactions on Automatic Control*, 45(9), pp. 1600–1619, 2000.

A. Isidori. *Nonlinear Control Systems, 3rd Edition*. Springer London, 1995.

O. Janin & C.H. Lamarque. Comparison of several numerical methods for mechanical systems with impacts. *International Journal of Numerical Methods in Engineering*, 51(9), pp. 1101–1132, 2001.

M. Jean. Unilateral contact and dry friction: time and space variables discretization. *Archives of Mechanics, Warszawa*, 40(1), pp. 677–691, 1988.

M. Jean. The non smooth contact dynamics method. *Computer Methods in Applied Mechanics and Engineering*, 177, pp. 235–257, 1999. Special issue on computational modeling of contact and friction, J.A.C. Martins and A. Klarbring, editors.

M. Jean & J.J. Moreau. Dynamics of elastic or rigid bodies with frictional contact and numerical methods. R. Blanc, P. Suquet & M. Raous, editor, *Publications du LMA*, pp. 9–29, 1991.

M. Jean & J.J. Moreau. Unilaterality and dry friction in the dynamics of rigid bodies collections. A. Curnier, editor, *Proceedings of Contact Mechanics International Symposium*, volume 1, pp. 31–48. Presses Polytechniques et Universitaires Romandes, 1992.

M. Jean & E. Pratt. A system of rigid bodies with dry friction. *International Journal of Engineering Science*, 23(23), pp. 497–513, 1985.

M. Jean & G. Touzot. Implementation of unilateral contact and dry friction in computer codes dealing with large deformations problems. *Journal de Mechanique Theorique et Appliquee*, 7(1), pp. 145–160, 1988.

M. Jean, V. Acary & Y. Monerie. Non-smooth contact dynamics approach of cohesive materials. *Philosophical Transactions of the Royal Society, Mathematical, Physical and Engineering Sciences*, 359(1789), pp. 2497–2518, 15 December 2001. Non-smooth Mechanics, A Theme Issue compiled and edited by F.G. Pfeiffer.

M. Johansson & A. Rantzer. Computation of piecewise quadratic Lyapunov functions for hybrid systems. *IEEE Transactions on Automatic Control*, 43(4), pp. 555–559, 1998.

N.H. Josephy. Newton's method for generalized equations. Technical report, Mathematics Research Center, University of Wisconsin, Madison, 1979.

F. Jourdan, P. Alart & M. Jean. A Gauss–Seidel like algorithm to solve frictional contact problems. *Computer Methods in Applied Mechanics and Engineering*, 155, pp. 31–47, 1998a.

F. Jourdan, M. Jean & P. Alart. An alternative method between implicit and explicit schemes devoted to frictional contact problems in deep drawing simulation. *Journal of Materials Processing Technology*, 80–81, pp. 257–262, 1998b.

C. Kanzow & H. Kleinmichel. A class of Newton-type methods for equality and inequality constrained equations. *Optimization Methods and Software*, 5, pp. 173–198, 1995.

N. K. Karmarkar. A new polynomial-time algorithm for linear programming. *Proceedings of the 16th Annual ACM Symposium on Theory of Computing*, pp. 302–311, 1984.

D. Karnopp. Computer simulation of stick–slip friction in mechanical dynamic systems. *ASME Journal of Dynamic Systems, Measurement and Control*, 107, pp. 100–103, 1985.

A. Kastner-Maresch. Implicit Runge–Kutta methods for differential inclusions. *Numerical Functional Analysis and Optimization*, 11(9–10), pp. 937–958, 1990–91.

A.E. Kastner-Maresch. The implicit midpoint rule applied to discontinuous differential equations. *Computing*, 49, pp. 45–62, 1992.

T. Kato. Accretive operators and nonlinear evolution equations in Banach spaces. *Nonlinear Functional Analysis, Proceedings of Symposia in Pure Mathematics 18, Part 1, Chicago 1968*, pp. 138–161, 1970.

D.M. Kaufman, T. Edmunds & D.K. Pai. Fast frictional dynamics for rigid bodies. *International Conference on Computer Graphics and Interactive Techniques, Proceedings of ACM SIGGRAPH 2005*, 24, pp. 946–956. AM Press, July 2005.

E.N. Khobotov. A modification of the extragradient method for the solution of variational inequalities and some optimization problems. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fisiki*, 27, pp. 1462–1473, 1987.

R. Kikuuwe, N. Takesue, A. Sano, H. Mochiyama & H. Fujimoto. Fixed-step friction simulation: from classical Coulomb model to modern continuous models. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005)*, pp. 1009–1016, Edmonton, Alberta, Canada, August 2005.

D. Kinderlehrer & G. Stampacchia. *An Introduction to Variational Inequalities.* Academic Press, New York, 1980.

D. Kinzebulatov. Systems with distributions and viability theorem. *Journal of Mathematical Analysis and Applications*, 331, pp. 1046–1067, 2007.

A. Klarbring. A mathematical programming approach to three-dimensional contact problem with friction. *Comptuter Methods in Applied Mechanics and Engineering*, 58, pp. 175–200, 1986a.

A. Klarbring. A mathematical programming approach to three-dimensional contact problems with friction. *Computer Methods in Applied Mechanics and Engineering*, 58, pp. 175–200, 1986b.

A. Klarbring & G. Björkman. A mathematical programming approach to contact problems with friction and varying contact surface. *Computers & Structures*, 30(5), pp. 1185–1198, 1988.

M. Kocvara & J. Zowe. An iterative two-step method for linear complementarity problems. *Numerische Mathematik*, 68, pp. 95–106, 1994.

M. Kojima, N. Megiddo, T. Noma & A. Yoshise. A unified approach to interior point algorithms for linear complementarity problems : a summary. *Operations Research Letters*, 10, pp. 247–254, 1991.

I.V. Konnov. Combined relaxation methods for finding equilibrium point and solving related problems. *Russian Mathematics*, 37, pp. 46–53, 1993.

G.M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12(747–756), 1976.

M.M. Kostreva & X.Q. Yang. Unified approaches for solvable and unsolvable linear complementarity problems. *European Journal of Operational Research*, 158, pp. 409–416, 2004.

V.V. Kozlov & D.V. Treshchev. *Billiards. A Genetic Introduction to the Dynamics of Systems with Collisions.* Izdatel'stvo Moskovskogo Universiteta, Moskva, 1991.

M. Kunze & M.D.P. Monteiro Marques. On parabolic quasi-variational inequalities and state-dependent sweeping processes. *Topological Methods in Nonlinear Analysis*, 12, pp. 179–191, 1998.

M. Kunze & M.D.P. Monteiro Marquès. An introduction to Moreau's sweeping process. B. Brogliato, editor, *Impact in Mechanical systems: Analysis and Modelling*, volume 551 of *Lecture Notes in Physics*, pp. 1–60. Springer, 2000.

C.H. Lamarque, J. Bastien & M. Holland. Study of a maximal monotone model with a delay term. *SIAM Journal of Numerical Analysis*, 41(4), pp. 1286–1300, 2003.

T. Larsson & M. Patriksson. A class of gap functions for variational inequalities. *Mathematical Programming*, 64(1), pp. 53–79, 1994.

T.A. Laursen & X.N. Meng. A new solution procedure for application of energy-conserving algorithms to general constitutive models in nonlinear elastodynamics. *Computer Methods in Applied Mechanics and Engineering*, 190, pp. 6309–6322, 2001.

T.A. Laursen & J.C. Simo. A continuum-based finite element formulation for the implicit solution of multibody, large deformation frictional contact problems. *International Journal for Numerical Methods in Engineering*, 36(20), pp. 3451–3485, 1993a.

T.A. Laursen & J.C. Simo. Algorithmic symmetrization of Coulomb frictional problems using augmented Lagrangians. *Computer Methods in Applied Mechanics and Engineering*, 108(1–2), pp. 133–146, 1993b.

C. Le Saux, R.I. Leine & C. Glocker. Dynamics of a rolling disk in the presence of dry friction. *Journal of Nonlinear Science*, 15, pp. 27–61, 2005.

F. Lebon & M. Raous. Multibody contact problem including friction in structure assembly. *Computers and Structures*, 43(5), pp. 925–934, 1992.

D.M.W. Leenaerts & W.M. Van Bokhoven. *Piecewise Linear Modeling and Analysis*. Kluwer Academic, Norwell, MA, 1998.

R.I. Leine & C. Glocker. A set-valued force law for spatial Coulomb-Contensou friction. *European Journal of Mechanics A/Solids*, 22, pp. 193–216, 2003.

R. Leine & H. Nijmeijer. *Dynamics and Bifurcations of Non-smooth Mechanical Systems*. Springer, *Lecture Notes in Applied and Computational Mechanics* 18, 2004.

R. Leine & N. van de Wouw. Stability properties of equilibrium sets of non-linear mechanical systems with dry friction and impact. *Nonlinear Dynamics*, 2007, DOI: 10.1007/s11071-007-9244-z.

R.I. Leine, D.H. van Campen, A. de Kraker & L. van den Steen. Stick–slip vibrations induced by alternate friction models. *Nonlinear Dynamics*, 16(1), pp. 41–54, May 1998.

R.I. Leine, B. Brogliato & H. Nijmeijer. Periodic motion and bifurcations induced by the Painlevé paradox. *European Journal of Mechanics A/Solids*, 21, pp. 869–896, 2002.

R. Leine, Ch. Glocker & D.H. van Campen. Nonlinear dynamics and modeling of various wooden toys with impact and friction. *Journal of Vibration and Control*, 9, pp. 25–78, 2003.

C.E. Lemke. Bimatrix equilibrium points and mathematical programming. *Management Science*, 11, pp. 681–689, 1965.

C.E. Lemke & J.T. Howson. Equilibrium points of bimatrix games. *SIAM Journal on Applied Mathematics*, 12, pp. 413–423, 1964.

F. Lempio. Difference methods for differential inclusions. *Lecture Notes in Economics and Mathematical Systems*, 378, pp. 236–273, 1992.

P. Lerat, A. Zervos & M. Jean. Déplacement et rotation des grains au sein d'une interface structure–milieu granulaire: résultats expérimlentaux et numériques. *GEO, 2ème réunion annuelle*, 1995.

A.Y.T. Leung, C. Guoqing & C. Wanji. Smoothing Newton method for solving two-and three-dimensional frictional contact problems. *International Journal for Numerical Methods in Engineering*, 41, pp. 1001–1027, 1998.

R.J. LeVeque. *Numerical Methods for Conservation Laws*. Birkhauser Verlag, Basel, 1990.

E.S. Levitin & B.T. Polyak. Constrained minimization problems. *USSR Computational Mathematics and Mathematical Physics*, 6, pp. 1–50, 1966.

D. Li. Morse decomposition for general dynamical systems and differential inclusions with applications to control systems. *SIAM Journal of Control and Optimization*, 46(1), pp. 35–60, 2007.

M.C. Lin & J.F. Canny. A fast algorithm for incremental distance calculation. *Proceedings of the 7th IEEE Conference on Robotics and Automation, Sacramento*, pp. 1008–1014, 1991.

J.L. Lions & G. Stampacchia. Variational inequalities. *Communications on Pure and Applied Mathematics*, XX, pp. 493–519, 1967.

G. Lippold. Error estimates for the implicit euler approximation of an evolution inequality. *Nonlinear Analysis: Theory, Methods and Applications*, 15(11), pp. 1077–1089, 1990.

C. Liu, Z. Zhao & B. Chen. The bouncing motion appearing in a robotic system with unilateral constraint. *Nonlinear Dynamics*, 49, pp. 217–232, 2007.

P. Lötstedt. Coulomb friction in two-dimensional rigid body systems. *Zeitschrift für Angewandte Mathematik und Mechanik*, 61(12), pp. 605–615, 1981.

P. Lötstedt. Mechanical systems of rigid bodies subject to unilateral constraints. *SIAM Journal of Applied Mathematics*, 42(2), pp. 281–296, 1982.

P. Lötstedt. Numerical simulation of time-dependent contact and friction problems in rigid body mechanics. *SIAM Journal of Scientific and Statistical Computing*, 5, pp. 370–393, 1984.

D. Luenberger. A double look at duality. *IEEE Transactions on Automatic Control*, 37(10), pp. 1474–1482, October 1992.

M. Mabrouk. A unified variational for the dynamics of perfect unilateral constraints. *European Journal of Mechanics– A/Solids*, 17, pp. 819–842, 1998.

J.L. Mancilla-Aguilar, R. Garcia, E. Sontag & Y. Wang. On the representation of switched systems with inputs by perturbed control systems. *Nonlinear Analysis TMA*, 60, pp. 1111–1150, 2005.

J. Mandel. A multilevel iterative method for symmetric, positive definite linear complementarity problems. *Applied Mathematics and Optimization*, 11(1), pp. 77–95, 1984.

O. Mangasarian. *Nonlinear Programming*. McGraw-Hill, 1969.

O.L. Mangasarian. Equivalence of the complementarity problem to a system of nonlinear equations. *SIAM Journal of Applied Mathematics*, 31, pp. 89–92, 1976.

O.L. Mangasarian & J.S. Pang. The extended linear complementarity problem. *SIAM Journal on Matrix Analysis and Applications*, 16, pp. 359–368, 1995.

O.L. Mangasarian & M.V. Solodov. Nonlinear complementarity as unconstrained and constrained minimization. *Mathematical Programming*, 62, pp. 277–298, 1993.

R. Mannshardt. One-step methods of any order for ordinary differential equations with discontinuous right-hand sides. *Numerische Mathematik*, 31, pp. 131–152, 1978.

P. Marcotte. A new algorithm for solving variational inequalities, with application to the traffic assignment problem. *Mathematical Programing*, 33(339–351), 1985.

P. Marcotte. Application of Khobotov's algorithm to variational inequalities and network equilibirum problems. *Information Systems and Operational Research*, 29(258–270), 1991.

P. Marcotte & J. Dussault. A note on a globally convergent Newton method for solving monotone variational inequalities, 1987.

P. Marcotte & J.H. Wu. On the convergence of projection methods: application to the decomposition of affine variational inequalities. *Journal of Optimization Theory and Applications*, 85, pp. 347–362, 1995.

B. Martinet. Régularisations d'inéquations variationelles par approximations succesives. *Revue Francaise d'Informatique et de Recherche Opérationnelle*, 1970.

S.E. Mattsson, M. Otter & H. Elmqvist. Modelica hybrid modeling and efficient simulation. *38th IEEE Conference on Decision and Control*, pp. 3502–3507, December 1999.

J.M. McCarthy. *Introduction to Theoretical Kinematics*. The MIT Press, 1990.

N.H. McClamroch & D. Wang. Feedback stabilization and tracking of constrained robots. *IEEE Transactions on Automatic Control*, 33(5), pp. 419–426, May 1988.

F. Mehrez. *Unilateral Contact and Friction Modelling for Numerical Simulation of Deep-Drawing*. PhD thesis, Université de Paris VI, 1991. (In French).

S. Mehrotra. On the implementation of a primal–dual interior point method. *SIAM Journal on Optimization*, 2(4), pp. 575–601, 1992.

B.V. Mirtich. *Impulse-Based Dynamics Simulation of Rigid Body Systems*. PhD thesis, University of Califormia at Berkeley, USA, 1997.

E.N. Mitsopoulou & I.N. Doudoumis. A contribution to the analysis of unilateral contact problems with friction. *Solid Mechanics Archives*, 12(3), pp. 165–186, 1987.

E.N. Mitsopoulou & I.N. Doudoumis. On the solution of the unilateral contact frictional problem for general static loading conditions. *Computers & Structures*, 30(5), pp. 1111–1126, 1988.

H.D. Mittelmann. On the approximate solution of nonlinear variational inequalities. *Numerische Mathematik*, 29, pp. 451–462, 1978.

H.D. Mittelmann. On the efficient solution of nonlinear finite element equations I. *Numerische Mathematik*, 35, pp. 277–291, 1980.

H.D. Mittelmann. On the efficient solution of nonlinear finite element equations II. Bound-constrained problems. *Numerische Mathematik*, 36, pp. 375–387, 1981a.

H.D. Mittelmann. On the numerical solution of contact problems. K. Glashoff, E. L. Allgower & H. O. Peitgen, editors, *Numerical Solution of Nonlinear Equations*, volume 878 of *Springer Lecture Notes in Mathematics*, pp. 259–274. Springer, 1981b.

S. Mizuno, M. Todd & Y. Ye. On adaptive step primal–dual interior-point algorithms for linear programming. *Mathematics of Operations Research*, 18, pp. 964–981, 1993.

C. Moler. Are we there yet? Zero crossing and event handling for differential equations. *Simulink 2 Special Edition. Matlab News and Notes*, pp. 16–17, 1997.

M. Möller & C. Glocker. Non-smooth modelling of electrical systems using the flux approach. *Nonlinear Dynamics*, 50, pp. 273–295, 2007.

Y. Monerie. *Fissuration des matériaux composites: rôle de l'interface fibre/matrice*. PhD thesis, Université d'Aix-Marseille II, Octobre 2000.

Y. Monerie & V. Acary. Formulation dynamique d'un modèle de zone cohésive tridimensionnel couplant endommagement et frottement. *Revue européenne des éléments finis*, 10(02–03–04), pp. 489–503, 2001.

Y. Monerie & M. Raous. A model coupling adhesion to friction for the interaction between a crack and a fiber/matrix interface. *Zeitschrift für Angewandte Mathematik und Mechanik*, 1999. Special issues "Annual Gesellschaft für Angewandte Mathematik und Mechanik Conference", April 12–16, Metz.

M.D.P. Monteiro Marques. Chocs in élastiques standards: un résultat d'existence. *Séminaire d'Analyse Convexe, exposé no 4*, volume 15, USTL, Montpellier, France, 1985.

M.D.P. Monteiro Marques. *Differential Inclusions in Nonsmooth Mechanical Problems. Shocks and Dry Friction*. Progress in Nonlinear Differential Equations and their Applications, vol. 9. Birkhauser, Basel, 1993.

I.-C Morărescu & B. Brogliato. Tracking control of nonsmooth lagrangian systems with time constraints. *6th Euromech Conference ENOC*, Saint Petersburg, Russia, June, 30-July, 4, 2008.

J.J. Moré. Classes of functions and feasibility conditions in nonlinear complementarity problems. *Mathematical Programming*, 6, pp. 327–338, 1974.

J.J. Moré & W.C. Rheinbolt. On the p- and s-functions and related classes of n-dimensional nonlinear mappings. *Linear Algebra and Its Applications*, 6, pp. 45–68, 1973.

J.J. Moré & G. Toraldo. On the solution of large quadratic convex programming problems with bound constraints. *SIAM Journal of Optimization*, 1(1), pp. 93–113, 1991.

J.J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93, pp. 273–299, 1965.

J.J. Moreau. Rafle par un convexe variable (première partie), exposé no 15. *Séminaire d'analyse convexe, University of Montpellier*, 43 pages, 1971.

J.J. Moreau. Rafle par un convexe variable (deuxième partie) exposé no 3. *Séminaire d'analyse convexe, University of Montpellier*, 36 pages, 1972.

J.J. Moreau. On unilateral constraints, friction and plasticity. G. Capriz & G. Stampacchia, editors, *New Variational Techniques in Mathematical Physics, CIME II ciclo 1973*, pp. 175–322. Edizioni Cremonese, 1974.

J.J. Moreau. Evolution problem associated with a moving convex set in a Hilbert space. *Journal of Differential Equations*, 26, pp. 347–374, 1977.

J.J. Moreau. Liaisons unilatérales sans frottement et chocs inélastiques. *Comptes Rendus de l'Académie des Sciences*, 296 serie II, pp. 1473–1476, 1983.

J.J. Moreau. Dynamique des systèmes à liaisons unilatérales avec frottement sec éventuel. Technical Report 85-1, Laboratoire de Mécanique et de Génie civil, Université des sciences et techniques du Languedoc, Montpellier, Mai 1985a.

J.J. Moreau. Standard inelastic shocks and the dynamics of unilateral constraints. G. Del Piero & F. Maceri, editors, *Unilateral Problems in Structural Analysis*, number 288 in *CISM Course and Lectures*, pp. 173–221. Springer, 1985b.

J.J. Moreau. Bounded variation in time. J.J Moreau, P.D. Panagiotopoulos & G. Strang, editors, *Topics in Nonsmooth Mechanics*, pp. 1–74. Bikhäuser, Basel, 1988a.

J.J. Moreau. Unilateral contact and dry friction in finite freedom dynamics. In Moreau & Panagiotopoulos, editors, pp. 1–82, 1988b.

J.J. Moreau. Numerical experiments in granular dynamics: vibration-induced size segregation. M. Raous, M. Jean & J.J. Moreau, editors, *2nd Contact mechanics International Symposium*, Carry-Le-Rouet, France, September 1994a. Plenum Press, New York.

J.J. Moreau. Some numerical methods in multibody dynamics: application to granular materials. *European Journal of Mechanics – A/Solids*, suppl. 4, pp. 93–114, 1994b.

J.J. Moreau. Some basics of unilateral dynamics. F. Pfeiffer, editor, *IUTAM Symposium on Multibody Dynamics*. Kluwer, August 3-7 1998.

J.J. Moreau. Numerical aspects of the sweeping process. *Computer Methods in Applied Mechanics and Engineering*, 177, pp. 329–349, 1999. Special issue on computational modeling of contact and friction, J.A.C. Martins and A. Klarbring, editors.

J.J. Moreau. Facing the plurality of solutions in nonsmooth mechanics. *Proceedings of the Nonsmooth/Nonconvex Mechanics with Applications in Engineering, Thessaloniki*, pp. 3–12, 2006.

J.J. Moreau & P.D. Panagiotopoulos, editors. *Nonsmooth Mechanics and Applications*. Number 302 in *CISM, Courses and Lectures*. Springer, Wien- New York, 1988.

R. Motro. *Tensegrity. Strutural Systems for the Future*. Butterworth-Heinemann, 2006.

T.S. Munson. *Algorithms and Environment for Complementarity*. PhD thesis, University of Wisconsin, Madison, 2000.

K.G. Murty. Note on bard-type scheme for solving the complementarity problem. *Opsearch*, 11, pp. 123–130, 1974.

K.G. Murty. *Linear Complementarity, Linear and Nonlinear Programming*. Heldermann, 1988.

K.G. Murty & S.N. Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 19, pp. 200–212, 1987.

B. Muth, G. Of, P. Eberhard & O. Steinbach. Collision detection for complicated polyhedra using the fast multipole method or ray crossing. *Archives of Applied Mechanics*, 77, pp. 503–521, 2007.

A. Nagurney. *Network Economics. A Variational Inequality Approach*. Kluwer Academic, Dordrecht, 1993.

A. Nagurney & D. Zhang. *Projected Dynamical Systems and Variational Inequalities with Applications*. Kluwer Academic, 1996.

Y.E. Nesterov & A.S. Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM Publication, Philadelphia, 1993.

J.W. Nieuwenhuis. Some remarks on set-valued dynamical systems. *Journal of Australian Mathematical Society*, 252, pp. 308–313, 1981.

S. Nineb, P. Alart & D. Dureissex. Multiscale nonsmooth analysis of tensegrity systems. E. Ramm, W.A. Wall, K.U. Bletzinger & M. Bischoff, editors, *5th International Conference on Computation of Shell and Spatial Structures*, Salzburg, Austria, June 2005.

S. Nineb, P. Alart & D. Dureisseix. Approche multi-échelle des systèmes de tenségrité. *Revue Européenne de Mécanique Numérique*, 15(1–3), pp. 319–328, 2006.

J. Nocedal & S.J. Wright. *Numerical Optimization*. Springer, New York, 1999.

A. Nordsieck. On the numerical integration of ordinary differential equations. *Mathcomp*, 16, pp. 22–49, 1962.

C. Nouguier, C. Bohatier, J.J. Moreau & F. Radjai. Force fluctuations in a pushed granular material. *Granular Matter*, 2(4), pp. 171–178, 2000.

F.Z. Nqi. *Etude Numérique de divers Problèmes Dynamiques avec Impact et leur Propriétés Qualitatives*. PhD thesis, Université C. Bernard Lyon I, France, Laboratoire d'Analyse Numérique, June 1997.

C.W. Oosterlee. On multigrid for linear complementarity problems with application to American-style options. *Electronic Transactions on Numerical Analysis*, 15, pp. 165–185, 2003.

Y. Orlov. Vibrocorrect differential equations with measure. *Mathematical Notes*, 38(1), pp. 110–119, 1985.

Y. Orlov. Instantaneous impulse response of nonlinear system. *IEEE Transactions on Automatic Control*, 45(5), pp. 999–1001, 2000.

Y. Orlov. Finite time stability and robust control synthesis of uncertain switched systems. *SIAM Journal of Control and Optimization*, 43(4), pp. 1253–1271, 2005.

M. Ortiz & L. Stainier. The variational formulation of viscoplastic constitutive updates. *Computational Methods in Applied Mechanics and Engineering*, 171, pp. 419–444, 1999.

B.E. Paden & S.S. Sastry. A calculus for computing Filippov's differential inclusion with application to the variable structure control of robot manipulators. *IEEE Transactions on Circuits and Systems*, 34(1), pp. 73–82, 1987.

C.C. Paige & M.A. Saunders. LSQR: an algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software*, 8, pp. 43–71, 1982.

J.S. Pang. Newton's method for B-differentiable equations. *Mathematics of Operations Research*, 15, pp. 311–341, 1990.

J.S. Pang. A B-differentiable equation based, globally and locally quadratically convergent algorithm for nonlinear programs, complementarity and variational inequality problems. *Mathematical Programming*, 51, pp. 101–132, 1991.

J.S. Pang & S.A Gabriel. NE/SQP: A robust algorithm for the nonlinear complementarity problem. *Mathematical Programming*, 60, pp. 295–338, 1993.

J.S. Pang & D.E. Stewart. A unified approach to frictional contact problem. *International Journal of Engineering Science*, 37, pp. 1747–1768, 1999.

J.-S. Pang & D. Stewart. Differential variational inequalities. *Mathematical Programming A*, in press.

J.S. Pang & J.C. Trinkle. Complementarity formulations and existence of solutions of dynamic multi-rigid-body contact problems with Coulomb friction. *Mathematical Programming*, 73, pp. 199–226, 1996.

L. Paoli. Continuous dependence on data for vibro-impact problems. *Mathematical Models and Methods in Applied Sciences*, 15(1), pp. 53–93, 2005a.

L. Paoli. An existence result for non-smooth vibro-impact problems. *Journal of Differential Equations*, 211, pp. 247–281, 2005b.

L. Paoli & M. Schatzman. Mouvement à nombre fini de degres de liberteé avec contraintes unilatérales. Cas avec perte d'énergie. *M*odelisation Mathématique et analyse Numérique, 27, pp. 673–717, 1993.

L. Paoli & M. Schatzman. A numerical scheme for impact problems. Preprint 300, MAPLY, Mathématiques appliquées de Lyon, UMR 5585, Lyon, 1999.

L. Paoli & M. Schatzman. A numerical scheme for impact problems I: the one-dimensional case. *SIAM Journal of Numerical Analysis*, 40(2), pp. 702–733, 2002a.

L. Paoli & M. Schatzman. A numerical scheme for impact problems II: the multi-dimensional case. *SIAM Journal of Numerical Analysis*, 40(2), pp. 734–768, 2002b.

J.K. Park & B.M. Kwak. Three dimensional frictional contact analysis using the homotopy method. *Journal of Applied Mechanics, Transactions of ASME*, 61, pp. 703–709, 1994.

M. Payr & C. Glocker. Oblique fritional impact of a bar: analysis and comparison of different impact laws. *Nonlinear Dynamics*, 41, pp. 361–383, 2005.

J.-M. Peng. Equivalence of variational inequality problems to unconstrained minimization. *Mathematical Programming*, 78, pp. 347–355, 1997.

J. Pérès. *Mécanique Générale*. Masson, 1953.

F. Pfeiffer & C. Glocker. *Multibody Dynamics with Unilateral Contacts. Non-linear Dynamics*. Wiley, New York, 1996.

F. Pfeiffer, M. Foerg & H. Ulbrich. Numerical aspects of non-smooth multibody dynamics. *Computer Methods in Applied Mechanics and Engineering*, 195, pp. 6891–6908, 2006.

A.Yu. Pogromsky, W.P.M.H. Heemels & H. Nijmeijer. On solution concepts and well-posedness of linear relay systems. *Automatica*, 39, pp. 2139–2147, 2003.

M.K. Ponamgi, D. Manosha & M.C. Lin. Incremental algorithm for collision detection between solid models. *Proceedings of 3rd ACM Symposium on Solid Modeling and Applications*, pp. 293–304. ACM Press, 1995.

F.A. Potra & X. Liu. Predictor–corrector methods for sufficient linear complementarity problems in a wide neighborhood of the central path. *Optimization Methods and Software*, 20(1), pp. 145–168, 2005.

F.A. Potra & R. Sheng. A large-step infeasible interior point method for the $P_\star$-matrix LCP. *SIAM Journal of Optimization*, 7(2), pp. 318–335, 1997.

F.A. Potra & S.J. Wright. Interior-point methods. *Journal of Computational and Applied Mathematics*, 124, pp. 281–302, 2000.

F. Potra & Y. Ye. Interior-point methods for nonlinear complementarity problems. *Journal of Optimization Theory and Applications*, 88(3), pp. 617–642, 1996.

F.A. Potra, M. Anitescu, B. Gavrea & J. Trinkle. A linearly implicit trapezoidal method for integrating stiff multibody dynamics with contact, joints, and friction. *International Journal for Numerical Methods in Engineering*, 66(7), pp. 1079–1124, 2006.

M.J.D. Powell. A method for nonlinear constraints in minimization problems. R. Fletcher, editor, *Optimization*, pp. 283–298. Academic Press, 1969.

E. Pratt, A. Leger & M. Jean. Critical oscillations of mass-spring systems due to nonsmooth friction. *Archives of Applied Mechanics*, 2007, in press.

A.A.B. Pritsker & N.R. Hunt. GASP IV, a combined continuous discrete FORTRAN-based simulation language. *Simulation*, pp. 65–70, 1973.

B.N. Pshenichny. *Convex Analysis and Optimization*. Nauka, Moscow, 1980. (In Russian).

L. Qi & J. Sun. A nonsmooth version of Newton's method. *Mathematical Programming*, 58, pp. 353–367, 1993.

L. Qi & Y.-F. Yang. NCP functions applied to Lagrangian globalization for the nonlinear complementarity problem. *Journal of Global Optimization*, 24(2), pp. 261–283, 2002.

F. Radjai. Multicontact dynamics of granular systems. *Computer Physics Communications*, 121–122, pp. 294–298, 1999.

F. Radjai & S. Roux. Turbulentlike fluctuations in a quasistatic flow of granular media. *Physical Review Letters*, 89, page 064302, 2002.

F. Radjai & D.E. Wolf. Features of static pressure in dense granular media. *Granular Matter*, 1(1), pp. 3–8, 1998.

F. Radjai, M. Jean, J.J. Moreau & S. Roux. Force distributions in dense two-dimensional granular systems. *Physical Review Letters*, 77, pp. 274–277, 1996.

F. Radjai, J. Schaefer, S. Dippel & D. Wolf. Collective friction of an array of particles: a crucial test for numerical algorithms. *Journal of Physics Int.*, 7, page 1053, 1997.

F. Radjai, D. Wolf, M. Jean & J.J. Moreau. Bimodal character of stress transmission in granular packing. *Physical Review Letters*, 90, pp. 61–64, 1998.

F. Radjai, S. Roux & J.J. Moreau. Contact forces in a granular packing. *Chaos*, 9(3), pp. 544–550, 1999.

D. Ralph. Global convergence of damped Newton's method for nonsmooth equations, via the path search. *Mathematics of Operations Research*, 9(2), pp. 352–389, 1994.

M. Raous, L. Cangémi & M. Cocu. Consistent model coupling adhesion, friction and unilateral contact. *Computer Methods in Applied Mechanics and Engineering*, 177 (3–4), pp. 383–399, 1999.

M. Renouf & P. Alart. Conjugate gradient type algorithms for frictional multicontact problems: applications to granular materials. *Computational Methods in Applied Mechanics and Engineering*, 194(18–20), pp. 2019–2041, 2004a.

M. Renouf & P. Alart. Solveurs parallèles pour la simulation de systèmes multicontacts. *Revue Européenne des Eléments Finis*, 13, pp. 691–702, 2004b.

M. Renouf, F. Dubois & P. Alart. A parallel version of the nonsmooth contact dynamics algorithm applied to the simulation of granular media. *Journal of Computational and Applied Mathematics*, 168, pp. 375–38, 2004.

M. Renouf, V. Acary & G. Dumont. 3D frictional contact and impact multibody dynamics. A comparison of algorithms suitable for real-time applications. *ECCOMAS Thematic Conference Multibody Dynamics 2005*, Madrid, June 2005a.

M. Renouf, D. Bonamy, P. Alart & F. Dubois. Influence of the lateral friction on the surface flow – a 3D numerical approach, *Powders and Grains*. Stuttgart, July 2005b.

M. Renouf, D. Bonamy, F. Dubois & Alart P. Numerical simulation of 2D steady granular flows in rotating drum: on surface flows rheology. *Physics of Fluids*, 17, pp. 103303/1–103303/12, 2005c.

S.M. Robinson. Generalized equations and their solutions. I. Basic theory. *Mathematical Programming Study*, 10, pp. 128–141, 1979.

S.M. Robinson. Strongly regular generalized equations. *Mathematics of Operations Research*, 5, pp. 43–62, 1980.

S.M. Robinson. Generalized equations and their solutions. II. Applications to nonlinear programming. *Mathematical Programming Study*, 19, pp. 200–221, 1982.

S.M. Robinson. Newton's method for a class of nonsmooth equations. Technical report, Department of Industrial Engineering, University of Wisconsin-Madison, 1988.

S.M. Robinson. Normal maps induced by linear transformations. *Set-Valued Analysis*, 2, pp. 291–305, 1992.

R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

R.T. Rockafellar. The multiplier method of Hestenes and Powell applied to convex programming. *Journal of Optimization Theory and Applications*, 126), pp. 555–562, 1973.

R.T. Rockafellar. Augmented Lagrange multiplier functions and duality in nonconvex programming. *SIAM Journal on Control*, 12, pp. 268–285, 1974.

R.T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations research*, 1(2), pp. 97–116, 1976a.

R.T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal of Control and Optimization*, 14, pp. 877–898, 1976b.

R.T. Rockafellar. Lagrange multipliers and variational inequalities. J.L. Lions, F. Gianessi & R.W. Cottle, editor, *Variational Inequalities and Complementarity Problem*. Wiley, New York, 1979.

R.T. Rockafellar. Lagrange multipliers and optimality. *SIAM Review*, 35(2), pp. 183–238, 1993.

R.T. Rockafellar & J.B. Wets. *Variational Analysis*. Springer, 1998.

J.B. Rosen. The gradient projection method for nonlinear programming. Part I. Linear constraints. *Journal of the Society for Industrial and Applied Mathematics*, 8(1), pp. 181–217, 1960.

J.B. Rosen. The gradient projection method for nonlinear programming. II. Nonlinear constraints. *Journal of the Society for Industrial and Applied Mathematics*, 9, pp. 514–532, 1961.

T. Rutherford. Miles: a mixed inequality and nonlinear equation solver, working paper, Deportment of Economics, University of Colorado Bolder, 1993.

P. Sannuti. Direct singular pertubation analysis of high gain and cheap control problem. *Automatica*, 19(1), pp. 424–440, 1983.

G. Saussine. *Contribution à la modèlisation de granulats tridimensionnels: Application au ballast*. PhD thesis, Laboratoire de Mécanique et de Génie Civil, Université de Montpellier II, Montpellier, France, 2004.

G. Saussine, C. Cholet, P.E. Gautier, F. Dubois, C. Bohatier & J.J Moreau. Modelling ballast under cyclic loading using discrete element method. *Proceedings of International Conference on Cyclic Behaviour of Soils and Liquefaction Phenomena*. Balkema, April 2004a.

G. Saussine, C. Cholet, P.E. Gautier, F. Dubois, C. Bohatier & J.J. Moreau. Modelling ballast behaviour using a three-dimensional polyhedral discrete element method. *XI International Congress of Theoretical and Applied Mechanics*, Warsaw, Poland, August 2004b.

G. Saussine, F. Dubois, C. Bohatier, C. Cholet, P.E. Gautier & J.J. Moreau. Modelling ballast behaviour under dynamic loading, part 1: a 2D polygonal discrete element method approach. *Computer Methods in Applied Mechanics and Engineering*, 195(19–22), pp. 2841–2859, 2006.

M. Schatzman. A class of nonlinear differential equations of second order in time. *Nonlinear Analysis, TMA*, 2(3), pp. 355–373, 1978.

L. Schwartz. *Analyse III, Calcul Intégral.*. Hermann, 1993.

L.F. Shampine, I. Gladwell & R.W. Brankin. Reliable solution of special event location problems for ODEs. *ACM Transactions on Mathematical Software*, 17(1), pp. 11–25, 1991.

D. Shevitz & B. Paden. Lyapunov stability theory of nonsmooth systems. *IEEE Transactions on Automatic Control*, 39(9), pp. 1910–1914, 1994.

G.E. Shilov & B.L. Gurevich. *Integral Measure and Derivative. A Unified Approach.* Prentice-Hall, Englewood Cliffs, NJ, 1966. Hermann, Paris, 1993.

M. Sibony. Méthodes itératives pour les équations et inéquations aux dérivées partielles non linéaires de type monotone. *Calcolo*, 7, pp. 65–183, 1970.

J.C. Simo & T.A. Laursen. An augmented Lagrangian treatment of contact problems involving friction. *Computers & Structures*, 42(1), pp. 97–116, 1992.

J.C. Simo & N. Tarnow. The discrete energy–momentum method. Conserving algorithms for nonlinear elastodynamics. *Zeitschrift für Angewandte Mathematik und Physik*, 43, pp. 757–792, 1992.

G.V. Smirnov. Discrete approximations and optimal solutions to differential inclusions. *Cybernetics*, 27, pp. 101–107, 1991.

G. Smirnov. *Introduction to the Theory of Differential Inclusions*, volume 41 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2002.

E.E. Soellner & C. Führer. *Numerical Methods in Multibody Mechanics*. Teubner, Stuttgart, 1998. Corrected reprint Lund, 2002.

M.V. Solodov & P. Tseng. Modified projection-type methods for monotone variational inequalities. *SIAM Journal on Control and Optimization*, 34(5), pp. 1814–1830, 1996.

W. Son, K. Kim, N.M. Amato & J.C. Trinkle. A generalized framework for interactive dynamic simulation for multirigid bodies. *IEEE Transactions on Systems, Man and Cybernetics–Part B: Cybernetics*, 34(2), pp. 912–924, April 2004.

D. Stewart. A high accuracy method for solving ODEs with discontinuous right-hand-side. *Numerische Mathematik*, 58, pp. 299–328, 1990.

D. Stewart. Convergence of a time-stepping scheme for rigid-body dynamics and resolution of Painlevé's problem. *Archives for Rational Mechanics and Analysis*, 145, pp. 215–260, 1998.

D. Stewart. Rigid body dynamics with friction and impact. *S.I.A.M. Review*, 42(1), pp. 3–39, 2000.

D. Stewart. Reformulations of measure differential inclusions and their closed graph property. *Journal of Differential Equations*, 175(1), pp. 108–129, 2001.

D.E. Stewart & M. Anitescu. Optimal control of systems with discontinuous differential equations. *Preprint ANL/MCS-P1258-0605, Math and Comp. Sci. Division, Argonne Natl. Lab.*, 2006.

D.E. Stewart & J.C. Trinkle. An implicit time-stepping scheme for rigid body dynamics with inelastic collisions and Coulomb friction. *International Journal for Numerical Methods in Engineering*, 39(15), pp. 2673–2691, 1996.

W.J. Stronge. *Impact Mechanics*. Cambridge University Press, 2000.

D. Sun & L. Qi. On NCP-functions. *Computational Optimization and Applications*, 13, pp. 201–220, 1999.

R. Sznajder & M.S. Gowda. Generalizations of P_0,P-properties; extended vertical and horizontal LCPs. *Linear Algebra and Its Applications*, 223–224, pp. 695–715, 1995.

K. Taji, M. Fukushima & T. Ibaraki. A globally convergent Newton method for solving strongly monotne variational inequalities. *Mathematical Programming*, 58, pp. 369–383, 1993.

K. Taubert. Converging multistep methods for initial value problems involving multivalued maps. *Computing*, 27, pp. 123–136, 1981.

L. Thibault. Sweeping process with regular and nonregular sets. *Journal of Differential Equations*, 193(1), pp. 1–26, 2003.

A. Tonnelier & W. Gerstner. Piecewise linear differential equations and integrate-and-fire neurons: insights from two-dimensional membrane models. *Physical Review E*, 67, pp. 21908–21924, 2003.

A.A. Transeth, R.I. Leine, Ch. Glocker & K.Y. Pettersen. Non-smooth 3D modeling of a snake robot with external objects. *Proceedings of the 2006 IEEE International Conference on Robotics and Biomimetics (ROBIO2006)*, Kunming, China, December 2006a.

A.A. Transeth, R.I. Leine, Ch. Glocker & K.Y. Pettersen. Non-smooth 3D modeling of a snake robot with frictional unilateral constraints. *Proceedings of the 2006 IEEE International Conference on Robotics and Biomimetics (ROBIO2006)*, Kunming, China, December 2006b.

J.C. Trinkle, J.S. Pang, S. Sudarsky & G. Lo. On dynamic multi-rigid-body contact problems with coulomb friction. *Zeitschrift für Angewandte Mathematik und Mechanik*, 77, pp. 267–279, 1997.

J.C. Trinkle, J.A. Tzitzouris & J.S. Pang. Dynamic multi-rigid-body systems with concurrent distributed contacts. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 359(1789), pp. 2575–2593, December 2001.

P. Tseng. Further applications of a splitting algorithm to decomposition in variational inequalities and convex programming. *Mathematical Programming*, 48(2), pp. 249–263, 1990.

P. Tseng. On linear convergence of iterative methods for the variational inequality problem. *Proceedings of the International Meeting on Linear/Nonlinear Iterative Methods and Verification of Solution*, pp. 237–252, Amsterdam, The Netherlands. Elsevier Science, The Netherlands, 1995.

J.D. Turner. On the simulation of discontinuous functions. *ASME Journal of Applied Mechanics*, 68, pp. 751–757, September 2001.

J.A. Tzitzouris & J.S. Pang. A time-stepping complementarity approach for frictionless systems of rigid bodies. *SIAM Journal on Optimization*, 12(3), pp. 834–860, 2002.

V.I. Utkin. Variable structure systems with sliding modes: a survey. *IEEE Transactions on Automatic Control*, 22, pp. 212–222, 1977.

H. Väliaho. $P_\star$-matrices are just sufficient. *Linear Algebra and Its Applications*, 239, pp. 103–108, 1996.

A.J. van der Schaft & J.M. Schumacher. The complementary-slackness class of hybrid systems. *Mathematics of Control, Signals, and Systems*, 9(3), pp. 266–301, 1996.

A.J. van der Schaft & J.M. Schumacher. Complementarity modeling of hybrid systems. *IEEE Transactions on Automatic Control*, 43(4), pp. 483–490, 1998.

A. van der Schaft & J.M. Schumacher. *An Introduction to Hybrid Dynamical Systems*. Springer, London, 2000.

V. Veliov. Second-order discrete approximation to linear differential inclusions. *SIAM Journal of Numerical Analysis*, 29(2), pp. 439–451, 1992.

D. Vola, E. Pratt, M. Jean & M. Raous. Consistent time discretization for dynamical frictional contact problems and complementarity techniques. *Revue Européenne des éléments finis*, 7(1-2-3), pp. 149–162, 1998.

B. von Herzen, A.H. Barr & H.R. Zatz. Geometric collisions for time-dependent parametric surfaces. *SIGGRAPH Conference Proceedings, ACM Press*, pp. 39–48, 1990.

L. Vu & D. Liberzon. Common Lyapunov functions for families of commuting nonlinear systems. *Systems and Control Letters*, 54, pp. 405–416, 2005.

Y. Wang. Dynamic modelling and stability analysis of mechanical systems with time-varying topologies. *ASME Journal of Mechanical Design*, 115, pp. 808–816, 1993.

Y. Wang & M.T. Mason. Two-dimensional rigid-body collisions with friction. *Journal of Applied Mechanics, Transactions of A.S.M.E*, 59, pp. 635–642, 1992.

D. Wang, C. Conti, P. Dehonbruex & Verlinden O. A computer-aided simulation approach for mechanisms with time-varying topology. *Computers and Structures*, 64, pp. 519–530, 1997.

D. Wang, C. Conti & D. Beale. Interference impact analysis of multibody systems. *ASME Journal of Mechanical Design*, 121, pp. 128–135, 1999.

S.J. Wright. Implementing a proximal point methods for linear programming. Technical Report MCS-P45-0189, Argonne National Laboratory., IL., 1989.

S.J. Wright. Implementing proximal point methods for linear programming. *Journal of Optimization Theory and Applications*, 65(3), pp. 531–554, 1990.

S.J. Wright. A path-following interior-point algorithm for linear and quadratic problems. *Annals of Operations Research*, 62, pp. 103–130, 1996a. Preprint MCS–P401–1293, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA, December 1993.

S.J. Wright. *Primal-Dual Interior-Point Methods*. SIAM, Philadelphia, 1996b.

S.C. Wu, S.M. Yang & E.J. Haug. Dynamics of mechanical systems with Coulomb friction, stiction, impact and constraint addition–deletion; ii–planar systems. *Mechanisms and Machine Theory*, 21, pp. 407–416, 1986.

L. Xuewen, A.-K. Soh & C. Wanji. A new nonsmooth model for three-dimensional frictional contact problems. *Computational Mechanics*, 26, pp. 528–535, 2000.

N. Yamashita, K. Taji & M. Fukushima. Unconstrained optimization reformulations of variational inequality problems. *Journal of Optimization Theory and Applications*, 92, pp. 439–456, 1997.

J.C. Yao. Variational inequalities with generalized monotone operators. *Mathematics of Operations Research*, 19(3), pp. 691–705, 1994.

Y. Ye. A fully polynomial-time approximation algorithm for computing a stationary point of the general linear complementarity problem. *Mathematics of Operations Research*, 18, pp. 334–345, 1993.

Y. Ye. *Interior Point Algorithms*. Series on Discrete Mathematics and Optimization. Wily, New York, 1997.

A. Zervos, I. Vardoulakis, M. Jean & P. Lerat. Numerical investigation of granular interfaces kinematics. *Mechanics of Cohesive-Frictional Materials*, 5(4), pp. 305–324, 2000.

Z. Zhao, C. Liu & W.M. Chen. Experimental investigation of the Painlevé paradox in a robotic system. ASME Journal of Applied Mechanics, to appear.

D.L. Zhu & P. Marcotte. Modified descents methods for solving the monotone variational inequality problem. *Operations Research Letters*, 14, pp. 111–120, 1993.

D.L. Zhu & P. Marcotte. An extended descent framework for monotne variational inequalities. *Journal of Optimization Theory and Applications*, 80, pp. 349–366, 1994.

Index

P–matrix, 352
P_0–matrix, 353
$P_K(\cdot)$, 478
$P_\star$–matrix, 354
θ-method, 256, 287
θ-method
  energy conservation, 279
$f = O(g)$, 213
$f = o(g)$, 213
AVI, 389
DI
  Filippov, 24
EVI, 13
Evolution Variational Inequality, 13
LCP, 7
Linear Complementarity Problem, 7
LCP
  feasible, 351
LCP existence and Uniqueness, 352
LCS, 6
Measure Differential Inclusion (MDI) , 44
NLP
  Gradient-projection methods , 350
NSGS, 437
QP
  Active-set methods, 339
  dual problem, 333
  Gradient–projection methods , 341
  Interior point methods, 343
  lagrangian relaxation, 333
Mixed Complementarity Problem (MCP), 383
VI
  evolution, 13
3-parameter impact law, 153

B-differentiability, 429

Accumulation of events, 194
Active-set methods, 339
Affine complementarity system, 270
Augmented lagrangian function, 347

B-differentiable Newton's method, 429
Backward Euler method, 269, 276
Ballast modelling, 308
Bilateral constraints, 109, 225
Bipotential function, 148
Bohl
  distributions, 168, 487
  functions, 168, 487
Bouncing ball, 41
Bounded variation
  function, 481
  multivalued function, 482
Brazil nuts effect, 307, 317
Bulk segregation, 307
BV, 481

C
  function, 385
Carathéodory
  solutions, 65
Catching-up algorithm, 271
Central path, 375
Chattering behaviour, 30
Collision detection, 128

Complementarity, 2
Complementarity
  function, 385
Complementarity problem
  and VI, 390
  extended, 356
  general form, 390
  generalized, 356
  geometric, 356
  horizontal, 355
  linear, 7, 37, 274, 351
  mixed, 383
  mixed linear, 26, 355
  nonlinear, 274, 379
  vertical, 356
Complementarity systems
  affine, 270
  and impulsive ODEs, 103
  differential, 83
  dynamical, 26, 85, 170
  equivalence to sweeping process, 170
  fixed points, 19
  gradient, 84
  jump rule, 168
  linear, 6, 19, 86, 95, 96, 275
  linear non autonomous, 167
  mixed linear, 26, 166
  non linear, 88
  nonlinear passive, 172
  passive linear, 168
  time-stepping scheme, 275
Cone
  dual, 83, 165, 390
  kinetic normal, 477
  normal, 4, 23, 73, 74, 79
  polar, 74
  tangent, 74, 79, 85, 91, 182
Conjugate function, 23
Consistency, 236
Consistency of the gap, 294
Constraint qualification, 79
Continuum media, 131
Convergence, 236
Convex process, 70
Coordiante system, 112
Coordinates reduction, 111
Copositive
  plus, 374
  strictly, 479
Copositive matrix, 210, 374, 479
Coulomb friction, 40, 81, 155, 283, 314
Coulomb friction
  cone faceting, 410, 417
  enhanced, 155
  Glocker's formulation, 420
  Masing's model, 157
  Prandtl's model, 157
  rheological models, 157
  weakness of cone faceting, 418
Coulomb's friction, 145

Dahl's model (modified), 157
DCS, 166
De Saxcé's bipotential, 148
Deep drawing, 309
Deformable bodies, 131
Delassus' example, 229
Delassus' operator, 297, 303
Differential inclusion
  unilateral, 77
Differential inclusions, 13, 57, 252
Differential inclusions
  hypomonotone, 73
  implicit, 78, 89
  in normal cones, 262
  Lipschitzian, 59, 252
  maximal monotone, 73, 78
  one-sided Lipschitz, 69, 255
  outer semi-continuous, 61, 254
  unilateral, 13, 77, 274
  upper semi-continuous, 61, 254
  with infinitely many solutions, 264
Differential measure, 44
Differential variational inequalities, 82
Differential variational inequalities
  boundary value problem, 281
  initial value problem, 280
Diode, 2
Diode
  ideal, 3
  Shockley's law, 3
  Zener, 21
Dirac measures
  in electrical circuits, 54
Discontinuity w.r.t.initial data, 189
Dissipation (maximal), 148
Distribution differential inclusion, 178
Distributional formalism, 183

DOPRI5 method, 240
Dorn's duality, 363
Duality, 39

Electrical circuits, 6, 87
Electrical circuits
with Dirac measures, 54
EMTS, 322
Equilibrium points, 19
Euler angles, 115
Euler method
implicit, 256
perturbed, 255
Event, 38
Event
nonsmooth, 38
Event detection
linear interpolation, 237
second-order interpolation, 237
third-order interpolation, 237
Event-driven methods, 10, 39
Event-driven methods
consistency, 236
order, 236
Evolution variational inequalities, 80, 169
Explicit Euler method, 252

Fenchel's equality, 149
Filippov
DI, 24, 63
solutions, 65
system, 24
Filippov's inclusion, 63, 95, 96, 194, 257
Filippov's inclusion
and generalized derivatives, 67
Event-driven scheme, 203
event-driven scheme, 205
in mechanics, 162
LCP of the active set, 209
piecewise active solution, 206
smoothing method, 267
spontaneous switches, 210
Stewart's method, 205
Fracture dynamics, 314
Frictional paroxysms, 192, 315
Function
conjugate, 23
Function
indicator, 3
one-sided differentiable, 257
relay, 25
right-sided continuously differentiable, 261
support, 23

Gap function, 125
Gap function
consistency of, 294
discretization, 294
Generalized equation, 396
Glocker's inclusion, 224
Gradient
of convex analysis, 4
Gradient–projection methods
QP , 341
Gradient-projection methods
NLP , 350
Goldstein–Levitin–Polyak, 350
Rosen, 351
Granular material, 307
Granular material
parallelization of algorithms, 309
size segregation, 307
Gronwall's Lemma, 487

Hausdorff continuity, 60
Hausdorff distance, 475
Higher order sweeping process, 177
Hille-Yosida theorem, 78
Horizontal monotone LCP, 374
HOSP, 177, 319
Huygens' theorem, 116
Hypomonotone, 72

Impact time calculation
linear interpolation, 237
second-order interpolation, 237
third-order interpolation, 237
Impact with friction, 151
Implicit $\theta-$method, 256
Implicit *vs* catching-up, 50
Impulsive ODEs, 100
Impulsive ODEs
and complementarity systems, 103
Inclusion, 5
Inconsistent states, 192
Indeterminate states, 192
Indicator function, 3, 23

Inertia matrix, 116
Infinite-dimensional systems, 131
Interior point methods, 343
Inverters, 464
Iteration matrix, 288

Kalman-Yakubovic-Popov LMI, 168, 276
Karush–Kuhn–Tucker conditions, 332, 346
Kato's theorem, 82
Kinetic energy of a solid, 117

Lagrange and Newton-Euler dynamics, 120
Lagrange multipliers, 332
LBV, 481
LCP, 7, 12, 15, 351
  horizontal, 355
LCP
  and QP, 359
  definition, 351
  extended, 356
  generalized, 356
  geometric, 356
  horizontal, 360
  monotone horizontal, 357
  pivoting methods, 367
  splitting–based methods, 363
  vertical, 355
LCS
  definition, 166
  discretization, 14
Left-limit, 481
Lexicographic inequalities, 373
Line–search
  Armijo, 350
Linear Complementarity System, 6
Linear matrix inequality, 276
Lipschitz continuity, 58, 71
Local kinematics, 123

Markov parameter, 97
Markov parameters, 174
Masing model, 157
Masonry structures, 309
Matrix
  P, 352
  P_0, 353
  $P_\star$, 354
  $P_\star$, 354
  column adequate, 354
  copositive, 479
  copositive plus, 374
  Q, 354
  strictly copositive, 479
matrix
  sufficient, 354
Maximal dissipation, 148
Maximal monotone, 61, 72, 73, 77, 81, 256, 396
Maximality, 22
MCP, 420
Measure
  differential, 44
  differential inclusion, 44
  Stieltjes, 44
Measure differential formalism, 183
Measure differential inclusion, 89
Mechanical systems
  Lagrange equations, 107
  linear time-invariant, 108
  Newton-Euler equations, 112
Merit function, 382
Mid-point rule, 256
Minimum norm solution, 264, 265
Minimum ratio test, 370
MLCP, 27, 355, 415
MLCP
  and QP, 359
MLCS, 27, 166
Modes, 39
Moment of inertia, 116
Monotone, 380
Monotonicity, 22, 357
Moreau's superpotential, 149
Moreau's sweeping process, 45, 74, 88, 137
MOS transistors, 464
Multifunction
  Lipschitz, 58, 71, 75
  one-sided Lipschitz, 68
  sign, 40
Multiple impacts, 141
Multistep method, 258
Murty's least index pivoting method, 369

Natural map, 397
NCP, 379
Newmark's method, 240
Newton's impact law, 144
Newton's Method

Alart & Curnier, 424
Newton-Euler and Lagrange dynamics, 120
Newton-Euler dynamics, 119
NLP, 345
KKT conditions, 346
NLP
augmented lagrangian function, 347
dual problem, 347
Nonlinear programming problem, 345
Normal cone, 4, 74, 79, 111
Normal cone in the kinetic metric, 477
Normal map, 397
NSCD method, 285
Numerical simulation
electrical circuits, 9, 14

Olinde-Rodrigues formula, 115
One-sided Lipschitz continuity, 68
One-sided Lipschitz DI, 255
Order, 236
Order of consistency, 261
Order of convergence, 261
Order of event-driven methods
with finite number of impacts, 239
with infinite number of impacts, 239
Ordinary DI, 71
OSNSP, 296
Outer semi-continuous, 61

P-function, 381
P-matrix, 7
Painlevé paradox, 192, 315
Parallelization of algorithms, 309
Passive systems, 276
PGS, 364
Piecewise affine systems, 94
Piecewise continuous systems, 97
Piecewise linear system, 9
Piecewise smooth systems, 95
Pontryagin's maximum principle, 178
Positive real system, 322
Practicable schemes, 260
Prandtl's model, 157
Principal minors, 352
Principle of maximal dissipation, 148
$\mathrm{proj}[K;\cdot]$, 478
$\mathrm{proj}_K(\cdot)$, 478
projected
Gauss–Seidel method, 364
Jacobi method, 364
Successive Overrelaxation method, 364
Projected dynamical systems, 84
Projection, 8
Projection (notation for), 478
PSOR, 364

QP, 332
KKT conditions, 332
QP
and LCP, 359
lagrangian function, 332
nonconvex, 334
strictly convex, 334
Quadratic programming problem, 332
Quasi-variational inequality, 390

RCLBV, 481
RCSLBV, 482
Reachable sets, 254
Real-time simulations, 311
Reduced coordinates, 111
Relative degree, 18, 173
Relative local velocity, 123
Relay function, 25, 27, 61, 73
Relay system, 269
Rheological models, 157
Right-limit, 481
RK24 method, 240
Runge–Kutta
DI in normal cones, 262
Runge–Kutta method, 259

Saddle-point problem, 394
Sannuti's canonical form, 179
SBV, 481
Screw
dynamic, 118
kinematic, 115
kinetic, 116
of acceleration, 118
the twist, 115
second–order cone complementarity problem, 295
second-order cone, 391
second-order cone complementarity problem, 151
Selection procedure, 252, 264, 273
Selection procedure

minimal variation, 265
minimum norm, 264
Set-valued mappings
hypomonotonicity, 72, 73
linear growth, 72
Lipschitz continuity, 59
maximal monotonicity, 61, 72, 73, 78
one-sided Lipschitz, 69
outer semi-continuity, 61
upper semi-continuity, 61
Sign multifunction, 40
Simulation of
assembly process, 241
clutch vibrations, 241
deep drawing systems, 309
drop tower hydraulics, 232
gear rattling, 241
granular material, 307
hammering in gears, 241
inverters, 464
masonry, 314
MOS transistors, 464
roller coaster, 232
rolling disk, 314
snake robot, 314
tensegrity structures, 309
turbine blade damper, 241
woodpecker toy, 241, 463
circuit breakers, 241
ship turning gear, 241
Size segregation, 307
Sliding local velocity, 124
Slow solutions, 264
Software packages
Humans, 443
LMGC90, 443
SICONOS, 443
Spatial displacement, 118
Special bounded variation, 481
Spontaneous switches, 63, 194
Stability, 236
Stieltjes matrix, 188
Stieltjes measure, 44
Stiff systems, 50
Strictly monotone, 380
Strongly monotone, 380
Subdifferential, 4, 22, 40
Subgradients, 4
Superpotential, 149
Support function, 23
Sweeping process
catching-up algorithm, 271
discretization, 46, 271
equivalence to LCS, 170
first order, 74
first order RCBV, 77
higher order, 177, 319
Lispchitz, 76
perturbed, 77, 170, 173
second order, 88
Switch model, 267
Switched systems, 98
Switching diagrams, 230
Switching logic, 230
Switching surface, 63

Tableau, 368
Tangent cone
inclusion in, 223
Tangent cone, 74, 79
Tangential impact, 193
Tangential restitution, 153
Tensegrity structures, 309
Time-stepping algorithm
θ-method, 256
Explicit Euler, 252
extended, 322
for mechanical systems, 285
for the HOSP, 319
linear multistep, 258
Moreau's catching-up, 271
Runge–Kutta, 259
selection procedure, 264
Time-stepping methods, 10, 28
Transversality conditions, 215
Tresca's friction, 434
Twist of a rigid body, 115

UDI, 13
Unilateral constraints, 110
Unilateral differential inclusion, 13
Unilateral differential inclusion
discretization, 257, 274
Upper semi-continuous, 61

var(f)
variation of f, 481
Variable

basic , 368
blocking, 370
driving, 370
nonbasic , 368
Variational inequalities
and CP, 390
differential, 82, 279
evolution, 80, 83, 169
hemi, 82
mixed, 476
of first kind, 476
of second kind, 476
quasi, 81, 390
Variational inequality, 5, 19, 389
Variational inequality
linearly constrained, 389
Viability problem, 178
Virtual work, 122

Woodpecker toy, 463
Wrench, 117

xLCP, 356

ZD state representation, 179
Zener diode, 21
Zeno phenomenon, 195
Zero dynamics, 179, 324