# Reachability Analysis and Deterministic Global Optimization of Differential-Algebraic Systems

by

Joseph Kirk Scott

B.S. Chemical Engineering, Wayne State University (2006)
M.S. Chemical Engineering Practice, Massachusetts Institute of Technology (2008)

Submitted to the Department of Chemical Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Chemical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2012

Author . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Department of Chemical Engineering
April 30, 2012

Certified by . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Paul I. Barton
Lammot du Pont Professor of Chemical Engineering
Thesis Supervisor

Accepted by . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Patrick S. Doyle
Chairman, Department Committee on Graduate Theses

# Reachability Analysis and Deterministic Global Optimization of Differential-Algebraic Systems

by

Joseph Kirk Scott

Submitted to the Department of Chemical Engineering
on April 30, 2012, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Chemical Engineering

## Abstract

Systems of differential-algebraic equations (DAEs) are used to model an incredible variety of dynamic phenomena. In the chemical process industry in particular, the numerical simulation of detailed DAE models has become a cornerstone of many core activities including, process development, economic optimization, control system design and safety analysis. In such applications, one is primarily interested in the behavior of the model solution with respect variations in the model inputs or uncertainties in the model itself. This thesis addresses two computational problems of general interest in this regard.

In the first, we are interested in computing a guaranteed enclosure of all solutions of a given DAE model subject to a specified set of inputs. This analysis has natural applications in uncertainty quantification and process safety verification, and is used for many important tasks in process control. However, for nonlinear dynamic systems, this task is very difficult. Existing methods apply only to ordinary differential equation (ODE) models, and either provide very conservative enclosures or require excessive computational effort. Here, we present new methods for computing interval bounds on the solutions of ODEs and DAEs. For ODEs, the focus is on efficient methods for using physical information that is often available in applications to greatly reduce the conservatism of existing methods. These methods are then extended for the first time to the class of semi-explicit index-one DAEs.

The latter portion of the thesis concerns the global solution of optimization problems constrained by DAEs. Such problems arise in optimal control of batch processes, determination of optimal start-up and shut-down procedures, and parameter estimation for dynamic models. In nearly all conceivable applications, there is significant economic and/or intellectual impetus to locate a globally optimal solution. Yet again, this problem has proven to be extremely difficult for nonlinear dynamic models. A small number of practical algorithms have been proposed, all of which are limited to ODE models and require significant computational effort. Here, we present improved lower-bounding procedures for ODE constrained problems and develop a complete deterministic algorithm for problems constrained by semi-explicit index-one DAEs for the first time.

Thesis Supervisor: Paul I. Barton
Title: Lammot du Pont Professor of Chemical Engineering

*To my Parents*

# Acknowledgments

I owe a great debt to Paul Barton for all of the time and energy he spent on the work in this thesis, as well as on my education and professional development. Throughout my thesis work, Paul has been attentive, interested and enthusiastic. He provided advice when I needed it, but also encouraged me to pursue my own ideas and develop as an independent researcher. I can be argumentative about research. Perhaps even stubborn. A thing is not true unless I understand why, regardless of who says it. One of the great things about working for Paul was that he always saw the value in debating openly. He does not take offense, and he does not placate. When we discussed my work, he treated my ideas and opinions with respect. When we disagreed, he always took the time to convince me of his view. For me, it was this attitude that made our work together so educational, productive and enjoyable.

Many other people have been instrumental to this thesis and to my professional development. I would like to thank my thesis committee, Professors George Stephanopoulos, Bill Green and Mike Henson, for all of their help and encouragement. I am very grateful to Professor Richard Braatz, for taking many hours of his time to teach me and give me career advice, and to Professor Benoit Chachuat, who has been a mentor and a friend to me throughout my thesis. Many thanks to the past and present members of the PSE laboratory for making it a fun and stimulating place to work. Special thanks to Benoit Chachuat, Matt Stuber, Spencer Schaber, Achim Wechsung and Kamil Khan for their contributions to this thesis and for enhancing my understanding of the material in countless ways through our many discussions.

I am ever indebted to all of my incredible friends in Detroit. My greatest ambitions were planted by them and they have never let me forget them. Special thanks to Brad Kelly for being an example and the metric by which I measure my own accomplishments. Many thanks to my sisters, and to Brad Elliot, who is like a brother to me, for all of their encouragement and support. I am endlessly grateful to Alisha Peterson, whom I largely ignored during the last hectic months of this writing, but who was nonetheless supportive, caring and encouraging. I can never thank her enough for

the hundreds of little ways she makes me a better, more productive, happier person, every day.

Finally, I am extremely grateful to my parents for their love and support, and for their dedication to my education. You always say you don't know where I got the brains from. What about the work ethic? The patience? The confidence that I can achieve anything? Do you know where I got those? I do.

# Contents

**1 Introduction** **19**

1.1 Overview . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19

1.2 Enclosing the Reachable Set . . . . . . . . . . . . . . . . . . . 21

1.2.1 Motivation . . . . . . . . . . . . . . . . . . . . . . . . . 22

1.2.2 Existing Approaches . . . . . . . . . . . . . . . . . . . . 23

1.2.3 Contributions . . . . . . . . . . . . . . . . . . . . . . . . 28

1.3 Global Dynamic Optimization . . . . . . . . . . . . . . . . . . 29

1.3.1 Motivation . . . . . . . . . . . . . . . . . . . . . . . . . 30

1.3.2 Existing Approaches for Local Dynamic Optimization . . . . . 32

1.3.3 Global Optimization of Standard NLPs . . . . . . . . . . . . 36

1.3.4 Existing Approaches for Global Dynamic Optimization . . . . 41

1.3.5 Contributions . . . . . . . . . . . . . . . . . . . . . . . . 45

**2 Factorable Functions, Interval Arithmetic and McCormick Relaxations** **49**

2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . 49

2.2 Factorable Functions . . . . . . . . . . . . . . . . . . . . . . . 51

2.3 Interval Analysis . . . . . . . . . . . . . . . . . . . . . . . . . 53

2.3.1 Inclusion Functions and Interval Extensions . . . . . . . . . . 54

2.3.2 Interval Arithmetic and the Natural Interval Extension . . . . 55

2.4 McCormick Analysis . . . . . . . . . . . . . . . . . . . . . . . 60

2.4.1 Relaxation Functions and McCormick Extensions . . . . . . . 62

2.4.2 McCormick Arithmetic and the Natural McCormick Extension 67

2.5 Regularity of Functions on $\mathbb{IR}^n$ and $\mathbb{MR}^n$ . . . . . . . . . . . . . . . . 81

2.5.1 $\mathbb{IR}^n$ and $\mathbb{MR}^n$ as Metric Spaces . . . . . . . . . . . . . . . . . 82

2.5.2 Lipschitz and Locally Lipschitz Functions . . . . . . . . . . . 85

2.5.3 Piecewise $C^1$ Functions . . . . . . . . . . . . . . . . . . . . . . . 91

2.5.4 Regularity of Factorable Functions . . . . . . . . . . . . . . . . 95

2.5.5 Regularity of Natural Interval Extensions . . . . . . . . . . . . 96

2.5.6 Regularity of Natural McCormick Extensions . . . . . . . . . 99

2.6 Standard McCormick Relaxations . . . . . . . . . . . . . . . . . . . . . 102

2.6.1 McCormick Relaxations on Sequences of Intervals . . . . . . . 103

2.7 Generalized McCormick Relaxations . . . . . . . . . . . . . . . . . . . 105

2.7.1 Partial Relaxations . . . . . . . . . . . . . . . . . . . . . . . . . . 110

2.7.2 Composite Relaxations . . . . . . . . . . . . . . . . . . . . . . . . 111

2.8 Univariate Interval and McCormick Extensions . . . . . . . . . . . . . 114

2.9 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 121

**3 State Bounding Theory for Parametric ODEs and Control Systems123**

3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 123

3.2 Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . 126

3.3 The Standard Differential Inequalities Method . . . . . . . . . . . . . 128

3.3.1 Proof of Theorem 3.3.2 . . . . . . . . . . . . . . . . . . . . . . . 133

3.3.2 Comments on Similar Results in the Literature . . . . . . . . 137

3.4 The Use of a Priori Enclosures in Comparison Theorems . . . . . . . 138

3.5 A General Comparison Theorem . . . . . . . . . . . . . . . . . . . . . 142

3.5.1 Specialization to State Bounds for ODEs . . . . . . . . . . . . 145

3.5.2 Computation of State Bounds . . . . . . . . . . . . . . . . . . . 149

3.5.3 Recovering Harrison's Method . . . . . . . . . . . . . . . . . . . 152

3.5.4 Extending $D_\Omega$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . 153

3.6 State Bounds with a Priori Enclosures . . . . . . . . . . . . . . . . . 154

3.6.1 An Interval Approach for General Enclosures . . . . . . . . . . 155

3.6.2 An Interval Approach for Convex Polyhedra . . . . . . . . . . 157

3.6.3 An Optimization Approach for Convex Polyhedra . . . . . . . 159
3.6.4 Comparison with Existing Results . . . . . . . . . . . . . . . . 160
3.7 Differential Inequalities with Switching Conditions . . . . . . . . . . . 161
3.7.1 Preliminaries . . . . . . . . . . . . . . . . . . . . . . . . . . . 162
3.7.2 A General Comparison Theorem with Switching Conditions . 164
3.7.3 Specialization to State Bounds for ODEs . . . . . . . . . . . . 168
3.7.4 Application to Convex Polyhedral a Priori Enclosures . . . . . 172
3.7.5 Application to Interval a priori Enclosures . . . . . . . . . . . 176
3.8 Conclusions and Future Work . . . . . . . . . . . . . . . . . . . . . . 177

**4 Bounding the Solutions of Chemical Kinetics Models 179**

4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 179
4.2 Physical Information in Reaction Models . . . . . . . . . . . . . . . . 180
4.2.1 Affine Reaction Invariants . . . . . . . . . . . . . . . . . . . . 181
4.2.2 Natural Bounds . . . . . . . . . . . . . . . . . . . . . . . . . . 182
4.2.3 A Polyhedral a Priori Enclosure . . . . . . . . . . . . . . . . . 184
4.3 Numerical Examples . . . . . . . . . . . . . . . . . . . . . . . . . . . 186
4.4 Reduced Kinetic Models . . . . . . . . . . . . . . . . . . . . . . . . . 194
4.4.1 Constructing Reduced Models . . . . . . . . . . . . . . . . . . 195
4.4.2 Reduced Space State Bounding . . . . . . . . . . . . . . . . . 200
4.5 Numerical Examples . . . . . . . . . . . . . . . . . . . . . . . . . . . 206
4.6 The Role of Redundancy . . . . . . . . . . . . . . . . . . . . . . . . . 219
4.7 Conclusions and Future Work . . . . . . . . . . . . . . . . . . . . . . 221

**5 State Bounding Theory for Semi-Explicit Index-One DAEs 223**

5.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 223
5.2 Preliminaries . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 226
5.3 Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . 227
5.3.1 Semi-explicit DAEs . . . . . . . . . . . . . . . . . . . . . . . . 227
5.3.2 Existence and Uniqueness . . . . . . . . . . . . . . . . . . . . 228
5.3.3 State Bounds . . . . . . . . . . . . . . . . . . . . . . . . . . . 230

5.3.4 Uniqueness Proofs . . . . . . . . . . . . . . . . . . . . . . . . 230
5.4 An Interval Inclusion Test for DAE Solutions . . . . . . . . . . . . . . 235
5.4.1 The Interval Hansen-Sengupta Method . . . . . . . . . . . . 235
5.4.2 An Interval Existence and Uniqueness Test for DAEs . . . . . 239
5.5 Bounding DAE Solutions using Differential Inequalities . . . . . . . . 242
5.6 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 250

**6 Computing State Bounds for Semi-Explicit Index-One DAEs 251**
6.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 251
6.2 Preliminaries . . . . . . . . . . . . . . . . . . . . . . . . . . . . 252
6.2.1 Extended Interval Functions . . . . . . . . . . . . . . . . . . 252
6.2.2 Regularity of Interval Functions . . . . . . . . . . . . . . . . 256
6.3 A Generic Two-Phase Algorithm . . . . . . . . . . . . . . . . . . . 258
6.4 Satisfying the Existence and Uniqueness Test Computationally . . . . 261
6.4.1 Specification of $\mathbf{C}_J$ and $\tilde{\mathbf{z}}_y$ . . . . . . . . . . . . . . . . . . . 266
6.4.2 Phase 1 Algorithm . . . . . . . . . . . . . . . . . . . . . . . 267
6.4.3 Phase 1 Refinement . . . . . . . . . . . . . . . . . . . . . . 269
6.5 Computing Refined Enclosures Using Differential Inequalities . . . . . 269
6.6 A Single-Phase Method . . . . . . . . . . . . . . . . . . . . . . . . 274
6.6.1 Computational Complexity . . . . . . . . . . . . . . . . . . . 281
6.7 Case Studies . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 283
6.8 Conclusions and Future Work . . . . . . . . . . . . . . . . . . . . 288

**7 State Relaxations for Parametric ODEs 293**
7.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 293
7.2 State Relaxations for a Generic Function . . . . . . . . . . . . . . . 297
7.3 Relaxation Amplifying Dynamics . . . . . . . . . . . . . . . . . . . 298
7.3.1 Proof of Theorem 7.3.4 . . . . . . . . . . . . . . . . . . . . . 301
7.4 Relaxation Preserving Dynamics . . . . . . . . . . . . . . . . . . . 305
7.5 Sufficiency of Relaxation Preserving Dynamics . . . . . . . . . . . . 309
7.5.1 Bounding properties . . . . . . . . . . . . . . . . . . . . . . 311

7.5.2 Convexity/concavity properties . . . . . . . . . . . . . . . . . . 318

7.6 State Relaxations for ODEs . . . . . . . . . . . . . . . . . . . . . . . . 321

7.6.1 Constructing Relaxation Amplifying Dynamics . . . . . . . . . . 323

7.6.2 Constructing Relaxation Preserving Dynamics . . . . . . . . . . 325

7.6.3 Implementation . . . . . . . . . . . . . . . . . . . . . . . . . . 328

7.7 Numerical Examples . . . . . . . . . . . . . . . . . . . . . . . . . . . 332

7.8 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 340

**8 State Relaxations for Semi-Explicit Index-One DAEs 345**

8.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 345

8.2 Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . 346

8.2.1 State Bounds and Related Assumptions . . . . . . . . . . . . . 346

8.2.2 The Auxiliary System . . . . . . . . . . . . . . . . . . . . . . 348

8.3 Relaxing the Algebraic States . . . . . . . . . . . . . . . . . . . . . . 349

8.3.1 Characterizing the Algebraic States . . . . . . . . . . . . . . . 349

8.3.2 An Iterative Relaxation Scheme . . . . . . . . . . . . . . . . . 352

8.4 Relaxation Amplifying Dynamics . . . . . . . . . . . . . . . . . . . . 363

8.5 Relaxation Preserving Dynamics . . . . . . . . . . . . . . . . . . . . . 369

8.6 Numerical Examples . . . . . . . . . . . . . . . . . . . . . . . . . . . 371

8.7 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 371

**9 Convex Polyhedral Enclosures of Reachable Sets 375**

9.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 375

9.2 Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . 376

9.3 Convex enclosures of reachable sets . . . . . . . . . . . . . . . . . . . 377

9.3.1 A dual representation of $A$ . . . . . . . . . . . . . . . . . . . 378

9.3.2 Computation of $A$ . . . . . . . . . . . . . . . . . . . . . . . . 379

9.4 Numerical Example . . . . . . . . . . . . . . . . . . . . . . . . . . . . 381

9.5 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 383

**10 Deterministic Global Optimization with DAEs Embedded 385**

10.1 Introduction . . . . . 385

10.2 Problem Statement . . . . . 386

10.3 A Global Optimization Algorithm . . . . . 389

10.3.1 The Upper-Bounding Procedure . . . . . 390

10.3.2 The Lower-Bounding Procedure . . . . . 390

10.3.3 Domain Reduction . . . . . 398

10.3.4 Generation Skipping . . . . . 398

10.3.5 Algorithm Statement . . . . . 399

10.4 Numerical Examples . . . . . 402

10.5 Conclusion . . . . . 416

**11 Convex Relaxations for Nonconvex Optimal Control Problems 419**

11.1 Introduction . . . . . 419

11.1.1 Basic Approach . . . . . 421

11.2 McCormick Relaxations on Vector Spaces . . . . . 422

11.3 Relaxing the Objective and Constraints . . . . . 425

11.4 Relaxing Integral Functionals . . . . . 427

11.5 State Relaxations for Control Systems . . . . . 429

11.5.1 Proof of Theorem 11.5.4 . . . . . 431

11.6 Conclusions . . . . . 435

**12 Concluding Remarks 437**

12.1 Summary of Contributions . . . . . 437

12.1.1 Algorithms . . . . . 437

12.1.2 Theoretical Contributions . . . . . 440

12.2 Outlook . . . . . 443

# List of Figures

3-1 Diverging state bounds computed by Harrison's method . . . . . . . . 131

3-2 Schematic of the bounds tightening procedure of Definition 3.6.3 . . . 158

4-1 Affine reaction invariants for A + B → C . . . . . . . . . . . . . . . . . 185

4-2 Comparison of bounding methods for Example 3.3.1 . . . . . . . . . . 188

4-3 Comparison of bounding methods for Example 4.3.2 . . . . . . . . . . 190

4-4 Comparison of bounding methods for Example 4.3.3 . . . . . . . . . . 193

4-5 Comparison of bounding methods for Example 4.3.3 . . . . . . . . . . 193

4-6 Geometric view of full-space vs. reduced-space state bounds . . . . . 196

4-7 Comparison of bounding methods for Example 4.5.1 . . . . . . . . . . 209

4-8 Comparison of bounding methods for Example 4.5.2 . . . . . . . . . . 211

4-9 Comparison of bounding methods for Example 4.5.3 . . . . . . . . . . 214

4-10 Comparison of bounding methods for Example 4.5.3 . . . . . . . . . . 215

4-11 Comparison of bounding methods for Example 4.5.4 . . . . . . . . . . 220

4-12 Comparison of bounding methods for Example 4.5.4 . . . . . . . . . . 220

6-1 Comparison of state bounds for Example 6.7.1 . . . . . . . . . . . . . 286

6-2 Comparison of state bounds for Example 6.7.1 . . . . . . . . . . . . . 287

6-3 Comparison of state bounds for Example 6.7.2 . . . . . . . . . . . . . 290

6-4 Comparison of state bounds for Example 6.7.2 . . . . . . . . . . . . . 291

6-5 Comparison of state bounds for Example 6.7.2 . . . . . . . . . . . . . 291

7-1 A parametric final time ODE solution (Example 7.7.1) . . . . . . . . 333

7-2 State relaxations (Example 7.7.1) . . . . . . . . . . . . . . . . . . . . 335

7-3 State relaxations on nested intervals (Example 7.7.1) . . . . . . . . . 336
7-4 State relaxations after a short time (Example 7.7.2) . . . . . . . . . . 338
7-5 State relaxations after a long time (Example 7.7.2) . . . . . . . . . . 339
7-6 State relaxations with narrow $P$ (Example 7.7.2) . . . . . . . . . . . 340
7-7 State relaxations with large $P$ (Example 7.7.3) . . . . . . . . . . . . . 341
7-8 Relaxations of absorbance with large $P$ (Example 7.7.3) . . . . . . . . 342
7-9 State relaxations with narrow $P$ (Example 7.7.3) . . . . . . . . . . . 342
7-10 Relaxations of absorbance with narrow $P$ (Example 7.7.3) . . . . . . 343

8-1 State relaxations (Example 8.6.1) . . . . . . . . . . . . . . . . . . . . 372
8-2 State relaxations (Example 8.6.1) . . . . . . . . . . . . . . . . . . . . 372

9-1 State relaxations (Example 9.4.1) . . . . . . . . . . . . . . . . . . . . 382
9-2 Convex polyhedral enclosure of the reachable set (Example 9.4.1) . . 383

10-1 Objective function (Example 10.4.1) . . . . . . . . . . . . . . . . . . 404
10-2 A representative convex relaxation of the objective (Example 10.4.1) . 404
10-3 Intervals generated by branch-and-bound (Example 10.4.1) . . . . . . 407
10-4 Intervals generated by branch-and-bound (Example 10.4.1) . . . . . . 408
10-5 Experimental data and optimal trajectory (Example 10.4.2) . . . . . 410
10-6 Branch-and-bound convergence (Example 10.4.2) . . . . . . . . . . . 412
10-7 Intervals generated by branch-and-bound (Example 10.4.2) . . . . . . 412
10-8 Intervals generated by branch-and-bound (Example 10.4.2) . . . . . . 413
10-9 State bounding and relaxation failures (Example 10.4.2) . . . . . . . 413
10-10Intervals fathomed before the onset of clustering (Example 10.4.2) . . 414
10-11Intervals fathomed before the onset of clustering (Example 10.4.2) . . 415
10-12Branch-and-bound convergence (Example 10.4.2) . . . . . . . . . . . 415

# List of Tables

4.1 Rate coefficients and activation energies for Example 4.5.3 . . . . . . 214

6.1 Computational complexity of DAE state bounding methods . . . . . 282

6.2 Definition of algorithm statistics for DAE state bounding methods . . 285

6.3 State bounding algorithm statistics for Example 6.7.1 . . . . . . . . . . 285

6.4 State bounding algorithm statistics for Example 6.7.2 . . . . . . . . . . 289

7.1 Factorization and computation of $f_1$, $u_1$ and $o_1$ (Example 7.7.1) . . . 334

7.2 Factorization and computation of $f_2$, $u_2$ and $o_2$ (Example 7.7.1) . . . 334

8.1 Factorization and computation of $\psi$, $u_\psi$ and $o_\psi$ (Example 8.3.2) . . . 359

8.2 Factorization and computation of $f$, $u_f$ and $o_f$ (Example 8.4.1) . . . 368

10.1 Tolerances for Algorithm 3 . . . . . . . . . . . . . . . . . . . . . . . . . 402

10.2 Shorthand definitions used in Tables 10.3 and 10.4 . . . . . . . . . . . . 405

10.3 Optimization results for Example 10.4.1 . . . . . . . . . . . . . . . . . . 405

10.4 Optimization results for Example 10.4.2 . . . . . . . . . . . . . . . . . . 411

# Chapter 1

# Introduction

## 1.1 Overview

Systems of differential-algebraic equations (DAEs) are used throughout the engineering disciplines and hard sciences to model an incredible variety of dynamic phenomena [96, 12]. DAEs include both systems of algebraic equations and systems of ordinary differential equations (ODEs) as special cases, and are commonly used to approximate partial differential equations (PDEs) through a number of numerical schemes. Consequently, DAEs provide an extremely flexible modeling framework, underlying the continuum theories of classical physics as well as complex engineering models of mechanical, electrical, aeronautical and chemical systems [27, 86, 134, 155].

Of particular interest here is that DAEs have undoubtedly become the modeling framework of choice in the chemical process industries, where they provide detailed dynamic descriptions of everything from chemical reactors and separation units to entire chemical plants [134, 18]. In large part, this is due to the advent of powerful numerical solution techniques [96, 12, 82, 175]. However, numerical simulation alone is rarely enough to solve engineering problems of practical interest. In addition to model solutions, modern dynamic simulators typically provide parametric sensitivities, which describe the behavior of the solution in response to perturbations in the model inputs [59, 114]. This technology allows one to analyze the behavior of model systems with respect to various operating conditions, control actions, disturbances and uncertainties in the model itself. Combined with numerical optimization techniques, this further enables one to search efficiently among a range of permissible process inputs for those that optimize a desired objective. Based on these capabilities, numerical simulation and optimization of detailed DAE models has become a cornerstone of many core engineering practices, including model development, process development, economic optimization, control system design, and safety analysis [94].

In this thesis, a number of advanced techniques are developed for analyzing and optimizing processes described by systems of differential-algebraic equations. The core contributions address two related problems:

1. Computing a guaranteed enclosure of all possible solutions that can result from a given range of inputs under a given DAE model,
2. Solving optimization problems constrained by DAEs to guaranteed global optimality.

Similar to parametric sensitivity analysis, the first problem above concerns the behavior of DAE models with respect to variations in the model inputs. However, sensitivity analysis provides information that is only locally valid. That is, the parametric sensitivities describe the variation of the model solution with respect to infinitesimal perturbations in the inputs. In contrast, by enclosing all possible solutions corresponding to a given range of inputs, the methods developed here provide *global* information. This analysis has a wealth of applications, including quite direct applications in uncertainty analysis and safety verification.

However, like sensitivity analysis, these techniques are much more useful when combined with optimization procedures to search among the permissible inputs for those that are optimal with respect to some desired objective. Given the local nature of the available information, standard optimization methods based on sensitivity analysis provide solutions to such problems that are at best optimal with respect to infinitesimal perturbations. However, with the ability to compute guaranteed enclosures of all model solutions, it is possible to solve optimization problems constrained by differential-algebraic models with a guarantee that the resulting solution is optimal among all permissible alternatives; i.e. it is globally optimal. Thus, we present a deterministic algorithm for Problem 2 above as an application of the techniques developed for Problem 1.

In the following two sections, Problems 1 and 2 above are described in more detail. Given the flexibility of DAEs as a modeling framework, the range of motivating application areas is truly vast. We review only those that are most closely related to chemical engineering, typically arising in parameter estimation and chemical process design and control. We also give some fairly informal mathematical problem statements and summarize the contributions of this thesis in the context of existing approaches.

## 1.2 Enclosing the Reachable Set

Consider the generic system of differential-algebraic equations

$$\mathbf{f}(t, \mathbf{u}(t), \mathbf{x}(t), \dot{\mathbf{x}}(t)) = \mathbf{0}, \quad \forall t \in [t, t_f], \quad (\mathbf{x}(t_0), \dot{\mathbf{x}}(t_0)) = \boldsymbol{\sigma}_0. \tag{1.1}$$

The independent variable is $t$, which will be referred to as time for convenience. The initial time is denoted by $t_0$ and the initial condition, which is assumed to be consistent, is denoted by $\boldsymbol{\sigma}_0$. The solution is denoted by $\mathbf{x}$, and its time-derivative by $\dot{\mathbf{x}}$. Finally, $\mathbf{u}$ is a potentially time-varying input to the system. Given a set of permissible (consistent) initial conditions $\Sigma_0$ and a set of permissible input functions $\mathcal{U}$, the *reachable set* of (1.1) at time $t$ is the set

$$\mathcal{R}(t) \equiv \{\mathbf{x}(t) : \mathbf{x} \text{ satisfies (1.1) on } [t_0, t] \text{ with some } (\mathbf{u}, \boldsymbol{\sigma}_0) \in \mathcal{U} \times \Sigma_0\}. \tag{1.2}$$

In words, $\mathcal{R}(t)$ is the set of points that can be reached at time $t$ by a solution of (1.1) corresponding to some permissible choice of the initial condition and input. Of course, Problem 1 in §1.1 is exactly the problem of computing an enclosure of $\mathcal{R}(t)$, for every $t$ in some time-horizon of interest.

Consider for example that $\mathbf{x}$ is a vector of concentrations of the chemical species present in a reactor, and that the DAE model (1.1) describes the time evolution of these concentrations as the reaction proceeds. Depending on the problem at hand, we may consider a variety of quantities as inputs $\mathbf{u}$, including control inputs, disturbances, or uncertain model parameters. The reachable set $\mathcal{R}(t)$ contains all possible compositions in the reactor that can be achieved at time $t$ by operating the reactor from an initial state $\boldsymbol{\sigma}_0 \in \Sigma_0$ and with a permissible input $\mathbf{u} \in \mathcal{U}$. This set is interesting because, for any number of reasons, some compositions will be less desirable than others. It may happen that, in some region of composition space, the reacting mixture becomes hazardous, or catalyst fouling is accelerated. In such cases, it is extremely useful to have some means of ensuring that such regions cannot be reached by the system dynamics provided that, for example, control actions are limited to a certain safe set, or that the true model parameters do not deviate by more than a certain amount from their estimates. Of course, these are questions concerning the reachable set, and can be answered, at least in the affirmative, by computing a guaranteed enclosure of it. In particular, if an enclosure of $\mathcal{R}(t)$ does not intersect the undesirable region of composition space, then it is guaranteed that no point of $\mathcal{R}(t)$ is in the undesirable region either.

### 1.2.1 Motivation

The study of reachable sets is intimately related to the theory of optimal control and has been of general mathematical interest in this context for decades [22]. In modern control, the computation of approximations or enclosures of reachable sets is an active area of research and finds quite extensive application. Such computations have been used, for example, for state estimation from online measurements in chemical and biological processes [138, 88, 71, 141], feedback controller synthesis [110, 132, 13], robust model predictive control [102, 139], and fault detection for chemical processes [106]. Reachable sets are also used for the formal verification of control systems [99, 40, 20, 41, 10, 176] and the related problem of formal safety verification [85]. In [120], the connection between reachable sets and the solutions of dynamic pursuit-evasion games is explored with application to aircraft collision avoidance. Finally, reachable sets are also closely related to so-called invariance and viability domains for dynamic systems, both of which find similarly broad applications in control [28, 13].

In many applications, one is simply concerned with understanding how uncertainties in a process or a model will effect its output. Real world models nearly always have significant uncertainty, at least in the model parameters if not in the structure of the model itself. A particular example of interest comes from models of chemical reaction kinetics, where the rate parameters are often only known to within an order of magnitude or worse [163]. This is particularly true of models of biological systems [152, 124]. Even if the model itself is known very accurately, there may be significant uncertainties in the process inputs in the from of disturbances, measurement errors in closed-loop systems [31, 154], or highly variable resource availability and consumer demand [179]. If the model in question is nonlinear, the effects of such uncertainty on the model solution can be extremely difficult to infer. However, reachable set enclosures provide a natural means to propagate uncertainty through dynamic models, and have been applied in this context for uncertain chemical kinetics models [163, 156], compartment models [76], ecology models [105, 75], and biological systems [71, 141]. Moreover, such a description of model uncertainty is naturally useful in parameter estimation and model discrimination problems [163, 93, 103].

### 1.2.2 Existing Approaches

Given the broad importance of reachable sets, it is not surprising that a huge variety of methods have been developed for computing approximations or enclosures of them. However, we are not aware of any methods capable of computing a guaranteed enclosure of the reachable set of the general DAEs (1.1). The vast majority of work in this area applies instead to the system of explicit ODEs

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{u}(t), \mathbf{x}(t)), \quad \forall t \in [t_0, t_f], \quad \mathbf{x}(t_0) = \mathbf{x}_0, \tag{1.3}$$

with permissible initial conditions $X_0$ and reachable set

$$\mathcal{R}(t) \equiv \{\mathbf{x}(t) : \mathbf{x} \text{ satisfies (1.3) on } [t_0, t] \text{ with some } (\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0\}. \tag{1.4}$$

For such systems, the reachable set can be characterized exactly through two classes of methods. In the first, the reachable set is characterized as the subzero level set of a so-called value function, which is the solution of a partial differential equation known as the Hamilton-Jacobi-Bellman equation [98, 120]. An extension of such methods to semi-explicit index-one DAEs with no input $\mathbf{u}$ has been proposed in [45]. In general, the Hamilton-Jacobi-Bellman equations are very difficult to solve, making the numerical methods resulting from this approach computationally intensive [176, 120]. The second class of methods describes the reachable set as the solution of a differential inclusion or a related integral-funnel equation [98, 133, 13]. Again, for nonlinear systems these characterizations do not generally result in computationally tractable methods. Moreover, both of these approaches are designed to provide an accurate approximation of the reachable set, rather than a guaranteed enclosure of it, which makes them inappropriate for some important applications, such as formal safety verification.

Very general enclosures of the reachable set of (1.3) can be characterized by the solutions of differential inclusions using viability theory [13]. However, practical computational techniques arising from such characterizations typically involve computing interval bounds on the reachable set. In this case, viability conditions reduce to componentwise differential inequalities, which are discussed further below and used extensively in the methods developed in this thesis. Some further general methods using Lyapunov theory and a variant of Pontryagin's minimum principle are described in [68], though no general computational methods are provided.

Considerably more methods are available for the case where the time-varying inputs $\mathbf{u}$ in (1.3) are replaced by time-invariant parameters $\mathbf{p}$ to give the parametric

ODEs

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{p}, \mathbf{x}(t)), \quad \forall t \in [t_0, t_f], \quad \mathbf{x}(t_0) = \mathbf{x}_0, \tag{1.5}$$

with the permissible set of parameter values $P \subset \mathbb{R}^{n_p}$ and reachable set

$$\mathcal{R}(t) \equiv \{\mathbf{x}(t) : \mathbf{x} \text{ satisfies (1.5) on } [t_0, t] \text{ with some } (\mathbf{p}, \mathbf{x}_0) \in P \times X_0\}. \tag{1.6}$$

Even in this case, enclosing the reachable set is a very difficult problem when $\mathbf{f}$ is nonlinear. In [39, 41], a convex polyhedral enclosure is constructed by computing supporting hyperplanes to $\mathcal{R}(t)$. For each hyperplane, one specifies the desired normal and then computes an appropriate intercept by solving an optimization problem constrained by (1.5). However, for nonlinear systems, nonconvexity of the reachable set leads to nonconvex optimization problems which must be solved to guaranteed global optimality. This makes the method prohibitively expensive compared to other approaches. In Chapter 9, we provide a variant of this method in which the required optimization problems are guaranteed to be convex.

A large body of work in the control literature considers the reachable set of (1.5) under further simplifications, typically addressing linear ODEs or discrete time models. The earliest contributions in this area apply to hybrid discrete-continuous models with very simple continuous dynamics (i.e. simple integrators), and are essentially extensions of methods for purely discrete systems originating in computer science [6, 10]. Subsequently, methods were developed for continuous dynamics described by linear ODEs. In this case, tractable methods are available using geometric programming [184], ellipsoidal bounding techniques [97, 99], and polyhedral bounding techniques [10, 72] (some of these apply to the linear version of (1.3)). Many of these methods have been extended to treat nonlinear ODEs by constructing local approximations of the ODEs by simpler (e.g. linear) dynamics on a partition of the state space and rigorously bounding the approximation error [79, 10, 72, 5]. In such methods, the computed bounds can be quite conservative and are improved only by refining this partition. In [80], it is reported that this technique is not only very costly, but presents serious numerical problems as well. The same work is also an early example of the application of interval bounding techniques based on interval analysis [125], and advocates such techniques based on their ability to handle nonlinearity much more flexibly.

Interval bounding techniques compute a time-varying interval enclosure of the reachable set, and typically apply to systems of the form (1.5). These methods are nearly as old as interval arithmetic itself, with the earliest method appearing in [125]. Subsequently, a large body of literature has emerged on this topic. One class of interval methods, the Taylor methods [130, 129], use Taylor series expansions and various interval techniques to approximate the ODE solutions and rigorously bound the approximation error. These methods are unique among bounding methods in that they produce *validated* enclosures, meaning that the enclosures are guaranteed even when computed on a finite precision machine. Indeed, the original application of these methods was for computing validated bounds on the error introduced through numerical integration of ODEs without parametric uncertainty. For certain applications involving unstable or oscillatory systems, the consideration of numerical error can be very important. However, when applying these methods to bound the reachable sets of parametric ODEs, it is often a minor consideration. This is in part because the models of interest are typically dissipative, tending towards a stable steady-state over relatively short integration times. Moreover, when the parametric uncertainty in the model is large, its effect on the resulting bounds is much more significant than the effect of numerical error.

Some Taylor methods can be implemented very efficiently. However, when applied to ODEs with significant parametric uncertainties, such methods tend to produce extremely conservative enclosures of the reachable set. This conservatism can be greatly mitigated by using high-order Taylor expansions, or by using more sophisticated inclusion algebras, such as Taylor model arithmetic [24, 113], in place of interval arithmetic [24, 105]. Unfortunately, these measures dramatically increase the computational cost, which in the latter case scales exponentially in the dimensions of $\mathbf{p}$ and $\mathbf{x}_0$ and the order of the Taylor model. A Taylor method that applies to the control system (1.3) has recently been proposed in [89], and Taylor methods that apply to implicit systems of parametric ODEs [83] and systems of parametric index-one DAEs [142] have been proposed, though with some defficiencies discussed in Chapter 5.

A second class of interval bounding methods are based on differential inequalities [182] and use interval arithmetic to derive an auxiliary system of ODEs describing bounding trajectories [75, 162, 156, 140, 141]. The primary advantage of differential inequalities approaches is that they can be implemented using interval arithmetic and state-of-the-art numerical integration codes, yielding bounds at a cost comparable to a single simulation of the original model (order $10^{-3}$–$10^{-2}$s for systems with few states). The resulting enclosures are mathematically valid, but do not account for numerical error in their computation. Given the accuracy of modern numerical integration codes, this is not problematic for stable systems. Moreover, this issue can be overcome using a slightly more involved hybrid formulation as in [140]. Like Taylor methods, differential inequalities approaches are typically applied to parametric ODEs. However, the same methods apply directly to the control system (1.3). This observation has been made by several authors [75, 93] and is proven here in Chapter 3.

As with Taylor methods, it is known that differential inequalities methods generally yield extremely conservative enclosures of the reachable set. For these methods, the problem is related to certain monotonicity properties of the ODE right-hand sides; the problematic systems are those that are not *quasi-monotone* [182] (or *cooperative* [165]). In [162], it was shown that this condition is frequently violated in applications. On the other hand, it was also shown that it is often possible, through physical arguments, to obtain a crude set $G$ which is independently known to contain the reachable set, and that greatly improved bounds can be computed by leveraging this information. A practical implementation was developed for the case where $G$ is an interval.

### 1.2.3 Contributions

In this thesis, Chapters 3, 4, 5, 6 and 9 are devoted to computing enclosures of the reachable sets of ODEs and DAEs. Chapter 3 considers the computation of interval bounds on the reachable set of the nonlinear control system (1.3) using a differential inequalities approach. As mentioned above, such techniques are very flexible and very efficient, but often suffer from large conservatism in the resulting bounds. Chapter 3 presents a number of results that characterize interval bounds through much weaker conditions than those required by the standard differential inequalities approach. These conditions are useful for applications in which one has some physical information concerning the possible solutions of (1.3), which is very common in practice. In particular, these conditions are used to derive improved interval bounding methods that make very effective use of general physical insights in order to compute much sharper enclosures without sacrificing the efficiency of the standard differential inequalities method. In Chapter 4, these methods are specialized to ODE models of a particular form that arise in chemical reaction kinetics. It is shown that a wealth of useful physical information can be obtained automatically for such systems, resulting in an efficient method for computing very sharp enclosures of the reachable sets.

In Chapters 5 and 6, two interval bounding methods are developed that apply to systems of semi-explicit index-one DAEs of the form

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{f}(t, \mathbf{u}(t), \mathbf{x}(t), \mathbf{y}(t)), \\ \mathbf{0} &= \mathbf{g}(t, \mathbf{u}(t), \mathbf{x}(t), \mathbf{y}(t)). \end{aligned} \tag{1.7}$$

Chapter 5 contains a number of theoretical contributions, including a computational test for existence and uniqueness of a DAE solution, and several extensions of differential inequalities results for (1.3) to the case of semi-explicit index-one DAEs. Chapter 6 presents two efficient numerical methods for computing an interval enclosure of the reachable set of (1.7) based on these developments. To the authors knowledge, these methods are the first differential inequalities based bounding techniques applicable to DAE models.

Finally, in Chapter 9, a method is presented for computing convex polyhedral enclosures of the reachable sets of both (1.5) and (1.7). This is done by a modification of the method in [39, 41]. Though a convex polyhedral set can potentially provide a much sharper enclosure of the reachable set than can an interval, the method proposed in [39, 41] requires solving several nonconvex dynamic optimization problems to global optimality, which is extremely computationally expensive. Using methods for global dynamic optimization problems (see §1.3) developed in Chapters 7 and 8, the method of Chapter 9 produces convex polyhedral enclosures of the reachable set by solving only convex dynamic optimization problems.

## 1.3 Global Dynamic Optimization

Consider the general optimization problem constrained by DAEs:

**Problem 1.3.1.**

$$\inf_{\mathbf{x},\mathbf{u},\boldsymbol{\sigma}_0} \quad \phi(\mathbf{u}(t_f),\mathbf{x}(t_f)) + \int_{t_0}^{t_f} \psi(s,\mathbf{u}(s),\mathbf{x}(s))ds \tag{1.8}$$

$$\text{s.t.} \quad \mathbf{h}(\mathbf{u}(t_f),\mathbf{x}(t_f)) + \int_{t_0}^{t_f} \boldsymbol{\ell}(s,\mathbf{u}(s),\mathbf{x}(s))ds \leq \mathbf{0} \tag{1.9}$$

$$\mathbf{f}(t,\mathbf{u}(t),\mathbf{x}(t),\dot{\mathbf{x}}(t)) = \mathbf{0}, \quad \forall t \in [t_0,t_f], \quad (\mathbf{x}(t_0),\dot{\mathbf{x}}(t_0)) = \boldsymbol{\sigma}_0 \tag{1.10}$$

$$\mathbf{u} \in \mathcal{U}, \quad \mathbf{x} \in \mathcal{X}, \quad \boldsymbol{\sigma}_0 \in \Sigma_0. \tag{1.11}$$

As in §1.2, the set $\mathcal{U}$ is the set of permissible input functions, referred to as controls in this context, and $\Sigma_0$ is the set of permissible initial conditions. The independent variable $t$ takes values in the interval $[t_0, t_f]$, and the set $\mathcal{X}$ is a subset of a suitable space of functions $\mathbf{x} : [t_0, t_f] \to \mathbb{R}^{n_x}$ containing putative solutions of the DAEs (1.10). A crucial feature of Problem 1.3 is that the decision variables $\mathbf{x}$ and $\mathbf{u}$ are functions. In other words, this is an optimization problem on a function space, and is therefore an infinite-dimensional problem. In general, optimization problems constrained by systems of differential equations such as ODEs, DAEs, or PDEs are referred to as *dynamic optimization problems* or *optimal control problems* [22, 173].

As an example of a problem of this type, consider finding the time-varying temperature profile for a batch reactor that maximizes the yield of a desired product. In this case, the scalar-valued control function $u$ represents the temperature, the DAEs (1.10) represent a dynamic model of the state of the reactor, $\mathbf{x}(t)$, including the temporal profiles of the concentrations of the various reacting species, and the objective function (1.8) is specified as the negative of the yield at the final time $t_f$, so that it is minimized when the yield is maximized. The set $\mathcal{U}$ may restrict the permissible temperature profiles based on a number of considerations, such as the requirement that the temperature never exceeds a threshold value, or the requirement that the temperature is not varied too abruptly, so that it can be practically implemented by a controller. Similarly, the constraints (1.9) may represent any number of considerations, such as purity specifications or safety requirements. Then, in words, Problem 1.3.1 is to find the initial condition, the temperature profile, and the time-varying state of the reactor that maximizes the yield at $t_f$ among all alternatives which obey the reactor model and satisfy the given constraints.

The work in this thesis concerns algorithms for solving optimal control problems to guaranteed global optimality. In particular, we present such an algorithm for Problem 1.3.1 in the case where (1.10) is a system of semi-explicit index-one DAEs and the controls $\mathbf{u}$ are approximated by a finite number of real parameters. This approximation is termed *control parameterization* [173] and is a very common in modern numerical methods (see §1.3.2). Moreover, we show that some of the crucial steps in the proposed algorithm are valid even without this approximation, at least in the case where (1.10) is replaced by the explicit system of ODEs (1.3).

### 1.3.1 Motivation

Given the flexibility of DAEs as a modeling framework, it is evident that a great variety of problems can be posed as optimal control problems subject to DAEs. In the chemical process industry, dynamic optimization techniques are routinely used to locate optimal process designs, operating conditions and control actions. For example, open-loop control of batch processes can be formulated as a dynamic optimization problem and has been widely studied in this context, particularly with application to high-value added industries such as specialty chemicals, pharmaceuticals and bioprocessing [111, 144, 34, 25, 167, 31]. Dynamic optimization problems also arise when considering processes with periodic dynamic behavior, such as pressure swing adsorption and simulated moving bed chromatography [95, 51]. Even for processes that are nominally operated at steady-state, several important problems require dynamic optimization, including the determination of optimal start-up and shut-down procedures [17], optimal policies for changeover from one product to another [61], and optimal catalyst blending in tubular reactors [108]. Another area in which dynamic optimization is essential is in process safety verification [1, 48, 85, 106].

A more fundamental application is the problem of estimating unknown parameters in a dynamic model from a given set of data [29, 103, 54, 163, 124]. Here, the model parameters are the decision variables, and the optimization algorithm finds those parameters which minimize the deviation of the model prediction from the measured data. This problem is extremely important, for example, for the determination of chemical reaction mechanisms from kinetic data [163, 124]. As a final illustration of the broad applicability of dynamic optimization problems, we note applications in the diverse areas of biological network design [2, 166] and optimal drug scheduling for chemotherapy [116, 35].

As discussed in the following section, most available algorithms for solving dynamic optimization problems search only for locally optimal solutions. Such algorithms can only guarantee global optimality under restrictive convexity assumptions which are often violated in practical applications. For example, it has been shown that optimization problems resulting from control parameterization of Problem 1.3.1 are nearly always nonconvex, especially for problems arising in chemical engineering [108, 109, 16, 124]. Nonetheless, there is strong impetus to compute global solutions stemming from numerous applications. One need only consider the problem of maximizing the profitability of a process. Clearly, a significant economic penalty may be incurred by designing and operating such a process according to a locally optimal solution [160]. However, other applications pose more serious problems. In parameter estimation problems, one is often interested in determining whether a model, equipped with its best fit parameter estimates, is consistent with measured data according to a statistical significance test. However, if only locally optimal parameter estimates are available, any conclusions drawn from such an analysis are dubious [163, 121]. For process safety verification, it is desirable to identify the worst-case behavior of a dynamic system over a range of inputs in order to determine whether the system will remain within some safe operating region. This too can be formulated as a dynamic optimization problem, but again a locally optimal solution will not suffice because it may not necessarily describe the worst-case scenario, potentially leading to a false conclusion of safe operation with dire consequences [85]. Finally, optimization problems are often used to solve energy minimization problems in order to describe the fundamental properties of a system, such as its equilibrium state. In these applications, the value of a locally optimal solution simply does not provide the desired information [100, 121].

### 1.3.2 Existing Approaches for Local Dynamic Optimization

The primary challenge in dynamic optimization is the fact that the optimization is performed over an infinite-dimensional space, i.e., the space of the control functions $\mathbf{u}$ and state functions $\mathbf{x}$. At the broadest level, solution methods for optimal control problems can be classified in terms of how the problem of infinite-dimensionality is addressed. Most modern numerical methods are based on some form of discretization, with the primary aim of reducing the problem to a finite-dimensional one and applying methods developed for optimization problems on Euclidean spaces. Methods of this type are referred to as *direct* methods. However, effectively using discretization techniques requires modern computers, and direct methods have therefore only become popular in recent decades. Historically, methods for optimal control problems have addressed the problem directly in the infinite-dimensional space. Such methods are based on satisfying necessary conditions of optimality for optimal control problems [22, 81, 87] and are commonly referred to as *indirect* methods[1]. This approach has its origins in the classical calculus of variations [47, 81], which deals more generally with optimization problems on function spaces and dates back to the late sixteen hundreds. By analogy to the standard gradient-based necessary conditions of optimality for optimization problems on Euclidean spaces, the calculus of variations provides necessary conditions of optimality for many classes of optimal control problems. The resulting conditions are the Euler-Lagrange equations, which take the form of a two-point boundary value problem [47, 81, 177]. In modern optimal control theory, the Euler-Lagrange equations are generalized by Pontryagin's maximum principle [177, 77]. Methods based on either of these formulations require repeated solution of boundary value problems with estimates of the optimal control that are iteratively refined by a number of methods [33, 38].

There are several serious drawbacks to these approaches. First, it is difficult to derive appropriate necessary conditions in the presence of certain types of constraints that arise in applications. Furthermore, even when such conditions are available, generating the corresponding boundary value problem computationally requires derivative or adjoint information, which can be costly to obtain for large systems. Second, in all but the simplest cases the resulting boundary value problems do not permit analytical solutions and are known to be extremely difficult to solve numerically. The reasons for these numerical problems are quite serious and include instability of the boundary value problem and issues related to the differential index of the system, especially in the presence of so-called singular arcs in the optimal control. Third, the vast majority of work on these methods does not directly apply to systems of DAEs, but rather to explicit systems of ODEs of the form (1.3). Finally, these methods are based on necessary conditions characterizing locally optimal solutions, and only become sufficient for global optimality under restrictive convexity assumptions that are violated or very difficult to verify in applications [177, 159].

[1] The seemingly backward designations *direct* and *indirect* do not refer to the space in which the optimization is carried out. In general, a direct optimization method is one which produces a feasible sequence of decisions with monotonically decreasing objective value, while an indirect method is one which is based on satisfying necessary conditions of optimality. In this regard, labelling all discretization-based methods as direct is somewhat misleading but nonetheless very common.

A related approach which also considers the optimal control problem in the original infinite-dimensional space is the dynamic programming approach, based on Bellman's principle of optimality [19]. This principle leads to necessary and sufficient conditions of optimality through the solution of a boundary value problem in PDEs known as the Hamilton-Jacobi-Bellman (HJB) equations. In the formulation of this PDE, the state variables $\mathbf{x}$ are treated as independent variables, making the HJB equations impractically difficult to solve for large systems. On the whole, this approach does not result in practical numerical methods outside of some very particular applications [173].

As mentioned above, direct methods for optimal control use discretization techniques in order to approximate the optimal control problem by a nonlinear program (NLP) on a finite-dimensional Euclidean space. These methods can be further classified in terms of the level of discretization used in this approximation. In the simultaneous approach, both the state and control functions are discretized, either by finite differencing, collocation, or more general basis set expansions, with collocation being the most common [178, 54, 46, 42]. This provides a representation of the state and control functions in terms of finitely many real parameters, so that the resulting optimization problem is a standard NLP on a Euclidean space with a large system of equality constraints approximating the original DAEs. The benefit of this approximation procedure is that it enables one to apply standard methods in nonlinear programming. On the other hand, the simultaneous approach produces very large-scale NLPs, so that in practice specialized algorithms are required [26].

In contrast, the sequential approach (also called control parameterization [173, 32]) considers only discretization of the control functions. The controls are approximated by an expansion in terms of a finite set of basis functions, resulting in, for example, piece-wise constant, piece-wise affine, or polynomial controls. With this approximation, the controls can be represented in terms of a finite number of real parameters, $\mathbf{p} \in \mathbb{R}^{n_p}$, so that $\mathbf{u}$ is now regarded as a known function of $t$ and $\mathbf{p}$. If, for every admissible parameter vector $\mathbf{p}$, the differential-algebraic system has a unique solution, then this approximation reduces the search space to a finite-dimensional space.

Applying control parameterization to the Problem 1.3.1 gives the following program, where $P \subset \mathbb{R}^{n_p}$ is the set of admissible values for $\mathbf{p}$.

**Problem 1.3.2.**

$$\inf_{\mathbf{p},\boldsymbol{\sigma}_0} \quad \phi(\mathbf{u}(t_f,\mathbf{p}),\mathbf{x}(t_f,\mathbf{p},\boldsymbol{\sigma}_0)) + \int_{t_0}^{t_f} \psi(s,\mathbf{u}(s,\mathbf{p}),\mathbf{x}(s,\mathbf{p},\boldsymbol{\sigma}_0))ds \tag{1.12}$$

$$\text{s.t.} \quad \mathbf{h}(\mathbf{u}(t_f,\mathbf{p}),\mathbf{x}(t_f,\mathbf{p},\boldsymbol{\sigma}_0)) + \int_{t_0}^{t_f} \boldsymbol{\ell}(s,\mathbf{u}(s,\mathbf{p}),\mathbf{x}(s,\mathbf{p},\boldsymbol{\sigma}_0))ds \leq \mathbf{0} \tag{1.13}$$

$$\mathbf{p} \in P, \quad \boldsymbol{\sigma}_0 \in \Sigma_0, \tag{1.14}$$

where, for every $(\mathbf{p},\boldsymbol{\sigma}_0) \in P \times \Sigma_0$, $\mathbf{x}(\cdot,\mathbf{p},\boldsymbol{\sigma}_0)$ is the unique solution of

$$\mathbf{f}(t,\mathbf{u}(t,\mathbf{p}),\mathbf{x}(t,\mathbf{p},\boldsymbol{\sigma}_0),\dot{\mathbf{x}}(t,\mathbf{p},\boldsymbol{\sigma}_0)) = \mathbf{0}, \quad \forall t \in [t_0,t_f], \tag{1.15}$$

$$(\mathbf{x}(t_0,\mathbf{p},\boldsymbol{\sigma}_0),\dot{\mathbf{x}}(t_0,\mathbf{p},\boldsymbol{\sigma}_0)) = \boldsymbol{\sigma}_0.$$

Note that the objective and constraint functions above are not known explicitly as functions of the decision variables. However, they are well-defined as such and can be evaluated numerically via numerical solution of the embedded DAEs (1.15). Due to the availability of robust dynamic simulation software [96, 12, 82, 175], one can find local optima for large-scale dynamic optimization problems quite effectively with the sequential approach.

There has been much discussion in the literature concerning the advantages and disadvantages of the simultaneous and sequential approaches. As compared to the simultaneous approach, the sequential approach has the drawback that every evaluation of the objective and constraints, along with their derivatives, requires numerical integration and sensitivity analysis of the embedded DAE system. Another drawback is that the sequential approach may fail if the embedded DAEs have unstable modes for some feasible choice of $\mathbf{p}$ and $\boldsymbol{\sigma}_0$. On the other hand, the simultaneous approach requires the solution of very large-scale NLPs, while the sequential approach does not. Moreover, accurate discretization of the states is often problematic in the simultaneous approach, as is the need to provide the NLP solver with accurate initial guesses for the discretized state variables. In the sequential approach, discretization of the state trajectories is handled internally by a dynamic simulation code using very mature adaptive procedures that have proven to be accurate, efficient and reliable. Moreover, there is no need to provide an initial guess for the state trajectory. For most problems, there is no unified consensus on which of these methods should be used. However, the discussion has led to an interesting compromise known as multiple-shooting, which is particularly advantageous for unstable systems and embedded boundary value problems [101].

### 1.3.3 Global Optimization of Standard NLPs

As with many local optimization techniques, the existing methods for solving dynamic optimization problems to global optimality can be viewed as an application of established methods for optimization on Euclidean spaces to the NLPs resulting from either the simultaneous or sequential approach. Before discussing these methods, it is helpful to review some basic concepts from global optimization on Euclidean spaces, in particular, the spatial-branch-and-bound algorithm. Both here and in the discussion of dynamic optimization problems in the next section, we restrict our attention to so-called *deterministic* global optimization algorithms. This excludes the class of stochastic search algorithms, including simulated annealing, genetic algorithms, tabu search, particle swarm optimization, harmony search, ant-colony algorithms, etc. [65, 52, 124]. While these algorithms are designed to find global minima for problems with multiple suboptimal local minima, these approaches are ad hoc. They not only fail to provide a guarantee that a global solution will be found, they are incapable of verifying optimality in case such a point has been found. In contrast, our interest here is in algorithms that are guaranteed to furnish a globally optimal solution after finitely many iterations.

Consider the standard NLP

$$\begin{aligned} \min_{\mathbf{p} \in P} \quad & J(\mathbf{p}) \\ \text{s.t.} \quad & \mathbf{G}(\mathbf{p}) \leq \mathbf{0}. \end{aligned} \tag{1.16}$$

where $P \subset \mathbb{R}^{n_p}$ is an $n_p$-dimensional compact interval and $J$ and $\mathbf{G}$ are continuous on $P$. To solve this problem to global optimality, the spatial branch-and-bound (B&B) method considers a sequence of subproblems in which (1.16) is restricted to a subinterval $P^l \subset P$:

$$\begin{aligned} \min_{\mathbf{p} \in P^l} \quad & J(\mathbf{p}) \\ \text{s.t.} \quad & \mathbf{G}(\mathbf{p}) \leq \mathbf{0}, \end{aligned} \tag{1.17}$$

The basic requirement for applying spatial B&B is that, for any subinterval $P^l \subset P$ (which may be $P$ itself), procedures are available that compute guaranteed upper and lower bounds on the optimal objective value of (1.17). These bounds are denoted by $UBD^l$ and $LBD^l$, respectively. Since the value of the objective function at any feasible point provides an upper bound on the optimal objective value of (1.17), $UBD^l$ can be computed by solving (1.16) to local optimality. Computing a lower bound is substantially more difficult and is the key step in the spatial B&B algorithm. Methods for accomplishing this are discussed below.

Supposing that upper and lower bounding procedures are available, the spatial B&B algorithm procedes as follows. First, upper and lower bounds are computed for the optimal objective value of (1.16). Since these bounds apply to the original problem of interest, rather than to the subproblem (1.17), they are denoted by $UBD$ and $LBD$, respectively. If it happens that $UBD - LBD$ is less than a specified tolerance $\varepsilon$, then the B&B algorithm terminates, having bracketed the optimal objective value of (1.16) within the given tolerance. An estimate of the solution value $\mathbf{p}^*$ is then given by the value which attained the upper bound $UBD$. If this termination test fails, then $P$ is partitioned into two subintervals, termed *branching*, typically by bisection in its dimension of largest width. These subintervals inherit the bounds $UBD$ and $LBD$, which are obviously valid for the corresponding subproblems (1.17) on account of being valid for (1.16). These two subintervals are then added to a stack $\Sigma$ of subintervals, or *nodes*, to be processed that is maintained throughout the algorithm.

At the beginning of a generic iteration of the algorithm, $UBD$ and $LBD$ are the best known upper and lower bounds on the optimal objective value of (1.16), respectively, and the stack $\Sigma$ contains a number of nodes $P^l$, each of which is equipped with upper and lower bounds $UBD^l$ and $LBD^l$ that have been inherited from the parent node from which it was generated through bisection. Collectively, the nodes $P^l$ may not form a partition of $P$, but the complement of $\cup_l P^l$ in $P$ will have been proven not to contain the optimal solution of (1.16) through the procedures below. The iteration proceeds by selecting from the stack a node $P^l$ for which $LBD^l = LBD$. The upper and lower bounds $UBD^l$ and $LBD^l$ are then refined by computing bounds on the optimal objective value of (1.17) using the procedures that we have assumed to be available. If it is found that (1.17) is infeasible, then $P^l$ is eliminated from further consideration and a new element is selected from the stack. In this case, we say that $P^l$ is *fathomed by infeasibility.* Otherwise, upper and lower bounds on the optimal objective value of the original problem (1.16) are updated according to

$$UBD := \min_k UBD^k \quad \text{and} \quad LBD := \min_k LBD^k, \tag{1.18}$$

where the min is taken over all elements of $\Sigma$. These assignments are valid because the complement of $\cup_k P^k$ in $P$ has been shown not to contain a global optimum of (1.16). Moreover, if $P^l$ was the only element of $\Sigma$ for which $LBD^l = LBD$ at the beginning of the iteration, and if $LBD^l$ was improved by the application of the lower bounding procedure to (1.17), then $LBD$ is improved by this assignment. If $UBD$ is improved by this assignment, then there is an opportunity to fathom some nodes in the stack. This is done by checking the inequality $LBD^k > UBD$ for every $P^k \in \Sigma$. If this is true for some $P^k$, then the optimal solution cannot lie in $P^k$ and $P^k$ is eliminated from further consideration. In this case, $P^k$ is said to be *fathomed by value dominance.*

If $P^l$ has not been fathomed either by infeasibility or by value dominance, then it is bisected and the two resulting nodes are added to the stack.

The iteration outlined above is repeated until either the stack becomes empty, indicating that (1.16) is infeasible, or it is found in some iteration that $UBD - LBD < \varepsilon$, indicating that a point $\mathbf{p}^*$ has been found which achieves an objective value within $\varepsilon$ of the globally optimal objective value. Roughly, if the lower bounding procedure has the property that it provides sharper bounds on smaller intervals $P^l$ and becomes exact in the limit as $P^l$ tends toward a singleton, then it can be shown that one of these outcomes will occur after finitely many iterations [84]. Due to the repeated partitioning of $P$, the spatial B&B algorithm exhibits worst-case exponential run-time with respect to the dimension of $\mathbf{p}$ and the magnitude of $1/\varepsilon$. In practice, the primary determinants of the run-time are the computational cost and the accuracy of the lower bounding procedure. In addition, a number of more advanced techniques have been developed which can greatly accelerate convergence through the use of constraint propagation techniques [147, 148, 149]. Thus, while it is true that the basic procedure outlined above can be prohibitively expensive, impressive results have been achieved for many challenging problems using advanced implementations of the method [146, 147, 171, 160].

Several methods are available for computing lower bounds on the optimal objective value of the subproblem (1.17). A simple approach is to compute interval bounds on the image of $P^l$ under $J$ using interval arithmetic [125]. Though many early implementations are based on this approach [90], the lower bounds computed in this way are relatively weak. Moreover, these bounds obey a first-order convergence rate property [125], while it has been demonstrated that at least second-order convergence is required to avoid serious convergence problems in spatial B&B algorithms [50]. In most modern implementations, lower bounds are computed by constructing and solving convex underestimating programs [118, 7, 57, 171]. Though there are many ways to accomplish this, a popular and illustrative approach is to construct a convex function $J^{cv} : P^l \to \mathbb{R}$ which underestimates $J$ on $P^l$, and a (componentwise) convex function $\mathbf{G}^{cv} : P^l \to \mathbb{R}^{n_c}$ which underestimates $\mathbf{G}$ on $P^l$. Such functions are termed *convex relaxations* of $J$ and $\mathbf{G}$ on $P^l$, respectively. A convex underestimating program is then given by

$$\begin{aligned} \min_{\mathbf{p}\in P^l} \quad & J^{cv}(\mathbf{p}) \\ \text{s.t.} \quad & \mathbf{G}^{cv}(\mathbf{p}) \le \mathbf{0}. \end{aligned} \tag{1.19}$$

In particular, this program is convex, and hence solvable to global optimality using standard local optimization techniques, and its optimal objective value is easily seen to underestimate that of (1.17).

There are several methods for constructing a convex relaxation of a function. Floudas et al. have constructed convex relaxations for twice differentiable functions by adding a sufficiently large quadratic term to the original function. This is accomplished by shifting the diagonal elements of the Hessian matrix by a parameter $\alpha$ [7]. Values of $\alpha$ which guarantee convexity of the resulting function can be found via interval arithmetic [4, 3]. It has recently been shown that $\alpha$BB relaxations have a second-order convergence rate [30].

Another approach due to McCormick [118] provides a method for computing convex relaxations of so-called *factorable functions* (this technique is presented in detail in Chapter 2). Roughly, a function is said to be factorable if it can be defined by the recursive application of basic operations including binary addition, binary multiplication, and composition with a library of simple univariate functions. In particular, any function that can be written explicitly in computer code is factorable. Given such a function, McCormick's technique constructs relaxations by the recursive application of relaxation rules for each of the basic operations defining the function. McCormick's technique is easily implemented using the operator overloading capabilities of object-oriented programming languages, and tends to produce much tighter relaxations than those produced by the $\alpha$BB method, particularly on wide intervals $P^l$. Moreover, it is shown in [30] that McCormick's relaxations also have second-order convergence subject to some implementation details. On the other hand, they are generally nonsmooth, which makes solving (1.17) more difficult. This difficulty has been addressed in [122], which provides rules for efficiently computing subgradients for McCormick's relaxations which can then be used by nonsmooth solvers such as bundle methods.

A related technique that generates a convex underestimating program for (1.16) in the case where $J$ and $\mathbf{G}$ are factorable is described in [171] and used in the popular code `BARON`. This technique uses a recursive procedure similar to that of McCormick, which in this case substitutes the result of each basic operation defining the objective and constraint functions with a dummy variable subject to one or more linear constraints. This procedure does not result in program of the form (1.19), but rather produces a linear program in a higher-dimensional space whose optimal objective value is guaranteed to underestimate that of (1.17). This method has the advantage that the underestimating program is linear and can therefore be solved more efficiently and reliably than the nonlinear convex underestimating programs derived from $\alpha$BB or McCormick's relaxation technique. On the other hand, these underestimating programs have many more variables than the original problem. The convergence rate of this method is unknown, but its successful implementation in `BARON` suggests that it is likely second-order.

A key feature of all of these methods, which has significant consequences for global dynamic optimization, is that the objective function and constraints in (1.16) must be factorable. That is, these functions must be given by explicit algebraic expressions. Of course, this is notably not the case for the NLP (1.3.2) derived by control parameterization of (1.3.1).

### 1.3.4 Existing Approaches for Global Dynamic Optimization

All of the available methods for deterministic global optimization of nonconvex dynamic optimization problems are extensions of the direct methods discussed in §1.3.2, using variants of the spatial branch-and-bound algorithm of the previous section. Obtaining a guarantee of global optimality from an indirect method is problematic for several reasons. First, these methods are intimately related to necessary conditions of optimality, which do not distinguish between locally and globally optimal solutions. However, this fact alone is not insurmountable. In Chapter 11, we shown that some of the key ideas used for solving dynamic optimization problems to global optimality using the sequential approach can actually be applied directly to dynamic optimization problems in the original infinite-dimensional space. Specifically, we show that it is possible to construct convex underestimating programs in this space. However, a much more serious problem is that there is no known method for exhaustively partitioning an infinite-dimensional space, which precludes the use of the spatial B&B framework.

In the case of the simultaneous approach, the extension to global optimization is apparent. Since total discretization of the infinite-dimensional problem results in a standard NLP on a finite-dimensional Euclidean space, the spatial branch-and-bound algorithm can be applied directly using standard methods for the lower bounding procedure. However, given the size of the NLPs generated through the simultaneous approach and the worst-case exponential run-time of the spatial B&B algorithm, this cannot be considered a practical approach to global dynamic optimization. Nonetheless, it has been attempted in the articles [55, 42]. In [55], comparisons show that the simultaneous global optimization approach is badly outperformed by an early method based on the sequential approach. In both articles, it is clear that an adequate discretization of the state variables creates problems which are too large to be solved in reasonable time by a global optimization routine, and coarser discretizations can not represent the original dynamics well enough to produce reliable results (the optimal objective value was found to depend strongly on the discretization).

As discussed in §1.3.2, the sequential approach avoids the dramatic increase in problem size characteristic of the simultaneous approach. Moreover, it reduces the dynamic optimization problem to a standard NLP on a Euclidean space, so that in principle the spatial B&B method can be applied. However, the objective and constraint functions in the resulting program (Problem 1.3.2) are not known explicitly, but rather are defined implicitly through the solution of the embedded dynamic system. That is, in order to write Problem (1.3.2) as the standard NLP

$$\begin{aligned} \min_{\mathbf{p}\in P,\ \boldsymbol{\sigma}_0\in\Sigma_0} \quad & J(\mathbf{p},\boldsymbol{\sigma}_0) \\ \text{s.t.} \quad & \mathbf{G}(\mathbf{p},\boldsymbol{\sigma}_0) \leq \mathbf{0}, \end{aligned} \tag{1.20}$$

we must make the definitions

$$J(\mathbf{p},\boldsymbol{\sigma}_0) \equiv \phi(\mathbf{u}(t_f,\mathbf{p}),\mathbf{x}(t_f,\mathbf{p},\boldsymbol{\sigma}_0)) + \int_{t_0}^{t_f} \psi(s,\mathbf{u}(s,\mathbf{p}),\mathbf{x}(s,\mathbf{p},\boldsymbol{\sigma}_0))ds, \tag{1.21}$$

$$\mathbf{G}(\mathbf{p},\boldsymbol{\sigma}_0) \equiv \mathbf{h}(\mathbf{u}(t_f,\mathbf{p}),\mathbf{x}(t_f,\mathbf{p},\boldsymbol{\sigma}_0)) + \int_{t_0}^{t_f} \boldsymbol{\ell}(s,\mathbf{u}(s,\mathbf{p}),\mathbf{x}(s,\mathbf{p},\boldsymbol{\sigma}_0))ds, \tag{1.22}$$

where $\mathbf{x}$ is the solutions of the embedded DAEs (1.15). As discussed in the previous section, this precludes the use of standard lower bounding procedures.

The first method for overcoming this problem was proposed by Esposito and Floudas in [54], where convex relaxations of the functions $J$ and $\mathbf{G}$ are computed by a dynamic extension of the $\alpha$BB method known as $\beta$BB. Recall that the $\alpha$BB method computes a convex relaxation of a given function by adding a sufficiently large quadratic term, where the required magnitude $\alpha$ of this term is inferred by analysis of the Hessian matrix. The key idea here is that the Hessian matrix of $J$, for example, can be evaluated by solving the second-order sensitivity system for the embedded DAEs. However, without an explicit functional form for the Hessian, $\alpha$ cannot be computed through the standard approach and is instead approximated via a finite sampling procedure. This not only makes constructing these relaxations very inefficient, but also precludes any guarantee that the relaxation is indeed convex.

A method for computing a valid $\alpha$ was later proposed by Papamichail and Adjiman [135], resulting in the first practical global dynamic optimization algorithm. Notably, this method applies only in the case where the embedded dynamic system is an explicit system of ODEs. In fact, this is true of every existing global dynamic optimization algorithm, excluding those based on the simultaneous approach. The method described in [135] uses results from differential inequalities (see §1.2.2) in order to bound the solutions of the embedded system of differential equations, as well as the second-order sensitivities. This yields an interval Hessian matrix which can be used to compute a value for $\alpha$ that ensures convexity. Though this approach is rigorous, the convex relaxations generated in this way tend to be extremely weak, likely due to a very conservative bound on the required $\alpha$ value.

A different approach, which is also applicable to dynamic optimization problems with explicit ODEs embedded, was proposed by Singer and Barton in [161, 164]. Using the recursive nature of certain relaxation techniques (McCormick's technique and the methods in [171]), it was shown that a convex underestimating program for (1.20) can be constructed given only a method for computing (componentwise) convex and concave relaxations of $\mathbf{x}(t, \cdot, \cdot)$ on $P \times \Sigma$, for all $t \in [t_0, t_f]$ (a concave relaxation is a concave function that *overestimates* the function of interest). This idea was first used in order to solve dynamic optimization problems involving a class of linear time-varying ODEs whose solutions are known to be affine, and hence both convex and concave, with respect to the decision variables [161]. The approach was then extended to problems with nonlinear ODEs embedded in [164]. In this case, a combination of McCormick's relaxation technique and differential inequalities was used to derive an auxiliary system of ODEs whose solutions are both affine in the decision variables and describe upper and lower bounds on the solution of the original ODEs [162], hence providing the required relaxations of $\mathbf{x}(t, \cdot, \cdot)$ for all $t \in [t_0, t_f]$. Computational results for this method demonstrate that the resulting lower bounding procedure requires less computational effort and provides much more accurate bounds as compared to the $\alpha$BB based method in [135].

Lin and Stadtherr have proposed a method for globally solving dynamic optimization problems with ODEs embedded which does not use convex relaxations [103, 104]. Rather, a sophisticated Taylor method (see §1.2.2) is used to compute very tight interval bounds on the solution of the embedded ODEs [105], which are then used to compute a lower bound for the optimal objective value of (1.20). Unlike lower bounding procedures based on standard interval arithmetic, this method does not suffer from slow convergence. This is because the required interval computations are done using Taylor model arithmetic, which is a much more accurate method based on high-order Taylor series expansions [113, 24]. For many test problems, solution times for this approach are substantially faster than those reported for any other deterministic algorithm. However, using Taylor Models for bounding the solution of the embedded ODEs is extremely costly. The number of Taylor coefficients that must be stored increases exponentially with the number of decision variables and the order of the Taylor expansion [74]. Hence, there is reasonable concern that methods of this type will prove to be inefficient or unusable for problems with many decisions, and/or problems for which a high-order Taylor expansion is required to capture the state dependence on the decision variables accurately.

Though no optimization results have yet been presented, two related methods for computing convex and concave relaxations of the solutions of parametric ODEs have recently been proposed by Sahlodin and Chachuat [151, 150]. These methods extend the technique in [105] for computing interval bounds on the solutions of parametric ODEs by applying McCormick's relaxation technique in place of weaker interval computations throughout the algorithm. These methods appear capable of providing very tight relaxations when a sufficiently high-order Taylor expansion is used. On the other hand, the use of high-order Taylor expansions again makes these approaches potentially very expensive for high dimensional problems, and the existence of an appropriate compromise in the context of global optimization remains an open question.

### 1.3.5 Contributions

Aside from intractable methods based on a total discretization approach, all of the available methods for global dynamic optimization apply only to problems with explicit ODEs embedded. In this thesis, we present the first method capable of solving problems with DAEs embedded. In particular, we consider the class of semi-explicit index-one DAEs of the form (1.7). Like methods for ODEs, this method is based on a spatial B&B algorithm, and the primary challenge in developing it was to derive a valid lower bounding procedure.

Following the work of Singer and Barton [164], the key ingredient in the lower bounding procedure is a method that computes convex and concave relaxations of the solutions of a parametric system of DAEs. In Chapter 7, the problem of relaxing the solutions of a dynamic system is analyzed in a general setting, resulting in two novel relaxation theories. Though these methods are ultimately applied to systems of DAEs, they can also be applied directly to explicit ODEs. In both cases, efficient numerical methods are developed using an extension of McCormick's relaxation technique developed in Chapter 2.

For systems of ODEs, the resulting relaxation methods are most closely related to the existing method of Singer and Barton [162]. The choice to pursue methods of this type was based on several considerations including their ease of use, favorable scaling and computational efficiency as compared to other competitive methods. Of the two methods developed in this thesis, the first is shown to have distinct drawbacks and is illustrative of some problems unique to relaxation methods for dynamic problems. On the other hand, the second method has very satisfactory performance and is shown to significantly outperform the method of [162] in numerical experiments. For systems of semi-explicit index-one DAEs, the relaxation methods developed here are the first available in the literature.

In Chapter 10, we present a basic spatial B&B algorithm for the deterministic global solution of dynamic optimization problems with semi-explicit index-one DAEs embedded. During the course of this thesis, the vast majority of work on global dynamic optimization has been directed at deriving relaxations for the solutions of ODEs and DAEs. Comparatively little effort has been dedicated to developing optimization algorithms to make use of them. Accordingly, the presented algorithm is basic in several respects, and analogy with global optimization techniques for standard NLPs suggests that the method should be quite computationally intensive. Though we do find the efficiency of this basic algorithm to be problematic, it is no more so here than for existing techniques for problems with explicit ODEs embedded. Hence, the algorithm successfully provides an extension of the current state-of-the-art to problems with DAEs embedded. We analyze the performace of the algorithm in the context of several case studies and take the opportunity to suggest some promising directions for future improvement, noting in particular the advanced techniques that have proven to be indispensable for practical global solution of standard NLPs.

Finally, in Chapter 11, we present the surprising result that the relaxation theory developed here can largely be applied to dynamic optimization problems in the original infinite-dimensional space. In particular, this allows one to construct convex underestimating programs for nonconvex optimal control problems, without the need to discretized either the state or the controls. Though this seems to provide a key step towards a global optimization method for nonconvex optimal control problems, a complete algorithm remains elusive because their seems no reasonable way to exhaustively partition an infinite-dimensional space.

# Chapter 2

# Factorable Functions, Interval Arithmetic and McCormick Relaxations

## 2.1 Introduction

In order to solve global optimization problems, one must have some means of inferring global information about the functions involved. In general, local characterizations of a function, such as its value or its derivative at a point, are not enough. Rather, one requires information about the behavior of the function on the entire domain of interest. An essential tool in this regard is the so-called factorable representation of a function, which will be heavily used throughout this thesis. Essentially, a function is factorable if it can be written as a finite sequence of simple operations, including basic arithmetic operations as well as intrinsic functions available on a computer, such as $\sqrt{x}$, $x^n$, $e^x$, $\sin x$, etc. For example, the function

$$f(x_1, x_2) = 10x_1 + x_1^2 e^{x_2} \tag{2.1}$$

is factorable because it can be evaluated for any $(x_1, x_2) \in \mathbb{R}^2$ by executing the following sequence of simple computations:

$$\begin{aligned}
v_1(x_1, x_2) &= x_1, \\
v_2(x_1, x_2) &= x_2, \\
v_3(x_1, x_2) &= 10v_1(x_1, x_2), \\
v_4(x_1, x_2) &= (v_1(x_1, x_2))^2, \\
v_5(x_1, x_2) &= \exp(v_2(x_1, x_2)), \\
v_6(x_1, x_2) &= v_4(x_1, x_2) \times v_5(x_1, x_2), \\
v_7(x_1, x_2) &= v_3(x_1, x_2) + v_6(x_1, x_2), \\
f(x_1, x_2) &= v_7(x_1, x_2).
\end{aligned}$$

Roughly, each of the intermediates $v_i$ is called a factor, and the factorable representation is the sequence $v_1, \ldots, v_7$. In essence, any function written explicitly in computer code will be factorable, so it is not at all restrictive to develop methods for this class of functions.

In this chapter, the class of factorable functions is defined formally, and two standard methods are introduced for obtaining useful global information about them. These methods are interval arithmetic [125] and McCormick's relaxation technique [118], which are used to compute interval enclosures and convex relaxations of factorable functions, respectively. The presentation of interval arithmetic is mostly standard, though some definitions are made more general and some new regularity results are developed. On the other hand, the analysis of McCormick's relaxation technique includes many generalizations and new results, leading to the the generalized McCormick relaxations of §2.7. As will be seen in later chapters, the application of these techniques to dynamic problems will require more out of both of these methods than do typical global optimization algorithms.

## 2.2 Factorable Functions

To formalize the notion of a factorable function, we must first define the set of operations that will be permissible in the sequence of computations defining such functions. Each element of this set will be a real-valued function on a Euclidean space. To avoid notational conflicts in later sections, it is prudent here to use the formal notation for a function as a triple $(o, B, R)$, where $B$ is the domain, $R$ is the range, and $o$ is a mapping from $B$ into $R$, $o : B \to \mathbb{R}$. Throughout this thesis, the set of permissible operations will contain the binary addition operation $(+, \mathbb{R}^2, \mathbb{R})$, and the binary multiplication operation $(\times, \mathbb{R}^2, \mathbb{R})$. In addition, it will include a *library of univariate functions*, which is a set $\mathcal{L}$ whose elements are univariate functions; $(u, B, \mathbb{R}) \in \mathcal{L}$ has $B \subset \mathbb{R}$. The elements of $\mathcal{L}$ will be used to represent functions such as $\sqrt{x}$, $x^n$, $e^x$, $\sin x$, etc. Furthermore, $\mathcal{L}$ should include the negative and reciprocal functions $-x$ and $1/x$, so that subtraction and division can be achieved by combination with $(+, \mathbb{R}^2, \mathbb{R})$ and $(\times, \mathbb{R}^2, \mathbb{R})$. In order for the class of factorable functions to be useful, it is necessary to require that certain information about each element of $\mathcal{L}$ is either known or easily computable, and that certain basic properties are satisfied. For now, it is only required that, for each $(u, B, \mathbb{R}) \in \mathcal{L}$, $u(x)$ can be evaluated computationally for any $x \in B$. Further requirements will be added throughout this chapter. For reference, they are Assumptions 2.3.8, 2.4.25, 2.5.29 2.5.39, and 2.5.33. In practice, these assumptions are not at all restrictive. The required information is readily available for a large variety of univariate functions, and all required properties can be shown to hold with only minor exceptions.

**Definition 2.2.1.** Let $n_i, n_o \in \mathbb{N}$. A *$\mathcal{L}$-computational sequence* with $n_i$ inputs and $n_o$ outputs is a pair $(\mathcal{S}, \pi_o)$:

1. $\mathcal{S}$ is a finite sequence $\{((o_k, B_k, \mathbb{R}), (\pi_k, \mathbb{R}^{k-1}, \mathbb{R}^{d_k}))\}_{k=n_i+1}^{n_f}$ with every element defined by one of the following options:

   (a) $(o_k, B_k, \mathbb{R})$ is either $(+, \mathbb{R}^2, \mathbb{R})$ or $(\times, \mathbb{R}^2, \mathbb{R})$ and $\pi_k : \mathbb{R}^{k-1} \to \mathbb{R}^2$ is defined by $\pi_k(\mathbf{v}) = (v_i, v_j)$ for some integers $i, j \in \{1, \ldots, k-1\}$.

(b) $(o_k, B_k, \mathbb{R}) \in \mathcal{L}$ and $\pi_k : \mathbb{R}^{k-1} \to \mathbb{R}$ is defined by $\pi_k(\mathbf{v}) = v_i$ for some integer $i \in \{1, \ldots, k-1\}$.

2. $\pi_o : \mathbb{R}^{n_f} \to \mathbb{R}^{n_o}$ is defined by $\pi_o(\mathbf{v}) = (v_{i(1)}, \ldots, v_{i(n_o)})$ for some integers $i(1), \ldots, i(n_o) \in \{1, \ldots, n_f\}$.

A computational sequence defines a function $\mathbf{f}_{\mathcal{S}} : D_{\mathcal{S}} \subset \mathbb{R}^{n_i} \to \mathbb{R}^{n_o}$ by the following construction.

**Definition 2.2.2.** Let $(\mathcal{S}, \pi_o)$ be a $\mathcal{L}$-computational sequence with $n_i$ inputs and $n_o$ outputs. Define the *sequence of factors* $\{(v_k, D_k, \mathbb{R})\}_{k=1}^{n_f}$, with $D_k \subset \mathbb{R}^{n_i}$, where

1. For $k = 1, \ldots, n_i$, $D_k = \mathbb{R}^{n_i}$ and $v_k(\mathbf{x}) = x_k$, $\forall \mathbf{x} \in D_k$,

2. For $k = n_i + 1, \ldots, n_f$, $D_k = \{\mathbf{x} \in D_{k-1} : \pi_k(v_1(\mathbf{x}), \ldots, v_{k-1}(\mathbf{x})) \in B_k\}$ and $v_k(\mathbf{x}) = o_k \circ \pi_k \circ (v_1(\mathbf{x}), \ldots, v_{k-1}(\mathbf{x}))$, $\forall \mathbf{x} \in D_k$.

The set $D_{\mathcal{S}} \equiv D_{n_f}$ is called the *natural domain* of $(\mathcal{S}, \pi_o)$, and the *natural function* $(\mathbf{f}_{\mathcal{S}}, D_{\mathcal{S}}, \mathbb{R}^{n_o})$ is defined by $\mathbf{f}_{\mathcal{S}}(\mathbf{x}) = \pi_o \circ (v_1(\mathbf{x}), \ldots, v_{n_f}(\mathbf{x}))$, $\forall \mathbf{x} \in D_{\mathcal{S}}$.

**Example 2.2.1.** Equation (2.1) defines a computational sequence with $n_i = 2$ inputs, $\mathbf{x} = (x_1, x_2)$, and $n_0 = 1$ output. In fact there are several computational sequences that describe this function, depending on the order in which the operations are applied. The computational sequence leading to the sequence of factors shown previously is:

$$\begin{array}{lll}
- & - & v_1 = x_1, \\
- & - & v_2 = x_2, \\
o_3 = 10\times & \pi_3(\mathbf{v}) = v_1 & v_3 = 10v_1, \\
o_4 = (\cdot)^2 & \pi_4(\mathbf{v}) = v_1 & v_4 = v_1^2, \\
o_5 = \exp & \pi_5(\mathbf{v}) = v_2 & v_5 = \exp(v_2), \\
o_6 = \times & \pi_6(\mathbf{v}) = (v_4, v_5) & v_6 = v_4 v_5, \\
o_7 = + & \pi_7(\mathbf{v}) = (v_3, v_6) & v_7 = v_3 + v_6, \\
- & \pi_o(\mathbf{v}) = v_7 & f_{\mathcal{S}} = v_7.
\end{array}$$

Since every univariate function appearing in the computational sequence above is defined on the entire real line, the natural domain is $D_{\mathcal{S}} = \mathbb{R}^2$.

**Definition 2.2.3** (Factorable function)**.** A function $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ is $\mathcal{L}$-*factorable* if there exists a $\mathcal{L}$-computational sequence $(\mathcal{S}, \pi_o)$ with $n$ inputs and $m$ outputs such that the natural function $(\mathbf{f}_{\mathcal{S}}, D_{\mathcal{S}}, \mathbb{R}^{n_o})$ satisfies $D \subset D_{\mathcal{S}}$ and $\mathbf{f} = \mathbf{f}_{\mathcal{S}}|_D$.

**Remark 2.2.4.** Again, note that the use of the term *factorable* in later chapters will imply Assumptions 2.3.8, 2.4.25, 2.5.29, 2.5.39, and 2.5.33.

## 2.3 Interval Analysis

For $a, b \in \mathbb{R}$, $a \leq b$, define the *interval* $[a, b]$ as the compact, connected set $\{x \in \mathbb{R} : a \leq x \leq b\}$. Interval analysis is the study of intervals as basic arithmetic objects on par with integers and real numbers. This concept will be extensively used to compute global information about factorable functions in the form of interval bounds on their range. In this section, the basics of interval analysis are presented, leading in particular to the concept of a *natural interval extension* of a factorable function. Definitive resources in this field are [125] and [131].

The set of all nonempty intervals is denoted $\mathbb{IR}$. Intervals are denoted by capital letters, $Z \in \mathbb{IR}$. Since $Z$ is a subset of $\mathbb{R}$, the notation $z \in Z$ is well-defined. The set of $n$-dimensional interval vectors is denoted $\mathbb{IR}^n$. In particular, $Z \in \mathbb{IR}^n$ has elements $Z_i \in \mathbb{IR}$, $i = 1, \ldots, n$. Every $Z \in \mathbb{IR}^n$ can be regarded as a subset of $\mathbb{R}^n$ defined by the Cartesian product $Z_1 \times \ldots \times Z_n$, so that $\mathbf{z} \in \mathbb{R}^n$ satisfies $\mathbf{z} \in Z$ if $z_i \in Z_i$, $i = 1, \ldots, n$. The set of $n \times m$ interval matrices is denoted $\mathbb{IR}^{n \times m}$ and defined analogously to $\mathbb{IR}^n$; $A \in \mathbb{IR}^{n \times m}$ has elements $A_{ij} \in \mathbb{IR}$, for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$, and, for any $\mathbf{A} \in \mathbb{R}^{n \times m}$ with elements $a_{ij}$, $\mathbf{A} \in A$ if $a_{ij} \in A_{ij}$ for all indices $i$ and $j$. For any $D \subset \mathbb{R}^n$, let $\mathbb{I}D$ denote the set $\{Z \in \mathbb{IR}^n : Z \subset D\}$. This notation is also used for $D \subset \mathbb{R}^{n \times m}$.

If $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ and $\mathbf{v} \leq \mathbf{w}$, then $[\mathbf{v}, \mathbf{w}]$ denotes the $n$-dimensional interval $[v_1, w_1] \times \ldots, \times [v_n, w_n]$. Moreover, for any $Z \in \mathbb{IR}$, the notation $\mathbf{z}^L, \mathbf{z}^U \in \mathbb{R}^n$ will be commonly used to denote the vectors such that $Z = [\mathbf{z}^L, \mathbf{z}^U]$. The notation $m(Z)$ denotes the *midpoint* of $Z$, $m(Z) \equiv 0.5(\mathbf{z}^L + \mathbf{z}^U)$, and $w(Z)$ denotes the *width* of $Z$, $w(Z) \equiv \mathbf{z}^U - \mathbf{z}^L$. For $A \in \mathbb{IR}^{n\times m}$, $m(A)$ and $w(A)$ are real-valued matrices defined analogously. For any $\mathbf{z} \in \mathbb{R}^n$, the singleton $[\mathbf{z}, \mathbf{z}]$ is called a *degenerate* interval.

### 2.3.1 Inclusion Functions and Interval Extensions

The central task in interval analysis is to compute an interval which encloses the range of a given function. This is the notion of an *inclusion function*, formalized below.

**Definition 2.3.1.** Let $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$, and for any $E \subset D$, let $\mathbf{f}(E)$ denote the image of $E$ under $\mathbf{f}$. A mapping $F : \mathfrak{D} \subset \mathbb{I}D \to \mathbb{IR}^m$ is an *inclusion function* for $\mathbf{f}$ on $\mathfrak{D}$ if $\mathbf{f}(X) \subset F(X)$, $\forall X \in \mathfrak{D}$.

Ideally, an inclusion function should be defined on all of $\mathbb{I}D$; i.e., an interval enclosure can be computed for the image of any $X \in \mathbb{I}D$ under $\mathbf{f}$. In practice, however, this is not always possible. This issue is discussed further after Theorem 2.3.11. Typically, inclusion functions are derived from a simpler object known as an *interval extension.*

**Definition 2.3.2.** Let $D \subset \mathbb{R}^n$. A set $\mathfrak{D} \subset \mathbb{IR}^n$ is an *interval extension of* $D$ if every $\mathbf{x} \in D$ satisfies $[\mathbf{x}, \mathbf{x}] \in \mathfrak{D}$. Let $\mathbf{f} : D \to \mathbb{R}^m$. A function $F : \mathfrak{D} \to \mathbb{IR}^m$ is an *interval extension of* $\mathbf{f}$ if $\mathfrak{D}$ is an interval extension of $D$ and, for every $\mathbf{x} \in D$, $F([\mathbf{x}, \mathbf{x}]) = [\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x})]$.

An *interval extension* will be an inclusion function if it is *inclusion monotonic.*

**Definition 2.3.3.** Let $F : \mathfrak{D} \subset \mathbb{IR}^n \to \mathbb{IR}^m$. $F$ is *inclusion monotonic* on $\mathfrak{D}$ if

$$X_1 \subset X_2 \implies F(X_1) \subset F(X_2), \quad \forall X_1, X_2 \in \mathfrak{D}. \tag{2.2}$$

**Theorem 2.3.4.** *Let* $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ *and let* $F : \mathfrak{D} \to \mathbb{IR}^m$ *be an interval extension of* $\mathbf{f}$. *If* $F$ *is inclusion monotonic on* $\mathfrak{D} \cap \mathbb{I}D$, *then* $F$ *is an inclusion function for* $\mathbf{f}$ *on* $\mathfrak{D} \cap \mathbb{I}D$.

*Proof.* Choose any $X \in \mathfrak{D} \cap \mathbb{I}D$ and any $\mathbf{x} \in X$. Since $\mathbf{x} \in D$, it follows that $[\mathbf{x}, \mathbf{x}] \in \mathfrak{D}$ and $\mathbf{f}(\mathbf{x}) \in [\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x})] = F([\mathbf{x}, \mathbf{x}]) \subset F(X)$. □

The following result is useful for constructing inclusion functions for complex functions from those of simpler functions.

**Lemma 2.3.5.** *Let $\mathbf{f}_1 : D_1 \subset \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbf{f}_2 : D_2 \subset \mathbb{R}^m \to \mathbb{R}^k$, and define $D_{12} \equiv \{\mathbf{x} \in D_1 : \mathbf{f}_1(\mathbf{x}) \in D_2\}$. Let $F_1 : \mathfrak{D}_1 \to \mathbb{IR}^m$ and $F_2 : \mathfrak{D}_2 \to \mathbb{IR}^k$ be interval extensions of $\mathbf{f}_1$ and $\mathbf{f}_2$, respectively. Then $\mathfrak{D}_{12} \equiv \{X \in \mathfrak{D}_1 : F_1(X) \in \mathfrak{D}_2\}$ is an interval extension of $D_{12}$, and $(F_2 \circ F_1, \mathfrak{D}_{12}, \mathbb{IR}^k)$ is an interval extension of $(\mathbf{f}_2 \circ \mathbf{f}_1, D_{12}, \mathbb{R}^k)$. If $F_1$ and $F_2$ are inclusion monotonic on $\mathfrak{D}_1$ and $\mathfrak{D}_2$, respectively, then $F_2 \circ F_1$ is inclusion monotonic on $\mathfrak{D}_{12}$.*

*Proof.* First it is shown that $\mathbf{x} \in D_{12}$ implies $[\mathbf{x}, \mathbf{x}] \in \mathfrak{D}_{12}$. For any $\mathbf{x} \in D_{12}$, $\mathbf{x} \in D_1$ implies that $[\mathbf{x}, \mathbf{x}] \in \mathfrak{D}_1$ and $F_1([\mathbf{x}, \mathbf{x}]) = [\mathbf{f}_1(\mathbf{x}), \mathbf{f}_1(\mathbf{x})]$. Then $\mathbf{f}_1(\mathbf{x}) \in D_2$ implies that $F_1([\mathbf{x}, \mathbf{x}]) \in \mathfrak{D}_2$, so that $[\mathbf{x}, \mathbf{x}] \in \mathfrak{D}_{12}$.

To show that $(F_2 \circ F_1, \mathfrak{D}_{12}, \mathbb{IR}^k)$ is an interval extension of $(\mathbf{f}_2 \circ \mathbf{f}_1, D_{12}, \mathbb{R}^k)$, choose any $\mathbf{x} \in D_{12}$. Since $\mathfrak{D}_{12}$ is an interval extension of $D_{12}$, $[\mathbf{x}, \mathbf{x}] \in \mathfrak{D}_{12}$. Then, $F_2(F_1([\mathbf{x}, \mathbf{x}])) = F_2([\mathbf{f}_1(\mathbf{x}), \mathbf{f}_1(\mathbf{x})]) = [\mathbf{f}_2(\mathbf{f}_1(\mathbf{x})), \mathbf{f}_2(\mathbf{f}_1(\mathbf{x}))]$.

It remains to show that $F_2 \circ F_1$ is inclusion monotonic on $\mathfrak{D}_{12}$. Choose any $X, \hat{X} \in \mathfrak{D}_{12}$ such that $X \subset \hat{X}$. Then $F_1(X) \subset F_1(\hat{X})$, and both intervals are in $\mathfrak{D}_2$, so that $F_2(F_1(X)) \subset F_2(F_1(\hat{X}))$. □

## 2.3.2 Interval Arithmetic and the Natural Interval Extension

Just as one adds, multiplies and performs other simple operations on real numbers, these operations are defined for elements of $\mathbb{IR}$ as well. The basic property of these interval operations is that they are inclusion functions for the corresponding real operation. Using this system, termed *interval arithmetic*, one can compute an inclusion function for any factorable function in a very natural way.

**Definition 2.3.6.** Define $(+, \mathbb{IR}^2, \mathbb{IR})$ and $(\times, \mathbb{IR}^2, \mathbb{IR})$ by

$$
\begin{aligned}
&+(X, Y) = X + Y = [x^L + y^L, x^U + y^U], \\
&\times(X, Y) = XY = [\min(x^L y^L, x^L y^U, x^U y^L, x^U y^U), \max(x^L y^L, x^L y^U, x^U y^L, x^U y^U)].
\end{aligned}
$$

**Theorem 2.3.7.** $(+, \mathbb{IR}^2, \mathbb{IR})$ *and* $(\times, \mathbb{IR}^2, \mathbb{IR})$ *are interval extensions of* $(+, \mathbb{R}^2, \mathbb{R})$ *and* $(\times, \mathbb{R}^2, \mathbb{R})$*, respectively, and they are inclusion monotonic on* $\mathbb{IR}^2$*.*

*Proof.* Clearly, $\mathbb{IR}^2$ is an interval extension of $\mathbb{R}^2$. For any $x \in X$ and $y \in Y$, $[x, x] + [y, y] = [x + y, x + y]$ and $[x, x][y, y] = [\min(xy, xy, xy, xy), \max(xy, xy, xy, xy)] = [xy, xy]$. For inclusion monotonicity, see [125], §3.3. □

Of course, the previous theorem implies that $(+, \mathbb{IR}^2, \mathbb{IR})$ and $(\times, \mathbb{IR}^2, \mathbb{IR})$ are inclusion functions for $(+, \mathbb{R}^2, \mathbb{R})$ and $(\times, \mathbb{R}^2, \mathbb{R})$ on $\mathbb{IR}^2$. In particular, for any $X, Y \in \mathbb{IR}^n$, we have $x + y \in X + Y$ and $xy \in XY$, for all $x \in X$ and $y \in Y$. Furthermore, Lemma 2.3.5 implies that these functions may be composed to conclude, for example, that $x + xy \in X + XY$, for all $x \in X$ and $y \in Y$. Our aim is to extend this recursion to arbitrary $\mathcal{L}$-computational sequences. However, the ability to do so depends on $\mathcal{L}$. In particular, it requires the following.

**Assumption 2.3.8.** For every $(u, B, \mathbb{R}) \in \mathcal{L}$, an interval extension $(u, \mathbb{I}B, \mathbb{IR})$ is known and can be evaluated computationally. Furthermore, this interval extension is inclusion monotonic on $\mathbb{I}B$.

**Remark 2.3.9.** In the assumption above, the notation $u$ is used to denote both the original univariate function and its interval extension. The ambiguity in this convention is removed by specifying the domain and codomain of the function, which is the purpose of using the triplet notation for functions throughout this chapter. Overloading the notation $u$ in this manner has the advantage that we may write, for example, $\exp(X)$ for some $X \in \mathbb{IR}$ directly, without defining additional notation for the interval extension of the exponential.

Interval extensions for a wide variety of univariate functions are compiled in §2.8. Note that there is no need to define interval subtraction and division explicitly, since these operations can be achieved by combining addition and multiplication with univariate negative and reciprocal functions.

Suppose Assumption 2.3.8 holds and $(\mathcal{S}, \pi_o)$ is a $\mathcal{L}$-computational sequence. Then, to any element $((o_k, B_k, \mathbb{R}), (\pi_k, \mathbb{R}^{k-1}, \mathbb{R}^{d_k}))$ of $\mathcal{S}$, there corresponds an inclusion monotonic interval extension $(o_k, \mathbb{I}B_k, \mathbb{IR})$. Further, the functions $(\pi_k, \mathbb{IR}^{k-1}, \mathbb{IR}^2)$ (or $(\pi_k, \mathbb{IR}^{k-1}, \mathbb{IR})$) may be defined in the natural way, so that for example $\pi_k(V) = (V_i, V_j)$ if $\pi_k(\mathbf{v}) = (v_i, v_j)$. Then, the natural interval extension of $(\mathcal{S}, \pi_o)$ is defined as follows.

**Definition 2.3.10.** For every $\mathcal{L}$-computational sequence $(\mathcal{S}, \pi_o)$, with $n_i$ inputs and $n_o$ outputs, define the *sequence of inclusion factors* $\{(V_k, \mathfrak{D}_k, \mathbb{IR})\}_{k=1}^{n_f}$ where

1. For all $k = 1, \ldots, n_i$, $\mathfrak{D}_k = \mathbb{IR}^{n_i}$ and $V_k(X) = X_k$, $\forall X \in \mathfrak{D}_k$,
2. For all $k = n_i + 1, \ldots, n_f$, $\mathfrak{D}_k = \{X \in \mathfrak{D}_{k-1} : \pi_k \circ (V_1(X), \ldots, V_{k-1}(X)) \in \mathbb{I}B_k\}$ and $V_k(X) = o_k \circ \pi_k \circ (V_1(X), \ldots, V_{k-1}(X))$, $\forall X \in \mathfrak{D}_k$.

The *natural interval extension* of $(\mathcal{S}, \pi_o)$ is the function $(F_{\mathcal{S}}, \mathfrak{D}_{\mathcal{S}}, \mathbb{IR}^{n_o})$ defined by $\mathfrak{D}_{\mathcal{S}} \equiv \mathfrak{D}_{n_f}$ and $F_{\mathcal{S}}(X) = \pi_o \circ (V_1(X), \ldots, V_{n_f}(X))$, $\forall X \in \mathfrak{D}_{\mathcal{S}}$.

**Theorem 2.3.11.** *Let $(\mathcal{S}, \pi_o)$ be a $\mathcal{L}$-computational sequence with natural function $(\mathbf{f}_{\mathcal{S}}, D_{\mathcal{S}}, \mathbb{R}^{n_o})$. The natural interval extension $(F_{\mathcal{S}}, \mathfrak{D}_{\mathcal{S}}, \mathbb{IR}^{n_o})$ is an interval extension of $(\mathbf{f}_{\mathcal{S}}, D_{\mathcal{S}}, \mathbb{R}^{n_o})$, and is inclusion monotonic on $\mathfrak{D}_{\mathcal{S}}$.*

*Proof.* Consider the sequence of factors $\{(v_k, D_k, \mathbb{R})\}_{k=1}^{n_f}$ and the sequence of inclusion factors $\{(V_k, \mathfrak{D}_k, \mathbb{IR})\}_{k=1}^{n_f}$. Choose any $K \in \{1, \ldots, n_f\}$ and suppose that $(V_k, \mathfrak{D}_k, \mathbb{IR})$ is an interval extension of $(v_k, D_k, \mathbb{R})$, and inclusion monotonic on $\mathfrak{D}_k$, for all $k \in \{1, \ldots, K-1\}$. If $K \leq n_i + 1$, this is true because, for any $k < K$, $\mathfrak{D}_k = \mathbb{IR}^{n_i}$ is an interval extension of $D_k = \mathbb{R}^{n_i}$, $V_k([\mathbf{x}, \mathbf{x}]) = [x_k, x_k] = [v_k(\mathbf{x}), v_k(\mathbf{x})]$, and $V_k$ is trivially inclusion monotonic on $\mathbb{IR}^{n_i}$.

Now, $(v_1, \ldots, v_{K-1})$ is a well-defined mapping from $D_{K-1}$ into $\mathbb{R}^{K-1}$. By the inductive hypothesis, $(V_1, \ldots, V_{K-1})$, as a mapping from $\mathfrak{D}_{K-1}$ into $\mathbb{IR}^{K-1}$, is an interval extension of $(v_1, \ldots, v_{K-1})$, and is inclusion monotonic on $\mathfrak{D}_{K-1}$. It follows that $\pi_k \circ (V_1, \ldots, V_{K-1})$ is an interval extension of $\pi_k \circ (v_1, \ldots, v_{K-1})$, and is inclusion monotonic on $\mathfrak{D}_{K-1}$. By Theorem 2.3.7 and Assumption 2.3.8, $(o_K, \mathbb{I}B_K, \mathbb{IR})$ is an interval extension of $(o_K, B_K, \mathbb{R})$, and is inclusion monotonic on $\mathbb{I}B_K$. Then, Lemma 2.3.5 shows that $(V_K, \mathfrak{D}_K, \mathbb{IR})$ is an interval extension of $(v_K, D_K, \mathbb{R})$, and it is inclusion monotonic on $\mathfrak{D}_K$. By induction, this holds for every $K \in \{1, \ldots, n_f\}$, and the theorem follows from the definition of $(F_{\mathcal{S}}, \mathfrak{D}_{\mathcal{S}}, \mathbb{IR}^{n_o})$. □

Ideally, $\mathfrak{D}_{\mathcal{S}}$ should contain all of $\mathbb{I}D_{\mathcal{S}}$, so that for any $X \in \mathbb{I}D_{\mathcal{S}}$ the natural interval extension provides an interval enclosure of the image of $X$ under $\mathbf{f}$. From Definition 2.3.10, it is clear that $\mathfrak{D}_{\mathcal{S}}$ will only fail to be the whole of $\mathbb{I}D_{\mathcal{S}}$ if, for some $X \in \mathbb{I}D_{\mathcal{S}}$, a domain violation occurs when evaluating the interval extension of some univariate function in the computational sequence. Even though $\mathbf{f}_{\mathcal{S}}$ is well-defined on $D_{\mathcal{S}}$, this is possible because the value of an inclusion factor $V_k(X)$ may overestimate the image of $X$ under the corresponding factor $v_k$. However, Definition 2.3.10 and inclusion monotonicity of the inclusion factors immediately imply the following useful property.

**Lemma 2.3.12.** *Let $(\mathcal{S}, \pi_o)$ be a $\mathcal{L}$-computational sequence with natural interval extension $(F_{\mathcal{S}}, \mathfrak{D}_{\mathcal{S}}, \mathbb{IR}^{n_o})$. For any $X \in \mathbb{IR}^{n_i}$, $X \in \mathfrak{D}_{\mathcal{S}}$ implies that $\mathbb{I}X \subset \mathfrak{D}_{\mathcal{S}}$.*

**Definition 2.3.13.** Let $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ be a $\mathcal{L}$-factorable function. Then, for any $\mathcal{L}$-computational sequence describing $\mathbf{f}$, the natural interval extension $(F_{\mathcal{S}}, \mathfrak{D}_{\mathcal{S}}, \mathbb{IR}^m)$ is called a natural interval extension of $\mathbf{f}$.

It is apparent from Theorem 2.3.11 that a natural interval extension of a $\mathcal{L}$-factorable function is indeed an interval extension, and is inclusion monotonic on $\mathfrak{D}_{\mathcal{S}}$. More importantly, it is an inclusion function for $\mathbf{f}$ on $\mathfrak{D}_{\mathcal{S}} \cap \mathbb{I}D$. Moving forward, the notation $([\mathbf{f}], \mathfrak{D}, \mathbb{IR}^m)$ will be used to denote a natural interval extension of $(\mathbf{f}, D, \mathbb{R}^m)$.

**Example 2.3.1.** Consider again the function (2.1), and the computational sequence discussed in Example 2.2.1. Interval extensions of all of the univariate functions involved in this sequence are known and in fact quite intuitive. Consider computing an enclosure of the range of this function on the interval $X_1 \times X_2 = [-1, 3] \times [.4, 1]$. To do this using the natural interval extension, the sequence of inclusion factors is evaluated as follows:

$$\begin{aligned}
V_1(X_1, X_2) &= X_1 = [-1, 3],\\
V_2(X_1, X_2) &= X_2 = [.4, 1],\\
V_3(X_1, X_2) &= 10V_1(X_1, X_2) = 10[-1, 3] = [-10, 30],\\
V_4(X_1, X_2) &= (V_1(X_1, X_2))^2 = [-1, 3]^2 = [0, 9],\\
V_5(X_1, X_2) &= \exp(V_2(X_1, X_2)) = \exp([.4, 1]) = [\exp(.4), e],\\
V_6(X_1, X_2) &= V_4(X_1, X_2) \times V_5(X_1, X_2) = [0, 9][\exp(.4), e] = [0, 9e],\\
V_7(X_1, X_2) &= V_3(X_1, X_2) + V_6(X_1, X_2) = [-10, 30] + [0, 9e] = [-10, 30 + 9e],\\
F(X_1, X_2) &= V_7(X_1, X_2) = [-10, 54.5].
\end{aligned}$$

By Theorem 2.3.1 and Lemma 2.3.5, it is now guaranteed that the value of (2.1) lies in interval $[-10, 54.5]$, for any $(x_1, x_2) \in [-1, 3] \times [.4, 1]$.

From the previous example, it should be clear that computing natural interval extensions is easily automatable, and hardly more computationally demanding than executing the same sequence of computations in real arithmetic. Many libraries are available for computing interval extensions automatically using the operator overloading functionality of object oriented programming languages such as `C++` (Profil: `http://www.ti3.tu-harburg.de/keil/profil/index_e.html`; Boost: `http://www.boost.org/doc/libs/1_37_0/libs/numeric/interval/doc/interval.htm`).

The price that one pays for the efficiency and simplicity of interval arithmetic is that it often provides very conservative enclosures. Essentially, this is because the procedure is memoryless; the interval addition defining $V_7$ in example 2.3.1 takes no account of the fact that both $V_3$ and $V_6$ depend on $X_1$, so that $v_3$ and $v_6$ may not vary within $V_3$ and $V_6$ independently. This well-known shortcoming is termed the *dependency problem.* On the other hand, the interval arithmetic operations, and hence natural interval extensions under appropriate assumptions on $\mathcal{L}$, have the property that the computed interval bound becomes less conservative as the input interval is decreased in width [125].

## 2.4 McCormick Analysis

In this section, we begin the development of McCormick's relaxation technique, which is completed in Sections 2.6 and 2.7. McCormick's technique provides a means to compute convex and concave relaxations of $\mathcal{L}$-factorable functions. Let $D \subset \mathbb{R}^n$ be convex. A vector function $\mathbf{g} : D \to \mathbb{R}^m$ is called convex if each component is convex; i.e.,

$$\mathbf{g}(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \leq \lambda\mathbf{g}(\mathbf{x}_1) + (1-\lambda)\mathbf{g}(\mathbf{x}_2), \quad \forall(\lambda, \mathbf{x}_1, \mathbf{x}_2) \in [0,1] \times D \times D,$$

and it is called concave if the opposite (weak) inequality holds.

**Definition 2.4.1.** Let $D$ be a convex set in $\mathbb{R}^n$ and $\mathbf{f} : D \to \mathbb{R}^m$. A function $\mathbf{f}^{cv} : D \to \mathbb{R}^m$ is a *convex relaxation*, or *convex underestimator*, of $\mathbf{f}$ on $D$ if $\mathbf{f}^{cv}$ is convex on $D$ and $\mathbf{f}^{cv}(\mathbf{x}) \leq \mathbf{f}(\mathbf{x})$, $\forall\mathbf{x} \in D$. Similarly, a function $\mathbf{f}^{cc} : D \to \mathbb{R}^m$ is a *concave relaxation*, or *concave overestimator*, of $\mathbf{f}$ on $D$ if $\mathbf{f}^{cc}$ is concave on $D$ and $\mathbf{f}^{cc}(\mathbf{x}) \geq \mathbf{f}(\mathbf{x})$, $\forall\mathbf{x} \in D$.

Suppose $\mathbf{f} : D \to \mathbb{R}$ is $\mathcal{L}$-factorable with the $\mathcal{L}$-computational sequence $(\mathcal{S}, \pi_o)$, $\mathcal{S} = \{((u_k, B_k, \mathbb{R}), (\pi_k, \mathbb{R}^{k-1}, \mathbb{R}^{d_k}))\}_{k=n_i+1}^{n_f}$, and the sequence of factors $\{(v_k, D_k, \mathbb{R})\}_{k=1}^{n_f}$. McCormick's relaxation technique can be thought of as computing a *natural relaxation* similar to the natural interval extension of the previous section. When evaluating the natural interval extension of $\mathbf{f}$ on $X$, the interval $X$ is taken as input and an interval $V_k(X)$ is computed for each factor $v_k$ sequentially. In particular, this is done by interval versions of each operation $o_k$ taking intervals as inputs and returning intervals as outputs. Thus, in the evaluation of the interval extension, the basic unit of information passed from one operation to the next is the interval. In contrast, McCormick's procedure takes an interval $X$ and a point $\mathbf{x} \in X$ as input, and associates to each factor an interval $V_k(X)$ and two additional numbers $v^{cv}(X, \mathbf{x})$ and $v^{cc}(X, \mathbf{x})$. The interpretation of the interval is the same; it encloses the image of $X$ under $v_k$. The numbers $v^{cv}(X, \mathbf{x})$ and $v^{cc}(X, \mathbf{x})$, respectively, represent the values of convex and concave relaxations of $v_k$ on $X$ evaluated at $\mathbf{x}$. In effect, what is required for Mc-

Cormick's procedure is that each operation $o_k$ can be replaced with a mapping that takes an interval, as well as two relaxation values, as input, and returns the same as output. Thus, there is a direct analogy between McCormick's relaxation technique and interval arithmetic. However, the basic unit of information is more complex. It is the element of the set

$$\mathbb{MR}^n \equiv \{(Z^B, Z^C) \in \mathbb{IR}^n \times \mathbb{IR}^n : Z^B \cap Z^C \neq \emptyset\}. \tag{2.3}$$

Elements of $\mathbb{MR}^n$ are denoted by script capitals, $\mathscr{Z} \in \mathbb{MR}^n$. For any such $\mathscr{Z}$, the notations $Z^B, Z^C \in \mathbb{IR}^n$ and $(\mathbf{z}^L, \mathbf{z}^U, \mathbf{z}^{cv}, \mathbf{z}^{cc}) \in \mathbb{R}^n$ will commonly be used to denote the intervals and vectors satisfying $\mathscr{Z} = (Z^B, Z^C) = ([\mathbf{z}^L, \mathbf{z}^U], [\mathbf{z}^{cv}, \mathbf{z}^{cc}])$.

The representation of the relaxation values $\mathbf{z}^{cv}$ and $\mathbf{z}^{cc}$ as an interval of course imposes the basic requirement that $\mathbf{z}^{cv} \leq \mathbf{z}^{cc}$, which is natural given the interpretation above. The further requirement that $Z^B \cap Z^C$ be nonempty is also natural since the interval bounds $Z^B$ and the relaxation values $Z^C$ are intended to bound the same value. Some desirable properties of McCormick's relaxation procedure will further require that one works with objects for which $\mathbf{z}^{cv}, \mathbf{z}^{cc} \in Z^B$. Therefore, we make the following definitions.

**Definition 2.4.2.** $\mathscr{Z} \in \mathbb{MR}^n$ is called *proper* if $Z^C \subset Z^B$. The set of all proper elements of $\mathbb{MR}^n$ is denoted $\mathbb{MR}^n_{\text{prop}}$.

**Definition 2.4.3.** The function $\text{Cut} : \mathbb{MR}^n \to \mathbb{MR}^n_{\text{prop}}$ is defined by

$$\text{Cut}(\mathscr{Z}) \equiv (Z^B, Z^B \cap Z^C), \quad \forall \mathscr{Z} \in \mathbb{MR}^n. \tag{2.4}$$

**Definition 2.4.4.** For any $\mathbf{z} \in \mathbb{R}^n$, the element $([\mathbf{z}, \mathbf{z}], [\mathbf{z}, \mathbf{z}]) \in \mathbb{MR}^n$ is called *degenerate.*

Unlike elements of $\mathbb{IR}$, elements of $\mathbb{MR}$ are not subsets of $\mathbb{R}$, though it will be useful to interpret them as such. To do so unambiguously, we define the *enclosure* function.

**Definition 2.4.5.** The function Enc : $\mathbb{MR}^n \to \mathbb{IR}^n$ is defined by

$$\text{Enc}(\mathcal{Z}) \equiv Z^B \cap Z^C, \quad \forall \mathcal{Z} \in \mathbb{MR}^n. \tag{2.5}$$

According to the previous definition, the notation $\mathbf{z} \in \text{Enc}(\mathcal{Z})$ in well-defined, while $\mathbf{z} \in \mathcal{Z}$ is not. On the other hand, as elements of $\mathbb{IR}^n \times \mathbb{IR}^n$, elements of $\mathbb{MR}^n$ are subsets of $\mathbb{R}^n \times \mathbb{R}^n$, and the inclusion relation for $\mathcal{Z}_1, \mathcal{Z}_2 \in \mathbb{MR}^n$ is defined accordingly.

**Definition 2.4.6.** For any $\mathcal{Z}_1, \mathcal{Z}_2 \in \mathbb{MR}^n$, the inclusion $\mathcal{Z}_1 \subset \mathcal{Z}_2$ holds if and only if $Z_1^B \subset Z_2^B$ and $Z_1^C \subset Z_2^C$.

As with intervals, the set $\mathbb{MR}^{n\times m}$ can be defined analogously to $\mathbb{MR}^n$; $A \in \mathbb{MR}^{n\times m}$ has elements $A_{ij} \in \mathbb{MR}$, for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$. For any $D \subset \mathbb{R}^n$, let $\mathbb{M}D$ denote the set $\{\mathcal{Z} \in \mathbb{MR}^n : Z^B \subset D\}$. This notation is also used for $D \subset \mathbb{R}^{n\times m}$.

In what follows, McCormick's technique is formalized by defining operations on $\mathbb{MR}^n$, leading to a *relaxation function* analogous to the inclusion function of interval analysis. This presentation is not standard. However, the resulting method is equivalent and there are numerous advantages. First, the notation is much more compact and bears a direct relationship with the standard computational implementation of the method. Second, more precise statements of certain properties are achieved. Finally, the construction of the *generalized McCormick relaxations* presented in §2.7 becomes evident and is achieved with minimal additional effort.

### 2.4.1 Relaxation Functions and McCormick Extensions

By analogy to the inclusion function of §2.3.1, the *relaxation function* is defined here as the fundamental object that we wish to compute for a given function $\mathbf{f} : D \to \mathbb{R}^m$. As described above, McCormick's technique takes an interval $X \in \mathbb{I}D$ and a point $\mathbf{x} \in X$ as input and returns an interval $F(X)$ and relaxation values $\mathbf{f}^{cv}(X, \mathbf{x})$ as $\mathbf{f}^{cc}(X, \mathbf{x})$ as output. Accordingly, it is sensible to define our notion of a relaxation function as a mapping $\mathcal{F} : \mathbb{I}D \times D \to \mathbb{MR}^n$, with some appropriate convexity and enclosure properties. Of course, the interpretation of the output is $\mathcal{F}(X,\mathbf{x}) = (F(X),[\mathbf{f}^{cv}(X,\mathbf{x}),\mathbf{f}^{cc}(X,\mathbf{x})])$. However, a much more useful object is the mapping $\mathcal{F} : \mathbb{M}D \to \mathbb{MR}^n$. The same interpretation can be recovered using arguments of the form $\mathcal{X} = (X,[\mathbf{x},\mathbf{x}])$. At the same time, more general inputs are allowed, which leads directly to the notion of a generalized McCormick relaxation. In particular, mappings of this form are composable.

Relaxation functions are defined below, after some preliminary concepts are introduced.

**Definition 2.4.7.** Let $\mathcal{X},\mathcal{Y} \in \mathbb{MR}^n$. $\mathcal{X}$ and $\mathcal{Y}$ are *coherent*, or $\mathcal{X}$ is *coherent to* $\mathcal{Y}$, if $X^B = Y^B$. A set $\mathcal{D} \subset \mathbb{MR}^n$ is *closed under coherence* if, for every coherent $\mathcal{X},\mathcal{Y} \in \mathbb{MR}^n$, $\mathcal{X} \in \mathcal{D}$ implies $\mathcal{Y} \in \mathcal{D}$. If $\mathcal{D}$ is closed under coherence, then $Q \in \mathbb{IR}^n$ is said to be *represented in* $\mathcal{D}$ if there exists $\mathcal{X} \in \mathcal{D}$ with $X^B = Q$. A function $\mathcal{F} : \mathcal{D} \to \mathbb{MR}^m$ is coherent if $\mathcal{D}$ is closed under coherence and $\mathcal{F}(\mathcal{X})$ is coherent to $\mathcal{F}(\mathcal{Y})$ for every coherent $\mathcal{X},\mathcal{Y} \in \mathcal{D}$.

It is easy to see that any set of the form $\mathbb{M}D$, with $D \subset \mathbb{R}^n$, is closed under coherence, and any $Q \in \mathbb{I}D$ is represented in $\mathbb{M}D$. In order to impose an appropriate convexity/concavity condition on relaxation functions, it is necessary to define convex combinations of coherent elements of $\mathbb{MR}^n$. Unfortunately, the addition and scalar multiplication operations on $\mathbb{MR}^n$, defined in the next section, are not suitable for this task because these operations are designed to propagate relaxation information, not to act as vector space operations. Therefore, for any coherent $\mathcal{X}_1,\mathcal{X}_2 \in \mathbb{MR}^n$ with common interval part $Q$, we define

$$\text{Conv}(\lambda,\mathcal{X}_1,\mathcal{X}_2) \equiv (Q, \lambda X_1^C + (1-\lambda)X_2^C), \quad \forall \lambda \in [0,1].$$

Note in particular that $\text{Conv}(\lambda,\mathcal{X}_1,\mathcal{X}_2)$ is coherent to both $\mathcal{X}_1$ and $\mathcal{X}_2$, so that $\mathcal{X}_1,\mathcal{X}_2 \in \mathcal{D}$ implies that $\text{Conv}(\lambda,\mathcal{X}_1,\mathcal{X}_2) \in \mathcal{D}$ for any $\mathcal{D}$ that is closed under coherence.

**Definition 2.4.8.** A function $\mathcal{F} : \mathcal{D} \to \mathbb{MR}^m$ is *coherently concave* on $\mathcal{D}$ if it is coherent and, for every coherent $\mathcal{X}_1, \mathcal{X}_2 \in \mathcal{D}$,

$$\mathcal{F}(\mathrm{Conv}(\lambda, \mathcal{X}_1, \mathcal{X}_2)) \supset \mathrm{Conv}(\lambda, \mathcal{F}(\mathcal{X}_1), \mathcal{F}(\mathcal{X}_2)), \quad \forall \lambda \in [0, 1].$$

**Remark 2.4.9.** In the previous definition, the term coherently concave is used instead of coherently convex because of the direction of the required inclusion. If $\mathbb{MR}^n$ were a vector space and one considered the partial ordering imposed by the inclusion relation (i.e., $\leq = \subset$ and $\geq = \supset$), then a definition of concavity through the inclusion above would be consistent with the standard definition of concavity on a vector space. As mentioned above, $\mathbb{MR}^n$ is not a vector space, but we choose the term concave nonetheless.

**Definition 2.4.10.** Let $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$. A mapping $\mathcal{F} : \mathcal{D} \to \mathbb{MR}^m$ is a *relaxation function* for $\mathbf{f}$ on $\mathcal{D}$ if it is coherently concave on $\mathcal{D}$, and every $\mathcal{X} \in \mathcal{D}$ satisfies, $\mathbf{f}(\mathbf{x}) \in \mathrm{Enc}(\mathcal{F}(\mathcal{X}))$, $\forall \mathbf{x} \in \mathrm{Enc}(\mathcal{X})$.

The following lemma shows that this definition indeed provides convex and concave relaxations of $\mathbf{f}$. It uses the notation $\mathcal{F}(\mathcal{X}) = (F^B(\mathcal{X}), [\mathbf{f}^{cv}(\mathcal{X}), \mathbf{f}^{cc}(\mathcal{X})])$.

**Lemma 2.4.11.** *Let* $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ *and let* $\mathcal{F} : \mathcal{D} \to \mathbb{MR}^m$ *be a* relaxation function *for* $\mathbf{f}$ *on* $\mathcal{D}$*. For any* $X \in \mathbb{I}D$ *that is represented in* $\mathcal{D}$*, define the functions* $\mathcal{U}, \mathcal{O} : X \to \mathbb{R}^m$ *by*

$$\mathcal{U}(\mathbf{x}) = \mathbf{f}^{cv}((X, [\mathbf{x}, \mathbf{x}])) \quad \text{and} \quad \mathcal{O}(\mathbf{x}) = \mathbf{f}^{cc}((X, [\mathbf{x}, \mathbf{x}])) \tag{2.6}$$

*for all* $\mathbf{x} \in X$*. Then* $\mathcal{U}$ *is convex on* $X$*,* $\mathcal{O}$ *is concave on* $X$*, and* $\mathcal{U}(\mathbf{x}) \leq \mathbf{f}(\mathbf{x}) \leq \mathcal{O}(\mathbf{x})$*,* $\forall \mathbf{x} \in X$*.*

*Proof.* Choose any $\mathbf{x} \in X$. Since $\mathcal{D}$ is closed under coherence, $(X, [\mathbf{x}, \mathbf{x}]) \in \mathcal{D}$ for all $\mathbf{x} \in X$. Noting that $\mathbf{x} \in \mathrm{Enc}((X, [\mathbf{x}, \mathbf{x}]))$, it follows that $\mathbf{f}(\mathbf{x}) \in \mathrm{Enc}(\mathcal{F}((X, [\mathbf{x}, \mathbf{x}])))$. In particular, $\mathcal{U}(\mathbf{x}) = \mathbf{f}^{cv}((X, [\mathbf{x}, \mathbf{x}])) \leq \mathbf{f}(\mathbf{x}) \leq \mathbf{f}^{cc}((X, [\mathbf{x}, \mathbf{x}])) = \mathcal{O}(\mathbf{x})$.

Choose any $\mathbf{x}_1, \mathbf{x}_2 \in X$ and any $\lambda \in [0, 1]$. Then

$$\begin{aligned}\mathcal{F}(\mathrm{Conv}(\lambda, &(X, [\mathbf{x}_1, \mathbf{x}_1]), (X, [\mathbf{x}_2, \mathbf{x}_2]))) \\ &\supset \mathrm{Conv}(\lambda, \mathcal{F}((X, [\mathbf{x}_1, \mathbf{x}_1])), \mathcal{F}((X, [\mathbf{x}_2, \mathbf{x}_2]))).\end{aligned}$$

In particular

$$\begin{aligned}\mathcal{U}(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) &= \mathbf{f}^{cv}((X, [\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2, \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2])), \\ &\leq \lambda\mathbf{f}^{cv}((X, [\mathbf{x}_1, \mathbf{x}_1])) + (1-\lambda)\mathbf{f}^{cv}((X, [\mathbf{x}_2, \mathbf{x}_2])), \\ &= \lambda\mathcal{U}(\mathbf{x}_1) + (1-\lambda)\mathcal{U}(\mathbf{x}_2).\end{aligned}$$

Concavity of $\mathcal{O}$ follows analogously. □

As with inclusion functions, the enclosure property of a relaxation function will be achieved through a simpler construction, the *McCormick extension*, with the help of a monotonicity property.

**Definition 2.4.12.** Let $D \subset \mathbb{R}^n$. A set $\mathcal{D} \subset \mathbb{MR}^n$ is a McCormick extension of $D$ if every $\mathbf{x} \in D$ satisfies $([\mathbf{x}, \mathbf{x}], [\mathbf{x}, \mathbf{x}]) \in \mathcal{D}$. Let $\mathbf{f} : D \to \mathbb{R}^m$. A mapping $\mathcal{F} : \mathcal{D} \to \mathbb{MR}^m$ is an *McCormick extension* of $\mathbf{f}$ if $\mathcal{D}$ is a McCormick extension of $D$, and $\mathcal{F}(([\mathbf{x}, \mathbf{x}], [\mathbf{x}, \mathbf{x}])) = ([\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x})], [\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x})])$, $\forall \mathbf{x} \in D$.

Note that, for any $D \subset \mathbb{R}^n$, $\mathbb{M}D$ is a McCormick extension of $D$.

**Definition 2.4.13.** Let $\mathcal{F} : \mathcal{D} \subset \mathbb{MR}^n \to \mathbb{MR}^m$. $\mathcal{F}$ is *inclusion monotonic* on $\mathcal{D}$ if $\mathcal{X}_1 \subset \mathcal{X}_2 \implies \mathcal{F}(\mathcal{X}_1) \subset \mathcal{F}(\mathcal{X}_2)$, $\forall \mathcal{X}_1, \mathcal{X}_2 \in \mathcal{D}$.

**Theorem 2.4.14.** *Let* $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ *and let* $\mathcal{F} : \mathcal{D} \to \mathbb{MR}^m$ *be a McCormick extension of* $\mathbf{f}$. *If* $\mathcal{F}$ *is inclusion monotonic on* $\mathcal{D} \cap \mathbb{M}D$, *then every* $\mathcal{X} \in \mathcal{D} \cap \mathbb{M}D$ *satisfies* $\mathbf{f}(\mathbf{x}) \in \mathrm{Enc}(\mathcal{F}(\mathcal{X}))$, $\forall \mathbf{x} \in \mathrm{Enc}(\mathcal{X})$.

*Proof.* Choose any $\mathcal{X} \in \mathcal{D} \cap \mathbb{M}D$ and any $\mathbf{x} \in \mathrm{Enc}(\mathcal{X})$. Then $\mathbf{x} \in X^B \subset D$ and hence $([\mathbf{x}, \mathbf{x}], [\mathbf{x}, \mathbf{x}]) \in \mathcal{D}$ and $\mathcal{F}(([\mathbf{x}, \mathbf{x}], [\mathbf{x}, \mathbf{x}])) = ([\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x})], [\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x})])$. Then, by inclusion monotonicity, $\mathbf{f}(\mathbf{x}) \in \mathrm{Enc}(\mathcal{F}(([\mathbf{x}, \mathbf{x}], [\mathbf{x}, \mathbf{x}]))) \subset \mathrm{Enc}(\mathcal{F}(\mathcal{X}))$. □

The following composition results are useful for constructing relaxation functions for complex functions from those of simpler functions.

**Lemma 2.4.15.** *Let $\mathcal{D}_1 \subset \mathbb{MR}^n$ and $\mathcal{D}_2 \subset \mathbb{MR}^m$ be closed under coherence, and let $\mathcal{F}_1 : \mathcal{D}_1 \to \mathbb{MR}^m$ and $\mathcal{F}_2 : \mathcal{D}_2 \to \mathbb{MR}^k$ be coherently concave and inclusion monotonic on $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. Then the set $\mathcal{D}_{12} \equiv \{\mathcal{X} \in \mathcal{D}_1 : \mathcal{F}_1(\mathcal{X}) \in \mathcal{D}_2\}$ is closed under coherence and $\mathcal{F}_2 \circ \mathcal{F}_1$ is coherently concave and inclusion monotonic on $\mathcal{D}_{12}$.*

*Proof.* Let $\mathcal{X}, \mathcal{Y} \in \mathbb{MR}^n$ be coherent and suppose that $\mathcal{X} \in \mathcal{D}_{12}$. To show that $\mathcal{D}_{12}$ is closed under coherence, it is shown that $\mathcal{Y} \in \mathcal{D}_{12}$. Since $\mathcal{X}$ is in $\mathcal{D}_{12}$, it is also in $\mathcal{D}_1$, and since $\mathcal{D}_1$ is closed under coherence, $\mathcal{Y} \in \mathcal{D}_1$. Since $\mathcal{F}_1$ is coherently concave, $\mathcal{F}_1(\mathcal{X})$ and $\mathcal{F}_1(\mathcal{Y})$ are coherent. But $\mathcal{F}_1(\mathcal{X}) \in \mathcal{D}_2$ because $\mathcal{X} \in \mathcal{D}_{12}$, and hence $\mathcal{F}_1(\mathcal{Y}) \in \mathcal{D}_2$ because $\mathcal{D}_2$ is closed under coherence. Then, by definition, $\mathcal{Y} \in \mathcal{D}_{12}$, so $\mathcal{D}_{12}$ is closed under coherence.

Choose any $\lambda \in [0, 1]$. Because $\mathcal{F}_1$ is coherently concave,

$$\mathcal{F}_1(\mathrm{Conv}(\lambda, \mathcal{X}, \mathcal{Y})) \supset \mathrm{Conv}(\lambda, \mathcal{F}_1(\mathcal{X}), \mathcal{F}_1(\mathcal{Y})), \tag{2.7}$$

and $\mathcal{F}_1(\mathcal{X})$ and $\mathcal{F}_1(\mathcal{Y})$ are coherent. Because $\mathcal{F}_2$ is coherently concave,

$$\mathcal{F}_2(\mathrm{Conv}(\lambda, \mathcal{F}_1(\mathcal{X}), \mathcal{F}_1(\mathcal{Y}))) \supset \mathrm{Conv}(\lambda, \mathcal{F}_2(\mathcal{F}_1(\mathcal{X})), \mathcal{F}_2(\mathcal{F}_1(\mathcal{Y}))) \tag{2.8}$$

and $\mathcal{F}_2(\mathcal{F}_1(\mathcal{X}))$ and $\mathcal{F}_2(\mathcal{F}_1(\mathcal{Y}))$ are coherent. Since $\mathcal{F}_1(\mathrm{Conv}(\lambda, \mathcal{X}, \mathcal{Y}))$ is coherent to $\mathcal{F}_1(\mathcal{X})$, it is an element of $\mathcal{D}_2$. Since, $\mathcal{F}_2$ is inclusion monotonic on $\mathcal{D}_2$, combining (2.7) and (2.8) shows that

$$\mathcal{F}_2(\mathcal{F}_1(\mathrm{Conv}(\lambda, \mathcal{X}, \mathcal{Y}))) \supset \mathrm{Conv}(\lambda, \mathcal{F}_2(\mathcal{F}_1(\mathcal{X})), \mathcal{F}_2(\mathcal{F}_1(\mathcal{Y}))), \tag{2.9}$$

which shows that $\mathcal{F}_2 \circ \mathcal{F}_1$ is coherently concave on $\mathcal{D}_{12}$.

It remains to show that $\mathcal{F}_2 \circ \mathcal{F}_1$ is inclusion monotonic on $\mathcal{D}_{12}$. Choose any $\mathcal{X}, \mathcal{Y} \in \mathcal{D}_{12}$ such that $\mathcal{X} \subset \mathcal{Y}$. Then $\mathcal{F}_1(\mathcal{X}) \subset \mathcal{F}_1(\mathcal{Y})$, and both are elements of $\mathcal{D}_2$, so that $\mathcal{F}_2(\mathcal{F}_1(\mathcal{X})) \subset \mathcal{F}_2(\mathcal{F}_1(\mathcal{Y}))$. □

**Remark 2.4.16.** Note that only inclusion monotonicity of $\mathcal{F}_2$ was required to recover coherent concavity of the composition $\mathcal{F}_2 \circ \mathcal{F}_1$. This is analogous to standard composition results for convex and concave functions, where one must assume a monotonicity property for the outer function.

**Lemma 2.4.17.** *Let* $\mathbf{f}_1 : D_1 \subset \mathbb{R}^n \to \mathbb{R}^m$ *and* $\mathbf{f}_2 : D_2 \subset \mathbb{R}^m \to \mathbb{R}^k$, *and define* $D_{12} \equiv \{\mathbf{x} \in D_1 : \mathbf{f}_1(\mathbf{x}) \in D_2\}$. *Let* $\mathcal{F}_1 : \mathcal{D}_1 \to \mathbb{MR}^m$ *and* $\mathcal{F}_2 : \mathcal{D}_2 \to \mathbb{MR}^k$ *be McCormick extensions of* $\mathbf{f}_1$ *and* $\mathbf{f}_2$, *respectively. Then* $\mathcal{D}_{12} \equiv \{\mathcal{X} \in \mathcal{D}_1 : \mathcal{F}_1(\mathcal{X}) \in \mathcal{D}_2\}$ *is a McCormick extension of* $D_{12}$, *and* $(\mathcal{F}_2 \circ \mathcal{F}_1, \mathcal{D}_{12}, \mathbb{MR}^k)$ *is a McCormick extension of* $(\mathbf{f}_2 \circ \mathbf{f}_1, D_{12}, \mathbb{R}^k)$.

*Proof.* First it is shown that $\mathbf{x} \in D_{12}$ implies $([\mathbf{x},\mathbf{x}],[\mathbf{x},\mathbf{x}]) \in \mathcal{D}_{12}$. For any $\mathbf{x} \in D_{12}$, $\mathbf{x} \in D_1$ implies that $([\mathbf{x},\mathbf{x}],[\mathbf{x},\mathbf{x}]) \in \mathcal{D}_1$ because $\mathcal{D}_1$ is a McCormick extension of $D_1$. Because $\mathcal{F}_1$ is a McCormick extension of $\mathbf{f}_1$, $\mathcal{F}_1([\mathbf{x},\mathbf{x}],[\mathbf{x},\mathbf{x}]) = ([\mathbf{f}_1(\mathbf{x}),\mathbf{f}_1(\mathbf{x})],[\mathbf{f}_1(\mathbf{x}),\mathbf{f}_1(\mathbf{x})])$. Since $\mathbf{x} \in D_{12}$, we have $\mathbf{f}_1(\mathbf{x}) \in D_2$, which implies that $\mathcal{F}_1([\mathbf{x},\mathbf{x}],[\mathbf{x},\mathbf{x}]) \in \mathcal{D}_2$ because $\mathcal{D}_2$ is a McCormick extension of $D_2$. By definition, this implies that $([\mathbf{x},\mathbf{x}],[\mathbf{x},\mathbf{x}]) \in \mathcal{D}_{12}$.

To show that $(\mathcal{F}_2 \circ \mathcal{F}_1, \mathcal{D}_{12}, \mathbb{MR}^k)$ is a McCormick extension of $(\mathbf{f}_2 \circ \mathbf{f}_1, D_{12}, \mathbb{R}^k)$, choose any $\mathbf{x} \in D_{12}$. Since $\mathcal{D}_{12}$ is a McCormick extension of $D_{12}$, $([\mathbf{x},\mathbf{x}],[\mathbf{x},\mathbf{x}]) \in \mathcal{D}_{12}$. Then,

$$\begin{aligned}\mathcal{F}_2(\mathcal{F}_1(([\mathbf{x},\mathbf{x}],[\mathbf{x},\mathbf{x}]))) &= \mathcal{F}_2(([\mathbf{f}_1(\mathbf{x}),\mathbf{f}_1(\mathbf{x})],[\mathbf{f}_1(\mathbf{x}),\mathbf{f}_1(\mathbf{x})])),\\ &= ([\mathbf{f}_2(\mathbf{f}_1(\mathbf{x})),\mathbf{f}_2(\mathbf{f}_1(\mathbf{x}))],[\mathbf{f}_2(\mathbf{f}_1(\mathbf{x})),\mathbf{f}_2(\mathbf{f}_1(\mathbf{x}))]).\end{aligned}$$

□

### 2.4.2 McCormick Arithmetic and the Natural McCormick Extension

In this section, the basic operations defining $\mathcal{L}$-factorable functions are extended to $\mathbb{MR}$. Aside from some minor differences discussed below, these extensions are the addition, multiplication and composition rules of McCormick's original work [118].

**Definition 2.4.18.** Define $(+, \mathbb{MR}^2, \mathbb{MR})$ by

$$+(\mathcal{X}, \mathcal{Y}) = \mathcal{X} + \mathcal{Y} = (X^B + Y^B, (X^B \cap X^C) + (Y^B \cap Y^C)). \tag{2.10}$$

In the following results, it is shown that $(+, \mathbb{MR}^2, \mathbb{MR})$ is a McCormick extension of $(+, \mathbb{R}^2, \mathbb{R})$, and is coherently concave and inclusion monotonic on $\mathbb{MR}^2$.

**Theorem 2.4.19.** *For any coherent* $\mathcal{X}_1, \mathcal{X}_2 \in \mathbb{MR}^n$ *with common interval part* $Q$*,* $Q \cap (\lambda X_1^C + (1-\lambda)X_2^C) \supset \lambda(Q \cap X_1^C) + (1-\lambda)(Q \cap X_2^C)$*,* $\forall \lambda \in [0,1]$*.*

*Proof.* Letting $Q = [q^L, q^U]$, it suffices to show that

$$\begin{aligned} \max(q^L, \lambda x_1^{cv} + (1-\lambda)x_2^{cv}) &\leq \lambda \max(q^L, x_1^{cv}) + (1-\lambda)\max(q^L, x_2^{cv}), \\ \min(q^U, \lambda x_1^{cc} + (1-\lambda)x_2^{cc}) &\geq \lambda \min(q^U, x_1^{cc}) + (1-\lambda)\min(q^U, x_2^{cc}). \end{aligned}$$

But this follows directly from the fact that $\max(q^L, \cdot)$ and $\min(q^U, \cdot)$ are convex and concave on $\mathbb{R}$, respectively. □

**Theorem 2.4.20.** $(+, \mathbb{MR}^2, \mathbb{MR})$ *is a McCormick extension of* $(+, \mathbb{R}^2, \mathbb{R})$*. Furthermore, it is coherently concave and inclusion monotonic on* $\mathbb{MR}^2$*.*

*Proof.* $\mathbb{MR}^2$ is clearly a McCormick extension of $\mathbb{R}^2$, and for any $(x, y) \in \mathbb{R}^2$,

$$\begin{aligned} ([x,x],[x,x]) + ([y,y],[y,y]) &= ([x,x] + [y,y], [x,x] + [y,y]), & (2.11) \\ &= ([x+y, x+y], [x+y, x+y]). & (2.12) \end{aligned}$$

To show that $(+, \mathbb{MR}^2, \mathbb{MR})$ is inclusion monotonic, let $(\mathcal{X}_1, \mathcal{Y}_1), (\mathcal{X}_2, \mathcal{Y}_2) \in \mathbb{MR}^2$ and suppose that $\mathcal{X}_2 \subset \mathcal{X}_1$ and $\mathcal{Y}_2 \subset \mathcal{Y}_1$. Then $\mathcal{X}_2 + \mathcal{Y}_2 = ([X_2^B + Y_2^B], [(X_2^B \cap X_2^C) + (Y_2^B \cap Y_2^C)]) \subset ([X_1^B + Y_1^B], [(X_1^B \cap X_1^C) + (Y_1^B \cap Y_1^C)]) = \mathcal{X}_1 + \mathcal{Y}_1$.

It remains to show that $(+, \mathbb{MR}^2, \mathbb{MR})$ is coherently concave on $\mathbb{MR}^2$. Clearly, $\mathbb{MR}^2$ is closed under coherence. Choose any coherent $(\mathcal{X}_1, \mathcal{Y}_1), (\mathcal{X}_2, \mathcal{Y}_2) \in \mathbb{MR}^2$ and let $Q \times R$ denote their common interval part. It is clear that $\mathcal{Z}_1 = \mathcal{X}_1 + \mathcal{Y}_1$ and $\mathcal{Z}_2 = \mathcal{X}_2 + \mathcal{Y}_2$ are coherent with interval $Q + R$. Choose any $\lambda \in [0,1]$ and define

$\hat{\mathcal{X}} = \mathrm{Conv}(\lambda, \mathcal{X}_1, \mathcal{X}_2)$, $\hat{\mathcal{Y}} = \mathrm{Conv}(\lambda, \mathcal{Y}_1, \mathcal{Y}_2)$, and $\hat{\mathcal{Z}} = \hat{\mathcal{X}} + \hat{\mathcal{Y}}$. Then, using Theorem 2.4.19,

$$\begin{aligned}
\hat{Z}^C &= (Q \cap \hat{X}^C) + (R \cap \hat{Y}^C), \qquad &(2.13)\\
&= (Q \cap (\lambda X_1^C + (1-\lambda)X_2^C)) + (R \cap (\lambda Y_1^C + (1-\lambda)Y_2^C)), &(2.14)\\
&\supset \lambda(Q \cap X_1^C) + (1-\lambda)(Q \cap X_2^C) + \lambda(R \cap Y_1^C) + (1-\lambda)(R \cap Y_2^C), &(2.15)\\
&= \lambda\left[(Q \cap X_1^C) + (Q \cap Y_1^C)\right] + (1-\lambda)\left[(R \cap X_2^C) + (R \cap Y_2^C)\right], &(2.16)\\
&= \lambda Z_1^C + (1-\lambda)Z_2^C. &(2.17)
\end{aligned}$$

It follows that $\hat{\mathcal{Z}} \supset \mathrm{Conv}(\lambda, \mathcal{Z}_1, \mathcal{Z}_2)$, which is the desired result. □

**Definition 2.4.21.** Define $(\times, \mathbb{MR}^2, \mathbb{MR})$ by

$$\times(\mathcal{X}, \mathcal{Y}) = \mathcal{X}\mathcal{Y} = (X^B Y^B, [z^{cv}, z^{cc}]), \tag{2.18}$$

where

$$\begin{aligned}
z^{cv} &= \max\left(\left[y^L \bar{X}^C + x^L \bar{Y}^C - x^L y^L\right]^L, \left[y^U \bar{X}^C + x^U \bar{Y}^C - x^U y^U\right]^L\right), \qquad &(2.19)\\
z^{cc} &= \min\left(\left[y^L \bar{X}^C + x^U \bar{Y}^C - y^L x^U\right]^U, \left[y^U \bar{X}^C + x^L \bar{Y}^C - y^U x^L\right]^U\right). &(2.20)
\end{aligned}$$

and $\bar{\mathcal{X}} = \mathrm{Cut}(\mathcal{X})$ and $\bar{\mathcal{Y}} = \mathrm{Cut}(\mathcal{Y})$.

In the previous definition, the algebraic expressions in square brackets evaluate to intervals, and the superscript $L$ or $U$ indicates the lower or upper bound of that interval, respectively. This definition is based on the convex and concave envelopes of the bilinear term $xy$ on the intervals $X^B$ and $Y^B$, given by

$$\begin{aligned}
&\max(y^L x + x^L y - y^L x^L, y^U x + x^U y - y^U x^U)\\
&\qquad\qquad \le xy \le \min(y^L x + x^U y - y^L x^U, y^U x + x^L y - y^U x^L).
\end{aligned}$$

From this, it is simple to show that $z^{cv} \le z^{cc}$, $z^{cv} \le z^U$ and $z^L \le z^{cc}$, so that $\mathcal{X}\mathcal{Y}$ is indeed an element of $\mathbb{MR}$.

The notation above is not typically used to define McCormick multiplication. However, expanding $z^{cv}$, for example, gives

$$\begin{aligned} z^{cv} &= \max([y^L\bar{X}^C + x^L\bar{Y}^C - x^Ly^L]^L, [y^U\bar{X}^C + x^U\bar{Y}^C - x^Uy^U]^L), \qquad (2.21)\\ &= \max([y^L\bar{X}^C]^L + [x^L\bar{Y}^C]^L - x^Ly^L, [y^U\bar{X}^C]^L + [x^U\bar{Y}^C]^L - x^Uy^U),\\ &= \max(\min(y^L\bar{x}^{cv}, y^L\bar{x}^{cc}) + \min(x^L\bar{y}^{cv}, x^L\bar{y}^{cc}) - x^Ly^L,\\ &\qquad \min(y^U\bar{x}^{cv}, y^U\bar{x}^{cc}) + \min(x^U\bar{y}^{cv}, x^U\bar{y}^{cc}) - x^Uy^U). \end{aligned}$$

Readers familiar with the standard definition will now see that the definition above is equivalent, with the exception that the Cut operation is not applied to $\mathcal{X}$ and $\mathcal{Y}$ in McCormick's original work [118]. Note also that this operation also appears in the definition of $(+, \mathbb{MR}^2, \mathbb{MR})$ (written out explicitly in this case), though not in McCormick's original definition. In general, this step potentially makes the results of these operations sharper. It also makes $\bar{\mathcal{X}}$ and $\bar{\mathcal{Y}}$ proper, which has important consequences for the inclusion monotonicity of McCormick multiplication, as discussed below.

**Theorem 2.4.22.** $(\times, \mathbb{MR}^2, \mathbb{MR})$ *is a McCormick extension of* $(\times, \mathbb{R}^2, \mathbb{R})$.

*Proof.* Let $(x, y) \in \mathbb{MR}^2$. Multiplying $([x, x], [x, x])$ and $([y, y], [y, y])$ as per Definition 2.4.21, the conclusion follows from the observations

$$\begin{aligned} z^{cv} &= \max([y[x, x] + x[y, y] - xy]^L, [y[x, x] + x[y, y] - xy]^L) = xy,\\ z^{cc} &= \min([y[x, x] + x[y, y] - xy]^U, [y[x, x] + x[y, y] - xy]^U) = xy. \end{aligned}$$

□

**Theorem 2.4.23.** $(\times, \mathbb{MR}^2, \mathbb{MR})$ *is inclusion monotonic on* $\mathbb{MR}^2$.

*Proof.* Let $\mathcal{X}_1, \mathcal{Y}_1, \mathcal{X}_2, \mathcal{Y}_2 \in \mathbb{MR}$ and suppose that $\mathcal{X}_2 \subset \mathcal{X}_1$ and $\mathcal{Y}_2 \subset \mathcal{Y}_1$. It follows that $X_2^B \subset X_1^B$, $Y_2^B \subset Y_1^B$, $\bar{X}_2^C \subset \bar{X}_1^C$ and $\bar{Y}_2^C \subset \bar{Y}_1^C$. By Theorem 2.3.7, $X_2^BY_2^B \subset X_1^BY_1^B$. It remains to show that $[z_2^{cv}, z_2^{cc}] \subset [z_1^{cv}, z_1^{cc}]$, where $z_2^{cv}$, $z_2^{cc}$, $z_1^{cv}$ and $z_1^{cc}$ are defined as in Definition 2.4.21. It will be shown that

$$\begin{aligned}
z_1^{cv} &= \max([y_1^L \bar{X}_1^C + x_1^L \bar{Y}_1^C - x_1^L y_1^L]^L, [y_1^U \bar{X}_1^C + x_1^U \bar{Y}_1^C - x_1^U y_1^U]^L), \\
&\leq \max([y_1^L \bar{X}_2^C + x_1^L \bar{Y}_2^C - x_1^L y_1^L]^L, [y_1^U \bar{X}_2^C + x_1^U \bar{Y}_2^C - x_1^U y_1^U]^L), \\
&\leq \max([y_2^L \bar{X}_2^C + x_2^L \bar{Y}_2^C - x_2^L y_2^L]^L, [y_2^U \bar{X}_2^C + x_2^U \bar{Y}_2^C - x_2^U y_2^U]^L), \\
&= z_2^{cv}.
\end{aligned}$$

The proof that $z_1^{cc} \geq z_2^{cc}$ is analogous. In general, $\max(a, b) \leq \max(a', b')$ if $a \leq a'$ and $b \leq b'$. It will be shown that

$$\begin{aligned}
[y_1^L \bar{X}_1^C + x_1^L \bar{Y}_1^C - x_1^L y_1^L]^L &\leq [y_1^L \bar{X}_2^C + x_1^L \bar{Y}_2^C - x_1^L y_1^L]^L \qquad (2.22) \\
&\leq [y_2^L \bar{X}_2^C + x_2^L \bar{Y}_2^C - x_2^L y_2^L]^L.
\end{aligned}$$

The remaining inequality is proven analogously. The first inequality in (2.22) follows directly by inclusion monotonicity of interval multiplication and addition, and the fact that $\bar{X}_1^C \supset \bar{X}_2^C$ and $\bar{Y}_1^C \supset \bar{Y}_2^C$. Consider the second inequality in (2.22). First, it is shown that

$$\begin{aligned}
[y_1^L \bar{X}_2^C + x_1^L \bar{Y}_2^C - x_1^L y_1^L]^L &= [y_1^L (\bar{X}_2^C - x_1^L) + x_1^L \bar{Y}_2^C]^L, \qquad (2.23) \\
&= [y_1^L (\bar{X}_2^C - x_1^L)]^L + [x_1^L \bar{Y}_2^C]^L, \qquad (2.24) \\
&\leq [y_2^L (\bar{X}_2^C - x_1^L)]^L + [x_1^L \bar{Y}_2^C]^L. \qquad (2.25)
\end{aligned}$$

Since $\bar{X}_2^C \subset X_2^B \subset X_1^B$, the interval $(\bar{X}_2^C - x_1^L)$ contains no negative elements. Since $y_1^L \leq y_2^L$, it follows that $[y_1^L (\bar{X}_2^C - x_1^L)]^L \leq [y_2^L (\bar{X}_2^C - x_1^L)]^L$, which implies the inequality above. Then, using an identical argument,

$$\begin{aligned}
[y_1^L \bar{X}_2^C + x_1^L \bar{Y}_2^C - x_1^L y_1^L]^L &\leq [y_2^L(\bar{X}_2^C - x_1^L) + x_1^L \bar{Y}_2^C]^L, \qquad &(2.26)\\
&= [y_2^L \bar{X}_2^C + x_1^L \bar{Y}_2^C - y_2^L x_1^L]^L, &(2.27)\\
&= [y_2^L \bar{X}_2^C + x_1^L(\bar{Y}_2^C - y_2^L)]^L, &(2.28)\\
&\leq [y_2^L \bar{X}_2^C + x_2^L(\bar{Y}_2^C - y_2^L)]^L, &(2.29)\\
&= [y_2^L \bar{X}_2^C + x_2^L \bar{Y}_2^C - x_2^L y_2^L]^L. &(2.30)
\end{aligned}$$

This proves the second inequality in (2.22). □

**Theorem 2.4.24.** $(\times, \mathbb{MR}^2, \mathbb{MR})$ *is coherently concave on* $\mathbb{MR}^2$*.*

*Proof.* Choose any coherent $(\mathcal{X}_1, \mathcal{Y}_1), (\mathcal{X}_2, \mathcal{Y}_2) \in \mathbb{MR}^2$ with common interval part $Q \times R$. It is clear that $\mathcal{X}_1\mathcal{Y}_1$ and $\mathcal{X}_2\mathcal{Y}_2$ are coherent with common interval part $QR$. Choose any $\lambda \in [0,1]$ and let $\hat{\mathcal{X}} = \text{Conv}(\lambda, \mathcal{X}_1, \mathcal{X}_2)$ and $\hat{\mathcal{Y}} = \text{Conv}(\lambda, \mathcal{Y}_1, \mathcal{Y}_2)$. By Lemma 2.4.19, $\lambda \bar{X}_1^C + (1-\lambda)\bar{X}_2^C \subset (Q \cap \hat{X}^C)$ and $\lambda \bar{Y}_1^C + (1-\lambda)\bar{Y}_2^C \subset (R \cap \hat{Y}^C)$. It follows that

$$\begin{aligned}
\lambda r \bar{X}_1^C + (1-\lambda) r \bar{X}_2^C &\subset r(Q \cap \hat{X}^C), \qquad &(2.31)\\
\lambda r \bar{Y}_1^C + (1-\lambda) r \bar{Y}_2^C &\subset r(R \cap \hat{Y}^C), &(2.32)
\end{aligned}$$

for any $r \in \mathbb{R}$. Then

$$\begin{aligned}
&[y^L(Q \cap \hat{X}^C) + x^L(R \cap \hat{Y}^C) - y^L x^L]^L\\
&\qquad = [y^L(Q \cap \hat{X}^C)]^L + [x^L(R \cap \hat{Y}^C)]^L - y^L x^L\\
&\qquad \leq [\lambda y^L \bar{X}_1^C + (1-\lambda) y^L \bar{X}_2^C]^L + [\lambda x^L \bar{Y}_1^C + (1-\lambda) x^L \bar{Y}_2^C]^L - y^L x^L,\\
&\qquad = \lambda [y^L \bar{X}_1^C]^L + (1-\lambda)[y^L \bar{X}_2^C]^L + \lambda [x^L \bar{Y}_1^C]^L + (1-\lambda)[x^L \bar{Y}_2^C]^L - y^L x^L,\\
&\qquad = \lambda([y^L \bar{X}_1^C]^L + [x^L \bar{Y}_1^C]^L) + (1-\lambda)([y^L \bar{X}_2^C]^L + [x^L \bar{Y}_2^C]^L) - y^L x^L,\\
&\qquad = \lambda([y^L \bar{X}_1^C + x^L \bar{Y}_1^C - y^L x^L]^L) + (1-\lambda)([y^L \bar{X}_2^C + x^L \bar{Y}_2^C - y^L x^L]^L).
\end{aligned}$$

By an analogous sequence of manipulations it can be shown that

$$\begin{aligned}&[y^U(Q \cap \hat{X}^C) + x^U(R \cap \hat{Y}^C) - y^U x^U]^L \\ &\qquad \leq \lambda([y^U \bar{X}_1^C + x^U \bar{Y}_1^C - y^U x^U]^L) + (1-\lambda)([y^U \bar{X}_2^C + x^U \bar{Y}_2^C - y^U x^U]^L).\end{aligned}$$

By convexity of max on $\mathbb{R}^2$, it follows that

$$\begin{aligned}&\max([y^L(Q \cap \hat{X}^C) + x^L(R \cap \hat{Y}^C) - y^L x^L]^L, [y^U(Q \cap \hat{X}^C) + x^U(R \cap \hat{Y}^C) - y^U x^U]^L) \\ &\leq \max(\lambda([y^L \bar{X}_1^C + x^L \bar{Y}_1^C - y^L x^L]^L) + (1-\lambda)([y^L \bar{X}_2^C + x^L \bar{Y}_2^C - y^L x^L]^L), \\ &\qquad \lambda([y^U \bar{X}_1^C + x^U \bar{Y}_1^C - y^U x^U]^L) + (1-\lambda)([y^U \bar{X}_2^C + x^U \bar{Y}_2^C - y^U x^U]^L)), \\ &\leq \lambda \max([y^L \bar{X}_1^C + x^L \bar{Y}_1^C - y^L x^L]^L, [y^U \bar{X}_1^C + x^U \bar{Y}_1^C - y^U x^U]^L) \\ &\qquad + (1-\lambda)\max([y^L \bar{X}_2^C + x^L \bar{Y}_2^C - y^L x^L]^L, [y^U \bar{X}_2^C + x^U \bar{Y}_2^C - y^U x^U]^L).\end{aligned}$$

Letting $z_1^{cv}$, $z_2^{cv}$ and $\hat{z}^{cv}$ be as in Definition 2.4.21, this last inequality is exactly $\hat{z}^{cv} \leq \lambda z_1^{cv} + (1-\lambda) z_2^{cv}$, and an analogous argument shows that $\hat{z}^{cc} \geq \lambda z_1^{cc} + (1-\lambda) z_2^{cc}$. Combined, these imply that $\hat{Z}^C \supset \lambda Z_1^C + (1-\lambda) Z_2^C$. □

By Theorem 2.4.14, it has now been established that the functions $(+, \mathbb{MR}^2, \mathbb{MR})$ and $(\times, \mathbb{MR}^2, \mathbb{MR})$ are relaxation functions for $(+, \mathbb{R}^2, \mathbb{R})$ and $(\times, \mathbb{R}^2, \mathbb{R})$ on $\mathbb{MR}^2$, respectively, and are moreover inclusion monotonic there. It should be noted that $(\times, \mathbb{MR}^2, \mathbb{MR})$ can be proven to be a relaxation function $(\times, \mathbb{R}^2, \mathbb{R})$ on $\mathbb{MR}^2$ directly, without first showing inclusion monotonicity. This is the standard development, in particular because $(\times, \mathbb{MR}^2, \mathbb{MR})$ is not inclusion monotonic without the use of the Cut operation in Definition 2.4.21. This is demonstrated by the following example.

**Example 2.4.1.** Let $\mathcal{X}_1 = \mathcal{Y}_1 = ([-1, 1], [-3, 1])$ and $\mathcal{X}_2 = \mathcal{Y}_2 = ([0.7, 1], [-2.5, 1])$, and note that $\mathcal{X}_2 \subset \mathcal{X}_1$ and $\mathcal{Y}_2 \subset \mathcal{Y}_1$. Despite these inclusion, it will be shown that

$\mathcal{X}_2\mathcal{Y}_2 \not\subset \mathcal{X}_1\mathcal{Y}_1$, *if the* Cut *operations are not used in Definition 2.4.21.* Let

$$\begin{aligned}
z_1^{cv} &\equiv \max\left(\left[y_1^L X_1^C + x_1^L Y_1^C - x_1^L y_1^L\right]^L, \left[y_1^U X_1^C + x_1^U Y_1^C - x_1^U y_1^U\right]^L\right) \\
&= \max\Big([(-1)[-3,1] + (-1)[-3,1] - (-1)(-1)]^L, \\
&\qquad\qquad [(1)[-3,1] + (1)[-3,1] - (1)(1)]^L\Big), \\
&= \max\Big([[-1,3] + [-1,3] - 1]^L, [[-3,1] + [-3,1] - 1]^L\Big), \\
&= \max\Big([[-2,6] - 1]^L, [[-6,2] - 1]^L\Big), \\
&= \max(-3, -7) = -3. \\
z_2^{cv} &\equiv \max\left(\left[y_2^L X_2^C + x_2^L Y_2^C - x_2^L y_2^L\right]^L, \left[y_2^U X_2^C + x_2^U Y_2^C - x_2^U y_2^U\right]^L\right) \\
&= \max\Big([(0.7)[-2.5,1] + (0.7)[-2.5,1] - (0.7)(0.7)]^L, \\
&\qquad\qquad [(1)[-2.5,1] + (1)[-2.5,1] - (1)(1)]^L\Big), \\
&= \max\Big([[-1.75,0.7] + [-1.75,0.7] - 0.49]^L, [[-2.5,1] + [-2.5,1] - 1]^L\Big), \\
&= \max\Big([[-3.5,1.4] - 0.49]^L, [[-5,2] - 1]^L\Big), \\
&= \max(-3.99, -6) = -3.99.
\end{aligned}$$

With these definitions, $z_2^{cv} < z_1^{cv}$, so that inclusion monotonicity is violated.

Since $(\times, \mathbb{MR}^2, \mathbb{MR})$ has been proven to be inclusion monotonic when the Cut operation is used, but not otherwise, it follows that it is inclusion monotonic on $\mathbb{MR}_{\text{prop}}$ in either case. However, one cannot rely on always operating on $\mathbb{MR}_{\text{prop}}$. In the next example, it is shown that $(\times, \mathbb{MR}^2_{\text{prop}}, \mathbb{MR})$ itself may produce elements of $\mathbb{MR}$ that are not proper. That is, $\mathbb{MR}_{\text{prop}}$ is not closed under multiplication.

**Example 2.4.2.** Let $\mathcal{X} = \mathcal{Y} = ([-1,1],[-1,1])$. Clearly, $\mathcal{X}, \mathcal{Y} \in \mathbb{MR}_{\text{prop}}$. Using the notation of Definition 2.4.21,

$$\begin{aligned}
z^{cv} &= \max([y^L\bar{X}^C + x^L\bar{Y}^C - x^Ly^L]^L, [y^U\bar{X}^C + x^U\bar{Y}^C - x^Uy^U]^L),\\
&= \max([[-1,1] + [-1,1] - (-1)(-1)]^L, [[-1,1] + [-1,1] - (1)(1)]^L),\\
&= \max([[-2,2] - 1]^L, [[-2,2] - 1]^L),\\
&= \max([-3,1]^L, [-3,1]^L),\\
&= -3.
\end{aligned}$$

But $Z^B = [-1,1] \times [-1,1] = [-1,1]$. Therefore, $\mathcal{XY} \notin \mathbb{MR}_{\text{prop}}$.

We now define the univariate functions in $\mathcal{L}$ on $\mathbb{MR}$. The key contribution of McCormick's original work is the *McCormick composition rule*, which essentially shows how an inclusion monotonic relaxation function $(u, \mathbb{M}B, \mathbb{MR})$ can be constructed for any $(u, B, \mathbb{R}) \in \mathcal{L}$, provided that convex and concave relaxations for $u$ can be computed over a given interval $X$.

**Assumption 2.4.25.** For every $(u, B, \mathbb{R}) \in \mathcal{L}$, functions $u^{cv}, u^{cc} : \bar{B} \to \mathbb{R}$, where $\bar{B} \equiv \{(X, x) \in \mathbb{I}B \times B : x \in X\}$, and $x^{\min}, x^{\max} : \mathbb{I}B \to \mathbb{R}$ are known such that

1. For every $X \in \mathbb{I}B$, $u^{cv}(X, \cdot)$ and $u^{cc}(X, \cdot)$ are convex and concave relaxations of $u$ on $X$, respectively.

2. $x^{\min}(X)$ and $x^{\max}(X)$ are a minimum of $u^{cv}(X, \cdot)$ on $X$ and a maximum of $u^{cc}(X, \cdot)$ on $X$, respectively.

3. For any $X_1, X_2 \in \mathbb{IR}$ with $X_2 \subset X_1$, $u^{cv}(X_1, x) \leq u^{cv}(X_2, x)$ and $u^{cc}(X_1, x) \geq u^{cc}(X_2, x)$ for all $x \in X_2$.

4. $u^{cv}([x, x], x) = u^{cc}([x, x], x)$ for every $x \in B$.

Appropriate definitions of $u^{cv}$, $u^{cc}$, $x^{\min}$ and $x^{\max}$ are compiled for many univariate functions in §2.8. In most cases, it is simple to formulate the convex and concave envelopes of univariate functions. When these are used, Conditions 1, 3 and 4 of Assumption 2.4.25 hold by definition.

McCormick's composition rule now defines relaxation functions for the elements of $\mathcal{L}$ as follows.

**Definition 2.4.26.** For every $(u, B, \mathbb{R}) \in \mathcal{L}$, define $(u, \mathbb{M}B, \mathbb{M}\mathbb{R})$ by

$$u(\mathcal{X}) = \big(u(X^B), \big[u^{cv}(X^B, \mathrm{mid}(x^{cv}, x^{cc}, x^{\min}(X^B))), \\ u^{cc}(X^B, \mathrm{mid}(x^{cv}, x^{cc}, x^{\max}(X^B)))\big]\big),$$

where $u(X^B)$ is the value of $(u, \mathbb{I}B, \mathbb{I}\mathbb{R})$ at $X^B$.

Note that $\mathcal{X} \in \mathbb{M}B$ implies that either $x^{cv} \in X^B$ or $x^{cc} \in X^B$, or both. By definition $x^{\min}(X^B), x^{\max}(X^B) \in X^B$, so that, in both uses of the mid function above, at least two of the three arguments lie in $X^B$. It follows that the mid function chooses an element of $X^B$, and hence of $B$, in both cases, so that $u(\mathcal{X})$ is well-defined.

**Theorem 2.4.27.** *$(u, \mathbb{M}B, \mathbb{M}\mathbb{R})$ is a McCormick extension of $(u, B, \mathbb{R})$.*

*Proof.* Choose any $x \in B$. By Assumption 2.3.8, $u([x, x]) = [u(x), u(x)]$, and by Conditions 1 and 4 of Assumption 2.4.25, $u^{cv}([x, x], x) = u^{cc}([x, x], x) = u(x)$. □

Proving inclusion monotonicity requires the following lemma.

**Lemma 2.4.28.** *Suppose $g$ is a convex function on an interval $[x^L, x^U] \subset \mathbb{R}$ and $g$ attains its infimum at $x^{\min} \in [x^L, x^U]$. Then $g$ is monotone decreasing on $[x^L, x^{\min}]$ and monotone increasing on $[x^{\min}, x^U]$. Similarly, if $g$ is concave on $[x^L, x^U]$ and attains its supremum at $x^{\max} \in [x^L, x^U]$, then $g$ is monotone increasing on $[x^L, x^{\max}]$ and monotone decreasing on $[x^{\max}, x^U]$.*

*Proof.* The proof is elementary. □

**Theorem 2.4.29.** *$(u, \mathbb{M}B, \mathbb{M}\mathbb{R})$ is inclusion monotonic on $\mathbb{M}B$.*

*Proof.* Let $\mathcal{X}_1, \mathcal{X}_2 \in \mathbb{M}\mathbb{B}$, suppose that $\mathcal{X}_2 \subset \mathcal{X}_1$, and let $\mathcal{Z}_1 = u(\mathcal{X}_1)$ and $\mathcal{Z}_2 = u(\mathcal{X}_2)$. By Assumption 2.3.8, it suffices to show that $[z_2^{cv}, z_2^{cc}] \subset [z_1^{cv}, z_1^{cc}]$. It will be shown that $z_1^{cv} \leq z_2^{cv}$. The proof that $z_1^{cc} \geq z_2^{cc}$ is analogous.

Denote $h^{\min}(\mathscr{X}_i) = \text{mid}(x_i^{cv}, x_i^{cc}, x^{\min}(X_i^B))$, $i \in \{1, 2\}$. To show that $z_1^{cv} \leq z_2^{cv}$, It will be shown that

$$u^{cv}(X_1^B, h^{\min}(\mathscr{X}_1)) \leq u^{cv}(X_1^B, h^{\min}(\mathscr{X}_2)) \leq u^{cv}(X_2^B, h^{\min}(\mathscr{X}_2)). \tag{2.33}$$

It was argued above that $h^{\min}(\mathscr{X}_1) \in X_1^B$ and $h^{\min}(\mathscr{X}_2) \in X_2^B$. Since $X_2^B \subset X_1^B$, it follows that $h^{\min}(\mathscr{X}_2) \in X_1^B$, and hence the second inequality in (2.33) follows from Condition 3 of Assumption 2.4.25. It remains to show the first.

By definition, $x^{\min}(X_1^B)$ is a minimum of $u^{cv}(X_1^B, \cdot)$ on $X_1^B$. If $h^{\min}(\mathscr{X}_1) = x^{\min}(X_1^B)$, then the first inequality in (2.33) must be satisfied because $h^{\min}(\mathscr{X}_2) \in X_1^B$.

Suppose $h^{\min}(\mathscr{X}_1) = x_1^{cv}$. The definition of the mid function and the fact that $x_1^{cv} \leq x_1^{cc}$ require that $x^{\min}(X_1^B) \leq x_1^{cv} \leq x_1^{cc}$, so $h^{\min}(\mathscr{X}_1)$ is to the right of $x^{\min}(X_1^B)$. Since $u^{cv}(X_1^B, \cdot)$ is convex on $X_1^B$, it is monotonically increasing to the right of $x^{\min}(X_1^B)$ by Lemma 2.4.28. But $x_1^{cv} \leq x_2^{cv} \leq x_2^{cc}$, so if $h^{\min}(\mathscr{X}_2)$ is $x_2^{cv}$ or $x_1^{cc}$, then the first inequality in (2.33) holds. Further, if $h^{\min}(\mathscr{X}_2) = x^{\min}(X_2^B)$, the definition of the mid function requires that $x_2^{cv} \leq x^{\min}(X_2^B) \leq x_2^{cc}$, so $x^{\min}(X_2^B)$ is to the right of $h^{\min}(\mathscr{X}_1)$ and the first inequality in (2.33) still holds.

Now suppose that $h^{\min}(\mathscr{X}_1) = x_1^{cc}$. The definition of the mid function and the fact that $x_1^{cv} \leq x_1^{cc}$ require that $x^{\min}(X_1^B) \geq x_1^{cc} \geq x_1^{cv}$, so $h^{\min}(\mathscr{X}_1)$ is now to the left of $x^{\min}(X_1^B)$. By the convexity of $u^{cv}(X_1^B, \cdot)$, it is monotonically decreasing to the left of $x^{\min}(X_1^B)$ by Lemma 2.4.28. But, by hypothesis, $x_1^{cc} \geq x_2^{cc} \geq x_2^{cv}$, so if $h^{\min}(\mathscr{X}_2)$ is $x_2^{cv}$ or $x_2^{cc}$, then the first inequality in (2.33) holds. Further, if $h^{\min}(\mathscr{X}_2) = x^{\min}(X_2^B)$, the definition of the mid function requires that $x_2^{cv} \leq x^{\min}(X_2^B) \leq x_2^{cc}$, so $x^{\min}(X_2^B)$ is to the left of $h^{\min}(\mathscr{X}_1)$ and the first inequality in (2.33) still holds. □

**Theorem 2.4.30.** *$(u, \mathbb{M}B, \mathbb{MR})$ is coherently concave on $\mathbb{M}B$.*

*Proof.* Choose any coherent $\mathscr{X}_1, \mathscr{X}_2 \in \mathbb{M}B$ with common interval part $Q$, any $\lambda \in [0, 1]$ and let $\hat{\mathscr{X}} = \text{Conv}(\lambda, \mathscr{X}_1, \mathscr{X}_2)$. Let $\mathscr{Z}_1 = u(\mathscr{X}_1)$, $\mathscr{Z}_2 = u(\mathscr{X}_2)$ and $\hat{\mathscr{Z}} = u(\hat{\mathscr{X}})$. It will be shown that $\hat{z}^{cv} \leq \lambda z_1^{cv} + (1 - \lambda) z_2^{cv}$. The proof that $\hat{z}^{cc} \geq \lambda z_1^{cc} + (1 - \lambda) z_2^{cc}$ is analogous.

Let $w = u^{cv}(Q, \cdot)$ and $x^{\min} = x^{\min}(Q)$. Since $w$ is convex, it can be decomposed

[118] into a constant part, $A \equiv w(x^{\min})$, a convex, non-increasing part, $w_D(x) = w(\min(x, x^{\min})) - A$, and a convex, non-decreasing part, $w_I(x) = w(\max(x, x^{\min})) - A$, such that $w(x) = w_I(x) + w_D(x) + A$, $\forall x \in Q$.

Since $x_1^{cv} \leq x_1^{cc}$, there are three possible orderings of the numbers $x_1^{cv}$, $x_1^{cc}$ and $x^{\min}$. Assuming any of these, it is easy to see that one of the numbers $\max(x_1^{cv}, x^{\min})$ and $\min(x_1^{cc}, x^{\min})$ is equal to $x^{\min}$, and the other is equal to $\text{mid}(x_1^{cv}, x_1^{cc}, x^{\min})$. Then,

$$\begin{aligned} w(\text{mid}(x_1^{cv}, x_1^{cc}, x^{\min})) &= w(\max(x_1^{cv}, x^{\min})) + w(\min(x_1^{cc}, x^{\min})) - A \qquad (2.34) \\ &= w(\max(x_1^{cv}, q^L, x^{\min})) + w(\min(x_1^{cc}, q^U, x^{\min})) - A \\ &= w_I(\max(x_1^{cv}, q^L)) + w_D(\min(x_1^{cc}, q^U)) + A, \end{aligned}$$

and by the same arguments

$$w(\text{mid}(x_2^{cv}, x_2^{cc}, x^{\min})) = w_I(\max(x_2^{cv}, q^L)) + w_D(\min(x_2^{cc}, q^U)) + A, \qquad (2.35)$$

$$w(\text{mid}(\hat{x}^{cv}, \hat{x}^{cc}, x^{\min})) = w_I(\max(\hat{x}^{cv}, q^L)) + w_D(\min(\hat{x}^{cc}, q^U)) + A. \qquad (2.36)$$

Observing that $\max(\cdot, q^L)$ is convex on $\mathbb{R}$,

$$\max(\hat{x}^{cv}, q^L) \leq \lambda \max(x_1^{cv}, q^L) + (1 - \lambda) \max(x_2^{cv}, q^L),$$

and since $w_I^c$ is convex and non-decreasing

$$\begin{aligned} w_I(\max(\hat{x}^{cv}, q^L)) &\leq w_I([\lambda \max(x_1^{cv}, q^L) + (1 - \lambda) \max(x_2^{cv}, q^L)]), \\ &\leq \lambda w_I(\max(x_1^{cv}, q^L)) + (1 - \lambda) w_I(\max(x_2^{cv}, q^L)). \end{aligned}$$

Applying analogous arguments to the term $w_D(\min(\hat{x}^{cc}, q^U))$, it follows that

$$w_D(\min(\hat{x}^{cc}, q^U)) \leq \lambda w_D(\min(x_1^{cc}, q^U)) + (1 - \lambda) w_D(\min(x_2^{cc}, q^U)).$$

Now, applying (2.34), (2.35) and (2.36),

$$\begin{aligned}
w(\mathrm{mid}(\hat{x}^{cv}, \hat{x}^{cc}, x^{\min})) &\le \lambda w_I(\max(x_1^{cv}, q^L)) + (1-\lambda) w_I(\max(x_2^{cv}, q^L)) \\
&\quad + \lambda w_D(\min(x_1^{cc}, q^U)) + (1-\lambda) w_D(\min(x_2^{cc}, q^U)) + A \\
&= \lambda[w_I(\max(x_1^{cv}, q^L)) + w_D(\min(x_1^{cc}, q^U)) + A] \\
&\quad + (1-\lambda)[w_I(\max(x_2^{cv}, q^L)) + w_D(\min(x_2^{cc}, q^U)) + A] \\
&= \lambda w(\mathrm{mid}(x_1^{cv}, x_1^{cc}, x^{\min})) + (1-\lambda) w(\mathrm{mid}(x_2^{cv}, x_2^{cc}, x^{\min})).
\end{aligned}$$

But this last inequality is exactly $\hat{z}^{cv} \le \lambda z_1^{cv} + (1-\lambda) z_2^{cv}$. □

By Theorems 2.4.27, 2.4.29, 2.4.30 and 2.4.14, it now follows that each $(u, \mathbb{M}B, \mathbb{MR})$ is a relaxation function of the corresponding $(u, B, \mathbb{R}) \in \mathcal{L}$. As with the McCormick multiplication operation, it can be shown directly that $(u, \mathbb{M}B, \mathbb{MR})$ is a relaxation function without proving that it is a McCormick extension or that it is inclusion monotonic. This is a more standard development, and it does not require Conditions 3 and 4 in Assumption 2.4.25. However, for application to global optimization, both inclusion monotonicity and the condition for degenerate inputs dictated by the definition of a McCormick extension are very important, and will not necessarily hold without these additional assumptions.

We now define the natural McCormick extension of a $\mathcal{L}$-computational sequence.

**Definition 2.4.31.** For every $\mathcal{L}$-computational sequence $(\mathcal{S}, \pi_o)$, with $n_i$ inputs and $n_o$ outputs, define the *sequence of relaxation factors* $\{(\mathcal{V}_k, \mathcal{D}_k, \mathbb{MR})\}_{k=1}^{n_f}$ where

1. For all $k = 1, \ldots, n_i$, $\mathcal{D}_k = \mathbb{MR}^{n_i}$ and $\mathcal{V}_k(\mathcal{X}) = \mathcal{X}_k$, $\forall \mathcal{X} \in \mathcal{D}_k$,
2. For all $k = n_i+1, \ldots, n_f$, $\mathcal{D}_k = \{\mathcal{X} \in \mathcal{D}_{k-1} : \pi_k \circ (\mathcal{V}_1(\mathcal{X}), \ldots, \mathcal{V}_{k-1}(\mathcal{X})) \in \mathbb{M}B_k\}$ and $\mathcal{V}_k(\mathcal{X}) = o_k \circ \pi_k \circ (\mathcal{V}_1(\mathcal{X}), \ldots, \mathcal{V}_{k-1}(\mathcal{X}))$, $\forall \mathcal{X} \in \mathcal{D}_k$.

The *natural McCormick extension* of $(\mathcal{S}, \pi_o)$ is the function $(\mathcal{F}_\mathcal{S}, \mathcal{D}_\mathcal{S}, \mathbb{MR}^{n_o})$ defined by $\mathcal{D}_\mathcal{S} \equiv \mathcal{D}_{n_f}$ and $\mathcal{F}(\mathcal{X}) = \pi_o \circ (\mathcal{V}_1(\mathcal{X}), \ldots, \mathcal{V}_{n_f}(\mathcal{X}))$, $\forall \mathcal{X} \in \mathcal{D}_\mathcal{S}$.

**Theorem 2.4.32.** *Let $(\mathcal{S}, \pi_o)$ be a $\mathcal{L}$-computational sequence with natural function $(\mathbf{f}_\mathcal{S}, D_\mathcal{S}, \mathbb{R}^{n_o})$. The natural McCormick extension $(\mathcal{F}_\mathcal{S}, \mathcal{D}_\mathcal{S}, \mathbb{MR}^{n_o})$ is a McCormick* *extension of* $(\mathbf{f}_{\mathcal{S}}, D_{\mathcal{S}}, \mathbb{R}^{n_o})$*, and it is coherently concave and inclusion monotonic on* $\mathcal{D}_{\mathcal{S}}$*.*

*Proof.* Consider the sequence of factors $\{(v_k, D_k, \mathbb{R})\}_{k=1}^{n_f}$ and the sequence of relaxation factors $\{(\mathcal{V}_k, \mathcal{D}_k, \mathbb{MR})\}_{k=1}^{n_f}$. Choose any $K \in \{1, \dots, n_f\}$ and suppose that $(\mathcal{V}_k, \mathcal{D}_k, \mathbb{MR})$ is a McCormick extension of $(v_k, D_k, \mathbb{R})$, and coherently concave and inclusion monotonic on $\mathcal{D}_k$, for all $k \in \{1, \dots, K-1\}$. If $K \leq n_i + 1$, this is true because, for any $k < K$, $\mathcal{D}_k = \mathbb{MR}^{n_i}$ is a McCormick extension of $D_k = \mathbb{R}^{n_i}$, $\mathcal{V}_k(([\mathbf{x}, \mathbf{x}], [\mathbf{x}, \mathbf{x}])) = ([x_k, x_k], [x_k, x_k]) = ([v_k(\mathbf{x}), v_k(\mathbf{x})], [v_k(\mathbf{x}), v_k(\mathbf{x})])$ for any $\mathbf{x} \in D_k$, and $\mathcal{V}_k$ is trivially inclusion monotonic and coherently concave on $\mathbb{MR}^{n_i}$.

Now, $(v_1, \dots, v_{K-1})$ is a well-defined mapping from $D_{K-1}$ into $\mathbb{R}^{K-1}$. By the inductive hypothesis, $(\mathcal{V}_1, \dots, \mathcal{V}_{K-1})$, as a mapping from $\mathcal{D}_{K-1}$ into $\mathbb{MR}^{K-1}$, is a McCormick extension of $(v_1, \dots, v_{K-1})$, and is inclusion monotonic and coherently concave on $\mathcal{D}_{K-1}$. It follows that $\pi_k \circ (\mathcal{V}_1, \dots, \mathcal{V}_{K-1})$ is a McCormick extension of $\pi_k \circ (v_1, \dots, v_{K-1})$, and is inclusion monotonic and coherently concave on $\mathcal{D}_{K-1}$. By Theorems 2.4.20, 2.4.22 and 2.4.27, $(o_K, \mathbb{M}B_K, \mathbb{MR})$ is a McCormick extension of $(o_K, B_K, \mathbb{R})$, and is coherently concave and inclusion monotonic on $\mathbb{M}B_K$ by Theorems 2.4.20, 2.4.23, 2.4.29, 2.4.24 and 2.4.30. Then, Lemma 2.4.17 shows that $(\mathcal{V}_K, \mathcal{D}_K, \mathbb{MR})$ is a McCormick extension of $(v_K, D_K, \mathbb{R})$, and Lemma 2.4.15 shows that it is coherently concave and inclusion monotonic on $\mathcal{D}_K$. By induction, this holds for every $K \in \{1, \dots, n_f\}$, and the theorem follows immediately from the definition of $(\mathcal{F}_{\mathcal{S}}, \mathcal{D}_{\mathcal{S}}, \mathbb{IR}^{n_o})$. □

Similar to the situation for natural interval extensions, it is generally not possible to define a natural McCormick extension on all of $\mathbb{M}D_{\mathcal{S}}$. However, the situation for natural McCormick extensions is no more restrictive than that for natural interval extensions because the domain $\mathbb{M}B$ of a univariate function only restricts the interval part of its argument. In particular, it is easily seen that any $X \in \mathfrak{D}_{\mathcal{S}}$ is represented in $\mathcal{D}_{\mathcal{S}}$. Moreover, inclusion monotonicity of the relaxation factors immediately implies the following.

**Lemma 2.4.33.** *Let $(\mathcal{S}, \pi_o)$ be a $\mathcal{L}$-computational sequence with natural McCormick extension $(\mathcal{F}_{\mathcal{S}}, \mathcal{D}_{\mathcal{S}}, \mathbb{MR}^{n_o})$. If an interval $X \in \mathbb{IR}^{n_i}$ is represented in $\mathcal{D}_{\mathcal{S}}$, then every element of $\mathbb{I}X$ is represented in $\mathcal{D}_{\mathcal{S}}$; i.e. $\mathbb{M}X \subset \mathcal{D}_{\mathcal{S}}$.*

**Definition 2.4.34.** Let $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ be a $\mathcal{L}$-factorable function. Then, for any $\mathcal{L}$-computational sequence describing $\mathbf{f}$, the natural McCormick extension $(\mathcal{F}_{\mathcal{S}}, \mathcal{D}_{\mathcal{S}}, \mathbb{MR}^m)$ is called a natural McCormick extension of $\mathbf{f}$.

It is apparent from Theorem 2.4.32 that a natural McCormick extension of a $\mathcal{L}$-factorable function is indeed a McCormick extension, and is coherently concave and inclusion monotonic on $\mathcal{D}_{\mathcal{S}}$. More importantly, it is a relaxation function for $\mathbf{f}$ on $\mathcal{D}_{\mathcal{S}} \cap \mathbb{M}D$. In fact, the standard McCormick relaxations of $\mathbf{f}$ can be defined from a natural McCormick extension of $\mathbf{f}$ exactly as in Lemma 2.4.11 (see §2.6). Moving forward, the notation $(\{\mathbf{f}\}, \mathcal{D}, \mathbb{MR}^m)$ will be used to denote a natural McCormick extension of $(\mathbf{f}, D, \mathbb{R}^m)$. Furthermore, we denote $\{\mathbf{f}\}(\mathcal{X}) = (F^B(\mathcal{X}), [\{\mathbf{f}\}^{cv}(\mathcal{X}), \{\mathbf{f}\}^{cc}(\mathcal{X})])$.

As with natural interval extensions, the evaluation of the natural McCormick extension of a sequence of computations can be easily automated and is only marginally more computationally demanding than executing the same sequence of computations in real arithmetic. Throughout this thesis, natural McCormick extensions are computed using the library `MC++` (`http://www3.imperial.ac.uk/people/b.chachuat/research`). `MC++` is the successor of `libMC`, which is described in detail in [122].

## 2.5 Regularity of Functions on $\mathbb{IR}^n$ and $\mathbb{MR}^n$

It should not be surprising that the regularity of a $\mathcal{L}$-factorable function, i.e., whether it is continuous, Lipschitz, differentiable, etc., depends on the corresponding properties of the univariate functions in $\mathcal{L}$. In later chapters, it will be very useful to recognize that natural interval and McCormick extensions also enjoy some regularity properties, again inherited from the properties of the univariate interval and McCormick extensions $(u, \mathbb{I}B, \mathbb{IR})$ and $(u, \mathbb{M}B, \mathbb{MR})$. In this section, several notions of regularity are extended to include functions to or from the sets $\mathbb{IR}^n$ and $\mathbb{MR}^n$

and shown to hold for factorable functions, natural interval extensions, and natural McCormick extensions under mild assumptions. Among these, the piecewise differentiability of natural interval extensions and all properties of natural McCormick extensions are new contributions. The properties of factorable functions are apparent and a Lipschitz condition for natural interval extensions has been previously demonstrated in [131].

### 2.5.1 $\mathbb{IR}^n$ and $\mathbb{MR}^n$ as Metric Spaces

Let $Z, Y \subset \mathbb{R}^n$. The *Hausdorff distance* between $Z$ and $Y$, induced by the infinity-norm distance on $\mathbb{R}^n$, is defined by

$$d_H(Z, Y) = \max\left(\sup_{\mathbf{y}\in Y}\inf_{\mathbf{z}\in Z}\|\mathbf{z}-\mathbf{y}\|_\infty, \sup_{\mathbf{z}\in Z}\inf_{\mathbf{y}\in Y}\|\mathbf{z}-\mathbf{y}\|_\infty\right). \tag{2.37}$$

Let $\mathbb{KR}^n$ denote the set of all nonempty compact subsets of $\mathbb{R}^n$. It is well-known that $\mathbb{KR}^n$ is a complete metric space under $d_H$. In this context $d_H$ will be referred to as the *Hausdorff metric*. Since $\mathbb{IR}^n \subset \mathbb{KR}^n$, it follows that $\mathbb{IR}^n$ is also a metric space under $d_H$. If $Z, Y \in \mathbb{IR}^n$, $Z \equiv [\mathbf{z}^L, \mathbf{z}^U]$ and $Y \equiv [\mathbf{y}^L, \mathbf{y}^U]$, then the Hausdorff metric on $\mathbb{IR}^n$ is equivalently expressed as

$$d_H(Z, Y) = \max\left(\max_i |z_i^L - y_i^L|, \max_i |z_i^U - y_i^U|\right). \tag{2.38}$$

Recall that any subset of a metric space is itself a metric space with the same metric. Then, since $\mathbb{MR}^n$ is a subset of $\mathbb{IR}^{2n}$, it too is a metric space under the distance

$$\begin{aligned} d_M(\mathcal{Z}, \mathcal{Y}) &= d_H(Z^B \times Z^C, Y^B \times Y^C) \\ &= \max\left(d_H(Z^B, Y^B), d_H(Z^C, Y^C)\right), \quad \forall \mathcal{Z}, \mathcal{Y} \in \mathbb{MR}^n. \end{aligned} \tag{2.39}$$

In general, the set $\mathbb{R}^k \times \mathbb{IR}^n \times \mathbb{MR}^m$ is a metric space, for any $k, n, m \in \mathbb{N}$, with the metric

$$d_\infty\left((\mathbf{x}_1, Z_1, \mathcal{Y}_1), (\mathbf{x}_2, Z_2, \mathcal{Y}_2)\right) = \max\left(\|\mathbf{x}_1 - \mathbf{x}_2\|_\infty, d_H(Z_1, Z_2), d_M(\mathcal{Y}_1, \mathcal{Y}_2)\right). \quad (2.40)$$

Thus, open and closed subsets of $\mathbb{R}^k \times \mathbb{IR}^n \times \mathbb{MR}^m$ are defined in the standard way. Moreover, for functions mapping to and/or from this space, continuity is defined by the standard $\epsilon$-$\delta$ condition, or equivalently by the condition that the inverse images of open sets are open.

The practical reason for viewing $\mathbb{R}^k \times \mathbb{IR}^n \times \mathbb{MR}^m$ as a metric space is to analyze the regularity of natural interval and McCormick extensions. In later chapters, some developments will require that interval and/or McCormick extensions are continuous, or even Lipschitz, with respect to the bound and relaxation values taken as input. Consider an interval function $F : \mathbb{IR}^n \to \mathbb{IR}^m$. When discussing the regularity of $F$, it is often convenient to think about the dependence of, say, the lower bound $F^L$ on the real vectors $\mathbf{z}^L$ and $\mathbf{z}^U$ describing the input interval $Z \equiv [\mathbf{z}^L, \mathbf{z}^U]$. Other times, it will be more convenient to think of continuity on the metric space $\mathbb{IR}^n$ according to the standard definition. In the remainder of this section, it is shown that these notions are equivalent, so that we may use whichever is most convenient for the task at hand. The continuity of natural interval and McCormick extensions is demonstrated in §2.5.5 and §2.5.6, respectively.

**Definition 2.5.1.** Define the set

$$\begin{aligned}\mathbb{H}^{(k,n,m)} \equiv \{(\mathbf{x}, \mathbf{z}^L, \mathbf{z}^U, \mathbf{y}^L, \mathbf{y}^U, \mathbf{y}^{cv}, \mathbf{y}^{cc}) \in \mathbb{R}^{k+2n+4m} :& \qquad (2.41)\\ \mathbf{z}^L \leq \mathbf{z}^U,\ \mathbf{y}^L \leq \mathbf{y}^U,\ \mathbf{y}^{cv} \leq \mathbf{y}^{cc},\ [\mathbf{y}^L, \mathbf{y}^U] \cap [\mathbf{y}^{cv}, \mathbf{y}^{cc}] \neq \emptyset\}.&\end{aligned}$$

Furthermore, let $i_\mathbb{R} : \mathbb{R}^k \times \mathbb{IR}^n \times \mathbb{MR}^m \to \mathbb{H}^{(k,n,m)}$ be defined by

$$i_\mathbb{R}(\mathbf{x}, Z, \mathcal{Y}) = (\mathbf{x}, \mathbf{z}^L, \mathbf{z}^U, \mathbf{y}^L, \mathbf{y}^U, \mathbf{y}^{cv}, \mathbf{y}^{cc}), \quad (2.42)$$

where $[\mathbf{z}^L, \mathbf{z}^U] = Z$ and $([\mathbf{y}^L, \mathbf{y}^U], [\mathbf{y}^{cv}, \mathbf{y}^{cc}]) = \mathcal{Y}$.

The mapping $i_{\mathbb{R}}$ identifies its argument with an element of a Euclidean space in the natural way. It is defined as a mapping into $\mathbb{H}^{(k,n,m)}$ so that it is bijective, and hence invertible. The following lemma follows directly from this definition.

**Lemma 2.5.2.** $(i_{\mathbb{R}}, \mathbb{R}^k \times \mathbb{IR}^n \times \mathbb{MR}^m, \mathbb{H}^{(k,n,m)})$ *is bijective and isometric; i.e.,*

$$\|i_{\mathbb{R}}(\mathbf{x}_1, Z_1, \mathcal{Y}_1) - i_{\mathbb{R}}(\mathbf{x}_2, Z_2, \mathcal{Y}_2)\|_\infty = d_\infty\left((\mathbf{x}_1, Z_1, \mathcal{Y}_1), (\mathbf{x}_2, Z_2, \mathcal{Y}_2)\right), \tag{2.43}$$

*for any* $(\mathbf{x}_1, Z_1, \mathcal{Y}_1), (\mathbf{x}_2, Z_2, \mathcal{Y}_2) \in \mathbb{R}^k \times \mathbb{IR}^n \times \mathbb{MR}^m$*.*

**Theorem 2.5.3.** *Let* $\mathcal{M} : \mathcal{D} \subset \mathbb{R}^k \times \mathbb{IR}^n \times \mathbb{MR}^m \to \mathbb{R}^l \times \mathbb{IR}^q \times \mathbb{MR}^r$*. The following conditions are equivalent:*

1. $\mathcal{M}$ *is continuous on* $\mathcal{D}$*.*
2. $i_{\mathbb{R}} \circ \mathcal{M}$ *is continuous on* $\mathcal{D}$*.*
3. $\mathcal{M} \circ i_{\mathbb{R}}^{-1}$ *is continuous on* $Q \equiv i_{\mathbb{R}}(\mathcal{D})$*.*
4. $i_{\mathbb{R}} \circ \mathcal{M} \circ i_{\mathbb{R}}^{-1}$ *is continuous on* $Q \equiv i_{\mathbb{R}}(\mathcal{D})$*.*

*Proof.* Since $i_{\mathbb{R}}$ is an isometry, it follows that both $i_{\mathbb{R}}$ and $i_{\mathbb{R}}^{-1}$ are continuous. Then, since the composition of continuous functions is continuous, 1 implies 2, 3 and 4, 2 implies 4, and 3 implies 4. To prove the remaining results, it suffices to show that 4 implies 1.

Suppose that 4 holds but 1 does not. Then there exists an open set $O \subset \mathbb{R}^l \times \mathbb{IR}^q \times \mathbb{MR}^r$ such that $\mathcal{M}^{-1}(O)$ is not open in $\mathcal{D}$. Consider the image

$$i_{\mathbb{R}}(O) = \{i_{\mathbb{R}}(a) : a \in O\} = \{b \in \mathbb{H}^{(l,q,r)} : i_{\mathbb{R}}^{-1}(b) \in O\}. \tag{2.44}$$

By the last equality, it follows that $i_{\mathbb{R}}(O)$ is open in $\mathbb{H}^{(l,q,r)}$ because it is the inverse image of the open set $O$ under the continuous mapping $i_{\mathbb{R}}^{-1}$. Then, by 4, the inverse image $(i_{\mathbb{R}} \circ \mathcal{M} \circ i_{\mathbb{R}}^{-1})^{-1}(i_{\mathbb{R}}(O))$ is open in $Q$. But

$$(i_{\mathbb{R}} \circ \mathcal{M} \circ i_{\mathbb{R}}^{-1})^{-1}(i_{\mathbb{R}}(O)) = i_{\mathbb{R}}(\mathcal{M}^{-1}(i_{\mathbb{R}}^{-1}(i_{\mathbb{R}}(O)))) = i_{\mathbb{R}}(\mathcal{M}^{-1}(O)), \tag{2.45}$$

so that $i_{\mathbb{R}}(\mathcal{M}^{-1}(O))$ is open in $Q$. Since $i_{\mathbb{R}}$ is continuous as a mapping from $\mathcal{D}$ into $Q = i_{\mathbb{R}}(\mathcal{D})$, it follows that $i_{\mathbb{R}}^{-1}(i_{\mathbb{R}}(\mathcal{M}^{-1}(O)))$ is open in $\mathcal{D}$. But $i_{\mathbb{R}}^{-1}(i_{\mathbb{R}}(\mathcal{M}^{-1}(O))) = \mathcal{M}^{-1}(O)$, so this contradicts the hypothesis that $\mathcal{M}^{-1}(O)$ is not open in $\mathcal{D}$. □

Returning to the example of the function $F : \mathbb{IR}^n \to \mathbb{IR}^m$, Theorem 2.5.3 shows that it is equivalent to speak of the continuity of $F$ as a mapping from $\mathbb{IR}^n$ to $\mathbb{IR}^m$ (Condition 1) and the continuity of $F^L$ and $F^U$ as functions $\mathbf{z}^L$ and $\mathbf{z}^U$ on the set $\{(\mathbf{z}^L, \mathbf{z}^U) \in \mathbb{R}^n \times \mathbb{R}^n : \mathbf{z}^L \leq \mathbf{z}^U\}$ (Condition 4).

### 2.5.2 Lipschitz and Locally Lipschitz Functions

In this section, Lipschitz and locally Lipschitz conditions are defined and some standard results are presented. For functions to and/or from the spaces $\mathbb{IR}^n$ or $\mathbb{MR}^n$, it is shown that the analogues of Theorem 2.5.3 hold for these Lipschitz properties as well.

**Definition 2.5.4.** Let $(X, d_X)$ and $(Y, d_Y)$ be metric spaces and let $f : X \to Y$. $f$ is *Lipschitz* on $X$ if $\exists L > 0$ such that

$$d_Y(f(x_1), f(x_2)) \leq L d_X(x_1, x_2), \quad \forall x_1, x_2 \in X. \tag{2.46}$$

$f$ is *locally Lipschitz* on $X$ if, for every $\hat{x} \in X$, $\exists \eta, L > 0$ such that

$$d_Y(f(x_1), f(x_2)) \leq L d_X(x_1, x_2), \quad \forall x_1, x_2 \in B_\eta(\hat{x}), \tag{2.47}$$

where $B_\eta(\hat{x}) \equiv \{x \in X : d_X(x, \hat{x}) < \eta\}$ is the open ball in $X$ of radius $\eta$ about $\hat{x}$.

If $f$ is Lipschitz on $X$, then it is locally Lipschitz on $X$. Moreover, if $f$ is locally Lipschitz on $X$, then it is uniformly continuous on $X$. Neither of the converses are true. Affine functions are Lipschitz, as are finite sums of Lipschitz functions, and the same is true of locally Lipschitz functions. Compositions and products are discussed below.

Recall that any subset of a metric space is again a metric space with the same metric. Then, it is sensible for a function to be Lipschitz or locally Lipschitz on a subset. In the latter case, however, one must take care that the open ball is interpreted as open with respect to the subset. For example, let $(Y, d_Y)$ be a metric space and let $f : \mathbb{R}^n \to Y$. Then, $f$ is locally Lipschitz on $E \subset \mathbb{R}^n$ if, for every $\hat{\mathbf{x}} \in E$, $\exists \eta, L > 0$ such that

$$d_Y(f(\mathbf{x}_1), f(\mathbf{x}_2)) \leq L\|\mathbf{x}_1 - \mathbf{x}_2\|, \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in E \cap B_\eta(\hat{\mathbf{x}}). \tag{2.48}$$

Here, $B_\eta(\hat{\mathbf{x}}) = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \hat{\mathbf{x}}\| < \eta\}$ is the standard open ball in $\mathbb{R}^n$ of radius $\eta$ about $\hat{\mathbf{x}}$, so that $E \cap B_\eta(\hat{\mathbf{x}}) = \{\mathbf{x} \in E : \|\mathbf{x} - \hat{\mathbf{x}}\| < \eta\}$ is the open ball *in the metric space* $E$ of radius $\eta$ about $\hat{\mathbf{x}}$.

**Lemma 2.5.5.** *Let $(X, d_X)$ and $(Y, d_Y)$ be metric spaces and let $f : X \to Y$ be locally Lipschitz on $X$. Then $f$ is locally Lipschitz on any $E \subset X$.*

*Proof.* Choose any $\hat{x} \in E$. Since $f$ is locally Lipschitz on $X$, $\exists \eta, L > 0$ such that $d_Y(f(x_1), f(x_2)) \leq L d_X(x_1, x_2)$ for all $x_1, x_2 \in \{x \in X : d_X(x, \hat{x}) < \eta\}$. But since $E \subset X$, the same inequality must hold for all $x_1, x_2 \in \{x \in E : d_X(x, \hat{x}) < \eta\}$. □

It is very simple to show that the composition of two Lipschitz functions is Lipschitz. This also holds true for locally Lipschitz functions.

**Theorem 2.5.6.** *Let $(X, d_X)$, $(Y, d_Y)$ and $(Z, d_Z)$ be metric spaces and let $f : E_X \subset X \to Y$ and $g : E_Y \subset Y \to Z$ be locally Lipschitz on $E_X$ and $E_Y$, respectively. Then $g \circ f$ is locally Lipschitz on $E_{XY} \equiv \{x \in E_X : f(x) \in E_Y\}$.*

*Proof.* Choose any $\hat{x} \in E_{XY}$ and let $\hat{y} = f(\hat{x}) \in E_Y$. Since $g$ is locally Lipschitz on $E_Y$, $\exists \eta_g, L_g > 0$ such that $d_Z(g(y_1), g(y_2)) \leq L_g d_Y(y_1, y_2)$ for all $y_1, y_2 \in B_{\eta_g}(\hat{y})$, where $B_{\eta_g}(\hat{y}) = \{y \in E_Y : d_Y(y, \hat{y}) < \eta_g\}$. Since $f : E_{XY} \to E_Y$ is continuous, $Q = f^{-1}(B_{\eta_g}(\hat{y}))$ is open in $E_{XY}$ and contains $\hat{x}$. Since $f$ is locally Lipschitz on $E_{XY}$, $\exists \eta_f, L_f > 0$ such that $d_Y(f(x_1), f(x_2)) \leq L_f d_X(x_1, x_2)$ for all $x_1, x_2 \in B_{\eta_f}(\hat{x})$, where $B_{\eta_f}(\hat{x}) = \{x \in E_{XY} : d(x, \hat{x}) < \eta_f\}$. Since $Q$ is open in $E_{XY}$, $\eta \in (0, \eta_f]$ can be chosen small enough that the open ball $B_\eta(\hat{x})$ (again in $E_{XY}$) is a subset of $Q$, and hence $f(B_\eta(\hat{x})) \subset B_{\eta_g}(\hat{y})$. Then, for any $x_1, x_2 \in B_\eta(\hat{x})$

$$d_Z(g \circ f(x_1), g \circ f(x_2)) \leq L_g d_Y(f(x_1), f(x_2)) \leq L_g L_f d_X(x_1, x_2). \tag{2.49}$$

□

The following theorem is a very useful fact about locally Lipschitz functions. Let a *compact neighborhood* of $x$ in $X$ be a compact subset of $X$ with $x$ in its interior. Recall that a metric space $(X, d_X)$ is said to be *locally compact* if there exists a compact neighborhood of every $x \in X$.

**Theorem 2.5.7.** *Let $(X, d_X)$ and $(Y, d_Y)$ be metric spaces and let $f : X \to Y$. If $f$ is locally Lipschitz on $X$ then $f$ is Lipschitz on every compact $K \subset X$. If $(X, d_X)$ is locally compact, then the converse holds.*

Clearly, $\mathbb{R}^n$ is locally compact, since for any $\mathbf{x} \in \mathbb{R}^n$, the closure of any open ball about $\mathbf{x}$ in $\mathbb{R}^n$, $\overline{B}_\epsilon(\mathbf{x})$, is a compact neighborhood of $\mathbf{x}$ in $\mathbb{R}^n$. On the other hand, not all subsets $E \subset \mathbb{R}^n$ are locally compact metric spaces. If $E$ is open, then it is locally compact because $\overline{B}_\epsilon(\mathbf{x})$ is a compact neighborhood of $\mathbf{x}$ in $E$ for small enough $\epsilon > 0$. If $E$ is closed, it is also locally compact because $E \cap \overline{B}_\epsilon(\mathbf{x})$ is a compact neighborhood of $\mathbf{x}$ in $E$. If $E$ is neither open nor closed, then it may not be locally compact because $E \cap \overline{B}_\epsilon(\mathbf{x})$ may fail to be closed and its closure in $\mathbb{R}^n$ may fail to be a subset of $E$.

**Corollary 2.5.8.** *Let $(Y, d_Y)$ be a metric space and let $f : E \subset \mathbb{R}^n \to Y$. If $f$ is locally Lipschitz on $E$ then $f$ is Lipschitz on every compact $K \subset E$. If $E$ is either open or closed, then the converse holds.*

In general, products of Lipschitz functions are not Lipschitz. However, we have the following theorem.

**Theorem 2.5.9.** *Let $(X, d_X)$ be a metric space and let $(Y, d_Y)$ be a normed space, and hence a metric space with $d_Y(y_1, y_2) = \|y_1 - y_2\|$, $\forall y_1, y_2 \in Y$. If $f, g : X \to Y$ are Lipschitz and bounded on $X$, then $fg$ is Lipschitz on $X$.*

*Proof.* Choose any $x_1, x_2 \in X$. Then, using the triangle inequality,

$$\begin{aligned}
&\|f(x_1)g(x_1) - f(x_2)g(x_2)\| \qquad (2.50)\\
&\qquad \leq \|f(x_1)g(x_1) - f(x_1)g(x_2)\| + \|f(x_1)g(x_2) - f(x_2)g(x_2)\|,\\
&\qquad \leq \|f(x_1)\|\|g(x_1) - g(x_2)\| + \|g(x_2)\|\|f(x_1) - f(x_2)\|,\\
&\qquad \leq \sup_{x\in X} \|f(x)\| L_g d_X(x_1, x_2) + \sup_{x\in X} \|g(x)\| L_f d_X(x_1, x_2),\\
&\qquad = \left[\sup_{x\in X} \|f(x)\| L_g + \sup_{x\in X} \|g(x)\| L_f\right] d_X(x_1, x_2).
\end{aligned}$$

□

**Corollary 2.5.10.** *Let $(X, d_X)$ be a locally compact metric space and let $(Y, d_Y)$ be a normed space, and hence a metric space with $d_Y(y_1, y_2) = \|y_1 - y_2\|$, $\forall y_1, y_2 \in Y$. If $f, g : X \to Y$ are locally Lipschitz on $X$, then $fg$ is locally Lipschitz on $X$.*

*Proof.* Choose any $x \in X$. Since $X$ is locally compact, there exists a compact neighborhood $K_x$ of $x$ in $X$. Since both $f$ and $g$ are locally Lipschitz on $X$, there exists $\eta > 0$ sufficiently small that both $f$ and $g$ are Lipschitz on $B_\eta(x)$. Choosing $\eta > 0$ small enough that $B_\eta(x) \subset K_x$, continuity ensures that $f$ and $g$ are also bounded on $B_\eta(x)$. Then Theorem 2.5.9 implies that $fg$ is Lipschitz on $B_\eta(x)$. □

Since $\mathbb{R}^k \times \mathbb{IR}^n \times \mathbb{MR}^m$ is a metric space, Lipschitz and locally Lipschitz functions are well-defined on this space. The following theorems extend the observations of Theorem 2.5.3 to these classes of functions as well.

**Theorem 2.5.11.** *Let $\mathcal{M} : \mathcal{D} \subset \mathbb{R}^k \times \mathbb{IR}^n \times \mathbb{MR}^m \to \mathbb{R}^l \times \mathbb{IR}^q \times \mathbb{MR}^r$. The following conditions are equivalent:*

1. *$\mathcal{M}$ is Lipschitz on $\mathcal{D}$.*

2. *$i_{\mathbb{R}} \circ \mathcal{M}$ is Lipschitz on $\mathcal{D}$.*

3. *$\mathcal{M} \circ i_{\mathbb{R}}^{-1}$ is Lipschitz on $Q \equiv i_{\mathbb{R}}(\mathcal{D})$.*

4. *$i_{\mathbb{R}} \circ \mathcal{M} \circ i_{\mathbb{R}}^{-1}$ is Lipschitz on $Q \equiv i_{\mathbb{R}}(\mathcal{D})$.*

*Proof.* Since $i_{\mathbb{R}}$ is an isometry, it follows that both $i_{\mathbb{R}}$ and $i_{\mathbb{R}}^{-1}$ are Lipschitz. Then, since the composition of Lipschitz functions is Lipschitz, 1 implies 2, 3 and 4, 2 implies 4, and 3 implies 4. Then, it suffices to show that 4 implies 1.

Suppose that 4 holds and 1 does not. By 4, $\exists L > 0$ such that

$$\|i_{\mathbb{R}} \circ \mathcal{M} \circ i_{\mathbb{R}}^{-1}(\tilde{a}) - i_{\mathbb{R}} \circ \mathcal{M} \circ i_{\mathbb{R}}^{-1}(\hat{a})\|_{\infty} \leq L\|\tilde{a} - \hat{a}\|_{\infty}, \quad \forall \tilde{a}, \hat{a} \in Q. \tag{2.51}$$

Since 1 fails, there exist points $(\tilde{\mathbf{x}}, \tilde{Z}, \tilde{\mathcal{Y}})$ and $(\hat{\mathbf{x}}, \hat{Z}, \hat{\mathcal{Y}})$ in $\mathcal{D}$ such that

$$d_{\infty}\left(\mathcal{M}(\tilde{\mathbf{x}}, \tilde{Z}, \tilde{\mathcal{Y}}), \mathcal{M}(\hat{\mathbf{x}}, \hat{Z}, \hat{\mathcal{Y}})\right) > L d_{\infty}\left((\tilde{\mathbf{x}}, \tilde{Z}, \tilde{\mathcal{Y}}), (\hat{\mathbf{x}}, \hat{Z}, \hat{\mathcal{Y}})\right). \tag{2.52}$$

By (2.43), this implies that

$$\|i_{\mathbb{R}} \circ \mathcal{M}(\tilde{\mathbf{x}}, \tilde{Z}, \tilde{\mathcal{Y}}) - i_{\mathbb{R}} \circ \mathcal{M}(\hat{\mathbf{x}}, \hat{Z}, \hat{\mathcal{Y}})\|_{\infty} > L d_{\infty}\left((\tilde{\mathbf{x}}, \tilde{Z}, \tilde{\mathcal{Y}}), (\hat{\mathbf{x}}, \hat{Z}, \hat{\mathcal{Y}})\right). \tag{2.53}$$

Now, define $\tilde{a} \equiv i_{\mathbb{R}}(\tilde{\mathbf{x}}, \tilde{Z}, \tilde{\mathcal{Y}})$ and $\hat{a} \equiv i_{\mathbb{R}}(\hat{\mathbf{x}}, \hat{Z}, \hat{\mathcal{Y}})$. Then this inequality becomes

$$\|i_{\mathbb{R}} \circ \mathcal{M} \circ i_{\mathbb{R}}^{-1}(\tilde{a}) - i_{\mathbb{R}} \circ \mathcal{M} \circ i_{\mathbb{R}}^{-1}(\hat{a})\|_{\infty} > L d_{\infty}\left(i_{\mathbb{R}}^{-1}(\tilde{a}), i_{\mathbb{R}}^{-1}(\hat{a})\right). \tag{2.54}$$

Again, (2.43) implies that

$$\begin{aligned} \|i_{\mathbb{R}} \circ \mathcal{M} \circ i_{\mathbb{R}}^{-1}(\tilde{a}) - i_{\mathbb{R}} \circ \mathcal{M} \circ i_{\mathbb{R}}^{-1}(\hat{a})\|_{\infty} &> L\|i_{\mathbb{R}}(i_{\mathbb{R}}^{-1}(\tilde{a})) - i_{\mathbb{R}}(i_{\mathbb{R}}^{-1}(\hat{a}))\|_{\infty}, \\ &= L\|\tilde{a} - \hat{a}\|_{\infty}. \end{aligned} \tag{2.55}$$

But $\tilde{a}, \hat{a} \in Q$, so this contradicts (2.51). □

**Theorem 2.5.12.** *Let $\mathcal{M} : \mathcal{D} \subset \mathbb{R}^k \times \mathbb{IR}^n \times \mathbb{MR}^m \to \mathbb{R}^l \times \mathbb{IR}^q \times \mathbb{MR}^r$. The following conditions are equivalent:*

1. *$\mathcal{M}$ is locally Lipschitz on $\mathcal{D}$.*

2. *$i_{\mathbb{R}} \circ \mathcal{M}$ is locally Lipschitz on $\mathcal{D}$.*

3. *$\mathcal{M} \circ i_{\mathbb{R}}^{-1}$ is locally Lipschitz on $Q \equiv i_{\mathbb{R}}(\mathcal{D})$.*

*4.* $i_{\mathbb{R}} \circ \mathcal{M} \circ i_{\mathbb{R}}^{-1}$ *is locally Lipschitz on* $Q \equiv i_{\mathbb{R}}(\mathcal{D})$*.*

*Proof.* It follows directly from (2.43) that both $i_{\mathbb{R}}$ and $i_{\mathbb{R}}^{-1}$ are Lipschitz. Then, since the composition of locally Lipschitz functions is locally Lipschitz, 1 implies 2, 3 and 4, 2 implies 4, and 3 implies 4. Then, it suffices to show that 4 implies 1.

Suppose that 4 holds and that $\mathcal{M}$ is not locally Lipschitz at $(\mathbf{x}, Z, \mathcal{Y}) \in \mathcal{D}$. Then, for any $\eta, L > 0$, there exist distinct points $(\tilde{\mathbf{x}}, \tilde{Z}, \tilde{\mathcal{Y}})$ and $(\hat{\mathbf{x}}, \hat{Z}, \hat{\mathcal{Y}})$ in $\mathcal{D}$, both within $\eta$ of $(\mathbf{x}, Z, \mathcal{Y})$, such that

$$\frac{d_\infty\left(\mathcal{M}(\tilde{\mathbf{x}}, \tilde{Z}, \tilde{\mathcal{Y}}), \mathcal{M}(\hat{\mathbf{x}}, \hat{Z}, \hat{\mathcal{Y}})\right)}{d_\infty\left((\tilde{\mathbf{x}}, \tilde{Z}, \tilde{\mathcal{Y}}), (\hat{\mathbf{x}}, \hat{Z}, \hat{\mathcal{Y}})\right)} > L. \tag{2.56}$$

In particular, there must exists two sequences of points in $\mathcal{D}$, denoted by $(\tilde{\mathbf{x}}_k, \tilde{Z}_k, \tilde{\mathcal{Y}}_k)$ and $(\hat{\mathbf{x}}_k, \hat{Z}_k, \hat{\mathcal{Y}}_k)$, both converging to $(\mathbf{x}, Z, \mathcal{Y})$, such that $(\tilde{\mathbf{x}}_k, \tilde{Z}_k, \tilde{\mathcal{Y}}_k) \neq (\hat{\mathbf{x}}_k, \hat{Z}_k, \hat{\mathcal{Y}}_k)$ for all $k \in \mathbb{N}$ and

$$\limsup_{k\to\infty} \frac{d_\infty\left(\mathcal{M}(\tilde{\mathbf{x}}_k, \tilde{Z}_k, \tilde{\mathcal{Y}}_k), \mathcal{M}(\hat{\mathbf{x}}_k, \hat{Z}_k, \hat{\mathcal{Y}}_k)\right)}{d_\infty\left((\tilde{\mathbf{x}}_k, \tilde{Z}_k, \tilde{\mathcal{Y}}_k), (\hat{\mathbf{x}}_k, \hat{Z}_k, \hat{\mathcal{Y}}_k)\right)} = +\infty. \tag{2.57}$$

By (2.43), this implies that

$$\limsup_{k\to\infty} \frac{\|i_{\mathbb{R}} \circ \mathcal{M}(\tilde{\mathbf{x}}_k, \tilde{Z}_k, \tilde{\mathcal{Y}}_k) - i_{\mathbb{R}} \circ \mathcal{M}(\hat{\mathbf{x}}_k, \hat{Z}_k, \hat{\mathcal{Y}}_k)\|_\infty}{d_\infty\left((\tilde{\mathbf{x}}_k, \tilde{Z}_k, \tilde{\mathcal{Y}}_k), (\hat{\mathbf{x}}_k, \hat{Z}_k, \hat{\mathcal{Y}}_k)\right)} = +\infty. \tag{2.58}$$

Now, for every $k \in \mathbb{N}$, define $\tilde{a}_k \equiv i_{\mathbb{R}}(\tilde{\mathbf{x}}_k, \tilde{Z}_k, \tilde{\mathcal{Y}}_k)$ and $\hat{a}_k \equiv i_{\mathbb{R}}(\hat{\mathbf{x}}_k, \hat{Z}_k, \hat{\mathcal{Y}}_k)$, so that

$$\limsup_{k\to\infty} \frac{\|i_{\mathbb{R}} \circ \mathcal{M} \circ i_{\mathbb{R}}^{-1}(\tilde{a}_k) - i_{\mathbb{R}} \circ \mathcal{M} \circ i_{\mathbb{R}}^{-1}(\hat{a}_k)\|_\infty}{d_\infty\left(i_{\mathbb{R}}^{-1}(\tilde{a}_k), i_{\mathbb{R}}^{-1}(\hat{a}_k)\right)} = +\infty. \tag{2.59}$$

Using (2.43) again, this implies that

$$\limsup_{k\to\infty} \frac{\|i_{\mathbb{R}} \circ \mathcal{M} \circ i_{\mathbb{R}}^{-1}(\tilde{a}_k) - i_{\mathbb{R}} \circ \mathcal{M} \circ i_{\mathbb{R}}^{-1}(\hat{a}_k)\|_\infty}{\|\tilde{a}_k - \hat{a}_k\|_\infty} = +\infty. \tag{2.60}$$

But for every $k \in \mathbb{N}$, $\tilde{a}_k, \hat{a}_k \in Q$. Furthermore, the fact that $i_{\mathbb{R}}$ is an isometry implies that $\{\tilde{a}_k\}$ and $\{\hat{a}_k\}$ converge to $a \equiv i_{\mathbb{R}}(\mathbf{x}, Z, \mathcal{Y})$. Then, (2.60) contradicts 4. $\square$

### 2.5.3 Piecewise $C^1$ Functions

Definition 4.5.1 in [56] introduces the class of piecewise $C^1$ functions, which is extended to interval functions here. The formal definition of this class of functions is not important here. Only the following known facts will be used:

**Lemma 2.5.13.** *Let $E_f \subset \mathbb{R}^n$ and $E_g \subset \mathbb{R}^m$ be open.*

1. *If $\mathbf{f} \in C^1(E_f, \mathbb{R}^m)$, then $\mathbf{f}$ is piecewise $C^1$ on $E_f$.*
2. *Let $\mathbf{f}_1, \mathbf{f}_2 : E_f \subset \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbf{g} : E_g \to \mathbb{R}^q$ be piecewise $C^1$ on $E_f$ and $E_g$, respectively.*
   - *(a) $\mathbf{f}_1 + \mathbf{f}_2$ is piecewise $C^1$ on $E_f$.*
   - *(b) $\mathbf{g} \circ \mathbf{f}_1$ is piecewise $C^1$ on the open set $E_{fg} \equiv \{\mathbf{z} \in E_f : \mathbf{f}_1(\mathbf{z}) \in E_g\}$.*
   - *(c) If $m = 1$, then $f_1 f_2$, $\min(f_1, f_2)$ and $\max(f_1, f_2)$ are piecewise $C^1$ on $E_f$.*
3. *If $\mathbf{f} : E_f \to \mathbb{R}^m$ is piecewise $C^1$ on $E_f$, then $\mathbf{f}$ is locally Lipschitz on $E_f$.*
4. *If $\mathbf{f} : E_f \to \mathbb{R}^m$ is piecewise $C^1$ on $E_f$, then $\mathbf{f}$ is Frechet differentiable everywhere in $E_f$ except on a subset of Lebesgue measure zero.*

*Proof.* For Conclusions 1 and 2, see p. 92 of [153]. Conclusion 3 is Corollary 4.1.1 in [153], and Conclusion 4 follows from Theorem 3.1.1 in [56]. $\square$

The notion of a piecewise $C^1$ function is now extended to interval-valued mappings. By Theorem 2.5.3, a mapping $\phi : E \subset \mathbb{R}^n \to \mathbb{IR}^m$ is continuous on $E$ if and only if $i_{\mathbb{R}} \circ \phi$ is continuous on $E$. Then, the following definition is consistent with other notions of regularity for interval-valued mappings.

**Definition 2.5.14.** Let $E \subset \mathbb{R}^n$ be open and let $\phi : E \to \mathbb{IR}^m$. The mapping $\phi$ is called piecewise $C^1$ on $E$ if $i_{\mathbb{R}} \circ \phi$ is piecewise $C^1$ on $E$.

From the discussion above, it follows that if $\phi$ is piecewise $C^1$ on $E$, then it is continuous as a mapping from $E$ to $\mathbb{IR}^m$. This leads to the following lemma, which is required for further results to be well-posed.

**Lemma 2.5.15.** *Let $\phi : E \subset \mathbb{R}^n \to \mathbb{IR}^m$ be piecewise $C^1$ on $E$. If $\mathfrak{D} \subset \mathbb{IR}^m$ is open, then*

$$E_{\mathfrak{D}} \equiv \{\mathbf{z} \in E : \phi(\mathbf{z}) \in \mathfrak{D}\} \tag{2.61}$$

*is open.*

*Proof.* Since $\phi$ is piecewise $C^1$ on $E$, it is continuous on $E$. Therefore, $E_{\mathfrak{D}}$ is the inverse image in $E$ of the open set $\mathfrak{D}$ under a continuous mapping, and hence it is open with respect to $E$. Since $E$ is itself open, $E_{\mathfrak{D}}$ is open. □

The definition of a piecewise $C^1$ interval-valued mapping can now be extended to mappings from $\mathbb{IR}^m$ to $\mathbb{IR}^q$ as follows.

**Definition 2.5.16.** Let $\mathfrak{D} \subset \mathbb{IR}^m$ be open and let $M : \mathfrak{D} \to \mathbb{IR}^q$. $M$ is called piecewise $C^1$ on $\mathfrak{D}$ if, for every piecewise $C^1$ function $\phi : E \subset \mathbb{R}^n \to \mathbb{IR}^m$, the mapping

$$E_{\mathfrak{D}} \ni \mathbf{z} \longmapsto M(\phi(\mathbf{z})) \in \mathbb{IR}^q \tag{2.62}$$

is piecewise $C^1$ on the open set $E_{\mathfrak{D}} \equiv \{\mathbf{z} \in E : \phi(\mathbf{z}) \in \mathfrak{D}\}$.

As with real-valued functions, a piecewise $C^1$ interval function is locally Lipschitz. Proving this claim requires the following function, which is important in later chapters as well.

**Definition 2.5.17.** Let $\square : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{IR}^n$ be defined by

$$\square(\mathbf{v}, \mathbf{w}) \equiv \left[\mathbf{v} - \max\left(\mathbf{0}, \frac{1}{2}(\mathbf{v} - \mathbf{w})\right), \mathbf{w} + \max\left(\mathbf{0}, \frac{1}{2}(\mathbf{v} - \mathbf{w})\right)\right]. \tag{2.63}$$

Interpretation of $\square$ is provided by the following lemma. The proof is trivial.

**Lemma 2.5.18.** *Let* $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$.

1. *If* $\mathbf{v} \leq \mathbf{w}$, *then* $\square(\mathbf{v}, \mathbf{w}) = [\mathbf{v}, \mathbf{w}]$.

2. *For every* $i$ *with* $v_i > w_i$, $\square(v_i, w_i)$ *is the singleton* $\{m([w_i, v_i])\}$.

**Lemma 2.5.19.** $\square$ *is piecewise* $C^1$ *on* $\mathbb{R}^n \times \mathbb{R}^n$.

*Proof.* The result follows from Definition 2.5.17 and Conclusions 1 and 2 of Lemma 2.5.13. □

**Lemma 2.5.20.** *Let* $\mathfrak{D} \subset \mathbb{IR}^m$. *If* $M : \mathfrak{D} \to \mathbb{IR}^q$ *is piecewise* $C^1$ *on* $\mathfrak{D}$, *then it is locally Lipschitz on* $\mathfrak{D}$.

*Proof.* Choosing $\phi = \square$ in Definition 2.5.16, the hypothesis on $M$ implies that the function $i_{\mathbb{R}} \circ M \circ \square$ is piecewise $C^1$ on $E_{\mathfrak{D}} \equiv \{(\mathbf{v}, \mathbf{w}) \in \mathbb{R}^n \times \mathbb{R}^n : \square(\mathbf{v}, \mathbf{w}) \in \mathfrak{D}\}$. By Conclusion 3 of Lemma 2.5.13, this implies that $i_{\mathbb{R}} \circ M \circ \square$ is locally Lipschitz on $E_{\mathfrak{D}}$.

Recall from Definition 2.5.1 that $\mathbb{H}^{(0,n,0)} = \{(\mathbf{v}, \mathbf{w}) \in \mathbb{R}^{2n} : \mathbf{v} \leq \mathbf{w}\}$, and for any $(\mathbf{v}, \mathbf{w}) \in \mathbb{H}^{(0,n,0)}$, $i_{\mathbb{R}}^{-1}(\mathbf{v}, \mathbf{w}) = [\mathbf{v}, \mathbf{w}]$. Then, the restriction of $\square$ to $\mathbb{H}^{(0,n,0)}$ is exactly $i_{\mathbb{R}}^{-1}$. This implies that $i_{\mathbb{R}} \circ M \circ i_{\mathbb{R}}^{-1}$ is locally Lipschitz on $Q = \{(\mathbf{v}, \mathbf{w}) \in \mathbb{H}^{(0,n,0)} : \square(\mathbf{v}, \mathbf{w}) \in \mathfrak{D}\} = \{(\mathbf{v}, \mathbf{w}) \in \mathbb{H}^{(0,n,0)} : i_{\mathbb{R}}^{-1}(\mathbf{v}, \mathbf{w}) \in \mathfrak{D}\} = i_{\mathbb{R}}(E_{\mathfrak{D}})$. Now Theorem 2.5.12 shows that $M$ is locally Lipschitz on $E_{\mathfrak{D}}$. □

The composition of piecewise $C^1$ interval functions is again piecewise $C^1$, as shown by the following lemma.

**Lemma 2.5.21.** *Let* $\mathfrak{D}_1 \subset \mathbb{IR}^m$ *and* $\mathfrak{D}_2 \subset \mathbb{IR}^k$ *be open and let* $M_1 : \mathfrak{D}_1 \to \mathbb{IR}^k$ *and* $M_2 : \mathfrak{D}_2 \to \mathbb{IR}^q$ *be piecewise* $C^1$ *on* $\mathfrak{D}_1$ *and* $\mathfrak{D}_2$, *respectively. The set* $\mathfrak{D}_{12} \equiv \{Z \in \mathfrak{D}_1 : M_1(Z) \in \mathfrak{D}_2\}$ *is open and* $M_2 \circ M_1$ *is piecewise* $C^1$ *on* $\mathfrak{D}_{12}$.

*Proof.* Since $M_1$ is piecewise $C^1$ on $\mathfrak{D}_1$, it is locally Lipschitz and hence continuous there. Then, the set $\mathfrak{D}_{12}$ is the inverse image in $\mathfrak{D}_1$ of the open set $\mathfrak{D}_2$ under a continuous mapping. Therefore, $\mathfrak{D}_{12}$ is open with respect to $\mathfrak{D}_1$. Since $\mathfrak{D}_1$ is open in $\mathbb{IR}^m$, so is $\mathfrak{D}_{12}$.

Choose any piecewise $C^1$ mapping $\phi : E \subset \mathbb{R}^n \to \mathbb{IR}^m$ and define $E_{\mathfrak{D}_1} \equiv \{\mathbf{z} \in E : \phi(\mathbf{z}) \in \mathfrak{D}_1\}$. Now define $\phi' : E_{\mathfrak{D}_1} \to \mathbb{IR}^k$ by

$$\phi'(\mathbf{z}) = M_1(\phi(\mathbf{z})), \quad \forall \mathbf{z} \in E_{\mathfrak{D}_1}. \tag{2.64}$$

Since $M_1$ is piecewise $C^1$ on $\mathfrak{D}_1$, $\phi'$ is piecewise $C^1$ on $E_{\mathfrak{D}_1}$. But since $M_2$ is piecewise $C^1$ on $\mathfrak{D}_2$, this implies that

$$\mathbf{z} \longmapsto M_2(\phi'(\mathbf{z})) = M_2(M_1(\phi(\mathbf{z}))) \tag{2.65}$$

is piecewise $C^1$ on the set

$$\begin{aligned} \{\mathbf{z} \in E_{\mathfrak{D}_1} : \phi'(\mathbf{z}) \in \mathfrak{D}_2\} &= \{\mathbf{z} \in E : \phi(\mathbf{z}) \in \mathfrak{D}_1 \text{ and } M_1(\phi(\mathbf{z})) \in \mathfrak{D}_2\}, & (2.66) \\ &= \{\mathbf{z} \in E : \phi(\mathbf{z}) \in \mathfrak{D}_{12}\}. & (2.67) \end{aligned}$$

But $\phi$ was chosen arbitrarily, so $M_2 \circ M_1$ is piecewise $C^1$ on $\mathfrak{D}_{12}$. □

Before leaving this section, we introduce the *extended intersection* of intervals, which is a useful in later sections, and establish its regularity.

**Definition 2.5.22.** Let $\tilde{\cap} : \mathbb{IR}^n \times \mathbb{IR}^n \to \mathbb{IR}^n$ be defined by

$$\tilde{\cap}([\mathbf{z}^L, \mathbf{z}^U], [\hat{\mathbf{z}}^L, \hat{\mathbf{z}}^U]) \equiv [\text{mid}(\mathbf{z}^L, \mathbf{z}^U, \hat{\mathbf{z}}^L), \text{mid}(\mathbf{z}^L, \mathbf{z}^U, \hat{\mathbf{z}}^U)]. \tag{2.68}$$

Furthermore, define the standard notation $Z \tilde{\cap} \hat{Z} \equiv \tilde{\cap}(Z, \hat{Z})$, $\forall Z, \hat{Z} \in \mathbb{IR}^n$.

An interpretation of this function is given by the following lemma.

**Lemma 2.5.23.** *Let $Z, \hat{Z} \in \mathbb{IR}^n$.*

1. *If $Z \cap \hat{Z} \neq \emptyset$, then $Z \tilde{\cap} \hat{Z} = Z \cap \hat{Z}$.*

2. *For all $i$ such that $Z_i \cap \hat{Z}_i = \emptyset$, $Z_i \tilde{\cap} \hat{Z}_i$ is either $\{z_i^L\}$ or $\{z_i^U\}$.*

3. *$Z \tilde{\cap} \hat{Z} \subset Z$.*

The proof of the preceding lemma is straightforward and is omitted.

**Lemma 2.5.24.** *For any* $Q = [\mathbf{q}^L, \mathbf{q}^U] \in \mathbb{IR}^n$*, the mapping* $\tilde{\cap}(Q, \cdot)$ *is an inclusion monotonic interval extension of* $\mathrm{mid}(\mathbf{q}^L, \mathbf{q}^U, \cdot)$.

*Proof.* Let $\mathbf{z} \in \mathbb{R}^n$. Then, by definition,

$$\tilde{\cap}(Q, [\mathbf{z}, \mathbf{z}]) = [\mathrm{mid}(\mathbf{q}^L, \mathbf{q}^U, \mathbf{z}), \mathrm{mid}(\mathbf{q}^L, \mathbf{q}^U, \mathbf{z})].$$

Therefore, $\tilde{\cap}(Q, \cdot)$ is an interval extension of $\mathrm{mid}(\mathbf{q}^L, \mathbf{q}^U, \cdot)$. If $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^n$ and $\mathbf{z}_1 \leq \mathbf{z}_2$, then it is obvious that $\mathrm{mid}(\mathbf{q}^L, \mathbf{q}^U, \mathbf{z}_1) \leq \mathrm{mid}(\mathbf{q}^L, \mathbf{q}^U, \mathbf{z}_2)$. By the definition of $\tilde{\cap}(Q, \cdot)$, inclusion monotonicity must follow. □

**Lemma 2.5.25.** $\tilde{\cap}$ *is piecewise* $C^1$ *on* $\mathbb{IR}^n \times \mathbb{IR}^n$.

*Proof.* If $\mathbf{z}^L, \mathbf{z}^U \in \mathbb{R}^n$ and $\mathbf{z}^L \leq \mathbf{z}^U$, it is easily verified that $\mathrm{mid}(\mathbf{z}^L, \mathbf{z}^U, \hat{\mathbf{z}})$ is equivalent to $\max(\mathbf{z}^L, \min(\mathbf{z}^U, \hat{\mathbf{z}}))$ for all $\hat{\mathbf{z}} \in \mathbb{R}^n$. The result now follows from Definition 2.5.22 and Conclusion 2 of Lemma 2.5.13. □

### 2.5.4 Regularity of Factorable Functions

The natural function of a $\mathcal{L}$-computational sequence, and hence any $\mathcal{L}$-factorable function described by that sequence, inherits some nice properties from the elements of $\mathcal{L}$.

**Theorem 2.5.26.** *Let* $(\mathcal{S}, \pi_o)$ *be a* $\mathcal{L}$*-computational sequence with natural function* $(\mathbf{f}_{\mathcal{S}}, D_{\mathcal{S}}, \mathbb{R}^m)$*. If* $u$ *is continuous (resp. locally Lipschitz, Lipschitz,* $k$ *times continuously differentiable) on* $B$ *for every* $(u, B, \mathbb{R}) \in \mathcal{L}$*, then* $\mathbf{f}_{\mathcal{S}}$ *is continuous (resp. locally Lipschitz, Lipschitz,* $k$ *times continuously differentiable) on* $D_{\mathcal{S}}$*. If* $u$ *is continuous on* $B$ *and* $B$ *is open for every* $(u, B, \mathbb{R}) \in \mathcal{L}$*, then* $D_{\mathcal{S}}$ *is open.*

*Proof.* Suppose $u$ is continuous on $B$ and $B$ is open for every $(u, B, \mathbb{R}) \in \mathcal{L}$. For $K = n_i + 1$, it is clear that $D_k$ is open and $v_k$ is continuous on $D_k$ for all $k < K$. Suppose this is true of some arbitrary $K \in \{n_i+1, \ldots, n_f\}$. Then $D_K$ is the inverse image of an open set under a continuous mapping, and is hence open, and $v_K$ is continuous on $D_K$

by Theorem 4.7 in [145]. Finite induction shows that $D_{\mathcal{S}}$ is open and $\mathbf{f}_{\mathcal{S}}$ is continuous there. Since Theorem 4.7 in [145] does not require openness, the same argument shows that $\mathbf{f}_{\mathcal{S}}$ is continuous on $D_{\mathcal{S}}$ if only $u$ is continuous on $B$ for every $(u, B, \mathbb{R}) \in \mathcal{L}$. Furthermore, the same argument shows the claims for local Lipschitz continuity and Lipschitz continuity using well-known composition results (see Theorem 2.5.6), and for $k$ times continuous differentiability in light of the composition result on p. 199 of [127]. □

### 2.5.5 Regularity of Natural Interval Extensions

In this section, it is shown that natural interval extensions are locally Lipschitz and piecewise $C^1$ under appropriate assumptions on the elements of $\mathcal{L}$.

**Theorem 2.5.27.** $(+, \mathbb{IR}^2, \mathbb{IR})$ *and* $(\times, \mathbb{IR}^2, \mathbb{IR})$ *are piecewise* $C^1$ *on* $\mathbb{IR}^2$*.*

*Proof.* Let $(\phi_1, \phi_2) : E \subset \mathbb{R}^n \to \mathbb{IR} \times \mathbb{IR}$ be piecewise $C^1$ on $E$, and let $F = \phi_1 + \phi_2$. Then $F^L(\mathbf{x}) = \phi_1^L(\mathbf{x}) + \phi_2^L(\mathbf{x})$. By Definition 2.5.14, this a sum of two piecewise $C^1$ functions, and is itself piecewise $C^1$ by Condition 2 of Lemma 2.5.13. Using an analogous argument for $F^U$, it follows that $\phi_1 + \phi_2$ is piecewise $C^1$ on $E$. Since $(\phi_1, \phi_2)$ was chosen arbitrarily, this implies that $(+, \mathbb{IR}^2, \mathbb{IR})$ is piecewise $C^1$ on $\mathbb{IR}^2$.

Now let $F = \phi_1\phi_2$. Then

$$F^L(\mathbf{x}) = \min(\phi_1^L(\mathbf{x})\phi_2^L(\mathbf{x}), \phi_1^L(\mathbf{x})\phi_2^U(\mathbf{x}), \phi_1^U(\mathbf{x})\phi_2^L(\mathbf{x}), \phi_1^U(\mathbf{x})\phi_2^U(\mathbf{x})).$$

Thus, $F^L$ is piecewise $C^1$ on $E$ by Condition 2 of Lemma 2.5.13. Using an analogous argument for $F^U$, it follows that $\phi_1\phi_2$ is piecewise $C^1$ on $E$. Since $(\phi_1, \phi_2)$ was chosen arbitrarily, this implies that $(\times, \mathbb{IR}^2, \mathbb{IR})$ is piecewise $C^1$ on $\mathbb{IR}^2$. □

By Theorem 2.5.20, the previous result implies that $(+, \mathbb{IR}^2, \mathbb{IR})$ and $(\times, \mathbb{IR}^2, \mathbb{IR})$ are locally Lipschitz on $\mathbb{IR}^2$. The next theorem shows that $(+, \mathbb{IR}^2, \mathbb{IR})$ is in fact Lipschitz on $\mathbb{IR}^2$.

**Theorem 2.5.28.** $(+, \mathbb{IR}^2, \mathbb{IR})$ *is Lipschitz on* $\mathbb{IR}^2$*.*

*Proof.* Let $f^L, f^U : \mathbb{H}^{(0,2,0)} \to \mathbb{R}$ be defined by

$$[f^L(x^L, y^L, x^U, y^U), f^U(x^L, y^L, x^U, y^U)] = [x^L, x^U] + [y^L, y^U] \tag{2.69}$$

where $\mathbb{H}^{(0,2,0)} = \{(x^L, y^L, x^U, y^U) \in \mathbb{R}^4 : x^L \leq x^U,\ y^L \leq y^U\}$. By Theorem 2.5.11, it suffices to show that both $f^L$ and $f^U$ are Lipschitz on $\mathbb{H}^{(0,2,0)}$. By definition, $f^L(x^L, y^L, x^U, y^U) = x^L + y^L$ and $f^U(x^L, y^L, x^U, y^U) = x^U + y^U$, so this is clearly true. □

**Assumption 2.5.29.** For every $(u, B, \mathbb{R}) \in \mathcal{L}$, the interval extension $(u, \mathbb{I}B, \mathbb{IR})$ is locally Lipschitz on $\mathbb{I}B$.

**Theorem 2.5.30.** *Let $(\mathcal{S}, \pi_o)$ be a $\mathcal{L}$-computational sequence. The natural interval extension $(F_{\mathcal{S}}, \mathfrak{D}_{\mathcal{S}}, \mathbb{IR}^{n_o})$ is locally Lipschitz on $\mathfrak{D}_{\mathcal{S}}$.*

*Proof.* Consider the sequence of inclusion factors $\{(V_k, \mathfrak{D}_k, \mathbb{IR})\}_{k=1}^{n_f}$. Choose any $K \in \{1, \ldots, n_f\}$ and suppose that $(V_k, \mathfrak{D}_k, \mathbb{IR})$ is locally Lipschitz on $\mathfrak{D}_k$, for all $k \in \{1, \ldots, K-1\}$. If $K \leq n_i + 1$, this is obviously true. By Theorem 2.5.27 and Assumption 2.5.29, $(o_K, \mathbb{I}B_K, \mathbb{IR})$ must be locally Lipschitz on $\mathbb{I}B_K$. Then, since the composition of locally Lipschitz functions is again locally Lipschitz (Theorem 2.5.6), $(V_K, \mathfrak{D}_K, \mathbb{IR})$ is locally Lipschitz on $\mathfrak{D}_K$. By induction, this holds for every $K \in \{1, \ldots, n_f\}$, and the theorem follows from the definition of $(F_{\mathcal{S}}, \mathfrak{D}_{\mathcal{S}}, \mathbb{IR}^{n_o})$. □

**Corollary 2.5.31.** *Let $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ be a $\mathcal{L}$-factorable function. Then, every natural interval extension of $\mathbf{f}$, $([\mathbf{f}], \mathfrak{D}, \mathbb{IR}^m)$ is locally Lipschitz on $\mathfrak{D}$.*

Of course, obtaining piecewise continuous differentiability of natural interval extensions requires a stronger assumption on the elements of $\mathcal{L}$. The following lemma is required to make this assumption well-posed.

**Lemma 2.5.32.** *If $D \subset \mathbb{R}^n$ is open, then $\mathbb{I}D$ is open in $\mathbb{IR}^n$.*

*Proof.* If $\mathbb{I}D$ is empty, then it is trivially open. Otherwise, choose $Z \in \mathbb{I}D$. Then, $Z \subset D$, and since $D$ is open, $\exists \epsilon > 0$ such that $\hat{\mathbf{z}} \in D$ if $\|\hat{\mathbf{z}} - \mathbf{z}\|_\infty \leq \epsilon$ and $\mathbf{z} \in Z$ (uniformity of $\epsilon$ for every $\mathbf{z} \in Z$ results from the compactness of $Z$, as per Theorem

4.6 in [127]). Let $\hat{Z} \in \mathbb{IR}^n$ satisfy $d_H(Z, \hat{Z}) \leq \epsilon$. By the definition of $d_H$, this implies that, for any $\hat{\mathbf{z}} \in \hat{Z}$, there exists $\mathbf{z} \in Z$ such that $\|\hat{\mathbf{z}} - \mathbf{z}\|_\infty \leq \epsilon$. But this implies that $\hat{Z} \subset D$ or, equivalently, $\hat{Z} \in \mathbb{I}D$. Hence, $Z$ is an interior point of $\mathbb{I}D$ and, since $Z$ was chosen arbitrarily, $\mathbb{I}D$ is open. □

The following assumption is stronger than necessary for most uses of natural interval extensions in this thesis and will be stated explicitly wherever it is needed.

**Assumption 2.5.33.** For every $(u, B, \mathbb{R}) \in \mathcal{L}$, $B$ is open and the interval extension $(u, \mathbb{I}B, \mathbb{IR})$ is piecewise $C^1$ on $\mathbb{I}B$.

**Theorem 2.5.34.** *Let $(\mathcal{S}, \pi_o)$ be a $\mathcal{L}$-computational sequence with natural interval extension $(F_\mathcal{S}, \mathfrak{D}_\mathcal{S}, \mathbb{IR}^{n_o})$. If Assumption 2.5.33 holds, then $\mathfrak{D}_\mathcal{S}$ is open and $F_\mathcal{S}$ is piecewise $C^1$ on $\mathfrak{D}_\mathcal{S}$.*

*Proof.* Consider the sequence of inclusion factors $\{(V_k, \mathfrak{D}_k, \mathbb{IR})\}_{k=1}^{n_f}$. Choose any $K \in \{1, \dots, n_f\}$ and suppose that $\mathfrak{D}_k$ is open and $V_k$ is piecewise $C^1$ on $\mathfrak{D}_k$, for all $k \in \{1, \dots, K-1\}$. If $K \leq n_i + 1$, this is obviously true. Since $B_K$ is open, $\mathbb{I}B_K$ is open by Lemma 2.5.32. Then $\mathfrak{D}_K$ is the inverse image of an open set under a continuous mapping, and is therefore open. By Theorem 2.5.27 and Assumption 2.5.33, $(o_K, \mathbb{I}B_K, \mathbb{IR})$ must be piecewise $C^1$ on $\mathbb{I}B_K$. Then, since the composition of piecewise $C^1$ functions is again piecewise $C^1$ by Conclusion 2 of Lemma 2.5.13, $(V_K, \mathfrak{D}_K, \mathbb{IR})$ is piecewise $C^1$ on $\mathfrak{D}_K$. By induction, this holds for every $K \in \{1, \dots, n_f\}$, and the theorem follows from the definition of $(F_\mathcal{S}, \mathfrak{D}_\mathcal{S}, \mathbb{IR}^{n_o})$. □

**Corollary 2.5.35.** *Let $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ be a $\mathcal{L}$-factorable function. For every natural interval extension $([\mathbf{f}], \mathfrak{D}, \mathbb{IR}^m)$ of $\mathbf{f}$ satisfying Assumption 2.5.33, $\mathfrak{D}$ is open and $[\mathbf{f}]$ is piecewise $C^1$ on $\mathfrak{D}$.*

**Remark 2.5.36.**

1. It is clear from the proof of Theorem 2.5.30 that natural interval extensions remain continuous if the local Lipschitz condition on the univariate interval extensions in Assumption 2.5.29 is replaced with a continuity assumption.

2. The conclusion in Theorem 2.5.34 that $\mathfrak{D}_{\mathcal{S}}$ is open does not require that the univariate interval extensions are piecewise $C^1$, as in Assumption 2.5.33. An analogous proof shows that this conclusion holds if only $\mathbb{I}B$ is open and $(u, \mathbb{I}B, \mathbb{IR})$ is continuous on $\mathbb{I}B$ for every $(u, B, \mathbb{R}) \in \mathcal{L}$.

### 2.5.6 Regularity of Natural McCormick Extensions

In this section, it is shown that natural McCormick extensions are locally Lipschitz under appropriate assumptions on the elements of $\mathcal{L}$.

**Theorem 2.5.37.** $(\mathrm{Cut}, \mathbb{MR}, \mathbb{MR})$ *is Lipschitz on* $\mathbb{MR}$*.*

*Proof.* Let $f^L, f^U, f^{cv}, f^{cc} : \mathbb{H}^{(0,0,1)} \to \mathbb{R}$ be defined by

$$\begin{aligned}&([f^L(x^L, x^U, x^{cv}, x^{cc}), f^U(x^L, x^U, x^{cv}, x^{cc})],\\ &\quad [f^{cv}(x^L, x^U, x^{cv}, x^{cc}), f^{cc}(x^L, x^U, x^{cv}, x^{cc})]) = \mathrm{Cut}(([x^L, x^U], [x^{cv}, x^{cc}])),\end{aligned} \tag{2.70}$$

where

$$\begin{aligned}\mathbb{H}^{(0,0,1)} = \{&(x^L, x^U, x^{cv}, x^{cc}) \in \mathbb{R}^4 :\\ &x^L \leq x^U,\ x^{cv} \leq x^{cc},\ [x^L, x^U] \cap [x^{cv}, x^{cc}] \neq \emptyset\}.\end{aligned} \tag{2.71}$$

By Theorem 2.5.11, it suffices to show that $f^L$, $f^U$, $f^{cv}$ and $f^{cc}$ are Lipschitz on $\mathbb{H}^{(0,0,1)}$. For $f^L(x^L, x^U, x^{cv}, x^{cc}) = x^L$ and $f^U(x^L, x^U, x^{cv}, x^{cc}) = x^U$, this is obvious. For $f^{cv}(x^L, x^U, x^{cv}, x^{cc}) = \max(x^L, x^{cv})$ and $f^{cc}(x^L, x^U, x^{cv}, x^{cc}) = \min(x^U, x^{cc})$, it follows from the fact that min and max are Lipschitz on $\mathbb{R}^2$. □

**Theorem 2.5.38.** $(+, \mathbb{MR}^2, \mathbb{MR})$ *is Lipschitz on* $\mathbb{MR}^2$ *and* $(\times, \mathbb{MR}^2, \mathbb{MR})$ *is locally Lipschitz on* $\mathbb{MR}^2$*.*

*Proof.* Let $f^L, f^U, f^{cv}, f^{cc} : \mathbb{H}^{(0,0,2)} \to \mathbb{R}$ be defined by (omitting arguments)

$$([f^L, f^U], [f^{cv}, f^{cc}]) = ([x^L, x^U], [x^{cv}, x^{cc}]) + ([y^L, y^U], [y^{cv}, y^{cc}]), \tag{2.72}$$

where

$$\begin{aligned}\mathbb{H}^{(0,0,2)} = \{(x^L, y^L, x^U, y^U, x^{cv}, y^{cv}, x^{cc}, y^{cc}) \in \mathbb{R}^8 : \\ x^L \leq x^U,\ y^L \leq y^U,\ x^{cv} \leq x^{cc},\ y^{cv} \leq y^{cc}, \\ [x^L, x^U] \cap [x^{cv}, x^{cc}] \neq \emptyset,\ [y^L, y^U] \cap [y^{cv}, y^{cc}] \neq \emptyset\}.\end{aligned} \tag{2.73}$$

By Theorem 2.5.11, the claim for $(+, \mathbb{MR}^2, \mathbb{MR})$ holds provided that $f^L$, $f^U$, $f^{cv}$ and $f^{cc}$ are Lipschitz on $\mathbb{H}^{(0,0,2)}$. By Theorem 2.5.27, this is true of $f^L$ and $f^U$. By definition (again omitting arguments),

$$f^{cv} = \max(x^L, x^{cv}) + \max(y^L, y^{cv}) \quad \text{and} \quad f^{cc} = \min(x^U, x^{cc}) + \min(y^U, y^{cc}). \tag{2.74}$$

From this it is clear that $f^{cv}$ and $f^{cc}$ are Lipschitz on $\mathbb{H}^{(0,0,2)}$ because min and max are Lipschitz on $\mathbb{R}^2$.

Now let $f^L, f^U, f^{cv}, f^{cc} : \mathbb{H}^{(0,0,2)} \to \mathbb{R}$ be defined by (omitting arguments)

$$([f^L, f^U], [f^{cv}, f^{cc}]) = ([x^L, x^U], [x^{cv}, x^{cc}]) \times ([y^L, y^U], [y^{cv}, y^{cc}]). \tag{2.75}$$

By Theorem 2.5.12, the claim for $(\times, \mathbb{MR}^2, \mathbb{MR})$ holds provided that $f^L$, $f^U$, $f^{cv}$ and $f^{cc}$ are locally Lipschitz on $\mathbb{H}^{(0,2,0)}$. By Theorem 2.5.27, this is true of $f^L$ and $f^U$. The upper and lower bounds of the intervals $\bar{X}^C$ and $\bar{Y}^C$ defined in Definition 2.4.21 vary in a Lipschitz manner on $\mathbb{H}^{(0,0,2)}$ by Theorem 2.5.37. Then, by composition, it suffices to show that the expressions

$$\max\left(\left[y^L\bar{X}^C + x^L\bar{Y}^C - x^Ly^L\right]^L, \left[y^U\bar{X}^C + x^U\bar{Y}^C - x^Uy^U\right]^L\right), \tag{2.76}$$

$$\min\left(\left[y^L\bar{X}^C + x^U\bar{Y}^C - y^Lx^U\right]^U, \left[y^U\bar{X}^C + x^L\bar{Y}^C - y^Ux^L\right]^U\right), \tag{2.77}$$

are locally Lipschitz with respect to the bounds of $X^B$, $Y^B$, $\bar{X}^C$ and $\bar{Y}^C$. But this is apparent from Theorem 2.5.27 and the fact that min and max are Lipschitz on $\mathbb{R}^2$. □

**Assumption 2.5.39.** For every function $(u, B, \mathbb{R}) \in \mathcal{L}$, the McCormick extension

$(u, \mathbb{M}B, \mathbb{MR})$ is locally Lipschitz on $\mathbb{M}B$.

**Theorem 2.5.40.** *Let $(\mathcal{S}, \pi_o)$ be a $\mathcal{L}$-computational sequence. The natural McCormick extension $(\mathcal{F}_{\mathcal{S}}, \mathcal{D}_{\mathcal{S}}, \mathbb{MR}^{n_o})$ is locally Lipschitz on $\mathcal{D}_{\mathcal{S}}$.*

*Proof.* Consider the sequence of relaxation factors $\{(\mathcal{V}_k, \mathcal{D}_k, \mathbb{MR})\}_{k=1}^{n_f}$. Choose any $K \in \{1, \ldots, n_f\}$ and suppose that $(\mathcal{V}_k, \mathcal{D}_k, \mathbb{MR})$ is locally Lipschitz on $\mathcal{D}_k$, for all $k \in \{1, \ldots, K-1\}$. If $K \leq n_i + 1$, this is obviously true. By Theorem 2.5.38 and Assumption 2.5.39, $(o_K, \mathbb{M}B_K, \mathbb{MR})$ is locally Lipschitz on $\mathbb{M}B_K$. Since the composition of locally Lipschitz functions is locally Lipschitz (Theorem 2.5.6), $(\mathcal{V}_K, \mathcal{D}_K, \mathbb{MR})$ is locally Lipschitz on $\mathcal{D}_K$. By induction, this holds for every $K \in \{1, \ldots, n_f\}$, and the theorem follows immediately from the definition of $(\mathcal{F}_{\mathcal{S}}, \mathcal{D}_{\mathcal{S}}, \mathbb{IR}^{n_o})$. □

**Corollary 2.5.41.** *Let $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ be a $\mathcal{L}$-factorable function. Then, every natural McCormick extension $(\{\mathbf{f}\}, \mathcal{D}, \mathbb{MR}^m)$ is locally Lipschitz on $\mathcal{D}$.*

**Remark 2.5.42.** As was the case for natural interval extensions, it is clear from the proof of Theorem 2.5.40 that continuity of natural McCormick extensions is achieved if only continuity of the univariate McCormick extensions is assumed in place of the local Lipschitz condition of Assumption 2.5.39.

To conclude this section, we collect some results that are useful for establishing a Lipschitz condition for $\{\mathbf{f}\}$ on certain subsets of $\mathcal{D}_{\mathcal{S}}$.

**Corollary 2.5.43.** *Let $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ be $\mathcal{L}$-factorable and let $\{\mathbf{f}\} : \mathcal{D} \subset \mathbb{MR}^n \to \mathbb{MR}^m$ be a natural McCormick extension of $\mathbf{f}$. The function $i_{\mathbb{R}} \circ \{\mathbf{f}\} \circ i_{\mathbb{R}}^{-1}$ is Lipschitz on any compact $K \subset i_{\mathbb{R}}(\mathcal{D})$.*

*Proof.* Since $K \subset i_{\mathbb{R}}(\mathcal{D})$, $i_{\mathbb{R}} \circ \{\mathbf{f}\} \circ i_{\mathbb{R}}^{-1}$ is locally Lipschitz on $K$ by Corollary 2.5.41 and Theorem 2.5.12. Since $K$ is compact, $\{\mathbf{f}\}$ is Lipschitz on $K$ by Theorem 2.5.7. □

**Lemma 2.5.44.** *Let $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ be $\mathcal{L}$-factorable and let $\{\mathbf{f}\} : \mathcal{D} \subset \mathbb{MR}^n \to \mathbb{MR}^m$ be a natural McCormick extension of $\mathbf{f}$. If $X^0$ is represented in $\mathcal{D}$, then the set $K \equiv \{(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc}) \in \mathbb{H}^{(0,0,n)} : [\mathbf{x}^L, \mathbf{x}^U] \subset X^0,\ i_{\mathbb{R}}^{-1}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc}) \in \mathcal{D}\}$ is a compact subset of $i_{\mathbb{R}}(\mathcal{D})$.*

*Proof.* By Lemma 2.4.33, every $X \subset X^0$ is represented in $\mathcal{D}$. Moreover, $\mathcal{D}$ is closed under coherence by definition. It follows that $K = \{(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc}) \in \mathbb{H}^{(0,0,n)} : [\mathbf{x}^L, \mathbf{x}^U] \subset X^0\}$. Since $\mathbb{H}^{(0,0,n)}$ is closed, this set is clearly compact. □

## 2.6 Standard McCormick Relaxations

In this section, standard McCormick relaxations [118] are defined in terms of natural McCormick extensions of $\mathcal{L}$-factorable functions. Though this presentation is not standard, the resulting relaxations are the same as those defined in McCormick's original work, with the caveat that the McCormick addition and multiplication rules are modified, as discussed in §2.4.2. However, using the results for natural McCormick extensions in 2.5.6, some new regularity and convergence results for McCormick's relaxations are proven here.

**Definition 2.6.1.** Let $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ be $\mathcal{L}$-factorable and let $\{\mathbf{f}\} : \mathcal{D} \subset \mathbb{MR}^n \to \mathbb{MR}^m$ be a natural McCormick extension of $\mathbf{f}$. For any $X \in \mathbb{I}D$ that is represented in $\mathcal{D}$, define $\mathcal{U}, \mathcal{O} : X \to \mathbb{R}^m$ by

$$\mathcal{U}(\mathbf{x}) = \{\mathbf{f}\}^{cv}((X, [\mathbf{x}, \mathbf{x}])) \quad \text{and} \quad \mathcal{O}(\mathbf{x}) = \{\mathbf{f}\}^{cc}((X, [\mathbf{x}, \mathbf{x}])). \tag{2.78}$$

The functions $\mathcal{U}$ and $\mathcal{O}$ are called *standard McCormick relaxations of* $\mathbf{f}$ *on* $X$.

By Lemma 2.4.11, it follows immediately that $\mathcal{U}$ and $\mathcal{O}$ are convex and concave relaxations of $\mathbf{f}$ on $X$, respectively.

**Corollary 2.6.2.** *Let* $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ *be* $\mathcal{L}$*-factorable and let* $\mathcal{U}, \mathcal{O} : X \to \mathbb{R}^m$ *be standard McCormick relaxations of* $\mathbf{f}$ *on* $X$*.* $U$ *and* $O$ *are Lipschitz on* $X$*.*

*Proof.* Let $(\{\mathbf{f}\}, \mathcal{D}_S, \mathbb{MR}^m)$ be the natural McCormick relaxation of $\mathbf{f}$ defining $\mathcal{U}$ and $\mathcal{O}$, and let $K$ be defined as in Lemma 2.5.44 with $X^0 \equiv X$. By Corollary 2.5.43, $\{\mathbf{f}\}^{cv} \circ i_{\mathbb{R}}^{-1}$ and $\{\mathbf{f}\}^{cc} \circ i_{\mathbb{R}}^{-1}$ are Lipschitz on $K$. But for every $\mathbf{x} \in X$, $(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}, \mathbf{x})$ is in $K$, and it follows that $\mathcal{U}$ and $\mathcal{O}$ are Lipschitz on $X$. □

### 2.6.1 McCormick Relaxations on Sequences of Intervals

A primary motivation for constructing convex and concave relaxations is for their use in branch-and bound global optimization algorithms [84, 171]. There, convex and concave relaxations are used to obtain lower and/or upper bounds on the range of a nonconvex function on an interval $X$. These bounds are then successively refined by partitioning the interval into a number of subintervals and constructing convex and concave relaxations valid on each of these subintervals. In such applications, it is important to understand the relationship between relaxations generated on a nested and convergent sequence of subintervals of $X$. From these relationships, one can infer the limiting behavior of the relaxations when the partition of $X$ is refined infinitely, which has important consequences for the convergence of global optimization algorithms.

Let $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$, let $\{\mathbf{f}\} : \mathcal{D} \subset \mathbb{MR}^n \to \mathbb{MR}^m$ be a natural McCormick extension of $\mathbf{f}$, and let $X^0 \in \mathbb{IR}^n$ be represented in $\mathcal{D}$. In this section, standard McCormick relaxations of $\mathbf{f}$ on subintervals of $X^0$ are investigated. The superscript $\ell$ is used to index subintervals $X^\ell \subset X^0$, and also relaxations valid on subintervals of $X$; i.e. $\mathcal{U}^\ell$ and $\mathcal{O}^\ell$ denote the McCormick relaxations of $\mathbf{f}$ constructed as in Definition 2.6.1 with $X^\ell$ in place of $X$. We consider a nested and convergent sequence of subintervals, $\{X^\ell\} \to X^*$, where $X^*$ is by necessity a subinterval of $X^0$. The aim of this analysis is to prove that a branch-and-bound global optimization algorithm with a bounding operation based on McCormick convex and/or concave relaxations converges to within a specified tolerance finitely. The reader is referred to Chapter IV in [84] for a detailed discussion of the convergence of branch-and-bound algorithms and the requisite properties of bounding operations (see Definition IV.4 and Theorem IV.3). Here, we claim that the following properties of convex and concave relaxations are sufficient for this application.

**Definition 2.6.3** (Partition monotonic)**.** A procedure which, given any subinterval $X^\ell \subset X^0$, generates convex and concave relaxations of $\mathbf{f}$ on $X^\ell$, respectively $\mathcal{U}^\ell$ and $\mathcal{O}^\ell$, is *partition monotonic* if, for any subintervals $X^2 \subset X^1 \subset X$, $\mathcal{U}^2(\mathbf{x}) \geq \mathcal{U}^1(\mathbf{x})$ and $\mathcal{O}^2(\mathbf{x}) \leq \mathcal{O}^1(\mathbf{x})$, $\forall \mathbf{x} \in X^2$.

**Definition 2.6.4** (Partition convergent, degenerate perfect)**.** A procedure as in Definition 2.6.3 is *partition convergent* if, for any nested and convergent sequence of subintervals of $X^0$, $\{X^\ell\} \to X^*$, the sequences $\{\mathcal{U}^\ell\}$ and $\{\mathcal{O}^\ell\}$ converge to $\mathcal{U}^*$ and $\mathcal{O}^*$ uniformly on $X^*$, where $\mathcal{U}^*$ and $\mathcal{O}^*$ denote the relaxations generated on $X^*$. A procedure is *degenerate perfect* if the condition $X^* = [\mathbf{x}, \mathbf{x}]$ for any $\mathbf{x} \in X^0$ implies that $\mathcal{U}^*(\mathbf{x}) = \mathbf{f}(\mathbf{x}) = \mathcal{O}^*(\mathbf{x})$.

Below, it is shown that standard McCormick relaxations are partition monotonic, partition convergent and degenerate perfect.

**Theorem 2.6.5.** *Standard McCormick relaxations are partition monotonic.*

*Proof.* Choose any subintervals $X^2 \subset X^1 \subset X^0$ and any $\mathbf{x} \in X^2$. Let $\mathcal{X}^1 = (X^1, [\mathbf{x}, \mathbf{x}])$ and $\mathcal{X}^2 = (X^2, [\mathbf{x}, \mathbf{x}])$. Then, $\mathcal{X}^2 \subset \mathcal{X}^1$. Since $X^0$ is represented in $\mathcal{D}$, so are $X^1$ and $X^2$ (Lemma 2.4.33). Then, $\mathcal{X}^1, \mathcal{X}^2 \in \mathcal{D}$. By Theorem 2.4.32, $\{\mathbf{f}\}$ is inclusion monotonic on $\mathcal{D}$, which implies that $\{\mathbf{f}\}(\mathcal{X}^2) \subset \{\mathbf{f}\}(\mathcal{X}^1)$. From this, $\mathcal{U}^2(\mathbf{x}) = \{\mathbf{f}\}^{cv}(\mathcal{X}^2) \geq \{\mathbf{f}\}^{cv}(\mathcal{X}^1) = \mathcal{U}^1(\mathbf{x})$ and $\mathcal{O}^2(\mathbf{x}) = \{\mathbf{f}\}^{cc}(\mathcal{X}^2) \leq \{\mathbf{f}\}^{cc}(\mathcal{X}^1) = \mathcal{O}^1(\mathbf{x})$. □

**Theorem 2.6.6.** *Standard McCormick relaxations are degenerate perfect.*

*Proof.* Choose any $\mathbf{x} \in X^0$. By Theorem 2.4.32, $\{\mathbf{f}\}$ is a McCormick extension of $\mathbf{f}$, so that $([\mathbf{x}, \mathbf{x}], [\mathbf{x}, \mathbf{x}]) \in \mathcal{D}$ and $\{\mathbf{f}\}([\mathbf{x}, \mathbf{x}], [\mathbf{x}, \mathbf{x}]) = ([\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x})], [\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x})])$. It follows that $\mathcal{U}^*(\mathbf{x}) = \mathbf{f}(\mathbf{x}) = \mathcal{O}^*(\mathbf{x})$. □

**Lemma 2.6.7.** *Choose any two subintervals of $X^0$ with nonempty intersection, $X^1$ and $X^2$. Given any $\epsilon > 0$, there exists $\delta > 0$ independent of $\mathbf{x}$ such that $|\mathcal{U}^1(\mathbf{x}) - \mathcal{U}^2(\mathbf{x})| \leq \epsilon$ and $|\mathcal{O}^1(\mathbf{x}) - \mathcal{O}^2(\mathbf{x})| \leq \epsilon$, for all $\mathbf{x} \in X^1 \cap X^2$, provided that $\max(||\mathbf{x}^{L,1} - \mathbf{x}^{L,2}||_\infty, ||\mathbf{x}^{U,1} - \mathbf{x}^{U,2}||_\infty) \leq \delta$.*

*Proof.* By Theorem 2.5.40, $\{\mathbf{f}\}$ is locally Lipschitz on $\mathcal{D}$. It follows that $\{\mathbf{f}\}$ is uniformly continuous on $\mathcal{D}$. Then, for any $\epsilon > 0$, there exists $\delta > 0$ such that $d_M(\{\mathbf{f}\}(\mathcal{X}^1), \{\mathbf{f}\}(\mathcal{X}^2)) \leq \epsilon$ for every $\mathcal{X}^1, \mathcal{X}^2 \in \mathcal{D}$ with $d_M(\mathcal{X}^1, \mathcal{X}^2) \leq \delta$. For any $\mathbf{x} \in X^1 \cap X^2$, choosing $\mathcal{X}_1 = (X^1, [\mathbf{x}, \mathbf{x}])$ and $\mathcal{X}_2 = (X^2, [\mathbf{x}, \mathbf{x}])$ gives $\|\mathcal{U}^1(\mathbf{x}) - \mathcal{U}^2(\mathbf{x})\|_\infty \leq \epsilon$

and $\|\mathcal{O}^1(\mathbf{x}) - \mathcal{O}^2(\mathbf{x})\|_\infty \leq \epsilon$, provided that $\max(||\mathbf{x}^{L,1} - \mathbf{x}^{L,2}||_\infty, ||\mathbf{x}^{U,1} - \mathbf{x}^{U,2}||_\infty) \leq \delta$. □

**Theorem 2.6.8.** *Standard McCormick relaxations are partition convergent.*

*Proof.* Choose any nested and convergent sequence of subintervals of $X^0$, $\{X^\ell\} \to X^*$. Given any $\epsilon > 0$, Lemma 2.6.7 provides $\delta$ such that $|\mathcal{U}^\ell(\mathbf{x}) - \mathcal{U}^*(\mathbf{x})| \leq \epsilon$ for all $\mathbf{x} \in X^*$, provided that $\max(||\mathbf{x}^{L,\ell} - \mathbf{x}^{L,*}||_\infty, ||\mathbf{x}^{U,\ell} - \mathbf{x}^{U,*}||_\infty) \leq \delta$. By the convergence of $\{X^\ell\}$, this condition must be satisfied for every $\ell$ greater than some $N$, which implies that $\{\mathcal{U}^\ell\} \to \mathcal{U}^*$ uniformly on $X^*$. The exact same proof applies to $\{\mathcal{O}^\ell\}$. □

**Remark 2.6.9.** Partition convergence was not addressed in McCormick's original work [118] and is in fact stronger than what is necessary to ensure the convergence of spatial branch-and-bound algorithms using standard McCormick relaxations [84].

## 2.7 Generalized McCormick Relaxations

In this section, the concept of a generalized McCormick relaxation is introduced. Given a function $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$, the standard McCormick relaxations essentially take an interval $X$ and a point $\mathbf{x} \in X$, and return values of convex and concave relaxations of $\mathbf{f}$ on $X$, evaluated at $\mathbf{x}$. In the development so far, this has been represented by initializing the McCormick evaluation of the computational sequence describing $\mathbf{f}$ with an element of $\mathbb{MR}^n$ of the form $(X, [\mathbf{x}, \mathbf{x}])$. Within this context, the key observation behind generalized McCormick relaxations is that this is not the only useful initialization. Of course, the idea that we may evaluate a natural McCormick extension beginning from any element of $\mathbb{MR}^n$ is no surprise; it is the definition. What is perhaps more surprising is that, in some particular cases, very useful interpretations can be attached to these alternate initializations. This section details two simple applications, and others will appear in later chapters when the task of relaxing the solutions of dynamic systems is taken up. Of the two topics below, the notion of composite relaxations is the essential contribution.

**Definition 2.7.1.** Define MC : $\mathbb{R}^{4n} \to \mathbb{MR}^n$ be defined by

$$\mathrm{MC}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc}) \equiv (\square(\mathbf{x}^L, \mathbf{x}^U), \square(\mathbf{x}^L, \mathbf{x}^U) \tilde{\cap} \square(\mathbf{x}^{cv}, \mathbf{x}^{cc})). \tag{2.79}$$

**Definition 2.7.2.** Let $(\mathcal{S}, \pi_o)$ be a $\mathcal{L}$-computational sequence with natural function $(\mathbf{f}_{\mathcal{S}}, D_{\mathcal{S}}, \mathbb{R}^m)$ and natural McCormick extension $(\mathcal{F}_{\mathcal{S}}, \mathcal{D}_{\mathcal{S}}, \mathbb{MR}^{n_o})$. Define

$$\tilde{\Phi}_{\mathcal{S}} \equiv \{(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc}) \in \mathbb{R}^{4n_i} : [\mathbf{x}^L, \mathbf{x}^U] \in \mathbb{I}D_{\mathcal{S}} \text{ is represented in } \mathcal{D}_{\mathcal{S}}\}. \tag{2.80}$$

Define the functions $\tilde{\mathcal{U}}, \tilde{\mathcal{O}} : \tilde{\Phi} \to \mathbb{R}^n$ by

$$\tilde{\mathcal{U}}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc}) = \mathcal{F}^{cv}_{\mathcal{S}}(\mathrm{MC}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc})), \tag{2.81}$$

$$\tilde{\mathcal{O}}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc}) = \mathcal{F}^{cc}_{\mathcal{S}}(\mathrm{MC}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc})). \tag{2.82}$$

The functions $\tilde{\mathcal{U}}$ and $\tilde{\mathcal{O}}$ are called the *generalized McCormick relaxations of* $(\mathcal{S}, \pi_o)$.

Before considering specific applications in Sections 2.7.1 and 2.7.2, we collect the fundamental properties of generalized McCormick relaxations below.

**Lemma 2.7.3.** *Let $(\mathcal{S}, \pi_o)$ be a $\mathcal{L}$-computational sequence with the natural function $(\mathbf{f}_{\mathcal{S}}, D_{\mathcal{S}}, \mathbb{R}^{n_o})$ and natural McCormick extension $(\mathcal{F}_{\mathcal{S}}, \mathcal{D}_{\mathcal{S}}, \mathbb{MR}^{n_o})$. Let $X = [\mathbf{x}^L, \mathbf{x}^U] \subset \mathbb{I}D_{\mathcal{S}}$ be represented in $\mathcal{D}_{\mathcal{S}}$ and let $\mathbf{x}^{cv}, \mathbf{x}^{cc} \in \mathbb{R}^{n_i}$ satisfy $\mathbf{x}^{cv} \leq \mathbf{x}^{cc}$. Then*

$$\tilde{\mathcal{U}}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc}) \leq \mathbf{f}_{\mathcal{S}}(\mathbf{x}) \leq \tilde{\mathcal{O}}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc}), \quad \forall \mathbf{x} \in X \cap [\mathbf{x}^{cv}, \mathbf{x}^{cc}].$$

*Proof.* Let $\mathcal{X} \equiv \mathrm{MC}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc}) = (X, X \tilde{\cap} [\mathbf{x}^{cv}, \mathbf{x}^{cc}])$. If $X \cap [\mathbf{x}^{cv}, \mathbf{x}^{cc}]$ is empty, then the conclusion trivially holds. Assuming otherwise, it follows that

$$X \cap [\mathbf{x}^{cv}, \mathbf{x}^{cc}] = X \tilde{\cap} [\mathbf{x}^{cv}, \mathbf{x}^{cc}] \tag{2.83}$$

$$= [\mathrm{mid}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}), \mathrm{mid}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cc})] \tag{2.84}$$

$$= [\max(\mathbf{x}^L, \mathbf{x}^{cv}), \min(\mathbf{x}^U, \mathbf{x}^{cc})], \tag{2.85}$$

Choosing any $\mathbf{x} \in X \cap [\mathbf{x}^{cv}, \mathbf{x}^{cc}]$, it follows that $\mathbf{x} \in \mathrm{Enc}(\mathcal{X})$. By Theorem 2.4.32,

$\mathcal{F}_{\mathcal{S}}$ is a natural McCormick extension of $\mathbf{f}_{\mathcal{S}}$, and is inclusion monotonic on $\mathcal{D}_{\mathcal{S}}$. Since $X \in \mathbb{I}D_{\mathcal{S}}$ is represented in $\mathcal{D}_{\mathcal{S}}$, it follows that $\mathcal{X} \in \mathcal{D}_{\mathcal{S}} \cap \mathbb{M}D_{\mathcal{S}}$. Then, since $\mathbf{x} \in \mathrm{Enc}(\mathcal{X})$, Theorem 2.4.14 implies that $\mathbf{f}_{\mathcal{S}}(\mathbf{x}) \in \mathrm{Enc}(\mathcal{F}_{\mathcal{S}}(\mathcal{X}))$, which gives the desired inequalities. □

**Lemma 2.7.4.** *Let $(\mathcal{S}, \pi_o)$ be a $\mathcal{L}$-computational sequence with the natural function $(\mathbf{f}_{\mathcal{S}}, D_{\mathcal{S}}, \mathbb{R}^{n_o})$ and the natural McCormick extension $(\mathcal{F}_{\mathcal{S}}, \mathcal{D}_{\mathcal{S}}, \mathbb{MR}^{n_o})$. Let $X = [\mathbf{x}^L, \mathbf{x}^U] \subset \mathbb{I}D_{\mathcal{S}}$ be represented in $\mathcal{D}_{\mathcal{S}}$. Let $\mathbf{x}_i^{cv}, \mathbf{x}_i^{cc} \in \mathbb{R}^{n_i}$, $i \in \{1,2,3\}$, satisfy $\mathbf{x}_i^{cv} \leq \mathbf{x}_i^{cc}$ and $X \cap [\mathbf{x}_i^{cv}, \mathbf{x}_i^{cc}] \neq \emptyset$, $\forall i \in \{1,2,3\}$. Choose any $\lambda \in [0,1]$ and suppose that*

$$\mathbf{x}_3^{cv} \leq \lambda \mathbf{x}_1^{cv} + (1-\lambda)\mathbf{x}_1^{cv} \quad \text{and} \quad \mathbf{x}_3^{cc} \geq \lambda \mathbf{x}_1^{cc} + (1-\lambda)\mathbf{x}_1^{cc}. \tag{2.86}$$

*Then*

$$\tilde{\mathcal{U}}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}_3^{cv}, \mathbf{x}_3^{cc}) \leq \lambda \tilde{\mathcal{U}}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}_1^{cv}, \mathbf{x}_1^{cc}) + (1-\lambda)\tilde{\mathcal{U}}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}_2^{cv}, \mathbf{x}_2^{cc}), \tag{2.87}$$

$$\tilde{\mathcal{O}}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}_3^{cv}, \mathbf{x}_3^{cc}) \geq \lambda \tilde{\mathcal{O}}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}_1^{cv}, \mathbf{x}_1^{cc}) + (1-\lambda)\tilde{\mathcal{O}}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}_2^{cv}, \mathbf{x}_2^{cc}). \tag{2.88}$$

*Proof.* Define $\bar{\mathbf{x}}_i^{cv} = \mathrm{mid}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}_i^{cv})$ and $\bar{\mathbf{x}}_i^{cc} = \mathrm{mid}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}_i^{cc})$, $\forall i \in \{1,2,3\}$. Under the given hypotheses,

$$X \tilde{\cap} [\mathbf{x}_i^{cv}, \mathbf{x}_i^{cc}] = [\bar{\mathbf{x}}_i^{cv}, \bar{\mathbf{x}}_i^{cc}] = [\max(\mathbf{x}^L, \mathbf{x}_i^{cv}), \min(\mathbf{x}^U, \mathbf{x}_i^{cc})], \quad \forall i \in \{1,2,3\}. \tag{2.89}$$

From (2.86) and the fact that $\max(\mathbf{x}^L, \cdot)$ and $\min(\mathbf{x}^U, \cdot)$ are monotonic and, respectively, convex and concave on $\mathbb{R}$, it follows that

$$\bar{\mathbf{x}}_3^{cv} \leq \lambda \bar{\mathbf{x}}_1^{cv} + (1-\lambda)\bar{\mathbf{x}}_1^{cv} \quad \text{and} \quad \bar{\mathbf{x}}_3^{cc} \geq \lambda \bar{\mathbf{x}}_1^{cc} + (1-\lambda)\bar{\mathbf{x}}_1^{cc}. \tag{2.90}$$

Defining $\mathcal{X}_i = \mathrm{MC}(\mathbf{x}_i^L, \mathbf{x}_i^U, \mathbf{x}_i^{cv}, \mathbf{x}_i^{cc})$ for all $i \in \{1,2,3\}$, this further implies that $\mathcal{X}_3 \supset \mathrm{Conv}(\lambda, \mathcal{X}_1, \mathcal{X}_2)$. By Theorem 2.4.32, $\mathcal{F}_{\mathcal{S}}$ is inclusion monotonic and coherently concave on $\mathcal{D}_{\mathcal{S}}$. Then,

$$\mathcal{F}_{\mathcal{S}}(\mathcal{X}_3) \supset \mathcal{F}_{\mathcal{S}}(\mathrm{Conv}(\lambda, \mathcal{X}_1, \mathcal{X}_2)) \supset \mathrm{Conv}(\lambda, \mathcal{F}_{\mathcal{S}}(\mathcal{X}_1), \mathcal{F}_{\mathcal{S}}(\mathcal{X}_2)). \tag{2.91}$$

But this implies (2.88) and (2.87). □

**Theorem 2.7.5.** *Let $(\mathcal{S}, \pi_o)$ be a $\mathcal{L}$-computational sequence with the natural function $(\mathbf{f}_{\mathcal{S}}, D_{\mathcal{S}}, \mathbb{R}^{n_o})$ and the natural McCormick extension $(\mathcal{F}_{\mathcal{S}}, \mathcal{D}_{\mathcal{S}}, \mathbb{MR}^{n_o})$. Let $P \subset \mathbb{R}^{n_p}$ be convex, let $X = [\mathbf{x}^L, \mathbf{x}^U] \subset \mathbb{I}D_{\mathcal{S}}$, and let $\mathbf{x}, \mathbf{x}^c, \mathbf{x}^C : P \to \mathbb{R}^{n_i}$ be such that $\mathbf{x}(P) \subset X$ and $\mathbf{x}^c$ and $\mathbf{x}^C$ are, respectively, convex and concave relaxations of $\mathbf{x}$ on $P$. If $X$ is represented in $\mathcal{D}_{\mathcal{S}}$, then the functions $\mathcal{O}, \mathcal{U} : P \to \mathbb{R}^{n_o}$ defined by*

$$\mathcal{U}(\mathbf{p}) = \tilde{\mathcal{U}}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}(\mathbf{p}), \mathbf{x}^{cc}(\mathbf{p})), \tag{2.92}$$

$$\mathcal{O}(\mathbf{p}) = \tilde{\mathcal{O}}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}(\mathbf{p}), \mathbf{x}^{cc}(\mathbf{p})), \tag{2.93}$$

*are convex and concave relaxations of $\mathbf{f}_{\mathcal{S}} \circ \mathbf{x}$ on $P$, respectively.*

*Proof.* From the hypotheses, $\mathbf{x}(\mathbf{p}) \in X \cap [\mathbf{x}^{cv}(\mathbf{p}), \mathbf{x}^{cc}(\mathbf{p})]$ and $\mathbf{x}^{cv}(\mathbf{p}) \leq \mathbf{x}^{cc}(\mathbf{p})$, $\forall \mathbf{p} \in P$. Using the hypotheses on $X$, Lemma 2.7.3 gives

$$\tilde{\mathcal{U}}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}(\mathbf{p}), \mathbf{x}^{cc}(\mathbf{p})) \leq \mathbf{f}_{\mathcal{S}}(\mathbf{x}(\mathbf{p})) \leq \tilde{\mathcal{O}}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}(\mathbf{p}), \mathbf{x}^{cc}(\mathbf{p})), \quad \forall \mathbf{p} \in P.$$

Since $\mathbf{x}^{cv}$ and $\mathbf{x}^{cc}$ are,respectively, convex and concave, convexity and concavity of $\tilde{\mathcal{U}}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}(\cdot), \mathbf{x}^{cc}(\cdot))$ and $\tilde{\mathcal{O}}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}(\cdot), \mathbf{x}^{cc}(\cdot))$ follows from Lemma 2.7.4. □

The following regularity result is a consequence of Theorem 2.5.40.

**Corollary 2.7.6.** *Let $(\mathcal{S}, \pi_o)$ be a $\mathcal{L}$-computational sequence. The generalized McCormick relaxations of $(\mathcal{S}, \pi_o)$ are locally Lipschitz on $\tilde{\Phi}_{\mathcal{S}}$.*

*Proof.* By Lemmas 2.5.19 and 2.5.25, it is clear that MC is locally Lipschitz on $\mathbb{R}^{4n_i}$. For any $(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc}) \in \tilde{\Phi}_{\mathcal{S}}$, $[\mathbf{x}^L, \mathbf{x}^U]$ is represented in $\mathcal{D}_{\mathcal{S}}$. It follows that MC maps $\tilde{\Phi}_{\mathcal{S}}$ into $\mathcal{D}_{\mathcal{S}}$. By Theorem 2.5.6, the composition $\mathcal{F}_{\mathcal{S}} \circ \text{MC}$ must also be locally Lipschitz on $\tilde{\Phi}_{\mathcal{S}}$. By the definition of the generalized McCormick relaxations, this establishes the result. □

The following corollary now follows immediately from Theorem 2.5.7.

**Corollary 2.7.7.** *Let $(\mathcal{S}, \pi_o)$ be a $\mathcal{L}$-computational sequence. The generalized McCormick relaxations of $(\mathcal{S}, \pi_o)$ are Lipschitz on any compact subset of $\tilde{\Phi}_{\mathcal{S}}$.*

**Corollary 2.7.8.** *Let $(\mathcal{S}, \pi_o)$ be a $\mathcal{L}$-computational sequence. Let $K^B \subset \mathbb{R}^{2n}$ be compact and let*

$$K \equiv \{(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc}) \in \tilde{\Phi}_{\mathcal{S}} : (\mathbf{x}^L, \mathbf{x}^U) \in K_B\}.$$

*Then the generalized McCormick relaxations of $(\mathcal{S}, \pi_o)$ are Lipschitz on $K$.*

*Proof.* Let $(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc}) \in K$. By definition $\mathbf{x}^L \leq \mathbf{x}^U$ and hence $\mathrm{MC}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc}) = ([\mathbf{x}^L, \mathbf{x}^U], [\hat{\mathbf{x}}^{cv}, \hat{\mathbf{x}}^{cc}])$ for some $\hat{\mathbf{x}}^{cv}, \hat{\mathbf{x}}^{cc} \in \mathbb{R}^n$. By the definition of MC, it follows that $\hat{\mathbf{x}}^{cv} \leq \hat{\mathbf{x}}^{cc}$ and $[\hat{\mathbf{x}}^{cv}, \hat{\mathbf{x}}^{cc}] \subset [\mathbf{x}^L, \mathbf{x}^U]$. But for any such values, the definition of MC further shows that $\mathrm{MC}(\mathbf{x}^L, \mathbf{x}^U, \hat{\mathbf{x}}^{cv}, \hat{\mathbf{x}}^{cc}) = ([\mathbf{x}^L, \mathbf{x}^U], [\hat{\mathbf{x}}^{cv}, \hat{\mathbf{x}}^{cc}])$. From this, it follows that, for any $(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc}) \in K$,

$$\mathrm{MC} \circ i_{\mathbb{R}} \circ \mathrm{MC}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc}) = \mathrm{MC}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc}). \tag{2.94}$$

Then,

$$\mathcal{F}_{\mathcal{S}} \circ \mathrm{MC} \circ i_{\mathbb{R}} \circ \mathrm{MC}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc}) = \mathcal{F}_{\mathcal{S}} \circ \mathrm{MC}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc}), \tag{2.95}$$

and hence it suffice to show that $\mathcal{F}_{\mathcal{S}} \circ \mathrm{MC}$ is Lipschitz on

$$K_{\mathrm{MC}} \equiv \{i_{\mathbb{R}} \circ \mathrm{MC}(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc}) : (\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc}) \in K\} \tag{2.96}$$

But

$$K_{\mathrm{MC}} \subset \{(\mathbf{x}^L, \mathbf{x}^U, \mathbf{x}^{cv}, \mathbf{x}^{cc}) \in \tilde{\Phi}_{\mathcal{S}} : (\mathbf{x}^L, \mathbf{x}^U) \in K_B,\ \mathbf{x}^{cv}, \mathbf{x}^{cc} \in [\mathbf{x}^L, \mathbf{x}^U]\}, \tag{2.97}$$

and this latter set is closed and bounded and hence compact. Then $\mathcal{F}_{\mathcal{S}} \circ \mathrm{MC}$ is Lipschitz on $K_{\mathrm{MC}}$ by Corollary 2.7.7. □

### 2.7.1 Partial Relaxations

For some applications, it is desirable to relax a function with respect to only some of its arguments. Such relaxations are defined as follows.

**Definition 2.7.9.** Let $S \subset \mathbb{R}^{n_s}$, $D \subset \mathbb{R}^{n_y}$ and suppose that $\mathbf{f} : S \times D \to \mathbb{R}^m$. For any convex $Y \subset D$, two functions $\mathcal{U}, \mathcal{O} : S \times Y \to \mathbb{R}^m$ are called *partial relaxations* of $\mathbf{f}$ on $S \times Y$ if, for every $\mathbf{s} \in S$, $\mathcal{U}(\mathbf{s}, \cdot)$ and $\mathcal{O}(\mathbf{s}, \cdot)$ are convex and concave relaxations of $\mathbf{f}(\mathbf{s}, \cdot)$ on $Y$, respectively.

When $\mathbf{f}$ is $\mathcal{L}$-factorable and $Y$ is an interval, partial relaxations can be readily obtained from a natural McCormick extension.

**Theorem 2.7.10.** *Let $S \subset \mathbb{R}^{n_s}$, $D \subset \mathbb{R}^{n_y}$ and suppose that $\mathbf{f} : S \times D \to \mathbb{R}^m$ is $\mathcal{L}$-factorable. Let $\{\mathbf{f}\} : \mathcal{D} \to \mathbb{MR}^m$ be a natural McCormick extension of $\mathbf{f}$. For any $Y = [\mathbf{y}^L, \mathbf{y}^U] \in \mathbb{I}D$ such that $[\mathbf{s}, \mathbf{s}] \times Y$ is represented in $\mathcal{D}$ for every $\mathbf{s} \in S$, the functions $\mathcal{O}, \mathcal{U} : S \times Y \to \mathbb{R}^m$ defined by*

$$\mathcal{U}(\mathbf{s}, \mathbf{y}) \equiv \tilde{\mathcal{U}}(\mathbf{s}, \mathbf{y}^L, \mathbf{s}, \mathbf{y}^U, \mathbf{s}, \mathbf{y}, \mathbf{s}, \mathbf{y}) = \{\mathbf{f}\}^{cv}(([\mathbf{s}, \mathbf{s}], [\mathbf{s}, \mathbf{s}]), (Y, [\mathbf{y}, \mathbf{y}])), \tag{2.98}$$

$$\mathcal{O}(\mathbf{s}, \mathbf{y}) \equiv \tilde{\mathcal{O}}(\mathbf{s}, \mathbf{y}^L, \mathbf{s}, \mathbf{y}^U, \mathbf{s}, \mathbf{y}, \mathbf{s}, \mathbf{y}) = \{\mathbf{f}\}^{cc}(([\mathbf{s}, \mathbf{s}], [\mathbf{s}, \mathbf{s}]), (Y, [\mathbf{y}, \mathbf{y}])), \tag{2.99}$$

*are partial relaxations of $\mathbf{f}$ on $S \times Y$.*

*Proof.* The result follows directly from Theorem 2.7.5. In particular, fix any $\hat{\mathbf{s}} \in S$ and consider the definitions $P = [\hat{\mathbf{s}}, \hat{\mathbf{s}}] \times Y$, $X = [\hat{\mathbf{s}}, \hat{\mathbf{s}}] \times Y$ and $\mathbf{x}(\mathbf{s}, \mathbf{y}) = \mathbf{x}^{cv}(\mathbf{s}, \mathbf{y}) = \mathbf{x}^{cc}(\mathbf{s}, \mathbf{y}) = (\mathbf{s}, \mathbf{y})$, $\forall (\mathbf{s}, \mathbf{y}) \in [\hat{\mathbf{s}}, \hat{\mathbf{s}}] \times Y$. □

**Corollary 2.7.11.** *Define $\mathcal{U}, \mathcal{O} : S \times Y \to \mathbb{R}^m$ as in Theorem 2.7.10. Then $\mathcal{U}$ and $\mathcal{O}$ are Lipschitz on $S \times X$.*

*Proof.* Let $(\{\mathbf{f}\}, \mathcal{D}, \mathbb{MR}^m)$ be the natural McCormick relaxation of $\mathbf{f}$ defining $\tilde{\mathcal{U}}$ and $\tilde{\mathcal{O}}$, and let

$$K \equiv \{(\mathbf{s}^L, \mathbf{y}^L, \mathbf{s}^U, \mathbf{y}^U, \mathbf{s}^{cv}, \mathbf{y}^{cv}, \mathbf{s}^{cc}, \mathbf{y}^{cc}) \in \mathbb{R}^{4(n_s+n_y)} : \tag{2.100}$$
$$\mathbf{s}^L = \mathbf{s}^U = \mathbf{s}^{cv} = \mathbf{s}^{cc} \in S,\ \mathbf{y}^{cv}, \mathbf{y}^{cc} \in [\mathbf{y}^L, \mathbf{y}^U] \subset Y\}.$$

$K$ is clearly compact, and it is a subset of $\tilde{\Phi}$ by the assumption that $[\mathbf{s}, \mathbf{s}] \times Y$ is represented in $\mathcal{D}$ for every $\mathbf{s} \in S$. Then, by Corollary 2.7.7, $\tilde{U}$ and $\tilde{O}$ are Lipschitz on $K$. But for every $(\mathbf{s}, \mathbf{y}) \in S \times Y$, $(\mathbf{s}, \mathbf{y}^L, \mathbf{s}, \mathbf{y}^U, \mathbf{s}, \mathbf{y}, \mathbf{s}, \mathbf{y})$ is in $K$, and it follows that $\mathcal{U}$ and $\mathcal{O}$ are Lipschitz on $S \times Y$. □

The key result above is the Lipschitz dependence on $S$. Since changing $\mathbf{s}$ requires changing the bounds $\mathbf{s}^L$ and $\mathbf{s}^U$ in the standard relaxation, this result cannot be proven directly by the continuity result in the standard framework.

### 2.7.2 Composite Relaxations

In this section, we demonstrate the use of generalized McCormick relaxations to compute convex and concave relaxations of composite functions. This method is used extensively to derive relaxations for the solutions of dynamic systems in later chapters.

**Definition 2.7.12.** Let $P \subset \mathbb{R}^{n_p}$ be convex, $D \subset \mathbb{R}^{n_y}$ and $\mathbf{f} : P \times D \to \mathbb{R}^m$. For any set $Y \subset D$, functions $\mathbf{u_f}, \mathbf{o_f} : P \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} \to \mathbb{R}$ are called *convex and concave composite relaxations of* $\mathbf{f}$ on $P \times Y$ if the following condition holds: For any $\mathbf{y}, \mathbf{y}^c, \mathbf{y}^C : P \to \mathbb{R}^{n_y}$ with $\mathbf{y}(P) \subset Y$, convex and concave relaxations of the composite function

$$P \ni \mathbf{p} \longmapsto \mathbf{g}(\mathbf{p}) \equiv \mathbf{f}(\mathbf{p}, \mathbf{y}(\mathbf{p})) \tag{2.101}$$

on $P$ are given by the composite mappings

$$\begin{aligned} P \ni \mathbf{p} &\longmapsto \mathbf{g}^{cv}(\mathbf{p}) \equiv \mathbf{u_f}(\mathbf{p}, \mathbf{y}^c(\mathbf{p}), \mathbf{y}^C(\mathbf{p})) \\ P \ni \mathbf{p} &\longmapsto \mathbf{g}^{cc}(\mathbf{p}) \equiv \mathbf{o_f}(\mathbf{p}, \mathbf{y}^c(\mathbf{p}), \mathbf{y}^C(\mathbf{p})) \end{aligned} \tag{2.102}$$

provided that $\mathbf{y}^c$ and $\mathbf{y}^C$ are, respectively, convex and concave relaxations of $\mathbf{y}$ on $P$.

When $\mathbf{f}$ is $\mathcal{L}$-factorable and $P$ and $Y$ are intervals, composite relaxations can be readily obtained from a natural McCormick extension as follows.

**Theorem 2.7.13.** *Let $P \in \mathbb{IR}^{n_p}$ and $D \subset \mathbb{R}^{n_y}$. Suppose that $\mathbf{f} : P \times D \to \mathbb{R}^m$ is $\mathcal{L}$-factorable and let $\{\mathbf{f}\} : \mathcal{D} \to \mathbb{MR}^m$ be a natural McCormick extension of $\mathbf{f}$. For any $Y \in \mathbb{I}D$ such that $P \times Y$ is represented in $\mathcal{D}$, the functions $\mathbf{u_f}, \mathbf{o_f} : P \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} \to \mathbb{R}^m$ defined by*

$$\begin{aligned}\mathbf{u_f}(\mathbf{p}, \mathbf{z}^{cv}, \mathbf{z}^{cc}) &\equiv \tilde{\mathcal{U}}(\mathbf{p}^L, \mathbf{y}^L, \mathbf{p}^U, \mathbf{y}^U, \mathbf{p}, \mathbf{z}^{cv}, \mathbf{p}, \mathbf{z}^{cc}), \qquad (2.103)\\ &= \{\mathbf{f}\}^{cv}((P, [\mathbf{p}, \mathbf{p}]), \mathrm{MC}(\mathbf{y}^L, \mathbf{y}^U, \mathbf{z}^{cv}, \mathbf{z}^{cc})),\\ \mathbf{o_f}(\mathbf{p}, \mathbf{z}^{cv}, \mathbf{z}^{cc}) &\equiv \tilde{\mathcal{O}}(\mathbf{p}^L, \mathbf{y}^L, \mathbf{p}^U, \mathbf{y}^U, \mathbf{p}, \mathbf{z}^{cv}, \mathbf{p}, \mathbf{z}^{cc}),\\ &= \{\mathbf{f}\}^{cc}((P, [\mathbf{p}, \mathbf{p}]), \mathrm{MC}(\mathbf{y}^L, \mathbf{y}^U, \mathbf{z}^{cv}, \mathbf{z}^{cc})),\end{aligned}$$

*are composite relaxations of $\mathbf{f}$ on $P \times Y$.*

*Proof.* Let $\mathbf{y}, \mathbf{y}^{cv}, \mathbf{y}^{cc} : P \to \mathbb{R}^{n_y}$ be such that $\mathbf{y}(P) \subset Y$ and $\mathbf{y}^{cv}$ and $\mathbf{y}^{cc}$ are, respectively, convex and concave relaxations of $\mathbf{y}$ on $P$. Then, the result follows directly from Theorem 2.7.5 with the definitions $P = P$, $X = P \times Y$, $\mathbf{x}(\mathbf{p}) = (\mathbf{p}, \mathbf{y}(\mathbf{p}))$, $\mathbf{x}^{cv}(\mathbf{p}) = (\mathbf{p}, \mathbf{y}^{cv}(\mathbf{p}))$ and $\mathbf{x}^{cc}(\mathbf{p}) = (\mathbf{p}, \mathbf{y}^{cc}(\mathbf{p}))$. □

As with standard McCormick relaxations, it can be shown that generating relaxations of the composite function $\mathbf{g}$ in (2.101) via Theorem 2.7.13 is a partition monotonic, partition convergent and degenerate perfect procedure, provided that the relaxations $\mathbf{y}^{cv}$ and $\mathbf{y}^{cc}$ are generated by a partition monotonic, partition convergent and degenerate perfect procedure. Let $P \in \mathbb{IR}^{n_p}$, $D \subset \mathbb{R}^{n_y}$, $\mathbf{f} : P \times D \to \mathbb{R}^m$ and $\mathbf{y} : P \to D$, and define $\mathbf{g} : P \to \mathbb{R}^m$ as in (2.101). Let $\{\mathbf{f}\} : \mathcal{D} \to \mathbb{MR}^m$ be a natural McCormick extension of $\mathbf{f}$, and let $Y \in \mathbb{I}D$ be such that $P \times Y$ is represented in $\mathcal{D}$. Now, consider a nested and convergent sequence of subintervals of $P$, $\{P^\ell\} \to P^*$. The following assumption is required.

**Assumption 2.7.14.** For any subinterval $P^\ell \subset P$, valid bounds for $\mathbf{y}$ on $P^\ell$, $Y^\ell = [\mathbf{y}^{L,\ell}, \mathbf{y}^{U,\ell}]$, are available. Moreover, for any nested and convergent sequence of subintervals of $P$, $\{P^\ell\} \to P^*$,

1. $Y^{\ell+1} \subset Y^\ell \subset Y$, for any $\ell \in \mathbb{N}$,

2. $\{Y^\ell\} \to Y^*$, and
3. $P^* = [\mathbf{p}, \mathbf{p}]$ for some $\mathbf{p} \in P$ implies that $Y^* = [\mathbf{y}(\mathbf{p}), \mathbf{y}(\mathbf{p})]$.

For any $P^\ell$ ($P^*$), it is now sensible to define the functions $\mathbf{g}^{cv,\ell}$ and $\mathbf{g}^{cc,\ell}$ ($\mathbf{g}^{cv,*}$ and $\mathbf{g}^{cc,*}$) as in (2.102) and (2.103) with $P^\ell$ and $Y^\ell$ ($P^*$ and $Y^*$) in place of $P$ and $Y$.

**Theorem 2.7.15.** *Suppose that, given any interval $P^\ell \subset P$, convex and concave relaxations of $\mathbf{y}$ on $P^\ell$, $\mathbf{y}^{cv,\ell}$ and $\mathbf{y}^{cc,\ell}$, respectively, are available through a procedure which is partition monotonic. Then generating convex and concave relaxations of $\mathbf{g}$ on $P^\ell$ by* (2.102) *and* (2.103) *is a partition monotonic procedure.*

*Proof.* Choose any subintervals $P^2 \subset P^1 \subset P$ and any $\mathbf{p} \in P^2$. For $\ell \in \{1, 2\}$, define

$$\mathcal{P}^\ell = (P^\ell, [\mathbf{p}, \mathbf{p}]), \quad \mathcal{Y}^\ell = \mathrm{MC}(\mathbf{y}^{L,\ell}, \mathbf{y}^{L,\ell}, \mathbf{y}^{cv,\ell}(\mathbf{p}), \mathbf{y}^{cc,\ell}(\mathbf{p})).$$

Clearly, $\mathcal{P}^2 \subset \mathcal{P}^1$. By Condition 1 of Assumption 2.7.14, $Y^2 \subset Y^1$. Furthermore, since $\mathbf{y}^{cv,\ell}$ and $\mathbf{y}^{cc,\ell}$ are generated by a partition monotonic procedure, $\mathbf{y}^{cv,2}(\mathbf{p}) \geq \mathbf{y}^{cv,1}(\mathbf{p})$ and $\mathbf{y}^{cc,2}(\mathbf{p}) \leq \mathbf{y}^{cc,1}(\mathbf{p})$. Then, it is not difficult to see that $\mathcal{Y}^2 \subset \mathcal{Y}^1$.

Since $P \times Y$ is represented in $\mathcal{D}$, so are $P^1 \times Y^1$ and $P^1 \times Y^2$ (Lemma 2.4.33). Then, $(\mathcal{P}^1, \mathcal{Y}^1), (\mathcal{P}^2, \mathcal{Y}^2) \in \mathcal{D}$. By Theorem 2.4.32, $\{\mathbf{f}\}$ is inclusion monotonic on $\mathcal{D}$, which implies that $\{\mathbf{f}\}(\mathcal{P}^2, \mathcal{Y}^2) \subset \{\mathbf{f}\}(\mathcal{P}^1, \mathcal{Y}^1)$. From this, $\mathbf{g}^{cv,2}(\mathbf{p}) = \{\mathbf{f}\}^{cv}(\mathcal{P}^2, \mathcal{Y}^2) \geq \{\mathbf{f}\}^{cv}(\mathcal{P}^1, \mathcal{Y}^1) = \mathbf{g}^{cv,1}(\mathbf{p})$ and $\mathbf{g}^{cc,2}(\mathbf{p}) = \{\mathbf{f}\}^{cc}(\mathcal{P}^2, \mathcal{Y}^2) \leq \{\mathbf{f}\}^{cc}(\mathcal{P}^1, \mathcal{Y}^1) = \mathbf{g}^{cc,1}(\mathbf{p})$. □

**Theorem 2.7.16.** *Suppose that, given any interval $Y^\ell \subset Y$, convex and concave relaxations of $\mathbf{y}$ on $Y^\ell$, $\mathbf{y}^{c,\ell}$ and $\mathbf{y}^{C,\ell}$, respectively, are available through a procedure which is partition convergent and degenerate perfect. Then generating convex and concave relaxations of $\mathbf{g}$ on $P^\ell$ by* (2.102) *and* (2.103) *is a partition convergent and degenerate perfect procedure.*

*Proof.* Consider any nested and convergent sequence of subintervals of $P$, $\{P^\ell\} \to P^*$. Choose any $\mathbf{p} \in P^*$ and, for each $\ell \in \mathbb{N}$ (and $\ell = *$), let $\mathcal{P}^\ell = (P^\ell, [\mathbf{p}, \mathbf{p}])$ and $\mathcal{Y}^\ell = \mathrm{MC}(\mathbf{y}^{L,\ell}, \mathbf{y}^{U,\ell}, \mathbf{y}^{cv,\ell}(\mathbf{p}), \mathbf{y}^{cc,\ell}(\mathbf{p}))$. By the assumption that $P \times Y$ is represented in $\mathcal{D}$, it follows that $(\mathcal{P}^\ell, \mathcal{Y}^\ell) \in \mathcal{D}$ for every $\ell \in \mathbb{N}$. Now, for each $\mathbf{p} \in P^*$, the sequence $\{(\mathcal{P}^\ell, \mathcal{Y}^\ell)\}$ must converge to $(\mathcal{P}^*, \mathcal{Y}^*)$ by the convergence of the sequences $\{P^\ell\}$, $\{Y^\ell\}$, $\{\mathbf{y}^{cc,\ell}(\mathbf{p})\}$ and $\{\mathbf{y}^{cc,\ell}(\mathbf{p})\}$. Given any $\delta > 0$, each of these sequences has some integer $N$ above which every element deviates from its limit by less than $\delta$ in the appropriate norm. Further, these integers can be chosen independently of the point $\mathbf{p} \in P^*$ because $\{\mathbf{y}^{cv,\ell}\}$ and $\{\mathbf{y}^{cc,\ell}\}$ are assumed to converge uniformly on $P^*$. Taking the largest of these integers, this implies that, given any $\delta > 0$, it is possible to find an integer $N$ for which $d_M((\mathcal{P}^\ell, \mathcal{Y}^\ell), (\mathcal{P}^*, \mathcal{Y}^*)) < \delta$ for all $\ell \geq N$ and all $\mathbf{p} \in P^*$. Now, by the uniform continuity of $\{\mathbf{f}\}$ on $\mathcal{D}$, this implies that, given any $\epsilon > 0$, there exists $\delta$ such that $d_M(\{\mathbf{f}\}(\mathcal{P}^\ell, \mathcal{Y}^\ell), \{\mathbf{f}\}(\mathcal{P}^*, \mathcal{Y}^*)) < \epsilon$ if $d_M((\mathcal{P}^\ell, \mathcal{Y}^\ell), (\mathcal{P}^*, \mathcal{Y}^*)) < \delta$, regardless of $\mathbf{p} \in P^*$. But this condition must be satisfied for large enough $\ell$. Then, for any $\epsilon > 0$, there exists $\ell \in \mathbb{N}$ large enough that $\|\mathbf{g}^{cv,\ell}(\mathbf{p}) - \mathbf{g}^{cv,*}(\mathbf{p})\|_\infty < \epsilon$ and $\|\mathbf{g}^{cc,\ell}(\mathbf{p}) - \mathbf{g}^{cc,*}(\mathbf{p})\|_\infty < \epsilon$, for all $\mathbf{p} \in P$, which is the desired result.

Now, if $P^* = [\mathbf{p}, \mathbf{p}]$ for some $\mathbf{p} \in P$, Condition 3 of Assumption 2.7.14 ensures that $Y^* = [\mathbf{y}(\mathbf{p}), \mathbf{y}(\mathbf{p})]$. Then, $\mathcal{P}^* = ([\mathbf{p}, \mathbf{p}], [\mathbf{p}, \mathbf{p}])$ and $\mathcal{Y}^* = ([\mathbf{y}(\mathbf{p}), \mathbf{y}(\mathbf{p})], [\mathbf{y}(\mathbf{p}), \mathbf{y}(\mathbf{p})])$, and the conclusion follows from the fact that $\{\mathbf{f}\}$ is a McCormick extension of $\mathbf{f}$. □

## 2.8 Univariate Interval and McCormick Extensions

In order to derive natural interval and McCormick extensions, it was necessary to assume that interval extensions and convex and concave relaxations are available for the univariate functions in $\mathcal{L}$ (Assumptions 2.3.8 and 2.4.25). In this section, this information is compiled for many of the most common univariate functions. Univariate functions not listed here can certainly be used in the methods described in this work, provided that they can be shown to satisfy all assumptions. For information on constructing convex and concave envelopes for univariate functions of interest, see [118].

### Addition of a constant

$$u(x) = x + c, \quad B = \mathbb{R}$$

$$u^L(X) = x^L + c, \quad u^U(X) = x^U + c$$

$$u^{cv}(X,x) = x + c, \quad x^{\min}(X) = x^L$$

$$u^{cc}(X,x) = x + c, \quad x^{\max}(X) = x^U$$

**Multiplication by a positive constant**

$$u(x) = cx,\ c > 0, \quad B = \mathbb{R}$$

$$u^L(X) = cx^L, \quad u^U(X) = cx^U$$

$$u^{cv}(X,x) = cx, \quad x^{\min}(X) = x^L$$

$$u^{cc}(X,x) = cx, \quad x^{\max}(X) = x^U$$

**Negative**

$$u(x) = -x, \quad B = \mathbb{R}$$

$$u^L(X) = -x^U, \quad u^U(X) = -x^L$$

$$u^{cv}(X,x) = -x, \quad x^{\min}(X) = x^U$$

$$u^{cc}(X,x) = -x, \quad x^{\max}X(X) = x^L$$

**Reciprocal**

$$u(x) = \frac{1}{x}, \quad B = \mathbb{R} - \{0\}$$

$$u^L(X) = \frac{1}{x^U}, \quad u^U(X) = \frac{1}{x^L}$$

$$u^{cv}(X,x) = \begin{cases} \frac{1}{x} & \text{if} \quad x^L > 0 \\ \frac{1}{x^L} + \frac{1/x^U - 1/x^L}{x^U - x^L}(x - x^L) & \text{if} \quad x^U < 0 \end{cases}$$

$$u^{cc}(X,x) = \begin{cases} \frac{1}{x} & \text{if} \quad x^U < 0 \\ \frac{1}{x^L} + \frac{1/x^U - 1/x^L}{x^U - x^L}(x - x^L) & \text{if} \quad x^L > 0 \end{cases}$$

$$x^{\min}(X) = x^U, \quad x^{\max}(X) = x^L$$

### Exponential

$$u(x) = e^x, \quad B = \mathbb{R}$$

$$u^L(X) = e^{x^L}, \quad u^U(X) = e^{x^U}$$

$$u^{cv}(X, x) = e^x$$

$$u^{cc}(X, x) = e^{x^L} + \frac{e^{x^U} - e^{x^L}}{x^U - x^L}(x - x^L)$$

$$x^{\min}(X) = x^L, \quad x^{\max}(X) = x^U$$

### Natural log

$$u(x) = \ln(x), \quad B = (0, +\infty)$$

$$u^L(X) = \ln(x^L), \quad u^U(X) = \ln(x^U)$$

$$u^{cv}(X, x) = \ln(x^L) + \frac{\ln(x^U) - \ln(x^L)}{x^U - x^L}(x - x^L)$$

$$u^{cc}(X, x) = \ln(x)$$

$$x^{\min}(X) = x^L, \quad x^{\max}(X) = x^U$$

### x*ln(x)

$$u(x) = x\ln(x), \quad B = (0, +\infty)$$

$$u^L(X) = x^L\ln(x^L), \quad u^U(X) = x^U\ln(x^U)$$

$$u^{cv}(X, x) = x\ln(x)$$

$$u^{cc}(X, x) = x^L\ln(x^L) + \frac{x^U\ln(x^U) - x^L\ln(x^L)}{x^U - x^L}(x - x^L)$$

$$x^{\min}(X) = x^L, \quad x^{\max}(X) = x^U$$

### Square root

$$u(x) = \sqrt{x}, \quad B = (0, +\infty)$$

The set $B$ must be restricted from $[0, +\infty)$ to $(0, +\infty)$ because $\sqrt{x}$ is not Lipschitz on any interval containing zero.

$$u^L(X) = \sqrt{x^L}, \quad u^U(X) = \sqrt{x^U}$$

$$u^{cv}(X, x) = \sqrt{x^L} + \frac{\sqrt{x^U} - \sqrt{x^L}}{x^U - x^L}(x - x^L)$$

$$u^{cc}(X, x) = \sqrt{x}$$

$$x^{\min}(X) = x^L, \quad x^{\max}(X) = x^U$$

**Even integer powers**

$$u(x) = x^n, \; n = 2, 4, \ldots, \quad B = \mathbb{R}$$

$$u^L(X) = \begin{cases} 0 & \text{if} \quad 0 \in [x^L, x^U] \\ \min((x^L)^n, (x^U)^n) & \text{otherwise} \end{cases}$$

$$u^U(X) = \max((x^L)^n, (x^U)^n)$$

$$u^{cv}(X, x) = x^n$$

$$u^{cc}(X, x) = (x^L)^n + \frac{(x^U)^n - (x^L)^n}{x^U - x^L}(x - x^L)$$

$$x^{\min}(X) = \begin{cases} 0 & \text{if} \quad 0 \in [x^L, x^U] \\ \arg\min((x^L)^n, (x^U)^n) & \text{otherwise} \end{cases}$$

$$x^{\max}(X) = \arg\max((x^L)^n, (x^U)^n)$$

**Odd integer powers**

$$u(x) = x^n, \; n = 1, 3, \ldots, \quad B = \mathbb{R}$$

$$u^L(X) = (x^L)^n$$

$$u^U(X) = (x^U)^n$$

The convex envelope is

$$u^{cv}(X,x) = \begin{cases} x^n & \text{if} \quad x \in [x^*, x^U] \\ (x^L)^n + \frac{(x^*)^n - (x^L)^n}{x^* - x^L}(x - x^L) & \text{otherwise} \end{cases},$$

where

$$x^* = \begin{cases} x^U & \text{if} \quad x^U \leq 0 \\ x^L & \text{if} \quad x^L \geq 0 \\ x' & \text{otherwise} \end{cases},$$

and $x'$ is the solution of

$$(n-1)(x')^n - nx^L(x')^{n-1} + (x^L)^n = 0.$$

The concave envelope is

$$u^{cc}(X,x) = \begin{cases} x^n & \text{if} \quad x \in [x^L, x^{**}] \\ (x^{**})^n + \frac{(x^U)^n - (x^{**})^n}{x^U - x^{**}}(x - x^{**}) & \text{otherwise} \end{cases},$$

where

$$x^{**} = \begin{cases} x^U & \text{if} \quad x^U \leq 0 \\ x^L & \text{if} \quad x^L \geq 0 \\ x'' & \text{otherwise} \end{cases},$$

and $x''$ is the solution of

$$(n-1)(x'')^n - nx^U(x'')^{n-1} + (x^U)^n = 0.$$

Finally,

$$x^{\min}(X) = x^L,$$

and

$$x^{\max}(X) = x^U.$$

**Sin**

$$u(x) = \sin(x), \quad B = \mathbb{R}$$

The convex envelope is first formulated for the case where $[x^L, x^U] \subset [3\pi/2, 7\pi/2]$. This requires definition of the following points. Let

$$x_{\text{infx},1} = 2\pi, \quad x_{\text{infx},2} = 3\pi, \quad x_{\min,1} = \tfrac{3\pi}{2}, \quad \text{and} \quad x_{\min,2} = \tfrac{7\pi}{2}.$$

Next, define

$$\begin{aligned} x^* &= \begin{cases} x_{\text{infx},1} & \text{if} \quad x^U \leq x_{\text{infx},1} \\ x' & \text{otherwise} \end{cases}, \\ x^{**} &= \begin{cases} x_{\text{infx},2} & \text{if} \quad x^L \geq x_{\text{infx},2} \\ x'' & \text{otherwise} \end{cases}, \end{aligned}$$

where $x'$ is the solution of

$$\sin(x^U) - \sin(x') = (x^U - x')\cos x'$$

on $[x_{\min,1}, x_{\text{infx},1}]$ and $x''$ is the solution of

$$\sin(x'') - \sin(x^L) = (x'' - x^L)\cos x''$$

on $[x_{\text{infx},2}, x_{\min,2}]$. Now let $x_1$ and $x_2$ be defined by

$$x_1 = \text{mid}(x^L, x^U, x^*), \quad x_2 = \text{mid}(x^L, x_U, x^{**}).$$

Consider the function

$$\eta(X, x) = \begin{cases} \sin(x) & \text{for} \quad x \in [x^L, x_1] \\ \sin(x_1) + \frac{\sin(x_2) - \sin(x_1)}{x_2 - x_1}(x - x_1) & \text{for} \quad x \in (x_1, x_2] \\ \sin(x) & \text{for} \quad x \in (x_2, x^U] \end{cases}.$$

It can be verified that $\eta(X,\cdot)$ is the convex envelope of sin on $[x^L, x^U]$ provided that $[x^L, x^U] \subset [3\pi/2, 7\pi/2]$. The convex envelope on any other interval can be obtained by simple variable transformations and applications of $\eta$, as follows. Let $n(x) = \frac{1}{2\pi}x + \frac{1}{4}$ and define $n_1 = \lfloor n(x^L) \rfloor$. Further, let $n_2$ equal $n(x^U) - 1$ if $n(x^U)$ is an integer and $\lfloor n(x^U) \rfloor$ otherwise. Finally, define

$$\begin{aligned}
z^L &= x^L - 2(n_1 - 1)\pi, \\
z^U &= \min(x^U - 2(n_1 - 1)\pi, x_{\min,2}), \\
y^L &= x_{\min,1}, \\
y^U &= x^U - 2(n_2 - 1)\pi.
\end{aligned}$$

The convex envelope of sin on an arbitrary interval is now stated as

$$u^{cv}(X, x) = \begin{cases} \eta(Z, x - 2(n_1 - 1)\pi) & \text{if} \quad x - 2(n_1 - 1)\pi \le x_{\min,2} \\ \eta(Y, x - 2(n_2 - 1)\pi) & \text{if} \quad x - 2(n_2 - 1)\pi \ge x_{\min,1} \\ -1 & \text{otherwise} \end{cases}.$$

Similarly, the lower bound on an arbitrary interval and a minimum of $u^{cv}(X,\cdot)$ are given by

$$u^L(X) = \begin{cases} -1 & \text{if} \quad x^U - 2(n_1 - 1)\pi \ge x_{\min,2} \\ \min(\sin(x^L), \sin(x^U)) & \text{otherwise} \end{cases}$$

and

$$x^{\min}(X) = \begin{cases} x_{\min,2} + 2(n_1 - 1)\pi & \text{if} \quad x^U - 2(n_1 - 1)\pi \ge x_{\min,2} \\ \arg\min(\sin(x^L), \sin(x^U)) & \text{otherwise} \end{cases}.$$

Finally, the upper bound, the concave envelope and a maximum of the concave envelope are given by the symmetry relations

$$\begin{aligned}
u^U(X) &= -u^L(-X), \\
u^{cc}(X, x) &= -u^{cv}(-X, -x), \\
x^{\max}(X) &= -x^{\min}(-X).
\end{aligned}$$

**Cos**

The bounds and relaxations for cos can be obtained from the rules for sin and the identity

$$\cos(x) = \sin(x + \frac{\pi}{2}), \quad \forall x \in \mathbb{R}.$$

## 2.9 Conclusion

In this chapter, the class of factorable functions was introduced, and it was shown that useful global information about such functions can be automatically computed. Two methods for obtaining such information were presented. The first, interval arithmetic, provides guaranteed interval bounds on the range of a factorable function over an interval of inputs. The second, McCormick's relaxation technique, provides convex and concave relaxations of factorable functions. These methods were then analyzed in detail to establish several new regularity and convergence properties that will be required in later chapters. Finally, a generalized form of McCormick's relaxation technique was introduced which extends the applicability of McCormick-type relaxations greatly. It was shown here that this technique provides relaxations of composite functions. In Chapters 7 and 8, it will be shown that this technique is also essential for relaxing the solutions of dynamic systems.

# Chapter 3

# State Bounding Theory for Parametric ODEs and Control Systems

## 3.1 Introduction

In this chapter and the next, methods are developed for efficiently computing sharp interval enclosures of the solutions of nonlinear control systems, subject to permissible sets of inputs and initial conditions. This set of solution values is called the *reachable set*, and the computed interval bounds on this set are called *state bounds*. Enclosures of the reachable sets of dynamic systems are useful in many applications, including uncertainty quantification [75], state and parameter estimation [163, 93, 103, 138, 88], safety verification [85], fault detection [106] and controller synthesis [110]. The primary motivation for computing state bounds here, however, is for their use in algorithms for global optimization of dynamic systems [135, 164, 104]. Such algorithms embed the overestimation of reachable sets as a frequently called subroutine. Accordingly, we are interested in methods that can provide enclosures quickly (order $10^{-1}$s), and focus on mitigating the overestimation that such methods are prone to. Computing sharp bounds on this time scale is a problem of general interest, both for other algorithms that embed reachable set computations and for online applications such as state estimation [138, 88] and robust model predictive control [102].

For general nonlinear control systems, theoretical characterizations of the reachable set are available in terms of invariance domains [13], solutions of integral funnel equations [133], and solutions of Hamilton-Jacobi-Bellman equations [120]. Despite this rich body of theory, methods derived from these formulations are computationally demanding. Several more tractable approaches enclose the reachable set within polytopes or zonotopes. In [41], hyperplanes supporting the reachable set are computed by solving dynamic optimization problems. A variant that produces weaker enclosures but guarantees convex optimization problems is presented in Chapter 9. Other methods involve abstraction of the nonlinear system by a hybrid system with simplified (i.e., linearized) continuous dynamics in modes corresponding to a partition of the state space [72, 10, 5]. An enclosure for this simplified system is then augmented by a bound on the abstraction error. Refinement of the partition leads to sharper enclosures, but higher computational cost. As a representative example, enclosures computed in [5] took on the order of $10^1$s, making them inappropriate for the applications of interest here, though they are indeed very sharp.

A less expensive approach is to enclose the reachable set within time-varying interval bounds. Methods of this type are either based on Taylor approximations with rigorous error bounds [130], or on viability type conditions, which in the case of interval enclosures reduce to componentwise differential inequalities [75, 162, 156, 140, 141]. A unique feature of Taylor methods is that they produce *validated* enclosures, meaning that the enclosures are guaranteed even when computed on a finite precision machine. Unfortunately, these methods apply only to ODEs that depend on real parameters rather than controls, and produce very conservative bounds when the range of parameters is large (see [140] for comparison with differential inequalities). This conservatism can be greatly mitigated by using high-order Taylor expansions, or by using more sophisticated inclusion algebras, such as Taylor model arithmetic [24, 105], in place of interval arithmetic. Unfortunately, these measures dramatically increase the computational cost, which in the latter case scales exponentially in the number of uncertain initial conditions and parameters.

The primary advantage of differential inequalities approaches is that they can be implemented using interval arithmetic and numerical integration codes, yielding bounds at a cost comparable to a single model simulation (order $10^{-4}$–$10^{-1}$s for systems with a few states). While the enclosures produced by these methods are mathematically guaranteed, they are not validated. Therefore, they are inappropriate for investigating long-time behavior of unstable or oscillatory systems. Given the accuracy of modern numerical integration codes, however, these methods are effective for stable systems over modest integration times, especially when the reachable set is large compared to the expected numerical error owing to large parameter ranges. Moreover, this issue can be overcome using a slightly more involved hybrid formulation as in [140]. Like Taylor methods, differential inequalities approaches are typically applied to parametric ODEs, but the extension to controls is less problematic (See §3.3.2). A more difficult issue is that they are known to yield extremely conservative enclosures for ODEs that are not *quasi-monotone* [182] (or *cooperative* [165]). In [162], it was shown that this condition is frequently violated in applications. On the other hand, it was also shown that it is often possible, through physical arguments, to obtain a crude set $G$ which is independently known to contain the reachable set, and that greatly improved bounds can be computed by leveraging this information. A practical implementation was developed for the case where $G$ is an interval.

In this chapter, we develop a framework for effectively using general physical information in differential inequalities bounding methods, without a significant loss of efficiency. First, the basic differential inequalities bounding method, which does not make use of physical information, is presented and its advantages and disadvantages are discussed in the context of a simple example (§3.3). The use of physical information is discussed in detail in §3.4. Unfortunately, it happens that the most intuitive usage is not always valid, and we present some choice counterexamples in order to elucidate the fundamental problems. This discussion then motivates the central part of the chapter, comprising §3.5-§3.7, which contains a detailed analysis of the use of physical information in differential inequalities through a mathematical abstraction in terms of set-valued mappings. This results in an abstract bounding theory, which clearly isolates the fundamental problems of the conceptual discussion in §3.4 and prescribes conditions under which one yet arrives at a correct bounding procedure. From these general principles, we then derive several new bounding methods making use of physical information in various forms. Some illustrative numerical examples can be found throughout, and more thorough case studies follow in Chapter §4.

## 3.2 Problem Statement

For any measurable $I \subset \mathbb{R}$, the space of Lebesgue integrable functions $u : I \to \mathbb{R}$ is denoted by $L^1(I)$. A vector function $\mathbf{u} : I \to \mathbb{R}^n$ is said to be measurable if each scalar function $u_i$ is measurable, and is said to be Lebesgue integrable if each $u_i$ is an element of $L^1(I)$. The space of Lebesgue integrable vector functions is denoted by $(L^1(I))^n$.

Let $I_0 \subset \mathbb{R}$ be open, $I = [t_0, t_f] \subset I_0$, let $U \subset \mathbb{R}^{n_u}$ be compact, and define the set of admissible controls

$$\mathcal{U} \equiv \{\mathbf{u} \in (L^1(I))^{n_u} : \mathbf{u}(t) \in U \text{ for a.e. } t \in I\}. \tag{3.1}$$

Let the set of admissible initial conditions be a compact set $X_0 \subset \mathbb{R}^{n_x}$. Finally, let $D \supset X_0$ and let $\mathbf{f} : I_0 \times U \times D \to \mathbb{R}^{n_x}$. Consider the initial value problem in ODEs

$$\begin{aligned} \dot{\mathbf{x}}(t, \mathbf{u}, \mathbf{x}_0) &= \mathbf{f}(t, \mathbf{u}(t), \mathbf{x}(t, \mathbf{u}, \mathbf{x}_0)), \qquad (3.2) \\ \mathbf{x}(t_0, \mathbf{u}, \mathbf{x}_0) &= \mathbf{x}_0, \end{aligned}$$

where a solution is any mapping $\mathbf{x} : I \times \mathcal{U} \times X_0 \to D$ such that, for each $(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$, the mapping $\mathbf{x}(\cdot, \mathbf{u}, \mathbf{x}_0)$ is absolutely continuous and satisfies (3.2) a.e. on $I$. The following assumptions hold throughout this chapter.

**Assumption 3.2.1.** Assume that

1. For any $(\mathbf{p}, \mathbf{z}) \in U \times D$, $\mathbf{f}(\cdot, \mathbf{p}, \mathbf{z})$ is measurable on $I$,

2. For a.e. $t \in I$, $\mathbf{f}(t, \cdot, \cdot)$ is continuous on $U \times D$,

3. For every compact $K \subset D$, $\exists \alpha_K \in L^1(I)$ such that, for a.e. $t \in I$,

$$\|\mathbf{f}(t, \mathbf{p}, \mathbf{z})\|_1 \leq \alpha_K(t), \quad \forall (\mathbf{p}, \mathbf{z}) \in U \times K.$$

**Assumption 3.2.2.** For any $\mathbf{z} \in D$, there exists $\eta > 0$ and $\alpha \in L^1(I)$ such that, for a.e. $t \in I$ and every $\mathbf{p} \in U$,

$$\|\mathbf{f}(t, \mathbf{p}, \tilde{\mathbf{z}}) - \mathbf{f}(t, \mathbf{p}, \hat{\mathbf{z}})\|_\infty \leq \alpha(t)\|\tilde{\mathbf{z}} - \hat{\mathbf{z}}\|_\infty,$$

for every $\tilde{\mathbf{z}}, \hat{\mathbf{z}} \in B_\eta(\mathbf{z}) \cap D$.

Above, $B_\eta(\mathbf{z})$ denotes the open ball of radius $\eta$ around $\mathbf{z}$. In case $D$ is open, Assumptions 3.2.1 and 3.2.2 ensure that a unique solution of (3.2) exists locally [62]. In any case, it is always assumed that, for each $(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$, there exists a unique solution of (3.2) on all of $I$. We are interested in computing the following.

**Definition 3.2.3.** Two continuous functions $\mathbf{v}, \mathbf{w} : I \to \mathbb{R}^{n_x}$ are called *state bounds* for (3.2) if $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \in [\mathbf{v}(t), \mathbf{w}(t)]$, $\forall (t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$.

**Remark 3.2.4.** In many cases, we are interested in computing bounds on the solutions of an ODE subject to parametric uncertainty. This is simply a special case of the problem above, since a parameter vector $\mathbf{p} \in \mathbb{R}^{n_p}$ taking values in a compact set $P$ can simply be regarded as a vector of constant controls, $\mathbf{u}(t) = \mathbf{p}$ for all $t \in I$, taking values in $U \equiv P$. The only disadvantage of this approach is that the bounds will be valid for any solution that could result from a time-varying parameter vector taking values in $P$, while the solutions of interest are only those corresponding to constant parameter values. Thus, an additional source of conservatism is introduced. This observation does not seem to be generally appreciated, and this reformulation is routinely used in the standard differential inequalities method [135, 162]. It will be used here as well, since a better method is not available.

## 3.3 The Standard Differential Inequalities Method

In this section, a standard method for computing state bounds using the theory of differential inequalities [182, 170] is presented. The key result is Theorem 3.3.2 below, which gives a set of sufficient conditions under which two functions are guaranteed to bound the solutions of (3.2) pointwise in $t$. The statement here differs from statements in the literature in technical details. Its proof and a discussion of these differences can be found in §3.3.1 and §3.3.2, respectively.

**Definition 3.3.1.** Let $\mathcal{B}_i^L, \mathcal{B}_i^U : \mathbb{IR}^{n_x} \to \mathbb{IR}^{n_x}$ be defined by $\mathcal{B}_i^L([\mathbf{v}, \mathbf{w}]) = \{\mathbf{z} \in [\mathbf{v}, \mathbf{w}] : z_i = v_i\}$ and $\mathcal{B}_i^U([\mathbf{v}, \mathbf{w}]) = \{\mathbf{z} \in [\mathbf{v}, \mathbf{w}] : z_i = w_i\}$, for every $i = 1, \ldots, n_x$.

**Theorem 3.3.2.** *Let $\mathbf{v}, \mathbf{w} : I \to \mathbb{R}^{n_x}$ be absolutely continuous functions satisfying*

(EX): *For every $t \in I$ and each index $i$,*

1. $\mathbf{v}(t) \leq \mathbf{w}(t)$*,*
2. $\mathcal{B}_i^L([\mathbf{v}(t), \mathbf{w}(t)]) \subset D$*,* $\mathcal{B}_i^U([\mathbf{v}(t), \mathbf{w}(t)]) \subset D$*.*

(IC): $\mathbf{v}(t_0) \leq \mathbf{x}_0 \leq \mathbf{w}(t_0)$*,* $\forall \mathbf{x}_0 \in X_0$*.*

(RHS): *For a.e. $t \in I$ and each index $i$,*

1. $\dot{v}_i(t) \leq f_i(t, \mathbf{p}, \mathbf{z})$*, for all* $\mathbf{z} \in \mathcal{B}_i^L([\mathbf{v}(t), \mathbf{w}(t)])$ *and* $\mathbf{p} \in U$*,*
2. $\dot{w}_i(t) \geq f_i(t, \mathbf{p}, \mathbf{z})$*, for all* $\mathbf{z} \in \mathcal{B}_i^U([\mathbf{v}(t), \mathbf{w}(t)])$ *and* $\mathbf{p} \in U$*.*

*Then* $\mathbf{v}(t) \leq \mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \leq \mathbf{w}(t)$*,* $\forall (t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$*.*

Conceptually, the key hypotheses of Theorem 3.3.2 are (IC) and (RHS). Hypothesis (IC) simply requires that the bounding trajectories $\mathbf{v}$ and $\mathbf{w}$ are bounds at $t_0$. The conditions of (RHS) are the *differential inequalities*. The purpose of these conditions is to ensure that the solutions of (3.2) cannot cross $\mathbf{v}$ and $\mathbf{w}$ to the right of $t_0$. We will have much more to say about these conditions in Section 3.3.1. Theorems such as Theorem 3.3.2, which establish inequalities between the solution of a dynamic systems and other trajectories, are sometimes referred to as *comparison theorems*.

The hypotheses of Theorem 3.3.2 can be satisfied computationally using interval arithmetic. Suppose that $U$ and $X_0$ are $n_u$ and $n_x$-dimensional intervals, respectively, and that $\mathbf{f}$ is factorable. State bounds can then be computed by solving the ODEs

$$\begin{aligned}\dot{v}_i(t) &= [f_i]^L([t,t], U, \mathcal{B}_i^L([\mathbf{v}(t), \mathbf{w}(t)])), \\ \dot{w}_i(t) &= [f_i]^U([t,t], U, \mathcal{B}_i^U([\mathbf{v}(t), \mathbf{w}(t)])), \\ [v_i(t_0), w_i(t_0)] &= X_{0,i},\end{aligned} \tag{3.3}$$

for a.e. $t \in I$ and each index $i$. By construction, the intervals over which the interval extensions of each $f_i$ are taken in the right-hand sides of (3.3) are exactly the sets over which the differential inequalities in Hypothesis (RHS) are required to hold. Thus, any solutions $\mathbf{v}$ and $\mathbf{w}$ of (3.3) must satisfy (RHS). Furthermore, Hypothesis (IC) is satisfied by the choice of initial conditions in (3.3). Then, provided that (EX) holds, Theorem 3.3.2 ensures that $\mathbf{v}(t) \leq \mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \leq \mathbf{w}(t)$ for all $(t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$. The properties of bounding systems such as (3.3) are analyzed further in §3.5.2, and it is shown that a unique solution exists, at least locally about $t_0$, and indeed satisfies (EX). The implementation of Theorem 3.3.2 through (3.3) is due to Harrison [75], and will be referred to as Harrison's method throughout.

**Example 3.3.1.** Consider the reversible chemical reaction

$$\mathrm{A} + \mathrm{B} \rightleftharpoons \mathrm{C} \tag{3.4}$$

with forward and reverse rate constants $k_f$ and $k_r$, respectively, taking place in an isothermal batch reactor. The time evolution of the species concentrations $x_A$, $x_B$ and $x_C$ are described by a system of ODEs of the form (3.2), where $\mathbf{x} \equiv (x_A, x_B, x_C)$, $\mathbf{u} \equiv (k_f, k_r)$, and the right-hand side is defined by $\mathbf{f} = \mathbf{S}\mathbf{r}$, where

$$\mathbf{S} \equiv \begin{bmatrix} -1 & 1 \\ -1 & 1 \\ 1 & -1 \end{bmatrix}, \quad \mathbf{r}(t, \mathbf{p}, \mathbf{z}) \equiv \begin{bmatrix} p_f z_A z_B \\ p_r z_C \end{bmatrix}.$$

Consider computing state bounds on $I = [0, 0.05]$ min with the fixed initial condition $\mathbf{x}_0 = (1.5, 0.5, 0)$ (M) and the $(k_f, k_r)$ in the interval $U \equiv [100, 500] \times [0.001, 0.01]$ (M$^{-1}$min$^{-1}$, min$^{-1}$). That is, the set of admissible initial conditions is the singleton $X_0 = \{\mathbf{x}_0\}$ and the set of admissible controls is

$$\mathcal{U} = \{(k_f, k_r) \in (L^1(I))^2 : (k_f(t), k_r(t)) \in [100, 500] \times [0.001, 0.01] \text{ for a.e. } t \in I\}.$$

Here, the solutions of interest correspond to $k_f$ and $k_r$ that are constant in time, though the computed bounds will nonetheless be valid for time-varying rate constants, as discussed in Remark 3.2.4.

Consider the ODE describing $x_C$, which is given by

$$\dot{x}_C(t, \mathbf{u}, \mathbf{x}_0) = k_f(t)x_A(t, \mathbf{u}, \mathbf{x}_0)x_B(t, \mathbf{u}, \mathbf{x}_0) - k_r(t)x_C(t, \mathbf{u}, \mathbf{x}_0). \tag{3.5}$$

Denoting $U = [k_f^L, k_f^U] \times [k_r^L, k_r^U]$ and taking natural interval extensions of the right-hand side function, the bounding differential equations (3.3) for $x_C$ are given by

$$\begin{aligned} \dot{v}_C(t) &= k_f^L v_A(t)v_B(t) - k_r^U v_C(t), \\ \dot{w}_C(t) &= k_f^U w_A(t)w_B(t) - k_r^L w_C(t), \\ v_C(t_0) &= w_C(t_0) = x_{0,C}. \end{aligned} \tag{3.6}$$

The form of these equations result from the fact that all intervals are guaranteed to be positive, so that the choice of upper or lower bound for each variable is dictated simply by the sign of the term in which it appears. For example, the lower bound for the right-hand side for $\dot{x}_C$, $p_f z_A z_B - p_r z_C$, is computed by taking the lower bound for every variable in the first term and the upper bound for every variable in the second term. Note in particular that, in the second term, $v_C(t)$ is used in place of $w_C(t)$ since the natural interval extension in this case is taken over $\mathcal{B}_i^L([\mathbf{v}(t), \mathbf{w}(t)])$, not $[\mathbf{v}(t), \mathbf{w}(t)]$. The bounding differential equations for $x_A$ and $x_B$ are derived analogously. The solution of these bounding differential equations then gives the state bounds $[\mathbf{v}(t), \mathbf{w}(t)]$, for all $t \in I$. These equations were solved numerically using

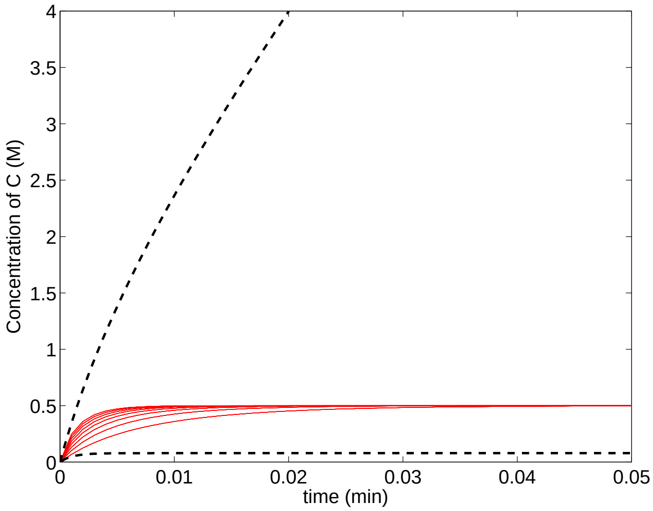

Figure 3-1: State bounds for the species concentration $x_C$ from Example 3.3.1. Solid curves are true model solutions computed for 64 constant vectors $\mathbf{u} = (k_f, k_r)$ on a uniform grid over $U = [100, 500] \times [0.001, 0.01]$ ($\mathrm{M}^{-1}\mathrm{min}^{-1}$, $\mathrm{min}^{-1}$). Dashed lines are state bounds computed by solving (3.3).

CVODE [44] with absolute and relative tolerances of $10^{-5}$. The state bounds on $x_C$ are shown by the dashed curves in Figure 3-1. The solid curves in Figure 3-1 are solutions $x_C(t, \mathbf{u}, \mathbf{x}_0)$ computed for 64 sampled $\mathbf{u} = (k_f, k_r)$ taking constant values on a uniform grid over $U = [100, 500] \times [0.001, 0.01]$ ($\mathrm{M}^{-1}\mathrm{min}^{-1}$, $\mathrm{min}^{-1}$).

It is clear from the figure that the computed trajectories do indeed bound the model solutions, but they are very conservative and do not represent the true set of solutions accurately. This issue is especially pronounced for the upper bound. On the other hand, the computation of these bounds required only $1.8 \times 10^{-4}$ CPU seconds, less than twice the cost of integrating a single model trajectory, $1.1 \times 10^{-4}$ s. These computations were done on a Dell Precision T3400 workstation with a 2.83 GHz Intel Core2 Quad CPU. One core and 512 MB of memory were dedicated to the job.

The previous example shows that Harrison's method can potentially produce very weak bounds. Unfortunately, these bounds are representative of the behavior of Harrison's method for many problems. On the other hand, the implementation of Theorem 3.3.2 using interval arithmetic and a state-of-the-art numerical integration routine is very inexpensive. Thus, the aim of this chapter and the next is to find ways to reduce the conservatism of Harrison's method, hopefully very significantly, without compromising its efficiency.

As discussed in the introduction, this will be done by incorporating physical information into the procedure at a very fundamental level. Examining the results of Example 3.3.1, it is easy to see that the bounds computed by Harrison's method disregard intuitive physical limitations. Given the reaction stoichiometry and the specified initial condition, simple conservation laws demand that $x_{\mathrm{C}}$ remains less than 0.5 M for all time, regardless of $\mathbf{u}$. Yet, the computed upper bound diverges toward $+\infty$. This suggests that even simple physical observations could be leveraged in order to compute much sharper bounds. This idea was first suggested in [162], where physical upper and lower bounds on each state variable, termed *natural bounds*, were used in a modified form of Harrison's method. In general, state bounds resulting from that method do not demonstrate catastrophic divergence, but still largely fail to provide an accurate enclosure of the reachable set throughout time.

In the next chapter, it will be shown that, for a very important class of ODE models in chemical engineering, including that of Example 3.3.1, the physical information used in [162], and in fact much more, is readily available and can often put massive restrictions on the regions of state space that must be considered during a state bounding computation. In the remainder of this chapter, we develop the theory required to use this information effectively, while still maintaining an efficient computational implementation. In comparison to [162], the methods developed here differ in that arbitrary physical information is considered instead of only natural bounds. This generalization is very challenging, both theoretically and from an implementation standpoint, but in the end results in vastly superior bounds for problems where rich physical information is available.

### 3.3.1 Proof of Theorem 3.3.2

#### Preliminaries

The proof of Theorem 3.3.2 involves some standard facts about absolutely continuous functions that can be found in [180]. Two important results are stated below. Denote the space of absolutely continuous functions from $[a, b]$ into $\mathbb{R}$ by $\mathcal{AC}([a, b], \mathbb{R})$. Recall that any $\phi \in \mathcal{AC}([a, b], \mathbb{R})$ is differentiable at almost every $t \in [a, b]$. The abbreviation "a.e. $t \in [a, b]$" is used throughout.

**Theorem 3.3.3.** *If $\phi \in \mathcal{AC}([a, b], \mathbb{R})$ satisfies $\dot{\phi}(t) \leq 0$ for a.e. $t \in [a, b]$, then $\phi$ is non-increasing on $[a, b]$.*

*Proof.* See Theorem 3.1 in [170]. □

**Lemma 3.3.4.** *For any $\epsilon > 0$ and any $\beta \in L^1([a, b])$, $\exists \rho \in \mathcal{AC}([a, b], \mathbb{R})$, non-decreasing, and satisfying*

$$0 < \rho(t) \leq \epsilon, \quad \forall t \in [a, b], \quad \textit{and} \quad \dot{\rho}(t) > |\beta(t)|\rho(t), \quad \textit{a.e. } t \in [a, b]. \tag{3.7}$$

*Proof.* Choose $\gamma > 0$, let $B(t) = \int_b^t (|\beta(s)| + \gamma)\, ds$ and let $\rho(t) = \epsilon e^{B(t)}$. Clearly, $\rho > 0$ and $\rho(b) = \epsilon$. $B$ is absolutely continuous and hence differentiable a.e. on $[a, b]$ with $\dot{B}(t) = |\beta(t)| + \gamma$. Since $B$ is absolutely continuous and $a \mapsto \epsilon e^a$ is locally Lipschitz, $\rho$ is absolutely continuous (See [119]) and, for a.e. $t \in [a, b]$, the chain rule gives

$$\dot{\rho}(t) = \epsilon e^{B(t)}(\dot{B}(t)) = \rho(t)\left(|\beta(t)| + \gamma\right) > |\beta(t)|\rho(t).$$

Theorem 3.3.3 shows that $\rho$ is non-decreasing. □

The proof of Theorem 3.3.2, and similar results in later sections, require a construction that is summarized in the following lemma and corollary.

**Lemma 3.3.5.** *Let $\boldsymbol{\delta} : I \to \mathbb{R}^n$ be a continuous function with $\boldsymbol{\delta}(t_0) \leq \mathbf{0}$. Suppose $\exists t \in I$ such that $\delta_i(t) > 0$ for at least one $i \in \{1, \ldots, n\}$, and define $t_1 \equiv \inf\{t \in I : \boldsymbol{\delta}(t) \not\leq \mathbf{0}\}$. Then*

*1. $t_0 \leq t_1 < t_f$ and $\boldsymbol{\delta}(t) \leq \mathbf{0}$, $\forall t \in [t_0, t_1]$.*

*2. The set $\mathcal{V} \equiv \{i : \forall \gamma > 0,\ \exists t \in (t_1, t_1 + \gamma]$ s.t. $\delta_i(t) > 0\}$ is nonempty.*

*Let $t_4 \in (t_1, t_f]$, $\epsilon > 0$ and $\beta \in L^1([t_1, t_4])$. Then there exists an index $j \in \{1, \ldots, n\}$, a non-decreasing function $\rho \in \mathcal{AC}([t_1, t_4], \mathbb{R})$ satisfying (3.7) on $[t_1, t_4]$, and numbers $t_2, t_3 \in [t_1, t_4]$ with $t_2 < t_3$ such that the following inequalities hold:*

$$\begin{aligned} \boldsymbol{\delta}(t) &< \mathbf{1}\rho(t), \quad \forall t \in [t_2, t_3), \\ 0 &< \delta_j(t), \quad \forall t \in (t_2, t_3), \\ \delta_j(t_3) &= \rho(t_3), \\ \delta_j(t_2) &= 0. \end{aligned} \tag{3.8}$$

*Proof.* By hypothesis, the set $\{t \in I : \boldsymbol{\delta}(t) \not\leq \mathbf{0}\}$ is nonempty. Since $t_1$ is a lower bound, $\boldsymbol{\delta}(t) \leq \mathbf{0}$ for all $t \in I$ such that $t < t_1$. If $t_1 > t_0$, then continuity ensures that this also holds at $t_1$, so that $\boldsymbol{\delta}(t) \leq \mathbf{0}$, $\forall t \in [t_0, t_1]$. If $t_1 = t_0$, then the same conclusion holds because $\boldsymbol{\delta}(t_0) \leq \mathbf{0}$. By the assumption that $\boldsymbol{\delta}(t) \not\leq \mathbf{0}$ for some $t \in I$, it follows that $t_1 < t_f$. Since $t_1$ is the greatest lower bound, it follows that the inequality $\boldsymbol{\delta}(t) \leq \mathbf{0}$ is violated arbitrarily close to the right of $t_1$. Then, since $\boldsymbol{\delta}$ is finite dimensional, there must be at least one $i$ such that $\delta_i(t) > 0$ arbitrarily close to the right of $t_1$. Thus, $\mathcal{V} \neq \emptyset$.

Choose any $t_4 \in (t_1, t_f]$, $\epsilon > 0$ and $\beta \in L^1([t_1, t_4])$. Choose $m$ so that $\exists t \in [t_1, t_4]$ with $\delta_i(t) \geq m > 0$, for some $i$. This must be possible since $\mathcal{V}$ is nonempty. By Lemma 3.3.4, there exists a non-decreasing function $\rho \in \mathcal{AC}([t_1, t_4], \mathbb{R})$ satisfying

$$0 < \rho(t) \leq \min(m/2, \epsilon), \quad \forall t \in [t_1, t_4], \quad \text{and} \quad \dot{\rho}(t) > |\beta(t)|\rho(t), \quad \text{a.e. } t \in [t_1, t_4].$$

Let $t_3 \equiv \inf\{t \in [t_1, t_4] : \delta_i(t) \geq \rho(t)$ for at least one $i\}$. Since $\rho < m$, this set is nonempty. Because $t_3$ is a lower bound, $\boldsymbol{\delta}(t) < \mathbf{1}\rho(t)$ for all $t \in [t_1, t_4]$ with $t < t_3$. Since $t_3$ is the greatest lower bound, $\delta_j(t_3) = \rho(t_3)$ for at least one $j$. Since $\boldsymbol{\delta}(t_1) \leq \mathbf{0}$, it follows that $t_3 \in (t_1, t_4]$.

Fix any $j$ such that $\delta_j(t_3) = \rho(t_3)$ and let $t_2 \equiv \sup\{t \in [t_1, t_3] : \delta_j(t) \leq 0\}$. Since $\delta_j(t_1) \leq 0$, this set is nonempty. Because $t_2$ is an upper bound, $\delta_j(t) > 0$ for all $t \in [t_1, t_3]$ with $t > t_2$. Because it is the least upper bound, $\delta_j(t_2) = 0$. It follows that $t_2 \in [t_1, t_3)$. □

**Corollary 3.3.6.** *Let $\boldsymbol{\phi}, \mathbf{v}, \mathbf{w} : I \to \mathbb{R}^{n_x}$ be continuous and satisfy $\mathbf{v}(t_0) \leq \boldsymbol{\phi}(t_0) \leq \mathbf{w}(t_0)$. Suppose $\exists t \in I$ such that either $\phi_i(t) < v_i(t)$ or $\phi_i > w_i(t)$, for at least one $i \in \{1, \ldots, n_x\}$, and define*

$$t_1 \equiv \inf\{t \in I : \phi_i(t) < v_i(t) \textit{ or } \phi_i > w_i(t), \textit{ for at lease one } i\}. \tag{3.9}$$

*Then*

1. *$t_0 \leq t_1 < t_f$ and $\mathbf{v}(t) \leq \boldsymbol{\phi}(t) \leq \mathbf{w}(t)$, $\forall t \in [t_0, t_1]$.*

2. *At least one of the sets*

$$\mathcal{V}^L \equiv \{i : \forall \gamma > 0,\ \exists t \in (t_1, t_1 + \gamma] \textit{ s.t. } \phi_i(t) < v_i(t)\},$$
$$\mathcal{V}^U \equiv \{i : \forall \gamma > 0,\ \exists t \in (t_1, t_1 + \gamma] \textit{ s.t. } \phi_i(t) > w_i(t)\},$$

   *is nonempty.*

*Let $t_4 \in (t_1, t_f]$, $\epsilon > 0$ and $\beta \in L^1([t_1, t_4])$. Then there exists an index $j \in \{1, \ldots, n_x\}$, a non-decreasing function $\rho \in \mathcal{AC}([t_1, t_4], \mathbb{R})$ satisfying* (3.7) *on $[t_1, t_4]$, and numbers $t_2, t_3 \in [t_1, t_4]$ with $t_2 < t_3$ such that*

$$\mathbf{v}(t) - \mathbf{1}\rho(t) < \boldsymbol{\phi}(t) < \mathbf{w}(t) + \mathbf{1}\rho(t), \quad \forall t \in [t_2, t_3) \tag{3.10}$$

*and*

$$\phi_j(t_2) = v_j(t_2), \quad \phi_j(t_3) = v_j(t_3) - \rho(t_3), \quad \textit{and} \quad \phi_j(t) < v_j(t), \tag{3.11}$$
$$\Big( \textit{ or } \phi_j(t_2) = w_j(t_2), \quad \phi_j(t_3) = w_j(t_3) + \rho(t_3), \quad \textit{and} \quad \phi_j(t) > w_j(t), \Big) \tag{3.12}$$

*for all $t \in (t_2, t_3)$.*

*Proof.* Define $\boldsymbol{\delta} : I \to \mathbb{R}^{2n_x}$ by $\boldsymbol{\delta}(t) \equiv (\mathbf{v}(t) - \boldsymbol{\phi}(t), \boldsymbol{\phi}(t) - \mathbf{w}(t))$, $\forall t \in I$. By hypothesis, $\boldsymbol{\delta}(t_0) \leq \mathbf{0}$, and $\exists t \in I$ such that $\delta_i(t) > 0$ for at least one $i$. The conclusion now follows from Lemma 3.3.5. □

We now proceed to the proof of Theorem 3.3.2.

### Proof

Choose any $(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$ and let $\mathbf{x}(t) \equiv \mathbf{x}(t, \mathbf{u}, \mathbf{x}_0)$ for convenience. Suppose that $\exists t \in I$ such that $\mathbf{x}(t) \notin [\mathbf{v}(t), \mathbf{w}(t)]$. We prove a contradiction.

Define $t_1$ as in (3.9) with $\boldsymbol{\phi} \equiv \mathbf{x}$, and define $\bar{\mathbf{x}}(t) \equiv \text{mid}(\mathbf{v}(t), \mathbf{w}(t), \mathbf{x}(t))$. Noting that the hypotheses of Corollary 3.3.6 are satisfied with $\boldsymbol{\phi} \equiv \mathbf{x}$, Conclusion 1 of that corollary implies that $\bar{\mathbf{x}}(t_1) = \mathbf{x}(t_1)$. Let $\eta > 0$ and $\alpha \in L^1(I)$ satisfy Assumption 3.2.2 with $\mathbf{z} \equiv \mathbf{x}(t_1)$. Choose $t_4 \in (t_1, t_f]$ small enough that $\mathbf{x}(t), \bar{\mathbf{x}}(t) \in B_\eta(\mathbf{x}(t_1))$, $\forall t \in [t_1, t_4]$.

Applying Corollary 3.3.6 with $t_4$, $\beta \equiv \alpha$ and arbitrary $\epsilon > 0$ yields an index $j \in \{1, \ldots, n_x\}$, a non-decreasing function $\rho \in \mathcal{AC}([t_1, t_4], \mathbb{R})$ satisfying (3.7) on $[t_1, t_4]$, and numbers $t_2, t_3 \in [t_1, t_4]$ with $t_2 < t_3$ such that (3.10) and (3.12) hold with $\boldsymbol{\phi} \equiv \mathbf{x}$ (the proof is analogous if instead (3.11) holds).

It will be shown that Hypothesis (RHS).2 can be applied at the point $(t, \mathbf{u}(t), \bar{\mathbf{x}}(t))$ for a.e. $t \in [t_2, t_3]$. By definition, it is clear that $\bar{\mathbf{x}}(t) \in [\mathbf{v}(t), \mathbf{w}(t)]$. Hypothesis (EX).1 and (3.12) show that $\bar{x}_j(t) = \text{mid}(v_j(t), w_j(t), x_j(t)) = w_j(t)$ and hence $\bar{\mathbf{x}}(t) \in \mathcal{B}_j^U([\mathbf{v}(t), \mathbf{w}(t)])$. By Hypothesis (EX).2, this implies that $\bar{\mathbf{x}}(t) \in D$.

Now, for a.e. $t \in [t_2, t_3]$, Hypothesis (RHS).2 gives

$$\dot{w}_j(t) \geq f_j(t, \mathbf{u}(t), \bar{\mathbf{x}}(t)) \geq f_j(t, \mathbf{u}(t), \mathbf{x}(t)) - \alpha(t)\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|_\infty.$$

By (3.10), $\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|_\infty \leq \rho(t)$, so that, for a.e. $t \in [t_2, t_3]$,

$$\begin{aligned}\dot{w}_j(t) + \dot{\rho}(t) &\geq f_j(t, \mathbf{u}(t), \mathbf{x}(t)) - \alpha(t)\rho(t) + \dot{\rho}(t),\\ &> f_j(t, \mathbf{u}(t), \mathbf{x}(t)),\\ &= \dot{x}_j(t).\end{aligned}$$

The second inequality above follows from (3.7). By Theorem 3.3.3, this implies that $(x_j - w_j - \rho)$ is non-increasing on $[t_2, t_3]$, so that

$$x_j(t_3) - w_j(t_3) - \rho(t_3) \leq x_j(t_2) - w_j(t_2) - \rho(t_2).$$

But, by (3.12), this implies that $0 \leq -\rho(t_2)$, which contradicts (3.7).

### 3.3.2 Comments on Similar Results in the Literature

Similar results for ODEs without controls [182] originate from the existence theorem of Müller [126]. The extension to ODEs with real parameter dependence is apparent and is discussed in [162]. The extension to ODEs with time-varying inputs has been stated by several authors [75, 93] and is indeed apparent from Müller's result in the case of continuous inputs. The present result holds also for $L^1$ controls. Its proof requires a different approach and was influenced by Theorem 3.1 in [170], which applies to quasi-monotone systems under Carathéodory hypotheses. This approach is required in order to treat weak solutions of (3.2); i.e., solutions which only satisfy (3.2) for a.e. $t \in I$.

Compared to the statements in [93, 140], note that we require absolute continuity of the bounds instead of continuity, and require that the differential inequalities hold almost everywhere with true derivatives, as opposed to everywhere with Dini derivatives. This is again related to the fact that the present result holds for $L^1$ controls, and hence weak solutions of (3.2). We also note that Hypothesis (EX), which is inherent in Müller's formulation, is notably omitted from the statement in [75]. This error originates from Remark 12.X($\beta$) in [182] (stated with incomplete proof) and is common in the literature. It is easy to see that (EX).2 is necessary in Theorem 3.3.2 because the hypothesis (RHS) is not well-posed without it. Moreover, (EX).1 is easily motivated by Example 3.3.2 below and is unrelated to the presence of controls in Theorem 3.3.2. Finally, we note that Theorem 3.3.2 is a special case of a general characterization of invariant tubes for differential inclusions given in [13]. However, the presented form is amenable to efficient interval computation whereas the general form is not.

**Example 3.3.2.** Let $I = [0, 1]$ and $D = \mathbb{R}^2$. Consider the 2-dimensional ODE with no controls defined by $f_1(t, \mathbf{z}) = z_1 - z_2$, $f_2(t, \mathbf{z}) = z_2 - z_1$, and $\mathbf{x}_0 = [1\ 1]^{\mathrm{T}}$. $\mathbf{f}$ clearly satisfies Assumptions 3.2.1 and 3.2.2. Furthermore, with these definitions, it is clear that $\mathbf{x}(t) = \mathbf{x}_0 = [1\ 1]^{\mathrm{T}}$ is the unique solution of (3.2) on $I$. Now consider the functions $\mathbf{v}$ and $\mathbf{w}$ given by $v_1(t) = v_2(t) = t^2 + 1$ and $w_1(t) = w_2(t) = -t^2 + 1$. By straightforward computation, $\dot{v}_1(t) = \dot{v}_2(t) = 2t$ and $\dot{w}_1(t) = \dot{w}_2(t) = -2t$.

Omitting (EX).1, the remaining hypothesis of Theorem 3.3.2 are verified as follows. Hypothesis (EX).2 is trivial by the choice of $D$. (IC) is true because $\mathbf{v}(0) = \mathbf{x}_0 = \mathbf{w}(0) = [1\ 1]^{\mathrm{T}}$. (RHS).1 states that, for $i \in \{1, 2\}$ and a.e. $t \in [0, 1]$, $v_i(t)$ must satisfy the stated inequality *if* $\mathbf{v}(t) \leq \mathbf{z} \leq \mathbf{w}(t)$ and $z_i = v_i(t)$. But, for any $t \in (0, 1]$, $\mathbf{w}(t) < \mathbf{v}(t)$, so there does not exist any $\mathbf{z}$ satisfying these conditions. Therefore, (RHS).1 is trivially satisfied. By an analogous argument, (RHS).2 is also satisfied.

On the other hand, it is clear that $\mathbf{v}$ and $\mathbf{w}$ do not satisfy the conclusion of Theorem 3.3.2 because $\mathbf{x}(t) = \mathbf{x}_0$ for all $t \in [0, 1]$ and $t^2 + 1 > 1 > -t^2 + 1$ for all $t \in (0, 1]$, which implies that $\mathbf{v}(t) > \mathbf{x}(t) > \mathbf{w}(t)$ on $(0, 1]$.

## 3.4 The Use of a Priori Enclosures in Comparison Theorems

This section provides a conceptual discussion of the use of physical information in the context of differential inequalities bounding methods. Throughout, we assume that some set $G \subset \mathbb{R}^{n_x}$ is available such that $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \in G$, $\forall (t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$. The set $G$ is called an *a priori* enclosure, since we assume that it is known *prior* to the application of a state bounding method. It will be shown that the most natural use of $G$ in the context of differential inequalities is not valid in general, and that valid uses can be non-intuitive and depend on the specific form of $G$ (interval, polyhedral, etc.).

Recall the central hypothesis of Theorem 3.3.2:

(RHS): For a.e. $t \in I$ and each index $i$,

1. $\dot{v}_i(t) \leq f_i(t, \mathbf{p}, \mathbf{z})$ for all $\mathbf{z} \in \mathcal{B}_i^L([\mathbf{v}(t), \mathbf{w}(t)])$ and $\mathbf{p} \in U$,
2. $\dot{w}_i(t) \geq f_i(t, \mathbf{p}, \mathbf{z})$ for all $\mathbf{z} \in \mathcal{B}_i^U([\mathbf{v}(t), \mathbf{w}(t)])$ and $\mathbf{p} \in U$.

Conceptually, (RHS) relates $\dot{v}_i(t)$ and $\dot{w}_i(t)$ to possible values of the derivatives of solutions of (3.2) at $t$, through the values of $f_i(t, \cdot, \cdot)$. However, it is clear that the only values of $f_i(t, \cdot, \cdot)$ which are related to the derivatives of solutions of (3.2) are those which $f_i(t, \cdot, \cdot)$ takes at the points $(\mathbf{u}(t), \mathbf{x}(t, \mathbf{u}, \mathbf{x}_0))$ with $(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$. Then, considering that $G$ satisfies, by definition, $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \in G$, $\forall (t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$, it seems reasonable to expect that the sets over which the differential inequalities in (RHS) are required to hold could be restricted in some way by $G$. Of course, the most natural restriction is obtained by simply taking the intersection with $G$ to arrive at:

(RHSa): For a.e. $t \in I$ and each index $i$,

1. $\dot{v}_i(t) \leq f_i(t, \mathbf{p}, \mathbf{z})$ for all $\mathbf{z} \in \mathcal{B}_i^L([\mathbf{v}(t), \mathbf{w}(t)]) \cap G$ and $\mathbf{p} \in U$,
2. $\dot{w}_i(t) \geq f_i(t, \mathbf{p}, \mathbf{z})$ for all $\mathbf{z} \in \mathcal{B}_i^U([\mathbf{v}(t), \mathbf{w}(t)]) \cap G$ and $\mathbf{p} \in U$.

It should be clear that (RHSa) is a weaker hypothesis than (RHS), and thus potentially enables one to characterize sharper bounds through Theorem 3.3.2.

Surprisingly, Theorem 3.3.2 is not generally valid with (RHSa) in place of (RHS). This claim is contrary to Remark 2.4 in [162] and is proven by the counterexamples below. These examples show two fundamentally different complications inherent in (RHSa), which are subsequently discussed.

**Example 3.4.1** (Encountering the Empty Set in (RHSa))**.** Let $I = [0, 1]$, $D = \mathbb{R}$ and consider the scalar ODE with no controls defined by $f(t, z) = z$ and $x_0 = 0$. Clearly, the unique solution of (3.2) with these definitions is given by $x(t) = 0$, $\forall t \in I$. Assumptions 3.2.1 and 3.2.2 are obviously satisfied.

Choose $G = [0, 0]$, and let $v(t) = w(t) = t^2$ for all $t \in I$. Clearly this satisfies Hypothesis (EX) of Theorem 3.3.2. Furthermore, (IC) clearly holds, and (RHSa) is trivially satisfied because, for any $t \in (0, 1]$, the set $[\mathbf{v}(t), \mathbf{w}(t)] \cap G = [t^2, t^2] \cap [0, 0] = \emptyset$. Therefore, all of the hypotheses of Theorem 3.3.2 are satisfied, with (RHSa) in place of (RHS), and the conclusion of that theorem is clearly false because $x(t) = 0 < t^2 = v(t)$ on $(0, 1]$.

**Example 3.4.2** (A Regularity Problem on the Boundary of $G$ in (RHSa))**.** Let $I = [0, 0.5]$, $D = (-0.51, 0.51) \times (-2.1, 2.1)$ and consider the 2-dimensional ODE with no controls defined by $f_1(t, \mathbf{z}) = -1$ and $f_2(t, \mathbf{z}) = z_1/\sqrt{1 - z_1^2}$. Assumption 3.2.1 is easily verified. Further, it can be shown that each $f_i$ is Lipschitz on $I \times D$ by simply checking that the partial derivatives with respect to $\mathbf{z}$ are bounded on $D$ (though not on $\mathbb{R}^2$), and Assumption 3.2.2 follows. Letting $\mathbf{x}_0 = [0 \ 1]^{\mathrm{T}}$, it is easily verified that the unique solution of (3.2) is given by $x_1(t) = -t$ and $x_2(t) = \sqrt{1 - t^2}$. Let $G = \{\mathbf{z} : z_1^2 + z_2^2 \leq 1\}$. Note that $\mathbf{x}(t) \in G$ for all $t \in I$.

Now consider the functions $\mathbf{v}, \mathbf{w} : I \to \mathbb{R}^{n_x}$ defined by $v_1(t) = -t$, $w_1(t) = t$, $v_2(t) = 1$ and $w_2(t) = 2$. Hypotheses (EX) and (IC) of Theorem 3.3.2 are easily verified. Moreover, for any $t \in (0, 0.5]$, the set $\mathcal{B}_i^U([\mathbf{v}(t), \mathbf{w}(t)]) \cap G$ is empty for every $i$ and the set $\mathcal{B}_i^L([\mathbf{v}(t), \mathbf{w}(t)]) \cap G$ is empty for $i = 1$ and contains the single point $\mathbf{z} = [0 \ 1]^{\mathrm{T}}$ for $i = 2$. Thus, (RHSa).2 is trivially satisfied, and so is (RHSa).1 when $i = 1$. (RHSa).1 is satisfied for $i = 2$ because $\dot{v}_2(t) = 0 = f_2(t, [0 \ 1]^{\mathrm{T}})$, $\forall t \in [0, 0.5]$. Of course, the conclusion of Theorem 3.3.2 does not hold since $x_2(0.5) = \sqrt{0.75} \notin [1, 2] = [v_2(0.5), w_2(0.5)]$.

Despite these pessimistic results, it will be shown in the following sections that hypotheses very similar to (RHSa) can in fact be used to derive strengthened comparison theorems and very effective bounding methods. To do so, however, it is necessary to dispense with the flawed conceptual idea leading to (RHSa), and come to grips with the mathematical requirements that a (RHS)-type hypothesis must satisfy. Conceptually, it is tempting to interpret the Hypothesis (RHS) in the following way: if at any $t \in I$ and for any $(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$, it happens that some solution $x_i \equiv x_i(\cdot, \mathbf{u}, \mathbf{x}_0)$ runs into a bound, say $v_i(t)$, for the first time, then $(t, \mathbf{u}(t), \mathbf{x}(t))$ is feasible in (RHS).1. Therefore, $\dot{v}_i(t) \leq \dot{x}_i(t)$. Moreover, since $\mathbf{x}(t) \in G$, the same argument shows that $\dot{v}_i(t) \leq \dot{x}_i(t)$ if we have (RHSa) instead of (RHS). So far, this argument is correct. What is false is the idea that this differential inequality implies that $v_i \leq x_i$ to the right of $t$. This implication fails for $v$ and $x$ at $t_0$ in Example 3.4.1.

Examining the proof of Theorem 3.3.2, the hypothesis (RHS) is used in quite a different way than the intuitive explanation above would suggest. In fact, the entire proof occurs in the hypothetical situation where $\mathbf{x}(t)$ is not in $[\mathbf{v}(t), \mathbf{w}(t)]$. The hypothesis (RHS) is never applied to the point $(t, \mathbf{u}(t), \mathbf{x}(t))$, because this point is not in the required set by construction. Instead (RHS) is applied to a nearby point, $(t, \mathbf{u}(t), \bar{\mathbf{x}}(t))$, that does satisfy the required conditions. Specifically, $(t, \mathbf{u}(t), \bar{\mathbf{x}}(t))$ is *nearby* in the sense that $\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|_\infty \leq \rho(t)$ for a.e. $t \in [t_2, t_3]$, and the usefulness of applying (RHS) at this point to get information about $\dot{x}_i(t)$ critically depends on the Lipschitz condition on $f_i$, as per Assumption 3.2.2.

Lets now consider how (RHSa) fails in the preceding examples, and how this relates to the proof of Theorem 3.3.2. One fundamental difference between (RHS) and (RHSa) is that, in the latter, it is possible for the set over which the differential inequalities are required to hold to be empty. This is exactly the circumstance leading to the counterexample Example 3.4.1, and it is fairly easy to see how this situation interrupts the proof of Theorem 3.3.2. Specifically, there is no point $\bar{\mathbf{x}}(t)$, nearby or otherwise, at which (RHSa) can be applied. In Example 3.4.2, empty sets also occur, but these are not the critical problem. (RHSa) does indeed impose a nontrivial condition on $\dot{v}_2(t)$, for all $t \in I$. However, the only point $\mathbf{z}$ for which we must have $\dot{v}_2(t) \leq f_2(t, \mathbf{z})$, according to (RHSa), is not nearby $\mathbf{x}(t)$ in the sense above. In essence, the shape of the set $G$ introduces non-Lipschitz behavior, despite the fact that $f_2$ is Lipschitz.

A final rather serious problem with (RHSa) is that the efficient implementation of the standard differential inequalities method is no longer sensible. The set $\mathcal{B}_i^L([\mathbf{v}(t), \mathbf{w}(t)]) \cap G$ is not necessarily an interval if $G$ is not an interval, so the hypotheses of (RHSa) cannot be satisfied efficiently through interval arithmetic. In fact, it turns out that the case where $G$ is not an interval is of significant interest in applications.

Roughly speaking, the solution to all of the problems discussed above is to replace the intersections with $G$ in (RHSa) with some weaker operations. Conceptually, these operations overestimate the set $\mathcal{B}_i^{L/U}([\mathbf{v}(t), \mathbf{w}(t)]) \cap G$ at each point in time. Moreover, they return nonempty sets and obey a certain Lipschitz condition. Finally, these operations can be chosen in order to return intervals or other types of sets that permit an efficient computational implementation.

In general, what constitutes a valid weaker form of $\mathcal{B}_i^{L/U}([\mathbf{v}(t), \mathbf{w}(t)]) \cap G$ will depend on the particular form of $G$ (interval, polyhedral, etc.). Moreover, this choice is not unique. Finally, in many cases the difference between $\mathcal{B}_i^{L/U}([\mathbf{v}(t), \mathbf{w}(t)]) \cap G$ and this weaker form are subtle. All of this then begs the question, what are the general principles that distinguish a valid usage of $G$ in a comparison theorem from the invalid use of (RHSa)?

To answer this question, these weaker operations are formalized in a general setting in the next section. Strictly, the requirement that these operations never return the empty set is not absolutely necessary. Nonetheless, it will be inherent in the development of the following section. A yet more general presentation permitting empty sets is given in §3.7, though the resulting methods are more difficult to implement and therefore not as useful in general.

## 3.5 A General Comparison Theorem

In this section, a comparison theorem is proven in a very general setting. In light of the complications discussed in the previous section, the purpose of this abstract analysis is to understand the fundamental requirements that one must impose on (RHS)-type hypotheses in order to arrive at a valid comparison theorem. The approach, then, is essentially to assume precisely what is required by the method of proof used in §3.3.1, and work backwards toward implementable methods. First, the problem of bounding an arbitrary function $\boldsymbol{\phi} \in \mathcal{AC}(I, \mathbb{R}^n)$ by two functions $\mathbf{v}, \mathbf{w} \in \mathcal{AC}(I, \mathbb{R}^n)$ is considered. State bounds for the ODEs (3.2) are considered explicitly in §3.5.1.

Let $D_\Pi \subset I \times \mathbb{R}^n \times \mathbb{R}^n$ and, for every $i \in \{1, \ldots, n\}$, let $\Pi_i^L, \Pi_i^U : D_\Pi \to \mathcal{P}(\mathbb{R})$. That is, for every $(t, \mathbf{v}, \mathbf{w}) \in D_\Pi$, $\Pi_i^L(t, \mathbf{v}, \mathbf{w})$ and $\Pi_i^U(t, \mathbf{v}, \mathbf{w})$ are subsets of $\mathbb{R}$. The following hypothesis provides a very minimal set of requirements relating the mappings $\Pi_i^{L/U}$ to the function $\boldsymbol{\phi}$ in such a way that Theorem 3.5.1 below holds.

**Hypothesis 3.5.1.** Suppose that $(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}}) \in I \times \mathbb{R}^n \times \mathbb{R}^n$ satisfies $\hat{\mathbf{v}} \leq \boldsymbol{\phi}(\hat{t}) \leq \hat{\mathbf{w}}$ and either $\phi_i(\hat{t}) = \hat{v}_i$ or $\phi_i(\hat{t}) = \hat{w}_i$ for at least one $i \in \{1, \ldots, n\}$. Then there exists $\eta > 0$ and $\alpha \in L^1(I)$ such that the following conditions hold for every $(\mathbf{v}, \mathbf{w}) \in B_\eta((\hat{\mathbf{v}}, \hat{\mathbf{w}}))$ and a.e. $t \in [\hat{t}, \hat{t} + \eta)$ such that $(t, \mathbf{v}, \mathbf{w}) \in D_\Pi$:

1. If $\phi_i(t) < v_i$, then $\exists \sigma \in \Pi_i^L(t, \mathbf{v}, \mathbf{w})$ such that

$$|\sigma - \dot{\phi}_i(t)| \leq \alpha(t) \max\left(\|\max(\mathbf{0}, \mathbf{v} - \boldsymbol{\phi}(t))\|_\infty, \|\max(\mathbf{0}, \boldsymbol{\phi}(t) - \mathbf{w})\|_\infty\right). \quad (3.13)$$

2. If $\phi_i(t) > w_i$, then $\exists \sigma \in \Pi_i^U(t, \mathbf{v}, \mathbf{w})$ such that (3.13) holds.

**Theorem 3.5.1.** *Let $\boldsymbol{\phi}, \mathbf{v}, \mathbf{w} \in \mathcal{AC}(I, \mathbb{R}^n)$ satisfy*

(EX): $(t, \mathbf{v}(t), \mathbf{w}(t)) \in D_\Pi$, $\forall t \in I$.

(IC): $\mathbf{v}(t_0) \leq \boldsymbol{\phi}(t_0) \leq \mathbf{w}(t_0)$.

(RHS): *For a.e. $t \in I$ and each index $i$,*

1. *$\dot{v}_i(t) \leq \sigma$ for all $\sigma \in \Pi_i^L(t, \mathbf{v}(t), \mathbf{w}(t))$,*
2. *$\dot{w}_i(t) \geq \sigma$ for all $\sigma \in \Pi_i^U(t, \mathbf{v}(t), \mathbf{w}(t))$.*

*If Hypothesis 3.5.1 holds, then $\mathbf{v}(t) \leq \boldsymbol{\phi}(t) \leq \mathbf{w}(t)$, $\forall t \in I$.*

*Proof.* Suppose that $\exists t \in I$ such that $\phi_i(t) < v_i(t)$ or $\phi_i(t) > w_i(t)$, for at least one $i \in \{1, \ldots, n\}$. It will be shown that this results in a contradiction.

Note that the hypotheses of Corollary 3.3.6 are satisfied and define $t_1$ as in (3.9). By Conclusion 1 of that corollary, $\mathbf{v}(t_1) \leq \boldsymbol{\phi}(t_1) \leq \mathbf{w}(t_1)$. By continuity and Conclusion 2, there must exist at least one $i$ such that either $\phi_i(t_1) = v_i(t_1)$ or $\phi_i(t_1) = w_i(t_1)$. Let $\eta > 0$ and $\alpha \in L^1(I)$ satisfy Hypothesis 3.5.1 with $(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}}) \equiv (t_1, \mathbf{v}(t_1), \mathbf{w}(t_1))$. Choose $t_4 \in (t_1, t_f]$ small enough that

$$t \in [t_1, t_1 + \eta) \quad \text{and} \quad (\mathbf{v}(t), \mathbf{w}(t)) \in B_\eta((\mathbf{v}(t_1), \mathbf{w}(t_1))), \quad \forall t \in [t_1, t_4]. \tag{3.14}$$

Noting that $(t, \mathbf{v}(t), \mathbf{w}(t)) \in D_\Pi$ for all $t \in [t_1, t_4]$ by Hypothesis (EX), we are now guaranteed the conditions of Hypothesis 3.5.1 with $(t, \mathbf{v}, \mathbf{w}) \equiv (t, \mathbf{v}(t), \mathbf{w}(t))$, for a.e. $t \in [t_1, t_4]$.

We now apply Corollary 3.3.6 with $t_4$, arbitrary $\epsilon > 0$ and $\beta = \alpha$. This furnishes an index $j \in \{1, \ldots, n\}$, a non-decreasing function $\rho \in \mathcal{AC}([t_1, t_4], \mathbb{R})$ satisfying (3.7) on $[t_1, t_4]$, and numbers $t_2, t_3 \in [t_1, t_4]$ with $t_2 < t_3$ such that (3.10)-(3.11) hold (the proof is analogous if (3.12) holds instead).

For a.e. $t \in [t_2, t_3]$, (3.11) states that $\phi_j(t) < v_j(t)$. Then, combining Condition 1 of Hypothesis 3.5.1 and Hypotheses (RHS).1 shows that

$$\dot{v}_j(t) - \dot{\phi}_j(t) \leq \alpha(t) \max\left(\| \max(\mathbf{0}, \mathbf{v}(t) - \boldsymbol{\phi}(t))\|_\infty, \| \max(\mathbf{0}, \boldsymbol{\phi}(t) - \mathbf{w}(t))\|_\infty\right), \tag{3.15}$$

for a.e. $t \in [t_2, t_3]$. But by (3.10),

$$\dot{v}_j(t) - \dot{\phi}_j(t) < \alpha(t)\rho(t), \quad \text{a.e. } t \in [t_2, t_3]. \tag{3.16}$$

Finally, using (3.7) and recalling that we have used $\beta = \alpha$, this implies that

$$\dot{v}_j(t) - \dot{\phi}_j(t) - \dot{\rho}(t) < \alpha(t)\rho(t) - \dot{\rho}(t) < 0, \quad \text{a.e. } t \in [t_2, t_3]. \tag{3.17}$$

By Theorem 3.3.3, this implies that $(v_j - \phi_j - \rho)$ is non-increasing on $[t_2, t_3]$, so that $v_j(t_3) - \phi_j(t_3) - \rho(t_3) \leq v_j(t_2) - \phi_j(t_2) - \rho(t_2)$. But by (3.11), this implies that $0 \leq -\rho(t_2)$, which contradicts (3.7). □

### 3.5.1 Specialization to State Bounds for ODEs

Let $D_\Omega \subset I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ and, for every $i \in \{1, \ldots, n_x\}$, let $\Omega_i^L, \Omega_i^U : D_\Omega \to \mathcal{P}(\mathbb{R}^{n_x})$. To specialize Theorem 3.5.1 to the task of characterizing state bounds for (3.2), let $\boldsymbol{\phi} \equiv \mathbf{x}(\cdot, \mathbf{u}, \mathbf{x}_0)$, for some $(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$, and let $\Pi_i^L$ and $\Pi_i^U$ take the form

$$\Pi_i^L(t, \mathbf{v}, \mathbf{w}) \equiv \{f_i(t, \mathbf{p}, \mathbf{z}) : \mathbf{p} \in U,\ \mathbf{z} \in \Omega_i^L(t, \mathbf{v}, \mathbf{w})), \tag{3.18}$$

$$\Pi_i^U(t, \mathbf{v}, \mathbf{w}) \equiv \{f_i(t, \mathbf{p}, \mathbf{z}) : \mathbf{p} \in U,\ \mathbf{z} \in \Omega_i^U(t, \mathbf{v}, \mathbf{w})), \tag{3.19}$$

for all $(t, \mathbf{v}, \mathbf{w})$ in the set

$$D_\Pi \equiv \{(t, \mathbf{v}, \mathbf{w}) \in D_\Omega : \Omega_i^{L/U}(t, \mathbf{v}, \mathbf{w}) \subset D,\ i = 1, \ldots, n_x\}. \tag{3.20}$$

It will be shown that Hypothesis 3.5.1 is ensured by imposing the following conditions on $\Omega_i^L$ and $\Omega_i^U$:

**Hypothesis 3.5.2.** Let $(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}}) \in I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ and suppose that $\exists (\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$ such that $\mathbf{x} \equiv \mathbf{x}(\cdot, \mathbf{u}, \mathbf{x}_0)$ satisfies $\hat{\mathbf{v}} \leq \mathbf{x}(\hat{t}) \leq \hat{\mathbf{w}}$ and either $x_i(\hat{t}) = \hat{v}_i$ or $x_i(\hat{t}) = \hat{w}_i$ for at least one $i \in \{1, \ldots, n_x\}$. Then there exist $\eta, L > 0$ such that the following conditions hold for every $(\mathbf{v}, \mathbf{w}) \in B_\eta((\hat{\mathbf{v}}, \hat{\mathbf{w}}))$ and a.e. $t \in [\hat{t}, \hat{t}+\eta)$ such that $(t, \mathbf{v}, \mathbf{w}) \in D_\Omega$:

1. If $x_i(t) < v_i$, then $\exists \mathbf{z} \in \Omega_i^L(t, \mathbf{v}, \mathbf{w})$ such that

$$\|\mathbf{x}(t) - \mathbf{z}\|_\infty \leq L \max\left(\|\max(\mathbf{0}, \mathbf{v} - \mathbf{x}(t))\|_\infty, \|\max(\mathbf{0}, \mathbf{x}(t) - \mathbf{w})\|_\infty\right). \tag{3.21}$$

2. If $x_i(t) > w_i$, then $\exists \mathbf{z} \in \Omega_i^U(t, \mathbf{v}, \mathbf{w})$ such that (3.21) holds.

**Lemma 3.5.2.** *Suppose that Hypothesis 3.5.2 holds. Then, for any* $(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$, *Hypothesis 3.5.1 holds with* $\boldsymbol{\phi} \equiv \mathbf{x}(\cdot, \mathbf{u}, \mathbf{x}_0)$ *and the definitions* (3.18)*,* (3.19) *and* (3.20)*.*

*Proof.* Choose any $(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$ and define $\boldsymbol{\phi} \equiv \mathbf{x}(\cdot, \mathbf{u}, \mathbf{x}_0)$. Let $(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}}) \in I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ and suppose that $\hat{\mathbf{v}} \leq \boldsymbol{\phi}(\hat{t}) \leq \hat{\mathbf{w}}$ and either $\phi_i(\hat{t}) = \hat{v}_i$ or $\phi_i(\hat{t}) = \hat{w}_i$ for at least one $i \in \{1, \ldots, n_x\}$. Noting that $(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}})$ satisfies the required properties, let $L_\Omega, \eta_\Omega > 0$ be constants satisfying Hypothesis 3.5.2.

Let $\eta_f > 0$ and $\alpha_f \in L^1(I)$ be given by Assumption 3.2.2 with $\mathbf{z} = \boldsymbol{\phi}(\hat{t})$. Define $\alpha \equiv L_\Omega \alpha_f$ and choose $\eta \in (0, \min(\eta_f, \eta_\Omega)]$ small enough that

$$\|\boldsymbol{\phi}(t) - \boldsymbol{\phi}(\hat{t})\|_\infty < \eta_f/2, \tag{3.22}$$

$$L_\Omega \max\left(\|\max(\mathbf{0}, \mathbf{v} - \boldsymbol{\phi}(t))\|_\infty, \|\max(\mathbf{0}, \boldsymbol{\phi}(t) - \mathbf{w})\|_\infty\right) < \eta_f/2, \tag{3.23}$$

for all $t \in [\hat{t}, \hat{t} + \eta)$ and every $(\mathbf{v}, \mathbf{w}) \in B_\eta((\hat{\mathbf{v}}, \hat{\mathbf{w}}))$. It will be shown that Hypothesis 3.5.1 holds with these definitions.

Choose any $(\mathbf{v}, \mathbf{w}) \in B_\eta((\hat{\mathbf{v}}, \hat{\mathbf{w}}))$. For a.e. $t \in [\hat{t}, \hat{t} + \eta)$ such that $(t, \mathbf{v}, \mathbf{w}) \in D_\Pi$, the conditions of Hypothesis 3.5.2 hold because $\eta \leq \eta_\Omega$ and $D_\Pi \subset D_\Omega$. Suppose that $\phi_i(t) < v_i$. By Condition 1 of Hypothesis 3.5.2, $\exists \mathbf{z} \in \Omega_i^L(t, \mathbf{v}, \mathbf{w}) \subset D$ such that (3.21) holds with $L = L_\Omega$ and $\mathbf{x} = \boldsymbol{\phi}$. Combining this with (3.23) implies that $\|\boldsymbol{\phi}(t) - \mathbf{z}\|_\infty < \eta_f/2$. By (3.22) and the triangle inequality, it follows that $\mathbf{z} \in B_{\eta_f}(\boldsymbol{\phi}(\hat{t}))$. This implies that the inequality of Assumption 3.2.2 can be applied to the points $\mathbf{z}$ and $\boldsymbol{\phi}(t)$.

Let $\sigma \equiv f_i(t, \mathbf{u}(t), \mathbf{z})$. By definition, $\sigma \in \Pi_i^L(t, \mathbf{v}, \mathbf{w})$. Moreover,

$$|\sigma - \dot{\phi}_i(t)| = |f_i(t, \mathbf{u}(t), \mathbf{z}) - f_i(t, \mathbf{u}(t), \boldsymbol{\phi}(t))|, \tag{3.24}$$

$$\leq \alpha_f(t)\|\boldsymbol{\phi}(t) - \mathbf{z}\|_\infty, \tag{3.25}$$

$$\leq \alpha(t) \max\left(\|\max(\mathbf{0}, \mathbf{v} - \boldsymbol{\phi}(t))\|_\infty, \|\max(\mathbf{0}, \boldsymbol{\phi}(t) - \mathbf{w})\|_\infty\right). \tag{3.26}$$

This proves Condition 1 of Hypothesis 3.5.1, and Condition 2 follows by an analogous argument. □

It is important to note that Hypothesis 3.5.2 only implies Hypothesis 3.5.1 when

$\mathbf{f}$ satisfies the Lipschitz condition of Assumption 3.2.2. The following Hypothesis is an alternative to Hypothesis 3.5.2 that is sometimes easier to confirm.

**Hypothesis 3.5.3.** The following conditions hold for all $i \in \{1, \ldots, n_x\}$:

1. Let $(t, \mathbf{v}, \mathbf{w}) \in I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$. If $\exists(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$ satisfying $\mathbf{v} \leq \mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \leq \mathbf{w}$ and $x_i(t, \mathbf{u}, \mathbf{x}_0) = v_i$, then $(t, \mathbf{v}, \mathbf{w}) \in D_\Omega$ and $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \in \Omega_i^L(t, \mathbf{v}, \mathbf{w})$. If $\exists(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$ satisfying $\mathbf{v} \leq \mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \leq \mathbf{w}$ and $x_i(t, \mathbf{u}, \mathbf{x}_0) = w_i$, then $(t, \mathbf{v}, \mathbf{w}) \in D_\Omega$ and $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \in \Omega_i^U(t, \mathbf{v}, \mathbf{w})$.

2. Let $(t, \mathbf{v}, \mathbf{w}) \in D_\Omega$. $\Omega_i^L(t, \mathbf{v}, \mathbf{w})$ and $\Omega_i^U(t, \mathbf{v}, \mathbf{w})$ are nonempty and compact.

3. Let $(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}}) \in D_\Omega$. There exists $\eta, L > 0$ such that

$$d_H(\Omega_i^L(t, \mathbf{v}_1, \mathbf{w}_1), \Omega_i^L(t, \mathbf{v}_2, \mathbf{w}_2)) \leq L \max\left(\|\mathbf{v}_1 - \mathbf{v}_2\|_\infty, \|\mathbf{w}_1 - \mathbf{w}_2\|_\infty\right),$$

for every $(\mathbf{v}_1, \mathbf{w}_1), (\mathbf{v}_2, \mathbf{w}_2) \in B_\eta((\hat{\mathbf{v}}, \hat{\mathbf{w}}))$ and a.e. $t \in [\hat{t}, \hat{t}+\eta)$ such that $(t, \mathbf{v}_1, \mathbf{w}_1), (t, \mathbf{v}_2, \mathbf{w}_2) \in D_\Omega$. The analogous condition holds for $\Omega_i^U$.

**Lemma 3.5.3.** *Hypothesis 3.5.3 implies Hypothesis 3.5.2.*

*Proof.* Suppose that Hypothesis 3.5.3 holds. Choose any $(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}}) \in I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ and $(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$ such that $\mathbf{x} \equiv \mathbf{x}(\cdot, \mathbf{u}, \mathbf{x}_0)$ satisfies $\hat{\mathbf{v}} \leq \mathbf{x}(\hat{t}) \leq \hat{\mathbf{w}}$ and either $x_i(\hat{t}) = \hat{v}_i$ or $x_i(\hat{t}) = \hat{w}_i$ for at least one $i \in \{1, \ldots, n_x\}$. By Condition 1 of Hypothesis 3.5.3, we must have $(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}}) \in D_\Omega$. Then, let $\eta_\Omega, L_\Omega > 0$ be constants satisfying Condition 3 of Hypothesis 3.5.3.

Let $L = L_\Omega$. Noting that $\min(\hat{\mathbf{v}}, \mathbf{x}(\hat{t})) = \hat{\mathbf{v}}$ and $\max(\hat{\mathbf{w}}, \mathbf{x}(\hat{t})) = \hat{\mathbf{w}}$, choose $\eta \in (0, \eta_\Omega]$ small enough that

$$(\min(\mathbf{v}, \mathbf{x}(t)), \max(\mathbf{w}, \mathbf{x}(t))) \in B_{\eta_\Omega}((\hat{\mathbf{v}}, \hat{\mathbf{w}})), \tag{3.27}$$

for all $(\mathbf{v}, \mathbf{w}) \in B_\eta((\hat{\mathbf{v}}, \hat{\mathbf{w}}))$ and every $t \in [\hat{t}, \hat{t} + \eta)$. It will be shown that Hypothesis 3.5.2 holds with these definitions.

Choose any $(\mathbf{v}, \mathbf{w}) \in B_\eta((\hat{\mathbf{v}}, \hat{\mathbf{w}}))$. To show Condition 1 of Hypothesis 3.5.2, choose any $t \in [\hat{t}, \hat{t} + \eta)$ such that $(t, \mathbf{v}, \mathbf{w}) \in D_\Omega$ and suppose that $x_i(t) < v_i$. It follows that

$x_i(t) = \min(v_i, x_i(t))$. Noting that $\min(\mathbf{v}, \mathbf{x}(t)) \leq \mathbf{x}(t) \leq \max(\mathbf{w}, \mathbf{x}(t))$, Condition 1 of Hypothesis 3.5.3 implies that $(t, \min(\mathbf{v}, \mathbf{x}(t)), \max(\mathbf{w}, \mathbf{x}(t))) \in D_\Omega$ and $\mathbf{x}(t) \in \Omega_i^L(t, \min(\mathbf{v}, \mathbf{x}(t)), \max(\mathbf{w}, \mathbf{x}(t)))$.

Condition 3 of Hypothesis 3.5.3 can now be applied with $(\mathbf{v}_1, \mathbf{w}_1) = (\mathbf{v}, \mathbf{w})$ and $(\mathbf{v}_2, \mathbf{w}_2) = (\min(\mathbf{v}, \mathbf{x}(t)), \max(\mathbf{w}, \mathbf{x}(t)))$ to give

$$\begin{aligned} d_H(\Omega_i^L(t, \mathbf{v}, \mathbf{w}), &\Omega_i^L(t, \min(\mathbf{v}, \mathbf{x}(t)), \max(\mathbf{w}, \mathbf{x}(t)))), \\ &\leq L_\Omega \max\left(\|\mathbf{v} - \min(\mathbf{v}, \mathbf{x}(t))\|_\infty, \|\mathbf{w} - \max(\mathbf{w}, \mathbf{x}(t))\|_\infty\right), \\ &= L_\Omega \max\left(\|\max(\mathbf{0}, \mathbf{v} - \mathbf{x}(t))\|_\infty, \|\max(\mathbf{0}, \mathbf{x}(t) - \mathbf{w})\|_\infty\right). \end{aligned} \tag{3.28}$$

It was argued above that $\mathbf{x}(t) \in \Omega_i^L(t, \min(\mathbf{v}, \mathbf{x}(t)), \max(\mathbf{w}, \mathbf{x}(t)))$. Moreover, $\Omega_i^L(t, \mathbf{v}, \mathbf{w})$ is nonempty and compact by Condition 2 of Hypothesis 3.5.3. It then follows from the definition of the Hausdorff metric that $\exists \mathbf{z} \in \Omega_i^L(t, \mathbf{v}, \mathbf{w})$ such that

$$\|\mathbf{x}(t) - \mathbf{z}\|_\infty \leq L_\Omega \max\left(\|\max(\mathbf{0}, \mathbf{v} - \mathbf{x}(t))\|_\infty, \|\max(\mathbf{0}, \mathbf{x}(t) - \mathbf{w})\|_\infty\right). \tag{3.29}$$

This establishes Condition 1 of Hypothesis 3.5.2. Condition 2 is proven analogously.

□

In light of Theorem 3.5.1 and the previous two lemmas, the following result is now apparent.

**Theorem 3.5.4.** *Let* $\mathbf{v}, \mathbf{w} \in \mathcal{AC}(I, \mathbb{R}^{n_x})$ *satisfy*

(EX): *For every* $t \in I$ *and every index* $i$*,*

1. $(t, \mathbf{v}(t), \mathbf{w}(t)) \in D_\Omega$*,*
2. $\Omega_i^L(t, \mathbf{v}(t), \mathbf{w}(t)) \subset D$ *and* $\Omega_i^U(t, \mathbf{v}(t), \mathbf{w}(t)) \subset D$*.*

(IC): $\mathbf{v}(t_0) \leq \mathbf{x}_0 \leq \mathbf{w}(t_0)$, $\forall \mathbf{x}_0 \in X_0$.

(RHS): *For a.e.* $t \in I$ *and each index* $i$*,*

1. $\dot{v}_i(t) \leq f_i(t, \mathbf{p}, \mathbf{z})$ *for all* $\mathbf{z} \in \Omega_i^L(t, \mathbf{v}(t), \mathbf{w}(t))$ *and* $\mathbf{p} \in U$*,*

*2. $\dot{w}_i(t) \geq f_i(t, \mathbf{p}, \mathbf{z})$ for all $\mathbf{z} \in \Omega_i^U(t, \mathbf{v}(t), \mathbf{w}(t))$ and $\mathbf{p} \in U$.*

*If either Hypothesis 3.5.2 or Hypothesis 3.5.3 holds, then $\mathbf{v}(t) \leq \mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \leq \mathbf{w}(t)$, $\forall (t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$.*

### 3.5.2 Computation of State Bounds

This section briefly describes how state bounds can be computed using Theorem 3.5.4. The formulations presented here will be made more precise for the specific instances of Theorem 3.5.4 given in §3.6.

For each index $i$, let $\underline{f}_i, \overline{f}_i : I \times U \times D \to \mathbb{R}$ and consider the coupled system of ODEs described by

$$\begin{aligned}
\dot{v}_i(t) = &\min_{(\mathbf{p},\mathbf{z})} \underline{f}_i(t, \mathbf{p}, \mathbf{z}) & , \qquad v_i(t_0) = \min_{\mathbf{z} \in X_0} z_i, \\
&\text{s.t.} \quad \mathbf{z} \in \Omega_i^L(t, \mathbf{v}(t), \mathbf{w}(t)), \quad \mathbf{p} \in U & \\
\dot{w}_i(t) = &\max_{(\mathbf{p},\mathbf{z})} \overline{f}_i(t, \mathbf{p}, \mathbf{z}) & , \qquad w_i(t_0) = \max_{\mathbf{z} \in X_0} z_i, \\
&\text{s.t.} \quad \mathbf{z} \in \Omega_i^U(t, \mathbf{v}(t), \mathbf{w}(t)), \quad \mathbf{p} \in U &
\end{aligned} \tag{3.30}$$

for a.e. $t \in I$ and every index $i$. Of course, some regularity will be required of $\underline{f}_i$ and $\overline{f}_i$, as well as $\Omega_i^L$ and $\Omega_i^U$, in order for this system to have a well-defined solution. However, if (3.30) does permit a solution, and $\underline{f}_i$ and $\overline{f}_i$ are chosen appropriately, then this solution provides state bounds for (3.2).

**Corollary 3.5.5.** *Suppose that $\mathbf{v}, \mathbf{w} \in \mathcal{AC}(I, \mathbb{R}^{n_x})$ satisfy (3.30) for a.e. $t \in I$. Further, suppose that, for a.e. $t \in I$ and every index $i$, the functions $\underline{f}_i$ and $\overline{f}_i$ are such that $\underline{f}_i(t, \mathbf{p}, \mathbf{z}) \leq f_i(t, \mathbf{p}, \mathbf{z})$, $\forall (\mathbf{p}, \mathbf{z}) \in U \times \Omega_i^L(t, \mathbf{v}(t), \mathbf{w}(t))$ and $f_i(t, \mathbf{p}, \mathbf{z}) \leq \overline{f}_i(t, \mathbf{p}, \mathbf{z})$, $\forall (\mathbf{p}, \mathbf{z}) \in U \times \Omega_i^U(t, \mathbf{v}(t), \mathbf{w}(t))$. If either Hypothesis 3.5.2 or Hypothesis 3.5.3 holds, then $\mathbf{v}(t) \leq \mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \leq \mathbf{w}(t)$, $\forall (t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$.*

*Proof.* It follows immediately from (3.30) and the assumptions on the functions $\underline{f}_i$ and $\overline{f}_i$ that Hypotheses (IC) and (RHS) of Theorem 3.5.4 are satisfied. Furthermore, if $\mathbf{v}$ and $\mathbf{w}$ satisfy (3.30) on $I$, then they must remain in the domains of definition of the functions appearing in the right-hand sides of (3.30) . In particular, $(t, \mathbf{v}(t), \mathbf{w}(t)) \in D_\Omega$ and $\Omega_i^{L/U}(t, \mathbf{v}(t), \mathbf{w}(t)) \subset D$ for all $t \in I$ and every index $i$. Then $\mathbf{v}$ and $\mathbf{w}$ also satisfy Hypothesis (EX) of Theorem 3.5.4, and the conclusion follows. $\square$

Note that one possible choice of $\underline{f}_i$ and $\overline{f}_i$ that is guaranteed to satisfy Corollary 3.5.5 is $\underline{f}_i = \overline{f}_i = f_i$, for each $i$. However, this makes solving the optimization problems defining the right-hand sides of (3.30) prohibitively expensive in general. As with Harrison's method, it is possible to greatly simplify (3.30) through the use of interval extensions. For this implementation, the following assumptions are required.

**Assumption 3.5.6.**

1. $U$ and $X_0$ are $n_u$ and $n_x$-dimensional intervals, respectively.
2. An inclusion monotonic interval extension for $\mathbf{f}$, $[\mathbf{f}] : \mathfrak{D}_f \subset \mathbb{I}I \times \mathbb{I}U \times \mathbb{I}D \to \mathbb{IR}^{n_x}$ is available.

**Assumption 3.5.7.** $\Omega_i^L, \Omega_i^U : D_\Omega \to \mathbb{IR}^{n_x}$ for all $i \in \{1, \ldots, n_x\}$.

Under Assumptions 3.5.6 and 3.5.7, the basic interval implementation of Theorem 3.5.4 is given by the ODEs

$$\begin{aligned} \dot{v}_i(t) &= [f_i]^L([t,t], U, \Omega_i^L(t, \mathbf{v}(t), \mathbf{w}(t))), \\ \dot{w}_i(t) &= [f_i]^U([t,t], U, \Omega_i^U(t, \mathbf{v}(t), \mathbf{w}(t))), \\ [v_i(t_0), w_i(t_0)] &= X_{0,i}, \end{aligned} \tag{3.31}$$

for a.e. $t \in I$ and each index $i$.

**Corollary 3.5.8.** *Suppose that Assumptions 3.5.6 and 3.5.7 hold and let* $\mathbf{v}, \mathbf{w} \in \mathcal{AC}(I, \mathbb{R}^{n_x})$ *satisfy* (3.31) *a.e. on* $I$*. If either Hypothesis 3.5.2 or Hypothesis 3.5.3 holds, then* $\mathbf{v}(t) \leq \mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \leq \mathbf{w}(t)$*,* $\forall (t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$*.*

*Proof.* Since $\mathbf{v}$ and $\mathbf{w}$ satisfy (3.31) on $I$, they must remain in the domains of definition of the right-hand side functions. It follows that $(t, \mathbf{v}(t), \mathbf{w}(t)) \in D_\Omega$ and $([t,t], U, \Omega_i^{L/U}(t, \mathbf{v}(t), \mathbf{w}(t))) \in \mathfrak{D}_f$, $\forall t \in I$ and every $i$. The latter implies that

$\Omega_i^{L/U}(t, \mathbf{v}(t), \mathbf{w}(t)) \subset D$, $\forall t \in I$ and every $i$, and hence Hypothesis (EX) of Theorem 3.5.4 holds. Hypotheses (IC) of Theorem 3.5.4 is satisfied by (3.31). Finally, Hypothesis (RHS) is satisfied by (3.31) and the enclosure property of inclusion monotonic interval extensions (Theorem 2.3.4). The conclusion now follows from Theorem 3.5.4. □

The existence of a unique solution of (3.31) can be guaranteed, at least locally about $t_0$, provided that the following regularity assumptions hold.

**Assumption 3.5.9.**

1. $[\mathbf{f}]$ is continuous on $\mathfrak{D}_f$.

2. Let $i \in \{1, \ldots, n_x\}$ and let $(\hat{t}, \hat{Z}) \in I \times \mathbb{IR}^{n_x}$ satisfy $([\hat{t}, \hat{t}], U, \hat{Z}) \in \mathfrak{D}_f$. There exists $\eta, L > 0$ such that

$$d_H([f_i]([t,t], U, Z_1), [f_i]([t,t], U, Z_2)) \leq L d_H(Z_1, Z_2),$$

$\forall (Z_1, Z_2) \in B_\eta(\hat{Z})$ and every $t \in [\hat{t}, \hat{t} + \eta)$ such that $([t,t], U, Z_1), ([t,t], U, Z_2) \in \mathfrak{D}_f$.

**Assumption 3.5.10.** For all $i \in \{1, \ldots, n_x\}$, $\Omega_i^L, \Omega_i^U : D_\Omega \to \mathbb{IR}^{n_x}$ are continuous.

**Lemma 3.5.11.** *Suppose that Assumptions 3.5.6, 3.5.9, 3.5.10 and Hypothesis 3.5.3 hold. If there exists an open set $B \subset D$, a number $\epsilon > 0$, and an interval $J \equiv [t_0, t_0 + \epsilon]$ satisfying*

1. $J \times B_\epsilon((\mathbf{x}_0^L, \mathbf{x}_0^U)) \subset D_\Omega$,

2. $\mathbb{I}J \times \mathbb{I}U \times \mathbb{I}B \subset \mathfrak{D}_f$,

3. $\Omega_i^{L/U}(t_0, \mathbf{x}_0^L, \mathbf{x}_0^U) \subset B$ *for all* $i \in \{1, \ldots, n_x\}$,

*then there exists $I' = [t_0, t_0 + \eta] \subset I$, $\eta > 0$, and two functions $\mathbf{v}, \mathbf{w} \in \mathcal{AC}(I', \mathbb{R}^{n_x})$ satisfying* (3.31) *for a.e. $t \in I'$. Moreover, this solution is unique.*

*Proof.* Choose any $i \in \{1, \dots, n_x\}$ and let $\eta_\Omega, L_\Omega > 0$ be the constants of Condition 3 of Hypothesis 3.5.3 with $(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}}) \equiv (t_0, \mathbf{x}_0^L, \mathbf{x}_0^U)$. By hypothesis, $\Omega_i^L(t_0, \mathbf{x}_0^L, \mathbf{x}_0^U) \subset B$ and hence $([t_0, t_0], U, \Omega_i^L(t_0, \mathbf{x}_0^L, \mathbf{x}_0^U)) \in \mathfrak{D}_f$. Let $\eta_f, L_f > 0$ be the constants of Condition 2 of Assumption 3.5.9 with $(\hat{t}, \hat{Z}) \equiv (t_0, \Omega_i^L(t_0, \mathbf{x}_0^L, \mathbf{x}_0^U))$.

By Assumption 3.5.10, we may choose $\gamma \in (0, \min(\epsilon, \eta_\Omega, \eta_f)]$ so small that $\Omega_i^L$ maps $[t_0, t_0 + \gamma] \times B_\gamma((\mathbf{x}_0^L, \mathbf{x}_0^U))$ into $B \cap B_{\eta_f}(\Omega_i^L(t_0, \mathbf{x}_0^L, \mathbf{x}_0^U))$. Then, Condition 1 of Assumption 3.5.9 implies that the mapping $(t, \mathbf{v}, \mathbf{w}) \longmapsto [f_i]^L([t, t], U, \Omega_i^L(t, \mathbf{v}, \mathbf{w}))$ is defined and continuous on $[t_0, t_0 + \gamma] \times B_\gamma((\mathbf{x}_0^L, \mathbf{x}_0^U))$. Moreover,

$$\begin{aligned}
|[f_i]^L([t,t], U, \Omega_i^L(t, \mathbf{v}_1, \mathbf{w}_1)) - &[f_i]^L([t,t], U, \Omega_i^L(t, \mathbf{v}_2, \mathbf{w}_2))| \\
&\leq L_f d_H(\Omega_i^L(t, \mathbf{v}_1, \mathbf{w}_1), \Omega_i^L(t, \mathbf{v}_2, \mathbf{w}_2)), \\
&\leq L_f L_\Omega \max\left(\|\mathbf{v}_1 - \mathbf{v}_2\|_\infty, \|\mathbf{w}_1 - \mathbf{w}_2\|_\infty\right),
\end{aligned}$$

for every $(\mathbf{v}_1, \mathbf{w}_1), (\mathbf{v}_2, \mathbf{w}_2) \in B_\gamma((\mathbf{x}_0^L, \mathbf{x}_0^U))$ and a.e. $t \in [t_0, t_0 + \gamma]$.

Repeating this argument for $\Omega_i^U$ and all $i \in \{1, \dots, n_x\}$, it is possible to choose $\gamma$ so small that the right-hand sides of the ODEs (3.31) are defined and continuous on $[t_0, t_0 + \gamma] \times B_\gamma((\mathbf{x}_0^L, \mathbf{x}_0^U))$, and Lipschitz on $B_\gamma((\mathbf{x}_0^L, \mathbf{x}_0^U))$ uniformly on $[t_0, t_0 + \gamma]$. Then, the existence and uniqueness of a solution of (3.31) on some $[t_0, t_0 + \eta] \subset I$ follows from Theorem 3.1 in [91]. □

### 3.5.3 Recovering Harrison's Method

Consider again the standard case where no *a priori* enclosure is available. To recover Theorem 3.3.2 and Harrison's method from Theorem 3.5.4 and (3.31), we need only define $D_\Omega$, $\Omega_i^L$ and $\Omega_i^U$ appropriately and check Hypothesis 3.5.3.

Consider the definitions

$$\begin{aligned}
&D_\Omega \equiv \{(t, \mathbf{v}, \mathbf{w}) \in I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} : \mathbf{v} \leq \mathbf{w}\}, \qquad (3.32) \\
&\Omega_i^L(t, \mathbf{v}, \mathbf{w}) \equiv \mathcal{B}_i^L([\mathbf{v}, \mathbf{w}]), \\
&\Omega_i^U(t, \mathbf{v}, \mathbf{w}) \equiv \mathcal{B}_i^U([\mathbf{v}, \mathbf{w}]).
\end{aligned}$$

Let $(t, \mathbf{v}, \mathbf{w}) \in I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ and suppose there exists $(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$ satisfying $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \in [\mathbf{v}, \mathbf{w}]$ and $x_i(t, \mathbf{u}, \mathbf{x}_0) = v_i$. Then clearly $\mathbf{v} \leq \mathbf{w}$, and hence $(t, \mathbf{v}, \mathbf{w}) \in D_\Omega$. Furthermore, $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \in \mathcal{B}_i^L([\mathbf{v}, \mathbf{w}])$. Thus, Condition 1 of Hypothesis 3.5.3 holds. Since each $\Omega_i^L(t, \mathbf{v}, \mathbf{w})$ and $\Omega_i^U(t, \mathbf{v}, \mathbf{w})$ maps into $\mathbb{IR}^{n_x}$, Condition 2 holds as well. Condition 3 holds since

$$d_H\left(\Omega_i^L(t, \mathbf{v}, \mathbf{w}), \Omega_i^L(t, \mathbf{v}', \mathbf{w}')\right) = \max\left(\max_j |v_j - v_j'|, \max_{j \neq i} |w_j - w_j'|\right),$$

for all $(t, \mathbf{v}, \mathbf{w}), (t, \mathbf{v}', \mathbf{w}') \in D_\Omega$ (analogous arguments hold for $\Omega_i^U$). Now, Theorem 3.5.4 reduces to Theorem 3.3.2, and the interval implementation (3.31) reduces to Harrison's method.

### 3.5.4 Extending $D_\Omega$

In the definitions (3.32), $D_\Omega$ is not open with respect to variations in $(\mathbf{v}, \mathbf{w})$ with $t$ fixed. This also turns out to be the case for many of the more obvious definitions of $D_\Omega$, $\Omega_i^L$ and $\Omega_i^U$ making use of *a priori* enclosures in §3.6. In general, this is undesirable for two reasons. First, Hypothesis 1 of Lemma 3.5.11 will not hold in general, so this result cannot be used to guarantee that the ODEs (3.31) have a solution. Second, it potentially causes problems when solving (3.31) numerically. Fortunately, Hypothesis 3.5.3 allows considerable freedom in the choice of $D_\Omega$, $\Omega_i^L$ and $\Omega_i^U$, so that this problem can almost always be avoided. In the case where no *a priori* enclosure is used, a better definition can be obtained through the use of the $\Box$ mapping defined in Definition 2.5.17.

Consider the definitions

$$\begin{aligned} D_\Omega &\equiv I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}, \\ \Omega_i^L(t, \mathbf{v}, \mathbf{w}) &\equiv \mathcal{B}_i^L(\Box(\mathbf{v}, \mathbf{w})), \\ \Omega_i^U(t, \mathbf{v}, \mathbf{w}) &\equiv \mathcal{B}_i^U(\Box(\mathbf{v}, \mathbf{w})). \end{aligned} \tag{3.33}$$

As with the definitions (3.32), it is straightforward to show that Hypothesis 3.5.3 holds with the definitions (3.33). Then, using (3.31), these definitions provide a variant of Harrison's method that is theoretically and numerically better behaved. They also provide an interesting variant of Theorem 3.3.2 where, notably, Hypothesis (EX) no longer requires that $\mathbf{v}(t) \leq \mathbf{w}(t)$, $\forall t \in I$. This does not contradict Example 3.3.2 because the hypothesis (RHS) is strengthened under the definitions (3.33).

**Corollary 3.5.12.** *Let $D_\Omega$, $\Omega_i^L$ and $\Omega_i^U$ be defined by* (3.33)*. Let Assumptions 3.5.6 and 3.5.9 hold. If there exists an open set $B \subset D$, a number $\epsilon > 0$, and an interval $J \equiv [t_0, t_0 + \epsilon]$ satisfying*

1. $\mathbb{I}J \times \mathbb{I}U \times \mathbb{I}B \subset \mathfrak{D}_f$,

2. $\Omega_i^{L/U}(t_0, \mathbf{x}_0^L, \mathbf{x}_0^U) \subset B$ *for all* $i \in \{1, \ldots, n_x\}$,

*then there exists $I' = [t_0, t_0 + \eta] \subset I$, $\eta > 0$, and a unique solution of* (3.31) *on $t \in I'$ satisfying $\mathbf{v}(t) \leq \mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \leq \mathbf{w}(t)$, $\forall (t, \mathbf{u}, \mathbf{x}_0) \in I' \times \mathcal{U} \times X_0$.*

*Proof.* Assumption 3.5.10 and Hypotheses 1 of Lemma 3.5.11 both clearly hold. Then, existence and uniqueness follows from Lemma 3.5.11, and the bounding property from Corollary 3.5.8. □

## 3.6 State Bounds with a Priori Enclosures

In this section, it is again assumed that, by physical or mathematical arguments, $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0)$ is known *a priori* to lie in some crude enclosure $G \subset \mathbb{R}^{n_x}$, for all $(t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$. We consider the use of such information in the context of Theorem 3.5.4 to derive state bounds under much weaker hypotheses than those required by Theorem 3.3.2. In most of the cases considered, efficient methods for computing these improved bounds follow directly from Corollary 3.5.8.

Because the functions $\Omega_i^L$ and $\Omega_i^U$ in Theorem 3.5.4 are permitted to vary with $t$, it is possible to handle the more general situation where $G : I \to \mathcal{P}(\mathbb{R}^{n_x})$ and $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \in G(t)$, for all $(t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$. One example of this is considered in §3.7.

### 3.6.1 An Interval Approach for General Enclosures

Consider an arbitrary *a priori* enclosure $G$ and suppose that the following mapping is available.

**Definition 3.6.1.** Let $D_{\mathcal{I}} \subset \mathbb{IR}^{n_x}$ be such that $\{Z \in \mathbb{IR}^{n_x} : Z \cap G \neq \emptyset\} \subset D_{\mathcal{I}}$, and let $\mathcal{I}_G : D_{\mathcal{I}} \to \mathbb{IR}^{n_x}$ satisfy

1. $\mathcal{I}_G(Z) \subset Z$ for all $Z \in D_{\mathcal{I}}$ with $Z \cap G \neq \emptyset$,

2. for any $Z \in D_{\mathcal{I}}$, if $\mathbf{z} \in Z$ and $\mathbf{z} \notin \mathcal{I}_G(Z)$, then $\mathbf{z} \notin G$,

3. for every $\hat{Z} \in D_{\mathcal{I}}$, $\exists \eta, L > 0$ such that $d_H(\mathcal{I}_G(Z_1), \mathcal{I}_G(Z_2)) \leq L_{\mathcal{I}} d_H(Z_1, Z_2)$, for all $Z_1, Z_2 \in D_{\mathcal{I}} \cap B_\eta(\hat{Z})$.

In words, $\mathcal{I}_G$ is a locally Lipschitz interval mapping which tightens a given interval $Z$ by discarding points which are not in $G$. We show that Hypothesis 3.5.3 holds with

$$\begin{aligned} D_\Omega &\equiv \{(t, \mathbf{v}, \mathbf{w}) \in I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} : \square(\mathbf{v}, \mathbf{w}) \in D_{\mathcal{I}}\}, \\ \Omega_i^L(t, \mathbf{v}, \mathbf{w}) &\equiv \mathcal{B}_i^L(\mathcal{I}_G(\square(\mathbf{v}, \mathbf{w}))), \\ \Omega_i^U(t, \mathbf{v}, \mathbf{w}) &\equiv \mathcal{B}_i^U(\mathcal{I}_G(\square(\mathbf{v}, \mathbf{w}))). \end{aligned} \tag{3.34}$$

To show Condition 1 of Hypothesis 3.5.3, let $(t, \mathbf{v}, \mathbf{w}) \in I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ and suppose that there exists $(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$ such that $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \in [\mathbf{v}, \mathbf{w}]$. Then $[\mathbf{v}, \mathbf{w}] \cap G \neq \emptyset$ and hence $[\mathbf{v}, \mathbf{w}] \in D_{\mathcal{I}}$, so that $(t, \mathbf{v}, \mathbf{w}) \in D_\Omega$. Further, $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \in [\mathbf{v}, \mathbf{w}] \cap G$ implies that $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \in \mathcal{I}_G(\square(\mathbf{v}, \mathbf{w}))$ by the contrapositive of Condition 2 in Definition 3.6.1. If in addition $x_i(t, \mathbf{u}, \mathbf{x}_0) = v_i$ for some $i$, then $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \in \mathcal{B}_i^L(\mathcal{I}_G(\square(\mathbf{v}, \mathbf{w}))) = \Omega_i^L(\mathbf{v}, \mathbf{w})$ by Condition 1 in Definition 3.6.1. Condition 2 of Hypothesis 3.5.3 is true because each $\Omega_i^L$ and $\Omega_i^U$ maps into $\mathbb{IR}^{n_x}$. By Lemma 2.5.19 and Condition 3 in Definition 3.6.1, it is clear that each $\Omega_i^L$ and $\Omega_i^U$ is a composition of locally Lipschitz functions, so that Condition 3 of Hypothesis 3.5.3 holds as well.

By Corollary 3.5.8, state bounds for (3.2) are given by the solutions of (3.31) with the definitions (3.34). Thus, if a suitable mapping $\mathcal{I}_G$ can be derived, an enclosure of the reachable set of (3.2) which takes advantage of an arbitrary *a priori* enclosure can be computed efficiently using interval computations.

**Remark 3.6.2.** Solving (3.31) with the definitions (3.34) should be distinguished from the naïve approach of solving (3.31) with the definitions (3.33) and subsequently applying $\mathcal{I}_G$ (or simply intersecting with $G$). In the former, $G$ is used to prevent conservatism in the interval enclosure from propagating forward in time, resulting in a much tighter enclosure. Interested readers should also note that, in contrast to Harrison's method, the validity of the method presented here does not follow readily from the standard results of viability theory, since it was not required that $G$ be an invariance domain and hence no assumption was made concerning the values of $\mathbf{f}$ on $\partial([\mathbf{v}(t), \mathbf{w}(t)] \cap G)$. These observations hold equally for all methods in the remainder of §3.6.

If $\mathcal{I}_G$ is defined on all of $\mathbb{IR}^{n_x}$, then another valid bounding method results from inverting the order of the operations $\mathcal{B}_i^{L/U}$ and $\mathcal{I}_G$ in (3.34). To verify this, we need only show that Hypotheses 3.5.3 holds with

$$\begin{aligned} &D_\Omega \equiv I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}, \\ &\Omega_i^L(t, \mathbf{v}, \mathbf{w}) \equiv \mathcal{I}_G(\mathcal{B}_i^L(\square(\mathbf{v}, \mathbf{w}))), \\ &\Omega_i^U(t, \mathbf{v}, \mathbf{w}) \equiv \mathcal{I}_G(\mathcal{B}_i^U(\square(\mathbf{v}, \mathbf{w}))). \end{aligned} \tag{3.35}$$

The mapping $\mathcal{I}_G(\mathcal{B}_i^{L/U}(\cdot))$ is defined on $\mathbb{IR}^{n_x}$ and maps into $\mathbb{IR}^{n_x}$ in a locally Lipschitz manner by Condition 3 of Definition 3.6.1. Further, for any $(t, \mathbf{v}, \mathbf{w}) \in I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$, if there exists $(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$ satisfying $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \in [\mathbf{v}, \mathbf{w}]$ and $x_i(t, \mathbf{u}, \mathbf{x}_0) = v_i$ for some $i$, then $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0)$ is in $\mathcal{B}_i^L(\square(\mathbf{v}, \mathbf{w}))$ and hence in $\mathcal{I}_G(\mathcal{B}_i^L(\square(\mathbf{v}, \mathbf{w})))$ by Condition 2 of Definition 3.6.1. Thus, Hypotheses 3.5.3 holds and Corollary 3.5.8 shows that the solutions of (3.31) with the definitions (3.35) are state bounds for (3.2).

Evaluating $\Omega_i^{L/U}$ in (3.35) requires $2n_x$ evaluations of $\mathcal{I}_G$, as opposed to only one for the definitions in (3.34). However, the former is much more effective because $\mathcal{I}_G$ operates on each face of $[\mathbf{v}(t), \mathbf{w}(t)]$ independently.

### 3.6.2 An Interval Approach for Convex Polyhedra

Suppose that $G \equiv \{\mathbf{z} : \mathbf{A}\mathbf{z} \leq \mathbf{b}\}$, with $\mathbf{A} \in \mathbb{R}^{m \times n_x}$ and $\mathbf{b} \in \mathbb{R}^m$. One possible mapping $\mathcal{I}_G$ is constructed as follows. Denote the $k^{\text{th}}$ row of $\mathbf{A}$ by $\mathbf{A}_k$ and the elements by $a_{k,i}$. It is desirable to tighten a given interval $[\mathbf{v}, \mathbf{w}]$ by excluding only points which violate $\mathbf{A}_k\mathbf{z} - b_k \leq 0$ for at least one $k$. Supposing that $a_{k,i} > 0$, rearranging this inequality for $z_i$ and applying interval arithmetic to bound the right-hand side from above gives

$$z_i \leq \frac{1}{a_{k,i}} \left( \sum_{j \neq i} \max\left(-a_{k,j}v_j, -a_{k,j}w_j\right) + b_k \right). \tag{3.36}$$

If (3.36) is satisfied with $w_i$ on the left-hand side, then $w_i$ cannot be tightened without excluding points which satisfy $\mathbf{A}_k\mathbf{z} - b_k \leq 0$. On the other hand, if (3.36) is false with $v_i$ on the left-hand side, then no element of $[\mathbf{v}, \mathbf{w}]$ satisfies $\mathbf{A}_k\mathbf{z} - b_k \leq 0$ and the assignment $w_i := v_i$ only eliminates points violating $\mathbf{A}_k\mathbf{z} - b_k \leq 0$ from the resulting interval. Finally, if (3.36) is false with $w_i$ on the left-hand side, then no vector $\mathbf{z} \in [\mathbf{v}, \mathbf{w}]$ with $z_i = w_i$ can possibly satisfy $\mathbf{A}_k\mathbf{z} - b_k \leq 0$, and the assignment $w_i := \frac{1}{a_{k,i}} \left( \sum_{j \neq i} \max\left(-a_{k,j}v_j, -a_{k,j}w_j\right) + b_k \right)$ only eliminates points violating $\mathbf{A}_k\mathbf{z} - b_k \leq 0$ from the resulting interval. Applying the same logic to the case where $a_{k,i} < 0$, it can be seen that Definition 3.6.1 is satisfied by the following mapping.

**Definition 3.6.3.** Define $\mathcal{I}_G$ for any $[\mathbf{v}, \mathbf{w}] \in D_{\mathcal{I}} \equiv \mathbb{IR}^{n_x}$ by the procedure:

1. Assign $[\hat{\mathbf{v}}, \hat{\mathbf{w}}] := [\mathbf{v}, \mathbf{w}]$, set $k = 1$ and set $i = 1$.

2. If $a_{k,i} = 0$, go to 3. Let $\gamma$ be the middle value of $\hat{v}_i$, $\hat{w}_i$ and $\frac{1}{a_{k,i}} \left( \sum_{j \neq i} \max(-a_{k,j}\hat{v}_j, -a_{k,j}\hat{w}_j) + b_k \right)$.
   If $a_{k,i} > 0$, set $\hat{w}_i := \gamma$. If $a_{k,i} < 0$, set $\hat{v}_i := \gamma$.

3. If $k < m$, set $k := k + 1$ and go to 2.

4. If $i < n_x$, set $k := 1$ and $i := i + 1$ and go to 2.

5. Set $\mathcal{I}_G([\mathbf{v}, \mathbf{w}]) := [\hat{\mathbf{v}}, \hat{\mathbf{w}}]$.

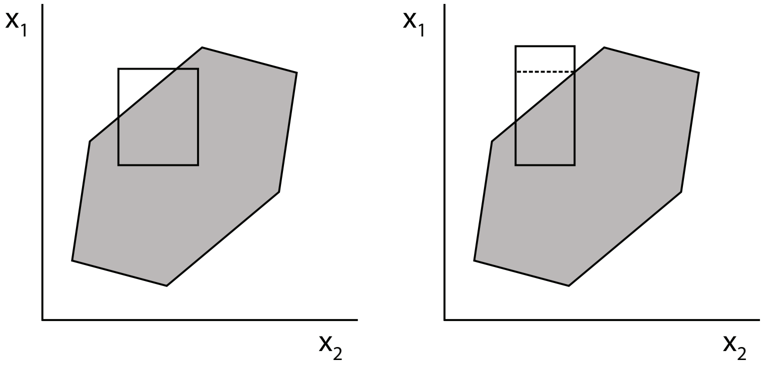

Figure 3-2: Schematic representation of the bounds tightening procedure described in Definition 3.6.3. Shaded regions depict $G$; boxes depict hypothetical intervals $[\mathbf{v}, \mathbf{w}]$. Left: $[\mathbf{v}, \mathbf{w}]$ is not entirely contained within the shaded region, yet no bound can be refined without excluding points in $[\mathbf{v}, \mathbf{w}] \cap G$. Right: $w_1$ may be reduced to the dashed line without excluding any point in $[\mathbf{v}, \mathbf{w}] \cap G$.

The bounds tightening procedure described in Definition 3.6.3 is represented schematically in Figure 3-2. In each panel, the shaded region depicts $G$, while the boxes depict hypothetical intervals $[\mathbf{v}, \mathbf{w}]$. On the left, $[\mathbf{v}, \mathbf{w}]$ is not entirely contained within the shaded region, yet no bound can be refined without excluding points in $[\mathbf{v}, \mathbf{w}] \cap G$. Alternatively, the right-hand schematic shows a situation where $w_1$ may be reduced to the dashed line without excluding any point in $[\mathbf{v}, \mathbf{w}] \cap G$.

With Definition 3.6.3, Conditions 1 and 2 of Definition 3.6.1 are satisfied by construction. Noting that the function $\text{mid}(a, b, c)$, which returns the middle value of its arguments, is Lipschitz on $\mathbb{R}^3$ with constant 1, Condition 3 can be verified by observing that $\mathcal{I}_G$ is computed by executing a finite number of operations on $\mathbf{v}$ and $\mathbf{w}$, each of which is clearly Lipschitz (addition, constant multiplication, mid, etc.). Thus, two bounding methods result from Definition 3.6.3; one through the definitions (3.34), and the other through the definitions (3.35). In practice, we find that the additional cost associated with (3.35) is far outweighed by the quality of the resulting enclosures. This method is demonstrated in Chapter 4.

### 3.6.3 An Optimization Approach for Convex Polyhedra

Consider again the case where $G$ is a convex polyhedral set, $G \equiv \{\mathbf{z} : \mathbf{A}\mathbf{z} \leq \mathbf{b}\}$. Another useful instance of Theorem 3.5.4 follows from the definitions

$$\begin{aligned} D_\Omega &\equiv \{(t, \mathbf{v}, \mathbf{w}) \in I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} : [\mathbf{v}, \mathbf{w}] \cap G \neq \emptyset\}, \qquad (3.37)\\ \Omega_i^L(t, \mathbf{v}, \mathbf{w}) &\equiv \left\{\mathbf{z} \in [\mathbf{v}, \mathbf{w}] \cap G : z_i = \min_{\boldsymbol{\psi} \in [\mathbf{v},\mathbf{w}] \cap G} \psi_i\right\},\\ \Omega_i^U(t, \mathbf{v}, \mathbf{w}) &\equiv \left\{\mathbf{z} \in [\mathbf{v}, \mathbf{w}] \cap G : z_i = \max_{\boldsymbol{\psi} \in [\mathbf{v},\mathbf{w}] \cap G} \psi_i\right\}. \end{aligned}$$

Let $(t, \mathbf{v}, \mathbf{w}) \in I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$, and define $v_i^*(t, \mathbf{v}, \mathbf{w}) \equiv \min_{\boldsymbol{\psi} \in [\mathbf{v},\mathbf{w}] \cap G} \psi_i$. Condition 1 of Hypothesis 3.5.3 holds because, if $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \in [\mathbf{v}, \mathbf{w}]$ for some $(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$, then $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \in [\mathbf{v}, \mathbf{w}] \cap G$ by the definition of $G$, so $(t, \mathbf{v}, \mathbf{w}) \in D_\Omega$. Further, combining $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \in [\mathbf{v}, \mathbf{w}] \cap G$ with $x_i(t, \mathbf{u}, \mathbf{x}_0) = v_i$ implies that $v_i^*(t, \mathbf{v}, \mathbf{w}) \leq x_i(t, \mathbf{u}, \mathbf{x}_0) = v_i \leq v_i^*(t, \mathbf{v}, \mathbf{w})$, so that $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \in \Omega_i^L(t, \mathbf{v}, \mathbf{w})$.

Since each $\Omega_i^L(t, \mathbf{v}, \mathbf{w})$ and $\Omega_i^U(t, \mathbf{v}, \mathbf{w})$ is a nonempty, bounded polyhedral set, Condition 2 of Hypothesis 3.5.3 also holds. To show Condition 3, the following Theorem is required.

**Theorem 3.6.4.** *Fix any $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{c} \in \mathbb{R}^n$ and, for each $\mathbf{b} \in \mathbb{R}^m$, define $S(\mathbf{b}) \equiv \{\mathbf{z} : \mathbf{A}\mathbf{z} \leq \mathbf{b}\}$ and $S^*(\mathbf{b}) \equiv \arg\min_{\mathbf{z} \in S(\mathbf{b})} \mathbf{c}^{\mathrm{T}}\mathbf{z}$. $\exists L \in \mathbb{R}_+$ such that $d_H(S(\mathbf{b}), S(\mathbf{b}')) \leq L\|\mathbf{b} - \mathbf{b}'\|_\infty$ and $d_H(S^*(\mathbf{b}), S^*(\mathbf{b}')) \leq L\|\mathbf{b} - \mathbf{b}'\|_\infty$, $\forall \mathbf{b}, \mathbf{b}' \in \mathbb{R}^m$, provided that these sets are nonempty.*

*Proof.* See Theorems 2.2 and 2.4 in [115]. □

Theorem 3.6.4 shows that $v_i^*(t, \mathbf{v}, \mathbf{w})$ is a Lipschitz mapping on $D_\Omega$ because $\mathbf{v}$ and $\mathbf{w}$ only effect the right-hand side data of the linear program $\min_{\boldsymbol{\psi} \in [\mathbf{v},\mathbf{w}] \cap G} \psi_i$. Then, noting that $\Omega_i^L(t, \mathbf{v}, \mathbf{w}) = \{\mathbf{z} \in [\mathbf{v}, \mathbf{w}] \cap G : z_i = v_i^*(t, \mathbf{v}, \mathbf{w})\}$, a second application of Theorem 3.6.4 gives

$$\begin{aligned} d_H(\Omega_i^L(t, \mathbf{v}, \mathbf{w}), \Omega_i^L(t, \mathbf{v}', \mathbf{w}')) &\leq L_1 \left(\|\mathbf{v} - \mathbf{v}'\|_\infty + \|\mathbf{w} - \mathbf{w}'\|_\infty + |v_i^*(t, \mathbf{v}, \mathbf{w}) - v_i^*(t, \mathbf{v}', \mathbf{w}')|\right)\\ &\leq L_1 L_2 \left(\|\mathbf{v} - \mathbf{v}'\|_\infty + \|\mathbf{w} - \mathbf{w}'\|_\infty\right), \end{aligned}$$

for all $(t, \mathbf{v}, \mathbf{w}), (t, \mathbf{v}', \mathbf{w}') \in D_\Omega$.

Now by Theorem 3.5.4 and Corollary 3.5.5, if the functions $\underline{f}_i$ and $\overline{f}_i$ are chosen appropriately, state bounds are given by the solutions, if any, of the system of ODEs:

$$\begin{aligned}
\dot{v}_i(t) = &\min_{(\mathbf{p},\mathbf{z})} \underline{f}_i(t, \mathbf{p}, \mathbf{z}) && , \qquad v_i(t_0) = \min_{\mathbf{z}\in X_0} z_i, \\
&\text{s.t.} \quad \mathbf{z} \in [\mathbf{v}(t), \mathbf{w}(t)] \cap G, \quad \mathbf{p} \in U \\
&\qquad z_i = \min_{\boldsymbol{\psi}\in[\mathbf{v}(t),\mathbf{w}(t)]\cap G} \psi_i \\
\dot{w}_i(t) = &\max_{(\mathbf{p},\mathbf{z})} \overline{f}_i(t, \mathbf{p}, \mathbf{z}) && , \qquad w_i(t_0) = \max_{\mathbf{z}\in X_0} z_i, \\
&\text{s.t.} \quad \mathbf{z} \in [\mathbf{v}(t), \mathbf{w}(t)] \cap G, \quad \mathbf{p} \in U \\
&\qquad z_i = \max_{\boldsymbol{\psi}\in[\mathbf{v}(t),\mathbf{w}(t)]\cap G} \psi_i
\end{aligned} \tag{3.38}$$

for a.e. $t \in I$ and each $i$. In the case where $U$ and $X_0$ are convex polyhedral sets and $\underline{f}_i$ and $\overline{f}_i$ are chosen as affine relaxations of $f_i$ for each $i$, evaluating the right-hand sides of (3.38) requires solving $2n_x$ bilevel linear programs. Thus, solving (3.38) computationally might seem impractical. On the other hand, the right-hand sides of (3.38) could in principle be reformulated as linear complementarity systems, for which efficient numerical solution seems possible. At present, there is no such numerical solver available, and the details of numerically implementing (3.38) are left for future consideration.

### 3.6.4 Comparison with Existing Results

As discussed in §3.4, the idea of including physical information in differential inequalities bounding methods is due to [162]. In that article, a method was developed for using interval *a priori* enclosures. To compare with the present developments, we let $G \equiv [\mathbf{g}^L, \mathbf{g}^U] \in \mathbb{IR}^{n_x}$, which is indeed a convex polyhedral set, and apply the methods of §3.6.2. With $\mathcal{I}_G$ defined as in Definition 3.6.3, it is easily verified that both (3.34) and (3.35) specify

$$\Omega_i^L(t,\mathbf{v},\mathbf{w}) \equiv \mathcal{B}_i^L([\mathbf{v},\mathbf{w}]) \cap G, \qquad (3.39)$$
$$\Omega_i^U(t,\mathbf{v},\mathbf{w}) \equiv \mathcal{B}_i^U([\mathbf{v},\mathbf{w}]) \cap G,$$

provided that $\mathbf{v} \leq \mathbf{w}$ and the above intersections are nonempty. These definitions also describe the method in [162]. However, while both (3.34) and (3.35) are well defined in the case of empty intersections, the proof in [162] is not clear on the appropriate action in this case. The text in [162] states that the choice of $\dot{v}_i$ (or $\dot{w}_i(t)$) is arbitrary in such cases. Though this is not justified in [162], it is proven in §3.7 below. Finally, note that the results in [162] were proven for parametric ODEs, while the present results provide the extension to control systems.

## 3.7 Differential Inequalities with Switching Conditions

Recall Hypothesis (RHSa) discussed in §3.4. In that section, it was shown that the standard comparison theorem, Theorem 3.3.2, does not hold with (RHSa) in place of (RHS), at least for some sets $G$. One of the primary complications leading to this situation is that the sets over which the differential inequalities in (RHSa) must hold can be empty. Theoretically, this causes problems because it trivializes the hypothesis; no meaningful condition is imposed on the corresponding $\dot{v}_i$ or $\dot{w}_i$ in such situations. However, it turns out that it is not necessary for all $2n_x$ of the conditions making up a (RHS) type hypothesis to hold for all $t \in I$.

In this section, we reproduce the derivation of the general comparison theorem of §3.5, only this time with an additional feature. It will be permissible that some or all of the $2n_x$ conditions in the (RHS) hypothesis, for at least some $t$, are *inactive*; i.e. simply do not hold. Given this possibility, we derive general requirements governing which of these conditions must hold, and when, so that a correct comparison theorem is nonetheless achieved.

### 3.7.1 Preliminaries

The following lemma and corollary are generalizations of Lemma 3.3.5 and Corollary 3.3.6.

**Lemma 3.7.1.** *Let $\boldsymbol{\delta} : I \to \mathbb{R}^n$ be a continuous function with $\boldsymbol{\delta}(t_0) \leq \mathbf{0}$. Suppose $\exists t \in I$ such that $\delta_i(t) > 0$ for at least one $i \in \{1, \ldots, n\}$, and define $t_1 \equiv \inf\{t \in I : \boldsymbol{\delta}(t) \nleq \mathbf{0}\}$. Then*

1. *$t_0 \leq t_1 < t_f$ and $\boldsymbol{\delta}(t) \leq \mathbf{0}$, $\forall t \in [t_0, t_1]$.*

2. *The set $\mathcal{V} \equiv \{i : \forall \gamma > 0,\ \exists t \in (t_1, t_1 + \gamma]$ s.t. $\delta_i(t) > 0\}$ is nonempty.*

*Let $t_4 \in (t_1, t_f]$, $\epsilon > 0$, $\beta \in L^1([t_1, t_4])$, and let $\mathcal{A}$ be a subset of $\{1, \ldots, n\}$ containing at least one element of $\mathcal{V}$. Then there exists an index $j \in \mathcal{A}$, a non-decreasing function $\rho \in \mathcal{AC}([t_1, t_4], \mathbb{R})$ satisfying (3.7) on $[t_1, t_4]$, and numbers $t_2, t_3 \in [t_1, t_4]$ with $t_2 < t_3$ such that the following inequalities hold:*

$$\begin{aligned} \delta_i(t) &< \rho(t), \quad \forall t \in [t_2, t_3),\ \forall i \in \mathcal{A}, \\ 0 &< \delta_j(t), \quad \forall t \in (t_2, t_3), \\ \delta_j(t_3) &= \rho(t_3), \\ \delta_j(t_2) &= 0. \end{aligned} \tag{3.40}$$

*Proof.* Conclusions 1 and 2 follow from Lemma 3.3.5. Choose any $t_4 \in (t_1, t_f]$, $\epsilon > 0$, $\beta \in L^1([t_1, t_4])$ and $\mathcal{A}$ as in the statement of the lemma. Choose $m$ so that $\exists t \in [t_1, t_4]$ with $\delta_i(t) \geq m > 0$, for some $i \in \mathcal{A}$. This must be possible since $\mathcal{A}$ contains an element of $\mathcal{V}$. By Lemma 3.3.4, there exists a non-decreasing function $\rho \in \mathcal{AC}([t_1, t_4], \mathbb{R})$ satisfying

$$0 < \rho(t) \leq \min(m/2, \epsilon), \quad \forall t \in [t_1, t_4], \quad \text{and} \quad \dot{\rho}(t) > |\beta(t)|\rho(t), \quad \text{a.e. } t \in [t_1, t_4].$$

Let $t_3 \equiv \inf\{t \in [t_1, t_4] : \delta_i(t) \geq \rho(t)$ for at least one $i \in \mathcal{A}\}$. Since $\rho < m$, this set is nonempty. Because $t_3$ is a lower bound, $\delta_i(t) < \rho(t)$, $\forall i \in \mathcal{A}$, for all $t \in [t_1, t_4]$ with

$t < t_3$. Since $t_3$ is the greatest lower bound, $\delta_j(t_3) = \rho(t_3)$ for at least one $j \in \mathcal{A}$. Since $\boldsymbol{\delta}(t_1) \leq \mathbf{0}$, it follows that $t_3 \in (t_1, t_4]$.

Fix any $j$ such that $\delta_j(t_3) = \rho(t_3)$ and let $t_2 \equiv \sup\{t \in [t_1, t_3] : \delta_j(t) \leq 0\}$. Since $\delta_j(t_1) \leq 0$, this set is nonempty. Because $t_2$ is an upper bound, $\delta_j(t) > 0$ for all $t \in [t_1, t_3]$ with $t > t_2$. Because it is the least upper bound, $\delta_j(t_2) = 0$. It follows that $t_2 \in [t_1, t_3)$. □

**Corollary 3.7.2.** *Let $\boldsymbol{\phi}, \mathbf{v}, \mathbf{w} : I \to \mathbb{R}^n$ be continuous and satisfy $\mathbf{v}(t_0) \leq \boldsymbol{\phi}(t_0) \leq \mathbf{w}(t_0)$. Suppose $\exists t \in I$ such that either $\phi_i(t) < v_i(t)$ or $\phi_i(t) > w_i(t)$, for at least one $i \in \{1, \ldots, n\}$, and define*

$$t_1 \equiv \inf\{t \in I : \phi_i(t) < v_i(t) \text{ or } \phi_i(t) > w_i(t), \text{ for at least one } i\}. \tag{3.41}$$

*Then*

1. *$t_0 \leq t_1 < t_f$ and $\mathbf{v}(t) \leq \boldsymbol{\phi}(t) \leq \mathbf{w}(t)$, $\forall t \in [t_0, t_1]$.*

2. *At least one of the sets*

$$\mathcal{V}^L \equiv \{i : \forall \gamma > 0, \ \exists t \in (t_1, t_1 + \gamma] \text{ s.t. } \phi_i(t) < v_i(t)\},$$
$$\mathcal{V}^U \equiv \{i : \forall \gamma > 0, \ \exists t \in (t_1, t_1 + \gamma] \text{ s.t. } \phi_i(t) > w_i(t)\},$$

   *is nonempty.*

*Let $t_4 \in (t_1, t_f]$, $\epsilon > 0$, $\beta \in L^1([t_1, t_4])$, and let $\mathcal{A}^L$ and $\mathcal{A}^U$ be subsets of $\{1, \ldots, n\}$ such that, either $\mathcal{A}^L \cap \mathcal{V}^L \neq \emptyset$ or $\mathcal{A}^U \cap \mathcal{V}^U \neq \emptyset$. Then there exists $j \in \mathcal{A}^L$ (or $j \in \mathcal{A}^U$), a non-decreasing function $\rho \in \mathcal{AC}([t_1, t_4], \mathbb{R})$ satisfying (3.7) on $[t_1, t_4]$, and numbers $t_2, t_3 \in [t_1, t_4]$ with $t_2 < t_3$ such that*

$$\phi_i(t) > v_i(t) - \rho(t), \quad \forall t \in [t_2, t_3), \ \forall i \in \mathcal{A}^L, \tag{3.42}$$

$$\phi_i(t) < w_i(t) + \rho(t), \quad \forall t \in [t_2, t_3), \ \forall i \in \mathcal{A}^U, \tag{3.43}$$

*and*

$$\phi_j(t_2) = v_j(t_2), \quad \phi_j(t_3) = v_j(t_3) - \rho(t_3), \quad \textit{and} \quad \phi_j(t) < v_j(t), \tag{3.44}$$

$$\Big( \textit{ or } \phi_j(t_2) = w_j(t_2), \quad \phi_j(t_3) = w_j(t_3) + \rho(t_3), \quad \textit{and} \quad \phi_j(t) > w_j(t), \Big) \tag{3.45}$$

*for all* $t \in (t_2, t_3)$*.*

*Proof.* Define $\boldsymbol{\delta} : I \to \mathbb{R}^{2n}$ by $\boldsymbol{\delta}(t) \equiv (\mathbf{v}(t) - \boldsymbol{\phi}(t), \boldsymbol{\phi}(t) - \mathbf{w}(t))$, $\forall t \in I$. By hypothesis, $\boldsymbol{\delta}(t_0) \leq \mathbf{0}$, and $\exists t \in I$ such that $\delta_i(t) > 0$ for at least one $i$. The conclusion now follows from Lemma 3.7.1. □

### 3.7.2 A General Comparison Theorem with Switching Conditions

Let $D_\Pi \subset I \times \mathbb{R}^n \times \mathbb{R}^n$ and, for every $i \in \{1, \ldots, n\}$, let $\Pi_i^L, \Pi_i^U : D_\Pi \to \mathcal{P}(\mathbb{R})$ and $s_i^L, s_i^U : D_\Pi \to \mathbb{R}$. Here, the mappings $\Pi_i^L$ and $\Pi_i^U$ will play exactly the same role as they did in §3.5. The new feature is the *switching conditions*, $s_i^L$ and $s_i^U$, the sign of which determines whether or not the corresponding differential inequality is required to hold. For any $(t, \mathbf{z}, \mathbf{v}, \mathbf{w}) \in I \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$, define the index sets

$$\begin{aligned}
\mathcal{V}^L(\mathbf{z}, \mathbf{v}, \mathbf{w}) &\equiv \{i : z_i < v_i\}, \\
\mathcal{V}^U(\mathbf{z}, \mathbf{v}, \mathbf{w}) &\equiv \{i : z_i > w_i\}, \\
\mathcal{A}^L(t, \mathbf{v}, \mathbf{w}) &\equiv \{i : s_i^L(t, \mathbf{v}, \mathbf{w}) > 0\}, \\
\mathcal{A}^U(t, \mathbf{v}, \mathbf{w}) &\equiv \{i : s_i^U(t, \mathbf{v}, \mathbf{w}) > 0\}.
\end{aligned}$$

The sets $\mathcal{V}^L$ and $\mathcal{V}^U$ are the sets of *violating indices*, respectively. The sets $\mathcal{A}^L$ and $\mathcal{A}^U$ are the sets of *active indices.*

Let $\boldsymbol{\phi} \in \mathcal{AC}(I, \mathbb{R}^n)$. As in §3.5, the problem of bounding $\boldsymbol{\phi}$ by two functions $\mathbf{v}, \mathbf{w} \in \mathcal{AC}(I, \mathbb{R}^n)$ is considered first. State bounds for the ODEs (3.2) are considered explicitly in §3.7.3. The following hypothesis gives a minimal set of conditions relating $\phi$ to the functions $s_i^L$, $s_i^U$, $\Pi_i^L$ and $\Pi_i^U$ in such a way that Theorem 3.7.4 below holds.

**Hypothesis 3.7.1.** Suppose that $(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}}) \in I \times \mathbb{R}^n \times \mathbb{R}^n$ satisfies $\hat{\mathbf{v}} \leq \boldsymbol{\phi}(\hat{t}) \leq \hat{\mathbf{w}}$ and either $\phi_i(\hat{t}) = \hat{v}_i$ or $\phi_i(\hat{t}) = \hat{w}_i$ for at least one $i \in \{1, \ldots, n\}$. Then there exists $\eta > 0$ and $\alpha \in L^1(I)$ such that the following conditions hold for every $(\mathbf{v}, \mathbf{w}) \in B_\eta((\hat{\mathbf{v}}, \hat{\mathbf{w}}))$ and a.e. $t \in [\hat{t}, \hat{t} + \eta)$ such that $(t, \mathbf{v}, \mathbf{w}) \in D_\Pi$:

1. If $\mathcal{V}^L(\boldsymbol{\phi}(t), \mathbf{v}, \mathbf{w}) \cup \mathcal{V}^U(\boldsymbol{\phi}(t), \mathbf{v}, \mathbf{w}) \neq \emptyset$, then at least one of the sets

$$\mathcal{Q}^L(t, \mathbf{v}, \mathbf{w}) \equiv \mathcal{A}^L(t, \mathbf{v}, \mathbf{w}) \cap \mathcal{V}^L(\boldsymbol{\phi}(t), \mathbf{v}, \mathbf{w}),$$
$$\mathcal{Q}^U(t, \mathbf{v}, \mathbf{w}) \equiv \mathcal{A}^U(t, \mathbf{v}, \mathbf{w}) \cap \mathcal{V}^U(\boldsymbol{\phi}(t), \mathbf{v}, \mathbf{w}),$$

   is nonempty.

2. If $i \in \mathcal{Q}^L(t, \mathbf{v}, \mathbf{w})$, then $\exists \sigma \in \Pi_i^L(t, \mathbf{v}, \mathbf{w})$ such that

$$|\sigma - \dot{\phi}_i(t)| \leq \alpha(t) \max\left( \max_{i \in \mathcal{Q}^L(t,\mathbf{v},\mathbf{w})} (v_i - \phi_i(t)), \max_{i \in \mathcal{Q}^U(t,\mathbf{v},\mathbf{w})} (\phi_i(t) - w_i) \right). \quad (3.46)$$

3. If $i \in \mathcal{Q}^U(t, \mathbf{v}, \mathbf{w})$, then $\exists \sigma \in \Pi_i^U(t, \mathbf{v}, \mathbf{w})$ such that (3.46) holds.

Theorem 3.7.4 requires one further technical assumption concerning *transition times*.

**Definition 3.7.3.** Let $\mathbf{v}, \mathbf{w} \in \mathcal{AC}(I, \mathbb{R}^n)$ satisfy $(t, \mathbf{v}(t), \mathbf{w}(t)) \in D_\Pi$, $\forall t \in I$. Call $t \in I$ a *transition time* for $(\mathbf{v}, \mathbf{w})$ if, for every $\delta > 0$, $\exists t', t'' \in B_\delta(t) \cap I$ such that either

$$\mathcal{A}^L(t', \mathbf{v}(t'), \mathbf{w}(t')) \neq \mathcal{A}^L(t'', \mathbf{v}(t''), \mathbf{w}(t'')), \quad \text{or}$$
$$\mathcal{A}^U(t', \mathbf{v}(t'), \mathbf{w}(t')) \neq \mathcal{A}^U(t'', \mathbf{v}(t''), \mathbf{w}(t'')).$$

A general comparison theorem can now be stated in terms of the mappings $\Pi_i^L$ and $\Pi_i^U$.

**Theorem 3.7.4.** *Let $\boldsymbol{\phi}, \mathbf{v}, \mathbf{w} \in \mathcal{AC}(I, \mathbb{R}^n)$ satisfy*

(EX): $(t, \mathbf{v}(t), \mathbf{w}(t)) \in D_\Pi$, $\forall t \in I$.

(IC): $\mathbf{v}(t_0) \leq \boldsymbol{\phi}(t_0) \leq \mathbf{w}(t_0)$.

(RHS): *For a.e.* $t \in I$ *and each index* $i$,

1. *If* $s_i^L(t, \mathbf{v}(t), \mathbf{w}(t)) > 0$, *then* $\dot{v}_i(t) \leq \sigma$ *for all* $\sigma \in \Pi_i^L(t, \mathbf{v}(t), \mathbf{w}(t))$,
2. *If* $s_i^U(t, \mathbf{v}(t), \mathbf{w}(t)) > 0$, *then* $\dot{w}_i(t) \geq \sigma$ *for all* $\sigma \in \Pi_i^U(t, \mathbf{v}(t), \mathbf{w}(t))$.

*If Hypotheses 3.7.1 holds and* $(\mathbf{v}, \mathbf{w})$ *has finitely many transition times in* $I$, *then* $\mathbf{v}(t) \leq \boldsymbol{\phi}(t) \leq \mathbf{w}(t)$, $\forall t \in I$.

*Proof.* Suppose that $\exists t \in I$ such that $\phi_i(t) < v_i(t)$ or $\phi_i(t) > w_i(t)$, for at least one $i \in \{1, \ldots, n\}$. We prove a contradiction.

Noting that the hypotheses of Corollary 3.7.2 are satisfied, define $t_1$ as in (3.41). By Conclusion 1 of Corollary 3.7.2, $\mathbf{v}(t_1) \leq \boldsymbol{\phi}(t_1) \leq \mathbf{w}(t_1)$. By continuity and Conclusion 2 of the same, there must exist at least one $i$ such that either $\phi_i(t_1) = v_i(t_1)$ or $\phi_i(t_1) = w_i(t_1)$. Let $\eta > 0$ and $\alpha \in L^1(I)$ satisfy Hypothesis 3.7.1 with $(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}}) \equiv (t_1, \mathbf{v}(t_1), \mathbf{w}(t_1))$. Choose $t_5 \in (t_1, t_f]$ small enough that

$$t \in [t_1, t_1 + \eta) \quad \text{and} \quad (\mathbf{v}(t), \mathbf{w}(t)) \in B_\eta((\mathbf{v}(t_1), \mathbf{w}(t_1))), \quad \forall t \in [t_1, t_5]. \tag{3.47}$$

Noting that $(t, \mathbf{v}(t), \mathbf{w}(t)) \in D_\Pi$ for all $t \in [t_1, t_5]$ by Hypothesis (EX), we are now guaranteed the conditions of Hypothesis 3.7.1 with $(t, \mathbf{v}, \mathbf{w}) \equiv (t, \mathbf{v}(t), \mathbf{w}(t))$, for a.e. $t \in [t_1, t_5]$.

By hypothesis, there are at most a finite number of transition times in $[t_1, t_5]$. Then, there must exist $t_4 \in (t_1, t_5]$ such that there are no transition times in $(t_1, t_4]$. Let $\mathcal{A}^L$ and $\mathcal{A}^U$ denote the constant sets $\mathcal{A}^L(t, \mathbf{v}(t), \mathbf{w}(t))$ and $\mathcal{A}^U(t, \mathbf{v}(t), \mathbf{w}(t))$ on $(t_1, t_4]$, respectively. Further, let $\mathcal{V}^L$ and $\mathcal{V}^U$ be as in Conclusion 2 of Corollary 3.7.2. In order to apply that corollary, it will now be shown that one of the sets $\mathcal{A}^L \cap \mathcal{V}^L$ or $\mathcal{A}^U \cap \mathcal{V}^U$ is nonempty.

If $i \notin \mathcal{V}^L$, then $t_4$ may be chosen small enough that $i \notin \mathcal{V}^L(\boldsymbol{\phi}(t), \mathbf{v}(t), \mathbf{w}(t))$, $\forall t \in (t_1, t_4]$. Using a similar argument for $\mathcal{V}^U$, choose $t_4$ small enough that

$$\mathcal{V}^L(\boldsymbol{\phi}(t), \mathbf{v}(t), \mathbf{w}(t)) \subset \mathcal{V}^L \quad \text{and} \quad \mathcal{V}^U(\boldsymbol{\phi}(t), \mathbf{v}(t), \mathbf{w}(t)) \subset \mathcal{V}^U, \quad \forall t \in (t_1, t_4]. \tag{3.48}$$

Now, Conclusion 2 of Corollary 3.7.2 implies that $\exists t \in (t_1, t_4]$ with at least one of $\mathcal{V}^L(\boldsymbol{\phi}(t), \mathbf{v}(t), \mathbf{w}(t))$ or $\mathcal{V}^U(\boldsymbol{\phi}(t), \mathbf{v}(t), \mathbf{w}(t))$ nonempty. Then, using Condition 1 of Hypothesis 3.7.1 and (3.48), it follows that at least one of the sets $\mathcal{A}^L \cap \mathcal{V}^L$ or $\mathcal{A}^U \cap \mathcal{V}^U$ is nonempty.

We now apply Corollary 3.7.2 with $t_4$, arbitrary $\epsilon > 0$, $\beta = \alpha$ and $\mathcal{A}^L$ and $\mathcal{A}^U$. This furnishes an index $j \in \mathcal{A}^L$ (or $j \in \mathcal{A}^U$), a non-decreasing function $\rho \in \mathcal{AC}([t_1, t_4], \mathbb{R})$ satisfying (3.7) on $[t_1, t_4]$, and numbers $t_2, t_3 \in [t_1, t_4]$ with $t_2 < t_3$ such that (3.42)-(3.43) and (3.44) (or (3.45)) hold. Assume that $j \in \mathcal{A}^L$, so that (3.44) holds. The proof is analogous if $j \in \mathcal{A}^U$ instead.

For a.e. $t \in [t_2, t_3]$, (3.44) implies that $j \in \mathcal{V}^L(\boldsymbol{\phi}(t), \mathbf{v}(t), \mathbf{w}(t))$. Furthermore, $j \in \mathcal{A}^L$ by construction. Then, let $\sigma \in \Pi_i^L(t, \mathbf{v}(t), \mathbf{w}(t))$ satisfy Condition 2 of Hypothesis 3.7.1. Using Hypotheses (RHS).1,

$$\begin{aligned} \dot{v}_j(t) - \dot{\phi}_j(t) &\leq \sigma - \dot{\phi}_j(t), \qquad (3.49)\\ &\leq |\sigma - \dot{\phi}_j(t)|,\\ &\leq \alpha(t) \max\left( \max_{i \in \mathcal{Q}^L(t,\mathbf{v}(t),\mathbf{w}(t))} (v_i(t) - \phi_i(t)), \max_{i \in \mathcal{Q}^U(t,\mathbf{v}(t),\mathbf{w}(t))} (\phi_i(t) - w_i(t)) \right), \end{aligned}$$

for a.e. $t \in [t_2, t_3]$. For any $i \in \mathcal{Q}^L(t, \mathbf{v}(t), \mathbf{w}(t))$, (3.42) ensures that $(v_i(t) - \phi_i(t)) < \rho(t)$. Using an analogous argument for $i \in \mathcal{Q}^U(t, \mathbf{v}(t), \mathbf{w}(t))$, (3.49) implies

$$\dot{v}_j(t) - \dot{\phi}_j(t) < \alpha(t)\rho(t), \quad \text{a.e. } t \in [t_2, t_3]. \qquad (3.50)$$

Finally, using (3.7) and recalling that we have used $\beta = \alpha$, this implies that

$$\begin{aligned} \dot{v}_j(t) - \dot{\phi}_j(t) - \dot{\rho}(t) &< \alpha(t)\rho(t) - \dot{\rho}(t), \qquad (3.51)\\ &< 0, \quad \text{a.e. } t \in [t_2, t_3]. \qquad (3.52) \end{aligned}$$

By Theorem 3.3.3, this implies that $(v_j - \phi_j - \rho)$ is non-increasing on $[t_2, t_3]$, so that $v_j(t_3) - \phi_j(t_3) - \rho(t_3) \leq v_j(t_2) - \phi_j(t_2) - \rho(t_2)$. But by (3.44), this implies that $0 \leq -\rho(t_2)$, which contradicts (3.7). □

The following hypothesis eliminates the need to assume a finite number of transition times in Theorem 3.7.4 by putting much more stringent requirements on the switching conditions.

**Hypothesis 3.7.2.** Suppose that $(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}}) \in I \times \mathbb{R}^n \times \mathbb{R}^n$ satisfies $\hat{\mathbf{v}} \leq \boldsymbol{\phi}(\hat{t}) \leq \hat{\mathbf{w}}$ and either $\phi_i(\hat{t}) = \hat{v}_i$ or $\phi_i(\hat{t}) = \hat{w}_i$ for at least one $i \in \{1, \ldots, n\}$. Then there exists $\eta > 0$ such that, for every $(\mathbf{v}, \mathbf{w}) \in B_\eta((\hat{\mathbf{v}}, \hat{\mathbf{w}}))$ and a.e. $t \in [\hat{t}, \hat{t}+\eta)$ such that $(t, \mathbf{v}, \mathbf{w}) \in D_\Pi$, $\mathcal{V}^L(\boldsymbol{\phi}(t), \mathbf{v}, \mathbf{w}) \subset \mathcal{A}^L(t, \mathbf{v}, \mathbf{w})$ and $\mathcal{V}^U(\boldsymbol{\phi}(t), \mathbf{v}, \mathbf{w}) \subset \mathcal{A}^U(t, \mathbf{v}, \mathbf{w})$.

**Theorem 3.7.5.** *Let $\boldsymbol{\phi}, \mathbf{v}, \mathbf{w} \in \mathcal{AC}(I, \mathbb{R}^n)$ satisfy*

(EX): $(t, \mathbf{v}(t), \mathbf{w}(t)) \in D_\Pi$, $\forall t \in I$.

(IC): $\mathbf{v}(t_0) \leq \boldsymbol{\phi}(t_0) \leq \mathbf{w}(t_0)$.

(RHS): *For a.e. $t \in I$ and each index $i$,*

1. *If $s_i^L(t, \mathbf{v}(t), \mathbf{w}(t)) > 0$, then $\dot{v}_i(t) \leq \sigma$ for all $\sigma \in \Pi_i^L(t, \mathbf{v}(t), \mathbf{w}(t))$,*
2. *If $s_i^U(t, \mathbf{v}(t), \mathbf{w}(t)) > 0$, then $\dot{w}_i(t) \geq \sigma$ for all $\sigma \in \Pi_i^U(t, \mathbf{v}(t), \mathbf{w}(t))$.*

*If Hypothesis 3.7.1 and Hypothesis 3.7.2 hold, then $\mathbf{v}(t) \leq \boldsymbol{\phi}(t) \leq \mathbf{w}(t)$, $\forall t \in I$.*

*Proof.* The proof is exactly the same as that of Theorem 3.5.1. It is only necessary to verify that the use of the (RHS) condition on $\dot{v}_j(t)$ is valid for a.e. $t \in [t_2, t_3]$. But by construction, $\phi_j(t) < v_j(t)$. Then, $j \in \mathcal{V}^L(\boldsymbol{\phi}(t), \mathbf{v}, \mathbf{w})$, and hence in $\mathcal{A}^L(t, \mathbf{v}, \mathbf{w})$ by Hypothesis 3.7.2. □

### 3.7.3 Specialization to State Bounds for ODEs

Let $D_\Omega \subset I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ and, for every $i \in \{1, \ldots, n_x\}$, let $s_i^L, s_i^U : D_\Omega \to \mathbb{R}$, $\Omega_i^L, \Omega_i^U : D_\Omega \to \mathcal{P}(\mathbb{R}^{n_x})$. To specialize Theorem 3.7.4 to the task of characterizing state bounds for (3.2), let $\boldsymbol{\phi} \equiv \mathbf{x}(\cdot, \mathbf{u}, \mathbf{x}_0)$, for some $(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$, and let $\Pi_i^L$ and $\Pi_i^U$ take the form

$$\Pi_i^L(t, \mathbf{v}, \mathbf{w}) \equiv \{f_i(t, \mathbf{p}, \mathbf{z}) : \mathbf{p} \in U,\ \mathbf{z} \in \Omega_i^L(t, \mathbf{v}, \mathbf{w})), \tag{3.53}$$

$$\Pi_i^U(t, \mathbf{v}, \mathbf{w}) \equiv \{f_i(t, \mathbf{p}, \mathbf{z}) : \mathbf{p} \in U,\ \mathbf{z} \in \Omega_i^U(t, \mathbf{v}, \mathbf{w})), \tag{3.54}$$

for all $(t, \mathbf{v}, \mathbf{w})$ in the set

$$D_\Pi \equiv \{(t, \mathbf{v}, \mathbf{w}) \in D_\Omega : \Omega_i^{L/U}(t, \mathbf{v}, \mathbf{w}) \subset D,\ i = 1, \ldots, n_x\}. \tag{3.55}$$

It will be shown that Hypothesis 3.7.1 is ensured by imposing the following conditions on $\Omega_i^L$ and $\Omega_i^U$:

**Hypothesis 3.7.3.** Let $(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}}) \in I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ and suppose that $\exists(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$ such that $\mathbf{x} \equiv \mathbf{x}(\cdot, \mathbf{u}, \mathbf{x}_0)$ satisfies $\hat{\mathbf{v}} \leq \mathbf{x}(\hat{t}) \leq \hat{\mathbf{w}}$ and either $x_i(\hat{t}) = \hat{v}_i$ or $x_i(\hat{t}) = \hat{w}_i$ for at least one $i \in \{1, \ldots, n_x\}$. Then there exist $\eta, L > 0$ such that the following conditions hold for every $(\mathbf{v}, \mathbf{w}) \in B_\eta((\hat{\mathbf{v}}, \hat{\mathbf{w}}))$ and a.e. $t \in [\hat{t}, \hat{t}+\eta)$ such that $(t, \mathbf{v}, \mathbf{w}) \in D_\Omega$:

1. If $\mathcal{V}^L(\mathbf{x}(t), \mathbf{v}, \mathbf{w}) \cup \mathcal{V}^U(\mathbf{x}(t), \mathbf{v}, \mathbf{w}) \neq \emptyset$, then at least one of the sets

$$\mathcal{Q}^L(t, \mathbf{v}, \mathbf{w}) \equiv \mathcal{A}^L(t, \mathbf{v}, \mathbf{w}) \cap \mathcal{V}^L(\mathbf{x}(t), \mathbf{v}, \mathbf{w}),$$
$$\mathcal{Q}^U(t, \mathbf{v}, \mathbf{w}) \equiv \mathcal{A}^U(t, \mathbf{v}, \mathbf{w}) \cap \mathcal{V}^U(\mathbf{x}(t), \mathbf{v}, \mathbf{w}),$$

   is nonempty.

2. If $i \in \mathcal{Q}^L(t, \mathbf{v}, \mathbf{w})$, then $\exists \mathbf{z} \in \Omega_i^L(t, \mathbf{v}, \mathbf{w})$ such that

$$\|\mathbf{x}(t) - \mathbf{z}\|_\infty \leq L \max\left( \max_{i \in \mathcal{Q}^L(t,\mathbf{v},\mathbf{w})} (v_i - x_i(t)), \max_{i \in \mathcal{Q}^U(t,\mathbf{v},\mathbf{w})} (x_i(t) - w_i) \right). \tag{3.56}$$

3. If $i \in \mathcal{Q}^U(t, \mathbf{v}, \mathbf{w})$, then $\exists \mathbf{z} \in \Omega_i^U(t, \mathbf{v}, \mathbf{w})$ such that (3.56) holds.

**Lemma 3.7.6.** *Suppose that Hypothesis 3.7.3 holds. Then, for any* $(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$*, Hypothesis 3.7.1 holds with* $\boldsymbol{\phi} \equiv \mathbf{x}(\cdot, \mathbf{u}, \mathbf{x}_0)$ *and the definitions* (3.53)*,* (3.54) *and* (3.55)*.*

*Proof.* Choose any $(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$ and define $\boldsymbol{\phi} \equiv \mathbf{x}(\cdot, \mathbf{u}, \mathbf{x}_0)$. Let $(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}}) \in I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ and suppose that $\hat{\mathbf{v}} \leq \boldsymbol{\phi}(\hat{t}) \leq \hat{\mathbf{w}}$ and either $\phi_i(\hat{t}) = \hat{v}_i$ or $\phi_i(\hat{t}) = \hat{w}_i$ for at least one $i \in \{1, \ldots, n_x\}$. Noting that $(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}})$ satisfies the required properties, let $L_\Omega, \eta_\Omega > 0$ be constants satisfying Hypothesis 3.7.3.

Let $\eta_f > 0$ and $\alpha_f \in L^1(I)$ be given by Assumption 3.2.2 with $\mathbf{z} = \boldsymbol{\phi}(\hat{t})$. Define $\alpha \equiv L_\Omega \alpha_f$ and choose $\eta \in (0, \min(\eta_f, \eta_\Omega)]$ small enough that

$$\|\boldsymbol{\phi}(t) - \boldsymbol{\phi}(\hat{t})\|_\infty < \eta_f/2, \tag{3.57}$$

$$L_\Omega \max\left(\|\max(\mathbf{0}, \mathbf{v} - \boldsymbol{\phi}(t))\|_\infty, \|\max(\mathbf{0}, \boldsymbol{\phi}(t) - \mathbf{w})\|_\infty\right) < \eta_f/2, \tag{3.58}$$

for all $t \in [\hat{t}, \hat{t} + \eta)$ and every $(\mathbf{v}, \mathbf{w}) \in B_\eta((\hat{\mathbf{v}}, \hat{\mathbf{w}}))$. It will be shown that Hypothesis 3.7.1 holds with these definitions.

Choose any $(\mathbf{v}, \mathbf{w}) \in B_\eta((\hat{\mathbf{v}}, \hat{\mathbf{w}}))$. For a.e. $t \in [\hat{t}, \hat{t} + \eta)$ such that $(t, \mathbf{v}, \mathbf{w}) \in D_\Pi$, the conditions of Hypothesis 3.7.3 hold because $\eta \leq \eta_\Omega$ and $D_\Pi \subset D_\Omega$. Condition 1 of Hypothesis 3.7.1 follows directly from Condition 1 of Hypothesis 3.7.3. Suppose that $i \in \mathcal{Q}^L(t, \mathbf{v}, \mathbf{w})$. By Condition 2 of Hypothesis 3.7.3, $\exists \mathbf{z} \in \Omega_i^L(t, \mathbf{v}, \mathbf{w}) \subset D$ such that (3.56) holds with $L = L_\Omega$ and $\mathbf{x} = \boldsymbol{\phi}$. Combining this with (3.58) implies that $\|\boldsymbol{\phi}(t) - \mathbf{z}\|_\infty < \eta_f/2$. By (3.57) and the triangle inequality, it follows that $\mathbf{z} \in B_{\eta_f}(\boldsymbol{\phi}(\hat{t}))$. This implies that the inequality of Assumption 3.2.2 can be applied to the points $\mathbf{z}$ and $\boldsymbol{\phi}(t)$.

Let $\sigma \equiv f_i(t, \mathbf{u}(t), \mathbf{z})$. By definition, $\sigma \in \Pi_i^L(t, \mathbf{v}, \mathbf{w})$. Moreover,

$$\begin{aligned}
|\sigma - \dot{\phi}_i(t)| &= |f_i(t, \mathbf{u}(t), \mathbf{z}) - f_i(t, \mathbf{u}(t), \boldsymbol{\phi}(t))|, && (3.59)\\
&\leq \alpha_f(t)\|\boldsymbol{\phi}(t) - \mathbf{z}\|_\infty, && (3.60)\\
&\leq \alpha_f(t) L_\Omega \max\left(\max_{i \in \mathcal{Q}^L(t,\mathbf{v},\mathbf{w})}(v_i - \phi_i(t)), \max_{i \in \mathcal{Q}^U(t,\mathbf{v},\mathbf{w})}(\phi_i(t) - w_i)\right). && (3.61)
\end{aligned}$$

This proves Condition 2 of Hypothesis 3.7.1, and Condition 3 follows by an analogous argument. □

It will also be convenient to have an analogue of Hypothesis 3.7.2 in terms of $\Omega_i^L$ and $\Omega_i^U$.

**Hypothesis 3.7.4.** Let $(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}}) \in I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ and suppose that $\exists (\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$ such that $\mathbf{x} \equiv \mathbf{x}(\cdot, \mathbf{u}, \mathbf{x}_0)$ satisfies $\hat{\mathbf{v}} \leq \mathbf{x}(\hat{t}) \leq \hat{\mathbf{w}}$ and either $x_i(\hat{t}) = \hat{v}_i$ or $x_i(\hat{t}) = \hat{w}_i$ for at least one $i \in \{1, \ldots, n_x\}$. Then there exist $\eta > 0$ such that, for every $(\mathbf{v}, \mathbf{w}) \in$

$B_\eta((\hat{\mathbf{v}}, \hat{\mathbf{w}}))$ and a.e. $t \in [\hat{t}, \hat{t}+\eta)$ such that $(t, \mathbf{v}, \mathbf{w}) \in D_\Omega$, $\mathcal{V}^L(\mathbf{x}(t), \mathbf{v}, \mathbf{w}) \subset \mathcal{A}^L(t, \mathbf{v}, \mathbf{w})$ and $\mathcal{V}^U(\mathbf{x}(t), \mathbf{v}, \mathbf{w}) \subset \mathcal{A}^U(t, \mathbf{v}, \mathbf{w})$.

**Lemma 3.7.7.** *Suppose that Hypothesis 3.7.4 holds. Then, for any $(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$, Hypothesis 3.7.2 holds with $\boldsymbol{\phi} \equiv \mathbf{x}(\cdot, \mathbf{u}, \mathbf{x}_0)$ and the definitions* (3.53)*,* (3.54) *and* (3.55)*.*

*Proof.* Choose any $(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$ and define $\boldsymbol{\phi} \equiv \mathbf{x}(\cdot, \mathbf{u}, \mathbf{x}_0)$. Let $(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}}) \in I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ and suppose that $\hat{\mathbf{v}} \leq \boldsymbol{\phi}(\hat{t}) \leq \hat{\mathbf{w}}$ and either $\phi_i(\hat{t}) = \hat{v}_i$ or $\phi_i(\hat{t}) = \hat{w}_i$ for at least one $i \in \{1, \ldots, n_x\}$. Noting that $(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}})$ satisfies the required properties, let $\eta > 0$ be the constant satisfying Hypothesis 3.7.4.

Choose any $(\mathbf{v}, \mathbf{w}) \in B_\eta((\hat{\mathbf{v}}, \hat{\mathbf{w}}))$. For a.e. $t \in [\hat{t}, \hat{t} + \eta)$ such that $(t, \mathbf{v}, \mathbf{w}) \in D_\Pi$, the condition of Hypothesis 3.7.4 holds because $D_\Pi \subset D_\Omega$. Then $\mathcal{V}^L(\boldsymbol{\phi}(t), \mathbf{v}, \mathbf{w}) \subset \mathcal{A}^L(t, \mathbf{v}, \mathbf{w})$ and $\mathcal{V}^U(\boldsymbol{\phi}(t), \mathbf{v}, \mathbf{w}) \subset \mathcal{A}^U(t, \mathbf{v}, \mathbf{w})$, which is the desired result. □

In light of Theorem 3.7.4 and the previous two lemmas, the following result is now apparent.

**Theorem 3.7.8.** *Let $\mathbf{v}, \mathbf{w} \in \mathcal{AC}(I, \mathbb{R}^{n_x})$ satisfy*

(EX): *For every $t \in I$ and every index $i$,*

1. *$(t, \mathbf{v}(t), \mathbf{w}(t)) \in D_\Omega$,*
2. *$\Omega_i^L(t, \mathbf{v}(t), \mathbf{w}(t)) \subset D$ and $\Omega_i^U(t, \mathbf{v}(t), \mathbf{w}(t)) \subset D$.*

(IC): $\mathbf{v}(t_0) \leq \mathbf{x}_0 \leq \mathbf{w}(t_0)$*, $\forall \mathbf{x}_0 \in X_0$.*

(RHS): *For a.e. $t \in I$ and each index $i$,*

1. *If $s_i^L(t, \mathbf{v}(t), \mathbf{w}(t)) > 0$, then $\dot{v}_i(t) \leq f_i(t, \mathbf{p}, \mathbf{z})$ for all $\mathbf{p} \in U$ and $\mathbf{z} \in \Omega_i^L(t, \mathbf{v}(t), \mathbf{w}(t))$.*
2. *If $s_i^U(t, \mathbf{v}(t), \mathbf{w}(t)) > 0$, then $\dot{w}_i(t) \geq f_i(t, \mathbf{p}, \mathbf{z})$ for all $\mathbf{p} \in U$ and $\mathbf{z} \in \Omega_i^U(t, \mathbf{v}(t), \mathbf{w}(t))$.*

*If Hypothesis 3.7.3 holds and* $(\mathbf{v}, \mathbf{w})$ *has a finite number of transition times in* $I$*, then* $\mathbf{v}(t) \leq \mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \leq \mathbf{w}(t)$*,* $\forall (t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$*. If Hypotheses 3.7.4 holds, then the assumption of finitely many transition times can be relaxed and the conclusion remains true.*

### 3.7.4 Application to Convex Polyhedral a Priori Enclosures

Let $G \subset \mathbb{R}^{n_x}$ satisfy $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \in G$, $\forall (t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$. It was shown in §3.4 that the Hypothesis (RHSa) is not generally permissible in Theorem 3.3.2. One of the many problems caused by such a hypothesis is that the set over which the differential inequalities in (RHSa) must hold can potentially be empty. In this section, it is shown that this is not problematic for the important class of convex polyhedral *a priori* enclosures.

Assume that $G \equiv \{\mathbf{z} : \mathbf{A}\mathbf{z} \leq \mathbf{b}\}$, with $\mathbf{A} \in \mathbb{R}^{m \times n_x}$ and $\mathbf{b} \in \mathbb{R}^m$, and consider the definitions

$$\begin{aligned} &D_\Omega \equiv \{(t, \mathbf{v}, \mathbf{w}) \in I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} : G \cap [\mathbf{v}, \mathbf{w}] \neq \emptyset\}, \\ &\Omega_i^L(t, \mathbf{v}, \mathbf{w}) \equiv G \cap \mathcal{B}_i^L([\mathbf{v}, \mathbf{w}]), \\ &\Omega_i^U(t, \mathbf{v}, \mathbf{w}) \equiv G \cap \mathcal{B}_i^U([\mathbf{v}, \mathbf{w}]). \end{aligned} \tag{3.62}$$

To check the validity of these definitions via Hypothesis 3.7.3, define

$$s_i^L(t, \mathbf{v}, \mathbf{w}) = \begin{cases} 1 & \text{if} \quad \Omega_i^L(t, \mathbf{v}, \mathbf{w}) \neq \emptyset \\ -1 & \text{otherwise} \end{cases}, \tag{3.63}$$

$$s_i^U(t, \mathbf{v}, \mathbf{w}) = \begin{cases} 1 & \text{if} \quad \Omega_i^U(t, \mathbf{v}, \mathbf{w}) \neq \emptyset \\ -1 & \text{otherwise} \end{cases}, \tag{3.64}$$

for all $i \in \{1, \ldots, n_x\}$. Hypothesis 3.7.3 is established through the following three lemmas.

**Lemma 3.7.9.** *Let* $(t, \mathbf{v}, \mathbf{w}) \in D_\Omega$ *and* $\mathbf{z} \in G$*. If* $\mathcal{V}^L(\mathbf{z}, \mathbf{v}, \mathbf{w}) \cup \mathcal{V}^U(\mathbf{z}, \mathbf{v}, \mathbf{w}) \neq \emptyset$*, then* *one of the sets*

$$\mathcal{V}^L(\mathbf{z},\mathbf{v},\mathbf{w})\cap\mathcal{A}^L(t,\mathbf{v},\mathbf{w}) \quad \textit{or} \quad \mathcal{V}^U(\mathbf{z},\mathbf{v},\mathbf{w})\cap\mathcal{A}^U(t,\mathbf{v},\mathbf{w})$$

*is nonempty.*

*Proof.* Choose any $\hat{\mathbf{z}} \in G \cap [\mathbf{v},\mathbf{w}]$ and consider the line segment

$$\mathbf{l}(\lambda) = \hat{\mathbf{z}} + \lambda(\mathbf{z} - \hat{\mathbf{z}}), \quad \lambda \in [0,1].$$

First note that $\mathbf{l}(\lambda) \in G$ for all $\lambda \in [0,1]$ by convexity. Now, by definition of the sets $\mathcal{V}^L(\mathbf{z},\mathbf{v},\mathbf{w})$ and $\mathcal{V}^U(\mathbf{z},\mathbf{v},\mathbf{w})$,

$$\begin{aligned}
&i \in \mathcal{V}^L(\mathbf{z},\mathbf{v},\mathbf{w}) \implies \exists\lambda^i \ : \ l_i(\lambda) \geq v_i, \quad \forall\lambda \in [0,\lambda^i] \quad \text{and} \quad l_i(\lambda^i) = v_i,\\
&i \in \mathcal{V}^U(\mathbf{z},\mathbf{v},\mathbf{w}) \implies \exists\lambda^i \ : \ l_i(\lambda) \leq w_i, \quad \forall\lambda \in [0,\lambda^i] \quad \text{and} \quad l_i(\lambda^i) = w_i,\\
&i \notin \mathcal{V}^L(\mathbf{z},\mathbf{v},\mathbf{w}) \cup \mathcal{V}^U(\mathbf{z},\mathbf{v},\mathbf{w}) \implies l_i(\lambda) \in [v_i, w_i], \quad \forall\lambda \in [0,1].
\end{aligned}$$

Suppose $\mathcal{V}^L(\mathbf{z},\mathbf{v},\mathbf{w}) \cup \mathcal{V}^U(\mathbf{z},\mathbf{v},\mathbf{w}) \neq \emptyset$ and let $\lambda^* \equiv \min_{i\in(\mathcal{V}^L(\mathbf{z},\mathbf{v},\mathbf{w})\cup\mathcal{V}^U(\mathbf{z},\mathbf{v},\mathbf{w}))} \lambda^i$. Then $\mathbf{l}(\lambda^*) \in G \cap [\mathbf{v},\mathbf{w}]$ and $l_i(\lambda^*) = v_i$ (or $l_i(\lambda^*) = w_i$) for some $i \in \mathcal{V}^L(\mathbf{z},\mathbf{v},\mathbf{w})$ (or $i \in \mathcal{V}^U(\mathbf{z},\mathbf{v},\mathbf{w})$). For any such $i$, $\Omega_i^L(t,\mathbf{v},\mathbf{w}) \neq \emptyset$ and hence $i \in \mathcal{A}^L(t,\mathbf{v},\mathbf{w})$ (or $\Omega_i^U(t,\mathbf{v},\mathbf{w}) \neq \emptyset$ and hence $i \in \mathcal{A}^U(t,\mathbf{v},\mathbf{w})$). □

**Lemma 3.7.10.** *Let* $(t,\mathbf{v},\mathbf{w}) \in D_\Omega$ *and define*

$$[\mathbf{v},\mathbf{w}]_{\mathcal{A}} \equiv \{\mathbf{z} \in \mathbb{R}^{n_x} : z_i \geq v_i, \forall i \in \mathcal{A}^L(t,\mathbf{v},\mathbf{w}), z_i \leq w_i, \forall i \in \mathcal{A}^U(t,\mathbf{v},\mathbf{w})\}. \quad (3.65)$$

*Then* $G \cap [\mathbf{v},\mathbf{w}] = G \cap [\mathbf{v},\mathbf{w}]_{\mathcal{A}}$.

*Proof.* It is clear that $(G \cap [\mathbf{v},\mathbf{w}]) \subset (G \cap [\mathbf{v},\mathbf{w}]_{\mathcal{A}})$. Suppose that the conclusion is false and choose any $\mathbf{z} \in G \cap [\mathbf{v},\mathbf{w}]_{\mathcal{A}}$ such that $\mathbf{z} \notin G \cap [\mathbf{v},\mathbf{w}]$. By this choice of $\mathbf{z}$, one of the sets $\mathcal{V}^L(\mathbf{z},\mathbf{v},\mathbf{w})$ or $\mathcal{V}^U(\mathbf{z},\mathbf{v},\mathbf{w})$ is nonempty, in which case Lemma 3.7.9 shows that either $\mathcal{V}^L(\mathbf{z},\mathbf{v},\mathbf{w}) \cap \mathcal{A}^L(t,\mathbf{v},\mathbf{w})$ or $\mathcal{V}^U(\mathbf{z},\mathbf{v},\mathbf{w}) \cap \mathcal{A}^U(t,\mathbf{v},\mathbf{w})$ is nonempty. This, however, implies that $\mathbf{z} \notin [\mathbf{v},\mathbf{w}]_{\mathcal{A}}$, which is a contradiction. □

**Lemma 3.7.11.** *Choose any* $(\mathbf{u},\mathbf{x}_0) \in \mathcal{U} \times X_0$ *and define* $\mathbf{x} \equiv \mathbf{x}(\cdot,\mathbf{u},\mathbf{x}_0)$. *There exists* $L > 0$ *such that the following conditions hold for every* $(t,\mathbf{v},\mathbf{w}) \in D_\Omega$*:*

1. *If* $i \in \mathcal{Q}^L(t,\mathbf{v},\mathbf{w})$, *then* $\exists \mathbf{z} \in \Omega_i^L(t,\mathbf{v},\mathbf{w})$ *such that*

$$\|\mathbf{x}(t)-\mathbf{z}\|_\infty \le L\max\left(\max_{i\in\mathcal{Q}^L(t,\mathbf{v},\mathbf{w})}(v_i - x_i(t)), \max_{i\in\mathcal{Q}^U(t,\mathbf{v},\mathbf{w})}(x_i(t)-w_i)\right), \quad (3.66)$$

   *where* $\mathcal{Q}^L$ *and* $\mathcal{Q}^U$ *are defined as in Hypothesis 3.7.3.*

2. *If* $i \in \mathcal{Q}^U(t,\mathbf{v},\mathbf{w})$, *then* $\exists \mathbf{z} \in \Omega_i^U(t,\mathbf{v},\mathbf{w})$ *such that* (3.66) *holds.*

*Proof.* Let $(t,\mathbf{v},\mathbf{w}) \in D_\Omega$ and suppose that $i \in \mathcal{Q}^L(t,\mathbf{v},\mathbf{w})$. Define the sets

$$\begin{aligned} M_1 &\equiv \{\mathbf{z}\in G : z_j \ge v_j,\ \forall j \in \mathcal{A}^L(t,\mathbf{v},\mathbf{w}), \quad z_j \le w_j,\ \forall j\in\mathcal{A}^U(t,\mathbf{v},\mathbf{w}), \quad z_i = v_i\}, \\ M_2 &\equiv \{\mathbf{z}\in G : z_j \ge \min(x_j(t),v_j),\ \forall j\in\mathcal{A}^L(t,\mathbf{v},\mathbf{w}), \\ &\qquad\qquad z_j \le \max(x_j(t),w_j),\ \forall j \in \mathcal{A}^U(t,\mathbf{v},\mathbf{w}), \quad z_i = \min(x_i(t),v_i)\}. \end{aligned}$$

Note that $\min(\mathbf{x}(t),\mathbf{v}) \le \mathbf{x}(t) \le \max(\mathbf{x}(t),\mathbf{w})$. Further, $i \in \mathcal{Q}^L(t,\mathbf{v},\mathbf{w})$ implies that $x_i(t) < v_i$ and hence $x_i(t) = \min(x_i(t),v_i)$. It follows that $\mathbf{x}(t) \in M_2$. Because $(t,\mathbf{v},\mathbf{w}) \in D_\Omega$, $G \cap [\mathbf{v},\mathbf{w}] \neq \emptyset$. Furthermore, $i \in \mathcal{Q}^L(t,\mathbf{v},\mathbf{w})$ implies that $G \cap \mathcal{B}_i^L([\mathbf{v},\mathbf{w}]) \neq \emptyset$, so that $M_1 \neq \emptyset$. Then, $M_1$ and $M_2$ are systems of linear inequalities which differ only in their right-hand side data, and both are nonempty. Using Theorem 3.6.4, this implies that there exists $L > 0$ satisfying

$$\begin{aligned} d_H(M_1,M_2) &\le L\max\left(\max_{i\in\mathcal{A}^L(t,\mathbf{v},\mathbf{w})}|v_i - \min(x_i(t),v_i)|, \max_{i\in\mathcal{A}^U(t,\mathbf{v},\mathbf{w})}|\max(x_i(t),w_i)-w_i|\right), \\ &\le L\max\left(\max_{i\in\mathcal{A}^L(t,\mathbf{v},\mathbf{w})}|\max(v_i - x_i(t),0)|, \max_{i\in\mathcal{A}^U(t,\mathbf{v},\mathbf{w})}|\max(x_i(t)-w_i,0)|\right), \\ &= L\max\left(\max_{i\in\mathcal{Q}^L(t,\mathbf{v},\mathbf{w})}(v_i - x_i(t)), \max_{i\in\mathcal{Q}^U(t,\mathbf{v},\mathbf{w})}(x_i(t)-w_i)\right). \end{aligned}$$

Since $\mathbf{x}(t) \in M_2$, the definition of the Hausdorff metric implies that there exists

$\mathbf{z} \in M_1$ such that

$$\|\mathbf{x}(t) - \mathbf{z}\|_\infty \leq L \max\left(\max_{i \in \mathcal{Q}^L(t,\mathbf{v},\mathbf{w})}(v_i - x_i(t)), \max_{i \in \mathcal{Q}^U(t,\mathbf{v},\mathbf{w})}(x_i(t) - w_i)\right). \tag{3.67}$$

But, by Lemma 3.7.10, $M_1 = \{\mathbf{z} \in G \cap [\mathbf{v},\mathbf{w}]_\mathcal{A} : z_i = v_i\} = \{\mathbf{z} \in G \cap [\mathbf{v},\mathbf{w}] : z_i = v_i\} = \Omega_i^L(t,\mathbf{v},\mathbf{w})$, and hence $\mathbf{z} \in \Omega_i^L(t,\mathbf{v},\mathbf{w})$. This proves Conclusion 1, and Conclusion 2 follows from an analogous argument. □

The previous lemmas imply the following comparison theorem for convex polyhedral *a priori* enclosures.

**Theorem 3.7.12.** *Let* $\mathbf{v},\mathbf{w} \in \mathcal{AC}(I,\mathbb{R}^{n_x})$ *satisfy*

(EX): *For every* $t \in I$ *and every index* $i$*,*

1. $G \cap [\mathbf{v}(t),\mathbf{w}(t)] \neq \emptyset$*,*
2. $G \cap \mathcal{B}_i^L(\mathbf{v}(t),\mathbf{w}(t)) \subset D$ *and* $G \cap \mathcal{B}_i^U(\mathbf{v}(t),\mathbf{w}(t)) \subset D$*.*

(IC): $\mathbf{v}(t_0) \leq \mathbf{x}_0 \leq \mathbf{w}(t_0)$*,* $\forall \mathbf{x}_0 \in X_0$*.*

(RHS): *For a.e.* $t \in I$ *and each index* $i$*,*

1. $\dot{v}_i(t) \leq f_i(t,\mathbf{p},\mathbf{z})$ *for all* $\mathbf{p} \in U$ *and* $\mathbf{z} \in G \cap \mathcal{B}_i^L(\mathbf{v}(t),\mathbf{w}(t))$*.*
2. $\dot{w}_i(t) \geq f_i(t,\mathbf{p},\mathbf{z})$ *for all* $\mathbf{p} \in U$ *and* $\mathbf{z} \in G \cap \mathcal{B}_i^U(\mathbf{v}(t),\mathbf{w}(t))$*.*

*If* $(\mathbf{v},\mathbf{w})$ *has finitely many transition times in* $I$*, then* $\mathbf{v}(t) \leq \mathbf{x}(t,\mathbf{u},\mathbf{x}_0) \leq \mathbf{w}(t)$*,* $\forall (t,\mathbf{u},\mathbf{x}_0) \in I \times \mathcal{U} \times X_0$*.*

*Proof.* By Theorem 3.7.8, it suffices to show that Hypothesis 3.7.3 holds. Let $(\hat{t},\hat{\mathbf{v}},\hat{\mathbf{w}}) \in I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ and suppose that $\exists(\mathbf{u},\mathbf{x}_0) \in \mathcal{U} \times X_0$ such that $\mathbf{x} \equiv \mathbf{x}(\cdot,\mathbf{u},\mathbf{x}_0)$ satisfies $\hat{\mathbf{v}} \leq \mathbf{x}(\hat{t}) \leq \hat{\mathbf{w}}$ and either $x_i(\hat{t}) = \hat{v}_i$ or $x_i(\hat{t}) = \hat{w}_i$ for at least one $i \in \{1,\ldots,n_x\}$. Choose an arbitrary $\eta > 0$ and let $L > 0$ be the constant of Lemma 3.7.11. It will be shown that Hypothesis 3.7.3 holds with these definitions.

Choose any $(\mathbf{v},\mathbf{w}) \in B_\eta((\hat{\mathbf{v}},\hat{\mathbf{w}}))$ and any $t \in [\hat{t},\hat{t}+\eta)$ such that $(t,\mathbf{v},\mathbf{w}) \in D_\Omega$. Condition 1 of Hypothesis 3.7.3 follows by applying Lemma 3.7.9 with $\mathbf{z} \equiv \mathbf{x}(t)$. Condition 2 follows by applying Lemma 3.7.11. □

### 3.7.5 Application to Interval a priori Enclosures

As a second application of the theory in this section, we prove a comparison theorem involving a time-varying interval *a priori* enclosure.

**Theorem 3.7.13.** *Let $X : I \to \mathbb{IR}^{n_x}$ satisfy $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \in X(t) \equiv [\mathbf{x}^L(t), \mathbf{x}^U(t)]$, $\forall (t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$, and assume that $D$ is open. Let $\mathbf{v}, \mathbf{w} \in \mathcal{AC}(I, \mathbb{R}^{n_x})$ satisfy*

(EX): $X(t) \cap [\mathbf{v}(t), \mathbf{w}(t)] \neq \emptyset$, $\forall t \in I$.

(IC): $\mathbf{v}(t_0) \leq \mathbf{x}_0 \leq \mathbf{w}(t_0)$, $\forall \mathbf{x}_0 \in X_0$.

(RHS): *For a.e. $t \in I$ and each index $i$,*

1. *If $v_i(t) > x_i^L(t)$, then $\dot{v}_i(t) \leq f_i(t, \mathbf{p}, \mathbf{z})$ for all $\mathbf{p} \in U$ and $\mathbf{z} \in D \cap X(t) \cap \mathcal{B}_i^L([\mathbf{v}(t), \mathbf{w}(t)])$,*
2. *If $w_i(t) < x_i^U(t)$, then $\dot{w}_i(t) \geq f_i(t, \mathbf{p}, \mathbf{z})$ for all $\mathbf{p} \in U$ and $\mathbf{z} \in D \cap X(t) \cap \mathcal{B}_i^U([\mathbf{v}(t), \mathbf{w}(t)])$.*

*Then $\mathbf{v}(t) \leq \mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \leq \mathbf{w}(t)$, $\forall (t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$.*

*Proof.* By Theorem 3.7.8, it suffices to show that Hypotheses 3.7.4 and 3.7.3 hold with the definitions

$$\begin{aligned}
&D_\Omega \equiv \{(t, \mathbf{v}, \mathbf{w}) \in I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} : X(t) \cap [\mathbf{v}, \mathbf{w}] \neq \emptyset\}, \qquad (3.68)\\
&\Omega_i^L(t, \mathbf{v}, \mathbf{w}) \equiv D \cap X(t) \cap \mathcal{B}_i^L([\mathbf{v}, \mathbf{w}]),\\
&\Omega_i^U(t, \mathbf{v}, \mathbf{w}) \equiv D \cap X(t) \cap \mathcal{B}_i^U([\mathbf{v}, \mathbf{w}]),\\
&s_i^L(t, \mathbf{v}, \mathbf{w}) \equiv v_i - x_i^L(t),\\
&s_i^U(t, \mathbf{v}, \mathbf{w}) \equiv x_i^U(t) - w_i.
\end{aligned}$$

Consider Hypothesis 3.7.4 first. Let $(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}}) \in I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ and suppose that there exists $(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$ such that $\mathbf{x} \equiv \mathbf{x}(\cdot, \mathbf{u}, \mathbf{x}_0)$ satisfies $\hat{\mathbf{v}} \leq \mathbf{x}(\hat{t}) \leq \hat{\mathbf{w}}$ and either $x_i(\hat{t}) = \hat{v}_i$ or $x_i(\hat{t}) = \hat{w}_i$ for at least one $i \in \{1, \ldots, n_x\}$. Choose any $\eta > 0$, any $(\mathbf{v}, \mathbf{w}) \in B_\eta((\hat{\mathbf{v}}, \hat{\mathbf{w}}))$ and any $t \in [\hat{t}, \hat{t} + \eta)$ such that $(t, \mathbf{v}, \mathbf{w}) \in D_\Omega$.

If $i \in \mathcal{V}^L(\mathbf{x}(t), \mathbf{v}, \mathbf{w})$, then $x_i(t) < v_i$ by definition. But then $s_i^L(t, \mathbf{v}, \mathbf{w}) > 0$, so that $i \in \mathcal{A}^L(t, \mathbf{v}, \mathbf{w})$. Applying and analogous argument for $i \in \mathcal{V}^U(\mathbf{x}(t), \mathbf{v}, \mathbf{w})$, this establishes Hypothesis 3.7.2.

To show Hypothesis 3.7.3, consider $(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}})$ as above. Making the definition $\mathbf{z}(t, \mathbf{v}, \mathbf{w}) \equiv \text{mid}(\mathbf{v}, \mathbf{w}, \mathbf{x}(t))$, choose $\eta > 0$ so small that $\mathbf{z}(t, \mathbf{v}, \mathbf{w}) \in B_\eta(\mathbf{x}(\hat{t})) \subset D$, for all $(\mathbf{v}, \mathbf{w}) \in B_\eta((\hat{\mathbf{v}}, \hat{\mathbf{w}}))$ and $t \in [\hat{t}, \hat{t} + \eta)$. Choose any $(\mathbf{v}, \mathbf{w}) \in B_\eta((\hat{\mathbf{v}}, \hat{\mathbf{w}}))$, any $t \in [\hat{t}, \hat{t} + \eta)$ such that $(t, \mathbf{v}, \mathbf{w}) \in D_\Omega$. Condition 1 of Hypothesis 3.7.3 follows immediately from Hypothesis 3.7.2. To show Condition 2, choose any $i \in \mathcal{Q}^L(t, \mathbf{v}, \mathbf{w})$. Since $x_i(t) < v_i \leq w_i$, $\mathbf{z}(t, \mathbf{v}, \mathbf{w}) \in \mathcal{B}_i^L([\mathbf{v}, \mathbf{w}])$. By the choice of $\eta$, $\mathbf{z}(t, \mathbf{v}, \mathbf{w}) \in D$. To show that $\mathbf{z}(t, \mathbf{v}, \mathbf{w}) \in X(t)$ as well, choose any $j$. If $z_j(t, \mathbf{v}, \mathbf{w}) = x_j(t)$ then $z_j(t, \mathbf{v}, \mathbf{w}) \in X_j(t)$ by definition. If $z_j(t, \mathbf{v}, \mathbf{w}) = w_j$, then $v_j \leq w_j \leq x_j(t) \leq x_j^U(t)$ and it follows from the fact that $[v_j, w_j] \cap X_j(t) \neq \emptyset$ that $z_j(t, \mathbf{v}, \mathbf{w}) \in X_j(t)$. Using an analogous argument for the case $z_j(t, \mathbf{v}, \mathbf{w}) = v_j$, it follows that $\mathbf{z}(t, \mathbf{v}, \mathbf{w}) \in \Omega_i^L(t, \mathbf{v}, \mathbf{w})$. Now, by the definition of $\mathbf{z}$, it follows that

$$\|\mathbf{x}(t) - \mathbf{z}\|_\infty \leq \max(\|\max(\mathbf{0}, \mathbf{v} - \mathbf{x}(t))\|_\infty, \|\max(\mathbf{0}, \mathbf{x}(t) - \mathbf{w}\|_\infty). \tag{3.69}$$

Applying Hypothesis 3.7.2, this is exactly

$$\|\mathbf{x}(t) - \mathbf{z}\|_\infty \leq \max\left(\max_{i \in \mathcal{Q}^L(t,\mathbf{v},\mathbf{w})}(v_i - x_i(t)), \max_{i \in \mathcal{Q}^U(t,\mathbf{v},\mathbf{w})}(x_i(t) - w_i)\right). \tag{3.70}$$

Thus, Condition 2 of Hypothesis 3.7.3 holds, and Condition 3 is proven analogously. □

## 3.8 Conclusions and Future Work

In this chapter, the problem of efficiently computing interval bounds on the solutions of parametric ODEs and control systems was considered. In particular, the use of known *a priori* enclosures, derived from physical information, was investigated as a means to enhance the performance of interval methods based on differential inequalities, while maintaining the ability to use efficient interval computations. Toward this end, a general comparison theorem was established in which the use of the *a priori* enclosure is abstracted in terms of set-valued mappings which are required to satisfy several key conditions. From these conditions, the basic requirements that an interval refinement operation $\mathcal{I}_G$ (based on an arbitrary *a priori* enclosure $G$) must satisfy in order to result in a valid bounding method were derived. An appropriate definition of this operation was given for the case when $G$ is a convex polyhedron, resulting in a novel computational method. This method is demonstrated for several numerical examples in the next chapter.

When $G$ is not a convex polyhedron, the framework of Section 3.6.1 still applies, but no valid definition of $\mathcal{I}_G$ is currently available. The use of interval Newton methods and constraint propagation techniques are promising in this regard and warrant future investigation. In addition to interval-based methods, the general comparison theorem derived here also suggests other approaches, such as the method of §3.6.3 using linear programming relaxations. This method can potentially describe sharper bounds than an interval-based method and also warrants further investigation into an efficient computational implementation.

In §3.7, some of the key restrictions imposed on the general comparison theorem of §3.5 were further relaxed. This analysis leads to some interesting results suggesting that sharper bounds could be obtained, at least in the case of convex polyhedral *a priori* enclosures. However, this theory is also lacking an efficient computational implementation. As opposed to the developments in §3.6, where state bounds could be described as the solutions of a system of ODEs, it seems that the state bounds derived in §3.7 would be more naturally described as the solutions of a hybrid system. At present, it is not clear whether this additional complexity would be justified by the resulting improvement in the bounds.

# Chapter 4

# Bounding the Solutions of Chemical Kinetics Models

## 4.1 Introduction

The previous chapter introduced several new methods for computing interval bounds on the solutions of parametric ODEs and control systems. In particular, these methods are able to use efficiently known physical information about the solutions of such systems as a means to reduce conservatism in the computed bounds. In this chapter, these methods are applied to ODE models of chemical reaction kinetics. Such models are very important in chemical engineering applications and are commonly cited as a primary motivation for state bounding methods [164, 135, 105] and related algorithms [163, 103, 106, 85, 164, 104, 37, 36, 123]. It will be shown that very rich physical information about the solutions of chemical kinetics models is available through a relatively simple analysis of the stoichiometry matrix. In particular, the solutions often obey affine reaction invariants, which are closely related to the notion of exact model reduction [63, 66, 181]. Through numerous examples, it is shown that using this information in conjunction with the bounding methods of the previous chapter results in state bounds that are substantially tighter than those computed by a similar methods that cannot make use of this physical information (i.e., Harrison's method). Moreover, this improvement is achieved at a small additional cost, making these methods appropriate for the class of problems that we set out to address in §3.1.

## 4.2 Physical Information in Reaction Models

Chemical reaction kinetics are most commonly modeled by a coupled system of ODEs [9] of the form

$$\dot{\mathbf{x}}(t, \mathbf{u}, \mathbf{x}_0) = \mathbf{S}\mathbf{r}(t, \mathbf{u}(t), \mathbf{x}(t, \mathbf{u}, \mathbf{x}_0)), \quad \mathbf{x}(t_0, \mathbf{u}, \mathbf{x}_0) = \mathbf{x}_0. \tag{4.1}$$

The state variables $\mathbf{x}$ represent the concentrations of chemical species, the *rate functions* $\mathbf{r} : I \times U \times D \to \mathbb{R}^{n_r}$ describe the rates of all possible reactions between species, and the *stoichiometry matrix* $\mathbf{S} \in \mathbb{R}^{n_x \times n_r}$ encodes the proportionalities by which the concentration of each species is effected by the occurence of each reaction. A simple model of this type has already been studied in Example 3.3.1. Throughout this chapter, it is assumed that Assumptions 3.2.1 and 3.2.2 hold with $\mathbf{f} = \mathbf{r}$. In this case, it is simple to show that (4.1) satisfies the requirements of §3.2, so that the bounding methods of Chapter 3 can be applied.

Information about the solutions of a chemical kinetics models is available in the form of *affine reaction invariants* and *natural bounds*. Both are obtained by a simple analysis of the stoichiometry matrix. To be clear, this information is available *prior* to the application of a differential inequalities bounding method of the type described in the previous chapter. The combination of natural bounds and reaction invariants often constitutes a massive restriction on the region of state space in which solutions potentially lie. This is demonstrated for the model of Example 3.3.1 below. Thus, the impact of leveraging such information in a state bounding method is potentially very significant. Of course, models of the form (4.1) are not restricted to chemical kinetics, so the ideas here likely apply in other important application areas as well (e.g., electrical circuit models [174]).

### 4.2.1 Affine Reaction Invariants

An affine reaction invariant is a linear combination of the state variables $\mathbf{x}$ which does not change as $\mathbf{x}$ evolves in time [181]. That is, a vector $\mathbf{m} \in \mathbb{R}^{n_x}$ is an affine reaction invariant if

$$\mathbf{m}^{\mathrm{T}}\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) = \mathbf{m}^{\mathrm{T}}\mathbf{x}_0, \quad \forall (t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0. \tag{4.2}$$

It is easily seen from (4.1) that the reaction invariants of a kinetic model include every vector $\mathbf{m}$ which lies in the left null space of the stoichiometry matrix, $\mathcal{N}(\mathbf{S}^{\mathrm{T}})$, since $\mathbf{m}^{\mathrm{T}}\dot{\mathbf{x}}(t, \mathbf{u}, \mathbf{x}_0) = \mathbf{m}^{\mathrm{T}}\mathbf{S}\mathbf{r}(t, \mathbf{u}(t), \mathbf{x}(t, \mathbf{u}, \mathbf{x}_0)) = 0$. Additional affine reaction invariants can exist if the range of the rate function $\mathbf{r}$ has dimension less than $n_r$ [60]. However, this is a kinetic phenomena, not a stoichiometric one, and it is difficult in practice to identify and make use of such invariants, so these are not considered here.

Since every vector in $\mathcal{N}(\mathbf{S}^{\mathrm{T}})$ must be an affine reaction invariant, it is clear that every kinetic model for which $\mathbf{S}$ is not full row rank must have at least one affine reaction invariant. In fact, the number of linearly independent affine reaction invariants is equal to the dimension of $\mathcal{N}(\mathbf{S}^{\mathrm{T}})$. Kinetic models very often have stoichiometry matrices which are not full row rank because of conservation laws which are implicit in the model, such as overall mass and atomic balances [63, 66, 181]. This is particularly true of models of biological reaction networks [58].

A basis for $\mathcal{N}(\mathbf{S}^{\mathrm{T}})$ provides a complete set of linearly independent reaction invariants. Such a basis is easily obtained from the singular value decomposition of $\mathbf{S}^{\mathrm{T}}$, or directly using the `MATLAB` routine `null`. Throughout, we denote by $\mathbf{M} \in \mathbb{R}^{m \times n_x}$ a matrix whose rows form a basis of $\mathcal{N}(\mathbf{S}^{\mathrm{T}})$, so that

$$\mathbf{M}\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) = \mathbf{M}\mathbf{x}_0, \quad \forall (t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0. \tag{4.3}$$

### 4.2.2 Natural Bounds

In addition to reaction invariants, physical considerations very often suggest a crude interval $X^N = [\mathbf{x}^{N,L}, \mathbf{x}^{N,U}]$ such that

$$\mathbf{x}^{N,L} \leq \mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \leq \mathbf{x}^{N,U}, \quad \forall (t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0. \tag{4.4}$$

These bounds are referred to as *natural bounds* [162], and arise from a number of considerations. Though it is not clear from (4.1), the solutions of chemical kinetics models are always nonnegative, provided that the initial conditions are nonnegative. Physically, this is because they represent concentrations of chemical species. Mathematically, it is because rate functions (i.e., components of $\mathbf{r}$) that act to decrease the concentration of some chemical species, $x_i$, are always zero if $x_i$ is zero. Combined with nonnegative initial conditions, this implies that $\mathbf{x}$ is nonnegative [13].

Other natural bounds may be implied by the directionality of reactions. If some of the reactions in a given reaction network are not reversible, the flow of mass or atomic elements throughout the network is hindered. For example, if a particular reactant is consumed but never generated, then a natural upper bound is given by its initial concentration.

Finally, additional natural bounds may be implied by conservation laws. If, for example, the volume and number of molecules in a reacting system is conserved, then the maximum concentration of any given species is bounded by the total concentration at the initial time. Bounds of this type are actually nothing more than the effects of nonnegativity constraints on other species, acting through an affine reaction invariant. In particular, the bounds $X^N$, which may at first contain only nonnegativity constraints, may always be refined by solving the programs

$$\begin{aligned} x_i^{N,L} \quad &:= \quad \inf_{\mathbf{z},\mathbf{x}_0} z_i && (4.5)\\ &\text{s.t.} \quad \mathbf{M}(\mathbf{z}-\mathbf{x}_0)=\mathbf{0} \\ &\qquad\quad \mathbf{z}\in X^N, \quad \mathbf{x}_0\in X_0 \\ x_i^{N,U} \quad &:= \quad \sup_{\mathbf{z},\mathbf{x}_0} z_i && (4.6)\\ &\text{s.t.} \quad \mathbf{M}(\mathbf{z}-\mathbf{x}_0)=\mathbf{0} \\ &\qquad\quad \mathbf{z}\in X^N, \quad \mathbf{x}_0\in X_0 \end{aligned}$$

for each $i = 1, \ldots, n_x$. If $X_0$ is a convex polyhedral set, then these are simple linear programs. Thus, natural bounds arising from complex stoichiometric relationships between species can be easily computed using the matrix $\mathbf{M}$. However, it should be noted that information based on the directionality of reactions is not contained in $\mathbf{S}$ (and hence in $\mathbf{M}$) and therefore cannot be ascertained by solving the linear programs above. Such observations should be included in the initial set of natural bounds, prior to refinement through (4.5) and (4.6).

**Example 4.2.1.** Consider again the reversible chemical reaction

$$\mathrm{A} + \mathrm{B} \rightleftharpoons \mathrm{C} \tag{4.7}$$

first considered in Example 3.3.1. Recall that the time evolution of the species concentrations $x_A$, $x_B$ and $x_C$ in an isothermal batch reactor is described by a kinetic model of the form (4.1) with $\mathbf{x} \equiv [x_A \ \ x_B \ \ x_C]^{\mathrm{T}}$, $\mathbf{u} \equiv [k_f \ \ k_r]^{\mathrm{T}}$, and

$$\mathbf{S} \equiv \begin{bmatrix} -1 & 1 \\ -1 & 1 \\ 1 & -1 \end{bmatrix}, \quad \mathbf{r}(t, \mathbf{u}, \mathbf{z}) \equiv \begin{bmatrix} k_f z_A z_B \\ k_r z_C \end{bmatrix}.$$

The stoichiometry matrix has dimension $3 \times 2$, yet is only rank 1, so the model must have two linearly independent reaction invariants. Computing the null space of $\mathbf{S}^{\mathrm{T}}$

using the MATLAB routine null gives the basis vectors

$$\mathbf{M} = \begin{bmatrix} -0.8165 & 0.40825 & -0.40825 \\ 0.0 & -0.7071 & -0.7071 \end{bmatrix}.$$

Though it is not necessary for further computations, one can obtain more physically meaningful reaction invariants through elementary row operations. Here, we find the vectors $\mathbf{m}_1^{\mathrm{T}} = [1\ 1\ 2]$ and $\mathbf{m}_2^{\mathrm{T}} = [1\ \ -1\ \ 0]$, so that

$$\begin{aligned} x_A + x_B + 2x_C &= x_{0,A} + x_{0,B} + 2x_{0,C}, \\ x_A - x_B &= x_{0,A} - x_{0,B}. \end{aligned} \tag{4.8}$$

Physically, the first invariant represents the overall mass balance for the system, while the second represents the proportionality between the species $A$ and $B$, which is maintained because they react with 1-to-1 stoichiometry.

The intersection of the subspaces of $\mathbb{R}^{n_x}$ with normals $\mathbf{m}_1$ and $\mathbf{m}_2$, translated by the initial condition vector, contain all points in $\mathbb{R}^{n_x}$ which satisfy the two invariants above, respectively. These planes, restricted to the positive orthant, are shown in Figure 4-1. From this it is determined that the only possible solutions of the kinetic model must lie in the intersection of these two planes, regardless of $(t, \mathbf{u}) \in I \times \mathcal{U}$. The direction $y$ depicted in the figure is a linear combination of $x_A$, $x_B$ and $x_C$ along which the solution vector $\mathbf{x}$ evolves in time as the reaction proceeds. We will have more to say about this coordinate in §4.4.

The meaning of the programs (4.5) and (4.6) is easily seen from Figure 4-1. Clearly, the combination of nonnegativity constraints and the plotted reaction invariants implies the natural bounds $X^N = [1, 1.5] \times [0, 0.5] \times [0, 0.5]$.

### 4.2.3 A Polyhedral a Priori Enclosure

Consider computing state bounds for (4.1) with given sets of admissible initial conditions and controls, $X_0$ and $\mathcal{U}$, respectively (see §3.2). If $X_0$ is an interval, then affine reaction invariants and natural bounds can be combined to give an *a priori* enclosure

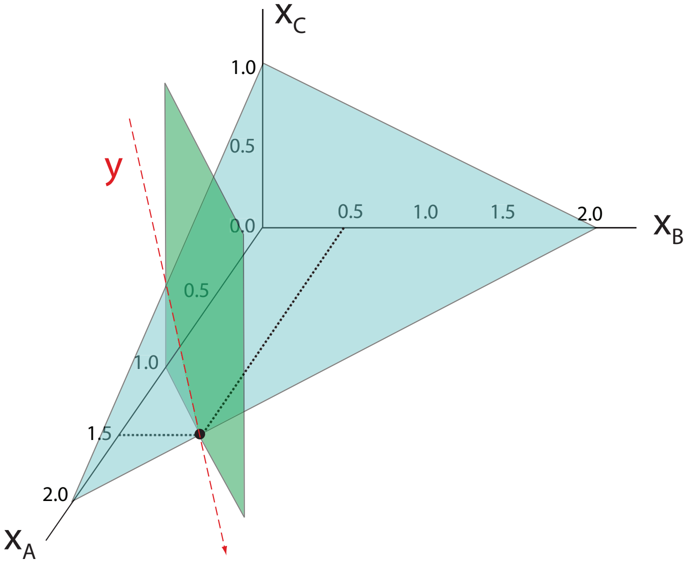

Figure 4-1: Planes in $\mathbb{R}^{n_x}$ containing all points in the positive orthant which satisfy the affine reaction invariants (4.8). All solutions lie on the intersection of these two planes, for all $(t, \mathbf{p}) \in I \times P$. The point at $(1.5, 0.5, 0)$ represents the initial condition. The direction $y$ is orthogonal to the normals of both planes, $\mathbf{m}_1$ and $\mathbf{m}_2$, and demonstrates an axis along which the time evolution of the reaction can be fully described (see Section 4.4).

for the solutions of (4.1) of the form

$$G \equiv \{\mathbf{z} \in \mathbb{R}^{n_x} : \mathbf{z} \in X^N,\ \mathbf{M}\mathbf{z} \in \mathbf{M}X_0\}. \tag{4.9}$$

Expanding the interval multiplication $\mathbf{M}X_0$, $G$ can be written as a convex polyhedral set in standard form, $G = \{\mathbf{z} \in \mathbb{R}^{n_x} : \mathbf{A}\mathbf{z} \leq \mathbf{b}\}$. Thus, the bounding methods of §3.6 are applicable. From Figure 4-1, it is evident that $G$ can potentially put a large restriction on the regions of state space that must be considered when computing state bounds. In the following section, the method of §3.6.2 is applied to three examples and shown to make very effective use of this restriction.

## 4.3 Numerical Examples

All numerical experiments in this section were performed on a Dell Precision T3400 workstation with a 2.83 GHz Intel Core2 Quad CPU. One core and 512 MB of memory were dedicated to each job. Numerical integration was carried out using the software `CVODE` [44] with absolute and relative tolerances of $10^{-5}$. Interval extensions were computed automatically using the library `MC++` (`http://www3.imperial.ac.uk/people/b.chachuat/research`). `MC++` is the successor of `libMC`, which is described in detail in [122].

For ease of comparison, recall that Harrison's method refers to the state bounding method given by solving (3.31) with the definitions (3.33). The method given by solving (3.31) with (3.35), $\mathcal{I}_G$ as in Definition 3.6.3 and $G$ a convex polyhedral set will be called the full-space invariant (FSI) method. Later, this will be contrasted with the so-called reduced-space methods developed in §4.4. In the special case where $G \equiv X^N$, the FSI method will be called Singer's method, after the author of [162]. Note, however, that this may not be identically the implementation used in [162], as discussed in §3.6.4.

**Example 4.3.1.** Consider again the model of Example 3.3.1. There, state bounds were computed using Harrison's method with $I = [0, 0.05]$ min, the set of admissible controls

$$\mathcal{U} = \{(k_f, k_r) \in (L^1(I))^2 : (k_f(t), k_r(t)) \in [100, 500] \times [0.001, 0.01] \text{ for a.e. } t \in I\},$$

and the singleton set of admissible initial conditions $X_0 = \{\mathbf{x}_0\}$ with $\mathbf{x}_0 = (1.5, 0.5, 0)$ (M). Here, state bounds are computed by Singer's method and the FSI method and compared. For ease of reference, the results of Harrison's method are reproduced as the dashed curves in Figure 4-2 below. In all figures in this section, solid curves represent true model solutions computed for sampled points $(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$.

In Example 4.2.1, it was shown that this model has two reaction invariants,

$$\mathbf{M} = \begin{bmatrix} -0.8165 & 0.40825 & -0.40825 \\ 0.0 & -0.7071 & -0.7071 \end{bmatrix},$$

and natural bounds $X^N = [1, 1.5] \times [0, 0.5] \times [0, 0.5]$. Given this information, the results of Singer's method are shown by the crosses in Figure 4-2. The blue circles are the results of the FSI method applied with

$$G \equiv \{\mathbf{z} \in \mathbb{R}^3 : \mathbf{M}\mathbf{z} = \mathbf{M}\mathbf{x}_0,\ \mathbf{z} \in X^N\}. \tag{4.10}$$

From Figure 4-2, it is clear that both methods provide much more reasonable bounds than Harrison's method, which does not make use of any physical information. However, while Singer's method prevents divergence of the upper bound, it still fails to provide an accurate enclosure of the model solutions throughout time. On the other hand, the state bounds computed using the FSI method are exact for this problem. That is, it is possible to realize the bounding trajectories with true model solutions.

Recall from Example 3.3.1 that Harrison's method produces bounds in $1.8 \times 10^{-4}$s, while integration of a single trajectory requires $1.1 \times 10^{-4}$s. Singer's method requires $3.1 \times 10^{-4}$s, while the FSI method requires $1.72 \times 10^{-3}$s. Thus, using natural bounds nearly doubles the cost of Harrison's method and produces bounds that are reasonable but weak, while obtaining exact bounds using natural bounds and reaction invariants

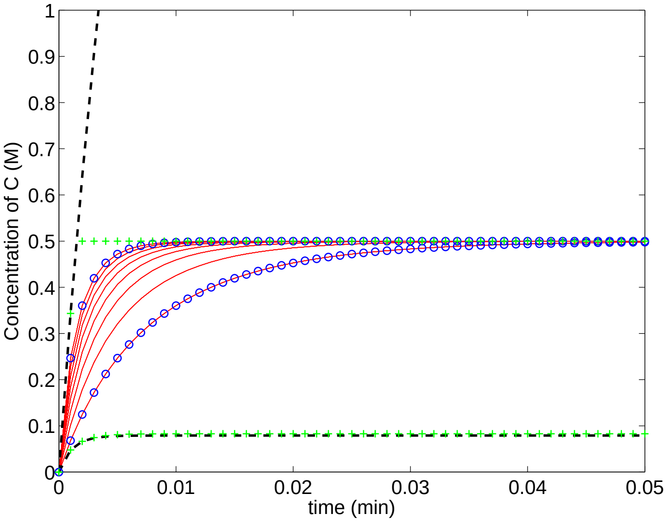

Figure 4-2: State bounds on $x_C$ from Example 3.3.1 computed by Harrison's method (dashed), Singer's method (crosses), and the FSI method (circles). Solid curves are true model solutions.

increases the cost by a factor of about 10. Though this latter increase is substantial, the absolute cost remains small. For slightly less than the cost of sampling trajectories on a 4 × 4 grid over $U$, we obtain a sharp, guaranteed enclosure.

**Example 4.3.2.** Consider the chemical reaction network

$$\mathrm{A} \rightarrow \mathrm{B} \rightarrow \mathrm{C}.$$

Assuming elementary reactions and Arrhenius rate constants, the concentrations of the chemical species, denoted $x_{\mathrm{A}}$, $x_{\mathrm{B}}$ and $x_{\mathrm{C}}$, in a closed system with temperature control are given by a kinetic model of the form (4.1), where $\mathbf{x} \equiv (x_{\mathrm{A}}, x_{\mathrm{B}}, x_{\mathrm{C}})$, $u \equiv T$, and

$$\mathbf{S} \equiv \begin{bmatrix} -1 & 0 \\ 1 & -1 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{r}(t, p, \mathbf{z}) \equiv \begin{bmatrix} A_1 e^{-E_1/(Rp)} z_{\mathrm{A}} \\ A_2 e^{-E_2/(Rp)} z_{\mathrm{B}} \end{bmatrix}.$$

Above, $R = 8.314 \, \frac{J}{\mathrm{mol} \cdot K}$ is the universal gas constant, $A_1 = 2400 \; s^{-1}$, $A_2 = 8800 \; s^{-1}$, $E_1 = 6.9 \times 10^3$ $(J/\mathrm{mol})$ and $E_2 = 1.69 \times 10^4$ $(J/\mathrm{mol})$. Note that this system is neither linear nor control-linear because the right-hand side functions contain products of the states and nonlinear functions of the control.

Consider bounding the reachable set on $I = [0, 0.08]$ $(s)$ with the fixed initial condition $\mathbf{x}_0 = (1.5, 0.5, 0.0)$ (M) and the temperature bounded between 300 and 600 $K$. That is, the set of admissible initial conditions is the singleton $X_0 = \{\mathbf{x}_0\}$ and the set of admissible controls is

$$\mathcal{U} = \{T \in L^1(I) : T(t) \in [300, 600] \ (K) \text{ for a.e. } t \in I\}.$$

With no further information, state bounds can be computed by Harrison's method. The resulting bounds on $x_{\text{B}}$ are shown in Figure 4-3, along with several model solutions for temperature profiles in $\mathcal{U}$. These solutions correspond to piecewise constant temperature profiles with 8 epochs of length $0.01s$, taking one of 8 possible temperature values in the first epoch, spaced evenly in the interval $[300, 600]$ $K$, and one of two possible values in each remaining epoch, $300K$ or $600K$. Clearly, the method provides valid bounds on all model solutions shown. Moreover, the choice of piecewise constant controls was simply for computational convenience; by Corollary 3.5.8 and the discussion in §3.5.3, the solutions of (3.31) are guaranteed to bound the model solutions with *any* $T \in \mathcal{U}$.

For this example, the cost of integrating a single trajectory is $1.7 \times 10^{-4}$s. Harrison's method requires only $4.6 \times 10^{-4}$s, but again produces very conservative bounds. However, a valid *a priori* enclosure can be obtained as follows. First, since $\mathbf{x}_0 \geq \mathbf{0}$, it follows that all model solutions are nonnegative for all $(t, T) \in I \times \mathcal{U}$. Furthermore, $x_{\text{A}}$ cannot be generated, so it is bounded above by 1.5 (this is an example of a bound based on the directionality of reactions that cannot be inferred from $\mathbf{S}$, as discussed in §4.2). The stoichiometry matrix has rank 2, so there is one linearly independent reaction invariant, which is easily seen to be $\mathbf{m}^{\text{T}} = \mathbf{1}^{\text{T}}$. Combining these observations, the programs (4.5) and (4.6) give the refined natural bounds $X^N = [1.5, 2] \times [0, 2] \times [0, 2]$.

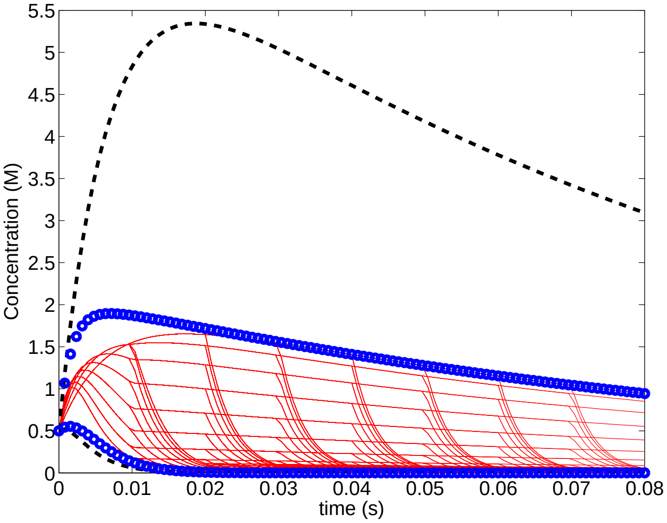

Figure 4-3: State bounds on $x_{\rm B}$ in Example 4.3.2 computed by Harrison's method (dashed) and the FSI method (circles), along with true model solutions for several piecewise constant temperature profiles (solid).

Then, a second set of state bounds can be computed using the FSI method with

$$G \equiv \{\mathbf{z} \in \mathbb{R}^3 : (0,0,0) \leq \mathbf{z} \leq (1.5, 2, 2),\ \mathbf{1}^{\rm T}\mathbf{z} = \mathbf{1}^{\rm T}\mathbf{x}_0\}.$$

The resulting bound, shown by the circles in Figure 4-3, are very tight. Moreover, this computation took only $2.7 \times 10^{-3}$s; less than 6 times longer than the standard method with no physical information.

**Example 4.3.3.** Consider the enzymatic reaction network with 6 states [2]:

$$\begin{aligned} \mathrm{A} + \mathrm{F} &\rightleftharpoons \mathrm{F:A} \rightarrow \mathrm{F} + \mathrm{A}', \\ \mathrm{A}' + \mathrm{R} &\rightleftharpoons \mathrm{R:A}' \rightarrow \mathrm{R} + \mathrm{A}. \end{aligned}$$

With $\mathbf{x} \equiv (x_{\rm A}, x_{\rm F}, x_{\rm F:A}, x_{\rm A'}, x_{\rm R}, x_{\rm R:A'})$ and the controls $\mathbf{u} = (k_1, \ldots, k_6)$ representing the rate constants for all six reactions, the dynamics in a closed system are described by a kinetic model of the form (4.1) with

$$\mathbf{S} \equiv \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 1 \\ -1 & 1 & 1 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 & 1 \\ 0 & 0 & 0 & 1 & -1 & -1 \end{bmatrix}, \quad \mathbf{r}(t, \mathbf{p}, \mathbf{z}) \equiv \begin{bmatrix} p_1 z_{\text{A}} z_{\text{F}} \\ p_2 z_{\text{F:A}} \\ p_3 z_{\text{F:A}} \\ p_4 z_{\text{A}'} z_{\text{R}} \\ p_5 z_{\text{R:A}'} \\ p_6 z_{\text{R:A}'} \end{bmatrix}.$$

Consider computing state bounds for this model with $I = [0, 0.04]$ $(s)$, the fixed initial condition $\mathbf{x}_0 = (20, 34, 0, 0, 16, 0)$ (M) and uncertain rate parameters $k_i \in [\hat{k}_i, 10\hat{k}_i]$, where

$$\hat{\mathbf{k}} = (0.1, 0.033, 16, 5, 0.5, 0.3).$$

That is, $X_0 = \{\mathbf{x}_0\}$ and the set of admissible controls is

$$\mathcal{U} = \{\mathbf{u} \in (L^1(I))^6 : \mathbf{u}(t) \in [\hat{\mathbf{k}}, 10\hat{\mathbf{k}}] \text{ for a.e. } t \in I\}.$$

The results of Harrison's method are shown in Figures 4-4 and 4-5, along with true model solutions corresponding to constant $\mathbf{u}$ taking values on a uniform grid with three points in each of the six dimensions. As discussed in Remark 3.2.4, the computed bounds are also valid for time-varying $\mathbf{u} \in \mathcal{U}$. Due to the large number of parameters considered, sampling true model solutions becomes unmanageable for piece-wise constant $\mathbf{u}$, even with only 3 epochs. Some such trajectories were explored manually and none were found to lie outside of the set reachable with constant $\mathbf{u}$.

For this example, integration of a single model solution required $1.7 \times 10^{-4}$s, while Harrison's method required $2.37 \times 10^{-3}$s. The resulting bounds diverge rapidly, providing no useful information about the reachable set. In fact, the divergence is so rapid in this case that numerical integration is slower than one might expect. In previous examples, the CPU time for Harrison's method compared more favorably with that of integrating a single model solution.

An *a priori* enclosure is derived as follows. Since $\mathbf{x}_0 \geq \mathbf{0}$, it follows that all model solutions must be nonnegative for all $(t, \mathbf{u}) \in I \times \mathcal{U}$. The stoichiometry matrix has rank 3, indicating that there are 3 linearly independent affine reaction invariants. Applying the `MATLAB` routine `null` to $\mathbf{S}^{\mathrm{T}}$, a basis for the left null space of $\mathbf{S}$, and hence a complete set of linearly independent reaction invariants, is given by the rows of

$$\mathbf{M} = \begin{bmatrix} -0.5743 & 0.2150 & -0.3593 & -0.5743 & 0.2872 & -0.2872 \\ -0.0589 & 0.7329 & 0.6740 & -0.0589 & 0.0295 & -0.0295 \\ 0 & 0.0000 & 0.0000 & 0.0000 & 0.7071 & 0.7071 \end{bmatrix}.$$

Through elementary row operations, it is not difficult to show that the basis

$$\mathbf{M} = \begin{bmatrix} 0 & -1 & -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 1 & -1 & 0 \\ -1 & 1 & 0 & -1 & 0 & -1 \end{bmatrix}.$$

defines the same subspace. Physically, the rows of $\mathbf{M}$ describe stoichiometric relationships between species which result from the cyclic structure of the reaction network. Such cycles are very common in biological networks, where they are referred to as metabolic pools [58]. Combining this with nonnegativity through (4.5) and (4.6) gives the natural bounds $X^N = [0, 20] \times [0, 24] \times [0, 20] \times [0, 24] \times [0, 16] \times [0, 16]$.

A second set of state bounds can now be computed using the FSI method with

$$G \equiv \{\mathbf{z} \in \mathbb{R}^6 : \mathbf{z} \in X^N, \ \mathbf{M}\mathbf{z} = \mathbf{M}\mathbf{x}_0\}.$$

The resulting bounds are shown in Figures 4-4 and 4-5. Clearly, the bounds produced by this approach do not diverge, and in fact track the true solution set very accurately. Moreover, this computation takes only $9.11 \times 10^{-3}$s; about 4 times longer than the standard approach without using physical information.

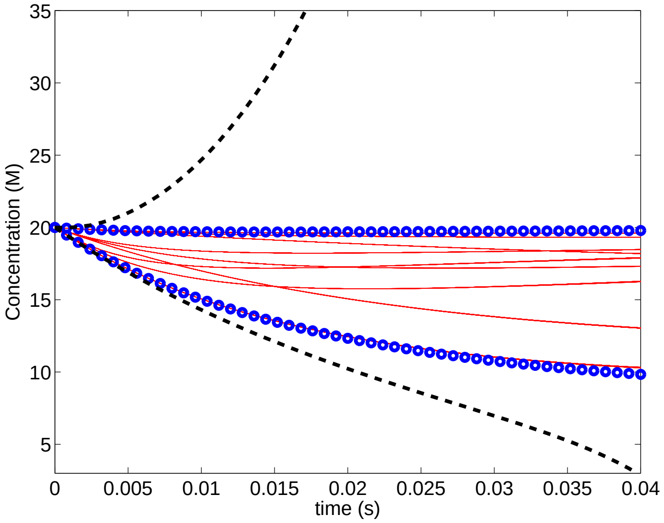

Figure 4-4: State bounds on $x_{\mathrm{A}}$ in Example 4.3.3 computed by Harrison's method (dashed) and the FSI method (circles), along with true model solutions for constant $\mathbf{u}$ on a uniform grid (solid).

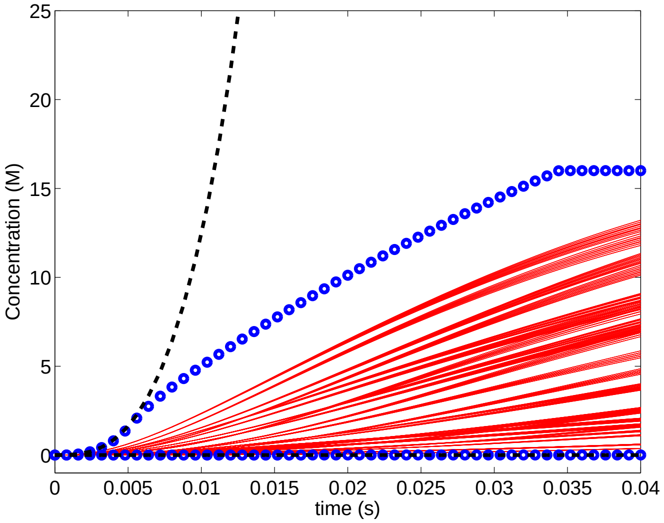

Figure 4-5: State bounds on $x_{\mathrm{R:A'}}$ in Example 4.3.3 computed by Harrison's method (dashed) and the FSI method (circles), along with true model solutions for constant $\mathbf{u}$ on a uniform grid (solid).

## 4.4 Reduced Kinetic Models

In the previous section, it was shown that using natural bounds and affine reaction invariants in conjunction with the state bounding methods developed in Chapter 3 results in dramatic improvements over existing bounding methods. In the remainder of this chapter, we investigate an alternative method that uses reaction invariants in a very different way. It is well known that the solution of (4.1) can be fully described by a reduced system of ODEs of dimension $n_x - m$, where $m$ is the number of linearly independent affine reaction invariants [181]. Then, the alternative approach is to construct such a reduced system, bound its solutions, and then recover state bounds for the original system through an affine transformation.

To illustrate this idea, recall Figure 4-1 from Example 4.2.1. The Figure shows that, though the kinetic model in question has three state variables, the affine reaction invariants render the evolution of the system essentially one dimensional. That is, the time evolution of the system may be described completely by its progression along the direction $y$ shown in the Figure. Accordingly, we consider computing bounds on the linear combination $y$, rather than on each of the species concentrations $x_A$, $x_B$ and $x_C$. This is illustrated in Figure 4-6. The image on the upper left shows a box in $(x_A, x_B, x_C)$-space which represents hypothetical state bounds computed for some $t' \in I$. The bounds are represented by a box, or 3-dimensional interval, simply because the bounding procedures discussed so far compute an upper and lower bound for each of the three state variables. However, it has been shown that all model solutions must lie on the $y$-axis, so every point inside of the depicted box that does not lie on the $y$-axis, and hence violates at least one of the reaction invariants, cannot possibly be a solution. Therefore, there is significant overestimation in this enclosure, simply on account of using a 3-dimensional interval. Moreover, integrating the bounding differential equations (3.31) forward from $t'$, using Harrison's method for example, requires taking the necessary natural interval extensions of the model right-hand side functions over the entire box, even though only points on the $y$-axis can be true model solutions. In this manner, the overestimation inherent in these state bounds propagates forward in time and weakens the computed bounds for every $t' \leq t \leq t_f$. Of course, the purpose of the FSI method is to refine this box using the reaction invariants, so that overestimation is prevented from propagating forward in time insofar as possible. Nonetheless, the end result is still a 3-dimensional interval that is geometrically forced to overestimate the true solution set significantly.

On the other hand, if a variable transformation is carried out which defines $y$ as a linear combination of $x_A$, $x_B$ and $x_C$, as depicted, and bounds are constructed for $y$, then the problem is alleviated. The brackets along the $y$-axis in the lower right of Figure 4-6 depict hypothetical bounds of this type. Though the bounds along the $y$-axis may overestimate the set of true model solutions, they may not include any points which violate the reaction invariants. Because the time evolution of the entire model can be recovered only from knowledge of the time evolution along the $y$-axis, valid bounds on $x_A$, $x_B$ and $x_C$ can be recovered from valid bounds on $y$. Of course, the resulting bounds will suffer from overestimation because, for any $t \in I$, the true set of model solutions is not a 3-dimensional interval. However, the major advantage of this approach is that this overestimation is prevented from propagating forward during integration of the bounding differential equations since only bounds on the linear combination $y$ are propagated forward in time.

Despite this geometric advantage, this approach is not generally superior to the FSI method demonstrated in the previous section. It requires much of the same theory, has similar cost, and typically produces bounds that are comparable but worse. However, often enough to encourage curiosity, the resulting bounds are sharper. Thus, there is an open opportunity to better understand and exploit the advantages of this approach in the future.

### 4.4.1 Constructing Reduced Models

As mentioned above, the solution of (4.1) can be fully described by a reduced system of ODEs of dimension $n_x - m$, where $m$ is the number of linearly independent affine reaction invariants. Such a reduced system can always be constructed by choosing an appropriate subset of the original state variables [11, 181], or using the Moore-

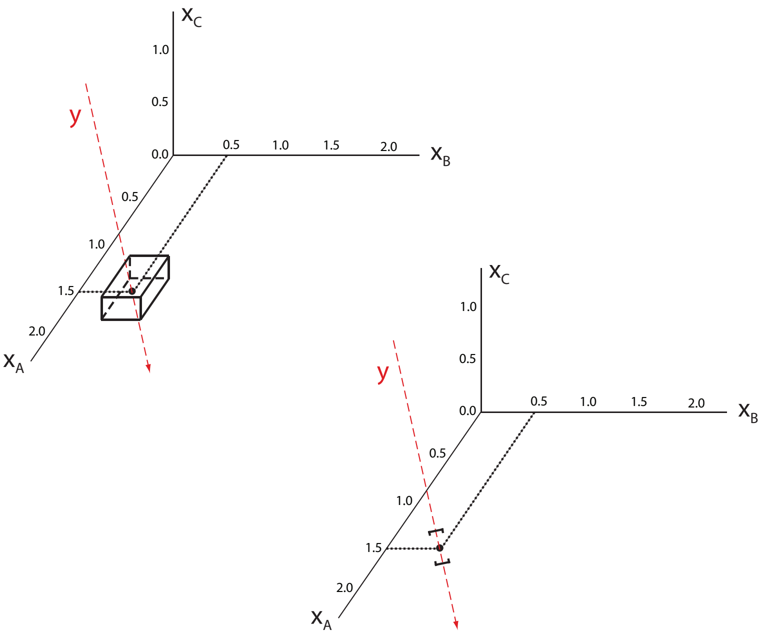

Figure 4-6: Geometric representation of state bounds in the full state space (upper left) and in the lower dimensional space (lower right) defined by the affine reaction invariants.

Penrose inverse of **S** [63, 181]. Theorem 4.4.6 below describes a more general family of reduced models based on $\{1,2\}$-inverses [21].

**Definition 4.4.1.** For any matrix $\mathbf{D} \in \mathbb{R}^{n\times m}$, $\mathbf{A} \in \mathbb{R}^{m\times n}$ is called a $\{1,2\}$-inverse of $\mathbf{D}$ if $\mathbf{DAD} = \mathbf{D}$ and $\mathbf{ADA} = \mathbf{A}$.

**Lemma 4.4.2.** *For any* $\mathbf{D} \in \mathbb{R}^{n\times m}$*, a* $\{1,2\}$*-inverse exists.*

*Proof.* See Theorem 1, Ch.1 Sec.2 and Lemma 3, Ch.1 Sec.4 in [21]. □

**Definition 4.4.3.** Let $L$ and $M$ be complementary subspaces of $\mathbb{R}^n$. Denote by $\mathcal{P}_{L,M} : \mathbb{R}^n \to \mathbb{R}^n$ the unique linear operator such that

$$\mathbf{z} = \mathcal{P}_{L,M}\mathbf{z} + (\mathbf{I} - \mathcal{P}_{L,M})\mathbf{z} = \mathbf{u} + \mathbf{v}, \tag{4.11}$$

where $\mathbf{u} \in L$ and $\mathbf{v} \in M$. $\mathcal{P}_{L,M}$ is referred to as the projector onto $L$ along $M$.

**Lemma 4.4.4.** $\mathcal{P}_{L,M}\mathbf{z} = \mathbf{z}$ *if and only if* $\mathbf{z} \in L$.

*Proof.* See Theorem 8 and Lemma 1(e), Sec. 4 Ch. 2 in [21]. □

**Lemma 4.4.5.** *If* $\mathbf{A} \in \mathbb{R}^{n\times m}$ *and* $\mathbf{D} \in \mathbb{R}^{m\times n}$ *are* $\{1,2\}$*-inverses of each other, then* $\mathbf{DA} = \mathcal{P}_{\mathcal{R}(\mathbf{D}),\mathcal{N}(\mathbf{A})}$ *and* $\mathbf{AD} = \mathcal{P}_{\mathcal{R}(\mathbf{A}),\mathcal{N}(\mathbf{D})}$.

*Proof.* See Corollary 7, Sec. 4 Ch. 2 in [21]. □

The following theorem defines a reduced system in terms of a pair of $\{1,2\}$-inverses and demonstrates that the solution of this system can be mapped uniquely to the solution of (4.1).

**Theorem 4.4.6.** *Consider the kinetic model* (4.1) *and suppose that the rank of* $\mathbf{S}$ *is* $n_y < n_x$. *Let* $\mathbf{A} \in \mathbb{R}^{n_y\times n_x}$ *and* $\mathbf{D} \in \mathbb{R}^{n_x\times n_y}$ *be* $\{1,2\}$*-inverses, and suppose that the range of* $\mathbf{D}$ *is equal the range of* $\mathbf{S}$, *i.e.,* $\mathcal{R}(\mathbf{D}) = \mathcal{R}(\mathbf{S})$. *If* (4.1) *has a unique solution, then there exists a unique solution of the reduced system*

$$\dot{\mathbf{y}}(t,\mathbf{u},\mathbf{x}_0) = \mathbf{ASr}(t,\mathbf{u}(t),\mathbf{Dy}(t,\mathbf{u},\mathbf{x}_0)+\mathbf{x}_0), \quad \mathbf{y}(t_0,\mathbf{u},\mathbf{x}_0) = \mathbf{0}. \tag{4.12}$$

*Moreover,* $\mathbf{y}$ *satisfies*

1. $\mathbf{y}(t,\mathbf{u},\mathbf{x}_0) = \mathbf{A}(\mathbf{x}(t,\mathbf{u},\mathbf{x}_0) - \mathbf{x}_0)$,
2. $\mathbf{x}(t,\mathbf{u},\mathbf{x}_0) = \mathbf{Dy}(t,\mathbf{u},\mathbf{x}_0) + \mathbf{x}_0$,

*for all* $(t,\mathbf{u},\mathbf{x}_0) \in I \times \mathcal{U} \times X_0$.

*Proof.* Consider the solution of (4.1), $\mathbf{x}$, and note that

$$\begin{aligned}(\mathbf{x}(t,\mathbf{u},\mathbf{x}_0) - \mathbf{x}_0) &= \int_{t_0}^{t} \mathbf{Sr}(s,\mathbf{u}(s),\mathbf{x}(s,\mathbf{u},\mathbf{x}_0))ds, \\ &= \mathbf{S}\int_{t_0}^{t} \mathbf{r}(s,\mathbf{u}(s),\mathbf{x}(s,\mathbf{u},\mathbf{x}_0))ds, \\ &\in \mathcal{R}(\mathbf{S}) = \mathcal{R}(\mathbf{D}),\end{aligned} \tag{4.13}$$

for all $(t,\mathbf{u},\mathbf{x}_0) \in I \times \mathcal{U} \times X_0$. Define $\mathbf{y}(t,\mathbf{u},\mathbf{x}_0) \equiv \mathbf{A}(\mathbf{x}(t,\mathbf{u},\mathbf{x}_0) - \mathbf{x}_0)$. Then $\mathbf{Dy}(t,\mathbf{u},\mathbf{x}_0) = \mathbf{DA}(\mathbf{x}(t,\mathbf{u},\mathbf{x}_0) - \mathbf{x}_0)$, and combining this with (4.13), Lemmas 4.4.4

and 4.4.5 imply that $\mathbf{D}\mathbf{y}(t,\mathbf{u},\mathbf{x}_0) = (\mathbf{x}(t,\mathbf{u},\mathbf{x}_0) - \mathbf{x}_0)$. Differentiating $\mathbf{y}$ now gives

$$\begin{aligned}\dot{\mathbf{y}}(t,\mathbf{u},\mathbf{x}_0) &= \mathbf{A}\dot{\mathbf{x}}(t,\mathbf{u},\mathbf{x}_0), \\ &= \mathbf{A}\mathbf{S}\mathbf{r}(t,\mathbf{u}(t),\mathbf{x}(t,\mathbf{u},\mathbf{x}_0)) \\ &= \mathbf{A}\mathbf{S}\mathbf{r}(t,\mathbf{u}(t),\mathbf{D}\mathbf{y}(t,\mathbf{u},\mathbf{x}_0)) + \mathbf{x}_0).\end{aligned} \tag{4.14}$$

Thus, (4.12) has at least one solution satisfying 1 and 2.

Now consider a second solution of (4.12), $\mathbf{z}$, and define $\boldsymbol{\phi}(t,\mathbf{u},\mathbf{x}_0) \equiv \mathbf{D}\mathbf{z}(t,\mathbf{u},\mathbf{x}_0) + \mathbf{x}_0$. By differentiation,

$$\begin{aligned}\dot{\boldsymbol{\phi}}(t,\mathbf{u},\mathbf{x}_0) &= \mathbf{D}\dot{\mathbf{z}}(t,\mathbf{u},\mathbf{x}_0), \\ &= \mathbf{D}\mathbf{A}\mathbf{S}\mathbf{r}(t,\mathbf{u}(t),\mathbf{D}\mathbf{z}(t,\mathbf{u},\mathbf{x}_0) + \mathbf{x}_0), \\ &= \mathbf{S}\mathbf{r}(t,\mathbf{u}(t),\boldsymbol{\phi}(t,\mathbf{u},\mathbf{x}_0)),\end{aligned} \tag{4.15}$$

for all $(t,\mathbf{u},\mathbf{x}_0) \in I \times \mathcal{U} \times X_0$. Therefore, uniqueness implies that $\boldsymbol{\phi} = \mathbf{x}$ on $I \times \mathcal{U} \times X_0$, and hence $\mathbf{x}(t,\mathbf{u},\mathbf{x}_0) = \mathbf{D}\mathbf{z}(t,\mathbf{u},\mathbf{x}_0) + \mathbf{x}_0$. In particular,

$$\mathbf{A}\mathbf{D}\mathbf{z}(t,\mathbf{u},\mathbf{x}_0) = \mathbf{A}(\mathbf{x}(t,\mathbf{u},\mathbf{x}_0) - \mathbf{x}_0), \quad \forall (t,\mathbf{u},\mathbf{x}_0) \in I \times \mathcal{U} \times X_0. \tag{4.16}$$

From the form of (4.12) and the fact that $\mathbf{z}(t_0,\mathbf{u},\mathbf{x}_0) = \mathbf{0} \in \mathcal{R}(\mathbf{A})$, it follows that $\mathbf{z}(t,\mathbf{u},\mathbf{x}_0) \in \mathcal{R}(\mathbf{A})$ for all $(t,\mathbf{u},\mathbf{x}_0) \in I \times \mathcal{U} \times X_0$. Combining this with (4.16), Lemmas 4.4.4 and 4.4.5 imply that $\mathbf{z}(t,\mathbf{u},\mathbf{x}_0) = \mathbf{A}(\mathbf{x}(t,\mathbf{u},\mathbf{x}_0) - \mathbf{x}_0)$. Then it has been shown that $\mathbf{z}(t,\mathbf{u},\mathbf{x}_0) = \mathbf{y}(t,\mathbf{u},\mathbf{x}_0)$ for all $(t,\mathbf{u},\mathbf{x}_0) \in I \times \mathcal{U} \times X_0$, so the solution of (4.12) is unique. □

Given a stoichiometry matrix $\mathbf{S}$ which is not full row rank, the hypotheses of Theorem 4.4.6 are easily satisfied. The matrix $\mathbf{D}$ may be formed simply by choosing any maximal set of linearly independent columns of $\mathbf{S}$. Algorithms for computing a $\{1,2\}$-inverse of $\mathbf{D}$ can be found in [21], so a suitable matrix $\mathbf{A}$ is readily available. For example, the well-known Moore-Penrose inverse is a $\{1,2\}$-inverse and can be computed using the `MATLAB` routine `pinv`. It is interesting to note that if $\mathbf{A}$ is the

Moore-Penrose inverse of $\mathbf{D}$, then $\mathbf{AD} = \mathbf{I}$ always holds. Thus, if $\mathbf{S}$ is full column rank, then $\mathbf{D} = \mathbf{S}$ is an appropriate choice and the reduced system (4.12) reduces to

$$\begin{aligned}\dot{\mathbf{y}}(t, \mathbf{u}, \mathbf{x}_0) &= \mathbf{ADr}(t, \mathbf{u}(t), \mathbf{Dy}(t, \mathbf{u}, \mathbf{x}_0) + \mathbf{x}_0),\\ &= \mathbf{r}(t, \mathbf{u}(t), \mathbf{Dy}(t, \mathbf{u}, \mathbf{x}_0) + \mathbf{x}_0),\end{aligned}$$

with $\mathbf{y}(t, \mathbf{u}, \mathbf{x}_0) = \mathbf{0}$. In this case $\mathbf{y}$ corresponds to the well-known *extents of reaction* representation of a kinetic model [63].

The connection between the reduced model (4.12) and the underlying reaction invariants follows from Conclusions 1 and 2 in Theorem 4.4.6. Combining these gives

$$\begin{aligned}(\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) - \mathbf{x}_0) &= \mathbf{Dy}(t, \mathbf{u}, \mathbf{x}_0), \qquad (4.17)\\ &= \mathbf{DA}(\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) - \mathbf{x}_0),\end{aligned}$$

which implies that $(\mathbf{I} - \mathbf{DA})\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) = (\mathbf{I} - \mathbf{DA})\mathbf{x}_0$. Thus, the rows of the matrix $(\mathbf{I} - \mathbf{DA})$ are affine reaction invariants of the original kinetic model and, of these, $n_x - n_y$ must be linearly independent.

**Example 4.4.1.** Consider again the kinetic model given in Example 4.2.1. Since $\mathbf{S}$ has rank 1, any one column spans its entire range. Choosing $\mathbf{D} \equiv [-1 \ \ -1 \ \ 1]^{\mathrm{T}}$, the Moore-Penrose inverse of $\mathbf{D}$ is computed by the `MATLAB` routine `pinv` as $\mathbf{A} = [-1/3 \ \ -1/3 \ \ 1/3]$. Since the hypotheses of Theorem 4.4.6 are satisfied, a reduced kinetic model is given by

$$\begin{aligned}\dot{y}(t, \mathbf{u}, \mathbf{x}_0) &= \mathbf{ASr}(t, \mathbf{u}(t), \mathbf{D}y(t, \mathbf{u}, \mathbf{x}_0) + \mathbf{x}_0) \qquad (4.18)\\ &= k_f(-y(t, \mathbf{u}, \mathbf{x}_0) + x_{0,A})(-y(t, \mathbf{u}, \mathbf{x}_0) + x_{0,B})\\ &\qquad - k_r(y(t, \mathbf{u}, \mathbf{x}_0) + x_{0,C}).\end{aligned}$$

A complete set of linearly independent reaction invariants for the original model is given by the rows of $(\mathbf{I} - \mathbf{DA})$, which evaluates to

$$(\mathbf{I} - \mathbf{DA}) = \begin{bmatrix} 2/3 & -1/3 & 1/3 \\ -1/3 & 2/3 & 1/3 \\ 1/3 & 1/3 & 2/3 \end{bmatrix}.$$

This matrix is rank 2, as expected, and it is easy to check that any 2 rows form a basis for $\mathcal{N}(\mathbf{S}^{\mathrm{T}})$. Moreover, the basis of Example 4.2.1 is easily obtained through elementary row operations.

### 4.4.2 Reduced Space State Bounding

Given an arbitrary kinetic model of the form (4.1), where $\mathbf{S}$ is not full row rank, Theorem 4.4.6 provides a means for constructing a reduced model which can describe the time evolution of the original model fully through Conclusion 2 of that theorem. Of course, any of the bounding methods described in Chapter 3 can be applied directly to this reduced model. To avoid confusion, denote state bounds for the original and reduced models by $\mathbf{x}^L, \mathbf{x}^U : I \to \mathbb{R}^{n_x}$ and $\mathbf{y}^L, \mathbf{y}^U : I \to \mathbb{R}^{n_y}$, respectively. Furthermore, assume that $X_0$ is an interval. Then, having computed $\mathbf{y}^L$ and $\mathbf{y}^U$ by any of the available methods, Conclusion 2 of Theorem 4.4.6 implies that state bounds for (4.1) are given by

$$[\mathbf{x}^L(t), \mathbf{x}^U(t)] = \mathbf{D}[\mathbf{y}^L(t), \mathbf{y}^U(t)] + X_0, \quad \forall t \in I. \tag{4.19}$$

In what follows, any method for computing state bounds for (4.1) through (4.19) is referred to as a *reduced-space method*, in contrast to the *full-space methods* that have been considered thus far.

The simplest reduced-space method results from applying Harrison's method to (4.12). Indeed, the affine reaction invariants of the original model have essentially been used to define the reduced model. Provided that the smallest possible reduction in dimension was made, the reduced model does not itself satisfy any affine reaction invariants. Therefore, there is seemingly no need to resort to the methods of §3.6 to compute state bounds for (4.12). However, it turns out that such a reduced-space Harrison's method does not produce sharp bounds. One reason for this is that the natural bounds $X^N$, in particular the nonnegativity constraints, are not exploited.

Ironically, the affine variable transformation from $\mathbf{x}$ to $\mathbf{y}$ that eliminates the need to deal directly with reaction invariants also convolutes the natural bounds. That is, the simple interval constraint $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \in X^N$ for all $I \times \mathcal{U} \times X_0$ can be written in terms of the reduced variables $\mathbf{y}$ only as the more complicated polyhedral constraint

$$\mathbf{x}^{N,L} \leq \mathbf{D}\mathbf{y}(t, \mathbf{u}, \mathbf{x}_0) + \mathbf{x}_0 \leq \mathbf{x}^{N,U}, \quad \forall (t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0. \tag{4.20}$$

Thus, to obtain sharp bounds from a reduced-space method, one is again forced to deal directly with *a priori* enclosures.

Define the *reduced-space natural bounds* $Y^N = [\mathbf{y}^{N,L}, \mathbf{y}^{N,U}]$ as the solutions of the following linear programs:

$$\begin{aligned} y_i^{N,L} \quad &:= \quad \inf_{\boldsymbol{\phi}, \mathbf{z}, \mathbf{x}_0} \phi_i && (4.21)\\ &\text{s.t.} \quad \boldsymbol{\phi} = \mathbf{A}(\mathbf{z} - \mathbf{x}_0) \\ &\qquad\ \ \mathbf{z} \in X^N, \quad \mathbf{x}_0 \in X_0 \\ y_i^{N,U} \quad &:= \quad \sup_{\boldsymbol{\phi}, \mathbf{z}, \mathbf{x}_0} \phi_i && (4.22)\\ &\text{s.t.} \quad \boldsymbol{\phi} = \mathbf{A}(\mathbf{z} - \mathbf{x}_0) \\ &\qquad\ \ \mathbf{z} \in X^N, \quad \mathbf{x}_0 \in X_0, \end{aligned}$$

for all $i = 1, \ldots, n_y$. Combining $Y^N$ with (4.20) gives the *a priori* enclosure for the solutions of (4.12),

$$G \equiv \{\mathbf{z} \in \mathbb{R}^{n_y} : \mathbf{z} \in Y^N, \ \mathbf{D}\mathbf{z} \in X^N - X_0\}, \tag{4.23}$$

Expanding the interval operations, it is easily seen that $G$ is a convex polyhedral set.

To compute state bounds on the solutions of (4.12), we will essentially apply the FSI method using the set $G$ above. To do this, it is necessary to ensure that the right-hand sides of (4.12) satisfy Assumptions and 3.2.1 and 3.2.2, and to derive an inclusion monotonic interval extension of these functions in order to use the efficient interval implementation (3.31). Actually, this will be done for a modified reduced system defined below, which has the advantage that $X^N$ can be further exploited in the interval extension of its right-hand side functions.

**Definition 4.4.7.** Let $D_y \subset \mathbb{R}^{n_y}$ and $\mathbf{h} : I \times U \times X_0 \times D_y \to \mathbb{R}^{n_y}$ be defined by

$$D_y \equiv \{\mathbf{y} \in \mathbb{R}^{n_y} : \text{mid}(\mathbf{x}^{N,L}, \mathbf{x}^{N,U}, \mathbf{D}\mathbf{y} + \mathbf{x}_0) \in D,\ \forall \mathbf{x}_0 \in X_0\},$$
$$\mathbf{h}(t, \mathbf{p}, \mathbf{x}_0, \mathbf{y}) \equiv \mathbf{A}\mathbf{S}\mathbf{r}(t, \mathbf{p}, \text{mid}(\mathbf{x}^{N,L}, \mathbf{x}^{N,U}, \mathbf{D}\mathbf{y} + \mathbf{x}_0)).$$

**Assumption 4.4.8.** Let $\mathbf{y}$ be the unique solution of (4.12). Then $\mathbf{y}(t, \mathbf{u}, \mathbf{x}_0) \in D_y$, $\forall (t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$.

**Corollary 4.4.9.** *Let the hypotheses of Theorem 4.4.6 hold. If Assumption 4.4.8 holds, then the solution of* (4.12) *is also a solution of*

$$\dot{\mathbf{y}}(t, \mathbf{u}, \mathbf{x}_0) = \mathbf{h}(t, \mathbf{u}(t), \mathbf{y}(t, \mathbf{u}, \mathbf{x}_0)), \quad \mathbf{y}(t_0, \mathbf{u}, \mathbf{x}_0) = \mathbf{0}, \tag{4.24}$$

*on* $I \times \mathcal{U} \times X_0$.

*Proof.* Let $\mathbf{y}$ be the unique solution of (4.12). By Conclusion 2 of Theorem 4.4.6, $\mathbf{D}\mathbf{y}(t, \mathbf{u}, \mathbf{x}_0) + \mathbf{x}_0 = \mathbf{x}(t, \mathbf{u}, \mathbf{x}_0)$, and hence

$$\text{mid}(\mathbf{x}^{N,L}, \mathbf{x}^{U,N}, \mathbf{D}\mathbf{y}(t, \mathbf{u}, \mathbf{x}_0) + \mathbf{x}_0) = \mathbf{D}\mathbf{y}(t, \mathbf{u}, \mathbf{x}_0) + \mathbf{x}_0, \tag{4.25}$$

for all $(t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$. Then, $\mathbf{y}$ is also a solution of (4.24). □

Next it is shown that the ODEs (4.24) satisfy the assumption of §3.2. This justifies applying the bounding methods developed in Chapter 3 to (4.24). It also implies that the solution of (4.24) is unique. For the following lemma, note that the initial condition of the reduced system (4.24) is always $\mathbf{0}$, and the initial condition vector for the full model $\mathbf{x}_0$ plays the role of a parameter in the right-hand side function.

**Lemma 4.4.10.** *If Assumptions 3.2.1 and 3.2.2 hold with* $\mathbf{f} \equiv \mathbf{r}$*, then they hold with* $\mathbf{f} \equiv \mathbf{h}$*, under the interpretation* $U \equiv U \times X_0$ *and* $D \equiv D_y$*.*

*Proof.* Fix any $(\mathbf{p}, \mathbf{x}_0, \mathbf{z}_y) \in U \times X_0 \times D_y$. Then $\mathrm{mid}(\mathbf{x}^{N,L}, \mathbf{x}^{N,U}, \mathbf{D}\mathbf{z}_y + \mathbf{x}_0) \in D$ and Condition 1 of Assumption 3.2.1 implies that $\mathbf{r}(\cdot, \mathbf{p}, \mathrm{mid}(\mathbf{x}^{N,L}, \mathbf{x}^{N,U}, \mathbf{D}\mathbf{z}_y + \mathbf{x}_0))$ is measurable on $I$. If follows that $\mathbf{ASr}(\cdot, \mathbf{p}, \mathrm{mid}(\mathbf{x}^{N,L}, \mathbf{x}^{N,U}, \mathbf{D}\mathbf{z}_y + \mathbf{x}_0))$ is also measurable on $I$, which establishes Condition 1 of Assumption 3.2.1 with $\mathbf{f} \equiv \mathbf{h}$.

For a.e. $t \in I$, Condition 2 of Assumption 3.2.1 states that $\mathbf{r}(t, \cdot, \cdot)$ is continuous on $P \times D$. It follows that $\mathbf{h}(t, \cdot, \cdot, \cdot)$ is continuous on $U \times X_0 \times D_y$. This establishes Condition 2 of Assumption 3.2.1 with $\mathbf{f} \equiv \mathbf{h}$.

Choose any compact $K_y \subset D_y$. The set $K \equiv \{\mathrm{mid}(\mathbf{x}^{N,L}, \mathbf{x}^{N,U}, \mathbf{D}\mathbf{z}_y + \mathbf{x}_0) : \mathbf{z}_y \in K_y,\ \mathbf{x}_0 \in X_0\}$ is compact, and $K \subset D$ by the definition of $D_y$. By Condition 3 of Assumption 3.2.1, $\exists \alpha \in L^1(I)$ such that $\|\mathbf{r}(t, \mathbf{p}, \mathbf{z}_x)\|_1 \leq \alpha(t)$ for a.e. $t \in I$ and every $(\mathbf{p}, \mathbf{z}_x) \in U \times K$. But this implies that $\|\mathbf{ASr}(t, \mathbf{p}, \mathrm{mid}(\mathbf{x}^{N,L}, \mathbf{x}^{N,U}, \mathbf{D}\mathbf{z}_y + \mathbf{x}_0))\|_1 \leq \|\mathbf{AS}\|_1 \alpha(t)$ for a.e. $t \in I$ and every $(\mathbf{p}, \mathbf{x}_0, \mathbf{z}_y) \in U \times X_0 \times K_y$. This proves Condition 3 of Assumption 3.2.1 with $\mathbf{f} \equiv \mathbf{h}$.

Choose any $\mathbf{z}_y \in D_y$ and any $\bar{\mathbf{x}}_0 \in X_0$. By the definition of $D_y$, the point $\mathbf{z}_x \equiv \mathrm{mid}(\mathbf{x}^{N,L}, \mathbf{x}^{N,U}, \mathbf{D}\mathbf{z}_y + \bar{\mathbf{x}}_0)$ is in $D$. Then, Assumption 3.2.2 furnishes $\eta > 0$ and $\bar{\alpha} \in L^1(I)$ such that, for a.e. $t \in I$ and every $\mathbf{p} \in U$,

$$\|\mathbf{r}(t, \mathbf{p}, \tilde{\mathbf{z}}_x) - \mathbf{r}(t, \mathbf{p}, \hat{\mathbf{z}}_x)\|_\infty \leq \bar{\alpha}(t) \|\tilde{\mathbf{z}}_x - \hat{\mathbf{z}}_x\|_\infty,$$

for every $\tilde{\mathbf{z}}, \hat{\mathbf{z}} \in B_\eta(\mathbf{z}) \cap D$. Choose $\bar{\epsilon} > 0$ small enough that $\mathrm{mid}(\mathbf{x}^{N,L}, \mathbf{x}^{N,U}, \mathbf{D}\hat{\mathbf{z}}_y + \mathbf{x}_0) \in B_\eta(\mathbf{z}_x)$ if $\hat{\mathbf{z}}_y \in B_{\bar{\epsilon}}(\mathbf{z}_y)$ and $\mathbf{x}_0 \in B_{\bar{\epsilon}}(\bar{\mathbf{x}}_0)$. Then, for a.e. $t \in I$ and any $(\mathbf{p}, \mathbf{x}_0) \in U \times B_{\bar{\epsilon}}(\bar{\mathbf{x}}_0)$,

$$\begin{aligned}\|\mathbf{ASr}(t, \mathbf{p}, \mathrm{mid}(\mathbf{x}^{N,L}, \mathbf{x}^{N,U}, \mathbf{D}\tilde{\mathbf{z}}_y + \mathbf{x}_0)) - \mathbf{ASr}(t, \mathbf{p}, \mathrm{mid}(\mathbf{x}^{N,L}, \mathbf{x}^{N,U}, \mathbf{D}\hat{\mathbf{z}}_y + \mathbf{x}_0))\|_\infty, \\ \leq \|\mathbf{AS}\| \bar{\alpha}(t) \|\mathbf{D}\| \|\tilde{\mathbf{z}}_y - \hat{\mathbf{z}}_y\|_\infty,\end{aligned}$$

for any $\tilde{\mathbf{z}}_y, \hat{\mathbf{z}}_y \in B_{\bar{\epsilon}}(\mathbf{z}_y) \cap D_y$, where the matrix norms are induced infinity-norms.

The previous construction provides a cover of $X_0$ by open balls. Since $X_0$ is compact, we may choose a finite subcover, $B_{\epsilon_1}(\mathbf{x}_0^1), \ldots, B_{\epsilon_K}(\mathbf{x}_0^K)$. Let $\epsilon \equiv \min_k \epsilon_k$ and define $\beta \in L^1(I)$ by $\beta(t) = \|\mathbf{AS}\|\|\mathbf{D}\| \max_k \alpha_k(t)$. Then, for any $(\mathbf{p}, \mathbf{x}_0) \in U \times X_0$, there is a $k$ such that $\mathbf{x}_0 \in B_{\epsilon_k}(\mathbf{x}_0^k)$, and hence

$$\begin{aligned}\|\mathbf{ASr}(t,\mathbf{p},\mathrm{mid}(\mathbf{x}^{N,L},\mathbf{x}^{N,U},\mathbf{D}\tilde{\mathbf{z}}_y+\mathbf{x}_0))-&\mathbf{ASr}(t,\mathbf{p},\mathrm{mid}(\mathbf{x}^{N,L},\mathbf{x}^{N,U},\mathbf{D}\hat{\mathbf{z}}_y+\mathbf{x}_0))\|_\infty\\ &\leq \|\mathbf{AS}\|\alpha_k(t)\|\mathbf{D}\|\|\tilde{\mathbf{z}}_y-\hat{\mathbf{z}}_y\|_\infty,\\ &\leq \beta(t)\|\tilde{\mathbf{z}}_y-\hat{\mathbf{z}}_y\|_\infty,\end{aligned}$$

for any $\tilde{\mathbf{z}}_y, \hat{\mathbf{z}}_y \in B_\epsilon(\mathbf{z}_y) \cap D_y$. Thus, Assumption 3.2.2 holds with $\mathbf{f} \equiv \mathbf{h}$. □

Next, an inclusion monotonic interval extension for $\mathbf{h}$ is derived. Suppose that Assumption 3.5.6 holds with $\mathbf{f} = \mathbf{r}$, so that an inclusion monotonic interval extension $[\mathbf{r}] : \mathfrak{D}_r \subset \mathbb{I}I \times \mathbb{I}U \times \mathbb{I}D \to \mathbb{R}^{n_r}$ is available. Then, an inclusion monotonic interval extension of $\mathbf{h}$ can be defined as follows.

**Definition 4.4.11.** Define $[\mathbf{h}] : \mathfrak{D}_H \to \mathbb{IR}^{n_y}$ by

$$\begin{aligned}&\mathfrak{D}_H \equiv \{(I', U', X_0', Y') \in \mathbb{I}I \times \mathbb{I}U \times \mathbb{I}X_0 \times \mathbb{I}D_y : (I', U', X^N \tilde{\cap} (\mathbf{D}Y' + X_0')) \in \mathfrak{D}_r\},\\ &[\mathbf{h}](I', U', X_0', Y') \equiv (\mathbf{AS})[\mathbf{r}](I', U', X^N \tilde{\cap} (\mathbf{D}Y' + X_0')).\end{aligned}$$

**Lemma 4.4.12.** *$([\mathbf{h}], \mathfrak{D}_H, \mathbb{IR}^{n_y})$ is an inclusion monotonic interval extension of $(\mathbf{h}, I \times U \times X_0 \times D_y, \mathbb{R}^{n_y})$.*

*Proof.* The lemma follows from Theorem 2.3.7, Lemma 2.5.24, and Lemma 2.3.5. □

The advantage of bounding the reduced system (4.24) as opposed to (4.12) is apparent from the previous definition and lemma. The intersection with $X^N$ in the definition of $[\mathbf{h}]$ would not be permitted otherwise. This particular usage of $X^N$ has a very profound impact on the resulting state bounds because it prevents the interval extension of the rate function $\mathbf{r}$ from being taken over intervals which contain non-physical points, such as negative species concentrations.

Another important point to note about the definition of $[\mathbf{h}]$ is the order of the multiplications $(\mathbf{AS})[\mathbf{r}]$. If this product is evaluated instead as $\mathbf{A}(\mathbf{S}[\mathbf{r}])$, a much weaker interval extension can result. For example suppose that $\mathbf{A}_i\mathbf{S}\mathbf{r}$ evaluates to something like $r_1 + r_2 + r_3 - r_2$. Then $(\mathbf{A}_i\mathbf{S})[\mathbf{r}] = [r_1] + [r_3]$, whereas $\mathbf{A}(\mathbf{S}[\mathbf{r}])$ gives the weaker enclosure $[r_1] + [r_2] + [r_3] - [r_2]$.

Now consider the bounding system

$$\begin{aligned} \dot{y}_i^L(t) &= [h_i]^L([t,t], U, \Omega_i^L(t, \mathbf{y}^L(t), \mathbf{y}^U(t))), \\ \dot{y}_i^U(t) &= [h_i]^U([t,t], U, \Omega_i^U(t, \mathbf{y}^L(t), \mathbf{y}^U(t))), \end{aligned} \tag{4.26}$$

for a.e. $t \in I$ and every $i \in \{1, \dots, n_y\}$, where

$$\begin{aligned} &D_\Omega \equiv I \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_y}, \\ &\Omega_i^L(t, \mathbf{y}^L, \mathbf{y}^U) \equiv \mathcal{B}_i^L(\mathcal{I}_G(\square(\mathbf{y}^L, \mathbf{y}^U))), \\ &\Omega_i^U(t, \mathbf{y}^L, \mathbf{y}^U) \equiv \mathcal{B}_i^U(\mathcal{I}_G(\square(\mathbf{y}^L, \mathbf{y}^U))). \end{aligned} \tag{4.27}$$

Above, $\mathcal{I}_G$ is defined as in Definition 3.6.3 and $G$ is defined by (4.23).

**Corollary 4.4.13.** *Suppose that $\mathbf{y}^L, \mathbf{y}^U \in \mathcal{AC}(I, \mathbb{R}^{n_x})$ satisfy (4.26) a.e. on $I$. If Assumption 4.4.8 holds, then $\mathbf{y}^L(t) \le \mathbf{y}(t, \mathbf{u}, \mathbf{x}_0) \le \mathbf{y}^U(t)$, $\forall(t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$. Moreover, let*

$$[\mathbf{x}^L(t), \mathbf{x}^U(t)] = X^N \cap \left(\mathbf{D}\mathcal{I}_G([\mathbf{y}^L(t), \mathbf{y}^U(t)]) + X_0\right), \quad \forall t \in I.$$

*Then $\mathbf{x}^L(t) \le \mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \le \mathbf{x}^U(t)$, $\forall(t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$.*

*Proof.* In light of Lemma 4.4.10, $\mathbf{y}^L(t) \le \mathbf{y}(t, \mathbf{u}, \mathbf{x}_0) \le \mathbf{y}^U(t)$, $\forall(t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$, provided that the hypotheses of Corollary 3.5.8 hold. Assumption 3.5.6 holds with $\mathbf{f} \equiv \mathbf{h}$ by Lemma 4.4.12. Assumption 3.5.7 follows directly from the definitions (4.27). Finally, Hypothesis 3.5.3 was proven in §3.6.1.

Since $\mathbf{y}(t, \mathbf{u}, \mathbf{x}_0) \in [\mathbf{y}^L(t), \mathbf{y}^U(t)]$, $\forall(t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$, it follows from Condition 2 of Definition 3.6.1 that $\mathbf{y}(t, \mathbf{u}, \mathbf{x}_0) \in \mathcal{I}_G([\mathbf{y}^L(t), \mathbf{y}^U(t)])$, $\forall(t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$. By Conclusion 2 of Theorem 4.4.6, $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) = \mathbf{D}\mathbf{y}(t, \mathbf{u}, \mathbf{x}_0) + \mathbf{x}_0$, $\forall(t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$, and hence $\mathbf{x}(t, \mathbf{u}, \mathbf{x}_0) \in \left(\mathbf{D}\mathcal{I}_G([\mathbf{y}^L(t), \mathbf{y}^U(t)]) + X_0\right)$, $\forall(t, \mathbf{u}, \mathbf{x}_0) \in I \times \mathcal{U} \times X_0$. By the definition of $X^N$, the corollary follows. □

## 4.5 Numerical Examples

State bounds for the case studies below were computed using a custom C++ code which takes as instance specific input $\mathbf{D}$, $\mathbf{A}$, $X^N$, $Y^N$, $\mathbf{S}$, $X_0$, $U$ and routines for evaluating the function $\mathbf{r}$. MATLAB was used to compute $\mathbf{A}$, $X^N$ and $Y^N$ from $\mathbf{D}$ as described in Sections 4.4.1, 4.2.2 and 4.4.2. The MATLAB routine pinv was used to compute $\mathbf{A}$, and linprog was used to compute $X^N$ and $Y^N$. From this input, it is trivial to evaluate the right-hand side function of the reduced model (4.24). All interval extensions were computed using the C++ library MC++ (http://www3.imperial.ac.uk/people/b.chachuat/research), and numerical integration was carried out using the package CVODE [44]. Hardware details can be found in §4.3.

Again, for ease of comparison, the method names introduced in §4.3 will be used here as well. In addition, let the reduced-space invariant (RSI) method refer to the method given by first solving (4.26) with the definitions (4.27), $G$ a polyhedral set and $\mathcal{I}_G$ as in Definition 3.6.3, and then computing state bounds for (4.1) exactly as in Corollary 4.4.13.

**Example 4.5.1.** Consider again the model of Example 4.3.1. There, state bounds were computed using Singer's method and the FSI method with $I = [0, 0.05]$ min, the set of admissible controls

$$\mathcal{U} = \{(k_f, k_r) \in (L^1(I))^2 : (k_f(t), k_r(t)) \in [100, 500] \times [0.001, 0.01] \text{ for a.e. } t \in I\},$$

and the singleton set of admissible initial conditions $X_0 = \{\mathbf{x}_0\}$ with $\mathbf{x}_0 = (1.5, 0.5, 0)$ (M). Here, state bounds are computed by the RSI method and compared. For ease of reference, the results of Singer's method are reproduced as the dashed curves in Figure 4-7 below. In all figures in this section, solid curves are true model solutions computed for sampled points $(\mathbf{u}, \mathbf{x}_0) \in \mathcal{U} \times X_0$.

From Example 4.4.1, a reduced model is given by (4.24) with $\mathbf{D} = [-1 \;\; -1 \;\; 1]^{\mathrm{T}}$

and $\mathbf{A} = [-1/3 \ \ -1/3 \ \ 1/3]$. Furthermore, natural bounds were derived in Example 4.2.1 as $X^N = [1, 1.5] \times [0, 0.5] \times [0, 0.5]$. Reduced space natural bounds $Y^N$ can be computed by solving the linear programs

$$\begin{aligned} y^{N,L} &:= \inf_{\boldsymbol{\phi},\mathbf{z}} \phi \qquad (4.28)\\ \text{s.t. } \ \phi &= -\frac{1}{3}(z_A - 1.5) - \frac{1}{3}(z_B - 0.5) + \frac{1}{3} z_C \\ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} &\leq \begin{bmatrix} z_A \\ z_B \\ z_C \end{bmatrix} \leq \begin{bmatrix} 1.5 \\ 0.5 \\ 0.5 \end{bmatrix} \\ y^{N,U} &:= \sup_{\boldsymbol{\phi},\mathbf{z}} \phi \qquad (4.29)\\ \text{s.t. } \ \phi &= -\frac{1}{3}(z_A - 1.5) - \frac{1}{3}(z_B - 0.5) + \frac{1}{3} z_C \\ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} &\leq \begin{bmatrix} z_A \\ z_B \\ z_C \end{bmatrix} \leq \begin{bmatrix} 1.5 \\ 0.5 \\ 0.5 \end{bmatrix} \end{aligned}$$

which yield $Y^N \equiv [0, 0.5]$.

Defining the set $G$ as in (4.23), the condition $z \in G$ is characterized by the following inequality constraints

$$\begin{aligned} y^{N,L} \leq & \quad z && \leq y^{N,U}, \\ (-x^U_{0,A} + x^{N,L}_A) \leq & \ -z && \leq (-x^L_{0,A} + x^{N,U}_A), \\ (-x^U_{0,B} + x^{N,L}_B) \leq & \ -z && \leq (-x^L_{0,B} + x^{N,U}_B), \\ (-x^U_{0,C} + x^{N,L}_C) \leq & \quad z && \leq (-x^L_{0,C} + x^{N,U}_C), \end{aligned}$$

where the bottom three lines result from the constraint $\mathbf{D}z \in X^N - X_0$. Therefore,

$\mathbf{G} \equiv \{z \in \mathbb{R} : \mathbf{J}z \leq \mathbf{b}\}$, where

$$\mathbf{J} = \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \\ -1 \\ 1 \\ 1 \\ -1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} y^{N,U} \\ -y^{N,L} \\ (-x^L_{0,A} + x^{N,U}_A) \\ -(-x^U_{0,A} + x^{N,L}_A) \\ (-x^L_{0,B} + x^{N,U}_B) \\ -(-x^U_{0,B} + x^{N,L}_B) \\ (-x^L_{0,C} + x^{N,U}_C) \\ -(-x^U_{0,C} + x^{N,L}_C) \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0 \\ 0 \\ 0.5 \\ 0 \\ 0.5 \\ 0.5 \\ 0 \end{bmatrix}. \tag{4.30}$$

Now, $\Omega^L_i$ and $\Omega^U_i$ as in (4.27) can be evaluated by directly applying $\mathcal{I}_G$ as per Definition 3.6.3.

The results of the RSI state bounding method are show for $x_{\mathrm{C}}$ by the circles in Figure 4-7. Clearly, the bounds generated through the RSI method are much tighter than those computed by Singer's method. The results of the FSI method presented in Example 4.2.1 are identical to those of the RSI method here. In fact, both methods produce the the best possible bounds for this problem.

**Example 4.5.2** (An initial condition problem)**.** Consider the reaction network

$$\begin{aligned} \mathrm{A} + \mathrm{B} &\rightarrow \mathrm{C} \\ \mathrm{A} + \mathrm{C} &\rightarrow \mathrm{D} \end{aligned} \tag{4.31}$$

with mass-action kinetics. Letting $\mathbf{x} = (x_A, x_B, x_C, x_D)$ and $u = k_1$, the dynamics of this system in an isothermal batch reactor are described by a kinetic model of the from (4.1) with

$$\mathbf{S} = \begin{bmatrix} -1 & -1 \\ -1 & 0 \\ 1 & -1 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{r}(t, p, \mathbf{z}) = \begin{bmatrix} p z_A z_B \\ 20 z_A z_C \end{bmatrix}, \tag{4.32}$$

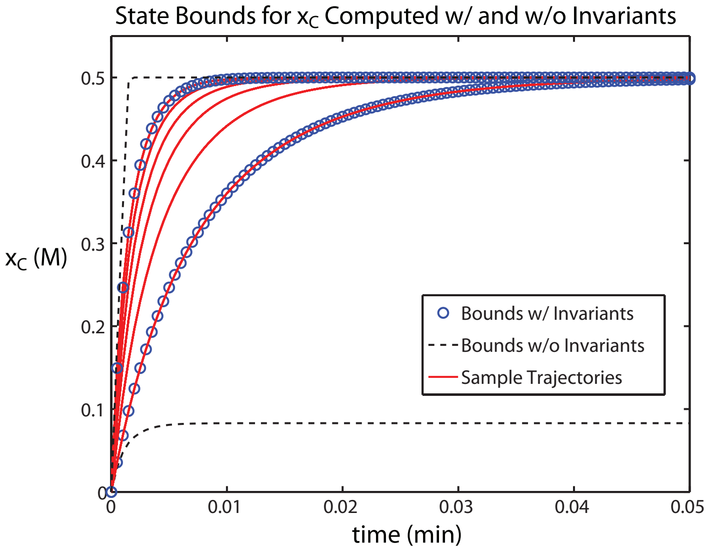

Figure 4-7: State bounds for the species concentration $x_C$ from Example 4.5.1 computed by Singer's method (dashed) and the RSI method (circles). Solid curves are true model solutions.

Let $I = [0, 0.1]$ s and suppose that $k_1$ is only known to within an order of magnitude, so that the single control $u = k_1$ is restricted to

$$\mathcal{U} = \{k_1 \in L^1(I) : k_1(t) \in [50, 500] \text{ M}^{-1}\text{s}^{-1} \text{ for a.e. } t \in I\}.$$

In addition, suppose that $x_{0,B}$ is measured as 1M with a ±5% error, so that the set of admissible initial conditions is $X_0 = [1, 1] \times [0.95, 1.05] \times [0, 0]$.

State bounds for this model were computed using Singer's method and the RSI method. Since $\mathbf{S}$ has rank 2, $\mathbf{D} = \mathbf{S}$ was chosen for the latter, and $\mathbf{A}$ was computed as the Moore-Penrose inverse of $\mathbf{D}$ by the `MATLAB` routine `pinv`. The result is

$$\mathbf{A} = \begin{bmatrix} -1/3 & -1/3 & 1/3 & 0 \\ -1/3 & 0 & -1/3 & 1/3 \end{bmatrix}.$$

From the reaction network, it can be seen that A and B are only consumed and C and D are limited by the amounts of A and B. Using these observations, the initial natural bounds were given as $X^N = [0, 1] \times [0, 1.05] \times [0, 1] \times [0, 1]$, and refined via the programs (4.5) and (4.6) to give $X^N = [0,1] \times [0,1.05] \times [0,1] \times [0,0.5]$. $Y^N$ was computed by solving the programs (4.21) and (4.22), resulting in $Y^N \equiv [0,1] \times [0,0.5]$.

Defining the set $G$ as in (4.23), the condition $\mathbf{z} \in G$ is characterized by the following inequality constraints

$$\begin{aligned}
y^{N,L} \leq &\quad z_1 &\leq y^{N,U}, \\
y^{N,L} \leq &\quad z_2 &\leq y^{N,U}, \\
(x_A^{N,U} - x_{0,A}^U) \leq &\quad -z_1 - z_2 &\leq (x_A^{N,L} - x_{0,A}^L), \\
(x_B^{N,U} - x_{0,B}^U) \leq &\quad -z_1 &\leq (x_B^{N,L} - x_{0,B}^L), \\
(x_C^{N,L} - x_{0,C}^U) \leq &\quad z_1 - z_2 &\leq (x_C^{N,U} - x_{0,C}^L), \\
(x_D^{N,L} - x_{0,D}^U) \leq &\quad z_2 &\leq (x_D^{N,U} - x_{0,D}^L)
\end{aligned}$$

Therefore, $\mathbf{G}$ can easily be put in the form $\{\mathbf{z} \in \mathbb{R}^2 : \mathbf{Jz} \leq \mathbf{b}\}$, and $\Omega_i^L$ and $\Omega_i^U$ as in (4.27) can be evaluated by applying $\mathcal{I}_G$ as per Definition 3.6.3.

State bounds computed for $x_B$ are shown in Figure 4-8. The dashed lines are the bounds computed by Singer's method, while the circles show the results of the RSI method. Clearly, the RSI method produces much sharper bounds than Singer's method.

**Example 4.5.3** (A PFR control problem)**.** Consider the reaction network

$$\begin{aligned}
\mathrm{R1} + \mathrm{R2} &\rightarrow \mathrm{I1} \\
\mathrm{R1} + \mathrm{I1} &\rightarrow \mathrm{A} \\
\mathrm{I1} &\rightarrow \mathrm{C} \\
\mathrm{C} + \mathrm{I1} &\rightleftharpoons \mathrm{I2} \\
\mathrm{Pt} &\rightarrow \mathrm{Pt}^*,
\end{aligned}$$

with rate coefficients $k_1, \ldots, k_6$ ($k_5$ denotes the reverse rate coefficient for the fourth reaction). All rate coefficients are temperature dependent through the standard Ar-

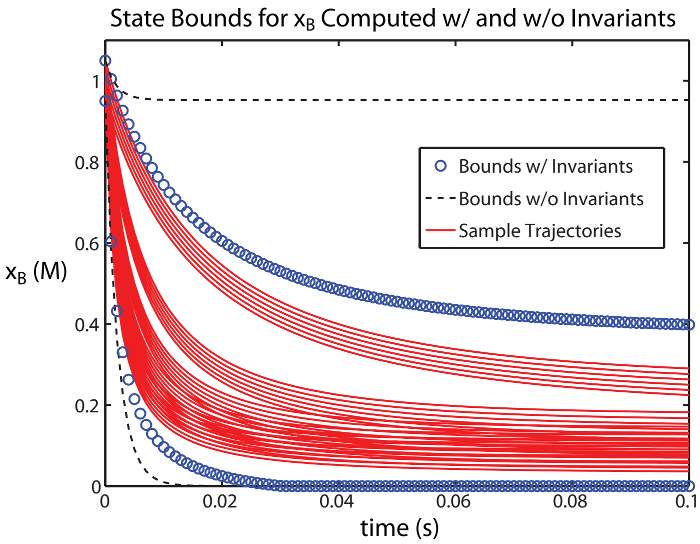

Figure 4-8: State bounds for the species concentration $x_B$ from Example 4.5.2 computed by Singer's method (dashed) and the RSI method (circles). Solid curves are true model solutions.

rhenius expression

$$k_i(T) = k_{0,i}e^{-E_i/RT},$$

where $R$ is the universal gas constant and values for $k_{0,i}$ and $E_i$ are listed in Table 4.1. Except for the first reaction, all reactions are considered to be elementary and obey mass-action kinetics. The first reaction takes place over a platinum catalyst with the rate expression given in (4.35), and deactivation of the catalyst is described by the final reaction. These reactions are considered in a plug flow reactor at steady state, and the response to various temperature profiles is bounded using Singer's method and the RSI method.

With $\mathbf{x} = (x_{R1}, x_{R2}, x_{I1}, x_{I2}, x_A, x_C, x_{Pt}, x_{Pt^*})$ and $u = T$, this system is described by a kinetic model of the form (4.1) with

$$\frac{d\mathbf{x}}{d\zeta}(\zeta, T, \mathbf{x}_0) = \mathbf{Sr}(\zeta, T(t), \mathbf{x}(\zeta, T, \mathbf{x}_0)), \quad \mathbf{x}(0, T, \mathbf{x}_0) = \mathbf{x}_0, \tag{4.33}$$

where $\zeta$ is the dimensionless reactor axial coordinate, $\mathbf{S}$ and $\mathbf{r}$ are given by

$$\mathbf{S} = \begin{bmatrix} -1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & -1 & 1 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{4.34}$$

and

$$\mathbf{r}(t, p, \mathbf{z}) = \begin{bmatrix} \tau k_1(p) z_{R1} z_{R2} z_{Pt}/(0.7 + z_{Pt}) \\ \tau k_2(p) z_{R1} z_{I1} \\ \tau k_3(p) z_{I1} \\ \tau k_4(p) z_{I1} x_C \\ \tau k_5(p) z_{I2} \\ \tau k_6(p) z_{Pt} \end{bmatrix}. \tag{4.35}$$

$\tau$ denotes the residence time, which is taken to be 1000 s. In contrast to the previous three examples, this system is not closed and does not obey mass action kinetics strictly. Nonetheless, the resulting kinetic model is of the form (4.1), so all of the methods presented are applicable.

We consider the solutions of this model for $\zeta \in I = [0, 1]$ and temperatures in the range $U = [350, 450]$ K, so that the set of admissible controls is

$$\mathcal{U} = \{T \in L^1(I) : T(\zeta) \in [350, 450] \text{ K for a.e. } \zeta \in I\}.$$

The set of admissible initial conditions is $X_0 = \{\mathbf{x}_0\}$ with $\mathbf{x}_0 = (516, 258, 0, 0, 0, 0, 1.1, 0)$ (mol/m$^3$).

To derive a reduced model, $\mathbf{D}$ was chosen as a maximal set of linearly independent columns of **S**,

$$\mathbf{D} = \begin{bmatrix} -1 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \tag{4.36}$$

and **A** chosen as the Moore-Penrose inverse of **D**, computed by the `MATLAB` `pinv` routine as

$$\mathbf{A} = \begin{bmatrix} -0.3333 & -0.5417 & 0.1250 & 0.2500 & -0.2083 & 0.125 & 0 & 0 \\ -0.3333 & 0.2083 & -0.125 & -0.25 & 0.5417 & -0.125 & 0 & 0 \\ 0 & -0.375 & -0.375 & 0.25 & -0.375 & 0.625 & 0 & 0 \\ 0 & -0.25 & -0.25 & 0.5 & -0.25 & -0.25 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0.5 & 0.5 \end{bmatrix}.$$

The natural bounds before refinement through (4.5) and (4.6) were

$$\begin{aligned} X^N = {} & [0{,}516] \times [0, 258] \times [0, 258] \times [0, 258] \\ & \times [0, 258] \times [0, 258] \times [0, 1.1] \times [0, 1.1] \text{ (mol/m}^3\text{)}, \end{aligned}$$

and (4.6) refined the fourth upper bound to 129 (mol/m$^3$).

State bounds computed for the species $x_{R1}$ and $x_C$ are shown in Figures 4-9 and 4-10, respectively, along with many solutions for several constant $T \in [350, 450]$. The dashed lines are the bounds computed by Singer's method, while the circles show the results of the RSI method. Clearly, the RSI method produces much sharper bounds than Singer's method. Among all species in the reaction network, the bounds on $x_{R1}$ in Figure 4-9 are typical, whereas the bounds on $x_C$ shown in Figure 4-10 are comparatively tight. There are other species in the model for which the bounds

Table 4.1: Values for the rate coefficients and activation energies, $k_{0,i}$ and $E_i$, for the kinetic model in Example 4.5.3.

| $i$ | $k_{0,i}$ (m$^3$/mol $\cdot$ s) or (1/s) | $E_i$ (J/mol) |
|---|---|---|
| 1 | $7.5 \times 10^4$ | 78240 |
| 2 | 1.01 | 45605 |
| 3 | $1.22 \times 10^{11}$ | 103345 |
| 4 | $3.58 \times 10^{-2}$ | 32217 |
| 5 | $7.33 \times 10^9$ | 91211 |
| 6 | $1.39 \times 10^{-4}$ | 0 |

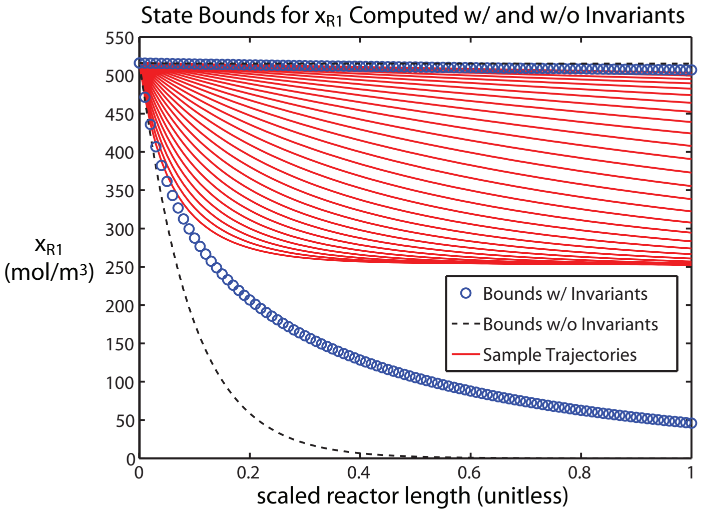

Figure 4-9: State bounds for the species concentration $x_{R1}$ from Example 4.5.3 computed by Singer's method (dashed) and the RSI method (circles). Solid curves are true model solutions.

computed by both methods are tight, and still others for which both methods give weak bounds (data not shown). In all cases, the bounds generated by the RSI method are superior to those generated using Singer's method. Finally, note that the bounds computed by Singer's method are often only reasonable because they are intersected with the natural bounds in the figures. For example, the upper bounding trajectory from Singer's method in Figure 4-10 carries almost no information that is not already known from the natural upper bound. In contrast, the upper bound computed from the RSI method tracks the upper limit of $x_C$ accurately along the entire length of the reactor.

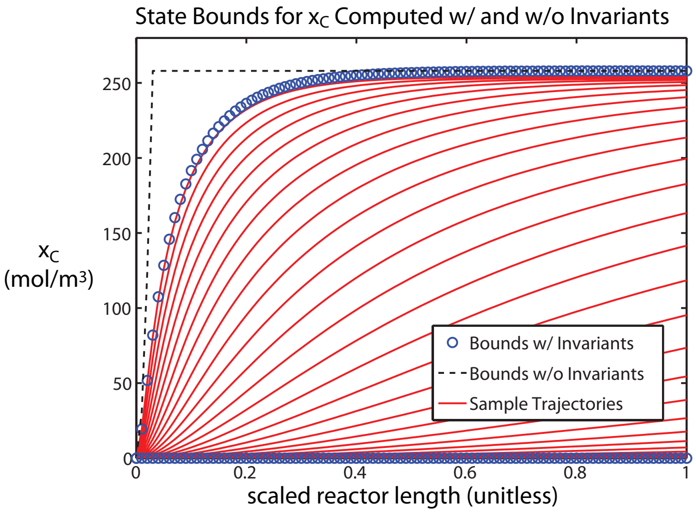

Figure 4-10: State bounds for the species concentration $x_C$ from Example 4.5.3 computed by Singer's method (dashed) and the RSI method (circles). Solid curves are true model solutions.

**Example 4.5.4** (A parameter estimation problem)**.** In [163], global parameter estimation and model verification is carried out for the proposed kinetic mechanism

$$\begin{gathered}
(\mathrm{CH_3})_3\mathrm{CO} + 1,4\text{-}\mathrm{C_6H_8} \rightarrow c\text{-}\mathrm{C_6H_7} + (\mathrm{CH_3})_3\mathrm{COH} \\
c\text{-}\mathrm{C_6H_7} + \mathrm{O_2} \rightleftharpoons p\text{-}\mathrm{C_6H_7OO} \\
c\text{-}\mathrm{C_6H_7} + \mathrm{O_2} \rightleftharpoons o\text{-}\mathrm{C_6H_7OO} \\
o\text{-}\mathrm{C_6H_7OO} \rightarrow \mathrm{C_6H_6} + \mathrm{HO_2} \\
2c\text{-}\mathrm{C_6H_7} \rightarrow \text{Products.}
\end{gathered}$$

Global parameter estimation requires the solution of a global dynamic optimization problem (least-squares minimization), which in turn requires the computation of state bounds for this model. In [163], this was done using Singer's method. Here, the RSI method is applied and the results are compared.

The unknown parameters are the forward rate constants for the second and third reactions, which are only known to lie in the interval $[100, 600]$ $(\mathrm{M}^{-1}\mu\mathrm{s}^{-1})$, and the forward rate constant for the fourth reaction, which is restricted to the interval $[0.001, 50]$ $(\mu\mathrm{s}^{-1})$. That is $\mathbf{u} = (k_2, k_3, k_4)$ and $U \equiv [100, 600] \times [100, 600] \times [0.001, 50]$ (wider parameter ranges are considered in [162], but the global optimization algorithm employed there also involves computing state bounds over subintervals of $U$, so this is a closely related problem, and the state bounds are more clearly illustrated over these ranges). Ordering the variables as

$$\begin{aligned}\mathbf{x} = (&x_{(CH_3)_3CO}, x_{1,4\text{-}C_6H_8}, x_{c\text{-}C_6H_7}, x_{(CH_3)_3COH}\\ &x_{O_2}, x_{p\text{-}C_6H_7OO}, x_{o\text{-}C_6H_7OO}, x_{C_6H_6}, x_{HO_2}, x_{\text{Products}}),\end{aligned}$$

The full kinetic model for this system in an isothermal batch reactor is now described by

$$\mathbf{S} = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & -2 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

and

$$\mathbf{r}(t, \mathbf{p}, \mathbf{z}) = \begin{bmatrix} k_1 z_1 z_2 \\ p_2(z_3 z_5 - (1/K_2) z_6) \\ p_3(z_3 z_5 - (1/K_3) z_7) \\ p_4 z_7 \\ k_5 z_3^2 \end{bmatrix},$$

where the values of the constants $k_1$, $k_5$, $K_2$ and $K_3$ are given in Table 2 of [163] for

298 $K$. The set of admissible initial conditions is the singleton containing

$$\mathbf{x}_0 = (1.53 \times 10^{-4}, 0.4, 0, 0, 0.0019, 0, 0, 0, 0, 0).$$

To construct a reduced model, $\mathbf{D} = \mathbf{S}$ was chosen since $\mathbf{S}$ is full column rank. $\mathbf{A}$ was again chosen as the Moore-Penrose inverse of $\mathbf{D}$, computed by the `MATLAB` `pinv` routine as

$$\mathbf{A} = \left[\begin{matrix} -0.3150 & -0.3150 & 0.0551 & 0.3150 & -0.0394 \\ -0.0157 & -0.0157 & -0.0472 & 0.0157 & -0.2520 \\ -0.0236 & -0.0236 & -0.0709 & 0.0236 & -0.3780 \\ -0.0079 & -0.0079 & -0.0236 & 0.0079 & -0.1260 \\ -0.1102 & -0.1102 & -0.3307 & 0.1102 & 0.2362 \end{matrix}\right.$$

$$\left.\begin{matrix} 0.0157 & 0.0157 & 0.0079 & 0.0079 & 0.1102 \\ 0.7008 & -0.2992 & -0.1496 & -0.1496 & -0.0945 \\ -0.4488 & 0.5512 & 0.2756 & 0.2756 & -0.1417 \\ -0.1496 & -0.1496 & 0.4252 & 0.4252 & -0.0472 \\ -0.0945 & -0.0945 & -0.0472 & -0.0472 & 0.3386 \end{matrix}\right].$$

The natural bounds before refinement by (4.5) and (4.6) were

$$\begin{aligned} X^N = {} & [0, 1.53 \times 10^{-4}] \times [0, 0.4] \times [0, 0.4025] \times [0, 0.4025] \\ & \times [0, 0.4025] \times [0, 0.4025] \times [0, 0.4025] \times [0, 0.4025] \\ & \times [0, 0.4025] \times [0, 0.4025] \text{ (M)}. \end{aligned}$$

The first two upper bounds result from the initial conditions and the fact that neither $(CH_3)_3CO$ or $1,4$-$C_6H_8$ are generated in the network. The remaining upper bounds are set according to the total number of moles in the system initially, since it is clear from the stoichiometry that the total number of moles will never exceed the initial number. In [163], parameter estimation is done by fitting the kinetic model to measured absorbance data. In Figures 4-11 and 4-12, state bounds are shown for two of the principle contributors to the measured absorbance, $x_{p\text{-}C_6H_7OO}$ and $x_{c\text{-}C_6H_7}$. State bounds computed using the RSI method are shown by circles, while those from Singer's method are shown by dashed lines. For $x_{p\text{-}C_6H_7OO}$ (Figure 4-11), the bounds computed using the RSI method are tighter. However, for $x_{c\text{-}C_6H_7}$ Singer's method produces tighter bounds than the RSI method. Theoretically, this is possible for two reasons. First, though the affine reaction invariants are enforced by the construction of the reduced model, it was shown in Section 4.4.2 that the natural bounds $X^N$ can only be enforced approximately. Secondly, the right-hand side functions of the reduced system (4.24) are potentially more involved than those of the original system due to the additional matrix multiplications. Accordingly, the natural interval extensions of the reduced system right-hand side functions may be weaker than those of the original model right-hand sides. As seen in Figure 4-12, these factors may overwhelm the geometric benefit of bounding the reduced system, and weak bounds result. When this occurs, it is very likely that a different reduced system may produce better bounds. That is, one can choose different matrices $\mathbf{A}$ and $\mathbf{D}$ and apply the RSI method again to obtain a different set of state bounds.

Consider the matrices

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & -1 & -1 & -1 \\ -0.5 & -0.5 & 0.5 & 0 & 0 \end{bmatrix},$$

and

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

It is not difficult to verify that $\mathbf{A}$ is a $\{1,2\}$-inverse of $\mathbf{D}$ (not the Moore-Penrose inverse in this case) and that $\mathcal{R}(\mathbf{S}) = \mathcal{R}(\mathbf{D})$. The state bounds computed by the RSI method using these matrices are shown by the crosses in Figures 4-11 and 4-12. In both figures, these state bounds agree exactly with those computed using Singer's method. Of course, this is an improvement for $x_{c\text{-}C_6H_7}$, but not for $x_{p\text{-}C_6H_7OO}$. This should not be surprising from the form of $\mathbf{A}$. Here, $\mathbf{A}$ was chosen so that the reduced variables are a subset of the original state variables, rather than linear combinations. These are often referred to as *reference components* in the literature [66]. If the chosen reference components are capable of describing the full system dynamics, then $\mathbf{D}$ can be computed as $\mathbf{D} = \mathbf{R}(\mathbf{AR})^{-1}$, where $\mathbf{R}$ is a maximal set of linearly independent columns of $\mathbf{S}$. It is worth noting that the state bounds computed using these matrices are at least as good as those computed by Singer's method for every species except $x_{\text{Products}}$.

## 4.6 The Role of Redundancy

In this chapter, we have not provided a direct comparison between the FSI and RSI methods; that is, between the efficacy of using affine reaction invariants in the form of a bounds tightening procedure in the full space, and using them to derive a reduced model for bounding. However, one general principle can be gleaned from Example 4.3.2, where Singer's method and the FSI method were compared, but not the RSI method. By inspection, one reduced model is obvious. If the relation $\mathbf{1}^{\mathrm{T}}(\mathbf{x}(t, T, \mathbf{x}_0) - \mathbf{x}_0) = 0$ is used to eliminate $x_{\mathrm{C}}$ from the model, then resulting reduced system is

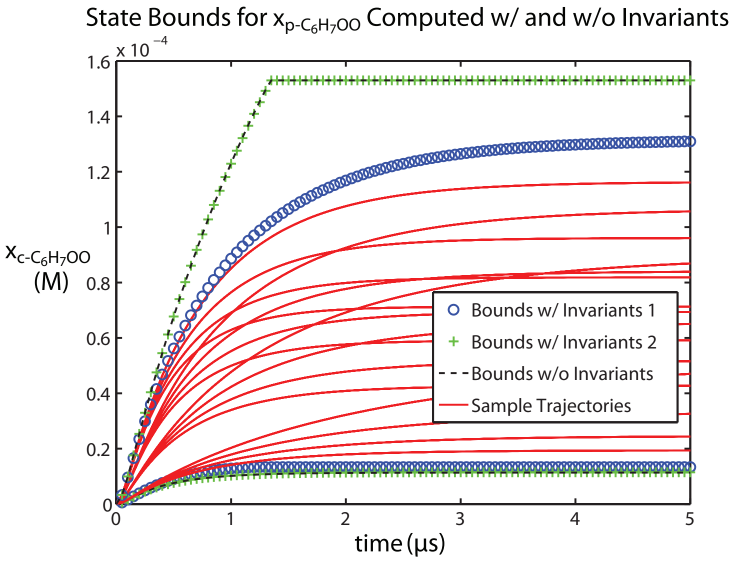

Figure 4-11: State bounds for the species concentration $x_{p\text{-}\mathrm{C_6H_7OO}}$ from Example 4.5.4. Dashed lines are state bounds computed by Singer's method. Two independent sets of state bounds were computed by applying the RSI method with two different pairs of **A** and **D** matrices as input. These bounds are shown by circles and crosses, respectively. Solid curves are true model solutions.

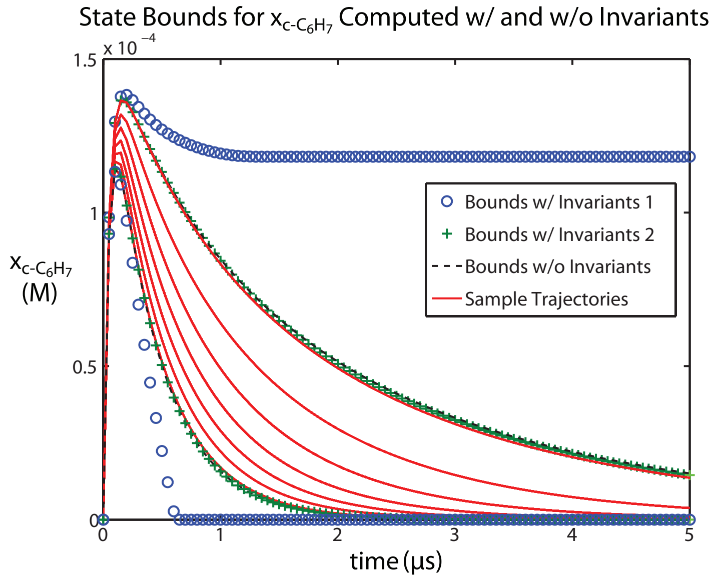

Figure 4-12: State bounds for the species concentration $x_{c\text{-}\mathrm{C_6H_7}}$ from Example 4.5.4. Dashed lines are state bounds computed by Singer's method. Two independent sets of state bounds were computed by applying the RSI method with two different pairs of **A** and **D** matrices as input. These bounds are shown by circles and crosses, respectively. Solid curves are true model solutions.

given simply by the ODEs for $x_{\mathrm{A}}$ and $x_{\mathrm{B}}$, unmodified. However, since elimination of $x_{\mathrm{C}}$ leaves the ODEs for $x_{\mathrm{A}}$ and $x_{\mathrm{B}}$ unchanged, bounding this reduced model will again produce the weak bounds shown by the dashed lines in Figure 4-3. Thus, for this example, FSI is clearly a better method.

One way to understand the efficacy of the FSI method for this example is to note that the reaction invariant is redundant with the three original ODEs. The FSI method makes use of all four redundant equations, whereas the RSI method eliminates one. In interval computations, it is generally known that multiple expressions for the same quantity may result in different enclosures, and that sharper enclosures result from using all available information. In line with this observation, the previous discussion shows that it is better to exploit redundancy in a kinetic model than to eliminate it by reformulation. This is a strong argument for FSI. On the other hand, it is worth exploring whether or not the reduced-space methods improve if some additional redundant equations from the original model are appended to the reduced system for the purposes of bounding. In essence, this leads to a bigger open question: is there any inherent advantage to using a variable transformation in state bounding methods? For Taylor methods, the answer is emphatically in the affirmative [128]. The question has apparently never been explored for differential inequalities methods.

## 4.7 Conclusions and Future Work

In this chapter, the state bounding methods developed in Chapter 3 were applied to ODE models of chemical reaction kinetics. It was shown that the special structure of such models affords a wealth of information constraining possible model solutions. This information takes the form of affine reaction invariants and natural bounds, both of which can be derived through a simple analysis of the stoichiometry matrix. Through several case studies, it was shown that the state bounding method developed for polyhedral *a priori* enclosures in §3.6.2, dubbed the FSI method, makes very effective use of this information. In particular, the FSI method produces bounds that are significantly tighter than those available through Harrison's method or Singer's method. More importantly, these bounds display only mild overestimation when compared to a large sample of true model solutions. Finally, the cost of this method remains small, increasing the cost of Harrison's method by no more than a factor of 10 for any problem considered.

The presence of affine reaction invariants in a chemical kinetics model also implies that the system can be described by a reduced model in a lower dimensional space. Accordingly, a further state bounding method was developed to investigate the advantages of computing bounds on such a reduced model. It was found that this approach is typically superior to Harrison's method and Singer's method, but not to the FSI method. However, a much more thorough comparison between these competing methods is in order, including CPU times. Another observation that warrants further investigation is that this reduced-space method is sensitive to the specific variable transformation used, with some reduced models producing substantially sharper bounds than others.

# Chapter 5

# State Bounding Theory for Semi-Explicit Index-One DAEs

## 5.1 Introduction

In this chapter and the next, two methods are developed for computing interval bounds on the solutions of nonlinear, semi-explicit index-one differential-algebraic equations (DAEs) subject to a given set of initial conditions and model parameters. These parameters may represent uncertain constants in the model, as well as parametrized control inputs or disturbances. As discussed in detail in Chapter 3, computing enclosures of the reachable sets of dynamic systems is a classical problem with a wide variety of applications, including propagating uncertainty through dynamic models [75, 162, 140, 141], solving state and parameter estimation problems [163, 103, 138, 88], safety verification and fault detection in dynamic systems [85, 106], global optimization of dynamic systems [164, 36, 104, 135], validated numerical integration [130], controller design and synthesis [132, 110], and verification of continuous and hybrid systems [176, 40, 70]. However, nearly all available methods apply only to systems of explicit ordinary differential equations (ODEs) (see Chapter 3 for a review of these methods). On the other hand, many dynamic systems encountered in applications are best modeled by DAEs [27, 117].

The state bouding methods developed in this chapter apply to the class of semi-explicit index-one DAEs. The fact that such DAEs are equivalent to an explicit system of ODEs, the so-called underlying ODEs (see Remark 5.3.3), suggests that methods for ODEs could be applied directly. Unfortunately, this turns out to be unworkable because ODE methods require that the right-hand side functions are factorable. For the underlying ODEs, this necessitates an explicit expression for the inverse of the Jacobian of the algebraic equations, which would be very difficult to obtain in general (this requires the construction of the cofactor matrix, which has a factorial number of terms [168]). Moreover, the theoretical reduction to explicit ODEs is only valid locally around a given solution trajectory. This proves problematic for ODE methods because the computed enclosures may come to contain regions of state space on which this reduction is invalid. For these reasons, it is necessary to develop a dedicated theory.

This chapter presents the theoretical developments requried to characterize state bounds for the solutions of DAEs, while Chapter 6 discusses numerical methods. The first theoretical contribution is an interval inclusion test that verifies the existence and uniqueness of a DAE solution within a given interval. This test combines a well-known interval inclusion test for solutions of ODEs (used in standard interval Taylor series bounding methods [130]) with an interval inclusion test for solutions of systems of nonlinear algebraic equations from the literature on interval Newton methods [131]. The second theoretical contribution is a pair of results using differential inequalities to derive bounding trajectories corresponding to the differential state variables; i.e., those state variables whose time derivatives are given explicitly by the DAE equations. Together, these contributions lead to the first bounding method proposed in Chapter 6. The final theoretical contribution is a result combining differential inequalities and interval Newton methods to compute bounds on both the differential and algebraic variables simultaneously. This result leads to the second method described in Chapter 6. Owing to the use of standard numerical integration codes in our implementation, the proposed methods produce enclosures that are mathematically guaranteed but not validated (i.e., they do not account for the numerical error in their computation). However, the existence and uniqueness test described above can be implemented in a validated manner, thus providing a key step towards validated bounding methods for DAEs.

A previous method for bounding the solutions of semi-explicit DAEs was proposed in [142]. This method is not based on differential inequalities, but it does involve an existence and uniqueness test based on an interval Newton method (the interval Krawczyk method). However, rather than combining the interval Krawczyk inclusion test with an interval inclusion tests for ODE solutions, as is done this work, the authors apply the interval Krawczyk inclusion test to the system of nonlinear integral equations obtained by replacing each instance of the differential variables in the original DAEs by the integrals of their time derivatives. The validity of this approach is unclear, since no justification is given for applying an inclusion test for real-valued solutions of algebraic equations to a system of functional equations defined on a function space.

The article [83] presents an algorithm for computing interval bounds on the solutions of implicit ODEs using Taylor models, which can be extended to treat DAEs as well. This method first computes a high-order polynomial approximation of the ODE solution, and then attempts to find a rigorous error bound by satisfying an inclusion test. Satisfying this inclusion test, which uses Taylor models rather than intervals, implies existence and uniqueness of an ODE solution near the polynomial approximation, i.e., within the validated error bound. This algorithm appears capable of computing very tight bounds, but requires the computation of a potentially very large number of Taylor coefficients. This method does not make use of differential inequalities. Furthermore, in addition to the use of Taylor models in place of intervals, the existence and uniqueness test proven in [83] is fundamentally different from the one presented here (and the one used in [142]) because it is derived through direct rearrangement of the implicit ODE equations into fixed-point form, rather than through application of the mean-value theorem, as is done in all interval Newton methods (see Remark 5.4.6).

Finally, in [45], a method for approximating the reachable sets of semi-explicit index-one DAEs is proposed, based on level set methods for ODEs [176]. Methods of this type are designed to provide an accurate approximation of the reachable set, rather than a rigorous enclosure of it. Accordingly, these methods are not appropriate for many applications of interest [163, 138, 85, 106, 164, 36].

The remainder this chapter is organized as follows. Notation and relevant background material is presented in §5.2. Section 5.3 formally describes the DAEs considered and presents basic results. In §5.4, an interval test for existence and uniqueness of solutions is described. Section 5.5 proves three results using differential inequalities to characterize bounding trajectories. Computational implementation of these results and case studies are presented in Chapter 6.

## 5.2 Preliminaries

For any open $D \subset \mathbb{R}^n$, $C^k(D, \mathbb{R}^m)$ denotes the set of $k$-times continuously differentiable mappings from $D$ into $\mathbb{R}^m$. For a general $D \subset \mathbb{R}^n$, $\boldsymbol{\phi} \in C^k(D, \mathbb{R}^m)$ if there exists an open set $\tilde{D} \supset D$ and a function $\tilde{\boldsymbol{\phi}} \in C^k(\tilde{D}, \mathbb{R}^m)$ such that $\tilde{\boldsymbol{\phi}}|_D = \boldsymbol{\phi}$.

The following result is standard ([127], p. 160).

**Lemma 5.2.1.** *Let $D \subset \mathbb{R}^n$ and $\boldsymbol{\phi} \in C^1(D, \mathbb{R}^m)$. Then, for any compact $K \subset D$, $\exists L_K \in \mathbb{R}_+$ such that $\|\boldsymbol{\phi}(\mathbf{z}) - \boldsymbol{\phi}(\hat{\mathbf{z}})\|_1 \leq L_K \|\mathbf{z} - \hat{\mathbf{z}}\|_1$, $\forall (\mathbf{z}, \hat{\mathbf{z}}) \in K \times K$.*

Let $D_s \subset \mathbb{R}^{n_s}$, $D_r \subset \mathbb{R}^{n_r}$, and $\boldsymbol{\ell} \in C^k(D_s \times D_r, \mathbb{R}^{n_r})$ with $k \geq 1$. For any $(\hat{\mathbf{s}}, \hat{\mathbf{r}}) \in D_s \times D_r$, the Jacobian matrix of the mapping $\boldsymbol{\ell}(\hat{\mathbf{s}}, \cdot)$ at $\hat{\mathbf{r}}$ is denoted by $\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}(\hat{\mathbf{s}}, \hat{\mathbf{r}})$. The implicit function theorem is required below and stated here for reference.

**Theorem 5.2.2** (Implicit Function Theorem)**.** *Let $D_s \subset \mathbb{R}^{n_s}$ and $D_r \subset \mathbb{R}^{n_r}$ be open and let $\boldsymbol{\ell} \in C^k(D_s \times D_r, \mathbb{R}^{n_r})$. Suppose that $(\mathbf{s}_0, \mathbf{r}_0) \in D_s \times D_r$ satisfies $\boldsymbol{\ell}(\mathbf{s}_0, \mathbf{r}_0) = \mathbf{0}$ and $\det \frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}(\mathbf{s}_0, \mathbf{r}_0) \neq 0$. Then there exists an open ball around $\mathbf{s}_0$, $V_0 \subset D_s$, an open ball around $\mathbf{r}_0$, $Q_0 \subset D_r$, and $\mathbf{h} \in C^k(V_0, Q_0)$ satisfying*

1. *$\mathbf{h}(\mathbf{s}_0) = \mathbf{r}_0$,*

2. *For any $\mathbf{s} \in V_0$, the vector $\mathbf{r} = \mathbf{h}(\mathbf{s})$ is the unique element of $Q_0$ satisfying $\boldsymbol{\ell}(\mathbf{s}, \mathbf{r}) = \mathbf{0}$,*

*3.* $\det \frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}(\mathbf{s},\mathbf{r}) \neq 0$, $\forall(\mathbf{s},\mathbf{r}) \in V_0 \times Q_0$.

*Proof.* See Theorem 9.2 in [127] and Theorem 9.28 in [145]. □

## 5.3 Problem Statement

In this section, the system of DAEs under consideration is defined and the problem of computing interval bounds is stated formally. Because we are interested in computing interval enclosures of the possible solutions of this system, it is necessary to have clear statements of the existence and uniqueness properties of these solutions. The basic local existence result is well-known [96] and is not proven here. On the other hand, certain arguments in this work require very particular properties related to uniqueness, so the relevant analysis is provided. Detailed proofs are relegated to §5.3.4.

### 5.3.1 Semi-explicit DAEs

Let $D_t \subset \mathbb{R}$, $D_p \subset \mathbb{R}^{n_p}$, $D_x \subset \mathbb{R}^{n_x}$ and $D_y \subset \mathbb{R}^{n_y}$ be open sets, and let $\mathbf{f} : D_t \times D_p \times D_x \times D_y \to \mathbb{R}^{n_x}$, $\mathbf{g} : D_t \times D_p \times D_x \times D_y \to \mathbb{R}^{n_y}$ and $\mathbf{x}_0 : D_p \to D_x$ be $C^1$ functions. Given some $t_0 \in D_t$, consider the initial value problem in semi-explicit differential-algebraic equations

$$\left.\begin{aligned} \dot{\mathbf{x}}(t,\mathbf{p}) &= \mathbf{f}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p})) \\ \mathbf{0} &= \mathbf{g}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p})) \end{aligned}\right\}, \tag{5.1a}$$

$$\mathbf{x}(t_0,\mathbf{p}) = \mathbf{x}_0(\mathbf{p}), \tag{5.1b}$$

where $t$ is the independent variable, $\mathbf{p}$ is a vector of problem parameters, $\dot{\mathbf{x}}(t,\mathbf{p})$ denotes the derivative of $\mathbf{x}(\cdot,\mathbf{p})$ at $t$, and $\mathbf{x}_0$ specifies the parametric initial conditions. A solution of (5.1) is defined below.

**Definition 5.3.1.** Define the sets

$$\begin{aligned}
\mathcal{G} &\equiv \{(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) \in D_t \times D_p \times D_x \times D_y : \mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}\}, \\
\mathcal{G}_0 &\equiv \{(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) \in \mathcal{G} : \mathbf{x}_0(\mathbf{p}) = \mathbf{z}_x\}, \\
\mathcal{G}_R &\equiv \{(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) \in D_t \times D_p \times D_x \times D_y : \det\frac{\partial \mathbf{g}}{\partial \mathbf{y}}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) \neq 0\}.
\end{aligned}$$

**Definition 5.3.2.** Let $I \subset D_t$ be connected, and let $P \subset D_p$. A function $(\mathbf{x}, \mathbf{y}) \in C^1(I \times P, D_x) \times C^1(I \times P, D_y)$ is called a *solution of* (5.1a) *on* $I \times P$ if (5.1a) holds for all $(t, \mathbf{p}) \in I \times P$. If in addition $(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p})) \in \mathcal{G}_R$, $\forall (t, \mathbf{p}) \in I \times P$, then $(\mathbf{x}, \mathbf{y})$ is called *regular*. When $t_0 \in I$ is specified and $\mathbf{x}$ also satisfies (5.1b), $(\mathbf{x}, \mathbf{y})$ it is called a (regular) solution of (5.1) on $I \times P$.

**Remark 5.3.3.** In this thesis, the assumption that (5.1) has differential index 1 is not stated directly, but rather implied by restricting our results to *regular* solutions, as defined above. Indeed, these notions are identical in this case, since, for any regular solution of (5.1) on $I \times P$, a single differentiation of the algebraic equations $\mathbf{g}$ gives the underlying ODEs

$$\dot{\mathbf{x}}(t, \mathbf{p}) = \mathbf{f}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p})), \tag{5.2}$$

$$\dot{\mathbf{y}}(t, \mathbf{p}) = -\left(\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\right)^{-1} \left(\frac{\partial \mathbf{g}}{\partial \mathbf{x}} \mathbf{f}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p})) + \frac{\partial \mathbf{g}}{\partial t}\right), \tag{5.3}$$

for all $(t, \mathbf{p}) \in I \times P$, where all partial derivatives of $\mathbf{g}$ are evaluated at $(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p}))$.

### 5.3.2 Existence and Uniqueness

Existence of a solution of (5.1) can of course only be guaranteed locally. The main result is stated in terms of local solutions, defined as follows.

**Definition 5.3.4.** For any $(t_0, \hat{\mathbf{p}}, \hat{\mathbf{x}}_0, \hat{\mathbf{y}}_0) \in \mathcal{G}_0$, a mapping $(\mathbf{x}, \mathbf{y}) \in C^1(I' \times P', D_x) \times C^1(I' \times P', D_y)$ is called a *solution* of (5.1) *local to* $(t_0, \hat{\mathbf{p}}, \hat{\mathbf{x}}_0, \hat{\mathbf{y}}_0)$ if $I'$ and $P'$ are open balls containing $t_0$ and $\hat{\mathbf{p}}$, respectively, $\mathbf{x}$ and $\mathbf{y}$ satisfy (5.1) on $I' \times P'$, and $\mathbf{y}(t_0, \hat{\mathbf{p}}) =$

$\hat{\mathbf{y}}_0$. If in addition $\mathbf{x}$ and $\mathbf{y}$ satisfy $(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p})) \in \mathcal{G}_R$, $\forall (t, \mathbf{p}) \in I' \times P'$, then $(\mathbf{x}, \mathbf{y})$ is called *regular*.

**Theorem 5.3.5.** *Let $(t_0, \hat{\mathbf{p}}, \hat{\mathbf{x}}_0, \hat{\mathbf{y}}_0) \in \mathcal{G}_0 \cap \mathcal{G}_R$. There exists a regular solution of* (5.1) *local to $(t_0, \hat{\mathbf{p}}, \hat{\mathbf{x}}_0, \hat{\mathbf{y}}_0)$.*

*Proof.* See Theorems 4.13 and 4.18 in [96]. □

For any $(\mathbf{x}, \mathbf{y}) \in C^1(I' \times P', D_x) \times C^1(I' \times P', D_y)$ satisfying (5.1), the initial value of $\mathbf{y}$ must obviously satisfy $\mathbf{g}(t_0, \mathbf{p}, \mathbf{x}(t_0, \mathbf{p}), \mathbf{y}(t_0, \mathbf{p})) = \mathbf{0}$ for each $\mathbf{p} \in P'$. Therefore, these values cannot be specified arbitrarily. On the other hand, this equation may have multiple solutions in $D_y$, so that in general more information (in addition to (5.1)) is required to specify a solution uniquely. As will be shown below, uniqueness of regular local solutions follows from the additional condition $\mathbf{y}(t_0, \hat{\mathbf{p}}) = \hat{\mathbf{y}}_0$ in Definition 5.3.4. The following example demonstrates that uniqueness is not guaranteed in the absence of this condition.

**Example 5.3.1.** Let $I \equiv [0, \delta] \subset D_t = \mathbb{R}$, $D_p = \emptyset$, $D_x = D_y = \mathbb{R}$, and define $g(t, z_x, z_y) = z_y^2 - z_x$. With fixed initial condition $x_0 = 1$ at $t_0 = 0$, there are two possible values for $y(t_0)$ satisfying $g(t_0, x(t_0), y(t_0)) = 0$; $y(t_0) = 1$ and $y(t_0) = -1$. Letting $f(t, z_x, z_y) = 1$, clearly $x(t) = 1 + t$ satisfies $\dot{x}(t) = 1 = f(t, x(t), y(t))$ for any $y : I \to \mathbb{R}$. However, both $y(t) = \sqrt{1+t}$ and $y(t) = -\sqrt{1+t}$ result in $g(t, x(t), y(t)) = (y(t))^2 - x(t) = 0$. In particular, $y(t) = \sqrt{1+t}$ is a solution of (5.1) local to $(t_0, \hat{x}_0, \hat{y}_0) = (0, 1, 1)$, while $y(t) = -\sqrt{1+t}$ is a solution of (5.1) local to $(t_0, \hat{x}_0, \hat{y}_0) = (0, 1, -1)$.

A detailed analysis of the uniqueness properties of solutions of (5.1) is given in §5.3.4. The most relevant conclusion is the following.

**Corollary 5.3.6.** *Let $(\mathbf{x}, \mathbf{y}) \in C^1(I \times P, D_x) \times C^1(I \times P, D_y)$ and $(\mathbf{x}^*, \mathbf{y}^*) \in C^1(\tilde{I} \times \tilde{P}, D_x) \times C^1(\tilde{I} \times \tilde{P}, D_y)$ be solutions of* (5.1) *on $I \times P$ and $\tilde{I} \times \tilde{P}$, respectively, with some $t_0 \in I \cap \tilde{I}$, and suppose that $(\mathbf{x}, \mathbf{y})$ is regular. If $\hat{P} \subset P \cap \tilde{P}$ is connected and $\exists \hat{\mathbf{p}} \in \hat{P}$ such that $\mathbf{y}(t_0, \hat{\mathbf{p}}) = \mathbf{y}^*(t_0, \hat{\mathbf{p}})$, then $\mathbf{x}(t, \mathbf{p}) = \mathbf{x}^*(t, \mathbf{p})$ and $\mathbf{y}(t, \mathbf{p}) = \mathbf{y}^*(t, \mathbf{p})$, $\forall (t, \mathbf{p}) \in (I \cap \tilde{I}) \times \hat{P}$.*

*Proof.* See §5.3.4. □

### 5.3.3 State Bounds

The primary aim of this chapter and the next is to compute interval bounds for the solutions of (5.1). Let $I = [t_0, t_f] \subset D_t$ and $P \subset D_p$ be intervals and suppose that $(\mathbf{x}, \mathbf{y}) \in C^1(I \times P, D_x) \times C^1(I \times P, D_y)$ is a regular solution of (5.1) on $I \times P$. Then, our objective is to compute functions $\mathbf{x}^L, \mathbf{x}^U : I \to \mathbb{R}^{n_x}$ and $\mathbf{y}^L, \mathbf{y}^U : I \to \mathbb{R}^{n_y}$ such that

$$\mathbf{x}^L(t) \leq \mathbf{x}(t, \mathbf{p}) \leq \mathbf{x}^U(t) \quad \text{and} \quad \mathbf{y}^L(t) \leq \mathbf{y}(t, \mathbf{p}) \leq \mathbf{y}^U(t), \quad \forall (t, \mathbf{p}) \in I \times P.$$

These functions are referred to as *state bounds* for the solution $(\mathbf{x}, \mathbf{y})$.

Recall that (5.1) may have multiple regular solutions on $I \times P$ corresponding to different solution branches of the algebraic equations (see Example 5.3.1). In the methods of this chapter, a single solution is specified for bounding through an interval, either provided as input or computed, which, for each $\mathbf{p} \in P$, contains exactly one initial condition for $\mathbf{y}$ which is consistent with $\mathbf{x}_0(\mathbf{p})$ (see Theorem 5.4.8). This interval specifies which solution branch defines $\mathbf{y}$ at $t_0$, and hence the solution is uniquely determined on $I \times P$ (Corollary 5.3.6). In principle, Theorem 5.5.2 provides bounds valid for all regular solutions of (5.1), but we do not pursue a method for computing such bounds.

In order to compute state bounds, we will make use of inclusion monotonic interval extensions of the functions $\mathbf{f}$, $\mathbf{g}$ and $\frac{\partial \mathbf{g}}{\partial \mathbf{y}}$. It will be assumed throughout that such functions are available and, for convenience, that they are defined on all of $\mathbb{I}D_t \times \mathbb{I}D_p \times \mathbb{I}D_x \times \mathbb{I}D_y$. Of course, if $\mathbf{f}$, $\mathbf{g}$ and $\frac{\partial \mathbf{g}}{\partial \mathbf{y}}$ are $\mathcal{L}$-factorable, then the natural interval extensions of §2.3.2 may be used.

### 5.3.4 Uniqueness Proofs

**Lemma 5.3.7.** *Let $E \subset \mathbb{R}^n$ be connected and let $\psi : E \to \mathbb{R}$ be continuous. If the set $\{\boldsymbol{\xi} \in E : \psi(\boldsymbol{\xi}) = 0\}$ is nonempty and open with respect to $E$, then $\psi(\boldsymbol{\xi}) = 0$, $\forall \boldsymbol{\xi} \in E$.*

*Proof.* Let $E_1 = \{\boldsymbol{\xi} \in E : \psi(\boldsymbol{\xi}) = 0\}$ and $E_2 = \{\boldsymbol{\xi} \in E : \psi(\boldsymbol{\xi}) \neq 0\}$, and note that $E_1 \cap E_2 = \emptyset$ and $E_1 \cup E_2 = E$. Since $E$ is connected, it cannot be written as the disjoint union of two nonempty open (w.r.t. $E$) sets. But $E_1$ is nonempty and open w.r.t. $E$ by hypothesis, and $E_2$ is open w.r.t. $E$ because it is the inverse image of an open set under a continuous mapping on $E$. Hence, $E_2 = \emptyset$ and $E_1 = E$. □

**Lemma 5.3.8.** *Let $(\mathbf{x}, \mathbf{y}) \in C^1(I \times P, D_x) \times C^1(I \times P, D_y)$ and $(\mathbf{x}^*, \mathbf{y}^*) \in C^1(\tilde{I} \times \tilde{P}, D_x) \times C^1(\tilde{I} \times \tilde{P}, D_y)$ be solutions of* (5.1a) *on $I \times P$ and $\tilde{I} \times \tilde{P}$, respectively, and suppose that $(\mathbf{x}, \mathbf{y})$ is regular. Then*

1. *For any $(t', \mathbf{p}') \in I \times P$, there exists an open ball around $(t', \mathbf{p}')$, $U' \subset D_t \times D_p$, an open ball around $(t', \mathbf{p}', \mathbf{x}(t', \mathbf{p}'))$, $V' \subset D_t \times D_p \times D_x$, an open ball around $\mathbf{y}(t', \mathbf{p}')$, $Q' \subset D_y$, and a function $\mathbf{h} \in C^1(V', Q')$ satisfying $(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})) \in V'$ and $\mathbf{y}(t, \mathbf{p}) = \mathbf{h}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})) \in Q'$, $\forall (t, \mathbf{p}) \in U' \cap (I \times P)$.*

2. *If $\hat{P} \subset P \cap \tilde{P}$ is connected and $\exists (t', \hat{\mathbf{p}}) \in (I \cap \tilde{I}) \times \hat{P}$ such that $\mathbf{x}(t', \mathbf{p}) = \mathbf{x}^*(t', \mathbf{p})$, $\forall \mathbf{p} \in \hat{P}$, and $\mathbf{y}(t', \hat{\mathbf{p}}) = \mathbf{y}^*(t', \hat{\mathbf{p}})$, then $\mathbf{y}(t', \mathbf{p}) = \mathbf{y}^*(t', \mathbf{p})$, $\forall \mathbf{p} \in \hat{P}$.*

*Proof.* Choose any $(t', \mathbf{p}') \in I \times P$. Since $(\mathbf{x}, \mathbf{y})$ is a regular solution of (5.1a) on $I \times P$, $(t', \mathbf{p}', \mathbf{x}(t', \mathbf{p}'), \mathbf{y}(t', \mathbf{p}')) \in \mathcal{G} \cap \mathcal{G}_R$. Then, by Theorem 5.2.2, there exists an open ball around $(t', \mathbf{p}', \mathbf{x}(t', \mathbf{p}'))$, $V' \subset D_t \times D_p \times D_x$, an open ball around $\mathbf{y}(t', \mathbf{p}')$, $Q' \subset D_y$, and a function $\mathbf{h} \in C^1(V', Q')$ such that $\mathbf{h}(t', \mathbf{p}', \mathbf{x}(t', \mathbf{p}')) = \mathbf{y}(t', \mathbf{p}')$ and, for each $(t, \mathbf{p}, \mathbf{z}_x) \in V'$, $\mathbf{h}(t, \mathbf{p}, \mathbf{z}_x)$ is the unique element of $Q'$ satisfying $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{h}(t, \mathbf{p}, \mathbf{z}_x)) = \mathbf{0}$. Now, by continuity, there exists an open ball $U'$ around the point $(t', \mathbf{p}')$ small enough that $(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})) \in V'$ for every $(t, \mathbf{p}) \in U' \cap (I \times P)$, and it follows that

$$\mathbf{g}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{h}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}))) = \mathbf{0}, \quad \forall (t, \mathbf{p}) \in U' \cap (I \times P). \tag{5.4}$$

Again by continuity, it is possible to choose $U'$ small enough that $\mathbf{y}(t, \mathbf{p}) \in Q'$ for all $(t, \mathbf{p}) \in U' \cap (I \times P)$, which implies, by the uniqueness property of $\mathbf{h}$ in $Q'$, that

$$\mathbf{y}(t, \mathbf{p}) = \mathbf{h}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})), \quad \forall (t, \mathbf{p}) \in U' \cap (I \times P). \tag{5.5}$$

This establishes the first conclusion of the lemma.

To prove the second conclusion, choose any $\hat{P}$, $\hat{\mathbf{p}}$ and $t'$ as in the hypothesis of the lemma and define

$$R \equiv \{\mathbf{p} \in \hat{P} : \|\mathbf{y}(t', \mathbf{p}) - \mathbf{y}^*(t', \mathbf{p})\| = 0\}. \tag{5.6}$$

By hypothesis, $\hat{\mathbf{p}} \in R$ so that $R$ is nonempty. It will be shown than $R$ is open with respect to $\hat{P}$. Choose any $\mathbf{p}' \in R$ and, corresponding to the point $(t', \mathbf{p}')$, let $U'$, $V'$, $Q'$ and $\mathbf{h}$ be as in the first conclusion of the lemma. By hypothesis, $(t', \mathbf{p}', \mathbf{x}^*(t', \mathbf{p}')) = (t', \mathbf{p}', \mathbf{x}(t', \mathbf{p}')) \in V'$, and by the definition of $R$, $\mathbf{y}^*(t', \mathbf{p}') = \mathbf{y}(t', \mathbf{p}') \in Q'$, so continuity implies that we may choose an open all around $\mathbf{p}'$, $J_{\mathbf{p}'}$, small enough that $J_{\mathbf{p}'} \times \{t'\} \subset U'$, and $(t', \mathbf{p}, \mathbf{x}^*(t', \mathbf{p})) \in V'$ and $\mathbf{y}^*(t', \mathbf{p}) \in Q'$, for all $\mathbf{p} \in J_{\mathbf{p}'} \cap \tilde{P}$. Then the first conclusion of the theorem gives

$$\mathbf{y}(t', \mathbf{p}) = \mathbf{h}(t', \mathbf{p}, \mathbf{x}(t', \mathbf{p})), \quad \forall \mathbf{p} \in J_{\mathbf{p}'} \cap \hat{P}, \tag{5.7}$$

and an identical argument shows that

$$\mathbf{y}^*(t', \mathbf{p}) = \mathbf{h}(t', \mathbf{p}, \mathbf{x}^*(t', \mathbf{p})), \quad \forall \mathbf{p} \in J_{\mathbf{p}'} \cap \hat{P}. \tag{5.8}$$

But $\mathbf{x}^*(t', \mathbf{p}) = \mathbf{x}(t', \mathbf{p})$, $\forall \mathbf{p} \in \hat{P}$ by hypothesis, so this implies that $\mathbf{y}^*(t', \mathbf{p}) = \mathbf{y}(t', \mathbf{p})$, $\forall \mathbf{p} \in J_{\mathbf{p}'} \cap \hat{P}$. Thus $R$ is open with respect to $\hat{P}$. Now, since $\hat{P}$ is connected by hypothesis and $R$ is nonempty and open with respect to $\hat{P}$, Lemma 5.3.7 shows that $R = \hat{P}$; i.e. $\mathbf{y}^*(t', \mathbf{p}) = \mathbf{y}(t', \mathbf{p})$, $\forall \mathbf{p} \in \hat{P}$. □

**Lemma 5.3.9.** *Let* $(\mathbf{x}, \mathbf{y}) \in C^1(I \times P, D_x) \times C^1(I \times P, D_y)$ *and* $(\mathbf{x}^*, \mathbf{y}^*) \in C^1(\tilde{I} \times \tilde{P}, D_x) \times C^1(\tilde{I} \times \tilde{P}, D_y)$ *be solutions of* (5.1a) *on* $I \times P$ *and* $\tilde{I} \times \tilde{P}$*, respectively, and suppose that* $(\mathbf{x}, \mathbf{y})$ *is regular. If* $\hat{P} \subset P \cap \tilde{P}$ *is connected and compact and* $\exists(\hat{t}, \hat{\mathbf{p}}) \in (I \cap \tilde{I}) \times \hat{P}$ *such that* $\mathbf{x}(\hat{t}, \mathbf{p}) = \mathbf{x}^*(\hat{t}, \mathbf{p})$*,* $\forall \mathbf{p} \in \hat{P}$*, and* $\mathbf{y}(\hat{t}, \hat{\mathbf{p}}) = \mathbf{y}^*(\hat{t}, \hat{\mathbf{p}})$*, then* $\mathbf{x}(t, \mathbf{p}) = \mathbf{x}^*(t, \mathbf{p})$ *and* $\mathbf{y}(t, \mathbf{p}) = \mathbf{y}^*(t, \mathbf{p})$*,* $\forall (t, \mathbf{p}) \in (I \cap \tilde{I}) \times \hat{P}$*.*

*Proof.* Choose any $\hat{P}$, $\hat{\mathbf{p}}$ and $\hat{t}$ as in the hypothesis of the lemma and define

$$R \equiv \{t \in I \cap \tilde{I} : \max_{\mathbf{p} \in \hat{P}} (\|\mathbf{x}(t, \mathbf{p}) - \mathbf{x}^*(t, \mathbf{p})\|) + \|\mathbf{y}(t, \hat{\mathbf{p}}) - \mathbf{y}^*(t, \hat{\mathbf{p}})\| = 0\}. \tag{5.9}$$

$R$ is nonempty since it contains $\hat{t}$. It will be shown that $R$ is open with respect to $I \cap \tilde{I}$. Choose any $t' \in R$. Applying the second conclusion of Lemma 5.3.8, we have $\mathbf{y}^*(t', \mathbf{p}) = \mathbf{y}(t', \mathbf{p})$, $\forall \mathbf{p} \in \hat{P}$. Choose any $\mathbf{p}' \in \hat{P}$ and, corresponding to the point $(t', \mathbf{p}')$, let $U'$, $V'$, $Q'$ and $\mathbf{h}$ be as in the first conclusion of Lemma 5.3.8. By the definition of $R$, $(t', \mathbf{p}', \mathbf{x}^*(t', \mathbf{p}')) = (t', \mathbf{p}', \mathbf{x}(t', \mathbf{p}')) \in V'$ and, by the argument above, $\mathbf{y}^*(t', \mathbf{p}') = \mathbf{y}(t', \mathbf{p}') \in Q'$. Then continuity implies that there exists an open ball around $t'$, $J_{t'}$, and an open ball around $\mathbf{p}'$, $J_{\mathbf{p}'}$, such that $J_{t'} \times J_{\mathbf{p}'} \subset U'$, and $(t, \mathbf{p}, \mathbf{x}^*(t, \mathbf{p})) \in V'$ and $\mathbf{y}^*(t, \mathbf{p}) \in Q'$, for all $(t, \mathbf{p}) \in (J_{t'} \times J_{\mathbf{p}'}) \cap (\tilde{I} \times \tilde{P})$. From Lemma 5.3.8, we have

$$\mathbf{y}(t, \mathbf{p}) = \mathbf{h}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})), \quad \forall (t, \mathbf{p}) \in (J_{t'} \times J_{\mathbf{p}'}) \cap (I \times \hat{P}), \tag{5.10}$$

and an identical argument using the uniqueness property of $\mathbf{h}$ in $Q'$ shows that

$$\mathbf{y}^*(t, \mathbf{p}) = \mathbf{h}(t, \mathbf{p}, \mathbf{x}^*(t, \mathbf{p})), \quad \forall (t, \mathbf{p}) \in (J_{t'} \times J_{\mathbf{p}'}) \cap (\tilde{I} \times \hat{P}). \tag{5.11}$$

Then, by definition,

$$\dot{\mathbf{x}}(t, \mathbf{p}) = \mathbf{f}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{h}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}))), \quad \forall (t, \mathbf{p}) \in (J_{t'} \times J_{\mathbf{p}'}) \cap (I \times \hat{P}), \tag{5.12}$$

$$\dot{\mathbf{x}}^*(t, \mathbf{p}) = \mathbf{f}(t, \mathbf{p}, \mathbf{x}^*(t, \mathbf{p}), \mathbf{h}(t, \mathbf{p}, \mathbf{x}^*(t, \mathbf{p}))), \quad \forall (t, \mathbf{p}) \in (J_{t'} \times J_{\mathbf{p}'}) \cap (\tilde{I} \times \hat{P}). \tag{5.13}$$

But $\mathbf{f}$ and $\mathbf{h}$ are continuously differentiable and hence the mapping $(t, \mathbf{p}, \mathbf{z}_x) \mapsto \mathbf{f}(t, \mathbf{p}, \mathbf{h}(t, \mathbf{p}, \mathbf{z}_x))$ is Lipschitz on $V'$ by Lemma 5.2.1. The definition of $R$ gives $\mathbf{x}(t', \mathbf{p}) = \mathbf{x}^*(t', \mathbf{p})$, $\forall \mathbf{p} \in \hat{P}$, so a standard application of Gronwall's inequality shows that $\mathbf{x}(t, \mathbf{p}) = \mathbf{x}^*(t, \mathbf{p})$, $\forall (t, \mathbf{p}) \in (J_{t'} \times J_{\mathbf{p}'}) \cap ((I \cap \tilde{I}) \times \hat{P})$. Furthermore, this implies that $\mathbf{y}(t, \mathbf{p}) = \mathbf{h}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})) = \mathbf{h}(t, \mathbf{p}, \mathbf{x}^*(t, \mathbf{p})) = \mathbf{y}^*(t, \mathbf{p})$, $\forall (t, \mathbf{p}) \in (J_{t'} \times J_{\mathbf{p}'}) \cap ((I \cap \tilde{I}) \times \hat{P})$.

Now, since $\mathbf{p}' \in \hat{P}$ was chosen arbitrarily, the preceding construction applies to every $\mathbf{p} \in \hat{P}$. Thus, to every $\mathbf{q} \in \hat{P}$, there corresponds an open ball around $t'$, $J_{t'}(\mathbf{q})$, and an open ball around $\mathbf{q}$, $J_{\mathbf{q}}$, such that $(\mathbf{x},\mathbf{y})(t,\mathbf{p}) = (\mathbf{x}^*,\mathbf{y}^*)(t,\mathbf{p})$, $\forall(t,\mathbf{p}) \in (J_{t'}(\mathbf{q}) \times J_{\mathbf{q}}) \cap ((I \cap \tilde{I}) \times \hat{P})$. Noting that the $J_{\mathbf{q}}$ constructed in this way form an open cover of $\hat{P}$, compactness of $\hat{P}$ implies that there exist finitely many elements of $\hat{P}$, $\mathbf{q}_1, \ldots, \mathbf{q}_n$, such that $\hat{P}$ is covered by $J_{\mathbf{q}_1} \cup \ldots \cup J_{\mathbf{q}_n}$. Let $J^*_{t'} \equiv J_{t'}(\mathbf{q}_1) \cap \ldots \cap J_{t'}(\mathbf{q}_n)$. Then, for every $\mathbf{p} \in \hat{P}$, there exists $i \in \{1, \ldots, n\}$ such that $\mathbf{p} \in J_{\mathbf{q}_i}$, which implies that $(\mathbf{x},\mathbf{y})(t,\mathbf{p}) = (\mathbf{x}^*,\mathbf{y}^*)(t,\mathbf{p})$, $\forall t \in J^*_{t'} \cap (I \cap \tilde{I})$. Therefore, $J^*_{t'} \cap (I \cap \tilde{I})$ is contained in $R$, so that $t'$ is an interior point of $R$ when viewed as a subset of $I \cap \tilde{I}$, and since $t' \in R$ was chosen arbitrarily, $R$ is open with respect to $I \cap \tilde{I}$. Since $I \cap \tilde{I}$ is connected and $R$ is nonempty and open with respect to $I \cap \tilde{I}$, Lemma 5.3.7 shows that $R = I \cap \tilde{I}$. But by definition, this implies that $\mathbf{x}(t,\mathbf{p}) = \mathbf{x}^*(t,\mathbf{p})$ and $\mathbf{y}(t,\hat{\mathbf{p}}) = \mathbf{y}^*(t,\hat{\mathbf{p}})$, $\forall(t,\mathbf{p}) \in (I \cap \tilde{I}) \times \hat{P}$. Finally, the second conclusion of Lemma 5.3.8 implies that $\mathbf{y}(t,\mathbf{p}) = \mathbf{y}^*(t,\mathbf{p})$, $\forall(t,\mathbf{p}) \in (I \cap \tilde{I}) \times \hat{P}$. □

**Theorem 5.3.10.** *Let $(\mathbf{x},\mathbf{y}) \in C^1(I \times P, D_x) \times C^1(I \times P, D_y)$ and $(\mathbf{x}^*,\mathbf{y}^*) \in C^1(\tilde{I} \times \tilde{P}, D_x) \times C^1(\tilde{I} \times \tilde{P}, D_y)$ be solutions of* (5.1a) *on $I \times P$ and $\tilde{I} \times \tilde{P}$, respectively, and suppose that $(\mathbf{x},\mathbf{y})$ is regular. If $\hat{P} \subset P \cap \tilde{P}$ is connected and $\exists(\hat{t},\hat{\mathbf{p}}) \in (I \cap \tilde{I}) \times \hat{P}$ such that $\mathbf{x}(\hat{t},\mathbf{p}) = \mathbf{x}^*(\hat{t},\mathbf{p})$, $\forall \mathbf{p} \in \hat{P}$, and $\mathbf{y}(\hat{t},\hat{\mathbf{p}}) = \mathbf{y}^*(\hat{t},\hat{\mathbf{p}})$, then $\mathbf{x}(t,\mathbf{p}) = \mathbf{x}^*(t,\mathbf{p})$ and $\mathbf{y}(t,\mathbf{p}) = \mathbf{y}^*(t,\mathbf{p})$, $\forall(t,\mathbf{p}) \in (I \cap \tilde{I}) \times \hat{P}$.*

*Proof.* Choose any $\mathbf{p} \in \hat{P}$. Clearly, $\{\mathbf{p}\} \subset P \cap \tilde{P}$ is compact and connected, and Lemma 5.3.8 guarantees that $\mathbf{y}(\hat{t},\mathbf{p}) = \mathbf{y}^*(\hat{t},\mathbf{p})$. Then Lemma 5.3.9 shows that $\mathbf{x}(t,\mathbf{p}) = \mathbf{x}^*(t,\mathbf{p})$ and $\mathbf{y}(t,\mathbf{p}) = \mathbf{y}^*(t,\mathbf{p})$, $\forall t \in I \cap \tilde{I}$. □

Corollary 5.3.6 is a simple consequence of these developments. By the definition of a solution of (5.1), we have $\mathbf{x}(t_0,\mathbf{p}) = \mathbf{x}^*(t_0,\mathbf{p})$, $\forall \mathbf{p} \in \hat{P}$, and $\mathbf{y}(t_0,\hat{\mathbf{p}}) = \mathbf{y}^*(t_0,\hat{\mathbf{p}})$ by hypothesis. Since $\hat{P}$ is connected, the result follows from Theorem 5.3.10.

## 5.4 An Interval Inclusion Test for DAE Solutions

This section presents an interval inclusion test which can computationally guarantee the existence and uniqueness of a solution of (5.1) over intervals $I'$ and $P'$ satisfying the test. When successful, the test provides intervals which are guaranteed to enclose the solutions $\mathbf{x}$ and $\mathbf{y}$ on $I' \times P'$. This test is very similar to the Phase 1 step of standard interval Taylor series bounding methods for ODEs [130]. The complicating factor here is of course the presence of the algebraic variables $\mathbf{y}$ and the fact that they are defined implicitly. To overcome this obstacle, a well-known interval inclusion test for existence and uniqueness of solutions of systems of nonlinear algebraic equations is used. This inclusion test is based on the interval Hansen-Sengupta method [131]. This method is described below, and its application to DAEs is discussed in §5.4.2.

### 5.4.1 The Interval Hansen-Sengupta Method

Let $D_s \subset \mathbb{R}^{n_s}$ and $D_r \subset \mathbb{R}^{n_r}$ be open, and let $\boldsymbol{\ell} \in C^k(D_s \times D_r, \mathbb{R}^{n_r})$. Furthermore, assume that inclusion monotonic interval extensions of $\mathbf{r}$ and $\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}$ are available, $[\mathbf{r}]$ and $\left[\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}\right]$, and are defined on all of $\mathbb{I}D_s \times \mathbb{I}D_r$. Given intervals $S \subset D_s$ and $R \subset D_r$, we are concerned with (i) determining if there exist points $\mathbf{r} \in R$ such that $\boldsymbol{\ell}(\mathbf{s}, \mathbf{r}) = \mathbf{0}$ for some $\mathbf{s} \in S$, and (ii) computing a refined interval $R' \subset R$ which contains all such $\mathbf{r}$. Conceptually, this is done by using the mean value theorem to characterize the zeros of $\boldsymbol{\ell}$. For any $(\mathbf{s}, \mathbf{r}) \in S \times R$ such that $\boldsymbol{\ell}(\mathbf{s}, \mathbf{r}) = \mathbf{0}$, any $\tilde{\mathbf{r}} \in R$, $\tilde{\mathbf{r}} \neq \mathbf{r}$, and any index $i$, the mean value theorem states that $\exists \boldsymbol{\xi}^{[i]} \in R$ such that $\boldsymbol{\xi}^{[i]} = \tilde{\mathbf{r}} + \lambda(\mathbf{r} - \tilde{\mathbf{r}})$ for some $\lambda \in (0, 1)$, and

$$\frac{\partial \ell_i}{\partial \mathbf{r}}(\mathbf{s}, \boldsymbol{\xi}^{[i]})\,(\mathbf{r} - \tilde{\mathbf{r}}) = -\ell_i(\mathbf{s}, \tilde{\mathbf{r}}). \tag{5.14}$$

Noting that $\boldsymbol{\xi}^{[i]} \in R$ because $\boldsymbol{\xi}^{[i]} = \tilde{\mathbf{r}} + \lambda(\mathbf{r} - \tilde{\mathbf{r}})$ and $\mathbf{r}, \tilde{\mathbf{r}} \in R$, consider the interval linear equations

$$\left[\frac{\partial \ell_i}{\partial \mathbf{r}}\right](S, R)\,(\mathbf{r} - \tilde{\mathbf{r}}) = -\left[\ell_i\right](S, \tilde{\mathbf{r}}), \tag{5.15}$$

which can be written in matrix form, preconditioned by any $\mathbf{C} \in \mathbb{R}^{n_r \times n_r}$, as

$$\mathbf{C}\left[\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}\right](S, R)\,(\mathbf{r} - \tilde{\mathbf{r}}) = -\mathbf{C}\,[\boldsymbol{\ell}]\,(S, \tilde{\mathbf{r}}). \tag{5.16}$$

The solution set of (5.16) is the set of all $\boldsymbol{\rho} \in \mathbb{R}^{n_r}$ such that $\mathbf{A}\boldsymbol{\rho} = \mathbf{b}$ for some $\mathbf{A} \in \mathbf{C}\left[\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}\right](S, R)$ and $\mathbf{b} \in -\mathbf{C}\,[\boldsymbol{\ell}]\,(S, \tilde{\mathbf{r}})$. Clearly, any $\mathbf{r} \in R$ satisfying $\boldsymbol{\ell}(\mathbf{s}, \mathbf{r}) = \mathbf{0}$ for some $\mathbf{s} \in S$ must correspond to an element $(\mathbf{r} - \tilde{\mathbf{r}}) = \boldsymbol{\rho}$ of this solution set. Thus, we are interested in computing an interval enclosure of the solution set of (5.16).

For $Q \subset \mathbb{R}$, let hull$(Q)$ denote the *interval hull* of $Q$; i.e, the smallest interval containing $Q$. To state the Hansen-Sengupta method formally, the following definition is useful.

**Definition 5.4.1.** For all $A, B, Z \in \mathbb{IR}$, let

$$\Gamma(A, B, Z) \equiv \text{hull}\left(\{z \in Z : az = b \text{ for some } (a, b) \in A \times B\}\right).$$

The following lemma provides a way to evaluate $\Gamma$ computationally.

**Lemma 5.4.2.** *For all $A, B, Z \in \mathbb{IR}$,*

$$\Gamma(A, B, Z) = \begin{cases} B/A \cap Z & \text{if } 0 \notin A \\ \text{hull}\left(Z \backslash \text{int}([b^L/a^L, b^L/a^U])\right) & \text{if } 0 \in A \text{ and } b^L > 0 \\ \text{hull}\left(Z \backslash \text{int}([b^U/a^U, b^U/a^L])\right) & \text{if } 0 \in A \text{ and } b^U < 0 \\ Z & \text{if } 0 \in A \text{ and } 0 \in B \end{cases}, \tag{5.17}$$

*where $B/A$ denotes interval division,*

$$B/A = [\min(b^L/a^L, b^U/a^L, b^L/a^U, b^U/a^U), \max(b^L/a^L, b^U/a^L, b^L/a^U, b^U/a^U)].$$

*Proof.* See Proposition 4.3.1 in [131]. □

For any $A, B, Z \in \mathbb{IR}$, either $\Gamma(A, B, Z) \in \mathbb{IR}$ or $\Gamma(A, B, Z) = \emptyset$. For convenience, the definition of $\Gamma$ is extended so that $\Gamma(A, B, Z) = \emptyset$ when any of $A$, $B$, or $Z$ is empty. Furthermore, we adopt the convention that any arithmetic operation between an element of $\mathbb{IR}$ and $\emptyset$ returns $\emptyset$, and any Cartesian product involving $\emptyset$ is equivalent to $\emptyset$. The following definition generalizes $\Gamma$ for application to $n$ dimensional linear systems.

**Definition 5.4.3.** For $A \in \mathbb{IR}^{n\times n}$, $B, Z \in \mathbb{IR}^n$, let

$$W_i \equiv \Gamma\left(A_{ii}, B_i - \sum_{j<i} A_{ij}W_j - \sum_{j>i} A_{ij}Z_j, Z_i\right),$$

for all $i = 1, \ldots, n$. Define $\Gamma(A, B, Z) \equiv W_1 \times \ldots \times W_n$.

Applying $\Gamma$ to (5.16) gives the following variant of the well-known result Theorem 5.1.8 in [131].

**Theorem 5.4.4.** *Let $S \in \mathbb{I}D_s$, $R \in \mathbb{I}D_r$, $\tilde{\mathbf{r}} \in R$, $\mathbf{C} \in \mathbb{R}^{n_r\times n_r}$, and let*

$$\mathcal{H}(S, R, \tilde{\mathbf{r}}, \mathbf{C}) \equiv \tilde{\mathbf{r}} + \Gamma\left(\mathbf{C}\left[\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}\right](S, R), -\mathbf{C}\,[\boldsymbol{\ell}]\,(S, \tilde{\mathbf{r}}), (R - \tilde{\mathbf{r}})\right).$$

*With $R' \equiv \mathcal{H}(S, R, \tilde{\mathbf{r}}, \mathbf{C})$, the following conclusions hold:*

1. *If $(\mathbf{s}, \mathbf{r}) \in S \times R$ satisfies $\boldsymbol{\ell}(\mathbf{s}, \mathbf{r}) = \mathbf{0}$, then $\mathbf{r} \in R'$.*

2. *If $R' = \emptyset$, then $\nexists(\mathbf{s}, \mathbf{r}) \in S \times R$ such that $\boldsymbol{\ell}(\mathbf{s}, \mathbf{r}) = \mathbf{0}$.*

3. *If $\tilde{\mathbf{r}} \in \mathrm{int}(R)$ and $\emptyset \neq R' \subset \mathrm{int}(R)$, then $\exists \mathbf{H} \in C^k(S, R')$ such that, for every $\mathbf{s} \in S$, $\mathbf{r} = \mathbf{H}(\mathbf{s})$ is the unique element of $R$ satisfying $\boldsymbol{\ell}(\mathbf{s}, \mathbf{r}) = \mathbf{0}$. Moreover, the interval matrix $\mathbf{C}\left[\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}\right](S, R)$ does not contain a singular matrix and does not contain zero in any of its diagonal elements.*

*Proof.* Suppose first that $S$ is a singleton, $S \equiv [\mathbf{s}, \mathbf{s}]$, for some $\mathbf{s} \in D_s$. Then, noting that $[\boldsymbol{\ell}]([\mathbf{s}, \mathbf{s}], \tilde{\mathbf{r}}) = \boldsymbol{\ell}(\mathbf{s}, \tilde{\mathbf{r}})$ by the definition of an interval extension, applying Corollary 5.1.5 and Theorem 5.1.8 in [131] to the function $\boldsymbol{\ell}(\mathbf{s}, \cdot)$ proves the theorem (the properties of $\mathbf{C}\left[\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}\right](S, R)$ in Conclusion 3 result from Theorem 4.4.5 (ii) in [131]). Next, suppose that $S$ is not a singleton. Fix any $\mathbf{s} \in S$ and suppose that $\mathbf{r} \in R$ satisfies $\boldsymbol{\ell}(\mathbf{s}, \mathbf{r}) = \mathbf{0}$. Since the theorem holds for $[\mathbf{s}, \mathbf{s}]$ as shown above, we must have $\mathbf{r} \in \mathcal{H}([\mathbf{s},\mathbf{s}], R, \tilde{\mathbf{r}}, \mathbf{C})$. But, by the inclusion monotonicity of natural interval extensions, $\mathbf{C}\left[\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}\right]([\mathbf{s},\mathbf{s}], R) \subset \mathbf{C}\left[\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}\right](S, R)$ and $-\mathbf{C}[\boldsymbol{\ell}]([\mathbf{s},\mathbf{s}], \tilde{\mathbf{r}}) \subset -\mathbf{C}[\boldsymbol{\ell}](S, \tilde{\mathbf{r}})$. Then Proposition 4.3.4 in [131] gives

$$\begin{aligned}
\mathcal{H}([\mathbf{s},\mathbf{s}], R, \tilde{\mathbf{r}}, \mathbf{C}) &= \tilde{\mathbf{r}} + \Gamma\left(\mathbf{C}\left[\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}\right]([\mathbf{s},\mathbf{s}], R), -\mathbf{C}[\boldsymbol{\ell}]([\mathbf{s},\mathbf{s}], \tilde{\mathbf{r}}), (R - \tilde{\mathbf{r}})\right), && (5.18)\\
&\subset \tilde{\mathbf{r}} + \Gamma\left(\mathbf{C}\left[\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}\right](S, R), -\mathbf{C}[\boldsymbol{\ell}](S, \tilde{\mathbf{r}}), (R - \tilde{\mathbf{r}})\right), && (5.19)\\
&= \mathcal{H}(S, R, \tilde{\mathbf{r}}, \mathbf{C}). && (5.20)
\end{aligned}$$

Therefore, $\mathbf{r} \in R'$, which proves 1, and 2 is an immediate consequence.

To prove Conclusion 3, suppose that $\tilde{\mathbf{r}} \in \text{int}(R)$, and $\emptyset \neq R' \subset \text{int}(R)$. Theorem 4.4.5 (ii) in [131] again establishes the properties of $\mathbf{C}\left[\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}\right](S, R)$. By Theorem 5.5.1 in [131] (see also Corollary 5.1.5), there exists a continuous function $\mathbf{H} : S \to R$ such that, for every $\mathbf{s} \in S$, $\mathbf{r} = \mathbf{H}(\mathbf{s})$ is the unique element of $R$ satisfying $\boldsymbol{\ell}(\mathbf{s}, \mathbf{r}) = \mathbf{0}$. By Conclusion 1 of the present theorem, $\mathbf{H} : S \to R'$. It only remains to show that $\mathbf{H} \in C^k(S, R')$.

Choosing any $\hat{\mathbf{s}} \in S$, Theorem 5.2.2 can be applied at the point $(\hat{\mathbf{s}}, \mathbf{H}(\hat{\mathbf{s}}))$ to conclude that there exists an open ball around $\hat{\mathbf{s}}$, $V_{\hat{\mathbf{s}}} \subset D_s$, an open ball around $\mathbf{H}(\hat{\mathbf{s}})$, $Q_{\hat{\mathbf{s}}}$, and $\mathbf{h}_{\hat{\mathbf{s}}} \in C^k(V_{\hat{\mathbf{s}}}, Q_{\hat{\mathbf{s}}})$ such that $\mathbf{h}_{\hat{\mathbf{s}}}(\hat{\mathbf{s}}) = \mathbf{H}(\hat{\mathbf{s}})$ and, for every $\mathbf{s} \in V_{\hat{\mathbf{s}}}$, $\mathbf{r} = \mathbf{h}_{\hat{\mathbf{s}}}(\mathbf{s})$ is the unique element of $Q_{\hat{\mathbf{s}}}$ satisfying $\boldsymbol{\ell}(\mathbf{s}, \mathbf{r}) = \mathbf{0}$. By continuity of $\mathbf{H}$, it is possible to choose an open ball $U_{\hat{\mathbf{s}}}$ around $\hat{\mathbf{s}}$ small enough that $\mathbf{H}$ maps $U_{\hat{\mathbf{s}}} \cap S$ into $Q_{\hat{\mathbf{s}}}$. Then, by the uniqueness property of $\mathbf{h}_{\hat{\mathbf{s}}}$ in $Q_{\hat{\mathbf{s}}}$, $\mathbf{H} = \mathbf{h}_{\hat{\mathbf{s}}}$ on $U_{\hat{\mathbf{s}}} \cap S$. The fact that $\mathbf{H} \in C^k(S, R')$ now follows from Lemma 23.1 in [127]. □

**Remark 5.4.5.** When applying Theorem 5.4.4, one should always choose a nonsingular preconditioner $\mathbf{C}$. In fact, the inclusion test $\emptyset \neq R' \subset \text{int}(R)$ in Conclusion 3 will never be satisfied if $\mathbf{C}$ is singular. However, the theorem holds in any case, so nonsingularity is not assumed.

**Remark 5.4.6.** The interval inclusion test given in part 3 of Theorem 5.4.4 is based on a characterization of the zeros of $\boldsymbol{\ell}$ derived from the mean-value theorem. Alternatively, an inclusion test can be derived from Brouwer's fixed point theorem without using the mean value theorem. This requires deriving a fixed point equation, $\mathbf{r} = \boldsymbol{\phi}(\mathbf{s}, \mathbf{r})$, with the same solutions as the original equations. For example, assuming that $\left(\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}\right)$ is nonsingular on $S \times R$, let

$$\boldsymbol{\phi}(\mathbf{s}, \mathbf{r}) \equiv \mathbf{r} - \left(\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}\right)^{-1}(\mathbf{s}, \mathbf{r})\boldsymbol{\ell}(\mathbf{s}, \mathbf{r}). \tag{5.21}$$

Brouwer's fixed point theorem can be used to show that the inclusion $[\boldsymbol{\phi}](S, R) \subset R$ guarantees the existence of $\mathbf{H} : S \to R$ satisfying $\mathbf{H}(\mathbf{s}) = \boldsymbol{\phi}(\mathbf{s}, \mathbf{H}(\mathbf{s}))$, and hence $\boldsymbol{\ell}(\mathbf{s}, \mathbf{H}(\mathbf{s})) = \mathbf{0}$, for all $\mathbf{s} \in S$. However, it is easily demonstrated that this inclusion will almost never be satisfied when the natural interval extension of $\boldsymbol{\phi}$ is used. Denoting the natural interval extension of the second term on the right-hand side of (5.21) over $S \times R$ by $M$, the natural interval extension of $\boldsymbol{\phi}$ is computed as $[\boldsymbol{\phi}](S, R) := R - M$. If $\exists(\mathbf{s}, \mathbf{r}) \in S \times R$ satisfying $\boldsymbol{\ell}(\mathbf{s}, \mathbf{r}) = \mathbf{0}$, then we must have $\mathbf{0} \in M$, and hence $[\boldsymbol{\phi}](S, R) \supset R$. Therefore, the desired inclusion will only hold when $[\boldsymbol{\phi}](S, R) = R$. This requires $M = [\mathbf{0}, \mathbf{0}]$, which can only occur in trivial cases.

### 5.4.2 An Interval Existence and Uniqueness Test for DAEs

Applying Theorem 5.4.4 to the algebraic equations in (5.1) gives the following corollary.

**Corollary 5.4.7.** *Let $(I, P, Z_x, Z_y) \in \mathbb{I}D_t \times \mathbb{I}D_p \times \mathbb{I}D_x \times \mathbb{I}D_y$, $\tilde{\mathbf{z}}_y \in Z_y$, $\mathbf{C} \in \mathbb{R}^{n_y \times n_y}$ and define*

$$\begin{aligned}&\mathcal{H}(I, P, Z_x, Z_y, \tilde{\mathbf{z}}_y, \mathbf{C})\\ &\qquad\equiv \tilde{\mathbf{z}}_y + \Gamma\left(\mathbf{C}\left[\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\right](I, P, Z_x, Z_y), -\mathbf{C}\left[\mathbf{g}\right](I, P, Z_x, \tilde{\mathbf{z}}_y), (Z_y - \tilde{\mathbf{z}}_y)\right).\end{aligned}$$

*With $Z'_y \equiv \mathcal{H}(I, P, Z_x, Z_y, \tilde{\mathbf{z}}_y, \mathbf{C})$, the following conclusions hold:*

1. *If $(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) \in I \times P \times Z_x \times Z_y$ satisfies $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$, then $\mathbf{z}_y \in Z'_y$.*

2. *If $Z'_y = \emptyset$, then $\nexists(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) \in I \times P \times Z_x \times Z_y$ such that $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$.*

3. *If $\tilde{\mathbf{z}}_y \in \mathrm{int}(Z_y)$ and $\emptyset \neq Z'_y \subset \mathrm{int}(Z_y)$, then $\exists \mathbf{H} \in C^1(I \times P \times Z_x, Z'_y)$ such that, for every $(t, \mathbf{p}, \mathbf{z}_x) \in I \times P \times Z_x$, $\mathbf{z}_y = \mathbf{H}(t, \mathbf{p}, \mathbf{z}_x)$ is the unique element of $Z_y$ satisfying $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$. Moreover, the interval matrix $\mathbf{C}\left[\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\right](I, P, Z_x, Z_y)$ does not contain a singular matrix and does not contain zero in any of its diagonal elements.*

*Proof.* The result follows immediately from Theorem 5.4.4. □

The following theorem is the main result of this section.

**Theorem 5.4.8.** *Let $(I, P, Z_x, Z_y) \in \mathbb{I}D_t \times \mathbb{I}D_p \times \mathbb{I}D_x \times \mathbb{I}D_y$, $\tilde{\mathbf{z}}_y \in Z_y$, $\mathbf{C} \in \mathbb{R}^{n_y \times n_y}$, and define $\mathcal{H}(I, P, Z_x, Z_y, \tilde{\mathbf{z}}_y, \mathbf{C})$ as in Corollary 5.4.7. Furthermore, let $X_0 \in \mathbb{IR}^{n_x}$ satisfy $\mathbf{x}_0(P) \subset X_0$ and denote $I = [t_0, t_f]$. If the inclusions*

$$\tilde{\mathbf{z}}_y \in \mathrm{int}(Z_y), \tag{5.22}$$

$$\emptyset \neq Z'_y \equiv \mathcal{H}(I, P, Z_x, Z_y, \tilde{\mathbf{z}}_y, \mathbf{C}) \subset \mathrm{int}(Z_y), \tag{5.23}$$

$$X_0 + [0, t_f - t_0]\,[\mathbf{f}]\,(I, P, Z_x, Z'_y) \subset Z_x, \tag{5.24}$$

*hold, then there exists a regular solution of (5.1) on $I \times P$ satisfying $(\mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p})) \in Z_x \times Z'_y$ for all $(t, \mathbf{p}) \in I \times P$. Furthermore, for any connected $\tilde{I} \subset I$ containing $t_0$, any connected $\tilde{P} \subset P$, and any solution $(\mathbf{x}^*, \mathbf{y}^*)$ of (5.1) on $\tilde{I} \times \tilde{P}$, either $(\mathbf{x}^*, \mathbf{y}^*) = (\mathbf{x}, \mathbf{y})$ on $\tilde{I} \times \tilde{P}$, or $\mathbf{y}^*(t_0, \mathbf{p}) \notin Z_y$, $\forall \mathbf{p} \in \tilde{P}$.*

*Proof.* By Conclusion 3 of Corollary 5.4.7, $\mathbf{C}\left[\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\right](I, P, Z_x, Z_y)$ contains no singular matrix and $\exists \mathbf{H} \in C^1(I \times P \times Z_x, Z'_y)$ such that, for every $(t, \mathbf{p}, \mathbf{z}_x) \in I \times P \times Z_x$, $\mathbf{z}_y = \mathbf{H}(t, \mathbf{p}, \mathbf{z}_x)$ is the unique element of $Z_y$ satisfying $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$.

Choose any $\mathbf{x}^0 \in C^1(I \times P, Z_x)$ and define the sequence $\{\mathbf{x}^k\}$ by

$$\mathbf{x}^{k+1}(t, \mathbf{p}) = \mathbf{x}_0(\mathbf{p}) + \int_{t_0}^{t} \mathbf{f}(s, \mathbf{p}, \mathbf{x}^k(s, \mathbf{p}), \mathbf{H}(s, \mathbf{p}, \mathbf{x}^k(s, \mathbf{p})))ds, \quad \forall (t, \mathbf{p}) \in I \times P. \tag{5.25}$$

If $\mathbf{x}^k \in C^1(I \times P, Z_x)$, which is true for $k = 0$, then $\mathbf{x}^{k+1}$ is well-defined and

$$\mathbf{x}^{k+1}(t, \mathbf{p}) \in X_0 + [0, t_f - t_0]\,[\mathbf{f}]\,(I, P, Z_x, Z'_y) \subset Z_x, \quad \forall (t, \mathbf{p}) \in I \times P. \tag{5.26}$$

Then, by induction, $\mathbf{x}^k \in C^1(I \times P, Z_x)$, $\forall k \in \mathbb{N}$.

Noting that both $\mathbf{f}$ and $\mathbf{H}$ are continuously differentiable, the mapping $(t, \mathbf{p}, \mathbf{z}_x) \mapsto \mathbf{f}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{H}(t, \mathbf{p}, \mathbf{z}_x))$ is Lipschitz on $I \times P \times Z_x$ by Lemma 5.2.1. Then, a standard inductive argument (see [78], Ch. II, Thm. 1.1) shows that $\{\mathbf{x}^k\}$ converges uniformly on $I \times P$ to a continuous limit function, denoted $\mathbf{x}$, and $\mathbf{x}$ satisfies

$$\dot{\mathbf{x}}(t, \mathbf{p}) = \mathbf{f}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{H}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}))), \quad \mathbf{x}(t_0, \mathbf{p}) = \mathbf{x}_0(\mathbf{p}), \quad \forall (t, \mathbf{p}) \in I \times P. \tag{5.27}$$

Since $\dot{\mathbf{x}}$ is continuous on $I \times P$, $\mathbf{x} \in C^1(I \times P, Z_x)$. Then, we may define $\mathbf{y} : I \times P \to D_y$ by $\mathbf{y}(t, \mathbf{p}) \equiv \mathbf{H}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}))$. With this definition, $\mathbf{y} \in C^1(I \times P, Z'_y)$ and

$$\mathbf{g}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p})) = \mathbf{g}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{H}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}))) = \mathbf{0}, \quad \forall (t, \mathbf{p}) \in I \times P. \tag{5.28}$$

Therefore, $(\mathbf{x}, \mathbf{y})$ is a solution of (5.1) on $I \times P$. Since $\mathbf{C}\left[\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\right](I, P, Z_x, Z_y)$, and hence $\left[\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\right](I, P, Z_x, Z_y)$, contains no singular matrix, $(\mathbf{x}, \mathbf{y})$ must be regular.

Now consider any connected $\tilde{I} \subset I$ containing $t_0$, any connected $\tilde{P} \subset P$, and any solution $(\mathbf{x}^*, \mathbf{y}^*)$ of (5.1) on $\tilde{I} \times \tilde{P}$. If $\mathbf{y}^*(t_0, \mathbf{p}) \in Z_y$ for some $\mathbf{p} \in \tilde{P}$, then the fact that $\mathbf{H}(t_0, \mathbf{p}, \mathbf{x}_0(\mathbf{p}))$ satisfies $\mathbf{g}(t_0, \mathbf{p}, \mathbf{x}_0(\mathbf{p}), \mathbf{H}(t_0, \mathbf{p}, \mathbf{x}_0(\mathbf{p}))) = \mathbf{0}$ uniquely among elements of $Z_y$ implies that $\mathbf{y}^*(t_0, \mathbf{p}) = \mathbf{H}(t_0, \mathbf{p}, \mathbf{x}_0(\mathbf{p})) = \mathbf{y}(t_0, \mathbf{p})$. Then the fact that $(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^*, \mathbf{y}^*)$ on $\tilde{I} \times \tilde{P}$ follows from Corollary 5.3.6. □

By checking some relatively simple inclusions, Theorem 5.4.8 provides a computational means to verify existence and uniqueness of a solution of (5.1) on given intervals $I \times P$, and provides a valid interval enclosure of this solution. In Chapter 6, an efficient numerical procedure for satisfying these inclusions is presented. In the following section, this result is used to develop computationally useful characterizations of bounding trajectories for the solutions of (5.1).

## 5.5 Bounding DAE Solutions using Differential Inequalities

This section presents three comparison theorems which provide sufficient conditions, in terms of differential inequalities, for mappings $\mathbf{v}, \mathbf{w} : I \to \mathbb{R}^{n_x}$ to satisfy

$$\mathbf{v}(t) \leq \mathbf{x}(t, \mathbf{p}) \leq \mathbf{w}(t), \quad \forall (t, \mathbf{p}) \in I \times P, \tag{5.29}$$

for some solution of (5.1) on $I \times P$. The first such theorem (Theorem 5.5.2) is very general, but does not suggest a complete computational bounding procedure for reasons discussed below. The remaining two results are modifications of Theorem 5.5.2 that address these issues. The following lemma is required to minimize repeated arguments.

**Lemma 5.5.1.** *Let $I = [t_0, t_f] \subset D_t$ and $P \subset D_p$ be intervals and let $(\mathbf{x}, \mathbf{y})$ be a regular solution of* (5.1) *on $I \times P$. Choose any continuous $\mathbf{v}, \mathbf{w} : I \to \mathbb{R}^{n_x}$ and any $\hat{\mathbf{p}} \in P$ and define*

$$\bar{\mathbf{x}}(t, \hat{\mathbf{p}}) \equiv \text{mid}(\mathbf{v}(t), \mathbf{w}(t), \mathbf{x}(t, \hat{\mathbf{p}})). \tag{5.30}$$

*For any $t_1 \in [t_0, t_f)$ such that $\bar{\mathbf{x}}(t_1, \hat{\mathbf{p}}) = \mathbf{x}(t_1, \hat{\mathbf{p}})$, there exists $t_4 \in (t_1, t_f]$, $L > 0$, and a continuous function $\bar{\mathbf{y}} : [t_1, t_4] \times P \to \mathbb{R}^{n_y}$ such that*

$$(\bar{\mathbf{x}}(t, \hat{\mathbf{p}}), \bar{\mathbf{y}}(t, \hat{\mathbf{p}})) \in D_x \times D_y, \tag{5.31}$$

$$\mathbf{g}(t, \hat{\mathbf{p}}, \bar{\mathbf{x}}(t, \hat{\mathbf{p}}), \bar{\mathbf{y}}(t, \hat{\mathbf{p}})) = \mathbf{0}, \tag{5.32}$$

$$\|\mathbf{y}(t, \hat{\mathbf{p}}) - \bar{\mathbf{y}}(t, \hat{\mathbf{p}})\|_\infty \leq L \|\mathbf{x}(t, \hat{\mathbf{p}}) - \bar{\mathbf{x}}(t, \hat{\mathbf{p}})\|_\infty, \tag{5.33}$$

$$\|\dot{\mathbf{x}}(t, \hat{\mathbf{p}}) - \mathbf{f}(t, \hat{\mathbf{p}}, \bar{\mathbf{x}}(t, \hat{\mathbf{p}}), \bar{\mathbf{y}}(t, \hat{\mathbf{p}}))\|_\infty \leq L \|\mathbf{x}(t, \hat{\mathbf{p}}) - \bar{\mathbf{x}}(t, \hat{\mathbf{p}})\|_\infty, \tag{5.34}$$

*for all $t \in [t_1, t_4]$.*

*Proof.* Since $(\mathbf{x},\mathbf{y})$ is regular, Theorem 5.2.2 may be applied to conclude that their exists an open ball around $(t_1,\hat{\mathbf{p}},\mathbf{x}(t_1,\hat{\mathbf{p}}))$, $V_1 \subset D_t \times D_p \times D_x$, and a function $\mathbf{h} \in C^1(V_1, D_y)$ such that $\mathbf{y}(t_1,\hat{\mathbf{p}}) = \mathbf{h}(t_1,\hat{\mathbf{p}},\mathbf{x}(t_1,\hat{\mathbf{p}}))$ and

$$\mathbf{g}(t,\mathbf{p},\mathbf{z}_x,\mathbf{h}(t,\mathbf{p},\mathbf{z}_x)) = \mathbf{0}, \quad \forall (t,\mathbf{p},\mathbf{z}_x) \in V_1. \tag{5.35}$$

Moreover, Lemma 5.3.8 shows that there exists an open ball around $(t_1,\hat{\mathbf{p}})$, $U_1 \subset D_t \times D_p$, such that $(t,\mathbf{p},\mathbf{x}(t,\mathbf{p})) \in V_1$ and $\mathbf{y}(t,\mathbf{p}) = \mathbf{h}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}))$, $\forall (t,\mathbf{p}) \in U_1 \cap (I \times P)$. Since $\bar{\mathbf{x}}(\cdot,\hat{\mathbf{p}})$ is continuous and $(t_1,\hat{\mathbf{p}},\bar{\mathbf{x}}(t_1,\hat{\mathbf{p}})) = (t_1,\hat{\mathbf{p}},\mathbf{x}(t_1,\hat{\mathbf{p}})) \in V_1$, $U_1$ may be chosen small enough that in addition $(t,\mathbf{p},\bar{\mathbf{x}}(t,\hat{\mathbf{p}})) \in V_1$, $\forall (t,\mathbf{p}) \in U_1 \cap (I \times P)$. Choosing $t_4 > t_1$ such that $[t_1,t_4] \times \{\hat{\mathbf{p}}\} \subset U_1 \cap (I \times P)$, define $\bar{\mathbf{y}}(t,\hat{\mathbf{p}}) \equiv \mathbf{h}(t,\hat{\mathbf{p}},\bar{\mathbf{x}}(t,\hat{\mathbf{p}}))$, $\forall t \in [t_1,t_4]$. Equation (5.31) now follows since $\mathbf{h}$ maps into $D_y$, and (5.32) follows from (5.35).

Since both $\mathbf{f}$ and $\mathbf{h}$ are continuously differentiable, the mappings

$$(t,\mathbf{p},\mathbf{z}_x) \mapsto \mathbf{h}(t,\mathbf{p},\mathbf{z}_x),$$
$$(t,\mathbf{p},\mathbf{z}_x) \mapsto \mathbf{f}(t,\mathbf{p},\mathbf{z}_x,\mathbf{h}(t,\mathbf{p},\mathbf{z}_x)),$$

are Lipschitz on any compact $K \subset V_1$ by Lemma 5.2.1. Let $K \equiv \{(t,\mathbf{p},\mathbf{z}_x) \in V_1 : t \in [t_1,t_4],\ \mathbf{p} = \hat{\mathbf{p}},\ \mathbf{z}_x = \mathbf{x}(t,\hat{\mathbf{p}}) \text{ or } \mathbf{z}_x = \bar{\mathbf{x}}(t,\hat{\mathbf{p}})\}$. Letting $L$ be the maximum of the corresponding Lipschitz constants, we arrive at (5.33) and (5.34). □

**Theorem 5.5.2.** *Let $I = [t_0,t_f] \subset D_t$ and $P \subset D_p$ be intervals and let $\mathbf{v},\mathbf{w} \in \mathcal{AC}(I,\mathbb{R}^{n_x})$ satisfy*

(EX): $\mathbf{v}(t) \leq \mathbf{w}(t)$, $\forall t \in I$.

(IC): $\mathbf{v}(t_0) \leq \mathbf{x}_0(\mathbf{p}) \leq \mathbf{w}(t_0)$, $\forall \mathbf{p} \in P$.

(RHS): *For a.e. $t \in I$ and each index $i$,*

1. $\dot{v}_i(t) \leq f_i(t,\mathbf{p},\mathbf{z}_x,\mathbf{z}_y)$ *for all* $(\mathbf{p},\mathbf{z}_x,\mathbf{z}_y) \in P \times D_x \times D_y$ *such that* $\mathbf{z}_x \in \mathcal{B}_i^L([\mathbf{v}(t),\mathbf{w}(t)])$ *and* $\mathbf{g}(t,\mathbf{p},\mathbf{z}_x,\mathbf{z}_y) = \mathbf{0}$,

2. *$\dot{w}_i(t) \geq f_i(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y)$ for all $(\mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) \in P \times D_x \times D_y$ such that $\mathbf{z}_x \in \mathcal{B}_i^U([\mathbf{v}(t), \mathbf{w}(t)])$ and $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$.*

*Then every regular solution of* (5.1) *on $I \times P$ satisfies $\mathbf{x}(t, \mathbf{p}) \in [\mathbf{v}(t), \mathbf{w}(t)]$, $\forall (t, \mathbf{p}) \in I \times P$.*

*Proof.* Let $(\mathbf{x}, \mathbf{y})$ be any regular solution of (5.1) on $I \times P$. Choose any $\hat{\mathbf{p}} \in P$ and suppose that there exists $t \in I$ such that $\mathbf{x}(t, \hat{\mathbf{p}}) \notin [\mathbf{v}(t), \mathbf{w}(t)]$. It will be shown that this results in a contradiction.

Define $t_1$ as in (3.9) with $\boldsymbol{\phi} = \mathbf{x}(\cdot, \hat{\mathbf{p}})$ and define $\bar{\mathbf{x}}$ as in (5.30). Noting that the hypotheses of Corollary 3.3.6 are satisfied, Conclusion 1 of Corollary 3.3.6 implies that $\bar{\mathbf{x}}(t_1, \hat{\mathbf{p}}) = \mathbf{x}(t_1, \hat{\mathbf{p}})$. Then, the hypotheses of Lemma 5.5.1 are verified, so that there exists $t_4 \in (t_1, t_f]$, $L > 0$ and $\bar{\mathbf{y}}$ satisfying (5.31)-(5.34). Applying Corollary 3.3.6 with $t_4$, $\beta = L$ and arbitrary $\epsilon > 0$ yields an index $j \in \{1, \ldots, n_x\}$, a non-decreasing function $\rho \in \mathcal{AC}([t_1, t_4], \mathbb{R})$ satisfying (3.7) on $[t_1, t_4]$, and numbers $t_2, t_3 \in [t_1, t_4]$ with $t_2 < t_3$ such that (3.10) and (3.11) hold with $\boldsymbol{\phi} = \mathbf{x}(\cdot, \hat{\mathbf{p}})$ (the proof is analogous if instead (3.12) holds).

It will now be shown that $\dot{v}_j(t) - \rho'(t) \leq \dot{x}_j(t, \hat{\mathbf{p}})$ for a.e. $t \in [t_2, t_3]$. Choose any $t \in (t_2, t_3)$. By (3.11) and Hypothesis (EX), we have $x_j(t, \hat{\mathbf{p}}) < v_j(t) \leq w_j(t)$. By definition, this implies that $\bar{\mathbf{x}}(t, \hat{\mathbf{p}}) \in \mathcal{B}_j^L([\mathbf{v}(t), \mathbf{w}(t)])$. Then, by (5.31) and (5.32), the point $(\hat{\mathbf{p}}, \bar{\mathbf{x}}(t, \hat{\mathbf{p}}), \bar{\mathbf{y}}(t, \hat{\mathbf{p}}))$ satisfies all of the of conditions of Hypothesis (RHS).1. Combining this with (5.34) gives

$$\dot{v}_j(t) \leq f_j(t, \hat{\mathbf{p}}, \bar{\mathbf{x}}(t, \hat{\mathbf{p}}), \bar{\mathbf{y}}(t, \hat{\mathbf{p}})) \leq \dot{x}_j(t, \hat{\mathbf{p}}) + L\|\mathbf{x}(t, \hat{\mathbf{p}}) - \bar{\mathbf{x}}(t, \hat{\mathbf{p}})\|_\infty, \tag{5.36}$$

for a.e. $t \in [t_2, t_3]$. By (3.10), $\|\mathbf{x}(t, \hat{\mathbf{p}}) - \bar{\mathbf{x}}(t, \hat{\mathbf{p}})\|_\infty$ is bounded by $\rho(t)$ for all $t \in [t_2, t_3)$. Then, since $\rho'(t) > L\rho(t)$ for a.e. $t \in [t_1, t_4]$,

$$\dot{v}_j(t) - \rho'(t) \leq \dot{x}_j(t, \hat{\mathbf{p}}) + L\rho(t) - \rho'(t) < \dot{x}_j(t, \hat{\mathbf{p}}), \tag{5.37}$$

for a.e. $t \in [t_2, t_3]$.

Applying Theorem 3.3.3, the function $v_j - \rho - x_j(\cdot, \hat{\mathbf{p}})$ is non-increasing on $(t_2, t_3)$, so that in particular,

$$v_j(t_3) - \rho(t_3) - x_j(t_3, \hat{\mathbf{p}}) \leq v_j(t_2) - \rho(t_2) - x_j(t_2, \hat{\mathbf{p}}). \tag{5.38}$$

Using (3.11), this implies that $0 \leq -\rho(t_2)$, which is a contradiction because $\rho(t) > 0$ for all $t \in [t_2, t_3]$. Thus, we must have $\mathbf{x}(t, \hat{\mathbf{p}}) \in [\mathbf{v}(t), \mathbf{w}(t)]$, $\forall t \in I$. In fact, since $\hat{\mathbf{p}} \in P$ was chosen arbitrarily, we have $\mathbf{x}(t, \mathbf{p}) \in [\mathbf{v}(t), \mathbf{w}(t)]$, $\forall (t, \mathbf{p}) \in I \times P$. □

Theorem 5.5.2 is very similar to the results for bounding the solutions of explicit ODEs presented in Chapter 3. There, it was shown that interval arithmetic can be used to derive an auxiliary system of ODEs whose solutions satisfy conditions analogous to (IC) and (RHS) in Theorem 5.5.2, and these ODEs can be solved efficiently using a state-of-the-art numerical integrator to provide bounds. We present similar approaches for DAEs in Chapter 6. However, there is a problem with using Theorem 5.5.2 directly. Using interval methods to satisfy (RHS) would require some procedure for computing bounds on the zeros of $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \cdot)$ with $(t, \mathbf{p}, \mathbf{z}_x)$ restricted to a given interval. Using the interval Hansen-Sengupta method, it is only possible to refine such an enclosure when provided with a guaranteed *a priori* enclosure.

A further complication is that Theorem 5.5.2 produces bounds that enclose *all* regular solutions of (5.1) on $I \times P$. However, in applications it is very likely that there will be a particular solution of interest, specified by a consistent initial condition $\mathbf{y}(t_0, \hat{\mathbf{p}})$ for some $\hat{\mathbf{p}} \in P$ (see Corollary 5.3.6). Theorem 5.5.2 provides no mechanism for restricting $\mathbf{v}$ and $\mathbf{w}$ based on this information because (RHS) requires that $\dot{v}_i$ and $\dot{w}_i$ bound $f_i(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y)$ for *all* $\mathbf{z}_y$ satisfying $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$. The following theorem shows that both of these problems can be avoided by modifying (RHS) in the case where intervals satisfying the conditions of Theorem 5.4.8 are available.

**Theorem 5.5.3.** *Let $(I, P, Z_x, Z_y, Z'_y) \in \mathbb{I}D_t \times \mathbb{I}D_p \times \mathbb{I}D_x \times \mathbb{I}D_y \times \mathbb{I}D_y$, $I = [t_0, t_f]$ and $Z'_y \subset Z_y$, and let $(\mathbf{x}, \mathbf{y}) \in C^1(I \times P, Z_x) \times C^1(I \times P, Z'_y)$ be a regular solution of (5.1) on $I \times P$. Suppose further that $\exists \mathbf{H} \in C^1(I \times P \times Z_x, Z'_y)$ such that, for every $(t, \mathbf{p}, \mathbf{z}_x) \in I \times P \times Z_x$, $\mathbf{z}_y = \mathbf{H}(t, \mathbf{p}, \mathbf{z}_x)$ is the unique element of $Z_y$ satisfying $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$. Let $\mathbf{v}, \mathbf{w} \in \mathcal{AC}(I, \mathbb{R}^{n_x})$ satisfy*

(EX): $\mathbf{v}(t) \leq \mathbf{w}(t)$ *and* $Z_x \cap [\mathbf{v}(t), \mathbf{w}(t)] \neq \emptyset$, $\forall t \in I$.

(IC): $\mathbf{v}(t_0) \leq \mathbf{x}_0(\mathbf{p}) \leq \mathbf{w}(t_0)$, $\forall \mathbf{p} \in P$.

(RHS): *For a.e.* $t \in I$ *and each index* $i$,

1. $\dot{v}_i(t) \leq f_i(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y)$ *for all* $(\mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) \in P \times Z_x \times Z'_y$ *such that* $\mathbf{z}_x \in \mathcal{B}_i^L(Z_x \cap [\mathbf{v}(t), \mathbf{w}(t)])$ *and* $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$,
2. $\dot{w}_i(t) \geq f_i(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y)$ *for all* $(\mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) \in P \times Z_x \times Z'_y$ *such that* $\mathbf{z}_x \in \mathcal{B}_i^U(Z_x \cap [\mathbf{v}(t), \mathbf{w}(t)])$ *and* $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$.

*Then* $\mathbf{x}(t, \mathbf{p}) \in [\mathbf{v}(t), \mathbf{w}(t)]$ *for all* $(t, \mathbf{p}) \in I \times P$.

*Proof.* Choose any $\hat{\mathbf{p}} \in P$ and suppose that there exists $t \in I$ such that $\mathbf{x}(t, \hat{\mathbf{p}}) \notin [\mathbf{v}(t), \mathbf{w}(t)]$. It will be shown that this results in a contradiction.

Define $\bar{\mathbf{x}}(t, \hat{\mathbf{p}})$ as in (5.30). Clearly, $\bar{\mathbf{x}}(t, \hat{\mathbf{p}}) \in [\mathbf{v}(t), \mathbf{w}(t)]$, $\forall t \in I$. Let $[\mathbf{z}_x^L, \mathbf{z}_x^U] \equiv Z_x$. Since $x_j(t, \hat{\mathbf{p}}) \in [z_{x,j}^L, z_{x,j}^U]$ by definition, it follows that $\bar{x}_j(t, \hat{\mathbf{p}}) \in [z_{x,j}^L, z_{x,j}^U]$ for any index $j$ such that $x_j(t, \hat{\mathbf{p}}) = \bar{x}_j(t, \hat{\mathbf{p}})$. Alternatively, for any $j$ such that $x_j(t, \hat{\mathbf{p}}) \neq \bar{x}_j(t, \hat{\mathbf{p}})$, we have $x_j(t, \hat{\mathbf{p}}) < v_j(t)$ (or $x_j(t, \hat{\mathbf{p}}) > w_j(t)$), which, combined with the fact that $Z_x \cap [\mathbf{v}(t), \mathbf{w}(t)]$ is nonempty by hypothesis, gives

$$z_{x,j}^L \leq x_j(t, \hat{\mathbf{p}}) < v_j(t) = \text{mid}(v_j(t), w_j(t), x_j(t, \hat{\mathbf{p}})) = \bar{x}_j(t, \hat{\mathbf{p}}) \leq z_{x,j}^U \tag{5.39}$$

$$\Big(\text{or} \quad z_{x,j}^U \geq x_j(t, \hat{\mathbf{p}}) > w_j(t) = \text{mid}(v_j(t), w_j(t), x_j(t, \hat{\mathbf{p}})) = \bar{x}_j(t, \hat{\mathbf{p}}) \geq z_{x,j}^L\Big). \tag{5.40}$$

Therefore $\bar{\mathbf{x}}(t, \hat{\mathbf{p}}) \in Z_x$.

Define $t_1$ as in (3.9) with $\boldsymbol{\phi} = \mathbf{x}(\cdot, \hat{\mathbf{p}})$, define $t_4 \equiv t_f$, and define $\bar{\mathbf{y}}(t, \hat{\mathbf{p}}) \equiv \mathbf{H}(t, \hat{\mathbf{p}}, \bar{\mathbf{x}}(t, \hat{\mathbf{p}}))$, $\forall t \in I$. By the definition of $\mathbf{H}$, it follows that $\bar{\mathbf{y}}(t, \hat{\mathbf{p}}) \in Z'_y$ for all $t \in [t_1, t_4]$ and (5.32) holds. Moreover, it can be shown that (5.34) holds by noting that the function

$$(t, \mathbf{p}, \mathbf{z}_x) \mapsto \mathbf{f}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{H}(t, \mathbf{p}, \mathbf{z}_x)),$$

is Lipschitz on compact subsets of $I \times P \times Z_x$, exactly as in Lemma 5.5.1. Applying Corollary 3.3.6 with $t_4$, $\beta = L$ and arbitrary $\epsilon > 0$ yields an index $j \in \{1, \dots, n_x\}$, a non-decreasing function $\rho \in \mathcal{AC}([t_1, t_4], \mathbb{R})$ satisfying (3.7) on $[t_1, t_4]$, and numbers $t_2, t_3 \in [t_1, t_4]$ with $t_2 < t_3$ such that (3.10) and (3.11) hold with $\boldsymbol{\phi} = \mathbf{x}(\cdot, \hat{\mathbf{p}})$ (the proof is analogous if instead (3.12) holds).

It will now be shown that (5.36) holds for a.e. $t \in [t_2, t_3]$. Choose any $t \in (t_2, t_3)$. It was argued above that $\bar{\mathbf{x}}(t, \hat{\mathbf{p}}) \in Z_x \cap [\mathbf{v}(t), \mathbf{w}(t)]$ and $\bar{\mathbf{y}}(t, \hat{\mathbf{p}}) \in Z'_y$. By (3.11) and Hypothesis (EX), we have $z^L_{x,j} \leq x_j(t, \hat{\mathbf{p}}) < v_j(t) = \text{mid}(v_j(t), w_j(t), x_j(t, \hat{\mathbf{p}})) = \bar{x}_j(t, \hat{\mathbf{p}})$, and therefore $\bar{\mathbf{x}}(t, \hat{\mathbf{p}}) \in \mathcal{B}^L_j(Z_x \cap [\mathbf{v}(t), \mathbf{w}(t)])$. Then, by (5.32), the point $(\hat{\mathbf{p}}, \bar{\mathbf{x}}(t, \hat{\mathbf{p}}), \bar{\mathbf{y}}(t, \hat{\mathbf{p}}))$ satisfies all of the conditions of Hypothesis (RHS).1. Combining this with (5.34) proves (5.36), and the remainder of the proof follows exactly as is the proof of Theorem 5.5.2. □

The final result below shows that the complications with Theorem 5.5.2 can also be avoided without having to first satisfy the conditions of Theorem 5.4.8, as in Theorem 5.5.3. Instead, we require satisfaction of (5.23) pointwise along the bounding trajectories $\mathbf{v}$ and $\mathbf{w}$, as in the following Hypothesis.

**Hypothesis 5.5.1.** Let $(I, P) \in \mathbb{I}D_t \times \mathbb{I}D_p$, $\mathbf{C} : I \to \mathbb{R}^{n_y \times n_y}$ and $\tilde{\mathbf{z}}_y : I \to \mathbb{R}^{n_y}$. Suppose that $\mathbf{z}^L_y, \mathbf{z}^U_y : I \to \mathbb{R}^{n_y}$ and $\mathbf{v}, \mathbf{w} : I \to \mathbb{R}^{n_x}$ are continuous and satisfy

(EX): $\mathbf{v}(t) \leq \mathbf{w}(t)$ and $\mathbf{z}^L_y(t) \leq \mathbf{z}^U_y(t)$, $\forall t \in I$.

(ALG): For all $t \in I$,

$$([\mathbf{v}(t), \mathbf{w}(t)], Z_y(t)) \in \mathbb{I}D_x \times \mathbb{I}D_y, \tag{5.41}$$

$$\tilde{\mathbf{z}}_y(t) \in \text{int}(Z_y(t)), \tag{5.42}$$

$$\emptyset \neq Z'_y(t) \equiv \mathcal{H}([t, t], P, [\mathbf{v}(t), \mathbf{w}(t)], Z_y(t), \tilde{\mathbf{z}}_y(t), \mathbf{C}(t)) \subset \text{int}(Z_y(t)), \tag{5.43}$$

where $Z_y(t) \equiv [\mathbf{z}^L_y(t), \mathbf{z}^U_y(t)]$ and $\mathcal{H}$ is defined as in Corollary 5.4.7.

**Lemma 5.5.4.** *Suppose Hypothesis 5.5.1 holds and define*

$$V \equiv \{(t, \mathbf{p}, \mathbf{z}_x) \in I \times P \times D_x : \mathbf{z}_x \in [\mathbf{v}(t), \mathbf{w}(t)]\}. \tag{5.44}$$

*There exists* $\mathbf{H} \in C^1(V, D_y)$ *such that, for every* $(t, \mathbf{p}, \mathbf{z}_x) \in V$, $\mathbf{z}_y = \mathbf{H}(t, \mathbf{p}, \mathbf{z}_x)$ *is an element of* $Z'_y(t)$ *and satisfies* $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$ *uniquely among elements of* $Z_y(t)$.

*Proof.* Choose any $t \in I$ and define $V_t \equiv [t, t] \times P \times [\mathbf{v}(t), \mathbf{w}(t)]$. By Hypothesis 5.5.1 and Conclusion 3 of Corollary 5.4.7, there exists $\mathbf{H}_t \in C^1(V_t, Z'_y(t))$ such that, for every $(t, \mathbf{p}, \mathbf{z}_x) \in V_t$, $\mathbf{z}_y = \mathbf{H}_t(t, \mathbf{p}, \mathbf{z}_x)$ is the unique element of $Z_y(t)$ satisfying $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$. Define $\mathbf{H} : V \to D_y$ by $\mathbf{H}(t, \mathbf{p}, \mathbf{z}_x) = \mathbf{H}_t(t, \mathbf{p}, \mathbf{z}_x)$. By the properties of each $H_t$ above, it only remains to show that $\mathbf{H} \in C^1(V, D_y)$.

By Lemma 23.1 in [127], it suffices to show that, for every $(\hat{t}, \hat{\mathbf{p}}, \hat{\mathbf{z}}_x) \in V$, there exists an open ball $\hat{U}$ and a function $\hat{\mathbf{h}} \in C^1(\hat{U}, D_y)$ that agrees with $\mathbf{H}$ on $\hat{U} \cap V$. Choose any such point and let $\hat{\mathbf{z}}_y = \mathbf{H}(\hat{t}, \hat{\mathbf{p}}, \hat{\mathbf{z}}_x)$. Applying Theorem 5.2.2 at the point $(\hat{t}, \hat{\mathbf{p}}, \hat{\mathbf{z}}_x, \hat{\mathbf{z}}_y)$ gives an open ball around $(\hat{t}, \hat{\mathbf{p}}, \hat{\mathbf{z}}_x)$, $\hat{V} \subset D_t \times D_p \times D_x$, an open ball around $\hat{\mathbf{z}}_y$, $\hat{Q} \subset D_y$, and $\hat{\mathbf{h}} \in C^1(\hat{V}, \hat{Q})$ such that $\hat{\mathbf{h}}(\hat{t}, \hat{\mathbf{p}}, \hat{\mathbf{z}}_x) = \hat{\mathbf{z}}_y$ and, for every $(t, \mathbf{p}, \mathbf{z}_x) \in \hat{V}$, $\mathbf{z}_y = \hat{\mathbf{h}}(t, \mathbf{p}, \mathbf{z}_x)$ is the unique element of $\hat{Q}$ satisfying $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x) = \mathbf{0}$. Noting that $\hat{\mathbf{z}}_y = \mathbf{H}(\hat{t}, \hat{\mathbf{p}}, \hat{\mathbf{z}}_x)$ is in $Z'_y(\hat{t})$, and hence in $\text{int}(Z_y(\hat{t}))$ by (5.43), choose an open ball $\hat{Q}'$ around $\hat{\mathbf{z}}_y$ such that its closure is contained in $\text{int}(Z_y(\hat{t}))$. By continuity of $\mathbf{z}^L_y$ and $\mathbf{z}^U_y$, $\exists \delta > 0$ such that $\hat{Q}' \subset \text{int}(Z_y(t))$, for all $t \in I$ with $|t - \hat{t}| < \delta$. By continuity of $\hat{\mathbf{h}}$, there exists an open ball around $(\hat{t}, \hat{\mathbf{p}}, \hat{\mathbf{z}}_x)$, $\hat{U} \subset \hat{V}$, so small that any $(t, \mathbf{p}, \mathbf{z}_x) \in \hat{U} \cap V$ has $|t - \hat{t}| < \delta$ and $\hat{\mathbf{h}}(t, \mathbf{p}, \mathbf{z}_x) \in \hat{Q}'$. Then, for any $(t, \mathbf{p}, \mathbf{z}_x) \in \hat{U} \cap V$, both $\hat{\mathbf{h}}(t, \mathbf{p}, \mathbf{z}_x)$ and $\mathbf{H}(t, \mathbf{p}, \mathbf{z}_x)$ are zeros of $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \cdot)$ in $Z_y(t)$, and hence $\hat{\mathbf{h}}(t, \mathbf{p}, \mathbf{z}_x) = \mathbf{H}(t, \mathbf{p}, \mathbf{z}_x)$. □

**Lemma 5.5.5.** *Suppose Hypothesis 5.5.1 holds and let* $(\mathbf{x}, \mathbf{y})$ *be a solution of* (5.1) *on* $I \times P$. *For any* $I' \equiv [t', t''] \subset I$ *and* $\mathbf{p}' \in P$, *the following implication holds:*

$$\left.\begin{array}{rl} \mathbf{x}(t, \mathbf{p}) & \in [\mathbf{v}(t), \mathbf{w}(t)], \quad \forall (t, \mathbf{p}) \in I' \times P \\ \mathbf{y}(t', \mathbf{p}') & \in Z_y(t') \end{array}\right\} \implies \begin{array}{c} \mathbf{y}(t, \mathbf{p}) \in Z'_y(t), \\ \forall (t, \mathbf{p}) \in I' \times P \end{array} \tag{5.45}$$

*Proof.* First, it is shown that the implication

$$(\mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p})) \in [\mathbf{v}(t), \mathbf{w}(t)] \times Z_y(t) \implies \mathbf{y}(t, \mathbf{p}) \in Z'_y(t) \tag{5.46}$$

holds for any $(t, \mathbf{p}) \in I \times P$. Let $V$ and $\mathbf{H}$ be as in Lemma 5.5.4 and suppose that the hypothesis of (5.46) holds. By definition $\mathbf{H}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}))$ is the unique zero of $\mathbf{g}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \cdot)$ in $Z_y(t)$. But $\mathbf{y}(t, \mathbf{p})$ is a zero of $\mathbf{g}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \cdot)$ in $Z_y(t)$, and hence $\mathbf{y}(t, \mathbf{p}) = \mathbf{H}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}))$. Noting that $\mathbf{H}$ maps into $Z'_y(t)$, (5.46) is established.

Under the hypotheses of (5.45), (5.46) implies that $\mathbf{y}(t', \mathbf{p}') \in Z'_y(t')$. If the conclusion of (5.45) fails, then there must exist $(t_2, \mathbf{p}_2) \in (t', t''] \times P$ such that $\mathbf{y}(t_2, \mathbf{p}_2) \notin Z'_y(t_2)$. Furthermore, this point must satisfy $\mathbf{y}(t_2, \mathbf{p}_2) \notin Z_y(t_2)$, since otherwise (5.46) provides a contradiction. Continuity of $\mathbf{y}$, $\mathbf{z}_y^L$ and $\mathbf{z}_y^U$ then imply that $\exists (t_1, \mathbf{p}_1) \in (t', t''] \times P$ such that $\mathbf{y}(t_1, \mathbf{p}_1)$ is an element of the boundary of $Z_y(t_1)$, and hence of $Z_y(t_1)$, but not an element of $Z'_y(t_1) \subset \text{int}(Z_y(t_1))$. Again, (5.46) provides a contradiction. □

**Theorem 5.5.6.** *Suppose Hypothesis 5.5.1 holds. Additionally, let $\mathbf{v}, \mathbf{w}$ be absolutely continuous and satisfy*

(IC): $\mathbf{v}(t_0) \leq \mathbf{x}_0(\mathbf{p}) \leq \mathbf{w}(t_0)$, $\forall \mathbf{p} \in P$.

(RHS): *For a.e. $t \in I$ and each index $i$,*

1. $\dot{v}_i(t) \leq f_i(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y)$ *for all* $(\mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) \in P \times D_x \times Z'_y(t)$ *such that* $\mathbf{z}_x \in \mathcal{B}_i^L([\mathbf{v}(t), \mathbf{w}(t)])$ *and* $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$,
2. $\dot{w}_i(t) \geq f_i(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y)$ *for all* $(\mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) \in P \times D_x \times Z'_y(t)$ *such that* $\mathbf{z}_x \in \mathcal{B}_i^U([\mathbf{v}(t), \mathbf{w}(t)])$ *and* $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$.

*Then every regular solution of* (5.1) *on $I \times P$ with $\mathbf{y}(t_0, \tilde{\mathbf{p}}) \in Z_y(t_0)$ for at least one $\tilde{\mathbf{p}} \in P$ must satisfy $(\mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p})) \in [\mathbf{v}(t), \mathbf{w}(t)] \times Z'_y(t)$ for all $(t, \mathbf{p}) \in I \times P$.*

*Proof.* Let $(\mathbf{x}, \mathbf{y})$ be a regular solution of (5.1) on $I \times P$ satisfying $\mathbf{y}(t_0, \tilde{\mathbf{p}}) \in Z_y(t_0)$ for some $\tilde{\mathbf{p}} \in P$. Choose any $\hat{\mathbf{p}} \in P$ and suppose that there exists $t \in I$ such that $\mathbf{x}(t, \hat{\mathbf{p}}) \notin [\mathbf{v}(t), \mathbf{w}(t)]$. It will be shown that this results in a contradiction.

Define $t_1$ as in (3.9) with $\boldsymbol{\phi} \equiv \mathbf{x}(\cdot, \hat{\mathbf{p}})$. Noting that the hypotheses of Corollary 3.3.6 are satisfied, Conclusion 1 of that corollary and (5.45) imply that $\mathbf{y}(t, \hat{\mathbf{p}}) \in Z'_y(t)$, $\forall t \in [t_0, t_1]$. Define $\bar{\mathbf{x}}$ as in Lemma 5.5.1. Noting that $\bar{\mathbf{x}}(t_1, \hat{\mathbf{p}}) = \mathbf{x}(t_1, \hat{\mathbf{p}})$ by Conclusion 1 of Corollary 3.3.6, Lemma 5.5.1 furnishes $t_4 \in (t_1, t_f]$, $L > 0$ and $\bar{\mathbf{y}}$

satisfying (5.31)-(5.34). By (5.33) and (5.43), $\bar{\mathbf{y}}(t_1, \hat{\mathbf{p}}) = \mathbf{y}(t_1, \hat{\mathbf{p}}) \in \text{int}(Z_y(t_1))$. By continuity of $\bar{\mathbf{y}}$, $\mathbf{z}_y^L$, $\mathbf{z}_y^U$, it is possible to restrict $t_4$ so that

$$\bar{\mathbf{y}}(t, \hat{\mathbf{p}}) \in Z_y(t), \quad \forall t \in [t_1, t_4]. \tag{5.47}$$

We now apply Corollary 3.3.6 with $t_4$, $\beta = L$ and arbitrary $\epsilon > 0$. This yields an index $j \in \{1, \ldots, n_x\}$, a non-decreasing function $\rho \in \mathcal{AC}([t_1, t_4], \mathbb{R})$ satisfying (3.7) on $[t_1, t_4]$, and numbers $t_2, t_3 \in [t_1, t_4]$ with $t_2 < t_3$ such that (3.10) and (3.11) hold with $\boldsymbol{\phi} \equiv \mathbf{x}(\cdot, \hat{\mathbf{p}})$ (the proof is analogous if instead (3.12) holds).

It will now be shown that (5.36) holds for a.e. $t \in [t_2, t_3]$. Choose any $t \in (t_2, t_3)$. By (3.11) and Hypothesis 5.5.1 (EX), we have $x_j(t, \hat{\mathbf{p}}) < v_j(t) \leq w_j(t)$. By definition, this implies that $\bar{\mathbf{x}}(t, \hat{\mathbf{p}}) \in \mathcal{B}_j^L([\mathbf{v}(t), \mathbf{w}(t)])$. Since $\bar{\mathbf{x}}(t, \hat{\mathbf{p}}) \in [\mathbf{v}(t), \mathbf{w}(t)]$ and $\bar{\mathbf{y}}(t, \hat{\mathbf{p}})$ is a zero of $\mathbf{g}(t, \hat{\mathbf{p}}, \bar{\mathbf{x}}(t, \hat{\mathbf{p}}), \cdot)$ by (5.32), Equation (5.47) and Corollary 5.4.7 show that $\bar{\mathbf{y}}(t, \hat{\mathbf{p}}) \in Z_y'(t)$. Then, by (5.31) and (5.32), the point $(\hat{\mathbf{p}}, \bar{\mathbf{x}}(t, \hat{\mathbf{p}}), \bar{\mathbf{y}}(t, \hat{\mathbf{p}}))$ satisfies all of the conditions of (RHS).1. Combining this with (5.34) proves (5.36) and, exactly as is the proof of Theorem 5.5.2, we conclude that $\mathbf{x}(t, \mathbf{p}) \in [\mathbf{v}(t), \mathbf{w}(t)]$, $\forall (t, \mathbf{p}) \in I \times P$. The theorem now follows from (5.45). □

## 5.6 Conclusions

We have presented a detailed analysis characterizing interval enclosures of the solutions of semi-explicit, index-one DAEs subject to uncertain initial conditions and parameters. The primary contributions are (1) a set of conditions guaranteeing existence and uniqueness of a solution and providing a crude enclosure, and (2) three theorems giving sufficient conditions for some functions to describe bounds on one or all solutions pointwise in the independent variable. What remains is to develop methods for satisfying these conditions computationally, thus leading to efficient, constructive procedures for computing bounds. We take up this task in Chapter 6.

# Chapter 6

# Computing State Bounds for Semi-Explicit Index-One DAEs

## 6.1 Introduction

In the previous chapter, several theoretical results were presented that provide computationally useful characterizations of interval bounds on the solutions of semi-explicit index-one DAEs. In this chapter, these results are used to derive two efficient numerical methods for computing such bounds. The first method proceeds in two-phases, as described in §6.3. In Phase 1, the interval inclusion test of §5.4 is applied to verify existence and uniqueness of a DAE solution, and to provide a crude enclosure of this solution. Unfortunately, this test is difficult to satisfy computationally because it involves implicit conditions. This challenge is addressed in §6.4. Using the crude enclosure from Phase 1, the second phase computes refined, time-varying bounds on the DAE solution using the results of §5.5. The implementation of Phase 2 involves numerical integration of an auxiliary system of ODEs whose solutions describe the desired bounds, and is described in §6.5.

The second proposed bounding method, which is described in §6.6, reduces the first method to a single phase based on Theorem 5.5.6 in §5.5. The computation of the resulting bounds is similar to Phase 2 of the first method, only here the auxiliary system to be solved is described by semi-explicit DAEs.

The two-phase framework described above is analogous to the two-phase approach used for validated integration of ODEs [130]. Indeed, Phase 1 of this approach provides a key step toward the development of validated methods for DAEs. In Phase 2, however, we deviate from this approach by using a standard numerical integration code to compute refined bounds via the theory of differential inequalities. The resulting bounds are mathematically guaranteed, but subject to the error of numerical integration. Therefore, this method is not validated, and the same is true of the single-phase method. On the other hand, the use of state-of-the-art numerical integration codes leads to a very effective implementation. In §6.7, both methods are applied to numerical examples and shown to produce accurate bounds very efficiently.

## 6.2 Preliminaries

### 6.2.1 Extended Interval Functions

The methods of this chapter will make extensive use of intervals and interval-valued functions. For computational reasons, it is often convenient to extend such functions outside their domains in a regular manner. For example, it is desirable to define the behavior of an interval function taking the argument $[\mathbf{v}, \mathbf{w}]$ if, by some numerical error, we have $v_i > w_i$ for some $i$. There is a large literature on interval implementations that account for numerical error in a conservative manner in order to avoid these types of issues altogether. However, as we will see, the proposed methods for DAEs present unique challenges. As a particular example, we will make use of an algebraic equation solver to locate $\mathbf{v}$ and $\mathbf{w}$ such that $[\mathbf{v}, \mathbf{w}]$ satisfies an implicit interval equation. Though the solution is guaranteed to satisfy $\mathbf{v} \leq \mathbf{w}$, this may not hold for some iterate produced by the solver. If no provisions are made for this situation, the solver will be forced to abort. On the other hand, if the participating interval functions are extended onto $\mathbb{R}^n \times \mathbb{R}^n$ in a regular manner, this situation poses no problem for the solver, which may eventually converge to a solution describing a proper interval.

Some basic extended interval operations have already been defined in previous chapters, including the $\Box$ function (Definition 2.5.17) and the extended intersection $\tilde{\cap}$ (Definition 2.5.22). Both of these will be used throughout this chapter. Moreover, we will make use of two modified forms of the interval function $\Gamma$ (Definition 5.4.3).

**Definition 6.2.1.** Let

$$\mathcal{D}^* \equiv \{(A, B, Z) \in \mathbb{IR}^{n\times n} \times \mathbb{IR}^n \times \mathbb{IR}^n : 0 \notin A_{ii}, \forall i = 1, \ldots, n\}, \tag{6.1}$$

and define $\Gamma^* : \mathcal{D}^* \to \mathbb{IR}^n$ by $\Gamma^*(A, B, Z) \equiv W_1^* \times \ldots \times W_n^*$, where

$$W_i^* = \frac{1}{A_{ii}}\left(B_i - \sum_{k<i} A_{ik}W_k^* - \sum_{k>i} A_{ik}Z_k\right), \quad \forall i \in \{1, \ldots, n\}. \tag{6.2}$$

**Definition 6.2.2.** Define $\Gamma^+ : \mathcal{D}^* \to \mathbb{IR}^n$ by $\Gamma^+(A, B, Z) \equiv W_1^+ \times \ldots \times W_{n_y}^+$, where

$$W_i^+ = Z_i \tilde{\cap} \frac{1}{A_{ii}}\left(B_i - \sum_{k<i} A_{ik}W_k^+ - \sum_{k>i} A_{ik}Z_k\right), \quad \forall i \in \{1, \ldots, n\}. \tag{6.3}$$

The functions $\Gamma^+$ and $\Gamma^*$ differ from $\Gamma$ in that they omit or extend the intersection with $Z$ in the definition of $\Gamma$. We have the following properties and relationships.

**Lemma 6.2.3.** *Let* $(A, B, Z) \in \mathbb{IR}^{n\times n} \times \mathbb{IR}^n \times \mathbb{IR}^n$ *and* $(\tilde{A}, \tilde{B}, \tilde{Z}) \in \mathbb{I}A \times \mathbb{I}B \times \mathbb{I}Z$*.*

1. *If* $(A, B, Z) \in \mathcal{D}^*$*, then* $(\tilde{A}, \hat{B}, \hat{Z}) \in \mathcal{D}^*$*,* $\forall \hat{B}, \hat{Z} \in \mathbb{IR}^n$*.*

2. *If* $(A, B, Z) \in \mathcal{D}^*$*, then* $\Gamma^*(\tilde{A}, \tilde{B}, \tilde{Z}) \subset \Gamma^*(A, B, Z)$*.*

3. *If* $(A, B, Z) \in \mathcal{D}^*$*, then* $\Gamma^+(A, B, Z) \subset Z$*.*

4. *If* $(A, B, Z) \in \mathcal{D}^*$ *and* $\Gamma(A, B, Z) \neq \emptyset$*, then* $\Gamma^+(A, B, Z) = \Gamma(A, B, Z)$*.*

5. *If* $(A, B, Z) \in \mathcal{D}^*$ *and* $\Gamma^*(A, B, Z) \subset Z$*, then* $\Gamma^*(A, B, Z) = \Gamma(A, B, Z)$*.*

6. *If* $\emptyset \neq \Gamma(A, B, Z) \subset \mathrm{int}(Z)$*, then* $(A, B, Z) \in \mathcal{D}^*$ *and* $\Gamma^*(A, B, Z) = \Gamma(A, B, Z)$*.*

*Proof.* Conclusion 1 is obvious and 2 follows from inclusion monotonicity of interval arithmetic. Conclusion 3 follows from Conclusion 3 of Lemma 2.5.23. To show 4 and

5, suppose $(A, B, Z) \in \mathcal{D}^*$ and denote $\Gamma(A, B, Z) \equiv W_1 \times \ldots \times W_n$,

$$W_i = Z_i \cap \frac{1}{A_{ii}}\bigg(B_i - \sum_{k<i} A_{ik}W_k - \sum_{k>i} A_{ik}Z_k\bigg), \quad \forall i = 1, \ldots, n. \tag{6.4}$$

Define $W_i^+$ as in (6.3), choose any $i \in \{1, \ldots, n\}$ and assume that $W_i = W_i^+$ for all $k < i$, which is trivially true if $i = 1$. Then, comparing (6.4) and (6.3), Conclusion 1 of Lemma 2.5.23 implies that $W_i = W_i^+$ if $W_i \neq \emptyset$. Then, Conclusion 4 follows by finite induction.

To show 5, define $W_i^*$ as in (6.2) and assume that $W_i = W_i^*$ for all $k < i$, which is again trivially true if $i = 1$. Comparing (6.4) and (6.2) yields $W_i = Z_i \cap W_i^*$. But the assumption that $\Gamma^*(A, B, Z) \subset Z$ implies that $W_i^* \subset Z_i$, and hence $W_i = W_i^*$. Therefore, Conclusion 5 also follows by finite induction.

To show 6, suppose $\emptyset \neq \Gamma(A, B, Z) \subset \text{int}(Z)$. Theorem 4.4.5 (ii) of [131] implies $(A, B, Z) \in \mathcal{D}^*$. Now denoting $\Gamma(A, B, Z) \equiv W_1 \times \ldots \times W_n$, (6.4) again holds. Assuming that $W_i = W_i^*$ for all $k < i$ (trivial for $i = 1$) and comparing (6.4) and (6.2) again yields $W_i = Z_i \cap W_i^*$. The assumption that $\Gamma(A, B, Z) \subset \text{int}(Z)$ implies that $W_i \subset \text{int}(Z_i)$, which is only possible if $W_i = W_i^*$. Then, Conclusion 6 follows by finite induction. □

The following definition formalizes the notation $\mathcal{H}$ from Corollary 5.4.7, with a slight modification to reflect the fact that, in the proposed methods, the reference point $\tilde{\mathbf{z}}_y$ is a function of $Z_y$ and does not need to be specified independently. Notation is also introduced for iterative application of $\mathcal{H}$, and extended forms based on $\Gamma^+$ and $\Gamma^*$ are defined.

**Definition 6.2.4.** Let $\tilde{\mathbf{z}}_y : \mathbb{I}D_y \to \mathbb{R}^{n_y}$, define $M_\Gamma : \mathbb{I}D_t \times \mathbb{I}D_p \times \mathbb{I}D_x \times \mathbb{I}D_y \times \mathbb{IR}^{n_y \times n_y} \to \mathbb{IR}^{n_y \times n_y} \times \mathbb{IR}^{n_y} \times \mathbb{IR}^{n_y}$ by

$$M_\Gamma(I, P, Z_x, Z_y, \mathbf{C}) \equiv \bigg(\mathbf{C}\left[\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\right](I, P, Z_x, Z_y), -\mathbf{C}\left[\mathbf{g}\right](I, P, Z_x, \tilde{\mathbf{z}}(Z_y)), Z_y - \tilde{\mathbf{z}}(Z_y)\bigg),$$

and define the set

$$\mathcal{D}^*_{\mathcal{H}} \equiv \big\{(I, P, Z_x, Z_y, \mathbf{C}) \in \mathbb{I}D_t \times \mathbb{I}D_p \times \mathbb{I}D_x \times \mathbb{I}D_y \times \mathbb{IR}^{n_y \times n_y} : \\ M_\Gamma(I, P, Z_x, Z_y, \mathbf{C}) \in \mathcal{D}^*\big\}.$$

For every $K \in \mathbb{N}$, let $\mathcal{H}^K : \mathbb{I}D_t \times \mathbb{I}D_p \times \mathbb{I}D_x \times \mathbb{I}D_y \times \mathbb{IR}^{n_y \times n_y} \to \mathbb{IR}^{n_y}$ be defined by $\mathcal{H}^K(I, P, Z_x, Z^0_y, \mathbf{C}) \equiv Z^K_y$, where $Z^{k+1}_y = \tilde{\mathbf{z}}(Z^k_y) + \Gamma\left(M_\Gamma(I, P, Z_x, Z^k_y, \mathbf{C})\right)$, $\forall k \in \{0, \ldots, K-1\}$. Furthermore, define $\mathcal{H}^{+,K} : \mathcal{D}^*_{\mathcal{H}} \to \mathbb{IR}^{n_y}$ exactly as $\mathcal{H}^K$ with $\Gamma^+$ in place of $\Gamma$, and define $\mathcal{H}^* : \mathcal{D}^*_{\mathcal{H}} \to \mathbb{IR}^{n_y}$ exactly as $\mathcal{H}^1$ with $\Gamma^*$ in place of $\Gamma$. Finally, define the set

$$\mathcal{D}^K_{\mathcal{H}} \equiv \left\{(I, P, Z_x, Z_y, \mathbf{C}) \in \mathcal{D}^*_{\mathcal{H}} : \mathcal{H}^K(I, P, Z_x, Z_y, \mathbf{C}) \neq \emptyset\right\}.$$

For simplicity, the superscript $K$ on $\mathcal{H}^K$ and $\mathcal{H}^{+,K}$ will be omitted when $K = 1$. When $K > 1$, some justification for Definition 6.2.4 is needed. For any $k \in \{0, \ldots, K-1\}$ with $Z^k_y \in \mathbb{I}D_y$, the definition of $\Gamma$ implies that $Z^{k+1}_y \subset Z^k_y$, and hence $Z^{k+1}_y \in \mathbb{I}D_y$. Then, a simple inductive argument shows that $\mathcal{H}^K$ is well-defined for any $K \in \mathbb{N}$. In the definition of $\mathcal{H}^{+,K}$, we similarly note that $(I, P, Z_x, Z^k_y, \mathbf{C}) \in \mathcal{D}^*_{\mathcal{H}}$ implies $Z^{k+1}_y \subset Z^k_y$ by Conclusion 3 of Lemma 6.2.3. It follows by Conclusion 1 of Lemma 6.2.5 below that $(I, P, Z_x, Z^{k+1}_y, \mathbf{C}) \in \mathcal{D}^*_{\mathcal{H}}$, so that again induction shows that $\mathcal{H}^{+,K}$ is well-defined.

In Definition 6.2.4, the preconditioner $\mathbf{C}$ is allowed to be an interval matrix. This makes $\mathcal{H}^*$, $\mathcal{H}^{+,K}$ and $\mathcal{H}^K$ pure interval functions and is only done for consistency with the results on regularity of interval functions in the next section. In the proposed methods, $\mathbf{C}$ will always be a real matrix. To conform with Definition 6.2.4, $\mathbf{C}$ is simply identified with the corresponding degenerate element of $\mathbb{IR}^{n_y \times n_y}$.

Specific definitions for $\tilde{\mathbf{z}}$ will be given when $\mathcal{H}^K$, $\mathcal{H}^{+,K}$ or $\mathcal{H}^*$ are used in later sections. The results in the remainder of this section are independent of this choice.

**Lemma 6.2.5.** *Let $K \in \mathbb{N}$, let $(I, P, Z_x, Z_y, \mathbf{C}) \in \mathbb{I}D_t \times \mathbb{I}D_p \times \mathbb{I}D_x \times \mathbb{I}D_y \times \mathbb{IR}^{n_y \times n_y}$ and let $(\tilde{I}, \tilde{P}, \tilde{Z}_x, \tilde{Z}_y, \tilde{\mathbf{C}}) \in \mathbb{I}I \times \mathbb{I}P \times \mathbb{I}Z_x \times \mathbb{I}Z_y \times \mathbb{I}\mathbf{C}$.*

1. *If* $(I, P, Z_x, Z_y, \mathbf{C}) \in \mathcal{D}^*_{\mathcal{H}}$, *then* $(\tilde{I}, \tilde{P}, \tilde{Z}_x, \tilde{Z}_y, \tilde{\mathbf{C}}) \in \mathcal{D}^*_{\mathcal{H}}$.
2. *If* $(I, P, Z_x, Z_y, \mathbf{C}) \in \mathcal{D}^*_{\mathcal{H}}$, *then* $\mathcal{H}^*(\tilde{I}, \tilde{P}, \tilde{Z}_x, Z_y, \tilde{\mathbf{C}}) \subset \mathcal{H}^*(I, P, Z_x, Z_y, \mathbf{C})$.
3. *If* $(I, P, Z_x, Z_y, \mathbf{C}) \in \mathcal{D}^*_{\mathcal{H}}$, *then* $\mathcal{H}^{+,K}(I, P, Z_x, Z_y, \mathbf{C}) \subset Z_y$.
4. *If* $(I, P, Z_x, Z_y, \mathbf{C}) \in \mathcal{D}^K_{\mathcal{H}}$, *then* $\mathcal{H}^K(I, P, Z_x, Z_y, \mathbf{C}) = \mathcal{H}^{+,K}(I, P, Z_x, Z_y, \mathbf{C})$.
5. *If* $(I, P, Z_x, Z_y, \mathbf{C}) \in \mathcal{D}^*_{\mathcal{H}}$ *and* $\mathcal{H}^*(I, P, Z_x, Z_y, \mathbf{C}) \subset Z_y$, *then*

$$\mathcal{H}(I, P, Z_x, Z_y, \mathbf{C}) = \mathcal{H}^*(I, P, Z_x, Z_y, \mathbf{C}). \tag{6.5}$$

6. *If* $\emptyset \neq \mathcal{H}(I, P, Z_x, Z_y, \mathbf{C}) \subset \text{int}(Z_y)$, *then* $(I, P, Z_x, Z_y, \mathbf{C}) \in \mathcal{D}^*_{\mathcal{H}}$ *and* (6.5) *holds.*
7. *If* $\emptyset \neq \mathcal{H}(I, P, Z_x, Z_y, \mathbf{C}) \subset \text{int}(Z_y)$, $\tilde{\mathbf{z}}_y(Z_y) \in \text{int}(Z_y)$, *and* $\mathbf{C}$ *is degenerate, then* $(\tilde{I}, \tilde{P}, \tilde{Z}_x, Z_y, \mathbf{C}) \in \mathcal{D}^K_{\mathcal{H}}$.

*Proof.* Conclusions 1 and 2 follow from inclusion monotonicity of interval arithmetic and the corresponding conclusions of Lemma 6.2.3 (it is essential in 2 that $Z_y$, and not $\tilde{Z}_y$, appears on the left, since otherwise $\tilde{\mathbf{z}}$ will be modified and inclusion monotonicity does not apply). Conclusion 3 was argued inductively in the discussion above. Conclusion 4 follows by inductive application of Conclusion 4 in Lemma 6.2.3. Conclusions 5 and 6 are direct applications of the corresponding conclusions of Lemma 6.2.3. Assume the hypotheses of 7. By Conclusion 3 of Corollary (5.4.7), to every $(t, \mathbf{p}, \mathbf{z}_x) \in I \times P \times Z_x$ there corresponds some $\mathbf{z}_y \in Z_y$ satisfying $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$. Choosing any $(t, \mathbf{p}, \mathbf{z}_x) \in \tilde{I} \times \tilde{P} \times \tilde{Z}_x$, Conclusion 1 of the same shows that the corresponding $\mathbf{z}_y$ must be in $\mathcal{H}^K(\tilde{I}, \tilde{P}, \tilde{Z}_x, Z_y, \mathbf{C})$. By Conclusion 1 of the present lemma, this implies $(\tilde{I}, \tilde{P}, \tilde{Z}_x, Z_y, \mathbf{C}) \in \mathcal{D}^K_{\mathcal{H}}$. □

### 6.2.2 Regularity of Interval Functions

Recall the interval extensions $[\mathbf{f}]$, $[\mathbf{g}]$ and $\left[\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\right]$. For certain computations required by the proposed bounding methods, these mappings, as well as others defined in the previous section, will be requried to be piecewise $C^1$ as defined in §2.5.3.

**Assumption 6.2.6.** Let $c : D_t \times D_p \times D_x \times D_y \to \mathbb{R}$ represent any of $f_i$, $g_j$ or $\frac{\partial g_j}{\partial y_k}$, with indices $i \in \{1, \ldots, n_x\}$ and $j, k \in \{1, \ldots, n_y\}$. The interval extension $[c]$ is piecewise $C^1$ on the open set $\mathbb{I}D_t \times \mathbb{I}D_p \times \mathbb{I}D_x \times \mathbb{I}D_y$.

**Remark 6.2.7.** When $c$ is $\mathcal{L}$-factorable and $[c]$ is the natural interval extension (as it is in our implementation), Assumption 6.2.6 holds under minor restrictions on the factors of $c$. As shown in §2.5.5, if the interval extension of each univariate function in $\mathcal{L}$ is piecewise $C^1$ on an open domain, then $[c]$ is piecewise $C^1$ by Theorem 2.5.34.

We now establish that several other interval mappings of interest are also piecewise $C^1$.

**Lemma 6.2.8.** *$\mathcal{D}^*$ is open and both $\Gamma^+$ and $\Gamma^*$ are piecewise $C^1$ on $\mathcal{D}^*$.*

*Proof.* Let $U \equiv \{(\mathbf{A}, \mathbf{b}, \mathbf{z}) \in \mathbb{R}^{n \times n} \times \mathbb{R}^n \times \mathbb{R}^n : \mathbf{A}_{ii} \neq 0, \forall i = 1, \ldots, n\}$. By definition, $\mathbb{I}U = \mathcal{D}^*$. Since $U$ is open, $\mathcal{D}^*$ is open by Lemma 2.5.32. It follows from (6.2), the rules of interval addition, subtraction, multiplication and division (see [131]), and Conclusion 2 of Lemma 2.5.13 that $\Gamma^*$ is piecewise $C^1$ on $\mathcal{D}^*$. For $\Gamma^+$, (6.3) leads to the same conclusion by additionally applying Lemmas 2.5.25 and 2.5.21. □

**Theorem 6.2.9.** *Suppose Assumption 6.2.6 holds and the function $\tilde{\mathbf{z}}_y$ in Definition 6.2.4 is piecewise $C^1$ on $\mathbb{I}D_y$. Then $\mathcal{D}^*_{\mathcal{H}}$ is open and $\mathcal{H}^{K,+}$ and $\mathcal{H}^*$ are piecewise $C^1$ on $\mathcal{D}^*_{\mathcal{H}}$.*

*Proof.* Under the stated hypotheses, it follows from the rules of interval addition, subtraction and multiplication and Conclusion 2 of Lemma 2.5.13 that $M_\Gamma$ in Definition 6.2.4 is piecewise $C^1$ on $\mathbb{I}D_t \times \mathbb{I}D_p \times \mathbb{I}D_x \times \mathbb{I}D_y \times \mathbb{IR}^{n_y \times n_y}$. By Lemma 6.2.8, $\Gamma^*$ and $\Gamma^+$ are piecewise $C^1$ on $\mathcal{D}^*$, which is open. Then Lemma 2.5.21 implies that $\mathcal{D}^*_{\mathcal{H}}$ is open and $\Gamma^* \circ M_\Gamma$ is piecewise $C^1$ there, so that $\mathcal{H}^*$ is piecewise $C^1$ on $\mathcal{D}^*_{\mathcal{H}}$ by the hypothesis on $\tilde{\mathbf{z}}_y$ and Conclusion 2 of Lemma 2.5.13. For $\mathcal{H}^{+,K}$, we additionally note that $(I, P, Z_x, Z^k_y, \mathbf{C}) \in \mathcal{D}^*_{\mathcal{H}}$ for all $k \in \{0, \ldots, K-1\}$ (see discussion following Definition 6.2.4). Then, the result follows by $K$ applications of Lemmas 2.5.21 and Lemma 2.5.13. □

## 6.3 A Generic Two-Phase Algorithm

In this section, we introduce the first bounding method of this chapter, which is based on a time-stepping framework outlined in Algorithm 1 below. In a generic time step $j$, the algorithm proceeds in two phases. The purpose of Phase 1 is to establishes existence and uniqueness of a solution $(\mathbf{x},\mathbf{y})$ of (5.1) on $I_j \times P$, for some time interval $I_j = [t_{j-1}, t_j]$, and to determine crude enclosures $Z_{x,j}$ and $Z'_{y,j}$ satisfying

$$(\mathbf{x}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p})) \in Z_{x,j} \times Z'_{y,j}, \quad \forall (t,\mathbf{p}) \in I_j \times P. \tag{6.6}$$

Subsequently, Phase 2 computes refined intervals $X_j \subset Z_{x,j}$ and $Y_j \subset Z'_{y,j}$ such that

$$(\mathbf{x}(t_j,\mathbf{p}),\mathbf{y}(t_j,\mathbf{p})) \in X_j \times Y_j, \quad \forall \mathbf{p} \in P. \tag{6.7}$$

In contrast to $Z_{x,j}$ and $Z'_{y,j}$, the refined bounds $X_j$ and $Y_j$ are valid only at $t_j$. The method for computing these refinements is not specified in Algorithm 1. Our approach is the subject of §6.5.

As input, Algorithm 1 takes intervals $I = [t_0, t_f] \subset D_t$, $P \subset D_p$ and $X_0 \subset D_x$ under the assumption that $\mathbf{x}_0(P) \subset X_0$, $\forall \mathbf{p} \in P$. The final input is a vector $\hat{\mathbf{y}}_0 \in D_y$ satisfying $\mathbf{g}(t_0, \hat{\mathbf{p}}, \mathbf{x}_0(\hat{\mathbf{p}}), \hat{\mathbf{y}}_0) = \mathbf{0}$ for some $\hat{\mathbf{p}} \in P$. The purpose of this vector is to specify a particular solution of interest in case the DAE in question permits multiple regular solutions (see Example 5.3.1). Phases 1 and 2 described above correspond to Steps 3 and 6, respectively. Finally, the algorithm makes use of the functions $\mathcal{H}^K$ and $\tilde{\mathbf{z}}_y$ from Definition 6.2.4, and is independent of the choice of $\tilde{\mathbf{z}}_y$. Choices for $\tilde{\mathbf{z}}_y$ and $\mathbf{C}$ are discussed in §6.4.1.

**Algorithm** 1 (Two-phase algorithm)

1. Input: $I = [t_0, t_f]$, $P$, $X_0$, $\hat{\mathbf{y}}_0$.

2. Initialize $j := 1$, $Y_0 := [\hat{\mathbf{y}}_0, \hat{\mathbf{y}}_0]$.

3. Find $I_j = [t_{j-1}, t_j]$, $Z_{x,j}$, $Z_{y,j}$ and $\mathbf{C}_j$ satisfying

$$(I_j, P, Z_{x,j}, Z_{y,j}, \mathbf{C}_j) \in \mathbb{I}D_t \times \mathbb{I}D_p \times \mathbb{I}D_x \times \mathbb{I}D_y \times \mathbb{R}^{n_y \times n_y}, \tag{6.8}$$

$$Y_{j-1} \subset Z_{y,j}, \tag{6.9}$$

$$\tilde{\mathbf{z}}_y(Z_{y,j}) \in \text{int}(Z_{y,j}), \tag{6.10}$$

$$\emptyset \neq Z'_{y,j} \equiv \mathcal{H}(I_j, P, Z_{x,j}, Z_{y,j}, \mathbf{C}_j) \subset \text{int}(Z_{y,j}), \tag{6.11}$$

$$X_{j-1} + [0, t_j - t_{j-1}]\,[\mathbf{f}]\,(I_j, P, Z_{x,j}, Z'_{y,j}) \subset Z_{x,j}. \tag{6.12}$$

4. Set $X_j := Z_{x,j}$ and $Y_j := Z'_{y,j}$. If $j = 1$, set $Y_0 := Z'_{y,j}$.

5. If $j = 1$, refine $Y_0$ (see §6.5).

6. Refine $X_j$ and $Y_j$ (see §6.5).

7. If $t_j \geq t_f$, terminate. Otherwise, set $j := j + 1$ and go to 3.

The behavior of Algorithm 1 is formalized in Corollary 6.3.2 below. Of course, this depends on the refinement procedures in Steps 5 and 6, which have not yet been specified. Therefore, we assume the following:

**Assumption 6.3.1.** Consider an iteration $J \in \mathbb{N}$ of Algorithm 1 and suppose that Steps 3-4 are complete. Let $(\mathbf{x}, \mathbf{y})$ be a regular solution of (5.1) on $[t_0, t_J] \times P$ satisfying (6.6) for all $j \in \{1, \ldots, J\}$. If $J = 1$, the refinement to $Y_0$ computed in Step 5 satisfies (6.7) with $j = 0$. Suppose that Step 5 is complete. If $(\mathbf{x}, \mathbf{y})$ additionally satisfies (6.7) for all $j \in \{0, \ldots, J-1\}$, then Step 6 produces $X_J$ and $Y_J$ satisfying (6.7) with $j = J$.

**Corollary 6.3.2.** *Let* $(I, P, X_0, \hat{\mathbf{y}}_0) \in \mathbb{I}D_t \times \mathbb{I}D_p \times \mathbb{I}D_x \times D_y$ *satisfy* $\mathbf{x}_0(\mathbf{p}) \in X_0$, $\forall \mathbf{p} \in P$, *and* $\mathbf{g}(t_0, \hat{\mathbf{p}}, \mathbf{x}_0(\hat{\mathbf{p}}), \hat{\mathbf{y}}_0) = \mathbf{0}$ *for some* $\hat{\mathbf{p}} \in P$. *Suppose that Algorithm 1 has completed* $J$ *iterations, furnishing the intervals* $Y_0$ *and*

$$I_j,\ Z_{x,j},\ Z_{y,j},\ Z'_{y,j},\ X_j,\ Y_j, \quad j = 1, \ldots, J. \tag{6.13}$$

*Then there exists a regular solution* $(\mathbf{x}, \mathbf{y})$ *of* (5.1) *on* $[t_0, t_J] \times P$ *with* $\mathbf{y}(t_0, \hat{\mathbf{p}}) = \hat{\mathbf{y}}_0$, *satisfying* (6.6) *for every* $j \in \{1, \ldots, J\}$ *and* (6.7) *for every* $j \in \{0, \ldots, J\}$.

*Furthermore, for any $\tilde{I} = [t_0, \tilde{t}] \subset [t_0, t_J]$, any connected $\tilde{P} \subset P$, and any solution $(\mathbf{x}^*, \mathbf{y}^*)$ of (5.1) on $\tilde{I} \times \tilde{P}$, either $(\mathbf{x}^*, \mathbf{y}^*) = (\mathbf{x}, \mathbf{y})$ on $\tilde{I} \times \tilde{P}$, or $\mathbf{y}^*(t_0, \mathbf{p}) \notin Z_{y,1}$, $\forall \mathbf{p} \in \tilde{P}$.*

*Proof.* Define $(\mathbf{x}^*, \mathbf{y}^*)$ as above and suppose that $\mathbf{y}^*(t_0, \mathbf{p}) \in Z_{y,1}$ for at least one $\mathbf{p} \in \tilde{P}$. Consider the following inductive hypotheses for $k \in \{1, \ldots, J\}$:

1. There exists a regular solution $(\mathbf{x}, \mathbf{y})$ of (5.1) on $[t_0, t_k] \times P$,

2. $(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^*, \mathbf{y}^*)$ on $[t_0, \min(t_k, \tilde{t})] \times \tilde{P}$,

3. (6.6) holds for $j \in \{1, \ldots, k\}$,

4. $\mathbf{y}(t_0, \hat{\mathbf{p}}) = \hat{\mathbf{y}}_0$,

5. (6.7) holds for $j \in \{0, \ldots, k\}$.

It suffices to show that these hypotheses hold with $k = J$.

Let $k = 1$. Since (6.8)-(6.12) hold with $j = 1$, Theorem 5.4.8 establishes Hypotheses 1-3. Because $\hat{\mathbf{y}}_0$ is a zero of $\mathbf{g}(t_0, \hat{\mathbf{p}}, \mathbf{x}_0(\hat{\mathbf{p}}), \cdot)$ and $\hat{\mathbf{y}}_0 \in Z_{y,1}$ by (6.9), Hypothesis 4 follows from Conclusion 3 of Corollary 5.4.7. Applying Assumption 6.3.1 with $J = 1$ proves Hypothesis 5.

Choose any $k \in \{1, \ldots, J-1\}$ and assume Hypotheses 1-5. Since $\mathbf{x}(t_k, P) \subset X_k$ and (6.8)-(6.12) hold with $j = k+1$, Theorem 5.4.8 furnishes a regular solution of (5.1a) on $I_{k+1} \times P$, $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in C^1(I_{k+1} \times P, Z_{x,k+1}) \times C^1(I_{k+1} \times P, Z'_{y,k+1})$, satisfying $\hat{\mathbf{x}}(t_k, \mathbf{p}) = \mathbf{x}(t_k, \mathbf{p})$, $\forall \mathbf{p} \in P$. Noting that both $\mathbf{y}(t_k, \mathbf{p})$ and $\hat{\mathbf{y}}(t_k, \mathbf{p})$ are zeros of $\mathbf{g}(t_k, \mathbf{p}, \mathbf{x}(t_k, \mathbf{p}), \cdot)$ and $\mathbf{y}(t_k, \mathbf{p}) \in Y_k \subset Z_{y,k+1}$ by (6.9), it follows from Conclusion 3 of Corollary 5.4.7 that $\mathbf{y}(t_k, \mathbf{p}) = \hat{\mathbf{y}}(t_k, \mathbf{p})$, $\forall \mathbf{p} \in P$. If $\tilde{t} \geq t_k$, Hypothesis 2 implies that we also have $\hat{\mathbf{x}}(t_k, \mathbf{p}) = \mathbf{x}^*(t_k, \mathbf{p})$ and $\hat{\mathbf{y}}(t_k, \mathbf{p}) = \mathbf{y}^*(t_k, \mathbf{p})$, $\forall \mathbf{p} \in \tilde{P}$, so that $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = (\mathbf{x}^*, \mathbf{y}^*)$ on $[t_k, \min(t_{k+1}, \tilde{t})] \times \tilde{P}$ by Theorem 5.4.8.

From the arguments above, $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ extends $(\mathbf{x}, \mathbf{y})$ onto all of $[t_0, t_{k+1}] \times P$, and this extension satisfies Hypothesis 1-4 with $k := k+1$. Applying Assumption 6.3.1 with $J = k+1$ establishes Hypotheses 5, and finite induction completes the proof. □

From Corollary 6.3.2, it is clear that Algorithm 1 produces bounds on a single, isolated solution of (5.1) specified by the input $\hat{\mathbf{y}}_0$. This input can be ignored by omitting (6.9) when $j = 1$. However, the algorithm still produces bounds on a unique solution dictated by the interval $Z_{y,1}$ found in the first time step. If one is interested in bounds on all solutions, then Algorithm 1 would need to be applied to each solution in turn, though it has no provisions for exhaustively enumerating solutions. This problem is not pursued in this thesis, though a good starting point is provided by Theorem 5.5.2. On the other hand, if there is a particular solution of interest, then Algorithm 1 avoids any unnecessary conservatism that would result from bounding other solutions as well.

## 6.4 Satisfying the Existence and Uniqueness Test Computationally (Phase 1)

In this section, the execution of Step 3 in a single time step $J$ of Algorithm 1 is considered. Based on the previous time step, it is assumed that there exists a regular solution $(\mathbf{x}, \mathbf{y})$ of (5.1) on $[t_0, t_{J-1}] \times P$ satisfying $\mathbf{y}(t_0, \hat{\mathbf{p}}) = \hat{\mathbf{y}}_0$ and $\mathbf{x}(t_{J-1}, P) \subset X_{J-1}$. The objective is to derive an automatic computational procedure for finding intervals $I_J$, $Z_{x,J}$, $Z_{y,J}$ and $\mathbf{C}_J$ satisfying (6.8)-(6.12). Though we present an effective method for this task, it is generally impossible to guarantee that such intervals can be found. This seems to be an inherent complication owing to the implicit nature of nonlinear DAEs, and hence of the inclusion (6.11), and it appears in much the same form in both of the methods in [142] and [83]. However, it is important to note that the validity of any intervals provided by Step 3 is guaranteed, regardless of the method used to find them. The proposed procedure will either succeed in satisfying (6.8)-(6.12), and hence (6.6) with $j = J$, or it will fail and report an error, forcing Algorithm 1 to terminate prematurely.

Since the implicit conditions (6.11) and (6.12) are the most challenging, they are addressed first. The key insight used to satisfy these conditions is that, once some putative $\mathbf{C}_J$ and $t_J$ have been chosen, intervals $Z_{x,J}$ and $Z_{y,J}$ satisfying (6.11) and (6.12) are related to solutions of a square system of real-valued algebraic equations that can be solved by standard methods with a few caveats. This approach is developed below. A complete algorithm for satisfying all of the conditions (6.8)-(6.12) is presented in §6.4.2.

**Lemma 6.4.1.** *The conditions* (6.8) *and* (6.11)*, with $j = J$, are equivalent to*

$$(I_J, P, Z_{x,J}, Z_{y,J}, \mathbf{C}_J) \in \mathcal{D}^*_{\mathcal{H}}, \tag{6.14}$$

$$\mathcal{H}^*(I_J, P, Z_{x,J}, Z_{y,J}, \mathbf{C}_J) \subset \mathrm{int}(Z_{y,J}), \tag{6.15}$$

*provided that $\mathbf{C}_J$ is degenerate.*

*Proof.* The result is a direct application of Conclusions 5 and 6 of Lemma 6.2.5. □

For the following result, denote $[\mathbf{x}^L_{J-1}, \mathbf{x}^U_{J-1}] \equiv X_{J-1}$ and

$$[\mathcal{H}^{*,L}(I, P, Z_x, Z_y, \mathbf{C}), \mathcal{H}^{*,U}(I, P, Z_x, Z_y, \mathbf{C})] \equiv \mathcal{H}^*(I, P, Z_x, Z_y, \mathbf{C}). \tag{6.16}$$

**Lemma 6.4.2.** *Let $I_J \equiv [t_{J-1}, t_J] \in \mathbb{I}D_t$, $P \in \mathbb{I}D_p$, $\mathbf{C}_J \in \mathbb{R}^{n_y \times n_y}$ and $\gamma > 0$. If the vectors $\mathbf{z}^L_x, \mathbf{z}^U_x \in \mathbb{R}^{n_x}$ and $\mathbf{z}^L_y, \mathbf{z}^U_y \in \mathbb{R}^{n_y}$ satisfy*

$$(I_J, P, \Box(\mathbf{z}^L_x, \mathbf{z}^U_x), \Box(\mathbf{z}^L_y, \mathbf{z}^U_y), \mathbf{C}_J) \in \mathcal{D}^*_{\mathcal{H}}, \tag{6.17}$$

$$\mathbf{z}'^{\,L}_y := \mathcal{H}^{*,L}(I_J, P, \Box(\mathbf{z}^L_x, \mathbf{z}^U_x), \Box(\mathbf{z}^L_y, \mathbf{z}^U_y), \mathbf{C}_J), \tag{6.18}$$

$$\mathbf{z}'^{\,U}_y := \mathcal{H}^{*,U}(I_J, P, \Box(\mathbf{z}^L_x, \mathbf{z}^U_x), \Box(\mathbf{z}^L_y, \mathbf{z}^U_y), \mathbf{C}_J), \tag{6.19}$$

$$\mathbf{0} = \mathbf{z}^L_y - \mathbf{z}'^{\,L}_y + \mathbf{1}\gamma, \tag{6.20}$$

$$\mathbf{0} = -\mathbf{z}^U_y + \mathbf{z}'^{\,U}_y + \mathbf{1}\gamma, \tag{6.21}$$

$$\mathbf{0} = \mathbf{z}^L_x - \mathbf{x}^L_{J-1} - [0, t_J - t_{J-1}][\mathbf{f}]^L(I_J, P, \Box(\mathbf{z}^L_x, \mathbf{z}^U_x), \Box(\mathbf{z}'^{\,L}_y, \mathbf{z}'^{\,U}_y)) + \mathbf{1}\gamma, \tag{6.22}$$

$$\mathbf{0} = -\mathbf{z}^U_x + \mathbf{x}^U_{J-1} + [0, t_J - t_{J-1}][\mathbf{f}]^U(I_J, P, \Box(\mathbf{z}^L_x, \mathbf{z}^U_x), \Box(\mathbf{z}'^{\,L}_y, \mathbf{z}'^{\,U}_y)) + \mathbf{1}\gamma, \tag{6.23}$$

*then $\mathbf{z}^L_x < \mathbf{z}^U_x$ and $\mathbf{z}^L_y < \mathbf{z}^U_y$, and $Z_{x,J} \equiv [\mathbf{z}^L_x, \mathbf{z}^U_x]$ and $Z_{y,J} \equiv [\mathbf{z}^L_y, \mathbf{z}^U_y]$ satisfy* (6.8)*,* (6.11) *and* (6.12) *with $j = J$. Furthermore, these conclusions remain true if the* *right-hand sides of* (6.20)-(6.23) *are componentwise less than* $\gamma$.

*Proof.* It suffices to prove the case where the right-hand sides of (6.20)-(6.23) are componentwise less than $\gamma$. Since $\mathcal{H}^*$ returns an interval, $\mathbf{z}'^{L}_{y} \leq \mathbf{z}'^{U}_{y}$ and hence

$$\mathbf{z}_y^L < \mathbf{z}'^{L}_{y} \leq \mathbf{z}'^{U}_{y} < \mathbf{z}_y^U. \tag{6.24}$$

An analogous argument shows that $\mathbf{z}_x^L < \mathbf{z}_x^U$.

Let $Z_{x,J}$ and $Z_{y,J}$ be as in the statement of the lemma, and let $Z'_y = [\mathbf{z}'^{L}_{y}, \mathbf{z}'^{U}_{y}]$. Then, (6.17) implies (6.14) and (6.24) implies (6.15). Then, (6.8) and (6.11) follow from Lemma 6.4.1. Again, an argument analogous to (6.24) shows that $X_{J-1}+[0, t_J - t_{J-1}][\mathbf{f}](I_J, P, Z_{x,J}, Z'_{y,J}) \subset \text{int}(Z_{x,J})$, which implies (6.12). □

Equations (6.20)-(6.23) form a system of nonlinear algebraic equations of the general form

$$\mathbf{L}(\mathbf{z}) = \mathbf{0}, \tag{6.25}$$

where $\mathbf{z}$ is a concatenation of the vectors $\mathbf{z}_x^L$, $\mathbf{z}_x^U$, $\mathbf{z}_y^L$ and $\mathbf{z}_y^U$, and the domain of $\mathbf{L}$ is specified by (6.17). To compute intervals satisfying the existence and uniqueness conditions (6.8), (6.11) and (6.12), (6.25) is solved using a Newton-type iteration of the form

$$\mathbf{z}^{k+1} := \mathbf{z}^k - \tilde{\mathbf{J}}^{-1}(\mathbf{z}^k)\mathbf{L}(\mathbf{z}^k) \tag{6.26}$$

(this should not be confused with the interval Newton method used to derive $\mathcal{H}^*$, and hence equations (6.20) and (6.21)). During this iteration, we may terminate whenever $\mathbf{L}(\mathbf{z}^k) < \mathbf{1}\gamma$ for some iterate, and Lemma 6.4.2 ensures that $\mathbf{z}^k$ furnishes the desired intervals. Using the definition of $\mathcal{H}^*$ and the rules of interval arithmetic, it is in principle possible to write out explicit expressions for the functions $\mathbf{L}$, though they may be very cumbersome. Then, the only complication with this approach is that $\mathbf{L}$ is in general nonsmooth owing to the rules of interval arithmetic. Even so, the developments of §6.2.2 imply sufficient regularity of $\mathbf{L}$ for a Newton-type method to be well motivated.

**Lemma 6.4.3.** *Let* $I_J \equiv [t_{J-1}, t_J] \in \mathbb{I}D_t$*,* $P \in \mathbb{I}D_p$*,* $\mathbf{C}_J \in \mathbb{IR}^{n_y \times n_y}$ *and* $\gamma > 0$*. Suppose Assumption 6.2.6 holds and the function* $\tilde{\mathbf{z}}_y$ *in Definition 6.2.4 is piecewise* $C^1$ *on* $\mathbb{I}D_y$*. Then the set*

$$E^*_{\mathcal{H}} \equiv \left\{(\mathbf{z}^L_x, \mathbf{z}^U_x, \mathbf{z}^L_y, \mathbf{z}^U_y) \in \mathbb{R}^{2(n_x+n_y)} : (I_J, P, \square(\mathbf{z}^L_x, \mathbf{z}^U_x), \square(\mathbf{z}^L_y, \mathbf{z}^U_y), \mathbf{C}_J) \in \mathcal{D}^*_{\mathcal{H}}\right\} \quad (6.27)$$

*is open and* $\mathbf{L}$ *is Frechet differentiable a.e. in* $E^*_{\mathcal{H}}$*.*

*Proof.* Define $\phi : \mathbb{R}^{2(n_x+n_y)} \to \mathbb{IR} \times \mathbb{IR}^{n_p} \times \mathbb{IR}^{n_x} \times \mathbb{IR}^{n_y} \times \mathbb{IR}^{n_y \times n_y}$ by

$$\phi(\mathbf{z}^L_x, \mathbf{z}^U_x, \mathbf{z}^L_y, \mathbf{z}^U_y) \equiv (I_J, P, \square(\mathbf{z}^L_x, \mathbf{z}^U_x), \square(\mathbf{z}^L_y, \mathbf{z}^U_y), \mathbf{C}_J). \quad (6.28)$$

By Lemma 2.5.19, $\phi$ is piecewise $C^1$ on $\mathbb{R}^{2(n_x+n_y)}$. By Theorem 6.2.9, $\mathcal{D}^*_{\mathcal{H}}$ is open and $\mathcal{H}^*$ is piecewise $C^1$ there. Then Lemma 2.5.15 shows that $E^*_{\mathcal{H}}$ is open by and it follows from Definition 2.5.16 that the right-hand sides of (6.20) and (6.21) are piecewise $C^1$ on $E^*_{\mathcal{H}}$. From Assumption 6.2.6, the same holds for (6.22) and (6.23). Then, Conclusion 4 of Lemma 2.5.13 implies differentiability a.e. in $E^*_{\mathcal{H}}$. □

To implement (6.26), the matrix $\tilde{\mathbf{J}}(\mathbf{z}^k)$ is computed by forward automatic differentiation [74]. Automatic differentiation (AD) provides exact derivative evaluations for factorable functions by propagating derivatives through the sequence of factors by repeated application of the addition, multiplication and chain rules of differentiation. As mentioned above, the right-hand sides of (6.20)-(6.23) may involve nonsmooth operations resulting from the rules of interval arithmetic. If these operations are piecewise $C^1$, as we have assumed, then AD can be easily extended to handle them as well. For example, consider the operation

$$c(\mathbf{z}) = \min(a(\mathbf{z}), b(\mathbf{z})), \quad (6.29)$$

which is ubiquitous in interval computations. To propagate derivatives through this operation, we simply let $\partial c/\partial \mathbf{z}$ equal $\partial a/\partial \mathbf{z}$ when $a(\mathbf{z}) \leq b(\mathbf{z})$, and $\partial b/\partial \mathbf{z}$ when $a(\mathbf{z}) > b(\mathbf{z})$. The value assigned to the derivative when $a(\mathbf{z}) = b(\mathbf{z})$ is arbitrary. Extending this approach to other simple piecewise $C^1$ functions, an in house C++ library has been developed that uses operator overloading to both do interval computations and compute such pseudo-derivatives of the resulting bounds. During the differentiation of $\mathbf{L}$ at some point $\mathbf{z}$, the evaluation of any operation at a nondifferentiable point (e.g., when $a(\mathbf{z}) = b(\mathbf{z})$ above) implies that $\mathbf{z}$ is a member of the set of measure zero in Lemma 6.4.3. For all other points, this scheme results in the true Jacobian.

A thorough survey of methods for solving nonsmooth equations is given in [56]. Among these, the *semi-smooth Newton methods*, which are based on the set-valued *generalized Jacobian*, provide the most satisfactory convergence properties, similar to those of a standard Newton iteration. Unfortunately, there is little work on computing an element of the generalized Jacobian. It is known that the directional derivatives of piecewise $C^1$ functions obey a chain rule, from which it follows that the forward mode of AD will give correct directional derivatives [73, 153]. On the other hand, the matrix formed by computing the directional derivatives in all coordinate directions is not necessarily an element of the generalized Jacobian [92]. From this, it follows that $\tilde{\mathbf{J}}$, as computed above, will not necessarily be an element of the generalized Jacobian, and hence (6.26) may not enjoy the properties of semi-smooth Newton methods. However, [92] also presents a modified forward mode AD algorithm that is guaranteed to generate an element of the generalized Jacobian for functions where the nonsmoothness arises from the absolute value function. Further work is underway to extend this method to a much broader class of functions. Thus, the prospects for improving the iteration (6.26) in the future are promising. Finally, we emphasize again that the use of this iteration is still valid. It will either succeed in satisfying (6.8)-(6.12), or it will fail and report an error. Under no circumstances will Algorithm 1 proceed with invalid bounds computed through the use of this iteration.

**Remark 6.4.4.** During the search for a computational means of satisfying (6.11), a significant amount of experimentation was done with methods that, modulo various heuristics, centered around the iteration

$$Z_{y,J} := \mathcal{H}^*(I_J, P, Z_{x,J}, Z_{y,J}, \mathbf{C}_J) + [-\mathbf{1}\gamma, \mathbf{1}\gamma] \tag{6.30}$$

(here, $Z_{x,J}$ is fixed, having been selected earlier by other means). Though this avoids evaluation and inversion of $\tilde{\mathbf{J}}$, we had only limited success. In hindsight, this approach can be viewed as an attempt to solve the system of equations (6.20)-(6.21) using a successive substitution algorithm. Even for the best heuristics found, our results were exactly what one should expect in light of this observation: slow convergence for some systems and disastrous divergence for others. In comparison, the iteration (6.26) is much more robust.

### 6.4.1 Specification of $\mathbf{C}_J$ and $\tilde{\mathbf{z}}_y$

In the Phase 1 implementation below, $\mathcal{H}^*$ is implemented with

$$\tilde{\mathbf{z}}_y(Z_y) \equiv m(Z_y), \quad \forall Z_y \in \mathbb{IR}^{n_y}. \tag{6.31}$$

Note in particular that this guarantees (6.10) for any $Z_{y,J}$ with nonempty interior.

In practice, the choice of preconditioner can have a large impact on the sharpness of the bounds $Z_{x,J}$ and $Z_{y,J}$, and even the ability to satisfy (6.11) and (6.12) at all. A good preconditioner for evaluating $\mathcal{H}^*(I, P, Z_x, Z_y, \mathbf{C})$ is the midpoint inverse

$$\mathbf{C} \equiv \left( m\left( \left[\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\right](I, P, Z_x, Z_y) \right) \right)^{-1}. \tag{6.32}$$

For efficiency reasons, however, it is desirable to compute a preconditioner only once per time step of Algorithm 1. Therefore, the definition

$$\mathbf{C}_J \equiv \left( m\left( \left[\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\right]([t_{J-1}, t_{J-1}], P, X_{J-1}, Y_{J-1}) \right) \right)^{-1} \tag{6.33}$$

is used instead. Thus, $\mathbf{C}_J$ is constant throughout the iteration (6.26). For $J > 1$, $X_{J-1}$ and $Y_{J-1}$ are subsets of $Z_{x,J-1}$ and $Z_{y,J-1}$, and these intervals will have satisfied (6.8)-(6.12) with $j = J - 1$ in the previous time step. It follows that the inverse in (6.33) exists because $\left[\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\right]([t_{J-2}, t_{J-1}], P, Z_{x,J-1}, Z_{y,J-1})$ cannot contain any singular matrices (Corollary 5.4.7). If invertibility fails for $J = 1$, then the inverse of $\frac{\partial \mathbf{g}}{\partial \mathbf{y}}(t_0, \hat{\mathbf{p}}, \mathbf{x}(\hat{\mathbf{p}}), \hat{\mathbf{y}}_0)$ is used instead. If this matrix is singular, then the corresponding solution of (5.1) is not regular and the method does not apply.

### 6.4.2 Phase 1 Algorithm

Algorithm 2 below describes the complete implementation of Step 3 of Algorithm 1. Algorithm 2 terminates with flag = 0 when (6.8)-(6.12) have been satisfied successfully, and returns flag = −1 otherwise. For the examples in §6.7, Algorithm 2 is implemented with $\gamma = 10^{-4}$, H_MAX = 1, H_MIN = $10^{-6}$ and PH1_MAX_ITER = 10.

**Algorithm** 2 (Phase 1)

1. Input: $[t_0, t_f]$, $P$, $\gamma$, $t_{J-1}$, $X_{J-1}$, $Y_{J-1}$, $\Delta t_{J-1}$.

2. Assign $\Delta t_J := \min(2\Delta t_{J-1}, \text{H\_MAX}, t_f - t_{J-1} + \text{H\_MIN})$ and $t_J := t_{J-1} + \Delta t_J$.

3. Assign $\mathbf{z}_x^L := \mathbf{x}_{J-1}^L - \mathbf{1}\gamma$, $\mathbf{z}_x^U := \mathbf{x}_{J-1}^U + \mathbf{1}\gamma$, $\mathbf{z}_y^L := \mathbf{y}_{J-1}^L - \mathbf{1}\gamma$, $\mathbf{z}_y^U := \mathbf{y}_{J-1}^U + \mathbf{1}\gamma$.

4. With initial guesses from 3, apply the iteration (6.26) described above.

   (a) If PH1_MAX_ITER iterations are taken without success, go to 6.

   (b) If any iterate violates (6.17), go to 6.

   (c) If $(\mathbf{z}_x^L, \mathbf{z}_x^U, \mathbf{z}_y^L, \mathbf{z}_y^U)$ is found such that the right-hand sides of (6.20)-(6.23) are componentwise less than $\gamma$, set $Z_{x,J} := [\mathbf{z}_x^L, \mathbf{z}_x^U]$ and $Z_{y,J} := [\mathbf{z}_y^L, \mathbf{z}_y^U]$ and go to 5.

5. If $Y_{J-1} \subset Z_{y,J}$, terminate with flag = 0. Otherwise, go to 6.

6. Assign $\Delta t_J := \Delta t_J/2$ and $t_J := t_{J-1} + \Delta t_J$. If $\Delta t_J \geq$ H_MIN go to 3. Otherwise, terminate with flag = −1.

Suppose that Algorithm 2 returns 0. By Step 4 and Lemma 6.4.2, (6.8), (6.11) and (6.12) are satisfied. Since (6.11) implies that $Z_{y,J}$ has nonempty interior, (6.10) is guaranteed by the choice of $\tilde{\mathbf{z}}_y$ in §6.4.1. Finally, (6.9) is verified by Step 5. Then, Phase 1 is complete. The only way Algorithm 2 can fail is if $\Delta t_J$ is reduced below H_MIN by repeated failure in Step 4 or 5. To avoid many such failures, $\Delta t_J$ is bounded by $2\Delta t_{J-1}$.

In practice, Step 4 succeeds reliably when the intervals $I_J$ and $P$ are narrow, and becomes less reliable as they are widened. This is natural given that (6.6) follows from (6.8)-(6.12). When $I_J$ and $P$ are narrow, (6.8)-(6.12) can potentially be satisfied by narrower intervals $Z_{x,J}$ and $Z_{y,J}$. Working with narrower intervals in turn reduces the overestimation incurred through interval computations, and reduces the likelihood of violating (6.8). Both of these factors make Step 4 more likely to succeed.

When Step 4 fails, the recourse is to half $\Delta t_J$ and try again. On the other hand, Algorithm 2 does not resort to partitioning $P$. Though algorithms for bisecting $P$ and propagating bounds valid on each partition element separately are easily conceivable, computational efficiency will be lost if many partitions are required, so this strategy is avoided. With $P$ fixed, one can create pathological problems for which it is impossible to satisfy (6.11), and therefore there is no theoretical guarantee that Step 4 will succeed. This happens, for example, if the algebraic equations permit multiple solution branches on $[t_{J-1}, t_{J-1}] \times P \times X_{J-1}$ and it is geometrically impossible to enclose one uniquely by an interval (see Corollary 5.4.7).

Though the condition (6.9) is checked in Step 5 of Algorithm 2, no special attempt is made to guarantee it. The condition (6.9) is merely a provision for the case where (5.1) permits multiple regular solutions. Its purpose is to ensure that the interval $Z_{y,J}$ computed in Step 4 encloses the solution of (5.1) that is consistent with the input $\hat{\mathbf{y}}_0$ in Algorithm 1, rather than jumping to some other solution (see the proof of Corollary 6.3.2). Since the initial guesses specified in Step 3 are in the vicinity of the solution of interest, (6.9) is likely to hold whenever Step 4 succeeds.

### 6.4.3 Phase 1 Refinement

Before moving on to Phase 2 of Algorithm 1, $Z_{x,J}$ and $Z_{y,J}$ may be refined by iteratively assigning

$$Z_{x,J} := (X_{J-1} + [0, t_J - t_{J-1}] \, [\mathbf{f}] \, (I_J, P, Z_{x,J}, Z_{y,J})) \cap Z_{x,J}, \tag{6.34}$$

$$Z_{y,J} := \mathcal{H}(I_J, P, Z_{x,J}, Z_{y,J}, \mathbf{C}_J). \tag{6.35}$$

By (6.7), it is clear that

$$\mathbf{x}(t, \mathbf{p}) = \mathbf{x}(t_{J-1}, \mathbf{p}) + \int_{t_{J-1}}^{t} \mathbf{f}(s, \mathbf{p}, \mathbf{x}(s, \mathbf{p}), \mathbf{y}(s, \mathbf{p})) ds, \tag{6.36}$$

$$\in X_{J-1} + [0, t - t_{J-1}] \, [\mathbf{f}] \, (I_J, P, Z_{x,J}, Z_{y,J}), \tag{6.37}$$

for all $(t, \mathbf{p}) \in I_J \times P$. Therefore, (6.6) remains valid after the assignment (6.34). By Conclusion 1 of Corollary 5.4.7, the same is true of the assignment (6.35). Note that these refinements are distinct from the refinements $X_J$ and $Y_J$ detailed in §6.5 in that (6.6) remains true. That is, the refined intervals still provide bounds on all of $I_J \times P$, rather than only at $t_J$, as in (6.7). For the examples in §6.7, (6.34) and (6.35) are applied with a maximum of 50 iterations, terminating early if the absolute or relative change between each bound in successive iterates is less that $10^{-8}$.

## 6.5 Computing Refined Enclosures Using Differential Inequalities (Phase 2)

In this section, we consider the implementation of Step 6 in a single time step $J$ of Algorithm 1. It is assumed that a solution $(\mathbf{x}, \mathbf{y})$ of (5.1) exists on $[t_0, t_J] \times P$, and that $Y_0$ and ($I_j$, $Z_{x,j}$, $Z_{y,j}$, $Z'_{y,j}$, $\mathbf{C}_j$, $X_j$, $Y_j$) are available and satisfy (6.6) and (6.8)-(6.12) for all $j \in \{1, \ldots, J\}$ and (6.7) for all $j \in \{0, \ldots, J-1\}$. The present task is to compute refined intervals $X_J \subset Z_{x,J}$ and $Y_J \subset Z'_{y,J}$ satisfying (6.7) with $j = J$.

By the assumption that (6.8)-(6.12) hold with $j = J$, Corollary 5.4.7 guarantees that $\exists \mathbf{H} \in C^1(I_J \times P \times Z_{x,J}, Z'_{y,J})$ such that, for every $(t, \mathbf{p}, \mathbf{z}_x) \in I_J \times P \times Z_{x,J}$, $\mathbf{z}_y = \mathbf{H}(t, \mathbf{p}, \mathbf{z}_x)$ is the unique element of $Z_{y,J}$ satisfying $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$. Therefore, we aim to apply Theorem 5.5.3 to derive time-varying bounds on $(\mathbf{x}, \mathbf{y})$ over $I_J$.

Choose any $K \in \mathbb{N}$ and, for every $i \in \{1, \ldots, n_x\}$, define

$$\begin{aligned}
\phi_i^L, \phi_i^U &: \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{IR} \times \mathbb{IR}^{n_p} \times \mathbb{IR}^{n_x} \times \mathbb{IR}^{n_y} \times \mathbb{IR}^{n_y \times n_y}, && (6.38)\\
\mathcal{Y}_i^L, \mathcal{Y}_i^U &: \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{IR}^{n_y}, && (6.39)\\
\psi_i^L, \psi_i^U &: \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{IR} \times \mathbb{IR}^{n_p} \times \mathbb{IR}^{n_x} \times \mathbb{IR}^{n_y}, && (6.40)
\end{aligned}$$

by

$$\begin{aligned}
\phi_i^L(t, \mathbf{v}, \mathbf{w}) &\equiv \left(I_J \tilde{\cap} [t,t], P, \mathcal{B}_i^L(Z_{x,J} \tilde{\cap} \square(\mathbf{v}, \mathbf{w})), Z'_{y,J}, \mathbf{C}_J\right), && (6.41)\\
\phi_i^U(t, \mathbf{v}, \mathbf{w}) &\equiv \left(I_J \tilde{\cap} [t,t], P, \mathcal{B}_i^U(Z_{x,J} \tilde{\cap} \square(\mathbf{v}, \mathbf{w})), Z'_{y,J}, \mathbf{C}_J\right), && (6.42)\\
\mathcal{Y}_i^L(t, \mathbf{v}, \mathbf{w}) &\equiv \mathcal{H}^{+,K}(\phi_i^L(t, \mathbf{v}, \mathbf{w})), && (6.43)\\
\mathcal{Y}_i^U(t, \mathbf{v}, \mathbf{w}) &\equiv \mathcal{H}^{+,K}(\phi_i^U(t, \mathbf{v}, \mathbf{w})), && (6.44)\\
\psi_i^L(t, \mathbf{v}, \mathbf{w}) &\equiv \left(I_J \tilde{\cap} [t,t], P, \mathcal{B}_i^L(Z_{x,J} \tilde{\cap} \square(\mathbf{v}, \mathbf{w})), \mathcal{Y}_i^L(t, \mathbf{v}, \mathbf{w})\right), && (6.45)\\
\psi_i^U(t, \mathbf{v}, \mathbf{w}) &\equiv \left(I_J \tilde{\cap} [t,t], P, \mathcal{B}_i^U(Z_{x,J} \tilde{\cap} \square(\mathbf{v}, \mathbf{w})), \mathcal{Y}_i^U(t, \mathbf{v}, \mathbf{w}),\right). && (6.46)
\end{aligned}$$

Now, consider the initial value problem in ODEs

$$\begin{aligned}
\dot{v}_i(t) &= [f_i]^L\left(\psi_i^L(t, \mathbf{v}(t), \mathbf{w}(t))\right), && (6.47)\\
\dot{w}_i(t) &= [f_i]^U\left(\psi_i^U(t, \mathbf{v}(t), \mathbf{w}(t))\right), && (6.48)
\end{aligned}$$

for all $i = 1, \ldots, n_x$, with initial conditions

$$[\mathbf{v}(t_{J-1}), \mathbf{w}(t_{J-1})] = X_{J-1}. \qquad (6.49)$$

The following results show that these ODEs are well-defined and have a unique solution describing the desired bounds. It is assumed thoughout that Assumption 6.2.6

holds and $\tilde{\mathbf{z}}_y$ is the midpoint, as in §6.4.1.

**Corollary 6.5.1.** *When viewed as functions of* $(t, \mathbf{v}, \mathbf{w})$*, the right-hand sides of* (6.47) *and* (6.48) *are defined and piecewise* $C^1$ *on* $\mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$*. Furthermore,*

$$\mathcal{Y}_i^L(t, \mathbf{v}, \mathbf{w}) = \mathcal{H}^K\left(\phi_i^L(t, \mathbf{v}, \mathbf{w})\right) \quad \textit{and} \quad \mathcal{Y}_i^U(t, \mathbf{v}, \mathbf{w}) = \mathcal{H}^K\left(\phi_i^U(t, \mathbf{v}, \mathbf{w})\right), \tag{6.50}$$

*for all* $(t, \mathbf{v}, \mathbf{w}) \in \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ *and every* $i = 1, \ldots, n_x$*.*

*Proof.* Choose any $i \in \{1, \ldots, n_x\}$ and any $(t, \mathbf{v}, \mathbf{w}) \in \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$. By Conclusion 3 of Lemma 2.5.23, $\phi_i^L(t, \mathbf{v}, \mathbf{w}) \subset (I_J, P, Z_{x,J}, Z_{y,J}, \mathbf{C}_J)$. Using (6.8), (6.10) and (6.11), Conclusion 7 of Lemma 6.2.5 implies that $\phi_i^L(t, \mathbf{v}, \mathbf{w}) \in \mathcal{D}_{\mathcal{H}}^K$. Then, $\mathcal{Y}_i^L(t, \mathbf{v}, \mathbf{w})$ is well-defined and Conclusion 4 of Lemma 6.2.5 shows (6.50) (an analogous argument holds for $\mathcal{Y}_i^U$).

Now (6.50) implies that $\mathcal{Y}_i^L(t, \mathbf{v}, \mathbf{w}) \subset Z'_{y,J}$. It follows that $\psi_i^L(t, \mathbf{v}, \mathbf{w})$ is in $\mathbb{I}D_t \times \mathbb{I}D_p \times \mathbb{I}D_x \times \mathbb{I}D_y$. Then, the right-hand side of (6.47) is defined on $\mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$.

By Lemmas 2.5.19 and 2.5.25 and Definition 3.3.1, it is clear that $\phi_i^L$ is piecewise $C^1$ on $\mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$, which is open. Theorem 6.2.9 shows that $\mathcal{H}^{+,K}$, and hence $\mathcal{Y}_i^L = \mathcal{H}^{+,K} \circ \phi_i^L$, is also piecewise $C^1$ on $\mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$. It follows that $\psi_i^L$ is piecewise $C^1$ on $\mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$. Finally, Assumption 6.2.6 implies that $[f_i]^L \circ \psi_i^L$ is piecewise $C^1$ on $\mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$, which is the desired result (an analogous argument holds for $[f_i]^U \circ \psi_i^U$). □

**Lemma 6.5.2.** *There exist* $\mathbf{v}, \mathbf{w} \in C^1(I_J, \mathbb{R}^{n_x})$ *satisfying the ODEs* (6.47)-(6.49)*. Moreover, this solution is unique and satisfies* $\mathbf{v}(t) \leq \mathbf{w}(t)$ *and* $[\mathbf{v}(t), \mathbf{w}(t)] \cap Z_{x,J} \neq \emptyset$*,* $\forall t \in I_J$*.*

*Proof.* Consider the ODEs

$$\dot{s}(t) = 1, \tag{6.51}$$

$$\dot{v}_i(t) = [f_i]^L\left(\psi_i^L(s(t), \mathbf{v}(t), \mathbf{w}(t))\right), \tag{6.52}$$

$$\dot{w}_i(t) = [f_i]^U\left(\psi_i^U(s(t), \mathbf{v}(t), \mathbf{w}(t))\right), \tag{6.53}$$

with initial conditions (6.49) and $s(t_0) = t_0$. This system simply describes the bounding ODEs (6.47) and (6.48) in autonomous form.

By Corollary 6.5.1 and Conclusion 3 of Lemma 2.5.13, the right-hand sides of (6.51)-(6.53) are locally Lipschitz continuous on $\mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$. Moreover, $\psi^L$ and $\psi^U$ are easily seen to map into subsets of $(I_J, P, Z_{x,J}, Z_{y,J})$. Thus, the right-hand sides of (6.51)-(6.53) are also bounded on $\mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ by

$$\max\left(1, \left|[f_i]^L(I_J, P, Z_{x,J}, Z_{y,J})\right|, \left|[f_i]^U(I_J, P, Z_{x,J}, Z_{y,J})\right|\right). \tag{6.54}$$

For any $(\hat{s}, \hat{\mathbf{v}}, \hat{\mathbf{w}}) \in \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ and any $i \in \{1, \ldots, n_x\}$, the definitions of $\Box$ and $\tilde{\cap}$ guarantee that

$$\begin{aligned}
\hat{v}_i = \hat{w}_i &\implies (Z_{x,J})_i \tilde{\cap} \Box(\hat{v}_i, \hat{w}_i) \text{ is a singleton}, && (6.55)\\
&\implies \mathcal{B}_i^L(Z_{x,J} \tilde{\cap} \Box(\hat{\mathbf{v}}, \hat{\mathbf{w}})) = \mathcal{B}_i^U(Z_{x,J} \tilde{\cap} \Box(\hat{\mathbf{v}}, \hat{\mathbf{w}})), && (6.56)\\
&\implies \mathcal{Y}_i^L(\hat{s}, \hat{\mathbf{v}}, \hat{\mathbf{w}}) = \mathcal{Y}_i^U(\hat{s}, \hat{\mathbf{v}}, \hat{\mathbf{w}}), && (6.57)\\
&\implies [f_i]^L(\psi_i^L(\hat{s}, \hat{\mathbf{v}}, \hat{\mathbf{w}})) \leq [f_i]^U(\psi_i^U(\hat{s}, \hat{\mathbf{v}}, \hat{\mathbf{w}})). && (6.58)
\end{aligned}$$

This implies that $K \equiv \{(\hat{s}, \hat{\mathbf{v}}, \hat{\mathbf{w}}) \in \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} : \hat{\mathbf{v}} \leq \hat{\mathbf{w}}\}$ is a viability domain for the ODEs (6.51)-(6.53) (Definition 1.1.5 in [13]). Combining this with continuity and boundedness of the right-hand sides, Nagumo's Theorem implies that there exist $s \in C^1(I_J, \mathbb{R}^n)$ and $\mathbf{v}, \mathbf{w} \in C^1(I_J, \mathbb{R}^{n_x})$ satisfying (6.51)-(6.53) and satisfying $(s(t), \mathbf{v}(t), \mathbf{w}(t)) \in K$, and hence $\mathbf{v}(t) \leq \mathbf{w}(t)$, $\forall t \in I_J$ (see Theorem 1.2.4 in [13]). Clearly, this $\mathbf{v}, \mathbf{w}$ also satisfies (6.47)-(6.49). Due to the local Lipschitz continuity of the ODE right-hand side functions on $\mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$, uniqueness follows by a standard application of Gronwall's inequality.

Let $[x_{J-1,i}^L, x_{J-1,i}^U]$ and $[{z'}_{y,J,i}^L, {z'}_{y,J,i}^U]$ denote the $i^{\text{th}}$ components of $X_{J-1}$ and $Z'_{y,J}$, respectively. By (6.12) and the integral form of (6.47),

$$v_i(t) = v_i(t_{J-1}) + \int_{t_{J-1}}^{t} [f_i]^L (\psi_i^L(s, \mathbf{v}(s), \mathbf{w}(s)))ds, \tag{6.59}$$

$$\geq x_{J-1,i}^L + \int_{t_{J-1}}^{t} [f_i]^L (I_J, P, Z_{x,J}, Z'_{y,J}), \tag{6.60}$$

$$\geq x_{J-1,i}^L + [0, t_J - t_{J-1}] [f_i]^L (I_J, P, Z_{x,J}, Z'_{y,J}) \geq z_{x,J,i}^L, \quad \forall t \in I_J. \tag{6.61}$$

Using an analogous argument for $w_i$, it follows that $[\mathbf{v}(t), \mathbf{w}(t)] \subset Z_{x,J}$, $\forall t \in I_J$. □

**Corollary 6.5.3.** *Let* $\mathbf{v}, \mathbf{w} \in C^1(I_J, \mathbb{R}^{n_x})$ *be the unique solutions of* (6.47)-(6.49). *Then*

$$\mathbf{x}(t, \mathbf{p}) \in [\mathbf{v}(t), \mathbf{w}(t)], \tag{6.62}$$

$$\mathbf{y}(t, \mathbf{p}) \in \mathcal{Y}(t, \mathbf{v}(t), \mathbf{w}(t)) \equiv \mathcal{H}^q \left([t, t], P, Z_{x,J} \cap [\mathbf{v}(t), \mathbf{w}(t)], Z_{y,J}\right), \tag{6.63}$$

*for all* $(t, \mathbf{p}) \in I_J \times P$ *and any* $q \in \mathbb{N}$.

*Proof.* To show (6.62), it suffices to establish the hypotheses of Theorem 5.5.3 with $(I, Z_x, Z'_y) = (I_J, Z_{x,J}, Z'_{y,J})$, $t_f = t_J$, $t_0 = t_{J-1}$ and $\mathbf{x}_0 = \mathbf{x}_{J-1} \equiv \mathbf{x}(t_{J-1}, \cdot)$. By (6.8)-(6.12) and Corollary 5.4.7, there exists $\mathbf{H} \in C^1(I_J \times P \times Z_{x,J}, Z'_{y,J})$ such that, for every $(t, \mathbf{p}, \mathbf{z}_x) \in I_J \times P \times Z_{x,J}$, $\mathbf{z}_y = \mathbf{H}(t, \mathbf{p}, \mathbf{z}_x)$ is the unique element of $Z_{y,J}$ satisfying $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$. Then, it only remains to satisfy the hypotheses (EX), (IC) and (RHS). (EX) holds by Lemma 6.5.2. By (6.49) and (6.7) with $j = J-1$, (IC) is clearly satisfied. Choose any $t \in I_J$. If there exists $(\mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) \in P \times Z_{x,J} \times Z'_{y,J}$ such that $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$ and $\mathbf{z}_x \in \mathcal{B}_i^L(Z_{x,J} \cap [\mathbf{v}(t), \mathbf{w}(t)])$, then (6.50) and Conclusion 1 of Corollary 5.4.7 ensure that $\mathbf{z}_y \in \mathcal{Y}_i^L(t, \mathbf{v}(t), \mathbf{w}(t))$. It follows that

$$f_i(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) \in [f_i]([t, t], P, \mathcal{B}_i^L(Z_{x,J} \cap [\mathbf{v}(t), \mathbf{w}(t)]), \mathcal{Y}_i^L(t, \mathbf{v}(t), \mathbf{w}(t))), \tag{6.64}$$

$$= [f_i](\psi_i^L(t, \mathbf{v}(t), \mathbf{w}(t))), \tag{6.65}$$

and hence (6.47) ensures that (RHS).1 is satisfied. Proof of (RHS).2 is analogous. Then, (6.62) holds, and (6.63) follows from Conclusion 1 of Corollary 5.4.7. □

According to Corollary 6.5.3, Step 6 of Algorithm 1 can be accomplished by solving (6.47)-(6.49) on $I_J$ and assigning $X_J := [\mathbf{v}(t_J), \mathbf{w}(t_J)]$ and $Y_J := \mathcal{Y}(t_J, \mathbf{v}(t_J), \mathbf{w}(t_J))$. Provided that numerical error is not a crucial concern, these ODEs can be solved numerically using any state of the art code. In the examples in §6.7, we use `CVODE` [44] with absolute and relative tolerances of $10^{-5}$. The evaluation of $\mathcal{Y}_i^L$ and $\mathcal{Y}_i^U$ for each $i$ can make evaluating the right-hand sides of (6.47)-(6.48) costly, so $K$ should be small (see §6.6.1). On the other hand, $q$ can be fairly large, because $\mathcal{Y}$ is evaluated after numerical integration is complete rather than within the right-hand sides of (6.47) and (6.48). Moreover, $\mathcal{Y}$ need only be evaluated at select points of interest in $I_J$, since only the value at $t_J$, which defines $Y_J$, will effect the next time step of Algorithm 1. In §6.7, we choose $K = 5$ and evaluate $\mathcal{Y}$ with $q = 50$ at all points shown in the plots there.

## 6.6 A Single-Phase Method

In this section, a single-phase method is presented which essentially combines the two phases of the previous approach. In short, time-varying bounds for both the differential and the algebraic state variables will be computed by satisfying the hypotheses of Theorem 5.5.6. As before, let $I = [t_0, t_f] \subset D_t$, $P \subset D_p$ and $X_0 \subset D_x$ be intervals and suppose that $\mathbf{x}_0(P) \subset X_0$.

For every $i \in \{1, \ldots, n_x\}$, let

$$\eta : \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} \to \mathbb{IR} \times \mathbb{IR}^{n_p} \times \mathbb{IR}^{n_x} \times \mathbb{IR}^{n_y} \tag{6.66}$$

$$\mathbf{C} : E_{\text{inv}} \to \mathbb{IR}^{n_y \times n_y}, \tag{6.67}$$

$$\phi, \phi_i^L, \phi_i^U : E_{\text{inv}} \to \mathbb{IR} \times \mathbb{IR}^{n_p} \times \mathbb{IR}^{n_x} \times \mathbb{IR}^{n_y} \times \mathbb{IR}^{n_y \times n_y}, \tag{6.68}$$

$$\mathcal{Y}_i^L, \mathcal{Y}_i^U : E_{\mathcal{H}}^* \to \mathbb{IR}^{n_y}, \tag{6.69}$$

$$\psi_i^L, \psi_i^U : E_{\mathcal{H}}^* \to \mathbb{IR} \times \mathbb{IR}^{n_p} \times \mathbb{IR}^{n_x} \times \mathbb{IR}^{n_y}, \tag{6.70}$$

where

$$E_{\mathbb{I}D} \equiv \big\{(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U) \in \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} : \tag{6.71}$$

$$\eta(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U) \in \mathbb{I}D_t \times \mathbb{I}D_p \times \mathbb{I}D_x \times \mathbb{I}D_y\big\}\,, \tag{6.72}$$

$$\mathcal{D}_{\text{inv}} \equiv \big\{Q \in \mathbb{IR}^{n_y \times n_y} : \det\left(m\left(Q\right)\right) \neq 0\big\}\,, \tag{6.73}$$

$$E_{\text{inv}} \equiv \left\{(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U) \in E_{\mathbb{I}D} : \left[\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\right](\eta(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U)) \in \mathcal{D}_{\text{inv}}\right\}, \tag{6.74}$$

$$E_{\mathcal{H}}^* \equiv \big\{(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U) \in E_{\text{inv}} : \phi(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U) \in \mathcal{D}_{\mathcal{H}}^*\big\}\,. \tag{6.75}$$

Choosing any $K \in \mathbb{N}$, define the functions in (6.66)-(6.70) by

$$\eta(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U) \equiv \big(I\tilde{\cap}[t,t], P, \square(\mathbf{v},\mathbf{w}), \square(\mathbf{z}_y^L,\mathbf{z}_y^U)\big)\,, \tag{6.76}$$

$$\mathbf{C}(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U) \equiv m\left(\left[\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\right](\eta(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U))\right)^{-1}, \tag{6.77}$$

$$\phi(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U) \equiv \big(I\tilde{\cap}[t,t], P, \square(\mathbf{v},\mathbf{w}), \square(\mathbf{z}_y^L,\mathbf{z}_y^U), \mathbf{C}(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U)\big)\,, \tag{6.78}$$

$$\phi_i^L(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U) \equiv \big(I\tilde{\cap}[t,t], P, \mathcal{B}_i^L(\square(\mathbf{v},\mathbf{w})), \square(\mathbf{z}_y^L,\mathbf{z}_y^U), \mathbf{C}(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U)\big)\,, \tag{6.79}$$

$$\phi_i^U(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U) \equiv \big(I\tilde{\cap}[t,t], P, \mathcal{B}_i^U(\square(\mathbf{v},\mathbf{w})), \square(\mathbf{z}_y^L,\mathbf{z}_y^U), \mathbf{C}(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U)\big)\,, \tag{6.80}$$

$$\mathcal{Y}_i^L(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U) \equiv \mathcal{H}^{+,K}(\phi_i^L(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U)), \tag{6.81}$$

$$\mathcal{Y}_i^U(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U) \equiv \mathcal{H}^{+,K}(\phi_i^U(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U)), \tag{6.82}$$

$$\psi_i^L(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U) \equiv \big(I\tilde{\cap}[t,t], P, \mathcal{B}_i^L(\square(\mathbf{v},\mathbf{w})), \mathcal{Y}_i^L(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U)\big)\,, \tag{6.83}$$

$$\psi_i^U(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U) \equiv \big(I\tilde{\cap}[t,t], P, \mathcal{B}_i^U(\square(\mathbf{v},\mathbf{w})), \mathcal{Y}_i^U(t,\mathbf{v},\mathbf{w},\mathbf{z}_y^L,\mathbf{z}_y^U)\big)\,. \tag{6.84}$$

For any continuous and pointwise positive $\gamma : I \to \mathbb{R}$, consider the initial value problem in DAEs

$$\dot{v}_i(t) = [f_i]^L\left(\psi_i^L(t,\mathbf{v}(t),\mathbf{w}(t),\mathbf{z}_y^L(t),\mathbf{z}_y^U(t))\right), \tag{6.85}$$

$$\dot{w}_i(t) = [f_i]^U\left(\psi_i^U(t,\mathbf{v}(t),\mathbf{w}(t),\mathbf{z}_y^L(t),\mathbf{z}_y^U(t))\right), \tag{6.86}$$

$$\mathbf{0} = \mathbf{z}_y^L(t) - \mathcal{H}^{*,L}(\phi(t,\mathbf{v}(t),\mathbf{w}(t),\mathbf{z}_y^L(t),\mathbf{z}_y^U(t))) + \mathbf{1}\gamma(t), \tag{6.87}$$

$$\mathbf{0} = -\mathbf{z}_y^U(t) + \mathcal{H}^{*,U}(\phi(t,\mathbf{v}(t),\mathbf{w}(t),\mathbf{z}_y^L(t),\mathbf{z}_y^U(t))) + \mathbf{1}\gamma(t), \tag{6.88}$$

for all $i = 1, \dots, n_x$, with initial conditions

$$[\mathbf{v}(t_0), \mathbf{w}(t_0)] = X_0. \tag{6.89}$$

In the following results, it will be shown that the solutions of these DAEs describe the desired bounds. It is assumed thoughout that Assumption 6.2.6 holds and $\tilde{\mathbf{z}}_y$ is the midpoint, as in §6.4.1.

**Corollary 6.6.1.** *$E^*_{\mathcal{H}}$ is open and, when viewed as functions of $(t, \mathbf{v}, \mathbf{w}, \mathbf{z}^L_y, \mathbf{z}^U_y)$, the right-hand sides of* (6.85)-(6.88) *are defined and piecewise $C^1$ on $E^*_{\mathcal{H}}$.*

*Proof.* By Lemmas 2.5.19 and 2.5.25, $\eta$ is piecewise $C^1$ on $\mathbb{I}D_t \times \mathbb{I}D_p \times \mathbb{I}D_x \times \mathbb{I}D_y$. Since this set is open, $E_{\mathbb{I}D}$ is open by Lemma 2.5.15. Moreover, the set of nonsingular matrices is open. Then, since $m(\cdot)$ is clearly a continuous function from $\mathbb{IR}^{n_y \times n_y}$ to $\mathbb{R}^{n_y \times n_y}$, $\mathcal{D}_{\text{inv}}$ is the inverse image of an open set under a continuous mapping, and is hence open. By Assumption 6.2.6, $\left[\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\right] \circ \eta$ is piecewise $C^1$ on $E_{\mathbb{I}D}$. Then, another application of Lemma 2.5.15 now shows that $E_{\text{inv}}$ is open. The fact that $\mathbf{C}$ is piecewise $C^1$ on $E_{\text{inv}}$ now follows from the definition of $m(\cdot)$ and the fact that the inverse of a matrix is a differentiable function of its elements. Combining this with Lemmas 2.5.19 and 2.5.25 shows that $\phi$, $\phi^L_i$ and $\phi^U_i$ are piecewise $C^1$ on $E_{\text{inv}}$, so that openness of $\mathcal{D}^*_{\mathcal{H}}$ and a final application of Lemma 2.5.15 show that $E^*_{\mathcal{H}}$ is open.

Choose any $i \in \{1, \dots, n_x\}$. By the definition of $E^*_{\mathcal{H}}$ and Conclusion 1 of Lemma 6.2.5,

$$\phi^L_i(t, \mathbf{v}, \mathbf{w}, \mathbf{z}^L_y, \mathbf{z}^U_y) \in \mathcal{D}^*_{\mathcal{H}}, \quad \forall (t, \mathbf{v}, \mathbf{w}, \mathbf{z}^L_y, \mathbf{z}^U_y) \in E^*_{\mathcal{H}}. \tag{6.90}$$

Theorem 6.2.9 shows that $\mathcal{H}^*$ and $\mathcal{H}^{+,K}$ are piecewise $C^1$ on $\mathcal{D}^*_{\mathcal{H}}$, and hence $\mathcal{H}^* \circ \phi$ and $\mathcal{Y}^L_i = \mathcal{H}^{+,K} \circ \phi^L_i$ are piecewise $C^1$ on $E^*_{\mathcal{H}}$. It follows that the right-hand side of (6.87) and $\psi^L_i$ are piecewise $C^1$ on $E^*_{\mathcal{H}}$. For any $(t, \mathbf{v}, \mathbf{w}, \mathbf{z}^L_y, \mathbf{z}^U_y) \in E^*_{\mathcal{H}}$, the definition of $\mathcal{H}^{+,K}$ implies that $\mathcal{Y}^L_i(t, \mathbf{v}, \mathbf{w}, \mathbf{z}^L_y, \mathbf{z}^U_y) \subset \square(\mathbf{z}^L_y, \mathbf{z}^U_y)$, and hence

$$\psi^L_i(t, \mathbf{v}, \mathbf{w}, \mathbf{z}^L_y, \mathbf{z}^U_y) \subset \eta(t, \mathbf{v}, \mathbf{w}, \mathbf{z}^L_y, \mathbf{z}^U_y) \subset D_t \times D_p \times D_x \times D_y. \tag{6.91}$$

Then, Assumption 6.2.6 implies that $[f_i]^L \circ \psi_i^L$ is piecewise $C^1$ on $E^*_{\mathcal{H}}$. Analogous arguments hold for the right-hand sides of (6.88) and (6.86). $\square$

In contrast to the analysis of the Phase 2 bounding ODEs in §6.5, existence and uniqueness of a solution of (6.85)-(6.89) does not follow from standard results because the participating functions are only piecewise $C^1$, rather than $C^1$. However, such a result seems quite plausible. From a variant of the implicit function theorem in [153], one can write an invertibility condition for the right-hand sides of (6.87)-(6.88) which guarantees the existence of a piecewise $C^1$ implicit function locally around a consistent initial condition. By Conclusion 3 of Lemma 2.5.13, this would imply that $\mathbf{v}$ and $\mathbf{w}$ are, locally, described by ODEs with locally Lipschitz continuous right-hand sides. Combining this with standard results for Lipschitz ODEs then implies that there exists a solution in a neighborhood of $t_0$ with $\mathbf{v}$ and $\mathbf{w}$ continuously differentiable and $\mathbf{z}_y^L$ and $\mathbf{z}_y^U$ piecewise $C^1$. We do not pursue this development formally here. Instead, we will assume that such a solution exists on an open set $I_0$ containing $I$ and demonstrate that it must describe the desired bounds.

**Lemma 6.6.2.** *Let* $(\mathbf{v}, \mathbf{w}, \mathbf{z}_y^L, \mathbf{z}_y^U)$ *be a solution of* (6.85)-(6.89). *Then* $\mathbf{v}(t) \le \mathbf{w}(t)$ *and* $\mathbf{z}_y^L(t) < \mathbf{z}_y^U(t)$ *for all* $t \in I$.

*Proof.* Arguing as in Lemma 6.4.2, it is clear from (6.87) and (6.88) that any solution must satisfy $\mathbf{z}_y^L(t) < \mathbf{z}_y^U(t)$ for all $t \in I$.

For a contradiction, suppose that $\{t \in I : v_i(t) > w_i(t) \text{ for at least one } i\}$ is nonempty and let $t_1 < t_f$ denote its infimum. Because $t_1$ is a lower bound, $\mathbf{v}(t) \le \mathbf{w}(t)$, $\forall t \in [t_0, t_1]$. Because $t_1$ is the greatest lower bound, it follows that $v_i(t) > w_i(t)$ for at least one $i$ for $t$ arbitrarily close to the right of $t_1$.

Now, treating $\mathbf{z}_y^L$ and $\mathbf{z}_y^U$ as known functions, consider the ODEs

$$\dot{s}(t) = 1, \tag{6.92}$$

$$\dot{v}_i^*(t) = [f_i]^L\left(\psi_i^L(s(t), \mathbf{v}^*(t), \mathbf{w}^*(t), \mathbf{z}_y^L(s(t)), \mathbf{z}_y^U(s(t)))\right), \tag{6.93}$$

$$\dot{w}_i^*(t) = [f_i]^U\left(\psi_i^U(s(t), \mathbf{v}^*(t), \mathbf{w}^*(t), \mathbf{z}_y^L(s(t)), \mathbf{z}_y^U(s(t)))\right), \tag{6.94}$$

for all $i = 1, \ldots, n_x$. Corollary 6.6.1 implies that the right-hand sides of these ODEs are piecewise $C^1$, and hence locally Lipschitz continuous, on the set

$$Q \equiv \left\{(\hat{s}, \hat{\mathbf{v}}, \hat{\mathbf{w}}) \in I_0 \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} : (\hat{s}, \hat{\mathbf{v}}, \hat{\mathbf{w}}, \mathbf{z}^L_y(\hat{s}), \mathbf{z}^U_y(\hat{s})) \in E^*_{\mathcal{H}}\right\}. \tag{6.95}$$

We refer to these ODEs as the reduced ODEs and consider them with initial contitions $(s(t_1), \mathbf{v}^*(t_1), \mathbf{w}^*(t_1)) = (t_1, \mathbf{v}(t_1), \mathbf{w}(t_1))$. Clearly, for any solution $(s, \mathbf{v}^*, \mathbf{w}^*)$ of the reduced ODEs on $[t_t, t_1 + \delta]$, $(s, \mathbf{v}, \mathbf{w})$ is also a solution.

For any $(\hat{s}, \hat{\mathbf{v}}, \hat{\mathbf{w}}) \in Q$ and any $i \in \{1, \ldots, n_x\}$,

$$\begin{aligned}
\hat{v}_i = \hat{w}_i &\implies \mathcal{B}^L_i(\square(\hat{\mathbf{v}}, \hat{\mathbf{w}})) = \mathcal{B}^U_i(\square(\hat{\mathbf{v}}, \hat{\mathbf{w}})), && (6.96)\\
&\implies \mathcal{Y}^L_i(\hat{s}, \hat{\mathbf{v}}, \hat{\mathbf{w}}, \mathbf{z}^L_y(\hat{s}), \mathbf{z}^U_y(\hat{s})) = \mathcal{Y}^U_i(\hat{s}, \hat{\mathbf{v}}, \hat{\mathbf{w}}, \mathbf{z}^L_y(\hat{s}), \mathbf{z}^U_y(\hat{s})), && (6.97)\\
&\implies [f_i]^L\left(\psi^L_i(\hat{s}, \hat{\mathbf{v}}, \hat{\mathbf{w}}, \mathbf{z}^L_y(\hat{s}), \mathbf{z}^U_y(\hat{s}))\right) \leq [f_i]^U\left(\psi^U_i(\hat{s}, \hat{\mathbf{v}}, \hat{\mathbf{w}}, \mathbf{z}^L_y(\hat{s}), \mathbf{z}^U_y(\hat{s}))\right). && (6.98)
\end{aligned}$$

This implies that $K \equiv \{(\hat{s}, \hat{\mathbf{v}}, \hat{\mathbf{w}}) \in I \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} : \hat{\mathbf{v}} \leq \hat{\mathbf{w}}\}$ is a viability domain for the reduced ODEs (Definition 1.1.5 in [13]). Combining this with continuity the right-hand sides, Nagumo's Theorem implies that there exist $\delta > 0$ $s \in C^1([t_1, t_1 + \delta], \mathbb{R})$ and $\mathbf{v}^*, \mathbf{w}^* \in C^1([t_1, t_1 + \delta], \mathbb{R}^{n_x})$ satisfying the reduced ODEs and satisfying $(s(t), \mathbf{v}^*(t), \mathbf{w}^*(t)) \in K$, and hence $\mathbf{v}^*(t) \leq \mathbf{w}^*(t)$, $\forall t \in [t_1, t_1 + \delta]$ (see Theorem 1.2.3 in [13]). But by the definition of $t_1$, $(s, \mathbf{v}, \mathbf{w})$ leaves $K$ immediately to the right of $t_1$. Therefore, $(s, \mathbf{v}, \mathbf{w}) \neq (s, \mathbf{v}^*, \mathbf{w}^*)$ on $[t_1, t_1 + \delta]$. But it has been shown above that the right-hand sides of the reduced ODEs are locally Lipschitz continuous, so a standard application of Gronwall's inequality yields a contradiction. □

**Corollary 6.6.3.** *Let* $(\mathbf{v}, \mathbf{w}, \mathbf{z}^L_y, \mathbf{z}^U_y)$ *be a solution of* (6.85)-(6.89) *on* $I$. *Then any regular solution* $(\mathbf{x}, \mathbf{y})$ *of* (5.1) *on* $I \times P$ *satisfying* $\mathbf{y}(t_0, \tilde{\mathbf{p}}) \in [\mathbf{z}^L_y(t_0), \mathbf{z}^U_y(t_0)]$ *for at least one* $\tilde{\mathbf{p}} \in P$ *also satisfies*

$$\mathbf{x}(t, \mathbf{p}) \in [\mathbf{v}(t), \mathbf{w}(t)], \tag{6.99}$$

$$\mathbf{y}(t, \mathbf{p}) \in \mathcal{Y}(t, \mathbf{v}(t), \mathbf{w}(t), \mathbf{z}^L_y(t), \mathbf{z}^U_y(t)) \equiv \mathcal{H}^q\left(\phi\left(t, \mathbf{v}(t), \mathbf{w}(t), \mathbf{z}^L_y(t), \mathbf{z}^U_y(t)\right)\right), \tag{6.100}$$

*for all* $(t, \mathbf{p}) \in I \times P$ *and any* $q \in \mathbb{N}$.

*Proof.* Consider Hypothesis 5.5.1. By Lemma 6.6.2, the condition (EX) holds. Since $(\mathbf{v}, \mathbf{w}, \mathbf{z}_y^L, \mathbf{z}_y^U)$ satisfy (6.87)-(6.88) on $I$, we must have $(t, \mathbf{v}(t), \mathbf{w}(t), \mathbf{z}_y^L(t), \mathbf{z}_y^U(t)) \in E_{\mathcal{H}}^*$, $\forall t \in I$. Then, by (6.87), (6.88) and Conclusion 5 of Lemma 6.2.5, the condition (ALG) in Hypothesis 5.5.1 also holds. Now, it suffices to establish Hypotheses (IC) and (RHS) of Theorem 5.5.6. (IC) holds by (6.89). To show (RHS).1, choose any $t \in I$ and suppose $\exists(\hat{\mathbf{p}}, \hat{\mathbf{z}}_x, \hat{\mathbf{z}}_y) \in P \times D_x \times [\mathbf{z}_y^L(t), \mathbf{z}_y^U(t)]$ such that $\mathbf{g}(t, \hat{\mathbf{p}}, \hat{\mathbf{z}}_x, \hat{\mathbf{z}}_y) = \mathbf{0}$ and $\hat{\mathbf{z}}_x \in \mathcal{B}_i^L([\mathbf{v}(t), \mathbf{w}(t)])$. By definition,

$$\phi_i^L(t, \mathbf{v}(t), \mathbf{w}(t), \mathbf{z}_y^L(t), \mathbf{z}_y^U(t)) \subset \phi(t, \mathbf{v}(t), \mathbf{w}(t), \mathbf{z}_y^L(t), \mathbf{z}_y^U(t)). \tag{6.101}$$

Then, by Conclusions 5 and 7 of Lemma 6.2.5, satisfaction of (6.87) and (6.88) implies that $\phi_i^L(t, \mathbf{v}(t), \mathbf{w}(t), \mathbf{z}_y^L(t), \mathbf{z}_y^U(t)) \in \mathcal{D}_{\mathcal{H}}^K$. By Conclusion 4 of the same,

$$\mathcal{Y}_i^L(t, \mathbf{v}(t), \mathbf{w}(t), \mathbf{z}_y^L(t), \mathbf{z}_y^U(t)) = \mathcal{H}^K(\phi_i^L(t, \mathbf{v}(t), \mathbf{w}(t), \mathbf{z}_y^L(t), \mathbf{z}_y^U(t))). \tag{6.102}$$

Then, Conclusion 1 of Corollary 5.4.7 ensures that $\hat{\mathbf{z}}_y \in \mathcal{Y}_i^L(t, \mathbf{v}(t), \mathbf{w}(t), \mathbf{z}_y^L(t), \mathbf{z}_y^U(t))$. It follows that

$$f_i(t, \hat{\mathbf{p}}, \hat{\mathbf{z}}_x, \hat{\mathbf{z}}_y) \in [f_i]([t,t], P, \mathcal{B}_i^L([\mathbf{v}(t), \mathbf{w}(t)]), \mathcal{Y}_i^L(t, \mathbf{v}(t), \mathbf{w}(t), \mathbf{z}_y^L(t), \mathbf{z}_y^U(t))), \tag{6.103}$$

$$= [f_i](\psi_i^L(t, \mathbf{v}(t), \mathbf{w}(t), \mathbf{z}_y^L(t), \mathbf{z}_y^U(t))), \tag{6.104}$$

and hence (6.85) ensures that (RHS).1 is satisfied. Proof of (RHS).2 is analogous. □

A primary distinction between the two-phase and single-phase methods thus far is that the former is able to verify existence of a solution, while this has been assumed for the latter. It is shown below that the conditions of Corollary 6.6.3 are in fact sufficient to assert existence as well.

**Theorem 6.6.4.** *Let* $(\mathbf{v}, \mathbf{w}, \mathbf{z}_y^L, \mathbf{z}_y^U)$ *be a solution of* (6.85)-(6.89) *on* $I$*. Then there exists a regular solution* $(\mathbf{x}, \mathbf{y})$ *of* (5.1) *on* $I \times P$ *satisfying* (6.99) *and* (6.100) *for all* $(t, \mathbf{p}) \in I \times P$ *and any* $q \in \mathbb{N}$.

*Proof.* Let $A$ be the set of points $(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}}, \hat{\mathbf{z}}_y^L, \hat{\mathbf{z}}_y^U) \in E_{\mathcal{H}}^*$ such that

$$\mathcal{H}^{*,L}(\phi(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}}, \hat{\mathbf{z}}_y^L, \hat{\mathbf{z}}_y^U)) > \hat{\mathbf{z}}_y^L \quad \text{and} \quad \mathcal{H}^{*,U}(\phi(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}}, \hat{\mathbf{z}}_y^L, \hat{\mathbf{z}}_y^U)) < \hat{\mathbf{z}}_y^U. \tag{6.105}$$

By Theorem (6.6.1), $A$ is open. Furthermore, $A \supset K$, where $K$ is the image of $I$ under $\phi(\cdot, \mathbf{v}(\cdot), \mathbf{w}(\cdot), \mathbf{z}_y^L(\cdot), \mathbf{z}_y^U(\cdot))$ because $(\mathbf{v}, \mathbf{w}, \mathbf{z}_y^L, \mathbf{z}_y^U)$ satisfy (6.87)-(6.88). Because $K$ is compact, $\exists \delta > 0$ such that $\mathbf{q} \in K$ and $\|\mathbf{q} - \mathbf{q}'\|_\infty \leq \delta$ implies $\mathbf{q}' \in A$. As a special case, this implies that (6.105) holds with $(\hat{t}, \hat{\mathbf{v}}, \hat{\mathbf{w}}, \hat{\mathbf{z}}_y^L, \hat{\mathbf{z}}_y^U) = (t, \mathbf{v}(t) - \mathbf{1}\delta, \mathbf{w}(t) + \mathbf{1}\delta, \mathbf{z}_y^L(t), \mathbf{z}_y^U(t))$ for every $t \in I$. Arguing as in Corollary 6.6.3, this implies that Hypothesis 5.5.1 is satisfied with $[\mathbf{v}(t) - \mathbf{1}\delta, \mathbf{w}(t) + \mathbf{1}\delta]$ in place of $[\mathbf{v}(t), \mathbf{w}(t)]$.

Define

$$V_\delta \equiv \{(t, \mathbf{p}, \mathbf{z}_x) \in I \times P \times D_x : \mathbf{z}_x \in [\mathbf{v}(t) - \mathbf{1}\delta, \mathbf{w}(t) + \mathbf{1}\delta]\}. \tag{6.106}$$

By Lemma 5.5.4, $\exists \mathbf{H}_\delta \in C^1(V_\delta, D_y)$ such that, for every $(t, \mathbf{p}, \mathbf{z}_x) \in V_\delta$, $\mathbf{z}_y = \mathbf{H}_\delta(t, \mathbf{p}, \mathbf{z}_x)$ is an element of $Z_y'(t)$ and satisfies $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$ uniquely among elements of $Z_y(t)$.

Now consider the system of ODEs

$$\dot{\mathbf{x}}(t, \mathbf{p}) = \mathbf{f}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{H}_\delta(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}))), \quad \mathbf{x}(t_0, \mathbf{p}) = \mathbf{x}_0(\mathbf{p}). \tag{6.107}$$

By the definition of $C^1$ functions (see §6.2), the right-hand side above is defined and $C^1$ on an open set $\tilde{V} \supset V_\delta$. Fixing any $\mathbf{p} \in P$, it follows that there exists a unique solution of (6.107), $\mathbf{x}(\cdot, \mathbf{p}) \in C^1([t_0, \tilde{t}], D_x)$, for some sufficiently small $\tilde{t} \in (t_0, t_f]$ (see [78], Ch. II, Thm. 1.1). Furthermore, this solution can be extended to a maximal interval of existence $[t_0, t^*)$ such that $(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})) \to \partial\tilde{V}$ as $t \to t^*$ (see [78], Ch. II, Thm. 3.1). Formally, this means that, for any compact $K \subset \tilde{V}$, there exists $\hat{t} \in (t_0, t^*)$ with $(\hat{t}, \mathbf{p}, \mathbf{x}(\hat{t}, \mathbf{p})) \notin K$.

Note that $V_\delta$ is compact and suppose that $t^* \leq t_f$. Then, since $(t_0, \mathbf{p}, \mathbf{x}_0(\mathbf{p})) \in V_\delta$, continuity ensures that $\exists t' \in (t_0, t_f)$ with $(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})) \in V_\delta$, $\forall t \in [t_0, t']$, and $\mathbf{x}(t', \mathbf{p}) \notin [\mathbf{v}(t'), \mathbf{w}(t')]$. Define $\mathbf{y}(t, \mathbf{p}) \equiv \mathbf{H}_\delta(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}))$, $\forall t \in [t_0, t']$. It follows from the properties of $\mathbf{H}_\delta$ on $V_\delta$ that $(\mathbf{x}, \mathbf{y})$ is a solution of (5.1) on $[t_0, t'] \times \{\mathbf{p}\}$. It further follows that $\mathbf{y}(t, \mathbf{p}) \in Z_y(t)$, $\forall t \in [t_0, t']$. Then, Conclusion 3 of Corollary 5.4.7 shows that this solution is regular. By Corollary 6.6.3, this implies that $\mathbf{x}(t', \mathbf{p}) \in [\mathbf{v}(t'), \mathbf{w}(t')]$, which is a contradiction. Therefore, $t^* > t_f$.

Since $\mathbf{p} \in P$ was arbitrary, the previous construction defines $(\mathbf{x}, \mathbf{y}) \in C^1(I \times P, D_x \times D_y)$, which is $C^1$ because $\mathbf{f}$ and $\mathbf{H}_\delta$ are. Arguing as above, this is a regular solution of (5.1) on $I \times P$ and satisfies (6.99) and (6.100) for all $(t, \mathbf{p}) \in I \times P$ and any $q \in \mathbb{N}$. □

In light of Theorem 6.6.4, the single-phase bounding method is simply to solve the DAEs (6.85)-(6.89). Provided that numerical error is not a critical concern, this can be done using any state-of-the-art DAE solver. In the case studies in §6.7 we use `IDA` [82] with absolute and relative tolerances of $10^{-5}$. Furthermore, we choose $K = 4$ and $\gamma(t) = 10^{-4}$, $\forall t \in I$. In addition to the function evaluators, IDA is provided with an additional routine to compute the system Jacobian. This is done using the forward mode AD scheme discussed in §6.4, with the exception that the contribution to the Jacobian owing to the dependence of $\mathbf{C}$ on $(\mathbf{v}, \mathbf{w}, \mathbf{z}_y^L, \mathbf{z}_y^U)$ is ignored.

### 6.6.1 Computational Complexity of the Single-Phase and Two-Phase Methods

Suppose that the cost of evaluating any of the functions $[f_i]$, $[g_j]$ or $[\frac{\partial g_j}{\partial y_k}]$ is $O(m)$, where $m$ can be interpreted as the number of bits required to store the longest code list describing one of these functions (i.e., the factorable representation of Chapter 2). Then complexity of a single evaluation of the right-hand sides of (6.85)-(6.88) is $O\left(n_x K \left(m n_y^2 + n_y^3\right)\right)$. The contributions to this figure are described in Table 6.1. From the table, it can be seen that the cost of a right-hand side evaluation is dominated by the evaluation of $\mathcal{Y}_i^{L/U}$ and hence $\mathcal{H}^{+,K}$. The complexity of this step derives from the $O(m n_y^2)$ evaluation of $[\frac{\partial \mathbf{g}}{\partial \mathbf{y}}]$ and the $O(n_y^3)$ multiplication $\mathbf{C}[\frac{\partial \mathbf{g}}{\partial \mathbf{y}}]$. In addition to right-hand side evaluations, numerical integration of (6.85)-(6.89) will require $O((n_x + n_y)^3)$ operations due to matrix factorization in the corrector iteration.

Table 6.1: Computational complexity of evaluating the right-hand sides of (6.85)-(6.88). The left portion shows the sequence of computations, from top to bottom, using the definitions (6.76)-(6.84). The right portion shows the complexity of evaluating each function on the left, assuming that values for all previous computations (i.e. all quantities directly above the function on the left portion of the table) are given. For functions with subscript $i$, the tabulated complexities are for all $i = 1, \ldots, n_x$ evaluations.

| | | | | | |
|---|---|---|---|---|---|
| $\eta$ | | | $n_x + n_y$ | | |
| $\mathbf{C}$ | | | $mn_y^2 + n_y^3$ | | |
| $\phi$ | $\phi_i^L$ | $\phi_i^U$ | $1$ | $n_x$ | $n_x$ |
| $\mathcal{H}^* \circ \phi$ | $\mathcal{Y}_i^L$ | $\mathcal{Y}_i^U$ | $mn_y^2 + n_y^3$ | $n_x K \left(mn_y^2 + n_y^3\right)$ | $n_x K \left(mn_y^2 + n_y^3\right)$ |
| | $\psi_i^L$ | $\psi_i^U$ | | $0$ | $0$ |
| | $[f_i]^L \circ \psi_i^L$ | $[f_i]^U \circ \psi_i^U$ | | $n_x m$ | $n_x m$ |

The complexity of the two-phase method is the same as that of the single-phase method. By a similar analysis, evaluation of the right-hand sides of (6.47) and (6.48) is $O\left(n_x K \left(mn_y^2 + n_y^3\right)\right)$, while numerical integration requires $O\left(n_x^3\right)$ operations. Phase 1 is dominated by Step 4 of Algorithm 2, which requires the $O((n_y + n_x)^3)$ factorization of $\tilde{\mathbf{J}}$. In practice, we find that the single-phase method is significantly more efficient than the two-phase method (see §6.7).

Table 6.1 suggests some target areas for efficiency gains in the single-phase method, and similar considerations also apply to the two-phase method. An approach that removes a factor of $n_x$ from the entries in the last two columns of the fourth row is to replace each $\mathcal{Y}_i^L$ and $\mathcal{Y}_i^U$ by $\mathcal{Y}(t, \mathbf{v}, \mathbf{w}) \equiv \mathcal{H}^{+,K}(\phi(t, \mathbf{v}, \mathbf{w}))$. It is not difficult to show that Corollary 6.6.3 remains true, and because $\mathcal{Y}$ is used for all $i$, $\mathcal{H}^{+,K}$ only needs to be evaluated once in order to compute the right-hand sides of the entire system. However, the resulting bounds are weaker, and our experience suggests that the original implementation is well worth the effort. Another approach is to eliminate the $n_y^3$ terms in the second and fourth rows of Table 6.1 by using a different preconditioning scheme and/or exploiting sparsity of $\partial \mathbf{g}/\partial \mathbf{y}$. For larger systems, this will become important not only for efficiency, but also because computing $\mathbf{C}$ by direct matrix inversion will become numerically unstable. We leave these considerations for future work.

## 6.7 Case Studies

The computations presented in this section were performed on a Dell Precision T3400 workstation with a 2.83 GHz Intel Core2 Quad CPU. All experiments had one core and 512 MB of memory dedicated to the job. All interval computations and differentiation of interval equations was done using an in house `C++` library based on operator overloading.

**Example 6.7.1** (A simple DAE with a singularity)**.** Consider the semi-explicit DAEs

$$\dot{x}(t,p) = -px(t,p) - 0.1y(t,p), \tag{6.108}$$
$$0 = y(t,p) - \frac{\sin{(p)}}{\sqrt{y(t,p)}} - 25x(t,p),$$

with initial condition $x_0 = 1$ at $t_0 = 0$ and $p \in P \equiv [0.5, 4.0]$. We note that the solutions $y(t, \mathbf{p})$ approach 0 for all $\mathbf{p} \in P$ (Figure 6-2). Since the algebraic equation is not defined at $y = 0$, this poses an interesting challenge for bounding because even slight conservatism in the bounds for $y$ will eventually enclose 0 and cause the methods to fail.

The results of applying the two proposed bounding approaches are shown in Figures 6-1 and 6-2. Note that the refined time-varying bounds computed in Phase 2 of the two-phase method are not shown because they are indistinguishable from those computed by the single-phase method (scrutiny shows that the latter are slightly sharper). The bounds produced by both methods are very sharp until roughly $t = 0.25$, where some slight overestimation becomes apparent. Computational times and other performace statistics are shown in Table 6.3 for various values of $t_f$ (see also Table 6.2).

With $t_f = 0.25$, neither method has any significant difficulty and both produce bounds very efficiently. As $t_f$ is increased to 0.30 and 0.33, the effort required of both methods increases significantly, with the increase for the two-phase method being more pronounced. For both methods, failure occurs around $t = 0.3313$ and bounds cannot be propagated further. For the single-phase method, IDA terminates after the corrector iteration fails to converge with minimum step size. Similarly for the two-phase method, repeated failures in Step 4 of Algorithm 2 cause the time step to be reduced below H_MIN (via Step 6). Indeed, the time steps taken by Algorithm 1 are evident from the staircase structure of the Phase 1 bounds in Figures 6-1 and 6-2, and are seen to shrink dramatically as $t$ approaches 0.3313.

The ultimate cause of failure is that the inclusion (6.11), and analogously the equations (6.87)-(6.88), becomes difficult to satisfy. For the two-phase approach, the statistic STP in Table 6.3 shows that the relative number of failed time steps is increasing with increasing final time. These correspond to failures in Step 4 of Algorithm 2, which are split evenly between cases (a) and (b), with (b) occurring because $0 \in \square(z_y^L, z_y^U)$ for some iterate. In the single-phase approach, the corrector iteration in IDA encounters the same problems. Table 6.3 shows disproportionate increases in both the number of time steps and the number of corrector iterations required by IDA as $t_f$ is increased, indicating that the solver is having trouble satisfying (6.87)-(6.88). Despite their eventual failures, both methods produce bounds over a longer time horizon than any other approaches tried (see Remark 6.4.4).

On the whole, the two bounding methods fail at nearly the same time and produce nearly identical bounds where they are successful. In cases where the two-phase method reaches the final time with few, large time steps, the CPU time is nearly equivalent to that of the single phase method. On the other hand, the single-phase method is significantly faster in the difficult experiments where $t_f$ approaches the failure time of 0.3313.

**Example 6.7.2** (Simple distillation)**.** Consider the simple distillation of a Benzene/ Toluene mixture. Following the analysis in [49], this process can be described by the

Table 6.2: Definition of algorithm statistics presented in Tables 6.3 and 6.4.

| | |
|---|---|
| CPU(s) | Both methods: Computational time for the complete bounding algorithm. |
| Ph1(s) | Two-phase method: Time spent in Phase 1 (Step 3 of Algorithm 1 as in §6.4). |
| Ph2(s) | Two-phase method: Time spent in Phase 2 (Step 6 of Algorithm 1 as in §6.5). |
| STP | Two-phase method: Number of time steps taken by Algorithm 1 over the number of attempted steps (the difference is the number of visits to Step 6 in Algorithm 2). Single-phase method: Number of times steps required by IDA [82] to solve (6.85)-(6.89). |
| CRI | Single-phase method: Cumulative number of corrector iterations during solution of (6.85)-(6.89) by IDA [82]. |

Table 6.3: Algorithm statistics for Example 6.7.1. Columns represents single experiments, which vary in the specified value of $t_f$.

| $t_f$ | 0.25 | 0.30 | 0.33 |
|---|---|---|---|
| Two-Phase Method Statistics | | | |
| CPU(s) | 0.0026 | 0.0055 | 0.0500 |
| Ph1(s) | 0.0007 | 0.0020 | 0.0280 |
| Ph2(s) | 0.0019 | 0.0034 | 0.0212 |
| STP | 4/5 | 11/25 | 100/214 |
| Single-Phase Method Statistics | | | |
| CPU(s) | 0.0020 | 0.0024 | 0.0089 |
| STP | 40 | 45 | 84 |
| CRI | 58 | 73 | 268 |

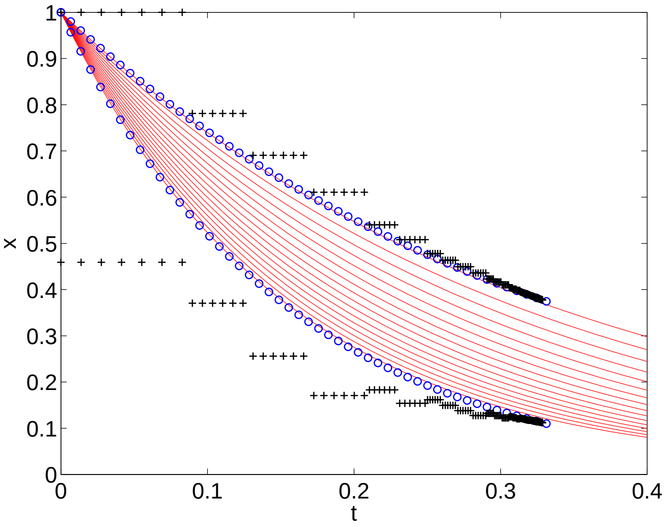

Figure 6-1: Solutions $x(t, p)$ of (6.108) for 16 values of $p \in [0.5, 4.0]$ (solid curves), along with bounds from the single-phase method (circles) and bounds from Phase 1 of the two-phase method (crosses). Bounds from Phase 2 of the two-phase method are indistinguishable from the single-phase bounds and are not shown.

system of semi-explicit index-one DAEs

$$\begin{aligned} \frac{d\phi_{\mathrm{B}}}{d\xi} &= \phi_{\mathrm{B}} - \psi_{\mathrm{B}}, \\ 0 &= \phi_{\mathrm{B}} + \phi_{\mathrm{T}} - 1, \\ 0 &= \psi_{\mathrm{B}} + \psi_{\mathrm{T}} - 1, \\ 0 &= \mathcal{P}\psi_{\mathrm{B}} - \mathcal{P}^{\mathrm{sat}}_{\mathrm{B}}(T)\phi_{\mathrm{B}}, \\ 0 &= \mathcal{P}\psi_{\mathrm{T}} - \mathcal{P}^{\mathrm{sat}}_{\mathrm{T}}(T)\phi_{\mathrm{T}}, \end{aligned} \tag{6.109}$$

where the subscripts B and T denote Benzene and Toluene, respectively, $\phi$ is a liquid phase mole fraction, $\psi$ is a vapor phase mole fraction, $T$ denotes temperature, $\mathcal{P}$ denotes pressure, and the vapor pressures $\mathcal{P}^{\mathrm{sat}}_{\mathrm{B}}(T)$ and $\mathcal{P}^{\mathrm{sat}}_{\mathrm{T}}(T)$ are given by the Antoine expression

$$\log_{10} \mathcal{P}^{\mathrm{sat}}_{i}(T) = A_i - \frac{B_i}{T + C_i}, \quad i \in \{\mathrm{B}, \mathrm{T}\}. \tag{6.110}$$

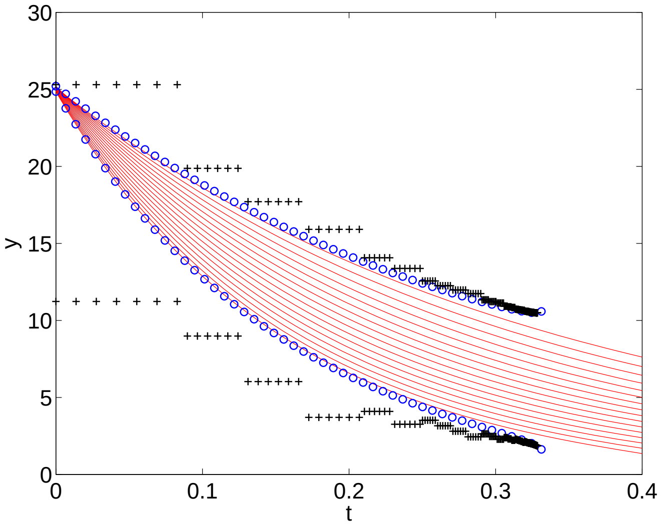

Figure 6-2: Solutions $y(t,p)$ of (6.108) for 16 values of $p \in [0.5, 4.0]$ (solid curves), along with bounds from the single-phase method (circles) and bounds from Phase 1 of the two-phase method (crosses). Bounds from Phase 2 of the two-phase method are indistinguishable from the single-phase bounds and are not shown.

The independent variable $\xi$ is a dimensionless *warped time* (see [49]). The last two equations in (6.109) are derived assuming that Benzene/ Toluene is an ideal mixture. Nominal values of the Antoine coefficients in (6.110) are given for temperature in degrees $C$ and pressures in mm HG in [53] as: $A_{\mathrm{B}} = 6.87987$, $B_{\mathrm{B}} = 1196.76$, $C_{\mathrm{B}} = 219.161$, $A_{\mathrm{T}} = 6.95087$, $B_{\mathrm{T}} = 1342.31$ and $C_{\mathrm{T}} = 219.187$. With $\mathcal{P} = 759.81$ mm Hg constant, we consider bounding the solutions of (6.109), $\mathbf{x} = \phi_{\mathrm{B}}$ and $\mathbf{y} = (\phi_{\mathrm{T}}, \psi_{\mathrm{B}}, \psi_{\mathrm{T}}, T)$, over the interval $\xi \in [0, 6]$, while considering various combinations of the Antoine coefficients as uncertain parameters. Computational times and algorithm statistics are presented in Table 6.4, where the first row indicates the Antoine coefficients which are considered to be uncertain, and the second row describes the interval $P$ as a percent deviation around the nominal values of these coefficients. Though the uncertainty ranges considered may seem small, they describe a wide range of solution behavior because the corresponding parameters appear inside of an exponential in the model equations. Indeed, within a 6% deviation from the nominal value of $A_{\mathrm{B}}$ alone, the most volatile component can switch from Benzene to Toluene.

In the case where $\mathbf{p} = (A_{\mathrm{B}}, B_{\mathrm{B}}, A_{\mathrm{T}}, B_{\mathrm{T}})$ and the deviation is ±0.2%, the results of both bounding methods are shown for $\phi_{\mathrm{B}}$, $\psi_{\mathrm{B}}$ and $T$ in Figures 6-3, 6-4 and 6-5, respectively. Again, the time-varying bounds computed in Phase 2 of the two-phase method are not shown because they are indistinguishable from the single-phase bounds. Both methods provide very tight bounds on $\phi_{\mathrm{B}}$ throughout the $\xi$ interval of interest, and very reasonable bounds on $\psi_{\mathrm{B}}$ and $T$, with tight bounds at the beginning and end of the integration time.

In contrast to the simple example of the previous section, Algorithm 1 is forced to take relatively small time steps here. In Figures 6-3, 6-4 and 6-5, every cross plotted marks the end of a single such step. For experiments requiring many time steps of Algorithm 1, most are taken between $\xi$ values of about 1.2 and 2.6. Within this interval, it is difficult to satisfy the inclusions of Step 3 and the step must be restricted often. In Figures 6-4 and 6-5, sharp jumps in the Phase 1 bounds can be observed at values of $\xi$ where a relatively large step has been achieved after a difficult period through which the step size has been kept small. These jumps reflect the fact that wider $Z_{x,j}$ and $Z_{y,j}$ are required to satisfy (6.11) and (6.12) over large steps. For the single-phase method, one similarly observes that IDA takes more time steps for $\xi \in [1.2, 2.6]$, where it is difficult to satisfy (6.87)-(6.88). When the parameter interval $P$ is sufficiently wide, neither algorithm is able to produce bounds through the difficult region between $\xi = 1.2$ and $\xi = 2.6$ (see Table 6.4). For example, when all six Antoine coefficients are considered as unknown with a ±0.2% deviation, both algorithms fail near $\xi = 1.53$.

As in the first example, the two bounding methods are equally robust and produce nearly identical bounds. However, the single-phase method is faster than the two-phase method in every experiment, with a factor varying between 3.5 to 7.

## 6.8 Conclusions and Future Work

Two methods have been proposed for computing interval bounds on the solutions of semi-explicit index-one DAEs over a range of initial conditions and problem pa-

Table 6.4: Algorithm statistics for Example 6.7.2. Each column represents a single experiment. The first row indicates the model parameters considered as uncertain, and the second row indicates the percent deviation considered around the nominal parameter values. The symbol † indicates that the algorithm terminated unsuccessfully before $\xi = 6.0$.

| | $[A_{\mathrm{B}}\ B_{\mathrm{T}}]$ | | $[A_{\mathrm{B}}\ B_{\mathrm{B}}\ A_{\mathrm{T}}\ B_{\mathrm{T}}]$ | | $[A_{\mathrm{B}}\ B_{\mathrm{B}}\ C_{\mathrm{B}}\ A_{\mathrm{T}}\ B_{\mathrm{T}}\ C_{\mathrm{T}}]$ | |
|---|---|---|---|---|---|---|
| | ±0.2% | ±0.4% | ±0.2% | ±0.3%† | ±0.1% | ±0.2%† |
| $\xi$ | 6.0 | 6.0 | 6.0 | 1.090 | 6.0 | 1.534 |
| Two-Phase Method Statistics | | | | | | |
| CPU(s) | 0.073 | 0.1610 | 0.1637 | 0.24 | 0.0929 | 0.22 |
| Ph1(s) | 0.0315 | 0.0746 | 0.0800 | 0.16 | 0.0413 | 0.15 |
| Ph2(s) | 0.0412 | 0.0862 | 0.0835 | 0.08 | 0.0516 | 0.07 |
| STP | 44/88 | 93/187 | 96/193 | 100/214 | 55/110 | 100/209 |
| Single-Phase Method Statistics | | | | | | |
| CPU(s) | 0.0204 | 0.0229 | 0.0241 | 0.06 | 0.0185 | 0.06 |
| STP | 77 | 83 | 103 | 89 | 77 | 91 |
| CRI | 110 | 132 | 160 | 259 | 103 | 244 |

rameters. The first method is a two-phase approach using an interval existence and uniqueness test in Phase 1 and a refinement procedure based on differential inequalities in Phase 2. Efficient implementations for both phases were presented using interval computations and a state-of-the-art ODE solver. The second method combines the two phases of the first method and requires numerical solution of a system of semi-explicit DAEs. Two case studies were considered, demonstrating that both methods produce sharp bounds very efficiently, with the single-phase method being consistently faster.

Several potential improvements to the presented algorithms remain to be explored. In the case of ODEs, it has been shown that problem specific physical information can often be incorporated into bounding methods based on differential inequalities to achieve significantly sharper bounds (see Chapters 3 and 4). The use of such information should be explored for sharpening the results in Theorems 5.4.8, 5.5.2, 5.5.3 and 5.5.6. The bounding methods presented here demonsrate that the presence of implicit equations can be overcome through the use of interval Newton methods. Thus, a further area of research is to extend these ideas to the problem of bounding

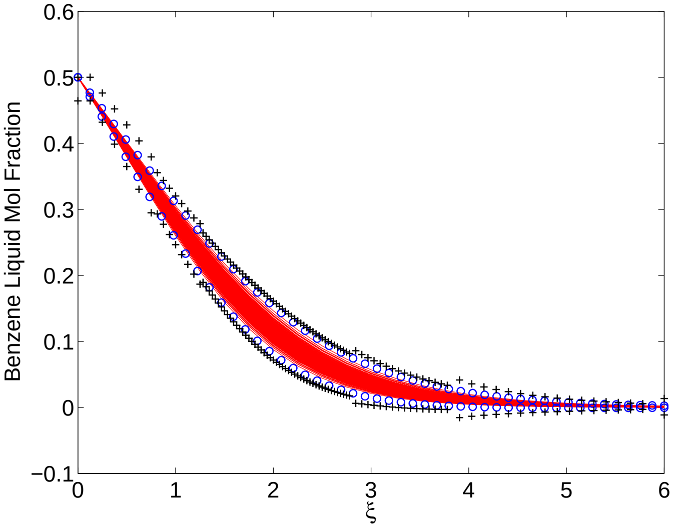

Figure 6-3: Solutions $\phi_{\mathrm{B}}(\xi, \mathbf{p})$ of (6.109) for $\mathbf{p} = (A_{\mathrm{B}}, B_{\mathrm{B}}, A_{\mathrm{T}}, B_{\mathrm{T}})$ uniformly sampled within a $\pm 0.2\%$ deviation from nominal values (solid curves), along with bounds from the single-phase method (circles) and bounds from Phase 1 of the two-phase method (crosses). Bounds from Phase 2 of the two-phase method are indistinguishable from the single-phase bounds and are not shown.

fully implicit DAEs. Finally, these ideas could also be used to compute bounds on the solutions of high-index systems by combining the approach here with the derivative array equations.

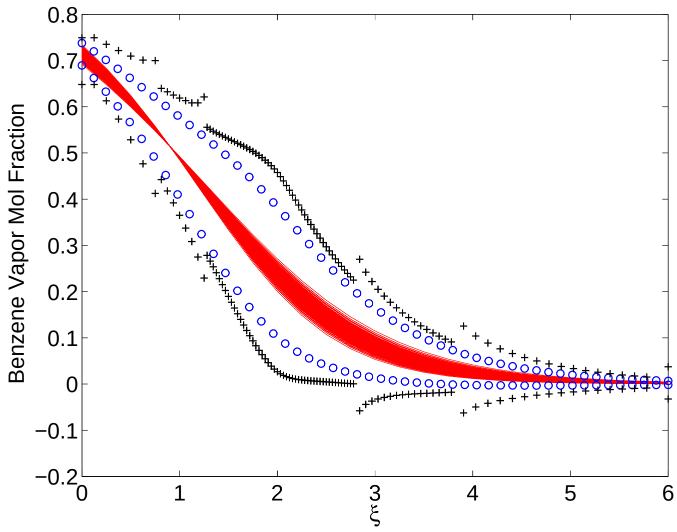

Figure 6-4: Solutions $\psi_{\mathrm{B}}(\xi, \mathbf{p})$ of (6.109) for $\mathbf{p} = (A_{\mathrm{B}}, B_{\mathrm{B}}, A_{\mathrm{T}}, B_{\mathrm{T}})$ uniformly sampled within a ±0.2% deviation from nominal values (solid curves), along with bounds from the single-phase method (circles) and bounds from Phase 1 of the two-phase method (crosses). Bounds from Phase 2 of the two-phase method are indistinguishable from the single-phase bounds and are not shown.

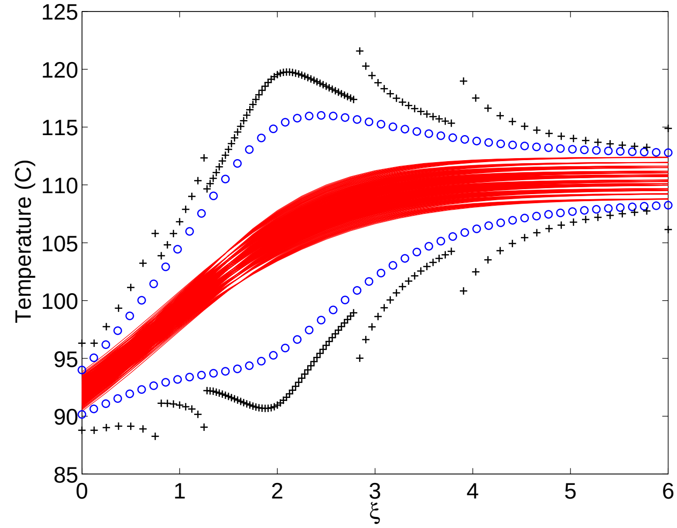

Figure 6-5: Solutions $T(\xi, \mathbf{p})$ of (6.109) for $\mathbf{p} = (A_{\mathrm{B}}, B_{\mathrm{B}}, A_{\mathrm{T}}, B_{\mathrm{T}})$ uniformly sampled within a ±0.2% deviation from nominal values (solid curves), along with bounds from the single-phase method (circles) and bounds from Phase 1 of the two-phase method (crosses). Bounds from Phase 2 of the two-phase method are indistinguishable from the single-phase bounds and are not shown.

# Chapter 7

# State Relaxations for Parametric ODEs

## 7.1 Introduction

In this chapter, two methods are developed for computing convex and concave relaxations of the parametric solutions of nonlinear ordinary differential equations (ODEs). In particular, a general system of ODEs is considered where both the initial conditions and the right-hand side functions depend on a real parameter vector. Given such a system, an auxiliary system of ODEs is derived which describes convex underestimators and concave overestimators for each of the state variables with respect to the parameters, pointwise in the independent variable. These relaxations are termed *state relaxations*.

There are two motivations for computing state relaxations here. First, they provide another method for enclosing the reachable sets of parametric ODEs, in addition to the methods for computing state bounds presented in Chapter 3. Since state relaxations are parameter dependent and are evaluated one parameter value at a time, deriving a useful enclosure from them is not entirely direct and is the subject of Chapter 9. The second motivation for computing state relaxations is for their use in deterministic global optimization algorithms for problems with ODEs embedded [135, 164, 104]. The computation of state relaxations for this application is the subject of several recent articles [162, 157, 151, 150]. Largely owing to the difficulty of this computation and the weaknesses of available methods, it remains an unfortunate fact that state-of-the-art deterministic methods for global dynamic optimization can only solve problems of modest size with reasonable computational effort, typically on the order of 5 state variables and 5 decisions. On the other hand, potential applications for such techniques are ubiquitous, including parameter estimation problems with dynamic models [163, 55, 42, 103], optimal control of batch processes [167, 34], safety verification problems [85], optimal catalyst blending [108], optimal drug scheduling [35, 116], etc. Moreover, representative case studies in the literature suggest that these applications commonly lead to problems with multiple suboptimal local minima, especially when the embedded dynamic system involves a model of chemical reaction kinetics [108, 55, 16]. Thus, the need for improved relaxation techniques is clear.

The first method for computing state relaxations was proposed by Esposito and Floudas [54] using a dynamic extension of the $\alpha BB$ convexification theory described in [7]. This method relies on a finite sampling step to bound the second-order sensitivities of the ODEs, and therefore cannot guarantee that the resulting relaxations are convex. In [135], bounds on these sensitivities are computed, resulting in guaranteed convex relaxations, yet these relaxations are typically very weak and the second-order sensitivities are costly to evaluate. Much more recently, two related approaches have been developed in which McCormick's relaxation technique is applied to a characterization of the ODE solution by a Taylor expansion with a rigorous enclosure of the truncation error [151, 150]. These methods extend interval bounding techniques based on a similar analysis [104] and appear capable of providing very tight relaxations when a sufficiently high-order expansion is used. On the other hand, computing relaxations of a high-order Taylor expansion is very expensive for high-dimensional problems, and the existence of an appropriate compromise in the context of global optimization remains an open question.

In this chapter, we consider methods of a third type [161, 162, 157, 158], where state relaxations are computed as the solutions of an auxiliary system of ODEs which are derived by relaxing the right-hand sides of the original ODEs in various manners. Methods of this type have the advantage that they are efficient and relatively simple to implement. Since the auxiliary ODEs can be solved by numerical integration, the cost of evaluating these relaxations is comparable to that of simulating the original dynamic system. This approach originates with the method in [162], which describes affine state relaxations. Here, two new classes of auxiliary ODEs are defined, and both are proven to describe valid state relaxations as their solutions. In contrast to the method in [162], both of these methods produce state relaxations that are potentially non-affine with respect to the ODE parameters (these will be called *nonlinear* for brevity).

In order to develop these methods, we consider in a general context the basic requirements that must be imposed on an auxiliary system in order to guarantee that it describes valid state relaxations as its solutions. We arrive at two independent sets of sufficient conditions, leading to the two proposed relaxation methods. The first set of conditions, termed *relaxation amplifying dynamics*, was developed first and can be shown to provide much tighter relaxations than the affine theory in [162] for highly nonlinear problems on large parameter ranges. This is due to the fact that the parametric ODE solution is itself highly nonlinear over large parameter ranges, and can therefore be better approximated by a nonlinear relaxation. On the other hand, preliminary numerical experiments show that the affine relaxations are often superior in the context of branch-and-bound global optimization because they provide tighter relaxations over small intervals, where the original ODE solution is only weakly nonlinear.

Motivated by these observations, we present a conceptual analysis of the conditions of relaxation amplifying dynamics and demonstrate that relaxations resulting from this theory necessarily have several undesirable properties. At the same time, this analysis suggests a much weaker set of conditions, termed *relaxation preserving dynamics*, which essentially integrate the most advantageous aspects of the original nonlinear relaxation theory and the affine theory in [162]. A second method is then derived based on these weaker conditions.

In §7.6, we develop numerical methods for computing state relaxations according to both of the theories discussed above. In both cases, auxiliary systems satisfying the required properties are derived through the use of natural McCormick extensions according to the generalized McCormick relaxation technique presented in Chapter 2. Like the affine relaxation theory in [162], evaluating these relaxations involves a single numerical integration of the auxiliary system. For the relaxations derived through the use of convexity amplifying dynamics, this simulation is straightforward. For the relaxations derived through the use of convexity preserving dynamics, the auxiliary system is slightly more complicated and must be simulated as a hybrid system with state events [136].

Several other seemingly related notions of convexity and relaxation appear in the literature on optimal control and ODE theory which are relevant to this work in varying degrees. In [15], sufficient conditions are given under which an optimal control problem on a general Hilbert space is convex, based on classical results on the composition of convex functions. If this Hilbert space is taken as a finite-dimensional real vector space, as would result from reformulation through control parameterization [173], this notion of convexity is equivalent to that in the work presented here. However, the conditions in [15] are extremely restrictive, and no constructive procedure is given for generating convex and concave relaxations of nonconvex problems. In more classical results regarding sufficient optimality conditions for optimal control problems [177, 159], convexity of the Hamiltonian is assumed with respect to the state variables and the controls. Convexity in this sense treats the states and controls as unrelated, whereas the purpose of this work is to approximate the parametric dependence of the state variables by convex and concave functions, so these notions are distinct. The article [143] (and the references therein) details conditions for the reachable set of a system of ODEs beginning from a ball of initial conditions to be convex. Again, this is an unrelated notion because a convex set in state space does not imply convex dependence on the initial conditions for each state variable, nor the converse. Finally, the term relaxation is often applied to optimal control and variational problems where the set of admissible controls is enlarged or embedded in a larger space (i.e., measure-valued controls), and/or the cost functional is underestimated by a lower semicontinuous functional [183, 64]. Though similar in spirit, the type of relaxations considered here are fundamentally different (see Definition 7.2.1).

## 7.2 State Relaxations for a Generic Function

Let $I = [t_0, t_f] \subset \mathbb{R}$ and $P \subset \mathbb{R}^{n_p}$ be compact intervals and let $\mathbf{x} : I \times P \to \mathbb{R}^{n_x}$ be a continuous function such that $\mathbf{x}(\cdot, \mathbf{p})$ is absolutely continuous on $I$ for every $\mathbf{p} \in P$. In this section and the following two, we consider computing state relaxations for such an arbitrary function. The purpose of this generality is to apply the same theory to ODEs and DAEs alike. State relaxations for $\mathbf{x}$ are defined as follows.

**Definition 7.2.1.** Continuous functions $\mathbf{x}^{cv}, \mathbf{x}^{cc} : I \times P \to \mathbb{R}^{n_x}$ are called *state relaxations* for $\mathbf{x}$ on $I \times P$ if, for every $t \in I$, $\mathbf{x}^{cv}(t, \cdot)$ is convex on $P$, $\mathbf{x}^{cc}(t, \cdot)$ is concave on $P$, and $\mathbf{x}^{cv}(t, \mathbf{p}) \leq \mathbf{x}(t, \mathbf{p}) \leq \mathbf{x}^{cc}(t, \mathbf{p})$, $\forall \mathbf{p} \in P$.

**Remark 7.2.2.** The requirement that $P$ is an $n_p$-dimensional compact interval is primarily for computational reasons. The theoretical developments to follow could deal just as easily with a more general compact, convex set in $\mathbb{R}^{n_p}$. In particular, McCormick's relaxation technique [118] requires that $P$ be an interval.

With one caveat, the approaches in this thesis compute state relaxations as the solutions of an auxiliary system of ODEs of the form

$$\begin{aligned} \dot{\mathbf{x}}^{cv}(t, \mathbf{p}) &= \mathbf{u}(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p})), \quad \mathbf{x}^{cv}(t_0, \mathbf{p}) = \mathbf{x}_0^{cv}(\mathbf{p}), \\ \dot{\mathbf{x}}^{cc}(t, \mathbf{p}) &= \mathbf{o}(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p})), \quad \mathbf{x}^{cc}(t_0, \mathbf{p}) = \mathbf{x}_0^{cc}(\mathbf{p}), \end{aligned} \tag{7.1}$$

where $\mathbf{x}_0^{cv}, \mathbf{x}_0^{cc} : P \to \mathbb{R}^{n_x}$ and $\mathbf{u}, \mathbf{o} : I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$. The following regularity assumption holds throughout this chapter.

**Assumption 7.2.3.** The ODEs (7.1) satisfy the following conditions:

1. $\mathbf{x}_0^{cv}$ and $\mathbf{x}_0^{cc}$ are continuous on $P$,

2. $\mathbf{u}$ and $\mathbf{o}$ are continuous on $I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$,

3. There exists $L \in \mathbb{R}_+$ such that

$$\begin{aligned}\|\mathbf{u}(t,\mathbf{p},\mathbf{z}^{cv},\mathbf{z}^{cc}) - \mathbf{u}(t,\mathbf{p},\hat{\mathbf{z}}^{cv},\hat{\mathbf{z}}^{cc})\|_\infty + \|\mathbf{o}(t,\mathbf{p},\mathbf{z}^{cv},\mathbf{z}^{cc}) - \mathbf{o}(t,\mathbf{p},\hat{\mathbf{z}}^{cv},\hat{\mathbf{z}}^{cc})\|_\infty& \\ \leq L\left(\|\mathbf{z}^{cv} - \hat{\mathbf{z}}^{cv}\|_\infty + \|\mathbf{z}^{cc} - \hat{\mathbf{z}}^{cc}\|_\infty\right)&\end{aligned}$$

for all $(t,\mathbf{p},\mathbf{z}^{cv},\mathbf{z}^{cc},\hat{\mathbf{z}}^{cv},\hat{\mathbf{z}}^{cc}) \in I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$.

It is always assumed that the functions $\mathbf{x}_0^{cv}$ and $\mathbf{x}_0^{cc}$ are, respectively, convex and concave relaxations of $\mathbf{x}(t_0,\cdot)$ on $P$. The conditions that are required of $\mathbf{u}$ and $\mathbf{o}$ in order to guarantee that (7.1) furnishes state relaxations as its solutions are of course more difficult to formulate and impose. This is the primary question addressed in this chapter, and two sets of sufficient conditions are presented.

Once these conditions have been formulated, they are applied to the case where $\mathbf{x}$ is the solution of a system of parametric ODEs in §7.6, and to the case where $\mathbf{x}$ is the solution of a system of semi-explicit DAEs in Chapter 8. In both of these cases, the required functions $\mathbf{u}$ and $\mathbf{o}$ are derived using the generalized McCormick relaxations of Chapter 2.

Both in the construction of $\mathbf{u}$ and $\mathbf{o}$ in these cases, and in the statement of the required conditions on these functions, we will make use of state bounds. These can of course be computed by any of the methods in Chapters 3-6.

**Assumption 7.2.4.** State bounds $\mathbf{x}^L, \mathbf{x}^U : I \to \mathbb{R}^{n_x}$ for $\mathbf{x}$ on $I \times P$ are available; i.e., $\mathbf{x}(t,\mathbf{p}) \in X(t) \equiv [\mathbf{x}^L(t), \mathbf{x}^U(t)]$, $\forall (t,\mathbf{p}) \in I \times P$.

## 7.3 Relaxation Amplifying Dynamics

In this section, we give the first set of conditions on $(\mathbf{u},\mathbf{o})$, under the name *relaxation amplifying dynamics*, that guarantee that (7.1) furnishes state relaxations of $\mathbf{x}$ as its solutions. The functions $(\mathbf{u},\mathbf{o})$ are said to describe relaxation amplifying dynamics if they describe both *bound amplifying dynamics* and *convexity amplifying dynamics*, defined below.

**Definition 7.3.1.** The functions $(\mathbf{u}, \mathbf{o})$ describe *bound amplifying dynamics* for $\mathbf{x}$ on $I \times P$ if, for arbitrary functions $\mathbf{z}^{cv}, \mathbf{z}^{cc} : I \times P \to \mathbb{R}^{n_x}$ and every $\mathbf{p} \in P$, the following condition holds: For a.e. $t \in I$ such that $\mathbf{z}^{cv}(t, \mathbf{p}) \leq \mathbf{x}(t, \mathbf{p}) \leq \mathbf{z}^{cc}(t, \mathbf{p})$, $\mathbf{u}$ and $\mathbf{o}$ satisfy

$$\mathbf{u}(t, \mathbf{p}, \mathbf{z}^{cv}(t, \mathbf{p}), \mathbf{z}^{cc}(t, \mathbf{p})) \leq \dot{\mathbf{x}}(t, \mathbf{p}) \leq \mathbf{o}(t, \mathbf{p}, \mathbf{z}^{cv}(t, \mathbf{p}), \mathbf{z}^{cc}(t, \mathbf{p})).$$

Note that the condition of the previous definition holds *pointwise* in $\mathbf{p}$; i.e., both the hypotheses and the conclusion need only hold at a single $\mathbf{p} \in P$. In order to make equally general statements of convexity/concavity assumptions on $(\mathbf{u}, \mathbf{o})$, the notion of a function being *consistent with convexity* at a point is useful.

**Definition 7.3.2.** A function $\mathbf{g} : P \to \mathbb{R}^n$ is *consistent with convexity* at a point $(\lambda, \mathbf{p}_1, \mathbf{p}_2) \in (0, 1) \times P \times P$ if

$$\mathbf{g}(\lambda \mathbf{p}_1 + (1 - \lambda)\mathbf{p}_2) \leq \lambda \mathbf{g}(\mathbf{p}_1) + (1 - \lambda)\mathbf{g}(\mathbf{p}_2).$$

It is consistent with concavity at $(\lambda, \mathbf{p}_1, \mathbf{p}_2)$ if the opposite (weak) inequality holds.

**Definition 7.3.3.** The functions $(\mathbf{u}, \mathbf{o})$ describe *convexity amplifying dynamics* for $\mathbf{x}$ on $I \times P$ if, for arbitrary functions $\mathbf{z}^{cv}, \mathbf{z}^{cc} : I \times P \to \mathbb{R}^{n_x}$ and every $(\lambda, \mathbf{p}_1, \mathbf{p}_2) \in (0, 1) \times P \times P$, the following condition holds: For a.e. $t \in I$ such that

1. $\mathbf{z}^{cv}(t, \cdot)$ is consistent with convexity at $(\lambda, \mathbf{p}_1, \mathbf{p}_2)$,

2. $\mathbf{z}^{cc}(t, \cdot)$ is consistent with concavity at $(\lambda, \mathbf{p}_1, \mathbf{p}_2)$,

3. $\mathbf{z}^{cv}(t, \mathbf{q}) \leq \mathbf{x}(t, \mathbf{q}) \leq \mathbf{z}^{cc}(t, \mathbf{q})$, $\forall \mathbf{q} \in \{\mathbf{p}_1, \mathbf{p}_2, \lambda \mathbf{p}_1 + (1 - \lambda)\mathbf{p}_2\}$,

the functions

$$P \ni \mathbf{p} \longmapsto \mathbf{u}(t, \mathbf{p}, \mathbf{z}^{cv}(t, \mathbf{p}), \mathbf{z}^{cc}(t, \mathbf{p})) \quad \text{and} \quad P \ni \mathbf{p} \longmapsto \mathbf{o}(t, \mathbf{p}, \mathbf{z}^{cv}(t, \mathbf{p}), \mathbf{z}^{cc}(t, \mathbf{p}))$$

are, respectively, consistent with convexity and consistent with concavity at $(\lambda, \mathbf{p}_1, \mathbf{p}_2)$.

To interpret these definitions, let $(\mathbf{x}^{cv}, \mathbf{x}^{cc})$ be a solution of (7.1) and suppose that, for some $\hat{t} \in I$, it is known that $\mathbf{x}^{cv}(\hat{t}, \cdot)$ and $\mathbf{x}^{cc}(\hat{t}, \cdot)$ are, respectively, convex and concave relaxations of $\mathbf{x}(\hat{t}, \cdot)$ on $P$. Then applying Definition 7.3.1 with $\mathbf{z}^{cv} \equiv \mathbf{x}^{cv}$ and $\mathbf{z}^{cc} \equiv \mathbf{x}^{cc}$, and using Equation (7.1),

$$\dot{\mathbf{x}}^{cv}(\hat{t}, \mathbf{p}) \leq \dot{\mathbf{x}}(\hat{t}, \mathbf{p}) \leq \dot{\mathbf{x}}^{cc}(\hat{t}, \mathbf{p}), \quad \forall \mathbf{p} \in P. \tag{7.2}$$

Moreover, choosing any $(\lambda, \mathbf{p}_1, \mathbf{p}_2) \in (0, 1) \times P \times P$ and letting $\bar{\mathbf{p}} \equiv \lambda \mathbf{p}_1 + (1 - \lambda)\mathbf{p}_2$, Definition 7.3.3 implies that

$$\begin{aligned} \dot{\mathbf{x}}^{cv}(\hat{t}, \bar{\mathbf{p}}) &\leq \lambda \dot{\mathbf{x}}^{cv}(\hat{t}, \mathbf{p}_1) + (1 - \lambda)\dot{\mathbf{x}}^{cv}(\hat{t}, \mathbf{p}_2), \\ \dot{\mathbf{x}}^{cc}(\hat{t}, \bar{\mathbf{p}}) &\geq \lambda \dot{\mathbf{x}}^{cc}(\hat{t}, \mathbf{p}_1) + (1 - \lambda)\dot{\mathbf{x}}^{cc}(\hat{t}, \mathbf{p}_2). \end{aligned} \tag{7.3}$$

Intuitively, these inequalities suggest that the bounding properties and the convexity and concavity properties of $(\mathbf{x}^{cv}, \mathbf{x}^{cc})$ should be preserved to the right of $\hat{t}$. Indeed, we have the following theorem.

**Theorem 7.3.4.** *Suppose that Assumption 7.2.3 holds. Let* $\mathbf{x}_0^{cv}, \mathbf{x}_0^{cc} : P \to \mathbb{R}^{n_x}$ *be, respectively, convex and concave relaxations of* $\mathbf{x}_0$ *on* $P$*, and let* $(\mathbf{u}, \mathbf{o})$ *describe relaxation amplifying dynamics for* $\mathbf{x}$ *on* $I \times P$*. Then* (7.1) *has a unique solution* $(\mathbf{x}^{cv}, \mathbf{x}^{cc})$ *on all of* $I \times P$*, and* $\mathbf{x}^{cv}$ *and* $\mathbf{x}^{cc}$ *are state relaxations for* $\mathbf{x}$ *on* $I \times P$*.*

This result shows that state relaxations for $\mathbf{x}$ can be evaluated by simply integrating any auxiliary system that satisfies Assumption 7.2.3 and describes relaxation amplifying dynamics. In §7.6, we show how to construct such a system automatically using generalized McCormick relaxations in the case where $\mathbf{x}$ is the solution of a system of parametric ODEs. Combining this with a state-of-the-art numerical integration code, Theorem 7.3.4 provides a simple and efficient means of computing state relaxations. However, this method also has some significant drawbacks that lead to the second, weaker set of conditions on $(\mathbf{u}, \mathbf{o})$ presented in §7.4.

### 7.3.1 Proof of Theorem 7.3.4

#### Preliminaries

The proof uses a standard construction in ODE theory known as successive approximations (or Picard iterates) [43], presented in the following theorem.

**Theorem 7.3.5.** *Let $I = [t_0, t_f] \subset \mathbb{R}$, $P \in \mathbb{IR}^{n_p}$, and let $\mathbf{v}_0 : P \to \mathbb{R}^{n_v}$ and $\mathbf{h} : I \times P \times \mathbb{R}^{n_v} \to \mathbb{R}^{n_v}$ be continuous functions. Furthermore, suppose $\exists L \in \mathbb{R}_+$ such that*

$$\|\mathbf{h}(t,\mathbf{p},\mathbf{z}) - \mathbf{h}(t,\mathbf{p},\hat{\mathbf{z}})\|_1 \leq L\|\mathbf{z} - \hat{\mathbf{z}}\|_1, \quad \forall (t,\mathbf{p},\mathbf{z},\hat{\mathbf{z}}) \in I \times P \times \mathbb{R}^{n_v} \times \mathbb{R}^{n_v}.$$

*Given any continuous function $\mathbf{v}^0 : I \times P \to \mathbb{R}^{n_v}$, the successive approximations defined recursively by*

$$\mathbf{v}^{k+1}(t,\mathbf{p}) = \mathbf{v}_0(\mathbf{p}) + \int_{t_0}^{t} \mathbf{h}(s,\mathbf{p},\mathbf{v}^k(s,\mathbf{p}))ds \tag{7.4}$$

*exist as continuous functions on $I \times P$ and converge uniformly to a solution of*

$$\dot{\mathbf{v}}(t,\mathbf{p}) = \mathbf{h}(t,\mathbf{p},\mathbf{v}(t,\mathbf{p})), \quad \mathbf{v}(t_0,\mathbf{p}) = \mathbf{v}_0(\mathbf{p}), \tag{7.5}$$

*on $I \times P$. Furthermore, this solution is unique.*

*Proof.* By hypothesis, $\mathbf{v}^0$ is defined and continuous on all of $I \times P$. Supposing this is true of $\mathbf{v}^k$, (7.4) defines $\mathbf{v}^{k+1}$ on all of $I \times P$ and continuity follows from the continuity of $\mathbf{v}_0$ and $\mathbf{h}$. Thus, induction shows that each $\mathbf{v}^k$ is defined and continuous on all of $I \times P$.

Now define

$$\gamma \equiv \max_{(t,\mathbf{p}) \in I \times P} \|\mathbf{h}(t,\mathbf{p},\mathbf{v}^1(t,\mathbf{p})) - \mathbf{h}(t,\mathbf{p},\mathbf{v}^0(t,\mathbf{p}))\|_1.$$

It will be shown that

$$\|\mathbf{v}^{k+1}(t,\mathbf{p}) - \mathbf{v}^{k}(t,\mathbf{p})\|_1 \leq \frac{\gamma L^k (t-t_0)^k}{Lk!}, \tag{7.6}$$

for all $(t,\mathbf{p}) \in I \times P$ and every $k \in \mathbb{N}$. For $k = 1$, (7.4) directly gives

$$\|\mathbf{v}^{2}(t,\mathbf{p}) - \mathbf{v}^{1}(t,\mathbf{p})\|_1 \leq \int_{t_0}^{t} \|\mathbf{h}(s,\mathbf{p},\mathbf{v}^1(s,\mathbf{p})) - \mathbf{h}(s,\mathbf{p},\mathbf{v}^0(s,\mathbf{p}))\|_1 ds \leq \gamma(t-t_0),$$

for all $(t,\mathbf{p}) \in I \times P$. Supposing that (7.6) holds for some arbitrary $k$, it must also hold for $k+1$ since

$$\begin{aligned}
\|\mathbf{v}^{k+2}(t,\mathbf{p}) - \mathbf{v}^{k+1}(t,\mathbf{p})\|_1 &\leq \int_{t_0}^{t} \|\mathbf{h}(s,\mathbf{p},\mathbf{v}^{k+1}(s,\mathbf{p})) - \mathbf{h}(s,\mathbf{p},\mathbf{v}^k(s,\mathbf{p}))\|_1 ds, \\
&\leq L \int_{t_0}^{t} \|\mathbf{v}^{k+1}(s,\mathbf{p}) - \mathbf{v}^{k}(s,\mathbf{p})\|_1 ds, \\
&\leq \frac{\gamma L^{k+1}}{Lk!} \int_{t_0}^{t} (s-t_0)^k ds, \\
&\leq \frac{\gamma L^{k+1}(t-t_0)^{k+1}}{L(k+1)!},
\end{aligned}$$

for all $(t,\mathbf{p}) \in I \times P$. Thus, induction proves (7.6). Now, for any $n, m \in \mathbb{N}$ with $m > n$, Equation (7.6) and the triangle inequality give

$$\begin{aligned}
\|\mathbf{v}^{n}(t,\mathbf{p}) - \mathbf{v}^{m}(t,\mathbf{p})\|_1 &\leq \|\mathbf{v}^{n+1}(t,\mathbf{p}) - \mathbf{v}^{n}(t,\mathbf{p})\|_1 + \ldots + \|\mathbf{v}^{m}(t,\mathbf{p}) - \mathbf{v}^{m-1}(t,\mathbf{p})\|_1, \\
&\leq \frac{\gamma L^n (t_f-t_0)^n}{Ln!} + \ldots + \frac{\gamma L^{m-1}(t_f-t_0)^{m-1}}{L(m-1)!}, \\
&\leq \sum_{k=n}^{\infty} \frac{\gamma L^k (t_f-t_0)^k}{Lk!},
\end{aligned}$$

for all $(t,\mathbf{p}) \in I \times P$. But

$$\sum_{k=0}^{\infty} \frac{\gamma L^k (t_f-t_0)^k}{Lk!} = \frac{\gamma}{L} e^{L(t_f-t_0)} < \infty,$$

and hence $\lim_{n\to\infty}\sum_{k=n}^{\infty}\frac{\gamma L^k(t_f-t_0)^k}{Lk!} = 0$, which implies that the sequence $\{\mathbf{v}^k\}$ is uniformly Cauchy on $I \times P$, and hence converges uniformly to a continuous limit function there.

Next it is shown that this limit function, denoted $\mathbf{v}$, is a solution of (7.5) on $I \times P$. From the Lipschitz condition on $\mathbf{h}$,

$$\|\int_{t_0}^{t}\mathbf{h}(s,\mathbf{p},\mathbf{v}^k(s,\mathbf{p}))ds - \int_{t_0}^{t}\mathbf{h}(s,\mathbf{p},\mathbf{v}(s,\mathbf{p}))ds\|_1 \leq L\int_{t_0}^{t}\|\mathbf{v}^k(s,\mathbf{p}) - \mathbf{v}(s,\mathbf{p})\|_1 ds,$$

for all $(t,\mathbf{p}) \in I \times P$, so that the uniform convergence of $\{\mathbf{v}^k\}$ to $\mathbf{v}$ on $I \times P$ implies that $\lim_{k\to\infty}\int_{t_0}^{t}\mathbf{h}(s,\mathbf{p},\mathbf{v}^k(s,\mathbf{p}))ds = \int_{t_0}^{t}\mathbf{h}(s,\mathbf{p},\mathbf{v}(s,\mathbf{p}))ds$, for all $(t,\mathbf{p}) \in I \times P$. Then, taking limits on both sides of (7.4) gives

$$\mathbf{v}(t,\mathbf{p}) = \mathbf{v}_0(\mathbf{p}) + \int_{t_0}^{t}\mathbf{h}(s,\mathbf{p},\mathbf{v}(s,\mathbf{p}))ds, \quad \forall(t,\mathbf{p}) \in I \times P,$$

which, by the fundamental theorem of calculus and continuity of the integrand, implies that $\mathbf{v}$ is a solution of (7.5). Uniqueness of $\mathbf{v}$ now follows (for each fixed $\mathbf{p} \in P$), by a standard application of Gronwall's inequality (Theorem 1.1, Ch. III, [78]). □

### Proof

Define $\mathbf{x}^{cv,0}(t,\mathbf{p}) = \mathbf{x}^L(t)$ and $\mathbf{x}^{cc,0}(t,\mathbf{p}) = \mathbf{x}^U(t)$, $\forall(t,\mathbf{p}) \in I \times P$, and consider the successive approximations defined recursively by

$$\begin{aligned}\mathbf{x}^{cv,k+1}(t,\mathbf{p}) &= \mathbf{x}_0^{cv}(\mathbf{p}) + \int_{t_0}^{t}\mathbf{u}(s,\mathbf{p},\mathbf{x}^{cv,k}(s,\mathbf{p}),\mathbf{x}^{cc,k}(s,\mathbf{p}))ds, \\ \mathbf{x}^{cc,k+1}(t,\mathbf{p}) &= \mathbf{x}_0^{cc}(\mathbf{p}) + \int_{t_0}^{t}\mathbf{o}(s,\mathbf{p},\mathbf{x}^{cv,k}(s,\mathbf{p}),\mathbf{x}^{cc,k}(s,\mathbf{p}))ds.\end{aligned} \tag{7.7}$$

Note that $\mathbf{u}$ and $\mathbf{o}$ are defined on $I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ and Lipschitz on all of $\mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ uniformly on $I \times P$ by Assumption 7.2.3. Thus, Theorem 7.3.5 may be applied to (7.1), which proves that the successive approximations $\mathbf{x}^{cv,k}$ and $\mathbf{x}^{cc,k}$ in (7.7) exist and converge uniformly to the unique solutions of (7.1), $\mathbf{x}^{cv}$ and $\mathbf{x}^{cc}$, on $I \times P$.

Next, note that $\mathbf{x}^{cv,0}(t,\cdot)$ and $\mathbf{x}^{cc,0}(t,\cdot)$ are trivially convex and concave relaxations of $\mathbf{x}(t,\cdot)$ on $P$, respectively, for each fixed $t \in I$. Suppose that the same is true of $\mathbf{x}^{cv,k}$ and $\mathbf{x}^{cc,k}$. Then, choosing any $\mathbf{p} \in P$, we may apply Definition 7.3.1 with $(\mathbf{z}^{cv}, \mathbf{z}^{cc}) \equiv (\mathbf{x}^{cv,k}, \mathbf{x}^{cc,k})$ to conclude that

$$\mathbf{u}(t,\mathbf{p},\mathbf{x}^{cv,k}(t,\mathbf{p}),\mathbf{x}^{cc,k}(t,\mathbf{p})) \leq \dot{\mathbf{x}}(t,\mathbf{p}) \leq \mathbf{o}(t,\mathbf{p},\mathbf{x}^{cv,k}(t,\mathbf{p}),\mathbf{x}^{cc,k}(t,\mathbf{p})),$$

for a.e. $t \in I$. Combining this with integral monotonicity,

$$\begin{aligned}\int_{t_0}^{t} \mathbf{u}(s,\mathbf{p},\mathbf{x}^{cv,k}(s,\mathbf{p}),\mathbf{x}^{cc,k}(s,\mathbf{p}))ds \;&\leq\; \int_{t_0}^{t} \dot{\mathbf{x}}(s,\mathbf{p})ds,\\ &\leq\; \int_{t_0}^{t} \mathbf{o}(s,\mathbf{p},\mathbf{x}^{cv,k}(s,\mathbf{p}),\mathbf{x}^{cc,k}(s,\mathbf{p}))ds,\end{aligned}$$

for all $(t,\mathbf{p}) \in I \times P$. But since $\mathbf{x}_0^{cv}(\mathbf{p}) \leq \mathbf{x}(t_0,\mathbf{p}) \leq \mathbf{x}_0^{cc}(\mathbf{p})$ for all $\mathbf{p} \in P$, (7.7) shows that

$$\mathbf{x}^{cv,k+1}(t,\mathbf{p}) \leq \mathbf{x}(t_0,\mathbf{p}) + \int_{t_0}^{t} \dot{\mathbf{x}}(s,\mathbf{p})ds \leq \mathbf{x}^{cc,k+1}(t,\mathbf{p}), \quad \forall (t,\mathbf{p}) \in I \times P,$$

which gives

$$\mathbf{x}^{cv,k+1}(t,\mathbf{p}) \leq \mathbf{x}(t,\mathbf{p}) \leq \mathbf{x}^{cc,k+1}(t,\mathbf{p}), \quad \forall (t,\mathbf{p}) \in I \times P.$$

For every $t \in I$, convexity of $\mathbf{x}^{cv,k}(t,\cdot)$ on $P$ implies that it is consistent with convexity at every $(\lambda,\mathbf{p}_1,\mathbf{p}_2) \in (0,1) \times P \times P$, and the analogous observation holds for $\mathbf{x}^{cc,k}(t,\cdot)$. Then, choosing any $(\lambda,\mathbf{p}_1,\mathbf{p}_2) \in (0,1) \times P \times P$, Conditions 1 and 2 of Definition 7.3.3 hold with $(\mathbf{z}^{cv}, \mathbf{z}^{cc}) \equiv (\mathbf{x}^{cv,k}, \mathbf{x}^{cc,k})$. Moreover, the fact $\mathbf{x}^{cv,k}(t,\mathbf{p}) \leq \mathbf{x}(t,\mathbf{p}) \leq \mathbf{x}^{cc,k}(t,\mathbf{p})$, $\forall (t,\mathbf{p}) \in I \times P$, implies that Condition 3 holds as well. Then, Definition 7.3.3 implies that

$$\begin{aligned}\mathbf{u}(t,\hat{\mathbf{p}},\mathbf{x}^{cv,k}(t,\hat{\mathbf{p}}),\mathbf{x}^{cc,k}(t,\hat{\mathbf{p}})) \leq &\lambda\mathbf{u}(t,\mathbf{p}_1,\mathbf{x}^{cv,k}(t,\mathbf{p}_1),\mathbf{x}^{cc,k}(t,\mathbf{p}_1))\\ &+ (1-\lambda)\mathbf{u}(t,\mathbf{p}_2,\mathbf{x}^{cv,k}(t,\mathbf{p}_2),\mathbf{x}^{cc,k}(t,\mathbf{p}_2)),\end{aligned}$$

for a.e. $t \in I$, where $\hat{\mathbf{p}} \equiv \lambda\mathbf{p}_1 + (1-\lambda)\mathbf{p}_2$. By monotonicity and linearity of the integral,

$$\begin{aligned}\int_{t_0}^{t}\mathbf{u}(s,\hat{\mathbf{p}},\mathbf{x}^{cv,k}(s,\hat{\mathbf{p}}),\mathbf{x}^{cc,k}(s,\hat{\mathbf{p}}))ds \leq&\lambda\int_{t_0}^{t}\mathbf{u}(s,\mathbf{p}_1,\mathbf{x}^{cv,k}(s,\mathbf{p}_1),\mathbf{x}^{cc,k}(s,\mathbf{p}_1))ds\\ &+(1-\lambda)\int_{t_0}^{t}\mathbf{u}(s,\mathbf{p}_2,\mathbf{x}^{cv,k}(s,\mathbf{p}_2),\mathbf{x}^{cc,k}(s,\mathbf{p}_2))ds,\end{aligned}$$

for every $t \in I$. Making the analogous concavity arguments for $\mathbf{o}$ and noting that the conditions of Definition 7.3.3 hold for all $(\lambda, \mathbf{p}_1, \mathbf{p}_2) \in (0,1) \times P \times P$, this implies that

$$\int_{t_0}^{t}\mathbf{u}(s,\cdot,\mathbf{x}^{cv,k}(s,\cdot),\mathbf{x}^{cc,k}(s,\cdot))ds \quad \text{and} \quad \int_{t_0}^{t}\mathbf{o}(s,\cdot,\mathbf{x}^{cv,k}(s,\cdot),\mathbf{x}^{cc,k}(s,\cdot))ds$$

are, respectively, convex and concave on $P$, for every fixed $t \in I$. Since $\mathbf{x}_0^{cv}$ and $\mathbf{x}_0^{cc}$ are respectively convex and concave by hypothesis, (7.7) shows that $\mathbf{x}^{cv,k+1}$ and $\mathbf{x}^{cc,k+1}$ are, respectively, convex and concave on $P$ for every fixed $t \in I$. Therefore, by induction, $\mathbf{x}^{cv,k}(t,\cdot)$ and $\mathbf{x}^{cc,k}(t,\cdot)$ are, respectively, convex and concave relaxations of $\mathbf{x}(t,\cdot)$ on $P$, for each fixed $t \in I$ and every $k \in \mathbb{N}$.

It was shown above that, as $k \to \infty$, $\mathbf{x}^{cv,k}$ and $\mathbf{x}^{cc,k}$ converge uniformly to the unique solutions of (7.1) on $I \times P$. Then, taking limits, it is clear that $\mathbf{x}^{cv}(t,\cdot)$ and $\mathbf{x}^{cc}(t,\cdot)$ are, respectively, convex and concave relaxations of $\mathbf{x}(t,\cdot)$ on $P$, for each fixed $t \in I$.

## 7.4 Relaxation Preserving Dynamics

In this section, a second set of sufficient conditions on $(\mathbf{u}, \mathbf{o})$ is developed. Through a conceptual discussion, it is shown that the requirement of relaxation amplifying dynamics (Definitions 7.3.1 and 7.3.3) imply two very undesirable properties of $(\mathbf{x}^{cv}, \mathbf{x}^{cc})$. We use these observations to motivate the weaker requirements of *relaxation preserving dynamics*, which potentially describe much tighter relaxations.

Suppose that $(\mathbf{u}, \mathbf{o})$ describe relaxation amplifying dynamics for $\mathbf{x}$ on $I \times P$ and let $\mathbf{x}^{cv}, \mathbf{x}^{cc} : I \times P \to \mathbb{R}^{n_x}$ be solutions of (7.1). Consider again Definition 7.3.1 and suppose that $\mathbf{x}^{cv}(t,\mathbf{p}) \leq \mathbf{x}(t,\mathbf{p}) \leq \mathbf{x}^{cc}(t,\mathbf{p})$, $\forall(t,\mathbf{p}) \in I \times P$, as desired. Then

Definition 7.3.1 implies that (7.2) holds for a.e. $\hat{t} \in I$, and it follows that, for example, the difference $\mathbf{x}(\cdot, \mathbf{p}) - \mathbf{x}^{cv}(\cdot, \mathbf{p})$ is non-decreasing on $I$, for every $\mathbf{p} \in P$ (See Theorem 3.3.3). In other words, once a certain level of conservatism in the underestimator $\mathbf{x}^{cv}(\cdot, \mathbf{p})$ has been established, it can only be amplified as $t$ increases (hence the name). Clearly, an analogous argument holds for the upper bound $\mathbf{x}^{cc}$.

Similarly, suppose that $\mathbf{x}^{cv}(t, \cdot)$ and $\mathbf{x}^{cc}(t, \cdot)$ are, respectively, convex and concave on $P$ for all $t \in I$, as desired. Then, choosing any $(\lambda, \mathbf{p}_1, \mathbf{p}_2) \in (0, 1) \times P \times P$ and letting $\bar{\mathbf{p}} \equiv \lambda \mathbf{p}_1 + (1 - \lambda)\mathbf{p}_2$, Definition 7.3.3 implies that (7.3) holds for all $\hat{t} \in I$. Again, it follows that the difference

$$[\lambda \mathbf{x}^{cv}(\cdot, \mathbf{p}_1) + (1 - \lambda)\mathbf{x}^{cv}(\cdot, \mathbf{p}_2)] - \mathbf{x}^{cv}(\cdot, \bar{\mathbf{p}})$$

is a non-decreasing on $I$; i.e., $\mathbf{x}^{cv}(t, \cdot)$ becomes in a sense *more* convex as $t$ increases. Similarly, $\mathbf{x}^{cc}(t, \cdot)$ becomes *more* concave with increasing $t$, and these trends are quite regardless of the parametric behavior of $\mathbf{x}$.

These observations motivate a more conservative set of conditions on $(\mathbf{u}, \mathbf{o})$. Consider, for example, the upper bounding property of $\mathbf{x}^{cc}$. Suppose that, for some $(\hat{t}, \mathbf{p}) \in I \times P$ and some $i \in \{1, \ldots, n_x\}$, it happens that $x_i(\hat{t}, \mathbf{p}) < x_i^{cc}(\hat{t}, \mathbf{p})$. Then, regardless of the values of $\dot{x}_i(\hat{t}, \mathbf{p})$ and $\dot{x}_i^{cc}(\hat{t}, \mathbf{p})$, continuity ensures that $\exists \delta > 0$ such that $x_i(t, \mathbf{p}) \leq x_i^{cc}(t, \mathbf{p})$, $\forall t \in [\hat{t}, \hat{t} + \delta]$. Thus, there is no reason to require that $\dot{x}_i(\hat{t}, \mathbf{p}) \leq \dot{x}_i^{cc}(\hat{t}, \mathbf{p})$, because $x_i(\cdot, \mathbf{p})$ and $x_i^{cc}(\cdot, \mathbf{p})$ are not in danger of crossing immediately to the right of $\hat{t}$. This suggests that it should only be necessary to require that $\dot{x}_i(\hat{t}, \mathbf{p}) \leq \dot{x}_i^{cc}(\hat{t}, \mathbf{p})$ in the situation where $x_i(\hat{t}, \mathbf{p}) = x_i^{cc}(\hat{t}, \mathbf{p})$ which leads to the notion of *bound preserving dynamics.*

**Definition 7.4.1.** The functions $(\mathbf{u}, \mathbf{o})$ describe *bound preserving dynamics* for $\mathbf{x}$ on $I \times P$ if, for arbitrary functions $\mathbf{z}^{cv}, \mathbf{z}^{cc} : I \times P \to \mathbb{R}^{n_x}$, every $\mathbf{p} \in P$ and every $i \in \{1, \ldots, n_x\}$, the following conditions hold:

1. For a.e. $t \in I$ such that $\mathbf{z}^{cv}(t, \mathbf{p}), \mathbf{z}^{cc}(t, \mathbf{p}) \in X(t)$, $\mathbf{z}^{cv}(t, \mathbf{p}) \leq \mathbf{z}^{cc}(t, \mathbf{p})$ and

$z_i^{cv}(t,\mathbf{p}) = z_i^{cc}(t,\mathbf{p})$, $u_i$ and $o_i$ satisfy

$$u_i(t,\mathbf{p},\mathbf{z}^{cv}(t,\mathbf{p}),\mathbf{z}^{cc}(t,\mathbf{p})) \leq o_i(t,\mathbf{p},\mathbf{z}^{cv}(t,\mathbf{p}),\mathbf{z}^{cc}(t,\mathbf{p})). \tag{7.8}$$

2. For a.e. $t \in I$ such that $\mathbf{z}^{cv}(t,\mathbf{p}), \mathbf{z}^{cc}(t,\mathbf{p}) \in X(t)$, $\mathbf{z}^{cv}(t,\mathbf{p}) \leq \mathbf{x}(t,\mathbf{p}) \leq \mathbf{z}^{cc}(t,\mathbf{p})$ and $x_i(t,\mathbf{p}) = z_i^{cv}(t,\mathbf{p})$, $u_i$ satisfies $u_i(t,\mathbf{p},\mathbf{z}^{cv}(t,\mathbf{p}),\mathbf{z}^{cc}(t,\mathbf{p})) \leq \dot{x}_i(t,\mathbf{p})$.

3. For a.e. $t \in I$ such that $\mathbf{z}^{cv}(t,\mathbf{p}), \mathbf{z}^{cc}(t,\mathbf{p}) \in X(t)$, $\mathbf{z}^{cv}(t,\mathbf{p}) \leq \mathbf{x}(t,\mathbf{p}) \leq \mathbf{z}^{cc}(t,\mathbf{p})$ and $x_i(t,\mathbf{p}) = z_i^{cc}(t,\mathbf{p})$, $o_i$ satisfies $o_i(t,\mathbf{p},\mathbf{z}^{cv}(t,\mathbf{p}),\mathbf{z}^{cc}(t,\mathbf{p})) \geq \dot{x}_i(t,\mathbf{p})$.

The intuitive principle behind this definition has its roots in viability theory and the study of differential inequalities [13, 182]. Indeed, this idea was used extensively in the state bounding results of Chapter 3. A very interesting result of the development here is that a similar observation can be used to weaken significantly the requirements of convexity amplifying dynamics. To see this, choose any $(\lambda, \mathbf{p}_1, \mathbf{p}_2) \in (0,1) \times P \times P$, let $\bar{\mathbf{p}} \equiv \lambda\mathbf{p}_1 + (1-\lambda)\mathbf{p}_2$, and suppose that, for some $\hat{t} \in I$ and some $i \in \{1,\ldots,n_x\}$, it happens that

$$x_i^{cv}(\hat{t},\bar{\mathbf{p}}) < \lambda x_i^{cv}(\hat{t},\mathbf{p}_1) + (1-\lambda)x_i^{cv}(\hat{t},\mathbf{p}_2). \tag{7.9}$$

Then, again, there is no need to require that

$$\dot{x}_i^{cv}(\hat{t},\bar{\mathbf{p}}) \leq \lambda \dot{x}_i^{cv}(\hat{t},\mathbf{p}_1) + (1-\lambda)\dot{x}_i^{cv}(\hat{t},\mathbf{p}_2), \tag{7.10}$$

since mere continuity ensures that $\exists\delta > 0$ with

$$x_i^{cv}(t,\bar{\mathbf{p}}) \leq \lambda x_i^{cv}(t,\mathbf{p}_1) + (1-\lambda)x_i^{cv}(t,\mathbf{p}_2), \quad \forall t \in [\hat{t},\hat{t}+\delta]. \tag{7.11}$$

This suggest that (7.10) need only hold in the case where

$$x_i^{cv}(\hat{t},\bar{\mathbf{p}}) = \lambda x_i^{cv}(\hat{t},\mathbf{p}_1) + (1-\lambda)x_i^{cv}(\hat{t},\mathbf{p}_2). \tag{7.12}$$

**Definition 7.4.2.** The functions $(\mathbf{u}, \mathbf{o})$ describe *convexity preserving dynamics* for $\mathbf{x}$ on $I \times P$ if, for arbitrary functions $\mathbf{z}^{cv}, \mathbf{z}^{cc} : I \times P \to \mathbb{R}^{n_x}$ and every $(\lambda, \mathbf{p}_1, \mathbf{p}_2) \in (0,1) \times P \times P$, the following condition holds: For every $i \in \{1, \ldots, n_x\}$ and a.e. $t \in I$ such that

1. $\mathbf{z}^{cv}(t, \cdot)$ is consistent with convexity at $(\lambda, \mathbf{p}_1, \mathbf{p}_2)$,
2. $\mathbf{z}^{cc}(t, \cdot)$ is consistent with concavity at $(\lambda, \mathbf{p}_1, \mathbf{p}_2)$,
3. $\mathbf{z}^{cv}(t, \mathbf{q}) \leq \mathbf{x}(t, \mathbf{q}) \leq \mathbf{z}^{cc}(t, \mathbf{q})$ and $\mathbf{z}^{cv}(t, \mathbf{q}), \mathbf{z}^{cc}(t, \mathbf{q}) \in X(t)$, $\forall \mathbf{q} \in \{\mathbf{p}_1, \mathbf{p}_2, \lambda\mathbf{p}_1 + (1-\lambda)\mathbf{p}_2\}$,

the functions $\mathbf{u}$ and $\mathbf{o}$ satisfy

1. If $z_i^{cv}(t, \bar{\mathbf{p}}) = \lambda z_i^{cv}(t, \mathbf{p}_1) + (1-\lambda) z_i^{cv}(t, \mathbf{p}_2)$, then the composite function $P \ni \mathbf{p} \mapsto u_i(t, \mathbf{p}, \mathbf{z}^{cv}(t, \mathbf{p}), \mathbf{z}^{cc}(t, \mathbf{p}))$ is consistent with convexity at $(\lambda, \mathbf{p}_1, \mathbf{p}_2)$,
2. If $z_i^{cc}(t, \bar{\mathbf{p}}) = \lambda z_i^{cc}(t, \mathbf{p}_1) + (1-\lambda) z_i^{cc}(t, \mathbf{p}_2)$, then the composite function $P \ni \mathbf{p} \mapsto o_i(t, \mathbf{p}, \mathbf{z}^{cv}(t, \mathbf{p}), \mathbf{z}^{cc}(t, \mathbf{p}))$ is consistent with concavity at $(\lambda, \mathbf{p}_1, \mathbf{p}_2)$.

If $(\mathbf{u}, \mathbf{o})$ describe both bound preserving dynamics and convexity preserving dynamics, then it will be said that they describe *relaxation preserving dynamics*. The main result of this chapter is the proof that, if $(\mathbf{u}, \mathbf{o})$ describe relaxation preserving dynamics for $\mathbf{x}$ on $I \times P$, then state relaxations for $\mathbf{x}$ on $I \times P$ are given by the solutions of a system of ODEs similar to (7.1). This is the subject of the next section. Though these conditions may seem cumbersome, it will be shown that they can be satisfied automatically through a construction based on generalized McCormick relaxations in the case where $\mathbf{x}$ is the solution of a system of ODEs or semi-explicit DAEs. In fact, this construction requires only a minor modification of the procedure used to construct functions satisfying relaxation amplifying dynamics. Thus, this results in a second method for computing state relaxations. In §7.7, it is shown that in the case of ODEs these relaxations offer significant improvements over relaxations derived by existing methods, as well as those derived through relaxation amplifying dynamics.

## 7.5 Sufficiency of Relaxation Preserving Dynamics

Even under the assumption that $(\mathbf{u}, \mathbf{o})$ describe relaxation preserving dynamics for $\mathbf{x}$ on $I \times P$, the solutions of (7.1) will not necessarily be state relaxations for $\mathbf{x}$ on $I \times P$. The reason is that the required properties of $(\mathbf{x}^{cv}, \mathbf{x}^{cc})$ only follow from the properties of $(\mathbf{u}, \mathbf{o})$ provided that $\mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p}) \in X(t)$, $\forall (t, \mathbf{p}) \in I \times P$ (Note the role of $X(t)$ in Definitions 7.4.1 and 7.4.2). Unfortunately, this inclusion is not guaranteed by (7.1).

Accordingly, it is necessary to modify (7.1). In doing this, it is assumed that the state bounds $\mathbf{x}^L$ and $\mathbf{x}^U$ are absolutely continuous functions. Recall that an absolutely continuous function is differentiable almost everywhere, so that $\dot{\mathbf{x}}^L(t)$ and $\dot{\mathbf{x}}^U(t)$ are well defined for a.e. $t \in I$. For the state bounding methods of Chapter 3, $\mathbf{x}^L$ and $\mathbf{x}^U$ are themselves given by the solution of an auxiliary system of ODEs, so that absolute continuity of these functions follows directly. Now, consider the auxiliary system of ODEs described by

$$\dot{x}_i^{cv}(t, \mathbf{p}) = \begin{cases} u_i(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p})) & \text{if} \quad x_i^{cv}(t, \mathbf{p}) \in [x_i^L(t), x_i^U(t)] \\ \max(\dot{x}_i^L(t), u_i(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p}))) & \text{if} \quad x_i^{cv}(t, \mathbf{p}) < x_i^L(t) \\ \min(\dot{x}_i^U(t), u_i(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p}))) & \text{if} \quad x_i^{cv}(t, \mathbf{p}) > x_i^U(t) \end{cases},$$

$$(7.13)$$

$$\dot{x}_i^{cc}(t, \mathbf{p}) = \begin{cases} o_i(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p})) & \text{if} \quad x_i^{cc}(t, \mathbf{p}) \in [x_i^L(t), x_i^U(t)] \\ \max(\dot{x}_i^L(t), o_i(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p}))) & \text{if} \quad x_i^{cc}(t, \mathbf{p}) < x_i^L(t) \\ \min(\dot{x}_i^U(t), o_i(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p}))) & \text{if} \quad x_i^{cc}(t, \mathbf{p}) > x_i^U(t) \end{cases},$$

$$x_i^{cv}(t_0, \mathbf{p}) = \max(x_i^L(t_0), x_{0,i}^{cv}(\mathbf{p})), \quad x_i^{cc}(t_0, \mathbf{p}) = \min(x_i^U(t_0), x_{0,i}^{cc}(\mathbf{p})),$$

for each $i = 1, \ldots, n_x$. It is shown in Lemma 7.5.3 below that the solutions of this system satisfy $\mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p}) \in X(t)$, $\forall (t, \mathbf{p}) \in I \times P$. Having established this, the fact that $(\mathbf{x}^{cv}, \mathbf{x}^{cc})$ are state relaxations for $\mathbf{x}$ on $I \times P$ is derived from the fact that $(\mathbf{u}, \mathbf{o})$ satisfy relaxation preserving dynamics for $\mathbf{x}$ on $I \times P$ in Sections 7.5.1 and 7.5.2.

To begin, it is necessary to define precisely what constitutes a solution of (7.13). We follow the classical definition of a solution for a system of ODEs with discontinuous right-hand sides from Chapter 2, §4 in [62], which in the case of (7.13) reduces to the following:

**Definition 7.5.1.** Two functions $\mathbf{x}^{cv}, \mathbf{x}^{cc} : I \times P \to \mathbb{R}^{n_x}$ are solutions of (7.13) on $I \times P$ if, for each $i$ and every $\mathbf{p} \in P$, $x_i^{cv}(\cdot, \mathbf{p})$ and $x_i^{cc}(\cdot, \mathbf{p})$ are absolutely continuous on $I$, the initial conditions $x_i^{cv}(t_0, \mathbf{p}) = \max(x_i^L(t_0), x_{0,i}^{cv}(\mathbf{p}))$ and $x_i^{cc}(t_0, \mathbf{p}) = \min(x_i^U(t_0), x_{0,i}^{cc}(\mathbf{p}))$ are satisfied, and, for a.e. $t \in I$, $\dot{x}_i^{cv}(t, \mathbf{p})$ satisfies (7.13) if $x_i^{cv}(t, \mathbf{p}) \neq x_i^L(t)$ and $x_i^{cv}(t, \mathbf{p}) \neq x_i^U(t)$, and otherwise satisfies

$$\begin{aligned}
\dot{x}_i^{cv}(t, \mathbf{p}) &\in [u_i(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p})), \max(\dot{x}_i^L(t), u_i(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p})))] \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{if} \quad x_i^{cv}(t, \mathbf{p}) = x_i^L(t), \\
\dot{x}_i^{cv}(t, \mathbf{p}) &\in [\min(\dot{x}_i^U(t), u_i(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p}))), u_i(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p}))] \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{if} \quad x_i^{cv}(t, \mathbf{p}) = x_i^U(t),
\end{aligned}$$

and $\dot{x}_i^{cc}(t, \mathbf{p})$ satisfies (7.13) if $x_i^{cc}(t, \mathbf{p}) \neq x_i^L(t)$ and $x_i^{cc}(t, \mathbf{p}) \neq x_i^U(t)$, and otherwise satisfies

$$\begin{aligned}
\dot{x}_i^{cc}(t, \mathbf{p}) &\in [o_i(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p})), \max(\dot{x}_i^L(t), o_i(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p})))] \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{if} \quad x_i^{cc}(t, \mathbf{p}) = x_i^L(t), \\
\dot{x}_i^{cc}(t, \mathbf{p}) &\in [\min(\dot{x}_i^U(t), o_i(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p}))), o_i(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p}))] \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{if} \quad x_i^{cc}(t, \mathbf{p}) = x_i^U(t).
\end{aligned}$$

**Remark 7.5.2.** The definition of a solution above is weak in the sense that it permits $\dot{x}_i^{cv}(t, \mathbf{p})$ and $\dot{x}_i^{cc}(t, \mathbf{p})$ to take a range of values whenever the solutions lie on potential points of discontinuity of the right-hand side functions in (7.13); i.e. $x_i^{cv}(t, \mathbf{p}) = x_i^L(t)$, $x_i^{cv}(t, \mathbf{p}) = x_i^U(t)$, $x_i^{cc}(t, \mathbf{p}) = x_i^L(t)$ or $x_i^{cc}(t, \mathbf{p}) = x_i^U(t)$. The advantage of this definition is that local existence and uniqueness of a solution of (7.13) in this sense follows from the classical results in [62] (see Theorem 1 in Chapter 2, §7.2, and Theorem 1 in Chapter 2, §10). On the other hand, this generality does not impede the arguments establishing that $(\mathbf{x}^{cv}, \mathbf{x}^{cc})$ are state relaxations for $\mathbf{x}$ on $I \times P$. In fact, it will be shown in the course of these developments that $(\mathbf{x}^{cv}, \mathbf{x}^{cc})$ actually satisfy a much simpler set of conditions, which enable $(\mathbf{x}^{cv}, \mathbf{x}^{cc})$ to be approximated using numerical integration with event detection (See Lemma 7.5.6).

It is now shown that the solutions of (7.13) satisfy $\mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p}) \in X(t)$, $\forall (t, \mathbf{p}) \in I \times P$.

**Lemma 7.5.3.** *Let* $(\mathbf{x}^{cv}, \mathbf{x}^{cc})$ *be a solution of* (7.13) *on* $I \times P$. *Then* $\mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p}) \in X(t)$, $\forall (t, \mathbf{p}) \in I \times P$.

*Proof.* Suppose there exists $\mathbf{p} \in P$, $i \in \{1, \ldots, n_x\}$ and $\hat{t} \in I$ for which $x_i^{cv}(\hat{t}, \mathbf{p}) < x_i^L(\hat{t})$. We show a contradiction (the proof is analogous if $x_i^{cv}(\hat{t}, \mathbf{p}) > x_i^U(\hat{t})$ or $x_i^{cc}(\hat{t}, \mathbf{p}) \notin [x_i^L(\hat{t}), x_i^U(\hat{t})]$). Let $t_1 \equiv \sup\{s \in [t_0, \hat{t}] : x_i^{cv}(s, \mathbf{p}) \geq x_i^L(s)\}$. Since $x_i^{cv}(t_0, \mathbf{p}) \geq x_i^L(t_0)$, this set is nonempty. Because $t_1$ is an upper bound, we have $x_i^{cv}(t, \mathbf{p}) < x_i^L(t)$ for all $t \in [t_0, \hat{t}]$ with $t > t_1$. Because $t_1$ is the least upper bound, we must have $x_i^{cv}(t, \mathbf{p}) \geq x_i^L(t)$ immediately to the left of $t_1$. By continuity, this implies that $x_i^{cv}(t_1, \mathbf{p}) = x_i^L(t_1)$, and hence $t_1 \in [t_0, \hat{t})$.

Then, for a.e. $t \in [t_1, \hat{t}]$, Definition 7.5.1 implies that

$$\dot{x}_i^{cv}(t, \mathbf{p}) = \max(\dot{x}_i^L(t), u_i(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p}))) \geq \dot{x}_i^L(t).$$

Applying Theorem 3.3.3, $(x_i^L - x_i^{cv}(\cdot, \mathbf{p}))$ is non-increasing on $[t_1, \hat{t}]$, and hence $0 = x_i^L(t_1) - x_i^{cv}(t_1, \mathbf{p}) \geq x_i^L(\hat{t}) - x_i^{cv}(\hat{t}, \mathbf{p})$. But then $x_i^{cv}(\hat{t}, \mathbf{p}) \geq x_i^L(\hat{t})$, which is a contradiction. □

### 7.5.1 Bounding properties

Under the assumption that $(\mathbf{u}, \mathbf{o})$ satisfy bound amplifying dynamics for $\mathbf{x}$ on $I \times P$, the fact that the solutions of (7.1) bound $\mathbf{x}$ on $I \times P$ is in essence a consequence of integral monotonicity. On the other hand, establishing this result for the solutions of (7.13), under the weaker assumption that $(\mathbf{u}, \mathbf{o})$ satisfy bound preserving dynamics for

$\mathbf{x}$ on $I \times P$, requires much more sophisticated arguments using differential inequalities [182]. We first establish that $\mathbf{x}^{cv}(t,\mathbf{p}) \leq \mathbf{x}^{cc}(t,\mathbf{p})$, $\forall (t,\mathbf{p}) \in I \times P$.

**Lemma 7.5.4.** *Let $(\mathbf{x}^{cv}, \mathbf{x}^{cc})$ be a solution of (7.13) on $I \times P$ and suppose that Assumption 7.2.3 holds. If $(\mathbf{u}, \mathbf{o})$ describe bound preserving dynamics, then $\mathbf{x}^{cv}(t,\mathbf{p}) \leq \mathbf{x}^{cc}(t,\mathbf{p})$, $\forall (t,\mathbf{p}) \in I \times P$.*

*Proof.* Choose any $\mathbf{p} \in P$ and suppose that $x_j^{cv}(\hat{t},\mathbf{p}) > x_j^{cc}(\hat{t},\mathbf{p})$ for some $\hat{t} \in I$ and some $j \in \{1,\ldots,n_x\}$. It will be shown that this results in a contradiction.

Define $\boldsymbol{\delta} : I \to \mathbb{R}^{n_x}$ by $\boldsymbol{\delta}(t) \equiv \mathbf{x}^{cv}(t,\mathbf{p}) - \mathbf{x}^{cc}(t,\mathbf{p})$, $\forall t \in I$. By hypothesis, $\delta_j(\hat{t}) > 0$ for at least one $j$, and $\boldsymbol{\delta}(t_0) \leq \mathbf{0}$ since

$$\mathbf{x}^{cv}(t_0,\mathbf{p}) = \max(\mathbf{x}^L(t_0), \mathbf{x}_0^{cv}(\mathbf{p})) \leq \mathbf{x}_0(\mathbf{p}) \leq \min(\mathbf{x}^U(t), \mathbf{x}_0^{cc}(\mathbf{p})) = \mathbf{x}^{cc}(t_0,\mathbf{p}).$$

Then, the hypotheses of Lemma 3.3.5 are satisfied. Define $t_1$ as in that lemma. Let $L \in \mathbb{R}^+$ be the Lipschitz constant of Assumption 7.2.3. Applying Lemma 3.3.5 with $t_4 \equiv t_f$, $\beta \equiv 2L$ and arbitrary $\epsilon > 0$ furnishes an index $j \in \{1,\ldots,n_x\}$, a non-decreasing function $\rho \in \mathcal{AC}([t_1,t_f],\mathbb{R})$ satisfying

$$0 < \rho(t), \quad \forall t \in [t_1,t_f], \quad \text{and} \quad \dot{\rho}(t) > 2L\rho(t), \quad \text{a.e. } t \in [t_1,t_f],$$

and numbers $t_2, t_3 \in [t_1,t_f]$ with $t_2 < t_3$ such that the following inequalities hold:

$$\begin{aligned} \mathbf{x}^{cv}(t,\mathbf{p}) &< \mathbf{x}^{cc}(t,\mathbf{p}) + \mathbf{1}\rho(t), \quad \forall t \in [t_2,t_3), \\ x_j^{cc}(t,\mathbf{p}) &< x_j^{cv}(t,\mathbf{p}), \quad \forall t \in (t_2,t_3), \\ x_j^{cv}(t_3,\mathbf{p}) &= x_j^{cc}(t_3,\mathbf{p}) + \rho(t_3), \\ x_j^{cv}(t_2,\mathbf{p}) &= x_j^{cc}(t_2,\mathbf{p}). \end{aligned} \tag{7.14}$$

Define $\mathbf{x}^{cv,\dagger}(t,\mathbf{p}) \equiv \min(\mathbf{x}^{cv}(t,\mathbf{p}), \mathbf{x}^{cc}(t,\mathbf{p}))$, $\forall t \in I$. By Lemma 7.5.3, $\mathbf{x}^{cc}(t,\mathbf{p}) \in X(t)$, $\forall t \in I$, so the second inequality in (7.14) shows that $x_j^{cv}(t,\mathbf{p}) > x_j^L(t)$, for a.e.

$t \in [t_2, t_3]$. Using Definition 7.5.1, Assumption 7.2.3 and the first inequality in (7.14),

$$\begin{aligned}\dot{x}_j^{cv}(t,\mathbf{p}) &\leq u_j(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})),\\ &\leq u_j(t,\mathbf{p},\mathbf{x}^{cv,\dagger}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})) + L\|\mathbf{x}^{cv}(t,\mathbf{p}) - \mathbf{x}^{cv,\dagger}(t,\mathbf{p})\|_\infty\\ &\leq u_j(t,\mathbf{p},\mathbf{x}^{cv,\dagger}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})) + L\rho(t),\end{aligned}$$

for a.e. $t \in [t_2, t_3]$. Next, note that $\mathbf{x}^{cc}(t,\mathbf{p}), \mathbf{x}^{cv,\dagger}(t,\mathbf{p}) \in X(t)$, $\forall t \in I$, by Lemma 7.5.3. Moreover, $x_j^{cv,\dagger}(t) = x_j^{cc}(t,\mathbf{p})$, $\forall t \in [t_2, t_3]$ by the second inequality in (7.14). Since $(\mathbf{u},\mathbf{o})$ describe bound preserving dynamics for $\mathbf{x}$ on $I \times P$, this implies that Condition 1 of Definition 7.4.1 may be applied with $(\mathbf{z}^{cv}, \mathbf{z}^{cc}) \equiv (\mathbf{x}^{cv,\dagger}, \mathbf{x}^{cc})$. This gives

$$\begin{aligned}\dot{x}_j^{cv}(t,\mathbf{p}) &\leq u_j(t,\mathbf{p},\mathbf{x}^{cv,\dagger}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})) + L\rho(t),\\ &\leq o_j(t,\mathbf{p},\mathbf{x}^{cv,\dagger}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})) + L\rho(t),\\ &\leq o_j(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})) + L\rho(t) + L\|\mathbf{x}^{cv,\dagger}(t,\mathbf{p}) - \mathbf{x}^{cv}(t,\mathbf{p}))\|_\infty,\\ &\leq o_j(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})) + 2L\rho(t),\end{aligned}$$

for a.e. $t \in [t_2, t_3]$. Because $x_j^{cv}(t,\mathbf{p}) > x_j^{cc}(t,\mathbf{p})$ for a.e. $t \in [t_2, t_3]$ by (7.14), it follows that $x_j^{cc}(t,\mathbf{p}) < x_j^U(t)$ (Lemma 7.5.3). Therefore, Definition 7.5.1 gives

$$\dot{x}_j^{cv}(t,\mathbf{p}) \leq \dot{x}_j^{cc}(t,\mathbf{p}) + 2L\rho(t) < \dot{x}_j^{cc}(t,\mathbf{p}) + \dot{\rho}(t), \quad \text{for a.e.} \quad t \in [t_2, t_3].$$

By Theorem 3.3.3, this implies that

$$x_j^{cv}(t_3,\mathbf{p}) - x_j^{cc}(t_3,\mathbf{p}) - \rho(t_3) \leq x_j^{cv}(t_2,\mathbf{p}) - x_j^{cc}(t_2,\mathbf{p}) - \rho(t_2).$$

By (7.14) this implies that $-\rho(t_2) \geq 0$, which is a contradiction. □

**Theorem 7.5.5.** *Let $(\mathbf{x}^{cv}, \mathbf{x}^{cc})$ be a solution of (7.13) on $I \times P$ and suppose that Assumption 7.2.3 holds. If $(\mathbf{u},\mathbf{o})$ describe bound preserving dynamics for $\mathbf{x}$ on $I \times P$, then $\mathbf{x}^{cv}(t,\mathbf{p}) \leq \mathbf{x}(t,\mathbf{p}) \leq \mathbf{x}^{cc}(t,\mathbf{p})$, $\forall (t,\mathbf{p}) \in I \times P$.*

*Proof.* Fix any $\mathbf{p} \in P$ and suppose $\exists t \in I$ such that $\mathbf{x}(t, \mathbf{p}) \notin [\mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p})]$. It will be shown that this results in a contradiction.

Noting that $\mathbf{x}^{cv}(t_0, \mathbf{p}) \leq \mathbf{x}(t_0, \mathbf{p}) \leq \mathbf{x}^{cc}(t_0, \mathbf{p})$, define $t_1$ as in Corollary 3.3.6. Let $L$ be the Lipschitz constant of Assumption 7.2.3. Applying Corollary 3.3.6 with $t_4 = t_f$, $\beta = L$ and arbitrary $\epsilon > 0$ gives an index $j \in \{1, \ldots, n_x\}$, a non-decreasing function $\rho \in \mathcal{AC}([t_1, t_f], \mathbb{R})$ satisfying

$$0 < \rho(t), \quad \forall t \in [t_1, t_f], \quad \text{and} \quad \dot{\rho}(t) > L\rho(t), \quad \text{a.e. } t \in [t_1, t_f],$$

and numbers $t_2, t_3 \in [t_1, t_f]$ with $t_2 < t_3$ such that the following inequalities hold:

$$\begin{aligned} \mathbf{x}^{cv}(t, \mathbf{p}) - \mathbf{1}\rho(t) < \mathbf{x}(t, \mathbf{p}) < \mathbf{x}^{cc}(t, \mathbf{p}) + \mathbf{1}\rho(t), \quad \forall t \in [t_2, t_3), \qquad (7.15) \\ x_j^{cc}(t, \mathbf{p}) < x_j(t), \quad \forall t \in (t_2, t_3), \\ x_j(t_3, \mathbf{p}) = x_j^{cc}(t_3, \mathbf{p}) + \rho(t_3), \\ x_j(t_2, \mathbf{p}) = x_j^{cc}(t_2, \mathbf{p}). \end{aligned}$$

Strictly, Corollary 3.3.6 permits an alternative set of inequalities, but the proof is analogous in that case.

For a.e. $t \in [t_2, t_3]$, the fact that $x_j^{cc}(t, \mathbf{p}) < x_j(t, \mathbf{p})$ implies that $x_j^{cc}(t, \mathbf{p}) < x_j^U(t)$. Then, Definition 7.5.1 gives

$$\dot{x}_j^{cc}(t, \mathbf{p}) \geq o_j(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p})), \quad \text{for a.e. } t \in [t_2, t_3]. \qquad (7.16)$$

Define $\tilde{\mathbf{x}}^{cv}(t) = \min(\mathbf{x}(t, \mathbf{p}), \mathbf{x}^{cv}(t, \mathbf{p}))$ and $\tilde{\mathbf{x}}^{cc}(t, \mathbf{p}) = \max(\mathbf{x}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p}))$ for all $t \in I$. Clearly, $\tilde{\mathbf{x}}^{cv}(t, \mathbf{p}) \leq \mathbf{x}(t, \mathbf{p}) \leq \tilde{\mathbf{x}}^{cc}(t, \mathbf{p})$ for all $t \in I$. Lemma 7.5.3 implies that $\tilde{\mathbf{x}}^{cv}(t, \mathbf{p}), \tilde{\mathbf{x}}^{cc}(t, \mathbf{p}) \in X(t)$, $\forall t \in I$. Finally, the second inequality in (7.15) implies that $\tilde{x}_j^{cc}(t, \mathbf{p}) = x_j(t, \mathbf{p})$ for a.e. $t \in [t_2, t_3]$. Then, since $(\mathbf{u}, \mathbf{o})$ describe bound preserving dynamics for $\mathbf{x}$ on $I \times P$, this implies that Condition 3 of Definition 7.4.1 can be applied with $(\mathbf{z}^{cv}, \mathbf{z}^{cc}) \equiv (\tilde{\mathbf{x}}^{cv}, \tilde{\mathbf{x}}^{cc})$ for a.e. $t \in [t_2, t_3]$. This gives

$$o_j(t, \mathbf{p}, \tilde{\mathbf{x}}^{cv}(t, \mathbf{p}), \tilde{\mathbf{x}}^{cc}(t, \mathbf{p})) \geq \dot{x}_j(t, \mathbf{p}), \quad \text{for a.e. } t \in [t_2, t_3]. \tag{7.17}$$

To combine (7.17) and (7.16), note that the first inequality in (7.15) implies that $\|\mathbf{x}^{cv}(t, \mathbf{p}) - \tilde{\mathbf{x}}^{cv}(t, \mathbf{p})\|_\infty + \|\mathbf{x}^{cc}(t, \mathbf{p}) - \tilde{\mathbf{x}}^{cc}(t, \mathbf{p})\|_\infty \leq \rho(t)$ for a.e. $t \in [t_2, t_3]$. Then,

$$\begin{aligned}
\dot{x}_j^{cc}(t, \mathbf{p}) &\geq o_j(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p})), && (7.18)\\
&\geq o_j(t, \mathbf{p}, \tilde{\mathbf{x}}^{cv}(t, \mathbf{p}), \tilde{\mathbf{x}}^{cc}(t, \mathbf{p})) - L\rho(t), && (7.19)\\
&\geq \dot{x}_j(t, \mathbf{p}) - L\rho(t), && (7.20)
\end{aligned}$$

for a.e. $t \in [t_2, t_3]$. Adding $\dot{\rho}(t)$ to both sides and recalling that $\dot{\rho}(t) > L\rho(t)$ for a.e. $t \in [t_2, t_3]$, it follows that

$$\dot{x}_j^{cc}(t, \mathbf{p}) + \dot{\rho}(t) \geq \dot{x}_j(t, \mathbf{p}) - L\rho(t) + \dot{\rho}(t) > \dot{x}_j(t), \tag{7.21}$$

for a.e. $t \in [t_2, t_3]$. By Theorem 3.3.3, this implies that $(x_j^{cc}(t, \mathbf{p}) + \rho(t) - x_j(t, \mathbf{p}))$ is non-decreasing on $[t_2, t_3]$, so that

$$x_j^{cc}(t_3, \mathbf{p}) + \rho(t_3) - x_j(t_3, \mathbf{p}) \geq x_j^{cc}(t_2, \mathbf{p}) + \rho(t_2) - x_j(t_2, \mathbf{p}).$$

But, by (7.15), this implies that $0 \geq \rho(t_2)$, which is a contradiction. □

Based on the results above, it is now possible to show that the solutions of (7.13) actually satisfy a simpler set of conditions than those given in Definition 7.5.1. These conditions show that, for almost every $t \in I$, the functions $\dot{x}_i^{cv}(\cdot, \mathbf{p})$ and $\dot{x}_i^{cc}(\cdot, \mathbf{p})$ take values which are consistent with those generated by simulating (7.13) as a hybrid system with state events, as described in §7.6.3.

**Lemma 7.5.6.** *Let $(\mathbf{x}^{cv}, \mathbf{x}^{cc})$ be a solution of (7.13) on $I \times P$ and fix any $\mathbf{p} \in P$ and any $i \in \{1, \ldots, n_x\}$. If $(\mathbf{u}, \mathbf{o})$ describe bound preserving dynamics for $\mathbf{x}$ on $I \times P$ and*

*Assumption 7.2.3 holds, then the sets*

$$
\begin{aligned}
S_1 &= \{t \in I : x_i^{cv}(t,\mathbf{p}) = x_i^L(t),\ \dot{x}_i^{cv}(t,\mathbf{p}) \neq \max(\dot{x}_i^L(t), u_i(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})))\}\\
S_2 &= \{t \in I : x_i^{cc}(t,\mathbf{p}) = x_i^U(t),\ \dot{x}_i^{cc}(t,\mathbf{p}) \neq \min(\dot{x}_i^U(t), o_i(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})))\}\\
S_3 &= \{t \in I : x_i^{cv}(t,\mathbf{p}) = x_i^U(t),\ \dot{x}_i^U(t) < u_i(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p}))\}\\
S_4 &= \{t \in I : x_i^{cc}(t,\mathbf{p}) = x_i^L(t),\ \dot{x}_i^L(t) > o_i(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p}))\}
\end{aligned}
$$

*have measure zero and hence, for a.e. $t \in I$,*

$$
\begin{aligned}
&\dot{x}_i^{cv}(t,\mathbf{p}) = u_i(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})) && \text{if} \quad x_i^{cv}(t,\mathbf{p}) \neq x_i^L(t), && (7.22)\\
&\dot{x}_i^{cv}(t,\mathbf{p}) = \max(\dot{x}_i^L(t), u_i(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p})\mathbf{x}^{cc}(t,\mathbf{p})) && \text{otherwise,} && (7.23)\\
&\dot{x}_i^{cc}(t,\mathbf{p}) = o_i(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})) && \text{if} \quad x_i^{cc}(t,\mathbf{p}) \neq x_i^U(t), && (7.24)\\
&\dot{x}_i^{cc}(t,\mathbf{p}) = \min(\dot{x}_i^U(t), o_i(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p}))) && \text{otherwise.} && (7.25)
\end{aligned}
$$

*Proof.* Choose any $\mathbf{p} \in P$ and any $i \in \{1,\ldots,n_x\}$. Let $Q = \{t \in I : x_i^{cv}(t,\mathbf{p}) = x_i^L(t)\}$. It will be shown that

$$\dot{x}_i^{cv}(t,\mathbf{p}) = \max(\dot{x}_i^L(t), u_i(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p}))) \tag{7.26}$$

for almost every $t \in Q$, which implies that $S_1$ has measure zero.

For any $t \in Q$, the fact that $x_i^{cv}(s,\mathbf{p}) \geq x_i^L(s)$, $\forall s \in I$, implies that

$$\frac{x_i^L(t) - x_i^L(s)}{t-s} = \frac{x_i^{cv}(t,\mathbf{p}) - x_i^L(s)}{t-s} \leq \frac{x_i^{cv}(t,\mathbf{p}) - x_i^{cv}(s,\mathbf{p})}{t-s}, \quad \forall s \in (t,t_f].$$

Since $x_i^L$ and $x_i^{cv}(\cdot,\mathbf{p})$ are differentiable at a.e. $t \in I$, taking limits above implies that $\dot{x}_i^L(t) \leq \dot{x}_i^{cv}(t,\mathbf{p})$ for a.e. $t \in Q$.

Now, by Definition 7.5.1, the fact that $x_i^L(t) = x_i^{cv}(t,\mathbf{p})$ implies that

$$\dot{x}_i^{cv}(t,\mathbf{p}) \in [u_i(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})), \max(\dot{x}_i^L(t), u_i(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})))] \tag{7.27}$$

for a.e. $t \in Q$. Suppose first that $t \in Q$ satisfies

$$\max(\dot{x}_i^L(t), u_i(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p}))) = \dot{x}_i^L(t). \tag{7.28}$$

Using (7.27),

$$\dot{x}_i^{cv}(t, \mathbf{p}) \leq \max(\dot{x}_i^L(t), u_i(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p}))) = \dot{x}_i^L(t),$$

and, since it was established above that $\dot{x}_i^L(t) \leq \dot{x}_i^{cv}(t, \mathbf{p})$ for a.e. $t \in Q$, it follows that (7.26) holds for a.e. $t \in Q$ satisfying (7.28).

On the other hand, suppose that $t \in Q$ satisfies

$$\max(\dot{x}_i^L(t), u_i(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p}))) = u_i(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p})). \tag{7.29}$$

Then the right-hand side of (7.27) is a singleton and it follows that (7.26) holds for a.e. $t \in Q$ satisfying (7.29). Since either (7.28) or (7.29) holds for every $t \in Q$, (7.26) holds for a.e. $t \in Q$ and hence $S_1$ has measure zero. The proof that $S_2$ has measure zero is analogous.

Next, consider $S_3$. For any $t \in S_3$, it follows from Theorem 7.5.5 that $x_i^{cv}(t, \mathbf{p}) = x_i(t, \mathbf{p}) = x_i^U(t)$. Because $(\mathbf{u}, \mathbf{o})$ describe bound preserving dynamics for $\mathbf{x}$ on $I \times P$, this implies that

$$\dot{x}_i^U(t) < u_i(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p})) \leq \dot{x}_i(t, \mathbf{p}), \quad \text{for a.e.} \quad t \in S_3.$$

On the other hand, for any $t \in S_3$, the fact that $x_i(s, \mathbf{p}) \leq x_i^U(s)$, $\forall s \in I$ implies that

$$\frac{x_i^U(t) - x_i^U(s)}{t - s} = \frac{x_i(t, \mathbf{p}) - x_i^U(s)}{t - s} \geq \frac{x_i(t, \mathbf{p}) - x_i(s, \mathbf{p})}{t - s}, \quad \forall s \in (t, t_f].$$

Since $x_i^U$ and $x_i(\cdot, \mathbf{p})$ are differentiable at a.e. $t \in I$, taking limits above implies that $\dot{x}_i^U(t) \geq \dot{x}_i(t, \mathbf{p})$ for a.e. $t \in S_3$. Thus $S_3$ has measure zero. The proof that $S_4$ has measure zero is analogous. □

### 7.5.2 Convexity/concavity properties

**Theorem 7.5.7.** *Let $(\mathbf{x}^{cv}, \mathbf{x}^{cc})$ be a solution of (7.13) on $I \times P$ and suppose that Assumption 7.2.3 holds. If $(\mathbf{u}, \mathbf{o})$ describe bound preserving dynamics and convexity preserving dynamics for $\mathbf{x}$ on $I \times P$, then $\mathbf{x}^{cv}(t, \cdot)$ and $\mathbf{x}^{cc}(t, \cdot)$ are, respectively, convex and concave on $P$, for every $t \in I$.*

*Proof.* Choose any $\mathbf{p}_1, \mathbf{p}_2 \in P$, any $\lambda \in (0, 1)$, and, for all $t \in I$, define

$$\begin{aligned}\bar{\mathbf{p}} &\equiv \lambda\mathbf{p}_1 + (1-\lambda)\mathbf{p}_2,\\ \bar{\mathbf{x}}^{cv}(t) &\equiv \lambda\mathbf{x}^{cv}(t, \mathbf{p}_1) + (1-\lambda)\mathbf{x}^{cv}(t, \mathbf{p}_2),\\ \bar{\mathbf{x}}^{cc}(t) &\equiv \lambda\mathbf{x}^{cc}(t, \mathbf{p}_1) + (1-\lambda)\mathbf{x}^{cc}(t, \mathbf{p}_2).\end{aligned}$$

To arrive at a contradiction, suppose that there exists $\hat{t} \in I$ such that either $x^{cv}_j(\hat{t}, \bar{\mathbf{p}}) > \bar{x}^{cv}_j(\hat{t})$ or $x^{cc}_j(\hat{t}, \bar{\mathbf{p}}) < \bar{x}^{cc}_j(\hat{t})$, for at least one index $j$. Define $\boldsymbol{\delta} : I \to \mathbb{R}^{2n_x}$ by

$$\boldsymbol{\delta}(t) \equiv \left(\mathbf{x}^{cv}(t, \bar{\mathbf{p}}) - \bar{\mathbf{x}}^{cv}(t), \bar{\mathbf{x}}^{cc}(t) - \mathbf{x}^{cc}(t, \bar{\mathbf{p}})\right), \quad \forall t \in I.$$

Then, $\delta_j(\hat{t}) > 0$ for at least one $j$, and $\boldsymbol{\delta}(t_0) \leq \mathbf{0}$ since $\mathbf{x}^{cv}(t_0, \cdot)$ and $\mathbf{x}^{cc}(t_0, \cdot)$ are convex and concave on $P$, respectively. Then, the hypotheses of Lemma 3.3.5 are satisfied. Define $t_1$ as in that lemma. Let $L \in \mathbb{R}^+$ be the Lipschitz constant of Assumption 7.2.3. Applying Lemma 3.3.5 with $t_4 \equiv t_f$, $\beta \equiv 2L$ and arbitrary $\epsilon > 0$ furnishes an index $j \in \{1, \ldots, n_x\}$, a non-decreasing function $\rho \in \mathcal{AC}([t_1, t_f], \mathbb{R})$ satisfying

$$0 < \rho(t), \quad \forall t \in [t_1, t_f], \quad \text{and} \quad \dot{\rho}(t) > 2L\rho(t), \quad \text{a.e. } t \in [t_1, t_f],$$

and numbers $t_2, t_3 \in [t_1, t_f]$ with $t_2 < t_3$ such that the following inequalities hold:

$$\begin{aligned}
\mathbf{x}^{cv}(t, \bar{\mathbf{p}}) &< \bar{\mathbf{x}}^{cv}(t) + \mathbf{1}\rho(t), \quad \forall t \in [t_2, t_3), \\
\mathbf{x}^{cc}(t, \bar{\mathbf{p}}) &> \bar{\mathbf{x}}^{cc}(t) - \mathbf{1}\rho(t), \quad \forall t \in [t_2, t_3), \\
\bar{x}_j^{cv}(t) &< x_j^{cv}(t, \bar{\mathbf{p}}), \quad \forall t \in (t_2, t_3), \\
x_j^{cv}(t_3, \bar{\mathbf{p}}) &= \bar{x}_j^{cv}(t_3) + \rho(t_3), \\
x_j^{cv}(t_2, \bar{\mathbf{p}}) &= \bar{x}_j^{cv}(t_2).
\end{aligned} \tag{7.30}$$

In fact, Lemma 3.3.5 permits the alternate possibility that the last three lines of (7.30) show analogous inequalities for a violation of the concavity of $x_j^{cc}(t, \cdot)$ on $[t_2, t_3]$ for some $j$; i.e., $x_j^{cc}(t, \bar{\mathbf{p}}) < \bar{x}_j^{cc}(t)$. The proof in this case is analogous and we proceed assuming that (7.30) holds.

Since $\mathbf{x}^{cv}(t, \mathbf{p}_1), \mathbf{x}^{cv}(t, \mathbf{p}_2) \in X(t)$, $\forall t \in I$ (Lemma 7.5.3), the same holds for $\bar{\mathbf{x}}^{cv}(t)$ and the third inequality in (7.30) implies that $x_j^{cv}(t, \bar{\mathbf{p}}) > x_j^L(t)$, $\forall t \in (t_2, t_3)$. Then, by Lemma 7.5.6,

$$\dot{x}_j^{cv}(t, \bar{\mathbf{p}}) = u_j(t, \bar{\mathbf{p}}, \mathbf{x}^{cv}(t, \bar{\mathbf{p}}), \mathbf{x}^{cc}(t, \bar{\mathbf{p}})), \quad \text{for a.e.} \quad t \in [t_2, t_3].$$

Define $\mathbf{x}^{cv,*}(t, \bar{\mathbf{p}}) = \min(\mathbf{x}^{cv}(t, \bar{\mathbf{p}}), \bar{\mathbf{x}}^{cv}(t))$ and $\mathbf{x}^{cc,*}(t, \bar{\mathbf{p}}) = \max(\mathbf{x}^{cc}(t, \bar{\mathbf{p}}), \bar{\mathbf{x}}^{cc}(t))$, $\forall t \in [t_2, t_3]$. Assumption 7.2.3 gives

$$\begin{aligned}
\dot{x}_j^{cv}(t, \bar{\mathbf{p}}) \leq{} & u_j(t, \bar{\mathbf{p}}, \mathbf{x}^{cv,*}(t, \bar{\mathbf{p}}), \mathbf{x}^{cc,*}(t, \bar{\mathbf{p}})) \\
& + L\left(\|\mathbf{x}^{cv}(t, \bar{\mathbf{p}}) - \mathbf{x}^{cv,*}(t, \bar{\mathbf{p}})\|_\infty + \|\mathbf{x}^{cc}(t, \bar{\mathbf{p}}) - \mathbf{x}^{cc,*}(t, \bar{\mathbf{p}})\|_\infty\right),
\end{aligned}$$

for a.e. $t \in [t_2, t_3]$. By the first and second inequalities in (7.30), it follows that

$$\begin{aligned}
\dot{x}_j^{cv}(t, \bar{\mathbf{p}}) &\leq u_j(t, \bar{\mathbf{p}}, \mathbf{x}^{cv,*}(t, \bar{\mathbf{p}}), \mathbf{x}^{cc,*}(t, \bar{\mathbf{p}})) + 2L\rho(t) \\
&< u_j(t, \bar{\mathbf{p}}, \mathbf{x}^{cv,*}(t, \bar{\mathbf{p}}), \mathbf{x}^{cc,*}(t, \bar{\mathbf{p}})) + \dot{\rho}(t), \quad \text{for a.e. } t \in [t_2, t_3].
\end{aligned}$$

Next, we use the fact that $(\mathbf{u}, \mathbf{o})$ describe convexity preserving dynamics for $\mathbf{x}$ on

$I \times P$ to show that

$$\begin{aligned} \dot{x}_j^{cv}(t, \bar{\mathbf{p}}) &< u_j(t, \bar{\mathbf{p}}, \mathbf{x}^{cv,*}(t, \bar{\mathbf{p}}), \mathbf{x}^{cc,*}(t, \bar{\mathbf{p}})) + \dot{\rho}(t) \\ &\leq \lambda u_j(t, \mathbf{p}_1, \mathbf{x}^{cv}(t, \mathbf{p}_1), \mathbf{x}^{cc}(t, \mathbf{p}_1)) \\ &\qquad + (1-\lambda) u_j(t, \mathbf{p}_2, \mathbf{x}^{cv}(t, \mathbf{p}_2), \mathbf{x}^{cc}(t, \mathbf{p}_2)) + \dot{\rho}(t), \end{aligned} \tag{7.31}$$

for a.e. $t \in [t_2, t_3]$. To justify this, note that Theorem 7.5.5 ensures that $\mathbf{x}^{cv}(t, \mathbf{p}_1) \leq \mathbf{x}(t, \mathbf{p}_1) \leq \mathbf{x}^{cc}(t, \mathbf{p}_1)$, $\mathbf{x}^{cv}(t, \mathbf{p}_2) \leq \mathbf{x}(t, \mathbf{p}_2) \leq \mathbf{x}^{cc}(t, \mathbf{p}_2)$ and

$$\mathbf{x}^{cv,*}(t, \bar{\mathbf{p}}) \leq \mathbf{x}^{cv}(t, \bar{\mathbf{p}}) \leq \mathbf{x}(t, \bar{\mathbf{p}}) \leq \mathbf{x}^{cc}(t, \bar{\mathbf{p}}) \leq \mathbf{x}^{cc,*}(t, \bar{\mathbf{p}}), \quad \forall t \in [t_2, t_3].$$

Moreover, noting that $\mathbf{x}^{cv}(t, \mathbf{q}), \mathbf{x}^{cc}(t, \mathbf{q}) \in X(t)$, $\forall \mathbf{q} \in \{\mathbf{p}_1, \mathbf{p}_2, \bar{\mathbf{p}}\}$, by Lemma 7.5.3, we must have $\bar{\mathbf{x}}^{cv}(t), \bar{\mathbf{x}}^{cc}(t) \in X(t)$ since these are convex combinations of elements of $X(t)$, and it follows that $\mathbf{x}^{cv,*}(t, \bar{\mathbf{p}}), \mathbf{x}^{cc,*}(t, \bar{\mathbf{p}}) \in X(t)$. Finally, the definitions of $\mathbf{x}^{cv,*}$ and $\mathbf{x}^{cc,*}$ imply that

$$\begin{aligned} \mathbf{x}^{cv,*}(t, \bar{\mathbf{p}}) &\leq \bar{\mathbf{x}}^{cv}(t) = \lambda \mathbf{x}^{cv}(t, \mathbf{p}_1) + (1-\lambda) \mathbf{x}^{cv}(t, \mathbf{p}_2), \\ \mathbf{x}^{cc,*}(t, \bar{\mathbf{p}}) &\geq \bar{\mathbf{x}}^{cc}(t) = \lambda \mathbf{x}^{cc}(t, \mathbf{p}_1) + (1-\lambda) \mathbf{x}^{cc}(t, \mathbf{p}_2), \end{aligned}$$

for all $t \in [t_2, t_3]$, and the third inequality in (7.30) provides

$$x_j^{cv,*}(t, \bar{\mathbf{p}}) = \bar{x}_j^{cv}(t) = \lambda x_j^{cv}(t, \mathbf{p}_1) + (1-\lambda) x_j^{cv}(t, \mathbf{p}_2), \quad \forall t \in [t_2, t_3].$$

Therefore, (7.31) results from applying Definition 7.4.2 with arbitrary functions $\mathbf{z}^{cv}, \mathbf{z}^{cc} : I \times P \to \mathbb{R}^{n_x}$ satisfying

$$\begin{aligned} \mathbf{z}^{cv}(t, \mathbf{p}_1) = \mathbf{x}^{cv}(t, \mathbf{p}_1), \quad \mathbf{z}^{cv}(t, \mathbf{p}_2) = \mathbf{x}^{cv}(t, \mathbf{p}_2), \quad \mathbf{z}^{cv}(t, \bar{\mathbf{p}}) = \mathbf{x}^{cv,*}(t, \bar{\mathbf{p}}), \\ \mathbf{z}^{cc}(t, \mathbf{p}_1) = \mathbf{x}^{cc}(t, \mathbf{p}_1), \quad \mathbf{z}^{cc}(t, \mathbf{p}_2) = \mathbf{x}^{cc}(t, \mathbf{p}_2), \quad \mathbf{z}^{cc}(t, \bar{\mathbf{p}}) = \mathbf{x}^{cc,*}(t, \bar{\mathbf{p}}), \end{aligned} \tag{7.32}$$

for a.e. $t \in [t_1, t_2]$.

Now, by Lemma 7.5.6,

$$\dot{x}_j^{cv}(t, \mathbf{p}_1) \geq u_j(t, \mathbf{p}_1, \mathbf{x}^{cv}(t, \mathbf{p}_1), \mathbf{x}^{cc}(t, \mathbf{p}_1)),$$
$$\dot{x}_j^{cv}(t, \mathbf{p}_2) \geq u_j(t, \mathbf{p}_2, \mathbf{x}^{cv}(t, \mathbf{p}_2), \mathbf{x}^{cc}(t, \mathbf{p}_2)),$$

for a.e. $t \in [t_2, t_3]$, so that (7.31) gives

$$\dot{x}_j^{cv}(t, \bar{\mathbf{p}}) < \lambda \dot{x}_j^{cv}(t, \mathbf{p}_1) + (1 - \lambda)\dot{x}_j^{cv}(t, \mathbf{p}_2) + \dot{\rho}(t) = \dot{\bar{x}}_j^{cv}(t) + \dot{\rho}(t), \tag{7.33}$$

for a.e. $t \in [t_2, t_3]$. By Theorem 3.3.3, this implies that $x_j^{cv}(t, \bar{\mathbf{p}}) - \bar{x}_j^{cv}(t) - \rho(t)$ is non-increasing on $[t_2, t_3]$, so that

$$x_j^{cv}(t_3, \bar{\mathbf{p}}) - \bar{x}_j^{cv}(t_3) - \rho(t_3) \leq x_j^{cv}(t_2, \bar{\mathbf{p}}) - \bar{x}_j^{cv}(t_2) - \rho(t_2).$$

By the last two equalities in (7.30), this implies that $0 \leq -\rho(t_2)$, which is a contradiction. Therefore,

$$\mathbf{x}^{cv}(t, \lambda\mathbf{p}_1 + (1 - \lambda)\mathbf{p}_2) = \mathbf{x}^{cv}(t, \bar{\mathbf{p}}) \leq \bar{\mathbf{x}}^{cv}(t) = \lambda\mathbf{x}^{cv}(t, \mathbf{p}_1) + (1 - \lambda)\mathbf{x}^{cv}(t, \mathbf{p}_2), \tag{7.34}$$
$$\mathbf{x}^{cc}(t, \lambda\mathbf{p}_1 + (1 - \lambda)\mathbf{p}_2) = \mathbf{x}^{cc}(t, \bar{\mathbf{p}}) \geq \bar{\mathbf{x}}^{cc}(t) = \lambda\mathbf{x}^{cc}(t, \mathbf{p}_1) + (1 - \lambda)\mathbf{x}^{cc}(t, \mathbf{p}_2), \tag{7.35}$$

for all $t \in I$. Since the choice of $\mathbf{p}_1, \mathbf{p}_2 \in P$ and $\lambda \in (0, 1)$ was arbitrary, (7.34) and (7.35) hold for all $\mathbf{p}_1, \mathbf{p}_2 \in P$ and $\lambda \in (0, 1)$. □

## 7.6 State Relaxations for ODEs

In this section, we apply the state relaxation theories of the previous sections to the case where $\mathbf{x}$ is the solution of a system of parametric ODEs. Let $I = [t_0, t_f] \subset \mathbb{R}$ and $P \subset \mathbb{R}^{n_p}$ be compact intervals, let $D \subset \mathbb{R}^{n_x}$ be open, and let $\mathbf{x}_0 : P \to D$ and $\mathbf{f} : I \times P \times D \to \mathbb{R}^{n_x}$ be continuous mappings. Consider the initial value problem

$$\dot{\mathbf{x}}(t, \mathbf{p}) = \mathbf{f}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})), \quad \mathbf{x}(t_0, \mathbf{p}) = \mathbf{x}_0(\mathbf{p}). \tag{7.36}$$

A solution of (7.36) is any continuous $\mathbf{x} : I \times P \to D$ such that, for every $\mathbf{p} \in P$, $\mathbf{x}(\cdot, \mathbf{p})$ is continuously differentiable and satisfies (7.36) on $I$ (with derivatives from the right and left at $t_0$ and $t_f$, respectively).

**Assumption 7.6.1.** For every compact $K \subset D$, $\exists L_K \in \mathbb{R}_+$ such that

$$\|\mathbf{f}(t, \mathbf{p}, \mathbf{z}) - \mathbf{f}(t, \mathbf{p}, \hat{\mathbf{z}})\|_\infty \leq L_K \|\mathbf{z} - \hat{\mathbf{z}}\|_\infty, \quad \forall (t, \mathbf{p}, \mathbf{z}, \hat{\mathbf{z}}) \in I \times P \times K \times K.$$

For any compact $K \subset D$, a function satisfying the inequality of Assumption 7.6.1 is said to be Lipschitz on $K$ uniformly on $I \times P$. Under Assumption 7.6.1, the existence of a unique solution to (7.36) can be ensured locally (by, for example, a straightforward extension of Theorem 3.1 in [91]). In what follows, it will always be assumed that Assumption 7.6.1 holds, and that the unique solution of (7.36) exists on all of $I \times P$.

Since the solution of (7.36) is continuously differentiable on $I$ for every $\mathbf{p} \in P$, it is also absolutely continuous on $I$ for every $\mathbf{p} \in P$. Then, the developments of the previous sections imply that any $(\mathbf{u}, \mathbf{o})$ which describe either relaxation amplifying or relaxation preserving dynamics for $\mathbf{x}$ on $I \times P$ can be used to compute state relaxations of $\mathbf{x}$ on $I \times P$ through the solution of either (7.1) of (7.13). It remains to develop a computational procedure for constructing and evaluating appropriate functions $\mathbf{x}_0^{cv}$, $\mathbf{x}_0^{cc}$, $\mathbf{u}$ and $\mathbf{o}$. This is done here using natural McCormick extensions. The following assumption is required.

**Assumption 7.6.2.** The functions $\mathbf{x}_0$ and $\mathbf{f}$ are $\mathcal{L}$-factorable with natural McCormick extensions $\{\mathbf{x}_0\} : \mathcal{D}_0 \to \mathbb{MR}^{n_x}$ and $\{\mathbf{f}\} : \mathcal{D}_f \to \mathbb{MR}^{n_x}$. Moreover, $P$ is represented in $\mathcal{D}_0$ and $[t, t] \times P \times X(t)$ is represented in $\mathcal{D}_f$ for every $t \in I$.

The initial condition functions $\mathbf{c}_0$ and $\mathbf{C}_0$ can now be constructed and evaluated by computing the standard McCormick relaxations of $\mathbf{x}_0$. Furthermore, these functions will satisfy Assumption 7.2.3 by Corollary 2.6.2.

### 7.6.1 Constructing Relaxation Amplifying Dynamics

Define $\mathbf{u}_f, \mathbf{o}_f : I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ by

$$\begin{aligned} \mathbf{u}_f(t, \mathbf{p}, \mathbf{z}^{cv}, \mathbf{z}^{cc}) &\equiv \{\mathbf{f}\}^{cv}(([t,t],[t,t]),(P,[\mathbf{p},\mathbf{p}]),\mathrm{MC}(\mathbf{x}^L(t),\mathbf{x}^U(t),\mathbf{z}^{cv},\mathbf{z}^{cc})), \\ \mathbf{o}_f(t, \mathbf{p}, \mathbf{z}^{cv}, \mathbf{z}^{cc}) &\equiv \{\mathbf{f}\}^{cc}(([t,t],[t,t]),(P,[\mathbf{p},\mathbf{p}]),\mathrm{MC}(\mathbf{x}^L(t),\mathbf{x}^U(t),\mathbf{z}^{cv},\mathbf{z}^{cc})). \end{aligned} \tag{7.37}$$

For the benefit of the next section, the lemmas below establish properties of $(\mathbf{u}_f, \mathbf{o}_f)$ that are stronger than required to show that they describe relaxation amplifying dynamics.

**Lemma 7.6.3.** *For arbitrary functions $\mathbf{z}^{cv}, \mathbf{z}^{cc} : I \times P \to \mathbb{R}^{n_x}$ and every $\mathbf{p} \in P$, the following conditions hold:*

1. *For a.e. $t \in I$ such that $\mathbf{z}^{cv}(t,\mathbf{p}) \leq \mathbf{z}^{cc}(t,\mathbf{p})$ and $X(t) \cap [\mathbf{z}^{cv}(t,\mathbf{p}), \mathbf{z}^{cc}(t,\mathbf{p})] \neq \emptyset$, $\mathbf{u}_f$ and $\mathbf{o}_f$ satisfy*

$$\mathbf{u}_f(t,\mathbf{p},\mathbf{z}^{cv}(t,\mathbf{p}),\mathbf{z}^{cc}(t,\mathbf{p})) \leq \mathbf{o}_f(t,\mathbf{p},\mathbf{z}^{cv}(t,\mathbf{p}),\mathbf{z}^{cc}(t,\mathbf{p})). \tag{7.38}$$

2. *For a.e. $t \in I$ such that $\mathbf{z}^{cv}(t,\mathbf{p}) \leq \mathbf{x}(t,\mathbf{p}) \leq \mathbf{z}^{cc}(t,\mathbf{p})$, $\mathbf{u}_f$ and $\mathbf{o}_f$ satisfy*

$$\mathbf{u}_f(t,\mathbf{p},\mathbf{z}^{cv}(t,\mathbf{p}),\mathbf{z}^{cc}(t,\mathbf{p})) \leq \dot{\mathbf{x}}(t,\mathbf{p}) \leq \mathbf{o}_f(t,\mathbf{p},\mathbf{z}^{cv}(t,\mathbf{p}),\mathbf{z}^{cc}(t,\mathbf{p})). \tag{7.39}$$

*Proof.* Choose arbitrary functions $\mathbf{z}^{cc}, \mathbf{z}^{cv} : I \times P \to \mathbb{R}^{n_x}$ and any $\mathbf{p} \in P$. For any $t \in I$, (7.38) follows from (7.37) and the fact that $\{\mathbf{f}\}$ takes values in $\mathbb{MR}^{n_x}$. Next, suppose that $t \in I$ is such that $\mathbf{z}^{cv}(t,\mathbf{p}) \leq \mathbf{x}(t,\mathbf{p}) \leq \mathbf{z}^{cc}(t,\mathbf{p})$. Since $[t,t] \times P \times X(t)$ is represented in $\mathcal{D}_f$, Lemma 2.7.3 may be applied with $\mathbf{x} \equiv (t,\mathbf{p},\mathbf{x}(t,\mathbf{p}))$, $\mathbf{x}^{cv} \equiv (t,\mathbf{p},\mathbf{z}^{cv}(t,\mathbf{p}))$ and $\mathbf{x}^{cc} \equiv (t,\mathbf{p},\mathbf{z}^{cc}(t,\mathbf{p}))$ to establish (7.39). □

**Corollary 7.6.4.** *The functions $(\mathbf{u}_f, \mathbf{o}_f)$ describe bound amplifying dynamics for $\mathbf{x}$ on $I \times P$.*

*Proof.* This follows immediately from Conclusion 2 of Lemma 7.6.3. □

**Lemma 7.6.5.** *For arbitrary functions $\mathbf{z}^{cv}, \mathbf{z}^{cc} : I \times P \to \mathbb{R}^{n_x}$ and every $(\lambda, \mathbf{p}_1, \mathbf{p}_2) \in (0, 1) \times P \times P$, the following condition holds: For a.e. $t \in I$ such that*

1. *$\mathbf{z}^{cv}(t, \cdot)$ is consistent with convexity at $(\lambda, \mathbf{p}_1, \mathbf{p}_2)$,*

2. *$\mathbf{z}^{cc}(t, \cdot)$ is consistent with concavity at $(\lambda, \mathbf{p}_1, \mathbf{p}_2)$,*

3. *$\mathbf{z}^{cv}(t, \mathbf{q}) \leq \mathbf{z}^{cc}(t, \mathbf{q})$ and $X(t) \cap [\mathbf{z}^{cv}(t, \mathbf{q}), \mathbf{z}^{cc}(t, \mathbf{q})] \neq \emptyset$, for all $\mathbf{q} \in \{\mathbf{p}_1, \mathbf{p}_2, \lambda\mathbf{p}_1 + (1 - \lambda)\mathbf{p}_2\}$,*

*the functions*

$$P \ni \mathbf{p} \longmapsto \mathbf{u}_f(t, \mathbf{p}, \mathbf{z}^{cv}(t, \mathbf{p}), \mathbf{z}^{cc}(t, \mathbf{p})) \quad \textit{and} \quad P \ni \mathbf{p} \longmapsto \mathbf{o}_f(t, \mathbf{p}, \mathbf{z}^{cv}(t, \mathbf{p}), \mathbf{z}^{cc}(t, \mathbf{p}))$$

*are, respectively, consistent with convexity and consistent with concavity at $(\lambda, \mathbf{p}_1, \mathbf{p}_2)$.*

*Proof.* Choose arbitrary functions $\mathbf{z}^{cc}, \mathbf{z}^{cv} : I \times P \to \mathbb{R}^{n_x}$, let $(\lambda, \mathbf{p}_1, \mathbf{p}_2) \in (0, 1) \times P \times P$, define $\mathbf{p}_3 \equiv \lambda\mathbf{p}_1 + (1 - \lambda)\mathbf{p}_2$, and suppose that $t \in I$ is such that Conditions 1-3 hold. Since $[t, t] \times P \times X(t)$ is represented in $\mathcal{D}_f$, Lemma 2.7.4 may be applied with $\mathbf{x}_i^{cv} \equiv (t, \mathbf{p}_i, \mathbf{z}^{cv}(t, \mathbf{p}_i))$ and $\mathbf{x}_i^{cc} \equiv (t, \mathbf{p}_i, \mathbf{z}^{cc}(t, \mathbf{p}_i))$, $\forall i \in \{1, 2, 3\}$. By (7.37), this gives the desired result. □

**Corollary 7.6.6.** *The functions $(\mathbf{u}_f, \mathbf{o}_f)$ describe convexity amplifying dynamics for $\mathbf{x}$ on $I \times P$.*

*Proof.* Choose arbitrary functions $\mathbf{z}^{cc}, \mathbf{z}^{cv} : I \times P :\to \mathbb{R}^{n_x}$, let $(\lambda, \mathbf{p}_1, \mathbf{p}_2) \in (0, 1) \times P \times P$, define $\mathbf{p}_3 \equiv \lambda\mathbf{p}_1 + (1 - \lambda)\mathbf{p}_2$, and suppose that $t \in I$ is such that Conditions 1-3 of Definition 7.3.3 hold. Since $\mathbf{x}(t, \mathbf{q}) \in X(t)$, $\forall \mathbf{q} \in \{\mathbf{p}_1, \mathbf{p}_2, \lambda\mathbf{p}_1 + (1 - \lambda)\mathbf{p}_2\}$, Conditions 1-3 of Lemma 7.6.5 are verified, and the conclusion follows. □

**Lemma 7.6.7.** *The functions $(\mathbf{u}_f, \mathbf{o}_f)$ satisfy Assumption 7.2.3.*

*Proof.* Consider the set

$$\begin{aligned} K^B \equiv \{(s^L, \mathbf{q}^L, \mathbf{z}^L, s^U, \mathbf{q}^U, \mathbf{z}^U) \in \mathbb{R}^{2(1+n_p+n_x)} : s^L = s^U \in I, \\ [\mathbf{q}^L, \mathbf{q}^U] \subset P,\ [\mathbf{z}^L, \mathbf{z}^U] \subset X(s^L)\}, \end{aligned}$$

and let

$$K \equiv \{(s^L, \mathbf{q}^L, \mathbf{z}^L, s^U, \mathbf{q}^U, \mathbf{z}^U, s^{cv}, \mathbf{q}^{cv}, \mathbf{z}^{cv}, s^{cc}, \mathbf{q}^{cc}, \mathbf{z}^{cc}) \in \mathbb{R}^{4(1+n_p+n_x)} : \\ (s^L, \mathbf{q}^L, \mathbf{z}^L, s^U, \mathbf{q}^U, \mathbf{z}^U) \in K^B\}. \tag{7.40}$$

Clearly, $K^B$ is closed and bounded, and hence compact. By the assumption that $[t,t] \times P \times X(t)$ is represented in $\mathcal{D}_f$ for every $t \in I$, Corollary 2.7.8 may be applied to conclude that $\{\mathbf{f}\}^{cv} \circ \text{MC}$ and $\{\mathbf{f}\}^{cc} \circ \text{MC}$ are Lipschitz on $K$. Denote the Lipschitz constant by $L$.

For any $(t, \mathbf{p}, \mathbf{z}^{cv}, \mathbf{z}^{cc}) \in I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$, it is easily verified that the point $(t, \mathbf{p}^L, \mathbf{x}^L(t), t, \mathbf{p}^U, \mathbf{x}^U(t), t, \mathbf{p}, \mathbf{z}^{cv}, t, \mathbf{p}, \mathbf{z}^{cc})$ is an element of $K$. By continuity of $\mathbf{x}^L$ and $\mathbf{x}^U$, it follows that $\mathbf{u}_f$ and $\mathbf{o}_f$ are continuous on $I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$. Furthermore, it follows that the Lipschitz condition of Assumption 7.2.3 is satisfied with the constant $L$. □

**Remark 7.6.8.** Note that the global Lipschitz condition of Assumption 7.2.3 is made possible by the use of the state bounds $X(t)$ and does not imply a global Lipschitz condition on $\mathbf{f}$. For fixed $(t, \mathbf{p}) \in I \times P$, the construction of $\mathbf{u}_f$ and $\mathbf{o}_f$ involves mapping any arguments $(\mathbf{z}^{cv}, \mathbf{z}^{cc}) \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ into $X(t) \times X(t)$ in a Lipschitz manner (using the MC function), so that Lipschitz continuity of $\mathbf{u}_f(t, \mathbf{p}, \cdot, \cdot)$ and $\mathbf{o}_f(t, \mathbf{p}, \cdot, \cdot)$ is really only required on this compact interval.

### 7.6.2 Constructing Relaxation Preserving Dynamics

Let $\tilde{\mathbf{u}}, \tilde{\mathbf{o}} : I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ be arbitrary functions satisfying the conditions of Lemmas 7.6.3 and 7.6.5. In this section, functions $(\mathbf{u}, \mathbf{o})$ describing relaxation preserving dynamics are derived from $(\tilde{\mathbf{u}}, \tilde{\mathbf{o}})$. In practice, we will always choose $(\tilde{\mathbf{u}}, \tilde{\mathbf{o}}) = (\mathbf{u}_f, \mathbf{o}_f)$. The notation $(\tilde{\mathbf{u}}, \tilde{\mathbf{o}})$ is used only to highlight the fact that the results below are not particular to the construction of §7.6.1.

**Definition 7.6.9.** For each $i \in \{1, \ldots, n_x\}$, define $\mathcal{R}_i^{cv}, \mathcal{R}_i^{cc} : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ by $\mathcal{R}_i^{cv}(\mathbf{z}^{cv}, \mathbf{z}^{cc}) = (\mathbf{z}^{cv}, \hat{\mathbf{z}}^{cc})$, where $\hat{z}_k^{cc} = z_k^{cc}$ if $k \neq i$ and $\hat{z}_i^{cc} = z_i^{cv}$, and $\mathcal{R}_i^{cc}(\mathbf{z}^{cv}, \mathbf{z}^{cc}) =$

$(\hat{\mathbf{z}}^{cv}, \mathbf{z}^{cc})$, where $\hat{z}_k^{cv} = z_k^{cv}$ if $k \neq i$ and $\hat{z}_i^{cv} = z_i^{cc}$.

For the remainder of this section, define $\mathbf{u}, \mathbf{o} : I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ by

$$u_i(t, \mathbf{p}, \mathbf{z}^{cv}, \mathbf{z}^{cc}) \equiv \tilde{u}_i(t, \mathbf{p}, \mathcal{R}_i^{cv}(\mathbf{z}^{cv}, \mathbf{z}^{cc})),$$
$$o_i(t, \mathbf{p}, \mathbf{z}^{cv}, \mathbf{z}^{cc}) \equiv \tilde{o}_i(t, \mathbf{p}, \mathcal{R}_i^{cc}(\mathbf{z}^{cv}, \mathbf{z}^{cc})),$$

for all $(t, \mathbf{p}, \mathbf{z}^{cv}, \mathbf{z}^{cc}) \in I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ and each $i \in \{1, \ldots, n_x\}$.

**Lemma 7.6.10.** *If $(\tilde{\mathbf{u}}, \tilde{\mathbf{o}})$ satisfy the conditions of Lemma 7.6.3, then $(\mathbf{u}, \mathbf{o})$ describe bound preserving dynamics for $\mathbf{x}$ on $I \times P$.*

*Proof.* To show that $(\mathbf{u}, \mathbf{o})$ describe bound preserving dynamics, let $\mathbf{z}^{cv}, \mathbf{z}^{cc} : I \times P \to \mathbb{R}^{n_x}$ be arbitrary functions and choose any $\mathbf{p} \in P$ and any $i \in \{1, \ldots, n_x\}$. To show Condition 1 of Definition 7.4.1, suppose $t \in I$ is such that $\mathbf{z}^{cv}(t, \mathbf{p}), \mathbf{z}^{cc}(t, \mathbf{p}) \in X(t)$, $\mathbf{z}^{cv}(t, \mathbf{p}) \leq \mathbf{z}^{cc}(t, \mathbf{p})$ and $z_i^{cv}(t, \mathbf{p}) = z_i^{cc}(t, \mathbf{p})$. Let $(\tilde{\mathbf{z}}^{cv}(t, \mathbf{p}), \tilde{\mathbf{z}}^{cc}(t, \mathbf{p})) \equiv \mathcal{R}_i^{cv}(\mathbf{z}^{cv}(t, \mathbf{p}), \mathbf{z}^{cc}(t, \mathbf{p}))$. Since $\mathbf{z}^{cv}(t, \mathbf{p}) \leq \mathbf{z}^{cc}(t, \mathbf{p})$, it follows that $\tilde{\mathbf{z}}^{cv}(t, \mathbf{p}) \leq \tilde{\mathbf{z}}^{cc}(t, \mathbf{p})$, and since $\mathbf{z}^{cv}(t, \mathbf{p}), \mathbf{z}^{cc}(t, \mathbf{p}) \in X(t)$, it follows that $\tilde{\mathbf{z}}^{cv}(t, \mathbf{p}), \tilde{\mathbf{z}}^{cc}(t, \mathbf{p}) \in X(t)$. Then Condition 1 of Lemma 7.6.3 can be applied to $(\tilde{\mathbf{z}}^{cv}(t, \mathbf{p}), \tilde{\mathbf{z}}^{cc}(t, \mathbf{p}))$ to conclude that

$$u_i(t, \mathbf{p}, \mathbf{z}^{cv}(t, \mathbf{p}), \mathbf{z}^{cc}(t, \mathbf{p})) = \tilde{u}_i(t, \mathbf{p}, \tilde{\mathbf{z}}^{cv}(t, \mathbf{p}), \tilde{\mathbf{z}}^{cc}(t, \mathbf{p})) \leq \tilde{o}_i(t, \mathbf{p}, \tilde{\mathbf{z}}^{cv}(t, \mathbf{p}), \tilde{\mathbf{z}}^{cc}(t, \mathbf{p})).$$

But, because $z_i^{cv}(t, \mathbf{p}) = z_i^{cc}(t, \mathbf{p})$, it follows that $(\tilde{\mathbf{z}}^{cv}(t, \mathbf{p}), \tilde{\mathbf{z}}^{cc}(t, \mathbf{p})) = \mathcal{R}_i^{cc}(\mathbf{z}^{cv}(t, \mathbf{p}), \mathbf{z}^{cc}(t, \mathbf{p}))$ as well, and hence

$$u_i(t, \mathbf{p}, \mathbf{z}^{cv}(t, \mathbf{p}), \mathbf{z}^{cc}(t, \mathbf{p})) \leq o_i(t, \mathbf{p}, \mathbf{z}^{cv}(t, \mathbf{p}), \mathbf{z}^{cc}(t, \mathbf{p})).$$

This proves Condition 1 of Definition 7.4.1.

To show Condition 2, suppose $t \in I$ is such that $\mathbf{z}^{cv}(t, \mathbf{p}), \mathbf{z}^{cc}(t, \mathbf{p}) \in X(t)$, $\mathbf{z}^{cv}(t, \mathbf{p}) \leq \mathbf{x}(t, \mathbf{p}) \leq \mathbf{z}^{cc}(t, \mathbf{p})$ and $x_i(t, \mathbf{p}) = z_i^{cv}(t, \mathbf{p})$. Let $(\tilde{\mathbf{z}}^{cv}(t, \mathbf{p}), \tilde{\mathbf{z}}^{cc}(t, \mathbf{p})) \equiv \mathcal{R}_i^{cv}(\mathbf{z}^{cv}(t, \mathbf{p}), \mathbf{z}^{cc}(t, \mathbf{p}))$. Since $x_i(t, \mathbf{p}) = z_i^{cv}(t, \mathbf{p})$, it follows that $\tilde{\mathbf{z}}^{cv}(t, \mathbf{p}) \leq \mathbf{x}(t, \mathbf{p}) \leq$

$\tilde{\mathbf{z}}^{cc}(t,\mathbf{p})$. Then Condition 2 of Lemma 7.6.3 can be applied to conclude that

$$u_i(t,\mathbf{p},\mathbf{z}^{cv}(t,\mathbf{p}),\mathbf{z}^{cc}(t,\mathbf{p})) = \tilde{u}_i(t,\mathbf{p},\tilde{\mathbf{z}}^{cv}(t,\mathbf{p}),\tilde{\mathbf{z}}^{cc}(t,\mathbf{p})) \leq \dot{x}_i(t,\mathbf{p}).$$

This proves Condition 2 of Definition 7.4.1, and Condition 3 is shown analogously. □

**Lemma 7.6.11.** *If* $(\tilde{\mathbf{u}},\tilde{\mathbf{o}})$ *satisfy the condition of Lemma 7.6.5, then* $(\mathbf{u},\mathbf{o})$ *describe convexity preserving dynamics for* $\mathbf{x}$ *on* $I\times P$*.*

*Proof.* To show that $(\mathbf{u},\mathbf{o})$ describe convexity preserving dynamics, let $\mathbf{z}^{cv},\mathbf{z}^{cc}: I\times P\to\mathbb{R}^{n_x}$ be arbitrary functions, choose any $(\lambda,\mathbf{p}_1,\mathbf{p}_2)\in(0,1)\times P\times P$ and define $\bar{\mathbf{p}}\equiv\lambda\mathbf{p}_1+(1-\lambda)\mathbf{p}_2$. Choose any $i\in\{1,\ldots,n_x\}$ and suppose $t\in I$ is such that

1. $\mathbf{z}^{cv}(t,\cdot)$ is consistent with convexity at $(\lambda,\mathbf{p}_1,\mathbf{p}_2)$,

2. $\mathbf{z}^{cc}(t,\cdot)$ is consistent with concavity at $(\lambda,\mathbf{p}_1,\mathbf{p}_2)$,

3. $\mathbf{z}^{cv}(t,\mathbf{q})\leq\mathbf{x}(t,\mathbf{q})\leq\mathbf{z}^{cc}(t,\mathbf{q})$ and $\mathbf{z}^{cv}(t,\mathbf{q}),\mathbf{z}^{cc}(t,\mathbf{q})\in X(t)$, $\forall\mathbf{q}\in\{\mathbf{p}_1,\mathbf{p}_2,\bar{\mathbf{p}}\}$.

Let $(\tilde{\mathbf{z}}^{cv}(t,\mathbf{p}),\tilde{\mathbf{z}}^{cc}(t,\mathbf{p}))\equiv\mathcal{R}_i^{cv}(\mathbf{z}^{cv}(t,\mathbf{p}),\mathbf{z}^{cc}(t,\mathbf{p}))$, $\forall\mathbf{p}\in P$. By the definition of $\mathcal{R}_i^{cv}$, it is clear that

$$\tilde{\mathbf{z}}^{cv}(t,\mathbf{q})\leq\tilde{\mathbf{z}}^{cc}(t,\mathbf{q}) \quad\text{and}\quad \tilde{\mathbf{z}}^{cv}(t,\mathbf{q}),\tilde{\mathbf{z}}^{cc}(t,\mathbf{q})\in X(t),\quad \forall\mathbf{q}\in\{\mathbf{p}_1,\mathbf{p}_2,\bar{\mathbf{p}}\}. \tag{7.41}$$

If $z_i^{cv}(t,\bar{\mathbf{p}})=\lambda z_i^{cv}(t,\mathbf{p}_1)+(1-\lambda)z_i^{cv}(t,\mathbf{p}_2)$, then

$$\begin{aligned}\tilde{z}_i^{cv}(t,\bar{\mathbf{p}})&=\lambda\tilde{z}_i^{cv}(t,\mathbf{p}_1)+(1-\lambda)\tilde{z}_i^{cv}(t,\mathbf{p}_2), \qquad (7.42)\\ \tilde{z}_i^{cc}(t,\bar{\mathbf{p}})&=\lambda\tilde{z}_i^{cc}(t,\mathbf{p}_1)+(1-\lambda)\tilde{z}_i^{cc}(t,\mathbf{p}_2),\end{aligned}$$

and, combining this with the hypotheses on $\mathbf{z}^{cv}$ and $\mathbf{z}^{cc}$ implies that $\tilde{\mathbf{z}}^{cv}$ and $\tilde{\mathbf{z}}^{cc}$ are, respectively, consistent with convexity and consistent with concavity at $(\lambda,\mathbf{p}_1,\mathbf{p}_2)$.

Then, since $(\tilde{\mathbf{u}}, \tilde{\mathbf{o}})$ satisfy the condition of Lemma 7.6.5,

$$\begin{aligned}
&u_i(t, \bar{\mathbf{p}}, \mathbf{z}^{cv}(t, \bar{\mathbf{p}}), \mathbf{z}^{cc}(t, \bar{\mathbf{p}})) \\
&\qquad = \tilde{u}_i(t, \bar{\mathbf{p}}, \tilde{\mathbf{z}}^{cv}(t, \bar{\mathbf{p}}), \tilde{\mathbf{z}}^{cc}(t, \bar{\mathbf{p}})) \\
&\qquad \leq \lambda \tilde{u}_i(t, \mathbf{p}_1, \tilde{\mathbf{z}}^{cv}(t, \mathbf{p}_1), \tilde{\mathbf{z}}^{cc}(t, \mathbf{p}_1)) + (1-\lambda)\tilde{u}_i(t, \mathbf{p}_2, \tilde{\mathbf{z}}^{cv}(t, \mathbf{p}_2), \tilde{\mathbf{z}}^{cc}(t, \mathbf{p}_2)) \\
&\qquad = \lambda u_i(t, \mathbf{p}_1, \mathbf{z}^{cv}(t, \mathbf{p}_1), \mathbf{z}^{cc}(t, \mathbf{p}_1)) + (1-\lambda) u_i(t, \mathbf{p}_2, \mathbf{z}^{cv}(t, \mathbf{p}_2), \mathbf{z}^{cc}(t, \mathbf{p}_2)).
\end{aligned}$$

On the other hand, letting $(\tilde{\mathbf{z}}^{cv}(t, \mathbf{p}), \tilde{\mathbf{z}}^{cc}(t, \mathbf{p})) \equiv \mathcal{R}_i^{cc}(\mathbf{z}^{cv}(t, \mathbf{p}), \mathbf{z}^{cc}(t, \mathbf{p}))$ and supposing that $z_i^{cc}(t, \bar{\mathbf{p}}) = \lambda z_i^{cc}(t, \mathbf{p}_1) + (1-\lambda) z_i^{cc}(t, \mathbf{p}_2)$, (7.41) and (7.42) again hold and hence

$$\begin{aligned}
&o_i(t, \bar{\mathbf{p}}, \mathbf{z}^{cv}(t, \bar{\mathbf{p}}), \mathbf{z}^{cc}(t, \bar{\mathbf{p}})) \\
&\qquad = \tilde{o}_i(t, \bar{\mathbf{p}}, \tilde{\mathbf{z}}^{cv}(t, \bar{\mathbf{p}}), \tilde{\mathbf{z}}^{cc}(t, \bar{\mathbf{p}})) \\
&\qquad \geq \lambda \tilde{o}_i(t, \mathbf{p}_1, \tilde{\mathbf{z}}^{cv}(t, \mathbf{p}_1), \tilde{\mathbf{z}}^{cc}(t, \mathbf{p}_1)) + (1-\lambda)\tilde{o}_i(t, \mathbf{p}_2, \tilde{\mathbf{z}}^{cv}(t, \mathbf{p}_2), \tilde{\mathbf{z}}^{cc}(t, \mathbf{p}_2)) \\
&\qquad = \lambda o_i(t, \mathbf{p}_1, \mathbf{z}^{cv}(t, \mathbf{p}_1), \mathbf{z}^{cc}(t, \mathbf{p}_1)) + (1-\lambda) o_i(t, \mathbf{p}_2, \mathbf{z}^{cv}(t, \mathbf{p}_2), \mathbf{z}^{cc}(t, \mathbf{p}_2)).
\end{aligned}$$

□

For the computations described in the following section, we will always use $(\tilde{\mathbf{u}}, \tilde{\mathbf{o}}) = (\mathbf{u}_f, \mathbf{o}_f)$, where $(\mathbf{u}_f, \mathbf{o}_f)$ are as described in §7.6.1. In addition to guaranteeing that $(\mathbf{u}, \mathbf{o})$ describe relaxation preserving dynamics, this choice also guarantees continuity of $(\mathbf{u}, \mathbf{o})$ on $I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$, as well as the global Lipschitz condition of Assumption 7.2.3. Since both of these properties hold for $(\mathbf{u}_f, \mathbf{o}_f)$, it is trivial to show that they hold for $(\mathbf{u}, \mathbf{o})$ as defined in this section.

### 7.6.3 Implementation

This section describes the computational implementation of the state relaxation theories developed for ODEs in the previous sections. Following the results in §7.6.2, we define $\mathbf{u}, \mathbf{o} : I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ by

$$\begin{aligned} u_i(t, \mathbf{p}, \mathbf{z}^{cv}, \mathbf{z}^{cc}) &\equiv u_{f,i}(t, \mathbf{p}, \mathcal{R}_i^{cv}(\mathbf{z}^{cv}, \mathbf{z}^{cc})), \\ o_i(t, \mathbf{p}, \mathbf{z}^{cv}, \mathbf{z}^{cc}) &\equiv o_{f,i}(t, \mathbf{p}, \mathcal{R}_i^{cc}(\mathbf{z}^{cv}, \mathbf{z}^{cc})), \end{aligned}$$

for each $i \in \{1, \ldots, n_x\}$, where $(\mathbf{u}_f, \mathbf{o}_f)$ are defined by (7.37).

For the numerical examples in §7.7, state bounds are computed using Harrison's method. As with all of the methods described in Chapter 3, Harrison's method describes $\mathbf{x}^L$ and $\mathbf{x}^U$ as the solutions of another auxiliary system of ODEs. Given $(t_f, \mathbf{p}) \in I \times P$ at which the values $\mathbf{x}^{cv}(t_f, \mathbf{p})$ and $\mathbf{x}^{cc}(t_f, \mathbf{p})$ are desired, the ODEs describing the state bounds are numerically integrated simultaneously with the auxiliary ODEs (either (7.1) or (7.13)) at $\mathbf{p}$, from $t_0$ to $t_f$. Thus, the values $\mathbf{x}^L(t)$, $\mathbf{x}^U(t)$, $\dot{\mathbf{x}}^L(t)$ and $\dot{\mathbf{x}}^U(t)$ are available at every time-step during numerical integration. To begin this computation, the initial conditions $\mathbf{x}_0^{cv}(\mathbf{p})$ and $\mathbf{x}_0^{cc}(\mathbf{p})$ are computed by taking standard McCormick relaxations of $\mathbf{x}_0$ on $P$, evaluated at $\mathbf{p}$. This is done using the `C++` library `MC++`, which automatically computes interval extensions and McCormick relaxations of factorable functions using operator overloading (http://www3.imperial.ac.uk/people/b.chachuat/research). `MC++` is the successor of `libMC`, which is described in detail in [122]. Whenever it is required to evaluate the right-hand side of (7.1) or (7.13), the functions $u_i$ and $o_i$ are evaluated automatically using `MC++` according to the discussion in the previous sections and the definitions in Chapter 2.

**Remark 7.6.12.** When using any of the more sophisticated state bounding methods developed in Chapter 3, one must take care that the derivatives $\dot{\mathbf{x}}^L(t)$ and $\dot{\mathbf{x}}^U(t)$ appearing in the right-hand sides of (7.13) correspond exactly to the bounds $\mathbf{x}^L(t)$ and $\mathbf{x}^U(t)$ used in the computation of $(\mathbf{u}, \mathbf{o})$ as per §7.6.1 and §7.6.2. For example, it is not valid to evaluate the natural McCormick extensions defining $(\mathbf{u}, \mathbf{o})$ over the interval resulting from a refinement of $X(t)$ based on some known physical information, unless $\dot{\mathbf{x}}^L(t)$ and $\dot{\mathbf{x}}^U(t)$ are adjusted accordingly.

For both state relaxation methods, numerical simulation of the auxiliary ODEs is done using CVODE [44] with relative and absolute tolerances of $1\times 10^{-8}$. The simulation of (7.1) is straightforward. To simulate a solution of (7.13), we numerically integrate the system

$$\dot{x}_i^{cv}(t,\mathbf{p}) = \begin{cases} u_i(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})) & \text{if } \;\; b_i^{cv}=0 \\ \max(\dot{x}_i^L(t), u_i(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p}))) & \text{if } \;\; b_i^{cv}=1 \end{cases}, \tag{7.43}$$

$$\dot{x}_i^{cc}(t,\mathbf{p}) = \begin{cases} o_i(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})) & \text{if } \;\; b_i^{cc}=0 \\ \min(\dot{x}_i^U(t), o_i(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p}))) & \text{if } \;\; b_i^{cc}=1 \end{cases},$$

with initial conditions as in (7.13), where $b_i^{cv}$ and $b_i^{cc}$ are Boolean variables which ideally satisfy

$$b_i^{cv} = \begin{cases} 0 & \text{if } \;\; x_i^{cv}(t,\mathbf{p}) > x_i^L(t) \\ 1 & \text{if } \;\; x_i^{cv}(t,\mathbf{p}) \le x_i^L(t) \end{cases}, \qquad b_i^{cc} = \begin{cases} 0 & \text{if } \;\; x_i^{cc}(t,\mathbf{p}) < x_i^U(t) \\ 1 & \text{if } \;\; x_i^{cc}(t,\mathbf{p}) \ge x_i^U(t) \end{cases}. \tag{7.44}$$

In practice, the values of each $b_i^{cv}$ and $b_i^{cc}$ are set according to an event detection scheme described below. Assuming that a solution $(\mathbf{x}^{cv},\mathbf{x}^{cc})$ of (7.13) exists on $I\times P$ (see Remark 7.5.2), Lemmas 7.5.3 and 7.5.6 imply that $(\mathbf{x}^{cv},\mathbf{x}^{cc})$ is a solution of (7.43) in the sense that, for each $\mathbf{p}\in P$, $\mathbf{x}^{cv}(\cdot,\mathbf{p})$ and $\mathbf{x}^{cc}(\cdot,\mathbf{p})$ satisfy (7.43) and (7.44) for a.e. $t\in I$. Furthermore, it follows from Assumption 7.2.3 and Theorem 1 in Chapter 2, §10 of [62] that this is the unique solution of (7.43). Thus, computing the solution of (7.43) furnishes the solution of (7.13), which guarantees that the computed $(\mathbf{x}^{cv},\mathbf{x}^{cc})$ are state relaxations for (7.36) on $I\times P$.

The numerical simulation of (7.43) is carried out as follows. At $t_0$, the variables $b_i^{cv}$ and $b_i^{cc}$ are set according to (7.44). With these variables fixed, numerical integration of (7.43) is initiated using CVODES [44]. CVODES offers a built-in feature which checks for zero crossings of user supplied event functions $\mathbf{g}(t)$ during each integration step, and locates event times $t_e$ at which $g_j(t_e)=0$ for some $j$ using a bisection algorithm.

For the integration of (7.43), we provide the event functions

$$\begin{aligned} g_i(t) &= x_i^{cv}(t,\mathbf{p}) - x_i^L(t) - b_i^{cv}\epsilon, \\ g_{n_x+i}(t) &= x_i^U(t) - x_i^{cc}(t,\mathbf{p}) - b_i^{cc}\epsilon, \end{aligned}$$

for $i \in \{1,\ldots,n_x\}$. Starting from $t_0$, the system (7.43) is integrated until a root of any of these event functions is located. Suppose that, for some $t_j \in I$ and some index $i$, $x_i^{cv}(t_j,\mathbf{p}) > x_i^L(t)$ and $b_i^{cv} = 0$ as desired. If, after the next integration step $[t_j, t_{j+1}]$, it is found that $x_i^{cv}(t_{j+1},\mathbf{p}) < x_i^L(t_{j+1})$, then a zero crossing of $g_i(t) = x_i^{cv}(t,\mathbf{p}) - x_i^L(t)$ has occurred and CVODES will search for a point $t_e \in [t_j, t_{j+1}]$ such that $x_i^{cv}(t_e,\mathbf{p}) = x_i^L(t_e)$. Then, integration is stopped at $t_e$, $b_i^{cv}$ is reset to 1 in order to specify the correct evolution of $x_i^{cv}(\cdot,\mathbf{p})$ to the right of $t_e$, and integration is resumed. In addition to specifying the correct evolution of $c(\cdot,\mathbf{p})$ to the right of $t_e$, setting $b_i^{cv} = 1$ also introduces an epsilon perturbation into the event function $g_i$. This is to avoid repeatedly 'finding' roots of $g_i$, since if $\dot{x}_i^L(t) > u_i(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p}))$ for some nontrivial period of time to the right of $t_e$, then the equality $x_i^{cv}(t,\mathbf{p}) = x_i^L(t)$ will persist (indeed, this is the intended behavior).

Next, suppose that, for some $t_j \in I$ and some index $i$, $x_i^{cv}(t_j,\mathbf{p}) = x_i^L(t)$ and $b_i^{cv} = 1$ as desired. Further, suppose that after the next integration step $[t_j, t_{j+1}]$, it happens that $x_i^{cv}(t_{j+1},\mathbf{p}) > x_i^L(t_{j+1})$. If $\epsilon > 0$ is sufficiently small, then a zero crossing of $g_i(t) = x_i^{cv}(t,\mathbf{p}) - x_i^L(t) + \epsilon$ will be detected and CVODES will search for a point $t_e \in [t_j, t_{j+1}]$ such that $g_i(t_e) = x_i^{cv}(t_e,\mathbf{p}) - x_i^L(t_e) - \epsilon = 0$. Again, integration is stopped in order to reset $b_i^{cv}$ to 0. On account of the $\epsilon$ perturbation in $g_i$, the event where $x_i^{cv}(\cdot,\mathbf{p})$ ceases to equal $x_i^L$ is not found precisely, and therefore the value of $\dot{x}_i^{cv}(\cdot,\mathbf{p})$ is not adjusted as per (7.43) until slightly too late. However, this is a minor issue for this particular system since the fact that such an event has occurred implies that $\dot{x}_i^L(t)$ has been strictly less than $u_i(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p}))$ for some nontrivial period of time before the root $g_i(t) = x_i^{cv}(t,\mathbf{p}) - x_i^L(t) - \epsilon = 0$ was detected, and hence

$$\begin{aligned}\dot{x}_i^{cv}(t,\mathbf{p}) &= \max(\dot{x}_i^L(t), u_i(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p}))),\\ &= u_i(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})),\end{aligned}$$

over that same period, exactly as if $b_i^{cv} = 0$. Of course, the variables $b_i^{cc}$ are managed in an analogous fashion.

An alternative approach to detecting this last event, i.e. where $x_i^{cv}(t_j,\mathbf{p}) = x_i^L(t_j)$ and $x_i^{cv}(\cdot,\mathbf{p}) - x_i^L$ first becomes positive at some $t_e \in [t_j, t_{j+1}]$, is to search for a zero crossing of the function $u_i(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})) - \dot{x}_i^L(t)$. Since this function must become positive immediately to the right of $t_e$, the event will be detected provided that this function is negative at $t_j$. We do not use this implementation here for the following reason. For the purposes of optimization, it may be desirable to evaluate $\mathbf{x}^{cv}(t_f,\cdot)$ and $\mathbf{x}^{cc}(t_f,\cdot)$ at several points in $P$. However, the state bounds need only be integrated once because $P$ is constant. In this case, it is beneficial to store the state bounding trajectories and evaluate $\mathbf{x}^L$ and $\mathbf{x}^U$ when needed by interpolation. Values for $\dot{\mathbf{x}}^L$ and $\dot{\mathbf{x}}^U$ can be computed by evaluating the right-hands sides of the bounding ODEs. However, in this scheme there is some numerical disagreement between the values $\dot{\mathbf{x}}^L(\hat{t})$ and $\dot{\mathbf{x}}^U(\hat{t})$ computed at some point $\hat{t}$ and the behavior of the interpolated values $\mathbf{x}^L(t)$ and $\mathbf{x}^U(t)$ for $t$ immediately to the right of $\hat{t}$. Though this inconsistency is minor, it does cause significant complications for an event detection scheme based on precise values of $\dot{\mathbf{x}}^L$ and $\dot{\mathbf{x}}^U$.

## 7.7 Numerical Examples

All numerical experiments in this section were performed on a Dell Precision T3400 workstation with a 2.83 GHz Intel Core2 Quad CPU. One core and 512 MB of memory were dedicated to each job.

**Example 7.7.1** (Relaxation Amplifying Dynamics)**.** Section 1.2.4 of [91] discusses a negative resistance circuit consisting of an inductor, a capacitor and a resistive

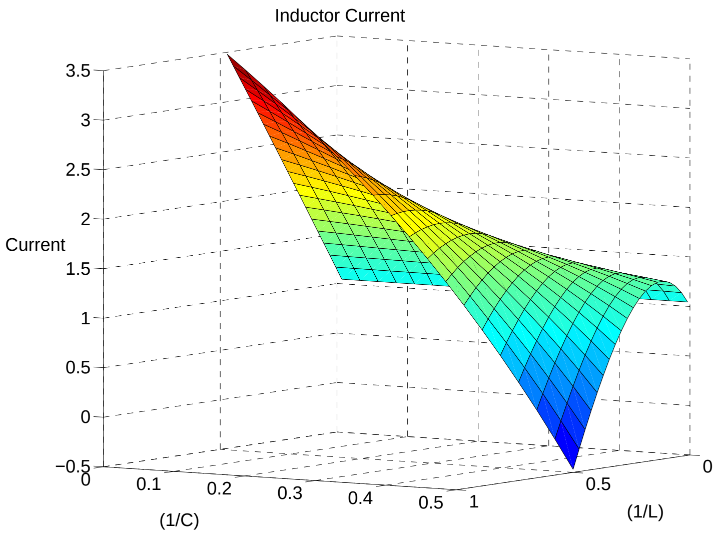

Figure 7-1: The parametric final time solution of the ODEs (7.45), $x_1(t_f, \cdot)$, on the interval $P = [0.01, 0.5]^2$.

element in parallel. The circuit can be described by the nonlinear ODEs

$$\dot{x}_1 = \frac{1}{L}x_2, \quad \dot{x}_2 \quad = \quad -\frac{1}{C}[x_1 - x_2 + \frac{1}{3}x_2^3], \tag{7.45}$$

where $L$ and $C$ are the inductance and capacitance respectively, $x_1$ is the current through the inductor, and $x_2$ is the voltage across the capacitor. It is assumed that time, $C$, $L$, $x_1$ and $x_2$ are scaled so that the equations above are dimensionless and all quantities are of order one with the possible exception of $(1/L)$ and $(1/C)$. Therefore, the initial value problem with $x_{0,1} = x_{0,2} = 1$, $t_0 = 0$ and $t_f = 5$ is considered. Letting the parameters be $p_1 = (1/C)$ and $p_2 = (1/L)$, the solution $x_1(t_f, \cdot)$ on the set $P = [p_1^L, p_1^U] \times [p_2^L, p_2^U] = [0.01, 0.5] \times [0.01, 0.5]$ is shown in Figure 7-1. The parametric final time solution is clearly nonconvex, with a single maximum at $(p_1, p_2) = (0.01, 0.5)$ and two local minima, the global minimum at $(p_1, p_2) = (0.5, 0.5)$, and a suboptimal local minimum at $(p_1, p_2) = (0.01, 0.01)$.

Since $\mathbf{x}_0$ is constant, appropriate convex and concave relaxations are simply $\mathbf{x}_0^{cv} =$

Table 7.1: Factorization and computation of $f_1$ at $(t, \mathbf{p}, \mathbf{x})$ and $u_1$ and $o_1$ at $(t, \mathbf{p}, \mathbf{z}, \mathbf{y})$.

| $i$ | $v_i$ | $V_i$ | $\mathcal{V}_i$ |
|---|---|---|---|
| 1 | $p_1$ | $P_1$ | $(P_1, [p_1, p_1])$ |
| 2 | $x_2$ | $X_2(t)$ | $\mathrm{MC}(x_2^L(t), x_2^U(t), z_2, y_2))$ |
| 3 | $v_1 v_2$ | $V_1 V_2$ | $\mathcal{V}_1 \mathcal{V}_2$ |

Table 7.2: Factorization and computation of $f_2$ at $(t, \mathbf{p}, \mathbf{x})$ and $u_2$ and $o_2$ at $(t, \mathbf{p}, \mathbf{z}, \mathbf{y})$.

| $i$ | $v_i$ | $V_i$ | $\mathcal{V}_i$ |
|---|---|---|---|
| 1 | $p_2$ | $P_2$ | $(P_2, [p_2, p_2])$ |
| 2 | $x_1$ | $X_1(t)$ | $\mathrm{MC}(x_1^L(t), x_1^U(t), z_1, y_1)$ |
| 3 | $x_2$ | $X_2(t)$ | $\mathrm{MC}(x_2^L(t), x_2^U(t), z_2, y_2)$ |
| 4 | $v_3^3$ | $V_3^3$ | $\mathcal{V}_3^3$ |
| 5 | $(1/3)v_4$ | $(1/3)V_4$ | $(1/3)\mathcal{V}_4$ |
| 6 | $-v_3$ | $-V_3$ | $-\mathcal{V}_3^L$ |
| 7 | $v_2 + v_6$ | $V_2 + V_6$ | $\mathcal{V}_2 + \mathcal{V}_6$ |
| 8 | $v_7 + v_5$ | $V_7 + V_5$ | $\mathcal{V}_7 + \mathcal{V}_5$ |
| 9 | $-v_1$ | $-V_1$ | $-\mathcal{V}_1$ |
| 10 | $v_8 v_9$ | $V_8 V_9$ | $\mathcal{V}_8 \mathcal{V}_9$ |

$\mathbf{x}_0^{cc} = \mathbf{x}_0$. Then, beginning from the function

$$\mathbf{f} = [f_1, f_2]^{\mathrm{T}} = \left[ p_1 x_2, \quad -p_2 \left( x_1 - x_2 + \frac{1}{3} x_2^3 \right) \right]^{\mathrm{T}}, \tag{7.46}$$

it remains to construct functions $\mathbf{u}$ and $\mathbf{o}$ satisfying relaxation amplifying dynamics.

For any $(t, \mathbf{p}, \mathbf{z}, \mathbf{y}) \in I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$, appropriate values for the functions $\mathbf{u}$ and $\mathbf{o}$ at $(t, \mathbf{p}, \mathbf{z}, \mathbf{y})$ can be computed by evaluating the natural McCormick extension $\{\mathbf{f}\}(([t,t],[t,t]), (P, [\mathbf{p}, \mathbf{p}]), \mathrm{MC}(\mathbf{x}^L(t), \mathbf{x}^U(t), \mathbf{z}, \mathbf{y}))$. This is implemented by the factorable representations of $f_1$ and $f_2$ shown in Tables 7.1 and 7.2, with factors $v_i$, inclusion factors $V_i$, and relaxation factors $\mathcal{V}_i$. Now $u_1(t, \mathbf{p}, \mathbf{z}, \mathbf{y})$ and $o_1(t, \mathbf{p}, \mathbf{z}, \mathbf{y})$ evaluate to $\mathcal{V}_3^{cv}$ and $\mathcal{V}_3^{cc}$ in Table 7.1, respectively, and $u_2(t, \mathbf{p}, \mathbf{z}, \mathbf{y})$ and $o_2(t, \mathbf{p}, \mathbf{z}, \mathbf{y})$ evaluate to $\mathcal{V}_{10}^{cv}$ and $\mathcal{V}_{10}^{cc}$ in Table 7.2, respectively.

Given the functions $\mathbf{x}_0^{cv}$, $\mathbf{x}_0^{cc}$, $\mathbf{u}$ and $\mathbf{o}$ as described above, convex and concave relaxations for the parametric solution of (7.45) were generated by application of Theorem 7.3.4. The resulting relaxations are shown in Figure 7-2. Clearly, the minimum of the convex relaxation underestimates the global minimum of $x_1(t_f, \cdot)$.

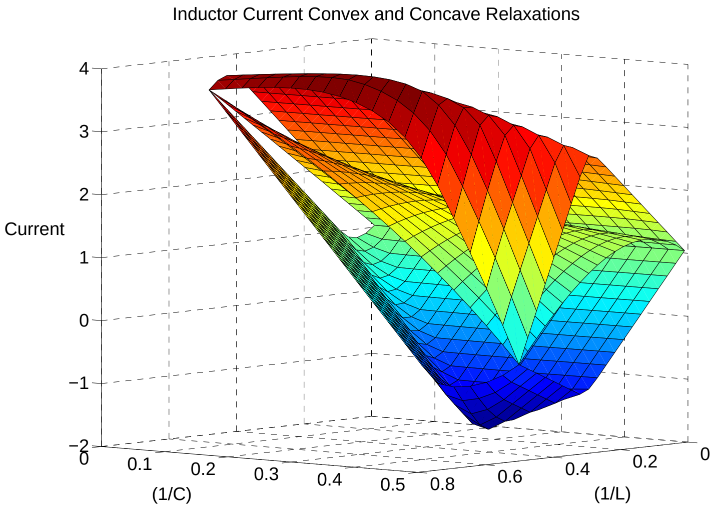

Figure 7-2: State relaxations for $x_1(t_f, \cdot)$, the solution of the ODEs (7.45), constructed over the interval $P = [0.01, 0.5]^2$.

Figure 7-3 shows a second pair of convex and concave relaxations, plotted with the first, constructed in exactly the same way over the subinterval $P^1 = [0.3, 0.5]^2$ (the solution of (7.45) has been omitted for clarity). Clearly, the relaxations become much tighter when taken over a subinterval of the original parameter interval $P$.

**Example 7.7.2** (Relaxation Preserving Dynamics)**.** Consider the nonlinear ODEs

$$\begin{aligned} \dot{x}_1(t,\mathbf{p}) &= -(2+\sin(p_1/3))(x_1(t,\mathbf{p}))^2 + p_2 x_1(t,\mathbf{p}) x_2(t,\mathbf{p}), \quad && x_1(t_0,\mathbf{p}) = 1, \\ \dot{x}_2(t,\mathbf{p}) &= \sin(p_1/3)(x_1(t,\mathbf{p}))^2 - p_2 x_1(t,\mathbf{p}) x_2(t,\mathbf{p}), && x_2(t_0,\mathbf{p}) = 0.5, \end{aligned} \tag{7.47}$$

where $\mathbf{p} = (p_1, p_2) \in P \equiv [-6.5, 6.5] \times [0.01, 0.5]$ and $I \equiv [t_0, t_f] = [0.0, 2.0]$. State relaxations for this system on $I \times P$ were computed using the affine relaxation method in [162], as well as the two nonlinear state relaxations methods presented here. For brevity, we will refer to state relaxations computed through the theory of relaxation amplifying dynamics as RAD relaxations, and those computed through relaxation preserving dynamics as RPD relaxations.

The results are shown in Figures 7-4 and 7-5. In both figures, the parametric solution $x_2(t, \cdot)$, along with convex and concave relaxations computed by all three

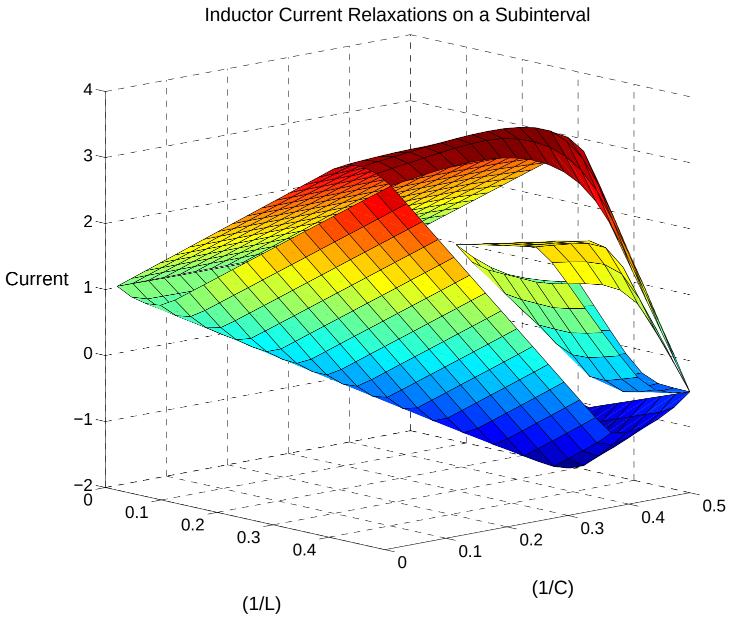

Figure 7-3: State relaxations for $x_1(t_f, \cdot)$, the solution of the ODEs (7.45), constructed over the interval $P = [0.01, 0.5]^2$ and the subinterval $P^1 = [0.3, 0.5]^2$.

methods, is plotted as a function of $p_1$, with $p_2 = 0.5$ fixed. Figure 7-4 displays these functions at $t = 0.1$, after only a short integration time, while Figure 7-5 shows them at $t = t_f = 2.0$. In the case of the affine and RAD relaxations, the figures actually display $\max(x_2^L(t), c_2(t, \cdot))$ and $\min(x_2^U(t), C_2(t, \cdot))$, in order to compare the best bounds that would be available from each method in a branch-and-bound setting (this makes the affine relaxations appear non-affine in some figures). Of course, the RPD always remain tighter than the state bounds by construction (Lemma 7.5.3).

After a short integration time, the nonlinear methods produce almost identical results (the curves nearly overlap in Figure 7-4). On the other hand, the affine relaxations are weaker for many values of $p_1$ because the parametric solution $x_2(t, \cdot)$ is highly nonlinear. After a long integration time, the RPD relaxations proposed here are significantly tighter than those resulting from either of the other two methods. The strength of the RAD relaxations apparently deteriorates with integration time, so that at $t = 2.0$ the advantages of nonlinearity are nearly lost. This is attributed to the fact that this method is based on bound amplifying dynamics, whereas the RPD relaxations are based on bound preserving dynamics. The bounding properties of the affine state relaxations in [162] are also essentially based on the principle of bound preserving dynamics. However, the affine nature of these relaxations is a consequence of linear systems theory, not convexity preserving dynamics.

In order to demonstrate the convergence behavior of these methods, Figure 7-6 shows state relaxations constructed on the interval $P^* \equiv [-4.9, -4.6] \times [0.45, 0.5]$, again plotted as functions of $p_1$ with $p_2 = 0.5$ fixed and $t = 2.0$. The parametric solution $x_2(2.0, \cdot)$ is much more nearly linear on $P^*$, so that the advantages of nonlinear relaxations are diminished. It is clear from this figure that the RAD relaxations are the weakest on small intervals. These relaxations are also much more convex/concave than seems warranted by the curvature of $x_2(2.0, \cdot)$. In contrast, the affine relaxations provide very reasonable bounds on this smaller interval, in part because $x_2(2.0, \cdot)$ is more nearly linear, but also because the bounding properties of these relaxations are based on bound preserving dynamics. However, the RPD relaxations are again the strongest. These relaxations not only make use of bound preserving dynamics, but also of convexity preserving dynamics, so that the advantages of nonlinearity are maintained without the excessive convexification demonstrated by the RAD theory.

**Example 7.7.3.** Consider again the chemical kinetics model of Example 4.5.4. Using experimental data from [172], globally optimal parameter estimates for this model were computed in [163] using the global dynamic optimization algorithm of [164], which is based on the affine state relaxation method in [162]. In this example, RAD and RPD state relaxations are computed for this problem and compared to the affine relaxations of [162].

The kinetic model takes the form

$$\dot{\mathbf{x}}(t, \mathbf{p}) = \mathbf{S}\mathbf{r}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})),$$

where $\mathbf{S}$ and $\mathbf{r}$ are give in Example 4.5.4 and $\mathbf{p} = (k_2, k_3, k_4)$. The parameters $k_2$ and $k_3$ are restricted to the interval $[1, 1200]$ ($\text{M}^{-1}\mu\text{s}^{-1}$), and $k_4$ is restricted to the interval $[0.001, 100]$ ($\mu\text{s}^{-1}$). That is $P \equiv [1, 1200] \times [1, 1200] \times [0.001, 100]$. The initial

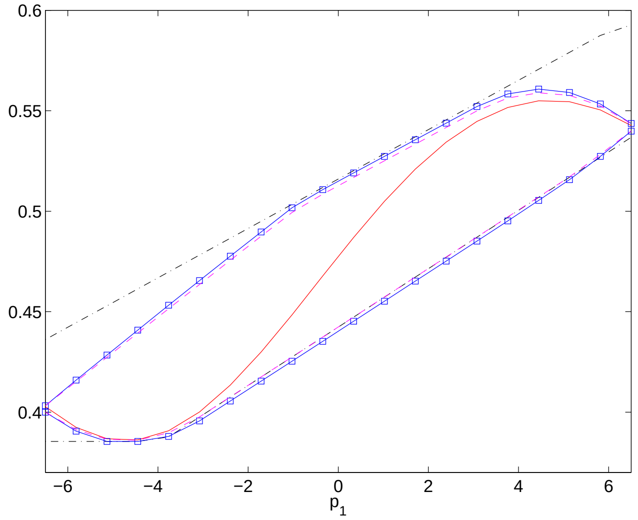

Figure 7-4: The parametric solution $x_2$ (solid), along with affine state relaxations (dot-dashed), RAD (squares) and RPD (dashed) state relaxations, constructed on $P$ and plotted as functions of $p_1$ with $p_2 = 0.5$ and $t = 0.1$ fixed.

conditions are the same as in Example 4.5.4.

State relaxations for this system were computed on the time interval $I \equiv [0, 4.5]$ $\mu$s. Due to the very fast time constants in this system, this time horizon is long enough for the system to nearly reach steady-state for all parameter values. For all three state relaxation methods, numerical integration of the auxiliary system was done using CVODES [44] with absolute and relative error tolerances of $10^{-10}$. These tolerances were used because some state variables in the original system take meaningful nonzero values on the order of $10^{-8}$ M.

In [163], parameter estimation for this system was done by fitting the model to measured absorbance data. The absorbance depends on the species concentrations according to

$$\text{Abs}(\mathbf{x}(t, \mathbf{p})) \equiv 1470 x_3(t, \mathbf{p}) + 140(x_6(t, \mathbf{p}) + x_7(t, \mathbf{p})). \tag{7.48}$$

Figures 7-7 and 7-8 below show relaxations of $x_6(t_f, \cdot)$ and $\text{Abs}(\mathbf{x}(t_f, \cdot))$ on $P$, respectively, for the final time $t_f = 4.5$ $\mu$s. Having computed state relaxations, convex and

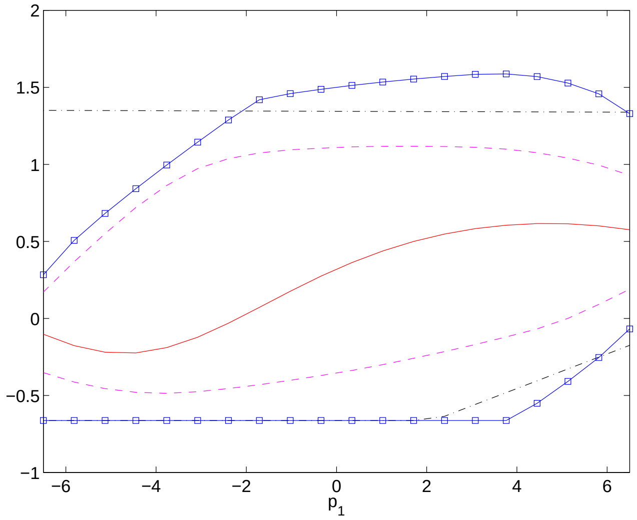

Figure 7-5: The parametric solution $x_2$ (solid), along with affine state relaxations (dot-dashed), RAD (squares) and RPD (dashed) state relaxations, constructed on $P$ and plotted as functions of $p_1$ with $p_2 = 0.5$ and $t = 2.0$ fixed.

concave relaxations for $\text{Abs}(\mathbf{x}(t_f, \cdot))$ are given by

$$u_{\text{Abs}}(\mathbf{x}^{cv}(t_f, \mathbf{p}), \mathbf{x}^{cc}(t_f, \mathbf{p})) \equiv 1470 x_3^{cv}(t, \mathbf{p}) + 140(x_6^{cv}(t, \mathbf{p}) + x_7^{cv}(t, \mathbf{p})),$$
$$o_{\text{Abs}}(\mathbf{x}^{cv}(t_f, \mathbf{p}), \mathbf{x}^{cc}(t_f, \mathbf{p})) \equiv 1470 x_3^{cc}(t, \mathbf{p}) + 140(x_6^{cc}(t, \mathbf{p}) + x_7^{cc}(t, \mathbf{p})).$$

To illustrate the convergence behavior of the three relaxation methods, Figures 7-9 and 7-10 show relaxations of $x_6(t_f, \cdot)$ and $\text{Abs}(\mathbf{x}(t_f, \cdot))$, respectively, on a much smaller interval containing the globally optimal parameter values from [163], $P^* \equiv [475, 550] \times [375, 425] \times [17, 21]$. In all figures, the depicted relaxations are constructed over the entire interval $P$ (or $P^*$), but plotted for clarity only as functions of $k_2$, with $(k_3, k_4)$ fixed at the globally optimal values from [163], $(403\ \text{M}^{-1}\mu\text{s}^{-1}, 19.2\mu\text{s}^{-1})$. In the case of the affine and first generation nonlinear relaxations, Figures 7-7 and 7-9 actually display $\max(x_6^L(t), x_6^{cv}(t, \cdot))$ and $\min(x_6^U(t), x_6^{cc}(t, \cdot))$. The relaxations in Figures 7-8 and 7-10 are similarly truncated at upper and lower bounds for $\text{Abs}(\mathbf{x}(t_f, \cdot))$ on $P$ (resp. $P^*$) computed by taking the natural interval extension of (7.48) using the state bounds. Though the parameter dependence of the true model solutions in these

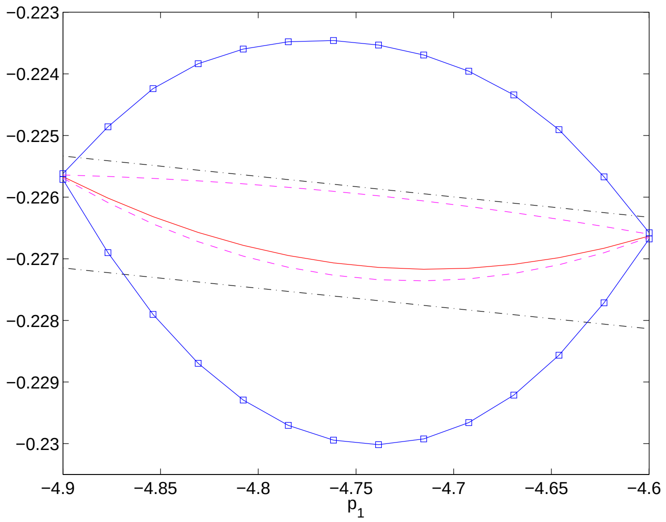

Figure 7-6: The parametric solution $x_2$ (solid), along with affine state state relaxations (dot-dashed), RAD (squares) and RPD (dashed) state relaxations, constructed on $P$ and plotted as functions of $p_1$ with $p_2 = 0.5$ and $t = 2.0$ fixed.

figures appears to be fairly simple, these figures show only the dependence on $k_2$ with the other two parameters fixed. In [163], it is shown that the corresponding parameter estimation problem is indeed nonconvex and has numerous suboptimal local minima.

As in the previous example, the RAD nonlinear relaxations are superior to the affine relaxations on the large interval $P$ due to their nonlinearity. However, on the smaller interval $P^*$ the situation is reversed, showing again that the RAD nonlinear relaxations converge much more slowly than do the affine relaxations. On the other hand, the RPD relaxations provide much tighter bounds than either of the other methods on both large and small intervals.

## 7.8 Conclusion

Given a nonlinear system of ODEs (7.36), two sets of sufficient conditions have been established for a system of auxiliary differential equations of the form (7.1) to describe convex and concave relaxations of each state variable with respect to the ODE parameters, pointwise in the independent variable. Furthermore, computational methods

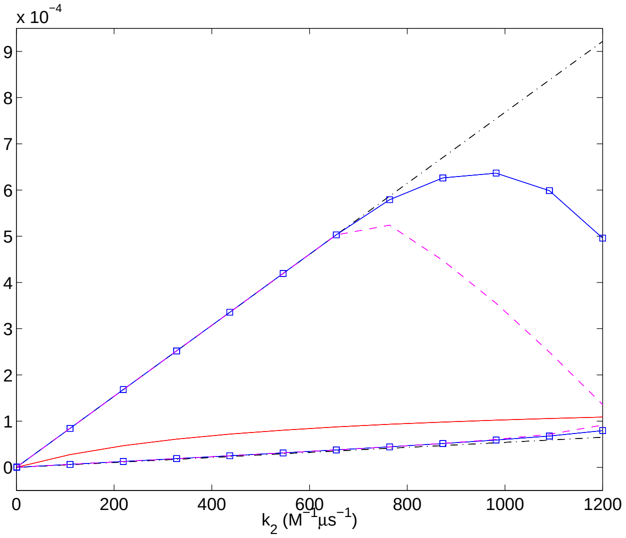

Figure 7-7: The parametric solution $x_6$ (solid), along with affine state relaxations (dot-dashed), RAD (squares) and RPD (dashed) state relaxations, constructed on $P$ and plotted as functions of $k_2$ with $k_3 = 403.0$ M$^{-1}\mu$s$^{-1}$, $k_4 = 19.2$ $\mu$s$^{-1}$ and $t = 4.5$ $\mu$s fixed.

were developed for automatically constructing and solving the required auxiliary system for both sets of conditions. Through a conceptual analysis, corroborated by numerical results, it has been shown that the second relaxation theory, based on the concept of relaxation preserving dynamics, is superior to the first. It has also been shown through numerical examples that this relaxation theory significantly outperforms a related existing method for computing affine state relaxations.

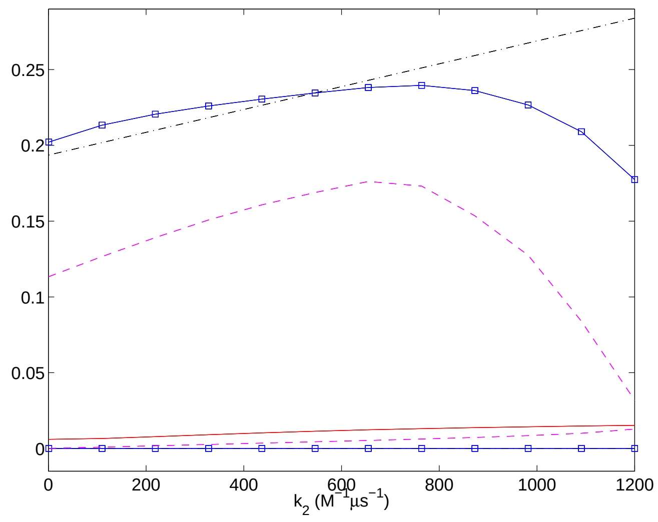

Figure 7-8: The absorbance Abs($\mathbf{x}(t_f, \cdot)$) (solid), along with affine state relaxations (dot-dashed), RAD (squares) and RPD (dashed) state relaxations, constructed on $P$ and plotted as functions of $k_2$ with $k_3 = 403.0$ M$^{-1}\mu$s$^{-1}$, $k_4 = 19.2$ $\mu$s$^{-1}$ and $t = 4.5$ $\mu$s fixed.

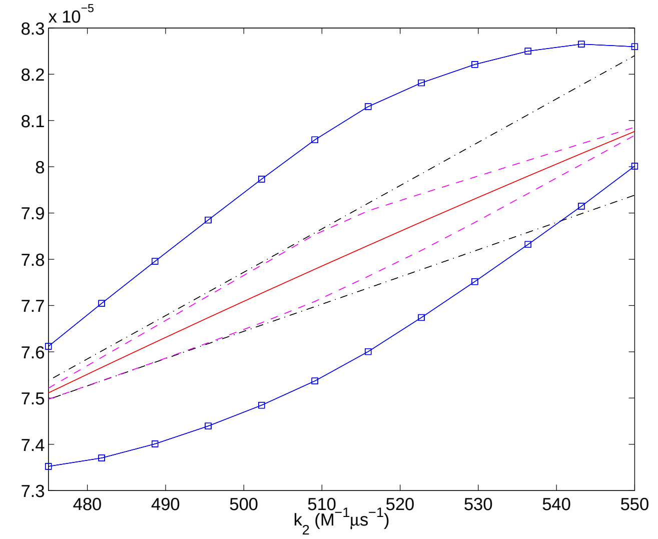

Figure 7-9: The parametric solution $x_6$ (solid), along with affine state relaxations (dot-dashed), RAD (squares) and RPD (dashed) state relaxations, constructed on $P^*$ and plotted as functions of $k_2$ with $k_3 = 403.0$ M$^{-1}\mu$s$^{-1}$, $k_4 = 19.2$ $\mu$s$^{-1}$ and $t = 4.5$ $\mu$s fixed.

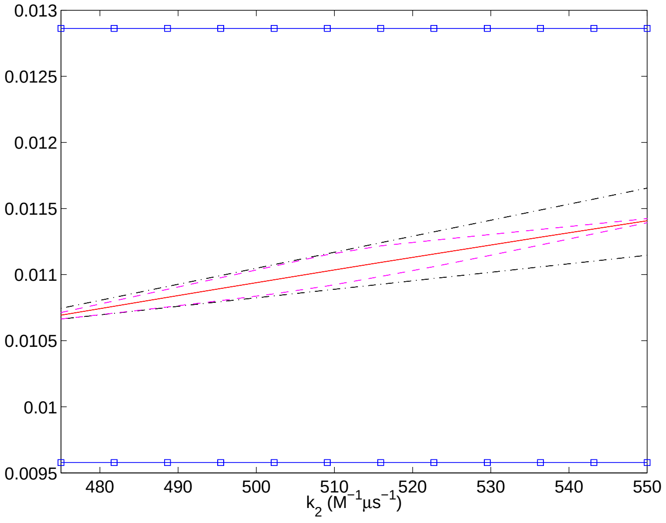

Figure 7-10: The absorbance $\mathrm{Abs}(\mathbf{x}(t_f, \cdot))$ (solid), along with affine state relaxations (dot-dashed), RAD (squares) and RPD (dashed) state relaxations, constructed on $P^*$ and plotted as functions of $k_2$ with $k_3 = 403.0$ M$^{-1}\mu$s$^{-1}$, $k_4 = 19.2$ $\mu$s$^{-1}$ and $t = 4.5$ $\mu$s fixed.

# Chapter 8

# State Relaxations for Semi-Explicit Index-One DAEs

## 8.1 Introduction

This chapter considers the computation of state relaxations for the solutions of a general system of nonlinear, semi-explicit index-one differential-algebraic equations (DAEs), which is parameterized in the governing equations and initial conditions by a real parameter vector. As in Chapter 7, there are two primary motivations for this construction. First, to provide another means to compute enclosures of the reachable sets of DAEs, in addition to the state bounding methods of Chapter 6, and second, for their use in deterministic global optimization algorithms for problems with DAEs embedded.

At present, there does not exist a fully deterministic algorithm for solving optimization problems with DAEs embedded to global optimality. In [42], problems of this type are addressed by discretizing the embedded DAEs by collocation on finite elements. This reduces the original dynamic optimization problem to a standard nonlinear program which can be solved by existing global optimization techniques. However, it was found that a fine discretization creates problems which are too large for global optimization routines to solve in reasonable time, and coarser discretization could not represent the original dynamics well enough to produce reliable results (the optimal objective value was found to depend strongly on the discretization). In [55], a method was proposed which does not require discretization. However, the method employs a sampling procedure to obtain global information concerning the embedded dynamics and hence cannot guarantee global optimality. Here, we provide two guaranteed methods for computing state relaxations for the solutions of semi-explicit DAEs. Thus, using a branch-and-bound framework as in [55] (see Chapter 1), the state relaxation methods developed here lead to a deterministic global optimization algorithm for problems with DAEs embedded.

## 8.2 Problem Statement

In this chapter, we apply the state relaxation methods developed in §7.3 and §7.4 to functions $(\mathbf{x}, \mathbf{y}) \in C^1(I, \times P, \mathbb{R}^{n_x}) \times C^1(I, \times P, \mathbb{R}^{n_y})$ that are the solutions of semi-explicit index-one systems of DAEs. The class of DAEs considered here is exactly the same as that considered in Chapter 5. The relevant assumptions are repeated here for convenience. Let $D_t \subset \mathbb{R}$, $D_p \subset \mathbb{R}^{n_p}$, $D_x \subset \mathbb{R}^{n_x}$ and $D_y \subset \mathbb{R}^{n_y}$ be open sets, and let $\mathbf{f} : D_t \times D_p \times D_x \times D_y \to \mathbb{R}^{n_x}$, $\mathbf{g} : D_t \times D_p \times D_x \times D_y \to \mathbb{R}^{n_y}$ and $\mathbf{x}_0 : D_p \to D_x$ be $C^1$ functions. Given $t_0 \in D_t$, consider the initial value problem

$$\left.\begin{aligned} \dot{\mathbf{x}}(t,\mathbf{p}) &= \mathbf{f}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p})) \\ \mathbf{0} &= \mathbf{g}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p})) \end{aligned}\right\}, \tag{8.1a}$$

$$\mathbf{x}(t_0,\mathbf{p}) = \mathbf{x}_0(\mathbf{p}). \tag{8.1b}$$

A solution of (8.1) is defined in Definition 5.3.2.

### 8.2.1 State Bounds and Related Assumptions

To derive state relaxations for some solution $(\mathbf{x}, \mathbf{y})$ of (8.1) on $I \times P$, state bounds on this solution will be required. This will be done using the single-phase method of Chapter 6. From the results there, it can be seen that successful completion of this method provides bounds and a preconditioning matrix satisfying several useful properties related to existence and uniqueness of a solution of (8.1) and invertability of $\frac{\partial \mathbf{g}}{\partial \mathbf{y}}$. These conditions are summarized in the following assumption, which holds in the remainder of the chapter.

**Assumption 8.2.1.** Let $I = [t_0, t_f] \subset D_t$ and $P \subset D_p$ be intervals. Continuous functions $\mathbf{C} : I \to \mathbb{R}^{n_y \times n_y}$, $\mathbf{x}^L, \mathbf{x}^U : I \to \mathbb{R}^{n_x}$ and $\mathbf{y}^L, \mathbf{y}^U : I \to \mathbb{R}^{n_y}$ are available and satisfy the following conditions with $X(t) \equiv [\mathbf{x}^L(t), \mathbf{x}^U(t)]$ and $Y(t) \equiv [\mathbf{y}^L(t), \mathbf{y}^U(t)]$:

1. There exists a regular solution of (8.1) on $I \times P$ satisfying

$$(\mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p})) \in X(t) \times Y(t), \quad \forall (t, \mathbf{p}) \in I \times P. \tag{8.2}$$

2. For any other solution $(\mathbf{x}^*, \mathbf{y}^*)$ of (8.1) on $I \times P$, $\mathbf{y}^*(t_0, \mathbf{p}) \notin Y(t_0)$, $\forall \mathbf{p} \in P$.

3. $X(t) \times Y(t) \subset D_x \times D_y$, $\forall t \in I$.

4. For every $t \in I$, the interval matrix $\mathbf{C}(t) \left[\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\right] (t, P, X(t), Y(t))$ does not contain any singular matrix and does not contain zero in any of its diagonal elements.

5. For every $(t, \mathbf{p}, \mathbf{z}_x) \in I \times P \times D_x$ with $\mathbf{z}_x \in X(t)$, there is a unique point $\mathbf{z}_y \in Y(t)$ such that $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$.

The bounding methods of Chapter 6 produce state bounds on a single regular solution $(\mathbf{x}, \mathbf{y})$ of (8.1), which is evident from Assumption 8.2.1. Of course, this implies that the state relaxations derived here will not be valid for *all* solutions of (8.1), but will be specific to the solution of Condition 1. For applications in which there is a single solution of interest, this has the advantage that we avoid unnecessary conservatism in the relaxations that might result from bounding multiple solutions. On the other hand, if one is interested in all possible solutions, then the relaxation method presented here would need to be combined with some procedure for exhaustively enumerating regular solutions of (8.1). We do not pursue such a procedure here, though analogous search methods for pure algebraic systems have been thoroughly studied [131]. In any case, it seems problematic to work with state relaxations that are valid for multiple solutions simultaneously, at least in the context of global optimization, since this would make the required convergence properties impossible (convergence conditions in the case of ODEs are given in [158]).

In the remainder of this chapter, the notations $I$, $P$, $(\mathbf{x}, \mathbf{y})$, $X$, $Y$ and $\mathbf{C}$ will refer to the quantities of Assumption 8.2.1.

### 8.2.2 The Auxiliary System

As in Chapter 7, state relaxations will be computed as the solutions of an auxiliary system of ODEs. Here, this system takes the form

$$\begin{aligned}
\dot{\mathbf{x}}^{cv}(t,\mathbf{p}) &= \mathbf{u}_f(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p}),\mathbf{y}^{cv}(t,\mathbf{p}),\mathbf{y}^{cc}(t,\mathbf{p})), \qquad (8.3)\\
\dot{\mathbf{x}}^{cc}(t,\mathbf{p}) &= \mathbf{o}_f(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p}),\mathbf{y}^{cv}(t,\mathbf{p}),\mathbf{y}^{cc}(t,\mathbf{p})),\\
\mathbf{y}^{cv}(t,\mathbf{p}) &= \bar{\mathbf{u}}^K_\psi(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})),\\
\mathbf{y}^{cc}(t,\mathbf{p}) &= \bar{\mathbf{o}}^K_\psi(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})),\\
\mathbf{x}^{cv}(t_0,\mathbf{p}) &= \mathbf{x}^{cv}_0(\mathbf{p}),\\
\mathbf{x}^{cc}(t_0,\mathbf{p}) &= \mathbf{x}^{cc}_0(\mathbf{p}).
\end{aligned}$$

The reason for the particular notations here will become clear in later sections.

Though this system has algebraic equations, they are explicit, so that we may consider it as a system of ODEs for $\mathbf{x}^{cv}$ and $\mathbf{x}^{cc}$. Our approach then, is to derive functions $\mathbf{u}_f$, $\mathbf{o}_f$, $\bar{\mathbf{u}}^K_\psi$ and $\bar{\mathbf{o}}^K_\psi$ such that the system (8.3) describes relaxation amplifying dynamics for $\mathbf{x}$ on $I \times P$. Then, a modification exactly analogous to that in Chapter 7 will be applied to arrive at a system describing relaxation preserving dynamics for $\mathbf{x}$ on $I \times P$. Again, this system will not be exactly of the form (8.3) and will involve state events.

In constructing appropriate functions $\mathbf{u}_f$, $\mathbf{o}_f$, $\bar{\mathbf{u}}^K_\psi$ and $\bar{\mathbf{o}}^K_\psi$, a crucial step is to compute convex and concave relaxations for the algebraic variables $\mathbf{y}$ given state relaxations for $\mathbf{x}$. This is essentially what the functions $\bar{\mathbf{u}}^K_\psi$ and $\bar{\mathbf{o}}^K_\psi$ accomplish. This then serves the added purpose of providing a means to compute state relaxations for $\mathbf{y}$ after solving (8.3).

Of course, natural McCormick extensions will play a key role in defining the required functions. Therefore, the following assumption is required throughout.

**Assumption 8.2.2.** The functions $\mathbf{x}_0$, $\mathbf{f}$, $\mathbf{g}$ and $\frac{\partial \mathbf{g}}{\partial \mathbf{y}}$ are $\mathcal{L}$-factorable with natural McCormick extensions $\mathbf{x}_0 : \mathcal{D}_0 \to \mathbb{MR}^{n_x}$, $\{\mathbf{f}\} : \mathcal{D} \to \mathbb{MR}^{n_x}$, $\{\mathbf{g}\} : \mathcal{D} \to \mathbb{MR}^{n_y}$ and $\{\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\} : \mathcal{D} \to \mathbb{MR}^{n_y \times n_y}$. Furthermore, $P$ is represented in $\mathcal{D}_0$ and the interval $[t,t] \times P \times X(t) \times Y(t)$ is represented in $\mathcal{D}$ for every $t \in I$.

## 8.3 Relaxing the Algebraic States

In this section, appropriate functions $\bar{\mathbf{u}}^K_\psi$ and $\bar{\mathbf{o}}^K_\psi$ are derived. Essentially, these functions compute relaxations of $\mathbf{y}(t, \cdot)$ on $P$, for each fixed $t \in I$, when provided with state relaxations for $\mathbf{x}$ as input. Conceptually, this is accomplished by deriving from $\mathbf{g}$ a semi-explicit expression for $\mathbf{y}$ which can be iteratively relaxed by McCormick's relaxation procedure.

As the development proceeds, we will periodically stop to illustrate the proposed methods for the simple DAEs

$$\begin{aligned} \dot{x}(t,p) &= -\frac{1}{2}(y(t,p) - \frac{1}{2}p)x(t,p), \quad x_0(p) = 1, \\ 0 &= y(t,p) - \frac{2\sin(p)}{\sqrt{y(t,p)}} - 7x(t,p), \end{aligned} \tag{8.4}$$

where $t \in I = [0, 0.2]$ and $p \in P \equiv [-1, 2.5]$. Applying, the single-phase method of Chapter 6, state bounds were computed for the unique regular solution of these DAEs satisfying the consistent initial condition $(1, 7.354) \in X(t_0) \times Y(t_0)$ for $p = 0.5$. Numerical results for this example are presented in §8.6.

### 8.3.1 Characterizing the Algebraic States

Below, the solutions of $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$ are characterized through an application of the mean value theorem. The following notation is convenient.

**Definition 8.3.1.** For all $(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) \in D_t \times D_p \times D_x \times D_y$, define the matrix $\mathbf{M}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) \equiv \mathbf{C}(t)\frac{\partial \mathbf{g}}{\partial \mathbf{y}}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y)$, and denote the elements of $\mathbf{M}$ by $m_{ij}$.

**Definition 8.3.2.** Let $\tilde{\mathbf{y}} : I \times P \to \mathbb{R}^{n_y}$ denote an arbitrary function that is affine on $P$ for every fixed $t \in I$ and satisfies $\tilde{\mathbf{y}}(t, \mathbf{p}) \in Y(t)$, $\forall (t, \mathbf{p}) \in I \times P$.

The function $\tilde{\mathbf{y}}$ is essentially used as a reference point in the application of the mean value theorem below. The assumption that it is affine on $P$ for every fixed $t \in I$ is not required for this purpose, but is required for the relaxation scheme described in §8.3.2.

**Theorem 8.3.3.** *Let $\mathbf{z}_x : I \times P \to \mathbb{R}^{n_x}$ satisfy $\mathbf{z}_x(t, \mathbf{p}) \in X(t)$, $\forall (t, \mathbf{p}) \in I \times P$, and let $\mathbf{z}_y : I \times P \to \mathbb{R}^{n_y}$ be the unique function satisfying $\mathbf{z}_y(t, \mathbf{p}) \in Y(t)$ and $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x(t, \mathbf{p}), \mathbf{z}_y(t, \mathbf{p})) = \mathbf{0}$, $\forall (t, \mathbf{p}) \in I \times P$ (see Condition 5 of Assumption 8.2.1). There exists $\boldsymbol{\lambda} : I \times P \to [0, 1]^{n_y}$ such that the definition $\boldsymbol{\xi}^i(t, \mathbf{p}) \equiv \tilde{\mathbf{y}}(t, \mathbf{p}) + \lambda_i(t, \mathbf{p})(\mathbf{z}_y(t, \mathbf{p}) - \tilde{\mathbf{y}}(t, \mathbf{p}))$ satisfies*

$$\begin{aligned} z_{y,i}(t, \mathbf{p}) = \tilde{y}_i(t, \mathbf{p}) - \frac{1}{m_{ii}(t, \mathbf{p}, \mathbf{z}_x(t, \mathbf{p}), \boldsymbol{\xi}^i(t, \mathbf{p}))} \Big[ & \mathbf{C}_i(t)\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x(t, \mathbf{p}), \tilde{\mathbf{y}}(t, \mathbf{p})) \\ & + \sum_{j \neq i} m_{ij}(t, \mathbf{p}, \mathbf{z}_x(t, \mathbf{p}), \boldsymbol{\xi}^i(t, \mathbf{p}))(z_{y,j}(t, \mathbf{p}) - \tilde{y}_j(t, \mathbf{p})) \Big], \end{aligned} \tag{8.5}$$

*for all $(t, \mathbf{p}) \in I \times P$ and every $i \in \{1, \ldots, n_y\}$, where $\mathbf{C}_i$ denotes the $i^{\text{th}}$ row of $\mathbf{C}$.*

*Proof.* Fix any $(t, \mathbf{p}) \in I \times P$ and note that $\mathbf{z}_y(t, \mathbf{p}), \tilde{\mathbf{y}}(t, \mathbf{p}) \in Y(t)$. Since $\mathbf{g} \in C^1(D_t \times D_p \times D_x \times D_y, \mathbb{R}^{n_y})$, it is clear that $\mathbf{C}_i(t)\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x(t, \mathbf{p}), \cdot)$ is differentiable on $D_y$. Since $Y(t) \subset D_y$ is convex, the mean value theorem asserts that there exists $\lambda_i(t, \mathbf{p}) \in [0, 1]$ such that $\boldsymbol{\xi}^i(t, \mathbf{p}) \equiv \tilde{\mathbf{y}}(t, \mathbf{p}) + \lambda_i(t, \mathbf{p})(\mathbf{z}_y(t, \mathbf{p}) - \tilde{\mathbf{y}}(t, \mathbf{p}))$ satisfies

$$\begin{aligned} & \mathbf{C}_i(t)\Big(\frac{\partial \mathbf{g}}{\partial \mathbf{y}}(t, \mathbf{p}, \mathbf{z}_x(t, \mathbf{p}), \boldsymbol{\xi}^i(t, \mathbf{p}))\Big)(\mathbf{z}_y(t, \mathbf{p}) - \tilde{\mathbf{y}}(t, \mathbf{p})) \\ & \qquad = \mathbf{C}_i(t)\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x(t, \mathbf{p}), \mathbf{z}_y(t, \mathbf{p})) - \mathbf{C}_i(t)\mathbf{g}(t, \mathbf{p}, \mathbf{z}_y(t, \mathbf{p}), \tilde{\mathbf{y}}(t, \mathbf{p})), \\ & \qquad = -\mathbf{C}_i(t)\mathbf{g}(t, \mathbf{p}, \mathbf{z}_y(t, \mathbf{p}), \tilde{\mathbf{y}}(t, \mathbf{p})), \end{aligned}$$

where the last equality follows from the fact that $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x(t, \mathbf{p}), \mathbf{z}_y(t, \mathbf{p})) = \mathbf{0}$. This is equivalent to

$$\mathbf{M}_i(t, \mathbf{p}, \mathbf{z}_x(t, \mathbf{p}), \boldsymbol{\xi}^i(t, \mathbf{p}))(\mathbf{z}_y(t, \mathbf{p}) - \tilde{\mathbf{y}}(t, \mathbf{p})) = -\mathbf{C}_i(t)\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x(t, \mathbf{p}), \tilde{\mathbf{y}}(t, \mathbf{p})),$$

where $\mathbf{M}_i$ denotes the $i^{\text{th}}$ row of $\mathbf{M}$. Since $m_{ii}(t, \mathbf{p}, \mathbf{z}_x(t, \mathbf{p}), \boldsymbol{\xi}^i(t, \mathbf{p})) \neq 0$ by Condition 4 of Assumption 8.2.1, $z_{y,i}(t, \mathbf{p})$ can be isolated on the left to give (8.5). Then, it has been shown that, for arbitrary, $(t, \mathbf{p})$ and $i$, there exists $\lambda_i(t, \mathbf{p}) \in [0, 1]$ such that (8.5) is satisfied with $\boldsymbol{\xi}^i(t, \mathbf{p}) \equiv \tilde{\mathbf{y}}(t, \mathbf{p}) + \lambda_i(t, \mathbf{p})(\mathbf{z}_y(t, \mathbf{p}) - \tilde{\mathbf{y}}(t, \mathbf{p}))$. Accordingly, $\exists \boldsymbol{\lambda} : I \times P \to [0, 1]^{n_y}$ satisfying the theorem. □

The following definitions simplify Theorem 8.3.3 to give Corollary 8.3.6. This Corollary provides the characterization of $\mathbf{y}$ required to construct the desired relaxations.

**Definition 8.3.4.** Define the set

$$\Phi^+ \equiv \{(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y, \tilde{\mathbf{z}}_y, \boldsymbol{\lambda}, \hat{\mathbf{C}}) : (t, \mathbf{p}, \boldsymbol{\lambda}) \in I \times P \times [0, 1]^{n_y},\ \mathbf{z}_x \in X(t),\\ \mathbf{z}_y, \tilde{\mathbf{z}}_y \in Y(t),\ \hat{\mathbf{C}} = \mathbf{C}(t)\}.$$

**Definition 8.3.5.** Define $\boldsymbol{\psi} : \Phi^+ \to \mathbb{R}^{n_y}$ elementwise, for each $i \in \{1, \ldots, n_y\}$, by

$$\psi_i(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y, \tilde{\mathbf{z}}_y, \boldsymbol{\lambda}, \hat{\mathbf{C}}) = \tilde{z}_{y,i} - \frac{1}{m_{ii}(t, \mathbf{p}, \mathbf{z}_x, \boldsymbol{\xi}^i)}\Bigg[\hat{\mathbf{C}}_i\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \tilde{\mathbf{z}}_y)\\ + \sum_{j \neq i} m_{ij}(t, \mathbf{p}, \mathbf{z}_x, \boldsymbol{\xi}^i)(z_{y,j} - \tilde{z}_{y,j})\Bigg], \tag{8.6}$$

where $\boldsymbol{\xi}^i \equiv \tilde{\mathbf{z}}_y + \lambda_i(\mathbf{z}_y - \tilde{\mathbf{z}}_y)$ and $m_{ij}$ is the $ij^{\text{th}}$ element of $\hat{\mathbf{C}}\frac{\partial \mathbf{g}}{\partial \mathbf{y}}$.

**Corollary 8.3.6.** *Let* $\mathbf{z}_x : I \times P \to \mathbb{R}^{n_x}$ *satisfy* $\mathbf{z}_x(t, \mathbf{p}) \in X(t)$, $\forall (t, \mathbf{p}) \in I \times P$, *and let* $\mathbf{z}_y : I \times P \to \mathbb{R}^{n_y}$ *be the unique function satisfying* $\mathbf{z}_y(t, \mathbf{p}) \in Y(t)$ *and* $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x(t, \mathbf{p}), \mathbf{z}_y(t, \mathbf{p})) = \mathbf{0}$, $\forall (t, \mathbf{p}) \in I \times P$. *There exists* $\boldsymbol{\lambda} : I \times P \to [0, 1]^{n_y}$ *such that*

$$\mathbf{z}_y(t, \mathbf{p}) = \boldsymbol{\psi}(t, \mathbf{p}, \mathbf{z}_x(t, \mathbf{p}), \mathbf{z}_y(t, \mathbf{p}), \tilde{\mathbf{y}}(t, \mathbf{p}), \boldsymbol{\lambda}(t, \mathbf{p}), \mathbf{C}(t)), \quad \forall (t, \mathbf{p}) \in I \times P. \tag{8.7}$$

*In particular, there exists* $\boldsymbol{\lambda} : I \times P \to [0,1]^{n_y}$ *such that*

$$\mathbf{y}(t,\mathbf{p}) = \boldsymbol{\psi}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p}),\tilde{\mathbf{y}}(t,\mathbf{p}),\boldsymbol{\lambda}(t,\mathbf{p}),\mathbf{C}(t)), \quad \forall(t,\mathbf{p}) \in I \times P. \tag{8.8}$$

**Example 8.3.1.** Consider the algebraic equation in (8.4), defined by the function

$$g(t,p,z_x,z_y) = z_y - \frac{2\sin(p)}{\sqrt{z_y}} - 7z_x.$$

Differentiating and applying the mean-value theorem as in Theorem 8.3.3, $\psi$ is given by

$$\psi(t,p,z_x,z_y,\tilde{z}_y,\lambda,\hat{c}) = \tilde{z}_y - \frac{1}{1+\frac{\sin(p)}{(\tilde{z}_y+\lambda(z_y-\tilde{z}_y))^{\frac{3}{2}}}}\left(\tilde{z}_y - \frac{2\sin(p)}{\sqrt{\tilde{z}_y}} - 7z_x\right). \tag{8.9}$$

Note that the matrix $\mathbf{C}$ of Assumption 8.2.1 is $1 \times 1$ in this case and therefore cancels out in the definition of $\psi$. In order to characterize the solution $y$ through Corollary 8.3.6, a reference trajectory $\tilde{y}$ must be specified. For the numerical results in §8.6, $\tilde{y}(t,p) = 0.5(y^L(t) + y^U(t))$ was chosen for all $t \in I$.

### 8.3.2 An Iterative Relaxation Scheme

The characterization of $\mathbf{y}$ given in Corollary 8.3.6 can be used to relax $\mathbf{y}$ by an iterative scheme. First, note that Definition 8.3.5 and Assumption 8.2.2 guarantee that $\boldsymbol{\psi}$ is $\mathcal{L}$-factorable. Let $\{\boldsymbol{\psi}\} : \mathcal{D}_\psi \to \mathbb{MR}^{n_y}$ be a natural McCormick extension. Since the set $[t,t] \times P \times X(t) \times Y(t)$ is represented in $\mathcal{D}$ by Assumption 8.2.2, it follows that $[t,t] \times P \times X(t) \times Y(t) \times \tilde{Y}(t) \times [\mathbf{0},\mathbf{1}] \times [\mathbf{C}(t),\mathbf{C}(t)]$ is represented in $\mathcal{D}_\psi$ provided that no division by an interval containing zero occurs. But by Condition 4 of Assumption 8.2.1, such a division is impossible.

Evaluating the natural McCormick extension of $\boldsymbol{\psi}$ requires bounds on all of its arguments. By definition, $Y(t)$ is a valid bound on the reference point $\tilde{\mathbf{y}}(t,\mathbf{p})$, $\forall\mathbf{p} \in P$. However, sharper bounds will usually be available, as in Example 8.3.1 above, where the reference point is constant with respect to $\mathbf{p}$. Therefore, we define independent bounds for $\tilde{\mathbf{y}}$ below.

**Definition 8.3.7.** Let $\tilde{\mathbf{y}}^L, \tilde{\mathbf{y}}^U : I \to \mathbb{R}^{n_y}$ be functions satisfying $\tilde{\mathbf{y}}(t,\mathbf{p}) \in \tilde{Y}(t) \equiv [\tilde{\mathbf{y}}^L(t), \tilde{\mathbf{y}}^U(t)] \subset Y(t)$, $\forall (t,\mathbf{p}) \in I \times P$.

**Definition 8.3.8.** Let $\mathbf{u}_\psi, \mathbf{o}_\psi : I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} \to \mathbb{R}^{n_y}$ be defined by

$$\begin{aligned}
\mathbf{u}_\psi(t,\mathbf{p},\mathbf{z}_x^{cv},\mathbf{z}_x^{cc},\mathbf{z}_y^{cv},\mathbf{z}_y^{cc}) &= \max(\mathbf{z}_y^{cv}, \{\boldsymbol{\psi}\}^{cv}(\mathcal{T},\mathcal{P},\mathcal{X},\mathcal{Y},\tilde{\mathcal{Y}},\mathcal{L},\mathcal{C}))\\
\mathbf{o}_\psi(t,\mathbf{p},\mathbf{z}_x^{cv},\mathbf{z}_x^{cc},\mathbf{z}_y^{cv},\mathbf{z}_y^{cc}) &= \min(\mathbf{z}_y^{cc}, \{\boldsymbol{\psi}\}^{cc}(\mathcal{T},\mathcal{P},\mathcal{X},\mathcal{Y},\tilde{\mathcal{Y}},\mathcal{L},\mathcal{C}))
\end{aligned}$$

where

$$\begin{aligned}
&\mathcal{X} = \mathrm{MC}(\mathbf{x}^L(t), \mathbf{x}^U(t), \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}) && \mathcal{T} = \mathrm{MC}(t,t,t,t),\\
&\mathcal{Y} = \mathrm{MC}(\mathbf{y}^L(t), \mathbf{y}^U(t), \mathbf{z}_y^{cv}, \mathbf{z}_y^{cc}), && \mathcal{P} = \mathrm{MC}(\mathbf{p},\mathbf{p},\mathbf{p},\mathbf{p}),\\
&\tilde{\mathcal{Y}} = \mathrm{MC}(\tilde{\mathbf{y}}^L(t), \tilde{\mathbf{y}}^U(t), \tilde{\mathbf{y}}(t,\mathbf{p}), \tilde{\mathbf{y}}(t,\mathbf{p})), && \mathcal{L} = \mathrm{MC}(\mathbf{0},\mathbf{1},\mathbf{0},\mathbf{1}),\\
&\mathcal{C} = \mathrm{MC}(\mathbf{C}(t),\mathbf{C}(t),\mathbf{C}(t),\mathbf{C}(t)).
\end{aligned}$$

It will be shown that the previous definition provides a means to compute state relaxations for $\mathbf{y}$ on $I \times P$ as a refinement of the state bounds. In particular, the following theorem holds.

**Theorem 8.3.9.** *Let $\mathbf{x}^{cv}, \mathbf{x}^{cc} : I \times P \to \mathbb{R}^{n_x}$ be state relaxations for $\mathbf{x}$ on $I \times P$. Then state relaxations for $\mathbf{y}$ on $I \times P$, $\mathbf{y}^{cv}, \mathbf{y}^{cc} : I \times P \to \mathbb{R}^{n_y}$ are given by the definitions*

$$\begin{aligned}
\mathbf{y}^{cv}(t,\mathbf{p}) &= \mathbf{u}_\psi(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p}),\mathbf{y}^L(t),\mathbf{y}^U(t)),\\
\mathbf{y}^{cc}(t,\mathbf{p}) &= \mathbf{o}_\psi(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p}),\mathbf{y}^L(t),\mathbf{y}^U(t)).
\end{aligned}$$

This theorem is proven through a series of more fundamental lemmas that will be required to show that (8.3) satisfies the conditions of relaxation amplifying dynamics in 8.4.

**Lemma 8.3.10.** *Let $(t,\mathbf{p},\mathbf{z}_x) \in I \times P \times \mathbb{R}^{n_x}$ satisfy $\mathbf{z}_x \in X(t)$, and let $\mathbf{z}_y \in \mathbb{R}^{n_y}$ be the unique point satisfying $\mathbf{z}_y \in Y(t)$ and $\mathbf{g}(t,\mathbf{p},\mathbf{z}_x,\mathbf{z}_y) = \mathbf{0}$. For any $\mathbf{z}_x^{cv}, \mathbf{z}_x^{cc} \in \mathbb{R}^{n_x}$*

*and* $\mathbf{z}_y^{cv}, \mathbf{z}_y^{cc} \in \mathbb{R}^{n_y}$ *satisfying* $\mathbf{z}_x^{cv} \leq \mathbf{z}_x \leq \mathbf{z}_x^{cc}$ *and* $\mathbf{z}_y^{cv} \leq \mathbf{z}_y \leq \mathbf{z}_y^{cc}$*,*

$$\mathbf{u}_\psi(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}, \mathbf{z}_y^{cv}, \mathbf{z}_y^{cc}) \leq \mathbf{z}_y \leq \mathbf{o}_\psi(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}, \mathbf{z}_y^{cv}, \mathbf{z}_y^{cc}). \tag{8.10}$$

*Proof.* Choose any $(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) \in I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_y}$ as in the statement. By Corollary 8.3.6, there exists $\boldsymbol{\lambda} \in [0,1]^{n_y}$ such that

$$\mathbf{z}_y = \boldsymbol{\psi}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y, \tilde{\mathbf{y}}(t, \mathbf{p}), \boldsymbol{\lambda}, \mathbf{C}(t)). \tag{8.11}$$

Choose any $\mathbf{z}_x^{cv}, \mathbf{z}_x^{cc} \in \mathbb{R}^{n_x}$ and $\mathbf{z}_y^{cv}, \mathbf{z}_y^{cc} \in \mathbb{R}^{n_y}$ and suppose that $\mathbf{z}_x^{cv} \leq \mathbf{z}_x \leq \mathbf{z}_x^{cc}$ and $\mathbf{z}_y^{cv} \leq \mathbf{z}_y \leq \mathbf{z}_y^{cc}$. Using the notation of Definition 8.3.8, define

$$\begin{aligned} \mathbf{u}'_\psi(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}, \mathbf{z}_y^{cv}, \mathbf{z}_y^{cc}) &\equiv \{\psi\}^{cv}(\mathcal{T}, \mathcal{P}, \mathcal{X}, \mathcal{Y}, \tilde{\mathcal{Y}}, \mathcal{L}, \mathcal{C}), \\ \mathbf{o}'_\psi(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}, \mathbf{z}_y^{cv}, \mathbf{z}_y^{cc}) &\equiv \{\psi\}^{cc}(\mathcal{T}, \mathcal{P}, \mathcal{X}, \mathcal{Y}, \tilde{\mathcal{Y}}, \mathcal{L}, \mathcal{C}). \end{aligned} \tag{8.12}$$

That is, $\mathbf{u}'_\psi$ and $\mathbf{o}'_\psi$ are the same as $\mathbf{u}_\psi$ and $\mathbf{o}_\psi$ up to the application of the min and max functions. By hypothesis,

$$\mathbf{z}_x \in X(t) \cap [\mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}], \quad \mathbf{z}_y \in Y(t) \cap [\mathbf{z}_y^{cv}, \mathbf{z}_y^{cc}]. \tag{8.13}$$

Additionally, we make the trivial observations

$$\begin{aligned} &t \in [t,t] \cap [t,t], && \tilde{\mathbf{y}}(t,\mathbf{p}) \in \tilde{Y}(t) \cap [\tilde{\mathbf{y}}(t,\mathbf{p}), \tilde{\mathbf{y}}(t,\mathbf{p})], \\ &\mathbf{p} \in P \cap [\mathbf{p}, \mathbf{p}], && \mathbf{C}(t) \in [\mathbf{C}(t), \mathbf{C}(t)] \cap [\mathbf{C}(t), \mathbf{C}(t)], \\ &\boldsymbol{\lambda} \in [\mathbf{0}, \mathbf{1}] \cap [\mathbf{0}, \mathbf{1}]. \end{aligned} \tag{8.14}$$

By Assumption 8.2.2 and Condition 4 of Assumption 8.2.1, the interval

$$[t,t] \times P \times X(t) \times Y(t) \times \tilde{Y}(t) \times [\mathbf{0}, \mathbf{1}] \times [\mathbf{C}(t), \mathbf{C}(t)], \tag{8.15}$$

is represented in $\mathcal{D}_\psi$. Then, by (8.13) and (8.14), we may apply Lemma 2.7.3 with

$$\begin{aligned}\mathbf{x} &:= (t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y, \tilde{\mathbf{y}}(t, \mathbf{p}), \boldsymbol{\lambda}, \mathbf{C}(t)), \\ \mathbf{x}^{cv} &:= (t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_y^{cv}, \tilde{\mathbf{y}}(t, \mathbf{p}), \mathbf{0}, \mathbf{C}(t)), \\ \mathbf{x}^{cc} &:= (t, \mathbf{p}, \mathbf{z}_x^{cc}, \mathbf{z}_y^{cc}, \tilde{\mathbf{y}}(t, \mathbf{p}), \mathbf{1}, \mathbf{C}(t)).\end{aligned} \tag{8.16}$$

This gives

$$\mathbf{u}'_\psi(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}, \mathbf{z}_y^{cv}, \mathbf{z}_y^{cc}) \le \boldsymbol{\psi}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y, \tilde{\mathbf{y}}(t, \mathbf{p}), \boldsymbol{\lambda}, \mathbf{C}(t)) \le \mathbf{o}'_\psi(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}, \mathbf{z}_y^{cv}, \mathbf{z}_y^{cc}).$$

Combining this with (8.11) yields

$$\mathbf{u}'_\psi(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}, \mathbf{z}_y^{cv}, \mathbf{z}_y^{cc}) \le \mathbf{z}_y \le \mathbf{o}'_\psi(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}, \mathbf{z}_y^{cv}, \mathbf{z}_y^{cc}). \tag{8.17}$$

Combining this with $\mathbf{z}_y^{cv} \le \mathbf{z}_y \le \mathbf{z}_y^{cc}$ gives (8.10). □

**Lemma 8.3.11.** *Let $\mathbf{z}_x : I \times P \to \mathbb{R}^{n_x}$ and $\mathbf{z}_y : I \times P \to \mathbb{R}^{n_y}$ satisfy $\mathbf{z}_x(t, \mathbf{p}) \in X(t)$ and $\mathbf{z}_y(t, \mathbf{p}) \in Y(t)$, $\forall (t, \mathbf{p}) \in I \times P$. Let $\mathbf{z}_x^{cv}, \mathbf{z}_x^{cc} : I \times P \to \mathbb{R}^{n_x}$ and $\mathbf{z}_y^{cv}, \mathbf{z}_y^{cc} : I \times P \to \mathbb{R}^{n_y}$ and choose any $(\lambda, \mathbf{p}_1, \mathbf{p}_2) \in (0, 1) \times P \times P$. For any $t \in I$, if*

1. *$\mathbf{z}_x^{cv}(t, \cdot)$ and $\mathbf{z}_y^{cv}(t, \cdot)$ are consistent with convexity at $(\lambda, \mathbf{p}_1, \mathbf{p}_2)$,*
2. *$\mathbf{z}_x^{cc}(t, \cdot)$ and $\mathbf{z}_y^{cc}(t, \cdot)$ are consistent with concavity at $(\lambda, \mathbf{p}_1, \mathbf{p}_2)$,*
3. *$\mathbf{z}_x^{cv}(t, \mathbf{q}) \le \mathbf{z}_x(t, \mathbf{q}) \le \mathbf{z}_x^{cc}(t, \mathbf{q})$ and $\mathbf{z}_y^{cv}(t, \mathbf{q}) \le \mathbf{z}_y(t, \mathbf{q}) \le \mathbf{z}_y^{cc}(t, \mathbf{q})$ for all $\mathbf{q} \in \{\mathbf{p}_1, \mathbf{p}_2, \lambda\mathbf{p}_1 + (1 - \lambda)\mathbf{p}_2\}$,*

*then the functions*

$$\begin{aligned}P \ni \mathbf{p} &\longmapsto \mathbf{u}_\psi(t, \mathbf{p}, \mathbf{z}_x^{cv}(t, \mathbf{p}), \mathbf{z}_x^{cc}(t, \mathbf{p}), \mathbf{z}_y^{cv}(t, \mathbf{p}), \mathbf{z}_y^{cc}(t, \mathbf{p})), \\ P \ni \mathbf{p} &\longmapsto \mathbf{o}_\psi(t, \mathbf{p}, \mathbf{z}_x^{cv}(t, \mathbf{p}), \mathbf{z}_x^{cc}(t, \mathbf{p}), \mathbf{z}_y^{cv}(t, \mathbf{p}), \mathbf{z}_y^{cc}(t, \mathbf{p})),\end{aligned} \tag{8.18}$$

*are, respectively, consistent with convexity and consistent with concavity at $(\lambda, \mathbf{p}_1, \mathbf{p}_2)$.*

*Proof.* Choose any $(\lambda, \mathbf{p}_1, \mathbf{p}_2) \in (0,1) \times P \times P$, define $\mathbf{p}_3 \equiv \lambda\mathbf{p}_1 + (1-\lambda)\mathbf{p}_2$, and suppose that $t \in I$ is such that Conditions 1-3 hold.

By Assumption 8.2.2 and Condition 4 of Assumption 8.2.1, the interval

$$X := [t,t] \times P \times X(t) \times Y(t) \times \tilde{Y}(t) \times [\mathbf{0}, \mathbf{1}] \times [\mathbf{C}(t), \mathbf{C}(t)], \tag{8.19}$$

is represented in $\mathcal{D}_\psi$. We will apply Lemma 2.7.4 with

$$\begin{aligned}
\mathbf{x}_i^{cv} &:= (t, \mathbf{p}_i, \mathbf{z}_x^{cv}(t,\mathbf{p}_i), \mathbf{z}_y^{cv}(t,\mathbf{p}_i), \tilde{\mathbf{y}}(t,\mathbf{p}_i), \mathbf{0}, \mathbf{C}(t)), \\
\mathbf{x}_i^{cc} &:= (t, \mathbf{p}_i, \mathbf{z}_x^{cc}(t,\mathbf{p}_i), \mathbf{z}_y^{cc}(t,\mathbf{p}_i), \tilde{\mathbf{y}}(t,\mathbf{p}_i), \mathbf{1}, \mathbf{C}(t)),
\end{aligned}$$

and $i \in \{1,2,3\}$. To verify the hypotheses of that lemma, we first show that $X \cap [\mathbf{x}_i^{cv}, \mathbf{x}_i^{cc}] \neq \emptyset$ for all $i \in \{1,2,3\}$. By hypothesis,

$$X(t) \cap [\mathbf{z}_x^{cv}(t,\mathbf{p}_i), \mathbf{z}_x^{cc}(t,\mathbf{p}_i)] \neq \emptyset, \quad Y(t) \cap [\mathbf{z}_y^{cv}(t,\mathbf{p}_i), \mathbf{z}_y^{cc}(t,\mathbf{p}_i)] \neq \emptyset, \tag{8.20}$$

for all $i \in \{1,2,3\}$, because these intervals contain $\mathbf{z}_x(t,\mathbf{p}_i)$ and $\mathbf{z}_y(t,\mathbf{p}_i)$, respectively. Additionally, we make the trivial observations

$$\begin{aligned}
&[t,t] \cap [t,t] \neq \emptyset, && \tilde{Y}(t) \cap [\tilde{\mathbf{y}}(t,\mathbf{p}_i), \tilde{\mathbf{y}}(t,\mathbf{p}_i)] \neq \emptyset, \\
&P \cap [\mathbf{p}_i, \mathbf{p}_i] \neq \emptyset, && [\mathbf{C}(t), \mathbf{C}(t)] \cap [\mathbf{C}(t), \mathbf{C}(t)] \neq \emptyset, \\
&[\mathbf{0}, \mathbf{1}] \cap [\mathbf{0}, \mathbf{1}] \neq \emptyset,
\end{aligned} \tag{8.21}$$

for all $i \in \{1,2,3\}$.

Now, when considered as functions on $P$, the constant values $t$, $\mathbf{0}$, $\mathbf{1}$ and $\mathbf{C}(t)$, as well as the identity function $\mathbf{p}$, are all both consistent with convexity and consistent with concavity at $(\lambda, \mathbf{p}_1, \mathbf{p}_2)$. Moreover, by the assumption that $\tilde{\mathbf{y}}(t,\cdot)$ is affine on $P$, it is also both consistent with convexity and consistent with concavity at $(\lambda, \mathbf{p}_1, \mathbf{p}_2)$. Combining this with Hypotheses 1 and 2 of the present lemma, Theorem 2.7.4 now shows that

$$\begin{aligned} &P \ni \mathbf{p} \longmapsto \mathbf{u}'_\psi(t,\mathbf{p},\mathbf{z}_x^{cv}(t,\mathbf{p}),\mathbf{z}_x^{cc}(t,\mathbf{p}),\mathbf{z}_y^{cv}(t,\mathbf{p}),\mathbf{z}_y^{cc}(t,\mathbf{p})), \qquad (8.22)\\ &P \ni \mathbf{p} \longmapsto \mathbf{o}'_\psi(t,\mathbf{p},\mathbf{z}_x^{cv}(t,\mathbf{p}),\mathbf{z}_x^{cc}(t,\mathbf{p}),\mathbf{z}_y^{cv}(t,\mathbf{p}),\mathbf{z}_y^{cc}(t,\mathbf{p})), \end{aligned}$$

are, respectively, consistent with convexity and consistent with concavity at $(\lambda,\mathbf{p}_1,\mathbf{p}_2)$ ($\mathbf{u}'_\psi$ and $\mathbf{o}'_\psi$ are defined as in (8.12)).

Since the min of two convex functions is convex, Definition 8.3.8, Hypothesis 1 and (8.22) imply that

$$P \ni \mathbf{p} \longmapsto \mathbf{u}_\psi(t,\mathbf{p},\mathbf{z}_x^{cv}(t,\mathbf{p}),\mathbf{z}_x^{cc}(t,\mathbf{p}),\mathbf{z}_y^{cv}(t,\mathbf{p}),\mathbf{z}_y^{cc}(t,\mathbf{p})) \qquad (8.23)$$

is consistent with convexity at $(\lambda,\mathbf{p}_1,\mathbf{p}_2)$. Arguing analogously for $\mathbf{o}_\psi$, this proves the lemma. □

**Lemma 8.3.12.** *$\mathbf{u}_\psi$ and $\mathbf{o}_\psi$ are continuous on $I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_y}$. Moreover, $\exists L \in \mathbb{R}_+$ such that*

$$\begin{aligned} &\|\mathbf{u}_\psi(t,\mathbf{p},\mathbf{z}_x^{cv},\mathbf{z}_x^{cc},\mathbf{z}_y^{cv},\mathbf{z}_y^{cc}) - \mathbf{u}_\psi(t,\mathbf{p},\hat{\mathbf{z}}_x^{cv},\hat{\mathbf{z}}_x^{cc},\hat{\mathbf{z}}_y^{cv},\hat{\mathbf{z}}_y^{cc})\|_\infty\\ &+\|\mathbf{o}_\psi(t,\mathbf{p},\mathbf{z}_x^{cv},\mathbf{z}_x^{cc},\mathbf{z}_y^{cv},\mathbf{z}_y^{cc}) - \mathbf{o}_\psi(t,\mathbf{p},\hat{\mathbf{z}}_x^{cv},\hat{\mathbf{z}}_x^{cc},\hat{\mathbf{z}}_y^{cv},\hat{\mathbf{z}}_y^{cc})\|_\infty\\ &\leq L(\|\mathbf{z}_x^{cv}-\hat{\mathbf{z}}_x^{cv}\|_\infty + \|\mathbf{z}_x^{cc}-\hat{\mathbf{z}}_x^{cc}\|_\infty + \|\mathbf{z}_y^{cv}-\hat{\mathbf{z}}_y^{cv}\|_\infty + \|\mathbf{z}_y^{cc}-\hat{\mathbf{z}}_y^{cc}\|_\infty) \end{aligned}$$

*for all $(t,\mathbf{p},\mathbf{z}_x^{cv},\mathbf{z}_x^{cc},\hat{\mathbf{z}}_x^{cv},\hat{\mathbf{z}}_x^{cc},\mathbf{z}_y^{cv},\mathbf{z}_y^{cc},\hat{\mathbf{z}}_y^{cv},\hat{\mathbf{z}}_y^{cc}) \in I \times P \times \mathbb{R}^{4n_x} \times \mathbb{R}^{4n_y}$.*

*Proof.* This assertions follows from Corollary 2.7.8 by a construction exactly analogous to the proof of Lemma 7.6.7. □

Theorem 8.3.9 can now be proven by noting that, for any $t \in I$, $\mathbf{y}^L(t)$ and $\mathbf{y}^U(t)$ are, respectively convex and concave relaxations of $\mathbf{y}(t,\cdot)$ on $P$. Then, the conclusion follows at once from Lemmas 8.3.10 and 8.3.11. Thus, Theorem 8.3.9 gives a simple method for refining the state bounds $\mathbf{y}^L$ and $\mathbf{y}^U$ to obtain state relaxations for $\mathbf{y}$. An example of this construction is given in Example 8.3.2 below.

**Example 8.3.2.** Recall the definition of $\psi$ in (8.9) for the example DAE (8.4). The construction of $u_\psi$ and $o_\psi$ for this example is shown here. Note, however, that this is only illustrative. The entire procedure can be easily automated using operator overloading, as in `MC++` (http://www3.imperial.ac.uk/people/b.chachuat/research).

First, bounds on the reference trajectory, $\tilde{y}^L$ and $\tilde{y}^U$, are required as per Definition 8.3.7. Since $\tilde{y}$ was chosen to be constant with respect to $p$ in Example 8.3.1, we choose $\tilde{y}^L(t) = \tilde{y}^U(t) = \tilde{y}(t)$, $\forall t \in I$. Furthermore, note that $\tilde{y}(t, \cdot)$ is trivially affine for each $t \in I$.

For any $(t, p, z_x^{cv}, z_x^{cc}, z_y^{cv}, z_y^{cc}) \in I \times P \times \mathbb{R}^4$, appropriate values for $u_\psi$ and $o_\psi$ are computed by evaluating the natural McCormick extension of $\psi$ with the initializations given in Definition 8.3.8. This is implemented by the factorization of $\psi$ shown in Table 8.1 with factors $v_k$, inclusion factors $V_k$, and relaxation factors, $\mathcal{V}_k$. The values $u_\psi(t, p, z_x^{cv}, z_x^{cc}, z_y^{cv}, z_y^{cc})$ and $o_\psi(t, p, z_x^{cv}, z_x^{cc}, z_y^{cv}, z_y^{cc})$ are given by $\mathcal{V}_{25}^{cv}$ and $\mathcal{V}_{25}^{cc}$ in Table 8.1, respectively.

The state relaxations given by Theorem 8.3.9 can obviously be refined iteratively. That is, sequences of state relaxations for $\mathbf{y}$, $\{\mathbf{y}^{cv,k}\}$ and $\{\mathbf{y}^{cc,k}\}$, can be computed by recursive application of $\mathbf{u}_\psi$ and $\mathbf{o}_\psi$. However, a direct recursive application of $\mathbf{u}_\psi$ and $\mathbf{o}_\psi$ is not the most efficient way to accomplish such an iterative refinement. In particular, this would update each $y_j^{cv,k+1}$ and $y_j^{cc,k+1}$ based on the relaxations $\mathbf{y}^{cv,k}$ and $\mathbf{y}^{cc,k}$, regardless of $j$. However, if the sequence of computations updates $y_1^{cv,k}$ and $y_1^{cc,k}$ before $y_2^{cv,k}$ and $y_2^{cc,k}$, for example, then the updated relaxations $y_1^{cv,k+1}$ and $y_1^{cc,k+1}$ can be used in the subsequent computation of $y_2^{cv,k+1}$ and $y_2^{cc,k+1}$. This accelerated updating scheme is analogous to that used in the Gauss-Seidel algorithm for iteratively solving systems of equations. The following functions describes this procedure.

**Definition 8.3.13.** For any $K \in \mathbb{N}$, define the functions $\bar{\mathbf{u}}_\psi^K, \bar{\mathbf{o}}_\psi^K : I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_y}$ by

$$\bar{\mathbf{u}}_\psi^K(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}) \equiv \mathbf{z}_y^{cv,K} \quad \text{and} \quad \bar{\mathbf{o}}_\psi^K(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}) \equiv \mathbf{z}_y^{cc,K},$$

Table 8.1: Factorization and computation of $\psi$ at $(t, p, z_x, z_y, \tilde{z}_y, \lambda)$ and $u_\psi$ and $o_\psi$ at $(t, p, z_x^{cv}, z_x^{cc}, z_y^{cv}, z_y^{cc})$.

| $k$ | $v_k$ | $V_k$ | $\mathcal{V}_k$ |
|---|---|---|---|
| 1 | $p$ | $P$ | $\text{MC}(p, p, p, p)$ |
| 2 | $z_x$ | $X(t)$ | $\text{MC}(x^L(t), x^U(t), z_x^{cv}, z_x^{cc})$ |
| 3 | $z_y$ | $Y(t)$ | $\text{MC}(y^L(t), y^U(t), z_y^{cv}, z_y^{cc})$ |
| 4 | $\tilde{z}_y$ | $\tilde{Y}(t)$ | $\text{MC}(\tilde{y}^L(t), \tilde{y}^U(t), \tilde{y}(t, \mathbf{p}), \tilde{y}(t, \mathbf{p}))$ |
| 5 | $\lambda$ | $[0, 1]$ | $\text{MC}(0, 1, 0, 1)$ |
| 6 | $-v_4$ | $-V_4$ | $-\mathcal{V}_4$ |
| 7 | $v_3 + v_6$ | $V_3 + V_6$ | $\mathcal{V}_3 + \mathcal{V}_6$ |
| 8 | $v_5 v_7$ | $V_5 V_7$ | $\mathcal{V}_5 \mathcal{V}_7$ |
| 9 | $v_4 + v_8$ | $V_4 + V_8$ | $\mathcal{V}_4 + \mathcal{V}_8$ |
| 10 | $(v_9)^{3/2}$ | $(V_9)^{3/2}$ | $(\mathcal{V}_9)^{3/2}$ |
| 11 | $1/v_{10}$ | $1/V_{10}$ | $1/\mathcal{V}_{10}$ |
| 12 | $\sin(v_1)$ | $\sin(V_1)$ | $\sin(\mathcal{V}_1)$ |
| 13 | $v_{11} v_{12}$ | $V_{11} V_{12}$ | $\mathcal{V}_{11} \mathcal{V}_{12}$ |
| 14 | $1 + v_{13}$ | $1 + V_{13}$ | $1 + \mathcal{V}_{13}$ |
| 15 | $1/v_{14}$ | $1/V_{14}$ | $1/\mathcal{V}_{14}$ |
| 16 | $\sqrt{v_4}$ | $\sqrt{V_4}$ | $\sqrt{\mathcal{V}_4}$ |
| 17 | $1/v_{16}$ | $1/V_{16}$ | $1/\mathcal{V}_{16}$ |
| 18 | $-2v_{12}$ | $-2V_{12}$ | $-2\mathcal{V}_{12}$ |
| 19 | $v_{17} v_{18}$ | $V_{17} V_{18}$ | $\mathcal{V}_{17} \mathcal{V}_{18}$ |
| 20 | $v_4 + v_{19}$ | $V_4 + V_{19}$ | $\mathcal{V}_4 + \mathcal{V}_{19}$ |
| 21 | $-7v_2$ | $-7V_2$ | $-7v_2$ |
| 22 | $v_{20} + v_{21}$ | $V_{20} + V_{21}$ | $\mathcal{V}_{20} + \mathcal{V}_{21}$ |
| 23 | $v_{15} v_{22}$ | $V_{15} V_{22}$ | $\mathcal{V}_{15} \mathcal{V}_{22}$ |
| 24 | $-v_{23}$ | $-V_{23}$ | $-\mathcal{V}_{23}$ |
| 25 | $v_4 + v_{24}$ | $V_4 + V_{24}$ | $\mathcal{V}_4 + \mathcal{V}_{24}$ |

where $\mathbf{z}_y^{cv,0} = \mathbf{y}^L(t)$, $\mathbf{z}_y^{cc,0} = \mathbf{y}^U(t)$, and

$$\begin{aligned}
\boldsymbol{\gamma}_i^{cv,k} &= (z_{y,1}^{cv,k+1}, \ldots, z_{y,i-1}^{cv,k+1}, z_{y,i}^{cv,k}, \ldots, z_{y,n_y}^{cv,k}) \\
\boldsymbol{\gamma}_i^{cc,k} &= (z_{y,1}^{cc,k+1}, \ldots, z_{y,i-1}^{cc,k+1}, z_{y,i}^{cc,k}, \ldots, z_{y,n_y}^{cc,k}) \\
z_{y,i}^{cv,k+1} &= u_{\psi,i}(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}, \boldsymbol{\gamma}_i^{cv,k}, \boldsymbol{\gamma}_i^{cc,k}) \\
z_{y,i}^{cc,k+1} &= o_{\psi,i}(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}, \boldsymbol{\gamma}_i^{cv,k}, \boldsymbol{\gamma}_i^{cc,k})
\end{aligned}$$

for all $i = 1, \ldots, n_y$ and all $0 \leq k < K$.

**Theorem 8.3.14.** *Let $\mathbf{x}^{cv}, \mathbf{x}^{cc} : I \times P \to \mathbb{R}^{n_x}$ be state relaxations for $\mathbf{x}$ on $I \times P$. Then state relaxations for $\mathbf{y}$ on $I \times P$, $\mathbf{y}^{cv}, \mathbf{y}^{cc} : I \times P \to \mathbb{R}^{n_y}$ are given by the definitions*

$$\begin{aligned}
\mathbf{y}^{cv}(t, \mathbf{p}) &= \bar{\mathbf{u}}_\psi^K(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p})), \\
\mathbf{y}^{cc}(t, \mathbf{p}) &= \bar{\mathbf{o}}_\psi^K(t, \mathbf{p}, \mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p})).
\end{aligned}$$

Again, this result is proven through a series of more fundamental lemmas.

**Lemma 8.3.15.** *For any $(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}) \in I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ satisfying $X(t) \cap [\mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}] \neq \emptyset$,*

$$\mathbf{y}^L(t) \leq \bar{\mathbf{u}}_\psi^K(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}) \leq \bar{\mathbf{o}}_\psi^K(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}) \leq \mathbf{y}^U(t). \tag{8.24}$$

*Moreover, choosing any $\mathbf{z}_x \in X(t) \cap [\mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}]$,*

$$\bar{\mathbf{u}}_\psi^K(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}) \leq \mathbf{z}_y \leq \bar{\mathbf{o}}_\psi^K(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}), \tag{8.25}$$

*where $\mathbf{z}_y \in \mathbb{R}^{n_y}$ is the unique point satisfying $\mathbf{z}_y \in Y(t)$ and $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$.*

*Proof.* Choose any $(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}) \in I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ and suppose that $X(t) \cap [\mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}] \neq \emptyset$. Define the quantities $\mathbf{z}_y^{cv,k}$, $\mathbf{z}_y^{cc,k}$, $\boldsymbol{\gamma}_i^{cv,k}$ and $\boldsymbol{\gamma}_i^{cc,k}$ as in Definition 8.3.13. Since $\mathbf{z}_y^{cv,0} = \mathbf{y}^L(t)$ and $\mathbf{z}_y^{cc,0} = \mathbf{y}^U(t)$ by definition, the inequalities $\mathbf{y}^L(t) \leq$

$\bar{\mathbf{u}}^K_\psi(t, \mathbf{p}, \mathbf{z}^{cv}_x, \mathbf{z}^{cc}_x)$ and $\bar{\mathbf{o}}^K_\psi(t, \mathbf{p}, \mathbf{z}^{cv}_x, \mathbf{z}^{cc}_x) \leq \mathbf{y}^U(t)$ follow from the use of the min and max functions in the definition of $\mathbf{u}_\psi$ and $\mathbf{o}_\psi$ using a trivial inductive argument. Since, $\exists \mathbf{z}_x \in X(t) \cap [\mathbf{z}^{cv}_x, \mathbf{z}^{cc}_x]$, it suffices to show (8.25), since this then implies (8.24).

Choose any $\mathbf{z}_x \in X(t) \cap [\mathbf{z}^{cv}_x, \mathbf{z}^{cc}_x] \neq \emptyset$ and define $\mathbf{z}_y$ accordingly. By definition, $\mathbf{z}_y \in [\mathbf{z}^{cv,0}_y, \mathbf{z}^{cc,0}_y]$. Suppose that this is true for some arbitrary $k \geq 0$. Since $\boldsymbol{\gamma}^{cv,k}_1 = \mathbf{z}^{cv,k}_y$ and $\boldsymbol{\gamma}^{cc,k}_1 = \mathbf{z}^{cc,k}_y$, $\mathbf{z}_y \in [\boldsymbol{\gamma}^{cv,k}_1, \boldsymbol{\gamma}^{cc,k}_1]$. Then, Lemma 8.3.10 implies that $z_{y,1} \in [z^{cv,k+1}_{y,1}, z^{cc,k+1}_{y,1}]$. Suppose that, for some $1 < \ell \leq n_y$, $z_{y,i} \in [z^{cv,k+1}_{y,i}, z^{cc,k+1}_{y,i}]$, for all $i < \ell$. By definition, it follows that $\mathbf{z}_y \in [\boldsymbol{\gamma}^{cv,k}_\ell, \boldsymbol{\gamma}^{cc,k}_\ell]$. By Lemma 8.3.10, this implies that $z_{y,\ell} \in [z^{cv,k+1}_{y,\ell}, z^{cc,k+1}_{y,\ell}]$. Now, applying finite induction over $\ell$ shows that $\mathbf{z}_y \in [\mathbf{z}^{cv,k+1}_y, \mathbf{z}^{cc,k+1}_y]$. Induction over $k$ shows that this conclusion holds for all $k \in \mathbb{N}$. $\square$

**Lemma 8.3.16.** *Let $\mathbf{z}^{cv}_x, \mathbf{z}^{cc}_x : I \times P \to \mathbb{R}^{n_x}$ and choose any $(\lambda, \mathbf{p}_1, \mathbf{p}_2) \in (0,1) \times P \times P$. For every $t \in I$ such that*

1. *$\mathbf{z}^{cv}_x(t, \cdot)$ is consistent with convexity at $(\lambda, \mathbf{p}_1, \mathbf{p}_2)$,*
2. *$\mathbf{z}^{cc}_x(t, \cdot)$ is consistent with concavity at $(\lambda, \mathbf{p}_1, \mathbf{p}_2)$,*
3. *$X(t) \cap [\mathbf{z}^{cv}_x(t, \mathbf{q}), \mathbf{z}^{cc}_x(t, \mathbf{q})] \neq \emptyset$, $\forall \mathbf{q} \in \{\mathbf{p}_1, \mathbf{p}_2, \lambda\mathbf{p}_1 + (1-\lambda)\mathbf{p}_2\}$,*

*the functions*

$$P \ni \mathbf{p} \longmapsto \bar{\mathbf{u}}^K_\psi(t, \mathbf{p}, \mathbf{z}^{cv}_x(t, \mathbf{p}), \mathbf{z}^{cc}_x(t, \mathbf{p})),$$
$$P \ni \mathbf{p} \longmapsto \bar{\mathbf{o}}^K_\psi(t, \mathbf{p}, \mathbf{z}^{cv}_x(t, \mathbf{p}), \mathbf{z}^{cc}_x(t, \mathbf{p})),$$

*are, respectively, consistent with convexity and consistent with concavity at $(\lambda, \mathbf{p}_1, \mathbf{p}_2)$.*

*Proof.* Considering the evaluation of $\bar{\mathbf{u}}^K_\psi$ and $\bar{\mathbf{o}}^K_\psi$ at $(t, \mathbf{p}, \mathbf{z}^{cv}_x(t, \mathbf{p}), \mathbf{z}^{cc}_x(t, \mathbf{p}))$, define the quantities $\mathbf{z}^{cv,k}_y(t, \mathbf{p})$, $\mathbf{z}^{cc,k}_y(t, \mathbf{p})$, $\boldsymbol{\gamma}^{cv,k}_i(t, \mathbf{p})$ and $\boldsymbol{\gamma}^{cc,k}_i(t, \mathbf{p})$ according to Definition 8.3.13 for all $(t, \mathbf{p}) \in I \times P$, all $k \in \{1, \ldots, K\}$ and all $i \in \{1, \ldots, n_y\}$.

Choose any $(\lambda, \mathbf{p}_1, \mathbf{p}_2) \in (0,1) \times P \times P$, define $\mathbf{p}_3 \equiv \lambda\mathbf{p}_1 + (1-\lambda)\mathbf{p}_2$, and suppose that $t \in I$ is such that 1-3 hold. By Hypothesis 3, it is possible to choose a function $\mathbf{z}_x : I \times P \to \mathbb{R}^{n_x}$ such that $\mathbf{z}_x(s, \mathbf{p}) \in X(s)$, $\forall (s, \mathbf{p}) \in I \times P$ and $\mathbf{z}_x(t, \mathbf{p}_i) \in$

$[\mathbf{z}_x^{cv}(t,\mathbf{p}_i),\mathbf{z}_x^{cc}(t,\mathbf{p}_i)]$, $\forall i\in\{1,2,3\}$. Let $\mathbf{z}_y : I\times P\to\mathbb{R}^{n_y}$ be the unique function satisfying $\mathbf{z}_y(s,\mathbf{p})\in Y(s)$ and $\mathbf{g}(s,\mathbf{p},\mathbf{z}_x(s,\mathbf{p}),\mathbf{z}_y(s,\mathbf{p}))=\mathbf{0}$, $\forall(s,\mathbf{p})\in I\times P$.

For arbitrary functions $\mathbf{r}^{cv},\mathbf{r}^{cc}: I\times P\to\mathbb{R}^{n_y}$, consider the hypotheses:

1. $\mathbf{r}^{cv}(t,\cdot)$ is consistent with convexity at $(\lambda,\mathbf{p}_1,\mathbf{p}_2)$,

2. $\mathbf{r}^{cv}(t,\cdot)$ is consistent with concavity at $(\lambda,\mathbf{p}_1,\mathbf{p}_2)$,

3. $\mathbf{r}^{cv}(t,\mathbf{p}_i)\leq\mathbf{z}_y(t,\mathbf{p}_i)\leq\mathbf{r}^{cc}(t,\mathbf{p}_i)$, $\forall i\in\{1,2,3\}$.

By definition, the Hypotheses 1-3 above hold with $(\mathbf{r}^{cv},\mathbf{r}^{cc})=(\mathbf{z}_y^{cv,0},\mathbf{z}_y^{cc,0})$. As an inductive hypothesis, suppose that this is true for some $k\geq 0$.

Since $\boldsymbol{\gamma}_1^{cv,k}=\mathbf{z}_y^{cv,k}$ and $\boldsymbol{\gamma}_1^{cc,k}=\mathbf{z}_y^{cc,k}$, 1-3 hold with $(\mathbf{r}^{cv},\mathbf{r}^{cc})=(\boldsymbol{\gamma}_1^{cv,k},\boldsymbol{\gamma}_1^{cc,k})$. Suppose that, for some $1\leq\ell<n_y$, Hypotheses 1-3 hold with $(\mathbf{r}^{cv},\mathbf{r}^{cc})=(\boldsymbol{\gamma}_\ell^{cv,k},\boldsymbol{\gamma}_\ell^{cc,k})$. Then, Lemma 8.3.11 may be applied with $(\mathbf{z}_y^{cv},\mathbf{z}_y^{cc}):=(\boldsymbol{\gamma}_\ell^{cv,k},\boldsymbol{\gamma}_\ell^{cc,k})$. This implies that $z_{y,\ell}^{cv,k+1}(t,\cdot)$ and $z_{y,\ell}^{cc,k+1}(t,\cdot)$ are, respectively, consistent with convexity and consistent with concavity at $(\lambda,\mathbf{p}_1,\mathbf{p}_2)$. For each $i\in\{1,2,3\}$, applying Lemma 8.3.10 with

$$(\mathbf{z}_x,\mathbf{z}_x^{cv},\mathbf{z}_x^{cv}):=(\mathbf{z}_x(t,\mathbf{p}_i),\mathbf{z}_x^{cv}(t,\mathbf{p}_i),\mathbf{z}_x^{cc}(t,\mathbf{p}_i)),\tag{8.26}$$

$$(\mathbf{z}_y,\mathbf{z}_y^{cv},\mathbf{z}_y^{cv}):=(\mathbf{z}_y(t,\mathbf{p}_i),\boldsymbol{\gamma}_\ell^{cv,k}(t,\mathbf{p}_i),\boldsymbol{\gamma}_\ell^{cc,k}(t,\mathbf{p}_i)),\tag{8.27}$$

proves that $z_{y,\ell}^{cv,k+1}(t,\mathbf{p}_i)\leq z_{y,\ell}(t,\mathbf{p}_i)\leq z_{y,\ell}^{cc,k+1}(t,\mathbf{p}_i)$. It follows that Hypotheses 1-3 hold with $(\mathbf{r}^{cv},\mathbf{r}^{cc})=(\boldsymbol{\gamma}_{\ell+1}^{cv,k},\boldsymbol{\gamma}_{\ell+1}^{cc,k})$. Finite induction over $\ell$ shows that 1-3 hold with $(\mathbf{r}^{cv},\mathbf{r}^{cc})=(\boldsymbol{\gamma}_{n_y}^{cv,k},\boldsymbol{\gamma}_{n_y}^{cc,k})$. Then, one more application of the inductive step above shows that $z_{y,n_y}^{cv,k+1}(t,\cdot)$ and $z_{y,n_y}^{cc,k+1}(t,\cdot)$ are, respectively, consistent with convexity and consistent with concavity at $(\lambda,\mathbf{p}_1,\mathbf{p}_2)$, and that $z_{y,n_y}^{cv,k+1}(t,\mathbf{p}_i)\leq z_{y,n_y}(t,\mathbf{p}_i)\leq z_{y,n_y}^{cc,k+1}(t,\mathbf{p}_i)$, $\forall i\in\{1,2,3\}$. Combining this with the fact that 1-3 hold with $(\mathbf{r}^{cv},\mathbf{r}^{cc})=(\boldsymbol{\gamma}_{n_y}^{cv,k},\boldsymbol{\gamma}_{n_y}^{cc,k})$, it follows that 1-3 hold with $(\mathbf{r}^{cv},\mathbf{r}^{cc})=(\mathbf{z}^{cv,k+1},\mathbf{z}^{cc,k+1})$. Induction over $k$ now shows that this conclusion holds for all $0\leq k\leq K$. In particular, Hypotheses 1 and 2 hold with $(\mathbf{r}^{cv},\mathbf{r}^{cc})=(\mathbf{z}^{cv,K},\mathbf{z}^{cc,K})$, which is the desired result. □

Theorem 8.3.14 now follows directly from Lemmas 8.3.15 and 8.3.16.

## 8.4 Relaxation Amplifying Dynamics

In the previous section, the functions $\bar{\mathbf{u}}_{\psi}^{K}$ and $\bar{\mathbf{o}}_{\psi}^{K}$ were defined. To specify fully the auxiliary system (8.3), it remains to define $\mathbf{x}_0^{cv}$, $\mathbf{x}_0^{cc}$, $\mathbf{u}_f$ and $\mathbf{o}_f$. These functions are defined below, and it is then shown that (8.3) furnishes state relaxations as its solutions by appealing to the theory of relaxation amplifying dynamics (§7.3).

**Definition 8.4.1.** Let $\mathbf{u}_f, \mathbf{o}_f : I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} \to \mathbb{R}^{n_y}$ and $\mathbf{x}_0^{cv}, \mathbf{x}_0^{cc} : P \to \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ be defined by

$$\begin{aligned}
\mathbf{u}_f(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}, \mathbf{z}_y^{cv}, \mathbf{z}_y^{cc}) &= \{\mathbf{f}\}^{cv}(\mathcal{T}, \mathcal{P}, \mathcal{X}, \mathcal{Y}), \quad \mathbf{x}_0^{cv}(\mathbf{p}) = \{\mathbf{x}_0\}^{cv}(\mathcal{P}),\\
\mathbf{o}_f(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}, \mathbf{z}_y^{cv}, \mathbf{z}_y^{cc}) &= \{\mathbf{f}\}^{cc}(\mathcal{T}, \mathcal{P}, \mathcal{X}, \mathcal{Y}), \quad \mathbf{x}_0^{cc}(\mathbf{p}) = \{\mathbf{x}_0\}^{cc}(\mathcal{P}),
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{X} &= \mathrm{MC}(\mathbf{x}^L(t), \mathbf{x}^U(t), \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}), \quad \mathcal{T} = \mathrm{MC}(t, t, t, t).\\
\mathcal{Y} &= \mathrm{MC}(\mathbf{y}^L(t), \mathbf{y}^U(t), \mathbf{z}_y^{cv}, \mathbf{z}_y^{cc}), \quad \mathcal{P} = \mathrm{MC}(\mathbf{p}, \mathbf{p}, \mathbf{p}, \mathbf{p}).
\end{aligned}$$

Because the algebraic equations in the auxiliary system (8.3) are explicit, it can be viewed as a system of explicit ODEs with right-hand side functions $\mathbf{u}, \mathbf{o} : I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ defined by

$$\begin{aligned}
\mathbf{u}(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}) &\equiv \mathbf{u}_f(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}, \bar{\mathbf{u}}_{\psi}^{K}(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}), \bar{\mathbf{o}}_{\psi}^{K}(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc})), \qquad (8.28)\\
\mathbf{o}(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}) &\equiv \mathbf{o}_f(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}, \bar{\mathbf{u}}_{\psi}^{K}(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}), \bar{\mathbf{o}}_{\psi}^{K}(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc})).
\end{aligned}$$

In the following results, it is established that $(\mathbf{u}, \mathbf{o})$ defined in this way describe relaxation amplifying dynamics for $\mathbf{x}$ on $I \times P$.

**Lemma 8.4.2.** *For arbitrary functions $\mathbf{z}_x^{cv}, \mathbf{z}_x^{cc} : I \times P \to \mathbb{R}^{n_x}$ and every $\mathbf{p} \in P$, the following conditions hold:*

1. *For a.e. $t \in I$ such that $\mathbf{z}_x^{cv}(t, \mathbf{p}) \leq \mathbf{z}_x^{cc}(t, \mathbf{p})$ and $X(t) \cap [\mathbf{z}_x^{cv}(t, \mathbf{p}), \mathbf{z}_x^{cc}(t, \mathbf{p})] \neq \emptyset$,*

$\mathbf{u}$ *and* $\mathbf{o}$ *satisfy*

$$\mathbf{u}(t, \mathbf{p}, \mathbf{z}_x^{cv}(t, \mathbf{p}), \mathbf{z}_x^{cc}(t, \mathbf{p})) \leq \mathbf{o}(t, \mathbf{p}, \mathbf{z}_x^{cv}(t, \mathbf{p}), \mathbf{z}_x^{cc}(t, \mathbf{p})). \tag{8.29}$$

2. *For a.e.* $t \in I$ *such that* $\mathbf{z}_x^{cv}(t, \mathbf{p}) \leq \mathbf{x}(t, \mathbf{p}) \leq \mathbf{z}_x^{cc}(t, \mathbf{p})$*,* $\mathbf{u}$ *and* $\mathbf{o}$ *satisfy*

$$\mathbf{u}(t, \mathbf{p}, \mathbf{z}_x^{cv}(t, \mathbf{p}), \mathbf{z}_x^{cc}(t, \mathbf{p})) \leq \dot{\mathbf{x}}(t, \mathbf{p}) \leq \mathbf{o}(t, \mathbf{p}, \mathbf{z}_x^{cv}(t, \mathbf{p}), \mathbf{z}_x^{cc}(t, \mathbf{p})). \tag{8.30}$$

*Proof.* Choose arbitrary functions $\mathbf{z}_x^{cv}, \mathbf{z}_x^{cc} : I \times P \to \mathbb{R}^{n_x}$ and let $\mathbf{p} \in P$. For any $t \in I$, (8.29) follows from Definition 8.4.1 and the fact that $\{\mathbf{f}\}$ takes values in $\mathbb{MR}^{n_x}$. Suppose that $t \in I$ is such that $\mathbf{z}_x^{cv}(t, \mathbf{p}) \leq \mathbf{x}(t, \mathbf{p}) \leq \mathbf{z}_x^{cc}(t, \mathbf{p})$. Define

$$\begin{aligned} \mathbf{z}_y^{cv}(t, \mathbf{p}) &\equiv \bar{\mathbf{u}}_\psi^K(t, \mathbf{p}, \mathbf{z}_x^{cv}(t, \mathbf{p}), \mathbf{z}_x^{cc}(t, \mathbf{p})), \\ \mathbf{z}_y^{cc}(t, \mathbf{p}) &\equiv \bar{\mathbf{o}}_\psi^K(t, \mathbf{p}, \mathbf{z}_x^{cv}(t, \mathbf{p}), \mathbf{z}_x^{cc}(t, \mathbf{p})). \end{aligned} \tag{8.31}$$

Applying Lemma 8.3.15 with $(\mathbf{z}_x, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}) := (\mathbf{x}(t, \mathbf{p}), \mathbf{z}_x^{cv}(t, \mathbf{p}), \mathbf{z}_x^{cc}(t, \mathbf{p}))$ proves that

$$\mathbf{z}_y^{cv}(t, \mathbf{p}) \leq \mathbf{y}(t, \mathbf{p}) \leq \mathbf{z}_y^{cc}(t, \mathbf{p}). \tag{8.32}$$

By Assumption 8.2.2, the interval $[t, t] \times P \times X(t) \times Y(t)$ is represented in $\mathcal{D}$. Thus, Lemma 2.7.3 may be applied with

$$\begin{aligned} \mathbf{x} &:= (t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p})), \\ \mathbf{x}^{cv} &:= (t, \mathbf{p}, \mathbf{z}_x^{cv}(t, \mathbf{p}), \mathbf{z}_y^{cv}(t, \mathbf{p})), \\ \mathbf{x}^{cc} &:= (t, \mathbf{p}, \mathbf{z}_x^{cc}(t, \mathbf{p}), \mathbf{z}_y^{cc}(t, \mathbf{p})), \end{aligned} \tag{8.33}$$

to conclude that

$$\begin{aligned}\mathbf{u}_f(t,\mathbf{p},\mathbf{z}_x^{cv}(t,\mathbf{p}),\mathbf{z}_x^{cc}(t,\mathbf{p}),\mathbf{z}_y^{cv}(t,\mathbf{p}),\mathbf{z}_y^{cc}(t,\mathbf{p}))& \\ \leq \mathbf{f}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p}))& \\ \leq \mathbf{o}_f(t,\mathbf{p},\mathbf{z}_x^{cv}(t,\mathbf{p}),\mathbf{z}_x^{cc}(t,\mathbf{p}),\mathbf{z}_y^{cv}(t,\mathbf{p}),\mathbf{z}_y^{cc}(t,\mathbf{p})),&\end{aligned} \tag{8.34}$$

which is the desired inequality. □

**Corollary 8.4.3.** *The functions* $(\mathbf{u},\mathbf{o})$ *describe bound amplifying dynamics for* $\mathbf{x}$ *on* $I \times P$.

*Proof.* This follows immediately from Conclusion 2 of Lemma 8.4.2. □

**Lemma 8.4.4.** *Let* $\mathbf{z}_x^{cv},\mathbf{z}_x^{cc} : I \times P \to \mathbb{R}^{n_x}$ *and choose any* $(\lambda,\mathbf{p}_1,\mathbf{p}_2) \in (0,1)\times P \times P$. *For every* $t \in I$ *such that*

1. $\mathbf{z}_x^{cv}(t,\cdot)$ *is consistent with convexity at* $(\lambda,\mathbf{p}_1,\mathbf{p}_2)$,

2. $\mathbf{z}_x^{cc}(t,\cdot)$ *is consistent with concavity at* $(\lambda,\mathbf{p}_1,\mathbf{p}_2)$,

3. $X(t)\cap[\mathbf{z}_x^{cv}(t,\mathbf{q}),\mathbf{z}_x^{cc}(t,\mathbf{q})] \neq \emptyset$, $\forall \mathbf{q} \in \{\mathbf{p}_1,\mathbf{p}_2,\lambda\mathbf{p}_1+(1-\lambda)\mathbf{p}_2\}$,

*the functions*

$$\begin{aligned}P \ni \mathbf{p} &\longmapsto \mathbf{u}(t,\mathbf{p},\mathbf{z}_x^{cv}(t,\mathbf{p}),\mathbf{z}_x^{cc}(t,\mathbf{p})),\\ P \ni \mathbf{p} &\longmapsto \mathbf{o}(t,\mathbf{p},\mathbf{z}_x^{cv}(t,\mathbf{p}),\mathbf{z}_x^{cc}(t,\mathbf{p})),\end{aligned}$$

*are, respectively, consistent with convexity and consistent with concavity at* $(\lambda,\mathbf{p}_1,\mathbf{p}_2)$.

*Proof.* Define $\mathbf{z}_y^{cv}(t,\mathbf{p})$ and $\mathbf{z}_y^{cc}(t,\mathbf{p})$ as in (8.31), for all $(t,\mathbf{p}) \in I\times P$. Let $(\lambda,\mathbf{p}_1,\mathbf{p}_2) \in (0,1)\times P\times P$, define $\mathbf{p}_3 \equiv \lambda\mathbf{p}_1+(1-\lambda)\mathbf{p}_2$, and suppose that $t \in I$ is such that Conditions 1-3 hold. Under these hypotheses, Lemma 8.3.16 implies that $\mathbf{z}_y^{cv}(t,\cdot)$ and $\mathbf{z}_y^{cc}(t,\cdot)$ are, respectively, consistent with convexity and consistent with concavity at $(\lambda,\mathbf{p}_1,\mathbf{p}_2)$. Moreover, for each $i \in \{1,2,3\}$, applying Lemma 8.3.15 with $(\mathbf{z}_x^{cv},\mathbf{z}_x^{cc}) :=$

$(\mathbf{z}_x^{cv}(t,\mathbf{p}_i),\mathbf{z}_x^{cc}(t,\mathbf{p}_i))$ proves that $\mathbf{z}_y^{cv}(t,\mathbf{p}_i) \leq \mathbf{z}_y^{cc}(t,\mathbf{p}_i)$ and $[\mathbf{z}_y^{cv}(t,\mathbf{p}_i),\mathbf{z}_y^{cc}(t,\mathbf{p}_i)] \subset Y(t)$, and hence $Y(t) \cap [\mathbf{z}_y^{cv}(t,\mathbf{p}_i),\mathbf{z}_y^{cc}(t,\mathbf{p}_i)] \neq \emptyset$.

By Assumption 8.2.2, the interval $[t,t] \times P \times X(t) \times Y(t)$ is represented in $\mathcal{D}$. Thus, Lemma 2.7.4 may be applied with

$$\mathbf{x}_i^{cv} \equiv (t,\mathbf{p}_i,\mathbf{z}_x^{cv}(t,\mathbf{p}_i),\mathbf{z}_y^{cv}(t,\mathbf{p}_i)), \tag{8.35}$$

$$\mathbf{x}_i^{cc} \equiv (t,\mathbf{p}_i,\mathbf{z}_x^{cc}(t,\mathbf{p}_i),\mathbf{z}_y^{cc}(t,\mathbf{p}_i)), \tag{8.36}$$

for all $i \in \{1,2,3\}$, to conclude that the functions

$$\begin{aligned} &P \ni \mathbf{p} \longmapsto \mathbf{u}_f(t,\mathbf{p},\mathbf{z}_x^{cv}(t,\mathbf{p}),\mathbf{z}_x^{cc}(t,\mathbf{p}),\mathbf{z}_y^{cv}(t,\mathbf{p}),\mathbf{z}_y^{cc}(t,\mathbf{p})), \qquad (8.37)\\ &P \ni \mathbf{p} \longmapsto \mathbf{o}_f(t,\mathbf{p},\mathbf{z}_x^{cv}(t,\mathbf{p}),\mathbf{z}_x^{cc}(t,\mathbf{p}),\mathbf{z}_y^{cv}(t,\mathbf{p}),\mathbf{z}_y^{cc}(t,\mathbf{p})), \end{aligned}$$

are, respectively, consistent with convexity and consistent with concavity at $(\lambda,\mathbf{p}_1,\mathbf{p}_2)$. By (8.28) and (8.31), this is the desired result. □

**Corollary 8.4.5.** *The functions* $(\mathbf{u},\mathbf{o})$ *describe convexity amplifying dynamics for* $\mathbf{x}$ *on* $I \times P$.

*Proof.* Choose arbitrary functions $\mathbf{z}_x^{cv},\mathbf{z}_x^{cc} : I \times P \to \mathbb{R}^{n_x}$, let $(\lambda,\mathbf{p}_1,\mathbf{p}_2) \in (0,1) \times P \times P$, and suppose that $t \in I$ is such that Conditions 1-3 of Definition 7.3.3 hold. This immediately implies that Conditions 1-3 of Lemma 8.4.4 are satisfied, which gives the desired conclusion. □

It has now been shown that $(\mathbf{u},\mathbf{o})$ describe relaxation amplifying dynamics for $\mathbf{x}$ on $I \times P$. In order to guarantee that (8.3) describes state relaxations for $\mathbf{x}$ on $I \times P$ as its solutions, the conditions of Assumption 7.2.3 must be verified as well. Since the initial conditions in (8.3) are defined to be the standard McCormick relaxations of $\mathbf{x}_0$ on $P$, they are Lipschitz on $P$ on account of Corollary 2.6.2. Then, Assumption 7.2.3 holds in light of the following lemma.

**Lemma 8.4.6.** *The functions* $\mathbf{u}$ *and* $\mathbf{o}$ *are continuous on* $I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$. *Moreover,*

*$\exists L \in \mathbb{R}_+$ such that*

$$\begin{aligned}\|\mathbf{u}(t,\mathbf{p},\mathbf{z}^{cv}_x,\mathbf{z}^{cc}_x) - \mathbf{u}(t,\mathbf{p},\hat{\mathbf{z}}^{cv}_x,\hat{\mathbf{z}}^{cc}_x)\|_\infty + \|\mathbf{o}(t,\mathbf{p},\mathbf{z}^{cv}_x,\mathbf{z}^{cc}_x) - \mathbf{o}(t,\mathbf{p},\hat{\mathbf{z}}^{cv}_x,\hat{\mathbf{z}}^{cc}_x)\|_\infty \\ \leq L(\|\mathbf{z}^{cv}_x - \hat{\mathbf{z}}^{cv}_x\|_\infty + \|\mathbf{z}^{cc}_x - \hat{\mathbf{z}}^{cc}_x\|_\infty),\end{aligned} \tag{8.38}$$

*for all $(t,\mathbf{p},\mathbf{z}^{cv}_x,\mathbf{z}^{cc}_x,\hat{\mathbf{z}}^{cv}_x,\hat{\mathbf{z}}^{cc}_x) \in I \times P \times \mathbb{R}^{4n_x}$.*

*Proof.* Using Corollary 2.7.8 and a construction exactly analogous to the proof of Lemma 7.6.7, it is straightforward to show that $\mathbf{u}_f$ and $\mathbf{o}_f$ are continuous on $I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_y}$, and $\exists L_f \in \mathbb{R}_+$ such that

$$\|\mathbf{u}_f(t,\mathbf{p},\mathbf{z}^{cv}_x,\mathbf{z}^{cc}_x,\mathbf{z}^{cv}_y,\mathbf{z}^{cc}_y) - \mathbf{u}_f(t,\mathbf{p},\hat{\mathbf{z}}^{cv}_x,\hat{\mathbf{z}}^{cc}_x,\hat{\mathbf{z}}^{cv}_y,\hat{\mathbf{z}}^{cc}_y)\|_\infty \tag{8.39}$$

$$+ \|\mathbf{o}_f(t,\mathbf{p},\mathbf{z}^{cv}_x,\mathbf{z}^{cc}_x,\mathbf{z}^{cv}_y,\mathbf{z}^{cc}_y) - \mathbf{o}_f(t,\mathbf{p},\hat{\mathbf{z}}^{cv}_x,\hat{\mathbf{z}}^{cc}_x,\hat{\mathbf{z}}^{cv}_y,\hat{\mathbf{z}}^{cc}_y)\|_\infty \tag{8.40}$$

$$\leq L_f(\|\mathbf{z}^{cv}_x - \hat{\mathbf{z}}^{cv}_x\|_\infty + \|\mathbf{z}^{cc}_x - \hat{\mathbf{z}}^{cc}_x\|_\infty + \|\mathbf{z}^{cv}_y - \hat{\mathbf{z}}^{cv}_y\|_\infty + \|\mathbf{z}^{cc}_y - \hat{\mathbf{z}}^{cc}_y\|_\infty), \tag{8.41}$$

for all $(t,\mathbf{p},\mathbf{z}^{cv}_x,\mathbf{z}^{cc}_x,\hat{\mathbf{z}}^{cv}_x,\hat{\mathbf{z}}^{cc}_x,\mathbf{z}^{cv}_y,\mathbf{z}^{cc}_y,\hat{\mathbf{z}}^{cv}_y,\hat{\mathbf{z}}^{cc}_y) \in I \times P \times \mathbb{R}^{4n_x} \times \mathbb{R}^{4n_y}$.

Sine the composition of continuous functions is continuous, it follows from Lemma 8.3.12 that $\bar{\mathbf{u}}^K_\psi$ and $\bar{\mathbf{o}}^K_\psi$ are continuous on $I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$. Moreover, since the composition of Lipschitz functions is Lipschitz, it further follows from Lemma 8.3.12 that $\exists L_\psi \in \mathbb{R}_+$ such that

$$\begin{aligned}\|\bar{\mathbf{u}}^K_\psi(t,\mathbf{p},\mathbf{z}^{cv}_x,\mathbf{z}^{cc}_x) - \bar{\mathbf{u}}^K_\psi(t,\mathbf{p},\hat{\mathbf{z}}^{cv}_x,\hat{\mathbf{z}}^{cc}_x)\|_\infty + \|\bar{\mathbf{o}}^K_\psi(t,\mathbf{p},\mathbf{z}^{cv}_x,\mathbf{z}^{cc}_x) - \bar{\mathbf{o}}^K_\psi(t,\mathbf{p},\hat{\mathbf{z}}^{cv}_x,\hat{\mathbf{z}}^{cc}_x)\|_\infty \\ \leq L_\psi(\|\mathbf{z}^{cv}_x - \hat{\mathbf{z}}^{cv}_x\|_\infty + \|\mathbf{z}^{cc}_x - \hat{\mathbf{z}}^{cc}_x\|_\infty),\end{aligned}$$

for all $(t,\mathbf{p},\mathbf{z}^{cv}_x,\mathbf{z}^{cc}_x,\hat{\mathbf{z}}^{cv}_x,\hat{\mathbf{z}}^{cc}_x) \in I \times P \times \mathbb{R}^{4n_x}$. Again using composition results, it follows that $\mathbf{u}$ and $\mathbf{o}$ are continuous, and that (8.38) holds with $L = L_f L_\psi$. □

**Corollary 8.4.7.** *The auxiliary system* (8.3) *has a unique solution* $(\mathbf{x}^{cv},\mathbf{x}^{cc},\mathbf{y}^{cv},\mathbf{y}^{cc})$ *on all of $I \times P$, and $\mathbf{x}^{cv}$, $\mathbf{x}^{cc}$, $\mathbf{y}^{cv}$ and $\mathbf{y}^{cc}$ are state relaxations for* $(\mathbf{x},\mathbf{y})$ *on $I \times P$.*

*Proof.* Existence of the solution and the fact that $\mathbf{x}^{cv}$ and $\mathbf{x}^{cc}$ are state relaxations of $\mathbf{x}$ on $I \times P$ follows from Theorem 7.3.4. The fact that $\mathbf{y}^{cv}$ and $\mathbf{y}^{cc}$ are state relaxations of $\mathbf{y}$ on $I \times P$ follows from Theorem 8.3.9. □

Table 8.2: Factorization and computation of $f$ at $(t, p, z_x, z_y)$ and $u_f$ and $o_f$ at $(t, p, z_x^{cv}, z_x^{cc}, z_y^{cv}, z_y^{cc})$.

| $k$ | $v_k$ | $V_k$ | $\mathcal{V}_k$ |
|---|---|---|---|
| 1 | $p$ | $P$ | $\text{MC}(p, p, p, p)$ |
| 2 | $z_x$ | $X(t)$ | $\text{MC}(x^L(t), x^U(t), z_x^{cv}, z_x^{cc})$ |
| 3 | $z_y$ | $Y(t)$ | $\text{MC}(y^L(t), y^U(t), z_y^{cv}, z_y^{cc})$ |
| 4 | $-(1/2)v_1$ | $-(1/2)V_1$ | $-(1/2)\mathcal{V}_1$ |
| 5 | $v_3 + v_4$ | $V_3 + V_4$ | $\mathcal{V}_3 + \mathcal{V}_4$ |
| 6 | $-(1/2)v_5$ | $-(1/2)V_5$ | $-(1/2)\mathcal{V}_5$ |
| 7 | $v_6 v_2$ | $V_6 V_2$ | $\mathcal{V}_6 \mathcal{V}_2$ |

According to the previous Corollary, state relaxations of $(\mathbf{x}, \mathbf{y})$ on $I \times P$ can be computed by constructing the auxiliary system of DAEs (8.3) and solving it using any standard numerical integration technique. For the DAEs (8.4), the construction of this system was initiated in Examples 8.3.1 and 8.3.2, and is completed in the following example.

**Example 8.4.1.** Consider the functions

$$f(t, p, z_x, z_y) = -\frac{1}{2}(z_y - \frac{1}{2}p)z_x, \quad x_0(p) = 1,$$

from the example DAEs (8.4). Here, we demonstrate the computation of $x_0^{cv}$, $x_0^{cc}$, $u_f$ and $o_f$ for this example, as per Definition 8.4.1. Since $x_0$ is constant, appropriate convex and concave relaxations are simply $x_0^{cv} = x_0^{cc} = x_0$.

Now consider $u_f$ and $o_f$. For any $(t, p, z_x^{cv}, z_x^{cc}, z_y^{cv}, z_y^{cc}) \in I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_y}$, appropriate values for the functions $u_f$ and $o_f$ at $(t, p, z_x^{cv}, z_x^{cc}, z_y^{cv}, z_y^{cc})$ are computed by evaluating the natural McCormick extension of $\psi$ with the initializations given in Definition 8.4.1. This is implemented by the factorization of $f$ shown in Table 8.2 with factors $v_k$, inclusion factors $V_k$, and relaxation factors, $\mathcal{V}_k$. The values of $u_f(t, p, z_x^{cv}, z_x^{cc}, z_y^{cv}, z_y^{cc})$ and $o_f(t, p, z_x^{cv}, z_x^{cc}, z_y^{cv}, z_y^{cc})$ are given by $\mathcal{V}_7^{cv}$ and $\mathcal{V}_7^{cc}$ in Table 8.2, respectively.

## 8.5 Relaxation Preserving Dynamics

In this section, a modified auxiliary system is defined which provides sharper state relaxations for $(\mathbf{x}, \mathbf{y})$ by appealing to the theory of relaxation preserving dynamics (§7.4). Given the properties already established for the functions $(\mathbf{u}, \mathbf{o})$ defined by (8.28), functions describing relaxation preserving dynamics for $\mathbf{x}$ on $I \times P$ can be derived through the use of the functions $\mathcal{R}_i^{cv}$ and $\mathcal{R}_i^{cc}$ (Definition 7.6.9), exactly as in §7.6.2.

For the remainder of this section, define $\mathbf{u}, \mathbf{o} : I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ by

$$\begin{aligned} u_i(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}) &= u_{f,i}(t, \mathbf{p}, \mathcal{R}_i^{cv}(\mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}), \bar{\mathbf{u}}_\psi^K(t, \mathbf{p}, \mathcal{R}_i^{cv}(\mathbf{z}_x^{cv}, \mathbf{z}_x^{cc})), \\ &\qquad \bar{\mathbf{o}}_\psi^K(t, \mathbf{p}, \mathcal{R}_i^{cv}(\mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}))), \\ o_i(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}) &= o_{f,i}(t, \mathbf{p}, \mathcal{R}_i^{cc}(\mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}), \bar{\mathbf{u}}_\psi^K(t, \mathbf{p}, \mathcal{R}_i^{cc}(\mathbf{z}_x^{cv}, \mathbf{z}_x^{cc})), \\ &\qquad \bar{\mathbf{o}}_\psi^K(t, \mathbf{p}, \mathcal{R}_i^{cc}(\mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}))), \end{aligned} \tag{8.42}$$

for all $(t, \mathbf{p}, \mathbf{z}_x^{cv}, \mathbf{z}_x^{cc}) \in I \times P \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ and each $i \in \{1, \dots, n_x\}$.

**Corollary 8.5.1.** *Define* $(\mathbf{u}, \mathbf{o})$ *as in* (8.42). *The functions* $(\mathbf{u}, \mathbf{o})$ *describe bound preserving dynamics for* $\mathbf{x}$ *on* $I \times P$.

*Proof.* This is an immediate consequence of Lemmas 8.4.2 and 7.6.10. □

**Corollary 8.5.2.** *Define* $(\mathbf{u}, \mathbf{o})$ *as in* (8.42). *The functions* $(\mathbf{u}, \mathbf{o})$ *describe convexity preserving dynamics for* $\mathbf{x}$ *on* $I \times P$.

*Proof.* This is an immediate consequence of Lemmas 8.4.4 and 7.6.11. □

Given the previous two Corollaries, it follows from Theorems 7.5.5 and 7.5.7 in §7.5 that state relaxations for $\mathbf{x}$ on $I \times P$ are given by the solutions of (7.13) with the definitions (8.42). Using the simplification of Lemma 7.5.6, this system is defined by

$$
\begin{aligned}
\dot{x}_i^{cv}(t,\mathbf{p}) &= \begin{cases} u_{f,i}(t,\mathbf{p},\mathcal{R}_i^{cv}(\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})),\mathbf{y}_{i,cv}^{cv}(t,\mathbf{p}),\mathbf{y}_{i,cv}^{cc}(t,\mathbf{p})) \\ \qquad \text{if} \quad x_i^{cv}(t,\mathbf{p}) \neq x_i^L(t) \\ \max(\dot{x}_i^L(t), u_{f,i}(t,\mathbf{p},\mathcal{R}_i^{cv}(\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})),\mathbf{y}_{i,cv}^{cv}(t,\mathbf{p}),\mathbf{y}_{i,cv}^{cc}(t,\mathbf{p}))) \\ \qquad \text{otherwise} \end{cases}, \\
\mathbf{y}_{i,cv}^{cv}(t,\mathbf{p}) &= \bar{\mathbf{u}}_\psi^K(t,\mathbf{p},\mathcal{R}_i^{cv}(\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p}))), \\
\mathbf{y}_{i,cv}^{cc}(t,\mathbf{p}) &= \bar{\mathbf{o}}_\psi^K(t,\mathbf{p},\mathcal{R}_i^{cv}(\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p}))), \\
\dot{x}_i^{cc}(t,\mathbf{p}) &= \begin{cases} o_{f,i}(t,\mathbf{p},\mathcal{R}_i^{cc}(\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})),\mathbf{y}_{i,cc}^{cv}(t,\mathbf{p}),\mathbf{y}_{i,cc}^{cc}(t,\mathbf{p})) \\ \qquad \text{if} \quad x_i^{cc}(t,\mathbf{p}) \neq x_i^U(t) \\ \min(\dot{x}_i^U(t), o_{f,i}(t,\mathbf{p},\mathcal{R}_i^{cc}(\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})),\mathbf{y}_{i,cc}^{cv}(t,\mathbf{p}),\mathbf{y}_{i,cc}^{cc}(t,\mathbf{p}))) \\ \qquad \text{otherwise} \end{cases}, \\
\mathbf{y}_{i,cc}^{cv}(t,\mathbf{p}) &= \bar{\mathbf{u}}_\psi^K(t,\mathbf{p},\mathcal{R}_i^{cc}(\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p}))), \\
\mathbf{y}_{i,cc}^{cc}(t,\mathbf{p}) &= \bar{\mathbf{o}}_\psi^K(t,\mathbf{p},\mathcal{R}_i^{cc}(\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p}))), \\
x_i^{cv}(t_0,\mathbf{p}) &= \max(x_i^L(t_0), x_{0,i}^{cv}(\mathbf{p})), \quad x_i^{cc}(t_0,\mathbf{p}) = \min(x_i^U(t_0), x_{0,i}^{cc}(\mathbf{p})),
\end{aligned} \tag{8.43}
$$

for each $i = 1, \ldots, n_x$. Note that the explicit equations for the algebraic relaxations are composed with $\mathcal{R}_i^{cv}$ and $\mathcal{R}_i^{cc}$, so that there are $2n_x$ complete sets of $n_y$ algebraic variables involved in this system. In general, it is not necessary for either of $(\mathbf{y}_{i,cv}^{cv}, \mathbf{y}_{i,cv}^{cc})$ or $(\mathbf{y}_{i,cc}^{cv}, \mathbf{y}_{i,cc}^{cc})$ to be state relaxations for $\mathbf{y}$ on $I \times P$, for any $i$. However, state relaxations for $\mathbf{y}$ on $I \times P$ can be computed after the solution of (8.43) through the definitions

$$
\begin{aligned}
\mathbf{y}^{cv}(t,\mathbf{p}) &= \bar{\mathbf{u}}_\psi^K(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})), \qquad (8.44) \\
\mathbf{y}^{cc}(t,\mathbf{p}) &= \bar{\mathbf{o}}_\psi^K(t,\mathbf{p},\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})),
\end{aligned}
$$

as per Theorem 8.3.14.

## 8.6 Numerical Examples

All numerical experiments in this section were performed on a Dell Precision T3400 workstation with a 2.83 GHz Intel Core2 Quad CPU. One core and 512 MB of memory were dedicated to each job.

**Example 8.6.1.** Numerical results for the DAEs (8.4) are shown in Figures 8-1 and 8-2. The figures show the parametric final time solutions $x(t, \cdot)$ and $y(t, \cdot)$, respectively, which are both nonconvex (solid curves). The figures also show the state bounds at $t_f$, computed using the single-phase method of Chapter 6 (circles). Finally, Figures 8-1 and 8-2 show state relaxations for $(x, y)$, computed by deriving and solving the systems (8.3) (squares) and (8.43) (dashed). In constructing $\bar{u}^K_\psi$ and $\bar{o}^K_\psi$ by definition 8.3.13, the value $K = 10$ was used. For numerical solution, (8.3) was regarded as an explicit system of ODEs and integrated using the BDF method in `CVODES` [44] with relative and absolute tolerances of $1 \times 10^{-6}$. The system (8.43) was also solved using `CVODES` through the event detection scheme described in §7.6.3. Clearly, the state relaxations derived through the theory of relaxation preserving dynamics are much tighter than those derived through relaxation amplifying dynamics.

## 8.7 Conclusion

Two numerical method has been presented for computing convex and concave relaxations of the parametric solutions of a system of nonlinear, semi-explicit, index-one DAEs. Relaxations of the algebraic variables are computed by iterative refinement of known interval bounds which are available from the methods of Chapter 6. This procedure is then used in the definition of an auxiliary system of DAEs, the solutions of which provide the desired relaxations of both the differential and algebraic variables. This relaxation procedure was demonstrated for a simple example problem, and the computed relaxations were shown to provide tight approximations to the original DAE solutions. Analogous to the results for ODEs in Chapter 7, it was observed that state relaxations computed based on the concept of relaxation preserv-

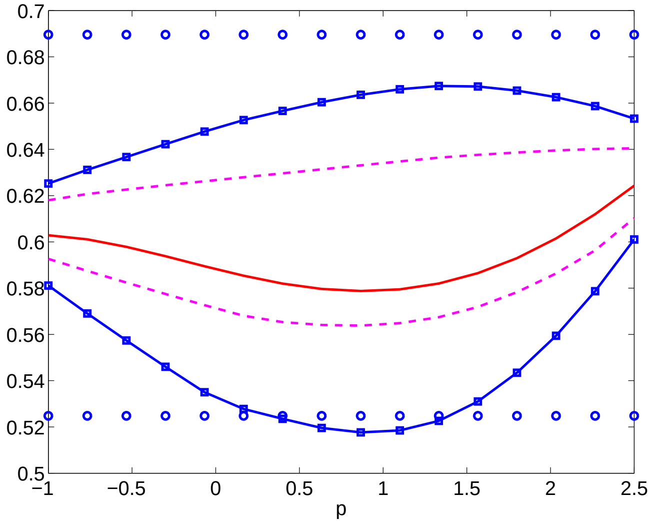

Figure 8-1: Parametric final time solution of (8.4), $x(t_f, \cdot)$ (solid line), along with interval bounds (circles) and convex and concave relaxations, $x^{cv}(t_f, \cdot)$ and $x^{cc}(t_f, \cdot)$, computed by solving (8.3) (squares) and (8.43) (dashed lines).

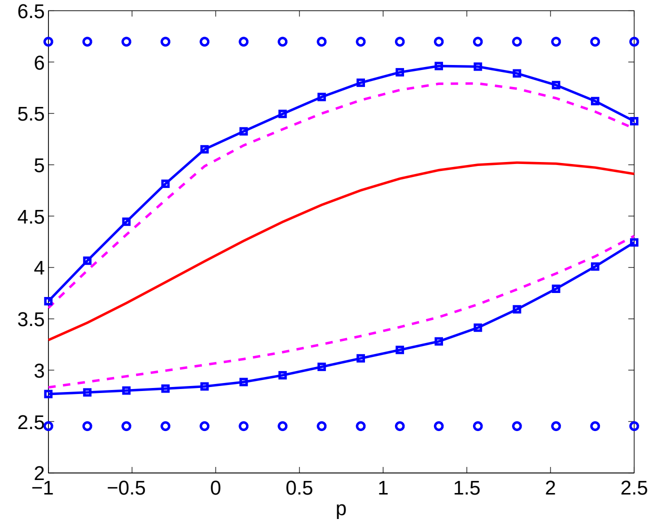

Figure 8-2: Parametric final time solution of (8.4), $y(t_f, \cdot)$ (solid line), along with interval bounds (circles) and convex and concave relaxations, $y^{cv}(t_f, \cdot)$ and $y^{cc}(t_f, \cdot)$, computed by solving (8.3) (squares) and (8.43) (dashed lines).

ing dynamics are significantly sharper than those computed based on the concept of relaxation amplifying dynamics. Finally, note that no discretization of the original DAEs is required in order to construct these relaxations, aside from that inherent in the numerical solution of the auxiliary DAEs.

The primary motivation for constructing convex and concave relaxations of DAE solutions is for their use in deterministic global optimization algorithms for problems with DAEs embedded. Based on existing methods for problems with ODEs embedded, the ability to construct relaxations of DAE solutions leads directly to such an algorithm.

# Chapter 9

# Convex Polyhedral Enclosures of Reachable Sets

## 9.1 Introduction

In this chapter, an efficient method is presented for computing convex polyhedral enclosures of the reachable sets of parametric ODEs and semi-explicit index-one DAEs. Informally, the reachable set of a dynamic system at some fixed time is the set of states which can be attained at the given time by solutions of the system with initial conditions, parameters, controls and disturbances in some specified sets. The computation of reachable sets, or conservative approximations of them, is a classical problem with a long history [68]. Reachability analysis is also closely related to the construction of discrete abstractions and plays a central role in problems in controller design and synthesis [132, 110], fault detection [106] and verification of continuous and hybrid systems [85, 99, 40, 20, 184].

A variety of methods exist for computing conservative approximations of the reachable sets of dynamic systems. Many of these methods are only possible or tractable for linear systems [20, 97], while others require the costly solution of Hamilton-Jacobi-Bellman equations [98, 132]. If an interval enclosure is sufficient, a number of more efficient methods are available, including those developed in Chapters 3 - 6. However, interval methods can produce quite conservative enclosures, especially if no additional physical information can be provided to the method (see Chapter 4). Finally, there are several methods which compute approximations of the reachable set by constructing supporting hyperplanes. These hyperplanes are obtained either from the solutions of adjoint equations [68, 143], or by specifying the normal to the desired hyperplane and computing an appropriate intercept by solving a dynamic optimization problem [39, 41]. Unfortunately, nonconvexity of the reachable set makes implementation impractical in both cases. In the latter case, nonconvexity of the reachable set leads to nonconvex dynamic optimization problems which must be solved to guaranteed global optimality. In the former case, the resulting hyperplanes are not guaranteed to support the reachable set when it is nonconvex. In response to this issue, some authors have developed conditions for the convexity of reachable sets of nonlinear dynamic systems [143, 137]. Unfortunately, these results involve bounds on the size of the sets of permissible initial states and controls and/or bounds on the time horizon which are extremely restrictive in the general case.

Here, it is shown that a convex enclosure of the reachable set for some fixed time can be computed efficiently, regardless of whether or not the reachable set is itself convex. The method for doing this relies on the computation of state relaxations, described in Chapters 7 and 8. Using these relaxations, a convex enclosure of the reachable set can be expressed as an infinite intersection of halfspaces, and a valid convex polyhedral outer approximation of this set is given by considering any finite subset of these halfspaces. As in [39], each halfspace is defined by a hyperplane computed by first specifying its normal and subsequently computing a suitable intercept through the solution of a dynamic optimization problem. However, unlike the method presented in [39], these optimization problems are guaranteed to be convex, even when the reachable set is nonconvex.

## 9.2 Problem Statement

Let $I = [t_0, t_f] \subset \mathbb{R}$ and $P \subset \mathbb{R}^{n_p}$ be compact intervals and let $\mathbf{x} : I \times P \to \mathbb{R}^{n_x}$ be a continuous function such that $\mathbf{x}(\cdot, \mathbf{p})$ is absolutely continuous on $I$ for every $\mathbf{p} \in P$.

For the developments in this chapter, it is irrelevant whether this function is the solution of a system of parametric ODEs, as in Chapter 7, or the solution of a system of semi-explicit DAEs, as in Chapter 8 (in the latter case we interpret $\mathbf{x}$ as the complete vector of DAE states $(\mathbf{x}, \mathbf{y})$). We only assume that, by one of the methods in those chapters, state relaxations $\mathbf{x}^{cv}, \mathbf{x}^{cc} : I \times P \to \mathbb{R}^{n_x}$ are available; i.e., for every $t \in I$, $\mathbf{x}^{cv}(t, \cdot)$ is convex on $P$, $\mathbf{x}^{cc}(t, \cdot)$ is concave on $P$, and $\mathbf{x}^{cv}(t, \mathbf{p}) \leq \mathbf{x}(t, \mathbf{p}) \leq \mathbf{x}^{cc}(t, \mathbf{p})$, $\forall \mathbf{p} \in P$. The objective of this chapter is to solve the following problem.

**Problem 9.2.1.** Given any fixed $t \in I$, compute a convex set $A \subset \mathbb{R}^{n_x}$ such that the image of the interval $P$ under $\mathbf{x}(t, \cdot)$ is contained in $A$.

## 9.3 Convex enclosures of reachable sets

In this section, state relaxations are used in order to construct a convex enclosure of the image $\mathbf{x}(t, P)$ for some fixed $t \in I$. As proven in the following theorem, the desired convex enclosure of the image $\mathbf{x}(t, P)$ is the set

$$A \equiv \bigcup_{\mathbf{p} \in P} [\mathbf{x}^{cv}(t, \mathbf{p}), \mathbf{x}^{cc}(t, \mathbf{p})]. \tag{9.1}$$

Unfortunately, this set is not immediately useful for computations and further derivations will be required to arrive at a more useful formulation.

**Theorem 9.3.1.** *$A$ is convex and contains $\mathbf{x}(t, P)$.*

*Proof.* Given any $\mathbf{p} \in P$, $\mathbf{x}^{cv}(t, \mathbf{p}) \leq \mathbf{x}(t, \mathbf{p}) \leq \mathbf{x}^{cc}(t, \mathbf{p})$ by the definition of $\mathbf{x}^{cv}$ and $\mathbf{x}^{cc}$, and hence $\mathbf{x}(t, \mathbf{p}) \in A$. It remains to show that $A$ is convex. Let $\mathbf{z}_1, \mathbf{z}_2 \in A$ and choose any $\lambda \in [0, 1]$. By definition, $\exists \mathbf{p}_1, \mathbf{p}_2 \in P$ such that $\mathbf{z}_1 \in [\mathbf{x}^{cv}(t, \mathbf{p}_1), \mathbf{x}^{cc}(t, \mathbf{p}_1)]$ and $\mathbf{z}_2 \in [\mathbf{x}^{cv}(t, \mathbf{p}_2), \mathbf{x}^{cc}(t, \mathbf{p}_2)]$. Using these inclusions along with the convexity and concavity of $\mathbf{x}^{cv}(t,\cdot)$ and $\mathbf{x}^{cc}(t,\cdot)$ on $P$, respectively,

$$\begin{aligned}\mathbf{x}^{cv}(t,\lambda\mathbf{p}_1+(1-\lambda)\mathbf{p}_2) &\leq \lambda\mathbf{x}^{cv}(t,\mathbf{p}_1)+(1-\lambda)\mathbf{x}^{cv}(t,\mathbf{p}_2)\\ &\leq \lambda\mathbf{z}_1+(1-\lambda)\mathbf{z}_2\\ &\leq \lambda\mathbf{x}^{cc}(t,\mathbf{p}_1)+(1-\lambda)\mathbf{x}^{cc}(t,\mathbf{p}_2)\\ &\leq \mathbf{x}^{cc}(t,\lambda\mathbf{p}_1+(1-\lambda)\mathbf{p}_2).\end{aligned}$$

But $\lambda\mathbf{p}_1+(1-\lambda)\mathbf{p}_2 \in P$, which implies that $\lambda\mathbf{z}_1+(1-\lambda)\mathbf{z}_2 \in A$ and hence $A$ is convex. □

Though $A$ is in fact a convex enclosure of $\mathbf{x}(t,P)$ as desired, it is not immediately useful for computation because it is expressed in terms of an infinite union of intervals. In the following section, it shown that $A$ can be expressed more usefully as an infinite intersection of halfspaces which can be computed efficiently.

### 9.3.1 A dual representation of $A$

Let $S^{n_x}$ denote the unit sphere in $\mathbb{R}^{n_x}$ with respect to the one norm, and define

$$d^*(\boldsymbol{\mu}) = \min_{\mathbf{z}\in A} \boldsymbol{\mu}^{\mathrm{T}}\mathbf{z}, \quad \forall \boldsymbol{\mu} \in S^{n_x}. \tag{9.2}$$

This function is well defined on account of the following lemma. Also, note that, for each $\boldsymbol{\mu} \in S^{n_x}$, the optimization problem defining $d^*(\boldsymbol{\mu})$ is guaranteed to be convex by Theorem 9.3.1.

**Lemma 9.3.2.** *A is compact*

*Proof.* Since $P$ is compact and $\mathbf{x}^{cv}(t,\cdot)$ and $\mathbf{x}^{cc}(t,\cdot)$ are continuous on $P$, these functions are bounded on $P$ and it follows that $A$ is also bounded. To show that $A$ is closed, consider any convergent sequence of elements of $A$, $\{\mathbf{z}_n\} \to \mathbf{z}^*$. By the definition of $A$, there exists a corresponding sequence $\{\mathbf{p}_n\}$ in $P$ such that $\mathbf{z}_n \in [\mathbf{x}^{cv}(t,\mathbf{p}_n),\mathbf{x}^{cc}(t,\mathbf{p}_n)]$, $\forall n \in \mathbb{N}$. Compactness of $P$ then implies that there exists a convergent subsequence $\{\mathbf{p}_{n_k}\} \to \mathbf{p}^* \in P$, and by taking limits, the continuity of $\mathbf{x}^{cv}(t,\cdot)$ and $\mathbf{x}^{cc}(t,\cdot)$ on $P$

ensures that $\mathbf{z}^* \in [\mathbf{x}^{cv}(t,\mathbf{p}^*), \mathbf{x}^{cc}(t,\mathbf{p}^*)] \subset A$. Thus, $A$ is also closed and compactness follows from the Heine-Borel Theorem. □

**Corollary 9.3.3.** *For every $\boldsymbol{\mu} \in S^{n_x}$, $\exists \mathbf{z}^* \in A$ such that $d^*(\boldsymbol{\mu}) = \boldsymbol{\mu}^{\mathrm{T}}\mathbf{z}^*$.*

It is now possible to formulate an alternate representation of $A$ as an infinite intersection of halfspaces. Define the halfspaces

$$H^+(\boldsymbol{\mu}) \equiv \{\mathbf{z} \in \mathbb{R}^{n_x} : \boldsymbol{\mu}^{\mathrm{T}}\mathbf{z} \geq d^*(\boldsymbol{\mu})\},$$

for all $\boldsymbol{\mu} \in S^{n_x}$. Now let

$$A^* \equiv \bigcap_{\boldsymbol{\mu} \in S^{n_x}} H^+(\boldsymbol{\mu}). \tag{9.3}$$

**Theorem 9.3.4.** *$A^* = A$.*

*Proof.* For any $\boldsymbol{\mu} \in S^{n_x}$, the definition of $d^*(\boldsymbol{\mu})$ ensures that $\boldsymbol{\mu}^{\mathrm{T}}\mathbf{z} \geq d^*(\boldsymbol{\mu})$, $\forall \mathbf{z} \in A$. Therefore, $A \subset H^+(\boldsymbol{\mu})$, $\forall \boldsymbol{\mu} \in S^{n_x}$, and hence $A \subset A^*$. To conclude that $A^* \subset A$, it is assumed that $\hat{\mathbf{z}} \notin A$ and shown that $\hat{\mathbf{z}} \notin A^*$. Because $A$ is closed and convex, the separating hyperplane theorem furnishes $\boldsymbol{\sigma}$ such that $\boldsymbol{\sigma}^{\mathrm{T}}\hat{\mathbf{z}} < \boldsymbol{\sigma}^{\mathrm{T}}\mathbf{z}$, $\forall \mathbf{z} \in A$ (Proposition B.14 in [23]). Letting, $\boldsymbol{\mu} = \boldsymbol{\sigma}/\|\boldsymbol{\sigma}\|_1$, we have $\boldsymbol{\mu}^{\mathrm{T}}\hat{\mathbf{z}} < \boldsymbol{\mu}^{\mathrm{T}}\mathbf{z}$, $\forall \mathbf{z} \in A$, which implies that $\hat{\mathbf{z}} \notin H^+(\boldsymbol{\mu})$ and hence $\hat{\mathbf{z}} \notin A^*$. □

### 9.3.2 Computation of $A$

Given the alternate representation of $A$ as the infinite intersection of halfspaces $A^*$, it is possible to compute a convex polyhedral enclosure of $A$, and hence of $\mathbf{x}(t,P)$, by considering any finite number of these halfspaces. In particular, choosing any $\boldsymbol{\mu}^{[1]}, \ldots, \boldsymbol{\mu}^{[m]} \in S^{n_x}$, a convex polyhedral enclosure of $\mathbf{x}(t,P)$ is given by

$$\mathcal{P}_A(\boldsymbol{\mu}^{[1]}, \ldots, \boldsymbol{\mu}^{[m]}) \equiv \bigcap_{j=1}^{m} H^+(\boldsymbol{\mu}^{[j]}). \tag{9.4}$$

In order to characterize the set $\mathcal{P}_A(\boldsymbol{\mu}^{[1]}, \ldots, \boldsymbol{\mu}^{[m]})$ completely, it is necessary to compute $d^*(\boldsymbol{\mu}^{[j]})$ for all $j = 1, \ldots, m$. This task is simplified by the following lemma.

**Lemma 9.3.5.** *For any $\boldsymbol{\mu} \in S^{n_x}$,*

$$d^*(\boldsymbol{\mu}) = \min_{\mathbf{p}\in P} \sum_{i=1}^{n_x} \min\left(\mu_i x_i^{cv}(t,\mathbf{p}), \mu_i x_i^{cc}(t,\mathbf{p})\right).$$

*Proof.* Choose any $\boldsymbol{\mu} \in S^{n_x}$ and let $\mathbf{z}^* \in A$ be such that $d^*(\boldsymbol{\mu}) = \boldsymbol{\mu}^{\mathrm{T}}\mathbf{z}^*$ (Corollary 9.3.3). Since $\mathbf{z}^* \in A$, $\exists \mathbf{p}^* \in P$ such that $\mathbf{z}^* \in [\mathbf{x}^{cv}(t,\mathbf{p}^*), \mathbf{x}^{cc}(t,\mathbf{p}^*)]$. For any such $\mathbf{p}^*$,

$$\begin{aligned}\min_{\mathbf{z}\in[\mathbf{x}^{cv}(t,\mathbf{p}^*),\mathbf{x}^{cc}(t,\mathbf{p}^*)]} \boldsymbol{\mu}^{\mathrm{T}}\mathbf{z} &\le \boldsymbol{\mu}^{\mathrm{T}}\mathbf{z}^* \\ &= d^*(\boldsymbol{\mu}) \\ &\le \min_{\mathbf{z}\in[\mathbf{x}^{cv}(t,\mathbf{p}^*),\mathbf{x}^{cc}(t,\mathbf{p}^*)]} \boldsymbol{\mu}^{\mathrm{T}}\mathbf{z},\end{aligned}$$

where the first inequality follows from feasibility of $\mathbf{z}^*$ in $[\mathbf{x}^{cv}(t,\mathbf{p}^*), \mathbf{x}^{cc}(t,\mathbf{p}^*)]$ and the second holds because $\mathbf{z}^*$ is optimal in $A \supset [\mathbf{x}^{cv}(t,\mathbf{p}^*), \mathbf{x}^{cc}(t,\mathbf{p}^*)]$. Clearly, these inequalities imply that

$$\begin{aligned}d^*(\boldsymbol{\mu}) &= \min_{\mathbf{z}\in[\mathbf{x}^{cv}(t,\mathbf{p}^*),\mathbf{x}^{cc}(t,\mathbf{p}^*)]} \boldsymbol{\mu}^{\mathrm{T}}\mathbf{z} \\ &= \sum_{i=1}^{n_x} \min\left(\mu_i x_i^{cv}(t,\mathbf{p}^*), \mu_i x_i^{cc}(t,\mathbf{p}^*)\right).\end{aligned}$$

Finally, if $\exists \hat{\mathbf{p}} \in P$ such that

$$\begin{aligned}&\sum_{i=1}^{n_x} \min\left(\mu_i x_i^{cv}(t,\hat{\mathbf{p}}), \mu_i x_i^{cc}(t,\hat{\mathbf{p}})\right) \\ &\qquad < \sum_{i=1}^{n_x} \min\left(\mu_i x_i^{cv}(t,\mathbf{p}^*), \mu_i x_i^{cc}(t,\mathbf{p}^*)\right),\end{aligned}$$

then the vector $\hat{\mathbf{z}}$ defined by $\hat{z}_i = x_i^{cv}(t,\hat{\mathbf{p}})$ if $\mu_i \ge 0$ and $\hat{z}_i = x_i^{cc}(t,\hat{\mathbf{p}})$ otherwise is an element of $[\mathbf{x}^{cv}(t,\hat{\mathbf{p}}), \mathbf{x}^{cc}(t,\hat{\mathbf{p}})]$, and hence of $A$, for which $\boldsymbol{\mu}^{\mathrm{T}}\hat{\mathbf{z}} < \boldsymbol{\mu}^{\mathrm{T}}\mathbf{z}^*$, which is a contradiction. Therefore, since $\mathbf{p}^* \in P$,

$$d^*(\boldsymbol{\mu}) = \min_{\mathbf{p}\in P} \sum_{i=1}^{n_x} \min\left(\mu_i x_i^{cv}(t,\mathbf{p}), \mu_i x_i^{cc}(t,\mathbf{p})\right).$$

□

Fixing any $\boldsymbol{\mu} \in S^{n_x}$, the previous lemma defines $d^*(\boldsymbol{\mu})$ as the solution value of a convex dynamic optimization problem; convexity follows from the sign of each $\mu_i$ and the convexity and concavity, respectively, of $\mathbf{x}^{cv}(t, \cdot)$ and $\mathbf{x}^{cc}(t, \cdot)$ on $P$, while the program is dynamic because evaluating the state relaxations $\mathbf{x}^{cv}$ and $\mathbf{x}^{cc}$ requires the solution of an auxiliary system of ODEs (see Chapters 7 and 8). Programs of this type are easily solved using modern dynamic simulation techniques in conjunction with a local NLP solver. Thus, computation of the enclosure $\mathcal{P}_A(\boldsymbol{\mu}^{[1]}, \ldots, \boldsymbol{\mu}^{[m]})$ requires the solution of $m$ convex dynamic optimization problems. Owing to the use of the state relaxations $\mathbf{x}^{cv}$ and $\mathbf{x}^{cc}$, the convexity of these programs holds even when the image $\mathbf{x}(t, P)$ is nonconvex.

## 9.4 Numerical Example

All numerical experiments in this section were performed on a Dell Precision T3400 workstation with a 2.83 GHz Intel Core2 Quad CPU. One core and 512 MB of memory were dedicated to each job.

**Example 9.4.1.** Consider again the system of parametric ODEs given in Example 7.7.1:

$$\dot{x}_1 = \frac{1}{L}x_2, \quad \dot{x}_2 = -\frac{1}{C}[x_1 - x_2 + \frac{1}{3}x_2^3], \tag{9.5}$$

with $x_{0,1} = x_{0,2} = 1$, $p_1 = (1/C)$, $p_2 = (1/L)$, $t_0 = 0$ and $t_f = 3.5$. In order to compute a polyhedral enclosure of the reachable set of (9.5) at $t_f$, state relaxations were computed using convexity amplifying dynamics as described in Chapter 7. For implementation details, see Example 7.7.1. The solution $x_1(t_f, \cdot)$ and the corresponding state relaxations on the set $P = [0.01, 0.5] \times [0.01, 0.5]$ are shown in Fig. 9-1.

Using the state relaxations shown in Fig. 9-1, a convex enclosure of the reachable set of (9.5) can be computed as described in §9.3. The diamonds in Fig. 9-2 show points sampled from the reachable set at $t_f$ by evaluating $\mathbf{x}(t_f, \mathbf{p})$ for $\mathbf{p}$ on a

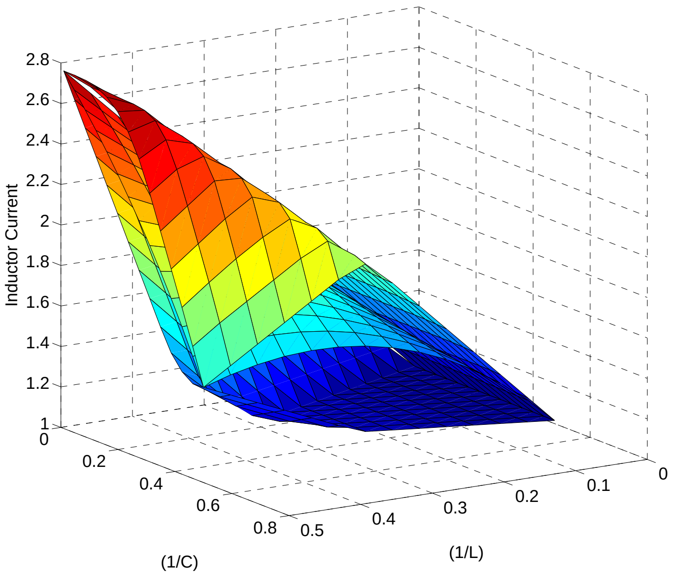

Figure 9-1: State relaxations of the solution of the ODEs (9.5), $x_1(t_f, \cdot)$, on the interval $P = [0.01, 0.5] \times [0.01, 0.5]$.

uniform grid over $P = [0.01, 0.5] \times [0.01, 0.5]$. From these sampled points, it can be seen that the reachable set is almost certainly nonconvex. The circles in the same figure show points sampled from the set $A$ (see (9.1)) by again considering a uniform grid over $P$, and for each $\mathbf{p}$ on this grid, sampling several points from the interval $[\mathbf{x}^{cv}(t_f, \mathbf{p}), \mathbf{x}^{cc}(t_f, \mathbf{p})]$. Of course, it is impossible to compute $A$ finitely using the representation (9.1). On the other hand, the dual representation of $A$ ($A^*$ in §9.3.1) can be used to compute a polyhedral enclosure of $A$, and hence $\mathbf{x}(t, P)$, of the form (9.4). Such an enclosure is shown by the solid lines in Fig. 9-2, which correspond to the multipliers $\boldsymbol{\mu}^{[1]} = [1\ 0]^{\mathrm{T}}$, $\boldsymbol{\mu}^{[2]} = [0\ 1]^{\mathrm{T}}$, $\boldsymbol{\mu}^{[3]} = [0.5\ 0.5]^{\mathrm{T}}$, $\boldsymbol{\mu}^{[4]} = [0.5\ {-0.5}]^{\mathrm{T}}$, $\boldsymbol{\mu}^{[5]} = [-0.75\ 0.25]^{\mathrm{T}}$, $\boldsymbol{\mu}^{[6]} = [0.95\ 0.05]^{\mathrm{T}}$ and $\boldsymbol{\mu}^{[6+j]} = -\boldsymbol{\mu}^{[j]}$ for $j = 1, \ldots, 4$. Clearly, the resulting convex polyhedral set encloses the nonconvex image $\mathbf{x}(t_f, P)$.

Due to the size of this example, the cost of adequately sampling $\mathbf{x}(t_f, P)$ is comparable to that of computing the convex polyhedral enclosure in Figure 9-2. However, the cost of sampling grows exponentially with the number of parameters, while the proposed procedure involves only numerical integration and convex optimization, so is polynomial time.

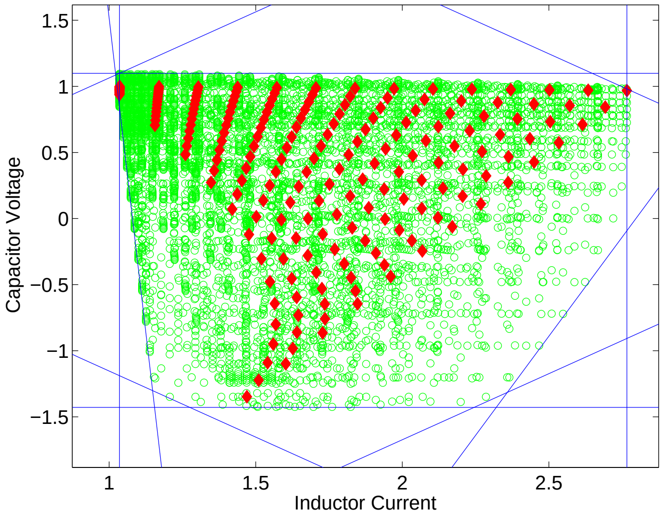

Figure 9-2: Sampled points from the image $\mathbf{x}(t_f, P)$ (diamonds), sampled points from the set $A$ (circles), and supporting hyperplanes to $A$ forming a polyhedral enclosure of $\mathbf{x}(t_f, P)$ (solid lines).

## 9.5 Conclusions

A method has been described which uses state relaxations to compute efficiently a convex polyhedral enclosure of the reachable set of a dynamic system. Given the methods for computing state relaxations in Chapters 7 and 8, this method can be directly applied to systems of parametric ODEs and systems of semi-explicit index one DAEs. Given state relaxations, a convex enclosure of the reachable set for any fixed time is easily formulated, but is expressed in terms of an infinite union of intervals. It was shown that this enclosure can be equivalently expressed as an infinite intersection of halfspaces, so that a convex polyhedral enclosure of the reachable set can be computed by considering only some finite number $m$ of these halfspaces. Computing an appropriate intercept for each halfspace requires the solution of one convex dynamic optimization problem. Unlike other similar methods in the literature, the use of state relaxations ensures that this optimization problem is convex even when the reachable set is nonconvex. Accordingly, a valid convex polyhedral enclosure is obtained by solving $m$ convex dynamic optimization problems, even in the case where the reachable set is itself nonconvex. This procedure was demonstrated for a small example with a nonconvex reachable set and a valid convex enclosure was obtained.

# Chapter 10

# Deterministic Global Optimization with DAEs Embedded

## 10.1 Introduction

In this chapter, we present a deterministic global optimization algorithm for solving problems with semi-explicit index-one DAEs embedded. This problem has been addressed previously in two articles [55, 42]. In both articles, the authors propose methods based on the simultaneous approach to dynamic optimization. That is, these methods apply a total discretization approach, resulting in a large-scale NLP with equality constraints approximating the original dynamics. To solve this NLP to global optimality, a spatial branch-and-bound (B&B) algorithm is used, as described in §1.3.3. However, given the size of the NLPs generated through the simultaneous approach and the worst-case exponential run-time of the spatial B&B algorithm, this cannot be considered a practical approach to global dynamic optimization. In both articles, it is clear that an adequate discretization of the state variables creates problems which are too large to be solved in reasonable time by a global optimization routine, and coarser discretizations can not represent the original dynamics well enough to produce reliable results (the optimal objective value was found to depend strongly on the discretization).

In [55], a second method was proposed based on the sequential approach to dynamic optimization, and shown to significantly outperform the simultaneous approach for several numerical examples. This method is also based on a spatial B&B procedure. However, the lower bounding procedure is based on a finite sampling step, and therefore this algorithm must be considered heuristic rather than deterministic. For optimization problems with explicit ODEs embedded, this deficiency has been overcome by the method in [135]. Subsequently, other methods have emerged for solving problems with ODEs embedded to global optimality using the sequential approach [164, 104]. However, for problems with DAEs embedded, a deterministic global optimization algorithm based on the sequential approach has not previously been achieved. We accomplish this task here using the relaxation techniques described in Chapter 8.

## 10.2 Problem Statement

In this section, the dynamic optimization problem under consideration is stated formally. The embedded system of DAEs is exactly the same as that considered in Chapters 5, 6 and 8, with the exception that one additional specification is made for the algebraic variables at the initial time. As discussed below, this specification guarantees uniqueness of the DAE solution, so that the dynamic optimization problem is well-posed.

Let $D_t \subset \mathbb{R}$, $D_p \subset \mathbb{R}^{n_p}$, $D_x \subset \mathbb{R}^{n_x}$ and $D_y \subset \mathbb{R}^{n_y}$ be open sets, and let $\mathbf{f} : D_t \times D_p \times D_x \times D_y \to \mathbb{R}^{n_x}$, $\mathbf{g} : D_t \times D_p \times D_x \times D_y \to \mathbb{R}^{n_y}$ and $\mathbf{x}_0 : D_p \to D_x$ be $C^1$ functions. Furthermore, let $I \equiv [t_0, t_f] \subset D_t$ and $P \in \mathbb{I}D_p$. Finally, let $\hat{\mathbf{p}} \in P$ and $\hat{\mathbf{y}}_0 \in D_y$ satisfy $\mathbf{g}(t_0, \hat{\mathbf{p}}, \mathbf{x}_0(\hat{\mathbf{p}}), \hat{\mathbf{y}}_0) = \mathbf{0}$. Now, the embedded system of semi-explicit DAEs is given by

$$\left.\begin{aligned} \dot{\mathbf{x}}(t,\mathbf{p}) &= \mathbf{f}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p})) \\ \mathbf{0} &= \mathbf{g}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p})) \end{aligned}\right\}, \tag{10.1a}$$

$$\mathbf{x}(t_0,\mathbf{p}) = \mathbf{x}_0(\mathbf{p}). \tag{10.1b}$$

$$\mathbf{y}(t_0,\hat{\mathbf{p}}) = \hat{\mathbf{y}}_0. \tag{10.1c}$$

A function $(\mathbf{x},\mathbf{y}) \in C^1(I\times P, D_x)\times C^1(I\times P, D_y)$ is a solution of (10.1) on $I\times P$ if (10.1c) holds, (10.1b) is satisfied for all $\mathbf{p}\in P$, and (10.1a) holds for all $(t,\mathbf{p})\in I\times P$. If in addition $\det\frac{\partial \mathbf{g}}{\partial \mathbf{y}}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p})) \neq 0$, $\forall(t,\mathbf{p})\in I\times P$, then $(\mathbf{x},\mathbf{y})$ is called *regular*. If $\hat{\mathbf{p}}$ and $\hat{\mathbf{y}}_0$ satisfy $\det\frac{\partial \mathbf{g}}{\partial \mathbf{y}}(t_0,\hat{\mathbf{p}},\mathbf{x}_0(\hat{\mathbf{p}}),\hat{\mathbf{y}}_0)\neq 0$, then the existence of a regular solution of (10.1a)-(10.1b) local to $(t_0,\hat{\mathbf{p}},\mathbf{x}_0(\hat{\mathbf{p}}),\hat{\mathbf{y}}_0)$ (Definition 5.3.4) is guaranteed by Theorem 5.3.5. Throughout this chapter, we assume the following.

**Assumption 10.2.1.** A regular solution $(\mathbf{x},\mathbf{y})$ of (10.1) exists on all of $I\times P$.

It was shown in Example 5.3.1 that there may be multiple regular solutions of (10.1a)-(10.1b). However, the specification (10.1c) ensures that the solution of Assumption 10.2.1 is unique. This fact follows directly from Corollary 5.3.6. In the remainder of this chapter, the notation $(\mathbf{x},\mathbf{y})$ refers specifically to this solution.

Let $(\phi,\mathbf{h}): D_p\times D_x\times D_y \to \mathbb{R}\times\mathbb{R}^{n_c}$ be continuous. The dynamic optimization problem addressed in this chapter is stated as follows:

**Problem 10.2.1.**

$$\begin{aligned} \min_{\mathbf{p}\in P} \quad & \phi(\mathbf{p},\mathbf{x}(t_f,\mathbf{p}),\mathbf{y}(t_f,\mathbf{p})) \\ \text{s.t.} \quad & \mathbf{h}(\mathbf{p},\mathbf{x}(t_f,\mathbf{p}),\mathbf{y}(t_f,\mathbf{p})) \le \mathbf{0}, \end{aligned} \tag{10.2}$$

where $(\mathbf{x},\mathbf{y})$ is the unique solution of (10.1) on $I\times P$.

Note that Problem 10.2.1 is an optimization problem on a Euclidean space. In particular, the state variables are not considered as decisions, because they are uniquely specified for every $(t,\mathbf{p})\in I\times P$. The ability to pose the problem in this way is crucial to the solution method described in the next section, and is made possible by the fact that the DAE solution $(\mathbf{x},\mathbf{y})$ is unique. Again, this uniqueness is a result of the specification (10.1c). In most applications, there is a consistent initial condition of interest, so that this specification is easily made. On the other hand, if one is interested in an optimization problem that considers all possible solutions of (10.1), then some additional method will be required for exhaustively enumerating such solutions. We do not pursue such an algorithm here. Several simple extensions of Problem 10.2.1 are discussed in the following remark.

**Remark 10.2.2.**

1. The optimization formulation above does not include integral terms in the objective and constraints; i.e.,

   $$\begin{aligned} \min_{\mathbf{p}\in P} \quad & \phi(\mathbf{p}, \mathbf{x}(t_f, \mathbf{p}), \mathbf{y}(t_f, \mathbf{p})) + \int_{t_0}^{t_f} \psi(s, \mathbf{p}, \mathbf{x}(s, \mathbf{p}), \mathbf{y}(s, \mathbf{p}))ds \\ \text{s.t.} \quad & \mathbf{h}(\mathbf{p}, \mathbf{x}(t_f, \mathbf{p}), \mathbf{y}(t_f, \mathbf{p})) + \int_{t_0}^{t_f} \boldsymbol{\ell}(s, \mathbf{p}, \mathbf{x}(s, \mathbf{p}), \mathbf{y}(s, \mathbf{p}))ds \leq \mathbf{0}. \end{aligned} \tag{10.3}$$

   However, problems with integral terms can always be recast in the form of Problem 10.2.1 by introducing quadrature variables $(z_\psi, \mathbf{z}_\ell) : I \times P \to \mathbb{R} \times \mathbb{R}^{n_c}$ satisfying the differential equations

   $$\begin{aligned} \dot{z}_\psi(t, \mathbf{p}) &= \psi(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p})), \quad z_\psi(t_0, \mathbf{p}) = 0, \\ \dot{\mathbf{z}}_\ell(t, \mathbf{p}) &= \boldsymbol{\ell}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p})), \quad \mathbf{z}_\ell(t_0, \mathbf{p}) = \mathbf{0}. \end{aligned} \tag{10.4}$$

   From these definitions, it follows that

   $$\begin{aligned} z_\psi(t_f, \mathbf{p}) &= \int_{t_0}^{t_f} \psi(s, \mathbf{p}, \mathbf{x}(s, \mathbf{p}), \mathbf{y}(s, \mathbf{p}))ds, \\ \mathbf{z}_\ell(t_f, \mathbf{p}) &= \int_{t_0}^{t_f} \boldsymbol{\ell}(s, \mathbf{p}, \mathbf{x}(s, \mathbf{p}), \mathbf{y}(s, \mathbf{p}))ds. \end{aligned} \tag{10.5}$$

   Then, the dynamic optimization problem (10.3) can be written in the form of Problem 10.2.1 by augmenting the embedded DAEs (10.1) with the equations (10.4).

2. In parameter estimation problems, the objective and constraints typically depend on the values of the DAE solution at several points in the time interval $I$;

i.e.,

$$\begin{aligned} \min_{\mathbf{p}\in P} \quad & \phi(\mathbf{p},\mathbf{x}(t_0,\mathbf{p}),\ldots,\mathbf{x}(t_m,\mathbf{p}),\mathbf{y}(t_0,\mathbf{p}),\ldots,\mathbf{y}(t_m,\mathbf{p})) \qquad (10.6)\\ \text{s.t.} \quad & \mathbf{h}(\mathbf{p},\mathbf{x}(t_0,\mathbf{p}),\ldots,\mathbf{x}(t_m,\mathbf{p}),\mathbf{y}(t_0,\mathbf{p}),\ldots,\mathbf{y}(t_m,\mathbf{p})) \leq \mathbf{0}. \end{aligned}$$

The algorithm for solving Problem 10.2.1 presented below is easily extended to this case. The restriction to final time terms only simplifies the notation.

In order to use natural McCormick extensions in the proposed optimization algorithm, the following assumption is required throughout.

**Assumption 10.2.3.** The functions $\mathbf{x}_0$, $\mathbf{f}$, $\mathbf{g}$, $\frac{\partial \mathbf{g}}{\partial \mathbf{y}}$, $\phi$ and $\mathbf{h}$ are $\mathcal{L}$-factorable with natural McCormick extensions $\mathbf{x}_0 : \mathcal{D}_0 \to \mathbb{MR}^{n_x}$, $\{\mathbf{f}\} : \mathcal{D} \to \mathbb{MR}^{n_x}$, $\{\mathbf{g}\} : \mathcal{D} \to \mathbb{MR}^{n_y}$, $\{\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\} : \mathcal{D} \to \mathbb{MR}^{n_y \times n_y}$, $\{\phi\} : \mathcal{E} \to \mathbb{MR}$ and $\{\mathbf{h}\} : \mathcal{E} \to \mathbb{MR}^{n_c}$.

## 10.3 A Global Optimization Algorithm

In this section, a deterministic algorithm is described for solving Problem 10.2.1 to global optimality. Since Problem 10.2.1 is formulated as an optimization problem on a Euclidean space, the basic approach is to apply a standard spatial branch-and-bound algorithm, as discussed in §1.3.3. For each node visited by the algorithm, it is necessary to provide upper and lower bounds on the globally optimal objective value of the following subproblem, where $P^\ell \in \mathbb{I}P$:

**Problem 10.3.1.**

$$\min_{\mathbf{p}\in P^\ell} \quad \phi(\mathbf{p},\mathbf{x}(t_f,\mathbf{p}),\mathbf{y}(t_f,\mathbf{p})) \qquad (10.7)$$

$$\text{s.t.} \quad \mathbf{h}(\mathbf{p},\mathbf{x}(t_f,\mathbf{p}),\mathbf{y}(t_f,\mathbf{p})) \leq \mathbf{0}, \qquad (10.8)$$

where $(\mathbf{x},\mathbf{y})$ is the unique solution of (10.1) on $I \times P^\ell$.

In the following sections, the computation of these bounds is described in detail. A complete statement of the proposed algorithm is given in §10.3.5.

### 10.3.1 The Upper-Bounding Procedure

To compute an upper bound on the globally optimal objective value of Problem 10.3.1, this problem is solved to local optimality using the sequential approach. The optimization is done using the package SNOPT [69]. SNOPT uses a sparse sequential-quadratic-programming algorithm with quasi-Newton approximations of the Hessian, and is specialized to problems where the objective and constraints, and their gradients, are expensive to evaluate. This is true for Problem 10.3.1 because these evaluations require numerical integration and sensitivity analysis of the embedded DAEs. Numerical integration is done using the package IDAS [82]. For any given $\mathbf{p} \in P^{\ell}$, the initial condition $\mathbf{y}(t_0, \mathbf{p})$ is computed using a consistent initialization routine for semi-explicit DAEs provided in IDAS. An initial guess function $\mathbf{y}_0^{\text{guess}} : P \to \mathbb{R}^{n_y}$ must be provided by the user such that the consistent initialization solver converges to $\mathbf{y}(t_0, \mathbf{p})$ from $\mathbf{y}_0^{\text{guess}}(\mathbf{p})$, for all $\mathbf{p} \in P$. IDAS provides parametric sensitivities for the solution of the embedded DAEs automatically, which are used to evaluate the gradients of the objective and constraints in Problem 10.3.1. Differentiation of the objective and constraint functions and evaluation of the right-hand sides of the sensitivity system are done by forward mode automatic differentiation using the package FADBAD++ (http://www.fadbad.com). All solver tolerances are given in §10.4.

### 10.3.2 The Lower-Bounding Procedure

To compute a lower bound on the optimal objective function value of Problem 10.3.1, we construct and solve a convex underestimating program. As discussed in §1.3.4, the primary complication in doing this is that the objective and constraint functions in Problem 10.3.1 are not $\mathcal{L}$-factorable functions of $\mathbf{p}$, so standard relaxation techniques cannot be applied. Of course, this problem is circumvented using the state bounding and relaxation techniques developed throughout this thesis.

### Computing State Bounds

The first step in the lower-bounding procedure is to compute state bounds for $(\mathbf{x}, \mathbf{y})$ on $I \times P^{\ell}$. Using the single-phase method described in Chapter 6, this is accomplished by numerically integrating the systems of DAEs (6.85)-(6.88), with $P^{\ell}$ in place of $P$ in (6.76)-(6.84). For all of the numerical experiments in this chapter, we set $\gamma(t) = 10^{-4}$, $\forall t \in I$ (see (6.87)-(6.88)). The integer $K$, which determines how many Hansen-Sengupta iterations are done when evaluating the right-hand sides of (6.85) and (6.86), is an important parameter in the proposed optimization algorithm. The value of this parameter and its effect on the performance of the algorithm is discussed for specific numerical examples in §10.4.

The DAEs (6.85)-(6.88) are solved using `IDAS` [82]. The initial conditions for the bounds on the differential variables $\mathbf{x}$ are given by (6.89). The initial conditions for the algebraic bounds are found by solving (6.87)-(6.88) at $t_0$ using the consistent initialization routine provided in `IDAS`. The initial guess for this computation is specified as $\mathbf{z}_y^L(t_0) = \mathbf{z}_y^U(t_0) = \mathbf{y}(t_0, m(P^{\ell}))$, where $\mathbf{y}(t_0, m(P^{\ell}))$ is computed as in §10.3.1. The algebraic bounds provided by this initialization problem must contain $\mathbf{y}(t_0, m(P^{\ell}))$. If this is false, then the computed initial bounds pertain to a different regular solution of the embedded DAEs than that specified by $\hat{\mathbf{y}}_0$, and the state bounding algorithm will terminate with an error flag. Otherwise, the solution of (6.85)-(6.89) provides state bounds for $(\mathbf{x}, \mathbf{y})$ on $I \times P^{\ell}$ by Corollary 6.6.3. After numerical integration, the computed state bounds for the algebraic variables are refined by $q = K$ further Hansen-Sengupta iterations, as in the conclusion of Corollary 6.6.3. In practice, this iteration is only done at $t_f$ because only the state bounds at $t_f$ will effect the objective and constraints of the lower bounding problem derived below.

In the remainder of this chapter, the state bounds for $(\mathbf{x}, \mathbf{y})$ on $I \times P^{\ell}$ computed by the procedure above will be denoted by $\mathbf{x}^{L,\ell}, \mathbf{x}^{U,\ell} : I \to \mathbb{R}^{n_x}$ and $\mathbf{y}^{L,\ell}, \mathbf{y}^{U,\ell} : I \to \mathbb{R}^{n_x}$. Furthermore, we define $X^{\ell}(t) \equiv [\mathbf{x}^{L,\ell}(t), \mathbf{x}^{U,\ell}(t)]$ and $Y^{\ell}(t) \equiv [\mathbf{y}^{L,\ell}(t), \mathbf{y}^{U,\ell}(t)]$.

### The Convex Underestimating Subproblem

Once state bounds have been computed, we may derived a convex underestimating program for Problem 10.3.1. In order to use natural McCormick extensions, we make the following assumption:

**Assumption 10.3.1.** The interval $P^\ell \times X^\ell(t_f) \times Y^\ell(t_f)$ is represented in $\mathcal{E}$.

Under Assumption 10.3.1, we may define the functions $(u^\ell_\phi, \mathbf{u}^\ell_h) : P^\ell \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} \to \mathbb{R} \times \mathbb{R}^{n_h}$ by

$$\begin{aligned} u^\ell_\phi(\mathbf{p}, \mathbf{z}^{cv}_x, \mathbf{z}^{cc}_x, \mathbf{z}^{cv}_y, \mathbf{z}^{cc}_y) \equiv \{\phi\}^{cv}(&\mathrm{MC}(\mathbf{p}^{L,\ell}, \mathbf{p}^{U,\ell}, \mathbf{p}, \mathbf{p}), \\ &\mathrm{MC}(\mathbf{x}^{L,\ell}(t_f), \mathbf{x}^{U,\ell}(t_f), \mathbf{z}^{cv}_x, \mathbf{z}^{cc}_x), \\ &\mathrm{MC}(\mathbf{y}^{L,\ell}(t_f), \mathbf{y}^{U,\ell}(t_f), \mathbf{z}^{cv}_y, \mathbf{z}^{cc}_y)), \end{aligned} \tag{10.9}$$

$$\begin{aligned} \mathbf{u}^\ell_h(\mathbf{p}, \mathbf{z}^{cv}_x, \mathbf{z}^{cc}_x, \mathbf{z}^{cv}_y, \mathbf{z}^{cc}_y) \equiv \{\mathbf{h}\}^{cv}(&\mathrm{MC}(\mathbf{p}^{L,\ell}, \mathbf{p}^{U,\ell}, \mathbf{p}, \mathbf{p}), \\ &\mathrm{MC}(\mathbf{x}^{L,\ell}(t_f), \mathbf{x}^{U,\ell}(t_f), \mathbf{z}^{cv}_x, \mathbf{z}^{cc}_x), \\ &\mathrm{MC}(\mathbf{y}^{L,\ell}(t_f), \mathbf{y}^{U,\ell}(t_f), \mathbf{z}^{cv}_y, \mathbf{z}^{cc}_y)). \end{aligned} \tag{10.10}$$

A convex underestimating program for Problem 10.3.1 is now given by:

**Problem 10.3.2.**

$$\begin{aligned} \min_{\mathbf{p}\in P^\ell} \quad & u^\ell_\phi(\mathbf{p}, \mathbf{x}^{cv,\ell}(t_f,\mathbf{p}), \mathbf{x}^{cc,\ell}(t_f,\mathbf{p}), \mathbf{y}^{cv,\ell}(t_f,\mathbf{p}), \mathbf{y}^{cc,\ell}(t_f,\mathbf{p})) \\ \text{s.t.} \quad & \mathbf{u}^\ell_h(\mathbf{p}, \mathbf{x}^{cv,\ell}(t_f,\mathbf{p}), \mathbf{x}^{cc,\ell}(t_f,\mathbf{p}), \mathbf{y}^{cv,\ell}(t_f,\mathbf{p}), \mathbf{y}^{cc,\ell}(t_f,\mathbf{p})) \leq \mathbf{0}, \end{aligned} \tag{10.11}$$

where $\mathbf{x}^{cv,\ell}$, $\mathbf{x}^{cc,\ell}$, $\mathbf{y}^{cv,\ell}$ and $\mathbf{y}^{cc,\ell}$ are state relaxations of $(\mathbf{x}, \mathbf{y})$ on $I \times P^\ell$.

The fact that the objective and constraints of Problem 10.3.2 are convex relaxations of the objective and constraints of Problem 10.3.1 on $P^\ell$, respectively, follows from Theorem 2.7.13.

To compute a lower bound on the optimal objective function value of Problem 10.3.1, Problem 10.3.2 is solved to global optimality. The optimal solution found is denoted by $\check{\mathbf{p}}^\ell$. Due to the use of McCormick's relaxation technique, it is possible that the objective and constraints in Problem 10.3.2 are non-differentiable. However, given subgradients for the state relaxations (see below), subgradients for the objective and constraint functions are easily computed using the subgradient propagation rules for McCormick relaxations developed in [122]. In our implementation, the computation of the natural McCormick extensions in Problem 10.3.2, and their subgradients, is done automatically using the library `MC++` (`http://www3.imperial.ac.uk/people/b.chachuat/research`). Because Problem 10.3.1 is a potentially nonsmooth convex optimization problem, it would be best to solve it using a specialized nonsmooth solver, such as a bundle method [112, 107]. However, these methods are not as mature as those for differentiable problems, and the available solvers of this type remain problematic. For the time being, we have implemented the code `SNOPT` to solve Problem 10.3.2. In lieu of gradient information, `SNOPT` is provided with subgradients as described above. While nonsmoothness in Problem 10.3.2 should be expected to lead to some inefficiency and numerical difficulties in `SNOPT`, this did not cause serious complications for the numerical examples in §10.4.

For each $\mathbf{p} \in P^\ell$ visited by the optimizer during the solution of Problem 10.3.2, state relaxations and subgradients must be evaluated at $(t_f, \mathbf{p})$. State relaxations are computed using the theory of relaxation preserving dynamics developed in Chapter 8. Let $\tilde{\mathbf{x}}^{cv,\ell}$, $\tilde{\mathbf{x}}^{cc,\ell}$, $\tilde{\mathbf{y}}^{cv,\ell}$, and $\tilde{\mathbf{y}}^{cc,\ell}$ denote the solutions of the auxiliary system (8.43), derived with $P^\ell$ in place of $P$. For the state relaxations in Problem 10.3.2, we consider two alternatives:

1. Directly use $(\mathbf{x}^{cv,\ell}, \mathbf{x}^{cc,\ell}, \mathbf{y}^{cv,\ell}, \mathbf{y}^{cc,\ell}) = (\tilde{\mathbf{x}}^{cv,\ell}, \tilde{\mathbf{x}}^{cc,\ell}, \tilde{\mathbf{y}}^{cv,\ell}, \tilde{\mathbf{y}}^{cc,\ell})$,
2. Define $(\mathbf{x}^{cv,\ell}, \mathbf{x}^{cc,\ell}, \mathbf{y}^{cv,\ell}, \mathbf{y}^{cc,\ell})$ as the affine state relaxations specified by the values and subgradients of $(\tilde{\mathbf{x}}^{cv,\ell}, \tilde{\mathbf{x}}^{cc,\ell}, \tilde{\mathbf{y}}^{cv,\ell}, \tilde{\mathbf{y}}^{cc,\ell})$ at $(t_f, m(P^\ell))$.

Clearly, the first option will result in tighter relaxations, and hence a sharper lower bound. On the other hand, this option requires the solution of the auxiliary system (8.43) for every $\mathbf{p} \in P^\ell$ visited by the optimizer during the solution of Problem 10.3.2. In contrast, the second option requires only a single numerical integration of (8.43). In practice, numerical integration of the state bounds and state relaxations dominates the cost of the lower-bounding procedure, even when the second option is used. This makes the first option impractical, so the second option is used for all of the numerical examples in §10.4. The additional conservatism introduced by this linearization when $P^\ell$ is wide is not expected to be as problematic as it is in the affine relaxation method for ODEs [162], as discussed in Chapter 7. This is because here the linearization is applied to the solution of the auxiliary system, whereas the method in [162] uses linearization in the definition of the auxiliary system itself. In the latter case, the conservatism of linearization effects the state relaxations at early times and propagates forward, weakening the state relaxations at later times. In the method here, the conservatism of linearization is introduced only after the solution of the auxiliary system, and does not effect that solution in any way.

To compute state relaxations according to Option 2 above, the auxiliary system (8.43) is solved once at $m(P^\ell)$ to evaluate $\tilde{\mathbf{x}}^{cv,\ell}(t_f, m(P^\ell))$, $\tilde{\mathbf{x}}^{cc,\ell}(t_f, m(P^\ell))$, $\tilde{\mathbf{y}}^{cv,\ell}(t_f, m(P^\ell))$ and $\tilde{\mathbf{y}}^{cc,\ell}(t_f, m(P^\ell))$. Using the state bounds computed by the single phase method of Chapter 6 as described in the previous section, Assumption 8.2.1 holds. Supposing further that the factorability Assumption 8.2.2 holds, the auxiliary system (8.43) is well-defined and all of the participating functions are evaluated by taking natural McCormick extensions. In our implementation, this is done automatically using `MC++`. The integer $K$ in the auxiliary system (8.43), which determines how many refinement iterations are applied to the algebraic state relaxations when evaluating the system right-hand side functions, is an important parameter in the proposed optimization algorithm. The value of this parameter and its effect on the performance of the algorithm is discussed for specific numerical examples in §10.4. The auxiliary system is solved numerically as an explicit system of ODEs with state events as described in §7.6.3.

Evaluating the right-hand side functions of the auxiliary system (8.43) requires values for the state bounds and the time-varying preconditioning matrix $\mathbf{C}$ computed during integration of the state bounds. These quantities are evaluated whenever they are required by interpolation from stored data. Time derivatives of $\mathbf{x}^{L,\ell}$ and $\mathbf{x}^{U,\ell}$ are computed when required by evaluating the right-hand sides of (6.85) and (6.86). This scheme requires that the values of the state bounds, the value of the preconditioner $\mathbf{C}$, and the order of the integration method are stored at every time point visited during numerical integration of the state bounds. An alternative implementation is to integrate the state bounds and state relaxations simultaneously. This was avoided in order to facilitate comparison with the first state relaxation option discussed above, which requires multiple integrations of the state relaxations but only one integration of the state bounds.

It remains to compute subgradients for the state relaxations $\tilde{\mathbf{x}}^{cv,\ell}(t_f,\cdot)$, $\tilde{\mathbf{x}}^{cc,\ell}(t_f,\cdot)$, $\tilde{\mathbf{y}}^{cv,\ell}(t_f,\cdot)$ and $\tilde{\mathbf{y}}^{cc,\ell}(t_f,\cdot)$ at $m(P^\ell)$. First, note that $\tilde{\mathbf{y}}^{cv,\ell}(t_f,\cdot)$ and $\tilde{\mathbf{y}}^{cc,\ell}(t_f,\cdot)$ are given explicitly as functions of $\tilde{\mathbf{x}}^{cv,\ell}(t_f,\cdot)$ and $\tilde{\mathbf{x}}^{cc,\ell}(t_f,\cdot)$ by the iterative refinement in (8.44). Then, it suffices to compute subgradients for $\tilde{\mathbf{x}}^{cv,\ell}(t_f,\cdot)$ and $\tilde{\mathbf{x}}^{cc,\ell}(t_f,\cdot)$ at $m(P^\ell)$. From these, subgradients for $\tilde{\mathbf{y}}^{cv,\ell}(t_f,\cdot)$ and $\tilde{\mathbf{y}}^{cc,\ell}(t_f,\cdot)$ are computed by applying the rules for subgradient propagation for McCormick relaxations to the equations (8.44) using `MC++` [122].

Subgradients for the functions $\tilde{\mathbf{x}}^{cv,\ell}(t_f,\cdot)$ and $\tilde{\mathbf{x}}^{cc,\ell}(t_f,\cdot)$ are computed by sensitivity analysis. Recall that the auxiliary system (8.43) is solved for $\tilde{\mathbf{x}}^{cv,\ell}$ and $\tilde{\mathbf{x}}^{cc,\ell}$ as an explicit system of ODEs with state events. We consider first the computation of subgradients by sensitivity analysis on an interval of time $[t_{e1},t_{e2}]$ during which the mode of the auxiliary system does not change. That is, the Boolean variables $b_i^{cv}$ and $b_i^{cc}$ in (8.43), for $i=1,\ldots,n_x$, are constant. In any such mode, $\tilde{\mathbf{x}}^{cv,\ell}$ and $\tilde{\mathbf{x}}^{cc,\ell}$ evolve according to a system of explicit ODEs with continuous right-hand side functions of the general form

$$\dot{\tilde{\mathbf{x}}}^{cv}(t,\mathbf{p}) = \mathbf{w}^{cv}(t,\mathbf{p},\tilde{\mathbf{x}}^{cv}(t,\mathbf{p}),\tilde{\mathbf{x}}^{cc}(t,\mathbf{p})), \tag{10.12}$$

$$\dot{\tilde{\mathbf{x}}}^{cc}(t,\mathbf{p}) = \mathbf{w}^{cc}(t,\mathbf{p},\tilde{\mathbf{x}}^{cv}(t,\mathbf{p}),\tilde{\mathbf{x}}^{cc}(t,\mathbf{p})). \tag{10.13}$$

In the case where $\tilde{\mathbf{x}}^{cv,\ell}(t_{1e},\cdot)$ and $\tilde{\mathbf{x}}^{cc,\ell}(t_{1e},\cdot)$ are differentiable at $m(P^\ell)$ and the functions

$$\mathbf{w}^{cv}(t,\cdot,\tilde{\mathbf{x}}^{cv}(t,\cdot),\tilde{\mathbf{x}}^{cc}(t,\cdot)) \quad \text{and} \quad \mathbf{w}^{cc}(t,\cdot,\tilde{\mathbf{x}}^{cv}(t,\cdot),\tilde{\mathbf{x}}^{cc}(t,\cdot)) \tag{10.14}$$

are continuously differentiable at $m(P^\ell)$, for every $t \in [t_{e1}, t_{e2}]$, it is well know that the solutions $\tilde{\mathbf{x}}^{cv,\ell}(t_{2e}, \cdot)$ and $\tilde{\mathbf{x}}^{cc,\ell}(t_{2e}, \cdot)$ are differentiable at $m(P^\ell)$ as well. By convexity (resp. concavity), these derivatives are equivalent to the unique subgradients of $\tilde{\mathbf{x}}^{cv,\ell}(t_{2e}, \cdot)$ and $\tilde{\mathbf{x}}^{cc,\ell}(t_{2e}, \cdot)$ at $m(P^\ell)$. Moreover, these derivatives can be computed as the final time solutions of a sensitivity system; i.e., a system of explicit ODEs, coupled to (10.12), whose initial conditions are the derivatives of $\tilde{\mathbf{x}}^{cv,\ell}(t_{1e}, \cdot)$ and $\tilde{\mathbf{x}}^{cc,\ell}(t_{1e}, \cdot)$ at $m(P^\ell)$ and whose right-hand side functions map the derivatives of $\tilde{\mathbf{x}}^{cv,\ell}(t, \cdot)$ and $\tilde{\mathbf{x}}^{cc,\ell}(t, \cdot)$ at $m(P^\ell)$ to the derivatives of the functions (10.14) at $m(P^\ell)$, for any $t \in [t_{e1}, t_{e2}]$. However, due to the use of McCormick relaxations in the auxiliary system (8.43), the functions in (10.14) are potentially non-differentiable at $m(P^\ell)$. Therefore, we define the sensitivity system for (8.43) by specifying the initial conditions as subgradients of $\tilde{\mathbf{x}}^{cv,\ell}(t_{1e}, \cdot)$ and $\tilde{\mathbf{x}}^{cc,\ell}(t_{1e}, \cdot)$ at $m(P^\ell)$, and by defining the right-hand sides as the functions that map a given pair of subgradients for $\tilde{\mathbf{x}}^{cv,\ell}(t, \cdot)$ and $\tilde{\mathbf{x}}^{cc,\ell}(t, \cdot)$ at $m(P^\ell)$ to subgradients of the functions (10.14) at $m(P^\ell)$ according to the subgradient propagation rules for McCormick relaxations [122].

A formal proof that the modified sensitivity system described above furnishes subgradients of $\tilde{\mathbf{x}}^{cv,\ell}(t_{2e}, \cdot)$ and $\tilde{\mathbf{x}}^{cc,\ell}(t_{2e}, \cdot)$ at $m(P^\ell)$ as its final time solution is left for future work. We note, however, that it may be possible in many cases to ensure differentiability of $\tilde{\mathbf{x}}^{cv,\ell}(t_{2e}, \cdot)$ and $\tilde{\mathbf{x}}^{cc,\ell}(t_{2e}, \cdot)$ at $m(P^\ell)$, in which case the validity of our approach follows directly. In particular, it is known that $\tilde{\mathbf{x}}^{cv,\ell}(t_{2e}, \cdot)$ and $\tilde{\mathbf{x}}^{cc,\ell}(t_{2e}, \cdot)$ will be differentiable at $m(P^\ell)$ if the points in time at which the functions (10.14) are non-differentiable at $m(P^\ell)$ form a set of measure zero in $[t_{1e}, t_{2e}]$ [185]. In this case, the derivatives of $\tilde{\mathbf{x}}^{cv,\ell}(t_{2e}, \cdot)$ and $\tilde{\mathbf{x}}^{cc,\ell}(t_{2e}, \cdot)$ at $m(P^\ell)$ are again given as the final time solutions of a standard sensitivity system with one exception; the value of the right-hand sides of the sensitivity system may take arbitrary values at all points $t \in [t_{1e}, t_{2e}]$ for which the functions in (10.14) are non-differentiable at $m(P^\ell)$. At points of differentiability, the functions (10.14) have a unique subgradient that is equal to the derivative. It follows that the sensitivity approach described above will furnish the true derivatives $\tilde{\mathbf{x}}^{cv,\ell}(t_{2e}, \cdot)$ and $\tilde{\mathbf{x}}^{cc,\ell}(t_{2e}, \cdot)$ at $m(P^\ell)$ in this case. Given the sources of nonsmoothness in McCormick's relaxation technique, it seems unlikely that the right-hand side functions of the auxiliary system, within a single mode, will be non-differentiable other than at a finite number of points in $I$. An important area for future work is to formalize verifiable conditions under which this is the case.

From the discussion above, we can derive a sensitivity system corresponding to every mode of the auxiliary system (8.43). Whenever an event occurs during the simulation of (8.43), there is a change in the mode of the system, and a corresponding change in the sensitivity system. Moreover, the sensitivity values themselves may be reset at the event time, since the final time sensitivities computed in the previous mode will not necessarily be valid initial conditions for the sensitivity equations in the new mode. Suppose for example that an event occurs at $t_e \in I$ in which $b_i^{cv}$ changes. In such an event, the right-hand side function describing $\dot{\tilde{x}}_i^{cv,\ell}$ changes discontinuously at $t_e$, and the equations for the sensitivities for $\tilde{x}_i^{cv,\ell}$ are changed accordingly. In addition, the sensitivities for $x_i^{cv,\ell}$ are reset before integration is resumed. If $b_i^{cv}$ changes from 0 to 1, then this event signifies that $\tilde{x}_i^{cv,\ell}(t_e, m(P^\ell))$ has reached the lower bound $x_i^L(t_e)$ and will slide along this bound to the right of $t_e$. In this case, the sensitivities for $\tilde{x}_i^{cv,\ell}(t_e, \cdot)$ are reset to $\mathbf{0}$ before integration is resumed. Since it was proven in Chapter 8 that $\tilde{x}_i^{cv,\ell}(t, m(P^\ell)) \geq x_i^L(t)$ for all $t \in I$, $\mathbf{0}$ is a valid subgradient for $\tilde{x}_i^{cv,\ell}(t, m(P^\ell))$ whenever $\tilde{x}_i^{cv,\ell}(t, m(M^\ell)) = x_i^L(t)$. In the opposite event, where $b_i^{cv}$ is changed from 1 to 0, the sensitivities for $\tilde{x}_i^{cv,\ell}$ will already be $\mathbf{0}$, so that no reset is required. In both cases, it can be shown that these reinitializations are consistent with the sensitivity theory for hybrid systems developed in [67]. For all other relaxations aside from $\tilde{x}_i^{cv,\ell}$, the corresponding right-hand side functions in the auxiliary system do not suffer a discontinuity at $t_e$, so no changes are required either in the sensitivities or their right-hand side functions for these relaxations.

Using the scheme outlined above, numerical integration of the auxiliary system and the appended sensitivity system is done by the code `CVODES` [82], using built-in sensitivity analysis and event detection features. Solver tolerances are given in §10.4. Rarely, very long integration times are observed for the auxiliary system due to chattering in the event detection scheme. To avoid this cost during optimization, the number of events is limited to $8n_x$. If this number is exceeded, the lower-bounding procedure is aborted and the lower bound for the problematic node is not updated. The maximum number of integration steps between events or times $t_i$ at which the state relaxation values are required for evaluating the objective or constraints of Problem 10.3.2 (see Remark 10.2.2 (2)) is set to the CVODES default of 500.

### 10.3.3 Domain Reduction

Upon successful solution of the lower bounding problem, SNOPT provides a vector $\boldsymbol{\mu} \in \mathbb{R}^{n_p}$ of duality multipliers for the constraints $\mathbf{p}^{L,\ell} \leq \mathbf{p} \leq \mathbf{p}^{U,\ell}$. The $i^{\text{th}}$ multiplier is positive if the $i^{\text{th}}$ lower bound is active, negative if the $i^{\text{th}}$ upper bound is active, and zero otherwise. If the node in question is not fathomed by value dominance (see §10.3.5), then these multipliers can be used to refine the interval $P^\ell$ by a standard procedure [146]. The refinement is given by

$$\begin{aligned} p_i^{L,\ell} &:= \max\left(p_i^{L,\ell}, p_i^{U,\ell} - \frac{LBD^\ell - UBD + \epsilon}{\mu_i}\right) \quad \text{if} \quad \mu_i < 0, \\ p_i^{U,\ell} &:= \min\left(p_i^{U,\ell}, p_i^{L,\ell} - \frac{LBD^\ell - UBD + \epsilon}{\mu_i}\right) \quad \text{if} \quad \mu_i > 0, \end{aligned} \tag{10.15}$$

for $i = 1, \ldots, n_p$, where $LBD^\ell$ is the optimal objective value for the lower bounding problem in the current node, $UBD$ is the incumbent upper bound, and $\epsilon$ is the absolute branch-and-bound tolerance (see §10.3.5). Though this refinement can be applied iteratively, in the case studies in this chapter it is applied only once per node, in a loop from $i = 1$ to $i = n_p$.

### 10.3.4 Generation Skipping

For many numerical examples, we find that the proposed lower-bounding procedure often fails because the numerical integration of the state bounds fails. As discussed in detail in Chapter 6, this failure is related to an inability to guarantee existence and uniqueness of a solution of the original DAE model, and is especially problematic on very wide intervals $P^\ell$. When the state bounding procedure fails, a lower bound cannot be computed. The problematic node is simply partitioned and its children are placed on the stack. This sequence of events is then repeated until branching produces intervals narrow enough for the state bounding procedure to succeed, whereupon lower bounds become available.

For some problems, this process can account for a large portion of the overall runtime. In particular, a failed attempt to integrate state bounds can be significantly more expensive than a successful one because the integrator may take may steps in its attempts to succeed (the maximum number of integration steps is limited to 1000 for this reason). In such situations, it is clearly advantageous to initialize the stack with a sufficiently fine partition of $P$ and thereby avoid the cost of repeated failures in the state bounding algorithm. There are two complications with this idea in general. First, examples show that a *sufficiently fine partition* can be highly nonuniform. Secondly, the required partition is not known in advance. For these reasons, it is desirable to derive a heuristic which generates an appropriate partition dynamically.

In the algorithm below, this is optionally accomplished by a *generation skipping* heuristic. Simply, if the state bounding procedure fails in a given node, then it is not attempted for any of the children of that node out to $N_{\text{GS}}$ generations, where $N_{\text{GS}}$ is a user specified integer. When such a child is popped from the stack, the upper bounding problem is solved, the node is branched, and its children are returned to the stack. Of course, when $N_{\text{GS}} = 0$, we recover the standard B&B algorithm. The heuristic is off. When $N_{\text{GS}} > 0$, then $P$ is selectively and aggressively partitioned in areas of the search space in which the state bounding procedure has difficulties. There is an obvious tradeoff to this heuristic. When $N_{\text{GS}}$ is small, many expensive failures of the state bounding procedure may occur with no gain of information. When $N_{\text{GS}}$ is large, aggressive partitioning may lead to a large number of nodes representing regions of the search space on which adequate bounds could have been achieved with many fewer nodes.

### 10.3.5 Algorithm Statement

The proposed B&B algorithm is formally stated below. The stack is denoted by $\Sigma$, and has elements of the form $(P^{\ell}, LBD^{\ell}, N^{\ell}_{\text{GS}})$ where $P^{\ell}$ is a subinterval of $P$, $LBD^{\ell}$

is a lower bound on the optimal objective value of the Subproblem 10.3.1, and $N_{\text{GS}}^{\ell}$ is an integer related to the generation skipping heuristic discussed in the previous section. The inputs to the algorithm are $P$, the absolute B&B convergence tolerance $\epsilon > 0$, the integer $N_{\text{GS}}$ defined in §10.3.4, and an integer $N_{\text{MS}}$ defining the mesh size used for computing an initial upper bound using multistart. Upon successful termination, the algorithm produces an interval $[UBD, UBD - \epsilon]$ guaranteed to contain the optimal objective value of Problem 10.2.1, and a feasible point $\mathbf{p}^* \in P$ satisfying $\phi(\mathbf{p}^*, \mathbf{x}(t_f, \mathbf{p}^*), \mathbf{y}(t_f, \mathbf{p}^*)) = UBD$.

**Algorithm** 3 (Global Dynamic Optimization with DAEs Embedded)

1. Input: $P$, $\epsilon$, $N_{\text{GS}}$, $N_{\text{MS}}$.

2. Initialization

   (a) Set $\Sigma = \{(P, -\infty, 0)\}$, $LBD = -\infty$, $UBD = +\infty$, $\mathbf{p}^* = m(P)$.

3. Multistart

   (a) Solve Problem 10.2.1 to local optimality from $(N_{\text{MS}})^{n_p}$ initial guesses on a uniform grid over $P$.

   (b) Set $UBD$ to the lowest objective value found and set $\mathbf{p}^*$ to the corresponding solution value.

4. Termination

   (a) Delete from $\Sigma$ all nodes $(P^{\ell}, LBD^{\ell}, N_{\text{GS}}^{\ell})$ with $LBD^{\ell} \geq UBD - \epsilon$.

   (b) If $\Sigma = \emptyset$, terminate. If $UBD = +\infty$, the instance is infeasible. Otherwise, the optimal objective value lies in $[UBD, UBD - \epsilon]$ and $\mathbf{p}^*$ is a feasible point satisfying $\phi(\mathbf{p}^*, \mathbf{x}(t_f, \mathbf{p}^*), \mathbf{y}(t_f, \mathbf{p}^*)) = UBD$.

5. Node Selection

   (a) Pop and delete a node $(P^{\ell}, LBD^{\ell}, N_{\text{GS}}^{\ell})$ from $\Sigma$ such that $LBD^{\ell}$ is less than or equal to the lower bound of every other node in $\Sigma$. Set $LBD := \max(LBD, LBD^{\ell})$.

6. Generation Skipping

    (a) If $N_{\text{GS}}^{\ell} \neq 0$, go to 9.

7. Lower-Bounding Procedure

    (a) Compute the state bounds $X^{\ell}$ and $Y^{\ell}$.

        i. If this fails, set $N_{\text{GS}}^{\ell} := N_{\text{GS}} + 1$ and go to 9.

        ii. Set $LBD^{\ell} := \max(LBD^{\ell}, [\phi]^{L}(P^{\ell}, X^{\ell}(t_f), Y^{\ell}(t_f)))$.

        iii. (Fathom by infeasibility) If $[h_i]^{L}(P^{\ell}, X^{\ell}(t_f), Y^{\ell}(t_f))) > 0$ for any $i$, go to 4.

        iv. (Fathom by value dominance) If $LBD^{\ell} \geq UBD - \epsilon$, go to 4.

    (b) Solve Problem 10.3.2 to global optimality.

        i. If this fails, go to 9.

        ii. If an optimal solution $\check{\mathbf{p}}$ is found, set

$$\begin{aligned} LBD^{\ell} := \max \big( & LBD^{\ell}, u_{\phi}^{\ell}(\check{\mathbf{p}}, \mathbf{x}^{cv,\ell}(t_f, \check{\mathbf{p}}), \mathbf{x}^{cc,\ell}(t_f, \check{\mathbf{p}}), \\ & \mathbf{y}^{cv,\ell}(t_f, \check{\mathbf{p}}), \mathbf{y}^{cc,\ell}(t_f, \check{\mathbf{p}}))\big). \end{aligned}$$

        iii. (Fathom by infeasibility) If Problem 10.3.2 is infeasible, go to 4.

        iv. (Fathom by value dominance) If $LBD^{\ell} \geq UBD - \epsilon$, go to 4.

8. Domain Reduction (Optional)

    (a) Refine $P^{\ell}$ by executing the assignments (10.15) in a single loop from $i = 1$ to $i = n_p$.

9. Upper-Bounding Procedure

    (a) Solve Problem 10.3.1 to local optimality.

        i. If this fails, go to 10.

        ii. If a solution $\hat{\mathbf{p}}$ is found with objective value $UBD^{\ell} < UBD$, set $UBD := UBD^{\ell}$ and $\mathbf{p}^{*} := \hat{\mathbf{p}}$.

Table 10.1: Tolerances for Algorithm 3 used in numerical examples.

| Task | Solver | abstol | reltol |
|---|---|---|---|
| State integration | IDAS | $1 \times 10^{-8}$ | $1 \times 10^{-7}$ |
| Upper bounding problem | SNOPT | $1 \times 10^{-5}$ | $1 \times 10^{-5}$ |
| State bounds integration | IDAS | $1 \times 10^{-6}$ | $1 \times 10^{-6}$ |
| State relaxations integration | CVODES | $1 \times 10^{-6}$ | $1 \times 10^{-6}$ |
| Lower bounding problem | SNOPT | $1 \times 10^{-5}$ | $1 \times 10^{-5}$ |
| B&B | Alg. 3 | $1 \times 10^{-3}$ | – |

10. Branching

   (a) Compute $j$, the smallest integer in $\arg \max_{i=1,\ldots,n_p} w(P_i^\ell)$.

   (b) Create intervals $P^{\ell'}$ and $P^{\ell''}$ by bisecting $P^\ell$ in the $j^{\text{th}}$ coordinate direction.

   (c) Set $N_{\text{GS}}^{\ell'} = N_{\text{GS}}^{\ell''} = \max(0, N_{\text{GS}}^{\ell} - 1)$.

   (d) Push the nodes $(P^{\ell'}, LBD^\ell, N_{\text{GS}}^{\ell'})$ and $(P^{\ell''}, LBD^\ell, N_{\text{GS}}^{\ell''})$ onto the stack $\Sigma$.

   (e) Go to 4.

## 10.4 Numerical Examples

All numerical experiments in this section were performed on a Dell Precision T3400 workstation with a 2.83 GHz Intel Core2 Quad CPU. One core and 512 MB of memory were dedicated to each job.

**Example 10.4.1** (Mathematical Example)**.** We first consider a mathematical example that is highly nonlinear and nonconvex:

$$\min_{\mathbf{p}\in P} \quad 10x(t_f, \mathbf{p}) - y(t_f, \mathbf{p}) + 0.5\sin(8p_2 - 0.5) \tag{10.16}$$

where $(x, y)$ is the unique solution of

$$\dot{x}(t_f, \mathbf{p}) = -(0.1y(t_f, \mathbf{p}) - 3p_1 e^{-5t})(x(t_f, \mathbf{p}) - 0.5p_2), \tag{10.17}$$

$$0 = y(t_f, \mathbf{p}) - \frac{2\sin(5p_1 + 1)}{\sqrt{y(t_f, \mathbf{p})}} - (15 + 2p_2)x(t_f, \mathbf{p}), \tag{10.18}$$

$$x_0(t_f, \mathbf{p}) = 1, \tag{10.19}$$

$$y(t_0, \hat{\mathbf{p}}) = 16.415, \quad \hat{\mathbf{p}} = (0, 0.5), \tag{10.20}$$

on $I \times P$.

Above, $I = [t_0, t_f] = [0, 1]$ and $P = [-1, 1] \times [0, p_2^U]$. We will consider both the case where $p_2^U = 1.0$ and $p_2^U = 1.1$. The objective function is plotted on the larger interval in Figure 10-1. The objective function is clearly nonconvex and has nine isolated local minima in both cases. The important difference between the problem with $p_2^U = 1.0$ and that with $p_2^U = 1.1$ is that when $p_2^U = 1.0$ the global minima is unconstrained, occurring at $\mathbf{p}^* = (0.1469, 0.7438)$ with an objective value of $\phi^* = -4.6674366$. In contrast, when $p_2^U = 1.1$, the global minimum is constrained, occurring at $\mathbf{p}^* = (0.14155058, 1.1)$ with an objective value of $\phi^* = -4.9324536$.

Optimization results are given in Table 10.3. In addition to $p_2^U$, several parameters and options in Algorithm 3 were varied to investigate their influence on the overall performance. Table 10.2 defines the shorthand used to display these results in Table 10.3. For all experiments $N_{\text{MS}} = 2$ and the generation skipping heuristic was not used (i.e., $N_{\text{GS}} = 0$). The correct optimal solution was located for every experiment in Table 10.3. In the best cases, the proposed algorithm solved the problem in 1.33 s and 115 nodes with $p_2^U = 1.0$, and in 0.67 s and 53 nodes with $p_2^U = 1.1$. A representative convex relaxation of the objective function is shown in Figure 10-2.

In Runs 1 and 3, the advantage of computing state relaxations is illustrated by comparing the proposed algorithm to a simpler version in which the lower bound is computed using only state bounds and interval arithmetic (Step 7b in Algorithm 3 is omitted). Though the cost per node increases by a factor of 4 when the full lower bounding procedure is used, this is dramatically outweighed by a reduction in the

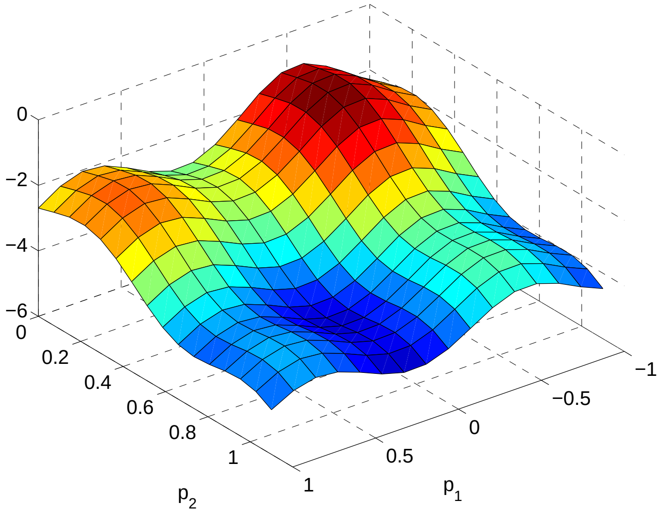

Figure 10-1: Objective function for Example 10.4.1 on $P = [-1, 1] \times [0, 1.1]$.

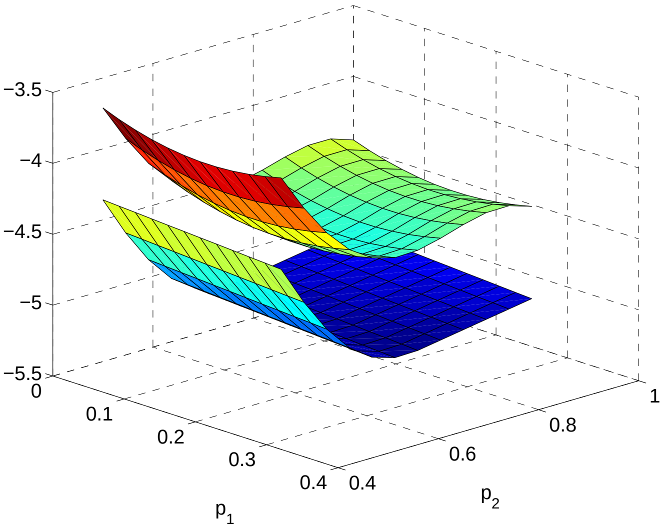

Figure 10-2: Convex relaxation of the objective function for Example 10.4.1 on $P^\ell = [0, 0.25] \times [0.5, 1]$. This interval contains the unconstrained global solution on $P = [-1, 1] \times [0, 1]$ and corresponds to the $18^{th}$ node processed in Run 3 of Table 10.3.

Table 10.2: Shorthand definitions used in Tables 10.3 and 10.4

| | |
|---|---|
| LBP | Lower bounding procedure. Either Step 7 is done as written (R), or Step 7b is skipped (I). |
| DR | Indicates whether domain reduction is used (Step 10.3.3 in Algorithm 3). |
| K | Number of refinement iterations in the right-hand sides of the auxiliary systems defining the state bounds ($K$ in (6.85) and (6.86)) and state relaxations ($K$ in (8.43)). The same number is used for both. |
| GS | Integer used for the generation skipping heuristic ($N_{\mathrm{GS}}$ in Algorithm 3). |
| CPU(s) | Total CPU time for Algorithm 3. |
| Nodes | Number of nodes processed by Algorithm 3. |
| s/N | Cost per node (CPU(s)/Nodes). |
| BFail | Number of nodes visited by Algorithm 3 for which the state bounding computation failed. |
| Rfail | Number of nodes visited by Algorithm 3 for which the state relaxation computation failed. |

Table 10.3: Optimization results for Example 10.4.1.

| Run | $p_2^U$ | DR | LBP | K | CPU(s) | Nodes | s/N | BFail | RFail |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.0 | N | I | 1 | 217.1 | 74,623 | 0.003 | 5 | – |
| 2 | 1.0 | N | I | 3 | 325.9 | 74,623 | 0.004 | 5 | – |
| 3 | 1.0 | N | R | 1 | 1.67 | 145 | 0.012 | 5 | 0 |
| 4 | 1.0 | N | R | 3 | 3.18 | 145 | 0.022 | 5 | 0 |
| 5 | 1.0 | Y | R | 1 | 1.33 | 115 | 0.012 | 5 | 0 |
| 6 | 1.0 | Y | R | 3 | 2.60 | 115 | 0.023 | 5 | 1 |
| 7 | 1.1 | N | I | 1 | 4.3 | 1,253 | 0.003 | 4 | – |
| 8 | 1.1 | N | I | 3 | 6.1 | 1,251 | 0.005 | 4 | – |
| 9 | 1.1 | N | R | 1 | 1.06 | 87 | 0.012 | 4 | 0 |
| 10 | 1.1 | N | R | 3 | 1.92 | 87 | 0.022 | 4 | 0 |
| 11 | 1.1 | Y | R | 1 | 0.67 | 53 | 0.013 | 4 | 0 |
| 12 | 1.1 | Y | R | 3 | 1.23 | 53 | 0.023 | 4 | 0 |

required number of nodes of 3 orders of magnitude, and a reduction in the CPU time of 2 orders of magnitude.

Comparing Runs 3 and 4, it is found that the number of refinement iterations $K$ has a rather significant effect on the cost per node. For this example, iterations beyond the first do not effect the node count and are not worth the additional effort. This observation is repeated throughout the table, and holds even in the case of an interval lower-bounding procedure (see Runs 1 and 2).

The effect of using domain reduction is seen by comparing Runs 3 and 5. For this example, the additional cost per node in insignificant. In general, it will be true for dynamic optimization problems that the cost of a simple domain reduction scheme like the one used here will be dominated by the cost of numerical integration in the lower-bounding problem. Using domain reduction reduces the number of nodes and the CPU time both by about 20%. The action of the domain reduction procedure can be seen more clearly in Figure 10-3, which shows all of the nodes fathomed in Run 5 as shaded subintervals of $P$. Without domain reduction, these intervals would form a partition of $P$ upon termination of Algorithm 3. Accordingly, white space in the figure corresponds to regions that were eliminated through domain reduction. The global minimum is indicated by the red diamond.

Runs 7-12 in Table 10.3 are exactly analogous to Runs 1-6, except that $p_2^U = 1.1$, and hence the global minimum is now constrained with $p_2^* = p_2^U$. In general, problems with unconstrained solutions are more difficult for spatial-B&B algorithms because the objective function is necessarily flat in the vicinity of such a solution. Because of this, nodes that contain points nearby an unconstrained solution, but do not contain the solution itself, cannot be fathomed by value dominance unless the lower-bound is very accurate. This causes the B&B procedure to generate a large number of nodes with diminishing interval width in the vicinity of the unconstrained solution, termed the cluster effect [50]. The severity of this problem is known to be related to the rate of convergence of the lower-bounding procedure.

For this example, the results in Table 10.3 indeed show that global optimization is significantly more efficient when the global minimum lies on an active bound con-

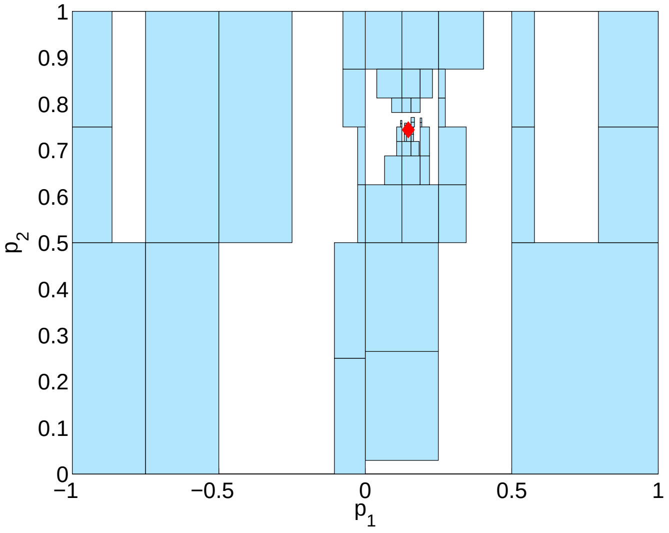

Figure 10-3: Intervals in the search space $P = [-1, 1] \times [0, 1]$ that were fathomed by value dominance (shaded boxes) in Example 10.4.1 (Run 5). White space indicates regions that were eliminated by domain reduction. The global minimum is marked by the red diamond.

straint. This is most dramatically true when the interval lower-bounding procedure is used. Comparing Runs 1 and 7, it is clear that the interval lower-bounding procedure suffers severe clustering in the case of an unconstrained solution. Comparatively, Runs 3 and 9 suggest that the effect of clustering is much less serious for the lower-bounding problem using state relaxations. This lends evidence to presumption that the state relaxations have a higher-order of convergence than do the state bounds.

Comparing Runs 9 and 11, it is seen that domain reduction reduces the number of nodes by nearly 40% in the case of a constrained solution, as compared to 20% in the unconstrained case. The action of the domain reduction procedure in the former case is illustrated in Figure 10-4.

**Example 10.4.2** (Kinetic Parameter Estimation with PSSA)**.** In this example, we consider a parameter estimation problem posed in [55]. The DAEs in this problem model a chemical reaction network converting methanol to various hydrocarbons, and the parameters $\mathbf{p} = (p_1, \dots, p_5)$ to be determined are related to reaction rate constants. The single algebraic equation in the model results from a pseudo-steady

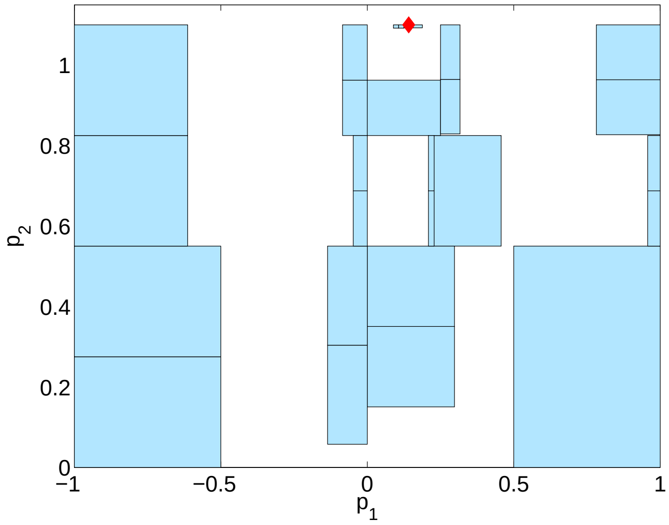

Figure 10-4: Intervals in the search space $P = [-1, 1] \times [0, 1.1]$ that were fathomed by value dominance (shaded boxes) in Example 10.4.1 (Run 11). White space indicates regions that were eliminated by domain reduction. The global minimum is marked by the red diamond.

state approximation applied to a reactive intermediate. For a detailed derivation of this model, see [55].

The problem is stated mathematically as

$$\min_{\mathbf{p}\in P} \quad \sum_{i=1}^{3}\sum_{k=1}^{16}(\hat{x}_i^k - x_i(t_k, \mathbf{p}))^2, \tag{10.21}$$

where, omitting arguments for clarity, $(\mathbf{x}, y)$ is the unique solution of

$$\dot{x}_1 = -x_1(2p_1 + p_3 + p_4) + x_2 y, \tag{10.22}$$

$$\dot{x}_2 = x_1(p_3 + p_2 y) - x_2 y, \tag{10.23}$$

$$\dot{x}_3 = x_1(p_4 + p_5 y) + x_2 y, \tag{10.24}$$

$$0 = x_1(p_1 - y(p_2 + p_5)) - x_2 y, \tag{10.25}$$

with $\mathbf{x}(t_0, \mathbf{p}) = (1, 0, 0)$, $\mathbf{y}(t_0, \hat{\mathbf{p}}) = 0.952$ and $\hat{\mathbf{p}} = (10, 10.5, 0, 0, 0)$.

Above, the constants $\hat{x}_i^k$ are experimental measurements of $x_i$ taken at 16 time points $t_k$ in the interval $I = [t_0, t_f] = [0, 1.2]$. These measurements are tabulated in [55]. In the original problem statement in [55], each $p_i$ is assumed to lie in the interval $[0, 20]$, so that $P = [0, 20]^5$, and the problem was solved with the constraint $0.1 \leq p_2 + p_5$ in order to avoid singularity of $\frac{\partial g}{\partial y}$.

As discussed in §10.1, the method used to solve this problem in [55] is not rigorous. Since the crucial step in the lower-bounding procedure relies on a finite sampling step, there is no possibility for significant conservatism in the lower bound. As a result, the method in [55] is much less computationally intensive than the method proposed here. In particular, we find that it is not possible to solve the full five dimensional problem with Algorithm 3 in reasonable time. On the other hand, the method in [55] does not provide a guarantee of global optimality, and therefore it is not meaningful to compare the performance of Algorithm 3 to the method in [55].

Here, we solve two simplified instances of the problem with Algorithm 3. In the first, we consider the two parameter problem given by setting $P = [0, 20] \times [1, 20] \times [0, 0] \times [0, 0] \times [0, 0]$ (from the results in [55], it is known that $p_3 = p_4 = p_5 = 0$ at the global solution). In the second case, we consider the three parameter problem given by setting $P = [0, 20] \times [1, 20] \times [0, 20] \times [0, 0] \times [0, 0]$. We do not use the constraint $0.1 \leq p_2 + p_5$ in either case. Rather, $p_2^L$ is set to 1 instead of 0 to avoid the singularity in the model. This is because only the interval $P$, and not the constraint $0.1 \leq p_2 + p_5$, is used during the computation of state bounds. Since this computation will fail if the index-one assumption fails in $P$, the interval $P$ must be restricted.

The reader may have noted that the algebraic equation in the embedded DAEs can be explicitly rearranged for $y$ provided that $x_1(p_2 + p_5) + x_2$ is nonzero. Carrying out this rearrangement and substituting throughout the system clearly results in an explicit system of ODEs. We only solve the problem as a DAE here because it has been posed as a benchmark problem in this form in the literature [55]. In fact, between the two articles which have previously presented methods for solving global optimization problems with DAEs embedded [55, 42], this is the only problem in which the embedded system was not written as an explicit system of ODEs.

Optimization results are shown in Table 10.4 (see definitions in Table 10.2). For

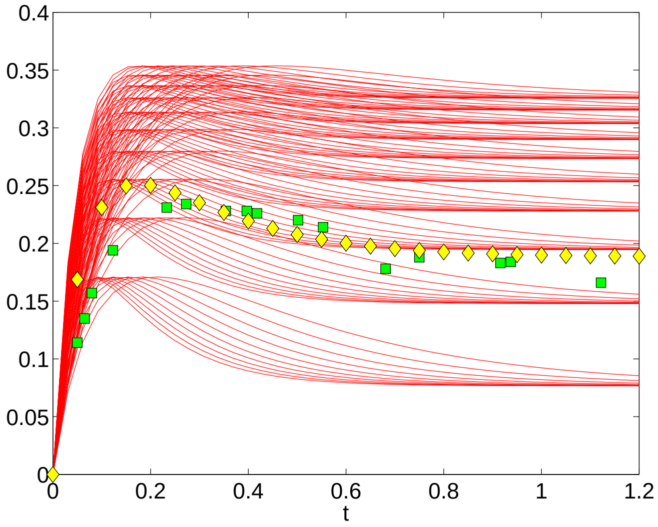

Figure 10-5: Experimental data (green squares) and the optimal state trajectory (yellow triangles) for $x_2$ in Example 10.4.2, superimposed on 100 trajectories for parameters on a uniform grid over $P$ (red solid lines).

this example, the full lower-bounding procedure was used in every experiment, and the global solution was correctly located in every case as $\mathbf{p}^* = (5.2407, 1.2176, 0, 0, 0)$ with objective value $\phi^* = 0.1069$. For the two parameter problem, the best performance in terms of CPU time required 425s. For the three parameter problem, the fastest solution time required 8,431s (2h 20m 31s). In both cases, this was achieved with a generation skipping heuristic, so the number of nodes is inflated. The optimal trajectory and measured data for $x_2$ are shown in Figure 10-5, overlaid on a large sample of feasible trajectories.

It is immediately evident from Table 10.4 that this problem is much more difficult than Example 10.4.1, even in the two parameter case. The convergence behavior of the upper and lower bounds for Run 2 is shown in Figure 10-6. From this plot, one can attribute the large number of nodes required to solve this problem to two sources. Firstly, a finite lower bound was not achieved until after 6,109 nodes were processed. This is due to repeated failures in the state bounding procedure, of which there were 3,053. Secondly, the rate at which the lower bounds converge toward the upper bounds in Figure 10-6 decreases abruptly after roughly 6262 nodes have been

Table 10.4: Optimization results for Example 10.4.2.

| Run | Decisions | DR | K | GS | CPU(s) | Nodes | s/N | BFail | RFail |
|---|---|---|---|---|---|---|---|---|---|
| 1 | $(p_1, p_2)$ | N | 2 | 0 | 593 | 7,611 | 0.078 | 3,053 | 220 |
| 2 | $(p_1, p_2)$ | Y | 2 | 0 | 593 | 7,599 | 0.078 | 3,053 | 221 |
| 3 | $(p_1, p_2)$ | Y | 6 | 0 | 1,200 | 6,725 | 0.178 | 2,482 | 339 |
| 4 | $(p_1, p_2)$ | Y | 2 | 1 | 425 | 9,067 | 0.047 | 1,280 | 162 |
| 5 | $(p_1, p_2)$ | Y | 2 | 2 | 571 | 16,221 | 0.035 | 1,094 | 133 |
| 6 | $(p_1, p_2)$ | Y | 2 | 3 | 604 | 20,037 | 0.030 | 627 | 92 |
| 7 | $(p_1, p_2, p_3)$ | Y | 2 | 0 | 12,404 | 154,911 | 0.080 | 70,677 | 9,342 |
| 8 | $(p_1, p_2, p_3)$ | Y | 2 | 1 | 8,431 | 211,689 | 0.04 | 33,240 | 7,978 |

processed. This is due to the cluster effect, as demonstrated below.

To better understand the performance of Algorithm 3 in Run 2, it is helpful to see the subintervals of $P$ generated by branching during the course of the algorithm. These are shown in Figure 10-7 and, zoomed in near the global solution, in Figure 10-8. In both figures, the shaded boxes are nodes fathomed by value dominance, while white space represents regions that were eliminated by domain reduction. Firstly, these figures show that the effect of domain reduction is minimal for this problem, which is corroborated by comparing Runs 1 and 2 in Table 10.4. These figures also clearly demonstrate the cluster effect, which is to be expected because no constraints are active at the global solution. From Figure 10-7, however, one also notes a very high density of small intervals all along the $p_2^L$ boundary, even quite far from the solution. The explanation for this is given in Figure 10-9, which shows nodes for which the state bounding procedure (white boxes) or the state relaxation procedure (shaded boxes with dashed outline) failed. From this figure, it is clear that the accumulation of small intervals along the bottom of Figure 10-7 is not due to the cluster effect, but rather results from repeated failure of the state bounding procedure in this region. The state bounding procedure evidently has a very difficult time verifying existence and uniqueness of the DAE solution in this region, though the ultimate reason for this is not understood. Failure of the state relaxation procedure is much less frequent and is believed to result from chattering in the event detection scheme (see §10.3.2).

To elucidate the source of the abrupt change in the slope of the lower bound curve

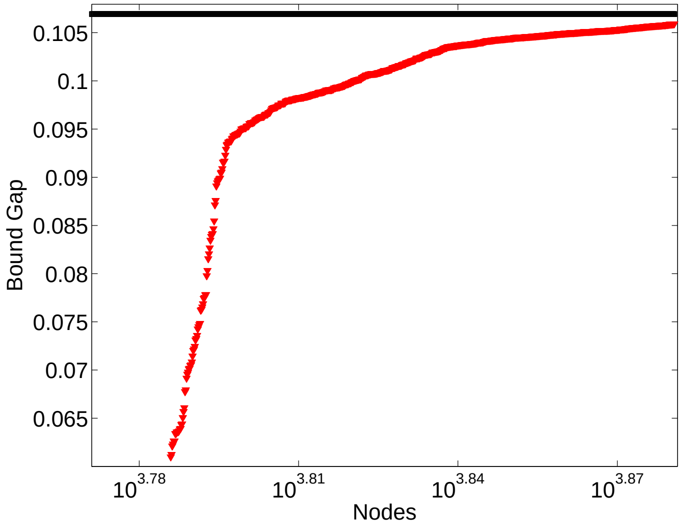

Figure 10-6: Convergence of the upper (black squares) and lower (red triangles) bounds on the globally optimal objective value for Example 10.4.2 (Run 2) as a function of the number of nodes processed.

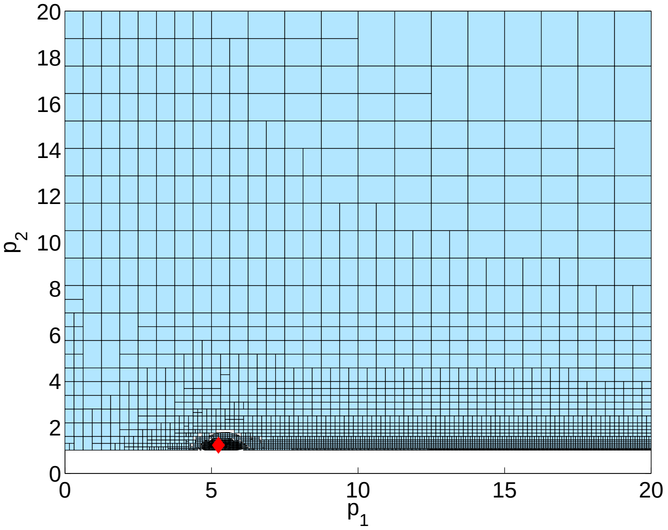

Figure 10-7: Intervals in the search space $P$ that were fathomed by value dominance (shaded boxes) in Example 10.4.2 (Run 2). White space indicates regions that were eliminated by domain reduction. The global minimum is marked by the red diamond.

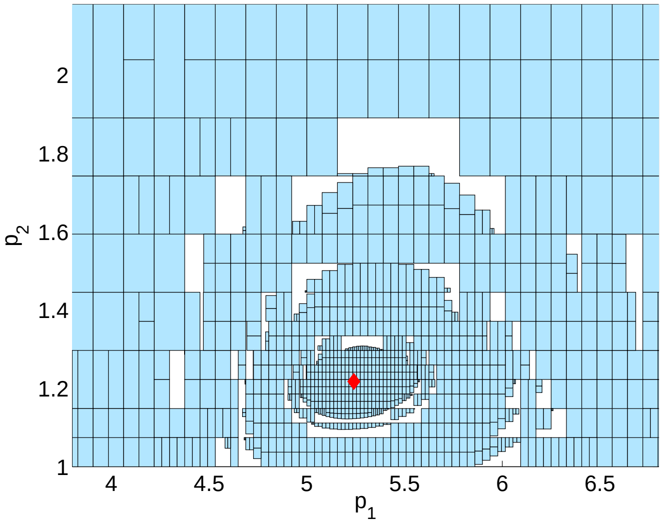

Figure 10-8: A closer look at intervals in the search space $P$ that were fathomed by value dominance (shaded boxes) in the vicinity of the global minimum (red diamond) in Example 10.4.2 (Run 2). White space indicates regions that were eliminated by domain reduction.

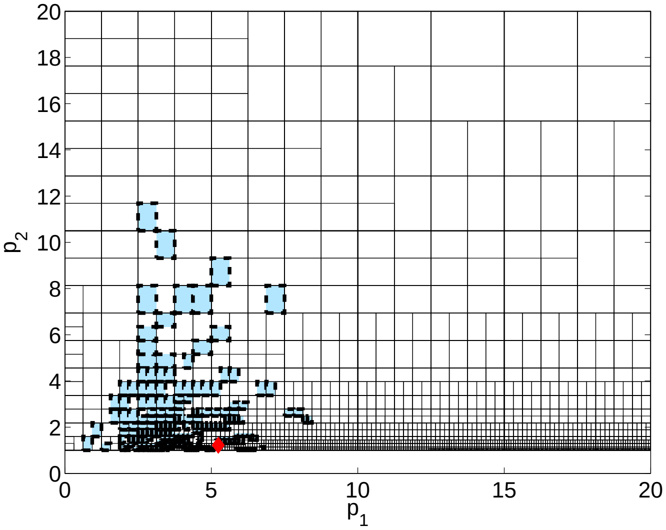

Figure 10-9: Intervals in $P$ visited by Algorithm 3 in Example 10.4.2 (Run 2) where either the computation of state bounds (white boxes) or state relaxations (shaded boxes with dashed outline) failed. Intervals are plotted in the order they were visited, so that smaller intervals cover larger intervals where failures may also have occurred. The global minimum is marked by the red diamond.

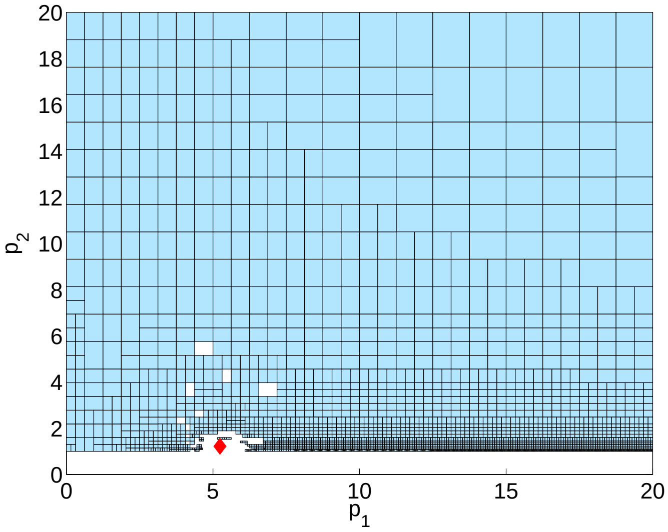

Figure 10-10: Intervals from the first 6,275 nodes visited by Algorithm 3 that were fathomed by value dominance (shaded boxes) in Example 10.4.2 (Run 2). White space corresponds to nodes remaining on the stack. The qualitative slope change in the lower bounds in Figure 10-6 occurs at this stage in the algorithm. The global minimum is marked by the red diamond.

in Figure 10-6, the nodes fathomed by Algorithm 3 prior to the change, which occurs near node $6,275$, are plotted in Figures 10-10 and 10-11. From these figures, it is clear that the reduced rate of convergence after 6,275 nodes is directly correlated with the onset of clustering around the global solution.

Considering Figure 10-6 and the BFail column in Table 10.4, it is evident that the repeated failure of the state bounding procedure is more problematic than the cluster effect for this problem. From Figure 10-6, one can see that the onset of clustering occurs only after more than 80% of the total nodes have been processed. In contrast, the state bounding procedure failed on 40% of the total nodes. Figure 10-12 shows that the bahavior of the problem with three decision variables (Run 7) is analogous, with only 6% of the total node count related to clustering, and failure of the state bounding procedure for some 46% of the total nodes.

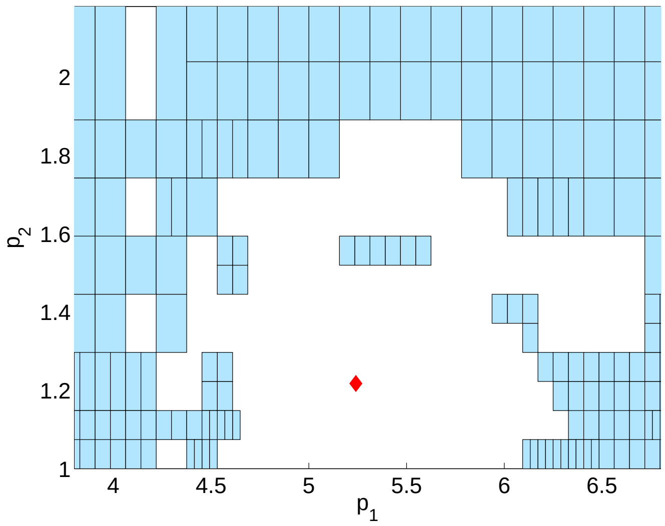

Figure 10-11: A closer look at intervals from the first 6,275 nodes visited by Algorithm 3 that were fathomed by value dominance (shaded boxes) in the vicinity of the global minimum (red diamond) in Example 10.4.2 (Run 2). White space corresponds to nodes remaining on the stack. The qualitative slope change in the lower bounds in Figure 10-6 occurs at this stage in the algorithm.

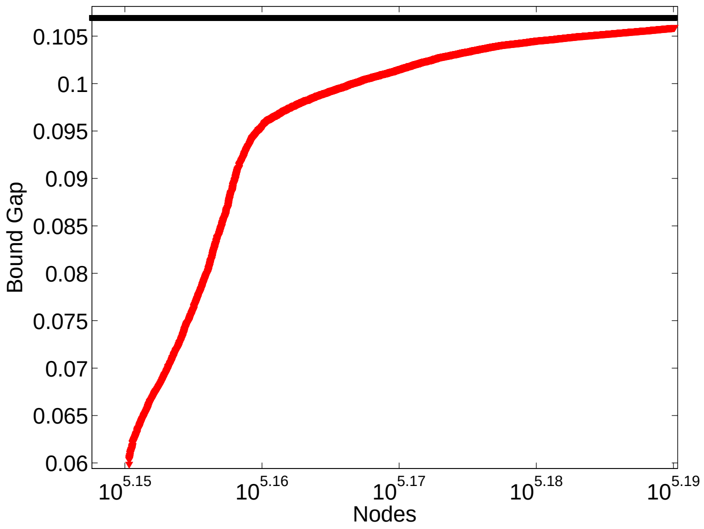

Figure 10-12: Convergence of the upper (black squares) and lower (red triangles) bounds on the globally optimal objective value for Example 10.4.2 (Run 7) as a function of the number of nodes processed.

## 10.5 Conclusion

In this chapter, a deterministic global optimization algorithm for problems with semi-explicit index-one DAEs embedded was developed. The algorithm is based on a spatial-B&B framework and uses the sequential approach to dynamic optimization. The lower-bounding procedure is enabled by the state bounding and relaxation techniques developed throughout this thesis.

The performance of the algorithm was demonstrated on two example problems. The first test problem was highly nonlinear and displayed multiple suboptimal local minima. Nonetheless, the proposed algorithm was shown to locate the global solution with only minor computational effort. By introducing a small perturbation in in the host interval, it was shown that the performance of the algorithm is significantly improved when the optimal solution lies on a constraint, all other things being roughly equal. This is due to the cluster effect and is typical of global optimization algorithms. However, the state relaxation method using relaxation preserving dynamics developed in Chapter 8 was shown to be effective for reducing the cluster effect, at least when compared to an interval lower-bounding procedure.

The second test problem was much more challenging. Again, the proposed algorithm was able to provide a guaranteed global solution, but with significant computational expense. The cluster effect was again a source of inefficiency for this problem, but was overshadowed by difficulties in the state bounding procedure. As discussed in detail in Chapter 6, failure of the single-phase state bounding method is caused by an inability to verify existence and uniqueness of a solution of the embedded DAEs on the given interval. From experiments with the single-phase method, it is known that the DAEs of Example 10.4.2 are particularly difficult from a state bounding prospective, though it is not clear why. Thus, more experiments are needed to determine whether the state bounding procedure will typically be the weak link in Algorithm 3, or if Example 10.4.2 is exceptional in this regard. In any case, a more robust bounding method should be considered a key target for future research.

In the proposed algorithm, we have implemented only one very simple form of domain reduction. However, it is known that efficient global optimization algorithms rely very heavily on a variety of domain reduction techniques. Therefore, a primary goal for future work is to implement more sophisticated domain reduction techniques, potentially specialized to dynamic problems. Moreover, advanced techniques for combating the cluster effect, such as convexity detection and the computation of exclusion regions, should also be pursued.

# Chapter 11

# Convex Relaxations for Nonconvex Optimal Control Problems

## 11.1 Introduction

Consider the open-loop optimal control problem informally stated as

$$\begin{aligned} &\inf_{\mathbf{u}\in\mathcal{U}} \phi(\mathbf{u}(t_f), \mathbf{x}(t_f, \mathbf{u})) \\ &\text{s.t.} \quad \mathbf{g}(\mathbf{u}(t_f), \mathbf{x}(t_f, \mathbf{u})) \leq \mathbf{0} \\ &\qquad\ \mathbf{q}(t, \mathbf{u}(t), \mathbf{x}(t, \mathbf{u})) \leq \mathbf{0}, \quad \text{a.e.} \quad t \in [t_0, t_f], \end{aligned} \tag{11.1}$$

where $\mathcal{U}$ is a subset of $(L^1([t_0, t_f]))^{n_u}$ and, for each $\mathbf{u} \in \mathcal{U}$, $\mathbf{x}(\cdot, \mathbf{u})$ is an absolutely continuous solution of

$$\begin{aligned} \dot{\mathbf{x}}(t, \mathbf{u}) &= \mathbf{f}(t, \mathbf{u}(t), \mathbf{x}(t, \mathbf{u})), \quad \text{a.e.} \quad t \in [t_0, t_f], \\ \mathbf{x}(t_0, \mathbf{u}) &= \mathbf{x}_0, \end{aligned} \tag{11.2}$$

which is assumed unique. This problem is of general interest and has been the subject of intense research for decades [22]. Nonetheless, (11.1) is an infinite dimensional problem and, as such, there is no general purpose algorithm for solving it to guaranteed global optimality. If the control functions **u** are approximated by a finite number of real parameters [173], then the resulting approximation of (11.1) can be solved to global optimality using any of the methods described in the articles [135, 164, 104]. However, this approach may be unsatisfactory for both theoretical and practical reasons. Theoretically, the solution so obtained is only globally optimal for an approximate problem. Practically, it often happens that many parameters are required to accurately approximate a single control function, making the approximate NLP large.

Based on these shortcomings, it is desirable to develop a method for solving (11.1) to guaranteed global optimality directly in the infinite-dimensional setting. In this chapter, we provide a crucial step towards accomplishing this through a branch-and-bound (B&B) approach. In particular, we present a method for computing a guaranteed lower bound on the optimal objective value of (11.1). For finite-dimensional optimization problems, providing a valid lower-bounding procedure is the most difficult aspect of applying the B&B framework, and is typically the key development required to extend B&B techniques to a new class of problems. However, infinite-dimensional problems introduce new complications, and therefore we cannot present a complete B&B global optimization algorithm for (11.1) at this time. Specifically, we have so far found no way to partition an infinite-dimensional set in a way that is both exhaustive and useful for refining the lower bound computed through the procedure given here.

To compute a lower bound on the optimal objective value of (11.1), we construct an auxiliary optimal control problem, called a relaxation of (11.1), with the properties (a) the optimal objective value is guaranteed to underestimate the infimum in (11.1), and (b) it is convex in the sense that the feasible set is a convex subset of $\mathcal{U}$ and the mapping taking $\mathbf{u}$ to the objective value is convex on this set. Because it is convex, this relaxed problem is in principle solvable to global optimality. For example, necessary and sufficient optimality conditions for such programs are derived in [15], and gradient based solution methods are proposed. Supposing that such a solution can be obtained, this procedure generates a guaranteed lower bound on the solution of (11.1). In [15], conditions were also studied under which (11.1) can be guaranteed to be convex, based on arguments similar to those presented in §11.5. In contrast, the method presented here is not used to verify convexity, but rather to construct a convex optimization problem which underestimates a given instance of (11.1), even when (11.1) is nonconvex.

By analogy to the global dynamic optimization methods presented in [135, 164, 104] (and the method for semi-explicit index-one DAEs presented in Chapter 10), the proposed lower-bounding procedure depends on the ability to compute state bounds and a form of state relaxations for the embedded control system (11.2). It has already been shown in Chapter 3 that state bounds for (11.2) can be computed without the need for control parameterization. The main result of this chapter is that the same is essentially true for state relaxations. By a suitable reinterpretation of McCormick's relaxation technique, it is shown that convex and concave relaxations of the solutions of (11.2) on a convex subset of $L^1([t_0, t_f])$ can be derived and evaluated computationally by exactly the same techniques already developed in Chapter 7.

### 11.1.1 Basic Approach

Let $I = [t_0, t_f] \subset \mathbb{R}$, let $\bar{U} \subset \mathbb{R}^{n_u}$ be compact, and let $U : I \to \mathbb{IR}^{n_u}$ be a continuous mapping such that $U(t) = [\mathbf{u}^L(t), \mathbf{u}^U(t)] \subset \bar{U}$, $\forall t \in I$. In the remainder of this chapter, the set of admissible controls is defined by

$$\mathcal{U} \equiv \{\mathbf{u} \in (L^1(I))^{n_u} : \mathbf{u}(t) \in U(t) \text{ a.e. in } I\} \tag{11.3}$$

and is assumed nonempty. It is trivial to verify that $\mathcal{U}$ is a convex subset of the vector space $(L^1(I))^{n_u}$. Finally, let $D \subset \mathbb{R}^{n_x}$ be open and suppose that the mappings in (11.1) have the form $\phi : \bar{U} \times D \to \mathbb{R}$, $\mathbf{g} : \bar{U} \times D \to \mathbb{R}^{n_g}$, and $\mathbf{q} : I \times \bar{U} \times D \to \mathbb{R}^{n_q}$. Assumptions regarding the control systems (11.2) are discussed in §11.5. We note here that the solution has the form $\mathbf{x} : I \times \mathcal{U} \to D$.

In order to construct a convex underestimating program for (11.1), convex underestimating functions are derived for the mappings

$$\mathcal{U} \ni \mathbf{u} \longmapsto \mathcal{F}_{\phi}(\mathbf{u}) \equiv \phi(\mathbf{u}(t_f), \mathbf{x}(t_f, \mathbf{u})), \tag{11.4}$$

$$\mathcal{U} \ni \mathbf{u} \longmapsto \mathcal{F}_{\mathbf{g}}(\mathbf{u}) \equiv \mathbf{g}(\mathbf{u}(t_f), \mathbf{x}(t_f, \mathbf{u})), \tag{11.5}$$

and the family of mappings

$$\mathcal{U} \ni \mathbf{u} \longmapsto \mathcal{F}_{\mathbf{q},t}(\mathbf{u}) \equiv \mathbf{q}(t, \mathbf{u}(t), \mathbf{x}(t, \mathbf{u})), \tag{11.6}$$

for a.e. $t \in I$. Defining the relaxed program with these convex underestimators in place of the mappings above, both convexity and the desired underestimation property follow from standard arguments [84, 15].

## 11.2 McCormick Relaxations on Vector Spaces

Convex relaxations for the mappings (11.4), (11.5) and (11.6) will be derived using McCormick's relaxation technique (see Chapter 2). The novelty in the present application is that these functions are not defined on $\mathbb{R}^n$, but rather on $\mathcal{U}$, which is a subset of the function space $L^1([t_0, t_f])$. To treat this case, it is shown here that the basic properties of McCormick relaxations are preserved when one considers an arbitrary vector space in place of $\mathbb{R}^n$. Considering the form of the mappings (11.4), (11.5) and (11.6), we are particularly interested in extending the composite relaxation technique of §2.7.2 to the case where the inner function is defined on an arbitrary vector space.

Let $V$ be a vector space. Clearly, convex combinations of the elements of $V$ are well-defined. Then, convexity of a subset $C \subset V$ is defined in the standard way. Moreover, convexity and concavity of functions mapping $C$ into $\mathbb{R}$ are defined by the standard inequalities. Relaxations in this context are defined as follows.

**Definition 11.2.1.** Let $V$ be a vector space, let $C \subset V$ be convex, and let $h, h^{cv}, h^{cc} : C \to \mathbb{R}$. The function $h^{cv}$ is called a *convex relaxation* of $h$ on $C$ if $h^{cv}$ is convex on $C$ and $h^{cv}(\mathbf{v}) \leq h(\mathbf{v})$, $\forall \mathbf{v} \in C$. Similarly, $h^{cc}$ is called a *concave relaxation* of $h$

on $C$ if $h^{cc}$ is concave on $C$ and $h^{cc}(\mathbf{v}) \geq h(\mathbf{v})$, $\forall \mathbf{v} \in C$. The terms convex and concave relaxation will also be used for vector functions when these conditions hold componentwise.

The definition of a composite relaxation from §2.7.2 can now be extended to functions on $V$.

**Definition 11.2.2.** Let $Q \subset \mathbb{R}^{n_y}$ and $\mathbf{w} : Q \to \mathbb{R}^m$. For any $Y \subset Q$, functions $\mathbf{u_w}, \mathbf{o_w} : \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} \to \mathbb{R}$ are called *convex and concave composite relaxations of* $\mathbf{w}$ on $Y$ if the following condition holds: For any vector space $V$, and convex $C \subset V$, and any $\mathbf{y}, \mathbf{y}^{cv}, \mathbf{y}^{cc} : C \to \mathbb{R}^{n_y}$ with $\mathbf{y}(C) \subset Y$, convex and concave relaxations of the composite function

$$C \ni \mathbf{v} \longmapsto \mathbf{h}(\mathbf{v}) \equiv \mathbf{w}(\mathbf{y}(\mathbf{v})) \tag{11.7}$$

on $C$ are given by the composite mappings

$$\begin{aligned} C \ni \mathbf{v} &\longmapsto \mathbf{h}^{cv}(\mathbf{v}) \equiv \mathbf{u_w}(\mathbf{y}^{cv}(\mathbf{v}), \mathbf{y}^{cc}(\mathbf{v})) \\ C \ni \mathbf{v} &\longmapsto \mathbf{h}^{cc}(\mathbf{v}) \equiv \mathbf{o_w}(\mathbf{y}^{cv}(\mathbf{v}), \mathbf{y}^{cc}(\mathbf{v})) \end{aligned} \tag{11.8}$$

provided that $\mathbf{y}^{cv}$ and $\mathbf{y}^{cc}$ are, respectively, convex and concave relaxations of $\mathbf{y}$ on $C$.

When $\mathbf{y}$ is $\mathcal{L}$-factorable and $Y$ is an interval, composite relaxations can be readily obtained from a natural McCormick extension as follows.

**Theorem 11.2.3.** *Let $Q \subset \mathbb{R}^{n_y}$ and let $\mathbf{w} : Q \to \mathbb{R}^m$ be $\mathcal{L}$-factorable with natural McCormick extension $\{\mathbf{w}\} : \mathcal{Q} \to \mathbb{MR}^m$. For any $Y \in \mathbb{I}Q$ such that $Y$ is represented in $\mathcal{Q}$, the functions $\mathbf{u_w}, \mathbf{o_w} : \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} \to \mathbb{R}^m$ defined by*

$$\begin{aligned} \mathbf{u_w}(\mathbf{z}^{cv}, \mathbf{z}^{cc}) &\equiv \{\mathbf{w}\}^{cv}(\mathrm{MC}(\mathbf{y}^L, \mathbf{y}^U, \mathbf{z}^{cv}, \mathbf{z}^{cc})), \\ \mathbf{o_w}(\mathbf{z}^{cv}, \mathbf{z}^{cc}) &\equiv \{\mathbf{w}\}^{cc}(\mathrm{MC}(\mathbf{y}^L, \mathbf{y}^U, \mathbf{z}^{cv}, \mathbf{z}^{cc})), \end{aligned} \tag{11.9}$$

*are composite relaxations of* $\mathbf{w}$ *on* $Y$*.*

*Proof.* Choose any vector space $V$, any convex $C \subset V$, and any $\mathbf{y}, \mathbf{y}^{cv}, \mathbf{y}^{cc} : C \to \mathbb{R}^{n_y}$ such that $\mathbf{y}(C) \subset Y$ and $\mathbf{y}^{cv}$ and $\mathbf{y}^{cc}$ are, respectively, convex and concave relaxations of $\mathbf{y}$ on $C$. For any $\mathbf{v} \in C$, these hypothesis ensure that $\mathbf{y}^{cv}(\mathbf{v}) \leq \mathbf{y}^{cc}(\mathbf{v})$ and $\mathbf{y}(\mathbf{v}) \in Y \cap [\mathbf{y}^{cv}(\mathbf{v}), \mathbf{y}^{cc}(\mathbf{v})]$. Then, applying Lemma 2.7.3 with the definitions $X := Y$, $\mathbf{x} := \mathbf{y}(\mathbf{x}^{cv})$, $\mathbf{x}^{cv} := \mathbf{y}^{cv}(\mathbf{v})$ and $\mathbf{x}^{cc} := \mathbf{y}^{cc}(\mathbf{v})$ gives the inequality

$$\mathbf{u}_{\mathbf{w}}(\mathbf{y}^{cv}(\mathbf{v}), \mathbf{y}^{cc}(\mathbf{v})) \leq \mathbf{w}(\mathbf{y}(\mathbf{v})) \leq \mathbf{o}_{\mathbf{w}}(\mathbf{y}^{cv}(\mathbf{v}), \mathbf{y}^{cc}(\mathbf{v})). \tag{11.10}$$

Then, using the definitions (11.7) and (11.8), the functions $\mathbf{h}^{cv}$ and $\mathbf{h}^{cc}$ underestimate and overestimate $\mathbf{h}$ on $C$, respectively.

Next, choose any $(\lambda, \mathbf{v}_1, \mathbf{v}_2) \in [0,1] \times C \times C$ and let $\mathbf{v}_3 = \lambda\mathbf{v}_1 + (1-\lambda)\mathbf{v}_2$. By hypothesis,

$$\begin{aligned}
\mathbf{y}^{cv}(\mathbf{v}_3) &\leq \lambda\mathbf{y}^{cv}(\mathbf{v}_1) + (1-\lambda)\mathbf{y}^{cv}(\mathbf{v}_2),\\
\mathbf{y}^{cc}(\mathbf{v}_3) &\geq \lambda\mathbf{y}^{cc}(\mathbf{v}_1) + (1-\lambda)\mathbf{y}^{cc}(\mathbf{v}_2),
\end{aligned}$$

and $Y \cap [\mathbf{y}^{cv}(\mathbf{v}_i), \mathbf{y}^{cc}(\mathbf{v}_i)] \neq \emptyset$, for all $i \in \{1,2,3\}$. Then, applying Lemma 2.7.4 with the definitions $X := Y$, $\mathbf{x}_i^{cv} := \mathbf{y}^{cv}(\mathbf{v}_i)$ and $\mathbf{x}_i^{cc} := \mathbf{y}^{cc}(\mathbf{v}_i)$, for $i \in \{1,2,3\}$, gives the inequalities

$$\begin{aligned}
\mathbf{u}_{\mathbf{w}}(\mathbf{y}^{cv}(\mathbf{v}_3), \mathbf{y}^{cc}(\mathbf{v}_3)) &\leq \lambda\mathbf{u}_{\mathbf{w}}(\mathbf{y}^{cv}(\mathbf{v}_1), \mathbf{y}^{cc}(\mathbf{v}_1)) + (1-\lambda)\mathbf{u}_{\mathbf{w}}(\mathbf{y}^{cv}(\mathbf{v}_2), \mathbf{y}^{cc}(\mathbf{v}_2)),\\
\mathbf{o}_{\mathbf{w}}(\mathbf{y}^{cv}(\mathbf{v}_3), \mathbf{y}^{cc}(\mathbf{v}_3)) &\geq \lambda\mathbf{o}_{\mathbf{w}}(\mathbf{y}^{cv}(\mathbf{v}_1), \mathbf{y}^{cc}(\mathbf{v}_1)) + (1-\lambda)\mathbf{o}_{\mathbf{w}}(\mathbf{y}^{cv}(\mathbf{v}_2), \mathbf{y}^{cc}(\mathbf{v}_2)).
\end{aligned}$$

Therefore, the functions $\mathbf{h}^{cv}$ and $\mathbf{h}^{cc}$ defined by (11.8) are convex and concave on $C$, respectively. □

**Remark 11.2.4.** Since the previous theorem holds for any $\mathcal{L}$-factorable outer function, it holds for the simple cases where the outer function is any of $(+, \mathbb{R}^2, \mathbb{R})$, $(\times, \mathbb{R}^2, \mathbb{R})$ or $(u, B, \mathbb{R}) \in \mathcal{L}$. Thus, the previous result establishes that McCormick's relaxation rules for these basic operations are applicable in an arbitrary vector space without modification.

## 11.3 Relaxing the Objective and Constraints

In this section, the results of §11.2 are applied to compute relaxations of the functions (11.4), (11.5) and (11.6), under the assumption that convex and concave relaxations of $\mathbf{x}(t,\cdot)$ on $\mathcal{U}$ are available for every $t \in I$ (see §11.5). The following assumptions are required.

**Assumption 11.3.1.** Functions $\mathbf{x}^L, \mathbf{x}^U : I \to \mathbb{R}^{n_x}$ are available such that $\mathbf{x}(t,\mathbf{u}) \in X(t) \equiv [\mathbf{x}^L(t), \mathbf{x}^U(t)]$, $\forall (t,\mathbf{u}) \in I \times \mathcal{U}$, and $X(t) \subset D$, $\forall t \in I$.

**Assumption 11.3.2.** The functions $\phi$, $\mathbf{g}$, $\mathbf{q}$ and $\mathbf{f}$ are $\mathcal{L}$-factorable with natural McCormick extensions $\{\phi\} : \mathcal{E} \to \mathbb{MR}$, $\{\mathbf{g}\} : \mathcal{E} \to \mathbb{MR}^{n_g}$, $\{\mathbf{q}\} : \mathcal{D} \to \mathbb{MR}^{n_q}$ and $\{\mathbf{f}\} : \mathcal{D} \to \mathbb{MR}^{n_f}$. Moreover, the interval $U(t_f) \times X(t_f)$ is represented in $\mathcal{E}$ and, for every $t \in I$, the interval $I \times U(t) \times X(t)$ is represented in $\mathcal{D}$.

Of course, the state bounds of Assumption 11.3.1 can be computed using any of the methods presented in Chapters 3 and 4. Using Assumption 11.3.2, we now derive a convex relaxation for the mapping $\mathcal{F}_{\mathbf{q},t}$. Define the function $\mathbf{u}_{\mathbf{q}} : I \times \bar{U} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ by

$$\begin{aligned}\mathbf{u}_{\mathbf{q}}(t,\mathbf{p},\mathbf{z}^{cv},\mathbf{z}^{cc}) \equiv \{\mathbf{q}\}^{cv}(&\mathrm{MC}(t,t,t,t), \mathrm{MC}(\mathbf{u}^L(t),\mathbf{u}^U(t),\mathbf{p},\mathbf{p}),\\ &\mathrm{MC}(\mathbf{x}^L(t),\mathbf{x}^U(t),\mathbf{z}^{cv},\mathbf{z}^{cc})).\end{aligned} \tag{11.11}$$

From Theorem 11.2.3, we have the following lemma.

**Lemma 11.3.3.** *For any $\boldsymbol{\psi}^c, \boldsymbol{\psi}^C : I \times \mathcal{U} \to \mathbb{R}^{n_x}$ and any $t \in I$, the function*

$$U(t) \times \mathcal{U} \ni (\mathbf{p},\mathbf{u}) \longmapsto \mathbf{u}_{\mathbf{q}}(t,\mathbf{p},\boldsymbol{\psi}^{cv}(t,\mathbf{u}),\boldsymbol{\psi}^{cc}(t,\mathbf{u}))$$

*is a convex relaxation of*

$$U(t) \times \mathcal{U} \ni (\mathbf{p},\mathbf{u}) \longmapsto \mathbf{q}(t,\mathbf{p},\mathbf{x}(t,\mathbf{u}))$$

*on $U(t) \times \mathcal{U}$, provided that $\boldsymbol{\psi}^{cv}(t,\cdot)$ and $\boldsymbol{\psi}^{cc}(t,\cdot)$ are, respectively, convex and concave relaxations of $\mathbf{x}(t,\cdot)$ on $\mathcal{U}$.*

*Proof.* The set $\mathbb{R}^{n_u} \times L^1(I)$ is a vector space, and for any fixed $t \in I$, the set $U(t) \times \mathcal{U}$ is convex. Then, the result follows from Theorem 11.2.3 using the definitions $Q := I \times \bar{U} \times D$, $\mathbf{h} := \mathbf{q}$, $V := \mathbb{R}^{n_u} \times L^1(I)$, $C := U(t) \times \mathcal{U}$, $Y := [t,t] \times U(t) \times X(t)$, and

$$\begin{aligned}
U(t) \times \mathcal{U} \ni (\mathbf{p},\mathbf{u}) &\longmapsto \mathbf{y}(\mathbf{p},\mathbf{u}) \equiv (t,\mathbf{p},\mathbf{x}(t,\mathbf{u})),\\
U(t) \times \mathcal{U} \ni (\mathbf{p},\mathbf{u}) &\longmapsto \mathbf{y}^{cv}(\mathbf{p},\mathbf{u}) \equiv (t,\mathbf{p},\boldsymbol{\psi}^{cv}(t,\mathbf{u})),\\
U(t) \times \mathcal{U} \ni (\mathbf{p},\mathbf{u}) &\longmapsto \mathbf{y}^{cc}(\mathbf{p},\mathbf{u}) \equiv (t,\mathbf{p},\boldsymbol{\psi}^{cc}(t,\mathbf{u})).
\end{aligned}$$

□

**Theorem 11.3.4.** *For any $\boldsymbol{\psi}^c, \boldsymbol{\psi}^C : I \times \mathcal{U} \to \mathbb{R}^{n_x}$ and a.e. $t \in I$, the mapping*

$$\mathcal{U} \ni \mathbf{u} \longmapsto \mathcal{F}^{cv}_{\mathbf{q},t}(\mathbf{u}) \equiv \mathbf{u}_{\mathbf{q}}(t,\mathbf{u}(t),\boldsymbol{\psi}^{cv}(t,\mathbf{u}),\boldsymbol{\psi}^{cc}(t,\mathbf{u}))$$

*is a convex relaxation of $\mathcal{F}_{\mathbf{q},t}$ on $\mathcal{U}$, provided that $\boldsymbol{\psi}^{cv}(t,\cdot)$ and $\boldsymbol{\psi}^{cc}(t,\cdot)$ are, respectively, convex and concave relaxations of $\mathbf{x}(t,\cdot)$ on $\mathcal{U}$.*

*Proof.* Choose any $(\lambda,\mathbf{u}_1,\mathbf{u}_2) \in [0,1] \times \mathcal{U} \times \mathcal{U}$ and let $\mathbf{u}_3 = \lambda\mathbf{u}_1 + (1-\lambda)\mathbf{u}_2$. Clearly, $\mathbf{u}_3(t) = \lambda\mathbf{u}_1(t) + (1-\lambda)\mathbf{u}_2(t)$ for all $t \in I$. For a.e. $t \in I$, $\mathbf{u}_1(t),\mathbf{u}_2(t),\mathbf{u}_3(t) \in U(t)$, and hence Lemma 11.3.3 shows that

$$\begin{aligned}
\mathbf{u}_{\mathbf{q}}(t,\mathbf{u}_3(t),\boldsymbol{\psi}^{cv}(t,\mathbf{u}_3),\boldsymbol{\psi}^{cc}(t,\mathbf{u}_3)) \leq& \lambda\mathbf{u}_{\mathbf{q}}(t,\mathbf{u}_1(t),\boldsymbol{\psi}^{cv}(t,\mathbf{u}_1),\boldsymbol{\psi}^{cc}(t,\mathbf{u}_1))\\
&+ (1-\lambda)\mathbf{u}_{\mathbf{q}}(t,\mathbf{u}_2(t),\boldsymbol{\psi}^{cv}(t,\mathbf{u}_2),\boldsymbol{\psi}^{cc}(t,\mathbf{u}_2))
\end{aligned}$$

and $\mathbf{u}_{\mathbf{q}}(t,\mathbf{u}_i(t),\boldsymbol{\psi}^{cv}(t,\mathbf{u}_i),\boldsymbol{\psi}^{cc}(t,\mathbf{u}_i)) \leq \mathbf{q}(t,\mathbf{u}_i(t),\mathbf{x}(t,\mathbf{u}_i))$ for all $i \in \{1,2,3\}$. □

It is not difficult to see that relaxations of $\mathcal{F}_\phi$ and $\mathcal{F}_{\mathbf{g}}$ can also be constructed by analogous procedures. Thus, the task of deriving a convex underestimating program for (11.1) has been reduced to that of deriving convex and concave relaxations for the end-point map of the control system (11.2). That development occupies the remainder of the chapter.

## 11.4 Relaxing Integral Functionals

Let $\bar{U}$, $U(t)$ and $\mathcal{U}$ be defined as in §11.1.1. In this section, relaxations of the functional

$$\mathcal{U} \ni \mathbf{u} \longmapsto \mathcal{H}(\mathbf{u}) \equiv \int_{t_0}^{t} \mathbf{h}(s, \mathbf{u}(s))ds, \tag{11.12}$$

are considered, where $\mathbf{h} : I \times \bar{U} \to \mathbb{R}^n$. Though no integral functionals appear in (11.1), the development in this section is required for relaxing the end-point maps of control systems in §11.5. Indeed, integral functionals have not been included in (11.1) because they can be treated by augmenting quadrature variables to the control system (11.2). For the benefit of §11.5, the following lemma is stated for a more general functionals than above.

**Lemma 11.4.1.** *Let* $\mathbf{h} : I \times \bar{U} \times \mathcal{U} \to \mathbb{R}^n$ *and suppose that the mapping* $t \mapsto \mathbf{h}(t, \mathbf{u}(t), \mathbf{u})$ *is in* $(L^1(I))^n$ *for every* $\mathbf{u} \in \mathcal{U}$*. If, for a.e.* $t \in I$*, the mapping*

$$U(t) \times \mathcal{U} \ni (\mathbf{p}, \mathbf{u}) \longmapsto \mathbf{h}(t, \mathbf{p}, \mathbf{u})$$

*is convex on* $U(t) \times \mathcal{U}$*, then the mapping*

$$\mathcal{U} \ni \mathbf{u} \longmapsto \mathcal{H}(\mathbf{u}) \equiv \int_{t_0}^{t} \mathbf{h}(s, \mathbf{u}(s), \mathbf{u})ds$$

*is convex on* $\mathcal{U}$*, for every* $t \in I$*.*

*Proof.* Choose any $(\lambda, \mathbf{u}_1, \mathbf{u}_2) \in [0, 1] \times \mathcal{U} \times \mathcal{U}$ and let $\mathbf{u}_3 = \lambda\mathbf{u}_1 + (1 - \lambda)\mathbf{u}_2$. Clearly, $\mathbf{u}_3(s) = \lambda\mathbf{u}_1(s) + (1 - \lambda)\mathbf{u}_2(s)$ for all $s \in I$. For a.e. $s \in I$, the hypothesis on $\mathbf{h}$ and the fact that $\mathbf{u}_1(s), \mathbf{u}_2(s) \in U(s)$ imply that

$$\mathbf{h}(s, \mathbf{u}_3(s), \mathbf{u}_3) \leq \lambda\mathbf{h}(s, \mathbf{u}_1(s), \mathbf{u}_1) + (1 - \lambda)\mathbf{h}(s, \mathbf{u}_2(s), \mathbf{u}_2).$$

Since this holds for a.e $s \in I$, linearity and monotonicity of the integral imply that, for any $t \in I$,

$$\int_{t_0}^{t} \mathbf{h}(s, \mathbf{u}_3(s), \mathbf{u}_3)ds \leq \lambda \int_{t_0}^{t} \mathbf{h}(s, \mathbf{u}_1(s), \mathbf{u}_1)ds + (1-\lambda) \int_{t_0}^{t} \mathbf{h}(s, \mathbf{u}_2(s), \mathbf{u}_2)ds.$$

The result follows since $\mathbf{u}_1, \mathbf{u}_2 \in \mathcal{U}$ and $\lambda \in (0, 1)$ are arbitrary. □

Lemma 11.4.1 will be used in its full generality in §11.5. For the moment, consider the functional $\mathcal{H}$ as defined in (11.12), with $\mathbf{h} : I \times \bar{U} \to \mathbb{R}^n$. Suppose further that $\mathbf{h}$ is $\mathcal{L}$-factorable with natural McCormick extension $\{\mathbf{h}\} : \mathcal{K} \to \mathbb{R}^n$, and that $U(t)$ is represented in $\mathcal{K}$ for every $t \in I$. Finally define the function $\mathbf{u_h} : I \times \mathbb{R}^{n_u} \to \mathbb{R}^n$ by

$$\mathbf{u_h}(t, \mathbf{p}) \equiv \{\mathbf{h}\}^{cv}(\mathrm{MC}(t, t, t, t), \mathrm{MC}(\mathbf{u}^L(t), \mathbf{u}^U(t), \mathbf{p}, \mathbf{p})). \tag{11.13}$$

Then, we have the following corollary of Lemma 11.4.1.

**Corollary 11.4.2.** *Suppose that the mapping $t \mapsto \mathbf{h}(t, \mathbf{u}(t))$ is in $(L^1(I))^n$ for every $\mathbf{u} \in \mathcal{U}$. Then the mapping*

$$\mathcal{U} \ni \mathbf{u} \longmapsto \mathcal{H}^{cv}(\mathbf{u}) \equiv \int_{t_0}^{t} \mathbf{u_h}(s, \mathbf{u}(s))ds \tag{11.14}$$

*is convex a convex relaxation of $\mathcal{H}$ on $\mathcal{U}$.*

*Proof.* For any $t \in I$, convexity of $\mathbf{u_h}(t, \cdot)$ on $U(t)$ follows from Theorem 2.7.10. Then, the result follows from Lemma 11.4.1, provided that the mapping $t \mapsto \mathbf{u_h}(t, \mathbf{u}(t))$ is in $(L^1(I))^n$ for every $\mathbf{u} \in \mathcal{U}$.

Since $U$ is continuous, $\mathbf{u_h}$ is continuous by Corollary 2.7.11. Choosing any $\mathbf{u} \in \mathcal{U}$, it follows that $t \mapsto \mathbf{u_h}(t, \mathbf{u}(t))$ is measurable (see [8]). Furthermore, $\mathbf{u_h}$ is bounded on $I \times \bar{U}$, which implies that $t \mapsto \mathbf{u_h}(t, \mathbf{u}(t))$ is bounded almost everywhere on $I$, and hence it is integrable on $I$. □

## 11.5 State Relaxations for Control Systems

Let $\bar{U}$, $U(t), \mathcal{U}$ and $D$ be defined as in §11.1.1, and consider the control system (11.2), where $\mathbf{f} : I \times \bar{U} \times D \to \mathbb{R}^{n_x}$. The following assumption holds throughout this section.

**Assumption 11.5.1.** $\mathbf{f}$ is continuous on $I \times \bar{U} \times D$ and, for every compact $K \subset D$, $\exists L_K \in \mathbb{R}_+$ such that

$$\|\mathbf{f}(t, \mathbf{p}, \mathbf{z}) - \mathbf{f}(t, \mathbf{p}, \hat{\mathbf{z}})\|_1 \leq L_K \|\mathbf{z} - \hat{\mathbf{z}}\|_1,$$

for every $(t, \mathbf{p}, \mathbf{z}, \hat{\mathbf{z}}) \in I \times \bar{U} \times K \times K$.

Under Assumption 11.5.1, it can be shown by standard methods that there exists a closed interval $I' \subset I$ such that, corresponding to each $\mathbf{u} \in \mathcal{U}$ there exists a unique, absolutely continuous solution of (11.2) on $I'$. It is assumed that such a solution exists on all of $I$; that is, there exists a unique mapping $\mathbf{x} : I \times \mathcal{U} \to D$ satisfying (11.2) a.e. in $I$ for every $\mathbf{u} \in \mathcal{U}$. The objective of this section is to derive relaxations for the family of mappings $\mathcal{X}_t(\mathbf{u}) \equiv \mathbf{x}(t, \mathbf{u})$ on $\mathcal{U}$, for each $t \in I$. It will be shown that these relaxations are given by the solutions of a suitable auxiliary control system which can be generated using McCormick's relaxation technique. The development here is analogous to the development of state relaxations for parametric ODEs using relaxation amplifying dynamics in Chapter 7. The development of a relaxation theory based on relaxation preserving dynamics for control systems is left for future work.

Let $\mathbf{u_f}, \mathbf{o_f} : I \times \bar{U} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ be defined by

$$\begin{aligned} \mathbf{u_f}(t, \mathbf{p}, \mathbf{z}^{cv}, \mathbf{z}^{cc}) = \{\mathbf{f}\}^{cv}(&\mathrm{MC}(t,t,t,t), \mathrm{MC}(\mathbf{u}^L(t), \mathbf{u}^U(t), \mathbf{p}, \mathbf{p}), \\ &\mathrm{MC}(\mathbf{x}^L(t), \mathbf{x}^U(t), \mathbf{z}^{cv}, \mathbf{z}^{cc})), \\ \mathbf{o_f}(t, \mathbf{p}, \mathbf{z}^{cv}, \mathbf{z}^{cc}) = \{\mathbf{f}\}^{cc}(&\mathrm{MC}(t,t,t,t), \mathrm{MC}(\mathbf{u}^L(t), \mathbf{u}^U(t), \mathbf{p}, \mathbf{p}), \\ &\mathrm{MC}(\mathbf{x}^L(t), \mathbf{x}^U(t), \mathbf{z}^{cv}, \mathbf{z}^{cc})). \end{aligned} \qquad (11.15)$$

The following properties of these functions will be required below.

**Corollary 11.5.2.** *For any* $\boldsymbol{\psi}^c, \boldsymbol{\psi}^C : I \times \mathcal{U} \to \mathbb{R}^{n_x}$ *and a.e.* $t \in I$*, the functions*

$$\begin{aligned} U(t) \times \mathcal{U} \ni (\mathbf{p}, \mathbf{u}) &\longmapsto \mathbf{u_f}(t, \mathbf{p}, \boldsymbol{\psi}^{cv}(t, \mathbf{u}), \boldsymbol{\psi}^{cc}(t, \mathbf{u})) \\ U(t) \times \mathcal{U} \ni (\mathbf{p}, \mathbf{u}) &\longmapsto \mathbf{o_f}(t, \mathbf{p}, \boldsymbol{\psi}^{cv}(t, \mathbf{u}), \boldsymbol{\psi}^{cc}(t, \mathbf{u})) \end{aligned}$$

*are, respectively, convex and concave relaxations of*

$$U(t) \times \mathcal{U} \ni (\mathbf{p}, \mathbf{u}) \longmapsto \mathbf{f}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{u}))$$

*on* $U(t) \times \mathcal{U}$*, provided that* $\boldsymbol{\psi}^{cv}(t, \cdot)$ *and* $\boldsymbol{\psi}^{cc}(t, \cdot)$ *are, respectively, convex and concave relaxations of* $\mathbf{x}(t, \cdot)$ *on* $\mathcal{U}$*.*

*Proof.* The result follows from Theorem 11.2.3, arguing exactly as in the proof of Lemma 11.3.3. □

**Corollary 11.5.3.** $\mathbf{u_f}$ *and* $\mathbf{o_f}$ *are continuous on* $I \times \bar{U} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$*, and* $\exists L \in \mathbb{R}_+$ *such that*

$$\begin{aligned} \|\mathbf{u_f}(t, \mathbf{p}, \mathbf{z}, \mathbf{y}) - \mathbf{u_f}(t, \mathbf{p}, \hat{\mathbf{z}}, \hat{\mathbf{y}})\|_1 + &\|\mathbf{o_f}(t, \mathbf{p}, \mathbf{z}, \mathbf{y}) - \mathbf{o_f}(t, \mathbf{p}, \hat{\mathbf{z}}, \hat{\mathbf{y}})\|_1 \\ &\leq L(\|\mathbf{z} - \hat{\mathbf{z}}\|_1 + \|\mathbf{y} - \hat{\mathbf{y}}\|_1) \end{aligned}$$

*for all* $(t, \mathbf{p}, \mathbf{z}, \mathbf{y}, \hat{\mathbf{z}}, \hat{\mathbf{y}}) \in I \times \bar{U} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$*.*

*Proof.* The result follows from Corollary 2.7.8 using an argument analogous to that in Lemma 7.6.7. □

Now, define the auxiliary control system by

$$\begin{aligned} \dot{\mathbf{x}}^{cv}(t, \mathbf{u}) &= \mathbf{u_f}(t, \mathbf{u}(t), \mathbf{x}^{cv}(t, \mathbf{u}), \mathbf{x}^{cc}(t, \mathbf{u})), \quad \mathbf{x}^{cv}(t_0, \mathbf{u}) = \mathbf{x}_0, \qquad (11.16) \\ \dot{\mathbf{x}}^{cc}(t, \mathbf{u}) &= \mathbf{o_f}(t, \mathbf{u}(t), \mathbf{x}^{cv}(t, \mathbf{u}), \mathbf{x}^{cc}(t, \mathbf{u})), \quad \mathbf{x}^{cc}(t_0, \mathbf{u}) = \mathbf{x}_0, \end{aligned}$$

for a.e. $t \in I$ and every $\mathbf{u} \in \mathcal{U}$. The main result of this section is the following:

**Theorem 11.5.4.** *The auxiliary system* (11.16) *has a unique solution* $(\mathbf{x}^{cv}, \mathbf{x}^{cc})$ *on all of* $I \times \mathcal{U}$*, and* $\mathbf{x}^{cv}(t, \cdot)$ *and* $\mathbf{x}^{cc}(t, \cdot)$ *are, respectively, convex and concave relaxations of* $\mathbf{x}(t, \cdot)$ *on* $\mathcal{U}$*, for every* $t \in I$*.*

According to the previous theorem, the desired relaxations of the endpoint map $\mathcal{X}_t$ are given by $\mathcal{X}_t^{cv}(\mathbf{u}) \equiv \mathbf{x}^{cv}(t, \mathbf{u})$ and $\mathcal{X}_t^{cc}(\mathbf{u}) \equiv \mathbf{x}^{cc}(t, \mathbf{u})$, $\forall (t, \mathbf{u}) \in I \times \mathcal{U}$. Combining these relaxations with the analysis in §11.1.1 and 11.3, the desired relaxation of (11.1) are readily derived.

## 11.5.1 Proof of Theorem 11.5.4

### Preliminaries

**Theorem 11.5.5.** *Consider the ODEs* (11.2) *and suppose that* $\mathbf{f}$ *is continuous on* $I \times \bar{U} \times \mathbb{R}^{n_x}$ *and* $\exists L \in \mathbb{R}_+$ *such that*

$$\|\mathbf{f}(t, \mathbf{p}, \mathbf{z}) - \mathbf{f}(t, \mathbf{p}, \hat{\mathbf{z}})\|_1 \leq L \|\mathbf{z} - \hat{\mathbf{z}}\|_1,$$

*for every* $(t, \mathbf{p}, \mathbf{z}, \hat{\mathbf{z}}) \in I \times \bar{U} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$*. Given any* $\mathbf{x}^0 : I \times \mathcal{U} \to \mathbb{R}^{n_x}$ *such that* $\mathbf{x}^0(\cdot, \mathbf{u})$ *is absolutely continuous on* $I$ *for any* $\mathbf{u} \in \mathcal{U}$*, the sequence of successive approximations defined recursively by*

$$\mathbf{x}^{k+1}(t, \mathbf{u}) = \mathbf{x}_0 + \int_{t_0}^{t} \mathbf{f}(s, \mathbf{u}(s), \mathbf{x}^k(s, \mathbf{u}))ds \tag{11.17}$$

*satisfies the following conditions:*

1. *For each* $\mathbf{u} \in \mathcal{U}$*, each* $\mathbf{x}^k$ *exists and is absolutely continuous on* $I$*,*

2. *For each* $\mathbf{u} \in \mathcal{U}$*, the sequence* $\{\mathbf{x}^k(\cdot, \mathbf{u})\}$ *converges uniformly on* $I$ *to an absolutely continuous limit function* $\mathbf{x}(\cdot, \mathbf{u})$ *satisfying* (11.2) *uniquely.*

*Proof.* Fix any $\mathbf{u} \in \mathcal{U}$. By hypothesis, $\mathbf{x}^0(\cdot, \mathbf{u})$ is absolutely continuous on $I$. Suppose this is true of $\mathbf{x}^k$. Continuity of $\mathbf{f}$ and measurability of $\mathbf{u}$ and $\mathbf{x}^k(\cdot, \mathbf{u})$ imply that $\mathbf{f}(\cdot, \mathbf{u}(\cdot), \mathbf{x}^k(\cdot, \mathbf{u}))$ is measurable (see [8]). Since this function is also bounded a.e. on

$I$, it is integrable and hence (11.17) defines $\mathbf{x}^{k+1}(\cdot, \mathbf{u})$ as an absolutely continuous function on $I$. Induction shows that this property holds for all $k \in \mathbb{N}$.

Define $\gamma(t) \equiv \|\mathbf{f}(t, \mathbf{u}(t), \mathbf{x}^1(t, \mathbf{u})) - \mathbf{f}(t, \mathbf{u}(t), \mathbf{x}^0(t, \mathbf{u}))\|_1$ and let $\bar{\gamma} = \text{ess sup}_{t \in I} \gamma(t)$. The assumption that $U(t) \subset \bar{U}$ for all $t \in I$, with $\bar{U}$ compact, along with the continuity of $\mathbf{f}$, $\mathbf{x}^1$ and $\mathbf{x}^0$, ensures that $\bar{\gamma}$ is finite. It will be shown that

$$\|\mathbf{x}^{k+1}(t, \mathbf{u}) - \mathbf{x}^k(t, \mathbf{u})\|_1 \leq \frac{\bar{\gamma} L^k (t - t_0)^k}{L k!}, \tag{11.18}$$

for all $t \in I$ and every $k \in \mathbb{N}$. For $k = 1$, (11.17) directly gives

$$\begin{aligned} \|\mathbf{x}^2(t, \mathbf{u}) - \mathbf{x}^1(t, \mathbf{u})\|_1 &\leq \int_{t_0}^{t} \|\mathbf{f}(s, \mathbf{u}(s), \mathbf{x}^1(s, \mathbf{u})) - \mathbf{f}(s, \mathbf{u}(s), \mathbf{x}^0(s, \mathbf{u}))\|_1 ds \\ &\leq \bar{\gamma}(t - t_0), \quad \forall t \in I. \end{aligned}$$

Supposing that (11.18) holds for some arbitrary $k$, the Lipschitz condition on $\mathbf{f}$ gives

$$\begin{aligned} \|\mathbf{x}^{k+2}(t, \mathbf{u}) - \mathbf{x}^{k+1}(t, \mathbf{u})\|_1 &\leq L \int_{t_0}^{t} \|\mathbf{x}^{k+1}(s, \mathbf{u}) - \mathbf{x}^k(s, \mathbf{u})\|_1 ds, \\ &\leq \frac{\bar{\gamma} L^{k+1}}{L k!} \int_{t_0}^{t} (s - t_0)^k ds, \\ &\leq \frac{\bar{\gamma} L^{k+1} (t - t_0)^{k+1}}{L(k+1)!}, \quad \forall t \in I. \end{aligned}$$

Thus, induction proves (11.18). Now, for any $n, m \in \mathbb{N}$ with $m > n$, expansion by the triangle inequality and application of Equation (11.18) gives

$$\|\mathbf{x}^m(t, \mathbf{u}) - \mathbf{x}^n(t, \mathbf{u})\|_1 \leq \sum_{k=n}^{m-1} \frac{\bar{\gamma} L^k (t_f - t_0)^k}{L k!}, \tag{11.19}$$

for all $t \in I$. But

$$\sum_{k=0}^{\infty} \frac{\bar{\gamma} L^k (t_f - t_0)^k}{L k!} = \frac{\bar{\gamma}}{L} e^{L(t_f - t_0)} < +\infty,$$

and hence

$$\lim_{n\to\infty}\sum_{k=n}^{\infty}\frac{\bar{\gamma}L^k(t_f-t_0)^k}{Lk!}=0, \tag{11.20}$$

which implies by (11.19) that the sequence $\{\mathbf{x}^k(\cdot,\mathbf{u})\}$ is uniformly Cauchy on $I$. Continuity implies that this sequence converges uniformly to a continuous limit function $\mathbf{x}(\cdot,\mathbf{u})$ on $I$.

Next, it is shown that $\mathbf{x}$ is a solution of (11.2) on $I\times\mathcal{U}$. For any $\mathbf{u}\in\mathcal{U}$, the Lipschitz condition on $\mathbf{f}$ gives,

$$\begin{aligned}&\|\int_{t_0}^{t}\mathbf{f}(s,\mathbf{u}(s),\mathbf{x}^k(s,\mathbf{u}))ds-\int_{t_0}^{t}\mathbf{f}(s,\mathbf{u}(s),\mathbf{x}(s,\mathbf{u}))ds\|_1\\&\leq L\int_{t_0}^{t}\|\mathbf{x}^k(s,\mathbf{u})-\mathbf{x}(s,\mathbf{u})\|_1 ds,\quad\forall t\in I,\end{aligned}$$

so uniform convergence of $\{\mathbf{x}^k(\cdot,\mathbf{u})\}$ to $\mathbf{x}(\cdot,\mathbf{u})$ on $I$ implies that

$$\lim_{k\to\infty}\int_{t_0}^{t}\mathbf{f}(s,\mathbf{u}(s),\mathbf{x}^k(s,\mathbf{u}))ds=\int_{t_0}^{t}\mathbf{f}(s,\mathbf{u}(s),\mathbf{x}(s,\mathbf{u}))ds,\quad\forall t\in I.$$

Then, taking limits on both sides of (11.17) gives

$$\mathbf{x}(t,\mathbf{u})=\mathbf{x}_0+\int_{t_0}^{t}\mathbf{f}(s,\mathbf{u}(s),\mathbf{x}(s,\mathbf{u}))ds,\quad\forall t\in I,$$

which implies that $\mathbf{x}(\cdot,\mathbf{u})$ is absolutely continuous and solves (11.2). Uniqueness of $\mathbf{x}$ now follows (for each fixed $\mathbf{u}\in\mathcal{U}$) by a standard application of Gronwall's inequality (Proposition 1, Ch. 2, Sec. 4, [14]). □

**Proof**

Choose any vectors $\mathbf{x}^L,\mathbf{x}^U\in\mathbb{R}^{n_x}$, such that $\mathbf{x}^L\leq\mathbf{x}(t,\mathbf{u})\leq\mathbf{x}^U$, $\forall(t,\mathbf{u})\in I\times\mathcal{U}$. Under Assumption 11.3.1, such vectors certainly exist. Let $\mathbf{x}^{cv,0}(t,\mathbf{u})=\mathbf{x}^L$ and $\mathbf{x}^{cc,0}(t,\mathbf{u})=\mathbf{x}^U$, $\forall(t,\mathbf{u})\in I\times\mathcal{U}$, and consider the successive approximations defined recursively by

$$\mathbf{x}^{cv,k+1}(t,\mathbf{u}) = \mathbf{x}_0 + \int_{t_0}^{t} \mathbf{u_f}(s,\mathbf{u}(s),\mathbf{x}^{cv,k}(s,\mathbf{u}),\mathbf{x}^{cc,k}(s,\mathbf{u}))ds, \tag{11.21}$$
$$\mathbf{x}^{cc,k+1}(t,\mathbf{u}) = \mathbf{x}_0 + \int_{t_0}^{t} \mathbf{o_f}(s,\mathbf{u}(s),\mathbf{x}^{cv,k}(s,\mathbf{u}),\mathbf{x}^{cc,k}(s,\mathbf{u}))ds.$$

Note that $\mathbf{u_f}$ and $\mathbf{o_f}$ are defined on $I \times \bar{U} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ and Lipschitz on all of $\mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ uniformly on $I \times \bar{U}$ by Corollary 11.5.3. Thus, Theorem 11.5.5 may be applied to (11.16), which shows that the successive approximations $\mathbf{x}^{cv,k}$ and $\mathbf{x}^{cc,k}$ in (11.21) exist and, for each fixed $\mathbf{u} \in \mathcal{U}$, converge uniformly on $I$ to the unique solutions of (11.16), $\mathbf{x}^{cv}(\cdot,\mathbf{u})$ and $\mathbf{x}^{cc}(\cdot,\mathbf{u})$.

Next, note that $\mathbf{x}^{cv,0}(t,\cdot)$ and $\mathbf{x}^{cc,0}(t,\cdot)$ are trivially convex and concave relaxations of $\mathbf{x}(t,\cdot)$ on $\mathcal{U}$, respectively, for each fixed $t \in I$. Suppose that the same is true of $\mathbf{x}^{cv,k}$ and $\mathbf{x}^{cc,k}$. Then, by Corollary 11.5.2,

$$U(t) \times \mathcal{U} \ni (\mathbf{p},\mathbf{u}) \longmapsto \mathbf{u_f}(t,\mathbf{p},\mathbf{x}^{cv,k}(t,\mathbf{u}),\mathbf{x}^{cc,k}(t,\mathbf{u}))$$
$$U(t) \times \mathcal{U} \ni (\mathbf{p},\mathbf{u}) \longmapsto \mathbf{o_f}(t,\mathbf{p},\mathbf{x}^{cv,k}(t,\mathbf{u}),\mathbf{x}^{cc,k}(t,\mathbf{u}))$$

are, respectively, convex and concave relaxations of

$$U(t) \times \mathcal{U} \ni (\mathbf{p},\mathbf{u}) \longmapsto \mathbf{f}(t,\mathbf{p},\mathbf{x}(t,\mathbf{u}))$$

on $U(t) \times \mathcal{U}$, for all $t \in I$. Lemma 11.4.1 shows that

$$\mathcal{U} \ni \mathbf{u} \longmapsto \int_{t_0}^{t} \mathbf{u_f}(s,\mathbf{u}(s),\mathbf{x}^{cv,k}(s,\mathbf{u}),\mathbf{x}^{cc,k}(s,\mathbf{u}))ds$$
$$\mathcal{U} \ni \mathbf{u} \longmapsto \int_{t_0}^{t} \mathbf{o_f}(s,\mathbf{u}(s),\mathbf{x}^{cv,k}(s,\mathbf{u}),\mathbf{x}^{cc,k}(s,\mathbf{u}))ds$$

are, respectively, convex and concave on $\mathcal{U}$, for every fixed $t \in I$. Then, (11.21) shows that $\mathbf{x}^{cv,k+1}(t,\cdot)$ and $\mathbf{x}^{cc,k+1}(t,\cdot)$ are, respectively, convex and concave on $\mathcal{U}$ for every fixed $t \in I$.

Now, considering the under and overestimating properties of the functions $\mathbf{u_f}$ and

$\mathbf{o_f}$ described above, for any $\mathbf{u} \in \mathcal{U}$ and a.e. $s \in I$, we have

$$\begin{aligned}\mathbf{u_f}(s, \mathbf{u}(s), \mathbf{x}^{cv,k}(s, \mathbf{u}), \mathbf{x}^{cc,k}(s, \mathbf{u})) &\leq \mathbf{f}(s, \mathbf{u}(s), \mathbf{x}(s, \mathbf{u})),\\ &\leq \mathbf{o_f}(s, \mathbf{u}(s), \mathbf{x}^{cv,k}(s, \mathbf{u}), \mathbf{x}^{cc,k}(s, \mathbf{u})).\end{aligned}$$

Combining this with integral monotonicity,

$$\begin{aligned}\int_{t_0}^{t} \mathbf{u_f}(s, \mathbf{u}(s), \mathbf{x}^{cv,k}(s, \mathbf{u}), \mathbf{x}^{cc,k}(s, \mathbf{u}))ds &\leq \int_{t_0}^{t} \mathbf{f}(s, \mathbf{u}(s), \mathbf{x}(s, \mathbf{u}))ds,\\ &\leq \int_{t_0}^{t} \mathbf{o_f}(s, \mathbf{u}(s), \mathbf{x}^{cv,k}(s, \mathbf{u}), \mathbf{x}^{cc,k}(s, \mathbf{u}))ds,\end{aligned}$$

for all $(t, \mathbf{u}) \in I \times \mathcal{U}$. Then, (11.21) shows that

$$\mathbf{x}^{cv,k+1}(t, \mathbf{u}) \leq \mathbf{x}_0 + \int_{t_0}^{t} \mathbf{f}(s, \mathbf{u}(s), \mathbf{x}(s, \mathbf{u}))ds \leq \mathbf{x}^{cc,k+1}(t, \mathbf{u}), \quad \forall (t, \mathbf{u}) \in I \times \mathcal{U},$$

which, by the integral form of (11.2), gives

$$\mathbf{x}^{cv,k+1}(t, \mathbf{u}) \leq \mathbf{x}(t, \mathbf{u}) \leq \mathbf{x}^{cc,k+1}(t, \mathbf{u}), \quad \forall (t, \mathbf{u}) \in I \times \mathcal{U}.$$

Therefore, by induction, $\mathbf{x}^{cv,k}(t, \cdot)$ and $\mathbf{x}^{cc,k}(t, \cdot)$ are, respectively, convex and concave relaxations of $\mathbf{x}(t, \cdot)$ on $\mathcal{U}$, for each fixed $t \in I$ and every $k \in \mathbb{N}$.

It was shown above that, as $k \to \infty$, $\mathbf{x}^{cv,k}$ and $\mathbf{x}^{cc,k}$ converge pointwise to the unique solutions of (11.16) on $I \times \mathcal{U}$. Then, taking limits, it is clear that $\mathbf{x}^{cv}(t, \cdot)$ and $\mathbf{x}^{cc}(t, \cdot)$ are, respectively, convex and concave relaxations of $\mathbf{x}(t, \cdot)$ on $\mathcal{U}$, for each fixed $t \in I$.

## 11.6 Conclusions

A method has been presented for computing a rigorous lower bound for the nonconvex optimal control problem (11.1). In particular, a constructive procedure was described, based on natural McCormick extensions, which produces a convex optimization problem whose solution is guaranteed to underestimate the infimum in

(11.1). Supposing that this convex program can be solved to global optimality, using for example the methods described in [15], a guaranteed lower bound on the infimum in (11.1) is obtained. Computing guaranteed lower bounds is a crucial step required by branch-and-bound global optimization algorithms. Thus, the method developed here provides a key ingredient required for branch-and-bound global optimization of nonconvex optimal control problems. Finally, the proposed lower bounding technique is distinguished from previous work in that it does not require control parameterization. The derived relaxations are valid in the original space of admissible control functions.

# Chapter 12

# Concluding Remarks

The work in this thesis has addressed two problems of broad interest concerning the behavior of nonlinear differential-algebraic process models subject to a range of permissible process inputs. In contrast to classical methods for treating such problems, the methods herein provide information that is *global* in nature. The first problem addressed was that of enclosing the set of all possible solutions of a given DAE model subject to a range of permissible input functions and/or model parameters. This analysis is useful in nearly any application in which process disturbances and/or model uncertainties are of significant concern, and has numerous practical applications in process control. The second problem addressed was that of solving optimization problems constrained by differential-algebraic equations to guaranteed global optimality. Such optimization problems arise in the design and control of transient processes, as well as in the important area of parameter estimation for dynamic process models.

## 12.1 Summary of Contributions

### 12.1.1 Algorithms

The most immediate and tangible results of this thesis are the algorithms that have been developed for the tasks outlined above. In Chapters 3 and 4, the full-space bounding (FSB) and reduced-space bounding (RSB) methods were developed for computing time-varying interval bounds on the solutions of explicit ODE models subject to a permissible set of input functions and/or model parameters. These methods extend a standard bounding technique which is known to be very efficient but also very conservative. For problems where it is possible through physical insight to provide an *a priori* convex polyhedral enclosure of the ODE solutions, the FSB and RSB methods are able to achieve a dramatic reduction in the conservatism of the resulting bounds by exploiting this information at a fundamental level within the bounding procedure. These methods are implemented using only efficient interval computations and a state-of-the-art numerical integration routine, so that improved enclosures are achieved without sacrificing the computational efficiency of the original method. Finally, it was demonstrated that the class of problems for which the required *a priori* information is available is by no means small or insignificant. Rather, it contains the important class of systems that can be regarded as describing fluxes through a network, notably including models of chemical reaction kinetics. For such systems, *a priori* physical information can be obtained and exploited easily and automatically by a simple matrix analysis.

In Chapters 5 and 6, the two-phase and single-phase methods were developed for computing interval bounds on the solutions of systems of nonlinear semi-explicit index-one DAEs subject to a range of model parameters. Again, an efficient numerical implementation of both methods was developed using an interval Newton type method and a state-of-the-art numerical integration routine. Numerical case studies for these algorithms demonstrate that they are capable of producing results with the same efficiency that makes differential inequalities methods for ODEs attractive. Moreover, as compared to the ODE methods, the methods developed for semi-explicit index-one DAEs provide very reasonable bounds, especially considering the fact that, at present, no *a priori* physical information is used in their computation. Thus, these methods extend the efficient class of differential inequalities based bounding techniques to systems of DAEs for the first time.

In Chapter 7, two algorithms were presented for computing nonlinear convex and concave relaxations of the parametric solutions of nonlinear ODEs. For both methods, relaxations are computed through the numerical solution of an auxiliary system of ODEs that is derived efficiently and automatically using the generalized McCormick relaxation technique developed in Chapter 2. Comparing these methods, the superior method was found to be method using relaxation preserving dynamics (RPD). Compared to the state-of-the-art relaxation method in [164], the RPD method was shown to provide significantly tighter relaxations. In Chapter 8, the methods of Chapter 7 were extended to handle systems of nonlinear semi-explicit index-one DAEs, providing convex and concave relaxations for the solutions of such systems for the first time.

In Chapter 9, a further algorithm was developed for enclosing the solutions of parametric ODEs and semi-explicit index-one DAEs, in this case within a convex polyhedral set rather than an interval. It was shown that the resulting enclosure can be significantly sharper than the interval enclosures produced by the methods of Chapters 3-6. On the other hand, obtaining a valid enclosure from this technique has previously only been possible through the global solution of several potentially nonconvex dynamic optimization problems, which is prohibitively expensive in general [39, 41]. Using the relaxation techniques developed in Chapters 7 and 8, the proposed algorithm is able to provide a guaranteed convex polyhedral enclosure while solving only convex dynamic optimization problems.

In Chapter 10, a deterministic global optimization algorithm was developed for problems with semi-explicit index-one DAEs embedded. This algorithm is based on a standard spatial branch-and-bound framework, where the lower bounding procedure is enabled by the relaxation techniques developed in Chapter 8. Aside from intractable methods based on a total discretization approach, this is the first method capable of solving optimization problems with DAEs embedded to guaranteed global optimality. The algorithm has been demonstrated for several numerical case studies and shown to perform comparably to state-of-the-art methods for optimization problems with ODEs embedded.

### 12.1.2 Theoretical Contributions

The objective of this thesis has always been to develop practical numerical methods for practical engineering problems. Even so, the final result is largely a piece of abstract mathematical analysis. Throughout this analysis, certain results, computational techniques, and basic principles have emerged that marked turning points in the thesis and seem to be significant contributions in their own right. A significant effort has been made throughout the thesis to present these ideas in the most general and broadly applicable way possible. To be sure, this has robbed some chapters of a degree of clarity and intuition that they might otherwise have had. On the other hand, it has also simplified matters in several places where distinct variations of a method could be proven by application of the same result, or where methods for ODEs and DAEs could be derived by a unified abstract development. However, the primary motivation for this generality is that global dynamic optimization is a field in flux. While many of the numerical results presented herein are promising and represent significant advances over the state-of-the-art, it is nonetheless clear that much work remains to be done before these methods can be routinely applied to practical engineering problems. Thus, it is worthwile to point out some of the fundamental theoretical contributions of the thesis that, as the field progresses, may prove to be useful beyond the specific methods that have so far been derived from them.

The first of these contributions is the abstract development of McCormick's relaxation technique in Chapter 2, leading to the notion of a natural McCormick extension, or equivalently, a generalized McCormick relaxation. The fundamental contribution of this analysis is the ability to construct relaxations of multivariate compositions (§2.7.2). This capability is essential for the proposed numerical implementation of the relaxation theories for ODEs and DAEs developed in Chapters 7 and 8. Furthermore, it is essential for making use of these relaxations for constructing convex underestimating programs in the context of global dynamic optimization. Notably, this procedure permits one to compute relaxations of non-factorable functions by a fairly direct application of McCormick's technique. As another example of this, generalized McCormick relaxations have also been used to compute convex and concave relaxations for the solutions of implicit systems of nonlinear algebraic equations [169]. On the whole, this procedure extends the reach of McCormick's relaxation technique, and hence global optimization, to the important class of optimization problems in which the objective and constraints are implicitly defined by the solutions of an embedded model.

Among the most important theoretical contributions of this thesis are the general comparison theorems of Chapter 3 (Theorems 3.5.1 and 3.5.4). From these results, we have developed very effective bounding procedures for parametric ODEs whose solutions are known to lie within convex polyhedral sets. However, Theorems 3.5.1 and 3.5.4 address a much broader issue. In particular, the conditions of these theorems formally delineate valid and invalid uses of *arbitrary* auxiliary information in the context of differential inequalities bounding techniques. The use of redundant information to refine conservative approximations is a well-established tenet of global optimization and rigorous computing at large, and the numerical results of Chapter 4 clearly demonstrate that this tool is no less essential for dynamic problems. What instead seems unique to dynamic problems is that the distinction between valid and invalid uses of such information, at least with regards to differential inequalities, is complicated to the point of mathematical pedantry. Thus the utility of Theorems 3.5.1 and 3.5.4 is that they allow one to check the validity of putative new bounding techniques with the relative ease of verifying a few simple conditions. A particularly useful incarnation of these conditions was presented in §3.6.1, where simple criteria are established for a valid use of efficient interval computations to exploit an arbitrary *a priori* enclosure. Among other applications, this suggests an avenue for exploiting nonlinear solution invariants in dynamic models using, for example, interval Newton type methods as in Chapter 5.

The final broadly important theoretical contribution of this thesis is the formulation of the conditions of relaxation amplifying dynamics (RAD) and relaxation preserving dynamics (RPD). Of course, these conditions underly the novel relaxation methods for ODEs and DAEs developed in Chapters 7 and 8. However, these conditions provide important insights beyond the implementations given here. The conditions of RAD have been shown to result in relaxations which accumulate conservatism and become in a sense more convex (or concave) as integration proceeds. These properties in turn result in unnecessarily weak relaxations and, at least empirically, a poor rate of convergence. Interestingly, these properties are a direct result of precisely the same conditions from which the method derives its validity in the first place. Thus, the conditions of RAD are illustrative of two fundamental issues for global optimization that are unique to dynamic problems and have not been previously understood. On the other hand, the conditions of RPD correct these problems and result in very satisfactory relaxation techniques. The conditions through which RPD type relaxations are guaranteed to underestimate and overestimate the function of interest are derived from standard arguments in the theory of differential inequalities. On the other hand, the principle through which the RPD conditions impart convexity and concavity on the resulting relaxations is entirely novel and has not previously been exploited by any method. As such, it provides a new principle for relaxing the solutions of dynamic systems that can be used to motivate and analyze future methods.

In both the state bounding theory of Chapter 3 and the relaxation theory of Chapter 7, the fundamental results have been proven for an arbitrary absolutely continuous function. That is, these results do not require that the function to be bounded or relaxed is a solution of a particular type of dynamic system. Though we have proposed complete methods only for systems of ODEs and semi-explicit index-one DAEs, the fundamental principles may be applied directly to the solutions of more complex systems including fully implicit DAEs, high-index DAEs, and hybrid discrete-continuous dynamical systems. Deriving practical computational methods from these conditions then only requires that one can construct appropriate auxiliary systems. For this task, it seems very likely that generalized McCormick relaxations will again prove to be a useful tool.

## 12.2 Outlook

It is difficult to appraise the state-of-the-art in reachability computations for nonlinear dynamic systems. Mostly, this is because the literature has historically developed into several pockets that remain largely isolated from one another. The most prominent of these are the literature on Taylor methods, originally developed for the purposes of validated numerical integration, the literature on differential inequalities, which were used only as a mathematical tool until fairly recently, and the literature in the process control community, where methods are largely rooted in linear systems theory and applied to nonlinear systems through the extensive use of local linear approximations. Because they are based on quite distinct ideas, these classes of methods each come with a unique list of advantageous and disadvantages with respect to accuracy and computational efficiency that make an intuitive comparison difficult. Moreover, given the lack of communication between these literatures and the complexity of implementing some of the available methods, there have been no adequate numerical comparisons between these classes of methods, and unfortunately too few good comparisons even within them.

What does seem to be generally agreed upon is that it is extremely difficult to compute accurate global information about the solutions of nonlinear dynamic models. In all the available literature, there are almost no examples demonstrating a reasonably sharp enclosure for a nonlinear system with more than three state variables and three uncertain parameters. Even for small systems, computing very sharp enclosures can require exorbitant computational effort. With the methods developed in this thesis, we have demonstrated accurate bounds on the solutions of substantially larger models subject to larger uncertainty. While these results represent significant advances, the models considered here are still a far cry from the size and complexity of those arising in many important applications.

Fortunately, there are many avenues for future research. Among these, probably few would be as enlightening as simply taking stock of the available methods and undertaking extensive numerical comparisons. Again, the classes of methods described above are based on fairly distinct ideas, and it is reasonable to suspect that such a comparison would enable one to incorporate the advantageous features of each into a new generation of methods. For example, the enabling features of early Taylor methods are the use of high-order Taylor series expansions of the states with respect to time, and effective control of the wrapping effect through, for example, Lohner's method [130]. Neither of these techniques has ever been leveraged in the context of differential inequalities, though it it seems very likely that they could be.

If there is one contribution of this thesis that should resonate throughout the field, it is the observation that using redundant information in a bounding technique can have a profound impact on the quality of the resulting enclosure. As we have mentioned previously, the use of redundant information to refine conservative approximations is a well-established and highly successful technique in the areas of constraint propagation and global optimization. However, to our knowledge, it has not been previously been used in the context of reachable set computations. The dramatic effect of using such information in the methods of this thesis makes a compelling case that similar methods will be a key ingredient for addressing larger and more complex models in the future. There are several topics to explore in this regard, such as the existence of other important classes of systems that we have not considered for which *a priori* information is readily available, and the question of whether or not it is effective to augment a system artificially with additional redundant equations in order to obtain a sharper enclosure through similar methods.

At the outset of this thesis, the state-of-the-art in global dynamic optimization was undoubtedly the method of Singer and Barton [164], which has been demonstrated to solve problems with up to three state variables and three decision variables in reasonable computational time. Since then, there as been quite a lot of activity. In addition to the methods developed in this thesis, impressive new methods for computing enclosures and relaxations of the solutions of ODEs have been proposed by Lin and Stadtherr [105, 104] and Sahlodin and Chachuat [151, 150]. When implemented in a complete global optimization algorithm [104], the method of Stadtherr and Lin has been shown to outperform that of Singer and Barton for some but not all test problems, and has been shown to solve one problem with up to five decisions.

Unfortunately, the improved bounding and relaxation methods for ODEs developed in this thesis have not yet been implemented in a complete optimization algorithm. However, compared to the methods used in [164], the bounding and relaxation methods developed here have been shown to be much tighter, which is expected to result in a significantly improved global optimization algorithm. Because they were developed only very recently, the methods of Sahlodin and Chachuat [151, 150] have also not yet been demonstrated within a global optimization algorithm. Furthermore, they have not been compared to the improved relaxation methods of this thesis, though they have been shown to compare very favorably to the method Stadtherr and Lin [105]. Thus, a fair assessment of the field at present is that there are many interesting and potentially powerful ideas available, and too few useful metrics and numerical comparisons by which to understand them and direct future efforts.

With the results of this thesis, global dynamic optimization methods have been extended to problems with semi-explicit index-one DAEs embedded for the first time. However, like existing methods for ODEs, the practical applicability of this method is limited to very small problems. Thus, much work remains to be done before this method can be applied to most problems of practical interest.

There are several target areas for future research in global optimization of ODEs an DAEs. In the case of DAEs, an obvious task is to extend the bounding methods based on *a priori* enclosures developed here to the case of DAEs. This alone should lead to a substantial improvement in the performance of the global optimization algorithm of Chapter 10.

Following the work in [30], there has recently been a renewed interest in the convergence rate of lower bounding procedures, which can have a very significant effect on the overall run-time of a spatial B&B algorithm. The issue of convergence rate has been almost completely ignored in the literature on global dynamic optimization to date, and at present there is no convergence rate analysis available for any of the available techniques. These analyses are therefore a primary target for future research in understanding and improving the available global dynamic optimization methods.

Another important area for future work is to develop effective constraint propagation and optimality based range reduction techniques for global dynamic optimization problems. For standard NLPs, these methods have revolutionized modern methods, which could not otherwise address practical problem instances. Such methods have been employed in the algorithms of Singer and Barton [164] and Lin and Stadtherr [104], but not nearly to the same extent. A related criticism of the field, and particularly of this thesis, is that the problem of enclosing and/or relaxing the solutions of the embedded dynamic system has largely been treated in isolation from the rest of the optimization problem. Certainly, these computations are the weak point of the available algorithms and account for their limited applicability. On the other hand, effective algorithms for standard NLPs suggest that all of the available information in a problem should be brought to bear in all aspects of the lower bounding computation. Why then, has as it always seemed appropriate to bound *all* solutions of the embedded dynamic system when we are only interested in those that are feasible and potentially optimal? This and other similar questions deserve serious consideration moving forward.

# Bibliography

[1] O. Abel and Wolfgang Marquardt. Scenario-integrated modeling and optimization of dynamic systems. *AIChE Journal*, 46(4):803–823, 2000.

[2] Bambang S. Adiwijaya, Paul I. Barton, and Bruce Tidor. Biological network design strategies: discovery through dynamic optimization. *Mol. BioSyst.*, 2:650–659, 2006.

[3] C. S. Adjiman, I. P. Androulakis, and C. A. Floudas. A global optimization method, $\alpha$BB, for general twice-differentiable constrained NLPs - II. implementation and computational results. *Computers and Chemical Engineering*, 22(9):1159–1179, 1998.

[4] C. S. Adjiman, S. Dallwig, C. A. Floudas, and A. Neumaier. A global optimization method, $\alpha$BB, for general twice-differentiable constrained NLPs - I. theoretical advances. *Computers and Chemical Engineering*, 22(9):1137–1158, 1998.

[5] M. Althoff, O. Stursberg, and M. Buss. Reachability analysis of nonlinear systems with uncertain parameters using conservative linearization. In *Proc. 47th IEEE Conference on Decision and Control*, pages 4042–4048, 2008.

[6] R. Alur, T. A. Henzinger, and P.-H. Ho. Automatic symbolic verification of embedded systems. *IEEE Transactions on Software Engineering*, 22(3):181–201, 1996.

[7] I. P. Androulakis, C. D. Maranas, and C. A. Floudas. $\alpha$BB: A global optimization method for general constrained nonconvex problems. *J. Glob. Optim.*, 7:337–363, 1995.

[8] Jurgen Appell and Petr P. Zabrejko. *Nonlinear superposition operators.* Cambridge University Press, Cambridge, 1990.

[9] Rutherford Aris. Prolegomena to the rational analysis of systems of chemical reactions. *Arch. Rational Mech. Anal.*, 19(2):81–99, 1965.

[10] E. Asarin, T. Dang, G. Frehse, A. Girard, C. Le Guernic, and O. Maler. Recent progress in continuous and hybrid reachability analysis. In *Proc. 2006 IEEE Conference on Computer Aided Control System Design*, pages 1582–1587, Munich, Germany, 2006.

[11] O. A. Asbjornsen and M. Fjeld. Response modes of continuous stirred tank reactors. *Chem. Eng. Sci.*, 25:1627–1636, 1970.

[12] Uri M. Ascher and Linda R. Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations.* SIAM, Philidelphia, 1998.

[13] Jean-Pierre Aubin. *Viability Theory.* Birkhauser, Boston, MA, 1991.

[14] Jean-Pierre Aubin and Arrigo Cellina. *Differential Inclusions.* Springer, Berlin, 1984.

[15] Vadim Azhmyakov and Jorg Raisch. Convex control systems and convex optimal control. *IEEE Trans. Automat. Contr.*, 53(4):993–998, 2008.

[16] J.R. Banga and W.D. Seider. Global optimization of chemical processes using stochastic algorithms. In C.A. Floudas and P.M. Pardalos, editors, *State of the Art in Global Optimization: Computational Methods and Applications.* Kluwer Academic Publishing, 1996.

[17] P.I. Barton, J.R. Banga, and S. Galan. Optimization of Hybrid discrete /continouos dynamic systems. *Computers and Chemical Engineering*, 4(9/10):2171–2182, 2000.

[18] P.I. Barton and C.C. Pantelides. Modeling of combined discrete /continuous processes. *AIChE Journal*, 40:966–979, 1994.

[19] R. Bellman. *Dynamic Programming.* Princeton University Press, New Jersey, 1957.

[20] A. Bemporad and M. Morari. Verification of hybrid systems via mathematical programming. In *Hybrid Systems: Computation and Control*, volume 1569 of *Lecture Notes in Computer Science*, pages 31–45, 1999.

[21] A. Ben-Israel and T.N.E. Greville. *Generalized Inverses: Theory and Applications.* Springer-Verlag, New York, 2 edition, 2003.

[22] Leonard D. Berkovitz. *Optimal Control Theory.* Springer, New York, 1974.

[23] Dimitri P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, MA, 2 edition, 1999.

[24] M. Berz and K. Makino. Verified integration of ODEs and flows using differential algebraic methods on high-order Taylor models. *Reliable Computing*, 4:361–369, 1998.

[25] T.K. Bhatia and L.T. Biegler. Dynamic optimization in the design and scheduling of multiproduct batch plants. *Ind Eng Chem Res*, 35:2234–2246, 1996.

[26] L. T. Biegler and V. M. Zavala. Large-scale nonlinear programming using IPOPT: An integrating framework for enterprise-wide dynamic optimization. *Comp. and Chem. Eng.*, 33(3):575–582, 2009.

[27] Wojciech Blajer. Index of differential-algebraic equations governing the dynamics of contrained mechanical systems. *Appl. Math. Modelling*, 16:70–77, 1992.

[28] Franco Blanchini. Set invariance in control. *Automatica*, 35:1747–1767, 1999.

[29] H.G. Bock. *Numerical treatment of inverse problems in chemical reaction kinetics*, pages 102–125. Springer Series in Chemical Physics. Springer Verlag, 1981.

[30] A. Bompadre and Alexander Mitsos. Convergence rate of McCormick relaxations. *J. Glob. Optim.*, 52(1):1–28, 2012.

[31] D. Bonvin. Optimal control of batch reactors - a personal view. *J. Proc. Cont.*, 8(5-6):355–368, 1998.

[32] R.G. Brusch and R.H. Schappelle. Solution of highly constrained optimal control problems using nonlinear programming. *AIAA Journal*, 11(2):135–136, 1973.

[33] A. E. Jr. Bryson and Y.-C. Ho. *Applied Optimal Control.* Hemisphere Publishing Corporation, Washington, 1975.

[34] E.F. Carrasco and J.R. Banga. Dynamic optimization of batch reactors using adaptive stochastic algorithms. *Ind Eng Chem Res*, 36(6):2252–2261, 1997.

[35] F. Castiglione and B. Piccoli. Cancer immunotherapy, mathematical modeling and optimal control. *J Theor Biol*, 247:723–732, 2007.

[36] Benoit Chachuat, Adam B. Singer, and Paul I. Barton. Global mixed-integer dynamic optimization. *AIChE Journal*, 51(8):2235–2253, 2005.

[37] Benoit Chachuat, Adam B. Singer, and Paul I. Barton. Global methods for dynamic optimization and mixed-integer dynamic optimization. *Ind. Eng. Chem. Res.*, 45:8373–8392, 2006.

[38] F. L. Chernousko and A. A. Lyubushin. Methods of successive approximations for solution of optimal control problems. *Optimal Control Applications and Methods*, 3:101–114, 1982.

[39] A. Chutinan and B.H. Krogh. Computing polyhedral approximations to flow pipes for dynamic systems. In *Proc. 37th IEEE Conference on Decision and Control*, volume 2, pages 2089–2094, Tampa, FL, Dec. 1998.

[40] A. Chutinan and B.H. Krogh. Verification of polyhedral-invariant hybrid automata using polygonal flow pipe approximations. In *Hybrid Systems: Computation and Control*, volume 1569 of *Lecture Notes in Computer Science*, pages 76–90, 1999.

[41] A. Chutinan and B.H. Krogh. Computational techniques for hybrid system verification. *IEEE Trans. Automat. Contr.*, 48(1):64–75, 2003.

[42] Michal Cizniar, Marian Podmajersky, Tomas Hirmajer, Miroslav Fikar, and Abderrazak M. Latifi. Global optimization for parameter estimation of differential-algebraic systems. *Chemical Papers*, 63(3):274–283, 2009.

[43] Earl A. Coddington and Norman Levinson. *Theory of Ordinary Differential Equations*. McGraw-Hill, New York, 1955.

[44] S. D. Cohen and A. C. Hindmarsh. CVODE, A Stiff/Nonstiff ODE Solver in C. *Computers in Physics*, 10(2):138–143, 1996.

[45] Elizabeth Ann Cross and Ian M. Mitchell. Level set methods for computing reachable sets of systems with differential algebraic equation dynamics. In *Proc. 2008 American Control Conference*, pages 2260–22–65, June 2008.

[46] J. E. Cuthrell and L. T. Biegler. On the optimization of differential-algebraic process systems. *AIChE Journal*, 33(8):1257–1270, 1987.

[47] Bernard Dacorogna. *Introduction to the Calculus of Variations*. Imperial College Press, London, 2004.

[48] V.D. Dimitriadis, N. Shah, and C.C. Pantelides. Modeling and safety verification of discrete /continuous processing systems. *AIChE Journal*, 43(4):1041–1059, 1997.

[49] Michael F. Doherty and Michael F. Malone. *Conceptual Design of Distillation Systems*. McGraw-Hill, New York, 2001.

[50] K. S. Du and R.B. Kearfott. The cluster problem in multivariate global optimization. *J. Glob. Optim.*, 5(3):253–265, 1994.

[51] G. Dunnebier, J. Fricke, and K. U. Klatt. Optimal design and operation of simulated moving bed chromatographic reactors. *Ind Eng Chem Res*, 39(7):2290–2304, 2000.

[52] E. Elbeltagi, T. Hagazy, and D. Grierson. Comparison among five evolutionary-based optimization algorithms. *Advanced Engineering Informatics*, 19(1):43–53, 2005.

[53] Richard J. Elliott and Carl T. Lira. *Introductory Chemical Engineering Thermodynamics*. Prentice-Hall, Upper Saddle River, NJ, 1999.

[54] William R. Esposito and Christodoulos A. Floudas. Deterministic global optimization in nonlinear optimal control problems. *J. Glob. Optim.*, 17:97–126, 2000.

[55] William R. Esposito and Christodoulos A. Floudas. Global optimization for the parameter estimation of differential-algebraic systems. *Ind. Eng. Chem. Res.*, 39:1291–1310, 2000.

[56] F. Facchinei and J. S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*, volume 1. Springer, New York, 2003.

[57] James E. Falk and Richard M. Soland. An algorithm for separable nonconvex programming problems. *Management Science*, 15(9):550–569, 1969.

[58] Iman Famili and Bernhard O. Palsson. The convex basis of the left null space of the stoichiometric matrix leads to the definition of metabolically meaningful pools. *Biophysical Journal*, 85:16–26, 2003.

[59] W.F. Feehery, J.E. Tolsma, and P.I. Barton. Efficient sensitivity analysis of large-scale differential-algebraic systems. *Applied Numerical Mathematics*, 25(1):41–54, 1997.

[60] Martin Feinberg and Friedrich J. M. Horn. Chemical mechanism structure and the coincidence of the stoichiometric and kinetic subspaces. *Arch. Rational Mech. Anal.*, 66(1):83–97, 1977.

[61] M. Fikar, M.A. Latifi, and Y. Creff. Optimal changeover profiles for an industrial depropanizer. *Chemical Engineering Science*, 54(13):2715–2120, 1999.

[62] A. F. Filippov. *Differential Equations with Discontinuous Righthand Sides.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988.

[63] M. Fjeld, O. A. Asbjornsen, and K. J. Astrom. Reaction invariants and their importance in the analysis of eigenvectors, state observability and controllability of the continuous stirred tank reactor. *Chem. Eng. Sci.*, 29:1917–1926, 1974.

[64] Irene Fonseca and Giovanni Leoni. *Modern Methods in the Calculus of Variations: $L^p$ Spaces.* Springer Monographs in Mathematics. Springer, New York, 2007.

[65] D. Fouskakis and D. Draper. Stochastic optimization: A review. *International Statistical Review*, 70(3):315–349, 2002.

[66] Sagar B. Gadewar, Michael F. Doherty, and Michael F. Malone. A systematic method for reaction invariants and mole balances for complex chemistries. *Computers and Chemical Engineering*, 25:1199–1217, 2001.

[67] S. Galan, W.F. Feehery, and P.I. Barton. Parametric sensitivity functions for hybrid discrete /continuous systems. *Applied Numerical Mathematics*, 31(1):17–48, 1999.

[68] Jonathan E. Gayek. A survey of techniques for approximating reachable and controllable sets. In *Proc. 30th IEEE Conference on Decision and Control*, pages 1724–1729, Brighton, England, Dec. 1991.

[69] P. E. Gill, W. Murray, and M. A. Saunders. SNOPT: An SQP algorithm for large-scale constrained optimization. *SIAM J Optim*, 12(4):979–1006, 2002.

[70] A Girard. Reachability of uncertain linear systems using zonotopes. In Manfred Morari and L. Thiele, editors, *Hybrid Systems: Computation and Control*, volume 3414 of *Lecture Notes in Computer Science*, pages 291–305, 2005.

[71] J. L. Gouze, A. Rapaport, and M. Z. Hadj-Sadok. Interval observers for uncertain biological systems. *Ecological Modelling*, 133:45–56, 2000.

[72] M. R. Greenstreet and Ian M. Mitchell. Reachability analysis using polygonal projections. In *Hybrid Systems: Computation and Control*, volume 1569 of *Lecture Notes in Computer Science*, pages 103–116, 1999.

[73] A. Griewank. Automatic directional differentiation of nonsmooth composite functions. In R. Durier and C. Michelot, editors, *Recent Developments in Optimization*, volume 429 of *Lecture Notes in Economics and Mathematical Systems*, pages 155–169, 1995.

[74] Andreas Griewank. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation.* SIAM, 3600 University City Science Center, Philadelphia, PA, 2000.

[75] G. W. Harrison. Dynamic models with uncertain parameters. In X.J.R. Avula, editor, *Proc. of the First International Conference on Mathematical Modeling*, volume 1, pages 295–304, 1977.

[76] G. W. Harrison. Compartmental models with uncertain flow rates. *Mathematical Biosciences*, 43:131–139, 1979.

[77] R. F. Hartl, S. P. Sethi, and R. G. Vickson. A survey of the maximum-principles for optimal-control problems with state constraints. *SIAM Review*, 37(2):181–218, 1995.

[78] Philip Hartman. *Ordinary differential equations.* SIAM, Philidelphia, PA, second edition, 2002.

[79] T. A. Henzinger, P.-H. Ho, and H. Wong-Toi. Algorithmic analysis of nonlinear hybrid systems. *IEEE Transactions on Automatic Control*, 43(4):540–554, 1998.

[80] T. A. Henzinger, B. Horowitz, R. Majumdar, and H. Wong-Toi. Beyond HyTech: Hybrid systems analysis using interval numerical methods. In *Hybrid Systems: Computation and Control*, volume 1790 of *Lecture Notes in Computer Science*, pages 130–144, 2000.

[81] Magnus R. Hestenes. *Calculus of Variations and Optimal Control Theory.* John Wiley and Sons, Inc., New York, 1966.

[82] A. C. Hindmarsh, P. N. Brown, K. E. Grant, S. L. Lee, R. Serban, D. E. Shumaker, and C. S. Woodward. SUNDIALS, suite of nonlinear and differential/algebraic equation solvers. *ACM Transactions on Mathematical Software*, 31:363–396, 2005.

[83] Jens Hoefkens, Martin Berz, and Kyoto Makino. Computing validated solutions of implicit differential equations. *Advances in Computational Mathematics*, 19:231–253, 2003.

[84] Reiner Horst and Hoang Tuy. *Global Optimization: Deterministic Approaches.* Springer, New York, third edition, 1996.

[85] Haitao Huang, C. S. Adjiman, and Nilay Shah. Quantitative framework for reliable safety analysis. *AIChE Journal*, 48(1):78–96, 2002.

[86] Satoru Iwata and Mizuyo Takamatsu. Index minimization of differential-algebraic equations in hybrid analysis for circuit simulation. *Math, Program., Ser. A*, 121:105–121, 2010.

[87] D.H. Jacobson, M.M. Lele, and J.L. Speyer. New necessary conditions of optimality for control problems with state-variable inequality constraints. *AIAA Journal*, 6(8):1488–1491, 1968.

[88] L. Jaulin. Nonlinear bounded-error state estimation of continuous-time systems. *Automatica*, 38:1079–1082, 2002.

[89] Tomasz Kapela and Piotr Zgliczynski. A Lohner-type algorithm for control systems and ordinary differential inclusions. *Discrete and Continuous Dynamic Systems - Series B*, 11(2):365–385, 2009.

[90] R.B. Kearfott. *Rigorous Global Search: Continuous Problems.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[91] K. Hassan Khalil. *Nonlinear Systems.* Prentice Hall, Upper Saddle River, NJ, third edition, 2002.

[92] K. A. Khan and P. I. Barton. Evaluating an element of the Clarke generalized Jacobian of a piecewise differentiable function. In *Proc. 6th International Conference on Automatic Differentiation*, page In Press., 2012.

[93] M. Kieffer, E. Walter, and I. Simeonov. Guaranteed nonlinear parameter estimation for continuous-time dynamical models. In *Proc. 14th IFAC Symposium on System Identification*, 2006.

[94] K. U. Klatt and W. Marquardt. Perspectives for process systems engineering - personal views from academia and industry. *Computers and Chemical Engineering*, 33(3):536–550, 2009.

[95] D. Ko, R. Siriwardane, and L. T. Biegler. Optimization of pressure-swing adsorption process using zeolite 13X for CO2 sequestration. *Ind. Eng. Chem. Res.*, 42(2):339–348, 2003.

[96] Peter Kunkel and Volker Mehrmann. *Differential-Algebraic Equations: Analysis and Numerical Solution.* European Mathematical Society, Zurich, Switzerland, 2006.

[97] Alexander B. Kurzhanski and Pravin Varaiya. Ellipsoidal techniques for reachability analysis. In *Hybrid Systems: Computation and Control*, volume 1790 of *Lecture Notes in Computer Science*, pages 202–214, 2000.

[98] Alexander B. Kurzhanski and Pravin Varaiya. Dynamic optimization for reachability problems. *J. Optim. Theory Appl.*, 108(2):227–251, 2001.

[99] Alexander B. Kurzhanski and Pravin Varaiya. On verification of controlled hybrid dynamics through ellipsoidal techniques. In *Proc. 44th IEEE Conference on Decision and Control*, pages 4682–4687, Seville, Spain, Dec. 2005.

[100] C. Lavor, L. Liberti, N. Maculan, and M. A. C. Nascimeto. Solving Hartree-Fock systems with global optimization methods. *Europhysics Letters*, 77(5):50006, 2007.

[101] D.B. Leineweber, I. Bauer, and H.G. Bock. An efficient multiple shooting based reduced SQP strategy for large-scale dynamic process optimization. Part 1: theoretical aspects. *Comp. and Chem. Eng.*, 27(2):157–166, 2003.

[102] D. Limon, J. M. Bravo, T. Alamo, and E. F. Camacho. Robust MPC of constrained nonlinear systems based on interval arithmetic. *IEEE Proc.-Control Theory Appl.*, 152(3):325–332, 2005.

[103] Youdong Lin and Mark A. Stadtherr. Deterministic global optimization for parameter estimation of dynamic systems. *Ind. Eng. Chem. Res.*, 45:8438–8448, 2006.

[104] Youdong Lin and Mark A. Stadtherr. Deterministic global optimization of nonlinear dynamic systems. *AIChE Journal*, 53(4):866–875, 2007.

[105] Youdong Lin and Mark A. Stadtherr. Validated Solutions of initial value problems for parametric ODEs. *Applied Numerical Mathematics*, 57:1145–1162, 2007.

[106] Youdong Lin and Mark A. Stadtherr. Fault detection in nonlinear continuous-time systems with uncertain parameters. *AIChE Journal*, 54(9):2335–2345, 2008.

[107] L. Luksan and J. Vlcek. Algorithm 811: NDA: algorithms for nondifferentiable optimization. *ACM Transactions on Mathematical Software*, 27(2):193–213, 2001.

[108] R. Luus, J. Dittrich, and F.J. Keil. Multiplicity of solutions in the optimization of a bifunctional catalyst blend in a tubular reactor. *Canadian Journal of Chemical Engineering*, 70:780–785, 1992.

[109] Rein Luus. *Iterative Dynamic Programming.* Chapman and Hall /CRC, Boca Raton, 2000.

[110] John Lygeros, Claire Tomlin, and Shankar Sastry. Controllers for reachability specifications for hybrid systems. *Automatica*, 35:349–370, 1999.

[111] D. L. Ma, S. H. Chung, and R. D. Braatz. Worst-case performance analysis of optimal batch control trajectories. *AIChE Journal*, 45(7):1496–1476, 1999.

[112] M. M. Makela. Survey of bundle methods for nonsmooth optimization. *Optimization Methods and Software*, 17(1):1–29, 2002.

[113] K. Makino and M. Berz. *Remainder differential algebras and their applications.* Computational Differentiation: Techniques, Applications, and Tools. SIAM, Philidelphia, PA, 1996.

[114] T. Maly and L.R. Petzold. Numerical methods and software for sensitivity analysis of differential-algabraic systems. *Applied Numerical Mathematics*, 20:57–79, 1996.

[115] O. L. Mangasarian and T.-H. Shiau. Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems. *SIAM J. Control Optim.*, 25(3):583–595, 1987.

[116] R.B. Martin. Optimal control drug scheduling of cancer chemotherapy. *Automatica*, 28(6):1113–1123, 1992.

[117] S. E. Mattsson. On modeling and differential algebraic systems. *Simulation*, 52(1):24–32, 1989.

[118] Garth P. McCormick. Computability of global solutions to factorable nonconvex programs: Part I - convex underestimating problems. *Math. Program.*, 10:147–175, 1976.

[119] Nelson Merentes. On the Composition Operator in $\mathcal{AC}[a, b]$. *Collect. Math.*, 42(3):237–243, 1991.

[120] Ian Mitchell, Alexandre M. Bayen, and Claire Tomlin. A time-dependent Hamilton-Jacobi formulation of reachable sets for continuous dynamic games. *IEEE Trans. Automat. Contr.*, 50(7):947–957, July 2005.

[121] A. Mitsos, G. M. Bollas, and P. I. Barton. Bilevel optimization formulation for parameter estimation in liquid-liquid phase equilibrium problems. *Chem. Eng. Sci.*, 64(3):548–559, 2009.

[122] Alexander Mitsos, Benoit Chachuat, and Paul I. Barton. McCormick-Based Relaxations of Algorithms. *SIAM J. on Optim.*, 20(2):573–601, 2009.

[123] Alexander Mitsos, Benoit Chachuat, and Paul I. Barton. Towards global bilevel dynamic optimization. *J. Glob. Optim.*, 45(1):63–93, 2009.

[124] C. G. Moles, P. Mendes, and J. R. Banga. Parameter estimation in biochemical pathways: A comparison of global optimization methods. *Genome Research*, 13(11):2467–2474, 2003.

[125] R. E. Moore. *Methods and Applications of Interval Analysis*. SIAM, Philadelphia, PA, 1979.

[126] Max Muller. Uber das Fundamentaltheorem in der Theorie der gewohnlichen Differentialgleichungen. *Math. Zeit.*, 26:619–645, 1926.

[127] James R. Munkres. *Analysis on Manifolds*. Westview Press, Cambridge, MA, 1991.

[128] N. S. Nedialkov. Some recent advances in validated methods for IVPs for ODEs. *Applied Numerical Mathematics*, 42:269–284, 2002.

[129] N. S. Nedialkov, K. R. Jackson, and G. F. Corliss. Validated solutions of initial value problems for ordinary differential equations. *Appl. Math. Comput.*, 105:21–68, 1999.

[130] M. Neher, K. R. Jackson, and N. S. Nedialkov. On Taylor Model Based Integration of ODEs. *SIAM J. on Numer. Anal.*, 45(1):236–262, 2007.

[131] A. Neumaier. *Interval Methods for Systems of Equations*. Cambridge University Press, Cambridge, 1990.

[132] Meeko Oishi, Ian Mitchell, Claire Tomlin, and Patrick Saint-Pierre. Computing viable sets and reachable sets to design feedback linearizing control laws under saturation. In *Proc. 45th IEEE Conference on Decision and Control*, pages 3801–3807, San Diego, CA, Dec. 2006.

[133] A. I. Panasyuk. Equations of attainable set dynamics, part 1: Integral funnel equations. *J. Optim. Theory Appl.*, 64(2):349–366, 1990.

[134] C.C. Pantelides, D. Gritsis, K. R. Morison, and R. W. H. Sargent. The mathematical modelling of transient systems using differential-algebraic equations. *Comp. and Chem. Eng.*, 12(5):449–454, 1988.

[135] I. Papamichail and C. S. Adjiman. A rigorous global optimization algorithm for problems with ordinary differential equations. *J. Glob. Optim.*, 24(1):1–33, 2002.

[136] T. Park and P.I. Barton. State event location in differential-algebraic models. *ACM Transactions on Modeling and Computer Simulation*, 6(2):137–165, 1996.

[137] B. T. Polyak. Convexity of the reachable sets of nonlinear systems under $L_2$ bounded controls. *Dynamics of Continuous, Discrete and Impulsive Systems Series A: Mathematical Analysis*, 11:255–267, 2004.

[138] T. Raissi, N. Ramdani, and Y. Candau. Set membership state and parameter estimation for systems described by nonlinear differential equations. *Automatica*, 40:1771–1777, 2004.

[139] S. V. Rakovic, A.R. Teel, D. Q. Mayne, and A. Astolfi. Simple robust control invariant tubes for some classes of nonlinear discrete time systems. Proceedings of the 45th IEEE Conference on Descision and Control, pages 6397–6402, 2006.

[140] N. Ramdani, N. Meslem, and Y. Candau. A hybrid bounding method for computing an over-approximation for the reachable set of uncertain nonlinear systems. *IEEE Trans. Automat. Contr.*, 54(10):2352–2364, 2009.

[141] A. Rapaport and D. Dochain. Interval observers for biochemical processes with uncertain kinetics and inputs. *Mathematical Biosciences*, 193:235–253, 2005.

[142] Andreas Rauh, Michael Brill, and Clemens Gunther. A novel interval arithmetic approach for solving differential-algebraic equations with Valencia-IVP. *Int. J. Appl. Math. Comput. Sci.*, 19(3):381–397, 2009.

[143] G. Reissig. Convexity of reachable sets of nonlinear ordinary differential equations. *Automation and Remote Control*, 68(9):64–78, 2007.

[144] D.W.T. Rippin. Simulation of single and multiproduct batch chemical plants for optimal design and operation. *Computers and Chemical Engineering*, 7:137–156, 1983.

[145] Walter Rudin. *Principles of Mathematical Analysis.* McGraw-Hill, New York, third edition, 1964.

[146] H.S. Ryoo and N.V. Sahinidis. Global optimization of nonconvex NLPs and MINLPs with application in process design. *Computers and Chemical Engineering*, 19(5):551–566, 1995.

[147] H.S. Ryoo and N.V. Sahinidis. A branch-and-reduce approach to global optimization. *J Glob Optim*, 2:107–139, 1996.

[148] N.V. Sahinidis and Mohit Tawarmalani. Accelerating branch-and-bound through a modeling language construct for relaxation-specific constraints. *J. Glob. Optim.*, 32(2):259–280, 2005.

[149] N.V. Sahinidis and Mohit Tawarmalani. A polyhedral branch-and-cut approach to global optimization. *Math. Program.*, 130(2):225–249, 2005.

[150] Ali M. Sahlodin and Benoit Chachuat. Convex/concave relaxations of parametric ODEs using Taylor models. *Comp. and Chem. Eng.*, 35:844–857, 2011.

[151] Ali M. Sahlodin and Benoit Chachuat. Discretize-then-relax approach for convex/concave relaxations of the solutions of parametric ODEs. *Applied Numerical Mathematics*, 61:803–820, 2011.

[152] J. Schaber, W. Liebermeister, and E. Klipp. Nested uncertainties in biochemical models. *IET Syst. Biol.*, 3(1):1–9, 2009.

[153] S. Scholtes. Introduction to piecewise differentiable equations, 1994. Habilitation Thesis, Institut für Statistik und Mathematische Wirtschaftstheorie, University of Karlsruhe.

[154] Karl Schugerl. Progress in monitoring, modeling and control of bioprocesses during the last 20 years. *J. Biotech.*, 85:149–173, 2001.

[155] D. A. Schwer, J. E. Tolsma, W. H. Green, and P. I. Barton. On upgrading the numerical in combustion chemistry codes. *Combustion and Flame*, 128(3):270–291, 2002.

[156] Joseph K. Scott and Paul I. Barton. Tight, efficient bounds on the solutions of chemical kinetics models. *Computers and Chemical Engineering*, 34:717–731, 2010.

[157] Joseph K. Scott, Benoit Chachuat, and Paul I. Barton. Nonlinear convex and concave relaxations for the solutions of parametric ODEs. *In Press: Optimal Control Applications and Methods*, 2012.

[158] Joseph K. Scott, Matthew D. Stuber, and Paul I. Barton. Generalized McCormick relaxations. *J. Glob. Optim.*, 51:569–606, 2011.

[159] Atle Seierstad and Knut Sydstaeter. Sufficient conditions in optimal control theory. *International Economic Review*, 18(2):367–391, 1977.

[160] Ajay Selot, Loi Kwong Kuok, Mark Robinson, Thomas Mason, and Paul I. Barton. A short-term operational planning model for natural gas production systems. *AIChE Journal*, 54(2):495–515, 2007.

[161] Adam B. Singer and Paul I. Barton. Global solution of optimization problems with parameter-embedded linear dynamic systems. *J. Optim. Theory Appl.*, 121:613–646, 2004.

[162] Adam B. Singer and Paul I. Barton. Bounding the solutions of parameter dependent nonlinear ordinary differential equations. *SIAM J. Sci. Comput.*, 27:2167–2182, 2006.

[163] Adam B. Singer and Paul I. Barton. Global dynamic optimization for parameter estimation in chemical kinetics. *J. Phys. Chem. A*, 110(3):971–976, 2006.

[164] Adam B. Singer and Paul I. Barton. Global optimization with nonlinear ordinary differential equations. *J. Glob. Optim.*, 34:159–190, 2006.

[165] Hal L. Smith. *Monotone Dynamical Systems: An Introduction to the Theory of Competitive and Cooperative Systems*, volume 41 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1995.

[166] A. Sorribas, C. Pozo, E. Vilaprinyo, and G. Guillen-Gosalbez. Optimization and evolution in metabolic pathways: Global optimization techniques in generalized mass action models. *J. Biotech.*, 149:141–153, 2010.

[167] B. Srinivasan, S. Palanki, and D. Bonvin. Dynamic optimization of batch processes - I. characterization of the nominal solution. *Comp. and Chem. Eng.*, 27(1):1–26, 2003.

[168] Gilbert Strang. *Linear Algebra and its Applications*. Thomson Brooks/Cole, Belmont, CA, 4 edition, 2006.

[169] M. D. Stuber, J. K. Scott, and P. I. Barton. Global optimization of implicit functions. *Submitted*, 2011.

[170] Jacek Szarski. *Differential Inequalities*. Polish Scientific Publishers, Warszawa, Poland, 1965.

[171] Mohit Tawarmalani and Nikolaos V. Sahinidis. *Convexification and Global Optimization in Continuous and Mixed-Integer Nonlinear Programming*. Kluwer Academic Publishers, 2002.

[172] J. W. Taylor, G. Ehlker, H.-H. Carstensen, L. Ruslen, R. W. Field, and W. H. Green. Direct measurement of the fast, reversible addition of oxygen to cyclohexadienyl radicals in nonpolar solvents. *J. Phys. Chem. A*, 108:7193–7203, 2004.

[173] K. L. Teo, G. Goh, and K. Wong. *A Unified Computational Approach to Optimal Control Problems*. John Wiley and Sons, Inc., New York, 1991.

[174] Jared Toettcher, Anya Catillo, Bruce Tidor, and Jacob White. Biochemical oscillator sensitivity analysis in the presence of conservation constraints. In *Proc. 48th ACM/IEEE/EDAC Design Automation Conference*, 2011.

[175] J.E. Tolsma and P.I. Barton. DAEPACK an open modeling environment for legacy models. *Industrial and Engineering Chemistry Research*, 39(6):1826–1839, 2000.

[176] Claire Tomlin, Ian Mitchell, Alexandre M. Bayen, and Meeko Oishi. Computational techniques for the verification of hybrid systems. *Proceedings of the IEEE*, 91(7):986–1001, 2003.

[177] John L. Troutman. *Variational Calculus and Optimal Control: Optimzation with Elementary Convexity*. Springer-Verlag, New York, second edition, 1996.

[178] T.H. Tsang, D.M. Himmelblau, and T.F. Edgar. Optimal control via collocation and nonlinear programming. *International Journal of Control*, 21:763–768, 1975.

[179] Pravin P. Varaiya, Felix F. Wu, and Janusz W. Bialek. Smart operation of smart grid: Risk limiting dispatch. *Proc. of the IEEE*, 99(1):40–57, 2011.

[180] Dale E. Varberg. On absolutely continuous functions. *Amer. Math. Monthly*, 72:831–941, 1965.

[181] Kurt V. Waller and Pertti M. Makila. Chemical reaction invariants and variants and their use in reactor modeling, simulation, and control. *Ind. Eng. Chem. Proc. Des. Dev.*, 20:1–11, 1981.

[182] W. Walter. *Differential and Integral Inequalities.* Springer-Verlag, New York, 1970.

[183] J. Warga. *Optimal Control of Differential and Functional Equations.* Academic Press, Inc., New York, 1972.

[184] H. Yazarel and G. J. Pappas. Geometric programming relaxations for linear system reachability. In *Proc. American Control Conference (2004)*, volume 1, pages 553–559, Boston, MA, July 2004.

[185] Mehmet Yunt. *Nonsmooth Dynamic Optimization of Systems with Varying Structure.* PhD thesis, MIT, 2011.