Nonsmooth Dynamic Optimization of Systems with Varying Structure

by

Mehmet Yunt

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

Doctor of Philosophy in Mechanical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2011

Author .....
Department of Mechanical Engineering
January 8, 2011

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

Certified by .....
Kamal Youcef-Toumi
Professor of Mechanical Engineering
Thesis Committee Chairman

Accepted by .....
David E. Hardt
Chairman, Department Committee on Graduate Theses

Nonsmooth Dynamic Optimization of Systems with Varying Structure

by

Mehmet Yunt

Submitted to the Department of Mechanical Engineering
on January 8, 2011, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Mechanical Engineering

Abstract

In this thesis, an open-loop numerical dynamic optimization method for a class of dynamic systems is developed. The structure of the governing equations of the systems under consideration change depending on the values of the states, parameters and the controls. Therefore, these systems are called systems with varying structure. Such systems occur frequently in the models of electric and hydraulic circuits, chemical processes, biological networks and machinery. As a result, the determination of parameters and controls resulting in the optimal performance of these systems has been an important research topic. Unlike dynamic optimization problems where the structure of the underlying system is constant, the dynamic optimization of systems with varying structure requires the determination of the optimal evolution of the system structure in time in addition to optimal parameters and controls. The underlying varying structure results in nonsmooth and discontinuous optimization problems.

The nonsmooth single shooting method introduced in this thesis uses concepts from nonsmooth analysis and nonsmooth optimization to solve dynamic optimization problems involving systems with varying structure whose dynamics can be described by locally Lipschitz continuous ordinary or differential-algebraic equations. The method converts the infinite-dimensional dynamic optimization problem into an nonlinear program by parameterizing the controls. Unlike the state of the art, the method does not enumerate possible structures explicitly in the optimization and it does not depend on the discretization of the dynamics. Instead, it uses a special integration algorithm to compute state trajectories and derivative information. As a result, the method produces more accurate solutions to problems where the underlying dynamics is highly nonlinear and/or stiff for less effort than the state of the art.

The thesis develops substitutes for the gradient and the Jacobian of a function in case these quantities do not exist. These substitutes are set-valued maps and an elements of these maps need to be computed for optimization purposes. Differential equations are derived whose solutions furnish the necessary elements. These differential equations have discontinuities in time. A numerical method for their solution is proposed based on state event location algorithms that detects these discontinuities. Necessary conditions of optimality

for nonlinear programs are derived using these substitutes and it is shown that nonsmooth optimization methods called bundle methods can be used to obtain solutions satisfying these necessary conditions. Case studies compare the method to the state of the art and investigate its complexity empirically.

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

Acknowledgments

This thesis is the summation of my toils and ingenuity, the academic guidance of my thesis committee, the financial support of my sponsors and the help of my family and friends.

First and foremost, I would like thank Professor Paul I. Barton for the support and freedom to work in a field of my own choosing. His critical and constructive feedback made me a better engineer and a competent researcher. I would like to thank Professor Kamal Youcef-Toumi for being the chairman of my committee; Professor Pieter J. Mosterman for the insightful discussions on the modeling of systems with varying structure; Professor Jonathan How and Professor Jean-Jacques Slotine for their suggestions and different viewpoints about my research.

I would like to thank past and present members of the Process Systems Engineering Laboratory for their help in matters of research and Leslie Regan for all her help in administrative matters.

I would like to thank the DoD Multidisciplinary University Research Initiative (MURI) Program, and BP for their financial support.

I could not have finished my studies without the support of my family. There are no words that can express my gratitude to my sisters, Zeynep and Elif. They interrupted their own onerous doctoral studies without hesitation to literally get me back on my feet after I suffered a spinal chord injury. They have always lent me their sympathetic ears and provided unconditional help. I am blessed to have them. I am grateful to my parents, Havva and Mustafa, for their love, encouragement and unfaltering support throughout the years. I would not have ever made it to MIT if it were not for the sacrifices they made to provide a world-class education for their children.

I am indebted to my long-time friend Keith Eisenhauer for reminding me that MIT and a PhD are neither necessary nor sufficient conditions for a meaningful life, for teaching me to smoke a briar pipe properly, for exposing me to various Americana, for his insight into human affairs and for his assistance during my illness. I am thankful to Dr. Derya Özyurt, and Professor Hande Özdinler for their friendship and their help to me and my sisters. Dr. Ajay Selot, my friend, my former colleague and former house-mate, deserves all the best for helping me out in my research and private life during the hardest times of my stay at MIT. Our discussions on topics ranging from politics to food have been a welcome daily respite from the drab graduate student experience. The house of Kıvanc Bas, another long-time friend, was a peaceful haven from the stressful life at MIT. I thank him and his family for welcoming me as a friend. I will always remember the Turkish Thanksgiving dinners, board gaming sessions and conversations. It took me six years to appreciate the wisdom of some of Kıvanc's advice. The last time I practiced hands-on nuts and bolts mechanical engineering was to build an underwater autonomous vehicle for Project ORCA. I am grateful to Dimitrios Tzeranis for making that a memorable experience and for being a dependable friend. Finally, I would like to thank Shuyin Sim for being my private top-tier B-level business and non-business management consultant.

This thesis is dedicated to my family and close friends.

Contents

1 Introduction **21**

1.1 Dynamic Systems with Varying Structure 21

1.2 Dynamic Optimization of Systems with Varying Structure 25

1.3 Nonsmooth Optimization of Dynamic Systems with Varying Structure 30

1.4 Overview 32

2 Preliminaries **39**

2.1 Notation 39

2.2 The Gâteaux, Partial, Fréchet and Strict Derivatives 40

2.2.1 Properties 43

2.2.2 Mean Value Theorem for Differentiable Functions 44

2.3 Set-valued Maps 44

2.3.1 Properties of Set-valued Maps 46

2.4 Elementary Convex Analysis 47

2.4.1 Convex Sets 47

2.4.2 Convex Functions 48

2.5 Locally Lipschitz Continuous Functions 49

2.5.1 Properties of Locally Lipschitz Continuous Functions 50

2.6 Nonsmooth Analysis for Locally Lipschitz Continuous Functions 51

2.6.1 The Generalized Directional Derivative 52
2.6.2 The Generalized Gradient 52
2.6.3 The Generalized Jacobian 56
2.6.4 The Partial Generalized Gradient and Jacobian 58
2.6.5 Implicit Function Theorem for Locally Lipschitz Continuous Functions 59
2.7 Piecewise Continuously Differentiable (PC^1) Functions 60
2.7.1 Properties of PC^1 Functions 60
2.7.2 Implicit Function Theorem for PC^1 Functions 60
2.8 Semismooth Functions 61
2.8.1 Bouligand Differentiable Functions 61
2.8.2 Scalar-Valued Semismooth Functions 63
2.8.3 Vector-valued Semismooth Functions 65
2.8.4 A Restricted Definition of Semismoothness 66
2.8.5 The Linear Newton Approximation 68
2.9 Examples 71

3 Parametric Sensitivity Analysis of Dynamic Systems using the Generalized Jacobian **87**

3.1 Preliminaries 90
3.1.1 Note on Notation and Assumptions 91
3.2 Ordinary Differential Equations 92
3.3 Differential-Algebraic Equations 101
3.4 Multistage Systems 108
3.5 Comments on the Numerical Computation of Forward and Adjoint Sensitivities 117
3.6 Examples 118

4 Parametric Sensitivity Analysis of Dynamic Systems using Linear Newton Approximations **125**

4.1 Preliminaries 126

4.1.1 Note on Notation and Assumptions 126

4.2 Ordinary Differential Equations 134

4.3 Differential-Algebraic Equations 149

4.4 Multistage Systems 156

4.5 Example 167

5 Numerical Computation of the Generalized Jacobians and Linear Newton Approximations **171**

5.1 Linear Newton Approximations of the Right-Hand Side 172

5.2 Singleton and Non-Singleton Trajectories 176

5.3 Computational Improvements 178

5.4 State Event Location and Detection of Non-singleton Trajectories 185

5.4.1 The State Event Location and Non-singleton Trajectory Detection Algorithm 188

5.5 Computational Methods Used in this Thesis 193

6 Bundle Method using Linear Newton Approximations **195**

6.1 Necessary Conditions of Optimality for Lipschitzian Optimization 196

6.2 Necessary Conditions of Optimality using Linear Newton Approximations . 200

6.3 Bundle Method using Linear Newton Approximations 214

6.3.1 Formal Statement of the Bundle Method with Linear Newton Approximations 224

6.3.2 Discussion and Convergence 229

6.3.3 Variants and Commentary 234

7 Nonsmooth Dynamic Optimization of Systems with Varying Structure 237

7.1 The Nonsmooth Single Shooting Method 238

7.1.1 Formulation and Assumptions 238

7.1.2 Computation of the elements of the linear Newton approximations of Program (7.1.2) 240

7.1.3 Description of the Method 242

7.2 Control Parameterization 243

7.2.1 Preliminaries 244

7.2.2 Description of Control Parameterization 246

7.3 Minimum Time Problems 252

7.4 Dynamic Optimization Literature Review and Comparison with the State of the Art 255

8 Case Studies 259

8.1 Electric Circuit 259

8.1.1 System Dynamics 259

8.1.2 Dynamic Optimization Formulation 262

8.1.3 Nonsmooth Single Shooting Method Results 265

8.1.4 MPEC Approach Results 266

8.1.5 A Variant Dynamic Optimization Formulation 272

8.1.6 Conclusion 277

8.2 Cascading Tanks: Empirical Complexity Analysis 279

8.2.1 System Dynamics 279

8.2.2 Dynamic Optimization Formulation 282

8.2.3 Scaling of the Nonsmooth Single Shooting Method with Respect to Number of Tanks and Number of Epochs 283

8.2.4 Conclusion 288
8.3 Cascading Tank: Comparison with the MPEC Approach 289
8.3.1 Conclusion 294
8.4 Cascaded Tank Example: Comparison with the MILP approach 294
8.4.1 MILP Formulation 295
8.4.2 MILP Approach Results 298
8.4.3 Conclusion 301
8.5 Tank Change Over 303
8.5.1 System Dynamics 304
8.5.2 Safe Operation Conditions for the Tank 307
8.5.3 Dynamic Optimization Formulation 308
8.5.4 Nonsmooth Single Shooting Method Results 310
8.5.5 Conclusion 315
8.6 Chemotherapy Scheduling Case Study 315
8.6.1 System Dynamics 315
8.6.2 Dynamic Optimization Formulation 318
8.6.3 Nonsmooth Single Shooting Method Results 319
8.6.4 Derivative Free Method Results 320
8.6.5 Conclusion 322
8.7 Notes 323

9 Conclusions and Future Directions for Research **325**

9.1 Future Directions for Research 329
9.1.1 Parametric Sensitivities, Their Computation and Use 329
9.1.2 Dynamic Optimization 331
9.1.3 Systems with Discontinuous Vector Fields 333

A Convergence Proof of the Modified Bundle Method **339**

A.1 Convergence Proof 339

Bibliography **357**

List of Figures

2-1 (Example 2.9.11) Plot of $f(x_1, x_2) = |x_1^2 - \sin(|x_2|)|$ and of its contours. . . . 81

2-2 (Example 2.9.12) Plot of f. 83

5-1 Examples of singleton and non-singleton trajectories. 177

7-1 Nonsmooth Single Shooting Method. 243

8-1 Electric Circuit: Configuration. 260

8-2 Electric Circuit: Voltage and current trajectories for different constant I_c. . . 263

8-3 Electric Circuit: Difference in voltage and current trajectories between FitzHugh-Nagumo and Electric Circuit Models if $I_c(t, \mathbf{p}) = 0.50\text{A}$, $\forall t \in [t_0, t_f]$. 263

8-4 Electric Circuit:Surface and contour plots of the dissipated energy by the diode264

8-5 Electric Circuit: Effect of nonzero complementarity deviations on voltage and current, $\Delta = 0.4304$ 270

8-6 Electric Circuit: Difference between MPEC predicted voltage and current trajectories and simulation for Case E and $n_t = 201$. 274

8-7 Electric Circuit: Difference between MPEC predicted voltage and current trajectories and simulation for Case E and $n_t = 401$. 274

8-8 Electric Circuit: Difference between MPEC predicted voltage and current trajectories and simulation for Case E and $n_t = 701$. 275

8-9 Electric Circuit: Difference between MPEC predicted voltage trajectories and simulation for Case E and $n_t = 201$ and $n_t = 801$. 277

8-10 Cascaded Tanks Configuration. 280

8-11 Cascading Tanks: Optimal valve openings for $n_T = 3.0$ and $n_e = 10$. 284

8-12 Cascading Tanks: Plot of optimal state trajectories for $n_T = 3.0$ and $n_e = 10$. 285

8-13 Cascading Tanks: Plot of $\bar{\tau}$ versus $\bar{n}_T$ and $\bar{\tau}$ versus $\bar{n}_e$ for the relaxed tolerance case. 288

8-14 Cascading Tanks: Plot of $\bar{\tau}$ versus $\bar{n}_T$ and $\bar{\tau}$ versus $\bar{n}_e$ tight tolerance case. . 288

8-15 Cascading Tanks: Plot of CPU time versus n_T and n_e for the relaxed tolerance case. 289

8-16 Cascading Tanks: Plot of CPU time versus n_T and n_e for the tight tolerance case. 289

8-17 Cascading Tanks: MPEC Approach vs. The Single Nonsmooth Shooting Method for different numbers of tanks. 293

8-18 Cascading Tanks: MPEC Approach vs. The Nonsmooth Single Shooting Method for different numbers of epochs. 294

8-19 Cascading Tanks: MILP Approach CPU times and optimal objective values for different numbers of epochs and $n_a = 1$, $n_b = 1$. 299

8-20 Cascading Tanks: MILP Approach CPU times and optimal objective values for different numbers of epochs and $n_a = 2$, $n_b = 2$. 299

8-21 Cascading Tanks: MILP Approach CPU times and optimal objective values for different numbers of epochs and $n_a = 3$, $n_b = 3$. 300

8-22 Cascading Tanks: MILP Approach CPU times and optimal objective values for different numbers of tanks and $n_a = 1$, $n_b = 1$. 300

8-23 Cascading Tanks: MILP Approach CPU times and optimal objective values for different numbers of tanks and $n_a = 2$, $n_b = 2$. 301

8-24 Cascading Tanks: MILP Approach CPU times and optimal objective values for different numbers of tanks and $n_a = 3$, $n_b = 3$. 301
8-25 Cascading Tanks: State trajectories for $n_T = 3$, $n_e = 60$, $n_a = 1$, $n_b = 1$. . . . 302
8-26 Cascading Tanks: State trajectories for $n_T = 3$, $n_e = 60$, $n_a = 2$, $n_b = 2$. . . . 302
8-27 Cascading Tanks: State trajectories for $n_T = 3$, $n_e = 60$, $n_a = 3$, $n_b = 3$. . . . 303
8-28 Cascading Tanks: Comparison of objective values and CPU times between the MILP approach and the nonsmooth single shooting method for varying epochs, $n_T = 3$, $n_a = 3$, $n_b = 3$. 303
8-29 Cascading Tanks: Comparison of objective values and CPU times between the MILP approach and the nonsmooth single shooting method for varying numbers of tanks, $n_e = 10$, $n_a = 3$, $n_b = 3$. 304
8-30 Tank Change Over: Configuration. 304
8-31 Tank Change Over: Final mole fraction profiles corresponding to the solution of (8.5.8) for $n_e = 3$. 312
8-32 Tank Change Over: Initial mole fraction profiles corresponding to parameters in Table 8.21 of program (8.5.8) for $n_e = 3$. 313
8-33 Tank Change Over: Plot of the path constraint and the mole fraction profiles corresponding to the solution of (8.5.8) for $n_e = 3$. 313
8-34 Chemotherapy Scheduling: Configuration 315
8-35 Chemotherapy Schedule: Optimal drug schedules. 320
8-36 Chemotherapy Schedule: Drug concentrations in the tissue. 321
8-37 Chemotherapy Schedule: Cell populations. 321

9-1 Transition times as a function of p. 334
9-2 $-x(3.0, p)$ as a function of p. 335
9-3 $\ln |J_p x(3.0, p)|$ as a function of p. 336

List of Tables

2.1 (Example 2.9.11) The selection functions of $|x_1^2 - \sin(|x_2|)|$. 82

2.2 (Example 2.9.11) $\nabla f(x_1, x_2)$ at points where it is defined. 82

3.1 Solution and Generalized Gradients of Example 3.6.1 119

8.1 Electric Circuit: Nonsmooth single shooting results 265

8.2 Electric Circuit: MPEC feasibility problem results, $I_{c,k} = 0.50, \forall k \in \mathcal{K}$. . . . 269

8.3 Electric Circuit: MPEC feasibility problem results, $I_{c,k} = 1.50, \forall k \in \mathcal{K}$. . . . 269

8.4 Electric Circuit: MPEC feasibility problem results, $I_{c,k} = 1.00$, initialization with simulation results. 270

8.5 Electric Circuit: MPEC method optimization results, $\mu = 5$, termination tolerance is 3.0×10^{-13} 272

8.6 Electric Circuit: MPEC method optimization results, $\mu = 1$, termination tolerance is 3.0×10^{-13} 272

8.7 Electric Circuit: MPEC method optimization results, $\mu = 5$, termination tolerance is 1.0×10^{-7} 273

8.8 Electric Circuit: MPEC method optimization results for various values of n_t for Case E, $\mu = 5$, termination tolerance is 1.0×10^{-7} 273

8.9 Electric Circuit: Nonsmooth single shooting results for variant problem. . . . 276

8.10 Electric Circuit: MPEC method optimization results for the variant dynamic optimization problem for various values of n_t for Case E, $\mu = 5$, termination tolerance is 1.0×10^{-7}. 277
8.11 Cascading Tanks: Simulation and optimization tolerances. 284
8.12 Cascading Tanks: Optimization run data for $n_e = 10$ and different numbers of tanks. 286
8.13 Cascading Tanks: Optimization run data for $n_T = 3$ and different numbers of epochs 287
8.14 Cascading Tanks: Curve fitting results for the natural logarithm of the normalized CPU times versus natural logarithm of the normalized number of states. 290
8.15 Cascading Tanks: Curve fitting results for the natural logarithm of the normalized CPU times versus natural logarithm of the normalized number of epochs. 290
8.16 Cascading Tanks: Curve fitting results for the CPU times versus number of epochs. 291
8.17 Cascading Tanks: Curve fitting results for the CPU times versus number of tanks. 291
8.18 Tank Change Over: Model parameter values. 307
8.19 Tank Change Over: Path constraint polynomial coefficients. 308
8.20 Tank Change Over: Solution of program (8.5.8) for $n_e = 3$. 312
8.21 Tank Change Over: Initial parameter values used to solve program (8.5.8) for $n_e = 3$. 312
8.22 Tank Change Over Case Study: Objective and constraint values of program (8.5.8) for $n_e = 3$. 312
8.23 Tank Change Over: Solution of program (8.5.8) for $n_e = 4$. 314

8.24 Tank Change Over: Initial parameter values used to solve program (8.5.8) with $n_e = 4$. 314
8.25 Tank Change Over: Objective and constraint values of program (8.5.8) with $n_e = 4$. 314
8.26 Chemotherapy Scheduling: Model Parameters 318
8.27 Chemotherapy Schedule:Cell populations at the beginning and end of treatment 320
8.28 Chemotherapy Schedule: Comparison of Nonsmooth Single Shooting Method and Derivative Free Method 322

Chapter 1

Introduction

A new dynamic optimization method is developed in this thesis that can be applied to a class of dynamic systems whose structure changes depending on the state, parameters, and controls. First, an overview of these systems and applicable dynamic optimization methods is presented. Then, the approach of this thesis is explained. Finally, the contents of the subsequent chapters are summarized.

1.1 Dynamic Systems with Varying Structure

A deterministic dynamic system whose governing equations change instantaneously when the system's states, parameters, and time satisfy certain conditions is called a *system with varying structure* in this thesis. Such systems occur frequently in the mathematical description of electrical circuits, hydraulic systems, machinery, chemical process plants, and biological networks. For example, the constitutive equations of a check valve change depending on the upstream and downstream pressures. If the downstream pressure is larger than the upstream pressure, there is no flow through the valve. Otherwise, the flow rate through the valve is a function of the pressure difference. Diodes in electric circuits display analogous behavior.

A mechanical example is backlash in gears. In this case, the governing equations change abruptly when gear teeth lose or gain contact. These systems also form when computer implemented logic rules and controllers are used to govern the behavior of physical systems.[1]

Two components make up the state of these systems. One component comprises *the continuous states* which evolve according to differential or difference equations in subsets of the Euclidean spaces. The other component comprises *the discrete states*, which evolve in discrete sets such as the integers.

Various modeling paradigms have been proposed to describe systems with varying structure. Hybrid automata [3, 38], hybrid bond graphs [73], state-transition networks [7], complementarity systems [99], differential variational inequalities [80], the unified framework for hybrid control [20], mixed logical dynamical systems [13], differential automata [104] and switching systems [62] are some of the paradigms available in the literature. Note that some of these paradigms can be analyzed using the theory of ordinary differential equations [26] or differential inclusions [5, 37].

In this thesis, the nomenclature of the hybrid system paradigm in [57, 56] is adopted for the general discussion of systems with varying structure. In this paradigm, a set of governing equations is called a *mode*. Each mode is associated with an integer index. The mode comprises ordinary differential equations or differential-algebraic equations that govern the evolution of the continuous states. The *hybrid mode trajectory* is a sequence of these indices in strict time order of the modes encountered during the evolution of the system. It is the trajectory of the discrete state of the system. *The length of the hybrid mode trajectory* is the number of elements in the hybrid mode trajectory. Given an element of this sequence, its *predecessor mode* is the mode immediately before and the *successor mode* is the mode immediately after in the hybrid mode trajectory. Corresponding to the hybrid mode trajectory, there is a *hybrid time trajectory* which contains subsets of the time interval

[1]For a comprehensive collection of examples see [32].

called *epochs*. For each element of the hybrid mode trajectory, there exists an epoch in the hybrid time trajectory. A mode is *active* and governs the evolution of the system during its corresponding epochs in the hybrid time trajectory. The value of the discrete state changes when a *transition* occurs. A transition occurs at the earliest time at which the corresponding logical condition called a *transition condition* on the states, parameters and controls of the system is satisfied. A transition is *explicit* if the timing, predecessor and successor modes are known before the state trajectories are computed. Otherwise, a transition is *implicit*. A transition is *autonomous* if the corresponding logical conditions of the transition depend on the states of the system. A transition is *controlled* if the transition occurs in response to a control input. *Transition functions* determine the successor mode and initial conditions of the continuous states for the evolution in the next epoch.

The mechanics of mode switching complicates the analysis of systems with varying structure. It is possible to observe situations where the discrete state changes infinitely often at a given value of time, preventing the further evolution of the continuous states. This phenomenon is called *deadlock*. It is also possible that the number of transitions eventually becomes infinite while time remains finite. This phenomenon is called *Zeno behavior*. Examples and detailed discussion of these behaviors can be found [38, 116, 52, 75]. The existence of solutions and continuous dependence of the solutions on initial conditions can be proven for special cases [65, 80, 110, 37, 26].

The discrete state helps determine the mode of the system. In some systems with varying structure, if the continuous states are known, then the active mode can be determined without knowledge of the discrete state. Example (1.1.1) describes a system where a discrete state is necessary and Example (1.1.2) contains a system where it is not. The accurate computation of the continuous state trajectories evolving according to continuous-time dynamics requires the detection of instants when the active mode changes. At these instants, the vector fields are possibly discontinuous and continuous-time integration algorithms either

fail or become inefficient when trying to integrate over these discontinuities while satisfying integration error tolerances. Special integration methods and numerical codes have been developed for this purpose. See [11, 74, 80] for further discussion.

Example 1.1.1 (Pressure Relief Valve). A pressure relief valve is used to reduce the pressure inside a vessel, P, to an acceptable maximum value. The relief valve opens if the pressure inside the vessel is higher than P_h and closes when the pressure inside the vessel is less than P_l. $P_h > P_l$ holds in order to prevent the relief valve from opening and closing too rapidly and wearing out unnecessarily. In this system, if the pressure of the vessel is such that $P_l < P < P_h$ holds, then it cannot be determined whether the valve is open or closed without the aid of the discrete state. The reason for this ambiguity is the fact that P can satisfy this condition irrespective of the state of the valve.

Example 1.1.2 (Tank with Outlet Flow). Consider a tank with an inlet flow at the bottom and an outlet flow at height H. Assume that there is a check valve at the inlet and the outlet flow discharges to the atmosphere through a valve. Let h be the liquid level in the tank. Then, the dynamics of the system can be written as

$$F_{in}(t,P) = \begin{cases} 0 & \text{if } P - h(t,P) \leq 0, \\ C_v \frac{P-h(t,P)}{\sqrt{|P-h(t,P)|+k_b}} & \text{otherwise,} \end{cases}$$

$$F_{out}(t,P) = \begin{cases} 0 & \text{if } h(t,P) \leq H, \\ C_v \frac{h(t,P)-H}{\sqrt{|h(t,P)-H|+k_b}} & \text{otherwise,} \end{cases}$$

$$\dot{h}(t,P) = (F_{in}(t,P) - F_{out}(t,P))/A, \forall t \in (t_0, t_f],\ h(t_0,P) = 0,$$

where P is the inlet pressure measured in the height of liquid, F_{in} is the inlet flow rate, F_{out} is the outlet flow rate, A is the cross section of the tank, C_v is the valve coefficient and k_b is a small regularization constant. The quantity, $z/(\sqrt{|z|+k_b})$ approximates $\sqrt{z}$ if $z \gg k_b$ and

$z/\sqrt{k_b}$ if $z \approx 0$ for $z \geq 0$. Unlike the square root function, it is continuous and differentiable at zero. This regularization is necessary to avoid theoretical and numerical issues caused by the behavior of the square root function at zero.

Suppose $h(t^*, P) = H$. A unique $\dot{h}(t^*, P)$ can still be computed because the system with varying structure is equivalent to an ordinary differential equation with a continuous right-hand side at this point.

1.2 Dynamic Optimization of Systems with Varying Structure

Systems with varying structure are ubiquitous in economically important engineering systems. Therefore, the development of dynamic optimization methods to determine optimal performance of these systems has been the subject of research for some time. Open loop and closed loop dynamic optimization methods have been developed. In this thesis, open loop methods are of primary interest. In this section, these methods are reviewed.

The dynamic optimization methods for systems with varying structure need to determine the optimal hybrid mode trajectory and the corresponding hybrid time trajectory in order to determine the optimal continuous state trajectories and controls. In general, the hybrid mode and time trajectories depend on the controls. This dependence causes nonsmoothness and discontinuous behavior [8, 84]. Therefore, standard dynamic optimization methods [17, 21] that assume continuous differentiability cannot be directly applied to these problems. Derivatives of the states with respect to the parameters of the system in the form of parametric forward sensitivities [44] and parametric adjoint sensitivities [24] may not always exist. An instance of nonsmoothness can be found in Example 1.2.1.

Example 1.2.1 (Nonsmooth Control Example). Consider the dynamic system

$$\dot{y}(t,p) = \begin{cases} -y(t,p) + u(t,p) & \text{if } y \leq 0 \\ y(t,p) + u(t,p) & \text{if } y > 0 \end{cases}, \ \forall t \in (0, t_f],$$

$$u(t,p) = \begin{cases} \frac{-y_0 e^{-t^*}}{1-e^{-t^*}} & \text{if } t \in [0, t^*) \\ p & \text{if } t \in [t^*, t_f] \end{cases},$$

$$y(0,p) = y_0, \ y_0 < 0, \ p \in \mathbb{R}.$$

For the given initial condition and $u(t,p)$, $y(t^*,p) = 0$ holds. The choice of dynamics at $t = t^*$, depends on the value of p. If $p = 0$, then $y(t,p) = 0$ holds for all $t \in [t^*, t_f]$. If $p < 0$, then $y(t,p) = -pe^{-(t-t^*)} + p = p(1 - e^{-(t-t^*)})$. If $p > 0$, then $y(t,p) = p(e^{(t-t^*)} - 1)$ for all $t \in [t^*, t^f]$. As a result, $y(t_f,p)$ is

$$y(t_f,p) = \begin{cases} p\left(1 - e^{-(t_f - t^*)}\right) & \text{if } p < 0, \\ 0 & \text{if } p = 0, \\ p\left(e^{(t_f - t^*)} - 1\right) & \text{if } p > 0. \end{cases}$$

Note that $y(t_f, \cdot)$ is continuous but not differentiable at $p = 0$.

Maximum principles [86, 103, 101] analogous to Pontryagin's Maximum Principle [21] have been developed for the case where the hybrid mode trajectory does not depend on the controls. Necessary conditions of optimality for special cases have been developed that allow the hybrid mode trajectory to vary [25, 111]. These apply only if the dynamics of the system can be expressed as ordinary differential equations satisfying a Lipschitz condition or a differential inclusion of certain structure. These conditions use elements of nonsmooth analysis [25, 92] which extend the concept of the derivative to nonsmooth functions. Except

for the conditions in [101], these conditions currently do not appear to be amenable to numerical computation to solve dynamic optimization problems.

In [101, 23], two-stage approaches are discussed where in the first stage the necessary conditions are used to solve a dynamic optimization problem for a constant hybrid mode trajectory. The first-stage resembles multistage problems considered in [21] and is a continuously differentiable problem. In the second stage, the hybrid mode trajectory is altered while the number of elements in the trajectory is kept constant. The first stage is repeated using the updated hybrid mode trajectory. The entire two-stage process is repeated until all possible hybrid mode trajectories are processed. This method is a combinatorial algorithm. If the number of elements in the hybrid mode trajectory is n_e and the number of possible modes is n_m, the algorithm processes $(n_m)^{n_e}$ mode sequences. Similar two-stage approaches have been proposed in [114] and [41].

Sufficient conditions for the existence of parametric forward sensitivities and parametric adjoint sensitivities are given in [94, 39, 95]. In addition to the constant hybrid mode trajectory requirement, a transversality condition is required to hold at each transition and only one transition can occur at any given time. It is also shown that the states are continuously differentiable functions of the parameters in this case [39]. Computation of parametric forward sensitivities derived in [39] can be carried out with the integration algorithms in [108, 36].

The most common approach to enumerate hybrid mode trajectory candidates of fixed length is to use integer variables. Systems whose modes consist of discrete-time linear systems are considered in [13] and systems whose modes consist of continuous-time linear systems are considered, for example, in [102, 106]. In these approaches, nonlinear dynamics are linearized. Continuous-time dynamics are discretized using a scheme such as the forward Euler method. Note that each linearization increases the number of candidate modes, n_m. These methods divide the time horizon into n_t subintervals. Hence, they consider hybrid mode

trajectories of length n_t. The sizes of these subintervals correspond to integration steps in case of continuous-time linear dynamics. For each mode candidate, a binary variable is used at each subinterval to keep track if a mode is active during that subinterval. The binary variable's value is one if the mode is active and zero otherwise. Additional constraints prevent more than one mode to be active on a subinterval. The final formulation is a *mixed-integer linear program*, (MILP), and can be solved to global optimality. Solvers for MILP problems enumerate candidate hybrid mode trajectories implicitly by taking advantage of the linear structure of the mathematical program [16]. In the case of continuous-time dynamics, the MILP solver also acts as the integration algorithm. The main drawback of this method in the case of continuous-time dynamics is the approximation error in the computed state trajectories caused by linearization and discretization. Note that features that enable integration algorithms to provide accurate solutions such as adjusting integration time steps to satisfy error tolerances cannot be implemented. In order to reduce this approximation error, relatively large n_t and n_m values have to be used. This adversely effects the solution times of the solver. The solution times of MILP problems scale worst-case exponentially with n_t and n_m [106]. Hence, this approach is not very suitable for problems with nonlinear continuous-time dynamics and nonlinear constraints. An example of this behavior can be seen in the Cascading Tank Case Study in Chapter 8. An attempt to handle nonlinear continuous-time dynamics is made in [7] without using linearization. The resultant formulation is a *mixed-integer nonlinear program,* (MINLP). This approach is not practical because current MINLP solvers cannot handle problems of the size obtained by this approach. Note that the MILP approach is used in closed-loop applications as well.

In [57, 56, 55] systems with varying structure capable of only explicit transitions are considered. Continuous-time linear systems constitute the modes. Integer variables are used, but the continuous-time linear dynamics are not discretized. The number of transitions is constant. Auxiliary dynamic systems are constructed to underestimate the objective

value. Using the parametric sensitivity results in [39], special integration algorithms, and the auxiliary dynamic systems, these nonconvex problems are solved deterministically to $\epsilon-$ global optimality. This approach does not suffer from approximation errors; however, it is currently limited to problems with a few states and parameters and explicit transitions.

An alternative approach to alter the hybrid mode trajectory is to formulate *mathematical programs with equilibrium constraints,* (MPEC) [12, 90]. Real-valued variables that satisfy special constraints called *complementarity conditions* are used. The time horizon is partitioned into n_t subintervals called finite elements. On each finite element, the continuous-time dynamics are discretized using Radau collocation. The active mode on each finite element is determined by the complementarity conditions at the ends of these subintervals. Since complementarity conditions violate optimization regularity conditions called *constraint qualifications*, special methods are required to solve these problems. In this approach, the nonlinear programming solver acts as the integration algorithm as well. This results in less accurate computation of the state trajectories. A relatively large value for n_t needs to be used if the underlying dynamics are nonlinear or stiff. As a result, this approach produces very large optimization problems mandating large-scale optimization solvers. The nonlinear nonconvex programs are solved to obtain stationary points. In this approach it is not clear what the value of n_t should be to obtain accurate solutions. Even though there are convergence results for the use of Radau collocation on finite elements when the dynamics are sufficiently continuously differentiable [53], it is an open question whether these results apply when used in conjunction with complementarity constraints. The Electrical Circuit Case Study in Chapter 8 illustrates this problem. The determination of a value for n_t and convergence is an issue in the MILP approach for continuous-time dynamics as well.

Numerical optimization methods that do not explicitly use derivative information such as derivative-free methods [6], genetic algorithms or stochastic methods [38] can be applied to the solution of dynamic optimization problems involving systems with varying structure, [38].

Most of these methods are heuristic and require more effort to provide a solution compared to derivative-based methods on problems where derivative information is available.

The MILP, MPEC and derivative-free approaches do not solve for the necessary conditions of optimality. Either these are not available or they are not amenable to numerical computation. They directly try to minimize the objective. Hence, they are direct dynamic optimization methods. In addition, in all these methods, the controls are parameterized.

Finally, a dynamic programming based approach can be found in [45]. This approach suffers from the curse of dimensionality and is not suitable for problems with more than three or four states.

1.3 Nonsmooth Optimization of Dynamic Systems with Varying Structure

In this thesis, an open loop dynamic optimization method for a class of dynamic systems with varying structure is developed that does not discretize the continuous-time dynamics as part of the optimization formulation and that does not enumerate candidate hybrid mode trajectories. The method does not assume any a priori information about the hybrid mode trajectory except that the length should be finite.

The method is applicable to problems where the controls are real-valued and the dynamics of the system with varying structure can be reduced to an ordinary differential equation satisfying local Lipschitz continuity and piecewise continuous differentiability[2] assumptions. Instances of such systems can be found in Example 1.1.2 and Chapter 8. Note that for this class of systems, crucial properties such as the existence and uniqueness of solutions and continuous dependence on initial conditions are established by classical theory [26].

The method is in the class of direct single shooting methods (see [17] for a classification

[2]In this thesis, piecewise continuously differentiable functions are continuous functions.

of numerical methods) and it is called *the nonsmooth single shooting method.* Instead of discretizing the dynamics and using an optimization algorithm to solve for the continuous state trajectories, a specialized and efficient numerical integration algorithm [108] is used to compute trajectories accurate within integration tolerances. Therefore, the method is a single shooting method. In this approach, the real-valued controls are approximated by functions depending on finitely many parameters. This enables the conversion of the dynamic optimization problem into a nonlinear program (NLP) as in [105, 40]. This approach allows the handling of path and point constraints in a unified manner. The resultant NLP is a nonsmooth mathematical program. Therefore, concepts from nonsmooth analysis [25, 92, 35] are used in place of the gradient where it does not exist and nonsmooth optimization [54, 66] methods are applied to solve the resultant NLP.

The resultant basic nonsmooth NLP formulation is

$$\begin{aligned}
\min_{\mathbf{p}\in\mathcal{P}} \; & J(\mathbf{p}) = \int_{t_0}^{t_f} h_0(t,\mathbf{p},\mathbf{u}(t,\mathbf{p}),\mathbf{x}(t,\mathbf{p}))dt + H_0(t_f,\mathbf{p},\mathbf{u}(t_f,\mathbf{p}),\mathbf{x}(t_f,\mathbf{p})) \qquad (1.3.1)\\
& \text{s.t.} \int_{t_0}^{t_f} h_i(t,\mathbf{p},\mathbf{u}(t,\mathbf{p}),\mathbf{x}(t,\mathbf{p}))dt + H_i(t_f,\mathbf{p},\mathbf{u}(t_f,\mathbf{p}),\mathbf{x}(t_f,\mathbf{p})) \leq 0,\ \forall i \in \{1,\ldots,n_c\},\\
& \dot{\mathbf{x}}(t,\mathbf{p}) = \mathbf{f}(t,\mathbf{p},\mathbf{u}(t,\mathbf{p}),\mathbf{x}(t,\mathbf{p})),\ \forall t \in (t_0,t_f],\\
& \mathbf{x}(t_0,\mathbf{p}) = \mathbf{f}_0(\mathbf{p}),
\end{aligned}$$

where $\mathbf{p}$ are the real-valued parameters; n_c is a finite positive integer, $\mathbf{u}$ are the controls; $\mathbf{x}$ are the continuous states; $\mathbf{f}$, $\mathbf{f}_0$, h_i and H_i are piecewise continuously differentiable functions for all $i \in \{0,\ldots,n_c\}$. Note that in this approach, inequality path constraints of the form $g(t,\mathbf{p},\mathbf{u}(t,\mathbf{p}),\mathbf{x}(t,\mathbf{p})) \leq 0,\ \forall t \in [t_0,t_f]$ are handled by converting them into the following point constraints,

$$\int_{t_0}^{t_f} \max(0, g(t,\mathbf{p},\mathbf{u}(t,\mathbf{p}),\mathbf{x}(t,\mathbf{p}))dt \leq 0$$

or

$$\int_{t_0}^{t_f} \max(0, g(t, \mathbf{p}, \mathbf{u}(t, \mathbf{p}), \mathbf{x}(t, \mathbf{p}))^2) dt \leq 0.$$

The approach developed in this thesis can handle multistage problems where at each stage the dynamics of the system are governed by disparate vector fields that satisfy the piecewise continuous differentiability requirement. In addition, the dynamics can be governed by certain classes of differential-algebraic equations. In the remainder of this thesis, the controls are omitted from the formulations of the dynamics and mathematical programs since they are functions of the parameters and time only.

1.4 Overview

The contents of the subsequent chapters of this thesis are:

Chapter 2: This chapter is a review of nonsmooth analysis necessary for the theoretical developments in this thesis. The *generalized gradient* [25] of a function and *the linear Newton approximation* [35] of a function are used at points where the gradient of the function does not exist. Unlike the gradient, these entities are set-valued maps. For instance, it can be shown that $\partial_p y(t_f, p)$, the generalized gradient of the function $y(t_f, \cdot)$ defined in Example 1.2.1 at p is,

$$\partial_p y(t_f, p) = \begin{cases} \left(1 - e^{-(t_f - t^*)}\right) & \text{if } p < 0, \\ \left[\left(1 - e^{-(t_f - t^*)}\right), \left(e^{(t_f - t^*)} - 1\right)\right] & \text{if } p = 0, \\ \left(e^{(t_f - t^*)} - 1\right) & \text{if } p > 0. \end{cases}$$

The importance of the generalized gradient stems from the fact that it can be used to formulate necessary conditions of optimality and determine descent directions in

numerical optimization methods for nonsmooth problems [25, 54]. The generalized Jacobian and gradient coincide with the usual Jacobian and gradient for continuously differentiable functions. The generalized gradient of a function is unique. The linear Newton approximation, on the other hand, is not unique. It represents a class of set-valued maps that contain the generalized gradient or generalized Jacobian. Its main use to date has been in the solution of nonsmooth algebraic equations. In this thesis, it is used to replace the generalized gradient/generalized Jacobian when these quantities cannot be computed for an optimization algorithm.

Specifically, Chapter 2 contains a review of derivatives, elementary set-valued and convex analysis. It contains the definitions, basic properties and calculus rules of the linear Newton approximation and generalized Jacobian/gradient. In addition, the classes of functions of interest are introduced. Implicit function theorems for these functions are stated. The chapter ends with demonstrative examples.

Chapter 3: The numerical solution of (1.3.1) requires that an element of the generalized Jacobians and linear Newton approximations of the objective and constraint functions be computable for each parameter value. In order to apply the calculus rules of the generalized Jacobian and the linear Newton approximation, an element of the generalized Jacobian or linear Newton approximation of the map $\boldsymbol{\eta} \mapsto \mathbf{x}(t, \boldsymbol{\eta})$ at $\mathbf{p}$ is required. However, the explicit form of this map is generally not known. Computing numerically an element of the generalized Jacobian or linear Newton approximation of the map $\boldsymbol{\eta} \mapsto \mathbf{x}(t, \boldsymbol{\eta})$ at $\mathbf{p}$ is the main challenge of this thesis.

In this chapter, sufficient conditions for the existence of the gradient of the map $\boldsymbol{\eta} \mapsto \mathbf{x}(t, \boldsymbol{\eta})$ are derived using the generalized Jacobian and results from [25]. The functions involved in (1.3.1) are assumed to be *locally Lipschitz continuous*. These sufficient conditions result in trajectories along which $\partial_{\mathbf{p}}\mathbf{x}(t, \mathbf{p})$, the generalized Jacobian of the map $\boldsymbol{\eta} \mapsto \mathbf{x}(t, \boldsymbol{\eta})$ at $\mathbf{p}$, is a singleton set. Loosely, the key condition is that the

trajectory $(\mathbf{p}, \mathbf{x}(\cdot, \mathbf{p}))$ visit the points of nondifferentiability in the domain of $\mathbf{f}$ only at times that constitute a measure zero subset of the time horizon $[t_0, t_f]$.

Forward and adjoint parametric sensitivity differential equations are derived. These differential equations resemble results in [44, 24] and [39, 95]. However, unlike results in [44, 24], the right-hand sides of these differential equations comprise functions that are discontinuous in time and their solutions exist in the sense of Carathéodory. Unlike results in [39, 95], invariance of the hybrid mode trajectory and transversality at each transition are not required. Also multiple transitions can occur at one time.

The results are extended to differential-algebraic equations using nonsmooth implicit function theorems and to multistage systems. The chapter ends with demonstrative examples.

Chapter 4: This chapter considers trajectories where the assumptions of Chapter 3 do not hold. In this case, one must consider differential inclusions. The solutions of these differential inclusions define sets which may or may not contain the desired generalized Jacobian information. Restricting the functions involved in (1.3.1) to the class of *semismooth* functions, sharper results are obtained using results from [81]. Note that semismooth functions include piecewise continuously differentiable functions. In this case, a linear Newton approximation can be defined whose value at a point contains the value of the generalized Jacobian of the map $\boldsymbol{\eta} \mapsto \mathbf{x}(t, \boldsymbol{\eta})$. The results of this chapter reduce to the results in Chapter 3 if the assumptions of that chapter in addition to the semismoothness assumption hold.

Elements of the linear Newton approximations can be computed using integration forward in time as in forward parametric sensitivities or using integration backwards in time as in adjoint parametric sensitivities. The results are extended to certain differential-algebraic systems using an implicit function theorem for semismooth functions. The extension to multistage systems are derived. Finally, a demonstrative

example is presented.

Chapter 5: A computational method to obtain an element of the linear Newton approximations defined in Chapter 4 is described in this chapter. The differential equations defining elements of the linear Newton approximations are possibly discontinuous at times when the state trajectory passes through points of nondifferentiability in the domain of the right-hand sides. In order to detect these discontinuities, a structural assumption is made that in essence makes all functions in (1.3.1) piecewise continuously differentiable functions. The structural assumption places the points of nondifferentiability on the boundaries of open sets which can be represented by the zero-level sets of certain functions. These functions are used in conjunction with *state event location algorithms* [83] to determine time points at which discontinuities occur. The structural assumption also allows the use of more efficient methods to compute an element of the linear Newton approximations. An implementation with available software to compute simultaneously the states and an element of the associated linear Newton approximations using integration forward in time is presented.

Chapter 6: It is known for some time that numerical algorithms for continuously differentiable optimization problems can get stuck at arbitrary nondifferentiable points or experience numerical difficulties when applied to nonsmooth optimization problems [61]. Furthermore, the stationarity conditions for continuously differentiable optimization problems do not hold for nonsmooth problems. Therefore special methods are required.

Bundle methods are nonsmooth optimization algorithms that use the generalized gradients of the objective and constraint functions to compute stationary points of nonconvex programs [54, 66]. The stationarity conditions are formulated using the generalized gradients of the objective and constraint functions and reduce to the well-known Karush-Kuhn-Tucker (KKT) conditions if the objective and constraints are continu-

ously differentiable [25, 54].

Bundle methods use a set of generalized gradients obtained at nearby points to compute a direction of descent and a specialized line search algorithm to construct this set efficiently. The set of generalized gradients is called the *the bundle.* The use of the extra information furnished by the bundle prevents these methods from getting stuck at arbitrary nondifferentiable points. In terms of convergence, bundle methods produce a sequence of iterates whose limit points are stationary.

In this chapter, extended stationary conditions are formulated using the linear Newton approximations defined in Chapter 4. This formulation is possible because the values of these linear Newton approximations contain the values of the generalized gradients. It is shown that using these linear Newton approximations instead of the generalized gradients results in a nonsmooth optimization algorithm that produces a sequence of iterates whose limits points satisfy the extended stationarity conditions. In essence, one set-valued map is replaced with another that has similar properties. The bundle method is formally stated.

The key result in the convergence proof is the finite termination of the specialized line search algorithm. The rest of the proof is the same as the proofs in [54] with the linear Newton approximation replacing the generalized gradient. Therefore, only a summary of the proof is placed in the Appendix.

Chapter 7: The nonsmooth single shooting method is formally developed in this chapter. The results of the previous chapters are used to assemble the method.

Convergence of the approximate controls to the solution of the original dynamic optimization problem is discussed. Using the results in [105, 40], it can be shown that if the optimal approximate controls convergence to a function as the number of parameters increases, then that function is an optimal control of the original dynamic optimization problem. Similarly, if the optimal objective values corresponding to the approximate

controls converge, the limit is the optimal objective value of the original problem.

The section also contains a technique to solve minimum time problems. The sensitivity results of the previous chapters deal with parameters of the dynamic system only. Minimum time problems can be solved by transforming time into a state variable. Then, the initial time and the difference between the final and initial times become parameters of the transformed system.

Chapter 8: Case studies are collected in this chapter. The performance of the MILP, MPEC and derivative-free methods are compared to the performance of the nonsmooth single shooting method. It is shown that the nonsmooth single shooting method provides more accurate solutions to problems involving systems whose dynamics are highly nonlinear and exhibit stiffness for less effort. An empirical complexity analysis is carried out. The results strongly suggest that the nonsmooth single shooting method scales polynomially with the number of states and number of parameters.

Chapter 9: The contributions of this thesis are summarized in this chapter. The main contribution of this thesis is the development of the nonsmooth single shooting method. The novelty of this method stems from the fact that explicit discretization of the dynamics in the optimization formulation and enumeration of the hybrid mode trajectory are not used. The parametric sensitivity results in Chapters 3 and 4 are new. The use of a bundle method in conjunction with linear Newton approximations is new. The detailed comparison of the MPEC approach, MILP approach and nonsmooth single shooting method is new. This is the first comparison that considers accuracy of the solutions in addition to the effort to obtain solutions. The empirical complexity analysis of the nonsmooth single shooting method is new.

The chapter also discusses possible future directions of research.

Chapter 2

Preliminaries

This chapter provides a brief summary of the necessary mathematical background for the developments presented in this thesis. The chapter focuses on results in nonsmooth analysis for locally Lipschitz continuous functions. These results depend on derivatives, convex sets, convex functions and set-valued maps. Therefore the chapter begins with a briéf review of results in differentiation, set-valued maps and convex analysis. The chapter concludes with examples to illustrate some of the reviewed concepts.

2.1 Notation

In this document, symbols printed in boldface represent vector and matrix-valued quantities.

Let O and S be sets in a metric space. $O \backslash S$ is $\{u \in O : u \notin S\}$. $\text{int}\,(O)$ is the interior and $\text{cl}\,(O)$ is the closure of O. $\text{bd}\,(O)$ is the boundary of O and it is equal to $\text{cl}\,(O) \backslash O$. O is a *singleton* if it has exactly one element.

Let n be a finite positive integer. $S \subset \mathbb{R}^n$ is a set of measure zero in $\mathbb{R}^n$ if it has Lebesgue measure zero.

Let O be the set $\{o_i\}_{i=1}^n$ where n is a positive integer (possibly equal to ∞) and o_i are

elements of an arbitrary set. Then, $ind(o_i, O)$, the *index of* o_i, is i and $s(O)$, the number of elements in O, is n. If $n = \infty$ then $\{o_i\}_{i=1}^{n}$ is equivalent to $\{o_i\}$.

Let $O \subset \mathbb{R}^{n\times m}$. If $\mathbf{A} \in O$, then $A_{i,j}$ represents the element occupying the ith row and and jth column of $\mathbf{A}$ where $i \in \{1,\ldots,n\}$ and $j \in \{1,\ldots,m\}$. $\mathbf{A}^{\mathrm{T}}$ is the transpose of $\mathbf{A}$. $\|\mathbf{A}\|$, the norm of $\mathbf{A}$, is $\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}|A_{i,j}|^2}$. Let $\mathbf{A} \in \mathbb{R}^{n\times m}$ and $\mathbf{B} \in \mathbb{R}^{m\times p}$, then $\mathbf{AB}$ is the $n \times p$ matrix that is the product of $\mathbf{A}$ and $\mathbf{B}$. $[\mathbf{A}^{\mathrm{T}}\ \mathbf{B}]$ is an $m \times (n+p)$ matrix such that for all $i = 1,\ldots,m$, $C_{i,j} = A_{i,j}^{\mathrm{T}}$ if $1 \leq j \leq n$ and $C_{i,j} = B_{i,(j-n)}$ if $n < j \leq n+p$. If $\mathbf{A} \in \mathbb{R}^{n\times n}$ is invertible, then $\mathbf{A}^{-1}$ represents the inverse.

The elements of $\mathbb{R}^n$ are column vectors. If $\mathbf{v} \in \mathbb{R}^n$ and $\mathbf{u} \in \mathbb{R}^m$, then $(\mathbf{v}, \mathbf{u})$ is equivalent to $[\mathbf{v}^{\mathrm{T}}\ \mathbf{u}^{\mathrm{T}}]^{\mathrm{T}}$.

$\mathbf{I}_n$ is the $n \times n$ identity matrix. $\mathbf{e}_i$ is the ith column of $\mathbf{I}_n$. $\mathbf{0}$ represents any matrix whose elements are all zero.

Let $Z \subset \mathbb{R}^{n\times m}$ and $\alpha \in \mathbb{R}$. Then αZ represents the set $\{\alpha\mathbf{z} : \mathbf{z} \in Z\}$. Let $Y \subset \mathbb{R}^{n\times m}$. Then $Z + Y$ represents the set $\{\mathbf{z}+\mathbf{y} : \mathbf{z} \in Z,\ \mathbf{y} \in Y\}$.

2.2 The Gâteaux, Partial, Fréchet and Strict Derivatives

The results in this section are from Chapter 3 in [76] unless otherwise stated.

Let m and n be finite positive integers and X be an open subset of $\mathbb{R}^n$. Let $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ denote the space of continuous linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$. Let $f_i : X \to \mathbb{R}$ for $i = 1,\ldots,m$ and $\mathbf{f} : X \to \mathbb{R}^m : \mathbf{y} \mapsto (f_i(\mathbf{y}),\ldots,f_m(\mathbf{y}))$. Let $\mathbf{x} \in X$, $\mathbf{x} = (x_i,\ldots,x_n)$ where $x_i \in \mathbb{R}$.

Definition 2.2.1 (The Directional Derivative). *Let* $\mathbf{v} \in \mathbb{R}^n$. *Then* $\mathbf{f}'(\mathbf{x};\mathbf{v})$, *the direc-*

tional derivative or Gâteaux differential of $\mathbf{f}$ *at* $\mathbf{x}$ *in the direction* $\mathbf{v}$*, is defined by*

$$\mathbf{f}'(\mathbf{x};\mathbf{v}) = \lim_{t \downarrow 0} \frac{\mathbf{f}(\mathbf{x}+t\mathbf{v}) - \mathbf{f}(\mathbf{x})}{t}.$$

$\mathbf{f}$ *is directionally differentiable at* $\mathbf{x}$ *if* $\mathbf{f}'(\mathbf{x};\mathbf{v})$ *exists for all* $\mathbf{v} \in \mathbb{R}^n$*.*

Definition 2.2.2 (The Gâteaux Derivative). *If* $\mathbf{f}'(\mathbf{x};\mathbf{v})$ *exists for all* $\mathbf{v} \in \mathbb{R}^n$ *at* $\mathbf{x}$ *and there exists a continuous linear transformation* $\mathbf{A}(\mathbf{x}) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ *such that* $\mathbf{f}'(\mathbf{x};\mathbf{v}) = \mathbf{A}(\mathbf{x})\mathbf{v}$, $\forall \mathbf{v} \in \mathbb{R}^n$ *then* $\mathbf{f}$ *is Gâteaux differentiable at* $\mathbf{x}$ *and* $\mathbf{A}(\mathbf{x})$ *is the unique Gâteaux derivative of* $\mathbf{f}$ *at* $\mathbf{x}$*.*

Equivalently $\mathbf{f}$ *is Gâteaux differentiable at* $\mathbf{x}$ *if there exists a continuous linear transformation* $\mathbf{A}(\mathbf{x}) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ *such that for any* $\mathbf{v} \in \mathbb{R}^n$*,*

$$\lim_{t \downarrow 0} \frac{\|\mathbf{f}(\mathbf{x}+t\mathbf{v}) - \mathbf{f}(\mathbf{x}) - t\mathbf{A}(\mathbf{x})\mathbf{v}\|}{t} = 0.$$

In the remainder of this document, if $\mathbf{f}$ is Gâteaux differentiable at $\mathbf{x}$, then it will be called differentiable at $\mathbf{x}$.

Definition 2.2.3 (The Jacobian, the Gradient and the Partial Derivative). $\mathbf{Jf}(\mathbf{x})$*, the Jacobian of* $\mathbf{f}$ *at* $\mathbf{x}$*, is an* $m \times n$ *matrix of the form*

$$\begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\mathbf{x}) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial f_m}{\partial x_n}(\mathbf{x}) \end{bmatrix}.$$

where each $\frac{\partial f_i}{\partial x_j}(\mathbf{x})$ *is called the partial derivative of* f_i *with respect to* x_j *at* $\mathbf{x}$ *and satisfies*

$$\frac{\partial f_i}{\partial x_j}(\mathbf{x}) = \lim_{t \downarrow 0} \frac{f_i(\mathbf{x}+t\mathbf{e}_j) - f_i(\mathbf{x})}{t}.$$

If $m = 1$, the gradient of f at $\mathbf{x}$ is $\nabla f(\mathbf{x})$ and it is equal to $\mathbf{J}f(\mathbf{x})^{\mathrm{T}}$.

If $\mathbf{f}$ is Gâteaux differentiable at $\mathbf{x} \in X$, then $\mathbf{A}(\mathbf{x}) = \mathbf{J}\mathbf{f}(\mathbf{x})$.

Let $O \subset X$ be a neighborhood of $\mathbf{x} \in X$. If $\mathbf{f}$ is Gâteaux differentiable at every $\mathbf{z} \in O$ and the mapping $\mathbf{z} \mapsto \mathbf{J}\mathbf{f}(\mathbf{z})$ is continuous on O, then $\mathbf{f}$ is *continuously Gâteaux differentiable* at $\mathbf{x}$. If $O = X$, then $\mathbf{f}$ is a *continuously Gâteaux differentiable function.* This is denoted by $\mathbf{f} \in \mathcal{C}^1(X)$. In the remainder of this document, continuous differentiability is a synonym for continuous Gâteaux differentiability.

Let $\{s_i\}_{i=1}^n$ be a set where each s_i and n are finite positive integers. Let X_i be an open subset of $\mathbb{R}^{s_i}$, $\mathbf{x}_i \in X_i$ for all $i \in \{1, \ldots, n\}$ and $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$. Let $\mathbf{f}$ be a function from $\Pi_{i=1}^n X_i$ to $\mathbb{R}^m$. If $k \in \{1, \ldots, n\}$, then $\mathbf{J}_k\mathbf{f}(\mathbf{x})$ is the Gâteaux derivative of the function $\mathbf{f}(\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}, \cdot, \mathbf{x}_{k+1}, \ldots, \mathbf{x}_n)$ at $\mathbf{x}$. If f is a scalar-valued function, then $\nabla_k f(\mathbf{x})$ is equivalent to $\mathbf{J}_k f(\mathbf{x})^{\mathrm{T}}$.

Definition 2.2.4 (The Fréchet Derivative). $\mathbf{f}$ *is Fréchet differentiable at* $\mathbf{x} \in X$ *if there exists a unique* $\mathbf{A}(\mathbf{x}) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ *such that,*

$$\lim_{\mathbf{v} \to \mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x} + \mathbf{v}) - \mathbf{f}(\mathbf{x}) - \mathbf{A}(\mathbf{x})\mathbf{v}\|}{\|\mathbf{v}\|} = 0.$$

holds. $\mathbf{A}(\mathbf{x})$ *is called the Fréchet derivative of* $\mathbf{f}$ *at* $\mathbf{x}$.

Definition 2.2.5 (The Strict Derivative). $\mathbf{f}$ *is strictly differentiable at* $\mathbf{x} \in X$ *if there exists a unique* $\mathbf{A}(\mathbf{x}) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ *such that,*

$$\lim_{(\mathbf{y}, \mathbf{v}) \to (\mathbf{x}, \mathbf{0})} \frac{\mathbf{f}(\mathbf{y} + \mathbf{v}) - \mathbf{f}(\mathbf{y}) - \mathbf{A}(\mathbf{x})\mathbf{v}}{\|\mathbf{v}\|} = 0.$$

holds. $\mathbf{A}(\mathbf{x})$ *is called the strict derivative of* $\mathbf{f}$ *at* $\mathbf{x}$ *(page 132 in [19]).*

Equivalently, $\mathbf{f}$ *is strictly differentiable at* $\mathbf{x} \in X$ *if there exists a unique* $\mathbf{A}(\mathbf{x}) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$

such that

$$\lim_{(\mathbf{y},\, t)\to(\mathbf{x},0^+)} \frac{\mathbf{f}(\mathbf{y}+t\mathbf{v}) - \mathbf{f}(\mathbf{y}) - t\mathbf{A}(\mathbf{x})\mathbf{v}}{t} = 0$$

holds for all $\mathbf{v} \in \mathbb{R}^n$ *and the convergence is uniform for* $\mathbf{v}$ *in compact sets (page 30 in [25]).*

Example 2.9.6 contains the proof of the equivalence of these two definitions.

2.2.1 Properties

The results in this section are from Chapter 3 in [76] unless otherwise stated.

1. Even if $\mathbf{f}$ is Gâteaux differentiable at $\mathbf{x} \in X$, it may not be continuous at $\mathbf{x}$. However, if $\mathbf{f}$ is Fréchet differentiable at $\mathbf{x}$, then it is continuous at $\mathbf{x}$.

2. The existence of the Jacobian at $\mathbf{x} \in X$ does not imply the existence of a Gâteaux, strict or Fréchet derivative. Only if the Jacobian is a continuous function in the neighborhood of $\mathbf{x}$, Fréchet differentiability can be deduced from Theorem 9.21 in [96].

3. If $\mathbf{f}$ is Fréchet differentiable at $\mathbf{x} \in X$, then it is also Gâteaux differentiable. The Fréchet and Gâteaux derivatives are equal in this case.

4. If $\mathbf{f}$ is strictly differentiable at $\mathbf{x} \in X$, then it is also Fréchet differentiable at $\mathbf{x}$. The strict and the Fréchet derivatives are the equal in this case.

5. If $\mathbf{f}$ is continuously Gâteaux differentiable at $\mathbf{x}$, then it is Fréchet differentiable at $\mathbf{x}$. This follows from the fact that the partial derivatives of $\mathbf{f}$ are continuous functions and per Theorem 9.21 in [96], Fréchet differentiability follows. Note that differentiability in [96] is equivalent to Fréchet differentiability.

6. Let $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbf{g} : \mathbb{R}^m \to \mathbb{R}^p$ be Fréchet differentiable at $\mathbf{x}$ and $\mathbf{f}(\mathbf{x})$, respectively. Then $\mathbf{g} \circ \mathbf{f}$ is Fréchet differentiable at $\mathbf{x}$. If $\mathbf{f}$ is only Gâteaux differentiable at $\mathbf{x}$ then $\mathbf{g} \circ \mathbf{f}$ is Gâteaux differentiable at $\mathbf{x}$.

2.2.2 Mean Value Theorem for Differentiable Functions

Theorem 2.2.6 (Mean Value Theorem for Gâteaux Differentiable Functions). *If $f : [a, b] \to \mathbb{R}$ is continuous and Gâteaux differentiable on (a, b), then there is a point $x \in (a, b)$ such that $f(b) - f(a) = (b - a)\nabla f(x)$ (Theorem 5.10 in [96]).*

2.3 Set-valued Maps

The results of this section are mainly from [35]. Let $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ in the remainder of this section.

Definition 2.3.1 (Set-valued Map). *A set-valued map $\mathcal{S} : X \rightrightarrows Y$ is a map from the set X to the subsets of the set Y.*

$gph(\mathcal{S})$, the graph of $\mathcal{S}$, is the set $\{(\mathbf{x}, \mathbf{y}) \in X \times Y : \mathbf{y} \in \mathcal{S}(\mathbf{x})\}$.

$dom(\mathcal{S})$, the domain of $\mathcal{S}$, is the set $\{\mathbf{x} \in X : \mathcal{S}(\mathbf{x}) \neq \emptyset\}$.

$rge(\mathcal{S})$, the range of $\mathcal{S}$, is the set $\{\mathbf{y} \in Y : \exists \mathbf{x} \in X$ with $\mathbf{y} \in \mathcal{S}(\mathbf{x})\}$.

Instead of considering $\mathcal{S} : X \rightrightarrows Y$, one can consider $\mathcal{S} : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ by defining $\mathcal{S}(\mathbf{x}) = \emptyset$ if $\mathbf{x}$ is not in X. The domain and the range do not change after this extension.

Definition 2.3.2 (Closed Set-valued Map). *Let $\mathcal{S}$ be a set-valued map from X to Y. Let $\{\mathbf{x}_i\} \subset \mathbb{R}^n$ be a sequence such that $\mathbf{x}_i \to \mathbf{x}$. Let $\{\mathbf{y}_i\} \subset \mathbb{R}^m$ be such that $\mathbf{y}_i \in \mathcal{S}(\mathbf{x}_i)$ and $\mathbf{y}_i \to \mathbf{y}$. If $\mathbf{y} \in \mathcal{S}(\mathbf{x})$ for any sequences $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_i\}$ such as described, then $\mathcal{S}$ is closed at $\mathbf{x}$. $\mathcal{S}$ is a closed set-valued map if it is closed at all $\mathbf{x} \in X$.*

Definition 2.3.3 (Locally Bounded Set-valued Map). *Let $\mathcal{S}$ be a set-valued map from X to Y. $\mathcal{S}$ is locally bounded at $\mathbf{x}$ if there exists, O, a neighborhood of $\mathbf{x}$ such that* $\bigcup_{\mathbf{z} \in O \cap dom(\mathcal{S})} \mathcal{S}(\mathbf{z})$ *is bounded. $\mathcal{S}$ is a locally bounded set-valued map if it is locally bounded at all $\mathbf{x} \in X$.*

Definition 2.3.4 (Upper Semicontinuity of Set-valued Maps). *Let $\mathcal{S}$ be a set-valued map from X to Y. $\mathcal{S}$ is upper semicontinuous at $\mathbf{x}$ if for all open sets $V \subset \mathbb{R}^m$ such that $\mathcal{S}(\mathbf{x}) \subset V$, there exists O, a neighborhood of $\mathbf{x}$ such that for all $\mathbf{z} \in O$, $\mathcal{S}(\mathbf{z}) \subset V$ holds. $\mathcal{S}$ is an upper semicontinuous set-valued map if $\mathcal{S}$ is upper semicontinuous at all $\mathbf{x} \in X$.*

Definition 2.3.5 (Lower Semicontinuity of Set-valued Maps). *Let $\mathcal{S}$ be a set-valued map from X to Y. $\mathcal{S}$ is lower semicontinuous at $\mathbf{x} \in \mathbb{R}^n$ if for every open set V such that $\mathcal{S}(\mathbf{x}) \cap V \neq \emptyset$, there exists O, a neighborhood of $\mathbf{x}$ such that for all $\mathbf{z} \in O$, $\mathcal{S}(\mathbf{z}) \cap V \neq \emptyset$ holds. $\mathcal{S}$ is a lower semicontinuous set-valued map if $\mathcal{S}$ is lower semicontinuous at all $\mathbf{x} \in X$.*

Definition 2.3.6 (Continuity of Set-valued maps). *Let $\mathcal{S}$ be a set-valued map from X to Y. $\mathcal{S}$ is continuous at $\mathbf{x} \in X$ if it is lower and upper semicontinuous at $\mathbf{x}$. $\mathcal{S}$ is a continuous set-valued map if $\mathcal{S}$ is continuous at all $\mathbf{x} \in X$.*

Remark 2.3.7 (Upper and Lower Semicontinuity of Functions). The concept of upper semicontinuity defined for set-valued maps does not coincide with the concept of upper and lower semicontinuity defined for functions.

Let X be a subset of $\mathbb{R}^n$ and $\mathbf{x}^*$ be a limit point of X.

The *limit inferior* of $f : X \to \mathbb{R}$ at $\mathbf{x}^*$ is

$$\liminf_{\mathbf{x} \to \mathbf{x}^*} f(\mathbf{x}) = \sup_{\sigma > 0} \inf\{f(\mathbf{x}) : \mathbf{x} \in X,\ 0 < \|\mathbf{x} - \mathbf{x}^*\| < \sigma\}.$$

The *limit superior* of $f : X \to \mathbb{R}$ is

$$\limsup_{\mathbf{x}\to\mathbf{x}^*} f(\mathbf{x}) = \inf_{\sigma>0} \sup\{f(\mathbf{x}) : \mathbf{x} \in X,\ 0 < \|\mathbf{x} - \mathbf{x}^*\| < \sigma\}.$$

Note that $\liminf_{\mathbf{x}\to\mathbf{x}^*} f(\mathbf{x}) \leq \limsup_{\mathbf{x}\to\mathbf{x}^*} f(\mathbf{x})$. Also f is continuous at $\mathbf{x}^*$ if and only if $\liminf_{\mathbf{x}\to\mathbf{x}^*} f(\mathbf{x}) = \limsup_{\mathbf{x}\to\mathbf{x}^*} f(\mathbf{x}) = f(\mathbf{x}^*)$.

Let $\mathbf{x}^* \in X$. $f : X \to \mathbb{R}$ is *upper semicontinuous at* $\mathbf{x}^*$ if $\limsup_{\mathbf{x}\to\mathbf{x}^*} f(\mathbf{x}) \leq f(\mathbf{x}^*)$ and *lower semicontinuous at* $\mathbf{x}^*$ if $\liminf_{\mathbf{x}\to\mathbf{x}^*} f(\mathbf{x}) \geq f(\mathbf{x}^*)$. A function that is both upper and lower semicontinuous at a point is *continuous* at that point.

If a function is continuous at a point, it is also continuous at that point as a set-valued map. If a function is upper or lower semicontinuous at a point then it is neither upper nor lower semicontinuous at that point as a set-valued map.

2.3.1 Properties of Set-valued Maps

Let $\mathcal{S}$ be a set-valued mapping from X to Y.

1. If $\mathcal{S}$ is closed and locally bounded at $\mathbf{x}$, then it is upper semicontinuous at $\mathbf{x}$.

2. $\mathcal{S}$ is closed if and only if its graph is a closed set.

3. If $\mathcal{S}$ is lower semicontinuous at $\mathbf{x}$, then for every $\{\mathbf{x}_i\} \subset X$ such that $\mathbf{x}_i \to \mathbf{x}$ and every $\mathbf{y} \in \mathcal{S}(\mathbf{x})$, there exists a sequence $\{\mathbf{y}_i\}$ such that $\mathbf{y}_i \to \mathbf{y}$ and $\mathbf{y}_i \in \mathcal{S}(\mathbf{x}_i)$.

4. If $\mathcal{S}$ is upper semicontinuous at $\mathbf{x} \in X$, then for every scalar ϵ, there exists O, a neighborhood of $\mathbf{x}$, such that for all $\mathbf{z} \in O$, $\mathcal{S}(\mathbf{z}) \subset \mathcal{S}(\mathbf{x}) + \epsilon\mathbb{B}(0,1)$ where $\mathbb{B}(0,1)$ is the unit ball in $\mathbb{R}^m$.

2.4 Elementary Convex Analysis

The results of this section are mainly from [15].

2.4.1 Convex Sets

Definition 2.4.1 (Convex Set). *A set $C \in \mathbb{R}^n$ is convex if for all $\alpha \in (0,1)$, $\alpha\mathbf{x}+(1-\alpha)\mathbf{y}$ is in C whenever $\mathbf{x}$ and $\mathbf{y}$ are in C. The empty set and any singleton set are convex.*

Properties of Convex Sets

1. Let $\mathcal{I} = \{1, \ldots, k\}$ where k is a positive integer (possibly ∞). Let C_i be a convex subset of $\mathbb{R}^n$ for all $i \in \mathcal{I}$. Then $\cap_{i=1}^{k} C_i$ is a convex set.

2. If $C \subset \mathbb{R}^n$ is convex, then $\operatorname{int}(C)$ and $\operatorname{cl}(C)$ are convex sets.

3. If $C_1 \subset \mathbb{R}^n$ and $C_2 \subset \mathbb{R}^n$ are convex sets, then $\alpha_1 C_1 + \alpha_2 C_2$ is a convex set where α_1 and α_2 are scalars.

Definition 2.4.2 (Convex Combination). *Let $\mathcal{I} = \{1, \ldots, k\}$ where k is a positive integer (possibly ∞). Let $\mathbf{x}_i \in \mathbb{R}^n$ for all $i \in \mathcal{I}$. A convex combination of the vectors $\{\mathbf{x}_i\}_{i=1}^{k}$ is $\sum_{i=1}^{k} \alpha_i \mathbf{x}_i$ where for all $i \in \mathcal{I}$, $\alpha_i \in \mathbb{R}$, $\alpha_i \geq 0$ and $\sum_{i=1}^{k} \alpha_i = 1$.*

Definition 2.4.3 (Convex Hull). *The convex hull of a nonempty set $C \subset \mathbb{R}^n$ is the intersection of all convex sets containing C and is denoted by* $\operatorname{conv}(C)$.

$\operatorname{conv}(C)$ *is also equal to the set* $\{\sum_{i=1}^{k} \alpha_i \mathbf{x}_i : \alpha_i \in \mathbb{R},\ \alpha_i \geq 0,\ \sum_{i=1}^{k} \alpha_i = 1,\ \mathbf{x}_i \in C,\ k \in \{1, 2, \ldots\}\}$.

Definition 2.4.4 (Closed Convex Hull). *The closed convex hull of a nonempty set $C \subset \mathbb{R}^n$ is the intersection of all closed convex sets containing C and is denoted by* $\operatorname{cl}(\operatorname{conv}(C))$.

The closed convex hull is also the closure of the convex combinations of the elements of $C \subset \mathbb{R}^n$. Note that the closure and convex hull operations are not in general interchangeable. If the set C is bounded, then the operations are interchangeable.

Theorem 2.4.5 (Carathéodory's Theorem). *Let C be a nonempty subset of $\mathbb{R}^n$. Then any $\mathbf{x} \in \text{conv}\,(C)$ can be represented as a convex combination of $n+1$ not necessarily different elements of C.*

Theorem 2.4.6 (Projection onto Closed Convex Sets). *The closest point of a convex set $C \subset \mathbb{R}^n$ to a point $\mathbf{x} \in \mathbb{R}^n$ is called the projection of $\mathbf{x}$ on C and is denoted by $\mathbf{p}_C(\mathbf{x})$. A unique minimizer, ($\{\mathbf{p}_C(\mathbf{x})\} = \underset{\mathbf{y} \in C}{\text{argmin}} \|\mathbf{x}-\mathbf{y}\|$), always exists if C is closed and nonempty. In addition, $\mathbf{z} = \mathbf{p}_C(\mathbf{x})$ if and only if $\langle \mathbf{x}-\mathbf{z}, \mathbf{v}-\mathbf{z} \rangle \leq 0$, $\forall \mathbf{v} \in C$ and $\|\mathbf{p}_C(\mathbf{x})-\mathbf{p}_C(\mathbf{y})\| \leq \|\mathbf{x}-\mathbf{y}\|$ holds for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.*

Definition 2.4.7 (Hyperplane). *$H_{\mathbf{a},b}$, a hyperplane in $\mathbb{R}^n$, is the set $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^{\mathrm{T}}\mathbf{x} = b\}$ where $\mathbf{a} \in \mathbb{R}^n$ and b is a scalar.*

The sets $\{\mathbf{x} : \mathbf{a}^{\mathrm{T}}\mathbf{x} > b\}$ and $\{\mathbf{x} : \mathbf{a}^{\mathrm{T}}\mathbf{x} \geq b\}$ are the open and closed positive halfspaces, respectively, associated with H. Analogously, $\{\mathbf{x} : \mathbf{a}^{\mathrm{T}}\mathbf{x} < b\}$ and $\{\mathbf{x} : \mathbf{a}^{\mathrm{T}}\mathbf{x} \leq b\}$ are the open and closed negative halfspaces, respectively, associated with H.

Theorem 2.4.8 (Supporting Hyperplane Theorem). *Let C be a nonempty convex subset of $\mathbb{R}^n$ and $\mathbf{x} \in C \backslash \text{int}\,(C)$. Then there exists a hyperplane $H_{\mathbf{a},b}$ such that $\mathbf{a}^{\mathrm{T}}\mathbf{x} = b$ and $\mathbf{a}^{\mathrm{T}}\mathbf{x} \leq \mathbf{a}^{\mathrm{T}}\mathbf{y}$, $\forall \mathbf{y} \in C$.*

2.4.2 Convex Functions

Definition 2.4.9 (Convex Function). *$f : C \to \mathbb{R}$ is a convex function if $C \subset \mathbb{R}^n$ is a convex set and for any $\mathbf{x} \in C$ and $\mathbf{y} \in C$ and all $\alpha \in (0,1)$, $f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y})$ holds.*

The epigraph of f, epi(f) is the set $\{(\mathbf{x}, y) : \mathbf{x} \in C,\ y \in \mathbb{R},\ f(\mathbf{x}) \leq y\}$.

Let $C \subset \mathbb{R}^n$ be a convex set. Then $f : C \to \mathbb{R}$ is a convex function if and only if epi(f) is a convex subset of $\mathbb{R}^{n+1}$.

Definition 2.4.10 (Strictly Convex Function). *$f : C \to \mathbb{R}$ is strictly convex function if it is a convex function and for any $\mathbf{x} \in C$ and $\mathbf{y} \in C$ and all $\alpha \in (0,1)$, $f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) < \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y})$ holds.*

Definition 2.4.11 (The Subgradient and Subdifferential). *Let C be a convex subset of $\mathbb{R}^n$ and $f : C \to \mathbb{R}$ be a convex function.*

$\mathbf{g} \in \mathbb{R}^n$ is a subgradient of f at $\mathbf{x} \in \text{int}\,(C)$ if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{x} - \mathbf{y} \rangle,\ \forall \mathbf{y} \in C.$$

holds. The existence of a subgradient is guaranteed by Theorem 2.4.8 if epi(f)is considered as a convex subset of $\mathbb{R}^{n+1}$. Note that it is possible to obtain $\mathbf{g}$ such that $\|\mathbf{g}\| = +\infty$ at $\mathbf{x} \notin \text{int}\,(C)$. Therefore only $\mathbf{x} \in \text{int}\,(C)$ are considered in the above definition.

$\partial f(\mathbf{x})$, the subdifferential of f at $\mathbf{x} \in \text{int}\,(C)$ is the set of all subgradients of f at $\mathbf{x}$. $\partial f(\mathbf{x})$ is a convex, compact subset of $\mathbb{R}^n$ and ∂f is an upper semicontinuous set-valued map at $\mathbf{x} \in \text{int}\,(C)$.

2.5 Locally Lipschitz Continuous Functions

Let n and m be finite positive integers. Let X be a subset of $\mathbb{R}^n$.

Definition 2.5.1 (Lipschitz Continuity). *$\mathbf{f} : X \to \mathbb{R}^m$ is a Lipschitz continuous function on X if there exists $K \in [0, +\infty)$ such that $\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| \leq K\|\mathbf{x} - \mathbf{y}\|$, $\forall \mathbf{x}, \mathbf{y} \in X$.*

Definition 2.5.2 (Local Lipschitz Continuity). *$\mathbf{f} : X \to \mathbb{R}^m$ is locally Lipschitz continuous at $\mathbf{x} \in \text{int}\,(X)$ if there exists a constant $K \in [0, +\infty)$ and O, a neighborhood of $\mathbf{x}$, such*

that $\|\mathbf{f}(\mathbf{z}) - \mathbf{f}(\mathbf{y})\| \leq K\|\mathbf{z} - \mathbf{y}\|$, $\forall \mathbf{z}, \mathbf{y} \in O$. $\mathbf{f}$ *is a locally Lipschitz continuous function if it is locally Lipschitz continuous at all* $\mathbf{x} \in X$.

2.5.1 Properties of Locally Lipschitz Continuous Functions

The following are standard results that can be found easily in the literature [25, 92].

1. Local Lipschitz continuity does not imply Lipschitz continuity. For example, $f : \mathbb{R} \to \mathbb{R} : x \mapsto x^2$ is a locally Lipschitz continuous function but not a Lipschitz continuous function.

2. Differentiable functions may not be locally Lipschitz continuous (See Example 2.9.7).

3. If $\mathbf{f} : X \to \mathbb{R}^m$ is continuously differentiable at $\mathbf{x} \in X$, then it is locally Lipschitz continuous at $\mathbf{x}$.

4. If $\mathbf{f} : X \to \mathbb{R}^m$ is locally Lipschitz continuous and Gâteaux differentiable at $\mathbf{x} \in X$, then it is also Fréchet differentiable.

5. If $\mathbf{f} : X \to \mathbb{R}^m$ is strictly differentiable at $\mathbf{x} \in X$, then it is locally Lipschitz continuous at $\mathbf{x}$ (Proposition 2.2.1 in [25]).

6. Let $\mathbf{f} : X \to \mathbb{R}^m$ and $\mathbf{g} : X \to \mathbb{R}^m$ be locally Lipschitz continuous at $\mathbf{x}$. Then $\mathbf{f} + \mathbf{g}$ is locally Lipschitz continuous at $\mathbf{x}$.

7. Let $\mathbf{f} : X \to \mathbb{R}^m$ and $\mathbf{g} : \mathbb{R}^m \to \mathbb{R}^l$ be locally Lipschitz continuous at $\mathbf{x}$ and $\mathbf{f}(\mathbf{x})$, respectively. Then $\mathbf{g} \circ \mathbf{f}$ is locally Lipschitz continuous at $\mathbf{x}$.

8. Let $f : X \to \mathbb{R}$ and $g : X \to \mathbb{R}$ be locally Lipschitz continuous at $\mathbf{x}$. Then fg is locally Lipschitz continuous at $\mathbf{x}$.

9. (Rademacher's Theorem, [92]) Let X be an open subset of $\mathbb{R}^n$. If $\mathbf{f} : X \to \mathbb{R}^m$ is a locally Lipschitz continuous function on X then it is differentiable at all $\mathbf{x} \in X \backslash S$ where S is a measure zero subset of X, and $X \backslash S$ is dense in X.

10. Let X be an open convex subset of $\mathbb{R}^n$. If $f : X \to \mathbb{R}$ is a convex and bounded function, then f is a locally Lipschitz continuous function on X.

11. $f : X \to \mathbb{R}$ is strictly differentiable in a neighborhood of $\mathbf{x}$ if and only if it is continuously differentiable on that neighborhood of $\mathbf{x}$ (Corollary of Proposition 2.2.4 in [25]).

12. $\mathbf{F} : X \to \mathbb{R}^m$ is strictly differentiable in a neighborhood of $\mathbf{x}$ if and only if it is continuously differentiable on that neighborhood of $\mathbf{x}$.

    Proof. Let $\mathbf{F}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$. If $\mathbf{F}$ is strictly differentiable in a neighborhood of $\mathbf{x}$, then each f_i are strictly differentiable on that neighborhood. Hence each f_i is continuously differentiable on that neighborhood per the previous item. As a result $\mathbf{F}$ is continuously differentiable on that neighborhood. If $\mathbf{F}$ is continuously differentiable on a neighborhood of $\mathbf{x}$, then f_i are continuously differentiable on that neighborhood. Hence f_i are strictly differentiable on that neighborhood. Strict differentiability of $\mathbf{F}$ on that neighborhood follows. □

2.6 Nonsmooth Analysis for Locally Lipschitz Continuous Functions

In this section, relevant results of nonsmooth analysis for finite dimensional Euclidean spaces are summarized. The results are mainly from [25].

In the remainder of this section, let $X \subset \mathbb{R}^n$ and $f : X \to \mathbb{R}$ be locally Lipschitz continuous at $\mathbf{x} \in X$ with Lipschitz constant K. Let O be the corresponding neighborhood of $\mathbf{x}$ and S be the measure zero subset of O such that if $\mathbf{z} \in S$, $\nabla f(\mathbf{z})$ does not exist.

2.6.1 The Generalized Directional Derivative

Definition 2.6.1 (The Generalized Directional Derivative). *Let* $\mathbf{v} \in \mathbb{R}^n$. $f^o(\mathbf{x}; \mathbf{v})$, *the generalized directional derivative at* $\mathbf{x}$ *in the direction* $\mathbf{v}$ *(page 25 in [25]) is*

$$f^o(\mathbf{x}; \mathbf{v}) = \limsup_{(\mathbf{y},\, t) \to (\mathbf{x}, 0^+)} \frac{f(\mathbf{y} + t\mathbf{v}) - f(\mathbf{y})}{t}. \tag{2.6.1}$$

Properties of Generalized Directional Derivatives

1. The mapping $\mathbf{v} \mapsto f^o(\mathbf{x}; \mathbf{v})$ is finite, convex and satisfies $|f^o(\mathbf{x}; \mathbf{v})| \leq K\|\mathbf{v}\|$ on $\mathbb{R}^n$. In addition, the mapping $\mathbf{v} \mapsto f^o(\mathbf{x}; \mathbf{v})$ is Lipschitz continuous with constant K on $\mathbb{R}^n$.

2. The mapping $(\mathbf{x}, \mathbf{v}) \mapsto f^o(\mathbf{x}; \mathbf{v})$ is an upper semicontinuous function.

3. $f^o(\mathbf{x}; -\mathbf{v}) = -f^o(\mathbf{x}; \mathbf{v})$.

4. $f^o(\mathbf{x}; \mathbf{0}) = 0$.

2.6.2 The Generalized Gradient

Definition 2.6.2 (The Generalized Gradient). *If* f *is locally Lipschitz continuous at* $\mathbf{x}$, *then the function* $\mathbf{v} \mapsto f^o(\mathbf{x}; \mathbf{v})$ *is a finite convex function from* $\mathbb{R}^n$ *to* $\mathbb{R}$. *Per the Supporting Hyperplane Theorem (Theorem 2.4.8), there exists a vector* $\boldsymbol{\zeta}$ *such that* $f^o(\mathbf{x}; \mathbf{v}) - f^o(\mathbf{x}; \mathbf{0}) \geq \langle \boldsymbol{\zeta}, \mathbf{v} \rangle, \forall \mathbf{v} \in \mathbb{R}^n$. $\partial f(\mathbf{x})$, *the generalized gradient at* $\mathbf{x}$, *is the set of all such* $\boldsymbol{\zeta} \in \mathbb{R}^n$ *(page 27*

in [25]). Formally, it is the set

$$\{\boldsymbol{\zeta} \in \mathbb{R}^n : \ f^o(\mathbf{x};\mathbf{v}) \geq \langle \boldsymbol{\zeta}, \mathbf{v} \rangle, \ \forall \mathbf{v} \in \mathbb{R}^n\}.$$

This definition can be used to define the generalized gradient of scalar functions whose domains are subsets of arbitrary Banach spaces using the Hahn-Banach Theorem.

An alternative definition applicable to functions whose domains are subsets of finite-dimensional Euclidean spaces uses the gradient of the functions.

Definition 2.6.3 (The Generalized Gradient II). *Let $Q \subset O$ be any set of measure zero. Let $\{\mathbf{x}_i\}$ be any sequence such that $\mathbf{x}_i \in O\backslash(S \cup Q)$ for all i and $\mathbf{x}_i \to \mathbf{x}$. Let $\{\nabla f(\mathbf{x}_i)\}$ be the corresponding sequence of gradients and $\{\lim_{i\to\infty} \nabla f(\mathbf{x}_i)\}$ be the set of the limits of all convergent sequences, $\{\nabla f(\mathbf{x}_i)\}$. Then $\partial f(\mathbf{x})$, the generalized gradient of f at $\mathbf{x}$, is the convex hull of the set $\{\lim_{i\to\infty} \nabla f(\mathbf{x}_i)\}$ (Theorem 2.5.1 in [25]). Formally,*

$$\partial f(\mathbf{x}) = \text{conv}\left(\{\lim_{i\to\infty} \nabla f(\mathbf{x}_i) : \ \mathbf{x}_i \to \mathbf{x}, \ \mathbf{x}_i \in O\backslash(S \cup Q)\}\right). \tag{2.6.2}$$

Properties of the Generalized Gradient

1. $\boldsymbol{\zeta} \in \partial f(\mathbf{x})$ if and only if $f^o(\mathbf{x};\mathbf{v}) \geq \langle \boldsymbol{\zeta}, \mathbf{v} \rangle, \ \forall \mathbf{v} \in \mathbb{R}^n$.

2. $\partial f(\mathbf{x})$ is a nonempty, convex and compact subset of $\mathbb{R}^n$ and if $\boldsymbol{\zeta} \in \partial f(\mathbf{x})$ then $\|\boldsymbol{\zeta}\| \leq K$.

3. For every $\mathbf{v} \in \mathbb{R}^n$, $f^o(\mathbf{x};\mathbf{v}) = \max_{\boldsymbol{\zeta} \in \partial f(\mathbf{x})} \{\langle \boldsymbol{\zeta}, \mathbf{v} \rangle\}$.

4. The set-valued map ∂f is locally bounded and uppersemicontinuous at $\mathbf{x}$.

5. If f is a convex and finite function on O, then the generalized gradient at $\mathbf{x} \in O$ is equal to the subdifferential at $\mathbf{x}$.

6. If f is differentiable at $\mathbf{x}$, then $\nabla f(\mathbf{x}) \in \partial f(\mathbf{x})$.

7. If f is strictly differentiable at $\mathbf{x}$, then $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.

8. If the directional derivative exists in the direction $\mathbf{v} \in \mathbb{R}^n$, then $f'(\mathbf{x}; \mathbf{v}) = \langle \boldsymbol{\zeta}, \mathbf{v} \rangle$ for some $\boldsymbol{\zeta} \in \partial f(\mathbf{x})$.

9. Let O be an open subset of $\mathbb{R}^n$ and $g : O \to \mathbb{R}$ be a locally Lipschitz continuous function which attains a minimum or maximum at $\mathbf{x} \in O$, then $\mathbf{0} \in \partial f(\mathbf{x})$ (Proposition 2.3.2 in [25]).

Mean Value Theorem for Locally Lipschitz Continuous Functions

Theorem 2.6.4 (Mean Value Theorem for Locally Lipschitz Continuous Functions). *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and f be locally Lipschitz continuous on an open set containing the line segment $L = \{\mathbf{u} : \mathbf{u} = \lambda \mathbf{x} + (1 - \lambda)\mathbf{y}, \lambda \in (0, 1)\}$. Then, there exists a point $\mathbf{u}^* \in L$ and $\boldsymbol{\zeta} \in \partial f(\mathbf{u}^*)$ such that $f(\mathbf{y}) - f(\mathbf{x}) = \langle \boldsymbol{\zeta}, \mathbf{y} - \mathbf{x} \rangle$ (Theorem 2.3.7 in [25]).*

Regularity

If $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz continuous at $\mathbf{x} \in X$, it is also *(Clarke) regular* at $\mathbf{x}$ if

1. for all $\mathbf{v}$, the directional derivative exists, and

2. for all $\mathbf{v}$, $f'(\mathbf{x}; \mathbf{v}) = f^o(\mathbf{x}; \mathbf{v})$.

If f is convex and finite in a neighborhood of $\mathbf{x} \in X$, then it is regular at $\mathbf{x}$. If f is strictly differentiable or continuously differentiable at $\mathbf{x} \in X$, then it is regular at $\mathbf{x}$.

Calculus Rules for the Generalized Gradient

Let $f : X \to \mathbb{R}$ and $g_i : X \to \mathbb{R}$ be locally Lipschitz continuous at $\mathbf{x} \in X$ for all $i \in \{1, \ldots, N\}$ where N is a finite integer.

1. If $\alpha \in \mathbb{R}$ then $\partial(\alpha f)(\mathbf{x}) = \alpha \partial f(\mathbf{x})$.

2. $\partial\left(\sum_{i=1}^{N} g_i\right)(\mathbf{x}) \subset \sum_{i=1}^{N} \partial g_i(\mathbf{x})$. Equality holds if all but at most one of the g_i are strictly differentiable at $\mathbf{x}$. Equality holds if all g_i are regular at $\mathbf{x}$.

3. If $\alpha_i \in \mathbb{R}$ for all $i \in \{1, \ldots, N\}$, then

$$\partial(\sum_{i=1}^{N} \alpha_i g_i)(\mathbf{x}) \subset \sum_{i=1}^{N} \alpha_i \partial g_i(\mathbf{x}).$$

Equality holds if all but at most one of the g_i are strictly differentiable at $\mathbf{x}$. Equality holds if all g_i are regular at $\mathbf{x}$ and each α_i is nonnegative.

4. $\partial(g_1 g_2)(\mathbf{x}) \subset g_2(\mathbf{x})\partial g_1(\mathbf{x}) + g_1(\mathbf{x})\partial g_2(\mathbf{x})$. If $g_2(\mathbf{x}) \geq 0$, $g_1(\mathbf{x}) \geq 0$ and g_1, g_2 are both regular at $\mathbf{x}$ then equality holds and $g_1 g_2$ is regular at $\mathbf{x}$.

5. Suppose $g_2(\mathbf{x}) \neq 0$, then

$$\partial\left(\frac{g_1}{g_2}\right)(\mathbf{x}) \subset \frac{g_2(\mathbf{x})\partial g_1(\mathbf{x}) - g_1(\mathbf{x})\partial g_2(\mathbf{x})}{g_2^2(\mathbf{x})}.$$

If $g_1(\mathbf{x}) \geq 0$, $g_2(\mathbf{x}) > 0$ and if g_1,$-g_2$ are both regular at $\mathbf{x}$ then equality holds and g_1/g_2 is regular at $\mathbf{x}$.

6. Let $h(\mathbf{x}) = \max\{g_i(\mathbf{x}), \ldots, g_N(\mathbf{x})\}$. Let $\mathcal{I}(\mathbf{x}) \subset \{1, \ldots, N\}$ denote the set of indices i for which $h(\mathbf{x}) = g_i(\mathbf{x})$. Then

$$\partial h(\mathbf{x}) \subset \operatorname{conv}\left(\bigcup_{i \in \mathcal{I}(\mathbf{x})} \partial g_i(\mathbf{x})\right). \tag{2.6.3}$$

If g_i are regular at $\mathbf{x}$ for all $i \in \mathcal{I}(\mathbf{x})$, then h is regular at $\mathbf{x}$ and

$$\partial h(\mathbf{x}) = \operatorname{conv}\left(\bigcup_{i \in \mathcal{I}(\mathbf{x})} \partial g_i(\mathbf{x})\right). \tag{2.6.4}$$

(Proposition 2.3.12 in [25]).

2.6.3 The Generalized Jacobian

Let X be an open subset of $\mathbb{R}^n$ and $\mathbf{F} : X \to \mathbb{R}^m$ be such that $\mathbf{F}(\mathbf{y}) = (f_i(\mathbf{y}), \ldots, f_m(\mathbf{y}))$ where each $f_i : X \to \mathbb{R}$ is a locally Lipschitz continuous function on X. Let O be a neighborhood of $\mathbf{x}$ such that the Lipschitz constant of each f_i on O is K_i. Then $\mathbf{F}$ is locally Lipschitz continuous at $\mathbf{x}$ with Lipschitz constant $K = \sqrt{\sum_{i=1}^{m} K_i^2}$ and $\mathbf{JF}(\mathbf{y})$ exists for all $\mathbf{y} \in O \backslash S$ where S is a measure zero subset of O per Rademacher's Theorem.

Definition 2.6.5 (The Generalized Jacobian). *Let $Q \subset O$ be any set of measure zero. Let $\{\mathbf{x}_i\}$ be any sequence such that $\mathbf{x}_i \in O \backslash (S \cup Q)$[†] for all i and $\mathbf{x}_i \to \mathbf{x}$. Let $\{\mathbf{JF}(\mathbf{x}_i)\}$ be the corresponding sequence of Jacobians and $\{\lim_{i\to\infty} \mathbf{JF}(\mathbf{x}_i)\}$ be the set of the limits of all convergent sequences, $\{\mathbf{JF}(\mathbf{x}_i)\}$. The generalized Jacobian of $\mathbf{F}$ at $\mathbf{x}$, $\partial\mathbf{F}(\mathbf{x})$, is the convex hull of $\{\lim_{i\to\infty} \mathbf{JF}(\mathbf{x}_i)\}$ (Definition 2.6.1 in [25]). In short,*

$$\partial\mathbf{F}(\mathbf{x}) = \operatorname{conv}\Big(\{\lim_{i\to\infty} \mathbf{JF}(\mathbf{x}_i) : \ \mathbf{x}_i \to \mathbf{x}, \ \mathbf{x}_i \in O \backslash (S \cup Q)\}\Big). \tag{2.6.5}$$

Properties of the Generalized Jacobian

1. $\partial\mathbf{F}(\mathbf{x})$, $\mathbf{x} \in X$ is a nonempty, compact and convex subset of $\mathbb{R}^{m\times n}$.

2. The set-valued mapping $\partial\mathbf{F}$ is locally bounded and uppersemicontinuous at $\mathbf{x} \in X$.

3. $\|\mathbf{Z}\| \leq K$ holds for all $\mathbf{Z} \in \partial\mathbf{F}(\mathbf{x})$ and $\mathbf{x} \in X$.

4. $\partial\mathbf{F}(\mathbf{x}) \subset \{\mathbf{A} : \ \mathbf{A} \in \mathbb{R}^{m\times n}, \ \mathbf{a}_i^{\mathrm{T}} \in \partial f_i(\mathbf{x}), \ i = 1, \ldots, m\}$ where $\mathbf{a}_i$ is the ith row of $\mathbf{A}$ and $\mathbf{x} \in X$.

[†]Clarke does not mention the set Q when defining the generalized Jacobian in [25]. The indifference of the generalized Jacobian to a set of measure zero is proven in [113].

5. If $m = 1$ then the set containing the transposes of all the elements of the generalized Jacobian of F at $\mathbf{x} \in X$ is the generalized gradient of F at $\mathbf{x} \in X$.

6. If $\mathbf{F}$ is differentiable at $\mathbf{x} \in X$, then $\mathbf{JF}(\mathbf{x}) \in \partial\mathbf{F}(\mathbf{x})$.

7. If $\mathbf{F}$ is strictly differentiable at $\mathbf{x} \in X$, then $\partial\mathbf{F}(\mathbf{x}) = \{\mathbf{JF}(\mathbf{x})\}$(Corollary 3.8 in [78]).

Chain Rules for the Generalized Jacobian and Gradient

Theorem 2.6.6. *Let $f = g \circ \mathbf{F}$ where $\mathbf{F} : \mathbb{R}^m \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}$ are locally Lipschitz continuous at $\mathbf{x}$ and $\mathbf{F}(\mathbf{x})$, respectively, then*

$$\partial f(\mathbf{x}) \subset \operatorname{conv}\left(\{(\boldsymbol{\zeta}^{\mathrm{T}}\mathbf{A})^{\mathrm{T}},\ \boldsymbol{\zeta} \in \partial g(\mathbf{F}(\mathbf{x})),\ \mathbf{A} \in \partial\mathbf{F}(\mathbf{x})\}\right).$$

If g is strictly differentiable at $\mathbf{F}(\mathbf{x})$ then equality holds (Theorem 2.6.6 in [25]).

Theorem 2.6.7. *Let $\mathbf{F} : \mathbb{R}^m \to \mathbb{R}^n$ and $\mathbf{G} : \mathbb{R}^n \to \mathbb{R}^p$ with $\mathbf{F}$ and $\mathbf{G}$ locally Lipschitz continuous at $\mathbf{x}$ and $\mathbf{F}(\mathbf{x})$, respectively, then*

$$\partial(\mathbf{G} \circ \mathbf{F})(\mathbf{x}) \subset \operatorname{conv}\left(\{\mathbf{AB},\ \mathbf{A} \in \partial\mathbf{G}(\mathbf{F}(\mathbf{x})),\ \mathbf{B} \in \partial\mathbf{F}(\mathbf{x})\}\right).$$

Equality holds if $\mathbf{G}$ is strictly differentiable at $\mathbf{x}$ (Theorem 4 in [48] and Theorem 4.3 in [78]).

Mean Value Theorem for Generalized Jacobians

Theorem 2.6.8 (Mean Value Theorem for Generalized Jacobians). *Let U be a convex open subset of $\mathbb{R}^n$ and $\mathbf{F} : U \to \mathbb{R}^m$ be a locally Lipschitz continuous function on U. Let $\mathbf{x} \in U$, $\mathbf{y} \in U$ and $\bar{L} = \{\mathbf{u} :\ \mathbf{u} = \lambda\mathbf{x} + (1-\lambda)\mathbf{y}, \lambda \in [0,1]\}$ Then $\mathbf{F}(\mathbf{y}) - \mathbf{F}(\mathbf{x}) \in \operatorname{conv}\left(\{\mathbf{Z}(\mathbf{y}-\mathbf{x}) :\ \mathbf{Z} \in \partial\mathbf{F}(\mathbf{u}),\ \mathbf{u} \in \bar{L}\}\right)$ (Proposition 2.6.5 in [25]).*

2.6.4 The Partial Generalized Gradient and Jacobian

Let $\{s_i\}_{i=1}^n$ be a set where each s_i and n are finite positive integers. Let $X_i \subset \mathbb{R}^{s_i}$ be an open set, $\mathbf{x}_i \in X_i$ for all $i \in \{1,\ldots,n\}$ and $\mathbf{x} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)$. Let $\mathbf{F}$ be a function from $\Pi_{i=1}^n X_i$ to $\mathbb{R}^m$, $k \in \{1,\ldots,n\}$ and $\mathbf{F}(\mathbf{x}_1,\ldots,\mathbf{x}_{k-1},\cdot,\mathbf{x}_{k+1},\ldots,\mathbf{x}_n)$ be a locally Lipschitz continuous function on O_k, a neighborhood of $\mathbf{x}_k$. Then

$$\partial_k \mathbf{F}(\mathbf{x}) = \text{conv}\left(\{\lim_{i\to\infty} \mathbf{J}_k\mathbf{F}(\mathbf{z}_i) : \mathbf{x}_{k,i} \to \mathbf{x}_k,\ \mathbf{x}_{k,i} \in O_k \backslash (S_k \cup Q_k)\}\right)$$

where $\mathbf{z}_i = (\mathbf{x}_1,\ldots,\mathbf{x}_{k-1},\mathbf{x}_{k,i},\mathbf{x}_{k+1},\ldots,\mathbf{x}_n)$, Q_k is any measure zero subset of O_k and S_k is the set of points in O_k such that $\mathbf{J}_k\mathbf{F}(\mathbf{x}_i,\ldots,\mathbf{x}_{k-1},\mathbf{u},\mathbf{x}_{k+1},\ldots,\mathbf{x}_n)$ does not exist if $\mathbf{u} \in S_k$.

If $m = 1$ then the set containing the transposes of all the elements of the generalized Jacobian, $\partial_k F(\mathbf{x})$, is the generalized gradient of the function $F(\mathbf{x}_1,\ldots,\mathbf{x}_{k-1},\cdot,\mathbf{x}_{k+1},\ldots,\mathbf{x}_n)$ at $\mathbf{x}$.

Definition 2.6.9 (The Projection of the Generalized Jacobian). *Let $\mathbf{F} : X_1 \times X_2 \to \mathbb{R}^p$ where X_1 and X_2 are open subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. Let $\mathbf{F}$ be locally Lipschitz continuous at $(\mathbf{x}_1,\mathbf{x}_2)$ where $\mathbf{x}_1 \in X_1$ and $\mathbf{x}_2 \in X_2$. Then $\pi_2\partial\mathbf{F}(\mathbf{x}_1,\mathbf{x}_2)$ is the set $\{\mathbf{M} \in \mathbb{R}^{p\times m} : \exists \mathbf{N} \in \mathbb{R}^{p\times n}$ such that $[\mathbf{N}\ \mathbf{M}] \in \partial\mathbf{F}(\mathbf{x}_1,\mathbf{x}_2)\}$. Analogously, $\pi_1\partial\mathbf{F}(\mathbf{x}_1,\mathbf{x}_2)$ is the set $\{\mathbf{M} \in \mathbb{R}^{p\times n} : \exists \mathbf{N} \in \mathbb{R}^{p\times m}$ such that $[\mathbf{M}\ \mathbf{N}] \in \partial\mathbf{F}(\mathbf{x}_1,\mathbf{x}_2)\}$.*

Theorem 2.6.10. *Let $\mathbf{F} : X_1 \times X_2 \to \mathbb{R}^p$ where X_1 and X_2 are open subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. Let $\mathbf{F}$ be locally Lipschitz continuous at $(\mathbf{x}_1,\mathbf{x}_2)$ where $\mathbf{x}_1 \in X_1$ and $\mathbf{x}_2 \in X_2$. Then $\partial_1\mathbf{F}(\mathbf{x}_1,\mathbf{x}_2) \subset \pi_1\partial\mathbf{F}(\mathbf{x}_1,\mathbf{x}_2)$.*

Proof. The result follows from Theorem 3.2 in [78]. Note that in the statement of this theorem, $\partial f(p)$ represents $\partial\mathbf{F}(\mathbf{x}_1,\mathbf{x}_2)$ and the subspace L represents $\mathbb{R}^n$. $\partial f(p)|_L$, the restriction of $\partial f(p)$ to the subspace L in this case corresponds to $\pi_1\partial\mathbf{F}(\mathbf{x}_1,\mathbf{x}_2)$. $\partial_L f(p)$ is an intermediate construct that contains $\partial_1\mathbf{F}(\mathbf{x}_1,\mathbf{x}_2)$ as stated on page 57 in [78]. □

Definition 2.6.11 (The Projection of the Generalized Gradient). *Let* $f : X_1 \times X_2 \to \mathbb{R}$ *where* X_1 *and* X_2 *are open subsets of* $\mathbb{R}^n$ *and* $\mathbb{R}^m$*, respectively. Let* f *be locally Lipschitz continuous at* $(\mathbf{x}_1, \mathbf{x}_2)$ *where* $\mathbf{x}_1 \in X_1$ *and* $\mathbf{x}_2 \in X_2$*. Then* $\pi_2 \partial f(\mathbf{x}_1, \mathbf{x}_2)$ *is the set* $\{\mathbf{M} \in \mathbb{R}^m : \exists \mathbf{N} \in \mathbb{R}^n$ *such that* $(\mathbf{N}, \mathbf{M}) \in \partial f(\mathbf{x}_1, \mathbf{x}_2)\}$*. Analogously,* $\pi_1 \partial f(\mathbf{x}_1, \mathbf{x}_2)$ *is set* $\{\mathbf{M} \in \mathbb{R}^n : \exists \mathbf{N} \in \mathbb{R}^m$ *such that* $(\mathbf{M}, \mathbf{N}) \in \partial f(\mathbf{x}_1, \mathbf{x}_2)\}$.

Theorem 2.6.12. *Let* $f : X_1 \times X_2 \to \mathbb{R}$ *where* X_1 *and* X_2 *are open subsets of* $\mathbb{R}^n$ *and* $\mathbb{R}^m$*, respectively. Let* f *be locally Lipschitz continuous at* $(\mathbf{x}_1, \mathbf{x}_2)$ *where* $\mathbf{x}_1 \in X_1$ *and* $\mathbf{x}_2 \in X_2$*. Then* $\partial_1 f(\mathbf{x}_1, \mathbf{x}_2) \subset \pi_1 \partial f(\mathbf{x}_1, \mathbf{x}_2)$ *(Proposition 2.3.16 in [25]).*

2.6.5 Implicit Function Theorem for Locally Lipschitz Continuous Functions

The next theorem is an implicit function theorem summarizing the necessary results for subsequent developments (Corollary of Theorem 7.1.1 on page 256 in [25] and Theorem 1.5 in [28]).

Theorem 2.6.13 (Implicit Function Theorem for Locally Lipschitz Functions). *Let* X_1 *and* X_2 *be open subsets of* $\mathbb{R}^n$ *and* $\mathbb{R}^m$*, respectively. Let* $\mathbf{x}_1 \in X_1$ *and* $\mathbf{x}_2 \in X_2$*. Let* $\mathbf{H} : X_1 \times X_2 \to \mathbb{R}^m$ *be locally Lipschitz continuous at* $(\mathbf{x}_1, \mathbf{x}_2)$.

Let $\pi_2 \partial \mathbf{H}(\mathbf{x}_1, \mathbf{x}_2)$ *be the set* $\{\mathbf{M} \in \mathbb{R}^{m \times m} : \exists \mathbf{N} \in \mathbb{R}^{m \times n}$ *such that* $[\mathbf{N}\ \mathbf{M}] \in \partial \mathbf{H}(\mathbf{x}_1, \mathbf{x}_2)\}$*. Let* $\pi_2 \partial \mathbf{H}(\mathbf{x}_1, \mathbf{x}_2)$ *be maximal, i.e., each element of* $\pi_2 \partial \mathbf{H}(\mathbf{x}_1, \mathbf{x}_2)$ *is invertible.*

If $\mathbf{H}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{0}$*, then there exists* O_1*, a neighborhood of* $\mathbf{x}_1$*, and a locally Lipschitz continuous function* $\mathbf{G}$ *from* O_1 *to* $\mathbb{R}^m$ *such that* $\mathbf{G}(\mathbf{x}_1) = \mathbf{x}_2$ *and* $\mathbf{H}(\mathbf{u}, \mathbf{G}(\mathbf{u})) = \mathbf{0}$*, for all* $\mathbf{u} \in O_1$*. If* $\partial \mathbf{H}(\mathbf{x}_1, \mathbf{x}_2) = \{[(\mathbf{J}_1 \mathbf{H}(\mathbf{x}_1, \mathbf{x}_2)\ \mathbf{J}_2(\mathbf{H}(\mathbf{x}_1, \mathbf{x}_2)]\}$*, then* $\partial \mathbf{G}(\mathbf{x}_1) = \{-\mathbf{J}_2 \mathbf{H}(\mathbf{x}_1, \mathbf{x}_2)^{-1} \mathbf{J}_1 \mathbf{H}(\mathbf{x}_1, \mathbf{x}_2)\}$.

2.7 Piecewise Continuously Differentiable (PC^1) Functions

The results in this section can be found in [91] and [98].

Definition 2.7.1 (PC^1 **Functions).** *Let X be an open subset of $\mathbb{R}^n$. $\mathbf{F} : X \to \mathbb{R}^m$ is a piecewise continuously differentiable function on X, denoted by $\mathbf{F} \in PC^1(X)$, if $\mathbf{F}$ is a continuous function on X and for every $\mathbf{x} \in X$ there exists a neighborhood $O \subset X$ and a finite set of selection functions, $\{\mathbf{F}_i : O \to \mathbb{R}^m,\ \mathbf{F}_i \in C^1(O)\}_{i=1}^k$, such that for all $\mathbf{y} \in O$, $\mathbf{F}(\mathbf{y}) \in \{\mathbf{F}_i(\mathbf{y})\}_{i=1}^k$. Let $O_i = \{\mathbf{y} \in O : \mathbf{F}_i(\mathbf{y}) = \mathbf{F}(\mathbf{y})\}$ for each $i \in \{1, \ldots, k\}$. A selection function, $\mathbf{F}_i$, is essentially active at $\mathbf{x}$ if $\mathbf{x} \in \mathrm{cl}\,(\mathrm{int}\,(O_i))$. $\mathcal{I}(\mathbf{F}, \mathbf{x})$, the set of essentially active function indices at $\mathbf{x}$ is the set of indices $i \in \{1, \ldots, k\}$ such that $\mathbf{F}_i$ is essentially active at $\mathbf{x}$.*

2.7.1 Properties of PC^1 Functions

1. If $\mathbf{F} \in PC^1(X)$, then there exists a set of selection functions that are essentially active at $\mathbf{x} \in X$ (Proposition 4.1.1 in [98]).

2. If $\mathbf{F} \in PC^1(X)$, then $\mathbf{F}$ is locally Lipschitz continuous at all $\mathbf{x} \in X$. The Lipschitz constant is the maximum of the Lipschitz constants of the essentially active selection functions.

3. If $\mathbf{F} \in PC^1(X)$, then $\partial \mathbf{F}(\mathbf{x}) = \mathrm{conv}\,(\{\mathbf{J}\mathbf{F}_i(\mathbf{x}) : i \in \mathcal{I}(\mathbf{F}, \mathbf{x})\})$.

2.7.2 Implicit Function Theorem for PC^1 Functions

Definition 2.7.2 (Complete Coherent Orientation). *Let Y_1 and Y_2 be open subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. Let $\mathbf{y}_1 \in Y_1$ and $\mathbf{y}_2 \in Y_2$. Let $\mathbf{H} : Y_1 \times Y_2 \to \mathbb{R}^m$ be a PC^1 function*

, *and let* $\{\mathbf{H}_i : i \in \mathcal{I}(\mathbf{H}, (\mathbf{y}_1, \mathbf{y}_2))\}$ *be the set of essentially active selection functions at* $(\mathbf{y}_1, \mathbf{y}_2)$. *Let* $\Lambda(\mathbf{y}_1, \mathbf{y}_2)$ *be the set of all* $m \times m$ *matrices* $\mathbf{M}$ *with the property that there exist matrices* $\mathbf{M}_k \in \{\mathbf{J}_2\mathbf{H}_i(\mathbf{y}_1, \mathbf{y}_2) : i \in \mathcal{I}(\mathbf{H}, (\mathbf{y}_1, \mathbf{y}_2))\}$ *where* $k = 1, \ldots, m$ *such that the* k*th row of* $\mathbf{M}$ *coincides with the* k*th row of* $\mathbf{M}_k$. *Then* $\mathbf{H}$ *is completely coherently oriented with respect to* Y_2 *at* $(\mathbf{y}_1, \mathbf{y}_2)$ *if all matrices* $\mathbf{M} \in \Lambda(\mathbf{y}_1, \mathbf{y}_2)$ *have the same non-vanishing determinantal sign (Definition 16 in [91]).*

Theorem 2.7.3 (Implicit Function Theorem for PC^1 Functions). *Let* Y_1 *and* Y_2 *be open subsets of* $\mathbb{R}^n$ *and* $\mathbb{R}^m$, *respectively. Let* $\mathbf{y}_1 \in Y_1$ *and* $\mathbf{y}_2 \in Y_2$. *Let* $\mathbf{H} : Y_1 \times Y_2 \to \mathbb{R}^m$ *be a* PC^1 *function that is completely coherently oriented with respect to* Y_2 *at* $(\mathbf{y}_1, \mathbf{y}_2)$. *If* $\mathbf{H}(\mathbf{y}_1, \mathbf{y}_2) = \mathbf{0}$, *then there exists a neighborhood,* O, *of* $\mathbf{y}_1$ *and a* PC^1 *function,* $\mathbf{G} : O \to \mathbb{R}^m$ *such that* $\mathbf{G}(\mathbf{y}_1) = \mathbf{y}_2$ *and* $\mathbf{H}(\mathbf{z}, \mathbf{G}(\mathbf{z})) = \mathbf{0}$ *for all* $\mathbf{z} \in O$ *(Corollary 20 in [91]).*

2.8 Semismooth Functions

Semismooth and related functions comprise a group of functions for which nonsmooth optimization methods with provable convergence can be devised. Nonsmooth Newton methods exist to solve nonsmooth equations involving vector-valued semismooth functions.

2.8.1 Bouligand Differentiable Functions

The results in this section can be found in [35].

Definition 2.8.1 (The Bouligand Derivative). *Let* X *be an open subset of* $\mathbb{R}^n$. *Then* $\mathbf{F} : X \to \mathbb{R}^m$ *is Bouligand differentiable (B-differentiable) at* $\mathbf{x} \in X$ *if* $\mathbf{F}$ *is locally Lipschitz continuous and directionally differentiable at* $\mathbf{x}$. *The function* $\mathbf{F}'(\mathbf{x}; \cdot)$ *is called the Bouligand derivative (B-derivative) of* $\mathbf{F}$ *at* $\mathbf{x}$ *(Definition 3.1.2 in [35]).*

Theorem 2.8.2. *Let X be an open subset of $\mathbb{R}^m$. $\mathbf{F} : X \to \mathbb{R}^m$ be B-differentiable at $\mathbf{x} \in X$. Then the limit*

$$\lim_{\mathbf{y}\to\mathbf{x}} \frac{\mathbf{F}(\mathbf{y}) - \mathbf{F}(\mathbf{x}) - \mathbf{F}'(\mathbf{x}; \mathbf{y} - \mathbf{x})}{\|\mathbf{x} - \mathbf{y}\|} = \mathbf{0}$$

holds (Proposition 3.1.3 in [35]).

Theorem 2.8.3 (Chain Rule for Bouligand Differentiable Functions). *Let X be an open subset of $\mathbb{R}^n$. Let $\mathbf{F} : X \to \mathbb{R}^m$ and $\mathbf{G} : \mathbb{R}^m \to \mathbb{R}^p$ be B-differentiable at $\mathbf{x} \in X$ and $\mathbf{F}(\mathbf{x})$ respectively. Then $\mathbf{H} \equiv \mathbf{G} \circ \mathbf{F}$ is B-differentiable at $\mathbf{x}$ and the B-derivative is*

$$\mathbf{H}'(\mathbf{x}; \mathbf{d}) = \mathbf{G}'(\mathbf{F}(\mathbf{x}); \mathbf{F}'(\mathbf{x}; \mathbf{d})), \ \forall \mathbf{d} \in \mathbb{R}^n.$$

(Proposition 3.1.6 in [35]).

Properties of Bouligand Differentiable Functions

Let X be an open subset of $\mathbb{R}^m$ in this section.

1. If $\mathbf{F} : X \to \mathbb{R}^m$ is B-differentiable at $\mathbf{x} \in X$, then $\mathbf{F}'(\mathbf{x}; \cdot)$ is a Lipschitz function from $\mathbb{R}^n$ to $\mathbb{R}^m$.

2. If $\mathbf{F} : X \to \mathbb{R}^m$ is B-differentiable at $\mathbf{x} \in X$, and $\mathbf{F}'(\mathbf{x}; \cdot)$ is a linear function, then $\mathbf{F}$ is Fréchet differentiable at $\mathbf{x}$.

3. If $\mathbf{F} : X \to \mathbb{R}^m \in PC^1(X)$, then it is B-differentiable at all $\mathbf{x} \in X$.

4. Let X be a convex set and $f : X \to \mathbb{R}$ be a finite convex function on X. Then f is B-differentiable at all $\mathbf{x} \in X$.

2.8.2 Scalar-Valued Semismooth Functions

In the remainder of this section, let X be an open subset of $\mathbb{R}^n$ and $f : X \to \mathbb{R}$ be locally Lipschitz continuous at $\mathbf{x} \in X$.

Definition 2.8.4 (Scalar-valued Semismooth Function). *f is semismooth at $\mathbf{x} \in X$ if for each $\mathbf{d} \in \mathbb{R}^n$ and for all sequences $\{t_k\} \subset \mathbb{R}$, $\{\mathbf{v}_k\} \subset \mathbb{R}^n$ and $\{\mathbf{g}_k\} \subset \mathbb{R}^n$ such that $t_k \neq 0$, $\forall k$, $t_k \downarrow 0$, $\mathbf{v}_k/t_k \to \mathbf{0}$ and $\mathbf{g}_k \in \partial f(\mathbf{x} + t_k\mathbf{d} + \mathbf{v}_k)$, the sequence $\{\langle \mathbf{g}_k, \mathbf{d}\rangle\}$ has exactly one accumulation point [71]. If f is semismooth for all $\mathbf{x} \in X$, then it is a semismooth function.*

Definition 2.8.5 (Weakly Upper Semismooth Functions). *f is weakly upper semismooth [70] at $\mathbf{x}$ if for each $\mathbf{d} \in \mathbb{R}^n$ and for any sequences $\{t_k\} \subset \mathbb{R}$ and $\{\mathbf{g}_k\} \subset \mathbb{R}^n$ such that $t_k > 0$, $\forall k$, $t_k \downarrow 0$ and $\mathbf{g}_k \in \partial f(\mathbf{x} + t_k\mathbf{d})$ the following holds:*

$$\liminf_{k\to\infty} \langle \mathbf{g}_k, \mathbf{d}\rangle \geq \limsup_{t_k \downarrow 0} \frac{f(\mathbf{x} + t_k\mathbf{d}) - f(\mathbf{x})}{t_k}. \tag{2.8.1}$$

Definition 2.8.6 (Upper Semidifferentiable Functions). *f is upper semidifferentiable at $\mathbf{x}$ if and only if for all $\mathbf{d} \in \mathbb{R}^n$, for all sequences $\{t_k\} \subset \mathbb{R}$ and $\{\mathbf{g}_k : \mathbf{g}_k \in \partial f(\mathbf{x} + t_k\mathbf{d})\}$ such that $t_k > 0$, $\forall k$ and $t_k \downarrow 0$, there exist subsequences whose indices are in the set $K \subset \mathbb{N}$ such that*

$$\liminf_{\substack{k\to\infty \\ k\in K}} \frac{f(\mathbf{x} + t_k\mathbf{d}) - f(\mathbf{x})}{t_k} - \langle \mathbf{g}_k, \mathbf{d}\rangle \leq 0 \tag{2.8.2}$$

holds [18].

Properties of Scalar-Valued Semismooth Functions

Let X be an open subset of $\mathbb{R}^n$ in the remainder of this section.

1. Definition 2.8.4 and the fact that the generalized gradient is a locally bounded set-valued mapping imply that all sequences $\{\langle \mathbf{g}_k, \mathbf{d}\rangle\}$ as described converge to the same limit.

   Let L be the accumulation point mentioned in Definition 2.8.4. The sequence $\{\langle \mathbf{g}_k, \mathbf{d}\rangle\}$ is bounded because the generalized gradient is locally bounded. Assume there exists a subsequence that does not converge to L. By the Bolzano-Weierstrass Theorem (Theorem 2.42 in [96]), this subsequence has a converging subsequence. If the limit of this subsequence is not L, then the semismoothness assumption is violated. Hence all sequences $\{\langle \mathbf{g}_k, \mathbf{d}\rangle\}$, converge to L.

2. If f is semismooth at $\mathbf{x} \in X$ then $f'(\mathbf{x}; \mathbf{d})$ exists for $\forall \mathbf{d} \in \mathbb{R}^n$ and is equal to $\lim_{k\to\infty} \langle \mathbf{g}_k, \mathbf{d}\rangle$ for any sequences as described in Definition 2.8.4 (Lemma 2 in [71]).

3. If f is semismooth in a neighborhood of $\mathbf{x} \in X$, then $f'(\mathbf{x}; \mathbf{d}) = \lim_{t_k \downarrow 0} f'(\mathbf{x}+t_k\mathbf{d}; \mathbf{d})$ for any sequence $\{t_k\}$ such that $t_k > 0$ for all k and $t_k \to 0$. Note that due to semismoothness in a neighborhood of $\mathbf{x}$, for small enough t_k, $f'(\mathbf{x}+t_k\mathbf{d}; \mathbf{d})$ exists and is equal to $\langle \mathbf{g}_k, \mathbf{d}\rangle$ where $\mathbf{g}_k \in \partial f(\mathbf{x} + t_k\mathbf{d})$. By semismoothness $\{\langle \mathbf{g}_k, \mathbf{d}\rangle\}$ converges to a limit which is $f'(\mathbf{x}; \mathbf{d})$.

4. Let $f : X \to \mathbb{R}$ and $g : X \to \mathbb{R}$ be locally Lipschitz continuous and semismooth functions on X. Then $g + f$ and αg where $\alpha \in \mathbb{R}$ are semismooth functions [71].

5. Let $\mathbf{F} : X \to \mathbb{R}^n : \mathbf{y} \mapsto (f_1(\mathbf{y}), \ldots, f_m(\mathbf{y}))$ where $f_i : X \to \mathbb{R}, i = 1, \ldots, m$ are locally Lipschitz continuous and semismooth at $\mathbf{x} \in X$. Let $g : \mathbb{R}^m \to \mathbb{R}$ be locally Lipschitz continuous and semismooth at $\mathbf{F}(\mathbf{x})$. Then $g \circ \mathbf{F}$ is locally Lipschitz continuous and semismooth at $\mathbf{x}$ (Theorem 5 in [71]).

6. If $f : X \to \mathbb{R}$ is a semismooth function then it is strictly differentiable for all $\mathbf{x} \in X \backslash S$ where S is a measure zero subset of X [88].

7. If $f : X \to \mathbb{R}$ is a finite and convex function in a neighborhood of $\mathbf{x} \in X$, then it is semismooth at $\mathbf{x}$.

8. If $f : X \to \mathbb{R}$ is semismooth at $\mathbf{x} \in X$, then it is weakly upper semismooth at $\mathbf{x}$ [70].

9. If $f : X \to \mathbb{R}$ is weakly upper semismooth at $\mathbf{x} \in X$, then it is directionally differentiable at $\mathbf{x}$ [70].

10. If f is weakly upper semismooth at $\mathbf{x}$, then it is upper semidifferentiable at $\mathbf{x}$. If $f : X \to \mathbb{R}$ is upper semidifferentiable at $\mathbf{x} \in X$, and is directionally differentiable at $\mathbf{x}$, then it is weakly upper semismooth at $\mathbf{x}$ [18].

11. Upper semidifferentiability is a sufficient condition for line search algorithms in nonsmooth optimization methods to terminate finitely [54, 66].

12. If $\mathbf{F} : X \to \mathbb{R} \in PC^1(X)$, then it is a semismooth function [35].

2.8.3 Vector-valued Semismooth Functions

The concept of semismoothness is extended to functions $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ using the generalized Jacobian [89].

In the remainder of this section, let X be an open subset of $\mathbb{R}^n$ and $\mathbf{F} : X \to \mathbb{R}^m$ be a locally Lipschitz continuous function.

Definition 2.8.7 (Vector-valued Semismooth Functions). *Let* $\mathbf{d} \in \mathbb{R}^n$, $\{t_k\}$ *and* $\{\mathbf{v}_k\}$ *be any sequences such that* $t_k \in \mathbb{R}$, $t_k > 0$ $\forall k$, $\mathbf{v}_k \in \mathbb{R}^n$, $t_k \downarrow 0$ *and* $\mathbf{v}_k/t_k \to \mathbf{0}$. *Let* $\mathbf{x}_k = \mathbf{x} + t_k\mathbf{d} + \mathbf{v}_k$ *and* $\mathbf{d}_k = \mathbf{d} + \mathbf{v}_k/t_k$. $\mathbf{F} : X \to \mathbb{R}^m$ *is semismooth at* $\mathbf{x} \in X$ *if for each* $\mathbf{d} \in \mathbb{R}^n$ *and for all sequences* $\{\mathbf{x}_k\}$ *and* $\{\mathbf{V}_k\}$ *such that for all* k, $\mathbf{V}_k \in \partial\mathbf{F}(\mathbf{x}_k)$, *the sequence* $\{\mathbf{V}_k\mathbf{d}_k\}$ *has exactly one accumulation point [89].* $\mathbf{F}$ *is a semismooth function if* $\mathbf{F}$ *is semismooth at all* $\mathbf{x} \in X$.

Properties of Vector-valued Semismooth Functions

1. As in the scalar-valued semismooth function case, all sequences $\{\mathbf{V}_k\mathbf{d}_k\}$ converge to the same limit because the generalized Jacobian is a locally bounded set-valued map.

2. If $\mathbf{F} : X \to \mathbb{R}^m$ is semismooth at $\mathbf{x}$, then the directional derivative exists for all $\mathbf{d} \in \mathbb{R}^n$ and $\mathbf{F}'(\mathbf{x}, \mathbf{d}) = \mathbf{V}\mathbf{d}$ where $\mathbf{V} \in \partial\mathbf{F}(\mathbf{x})$ [89].

3. If $\mathbf{F} : X \to \mathbb{R}^m$ is semismooth at $\mathbf{x}$

$$\lim_{k\to\infty} \frac{\mathbf{F}(\mathbf{x} + t_k\mathbf{d}_k) - \mathbf{F}(\mathbf{x})}{t_k} = \lim_{k\to\infty} \mathbf{V}_k\mathbf{d}_k = \mathbf{F}'(\mathbf{x}; \mathbf{d}) \tag{2.8.3}$$

   holds [89].

4. $\mathbf{F} : X \to \mathbb{R}^m$ is semismooth at $\mathbf{x}$ if and only if each element of $\mathbf{F}$ is semismooth [35].

5. Let $\mathbf{F} : X \to \mathbb{R}^m$ be semismooth at $\mathbf{x}$ and $\mathbf{G} : \mathbb{R}^m \to \mathbb{R}^p$ be semismooth at $\mathbf{F}(\mathbf{x})$. Then $\mathbf{G} \circ \mathbf{F}$ is semismooth at $\mathbf{x}$ [35].

6. If $\mathbf{F} : X \to \mathbb{R}^m \in PC^1(X)$, then it is a semismooth function [35].

2.8.4 A Restricted Definition of Semismoothness

In the subsequent chapters, a restricted definition of semismoothness is used. In order to be semismooth at $\mathbf{x} \in X$, $\mathbf{F} : X \to \mathbb{R}^m$ needs to be Bouligand differentiable on an open neighborhood of $\mathbf{x}$ in addition to satisfying conditions in Definitions 2.8.4 and 2.8.7. This restriction does not affect the results concerning semismooth functions presented so far. This restricted definition of semismoothness is automatically satisfied by the data used for this work. The reason for this restriction is to better align the exposition with key results from the literature on which the results in this document depend.

An equivalent definition under the stated restriction of semismoothness using Bouligand derivatives is as follows:

Definition 2.8.8. *Let X be an open subset of $\mathbb{R}^n$. Let $\mathbf{F} : X \to \mathbb{R}^m$ be a locally Lipschitz continuous function on O, a neighborhood of $\mathbf{x} \in X$. Let $\mathbf{F}$ be a directionally differentiable function on O. $\mathbf{F}$ is semismooth at $\mathbf{x}$ if there exists a function $\Delta : (0, +\infty) \to [0, +\infty)$ such that $\lim_{z \downarrow 0} \Delta(z) = 0$ and for any $\mathbf{y} \in O \backslash \{\mathbf{x}\}$*

$$\frac{\mathbf{F}'(\mathbf{y}; \mathbf{y} - \mathbf{x}) - \mathbf{F}'(\mathbf{x}; \mathbf{y} - \mathbf{x})}{\|\mathbf{y} - \mathbf{x}\|} \leq \Delta(\|\mathbf{y} - \mathbf{x})\|)$$

holds [35].

The following theorem establishes the connection between previous definitions of semismoothness and Definition 2.8.8.

Theorem 2.8.9. *Let $\mathbf{F} : X \to \mathbb{R}^m$ be a locally Lipschitz continuous and B-differentiable function on O, a neighborhood of $\mathbf{x}$. Then the following statements are equivalent (Theorem 7.4.3 in [35]).*

1. *$\mathbf{F}$ is semismooth at $\mathbf{x}$.*

2. *For $\mathbf{y} \in O \backslash \{\mathbf{x}\}$,*

$$\lim_{\mathbf{y} \to \mathbf{x}} \frac{\mathbf{F}'(\mathbf{y}; \mathbf{y} - \mathbf{x}) - \mathbf{F}'(\mathbf{x}; \mathbf{y} - \mathbf{x})}{\|\mathbf{y} - \mathbf{x}\|} = 0$$

   holds.

3. *Let $\{\mathbf{x}_k\} \in O \backslash \{\mathbf{x}\}$ be any sequence such that $\lim_{k \to \infty} \mathbf{x}_k \to \mathbf{x}$. Let $\{\mathbf{V}_k\}$ be any sequence such that for all k, $\mathbf{V}_k \in \partial \mathbf{F}(\mathbf{x}_k)$. Then for all sequences $\{\mathbf{x}_k\}$ and $\{\mathbf{V}_k\}$ as described,*

$$\lim_{k \to \infty} \frac{\mathbf{F}'(\mathbf{x}; \mathbf{x}_k - \mathbf{x}) - \mathbf{V}_k(\mathbf{x}_k - \mathbf{x})}{\|\mathbf{x} - \mathbf{x}_k\|} = 0$$

holds.

4. *For all sequences* $\{\mathbf{x}_k\}$ *and* $\{\mathbf{V}_k\}$ *as described in the previous item,*

$$\lim_{k\to\infty} \frac{\mathbf{F}(\mathbf{x}_k) - \mathbf{V}_k(\mathbf{x}_k - \mathbf{x}) - \mathbf{F}(\mathbf{x})}{\|\mathbf{x} - \mathbf{x}_k\|} = 0$$

holds.

Theorem 2.8.9 elucidates the most important properties of semismooth functions. Due to Bouligand differentiability, $\mathbf{F}'(\mathbf{x}; \mathbf{y} - \mathbf{x})$ provides a good approximation of $\mathbf{F}(\mathbf{y})$ for all $\mathbf{y}$ sufficiently close to $\mathbf{x}$. In addition, $\mathbf{F}'(\mathbf{x}; \mathbf{y} - \mathbf{x})$ can be approximated well using an element of $\partial\mathbf{F}(\mathbf{y})$.

2.8.5 The Linear Newton Approximation

Definition 2.8.10 (Newton Approximation). *Let* X *be an open subset of* $\mathbb{R}^n$. *Let* $\mathbf{F} : X \to \mathbb{R}^m$ *be a locally Lipschitz function on* X. $\mathbf{F}$ *has a Newton approximation at a point* $\mathbf{x} \in X$ *if there exists* $O \subset X$, *a neighborhood of* $\mathbf{x}$ *and a function* $\Delta : (0, +\infty) \to [0, +\infty)$ *with* $\lim_{z\to 0^+} \Delta(z) = 0$ *such that for every* $\mathbf{y} \in O$, *there is a family of functions* $\mathcal{A}(\mathbf{y})$ *called a Newton approximation, whose members map* $\mathbb{R}^n$ *to* $\mathbb{R}^m$ *and satisfy the following two properties:*

1. $A(\mathbf{y}, \mathbf{0}) = \mathbf{0}$ *for every* $A(\mathbf{y}, \cdot) \in \mathcal{A}(\mathbf{y})$.
2. *For any* $\mathbf{y} \in O \backslash \{\mathbf{x}\}$ *and for any* $A(\mathbf{y}, \cdot) \in \mathcal{A}(\mathbf{y})$

$$\frac{\|\mathbf{F}(\mathbf{y}) + A(\mathbf{y}, \mathbf{x} - \mathbf{y}) - \mathbf{F}(\mathbf{x})\|}{\|\mathbf{y} - \mathbf{x}\|} \leq \Delta(\|\mathbf{y} - \mathbf{x}\|)$$

holds (Definition 7.2.2 in [35]).

Definition 2.8.11 (The Linear Newton Approximation). *Let* X *be an open subset of* $\mathbb{R}^n$. *Let* $\mathbf{F} : X \to \mathbb{R}^m$ *be a locally Lipschitz function on* X. *Let* $\Gamma : X \rightrightarrows \mathbb{R}^{m\times n}$ *be an upper*

semicontinuous set-valued map at $\mathbf{x} \in X$ *and* $\Gamma(\mathbf{y})$ *be a compact set for all* $\mathbf{y} \in X$. *Assume there exists a function* Δ *as defined in Definition 2.8.10. If for any* $\mathbf{y} \in X \backslash \{\mathbf{x}\}$ *and for any* $\mathbf{M} \in \Gamma(\mathbf{y})$,

$$\frac{\|\mathbf{F}(\mathbf{y}) + \mathbf{M}(\mathbf{x} - \mathbf{y}) - \mathbf{F}(\mathbf{x})\|}{\|\mathbf{y} - \mathbf{x}\|} \leq \Delta(\|\mathbf{y} - \mathbf{x}\|)$$

holds, then $\Gamma(\mathbf{x})$ *is a linear Newton approximation of* $\mathbf{F}$ *at* $\mathbf{x}$ *(Definition 7.5.13 in [35]).*

It is possible that there exists more than one linear Newton approximation for a given function unlike the generalized Jacobian and generalized gradient. The linear Newton approximation construct is used to solve nonsmooth equations and its properties suffice to devise methods to solve these equations using Newton-type methods. The nonuniqueness of the Newton approximation helps overcome cases where the generalized Jacobian cannot be computed easily when solving nonsmooth equations and allows the development of different methods with varying properties to solve these equations. The linear Newton approximation by itself does not carry useful information for optimization purposes. The subsequent developments couple the generalized Jacobian and the Newton approximation to overcome situations where an element of the generalized Jacobian cannot not be computed to devise numerical optimization methods.

Similar to the generalized Jacobian, the linear Newton approximation has calculus and chain rules. Unlike the generalized Jacobian, these rules always involve equalities and not inclusions. In this respect, the linear Newton approximation behaves like the Jacobian.

Theorem 2.8.12 (Chain Rule for the Linear Newton Approximation). *Let* X *be an open subset of* $\mathbb{R}^n$. *Let* $\mathbf{F} : X \to \mathbb{R}^m$ *and* $\mathbf{G} : \mathbb{R}^m \to \mathbb{R}^p$ *be locally Lipschitz continuous at* $\mathbf{x} \in X$ *and* $\mathbf{F}(\mathbf{x})$, *respectively. Let* $\Gamma\mathbf{F}$ *and* $\Gamma\mathbf{G}$ *be the linear Newton approximations of* $\mathbf{F}$

and $\mathbf{G}$ *at* $\mathbf{x}$ *and* $\mathbf{F}(\mathbf{x})$*, respectively. Then*

$$\Gamma\mathbf{H} : X \rightrightarrows \mathbb{R}^{p\times n} : \mathbf{y} \mapsto \{\mathbf{AB} : \mathbf{A} \in \Gamma\mathbf{G}(\mathbf{F}(\mathbf{y})),\ \mathbf{B} \in \Gamma\mathbf{F}(\mathbf{y})\} \tag{2.8.4}$$

is a linear Newton approximation of $\mathbf{H} \equiv \mathbf{G} \circ \mathbf{F}$ *at* $\mathbf{x} \in X$*. (Theorem 7.5.17 in [35])* [†].

Properties of Linear Newton Approximation

1. Let X be an open subset of $\mathbb{R}^n$. Let $\mathbf{F} : X \to \mathbb{R}^m$ be locally Lipschitz function on X. Let $\Gamma\mathbf{F}$ be the linear Newton approximation of $\mathbf{F}$ at $\mathbf{x}$. Then $\mathcal{S} : X \rightrightarrows \mathbb{R}^{m\times n} : \mathbf{y} \mapsto \text{conv}\,(\Gamma\mathbf{F}(\mathbf{y}))$ is a linear Newton approximation of $\mathbf{F}$ at $\mathbf{x}$ (Lemma 9 in [81]).

2. Let $\mathbf{F} : X \to \mathbb{R}^m$ be locally Lipschitz continuous and semismooth at $\mathbf{x} \in X$. Then $\partial\mathbf{F}$ is a linear Newton approximation of $\mathbf{F}$ at $\mathbf{x}$ (Proposition 7.5.16 in [35]).

Calculus Rules for the Linear Newton Approximation

1. Let X be an open subset of $\mathbb{R}^n$. Let $\mathbf{F} : X \to \mathbb{R}^m$ and $\mathbf{G} : X \to \mathbb{R}^m$ be locally Lipschitz continuous at $\mathbf{x} \in X$. Let $\Gamma\mathbf{F} : X \rightrightarrows \mathbb{R}^{m\times n}$ and $\Gamma\mathbf{G} : X \rightrightarrows \mathbb{R}^{m\times n}$ be the linear Newton approximations of $\mathbf{F}$ and $\mathbf{G}$ at $\mathbf{x}$, respectively.

   Then $\Gamma\mathbf{H} : X \rightrightarrows \mathbb{R}^{m\times n} : \mathbf{y} \mapsto \alpha_1\Gamma\mathbf{F}(\mathbf{y}) + \alpha_2\Gamma\mathbf{G}(\mathbf{y})$ is a linear Newton approximation of the function $\alpha_1\mathbf{F} + \alpha_2\mathbf{G}$ at $\mathbf{x} \in X$ where α_1 and α_2 are scalars.

2. Let X be an open subset of $\mathbb{R}^n$. Let $f : X \to \mathbb{R}$ and $g : X \to \mathbb{R}$ be locally Lipschitz continuous at $\mathbf{x}$. Let $\Gamma f : X \rightrightarrows \mathbb{R}^{1\times n}$ and $\Gamma g : X \rightrightarrows \mathbb{R}^{1\times n}$ be the linear Newton approximations of f and g at $\mathbf{x}$ respectively. Then the following rules hold:

   (a) $\Gamma h : X \rightrightarrows \mathbb{R}^{1\times n} : \mathbf{y} \mapsto f(\mathbf{y})\Gamma f(\mathbf{y}) + g(\mathbf{y})\Gamma g(\mathbf{y})$ is a linear Newton approximation of the function $f + g$ at $\mathbf{x} \in X$.

[†]Theorem 7.5.17 considers the function $\mathbf{G} \circ \mathbf{F}$ where $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbf{G} : \mathbb{R}^m \to \mathbb{R}^n$. The proof holds for the general case where $\mathbf{G} : \mathbb{R}^m \to \mathbb{R}^p$.

(b) $\Gamma h : X \rightrightarrows \mathbb{R}^{2\times n} : \mathbf{y} \mapsto \Gamma f(\mathbf{y}) \times \Gamma g(\mathbf{y})$ is a linear Newton approximation of the function (f, g) at $\mathbf{x} \in X$.

(c) If $g(\mathbf{x}) \neq 0$, then

$$\Gamma h : X \rightrightarrows \mathbb{R}^{1\times n} : \mathbf{y} \mapsto \frac{g(\mathbf{y})\Gamma f(\mathbf{y}) - f(\mathbf{y})\Gamma g(\mathbf{y})}{g^2(\mathbf{y})}$$

is a linear Newton approximation of f/g at $\mathbf{x}$.

(d) If $\mathbf{y} \in X$, let

$$\Gamma(\mathbf{y}) = \begin{cases} \Gamma f(\mathbf{y}) & \text{if } f(\mathbf{y}) > g(\mathbf{y}) \\ \Gamma f(\mathbf{y}) \cup \Gamma g(\mathbf{y}) & \text{if } f(\mathbf{y}) = g(\mathbf{y}) \\ \Gamma g(\mathbf{y}) & \text{if } g(\mathbf{y}) > f(\mathbf{y}). \end{cases} \tag{2.8.5}$$

Then $\Gamma h : X \rightrightarrows \mathbb{R}^{1\times n} : \mathbf{y} \mapsto \Gamma(\mathbf{y})$ is a linear Newton approximation of the function $\max(f, g)$ at $\mathbf{x}$.

2.9 Examples

Example 2.9.1. Let $\mathcal{S} : \mathbb{R} \rightrightarrows \mathbb{R}$ be:

$$\mathcal{S}(x) = \begin{cases} [-1, 1] & \text{if } x = 0, \\ \{1\} & \text{if } x > 0, \\ \{-1\} & \text{if } x < 0. \end{cases}$$

$\mathcal{S}$ is upper semicontinuous at 0, but not lower semicontinuous. Let $O = (-0.5, 0.5)$. Even though $\mathcal{S}(0) \cap O \neq \emptyset$, $\mathcal{S}(y) \cap O = \emptyset$ for all $y \in \mathbb{R}\backslash\{0\}$. $\mathcal{S}$ is a locally bounded and closed set-valued map.

Example 2.9.2. Let $\mathcal{S} : \mathbb{R} \rightrightarrows \mathbb{R}$ be

$$\mathcal{S}(x) = \begin{cases} (-1,1) & \text{if } x = 0, \\ \{1\} & \text{if } x > 0, \\ \{-1\} & \text{if } x < 0. \end{cases}$$

$\mathcal{S}$ is not upper semicontinuous at zero. $\mathcal{S}$ is a locally bounded but not a closed set-valued map.

Example 2.9.3. Let $\mathcal{S} : \mathbb{R} \rightrightarrows \mathbb{R}$ be

$$\mathcal{S}(x) = \begin{cases} [-1,+\infty) & \text{if } x = 0, \\ \{1\} & \text{if } x > 0, \\ \{-1\} & \text{if } x < 0. \end{cases}$$

$\mathcal{S}$ is upper semicontinuous at 0, but not lower semicontinuous (see Example 2.9.1). $\mathcal{S}$ is a closed but not locally bounded set-valued map.

Example 2.9.4. Let $\mathcal{S} : \mathbb{R} \rightrightarrows \mathbb{R}$ be

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

$\mathcal{S}$ is not upper semicontinuous at zero but it is lower semicontinuous at zero. It is locally bounded at zero, but not closed at zero.

Example 2.9.5. Let $\mathcal{S} : \mathbb{R} \rightrightarrows \mathbb{R}$ be

$$\mathcal{S}(x) = \begin{cases} [1,-1] & \text{if } x = 0, \\ \emptyset & \text{if } 2 > |x| > 0, \\ \{1\} & \text{if } |x| \geq 2. \end{cases}$$

$\mathcal{S}$ is upper semicontinuous at zero, but is not lower semicontinuous at zero. It is locally bounded at zero. It is closed at zero.

Example 2.9.6. This example proves the equivalence of the two definitions of strict differentiability in Definition 2.2.5.

Let $\mathbf{f} : X \to \mathbb{R}^m$ where X is an open subset of $\mathbb{R}^n$. First assume that

$$\lim_{(\mathbf{y},\, t)\to(\mathbf{x},0^+)} \frac{\mathbf{f}(\mathbf{y}+t\mathbf{v}) - \mathbf{f}(\mathbf{y}) - t\mathbf{A}(\mathbf{x})\mathbf{v}}{t} = 0 \tag{2.9.1}$$

holds for all $\mathbf{v}$ and the convergence is uniform for $\mathbf{v}$ in compact sets.

Let $\{\bar{\mathbf{v}}_k\} \in \mathbb{R}^n$ be a sequence such that $\bar{\mathbf{v}}_k \to \mathbf{0}$ and for all k, $\bar{\mathbf{v}}_k \neq \mathbf{0}$. Let $t_k = \|\bar{\mathbf{v}}_k\|$, $t_k\mathbf{v}_k = \bar{\mathbf{v}}_k$, and $\mathbf{v}_k = \mathbf{v}_1 + \mathbf{w}_k$. Note that $\mathbf{v}_k$ is a unit vector. Let $\{\mathbf{y}_k\} \in \mathbb{R}^n$ be such that $\lim_{k\to\infty} \mathbf{y}_k = \mathbf{x}$ and $\mathbf{y}_k \neq \mathbf{x}$ for all k.

Then

$$\frac{\|\mathbf{f}(\mathbf{y}_k + \bar{\mathbf{v}}_k) - \mathbf{f}(\mathbf{y}_k) - \mathbf{A}(\mathbf{x})\bar{\mathbf{v}}_k\|}{\|\bar{\mathbf{v}}_k\|} =$$
$$\frac{\|\mathbf{f}(\mathbf{y}_k + t_k(\mathbf{v}_1 + \mathbf{w}_k)) - \mathbf{f}(\mathbf{y}_k) - t_k\mathbf{A}(\mathbf{x})(\mathbf{v}_1 + \mathbf{w}_k)\|}{t_k} \leq$$
$$\frac{\|\mathbf{f}(\mathbf{y}_k + t_k\mathbf{v}_1) - \mathbf{f}(\mathbf{y}_k) - t_k\mathbf{A}(\mathbf{x})(\mathbf{v}_1)\|}{t_k} + \frac{\|\mathbf{f}(\mathbf{y}_k + t_k(\mathbf{v}_1 + \mathbf{w}_k)) - \mathbf{f}(\mathbf{y}_k + t_k\mathbf{v}_1) - t_k\mathbf{A}(\mathbf{x})(\mathbf{w}_k)\|}{t_k}.$$

The first term converges to zero by assumptions as $k \to \infty$. Note that $\mathbf{v}_1 + \mathbf{w}_k$ and $\mathbf{v}_1$ are

elements of the set $C = \{\mathbf{u} : \|\mathbf{u}\| = 1,\ \mathbf{u} \in \mathbb{R}^n\}$ which is a compact set. In addition,

$$\frac{\|\mathbf{f}(\mathbf{y}_k + t_k(\mathbf{v}_1 + \mathbf{w}_k)) - \mathbf{f}(\mathbf{y}_k + t_k\mathbf{v}_1) - t_k\mathbf{A}(\mathbf{x})(\mathbf{w}_k)\|}{t_k} =$$
$$\left\| \frac{\mathbf{f}(\mathbf{y}_k + t_k(\mathbf{v}_1 + \mathbf{w}_k)) - \mathbf{f}(\mathbf{y}_k) - t_k\mathbf{A}(\mathbf{x})(\mathbf{v}_1 + \mathbf{w}_k)}{t_k} - \frac{\mathbf{f}(\mathbf{y}_k + t_k\mathbf{v}_1) - \mathbf{f}(\mathbf{y}_k) - t_k\mathbf{A}(\mathbf{x})(\mathbf{v}_1)}{t_k} \right\|.$$

This last quantity can be made arbitrarily small for all $\mathbf{v}_1 + \mathbf{w}_k \in C$ by picking a large enough k due to uniform convergence on compact sets. Hence

$$\lim_{(\mathbf{y},\mathbf{v})\to(\mathbf{x},\mathbf{0})} \frac{\mathbf{f}(\mathbf{y}+\mathbf{v}) - \mathbf{f}(\mathbf{y}) - \mathbf{A}(\mathbf{x})\mathbf{v}}{\|\mathbf{v}\|} = 0.$$

Conversely, assume

$$\lim_{(\mathbf{y},\mathbf{v})\to(\mathbf{x},\mathbf{0})} \frac{\mathbf{f}(\mathbf{y}+\mathbf{v}) - \mathbf{f}(\mathbf{y}) - \mathbf{A}(\mathbf{x})\mathbf{v}}{\|\mathbf{v}\|} = 0 \tag{2.9.2}$$

holds.

Let $\mathbf{v}_0 \in \mathbb{R}^n\backslash\{\mathbf{0}\}$. Let $\mathbf{v} = t\mathbf{v}_0$ where $t \in \mathbb{R}$ and $t > 0$. Then (2.9.2) becomes

$$\lim_{(\mathbf{y},t)\to(\mathbf{x},\mathbf{0}^+)} \frac{\mathbf{f}(\mathbf{y}+t\mathbf{v}_0) - \mathbf{f}(\mathbf{y}) - t\mathbf{A}(\mathbf{x})\mathbf{v}_0}{t\|\mathbf{v}_0\|} = 0 \tag{2.9.3}$$

which implies (2.9.1) since $\|\mathbf{v}_0\|$ is a positive constant.

Let $\mathbf{v}_1 \in \mathbb{R}^n\backslash\{\mathbf{0}\}$ and $\mathbf{v}_2 \in \mathbb{R}^n\backslash\{\mathbf{0}\}$. Assume $\mathbf{v}_1 \in C$ and $\mathbf{v}_2 \in C$ where C is a compact subset of $\mathbb{R}^n$. Consider

$$\frac{\mathbf{f}(\mathbf{y}+t\mathbf{v}_1) - \mathbf{f}(\mathbf{y}) - t\mathbf{A}(\mathbf{x})\mathbf{v}_1}{t} - \frac{\mathbf{f}(\mathbf{y}+t\mathbf{v}_2) - \mathbf{f}(\mathbf{y}) - t\mathbf{A}(\mathbf{x})\mathbf{v}_2}{t} =$$
$$\frac{\mathbf{f}(\mathbf{y}+t\mathbf{v}_1) - \mathbf{f}(\mathbf{y}+t\mathbf{v}_2) - t\mathbf{A}(\mathbf{x})(\mathbf{v}_1 - \mathbf{v}_2)}{t} =$$
$$\frac{\mathbf{f}(\bar{\mathbf{y}} + t(\mathbf{v}_1 - \mathbf{v}_2)) - \mathbf{f}(\bar{\mathbf{y}}) - t\mathbf{A}(\mathbf{x})(\mathbf{v}_1 - \mathbf{v}_2)}{t}.$$

Since (2.9.2) holds, then for any $\epsilon > 0$,

$$\left\| \frac{\mathbf{f}(\bar{\mathbf{y}} + t(\mathbf{v}_1 - \mathbf{v}_2)) - \mathbf{f}(\bar{\mathbf{y}}) - t\mathbf{A}(\mathbf{x})(\mathbf{v}_1 - \mathbf{v}_2)}{t} \right\| \leq 2\epsilon M$$

holds for t small enough and $\bar{\mathbf{y}}$ close enough to $\mathbf{x}$ where M is the bound on the magnitudes of the elements of C. This condition holds for any $\mathbf{v}_1$ and $\mathbf{v}_2$ in C. Hence (2.9.1) converges uniformly for $\mathbf{v}$ in a compact set.

Example 2.9.7. Let $h : \mathbb{R} \to \mathbb{R}$ be

$$h(x) = \begin{cases} x^2 \sin(1/x^2) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

h is a differentiable function and its derivative is

$$\nabla h(x) = \begin{cases} 2x \sin(1/x^2) - \frac{2}{x} \cos(1/x^2) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

However, h is not locally Lipschitz continuous at zero. Let $n = 1, 3, \ldots, \infty$ and $x_n = \sqrt{\frac{2}{\pi n}}$. Then $h(x_n) \in \{-\frac{2}{\pi n}, \frac{2}{\pi n}\}$. Let $x_{n+1} = \sqrt{\frac{2}{\pi(n+1)}}$. Note that $h(x_{n+1}) = 0$, $|h(x_n) - h(x_{n+1})| = \frac{2}{\pi n}$ and $|x_n - x_{n+1}| = \sqrt{\frac{2}{\pi}}(\frac{1}{\sqrt{n}} - \frac{1}{\sqrt{n+1}}) = \sqrt{\frac{2}{\pi}} \frac{1}{(\sqrt{n^2+n})(\sqrt{n+1}+\sqrt{n})}$. The ratio $|h(x_n) - h(x_{n+1})|/|x_n - x_{n+1}|$ is $\sqrt{\frac{2}{\pi}}\sqrt{(1+\frac{1}{n})}(\sqrt{n+1}+\sqrt{n})$. This ratio goes to infinity as $n \to \infty$ which shows that h is not locally Lipschitz continuous in a neighborhood of zero.

Example 2.9.8. Let $g : \mathbb{R} \to \mathbb{R}$ be

$$g(x) = \begin{cases} x^2 \sin(1/x) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

This function is differentiable at zero. The derivative is

$$\nabla g(x) = \begin{cases} 2x\sin(1/x) - \cos(1/x) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

g is locally Lipschitz continuous at zero. Let x_1 and x_2 be two points in an ϵ neighborhood of zero. By the mean value theorem for differentiable functions there exists an $x_3 \in [-\epsilon, \epsilon]$ such that $g(x_1) - g(x_2) = \nabla g(x_3)(x_1 - x_2)$. ∇g is bounded. Let $K = \sup_{y \in [-\epsilon,\epsilon]} \{\nabla g(y)\}$. Then $|g(x_1) - g(x_2)| = K|x_1 - x_2|$ holds on $(-\epsilon, \epsilon)$ and g is locally Lipschitz continuous at zero.

The generalized gradient of g obtained using Definition 2.6.3 is:

$$\partial g(x) = \begin{cases} \{2x\sin(1/x) - \cos(1/x)\} & \text{if } x \neq 0, \\ [-1, 1] & \text{if } x = 0. \end{cases}$$

The generalized gradient is not a singleton at zero where the function is differentiable. Hence g is not strictly differentiable at zero.

$g^o(0; v) = |v|$ because by definition, $\forall v \in \mathbb{R}$, $g^o(0; v) = \max_{\zeta \in \partial g(0)} \{\langle \zeta, v \rangle\}$. Since g is differentiable at zero, it is directionally differentiable at zero and $g'(0; v) = 0$. Therefore, g is not regular at zero.

Let $n = 1, \ldots, \infty$ and $x_n = \frac{1}{\pi n}$. Let $g_n \in \partial g(x_n)$ and $d = 1$. The sequence $\{\langle g_n, d \rangle\}$ is not convergent because $\langle g_n, d \rangle$ is 1 if n is odd and -1 if n is even. Hence, g is not semismooth at zero.

g is not weakly upper semismooth at zero. For the aforementioned sequence, the left-hand side of (2.8.1) is -1 and the right-hand side is 0 due to the existence of the derivative at 0.

g is not upper semidifferentiable at zero. Let $n = 1, 3, \ldots, \infty$ and $x_n = \frac{1}{\pi n}$. Let $g_n \in \partial g(x_n)$ and $d = 1$. Then the limit in (2.8.2) is 1.

Example 2.9.9. Let $f : \mathbb{R} \to \mathbb{R}$ be [18]

$$f(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ x \sin(\ln(\ln(1/x))) & \text{if } 0 < x \leq 1/2, \\ \frac{1}{2} \sin(\ln(\ln(1/2))) & \text{if } \frac{1}{2} < x. \end{cases}$$

f is locally Lipschitz continuous at all $x \in \mathbb{R}\backslash\{0\}$. In order to determine local Lipschitz continuity at zero, the following three cases are analyzed separately where $z, y \in (-\epsilon, \epsilon)$ and $\epsilon < 1/2$.

1. Case $z > 0$, $y \leq 0$. In this case, $|f(y) - f(z)| = |f(z)| \leq |z|$ because $|\sin(u)| \leq 1$. Since $|z - y| \geq |z|$, $|f(y) - f(z)| \leq |z - y|$ holds.

2. Case $z \leq 0$, $y \leq 0$. $0 = |f(y) - f(z)| \leq |z - y|$ holds trivially.

3. Case $z > 0$, $y > 0$. f is continuously differentiable on $(\epsilon, 0)$. The derivative is $\nabla f(x) = \sin(\ln(\ln(1/x))) - \cos(\ln(\ln(1/x)))(1/\ln(1/x))$. Hence $K = 2$ is a Lipschitz constant for f on this interval.

As a result, f is locally Lipschitz continuous at zero with Lipschitz constant $K = 2$.

f is not directionally differentiable at zero. For $0 < x < \epsilon$, consider the difference

$$\frac{x \sin(\ln(\ln(1/x))) - 0}{x - 0} = \sin(\ln(\ln(1/x))).$$

This difference does not converge to a limit as $x \to 0$ and therefore f is not directionally differentiable at zero. As a result it is not semismooth nor weakly upper semismooth at zero.

However, f is upper semidifferentiable at zero. Let the sequence $\{x_k\}$ be such that for all k, $0 < x_k < 1/2$ and $\lim_{k\to\infty} x_k = 0$. Let $d = 1$. Note that $\partial f(x_k) = \{\sin(\ln(\ln(1/x_k))) -$

$\cos(\ln(\ln(1/x_k)))(1/\ln(1/x_k))\}$. At zero, the limit (2.8.2) becomes

$$\lim_{k\to\infty} \frac{x_k \sin(\ln(\ln(1/x_k))) - 0}{x_k - 0} - \sin(\ln(\ln(1/x_k))) + \cos(\ln(\ln(1/x_k)))(1/\ln(1/x_k)) =$$

$$\lim_{k\to\infty} \cos(\ln(\ln(1/x_k)))(1/\ln(1/x_k)) = 0.$$

Since the above holds for any $d > 0$ and the conditions for upper semidifferentiability holds trivially at zero if $d < 0$, f is upper semidifferentiable at zero.

Example 2.9.10. Let $f : \mathbb{R} \to \mathbb{R} : x \mapsto 1 - e^{|x|}$. Note that $f \in PC^1(\mathbb{R})$ with selection functions; $f_1 : \mathbb{R} \to \mathbb{R} : x \mapsto 1 - e^x$ and $f_2 : \mathbb{R} \to \mathbb{R} : x \mapsto 1 - e^{-x}$. Therefore f is a locally Lipschitz continuous and semismooth function.

The generalized gradient of f is

$$\partial f(x) = \begin{cases} \{-e^x\} & \text{if } x > 0, \\ \{e^{-x}\} & \text{if } x < 0, \\ [-1, 1] & \text{if } x = 0. \end{cases}$$

In order to determine whether the function is regular at zero, the generalized directional derivative

$$f^o(0; v) = \limsup_{(y,\, t)\to(0,0^+)} \frac{(1 - e^{|y+tv|}) - (1 - e^{|y|})}{t}$$

needs to be calculated for $v \in \mathbb{R}$ and $y \in \mathbb{R}\backslash\{0\}$. The limit supremum is obtained as the supremum of the limits of sequences classified into four groups depending on the signs of $y + tv$ and y.

1. Case $y > 0$ and $y + tv > 0$. Let $f_t : (-\epsilon, \epsilon) \to \mathbb{R}$ be

$$f_t(y) = e^y \frac{1 - e^{tv}}{t}.$$

Note that for all $t > 0$, $\lim_{y \to 0} f_t(y) = (1 - e^{tv})/t$ holds. Let $f_0 : (-\epsilon, \epsilon) \to \mathbb{R}$ be $f_0(y) = -ve^y$. Then, for all $y \in (-\epsilon, \epsilon)$, $\lim_{t \downarrow 0} f_t(y) = f_0(y)$ holds per l'Hospital's rule. In addition, define M_t as;

$$M_t = \sup_{y \in [-\epsilon, \epsilon]} |f_t(y) - f_0(y)|$$
$$M_t = e^{\epsilon} \left| \frac{1 - e^{tv}}{t} + v \right|$$

and $\lim_{t \downarrow 0} M_t = 0$. As a result of this uniform convergence $\lim_{t \downarrow 0} \lim_{y \to 0} f_t(y) = \lim_{y \to 0} \lim_{t \downarrow 0} f_t(y) = -v$. Uniform convergence also implies $\lim_{(y,t) \to (0,0^+)} f_t(y) = -v$ because

$$|f_t(y) - f_0(0)| \leq |f_t(y) - f_0(y)| + |f_0(y) - f_0(0)|$$

and both terms on the right can be made arbitrarily small by letting t be small enough. Hence, if $y > 0$, $t > 0$ and $y + tv > 0$,

$$\limsup_{(y,\, t) \to (0,0^+)} \frac{(1 - e^{|y+tv|}) - (1 - e^{|y|})}{t} = -v.$$

2. Case $y < 0$ and $y + tv < 0$. In this case, let $f_t : (-\epsilon, \epsilon) \to \mathbb{R}$ be

$$f_t(y) = e^{-y} \frac{1 - e^{-tv}}{t}.$$

Then for all $t > 0$, $\lim_{y \to 0} f_t(y) = (1 - e^{-tv})/t$ holds. Let $f_0 : (-\epsilon, \epsilon) \to \mathbb{R}$ be $f_0(y) = ve^{-y}$. Note that for all $y \in (-\epsilon, \epsilon)$, $\lim_{t \downarrow 0} f_t(y) = f_0(y)$ holds per l'Hospital's rule.

Per similar analysis as in the previous case, it can be deduced that if $y < 0$, $t > 0$ and $y + tv < 0$,

$$\limsup_{(y,\, t)\to(0,0^+)} \frac{(1 - e^{|y+tv|}) - (1 - e^{|y|})}{t} = v.$$

3. Case $y < 0$ and $y + tv > 0$. Note that

$$\frac{e^{-y} - e^{y+tv}}{t} \leq \frac{e^{-y} - e^{-y-tv}}{t}.$$

Hence

$$\limsup_{(y,\, t)\to(0,0^+)} \frac{e^{-y} - e^{y+tv}}{t} \leq \limsup_{(y,\, t)\to(0,0^+)} \frac{e^{-y} - e^{-y-tv}}{t},$$
$$\limsup_{(y,\, t)\to(0,0^+)} \frac{e^{-y} - e^{y+tv}}{t} \leq v.$$

4. Case $y > 0$ and $y + tv < 0$. In this case

$$\frac{e^{y} - e^{-y-tv}}{t} \leq \frac{e^{y} - e^{y+tv}}{t}.$$

Hence

$$\limsup_{(y,\, t)\to(0,0^+)} \frac{e^{y} - e^{-y-tv}}{t} \leq \limsup_{(y,\, t)\to(0,0^+)} \frac{e^{y} - e^{y+tv}}{t},$$
$$\limsup_{(y,\, t)\to(0,0^+)} \frac{e^{y} - e^{-y-tv}}{t} \leq -v.$$

The supremum of the limits is $|v|$ for all cases. Therefore $f^o(0; v) = |v|$. The directional derivative exists and is $f'(0; v) = -v$ if $v \geq 0$ and $f'(0; v) = v$, if $v \leq 0$. The directional derivative is not equal to the generalized directional derivative. Hence f is not regular at

zero.

Usually, it is simpler to obtain the generalized gradient first and then obtain the generalized directional derivative using the generalized gradient.

Example 2.9.11. Let $x_1 \in \mathbb{R}$ and $x_2 \in \mathbb{R}$. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ from [43] which is defined as

$$f(x_1, x_2) = |x_1^2 - \sin(|x_2|)|. \tag{2.9.4}$$

f is plotted in Figure 2-1.

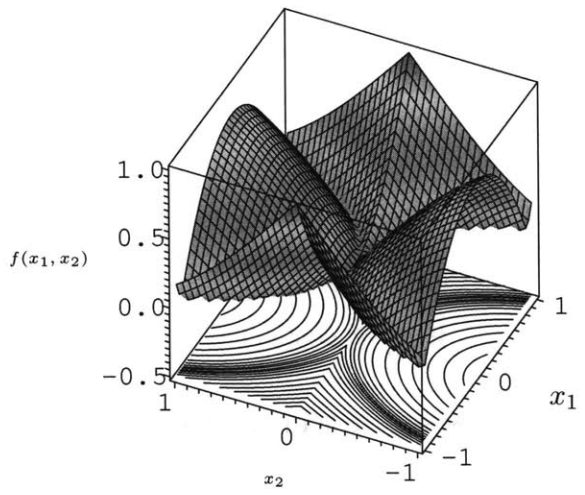


Figure 2-1: (Example 2.9.11) Plot of $f(x_1, x_2) = |x_1^2 - \sin(|x_2|)|$ and of its contours.

f is a locally Lipschitz continuous function because it is a composition of locally Lipschitz functions. Note that $f \in PC^1(\mathbb{R}^2)$. The selection functions of f are presented in Table 2.1. In order to calculate $\partial f(0,0)$, properties of PC^1 functions in Section 2.7.1 are used in conjunction with the data in Table 2.2 to obtain $\partial f(0,0) = \text{conv}\,(\{(0,1),(0,-1)\})$. Note that $(0,0) \in \partial f(0,0)$.

$f(x_1, x_2)$	$x_2 \geq 0$	$x_2 < 0$
$x_1^2 \geq \sin \lvert x_2 \rvert$	$x_1^2 - \sin x_2$	$x_1^2 + \sin x_2$
$x_1^2 < \sin \lvert x_2 \rvert$	$-x_1^2 + \sin x_2$	$-x_1^2 - \sin x_2$

Table 2.1: (Example 2.9.11) The selection functions of $\lvert x_1^2 - \sin(\lvert x_2 \rvert) \rvert$.

$\nabla f(x_1, x_2)$	$x_2 > 0$	$x_2 < 0$
$x_1^2 > \sin \lvert x_2 \rvert$	$(2x_1, -\cos x_2)$	$(2x_1, \cos x_2)$
$x_1^2 < \sin \lvert x_2 \rvert$	$(-2x_1, \cos x_2)$	$(-2x_1, -\cos x_2)$

Table 2.2: (Example 2.9.11) $\nabla f(x_1, x_2)$ at points where it is defined.

Example 2.9.12. Let $t_f = 5.0$. Let $x_1 \in (0, 2\pi)$ and $x_2 \in (0.5, 4)$ Consider the function $f : (0, 2\pi) \times (0.5, 4) \to \mathbb{R}$ [115] defined by

$$f(x_1, x_2) = \begin{cases} 2t_f \sin(x_1) & \text{if } \lvert \cos(x_1) \rvert < \frac{x_2}{2t_f}, \\ -x_2 \tan(x_1) e^{t_f + \frac{x_2}{2\cos(x_1)}} & \text{if } -\frac{x_2}{2t_f} \geq \cos(x_1), , \\ g(x_1, x_2) & \text{if } \frac{x_2}{2t_f} \leq \cos(x_1) \end{cases} \tag{2.9.5}$$

$$g(x_1, x_2) = \begin{cases} x_2 \tan(x_1) + 2(t_f - \frac{x_2}{2\cos(x_1)}) & \text{if } \tan(x_1) \geq 1, \\ x_2 \tan(x_1) e^{t_f - \frac{x_2}{2\cos(x_1)}} & \text{if } \tan(x_1) \leq -1, \\ h(x_1, x_2) & \text{if } -1 < \tan(x_1) < 1, \end{cases}$$

$$h(x_1, x_2) = \begin{cases} 2t_f - 4 \ln\left(\frac{\cos(x_1)}{\sin(x_1)}\right) - \frac{x_2}{\cos(x_1)} + x_2 & \text{if } \sqrt{e^{\frac{x_2}{2\cos(x_1)} - t_f}} \leq \tan(x_1) < 1, \\ x_2 \tan(x_1) e^{\frac{1}{2}(t_f - \frac{x_2}{2\cos(x_1)})} & \text{if } -\sqrt{e^{\frac{x_2}{2\cos(x_1)} - t_f}} < \tan(x_1) < \sqrt{e^{\frac{x_2}{2\cos(x_1)} - t_f}}, \\ -x_2 \tan^2(x_1) e^{(t_f - \frac{x_2}{2\cos(x_1)})} & \text{if } -1 < \tan(x_1) \leq -\sqrt{e^{\frac{x_2}{2\cos(x_1)} - t_f}}. \end{cases}$$

The plot of f is in Figure 2-2. In order to analyze f, first open sets that partition its domain will be constructed. Using these open sets, it will be shown that $f \in PC^1((0, 2\pi) \times (0.5, 4))$.

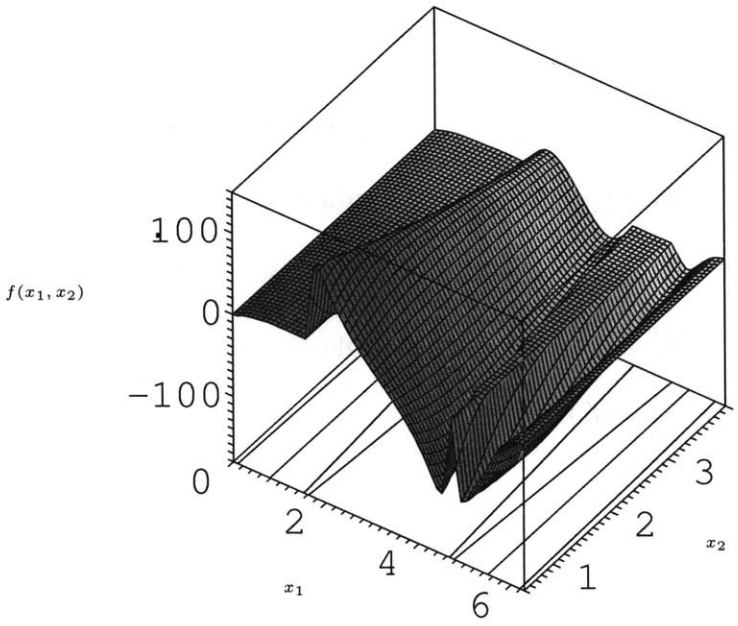


Figure 2-2: (Example 2.9.12) Plot of f.

Let $D = (0, 2\pi) \times (0.5, 4)$. Let ϵ be a small positive constant. Let

$$
\begin{aligned}
A_1' &= \{(x_1, x_2) \in D : \cos(x_1) - x_2/(2t_f) < \epsilon\}, \\
A_1'' &= \{(x_1, x_2) \in D : \cos(x_1) + x_2/(2t_f) > -\epsilon\}, \\
A_1 &= A_1' \cap A_1'', \\
A_2 &= \{(x_1, x_2) \in D : \cos(x_1) + x_2/(2t_f) < \epsilon\}, \\
A_3 &= \{(x_1, x_2) \in D : \cos(x_1) - x_2/(2t_f) > -\epsilon\}.
\end{aligned}
$$

The functions $g_1 : D \to \mathbb{R} : (x_1, x_2) \mapsto \cos(x_1) - x_2/t_f$ and $g_2 : D \to \mathbb{R} : (x_1, x_2) \mapsto \cos(x_1) + x_2/t_f$ are continuous functions. Let $(x_1, x_2) \in A_1'$, then using the continuity of g_1, it can be shown that, $O \subset D$, a neighborhood of (x_1, x_2), is a subset of A_1'. Therefore A_1' is an open set. Using similar reasoning, it can be shown that A_1'', A_1, A_2 and A_3 are open subsets of D.

Let

$$\begin{aligned}
B_1 =& \{(x_1, x_2) \in A_3 : \ \tan(x_1) > 1 + \epsilon\}, \\
B_2 =& \{(x_1, x_2) \in A_3 : \ \tan(x_1) < -1 - \epsilon\}, \\
B_3' =& \{(x_1, x_2) \in A_3 : \ \tan(x_1) < 1 + \epsilon\}, \\
B_3'' =& \{(x_1, x_2) \in A_3 : \ \tan(x_1) > -1 - \epsilon\}, \\
B_3 =& B_3' \cap B_3''.
\end{aligned}$$

Note that if $(x_1, x_2) \in A_3$, then $\cos(x_1) > 0$, hence $(x_1, x_2) \mapsto \tan(x_1)$ is a continuous function from A_3 to $\mathbb{R}$. Using similar reasoning as before, it can be deduced that B_1, B_2, B_3', B_3'' and B_3 are open sets.

Finally, let

$$\begin{aligned}
C_1 =& \{(x_1, x_2) \in B_3 : \ \sqrt{e^{\frac{x_2}{2\cos(x_1)} - t_f}} - \epsilon < \tan(x_1) < 1 + \epsilon\}, \\
C_2 =& \{(x_1, x_2) \in B_3 : \ -\sqrt{e^{\frac{x_2}{2\cos(x_1)} - t_f}} - \epsilon < \tan(x_1) < \sqrt{e^{\frac{x_2}{2\cos(x_1)} - t_f}} + \epsilon\}, \\
C_3 =& \{(x_1, x_2) \in B_3 : \ -1 - \epsilon < \tan(x_1) \leq -\sqrt{e^{\frac{x_2}{2\cos(x_1)} - t_f}} + \epsilon\}.
\end{aligned}$$

The sets C_1, C_2 and C_3 are open subsets of B_3 per the same arguments as before.

The functions

$$\begin{aligned}
h_1 &: C_1 \to \mathbb{R} : (x_1, x_2) \mapsto 2t_f - 4 \ln\left(\frac{\cos(x_1)}{\sin(x_1)}\right) - \frac{x_2}{\cos(x_1)} + x_2, \\
h_2 &: C_2 \to \mathbb{R} : (x_1, x_2) \mapsto x_2 \tan(x_1) e^{\frac{1}{2}(t_f - \frac{x_2}{2\cos(x_1)})}, \\
h_3 &: C_3 \to \mathbb{R} : (x_1, x_2) \mapsto -x_2 \tan^2(x_1) e^{(t_f - \frac{x_2}{2\cos(x_1)})}.
\end{aligned}$$

are continuously differentiable functions.

Let $(x_1^*, x_2^*) \in B_3$ be such that $\sqrt{e^{\frac{x_2^*}{2\cos(x_1^*)} - t_f}} = \tan(x_1^*)$. Then $(x_1^*, x_2^*) \in C_1 \cap C_2$ and $h_1(x_1^*, x_2^*) = h_2(x_1^*, x_2^*) = x_2$ for small enough ϵ. Hence h_1 and h_2 are selection functions of h on the neighborhoods of points $(x_1^*, x_2^*) \in B_3$ at which $\sqrt{e^{\frac{x_2^*}{2\cos(x_1^*)} - t_f}} = \tan(x_1^*)$ holds.

Let $(x_1^*, x_2^*) \in B_3$ be such that $-\sqrt{e^{\frac{x_2^*}{2\cos(x_1^*)} - t_f}} = \tan(x_1^*)$. Then $(x_1^*, x_2^*) \in C_2 \cap C_3$ and $h_2(x_1^*, x_2^*) = h_3(x_1^*, x_2^*) = -x_2$ for small enough ϵ. Hence h_2 and h_3 are selection functions of h on the neighborhoods of points $(x_1^*, x_2^*) \in B_3$ at which $-\sqrt{e^{\frac{x_2^*}{2\cos(x_1^*)} - t_f}} = \tan(x_1^*)$ holds.

As a result $h \in PC^1(B_3)$.

Define the functions

$$g_1 : B_1 \to \mathbb{R} : (x_1, x_2) \mapsto x_2 \tan(x_1) + 2\left(t_f - \frac{x_2}{2\cos(x_1)}\right),$$

$$g_2 : B_2 \to \mathbb{R} : (x_1, x_2) \mapsto x_2 \tan(x_1) e^{t_f - \frac{x_2}{2\cos(x_1)}},$$

$$g_3 : B_3 \to \mathbb{R} : (x_1, x_2) \mapsto h(x_1, x_2).$$

Note that g_1 and g_2 are continuously differentiable functions.

Let $(x_1^*, x_2^*) \in A_3$ be such that $\tan(x_1^*) = 1$. Then $(x_1^*, x_2^*) \in B_1 \cap B_3$ and $g_1(x_1^*, x_2^*) = g_3(x_1^*, x_2^*) = h_1(x_1^*, x_2^*) = x_2 + 2t_f - \frac{x_2}{\cos(x_1)}$ for small enough ϵ. Hence g_1 and h_1 are selection functions of g on the neighborhoods of points $(x_1^*, x_2^*) \in A_3$ at which $\tan(x_1^*) = 1$ holds.

Let $(x_1^*, x_2^*) \in A_3$ be such that $\tan(x_1^*) = -1$. Then $(x_1^*, x_2^*) \in B_2 \cap B_3$ and $g_2(x_1^*, x_2^*) = g_3(x_1^*, x_2^*) = h_3(x_1^*, x_2^*) = x_2 e^{\left(t_f - \frac{x_2}{2\cos(x_1)}\right)}$ for small enough ϵ. Hence h_3 and g_2 are selection functions of g on the neighborhoods of points $(x_1^*, x_2^*) \in A_3$ at which $\tan(x_1^*) = -1$ holds.

As a result, $g \in PC^1(A_3)$.

Finally, define the functions

$$f_1 : A_1 \to \mathbb{R} : (x_1, x_2) \mapsto 2t_f \sin(x_1),$$

$$f_2 : A_2 \to \mathbb{R} : (x_1, x_2) \mapsto -x_2 \tan(x_1) e^{t_f + \frac{x_2}{2\cos(x_1)}},$$

$$f_3 : A_3 \to \mathbb{R} : (x_1, x_2) \mapsto g(x_1, x_2).$$

Let $(x_1^*, x_2^*) \in D$ be such that $-\frac{x_2}{2t_f} = \cos(x_1)$. Then $(x_1^*, x_2^*) \in A_1 \cap A_2$ and $f_1(x_1^*, x_2^*) = f_2(x_1^*, x_2^*) = -x_2 \tan(x_1)$ for small enough ϵ. Hence f_1 and f_2 are selection functions of f on the neighborhoods of $(x_1^*, x_2^*) \in D$ at which $-\frac{x_2}{2t_f} = \cos(x_1)$ holds.

Let $(x_1^*, x_2^*) \in D$ be such that $\frac{x_2}{2t_f} = \cos(x_1)$. In this case $(x_1^*, x_2^*) \in A_1 \cap A_3$ and $f_1(x_1^*, x_2^*) = f_3(x_1^*, x_2^*) = x_2 \tan(x_1)$ for small enough ϵ. In order to compute $f_3(x_1^*, x_2^*)$, $g_1(x_1^*, x_2^*)$, $g_2(x_1^*, x_2^*)$, $h_2(x_1^*, x_2^*)$ need to be considered. The conditions $\sqrt{e^{\frac{x_2}{2\cos(x_1)} - t_f}} \leq \tan(x_1) < 1$ and $-1 < \tan(x_1) \leq -\sqrt{e^{\frac{x_2}{2\cos(x_1)} - t_f}}$ in (2.9.5) are violated in this case and therefore $h_1(x_1^*, x_2^*)$ and $h_2(x_1^*, x_2^*)$ need not be considered.

As a result of this analysis, $f \in PC^1(D)$.

Chapter 3

Parametric Sensitivity Analysis of Dynamic Systems using the Generalized Jacobian

The focus of this chapter is the existence of the derivative of the mapping $\boldsymbol{\eta} \mapsto \mathbf{x}(t_f, \boldsymbol{\eta})$ at $\mathbf{p} \in \mathcal{P}$, where $\mathbf{x} : [t_0, t_f] \times \mathcal{P} \to \mathcal{X}$ is the solution of the initial value problem:

$$\dot{\mathbf{x}}(t, \mathbf{p}) = \mathbf{f}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})),\ \forall t \in (t_0, t_f],\ \mathbf{x}(t_0, \mathbf{p}) = \mathbf{f}_0(\mathbf{p}),\ \forall \mathbf{p} \in \mathcal{P} \subset \mathbb{R}^{n_p}, \tag{3.0.1}$$

where $\mathbf{f} : \mathcal{T} \times \mathcal{P} \times \mathcal{X} \to \mathbb{R}^{n_x}$, $\mathbf{f}_0 : \mathcal{P} \to \tilde{\mathcal{X}}$ are locally Lipschitz continuous functions, $\mathcal{T}$ is an open subset of $\mathbb{R}$ such that $[t_0, t_f] \subset \mathcal{T}$, $\mathcal{X}$ is an open subset of $\mathbb{R}^{n_x}$, $\tilde{\mathcal{X}}$ is an open subset of $\mathcal{X}$, $\mathcal{P}$ is an open subset of $\mathbb{R}^{n_p}$, n_p and n_x are positive finite integers.

It is well known that the mapping $\boldsymbol{\eta} \mapsto \mathbf{x}(t_f, \boldsymbol{\eta})$ at $\mathbf{p} \in \mathcal{P}$ is a Lipschitz continuous function on O, some neighborhood of $\mathbf{p}$, and that it is differentiable for all $\boldsymbol{\eta} \in O \backslash S$ where S is a measure zero subset of O per Rademacher's Theorem. However, conditions on $\mathbf{x}(\cdot, \mathbf{p})$ that imply differentiability of $\boldsymbol{\eta} \mapsto \mathbf{x}(t_f, \boldsymbol{\eta})$ at $\mathbf{p}$ are not widely known.

If $\mathbf{f}$ were a continuously differentiable function on an open set containing $\{(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})) : t \in [t_0, t_f]\}$, then continuous differentiability would follow from Gronwall's classic result in [44]. This condition may not hold for the systems under consideration. Examples 3.6.1 and 3.6.2 consider cases where an open set with the desired properties does not exist, yet the mapping $\mathbf{p} \mapsto \mathbf{x}(t_f, \mathbf{p})$ is differentiable.

The sufficiency conditions follow from results in nonsmooth analysis and are based on the concepts of the *generalized gradient* and *Jacobian* [25]. A brief primer on nonsmooth analysis is presented in §3.1.

The results of this chapter define *forward* and *adjoint sensitivity* initial value problems to be solved to obtain the aforementioned derivative. The forward sensitivity initial value problem is a linear time-varying ordinary differential equation of the form

$$\dot{\boldsymbol{\zeta}}(t) = \mathbf{M}_0(t)\boldsymbol{\zeta}(t) + \mathbf{M}_1(t), \ \forall t \in (t_0, t_f], \ \boldsymbol{\zeta}(t_0) = \boldsymbol{\zeta}_0,$$

where $\mathbf{M}_1 : [t_0, t_f] \to \mathbb{R}^{n_x \times n_p}$ and $\mathbf{M}_0 : [t_0, t_f] \to \mathbb{R}^{n_x \times n_x}$ are measurable and $\boldsymbol{\zeta} : [t_0, t_f] \to \mathbb{R}^{n_x \times n_p}$ is an absolutely continuous function. This ordinary differential equation is solved simultaneously with (3.0.1). Thus, the derivative is obtained by integrating $n_x \times n_p$ additional equations. When the derivative of the mapping $\boldsymbol{\eta} \mapsto \int_{t_0}^{t_f} g(t, \boldsymbol{\eta}, \mathbf{x}(t, \boldsymbol{\eta}))dt$ with g a scalar-valued function is sought, the integration of the adjoint sensitivity initial value problem might be the computationally more efficient way to obtain the derivative. This is the case especially if $n_x \times n_p$ is significantly larger than $n_x + n_p$. The *adjoint*, $\boldsymbol{\lambda}$, is the solution of the *adjoint sensitivity initial value problem* of the form

$$\dot{\boldsymbol{\lambda}}(t) = \mathbf{A}_0(t)\boldsymbol{\lambda}(t) + \mathbf{A}_1(t), \ \forall t \in [t_0, t_f), \ \boldsymbol{\lambda}(t_f) = \boldsymbol{\lambda}_0$$

where $\mathbf{A}_0 : [t_0, t_f] \to \mathbb{R}^{n_x \times n_x}$ and $\mathbf{A}_1 : [t_0, t_f] \to \mathbb{R}^{n_x}$ are measurable functions and $\boldsymbol{\lambda} : [t_0, t_f] \to \mathbb{R}^{n_x}$ is an absolutely continuous function. The desired derivative is then computed

as the solution of an integral of the form $\int_{t_0}^{t_f} \mathbf{h}(t, \boldsymbol{\eta}, \mathbf{x}(t, \boldsymbol{\eta}), \boldsymbol{\lambda}(t))dt$. The adjoint method requires the integration of $n_x + n_p$ equations backwards in time including the quadrature of the integral. The development of forward and adjoint sensitivities in case $\mathbf{f}$ is continuously differentiable is well known and can be found in [24] and [26].

The adjoint and forward sensitivity results for (3.0.1) are derived in §3.2 and extended to a class of nonsmooth differential-algebraic equations in §3.3. Finally, results for a case where the evolution of the states is governed by disparate nonsmooth differential-algebraic equations in different time intervals is considered in §3.4.

The results of this chapter are most closely related to the works in [24], [93], [39], [95] in addition to [44]. The adjoint sensitivity initial value problems are derived for index-1 and index-2 differential-algebraic equations for sufficiently differentiable equations in [24]. In this chapter, forward and adjoint sensitivity initial value problems for index-1 differential-algebraic equations with locally Lipschitz equations are derived. The sensitivity and adjoint systems derived in this chapter have discontinuous right-hand sides, unlike the results in [24] and [44] and therefore require special treatment. In [93], the time interval in (3.0.1) is divided into finitely many subintervals and for each subinterval the evolution of the states is governed by different ordinary differential equations with continuously differentiable vector fields. The times at which the differential equations switch depend on the parameters; however, it is required that the number of subintervals and the order in which the equations are solved is independent of the parameters in some neighborhood of $\mathbf{p}$. The switching times are the solution of continuously differentiable equations of time, parameters and states. Discontinuities in the solution are allowed at switching times. Forward sensitivity equations for this case are derived in [93] and adjoint sensitivity equations are derived in [95]. The results in [93] are extended to differential-algebraic systems in [39]. In addition, an existence and uniqueness theory is developed. Existence and uniqueness results are developed by finite induction on the results in [44] and the implicit function theorem to compute the

jumps in the adjoint and forward sensitivity trajectories at switching times. This is not the approach used in this chapter. A subset of cases considered in this chapter can be analyzed using results from the aforementioned papers. However, the results in this chapter do not require invariance of the sequence of vector fields or a constant number of subintervals in a neighborhood of $\mathbf{p}$.

Implementation issues which are fully investigated in following chapters are summarized in §3.5. Examples in §3.6 conclude the chapter.

3.1 Preliminaries

Definition 3.1.1. *Let $t_0 \in \mathbb{R}$, $t_f \in \mathbb{R}$, $X_1 = [t_0, t_f]$ and $t \in X_1$. Let X_2 be an open subset of $\mathbb{R}^n$ and $\mathbf{x}_2 \in X_2$. Let $\mathbf{F} : X_1 \times X_2 \to \mathbb{R}^m$ be a function such that $\mathbf{F}(t, \cdot)$ is a locally Lipschitz continuous function for all $t \in [t_0, t_f]$. Let $\mathbf{w}_t : X_1 \to \mathbb{R}^{m \times n}$ be such that $\mathbf{w}_t(t) \in \partial_2 \mathbf{F}(t, \mathbf{x}_2)$ for all $t \in X_1 \backslash S$ where S is a measure zero subset of X_1. If the Lebesgue integral, $\int_{t_0}^{t_f} \mathbf{w}_t(t)dt$, exists, then $\mathbf{w}_t$ is a measurable selection of $\partial_2 \mathbf{F}(\cdot, \mathbf{x}_2)$ on $[t_0, t_f]$.*

A consequence of Theorem 2.7.2 in [25] is:

Theorem 3.1.2. *Let $g : X_1 \times X_2 \to \mathbb{R}$, where X_1 and X_2 are defined in Definition 3.1.1, satisfy the following conditions:*

1. *For each $\mathbf{x}_2 \in X_2$, $g(\cdot, \mathbf{x}_2)$ is a continuous function from X_1 to $\mathbb{R}$.*

2. *There exists a nonnegative Lebesgue integrable function, $k : X_1 \to \mathbb{R}$ such that for all $\mathbf{u}, \mathbf{v} \in X_2$, $|g(t, \mathbf{u}) - g(t, \mathbf{v})| \leq k(t)\|\mathbf{u} - \mathbf{v}\|$ for all $t \in X_1$.*

Let $f : X_2 \to \mathbb{R} : \mathbf{z} \mapsto \int_a^b g(t, \mathbf{z})dt$. Then f is locally Lipschitz continuous at all $\mathbf{x}_2 \in X_2$. Define W to be the set $\{\mathbf{w} \in \mathbb{R}^{n \times 1} : \mathbf{w} = \int_{t_0}^{t_f} \mathbf{w}_t(t)dt\}$ where $\mathbf{w}_t$ is any measurable selection of $\partial_2 g(\cdot, \mathbf{x}_2)$ on $[t_0, t_f]$. Then $\partial f(\mathbf{x}_2) \subset W$ holds.

The following is a pertinent restatement of Theorem 7.4.1 in [25].

Theorem 3.1.3. *Let X_1 and X_2 be open connected subsets of $\mathbb{R}$ and $\mathbb{R}^n$, respectively. Let $\tilde{X}_2$ be an open connected subset of X_2. Let $[t_0, t_f] \subset X_1$. Let $\mathbf{f} : X_1 \times X_2 \to \mathbb{R}^n$ be a locally Lipschitz continuous function.*

Let $\mathbf{x} : [t_0, t_f] \times \tilde{X}_2 \to X_2$ be such that $\mathbf{x}(\cdot, \mathbf{x}_2)$ is the only function that satisfies

$$\dot{\mathbf{x}}(t, \mathbf{x}_2) = \mathbf{f}(t, \mathbf{x}(t, \mathbf{x}_2)),\ \forall t \in (t_0, t_f],\ \mathbf{x}(t_0, \mathbf{x}_2) = \mathbf{x}_2,\ \forall \mathbf{x}_2 \in \tilde{X}_2.$$

Then $\mathbf{x}(t_f, \cdot)$ is locally Lipschitz continuous at all $\mathbf{x}_2 \in \tilde{X}_2$.

Let $\boldsymbol{\zeta} : [t_0, t_f] \times \tilde{X}_2 \to \mathbb{R}^{n \times n}$ be such that $\boldsymbol{\zeta}(\cdot, \mathbf{x}_2)$ is the solution of the differential equation

$$\dot{\boldsymbol{\zeta}}(t, \mathbf{x}_2) = \mathbf{M}(t)\boldsymbol{\zeta}(t, \mathbf{x}_2),\ \forall t \in (t_0, t_f],\ \boldsymbol{\zeta}(t_0, \mathbf{x}_2) = \mathbf{I}_n,\ \forall \mathbf{x}_2 \in \tilde{X}_2,$$

where $\mathbf{M}$ is any measurable selection of $\partial_2 \mathbf{f}(\cdot, \mathbf{x}(\cdot, \mathbf{x}_2))$, a set-valued mapping from $[t_0, t_f]$ to the subsets of $\mathbb{R}^{n \times n}$.

Let $\Phi(t_f, \mathbf{x}_2)$ be the set of $\boldsymbol{\zeta}(t_f, \mathbf{x}_2)$ obtained from all possible measurable selections, $\mathbf{M}$, and $R(t_f, \mathbf{x}_2)$ be the plenary hull of $\Phi(t_f, \mathbf{x}_2)$ i.e., $R(t_f, \mathbf{x}_2) = \{\mathbf{A} \in \mathbb{R}^{n \times n} : \mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{u} \leq \max\{\mathbf{v}^{\mathrm{T}}\mathbf{B}\mathbf{u} :\ \mathbf{B} \in \Phi(t_f, \mathbf{x}_2)\}, \forall (\mathbf{u}, \mathbf{v}) \in \mathbb{R}^n \times \mathbb{R}^n\}$.

Then $\partial_2 \mathbf{x}(t_f, \mathbf{x}_2) \subset R(t_f, \mathbf{x}_2)$. Let S be a measure zero subset of $[t_0, t_f]$. If $\partial_2 \mathbf{f}(t, \mathbf{x}(t, \mathbf{x}_2))$ is a singleton for all $t \in [t_0, t_f] \backslash S$, then $R(t_f, \mathbf{x}_2)$ is a singleton. Let the single element be $\mathbf{J}_2\mathbf{x}(t_f, \mathbf{x}_2)$. Then $\mathbf{x}(t_f, \cdot)$ is strictly differentiable at $\mathbf{x}_2$ and $\mathbf{J}_2\mathbf{x}(t_f, \mathbf{x}_2)$ is the strict derivative.

3.1.1 Note on Notation and Assumptions

In the remainder of this chapter, n_x, n_p and n_y represent finite positive integers, $t_0 \in \mathbb{R}$, $t_f \in \mathbb{R}$ and $t_0 \leq t_f$.

X_1, X_2, X_3, X_4, X_5 and X_6 are open connected subsets of $\mathbb{R}$, $\mathbb{R}^{n_p}$, $\mathbb{R}^{n_x}$, $\mathbb{R}^{n_y}$, $\mathbb{R}^{n_x}$ and $\mathbb{R}^{n_y}$

respectively. $X_7 = X_2 \times X_3$, $X_8 = X_2 \times X_3 \times X_4 \times X_5$ and $X_9 = X_4 \times X_5$. $T = [t_0, t_f] \subset X_1$.

In order to make the exposition more intuitive, the labels $\mathcal{T}$, $\mathcal{P}$, $\mathcal{X}$, $\mathcal{Y}$, $\dot{\mathcal{X}}$, $\mathcal{W}$ and $\mathcal{Q}$ will be used instead of X_1, X_2, X_3, X_4, X_5, X_6 and X_9. If the symbols $t, \mathbf{p}, \mathbf{x}, \mathbf{y}, \dot{\mathbf{x}}, \mathbf{w}, \mathbf{v}, \mathbf{u}$ and $\mathbf{q}$ appear as subscripts, they represent the indices $1, 2, 3, 4, 5, 6, 7, 8$ and 9.

3.2 Ordinary Differential Equations

In this section, sufficient conditions for the existence of adjoint and forward sensitivity trajectories are derived for the solutions of ordinary differential equations.

Assumption 3.2.1. *Let $\mathbf{f} : \mathcal{T} \times \mathcal{P} \times \mathcal{X} \to \dot{\mathcal{X}}$ and $\mathbf{f}_0 : \mathcal{P} \to \tilde{\mathcal{X}}$ be locally Lipschitz continuous functions where $\tilde{\mathcal{X}}$ is an open connected subset of $\mathcal{X}$. Let $\mathbf{x} : T \times \mathcal{P} \to \mathcal{X}$ be such that $\mathbf{x}(\cdot, \mathbf{p})$ is the unique solution of the initial value problem*

$$\dot{\mathbf{x}}(t, \mathbf{p}) = \mathbf{f}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})), \ \forall t \in (t_0, t_f], \ \mathbf{x}(t_0, \mathbf{p}) = \mathbf{f}_0(\mathbf{p}), \ \forall \mathbf{p} \in \mathcal{P}. \tag{3.2.1}$$

Remark 3.2.2. *Let $\mathbf{z} : T \times \mathcal{P} \to \mathcal{P} : (t, \mathbf{p}) \mapsto \mathbf{p}$ and $\mathbf{v} : T \times \mathcal{P} \to \mathcal{P} \times \mathcal{X} : (t, \mathbf{p}) \mapsto (\mathbf{z}(t, \mathbf{p}), \mathbf{x}(t, \mathbf{p}))$ for the remainder of this chapter.*

Theorem 3.2.3. *Let Assumption 3.2.1 hold. Assume $\mathbf{f}_0$ is strictly differentiable at $\mathbf{p} \in \mathcal{P}$. Let S be a measure zero subset of T. Assume $\partial_{\mathbf{v}} \mathbf{f}(t, \mathbf{v}(t, \mathbf{p}))$ is a singleton for all $t \in T \backslash S$.*

Then the mapping $\boldsymbol{\eta} \mapsto \mathbf{x}(t, \boldsymbol{\eta})$ is locally Lipschitz continuous and strictly differentiable at $\mathbf{p} \in \mathcal{P}$ for all $t \in T$. Hence $\partial_{\mathbf{p}} \mathbf{x}(t, \mathbf{p})$ is a singleton for all $t \in T$. Let $\partial_{\mathbf{p}} \mathbf{x}(t, \mathbf{p}) = \{\mathbf{J}_{\mathbf{p}} \mathbf{x}(t, \mathbf{p})\}$. Then, $\mathbf{J}_{\mathbf{p}} \mathbf{x}(\cdot, \mathbf{p})$ is the unique absolutely continuous function on T that satisfies

$$\begin{aligned} \dot{\mathbf{J}}_{\mathbf{p}} \mathbf{x}(t, \mathbf{p}) &= \hat{\mathbf{J}}_{\mathbf{x}} \mathbf{f}(t, \mathbf{v}(t, \mathbf{p})) \mathbf{J}_{\mathbf{p}} \mathbf{x}(t, \mathbf{p}) + \hat{\mathbf{J}}_{\mathbf{p}} \mathbf{f}(t, \mathbf{v}(t, \mathbf{p})), \ \forall t \in (t_0, t_f], \\ \mathbf{J}_{\mathbf{p}} \mathbf{x}(t_0, \mathbf{p}) &= \mathbf{J}_{\mathbf{p}} \mathbf{f}_0(\mathbf{p}). \end{aligned} \tag{3.2.2}$$

Proof. Let $\mathbf{g} : \mathcal{T} \times \mathcal{P} \times \mathcal{X} \to \mathbb{R}^{n_p} \times \dot{\mathcal{X}} : (t, \boldsymbol{\mu}) \mapsto (\mathbf{0}, \mathbf{f}(t, \boldsymbol{\mu}))$. $\mathbf{g}$ is a locally Lipschitz continuous function because it is the composition of locally Lipschitz continuous functions $\mathbf{f}$ and $\mathbf{h} : \mathbb{R}^{n_x} \to \mathbb{R}^{n_p+n_x} : \mathbf{u} \mapsto (\mathbf{0}, \mathbf{u})$. If $\partial_{\mathbf{v}}\mathbf{f}(t, \boldsymbol{\mu})$ is a singleton then $\partial_{\mathbf{v}}\mathbf{g}(t, \boldsymbol{\mu}))$ is a singleton per Theorem 2.6.7.

Consider the initial value problem:

$$\dot{\boldsymbol{\nu}}(t, \boldsymbol{\nu}_0) = \mathbf{g}(t, \boldsymbol{\nu}(t, \boldsymbol{\nu}_0)),\ \forall t \in (t_0, t_f],\ \boldsymbol{\nu}(t_0, \boldsymbol{\nu}_0) = \boldsymbol{\nu}_0,\ \boldsymbol{\nu}_0 \in \mathcal{P} \times \tilde{\mathcal{X}} \tag{3.2.3}$$

where $\boldsymbol{\nu} : T \times \mathcal{P} \times \tilde{\mathcal{X}} \to \mathcal{P} \times \mathcal{X}$. Per Theorem 3.1.3, the mapping $\boldsymbol{\nu}_0 \mapsto \boldsymbol{\nu}(t, \boldsymbol{\nu}_0)$ is locally Lipschitz continuous at $\bar{\boldsymbol{\nu}}_0 \in \mathcal{P} \times \tilde{\mathcal{X}}$ for all $t \in [t_0, t_f]$ if the solution $\boldsymbol{\nu}(\cdot, \bar{\boldsymbol{\nu}}_0)$ exists. $\partial_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(t, \bar{\boldsymbol{\nu}}_0)$ (here, the subscript 7 is replaced with $\boldsymbol{\nu}_0$) is contained in the plenary hull of the solutions of the family of initial value problems:

$$\dot{\boldsymbol{\zeta}}(t, \bar{\boldsymbol{\nu}}_0) = \mathbf{M}(t)\boldsymbol{\zeta}(t, \bar{\boldsymbol{\nu}}_0),\ \forall t \in (t_0, t_f],\ \boldsymbol{\zeta}(t_0, \bar{\boldsymbol{\nu}}_0) = \mathbf{I}_{n_p+n_x} \tag{3.2.4}$$

where $\mathbf{M}$ is any measurable selection of $\partial_{\mathbf{v}}\mathbf{g}(\cdot, \boldsymbol{\nu}(\cdot, \bar{\boldsymbol{\nu}}_0))$, a set-valued mapping from T to the subsets of $\mathbb{R}^{(n_p+n_x)\times(n_p+n_x)}$, of the form

$$\mathbf{M}(t) = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{M_p}(t) & \mathbf{M_x}(t) \end{bmatrix}$$

and $\mathbf{M_x} : T \to \mathbb{R}^{n_x \times n_x}$, $\mathbf{M_p} : T \to \mathbb{R}^{n_x \times n_p}$ are bounded measurable functions.

Suppose that $\bar{\boldsymbol{\nu}}_0 = (\mathbf{p}, \mathbf{x}_0)$ where $\mathbf{x}_0 \in \tilde{\mathcal{X}}$ is such that $\partial_{\mathbf{v}}\mathbf{g}(t, \boldsymbol{\nu}(t, \bar{\boldsymbol{\nu}}_0))$ is a singleton for all t in T except for a subset S of measure zero. Then by Theorem 3.1.3, $\partial_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(t, \bar{\boldsymbol{\nu}}_0)$ is a singleton for all $t \in T$. Let the single element and strict derivative of the mapping $\boldsymbol{\nu}_0 \mapsto \boldsymbol{\nu}(t, \boldsymbol{\nu}_0)$ at

$\bar{\boldsymbol{\nu}}_0$ be $\mathbf{J}_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(t, \bar{\boldsymbol{\nu}}_0)$. Then (3.2.4) can be written as

$$\dot{\mathbf{J}}_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(t, \bar{\boldsymbol{\nu}}_0) = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}(t, \boldsymbol{\nu}(t, \bar{\boldsymbol{\nu}}_0)) & \hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}(t, \boldsymbol{\nu}(t, \bar{\boldsymbol{\nu}}_0)) \end{bmatrix} \mathbf{J}_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(t, \bar{\boldsymbol{\nu}}_0), \ \forall t \in (t_0, t_f], \tag{3.2.5}$$

$$\mathbf{J}_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(t_0, \bar{\boldsymbol{\nu}}_0) = \mathbf{I}_{n_p+n_x}.$$

Note that $\hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}(\cdot, \boldsymbol{\nu}(\cdot, \bar{\boldsymbol{\nu}}_0))$ and $\hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}(\cdot, \boldsymbol{\nu}(\cdot, \bar{\boldsymbol{\nu}}_0))$ differ from any measurable selections $\mathbf{M}_{\mathbf{p}}$ and $\mathbf{M}_{\mathbf{x}}$ if $t \in S$ only. Therefore using these quantities instead of the measurable selections does not alter the value of $\mathbf{J}_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(t, \bar{\boldsymbol{\nu}}_0)$.

If $\bar{\boldsymbol{\nu}}_0 = (\mathbf{p}, \mathbf{f}_0(\mathbf{p}))$, then $\boldsymbol{\nu}(t, (\mathbf{p}, \mathbf{f}_0(\mathbf{p}))) = \mathbf{v}(t, \mathbf{p}), \ \forall (t, \mathbf{p}) \in T \times \mathcal{P}$ satisfies (3.2.3). In addition, $\partial_{\mathbf{v}}\mathbf{g}(t, \mathbf{v}(t, \mathbf{p}))$ is a singleton for all $t \in T \backslash S$, $\mathbf{M}_{\mathbf{x}}(t) = \mathbf{J}_{\mathbf{x}}\mathbf{f}(t, \mathbf{v}(t, \mathbf{p}))$ and $\mathbf{M}_{\mathbf{p}}(t) = \mathbf{J}_{\mathbf{p}}\mathbf{f}(t, \mathbf{v}(t, \mathbf{p}))$ for all $t \in T \backslash S$. Finally, $\partial_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(t, (\mathbf{p}, \mathbf{f}_0(\mathbf{p})))$ is a singleton for all $t \in T$.

Let $\mathbf{p} \in \mathcal{P}$ and $\bar{\mathbf{w}} \in \tilde{\mathcal{X}}$. The mapping $(\mathbf{u}, \mathbf{w}) \mapsto (\mathbf{u}, \mathbf{f}_0(\mathbf{u}))$ is strictly differentiable at $(\mathbf{p}, \bar{\mathbf{w}})$ and the derivative is

$$\mathbf{A}_0(\mathbf{p}, \bar{\mathbf{w}}) = \begin{bmatrix} \mathbf{I}_{n_p} & \mathbf{0} \\ \mathbf{J}\mathbf{f}_0(\mathbf{p}) & \mathbf{0} \end{bmatrix}$$

because $\mathbf{f}_0$ is strictly differentiable at $\mathbf{p}$. As a result, $(\mathbf{u}, \mathbf{f}_0(\mathbf{u})) \mapsto \boldsymbol{\nu}(t, (\mathbf{u}, \mathbf{f}_0(\mathbf{u})))$ is locally Lipschitz continuous at $(\mathbf{p}, \mathbf{f}_0(\mathbf{p}))$ and $\partial_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(t, (\mathbf{p}, \mathbf{f}_0(\mathbf{p})))$ is $\{\mathbf{J}_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(t, (\mathbf{p}, \mathbf{f}_0(\mathbf{p}))) \ \mathbf{A}_0(\mathbf{p}, \bar{\mathbf{w}})\}$ per Theorem 2.6.7.

Equation (3.2.5) is a linear ordinary differential equation that admits a matrix-valued function $\boldsymbol{\Gamma}(t, \tau)$ such that $\mathbf{J}_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(t, \bar{\boldsymbol{\nu}}_0) = \boldsymbol{\Gamma}(t, \tau)\mathbf{J}_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(\tau, \bar{\boldsymbol{\nu}}_0) = \boldsymbol{\Gamma}(t, t_0)\mathbf{I}_{n_p+n_x}$. Hence $\mathbf{J}_{\boldsymbol{\nu}_0}\boldsymbol{\nu}$ $(t, (\mathbf{p}, \mathbf{f}_0(\mathbf{p})))\mathbf{A}_0(\mathbf{p}, \bar{\mathbf{w}})$ is $\boldsymbol{\Gamma}(t, t_0)\mathbf{A}_0(\mathbf{p}, \bar{\mathbf{w}})$ and $\mathbf{J}_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(t, (\mathbf{p}, \mathbf{f}_0(\mathbf{p})))\mathbf{A}_0(\mathbf{p}, \bar{\mathbf{w}})$ is obtained as the solution of (3.2.5) with the initial conditions $\mathbf{J}_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(t_0, \bar{\boldsymbol{\nu}}_0) = \mathbf{A}_0(\mathbf{p}, \bar{\mathbf{w}})$ and $\bar{\boldsymbol{\nu}}_0 = (\mathbf{p}, \mathbf{f}_0(\mathbf{p}))$.

Let

$$\mathbf{J}_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(t,(\mathbf{p},\mathbf{f}_0(\mathbf{p}))) = \begin{bmatrix} \boldsymbol{\zeta}_A(t,\mathbf{p}) & \boldsymbol{\zeta}_B(t,\mathbf{p}) \\ \boldsymbol{\zeta}_C(t,\mathbf{p}) & \boldsymbol{\zeta}_D(t,\mathbf{p}) \end{bmatrix}, \ \forall (t,\mathbf{p}) \in T \times \mathcal{P},$$

where $\boldsymbol{\zeta}_A : T \times \mathcal{P} \to \mathbb{R}^{n_p \times n_p}$, $\boldsymbol{\zeta}_B : T \times \mathcal{P} \to \mathbb{R}^{n_p \times n_x}$, $\boldsymbol{\zeta}_C : T \times \mathcal{P} \to \mathbb{R}^{n_x \times n_p}$ and $\boldsymbol{\zeta}_D : T \times \mathcal{P} \to \mathbb{R}^{n_x \times n_x}$.

Multiplying out (3.2.5) and substituting $\mathbf{v}(t,\mathbf{p})$ for $\boldsymbol{\nu}(t,(\mathbf{p},\mathbf{f}_0(\mathbf{p})))$ results in

$$\begin{aligned}
&\dot{\boldsymbol{\zeta}}_A(t,\mathbf{p}) = \mathbf{0}, \forall t \in (t_0,t_f], \ \dot{\boldsymbol{\zeta}}_B(t,\mathbf{p}) = \mathbf{0}, \forall t \in (t_0,t_f], \\
&\dot{\boldsymbol{\zeta}}_C(t,\mathbf{p}) = \hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p})\boldsymbol{\zeta}_A(t,\mathbf{p}) + \hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))\boldsymbol{\zeta}_C(t,\mathbf{p}), \forall t \in (t_0,t_f], \\
&\dot{\boldsymbol{\zeta}}_D(t,\mathbf{p}) = \hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))\boldsymbol{\zeta}_B(t,\mathbf{p}) + \hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))\boldsymbol{\zeta}_D(t,\mathbf{p}), \forall t \in (t_0,t_f], \\
&\boldsymbol{\zeta}_A(t_0,\mathbf{p}) = \mathbf{I}_{n_p}, \ \boldsymbol{\zeta}_B(t_0,\mathbf{p}) = \mathbf{0}, \\
&\boldsymbol{\zeta}_C(t_0,\mathbf{p}) = \mathbf{J}\mathbf{f}_0(\mathbf{p}), \ \boldsymbol{\zeta}_D(t_0,\mathbf{p}) = \mathbf{0}.
\end{aligned}$$

Note that $\boldsymbol{\zeta}_C$ is the derivative of the map $\boldsymbol{\eta} \mapsto \mathbf{x}(t,\boldsymbol{\eta})$ at $\boldsymbol{\eta} = \mathbf{p}$. In addition, $\boldsymbol{\zeta}_A(t,\mathbf{p}) = \mathbf{I}_{n_p}$ for all $t \in T$. Renaming $\boldsymbol{\zeta}_C(t,\mathbf{p})$ as $\mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})$, the following desired result is obtained:

$$\begin{aligned}
&\dot{\mathbf{J}}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p}) = \hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))\mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p}) + \hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p})), \ \forall t \in (t_0,t_f], \\
&\mathbf{J}_{\mathbf{p}}\mathbf{x}(t_0,\mathbf{p}) = \mathbf{J}_{\mathbf{p}}\mathbf{f}_0(\mathbf{p}).
\end{aligned}$$

Note that $\hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}(\cdot,\mathbf{v}(\cdot,\mathbf{p}))$ and $\hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}(\cdot,\mathbf{v}(\cdot,\mathbf{p}))$ are bounded and measurable functions on T. Hence $\mathbf{J}_{\mathbf{p}}\mathbf{x}(\cdot,\mathbf{p})$ is absolutely continuous on T per Theorem 3 in [37]. □

The next two theorems contain adjoint sensitivity results that consider two cases. In the first theorem, the function $G : \mathcal{P} \to \mathbb{R}$ can be computed by integrating a locally Lipschitz function on T. The second theorem considers the case when such a computation is not

possible.

Theorem 3.2.4. *Let the hypotheses of Theorem 3.2.3 hold. Let $g : \mathcal{T} \times \mathcal{P} \times \mathcal{X} \to \mathbb{R}$ be a locally Lipschitz continuous function. Define $G : \mathcal{P} \to \mathbb{R}$ by*

$$G(\mathbf{p}) = \int_{t_0}^{t_f} g(t, \mathbf{v}(t, \mathbf{p}))dt.$$

Let Q be a measure zero subset of T. Let $\partial_{\mathbf{v}} g(t, \mathbf{v}(t, \mathbf{p}))$ be a singleton for all $t \in T \backslash Q$.

Let $\boldsymbol{\lambda} : T \to \mathbb{R}^{n_x}$ be a solution of the initial value problem,

$$\dot{\boldsymbol{\lambda}}(t) = -\hat{\mathbf{J}}_{\mathbf{x}} \mathbf{f}(t, \mathbf{v}(t, \mathbf{p}))^{\mathrm{T}} \boldsymbol{\lambda}(t) + \hat{\boldsymbol{\nabla}}_{\mathbf{x}} g(t, \mathbf{v}(t, \mathbf{p})), \ \forall t \in [t_0, t_f), \ \boldsymbol{\lambda}(t_f) = \mathbf{0}. \tag{3.2.6}$$

Then, $\boldsymbol{\lambda}$ is unique and absolutely continuous. In addition, G is locally Lipschitz continuous and strictly differentiable at $\mathbf{p}$ and the strict derivative is

$$\boldsymbol{\nabla} G(\mathbf{p}) = \int_{t_0}^{t_f} \hat{\boldsymbol{\nabla}}_{\mathbf{p}} g(t, \mathbf{v}(t, \mathbf{p})) - \hat{\mathbf{J}}_{\mathbf{p}} \mathbf{f}(t, \mathbf{v}(t, \mathbf{p}))^{\mathrm{T}} \boldsymbol{\lambda}(t) dt + \mathbf{J}_{\mathbf{p}} \mathbf{x}(t, \mathbf{p})^{\mathrm{T}} \boldsymbol{\lambda}(t) \Big|_{t_0}^{t_f}. \tag{3.2.7}$$

Proof. The proof consists of applying the results of Theorem 3.1.2 and Theorem 3.2.3 to the equivalent integral

$$G(\mathbf{p}) = \int_{t_0}^{t_f} g(t, \mathbf{v}(t, \mathbf{p})) - \boldsymbol{\lambda}(t)^{\mathrm{T}} \left(\mathbf{f}(t, \mathbf{v}(t, \mathbf{p})) - \dot{\mathbf{x}}(t, \mathbf{p}) \right) dt.$$

Let $\epsilon > 0$ and $\Gamma(\epsilon, \mathbf{p}) = \{(t, \tilde{\mathbf{v}}) \in T \times \mathbb{R}^{n_p + n_x} : \|\tilde{\mathbf{v}} - \mathbf{v}(t, \mathbf{p})\| < \epsilon\}$. Note that there exists an $\epsilon > 0$ such that $\Gamma(\epsilon, \mathbf{p}) \subset \mathcal{T} \times \mathcal{P} \times \mathcal{X}$ because $\mathcal{T} \times \mathcal{P} \times \mathcal{X}$ is open and $\{(t, \mathbf{v}(t, \mathbf{p})) : t \in T\}$ is a bounded subset of $\mathcal{T} \times \mathcal{P} \times \mathcal{X}$. Since $\Gamma(\epsilon, \mathbf{p})$ is bounded and $\mathbf{f}$ and g are locally Lipschitz continuous at all points in $\Gamma(\epsilon, \mathbf{p})$, there exists a Lipschitz constant, K, such that $\|g(t, \mathbf{v}_1) - g(t, \mathbf{v}_2)\| \leq K \|\mathbf{v}_1 - \mathbf{v}_2\|$ and $\|\mathbf{f}(t, \mathbf{v}_1) - \mathbf{f}(t, \mathbf{v}_2)\| \leq K \|\mathbf{v}_1 - \mathbf{v}_2\|$ for all $(t, \mathbf{v}_1) \in \Gamma(\epsilon, \mathbf{p})$, $(t, \mathbf{v}_2) \in \Gamma(\epsilon, \mathbf{p})$. In addition, $\|\hat{\boldsymbol{\nabla}}_{\mathbf{x}} g(t, \mathbf{v}(t, \mathbf{p}))\| \leq K$ and $\|\hat{\boldsymbol{\nabla}}_{\mathbf{p}} g(t, \mathbf{v}(t, \mathbf{p}))\| \leq K$ for

all $t \in T$.

Since $g(t,\cdot)$ is strictly differentiable for all $t \in T\backslash Q$,

$$\lim_{s_i \downarrow 0} \frac{g(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}) + s_i\mathbf{d}) - g(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}))}{s_i} = \hat{\nabla}_{\mathbf{x}} g(t,\mathbf{v}(t,\mathbf{p}))^{\mathrm{T}}\mathbf{d}$$

holds for all $t \in T\backslash Q$ and $\mathbf{d} \in \mathbb{R}^{n_x}$. The quotient on the left is a bounded continuous function of t for small enough s_i due to local Lipschitz continuity of g. Per the Lebesgue Dominated Convergence Theorem it can be shown that $\hat{\nabla}_{\mathbf{x}} g(\cdot,\mathbf{v}(\cdot,\mathbf{p}))$ is a measurable function from T to $\mathbb{R}^{n_x}$. Lebesgue integrability of $\hat{\nabla}_{\mathbf{p}} g(\cdot,\mathbf{v}(\cdot,\mathbf{p}))$ can be shown using the same arguments.

$\hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}(\cdot,\mathbf{v}(\cdot,\mathbf{p}))$ is a bounded measurable function of t because it differs from a bounded measurable function $\mathbf{M}_{\mathbf{x}}$ (as defined in Theorem 3.2.3) on a measure zero subset of T and $\hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))$ is zero for t in that set.

As a result, there exists a unique, bounded and absolutely continuous solution to (3.2.6) in the sense of Carathéodory per Theorem 3 in [37], employing the trivial extension of (3.2.6) from T to $\mathcal{T}$ obtained by defining $\dot{\boldsymbol{\lambda}}(t) = \mathbf{0}$ in case $t \notin T$.

Per Theorem 3.2.3, the mapping $\boldsymbol{\eta} \mapsto \mathbf{x}(t,\boldsymbol{\eta})$ is locally Lipschitz continuous. Therefore the mapping $\boldsymbol{\eta} \mapsto \mathbf{v}(t,\boldsymbol{\eta})$ is locally Lipschitz continuous at $\mathbf{p}$ for all $t \in T$. Let $O \subset \mathcal{P}$ be the open set on which local Lipschitz continuity holds. Due to the continuous dependence of $\mathbf{v}(t,\boldsymbol{\eta})$ on $\boldsymbol{\eta} \in O$, one can pick an $\epsilon_p > 0$ such that for all $\boldsymbol{\eta} \in O$ satisfying $\|\boldsymbol{\eta} - \mathbf{p}\| < \epsilon_p$, $\mathbf{v}(t,\boldsymbol{\eta}) \subset \Gamma(\epsilon,\mathbf{p})$.

Let $\mathbf{p}_1 \in O$ and $\mathbf{p}_2 \in O$ satisfy $\|\mathbf{p} - \mathbf{p}_1\| < \epsilon_p$ and $\|\mathbf{p} - \mathbf{p}_2\| < \epsilon_p$. Then

$$\begin{aligned}
\mathbf{x}(t,\mathbf{p}_1) - \mathbf{x}(t,\mathbf{p}_2) &= \int_{t_0}^{t} \mathbf{f}(\zeta,\mathbf{v}(\zeta,\mathbf{p}_1)) - \mathbf{f}(\zeta,\mathbf{v}(\zeta,\mathbf{p}_2))d\zeta + \mathbf{x}(t_0,\mathbf{p}_1) - \mathbf{x}(t_0,\mathbf{p}_2), \\
\|\mathbf{x}(t,\mathbf{p}_1) - \mathbf{x}(t,\mathbf{p}_2)\| &\leq \int_{t_0}^{t} \|\mathbf{f}(\zeta,\mathbf{v}(\zeta,\mathbf{p}_1)) - \mathbf{f}(\zeta,\mathbf{v}(\zeta,\mathbf{p}_2))\|d\zeta + K_0\|\mathbf{p}_1 - \mathbf{p}_2\|, \\
\|\mathbf{x}(t,\mathbf{p}_1) - \mathbf{x}(t,\mathbf{p}_2)\| &\leq \int_{t_0}^{t} K\|\mathbf{p}_2 - \mathbf{p}_2\| + K\|\mathbf{x}(\zeta,\mathbf{p}_1) - \mathbf{x}(\zeta,\mathbf{p}_2)\|d\zeta + K_0\|\mathbf{p}_1 - \mathbf{p}_2\|
\end{aligned}$$

hold where K_0 is a Lipschitz constant of $\mathbf{f}_0$ in a neighborhood of $\mathbf{p}$ that contains $\mathbf{p}_1$ and $\mathbf{p}_2$. Application of Gronwall's Lemma produces

$$\|\mathbf{x}(t,\mathbf{p}_1) - \mathbf{x}(t,\mathbf{p}_2)\| \leq (K(t-t_0)+K_0)\|\mathbf{p}_1 - \mathbf{p}_2\|e^{K(t-t_0)}.$$

Hence $\|g(t,\mathbf{v}(t,\mathbf{p}_1)) - g(t,\mathbf{v}(t,\mathbf{p}_2)\| \leq (K+K^2(t_f-t_0)+KK_0)e^{K(t_f-t_0)}\|\mathbf{p}_1-\mathbf{p}_2\|,\ \forall t \in T.$

The term $(\mathbf{f}(t,\mathbf{v}(t,\mathbf{p})) - \dot{\mathbf{x}}(t,\mathbf{p}))$ is identically zero for all values of $\mathbf{p}$ and t. As a result, the hypotheses of Theorem 3.1.2 hold. Hence, $\boldsymbol{\eta} \mapsto G(\boldsymbol{\eta})$ is locally Lipschitz continuous at $\mathbf{p}$ and $\partial G(\mathbf{p}) \subset W$ where $W = \{\mathbf{w} : \mathbf{w} = \int_{t_0}^{t_f} \mathbf{w}_t(t)dt\}$, $\mathbf{w}_t$ is a measurable selection of $\partial_{\mathbf{p}}\tilde{g}(\cdot,\mathbf{p})$ and $\tilde{g}(\cdot,\mathbf{p}) = g(\cdot,\mathbf{v}(\cdot,\mathbf{p})) - \boldsymbol{\lambda}(\cdot)^{\mathrm{T}}(\mathbf{f}(\cdot,\mathbf{v}(\cdot,\mathbf{p})) - \dot{\mathbf{x}}(\cdot,\mathbf{p}))$.

Let $\tilde{\mathbf{w}}_t(t) = \mathbf{J}_{\mathbf{p}}(\mathbf{f}(t,\mathbf{v}(t,\mathbf{p})) - \dot{\mathbf{x}}(t,\mathbf{p}))$. Then $\tilde{\mathbf{w}}_t(t) = \mathbf{0},\ \forall t \in T$ and

$$\mathbf{J}_{\mathbf{p}}\dot{\mathbf{x}}(t,\mathbf{p}) = \mathbf{J}_{\mathbf{x}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))\mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p}) + \mathbf{J}_{\mathbf{p}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p})) = \dot{\mathbf{J}}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p}),\ \forall t \in T\backslash S$$

per (3.2.2). Therefore, $\tilde{\mathbf{w}}_t(t)$ is $\hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))\mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p}) + \hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p})) - \dot{\mathbf{J}}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})$.

Any measurable selection $\mathbf{w}_t$ differs from $\hat{\mathbf{w}}_t$ only if $t \in S \cup Q$ where $\hat{\mathbf{w}}_t$ is

$$\hat{\mathbf{w}}_t(t) = \mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})^{\mathrm{T}}\hat{\boldsymbol{\nabla}}_{\mathbf{x}}g(t,\mathbf{v}(t,\mathbf{p})) + \hat{\boldsymbol{\nabla}}_{\mathbf{p}}g(t,\mathbf{v}(t,\mathbf{p})) - \tilde{\mathbf{w}}_t(t)^{\mathrm{T}}\boldsymbol{\lambda}(t).$$

The integral of $\hat{\mathbf{w}}_t(t)$ is

$$\int_{t_0}^{t_f} \mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})^{\mathrm{T}}\hat{\boldsymbol{\nabla}}_{\mathbf{x}}g(t,\mathbf{v}(t,\mathbf{p})) + \hat{\boldsymbol{\nabla}}_{\mathbf{p}}g(t,\mathbf{v}(t,\mathbf{p})) - \left(\hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))\mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p}) + \hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p})) - \dot{\mathbf{J}}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})\right)^{\mathrm{T}}\boldsymbol{\lambda}(t)dt. \tag{3.2.8}$$

Since $\boldsymbol{\lambda}$ and $\mathbf{J}_{\mathbf{p}}\mathbf{x}(\cdot,\mathbf{p})$ are absolutely continuous functions of t, integration by parts for the

Lebesgue integral produces

$$\int_{t_0}^{t_f} \dot{\mathbf{J}}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})^{\mathrm{T}}\boldsymbol{\lambda}(t)dt = \mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})^{\mathrm{T}}\boldsymbol{\lambda}(t)\Big|_{t_0}^{t_f} - \int_{t_0}^{t_f} \mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})^{\mathrm{T}}\dot{\boldsymbol{\lambda}}(t)dt. \tag{3.2.9}$$

Combining (3.2.9) with (3.2.8) results in

$$\begin{aligned}\int_{t_0}^{t_f} & \mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})^{\mathrm{T}}\hat{\boldsymbol{\nabla}}_{\mathbf{x}}g(t,\mathbf{v}(t,\mathbf{p})) + \hat{\boldsymbol{\nabla}}_{\mathbf{p}}g(t,\mathbf{v}(t,\mathbf{p})) - \\ & \mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})^{\mathrm{T}}\hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}(t) - \hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}(t) - \mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})^{\mathrm{T}}\dot{\boldsymbol{\lambda}}(t)dt + \\ & \mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})^{\mathrm{T}}\boldsymbol{\lambda}(t)\Big|_{t_0}^{t_f}.\end{aligned} \tag{3.2.10}$$

After collecting terms multiplying $\mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})$ in (3.2.10) and substituting the right-hand side expression in (3.2.6) for $\boldsymbol{\lambda}(t)$, the desired result in (3.2.7) is obtained. Strict differentiability follows from the fact that all possible measurable selections $\mathbf{w}_t$ differ from $\hat{\mathbf{w}}_t$ only if t is in a measure zero subset of T, and therefore W is a singleton. □

Theorem 3.2.5. *Let the hypotheses of Theorem 3.2.3 hold. Let $h : \mathcal{T}_0 \times \mathcal{P} \times \mathcal{X} \to \mathbb{R}$ be a locally Lipschitz continuous function where $\mathcal{T}_0$ is an open subset of $\mathcal{T}$ such that $t_f \in \mathcal{T}_0$.*

Let $G : \mathcal{P} \to \mathbb{R} : \boldsymbol{\eta} \mapsto h(t_f, \mathbf{v}(t_f, \boldsymbol{\eta}))$ and assume $\partial_{\mathbf{v}} h(t_f, \mathbf{v}(t_f, \mathbf{p}))$ is a singleton whose single element is $(\boldsymbol{\nabla}_{\mathbf{p}} h(t_f, \mathbf{v}(t_f, \mathbf{p}), \boldsymbol{\nabla}_{\mathbf{x}} h(t_f, \mathbf{v}(t_f, \mathbf{p}))$ where $\boldsymbol{\nabla}_{\mathbf{p}} h(t_f, \mathbf{v}(t_f, \mathbf{p}) \in \mathbb{R}^{n_p}$ and $\boldsymbol{\nabla}_{\mathbf{x}} h(t_f, \mathbf{v}(t_f, \mathbf{p})) \in \mathbb{R}^{n_x}$.

Let $\mathbf{C}_{\mathbf{x}} = \boldsymbol{\nabla}_{\mathbf{x}} h(t_f, \mathbf{v}(t_f, \mathbf{p}))$ and $\mathbf{C}_{\mathbf{p}} = \boldsymbol{\nabla}_{\mathbf{p}} h(t_f, \mathbf{v}(t_f, \mathbf{p}))$. Let $\boldsymbol{\lambda} : T \to \mathbb{R}^{n_x}$ be a solution of the initial value problem:

$$\dot{\boldsymbol{\lambda}}(t) = -\hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}(t), \forall t \in [t_0, t_f),\ \boldsymbol{\lambda}(t_f) = -\mathbf{C}_{\mathbf{x}}. \tag{3.2.11}$$

Then, it is unique and absolutely continuous. In addition, G is locally Lipschitz continuous

and strictly differentiable at $\mathbf{p}$ *and the derivative is*

$$\nabla G(\mathbf{p}) = \int_{t_0}^{t_f} -\hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}(t, \mathbf{v}(t, \mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}(t)dt - \mathbf{J}_{\mathbf{p}}\mathbf{x}(t_0, \mathbf{p})^{\mathrm{T}}\boldsymbol{\lambda}(t_0) + \mathbf{C}_{\mathbf{p}}. \tag{3.2.12}$$

Proof. The existence, uniqueness and absolute continuity of $\boldsymbol{\lambda}$ follows from similar arguments to those presented in Theorem 3.2.4.

G is locally Lipschitz continuous at $\mathbf{p}$ because it is the composition of locally Lipschitz continuous functions $h(t_f, \cdot)$ and the locally Lipschitz continuous function $\mathbf{v}(t_f, \cdot)$. Strict differentiability follows from the fact that $\partial_{\mathbf{v}} h(t_f, \mathbf{v}(t_f, \mathbf{p}))$ is a singleton and $\mathbf{v}(t_f, \cdot)$ is strictly differentiable at $\mathbf{p}$. The strict derivative is $\nabla G(\mathbf{p}) = \mathbf{J}_{\mathbf{p}}\mathbf{x}(t_f, \mathbf{p})^{\mathrm{T}}\mathbf{C}_{\mathbf{x}} + \mathbf{C}_{\mathbf{p}}$.

The expression

$$\begin{aligned}\int_{t_0}^{t_f} &\dot{\mathbf{J}}_{\mathbf{p}}\mathbf{x}(t, \mathbf{p})^{\mathrm{T}}\mathbf{C}_{\mathbf{x}} - \\ &(\hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}(t, \mathbf{v}(t, \mathbf{p}))\mathbf{J}_{\mathbf{p}}\mathbf{x}(t, \mathbf{p}) + \hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}(t, \mathbf{v}(t, \mathbf{p})) - \dot{\mathbf{J}}_{\mathbf{p}}\mathbf{x}(t, \mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}(t)dt + \\ &\mathbf{J}_{\mathbf{p}}\mathbf{x}(t_0, \mathbf{p})^{\mathrm{T}}\mathbf{C}_{\mathbf{x}}\end{aligned} \tag{3.2.13}$$

is equal to $\mathbf{J}_{\mathbf{p}}\mathbf{x}(t_f, \mathbf{p})^{\mathrm{T}}\mathbf{C}_{\mathbf{x}}$ regarding $\mathbf{C}_{\mathbf{x}}$ as a constant because the term multiplying $\boldsymbol{\lambda}(t)$ is identically zero as discussed in Theorem 3.2.4.

$\mathbf{J}_{\mathbf{p}}\mathbf{x}(\cdot, \mathbf{p})$ and $\boldsymbol{\lambda}$ are absolutely continuous functions from T to $\mathbb{R}^{n_x \times n_p}$ and $\mathbb{R}^{n_x}$, respectively, and therefore integration by parts for the Lebesgue integral produces

$$\int_{t_0}^{t_f} \dot{\mathbf{J}}_{\mathbf{p}}\mathbf{x}(t, \mathbf{p})^{\mathrm{T}}(\mathbf{C}_{\mathbf{x}} + \boldsymbol{\lambda}(t))dt = \mathbf{J}_{\mathbf{p}}\mathbf{x}(t, \mathbf{p})^{\mathrm{T}}(\mathbf{C}_{\mathbf{x}} + \boldsymbol{\lambda}(t))\Big|_{t_0}^{t_f} - \int_{t_0}^{t_f} \mathbf{J}_{\mathbf{p}}\mathbf{x}(t, \mathbf{p})^{\mathrm{T}}\dot{\boldsymbol{\lambda}}(t)dt.$$

Hence, the expression (3.2.13) can be written as

$$\int_{t_0}^{t_f} -(\hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}(t, \mathbf{v}(t, \mathbf{p}))\mathbf{J}_{\mathbf{p}}\mathbf{x}(t, \mathbf{p}) + \hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}(t, \mathbf{v}(t, \mathbf{p})))^{\mathrm{T}}\boldsymbol{\lambda}(t) - \mathbf{J}_{\mathbf{p}}\mathbf{x}(t, \mathbf{p})^{\mathrm{T}}\dot{\boldsymbol{\lambda}}(t)dt +$$

$$\mathbf{J_p x}(t,\mathbf{p})^{\mathrm{T}}(\mathbf{C_x}+\boldsymbol{\lambda}(t))\big|_{t_0}^{t_f}+\mathbf{J_p x}(t_0,\mathbf{p})^{\mathrm{T}}\mathbf{C_x},$$
$$\int_{t_0}^{t_f}\mathbf{J_p x}(t,\mathbf{p})^{\mathrm{T}}(-\hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}(t)-\dot{\boldsymbol{\lambda}}(t))-\hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}(t)dt+ \tag{3.2.14}$$
$$\mathbf{J_p x}(t_f,\mathbf{p})^{\mathrm{T}}(\mathbf{C_x}+\boldsymbol{\lambda}(t_f))-\mathbf{J_p x}(t_0,\mathbf{p})^{\mathrm{T}}\boldsymbol{\lambda}(t_0).$$

After substituting the right-hand side expression in (3.2.11) for $\dot{\boldsymbol{\lambda}}$, (3.2.14) becomes (3.2.12).

□

3.3 Differential-Algebraic Equations

Results in this section extend previous results to a subset of differential-algebraic equations.

Assumption 3.3.1. *Let* $\mathbf{F}:\mathcal{T}\times\mathcal{P}\times\mathcal{X}\times\mathcal{Y}\times\dot{\mathcal{X}}\to\mathbb{R}^{n_x+n_y}$ *and* $\mathbf{F}_0:\mathcal{P}\to\mathcal{X}$ *be locally Lipschitz continuous functions. Let* $\mathbf{x}:T\times\mathcal{P}\to\mathcal{X}$, $\mathbf{y}:T\times\mathcal{P}\to\mathcal{Y}$ *and* $\dot{\mathbf{x}}:T\times\mathcal{P}\to\dot{\mathcal{X}}$ *be such that they uniquely satisfy the initial value problem*

$$\mathbf{0}=\mathbf{F}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p}),\dot{\mathbf{x}}(t,\mathbf{p})),\ \forall t\in[t_0,t_f],\ \mathbf{x}(t_0,\mathbf{p})=\mathbf{F}_0(\mathbf{p}),\ \forall\mathbf{p}\in\mathcal{P},. \tag{3.3.1}$$

Let $\dot{\mathbf{x}}(t_0,\bar{\mathbf{p}})=\dot{\bar{\mathbf{x}}}$ *and* $\mathbf{y}(t_0,\bar{\mathbf{p}})=\bar{\mathbf{y}}$ *for some* $\bar{\mathbf{p}}\in\mathcal{P}$ *where* $\dot{\bar{\mathbf{x}}}$ *and* $\bar{\mathbf{y}}$ *are constants and satisfy* $\mathbf{F}(t_0,\bar{\mathbf{p}},\mathbf{x}(t_0,\bar{\mathbf{p}}),\mathbf{y}(t_0,\bar{\mathbf{p}}),\dot{\mathbf{x}}(t_0,\bar{\mathbf{p}}))=\mathbf{0}$. *Assume that this condition uniquely determines* $\mathbf{y}(t_0,\mathbf{p})$ *and* $\dot{\mathbf{x}}(t_0,\mathbf{p})$ *for all* $\mathbf{p}\in\mathcal{P}$.

Remark 3.3.2. Let $\mathbf{u}:T\times\mathcal{P}\to\mathcal{P}\times\mathcal{X}\times\mathcal{Y}\times\dot{\mathcal{X}}:(t,\mathbf{p})\mapsto(\mathbf{v}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p}),\dot{\mathbf{x}}(t,\mathbf{p}))$ for the remainder of this chapter.

Lemma 3.3.3. *Let* $\boldsymbol{\eta}_{\mathbf{p}}\in\mathcal{P}$, $\boldsymbol{\eta}_{\mathbf{x}}\in\mathcal{X}$, $\boldsymbol{\eta}_{\mathbf{y}}\in\mathcal{Y}$, $\boldsymbol{\eta}_{\dot{\mathbf{x}}}\in\dot{\mathcal{X}}$ *and* $\boldsymbol{\eta}=(\boldsymbol{\eta}_{\mathbf{p}},\boldsymbol{\eta}_{\mathbf{x}},\boldsymbol{\eta}_{\mathbf{y}},\boldsymbol{\eta}_{\dot{\mathbf{x}}})$.

Assume $\pi_{\mathbf{q}}\partial\mathbf{F}(t,\boldsymbol{\eta})$ *is maximal for all* $(t,\boldsymbol{\eta})\in\mathcal{T}\times\mathcal{P}\times\mathcal{X}\times\mathcal{Y}\times\dot{\mathcal{X}}$. *Then there exist locally Lipschitz continuous functions:* $\mathbf{f}:T\times\mathcal{P}\times\mathcal{X}\to\dot{\mathcal{X}}$ *and* $\mathbf{r}:T\times\mathcal{P}\times\mathcal{X}\to\mathcal{Y}$ *such that* $\mathbf{0}=\mathbf{F}(t,\boldsymbol{\eta}_{\mathbf{p}},\boldsymbol{\eta}_{\mathbf{x}},\boldsymbol{r}(t,\boldsymbol{\eta}_{\mathbf{p}},\boldsymbol{\eta}_{\mathbf{x}}),\mathbf{f}(t,\boldsymbol{\eta}_{\mathbf{p}},\boldsymbol{\eta}_{\mathbf{x}}))$ *holds.*

If $\partial \mathbf{F}(t, \boldsymbol{\eta})$ *is a singleton whose only member is*

$$\mathbf{JF}(t, \boldsymbol{\eta}) = \begin{bmatrix} \mathbf{J}_t\mathbf{F}(t, \boldsymbol{\eta}) & \mathbf{J}_\mathbf{p}\mathbf{F}(t, \boldsymbol{\eta}) & \mathbf{J}_\mathbf{x}\mathbf{F}(t, \boldsymbol{\eta}) & \mathbf{J}_\mathbf{y}\mathbf{F}(t\ \boldsymbol{\eta}) & \mathbf{J}_{\dot{\mathbf{x}}}\mathbf{F}(t, \boldsymbol{\eta}) \end{bmatrix}$$

then $\mathbf{f}$ *and* $\mathbf{r}$ *are strictly differentiable and the derivatives are the solutions of the equation,*

$$\begin{bmatrix} \mathbf{J}_\mathbf{y}\mathbf{F}(t, \boldsymbol{\eta}) & \mathbf{J}_{\dot{\mathbf{x}}}\mathbf{F}(t, \boldsymbol{\eta}) \end{bmatrix} \begin{bmatrix} \mathbf{J}_t\mathbf{r}(t, \boldsymbol{\eta}_\mathbf{p}, \boldsymbol{\eta}_\mathbf{x}) & \mathbf{J}_\mathbf{p}\mathbf{r}(t, \boldsymbol{\eta}_\mathbf{p}, \boldsymbol{\eta}_\mathbf{x}) & \mathbf{J}_\mathbf{x}\mathbf{r}(t, \boldsymbol{\eta}_\mathbf{p}, \boldsymbol{\eta}_\mathbf{x}) \\ \mathbf{J}_t\mathbf{f}(t, \boldsymbol{\eta}_\mathbf{p}, \boldsymbol{\eta}_\mathbf{x}) & \mathbf{J}_\mathbf{p}\mathbf{f}(t, \boldsymbol{\eta}_\mathbf{p}, \boldsymbol{\eta}_\mathbf{x}) & \mathbf{J}_\mathbf{x}\mathbf{f}(t, \boldsymbol{\eta}_\mathbf{p}, \boldsymbol{\eta}_\mathbf{x}) \end{bmatrix} =$$

$$- \begin{bmatrix} \mathbf{J}_t\mathbf{F}(t, \boldsymbol{\eta}) & \mathbf{J}_\mathbf{p}\mathbf{F}(t, \boldsymbol{\eta}) & \mathbf{J}_\mathbf{x}\mathbf{F}(t, \boldsymbol{\eta}) \end{bmatrix}.$$

Proof. The result follows from Theorem 2.6.13. □

Corollary 3.3.4. *Let Assumption 3.3.1 and the assumptions of Lemma 3.3.3 hold. Let* $\mathbf{u}(t, \bar{\mathbf{p}})$ *be the unique solution of* (3.3.1) *if* $\mathbf{p} = \bar{\mathbf{p}}$. *Then* $\mathbf{u}(t, \cdot)$ *is a locally Lipschitz continuous function at* $\bar{\mathbf{p}}$ *for all* $t \in T$.

Proof. Since the implicit function, $\mathbf{f}$ as defined in Lemma 3.3.3 is a locally Lipschitz continuous function, $\mathbf{v}(t, \cdot)$ is a locally Lipschitz continuous function at $\bar{\mathbf{p}}$ for all $t \in T$ per Theorem 3.1.3 considering the extended ODE in (3.2.3). The local Lipschitz continuity of $\mathbf{y}(t, \cdot)$ at $\bar{\mathbf{p}}$ follows from the local Lipschitz continuity of the implicit function $\mathbf{r}(t, \cdot)$ at $\mathbf{v}(t, \bar{\mathbf{p}})$ for all $t \in T$ and the local Lipschitz continuity of $\mathbf{v}(t, \cdot)$ at $\bar{\mathbf{p}}$ for all $t \in T$. The local Lipschitz continuity of $\dot{\mathbf{x}}(t, \cdot)$ at $\mathbf{p}$ follows from the same reasoning using $\mathbf{f}$ instead of $\mathbf{r}$. Since all elements of $\mathbf{u}(t, \cdot)$ are locally Lipschitz continuous at $\bar{\mathbf{p}}$ so is $\mathbf{u}(t, \cdot)$ for all $t \in T$. □

Lemma 3.3.5. *Let Assumption 3.3.1 and the assumptions of Lemma 3.3.3 hold. Let* $\partial_\mathbf{u}\mathbf{F}(t, \boldsymbol{\eta})$ *be a singleton whose element is*

$$\mathbf{J}_\mathbf{u}\mathbf{F}(t, \boldsymbol{\eta}) = \begin{bmatrix} \mathbf{J}_\mathbf{p}\mathbf{F}(t, \boldsymbol{\eta}) & \mathbf{J}_\mathbf{x}\mathbf{F}(t, \boldsymbol{\eta}) & \mathbf{J}_\mathbf{y}\mathbf{F}(t, \boldsymbol{\eta}) & \mathbf{J}_{\dot{\mathbf{x}}}\mathbf{F}(t, \boldsymbol{\eta}) \end{bmatrix}.$$

Then $\mathbf{f}(t,\cdot)$ *and* $\mathbf{r}(t,\cdot)$ *are strictly differentiable at* $(t,\boldsymbol{\eta})$. *The derivatives satisfy*

$$\begin{bmatrix}\mathbf{J_y F}(t,\boldsymbol{\eta})\ \mathbf{J_{\dot{x}} F}(t,\boldsymbol{\eta})\end{bmatrix}\begin{bmatrix}\mathbf{J_p r}(t,\boldsymbol{\eta_p},\boldsymbol{\eta_x}) & \mathbf{J_x r}(t,\boldsymbol{\eta_p},\boldsymbol{\eta_x})\\ \mathbf{J_p f}(t,\boldsymbol{\eta_p},\boldsymbol{\eta_x}) & \mathbf{J_x f}(t,\boldsymbol{\eta_p},\boldsymbol{\eta_x})\end{bmatrix} = -\begin{bmatrix}\mathbf{J_p F}(t,\boldsymbol{\eta})\ \mathbf{J_x F}(t,\boldsymbol{\eta})\end{bmatrix}. \quad (3.3.2)$$

Proof. If $\pi_{\mathbf{q}}\partial_{\mathbf{u}}\mathbf{F}(t,\boldsymbol{\eta})$ were maximal, then the result of the lemma would follow from Theorem 2.6.13. However, it is not obvious that if $\pi_{\mathbf{q}}\partial\mathbf{F}(t,\boldsymbol{\eta})$ is maximal, then $\pi_{\mathbf{q}}\partial_{\mathbf{u}}\mathbf{F}(t,\boldsymbol{\eta})$ is maximal.

In order to arrive at the desired result, Theorem 3.2 in [78] is used. $\partial f(p)$, L, and $\partial f(p)|_L$ (the restriction of $\partial f(p)$ to the subspace L), correspond to $\partial\mathbf{F}(t,\boldsymbol{\eta})$, $\mathbb{R}^{n_p\times n_x\times n_y\times n_x}$ and $\pi_{\mathbf{u}}\partial\mathbf{F}(t,\boldsymbol{\eta})$, respectively. $\partial_L f(p)$ is an intermediate construct that contains $\partial^C g(0)$ (see comment on page 57 in [78]). $\partial^C g(0)$ corresponds to $\partial_{\mathbf{u}}\mathbf{F}(t,\boldsymbol{\eta})$ in this case. Hence, $\partial_{\mathbf{u}}\mathbf{F}(t,\boldsymbol{\eta}) \subset \pi_{\mathbf{u}}\partial\mathbf{F}(t,\boldsymbol{\eta})$ and $\pi_{\mathbf{q}}(\pi_{\mathbf{u}}\partial\mathbf{F}(t,\boldsymbol{\eta})) = \pi_{\mathbf{q}}\partial\mathbf{F}(t,\boldsymbol{\eta}) \supset \pi_{\mathbf{q}}\partial_{\mathbf{u}}\mathbf{F}(t,\boldsymbol{\eta})$. As a result, $\pi_{\mathbf{q}}\partial_{\mathbf{u}}\mathbf{F}(t,\boldsymbol{\eta})$ is maximal. □

Theorem 3.3.6. *Let Assumption 3.3.1 and the assumptions of Lemma* 3.3.3 *hold. Let* S *be a measure zero subset of* T. *Let* $\partial_{\mathbf{u}}\mathbf{F}(t,\mathbf{u}(t,\mathbf{p}))$ *be a singleton for* $t\in T\backslash S$. *Assume* $\mathbf{F_0}$ *is strictly differentiable at* $\mathbf{p}$. *Then* $\partial_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})$ *is a singleton for all* $t\in T$. *Let the single element be* $\mathbf{J_p x}(t,\mathbf{p})$. *Then* $\mathbf{J_p x}(\cdot,\mathbf{p})$ *is the unique absolutely continuous function on* T *that satisfies*

$$\begin{aligned}
\partial_{\mathbf{u}}\mathbf{F}(t,\mathbf{u}(t,\mathbf{p})) &= \{[\boldsymbol{\pi_v}\mathbf{J_u F}(t,\mathbf{u}(t,\mathbf{p}))\ \boldsymbol{\pi_q}\mathbf{J_u F}(t,\mathbf{u}(t,\mathbf{p}))]\},\ \forall t\in T\backslash S\\
\mathbf{J}(t,\mathbf{p}) &= -\boldsymbol{\pi_q}\mathbf{J_u F}(t,\mathbf{u}(t,\mathbf{p}))^{-1}\boldsymbol{\pi_v}\mathbf{J_u F}(t,\mathbf{u}(t,\mathbf{p})),\ \forall t\in T\backslash S, \qquad (3.3.3)\\
\mathbf{J}(t,\mathbf{p}) &= \begin{bmatrix}\mathbf{J_p r}(t,\mathbf{v}(t,\mathbf{p})) & \mathbf{J_x r}(t,\mathbf{v}(t,\mathbf{p}))\\ \mathbf{J_p f}(t,\mathbf{v}(t,\mathbf{p})) & \mathbf{J_x f}(t,\mathbf{v}(t,\mathbf{p}))\end{bmatrix},\ \forall t\in T\backslash S,\\
\dot{\mathbf{J}}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p}) &= \hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))\mathbf{J_p x}(t,\mathbf{p}) + \hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p})),\ \forall t\in(t_0,t_f], \qquad (3.3.4)\\
\mathbf{J_p x}(t_0,\mathbf{p}) &= \mathbf{J_p F_0}(\mathbf{p})
\end{aligned}$$

where $\boldsymbol{\pi_v}\mathbf{J_u F}(t,\mathbf{u}(t,\mathbf{p})) \in \mathbb{R}^{(n_x+n_y)\times(n_p+n_x)}$, $\boldsymbol{\pi_q}\mathbf{J_u F}(t,\mathbf{u}(t,\mathbf{p})) \in \mathbb{R}^{(n_x+n_y)\times(n_y+n_x)}$,

$\mathbf{J_p f}(t, \mathbf{v}(t, \mathbf{p})) \in \mathbb{R}^{n_x \times n_p}$, $\mathbf{J_x f}(t, \mathbf{v}(t, \mathbf{p})) \in \mathbb{R}^{n_x \times n_x}$, $\mathbf{J_p r}(t, \mathbf{v}(t, \mathbf{p})) \in \mathbb{R}^{n_y \times n_p}$ *and* $\mathbf{J_x r}(t, \mathbf{v}(t, \mathbf{p}))$ $\in \mathbb{R}^{n_y \times n_x}$.

Proof. Result follows from Theorem 3.2.3, Lemma 3.3.3 and Lemma 3.3.5. □

Corollary 3.3.7. *Let the hypotheses of Theorem 3.3.6 hold. Then* $\mathbf{y}(\cdot, \mathbf{p})$ *and* $\dot{\mathbf{x}}(\cdot, \mathbf{p})$ *are strictly differentiable for all* $t \in T \backslash S$ *and the derivatives are*

$$\mathbf{J_p y}(t, \mathbf{p}) = \mathbf{J_x r}(t, \mathbf{v}(t, \mathbf{p}))\mathbf{J_p x}(t, \mathbf{p}) + \mathbf{J_p r}(t, \mathbf{v}(t, \mathbf{p})), \ \forall t \in T \backslash S,$$
$$\mathbf{J_p f}(t, \mathbf{p}) = \mathbf{J_x f}(t, \mathbf{v}(t, \mathbf{p}))\mathbf{J_p x}(t, \mathbf{p}) + \mathbf{J_p f}(t, \mathbf{v}(t, \mathbf{p})), \ \forall t \in T \backslash S,$$

where $\mathbf{J_p r}$, $\mathbf{J_x r}$, $\mathbf{J_p f}$ *and* $\mathbf{J_x f}$ *are as defined in Lemma 3.3.5.*

Proof. The result is obtained by applying Theorem 2.6.7 to the implicit functions $\mathbf{r}$ and $\mathbf{f}$. □

Remark 3.3.8. *The results of Theorems 3.2.4 and 3.2.5 hold for the initial value problem in* (3.3.1) *if* $\hat{\mathbf{J}}_\mathbf{x}\mathbf{f}$ *and* $\hat{\mathbf{J}}_\mathbf{p}\mathbf{f}$ *are obtained with* $\mathbf{J_x f}$ *and* $\mathbf{J_p f}$ *that are computed using* (3.3.3).

The next theorem is an extension of Theorem 3.2.4 where g is a function from $\mathcal{T} \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}}$ to $\mathbb{R}$ instead of $\mathcal{T} \times \mathcal{P} \times \mathcal{X}$ to $\mathbb{R}$. The extended result is obtained by replacing $\dot{\mathbf{x}}$ and $\mathbf{y}$ with the implicit functions $\mathbf{f}$ and $\mathbf{r}$.

Theorem 3.3.9. *Let the hypotheses of Theorem* 3.3.6 *hold. Let* $g : \mathcal{T} \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}} \to \mathbb{R}$ *be a locally Lipschitz continuous function. Let* $G : \mathcal{P} \to \mathbb{R}$ *be*

$$G(\mathbf{p}) = \int_{t_0}^{t_f} g(t, \mathbf{u}(t, \mathbf{p}))dt. \tag{3.3.5}$$

Let Q *be a measure zero subset of* T. *Let* $\partial_\mathbf{u} g(t, \mathbf{u}(t, \mathbf{p}))$ *be a singleton for all* $t \in T \backslash Q$.

Let

$$\begin{aligned}\mathbf{B}_{\mathbf{x}}(t,\mathbf{u}(t,\mathbf{p})) &= \hat{\boldsymbol{\nabla}}_{\mathbf{x}} g(t,\mathbf{u}(t,\mathbf{p})) + \hat{\mathbf{J}}_{\mathbf{x}}\mathbf{r}(t,\mathbf{v}(t,\mathbf{p}))^{\mathrm{T}}\hat{\boldsymbol{\nabla}}_{\mathbf{y}} g(t,\mathbf{u}(t,\mathbf{p})) + \\ &\quad \hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))^{\mathrm{T}}\hat{\boldsymbol{\nabla}}_{\dot{\mathbf{x}}} g(t,\mathbf{u}(t,\mathbf{p})), \\ \mathbf{B}_{\mathbf{p}}(t,\mathbf{u}(t,\mathbf{p})) &= \hat{\boldsymbol{\nabla}}_{\mathbf{p}} g(t,\mathbf{u}(t,\mathbf{p})) + \hat{\mathbf{J}}_{\mathbf{p}}\mathbf{r}(t,\mathbf{v}(t,\mathbf{p}))^{\mathrm{T}}\hat{\boldsymbol{\nabla}}_{\mathbf{y}} g(t,\mathbf{u}(t,\mathbf{p})) + \\ &\quad \hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))^{\mathrm{T}}\hat{\boldsymbol{\nabla}}_{\dot{\mathbf{x}}} g(t,\mathbf{u}(t,\mathbf{p})).\end{aligned}$$

Let $\boldsymbol{\lambda} : T \to \mathbb{R}^{n_x}$ *be a solution of the initial value problem*

$$\dot{\boldsymbol{\lambda}}(t) = -\hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}(t) + \mathbf{B}_{\mathbf{x}}(t,\mathbf{u}(t,\mathbf{p})),\ \forall t \in [t_0, t_f),\ \boldsymbol{\lambda}(t_f) = \mathbf{0}. \tag{3.3.6}$$

Then, it is unique and absolutely continuous. In addition, G is locally Lipschitz continuous and strictly differentiable at $\mathbf{p}$ *and the strict derivative is*

$$\boldsymbol{\nabla} G(\mathbf{p}) = \int_{t_0}^{t_f} \mathbf{B}_{\mathbf{p}}(t,\mathbf{u}(t,\mathbf{p})) - \hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}(t)dt\ + \mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})^{\mathrm{T}}\boldsymbol{\lambda}(t)\big|_{t_0}^{t_f}. \tag{3.3.7}$$

Proof. The proof proceeds along similar lines as the proof of Theorem 3.2.4 considering the equivalent integral

$$G(\mathbf{p}) = \int_{t_0}^{t_f} g(t,\mathbf{v}(t,\mathbf{p}),\mathbf{r}(t,\mathbf{v}(t,\mathbf{p})),\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))) - \boldsymbol{\lambda}(t)^{\mathrm{T}}(\mathbf{f}(t,\mathbf{v}(t,\mathbf{p})) - \dot{\mathbf{x}}(t,\mathbf{p}))dt.$$

The existence and uniqueness of solutions to (3.3.6) can be shown using arguments similar to those in Theorem 3.2.4.

The local Lipschitz continuity of G at $\mathbf{p}$ and the applicability of Theorem 3.1.2 follows from arguments similar to those presented in Theorem 3.2.4.

All measurable selections $\mathbf{w}_t$ differ from $\hat{\mathbf{w}}_t$ on a set of measure zero where

$$\begin{aligned}
\hat{\mathbf{w}}_t(t) = {} & \mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})^{\mathrm{T}}\hat{\boldsymbol{\nabla}}_{\mathbf{x}}g(t,\mathbf{u}(t,\mathbf{p})) + \hat{\boldsymbol{\nabla}}_{\mathbf{p}}g(t,\mathbf{u}(t,\mathbf{p})) + \\
& \mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})^{\mathrm{T}}\hat{\mathbf{J}}_{\mathbf{x}}\mathbf{r}(t,\mathbf{v}(t,\mathbf{p}))^{\mathrm{T}}\hat{\boldsymbol{\nabla}}_{\mathbf{y}}g(t,\mathbf{u}(t,\mathbf{p})) + \hat{\mathbf{J}}_{\mathbf{p}}\mathbf{r}(t,\mathbf{v}(t,\mathbf{p}))^{\mathrm{T}}\hat{\boldsymbol{\nabla}}_{\mathbf{y}}g(t,\mathbf{u}(t,\mathbf{p}))+ \\
& \mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})^{\mathrm{T}}\hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))^{\mathrm{T}}\hat{\boldsymbol{\nabla}}_{\dot{\mathbf{x}}}g(t,\mathbf{u}(t,\mathbf{p})) + \hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))^{\mathrm{T}}\hat{\boldsymbol{\nabla}}_{\dot{\mathbf{x}}}g(t,\mathbf{u}(t,\mathbf{p})) - \\
& \tilde{\mathbf{w}}_t(t)^{\mathrm{T}}\boldsymbol{\lambda}(t), \\
\hat{\mathbf{w}}_t(t) = {} & \mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})^{\mathrm{T}}\mathbf{B}_{\mathbf{x}}(t,\mathbf{u}(t,\mathbf{p})) + \mathbf{B}_{\mathbf{p}}(t,\mathbf{u}(t,\mathbf{p})) - \tilde{\mathbf{w}}_t(t)^{\mathrm{T}}\boldsymbol{\lambda}(t).
\end{aligned}$$

and $\tilde{\mathbf{w}}_t$ is defined in Theorem 3.2.4. Applying integration by parts to $\dot{\mathbf{J}}(t,\mathbf{p})^{\mathrm{T}}\boldsymbol{\lambda}(t)$ and collecting terms multiplying $\mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})$ results in (3.3.7) and (3.3.6). □

Since $\mathbf{y}$ and $\dot{\mathbf{x}}$ are strictly differentiable functions of the parameters only if $t \in T\backslash S$, the extension of Theorem 3.2.5 considers the case when $t_f \in T\backslash S$.

Theorem 3.3.10. *Let the hypotheses of Theorem 3.3.6 hold. Let* $h : \mathcal{T}_0 \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}} \to \mathbb{R}$ *be a locally Lipschitz continuous function where* $\mathcal{T}_0$ *is an open subset of* $\mathcal{T}$ *such that* $t_f \in \mathcal{T}_0$. *Assume* $t_f \notin S$.

Let $G : \mathcal{P} \to \mathbb{R} : \boldsymbol{\eta} \mapsto h(t_f, \mathbf{u}(t_f, \boldsymbol{\eta}))$ *and assume* $\partial_{\mathbf{u}}h(t_f, \mathbf{u}(t_f, \mathbf{p}))$ *is a singleton whose single element is* $(\boldsymbol{\nabla}_{\mathbf{p}}h(t_f, \mathbf{u}(t_f, \mathbf{p}), \boldsymbol{\nabla}_{\mathbf{x}}h(t_f, \mathbf{u}(t_f, \mathbf{p}), \boldsymbol{\nabla}_{\mathbf{y}}h(t_f, \mathbf{u}(t_f, \mathbf{p}), \boldsymbol{\nabla}_{\dot{\mathbf{x}}}h(t_f, \mathbf{u}(t_f, \mathbf{p}))$ *where* $\boldsymbol{\nabla}_{\mathbf{p}}h(t_f, \mathbf{u}(t_f, \mathbf{p}) \in \mathbb{R}^{n_p}$, $\boldsymbol{\nabla}_{\mathbf{x}}h(t_f, \mathbf{u}(t_f, \mathbf{p})) \in \mathbb{R}^{n_x}$, $\boldsymbol{\nabla}_{\mathbf{y}}h(t_f, \mathbf{u}(t_f, \mathbf{p})) \in \mathbb{R}^{n_y}$ *and* $\boldsymbol{\nabla}_{\dot{\mathbf{x}}}h(t_f, \mathbf{u}(t_f, \mathbf{p}))$ $\in \mathbb{R}^{n_x}$.

Let

$$\begin{aligned}
\mathbf{C}_{\mathbf{x}} = {} & \mathbf{J}_{\mathbf{x}}\mathbf{f}(t_f,\mathbf{v}(t_f,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\nabla}_{\dot{\mathbf{x}}}h(t_f,\mathbf{u}(t_f,\mathbf{p})) + \\
& \mathbf{J}_{\mathbf{x}}\mathbf{r}(t_f,\mathbf{v}(t_f,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\nabla}_{\mathbf{y}}h(t_f,\mathbf{u}(t_f,\mathbf{p})) + \boldsymbol{\nabla}_{\mathbf{x}}h(t_f,\mathbf{u}(t_f,\mathbf{p})), \\
\mathbf{C}_{\mathbf{p}} = {} & \mathbf{J}_{\mathbf{p}}\mathbf{f}(t_f,\mathbf{v}(t_f,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\nabla}_{\dot{\mathbf{x}}}h(t_f,\mathbf{u}(t_f,\mathbf{p})) + \\
& \mathbf{J}_{\mathbf{p}}\mathbf{r}(t_f,\mathbf{v}(t_f,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\nabla}_{\mathbf{y}}h(t_f,\mathbf{u}(t_f,\mathbf{p})) + \boldsymbol{\nabla}_{\mathbf{p}}h(t_f,\mathbf{u}(t_f,\mathbf{p})).
\end{aligned}$$

Let $\boldsymbol{\lambda} : T \to \mathbb{R}^{n_x}$ *be a solution of the initial value problem;*

$$\dot{\boldsymbol{\lambda}}(t) = -\hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}(t, \mathbf{v}(t, \mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}(t), \forall t \in [t_0, t_f), \boldsymbol{\lambda}(t_f) = -\mathbf{C}_{\mathbf{x}}. \tag{3.3.8}$$

Then, it is unique and absolutely continuous. In addition, G *is locally Lipschitz continuous and strictly differentiable at* $\mathbf{p}$ *and the derivative is*

$$\boldsymbol{\nabla} G(\mathbf{p}) = \int_{t_0}^{t_f} -\hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}(t, \mathbf{v}(t, \mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}(t)dt \; - \mathbf{J}_{\mathbf{p}}\mathbf{x}(t_0, \mathbf{p})^{\mathrm{T}}\boldsymbol{\lambda}(t_0) + \mathbf{C}_{\mathbf{p}}. \tag{3.3.9}$$

Proof. The existence, uniqueness and absolute continuity of $\boldsymbol{\lambda}$ follows from similar arguments to those presented in Theorem 3.2.4.

G is locally Lipschitz continuous at $\mathbf{p}$ because it is the composition of locally Lipschitz continuous functions $h(t_f, \cdot)$ and the locally Lipschitz continuous function $\mathbf{u}(t_f, \cdot)$. Strict differentiability follows from the fact that $\partial_{\mathbf{u}} h(t_f, \mathbf{u}(t_f, \mathbf{p}))$ is a singleton and $\mathbf{u}(t_f, \cdot)$ is strictly differentiable at $\mathbf{p}$. The strict derivative is

$$\begin{aligned}\boldsymbol{\nabla} G(\mathbf{p}) = \dot{\mathbf{J}}_{\mathbf{p}}\mathbf{x}(t_f, \mathbf{p})^{\mathrm{T}}\boldsymbol{\nabla}_{\dot{\mathbf{x}}} h(t_f, \mathbf{u}(t_f, \mathbf{p})) + \mathbf{J}_{\mathbf{p}}\mathbf{y}(t_f, \mathbf{p})^{\mathrm{T}}\boldsymbol{\nabla}_{\mathbf{y}} h(t_f, \mathbf{u}(t_f, \mathbf{p})) + \\ \mathbf{J}_{\mathbf{p}}\mathbf{x}(t_f, \mathbf{p})^{\mathrm{T}}\boldsymbol{\nabla}_{\mathbf{x}} h(t_f, \mathbf{u}(t_f, \mathbf{p})) + \boldsymbol{\nabla}_{\mathbf{p}} h(t_f, \mathbf{u}(t_f, \mathbf{p})).\end{aligned} \tag{3.3.10}$$

$\dot{\mathbf{J}}_{\mathbf{p}}\mathbf{x}(t_f, \mathbf{p})$ and $\mathbf{J}_{\mathbf{p}}\mathbf{y}(t_f, \mathbf{p})$ are equal to $\mathbf{J}_{\mathbf{x}}\mathbf{f}(t_f, \mathbf{v}(t_f, \mathbf{p}))\mathbf{J}_{\mathbf{p}}\mathbf{x}(t_f, \mathbf{p}) + \mathbf{J}_{\mathbf{p}}\mathbf{f}(t_f, \mathbf{v}(t_f, \mathbf{p}))$ and $\mathbf{J}_{\mathbf{x}}\mathbf{r}(t_f, \mathbf{v}(t_f, \mathbf{p}))\, \mathbf{J}_{\mathbf{p}}\mathbf{x}(t_f, \mathbf{p}) + \mathbf{J}_{\mathbf{p}}\mathbf{r}(t_f, \mathbf{v}(t_f, \mathbf{p}))$, respectively. Substituting these expressions into (3.3.10) and collecting the terms multiplying $\mathbf{J}_{\mathbf{p}}\mathbf{x}(t_f, \mathbf{p})$ results in

$$\boldsymbol{\nabla} G(\mathbf{p}) = \mathbf{J}_{\mathbf{p}}\mathbf{x}(t_f, \mathbf{p})^{\mathrm{T}}\mathbf{C}_{\mathbf{x}} + \mathbf{C}_{\mathbf{p}}.$$

The expression

$$\int_{t_0}^{t_f} \dot{\mathbf{J}}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})^{\mathrm{T}}\mathbf{C}_{\mathbf{x}} - (\hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))\mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p}) + \tag{3.3.11}$$
$$\hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p})) - \dot{\mathbf{J}}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}(t)dt + \mathbf{J}_{\mathbf{p}}\mathbf{x}(t_0,\mathbf{p})^{\mathrm{T}}\mathbf{C}_{\mathbf{x}}$$

is equal to $\mathbf{J}_{\mathbf{p}}\mathbf{x}(t_f,\mathbf{p})^{\mathrm{T}}\mathbf{C}_{\mathbf{x}}$ because the term multiplying $\boldsymbol{\lambda}(t)$ is identically zero as discussed in Theorem 3.2.4. $\mathbf{J}_{\mathbf{p}}\mathbf{x}(\cdot,\mathbf{p})$ and $\boldsymbol{\lambda}$ are absolutely continuous functions from T to $\mathbb{R}^{n_x\times n_p}$ and $\mathbb{R}^{n_x}$, respectively, and therefore integration by parts for the Lebesgue integral produces

$$\int_{t_0}^{t_f} \dot{\mathbf{J}}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})^{\mathrm{T}}(\mathbf{C}_{\mathbf{x}} + \boldsymbol{\lambda}(t))dt = \mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})^{\mathrm{T}}(\mathbf{C}_{\mathbf{x}} + \boldsymbol{\lambda}(t))\Big|_{t_0}^{t_f} - \int_{t_0}^{t_f} \mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})^{\mathrm{T}}\dot{\boldsymbol{\lambda}}(t)dt.$$

Hence, the expression (3.3.11) can be written as

$$\int_{t_0}^{t_f} \mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})^{\mathrm{T}}(-\hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}(t) - \dot{\boldsymbol{\lambda}}(t)) - \hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}(t,\mathbf{v}(t,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}(t)dt + \tag{3.3.12}$$
$$\mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})^{\mathrm{T}}(\mathbf{C}_{\mathbf{x}} + \boldsymbol{\lambda}(t))\Big|_{t_0}^{t_f} + \mathbf{J}_{\mathbf{p}}\mathbf{x}(t_0,\mathbf{p})^{\mathrm{T}}\mathbf{C}_{\mathbf{x}}.$$

After substituting the right-hand side expression in (3.3.8) for $\dot{\boldsymbol{\lambda}}$, (3.3.12) becomes the integral in (3.3.9). □

3.4 Multistage Systems

The previous forward and adjoint sensitivity results will be extended to dynamic systems whose evolutions are described by disparate differential-algebraic equations in consecutive intervals of time.

Assumption 3.4.1. *Let n_e be a finite positive integer and $\mathcal{I} = \{1,\ldots,n_e\}$. Let $\alpha_i \in \mathbb{R}$, $\beta_i \in \mathbb{R}$, $\alpha_i < \beta_i$, $\forall i \in \mathcal{I}$, $\alpha_{i+1} = \beta_i$, $\forall i \in \mathcal{I}\backslash\{n_e\}$, $-\infty < \alpha_1 < \beta_{n_e} < +\infty$. Let $T =$*

$\cup_{i=1}^{n_e}[\alpha_i, \beta_i]$ *and* $T \subset \mathcal{T}$. *Let* $\mathcal{T}_i$ *be an open subset of* $\mathcal{T}$ *such that* $[\alpha_i, \beta_i] \subset \mathcal{T}_i$ *for all* $i \in \mathcal{I}$. *Let* $\mathbf{x}_i : [\alpha_i, \beta_i] \times \mathcal{P} \to \mathcal{X}$, $\mathbf{y}_i : [\alpha_i, \beta_i] \times \mathcal{P} \to \mathcal{Y}$, $\dot{\mathbf{x}}_i : [\alpha_i, \beta_i] \times \mathcal{P} \to \dot{\mathcal{X}}$ *for all* $i \in \mathcal{I}$, $\mathbf{x} : T \times \mathcal{P} \to \mathcal{X}$, $\mathbf{y} : T \times \mathcal{P} \to \mathcal{Y}$ *and* $\dot{\mathbf{x}} : T \times \mathcal{P} \to \dot{\mathcal{X}}$. *Assume* $\mathbf{F}_i : \mathcal{T}_i \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}} \to \mathbb{R}^{n_x+n_y}$ *are locally Lipschitz continuous functions for all* $i \in \mathcal{I}$. *Assume* $\mathbf{F}_i^0 : \mathcal{P} \times \mathcal{X} \to \mathcal{X}$ *for all* $i \in \mathcal{I}\backslash\{1\}$ *and* $\mathbf{F}_1^0 : \mathcal{P} \to \mathcal{X}$ *are locally Lipschitz continuous functions.*

The parametric sensitivities associated with the solutions of the initial value problem,

$$\mathbf{0} = \mathbf{F}_i(t, \mathbf{p}, \mathbf{x}_i(t, \mathbf{p}), \mathbf{y}_i(t, \mathbf{p}), \dot{\mathbf{x}}_i(t, \mathbf{p})), \ \forall t \in [\alpha_i, \beta_i], \ \forall i \in \mathcal{I}, \tag{3.4.1}$$

$$\mathbf{0} = \mathbf{x}_1(\alpha_1, \mathbf{p}) - \mathbf{F}_1^0(\mathbf{p}),$$

$$\mathbf{0} = \mathbf{x}_i(\alpha_i, \mathbf{p}) - \mathbf{F}_i^0(\mathbf{p}, \mathbf{x}_{i-1}(\beta_{i-1}, \mathbf{p})), \ \forall i \in \mathcal{I}\backslash\{1\}, \tag{3.4.2}$$

$$\mathbf{0} = \mathbf{x}(t, \mathbf{p}) - \mathbf{x}_i(t, \mathbf{p}), \ \forall t \in [\alpha_i, \beta_i), \ \forall i \in \mathcal{I},$$

$$\mathbf{0} = \mathbf{x}(\beta_{n_e}, \mathbf{p}) - \mathbf{x}_{n_e}(\beta_{n_e}, \mathbf{p}),$$

$$\mathbf{0} = \mathbf{y}(t, \mathbf{p}) - \mathbf{y}_i(t, \mathbf{p}), \ \forall t \in [\alpha_i, \beta_i), \ \forall i \in \mathcal{I},$$

$$\mathbf{0} = \mathbf{y}(\beta_{n_e}, \mathbf{p}) - \mathbf{y}_{n_e}(\beta_{n_e}, \mathbf{p}),$$

$$\mathbf{0} = \dot{\mathbf{x}}(t, \mathbf{p}) - \dot{\mathbf{x}}_i(t, \mathbf{p}), \ \forall t \in [\alpha_i, \beta_i), \ \forall i \in \mathcal{I},$$

$$\mathbf{0} = \dot{\mathbf{x}}(\beta_{n_e}, \mathbf{p}) - \dot{\mathbf{x}}_{n_e}(\beta_{n_e}, \mathbf{p})$$

are the focus of this section.

Remark 3.4.2. $\mathbf{x}(\cdot, \mathbf{p})$, $\mathbf{y}(\cdot, \mathbf{p})$ *and* $\dot{\mathbf{x}}(\cdot, \mathbf{p})$ *might be discontinuous at* $t = \alpha_i$ *with* $i > 1$.

Remark 3.4.3. *The results derived in this section are applicable with slight modifications to the case where the number of states, number of algebraic variables as well as the domains of the functions* $\mathbf{F}_i$ *and* $\mathbf{F}_i^0$ *differ for each* $i \in \mathcal{I}$.

Assumption 3.4.4. *Let* $\dot{\mathbf{x}}(\alpha_i, \bar{\mathbf{p}}) = \bar{\dot{\mathbf{x}}}_i$ *and* $\mathbf{y}(\alpha_i, \bar{\mathbf{p}}) = \bar{\mathbf{y}}_i$ *for all* $i \in \mathcal{I}$ *where* $\bar{\dot{\mathbf{x}}}_i$ *and* $\bar{\mathbf{y}}_i$ *are constants. Assume that this condition is sufficient to uniquely determine* $\dot{\mathbf{x}}(\alpha_i, \mathbf{p})$ *and* $\mathbf{y}(\alpha_i, \mathbf{p})$ *uniquely for all* $i \in \mathcal{I}$ *and for all* $\mathbf{p} \in \mathcal{P}$.

Assumption 3.4.5. *Let* $(\mathbf{x}_i(\cdot,\mathbf{p}),\mathbf{y}_i(\cdot,\mathbf{p}),\dot{\mathbf{x}}_i(\cdot,\mathbf{p}))$, $\forall i\in\mathcal{I}$ *be the unique solution of* (3.4.1). *Let* $\mathbf{z}_i:[\alpha_i,\beta_i]\times\mathcal{P}\to\mathcal{P}:(t,\mathbf{p})\mapsto\mathbf{p}$, $\mathbf{v}_i:[\alpha_i,\beta_i]\times\mathcal{P}\to\mathcal{P}\times\mathcal{X}:(t,\mathbf{p})\mapsto((\mathbf{z}_i(t,\mathbf{p})),\mathbf{x}_i(t,\mathbf{p}))$ *and* $\mathbf{u}_i:[\alpha_i,\beta_i]\times\mathcal{P}\to\mathcal{P}\times\mathcal{X}\times\mathcal{Y}\times\dot{\mathcal{X}}:(t,\mathbf{p})\mapsto(\mathbf{v}_i(t,\mathbf{p}),\mathbf{y}_i(t,\mathbf{p}),\dot{\mathbf{x}}_i(t,\mathbf{p}))$. *Let* $\mathbf{u}:T\times\mathcal{P}\to\mathcal{P}\times\mathcal{X}\times\mathcal{Y}\times\dot{\mathcal{X}}$ *be such that* $\mathbf{u}(t,\mathbf{p})=\mathbf{u}_i(t,\mathbf{p})$ *for all* $t\in[\alpha_i,\beta_i)$ *and* $\mathbf{u}(\beta_{n_e},\mathbf{p})=\mathbf{u}_{n_e}(\beta_{n_e},\mathbf{p})$.

Corollary 3.4.6. *Let Assumptions 3.4.1 and 3.4.4 hold. Let the assumptions of Lemma* 3.3.3 *hold for all* $\mathbf{F}_i$, $i\in\mathcal{I}$. *Let* $(\mathbf{x}_i(\cdot,\mathbf{p}),\ \mathbf{y}_i(\cdot,\mathbf{p}),\ \dot{\mathbf{x}}_i(\cdot,\mathbf{p}))$, $\forall i\in\mathcal{I}$ *be the solution of* (3.4.1). *Then* $\mathbf{u}(t,\cdot)$ *is locally Lipschitz continuous at* $\mathbf{p}$ *for all* $t\in T$.

Proof. Let $n_e=1$. Then $\mathbf{u}_1(t,\cdot)$ is a locally Lipschitz continuous function at $\mathbf{p}$ for $t\in[\alpha_1,\beta_1]$ per Corollary 3.3.4. Since the composition of locally Lipschitz continuous functions is locally Lipschitz continuous and $\mathbf{F}_2^0$ is a locally Lipschitz continuous function, $\mathbf{u}_2(\alpha_2,\cdot)$ is locally Lipschitz continuous at $\mathbf{p}$ if $n_e=2$. Then $\mathbf{u}_2(t,\cdot)$ for all $t\in[\alpha_2,\beta_2]$ is locally Lipschitz continuous at $\mathbf{p}$ per Corollary 3.3.4. The final result follows from the repeated application of Corollary 3.3.4 and composition rule for locally Lipschitz continuous functions for $n_e>2$ as has been done for the case $n_e\leq 2$. □

Theorem 3.4.7. *Let Assumptions 3.4.1 and 3.4.4 hold. Let the assumptions of Lemma 3.3.3 hold for for all* $i\in\mathcal{I}$. *Let* S *be a measure zero subset of* T. *Let* $\partial_{\mathbf{u}}\mathbf{F}_i(t,\mathbf{u}_i(t,\mathbf{p}))$ *be a singleton for all* $t\in[\alpha_i,\beta_i]\backslash S$ *and for all* $i\in\mathcal{I}$. *Let* $\partial_{\mathbf{v}}\mathbf{F}_i^0(\mathbf{v}_i(\alpha_i,\mathbf{p}))$ *be a singleton for all* $i\in\mathcal{I}\backslash\{1\}$ *and* $\partial_{\mathbf{p}}\mathbf{F}_1^0(\mathbf{p})$ *be a singleton.*

Then $\mathbf{x}(t,\cdot)$ *is locally Lipschitz continuous and strictly differentiable at* $\mathbf{p}$ *for all* $t\in T$. $\mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})$, *the single element of* $\partial_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})$, *is the unique function that satisfies*

$$\begin{aligned}
\partial_{\mathbf{u}}\mathbf{F}_i(t,\mathbf{u}_i(t,\mathbf{p}))&=\{[\boldsymbol{\pi}_{\mathbf{v}}\mathbf{J}_{\mathbf{u}}\mathbf{F}_i(t,\mathbf{u}_i(t,\mathbf{p}))\ \boldsymbol{\pi}_{\mathbf{q}}\mathbf{J}_{\mathbf{u}}\mathbf{F}_i(t,\mathbf{u}_i(t,\mathbf{p}))]\},\ \forall t\in[\alpha_i,\beta_i]\backslash S,\\
-\mathbf{J}_i(t,\mathbf{p})&=\boldsymbol{\pi}_{\mathbf{q}}\mathbf{J}_{\mathbf{u}}\mathbf{F}_i(t,\mathbf{u}_i(t,\mathbf{p}))^{-1}\boldsymbol{\pi}_{\mathbf{v}}\mathbf{J}_{\mathbf{u}}\mathbf{F}_i(t,\mathbf{u}_i(t,\mathbf{p})),\ \forall t\in[\alpha_i,\beta_i]\backslash S,\qquad(3.4.3)\\
\mathbf{J}_i(t,\mathbf{p})&=\begin{bmatrix}\mathbf{J}_{\mathbf{p}}\mathbf{r}_i(t,\mathbf{v}_i(t,\mathbf{p}))&\mathbf{J}_{\mathbf{x}}\mathbf{r}_i(t,\mathbf{v}_i(t,\mathbf{p}))\\ \mathbf{J}_{\mathbf{p}}\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p}))&\mathbf{J}_{\mathbf{x}}\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p}))\end{bmatrix},\ \forall t\in[\alpha_i,\beta_i]\backslash S,
\end{aligned}$$

$$\dot{\mathbf{J}}_{\mathbf{p}}\mathbf{x}_i(t,\mathbf{p}) = \hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p}))\mathbf{J}_{\mathbf{p}}\mathbf{x}_i(t,\mathbf{p})+$$
$$\hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p})),\ \forall t\in(\alpha_i,\beta_i], \tag{3.4.4}$$
$$\mathbf{J}_{\mathbf{p}}\mathbf{x}_1(\alpha_1,\mathbf{p}) = \mathbf{J}_{\mathbf{p}}\mathbf{F}_1^0(\mathbf{p}),$$
$$\mathbf{J}_{\mathbf{p}}\mathbf{x}_i(\alpha_i,\mathbf{p}) = \mathbf{J}_{\mathbf{x}}\mathbf{F}_i^0(\mathbf{v}_{i-1}(\beta_{i-1},\mathbf{p}))\mathbf{J}_{\mathbf{p}}\mathbf{x}_{i-1}(\beta_{i-1},\mathbf{p})+$$
$$\mathbf{J}_{\mathbf{p}}\mathbf{F}_i^0(\mathbf{v}_{i-1}(\beta_{i-1},\mathbf{p})),\ \forall i\in\mathcal{I}\backslash\{1\}, \tag{3.4.5}$$
$$\mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p}) = \mathbf{J}_{\mathbf{p}}\mathbf{x}_i(t,\mathbf{p}),\ \forall t\in[\alpha_i,\beta_i),\ \mathbf{J}_{\mathbf{p}}\mathbf{x}(\beta_{n_e},\mathbf{p}) = \mathbf{J}_{\mathbf{p}}\mathbf{x}_{n_e}(\beta_{n_e},\mathbf{p}).$$

where $\boldsymbol{\pi}_{\mathbf{v}}\mathbf{J}_{\mathbf{u}}\mathbf{F}_i(t,\mathbf{u}(t,\mathbf{p}))\in\mathbb{R}^{(n_x+n_y)\times(n_p+n_x)}$, $\boldsymbol{\pi}_{\mathbf{q}}\mathbf{J}_{\mathbf{u}}\mathbf{F}_i(t,\mathbf{u}(t,\mathbf{p}))\in\mathbb{R}^{(n_x+n_y)\times(n_y+n_x)}$, $\mathbf{J}_{\mathbf{p}}\mathbf{f}_i\ (t,\mathbf{v}_i(t,\mathbf{p}))\in\mathbb{R}^{n_x\times n_p}$, $\mathbf{J}_{\mathbf{x}}\mathbf{f}_i(\ t,\mathbf{v}_i(t,\mathbf{p}))\in\mathbb{R}^{n_x\times n_x}$, $\mathbf{J}_{\mathbf{p}}\mathbf{r}_i\ (t,\mathbf{v}_i(t,\mathbf{p}))\in\mathbb{R}^{n_y\times n_p}$ *and* $\mathbf{J}_{\mathbf{x}}\mathbf{r}_i(t,\mathbf{v}_i(t,\mathbf{p}))\in\mathbb{R}^{n_y\times n_x}$, $\forall i\in\mathcal{I}$. $\mathbf{f}_i$ *and* $\mathbf{r}_i$ *are the locally Lipschitz continuous implicit functions that satisfy* $\mathbf{F}_i(t,\mathbf{p},\ \mathbf{x}_i(t,\mathbf{p}),\ \mathbf{r}_i(t,\mathbf{v}_i(t,\mathbf{p})),\ \mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p})))$ *for all* $i\in\mathcal{I}$. *Finally,* $\mathbf{J}_{\mathbf{p}}\mathbf{x}_i(\cdot,\mathbf{p})$ *are absolutely continuous functions on* $[\alpha_i,\beta_i]$

Proof. Let $n_e = 1$. Then the result holds per Theorem 3.3.6. If $n_e = 2$, then the strict derivative of the mapping $\boldsymbol{\eta}\mapsto\mathbf{x}_2(\alpha_2,\boldsymbol{\eta})$ at $\mathbf{p}$ is obtained after applying Theorem 2.6.7 to (3.4.2) and is (3.4.5). Equations (3.4.3), (3.4.4) hold for $i=2$ per Theorem 3.3.6 because $\partial_{\mathbf{u}}\mathbf{F}_i(t,\mathbf{u}_i(t,\mathbf{p}))$ is a singleton for all $t\in[\alpha_i,\beta_i]\backslash S$ and $\boldsymbol{\eta}\mapsto\mathbf{x}_2(\alpha_2,\boldsymbol{\eta})$ is strictly differentiable at $\mathbf{p}$. Hence the result holds for the case $n_e = 2$. The case for $n_e > 2$ can be proven similarly by repeatedly applying Theorem 3.3.6 and noting that the mappings $\boldsymbol{\eta}\mapsto\mathbf{x}_i(\alpha_i,\boldsymbol{\eta})$ are strictly differentiable at $\mathbf{p}$ for all $i\in\mathcal{I}$. □

Remark 3.4.8. $\mathbf{J}_{\mathbf{p}}\mathbf{x}(\cdot,\mathbf{p})$ *might be discontinuous at* $t=\alpha_i$ *with* $i\in\mathcal{I}\backslash\{1\}$.

Corollary 3.4.9. *Let the hypotheses of Theorem 3.4.7 hold. Then* $\mathbf{y}(t,\cdot)$ *and* $\dot{\mathbf{x}}(t,\cdot)$ *are strictly differentiable at* $\mathbf{p}$ *for all* $t\in T\backslash S$ *and the derivatives are*

$$\mathbf{J}_{\mathbf{p}}\mathbf{y}(t,\mathbf{p}) = \mathbf{J}_{\mathbf{x}}\mathbf{r}_i(t,\mathbf{v}_i(t,\mathbf{p}))\mathbf{J}_{\mathbf{p}}\mathbf{x}(t,\mathbf{p}) + \mathbf{J}_{\mathbf{p}}\mathbf{r}_i(t,\mathbf{v}_i(t,\mathbf{p})),\ \forall t\in[\alpha_i,\beta_i)\backslash S,\forall i\in\mathcal{I},$$

$$\mathbf{J_p}\dot{\mathbf{x}}(t,\mathbf{p}) = \mathbf{J_x}\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p}))\mathbf{J_p}\mathbf{x}(t,\mathbf{p}) + \mathbf{J_p}\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p})),\ \forall t \in [\alpha_i,\beta_i)\backslash S, \forall i \in \mathcal{I},$$

where $\mathbf{J_x}\mathbf{r}_i$, $\mathbf{J_p}\mathbf{r}_i$, $\mathbf{J_x}\mathbf{f}_i$, $\mathbf{J_p}\mathbf{f}_i$ *and* $\mathbf{J_p}\mathbf{x}$ *are as defined in Theorem 3.4.7. In addition, if* $\beta_{n_e} \notin S$, *then* $\mathbf{y}(\beta_{n_e},\cdot)$ *and* $\dot{\mathbf{x}}(\beta_{n_e},\cdot)$ *are strictly differentiable at* $\mathbf{p}$ *and the derivatives are*

$$\mathbf{J_p}\mathbf{y}(\beta_{n_e},\mathbf{p}) = \mathbf{J_x}\mathbf{r}_{n_e}(\beta_{n_e},\mathbf{v}_{n_e}(\beta_{n_e},\mathbf{p}))\mathbf{J_p}\mathbf{x}(\beta_{n_e},\mathbf{p}) + \mathbf{J_p}\mathbf{r}_{n_e}(\beta_{n_e},\mathbf{v}_{n_e}(\beta_{n_e},\mathbf{p})),$$
$$\mathbf{J_p}\dot{\mathbf{x}}(\beta_{n_e},\mathbf{p}) = \mathbf{J_x}\mathbf{f}_{n_e}(\beta_{n_e},\mathbf{v}_{n_e}(\beta_{n_e},\mathbf{p}))\mathbf{J_p}\mathbf{x}(\beta_{n_e},\mathbf{p}) + \mathbf{J_p}\mathbf{f}_{n_e}(\beta_{n_e},\mathbf{v}_{n_e}(\beta_{n_e},\mathbf{p}))$$

Proof. The result follows from Corollary 3.3.7. □

Remark 3.4.10. *Theorem 3.4.7 can be extended to the case where* $\mathbf{F}_i^0$ *are functions of* $\mathbf{u}_{i-1}$ *for* $i \in \mathcal{I}\backslash\{1\}$ *with slight modifications. In order to guarantee the strict differentiability of* $\mathbf{x}_i(\alpha_i,\cdot)$ *at* $\mathbf{p}$, $\mathbf{y}_{i-1}(\beta_i,\cdot)$ *and* $\dot{\mathbf{x}}_{i-1}(\beta_i,\cdot)$ *need to be strictly differentiable at* $\mathbf{p}$. *Hence,* $\beta_i \notin S$ *for all* $i \in \mathcal{I}\backslash\{1\}$ *needs to hold.*

The extensions of Theorems 3.3.9 and 3.3.10 follow next. The extensions require the introduction of additional variables in order to relate the adjoint equations for each separate time interval.

Theorem 3.4.11. *Let the hypotheses of Theorem 3.4.7 hold. Define* $G : \mathcal{P} \to \mathbb{R}$ *as*

$$G(\mathbf{p}) = \sum_{i=1}^{n_e} \int_{\alpha_i}^{\beta_i} g_i(t,\mathbf{u}_i(t,\mathbf{p}))dt$$

where $g_i : \mathcal{T}_i \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}} \to \mathbb{R}$ *are locally Lipschitz continuous functions for all* $i \in \mathcal{I}$. *Let* Q *be a measure zero subset of* T. *Let* $\partial_{\mathbf{u}} g_i(t,\mathbf{u}_i(t,\mathbf{p}))$ *be a singleton for all* $t \in [\alpha_i,\beta_i]\backslash Q$ *for all* $i \in \mathcal{I}$.

Define for each $i \in \mathcal{I}$:

$$\mathbf{B}_{\mathbf{x},i}(t,\mathbf{u}_i(t,\mathbf{p})) = \hat{\boldsymbol{\nabla}}_{\mathbf{x}} g_i(t,\mathbf{u}_i(t,\mathbf{p})) + \hat{\mathbf{J}}_{\mathbf{x}}\mathbf{r}_i(t,\mathbf{v}_i(t,\mathbf{p}))^{\mathrm{T}}\hat{\boldsymbol{\nabla}}_{\mathbf{y}} g_i(t,\mathbf{u}_i(t,\mathbf{p})) +$$

$$\hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p}))^{\mathrm{T}}\hat{\boldsymbol{\nabla}}_{\dot{\mathbf{x}}}g_i(t,\mathbf{u}_i(t,\mathbf{p})),\ \forall t\in[\alpha_i,\beta_i],$$

$$\mathbf{B}_{\mathbf{p},i}(t,\mathbf{u}_i(t,\mathbf{p})) = \hat{\boldsymbol{\nabla}}_{\mathbf{p}}g_i(t,\mathbf{u}_i(t,\mathbf{p})) + \hat{\mathbf{J}}_{\mathbf{p}}\mathbf{r}_i(t,\mathbf{v}_i(t,\mathbf{p}))^{\mathrm{T}}\hat{\boldsymbol{\nabla}}_{\mathbf{y}}g_i(t,\mathbf{u}_i(t,\mathbf{p})) +$$

$$\hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p}))^{\mathrm{T}}\hat{\boldsymbol{\nabla}}_{\dot{\mathbf{x}}}g_i(t,\mathbf{u}_i(t,\mathbf{p})),\ \forall t\in[\alpha_i,\beta_i].$$

Let $\boldsymbol{\lambda}_i : [\alpha_i,\beta_i]\to\mathbb{R}^{n_x}$ *be solutions of*

$$\dot{\boldsymbol{\lambda}}_i(t) = -\hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}_i(t) + \mathbf{B}_{\mathbf{x},i},\ \forall t\in[\alpha_i,\beta_i),\ \forall i\in\mathcal{I}, \tag{3.4.6}$$

$$\boldsymbol{\lambda}_i(\beta_i) = \mathbf{J}_{\mathbf{x}}\mathbf{F}^0_{i+1}(\mathbf{v}_i(t,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}_{i+1}(\beta_i),\ \forall i\in\mathcal{I}\backslash\{n_e\},\ \boldsymbol{\lambda}_{n_e}(\beta_{n_e}) = \mathbf{0}.$$

where $\mathbf{J}_{\mathbf{p}}\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p}))$, $\mathbf{J}_{\mathbf{x}}\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p}))$, $\mathbf{J}_{\mathbf{p}}\mathbf{r}_i(t,\mathbf{v}_i(t,\mathbf{p}))$, *and* $\mathbf{J}_{\mathbf{x}}\mathbf{r}_i(t,\mathbf{v}_i(t,\mathbf{p}))$ *are computed using* (3.4.3). *Then,* $\boldsymbol{\lambda}_i$ *are unique and absolutely continuous.*

In addition, G *is locally Lipschitz continuous and strictly differentiable at* $\mathbf{p}$ *and the derivative is*

$$\begin{aligned}\boldsymbol{\nabla}G(\mathbf{p}) = &\sum_{i=1}^{n_e}\int_{\alpha_i}^{\beta_i}\mathbf{B}_{\mathbf{p},i}(t,\mathbf{u}_i(t,\mathbf{p})) - \hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}_i(t)dt - \\ &\sum_{i=1}^{n_e-1}\boldsymbol{\lambda}_{i+1}(\beta_i)^{\mathrm{T}}\mathbf{J}_{\mathbf{p}}\mathbf{F}^0_{i+1}(\mathbf{v}_i(\beta_i,\mathbf{p})) + \\ &\mathbf{J}_{\mathbf{p}}\mathbf{x}_{n_e}(\beta_{n_e},\mathbf{p})^{\mathrm{T}}\boldsymbol{\lambda}_{n_e}(\beta_{n_e}) - \mathbf{J}_{\mathbf{p}}\mathbf{x}_1(\alpha_1,\mathbf{p})^{\mathrm{T}}\boldsymbol{\lambda}_1(\alpha_1).\end{aligned} \tag{3.4.7}$$

Proof. Consider the equivalent definition of G,

$$\begin{aligned}G(\mathbf{p}) = &\sum_{i=1}^{n_e}\int_{\alpha_i}^{\beta_i} g_i(t,\mathbf{u}_i(t,\mathbf{p})) - \boldsymbol{\lambda}_i(t)^{\mathrm{T}}(\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p})) - \dot{\mathbf{x}}_i(t,\mathbf{p}))dt + \\ &\sum_{i=1}^{n_e-1}\boldsymbol{\mu}_i^{\mathrm{T}}\left(\mathbf{x}_{i+1}(\beta_i,\mathbf{p}) - \mathbf{F}^0_{i+1}(\mathbf{v}_i(\beta_i,\mathbf{p}))\right)\end{aligned}$$

where $\boldsymbol{\mu}_i \in \mathbb{R}^{n_x}$ for all $i \in \mathcal{I}\backslash\{n_e\}$. Using Theorem 3.3.9, the relation

$$\begin{aligned}\boldsymbol{\nabla} G(\mathbf{p}) = &\sum_{i=1}^{n_e} \int_{\alpha_i}^{\beta_i} \mathbf{B}_{\mathbf{p}}(t, \mathbf{u}_i(t,\mathbf{p})) - \hat{\mathbf{J}}_{\mathbf{p}} \mathbf{f}_i(t, \mathbf{v}_i(t,\mathbf{p}))^{\mathrm{T}} \boldsymbol{\lambda}_i(t) dt + \\ &\sum_{i=1}^{n_e} \mathbf{J}_{\mathbf{p}} \mathbf{x}_i(t,\mathbf{p})^{\mathrm{T}} \boldsymbol{\lambda}_i(t) \Big|_{\alpha_i}^{\beta_i} + \\ &\sum_{i=1}^{n_e-1} \left(\mathbf{J}_{\mathbf{p}} \mathbf{x}_{i+1}(\beta_i,\mathbf{p}) - \mathbf{J}_{\mathbf{p}} \mathbf{F}_{i+1}^0(\mathbf{v}_i(\beta_i,\mathbf{p}))\right)^{\mathrm{T}} \boldsymbol{\mu}_i - \\ &\sum_{i=1}^{n_e-1} \left(\mathbf{J}_{\mathbf{x}} \mathbf{F}_{i+1}^0(\mathbf{v}_i(\beta_i,\mathbf{p})) \mathbf{J}_{\mathbf{p}} \mathbf{x}_i(\beta_i,\mathbf{p})\right)^{\mathrm{T}} \boldsymbol{\mu}_i\end{aligned}$$

is obtained. The results (3.4.7) and (3.4.6) follow after relating $\boldsymbol{\lambda}_i$ to $\boldsymbol{\lambda}_{i+1}$ by setting $\boldsymbol{\lambda}_i(\beta_i) = \mathbf{J}_{\mathbf{x}} \mathbf{F}_{i+1}^0(\mathbf{v}_i(\beta_i,\mathbf{p}))^{\mathrm{T}} \boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_i = \boldsymbol{\lambda}_{i+1}(\beta_i)$ for $i \in \mathcal{I}\backslash\{n_e\}$ and $\boldsymbol{\lambda}_{n_e}(\beta_{n_e}) = \mathbf{0}$. □

Theorem 3.4.12. *Let the hypotheses of Theorem 3.4.7 hold. Let $h : \mathcal{T}_0 \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}} \to \mathbb{R}$ be a locally Lipschitz continuous function where $\mathcal{T}_0$ is an open subset of $\mathcal{T}$ such that $t_f \in \mathcal{T}_0$. Assume $t_f \notin S$ (if $h : \mathcal{T}_0 \times \mathcal{P} \times \mathcal{X} \to \mathbb{R}$, this assumption is not necessary).*

Let $G : \mathcal{P} \to \mathbb{R} : \boldsymbol{\eta} \mapsto h(t_f, \mathbf{u}(t_f, \boldsymbol{\eta}))$ and assume $\partial_{\mathbf{u}} h(t_f, \mathbf{u}(t_f,\mathbf{p}))$ is a singleton with element $(\boldsymbol{\nabla}_{\mathbf{p}} h(t_f, \mathbf{u}(t_f,\mathbf{p}), \boldsymbol{\nabla}_{\mathbf{x}} h(t_f, \mathbf{u}(t_f,\mathbf{p}), \boldsymbol{\nabla}_{\mathbf{y}} h(t_f, \mathbf{u}(t_f,\mathbf{p}), \boldsymbol{\nabla}_{\dot{\mathbf{x}}} h(t_f, \mathbf{u}(t_f,\mathbf{p}))$ where $\boldsymbol{\nabla}_{\mathbf{p}} h(t_f, \mathbf{u}(t_f,\mathbf{p}) \in \mathbb{R}^{n_p}$, $\boldsymbol{\nabla}_{\mathbf{x}} h(t_f, \mathbf{u}(t_f,\mathbf{p})) \in \mathbb{R}^{n_x}$, $\boldsymbol{\nabla}_{\mathbf{y}} h(t_f, \mathbf{u}(t_f,\mathbf{p})) \in \mathbb{R}^{n_y}$ and $\boldsymbol{\nabla}_{\dot{\mathbf{x}}} h(t_f, \mathbf{u}(t_f,\mathbf{p})) \in \mathbb{R}^{n_x}$.

Let

$$\begin{aligned}\mathbf{C}_{\mathbf{x}} = &\mathbf{J}_{\mathbf{x}} \mathbf{f}_{n_e}(t_f, \mathbf{v}_{n_e}(t_f,\mathbf{p}))^{\mathrm{T}} \boldsymbol{\nabla}_{\dot{\mathbf{x}}} h(t_f, \mathbf{u}(t_f,\mathbf{p})) + \\ &\mathbf{J}_{\mathbf{x}} \mathbf{r}_{n_e}(t_f, \mathbf{v}_{n_e}(t_f,\mathbf{p}))^{\mathrm{T}} \boldsymbol{\nabla}_{\mathbf{y}} h(t_f, \mathbf{u}(t_f,\mathbf{p})) + \boldsymbol{\nabla}_{\mathbf{x}} h(t_f, \mathbf{u}(t_f,\mathbf{p})), \\ \mathbf{C}_{\mathbf{p}} = &\mathbf{J}_{\mathbf{p}} \mathbf{f}_{n_e}(t_f, \mathbf{v}_{n_e}(t_f,\mathbf{p}))^{\mathrm{T}} \boldsymbol{\nabla}_{\dot{\mathbf{x}}} h(t_f, \mathbf{u}(t_f,\mathbf{p})) + \\ &\mathbf{J}_{\mathbf{p}} \mathbf{r}_{n_e}(t_f, \mathbf{v}_{n_e}(t_f,\mathbf{p}))^{\mathrm{T}} \boldsymbol{\nabla}_{\mathbf{y}} h(t_f, \mathbf{u}(t_f,\mathbf{p})) + \boldsymbol{\nabla}_{\mathbf{p}} h(t_f, \mathbf{u}(t_f,\mathbf{p})).\end{aligned}$$

Let $\boldsymbol{\lambda}_i : [\alpha_i, \beta_i] \to \mathbb{R}^{n_x}$ *be solutions of the initial value problems:*

$$\begin{aligned} \dot{\boldsymbol{\lambda}}_i(t) = & -\hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}_i(t, \mathbf{v}_i(t,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}_i(t), \forall t \in (\alpha_i, \beta_i], \\ \boldsymbol{\lambda}_i(\beta_i) = & \mathbf{J}_{\mathbf{x}}\mathbf{F}^0_{i+1}(\mathbf{v}_i(t,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}_{i+1}(\alpha_{i+1}),\ \forall i \in \mathcal{I}\backslash\{n_e\},\ \boldsymbol{\lambda}_{n_e}(\beta_{n_e}) = -\mathbf{C}_{\mathbf{x}}. \end{aligned} \tag{3.4.8}$$

where $\mathbf{J}_{\mathbf{p}}\mathbf{f}_i(t, \mathbf{v}_i(t,\mathbf{p}))$, $\mathbf{J}_{\mathbf{x}}\mathbf{f}_i(t, \mathbf{v}_i(t,\mathbf{p}))$, $\mathbf{J}_{\mathbf{p}}\mathbf{r}_i(t, \mathbf{v}_i(t,\mathbf{p}))$, *and* $\mathbf{J}_{\mathbf{x}}\mathbf{r}_i(t, \mathbf{v}_i(t,\mathbf{p}))$ *are computed using* (3.4.3). *Then* $\boldsymbol{\lambda}_i$ *are unique and absolutely continuous.*

In addition, G *is locally Lipschitz continuous and strictly differentiable at* $\mathbf{p}$ *and the derivative is*

$$\begin{aligned} \boldsymbol{\nabla} G(\mathbf{p}) = & \sum_{i=1}^{n_e} \int_{\alpha_i}^{\beta_i} -\hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}_i(t, \mathbf{v}_i(t,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}_i(t)dt - \\ & \sum_{i=1}^{n_e-1} \mathbf{J}_{\mathbf{p}}\mathbf{F}^0_{i+1}(\mathbf{v}_i(\beta_i,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}_{i+1}(\beta_i) - \mathbf{J}_{\mathbf{p}}\mathbf{x}_1(\alpha_1,\mathbf{p})^{\mathrm{T}}\boldsymbol{\lambda}_1(\alpha_1) + \mathbf{C}_{\mathbf{p}}. \end{aligned}$$

Proof. As in Theorem 3.3.10, G is locally Lipschitz continuous at $\mathbf{p}$ because it is the composition of locally Lipschitz continuous functions $h(t_f, \cdot)$ and the locally Lipschitz continuous function $\mathbf{u}(t_f, \cdot)$. Strict differentiability follows from the fact that $\partial_{\mathbf{u}} h(t_f, \mathbf{u}(t_f, \mathbf{p}))$ is a singleton and $\mathbf{u}(t_f, \cdot)$ is strictly differentiable at $\mathbf{p}$. The strict derivative is

$$\begin{aligned} \boldsymbol{\nabla} G(\mathbf{p}) = & \dot{\mathbf{J}}_{\mathbf{p}}\mathbf{x}(t_f,\mathbf{p})^{\mathrm{T}}\boldsymbol{\nabla}_{\dot{\mathbf{x}}} h(t_f, \mathbf{u}(t_f,\mathbf{p})) + \mathbf{J}_{\mathbf{p}}\mathbf{y}(t_f,\mathbf{p})^{\mathrm{T}}\boldsymbol{\nabla}_{\mathbf{y}} h(t_f, \mathbf{u}(t_f,\mathbf{p})) + \\ & \mathbf{J}_{\mathbf{p}}\mathbf{x}(t_f,\mathbf{p})^{\mathrm{T}}\boldsymbol{\nabla}_{\mathbf{x}} h(t_f, \mathbf{u}(t_f,\mathbf{p})) + \boldsymbol{\nabla}_{\mathbf{p}} h(t_f, \mathbf{u}(t_f,\mathbf{p})). \end{aligned}$$

Replacing $\dot{\mathbf{J}}_{\mathbf{p}}\mathbf{x}(t_f,\mathbf{p})$ and $\mathbf{J}_{\mathbf{p}}\mathbf{y}(t_f,\mathbf{p})$ with the results in Corollary 3.4.9 produces

$$\boldsymbol{\nabla} G(\mathbf{p}) = \mathbf{J}_{\mathbf{p}}\mathbf{x}(t_f,\mathbf{p})^{\mathrm{T}}\mathbf{C}_{\mathbf{x}} + \mathbf{C}_{\mathbf{p}}.$$

Observe that $\mathbf{J}_{\mathbf{p}}\mathbf{x}(t_f,\mathbf{p})^{\mathrm{T}}\mathbf{C}_{\mathbf{x}}$ is the derivative of $\mathbf{x}(t_f,\cdot)^{\mathrm{T}}\mathbf{C}_{\mathbf{x}}$ treating $\mathbf{C}_{\mathbf{x}}$ as a constant at $\mathbf{p}$

and

$$\mathbf{x}(t_f,\mathbf{p})^{\mathrm{T}}\mathbf{C}_{\mathbf{x}} = \sum_{i=1}^{n_e}\int_{\alpha_i}^{\beta_i}\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p}))^{\mathrm{T}}\mathbf{C}_{\mathbf{x}}dt + \sum_{i=1}^{n_e-1}\left(\mathbf{x}_{i+1}(\alpha_{i+1},\mathbf{p}) - \mathbf{x}_i(\beta_i,\mathbf{p})\right)^{\mathrm{T}}\mathbf{C}_{\mathbf{x}} + \mathbf{x}_1(\alpha_1,\mathbf{p})^{\mathrm{T}}\mathbf{C}_{\mathbf{x}},$$

$$\mathbf{x}(t_f,\mathbf{p})^{\mathrm{T}}\mathbf{C}_{\mathbf{x}} = \sum_{i=1}^{n_e}\int_{\alpha_i}^{\beta_i}\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p}))^{\mathrm{T}}\mathbf{C}_{\mathbf{x}} - \boldsymbol{\lambda}_i(t)^{\mathrm{T}}(\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p})) - \dot{\mathbf{x}}_i(t,\mathbf{p}))dt + \sum_{i=1}^{n_e-1}\left(\mathbf{x}_{i+1}(\alpha_{i+1},\mathbf{p}) - \mathbf{x}_i(\beta_i,\mathbf{p})\right)^{\mathrm{T}}\mathbf{C}_{\mathbf{x}} + \mathbf{x}_1(\alpha_1,\mathbf{p})^{\mathrm{T}}\mathbf{C}_{\mathbf{x}}.$$

hold. Let $H : \mathcal{P} \to \mathbb{R} : \boldsymbol{\eta} \mapsto \mathbf{x}(t_f,\boldsymbol{\eta})^{\mathrm{T}}\mathbf{C}_{\mathbf{x}}$. Then

$$\nabla H(\mathbf{p}) = \sum_{i=1}^{n_e}\int_{\alpha_i}^{\beta_i}\dot{\mathbf{J}}_{\mathbf{p}}\mathbf{x}_i(t,\mathbf{p})^{\mathrm{T}}\mathbf{C}_{\mathbf{x}} - (\hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p}))\mathbf{J}_{\mathbf{p}}\mathbf{x}_i(t,\mathbf{p}) + \hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p})) - \dot{\mathbf{J}}_{\mathbf{p}}\mathbf{x}_i(t,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}_i(t)dt + \sum_{i=1}^{n_e-1}\left(\mathbf{J}_{\mathbf{p}}\mathbf{x}_{i+1}(\alpha_{i+1},\mathbf{p}) - \mathbf{J}_{\mathbf{p}}\mathbf{x}_i(\beta_i,\mathbf{p})\right)^{\mathrm{T}}\mathbf{C}_{\mathbf{x}} + \mathbf{J}_{\mathbf{p}}\mathbf{x}_1(\alpha_1,\mathbf{p})^{\mathrm{T}}\mathbf{C}_{\mathbf{x}}.$$

Collecting terms containing $\dot{\mathbf{J}}_{\mathbf{p}}\mathbf{x}_i(t,\mathbf{p})$ and using integration by parts as in Theorem 3.3.10 results in

$$\nabla H(\mathbf{p}) = \sum_{i=1}^{n_e}\int_{\alpha_i}^{\beta_i}\mathbf{J}_{\mathbf{p}}\mathbf{x}_i(t,\mathbf{p})^{\mathrm{T}}(-\hat{\mathbf{J}}_{\mathbf{x}}\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}_i(t) - \dot{\boldsymbol{\lambda}}_i(t)) - \hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}_i(t)dt + \mathbf{J}_{\mathbf{p}}\mathbf{x}_i(t,\mathbf{p})^{\mathrm{T}}(\mathbf{C}_{\mathbf{x}} + \boldsymbol{\lambda}_i(t))\Big|_{\alpha_i}^{\beta_i} + \sum_{i=1}^{n_e-1}\left(\mathbf{J}_{\mathbf{p}}\mathbf{x}_{i+1}(\alpha_{i+1},\mathbf{p}) - \mathbf{J}_{\mathbf{p}}\mathbf{x}_i(\beta_i,\mathbf{p})\right)^{\mathrm{T}}\mathbf{C}_{\mathbf{x}} + \mathbf{J}_{\mathbf{p}}\mathbf{x}_1(\alpha_1,\mathbf{p})^{\mathrm{T}}\mathbf{C}_{\mathbf{x}},$$

$$\nabla H(\mathbf{p}) = \sum_{i=1}^{n_e}\int_{\alpha_i}^{\beta_i} -\hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p}))^{\mathrm{T}}\boldsymbol{\lambda}_i(t)dt + \mathbf{J}_{\mathbf{p}}\mathbf{x}_i(t,\mathbf{p})^{\mathrm{T}}(\mathbf{C}_{\mathbf{x}} + \boldsymbol{\lambda}_i(t))\Big|_{\alpha_i}^{\beta_i} + \sum_{i=1}^{n_e-1}\left(\mathbf{J}_{\mathbf{p}}\mathbf{x}_{i+1}(\alpha_{i+1},\mathbf{p}) - \mathbf{J}_{\mathbf{p}}\mathbf{x}_i(\beta_i,\mathbf{p})\right)^{\mathrm{T}}\mathbf{C}_{\mathbf{x}} + \mathbf{J}_{\mathbf{p}}\mathbf{x}_1(\alpha_1,\mathbf{p})^{\mathrm{T}}\mathbf{C}_{\mathbf{x}},$$

$$\nabla H(\mathbf{p}) = \sum_{i=1}^{n_e} \int_{\alpha_i}^{\beta_i} -\hat{\mathbf{J}}_{\mathbf{p}}\mathbf{f}_i(t, \mathbf{v}_i(t,\mathbf{p}))^{\mathrm{T}} \boldsymbol{\lambda}_i(t) dt + \sum_{i=1}^{n_e-1} \left(-\mathbf{J}_{\mathbf{p}}\mathbf{x}_{i+1}(\alpha_{i+1},\mathbf{p})^{\mathrm{T}} \boldsymbol{\lambda}_{i+1}(\alpha_{i+1}) + \mathbf{J}_{\mathbf{p}}\mathbf{x}_i(\beta_i,\mathbf{p})^{\mathrm{T}} \boldsymbol{\lambda}_i(\beta_i)\right) + \mathbf{J}_{\mathbf{p}}\mathbf{x}_{n_e}(\beta_{n_e},\mathbf{p})^{\mathrm{T}} \left(\mathbf{C}_{\mathbf{x}} + \boldsymbol{\lambda}_{n_e}(\beta_{n_e})\right) - \mathbf{J}_{\mathbf{p}}\mathbf{x}_1(\alpha_1,\mathbf{p})^{\mathrm{T}} \boldsymbol{\lambda}_1(\alpha_1).$$

Setting $\boldsymbol{\lambda}_i(\beta_i) = \mathbf{J}_{\mathbf{x}}\mathbf{F}_{i+1}^0(\mathbf{v}_i(\beta_i,\mathbf{p}))^{\mathrm{T}} \boldsymbol{\lambda}_{i+1}(\alpha_{i+1})$ for all $i \in \mathcal{I}\backslash\{n_e\}$, $\boldsymbol{\lambda}_{n_e}(\beta_{n_e}) = -\mathbf{C}_{\mathbf{x}}$ and using (3.4.5) provides the desired result. □

3.5 Comments on the Numerical Computation of Forward and Adjoint Sensitivities

In this section, the discussion focuses on computational issues for the adjoint and forward sensitivities of solutions to the initial value problems (3.2.1) and (3.3.1).

In order to solve (3.3.1) where $\mathbf{F}$ is an arbitrary locally Lipschitz continuous function requires a method to solve nonsmooth equations (see [35] for examples of nonsmooth equation solvers) coupled with a numerical integration algorithm. Although, this is an interesting avenue of research, it is not pursued here because many systems of interest have special structures that allow the use of existing algorithms. Usually the right-hand side of (3.3.1) is continuously differentiable on open subsets whose closures partition the domain and the solution can be obtained using integration algorithms coupled with state event location algorithms [83].

The computation of the forward and adjoint sensitivities require the set $S = \{t \in T : \partial_{\mathbf{u}}\mathbf{F}(t, \mathbf{u}(t,\mathbf{p}))$ is not a singleton$\}$ to be determined. In general, when solutions of (3.3.1) are obtained numerically, it is not possible to determine whether S is a set of measure zero because the numerical solution comprises values computed at finitely many elements of T

which may or may not be elements of S. Another major issue is that the computation of the generalized Jacobian of an arbitrary locally Lipschitz continuous function using definition (2.6.5) is not computationally practical.

The initial value problem (3.2.1) has a continuous right-hand side; however, the corresponding adjoint and forward sensitivity initial value problems are equations with discontinuous right-hand sides (See Example 3.6.2). These discontinuities need to be detected and located using, for example, state event location [83], for the efficient and accurate computation of the adjoint and forward sensitivity trajectories. Similar observations apply to the adjoint and forward sensitivities of (3.3.1).

Finally, the adjoint and sensitivity initial value problems (3.3.4, 3.3.6, 3.3.8) and the integrals (3.3.7, 3.3.9) require the computation of the derivatives of the implicit functions $\mathbf{f}$ and $\mathbf{r}$. This computation is achieved by solving (3.3.2) at each function evaluation, which is computationally very costly. Ideally, auxiliary DAE systems analogous to those in [24] should be solved.

In order to overcome these computational issues, additional assumptions on the structure of $\mathbf{F}$ need to be imposed. For example, $\mathbf{F}$ can be continuously differentiable on open sets whose closures partition the domain of $\mathbf{F}$. The boundaries of these open sets can be the zero-level sets of certain functions. These assumptions and their numerical implications are discussed in the following chapters.

3.6 Examples

The first example is a case where the mapping $\boldsymbol{\eta} \mapsto \mathbf{x}(t_f, \boldsymbol{\eta})$ is strictly but not continuously differentiable at $\mathbf{p}$.

Example 3.6.1. Let $n_x = 1$, $n_p = 2$. Let $T = [t_0, t_f]$, $\mathcal{P} = \mathbb{R}^2$ and $\mathcal{X} = \mathbb{R}$ and $\Delta t = t_f - t_0$,

Table 3.1: Solution and Generalized Gradients of Example 3.6.1

Case	$x(t, \mathbf{p})$	$\partial_{\mathbf{p}} x(t, \mathbf{p})$
$p_1 > 0,\ p_2 > 0$	$p_1 \cdot e^{p_2 \Delta t}$	$\left\{\left(e^{p_2 \Delta t},\ p_1 \cdot \Delta t \cdot e^{p_2 \Delta t}\right)\right\}$
$p_1 > 0,\ p_2 < 0$	$p_1 \cdot e^{-p_2 \Delta t}$	$\left\{\left(e^{-p_2 \Delta t},\ -p_1 \cdot \Delta t \cdot e^{-p_2 \Delta t}\right)\right\}$
$p_1 < 0,\ p_2 > 0$	$p_1 \cdot e^{-p_2 \Delta t}$	$\left\{\left(e^{-p_2 \Delta t},\ -p_1 \cdot \Delta t \cdot e^{-p_2 \Delta t}\right)\right\}$
$p_1 < 0,\ p_2 < 0$	$p_1 \cdot e^{p_2 \Delta t}$	$\left\{\left(e^{p_2 \Delta t},\ p_1 \cdot \Delta t \cdot e^{p_2 \Delta t}\right)\right\}$
$p_1 = 0$	0	$\operatorname{conv}\left(\left\{\left(e^{p_2 \Delta t},\ 0\right), \left(e^{-p_2 \Delta t},\ 0\right)\right\}\right)$
$p_2 = 0$	p_1	$\operatorname{conv}\left(\left\{(1,\ p_1 \cdot \Delta t), (1,\ -p_1 \cdot \Delta t)\right\}\right)$
$p_1 = 0,\ p_2 = 0$	0	$\{(1, 0)\}$

$x : T \times \mathcal{P} \to \mathcal{X}$, $f : T \times \mathcal{P} \times \mathcal{X} : (t, \boldsymbol{\eta}, \mu) \mapsto |\eta_2 \cdot \mu|$. Consider the initial value problem

$$\dot{x}(t, \mathbf{p}) = f(t, \mathbf{p}, x(t, \mathbf{p})),\ \forall t \in (t_0, t_f],\ x(t_0, \mathbf{p}) = p_1.$$

Table 3.1 contains the solutions and generalized gradients as a function of the parameter values. Note that at $\mathbf{p} = (0, 0)$, $x(t, \cdot)$ is a strictly differentiable function. Gronwall's result [44] cannot be applied to conclude differentiability because the partial derivatives of f are not continuous in any open set containing $\{(t, x(t, \mathbf{p}), \mathbf{p}) : t \in [t_0, t_f]\}$=$\{(t, 0, 0, 0) : t \in [t_0, t_f]\}$. The results in [94], [95] and [39] are also not applicable in this case. In the neighborhood of $\mathbf{p} = (0, 0)$, the state evolves according to $\dot{x}(0, \mathbf{p}) = -p_2 x(0, \mathbf{p})$ or $\dot{x}(0, \mathbf{p}) = p_2 x(0, \mathbf{p})$ depending on the parameters. Hence the sequence of vector fields encountered is not invariant. Theorem 3.2.3 can be applied to deduce strict differentiability in this case.

The next example demonstrates the discontinuous nature of the sensitivity equations.

Example 3.6.2. Let $T = [0, t_f]$, $n_p = 2$, $n_x = 1$, $\mathcal{P} = \mathbb{R}^2$, $\mathcal{X} = \mathbb{R}$. Consider the dynamic

system

$$\dot{x}(t,\mathbf{p}) = \max(p_1 - x(t,\mathbf{p}),0) - \max(x(t,\mathbf{p}) - p_2, 0), \forall t \in (0,t_f],$$
$$x(0,\mathbf{p}) = 0,\ p_1 > p_2 > 0,\ \mathbf{p} \in \mathcal{P}.$$

where $x : T \times \mathcal{P} \to \mathcal{X}$. Let t^* be such that $x(t^*,\mathbf{p}) - p_2 = 0$. If $t^* \geq t_f$ then $x(t,\mathbf{p}) = p_1 \cdot (1 - e^{-t})$ for all $t \in [0,t_f]$. Let $\mathbf{p}$ be such that $0 < \ln \frac{p_1}{p_1 - p_2} \leq t_f$. Then $t^* = \ln \frac{p_1}{p_1 - p_2}$ and $x(t^*,\mathbf{p}) = p_2 = p_1 \cdot (1 - e^{-t^*})$.

In this case, $\hat{J}_{\mathbf{x}} f$ and $\hat{\mathbf{J}}_{\mathbf{p}} f$ in (3.2.2) are

$$\hat{J}_{\mathbf{x}} f(t,\mathbf{v}(t,\mathbf{p})) = \begin{cases} -1 & \text{if } t \in (t_0,t^*), \\ 0 & \text{if } t = t^*, \\ -2 & \text{if } t \in (t^*,t_f], \end{cases}, \quad \hat{\mathbf{J}}_{\mathbf{p}} f(t,\mathbf{v}(t,\mathbf{p})) = \begin{cases} (1,0) & \text{if } t \in (t_0,t^*), \\ (0,0) & \text{if } t = t^*, \\ (1,1) & \text{if } t \in (t^*,t_f]. \end{cases}$$

Hence the sensitivity equations have a discontinuity at t^*. The time of discontinuity depends on the parameter. Note that, in this case, S is a singleton set.

The next example involves piecewise continuously differentiable functions. These are locally Lipschitz continuous functions that are almost everywhere continuously differentiable and that have their own specific implicit function theorem [91, 98].

Example 3.6.3. Let $\mathcal{Y} = \{\mathbf{y} \in \mathbb{R}^{n_y} : y_i \geq 0, \forall i \in \{1,\ldots,n_y\}\}$ and $\mathcal{W} = \mathcal{Y}$. Let $\mathbf{x} : T \times \mathcal{P} \to \mathcal{X}$, $\dot{\mathbf{x}} : T \times \mathcal{P} \to \dot{\mathcal{X}}$, $\mathbf{y} : T \times \mathcal{P} \to \mathcal{Y}$, $\mathbf{w} : T \times \mathcal{P} \to \mathcal{W}$, $\mathbf{p} \in \mathcal{P}$ and $t \in \mathcal{T}$. Let $\mathbf{V} : \mathcal{T} \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \mathcal{W} \times \dot{\mathcal{X}} \to \mathbb{R}^{n_x}$, $\mathbf{Q} : \mathcal{T} \times \mathcal{P} \times \mathcal{X} \to \mathbb{R}^{n_y}$ and $\mathbf{V}_0 : \mathcal{P} \to \mathcal{X}$ be continuously differentiable functions. Let $\mathbf{M} \in \mathbb{R}^{n_y \times n_y}$ be a P-matrix, i.e., the determinant of every principal minor is positive. Let $\boldsymbol{\eta}_{\mathbf{p}} \in \mathcal{P}$, $\boldsymbol{\eta}_{\mathbf{x}} \in \mathcal{X}$, $\boldsymbol{\eta}_{\mathbf{y}} \in \mathcal{Y}$, $\boldsymbol{\eta}_{\dot{\mathbf{x}}} \in \dot{\mathcal{X}}$, $\boldsymbol{\eta}_{\mathbf{w}} \in \mathcal{W}$ and $\boldsymbol{\eta} = (\boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}}, \boldsymbol{\eta}_{\mathbf{y}}, \boldsymbol{\eta}_{\mathbf{w}}, \boldsymbol{\eta}_{\dot{\mathbf{x}}})$. Let $\mathbf{J}_{\dot{\mathbf{x}}} \mathbf{V}(t,\boldsymbol{\eta})$ be invertible for all $(t,\boldsymbol{\eta}) \in \mathcal{T} \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \mathcal{W} \times \dot{\mathcal{X}}$.

Consider the initial value problem

$$\mathbf{0} = \mathbf{V}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p}), \mathbf{w}(t, \mathbf{p}), \dot{\mathbf{x}}(t, \mathbf{p})), \ \forall t \in [t_0, t_f], \tag{3.6.1}$$

$$\mathbf{0} = \mathbf{w}(t, \mathbf{p}) - \mathbf{M}\mathbf{y}(t, \mathbf{p})) - \mathbf{Q}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})), \ \forall t \in [t_0, t_f], \tag{3.6.2}$$

$$0 \leq w_i(t, \mathbf{p}), \ 0 \leq y_i(t, \mathbf{p}), w_i(t, \mathbf{p}) y_i(t, \mathbf{p}) = 0, \forall i \in \mathcal{I}, \ \forall t \in [t_0, t_f], \tag{3.6.3}$$

$$\mathbf{0} = \mathbf{x}(t_0, \mathbf{p}) - \mathbf{V}_0(\mathbf{p}), \ \forall \mathbf{p} \in \mathcal{P}, \tag{3.6.4}$$

$$\mathcal{I} = \{1, \ldots, n_y\}.$$

Let $\boldsymbol{\mu}_\mathbf{y} \in \mathbb{R}^{n_y}$. The *linear complementarity problem* [27],

$$\boldsymbol{\eta}_\mathbf{w} = \mathbf{M}\boldsymbol{\eta}_\mathbf{y} + \boldsymbol{\mu}_\mathbf{y}, \eta_{\mathbf{w},i} \geq 0, \ \eta_{\mathbf{y},i} \geq 0, \ \eta_{\mathbf{w},i}\eta_{\mathbf{y},i} = 0, \ \forall i \in \mathcal{I} \tag{3.6.5}$$

has exactly one solution for each $\boldsymbol{\mu}_\mathbf{y} \in \mathbb{R}^{n_y}$ because $\mathbf{M}$ is a P-matrix. Define the functions $\mathbf{W} : \mathbb{R}^{n_y} \times \mathcal{Y} \to \mathbb{R}^{n_y}$ and $g_i : \mathcal{W} \times \mathcal{Y} \to \mathbb{R}$ as

$$W_i(\boldsymbol{\mu}_\mathbf{y}, \boldsymbol{\eta}_\mathbf{y}) = \min(\langle \mathbf{m}_i, \boldsymbol{\eta}_\mathbf{y} \rangle + \mu_{\mathbf{y},i}, \ \eta_{\mathbf{y},i}), \ \forall i \in \mathcal{I},$$

$$g_i(\boldsymbol{\eta}_\mathbf{w}, \boldsymbol{\eta}_\mathbf{y}) = \eta_{\mathbf{w},i} - \eta_{\mathbf{y},i}, \ \forall i \in \mathcal{I}.$$

where $\mathbf{m}_i$ is the ith row of $\mathbf{M}$. Then the linear complementarity problem (3.6.5) is equivalent to the equations:

$$\mathbf{0} = \mathbf{W}(\boldsymbol{\mu}_\mathbf{y}, \boldsymbol{\eta}_\mathbf{y}), \ \boldsymbol{\eta}_\mathbf{w} = \mathbf{M}\boldsymbol{\eta}_\mathbf{y} + \boldsymbol{\mu}_\mathbf{y}.$$

Note that $\mathbf{W}$ is a piecewise continuously differentiable function. If $\mathbf{W}(\bar{\boldsymbol{\mu}}_\mathbf{y}, \bar{\boldsymbol{\eta}}_\mathbf{y}) = \mathbf{0}$, then there exists a piecewise continuously differentiable function $\mathbf{H} : \mathbb{R}^{n_y} \to \mathcal{Y}$, such that $\mathbf{W}(\boldsymbol{\mu}_\mathbf{y}, \mathbf{H}(\boldsymbol{\mu}_\mathbf{y})) = \mathbf{0}, \ \forall \boldsymbol{\mu}_\mathbf{y} \in \mathbb{R}^{n_y}$ (See Example 17 in [91]). Let $\bar{\boldsymbol{\eta}}_\mathbf{w} = \mathbf{M}\bar{\boldsymbol{\eta}}_\mathbf{y} + \bar{\boldsymbol{\mu}}_\mathbf{y}$. De-

fine the sets

$$\mathcal{I}_1(\bar{\boldsymbol{\mu}}_{\mathbf{y}}) = \{i \in \mathcal{I} : g_i(\bar{\boldsymbol{\eta}}_{\mathbf{w}}, \bar{\boldsymbol{\eta}}_{\mathbf{y}}) > 0\},$$
$$\mathcal{I}_2(\bar{\boldsymbol{\mu}}_{\mathbf{y}}) = \{i \in \mathcal{I} : g_i(\bar{\boldsymbol{\eta}}_{\mathbf{w}}, \bar{\boldsymbol{\eta}}_{\mathbf{y}}) < 0\},$$
$$\mathcal{I}_3(\bar{\boldsymbol{\mu}}_{\mathbf{y}}) = \{i \in \mathcal{I} : g_i(\bar{\boldsymbol{\eta}}_{\mathbf{w}}, \bar{\boldsymbol{\eta}}_{\mathbf{y}}) = 0\}.$$

Then (3.6.5) can be written in the form

$$\mathbf{0} = \mathbf{R}(\mathcal{I}_1(\bar{\boldsymbol{\mu}}_{\mathbf{y}}), \mathcal{I}_2(\bar{\boldsymbol{\mu}}_{\mathbf{y}}), \mathcal{I}_3(\bar{\boldsymbol{\mu}}_{\mathbf{y}}))\bar{\boldsymbol{\eta}}_{\mathbf{y}} + \mathbf{K}(\mathcal{I}_1(\bar{\boldsymbol{\mu}}_{\mathbf{y}}), \mathcal{I}_2(\bar{\boldsymbol{\mu}}_{\mathbf{y}})\mathcal{I}_3(\bar{\boldsymbol{\mu}}_{\mathbf{y}}))\bar{\boldsymbol{\mu}}_{\mathbf{y}},$$
$$\bar{\boldsymbol{\eta}}_{\mathbf{w}} = \mathbf{M}\bar{\boldsymbol{\eta}}_{\mathbf{y}} + \bar{\boldsymbol{\mu}}_{\mathbf{y}},$$

where

$$\mathbf{R}_i(\mathcal{I}_1(\bar{\boldsymbol{\mu}}_{\mathbf{y}}), \mathcal{I}_2(\bar{\boldsymbol{\mu}}_{\mathbf{y}}), \mathcal{I}_3(\bar{\boldsymbol{\mu}}_{\mathbf{y}})) = \begin{cases} \mathbf{e}_i^{\mathrm{T}} & \text{if } i \in \mathcal{I}_1(\bar{\boldsymbol{\mu}}_{\mathbf{y}}), \\ \mathbf{m}_i & \text{if } i \in \mathcal{I}_2(\bar{\boldsymbol{\mu}}_{\mathbf{y}}), \\ \mathbf{e}_i^{\mathrm{T}} \text{ or } \mathbf{m}_i & \text{if } i \in \mathcal{I}_3(\bar{\boldsymbol{\mu}}_{\mathbf{y}}), \end{cases}$$
$$\mathbf{K}_i(\mathcal{I}_1(\bar{\boldsymbol{\mu}}_{\mathbf{y}}), \mathcal{I}_2(\bar{\boldsymbol{\mu}}_{\mathbf{y}}), \mathcal{I}_3(\bar{\boldsymbol{\mu}}_{\mathbf{y}})) = \begin{cases} \mathbf{0} & \text{if } i \in \mathcal{I}_1(\bar{\boldsymbol{\mu}}_{\mathbf{y}}), \\ \mathbf{e}_i^{\mathrm{T}} & \text{if } i \in \mathcal{I}_2(\bar{\boldsymbol{\mu}}_{\mathbf{y}}), \\ \mathbf{e}_i^{\mathrm{T}} \text{ or } \mathbf{0} & \text{if } i \in \mathcal{I}_3(\bar{\boldsymbol{\mu}}_{\mathbf{y}}), \end{cases}$$

$\mathbf{R}_i$ and $\mathbf{K}_i$ are the ith rows of $\mathbf{R}$ and $\mathbf{K}$, respectively. Observe that $\mathbf{R}(\mathcal{I}_1(\bar{\boldsymbol{\mu}}_{\mathbf{y}}), \mathcal{I}_2(\bar{\boldsymbol{\mu}}_{\mathbf{y}}), \mathcal{I}_3(\bar{\boldsymbol{\mu}}_{\mathbf{y}}))$ is invertible because $\mathbf{M}$ is a P-Matrix. If $\mathcal{I}_3(\bar{\boldsymbol{\mu}}_{\mathbf{y}})$ is empty, then for $(\boldsymbol{\mu}_{\mathbf{y}}, \boldsymbol{\eta}_{\mathbf{y}})$ in a neighborhood of $(\bar{\boldsymbol{\mu}}_{\mathbf{y}}, \bar{\boldsymbol{\eta}}_{\mathbf{y}})$

$$\mathbf{0} = \mathbf{R}(\mathcal{I}_1(\bar{\boldsymbol{\mu}}_{\mathbf{y}}), \mathcal{I}_2(\bar{\boldsymbol{\mu}}_{\mathbf{y}}), \mathcal{I}_3(\bar{\boldsymbol{\mu}}_{\mathbf{y}}))\boldsymbol{\eta}_{\mathbf{y}} + \mathbf{K}(\mathcal{I}_1(\bar{\boldsymbol{\mu}}_{\mathbf{y}}), \mathcal{I}_2(\bar{\boldsymbol{\mu}}_{\mathbf{y}}), \mathcal{I}_3(\bar{\boldsymbol{\mu}}_{\mathbf{y}}))\boldsymbol{\mu}_{\mathbf{y}}$$

holds due to the continuity of $\mathbf{W}$ and g_i. In this case, if $\bar{\boldsymbol{\mu}}_{\mathbf{y}} = \mathbf{Q}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}))$, $\bar{\boldsymbol{\eta}}_{\mathbf{y}} = \mathbf{y}(t, \mathbf{p})$, $\bar{\boldsymbol{\eta}}_{\mathbf{w}} = \mathbf{w}(t, \mathbf{p})$, then the system of equations (3.6.1) to (3.6.4) are a set of continuously differentiable hence strictly differentiable equations.

In the remainder, if $i \in \mathcal{I}_3(\boldsymbol{\mu}_{\mathbf{y}})$, then it is treated as if $i \in \mathcal{I}_2(\boldsymbol{\mu}_{\mathbf{y}})$ holds. Then $\mathcal{I}_1(\boldsymbol{\mu}_{\mathbf{y}}) \cup \mathcal{I}_2(\boldsymbol{\mu}_{\mathbf{y}}) = \mathcal{I}$. As a result, $\mathbf{R}$ and $\mathbf{K}$ can be written as $\mathbf{R}(\boldsymbol{\mu}_{\mathbf{y}})$ and $\mathbf{K}(\boldsymbol{\mu}_{\mathbf{y}})$. If $\boldsymbol{\mu}_{\mathbf{y}}$ depends on other variables, those variables are substituted for $\boldsymbol{\mu}_{\mathbf{y}}$.

Using the fact that the composition of locally Lipschitz continuous functions is locally Lipschitz continuous, the existence of $\mathbf{H}$ and the invertibility of $\mathbf{J}_{\dot{\mathbf{x}}}\mathbf{V}$, it can be shown that there exist locally Lipschitz continuous functions, $\mathbf{f} : \mathcal{T} \times \mathcal{P} \times \mathcal{X} \to \dot{\mathcal{X}}$ and $\mathbf{r} : \mathcal{T} \times \mathcal{P} \times \mathcal{X} \to \mathcal{Y}$, such that $\mathbf{0} = \mathbf{V}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}}, \mathbf{r}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}}), \mathbf{M}\mathbf{r}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}}) + \mathbf{Q}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}}), \mathbf{f}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}}))$. Then existence and uniqueness of solutions to equations (3.6.1) to (3.6.4) can be analyzed using results for ordinary differential equations to show that $(\mathbf{x}(\cdot, \mathbf{p}), \mathbf{y}(\cdot, \mathbf{p}), \mathbf{w}(\cdot, \mathbf{p}), \dot{\mathbf{x}}(\cdot, \mathbf{p}))$ is a continuous function on T.

Let $\mathbf{F} : T \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \mathcal{W} \times \dot{\mathcal{X}} \to \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_x}$ be

$$\mathbf{F}(t, \boldsymbol{\eta}) = \begin{bmatrix} \mathbf{V}(t, \boldsymbol{\eta}) \\ \boldsymbol{\eta}_{\mathbf{w}} - \mathbf{M}\boldsymbol{\eta}_{\mathbf{y}} - \mathbf{Q}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}}) \\ \mathbf{R}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}})\boldsymbol{\eta}_{\mathbf{y}} + \mathbf{K}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}})\mathbf{Q}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}}) \end{bmatrix}.$$

Let the mapping $\boldsymbol{\zeta} \mapsto \mathbf{F}(t, \boldsymbol{\zeta})$ be differentiable at $\boldsymbol{\eta}$, let the derivative be $\mathbf{J}_{\mathbf{u}}\mathbf{F}(t, \boldsymbol{\eta}) = [\mathbf{A}(t, \boldsymbol{\eta}) \ \mathbf{B}(t, \boldsymbol{\eta})]$ (the notation is modified here and the subscript $\mathbf{u}$ is associated with the space $\mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \mathcal{W} \times \dot{\mathcal{X}}$) where

$$\mathbf{B}(t, \boldsymbol{\eta}) = \begin{bmatrix} \mathbf{J}_{\mathbf{y}}\mathbf{V}(t, \boldsymbol{\eta}) & \mathbf{J}_{\mathbf{w}}\mathbf{V}(t, \boldsymbol{\eta}) & \mathbf{J}_{\dot{\mathbf{x}}}\mathbf{V}(t, \boldsymbol{\eta}) \\ -\mathbf{M} & \mathbf{I}_{n_y} & \mathbf{0} \\ \mathbf{R}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}}) & \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{3.6.6}$$

$$\mathbf{A}(t,\boldsymbol{\eta}) = \begin{bmatrix} \mathbf{J_p V}(t,\boldsymbol{\eta}) & \mathbf{J_x V}(t,\boldsymbol{\eta}) \\ -\mathbf{J_p Q}(t,\boldsymbol{\eta_p},\boldsymbol{\eta_x}) & -\mathbf{J_x Q}(t,\boldsymbol{\eta_p},\boldsymbol{\eta_x}) \\ \mathbf{K}(t,\boldsymbol{\eta_p},\boldsymbol{\eta_x})\mathbf{J_p Q}(t,\boldsymbol{\eta_p},\boldsymbol{\eta_x}) & \mathbf{K}(t,\boldsymbol{\eta_p},\boldsymbol{\eta_x})\mathbf{J_x Q}(t,\boldsymbol{\eta_p},\boldsymbol{\eta_x}) \end{bmatrix}.$$

Let the solution of the system of equations (3.6.1) to (3.6.4) be $(\mathbf{x}(\cdot,\mathbf{p}),\mathbf{y}(\cdot,\mathbf{p}),\ \mathbf{w}(\cdot,\mathbf{p}),\ \dot{\mathbf{x}}(\cdot,\mathbf{p}))$. Let $\mathbf{u}(t,\mathbf{p}) = (\mathbf{p},\mathbf{x}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p}),\mathbf{w}(t,\mathbf{p}),\dot{\mathbf{x}}(t,\mathbf{p}))$. Let S be the set $\{t : t \in T,\ g_i(\mathbf{w}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p})) = 0,\ \text{for some } i \in \mathcal{I}\}$. If S is a measure zero subset of T and $\mathbf{B}(t,\mathbf{u}(t,\mathbf{p}))$ is invertible for all $t \in T\backslash S$, then Theorem 3.3.6 can be used to obtain forward sensitivities. Finally observe that $\mathbf{K}$ and $\mathbf{R}$ are constant as long as $\mathcal{I}_1(t,\mathbf{v}(t,\mathbf{p}))$ and $\mathcal{I}_2(t,\mathbf{v}(t,\mathbf{p}))$ remain constant. If S is a set of measure zero, $a \in S$ and $b \in S$ and $t \notin S$ if $t \in (a,b)$, then due to the continuity of $\mathbf{u}(\cdot,\mathbf{p})$, $\mathbf{K}$ and $\mathbf{R}$ are constant for $t \in (a,b)$.

Chapter 4

Parametric Sensitivity Analysis of Dynamic Systems using Linear Newton Approximations

In Chapter 3, sufficient conditions for the existence of the strict derivative were analyzed for the mapping $\boldsymbol{\eta} \mapsto \mathbf{x}(t_f, \boldsymbol{\eta})$ at $\mathbf{p} \in \mathcal{P}$, where $\mathbf{x} : [t_0, t_f] \times \mathcal{P} \to \mathcal{X}$ was the solution of the initial value problem:

$$\dot{\mathbf{x}}(t, \mathbf{p}) = \mathbf{f}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})),\ \forall t \in (t_0, t_f],\ \mathbf{x}(t_0, \mathbf{p}) = \mathbf{f}_0(\mathbf{p}),\ \forall \mathbf{p} \in \mathcal{P} \subset \mathbb{R}^{n_p}, \tag{4.0.1}$$

where $\mathbf{f}$ and $\mathbf{f}_0$ were locally Lipschitz functions on their respective domains. Forward and adjoint sensitivity initial value problems were derived. The results were extended to DAEs that can be transformed into ODEs using the implicit function theorem for locally Lipschitz continuous functions (Theorem 2.6.13). Finally, the results were extended to multistage systems where at each stage the evolution of the states was governed by such a DAE.

The results in Chapter 3 required that the state trajectory visit points of nondifferen-

tiability in the domain of $\mathbf{f}$ only at times that constitute a measure zero subset of the time horizon, $[t_0, t_f]$. In this chapter, the case where this requirement is not met is analyzed. In this case, Theorem 3.1.3 states that $\partial_2 \mathbf{x}(t, \mathbf{x}_2) \subset R(t_f, \mathbf{x}_2)$, but provides no efficient means to calculate an element of $\partial_2 \mathbf{x}(t, \mathbf{x}_2)$. The theorem does not state whether $\Phi(t_f, \mathbf{x}_2)$, the set whose elements can be easily computed, contains $\partial_2 \mathbf{x}(t, \mathbf{x}_2)$ and it is not clear whether $R(t_f, \mathbf{x}_2)$ can be used as a surrogate for the generalized Jacobian.

In order to arrive at sharper results than Theorem 3.1.3 can provide, additional conditions on the functions involved in (4.0.1) are imposed. The functions are assumed to be semismooth in the restricted sense in addition to being locally Lipschitz continuous. As a consequence of this assumption, linear Newton approximations that contain the generalized Jacobian of the mapping $\boldsymbol{\eta} \mapsto \mathbf{x}(t_f, \boldsymbol{\eta})$ can be derived and equations can be formulated to calculate an element of these linear Newton approximations.

The results in this chapter depend on results in [42] and [81]. In §4.1, these results are summarized. In §4.2, results are derived for (4.0.1) assuming $\mathbf{f}$ and $\mathbf{f}_0$ are semismooth in the restricted sense. The results are extended to a set of DAEs using an implicit function theorem for semismooth functions derived from results in [42]. Then, multistage DAEs are considered. Finally, Example 3.6.3 is revisited.

4.1 Preliminaries

4.1.1 Note on Notation and Assumptions

In the remainder of this chapter, n_x, n_p and n_y represent finite positive integers, $t_0 \in \mathbb{R}$, $t_f \in \mathbb{R}$ and $t_0 \leq t_f$.

X_1, X_2, X_3, X_4, X_5 and X_6 are open connected subsets of $\mathbb{R}$, $\mathbb{R}^{n_p}$, $\mathbb{R}^{n_x}$, $\mathbb{R}^{n_y}$, $\mathbb{R}^{n_x}$ and $\mathbb{R}^{n_y}$, respectively. $X_7 = X_2 \times X_3$, $X_8 = X_2 \times X_3 \times X_4 \times X_5$, $X_9 = X_4 \times X_5$, $X_{10} = X_7 \times \mathbb{R}$ and $X_{11} = X_3 \times \mathbb{R}$. $T = [t_0, t_f] \subset X_1$.

In order to make the exposition more intuitive, the labels $\mathcal{T}$, $\mathcal{P}$, $\mathcal{X}$, $\mathcal{Y}$, $\dot{\mathcal{X}}$, $\mathcal{W}$ and $\mathcal{Q}$ will be used instead of X_1, X_2, X_3, X_4, X_5, X_6 and X_9. If the symbols $t, \mathbf{p}, \mathbf{x}, \mathbf{y}, \dot{\mathbf{x}}, \mathbf{w}, \mathbf{v}, \mathbf{u}, \mathbf{q}, \bar{\mathbf{v}}$ and $\bar{\mathbf{x}}$ appear as subscripts, they represent the indices $1, 2, 3, 4, 5, 6, 7, 8, 9, 10$ and 11.

In the remainder of this chapter, semismoothness means semismoothness in the restricted sense and if F is a scalar function, ∂F denotes its generalized Jacobian and not its generalized gradient to make the exposition simpler. The topological degree and related concepts are used to formulate conditions for the existence of implicit nonsmooth functions. Here, relevant properties of the topological degree are summarized (see [35] for a more detailed treatment of degree theory).

Definition 4.1.1 (Topological Degree). *Let* $\mathbf{F} : \mathrm{cl}\,(\Omega) \to \mathbb{R}^n$ *be a continuous function where* Ω *is a nonempty bounded open subset of* $\mathbb{R}^n$. *Let* $\mathrm{bd}\,(\Omega) = \mathrm{cl}\,(\Omega)\backslash\Omega$. *Let* $\mathbf{y} \in \mathbb{R}^n\backslash\mathbf{F}(\mathrm{bd}\,(\Omega))$. *The degree of* $\mathbf{F}$ *on* Ω *at* $\mathbf{y}$ *is denoted by* $\deg(\mathbf{F}, \Omega, \mathbf{y})$, *takes integer values and satisfies the following properties:*

1. *Let* $\mathbf{y} \in \mathbb{R}^n\backslash\mathbf{F}(\mathrm{bd}\,(\Omega))$. *If* $\deg(\mathbf{F}, \Omega, \mathbf{y}) \neq 0$, *then there exists,* $\mathbf{u}^* \in \Omega$, *a solution to the equation* $\mathbf{F}(\mathbf{u}) = \mathbf{y}$.
2. *If* $\mathbf{y} \notin \mathrm{cl}\,(\mathbf{F}(\Omega))$, $\deg(\mathbf{F}, \Omega, \mathbf{y}) = 0$.
3. $\deg(\mathbf{I}, \Omega, \mathbf{y}) = 1$ *if* $\mathbf{y} \in \Omega$ *where* $\mathbf{I}$ *is the identity map.*
4. $\deg(\mathbf{F}, \Omega, \mathbf{y}) = \deg(\mathbf{F}, \Omega_1, \mathbf{y}) + \deg(\mathbf{F}, \Omega_2, \mathbf{y})$ *if* Ω_1 *and* Ω_2 *are two disjoint open subsets of* Ω *and* $\mathbf{y} \notin \mathbf{F}(\mathrm{cl}\,(\Omega)\backslash(\Omega_1 \cup \Omega_2))$.
5. *Let* $\mathbf{H} : [0, 1] \times \mathrm{bd}\,(\Omega) \to \mathbb{R}^n$ *be a continuous function. Let* $\mathbf{H}(0, \mathbf{x}) = \mathbf{F}(\mathbf{x})$ *and* $\mathbf{H}(1, \mathbf{x}) = \mathbf{G}(\mathbf{x})$ *for all* $\mathbf{x} \in \mathrm{bd}\,(\Omega)$. *If* $\mathbf{y} \notin \mathbf{H}(t, \mathrm{bd}\,(\Omega))$ *for all* $t \in [0, 1]$, *then* $\deg(\mathbf{F}, \Omega, \mathbf{y}) = \deg(\mathbf{G}, \Omega, \mathbf{y})$.
6. *Let* $\mathbf{G}$ *be a continuous function on* $\mathrm{cl}\,(\Omega)$. $\deg(\mathbf{F}, \Omega, \mathbf{y}) = \deg(\mathbf{G}, \Omega, \mathbf{y})$ *if*

$$\max_{\mathbf{x}\in\mathrm{cl}(\Omega)} \|\mathbf{F}(\mathbf{x}) - \mathbf{G}(\mathbf{x})\|_\infty \leq \mathrm{dist}_\infty(\mathbf{y}, \mathbf{F}(\mathrm{bd}\,(\Omega)))$$

where $\|\cdot\|_\infty$ *is the max norm and*

$$\mathrm{dist}_\infty(\mathbf{y}, \mathbf{F}(\mathrm{bd}\,(\Omega)) = \inf_{\mathbf{z}\in\mathbf{F}(\mathrm{bd}(\Omega))} \|\mathbf{z}-\mathbf{y}\|_\infty.$$

7. *Let* $(\mathbf{y}_1, \mathbf{y}_2) \in \mathbb{R}^n \backslash \mathbf{F}(\mathrm{bd}\,(\Omega)) \times \mathbb{R}^n \backslash \mathbf{F}(\mathrm{bd}\,(\Omega))$. *Then,* $\deg(\mathbf{F}, \Omega, \mathbf{y}_1) = \deg(\mathbf{F}, \Omega, \mathbf{y}_2)$ *if*

$$\|\mathbf{y}_1 - \mathbf{y}_2\|_\infty \leq \mathrm{dist}_\infty(\mathbf{y}_1, \mathbf{F}(\mathrm{bd}\,(\Omega))).$$

8. $\deg(\mathbf{F}, \Omega, \mathbf{y}) = \deg(\mathbf{F}, \Omega_1, \mathbf{y})$ *for every open subset* Ω_1 *of* Ω *such that* $\mathbf{y} \notin \mathbf{F}(\Omega \backslash \Omega_1)$.
9. *Let* $\mathbf{y} \in \mathbb{R}^n \backslash \mathbf{F}(bd\,(\Omega))$. *If* Ω_1 *and* Ω_2 *are two disjoint open sets whose union is* Ω, *then* $\deg(\mathbf{F}, \Omega, \mathbf{y}) = \deg(\mathbf{F}, \Omega_1, \mathbf{y}) + \deg(\mathbf{F}, \Omega_2, \mathbf{y})$.
10. *Let* $\mathbf{y}_1 \in \mathbb{R}^n \backslash \mathbf{F}(\mathrm{bd}\,(\Omega))$. *Let* Ω' *be a nonempty bounded open subset of* $\mathbb{R}^m$. *Let* $\mathbf{G} : \mathrm{cl}\,(\Omega') \to \mathbb{R}^m$ *be a continuous function and* $\mathbf{y}_2 \in \mathbb{R}^m \backslash \mathbf{G}(\mathrm{bd}\,(\Omega'))$, *then*

$$\deg(\mathbf{F} \times \mathbf{G}, \Omega \times \Omega', (\mathbf{y}_1, \mathbf{y}_2)) = \deg(\mathbf{F}, \Omega, \mathbf{y}_1)\deg(\mathbf{G}, \Omega', \mathbf{y}_2).$$

Definition 4.1.2 (Index of a function). *Let* $\mathbf{F} : \mathrm{cl}\,(\Omega) \to \mathbb{R}^n$ *be a continuous function where* Ω *is a nonempty bounded open subset of* $\mathbb{R}^n$. *Let* $\mathbf{y}^* \in \mathbb{R}^n \backslash \mathbf{F}(\mathrm{bd}\,(\Omega))$. *Let* $\mathbf{x}^*$ *be an isolated solution of the equation* $\mathbf{F}(\mathbf{x}) = \mathbf{y}^*$, *i.e.,* $\mathbf{F}^{-1}(\mathbf{y}^*) \cap \mathrm{cl}\,(\Omega_1) = \{\mathbf{x}^*\}$ *where* Ω_1 *is an open subset of* Ω *containing* $\mathbf{x}^*$. *Then*

$$\deg(\mathbf{F}, \Omega_1, \mathbf{y}^*) = \deg(\mathbf{F}, \Omega_2, \mathbf{y}^*)$$

where Ω_2 *is a neighborhood of* $\mathbf{x}^*$ *such that* $\Omega_2 \subset \Omega_1$. *In this case, the index of* $\mathbf{F}$ *at* $\mathbf{x}^*$ *denoted by* $\mathrm{ind}(\mathbf{F}, \mathbf{x}^*)$ *takes the value* $\deg(\mathbf{F}, \Omega_1, \mathbf{y}^*)$ *and satisfies the following properties*

1. *If* $\mathbf{x}^*$ *is a Fréchet differentiable point of* $\mathbf{F}$, *then* $\mathrm{ind}(\mathbf{F}, \mathbf{x}^*)$ *is equal to the signum of the determinant of* $\mathbf{JF}(\mathbf{x}^*)$.

2. *If the equation* $\mathbf{F}(\mathbf{x}) = \mathbf{y}^*$ *has finitely many solutions* $\{\mathbf{x}_i^*\}_{i=1}^k$ *in* Ω, *then* $\deg(\mathbf{F}, \Omega, \mathbf{y}^*) = \sum_{i=1}^k \text{ind}(\mathbf{F}, \mathbf{x}_i^*)$.

Next, two extensions of the derivative similar to the generalized Jacobian are introduced. They appear as intermediate quantities when deriving the necessary relations to compute elements of the linear Newton approximations that contain the generalized Jacobian.

Definition 4.1.3 (B-Subdifferential). *Let* $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ *be a locally Lipschitz continuous function. Then, the B-subdifferential at* $\mathbf{x} \in \mathbb{R}$ *is defined by*

$$\partial^B \mathbf{F}(\mathbf{x}) = \bigcap_{\delta > 0} \mathbf{JF}(\text{cl}\left((\mathbf{x} + \delta \mathbb{B}(0,1))\right) \cap \Omega_{\mathbf{F}})$$

where $\Omega_{\mathbf{F}}$ is the set of points where $\mathbf{F}$ is differentiable and $\mathbb{B}(0,1)$ the open unit ball in $\mathbb{R}^n$. Equivalently,

$$\partial^B \mathbf{F}(\mathbf{x}) = \left\{ \lim_{i \to \infty} \mathbf{JF}(\mathbf{x}_i) : \mathbf{x}_i \to \mathbf{x},\ \mathbf{x}_i \in \Omega_{\mathbf{F}} \right\}.$$

Definition 4.1.4 (BN Generalized Jacobian). *Let* $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ *be a locally Lipschitz continuous function. Then, the BN generalized Jacobian at* $\mathbf{x} \in \mathbb{R}$ *is defined by*

$$\partial^{BN} \mathbf{F}(\mathbf{x}) = \bigcap_{N : |N| = 0} \bigcap_{\delta > 0} \mathbf{JF}(\text{cl}\left((\mathbf{x} + \delta \mathbb{B}(0,1))\right) \cap (\Omega_{\mathbf{F}} \backslash N))$$

where $|N|$ is the Lebesgue measure of set N and $\Omega_{\mathbf{F}}$ is the set of points where $\mathbf{F}$ is differentiable.

The following Lemma summarizes the properties of the B-subdifferential and BN generalized Jacobian. The results are from Lemma 5 in [81] and [35].

Lemma 4.1.5. *Let* $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ *be a locally Lipschitz continuous function. Let* $\mathbf{x} \in \mathbb{R}^n$. *Let* $\Omega_{\mathbf{F}}$ *be the set of points where* $\mathbf{F}$ *is differentiable. Then the following hold:*

1. $\partial^{BN}\mathbf{F}(\mathbf{x}) \subset \partial^{B}\mathbf{F}(\mathbf{x}) \subset \partial\mathbf{F}(\mathbf{x})$.
2. $\text{conv}\left(\partial^{B}\mathbf{F}(\mathbf{x})\right) = \partial\mathbf{F}(\mathbf{x})$.
3. $\text{conv}\left(\partial^{BN}\mathbf{F}(\mathbf{x})\right) = \partial\mathbf{F}(\mathbf{x})$.
4. $\mathbf{JF}(\mathbf{y}) \in \partial^{B}\mathbf{F}(\mathbf{y})$ *for all* $\mathbf{y} \in \Omega_{\mathbf{F}}$.
5. $\mathbf{JF}(\mathbf{y}) \in \partial^{BN}\mathbf{F}(\mathbf{y})$ *for all* $\mathbf{y} \in \Omega_{\mathbf{F}} \backslash S$ *where* S *is a measure zero subset of* $\mathbb{R}^n$.
6. $\partial^{B}\mathbf{F}$ *and* $\partial^{BN}\mathbf{F}$ *are uppersemicontinuous, nonempty and compact-valued set-valued maps from* $\mathbb{R}^n$ *to* $\mathbb{R}^{m\times n}$.
7. *If* $\mathbf{F} \in PC^1(O)$ *where* O *is an open subset of* $\mathbb{R}^n$, *then* $\partial^{B}\mathbf{F}(\mathbf{x}) = \{\mathbf{JF}_i(\mathbf{x}) : i \in \mathcal{I}(\mathbf{F},\mathbf{x})\}$ *for all* $\mathbf{x} \in O$ *where* $\mathcal{I}(\mathbf{F},\mathbf{x})$ *is the set of essentially active function indices at* $\mathbf{x}$ *defined in §2.7.*

Definition 4.1.6 (The Projection of the B-Subdifferential and BN Generalized Jacobian). *Let* $\mathbf{F} : X_1 \times X_2 \to \mathbb{R}^p$ *where* X_1 *and* X_2 *are open subsets of* $\mathbb{R}^n$ *and* $\mathbb{R}^m$, *respectively. Let* $\mathbf{F}$ *be locally Lipschitz continuous at* $(\mathbf{x}_1,\mathbf{x}_2)$ *where* $\mathbf{x}_1 \in X_1$ *and* $\mathbf{x}_2 \in X_2$. *Then* $\pi_2\partial^{B}\mathbf{F}(\mathbf{x}_1,\mathbf{x}_2)$ *is the set* $\{\mathbf{M} \in \mathbb{R}^{p\times m} : \exists \mathbf{N} \in \mathbb{R}^{p\times n}$ *such that* $[\mathbf{N}\ \mathbf{M}] \in \partial^{B}\mathbf{F}(\mathbf{x}_1,\mathbf{x}_2)\}$. *Analogously,* $\pi_1\partial^{B}\mathbf{F}(\mathbf{x}_1,\mathbf{x}_2)$ *is the set* $\{\mathbf{M} \in \mathbb{R}^{p\times n} : \exists \mathbf{N} \in \mathbb{R}^{p\times m}$ *such that* $[\mathbf{M}\ \mathbf{N}] \in \partial^{B}\mathbf{F}(\mathbf{x}_1,\mathbf{x}_2)\}$.

The projection of the BN Generalized Jacobian is similarly defined with the B-subdifferential replaced with the BN Generalized Jacobian.

The following is an implicit function theorem derived from Theorem 4 in [42] using Corollary 4 in [42].

Theorem 4.1.7. *Let* X_1 *be an open subset of* $\mathbb{R}^m$ *and* X_2 *be an open subset of* $\mathbb{R}^n$. *Let* $\mathbf{F} : X_1 \times X_2 \to \mathbb{R}^n$ *be a semismooth function. Let* $\mathbf{x}^* \in X_1$, $\mathbf{y}^* \in X_2$ *and* $\mathbf{z}^* = (\mathbf{x}^*,\mathbf{y}^*)$. *Let* $\mathbf{F}(\mathbf{z}^*) = \mathbf{0}$. *Assume the following hold:*

1. $\pi_2\partial^{B}\mathbf{F}(\mathbf{z}^*)$ *is coherently oriented, i.e, the determinants of all the matrices in* $\pi_2\partial^{B}\mathbf{F}(\mathbf{z}^*)$ *have the same nonzero sign* α.
2. $\text{ind}(\mathbf{h},\mathbf{y}^*) = \alpha$ *where* $\mathbf{h} : X_2 \to \mathbb{R}^n : \mathbf{y} \mapsto \mathbf{F}(\mathbf{x}^*,\mathbf{y})$.

Then, there exists, U, an open subset of $X_1 \times X_2$, W, an open subset of X_1 and a semismooth function $\mathbf{G} : W \to X_2$ such that $\mathbf{z}^* \in U$ and $\mathbf{x}^* \in W$, $\mathbf{F}(\mathbf{x}, \mathbf{G}(\mathbf{x})) = \mathbf{0}$, $(\mathbf{x}, \mathbf{G}(\mathbf{x})) \in U$ for all $\mathbf{x} \in W$.

The set-valued mapping $\Gamma\mathbf{G} : W \rightrightarrows \mathbb{R}^{n \times m}$ defined by

$$\Gamma\mathbf{G}(\mathbf{x}) = \operatorname{conv}\left(\{-\boldsymbol{\pi}_2\mathbf{M}^{-1}\boldsymbol{\pi}_1\mathbf{M} : \ [\boldsymbol{\pi}_1\mathbf{M}\ \boldsymbol{\pi}_2\mathbf{M}] \in \partial^B\mathbf{F}(\mathbf{x}, \mathbf{y}),\ \boldsymbol{\pi}_1\mathbf{M} \in \mathbb{R}^{n \times m},\ \boldsymbol{\pi}_2\mathbf{M} \in \mathbb{R}^{n \times n}\}\right)$$

is a linear Newton approximation of $\mathbf{G}$ at $\mathbf{x}$ such that $\partial\mathbf{G}(\bar{\mathbf{x}}) \subset \operatorname{conv}(\Gamma\mathbf{G}(\bar{\mathbf{x}}))$ holds for all $\bar{\mathbf{x}} \in W$.

Proof. The first part of the Theorem follows from Theorem 4 in [42]. Note that Theorem 4 provides an implicit function that is semismooth in the original sense (Definition 2.8.7). Semismoothness in the restricted sense follows from the fact that on an open neighborhood W containing $\mathbf{x}^*$, G, the implicit function is semismooth in the original sense which implies that the implicit function is B-differentiable on that set per the properties of semismooth functions.

In order to derive the linear Newton approximation $\Gamma\mathbf{G}$, the result in Corollary 4 in [42] is used as follows: Let $\mathbf{H} : X_1 \times X_2 \to X_1 \times X_2 : (\mathbf{x}, \mathbf{y}) \mapsto (\mathbf{x}, \mathbf{F}(\mathbf{x}, \mathbf{y}))$. Then $\mathbf{H}$ is a semismooth function on U as a composition of semismooth functions and

$$\partial^B\mathbf{H}(\mathbf{x}, \mathbf{y}) = \left\{ \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \boldsymbol{\pi}_1\mathbf{M} & \boldsymbol{\pi}_2\mathbf{M} \end{bmatrix} : [\boldsymbol{\pi}_1\mathbf{M}\ \boldsymbol{\pi}_2\mathbf{M}] \in \partial^B\mathbf{F}(\mathbf{x}, \mathbf{y}),\ \boldsymbol{\pi}_1\mathbf{M} \in \mathbb{R}^{n \times m},\ \boldsymbol{\pi}_2\mathbf{M} \in \mathbb{R}^{n \times n} \right\}$$

is obtained using the definition of the B-subdifferential. In addition, $\mathbf{H}$ has a semismooth inverse, $\mathbf{H}^{-1} : W \times V \to U$ where V is an open subset of $\mathbb{R}^n$ such that $\mathbf{0} \in V$. This can be

shown as in the proof of Theorem 4. Furthermore,

$$\partial^B \mathbf{H}^{-1}(\mathbf{x},\mathbf{0}) = \left\{ \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \boldsymbol{\pi}_1\mathbf{M} & \boldsymbol{\pi}_2\mathbf{M} \end{bmatrix}^{-1} : [\boldsymbol{\pi}_1\mathbf{M}\ \boldsymbol{\pi}_2\mathbf{M}] \in \partial^B \mathbf{F}(\mathbf{x},\mathbf{G}(\mathbf{x})),\ \boldsymbol{\pi}_1\mathbf{M} \in \mathbb{R}^{n\times m},\ \boldsymbol{\pi}_2\mathbf{M} \in \mathbb{R}^{n\times n} \right\}$$

holds for all $\mathbf{x} \in W$ per Corollary 4 in [42]. Observe that

$$\begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \boldsymbol{\pi}_1\mathbf{M} & \boldsymbol{\pi}_2\mathbf{M} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ -\boldsymbol{\pi}_2\mathbf{M}^{-1}\boldsymbol{\pi}_1\mathbf{M} & \boldsymbol{\pi}_2\mathbf{M}^{-1} \end{bmatrix}$$

and $\partial \mathbf{H}^{-1}(\mathbf{x},\mathbf{0}) = \operatorname{conv}\left(\partial^B \mathbf{H}^{-1}(\mathbf{x},\mathbf{0})\right)$.

The mapping $\boldsymbol{\eta} \mapsto \mathbf{G}(\boldsymbol{\eta})$ is equal to $\boldsymbol{\eta} \mapsto (\boldsymbol{\eta},\mathbf{0}) \mapsto \mathbf{H}^{-1}(\boldsymbol{\eta},\mathbf{0}) \mapsto \mathbf{G}(\boldsymbol{\eta})$. Then, using Theorem 2.6.7,

$$\partial \mathbf{G}(\mathbf{x}) \subset \operatorname{conv}\left(\left\{ \begin{bmatrix} \mathbf{0} & \mathbf{I}_n \end{bmatrix} \mathbf{N} \begin{bmatrix} \mathbf{I}_m \\ \mathbf{0} \end{bmatrix} : \mathbf{N} \in \partial \mathbf{H}^{-1}(\mathbf{x},\mathbf{0}) \right\} \right)$$

is obtained. Since $\partial \mathbf{H}^{-1}(\mathbf{x},\mathbf{0})$ is a linear Newton approximation of the mapping $(\mathbf{x},\mathbf{0}) \mapsto (\mathbf{x},\mathbf{G}(\mathbf{x}))$, the set $\Gamma\mathbf{G} : W \rightrightarrows \mathbb{R}^{n\times m}$ defined by

$$\Gamma\mathbf{G}(\mathbf{x}) = \left\{ \begin{bmatrix} \mathbf{0} & \mathbf{I}_n \end{bmatrix} \mathbf{N} \begin{bmatrix} \mathbf{I}_m \\ \mathbf{0} \end{bmatrix} : \mathbf{N} \in \partial \mathbf{H}^{-1}(\mathbf{x},\mathbf{0}) \right\}$$

is a linear Newton approximation of $\mathbf{G}$ at $\mathbf{x}$ per Theorem 2.8.12 such that $\partial \mathbf{G}(\mathbf{x}) \subset \operatorname{conv}(\Gamma\mathbf{G}(\mathbf{x}))$ holds for all $\mathbf{x} \in W$. The result

$$\Gamma\mathbf{G}(\mathbf{x}) = \operatorname{conv}\left(\{-\boldsymbol{\pi}_2\mathbf{M}^{-1}\boldsymbol{\pi}_1\mathbf{M} :\ [\boldsymbol{\pi}_1\mathbf{M}\ \boldsymbol{\pi}_2\mathbf{M}] \in \partial^B \mathbf{F}(\mathbf{x},\mathbf{G}(\mathbf{x})),\ \boldsymbol{\pi}_1\mathbf{M} \in \mathbb{R}^{n\times m},\ \boldsymbol{\pi}_2\mathbf{M} \in \mathbb{R}^{n\times n}\}\right)$$

can be obtained using the fact that $\partial \mathbf{H}^{-1}(\mathbf{x},\mathbf{0}) = \operatorname{conv}\left(\partial^B \mathbf{H}^{-1}(\mathbf{x},\mathbf{0})\right)$ and the definition of

$\partial^B \mathbf{H}^{-1}(\mathbf{x}, \mathbf{0})$. □

The following theorems are used to derive linear Newton approximations to the map $\boldsymbol{\eta} \mapsto \mathbf{x}(t, \boldsymbol{\eta})$.

Theorem 4.1.8. *Let X_1 and X_2 be open connected subsets of $\mathbb{R}$ and $\mathbb{R}^n$, respectively. Let $\tilde{X}_2$ be an open connected subset of X_2. Let $T = [t_0, t_f] \subset X_1$. Let $\mathbf{f} : X_2 \to \mathbb{R}^n$ be semismooth.*

Let $\mathbf{x} : [t_0, t_f] \times \tilde{X}_2 \to X_2$ be such that $\mathbf{x}(\cdot, \mathbf{x}_2)$ is the only function that satisfies

$$\dot{\mathbf{x}}(t, \mathbf{x}_2) = \mathbf{f}(\mathbf{x}(t, \mathbf{x}_2)),\ \forall t \in (t_0, t_f],\ \mathbf{x}(t_0, \mathbf{x}_2) = \mathbf{x}_2,\ \forall \mathbf{x}_2 \in \tilde{X}_2.$$

Then $\mathbf{x}(t, \cdot)$ is semismooth at $\mathbf{x}_2 \in \tilde{X}_2$ for all $t \in T$ (Theorem 8 in [81]).

Theorem 4.1.9. *Let the assumptions of Theorem 4.1.8 hold. Assume $\Gamma\mathbf{f} : X_2 \rightrightarrows \mathbb{R}^{n \times n}$ is a linear Newton approximation of $\mathbf{f}$ on X_2. Then, there exists a neighborhood $O \subset \tilde{X}_2$ such that, the set-valued map, $\Gamma_2 : T \times O \rightrightarrows \mathbb{R}^{n \times n}$ defined by*

$$\Gamma_2(t, \boldsymbol{\eta}) = \{\mathbf{Y}(t, \boldsymbol{\eta}) : \dot{\mathbf{Y}}(t, \boldsymbol{\eta}) \in \operatorname{conv}(\Gamma\mathbf{f}(\mathbf{x}(t, \boldsymbol{\eta})))\mathbf{Y}(t, \boldsymbol{\eta}),\ \forall t \in (t_0, t_f],\ \mathbf{Y}(t_0, \boldsymbol{\eta}) = \mathbf{I}_n\}$$

is a linear Newton approximation of $\mathbf{x}(t, \cdot)$ at $\mathbf{x}_2 \in O$ for all $t \in T$ (Theorem 11 in [81]).

Corollary 4.1.10. *Let assumptions and definitions of Theorem 4.1.9 hold. In addition, let $\Gamma\mathbf{f}(\mathbf{y}) = \partial\mathbf{f}(\mathbf{y}), \forall \mathbf{y} \in X_2$. Then, the result of Theorem 4.1.9 holds. Let $\Gamma_2(t, \mathbf{x}_2)$, $\mathbf{x}_2 \in O$ be computed as in Theorem 4.1.9 with $\Gamma\mathbf{f}(\mathbf{x}(t, \mathbf{x}_2)) = \partial\mathbf{f}(\mathbf{x}(t, \mathbf{x}_2))$. Then $\partial_2^{BN}\mathbf{x}(t, \mathbf{x}_2) \subset \Gamma_2(t, \mathbf{x}_2)$ where $\partial_2^{BN}\mathbf{x}(t, \mathbf{x}_2)$ is the BN generalized Jacobian of the mapping $\boldsymbol{\eta} \mapsto \mathbf{x}(t, \boldsymbol{\eta})$ at $\mathbf{x}_2$ (Corollary 12 in [81]). In addition, $\partial_2\mathbf{x}(t, \mathbf{x}_2) \subset \operatorname{conv}(\Gamma_2(t, \mathbf{x}_2))$ holds per Lemma 4.1.5.*

Remark 4.1.11. The results of Corollary 4.1.10 still hold if $\Gamma\mathbf{f}(\mathbf{x}(t, \boldsymbol{\eta}))$ is replaced with $\operatorname{conv}\left(\bar{\Gamma}\mathbf{f}(\mathbf{x}(t, \boldsymbol{\eta}))\right)$ where $\bar{\Gamma}\mathbf{f}$ is a linear Newton approximation of $\mathbf{f}$ that satisfies $\partial\mathbf{f}(\mathbf{y}) \subset \operatorname{conv}\left(\bar{\Gamma}\mathbf{f}(\mathbf{x}(\mathbf{y}))\right)$ for all $\mathbf{y} \in X_2$.

Definition 4.1.12 (The Projection of a Linear Newton Approximation). *Let* $\mathbf{F} : X_1 \times X_2 \to \mathbb{R}^p$ *where* X_1 *and* X_2 *are open subsets of* $\mathbb{R}^n$ *and* $\mathbb{R}^m$*, respectively. Let* $\mathbf{F}$ *be semismooth at* $(\mathbf{x}_1, \mathbf{x}_2)$ *where* $\mathbf{x}_1 \in X_1$ *and* $\mathbf{x}_2 \in X_2$*. Let* $\Gamma\mathbf{F}$ *be a linear Newton approximation of* $\mathbf{F}$ *in a neighborhood of* $(\mathbf{x}_1, \mathbf{x}_2)$*. Then* $\pi_2\Gamma\mathbf{F}(\mathbf{x}_1, \mathbf{x}_2)$ *is the set* $\{\mathbf{M} \in \mathbb{R}^{p\times m} : \exists \mathbf{N} \in \mathbb{R}^{p\times n}$ *such that* $[\mathbf{N}\ \mathbf{M}] \in \Gamma\mathbf{F}(\mathbf{x}_1, \mathbf{x}_2)\}$*. Analogously,* $\pi_1\Gamma\mathbf{F}(\mathbf{x}_1, \mathbf{x}_2)$ *is the set* $\{\mathbf{M} \in \mathbb{R}^{p\times n} : \exists \mathbf{N} \in \mathbb{R}^{p\times m}$ *such that* $[\mathbf{M}\ \mathbf{N}] \in \Gamma\mathbf{F}(\mathbf{x}_1, \mathbf{x}_2)\}$.

The following extends the definition of a measurable selection in Definition 3.1.1 to arbitrary set-valued maps with nonempty and closed images.

Definition 4.1.13 (Measurable Selection of Set-Valued Map). *Let* X *be a closed or open subset of* $\mathbb{R}^n$ *and* $\mathcal{S} : X \rightrightarrows \mathbb{R}^m$ *be set-valued map such that* $\mathcal{S}(\mathbf{x})$ *is a non-empty closed set for all* $\mathbf{x} \in X$*. Let* $\mathbf{s} : X \mapsto \mathbb{R}^m$ *be a Lebesgue measurable function on* X *such that* $\mathbf{s}(\mathbf{x}) \in \mathcal{S}(\mathbf{x}), \forall \mathbf{x} \in X$*. Then* $\mathbf{s}$ *is a measurable selection of* $\mathcal{S}$ *on* X*. This is denoted by* $\mathbf{s} \in \mathcal{L}(X, \mathcal{S})$.

The following result combines Theorem 8.1.3 and Proposition 8.2.1 in [5] and states sufficient conditions for the existence of a measurable selection.

Theorem 4.1.14 (Existence of Measurable Selections of Upper Semicontinuous Set-Valued Maps). *Let* X *be a closed or open subset of* $\mathbb{R}^n$ *and* $\mathcal{S} : X \rightrightarrows \mathbb{R}^m$ *be an upper semicontinuous set-valued map such that* $\mathcal{S}(\mathbf{x})$ *is a non-empty closed set for all* $\mathbf{x} \in X$*. Then there exists a measurable selection* $\mathbf{s} : X \to \mathbb{R}^m$ *of* $\mathcal{S}$ *on* X.

4.2 Ordinary Differential Equations

This section develops results for ordinary differential equations that satisfy the following conditions.

Assumption 4.2.1. *Let* $\mathbf{f} : \mathcal{T} \times \mathcal{P} \times \mathcal{X} \to \dot{\mathcal{X}}$ *and* $\mathbf{f}_0 : \mathcal{P} \to \tilde{\mathcal{X}}$ *be semismooth where* $\tilde{\mathcal{X}}$ *is an open connected subset of* $\mathcal{X}$. *Let* $\mathbf{x} : T \times \mathcal{P} \to \mathcal{X}$ *be such that* $\mathbf{x}(\cdot, \mathbf{p})$ *is the unique solution of the initial value problem*

$$\dot{\mathbf{x}}(t, \mathbf{p}) = \mathbf{f}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})), \ \forall t \in (t_0, t_f], \ \mathbf{x}(t_0, \mathbf{p}) = \mathbf{f}_0(\mathbf{p}), \ \forall \mathbf{p} \in \mathcal{P}. \tag{4.2.1}$$

Let $\Gamma\mathbf{f} : \mathcal{T} \times \mathcal{P} \times \mathcal{X} \rightrightarrows \mathbb{R}^{(n_x)\times(1+n_p+n_x)}$ *be a linear Newton approximation of* $\mathbf{f}$ *satisfying* $\partial\mathbf{f}(\boldsymbol{\mu}) \subset \text{conv}\,(\Gamma\mathbf{f}(\boldsymbol{\mu}))$ *for all* $\boldsymbol{\mu} \in \mathcal{T} \times \mathcal{P} \times \mathcal{X}$. *In addition, let* $\Gamma\mathbf{f}_0 : \mathcal{P} \rightrightarrows \mathbb{R}^{n_x \times n_p}$ *be a linear Newton approximation of* $\mathbf{f}_0$ *such that* $\partial\mathbf{f}_0(\boldsymbol{\eta}) \subset \text{conv}\,(\Gamma\mathbf{f}_0(\boldsymbol{\eta}))$ *for all* $\boldsymbol{\eta} \in \mathcal{P}$.

Remark 4.2.2. *Let* $\mathbf{z} : T \times \mathcal{P} \to \mathcal{P} : (t, \mathbf{p}) \mapsto \mathbf{p}$ *and* $\mathbf{v} : T \times \mathcal{P} \to \mathcal{P} \times \mathcal{X} : (t, \mathbf{p}) \mapsto (\mathbf{z}(t, \mathbf{p}), \mathbf{x}(t, \mathbf{p}))$ *for the remainder of this chapter.*

Theorem 4.2.3. *Let Assumption 4.2.1 hold. Then, the mapping* $\boldsymbol{\eta} \mapsto \mathbf{x}(t_f, \boldsymbol{\eta})$ *is semismooth at* $\mathbf{p}$. *Let* $\Gamma_{\mathbf{p}}\mathbf{x}(t_f, \mathbf{p})$ *be the set*

$$\begin{aligned}\{\mathbf{m}(t_f, \mathbf{p}) : \dot{\mathbf{m}}(t, \mathbf{p}) = \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t, \mathbf{p})\mathbf{m}(t, \mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t, \mathbf{p}), \ \forall t \in (t_0, t_f], \\ \mathbf{G}_f(\cdot, \mathbf{p}) \in \mathcal{L}(T, \pi_{\mathbf{v}}\text{conv}\,(\Gamma\mathbf{f}(\cdot, \mathbf{v}(\cdot, \mathbf{p})))), \\ \mathbf{G}_f(t, \mathbf{p}) = [\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t, \mathbf{p}) \ \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t, \mathbf{p})], \forall t \in T, \\ \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t, \mathbf{p}) \in \mathbb{R}^{n_x \times n_p}, \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t, \mathbf{p}) \in \mathbb{R}^{n_x \times n_x}, \ \forall t \in T, \\ \mathbf{m}(t_0, \mathbf{p}) \in \text{conv}\,(\Gamma\mathbf{f}_0(\mathbf{p}))\}.\end{aligned} \tag{4.2.2}$$

Then $\Gamma_{\mathbf{p}}\mathbf{x}(t_f, \mathbf{p})$ *is a linear Newton approximation of the map* $\boldsymbol{\eta} \mapsto \mathbf{x}(t_f, \boldsymbol{\eta})$ *at* $\mathbf{p}$ *and*

$$\partial_{\mathbf{p}}\mathbf{x}(t_f, \mathbf{p}) \subset \text{conv}\,(\Gamma_{\mathbf{p}}\mathbf{x}(t_f, \mathbf{p}))$$

holds.

Proof. Let $\Delta t = t_f - t_0$. Let $\tilde{\mathcal{T}}$ be a an open connected subset of $\mathcal{T}$ such that if $t \in \tilde{\mathcal{T}}$, then

$(t + \Delta t) \in \mathcal{T}$ and $t_0 \in \tilde{\mathcal{T}}$. Let $\bar{t}_0 \in \tilde{\mathcal{T}}$, $\bar{\mathbf{p}} \in \mathcal{P}$, $\bar{\mathbf{x}}_0 \in \tilde{\mathcal{X}}$ and $\boldsymbol{\nu}_0 = (\bar{t}_0, \bar{\mathbf{p}}, \bar{\mathbf{x}}_0)$.

Let $\boldsymbol{\nu}_t : [0, \Delta t] \times \tilde{\mathcal{T}} \times \mathcal{P} \times \mathcal{X} \to \mathcal{T}$, $\boldsymbol{\nu}_{\mathbf{p}} : [0, \Delta t] \times \tilde{\mathcal{T}} \times \mathcal{P} \times \mathcal{X} \to \mathcal{P}$ and $\boldsymbol{\nu}_{\mathbf{x}} : [0, \Delta t] \times \tilde{\mathcal{T}} \times \mathcal{P} \times \mathcal{X} \to \mathcal{X}$ be continuous functions and let $\boldsymbol{\nu}(\tau, \boldsymbol{\nu}_0) = (\boldsymbol{\nu}_t(\tau, \boldsymbol{\nu}_0), \boldsymbol{\nu}_{\mathbf{p}}(\tau, \boldsymbol{\nu}_0), \boldsymbol{\nu}_{\mathbf{x}}(\tau, \boldsymbol{\nu}_0))$. Let $\boldsymbol{\nu}_{\mathbf{v}}(\tau, \boldsymbol{\nu}_0) = (\boldsymbol{\nu}_{\mathbf{p}}(\tau, \boldsymbol{\nu}_0), \boldsymbol{\nu}_{\mathbf{x}}(\tau, \boldsymbol{\nu}_0))$. Let $\mathbf{g} : \mathcal{T} \times \mathcal{P} \times \mathcal{X} \to \mathbb{R} \times \mathbb{R}^{n_p} \times \dot{\mathcal{X}} : \boldsymbol{\mu} \mapsto (1, \mathbf{0}, \mathbf{f}(\boldsymbol{\mu}))$.

Consider the following augmented initial value problem:

$$\dot{\boldsymbol{\nu}}(\tau, \boldsymbol{\nu}_0) = \mathbf{g}(\boldsymbol{\nu}(\tau, \boldsymbol{\nu}_0)), \ \forall \tau \in (0, \Delta t], \ \boldsymbol{\nu}(0) = \boldsymbol{\nu}_0. \tag{4.2.3}$$

Note that:

$$\boldsymbol{\nu}_t(\tau, \boldsymbol{\nu}_0) = \bar{t}_0 + \tau, \forall \bar{t}_0 \in \tilde{\mathcal{T}}, \tag{4.2.4}$$

$$\boldsymbol{\nu}_{\mathbf{p}}(\tau, \boldsymbol{\nu}_0) = \bar{\mathbf{p}}, \forall \bar{\mathbf{p}} \in \mathcal{P}. \tag{4.2.5}$$

Observe that $\mathbf{g}$ is semismooth as a composition of functions that are semismooth. Therefore, the mapping $\boldsymbol{\nu}_0 \mapsto \boldsymbol{\nu}(\tau, \boldsymbol{\nu}_0)$ is semismooth per Theorem 4.1.8 if a unique solution $\boldsymbol{\nu}(\tau, \boldsymbol{\nu}_0)$ exists. As a result, the mapping $(\bar{\mathbf{p}}, \bar{\mathbf{x}}_0) \mapsto \boldsymbol{\nu}_{\mathbf{x}}(\tau, (\bar{t}_0, \bar{\mathbf{p}}, \bar{\mathbf{x}}_0))$ is semismooth for all $\bar{t}_0 \in \tilde{\mathcal{T}}$ and for all $\tau \in [0, \Delta t]$.

Let $\bar{\boldsymbol{\nu}}_0 = (t_0, \mathbf{p}, \mathbf{f}_0(\mathbf{p}))$. Then,

$$\boldsymbol{\nu}_{\mathbf{x}}(\tau, \bar{\boldsymbol{\nu}}_0) = \mathbf{x}(\tau + t_0, \mathbf{p}), \tag{4.2.6}$$

(4.2.4) and (4.2.5) satisfy (4.2.3) with $\bar{t}_0 = t_0$ and $\bar{\mathbf{p}} = \mathbf{p}$. Observe that for any given $\mathbf{p} \in \mathcal{P}$ and $\bar{t}_0 = t_0$, (4.2.6) holds.

The mapping $\boldsymbol{\eta} \mapsto \mathbf{f}_0(\boldsymbol{\eta})$ is semismooth per assumptions. Since the composition of semismooth functions is semismooth, the semismoothness of the mapping $\boldsymbol{\eta} \to \mathbf{x}(t, \boldsymbol{\eta})$ at $\mathbf{p}$ follows from the semismoothness of the mapping $(\bar{\mathbf{p}}, \bar{\mathbf{x}}_0) \mapsto (\boldsymbol{\nu}_{\mathbf{x}}(\tau, (t_0, \bar{\mathbf{p}}, \bar{\mathbf{x}}_0))$ and the equivalence in (4.2.6).

The generalized Jacobian of $\mathbf{g}$ at $\boldsymbol{\nu}(t, \boldsymbol{\nu}_0)$ is

$$\partial \mathbf{g}(\boldsymbol{\nu}(t, \boldsymbol{\nu}_0)) = \left\{ \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{M} \end{bmatrix}, \ \mathbf{M} \in \partial \mathbf{f}(\boldsymbol{\nu}(\tau, \boldsymbol{\nu}_0)) \right\}$$

per Theorem 2.6.7. Define the set-valued mapping $\Gamma\mathbf{g} : \tilde{\mathcal{T}} \times \mathcal{P} \times \mathcal{X} \rightrightarrows \mathbb{R}^{(1+n_p+n_x)\times(1+n_p+n_x)}$ by

$$\Gamma\mathbf{g}(\boldsymbol{\nu}(t, \boldsymbol{\nu}_0)) = \left\{ \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{M} \end{bmatrix}, \ \mathbf{M} \in \operatorname{conv}\left(\Gamma\mathbf{f}(\boldsymbol{\nu}(\tau, \boldsymbol{\nu}_0))\right) \right\}.$$

Note that $\Gamma\mathbf{g}$ is a linear Newton approximation of $\mathbf{g}$ per Theorem 2.8.12 and $\partial \mathbf{g}(\boldsymbol{\nu}(t, \boldsymbol{\nu}_0)) \subset \Gamma\mathbf{g}(\boldsymbol{\nu}(t, \boldsymbol{\nu}_0))$ holds because of the assumptions.

Then, the set-valued map $\Gamma_{\boldsymbol{\nu}_0}\boldsymbol{\nu} : [0, \Delta t] \times \tilde{\mathcal{T}} \times \mathcal{P} \times \mathcal{X} \rightrightarrows \mathbb{R}^{(1+n_p+n_x)\times(1+n_p+n_x)}$ defined by

$$\Gamma_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(\tau, \boldsymbol{\nu}_0) = \{\mathbf{Y}(\tau, \boldsymbol{\nu}_0) : \dot{\mathbf{Y}}(\epsilon, \boldsymbol{\nu}_0) \in \Gamma\mathbf{g}(\boldsymbol{\nu}(\epsilon, \boldsymbol{\nu}_0))\mathbf{Y}(\epsilon, \boldsymbol{\nu}_0), \forall \epsilon \in (0, \tau], \ \mathbf{Y}(0, \boldsymbol{\nu}_0) = \mathbf{I}_{1+n_p+n_x}\} \tag{4.2.7}$$

is a linear Newton approximation of the map $\boldsymbol{\mu} \mapsto \boldsymbol{\nu}(\tau, \boldsymbol{\mu})$ at $\boldsymbol{\nu}_0$ for all $\tau \in [0, \Delta t]$ and $\partial_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(\tau, \boldsymbol{\nu}_0) \subset \operatorname{conv}\left(\Gamma_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(\tau, \boldsymbol{\nu}_0)\right)$ per Corollary 4.1.10.

In order to derive a linear Newton approximation for the mapping $(\bar{\mathbf{p}}, \bar{\mathbf{x}}_0) \mapsto (\boldsymbol{\nu}_{\mathbf{p}}(\tau, \boldsymbol{\nu}_0), \boldsymbol{\nu}_{\mathbf{x}}(\tau, \boldsymbol{\nu}_0))$ for the case $\bar{t}_0 = t_0$, the composition of the following three maps is considered:

1. $(\bar{\mathbf{p}}, \bar{\mathbf{x}}_0) \mapsto (t_0, \bar{\mathbf{p}}, \bar{\mathbf{x}}_0)$. The generalized Jacobian of this map is a singleton and the single

element is

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{n_p+n_x} \end{bmatrix}.$$

The generalized Jacobian also defines a linear Newton approximation for this map because the map is semismooth.

2. $\boldsymbol{\nu}_0 \mapsto \boldsymbol{\nu}(\tau, \boldsymbol{\nu}_0)$. $\Gamma_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(\tau, \boldsymbol{\nu}_0)$ is the linear Newton approximation of this map.
3. $(\boldsymbol{\nu}_t(\tau, \boldsymbol{\nu}_0), \boldsymbol{\nu}_{\mathbf{p}}(\tau, \boldsymbol{\nu}_0), \boldsymbol{\nu}_{\mathbf{x}}(\tau, \boldsymbol{\nu}_0)) \mapsto (\boldsymbol{\nu}_{\mathbf{p}}(\tau, \boldsymbol{\nu}_0), \boldsymbol{\nu}_{\mathbf{x}}(\tau, \boldsymbol{\nu}_0))$. The generalized Jacobian of this map is a singleton set with the following element

$$\mathbf{B} = \begin{bmatrix} \mathbf{0} & \mathbf{I}_{n_p+n_x} \end{bmatrix}.$$

The generalized Jacobian also defines a linear Newton approximation for this map.

Per the chain rule for linear Newton approximations (Theorem 2.8.12), the set

$$\{\mathbf{BMA} : \mathbf{M} \in \Gamma_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(\tau, \boldsymbol{\nu}_0)\}$$

is a linear Newton approximation of the map $(\bar{\mathbf{p}}, \bar{\mathbf{x}}_0) \mapsto (\boldsymbol{\nu}_{\mathbf{p}}(\tau, \boldsymbol{\nu}_0), \boldsymbol{\nu}_{\mathbf{x}}(\tau, \boldsymbol{\nu}_0))$ (Note that $\mathbf{v}_0$ is used as an index instead of $\mathbf{v}$ here). Per the chain rule for the generalized Jacobian (Theorem 2.6.7) and the fact that $\partial_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(\tau, \boldsymbol{\nu}_0) \subset \text{conv}\,(\Gamma_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(\tau, \boldsymbol{\nu}_0))$,

$$\partial_{\mathbf{v}_0}\boldsymbol{\nu}_{\mathbf{v}}(\tau, \boldsymbol{\nu}_0) \subset \text{conv}\,(\{\mathbf{BMA} : \mathbf{M} \in \partial_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(\tau, \boldsymbol{\nu}_0)\}) \subset \text{conv}\,(\Gamma_{\mathbf{v}_0}\boldsymbol{\nu}_{\mathbf{v}}(\tau, \boldsymbol{\nu}_0)) \tag{4.2.8}$$

holds.

In order to compute an element of $\Gamma_{\mathbf{v}_0}\boldsymbol{\nu}_{\mathbf{v}}(\tau, \boldsymbol{\nu}_0)$ one can solve a reduced system of equations instead of (4.2.7). Let $\mathbf{Z} : [0, \Delta t] \to \mathbb{R}^{(1+n_p+n_x)\times(1+n_p+n_x)}$ be a measurable selection of $\Gamma\mathbf{g}(\boldsymbol{\nu}(\cdot, \boldsymbol{\nu}_0))$ on $[0, \Delta t]$. Then, $\mathbf{Y}(\tau, \boldsymbol{\nu}_0)$, an element of $\Gamma_{\boldsymbol{\nu}_0}\boldsymbol{\nu}(\tau, \boldsymbol{\nu}_0)$, can be computed by

solving

$$\dot{\mathbf{Y}}(\epsilon, \boldsymbol{\nu}_0) = \mathbf{Z}(\epsilon)\mathbf{Y}(\epsilon, \boldsymbol{\nu}_0), \ \forall \epsilon \in (0, \tau], \ \mathbf{Y}(0, \boldsymbol{\nu}_0) = \mathbf{I}_{1+n_p+n_x}. \tag{4.2.9}$$

The linear differential equation (4.2.9) admits a matrix-valued function $\boldsymbol{\Gamma}(\tau, \epsilon)$ such that $\mathbf{Y}(\tau, \boldsymbol{\nu}_0) = \boldsymbol{\Gamma}(\tau, 0)\mathbf{I}_{1+n_p+n_x}$. This implies that $\mathbf{Y}(\tau, \boldsymbol{\nu}_0)\mathbf{A} = \boldsymbol{\Gamma}(\tau, 0)\mathbf{A}$. This in turn implies that the product $\mathbf{Y}(\tau, \boldsymbol{\nu}_0)\mathbf{A}$ can be computed by solving

$$\dot{\mathbf{W}}(\epsilon, \boldsymbol{\nu}_0) = \mathbf{Z}(\epsilon)\mathbf{W}(\epsilon, \boldsymbol{\nu}_0), \ \forall \epsilon \in (0, \tau], \ \mathbf{W}(0, \boldsymbol{\nu}_0) = \mathbf{A}. \tag{4.2.10}$$

The product of $\mathbf{Z}(0^+)\mathbf{A}$ has zero first row because the first row of $\mathbf{Z}(\epsilon)$ is zero for all $\epsilon \in (0, \Delta t]$ and $\mathbf{A}$ has zero first row. As a result $\dot{\mathbf{W}}(0^+)$ has a zero first row. Since $\mathbf{W}(0)$ and $\dot{\mathbf{W}}(0^+)$ have zero first rows, $\mathbf{W}(\epsilon)$ has zero first row for any $\epsilon \in (0, \tau]$. As a result, the first column of $\mathbf{Z}(\epsilon)$ has no effect on the evolution of $\mathbf{W}$. In addition, the first row of $\mathbf{Z}(\epsilon)$ has no effect on the evolution of $\mathbf{W}$.

The pre-multiplication with $\mathbf{B}$ removes the zero first row from $\mathbf{W}(\tau)$ to produce an $(n_p + n_x) \times (n_p + n_x)$ matrix, $\mathbf{M}(\tau, \boldsymbol{\nu}_0)$, which can be computed by solving the differential equation

$$\dot{\mathbf{M}}(\epsilon, \boldsymbol{\nu}_0) = \tilde{\mathbf{Z}}(\epsilon)\mathbf{M}(\epsilon, \boldsymbol{\nu}_0), \ \forall \epsilon \in (0, \tau], \ \mathbf{M}(0, \boldsymbol{\nu}_0) = \mathbf{I}_{n_p+n_x}. \tag{4.2.11}$$

where

$$\tilde{\mathbf{Z}}(\epsilon) = \begin{bmatrix} \mathbf{0} \\ \mathbf{H}(\epsilon) \end{bmatrix},$$

and $\mathbf{H} : [0, \Delta t] \to \mathbb{R}^{n_x \times (n_p + n_x)}$ is a measurable selection of $\text{conv}\,(\pi_{\mathbf{v}} \Gamma \mathbf{f}(\boldsymbol{\nu}(\cdot, \boldsymbol{\nu}_0)))$. Hence

$$\Gamma_{\mathbf{v}_0} \boldsymbol{\nu}_{\mathbf{v}}(\tau, \boldsymbol{\nu}_0) = \{\mathbf{M}(\tau, \boldsymbol{\nu}_0) : \dot{\mathbf{M}}(\epsilon, \boldsymbol{\nu}_0) \in Z(\epsilon, \boldsymbol{\nu}_0)\mathbf{M}(\epsilon, \boldsymbol{\nu}_0),\ \forall \epsilon \in (0, \tau], \mathbf{M}(0, \boldsymbol{\nu}_0) = \mathbf{I}_{n_p + n_x}\}. \tag{4.2.12}$$

where

$$Z(\epsilon, \boldsymbol{\nu}_0) = \left\{ \begin{bmatrix} \mathbf{0} \\ \mathbf{N} \end{bmatrix} : \mathbf{N} \in \pi_{\mathbf{v}} \text{conv}\,(\Gamma \mathbf{f}(\boldsymbol{\nu}(\epsilon, \boldsymbol{\nu}_0))) \right\}.$$

In order to derive a linear Newton approximation for the mapping $\boldsymbol{\eta} \mapsto \mathbf{x}(t, \boldsymbol{\eta})$, the composition of the following functions is considered:

1. $\bar{\mathbf{p}} \mapsto (\bar{\mathbf{p}}, \mathbf{f}_0(\bar{\mathbf{p}}))$. A linear Newton approximation of this map is the set:

$$C = \left\{ \begin{bmatrix} \mathbf{I}_{n_p} \\ \mathbf{N} \end{bmatrix} : \mathbf{N} \in \text{conv}\,(\Gamma \mathbf{f}_0(\bar{\mathbf{p}})) \right\}$$

   per Theorem 2.8.12.

2. $(\bar{\mathbf{p}}, \bar{\mathbf{x}}_0) \mapsto \boldsymbol{\nu}_{\mathbf{v}}(\tau, \boldsymbol{\nu}_0)$. The linear Newton approximation for this map is $\Gamma_{\mathbf{v}_0} \boldsymbol{\nu}_{\mathbf{v}}(\tau, \boldsymbol{\nu}_0)$.

3. $\boldsymbol{\nu}_{\mathbf{v}}(\tau, \boldsymbol{\nu}_0) \mapsto \boldsymbol{\nu}_{\mathbf{x}}(\tau, \boldsymbol{\nu}_0)$. A linear Newton approximation of this mapping is the singleton set whose single element is

$$\mathbf{D} = \begin{bmatrix} \mathbf{0} & \mathbf{I}_{n_x} \end{bmatrix},$$

   which is also the element of the singleton generalized Jacobian.

It can be shown that

$$\partial_{\mathbf{p}} \boldsymbol{\nu}_{\mathbf{x}}(\tau, \boldsymbol{\nu}_0) \subset \text{conv}\,(\{\mathbf{DMN} : \mathbf{M} \in \partial_{\mathbf{v}_0} \boldsymbol{\nu}_{\mathbf{v}}(\tau, \boldsymbol{\nu}_0),\ \mathbf{N} \in C\}) \subset \text{conv}\,(\Gamma_{\mathbf{p}} \boldsymbol{\nu}_{\mathbf{x}}(\tau, \boldsymbol{\nu}_0))$$

where

$$\Gamma_{\mathbf{p}}\boldsymbol{\nu}_{\mathbf{x}}(\tau,\boldsymbol{\nu}_0)=\{\mathbf{DMN}:\ \mathbf{M}\in\Gamma_{\mathbf{v}_0}\boldsymbol{\nu}_{\mathbf{v}}(\tau,\boldsymbol{\nu}_0),\ \mathbf{N}\in C\}$$

using the relation (4.2.8), Theorem 2.6.7 and Theorem 2.8.12. Setting the initial condition to $\mathbf{N}\in C$ in (4.2.12) results in the set

$$\Gamma_{\mathbf{p}}\boldsymbol{\nu}_{\mathbf{v}}(\tau,\boldsymbol{\nu}_0)=\{\mathbf{M}(\tau,\boldsymbol{\nu}_0):\ \dot{\mathbf{M}}(\epsilon,\boldsymbol{\nu}_0)\in Z(\epsilon,\boldsymbol{\nu}_0)\mathbf{M}(\epsilon,\boldsymbol{\nu}_0),\ \forall\epsilon\in(0,\tau],\mathbf{M}(0,\boldsymbol{\nu}_0)=\mathbf{N},\ \mathbf{N}\in C\}$$

which is a linear Newton approximation of the map $\mathbf{p}\mapsto\boldsymbol{\nu}_{\mathbf{v}}(\tau,\boldsymbol{\nu}_0)$ per arguments similar to those used in the derivation of (4.2.10). Let $\mathbf{M}(\tau,\boldsymbol{\nu}_0)$ be an element of $\Gamma_{\mathbf{p}}\boldsymbol{\nu}_{\mathbf{v}}(\tau,\boldsymbol{\nu}_0)$. Pre-multiplication with $\mathbf{D}$ produces an $n_x\times n_p$ matrix that contains the last n_x rows of $\mathbf{M}(\tau,\boldsymbol{\nu}_0)$.

Note that the elements in the first n_p rows of $\mathbf{M}(\epsilon,\boldsymbol{\nu}_0)$ are constant for all $\epsilon\in(0,\tau]$ because the first n_p rows of any element of $Z(\epsilon,\boldsymbol{\nu}_0)$ constitute a zero matrix. Therefore, given $\mathbf{G}_f(\cdot,\mathbf{p})\in\mathcal{L}(T,\pi_{\mathbf{v}}\text{conv}\,(\Gamma\mathbf{f}(\boldsymbol{\nu}(\cdot,\boldsymbol{\nu}_0))))$, and $\mathbf{N}\in C$, an element of $\Gamma_{\mathbf{p}}\boldsymbol{\nu}_{\mathbf{v}}(\tau,\boldsymbol{\nu}_0)$ can be computed by

$$\begin{bmatrix}\dot{\mathbf{n}}(\epsilon,\boldsymbol{\nu}_0)\\ \dot{\mathbf{m}}(\epsilon,\boldsymbol{\nu}_0)\end{bmatrix}=\begin{bmatrix}\mathbf{0}\\ \mathbf{G}_f(\epsilon,\bar{\mathbf{p}})\end{bmatrix}\begin{bmatrix}\mathbf{n}(\epsilon,\boldsymbol{\nu}_0)\\ \mathbf{m}(\epsilon,\boldsymbol{\nu}_0)\end{bmatrix},\ \forall\epsilon\in(0,\Delta t],\ \mathbf{m}(0,\boldsymbol{\nu}_0)\in\text{conv}\,(\Gamma\mathbf{f}_0(\bar{\mathbf{p}})),\ \mathbf{n}(0,\boldsymbol{\nu}_0)=\mathbf{I}_{n_p},$$

where $\mathbf{n}(\epsilon,\boldsymbol{\nu}_0)\in\mathbb{R}^{n_p\times n_p}$ and $\mathbf{m}(\epsilon,\boldsymbol{\nu}_0)\in\mathbb{R}^{n_x\times n_p}$. Let $\mathbf{G}_f(\epsilon,\bar{\mathbf{p}})=[\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(\epsilon,\bar{\mathbf{p}})\ \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(\epsilon,\bar{\mathbf{p}})]$ where $\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(\epsilon,\bar{\mathbf{p}})\in\mathbb{R}^{n_x\times n_p}$ and $\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(\epsilon,\bar{\mathbf{p}})\in\mathbb{R}^{n_x\times n_x}$, then the evolution of $\mathbf{m}(\epsilon,\boldsymbol{\nu}_0)$ is governed by

$$\dot{\mathbf{m}}(\epsilon,\boldsymbol{\nu}_0)=\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(\epsilon,\bar{\mathbf{p}})\mathbf{m}(\epsilon,\boldsymbol{\nu}_0)+\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(\epsilon,\bar{\mathbf{p}}),\forall\epsilon\in(0,\Delta t],\ \mathbf{m}(\epsilon,\boldsymbol{\nu}_0)\in\text{conv}\,(\Gamma\mathbf{f}_0(\bar{\mathbf{p}})).$$

Hence

$$\Gamma_{\mathbf{p}}\boldsymbol{\nu}_{\mathbf{x}}(\tau,\boldsymbol{\nu}_0) = \{\mathbf{m}(\tau,\boldsymbol{\nu}_0):\ \dot{\mathbf{m}}(\epsilon,\boldsymbol{\nu}_0) = \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(\epsilon,\bar{\mathbf{p}})\mathbf{m}(\epsilon,\boldsymbol{\nu}_0) + \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(\epsilon,\bar{\mathbf{p}}),\ \forall\epsilon\in(0,\Delta t],$$
$$\mathbf{G}_f(\cdot,\bar{\mathbf{p}}) \in \mathcal{L}(T,\pi_{\mathbf{v}}\text{conv}\,(\Gamma\mathbf{f}(\boldsymbol{\nu}(\cdot,\boldsymbol{\nu}_0)))),$$
$$\mathbf{G}_f(\epsilon,\bar{\mathbf{p}}) = [\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(\epsilon,\bar{\mathbf{p}})\ \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(\epsilon,\bar{\mathbf{p}})],\ \forall\epsilon\in[0,\Delta t],$$
$$\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(\epsilon,\bar{\mathbf{p}}) \in \mathbb{R}^{n_x\times n_p},\ \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(\epsilon,\bar{\mathbf{p}}) \in \mathbb{R}^{n_x\times n_p},\ \forall\epsilon\in[0,\Delta t],$$
$$\mathbf{m}(0,\boldsymbol{\nu}_0) \in \text{conv}\,(\Gamma\mathbf{f}_0(\bar{\mathbf{p}}))\}.$$

If $\boldsymbol{\nu}_0 = (t_0,\mathbf{p},\mathbf{f}_0(\mathbf{p}))$, then $\boldsymbol{\nu}(\tau,\boldsymbol{\nu}_0) = (t,\mathbf{p},\mathbf{x}(t,\mathbf{p}))$ where $t = \tau+t_0$. Note that $\boldsymbol{\nu}_0$ is a function of $\mathbf{p}$ only. Renaming the quantities appropriately, the desired result (4.2.2) is obtained. □

Corollary 4.2.4. *The set*

$$H = \{\mathbf{m}(t_f,\mathbf{p}):\ \dot{\mathbf{m}}(t,\mathbf{p}) = \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t,\mathbf{p})\mathbf{m}(t,\mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t,\mathbf{p}),\ \forall t\in(t_0,t],$$
$$\mathbf{G}_f(\cdot,\mathbf{p}) \in \mathcal{L}(T,\partial_{\mathbf{v}}\mathbf{f}(\cdot,\mathbf{v}(\cdot,\mathbf{p}))),$$
$$\mathbf{G}_f(t,\mathbf{p}) = [\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t,\mathbf{p})\ \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t,\mathbf{p})],\ \forall t\in T,$$
$$\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t,\mathbf{p}) \in \mathbb{R}^{n_x\times n_p}, \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t,\mathbf{p}) \in \mathbb{R}^{n_x\times n_x},\ \forall t\in T,$$
$$\mathbf{m}(t_0,\mathbf{p}) \in \partial\mathbf{f}_0(\bar{\mathbf{p}})\}$$

is a subset of $\text{conv}\,(\Gamma_{\mathbf{p}}\mathbf{x}(t_f,\mathbf{p}))$.

Proof. The result follows from the fact that $\partial_{\mathbf{v}}\mathbf{f}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p})) \subset \pi_{\mathbf{v}}\partial\mathbf{f}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p})) \subset \pi_{\mathbf{v}}\text{conv}\,(\Gamma\mathbf{f}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p})))$ for all $t\in[t_0,t_f]$ per Theorem 2.6.10. □

Remark 4.2.5. Let S be some measure zero subset of $[t_0,t_f]$. Note that if $\pi_{\mathbf{v}}\partial\mathbf{f}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}))$ is a singleton for all $t\in[t_0,t_f]\backslash S$, then it can be deduced that $\Gamma_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})$ is a singleton and therefore $\partial_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})$ is a singleton. This differs from the result in Theorem 3.2.3 which states $\partial_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})$ is a singleton if $\partial_{\mathbf{v}}\mathbf{f}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}))$ is a singleton for all $t\in[t_0,t_f]\backslash S$. In order to

recover the result in Theorem 3.2.3 in case $\partial_{\mathbf{v}}\mathbf{f}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}))$ is a singleton for all $t \in [t_0,t_f]\backslash S$, the result in Corollary 4.2.4 will be applied and $\partial_{\mathbf{v}}\mathbf{f}(\cdot,\mathbf{p},\mathbf{x}(t,\mathbf{p}))$ will be used to compute an element of $\Gamma_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})$.

Theorem 4.2.6. *Let the assumptions of Theorem 4.2.3 hold. Let $g : \mathcal{T}\times\mathcal{P}\times\mathcal{X} \to \mathbb{R}$ be a semismooth function. Let $G : \mathcal{P} \to \mathbb{R}$ be defined by*

$$G(\boldsymbol{\eta}) = \int_{t_0}^{t_f} g(t,\boldsymbol{\eta},\mathbf{x}(t,\boldsymbol{\eta}))dt.$$

Then, the mapping $\boldsymbol{\eta} \mapsto G(\boldsymbol{\eta})$ is semismooth at $\mathbf{p}$.

Let $\Gamma g : \mathcal{T}\times\mathcal{P}\times\mathcal{X} \rightrightarrows \mathbb{R}^{1\times(1+n_x+n_p)}$ be a linear Newton approximation of g such that $\partial g(\boldsymbol{\mu}) \subset \operatorname{conv}(\Gamma g(\boldsymbol{\mu}))$ for all $\boldsymbol{\mu} \in \mathcal{T}\times\mathcal{P}\times\mathcal{X}$. Let $\Gamma G : \mathcal{P} \rightrightarrows \mathbb{R}^{1\times n_p}$ be the set

$$\begin{aligned}
\{\mathbf{n}(t_f,\mathbf{p}) \in \mathbb{R}^{1\times n_p} : \ & \dot{\mathbf{n}}(t,\mathbf{p}) = \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_g(t,\mathbf{p})\mathbf{m}(t,\mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_g(t,\mathbf{p}),\ \forall t \in (t_0,t_f],\ \mathbf{n}(t_0,\mathbf{p}) = \mathbf{0},\\
& \mathbf{G}_g(\cdot,\mathbf{p}) \in \mathcal{L}(T,\pi_{\mathbf{v}}\operatorname{conv}(\Gamma g(\cdot,\mathbf{v}(\cdot,\mathbf{p})))),\\
& \mathbf{G}_g(t,\mathbf{p}) = [\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_g(t,\mathbf{p})\ \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_g(t,\mathbf{p})],\ \forall t \in T,\\
& \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_g(t,\mathbf{p}) \in \mathbb{R}^{1\times n_p}, \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t,\mathbf{p}) \in \mathbb{R}^{1\times n_x}, \forall t \in T,\\
& \dot{\mathbf{m}}(t,\mathbf{p}) = \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t,\mathbf{p})\mathbf{m}(t,\mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t,\mathbf{p}),\ \forall t \in (t_0,t_f],\\
& \mathbf{G}_f(\cdot,\mathbf{p}) \in \mathcal{L}(T,\pi_{\mathbf{v}}\operatorname{conv}(\Gamma\mathbf{f}(\cdot,\mathbf{v}(\cdot,\mathbf{p})))),\ \forall t \in T,\\
& \mathbf{G}_f(t,\mathbf{p}) = [\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t,\mathbf{p})\ \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t,\mathbf{p})],\ \forall t \in T,\\
& \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t,\mathbf{p}) \in \mathbb{R}^{n_x\times n_p}, \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t,\mathbf{p}) \in \mathbb{R}^{n_x\times n_x},\ \forall t \in T,\\
& \mathbf{m}(t_0,\mathbf{p}) \in \operatorname{conv}(\Gamma\mathbf{f}_0(\mathbf{p}))\}.
\end{aligned}$$

Then ΓG is a linear Newton approximation of G at $\mathbf{p}$ and $\partial G(\mathbf{p}) \subset \operatorname{conv}(\Gamma G(\mathbf{p}))$ holds.

Proof. Let $x_g : T\times\mathcal{P} \mapsto \mathbb{R}$ be a continuous function. Let $\boldsymbol{\nu}(t,\mathbf{p}) = (\mathbf{x}(t,\mathbf{p}), x_g(t,\mathbf{p}))$. Let $\mathbf{h} : \mathcal{T}\times\mathcal{P}\times\mathcal{X}\times\mathbb{R} \mapsto \dot{\mathcal{X}}\times\mathbb{R} : (\bar{t},\bar{\mathbf{p}},\bar{\mathbf{x}},\bar{x}_g) \mapsto (\mathbf{f}(\bar{t},\bar{\mathbf{p}},\bar{\mathbf{x}}), g(\bar{t},\bar{\mathbf{p}},\bar{\mathbf{x}}))$. Note that $\mathbf{h}$ is semismooth as a

composition of semismooth functions g and $\mathbf{f}$. Define $\tilde{\Gamma}\mathbf{f} : \mathcal{T} \times \mathcal{P} \times \mathcal{X} \times \mathbb{R} \rightrightarrows \mathbb{R}^{(n_x)\times(1+n_p+n_x+1)}$ by

$$\tilde{\Gamma}\mathbf{f}(\bar{t}, \bar{\mathbf{p}}, \bar{\mathbf{x}}, \bar{x}_g) = \{[\mathbf{N}\ \mathbf{0}] :\ \mathbf{N} \in \Gamma\mathbf{f}(\bar{t}, \bar{\mathbf{p}}, \bar{\mathbf{x}})\}.$$

Then, $\tilde{\Gamma}\mathbf{f}$ is a linear Newton approximation of the mapping $(\bar{t}, \bar{\mathbf{p}}, \bar{\mathbf{x}}, \bar{x}_g) \mapsto \mathbf{f}(\bar{t}, \bar{\mathbf{p}}, \bar{\mathbf{x}})$ per Theorem 2.8.12. Define $\tilde{\Gamma}g : \mathcal{T} \times \mathcal{P} \times \mathcal{X} \times \mathbb{R} \rightrightarrows \mathbb{R}^{1\times(1+n_p+n_x+1)}$ by

$$\tilde{\Gamma}g(\bar{t}, \bar{\mathbf{p}}, \bar{\mathbf{x}}, \bar{x}_g) = \{[\mathbf{N}\ 0] :\ \mathbf{N} \in \Gamma g(\bar{t}, \bar{\mathbf{p}}, \bar{\mathbf{x}})\}.$$

Then, $\tilde{\Gamma}g$ is a linear Newton approximation of the mapping $(\bar{t}, \bar{\mathbf{p}}, \bar{\mathbf{x}}, \bar{x}_g) \mapsto g(\bar{t}, \bar{\mathbf{p}}, \bar{\mathbf{x}})$ per Theorem 2.8.12. Let $\Gamma\mathbf{h} : \mathcal{T} \times \mathcal{P} \times \mathcal{X} \times \mathbb{R} \rightrightarrows \mathbb{R}^{(n_x+1)\times(1+n_p+n_x+1)}$ be

$$\Gamma\mathbf{h}(\bar{t}, \bar{\mathbf{p}}, \bar{\mathbf{x}}, \bar{x}_g) = \operatorname{conv}\left(\left\{\begin{bmatrix}\mathbf{M}\\ \mathbf{N}\end{bmatrix} : \mathbf{M} \in \tilde{\Gamma}\mathbf{f}(\bar{t}, \bar{\mathbf{p}}, \bar{\mathbf{x}}, \bar{x}_g),\ \mathbf{N} \in \tilde{\Gamma}g(\bar{t}, \bar{\mathbf{p}}, \bar{\mathbf{x}}, \bar{x}_g)\right\}\right).$$

Then $\Gamma\mathbf{h}$ is a linear Newton approximation of $\mathbf{h}$ per Theorem 2.8.12 considering the chain of mappings $(\boldsymbol{\mu}) \mapsto (\mathbf{f}(\boldsymbol{\mu}), \boldsymbol{\mu}) \mapsto (\mathbf{f}(\boldsymbol{\mu}), g(\boldsymbol{\mu}))$. Note that in this case

$$\Gamma\mathbf{h}(\bar{t}, \bar{\mathbf{p}}, \bar{\mathbf{x}}, \bar{x}_g) = \left\{\begin{bmatrix}\mathbf{M}\\ \mathbf{N}\end{bmatrix} : \mathbf{M} \in \operatorname{conv}\left(\tilde{\Gamma}\mathbf{f}(\bar{t}, \bar{\mathbf{p}}, \bar{\mathbf{x}}, \bar{x}_g)\right),\ \mathbf{N} \in \operatorname{conv}\left(\tilde{\Gamma}g(\bar{t}, \bar{\mathbf{p}}, \bar{\mathbf{x}}, \bar{x}_g)\right)\right\}.$$

holds.

Consider the augmented ordinary differential equation:

$$\dot{\boldsymbol{\nu}}(t, \mathbf{p}) = \mathbf{h}(t, \mathbf{p}, \boldsymbol{\nu}(t, \mathbf{p})), \forall t \in (t_0, t_f],\ \boldsymbol{\nu}(t_0, \mathbf{p}) = (\mathbf{f}_0(\mathbf{p}), 0).$$

The mapping $\boldsymbol{\eta} \mapsto \boldsymbol{\nu}(t, \boldsymbol{\eta})$ is semismooth for all $t \in [t_0, t_f]$ per Theorem 4.2.3. As a conse-

quence, the mapping $\boldsymbol{\eta} \mapsto G(\boldsymbol{\eta})$ is semismooth at $\mathbf{p}$ because $G(\boldsymbol{\eta}) = x_g(t_f, \boldsymbol{\eta})$.

Let $\mathbf{G}(\cdot, \mathbf{p}) \in \mathcal{L}(T, \pi_{\bar{\mathbf{v}}} \Gamma \mathbf{h}(\cdot, \mathbf{p}, \boldsymbol{\nu}(\cdot, \mathbf{p})))$, $\mathbf{G}(t, \mathbf{p}) = [\boldsymbol{\pi}_{\mathbf{p}} \mathbf{G}(t, \mathbf{p}) \ \boldsymbol{\pi}_{\bar{\mathbf{x}}} \mathbf{G}(t, \mathbf{p})]$, $\boldsymbol{\pi}_{\mathbf{p}} \mathbf{G}(t, \mathbf{p}) \in \mathbb{R}^{(n_x+1) \times n_p}$, $\boldsymbol{\pi}_{\bar{\mathbf{x}}} \mathbf{G}(t, \mathbf{p}) \in \mathbb{R}^{(n_x+1) \times (n_x+1)}$ for all $t \in T$. Then, $\tilde{\mathbf{M}}(t, \mathbf{p})$, an element of $\Gamma_{\mathbf{p}} \boldsymbol{\nu}(t_f, \mathbf{p})$, can be computed by

$$\dot{\tilde{\mathbf{M}}}(t, \mathbf{p}) = \boldsymbol{\pi}_{\bar{\mathbf{x}}} \mathbf{G}(t, \mathbf{p}) \tilde{\mathbf{M}}(t, \mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}} \mathbf{G}(t, \mathbf{p}), \forall t \in (t_0, t_f], \ \tilde{\mathbf{M}}(t_0, \mathbf{p}) = \begin{bmatrix} \mathbf{N} \\ \mathbf{0} \end{bmatrix}, \ \mathbf{N} \in \operatorname{conv}(\Gamma \mathbf{f}_0(\mathbf{p}))$$

per Theorem 4.2.3. This equation can be written as

$$\begin{bmatrix} \dot{\mathbf{m}}(t, \mathbf{p}) \\ \dot{\mathbf{n}}(t, \mathbf{p}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\pi}_{\mathbf{x}} \mathbf{G}_f(t, \mathbf{p}) & \mathbf{0} \\ \boldsymbol{\pi}_{\mathbf{x}} \mathbf{G}_g(t, \mathbf{p}) & 0 \end{bmatrix} \begin{bmatrix} \mathbf{m}(t, \mathbf{p}) \\ \mathbf{n}(t, \mathbf{p}) \end{bmatrix} + \begin{bmatrix} \boldsymbol{\pi}_{\mathbf{p}} \mathbf{G}_f(t, \mathbf{p}) \\ \boldsymbol{\pi}_{\mathbf{p}} \mathbf{G}_g(t, \mathbf{p}) \end{bmatrix}, \forall t \in (t_0, t_f],$$
$$\mathbf{m}(t_0, \mathbf{p}) = \mathbf{N}, \ \mathbf{n}(t_0, \mathbf{p}) = \mathbf{0}, \ \mathbf{N} \in \operatorname{conv}(\Gamma \mathbf{f}_0(\mathbf{p})).$$

where $\mathbf{G}_f$ and $\mathbf{G}_g$ represent measurable selections of $\pi_{\bar{\mathbf{x}}} \operatorname{conv}(\Gamma \mathbf{f}(\cdot, \mathbf{p}, \mathbf{x}(\cdot, \mathbf{p})))$ and $\pi_{\mathbf{v}} \operatorname{conv}(\Gamma g(\cdot, \mathbf{p}, \mathbf{x}(\cdot, \mathbf{p})))$, respectively, on $[t_0, t_f]$. Due to the zero column in $\boldsymbol{\pi}_{\bar{\mathbf{v}}} \mathbf{G}(t, \mathbf{p})$, $\mathbf{n}(t, \mathbf{p})$ does not enter the computation of $\dot{\mathbf{m}}(\epsilon, \mathbf{p})$ and $\dot{\mathbf{n}}(\epsilon, \mathbf{p})$. Therefore, this equation can be rewritten as

$$\dot{\mathbf{m}}(t, \mathbf{p}) = \boldsymbol{\pi}_{\mathbf{x}} \mathbf{G}_f(t, \mathbf{p}) \mathbf{m}(t, \mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}} \mathbf{G}_f(t, \mathbf{p}), \forall t \in (t_0, t_f], \ \mathbf{m}(t_0, \mathbf{p}) \in \operatorname{conv}(\Gamma \mathbf{f}_0(\mathbf{p})), \quad (4.2.13)$$

$$\dot{\mathbf{n}}(t, \mathbf{p}) = \boldsymbol{\pi}_{\mathbf{x}} \mathbf{G}_g(t, \mathbf{p}) \mathbf{m}(t, \mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}} \mathbf{G}_g(t, \mathbf{p}), \ \forall t \in (t_0, t_f], \ \mathbf{n}(t_0, \mathbf{p}) = \mathbf{0}. \quad (4.2.14)$$

Note that $\partial_{\mathbf{p}} \boldsymbol{\nu}(t, \mathbf{p}) \subset \operatorname{conv}(\Gamma_{\mathbf{p}} \boldsymbol{\nu}(t, \mathbf{p}))$ per Theorem 4.2.3. Let $\mathbf{B} \in \mathbb{R}^{1 \times (n_x+1)}$ and let $B_{ij} = 0$ for $i = 1$ for all $j \in \{1, \ldots, n_x\}$. Let $B_{ij} = 1$ if $j = n_x + 1$ and $i = 1$. Then

$$\partial_{\mathbf{p}} x_g(t, \mathbf{p}) \subset \operatorname{conv}(\{\mathbf{BN} : \ \mathbf{N} \in \partial_{\mathbf{p}} \boldsymbol{\nu}(t, \mathbf{p})\}) \subset \operatorname{conv}(\{\mathbf{BN} : \ \mathbf{N} \in \Gamma_{\mathbf{p}} \boldsymbol{\nu}(t, \mathbf{p})\}). \quad (4.2.15)$$

The set $\{\mathbf{BN} : \mathbf{N} \in \Gamma_{\mathbf{p}}\boldsymbol{\nu}(t,\mathbf{p})\}$ corresponds to the set of all $\mathbf{n}(t,\mathbf{p})$ computed using (4.2.14) which corresponds to the set $\Gamma_{\mathbf{p}}x_g(t_f,\mathbf{p})$, a linear Newton approximation of the mapping $\boldsymbol{\eta} \mapsto x_g(t_f,\boldsymbol{\eta})$. Hence $\partial_{\mathbf{p}}x_g(t_f,\mathbf{p}) \subset \text{conv}\,(\Gamma_{\mathbf{p}}x_g(t_f,\mathbf{p}))$ and $\partial_{\mathbf{p}}x_g(t_f,\mathbf{p}) \subset \text{conv}\,(\Gamma G(\mathbf{p}))$ follows since $x_g(t_f,\mathbf{p}) = G(\mathbf{p})$. □

The next two theorems contain results analogous to Theorems 3.2.4 and 3.2.5.

Theorem 4.2.7. *Let the hypotheses and definitions of Theorem 4.2.3 and Theorem 4.2.6 hold. Let $\boldsymbol{\lambda} : T \to \mathbb{R}^{1\times n_x}$ be a solution of the initial value problem,*

$$\dot{\boldsymbol{\lambda}}(t) = -\boldsymbol{\lambda}(t)\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t,\mathbf{p}) + \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_g(t,\mathbf{p}),\ \forall t \in [t_0,t_f),\ \boldsymbol{\lambda}(t_f) = \mathbf{0}. \tag{4.2.16}$$

Then, $\boldsymbol{\lambda}$ is unique and absolutely continuous. Let $\mathbf{J} \in \mathbb{R}^{1\times n_p}$ be defined by

$$\mathbf{J} = \int_{t_0}^{t_f} \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_g(t,\mathbf{p}) - \boldsymbol{\lambda}(t)\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t,\mathbf{p})dt + \boldsymbol{\lambda}(t)\mathbf{m}(t,\mathbf{p})|_{t_0}^{t_f} \tag{4.2.17}$$

where $\mathbf{m}(t,\mathbf{p})$ is computed using (4.2.2). Then $\mathbf{J} \in \Gamma G(\mathbf{p})$.

Proof. The measurability of $\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(\cdot,\mathbf{p})$ and $\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_g(\cdot,\mathbf{p})$ on $[t_0,t_f]$ follows from assumptions. Their boundedness follows from the fact that $\Gamma\mathbf{f}$ and $\Gamma\mathbf{g}$ are bounded and upper semicontinuous set-valued mappings. Then $\boldsymbol{\lambda} : [t_0,t_f] \to \mathbb{R}^{1\times n_x}$ is the unique and absolutely continuous solution of (4.2.16) in the sense of Carathéodory per Theorem 3 in [37].

Redefine $\dot{\mathbf{n}}(t,\mathbf{p})$ in (4.2.14) by appending (4.2.13) to (4.2.14) using $\boldsymbol{\lambda}(t)$ to obtain

$$\begin{aligned}\dot{\mathbf{n}}(t,\mathbf{p}) = {} & \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_g(t,\mathbf{p})\mathbf{m}(t,\mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_g(t,\mathbf{p}) - \\ & \boldsymbol{\lambda}(t)(\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t,\mathbf{p})\mathbf{m}(t,\mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t,\mathbf{p}) - \dot{\mathbf{m}}(t,\mathbf{p})),\ \forall t \in (t_0,t_f], \\ & \mathbf{n}(t_0,\mathbf{p}) = \mathbf{0}.\end{aligned} \tag{4.2.18}$$

Equation (4.2.18) can be written in integral form as

$$\mathbf{n}(t_f, \mathbf{p}) = \int_{t_0}^{t_f} \boldsymbol{\pi}_{\mathbf{x}} \mathbf{G}_g(t, \mathbf{p})\mathbf{m}(t, \mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}} \mathbf{G}_g(t, \mathbf{p}) - \tag{4.2.19}$$
$$\boldsymbol{\lambda}(t)(\boldsymbol{\pi}_{\mathbf{x}} \mathbf{G}_f(t, \mathbf{p})\mathbf{m}(t, \mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}} \mathbf{G}_f(t, \mathbf{p}) - \dot{\mathbf{m}}(t, \mathbf{p}))dt.$$

Since $\boldsymbol{\lambda}$ and $\mathbf{m}(\cdot, \mathbf{p})$ are measurable functions on $[t_0, t_f]$, the application of integration by parts provides the relation

$$\int_{t_0}^{t_f} \boldsymbol{\lambda}(t)\dot{\mathbf{m}}(t, \mathbf{p})dt = \boldsymbol{\lambda}(t)\mathbf{m}(t, \mathbf{p})|_{t_0}^{t_f} - \int_{t_0}^{t_f} \dot{\boldsymbol{\lambda}}(t)\mathbf{m}(t, \mathbf{p})dt. \tag{4.2.20}$$

Using the relation (4.2.20) in (4.2.19) and collecting terms multiplying $\mathbf{m}(t, \mathbf{p})$ results in

$$\mathbf{n}(t_f, \mathbf{p}) = \int_{t_0}^{t_f} (\boldsymbol{\pi}_{\mathbf{x}} \mathbf{G}_g(t, \mathbf{p}) - \boldsymbol{\lambda}(t)\boldsymbol{\pi}_{\mathbf{x}} \mathbf{G}_f(t, \mathbf{p}) - \dot{\boldsymbol{\lambda}}(t))\mathbf{m}(t, \mathbf{p}) + \tag{4.2.21}$$
$$\boldsymbol{\pi}_{\mathbf{p}} \mathbf{G}_g(t, \mathbf{p}) - \boldsymbol{\lambda}(t)\boldsymbol{\pi}_{\mathbf{p}} \mathbf{G}_f(t, \mathbf{p})dt + \boldsymbol{\lambda}(t)\mathbf{m}(t, \mathbf{p})|_{t_0}^{t_f}.$$

The desired result is obtained after substituting the expression for $\dot{\boldsymbol{\lambda}}$ in (4.2.16) into (4.2.21).

□

Theorem 4.2.8. *Let the hypotheses and definitions of Theorem 4.2.3 hold. Let* $h : \mathcal{T}_0 \times \mathcal{P} \times \mathcal{X} \to \mathbb{R}$ *be a semismooth function where* $\mathcal{T}_0$ *is an open subset of* $\mathcal{T}$ *such that* $t_f \in \mathcal{T}_0$. *Let* $G : \mathcal{P} \to \mathbb{R} : \boldsymbol{\eta} \mapsto h(t_f, \boldsymbol{\eta}, \mathbf{x}(t_f, \boldsymbol{\eta}))$. *Then* G *is semismooth at* $\mathbf{p}$.

Let $\Gamma_{\mathbf{v}} h : \mathcal{P} \times \mathcal{X} \rightrightarrows \mathbb{R}^{1 \times (n_p + n_x)}$ *be a linear Newton approximation of the map* $(\bar{\mathbf{p}}, \bar{\mathbf{x}}) \mapsto h(t_f, \bar{\mathbf{p}}, \bar{\mathbf{x}})$ *such that* $\partial_{\mathbf{v}} h(t_f, \bar{\mathbf{p}}, \bar{\mathbf{x}}) \subset \operatorname{conv}(\Gamma_{\mathbf{v}} h(t_f, \bar{\mathbf{p}}, \bar{\mathbf{x}}))$ *in a neighborhood of* $(\mathbf{p}, \mathbf{x}(t_f, \mathbf{p}))$.

Let $\Gamma G : \mathbf{p} \rightrightarrows \mathbb{R}^{1+n_p}$ *be the set*

$$\{\mathbf{C}_{\mathbf{x}}\mathbf{N} + \mathbf{C}_{\mathbf{p}} : [\mathbf{C}_{\mathbf{p}}, \mathbf{C}_{\mathbf{x}}] \in \operatorname{conv}(\Gamma_{\mathbf{v}} h(t_f, \mathbf{v}(t_f, \mathbf{p}))),\ \mathbf{C}_{\mathbf{p}} \in \mathbb{R}^{1 \times n_p},$$
$$\mathbf{C}_{\mathbf{x}} \in \mathbb{R}^{1 \times n_x},\ \mathbf{N} \in \operatorname{conv}(\Gamma_{\mathbf{p}} \mathbf{x}(t_f, \mathbf{p}))\}.$$

Then, ΓG *is a linear Newton approximation of* G *at* $\mathbf{p}$ *per Theorem 2.8.12 such that* $\partial G(\mathbf{p}) \subset \operatorname{conv}(\Gamma G(\mathbf{p}))$ *holds. Let* $[\mathbf{C_p}, \mathbf{C_x}] \in \operatorname{conv}(\Gamma_{\mathbf{v}} h(t_f, \mathbf{v}(t_f, \mathbf{p})))$, $\mathbf{C_p} \in \mathbb{R}^{1 \times n_p}$, $\mathbf{C_x} \in \mathbb{R}^{1 \times n_x}$. *Let* $\boldsymbol{\lambda} : T \to \mathbb{R}^{1 \times n_x}$ *be a solution of the initial value problem,*

$$\dot{\boldsymbol{\lambda}}(t) = -\boldsymbol{\lambda}(t)\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t, \mathbf{p}),\ \forall t \in [t_0, t_f],\ \boldsymbol{\lambda}(t_f) = -\mathbf{C_x}. \tag{4.2.22}$$

Then, $\boldsymbol{\lambda}$ *is unique and absolutely continuous. Let* $\mathbf{J} \in \mathbb{R}^{1 \times n_p}$ *be defined by*

$$\mathbf{J} = \int_{t_0}^{t_f} -\boldsymbol{\lambda}(t)\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t, \mathbf{p})dt - \boldsymbol{\lambda}(t_0)\mathbf{m}(t_0, \mathbf{p}) + \mathbf{C_p} \tag{4.2.23}$$

where $\mathbf{m}(t, \mathbf{p})$ *is computed using* (4.2.2). *Then* $\mathbf{J} \in \Gamma G(\mathbf{p})$.

Proof. The semismoothness of G follows from the fact that the composition of semismooth functions is semismooth.

The measurability of $\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(\cdot, \mathbf{p})$ on $[t_0, t_f]$ follows from assumptions. Its boundedness follows from the fact that $\Gamma\mathbf{f}$ is a bounded and upper semicontinuous set-valued mapping. Then, $\boldsymbol{\lambda} : [t_0, t_f] \to \mathbb{R}^{1 \times n_x}$ is the unique and absolutely continuous solution of (4.2.22) in the sense of Carathéodory per Theorem 3 in [37].

The generalized Jacobian of G at $\mathbf{p}$ satisfies $\partial G(\mathbf{p}) \subset \operatorname{conv}(S)$ where S is the set

$$\{\mathbf{A_x}\mathbf{B} + \mathbf{A_p} : [\mathbf{A_p}, \mathbf{A_x}] \in \partial_{\mathbf{v}} h(t_f, \mathbf{v}(t_f, \mathbf{p})),\ \mathbf{A_p} \in \mathbb{R}^{1 \times n_p},\ \mathbf{A_x} \in \mathbb{R}^{1 \times n_x},\ \mathbf{B} \in \partial_{\mathbf{p}}\mathbf{x}(t_f, \mathbf{p})\},$$

due to Theorem 2.6.7. Note that, $\partial G(\mathbf{p}) \subset \operatorname{conv}(S) \subset \operatorname{conv}(\Gamma G(\mathbf{p}))$ because $\partial_{\mathbf{p}}\mathbf{x}(t_f, \mathbf{p}) \subset \operatorname{conv}(\Gamma_{\mathbf{p}}\mathbf{x}(t_f, \mathbf{p}))$ and $\partial_{\mathbf{v}} h(t_f, \mathbf{v}(t_f, \mathbf{p})) \subset \operatorname{conv}(\Gamma_{\mathbf{v}} h(t_f, \mathbf{v}(t_f, \mathbf{p})))$.

The product $\mathbf{C_x}\mathbf{N}$ can be written as

$$\int_{t_0}^{t_f} \mathbf{C_x}\dot{\mathbf{m}}(t, \mathbf{p})dt + \mathbf{C_x}\mathbf{m}(t_0, \mathbf{p})$$

where $\mathbf{m}(t, \mathbf{p})$ is computed using (4.2.2). Appending (4.2.13) to the integrand of this integral produces

$$\int_{t_0}^{t_f} \mathbf{C_x}\dot{\mathbf{m}}(t, \mathbf{p}) - \boldsymbol{\lambda}(t)(\boldsymbol{\pi}_\mathbf{x}\mathbf{G}_f(t, \mathbf{p})\mathbf{m}(t, \mathbf{p}) + \boldsymbol{\pi}_\mathbf{p}\mathbf{G}_f(t, \mathbf{p}) - \dot{\mathbf{m}}(t, \mathbf{p}))dt + \mathbf{C_x}\mathbf{m}(t_0, \mathbf{p}). \tag{4.2.24}$$

Since $\boldsymbol{\lambda}$ and $\mathbf{m}(\cdot, \mathbf{p})$ are measurable functions on $[t_0, t_f]$, the application of integration by parts provides the relation

$$\int_{t_0}^{t_f} (\mathbf{C_x} + \boldsymbol{\lambda}(t))\dot{\mathbf{m}}(t, \mathbf{p})dt = (\mathbf{C_x} + \boldsymbol{\lambda}(t))\mathbf{m}(t, \mathbf{p})|_{t_0}^{t_f} - \int_{t_0}^{t_f} \dot{\boldsymbol{\lambda}}(t)\mathbf{m}(t, \mathbf{p})dt. \tag{4.2.25}$$

Using the relation (4.2.25) in (4.2.24) and collecting terms multiplying $\mathbf{m}(t, \mathbf{p})$ results in

$$\int_{t_0}^{t_f} (-\boldsymbol{\lambda}(t)\boldsymbol{\pi}_\mathbf{x}\mathbf{G}_f(t, \mathbf{p}) - \dot{\boldsymbol{\lambda}}(t))\mathbf{m}(t, \mathbf{p}) - \boldsymbol{\lambda}(t)\boldsymbol{\pi}_\mathbf{p}\mathbf{G}_f(t, \mathbf{p})dt + \mathbf{C_x}\mathbf{m}(t_0, \mathbf{p}) + (\mathbf{C_x} + \boldsymbol{\lambda}(t))\mathbf{m}(t, \mathbf{p})|_{t_0}^{t_f}.$$

Substituting the expression for $\dot{\boldsymbol{\lambda}}$ provides the desired result. □

4.3 Differential-Algebraic Equations

The differential-algebraic equations considered in this section satisfy the following assumptions.

Assumption 4.3.1. *Let* $\mathbf{F} : \mathcal{T} \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}} \to \mathbb{R}^{n_x+n_y}$ *and* $\mathbf{F}_0 : \mathcal{P} \to \mathcal{X}$ *be semismooth functions. Let* $\mathbf{x} : T \times \mathcal{P} \to \mathcal{X}$, $\mathbf{y} : T \times \mathcal{P} \to \mathcal{Y}$ *and* $\dot{\mathbf{x}} : T \times \mathcal{P} \to \dot{\mathcal{X}}$ *be such that they uniquely*

satisfy the initial value problem

$$\mathbf{0} = \mathbf{F}(t, \mathbf{p}, \mathbf{x}(t,\mathbf{p}), \mathbf{y}(t,\mathbf{p}), \dot{\mathbf{x}}(t,\mathbf{p})),\ \forall t \in [t_0, t_f],\ \mathbf{x}(t_0,\mathbf{p}) = \mathbf{F}_0(\mathbf{p}),\ \forall \mathbf{p} \in \mathcal{P}. \tag{4.3.1}$$

Let $\Gamma\mathbf{F} : \mathcal{T}\times\mathcal{P}\times\mathcal{X}\times\mathcal{Y}\times\dot{\mathcal{X}} \rightrightarrows \mathbb{R}^{(n_x+n_y)\times(1+n_p+n_x+n_y+n_x)}$ *be a linear Newton approximation of* $\mathbf{F}$ *such that* $\partial\mathbf{F}(\boldsymbol{\mu}) \subset \operatorname{conv}(\Gamma\mathbf{F}(\boldsymbol{\mu}))$ *holds for all* $\boldsymbol{\mu} \in \mathcal{T}\times\mathcal{P}\times\mathcal{X}\times\mathcal{Y}\times\dot{\mathcal{X}}$. *Let* $\Gamma\mathbf{F}_0 : \mathcal{P} \rightrightarrows \mathbb{R}^{n_x}$ *be a linear Newton approximation of* $\mathbf{F}_0$ *such that* $\partial\mathbf{F}_0(\boldsymbol{\eta}) \subset \operatorname{conv}(\Gamma\mathbf{F}_0(\boldsymbol{\eta}))$ *holds for all* $\boldsymbol{\eta} \in \mathcal{P}$.

Let $\dot{\mathbf{x}}(t_0,\bar{\mathbf{p}}) = \dot{\bar{\mathbf{x}}}$ *and* $\mathbf{y}(t_0,\bar{\mathbf{p}}) = \bar{\mathbf{y}}$ *for some* $\bar{\mathbf{p}} \in \mathcal{P}$ *where* $\dot{\bar{\mathbf{x}}}$ *and* $\bar{\mathbf{y}}$ *are constants and satisfy* $\mathbf{F}(t_0,\bar{\mathbf{p}},\mathbf{x}(t_0,\bar{\mathbf{p}}),\mathbf{y}(t_0,\bar{\mathbf{p}}),\dot{\mathbf{x}}(t_0,\bar{\mathbf{p}})) = \mathbf{0}$. *Assume that this condition uniquely determines* $\mathbf{y}(t_0,\mathbf{p})$ *and* $\dot{\mathbf{x}}(t_0,\mathbf{p})$ *for all* $\mathbf{p} \in \mathcal{P}$.

In order to derive implicit functions for $\dot{\mathbf{x}}$ and $\mathbf{y}$, the following additional assumptions are made.

Assumption 4.3.2. *In addition to Assumption 4.3.1, let the following hold*

1. $\pi_{\mathbf{q}}\partial^B\mathbf{F}(\boldsymbol{\mu})$ *is coherently oriented, i.e., the determinants of all matrices in* $\pi_{\mathbf{q}}\partial^B\mathbf{F}(\boldsymbol{\mu})$ *have the same nonzero sign,* α, *for all* $\boldsymbol{\mu} \in \mathcal{T}\times\mathcal{P}\times\mathcal{X}\times\mathcal{Y}\times\dot{\mathcal{X}}$.
2. *Let* $\boldsymbol{\mu}^* = (t^*, \mathbf{v}^*, \mathbf{q}^*)$ *where* $t^* \in \mathcal{T}$, $\mathbf{v}^* \in \mathcal{P}\times\mathcal{X}$ *and* $\mathbf{q}^* \in \mathcal{Y}\times\dot{\mathcal{X}}$. *Let* $\mathbf{h}_{\boldsymbol{\mu}} : \dot{\mathcal{X}}\times\mathcal{Y} \to \mathbb{R}^{n_x+n_y} : \mathbf{q} \mapsto \mathbf{F}(t^*, \mathbf{v}^*, \mathbf{q})$. *If* $\mathbf{F}(\boldsymbol{\mu}^*) = \mathbf{0}$, *then* $\operatorname{index}(\mathbf{h}_{\boldsymbol{\mu}}, \mathbf{q}^*) = \alpha$.

Remark 4.3.3. Let $\mathbf{u} : T\times\mathcal{P} \to \mathcal{P}\times\mathcal{X}\times\mathcal{Y}\times\dot{\mathcal{X}} : (t,\mathbf{p}) \mapsto (\mathbf{v}(t,\mathbf{p}), \mathbf{y}(t,\mathbf{p}), \dot{\mathbf{x}}(t,\mathbf{p}))$ for the remainder of this chapter.

Lemma 4.3.4. *Let* $\boldsymbol{\eta}_{\mathbf{p}} \in \mathcal{P}$, $\boldsymbol{\eta}_{\mathbf{x}} \in \mathcal{X}$, $\boldsymbol{\eta}_{\mathbf{y}} \in \mathcal{Y}$, $\boldsymbol{\eta}_{\dot{\mathbf{x}}} \in \dot{\mathcal{X}}$ *and* $\boldsymbol{\eta} = (\boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}}, \boldsymbol{\eta}_{\mathbf{y}}, \boldsymbol{\eta}_{\dot{\mathbf{x}}})$. *Let Assumptions 4.3.1 and 4.3.2 hold. Then, there exist semismooth functions* $\mathbf{f} : \mathcal{T}\times\mathcal{P}\times\mathcal{X} \to \dot{\mathcal{X}}$ *and* $\mathbf{r} : \mathcal{T}\times\mathcal{P}\times\mathcal{X} \to \mathcal{Y}$ *such that* $\mathbf{0} = \mathbf{F}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}}, \mathbf{r}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}}), \mathbf{f}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}}))$ *holds. Let*

$$M = \operatorname{conv}\left(\{-\mathbf{B}^{-1}\mathbf{A},\ [\mathbf{A}\ \mathbf{B}] \in \partial^B\mathbf{F}(t,\boldsymbol{\eta}), \mathbf{A} \in \mathbb{R}^{(n_y+n_x)\times(1+n_p+n_x)},\ \mathbf{B} \in \mathbb{R}^{(n_y+n_x)\times(n_y+n_x)}\}\right).$$

Then,

$$\Gamma\mathbf{r}(t,\boldsymbol{\eta}_\mathbf{p},\boldsymbol{\eta}_\mathbf{x}) = \text{conv}\left(\left\{\begin{bmatrix}\mathbf{I}_{n_y} & \mathbf{0}\end{bmatrix}\mathbf{M} : \mathbf{M}\in M\right\}\right),$$
$$\Gamma\mathbf{f}(t,\boldsymbol{\eta}_\mathbf{p},\boldsymbol{\eta}_\mathbf{x}) = \text{conv}\left(\left\{\begin{bmatrix}\mathbf{0} & \mathbf{I}_{n_x}\end{bmatrix}\mathbf{M} : \mathbf{M}\in M\right\}\right),$$

are linear Newton approximations of $\mathbf{f}$ *and* $\mathbf{r}$ *such that* $\partial\mathbf{r}(t,\boldsymbol{\eta}_\mathbf{p},\boldsymbol{\eta}_\mathbf{x}) \subset \text{conv}\,(\Gamma\mathbf{r}(t,\boldsymbol{\eta}_\mathbf{p},\boldsymbol{\eta}_\mathbf{x}))$ *and* $\partial\mathbf{f}(t,\boldsymbol{\eta}_\mathbf{p},\boldsymbol{\eta}_\mathbf{x}) \subset \text{conv}\,(\Gamma\mathbf{f}(t,\boldsymbol{\eta}_\mathbf{p},\boldsymbol{\eta}_\mathbf{x}))$ *hold for all* $(t,\boldsymbol{\eta}_\mathbf{p},\boldsymbol{\eta}_\mathbf{x}) \in \mathcal{T}\times\mathcal{P}\times\mathcal{X}$.

Proof. The set M is a linear Newton approximation of the mapping $(t,\boldsymbol{\eta}_\mathbf{p},\boldsymbol{\eta}_\mathbf{x}) \mapsto (\mathbf{r}(t,\boldsymbol{\eta}_\mathbf{p},\boldsymbol{\eta}_\mathbf{x}),$ $\mathbf{f}(t,\boldsymbol{\eta}_\mathbf{p},\boldsymbol{\eta}_\mathbf{x}))$ per Theorem 4.1.7 such that the generalized Jacobian of this map is a subset of M. Then, the results follow from the application of Theorem 2.8.12 to M and Theorem 2.6.7 to the generalized Jacobian of the mapping $(t,\boldsymbol{\eta}_\mathbf{p},\boldsymbol{\eta}_\mathbf{x}) \mapsto (\mathbf{r}(t,\boldsymbol{\eta}_\mathbf{p},\boldsymbol{\eta}_\mathbf{x}),\mathbf{f}(t,\boldsymbol{\eta}_\mathbf{p},\boldsymbol{\eta}_\mathbf{x}))$ □

Corollary 4.3.5. *Let Assumptions 4.3.1 and 4.3.2 hold. Let* $\mathbf{u}(\cdot,\bar{\mathbf{p}})$ *be formed from the unique solution of* (4.3.1). *Then* $\mathbf{u}(t,\cdot)$ *is a semismooth function at* $\bar{\mathbf{p}}\in\mathcal{P}$ *for all* $t\in[t_0,t_f]$.

Proof. Since the implicit function, $\mathbf{f}$ as defined in Lemma 4.3.4 is a semismooth function, $\mathbf{v}(t,\cdot)$ is a semismooth function at $\bar{\mathbf{p}}$ for all $t\in T$ per Theorem 4.2.3. The semismoothness of $\mathbf{y}(t,\cdot)$ at $\bar{\mathbf{p}}$ follows from the semismoothness of the implicit function $\mathbf{r}(t,\cdot)$ at $\mathbf{v}(t,\bar{\mathbf{p}})$ for all $t\in T$ and the semismoothness of $\mathbf{v}(t,\cdot)$ at $\bar{\mathbf{p}}$ for all $t\in T$. The semismoothness of $\dot{\mathbf{x}}(t,\cdot)$ at $\mathbf{p}$ follows from the same reasoning using $\mathbf{f}$ instead of $\mathbf{r}$. Since all elements of $\mathbf{u}(t,\cdot)$ are semismooth at $\bar{\mathbf{p}}$ so is $\mathbf{u}(t,\cdot)$ for all $t\in T$. □

Theorem 4.3.6. *Let Assumptions 4.3.1 and 4.3.2 hold. Let* $\Gamma\mathbf{f} : T\times\mathcal{P}\times\mathcal{X} \rightrightarrows \mathbb{R}^{n_x\times(1+n_p+n_x)}$ *be as defined in Lemma 4.3.4. Let* $\Gamma_\mathbf{p}\mathbf{x}(t_f,\mathbf{p})$ *be the set*

$$\{\mathbf{m}(t_f,\mathbf{p}) : \dot{\mathbf{m}}(t,\mathbf{p}) = \boldsymbol{\pi}_\mathbf{x}\mathbf{G}_f(t,\mathbf{p})\mathbf{m}(t,\mathbf{p}) + \boldsymbol{\pi}_\mathbf{p}\mathbf{G}_f(t,\mathbf{p}),\ \forall t\in(t_0,t_f], \tag{4.3.2}$$
$$\mathbf{G}_f(\cdot,\mathbf{p}) \in \mathcal{L}(T,\pi_\mathbf{v}\text{conv}\,(\Gamma\mathbf{f}(\cdot,\mathbf{v}(\cdot,\mathbf{p})))),$$
$$\mathbf{G}_f(t,\mathbf{p}) = [\boldsymbol{\pi}_\mathbf{p}\mathbf{G}_f(t,\mathbf{p})\ \boldsymbol{\pi}_\mathbf{x}\mathbf{G}_f(t,\mathbf{p})], \forall t\in T,$$

$$\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t,\mathbf{p}) \in \mathbb{R}^{n_x \times n_p}, \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t,\mathbf{p}) \in \mathbb{R}^{n_x \times n_x}, \ \forall t \in T,$$

$$\mathbf{m}(t_0,\mathbf{p}) \in \text{conv}\,(\Gamma\mathbf{f}_0(\mathbf{p}))\},$$

where $\Gamma\mathbf{f}$ *is defined in Lemma 4.3.4. Then* $\Gamma_{\mathbf{p}}\mathbf{x}(t_f,\mathbf{p})$ *is a linear Newton approximation of the map* $\boldsymbol{\eta} \mapsto \mathbf{x}(t_f,\boldsymbol{\eta})$ *at* $\mathbf{p}$ *and*

$$\partial_{\mathbf{p}}\mathbf{x}(t_f,\mathbf{p}) \subset \text{conv}\,(\Gamma_{\mathbf{p}}\mathbf{x}(t_f,\mathbf{p}))$$

holds.

Proof. Result follows from Theorem 4.2.3, Lemma 4.3.4. □

Corollary 4.3.7. *Let Assumptions 4.3.1 and 4.3.2 hold.*

1. *Let* $\Gamma_{\mathbf{p}}\mathbf{y} : \mathcal{T} \times \mathcal{P} \rightrightarrows \mathbb{R}^{n_y \times n_p}$ *be defined by*

$$\begin{aligned}\Gamma_{\mathbf{p}}\mathbf{y}(t,\mathbf{p}) = \{\mathbf{n}(t,\mathbf{p}) :\ & \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_r(t,\mathbf{p})\mathbf{m}(t,\mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_r(t,\mathbf{p}),\\ & \mathbf{G}_r(t,\mathbf{p}) \in \text{conv}\,(\pi_{\mathbf{v}}\Gamma\mathbf{r}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}))),\\ & \mathbf{G}_r(t,\mathbf{p}) = [\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_r(t,\mathbf{p})\ \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_r(t,\mathbf{p})],\\ & \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_r(t,\mathbf{p}) \in \mathbb{R}^{n_y \times n_p},\\ & \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_r(t,\mathbf{p}) \in \mathbb{R}^{n_y \times n_x},\\ & \mathbf{m}(t,\mathbf{p}) \in \text{conv}\,(\Gamma_{\mathbf{p}}\mathbf{x}(t,\mathbf{p}))\},\ \forall t \in [t_0,t_f],\end{aligned}$$

   where $\Gamma_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})$ *is computed using* (4.3.2) *and* $\Gamma\mathbf{r}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}))$ *is defined in Lemma 4.3.4. Then,* $\Gamma_{\mathbf{p}}\mathbf{y}(t,\mathbf{p})$ *is a linear Newton approximation of the map* $\boldsymbol{\eta} \mapsto \mathbf{y}(t,\boldsymbol{\eta})$ *at* $\mathbf{p} \in \mathcal{P}$ *and* $\partial_{\mathbf{p}}\mathbf{y}(t,\mathbf{p}) \subset \text{conv}\,(\Gamma_{\mathbf{p}}\mathbf{y}(t,\mathbf{p}))$ *holds.*

2. *Let* $\Gamma_{\mathbf{p}}\dot{\mathbf{x}} : \mathcal{T} \times \mathcal{P} \rightrightarrows \mathbb{R}^{n_y \times n_p}$ *be defined by*

$$\Gamma_{\mathbf{p}}\dot{\mathbf{x}}(t,\mathbf{p}) = \{\mathbf{n}(t,\mathbf{p}) :\ \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t,\mathbf{p})\mathbf{m}(t,\mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t,\mathbf{p}),$$

$$\mathbf{G}_f(t,\mathbf{p}) \in \text{conv}\,(\pi_{\mathbf{v}}\Gamma\mathbf{f}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}))),$$

$$\mathbf{G}_f(t,\mathbf{p}) = [\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t,\mathbf{p})\ \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t,\mathbf{p})],$$

$$\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t,\mathbf{p}) \in \mathbb{R}^{n_x \times n_p},$$

$$\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t,\mathbf{p}) \in \mathbb{R}^{n_x \times n_x},$$

$$\mathbf{m}(t,\mathbf{p}) \in \text{conv}\,(\Gamma_{\mathbf{p}}\mathbf{x}(t,\mathbf{p}))\},\ \forall t \in [t_0,t_f],$$

where $\Gamma_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})$ *is computed using* (4.3.2) *and* $\Gamma\mathbf{f}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}))$ *is defined in Lemma 4.3.4. Then,* $\Gamma_{\mathbf{p}}\dot{\mathbf{x}}(t,\mathbf{p})$ *is a linear Newton approximation of the map* $\boldsymbol{\eta} \mapsto \dot{\mathbf{x}}(t,\boldsymbol{\eta})$ *at* $\mathbf{p} \in \mathcal{P}$ *and* $\partial_{\mathbf{p}}\dot{\mathbf{x}}(t,\mathbf{p}) \subset \text{conv}\,(\Gamma_{\mathbf{p}}\dot{\mathbf{x}}(t,\mathbf{p}))$ *holds.*

Proof. The first part of the results follow from applying Theorem 2.8.12 to the chain of mappings $\mathbf{p} \mapsto (\mathbf{p},\mathbf{x}(t,\mathbf{p})) \mapsto \mathbf{r}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}))$. The set

$$M = \left\{ \begin{bmatrix} \mathbf{I}_{n_p} \\ \mathbf{N} \end{bmatrix} : \mathbf{N} \in \text{conv}\,(\Gamma_{\mathbf{p}}\mathbf{x}(t,\mathbf{p})) \right\}$$

is a linear Newton approximation for the mapping $\mathbf{p} \mapsto (\mathbf{p},\mathbf{x}(t,\mathbf{p}))$. Using $\text{conv}\,(\Gamma\mathbf{r}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p})))$ as the linear Newton approximation for the map $(\mathbf{p},\mathbf{x}(t,\mathbf{p})) \mapsto \mathbf{r}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}))$ provides the desired result. The second part of the results can be derived using the same reasoning. □

Remark 4.3.8. The results in Theorems 4.2.6, 4.2.7 and 4.2.8 directly apply to the DAE in (4.3.1) with $\mathbf{G}_f(\cdot,\mathbf{p})$ a measurable selection of $\text{conv}\,(\Gamma\mathbf{f}(\cdot,\mathbf{p},\mathbf{x}(\cdot,\mathbf{p})))$ defined in Lemma 4.3.4.

The next three theorems extend the results in Theorems 3.3.9 and 3.3.10.

Theorem 4.3.9. *Let the hypotheses of Theorem 4.3.6 hold. Let* $g : \mathcal{T} \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}} \to \mathbb{R}$ *be a semismooth function. Let* $\Gamma g : \mathcal{T} \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}} \to \mathbb{R}^{1\times(1+n_p+n_x+n_y+n_x)}$ *be a linear Newton approximation of* g *such that* $\partial g(\boldsymbol{\mu}) \subset \text{conv}\,(\Gamma\mathbf{g}(\boldsymbol{\mu}))$ *for all* $\boldsymbol{\mu} \in \mathcal{T} \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}}$.

Let $G : \mathcal{P} \to \mathbb{R}$ be defined by

$$G(\mathbf{p}) = \int_{t_0}^{t_f} g(t, \mathbf{u}(t, \mathbf{p}))dt.$$

Then, G is a semismooth function at $\mathbf{p}$.

Let $\Gamma G : \mathcal{P} \rightrightarrows \mathbb{R}^{1\times n_p}$ be the set

$$\begin{aligned}
\{\mathbf{n}(t_f, \mathbf{p}) \in \mathbb{R}^{1\times n_p} : \dot{\mathbf{n}}(t, \mathbf{p}) &= \mathbf{B}_{\mathbf{x}}(t, \mathbf{u}(t, \mathbf{p}))\mathbf{m}(t, \mathbf{p}) + \mathbf{B}_{\mathbf{p}}(t, \mathbf{u}(t, \mathbf{p})),\ \forall t \in (t_0, t_f],\ \mathbf{n}(t_0, \mathbf{p}) = \mathbf{0},\\
&\mathbf{G}_g(\cdot, \mathbf{p}) \in \mathcal{L}(T, \pi_{\mathbf{u}}\text{conv}\,(\Gamma g(\cdot, \mathbf{u}(\cdot, \mathbf{p})))),\\
&\mathbf{G}_g(t, \mathbf{p}) = [\pi_{\mathbf{p}}\mathbf{G}_g(t, \mathbf{p})\ \pi_{\mathbf{x}}\mathbf{G}_g(t, \mathbf{p})\ \pi_{\mathbf{y}}\mathbf{G}_g(t, \mathbf{p})\ \pi_{\dot{\mathbf{x}}}\mathbf{G}_g(t, \mathbf{p})],\ \forall t \in T,\\
&\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_g(t, \mathbf{p}) \in \mathbb{R}^{1\times n_p},\ \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t, \mathbf{p}) \in \mathbb{R}^{1\times n_x}, \forall t \in T,\\
&\boldsymbol{\pi}_{\mathbf{y}}\mathbf{G}_g(t, \mathbf{p}) \in \mathbb{R}^{1\times n_y},\ \boldsymbol{\pi}_{\dot{\mathbf{x}}}\mathbf{G}_g(t, \mathbf{p}) \in \mathbb{R}^{1\times n_x},\ \forall t \in T,\\
&\dot{\mathbf{m}}(t, \mathbf{p}) = \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t, \mathbf{p})\mathbf{m}(t, \mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t, \mathbf{p}),\ \forall t \in (t_0, t_f],\\
&\mathbf{G}_f(\cdot, \mathbf{p}) \in \mathcal{L}(T, \pi_{\mathbf{v}}\text{conv}\,(\Gamma \mathbf{f}(\cdot, \mathbf{v}(\cdot, \mathbf{p})))),\\
&\mathbf{G}_f(t, \mathbf{p}) = [\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t, \mathbf{p})\ \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t, \mathbf{p})],\ \forall t \in T,\\
&\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t, \mathbf{p}) \in \mathbb{R}^{n_x\times n_p}, \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t, \mathbf{p}) \in \mathbb{R}^{n_x\times n_x},\ \forall t \in T,\\
&\mathbf{m}(t_0, \mathbf{p}) \in \text{conv}\,(\Gamma \mathbf{f}_0(\mathbf{p})),\\
&\mathbf{G}_r(\cdot, \mathbf{p}) \in \mathcal{L}(T, \pi_{\mathbf{v}}\text{conv}\,(\Gamma \mathbf{r}(\cdot, \mathbf{v}(\cdot, \mathbf{p})))),\ \forall t \in T,\\
&\mathbf{G}_r(t, \mathbf{p}) = [\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_r(t, \mathbf{p})\ \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_r(t, \mathbf{p})],\ \forall t \in T,\\
&\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_r(t, \mathbf{p}) \in \mathbb{R}^{n_y\times n_p}, \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_r(t, \mathbf{p}) \in \mathbb{R}^{n_y\times n_x},\ \forall t \in T,\\
&\mathbf{B}_{\mathbf{x}}(t, \mathbf{u}(t, \mathbf{p})) = \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_g(t, \mathbf{p}) +\\
&\qquad \boldsymbol{\pi}_{\mathbf{y}}\mathbf{G}_g(t, \mathbf{p})\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_r(t, \mathbf{p}) + \boldsymbol{\pi}_{\dot{\mathbf{x}}}\mathbf{G}_g(t, \mathbf{p})\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t, \mathbf{p}),\ \forall t \in T, \qquad (4.3.3)\\
&\mathbf{B}_{\mathbf{p}}(t, \mathbf{u}(t, \mathbf{p})) = \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_g(t, \mathbf{p}) +\\
&\qquad \boldsymbol{\pi}_{\mathbf{y}}\mathbf{G}_g(t, \mathbf{p})\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_r(t, \mathbf{p}) + \boldsymbol{\pi}_{\dot{\mathbf{x}}}\mathbf{G}_g(t, \mathbf{p})\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t, \mathbf{p}),\ \forall t \in T\} \qquad (4.3.4)
\end{aligned}$$

where $\Gamma\mathbf{f}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}))$ *and* $\Gamma\mathbf{r}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}))$ *is defined in Corollary 4.3.4. Then* ΓG *is a linear Newton approximation of* G *at* $\mathbf{p}$ *and* $\partial G(\mathbf{p}) \subset \text{conv}\,(\Gamma G(\mathbf{p}))$ holds.

Proof. Let $g(t,\mathbf{u}(t,\mathbf{p})) = g(t,\mathbf{v}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p}),\dot{\mathbf{x}}(t,\mathbf{p}))$. Then, the result follows from the application of Corollary 4.3.7 with Theorem 2.8.12, collecting terms multiplying $\mathbf{m}(t,\mathbf{p})$ and the application of Theorem 4.2.6. □

Theorem 4.3.10. *Let the hypotheses of Theorem 4.3.6 and Theorem 4.3.9 hold. Let* $\boldsymbol{\lambda} : T \to \mathbb{R}^{1\times n_x}$ *be a solution of the initial value problem,*

$$\dot{\boldsymbol{\lambda}}(t) = -\boldsymbol{\lambda}(t)\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t,\mathbf{p}) + \mathbf{B}_{\mathbf{x}}(t,\mathbf{u}(t,\mathbf{p})),\ \forall t \in [t_0,t_f),\ \boldsymbol{\lambda}(t_f) = \mathbf{0}.$$

Then, $\boldsymbol{\lambda}$ *is unique and absolutely continuous. Let* $\mathbf{J} \in \mathbb{R}^{1\times n_p}$ *be defined by*

$$\mathbf{J} = \int_{t_0}^{t_f} \mathbf{B}_{\mathbf{p}}(t,\mathbf{u}(t,\mathbf{p})) - \boldsymbol{\lambda}(t)\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t,\mathbf{p})dt + \boldsymbol{\lambda}(t)\mathbf{m}(t,\mathbf{p})\big|_{t_0}^{t_f}$$

where $\mathbf{m}(t,\mathbf{p})$ *is computed using* (4.3.2), $\mathbf{B}_{\mathbf{x}}(t,\mathbf{u}(t,\mathbf{p}))$ *and* $\mathbf{B}_{\mathbf{p}}(t,\mathbf{u}(t,\mathbf{p}))$ *are computed using* (4.3.3) *and* (4.3.4)*, respectively. Then* $\mathbf{J} \in \Gamma G(\mathbf{p})$.

Proof. The proof is similar to the proof of Theorem 4.2.7 where $\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_g(t,\mathbf{p})$ and $\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_g(t,\mathbf{p})$ are replaced with $\mathbf{B}_{\mathbf{x}}(t,\mathbf{u}(t,\mathbf{p}))$ and $\mathbf{B}_{\mathbf{p}}(t,\mathbf{u}(t,\mathbf{p}))$, respectively. □

Theorem 4.3.11. *Let the hypotheses of Theorem 4.3.6 hold. Let* $h : \mathcal{T}_0 \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}} \to \mathbb{R}$ *be a semismooth function where* $\mathcal{T}_0$ *is an open subset of* $\mathcal{T}$ *such that* $t_f \in \mathcal{T}_0$. *Let* $G : \mathcal{P} \to \mathbb{R} : \boldsymbol{\eta} \mapsto h(t_f,\mathbf{u}(t_f,\boldsymbol{\eta}))$. *Then* G *is semismooth at* $\mathbf{p}$.

Let $\Gamma_{\mathbf{u}}h : \mathcal{P}\times\mathcal{X}\times\mathcal{Y}\times\dot{\mathcal{X}} \rightrightarrows \mathbb{R}^{1\times(n_p+n_x+n_y+n_x)}$ *be a linear Newton approximation of the map* $\bar{\mathbf{u}} \mapsto h(t_f,\bar{\mathbf{u}})$ *such that* $\partial_{\mathbf{u}}h(t_f,\bar{\mathbf{u}}) \subset \text{conv}\,(\Gamma_{\mathbf{u}}h(t_f,\bar{\mathbf{u}}))$ *in a neighborhood of* $\mathbf{u}(t_f,\mathbf{p})$.

Let $\Gamma G : \mathbf{p} \rightrightarrows \mathbb{R}^{1+n_p}$ *be the set*

$$\{\mathbf{M}_{\mathbf{x}}\mathbf{N}_{\mathbf{x}} + \mathbf{M}_{\mathbf{y}}\mathbf{N}_{\mathbf{y}} + \mathbf{M}_{\dot{\mathbf{x}}}\mathbf{N}_{\dot{\mathbf{x}}} + \mathbf{M}_{\mathbf{p}} : [\mathbf{M}_{\mathbf{p}}\ \mathbf{M}_{\mathbf{x}}\ \mathbf{M}_{\mathbf{y}}\ \mathbf{M}_{\dot{\mathbf{x}}}] \in \text{conv}\,(\Gamma_{\mathbf{u}}h(t_f,\mathbf{u}(t_f,\mathbf{p})))\},$$

$$\mathbf{M_p} \in \mathbb{R}^{1\times n_p},\ \mathbf{M_x} \in \mathbb{R}^{1\times n_x},\ \mathbf{M_y} \in \mathbb{R}^{1\times n_y},\ \mathbf{M_{\dot{x}}} \in \mathbb{R}^{1\times n_x},$$

$$\mathbf{N_x} \in \text{conv}\,(\Gamma_{\mathbf{p}}\mathbf{x}(t_f,\mathbf{p})),\ \mathbf{N_y} \in \text{conv}\,(\Gamma_{\mathbf{p}}\mathbf{y}(t_f,\mathbf{p})),$$

$$\mathbf{N_{\dot{x}}} \in \text{conv}\,(\Gamma_{\mathbf{p}}\dot{\mathbf{x}}(t_f,\mathbf{p}))\}.$$

Then, ΓG *is a linear Newton approximation of* G *at* $\mathbf{p} \in \mathcal{P}$ *per Theorem 2.8.12 such that* $\partial G(\mathbf{p}) \subset \text{conv}\,(\Gamma G(\mathbf{p}))$ *holds.*

Let $\mathbf{N_x} = \mathbf{m}(t_f,\mathbf{p})$, $\mathbf{N_y} = \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_r(t_f,\mathbf{p})\mathbf{m}(t_f,\mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_r(t_f,\mathbf{p})$, $\mathbf{N_{\dot{x}}} = \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t_f,\mathbf{p})\mathbf{m}(t_f,\mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t_f,\mathbf{p})$, $\mathbf{C_x} = \mathbf{M_x} + \mathbf{M_y}\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_r(t_f,\mathbf{p}) + \mathbf{M_{\dot{x}}}\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t_f,\mathbf{p})$, $\mathbf{C_p} = \mathbf{M_p} + \mathbf{M_y}\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_r(t_f,\mathbf{p}) + \mathbf{M_{\dot{x}}}\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t_f,\mathbf{p})$ *where* $\mathbf{m}(t,\mathbf{p})$ *is computed using* (4.3.2), $\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_r(t_f,\mathbf{p})$, $\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_r(t_f,\mathbf{p})$, $\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t_f,\mathbf{p})$ *and* $\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t_f,\mathbf{p})$ *are defined in Corollary 4.3.7.*

Let $\boldsymbol{\lambda} : T \to \mathbb{R}^{1\times n_x}$ *be a solution of the initial value problem,*

$$\dot{\boldsymbol{\lambda}}(t) = -\boldsymbol{\lambda}(t)\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_f(t,\mathbf{p}),\ \forall t \in [t_0,t_f),\ \boldsymbol{\lambda}(t_f) = -\mathbf{C_x}.$$

Then, $\boldsymbol{\lambda}$ *is unique and absolutely continuous. Let* $\mathbf{J} \in \mathbb{R}^{1\times n_p}$ *be defined by*

$$\mathbf{J} = \int_{t_0}^{t_f} -\boldsymbol{\lambda}(t)\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_f(t,\mathbf{p})dt - \boldsymbol{\lambda}(t_0)\mathbf{m}(t_0,\mathbf{p}) + \mathbf{C_p}$$

where $\mathbf{m}(t,\mathbf{p})$ *is computed as in Theorem 4.3.6. Then* $\mathbf{J} \in \Gamma G(\mathbf{p})$.

Proof. The proof is the same as the proof of Theorem 4.2.8 except the redefined quantities $\mathbf{C_x}$ and $\mathbf{C_p}$. □

4.4 Multistage Systems

The previous results are extended to dynamic systems whose evolutions are described by disparate differential-algebraic equations in consecutive intervals of time.

Assumption 4.4.1. *Let n_e be a finite positive integer and $\mathcal{I} = \{1, \ldots, n_e\}$. Let $\alpha_i \in \mathbb{R}$, $\beta_i \in \mathbb{R}$, $\alpha_i < \beta_i$, $\forall i \in \mathcal{I}$, $\alpha_{i+1} = \beta_i$, $\forall i \in \mathcal{I}\backslash\{n_e\}$, $-\infty < \alpha_1 < \beta_{n_e} < +\infty$. Let $T = \cup_{i=1}^{n_e}[\alpha_i, \beta_i]$ and $T \subset \mathcal{T}$. Let $\mathcal{T}_i$ be an open subset of $\mathcal{T}$ such that $[\alpha_i, \beta_i] \subset \mathcal{T}_i$ for all $i \in \mathcal{I}$. Let $\mathbf{x}_i : [\alpha_i, \beta_i] \times \mathcal{P} \to \mathcal{X}$, $\mathbf{y}_i : [\alpha_i, \beta_i] \times \mathcal{P} \to \mathcal{Y}$, $\dot{\mathbf{x}}_i : [\alpha_i, \beta_i] \times \mathcal{P} \to \dot{\mathcal{X}}$ for all $i \in \mathcal{I}$, $\mathbf{x} : T \times \mathcal{P} \to \mathcal{X}$, $\mathbf{y} : T \times \mathcal{P} \to \mathcal{Y}$ and $\dot{\mathbf{x}} : T \times \mathcal{P} \to \dot{\mathcal{X}}$. Assume $\mathbf{F}_i : \mathcal{T}_i \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}} \to \mathbb{R}^{n_x+n_y}$ are semismooth functions for all $i \in \mathcal{I}$. Assume $\mathbf{F}_i^0 : \mathcal{P} \times \mathcal{X} \to \mathcal{X}$ for all $i \in \mathcal{I}\backslash\{1\}$ and $\mathbf{F}_1^0 : \mathcal{P} \to \mathcal{X}$ are semismooth functions. Assume there exists a linear Newton approximation $\Gamma\mathbf{F}_i : \mathcal{T}_i \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}} \rightrightarrows \mathbb{R}^{(n_x+n_y)\times(1+n_p+n_x+n_y+n_x)}$ such that $\partial\mathbf{F}_i(\boldsymbol{\mu}) \subset \text{conv}\,(\Gamma\mathbf{F}_i(\boldsymbol{\mu}))$ holds for all $\boldsymbol{\mu} \in \mathcal{T}_i \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}}$ for each $i \in \mathcal{I}$. Assume there exists a linear Newton approximation $\Gamma\mathbf{F}_i^0 : \mathcal{P} \times \mathcal{X} \rightrightarrows \mathbb{R}^{n_x\times(n_p+n_x)}$ such that $\partial\mathbf{F}_i^0(\boldsymbol{\mu}) \subset \text{conv}\,(\Gamma\mathbf{F}_i^0(\boldsymbol{\mu}))$ holds for all $\boldsymbol{\mu} \in \mathcal{P} \times \mathcal{X}$ for all $\mathcal{I}\backslash\{1\}$. Finally, assume there exists $\Gamma\mathbf{F}_1^0 : \mathcal{P} \rightrightarrows \mathbb{R}^{n_x\times n_p}$ such that $\partial\mathbf{F}_1^0(\boldsymbol{\eta}) \subset \text{conv}\,(\Gamma\mathbf{F}_1^0(\boldsymbol{\eta}))$ holds for all $\boldsymbol{\eta} \in \mathcal{P}$.*

The linear Newton approximations associated with the solutions of the initial value problem,

$$\mathbf{0} = \mathbf{F}_i(t, \mathbf{p}, \mathbf{x}_i(t, \mathbf{p}), \mathbf{y}_i(t, \mathbf{p}), \dot{\mathbf{x}}_i(t, \mathbf{p})), \ \forall t \in [\alpha_i, \beta_i], \ \forall i \in \mathcal{I}, \tag{4.4.1}$$

$$\mathbf{0} = \mathbf{x}_1(\alpha_1, \mathbf{p}) - \mathbf{F}_1^0(\mathbf{p}),$$

$$\mathbf{0} = \mathbf{x}_i(\alpha_i, \mathbf{p}) - \mathbf{F}_i^0(\mathbf{p}, \mathbf{x}_{i-1}(\beta_{i-1}, \mathbf{p})), \ \forall i \in \mathcal{I}\backslash\{1\}, \tag{4.4.2}$$

$$\mathbf{0} = \mathbf{x}(t, \mathbf{p}) - \mathbf{x}_i(t, \mathbf{p}), \ \forall t \in [\alpha_i, \beta_i), \ \forall i \in \mathcal{I},$$

$$\mathbf{0} = \mathbf{x}(\beta_{n_e}, \mathbf{p}) - \mathbf{x}_{n_e}(\beta_{n_e}, \mathbf{p}),$$

$$\mathbf{0} = \mathbf{y}(t, \mathbf{p}) - \mathbf{y}_i(t, \mathbf{p}), \ \forall t \in [\alpha_i, \beta_i), \ \forall i \in \mathcal{I},$$

$$\mathbf{0} = \mathbf{y}(\beta_{n_e}, \mathbf{p}) - \mathbf{y}_{n_e}(\beta_{n_e}, \mathbf{p}),$$

$$\mathbf{0} = \dot{\mathbf{x}}(t, \mathbf{p}) - \dot{\mathbf{x}}_i(t, \mathbf{p}), \ \forall t \in [\alpha_i, \beta_i), \ \forall i \in \mathcal{I},$$

$$\mathbf{0} = \dot{\mathbf{x}}(\beta_{n_e}, \mathbf{p}) - \dot{\mathbf{x}}_{n_e}(\beta_{n_e}, \mathbf{p})$$

are derived in this section.

Remark 4.4.2. $\mathbf{x}(\cdot, \mathbf{p})$, $\mathbf{y}(\cdot, \mathbf{p})$ *and* $\dot{\mathbf{x}}(\cdot, \mathbf{p})$ *might be discontinuous at* $t = \alpha_i$ *with* $i > 1$.

Remark 4.4.3. *The results derived in this section are applicable with slight modifications to the case where the number of states, number of algebraic variables as well as the domains of the functions* $\mathbf{F}_i$ *and* $\mathbf{F}_i^0$ *differ for each* $i \in \mathcal{I}$.

Assumption 4.4.4. *Let* $\dot{\mathbf{x}}(\alpha_i, \bar{\mathbf{p}}) = \dot{\bar{\mathbf{x}}}_i$ *and* $\mathbf{y}(\alpha_i, \bar{\mathbf{p}}) = \bar{\mathbf{y}}_i$ *for all* $i \in \mathcal{I}$ *where* $\dot{\bar{\mathbf{x}}}_i$ *and* $\bar{\mathbf{y}}_i$ *are constants. Assume that this condition is sufficient to uniquely determine* $\dot{\mathbf{x}}(\alpha_i, \mathbf{p})$ *and* $\mathbf{y}(\alpha_i, \mathbf{p})$ *uniquely for all* $i \in \mathcal{I}$ *and for all* $\mathbf{p} \in \mathcal{P}$.

Assumption 4.4.5. *Let* $(\mathbf{x}_i(\cdot, \mathbf{p}), \mathbf{y}_i(\cdot, \mathbf{p}), \dot{\mathbf{x}}_i(\cdot, \mathbf{p}))$, $\forall i \in \mathcal{I}$ *be the unique solution of* (4.4.1). *Let* $\mathbf{z}_i : [\alpha_i, \beta_i] \times \mathcal{P} \to \mathcal{P} : (t, \mathbf{p}) \mapsto \mathbf{p}$, $\mathbf{v}_i : [\alpha_i, \beta_i] \times \mathcal{P} \to \mathcal{P} \times \mathcal{X} : (t, \mathbf{p}) \mapsto ((\mathbf{z}_i(t, \mathbf{p})), \mathbf{x}_i(t, \mathbf{p}))$ *and* $\mathbf{u}_i : [\alpha_i, \beta_i] \times \mathcal{P} \to \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}} : (t, \mathbf{p}) \mapsto (\mathbf{v}_i(t, \mathbf{p}), \mathbf{y}_i(t, \mathbf{p}), \dot{\mathbf{x}}_i(t, \mathbf{p}))$. *Let* $\mathbf{u} : T \times \mathcal{P} \to \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}}$ *be such that* $\mathbf{u}(t, \mathbf{p}) = \mathbf{u}_i(t, \mathbf{p})$ *for all* $t \in [\alpha_i, \beta_i)$ *and* $\mathbf{u}(\beta_{n_e}, \mathbf{p}) = \mathbf{u}_{n_e}(\beta_{n_e}, \mathbf{p})$.

Corollary 4.4.6. *Let Assumptions 4.4.1 and 4.4.4 hold. Let the assumptions of Lemma* 4.3.4 *hold for all* $i \in \mathcal{I}$. *Let* $(\mathbf{x}_i(\cdot, \mathbf{p}), \mathbf{y}_i(\cdot, \mathbf{p}), \dot{\mathbf{x}}_i(\cdot, \mathbf{p}))$, $\forall i \in \mathcal{I}$ *be the solution of* (4.4.1). *Then,* $\mathbf{u}(t, \cdot)$ *is semismooth at* $\mathbf{p}$ *for all* $t \in T$.

Proof. Let $n_e = 1$. Then, $\mathbf{u}_1(t, \cdot)$ is a semismooth function at $\mathbf{p}$ for $t \in [\alpha_1, \beta_1]$ per Corollary 4.3.5. Since the composition of semismooth functions is semismooth and $\mathbf{F}_2^0$ is a semismooth function, $\mathbf{u}_2(\alpha_2, \cdot)$ is semismooth at $\mathbf{p}$ if $n_e = 2$. Then, $\mathbf{u}_2(t, \cdot)$ for all $t \in [\alpha_2, \beta_2]$ is semismooth at $\mathbf{p}$ per Corollary 4.3.5. The final result follows from the repeated application of Corollary 4.3.5 and the composition rule for semismooth functions as has been done for the case $n_e \leq 2$. □

Theorem 4.4.7. *Let Assumptions 4.4.1 and 4.4.4 hold. Let the hypotheses of Lemma 4.3.4 hold for all* $i \in \mathcal{I}$. *Let* $\mathbf{f}_i$ *and* $\mathbf{r}_i$ *be the semismooth implicit functions that satisfy* $\mathbf{F}_i(t, \mathbf{v}_i(t, \mathbf{p}), \mathbf{r}_i(t, \mathbf{v}_i(t, \mathbf{p})), \mathbf{f}_i(t, \mathbf{v}_i(t, \mathbf{p}))) = \mathbf{0}$ *for all* $t \in [\alpha_i, \beta_i]$ *and for all* $i \in \mathcal{I}$.

Let $\Gamma_{\mathbf{p}}\mathbf{x}_{0,1}(\mathbf{p}) = \Gamma\mathbf{F}_1^0(\mathbf{p})$. *Let* $\Gamma_{\mathbf{p}}\mathbf{x}_{0,i}(\mathbf{p})$ *be the set*

$$\begin{aligned}\{\mathbf{M}_{\mathbf{x},i}\mathbf{N}_{i-1} + \mathbf{M}_{\mathbf{p},i} :\ & [\mathbf{M}_{\mathbf{x},i}\ \mathbf{M}_{\mathbf{p},i}] \in \operatorname{conv}\left(\Gamma\mathbf{F}_i^0(\mathbf{p}, \mathbf{x}_{i-1}(\mathbf{p}, \beta_{i-1}))\right), \\ & \mathbf{M}_{\mathbf{x},i} \in \mathbb{R}^{n_x \times n_x},\ \mathbf{M}_{\mathbf{p},i} \in \mathbb{R}^{n_x \times n_p}, \mathbf{N}_{i-1} \in \operatorname{conv}\left(\Gamma_{\mathbf{p}}\mathbf{x}_{i-1}(\mathbf{p}, \beta_{i-1})\right)\}\end{aligned} \tag{4.4.3}$$

for all $i \in \mathcal{I}\backslash\{1\}$.

Let $\Gamma_{\mathbf{p}}\mathbf{x}_i(\beta_i, \mathbf{p})$ *be the set*

$$\begin{aligned}\{\mathbf{m}_i(\beta_i, \mathbf{p}) :\ & \dot{\mathbf{m}}_i(t, \mathbf{p}) = \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{f,i}(t, \mathbf{p})\mathbf{m}_i(t, \mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{f,i}(t, \mathbf{p}),\ \forall t \in (\alpha_i, \beta_i], \\ & \mathbf{G}_{f,i}(\cdot, \mathbf{p}) \in \mathcal{L}([\alpha_i, \beta_i], \pi_{\mathbf{v}}\operatorname{conv}\left(\Gamma\mathbf{f}_i(\cdot, \mathbf{v}_i(\cdot, \mathbf{p}))\right)), \\ & \mathbf{G}_{f,i}(t, \mathbf{p}) = [\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{f,i}(t, \mathbf{p})\ \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{f,i}(t, \mathbf{p})],\ \forall t \in [\alpha_i, \beta_i], \\ & \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{f,i}(t, \mathbf{p}) \in \mathbb{R}^{n_x \times n_p}, \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{f,i}(t, \mathbf{p}) \in \mathbb{R}^{n_x \times n_x},\ \forall t \in [\alpha_i, \beta_i], \\ & \mathbf{m}_i(\alpha_i, \mathbf{p}) \in \operatorname{conv}\left(\Gamma_{\mathbf{p}}\mathbf{x}_{0,i}(\mathbf{p})\right)\}.\end{aligned} \tag{4.4.4}$$

where $\Gamma\mathbf{f}_i$ *is defined in Lemma 4.3.4 for all* $i \in \mathcal{I}$.

Let $\Gamma_{\mathbf{p}}\mathbf{x}(t, \mathbf{p}) = \Gamma_{\mathbf{p}}\mathbf{x}_i(t, \mathbf{p})$ *for all* $t \in [\alpha_i, \beta_i)$ *for all* $i \in \mathcal{I}$ *and* $\Gamma_{\mathbf{p}}\mathbf{x}(\beta_{n_e}, \mathbf{p}) = \Gamma_{\mathbf{p}}\mathbf{x}_{n_e}(\beta_{n_e}, \mathbf{p})$. *Then,* $\Gamma_{\mathbf{p}}\mathbf{x}(t, \mathbf{p})$ *is a linear Newton approximation of the map* $\boldsymbol{\eta} \mapsto \mathbf{x}(t, \boldsymbol{\eta})$ *at* $\mathbf{p} \in \mathcal{P}$ *for all* $t \in [\alpha_i, \beta_i)$ *and* $\partial_{\mathbf{p}}\mathbf{x}(t, \mathbf{p}) \subset \operatorname{conv}\left(\Gamma_{\mathbf{p}}\mathbf{x}(t, \mathbf{p})\right)$ *holds for all* $t \in [\alpha_i, \beta_i)$.

Proof. If $n_e = 1$, then, the result follows from Theorem 4.3.6 by letting t_f in that theorem take values in $[\alpha_i, \beta_i)$ and setting $t_0 = \alpha_i$. In this case, $\partial_{\mathbf{p}}\mathbf{x}(t, \mathbf{p}) = \partial_{\mathbf{p}}\mathbf{x}_1(t, \mathbf{p}) \subset \operatorname{conv}\left(\Gamma_{\mathbf{p}}\mathbf{x}_1(t, \mathbf{p})\right)$ holds for all $t \in [\alpha_1, \beta_1]$. Per Theorem 2.8.12, $\Gamma_{\mathbf{p}}\mathbf{x}_{0,2}(\mathbf{p})$ is a linear Newton approximation of the map $\boldsymbol{\eta} \mapsto \mathbf{x}_2(\alpha_2, \boldsymbol{\eta})$ at $\mathbf{p}$. Per Theorem 2.6.7:

$$\partial_{\mathbf{p}}\mathbf{x}_2(\alpha_2, \mathbf{p}) \subset \operatorname{conv}\left(\{\mathbf{M}_{\mathbf{x}}\mathbf{N} + \mathbf{M}_{\mathbf{p}} : \mathbf{N} \in \partial_{\mathbf{p}}\mathbf{x}_1(\beta_1, \mathbf{p}),\ [\mathbf{M}_{\mathbf{x}}\ \mathbf{M}_{\mathbf{p}}] \in \partial\mathbf{F}_2^0(\mathbf{p}, \mathbf{x}_1(\beta_1, \mathbf{p}))\}\right)$$

holds where $\mathbf{M}_{\mathbf{p}} \in \mathbb{R}^{n_p \times n_x}$ and $\mathbf{M}_{\mathbf{x}} \in \mathbb{R}^{n_x \times n_x}$. Then, $\partial_{\mathbf{p}}\mathbf{x}_2(\alpha_2, \mathbf{p}) \subset \operatorname{conv}\left(\Gamma_{\mathbf{p}}\mathbf{x}_{0,2}(\mathbf{p})\right)$ follows

from the fact that $\partial_{\mathbf{p}}\mathbf{x}_1(\beta_1,\mathbf{p}) \subset \text{conv}\,(\Gamma_{\mathbf{p}}\mathbf{x}_1(\beta_1,\mathbf{p}))$ and $\partial\mathbf{F}_2^0(\mathbf{p},\mathbf{x}_1(\mathbf{p},\beta_1)) \subset \text{conv}\,(\Gamma\mathbf{F}_2^0(\beta_1,\mathbf{x}_1(\mathbf{p},\beta_1)))$. Application of Theorem 4.3.6 after setting $\mathbf{F}=\mathbf{F}_2$ and $\Gamma\mathbf{F}^0(\mathbf{p}) = \text{conv}\,(\Gamma_{\mathbf{p}}\mathbf{x}_{0,2}(\mathbf{p}))$ proves the result for $n_e = 2$. The case for larger n_e is proven by the repeated application of Theorem 4.3.6 and Theorem 2.8.12 as has been done for the case $n_e = 2$. □

Theorem 4.4.8. *Let the hypotheses of Theorem 4.4.7 hold. Define* $G_i : \mathcal{P} \to \mathbb{R}$ *as*

$$G_i(\mathbf{p}) = \int_{\alpha_i}^{\beta_i} g_i(t,\mathbf{u}_i(t,\mathbf{p}))dt$$

where $g_i : \mathcal{T}_i \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}} \mapsto \mathbb{R}$ *are semismooth functions for all* $i \in \mathcal{I}$. *Then,* G_i *are semismooth functions at* $\mathbf{p}$ *per Theorem 4.3.9.*

Let $G : \mathcal{P} \to \mathbb{R}$ *be defined by*

$$G(\boldsymbol{\eta}) = \sum_{i=1}^{n_e} G_i(\boldsymbol{\eta}).$$

Then, G *is a semismooth function at* $\mathbf{p} \in \mathcal{P}$ *since the sum of semismooth functions is semismooth.*

Let $\Gamma G_i : \mathcal{P} \rightrightarrows \mathbb{R}^{1\times n_p}$ *be the set*

$$\begin{aligned}
\{\mathbf{n}_i(\beta_i,\mathbf{p}) \in \mathbb{R}^{1\times n_p} :\ & \dot{\mathbf{n}}_i(t,\mathbf{p}) = \mathbf{B}_{\mathbf{x},i}(t,\mathbf{u}_i(t,\mathbf{p}))\mathbf{m}_i(t,\mathbf{p}) + \mathbf{B}_{\mathbf{p},i}(t,\mathbf{u}_i(t,\mathbf{p})),\ \forall t \in (\alpha_i,\beta_i],\ \mathbf{n}(\alpha_i,\mathbf{p}) = \mathbf{0},\\
& \mathbf{G}_{g,i}(\cdot,\mathbf{p}) \in \mathcal{L}(T, \pi_{\mathbf{u}}\text{conv}\,(\Gamma g_i(\cdot,\mathbf{u}_i(\cdot,\mathbf{p})))),\\
& \mathbf{G}_{g,i}(t,\mathbf{p}) = [\pi_{\mathbf{p}}\mathbf{G}_{g,i}(t,\mathbf{p})\ \pi_{\mathbf{x}}\mathbf{G}_{g,i}(t,\mathbf{p})\ \pi_{\mathbf{y}}\mathbf{G}_{g,i}(t,\mathbf{p})\ \pi_{\dot{\mathbf{x}}}\mathbf{G}_{g,i}(t,\mathbf{p})],\ \forall t \in [\alpha_i,\beta_i],\\
& \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{g,i}(t,\mathbf{p}) \in \mathbb{R}^{1\times n_p},\ \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{g,i}(t,\mathbf{p}) \in \mathbb{R}^{1\times n_x}, \forall t \in [\alpha_i,\beta_i],\\
& \boldsymbol{\pi}_{\mathbf{y}}\mathbf{G}_{g,i}(t,\mathbf{p}) \in \mathbb{R}^{1\times n_y},\ \boldsymbol{\pi}_{\dot{\mathbf{x}}}\mathbf{G}_{g,i}(t,\mathbf{p}) \in \mathbb{R}^{1\times n_x},\ \forall t \in [\alpha_i,\beta_i],\\
& \dot{\mathbf{m}}_i(t,\mathbf{p}) = \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{f,i}(t,\mathbf{p})\mathbf{m}_i(t,\mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{f,i}(t,\mathbf{p}),\ \forall t \in (\alpha_i,\beta_i],\\
& \mathbf{G}_{f,i}(\cdot,\mathbf{p}) \in \mathcal{L}(T, \pi_{\mathbf{v}}\text{conv}\,(\Gamma\mathbf{f}_i(\cdot,\mathbf{v}(\cdot,\mathbf{p})))),\\
& \mathbf{G}_{f,i}(t,\mathbf{p}) = [\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{f,i}(t,\mathbf{p})\ \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{f,i}(t,\mathbf{p})],\ \forall t \in [\alpha_i,\beta_i],
\end{aligned}$$

$$\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{f,i}(t,\mathbf{p}) \in \mathbb{R}^{n_x \times n_p}, \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{f,i}(t,\mathbf{p}) \in \mathbb{R}^{n_x \times n_x}, \ \forall t \in [\alpha_i, \beta_i],$$

$$\mathbf{m}_i(\alpha_i, \mathbf{p}) \in \text{conv}\,(\Gamma_{\mathbf{p}}\mathbf{x}_{0,i}(\mathbf{p})),$$

$$\mathbf{G}_{r,i}(\cdot, \mathbf{p}) \in \mathcal{L}(T, \pi_{\mathbf{v}}\text{conv}\,(\Gamma\mathbf{r}_i(\cdot, \mathbf{v}(\cdot, \mathbf{p}))), \ \forall t \in [\alpha_i, \beta_i],$$

$$\mathbf{G}_{r,i}(t,\mathbf{p}) = [\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{r,i}(t,\mathbf{p}) \ \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{r,i}(t,\mathbf{p})], \ \forall t \in [\alpha_i, \beta_i],$$

$$\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{r,i}(t,\mathbf{p}) \in \mathbb{R}^{n_y \times n_p}, \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{r,i}(t,\mathbf{p}) \in \mathbb{R}^{n_y \times n_x}, \ \forall t \in [\alpha_i, \beta_i],$$

$$\mathbf{B}_{\mathbf{x},i}(t, \mathbf{u}_i(t,\mathbf{p})) = \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{g,i}(t,\mathbf{p}) +$$

$$\boldsymbol{\pi}_{\mathbf{y}}\mathbf{G}_{g,i}(t,\mathbf{p})\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{r,i}(t,\mathbf{p}) + \boldsymbol{\pi}_{\dot{\mathbf{x}}}\mathbf{G}_{g,i}(t,\mathbf{p})\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{f,i}(t,\mathbf{p}), \ \forall t \in [\alpha_i, \beta_i], \tag{4.4.5}$$

$$\mathbf{B}_{\mathbf{p},i}(t, \mathbf{u}_i(t,\mathbf{p})) = \ \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{g,i}(t,\mathbf{p}) +$$

$$\boldsymbol{\pi}_{\mathbf{y}}\mathbf{G}_{g,i}(t,\mathbf{p})\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{r,i}(t,\mathbf{p}) + \boldsymbol{\pi}_{\dot{\mathbf{x}}}\mathbf{G}_{g,i}(t,\mathbf{p})\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{f,i}(t,\mathbf{p}), \ \forall t \in [\alpha_i, \beta_i]\} \tag{4.4.6}$$

where $\Gamma\mathbf{f}_i$ *and* $\Gamma\mathbf{r}_i$ *are defined in Lemma 4.3.4. Then* ΓG_i *is a linear Newton approximation of* G_i *at* $\mathbf{p}$ *and* $\partial G_i(\mathbf{p}) \subset \text{conv}\,(\Gamma G_i(\mathbf{p}))$ *holds for all* $i \in \mathcal{I}$ *per Theorem 4.3.9 and Theorem 4.4.7.*

Let $\Gamma G(\mathbf{p})$ *be*

$$\Gamma G(\mathbf{p}) = \sum_{i=1}^{n_e} \text{conv}\,(\Gamma G_i(\mathbf{p})).$$

Then, ΓG *is a linear Newton approximation of* G *and* $\partial G(\mathbf{p}) \subset \text{conv}\,(\Gamma G(\mathbf{p}))$ *holds since*

$$\partial G(\mathbf{p}) \subset \sum_{i=1}^{n_e} \partial G_i(\mathbf{p}).$$

For all $i \in \mathcal{I}$, *let* $\boldsymbol{\lambda}_i : [\alpha_i, \beta_i] \to \mathbb{R}^{1 \times n_x}$ *be a solution of the initial value problem*

$$\dot{\boldsymbol{\lambda}}_i(t) = \ -\boldsymbol{\lambda}_i(t)\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{f,i}(t,\mathbf{p}) + \mathbf{B}_{\mathbf{x},i}(t, \mathbf{u}(t,\mathbf{p}), \ \forall t \in (\alpha_i, \beta_i], \ \forall i \in \mathcal{I},$$

$$\boldsymbol{\lambda}_i(\beta_i) = \boldsymbol{\lambda}_{i+1}(\beta_i)\mathbf{M}_{\mathbf{x},i+1}, \forall i \in \mathcal{I}\backslash\{n_e\},\ \boldsymbol{\lambda}_{n_e}(\beta_{n_e}) = \mathbf{0}.$$

Then, $\boldsymbol{\lambda}_i$ is absolutely continuous and unique. Let $\mathbf{J}$ be

$$\sum_{i=1}^{n_e} \int_{\alpha_i}^{\beta_i} \mathbf{B}_{\mathbf{p},i}(t, \mathbf{u}_i(t,\mathbf{p})) - \boldsymbol{\lambda}_i(t)\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{f,i}(t,\mathbf{p})dt - \sum_{i=1}^{n_e-1} \boldsymbol{\lambda}_{i+1}(\beta_i)\mathbf{M}_{\mathbf{p},i+1} + \boldsymbol{\lambda}_{n_e}(\beta_{n_e})\mathbf{m}_{n_e}(\beta_{n_e},\mathbf{p}) - \boldsymbol{\lambda}_1(\alpha_1)\mathbf{m}_1(\alpha_1,\mathbf{p})$$

where $\mathbf{M}_{\mathbf{p},i+1}$, $\mathbf{M}_{\mathbf{x},i+1}$ and $\mathbf{m}_1(\alpha_1,\mathbf{p})$ are defined in Theorem 4.4.7. Then, $\mathbf{J} \in \Gamma G(\mathbf{p})$.

Proof. Applying Theorem 4.3.10 for each $i \in \mathcal{I}$:

$$\mathbf{J}_i = \int_{\alpha_i}^{\beta_i} \mathbf{B}_{\mathbf{p},i}(t, \mathbf{u}_i(t,\mathbf{p})) - \boldsymbol{\lambda}_i(t)\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{f,i}(t,\mathbf{p})dt + \boldsymbol{\lambda}_i(t)\mathbf{m}_i(t,\mathbf{p})|_{\alpha_i}^{\beta_i} \tag{4.4.7}$$

is obtained with $\mathbf{J}_i \in \Gamma G_i(\mathbf{p})$ where

$$\dot{\boldsymbol{\lambda}}_i(t) = -\boldsymbol{\lambda}_i(t)\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{f,i}(t,\mathbf{p}) + \mathbf{B}_{\mathbf{x},i}(t,\mathbf{u}_i(t,\mathbf{p})),\ \forall t \in [\alpha_i,\beta_i),\ \boldsymbol{\lambda}_i(\beta_i) = \mathbf{0}.$$

It can be seen from its derivation that (4.4.7) holds for any $\boldsymbol{\lambda}_0 \in \mathbb{R}^{1\times n_x}$ and $\boldsymbol{\lambda}_i(\beta_i) = \boldsymbol{\lambda}_0$. $\boldsymbol{\lambda}_0$ is set to the zero vector in order to avoid the computation of $\mathbf{m}_i(\beta_i,\mathbf{p})$.

Let $\mathbf{J} = \sum_{i=1}^{n_e} \mathbf{J}_i$. Clearly, $\mathbf{J} \in \Gamma G(\mathbf{p})$ because $\mathbf{J}_i \in \Gamma G_i(\mathbf{p})$. Then, $\mathbf{J}$ is equal to

$$\sum_{i=1}^{n_e} \int_{\alpha_i}^{\beta_i} \mathbf{B}_{\mathbf{p},i}(t, \mathbf{u}_i(t,\mathbf{p})) - \boldsymbol{\lambda}_i(t)\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{f,i}(t,\mathbf{p})dt + \boldsymbol{\lambda}_i(t)\mathbf{m}_i(t,\mathbf{p})|_{\alpha_i}^{\beta_i}. \tag{4.4.8}$$

Note that $\mathbf{m}_{i+1}(\alpha_{i+1},\mathbf{p})$ is $\mathbf{M}_{\mathbf{x},i+1}\mathbf{m}_i(\beta_i) + \mathbf{M}_{\mathbf{p},i+1}$ in view of (4.4.3) for some $\mathbf{M}_{\mathbf{x},i+1}$ and $\mathbf{M}_{\mathbf{p},i+1}$ where $\mathbf{M}_{\mathbf{x},i+1}$ and $\mathbf{M}_{\mathbf{p},i+1}$ are defined in Theorem 4.4.7. Hence

$$-\boldsymbol{\lambda}_{i+1}(\alpha_{i+1})\mathbf{m}_{i+1}(\alpha_{i+1},\mathbf{p}) = -\boldsymbol{\lambda}_{i+1}(\alpha_{i+1})(\mathbf{M}_{\mathbf{x},i+1}\mathbf{m}_i(\beta_i) + \mathbf{M}_{\mathbf{p},i+1}).$$

Then, setting

$$\boldsymbol{\lambda}_{i+1}(\alpha_{i+1})\mathbf{M}_{\mathbf{x},i+1} = \boldsymbol{\lambda}_i(\beta_i)$$

allows the cancellation of terms in (4.4.8) and results in

$$\sum_{i=1}^{n_e} \int_{\alpha_i}^{\beta_i} \mathbf{B}_{\mathbf{p},i}(t, \mathbf{u}_i(t,\mathbf{p})) - \boldsymbol{\lambda}_i(t)\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{f,i}(t,\mathbf{p})dt -$$
$$\sum_{i=1}^{n_e-1} \boldsymbol{\lambda}_{i+1}(\beta_i)\mathbf{M}_{\mathbf{p},i+1} + \boldsymbol{\lambda}_{n_e}(\beta_{n_e})\mathbf{m}_{n_e}(\beta_{n_e},\mathbf{p}) - \boldsymbol{\lambda}_1(\alpha_1)\mathbf{m}_1(\alpha_1,\mathbf{p}).$$

Setting $\boldsymbol{\lambda}_{n_e}(\beta_{n_e}) = \mathbf{0}$ provides the desired result. □

Corollary 4.4.9. *Let the hypotheses of Theorem 4.4.7 hold.*

1. *Let* $\Gamma_{\mathbf{p}}\mathbf{y} : \mathcal{T} \times \mathcal{P} \rightrightarrows \mathbb{R}^{n_y \times n_p}$ *be defined by*

$$\begin{aligned}
\Gamma_{\mathbf{p}}\mathbf{y}(t,\mathbf{p}) = \{\mathbf{n}(t,\mathbf{p}) : \; & \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{r,i}(t,\mathbf{p})\mathbf{m}_i(t,\mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{r,i}(t,\mathbf{p}), \\
& \mathbf{G}_{r,i}(t,\mathbf{p}) \in \operatorname{conv}(\pi_{\mathbf{v}}\Gamma\mathbf{r}_i(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}))), \\
& \mathbf{G}_{r,i}(t,\mathbf{p}) = [\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{r,i}(t,\mathbf{p}) \; \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{r,i}(t,\mathbf{p})], \\
& \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{r,i}(t,\mathbf{p}) \in \mathbb{R}^{n_y \times n_p}, \\
& \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{r,i}(t,\mathbf{p}) \in \mathbb{R}^{n_y \times n_x}, \\
& \mathbf{m}_i(t,\mathbf{p}) \in \operatorname{conv}(\Gamma_{\mathbf{p}}\mathbf{x}_i(\beta_{n_e},\mathbf{p}))\}, \; \forall t \in [\alpha_i,\beta_i), \forall i \in \mathcal{I}, \\
\Gamma_{\mathbf{p}}\mathbf{y}(\beta_{n_e},\mathbf{p}) = \{\mathbf{n}(\beta_{n_e},\mathbf{p}) : \; & \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{r,n_e}(\beta_{n_e},\mathbf{p})\mathbf{m}_{n_e}(\beta_{n_e},\mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{r,n_e}(\beta_{n_e},\mathbf{p}), \\
& \mathbf{G}_{r,n_e}(\beta_{n_e},\mathbf{p}) \in \operatorname{conv}(\pi_{\mathbf{v}}\Gamma\mathbf{r}_{n_e}(\beta_{n_e},\mathbf{p},\mathbf{x}(t,\mathbf{p}))), \\
& \mathbf{G}_{r,n_e}(\beta_{n_e},\mathbf{p}) = [\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{r,n_e}(\beta_{n_e},\mathbf{p}) \; \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{r,n_e}(\beta_{n_e},\mathbf{p})], \\
& \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{r,n_e}(\beta_{n_e},\mathbf{p}) \in \mathbb{R}^{n_y \times n_p}, \\
& \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{r,n_e}(\beta_{n_e},\mathbf{p}) \in \mathbb{R}^{n_y \times n_x},
\end{aligned}$$

$$\mathbf{m}_{n_e}(\beta_{n_e}, \mathbf{p}) \in \text{conv}\,(\Gamma_{\mathbf{p}}\mathbf{x}_{n_e}(\beta_{n_e}, \mathbf{p}))\}$$

where $\Gamma_{\mathbf{p}}\mathbf{x}_i(t, \mathbf{p})$ *is defined as in Theorem 4.4.7 and* $\Gamma\mathbf{r}_i$ *is defined using Lemma 4.3.4 for all* $i \in \mathcal{I}$. *Then,* $\Gamma_{\mathbf{p}}\mathbf{y}(t, \mathbf{p})$ *is a linear Newton approximation of the map* $\boldsymbol{\eta} \mapsto \mathbf{y}(t, \boldsymbol{\eta})$ *at* $\mathbf{p} \in \mathcal{P}$ *and* $\partial_{\mathbf{p}}\mathbf{y}(t, \mathbf{p}) \subset \text{conv}\,(\Gamma_{\mathbf{p}}\mathbf{y}(t, \mathbf{p}))$ *holds.*

2. *Let* $\Gamma_{\mathbf{p}}\dot{\mathbf{x}} : \mathcal{T} \times \mathcal{P} \rightrightarrows \mathbb{R}^{n_y \times n_p}$ *be defined by*

$$\begin{aligned}
\Gamma_{\mathbf{p}}\dot{\mathbf{x}}(t, \mathbf{p}) = \{\mathbf{n}(t, \mathbf{p}) : \; & \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{f,i}(t, \mathbf{p})\mathbf{m}_i(t, \mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{f,i}(t, \mathbf{p}), \\
& \mathbf{G}_{f,i}(t, \mathbf{p}) \in \text{conv}\,(\pi_{\mathbf{v}}\Gamma\mathbf{f}_i(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}))), \\
& \mathbf{G}_{f,i}(t, \mathbf{p}) = [\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{f,i}(t, \mathbf{p})\ \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{f,i}(t, \mathbf{p})], \\
& \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{f,i}(t, \mathbf{p}) \in \mathbb{R}^{n_x \times n_p}, \\
& \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{f,i}(t, \mathbf{p}) \in \mathbb{R}^{n_x \times n_x}, \\
& \mathbf{m}_i(t, \mathbf{p}) \in \text{conv}\,(\Gamma_{\mathbf{p}}\mathbf{x}_i(t, \mathbf{p}))\}, \; \forall t \in [\alpha_i, \beta_i), \forall i \in \mathcal{I}, \\
\Gamma_{\mathbf{p}}\dot{\mathbf{x}}(\beta_{n_e}, \mathbf{p}) = \{\mathbf{n}(\beta_{n_e}, \mathbf{p}) : \; & \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{f,n_e}(\beta_{n_e}, \mathbf{p})\mathbf{m}_{n_e}(\beta_{n_e}, \mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{f,n_e}(\beta_{n_e}, \mathbf{p}), \\
& \mathbf{G}_{f,n_e}(\beta_{n_e}, \mathbf{p}) \in \text{conv}\,(\pi_{\mathbf{v}}\Gamma\mathbf{f}_{n_e}(\beta_{n_e}, \mathbf{p}, \mathbf{x}(\beta_{n_e}, \mathbf{p}))), \\
& \mathbf{G}_{f,n_e}(\beta_{n_e}, \mathbf{p}) = [\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{f,n_e}(\beta_{n_e}, \mathbf{p})\ \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{f,n_e}(\beta_{n_e}, \mathbf{p})], \\
& \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{f,n_e}(\beta_{n_e}, \mathbf{p}) \in \mathbb{R}^{n_x \times n_p}, \\
& \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{f,n_e}(\beta_{n_e}, \mathbf{p}) \in \mathbb{R}^{n_x \times n_x}, \\
& \mathbf{m}_{n_e}(\beta_{n_e}, \mathbf{p}) \in \text{conv}\,(\Gamma_{\mathbf{p}}\mathbf{x}_{n_e}(\beta_{n_e}, \mathbf{p}))\}
\end{aligned}$$

where $\Gamma_{\mathbf{p}}\mathbf{x}_i(t, \mathbf{p})$ *is defined as in Theorem 4.4.7 and* $\Gamma\mathbf{f}_i$ *is defined using Lemma 4.3.4 for all* $i \in \mathcal{I}$. *Then,* $\Gamma_{\mathbf{p}}\dot{\mathbf{x}}(t, \mathbf{p})$ *is a linear Newton approximation of the map* $\boldsymbol{\eta} \mapsto \dot{\mathbf{x}}(t, \boldsymbol{\eta})$ *at* $\mathbf{p} \in \mathcal{P}$ *and* $\partial_{\mathbf{p}}\dot{\mathbf{x}}(t, \mathbf{p}) \subset \text{conv}\,(\Gamma_{\mathbf{p}}\dot{\mathbf{x}}(t, \mathbf{p}))$ *holds.*

Proof. Application of Theorem 2.8.12 to the implicit functions $\mathbf{r}_i$ and $\mathbf{f}_i$ produces $\Gamma_{\mathbf{p}}\mathbf{y}$ and $\Gamma_{\mathbf{p}}\dot{\mathbf{x}}$, respectively. Application of Theorem 2.6.7 to $\mathbf{r}_i$ and $\mathbf{f}_i$ provides sets M_r and M_f

such that $\partial_{\mathbf{p}}\mathbf{y}(t,\mathbf{p}) \subset \text{conv}\,(M_r)$ and $\partial_{\mathbf{p}}\dot{\mathbf{x}}(t,\mathbf{p}) \subset \text{conv}\,(M_f)$ hold. Then, it can be shown that $\text{conv}\,(M_r) \subset \text{conv}\,(\Gamma_{\mathbf{p}}\mathbf{y}(t,\mathbf{p}))$ and $\text{conv}\,(M_f) \subset \text{conv}\,(\Gamma\mathbf{p}\dot{\mathbf{x}}(t,\mathbf{p}))$ using the fact that $\partial\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p})) \subset \text{conv}\,(\Gamma\mathbf{f}_i(t,\mathbf{v}_i(t,\mathbf{p})))$, $\partial\mathbf{r}_i(t,\mathbf{v}_i(t,\mathbf{p})) \subset \text{conv}\,(\Gamma\mathbf{r}_i(t,\mathbf{v}_i(t,\mathbf{p})))$ and $\partial_{\mathbf{p}}\mathbf{x}(t,\mathbf{p}) \subset \text{conv}\,(\Gamma_{\mathbf{p}}\mathbf{x}(t,\mathbf{p}))$. □

Theorem 4.4.10. *Let the hypotheses of Theorem 4.4.7 hold. Let* $h : \mathcal{T}_0 \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}} \to \mathbb{R}$ *be a semismooth function where* $\mathcal{T}_0$ *is an open subset of* $\mathcal{T}$ *such that* $\beta_{n_e} \in \mathcal{T}_0$. *Let* $G : \mathcal{P} \to \mathbb{R} : \boldsymbol{\eta} \mapsto h(\beta_{n_e}, \mathbf{u}(\beta_{n_e}, \boldsymbol{\eta}))$. *Then,* G *is semismooth at* $\mathbf{p}$.

Let $\Gamma_{\mathbf{u}}h : \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}} \rightrightarrows \mathbb{R}^{1\times(n_p+n_x+n_y+n_x)}$ *be a linear Newton approximation of the map* $\bar{\mathbf{u}} \mapsto h(\beta_{n_e}, \bar{\mathbf{u}})$ *such that* $\partial_{\mathbf{u}}h(\beta_{n_e}, \bar{\mathbf{u}}) \subset \text{conv}\,(\Gamma_{\mathbf{u}}h(t_f, \bar{\mathbf{u}}))$ *in a neighborhood of* $\mathbf{u}(\beta_{n_e}, \mathbf{p})$.

Let $\Gamma G : \mathbf{p} \rightrightarrows \mathbb{R}^{1+n_p}$ *be the set*

$$\begin{aligned}\{\mathbf{A}_{\mathbf{x}}\mathbf{N}_{\mathbf{x}} + \mathbf{A}_{\mathbf{y}}\mathbf{N}_{\mathbf{y}} + \mathbf{A}_{\dot{\mathbf{x}}}\mathbf{N}_{\dot{\mathbf{x}}} + \mathbf{A}_{\mathbf{p}} : [\mathbf{A}_{\mathbf{p}}\ \mathbf{A}_{\mathbf{x}}\ \mathbf{A}_{\mathbf{y}}\ \mathbf{A}_{\dot{\mathbf{x}}}] \in \text{conv}\,(\Gamma_{\mathbf{u}}h(\beta_{n_e}, \mathbf{u}(\beta_{n_e}, \mathbf{p}))),\\ \mathbf{A}_{\mathbf{p}} \in \mathbb{R}^{1\times n_p},\ \mathbf{A}_{\mathbf{x}} \in \mathbb{R}^{1\times n_x},\ \mathbf{A}_{\mathbf{y}} \in \mathbb{R}^{1\times n_y},\ \mathbf{A}_{\dot{\mathbf{x}}} \in \mathbb{R}^{1\times n_x},\\ \mathbf{N}_{\mathbf{x}} \in \text{conv}\,(\Gamma_{\mathbf{p}}\mathbf{x}(\beta_{n_e}, \mathbf{p})),\ \mathbf{N}_{\mathbf{y}} \in \text{conv}\,(\Gamma_{\mathbf{p}}\mathbf{y}(\beta_{n_e}, \mathbf{p})),\\ \mathbf{N}_{\dot{\mathbf{x}}} \in \text{conv}\,(\Gamma_{\mathbf{p}}\dot{\mathbf{x}}(\beta_{n_e}, \mathbf{p}))\}\end{aligned}$$

where $\Gamma_{\mathbf{p}}\mathbf{x}(\beta_{n_e}, \mathbf{p})$ *is defined in Theorem 4.4.7,* $\Gamma_{\mathbf{p}}\mathbf{y}(\beta_{n_e}, \mathbf{p})$ *and* $\Gamma_{\mathbf{p}}\dot{\mathbf{x}}(\beta_{n_e}, \mathbf{p})$ *are defined in Corollary 4.4.9. Then,* ΓG *is a linear Newton approximation of* G *at* $\mathbf{p} \in \mathcal{P}$ *per Theorem 2.8.12 such that* $\partial G(\mathbf{p}) \subset \text{conv}\,(\Gamma G(\mathbf{p}))$ *holds.*

Let $\mathbf{N}_{\mathbf{x}} = \mathbf{m}_{n_e}(\beta_{n_e}, \mathbf{p})$, $\mathbf{N}_{\mathbf{y}} = \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{r,n_e}(\beta_{n_e}, \mathbf{p})\mathbf{m}_{n_e}(\beta_{n_e}, \mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{r,n_e}(\beta_{n_e}, \mathbf{p})$, $\mathbf{N}_{\dot{\mathbf{x}}} = \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{f,n_e}(\beta_{n_e}, \mathbf{p})\ \mathbf{m}_{n_e}(\beta_{n_e}, \mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{f,n_e}(\beta_{n_e}, \mathbf{p})$, $\mathbf{C}_{\mathbf{x}} = \mathbf{A}_{\mathbf{x}} + \mathbf{A}_{\mathbf{y}}\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{r,n_e}(\beta_{n_e}, \mathbf{p}) + \mathbf{A}_{\dot{\mathbf{x}}}\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{f,n_e}(\beta_{n_e}, \mathbf{p})$, $\mathbf{C}_{\mathbf{p}} = \mathbf{A}_{\mathbf{p}} + \mathbf{A}_{\mathbf{y}}\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{r,n_e}(\beta_{n_e}, \mathbf{p}) + \mathbf{A}_{\dot{\mathbf{x}}}\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{f,n_e}(\beta_{n_e}, \mathbf{p})$ *where* $\mathbf{m}_{n_e}(t, \mathbf{p})$ *is computed using Theorem 4.4.7,* $\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{r,n_e}(\beta_{n_e}, \mathbf{p})$, $\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{r,n_e}(\beta_{n_e}, \mathbf{p})$, $\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{f,n_e}(\beta_{n_e}, \mathbf{p})$ *and* $\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{f,n_e}(\beta_{n_e}, \mathbf{p})$ *are defined in Corollary 4.4.9.*

Let $\boldsymbol{\lambda}_i : [\alpha_i, \beta_i] \to \mathbb{R}^{1\times n_x}$ *be a solution of the initial value problem,*

$$\dot{\boldsymbol{\lambda}}_i(t) = -\boldsymbol{\lambda}_i(t)\boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_{f,i}(t,\mathbf{p}),\ \forall t \in [\alpha_i, \beta_i),\ \forall i \in \mathcal{I},$$

$$\boldsymbol{\lambda}_i(\beta_i) = \boldsymbol{\lambda}_{i+1}(\alpha_{i+1})\mathbf{M}_{\mathbf{x},i+1}, \forall i \in \mathcal{I}\backslash\{n_e\},\ \boldsymbol{\lambda}_{n_e}(\beta_{n_e}) = -\mathbf{C}_{\mathbf{x}}.$$

Then, $\boldsymbol{\lambda}$ *is unique and absolutely continuous. Let* $\mathbf{J} \in \mathbb{R}^{1\times n_p}$ *be defined by*

$$\mathbf{J} = \sum_{i=1}^{n_e}\int_{\alpha_i}^{\beta_i} -\boldsymbol{\lambda}_i(t)\boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_{f,i}(t,\mathbf{p})dt + \sum_{i=1}^{n_e-1}(-\boldsymbol{\lambda}_{i+1}(\alpha_{i+1})\mathbf{M}_{\mathbf{p},i+1}) - \boldsymbol{\lambda}_1(\alpha_1)\mathbf{m}_1(\alpha_1,\mathbf{p}) + \mathbf{C}_{\mathbf{p}}$$

where $\mathbf{m}_1(\alpha_1,\mathbf{p})$, $\mathbf{M}_{\mathbf{p},i+1}$ *and* $\mathbf{M}_{\mathbf{x},i+1}$ *are defined in Theorem 4.4.7. Then,* $\mathbf{J} \in \Gamma G(\mathbf{p})$.

Proof. Let N be the set

$$\{\mathbf{A}_{\mathbf{x}}\mathbf{N}_{\mathbf{x}} + \mathbf{A}_{\mathbf{y}}\mathbf{N}_{\mathbf{y}} + \mathbf{A}_{\dot{\mathbf{x}}}\mathbf{N}_{\dot{\mathbf{x}}} + \mathbf{A}_{\mathbf{p}} : [\mathbf{A}_{\mathbf{p}}\ \mathbf{A}_{\mathbf{x}}\ \mathbf{A}_{\mathbf{y}}\ \mathbf{A}_{\dot{\mathbf{x}}}] \in \partial_{\mathbf{u}}h(\beta_{n_e}, \mathbf{u}(\beta_{n_e},\mathbf{p})),$$
$$\mathbf{A}_{\mathbf{p}} \in \mathbb{R}^{1\times n_p},\ \mathbf{A}_{\mathbf{x}} \in \mathbb{R}^{1\times n_x},\ \mathbf{A}_{\mathbf{y}} \in \mathbb{R}^{1\times n_y},\ \mathbf{A}_{\dot{\mathbf{x}}} \in \mathbb{R}^{1\times n_x},$$
$$\mathbf{N}_{\mathbf{x}} \in \partial_{\mathbf{p}}\mathbf{x}(\beta_{n_e},\mathbf{p}),\ \mathbf{N}_{\mathbf{y}} \in \partial_{\mathbf{p}}\mathbf{y}(\beta_{n_e},\mathbf{p}),$$
$$\mathbf{N}_{\dot{\mathbf{x}}} \in \partial_{\mathbf{p}}\dot{\mathbf{x}}(\beta_{n_e},\mathbf{p})\}.$$

Then, using Theorem 2.6.7, it can be shown that $\partial G(\mathbf{p}) \subset \mathrm{conv}\,(N) \subset \mathrm{conv}\,(\Gamma G(\mathbf{p}))$ since $\partial_{\mathbf{p}}\mathbf{x}(\beta_{n_e},\mathbf{p}) \subset \mathrm{conv}\,(\Gamma_{\mathbf{p}}\mathbf{x}(\beta_{n_e},\mathbf{p}))$, $\partial_{\mathbf{p}}\mathbf{y}(\beta_{n_e},\mathbf{p}) \subset \mathrm{conv}\,(\Gamma_{\mathbf{p}}\mathbf{y}(\beta_{n_e},\mathbf{p}))$, $\partial_{\mathbf{p}}\dot{\mathbf{x}}(\beta_{n_e},\mathbf{p}) \subset \mathrm{conv}\,(\Gamma_{\mathbf{p}}\dot{\mathbf{x}}(\beta_{n_e},\mathbf{p}))$ and $\partial_{\mathbf{u}}h(\beta_{n_e},\mathbf{u}(\beta_{n_e},\mathbf{p})) \subset \mathrm{conv}\,(\Gamma_{\mathbf{u}}h(\beta_{n_e},\mathbf{u}(\beta_{n_e},\mathbf{p})))$ holds.

Let $\mathbf{J} = \mathbf{C}_{\mathbf{x}}\mathbf{m}_{n_e}(\beta_{n_e},\mathbf{p}) + \mathbf{C}_{\mathbf{p}}$. Then, $\mathbf{J} \in \Gamma G(\mathbf{p})$ by the definition of $\Gamma G(\mathbf{p})$.

Note that $\mathbf{C}_{\mathbf{x}}\mathbf{m}_{n_e}(\beta_{n_e},\mathbf{p})$ is equal to

$$\sum_{i=1}^{n_e}\int_{\alpha_i}^{\beta_i}\mathbf{C}_{\mathbf{x}}\dot{\mathbf{m}}_i(t,\mathbf{p})dt + \sum_{i=1}^{n_e-1}(\mathbf{C}_{\mathbf{x}}\mathbf{m}_{i+1}(\alpha_{i+1},\mathbf{p}) - \mathbf{C}_{\mathbf{x}}\mathbf{m}_i(\beta_i,\mathbf{p})) + \mathbf{C}_{\mathbf{x}}\mathbf{m}_1(\alpha_1,\mathbf{p}).$$

This can be written as

$$\sum_{i=1}^{n_e} \int_{\alpha_i}^{\beta_i} \mathbf{C_x}\dot{\mathbf{m}}_i(t,\mathbf{p}) - \boldsymbol{\lambda}_i(t)(\boldsymbol{\pi}_\mathbf{x}\mathbf{G}_{f,i}(t,\mathbf{p})\mathbf{m}_i(t,\mathbf{p}) + \boldsymbol{\pi}_\mathbf{p}\mathbf{G}_{f,i}(t,\mathbf{p}) - \dot{\mathbf{m}}_i(t,\mathbf{p}))dt +$$
$$\sum_{i=1}^{n_e-1}(\mathbf{C_x}\mathbf{m}_{i+1}(\alpha_{i+1},\mathbf{p}) - \mathbf{C_x}\mathbf{m}_i(\beta_i,\mathbf{p})) + \mathbf{C_x}\mathbf{m}_1(\alpha_1,\mathbf{p}).$$

Applying integration by parts and collecting terms as done in Theorem 4.2.8 results in

$$\sum_{i=1}^{n_e} \int_{\alpha_i}^{\beta_i} -\boldsymbol{\lambda}_i(t)\boldsymbol{\pi}_\mathbf{p}\mathbf{G}_{f,i}(t,\mathbf{p})dt + (\mathbf{C_x} + \boldsymbol{\lambda}_i(t))\mathbf{m}_i(t,\mathbf{p})|_{\alpha_i}^{\beta_i} +$$
$$\sum_{i=1}^{n_e-1}(\mathbf{C_x}\mathbf{m}_{i+1}(\alpha_{i+1},\mathbf{p}) - \mathbf{C_x}\mathbf{m}_i(\beta_i,\mathbf{p})) + \mathbf{C_x}\mathbf{m}_1(\alpha_1,\mathbf{p}).$$

This expression can be simplified to

$$\sum_{i=1}^{n_e} \int_{\alpha_i}^{\beta_i} -\boldsymbol{\lambda}_i(t)\boldsymbol{\pi}_\mathbf{p}\mathbf{G}_{f,i}(t,\mathbf{p})dt +$$
$$\sum_{i=1}^{n_e-1}(-\boldsymbol{\lambda}_{i+1}(\alpha_{i+1})\mathbf{m}_{i+1}(\alpha_{i+1},\mathbf{p}) + \boldsymbol{\lambda}_i(\beta_i)\mathbf{m}_i(\beta_i,\mathbf{p})) +$$
$$(\mathbf{C_x} + \boldsymbol{\lambda}_{n_e}(\beta_{n_e}))\mathbf{m}_{n_e}(\beta_{n_e},\mathbf{p}) - \boldsymbol{\lambda}_1(\alpha_1)\mathbf{m}_1(\alpha_1,\mathbf{p}).$$

Note that $\mathbf{m}_{i+1}(\alpha_{i+1},\mathbf{p}) = \mathbf{M}_{\mathbf{x},i+1}\mathbf{m}_i(\beta_i,\mathbf{p}) + \mathbf{M}_{\mathbf{p},i+1}$ where $\mathbf{M}_{\mathbf{x},i+1}$ and $\mathbf{M}_{\mathbf{p},i+1}$ are as defined in Theorem 4.4.7. Setting $\boldsymbol{\lambda}_i(\beta_i) = \boldsymbol{\lambda}_{i+1}(\alpha_{i+1})\mathbf{M}_{\mathbf{x},i+1}$, and $\boldsymbol{\lambda}_{n_e}(\beta_{n_e}) = -\mathbf{C_x}$, provides the desired result. □

4.5 Example

In this section, Example 3.6.3 is revisited. In order to analyze that example, the following corollary to Theorem 2.7.3 is required.

Corollary 4.5.1. *Let Y_1 and Y_2 be open subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. Let $\mathbf{y}_1 \in Y_1$ and $\mathbf{y}_2 \in Y_2$. Let $\mathbf{H} : Y_1 \times Y_2 \to \mathbb{R}^m$ be a PC^1 function that is completely coherently oriented with respect to Y_2 at $(\mathbf{y}_1, \mathbf{y}_2)$. Suppose $\mathbf{H}(\mathbf{y}_1, \mathbf{y}_2) = \mathbf{0}$. Let $\bar{\mathbf{H}} : Y_2 \to \mathbb{R}^m : \mathbf{y} \mapsto \mathbf{H}(\mathbf{y}_1, \mathbf{y})$. Let α be the sign of the determinants in $\pi_2 \partial^B \mathbf{H}(\mathbf{y}_1, \mathbf{y}_2) = \{\mathbf{J}_2 \mathbf{H}_i(\mathbf{y}_1, \mathbf{y}_2) : \ i \in \mathcal{I}(\mathbf{H}, (\mathbf{y}_1, \mathbf{y}_2))\}$. Then $\operatorname{ind}(\bar{\mathbf{H}}, \mathbf{y}_2) = \alpha$ and the conditions of Theorem 4.1.7 are satisfied.*

Proof. Let $\mathbf{F} : Y_1 \times Y_2 \to \mathbb{R}^{n+m} : (\mathbf{x}, \mathbf{y}) \mapsto (\mathbf{x}, \mathbf{H}(\mathbf{x}, \mathbf{y}))$. $\mathbf{F}$ is a PC^1 function in the neighborhood of $(\mathbf{y}_1, \mathbf{y}_2)$ with essentially active selection functions $\mathbf{F}_i : Y_1 \times Y_2 \to \mathbb{R}^{n+m} : (\mathbf{x}, \mathbf{y}) \mapsto (\mathbf{x}, \mathbf{H}_i(\mathbf{x}, \mathbf{y}))$, $i \in \mathcal{I}(\mathbf{H}, (\mathbf{y}_1, \mathbf{y}_2))$ and

$$\partial^B \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2) = \left\{ \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{J}_1 \mathbf{H}_i(\mathbf{y}_1, \mathbf{y}_2) & \mathbf{J}_2 \mathbf{H}_i(\mathbf{y}_1, \mathbf{y}_2) \end{bmatrix} : i \in \mathcal{I}(\mathbf{H}, (\mathbf{y}_1, \mathbf{y}_2)) \right\}.$$

Note that the elements of $\partial^B \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2)$ are coherently oriented because the determinant of each element is equal to $\det(\mathbf{I}_m) \times \det(\mathbf{J}_2 \mathbf{H}_i(\mathbf{y}_1, \mathbf{y}_2))$ due to the special structure of the elements. Let the sign of $\det(\mathbf{J}_2 \mathbf{H}_i(\mathbf{y}_1, \mathbf{y}_2))$ be α which is constant and nonzero for all $i \in \mathcal{I}$ by the complete coherent orientation of $\mathbf{H}$ with respect to Y_2. Since $\mathbf{H}$ is completely coherently oriented with respect to Y_2, $\mathbf{F}$ is completely coherently oriented with respect to $Y_1 \times Y_2$.

The B-derivative $\mathbf{F}'(\mathbf{y}_1, \mathbf{y}_2; \cdot)$ is a continuous piecewise linear function such that

$$\partial^B \mathbf{F}'(\mathbf{y}_1, \mathbf{y}_2; \mathbf{0}) \subset \partial^B \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2)$$

holds per Lemma 2 in [79]. Therefore, $\mathbf{F}'(\mathbf{y}1, \mathbf{y}2; \cdot)$ is completely coherently oriented at $\mathbf{0}$ with respect to $\mathbb{R}^{n+m}$. Per Corollary 19 in [91], $\mathbf{F}'(\mathbf{y}1, \mathbf{y}2; \cdot)$ is invertible at $\mathbf{0}$. Per the equivalence of Conditions (ii) and (iii) in Theorem 5 in [91], $\mathbf{F}$ is invertible at $\mathbf{y}$ and $\operatorname{ind}(\mathbf{F}, (\mathbf{y}_1, \mathbf{y}_2)) = \pm 1$. It can be shown as in the proof of Theorem 2 in [42], that the sign of $\operatorname{ind}(\mathbf{F}, (\mathbf{y}_1, \mathbf{y}_2))$ is α. Finally, it can be shown that $\operatorname{ind}(\mathbf{F}, (\mathbf{y}_1, \mathbf{y}_2)) = \operatorname{ind}(\bar{\mathbf{H}}, \mathbf{y}_2)$ as in the proof of Theorem 4 in

[42]. □

Example 4.5.2. Consider the dynamic system introduced in Example 3.6.3 where the computation of the strict derivative of the mapping $\boldsymbol{\eta} \mapsto \mathbf{x}(t, \boldsymbol{\eta})$ is discussed. Here, however, the computation of an element of $\Gamma_{\mathbf{p}}\mathbf{x}(t, \mathbf{p})$ is discussed.

First note that the matrix $\mathbf{B}(t, \boldsymbol{\eta})$ defined in (3.6.6) has nonzero determinant because of the triangular structure of the matrix and the invertibility of $\mathbf{J}_{\dot{\mathbf{x}}}\mathbf{V}(t, \boldsymbol{\eta})$, $\mathbf{I}_{n_y}$ and $\mathbf{R}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}})$. The sign of $\det(\mathbf{B}(t, \boldsymbol{\eta}))$ is equal to the sign of $\det(\mathbf{J}_{\dot{\mathbf{x}}}\mathbf{V}(t, \boldsymbol{\eta}))$ because the determinant of $\mathbf{I}_{n_y}$ and the sign of the determinant of any possible $\mathbf{R}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}})$ are both one. Assume that the sign of the determinant of $\mathbf{J}_{\dot{\mathbf{x}}}\mathbf{V}(t, \boldsymbol{\eta})$ is nonzero and constant on it domain.

Let

$$\mathbf{C}(t, \boldsymbol{\eta}) = \begin{bmatrix} \mathbf{J}_t\mathbf{V}(t, \boldsymbol{\eta}) & \mathbf{J}_{\mathbf{p}}\mathbf{V}(t, \boldsymbol{\eta}) & \mathbf{J}_{\mathbf{x}}\mathbf{V}(t, \boldsymbol{\eta}) \\ -\mathbf{J}_t\mathbf{Q}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}}) & -\mathbf{J}_{\mathbf{p}}\mathbf{Q}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}}) & -\mathbf{J}_{\mathbf{x}}\mathbf{Q}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}}) \\ \mathbf{K}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}})\mathbf{J}_t\mathbf{Q}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}}) & \mathbf{K}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}})\mathbf{J}_{\mathbf{p}}\mathbf{Q}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}}) & \mathbf{K}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}})\mathbf{J}_{\mathbf{x}}\mathbf{Q}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}}) \end{bmatrix}$$

at points where $\mathbf{F}$ is differentiable.

Every element of $\partial^B\mathbf{F}(t, \boldsymbol{\eta})$ has the same structure as $[\mathbf{C}(t, \boldsymbol{\eta})\ \mathbf{B}(t, \boldsymbol{\eta})]$ but with different $\mathbf{R}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}})$ and $\mathbf{K}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}})$ matrices due to the PC^1 nature of $\mathbf{F}$. It can be shown that $\mathbf{F}$ is completely coherently oriented at $(t, \boldsymbol{\eta})$ with respect to $\mathcal{Y} \times \mathcal{W} \times \dot{\mathcal{X}}$ using the fact that any matrices in the set $\Lambda(\mathbf{y}_1, \mathbf{y}_2)$ defined in Definition 2.7.2 differ only in the $\mathbf{R}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}})$ matrices they contain. Since all possible $\mathbf{R}(t, \boldsymbol{\eta}_{\mathbf{p}}, \boldsymbol{\eta}_{\mathbf{x}})$ have positive determinantal sign, complete coherent orientation follows from the structure of $\mathbf{B}(t, \boldsymbol{\eta})$.

Per Corollary 4.5.1, the conditions for Lemma 4.3.4 hold. As a result, Theorem 4.3.6 can be used to compute an element of $\Gamma_{\mathbf{p}}\mathbf{x}(t, \mathbf{p})$.

Chapter 5

Numerical Computation of the Generalized Jacobians and Linear Newton Approximations

In this chapter, the numerical computation of elements of the linear Newton approximations and generalized Jacobians derived for the mapping $\boldsymbol{\eta} \mapsto (\mathbf{x}(t_f, \boldsymbol{\eta}), \mathbf{y}(t_f, \boldsymbol{\eta}), \dot{\mathbf{x}}(t_f, \boldsymbol{\eta}))$ in Chapters 3 and 4 is discussed. Two main issues need to be addressed in order to obtain accurate numerical values. The first one is the computation of an element of the generalized Jacobians or the linear Newton approximations of the functions in the right-hand sides of (4.2.1), (4.3.1) and (4.4.1). The second one is the accurate numerical integration of the auxiliary equations that define the generalized Jacobians and linear Newton approximations of the mapping $\boldsymbol{\eta} \mapsto (\mathbf{x}(t_f, \boldsymbol{\eta}), \mathbf{y}(t_f, \mathbf{p}), \dot{\mathbf{x}}(t_f, \mathbf{p}))$. As shown in Example 3.6.2, these auxiliary equations have discontinuous right-hand sides. ODE/DAE solvers that do not take into account the discontinuous nature of these equations either cannot integrate these equations correctly or become inefficient because they have to take too many time steps to satisfy integration tolerances when discontinuities are encountered.

An assumption on the structure of the right-hand side functions is made in the rest of this thesis. The right-hand side functions become PC^1 functions as a result and an element of the generalized Jacobian of these functions can be readily computed using the properties of PC^1 functions (§2.7.1). The structural assumption divides the domain of the right-hand side functions into subsets. The right-hand side functions are continuously differentiable on the interior of these subsets and possibly nondifferentiable on the boundaries of these subsets. This regularity in the placement of the nondifferentiable points allows the detection of the discontinuities in the auxiliary equations using state event location algorithms [83].

The first section introduces the aforementioned structural assumption and discusses the computation of the elements of the linear Newton approximations and generalized Jacobians of the right-hand side functions. The second section discusses modifications to the auxiliary equations in case the underlying dynamics is of the form (4.3.1) and (4.4.1) to improve the efficiency of computation. In this case, the inversion of matrices is required to compute an element of the linear Newton approximation or the strict derivative. This is undesirable because it is computationally very inefficient. The structural assumption allows the use of more efficient techniques that do not explicitly invert the matrices. The final section reviews relevant aspects of state event location. Results in Chapter 3 require that the time points at which the state trajectory visits nondifferentiable points in the domains of the right-hand side functions constitute a set of measure zero. A numerical method to check this condition is introduced.

5.1 Linear Newton Approximations of the Right-Hand Side

Assumption 5.1.1. *(Property M) Let X be a connected open subset of $\mathbb{R}^n$. A locally Lipschitz continuous function $\mathbf{F} : X \to \mathbb{R}^m$ satisfies property M if the following hold:*

1. $n_D(\mathbf{F})$ *is a finite positive integer,* $\mathcal{D}(\mathbf{F}) = \{1, \ldots, n_D(\mathbf{F})\}$.
2. $\mathrm{cl}\,(X) = \cup_{i=1}^{n_D(\mathbf{F})} \mathrm{cl}\,(U_i)$ *where* U_i *are open subsets of* $\mathbb{R}^n$ *such that if* $i' \neq i''$, *then* $U_{i'} \cap U_{i''} = \emptyset$ *for all* $i', i'' \in \mathcal{D}(\mathbf{F})$.
3. O_i *are open subsets of* $\mathbb{R}^n$ *such that* $\mathrm{cl}\,(U_i) \subset O_i$ *for all* $i \in \mathcal{D}(\mathbf{F})$.
4. $\mathrm{int}\,(\mathrm{cl}\,(U_i)) = U_i,\ \forall i \in \mathcal{D}(\mathbf{F})$.
5. $\mathcal{J}_i(\mathbf{F}) = \{1, \ldots, n_i(\mathbf{F})\}$ *where* $n_i(\mathbf{F})$ *is a finite positive integer for all* $i \in \mathcal{D}(\mathbf{F})$.
6. $g_{i,j} : O_i \to \mathbb{R} \in \mathcal{C}^1(O_i)$ *for all* $i \in \mathcal{D}(\mathbf{F})$ *and for all* $j \in \mathcal{J}_i(\mathbf{F})$.
7. $g_i : O_i \to \mathbb{R} : \boldsymbol{\eta} \mapsto \max\{g_{i,j}(\boldsymbol{\eta}),\ j \in \mathcal{J}_i(\mathbf{F})\}$ *for all* $i \in \mathcal{D}(\mathbf{F})$.
8. *For all* $i \in \mathcal{D}(\mathbf{F})$, $g_i(\boldsymbol{\eta}) < 0,\ \forall \boldsymbol{\eta} \in U_i$, $g_i(\boldsymbol{\eta}) = 0,\ \forall \boldsymbol{\eta} \in \mathrm{cl}\,(U_i) \backslash U_i$ *and* $g_i(\boldsymbol{\eta}) > 0,\ \forall \boldsymbol{\eta} \in O_i \backslash \mathrm{cl}\,(U_i)$.
9. *The set* $\{\boldsymbol{\eta} \in O_i : g_i(\boldsymbol{\eta}) = 0\}$ *constitutes a piecewise continuously differentiable manifold of dimension* $n - 1$ *for all* $i \in \mathcal{D}(\mathbf{F})$.
10. *For each* $i \in \mathcal{D}(\mathbf{F})$, *there exists a function,* $\mathbf{F}_i : O_i \to \mathbb{R}^m$ *such that* $\mathbf{F}_i \in \mathcal{C}^1(O_i)$ *and* $\mathbf{F}(\boldsymbol{\eta}) = \mathbf{F}_i(\boldsymbol{\eta})$ *for all* $\boldsymbol{\eta} \in \mathrm{cl}\,(U_i)$.

Remark 5.1.2. The functions $g_{i,j}$ defined in Assumption 5.1.1 are called *discontinuity* or *zero-crossing* functions.

Remark 5.1.3. Item 4 is necessary to exclude any point, $\boldsymbol{\eta} \in \mathrm{cl}\,(U_i) \backslash U_i$ that has no neighborhood containing points not in U_i.

Corollary 5.1.4. *Suppose that* $\mathbf{F} : X \to \mathbb{R}^m$ *satisfies property M. Then* $\mathbf{F}$ *is a* PC^1 *function on* X. *Let* $\boldsymbol{\eta} \in X$. *If* $g_j(\boldsymbol{\eta}) < 0$ *for some* $j \in \mathcal{D}(\mathbf{F})$, *then, the essentially active function indices at* $\boldsymbol{\eta}$, $\mathcal{I}(\mathbf{F}, \boldsymbol{\eta})$, *is* $\{j\}$ *and* $\partial\mathbf{F}(\boldsymbol{\eta}) = \{\mathbf{J}\mathbf{F}_j(\boldsymbol{\eta})\}$. *If* $g_j(\boldsymbol{\eta}) = 0$, *for some* $j \in \mathcal{D}(\mathbf{F})$, *then, there exists at least one more index* $k \neq j$ *such that* $g_k(\boldsymbol{\eta}) = 0$. *Let* $\mathcal{K} \subset \mathcal{D}(\mathbf{F})$ *be the set containing all indices* $i \in \mathcal{K}$ *such that* $g_i(\boldsymbol{\eta}) = 0$ *holds. Then* $\mathcal{I}(\mathbf{F}, \boldsymbol{\eta}) = \mathcal{K}$ *and* $\partial\mathbf{F}(\boldsymbol{\eta}) = \mathrm{conv}\,(\{\mathbf{J}\mathbf{F}_i(\boldsymbol{\eta}), i \in \mathcal{K}\})$.

Proof. If $g_j(\boldsymbol{\eta}) < 0$ for some $j \in \mathcal{D}(\mathbf{F})$, then $\boldsymbol{\eta} \in U_j$. Since U_j is open, there exists O, a

neighborhood of $\boldsymbol{\eta}$ such that $O \subset U_j$. Hence $\mathcal{I}(\mathbf{F}, \boldsymbol{\eta}) = \{j\}$ and $\partial \mathbf{F}(\boldsymbol{\eta}) = \{\mathbf{J}\mathbf{F}_j(\boldsymbol{\eta})\}$. If $\mathbf{g}_j(\boldsymbol{\eta}) = 0$, let O be any neighborhood of $\boldsymbol{\eta}$ such that $O \subset X \subset \cup_{i=1}^{n_D(\mathbf{F})} \mathrm{cl}\,(U_i)$. $O \cap U_j$ is nonempty because $\boldsymbol{\eta} \in \mathrm{cl}\,(U_j)$. Therefore, there exists no O which is a subset of any U_i for $i \in \mathcal{D}(\mathbf{F}) \backslash \{j\}$. Otherwise, $U_j \cap U_i$ would be nonempty. This also implies that there is at least one more index, $k \in \mathcal{D}(\mathbf{F})$ such that $k \neq j$ and $g_k(\boldsymbol{\eta}) = 0$. Assume otherwise. Note that $\boldsymbol{\eta} \in X$ and X is open. Since $g_j(\boldsymbol{\eta}) = 0$, $\boldsymbol{\eta} \in \mathrm{cl}\,(U_j) \backslash U_j$ per property M. Per property 4, every neighborhood of $\boldsymbol{\eta}$ contains points not in U_j. Therefore, there exists no O, a neighborhood of $\boldsymbol{\eta}$, such that $O \subset U_j$. Hence $\boldsymbol{\eta} \notin X$ because X is open but $\boldsymbol{\eta}$ is not an interior point of X.

Let $\mathcal{K} \subset \mathcal{D}(\mathbf{F})$ contain all indices such that $g_i(\boldsymbol{\eta}) = 0$ if $i \in \mathcal{K}$. Note that $O \cap U_i$ is a nonempty open set for any neighborhood, O, of $\boldsymbol{\eta}$ for all $i \in \mathcal{K}$ because both sets are open. Let $X_i = \{\mathbf{u} \in O : \mathbf{F}_i(\mathbf{u}) = \mathbf{F}(\mathbf{u})\}$ for each $i \in \mathcal{K}$. Then, $O \cap U_i \subset \mathrm{int}\,(X_i)$. Note that $\boldsymbol{\eta} \in \mathrm{cl}\,(U_i)$, and therefore $\boldsymbol{\eta}$ is a limit point of $O \cap U_i$. Hence $\boldsymbol{\eta} \in \mathrm{cl}\,(O \cap U_i)$. Since $\mathrm{cl}\,(O \cap U_i) \subset \mathrm{cl}\,(\mathrm{int}\,(X_i))$ holds, $\boldsymbol{\eta} \in \mathrm{cl}\,(\mathrm{int}\,(X_i))$. Therefore, $\mathcal{I}(\mathbf{F}, \boldsymbol{\eta}) = \mathcal{K}$ and $\partial \mathbf{F}(\boldsymbol{\eta}) = \mathrm{conv}\,(\{\mathbf{J}\mathbf{F}_i(\boldsymbol{\eta}), i \in \mathcal{K}\})$ per the definition of essentially active function indices. □

Remark 5.1.5. In the remainder of this thesis, let $\Omega_m(\mathbf{F}) = (\cup_{i=1}^{n_D(\mathbf{F})}(\mathrm{cl}\,(U_i) \backslash U_i)) \cap X = X \backslash \cup_{i=1}^{n_D(\mathbf{F})} U_i$.

Consider (4.2.1). Assume $\mathbf{f}$ and $\mathbf{f}_0$ satisfy property M. Then $\mathbf{f}$ and $\mathbf{f}_0$ are PC^1 functions on their respective domains. Therefore,

$$\partial \mathbf{f}(\boldsymbol{\eta}) = \mathrm{conv}\,(\{\mathbf{J}\mathbf{f}_i(\boldsymbol{\eta}), i \in \mathcal{I}(\mathbf{f}, \boldsymbol{\eta})\}),\ \forall \boldsymbol{\eta} \in \mathcal{T} \times \mathcal{P} \times \mathcal{X},$$
$$\partial \mathbf{f}_0(\mathbf{p}) = \mathrm{conv}\,(\{\mathbf{J}\mathbf{f}_{0,i}(\mathbf{p}), i \in \mathcal{I}(\mathbf{f}_0, \mathbf{p})\}),\ \forall \mathbf{p} \in \mathcal{P},$$

where $\mathbf{f}_i$ and $\mathbf{f}_{0,i}$ correspond to the functions $\mathbf{F}_i$ defined in Assumption 5.1.1. In this case, Assumption 4.2.1 holds with $\Gamma \mathbf{f} = \partial \mathbf{f}$ and $\Gamma \mathbf{f}_0 = \partial \mathbf{f}_0$ since $\mathbf{f}$ and $\mathbf{f}_0$ are semismooth functions (§2.8.5).

Note that $\partial_{\mathbf{v}} \mathbf{f}(\boldsymbol{\eta}) \subset \pi_{\mathbf{v}} \partial \mathbf{f}(\boldsymbol{\eta})$ per Theorem 2.6.10. The assumptions of Theorem 3.2.3

hold if $S = \{t \in [t_0, t_f] : (t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})) \in \Omega_m(\mathbf{f})\}$ is a measure zero subset of $[t_0, t_f]$ and $\mathbf{p} \in \mathcal{P} \backslash \Omega_m(\mathbf{f}_0)$. If $t \in T \backslash S$, then $(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})) \in U_i$ for some $i \in \mathcal{I}(\mathbf{f}, (t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})))$ and $\partial_{\mathbf{v}} \mathbf{f}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})) = \pi_{\mathbf{v}} \partial \mathbf{f}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})) = \{\mathbf{J}_{\mathbf{v}} \mathbf{f}_i(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}))\}$ holds because $\mathbf{f}$ is continuously hence strictly differentiable in this case.

Consider (4.3.1). Assume $\mathbf{F}$ and $\mathbf{F}_0$ satisfy property M. Then, $\mathbf{F}$ and $\mathbf{F}_0$ are PC^1 functions on their respective domains as in the ODE case and

$$\partial \mathbf{F}(\boldsymbol{\eta}) = \operatorname{conv}(\{\mathbf{J}\mathbf{F}_i(\boldsymbol{\eta}), i \in \mathcal{I}(\mathbf{F}, \boldsymbol{\eta})\}), \ \forall \boldsymbol{\eta} \in \mathcal{T} \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}},$$
$$\partial \mathbf{F}_0(\mathbf{p}) = \operatorname{conv}(\{\mathbf{J}\mathbf{F}_{0,i}(\mathbf{p}), i \in \mathcal{I}(\mathbf{F}_0, \mathbf{p})\}), \ \forall \mathbf{p} \in \mathcal{P}.$$

Then, Assumption 4.3.1 holds with $\Gamma \mathbf{F} = \partial \mathbf{F}$ and $\Gamma \mathbf{F}_0 = \partial \mathbf{F}_0$ since $\mathbf{F}$ and $\mathbf{F}_0$ are semismooth functions (§2.8.5). In this case

$$\partial^B \mathbf{F}(\boldsymbol{\eta}) = \{\mathbf{J}\mathbf{F}_i(\boldsymbol{\eta}), i \in \mathcal{I}(\mathbf{F}, \boldsymbol{\eta})\}$$

because $\mathbf{F}$ is a PC^1 function. The assumptions of Theorem 3.3.6 hold if $S = \{t \in [t_0, t_f] : (t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p}), \dot{\mathbf{x}}(t, \mathbf{p})) \in \Omega_m(\mathbf{F})\}$ is a measure zero subset of $[t_0, t_f]$ and $\mathbf{p} \in \mathcal{P} \backslash \Omega_m(\mathbf{F}_0)$. As in the ODE case, if $t \in T \backslash S$, then $\partial_{\mathbf{u}} \mathbf{F}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p}), \dot{\mathbf{x}}(t, \mathbf{p})) = \pi_{\mathbf{u}} \partial \mathbf{F}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p}), \dot{\mathbf{x}}(t, \mathbf{p})) = \{\mathbf{J}_{\mathbf{u}} \mathbf{F}_i(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p}), \dot{\mathbf{x}}(t, \mathbf{p}))\}$ holds for some $i \in \mathcal{I}(\mathbf{F}, (t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p}), \dot{\mathbf{x}}(t, \mathbf{p})))$.

Assume g as defined in Theorem 4.3.10 satisfies property M. Then

$$\partial g(\boldsymbol{\eta}) = \operatorname{conv}(\{J g_i(\boldsymbol{\eta}), i \in \mathcal{I}(g, \boldsymbol{\eta})\}), \ \forall \boldsymbol{\eta} \in \mathcal{T} \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}}$$

where the functions g_i correspond to the functions $\mathbf{F}_i$ in the statement of Assumption 5.1.1 and the assumptions of Theorem 4.3.10 hold with $\Gamma g = \partial g$. The assumptions of Theorem 3.3.9 hold if $S = \{t \in [t_0, t_f] : (t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p}), \dot{\mathbf{x}}(t, \mathbf{p})) \in \Omega_m(g) \cup \Omega_m(\mathbf{F})\}$ is a measure

zero subset of $[t_0, t_f]$ and $\mathbf{p} \in \mathcal{P} \backslash \Omega_m(\mathbf{F}_0)$. If $t \in T \backslash S$, then $\partial_{\mathbf{u}} g(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p}), \dot{\mathbf{x}}(t, \mathbf{p})) = \pi_{\mathbf{u}} \partial g(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p}), \dot{\mathbf{x}}(t, \mathbf{p})) = \{J_{\mathbf{u}} g_i(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p}), \dot{\mathbf{x}}(t, \mathbf{p}))\}$ for some $i \in \mathcal{I}(g, (t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p}), \dot{\mathbf{x}}(t, \mathbf{p})))$.

Let h as defined in Theorem 4.3.11 satisfy property M. The assumptions of Theorem 4.3.11 hold if $\Gamma_{\mathbf{u}} h = \pi_{\mathbf{u}} \partial h$ where

$$\partial h(\boldsymbol{\eta}) = \operatorname{conv}(\{Jh_i(\boldsymbol{\eta}), i \in \mathcal{I}(h, \boldsymbol{\eta})\}), \ \forall \boldsymbol{\eta} \in \mathcal{T}_0 \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}}$$

and the h_i correspond to the functions $\mathbf{F}_i$ in the statement of Assumption 5.1.1. Note that ∂h is a linear Newton approximation of h. Then using Theorem 2.8.12, it can be shown that $\pi_{\mathbf{u}} \partial h$ is a linear Newton approximation of the mapping $\boldsymbol{\eta} \mapsto h(t_f, \boldsymbol{\eta})$.

The discussion for the multistage systems in (4.4.1) is similar to the DAE case and therefore it is omitted.

5.2 Singleton and Non-Singleton Trajectories

Assume $\mathbf{f}$ and $\mathbf{f}_0$ in (4.2.1) satisfy property M. Let $\mathbf{p} \in \mathcal{P}$. Let $\mathbf{x}(t, \mathbf{p})$ be the solution of (4.2.1). Let $S = \{t \in [t_0, t_f] : (t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})) \in \Omega_m(\mathbf{f})\}$. If S is a set of measure zero and $\mathbf{p} \in \mathcal{P} \backslash \Omega_m(\mathbf{f}_0)$, then the solution is called a *singleton trajectory* otherwise it is called a *non-singleton trajectory.* If the solution is a singleton trajectory, then the assumptions of Theorem 3.2.3 are satisfied and $\partial_{\mathbf{p}} \mathbf{x}(t_f, \mathbf{p})$ is a singleton set.

Figure 5-1 depicts examples of singleton trajectories and a non-singleton trajectory. For clarity, the solutions of an ordinary differential equation of the form $\dot{x} = f(t, x), \ x(0) = p$ are shown where f satisfies property M. Different shaded areas represent the sets, $\{U_i\}_{i=1}^5$ as described in the statement of Assumption 5.1.1. The functions $\{g_i\}_{i=1}^5$ are as defined in Assumption 5.1.1. The functions $\{f_i\}_{i=1}^5$ correspond to the functions $\{\mathbf{F}_i\}_{i=1}^{n_D(\mathbf{F})}$.

Trajectory A is a non-singleton trajectory because it tracks the boundary of U_5 defined by g_5. Trajectory B and C are singleton trajectories. Trajectory C crosses one boundary at a time whereas B crosses multiple boundaries at a time. In addition, trajectory B has a point where the trajectory is tangent to the boundary of U_5 at a point. Solutions of (4.2.1) display

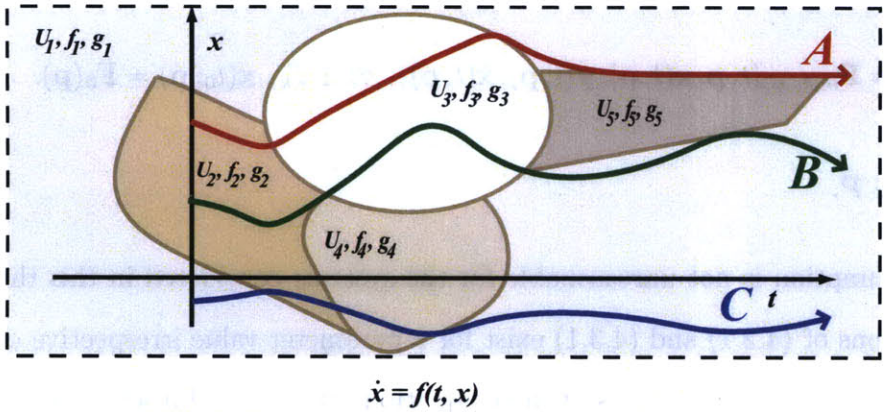


Figure 5-1: Examples of singleton and non-singleton trajectories.

the same behavior as shown in Figure 5-1. The forward sensitivity equations in Theorem 3.2.3 and the differential equations in Theorem 4.2.3 are possibly discontinuous when the trajectory crosses boundaries (Trajectory B and C). These equations have continuous right-hand sides when $\mathbf{x}(t, \mathbf{p}) \in U_i$ for some $i \in \mathcal{D}(\mathbf{f})$ per property M. If the trajectory tracks the boundary of U_i, using $\mathbf{J_v f}_i(\cdot, \mathbf{p}, \mathbf{x}(\cdot, \mathbf{p}))$ as $\mathbf{G}_f(\cdot, \mathbf{p})$ satisfies the requirements of Theorem 4.2.3. Since $\mathbf{J_v f}_i$ is continuous, the right-hand sides of the differential equations defined by 4.2.2 are continuous as long as the trajectory tracks the boundary of a U_i under this choice.

In order to integrate these equations accurately, the time points at which $(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})) \in \Omega_m(\mathbf{f})$ hold need to be determined. At these time points, $g_i(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}) = 0$ holds for some i. *State event location algorithms* can be used to detect these time points of discontinuity using the discontinuity functions. At the end of this chapter, a state event location algorithm is reviewed that is used in the remainder of this thesis.

In the remainder of this thesis, following assumption holds.

Assumption 5.2.1. *Let* $(\mathbf{x}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p}),\dot{\mathbf{x}}(t,\mathbf{p}))$ *be the solution of* (4.3.1). *Then,* $(t_0,t_f] = \bigcup_{k=1}^{n(\mathbf{p})} T_k$ *where* $n(\mathbf{p})$ *is an integer and depends on* $\mathbf{p}$; $T_k = (\alpha_k,\beta_k]$ *and* $\alpha_k < \beta_k$ *for all* $k \in \{1,\ldots,n(\mathbf{p})\}$; $\alpha_1 = t_0$, $\beta_{n(\mathbf{p})} = t_f$; $\alpha_k = \beta_{k-1}$ *for all* $k \in \{2,\ldots,n(\mathbf{p})\}$ *and* $m(t,\mathbf{p}) = s_k$, $\forall t \in T_k$, $s_k \in \mathcal{D}(\mathbf{F})$ *holds and the solution of* (4.3.1) *satisfies*

$$\mathbf{0} = \mathbf{F}_{m(t_k,\mathbf{p})}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p}),\dot{\mathbf{x}}(t,\mathbf{p})),\ \forall t \in T_k,\ \mathbf{x}(t_0,\mathbf{p}) = \mathbf{F}_0(\mathbf{p}). \tag{5.2.1}$$

for each $\mathbf{p} \in \mathcal{P}$.

This assumption is not unreasonable for the systems considered in this thesis. Furthermore, solutions of (4.2.1) and (4.3.1) exist for a parameter value irrespective of Assumption 5.2.1 because the right-hand sides of these equations are locally Lipschitz continuous. Issues that arise with discontinuous systems such as Zeno or chattering behavior [10] do not arise in these systems.

Remark 5.2.2. The quantities $n(\mathbf{p})$, $\{T_k\}_{k=1}^{n(\mathbf{p})}$ and $\{s_k\}_{k=1}^{n(\mathbf{p})}$ are not known *a priori*. They are determined by the state event location and non-singleton trajectory detection algorithm, which is discussed later in this chapter, during the integration of (4.3.1).

In the next section, computational improvements made possible by this assumption are discussed and then the state event location and non-singleton trajectory detection algorithm is presented.

5.3 Computational Improvements

In this section, computational improvements to the integration of an equation defined in (4.3.2) are discussed. The improvements apply to (3.3.4) as well. In the remainder of this

thesis, Assumption 5.2.1 is assumed to hold and all functions satisfy property M. In addition, Assumptions 4.3.1 and 4.3.2 hold.

Let k, s_k and $m(\cdot, \mathbf{p})$ be as defined in Assumption 5.2.1. Let $\{s'_k\}_{k=1}^{n(\mathbf{p})}$ be such that $s'_k \in \mathcal{D}(g)$ for all $k \in \{1, \ldots, n(\mathbf{p})\}$. Let $\mathbf{u}(t, \mathbf{p})$ be as defined in Remark 4.3.3. Then

$$\mathbf{JF}_{m(t,\mathbf{p})}(t, \mathbf{u}(t, \mathbf{p})) \in \partial^B \mathbf{F}(t, \mathbf{u}(t, \mathbf{p})) \tag{5.3.1}$$

holds per property M. Note that on T_k, $\mathbf{JF}_{s_k}(\cdot, \mathbf{u}(\cdot, \mathbf{p}))$ is a continuous function of time because $\mathbf{x}(\cdot, \mathbf{p})$, $\dot{\mathbf{x}}(\cdot, \mathbf{p})$ and $\mathbf{y}(\cdot, \mathbf{p})$ are continuous functions of time and $\mathbf{JF}_{s_k}$ is continuous. Let $M : T \times \mathcal{P} \Rightarrow \mathbb{R}^{(n_y+n_x)\times(1+n_p+n_x)}$ be defined by

$$M(t, \mathbf{p}) = \text{conv}\left(\{-\mathbf{B}^{-1}\mathbf{A},\ [\mathbf{A}\ \mathbf{B}] \in \partial^B \mathbf{F}(t, \mathbf{u}(t, \mathbf{p})), \mathbf{A} \in \mathbb{R}^{(n_y+n_x)\times(1+n_p+n_x)},\ \mathbf{B} \in \mathbb{R}^{(n_y+n_x)\times(n_y+n_x)}\}\right).$$

Let the subscript $\tilde{\mathbf{v}}$ be associated with the host space of $\mathcal{T} \times \mathcal{P} \times \mathcal{X}$. If $t \in T_k$,

$$\begin{aligned}
-\mathbf{J_q F}_{s_k}(t, \mathbf{u}(t, \mathbf{p}))^{-1} \mathbf{J}_{\tilde{\mathbf{v}}} \mathbf{F}_{s_k}(t, \mathbf{u}(t, \mathbf{p})) &\in M(t, \mathbf{p}), \\
-[\mathbf{0}\ \mathbf{I}_{n_x}] \mathbf{J_q F}_{s_k}(t, \mathbf{u}(t, \mathbf{p}))^{-1} \mathbf{J}_{\tilde{\mathbf{v}}} \mathbf{F}_{s_k}(t, \mathbf{u}(t, \mathbf{p})) &\in \Gamma\mathbf{f}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})), \qquad (5.3.2) \\
-[\mathbf{I}_{n_y}\ \mathbf{0}] \mathbf{J_q F}_{s_k}(t, \mathbf{u}(t, \mathbf{p}))^{-1} \mathbf{J}_{\tilde{\mathbf{v}}} \mathbf{F}_{s_k}(t, \mathbf{u}(t, \mathbf{p})) &\in \Gamma\mathbf{r}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})) \qquad (5.3.3)
\end{aligned}$$

hold where $\Gamma\mathbf{f}$, $\Gamma\mathbf{r}$ are as defined in Lemma 4.3.4, $\mathbf{J_q F}_{s_k}(t, \mathbf{u}(t, \mathbf{p})) \in \mathbb{R}^{(n_y+n_x)\times(n_y+n_x)}$ and $\mathbf{J}_{\tilde{\mathbf{v}}} \mathbf{F}_{s_k}(t, \mathbf{u}(t, \mathbf{p})) \in \mathbb{R}^{(n_y+n_x)\times(1+n_p+n_x)}$. Note that $\mathbf{J_q F}_{s_k}(\cdot, \mathbf{u}(\cdot, \mathbf{p}))^{-1} \mathbf{J}_{\tilde{\mathbf{v}}} \mathbf{F}_{s_k}(t, \mathbf{u}(\cdot, \mathbf{p}))$ is a continuous function on T_k. The inverse of $\mathbf{J_q F}_{s_k}(t, \mathbf{u}(t, \mathbf{p}))^{-1}$ exists per Assumption 4.3.2. $\mathbf{J_q F}_{s_k}(\cdot, \mathbf{u}(\cdot, \mathbf{p}))^{-1}$ is a continuous function of t because $\mathbf{J_q F}_{s_k}(\cdot, \mathbf{u}(\cdot, \mathbf{p}))$ is continuous on T_k (Theorem 9.8 in [96]). Hence, it is a measurable selection on T_k. Consider the set of equations

$$\dot{\mathbf{m}}(t, \mathbf{p}) = \boldsymbol{\pi}_{\mathbf{x}} \mathbf{G}_f(t, \mathbf{p}) \mathbf{m}(t, \mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}} \mathbf{G}_f(t, \mathbf{p}),$$

$$\mathbf{n}(t,\mathbf{p}) = \boldsymbol{\pi}_{\mathbf{x}}\mathbf{G}_r(t,\mathbf{p})\mathbf{m}(t,\mathbf{p}) + \boldsymbol{\pi}_{\mathbf{p}}\mathbf{G}_r(t,\mathbf{p}),$$

where the elements of these equations are defined in Theorem 4.2.3 and Corollary 4.3.7. Using the left-hand sides of (5.3.2) and (5.3.3) to define $\mathbf{G}_f$ and $\mathbf{G}_r$, the following is obtained

$$\begin{bmatrix} \dot{\mathbf{m}}(t,\mathbf{p}) \\ \mathbf{n}(t,\mathbf{p}) \end{bmatrix} = \boldsymbol{\pi}_{\mathbf{v}}\mathbf{H}(t,\mathbf{p}) \begin{bmatrix} \mathbf{m}(t,\mathbf{p}) \\ \mathbf{I}_{n_p} \end{bmatrix} \tag{5.3.4}$$

where $\mathbf{H}(t,\mathbf{p}) = -\mathbf{J}_{\mathbf{q}}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p}))^{-1}\mathbf{J}_{\tilde{\mathbf{v}}}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p}))$ and $\boldsymbol{\pi}_{\mathbf{v}}\mathbf{H}(t,\mathbf{p})$ represents the last $n_p + n_x$ columns of $\mathbf{H}(t,\mathbf{p})$. Pre-multiplying both sides with $\mathbf{J}_{\mathbf{q}}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p}))$, the following is obtained

$$\mathbf{J}_{\mathbf{q}}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p})) \begin{bmatrix} \dot{\mathbf{m}}(t,\mathbf{p}) \\ \mathbf{n}(t,\mathbf{p}) \end{bmatrix} = -\boldsymbol{\pi}_{\mathbf{v}}\mathbf{J}_{\tilde{\mathbf{v}}}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p})) \begin{bmatrix} \mathbf{m}(t,\mathbf{p}) \\ \mathbf{I}_{n_p} \end{bmatrix}$$

where $\boldsymbol{\pi}_{\mathbf{v}}\mathbf{J}_{\tilde{\mathbf{v}}}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p}))$ is the last $n_p + n_x$ columns of $\mathbf{J}_{\tilde{\mathbf{v}}}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p}))$. The final form of the equations is

$$\begin{aligned} \mathbf{J}_{\mathbf{y}}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p}))\mathbf{n}(t,\mathbf{p}) + \mathbf{J}_{\dot{\mathbf{x}}}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p}))\dot{\mathbf{m}}(t,\mathbf{p}) + \\ \mathbf{J}_{\mathbf{x}}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p}))\mathbf{m}(t,\mathbf{p}) + \mathbf{J}_{\mathbf{p}}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p})) = \mathbf{0}. \end{aligned} \tag{5.3.5}$$

Equation (5.3.5) can be solved efficiently for the unknowns $\mathbf{n}(t,\mathbf{p})$ and $\dot{\mathbf{m}}(t,\mathbf{p})$ without explicitly inverting $\mathbf{J}_{\mathbf{q}}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p}))$ using the numerical method described in [36].

Reverse integration of the quantities in Theorems 4.3.10 and 4.3.11 can be achieved without explicitly inverting matrices as well.

Consider the integral in Theorem 4.3.9. An element of ΓG defined in Theorem 4.3.9 can

be computed using the equation

$$\dot{\mathbf{z}}(t,\mathbf{p}) = \mathbf{J}_{\mathbf{x}} g_{s'_k}(t,\mathbf{u}(t,\mathbf{p}))\mathbf{m}(t,\mathbf{p}) + \mathbf{J}_{\mathbf{y}} g_{s'_k}(t,\mathbf{u}(t,\mathbf{p}))\mathbf{n}(t,\mathbf{p}) + \tag{5.3.6}$$

$$\mathbf{J}_{\dot{\mathbf{x}}} g_{s'_k}(t,\mathbf{u}(t,\mathbf{p}))\dot{\mathbf{m}}(t,\mathbf{p}) + \mathbf{J}_{\mathbf{p}} g_{s'_k}(t,\mathbf{u}(t,\mathbf{p})),\ \forall t \in T_k,\ \forall k \in \{1,\ldots,n(\mathbf{p})\},$$

$$\mathbf{z}(t_0,\mathbf{p}) = \mathbf{0},$$

where $\mathbf{m}$, $\mathbf{n}$ and $\dot{\mathbf{m}}$ are computed using (5.3.5). Note that substituting the expressions for $\mathbf{n}$ and $\dot{\mathbf{m}}$, equations of the form in Theorem 4.3.9 are obtained.

Consider the following integral

$$\int_{\alpha_k}^{\beta_k} [\mathbf{J}_{\mathbf{x}} g_{s'_k}(t,\mathbf{u}(t,\mathbf{p}))\ \mathbf{J}_{\mathbf{y}} g_{s'_k}(t,\mathbf{u}(t,\mathbf{p})] \begin{bmatrix} \mathbf{m}(t,\mathbf{p}) \\ \mathbf{n}(t,\mathbf{p}) \end{bmatrix} + \tag{5.3.7}$$

$$[\mathbf{J}_{\dot{\mathbf{x}}} g_{s'_k}(t,\mathbf{u}(t,\mathbf{p}))\ \mathbf{J}_{\dot{\mathbf{y}}} g_{s'_k}(t,\mathbf{u}(t,\mathbf{p})] \begin{bmatrix} \dot{\mathbf{m}}(t,\mathbf{p}) \\ \dot{\mathbf{n}}(t,\mathbf{p}) \end{bmatrix} + \mathbf{J}_{\mathbf{p}} g_{s'_k}(t,\mathbf{u}(t,\mathbf{p}))dt$$

where the integrand is obtained by the application of Theorem 2.8.12 to $g_{s'_k}$ and $\mathbf{m}$, $\dot{\mathbf{m}}$ and $\mathbf{n}$ are computed using (5.3.5). Note that $\dot{\mathbf{n}} : T \times \mathcal{P} \to \mathbb{R}^{n_y \times n_p}$ and $\mathbf{J}_{\dot{\mathbf{y}}} g_{s'_k} : T \to \mathbb{R}^{1 \times n_y}$ are used as place holders. In addition $\mathbf{J}_{\dot{\mathbf{y}}} g_{s'_k}(t) = \mathbf{0}, \forall t \in T_k$ since g does not depend on $\dot{\mathbf{y}}$. Let $\boldsymbol{\lambda}_k : [\alpha_k, \beta_k] \to \mathbb{R}^{1 \times (n_x + n_y)}$ be an absolutely continuous function. Let $\boldsymbol{\lambda}_k(t) = (\boldsymbol{\lambda}_{k,x}(t), \boldsymbol{\lambda}_{k,y}(t))$ where $\boldsymbol{\lambda}_{k,x}(t) \in \mathbb{R}^{1 \times n_x}$ and $\boldsymbol{\lambda}_{k,y}(t) \in \mathbb{R}^{1 \times n_y}$. Appending (5.3.5) to (5.3.7) using $\boldsymbol{\lambda}_k$, the following is obtained:

$$\int_{\alpha_k}^{\beta_k} [\mathbf{J}_{\mathbf{x}} g_{s'_k}(t,\mathbf{u}(t,\mathbf{p}))\ \mathbf{J}_{\mathbf{y}} g_{s'_k}(t,\mathbf{u}(t,\mathbf{p}))] \begin{bmatrix} \mathbf{m}(t,\mathbf{p}) \\ \mathbf{n}(t,\mathbf{p}) \end{bmatrix} + \tag{5.3.8}$$

$$[\mathbf{J}_{\dot{\mathbf{x}}} g_{s'_k}(t,\mathbf{u}(t,\mathbf{p}))\ \mathbf{J}_{\dot{\mathbf{y}}} g_{s'_k}(t,\mathbf{u}(t,\mathbf{p}))] \begin{bmatrix} \dot{\mathbf{m}}(t,\mathbf{p}) \\ \dot{\mathbf{n}}(t,\mathbf{p}) \end{bmatrix} + \mathbf{J}_{\mathbf{p}} g_{s'_k}(t,\mathbf{u}(t,\mathbf{p})) -$$

$$\boldsymbol{\lambda}_k(t)[\mathbf{J}_{\mathbf{x}}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p}))\ \mathbf{J}_{\mathbf{y}}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p}))]\begin{bmatrix}\mathbf{m}(t,\mathbf{p})\\ \mathbf{n}(t,\mathbf{p})\end{bmatrix} -$$

$$\boldsymbol{\lambda}_k(t)[\mathbf{J}_{\dot{\mathbf{x}}}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p}))\ \mathbf{J}_{\dot{\mathbf{y}}}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p}))]\begin{bmatrix}\dot{\mathbf{m}}(t,\mathbf{p})\\ \dot{\mathbf{n}}(t,\mathbf{p})\end{bmatrix} -$$

$$\boldsymbol{\lambda}_k(t)\mathbf{J}_{\mathbf{p}}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p}))dt.$$

Note that $\mathbf{J}_{\dot{\mathbf{y}}}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p})) = \mathbf{0}$ because $\mathbf{F}_{s_k}$ is not a function of $\dot{\mathbf{y}}$. Using integration by parts (the arguments of the functions are omitted for clarity) results in

$$\int_{\alpha_k}^{\beta_k}[\mathbf{J}_{\dot{\mathbf{x}}}g_{s'_k}\ \mathbf{J}_{\dot{\mathbf{y}}}g_{s'_k}]\begin{bmatrix}\dot{\mathbf{m}}\\ \dot{\mathbf{n}}\end{bmatrix} - \boldsymbol{\lambda}_k[\mathbf{J}_{\dot{\mathbf{x}}}\mathbf{F}_{s_k}\ \mathbf{J}_{\dot{\mathbf{y}}}\mathbf{F}_{s_k}]\begin{bmatrix}\dot{\mathbf{m}}\\ \dot{\mathbf{n}}\end{bmatrix}dt = \tag{5.3.9}$$

$$\int_{\alpha_k}^{\beta_k}([\mathbf{J}_{\dot{\mathbf{x}}}g_{s'_k}\ \mathbf{J}_{\dot{\mathbf{y}}}g_{s'_k}] - \boldsymbol{\lambda}_k[\mathbf{J}_{\dot{\mathbf{x}}}\mathbf{F}_{s_k}\ \mathbf{J}_{\dot{\mathbf{y}}}\mathbf{F}_{s_k}])\begin{bmatrix}\dot{\mathbf{m}}\\ \dot{\mathbf{n}}\end{bmatrix}dt = \tilde{\boldsymbol{\lambda}}_k\begin{bmatrix}\mathbf{m}\\ \mathbf{n}\end{bmatrix}\Bigg|_{\alpha_k}^{\beta_k} - \int_{\alpha_k}^{\beta_k}\dot{\tilde{\boldsymbol{\lambda}}}_k\begin{bmatrix}\mathbf{m}\\ \mathbf{n}\end{bmatrix}dt. \tag{5.3.10}$$

where $\tilde{\boldsymbol{\lambda}}_k = ([\mathbf{J}_{\dot{\mathbf{x}}}g_{s'_k}\ \mathbf{J}_{\dot{\mathbf{y}}}g_{s'_k}] - \boldsymbol{\lambda}_k[\mathbf{J}_{\dot{\mathbf{x}}}\mathbf{F}_{s_k}\ \mathbf{J}_{\dot{\mathbf{y}}}\mathbf{F}_{s_k}])$. Let $\tilde{\boldsymbol{\lambda}}_k(t) = (\tilde{\boldsymbol{\lambda}}_{k,x}(t),\ \tilde{\boldsymbol{\lambda}}_{k,y}(t))$ where $\tilde{\boldsymbol{\lambda}}_{k,x}(t) \in \mathbb{R}^{1\times n_x}$ and $\tilde{\boldsymbol{\lambda}}_{k,y}(t) \in \mathbb{R}^{1\times n_y}$. Substituting the expression in (5.3.10) into (5.3.8) produces

$$\int_{\alpha_k}^{\beta_k}([\mathbf{J}_{\mathbf{x}}g_{s'_k}\ \mathbf{J}_{\mathbf{y}}g_{s'_k}] - \boldsymbol{\lambda}_k[\mathbf{J}_{\mathbf{x}}\mathbf{F}_{s_k}\ \mathbf{J}_{\mathbf{y}}\mathbf{F}_{s_k}] - \dot{\tilde{\boldsymbol{\lambda}}}_k)\begin{bmatrix}\mathbf{m}\\ \mathbf{n}\end{bmatrix} + \mathbf{J}_{\mathbf{p}}g_{s'_k} - \boldsymbol{\lambda}_k\mathbf{J}_{\mathbf{p}}\mathbf{F}_{s_k}dt + \tilde{\boldsymbol{\lambda}}_k\begin{bmatrix}\mathbf{m}\\ \mathbf{n}\end{bmatrix}\Bigg|_{\alpha_k}^{\beta_k}.$$

This expression is simplified by setting

$$\dot{\tilde{\boldsymbol{\lambda}}}_k = -\boldsymbol{\lambda}_k[\mathbf{J}_{\mathbf{x}}\mathbf{F}_{s_k}\ \mathbf{J}_{\mathbf{y}}\mathbf{F}_{s_k}] + [\mathbf{J}_{\mathbf{x}}g_{s'_k}\ \mathbf{J}_{\mathbf{y}}g_{s'_k}],\ \forall t \in [\alpha_k, \beta_k) \tag{5.3.11}$$

$$\dot{\tilde{\boldsymbol{\lambda}}}_{k,x} = -\boldsymbol{\lambda}_{k,x}\mathbf{J}_{\mathbf{x}}\mathbf{F}_{s_k} + \mathbf{J}_{\mathbf{x}}g_{s'_k},\ \forall t \in [\alpha_k, \beta_k)$$

$$\dot{\tilde{\boldsymbol{\lambda}}}_{k,y} = -\boldsymbol{\lambda}_{k,y}\mathbf{J}_{\mathbf{y}}\mathbf{F}_{s_k} + \mathbf{J}_{\mathbf{y}}g_{s'_k} = \mathbf{0},\ \forall t \in [\alpha_k, \beta_k).$$

As a result, the equations

$$\tilde{\boldsymbol{\lambda}}_{k,x} - \mathbf{J}_{\dot{\mathbf{x}}} g_{s'_k} = \boldsymbol{\lambda}_{k,x} \mathbf{J}_{\dot{\mathbf{x}}} \mathbf{F}_{s_k},$$
$$\mathbf{J}_{\mathbf{y}} g_{s'_k} = \boldsymbol{\lambda}_{k,y} \mathbf{J}_{\mathbf{y}} \mathbf{F}_{s_k},$$

are solved to determine $\boldsymbol{\lambda}_k$. Since $[\mathbf{J}_{\mathbf{y}} \mathbf{F}_{s_k} \ \mathbf{J}_{\dot{\mathbf{x}}} \mathbf{F}_{s_k}]$ is nonsingular per Assumption 5.1.1, the above equations have a unique solution which is

$$\boldsymbol{\lambda}_k = [\tilde{\boldsymbol{\lambda}}_{k,x} - \mathbf{J}_{\dot{\mathbf{x}}} g_{s'_k} \ \mathbf{J}_{\mathbf{y}} g_{s'_k}][\mathbf{J}_{\dot{\mathbf{x}}} \mathbf{F}_{s_k} \ \mathbf{J}_{\mathbf{y}} \mathbf{F}_{s_k}]^{-1}.$$

Substituting the solution into (5.3.11), the equation

$$\dot{\tilde{\boldsymbol{\lambda}}}_k = [\tilde{\boldsymbol{\lambda}}_{k,x} - \mathbf{J}_{\dot{\mathbf{x}}} g_{s'_k} \ \mathbf{J}_{\mathbf{y}} g_{s'_k}][\mathbf{J}_{\dot{\mathbf{x}}} \mathbf{F}_{s_k} \ \mathbf{J}_{\mathbf{y}} \mathbf{F}_{s_k}]^{-1}[\mathbf{J}_{\mathbf{x}} \mathbf{F}_{s_k} \ \mathbf{J}_{\mathbf{y}} \mathbf{F}_{s_k}] + [\mathbf{J}_{\mathbf{x}} g_{s'_k} \ \mathbf{J}_{\mathbf{y}} g_{s'_k}]$$

is obtained. Multiplying out the terms, a linear ordinary differential equation for $\tilde{\boldsymbol{\lambda}}_{k,x}$ is obtained with continuous right-hand side. Hence $\tilde{\boldsymbol{\lambda}}_{k,x}$ is an absolutely continuous function. Therefore $\tilde{\boldsymbol{\lambda}}_k$ is an absolutely continuous function. As a result, $\boldsymbol{\lambda}_k$ is absolutely continuous as defined.

If $k = n(\mathbf{p})$, let $\tilde{\boldsymbol{\lambda}}_k(\beta_k) = \mathbf{0}$, otherwise let $\tilde{\boldsymbol{\lambda}}_k(\beta_k) = \tilde{\boldsymbol{\lambda}}_{k+1}(\alpha_{k+1})$. Then (5.3.6) can be written as

$$\dot{\mathbf{z}}(t, \mathbf{p}) = \mathbf{J}_{\mathbf{p}} g_{s'_k}(t, \mathbf{u}(t, \mathbf{p})) - \tag{5.3.12}$$
$$\boldsymbol{\lambda}_k(t) \mathbf{J}_{\mathbf{p}} \mathbf{F}_{s_k}(t, \mathbf{u}(t, \mathbf{p})), \ \forall t \in (\alpha_k, \beta_k], \ \forall k \in \{1, \ldots, n(\mathbf{p})\}, \ \mathbf{z}(t_0, \mathbf{p}) = \mathbf{0},$$
$$\dot{\tilde{\boldsymbol{\lambda}}}_k(t) = -\boldsymbol{\lambda}_k(t)[\mathbf{J}_{\mathbf{x}} \mathbf{F}_{s_k}(t, \mathbf{u}(t, \mathbf{p})) \ \mathbf{J}_{\mathbf{y}} \mathbf{F}_{s_k}(t, \mathbf{u}(t, \mathbf{p}))] +$$
$$[\mathbf{J}_{\mathbf{x}} g_{s'_k}(t, \mathbf{u}(t, \mathbf{p})) \ \mathbf{J}_{\mathbf{y}} g_{s'_k}(t, \mathbf{u}(t, \mathbf{p}))], \ \forall t \in [\alpha_k, \beta_k), \ \forall k \in \{1, \ldots, n(\mathbf{p})\},$$

$$\tilde{\boldsymbol{\lambda}}_k(\beta_k) = \tilde{\boldsymbol{\lambda}}_{k+1}(\alpha_{k+1}),\ \forall k \in \{1,\ldots,n(\mathbf{p})-1\},\ \tilde{\boldsymbol{\lambda}}_{n(\mathbf{p})}(\beta_{n(\mathbf{p})}) = \mathbf{0},$$

$$\tilde{\boldsymbol{\lambda}}_{k,x}(t) = \boldsymbol{\lambda}_{k,x}(t)\mathbf{J}_{\dot{\mathbf{x}}}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p})) + \mathbf{J}_{\dot{\mathbf{x}}}g_{s'_k}(t,\mathbf{u}(t,\mathbf{p})),\ \forall t \in [\alpha_k,\beta_k),\ \forall k \in \{1,\ldots,n(\mathbf{p})\},$$

$$\mathbf{0} = \boldsymbol{\lambda}_{k,y}(t)\mathbf{J}_{\mathbf{y}}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p})) - \mathbf{J}_{\mathbf{y}}g_{s'_k}(t,\mathbf{u}(t,\mathbf{p})),\ \forall t \in [\alpha_k,\beta_k),\ \forall k \in \{1,\ldots,n(\mathbf{p})\}.$$

Consider the case in Theorem 4.3.11. Let $\mathbf{M_p}$, $\mathbf{M_x}$, $\mathbf{M_y}$ and $\mathbf{M_{\dot{x}}}$ be constants as defined in Theorem 4.3.11. Consider the integral

$$\mathbf{z}(t_f,\mathbf{p}) = \int_{t_0}^{t_f} \mathbf{M_x}\mathbf{m}(t,\mathbf{p}) + \mathbf{M_y}\mathbf{n}(t,\mathbf{p}) + \mathbf{M_{\dot{x}}}\dot{\mathbf{m}}(t,\mathbf{p}) + \mathbf{M_p}dt$$

where $\mathbf{m}$, $\mathbf{n}$ and $\dot{\mathbf{m}}$ are computed using (5.3.5). Note that $\dot{\mathbf{z}}(t_f,\mathbf{p})$ is the quantity of interest which is an element of ΓG defined in Theorem 4.3.11. Converting the integral to the form in (5.3.8), the following integral on T_k is obtained

$$\begin{aligned}\int_{\alpha_k}^{\beta_k} &[\mathbf{M_x}\ \mathbf{M_y}]\begin{bmatrix}\mathbf{m}(t,\mathbf{p})\\ \mathbf{n}(t,\mathbf{p})\end{bmatrix} + [\mathbf{M_{\dot{x}}}\ \mathbf{M_{\dot{y}}}]\begin{bmatrix}\dot{\mathbf{m}}(t,\mathbf{p})\\ \dot{\mathbf{n}}(t,\mathbf{p})\end{bmatrix} + \mathbf{M_p} - \\ &\boldsymbol{\lambda}_k(t)[\mathbf{J_x}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p}))\ \mathbf{J_y}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p}))]\begin{bmatrix}\mathbf{m}(t,\mathbf{p})\\ \mathbf{n}(t,\mathbf{p})\end{bmatrix} - \\ &\boldsymbol{\lambda}_k(t)[\mathbf{J_{\dot{x}}}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p}))\ \mathbf{J_{\dot{y}}}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p}))]\begin{bmatrix}\dot{\mathbf{m}}(t,\mathbf{p})\\ \dot{\mathbf{n}}(t,\mathbf{p})\end{bmatrix} - \\ &\boldsymbol{\lambda}_k(t)\mathbf{J_p}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p}))dt,\end{aligned} \tag{5.3.13}$$

where $\mathbf{M_{\dot{y}}} \in \mathbb{R}^{1\times n_y}$, $\mathbf{M_{\dot{y}}} = \mathbf{0}$ and $\mathbf{J_{\dot{y}}}\mathbf{F}_{s_k}(\cdot,\mathbf{u}(\cdot,\mathbf{p})) = \mathbf{0}$ for all $t \in T_k$. Following the same procedure as before, the following equations are obtained:

$$\dot{\mathbf{z}}(t,\mathbf{p}) = \mathbf{M_p} - \boldsymbol{\lambda}_k(t)\mathbf{J_p}\mathbf{F}_{s_k}(t,\mathbf{u}(t,\mathbf{p})),\ \forall t \in (\alpha_k,\beta_k],\ \forall k \in \{1,\ldots,n(\mathbf{p})\}, \tag{5.3.14}$$

$$\mathbf{z}(t_0, \mathbf{p}) = \mathbf{0},$$

$$\dot{\tilde{\boldsymbol{\lambda}}}_k(t) = -\boldsymbol{\lambda}_k(t)[\mathbf{J_x F}_{s_k}(t, \mathbf{u}(t, \mathbf{p}))\ \mathbf{J_y F}_{s_k}(t, \mathbf{u}(t, \mathbf{p}))] +$$

$$[\mathbf{M_x}\ \mathbf{M_y}],\ \forall t \in [\alpha_k, \beta_k),\ \forall k \in \{1, \ldots, n(\mathbf{p})\},$$

$$\tilde{\boldsymbol{\lambda}}_k(\beta_k) = \tilde{\boldsymbol{\lambda}}_{k+1}(\alpha_{k+1}),\ \forall k \in \{1, \ldots, n(\mathbf{p}) - 1\},\ \tilde{\boldsymbol{\lambda}}_{n(\mathbf{p})}(\beta_{n(\mathbf{p})}) = \mathbf{0},$$

$$\tilde{\boldsymbol{\lambda}}_{k,x}(t) = \boldsymbol{\lambda}_{k,x}(t)\mathbf{J}_{\dot{\mathbf{x}}}\mathbf{F}_{s_k}(t, \mathbf{u}(t, \mathbf{p})) + \mathbf{M}_{\dot{\mathbf{x}}},\ \forall t \in [\alpha_k, \beta_k),\ \forall k \in \{1, \ldots, n(\mathbf{p})\},$$

$$\mathbf{M_y} = \boldsymbol{\lambda}_{k,y}(t)\mathbf{J_y F}_{s_k}(t, \mathbf{u}(t, \mathbf{p})),\ \forall t \in [\alpha_k, \beta_k),\ \forall k \in \{1, \ldots, n(\mathbf{p})\}.$$

5.4 State Event Location and Detection of Non-singleton Trajectories

The governing equations for systems with varying structure are implemented in a programming language using logical expressions and if-then-else statements in order to compute numerical values. Each logical condition corresponds to a discontinuity function. If the value of the discontinuity function is less than or equal to zero, the logical condition is assumed to be true and false otherwise. In order to determine $m(t_k, \mathbf{p})$, the state of these logical conditions need to be determined.

The *set of active discontinuity functions* is the set of discontinuity functions corresponding to the logical expressions that need to be checked in order to evaluate the right-hand sides of (3.3.1) or (4.3.1) given $(t, \mathbf{u}(t, \mathbf{p}))$. The *set of active discontinuity function indices* is the set of pairs corresponding to the indices of these functions and is denoted by $\mathcal{A}$. The set of active discontinuity function indices is constant for all $t \in T_k$. Therefore, the set of active function indices corresponding to T_k is denoted by $\mathcal{A}_k$.

Let $T_k = (t_k, t_{k+1}]$. When $t = t_k$, $\mathcal{A}_k$ can be determined. However, at this point in time,

t_{k+1} is not known. t_{k+1} is the time at which $\mathcal{A}_k$ changes and it is not known a priori. In order to determine t_{k+1}, the dynamics of the system are integrated for a time step assuming $\mathcal{A}_k$ is constant. The integration of the dynamics assuming $\mathcal{A}_k$ is constant is called *discontinuity-locked integration* [83]. Then, the discontinuity functions are analyzed in order to determine whether the state of any of the logical conditions determining $\mathcal{A}_k$ changed during this time step. The state of a logical condition changes if the corresponding discontinuity function crosses zero. If any of these discontinuity functions crosses zero, the infimum of the times at which such a crossing occurs is determined. The infimum corresponds to one of the roots of the active discontinuity functions. This infimum is the value of t_{k+1}. The process of determining this infimum value is called *state event location.*

Consider the time interval, $[t_k, t_{k+1}]$, during the integration of the dynamics. After appending (5.3.5) and the active discontinuity functions to the system dynamics,

$$\mathbf{0} = \mathbf{F}_{m(t_k,\mathbf{p})}(t, \mathbf{u}(t,\mathbf{p})),\ \forall t \in (t_k, t_{k+1}] \tag{5.4.1}$$

$$z_{v,w}(t,\mathbf{p}) = g_{v,w}(t, \mathbf{u}(t,\mathbf{p})), \forall (v,w) \in \mathcal{A}_k,$$

$$\mathbf{0} = \mathbf{J}_{\mathbf{y}}\mathbf{F}_{m(t_k,\mathbf{p})}(t, \mathbf{u}(t,\mathbf{p}))\mathbf{n}(t,\mathbf{p}) + \mathbf{J}_{\dot{\mathbf{x}}}\mathbf{F}_{m(t_k,\mathbf{p})}(t, \mathbf{u}(t,\mathbf{p}))\dot{\mathbf{m}}(t_k,\mathbf{p}) +$$

$$\mathbf{J}_{\mathbf{x}}\mathbf{F}_{m(t_k,\mathbf{p})}(t, \mathbf{u}(t,\mathbf{p}))\mathbf{m}(t,\mathbf{p}) + \mathbf{J}_{\mathbf{p}}\mathbf{F}_{m(t_k,\mathbf{p})}(t, \mathbf{u}(t,\mathbf{p})),\ \forall t \in (t_k, t_{k+1}]$$

is obtained. This set of equations is used for discontinuity-locked integration.

A variant of the state event location algorithm in [83] is used in this thesis to detect zero crossings. The only difference is the root finding algorithm used. Note that, the right-hand sides of the equations defining the elements of the linear Newton approximations and generalized Jacobians are discontinuous at these zero crossings.

The state event location algorithm in [83] makes use of the properties of the integration algorithms for differential-algebraic systems. In this thesis, *the backward differentiation formula* (BDF) ([4]) family of numerical integration algorithms is used. This family of inte-

grators use polynomials to interpolate the state trajectories and all $z_{v,w}(\cdot, \mathbf{p})$. Let q be the order of the polynomials used. Let $\tilde{z}_{v,w}$ be the polynomial that approximates $z_{v,w}(\cdot, \mathbf{p})$, on the time interval $[t_k, t_{k+1}]$. This polynomial is ([83])

$$\tilde{z}_{v,w}(t) = \sum_{l=0}^{q} \frac{h_k^l \nabla^l \tilde{z}_{v,w}(t_{k+1})}{l!} \left(\frac{t - t_{k+1}}{h_k} \right)^l, \ \forall (v,w) \in \mathcal{A}_k \tag{5.4.2}$$

where $h_k = t_{k+1} - t_k$ is the integration time step which is determined by the integration algorithm, ∇^l is the backward difference operator defined recursively such that $\nabla^0 z_{v,w}(t_{k+1}) = z_{v,w}(t_{k+1})$, $\nabla z_{v,w}(t_{k+1}) = z_{v,w}(t_{k+1}) - z_{v,w}(t_k)$ and $\nabla^l z_{v,w}(t_{k+1}) = \nabla^{l-1} z_{v,w}(t_{k+1}) - \nabla^{l-1} z_{v,w}(t_k)$. The state event location algorithm uses the roots of these interpolating polynomials to determine zero crossing times. The algorithm is discussed in detail in the following subsection.

The detection of non-singleton trajectories depends on the analysis of discontinuity functions on the interval $[t_k, t_{k+1}]$. If there exists an interval of time in $[t_k, t_{k+1}]$ such that one of the active discontinuity functions is zero for all t in that interval, then the solution is a non-singleton trajectory. Since the discontinuity functions are approximated by the interpolating polynomials defined in (5.4.2) on $[t_k, t_{k+1}]$, these polynomials are used instead of the discontinuity functions. A relaxed criterion introduced next is used to determine non-singleton trajectories because these polynomials are only approximations to the discontinuity functions.

Definition 5.4.1 (Numerically Non-singleton Trajectory). *The solution of* (4.3.1) *is a numerically non-singleton trajectory if there exists a* $k \in \{1, \ldots, n(\mathbf{p}) - 1\}$ *and interval* $\Delta \subset \mathrm{cl}\,(T_k)$, *such that for some* $(v,w) \in \mathcal{A}_k$, $|\tilde{z}_{v,w}(t, \mathbf{p})| \leq \epsilon_a$ *and* $|\dot{\tilde{z}}_{v,w}(t, \mathbf{p})| \leq \epsilon_a$ *hold for all* $t \in \Delta$ *where* ϵ_a *is the absolute error tolerance used in the BDF algorithm.*

The non-singleton trajectory detection algorithm is used to determine if the solution is a numerically non-singleton trajectory. The algorithm is discussed in Step 4 of the main algorithm presented in the next section.

5.4.1 The State Event Location and Non-singleton Trajectory Detection Algorithm

Step 1: Determine $m(t_k, \mathbf{p})$, $\mathcal{A}_k$ and t_{k+1} using the BDF algorithm and execute discontinuity locked integration of (5.4.1) to form the interpolating polynomials defined by (5.4.2).

Step 2: In order to determine if $m(t_k, \mathbf{p})$ needs to change at some $t \in [t_k, t_{k+1}]$, the real roots of the polynomials $\{\tilde{z}_{v,w} : (v, w) \in \mathcal{A}_k\}$ are used. The minimum of the real roots of these polynomials is used as a candidate time at which $m(t_k, \mathbf{p})$ changes on $[t_k, t_{k+1}]$. There are two classes of polynomials to consider:

(a) $\mathcal{A}_k^0 = \{(v, w) \in \mathcal{A}_k : \tilde{z}_{v,w}$ is a zero polynomial$\}$.

(b) $\mathcal{A}_k^1 = \{(v, w) \in \mathcal{A}_k : \tilde{z}_{v,w}$ is not a zero polynomial$\}$.

Let $\tau_{ind((v,w),\mathcal{A}_k)}$ represent the minimum of the roots of the polynomial $\tilde{z}_{v,w}$. Let $\tau_{ind((v,w),\mathcal{A}_k)} = +\infty$ for all $(v, w) \in \mathcal{A}_k^0$. These discontinuity functions imply that $(t, \mathbf{u}(t, \mathbf{p}))$ lies on the boundary of an open set U_i as defined in Assumption 5.1.1 for $t \in [t_k, t_{k+1}]$ and therefore are ignored. If $\mathcal{A}_k^1$ is empty, then go to Step 4.

Determining $\tau_{ind((v,w),\mathcal{A}_k)}$ for $(v, w) \in \mathcal{A}_k^1$ is more complicated. In theory, one can use the Jenkins-Traub algorithm [51, 50] to compute all the roots of $\tilde{z}_{v,w}$, $(v, w) \in \mathcal{A}_k^1$. However, most of the time $\tilde{z}_{v,w}$ does not have any real roots or it has a unique zero and the application of the Jenkins-Traub algorithm incurs unnecessary computational cost. The algorithm described in this section uses elements from interval arithmetic [1] to reduce the number of times the Jenkins-Traub algorithm is applied. Note that the degree of $\tilde{z}_{v,w}$ can be at most 5 in this thesis because the BDF algorithm uses polynomials whose degree is at most 5.

The algorithm scales the polynomials so that their domains are $[0, 1]$ instead of $[t_k, t_{k+1}]$. Let the corresponding scaled polynomials be $s_{v,w} : [0, 1] \to \mathbb{R}$, $(v, w) \in \mathcal{A}_k^1$. Let q be the order of the polynomials and $\{C_n\}_{n=0}^q$ be their coefficients.

The root-finding algorithm applied to all $(v, w) \in \mathcal{A}_k^1$ is:

R.1. Let $0 < \Delta < 1$. Let $l = 1$, $a_l = 0$, $b_l = 1$, $D_l = [a_l, b_l]$ and $S = \{D_l\}$.

R.2. If $l > n(S)$, then stop. All intervals in S are analyzed.

R.3. Let

$$\mathbf{s}_{v,w}(D_l) = \sum_{n=0}^{q} C_n D_l^n$$

be the enclosure of $s_{v,w}$. The enclosure contains the range of $s_{v,w}$ for the domain D_l. If $0 \notin \mathbf{s}_{v,w}(D_l)$, then there is no root of $s_{v,w}$ on D_l. In this case, go to Step R.9. Otherwise, go to Step R.4.

R.4. Let

$$\dot{\mathbf{s}}_{v,w}(D_l) = \sum_{n=1}^{q} n \cdot C_n \cdot D_l^{n-1}$$

be the enclosure of $\dot{s}_{v,w}$. If $0 \in \dot{\mathbf{s}}_{v,w}(D_l)$, then the zeros of $s_{v,w}$ may not be regular i.e. they may have multiplicity greater than one. In this case, if $b_l - a_l \leq \Delta$, go to Step R.7 and if $b_l - a_l \geq \Delta$, go to Step R.8. If $0 \notin \dot{\mathbf{s}}_{v,w}(D_l)$, then go to Step R.5.

R.5. Consider the Krawcyk operator [72];

$$\mathbf{K}(D_l) = \text{mid}(D_l) - \frac{s_{v,w}(\text{mid}(D_l))}{\text{mid}(\dot{\mathbf{s}}_{v,w}(D_l))} + \left(1 - \frac{\dot{\mathbf{s}}_{v,w}(D_l)}{\text{mid}(\dot{\mathbf{s}}_{v,w}(D_l))}\right)(D_l - \text{mid}(D_l))$$

where mid is the midpoint operator to an interval. If $\mathbf{K}(D_l) \cap D_l = \emptyset$, then there exists no zero of the polynomial $s_{v,w}$ in D_l. In this case, go to Step R.9. If $\mathbf{K}(D_l) \subset \text{int}\,(D_l)$, then there exists a unique zero of $s_{v,w}$ in D_l. In this case, go to Step R.6.

R.6. Apply Newton's method to determine the unique zero in D_l. Go to Step R.9.

R.7. Apply the Jenkins-Traub algorithm to find all zeros of $s_{v,w}$ in D_l. Go to Step R.9.

R.8. Let $D_{n(S)+1} = [a_l, (a_l + b_l)/2]$, $D_{n(S)+2} = [(a_l + b_l)/2, b_l]$ and set $S = S \cup \{D_{n(S)+1}, D_{n(S)+2}\}$. Go to Step R.9.

R.9. Let $l = l + 1$ and go to Step R.2.

Set $\tau_{ind((v,w),\mathcal{A}_k)}$ to the minimum of the roots with odd multiplicity. If there are no such real roots, set $\tau_{ind((v,w),\mathcal{A}_k)} = +\infty$. Roots with even multiplicity imply that $(t, \mathbf{u}(t, \mathbf{p}))$ touches but does not cross the boundary of some U_i where U_i is as defined in Assumption 5.1.1.

Step 3: Let $\bar{\tau} = \min_{(v,w)\in\mathcal{A}_k} \{\tau_{ind((v,w),\mathcal{A}_k)}\}$. If $\bar{\tau} = +\infty$, set the event time, τ^* to t_{k+1} and go to Step 4. Otherwise, let $(\bar{v}, \bar{w})$ be such that $\tau_{ind((\bar{v},\bar{w}),\mathcal{A}_k)} = \bar{\tau}$. Compute a consistent event time, τ^* and $\boldsymbol{\eta} = (\tau^*, \boldsymbol{\eta}_\mathbf{p}, \boldsymbol{\eta}_\mathbf{x}, \boldsymbol{\eta}_\mathbf{y}, \boldsymbol{\eta}_{\dot{\mathbf{x}}})$ that satisfy

$$
\begin{aligned}
\mathbf{0} &= \mathbf{F}_{m(t_k,\mathbf{p})}(\boldsymbol{\eta}), \\
\boldsymbol{\eta}_\mathbf{p} &= \mathbf{p}, \\
\boldsymbol{\eta}_{\dot{\mathbf{x}}} &= \frac{\alpha_0 \boldsymbol{\eta}_\mathbf{x}}{\tau^* - t_k} + \sum_{i=1}^{q} \frac{\alpha_i \mathbf{x}(t_{k+1-i}, \mathbf{p})}{\tau^* - t_k}, \\
\pm\epsilon_g &= g_{\bar{v},\bar{w}}(\boldsymbol{\eta})
\end{aligned}
$$

as in [83] where $\boldsymbol{\eta}_\mathbf{p} \in \mathbb{R}^{n_p}$, $\boldsymbol{\eta}_\mathbf{x} \in \mathbb{R}^{n_x}$, $\boldsymbol{\eta}_\mathbf{y} \in \mathbb{R}^{n_y}$, $\boldsymbol{\eta}_{\dot{\mathbf{x}}} \in \mathbb{R}^{n_x}$, $\{\alpha_i\}_{i=1}^q$ are the coefficients of the BDF method, ϵ_g and its sign are determined as in [83]. Set $t_{k+1} = \tau^*$. Update the polynomials $\tilde{z}_{v,w}$ for all $(v, w) \in \mathcal{A}_k$ using $\boldsymbol{\eta}$. Go to Step 2.

Step 4: If $\mathcal{A}_k^0$ is nonempty, then the trajectory is labeled as numerically non-singleton. Otherwise, the active discontinuity functions are analyzed to determine if the conditions of Definition 5.4.1 are satisfied for any of the active discontinuity functions.

The algorithm to check the conditions of Definition 5.4.1 employs interval arithmetic,

bisection and the Jenkins-Traub algorithm. The algorithm scales the polynomials so that their domains are $[0,1]$ instead of $[t_k, t_{k+1}]$. Let the corresponding scaled polynomials be $s_{v,w} : [0,1] \to \mathbb{R}$, $(v,w) \in \mathcal{A}_k^1$. Let q be the order of the polynomials and $\{C_n\}_{n=0}^q$ be their coefficients. The following algorithm is applied to all elements of $\{\tilde{z}_{v,w} : (v,w) \in \mathcal{A}_k^1\}$ to determine whether the solution is numerically non-singleton.

N.1 Let $0 < \Delta < 1$. Let $l = 1$, $a_l = 0$, $b_l = 1$, $D_l = [a_l, b_l]$ and $S = \{D_l\}$.

N.2 If $l > n(S)$, then stop. All intervals in S are analyzed.

N.3 Let

$$\mathbf{s}_{v,w}(D_l) = \sum_{n=0}^{q} C_n D_l^n$$

be the enclosure of $s_{v,w}$. There are three cases to consider

- $\mathbf{s}_{v,w}(D_l) \subset [-\epsilon_a, \epsilon_a]$. In this case, the condition on $s_{v,w}$ holds. Go to Step N.5 to check if the condition on the derivative holds.
- $[-\epsilon_a, \epsilon_a] \cap \mathbf{s}_{v,w}(D_l) = \emptyset$. In this case, the range of the polynomial corresponding to the domain D_l does not intersect $[-\epsilon_a, \epsilon_a]$. Go to Step N.7.
- $[-\epsilon_a, \epsilon_a] \cap \mathbf{s}_{v,w}(D_l) \neq \emptyset$ and $\mathbf{s}_{v,w}(D_l) \backslash [-\epsilon_a, \epsilon_a] \neq \emptyset$. In this case, the enclosure does not furnish enough information. If $b_l - a_l > \Delta$, go to Step N.6, otherwise go to go to Step N.4.

N.4 In this step, the maximum and minimum values attained by $s_{v,w}$ on D_l are analyzed. The extremal values are attained at $t = a_l$, $t = b_l$ and/or at $t \in D_l$ such that the necessary condition of optimality, $\dot{s}_{v,w}(t) = 0$ holds. In order to determine such t, the Jenkins-Traub algorithm is used to find the real roots of $\dot{s}_{v,w}$. Let $\bar{s}_{v,w}$ be the maximum and $\underline{s}_{v,w}$ denote the minimum values attained on the interval D_l.

If $[-\epsilon_a, \epsilon_a] \cap [\underline{s}_{v,w}, \bar{s}_{v,w}] \neq \emptyset$ go to Step N.5. Note that, this conditions is more

relaxed than the condition in Definition 5.4.1 for $z_{v,w}$. Otherwise, go to Step N.7.

N.5 The actions taken for $s_{v,w}$ are repeated in this step for $\dot{s}_{v,w}$.

N.5.1 Let $m = 1$, $c_m = a_l$, $d_m = b_l$, $D_{l,m} = [c_m, d_m]$ and $S_l = \{D_{l,m}\}$.

N.5.2 If $m > n(S_l)$, then stop. All intervals in S_l are analyzed.

N.5.3 Let

$$\dot{\mathbf{s}}_{v,w}(D_{l,m}) = \sum_{n=1}^{q} n \cdot C_n \cdot D_{l,m}^{n-1}$$

be the enclosure of $\dot{s}_{v,w}$. There are three cases to consider

- If $\dot{\mathbf{s}}_{v,w}(D_{l,m}) \subset [-\epsilon_a, \epsilon_a]$, the trajectory is numerically non-singleton.
- If $[-\epsilon_a, \epsilon_a] \cap \dot{\mathbf{s}}_{v,w}(D_{l,m}) = \emptyset$, go to N.5.6.
- If $[-\epsilon_a, \epsilon_a] \cap \dot{\mathbf{s}}_{v,w}(D_l) \neq \emptyset$ and $\dot{\mathbf{s}}_{v,w}(D_{l,m}) \backslash [-\epsilon_a, \epsilon_a] \neq \emptyset$, then there are two options. If $d_m - c_m > \Delta$, Go to Step N.5.5, otherwise go to N.5.4.

N.5.4 In this step, the maximum and minimum values attained by $\dot{s}_{v,w}$ on $D_{l,m}$ are analyzed. The extremal values are attained at $t = c_m$, $t = d_m$ and/or at $t \in D_{l,m}$ such that the necessary condition of optimality, $\ddot{s}_{v,w}(t) = 0$ holds. In order to determine such t, the Jenkins-Traub algorithm is used to find the real roots of $\ddot{s}_{v,w}$. Let $\bar{\dot{s}}_{v,w}$ be the maximum and $\underline{\dot{s}}_{v,w}$ denote the minimum values attained on the interval $D_{l,m}$.

If $[-\epsilon_a, \epsilon_a] \cap [\underline{\dot{s}}_{v,w}, \bar{\dot{s}}_{v,w}] \neq \emptyset$, then label the solution as numerically non-singleton. Note that, this conditions is more relaxed than the condition in Definition 5.4.1 for $\dot{z}_{v,w}$. Otherwise, go to Step N.7.

N.5.5 Let $D_{l,n(S)+1} = [c_m, (c_m + d_m)/2]$, $D_{l,n(S)+2} = [(c_m + d_m)/2, d_m]$ and set $S = S \cup \{D_{l,n(S)+1}, D_{l,n(S)+2}\}$. Go to Step N.5.6.

N.5.6 Let $m = m + 1$. Go to Step N.5.2.

N.6 Let $D_{n(S)+1} = [a_l, (a_l + b_l)/2]$, $D_{n(S)+2} = [(a_l + b_l)/2, b_l]$ and set $S = S \cup$

$\{D_{n(S)+1}, D_{n(S)+2}\}$. Go to Step N.7.

N.7 Let $l = l + 1$. Go to Step N.2.

Go to Step 5.

Step 5: Let $k = k + 1$. Go to Step 1.

The discussion on state event location and non-singleton trajectory detection directly applies to the multistage case in (4.4.1). It also directly applies to the integral in Theorem 4.3.9. This integral can be computed by appending the integrand to (4.3.1) and considering the augmented system.

5.5 Computational Methods Used in this Thesis

In the remainder of this thesis, the algorithm described in the previous section based on the algorithm in [83] is used for state event location. The dynamics in (5.2.1) and the auxiliary equations in (5.3.5) are integrated using the integration code DSL48SE ([36, 107, 108]). The quantities in (5.3.5) are derived from (4.3.1) using automatic differentiation [43] code DAEPACK [107]. The FORTRAN 77 code implementation of the equations in (4.3.1) is processed by DAEPACK to generate the FORTRAN 77 code representing the equations in (5.3.5) as well as the additional code to keep track of the states of the logical conditions in the code discussed in §5.4. The implementation of the infrastructure to solve (5.3.12) and (5.3.14) by reverse integration is a topic for future research. Implementations exist when the data of the problem is sufficiently smooth [24]; however, there are no software implementations that combine state event location and reverse integration.

Chapter 6

Bundle Method using Linear Newton Approximations

This chapter contains the development of a bundle-type nonsmooth optimization algorithm to numerically solve the mathematical program

$$J(\mathbf{p}) = \min_{\mathbf{p}\in\mathcal{P}} f(\mathbf{p}), \text{ s.t. } g_i(\mathbf{p}) \leq 0,\ i = 1, \ldots, n_c. \tag{6.0.1}$$

The bundle method developed in this chapter takes into account the fact that the generalized gradient of f and g_i cannot be computed at all $\mathbf{p} \in \mathcal{P}$. The algorithm uses linear Newton approximations (Definition 2.8.11) where the generalized gradient is not available.

In the remainder of this chapter, first, the necessary conditions of optimality for the Lipschitzian optimization problem (6.0.1) are summarized. Then, a bundle method using linear Newton approximations is developed.

6.1 Necessary Conditions of Optimality for Lipschitzian Optimization

Assumption 6.1.1. *In the remainder of this chapter, program* (6.0.1) *satisfies the following:*

1. n_c *and* n_p *are finite positive integers;*
2. $\mathcal{P}$ *is an open convex subset of* $\mathbb{R}^n$*;*
3. $f : \mathcal{P} \to \mathbb{R}$, *and* $g_i : \mathcal{P} \to \mathbb{R}$, $i = 1, \ldots, n_c$ *are locally Lipschitz continuous and semismooth functions.*

Definition 6.1.2 (Unconstrained Local and Global Minima). *The point* $\mathbf{p}^* \in \mathcal{P}$ *is an unconstrained local minimum of the program*

$$\min_{\mathbf{p} \in \mathcal{P}} f(\mathbf{p})$$

if there exists an $\epsilon > 0$ *such that* $f(\mathbf{p}^*) \leq f(\mathbf{p})$ *holds for all* $\mathbf{p} \in \mathcal{P}$ *satisfying* $\|\mathbf{p} - \mathbf{p}^*\| < \epsilon$. $\mathbf{p}^* \in \mathcal{P}$ *is an unconstrained global minimum if* $f(\mathbf{p}^*) \leq f(\mathbf{p})$ *holds for all* $\mathbf{p} \in \mathcal{P}$.

Theorem 6.1.3 (Necessary Conditions of Optimality for Unconstrained Optimization Problems). *Let Assumption 6.1.1 hold. If* $\mathbf{p}^*$ *is an unconstrained local minimum of*

$$\min_{\mathbf{p} \in \mathcal{P}} f(\mathbf{p}),$$

then $\mathbf{0} \in \partial f(\mathbf{p}^*)$ *holds (Proposition 2.3.2 in [25]).*

Theorem 6.1.4 (Direction of Descent for Unconstrained Optimization Problems). *Let Assumption 6.1.1 hold. Let* $\{\mathbf{v}\} = \arg\min\{\|\boldsymbol{\zeta}\| : \boldsymbol{\zeta} \in \partial f(\mathbf{p})\}$.[1] *Assume* $\mathbf{v} \neq \mathbf{0}$. *Let* $\mathbf{d} = -\mathbf{v}$. *Then* $f(\mathbf{p} + t\mathbf{d}) < f(\mathbf{p})$ *holds for small enough positive* t *(Lemmas 2.10, 2.11 and 2.12 in [54]).*

[1]Note that the solution of this program is the projection of the zero vector on the nonempty compact convex set $\partial f(\mathbf{p})$. Therefore, the solution exists and is unique per Theorem 2.4.6.

Definition 6.1.5 (Feasible Set). *The set of feasible points, S, is*

$$S = \{\mathbf{p} \in \mathcal{P} : \ g_i(\mathbf{p}) \leq 0, \ \forall i \in \{1, \ldots, n_c\}\}.$$

Assumption 6.1.6. *The feasible set, S, is nonempty.*

Definition 6.1.7 (Set of Active Constraints). *Let $\mathbf{p} \in S$. Then $\mathcal{I}(\mathbf{p})$, the set of active constraints at $\mathbf{p}$, is $\{i : \ g_i(\mathbf{p}) = 0, \ i \in \{i, \ldots, n_c\}\}$.*

Definition 6.1.8 (Constrained Local and Global Minima). *A point $\mathbf{p}^* \in S$ is a constrained local minimum of (6.0.1) if there exists an $\epsilon > 0$ such that $f(\mathbf{p}^*) \leq f(\mathbf{p})$ for all $\mathbf{p} \in S$ satisfying $\|\mathbf{p} - \mathbf{p}^*\| < \epsilon$ holds. $\mathbf{p}^* \in \mathcal{P}$ is a constrained global minimum if $f(\mathbf{p}^*) \leq f(\mathbf{p})$ holds for all $\mathbf{p} \in S$.*

Theorem 6.1.9 (Necessary Conditions of Optimality for Constrained Optimization Problems). *Let Assumption 6.1.1 hold. If $\mathbf{p}^*$ is a constrained local minimum of (6.0.1), then there exist numbers and μ_i, $i = 0, \ldots, n_c$ such that*

$$\mathbf{0} \in \mu_0 \partial f(\mathbf{p}^*) + \sum_{i=1}^{n_c} \mu_i \partial g_i(\mathbf{p}^*), \tag{6.1.1}$$

$$0 \neq |\mu_0| + \sum_{i=1}^{n_c} |\mu_i|, \tag{6.1.2}$$

$$0 = \mu_i g_i(\mathbf{p}^*), \ \forall i \in \{1, \ldots, n_c\}, \tag{6.1.3}$$

$$0 \leq \mu_i, \ \forall i \in \{0, \ldots, n_c\} \tag{6.1.4}$$

hold (Theorem 6.1.1 in [25]).

Since (6.1.2) holds, the additional condition

$$\sum_{i=0}^{n_c} \mu_i = 1$$

can be imposed. This condition, (6.1.1) and (6.1.4) clearly show that the zero vector is an element of the convex combinations of the elements of the generalized gradients of the constraints and the objective at a local minimum of (6.0.1).

In order to exclude cases where $\mu_0 = 0$ in Theorem 6.1.9, *constraint qualifications* need to be imposed on problem (6.0.1). If $\mu_0 = 0$, then the necessary conditions in Theorem 6.1.9 furnish no information about f. Two constraint qualifications relevant to the work in this thesis are as follows.

Definition 6.1.10 (Cottle Constraint Qualification). *Let $\bar{\mathbf{p}} \in S$. Then the Cottle constraint qualification holds at $\bar{\mathbf{p}}$ if either $g_i(\bar{\mathbf{p}}) < 0, \quad \forall i \in \{1, \ldots, n_c\}$ or $\mathbf{0} \notin \text{conv}\left(\bigcup_{i \in \mathcal{I}(\bar{\mathbf{p}})} \partial g_i(\bar{\mathbf{p}})\right)$.*

Definition 6.1.11 (Slater Constraint Qualification). *Let $\bar{\mathbf{p}} \in S$. Then the Slater constraint qualification holds for (6.0.1) at $\bar{\mathbf{p}}$ if g_i are convex functions for all $i \in \mathcal{I}(\bar{\mathbf{p}})$ and there exists a $\tilde{\mathbf{p}} \in S$ such that $g_i(\tilde{\mathbf{p}}) < 0$ for all $i \in \mathcal{I}(\bar{\mathbf{p}})$.*

The constrained optimization program (6.0.1) is transformed locally into an unconstrained optimization program using *the total constraint function* and *the improvement function.*

Definition 6.1.12 (Total Constraint Function). *The total constraint function, $G : \mathcal{P} \to \mathbb{R}$, is defined by*

$$G(\mathbf{p}) = \max\{g_i(\mathbf{p}), \ldots, g_{n_c}(\mathbf{p})\}.$$

It is a locally Lipschitz continuous function and its generalized gradient at $\mathbf{p}$ satisfies

$$\partial G(\mathbf{p}) \subset \partial \tilde{G}(\mathbf{p})$$

where $\partial \tilde{G}(\mathbf{p}) = \text{conv}\left(\bigcup_{i \in \mathcal{I}(\mathbf{p})} \partial g_i(\mathbf{p})\right)$ per (2.6.3).

Definition 6.1.13 (Improvement Function). *Let $\bar{\mathbf{p}} \in S$. Let the improvement function at $\bar{\mathbf{p}}$, $H : \mathcal{P} \to \mathbb{R}$, be defined by*

$$H(\mathbf{p}; \bar{\mathbf{p}}) = \max\{f(\mathbf{p}) - f(\bar{\mathbf{p}}), G(\mathbf{p})\}. \tag{6.1.5}$$

Then, H is a locally Lipschitz continuous function and

$$\partial H(\bar{\mathbf{p}}; \bar{\mathbf{p}}) \subset \tilde{M}(\bar{\mathbf{p}})$$

holds [54] where

$$\tilde{M}(\bar{\mathbf{p}}) = \begin{cases} \partial f(\bar{\mathbf{p}}) & \text{if } G(\bar{\mathbf{p}}) < 0 \\ \operatorname{conv}\left(\partial f(\bar{\mathbf{p}}) \cup \partial \tilde{G}(\bar{\mathbf{p}})\right) & \text{if } G(\bar{\mathbf{p}}) = 0 \\ \partial \tilde{G}(\bar{\mathbf{p}}) & \text{if } G(\bar{\mathbf{p}}) > 0. \end{cases}$$

Definition 6.1.14 (Stationary Point). *A point $\bar{\mathbf{p}} \in S$ that satisfies* (6.1.1), (6.1.2), (6.1.3) *and* (6.1.4) *or equivalently satisfies $\mathbf{0} \in \tilde{M}(\bar{\mathbf{p}})$ is called a stationary point of problem* (6.0.1) *[54].*

Theorem 6.1.15. *Let $\mathbf{p}^*$ be a constrained local minimum of* (6.0.1). *Then $\mathbf{0} \in \partial H(\mathbf{p}^*; \mathbf{p}^*)$ and $\mathbf{0} \in \tilde{M}(\mathbf{p}^*)$ hold. In addition, there exist scalars $\mu_i, i = 0, \ldots, n_c$ such that* (6.1.1), (6.1.2), (6.1.3) *and* (6.1.4) *hold (Lemma 2.15 in [54]).*

Corollary 6.1.16 (Descent Direction for the Improvement Function). *Let* $\{\mathbf{v}\} = \arg\min\{\|\boldsymbol{\zeta}\| : \boldsymbol{\zeta} \in \partial H(\bar{\mathbf{p}}; \bar{\mathbf{p}})\}$[2]. *Assume $\mathbf{v} \neq \mathbf{0}$. Let $\mathbf{d} = -\mathbf{v}$. Then for small enough positive t, $H(\bar{\mathbf{p}} + t\mathbf{d}; \bar{\mathbf{p}}) < H(\bar{\mathbf{p}}; \bar{\mathbf{p}})$ holds.*

The following is a restatement of Lemma 2.16 in [54].

[2]Note that the solution of this program is the projection of the zero vector on the nonempty compact convex set $\partial H(\bar{\mathbf{p}}; \bar{\mathbf{p}})$. Therefore, the solution exists and is unique per Theorem 2.4.6.

Theorem 6.1.17 (Necessary and Sufficient Conditions of Optimality for Convex Programs). *Let problem* (6.0.1) *be a convex program, i.e., f and g_i for $i = 1, \ldots, n_c$ are convex functions. Assume that the Slater constraint qualification (6.1.11) holds at $\bar{\mathbf{p}} \in S$. Then, the following are equivalent:*

1. *$\bar{\mathbf{p}}$ is a constrained global minimum of* (6.0.1);
2. *$\mathbf{0} \in \partial H(\bar{\mathbf{p}}; \bar{\mathbf{p}}) = \tilde{M}(\bar{\mathbf{p}})$ holds;*
3. *$\bar{\mathbf{p}}$ is a stationary point of* (6.0.1);
4. *the necessary conditions of optimality in Theorem* (6.1.9) *hold with $\mu_0 \neq 0$.*

6.2 Necessary Conditions of Optimality using Linear Newton Approximations

In this section, necessary conditions of optimality using linear Newton approximations are presented. In addition, a descent direction is obtained using linear Newton approximations.

Assumption 6.2.1. *Assume program* (6.0.1) *satisfies the following:*

1. *f and each g_i for $i = 1, \ldots, n_c$ are semismooth functions in the restricted sense as defined in Section 2.8.4.*

2. *There exists Γf, a linear Newton approximation of f, such that $\partial f(\mathbf{p}) \subset \Gamma f(\mathbf{p})$ holds for all $\mathbf{p} \in \mathcal{P}$. In addition, $\Gamma f(\mathbf{p})$ is a convex set for all $\mathbf{p} \in \mathcal{P}$.*

3. *For each g_i, there exists Γg_i, a linear Newton approximation of g_i such that $\partial g_i(\mathbf{p}) \subset \Gamma g_i(\mathbf{p})$ holds for all $\mathbf{p} \in \mathcal{P}$. In addition, $\Gamma g_i(\mathbf{p})$ is a convex set for all $\mathbf{p} \in \mathcal{P}$.*

Theorem 6.2.2 (Necessary Conditions of Optimality for Unconstrained Optimization Problems using Linear Newton Approximations). *Let Assumptions 6.1.1 and*

6.2.1 hold. If $\mathbf{p}^*$ *is a local minimum of the unconstrained optimization problem*

$$\min_{\mathbf{p}\in\mathcal{P}} f(\mathbf{p})$$

then $\mathbf{0} \in \Gamma f(\mathbf{p}^*)$ *holds.*

Proof. The result follows from Theorem 6.1.3 and Assumption 6.2.1 noting that $\partial f(\mathbf{p}^*) \subset \Gamma f(\mathbf{p}^*)$. □

Theorem 6.2.3 (Direction of Descent for Unconstrained Optimization Problems using Linear Newton Approximations). *Let Assumptions 6.1.1 and 6.2.1 hold. Let* $\{\mathbf{v}\} = \arg\min\{\|\boldsymbol{\zeta}\| : \boldsymbol{\zeta} \in \Gamma f(\mathbf{p})\}$. *Assume* $\mathbf{v} \neq \mathbf{0}$. *Let* $\mathbf{d} = -\mathbf{v}$. *Then for small enough positive* t, $f(\mathbf{p} + t\mathbf{d}) < f(\mathbf{p})$ *holds.*

Proof. $\Gamma f(\mathbf{p})$ is a convex and compact set therefore the element of minimum norm, $\mathbf{v}$, is well-defined, unique and satisfies the following per Theorem 2.4.6

$$\begin{aligned} &\mathbf{v}^{\mathrm{T}}\mathbf{u} \geq \mathbf{v}^{\mathrm{T}}\mathbf{v}, \ \forall \mathbf{u} \in \Gamma f(\mathbf{p}), \\ &\mathbf{d}^{\mathrm{T}}\mathbf{u} \leq \mathbf{d}^{\mathrm{T}}\mathbf{v} = -\|\mathbf{d}\|^2, \ \forall \mathbf{u} \in \Gamma f(\mathbf{p}). \end{aligned} \tag{6.2.1}$$

Let $t > 0$. Then

$$f(\mathbf{p} + t\mathbf{d}) - f(\mathbf{p}) = t\boldsymbol{\zeta}^{*\mathrm{T}}\mathbf{d} \tag{6.2.2}$$

holds for some $\boldsymbol{\zeta}^* \in \partial f(\mathbf{p}^*)$ where $\mathbf{p}^* = \mathbf{p} + \alpha\mathbf{d}$ and $0 < \alpha < t$ per the Mean Value Theorem for locally Lipschitz functions (Theorem 2.6.4).

Note that (6.2.2) implies

$$f(\mathbf{p} + t\mathbf{d}) - f(\mathbf{p}) \in \{t\boldsymbol{\zeta}^{\mathrm{T}}\mathbf{d} : \boldsymbol{\zeta} \in \partial f(\mathbf{p}^*)\}. \tag{6.2.3}$$

Since Γf is compact-valued and upper semicontinuous, for small enough t,

$$\Gamma f(\mathbf{p}^*) \subset \Gamma f(\mathbf{p}) + \frac{\|\mathbf{d}\|}{2}\mathbb{B}(0,1) \tag{6.2.4}$$

holds where $\mathbb{B}(0,1)$ is the unit ball in $\mathbb{R}^n$.

Let $\boldsymbol{\zeta} \in \partial f(\mathbf{p}^*) \subset \Gamma f(\mathbf{p}^*)$ and consider $\boldsymbol{\zeta}^{\mathrm{T}}\mathbf{d}$. For any such $\boldsymbol{\zeta}$, there exist a $\mathbf{u} \in \Gamma f(\mathbf{p})$ per (6.2.4) such that

$$\boldsymbol{\zeta}^{\mathrm{T}}\mathbf{d} \leq \mathbf{u}^{\mathrm{T}}\mathbf{d} + \|\mathbf{d}\|^2/2$$

holds. Using (6.2.1),

$$\boldsymbol{\zeta}^{\mathrm{T}}\mathbf{d} \leq -\|\mathbf{d}\|^2 + \|\mathbf{d}\|^2/2, \forall \boldsymbol{\zeta} \in \partial f(\mathbf{p}^*)$$

is obtained. Combining with (6.2.3), the desired result;

$$f(\mathbf{p} + t\mathbf{d}) - f(\mathbf{p}) \leq -t\frac{\|\mathbf{d}\|^2}{2}$$

is obtained for small enough t. Hence $\mathbf{d}$ is a descent direction. □

Remark 6.2.4. The above proof can be carried out using $\tilde{M}(\mathbf{p})$ as defined in Definition 6.1.13 to show that the element of minimum norm of $\tilde{M}(\mathbf{p})$ defines a descent direction.

The following is a technical lemma related to the compactness of the convex hull of a compact set. It will be used repeatedly in the remainder of this chapter.

Lemma 6.2.5 (Convex Hull of a Compact Set). *Let $A \in \mathbb{R}^n$ be a compact set. Then* conv (A) *is a compact set.*

Proof. The result holds trivially if A is empty or a singleton set. The boundedness of conv (A) follows from the definition of the convex hull and the boundedness of A. In the

remainder of the proof, it is shown that conv (A) is a closed set. Assume A is not empty. Let $\bar{\mathbf{x}}$ be a limit point of A. Let $\{\mathbf{x}_k : \mathbf{x}_k \in \text{conv}\,(A)\}_{k=1}^{\infty}$ be a sequence such that $\lim_{k\to\infty} \mathbf{x}_k = \bar{\mathbf{x}}$. Per Carathéodory's Theorem (Theorem 2.4.5), for all k, $\mathbf{x}_k = \sum_{i=1}^{n+1} \alpha_{k,i}\mathbf{y}_{k,i}$, $\sum_{i=1}^{n+1} \alpha_{k,i} = 1$, $\alpha_{k,i} \geq 0, \forall i \in \{1,\ldots,n+1\}$, $\mathbf{y}_{k,i} \in A, \forall i \in \{1,\ldots,n+1\}$ holds. Since $\{\alpha_{k,i}\}_{k=1}^{\infty}$ and $\{\mathbf{y}_{k,i}\}_{k=1}^{\infty}$ are bounded sequences in $\mathbb{R}$ and $\mathbb{R}^n$, there exists an infinite set $\mathcal{J} \subset \{1,\ldots,\infty\}$ such that for all $i \in \{1,\ldots,n+1\}$, $\lim_{j\to\infty} \alpha_{j,i} = \bar{\alpha}_i$, $j \in \mathcal{J}$ and $\lim_{j\to\infty} \mathbf{y}_{j,i} = \bar{\mathbf{y}}_i$, $j \in \mathcal{J}$ holds per the Bolzano-Weierstrass Theorem. Note that $\sum_{i=1}^{n+1} \bar{\alpha}_i = 1$ since $\sum_{i=1}^{n+1} \alpha_{j,i} = 1$ for all $j \in \mathcal{J}$. In addition, $\bar{\mathbf{y}}_i \in A$ for all $i \in \{1,\ldots,n+1\}$ because A is compact. Then for $j \in \mathcal{J}$, $\bar{\mathbf{x}} = \lim_{j\to\infty} \mathbf{x}_j = \sum_{i=1}^{n+1} \bar{\alpha}_i\bar{\mathbf{y}}_i$ holds. This proves that $\bar{\mathbf{x}} \in \text{conv}\,(A)$. Hence conv (A) is compact. □

Theorem 6.2.6 (Necessary Conditions of Optimality for Constrained Optimization Problems using Linear Newton Approximations). *Let Assumptions 6.1.1 and 6.2.1 hold. If* $\mathbf{p}^*$ *is a constrained local solution of program* (6.0.1)*, then there exist numbers* μ_0 *and* μ_i, $i = 1,\ldots,n_c$ *such that*

$$\mathbf{0} \in \mu_0 \Gamma f(\mathbf{p}^*) + \sum_{i=1}^{n_c} \mu_i \Gamma g_i(\mathbf{p}^*), \tag{6.2.5}$$

$$0 \neq |\mu_0| + \sum_{i=1}^{n_c} |\mu_i|, \tag{6.2.6}$$

$$0 = \mu_i g_i(\mathbf{p}^*),\ \forall i \in \{1,\ldots,n_c\}, \tag{6.2.7}$$

$$0 \leq \mu_i,\ \forall i \in \{0,\ldots,n_c\} \tag{6.2.8}$$

hold.

Proof. The result follows from Theorem 6.1.9 and Assumption 6.2.1. □

Definition 6.2.7 (Stationary Point in the Extended Sense). *A point* $\mathbf{p} \in S$ *that satisfies* (6.2.5), (6.2.6), (6.2.7) *and* (6.2.8) *is called a stationary point in the extended sense of problem* (6.0.1).

In the remainder, the following constraint qualification similar to the Cottle constraint qualification (Definition 6.1.10) is assumed to hold.

Definition 6.2.8 (Extended Cottle Constraint Qualification). *Let $\bar{\mathbf{p}} \in S$. Then the extended Cottle constraint qualification holds at $\bar{\mathbf{p}}$ if either $g_i(\bar{\mathbf{p}}) < 0 \; \forall i \in \{1, \ldots, n_c\}$ or $\mathbf{0} \notin \operatorname{conv}\left(\bigcup_{i \in \mathcal{I}(\bar{\mathbf{p}})} \Gamma g_i(\bar{\mathbf{p}})\right)$.*

Remark 6.2.9. Since $\operatorname{conv}\left(\bigcup_{i \in \mathcal{I}(\bar{\mathbf{p}})} \Gamma g_i(\bar{\mathbf{p}})\right)$ contains $\operatorname{conv}\left(\bigcup_{i \in \mathcal{I}(\bar{\mathbf{p}})} \partial g_i(\bar{\mathbf{p}})\right)$, the extended Cottle constraint qualification implies the Cottle constraint qualification. If the constraint functions are convex, the Cottle constraint qualification implies the Slater constraint qualification [54].

In order to relate the improvement function to the necessary conditions of optimality, the linear Newton approximation of the improvement function needs to be derived.

Corollary 6.2.10 (Linear Newton Approximation of the Total Constraint Function). *Let G be as defined in Definition 6.1.12. Then $\Gamma G : \mathcal{P} \rightrightarrows \mathbb{R}^n$ defined by*

$$\Gamma G(\mathbf{p}) = \operatorname{conv}\left(\bigcup_{i \in \mathcal{I}(\mathbf{p})} \Gamma g_i(\mathbf{p})\right)$$

is a linear Newton approximation of G.

Proof. The result follows from the repeated application of (2.8.5) and the properties of linear Newton approximations. □

Corollary 6.2.11 (Linear Newton Approximation of the Improvement Function). *Let H be defined as in Definition 6.1.13. Then $\Gamma H : \mathcal{P} \rightrightarrows \mathbb{R}$ defined by*

$$\Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}}) = \begin{cases} \Gamma f(\bar{\mathbf{p}}) & \text{if } G(\bar{\mathbf{p}}) < 0 \\ \operatorname{conv}\left(\Gamma f(\bar{\mathbf{p}}) \cup \Gamma G(\bar{\mathbf{p}})\right) & \text{if } G(\bar{\mathbf{p}}) = 0 \\ \Gamma G(\bar{\mathbf{p}}) & \text{if } G(\bar{\mathbf{p}}) > 0 \end{cases} \tag{6.2.9}$$

is a linear Newton approximation of H at $\bar{\mathbf{p}}$.

Proof. Using (2.8.5), $\tilde{\Gamma}H : \mathcal{P} \rightrightarrows \mathbb{R}$, a linear Newton approximation at $\bar{\mathbf{p}}$, defined by

$$\tilde{\Gamma}H(\bar{\mathbf{p}};\bar{\mathbf{p}}) = \begin{cases} \Gamma f(\bar{\mathbf{p}}) & \text{if } G(\bar{\mathbf{p}}) < 0 \\ \Gamma f(\bar{\mathbf{p}}) \cup \Gamma G(\bar{\mathbf{p}}) & \text{if } G(\bar{\mathbf{p}}) = 0 \\ \Gamma G(\bar{\mathbf{p}}) & \text{if } G(\bar{\mathbf{p}}) > 0 \end{cases}$$

is obtained. The desired result follows from $\Gamma H(\bar{\mathbf{p}};\bar{\mathbf{p}}) = \text{conv}\left(\tilde{\Gamma}H(\bar{\mathbf{p}};\bar{\mathbf{p}})\right)$ for all $\bar{\mathbf{p}} \in \mathcal{P}$. □

Corollary 6.2.12. *Let $\mathbf{p}^*$ be a constrained local minimum of* (6.0.1). *Then $\mathbf{0} \in \Gamma H(\mathbf{p}^*;\mathbf{p}^*)$ holds. In addition, there exist scalars $\mu_i, i = 0,\ldots,n_c$ such that* (6.2.5), (6.2.6), (6.2.7) *and* (6.2.8) *hold.*

Proof. Since $\mathbf{p}^*$ is a constrained local minimum, $\mathbf{0} \in \partial H(\mathbf{p}^*,\mathbf{p}^*)$ holds per Theorem 6.1.15. Since $\partial H(\mathbf{p}^*,\mathbf{p}^*) \subset \tilde{M}(\mathbf{p}^*) \subset \Gamma H(\mathbf{p}^*;\mathbf{p}^*)$ holds per Definition 6.1.13, Assumption 6.2.1 and Corollary 6.2.11, the rest of the results follow using the set $\{\mu_i\}_{i=0}^{n_c}$ whose existence is stated in Theorem 6.1.15 in the expressions (6.2.5), (6.2.6), (6.2.7) and (6.2.8). □

Corollary 6.2.13. *Let $\bar{\mathbf{p}} \in S$. Assume $\mathbf{0} \in \Gamma H(\bar{\mathbf{p}};\bar{\mathbf{p}})$. Then $\bar{\mathbf{p}}$ is a stationary point in the extended sense of the program* (6.0.1).

Proof. By definition of $\Gamma H(\bar{\mathbf{p}};\bar{\mathbf{p}})$,

$$\mathbf{0} \in \text{conv}\left(\Gamma f(\bar{\mathbf{p}}) \cup \left(\cup_{i\in\mathcal{I}(\bar{\mathbf{p}})} \Gamma g_i(\bar{\mathbf{p}})\right)\right)$$

holds. Hence, there exist μ_i for $i \in \{0\} \cup \mathcal{I}(\bar{\mathbf{p}})$ such that

$$\mathbf{0} \in \mu_0 \Gamma f(\bar{\mathbf{p}}) + \sum_{i\in\mathcal{I}} \mu_i \Gamma g_i(\bar{\mathbf{p}}),$$

$$\mu_i \geq 0, \ \forall i \in \{0\} \cup \mathcal{I}(\bar{\mathbf{p}}),$$

$$\sum_{i \in \{0\} \cup \mathcal{I}(\bar{\mathbf{p}})} \mu_i = 1$$

hold. Let $\mu_i = 0$ for all $i \in \{1, \ldots, n_c\} \backslash \mathcal{I}(\bar{\mathbf{p}})$. Then

$$\mu_i \mathbf{g}_i(\bar{\mathbf{p}}) = 0, \ \forall i \in \{1, \ldots, n_c\}$$

holds because if $i \in \mathcal{I}(\bar{\mathbf{p}})$, then $\mathbf{g}_i(\bar{\mathbf{p}}) = 0$ holds. As a result, there exist scalars μ_i, $i \in \{0, \ldots, n_c\}$ satisfying conditions (6.2.5), (6.2.6), (6.2.7) and (6.2.8). □

The following descent direction for the improvement function is a result of Theorem 6.2.3.

Corollary 6.2.14 (Descent Direction for the Improvement Function). *Let* $\{\mathbf{v}\} = \arg \min\{\|\boldsymbol{\zeta}\| : \boldsymbol{\zeta} \in \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})\}$.[3]. *Assume* $\mathbf{v} \neq \mathbf{0}$. *Let* $\mathbf{d} = -\mathbf{v}$. *Then for small enough positive* t, $H(\bar{\mathbf{p}} + t\mathbf{d}; \bar{\mathbf{p}}) < H(\bar{\mathbf{p}}; \bar{\mathbf{p}})$ *holds.*

In order to compute a descent direction, an approximation of the improvement function will be used in bundle methods. In this thesis, the approximation is a convex function that has a subdifferential at $\mathbf{p}$ equal to $\Gamma H(\mathbf{p}; \mathbf{p})$.

Discussion of the properties of this approximation requires the following theorem (Lemma 2.5 in [54] which is a direct result of Theorems 2.8.2 and 2.8.6 in [25]).

Theorem 6.2.15 (Pointwise Maximum of Functions). *Let* Z *be a compact subset of* $\mathbb{R}^n$. *Let* $\mathbf{z} \in Z$. *Let* $h_{\mathbf{z}} : \mathcal{P} \to \mathbb{R}$ *be a member of a family of functions parameterized by* $\mathbf{z}$. *Let* $h : \mathcal{P} \to \mathbb{R}$ *be defined by*

$$h(\mathbf{p}) = \max\{h_{\mathbf{z}}(\mathbf{p}), \ \mathbf{z} \in Z\}.$$

[3]Note that the solution of this program is the projection of the zero vector on the nonempty compact convex set $\Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})$ Therefore, the solution exists and is unique per Theorem 2.4.6.

Let $\mathcal{M}(\bar{\mathbf{p}}) = \{\mathbf{z} \in Z : h(\bar{\mathbf{p}}) = h_{\mathbf{z}}(\bar{\mathbf{p}})\}$. *Let* O *be a neighborhood of* $\bar{\mathbf{p}}$. *Assume:*

1. *For all* $\mathbf{z} \in \mathcal{M}(\bar{\mathbf{p}})$, $h_{\mathbf{z}}$ *is a Lipschitz continuous function on* O *with the same Lipschitz constant* K.
2. h *is finite at some* $\mathbf{p} \in O$.
3. $h_{\mathbf{z}}$ *is a continuous function from* $Z \times O$ *to* $\mathbb{R}$.
4. $\partial h_{\mathbf{z}}$ *is an upper semicontinuous set-valued mapping from* $Z \times O$ *to* $\mathbb{R}^{n_p}$.

Then:

1. h *is locally Lipschitz continuous at* $\bar{\mathbf{p}}$.
2. $\partial h(\bar{\mathbf{p}}) \subset \text{conv}\left(\cup_{\mathbf{z} \in \mathcal{M}(\bar{\mathbf{p}})} \partial h_{\mathbf{z}}(\bar{\mathbf{p}})\right)$.
3. *If* $h_{\mathbf{z}}$ *for each* $\mathbf{z} \in \mathcal{M}(\bar{\mathbf{p}})$ *is regular at* $\bar{\mathbf{p}}$, *then* h *is regular at* $\bar{\mathbf{p}}$ *and* $\partial h(\bar{\mathbf{p}}) = \text{conv}\left(\cup_{z \in \mathcal{M}(\bar{\mathbf{p}})} \partial h_{\mathbf{z}}(\bar{\mathbf{p}})\right)$. *In addition,* $h'(\bar{\mathbf{p}}; \mathbf{d}) = \max\{\langle \boldsymbol{\zeta}, \mathbf{d} \rangle : \boldsymbol{\zeta} \in \partial h(\bar{\mathbf{p}})\}$ *for all* $\mathbf{d} \in \mathbb{R}^{n_p}$.

Theorem 6.2.16 (Convex Approximation of the Improvement Function). *Let* $\bar{\mathbf{p}} \in \mathcal{P}$ *and* O *be a neighborhood of* $\bar{\mathbf{p}}$. *Define*

1. $\bar{f}_{\boldsymbol{\zeta}} : O \to \mathbb{R} : \mathbf{p} \mapsto f(\bar{\mathbf{p}}) + \langle \boldsymbol{\zeta}, \mathbf{p} - \bar{\mathbf{p}} \rangle$ *for each* $\boldsymbol{\zeta} \in \Gamma f(\bar{\mathbf{p}})$.
2. $\bar{f} : O \to \mathbb{R} : \mathbf{p} \mapsto \max\{\bar{f}_{\boldsymbol{\zeta}}(\mathbf{p}) : \boldsymbol{\zeta} \in \Gamma f(\bar{\mathbf{p}})\}$.
3. $\bar{g}_{i,\boldsymbol{\zeta}} : O \to \mathbb{R} : \mathbf{p} \mapsto g_i(\bar{\mathbf{p}}) + \langle \boldsymbol{\zeta}, \mathbf{p} - \bar{\mathbf{p}} \rangle$ *for each* $\boldsymbol{\zeta} \in \Gamma g_i(\bar{\mathbf{p}})$ *and for all* $i \in \{1, \ldots, n_c\}$.
4. $\bar{g}_i : O \to \mathbb{R} : \mathbf{p} \mapsto \max\{\bar{g}_{i,\boldsymbol{\zeta}}(\mathbf{p}) : \boldsymbol{\zeta} \in \Gamma g_i(\bar{\mathbf{p}})\}$, *for all* $i \in \{1, \ldots, n_c\}$.
5. $\bar{G} : O \to \mathbb{R} : \mathbf{p} \mapsto \max\{\bar{g}_i(\mathbf{p}) : i \in \{1, \ldots, n_c\}\}$.
6. $\bar{H} : O \to \mathbb{R} : \mathbf{p} \mapsto \max\{\bar{f}(\mathbf{p}) - \bar{f}(\bar{\mathbf{p}}), \bar{G}(\mathbf{p})\}$. $\bar{H}$ *is called the convex approximation of the improvement function.*

Then:

1. $\bar{f}$ *is a convex function on* O *and* $\partial \bar{f}(\bar{\mathbf{p}}) = \Gamma f(\bar{\mathbf{p}})$.
2. $\bar{g}_i$ *is a convex function on* O *and* $\partial \bar{g}_i(\bar{\mathbf{p}}) = \Gamma g_i(\bar{\mathbf{p}})$ *for all* $i \in \{1, \ldots, n_c\}$.
3. $\bar{G}$ *is a convex function on* O *and* $\partial \bar{G}(\bar{\mathbf{p}}) = \text{conv}\left(\cup_{i \in \mathcal{Y}(\bar{\mathbf{p}})} \Gamma \bar{g}_i(\bar{\mathbf{p}})\right)$ *where* $\mathcal{Y}(\bar{\mathbf{p}}) = \{i : \bar{g}_i(\bar{\mathbf{p}}) = \bar{G}(\bar{\mathbf{p}}),\ i \in \{1, \ldots, n_c\}\}$.
4. $\bar{H}$ *is a convex function on* O *and* $\partial \bar{H}(\bar{\mathbf{p}}) = \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})$.

5. *Let* $\mathbf{d} = -\mathbf{v}$ *where* $\{\mathbf{v}\} = \arg\min\{\|\boldsymbol{\zeta}\| : \boldsymbol{\zeta} \in \partial \bar{H}(\bar{\mathbf{p}})\}$. *If* $\mathbf{v} \neq \mathbf{0}$, *then* $\mathbf{d}$ *is a descent direction of* $\bar{H}$ *at* $\bar{\mathbf{p}}$.

Proof. Let $K = \max\left\{\|\boldsymbol{\zeta}\| : \boldsymbol{\zeta} \in \Gamma f(\bar{\mathbf{p}}) \cup \left(\cup_{i\in\{1,\ldots,n_c\}} \Gamma g_i(\bar{\mathbf{p}})\right)\right\}$. K is finite and well-defined since the linear Newton approximations in the definition are all compact subsets of $\mathbb{R}^{n_p}$ and n_c is finite. Note that K is a Lipschitz constant for all $\bar{f}_{\boldsymbol{\zeta}}$ and for all $\bar{g}_{i,\boldsymbol{\zeta}}$.

Observe that $\partial_{\boldsymbol{\zeta}} \bar{f}(\mathbf{p}) = \{\boldsymbol{\zeta}\}, \boldsymbol{\zeta} \in \Gamma f(\bar{\mathbf{p}})$ and $\partial \bar{g}_{i,\boldsymbol{\zeta}}(\mathbf{p}) = \{\boldsymbol{\zeta}\}, \boldsymbol{\zeta} \in \Gamma g_i(\bar{\mathbf{p}})$ for all $i \in \{i, \ldots, n_c\}$. $(\boldsymbol{\zeta}, \mathbf{p}) \mapsto f_{\boldsymbol{\zeta}}(\mathbf{p})$ is a continuous function and $(\boldsymbol{\zeta}, \mathbf{p}) \Rightarrow \partial f_{\boldsymbol{\zeta}}(\mathbf{p})$ is an upper semi-continuous set-valued map because $\bar{f}_{\boldsymbol{\zeta}}$ and $\boldsymbol{\nabla} \bar{f}_{\boldsymbol{\zeta}}$ are continuous functions of $(\boldsymbol{\zeta}, \mathbf{p})$. Similarly, $(\boldsymbol{\zeta}, \mathbf{p}) \mapsto g_{i,\boldsymbol{\zeta}}(\mathbf{p})$ is a continuous and $(\boldsymbol{\zeta}, \mathbf{p}) \Rightarrow \partial g_{i,\boldsymbol{\zeta}}(\mathbf{p})$ is an upper semicontinuous map for all $i \in \{1, \ldots, n_c\}$. Regularity of $\bar{f}_{\boldsymbol{\zeta}}$ and $\bar{g}_{i,\boldsymbol{\zeta}}$ follows from their continuous differentiability. Convexity of $\bar{f}_{\boldsymbol{\zeta}}$ and $\bar{g}_{i,\boldsymbol{\zeta}}$ is a result of $\bar{f}_{\boldsymbol{\zeta}}$ and $\bar{g}_{i,\boldsymbol{\zeta}}$ being affine functions.

Hence $\bar{f}$ and $\bar{g}_i$ are Lipschitz continuous functions on O, regular at $\bar{\mathbf{p}}$, $\partial \bar{f}(\bar{\mathbf{p}}) = \text{conv}\,(\Gamma f(\bar{\mathbf{p}}))$ and $\partial \bar{g}_i(\bar{\mathbf{p}}) = \text{conv}\,(\Gamma g_i(\bar{\mathbf{p}}))$ per Theorem 6.2.15. Since $\Gamma f(\bar{\mathbf{p}})$ and $\Gamma g_i(\bar{\mathbf{p}})$ are convex sets per Assumption 6.2.1, $\partial \bar{f}(\bar{\mathbf{p}}) = \Gamma f(\bar{\mathbf{p}})$ and $\partial \bar{g}_i(\bar{\mathbf{p}}) = \Gamma g_i(\bar{\mathbf{p}})$ follows.

Let $\mathbf{p}_1 \in O$ and $\mathbf{p}_2 \in O$. Let $\mathbf{p}_3 \in \{\mathbf{p} : \mathbf{p} = \alpha \mathbf{p}_1 + (1-\alpha)\mathbf{p}_2, \alpha \in (0,1)\}$. Let $\boldsymbol{\zeta}_3 \in \partial \bar{f}(\bar{\mathbf{p}})$ be such that $f(\bar{\mathbf{p}}) + \langle \boldsymbol{\zeta}_3, \mathbf{p}_3 - \bar{\mathbf{p}} \rangle = \bar{f}(\mathbf{p}_3)$. Then $f(\bar{\mathbf{p}}) + \langle \boldsymbol{\zeta}_3, \mathbf{p}_3 - \bar{\mathbf{p}} \rangle = \alpha f(\bar{\mathbf{p}}) + \alpha \langle \boldsymbol{\zeta}_3, \mathbf{p}_1 - \bar{\mathbf{p}} \rangle + (1-\alpha) f(\bar{\mathbf{p}}) + (1-\alpha)\langle \boldsymbol{\zeta}_3, \mathbf{p}_2 - \bar{\mathbf{p}} \rangle$ holds. $\bar{f}(\mathbf{p}_3) \leq \alpha \bar{f}(\mathbf{p}_1) + (1-\alpha)\bar{f}(\mathbf{p}_2)$ follows from $f(\bar{\mathbf{p}}) + \langle \boldsymbol{\zeta}_3, \mathbf{p}_1 - \bar{\mathbf{p}} \rangle \leq \bar{f}(\mathbf{p}_1)$ and $f(\bar{\mathbf{p}}) + \langle \boldsymbol{\zeta}_3, \mathbf{p}_2 - \bar{\mathbf{p}} \rangle \leq \bar{f}(\mathbf{p}_2)$. Hence $\bar{f}$ is a convex function. The convexity of $\bar{g}_i$ follows from the same reasoning.

Per (2.6.4), $\bar{G}$ is regular at $\bar{\mathbf{p}}$, Lipschitz continuous on O and $\partial \bar{G}(\bar{\mathbf{p}}) = \text{conv}\left(\cup_{i\in\mathcal{Y}(\bar{\mathbf{p}})} \Gamma \bar{g}_i(\bar{\mathbf{p}})\right)$. Convexity follows from the fact that the maximum of a finite number of convex functions is convex.

Similar to the case $\bar{G}$, $\bar{H}$ is regular at $\bar{\mathbf{p}}$, Lipschitz continuous and convex on O. Note

that $\bar{H}(\bar{\mathbf{p}}) = \max\{0, \bar{G}(\bar{\mathbf{p}})\}$. Per (2.6.4),

$$\partial \bar{H}(\bar{\mathbf{p}}) = \begin{cases} \partial \bar{f}(\bar{\mathbf{p}}) & \text{if } \bar{G}(\bar{\mathbf{p}}) < 0 \\ \text{conv}\left(\partial \bar{f}(\bar{\mathbf{p}}) \cup \left(\cup_{i \in \mathcal{Y}(\bar{\mathbf{p}})} \partial \bar{g}_i(\bar{\mathbf{p}})\right)\right) & \text{if } \bar{G}(\bar{\mathbf{p}}) = 0 \\ \text{conv}\left(\cup_{i \in \mathcal{Y}(\bar{\mathbf{p}})} \partial \bar{g}_i(\bar{\mathbf{p}})\right) & \text{if } \bar{G}(\bar{\mathbf{p}}) > 0. \end{cases} \tag{6.2.10}$$

which is equal to (6.2.9).

$\mathbf{d}$ is a descent direction per Theorem 6.1.4. □

Lemma 6.2.17. *Let $\bar{H}$ be as defined in Theorem 6.2.16. Then $\bar{H}(\bar{\mathbf{p}}+\mathbf{d}) = \max(\langle \boldsymbol{\zeta}, \mathbf{d}\rangle,\ \boldsymbol{\zeta} \in \Gamma \bar{H}(\bar{\mathbf{p}}; \bar{\mathbf{p}}))$ holds for $\bar{\mathbf{p}} \in S$ and for all $\mathbf{d} \in \mathbb{R}^{n_p}$.*

Proof. Let $\Delta = \bar{H}(\bar{\mathbf{p}} + \mathbf{d})$. First, assume $\bar{f}(\bar{\mathbf{p}} + \mathbf{d}) - \bar{f}(\bar{\mathbf{p}}) > \bar{G}(\bar{\mathbf{p}} + \mathbf{d})$. Then, from the definition of $\bar{f}$, it can deduced that $\Delta = \max(\langle \boldsymbol{\zeta}, \mathbf{d}\rangle,\ \boldsymbol{\zeta} \in \partial \bar{f}(\bar{\mathbf{p}}))$. Now assume that for some $i \in \mathcal{Y}(\bar{\mathbf{p}})$, $\bar{g}_i(\mathbf{p}) > \bar{g}_j(\mathbf{p})$ holds for all $j \in \mathcal{Y}(\bar{\mathbf{p}})\backslash\{i\}$ and $\bar{g}_i(\mathbf{p}) > \bar{f}(\bar{\mathbf{p}} + \mathbf{d}) - \bar{f}(\bar{\mathbf{p}})$. Then, using the definition of $\bar{g}_i$ and the fact that $\bar{g}_i(\bar{\mathbf{p}}) = 0$, it can be shown that $\Delta = \max(\langle \boldsymbol{\zeta}, \mathbf{d}\rangle,\ \boldsymbol{\zeta} \in \Gamma g_i(\bar{\mathbf{p}}))$.

Let $\mathcal{J} \subset \mathcal{Y}(\bar{\mathbf{p}})$ be such that $\Delta = \bar{g}_j(\mathbf{p}),\ \forall j \in \mathcal{J}$ and $\Delta > \bar{g}_i(\mathbf{p}),\ \forall i \in \mathcal{Y}(\bar{\mathbf{p}})\backslash\mathcal{J}$. Then $\Delta = \sum_{i \in \mathcal{Y}(\bar{\mathbf{p}})} \alpha_i \max(\langle \boldsymbol{\zeta}, \mathbf{d}\rangle,\ \boldsymbol{\zeta} \in \Gamma g_i(\bar{\mathbf{p}}))$ holds where $\alpha_i \geq 0,\ \sum_{i \in \mathcal{Y}(\bar{\mathbf{p}})} \alpha_i = 1$. Note that $\alpha_i = 0,\ \forall i \in \mathcal{Y}(\bar{\mathbf{p}})\backslash\mathcal{J}$. This can be written as

$$\Delta = \max\left(\left\langle\left(\sum_{i \in \mathcal{Y}(\bar{\mathbf{p}})} \alpha_i \boldsymbol{\zeta}_i\right), \mathbf{d}\right\rangle,\ \boldsymbol{\zeta}_i \in \Gamma g_i(\bar{\mathbf{p}}),\ 0 \leq \alpha_i \leq 1,\ \sum_{i \in \mathcal{Y}(\bar{\mathbf{p}})} \alpha_i = 1,\ i \in \mathcal{Y}(\mathbf{p})\right)$$

or equivalently $\Delta = \max\left(\langle \boldsymbol{\zeta}, \mathbf{d}\rangle,\ \boldsymbol{\zeta} \in \partial \bar{G}(\bar{\mathbf{p}})\right)$. Note that if $\boldsymbol{\zeta} \in \partial \bar{G}(\bar{\mathbf{p}})$ such that $\Delta = \langle \boldsymbol{\zeta}, \mathbf{d}\rangle$, then $\boldsymbol{\zeta} \in \text{conv}\left(\cup_{i \in \mathcal{J}} \Gamma g_i(\bar{\mathbf{p}})\right)$ otherwise $\Delta > \langle \boldsymbol{\zeta}, \mathbf{d}\rangle$ has to hold.

If $\Delta = \bar{f}(\bar{\mathbf{p}}+\mathbf{d}) - \bar{f}(\bar{\mathbf{p}})$ as well, then $\Delta = \lambda_1 \max(\langle \boldsymbol{\zeta}, \mathbf{d}\rangle,\ \boldsymbol{\zeta} \in \partial \bar{f}(\bar{\mathbf{p}})) + \lambda_2 \max\left(\langle \boldsymbol{\zeta}, \mathbf{d}\rangle,\ \boldsymbol{\zeta}_i \in \partial \bar{G}(\bar{\mathbf{p}})\right)$ holds where $\lambda_1 \geq 0,\ \lambda_2 \geq 0$ and $\lambda_1 + \lambda_2 = 1$. This can be shown to be equivalent

to $\max(\langle \boldsymbol{\zeta}, \mathbf{d} \rangle,\ \boldsymbol{\zeta} \in \Gamma \bar{H}(\bar{\mathbf{p}}; \bar{\mathbf{p}}))$ using the definition of $\partial \bar{H}(\bar{\mathbf{p}})$ and the results in Theorem 6.2.16. □

The following theorem elucidates the relationship between $\bar{H}'(\bar{\mathbf{p}}; \mathbf{d})$ and $H'(\bar{\mathbf{p}}; \bar{\mathbf{p}}; \mathbf{d})$, the directional derivative of the improvement function. Note that $H'(\bar{\mathbf{p}}; \bar{\mathbf{p}}; \mathbf{d})$ exists for all $\mathbf{d} \in \mathbb{R}^{n_p}$ because H is a composition of the max function, f and g_i, $i \in \{1, \ldots, n_c\}$ which are all Bouligand differentiable at all $\mathbf{p} \in \mathcal{P}$ as a result of Assumption 6.2.1. $\bar{H}'(\bar{\mathbf{p}}; \mathbf{d})$ exists for all $\mathbf{d} \in \mathbb{R}^{n_p}$ because $\bar{H}$ is a finite convex function at $\mathbf{p}$.

Corollary 6.2.18. *Let Assumptions 6.1.1 and 6.2.1 hold. Let H be as defined in Definition 6.1.13 and $\bar{H}$ as defined in Theorem 6.2.16. Let $\bar{\mathbf{p}} \in S$. Then*

$$H'(\bar{\mathbf{p}}; \bar{\mathbf{p}}; \mathbf{d}) \leq \bar{H}'(\bar{\mathbf{p}}; \mathbf{d}),\ \forall \mathbf{d} \in \mathbb{R}^{n_p},$$
$$H(\bar{\mathbf{p}} + t\mathbf{d}; \bar{\mathbf{p}}) \leq H(\bar{\mathbf{p}}; \bar{\mathbf{p}}) + t\bar{H}'(\bar{\mathbf{p}}; \mathbf{d}) + o(t),$$

where t is a positive scalar and $o(t)/t \to 0$ as $t \downarrow 0$.

Proof. $H'(\bar{\mathbf{p}}; \bar{\mathbf{p}}; \mathbf{d}) \leq H^o(\bar{\mathbf{p}}; \bar{\mathbf{p}}; \mathbf{d})$ per the definition of the generalized directional derivative (Definition 2.6.1). Note that $H^o(\bar{\mathbf{p}}; \bar{\mathbf{p}}; \mathbf{d}) = \max\{\langle \boldsymbol{\zeta}, \mathbf{d} \rangle, \boldsymbol{\zeta} \in \partial H(\bar{\mathbf{p}}; \bar{\mathbf{p}})\}$ and $\bar{H}'(\bar{\mathbf{p}}; \mathbf{d}) = \bar{H}^o(\bar{\mathbf{p}}; \mathbf{d}) = \max\{\langle \boldsymbol{\zeta}, \mathbf{d} \rangle, \boldsymbol{\zeta} \in \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})\}$ since $\bar{H}$ is regular at $\bar{\mathbf{p}}$. $H'(\bar{\mathbf{p}}; \bar{\mathbf{p}}; \mathbf{d}) \leq \bar{H}'(\bar{\mathbf{p}}; \mathbf{d})$, $\forall \mathbf{d} \in \mathbb{R}^{n_p}$ follows because $\partial H(\bar{\mathbf{p}}; \bar{\mathbf{p}}) \subset \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})$.

$H(\bar{\mathbf{p}} + t\mathbf{d}; \bar{\mathbf{p}}) = H(\bar{\mathbf{p}}; \bar{\mathbf{p}}) + tH'(\bar{\mathbf{p}}; \bar{\mathbf{p}}; \mathbf{d}) + o(t)$ holds per the definition of the directional derivative. The result follows after substituting $\bar{H}'(\bar{\mathbf{p}}; \mathbf{d})$ for $H'(\bar{\mathbf{p}}; \bar{\mathbf{p}}; \mathbf{d})$. □

Definition 6.2.19 (Feasible Descent Direction). *$\mathbf{d} \in \mathbb{R}^{n_p}$ is a feasible descent direction for f at $\mathbf{p}$ with respect to S if $(\mathbf{p} + t\mathbf{d}) \in S$ for small enough positive t and $\mathbf{d}$ is a descent direction with respect to f.*

The next corollary motivates searching for a descent direction for H using $\bar{H}$.

Corollary 6.2.20 (Feasible Descent Direction of H obtained from $\bar{H}$). *Let Assumptions 6.1.1 and 6.2.1 hold. Let H be as defined in Definition 6.1.13 and $\bar{H}$ as defined in Theorem 6.2.16. Let $\mathbf{d}$ be a descent direction for $\bar{H}$ at $\bar{\mathbf{p}} \in S$, i.e $\bar{H}(\bar{\mathbf{p}} + t\mathbf{d}) < \bar{H}(\bar{\mathbf{p}}) = 0$ for small enough t. Then $\mathbf{d}$ is a feasible descent direction for f at $\bar{\mathbf{p}}$ relative to S.*

Proof. Note that $\bar{\mathbf{p}} \in S$, $H(\bar{\mathbf{p}}; \bar{\mathbf{p}}) = 0$. Since $\mathbf{d}$ is a descent direction $\bar{H}'(\bar{\mathbf{p}}; \mathbf{d}) < 0$. Per Corollary 6.2.18,

$$H(\bar{\mathbf{p}} + t\mathbf{d}; \bar{\mathbf{p}}) \leq H(\bar{\mathbf{p}}; \bar{\mathbf{p}}) + t(\bar{H}'(\bar{\mathbf{p}}; \mathbf{d}) + o(t)/t) \tag{6.2.11}$$

holds. Since $o(t)/t \to 0$ as $t \downarrow 0$, for small enough t, $H(\bar{\mathbf{p}} + t\mathbf{d}; \bar{\mathbf{p}}) < H(\bar{\mathbf{p}}; \bar{\mathbf{p}}) = 0$.

Since $\bar{\mathbf{p}} \in S$, $G(\bar{\mathbf{p}}) \leq 0$. Note that $H(\bar{\mathbf{p}}; \bar{\mathbf{p}}) = \max(f(\bar{\mathbf{p}}) - f(\bar{\mathbf{p}}), G(\bar{\mathbf{p}})) = \max(0, G(\bar{\mathbf{p}})) = 0$. In order for $H(\bar{\mathbf{p}} + t\mathbf{d}; \bar{\mathbf{p}}) < 0$ to hold for sufficiently small t, $f(\bar{\mathbf{p}} + t\mathbf{d}) < f(\bar{\mathbf{p}})$ and $G(\bar{\mathbf{p}} + t\mathbf{d}) \leq 0$ have to hold simultaneously, proving the claim. □

If all the elements of $\partial\bar{H}(\mathbf{p})$ were available, the following quadratic problem would furnish a direction of descent.

Theorem 6.2.21. *Let Assumptions 6.1.1 and 6.2.1 hold. Let H be as defined in Definition 6.1.13 and $\bar{H}$ as defined in Theorem 6.2.16.*

Let $\bar{\mathbf{p}} \in S$ and let $\{\mathbf{v}\} = \arg\min\{\|\boldsymbol{\zeta}\| : \boldsymbol{\zeta} \in \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})\}$.[4] Let $\mathbf{d}^*$ be a solution of

$$\min_{\mathbf{d} \in \mathbb{R}^{n_p}} \bar{H}(\bar{\mathbf{p}} + \mathbf{d}) + \frac{1}{2}\|\mathbf{d}\|^2. \tag{6.2.12}$$

Then:

1. *$\mathbf{d}^*$ exists and is unique.*
2. *$\mathbf{d}^* = -\mathbf{v}$.*

[4] $\mathbf{v}$ is well-defined per Theorem 2.4.6 because $\Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})$ is a nonempty compact and convex set and $\mathbf{v}$ is the unique projection of the zero vector onto $\Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})$.

3. $\bar{H}(\bar{\mathbf{p}} + \mathbf{d}^*) = \bar{H}(\bar{\mathbf{p}}) - \|\mathbf{d}^*\|^2$.

4. $\bar{H}(\bar{\mathbf{p}} + t\mathbf{d}^*) \leq \bar{H}(\bar{\mathbf{p}}) - t\|\mathbf{d}^*\|^2$ *for all* $t \in [0, 1]$.

5. $\mathbf{d}^* \neq \mathbf{0}$ *if and only if* $\mathbf{0} \notin \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})$.

6. $\bar{\mathbf{p}}$ *is an unconstrained global minimum of the function* $\bar{H}$ *if and only if* $\mathbf{d}^* = \mathbf{0}$.

7. *If the extended Cottle constraint is satisfied at* $\bar{\mathbf{p}}$ *for problem* (6.0.1) *then* $\bar{\mathbf{p}}$ *is stationary in the extended sense for* (6.0.1) *if and only if* $\bar{\mathbf{p}}$ *is a constrained global minimum of*

$$\min_{\mathbf{p} \in \mathcal{P}} \bar{f}(\mathbf{p}), \text{ s.t. } \bar{G}(\mathbf{p}) \leq 0. \tag{6.2.13}$$

8. *Problem* (6.2.12) *is equivalent to the problem*

$$\begin{aligned} &\min_{\Delta, \mathbf{d}} \Delta + \frac{1}{2}\|\mathbf{d}\|^2 \\ &\quad \text{s.t. } H(\bar{\mathbf{p}}; \bar{\mathbf{p}}) + \langle \boldsymbol{\zeta}, \mathbf{d} \rangle \leq \Delta, \ \forall \boldsymbol{\zeta} \in \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}}), \\ &\quad \Delta \in \mathbb{R}, \ \mathbf{d} \in \mathbb{R}^{n_p}. \end{aligned} \tag{6.2.14}$$

Proof. Let $V : \mathbb{R}^{n_p} \to \mathbb{R} : \mathbf{d} \mapsto \max\{\langle \boldsymbol{\zeta}, \mathbf{d} \rangle : \ \boldsymbol{\zeta} \in \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})\}$ and $J : \mathbb{R}^{n_p} \to \mathbb{R} : \mathbf{d} \mapsto \bar{H}(\bar{\mathbf{p}} + \mathbf{d}) + \frac{1}{2}\|\mathbf{d}\|^2$. Note that

$$\begin{aligned} J(\mathbf{d}) &= \bar{H}(\bar{\mathbf{p}}) + V(\mathbf{d}) + \frac{1}{2}\|\mathbf{d}\|^2 = V(\mathbf{d}) + \frac{1}{2}\|\mathbf{d}\|^2, \\ V(\mathbf{d}) + \frac{1}{2}\|\mathbf{d}\|^2 &\geq -M\|\mathbf{d}\| + \frac{1}{2}\|\mathbf{d}\|^2 \end{aligned}$$

where $M = \max\{\|\boldsymbol{\zeta}\| : \ \boldsymbol{\zeta} \in \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})\}$ per Lemma 6.2.17. Since $-M\|\mathbf{d}\| + \frac{1}{2}\|\mathbf{d}\|^2 \to +\infty$ as $\|\mathbf{d}\| \to +\infty$, $J(\mathbf{d}) \to +\infty$ as $\|\mathbf{d}\| \to +\infty$. Therefore the minimum of (6.2.12) and a $\mathbf{d}^*$ exists.

$\bar{H}(\bar{\mathbf{p}}) + V(\mathbf{d}) + \frac{1}{2}\|\mathbf{d}\|^2$ is a strictly convex function because $\|\mathbf{d}\|^2 = \mathbf{d}^{\mathrm{T}}\mathbf{d}$ is strictly convex and $\max\{\langle \boldsymbol{\zeta}, \mathbf{d} \rangle, \ \boldsymbol{\zeta} \in \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})\}$ is convex. Hence $\mathbf{d}^*$ is unique.

Note that V and $\|\cdot\|^2$ are convex hence regular functions. As a result, the sum rules in §2.6.2 hold with equality. Hence $\partial J(\mathbf{d}^*) = \partial V(\mathbf{d}^*) + \mathbf{d}^*$ holds. In addition, at $\mathbf{d}^*$, $\mathbf{0} \in \partial J(\mathbf{d}^*)$ has to hold per Theorem 6.1.3. Therefore $-\mathbf{d}^* \in \partial V(\mathbf{d}^*)$. Since $\partial V(\mathbf{d}^*) = \{\boldsymbol{\zeta} \in \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}}) : V(\mathbf{d}^*) = \langle \boldsymbol{\zeta}, \mathbf{d}^* \rangle\}$ per Theorem 6.2.15, $-\mathbf{d}^* \in \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})$. Note that $\{\boldsymbol{\zeta} \in \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}}) : V(\mathbf{d}^*) = \langle \boldsymbol{\zeta}, \mathbf{d}^* \rangle\}$ is a convex set. Let $\boldsymbol{\zeta}_1 \in \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})$ and $\boldsymbol{\zeta}_2 \in \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})$ such that $V(\mathbf{d}^*) = \langle \boldsymbol{\zeta}_1, \mathbf{d}^* \rangle = \langle \boldsymbol{\zeta}_2, \mathbf{d}^* \rangle$ holds. Then $V(\mathbf{d}^*) = \langle \alpha_1 \boldsymbol{\zeta}_1 + \alpha_2 \boldsymbol{\zeta}_2, \mathbf{d}^* \rangle$ holds for all $\alpha_1 \geq 0$, $\alpha_2 \geq 0$ satisfying $\alpha_1 + \alpha_2 = 1$. Note that $\alpha_1 \boldsymbol{\zeta}_1 + \alpha_2 \boldsymbol{\zeta}_2 \in \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})$ because $\Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})$ is a convex set.

Let $\bar{\mathbf{v}} = -\mathbf{d}^*$. Note that $V(\mathbf{d}^*) = \langle \bar{\mathbf{v}}, \mathbf{d}^* \rangle = \max\{\langle \boldsymbol{\zeta}, \mathbf{d}^* \rangle : \boldsymbol{\zeta} \in \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})\} = -\|\bar{\mathbf{v}}\|^2$. Hence $\langle \boldsymbol{\zeta}, \bar{\mathbf{v}} \rangle \geq \|\bar{\mathbf{v}}\|^2$ for all $\boldsymbol{\zeta} \in \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})$ and $\bar{\mathbf{v}} = \mathbf{v}$ per Theorem 2.4.6.

Since $V(\mathbf{d}^*) = -\|\mathbf{d}^*\|^2$, $\bar{H}(\bar{\mathbf{p}} + \mathbf{d}^*) = \bar{H}(\bar{\mathbf{p}}) - \|\mathbf{d}^*\|^2$. Note that $V(t\mathbf{d}^*) = tV(\mathbf{d}^*)$. Therefore $\bar{H}(\bar{\mathbf{p}} + t\mathbf{d}^*) = \bar{H}(\bar{\mathbf{p}}) - t\|\mathbf{d}^*\|^2$.

Item 5 follows from the fact that $\mathbf{d}^* = -\mathbf{v}$. Item 6 follows from the fact if $\mathbf{d}^* = 0$, then $\mathbf{0} \in \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})$ and this implies that $\mathbf{0} \in \partial \bar{H}(\bar{\mathbf{p}})$. Since $\bar{H}$ is a convex function, this condition is necessary and sufficient for $\bar{\mathbf{p}}$ to be an unconstrained global minimum per Theorem 6.1.17.

Program (6.2.13) is a convex optimization problem. The necessary conditions of optimality at $\bar{\mathbf{p}}$ can be shown to be $\mathbf{0} \in \partial \bar{H}(\bar{\mathbf{p}})$ or equivalently $\mathbf{0} \in \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})$. If the extended Cottle constraint is satisfied at $\bar{\mathbf{p}}$ for problem (6.0.1), then the Cottle constraint qualification (also the Slater constraint qualification) is satisfied for (6.2.13). Then $\bar{\mathbf{p}} \in S$ is a constrained global minimum of (6.2.13) if and only if $\mathbf{0} \in \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})$ per Theorem 6.1.17.

Note that $\bar{H}(\bar{\mathbf{p}} + \mathbf{d}) = V(\mathbf{d})$. Let $\mathbf{d}^*$ be the solution of (6.2.12). Let $\Delta^* = V(\mathbf{d}^*)$. Note that $(\Delta^*, \mathbf{d}^*)$ is a feasible point for (6.2.14) and that (6.2.14) can be written as

$$\min_{\Delta, \mathbf{d}} \Delta + \frac{1}{2}\|\mathbf{d}\|^2 \text{ s.t. } \max\{\langle \boldsymbol{\zeta}, \mathbf{d} \rangle,\ \boldsymbol{\zeta} \in \Gamma H(\bar{\mathbf{p}}, \bar{\mathbf{p}})\} - \Delta \leq 0 \tag{6.2.15}$$

which is a convex program. The generalized gradient of the objective $(\Delta, \mathbf{d}) \mapsto \Delta + \frac{1}{2}\|\mathbf{d}\|^2$

is $\{(1, \mathbf{d})\}$ because the objective is a continuously differentiable function. The generalized gradient of the constraint $(\Delta, \mathbf{d}) \mapsto \max\{\langle \boldsymbol{\zeta}, \mathbf{d}\rangle,\ \boldsymbol{\zeta} \in \Gamma H(\bar{\mathbf{p}}, \bar{\mathbf{p}})\} - \Delta$ is $\{(-1, \boldsymbol{\zeta}) : \langle \boldsymbol{\zeta}, \mathbf{d}\rangle = V(\mathbf{d}), \boldsymbol{\zeta} \in \Gamma H(\bar{\mathbf{p}}, \bar{\mathbf{p}})\}$. Then the necessary condition of optimality at $\bar{\mathbf{p}}$ is

$$\mathbf{0} \in \operatorname{conv}\left(\{(1, \mathbf{d})\}, \{(-1, \boldsymbol{\zeta}) : \boldsymbol{\zeta} \in \Gamma H(\bar{\mathbf{p}}, \bar{\mathbf{p}}),\ \langle \boldsymbol{\zeta}, \mathbf{d}\rangle = V(\mathbf{d})\}\right). \tag{6.2.16}$$

Note that $(2M\|\mathbf{d}\|, \mathbf{d}),\ \mathbf{d} \in \mathbb{R}^{n_p}$ where $M = \max\{\|\boldsymbol{\zeta}\| : \boldsymbol{\zeta} \in \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})\}$ satisfies the constraint in (6.2.15) with strict inequality. Hence, the Slater constraint qualification holds for (6.2.15). Then, the necessary conditions are also sufficient for optimality. Since $-\mathbf{d}^* \in \{\boldsymbol{\zeta} \in \Gamma H(\bar{\mathbf{p}}, \bar{\mathbf{p}}) :\ \langle \boldsymbol{\zeta}, \mathbf{d}^*\rangle = \Delta^*\}$, $(\Delta^*, \mathbf{d}^*)$ is a constrained global minimum of (6.2.15) and (6.2.14).

Assume $(\bar{\Delta}, \bar{\mathbf{d}})$ is another constrained global minimum of (6.2.15) where $\bar{\mathbf{d}} \neq \mathbf{d}^*$. Then from the necessary condition of optimality (6.2.16), it can be deduced that $-\bar{\mathbf{d}} \in \Gamma H(\bar{\mathbf{p}}, \bar{\mathbf{p}})$ and $V(\bar{\mathbf{d}}) = -\|\bar{\mathbf{d}}\|^2$. Then the optimal solution value is $V(\bar{\mathbf{d}}) + \frac{1}{2}\|\bar{\mathbf{d}}\|^2 = -\frac{1}{2}\|\bar{\mathbf{d}}\|^2$. Note that, the optimal solution value is also $V(\mathbf{d}^*) + \frac{1}{2}\|\mathbf{d}^*\|^2 = -\frac{1}{2}\|\mathbf{d}^*\|^2$. $\|\mathbf{d}^*\| < \|\bar{\mathbf{d}}\|$ because $-\mathbf{d}^*$ is the element of minimum norm in $\Gamma H(\bar{\mathbf{p}}, \bar{\mathbf{p}})$ per item 2 and it is unique per Theorem 2.4.6. Hence, (6.2.15) and (6.2.14) have a unique constrained global minimum which is $(\Delta^*, \mathbf{d}^*)$. As a result, programs (6.2.14) and (6.2.12) are equivalent. □

6.3 Bundle Method using Linear Newton Approximations

This section discusses a bundle method algorithm using linear Newton approximations in detail. The algorithm is an extension of Algorithm 3.1 in Chapter 6 in [54] that uses linear Newton approximations which satisfy Assumption 6.2.1. The proof of convergence follows closely the proof of convergence of Algorithm 3.1 in Chapter 6 in [54] making modifications that take into account the use of linear Newton approximations.

The bundle method is an iterative algorithm. It requires an initial point $\mathbf{p}_1 \in S$ and produces an infinite sequence, $\{\mathbf{p}_k\}$, such that $\{\mathbf{p}_k\} \in S$ and if $\mathbf{p}_k \neq \mathbf{p}_{k+1}$, then $f(\mathbf{p}_{k+1}) < f(\mathbf{p}_k)$. The algorithm produces a sequence $\{\mathbf{p}_k\}$ such that if $\bar{\mathbf{p}}$ is an accumulation point of $\{\mathbf{p}_k\}$, then $\bar{\mathbf{p}} \in S$ and $\mathbf{0} \in \Gamma H(\bar{\mathbf{p}}; \bar{\mathbf{p}})$ holds, i.e $\bar{\mathbf{p}}$ is stationary in the extended sense per Corollary 6.2.13. The algorithm requires that an element of Γf and ΓG be computable for all $\mathbf{p} \in \mathcal{P}$.

Ideally, in order to generate $\mathbf{p}_{k+1}$ at the kth iteration, (6.2.12) should be solved to obtain a descent direction and then a line search should be executed along this direction of descent. In practice, given $\mathbf{p} \in \mathcal{P}$, every element of $\Gamma H(\mathbf{p}; \mathbf{p})$ is not known. Usually only a single element can be computed. Therefore problem (6.2.12) cannot be solved to obtain a descent direction in most applications. Instead an approximation of (6.2.12) is formulated using elements from linear Newton approximations of neighboring points that are close enough. These neighboring points comprise the second sequence of points $\{\mathbf{y}_k\} \in \mathcal{P}$ with $\mathbf{y}_1 = \mathbf{p}_1$.

The candidate direction obtained from the approximation of (6.2.12) is tested in a special line search algorithm. The line search algorithm returns $\mathbf{p}_{k+1}$ and another point $\mathbf{y}_{k+1}$ which is used to further improve the approximation to (6.2.12). It is possible that $\mathbf{p}_k = \mathbf{p}_{k+1}$ in which case, the candidate direction is not a descent direction.

At the kth iteration not all points $\mathbf{y}_j$, $j \in \{1, \ldots, k\}$ are used to approximate (6.2.12). First, as k increases, the distance between $\mathbf{p}_k$ and $\mathbf{y}_j$ with smaller indices j may be become too large for a good approximation. Second, storing the information for all k points is costly. At each iteration k, the algorithm uses two sets of indices M_k^f and M_k^G such that $M_k^f \subset \{1, \ldots, k\}$ and $M_k^G \subset \{1, \ldots, k\}$. In addition, $k \in M_k^f$ and $k \in M_k^G$. These sets define bundles of points, hence the name bundle method. The method uses $\boldsymbol{\zeta}_j^f$, an element $\Gamma f(\mathbf{y}_j)$, $j \in M_k^f$ and $\boldsymbol{\zeta}_j^G$, an element of $\Gamma G(\mathbf{y}_j)$, $j \in M_k^G$ to approximate (6.2.12).

Define:

$$f_{k,j} = f(\mathbf{y}_j) + \langle \boldsymbol{\zeta}_j^f, \mathbf{p}_k - \mathbf{y}_j \rangle, \ \forall j \in \{1, \ldots, k\}, \tag{6.3.1}$$

$$G_{k,j} = G(\mathbf{y}_j) + \langle \boldsymbol{\zeta}_j^G, \mathbf{p}_k - \mathbf{y}_j \rangle, \ \forall j \in \{1, \ldots, k\}, \tag{6.3.2}$$

$$s_{k,j} = \|\mathbf{y}_j - \mathbf{p}_k\|, \forall j \in M_k^f \cup M_k^G \tag{6.3.3}$$

$$\alpha_{k,j}^f = \max\{|f(\mathbf{p}_k) - f_{k,j}|, \gamma_f (s_{k,j})^2\}, \ \forall j \in \{1, \ldots, k\}, \tag{6.3.4}$$

$$\alpha_{k,j}^G = \max\{|G_{k,j}|, \gamma_G (s_{k,j})^2\}, \ \forall j \in \{1, \ldots, k\}, \tag{6.3.5}$$

where $\gamma_G > 0$ and $\gamma_f > 0$ are constants. Equations (6.3.1) and (6.3.2) define linearizations of f and G at nearby points of $\mathbf{p}_k$. The quantities $f_{k,j}$ and $G_{k,j}$ are called *linearization values.* Equations (6.3.4) and (6.3.5) define a measure of the goodness of the linearizations. These quantities are called *locality measures.* The smaller the locality measure for a given linearization, the better the linearization approximates the improvement function in the neighborhood of $\mathbf{p}_k$.

The sets M_k^f and M_k^G are constructed such that

$$M_k^f = \{j \in \{1, \ldots, k\} : \ \alpha_{k,j}^f < \bar{a}\} \tag{6.3.6}$$

$$M_k^G = \{j \in \{1, \ldots, k\} : \ \alpha_{k,j}^G < \bar{a}\} \tag{6.3.7}$$

hold where $\bar{a}$ is a finite positive number. At the kth iteration of the algorithm, a $\mathbf{p}_{k+1}$ and $\mathbf{y}_{k+1}$ are calculated. $\tilde{M}_k^f$ and $\tilde{M}_k^G$, subsets of M_k^f and M_k^G, respectively are determined such that the sets $M_{k+1}^f = \tilde{M}_k^f \cup \{k+1\}$ and $M_{k+1}^G = \tilde{M}_k^G \cup \{k+1\}$ satisfy (6.3.6) and (6.3.7), respectively. The process of removing elements from the bundle is called *distance resetting.*

The following assumption is necessary to make sure that the sequences $\{\mathbf{p}_k\}$, $\{\mathbf{y}_k\}$ and their limit points are subsets of $\mathcal{P}$.

Assumption 6.3.1. *Let $\bar{a}$ be a positive constant. Let $X = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{p}\|^2 \leq \bar{a}, \mathbf{p} \in S\}$.*

Then $\mathrm{cl}\,(X) \subset \mathcal{P}$ holds.

The approximation of (6.2.12) is

$$\begin{aligned} &\min_{\Delta,\mathbf{d}} \Delta + \frac{1}{2}\|\mathbf{d}\|^2 \\ &\text{s.t.} - \alpha_{k,j}^f + \langle \boldsymbol{\zeta}_j^f, \mathbf{d}\rangle \leq \Delta,\ \forall j \in M_k^f, \\ &\quad - \alpha_{k,j}^G + \langle \boldsymbol{\zeta}_j^G, \mathbf{d}\rangle \leq \Delta,\ \forall j \in M_k^G, \\ &\quad \mathbf{d} \in \mathbb{R}^{n_p},\ \Delta \in \mathbb{R}. \end{aligned} \tag{6.3.8}$$

The properties of problem (6.3.8) are summarized in the following lemmas.

Lemma 6.3.2. *The program* (6.3.8) *satisfies the Slater constraint qualification for all* $k = 1, \ldots, \infty$.

Proof. Given $\mathbf{p}_k$, the points $\{\mathbf{y}_j : j \in M_k^f \cup M_k^G\}$ satisfy $\|\mathbf{p}_k - \mathbf{y}_j\| \leq \sqrt{\bar{a}}$ for all $j \in M_k^f \cup M_k^G$ per (6.3.6) and (6.3.7).

Let $Z = \mathbf{z} \in \mathbb{R}^{n_p} : \|\mathbf{p}_k - \mathbf{z}\| \leq \sqrt{\bar{a}}$. Note that Z is a compact set. Therefore, due to the upper semicontinuity of the linear Newton approximations and their compact-valuedness, one can find a constant C_k such that $\|\boldsymbol{\zeta}\| \leq C_k$ holds for all $\boldsymbol{\zeta} \in \Gamma f(\mathbf{z})$ and $\boldsymbol{\zeta} \in \Gamma g_i(\mathbf{z})$, $i = 1, \ldots, n_c$ for all $\mathbf{z} \in Z$. Assume that such constant does not exist. Let there exist an infinite sequence $\{\mathbf{z}_i\} \subset Z$ such that $\|\boldsymbol{\zeta}_i\| > i$ where $\boldsymbol{\zeta}_i \in \Gamma f(\mathbf{z}_i)$. The choice of Γf is arbitrary. The proof holds if any of the Γg_i are used instead. Since Z is compact, $\{\mathbf{z}_i\}$ has a convergent subsequence in Z by the Bolzano-Weierstrass Theorem. Let the limit point be $\bar{\mathbf{z}}$ and the converging subsequence be $\{\mathbf{z}_j\}$. Let O be a bounded open set such that $\Gamma f(\bar{\mathbf{z}}) \subset O$. By upper semicontinuity of Γf, there exists a neighborhood of $\bar{\mathbf{z}}$ such that $\boldsymbol{\zeta}_j \in O$. This contradicts the fact that $\lim_{j\to\infty} \|\boldsymbol{\zeta}_j\| = +\infty$. Hence, there exists a C_k as described.

It is obvious that (6.3.8) is a convex optimization problem. Let $\mathbf{d} \in \mathbb{R}^{n_p}$. The point $(2C_k\|\mathbf{d}\|, \mathbf{d})$ is a feasible point of (6.3.8) since $\|\langle \boldsymbol{\zeta}_j^f, \mathbf{d}\rangle\| \leq C_k\|\mathbf{d}\|$ and $\|\langle \boldsymbol{\zeta}_j^G, \mathbf{d}\rangle\| \leq C_k\|\mathbf{d}\|$

hold for all $j \in M_k^f \cup M_k^G$ and all locality measures are nonnegative. Note that in this case, the inequalities in the constraints are satisfied strictly. Hence, the Slater constraint qualification is satisfied. □

Lemma 6.3.3. *Problem* (6.3.8) *is a convex optimization problem. It has a unique solution. Let the solution be* $(\Delta_k, \mathbf{d}_k)$. *Furthermore, there exist* $\lambda_{k,j}$, $j \in M_k^f$, *and* $\mu_{k,j}$, $j \in M_k^G$ *satisfying*

$$(\lambda_{k,j})(-\alpha_{k,j}^f + \langle \boldsymbol{\zeta}_j^f, \mathbf{d}_k \rangle - \Delta_k) = 0, \ \forall j \in M_k^f, \tag{6.3.9}$$

$$(\mu_{k,j})(-\alpha_{k,j}^G + \langle \boldsymbol{\zeta}_j^G, \mathbf{d}_k \rangle - \Delta_k) = 0, \ \forall j \in M_k^G, \tag{6.3.10}$$

$$\sum_{j \in M_k^f} \lambda_{k,j} + \sum_{j \in M_G^f} \mu_{k,j} = 1, \tag{6.3.11}$$

$$\lambda_{k,j} \geq 0, \ \forall j \in M_k^f, \ \mu_{k,j} \geq 0, \ \forall j \in M_k^G, \tag{6.3.12}$$

$$\mathbf{d}_k = -\left(\sum_{j \in M_k^f} \lambda_{k,j} \boldsymbol{\zeta}_j^f + \sum_{j \in M_k^G} \mu_{k,j} \boldsymbol{\zeta}_j^G \right), \tag{6.3.13}$$

$$\Delta_k = -\left(\|\mathbf{d}_k\|^2 + \sum_{j \in M_k^f} \lambda_{k,j} \alpha_{k,j}^f + \sum_{j \in M_k^G} \mu_{k,j} \alpha_{k,j}^G \right). \tag{6.3.14}$$

Finally, $\lambda_{k,j}$, $j \in M_k^f$, *and* $\mu_{k,j}$, $j \in M_k^G$ *satisfy the above conditions if and only if they constitute a solution of the dual problem*

$$\min_{\boldsymbol{\lambda}, \boldsymbol{\mu}} \ \frac{1}{2} \|\tilde{\mathbf{d}}_k\|^2 + \sum_{j \in M_k^f} \lambda_j \alpha_{k,j}^f + \sum_{j \in M_k^G} \mu_j \alpha_{k,j}^G \tag{6.3.15}$$

$$\text{s.t.} \sum_{j \in M_k^f} \lambda_j + \sum_{j \in M_G^f} \mu_j = 1, \tag{6.3.16}$$

$$\lambda_j \geq 0, \ \forall j \in M_k^f, \ \mu_j \geq 0, \ \forall j \in M_k^G,$$

where

$$\tilde{\mathbf{d}}_k = -\left(\sum_{j\in M_k^f} \lambda_j \boldsymbol{\zeta}_j^f + \sum_{j\in M_k^G} \mu_j \boldsymbol{\zeta}_j^G\right). \tag{6.3.17}$$

Proof. The convex nature of (6.3.8) is obvious. Since the Slater constraint qualification holds per Lemma 6.3.2, equations (6.3.9), (6.3.10), (6.3.11), (6.3.12) and (6.3.13) follow from the necessary and sufficient conditions of optimality for this problem. Equation (6.3.14) is obtained after substituting (6.3.13) into (6.3.9) and (6.3.10) and summing up over all $j \in M_k^f \cup M_k^G$.

The solution $(\Delta_k, \mathbf{d}_k)$ is unique. Let

$$V_f(\mathbf{d}) = \max(-\alpha_{k,j}^f + \langle \boldsymbol{\zeta}_j^f, \mathbf{d}\rangle,\ j \in M_k^f)$$
$$V_G(\mathbf{d}) = \max(-\alpha_{k,j}^G + \langle \boldsymbol{\zeta}_j^G, \mathbf{d}\rangle,\ j \in M_k^G).$$

Then, problem (6.3.8) can be written as

$$\min_{\mathbf{d}} \frac{1}{2}\|\mathbf{d}\|^2 + \max(V_G(\mathbf{d}), V_f(\mathbf{d}))$$

where $\Delta_k = \max(V_G(\mathbf{d}_k), V_f(\mathbf{d}_k))$. This reformulation is possible because an optimal solution of (6.3.8) has to satisfy at least one of the constraints with equality. Otherwise, the optimal Δ_k can be further decreased, violating optimality. Note that V_G and V_f are convex functions. The maximum of two convex functions is convex. Since $\|\mathbf{d}\|^2$ is strictly convex, $(\Delta_k, \mathbf{d}_k)$ is unique.

The Lagrangian of (6.3.8) is

$$\mathcal{L}(\Delta, \mathbf{d}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \Delta + \frac{1}{2}\|\mathbf{d}\|^2 + \sum_{j\in M_k^f} (\lambda_j)(-\alpha_{k,j}^f + \langle \boldsymbol{\zeta}_j^f, \mathbf{d}\rangle - \Delta) + \tag{6.3.18}$$

$$\sum_{j \in M_k^G} (\mu_j)(-\alpha_{k,j}^f + \langle \boldsymbol{\zeta}_j^G, \mathbf{d} \rangle - \Delta).$$

The dual function q is

$$q(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{\Delta, \mathbf{d}} \mathcal{L}(\Delta, \mathbf{d}, \boldsymbol{\lambda}, \boldsymbol{\mu}), \text{ s.t. } \lambda_j \geq 0, \ \mu_j \geq 0, \ \forall j \in M_k^f \cup M_k^G.$$

Note that $q(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is finite only if (6.3.16) holds. Otherwise, one can pick an arbitrarily large negative Δ to obtain $q(\boldsymbol{\lambda}, \boldsymbol{\mu}) = -\infty$. If (6.3.16) holds, then $\mathcal{L}(\Delta, \mathbf{d}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ becomes a strictly convex function of $\mathbf{d}$ and the infimum becomes an unconstrained optimization problem. Due to the strict convexity, there exists a unique solution, $\tilde{\mathbf{d}}_k$.

Note that,

$$\nabla_{\mathbf{d}} \mathcal{L}(\Delta, \mathbf{d}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathbf{d} + \sum_{j \in M_k^f} \lambda_j \boldsymbol{\zeta}_j^f + \sum_{j \in M_k^f} \mu_j \boldsymbol{\zeta}_j^G.$$

Then the necessary condition of optimality $\nabla_{\mathbf{d}} \mathcal{L}(\Delta, \mathbf{d}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathbf{0}$ produces (6.3.17).

The dual problem becomes

$$\begin{aligned} \max_{\boldsymbol{\lambda}, \boldsymbol{\mu}} \ & -\frac{1}{2} \|\tilde{\mathbf{d}}_k\|^2 - \sum_{j \in M_k^f} \lambda_j \alpha_{k,j}^f - \sum_{j \in M_k^G} \mu_j \alpha_{k,j}^G \\ \text{s.t.} \ & \sum_{j \in M_k^f} \lambda_j + \sum_{j \in M_G^f} \mu_j = 1, \\ & \lambda_j \geq 0, \ \forall j \in M_k^f, \ \mu_j \geq 0, \ \forall j \in M_k^G, \end{aligned}$$

once (6.3.17) is substituted into the expression for $\mathbf{q}(\boldsymbol{\lambda}, \boldsymbol{\mu})$. Replacing the max with the min, provides the desired result.

The optimal solution value of (6.3.8) and (6.3.15) are equal by strong duality (Proposition 5.3.1 in [14]) and the solution of (6.3.15) are the multipliers satisfying (6.3.9), (6.3.10),

(6.3.11), (6.3.12) and (6.3.13). □

Lemma 6.3.3 shows that $-\mathbf{d}_k$ is in the convex hull of the elements of the set of linear Newton approximations. The objective of the dual problem implies that the effect of each element is weighed by the corresponding locality measure. In order to minimize the objective of the dual problem, $\lambda_{k,j}$ and $\mu_{k,j}$ corresponding to larger values of locality measures should be made as small as possible.

A new linearization needs to be added in case a direction of descent cannot be determined with the available linearizations. The new linearization needs to significantly change $\mathbf{d}_{k+1}$ obtained as the solution of (6.3.15). Mathematical discussion of what is meant by significant change is deferred for the latter part of this section where the convergence of the bundle method is discussed.

If $\mathbf{p}_{k+1} \neq \mathbf{p}_k$, then the linearization values and locality measures need to be updated. The linearization values can be updated as follows

$$f_{k+1,j} = f_{k,j} + \langle \boldsymbol{\zeta}_j^f, \mathbf{p}_{k+1} - \mathbf{p}_k \rangle, \ \forall j \in M_k^f,$$
$$G_{k+1,j} = G_{k+1,j} + \langle \boldsymbol{\zeta}_j^G, \mathbf{p}_{k+1} - \mathbf{p}_k \rangle, \ \forall j \in M_k^G.$$

In order to avoid storing the points $\{\mathbf{y}_k\}$, the locality measures are updated using

$$s_{k+1,j} = s_{k,j} + \|\mathbf{p}_k - \mathbf{p}_{k+1}\| \text{ if } j \in M_k^f \cup M_k^G,$$
$$s_{k+1,k+1} = \|\mathbf{y}_{k+1} - \mathbf{p}_{k+1}\|.$$

Note that $\|\mathbf{p}_{k+1} - \mathbf{y}_j\| \leq s_{k+1,j}$ for all $j \in M_{k+1}^f \cup M_{k+1}^G$.

The number of $\boldsymbol{\zeta}_j^f$ and $\boldsymbol{\zeta}_j^G$ stored can be kept at a minimum using *aggregation*. Aggregation combines the active constraints in the solution of (6.3.8) into two linear constraints, one for the constraints using $\boldsymbol{\zeta}_j^f$ and one for the constraints using $\boldsymbol{\zeta}_j^G$. Let $\boldsymbol{\lambda}_k$ and $\boldsymbol{\mu}_k$ be

the Lagrange multipliers that solve (6.3.15) at iteration $k > 1$. Assume distance resetting has occurred at iteration $k-1$. Then, the *aggregate subgradients*, $\mathbf{v}_k^f$ and $\mathbf{v}_k^G$, are computed as the convex combination of the $\boldsymbol{\zeta}_j^f$ and $\boldsymbol{\zeta}_j^G$ using the solution of (6.3.15). Since $\mathbf{v}_k^f$ and $\mathbf{v}_k^G$ are convex combinations of the $\boldsymbol{\zeta}_j^f$ and $\boldsymbol{\zeta}_j^G$, they satisfy the bound discussed in Lemma 6.3.2. The *aggregate linearization values*, $\tilde{f}_k^v$ and $\tilde{G}_k^v$, computed similarly. The *aggregate locality measures*, $\tilde{\alpha}_k^{f,v}$ and $\tilde{\alpha}_k^{G,v}$, can be considered as the locality measures of imaginary points associated with aggregate subgradients and aggregate linearization values. The exact computation of these values is described in the next section as well as the method to update them from iteration to iteration.

Once the aggregate quantities are computed, they can be used in the next iteration of the algorithm if distance resetting does not occur. If distance resetting occurs at the kth iteration, then the aggregate quantities need to be discarded and (6.3.8) is again solved during the next iteration. The aggregate quantities are discarded because they may be derived from data obtained at points $\mathbf{y}_j$ such that $j \notin M_{k+1}^f \cup M_{k+1}^G$. These points are too far away from the next iterate, $\mathbf{p}_{k+1}$ and their linearizations do not carry relevant information anymore.

It is sufficient to consider the aggregate constraints in the $(k+1)$th iteration and the constraints associated with $\mathbf{y}_{k+1}$ to determine a descent direction as long as there is no distance resetting. Hence at every iteration, (6.3.8) can be formulated using at most four constraints. If distance resetting occurs in this case, then (6.3.8) has only two constraints and one has effectively restarted the algorithm. The parameter $\bar{a}$ can be chosen large enough to prevent excessive distance resetting. Furthermore, arbitrary subsets of M_k^f and M_k^G can be incorporated into M_{k+1}^f and M_{k+1}^G to minimize information loss due to distance resetting.

Let $\bar{k}$ be the index of the most recent iteration during which distance resetting occurred. Let $k \geq \bar{k}+1$. Let $r_k^a = 1$ if distance resetting occurred in the previous iteration and $r_k^a = 0$

otherwise. Then using aggregate quantities at iteration $k+1$, problem (6.3.8) becomes

$$\begin{aligned}
&\min_{\Delta,\mathbf{d}} \Delta + \frac{1}{2}\|\mathbf{d}\|^2 \qquad (6.3.19)\\
&\text{s.t.} - \alpha^f_{k+1,j} + \langle \boldsymbol{\zeta}^f_j, \mathbf{d}\rangle \le \Delta,\ \forall j \in M^f_{k+1},\\
&\quad - \alpha^{f,v}_{k+1} + \langle \mathbf{v}^f_k, \mathbf{d}\rangle \text{ if } r^a_{k+1} = 0,\\
&\quad - \alpha^G_{k+1,j} + \langle \boldsymbol{\zeta}^G_j, \mathbf{d}\rangle \le \Delta,\ \forall j \in M^G_{k+1},\\
&\quad - \alpha^{G,v}_{k+1} + \langle \mathbf{v}^G_k, \mathbf{d}\rangle \le \Delta, \text{ if } r^a_{k+1} = 0,\\
&\quad \mathbf{d} \in \mathbb{R}^{n_p},\ \Delta \in \mathbb{R},
\end{aligned}$$

where $M^f_{k+1} = \{k+1\} \cup \tilde{M}^f_k$, $M^G_{k+1} = \{k+1\} \cup \tilde{M}^G_k$, $\tilde{M}^f_k$ and $\tilde{M}^G_k$ are arbitrary subsets of M^f_k and M^G_k respectively and

$$\begin{aligned}
f^v_{k+1} &= \tilde{f}^v_k + \langle \mathbf{v}^f_k, \mathbf{p}_{k+1} - \mathbf{p}_k\rangle,\\
G^v_{k+1} &= \tilde{G}^v_k + \langle \mathbf{v}^G_k, \mathbf{p}_{k+1} - \mathbf{p}_k\rangle,\\
s^f_{k+1} &= \tilde{s}^f_k + \|\mathbf{p}_{k+1} - \mathbf{p}_k\|,\\
s^G_{k+1} &= \tilde{s}^G_k + \|\mathbf{p}_{k+1} - \mathbf{p}_k\|,\\
\alpha^{f,v}_{k+1} &= \max\{|f(\mathbf{p}_k) - f^v_{k+1}|, \gamma_f (s^f_{k+1})^2\},\\
\alpha^{G,v}_{k+1} &= \max\{|G^v_k|, \gamma_G (s^G_{k+1})^2\}.
\end{aligned}$$

The dual of (6.3.19) is

$$\begin{aligned}
\min_{\boldsymbol{\lambda}_k,\, \boldsymbol{\mu},\, \lambda^v,\, \mu^v} \frac{1}{2}\| &\sum_{j \in M^f_{k+1}} \lambda_j \boldsymbol{\zeta}^f_j + \lambda^v \mathbf{v}^f_k + \sum_{j \in M^G_{k+1}} \mu_j \boldsymbol{\zeta}^G_j + \mu^v \mathbf{v}^G_{k-1}\|^2 + \qquad (6.3.20)\\
&\sum_{j \in M^f_{k+1}} \lambda_j \alpha^f_{k+1,j} + \lambda^v \alpha^{f,v}_{k+1} + \sum_{j \in M^G_{k+1}} \mu_{k+1} \alpha^G_{k+1,j} + \mu^v_k \alpha^{G,v}_{k+1}
\end{aligned}$$

$$\text{s.t.} \sum_{j \in M_{k+1}^f} \lambda_j + \lambda^v + \sum_{j \in M_{k+1}^G} \mu_j + \mu^v = 1,$$

$$\lambda^v \geq 0,\ \mu^v \geq 0,$$

$$\lambda_j \geq 0, \forall j \in M_{k+1}^f,\ \mu_{k,j} \geq 0, \forall j \in M_{k+1}^G,$$

$$\lambda_k^v = \mu_k^v = 0 \text{ if } r_{k+1}^a = 1.$$

The bundle method described formally in the next section uses the aggregation technique.

6.3.1 Formal Statement of the Bundle Method with Linear Newton Approximations

Step 0: Initialize.

(a) Let $\epsilon_s \geq 0$.

(b) Let m_L, m_R, $\bar{a}$, $\bar{t}$ be such that $0 < m_L < m_R < 1$, $\bar{a} > 0$, $0 < \bar{t} \leq 1$.

(c) Let $\gamma_f > 0$ and $\gamma_G > 0$.

(d) Set $k = 1$, $j = 1$.

(e) Set $r_k^a = 1$.

(f) Let $\mathbf{p}_k \in S$.

(g) Set $\mathbf{y}_k = \mathbf{p}_k$, $s_{k,j} = s_k^f = s_k^G = 0$.

(h) Set $M_k^f = \{j\}$, $\boldsymbol{\zeta}_j^f \in \Gamma f(\mathbf{p}_k)$, $\mathbf{v}_{k-1}^f = \boldsymbol{\zeta}_j^f$, $f_{k,j} = f_k^v$.

(i) Set $M_k^G = \{j\}$, $\boldsymbol{\zeta}_j^G \in \Gamma G(\mathbf{p}_k)$, $\mathbf{v}_{k-1}^G = \boldsymbol{\zeta}_j^G$, $G_{k,j} = G_k^v$.

Step 1: Find Direction of Descent. Let $\boldsymbol{\lambda}_k$, $\boldsymbol{\mu}_k$, λ_k^v and μ_k^v be the solution of the following quadratic problem;

$$\min_{\boldsymbol{\lambda},\, \boldsymbol{\mu},\, \lambda^v,\, \mu^v} \frac{1}{2} \left\| \sum_{j \in M_k^f} \lambda_j \boldsymbol{\zeta}_j^f + \lambda^v \mathbf{v}_{k-1}^f + \sum_{j \in M_k^G} \mu_j \boldsymbol{\zeta}_j^G + \mu^v \mathbf{v}_{k-1}^G \right\|^2 + \tag{6.3.21}$$

$$\sum_{j\in M_k^f}\lambda_j\alpha_{k,j}^f+\lambda^v\alpha_k^{f,v}+\sum_{j\in M_k^G}\mu_j\alpha_{k,j}^G+\mu^v\alpha_k^{G,v} \tag{6.3.22}$$

$$\text{s.t.}\sum_{j\in M_k^f}\lambda_j+\lambda^v+\sum_{j\in M_k^G}\mu_j+\mu^v=1, \tag{6.3.23}$$

$$\lambda^v\geq 0,\ \mu^v\geq 0,\ \lambda_j\geq 0,\forall j\in M_k^f,\ \mu_j\geq 0,\forall j\in M_k^G,$$

$$\lambda^v=\mu^v=0 \text{ if } r_k^a=1,$$

where

$$\alpha_{k,j}^f=\max\{|f(\mathbf{p}_k)-f_{k,j}|,\gamma_f(s_{k,j})^2\},\ j\in M_k^f,$$

$$\alpha_{k,j}^G=\max\{|G_{k,j}|,\gamma_G(s_{k,j})^2\},\ j\in M_k^G,$$

$$\alpha_k^{f,v}=\max\{|f(\mathbf{p}_k)-f_k^v|,\gamma_f(s_k^f)^2\},$$

$$\alpha_k^{G,v}=\max\{|G_k^v|,\gamma_G(s_k^G)^2\}.$$

Compute

$$\nu_k^f=\sum_{j\in M_k^f}\lambda_{k,j}+\lambda_k^v,\ \nu_k^G=\sum_{j\in M_k^G}\mu_{k,j}+\mu_k^v.$$

If $\nu_k^f\neq 0$, set $\tilde{\lambda}_{k,j}=\lambda_{k,j}/\nu_k^f,\ j\in M_k^f,\ \tilde{\lambda}_k^v=\lambda_k^v/\nu_k^f$.

If $\nu_k^f=0$, set $\tilde{\lambda}_{k,k}=1,\ \tilde{\lambda}_{k,j}=0,\ j\in M_k^f\backslash\{k\},\ \tilde{\lambda}_k^v=0$.

If $\nu_k^G\neq 0$, set $\tilde{\mu}_{k,j}=\mu_{k,j}/\nu_k^G,\ j\in M_k^G,\ \tilde{\mu}_k^v=\mu_k^v/\nu_k^G$.

If $\nu_k^G=0$, set $\tilde{\mu}_{k,k}=1,\ \tilde{\mu}_{k,j}=0,\ j\in M_k^G\backslash\{k\},\ \tilde{\mu}_k^v=0$.

If $\lambda_k^v=\mu_k^v=0$ then compute $a_k=\max\{s_{k,j}:j\in M_k^f\cup M_k^G\}$.

Let

$$(\mathbf{v}_k^f,\tilde{f}_k^v,\tilde{s}_k^f)=\sum_{j\in M_k^f}\tilde{\lambda}_{k,j}(\boldsymbol{\zeta}_j^f,f_{k,j},s_{k,j})+\tilde{\lambda}_k^v(\mathbf{v}_{k-1}^f,f_k^v,s_k^f),$$

$$
\begin{aligned}
(\mathbf{v}_k^G, \tilde{G}_k^v, \tilde{s}_k^G) &= \sum_{j \in M_k^G} \tilde{\mu}_{k,j}(\boldsymbol{\zeta}_j^G, G_{k,j}, s_{k,j}) + \tilde{\mu}_k^v(\mathbf{v}_{k-1}^G, G_k^v, s_k^f), \\
\mathbf{v}_k &= \nu_k^f \mathbf{v}_k^f + \nu_k^G \mathbf{v}_k^G, \\
\mathbf{d}_k &= -\mathbf{v}_k, \\
\tilde{\alpha}_k^{f,v} &= \max\{|f(\mathbf{p}_k) - \tilde{f}_k^v|, \gamma_f(\tilde{s}_k^f)^2\}, \\
\tilde{\alpha}_k^{G,v} &= \max\{|\tilde{G}_k^v|, \gamma_G(\tilde{s}_k^G)^2\}, \\
\tilde{\alpha}_k^v &= \nu_k^f \tilde{\alpha}_k^{f,v} + \nu_k^G \tilde{\alpha}_k^{G,v}, \\
\Delta_k &= -(\|\mathbf{d}_k\|^2 + \tilde{\alpha}_k^v).
\end{aligned}
$$

Step 2: Check Stopping Criterion. Set

$$
w_k = \frac{1}{2}\|\mathbf{v}_k\|^2 + \tilde{\alpha}_k^v. \tag{6.3.24}
$$

If $w_k \leq \epsilon_s$, terminate. Otherwise go to Step 3.

Step 3: Do Line Search Using Algorithm 6.3.1.1, find two step sizes t_k^L and t_k^R such that $t_k^L \leq 1$, $0 \leq t_k^L \leq t_k^R$, $\mathbf{p}_{k+1} = \mathbf{p}_k + t_k^L \mathbf{d}_k$, $\mathbf{y}_{k+1} = \mathbf{p}_k + t_k^R \mathbf{d}_k$ and

$$
\begin{aligned}
f(\mathbf{p}_{k+1}) &\leq f(\mathbf{p}_k) + m_L t_k^L \Delta_k, \\
0 &\geq G(\mathbf{p}_{k+1}), \\
t_k^R &= t_k^L \text{ if } t_k^L \geq \bar{t}, \\
m_R \Delta_k &\leq -\alpha(\mathbf{p}_{k+1}, \mathbf{y}_{k+1}) + \langle \boldsymbol{\zeta}(\mathbf{y}_{k+1}), \mathbf{d}_k \rangle \text{ if } t_k^L < \bar{t}, \\
\bar{a}/2 &\geq \|\mathbf{y}_{k+1} - \mathbf{p}_{k+1}\|,
\end{aligned}
$$

where $\boldsymbol{\zeta}(\cdot)$ and $\alpha : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ are defined as

$$\boldsymbol{\zeta}(\mathbf{y}) \in \begin{cases} \Gamma f(\mathbf{y}) & \text{if } G(\mathbf{y}) \leq 0, \\ \Gamma G(\mathbf{y}) & \text{if } G(\mathbf{y}) > 0, \end{cases}$$

$$\alpha(\mathbf{x}, \mathbf{y}) = \begin{cases} \max\{|f(\mathbf{x}) - \bar{f}(\mathbf{x}; \mathbf{y})|, \gamma_f \|\mathbf{x} - \mathbf{y}\|^2\} & \text{if } G(\mathbf{y}) \leq 0, \\ \max\{|\bar{G}(\mathbf{x}; \mathbf{y})|, \gamma_G \|\mathbf{x} - \mathbf{y}\|^2\} & \text{if } G(\mathbf{y}) > 0, \end{cases}$$

$$\bar{f}(\mathbf{x}; \mathbf{y}) = f(\mathbf{y}) + \langle \boldsymbol{\zeta}(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \text{ if } G(\mathbf{y}) \leq 0,$$

$$\bar{G}(\mathbf{x}; \mathbf{y}) = G(\mathbf{y}) + \langle \boldsymbol{\zeta}(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \text{ if } G(\mathbf{y}) > 0,$$

hold.

Step 4: Update Linearizations Select $\tilde{M}_k^f \subset M_k^f$ and $\tilde{M}_k^G \subset M_k^G$. Let $M_{k+1}^f = \tilde{M}_k^f \cup \{k+1\}$ and $M_{k+1}^G = \tilde{M}_k^G \cup \{k+1\}$. Let

$$\begin{aligned} \boldsymbol{\zeta}_{k+1}^f &\in \Gamma f(\mathbf{y}_{k+1}), \\ \boldsymbol{\zeta}_{k+1}^G &\in \Gamma G(\mathbf{y}_{k+1}), \\ f_{k+1,k+1} &= f(\mathbf{y}_{k+1}) + \langle \boldsymbol{\zeta}_{k+1}^f, \mathbf{p}_{k+1} - \mathbf{y}_{k+1} \rangle, \\ f_{k+1,j} &= f_{k,j} + \langle \boldsymbol{\zeta}_j^f, \mathbf{p}_{k+1} - \mathbf{p}_k \rangle, \ \forall j \in \tilde{M}_k^f, \\ f_{k+1}^v &= \tilde{f}_k^v + \langle \mathbf{v}_k^f, \mathbf{p}_{k+1} - \mathbf{p}_k \rangle, \\ G_{k+1,k+1} &= G(\mathbf{y}_{k+1}) + \langle \boldsymbol{\zeta}_{k+1}^G, \mathbf{p}_{k+1} - \mathbf{y}_{k+1} \rangle, \\ G_{k+1,j} &= G_{k,j} + \langle \boldsymbol{\zeta}_j^G, \mathbf{p}_{k+1} - \mathbf{p}_k \rangle, \ \forall j \in \tilde{M}_k^G, \\ G_{k+1}^v &= \tilde{G}_k^v + \langle \mathbf{v}_k^G, \mathbf{p}_{k+1} - \mathbf{p}_k \rangle, \\ s_{k+1,k+1} &= \|\mathbf{y}_{k+1} - \mathbf{p}_{k+1}\|, \\ s_{k+1,j} &= s_{k,j} + \|\mathbf{p}_{k+1} - \mathbf{p}_k\|, \ \forall j \in \tilde{M}_k^f \cup \tilde{M}_k^G, \\ s_{k+1}^f &= \tilde{s}_k^f + \|\mathbf{p}_{k+1} - \mathbf{p}_k\|, \end{aligned}$$

$$s_{k+1}^G = \tilde{s}_k^G + \|\mathbf{p}_{k+1} - \mathbf{p}_k\|.$$

Step 5: Check Distance Resetting Criterion. Let $a_{k+1} = \max\{a_k + \|\mathbf{p}_{k+1} - \mathbf{p}_k\|, s_{k+1,k+1}\}$. If $a_{k+1} \leq \bar{a}$ then set $r_{k+1}^a = 0$ and go to Step 7. Otherwise set $r_{k+1}^a = 1$ and go to Step 6.

Step 6: Do Distance Reset. Remove the indices with smallest values in $\tilde{M}_k^f$ and $\tilde{M}_k^G$ until

$$a_{k+1} = \max\{s_{k+1,j} : \ j \in \tilde{M}_k^f \cup \tilde{M}_k^G\} \leq \bar{a}/2$$

holds.

Step 7: Increment Counter. Set $k = k + 1$ and go to Step 7.

Algorithm 6.3.1.1 Line Search Algorithm Using Linear Newton Approximations

---

Require: $\Delta_k < 0$
Require: $0 < m_L < m_R < 1$, $\bar{a} > 0$, $0 < \bar{t} \leq 1$.
Require: $\beta \in (0, 0.5)$.
Let $t_k^L = 0$ and $t = t^u = 1$.
while $t_k^L \neq t^u$ **do**
  if $f(\mathbf{p}_k + t\mathbf{d}) \leq f(\mathbf{p}_k) + m_L t\Delta_k$ and $G(\mathbf{p}_k + t\mathbf{d}) \leq 0$ **then**
    Set $t_k^L = t$.
  else
    Set $t^u = t$.
  end if
  if $t_k^L \geq \bar{t}$ **then**
    Set $t_k^R = t_k^L$ and return.
  else if $-\alpha(\mathbf{p}_k + t_k^L\mathbf{d}_k, \mathbf{p}_k + t\mathbf{d}_k) + \langle \boldsymbol{\zeta}(\mathbf{p}_k + t\mathbf{d}_k), \mathbf{d}_k \rangle \geq m_R\Delta_k$ and $t_k^L \leq \bar{t}$ and $(t - t_k^L)\|\mathbf{d}_k\| \leq \bar{a}/2$ **then**
    Set $t_k^R = t$ and return.
  else
    Let $t \in [t_k^L + \beta(t^u - t_k^L), t^u - \beta(t^u - t_k^L)]$.
  end if
end while

6.3.2 Discussion and Convergence

The proposed algorithm terminates when $w_k \leq \epsilon_S$. w_k can be considered as a quantity that measures the goodness of the approximation (6.2.12) via $\tilde{\alpha}_k^v$ and the size of the element of minimum norm via $\|\mathbf{v}_k\|^2$ in $\Gamma H(\mathbf{p}_k; \mathbf{p}_k)$. If $w_k = 0$, then $\mathbf{0} \in \Gamma H(\mathbf{p}_k; \mathbf{p}_k)$ holds. The convergence proof shows that all accumulation points of $\{\mathbf{p}_k\}$ are stationary in case $\epsilon_s = 0$. The proof is the same as the proof of Algorithm 3.1 in Chapter 6 in [54] where the linear Newton approximation replaces the generalized gradient. Therefore, a summary of it, is placed in the Appendix A.

In this chapter, the finite termination of the line search procedure is proved and the main results are stated.

Line Search

In order to show that any accumulation point of the sequence $\{\mathbf{p}_k\}$ is stationary, the finite termination of the line search algorithm needs to be shown. The line search algorithm differs from that of Algorithm 3.1 in Chapter 6 in [54] because linear Newton approximations are used instead of the generalized gradient. The next lemma proves an important property of linear Newton approximations that enables finite termination of Algorithm 6.3.1.1.

Lemma 6.3.4. *Let Assumption 6.2.1 hold. Let* $\mathbf{p} \in \mathcal{P}$. *Let* $\mathbf{d} \in \mathbb{R}^{n_p} \backslash \{\mathbf{0}\}$. *Let* $\{t_k\} \subset \mathbb{R}$ *be such that* $t_k > 0$, $\forall k$, $t_k \downarrow 0$ *and* $\boldsymbol{\zeta}_k^f \in \Gamma f(\mathbf{p} + t_k \mathbf{d})$. *Then*

$$f'(\mathbf{p}; \mathbf{d}) = \lim_{k \to \infty} \langle \boldsymbol{\zeta}_k^f, \mathbf{d} \rangle. \tag{6.3.25}$$

Let $\boldsymbol{\zeta}_k^G \in \Gamma G(\mathbf{p} + t_k \mathbf{d})$. *Then*

$$G'(\mathbf{p}; \mathbf{d}) = \lim_{k \to \infty} \langle \boldsymbol{\zeta}_k^G, \mathbf{d} \rangle. \tag{6.3.26}$$

Proof. Since Γf is a linear Newton approximation, for small enough t

$$\left\| f(\mathbf{p}+t_k\mathbf{d}) - f(\mathbf{p}) - \langle \zeta_k^f, t_k\mathbf{d}\rangle \right\| \leq \|t_k\mathbf{d}\|\Delta(\|t_k\mathbf{d}\|)$$

and therefore

$$\left\| \frac{f(\mathbf{p}+t_k\mathbf{d}) - f(\mathbf{p})}{t_k} - \langle \zeta_k^f, \mathbf{d}\rangle \right\| \leq \|\mathbf{d}\|\Delta(\|t_k\mathbf{d}\|)$$

holds. Since

$$\lim_{k\to\infty} \frac{f(\mathbf{p}+t_k\mathbf{d}) - f(\mathbf{p})}{t_k} = f'(\mathbf{p};\mathbf{d}),$$

and

$$\lim_{k\to\infty} \left\| \frac{f(\mathbf{p}+t_k\mathbf{d}) - f(\mathbf{p})}{t_k} - \langle \zeta_k^f, \mathbf{d}\rangle \right\| = 0 \tag{6.3.27}$$

(6.3.25) follows. Since ΓG is a linear Newton approximation of G (Corollary 6.2.10), the same reasoning proves (6.3.26). □

From the definitions of w_k and Δ_k it can be seen that $\Delta_k \leq -w_k$. Since $w_k \geq 0$, $\Delta_k \leq 0$. Note that if $\Delta_k = 0$, then $w_k = 0$ and the bundle method terminates before entering the line search algorithm. Hence, the line search is always entered with $\Delta_k < 0$. Note that it is possible that $\mathbf{d}_k = \mathbf{0}$ when entering the line search algorithm.

Theorem 6.3.5 (Convergence of Line Search Algorithm 6.3.1.1). *Assume $\Delta_k < 0$. Then the line search algorithm 6.3.1.1 terminates with two step sizes t_k^L and t_k^R such that $t_k^L \leq 1$, $0 \leq t_k^L \leq t_k^R$ and the points $\mathbf{p}_{k+1} = \mathbf{p}_k + t_k^L\mathbf{d}_k$ and $\mathbf{y}_{k+1} = \mathbf{p}_k + t_k^R\mathbf{d}_k$ satisfy*

$$f(\mathbf{p}_{k+1}) \leq f(\mathbf{p}_k) + m_L t_k^L \Delta_k, \tag{6.3.28}$$

$$
\begin{aligned}
0 &\geq G(\mathbf{p}_{k+1}), \\
t_k^R &= t_k^L \ \textit{if}\ t_k^L \geq \bar{t}, \\
m_R \Delta_k &\leq -\alpha(\mathbf{p}_{k+1}, \mathbf{y}_{k+1}) + \langle \boldsymbol{\zeta}(\mathbf{y}_{k+1}), \mathbf{d}_k \rangle \ \textit{if}\ t_k^L < \bar{t}, \\
\bar{a}/2 &\geq \|\mathbf{y}_{k+1} - \mathbf{p}_{k+1}\|,
\end{aligned}
\tag{6.3.29}
$$

where

$$
\boldsymbol{\zeta}(\mathbf{y}) \in \begin{cases} \Gamma f(\mathbf{y}) & \textit{if}\ G(\mathbf{y}) \leq 0, \\ \Gamma G(\mathbf{y}) & \textit{if}\ G(\mathbf{y}) > 0, \end{cases}
$$

$$
\alpha(\mathbf{x}, \mathbf{y}) = \begin{cases} \max\{|f(\mathbf{x}) - \bar{f}(\mathbf{x}; \mathbf{y})|, \gamma_f \|\mathbf{x} - \mathbf{y}\|^2\} & \textit{if}\ G(\mathbf{y}) \leq 0, \\ \max\{|\bar{G}(\mathbf{x}; \mathbf{y})|, \gamma_G \|\mathbf{x} - \mathbf{y}\|^2\} & \textit{if}\ G(\mathbf{y}) > 0, \end{cases}
$$

$$
\bar{f}(\mathbf{x}; \mathbf{y}) = f(\mathbf{y}) + \langle \boldsymbol{\zeta}(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \ \textit{if}\ G(\mathbf{y}) \leq 0,
$$

$$
\bar{G}(\mathbf{x}; \mathbf{y}) = G(\mathbf{y}) + \langle \boldsymbol{\zeta}(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \ \textit{if}\ G(\mathbf{y}) > 0.
$$

Proof. First assume that $\mathbf{d} = \mathbf{0}$. Then the line search terminates immediately with $t_k^L = 0$, $t_k^R = 1$ and $\mathbf{p}_{k+1} = \mathbf{p}_k = \mathbf{y}_{k+1}$.

Consider the case $\mathbf{d} \neq \mathbf{0}$ and assume that the line search does not terminate for contradiction purposes. Let t_j, $t_{k,j}^L$ and t_j^u represent the values of t, t_k^L and t^u after the jth execution of Step 3. Then $t_j = t_{k,j}^L$ or $t_j = t_j^u$.

Note that $(t_{j+1}^u - t_{k,j+1}^L) \leq (1-\beta)(t_j^u - t_{k,j}^L)$, $\{t_j^u\}$ is a monotonically decreasing and $\{t_{k,j}^L\}$ is a monotonically increasing sequence such that $t_{k,j}^L < t_j^u$. Hence there exists a $\tilde{t}$ such that $t_{k,j}^L \uparrow \tilde{t}$ and $t_j^u \downarrow \tilde{t}$.

Note that $\tilde{t} \leq \bar{t}$ because $t_{k,j}^L < \bar{t}$ for all j. Let $TL = \{t \geq 0 : \ f(\mathbf{p}_k + t\mathbf{d}_k) \leq f(\mathbf{p}_k) + m_L t \Delta_k, \ G(\mathbf{p}_k + t\mathbf{d}_k) \leq 0\}$. Since $\{t_{k,j}^L\} \subset TL$ and TL is a closed set due to the continuity

of f and G, $\tilde{t} \in TL$. Hence

$$f(\mathbf{p}_k + \tilde{t}\mathbf{d}) - f(\mathbf{p}_k) \leq m_L \tilde{t} \Delta_k. \tag{6.3.30}$$

and $G(\mathbf{p}_k + \tilde{t}\mathbf{d}) \leq 0$.

Since $t_j^u \notin TL$, $t_j^u \downarrow \tilde{t}$ and $t_{k,j}^L \to t_j^u$, there exists an infinite set $L \subset \{1, \ldots, \infty\}$, such that $t_j^u = t_j > \tilde{t}$ for all $j \in L$ and

$$f(\mathbf{p}_k + t_j\mathbf{d}) - f(\mathbf{p}_k) > m_L t_j \Delta_k, \ \forall j \in L. \tag{6.3.31}$$

Subtracting (6.3.31) from (6.3.30),

$$\frac{f(\mathbf{p}_k + t_j\mathbf{d}_k) - f(\mathbf{p}_k + \tilde{t}\mathbf{d}_k)}{(t_j - \tilde{t})} > m_L \Delta_k, \ \forall j \in L$$

is obtained. Taking the limit as $j \to \infty$ results in $f'(\mathbf{p}_k + \tilde{t}\mathbf{d}_k; \mathbf{d}_k) \geq m_L \Delta_k$.

For large enough $j \in L$, $(t_j - t_{k,j}^L)\|\mathbf{d}_k\| < \bar{a}/2$ because $t_j \to \tilde{t}$ and $t_{k,j}^L \to \tilde{t}$.

First assume $G(\mathbf{p}_k + t_j\mathbf{d}_k) \leq 0$ for all $j \in L$. Then $\boldsymbol{\zeta}(\mathbf{p}_k + t_j\mathbf{d}_k) \in \Gamma f(\mathbf{p}_k + t_j\mathbf{d}_k)$ and

$$\begin{aligned} \alpha(\mathbf{p}_k + \tilde{t}\mathbf{d}_k, \mathbf{p}_k + t_j\mathbf{d}_k) = \max\{&f(\mathbf{p}_k + \tilde{t}\mathbf{d}_k) - f(\mathbf{p}_k + t_j\mathbf{d}_k) - (\tilde{t} - t_j)\langle \boldsymbol{\zeta}(\mathbf{p}_k + t_j\mathbf{d}_k), \mathbf{d}_k \rangle, \\ &\gamma_f (t_j - \tilde{t})^2 \|\mathbf{d}_k\|^2\} \end{aligned}$$

for all $j \in L$.

Since the algorithm does not terminate $-\alpha(\mathbf{p}_k + \tilde{t}\mathbf{d}_k, \mathbf{p}_k + t_j\mathbf{d}_k) + \langle \boldsymbol{\zeta}(\mathbf{p}_k + t_j\mathbf{d}_k), \mathbf{d}_k \rangle < m_R \Delta_k$ for all $j \in L$. Note that $\lim_{j \to \infty} \alpha(\mathbf{p}_k + \tilde{t}\mathbf{d}_k, \mathbf{p}_k + t_j\mathbf{d}_k) = 0$ because $t_j \to \tilde{t}$ and Γf is locally bounded. Therefore

$$\lim_{j \to \infty} \langle \boldsymbol{\zeta}(\mathbf{p}_k + t_j\mathbf{d}_k), \mathbf{d}_k \rangle \leq m_R \Delta_k$$

for $j \in L$. This implies that $f'(\mathbf{p}_k + \tilde{t}\mathbf{d}_k; \mathbf{d}_k) \geq m_L\Delta_k > m_R\Delta_k \geq \lim_{j\to\infty}\langle \boldsymbol{\zeta}(\mathbf{p}_k + t_j\mathbf{d}_k), \mathbf{d}_k\rangle$ since $0 < m_L < m_R < 1$ and $\Delta_k < 0$. This is a contradiction with Lemma 6.3.4.

Assume $G(\mathbf{p}_k + t_j\mathbf{d}_k) > 0$ for all $j \in L$. Then $\boldsymbol{\zeta}(\mathbf{p}_k + t_j\mathbf{d}_k) \in \Gamma G(\mathbf{p}_k + t_j\mathbf{d}_k)$ and

$$\alpha(\mathbf{p}_k + \tilde{t}\mathbf{d}_k, \mathbf{p}_k + t_j\mathbf{d}_k) = \max\{G(\mathbf{p}_k + \tilde{t}\mathbf{d}_k) - G(\mathbf{p}_k + t_j\mathbf{d}_k) - (\tilde{t} - t_j)\langle \boldsymbol{\zeta}(\mathbf{p}_k + t_j\mathbf{d}_k), \mathbf{d}_k\rangle, \gamma_G(t_j - \tilde{t})^2\|\mathbf{d}_k\|^2\}.$$

for all $j \in L$. As before,

$$\lim_{j\to\infty}\langle \boldsymbol{\zeta}(\mathbf{p}_k + t_j\mathbf{d}_k), \mathbf{d}_k\rangle \leq m_R\Delta_k$$

holds. However; since $G(\mathbf{p}_k + \tilde{t}\mathbf{d}_k) \leq 0$,

$$\lim_{j\to\infty}\frac{G(\mathbf{p}_k + t_j\mathbf{d}_k) - G(\mathbf{p}_k + \tilde{t}\mathbf{d}_k)}{t_j - \tilde{t}} \geq 0 > m_L\Delta_k > m_R\Delta_k.$$

Again this contradicts with Lemma 6.3.4. Hence the line search terminates in a finite number of iterations. □

When the line search terminates, one of the following cases hold:

1. $t_k^L \geq \bar{t}$. This case is *the long serious step.*
2. $0 < t_k^L < \bar{t}$. This case is *the short serious step.*
3. $t_k^L = 0$. This case is *the null step.*

Main Convergence Results

Theorem 6.3.6. *Each accumulation point of the sequence* $\{\mathbf{p}_k\}$ *generated by the bundle method is stationary in the extended sense.*

Corollary 6.3.7. *If the level set* $P = \{\mathbf{p} \in \mathcal{P} : f(\mathbf{p}) \leq \mathbf{f}(\mathbf{p}_1)\}$ *is bounded,* $\mathrm{cl}\,(P) \subset \mathcal{P}$,

and the final accuracy tolerance ϵ_s is positive, then the bundle method terminates in a finite number of iterations.

6.3.3 Variants and Commentary

1. The generalized gradient can be replaced with linear Newton approximations satisfying Assumption 6.2.1 in all the algorithms developed in [54] to produce algorithms that converge to stationary points in the extended sense. Note that, one does not have to alter a given bundle code to make it work with linear Newton approximations.
2. In the remainder of the thesis, *the proximal bundle method* ([64], [68]), a variant of the developed algorithm is used to obtain numerical results. In the proximal bundle algorithm $\bar{a}$ is set to a large number to prevent reset and the quadratic problem

$$\begin{aligned}
\min_{\boldsymbol{\lambda},\, \boldsymbol{\mu},\, \lambda^v,\, \mu^v} \frac{1}{2\sigma}\| \sum_{j\in M_k^f} \lambda_j \boldsymbol{\zeta}_j^f + \lambda^v \mathbf{v}_{k-1}^f + \sum_{j\in M_k^G} \mu_j \boldsymbol{\zeta}_j^G + \mu^v \mathbf{v}_{k-1}^G\|^2 + \\
\sum_{j\in M_k^f} \lambda_j \alpha_{k,j}^f + \lambda^v \alpha_k^{f,v} + \sum_{j\in M_k^G} \mu_j \alpha_{k,j}^G + \mu^v \alpha_k^{G,v} \\
\text{s.t.} \sum_{j\in M_k^f} \lambda_j + \lambda^v + \sum_{j\in M_k^G} \mu_j + \mu^v = 1, \\
\lambda^v \geq 0,\ \mu^v \geq 0,\ \lambda_j \geq 0, \forall j \in M_k^f,\ \mu_j \geq 0, \forall j \in M_k^G, \\
\lambda^v = \mu^v = 0 \text{ if } r_k^a = 1,
\end{aligned} \tag{6.3.32}$$

is solved instead of (6.3.22). As a result, the candidate descent direction, $\mathbf{d}_k$ and Δ_k become

$$\begin{aligned}
\mathbf{d}_k &= -\frac{1}{\sigma}\mathbf{v}_k, \\
\Delta_k &= -\left(\frac{1}{\sigma}\|\mathbf{v}_k\|^2 + \tilde{\alpha}_k^v\right).
\end{aligned}$$

Hence, the extra parameter σ allows control over the step size taken and is adjusted depending on the progress of the bundle method algorithm [64]. Specifically, the software described in [64] is used to produce numerical results.

3. Using the linear Newton approximation instead of the generalized gradient leads to the loss of sharper results that can be obtained for bundle methods in case (6.0.1) is a convex program. The relationship between convexity and linear Newton approximations requires further research.

Chapter 7

Nonsmooth Dynamic Optimization of Systems with Varying Structure

This chapter describes a numerical method to solve nonsmooth nonlinear optimization problems where systems described by (4.4.1) are embedded as constraints.

The first section describes the numerical algorithm. The algorithm is assembled using results presented in the previous chapters. The second section discusses an extension of *control parameterization* ([40], [105]) from continuously differentiable vector fields to vector fields satisfying Assumption 5.1.1. In this approach, an open loop optimal control problem whose solution is a bounded measurable function is approximated by a sequence of nonlinear programs whose solutions consist of parameters defining piecewise constant functions in time. The convergence of the optimal solution values of the approximate problems to the optimal solution value of the optimal control problem as well as the convergence of the approximate solutions to the optimal control problem solution are discussed. The third section contains a technique with which minimum time problems can be solved. The results in the previous chapters can be applied to the solution of such problems once time is redefined as a continuous state of the system and the time horizon a parameter. The final section contains a review

of optimal control techniques related to the work in this thesis and a comparison with the presented numerical method.

7.1 The Nonsmooth Single Shooting Method

7.1.1 Formulation and Assumptions

The numerical method aims to solve the program:

$$\begin{aligned}
\min_{\mathbf{p}\in\mathcal{P}} J(\mathbf{p}) = &\sum_{k=1}^{n_e} \int_{\alpha_k}^{\beta_k} h_{0,k}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p}),\dot{\mathbf{x}}(t,\mathbf{p}))dt + \\
&H_0(\mathbf{p},\mathbf{x}(\beta_{n_e},\mathbf{p}),\mathbf{y}(\beta_{n_e},\mathbf{p}),\dot{\mathbf{x}}(\beta_{n_e},\mathbf{p})) \qquad (7.1.1)\\
\text{s.t.} \sum_{k=1}^{n_e} &\int_{\alpha_k}^{\beta_k} h_{i,k}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p}),\dot{\mathbf{x}}(t,\mathbf{p}))dt + \\
&H_i(\mathbf{p},\mathbf{x}(\beta_{n_e},\mathbf{p}),\mathbf{y}(\beta_{n_e},\mathbf{p}),\dot{\mathbf{x}}(\beta_{n_e},\mathbf{p})) \leq 0,\ \forall i \in \{1,\ldots,n_c\},\\
\mathbf{0} &= \mathbf{F}_k(t,\mathbf{p},\mathbf{x}_k(t,\mathbf{p}),\mathbf{y}_k(t,\mathbf{p}),\dot{\mathbf{x}}_k(t,\mathbf{p})),\ \forall t \in [\alpha_k,\beta_k],\ \forall k \in \mathcal{K},\\
\mathbf{0} &= \mathbf{x}_1(\alpha_1,\mathbf{p}) - \mathbf{F}_1^0(\mathbf{p}),\\
\mathbf{0} &= \mathbf{x}_k(\alpha_k,\mathbf{p}) - \mathbf{F}_k^0(\mathbf{p},\mathbf{x}_{i-1}(\beta_{i-1},\mathbf{p})),\ \forall k \in \mathcal{K}\backslash\{1\},\\
\mathbf{0} &= \mathbf{x}(t,\mathbf{p}) - \mathbf{x}_k(t,\mathbf{p}),\ \forall t \in [\alpha_k,\beta_k),\ \forall k \in \mathcal{K},\\
\mathbf{0} &= \mathbf{x}(\beta_{n_e},\mathbf{p}) - \mathbf{x}_{n_e}(\beta_{n_e},\mathbf{p}),\\
\mathbf{0} &= \mathbf{y}(t,\mathbf{p}) - \mathbf{y}_k(t,\mathbf{p}),\ \forall t \in [\alpha_k,\beta_k),\ \forall k \in \mathcal{K},\\
\mathbf{0} &= \mathbf{y}(\beta_{n_e},\mathbf{p}) - \mathbf{y}_{n_e}(\beta_{n_e},\mathbf{p}),\\
\mathbf{0} &= \dot{\mathbf{x}}(t,\mathbf{p}) - \dot{\mathbf{x}}_k(t,\mathbf{p}),\ \forall t \in [\alpha_k,\beta_k),\ \forall k \in \mathcal{K},\\
\mathbf{0} &= \dot{\mathbf{x}}(\beta_{n_e},\mathbf{p}) - \dot{\mathbf{x}}_{n_e}(\beta_{n_e},\mathbf{p}),
\end{aligned}$$

where

- n_c is a finite positive integer;
- n_e is a finite positive integer, $\mathcal{K} = \{1, \ldots, n_e\}$;
- α_k, β_k for all $k \in \mathcal{K}$ are as defined in Assumption 4.4.1;
- $\mathcal{T}_k$, $\mathcal{P}$, $\mathcal{X}$, $\mathcal{Y}$, $\dot{\mathcal{X}}$ are as defined in Assumption 4.4.1;
- The functions $h_{i,k} : \mathcal{T}_k \times \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}} \to \mathbb{R}$ and $H_i : \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \times \dot{\mathcal{X}} \to \mathbb{R}$ satisfy Assumption 5.1.1 for all $i \in \{0, \ldots, n_c\}$ and for all $k \in \mathcal{K}$;
- $\mathbf{x}$, $\dot{\mathbf{x}}$, $\mathbf{y}$, $\mathbf{x}_k$, $\dot{\mathbf{x}}_k$, $\mathbf{y}_k$ for all $k \in \mathcal{K}$ are as defined in Assumption 4.4.1;
- $\mathbf{F}_k$, $\mathbf{F}_k^0$ are as defined in Assumption 4.4.1 and satisfy Assumption 5.1.1 and the assumptions of Lemma 4.3.4 for all $k \in \mathcal{K}$.

Assumption 7.1.1. *For all* $\mathbf{p} \in \mathcal{P}$, *the solution* $(\mathbf{x}(\cdot, \mathbf{p}), \mathbf{y}(\cdot, \mathbf{p}), \dot{\mathbf{x}}(\cdot, \mathbf{p}))$ *exists.*

It can easily be shown that (7.1.1) is a locally Lipschitz continuous and semismooth optimization program using the properties of locally Lipschitz continuous and semismooth functions, Theorems 4.4.7, 4.4.8, 4.4.10, 3.4.7, 3.4.11 and 3.4.12.

In essence, Program (7.1.1) can be rewritten as

$$\min_{\mathbf{p} \in \mathcal{P}} f_0(\mathbf{p}) \text{ s.t. } f_i(\mathbf{p}) \leq 0, \ \forall i \in \{1, \ldots, n_c\}, \tag{7.1.2}$$

where f_i are locally Lipschitz continuous and semismooth functions of $\mathbf{p}$ and

$$f_i(\mathbf{p}) = \sum_{k=1}^{n_e} \int_{\alpha_k}^{\beta_k} h_{i,k}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p}), \dot{\mathbf{x}}(t, \mathbf{p}))dt + H_i(\mathbf{p}, \mathbf{x}(t_f, \mathbf{p}), \mathbf{y}(t, \mathbf{p}), \dot{\mathbf{x}}(t_f, \mathbf{p}))$$

holds for all $i \in \{0, \ldots, n_c\}$. Therefore, the bundle method described in Chapter 6 can be used to solve (7.1.1). In order to apply this bundle method, an element of the linear Newton approximations, Γf_0 and Γf_i, satisfying Assumption 6.2.1 need to be computed at $\mathbf{p}$. This in return, requires the computation of an element of $\Gamma_{\mathbf{p}}\mathbf{x}(t, \mathbf{p})$, $\Gamma_{\mathbf{p}}\mathbf{y}(t, \mathbf{p})$ and $\Gamma_{\mathbf{p}}\dot{\mathbf{x}}(t, \mathbf{p})$ as defined in Theorem 4.4.7 and Corollary 4.4.9. The details of these computations are discussed

next.

7.1.2 Computation of the elements of the linear Newton approximations of Program (7.1.2)

Let

$$Z_i(\mathbf{p}) = \sum_{k=1}^{n_e} \int_{\alpha_k}^{\beta_k} h_{i,k}(t, \mathbf{p}, \mathbf{x}(t,\mathbf{p}), \mathbf{y}(t,\mathbf{p}), \dot{\mathbf{x}}(t,\mathbf{p})), \ \forall i \in \{0,\ldots,n_c\}. \tag{7.1.3}$$

Then an element of $\Gamma_{\mathbf{p}} Z_i(\mathbf{p})$, a linear Newton approximation that satisfies $\partial_{\mathbf{p}} Z_i(\mathbf{p}) \subset \text{conv}\,(\Gamma_{\mathbf{p}} Z_i(\mathbf{p}))$, can be computed using Theorem 4.4.8. Another approach is to define the additional states $z_i : T \times \mathcal{P} \to \mathbb{R}$ that evolve in time according to

$$\begin{aligned} &\dot{z}_i(t,\mathbf{p}) = h_{i,k}(t, \mathbf{p}, \mathbf{x}(t,\mathbf{p}), \mathbf{y}(t,\mathbf{p}), \dot{\mathbf{x}}(t,\mathbf{p})), \forall t \in (\alpha_k, \beta_k], \ \forall k \in \mathcal{K}, \ \forall i \in \{0,\ldots,n_c\}, \\ &z_i(\alpha_1, \mathbf{p}) = 0, \ \forall i \in \{0,\ldots,n_c\}. \end{aligned} \tag{7.1.4}$$

The additional states $\{z_i\}_{i=0}^{n_c}$ and the corresponding equations in (7.1.4) can be appended to the system states $\mathbf{x}$ to obtain the augmented states, $\tilde{\mathbf{x}}$. Let $\tilde{\mathbf{x}}_{n_x+i+1} = z_i$ for all $i = 0,\ldots,n_c$. Theorem 4.4.7 can be used to compute an element of $\Gamma_{\mathbf{p}} \tilde{\mathbf{x}}(t_f, \mathbf{p})$, a linear Newton approximation which satisfies $\partial_{\mathbf{p}} \tilde{\mathbf{x}}(t_f,\mathbf{p}) \subset \text{conv}\,(\Gamma_{\mathbf{p}} \tilde{\mathbf{x}}(t_f,\mathbf{p}))$. Let $\mathbf{M}_i \in \mathbb{R}^{1\times(n_x+n_c+1)}$ for all $i \in \{0,\ldots,n_c\}$ be such that $M_{i,j} = 0$ for all $j \in \{1,\ldots,n_x+n_c+1\}\backslash\{n_x+i\}$ and $M_{i,n_x+i} = 1$. Note that $\mathbf{M}_i$ is the Jacobian of the mapping $(\mathbf{x}(t_f,\mathbf{p}), (z_0(t_f,\mathbf{p}),\ldots,z_{n_c}(t_f,\mathbf{p}))) \mapsto z_i(t_f,\mathbf{p})$. Let $\Gamma_{\mathbf{p}} z_i(t_f,\mathbf{p})$ be the set

$$\{\mathbf{M}_i\mathbf{A} : \mathbf{A} \in \text{conv}\,(\Gamma_{\mathbf{p}} \tilde{\mathbf{x}}(t_f,\mathbf{p}))\}.$$

Then per Theorem 2.8.12, it is a linear Newton approximation of the mapping $\boldsymbol{\eta} \mapsto z_i(t_f, \boldsymbol{\eta})$ at $\mathbf{p} \in \mathcal{P}$. Since $\partial_{\mathbf{p}} \tilde{\mathbf{x}}(t_f, \mathbf{p}) \subset \text{conv}\,(\Gamma_{\mathbf{p}} \tilde{\mathbf{x}}(t_f, \mathbf{p}))$,

$$\partial_{\mathbf{p}} z_i(t_f, \mathbf{p}) \subset \{\mathbf{M}_i \mathbf{A} : \mathbf{A} \in \text{conv}\,(\Gamma_{\mathbf{p}} \tilde{\mathbf{x}}(t_f, \mathbf{p}))\}$$

holds per Theorem 2.6.7. In the rest of the thesis, this approach is used to compute elements of the linear Newton approximations of z_i.

Consider the mapping $(\mathbf{x}(t_f, \mathbf{p}), (z_0(t_f, \mathbf{p}), \ldots, z_{n_c}(t_f, \mathbf{p}))) \mapsto \mathbf{x}(t_f, \mathbf{p})$. Let $\mathbf{N} \in \mathbb{R}^{n_x \times (n_x + n_c + 1)}$ and $\mathbf{N} = [\mathbf{I}_{n_x}\ \mathbf{0}]$. Then, $\mathbf{N}$ is the Jacobian and therefore a linear Newton approximation of this mapping. Let $\Gamma_{\mathbf{p}} \mathbf{x}(t_f, \mathbf{p})$ be the set

$$\{\mathbf{N}\mathbf{A} : \mathbf{A} \in \text{conv}\,(\Gamma_{\mathbf{p}} \tilde{\mathbf{x}}(t_f, \mathbf{p}))\}.$$

Then $\Gamma_{\mathbf{p}} \mathbf{x}(t_f, \mathbf{p})$ is a linear Newton approximation of the mapping $\boldsymbol{\eta} \mapsto \mathbf{x}(t_f, \boldsymbol{\eta})$ at $\mathbf{p} \in \mathcal{P}$ per Theorem 2.8.12. It can be shown that $\partial_{\mathbf{p}} \mathbf{x}(t_f, \mathbf{p}) \subset \text{conv}\,(\Gamma_{\mathbf{p}} \mathbf{x}(t_f, \mathbf{p}))$ using Theorem 2.6.7 and the fact that $\partial_{\mathbf{p}} \tilde{\mathbf{x}}(t_f, \mathbf{p}) \subset \text{conv}\,(\Gamma_{\mathbf{p}} \tilde{\mathbf{x}}(t_f, \mathbf{p}))$.

Hence once an element of $\Gamma_{\mathbf{p}} \tilde{\mathbf{x}}(t_f, \mathbf{p})$ is computed, an element of $\Gamma_{\mathbf{p}} \mathbf{x}(t_f, \mathbf{p})$ can be recovered using Theorem 2.8.12. An element of $\Gamma_{\mathbf{p}} \dot{\mathbf{x}}(t_f, \mathbf{p})$ and an element of $\Gamma_{\mathbf{p}} \mathbf{y}(t_f, \mathbf{p})$ can be computed using Corollary 4.4.9.

The necessary linear Newton approximation information for H_i can be computed using Theorem 4.4.10. Alternatively, the chain rule for linear Newton approximations, $\Gamma_{\mathbf{p}} \mathbf{x}(t_f, \mathbf{p})$, $\Gamma_{\mathbf{p}} \mathbf{y}(t_f, \mathbf{p})$ and $\Gamma_{\mathbf{p}} \dot{\mathbf{x}}(t_f, \mathbf{p})$ can be used. It can be shown that the set

$$S = \left\{ \begin{bmatrix} \mathbf{I}_{n_p} \\ \mathbf{A} \\ \mathbf{B} \\ \mathbf{C} \end{bmatrix} : \mathbf{A} \in \text{conv}\,(\Gamma_{\mathbf{p}} \mathbf{x}(t_f, \mathbf{p})),\ \mathbf{B} \in \text{conv}\,(\Gamma_{\mathbf{p}} \mathbf{y}(t_f, \mathbf{p})),\ \mathbf{C} \in \text{conv}\,(\Gamma_{\mathbf{p}} \dot{\mathbf{x}}(t_f, \mathbf{p})) \right\}$$

is a linear Newton approximation of the mapping $\boldsymbol{\eta} \mapsto (\boldsymbol{\eta}, \mathbf{x}(t_f, \boldsymbol{\eta}), \mathbf{y}(t_f, \boldsymbol{\eta}), \dot{\mathbf{x}}(t_f, \boldsymbol{\eta}))$ at $\mathbf{p} \in \mathcal{P}$ by applying Theorem 2.8.12. Theorem 2.6.7 and the fact that the aforementioned linear Newton approximations contain the related generalized Jacobians can be used to show that conv (S) contains the generalized Jacobian of the mapping $\boldsymbol{\eta} \mapsto (\boldsymbol{\eta}, \mathbf{x}(t_f, \boldsymbol{\eta}), \mathbf{y}(t_f, \boldsymbol{\eta}), \dot{\mathbf{x}}(t_f, \boldsymbol{\eta}))$ at $\mathbf{p} \in \mathcal{P}$.

Let $w_i : \mathcal{P} \to \mathbb{R} : \boldsymbol{\eta} \mapsto H_i(\boldsymbol{\eta}, \mathbf{x}(t_f, \boldsymbol{\eta}), \mathbf{y}(t_f, \boldsymbol{\eta}), \dot{\mathbf{x}}(t_f, \boldsymbol{\eta}))$. Let $\Gamma_{\mathbf{p}} w_i(\mathbf{p})$ be the set

$$\{\mathbf{AB} : \mathbf{A} \in \partial H_i(\mathbf{p}, \mathbf{x}(t_f, \mathbf{p}), \mathbf{y}(t_f, \mathbf{p}), \dot{\mathbf{x}}(t_f, \mathbf{p})),\ \mathbf{B} \in \text{conv}\,(S)\}.$$

∂H_i is the generalized Jacobian and a linear Newton approximation of the function H_i because H_i is a semismooth function per Assumption 5.1.1. Hence $\Gamma_{\mathbf{p}} w_i(\mathbf{p})$ is a linear Newton approximation of the map $\boldsymbol{\eta} \mapsto H_i(\boldsymbol{\eta}, \mathbf{x}(t_f, \boldsymbol{\eta}), \mathbf{y}(t_f, \boldsymbol{\eta}), \dot{\mathbf{x}}(t_f, \boldsymbol{\eta}))$ at $\mathbf{p} \in \mathcal{P}$. The fact that $\partial_{\mathbf{p}} w_i(\mathbf{p}) \subset \Gamma_{\mathbf{p}} w_i(\mathbf{p})$ follows from Theorem 2.6.7 and the properties of S.

An element of $\partial H_i(\mathbf{p}, \mathbf{x}(t_f, \mathbf{p}), \mathbf{y}(t_f, \mathbf{p}), \dot{\mathbf{x}}(t_f, \mathbf{p}))$ can be computed using the fact that under Assumption 5.1.1, H_i is a PC^1 function and the properties listed in §2.7.1.

Finally, Γf_i defined by $\Gamma f_i(\mathbf{p}) = \Gamma_{\mathbf{p}} z_i(t, \mathbf{p}) + \Gamma w_i(\mathbf{p})$ is a linear Newton approximation of f_i satisfying Assumptions 6.2.1 for all $i \in \{0, \ldots, n_c\}$ per the calculus rules for the linear Newton approximation (see §2.8.5).

7.1.3 Description of the Method

The nonsmooth single shooting method is an iterative method consisting of two main elements (Figure 7-1).

1. The Modified Bundle Method: During iteration k, the modified bundle method uses the Objective and Constraint Evaluator to obtain $f_i(\mathbf{p}_k)$ and an element of $\Gamma f_i(\mathbf{p}_k)$ for all $i \in \{0, \ldots, n_c\}$. Then, the bundle method determines if $\mathbf{p}_k$ satisfies the stationary conditions in the extended sense. If it does, the single shooting method terminates.

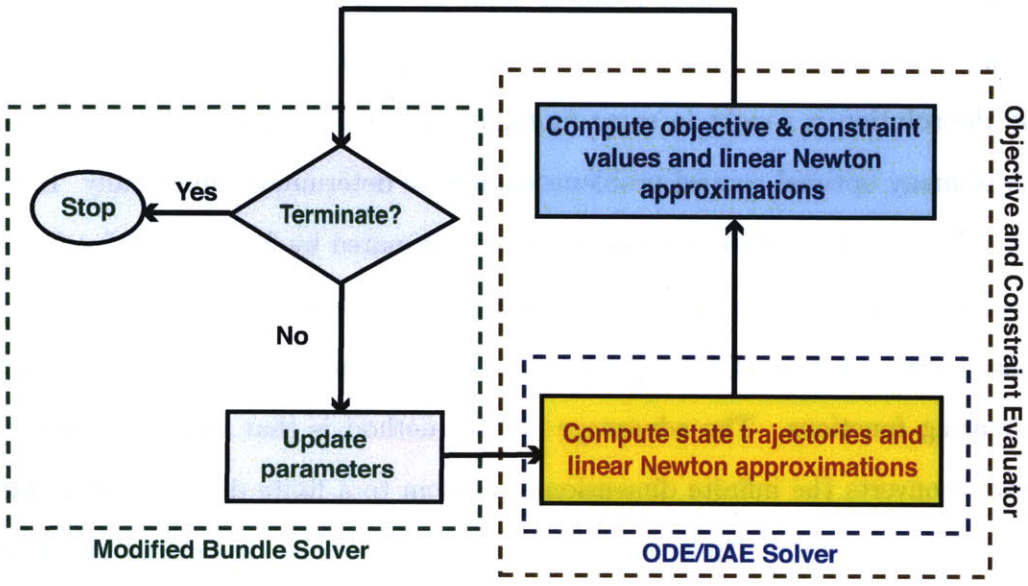


Figure 7-1: Nonsmooth Single Shooting Method.

Otherwise the modified bundle method updates $\mathbf{p}_k$ to obtain $\mathbf{p}_{k+1}$ and the iterative procedure continues.

2. The Objective and Constraint Evaluator: The ODE/DAE solver is used to compute $\tilde{\mathbf{x}}(t_f, \mathbf{p})$ and an element of $\Gamma_{\mathbf{p}}\tilde{\mathbf{x}}(t_f, \mathbf{p})$, $\Gamma_{\mathbf{p}}\mathbf{y}(t_f, \mathbf{p})$ and $\Gamma_{\mathbf{p}}\dot{\mathbf{x}}(t_f, \mathbf{p})$ each, using methods presented in Chapter 5. Then an element of $\Gamma f_i(\mathbf{p}_k)$ for each $\{0, \ldots, n_c\}$ is computed as described in §7.1.2.

7.2 Control Parameterization

Program (7.1.1) can be used to provide approximate solutions to certain open loop optimal control problems similar to those discussed in [40] and [105], where continuously differentiable vector fields are considered. In this section, the main results in [40] and Chapter 6 in [105] are extended to optimal control problems involving vector fields that satisfy Assumption (5.1.1).

An open loop optimal control problem seeks to find functions of time (controls) that satisfy the constraints and minimize the objective. It is an infinite dimensional problem because the solution is sought in some function space rather than in $\mathbb{R}^n$. In practice, the solution of many optimal control problems cannot be determined numerically. Instead, the elements of the function space of interest are approximated by functions defined in terms of a finite set of parameters, e.g., measurable functions are approximated by piecewise constant functions.[1] The solution of the optimal control problem is then sought in this class of approximating functions. The advantage of this method is that the use of finitely many parameters converts the infinite dimensional problem to a finite dimensional problem that can be solved numerically. Under certain conditions, it can be shown that the solutions of these finite dimensional problems converge to the optimal solution of the original problem as the number of parameters is increased.

The section begins with a brief summary of the necessary background information. Then, the control parameterization method is described. Finally, some results on the convergence of the approximate controls to the original optimal control are presented.

7.2.1 Preliminaries

Let I be a subset of $\mathbb{R}$ with positive measure. Let n and p be positive integers. Let $\mathcal{L}_p(I, \mathbb{R}^n)$ denote the class of measurable functions from I to $\mathbb{R}^n$ for which the quantity

$$\left(\int_I \|\mathbf{f}(t)\|^p dt \right)^{1/p}$$

[1] The controls can be approximated by other functions than piecewise constant.

is finite where $\mathbf{f}(t) = (f_1(t), \ldots, f_n(t))$ and $\|\mathbf{f}(t)\| = \sqrt{\sum_{i=1}^{n} f_i(t)^2}$. Note that $\mathcal{L}_p(I, \mathbb{R}^n)$ is a Banach space with respect to the norm:

$$\|\mathbf{f}\|_p = \left(\int_I \|\mathbf{f}(t)\|^p dt \right)^{\frac{1}{p}}.$$

A measurable function $\mathbf{f} : I \to \mathbb{R}^n$ is *essentially bounded* if there exists a positive number $C < +\infty$ such that the set $S = \{t \in I : \|\mathbf{f}(t)\| > C\}$ has Lebesgue measure zero. Let

$$\|\mathbf{f}\|_\infty = \text{ess sup}\{\|\mathbf{f}(t)\| : t \in I\}$$

where ess sup$\{\|\mathbf{f}(t)\| : t \in I\}$ denotes the smallest C such that the set $S = \{t \in I : \|\mathbf{f}(t)\| > C\}$ has Lebesgue measure zero. $\mathcal{L}_\infty(I, \mathbb{R}^n)$ denotes the class of essentially bounded measurable functions from I to $\mathbb{R}^n$. Note that $\mathcal{L}_\infty(I, \mathbb{R}^n)$ is a Banach space with respect to the norm $\|\mathbf{f}\|_\infty$. In addition, $\mathcal{L}_\infty(I, \mathbb{R}^n) \subset \mathcal{L}_1(I, \mathbb{R}^n)$.

A measurable function $\mathbf{f}$ belongs to $\mathcal{L}_1^{loc}$ if

$$\int_I \|\mathbf{f}(t)\| dt < +\infty \tag{7.2.1}$$

holds for all bounded I.

Lemma 7.2.1 (Gronwall's Lemma). *Let $f : [0, +\infty) \to \mathbb{R}$ and $\alpha : [0, +\infty) \to \mathbb{R}$ be continuous functions. Let $f(t) \geq 0$ and $\alpha(t) \geq 0$. Let $K \in \mathcal{L}_1^{loc}$ and $K(t) \geq 0$ a.e. Assume:*

$$f(t) \leq \alpha(t) + \int_0^t K(\tau) f(\tau) d\tau.$$

Then, for $0 \leq t < +\infty$,

$$f(t) \leq \alpha(t) + \int_0^t e^{\int_s^t K(\tau)d\tau} K(s)\alpha(s)ds$$

holds.

7.2.2 Description of Control Parameterization

The open loop optimal control problem of interest is:

$$\min_{\tilde{\mathbf{u}} \in \mathcal{L}_\infty(T,U)} \mathcal{J}(\tilde{\mathbf{u}}) = \int_{t_0}^{t_f} h_0(t, \tilde{\mathbf{u}}(t), \tilde{\mathbf{x}}(t, \tilde{\mathbf{u}}))dt + H_0(\tilde{\mathbf{x}}(t_f, \tilde{\mathbf{u}})) \tag{7.2.2}$$

$$\text{s.t.} \int_{t_0}^{t_f} h_i(t, \tilde{\mathbf{u}}(t), \tilde{\mathbf{x}}(t, \tilde{\mathbf{u}}))dt + H_i(\tilde{\mathbf{x}}(t_f, \tilde{\mathbf{u}})) \leq 0, \ \forall i \in \{1, \ldots, n_c\},$$

$$\dot{\tilde{\mathbf{x}}}(t, \tilde{\mathbf{u}}) = \mathbf{f}(t, \tilde{\mathbf{u}}(t), \tilde{\mathbf{x}}(t, \tilde{\mathbf{u}})), \ \forall t \in (t_0, t_f] \backslash S, \tag{7.2.3}$$

$$\tilde{\mathbf{x}}(t_0, \tilde{\mathbf{u}}) = \mathbf{x}_0, \tag{7.2.4}$$

where

- $T = [t_0, t_f]$,
- $\mathcal{T}$ is an open subset of $\mathbb{R}$ such that $T \subset \mathcal{T}$,
- n_q, n_x, n_c are finite positive integers,
- $U \subset \mathbb{R}^{n_q}$, $U = \{\mathbf{w} : \ c_j^L \leq w_j \leq c_j^U, \ -\infty < c_j^L < c_j^U < +\infty, \ \forall j \in \{1, \ldots, n_q\}\}$,
- $\mathcal{U}$ is an open subset of $\mathbb{R}^{n_q}$ such that $U \subset \mathcal{U}$ holds,
- $\mathcal{L}_\infty(T, U)$ is the set of essentially bounded measurable functions from T to U,
- S is a set of measure zero subset of T,
- $\tilde{\mathbf{x}} : T \times \mathcal{L}_\infty(T, U) \to \mathbb{R}^{n_x}$ is the continuous state of the system,
- $\mathbf{f} : \mathcal{T} \times \mathcal{U} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$, $h_i : \mathcal{T} \times \mathcal{U} \times \mathbb{R}^{n_x} \to \mathbb{R}$ and $H_i : \mathbb{R}^{n_x} \to \mathbb{R}$ satisfy Assumption 5.1.1 for all $i \in \{0, \ldots, n_c\}$.

Remark 7.2.2. Path constraints of the form $g(t, \tilde{\mathbf{u}}(t), \tilde{\mathbf{x}}(t, \tilde{\mathbf{u}})) \leq 0, \ \forall t \in [t_0, t_f]$ can be incorporated into this formulation by considering the constraints

$$\int_{t_0}^{t_f} \max(0, g(t, \tilde{\mathbf{u}}(t), \tilde{\mathbf{x}}(t, \tilde{\mathbf{u}}))dt \leq 0$$

or

$$\int_{t_0}^{t_f} \max(0, g(t, \tilde{\mathbf{u}}(t), \tilde{\mathbf{x}}(t, \tilde{\mathbf{u}}))^2)dt \leq 0.$$

The first constraint is nonsmooth whereas the second one is continuously differentiable if $g(t, \cdot)$ is continuously differentiable for all $t \in [t_0, t_f]$.

The solution of (7.2.2) if it exists is a measurable function which usually cannot be obtained numerically. Even if a measurable solution is known, it may not make sense to implement it in practice. Therefore, the measurable controls, $\tilde{\mathbf{u}}$ are approximated by piecewise constant functions[2] $\mathbf{u}^{n_e} : T \times \mathcal{P}^{n_e} \to U$ of the form

$$u_j^{n_e}(t, \mathbf{p}) = p_{k,j}, \ \forall t \in [\alpha_k, \beta_k), \ c_j^L \leq p_{k,j} \leq c_j^U, \ \forall k \in \{1, \ldots, n_e\}, \ \forall j \in \{1, \ldots, n_q\},$$
$$u_j^{n_e}(t_f, \mathbf{p}) = p_{n_e,j}, \ \ \forall j \in \{1, \ldots, n_q\},$$

where n_e, $\{\alpha_k\}_{k=1}^{n_e}$, $\{\beta_k\}_{k=1}^{n_e}$ are as defined in Assumption 4.4.1 and $\mathcal{P}^{n_e}$ is an open subset of $\mathbb{R}^{n_q \times n_e}$. Substituting $\mathbf{u}^{n_e}$ for $\tilde{\mathbf{u}}$,

$$\begin{aligned}
\min_{\mathbf{p} \in \mathcal{P}^{n_e}} \ J^{n_e}(\mathbf{p}) = & \int_{t_0}^{t_f} h_0(t, \mathbf{u}^{n_e}(t, \mathbf{p}), \mathbf{x}(t, \mathbf{p}))dt + H_0(\mathbf{x}(t_f, \mathbf{p})) \\
\text{s.t.} & \int_{t_0}^{t_f} h_i(t, \mathbf{u}^{n_e}(t, \mathbf{p}), \mathbf{x}(t, \mathbf{p}))dt + H_i(\mathbf{x}(t_f, \mathbf{p})) \leq 0, \ \forall i \in \{1, \ldots, n_c\}, \\
& \dot{\mathbf{x}}(t, \mathbf{u}^{n_e}(t, \mathbf{p})) = \mathbf{f}(t, \mathbf{u}^{n_e}(t, \mathbf{p}), \mathbf{x}(t, \mathbf{p})), \ \forall t \in (t_0, t_f] \backslash S,
\end{aligned}$$

[2]The measurable controls can be approximated by functions other than piecewise constant functions.

$$\mathbf{x}(t_0, \mathbf{u}^{n_e}(t, \mathbf{p})) = \mathbf{x}_0,$$

is obtained. This program can be written in the form

$$\begin{aligned} \min_{\mathbf{p}\in\mathcal{P}^{n_e}} \; J^{n_e}(\mathbf{p}) &= \int_{t_0}^{t_f} h_0(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}))dt + H_0(\mathbf{x}(t_f, \mathbf{p})) \\ \text{s.t.} &\int_{t_0}^{t_f} h_i(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}))dt + H_i(\mathbf{x}(t_f, \mathbf{p})) \leq 0, \; \forall i \in \{1, \ldots, n_c\}, \\ &\dot{\mathbf{x}}(t, \mathbf{p}) = \mathbf{f}_k(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})), \; \forall t \in (\alpha_k, \beta_k]\backslash S, \; \forall k \in \{1, \ldots, n_e\}, \\ &\mathbf{x}(t_0, \mathbf{p}) = \mathbf{x}_0 \end{aligned} \tag{7.2.5}$$

and can be solved using the nonsmooth single shooting method described in §7.1.

Convergence Results

In this section, the relationship between the approximate problem (7.2.5) and the original formulation (7.2.2) is analyzed. The results are similar to the results presented in Chapter 6 in [105] and in [40].

In this section, the following additional assumption holds.

Assumption 7.2.3.

For any compact set $\Omega \subset \mathcal{U}$, there exists a positive constant K such that

$$\|\mathbf{f}(t, \mathbf{v}, \mathbf{w})\| \leq K(1 + \|\mathbf{w}\|) \tag{7.2.6}$$

holds for all $(t, \mathbf{v}, \mathbf{w}) \in [t_0, t_f] \times \Omega \times \mathbb{R}^{n_x}$.

The following assumptions are made in [105] unlike the ones in this section:

1. For any compact set $\Omega \subset \mathcal{U}$, there exists a positive constant K such that

$$\|\mathbf{f}(t, \mathbf{v}, \mathbf{w})\| \leq K(1 + \|\mathbf{w}\|)$$

holds for all $(t, \mathbf{v}, \mathbf{w}) \in [t_0, t_f] \times \Omega \times \mathcal{X}$.

2. $\mathbf{f}(\cdot, \mathbf{v}, \mathbf{w})$ and $h_i(\cdot, \mathbf{v}, \mathbf{w})$ are piecewise continuous on $[t_0, t_f]$ for each $(\mathbf{v}, \mathbf{w}) \in \mathcal{U} \times \mathbb{R}^{n_x}$ for all $i \in \{0, \ldots, n_c\}$. $\mathbf{f}(t, \cdot)$, $h_i(t, \cdot)$ and H_i are continuously differentiable on $\mathcal{U} \times \mathbb{R}^{n_x}$ for all $t \in [t_0, t_f]$ and for all $i \in \{0, \ldots, n_c\}$.

The main result obtained in Chapter 6 in [105] can be obtained for the systems considered in this section. First, the necessary terminology to state the main results in Chapter 6 in [105] is introduced. Then, the main convergence results are stated. Later, the lemmas that differ in the proofs are presented.

Let $\mathbf{p}^{n_e,*}$ be an optimal solution of (7.2.5), $J^{n_e}(\mathbf{p}^{n_e,*})$ be the corresponding optimal solution value and $\mathbf{u}^{n_e,*}$ be the corresponding control. The convergence analysis makes use of the ϵ-relaxed problem

$$\begin{aligned} \min_{\mathbf{p} \in \mathcal{P}^{n_e}} \; J_\epsilon^{n_e}(\mathbf{p}) &= \int_{t_0}^{t_f} h_0(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}))dt + H_0(\mathbf{x}(t_f, \mathbf{p})) \\ \text{s.t.} \int_{t_0}^{t_f} & h_i(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}))dt + H_i(\mathbf{x}(t_f, \mathbf{p})) \leq \epsilon, \; \forall i \in \{1, \ldots, n_c\}, \\ & \dot{\mathbf{x}}(t, \mathbf{p}) = \mathbf{f}_k(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})), \; \forall t \in (\alpha_k, \beta_k] \backslash S, \; \forall k \in \{1, \ldots, n_e\}, \\ & \mathbf{x}(t_0, \mathbf{p}) = \mathbf{x}_0, \end{aligned} \tag{7.2.7}$$

where ϵ is a positive real number. Let $\mathbf{p}^{n_e,\epsilon,*}$ be an optimal solution of (7.2.7) for a given n_e and ϵ, $J_\epsilon^{n_e}(\mathbf{p}^{n_e,\epsilon,*})$ be the corresponding optimal solution value and $\mathbf{u}^{n_e,\epsilon,*}$ be the corresponding control. Results are obtained under the following regularity assumption:

Assumption 7.2.4. *There exists an integer $\bar{n}_e$ such that*

$$\lim_{\epsilon \downarrow 0} J_\epsilon^{n_e}(\mathbf{p}^{n_e,\epsilon,*}) = J^{n_e}(\mathbf{p}^{n_e,*}),$$

uniformly with respect to $n_e \geq \bar{n}_e$.

The next two theorems are the main convergence results.

Theorem 7.2.5. *Let Assumption* (7.2.4) *hold and suppose there exists an optimal solution $\tilde{\mathbf{u}}^*$ to problem* (7.2.2). *Then*

$$\lim_{n_e \to \infty} \mathcal{J}(\mathbf{u}^{n_e,*}) = \mathcal{J}(\tilde{\mathbf{u}}^*).$$

Theorem 7.2.6. *Let Assumption* (7.2.4) *hold. Suppose that*

$$\lim_{n_e \to \infty} \mathbf{u}^{n_e,*}(t, \mathbf{p}^{n_e,*}) \to \bar{\mathbf{u}}(t), \ \forall t \in [t_0, t_f] \backslash S,$$

where S is a measure zero subset of $[t_0, t_f]$. Then $\bar{\mathbf{u}}$ is an optimal solution of (7.2.2).

The proofs of the results are the same as the proof of Theorem 6.5.1 and Theorem 6.5.2 in [105] with Lemma 7.2.7 replacing Lemma 6.4.2 in [105] and Lemma 7.2.8 replacing Lemma 6.4.3 in [105].

Lemma 7.2.7. *Let $\{\mathbf{u}^{n_e}\}_{n_e=1}^{\infty}$ be a bounded sequence of functions in $\mathcal{L}_\infty(T, U)$. Then, the sequence $\{\tilde{\mathbf{x}}(\cdot, \mathbf{u}^{n_e})\}_{n_e=1}^{\infty}$ of corresponding solutions to* (7.2.3) *and* (7.2.4) *is also bounded and in $\mathcal{L}_\infty(T, \mathbb{R}^{n_x})$.*

Proof. Equations (7.2.3) and (7.2.4) can be stated as

$$\tilde{\mathbf{x}}(t, \mathbf{u}^{n_e}) = \mathbf{x}_0 + \int_{t_0}^{t} \mathbf{f}(\tau, \mathbf{u}^{n_e}(t, \mathbf{p}^{n_e}), \tilde{\mathbf{x}}(\tau, \mathbf{u}^{n_e}))d\tau \tag{7.2.8}$$

for all $t \in [t_0, t_f]$. Using property (7.2.6), one obtains

$$\|\tilde{\mathbf{x}}(t, \mathbf{u}^{n_e})\| \leq \|\mathbf{x}_0\| + \int_{t_0}^{t} K(1 + \|\tilde{\mathbf{x}}(\tau, \mathbf{u}^{n_e})\|)d\tau,$$
$$\|\tilde{\mathbf{x}}(t, \mathbf{u}^{n_e})\| \leq \|\mathbf{x}_0\| + K(t_f - t_0) + K\int_{t_0}^{t} \|\tilde{\mathbf{x}}(\tau, \mathbf{u}^{n_e})\|d\tau.$$

Application of Gronwall's Lemma results in

$$\|\tilde{\mathbf{x}}(t, \mathbf{u}^{n_e})\| \leq (\|\mathbf{x}_0\| + K(t_f - t_0))e^{K(t_f - t_0)}, \ \forall t \in [t_0, t_f].$$

□

Lemma 7.2.8. *Let $\{\mathbf{u}^{n_e}\}_{n_e=1}^{\infty}$ be a bounded sequence of functions in $\mathcal{L}_\infty(T, U)$ that converge to a function $\bar{\mathbf{u}}$ a.e. on $[t_0, t_f]$. Let $\{\tilde{\mathbf{x}}(\cdot, \mathbf{u}^{n_e})\}_{n_e=1}^{\infty}$ be the corresponding solutions to* (7.2.3) *and* (7.2.4)*. Then*

$$\lim_{n_e \to \infty} \|\tilde{\mathbf{x}}(\cdot, \mathbf{u}^{n_e}) - \tilde{\mathbf{x}}(\cdot, \bar{\mathbf{u}})\|_\infty = 0 \tag{7.2.9}$$

and for each $t \in [t_0, t_f]$,

$$\lim_{n_e \to \infty} \|\tilde{\mathbf{x}}(t, \mathbf{u}^{n_e}) - \tilde{\mathbf{x}}(t, \bar{\mathbf{u}})\| = 0$$

holds.

Proof. Let C_0 satisfy $\|\mathbf{u}^{n_e}\|_\infty \leq C_0$ for all $n_e \geq 1$. Per Lemma 7.2.7, there exists a constant C_1 such that $\|\tilde{\mathbf{x}}(\cdot, \mathbf{u}^{n_e})\|_\infty \leq C_1$ for all $n_e \geq 1$. Let $\bar{\mathcal{X}} = \{\mathbf{v} \in \mathbb{R}^{n_x} : \ \|\mathbf{v}\| \leq C_1\}$.

Since $\mathbf{f}$ is a locally Lipschitz continuous function due to Assumption 5.1.1, U and $\bar{\mathcal{X}}$ are

compact sets, there exists K_1 such that the difference

$$\|\tilde{\mathbf{x}}(t, \mathbf{u}^{n_e}) - \tilde{\mathbf{x}}(t, \bar{\mathbf{u}})\| \leq \int_{t_0}^{t_f} \|\mathbf{f}(\tau, \mathbf{u}^{n_e}(\tau, \mathbf{p}^{n_e}), \tilde{\mathbf{x}}(\tau, \mathbf{u}^{n_e})) - \mathbf{f}(\tau, \bar{\mathbf{u}}(\tau), \tilde{\mathbf{x}}(\tau, \bar{\mathbf{u}}))\| d\tau$$

satisfies

$$\|\tilde{\mathbf{x}}(t, \mathbf{u}^{n_e}) - \tilde{\mathbf{x}}(t, \bar{\mathbf{u}})\| \leq \int_{t_0}^{t_f} K_1 \|\tilde{\mathbf{x}}(\tau, \mathbf{u}^{n_e})) - \tilde{\mathbf{x}}(\tau, \bar{\mathbf{u}})\| + K_1 \|\mathbf{u}^{n_e}(\tau, \mathbf{p}^{n_e}) - \bar{\mathbf{u}}(\tau)\| d\tau.$$

Applying Gronwall's Lemma,

$$\|\tilde{\mathbf{x}}(t, \mathbf{u}^{n_e}) - \tilde{\mathbf{x}}(t, \bar{\mathbf{u}})\| \leq K_1 e^{K_1(t_f - t_0)} \int_{t_0}^{t_f} \|\mathbf{u}^{n_e}(\tau, \mathbf{p}^{n_e}) - \bar{\mathbf{u}}(\tau)\| d\tau$$

is obtained and the desired results follow from the fact that $\mathbf{u}^{n_e}(t, \mathbf{p}^{n_e}) \to \bar{\mathbf{u}}(t)$ for all $t \in [t_0, t_f] \backslash S$ where S is a measure zero subset of $[t_0, t_f]$. □

Remark 7.2.9. Assumption 7.2.3 is required to prove that the state trajectories remain in a bounded subset of $\mathbb{R}^{n_x}$ as proven in Lemma 7.2.7. Therefore this assumption can be replaced with any other condition ensuring boundedness of the trajectories.

Remark 7.2.10. In practice, a suitable n_e is determined by solving (7.2.5) repeatedly using an increasing sequence of values for n_e until $J^{n_e}(\mathbf{p}^{n_e,*})$ stops changing significantly.

7.3 Minimum Time Problems

Formulation (7.1.1) does not cover situations where the duration of the time horizon needs to be minimized. In addition, results in Chapters 3 and 4 do not directly apply to such problems. In order to apply these results, the dynamics need to be transformed so that time is a state.

Consider the ordinary differential equation

$$\dot{\mathbf{x}}(t,\mathbf{p}) = \mathbf{f}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p})),\ \forall t \in (t_0,t_f],\ \mathbf{x}(t_0,\mathbf{p}) = \mathbf{f}_0(\mathbf{p}),\ \forall \mathbf{p} \in \mathcal{P}, \tag{7.3.1}$$

where the equation satisfies Assumptions 4.2.1 and 5.1.1.

Redefine t as $t : [0,1] \to \mathbb{R}$. Let $\Delta T = t_f - t_0$. Let $\hat{\mathbf{p}} = (\Delta T, t_0)$. Redefine $\mathbf{p} = (\mathbf{p}, \hat{\mathbf{p}})$. Let t be the solution of the following initial value problem

$$\frac{dt}{d\tau}(\tau,\mathbf{p}) = \Delta T,\ \forall \tau \in (0,1],\ t(0,\mathbf{p}) = t_0.$$

Hence $t(\tau,\mathbf{p}) = \tau\Delta T + t_0$. Note that

$$\frac{d\mathbf{x}}{d\tau}(t(\tau,\mathbf{p}),\mathbf{p}) = \dot{\mathbf{x}}(t(\tau,\mathbf{p}),\mathbf{p}) \cdot \frac{dt}{d\tau}(\tau,\mathbf{p})$$

holds per the chain rule for derivatives where $t(\tau,\mathbf{p}) = \tau\Delta T + t_0$. Then, the equations in (7.3.1) can be written as

$$\begin{aligned}
\frac{d\mathbf{x}}{d\tau}(t(\tau,\mathbf{p}),\mathbf{p}) &= \Delta T \cdot \mathbf{f}(t(\tau,\mathbf{p}),\mathbf{p},\mathbf{x}(t(\tau,\mathbf{p}),\mathbf{p})),\ \forall \tau \in (0,1], \\
\frac{dt}{d\tau}(\tau,\mathbf{p}) &= \Delta T,\ \forall \tau \in (0,1], \\
t(0,\mathbf{p}) = t_0,\ \mathbf{x}(t(0,\mathbf{p}),\mathbf{p}) &= \mathbf{f}_0(\mathbf{p}),\ \forall \mathbf{p} \in \mathcal{P}.
\end{aligned}$$

Let $\hat{\mathbf{x}} : [0,1] \times \mathcal{P} \to \mathcal{X}$ be defined by $\hat{\mathbf{x}}(\tau,\mathbf{p}) = \mathbf{x}(t(\tau,\mathbf{p}),\mathbf{p})$. Then, the final form of the equations becomes

$$\begin{aligned}
\frac{d\hat{\mathbf{x}}}{d\tau}(\tau,\mathbf{p}) &= \Delta T \cdot \mathbf{f}(t(\tau,\mathbf{p}),\mathbf{p},\hat{\mathbf{x}}(\tau,\mathbf{p})),\ \forall \tau \in (0,1], \\
\frac{dt}{d\tau}(\tau,\mathbf{p}) &= \Delta T,\ \forall \tau \in (0,1], \\
t(0,\mathbf{p}) = t_0,\ \hat{\mathbf{x}}(0,\mathbf{p}) &= \mathbf{f}_0(\mathbf{p}),\ \forall \mathbf{p} \in \mathcal{P}.
\end{aligned}$$

Let $\mathbf{g} : \mathcal{T} \times \mathcal{P} \times \mathcal{X} \to \mathbb{R}^{n_x+1}$ be defined by $\mathbf{g}(\eta_t, \boldsymbol{\eta}_\mathbf{p}, \boldsymbol{\eta}_\mathbf{x}) = (\Delta T \cdot \mathbf{f}(\eta_t, \boldsymbol{\eta}_\mathbf{p}, \boldsymbol{\eta}_\mathbf{x}), \Delta T)$. It can easily be shown that $\mathbf{g}$ satisfies Assumption 5.1.1. Note that this form of the dynamics is amenable to be used in minimum time problems because the duration and initial time are parameters of the ordinary differential equation.

This transformation can also be used on systems of the form

$$\mathbf{0} = \mathbf{F}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p}), \dot{\mathbf{x}}(t, \mathbf{p})), \ \forall t \in [t_0, t_f], \ \mathbf{x}(t_0, \mathbf{p}) = \mathbf{F}_0(\mathbf{p}), \ \forall \mathbf{p} \in \mathcal{P}, \tag{7.3.2}$$

satisfying Assumptions 4.3.1, 4.3.2 and 5.1.1. The transformed dynamics are

$$\begin{aligned}
\mathbf{0} &= \mathbf{F}(t(\tau, \mathbf{p}), \mathbf{p}, \hat{\mathbf{x}}(\tau, \mathbf{p}), \hat{\mathbf{y}}(\tau, \mathbf{p}), \dot{\hat{\mathbf{x}}}(\tau, \mathbf{p})), \ \forall \tau \in [0, 1], \\
\frac{d\hat{\mathbf{x}}}{d\tau}(\tau, \mathbf{p}) &= \Delta T \cdot \dot{\hat{\mathbf{x}}}(\tau, \mathbf{p}), \ \forall \tau \in (0, 1], \\
\frac{dt}{d\tau}(\tau, \mathbf{p}) &= \Delta T, \ \forall \tau \in (0, 1], \\
t(0, \mathbf{p}) &= t_0, \ \hat{\mathbf{x}}(0, \mathbf{p}) = \mathbf{F}_0(\mathbf{p}), \ \forall \mathbf{p} \in \mathcal{P},
\end{aligned}$$

where $\hat{\mathbf{x}}(\tau, \mathbf{p}) = \mathbf{x}(t(\tau, \mathbf{p}), \mathbf{p})$, $\hat{\mathbf{y}}(\tau, \mathbf{p}) = \mathbf{y}(t(\tau, \mathbf{p}), \mathbf{p})$, $\dot{\hat{\mathbf{x}}}(\tau, \mathbf{p}) = \dot{\mathbf{x}}(t(\tau, \mathbf{p}), \mathbf{p})$ and $\mathbf{p}$ is redefined as in the ODE case.

In the multistage case, the final transformed equations become

$$\begin{aligned}
\mathbf{0} &= \mathbf{F}_i(t(\tau, \mathbf{p}), \mathbf{p}, \hat{\mathbf{x}}_i(\tau, \mathbf{p}), \hat{\mathbf{y}}_i(\tau, \mathbf{p}), \dot{\hat{\mathbf{x}}}_i(\tau, \mathbf{p})), \ \forall \tau \in [i-1, i], \ \forall i \in \mathcal{I}, \\
\frac{d\hat{\mathbf{x}}_i}{d\tau}(\tau, \mathbf{p}) &= \Delta T_i \cdot \dot{\hat{\mathbf{x}}}_i(\tau, \mathbf{p}), \ \forall \tau \in (i-1, i], \ \forall i \in \mathcal{I}, \\
\frac{dt}{d\tau}(\tau, \mathbf{p}) &= \Delta T_i, \ \forall \tau \in (i-1, i], \ \forall i \in \mathcal{I}, \\
t(0, \mathbf{p}) &= t_0, \\
\mathbf{0} &= \hat{\mathbf{x}}_1(0, \mathbf{p}) - \mathbf{F}_1^0(\mathbf{p}), \\
\mathbf{0} &= \hat{\mathbf{x}}_i(i-1, \mathbf{p}) - \mathbf{F}_i^0(\mathbf{p}, \hat{\mathbf{x}}_{i-1}(i-1, \mathbf{p})), \ \forall i \in \mathcal{I} \backslash \{1\},
\end{aligned}$$

$$
\begin{aligned}
\mathbf{0} &= \hat{\mathbf{x}}(\tau, \mathbf{p}) - \hat{\mathbf{x}}_i(\tau, \mathbf{p}),\ \forall \tau \in [i-1, i),\ \forall i \in \mathcal{I}, \\
\mathbf{0} &= \hat{\mathbf{x}}(n_e, \mathbf{p}) - \hat{\mathbf{x}}_{n_e}(n_e, \mathbf{p}), \\
\mathbf{0} &= \hat{\mathbf{y}}(\tau, \mathbf{p}) - \hat{\mathbf{y}}_i(\tau, \mathbf{p}),\ \forall \tau \in [i-1, i),\ \forall i \in \mathcal{I}, \\
\mathbf{0} &= \hat{\mathbf{y}}(n_e, \mathbf{p}) - \hat{\mathbf{y}}_{n_e}(n_e, \mathbf{p}), \\
\mathbf{0} &= \dot{\hat{\mathbf{x}}}(\tau, \mathbf{p}) - \dot{\hat{\mathbf{x}}}_i(\tau, \mathbf{p}),\ \forall \tau \in [i-1, i),\ \forall i \in \mathcal{I}, \\
\mathbf{0} &= \dot{\hat{\mathbf{x}}}(n_e, \mathbf{p}) - \dot{\hat{\mathbf{x}}}_{n_e}(n_e, \mathbf{p}),
\end{aligned}
$$

where $\Delta T_i = \beta_i - \alpha_i$ for all $i \in \mathcal{I}$, the equations satisfy Assumption 4.4.1, the right-hand side functions satisfy Assumption 5.1.1 and $\mathbf{F}_i$ satisfy Assumption 4.3.2 for all $i \in \mathcal{I}$. Now, formulation (7.1.1) can be rewritten using the transformed dynamics to solve minimum time problems.

7.4 Dynamic Optimization Literature Review and Comparison with the State of the Art

In the previous sections, the nonsmooth single shooting method was introduced and its application to dynamic optimization problems in the context of control parameterization was presented. In this section, the place of this numerical algorithm within the state of the art is discussed. The reader is referred to [17] for an excellent overview of the available numerical methods in case the data of the problem is continuously differentiable.

Methods that solve the necessary conditions of optimality to determine an optimal control are called *indirect* methods. Necessary conditions of optimality exist for the case where the data of the problem is locally Lipschitz continuous ([25]). These conditions involve the generalized gradient. As a result, the equations defining the optimal control are differential inclusions. Currently, these conditions are not amenable to numerical computation. Hence,

there are no indirect methods applicable to the dynamic optimization problems that can be solved by the nonsmooth single shooting method.

Direct methods solve the optimal control problem by directly minimizing the objective. The nonsmooth single shooting method introduced in §7.1 is therefore a direct method. Direct methods usually convert the optimal control problem into a nonlinear mathematical program similar to (7.2.5) using control parameterization and apply nonlinear optimization techniques to obtain a solution. It is possible to use optimization methods that do not require gradient information (derivative-free methods) in this approach. However, derivative-free methods require significantly more time to solve problems compared to methods that make use of gradient information where this information is available. An instance of this behavior is presented in the Chemotherapy Scheduling Case Study in Chapter 8. Lastly, these methods mostly lack proven convergence properties.

Numerical methods that only use an ODE/DAE solver to compute the state trajectories and the derivative information are called *sequential (single shooting)* methods. Therefore, the method in §7.1 is called a single shooting method. In these methods, most of the computational time is spent solving the embedded dynamics and derivative information with an ODE/DAE solver. The use of such a solver guarantees that the state trajectories always satisfy the initial value problem for any parameter value. The number of variables in the optimization problem is the smallest for this approach compared to other approaches. However, single shooting methods can only solve problems whose embedded initial value problems can be solved using an ODE/DAE solver. Hence, problems involving unstable dynamics cannot be reliably solved using a single shooting method because the integration error in the ODE/DAE solver grows unbounded.

The rest of the methods are called *simultaneous* methods because the integration of the dynamics is accomplished with the aid of the optimization algorithm. Simultaneous methods that discretize the embedded initial value problem are called *transcription (collocation)*

methods. In this approach, a discretization scheme such as Radau collocation on finite elements is used to approximate the state trajectories. The discretization scheme results in additional optimization variables that represent the values of the states at each time point of the discretization. The optimization method determines the value of these variables as well as those of the original ones. Therefore, these methods result in large optimization problems even for systems with a small to medium number of states. The number of discretization time points and grid that yields a sufficiently accurate approximation of the trajectories is not *a priori* known. Therefore, the state trajectories obtained as a solution need to be checked by comparing them to trajectories obtained with an ODE/DAE solver. In practice, however, the number of discretization points is increased until the trajectories obtained stop changing significantly. For problems involving stiff or highly nonlinear dynamics, this approach leads to very large optimization problems. In addition, the large number of variables complicates the determination of the initial values for the parameters. An example of this behavior can be seen in the Electric Circuit Case Study in Chapter 8. *Multiple shooting* methods, which are simultaneous methods, partition the time horizon into smaller intervals called epochs. The initial conditions for the states used in the numerical integration of the dynamics for each epoch become decision variables. The state trajectories over each epoch are computed using an ODE/DAE solver using values for theses variables. Then, the optimization method adjusts the values of these variables so that the states obtained at the end of each epoch are consistent with the values of the variables that are the initial conditions for the following epoch. Note that a direct (indirect) method can be sequential or simultaneous. Multiple shooting methods were invented to overcome some of the drawbacks of single shooting methods. Unstable dynamic systems can be handled because the integration is carried out over shorter intervals of time preventing the integration error from growing unbounded.

Finally, there are approaches based on dynamic programming. However, these approaches require a lot of memory and computational effort. These approaches are suitable only for

problems where the number of state variables is small or where a special structure is present.

Dynamic optimization problems of systems with varying structure that can be solved using the nonsmooth single shooting method, can also be solved by direct single shooting methods that use derivative-free optimization algorithms and some direct transcription methods ([102, 106], [12]). In transcription methods for systems with varying structure, the vector field that determines the evolution of the system between two time points of the discretization needs to be determined. This is a selection problem that can be handled using integer variables as in [102, 106] or complementarity conditions as in [12]. In order to solve practical problems with the integer approach, the underlying dynamics need to be linear because the solvers can handle only linear constraints effectively. Nonlinear dynamics result in nonlinear constraints. Hence, any nonlinearity in the system equations needs to be linearized. This degrades the quality of the approximation of the state trajectories. The resultant mathematical program is a mixed-integer linear program, (MILP). The MPEC (Mathematical Programs with Equilibrium Constraints) approach in [12] uses special constraints called complementarity conditions. This method can deal with nonlinear dynamics. However, the complementarity conditions violate certain regularity conditions called constraint qualifications. Therefore, the resultant mathematical programs require special handling and optimization methods. Both approaches suffer from the shortcomings of transcription methods. Examples of the MPEC approach are presented in the Electric Circuit and the Cascading Tanks Case Studies. An example of the use of integer variables is presented in the Cascading Tanks Case Study. The performance of these transcription methods are compared to the nonsmooth single shooting method in Chapter 8.

Finally, the results in the previous chapters can be used to devise a nonsmooth multiple shooting method. The rigorous development of this method is part of future research.

Chapter 8

Case Studies

8.1 Electric Circuit

In this case study, the behavior of the nonsmooth single shooting method and the MPEC approach presented in [12] is compared on a dynamic system exhibiting significant nonlinear behavior and stiffness.

8.1.1 System Dynamics

The system is an the electrical circuit (Figure 8-1) consisting of the well-known FitzHugh-Nagumo ([49]) electrical circuit used in modeling neurons connected in parallel with a diode and resistor. The elements of the model are:

- t_0: initial time in seconds; $t_0 = 0.0$ s.
- t_f: final time in seconds; $t_f = 60.0$ s.
- T: time horizon in seconds; $T = [t_0, t_f]$
- ΔT: the duration of time in seconds; $\Delta T = 60.0$ s.
- n_e: number of epochs used in control vector parameterization; $n_e = 2$.
- $\mathcal{K}$: the index set for the epoch; $\mathcal{K} = \{1, \ldots, n_e\}$.

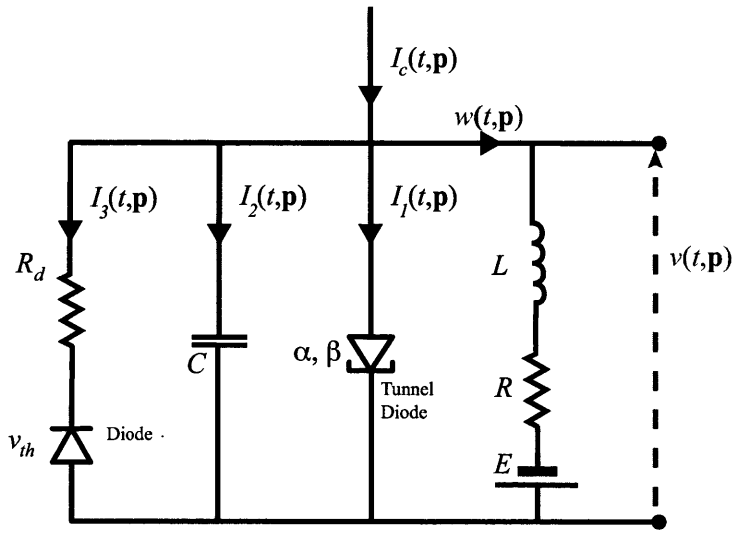


Figure 8-1: Electric Circuit: Configuration.

- $\{T_k\}_{k=1}^{n_e}$: the time intervals for each epoch. $T_k = [a_k, b_k)$ where $\mathbf{a}_k = t_0 + \frac{\Delta T \cdot (k-1)}{n_e}$ and $\mathbf{b}_k = t_0 + \frac{\Delta T \cdot (k)}{n_e}$ for all k in $\mathcal{K}$.
- $\{I_{c,k}\}_{k=1}^{n_e}$: control parameters of the system in amperes.
- $\mathbf{p}$: the parameters to be adjusted. $\mathbf{p} = (I_{c,1}, \ldots, I_{c,n_e})$.
- $\mathcal{P}$: the parameter set; $\mathcal{P}$ is an open subset of $\mathbb{R}^{n_e}$ such that $[0.00, 1.50]^{n_e} \subset \mathcal{P}$.
- $v : T \times \mathcal{P} \to \mathbb{R}$: the voltage of the circuit in volts.
- $w : T \times \mathcal{P} \to \mathbb{R}$: the current through the inductor in amperes. In the remainder of the case study unless otherwise noted, current refers to the current through the inductor.
- $\mathbf{x} : T \times \mathcal{P} \to \mathbb{R}^2 : (t, \mathbf{p}) \mapsto (v(t, \mathbf{p}), w(t, \mathbf{p}))$: is the continuous state of the system.
- $\mathcal{X}$: the state space; $\mathcal{X} = \mathbb{R}^2$.
- v_{th}: threshold voltage of the diode in volts; $v_{th} = -1.20\text{V}$.
- R_d: internal resistance of the diode; $R_d = 0.05\Omega$.
- C: capacitance of the capacitor in farads; $C = 1.00\text{F}$.
- (α, β): parameters of the tunnel diode; $\alpha = 1.00\ \frac{\text{A}}{\text{V}}$ and $\beta = \frac{1.00}{3.00}\frac{\text{A}}{\text{V}^3}$.
- L: inductance of the inductor in henries; $L = 12.5\text{H}$.
- R: resistance of the resistor in series with the inductor; $R = 0.8\Omega$.

- E: potential difference across the terminals of the battery in volts; $E = 0.7$V.
- $I_1 : T \times \mathcal{P} \rightarrow \mathbb{R}$: the current through the tunnel diode in amperes.
- $I_2 : T \times \mathcal{P} \rightarrow \mathbb{R}$: the current through the capacitor in amperes.
- $I_3 : T \times \mathcal{P} \rightarrow \mathbb{R}$: the current through the diode in amperes.
- $I_c : T \times \mathcal{P} \rightarrow \mathbb{R}$ the external current applied to the circuit. $I_c(t, \mathbf{p}) = I_{c,k}, \forall t \in \Delta T_k$ and $I_c(t_f, \mathbf{p}) = I_{c,n_e}$.

The numerical values of the parameters except the diode's threshold voltage and its internal resistance are the original numerical values of the FitzHugh-Nagumo model, although the units may be different.

The tunnel diode in Figure 8-1 is a passive circuit element whose voltage and current relationship is approximated by a third order polynomial. In the electrical circuit (Figure in 8-1), $I_1(t, \mathbf{p})$ and $v(t, \mathbf{p})$ satisfy

$$I_1(t, \mathbf{p}) = \beta v^3(t, \mathbf{p}) - \alpha v(t, \mathbf{p}). \tag{8.1.1}$$

The diode allows current to flow in one direction when the voltage across it is less than v_{th}. It determines $I_3(t, \mathbf{p})$ by

$$I_3(t, \mathbf{p}) = \min\left(\frac{v(t, \mathbf{p}) - v_{th}}{R_d}, 0\right). \tag{8.1.2}$$

Finally, $I_2(t, \mathbf{p})$ satisfies

$$I_2(t, \mathbf{p}) = C\dot{v}(t, \mathbf{p}). \tag{8.1.3}$$

Kirchoff's laws yield

$$I_c(t,\mathbf{p}) = w(t,\mathbf{p}) + I_1(t,\mathbf{p}) + I_2(t,\mathbf{p}) + I_3(t,\mathbf{p}), \tag{8.1.4}$$

$$v(t,\mathbf{p}) = L\dot{w}(t,\mathbf{p}) + Rw(t,\mathbf{p}) - E. \tag{8.1.5}$$

Substituting (8.1.1), (8.1.2) and (8.1.3) into (8.1.4) and (8.1.5) and solving for $\dot{v}$ and $\dot{w}$ produces the initial value problem

$$\dot{v}(t,\mathbf{p}) = \frac{1}{C}\left(I_c(t,\mathbf{p}) - \min\left(\frac{v(t,\mathbf{p}) - v_{th}}{R_{th}}, 0\right) - (\beta v^3(t,\mathbf{p}) - \alpha v(t,\mathbf{p})) - w(t,\mathbf{p})\right),\ \forall t \in (t_0, t_f],$$

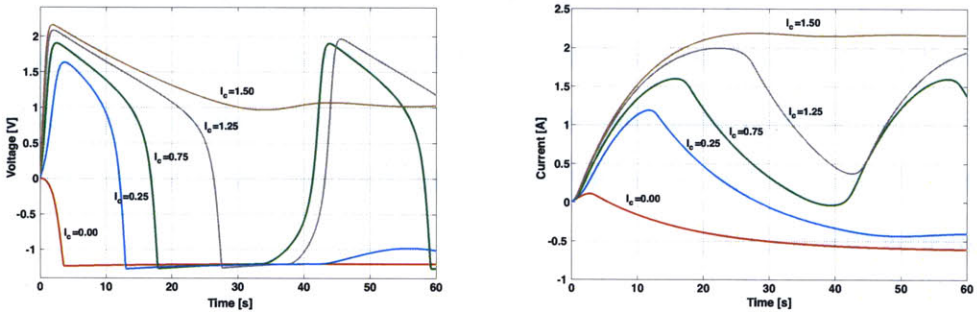
$$\dot{w}(t,\mathbf{p}) = \frac{1}{L}\left(-Rw(t,\mathbf{p}) + v(t,\mathbf{p}) + E\right),\ \forall t \in (t_0, t_f],$$

$$v(t_0,\mathbf{p}) = 0.0,\ w(t_0,\mathbf{p}) = 0.0.$$

The electric circuit exhibits different behavior for different current input as shown in Figures 8-2a and 8-2b. For low values of the input current, the circuit voltage spikes rapidly and then decays rapidly to a value around -1.20V. For larger input currents, the system shows oscillatory behavior. The presence of the diode causes a rapid change in the time derivative of the voltage when the voltage drops below -1.20V. The difference caused in the evolution of the states by the diode can be seen in in Figures 8-3a and 8-3b. Finally, for large enough current values, oscillations vanish and the voltage reaches a value close to 1.00V. The evolution of the current occurs relatively slow compared to the evolution of the voltage. This difference is especially pronounced at higher current values.

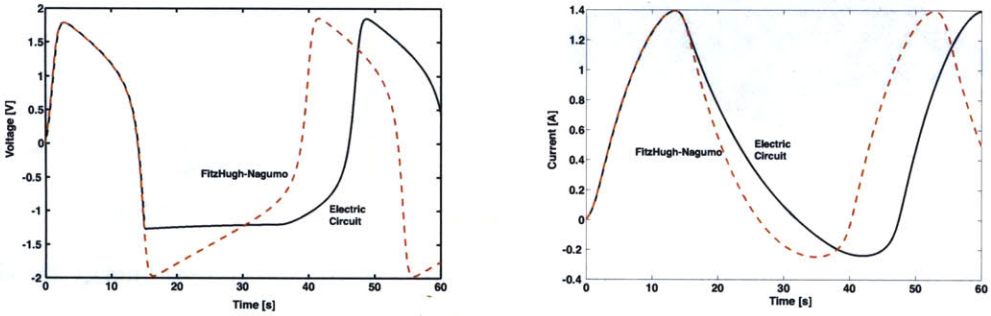
8.1.2 Dynamic Optimization Formulation

A dynamic optimization problem is solved to maximize the energy dissipated by the diode by adjusting $\{I_{c,k}\}_{k=1}^{n_e}$.



(a) Voltage (b) Current

Figure 8-2: Electric Circuit: Voltage and current trajectories for different constant I_c.



(a) Voltage (b) Current

Figure 8-3: Electric Circuit: Difference in voltage and current trajectories between FitzHugh-Nagumo and Electric Circuit Models if $I_c(t, \mathbf{p}) = 0.50\text{A}, \ \forall t \in [t_0, t_f]$.

The power dissipated by the diode is

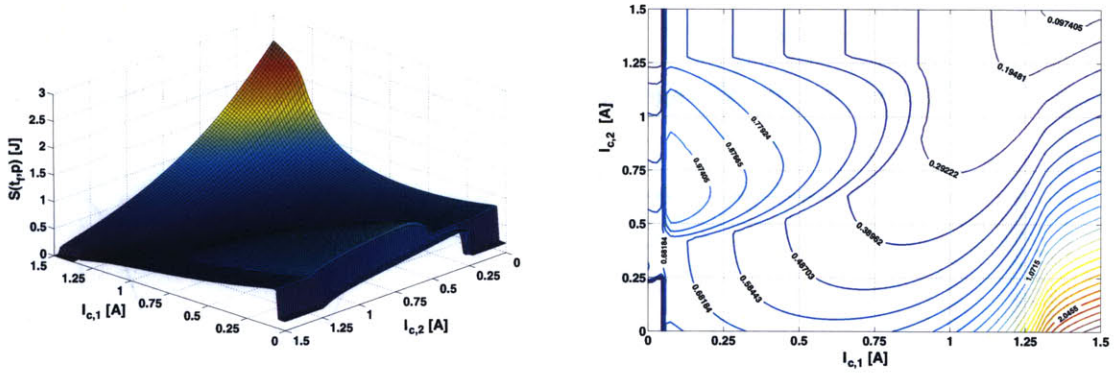
$$P(t, \mathbf{p}) = I_3(t, \mathbf{p})^2 R_d = \min\left(\frac{v(t, \mathbf{p}) - v_{th}}{R_d}, 0\right)(v(t, \mathbf{p}) - v_{th}).$$

Hence, the energy dissipated by the diode is computed by

$$\dot{S}(t, \mathbf{p}) = \min\left(\frac{v(t, \mathbf{p}) - v_{th}}{R_d}, 0\right)(v(t, \mathbf{p}) - v_{th}), \forall t \in (t_0, t_f], \ S(t_0, \mathbf{p}) = 0.0.$$

The plot of $S(t_f, \cdot)$ is in Figure 8-4. $S(t_f, \cdot)$ is a nonconvex function of the parameters. The function is fairly flat around $(0, 0)$, $(0, 1.5)$ and $(1.5, 1.5)$. The function values are zero in the

neighborhood of $(1.50, 1.50)$. The voltage never becomes less than v_{th} for parameter values close to $(1.50, 1.50)$. The function changes rapidly in the neighborhood of points $(0.05, y)$ where $y \in [0, 1.5]$. In this region, the evolution of the states transitions from non-oscillatory to oscillatory behavior. The function has a global maximum at $(1.5, 0.0)$ and a few local maxima such as $(0.076, 0.625)$.



(a) Surface Plot

(b) Contour Plot

Figure 8-4: Electric Circuit:Surface and contour plots of the dissipated energy by the diode

The formal statement of the dynamic optimization is

$$\max_{\mathbf{p}\in\mathcal{P}} J(\mathbf{p}) = S(t_f, \mathbf{p}) \tag{8.1.6}$$

$$\text{s.t. } 0.00 \leq I_{c,k} \leq 1.50,\ I_c(t, \mathbf{p}) = I_{c,k},\ t \in \Delta T_k,\ k \in \mathcal{K}, I_c(t_f, \mathbf{p}) = I_{c,k},$$

$$\dot{v}(t, \mathbf{p}) = \frac{1}{C}\left(I_c(t, \mathbf{p}) - \min\left(\frac{v(t, \mathbf{p}) - v_{th}}{R_d}, 0\right)\right) - \frac{1}{C}\left(\beta\, v^3(t, \mathbf{p}) + \alpha v(t, \mathbf{p}) - w(t, \mathbf{p})\right),\ \forall t \in (t_0, t_f],$$

$$\dot{w}(t, \mathbf{p}) = \frac{1}{L}\left(-Rw(t, \mathbf{p}) + v(t, \mathbf{p}) + E\right),\ \forall t \in (t_0, t_f],$$

$$\dot{S}(t, \mathbf{p}) = \min\left(\frac{v(t, \mathbf{p}) - v_{th}}{R_d}, 0\right)(v(t, \mathbf{p}) - v_{th}),\ \forall t \in (t_0, t_f],$$

$$v(t_0, \mathbf{p}) = 0.0,\ w(t_0, \mathbf{p}) = 0.0,\ S(t_0, \mathbf{p}) = 0.$$

This problem is solved using the nonsmooth single shooting method developed in this thesis

and the MPEC approach presented in [12].

8.1.3 Nonsmooth Single Shooting Method Results

The dynamic optimization problem (8.1.6) is solved with the method proposed in Chapter 7 for various initial guesses of for $I_{c,k}$ for all $k \in \{1, 2\}$. For the integration of the dynamics and auxiliary equations yielding derivative information DSL48SE ([108, 109, 36]) is used with code generated by DAEPACK ([107]). The nonlinear program is solved using the proximal bundle solver in [64] on a SUSE Linux 10 Virtual Machine with 1 GB of RAM and a 2.4 GHz Intel Core Duo CPU. The absolute and relative tolerances used in simulation are 1×10^{-8} and 1×10^{-8}, respectively. The optimality tolerance for the bundle solver is set to 1×10^{-6}.

Problem (8.1.6) is solved using various initial guesses for $\mathbf{p} = (I_{c,1}, I_{c,2})$. Table 8.1 summarizes the results for the test cases used in this study. The first column contains the label of the case. "$\mathbf{p}_0$" and "$\mathbf{p}^*$" represent the initial guess and final converged parameter values, respectively. The values are tabulated in columns 5 and 6. Columns 3 and 4 contain the initial and final objective values, respectively. Column 2 summarizes the termination status of the bundle solver. The column CPU contains the times taken to solve problem (8.1.6) numerically. NIT is the number of iterations done by the bundle solver and NFV is the number of times the dynamics are simulated. The behavior of the nonsmooth single

Case	Status	$J(\mathbf{p}_0)$	$J(\mathbf{p}^*)$	$\mathbf{p}_0$	$\mathbf{p}^*$	CPU [s]	NIT	NFV
A	Optimal	0.1786	0.1786	(0.000, 0.000)	(0.000, 0.000)	0.05	1	1
B	Optimal	0.1919	2.5325	(0.050, 0.050)	(1.500, 0.000)	0.36	4	7
C	Optimal	0.1166	2.5325	(1.400, 1.400)	(1.500, 0.000)	0.43	6	6
D	Optimal	0.3919	1.0649	(0.750, 1.250)	(0.076, 0.625)	1.24	12	12
E	Optimal	0.2590	1.0649	(1.000, 1.000)	(0.076, 0.625)	3.60	14	21
F	Optimal	0.7350	1.0649	(0.100, 1.250)	(0.076, 0.625)	2.26	16	24

Table 8.1: Electric Circuit: Nonsmooth single shooting results

shooting method is as expected. Due to the nonconvex nature of the objective, not all initial guesses for $\mathbf{p}$ lead to the final parameters equal to $(1.50, 0.00)$. The $\mathbf{p}^*$ obtained correspond to local maxima of the objective function.

8.1.4 MPEC Approach Results

MPEC Formulation

The transcription technique in [12] produces the mathematical program;

$$\max_X J(X) = S_{n_e,n_t} - \mu\Delta \tag{8.1.7}$$

$$\text{s.t: } I_{3,k,i+1} = (1-\lambda_{k,i+1})\left(\frac{v^-_{k,i+1}}{R_d}\right),\ \forall k \in \mathcal{K}, \forall i \in \mathcal{I}\backslash\{n_t\}, \tag{8.1.8}$$

$$\dot{w}_{k,i+1} = \frac{1}{L}\left(v_{k,i+1} - Rw_{k,i+1} + E\right),\ \forall k \in \mathcal{K}, \forall i \in \mathcal{I}\backslash\{n_t\}, \tag{8.1.9}$$

$$\dot{v}_{k,i+1} = \frac{1}{C}\left(\alpha v_{k,i+1} - \beta v^3_{k,i+1} + I_{c,k} + I_{3,k,i+1}\right),\ \forall k \in \mathcal{K}, \forall i \in \mathcal{I}\backslash\{n_t\}, \tag{8.1.10}$$

$$\dot{S}_{k,i+1} = I^2_{3,k,i+1}R_d,\ \forall k \in \mathcal{K}, \forall i \in \mathcal{I}\backslash\{n_t\}, \tag{8.1.11}$$

$$w_{k,i+1} - w_{k,i} = h_{k,i}\dot{w}_{k,i+1},\ \forall k \in \mathcal{K}, \forall i \in \mathcal{I}\backslash\{n_t\}, \tag{8.1.12}$$

$$v_{k,i+1} - v_{k,i} = h_{k,i}\dot{v}_{k,i+1},\ \forall k \in \mathcal{K}, \forall i \in \mathcal{I}\backslash\{n_t\}, \tag{8.1.13}$$

$$S_{k,i+1} - S_{k,i} = h_{k,i}\dot{S}_{k,i+1},\ \forall k \in \mathcal{K}, \forall i \in \mathcal{I}\backslash\{n_t\}, \tag{8.1.14}$$

$$v_{k+1,1} = v_{k,n_t},\ \forall k \in \mathcal{K}\backslash\{n_e\}, \tag{8.1.15}$$

$$w_{k+1,1} = w_{k,n_t},\ \forall k \in \mathcal{K}\backslash\{n_e\}, \tag{8.1.16}$$

$$S_{k+1,1} = S_{k,n_t},\ \forall k \in \mathcal{K}\backslash\{n_e\}, \tag{8.1.17}$$

$$v_{k,i+1} - v_{th} = v^+_{k,i+1} - v^-_{k,i+1},\ \forall k \in \mathcal{K}, \forall i \in \mathcal{I}\backslash\{n_t\}, \tag{8.1.18}$$

$$z^-_{k,i+1} = \lambda_{k,i+1}v^-_{k,i+1},\ \forall k \in \mathcal{K}, \forall i \in \mathcal{I}\backslash\{n_t\}, \tag{8.1.19}$$

$$z^+_{k,i+1} = (1-\lambda_{k,i+1})v^+_{k,i+1},\ \forall k \in \mathcal{K}, \forall i \in \mathcal{I}\backslash\{n_t\}, \tag{8.1.20}$$

$$\Delta = \sum_{k=1}^{n_e} \sum_{i=2}^{n_t} (z_{k,i}^- + z_{k,i}^+), \tag{8.1.21}$$

$$0 \leq v_{k,i}^+,\ 0 \leq v_{k,i}^-,\ 0 \leq \lambda_{k,i} \leq 1.0, \forall k \in \mathcal{K}, \forall i \in \mathcal{I}, \tag{8.1.22}$$

$$0 \leq z_{k,i}^-,\ 0 \leq z_{k,i}^+,\ \forall k \in \mathcal{K}, \forall i \in \mathcal{I}, \tag{8.1.23}$$

$$v_{1,1} = 0,\ w_{1,1} = 0,\ S_{1,1} = 0, \tag{8.1.24}$$

$$0 \leq I_{c,k} \leq 1.50,\ \forall k \in \mathcal{K}, \tag{8.1.25}$$

$$1 \times 10^{-6} \leq h_{k,i} \leq \frac{5 \times 60.0}{n_e \cdot (n_t - 1)},\ \forall k \in \mathcal{K}, \forall i \in \mathcal{I} \backslash \{n_t\}, \tag{8.1.26}$$

$$\sum_{i=1}^{n_t - 1} h_{k,i} = 30.0,\ \forall k \in \mathcal{K}, . \tag{8.1.27}$$

where

- μ is the positive penalty parameter;
- n_t is the number of finite elements in each epoch;
- $\mathcal{I} = \{1, \ldots, n_t\}$ is the index set for the finite elements;
- $\{\{z_{k,i}^-\}_{i=1}^{n_t}\}_{k=1}^{n_e}$ and $\{\{z_{k,i}^+\}_{i=1}^{n_t}\}_{k=1}^{n_e}$ are the deviations from the complementarity conditions in (8.1.19), (8.1.20), (8.1.22) and satisfy (8.1.23);
- Δ is the total deviation from the complementarity conditions and is computed using (8.1.21);
- $\{\{\lambda_{k,i}\}_{i=1}^{n_t}\}_{k=1}^{n_e}$, $\{\{v_{k,i}^+\}_{i=1}^{n_t}\}_{k=1}^{n_e}$ and $\{\{v_{k,i}^-\}_{i=1}^{n_t}\}_{k=1}^{n_e}$ are the complementarity variables;
- $\{\{h_{k,i}\}_{i=1}^{n_t}\}_{k=1}^{n_e}$ are the time steps;
- $\{\{v_{k,i}\}_{i=1}^{n_t}\}_{k=1}^{n_e}$ are the values of the voltage in epoch k and finite element i;
- $\{\{w_{k,i}\}_{i=1}^{n_t}\}_{k=1}^{n_e}$ are the values of the current in epoch k and finite element i;
- $\{\{S_{k,i}\}_{i=1}^{n_t}\}_{k=1}^{n_e}$ are the values of the energy dissipated by the resistor in epoch k and finite element i;
- and $\{\{I_{3,k,i}\}_{k=1}^{n_t}\}_{i=1}^{n_e}$ are the values of the $I_3(t, \mathbf{p})$ in epoch k and finite element i;
- $X = \{\{I_{c,k}\}_{k=1}^{n_e}, \{\{z_{k,i}^-\}_{i=1}^{n_t}\}_{k=1}^{n_e}, \{\{z_{k,i}^+\}_{i=1}^{n_t}\}_{k=1}^{n_e}, \{\{\lambda_{k,i}\}_{i=1}^{n_t}\}_{k=1}^{n_e}, \{\{v_{k,i}^+\}_{i=1}^{n_t}\}_{k=1}^{n_e},$

$$\{\{v^-_{k,i}\}_{i=1}^{n_t}\}_{k=1}^{n_e},\ \{\{w_{k,i}\}_{i=1}^{n_t}\}_{k=1}^{n_e},\ \{\{S_{k,i}\}_{i=1}^{n_t}\}_{k=1}^{n_e},\ \{\{I_{3,k,i}\}_{i=1}^{n_t}\}_{k=1}^{n_e}, \{\{h_{k,i}\}_{i=1}^{n_t}\}_{k=1}^{n_e}\}.$$

The differential equations of the circuit are discretized using an implicit Euler scheme and are in (8.1.9), (8.1.10), (8.1.11), (8.1.12), (8.1.13) and (8.1.14). Equations (8.1.15), (8.1.16) and (8.1.17) ensure continuity of the continuous states between epochs. Initial conditions are in (8.1.24). The time steps are part of the solution of the mathematical program and satisfy (8.1.26). In addition, the time steps in each epoch have to sum up to the epoch duration. This requirement is stated in (8.1.27). The constraints on $I_{c,k}$ are in (8.1.25).

Determination of n_t

The number of finite elements in each epoch, n_t determines the accuracy of approximation of the implicit Euler discretization scheme. If n_t is too small, the approximations to the state trajectories obtained as the solution of (8.1.7) are not accurate. If n_t is yet smaller, there may not even exist a feasible solution.

A feasibility problem is solved to determine a suitable n_t and obtain a feasible starting point, X_0. The constraint (8.1.25) is replaced with

$$I_{c,k} = 0.5,\ k \in \mathcal{K}, \tag{8.1.28}$$

and the objective of (8.1.7) is replaced with

$$\min_X J(X) = \Delta. \tag{8.1.29}$$

The feasibility problem is implemented in GAMS 23.1 and solved with the nonlinear programming solver CONOPT [29, 30] to a final tolerance of 3.0×10^{-13} on a SUSE Linux 10 Virtual Machine with 1 GB of memory and a 2.4 GHz Intel Core Duo CPU. Solution of the MPEC formulation has also been solved with the nonlinear programming solver IPOPT [112]. However, the CPU times obtained are significantly worse than the ones obtained using

CONOPT. Hence, they are omitted.

The feasibility problem is solved for different values of n_t and various initial values of X_0. It is also solved for different values of $I_{c,k}$. Sample results are in Tables 8.2 to 8.3. Status is the termination criteria of the solver. NVAR is the number of elements in X. NEQ is the number of equations in (8.1.7). The column labeled CPU contains the amount of time taken by the computer to solve the problem.

n_t	Status	Δ	NVAR	NEQ	CPU $[s]$
101	Optimal	0.5173	2012	1407	4.73
201	Optimal	0.2599	4012	2807	9.47
301	Optimal	0.3365	6012	4207	15.07
401	Optimal	0.4038	8012	5607	27.28
501	Optimal	0.4304	10012	7007	34.58

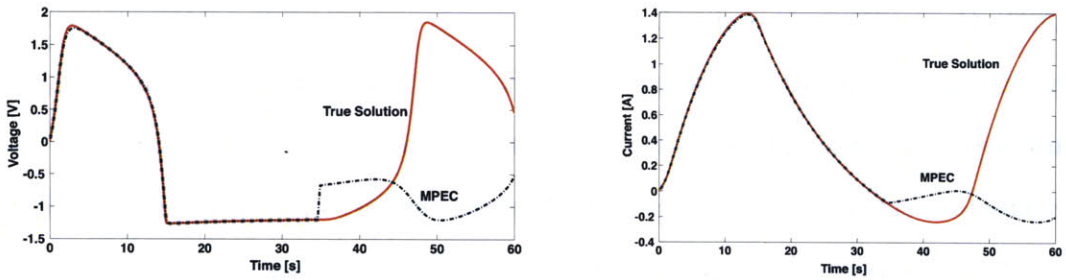
Table 8.2: Electric Circuit: MPEC feasibility problem results, $I_{c,k} = 0.50, \forall k \in \mathcal{K}$.

n_t	Status	Δ	NVAR	NEQ	CPU $[s]$
101	Optimal	1.4362	2012	1407	6.99
201	Optimal	26.0653	4012	2807	12.03
301	Optimal	0.8762	6012	4207	52.07
401	Optimal	0.9754	8012	5607	69.02
501	Optimal	1.1229	10012	7007	135.67

Table 8.3: Electric Circuit: MPEC feasibility problem results, $I_{c,k} = 1.50, \forall k \in \mathcal{K}$.

It is imperative that Δ is zero. Even a small violation can result in grossly erroneous state trajectories as can be seen in Figures 8-5a and 8-5b. This is the reason for the very small termination tolerance.

In this study a zero Δ could not be obtained using arbitrary X_0. Therefore, an X_0 is derived from state trajectories obtained by simulating the dynamics for $I_{c,k} = 1.0$ for all



(a) Voltage (b) Current

Figure 8-5: Electric Circuit: Effect of nonzero complementarity deviations on voltage and current, $\Delta = 0.4304$

$k \in \mathcal{K}$ and $h_{k,i}$ are set to $30/(n_t - 1)$ where n_t is varied. A feasibility problem is solved for each n_t. The results are tabulated in Table 8.4. Even if simulation results are used to initialize the variables, the feasibility problem may have nonzero objective value. The state

n_t	Status	Δ	NVAR	NEQ	CPU $[s]$
201	Optimal	0.0000	4012	2807	4.16
301	Optimal	0.0000	6012	4207	9.28
401	Optimal	0.0956	8012	5607	21.85
501	Optimal	0.6235	10012	7007	43.24
601	Optimal	0.1632	12012	8407	34.70
701	Optimal	0.0000	14012	9807	41.33

Table 8.4: Electric Circuit: MPEC feasibility problem results, $I_{c,k} = 1.00$, initialization with simulation results.

trajectories obtained for the cases $n_t = 201$, $n_t = 301$ and $n_t = 701$ are visually compared to simulation results. It is found that if n_t is set to 701, then the approximations to the state trajectories obtained as the solution of the feasibility problem agree with simulation results sufficiently.

Optimization Results

The penalty parameter μ is set to 5. This value is determined by trial and error. If μ is too small, then Δ may not be zero at the optimal solution.

It is imperative that the initial values supplied to the nonlinear programming solver represent a feasible or an almost feasible point. Otherwise, the solver is not able to provide a feasible solution or a solution X such that Δ is zero.

The optimization problem is implemented in GAMS 23.1 and solved with the nonlinear programming solver CONOPT [29, 30] to a final tolerance of 3.0×10^{-13} on a SUSE Linux 10 Virtual Machine with 1 GB of memory and a 2.4 GHz Intel Core Duo CPU. Solution of the MPEC formulation has also been solved with the nonlinear programming solver IPOPT [112]. However, the CPU times obtained are significantly worse than the ones obtained using CONOPT. Hence, they are omitted.

The results of the optimization runs are in Table 8.5. X_0 is computed from data obtained from the simulation of the dynamics with the given $\mathbf{p}_0$ values. $\bar{J}(\mathbf{p}^*)$ is the value of the objective obtained from the MPEC formulation. Table 8.6 contains the result of the optimization runs with $\mu = 1$. Note that Δ is nonzero for Case D. For this example, the value of μ does not affect the CPU time significantly. The CPU times are significantly more than the CPU times in Table 8.1. The CPU times do not depend strongly on the termination tolerance of the solver. Table 8.7 contains the data for the optimization runs with a termination tolerance of 1×10^{-7}. The CPU times strongly depend on n_t. Table 8.8 shows the change in the solution times for Case E for different values of n_t.

The quality of the approximation of the state trajectories also depend strongly on n_t. Figures 8-6, 8-7 and 8-8 show the effect of n_t on the voltage and current trajectories. The optimal parameters in Table 8.8 are simulated for the cases $n_t = 201$, $n_t = 401$ and $n_t = 701$. The curves marked as "Simulation" are the results obtained by simulation. The curves marked as "MPEC" are the approximations obtained from the solution of (8.1.7). It is

Case	Status	$\mathbf{p}_0$	$\mathbf{p}^*$	$\bar{J}(\mathbf{p}^*)$	Δ	CPU $[s]$
A	Optimal	(0.0000, 0.0000)	(0.0000, 0.0000)	0.1680	0.0000	22.16
B	Optimal	(0.0500, 0.0500)	(0.0657, 0.0000)	0.7902	0.0000	47.43
C	Optimal	(1.4000, 1.4000)	(1.500, 0.0000)	2.5288	0.0000	168.17
D	Optimal	(0.7500, 1.2500)	(0.0742, 0.6140)	1.0643	0.0000	41.95
E	Optimal	(1.0000, 1.0000)	(0.0639, 0.0000)	0.8021	0.0000	176.00
F	Optimal	(0.1000, 1.2500)	(0.0743, 0.6122)	1.0635	0.0000	24.44

Table 8.5: Electric Circuit: MPEC method optimization results, $\mu = 5$, termination tolerance is 3.0×10^{-13}

Case	Status	$\mathbf{p}_0$	$\mathbf{p}^*$	$\bar{J}(\mathbf{p}^*)$	Δ	CPU$[s]$
A	Optimal	(0.0000, 0.0000)	(0.0000, 0.0000)	0.1680	0.0000	15.29
B	Optimal	(0.0500, 0.0500)	(0.0657, 0.0000)	0.7902	0.0000	87.00
C	Optimal	(1.4000, 1.4000)	(1.5000, 0.0000)	2.5288	0.0000	137.00
D*	Optimal	(0.7500, 1.2500)	(0.0733, 0.3144)	1000.0	25.1260	64.46
E	Optimal	(1.0000, 1.0000)	(0.0639, 0.0000)	0.8021	0.0000	150.65
F	Optimal	(0.1000, 1.2500)	(0.0743, 0.6122)	1.0635	0.0000	29.57

Table 8.6: Electric Circuit: MPEC method optimization results, $\mu = 1$, termination tolerance is 3.0×10^{-13}

observed that a low n_t is not sufficient to approximate the state trajectories accurately. The choice of $n_t = 701$ is justified if a high quality of approximation is important.

8.1.5 A Variant Dynamic Optimization Formulation

In this subsection, a variant of program (8.1.6) is solved. The constraint;

$$-0.0001 \leq v(t_f, \mathbf{p}) \leq 0.0001 \tag{8.1.30}$$

Case	Status	$\mathbf{p}_0$	$\mathbf{p}^*$	$\bar{J}(\mathbf{p}^*)$	Δ	CPU[s]
A	Optimal	(0.0000, 0.0000)	(0.0000, 0.0000)	0.1680	0.0000	19.97
B	Optimal	(0.0500, 0.0500)	(0.0657, 0.0000)	0.7902	0.0000	45.67
C	Optimal	(1.4000, 1.4000)	(1.5000, 0.0000)	2.5288	0.0000	158.28
D	Optimal	(0.7500, 1.2500)	(0.0742, 0.6140)	1.0643	0.0000	41.25
E	Optimal	(1.0000, 1.0000)	(0.0639, 0.0000)	0.8021	0.0000	159.20
F	Optimal	(0.1000, 1.2500)	(0.0743, 0.6122)	1.0635	0.0000	23.04

Table 8.7: Electric Circuit: MPEC method optimization results, $\mu = 5$, termination tolerance is 1.0×10^{-7}

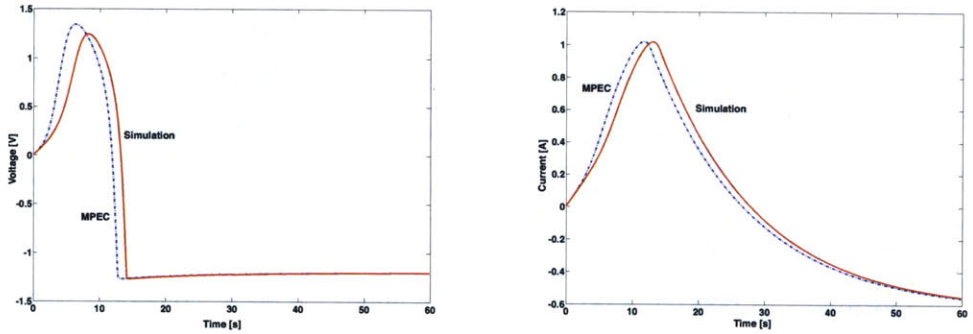
n_t	Status	$\mathbf{p}_0$	$\mathbf{p}^*$	Δ	CPU[s]
201	Optimal	(1.0000, 1.0000)	(0.0623, 0.0695)	0.0000	7.69
301	Optimal	(1.0000, 1.0000)	(0.0648, 0.0706)	0.0000	11.69
401	Optimal	(1.0000, 1.0000)	(0.0654, 0.0718)	0.0000	14.74
501	Optimal	(1.0000, 1.0000)	(0.0633, 0.0000)	0.0000	56.65
601	Optimal	(1.0000, 1.0000)	(0.0636, 0.0000)	0.0000	92.10
701	Optimal	(1.0000, 1.0000)	(0.0639, 0.6122)	0.0000	159.20

Table 8.8: Electric Circuit: MPEC method optimization results for various values of n_t for Case E, $\mu = 5$, termination tolerance is 1.0×10^{-7}

is added to problem (8.1.6) and the corresponding constraint

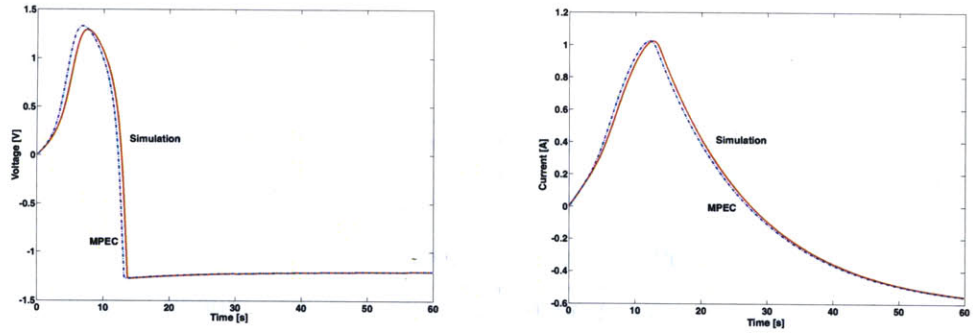
$$-0.0001 \leq v_{n_e,n_t} \leq 0.0001$$

is added to (8.1.7). The resultant programs are solved using the nonsmooth single shooting method and the MPEC approach.



(a) Voltage (b) Current

Figure 8-6: Electric Circuit: Difference between MPEC predicted voltage and current trajectories and simulation for Case E and $n_t = 201$.



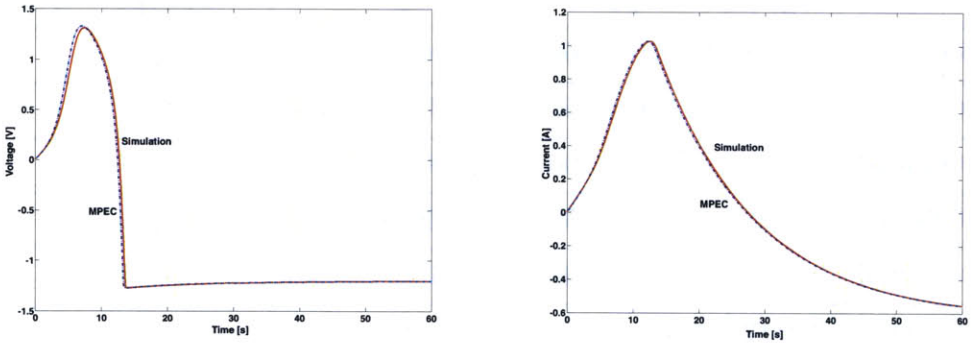
(a) Voltage (b) Current

Figure 8-7: Electric Circuit: Difference between MPEC predicted voltage and current trajectories and simulation for Case E and $n_t = 401$.

Nonsmooth Single Shooting Method

The dynamic optimization problem is solved using an exact penalty formulation since the bundle solver used does not support nonlinear constraints directly. Further details of the exact penalty approach in nonsmooth optimization can be found in [87]. The penalized objective of (8.1.6) is

$$J(\mathbf{p}) = S(t_f, \mathbf{p}) + \mu \max(0, v(t_f, \mathbf{p}) - 0.0001) + \mu \max(-0.0001 - v(t_f, \mathbf{p}), 0)$$



(a) Voltage (b) Current

Figure 8-8: Electric Circuit: Difference between MPEC predicted voltage and current trajectories and simulation for Case E and $n_t = 701$.

where μ is the penalty parameter. A sequence of dynamic optimization programs is solved where the penalty parameter is increased. Let l be the index of the optimization problem solved. The penalty parameter for problem l is $\mu(l) = 40 \cdot l$. Problem l is solved to optimality. If the solution is not a feasible point for (8.1.6) with the additional constraint (8.1.30), then l is set to $l + 1$, and the next program is solved. Otherwise, the solution is stationary in the extended sense for the problem (8.1.6) with the additional constraint (8.1.30). This can be derived as in [87] using the linear Newton approximations of the functions instead of the generalized gradients and the extended Cottle constraint qualification introduced in Chapter 6. The existence of a finite penalty parameter requires further research.

For the integration of the dynamics and auxiliary equations DSL48SE ([108], [109] [36]) is used with code generated by DAEPACK [107]. The nonlinear program is solved by the proximal bundle solver in [64] on a SUSE Linux 10 Virtual Machine with 1 GB of RAM and a 2.4 GHz Intel Core Duo CPU. The absolute and relative tolerances used in simulation are 1×10^{-8} and 1×10^{-8}, respectively. The optimality tolerance for the bundle solver is set to 1×10^{-6}. The results are summarized in Table 8.9.

Case	Status	$J(\mathbf{p}^*)$	$\mathbf{p}_0$	$\mathbf{p}^*$	$v(t_f, \mathbf{p})$	CPU $[s]$	NIT	NFV
A	Optimal	0.1738	(0.000, 0.000)	(0.0000, 1.2782)	0.0000	8.48	77	138
B	Optimal	0.7122	(0.050, 0.050)	(0.07179, 1.3036)	0.0000	13.33	120	152
C	Optimal	0.1738	(1.400, 1.400)	(0.0000, 1.2782)	0.0000	3.86	56	60
D	Optimal	0.7128	(0.750, 1.250)	(0.07369, 0.4135)	1.5230×10^{-5}	49.64	343	540
E	Optimal	0.7122	(1.000, 1.000)	(0.07150, 1.3036)	0.0000	4.13	38	48
F	Optimal	0.7112	(0.100, 1.250)	(0.07820, 0.4121)	4.372×10^{-5}	21.55	203	236

Table 8.9: Electric Circuit: Nonsmooth single shooting results for variant problem.

MPEC Approach

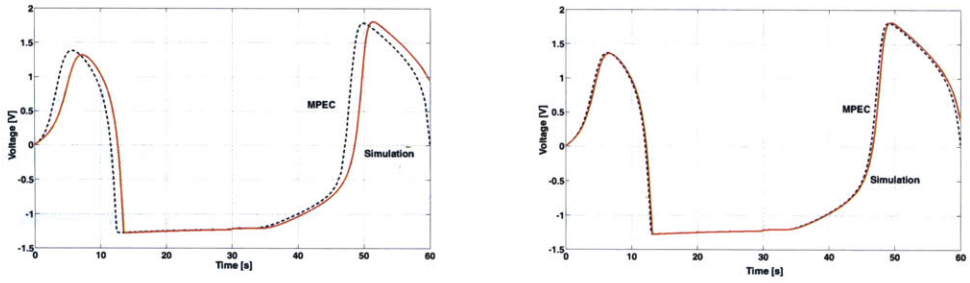
For the MPEC approach, the modified program is solved for various values of n_t and the MPEC predicted trajectories are compared to the trajectories obtained from the simulation of the optimal parameters furnished by the MPEC approach. The initial parameter guess is (1.00, 1.00) for all optimization runs. The variables are initialized using simulation data. The optimization problem is implemented in GAMS 23.1 and solved with the nonlinear programming solver CONOPT [29, 30] to a final tolerance of 1.0×10^{-7} on a SUSE Linux 10 Virtual Machine with 1 GB of memory and a 2.4 GHz Intel Core Duo CPU. Solution of the MPEC formulation has also been solved with the nonlinear programming solver IPOPT [112]. However, the CPU times obtained are significantly worse than the ones obtained using CONOPT. Hence, they are omitted.

The results are in Table 8.10. The column labeled "$v(t_f, \mathbf{p})$" contains the final voltage obtained by simulation of the dynamics using the optimal parameter values in the column labeled "$\mathbf{p}^*$". Note that there is significant difference between v_{n_e,n_t} and $v(t, \mathbf{p})$. The difference between the final voltage predicted by the MPEC approach and the simulation decreases as the n_t increases. Even though the complementary condition deviations are zero, there is gross error in estimating the voltage trajectory as can be seen in Figure 8-9. Unless a com-

n_t	Status	$\mathbf{p}^*$	$\bar{J}(\mathbf{p}^*)$	Δ	v_{n_e,n_t}	$v(t_f,\mathbf{p}^*)$	CPU $[s]$
201	Optimal	$(0.06477, 0.3878)$	0.7103	0.00	0.0000	0.9460	7.19
401	Optimal	$(0.06824, 0.4001)$	0.7108	0.00	0.0000	0.6400	39.89
701	Optimal	$(0.06955, 0.4064)$	0.7118	0.00	0.0000	0.4240	62.01
801	Optimal	$(0.06979, 0.4075)$	0.7116	0.00	0.0000	0.3800	87.38
1001	Optimal	$(0.07005, 0.4088)$	0.7121	0.00	0.0000	0.3177	195.70

Table 8.10: Electric Circuit: MPEC method optimization results for the variant dynamic optimization problem for various values of n_t for Case E, $\mu = 5$, termination tolerance is 1.0×10^{-7}.

parison with a simulation is carried out, the error in the MPEC predicted state trajectories cannot be detected.



(a) $n_t = 201$ (b) $n_t = 801$

Figure 8-9: Electric Circuit: Difference between MPEC predicted voltage trajectories and simulation for Case E and $n_t = 201$ and $n_t = 801$.

Finally, an MPEC formulation was developed using third order Radau collocation. CONOPT was not able to solve this formulation.

8.1.6 Conclusion

The Electric Circuit Case Study involves a relatively simple dynamic optimization problem. The effort to solve this problem using the MPEC approach was significantly more than the nonsmooth single shooting method.

The nonsmooth single shooting method required coding the model in FORTRAN77 and using the automatic code generation tool DAEPACK to generate the additional equations to obtain derivative information. Work had to be done to integrate the bundle solver and the routines used to simulate the dynamics. Running the optimization solver required no additional significant effort. It was observed that more than 95% of the solution times of the nonsmooth single shooting method were used to solve the initial value problem and compute an element of the linear Newton approximations.

The MPEC approach required the manual discretization of the dynamics and implementation in the GAMS environment. Determination of n_t required substantial time. The determination of a value for n_t guaranteeing a good approximation to the state trajectories does not appear to be possible without comparison to the simulation output for problems with highly nonlinear dynamics. In this regard, the MPEC approach does not seem to be a standalone method. Finding a good initial starting point for the solver was not possible by solving a feasibility problem. Instead, initial starting points were derived from simulation data. It was observed that if a good starting point is not provided to the solver, a feasible solution or a solution with zero Δ could not be obtained. A significant amount of time was spent to provide good starting points derived from simulation data.

The nonlinear and relatively stiff dynamics mandated the use of a relatively large n_t to approximate the state trajectories acceptably. This led to longer solution times than the solution times of the nonsmooth shooting method. It was observed that if Δ could not be driven to zero, gross errors in the approximation of the state trajectories could occur even for small values of Δ. In this regard, the MPEC approach is fragile. It is known that collocation methods produce state trajectories that are not realistic if constraints in the mathematical program representing dynamics are not satisfied. In the case of varying dynamics, the error incurred may be even more because the incorrect set of equations governing the dynamics between two time points may be selected. Even if a good starting point is provided, the

MPEC method may fail to produce realistic state trajectories and fail as can be seen in Table 8.6.

The MPEC approach may provide misleading results as can be seen in the solution of the variant dynamic optimization problem. Even though the voltage trajectories predicted by the MPEC approach looked reasonable and correct, comparison to simulation revealed gross errors. The nonsmooth shooting method should be used in dynamic optimization problems involving highly nonlinear and stiff dynamics and constraints on the transient behavior. In these problems, the accuracy obtained in the computation of the state trajectories using an initial value problem solver is crucial in obtaining a correct result.

8.2 Cascading Tanks: Empirical Complexity Analysis

The empirical scaling of the CPU times required by the nonsmooth single shooting algorithm introduced in Chapter 7 is investigated using a literature example described in [106].

8.2.1 System Dynamics

The system considered in this case study is originally presented in [106] for the empirical complexity analysis of the MILP approach to solving dynamic optimization problems involving systems with varying structure. A set of prismatic tanks with constant cross-sections connected in series using check valves to prevent flow in the reverse direction (Figure 8-10) constitute the model. The detailed listing of the elements of the system and the dynamic optimization formulation is as follows:

- t_0: initial time in seconds, $t_0 = 0.0$ s.
- t_f: final time in seconds, $t_f = 100.0$ s.
- T: time interval in seconds; $T = [t_0, t_f]$
- ΔT: the duration in seconds; $\Delta T = 100.0$ s.

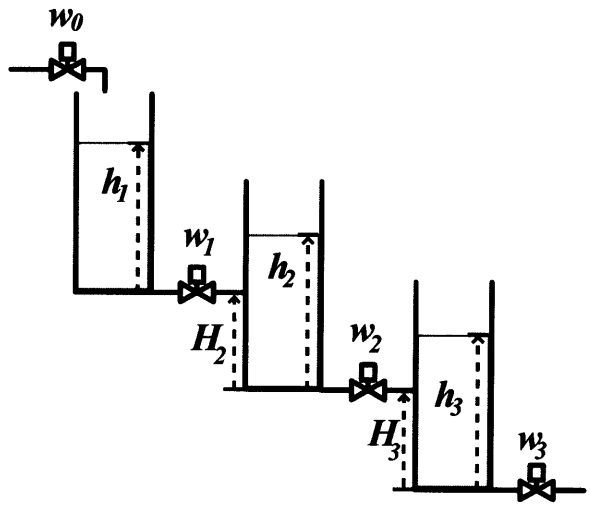


Figure 8-10: Cascaded Tanks Configuration.

- n_T: the number of tanks connected in series; $3 \leq n_T \leq 16$.
- $\mathcal{I}$: the set of tank indices, $\mathcal{I} = \{1, \ldots, n_T\}$. Tank $i+1$ is downstream of tank i and flow from tank $i+1$ to tank i is prevented by the connecting check valve.
- $\{A_i\}_{i=1}^{n_T}$: the cross sectional area of the tanks in m^2. $A_i = 3/n_T$ m^2 for all i in $\mathcal{I}$.
- $\{H_i\}_{i=2}^{n_T}$: the height of the feed pipe connecting tank $i-1$ to tank i. $H_i = 0.5\mathrm{m}$ for all $i \in \mathcal{I} \backslash \{1\}$.
- n_e: number of epochs used in the control vector parameterization; $6 \leq n_e \leq 100$.
- $\mathcal{K}$: the set of epoch indices; $\mathcal{K} = \{1, \ldots, n_e\}$.
- $\{T_k\}_{k=1}^{n_e}$: the time intervals for each epoch. $T_k = [\alpha_k, \beta_k)$ where $\alpha_k = t_0 + \frac{\Delta T(k-1)}{n_e}$ and $\beta_k = t_0 + \frac{\Delta T(k)}{n_e}$ for all k in $\mathcal{K}$.
- $\{\{w_{k,i}\}_{k=1}^{n_e}\}_{i=0}^{n_T}$: the valve openings. $0.25 \leq w_{k,i} \leq 1.25$ for all $k \in \mathcal{K}$ and for all $i \in \mathcal{I} \cup \{0\}$.
- n_p: number of parameters; $n_p = n_e \cdot (n_T + 1)$.
- $\mathbf{p}$: the parameter vector. $\mathbf{p} \in \mathbb{R}^{n_p}$ and $w_{k,i} = \mathbf{p}_{i \cdot n_e + k}$, $k \in \mathcal{K}$ and $i \in \mathcal{I} \cup \{0\}$.
- $\mathcal{P}$: the parameter set. $\mathcal{P}$ is an open subset of $\mathbb{R}^{n_p}$ such that $[0.25, 1.25]^{n_p} \subset \mathcal{P}$.
- $w_i : T \times \mathcal{P} \rightarrow \mathbb{R}$: the controls of the system for all $i \in \mathcal{I} \cup \{0\}$. $w_i(t, \mathbf{p}) = w_{i,k}$, $\forall t \in$

T_k, $\forall i \in \mathcal{I} \cup \{0\}$ and $w_i(t_f, \mathbf{p}) = w_{n_e,i}$, $\forall i \in \mathcal{I} \cup \{0\}$.

- $h_i : T \times \mathcal{P} \to \mathbb{R}$: the liquid height in the tank with index i for all $i \in \mathcal{I}$ in m.
- $\mathbf{h} : T \times \mathcal{P} \to \mathbb{R}^{n_T}$: the vector of liquid heights which is the continuous state of the system, $\mathbf{h}(t, \mathbf{p}) = (h_i(t, \mathbf{p}), \ldots, h_{n_T}(t, \mathbf{p}))$.
- $\mathcal{X}$: the state space. $\mathcal{X} = \mathbb{R}^{n_T}$.
- C_0: valve constant of valve 0 in m^3/s; $C_0 = 0.1\ \mathrm{m}^3/\mathrm{s}$.
- $\{C_i\}_{i=1}^{n_T}$: valve constant of valve i in $\mathrm{m}^{2.5}/\mathrm{s}$; $C_i = 0.1\ \mathrm{m}^{2.5}/\mathrm{s}$.
- $F_i : T \times \mathcal{P} \times \mathcal{X} \to \mathbb{R}$, the outlet flow rate from tank i in m^3/s for all $i \in \mathcal{I}$.
- $F_0 : T \times \mathcal{P} \to \mathbb{R}$, the inlet flow rate to tank 1 in m^3/s.
- k_r: a positive regularization constant, $k_r = 0.0001$ m.
- h_L: the lower bound on acceptable liquid heights in m, $h_L = 0.7$ m.
- h_U: the upper bound on acceptable liquid heights in m, $h_U = 0.8$ m.

The numerical values except for k_r are from [106].

The equations that govern the evolution of the cascaded tanks system are:

$$\dot{h}_i(t, \mathbf{p}) = \frac{1}{A_i}(F_{i-1}(t, \mathbf{p}) - F_i(t, \mathbf{p})), \forall t \in (t_0, t_f], \tag{8.2.1}$$

$$F_0(t, \mathbf{p}) = C_0 w_0(t, \mathbf{p}),\ \forall t \in [t_0, t_f], \tag{8.2.2}$$

$$F_i(t, \mathbf{p}) = C_i w_i(t, \mathbf{p}) \frac{\Delta h_i(t, \mathbf{p})}{\sqrt{|\Delta h_i(t, \mathbf{p})| + k_r}}, \forall t \in [t_0, t_f], \tag{8.2.3}$$

$$\Delta h_i(t, \mathbf{p}) = \begin{cases} \bar{\Delta} h_i(t, \mathbf{p}) & \text{if } \bar{\Delta} h_i(t, \mathbf{p}) \geq 0, \\ 0 & \text{if } \bar{\Delta} h_i(t, \mathbf{p}) < 0, \end{cases}, \ \forall i \in \mathcal{I} \backslash \{n_T\}, \tag{8.2.4}$$

$$\bar{\Delta} h_i(t, \mathbf{p}) = \begin{cases} h_i(t, \mathbf{p}) - (h_{i+1}(t, \mathbf{p}) - H_{i+1}) & \text{if } h_{i+1}(t, \mathbf{p}) > H_{i+1}, \\ h_i(t, \mathbf{p}) & \text{if } h_{i+1}(t, \mathbf{p}) \leq H_{i+1}, \end{cases}, \ \forall i \in \mathcal{I} \backslash \{n_T\}, \tag{8.2.5}$$

$$\Delta h_{n_e}(t, \mathbf{p}) = h_{n_e}(t, \mathbf{p}), \tag{8.2.6}$$

$$h_i(t_0, \mathbf{p}) = 0.1, \forall i \in \mathcal{I}. \tag{8.2.7}$$

Equation (8.2.1) represents conservation of mass assuming constant liquid density. Equation (8.2.2) governs the flow through the inlet valve. The difference between the liquid heights in consecutive tanks determines the flow between consecutive tanks. If the liquid height in the downstream tank is less than the height of the feed pipe, then the liquid height in the downstream tank does not affect the flow rate. This phenomena is captured in (8.2.5) and (8.2.6). It is possible that for some initial conditions that the downstream liquid height is large enough to force a flow in the reverse direction. In this case $\Delta h_i(t, \mathbf{p})$ is negative. Since the valves connecting the tanks are check valves, there can be no reverse flow. This situation is captured in (8.2.4). The flow relation used in (8.2.3) is an approximation of the square root function. If $\Delta h_i(t, \mathbf{p})$ is much larger than k_r then the flow is approximately proportional to $\sqrt{\Delta h_i(t, \mathbf{p})}$. When $\Delta h_i(t, \mathbf{p})$ is very small, the flow relation becomes approximately $\Delta h_i(t, \mathbf{p})/\sqrt{k_r}$. As a result, the function $y \mapsto y/(\sqrt{|y| + k_r})$ is continuously differentiable in the neighborhood of 0. The equations are not locally Lipschitz continuous if this regularization is not made. The model equations differ in this point from those presented in [106]. Finally, (8.2.7) determines the initial conditions of the state.

8.2.2 Dynamic Optimization Formulation

The aim of the dynamic optimization problem is to bring the liquid heights to values between h_L and h_U and minimize the deviation of the liquid heights from this range. Given $h_i(t, \mathbf{p})$, the deviation is $\max(0, h_L - h_i(t, \mathbf{p}), h_i(t, \mathbf{p}) - h_U)$. The total deviation of liquid height i is the integral of the deviation at t over the interval $[t_0, t_f]$. Therefore, the dynamic optimization problem is:

$$\min_{\mathbf{p} \in \mathcal{P}} J(\mathbf{p}) = \sum_{i=1}^{n_T} \int_{t_0}^{t_f} \max(0, h_L - h_i(t, \mathbf{p}), h_i(t, \mathbf{p}) - h_U) dt \tag{8.2.8}$$

$$\text{s.t.} 0.25 \leq w_{k,i} \leq 1.25, \ \forall k \in \mathcal{K}, \forall i \in \mathcal{I} cup\{0\}.$$

where $h_i(t, \mathbf{p})$ are computed using (8.2.1)-(8.2.7). Note that the objective function can be computed by adding auxiliary states $z_i : T \times \mathcal{P} \rightarrow \mathbb{R}$ to (8.2.1)-(8.2.7) whose evolutions are governed by

$$\dot{z}_i(t, \mathbf{p}) = \max(0, h_L - h_i(t, \mathbf{p}), h_i(t, \mathbf{p}) - h_U), \ \forall t \in (t_0, t_f], \ \forall i \in \mathcal{I}, \tag{8.2.9}$$

$$z_i(t_0, \mathbf{p}) = 0, \forall i \in \mathcal{I}. \tag{8.2.10}$$

As a result $\mathcal{X}$ becomes $\mathbb{R}^{2n_T}$.

The final form of the optimization problem is

$$\min_{\mathbf{p} \in \mathcal{P}} J(\mathbf{p}) = \sum_{i=1}^{n_T} z_i(t_f, \mathbf{p}) \tag{8.2.11}$$

$$\text{s.t} : 0.25 \leq w_{i,k} \leq 1.25, \ \forall k \in \mathcal{K}, \forall i \in \mathcal{I}, .$$

where z_i are computed using (8.2.1)-(8.2.7) and (8.2.9)-(8.2.10).

8.2.3 Scaling of the Nonsmooth Single Shooting Method with Respect to Number of Tanks and Number of Epochs

In this section empirical complexity analysis results are presented. Theoretical complexity analysis of bundle methods does not currently exist. Therefore, the complexity of the nonsmooth shooting method is analyzed empirically.

Problem (8.2.8) is solved for different values of n_T and n_e values to determine empirically how the solution times of the nonsmooth shooting method scales. The study is similar to the empirical complexity analysis in [106] carried out for the MILP approach that can be used to solve dynamic problem (8.2.8).

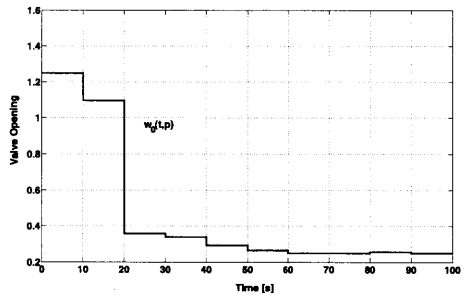
For the integration of the dynamics and the auxiliary equations, DSL48SE ([108, 109, 36])

is used with code generated by DAEPACK ([107]). The nonlinear program is solved by the proximal bundle solver in [64] on a SUSE Linux 10 Virtual Machine with 1 GB of RAM and a 2.4 GHz Intel Core Duo CPU using two sets of tolerances summarized in Table 8.11. The valve openings are initialized at the lower bound value of 0.25. All optimization runs

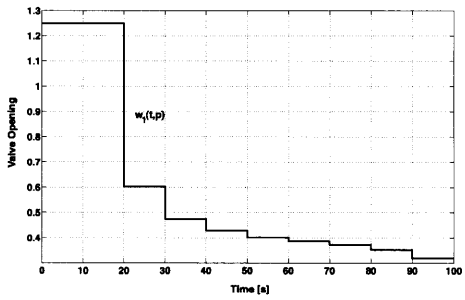
Label	Absolute Tolerance	Relative Tolerance	Optimality Tolerance
R	1.0×10^{-6}	1.0×10^{-6}	1.0×10^{-4}
T	1.0×10^{-7}	1.0×10^{-7}	1.0×10^{-5}

Table 8.11: Cascading Tanks: Simulation and optimization tolerances.

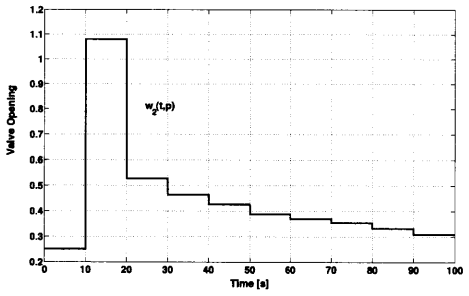
terminate satisfying the optimality tolerance. The solution obtained for the case $n_e = 10$ and $n_T = 3$ and the corresponding state trajectories are shown in Figures 8-11 and 8-12, respectively. The raw data obtained from the multiple optimization runs are documented



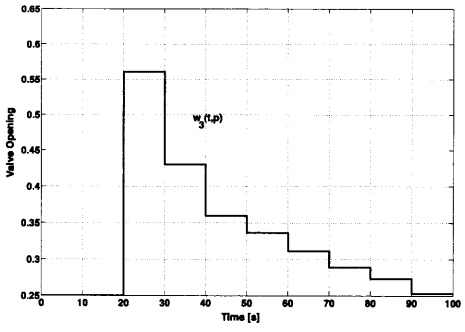
(a) $w_0(t, \mathbf{p})$



(b) $w_1(t, \mathbf{p})$



(c) $w_2(t, \mathbf{p})$



(d) $w_3(t, \mathbf{p})$

Figure 8-11: Cascading Tanks: Optimal valve openings for $n_T = 3.0$ and $n_e = 10$.

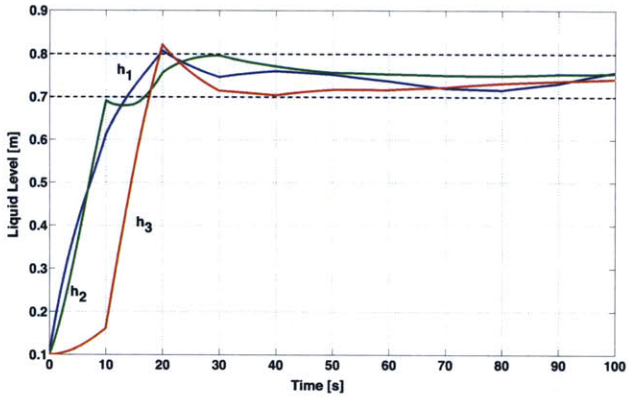


Figure 8-12: Cascading Tanks: Plot of optimal state trajectories for $n_T = 3.0$ and $n_e = 10$.

in Tables 8.12 and 8.13. The column with the label "total" contains the total number of equations integrated. This number is the sum of the number of original equations of the system and the associated sensitivity equations. The column with label "ndf" contains the number of discontinuity functions. Note that for each state z_i, two discontinuity functions are required. The first tank requires no discontinuity function and the last one requires only one. All other tanks require two discontinuity functions to compute the associated sensitivity equations. The total number of equations equals $(2n_T + \text{ndf}) \cdot (n_p + 1)$ where ndf stands for the number of discontinuity functions. This case study has a special structure. The total number of equations is a quadratic function of the number of tanks and a linear function of the parameters because $n_p = n_e \cdot (n_T + 1)$. The column with the label "$J(\mathbf{p}^*)$" contains the optimal solution values. The seconds taken to solve the dynamic optimization problem is in the "CPU" column. The "NIT" column contains the number of iterations carried out by the bundle solver. Finally, the "NFV" column contains the number of times the integrator is called to solve the problem. R and T denote the two sets of tolerances used. Note that the integration method used exploits the block lower triangular structure of the state and sensitivity equations ([36]) to efficiently solve these equations.

The optimal solution values do not change appreciably with the tolerances. On the other

n_T	n_p	NEQ		$J(\mathbf{p}^*)$		CPU $[s]$		NIT		NFV	
		total	ndf	R	T	R	T	R	T	R	T
3	40	656	10	14.41	14.42	11.54	13.99	51	47	52	48
4	50	1122	14	19.30	19.29	24.32	39.65	56	73	57	74
5	60	1708	18	24.48	24.45	31.06	53.34	46	63	47	64
6	70	2414	22	29.67	29.67	64.66	99.27	65	78	66	79
7	80	3240	26	35.08	35.07	89.38	145.64	61	80	66	81
8	90	4186	30	40.68	40.67	121.63	177.97	68	75	69	76
9	100	5252	34	46.31	46.33	205.98	214.95	76	64	77	65
10	110	6438	38	52.13	52.13	278.79	430.00	79	104	80	106
11	120	7744	42	58.01	58.00	339.81	550.53	75	105	76	106
12	130	9170	46	63.90	63.88	507.94	718.60	95	111	96	112
13	140	10716	50	69.87	69.88	636.29	790.66	92	92	93	93
14	150	12382	54	75.98	75.93	789.95	1538.94	94	156	95	157
15	160	14168	58	82.10	82.10	1091.52	1488.33	125	125	126	126
16	170	16074	62	88.35	88.31	1288.12	2422.15	105	174	106	175

Table 8.12: Cascading Tanks: Optimization run data for $n_e = 10$ and different numbers of tanks.

hand, the CPU times differ significantly with tolerances. The first reason is the increased amount of time to simulate the dynamics using tighter simulation tolerances. The second reason is the additional bundle solver iterations required to satisfy tighter optimality tolerances.

The functional dependence of CPU times on the number of epochs and number of tanks is estimated by fitting functions to the data in Tables 8.12 and 8.13. It is determined that the dependence of CPU times on n_T and n_e is not exponential by investigating the mappings $n_T \mapsto \ln(\tau(n_T, n_e))$ for a fixed value of n_e and $n_e \mapsto \ln(\tau(n_T, n_e))$ for a fixed value of n_T where $\tau(n_T, n_e)$ represents the CPU times. It is found that the growth of these mappings is slower than linear, implying that the CPU time growth is slower than exponential growth.

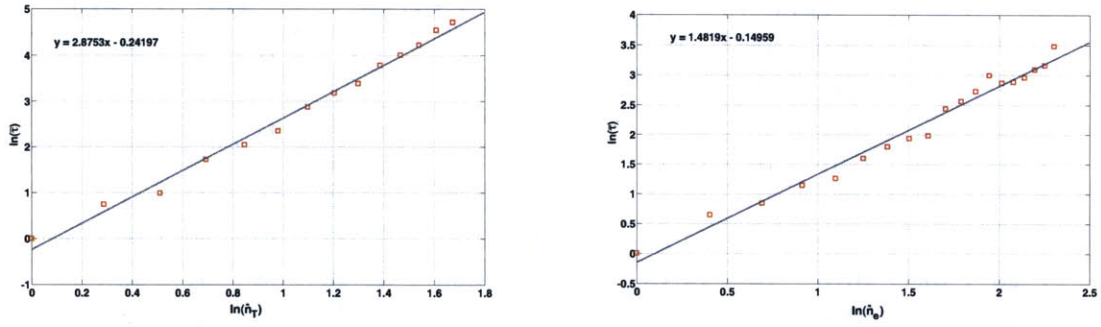
It is assumed that the CPU time grows polynomially with n_e and n_T. The mappings $\ln(\bar{n}_T) \mapsto \ln(\bar{\tau}(n_T, n_e))$ for fixed n_e and $\ln(\bar{n}_e) \mapsto \ln(\bar{\tau}(n_T, n_e))$ for fixed n_T are investigated where $\bar{n}_T = n_T/3$, $\bar{n}_e = n_e/10$ and $\bar{\tau}(n_T, n_e) = \tau(n_T, n_e)/\tau(3, 10)$ in order to determine the degree of the polynomial.

n_e	n_p	NEQ		$J(\mathbf{p}^*)$		CPU $[s]$		NIT		NFV	
		total	ndf	R	T	R	T	R	T	R	T
5	20	336	10	14.75	14.77	6.90	9.71	56	60	57	61
8	32	528	10	14.97	14.88	7.36	25.19	41	109	42	110
10	40	656	10	14.41	14.42	11.54	13.96	51	47	52	48
15	60	976	10	14.25	14.22	22.29	40.23	59	83	60	84
20	80	1296	10	14.22	14.22	27.13	54.83	49	80	50	81
25	100	1616	10	14.19	14.19	36.56	47.49	48	50	49	51
30	120	1936	10	14.20	14.19	41.14	70.46	41	57	42	58
35	140	2256	10	14.19	14.19	57.63	65.51	46	43	47	44
40	160	2576	10	14.17	14.19	70.25	106.52	44	56	45	57
45	180	2896	10	14.18	14.18	80.84	121.34	43	54	44	55
50	200	3216	10	14.22	14.17	84.80	174.51	38	65	39	66
55	220	3536	10	14.17	14.17	133.35	175.39	53	56	54	57
60	240	3856	10	14.20	14.17	150.52	248.68	51	70	52	71
65	260	4176	10	14.17	14.19	178.26	203.92	54	50	55	51
70	280	4496	10	14.17	14.18	232.47	246.03	58	49	59	50
75	300	4816	10	14.17	14.17	204.93	295.53	47	53	48	55
80	320	5136	10	14.17	14.17	207.73	301.07	42	50	43	51
85	340	5456	10	14.18	14.18	224.83	302.09	42	44	43	45
90	360	5776	10	14.17	14.17	256.50	453.00	44	61	45	62
95	380	6096	10	14.25	14.17	275.62	530.23	43	65	44	66
100	400	6416	10	14.17	14.17	379.66	617.44	55	71	56	72

Table 8.13: Cascading Tanks: Optimization run data for $n_T = 3$ and different numbers of epochs

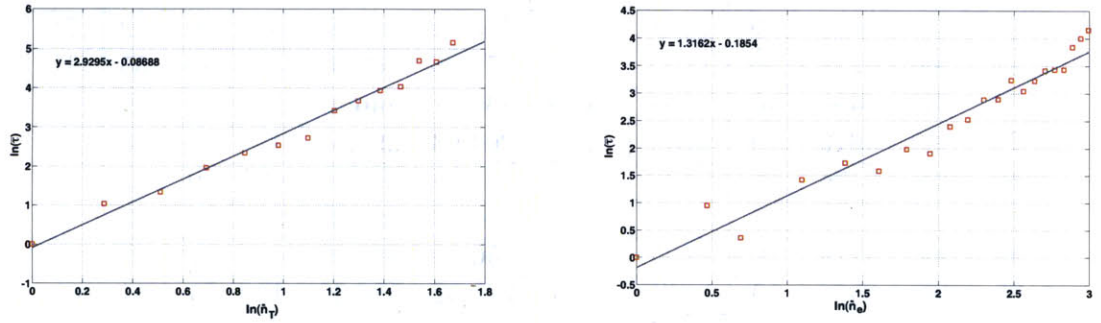
The curve fitting results are shown in Figures 8-13 and 8-14. Detailed information can be found in Tables 8.15 and 8.14. The column with the label "SSE" contains the sum of the squared errors, the columns with labels "R^2" and "$\bar{R}^2$" contain the R-squared and adjusted R-squared values, respectively. Finally the "RMSE" contains the root mean squared error.

The results suggest that the CPU time to obtain a solution is a cubic function of the number of tanks and at most a quadratic function of the number of epochs. The polynomials fitted to the data are shown in Figures 8-15 and 8-16. Additional information can be found in Tables 8.16 and 8.17.



(a) $\bar{n}_T$ (b) $\bar{n}_e$

Figure 8-13: Cascading Tanks: Plot of $\bar{\tau}$ versus $\bar{n}_T$ and $\bar{\tau}$ versus $\bar{n}_e$ for the relaxed tolerance case.

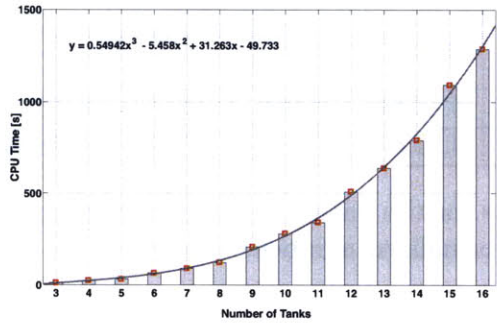


(a) $\bar{n}_T$ (b) $\bar{n}_e$

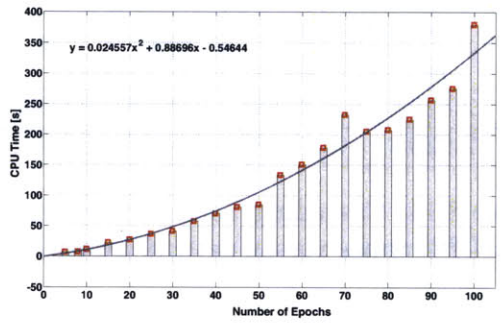
Figure 8-14: Cascading Tanks: Plot of $\bar{\tau}$ versus $\bar{n}_T$ and $\bar{\tau}$ versus $\bar{n}_e$ tight tolerance case.

8.2.4 Conclusion

The complexity of the nonsmooth single shooting method is investigated using a literature problem introduced in [106]. It is observed that the CPU times strongly depend on simulation and optimality tolerances used. In addition, for this literature example, it is determined that the time required to solve the dynamic optimization problem grows polynomially with the number of tanks hence states and the number of epochs. The dependence on the number of tanks is cubic and the dependence on the number of epochs is quadratic. The number of equations integrated depends quadratically on the number of tanks hence the number of states and the number of equations integrated depends linearly on the number of parameters

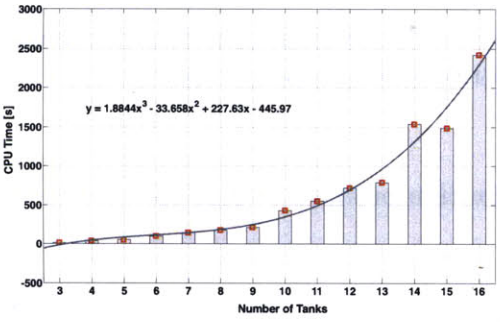


(a) n_T

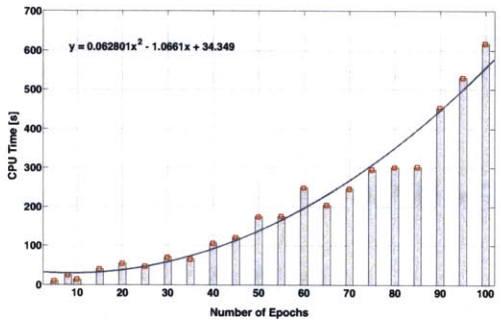


(b) n_e

Figure 8-15: Cascading Tanks: Plot of CPU time versus n_T and n_e for the relaxed tolerance case.



(a) n_T



(b) n_e

Figure 8-16: Cascading Tanks: Plot of CPU time versus n_T and n_e for the tight tolerance case.

for this example.

8.3 Cascading Tank: Comparison with the MPEC Approach

The dynamic optimization problem (8.2.8) is solved using the MPEC approach [12] and the solution times are compared to the nonsmooth shooting method solution times.

$y = p_1 x + p_2$								
	Coefficients		SSE	R^2	$\bar{R}^2$	RMSE	99% Confidence Intervals	
	p_1	p_2					p_1	p_2
R	2.8753	−0.24197	0.2796	0.9903	0.9895	0.1526	[2.6248, 3.1257]	[−0.5314, 0.0474]
T	2.9295	−0.08688	0.5407	0.9821	0.9806	0.2123	[2.5811, 3.2779]	[−0.4893, 0.3156]

Table 8.14: Cascading Tanks: Curve fitting results for the natural logarithm of the normalized CPU times versus natural logarithm of the normalized number of states.

$y = p_1 x + p_2$								
	Coefficients		SSE	R^2	$\bar{R}^2$	RMSE	99% Confidence Intervals	
	p_1	p_2					p_1	p_2
R	1.4819	−0.14959	0.3727	0.9788	0.9775	0.1481	[1.3399, 1.6465]	[−0.4450, 0.0652]
T	1.3162	−0.01854	1.2750	0.9536	0.9512	0.2591	[1.1257, 1.5068]	[−0.6061, 0.2353]

Table 8.15: Cascading Tanks: Curve fitting results for the natural logarithm of the normalized CPU times versus natural logarithm of the normalized number of epochs.

The MPEC formulation of the dynamic optimization problem (8.2.11) is

$$\min_X J(X) = \sum_{k=1}^{n_e} \sum_{j=2}^{n_t} \sum_{i=1}^{n_T} dt(u^L_{k,j,i} + u^H_{k,j,i}) + \mu\Delta \tag{8.3.1}$$

$$\text{s.t: } h_{k,j,i} = h_L - u^L_{k,j,i},\ \forall k \in \mathcal{K},\ \forall j \in \mathcal{J}, \forall i \in \mathcal{I}, \tag{8.3.2}$$

$$h_{k,j,i} = h_U + u^H_{k,j,i},\ \forall k \in \mathcal{K},\ \forall j \in \mathcal{J}, \forall i \in \mathcal{I}, \tag{8.3.3}$$

$$u^H_{k,j,i} \geq 0,\ u^L_{k,j,i},\ \geq 0,\ \forall k \in \mathcal{K},\ \forall j \in \mathcal{J}, \forall i \in \mathcal{I}, \tag{8.3.4}$$

$$h_{k,j+1,i+1} - H_{i+1} = \Delta\bar{h}^+_{k,j+1,i+1} - \Delta\bar{h}^-_{k,j+1,i+1},\ \forall k \in \mathcal{K},\ \forall j \in \bar{\mathcal{J}}, \forall i \in \bar{\mathcal{I}}, \tag{8.3.5}$$

$$p_{1,k,j+1,i+1} = (1 - \lambda_{k,j+1,i+1})\Delta\bar{h}^+_{k,j+1,i+1},\ \forall k \in \mathcal{K},\ \forall j \in \bar{\mathcal{J}}, \forall i \in \bar{\mathcal{I}}, \tag{8.3.6}$$

$$p_{2,k,j+1,i+1} = \lambda_{k,j+1,i+1}\Delta\bar{h}^-_{k,j+1,i+1},\ \forall k \in \mathcal{K},\ \forall j \in \bar{\mathcal{J}}, \forall i \in \bar{\mathcal{I}}, \tag{8.3.7}$$

$$\Delta h_{k,j+1,i} = h_{k,j+1,i} - \lambda_{k,j+1,i+1}\Delta\bar{h}^+_{k,j+1,i+1},\ \forall k \in \mathcal{K},\ \forall j \in \bar{\mathcal{J}}, \forall i \in \bar{\mathcal{I}}, \tag{8.3.8}$$

$$\Delta\bar{h}^+_{k,j,i} \geq 0,\ \Delta\bar{h}^-_{k,j,i} \geq 0, \forall k \in \mathcal{K},\ \forall j \in \mathcal{J},\ \forall i \in \mathcal{I}, \tag{8.3.9}$$

$$0.0 \leq \lambda_{k,j,i} \leq 1.0, \forall k \in \mathcal{K},\ \forall j \in \mathcal{J},\ \forall i \in \mathcal{I}, \tag{8.3.10}$$

$$\Delta h_{k,j+1,i} = \Delta h^+_{k,j+1,i} - \Delta h^-_{k,j+1,i},\ \forall k \in \mathcal{K},\ \forall j \in \bar{\mathcal{J}}, \forall i \in \bar{\mathcal{I}}, \tag{8.3.11}$$

$y = p_1x^2 + p_2x + p_3$							
	Coefficients			SSE	R^2	$\bar{R}^2$	RMSE
	p_1	p_2	p_3				
R	0.024557	0.88696	−0.54644	8263	0.9640	0.9600	21.43
T	0.06280	−1.0661	34.35	2.334×10^4	0.9613	0.9570	36.01
	99% Confidence Intervals						
	p_1		p_2		p_3		
R	[0.0068, 0.0423]		[−0.9836, 2.7525]		[−41.1879, 40.0951]		
T	[0.0330, 0.09260]		[−4.2100, 2.0777]		[−0.3396, 1.0266]		

Table 8.16: Cascading Tanks: Curve fitting results for the CPU times versus number of epochs.

$y = p_1x^3 + p_2x^2 + p_3x + p_4$								
	Coefficients				SSE	R^2	$\bar{R}^2$	RMSE
	p_1	p_2	p_3	p_4				
R	0.54942	5.458	31.26	−49.733	5278	0.9977	0.9970	22.97
T	1.8844	33.658	227.63	−445.97	1.8143×10^5	0.9734	0.9734	134.6
	99% Confidence Intervals							
	p_1		p_2		p_3		p_4	
R	[0.1603, 0.9386]		[−16.6306, 5.7147]		[−66.3452, 128.872]		[−302.0056, 202.5392]	
T	[−0.3963, 4.1652]		[−99.1404, 31.8236]		[−344.44, 799.70]		[−1924.5, 1032.0]	

Table 8.17: Cascading Tanks: Curve fitting results for the CPU times versus number of tanks.

$$p_{3,k,j+1,i} = (1 - \omega_{i,j+1,k})\Delta h^+_{k,j+1,i},\ \forall k \in \mathcal{K},\ \forall j \in \bar{\mathcal{J}}, \forall i \in \bar{\mathcal{I}}, \tag{8.3.12}$$

$$p_{4,k,j+1,i} = \omega_{i,j+1,k}\Delta h^-_{k,j+1,i},\ \forall k \in \mathcal{K},\ \forall j \in \bar{\mathcal{J}}, \forall i \in \bar{\mathcal{I}}, \tag{8.3.13}$$

$$\Delta h^+_{k,j+1,n_T} = h_{k,j+1,n_T},\ \Delta h^-_{k,j+1,n_T} = 0,\ \forall k \in \mathcal{K},\ \forall j \in \bar{\mathcal{J}}, \tag{8.3.14}$$

$$\omega_{k,j+1,n_T} = 1,\ \forall k \in \mathcal{K},\ \forall j \in \bar{\mathcal{J}}, \tag{8.3.15}$$

$$\Delta h^+_{k,j,i} \geq 0,\ \Delta h^-_{k,j,i} \geq 0, \forall k \in \mathcal{K},\ \forall j \in \mathcal{J},\ \forall i \in \mathcal{I}, \tag{8.3.16}$$

$$0.0 \leq \omega_{k,j,i} \leq 1.0, \forall k \in \mathcal{K},\ \forall j \in \mathcal{J},\ \forall i \in \mathcal{I}, \tag{8.3.17}$$

$$F_{k,j+1,i} = C_i w_{k,i} \frac{\omega_{i,j+1,k}\Delta h^+_{k,j+1,i}}{\sqrt{\omega_{k,j+1,i}\Delta h^+_{k,j+1,i} + k_r}},\ \forall k \in \mathcal{K},\ \forall j \in \bar{\mathcal{J}}, \forall i \in \mathcal{I}, \tag{8.3.18}$$

$$F_{k,j+1,0} = C_0 w_{k,0};\ \forall k \in \mathcal{K},\ \forall j \in \bar{\mathcal{J}}, \tag{8.3.19}$$

$$h_{k,j+1,i} = h_{k,j,i} + dt\left(\frac{F_{k,j+1,i-1} - F_{k,j+1,i}}{A_i}\right),\ \forall k \in \mathcal{K},\ \forall j \in \bar{\mathcal{J}},\ \forall i \in \mathcal{I}, \tag{8.3.20}$$

$$h_{k+1,1,i} = h_{k,n_t,i},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I}, \tag{8.3.21}$$

$$\Delta = \sum_{l=1}^{4}\sum_{k=1}^{n_e}\sum_{j=2}^{n_t}\sum_{i=1}^{n_T} p_{l,k,j,i},$$

$$p_{l,k,j,i} \geq 0,\ \forall l \in \mathcal{L},\ \forall k \in \mathcal{K},\ \forall j \in \mathcal{J},\ \forall i \in \mathcal{I},$$

$$0.25 \leq w_{k,i} \leq 1.25,\ \forall k \in \mathcal{K}, \forall i \in \mathcal{I},$$

$$h_{1,1,i} = 0.1,\ \forall i \in \mathcal{I}, \tag{8.3.22}$$

where

- n_t is the number of finite elements in each epoch;
- μ is the penalty parameter;
- $\bar{\mathcal{I}} = \mathcal{I}\backslash\{1\}$, $\bar{\mathcal{J}} = \mathcal{J}\backslash\{n_t\}$, $\bar{\mathcal{K}} = \mathcal{K}\backslash\{n_e\}$, $\mathcal{L} = \{1, \ldots, 4\}$;
- $\{\{\{u^L_{k,j,i}\}_{i=1}^{n_T}\}_{j=1}^{n_t}\}_{k=1}^{n_e}$ and $\{\{\{u^H_{k,j,i}\}_{i=1}^{n_T}\}_{j=1}^{n_t}\}_{k=1}^{n_e}$ are the deviations of the liquid heights from the desired interval;
- $\{\{\{h_{k,j,i}\}_{i=1}^{n_T}\}_{j=1}^{n_t}\}_{k=1}^{n_e}$ are the values of the liquid levels at epoch k and finite element j;
- $\{\{\{\{p_{k,j,i}\}_{i=1}^{n_T}\}_{j=1}^{n_t}\}_{k=1}^{n_e}\}_{l=1}^{4}$ are the deviations from the complementarity conditions;
- Δ is the sum of the deviations from the complementarity conditions;
- $\{\{\{\Delta h^+_{k,j,i}\}_{i=1}^{n_T}\}_{j=1}^{n_t}\}_{k=1}^{n_e}$, $\{\{\{\Delta h^-_{k,j,i}\}_{i=1}^{n_T}\}_{j=1}^{n_t}\}_{k=1}^{n_e}$, $\{\{\{\lambda^+_{k,j,i}\}_{i=1}^{n_T}\}_{j=1}^{n_t}\}_{k=1}^{n_e}$, $\{\{\{\Delta \bar{h}^+_{k,j,i}\}_{i=1}^{n_T}\}_{j=1}^{n_t}\}_{k=1}^{n_e}$, $\{\{\{\Delta \bar{h}^-_{k,j,i}\}_{i=1}^{n_T}\}_{j=1}^{n_t}\}_{k=1}^{n_e}$, $\{\{\{\omega^+_{k,j,i}\}_{i=1}^{n_T}\}_{j=1}^{n_t}\}_{k=1}^{n_e}$ are the complementarity variables;
- $\{\{\{F_{k,j,i}\}_{k=1}^{n_e}\}_{j=1}^{n_t}\}_{i=0}^{n_T}$ are the inlet and outlet flows of the tanks at epoch k and finite element j;
- dt is the time step;
- $\{\{w_{k,i}\}_{k=1}^{n_e}\}_{i=0}^{n_T}$ are the valve openings;
- X is the set $\{\{\{w_{k,i}\}_{k=1}^{n_e}\}_{i=0}^{n_T}$, $\{\{\{F_{k,j,i}\}_{k=1}^{n_e}\}_{j=1}^{n_t}\}_{i=0}^{n_T}$, $\{\{\{\Delta h^+_{k,j,i}\}_{i=1}^{n_T}\}_{j=1}^{n_t}\}_{k=1}^{n_e}$, $\{\{\{\Delta h^-_{k,j,i}\}_{i=1}^{n_T}\}_{j=1}^{n_t}\}_{k=1}^{n_e}$, $\{\{\{\lambda^+_{k,j,i}\}_{i=1}^{n_T}\}_{j=1}^{n_t}\}_{k=1}^{n_e}$, $\{\{\{\Delta \bar{h}^+_{k,j,i}\}_{i=1}^{n_T}\}_{j=1}^{n_t}\}_{k=1}^{n_e}$,

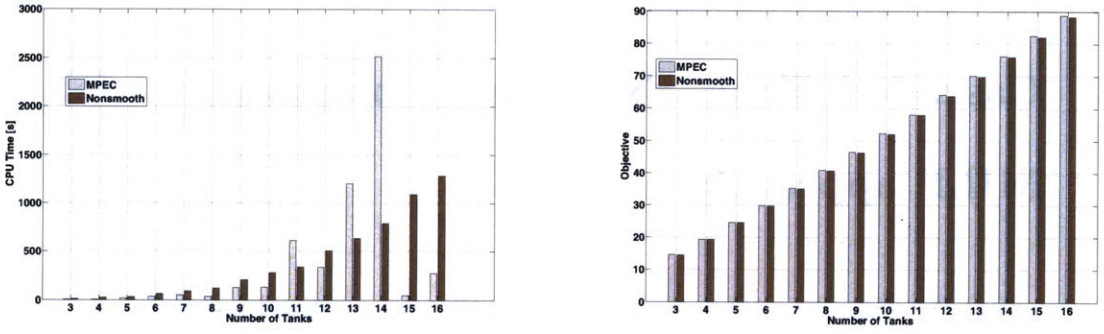
$\{\{\{\Delta\bar{h}^{-}_{k,j,i}\}_{i=1}^{n_T}\}_{j=1}^{n_t}\}_{k=1}^{n_e}$, $\{\{\{\omega^{+}_{k,j,i}\}_{i=1}^{n_T}\}_{j=1}^{n_t}\}_{k=1}^{n_e}$, $\{\{\{\{p_{k,j,i}\}_{i=1}^{n_T}\}_{j=1}^{n_t}\}_{k=1}^{n_e}\}_{l=1}^{4}$,

Δ, $\{\{\{h_{k,j,i}\}_{i=1}^{n_T}\}_{j=1}^{n_t}\}_{k=1}^{n_e}$, $\{\{\{u^{L}_{k,j,i}\}_{i=1}^{n_T}\}_{j=1}^{n_t}\}_{k=1}^{n_e}$, $\{\{\{u^{H}_{k,j,i}\}_{i=1}^{n_T}\}_{j=1}^{n_t}\}_{k=1}^{n_e}\}$.

The dynamics are discretized using an implicit Euler scheme and are represented by (8.3.20). The inlet and outlet flows computed using (8.3.18) and (8.3.19). Continuity of the state variables is ensured by (8.3.21). The complementarity conditions determining the height are in (8.3.5)-(8.3.17). Equations (8.3.2) (8.3.3) and (8.3.4) determine the deviations of liquid heights from the desired interval. Initial conditions for the states are in (8.3.22).

The program (8.3.1) is implemented in GAMS 23.1 and solved with the nonlinear programming solver IPOPT ([112]) as is done in [12] to a final tolerance of 1.0×10^{-6} on a SUSE Linux 10 Virtual Machine with 1 GB of memory and a 2.4 GHz Intel Core Duo CPU. The program is solved for various numbers of tanks and epochs. The initial X_0 is computed from data obtained by simulating the dynamics with all valve openings equal to 0.25. The number of finite elements, n_t is set to 10.

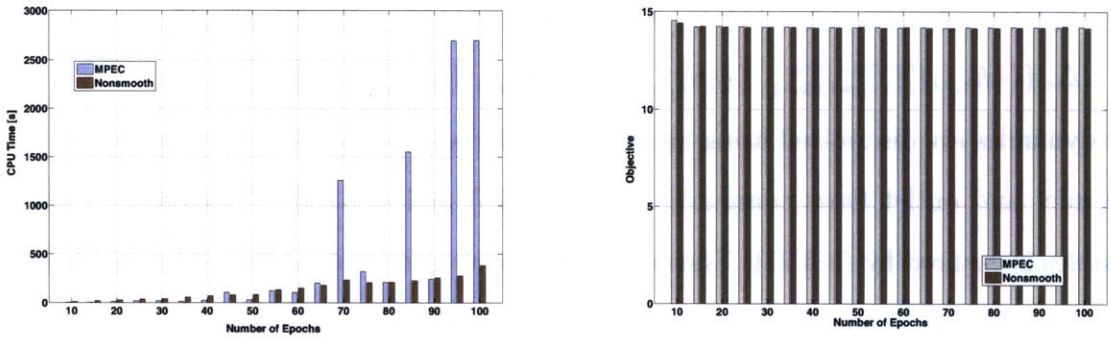
The CPU Times and the objective values are compared to the results obtained using the nonsmooth shooting method with relaxed tolerances. Figures 8-17 and 8-18 compare the CPU times and objective values. The objective values for the MPEC are the values computed by simulating the valve openings obtained as the solution of (8.3.1).



(a) CPU Time

(b) Objective

Figure 8-17: Cascading Tanks: MPEC Approach vs. The Single Nonsmooth Shooting Method for different numbers of tanks.



(a) CPU Time

(b) Objective

Figure 8-18: Cascading Tanks: MPEC Approach vs. The Nonsmooth Single Shooting Method for different numbers of epochs.

8.3.1 Conclusion

The objective values do not differ appreciably. The CPU times in case of varying tanks favor the MPEC approach slightly. The CPU times in case of varying epoch numbers favor the nonsmooth shooting method slightly. The cascading tank system has dynamics less nonlinear and stiff than the electric circuit considered in previous sections. For this example, the nonsmooth shooting method and the MPEC approach perform comparably.

8.4 Cascaded Tank Example: Comparison with the MILP approach

In this section, the cascading tanks example is solved using a mixed integer linear program (MILP) approach. Similar to the MPEC approach, the MILP approach discretizes the dynamics. Unlike the MPEC approach, binary variables are used to select the vector field with which the states evolve between two time points and the dynamics are linearized.

8.4.1 MILP Formulation

In order to apply an MILP approach similar to the one described in [106], the valve equation in (8.2.3) needs to be linearized. The set $[0.25, 1.25] \times [0.0, 1.0]$ is partitioned into subsets $D_{p,q} = [\delta w_q^L, \delta w_q^U] \times [\delta h_p^L, \delta h_p^U]$ such that:

- $\mathcal{Q} = \{1, \ldots, n_b\}$, $\mathcal{P} = \{1, \ldots, n_a\}$, $q \in \mathcal{Q}$, $p \in \mathcal{P}$;
- $\delta w_q^L = 0.25 + \frac{1.00(q-1)}{n_b}$, $\delta w_q^U = 0.25 + \frac{1.00(q)}{n_b}$, for all $q \in \mathcal{Q}$;
- $\delta h_p^L = \frac{1.00(p-1)}{n_a}$, $\delta h_p^U = \frac{1.00(p)}{n_a}$, for all $p \in \mathcal{P}$;
- $w_q^0 = (\delta w_q^L + \delta w_q^U)/2$, for all $q \in \mathcal{Q}$;
- $h_p^0 = (\delta h_p^L + \delta h_p^U)/2$, for all $p \in \mathcal{P}$;
- $F : X_1 \times X_2 \to \mathbb{R} : (x_1, x_2) \mapsto 0.1 x_1 \frac{x_2}{\sqrt{|x_2| + k_r}}$ where $X_1 = \mathbb{R}$ and $X_2 = \mathbb{R}$;
- $F_{p,q}^0 = F(h_p^0, w_q^0)$, for all $q \in \mathcal{Q}$, for all $p \in \mathcal{P}$;
- $J_w F_{p,q} = J_1 F(h_p^0, w_q^0)$, for all $q \in \mathcal{Q}$, for all $p \in \mathcal{P}$;
- $J_h F_{p,q} = J_2 F(h_p^0, w_q^0)$ for all $q \in \mathcal{Q}$, for all $p \in \mathcal{P}$.

The valve equation is approximated by the linearization:

$$F(x_1, x_2) \approx F_{p,q}^0 + J_w F_{p,q}(x_1 - w_q^0) + J_h F_{p,q}(x_2 - h_p^0) \text{ if } (x_1, x_2) \in D_{p,q}. \tag{8.4.1}$$

Note that the approximation is a discontinuous mapping on D. The approximation is multi-valued on the intersections of the boundaries of the $D_{p,q}$. Continuous approximations are possible but require more partitions of the domain, leading to more binary variables. The following MILP uses these linearizations and ensures that if the liquid level difference Δh_i is zero, there is no flow irrespective of what the linearizations predict:

$$\min_X J(X) = \sum_{k=1}^{n_e} \sum_{i=1}^{n_T} dt(u_{k,i}^L + u_{k,i}^H), \tag{8.4.2}$$

$$h_{k,i} = h_L - u_{k,i}^L, \ \forall k \in \mathcal{K}, \ \forall i \in \mathcal{I},$$

$$h_{k,i} = h_U + u^H_{k,i},\ \forall k \in \mathcal{K},\ \forall i \in \mathcal{I},$$

$$h_{k,i+1} - H_i = \Delta\bar{h}^+_{k,i} + \Delta\bar{h}^-_{k,i},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \bar{\mathcal{I}},$$

$$0.0 \geq \Delta\bar{h}^+_{k,i} - \beta_{k,i} h_{\max},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \bar{\mathcal{I}},$$

$$0.0 \leq \Delta\bar{h}^-_{k,i} + (1 - \beta_{k,i}) h_{\max},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \bar{\mathcal{I}},$$

$$\Delta\bar{h}^+_{k,n_T} = 0.0,\ \Delta\bar{h}^-_{k,n_T} = 0.0,\ \beta_{k,n_T} = 0,$$

$$\Delta\bar{h}^+_{k,i} \geq 0, 0,\ \Delta\bar{h}^-_{k,i} \leq 0.0,\ \beta_{k,i} \in \{0,1\},\ \forall k \in \bar{\mathcal{K}},$$

$$\Delta h_{k,i} = h_{k,i} - \Delta\bar{h}^+_{k,i},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I},$$

$$\Delta h_{k,i} \geq 0.0,\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \bar{\mathcal{I}},$$

$$\Delta h_{k,i} = \Delta h^+_{k,i} + \Delta h^-_{k,i},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I}, \tag{8.4.3}$$

$$0.0 \geq \Delta h^+_{k,i} - \alpha_{k,i} h_{\max},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I}, \tag{8.4.4}$$

$$0.0 \leq \Delta h^-_{k,i} + (1 - \alpha_{k,i}) h_{\max},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I}, \tag{8.4.5}$$

$$\Delta h^+_{k,i} \geq 0.0,\ \Delta h^-_{k,i} \leq 0.0,\ \alpha_{k,i} \in \{0,1\},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I}, \tag{8.4.6}$$

$$\bar{F}_{k,i,p,q} = F^0_{p,q} + J_w F_{p,q}(w_{k,i} - w^0_q) + J_h F_{p,q}(\Delta h^+_{k,i} - h^0_p), \tag{8.4.7}$$

$$\forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I},\ \forall p \in \mathcal{P},\ \forall q \in \mathcal{Q}, \tag{8.4.8}$$

$$\tilde{F}_{k,i} \leq \bar{F}_{k,i,p,q} + (1 - \gamma_{k,i,p}) F_{max} + (1 - \eta_{k,i,p}) F_{max}, \tag{8.4.9}$$

$$\forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I},\ \forall p \in \mathcal{P},\ \forall q \in \mathcal{Q}, \tag{8.4.10}$$

$$\tilde{F}_{k,i} \geq \bar{F}_{k,i,p,q} - (1 - \gamma_{k,i,p}) F_{max} - (1 - \eta_{k,i,p}) F_{max}, \tag{8.4.11}$$

$$\forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I},\ \forall p \in \mathcal{P},\ \forall q \in \mathcal{Q}, \tag{8.4.12}$$

$$\delta h_{k,i,p} \geq \delta h^L_p \gamma_{k,i,p},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I},\ \forall p \in \mathcal{P}, \tag{8.4.13}$$

$$\delta h_{k,i,p} \leq \delta h^U_p \gamma_{k,i,p},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I},\ \forall p \in \mathcal{P}, \tag{8.4.14}$$

$$\delta h_{k,i,p} \leq h_{\max} \gamma_{k,i,p},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I},\ \forall p \in \mathcal{P}, \tag{8.4.15}$$

$$\delta h_{k,i,p} \geq 0.0,\ \gamma_{k,i,p} \in \{0,1\},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I},\ \forall p \in \mathcal{P}, \tag{8.4.16}$$

$$\Delta h_{k,i} = \sum_{p=1}^{n_a} \delta h_{k,i,p}, \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I}, \tag{8.4.17}$$

$$1 = \sum_{p=1}^{n_a} \gamma_{k,i,p}, \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I}, \tag{8.4.18}$$

$$\delta w_{k,i,q} \geq \delta w_q^L \eta_{k,i,q},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I},\ \forall q \in \mathcal{Q}, \tag{8.4.19}$$

$$\delta w_{k,i,q} \leq \delta w_q^U \eta_{k,i,q},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I},\ \forall q \in \mathcal{Q}, \tag{8.4.20}$$

$$\delta w_{k,i,q} \leq w_{max} \eta_{k,i,q},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I},\ \forall q \in \mathcal{Q}, \tag{8.4.21}$$

$$\delta w_{k,i,q} \geq 0.0,\ \eta_{k,i,q} \in \{0,1\},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I},\ \forall q \in \mathcal{Q}, \tag{8.4.22}$$

$$w_{k,i} = \sum_{q=1}^{n_b} \delta w_{k,i,q},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I}, \tag{8.4.23}$$

$$1 = \sum_{q=1}^{n_b} \eta_{k,i,q},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I}, \tag{8.4.24}$$

$$0.25 \leq w_{k,i} \leq 1.25,\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I} \cup \{0\},$$

$$\tilde{F}_{k,i} = F_{k,i}^+ + F_{k,i}^-,\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I}, \tag{8.4.25}$$

$$0.0 \geq F_{k,i}^+ - \mu_{k,i} F_{max},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I}, \tag{8.4.26}$$

$$0.0 \leq F_{k,i}^- + (1 - \mu_{k,i}) F_{max},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I}, \tag{8.4.27}$$

$$F_{k,i}^+ \geq 0.0,\ F_{k,i}^- \leq 0.0,\ \mu_{k,i} \in \{0,1\},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I}, \tag{8.4.28}$$

$$F_{k,i} \leq \alpha_{k,i} K h_{max},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I}, \tag{8.4.29}$$

$$F_{k,i} \leq K \Delta h_{k,i}^+,\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I}, \tag{8.4.30}$$

$$F_{k,i} \leq F_{k,i}^+ + (1 - \alpha_{k,i}) K h_{max},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I}, \tag{8.4.31}$$

$$F_{k,i} \geq F_{k,i}^+ - (1 - \alpha_{k,i}) K h_{max},\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I}, \tag{8.4.32}$$

$$F_{k,0} = C_0 \cdot w_{k,0},\ \forall k \in \bar{\mathcal{K}},$$

$$h_{k+1,i} = h_{k,i} + \frac{dt}{A_i}(F_{k,i-1} - F_{k,i}),\ \forall k \in \bar{\mathcal{K}},\ \forall i \in \mathcal{I},$$

$$0.0 \leq h_{k,i} \leq 1.00,$$

where

- $\bar{\mathcal{K}} = \mathcal{K} \backslash \{n_e\}$, $\bar{\mathcal{I}} = \mathcal{I} \backslash \{n_T\}$;
- $\{\{h_{k,i}\}_{k=1}^{n_e}\}_{i=1}^{n_T}$ are the liquid levels at epoch k for tank i;
- $\{\{u_{k,i}^L\}_{k=1}^{n_e}\}_{i=1}^{n_T}$ and $\{\{u_{k,i}^H\}_{k=1}^{n_e}\}_{i=1}^{n_T}$ are the deviations of the liquid levels from the desired interval at epoch k for tank i;
- $\{\{w_{k,i}\}_{k=1}^{n_e}\}_{i=0}^{n_T}$ are the valve openings;
- F_{max} and K are large numbers, $h_{max} = 1.00$;
- X represents all the unknown variables in program (8.4.2).

The constraint (8.4.3) decomposes the liquid level difference, $\Delta h_i(t, \mathbf{p})$ at epoch k, into a nonnegative number, $\Delta h_{k,i}^+$ and a nonpositive number, $\Delta h_{k,i}^-$. Flow through valve i occurs only if $\Delta h_{k,i}^+ \geq 0$. If $\Delta h_{k,i} > 0$, then, the constraints (8.4.3)- (8.4.6) are satisfied only if $\Delta h_{k,i}^- = 0$, $\Delta h_{k,i}^+ = \Delta h_{k,i}$ and $\alpha_{k,i} = 1$. If $\Delta h_{k,i} < 0$, then, these constraints are satisfied only if $\Delta h_{k,i}^+ = 0$, $\Delta h_{k,i}^- = \Delta h_{k,i}$ and $\alpha_{k,i} = 0$. In case, $\Delta h_{k,i} = 0$, these constraints are satisfied if $\Delta h_{k,i}^+ = \Delta h_{k,i}^- = 0$ and $\alpha_{k,i} \in \{0, 1\}$.

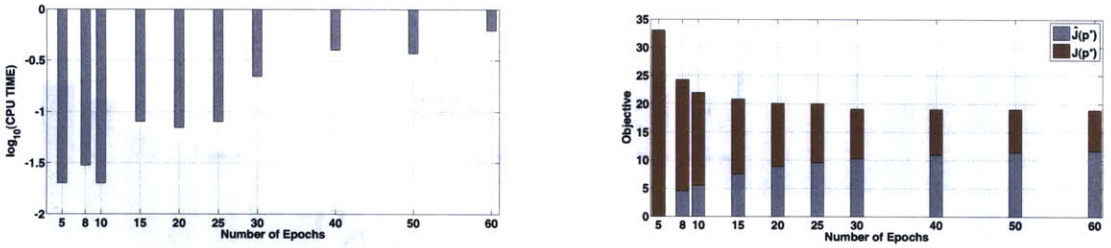
The constraints (8.4.13)-(8.4.24) determine which linearization to use by determining in which subdomain $(w_{k,i}, \Delta h_{k,i})$ is located. If $(w_{k,i}, \Delta h_{k,i}) \in D_{\bar{p},\bar{q}}$, then $\gamma_{k,i,\bar{p}} = 1$, $\eta_{k,i,\bar{p}} = 1$. The constraints (8.4.18) and (8.4.24) ensure that there is only one $\bar{p}$ and one $\bar{q}$ such that $\gamma_{k,i,\bar{p}} = 1$, $\eta_{k,i,\bar{p}} = 1$ hold given $i \in \mathcal{I}$ and $k \in \bar{\mathcal{K}}$. Using these $\gamma_{k,i,\bar{p}}$ and $\eta_{k,i,\bar{q}}$, constraints (8.4.9) and (8.4.11) determine the flows from the appropriate linearizations.

Constraints (8.4.25)-(8.4.32) ensure that the flow $F_{k,i}$ is positive or zero in case $\Delta h_{k,i}^+ = 0$. These constraints also enforce $\alpha_{k,i} = 0$ if $\Delta h_{k,i} = 0$.

8.4.2 MILP Approach Results

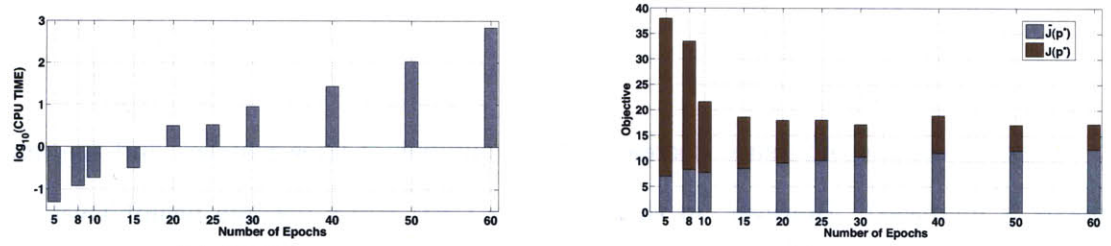
The MILP formulation (8.4.2) was implemented in GAMS 23.1 and solved with the MILP solver CPLEX using a relative optimality gap of 0.01. The optimization runs were terminated if the CPU times exceeded 10,000 seconds. The formulation was solved for different numbers

of tanks and epoch numbers using different numbers of linearizations. The CPU times of these runs as well as the objective values obtained are in Figures 8-19, 8-20, 8-21, 8-22, 8-23 and 8-24. The logarithm of the CPU times is plotted because the CPU times vary over a wide range. Note that in case $n_a = 3$ and $n_b = 3$, some optimization runs had to be terminated because the CPU times exceeded $10,000$ seconds. It can be seen that the CPU times scale exponentially with the number of epochs and states. This behavior is expected in the MILP approach.



(a) CPU Time (b) Objective

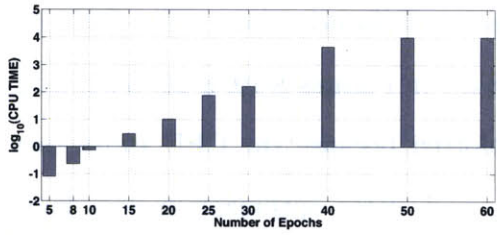
Figure 8-19: Cascading Tanks: MILP Approach CPU times and optimal objective values for different numbers of epochs and $n_a = 1$, $n_b = 1$.



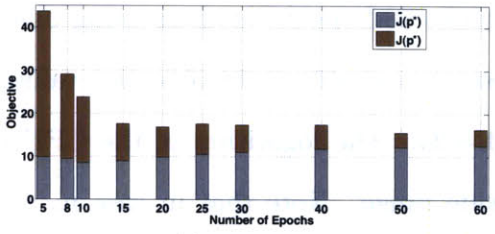
(a) CPU Time (b) Objective

Figure 8-20: Cascading Tanks: MILP Approach CPU times and optimal objective values for different numbers of epochs and $n_a = 2$, $n_b = 2$.

In the objective plots $\bar{J}(\mathbf{p}^*)$ denotes the objective values predicted by the MILP formulation and $\bar{J}(\mathbf{p})$ denotes the objective values obtained by simulating the valve openings obtained as a part of the solution of (8.4.2). The difference between $\bar{J}(\mathbf{p}^*)$ and $J(\mathbf{p}^*)$ is due to the approximation of the original nonlinear dynamics using linearizations and discretiza-

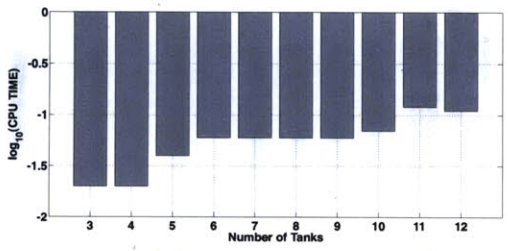


(a) CPU Time

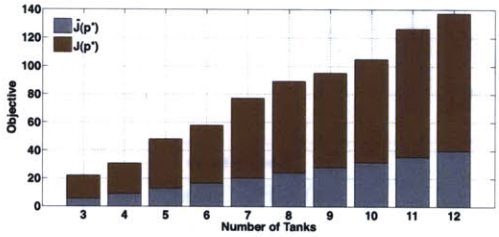


(b) Objective

Figure 8-21: Cascading Tanks: MILP Approach CPU times and optimal objective values for different numbers of epochs and $n_a = 3$, $n_b = 3$.



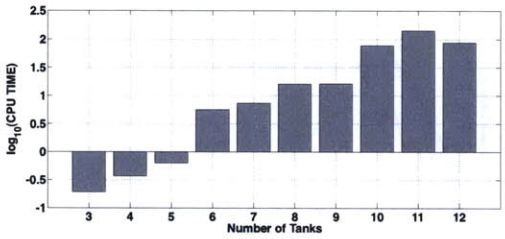
(a) CPU Time



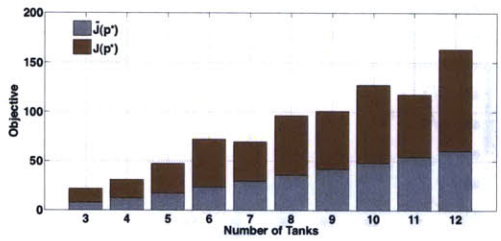
(b) Objective

Figure 8-22: Cascading Tanks: MILP Approach CPU times and optimal objective values for different numbers of tanks and $n_a = 1$, $n_b = 1$.

tion. As expected, the difference decreases if the number of linearizations and the number of epochs is increased. However, in this case, the CPU times become prohibitively large. The MILP predicted state trajectories and the actual trajectories obtained using simulation of the optimal valve openings are in Figures 8-25, 8-26 and 8-27 for various number of linearizations and $n_e = 60$. It can be seen that if $n_a = 3$, $n_b = 3$, $n_e = 60$, the dynamics are approximated well enough for the case $n_T = 3$. The comparison of objective values and CPU times for the case $n_a = 3$ and $n_b = 3$ with the nonsmooth single shooting method results using relaxed tolerances are in Figures 8-28 and 8-29. It is clear that for this example, the nonsmooth single shooting method performs better.

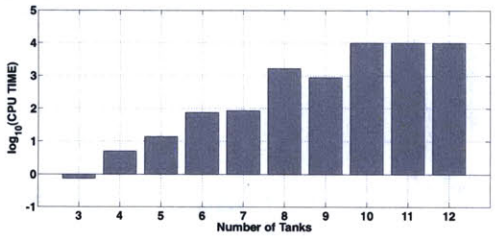


(a) CPU Time

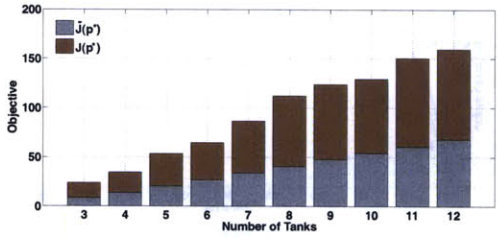


(b) Objective

Figure 8-23: Cascading Tanks: MILP Approach CPU times and optimal objective values for different numbers of tanks and $n_a = 2$, $n_b = 2$.



(a) CPU Time

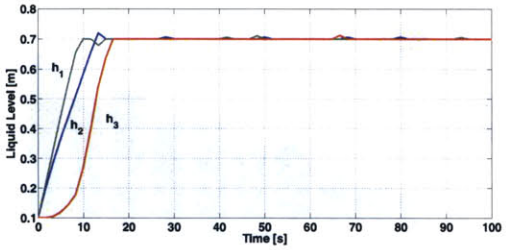


(b) Objective

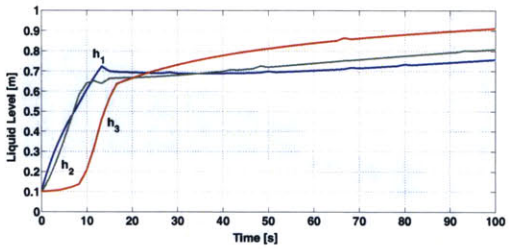
Figure 8-24: Cascading Tanks: MILP Approach CPU times and optimal objective values for different numbers of tanks and $n_a = 3$, $n_b = 3$.

8.4.3 Conclusion

The MILP approach requires comparable to or more CPU time than the nonsmooth single shooting method to produce acceptable approximations to the state trajectories of the dynamic system. The CPU times of the MILP approach seem to scale exponentially with the number of tanks and epochs as expected from the branch and bound algorithm used by the MILP solver. The MILP approach can in theory find the global minimum of (8.4.2). However, due to the CPU times required, the MILP approach can only be applied to small numbers of tanks and epochs. The main issue is the large number of linearizations and large number of epochs required to approximate the state trajectories reasonably well. For problems with nonlinear constraints such as the one considered in the Tank Change Over Case Study, not only the dynamics but also the nonlinear constraints need to be linearized, leading to problems that are intractable in a reasonable amount of time.

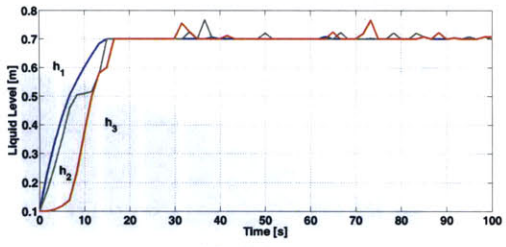


(a) MILP

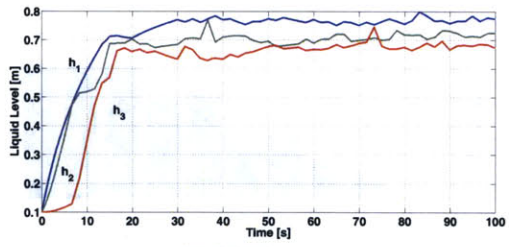


(b) Simulation

Figure 8-25: Cascading Tanks: State trajectories for $n_T = 3$, $n_e = 60$, $n_a = 1$, $n_b = 1$.



(a) MILP

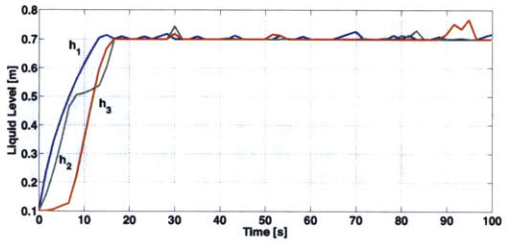


(b) Simulation

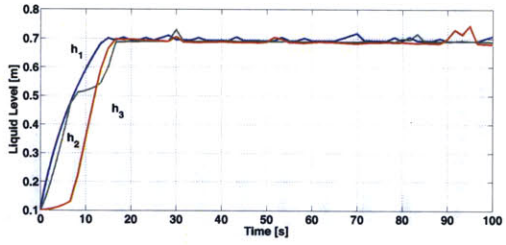
Figure 8-26: Cascading Tanks: State trajectories for $n_T = 3$, $n_e = 60$, $n_a = 2$, $n_b = 2$.

The program (8.4.2) differs from the formulation discussed in [106]. In this study, the big M method is used whereas in [106] a disjunctive formulation approach is applied. The big M method was chosen because it is conceptually simpler and easier to implement. It is known that the disjunctive formulation may result in fewer binary variables. However, possible exponential growth of the CPU time with respect to number of tanks and number of epochs is still observed in [106] even when the disjunctive formulation is used. The study in [106] does not consider the quality of the approximation of the state trajectories and does not compare simulation results to the solution of the formulated MILP. Even if the disjunctive formulation is used, the question of how to approximate the nonlinear elements of the model with linearizations well remains unanswered. A large number of epochs and linearization points is possibly still required.

In [106], the dynamics of the system are linearized without taking into account physical behavior. The linearizations used in [106] do not try to account for the fact that if $\Delta h_{k,i} = 0$

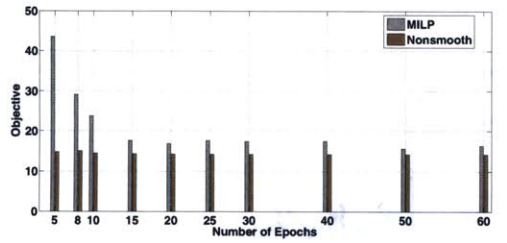


(a) MILP

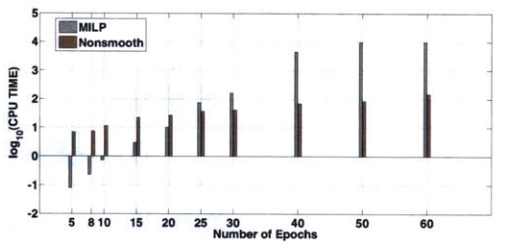


(b) Simulation

Figure 8-27: Cascading Tanks: State trajectories for $n_T = 3$, $n_e = 60$, $n_a = 3$, $n_b = 3$.



(a) Objective



(b) CPU Time

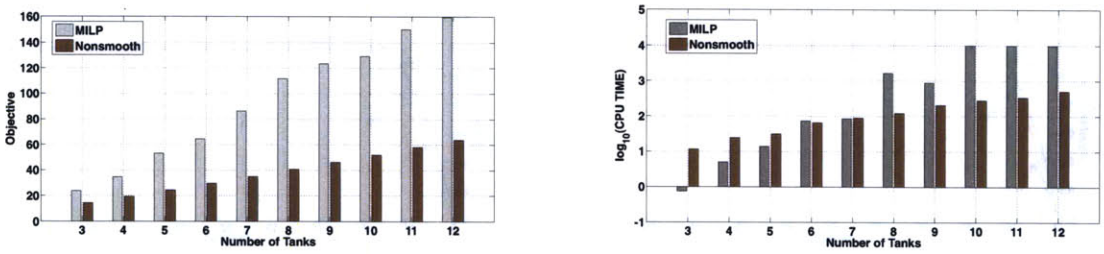
Figure 8-28: Cascading Tanks: Comparison of objective values and CPU times between the MILP approach and the nonsmooth single shooting method for varying epochs, $n_T = 3$, $n_a = 3$, $n_b = 3$.

for some $i \in \mathcal{I}$ and some $k \in \mathcal{K}$, then $F_{k,i}$ should be zero. Therefore, the modeling approach in [106] possibly incurs larger approximation error.

There have been attempts to use binary variables without linearizing the nonlinearities in the dynamics [7]. This approach produces large scale mixed-integer nonlinear programs (MINLPs). Currently, these programs cannot be solved within reasonable CPU times.

8.5 Tank Change Over

In this section, a dynamic optimization problem introduced in [9] is solved using the nonsmooth single shooting method using the transformation described in §7.3. The dynamic optimization problem aims to find the optimal schedule to change the contents of a vessel from CH_4 to O_2 in the shortest amount of time while avoiding an explosion and using N_2 to



(a) Objective (b) CPU Time

Figure 8-29: Cascading Tanks: Comparison of objective values and CPU times between the MILP approach and the nonsmooth single shooting method for varying numbers of tanks, $n_e = 10$, $n_a = 3$, $n_b = 3$.

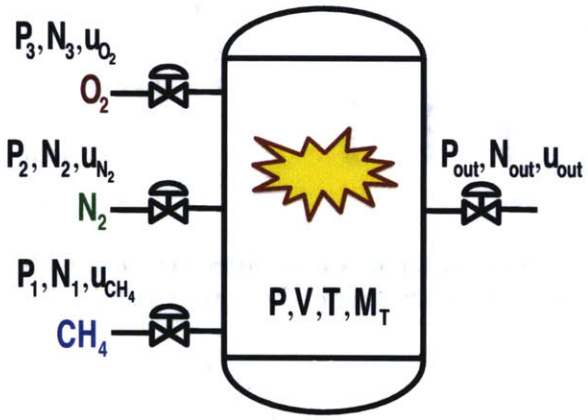


Figure 8-30: Tank Change Over: Configuration.

flush CH_4 out of the tank if necessary (Figure 8-30).

8.5.1 System Dynamics

The elements of the model are:

- t_0: initial time in seconds; $t_0 = 0.0$ s.
- t_f: final time in seconds; $t_0 \leq t_f \leq 800$ s.
- T: time interval in seconds; $T = [t_0, t_f]$.
- ΔT: the duration in seconds; $\Delta T = t_f - t_0$.
- n_e: number of epochs used in the control vector parametrization; $n_e = \{3, 4\}$.

- $\mathcal{K}$: the set of epoch indices; $\mathcal{K} = \{1, \ldots, n_e\}$.
- $\{T_k\}_{k=1}^{n_e}$: the time intervals for each epoch. $T_k = [\alpha_k, \beta_k)$ where $\alpha_1 = t_0$, $\beta_{n_e} = t_f$, $\alpha_k \leq \beta_k$, $\forall k \in \mathcal{K}$, $\beta_{k+1} = \alpha_k$, $\forall k \in \mathcal{K} \backslash \{n_e\}$.
- $\{\Delta T_k\}_{k=1}^{n_e}$: epoch durations; $\Delta T_k = \beta_k - \alpha_k$, $\forall k \in \mathcal{K}$, $0.0 \text{ s} \leq \Delta T_k \leq 200.0 \text{ s}$.
- $\mathcal{J} = \{\text{CH}_4 = 1, \text{N}_2 = 2, \text{O}_2 = 3\}$: the set of chemical species indices.
- $\bar{\mathcal{J}} = \{\text{CH}_4 = 1, \text{N}_2 = 2, \text{O}_2 = 3, \text{out} = 4\}$: the set of valve indices.
- $\{\{u_{k,j}\}_{k=1}^{n_e}\}_{j=1}^{4}$: the valve opening at epoch k for valve j, $0.0 \leq u_{k,j} \leq 1.0, \forall k \in \mathcal{K}$, $j \in \bar{\mathcal{J}}$.
- n_p: the number of parameters; $n_p = 4 \times n_e + n_e$.
- $\mathbf{p}$: the parameters to be adjusted, $\mathbf{p} = \{\{\{u_{k,j}\}_{k=1}^{n_e}\}_{j=1}^{4}, \{\Delta T_k\}_{k=1}^{n_e}\}$.
- $\mathcal{P}$: the parameter set. $\mathcal{P}$ is an open subset of $\mathbb{R}^{n_p}$ such that $[0.0, 1.0]^{4n_e} \times [0.0, 200.0]^{n_e} \subset \mathcal{P}$.
- $M_j : T \times \mathcal{P} \to \mathbb{R}$, $j \in \mathcal{J}$ are the number of moles of each chemical species in the tank.
- $\mathbf{x} : T \times \mathcal{P} \to \mathbb{R}^3$: the continuous state of the system; $\mathbf{x}(t, \mathbf{p}) = (M_{\text{CH}_4}(t, \mathbf{p}), M_{\text{N}_2}(t, \mathbf{p}), M_{\text{O}_2}(t, \mathbf{p}))$.
- $\mathcal{X}$: state space of the system. $\mathcal{X} = \mathbb{R}^3$.
- $P : T \times \mathcal{P} \to \mathbb{R}$: the pressure in the tank in bars.
- $u_j : T \times \mathcal{P} \to [0, 1]$, $j \in \mathcal{J}$: the controls of the system. $u_j(t, \mathbf{p}) = u_{k,j}$, $\forall t \in T_k$, $\forall j \in \bar{\mathcal{J}}$, $\forall k \in \mathcal{K}$. $u_j(t_f, \mathbf{p}) = u_{n_e,j}$, $\forall j \in \bar{\mathcal{J}}$.
- V: the volume of the tank in m^3.
- T: the temperature of the whole system in K.
- R: the universal gas constant in $\frac{\text{bar} \cdot \text{m}^3}{\text{mol} \cdot \text{K}}$.
- P_j, $j \in \bar{\mathcal{J}}$: supply and discharge pressures in bar.
- $C_{v,j}$, $j \in J$ and $C_{v,\text{out}}$: valve constants in $\frac{\text{mol}}{\text{s} \cdot \text{bar}}$.
- k_b: a small regularization constant in bar.

The values of the parameters of the system are given in Table 8.18.

The equations describing the dynamics are:

$$\dot{M}_j(t,\mathbf{p}) = N_j(t,\mathbf{p}) - N_{\text{out}}(t,\mathbf{p})y_j(t,\mathbf{p}),\ \forall j \in \mathcal{J},\ \forall t \in (t_0,t_f], \tag{8.5.1}$$

$$M_T(t,\mathbf{p}) = \sum_{j\in\mathcal{J}} M_j(t,\mathbf{p}),\ \forall t \in [t_0,t_f],$$

$$y_j(t,\mathbf{p}) = \frac{M_j(t,\mathbf{p})}{M_T(t,\mathbf{p})},\ \forall j \in \mathcal{J},\ \forall t \in [t_0,t_f],$$

$$P(t,\mathbf{p}) = M_T(t,\mathbf{p})\frac{RT}{V},\ \forall t \in [t_0,t_f], \tag{8.5.2}$$

$$N_j(t,\mathbf{p}) = \begin{cases} 0 & \text{if } \frac{P(t,\mathbf{p})}{P_j} \geq 1, \\ u_j(t,\mathbf{p})C_{v,j}\sqrt{\frac{P_j+P(t,\mathbf{p})}{2}}\frac{P_j-P(t,\mathbf{p})}{\sqrt{|P_j-P(t,\mathbf{p})|+k_bP_j}} & \text{if } 0.53 \leq \frac{P(t,\mathbf{p})}{P_j} < 1,\ \forall j \in \mathcal{J},\ \forall t \in [t_0,t_f], \\ u_j(t,\mathbf{p})C_kC_{v,j}\frac{P_j}{\sqrt{2}}0.85 & \text{if } \frac{P(t,\mathbf{p})}{P_j} < 0.53, \end{cases} \tag{8.5.3}$$

$$N_{\text{out}}(t,\mathbf{p}) = \begin{cases} 0 & \text{if } \frac{P_{\text{out}}}{P(t,\mathbf{p})} \geq 1, \\ u_{\text{out}}(t,\mathbf{p})C_{v,\text{out}}\sqrt{\frac{P(t,\mathbf{p})+P_{\text{out}}}{2}}\frac{P(t,\mathbf{p})-P_{\text{out}}}{\sqrt{|P(t,\mathbf{p})-P_{\text{out}}|+k_bP(t,\mathbf{p})}} & \text{if } 0.53 \leq \frac{P_{\text{out}}}{P(t,\mathbf{p})} < 1,\ \forall t \in [t_0,t_f], \\ u_{\text{out}}(t,\mathbf{p})C_kC_{v,\text{out}}\frac{P(t,\mathbf{p})}{\sqrt{2}}0.85 & \text{if } \frac{P_{\text{out}}}{P(t,\mathbf{p})} < 0.53, \end{cases} \tag{8.5.4}$$

$$C_k = \frac{0.47\cdot\sqrt{1.53}}{0.85\cdot\sqrt{0.47+k_b}},$$

$$M_{\text{CH}_4}(t_0,\mathbf{p}) = 900.0,\ M_{\text{O}_2}(t_0,\mathbf{p}) = 0.0,\ M_{\text{N}_2}(t_0,\mathbf{p}) = 0.0. \tag{8.5.5}$$

Equation (8.5.1) represents mass conservation. The gases in the tank are assumed to be perfectly mixed ideal gases and (8.5.2) is the ideal gas law. Equation (8.5.3) determines the inlet flow rates depending on the inlet and tank pressures. If the tank pressure is higher than the inlet pressure, there is no flow. If the tank pressure is low, then the flow is choked and depends only on the inlet pressure. Otherwise the flow is non-choked and depends both on the inlet and tank pressures. The flow out of the tank is governed by (8.5.4). The valve

T	300.0 K	V	3.0 m^3
C_{v,O_2}	8.0 $\frac{mol}{s \cdot bar}$	C_{v,CH_4}	8.0 $\frac{mol}{s \cdot bar}$
C_{v,N_2}	8.0 $\frac{mol}{s \cdot bar}$	$C_{v,out}$	8.0 $\frac{mol}{s \cdot bar}$
R	$8.314 \cdot 10^{-5} \frac{bar \cdot m^3}{mol \cdot K}$	P_{O_2}	12.0 bar
P_{CH_4}	10.0 bar	P_{N_2}	7.0 bar
P_{out}	2.0 bar	k_b	$1.0 \cdot 10^{-3}$ bar

Table 8.18: Tank Change Over: Model parameter values.

equations for non-choked flow are regularized using k_b in order to have PC^1 equations. C_k is a constant to ensure continuity of the valve equations. The valve equations differ from the equations presented in [9] due to the application of regularization. Finally, the initial conditions are in (8.5.5).

8.5.2 Safe Operation Conditions for the Tank

There is a chance of forming an explosive mixture during the change over operation. If the mole fractions satisfy the relation $h(y_{\mathrm{CH_4}}(t,\mathbf{p}), y_{\mathrm{O_2}}(t,\mathbf{p})) \leq 0$ where $h : [0,1] \times [0,1] \to \mathbb{R}$ is defined by

$$h(v,w) = \begin{cases} \sum_{i=1}^{5} \alpha_i (v \cdot 100.0)^{i-4} - (1-v-w) \cdot 100.0 \text{ if } v \in [0.03, 063] \\ 0 \text{ if otherwise} \end{cases}$$

then an explosion cannot occur. The coefficients are tabulated in Table 8.19. The function h is discontinuous at $v = 0.03$ and $v = 0.63$. However, for optimization purposes the points of interest are $S = \{(v,w) : h(v,w) \geq 0\}$ and they coincide with the points $S' = \{(v,w) : \max\{h(v,w), 0\} \geq 0\}$ where $\hat{h}(v,w) = \max\{h(v,w), 0\}$ is a PC^1 function.

α_1	α_2	α_3	α_4	α_5
−4761.168938	892.159351	−35.94512586	93.63386543	−1.480461088

Table 8.19: Tank Change Over: Path constraint polynomial coefficients.

8.5.3 Dynamic Optimization Formulation

The goal is to achieve the tank change over in the shortest duration without causing an explosion. The corresponding program is:

$$\min_{\mathbf{p}\in\mathcal{P}} J(\mathbf{p}) = t_f \tag{8.5.6}$$

$$\text{s.t. } \hat{h}(y_{CH_4}(t,\mathbf{p}), y_{O_2}(t,\mathbf{p})) \leq 0,\ \forall t \in [t_0, t_f], \tag{8.5.7}$$

$$y_{O_2}(t_f,\mathbf{p}) \geq 0.999,\ y_{CH_4}(t_f,\mathbf{p}) \leq 0.001,$$

$$0.0 \leq \Delta T_k \leq 200.0,\ \forall k \in \mathcal{K},$$

$$0.0 \leq u_{k,j} \leq 1.0,\ \forall k \in \mathcal{K},\ \forall j \in \bar{\mathcal{J}},$$

where the tank dynamics evolve according to equations discussed in §8.5.1.

The path constraint (8.5.7) needs to be enforced at all times. However, this is not practically possible. Instead, this constraint will be replaced with an end point constraint. Note that $\hat{h}$ is either zero or a positive number. Then if the constraint (8.5.7) holds for almost all $t \in [t_0, t_f]$ then $\int_0^{t_f} \hat{h}(y_{CH_4}(t,\mathbf{p}), y_{O_2}(t,\mathbf{p})) \leq 0$ except on a set of measure zero. The state trajectories of the dynamic system are absolutely continuous functions of time given piecewise constant controls. Since $\hat{h}$ is a continuous function of its arguments, $\int_0^{t_f} \hat{h}(y_{CH_4}(t,\mathbf{p}), y_{O_2}(t,\mathbf{p})) \leq 0$ implies that the path constraint is satisfied for all $t \in [t_0, t_f]$.

Formulation (8.5.6) is a nonsmooth optimization problem involving variable integration intervals. The problem is converted into an equivalent problem with fixed integration inter-

vals using the transformation in §7.3. The transformed dynamics are:

$$
\begin{aligned}
\dot{\tau}(\epsilon,\mathbf{p}) &= \Delta T_k,\ \forall \epsilon \in (k-1,k], \forall k \in \mathcal{K},\\
\dot{\hat{M}}_j(\epsilon,\mathbf{p}) &= \Delta T_k(\hat{N}_j(\epsilon,\mathbf{p}) - \hat{N}_{\text{out}}(\epsilon,\mathbf{p})\hat{y}_j(\epsilon,\mathbf{p})),\ \forall j \in \mathcal{J},\ \forall \epsilon \in (k-1,k],\ \forall k \in \mathcal{K},\\
\hat{M}_T(\epsilon,\mathbf{p}) &= \sum_{j\in\mathcal{J}} \hat{M}_j(\epsilon,\mathbf{p}),\ \forall \epsilon \in [0,n_e],\\
\hat{y}_j(\epsilon,\mathbf{p}) &= \frac{\hat{M}_j(\epsilon,\mathbf{p})}{\hat{M}_T(\epsilon,\mathbf{p})},\ \forall j \in \mathcal{J},\ \forall \epsilon \in [0,n_e],\\
\hat{P}(\epsilon,\mathbf{p}) &= \hat{M}_T(\epsilon,\mathbf{p})\frac{RT}{V},\ \forall \epsilon \in [0,n_e],\\
\hat{N}_j(\epsilon,\mathbf{p}) &= \begin{cases} 0 & \text{if } \frac{\hat{P}(\epsilon,\mathbf{p})}{P_j} \geq 1,\\ \hat{u}_j(\epsilon,\mathbf{p})C_{v,j}\sqrt{\frac{P_j+\hat{P}(\epsilon,\mathbf{p})}{2}}\frac{P_j-\hat{P}(\epsilon,\mathbf{p})}{\sqrt{|P_j-\hat{P}(\epsilon,\mathbf{p})|+k_bP_j}} & \text{if } 0.53 \leq \frac{\hat{P}(\epsilon,\mathbf{p})}{P_j} < 1,\ \forall j \in \mathcal{J},\ \forall \epsilon \in [0,n_e],\\ \hat{u}_j(\epsilon,\mathbf{p})C_kC_{v,j}\frac{P_j}{\sqrt{2}}0.85 & \text{if } \frac{\hat{P}(\epsilon,\mathbf{p})}{P_j} < 0.53, \end{cases}\\
\hat{N}_{\text{out}}(\epsilon,\mathbf{p}) &= \begin{cases} 0 & \text{if } \frac{P_{\text{out}}}{P'(\epsilon,\mathbf{p})} \geq 1,\\ \hat{u}_{\text{out}}(\epsilon,\mathbf{p})C_{v,\text{out}}\sqrt{\frac{\hat{P}(\epsilon,\mathbf{p})+P_{\text{out}}}{2}}\frac{\hat{P}(\epsilon,\mathbf{p})-P_{\text{out}}}{\sqrt{|\hat{P}(\epsilon,\mathbf{p})-P_{\text{out}}|+k_b\hat{P}(\epsilon,\mathbf{p})}} & \text{if } 0.53 \leq \frac{P_{\text{out}}}{\hat{P}(\epsilon,\mathbf{p})} < 1,\ \forall \epsilon \in [0,n_e],\\ \hat{u}_{\text{out}}(\epsilon,\mathbf{p})C_kC_{v,\text{out}}\frac{\hat{P}(\epsilon,\mathbf{p})}{\sqrt{2}}0.85 & \text{if } \frac{P_{\text{out}}}{\hat{P}(\epsilon,\mathbf{p})} < 0.53, \end{cases}\\
C_k &= \frac{0.47\cdot\sqrt{1.53}}{0.85\cdot\sqrt{0.47+k_b}}\\
\hat{M}_{\text{CH}_4}(0,\mathbf{p}) &= 900.0,\ \hat{M}_{\text{O}_2}(0,\mathbf{p}) = 0.0,\ \hat{M}_{\text{N}_2}(0,\mathbf{p}) = 0.0, \tau(0,\mathbf{p}) = t_0,
\end{aligned}
$$

where the following hold

$$
\begin{aligned}
M_j(\tau(\epsilon,\mathbf{p}),\mathbf{p}) &= \hat{M}_j(\epsilon,\mathbf{p}),\ \forall j \in \mathcal{J},\ \forall \epsilon \in [0,n_e],\\
N_j(\tau(\epsilon,\mathbf{p}),\mathbf{p}) &= \hat{N}_j(\epsilon,\mathbf{p}),\ \forall j \in \bar{\mathcal{J}},\ \forall \epsilon \in [0,n_e],\\
u_j(\tau(\epsilon,\mathbf{p}),\mathbf{p}) &= \hat{u}_j(\epsilon,\mathbf{p}), \forall j \in \bar{\mathcal{J}},\ \forall \epsilon \in [0,n_e],\\
\tau(n_e,\mathbf{p}) &= t_f = \sum_{k=1}^{n_e}\Delta T_k + t_0.
\end{aligned}
$$

The final optimization formulation is

$$J(\mathbf{p}) = \min_{\mathbf{p}\in\mathcal{P}} \sum_{i=1}^{n_e} \Delta T_k \tag{8.5.8}$$

$$\text{s.t.} \sum_{k=1}^{n_e-1} \int_k^{k+1} \Delta T_k \hat{h}(\hat{y}_{\mathrm{CH_4}}(\epsilon,\mathbf{p}), \hat{y}_{\mathrm{O_2}}(\epsilon,\mathbf{p}) d\epsilon \leq 0, \tag{8.5.9}$$

$$0.999 - \hat{y}_{\mathrm{O_2}}(n_e,\mathbf{p}) \leq 0, \tag{8.5.10}$$

$$\hat{y}_{\mathrm{CH_4}}(n_e,\mathbf{p}) - 0.001 \leq 0, \tag{8.5.11}$$

$$0.0 \leq \Delta T_k \leq 200.0, \ \forall k \in \mathcal{K},$$

$$0.0 \leq u_{k,j} \leq 1.0, \ \forall k \in \mathcal{K}, \ \forall j \in \bar{\mathcal{J}},$$

where the system evolves according to the transformed dynamics.

8.5.4 Nonsmooth Single Shooting Method Results

For the integration of the dynamics and associated sensitivity equations DSL48SE ([108, 109, 36]) is used with code generated by DAEPACK ([107]). The nonlinear program is solved by the proximal bundle solver in [64] on a SUSE Linux 10 Virtual Machine with 1 GB of RAM and a 2.4 GHz Intel Core Duo CPU. The absolute and relative integration tolerances are set to 1×10^{-8}. The proximal bundle solver optimality tolerance is set to 1×10^{-6}.

The constraints (8.5.9), (8.5.11) and (8.5.10) are appended to the objective using exact penalization to obtain the augmented objective

$$\sum_{i=1}^{n_e} \Delta T_k + \mu_1 \max(0.00, 0.999 - \hat{y}_{\mathrm{O_2}}(n_e,\mathbf{p})) + \mu_2 \max(0.00, \hat{y}_{\mathrm{CH_4}}(n_e,\mathbf{p}) - 0.001) +$$
$$\mu_3 \sum_{k=1}^{n_e-1} \int_k^{k+1} \Delta T_k \hat{h}(\hat{y}_{\mathrm{CH_4}}(\epsilon,\mathbf{p}), \hat{y}_{\mathrm{O_2}}(\epsilon,\mathbf{p})) d\epsilon.$$

where μ_1, μ_2 and μ_3 are positive penalty parameters. The dynamic optimization program

is solved repeatedly using monotonically increasing sequences of penalty parameters. The sequence of penalty parameters are $\mu_1(l) = 5000 \cdot l$, $\mu_2(l) = 4000 \cdot l$ and $\mu_3(l) = 1000 \cdot l$ where $l = 1, \ldots, \infty$ is the index of the dynamic optimization program solved. Each problem is solved to optimality. If the solution of the lth problem is not a feasible point of the original problem, l is set to $l + 1$ and the process is repeated. Otherwise, the solution of the lth problem is stationary for the original problem (see the Electric Circuit Case Study (§8.1) for more information on the exact penalty approach) and the process is terminated.

Three Epochs

The solutions of the program (8.5.8) for the case $n_e = 3$ are presented in this section. The initial parameter values are in Table 8.21 and the solution is in Table 8.20. The final objective and constraint values are in Table 8.22. The solution is obtained in 75.80 seconds. The total number of bundle solver iterations is 421 and the total number of times the dynamics are simulated is 425. The optimal tank change over time obtained is close to the optimal change over times reported in [9]. The number of times the dynamics are simulated are significantly less than those reported in [9]. This is mainly due to the use of exact derivative information.

The initial and final mole fraction profiles corresponding to the initial parameters and solution of program (8.5.8) are in Figures 8-32 and 8-31. The path constraint and the mole fractions of CH_4 and N_2 are shown in Figure 8-33. The initial mole fractions and final mole fractions are marked by arrows. Note that during operation, the system gets very close to the unsafe zone. In order to minimize the change over time, it is expected that the system operates close to the unsafe zone.

Four Epochs

This section contains the solution of the program (8.5.8) for the case $n_e = 4$. The program is solved in 536 seconds. The number of iterations carried out by the bundle solver is 1821

k	ΔT_k	$u_{k,\mathrm{CH_4}}$	$u_{k,\mathrm{O_2}}$	$u_{k,\mathrm{N_2}}$	$u_{k,\mathrm{out}}$
1	76.97 s	0.00	0.00	1.00	1.00
2	110.31 s	0.00	1.00	1.00	1.00
3	50.78 s	0.00	1.00	0.00	1.00

Table 8.20: Tank Change Over: Solution of program (8.5.8) for $n_e = 3$.

k	ΔT_k	$u_{k,\mathrm{CH_4}}$	$u_{k,\mathrm{O_2}}$	$u_{k,\mathrm{N_2}}$	$u_{k,\mathrm{out}}$
1	10.0 s	0.5	0.5	0.5	0.5
2	10.0 s	0.5	0.5	0.5	0.5
3	10.0 s	0.5	0.5	0.5	0.5

Table 8.21: Tank Change Over: Initial parameter values used to solve program (8.5.8) for $n_e = 3$.

	$J(\mathbf{p})$	Constraint (8.5.11)	Constraint (8.5.10)	Constraint (8.5.9)
Initial	30.00	0.5956	0.5950	10.3184
Final	238.06	-9.0×10^{-4}	1.0×10^{-4}	0.00

Table 8.22: Tank Change Over Case Study: Objective and constraint values of program (8.5.8) for $n_e = 3$.

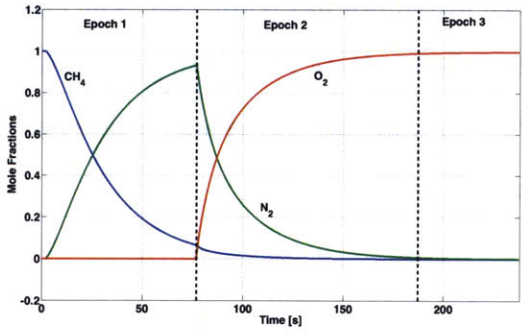


Figure 8-31: Tank Change Over: Final mole fraction profiles corresponding to the solution of (8.5.8) for $n_e = 3$.

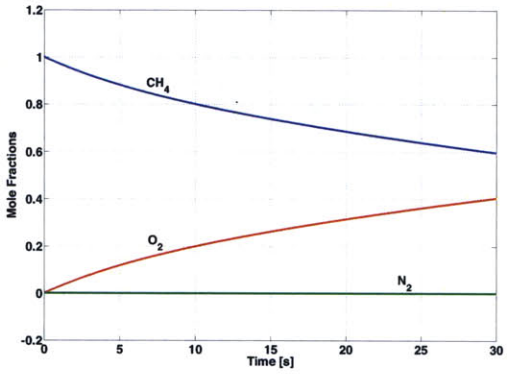


Figure 8-32: Tank Change Over: Initial mole fraction profiles corresponding to parameters in Table 8.21 of program (8.5.8) for $n_e = 3$.

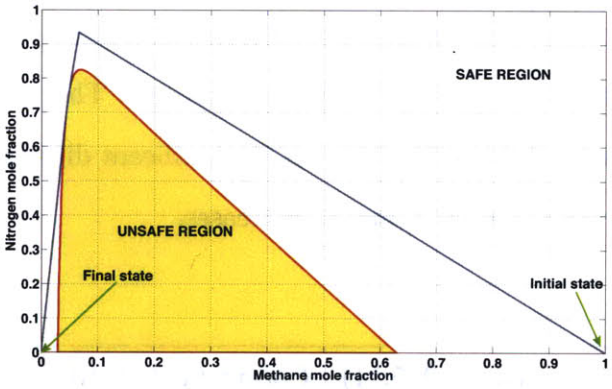


Figure 8-33: Tank Change Over: Plot of the path constraint and the mole fraction profiles corresponding to the solution of (8.5.8) for $n_e = 3$.

k	ΔT_k	$u_{k,\mathrm{CH_4}}$	$u_{k,\mathrm{O_2}}$	$u_{k,\mathrm{N_2}}$	$u_{k,\mathrm{out}}$
1	86.31 s	0.00	0.0240	0.6286	1.00
2	43.19 s	0.00	0.9791	0.0003	1.00
3	51.79 s	0.00	1.00	0.00	1.00
4	56.50 s	0.00	1.00	0.00	1.00

Table 8.23: Tank Change Over: Solution of program (8.5.8) for $n_e = 4$.

k	ΔT_k	$u_{k,\mathrm{CH_4}}$	$u_{k,\mathrm{O_2}}$	$u_{k,\mathrm{N_2}}$	$u_{k,\mathrm{out}}$
1	10.0 s	0.5	0.5	0.5	0.5
2	10.0 s	0.5	0.5	0.5	0.5
3	10.0 s	0.5	0.5	0.5	0.5
3	10.0 s	0.5	0.5	0.5	0.5

Table 8.24: Tank Change Over: Initial parameter values used to solve program (8.5.8) with $n_e = 4$.

and the number of times the dynamics were integrated is 1917. The solution and initial parameter values are in Tables 8.23 and 8.24, respectively. The initial and final objective and constraint values are in Table 8.25. There is no significant difference in the optimal tank change over times between the $n_e = 3$ and $n_e = 4$ cases.

	$J(\mathbf{p})$	Constraint (8.5.11)	Constraint (8.5.10)	Constraint (8.5.9)
Initial	40.00	0.5223	0.5213	116.71
Final	237.79	6.6×10^{-5}	9.7×10^{-5}	0.00

Table 8.25: Tank Change Over: Objective and constraint values of program (8.5.8) with $n_e = 4$.

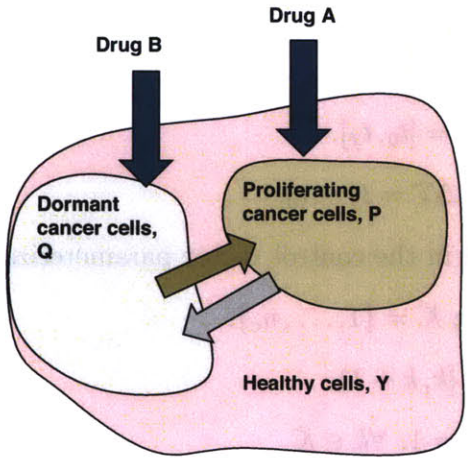


Figure 8-34: Chemotherapy Scheduling: Configuration

8.5.5 Conclusion

The time taken to solve (8.5.8) depends strongly on the policy with which the penalty parameters are updated. In this study, a very simple schedule of monotonically increasing parameters is used. It is expected that more sophisticated penalty updating policies will reduce the solution times.

8.6 Chemotherapy Scheduling Case Study

In this section, a modified version of the cell cycle specific chemotherapy model introduced in [82] is used to determine an optimal chemotherapy drug schedule. The example is used to compare the nonsmooth single shooting method with a derivative-free method.

8.6.1 System Dynamics

The elements of the model are:

- t_0: initial time in days; $t_0 = 1.0$ day.
- t_f: final time in days; $t_f = 31$ day.
- T: time interval in days; $T = [t_0, t_f]$.
- ΔT: the duration in days; $\Delta T = t_f - t_0$.
- n_e: number of epochs used in the control vector parameterization; $n_e = 30$.
- $\mathcal{K}$: the set of epoch indices; $\mathcal{K} = \{1, \ldots, n_e\}$.
- T_k: epoch k interval. $T_k = [k, k+1)$.
- ΔT_k: epoch duration, $\Delta T_k = 1$, $\forall k \in \mathcal{K}$.
- $\{u_{A,k}\}_{k=1}^{n_e}$: drug A dosage in units of drug concentration, $[D]$; $0 \leq u_{A,k} \leq 20.00$.
- $\{u_{B,k}\}_{k=1}^{n_e}$: drug B dosage in units of drug concentration, $[D]$; $0 \leq u_{B,k} \leq 20.00$.
- n_p: number of parameters, $n_p = 2n_e$.
- $\mathbf{p}$: the parameters to be adjusted. $\mathbf{p} = \{\{u_{A,k}\}_{k=1}^{n_e}, \{u_{B,k}\}_{k=1}^{n_e}\}$.
- $\mathcal{P}$: the parameter set, an open subset of $\mathbb{R}^{2n_e}$ such that $[0, 20]^{2n_e} \subset \mathcal{P}$.
- $u_A : T \times \mathcal{P} \to [0, 20]$: drug A schedule such that $u_A(t, \mathbf{p}) = u_{A,k}$, $\forall t \in t_k$, $u_A(t_f, \mathbf{p}) = u_{A,n_e}$.
- $u_B : T \times \mathcal{P} \to [0, 20]$: drug B schedule such that $u_B(t, \mathbf{p}) = u_{B,k}$, $\forall t \in t_k$, $u_B(t_f, \mathbf{p}) = u_{B,n_e}$.
- $P : T \times \mathcal{P} \to \mathbb{R}$: the size of the proliferating cancer cell population in the tissue.
- $Q : T \times \mathcal{P} \to \mathbb{R}$: the size of the quiescent cancer cell population in the tissue.
- $Y : T \times \mathcal{P} \to \mathbb{R}$: the size of the healthy cell population in the tissue.
- $v_A : T \times \mathcal{P} \to \mathbb{R}$: drug A concentration in the tissue.
- $v_B : T \times \mathcal{P} \to \mathbb{R}$: drug B concentration in the tissue.
- $\mathbf{x} : T \times \mathcal{P} \to \mathbb{R}^5$: the continuous state of the system, $\mathbf{x}(t, \mathbf{p}) = (P(t, \mathbf{p}), Q(t, \mathbf{p}), Y(t, \mathbf{p}), v_A(t, \mathbf{p}), v_B(t, \mathbf{p}))$.
- $\mathcal{X}$: the state space; $\mathcal{X} = \mathbb{R}^5$.

The equations governing the system evolution are:

$$\alpha = a - m - n,$$

$$\dot{P}(t,\mathbf{p}) = \alpha P(t,\mathbf{p}) + bQ(t,\mathbf{p}) - F_A(v_A(t,\mathbf{p}), P(t,\mathbf{p})), \forall t \in (t_0, t_f], \tag{8.6.1}$$

$$F_A(v_A(t,\mathbf{p}), P(t,\mathbf{p})) = \begin{cases} 0 & \text{if } v_A(t,\mathbf{p}) - \bar{v}_A \leq 0, \\ k_A(v_A(t,\mathbf{p}) - \bar{v}_A)P(t,\mathbf{p}) & \text{if } v_A(t,\mathbf{p}) - \bar{v}_A \geq 0, \end{cases}, \forall t \in [t_0, t_f],$$

$$\dot{Q}(t,\mathbf{p}) = mP(t,\mathbf{p}) - bQ(t,\mathbf{p}) - F_B(v_B(t,\mathbf{p}), Q(t,\mathbf{p})), \forall t \in (t_0, t_f], \tag{8.6.2}$$

$$F_B(v_B(t,\mathbf{p}), Q(t,\mathbf{p})) = \begin{cases} 0 & \text{if } v_B(t,\mathbf{p}) - \bar{v}_B \leq 0, \\ k_B(v_B(t,\mathbf{p}) - \bar{v}_B)Q(t,\mathbf{p}) & \text{if } v_B(t,\mathbf{p}) - \bar{v}_B \geq 0, \end{cases}, \forall t \in [t_0, t_f],$$

$$\dot{Y}(t,\mathbf{p}) = \sigma Y(t,\mathbf{p})(1 - Y(t,\mathbf{p})/K) - k_A v_A(t,\mathbf{p})Y - k_B v_B(t,\mathbf{p})Y(t,\mathbf{p})), \forall t \in (t_0, t_f], \tag{8.6.3}$$

$$\dot{v}_A(t,\mathbf{p}) = u_A(t,\mathbf{p}) - \gamma_A v_A(t,\mathbf{p}), \forall t \in (t_0, t_f], \tag{8.6.4}$$

$$\dot{v}_B(t,\mathbf{p}) = u_B(t,\mathbf{p}) - \gamma_B v_B(t,\mathbf{p}), \forall t \in (t_0, t_f], \tag{8.6.5}$$

$$P(t_0,\mathbf{p}) = 2.00 \times 10^{11},\ Q(t_0,\mathbf{p}) = 8.00 \times 10^{11},\ Y(t_0,\mathbf{p}) = 1.00 \times 10^{10}, \tag{8.6.6}$$

$$v_A(t_0,\mathbf{p}) = 0.0,\ v_B(t_0,\mathbf{p}) = 0.0. \tag{8.6.7}$$

The equations describe the behavior of tumor cells and healthy cells in human tissue under chemotherapy (Figure 8-34). The tissue comprises healthy cells, Y, proliferating tumor cells, P, and quiescent tumor cells, Q. Chemotherapy comprises two drugs; A and B. u_A and u_B are the chemotherapy drug schedules. v_A and v_B are the exponentially decaying drug concentrations in the tissue. Tumor cells develop resistance to drugs. As a result, drugs are effective against the tumor cells only if their concentrations in the tissue are above $\bar{v}_A$ and $\bar{v}_B$. A fraction, n, of proliferating cells die of natural causes and and a fraction, m, of proliferating cells become quiescent cells. The increase in proliferating cell population by

a	$0.500\ \text{day}^{-1}$	$\bar{v}_A$	10.000 [D]
m	$0.218\ \text{day}^{-1}$	$\bar{v}_B$	10.000 [D]
n	$0.477\ \text{day}^{-1}$	k_A	$8.400 \times 10^{-3}\ \text{day}^{-1}[D]^{-1}$
b	$0.100\ \text{day}^{-1}$	k_B	$8.400 \times 10^{-3}\ \text{day}^{-1}[D]^{-1}$
σ	$0.100\ \text{day}^{-1}$	K	1.0×10^{10} cells
γ_A	$0.100\ \text{day}^{-1}$	γ_B	$0.100\ \text{day}^{-1}$

Table 8.26: Chemotherapy Scheduling: Model Parameters

cell division is represented as another fraction, a, of the proliferating cell population. In addition, a fraction, b of quiescent cells become proliferating cells. The tumor cell dynamics are in (8.6.1) and (8.6.2). A logistic equation (8.6.3) governs the healthy cell population to ensure that the number of healthy cells does not exceed the carrying capacity, K. The drug concentrations in the tissue decrease with time according to (8.6.4) and (8.6.5). The initial cell populations and drug concentrations are in (8.6.6) and (8.6.7). Numerical values for the parameters are in Table 8.26. Most of the values are obtained from [31] where cell cycle specific chemotherapy with a single drug and without drug resistance is considered.

8.6.2 Dynamic Optimization Formulation

The goal is to kill as many tumor cells as possible during a 30-day chemotherapy session and still retain a minimum number of healthy cells at the end. The program:

$$\min_{\mathbf{p}\in\mathcal{P}} J(\mathbf{p}) = P(t_f, \mathbf{p}) + Q(t_f, \mathbf{p}) \tag{8.6.8}$$

$$\text{s.t. } Y(t_f, \mathbf{p}) \geq 1.0 \times 10^8, \tag{8.6.9}$$

$$v_A(t_f, \mathbf{p}) + v_B(t_f, \mathbf{p}) \leq 10.0, \tag{8.6.10}$$

$$0.0 \leq u_{A,k} \leq 20.0,\ k \in \mathcal{K},$$

$$0.0 \leq u_{B,k} \leq 20.0,\ k \in \mathcal{K},$$

is solved to determine such a schedule. Constraint (8.6.9) ensures that the size of final healthy cell population is above a certain size and constraint (8.6.10) ensures that the final drug concentration is at an acceptable level.

8.6.3 Nonsmooth Single Shooting Method Results

DSL48SE is the IVP solver ([108, 109, 36]) used to integrate the dynamics and the auxiliary equations to obtain an element of the linear Newton approximation. The auxiliary equations are obtained using automatic differentiation algorithms implemented in DAEPACK ([107]). The differential equations are integrated with an absolute tolerance of 1×10^{-8} and a relative tolerance of 1×10^{-7}.

The proximal bundle method in [64] is used to solve (8.6.8). A penalty approach to handle (8.6.9) and (8.6.10) is used because the algorithm in [64] handles only linear constraints on the decision variables. The objective of (8.6.8) is augmented with (8.6.9) and (8.6.10) to obtain the modified objective

$$P(t_f, \mathbf{p}) + Q(t_f, \mathbf{p}) + \mu \max(Y_{\min} - Y(t_f, \mathbf{p}), 0) + \mu \max(v_A(t_f, \mathbf{p}) + v_B(t_f, \mathbf{p}) - 10.0, 0)$$

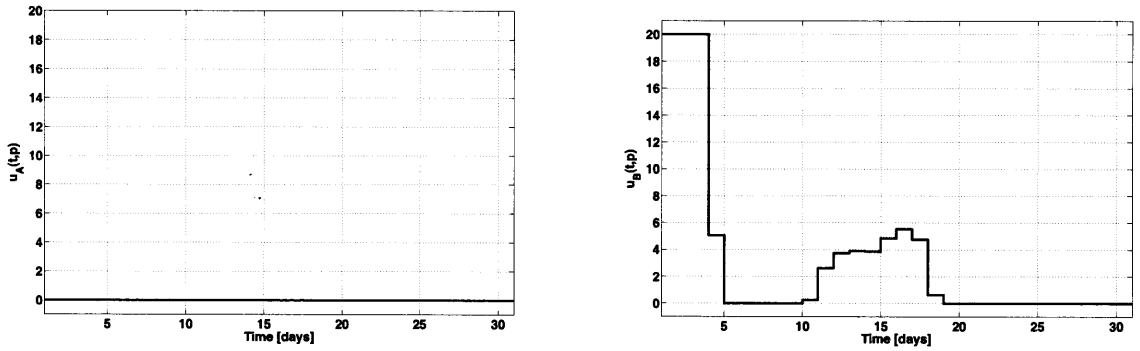
where μ is the penalty parameter. The modified program is successively solved two times with increasing penalty parameter to an optimality tolerance of 1×10^{-6}. The drug dosages are set to 2.0 initially. The penalty parameters used are $\{1000, 3000\}$. The total solution time was 40.0 seconds on a SUSE Linux 10 Virtual Machine with 1 GB of RAM and a 2.4 GHz Intel Core Duo CPU. The bundle solver required 71 iterations and 78 calls to the integrator.

The cell population numbers and constraint values at the beginning and end of the treatment are in Table 8.27. The tumor cell population is reduced to about 11 percent of its initial size. The drug schedules are shown in Figure 8-35a and Figure 8-35b. The preference

	Beginning of Treatment	End of Treatment
Y	1.00×10^{10} cells	1.00×10^{8} cells
Q	8.00×10^{11} cells	6.66×10^{10} cells
P	2.00×10^{11} cells	3.73×10^{10} cells
v_A	0.00 [D]	0.00 [D]
v_B	0.00 [D]	10.0 [D]

Table 8.27: Chemotherapy Schedule:Cell populations at the beginning and end of treatment

to use drug B is clearly seen. The effects of the drugs are proportional to the corresponding cell populations. Therefore using drug B results in more effective treatment as the population of quiescent cells is greater than that of proliferating cells. The drug concentrations are in 8-36a and 8-36b. The cell populations are in Figures 8-37a and 8-37b.

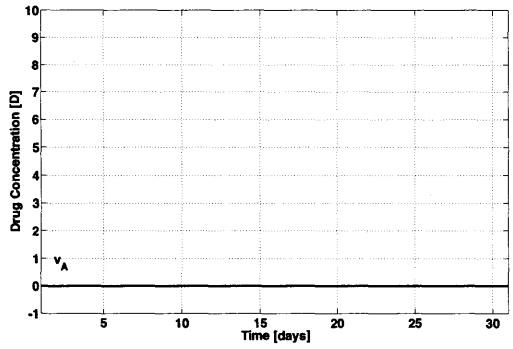


(a) Drug A (b) Drug B

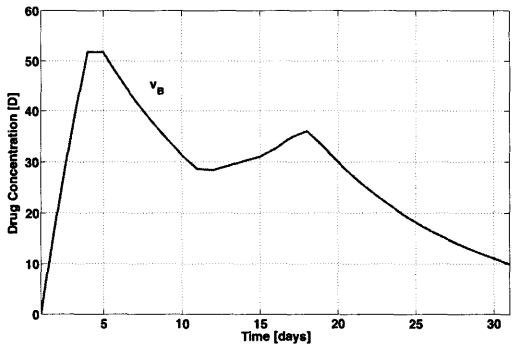
Figure 8-35: Chemotherapy Schedule: Optimal drug schedules.

8.6.4 Derivative Free Method Results

The Mesh Adaptive Direct Search Method [6] is used to solve program (8.6.8). This method does not require derivative information and unlike other derivative free methods there are some theoretical convergence results for locally Lipschitz functions. The algorithm is implemented in the software package NOMAD and the package can be found at *http: //www.gerad.ca /NOMAD.*

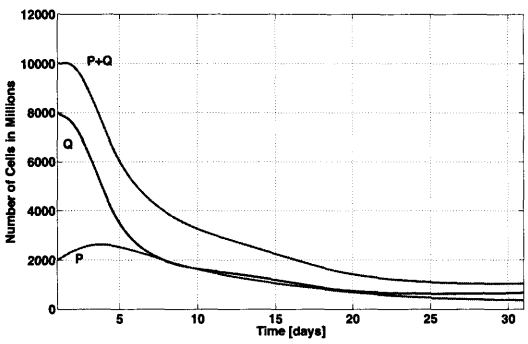


(a) Drug A

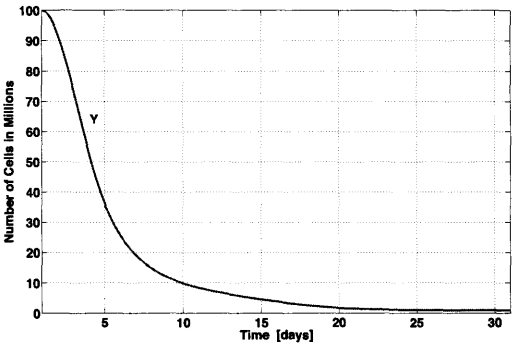


(b) Drug B

Figure 8-36: Chemotherapy Schedule: Drug concentrations in the tissue.



(a) Tumor Cells



(b) Healthy Cells

Figure 8-37: Chemotherapy Schedule: Cell populations.

Initially, the algorithm did not achieve significant progress towards a solution after 2000 evaluations of the objective and constraints with the initial drug dosages set to 2.0. Therefore the initial dosages were set to zero except the first 3 day dosages for drug B. These were set to 20. The algorithm was run for 7500 evaluations of the objective and constraints. The final tumor cell population obtained was 1.12×10^{11} cells, the final healthy cell population was 0.996×10^{8} cells and the final drug concentration in the healthy tissue was 10.19 [D]. The algorithm took 56 minutes to produce the results on a SUSE Linux 10 Virtual Machine with 1 GB of RAM and a 2.4 GHz Intel Core Duo CPU. The data is summarized in Table 8.28. NFV is the number of times the dynamics are simulated and CPU Time is the time

	Nonsmooth Single Shooting	Derivative Free
$Y(t_f, \mathbf{p}^*)$	1.00×10^8 cells	0.996×10^8 cells
$Q(t_f, \mathbf{p}^*) + P(t_f, \mathbf{p}^*)$	1.039×10^{11} cells	1.12×10^{11} cells
$v_A(t_f, \mathbf{p}^*) + v_B(t_f, \mathbf{p}^*)$	10.00 [D]	10.19 [D]
CPU Time [s]	40.0 s	3360.0 s
NFV	78	7500

Table 8.28: Chemotherapy Schedule: Comparison of Nonsmooth Single Shooting Method and Derivative Free Method

taken by the processor to solve the problem.

8.6.5 Conclusion

It was found that even though it is easier to set up and run the derivative free method, the solution time required to obtain an answer was significantly more. The nonsmooth shooting method took 40 seconds to terminate with a solution satisfying the stationarity condition where as the derivative free method took 56 minutes. The derivative-free method does not have a termination criteria based on stationarity. It terminates once the number of iterations or objective evaluations exceed their maximum values. In this case, the algorithm terminated once 2500 objective and constraint evaluations were made. The final solution provided by the derivative-free algorithm corresponds to a worse solution value than the one obtained using the bundle method.

From the performance of the derivative-free method, it is clear that derivative information should be used whenever available. Although, it is easier to set up a derivative-free method, the use of automatic differentiation tools makes the difference in the effort to set up the problem minimal. Therefore, for these problems, the use of derivative free methods is not warranted.

8.7 Notes

The performance of the nonsmooth single shooting method depends on the performance of the integration algorithm and the bundle solver. Currently, the integration routine DSL48SE ([108, 109, 36]) is the only available routine that incorporates state event location algorithms and the necessary infrastructure to integrate the sensivity equations. This integration routine uses sparse matrix algebra techniques to effciently handle systems with a lot states. However, the use of sparse matrix algebra techniques incurs setup costs. These costs are offset by the gain in efficiency obtained when dealing with systems with a lot states. For the case studies in this chapter, DSL48SE may not be the most efficient algorithm because the number of states of the systems considered in these studies is relatively small.

Chapter 9

Conclusions and Future Directions for Research

In this thesis, the nonsmooth single shooting method, an open loop dynamic optimization method for a class of systems with varying structure is developed. Unlike the state-of-the-art methods, this method does not explicitly enumerate the hybrid mode trajectories and it does not discretize the dynamics as a part of the optimization formulation. Instead a specialized and efficient numerical integration algorithm [108] is used to compute the continuous state trajectories accurate within integration tolerances.

The method converts the dynamic optimization problem into a nonlinear program by parametrizing the controls. The resultant program is a nonsmooth optimization problem due to the varying structure of the underlying dynamic systems. Therefore concepts from nonsmooth analysis and methods from nonsmooth optimization are used. The main challenge of this approach is determination of the replacement for the gradient. A custom set-valued map is defined using the generalized Jacobian [25] which turns out to be a linear Newton approximation [35]. Sensitivity initial value problems are derived to calculate an element of this set-valued map.

Stationarity conditions for optimization are defined in terms of this set-valued map. It is shown that bundle methods can be used to obtain solutions satisfying these stationarity conditions.

The performance of the nonsmooth single shooting method is compared to the state-of-the-art methods. The nonsmooth shooting method provides more accurate answers for equal or less effort than the state of the art in case the system dynamics are highly nonlinear and/or exhibit stiffness. This is the result of using a numerical integration algorithm instead of discretization as a part of the optimization formulation. An empirical investigation of complexity is performed. The results strongly suggest that the method scales polynomially with the number of states and parameters.

Finally, the thesis demonstrates that nonsmooth analysis and nonsmooth optimization methods can be used to solve practical dynamic optimizaton problems.

The breakdown of the contributions per chapter is:

Chapter 3 : In this chapter, sufficient conditions for the existence of the strict derivative (§2.2) of the map $\boldsymbol{\eta} \mapsto \mathbf{x}(t, \boldsymbol{\eta})$ in terms of the generalized Jacobian are derived where $\mathbf{x}$ represents the continuous states of the system with varying structure.

The first part considers dynamics described by the ordinary differential equations in (3.2.1). The forward and adjoint sensitivity initial value problems are derived to compute the aforementioned derivative. The second part extends these results to differential-algebraic systems described by (3.3.1) using the implicit function theorem for locally Lipschitz continuous functions. The strict derivatives of the maps $\boldsymbol{\eta} \mapsto \mathbf{y}(t, \boldsymbol{\eta})$ and $\boldsymbol{\eta} \mapsto \dot{\mathbf{x}}(t, \boldsymbol{\eta})$ are obtained where $\mathbf{y}$ are the algebraic variables. The results are extended to multistage systems.

The parametric sensitivity results in this chapter are new. The sufficiency conditions derived in this chapter are more general than those in [44]. They are more general than the conditions in [39, 95] in case the underlying dynamics are described by locally

Lipschitz continuous ODEs and DAEs.

Chapter 4 : It is not possible to compute the generalized Jacobian of the map $\boldsymbol{\eta} \mapsto \mathbf{x}(t, \boldsymbol{\eta})$ for all possible values of $\boldsymbol{\eta}$. Theorem 3.1.3 does not provide a means to compute the generalized Jacobian in case the sufficiency conditions in Chapter 3 are not satisfied. Therefore, a linear Newton approximation of the map $\boldsymbol{\eta} \mapsto \mathbf{x}(t, \boldsymbol{\eta})$ is derived under an additional semismoothness assumption on the right-hand side functions. Formulae are derived to compute an element of this linear Newton approximation using forward and reverse integration in time. Linear Newton approximations of the maps $\boldsymbol{\eta} \mapsto \mathbf{y}(t, \boldsymbol{\eta})$ and $\boldsymbol{\eta} \mapsto \dot{\mathbf{x}}(t, \boldsymbol{\eta})$ are derived. Results are extended to the multistage case. The values of the linear Newton approximations defined contain the values of the generalized Jacobians at all possible $\boldsymbol{\eta}$. The results reduce to the results in Chapter 3 in case the assumptions of that chapter hold in addition to the assumptions of this chapter.

The parametric sensitivity results based on linear Newton approximations are new.

Chapter 5 : The differential equations defining elements of the linear Newton approximations in Chapter 4 generally have right-hand sides that are discontinuous in time. The time at which a discontinuity occurs needs to be detected and located for efficient and accurate computation of these quantities using numerical integration. A numerical method is described to detect these discontinuities using the state event location algorithm in [83] and compute an element of the linear Newton approximations defined in Chapter 4 simultaneously with the states. This algorithm works if the functions of the right-hand side satisfy a structural assumption that in essence makes them PC^1 (§2.7) functions.

The numerical computation of linear Newton approximations using state event location algorithms is new. Note that a method based on time stepping [99] is described in [81].

Chapter 6 : In this chapter, bundle methods [54, 66] are modified to use the linear Newton approximations defined in Chapter 4. Extended stationarity conditions are defined

using these linear Newton approximations and it is shown that the bundle method produces a sequence of solutions whose limit points satisfy the extended stationary conditions. It is shown that a direction of descent can be computed and that the special line search algorithm of the bundle method converges if linear Newton approximations are substituted for the generalized gradients. In essence, it is shown that the generalized gradient can be replaced by the linear Newton approximations of Chapter 4.

The use of the linear Newton approximation in the context of nonsmooth optimization is new. The use of linear Newton approximations in conjunction with bundle methods is new.

Chapter 7 : The theoretical development of the nonsmooth single shooting method is in this chapter. The control parametrization approach in [105, 40] is extended to the case where the dynamics are governed by ordinary differential equations whose right-hand sides are PC^1 functions.

Chapter 8 : The performances and accuracy of solutions of the MILP, MPEC and nonsmooth single shooting methods are compared using literature examples. This comparison is the first of its kind to the best of the author's knowledge. It is found that the nonsmooth single shooting method provides the most accurate optimal state trajectories for less or comparable effort especially if the dynamics are highly nonlinear and/or stiff.

Empirical complexity analysis of the nonsmooth shooting method is performed. Currently, it is not possible to carry out a theoretical complexity analysis because theoretical complexity analysis results do not exist for bundle methods and numerical integration methods. The results strongly suggest that the method scales polynomially with the states and parameters.

9.1 Future Directions for Research

9.1.1 Parametric Sensitivities, Their Computation and Use

Existence, computation and use of the second derivative

The second derivative of the objective and the constraints in (1.3.1) with respect to the parameters is of practical interest. The second derivative can be used to improve the performance of bundle methods. In [67], it is shown that a bundle method using second derivative information [64, 63] outperforms the proximal bundle method in [64] significantly. Under certain conditions, superlinear convergence can be proven for this method.

There are results on the existence of the second derivative for systems with varying structure [2]. These results depend on conditions similar to those in [39]. It is an open question whether more general results can be achieved for piecewise twice continuously differentiable vector fields.

Computing the second derivative requires additional computational effort. Therefore, the efficient computation of the second derivative simultaneously with the first derivative is important.[1] It is an open question whether the advantages of using second derivatives offsets the additional computational burden.

In [67], the second derivative of a nearby point is used as an approximation in case the second derivative does not exist at a point. This might be a computationally expensive option for dynamic optimization problems. The existence of a suitable replacement is another open question.

[1] It is shown in [77] that directional second order derivatives are relatively cheap to compute.

Efficient implementation of reverse integration to compute parametric sensitivities of systems with varying structure

There are dynamic optimization problems where the number of parameters is very large. In this case, the linear Newton approximations can possibly be computed more efficiently using reverse integration in time. A smaller number of equations need to be integrated. However, reverse integration in time requires the storage of the state trajectories obtained using forward integration.

Currently, numerical methods exist that are applicable to sufficiently smooth ODEs and DAEs [85]. There exists no numerical method that uses the results in this thesis and [95]. The theoretical development and implementation of such a numerical method and the comparison of reverse integration to forward integration is another future direction of research.

Parametric sensitivities of linear program solutions with respect to the right-hand side vector

Consider the linear program:

$$\min_{\mathbf{x}} \mathbf{c}^{\mathrm{T}}\mathbf{x} \text{ s.t. } \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0}.$$

where $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{c} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $n > m$. Assume $\mathbf{b} \in \mathcal{B}$ where $\mathcal{B}$ is an open convex set. Assume that the solution set is a singleton for all $\mathbf{b} \in \mathcal{B}$. Let $\mathbf{x}^* : \mathcal{B} \to \mathbb{R}^n$ represent the optimal solution of this linear program as a function of $\mathbf{b}$. It is known that $\mathbf{x}^*$ is a locally Lipschitz continuous function in this case [69].

Dynamic systems with linear programs embedded occur when biological agents are modeled using the *flux balance analysis* (FBA) technique [58]. In this approach, the $\mathbf{A}$ matrix of the linear program encodes the chemical reactions that take place inside biological agents such as bacteria or yeast. $\mathbf{x}$ represents the rate at which each of the chemical species is

produced. Note that there are more chemical species than linearly independent reactions. The $\mathbf{b}$ vector represents the amount of material exchange between the organism and the environment. Usually, the objective is maximization of the growth of the organism. Depending on the $\mathbf{b}$ vector, the chemical reactions that occur inside the organism change. An example of a dynamic optimization problem with FBA models embedded can be found in [46].

The semismoothness property and the generalized Jacobian of the map $\boldsymbol{\eta} \mapsto \mathbf{x}^*(\boldsymbol{\eta})$ is of interest. If an element of the generalized Jacobian of the map $\boldsymbol{\eta} \mapsto \mathbf{x}^*(\boldsymbol{\eta})$ can be computed, then, the results in this thesis can be used to solve such problems. Currently, these problems are solved using the MPEC approach. Due to the discretization of the dynamics, the size of the FBA models used is limited, though.

9.1.2 Dynamic Optimization

Integer-valued controls

In [97], a method is described to handle integer-valued controls in continuous-time dynamic optimization problems. The method relaxes the original problem with respect to the integer-valued controls. The controls of the relaxed problem are approximated by functions described by finitely many parameters. The approximate integer-valued controls are recovered using special rounding off techniques from the solutions of the relaxed problems. The method is shown to approximate the solution of the original problem arbitrarily close. The systems considered do not contain autonomous transitions and for fixed integer-valued controls are continuously differentiable. Combining the nonsmooth single shooting method with the approach in [97] would enable the incorporation of integer-valued controls.

Nonsmooth multiple shooting method

Single shooting methods are not suitable for problems with unstable dynamics. In this case, integration errors grow without bound and the state trajectories computed are not reliable. In order to deal with such systems, multiple shooting methods have been devised [59, 60]. Multiple shooting methods can handle dynamic optimization problems of unstable systems with end point constraints.

The results in this thesis can be used to develop a multiple shooting method. In multiple shooting, the time horizon is partitioned into epochs. The dynamics are integrated separately on each epoch. This decoupling is achieved by making the initial conditions for each epoch, parameters of the dynamic optimization problem. Then, consistent parameters are obtained as a part of the solution. The challenge is to handle these additional parameters efficiently in the solution of the problem.

Optimization of convex programs

$\partial f(\mathbf{x}_2) = W$ holds in Theorem 3.1.2 if $g(t,\cdot)$ is a convex function (See Theorem 2.7.2 in [25]). In this case, W is equal to the subdifferential of f at $\mathbf{x}_2$. Consider the problem

$$\begin{aligned} \min_{\mathbf{p}\in\mathcal{P}} \; J(\mathbf{p}) = & \int_{t_0}^{t_f} h_0(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}))dt + H_0(\mathbf{p},\mathbf{x}(t_f,\mathbf{p})) \\ \text{s.t.} & \int_{t_0}^{t_f} h_i(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}))dt + H_i(\mathbf{p},\mathbf{x}(t_f,\mathbf{p})) \leq 0, \; \forall i \in \{1,\ldots,n_c\}, \\ & \dot{\mathbf{x}}(t,\mathbf{p}) = \mathbf{f}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p})), \; \forall t \in (t_0,t_f], \\ & \mathbf{x}(t_0,\mathbf{p}) = \mathbf{f}_0(\mathbf{p}). \end{aligned} \tag{9.1.1}$$

If the integrands in the objective and constraints can be shown to be convex functions of $\mathbf{p}$, then Theorem 2.7.2 in [25] can be used to compute an element of the subdifferential of the corresponding integrals. Cases when this holds are of interest because if H_i, for all

$i \in \{0, \ldots, n_c\}$ are convex functions, then this program is a convex program. In this case, bundle methods can be used to find the global minimum. The challenge is the computation of an element of the generalized gradients of the integrands with respect to $\mathbf{p}$ when assumptions of Theorem 3.2.3 do not hold.

Deterministic global optimization

The nonsmooth shooting method finds stationary points of the nonlinear programs. Problems involving few parameters may be solved to $\epsilon-$global optimally using the approach in [57, 56]. In this approach, convex nonlinear programs are constructed that underestimate the objective value of the original nonlinear program. Then, a deterministic global optimization method such as branch and bound [47] is used to obtain the $\epsilon-$global solution.

The construction of the underestimating convex programs need to be investigated using the ideas in [100]. Note that this research direction is linked to the optimization of convex programs.

Necessary conditions of optimality

The relationship between the solutions of the nonsmooth NLPs representing dynamic optimization problems and the necessary conditions of optimality in [25] are of interest.

9.1.3 Systems with Discontinuous Vector Fields

The extension of the nonsmooth single shooting method to discontinuous vector fields requires research in several fields. Dynamic optimization problems involving these systems may be discontinuous programs. Even if they are not discontinuous, they may not be locally Lipschitz continuous programs anymore.

Example 9.1.1 (Discontinuous Vector Field Example). Consider the dynamic system

$$g(p, x(t,p)) = x(t,p)^3 - 5x(t,p)^2 + 7x(t,p) - p,$$

$$\dot{x}(t,p) = \begin{cases} 4 - x(t,p) & \text{if } g(p, x(t,p)) \leq 0, \text{ Mode 1} \\ 0.7x(t,p) & \text{if } g(p, x(t,p)) \geq 0, \text{ Mode 2} \end{cases}, \quad \forall t \in (0, 3.0],$$

$$x(0,p) = -3.0, \; p \in [0, 7.5].$$

The vector field of this system is discontinuous at times when $g(p, x(t,p)) = 0$ holds. Note that the number of real roots of the polynomial depends on p. This dependence determines the number of transitions that occur during the evolution of the system. Figure 9-1 shows the transition times as a function of p. The system experiences up to three transitions during the time interval $[0, 3.0]$. If $p < 2$, only one transition occurs. The number of transitions eventually becomes three for $p \in [2, 3)$. At $p = 3$, two consecutive transitions occur instantaneously at around 0.75 s. The transversality condition is violated at this double transition. The number of transitions drops to one afterwards. Figure 9-2 depicts the

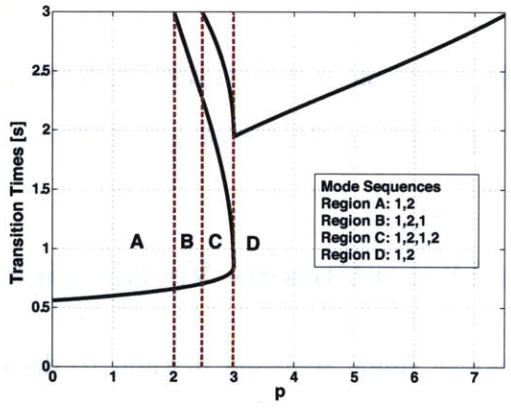


Figure 9-1: Transition times as a function of p.

dependence of the final state on the parameters. The nonsmoothness at $p = 3$ is obvious.

However, there exist 2 more points of nonsmoothness. These points correspond to the parameter values where the hybrid mode trajectory changes. The state is not a locally Lipschitz function at $p = 3$. This can be seen from the data in Figure 9-3. In Figure 9-3, the natural logarithm of the magnitude of the derivative of the mapping $\boldsymbol{\eta} \mapsto \mathbf{x}(3, \boldsymbol{\eta})$ is plotted. Note that the derivative is calculated using the results in [39]. The derivative is discontinuous whenever the hybrid mode trajectory changes. At $p = 3$, the limit from the left tends to infinity. This occurs because there is a division by zero in the computation of the sensitivities whenever the transversality condition does not hold. Initial tentative

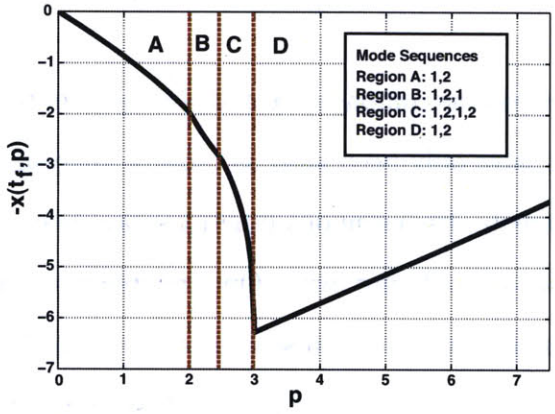


Figure 9-2: $-x(3.0, p)$ as a function of p.

results on sufficient conditions that guarantee local Lipschitz continuity and semismoothness when the dynamics are discontinuous can be found in [115].

In order to extend the results to systems with discontinuous dynamics requires advances in several fields:

1. Nonsmooth Analysis: The main issue is to determine a replacement for the generalized gradient/Jacobian and the linear Newton approximation. There are extensions of the generalized gradient to functions that are not locally Lipschitz continuous [25, 92]. Unlike the locally Lipschitz case, these set-valued maps can have empty images and

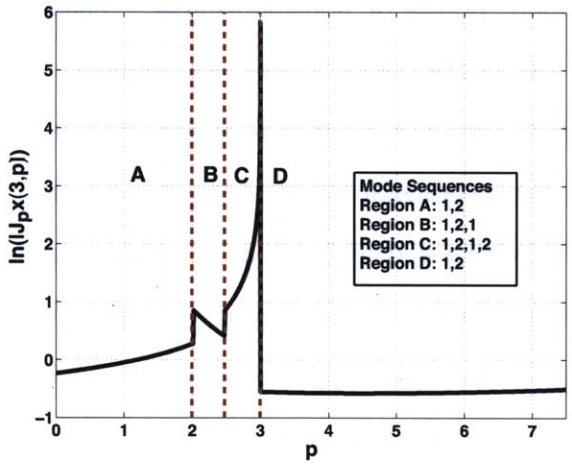


Figure 9-3: $\ln |J_p x(3.0, p)|$ as a function of p.

crucial properties such as compactness and convexity of the images are not guaranteed. These extensions have only been used as tools of analysis. Results pertaining to them are few. Extensions to discontinuous functions have been proposed [34]. Extensions of the generalized Jacobian do not exist. On a related note, functions that are not locally Lipschitz continuous have not been studied as extensively. It is not clear what classes of functions to expect when dealing with such systems. Experience suggests that the functions of interest are those that are continuously differentiable on open sets, but probably discontinuous or nonsmooth on the boundaries of these sets [22].

2. Nonsmooth Optimization: For efficient nonsmooth optimization, the use of a bundle to approximate the set-valued maps replacing the gradient appears to be necessary [22]. Bundle methods require semismoothness to operate efficiently. Replacement conditions need to be developed. Necessary conditions of optimality for the non-Lipschitz and discontinuous function cases exist [92, 33]. The practical applicability of these conditions need to be analyzed.
3. Parametric Sensitivity Analysis: The existence of auxiliary equations of the form developed in Chapters 3 and 4 needs to be determined.

Note that one can find examples in the literature where the MILP and MPEC approaches have been applied to such systems in conjunction with discretization of the dynamics. Currently, there exists no theoretical support that the approximations of the state trajectories converge to the state trajectories as the discretization gets finer grained.

Appendix A

Convergence Proof of the Modified Bundle Method

In this part, the convergence proof of the method described in Chapter 6 is summarized. The proof is very similar to the proof of convergence of Algorithm 3.1 in Chapter 6 in [54]. The main difference is that the generalized gradient is replaced with linear Newton approximations satisfying Assumption 6.2.1.

A.1 Convergence Proof

The following lemma proves an important property of the aggregate quantities.

Lemma A.1.1. *Let* $\mathbf{v}_k^f, f_{k+1}^v, s_{k+1}^f, \mathbf{v}_k^G, G_{k+1}^v, s_{k+1}^G$ *be as defined in §6.3.1. Then*

$$(\mathbf{v}_k^f, f_{k+1}^v, s_{k+1}^f) \in \text{conv}\left(\{(\boldsymbol{\zeta}_j^f, f_{k+1,j}, s_{k+1,j}) : j = 1, \ldots, k\}\right), \tag{A.1.1}$$

$$(\mathbf{v}_k^G, G_{k+1}^v, s_{k+1}^G) \in \text{conv}\left(\{(\boldsymbol{\zeta}_j^G, G_{k+1,j}, s_{k+1,j}) : j = 1, \ldots, k\}\right)$$

hold.

Proof. Assume that (A.1.1) holds for $k-1$. Let $\boldsymbol{\lambda}_k$, $\boldsymbol{\mu}_k$, λ_k^v and μ_k^v be the solutions of (6.3.22). Then

$$\mathbf{v}_k^f = \sum_{j \in M_k^f} \tilde{\lambda}_{k,j} \boldsymbol{\zeta}_j^f + \tilde{\lambda}_k^v \mathbf{v}_{k-1}^f, \ \mathbf{v}_k^G = \sum_{j \in M_k^G} \tilde{\mu}_{k,j} \boldsymbol{\zeta}_j^G + \tilde{\mu}_k^v \mathbf{v}_{k-1}^G.$$

Note that

$$\sum_{j \in M_k^f} \tilde{\lambda}_{k,j} + \tilde{\lambda}_k^v = 1, \ \sum_{j \in M_k^G} \tilde{\mu}_{k,j} + \tilde{\mu}_k^v = 1.$$

Hence $\mathbf{v}_k^f \in \mathrm{conv}\left(\{\boldsymbol{\zeta}_j^f : j \in M_k^f\}, \{\mathbf{v}_{k-1}^f\}\right)$ and $\mathbf{v}_k^G \in \mathrm{conv}\left(\{\boldsymbol{\zeta}_j^G : j \in M_k^G\}, \{\mathbf{v}_{k-1}^G\}\right)$. Since $\mathbf{v}_{k-1}^f \in \mathrm{conv}\left(\{\boldsymbol{\zeta}_j^f : j \in 1, \ldots, k-1\}\right)$ and $\mathbf{v}_{k-1}^G \in \mathrm{conv}\left(\{\boldsymbol{\zeta}_j^G : j \in 1, \ldots, k-1\}\right)$, $\mathbf{v}_k^f \in \mathrm{conv}\left(\{\boldsymbol{\zeta}_j^f : j \in 1, \ldots, k\}\right)$ and $\mathbf{v}_k^G \in \mathrm{conv}\left(\{\boldsymbol{\zeta}_j^G : j \in 1, \ldots, k\}\right)$ hold.

By definition,

$$\begin{aligned}
f_{k+1}^v &= \sum_{j \in M_k^f} \tilde{\lambda}_{k,j} f_{k,j} + \tilde{\lambda}_k^v f_k^v + \langle \mathbf{v}_k^f, \mathbf{p}_{k+1} - \mathbf{p}_k \rangle, \\
f_{k+1}^v &= \sum_{j \in M_k^f} \tilde{\lambda}_{k,j} f_{k,j} + \tilde{\lambda}_k^v f_k^v + \langle \sum_{j \in M_k^f} \tilde{\lambda}_{k,j} \boldsymbol{\zeta}_j^f + \tilde{\lambda}_k^v \mathbf{v}_{k-1}^f, \mathbf{p}_{k+1} - \mathbf{p}_k \rangle, \\
f_{k+1}^v &= \sum_{j \in M_k^f} \tilde{\lambda}_{k,j} (f_{k,j} + \langle \boldsymbol{\zeta}_j^f, \mathbf{p}_{k+1} - \mathbf{p}_k \rangle) + \tilde{\lambda}_k^v (f_k^v + \langle \mathbf{v}_{k-1}^f, \mathbf{p}_{k+1} - \mathbf{p}_k \rangle). \\
f_{k+1}^v &= \sum_{j \in M_k^f} \tilde{\lambda}_{k,j} f_{k+1,j} + \tilde{\lambda}_k^v (f_k^v + \langle \mathbf{v}_{k-1}^f, \mathbf{p}_{k+1} - \mathbf{p}_k \rangle).
\end{aligned}$$

Since $f_k^v \in \mathrm{conv}\left(\{f_{k,j} : 1, \ldots, k-1\}\right)$, $f_{k+1}^v \in \mathrm{conv}\left(\{f_{k,j} : 1, \ldots, k\}\right)$ holds. Similarly $G_{k+1}^v \in \mathrm{conv}\left(\{G_{k,j} : 1, \ldots, k\}\right)$ holds.

s_{k+1}^f is obtained as follows:

$$s_{k+1}^f = \tilde{s}_k^f + \|\mathbf{p}_{k-1} - \mathbf{p}_k\|,$$

$$s_{k+1}^f = \sum_{j \in M_k^f} \tilde{\lambda}_{k,j} s_{k,j}^f + \tilde{\lambda}_k^v s_k^f + \|\mathbf{p}_{k-1} - \mathbf{p}_k\|,$$

$$s_{k+1}^f = \sum_{j \in M_k^f} \tilde{\lambda}_{k,j} (s_{k,j}^f + \|\mathbf{p}_{k-1} - \mathbf{p}_k\|) + \tilde{\lambda}_k^v (s_k^f + \|\mathbf{p}_{k-1} - \mathbf{p}_k\|)$$

$$s_{k+1}^f = \sum_{j \in M_k^f} \tilde{\lambda}_{k,j} (s_{k+1,j}^f) + \tilde{\lambda}_k^v (s_k^f + \|\mathbf{p}_{k-1} - \mathbf{p}_k\|).$$

Since $s_k^f \in \text{conv}(\{s_{k,j} : 1, \ldots, k-1\})$, $s_{k+1}^f \in \text{conv}(\{s_{k,j} : 1, \ldots, k\})$ holds. Similarly $s_{k+1}^G \in \text{conv}(\{s_{k,j} : 1, \ldots, k\})$ holds.

As a result (A.1.1) holds if (A.1.1) holds for $k-1$. In order to complete the inductive proof, (A.1.1) needs to hold for $k = 1$.

Let $k = 1$. Then $\mathbf{v}_1^f = \boldsymbol{\zeta}_1^f$ and $\mathbf{v}_1^G = \boldsymbol{\zeta}_1^G$. $f_1^v = f_{1,1}$, $G_1^v = G_{1,1}$, $\tilde{s}_1^f = s_{1,1}$, $\tilde{s}_1^G = s_{1,1}$. Note that

$$f_2^v = f_{1,1} + \langle \boldsymbol{\zeta}_1^f, \mathbf{p}_2 - \mathbf{p}_1 \rangle = f_{2,1},$$

$$G_2^v = G_{1,1} + \langle \boldsymbol{\zeta}_1^G, \mathbf{p}_2 - \mathbf{p}_1 \rangle = G_{2,1},$$

$$s_2^f = s_{1,1} + \|\mathbf{p}_2 - \mathbf{p}_1\| = s_{2,1},$$

$$s_2^G = s_{1,1} + \|\mathbf{p}_2 - \mathbf{p}_1\| = s_{2,1}.$$

Hence (A.1.1) holds for $k = 1$. □

Definition A.1.2. *Let* $k_r = \max\{j : j \leq k,\ \lambda_j^v = \mu_j^v = 0\}$ *and* $\hat{M}_k^f = \{j : k_r < j \leq k\} \cup M_{k_r}^f$. $\hat{M}_k^G = \{j : k_r < j \leq k\} \cup M_{k_r}^G$.

The next lemma shows that the aggregate quantities are in the convex hull of the linear Newton approximations, locality measures and linearization values computed since the last reset.

Lemma A.1.3. *Let* $k \geq 1$ *and assume that the bundle method did not terminate before the*

kth iteration. Then, there exists numbers $\hat{\lambda}_j$ *and* $\hat{\mu}_j$ *satisfying*

$$(\mathbf{v}_k^f, \tilde{f}_k^v, \tilde{s}_k^f) = \sum_{j \in \hat{M}_k^f} \hat{\lambda}_j (\boldsymbol{\zeta}_j^f, f_{k,j}, s_{k,j}), \tag{A.1.2}$$

$$\hat{\lambda}_j \geq 0, \sum_{j \in \hat{M}_k^f} \hat{\lambda}_j = 1, \tag{A.1.3}$$

$$a_k = \max\{s_{k,j} : j \in \hat{M}_k^f\} \tag{A.1.4}$$

$$\|\mathbf{p}_k - \mathbf{y}_j\| \leq a_k \leq \bar{a}, \ \forall j \in \hat{M}_k^f, \tag{A.1.5}$$

and

$$(\mathbf{v}_k^G, \tilde{G}_k^v, \tilde{s}_k^G) = \sum_{j \in \hat{M}_k^G} \hat{\mu}_j (\boldsymbol{\zeta}_j^G, G_{k,j}, s_{k,j}), \tag{A.1.6}$$

$$\hat{\mu}_k \geq 0, \sum_{j \in \hat{M}_k^f} \hat{\mu}_k = 1, \tag{A.1.7}$$

$$a_k = \max\{s_{k,j} : j \in \hat{M}_k^G\} \tag{A.1.8}$$

$$\|\mathbf{p}_k - \mathbf{y}_j\| \leq a_k \leq \bar{a}, \ \forall j \in \hat{M}_k^G. \tag{A.1.9}$$

Proof. The proof for the results in (A.1.2), (A.1.3), (A.1.4) and (A.1.5) will be given. The proof of for the results in (A.1.6), (A.1.7),(A.1.8) and (A.1.9) follows the same reasoning.

Note that $M_k^f \subset \hat{M}_k^f$. Let $\tilde{\lambda}_{k,j} = 0, \ \forall j \in \hat{M}_k^f \backslash M_k^f$. Then

$$(\mathbf{v}_k^f, \tilde{f}_k^v, \tilde{s}_k^f) = \sum_{j \in \hat{M}_k^f} \tilde{\lambda}_{k,j} (\boldsymbol{\zeta}_j^f, f_{k,j}, s_{k,j}) + \tilde{\lambda}_k^v (\mathbf{v}_{k-1}^f, f_k^v, s_k^f), \tag{A.1.10}$$

holds. If $\lambda_k^v = 0$, then $M_k^f = \hat{M}_k^f$ and (A.1.2), (A.1.3), (A.1.4) and (A.1.5) hold trivially with $\hat{\lambda}_k = \tilde{\lambda}_k$. Since $r_1^a = 0$, $\lambda_1^v = 0$ and (A.1.2), (A.1.3), (A.1.4) and (A.1.5) hold for $k = 1$.

Assume that (A.1.2), (A.1.3), (A.1.4) and (A.1.5) hold for $k = \bar{k}$. Assume that $\tilde{\lambda}_{\bar{k}+1}^v > 0$.

Observe that $\hat{M}^f_{\bar{k}+1} = \hat{M}^f_{\bar{k}} \cup \{\bar{k}+1\}$. and

$$(\mathbf{v}^f_{\bar{k}}, \tilde{f}^v_{\bar{k}}, \tilde{s}^f_{\bar{k}}) = \sum_{j \in \hat{M}^f_{\bar{k}}} \hat{\lambda}_j (\boldsymbol{\zeta}^f_j, f_{\bar{k},j}, s_{\bar{k},j}).$$

The following can be shown to hold using the same reasoning in Lemma A.1.1:

$$(\mathbf{v}^f_{\bar{k}}, f^v_{\bar{k}+1}, s^f_{\bar{k}+1}) = \sum_{j \in \hat{M}^f_{\bar{k}}} \hat{\lambda}_j (\boldsymbol{\zeta}^f_j, f_{\bar{k}+1,j}, s_{\bar{k}+1,j}).$$

Note that

$$(\mathbf{v}^f_{\bar{k}+1}, \tilde{f}^v_{\bar{k}+1}, \tilde{s}^f_{\bar{k}+1}) = \sum_{j \in \hat{M}^f_{\bar{k}+1}} \tilde{\lambda}_{\bar{k}+1,j} (\boldsymbol{\zeta}^f_j, f_{\bar{k}+1,j}, s_{\bar{k}+1,j}) + \tilde{\lambda}^v_{\bar{k}+1} (\mathbf{v}^f_{\bar{k}}, f^v_{\bar{k}+1}, s^f_{\bar{k}+1}),$$

$$(\mathbf{v}^f_{\bar{k}+1}, \tilde{f}^v_{\bar{k}+1}, \tilde{s}^f_{\bar{k}+1}) = \sum_{j \in \hat{M}^f_{\bar{k}+1}} \tilde{\lambda}_{\bar{k}+1,j} (\boldsymbol{\zeta}^f_j, f_{\bar{k}+1,j}, s_{\bar{k}+1,j}) + \tilde{\lambda}^v_{\bar{k}+1} \sum_{j \in \hat{M}^f_{\bar{k}}} \hat{\lambda}_j (\boldsymbol{\zeta}^f_j, f_{\bar{k}+1,j}, s_{\bar{k}+1,j}).$$

Hence (A.1.2), (A.1.4), and (A.1.3) holds for $\bar{k}+1$. □

The following lemma uses Carathéodory's Theorem (Theorem 2.4.5) to keep the sizes of $\hat{M}^f_k$ and $\hat{M}^G_k$ less than $n_p + 3$.

Lemma A.1.4. *Let $k \geq 1$ and assume that the bundle method did not terminate before the kth iteration. Let $m = n_p + 3$. Then there exist two sets of nonnegative scalars $\{\hat{\lambda}_i\}_{i=1}^m$, $\{\hat{\mu}_i\}_{i=1}^m$, and two sets of vectors whose elements are not necessarily unique, $\{(\bar{\mathbf{y}}^f_{k,i}, \bar{\boldsymbol{\zeta}}^f_{k,i}, \bar{f}_{k,i}, \bar{s}^f_{k,i})\}_{i=1}^m \subset \mathbb{R}^{n_p} \times \mathbb{R}^{n_p} \times \mathbb{R} \times \mathbb{R}$, $\{(\bar{\mathbf{y}}^G_{k,i}, \bar{\boldsymbol{\zeta}}^G_{k,i}, \bar{G}_{k,i}, \bar{s}^G_{k,i})\}_{i=1}^m \subset \mathbb{R}^{n_p} \times \mathbb{R}^{n_p} \times \mathbb{R} \times \mathbb{R}$ such that*

$$(\mathbf{v}^f_k, \tilde{f}^v_k, \tilde{s}^f_k) = \sum_{i=1}^m \hat{\lambda}_i (\bar{\boldsymbol{\zeta}}^f_{k,i}, \bar{f}_{k,i}, \bar{s}^f_{k,i}), \ \sum_{i=1}^m \hat{\lambda}_i = 1,$$

$$\bar{\boldsymbol{\zeta}}^f_{k,i} \in \Gamma f(\bar{\mathbf{y}}^f_{k,i}), \ \forall i \in \{1, \ldots, m\},$$

$$(\bar{\mathbf{y}}^f_{k,i}, \bar{\boldsymbol{\zeta}}^f_{k,i}, \bar{f}_{k,i}, \bar{s}^f_{k,i}) \in \{(\mathbf{y}_j, \boldsymbol{\zeta}^f_j, f_{k,j}, s_{k,j}) : \ j \in \{1, \ldots, k\}\}, \ \forall i \in \{1, \ldots, m\},$$

$$\|\bar{\mathbf{y}}_{k,i}^f - \mathbf{p}_k\| \leq \bar{s}_{k,i}^f,\ \forall i \in \{1,\ldots,m\},$$
$$\max\{\bar{s}_{k,i}^f :\ i \in \{1,\ldots,m\}\} \leq a_k \leq \bar{a},$$

and

$$(\mathbf{v}_k^G, \tilde{G}_k^v, \tilde{s}_k^G) = \sum_{i=1}^m \hat{\lambda}_i(\bar{\boldsymbol{\zeta}}_{k,i}^G, \bar{G}_{k,i}, \bar{s}_{k,i}^G),\ \sum_{i=1}^m \hat{\mu}_i = 1,$$
$$\bar{\boldsymbol{\zeta}}_{k,i}^G \in \Gamma G(\bar{\mathbf{y}}_{k,i}^G),\ \forall i \in \{1,\ldots,m\},$$
$$(\bar{\mathbf{y}}_{k,i}^G, \bar{\boldsymbol{\zeta}}_{k,i}^G, \bar{G}_{k,i}, \bar{s}_{k,i}^G) \in \{(\mathbf{y}_j, \boldsymbol{\zeta}_j^G, G_{k,j}, s_{k,j}) :\ j \in \{1,\ldots,k\}\},\ \forall i \in \{1,\ldots,m\},$$
$$\|\bar{\mathbf{y}}_{k,i}^G - \mathbf{p}_k\| \leq \bar{s}_{k,i}^G,\ \forall i \in \{1,\ldots,m\},$$
$$\max\{\bar{s}_{k,i}^G :\ i \in \{1,\ldots,m\}\} \leq a_k \leq \bar{a},$$

Proof. The result follows from the previous lemma, Lemma A.1.1 and Carathéodory's Theorem. □

The next two lemmas discuss the asymptotic behavior of the representation in the previous lemmas.

Lemma A.1.5. *Let $m = n_p + 3$. Suppose at $\bar{\mathbf{p}} \in \mathcal{P}$ there exists a set of nonnegative scalars $\{\hat{\lambda}_i\}_{i=1}^m$, and a set of vectors $\{(\bar{\mathbf{y}}_i^f, \bar{\boldsymbol{\zeta}}_i^f, \bar{f}_i, \bar{s}_i^f)\}_{i=1}^m \subset \mathbb{R}^{n_p} \times \mathbb{R}^{n_p} \times \mathbb{R} \times \mathbb{R}$, satisfying*

$$(\bar{\mathbf{v}}^f, \bar{f}^v, \bar{s}^f) = \sum_{i=1}^m \hat{\lambda}_i(\bar{\boldsymbol{\zeta}}_i^f, \bar{f}_i, \bar{s}_i^f),\ \sum_{i=1}^m \lambda_i = 1, \tag{A.1.11}$$
$$\bar{\boldsymbol{\zeta}}_i^f \in \Gamma f(\bar{\mathbf{y}}_i^f),\ \forall i \in \{1,\ldots,m\},$$
$$\bar{f}_i = f(\bar{\mathbf{y}}_i^f) + \langle \bar{\boldsymbol{\zeta}}_i^f, \bar{\mathbf{p}} - \bar{\mathbf{y}}_i^f \rangle,$$
$$\|\bar{\mathbf{y}}_i^f - \bar{\mathbf{p}}\| \leq \bar{s}_i^f,\ \bar{s}_i^f \geq 0,\ \forall i \in \{1,\ldots,m\}, \tag{A.1.12}$$
$$\gamma_f \bar{s}^f = 0,\ \gamma_f > 0.$$

Then $\bar{\mathbf{v}}^f \in \Gamma f(\bar{\mathbf{p}})$.

Proof. Since $\bar{s}^f = 0$, and $\bar{s}_i^f \geq 0$, there exists i such that $\bar{s}_i^f = 0$. If $\bar{s}_i^f \neq 0$, then $\hat{\lambda}_i = 0$ per (A.1.11). $\bar{\mathbf{p}} = \bar{\mathbf{y}}_i^f$ if $\hat{\lambda}_i \neq 0$ per (A.1.12). Then, if $\hat{\lambda}_i > 0$, $\bar{\boldsymbol{\zeta}}_i^f \in \Gamma f(\bar{\mathbf{p}})$. Per (A.1.11) and convexity of $\Gamma f(\bar{\mathbf{p}})$, $\bar{\mathbf{v}}^f \in \Gamma f(\bar{\mathbf{p}})$. □

Lemma A.1.6. *Let* $m = n_p + 3$ *and suppose at* $\mathbf{p} \in \mathcal{P}$, *there exists a set of nonnegative scalars* $\{\hat{\mu}_i\}_{i=1}^m$, *and a set of vectors* $\{(\bar{\mathbf{y}}_i^G, \bar{\boldsymbol{\zeta}}_i^G, \bar{G}_i, \bar{s}_i^G)\}_{i=1}^m \subset \mathbb{R}^{n_p} \times \mathbb{R}^{n_p} \times \mathbb{R} \times \mathbb{R}$, *satisfying*

$$(\bar{\mathbf{v}}^G, \bar{G}^v, \bar{s}^G) = \sum_{i=1}^m \mu_i (\bar{\boldsymbol{\zeta}}_i^G, \bar{G}_i, \bar{s}_i^G), \tag{A.1.13}$$

$$\mu_i \geq 0, \forall i \in \{1, \ldots, m\}, \sum_{i=1}^m \mu_i = 1,$$

$$\bar{\boldsymbol{\zeta}}_i^G \in \Gamma G(\bar{\mathbf{y}}_i^G), \ \forall i \in \{1, \ldots, m\}, \tag{A.1.14}$$

$$\|\bar{\mathbf{y}}_i^G - \bar{\mathbf{p}}\| \leq \bar{s}_i^G, \ \bar{s}_i^G \geq 0, \ \forall i \in \{1, \ldots, m\}, \tag{A.1.15}$$

$$\bar{G}_i = G(\bar{\mathbf{y}}_i^G) + \langle \bar{\boldsymbol{\zeta}}_i^G, \bar{\mathbf{p}} - \bar{\mathbf{y}}_i^G \rangle, \ \forall i \in \{1, \ldots, m\}, \tag{A.1.16}$$

$$\max\{G(\bar{\mathbf{p}}), 0\} = \bar{G}^v, \tag{A.1.17}$$

$$\gamma_G \bar{s}^G = 0, \ \gamma_G > 0. \tag{A.1.18}$$

Then $\bar{\mathbf{v}}^G \in \Gamma G(\bar{\mathbf{p}})$ *and* $G(\mathbf{p}) \geq 0$.

Proof. Since $\bar{s}^G = 0$, and $\bar{s}_i^G \geq 0$, there exists i such that $\bar{s}_i^G = 0$. If $\bar{s}_i^G \neq 0$, then $\hat{\mu}_i = 0$ per (A.1.13). $\bar{\mathbf{p}} = \bar{\mathbf{y}}_i^G$ if $\hat{\mu}_i \neq 0$ per (A.1.15). Then, if $\mu_i > 0$, $\bar{\boldsymbol{\zeta}}_i^G \in \Gamma G(\bar{\mathbf{p}})$. Per (A.1.13) and convexity of $\Gamma G(\bar{\mathbf{p}})$, $\bar{\mathbf{v}}^G \in \Gamma G(\bar{\mathbf{p}})$.

Note that

$$0 = \max\{G(\bar{\mathbf{p}}), 0\} - \bar{G}^v = \sum_{i=1}^m \mu_i (\max\{G(\bar{\mathbf{p}}), 0\} - \bar{G}_i\})$$

$$0 = \sum_{i=1}^m \mu_i (\max\{G(\bar{\mathbf{p}}), 0\} - (G(\bar{\mathbf{y}}_i^G) + \langle \bar{\boldsymbol{\zeta}}_i^G, \bar{\mathbf{p}} - \bar{\mathbf{y}}_i^G \rangle)),$$

$$0 = \sum_{i=1}^{m} \mu_i (\max\{G(\bar{\mathbf{p}}), 0\} - G(\bar{\mathbf{p}})),$$

$$0 = \max\{G(\bar{\mathbf{p}}), 0\} - G(\bar{\mathbf{p}}),$$

because if $\bar{\mathbf{p}} \neq \bar{\mathbf{y}}_i^G$ holds for some i, then $s_i^G > 0$ must hold. This implies that $\hat{\mu}_i = 0$ in this case per (A.1.13), (A.1.15) and (A.1.18). Hence $G(\bar{\mathbf{p}}) \geq 0$. □

Following theorem discusses the case when the bundle method terminates after finitely many iterations with $\epsilon_s = 0$.

Theorem A.1.7. *If the modified bundle method terminates at the iteration k and $\epsilon_s = 0$, then the point $\mathbf{p}_k$ is stationary in the extended sense on S.*

Proof. Since $\mathbf{w}_k = 0$, $\mathbf{v}_k = \mathbf{0}$, $\nu_k^f \tilde{\alpha}_k^{f,v} = 0$, $\nu_k^G \tilde{\alpha}_k^{G,v} = 0$. The algorithm produces only $\mathbf{p}_k$ that are feasible. Therefore $\mathbf{p}_k$ is feasible and $G(\mathbf{p}_k) \leq 0$.

Assume $\nu_k^f \neq 0$. Then $\tilde{\alpha}_k^{f,v} = 0$. Lemmas A.1.4 and A.1.5 yield

$$\mathbf{v}_k^f = \mathbf{0},\ \mathbf{v}_k^f \in \Gamma f(\mathbf{p}_k).$$

Let $\nu_k^G \neq 0$ Then $\tilde{\alpha}_k^{G,v} = 0$. Lemmas A.1.4 and A.1.6 yield

$$\mathbf{v}_k^G = \mathbf{0},\ \mathbf{v}_k^G \in \Gamma G(\mathbf{p}_k),\ G(\mathbf{p}_k) = 0.$$

Then $\mathbf{v}_k \in \Gamma H(\mathbf{p}_k; \mathbf{p}_k)$. Note that $\nu_k^G = \mathbf{0}$ and $\nu_k^f = \mathbf{0}$ cannot occur due to the constraint (6.3.23). □

Remark A.1.8. In the remainder, it is assumed that $\epsilon = 0$ and that the modified bundle generates an infinite sequence $\{\mathbf{p}_k\}$ such that $w_k > 0$ for all k.

Lemma A.1.9. *Assume $\bar{\mathbf{p}} \in \mathcal{P}$ is a limit point of the sequence $\{\mathbf{p}_k\}$. Then there exists $K \subset \mathbb{N}$, $\bar{\mathbf{v}}^f \in \mathbb{R}^{n_p}$ and $\bar{\mathbf{v}}^G \in \mathbb{R}^{n_p}$ such that*

$$\lim_{k \in K,\, k \to \infty} \mathbf{p}_k = \bar{\mathbf{p}}, \quad \lim_{k \in K,\, k \to \infty} \mathbf{v}_k^f = \bar{\mathbf{v}}^f, \quad \lim_{k \in K,\, k \to \infty} \mathbf{v}_k^G = \bar{\mathbf{v}}^G. \tag{A.1.19}$$

In addition, if $\lim_{k \in K,\, k \to \infty} \tilde{\alpha}_k^{f,v} = 0$, then $\bar{\mathbf{v}}_k^f \in \Gamma f(\bar{\mathbf{p}})$. If $\lim_{k \in K,\, k \to \infty} \tilde{\alpha}_k^{G,v} = 0$, then $\bar{\mathbf{v}}_k^G \in \Gamma G(\bar{\mathbf{p}})$ and $G(\bar{\mathbf{p}}) \geq 0$.

Proof. There exists $K_0 \subset \mathbb{N}$ such that $\lim_{k \in K_0,\, k \to \infty} \mathbf{p}_k = \bar{\mathbf{p}}$ because $\bar{\mathbf{p}}$ is a limit point. Per Lemma A.1.4 and the fact that $\bar{a}$ is finite, it can be deduced that there exists an infinite subset $K_1 \subset K_0$ such that

$$\lim_{k \in K_1,\, k \to \infty} \bar{\mathbf{y}}_{k,i}^f = \bar{\mathbf{y}}_i^f, \quad \lim_{k \in K_1,\, k \to \infty} \bar{\mathbf{y}}_{k,i}^G = \bar{\mathbf{y}}_i^G$$

holds for all $i \in 1, \ldots, m$ where m, $\bar{\mathbf{y}}_{k,i}^f$ and $\bar{\mathbf{y}}_{k,i}^G$ are as defined in Lemma A.1.4. By the local boundedness and upper semicontinuity of Γf and ΓG, there exists and infinite set $K_2 \subset K_1$ such that

$$\lim_{k \in K_2,\, k \to \infty} \bar{\boldsymbol{\zeta}}_{k,i}^f = \bar{\boldsymbol{\zeta}}_i^f \in \Gamma f(\bar{\mathbf{y}}_i^f), \quad \lim_{k \in K_2,\, k \to \infty} \bar{\boldsymbol{\zeta}}_{k,i}^G = \bar{\boldsymbol{\zeta}}_i^G \in \Gamma G(\bar{\mathbf{y}}_i^f)$$

hold. This also implies that

$$\lim_{k \in K_2,\, k \to \infty} \bar{f}_{k,i} = \bar{f}_i, \quad \lim_{k \in K_2,\, k \to \infty} \bar{G}_{k,i} = \bar{G}_i$$

per the definition of $\bar{f}_{k,i}$ and $\bar{G}_{k,i}$. Since $\boldsymbol{\lambda}_k$ and $\boldsymbol{\mu}_k$ are bounded, there exists an infinite set $K_3 \subset K_2$ such that

$$\lim_{k \in K_3,\, k \to \infty} \lambda_{k,i} = \bar{\lambda}_i, \quad \lim_{k \in K_3,\, k \to \infty} \mu_{k,i} = \bar{\mu}_i$$

hold. Finally, there exists $K_4 \subset K_3$ such that

$$\lim_{k \in K_4,\, k \to \infty} s_{k,i}^f = \bar{s}_i^f, \quad \lim_{k \in K_4,\, k \to \infty} s_{k,i}^G = \bar{s}_i^G$$

because $s_{k,i}^f$ and $s_{k,i}^G$ are bounded. Letting $K = K_4$ and using Lemma A.1.4, (A.1.19) can be shown.

If $\lim_{k \in K,\, k \to \infty} \tilde{\alpha}_k^{f,v} = 0$, then $\bar{\mathbf{v}}_k^f \in \Gamma f(\bar{\mathbf{p}})$ follows from Lemma A.1.5. Similarly, if $\lim_{k \in K,\, k \to \infty} \tilde{\alpha}_k^{G,v} = 0$, then $\bar{\mathbf{v}}_k^G \in \Gamma G(\bar{\mathbf{p}})$ and $G(\bar{\mathbf{p}}) \geq 0$ follows from Lemma A.1.6. □

Lemma A.1.10. *Let $\bar{\mathbf{p}} \in \mathcal{P}$. Assume there exists an infinite sequence $\mathcal{J} \subset \mathbb{N}$ such that*

$$\lim_{j \in \mathcal{J},\, j \to \infty} \mathbf{p}_j = \bar{\mathbf{p}}, \quad \lim_{j \in \mathcal{J},\, j \to \infty} w_j = 0.$$

Then $\mathbf{0} \in \Gamma H(\bar{\mathbf{p}}, \bar{\mathbf{p}})$ and $\bar{\mathbf{p}} \in S$.

Proof. By (6.3.24) and the boundedness of ν_j^f and ν_j^G,

$$\lim_{j \in \mathcal{J},\, j \to \infty} \mathbf{v}_j = \mathbf{0}, \lim_{j \in \mathcal{J},\, j \to \infty} \tilde{\alpha}_j^{f,v} = 0, \lim_{j \in \mathcal{J},\, j \to \infty} \tilde{\alpha}_j^{G,v} = 0$$

hold. Then Lemma A.1.9 yields $\bar{\mathbf{v}}_j^f \in \Gamma f(\bar{\mathbf{p}})$ and $\bar{\mathbf{v}}_j^G \in \Gamma G(\bar{\mathbf{p}})$ and $G(\bar{\mathbf{p}}) \geq 0$. Since $\mathbf{p}_j \in S$ and S is closed, $\bar{\mathbf{p}} \in S$. □

In order to prove stationarity in the extended sense, given a sequence $\{\mathbf{p}_j\}$ converging to $\bar{\mathbf{p}}$, it is sufficient to show that the corresponding sequence $\{w_j\}$ converges to zero per Lemma A.1.10.

The next lemma shows that if you have a sequence $\{\mathbf{p}_j\}$ converging to $\bar{\mathbf{p}}$ by taking infinitely many long serious steps, then $\bar{\mathbf{p}}$ is stationary in the extended sense.

Lemma A.1.11. *Assume there exists an infinite set $\mathcal{J} \subset \mathbb{N}$ such that $\mathbf{p}_j \to \bar{\mathbf{p}}$ if $j \in \mathcal{J}$. Then $f(\mathbf{p}_k) \downarrow f(\bar{\mathbf{p}})$, and $t_k^L \Delta_k \to 0$ for $k \in \mathbb{N}$. In addition, if there exists an infinite sequence*

$\bar{\mathcal{J}} \subset \mathcal{J}$ such that $t_j^L \geq \hat{t} > 0$ for all $j \in \bar{\mathcal{J}}$, then $w_j \to 0$ for $j \in \bar{\mathcal{J}}$ holds and $\bar{\mathbf{p}}$ is stationary in the extended sense.

Proof. The bundle method produces an infinite sequence $\{\mathbf{p}_k\}$ such that $f(\mathbf{p}_{k+1}) \leq f(\mathbf{p}_k)$. Given any large enough k, there exists j_k^1 and j_k^2 in $\mathcal{J}$ such that $j_k^1 \leq k < j_k^2$. Then $f(\mathbf{p}_{j_k^1}) \geq f(\mathbf{p}_k) \geq f(\mathbf{p}_{j_k^2})$ holds. Since $\mathbf{p}_j \to \bar{\mathbf{p}}$ if $j \in \mathcal{J}$, $f(\mathbf{p}_{j_k^1}) - f(\mathbf{p}_{j_k^2}) \to 0$ and therefore $f(\mathbf{p}_k) - f(\mathbf{p}_{j_k^2}) \to 0$. Then $|f(\bar{\mathbf{p}}) - f(\mathbf{p}_k)| \leq |f(\bar{\mathbf{p}}) - f(\mathbf{p}_{j_k^2})| + |f(\mathbf{p}_{j_k^2}) - f(\mathbf{p}_k)|$ and the desired result follows as $k \to \infty$.

Observe that

$$0 \leq -t_k^L \Delta_k \leq \frac{f(\mathbf{p}_k) - f(\mathbf{p}_{k+1})}{m_L}$$

and $f(\mathbf{p}_k) - f(\mathbf{p}_{k+1}) \to 0$. Therefore $t_k^L \Delta_k \to 0$.

Since $t_j^L \Delta_j \to 0$ and $t_j^L \geq \hat{t} > 0$ for $j \in \bar{\mathcal{J}}$, $\Delta_j \to 0$ and therefore $w_j \to 0$ for $j \in \bar{\mathcal{J}}$. Stationarity follows from Lemma A.1.10. □

Corollary A.1.12. *Suppose there exist $\bar{\mathbf{p}} \in \mathcal{P}$ and an infinite sequence $\mathcal{J} \subset \mathbb{N}$ such that $\mathbf{p}_j \to \bar{\mathbf{p}}$ for $j \in \mathcal{J}$. Assume $\liminf_{k \to \infty} \max(\|\mathbf{p}_k - \bar{\mathbf{p}}\|, w_k) \geq \bar{\epsilon} > 0$. Then $t_j^L \to 0$ for $j \in \mathcal{J}$.*

Proof. The condition $\liminf_{k \to \infty} \max(\|\mathbf{p}_k - \bar{\mathbf{p}}\|, w_k) \geq \bar{\epsilon} > 0$ and $\mathbf{p}_j \to \bar{\mathbf{p}}$ for $j \in \mathcal{J}$ imply that $w_k > 0$ for k large enough. Per Lemma A.1.11, $t_j^L \Delta_j \to 0$. Since for large enough j, $w_j > 0$, $\Delta_j < -\bar{\epsilon}$ for large enough j. Therefore, $t_j^L \to 0$ for $j \in \mathcal{J}$. □

In order to prove that an accumulation point is stationary, it has to be shown that if $t_j^L \to 0$, $w_j \to 0$.

The following lemma relates the solution of the quadratic direction finding problem (6.3.22) to Δ_k and w_k.

Lemma A.1.13. *Define*

$$\begin{aligned}
\hat{w}_k &= \frac{1}{2}\|\mathbf{v}_k\|^2 + \hat{\alpha}_k, \\
\hat{\Delta}_k &= -\left(\|\mathbf{v}_k\|^2 + \hat{\alpha}_k\right), \\
\hat{\alpha}_k &= \nu_k^f \hat{\alpha}_k^{f,v} + \nu_k^G \hat{\alpha}_k^{G,v}, \\
\hat{\alpha}_k^{f,v} &= \sum_{j \in M_k^f} \tilde{\lambda}_{k,j} \alpha_{k,j}^f + \tilde{\lambda}_k^v \alpha_k^{f,v}, \\
\hat{\alpha}_k^{G,v} &= \sum_{j \in M_k^G} \tilde{\mu}_{k,j} \alpha_{k,j}^G + \tilde{\mu}_k^v \alpha_k^{G,v}.
\end{aligned}$$

Then

$$\begin{aligned}
&0 \leq \tilde{\alpha}_k^{f,v} \leq \hat{\alpha}_k^{f,v},\ 0 \leq \tilde{\alpha}_k^{G,v} \leq \hat{\alpha}_k^{G,v}, \\
&0 \leq w_k \leq \hat{w}_k,\ \Delta_k \leq -w_k \leq 0,\ \hat{\Delta}_k \leq \Delta_k \leq 0
\end{aligned}$$

hold.

Proof. Note that

$$\begin{aligned}
|f(\mathbf{p}_k) - \tilde{f}_k^v| &= |\sum_{j \in M_k^f} \tilde{\lambda}_j (f(\mathbf{p}_k) - f_{k,j}) + \tilde{\lambda}_k^v f_k^v|, \\
\gamma_f(\tilde{s}_k^f)^2 &= \gamma_f(\sum_{j \in M_k^f} \tilde{\lambda}_j s_{k,j} + \tilde{\lambda}_k^v s_k^f)^2 \leq \gamma_f(\sum_{j \in M_k^f} \tilde{\lambda}_j s_{k,j}^2 + \tilde{\lambda}_k^v (s_k^f)^2)
\end{aligned}$$

The last inequality follows from Jensen's Inequality. Hence

$$\tilde{\alpha}_k^{f,v} \leq \sum_{j \in M_k^f} \tilde{\lambda}_j \max(|f(\mathbf{p}_k) - f_{k,j}|, \gamma_f s_{k,j}^2) + \tilde{\lambda}_k^v \max(|f(\mathbf{p}_k) - f_k^v|, \gamma_f s_k^f)^2) = \hat{\alpha}_k^{f,v}$$

holds. The inequality $\tilde{\alpha}_v^{G,k} \leq \hat{\alpha}_k^{G,v}$ follows from similar reasoning. The rest of the inequalities

follow from the definitions of the quantities. □

The next lemma relates the termination criteria at iteration k to the termination criteria at iteration $k+1$ in case no reset occurs and no long serious step is taken.

Lemma A.1.14. *Suppose that $t_{k-1}^L < \bar{t}$ and $r_k^a = 0$ for some $k > 1$. Let*

$$\zeta_k = \begin{cases} \zeta_k^f & \text{if } \mathbf{y}_k \in S \\ \zeta_k^G & \text{otherwise} \end{cases}$$

$$\alpha_k^v = \nu_{k-1}^f \alpha_k^{f,v} + \nu_{k-1}^G \alpha_k^{G,v}.$$

Let $\Phi_C : \mathbb{R} \to \mathbb{R}$ be defined by

$$\Phi_C(x) = x - (1 - m_R)^2 \frac{x^2}{8C^2}$$

where C is any number that satisfying

$$C \geq \max(\|\mathbf{v}_{k-1}\|, \|\zeta_k\|, \tilde{\alpha}_{k-1}^v, 1). \tag{A.1.20}$$

Then

$$w_k \leq \hat{w}_k \leq \Phi_C(w_{k-1}) + |\alpha_k^v - \tilde{\alpha}_{k-1}^v|$$

holds.

Proof. The proof is the same as the proof of Lemma 4.7 in Chapter 6 in [54]. Therefore it is omitted for brevity. □

Lemma A.1.15. *For any $\epsilon_w > 0$ and $C > 0$, there exists numbers ϵ_a and $\bar{N} \geq 1$ such that*

for any sequences of numbers t_i satisfying

$$0 \leq t_{i+1} \leq \Phi_C(t_i) + \epsilon_a,\ i \geq 1,\ 0 \leq t_i \leq 4C^2, \tag{A.1.21}$$

$t_i \leq \epsilon_w$ *holds for all* $i \geq \bar{N}$.

Proof. Proof is the same as proof of Lemma 4.12 in Chapter 3 in [54]. □

Lemma A.1.14 and A.1.15 imply that $w_j \to 0$ if $t_j^L \to 0$ for some infinite set $\mathcal{J} \subset \mathbb{N}$ provided that for sufficiently many iterations $\bar{N}$,

1. a local bound of the form (A.1.20) exists;
2. no distance resetting occurs;
3. $|\alpha_j^v - \tilde{\alpha}_{j-1}^v| \leq \epsilon_a$
4. and $t_{j-1}^L \leq \bar{t}$ holds.

The next lemma provides a finite C so that (A.1.20) is satisfied.

Lemma A.1.16. *Let*

$$\zeta_k = \begin{cases} \zeta_k^f & \text{if } \mathbf{y}_k \in S, \\ \zeta_k^G & \text{otherwise,} \end{cases}$$

$$\alpha_k = \begin{cases} \max(|f(\mathbf{p}_k) - f_{k,k}|, \gamma_f \|\mathbf{p}_k - \mathbf{y}_k\|^2) & \text{if } \mathbf{y}_k \in S, \\ \max(|G_{k,k}|, \gamma_G \|\mathbf{p}_k - \mathbf{y}_k\|^2) & \text{otherwise,} \end{cases}$$

For each $k \geq 1$

$$\max\{\|\mathbf{v}_k\|, \tilde{\alpha}_k^v\} \leq \max\{\frac{1}{2}\|\zeta_k\|^2 + \alpha_k, \sqrt{\|\zeta_k\|^2 + 2\alpha_k}\}.$$

Let $\bar{\mathbf{p}} \in \mathcal{P}$ *and* $B = \{\mathbf{p} \in \mathcal{P} : \|\mathbf{p} - \bar{\mathbf{p}}\| \leq 2\bar{a}\}$. *Let*

$$
\begin{aligned}
C_g &= \sup\{\|\boldsymbol{\zeta}\| : \ \boldsymbol{\zeta} \in \Gamma H(\mathbf{p};\mathbf{p}),\ \mathbf{p} \in B\}, \\
C_\alpha &= \sup\{\alpha(\mathbf{p}_1,\mathbf{p}_2) : \mathbf{p}_1 \in B,\ \mathbf{p}_2 \in B\}, \\
C &= \max\left\{\frac{1}{2}C_g^2 + C_\alpha, \sqrt{C_g^2 + 2C_\alpha}, 1\right\}.
\end{aligned}
$$

Then C *is finite and*

$$
C \geq \max(\|\mathbf{v}_{k-1}\|, \|\boldsymbol{\zeta}_k\|, \tilde{\alpha}_{k-1}^v, 1).
$$

holds if $\|\mathbf{p}_k - \bar{\mathbf{p}}\| \leq \bar{a}$.

Proof. The proof of the lemma is almost identical to the proof of Lemma 5.4.8 on page 261 in [54]. Instead of using the local boundedness of the generalized gradient, the local boundedness of the linear Newton approximation is used to derive C_g. □

The next lemma states that the difference $|\alpha_j^v - \tilde{\alpha}_{j-1}^v|$ goes to zero under certain conditions.

Lemma A.1.17. *Suppose that there exists* $\bar{\mathbf{p}} \in \mathcal{P}$ *and an infinite set* $\mathcal{J} \subset \mathbb{N}$ *such that* $\mathbf{p}_j \to \bar{\mathbf{p}}$ *and* $\|\mathbf{p}_{j+1} - \mathbf{p}_j\| \to 0$ *if* $j \in \mathcal{J}$. *Then the sequences* $\{\mathbf{v}_j^f\}$ *and* $\{\mathbf{v}_j^G\}$ *are bounded for* $j \in \mathcal{J}$ *and*

$$
\begin{aligned}
\lim_{j\in\mathcal{J},\, j\to\infty} |\alpha_{j+1}^{f,v} - \tilde{\alpha}_j^{f,v}| &\to 0, \\
\lim_{j\in\mathcal{J},\, j\to\infty} |\alpha_{j+1}^{G,v} - \tilde{\alpha}_j^{G,v}| &\to 0, \\
\lim_{j\in\mathcal{J},\, j\to\infty} |\alpha_{j+1}^{v} - \tilde{\alpha}_j^{v}| &\to 0
\end{aligned}
$$

hold.

Proof. Proof is the same as in Lemma 5.4.9 on page 262 in [54] therefore it is omitted. □

The following lemma establishes the fact that if there exists $\bar{\mathbf{p}} \in \mathcal{P}$ and an infinite set $\mathcal{J} \subset \mathbb{N}$ such that $\mathbf{p}_j \to \bar{\mathbf{p}}$ for $j \in \mathcal{J}$, then $\|\mathbf{p}_j - \mathbf{p}_{j+1}\| \to 0$ for $j \in \mathcal{J}$.

Lemma A.1.18. *Suppose that there exists $\bar{\mathbf{p}} \in \mathcal{P}$ and an infinite set $\mathcal{J} \subset \mathbb{N}$ such that $\mathbf{p}_j \to \bar{\mathbf{p}}$ for $j \in \mathcal{J}$ and $\liminf_{k\to\infty} \max(\|\mathbf{p}_k - \bar{\mathbf{p}}\|, w_k) \geq \bar{\epsilon} > 0$. Then for any fixed integer $m \geq 0$ there exists a j_m such that for any integer $n \in [0, m]$*

$$\lim_{j\in\mathcal{J},\, j\to\infty} |\mathbf{p}_{j+n} - \bar{\mathbf{p}}\| = 0,$$

$$\lim_{j\in\mathcal{J},\, j\to\infty} t^L_{k+n} = 0,$$

$$w_{k+n} \geq \bar{\epsilon}/2,\ \forall j > j_m,\ j \in \mathcal{J}.$$

Moreover, for any numbers $\hat{j}$, $\hat{N}$ and ϵ_a, there exists a number $\tilde{j} \geq \hat{j}$, $\tilde{j} \in \mathcal{J}$, such that

$$w_k \geq \bar{\epsilon}/2 \qquad \text{for } \tilde{j} \leq k \leq \tilde{j} + \tilde{N}, \tag{A.1.22}$$

$$C \geq \max(\|\mathbf{v}_{k-1}\|, \|\boldsymbol{\zeta}_k\|, \tilde{\alpha}^v_{k-1}, 1) \qquad \text{for } \tilde{j} \leq k \leq \tilde{j} + \tilde{N}, \tag{A.1.23}$$

$$|\alpha^v_k - \tilde{\alpha}^v_{k-1}| \leq \epsilon_a \qquad \text{for } \tilde{j} \leq k \leq \tilde{j} + \tilde{N}, \tag{A.1.24}$$

$$t^L_k < \bar{t} \qquad \text{for } \tilde{j} \leq k \leq \tilde{j} + \tilde{N}. \tag{A.1.25}$$

where C is defined in Lemma A.1.16.

Proof. This is the same as Lemma 4.15 on page 119 in [54].The proof is omitted for brevity. □

Lemma A.1.19. *Suppose there exists $\bar{\mathbf{p}} \in \mathcal{P}$ and an infinite set $\mathcal{J} \subset \mathbb{N}$ such that $\mathbf{p}_j \to \bar{\mathbf{p}}$ for $j \in \mathcal{J}$, then $\liminf_{k\to\infty} \max(\|\mathbf{p}_k - \bar{\mathbf{p}}\|, w_k) \to 0$.*

Proof. Assume for contradiction purposes that $\liminf_{k\to\infty} \max(\|\mathbf{p}_k - \bar{\mathbf{p}}\|, w_k) \geq \bar{\epsilon} > 0$. Let $\epsilon_w = \bar{\epsilon}/2 > 0$ and choose ϵ_a and $\bar{N}$ as specified in Lemma A.1.15 where C is the constant

defined in Lemma A.1.16. Let $\tilde{N} = 10\bar{N}$. Using the previous lemma, choose $\tilde{j}$ satisfying (A.1.22)-(A.1.25) and

$$\sum_{k=\tilde{j}}^{\tilde{j}+\tilde{N}} \|\mathbf{p}_{k+1} - \mathbf{p}_k\| \leq \bar{a}/4. \tag{A.1.26}$$

Suppose there exists a number $\hat{k}$ satisfying $\tilde{j} \leq \hat{k} \leq \tilde{j} + \tilde{N} - 2\bar{N}$ such that $r_k^a = 0$ for all $k \in [\hat{k}, \hat{k} + \bar{N}]$. Then (A.1.23), (A.1.24) and (A.1.25), Lemma A.1.14, Lemma A.1.15 imply that $w_k \leq \epsilon_w = \bar{\epsilon}/2$ for some $k \in [\hat{k}, \hat{k} + \bar{N}]$ which contradicts (A.1.22) and the assumption that $r_k^a = 0$ for all $k \in [\hat{k}, \hat{k} + \bar{N}]$. Hence for any $\hat{k}$ such that $\tilde{j} \leq \hat{k} \leq \tilde{j} + \tilde{N} - 2\bar{N}$ holds, $r_k^a = 1$ for some $k \in [\hat{k}, \hat{k} + \bar{N}]$.

Let $\tilde{j} = \hat{k}$. Let $r_{k_l} = 1$ for some $k_l \in [\tilde{j}, \tilde{j}+\bar{N}]$. Then $a_{k_1} \leq \bar{a}/2$. Since $\|\mathbf{y}_{k+1} - \mathbf{p}_{k+1}\| \leq \bar{a}/2$ due to the line search rules and

$$a_{k+1} = \max\{a_k + \|\mathbf{p}_{k+1} - \mathbf{p}_k\|, \|\mathbf{y}_{k+1} - \mathbf{p}_{k+1}\|\} \leq \max\{a_k + \|\mathbf{p}_{k+1} - \mathbf{p}_k\|, \bar{a}/2\},$$

$a_k \leq 3/4\bar{a} \leq \bar{a}$ follows using (A.1.26). Hence no reset occurs for $k \in [k_l + 1, k_l + 1 + \bar{N}]$. However, $\hat{k} = k_l + 1$ satisfies $\tilde{j} \leq \hat{k} \leq \tilde{j} + \tilde{N} - 2\bar{N}$ and therefore there has to be a reset for $k \in [\hat{k}, \hat{k} + \bar{N}]$. This is a contradiction. Hence $\liminf_{k\to\infty} \max(\|\mathbf{p}_k - \bar{\mathbf{p}}\|, w_k) \to 0$ has to hold. □

Theorem A.1.20. *Each accumulation point of the sequence $\{\mathbf{p}_k\}$ generated by the bundle method is stationary in the extended sense.*

Proof. The proof follows from Lemma A.1.19 and Lemma A.1.10. □

Corollary A.1.21. *If the level set $P = \{\mathbf{p} \in \mathcal{P} : f(\mathbf{p}) \leq \mathbf{f}(\mathbf{p}_1)\}$ is bounded, $\mathrm{cl}\,(P) \subset \mathcal{P}$, and the final accuracy tolerance ϵ_s is positive, then the bundle method terminates in a finite number of iterations.*

Proof. The boundedness of the level set P implies that there exists an infinite set $\mathcal{J} \subset \mathbb{N}$ and $\bar{\mathbf{p}}$ such that $\mathbf{p}_j \to \bar{\mathbf{p}}$ if $j \in \mathcal{J}$ per the Bolzano-Weierstrass Theorem. Then per Lemma A.1.19, $w_j \to 0$ for $j \in \mathcal{J}$. Hence for large enough j, $w_j \leq \epsilon_s$. This implies that for large enough k, $w_k \leq \epsilon_s$. □

Bibliography

[1] *Interval Methods for Systems of Equations.* Cambridge University Press, Cambridge, England, 1990.

[2] M. U. Akhmet. On the smoothness of solutions of impulsive autonomous systems. *Nonlinear Analysis*, 60:311–324, 2005.

[3] R. Alur, C. Courcoubetis, N. Halbwachs, T. A. Henzinger, P. H. Ho, X. Nicollin, A. Olivero, J. Sifakis, and S. Yovine. The algorithmic analysis of hybrid systems. *Theoretical Computer Science*, 138(1):3–34, 1995.

[4] U. M. Ascher and L. R. Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations.* SIAM, Philadelphia, 1998.

[5] J. Aubin and H. Frankowska. *Set-Valued Analysis.* Birkhauser, Boston, 1990.

[6] C. Audet and J.E. Dennis JR. Analysis of Generalized Pattern Searches. *SIAM Journal of Control and Optimization*, 13(3):889–903, 2003.

[7] M. P. Avraam, N. Shah, and C.C. Pantelides. Modelling and Optimisation of General Hybrid Systems in the Continuous Time Domain. *Computers and Chemical Engineering*, 22(Suppl):S221–S228, 1998.

[8] P. I. Barton, J. R. Allgor, W. F. Feehery, and S. Galán. Dynamic optimization in a discontinuous world. *Industrial & Engineering Chemistry Research*, 37(3):966 – 981, 1998.

[9] P. I. Barton, J. R. Banga, and S. Galán. Optimization of hybrid discrete/continuous dynamic systems. *Computers & Chemical Engineering*, 24(9–10):2171 – 2182, 2000.

[10] P. I. Barton and C. K. Lee. Modeling, simulation, sensitivity analysis, and optimization of hybrid systems. *ACM Trans. Model. Comput. Simul.*, 12(4):256–289, 2002.

[11] P. I. Barton and C. C. Pantelides. Modeling of combined discrete/continuous processes. *AIChE Journal*, 40:966–979, 1994.

[12] B. T. Baumrucker and L. T. Biegler. MPEC strategies for the optimization of a class of hybrid dynamic systems. *Journal of Process Control*, 19:1248–1256, 2009.

[13] A. Bemporad and M. Morari. Control of systems integrating logic, dynamics, and constraints. *Automatica*, 35:407–427, 1999.

[14] D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, Massachusetts, 1995.

[15] D. P. Bertsekas. *Convex Optimization Theory.* Athena Scientific, Belmont, Massachusetts, 2009.

[16] D. Bertsimas and J. N. Tsiklis. *Introduction to Linear Optimization.* Athena Scientific, Belmont, Massachusetts, 1997.

[17] J. T. Betts. Survey of numerical methods for trajectory optimization. *Journal of Guidance Control and Dynamics*, 21(2):193–207, 1998.

[18] A. Bihain. Optimization of Upper Semidifferentiable Functions. *Journal of Optimization Theory and Applications*, 44(4):545–568, December 1984.

[19] J. M. Borwein and A. S. Lewis. *Convex analysis and nonlinear optimization: theory and examples.* New York, 2009.

[20] Michael S. Branicky, Vivek S. Borkar, and Sanjoy K. Mitter. A Unified Framework for Hybrid Control: Model and Optimal Control Theory. *IEEE Transactions on Automatic Control*, 43(1):31–45, January 1998.

[21] A. E. Bryson and Y. Ho. *Applied Optimal Control.* Taylor & Francis, Bristol, PA 19007 USA, 1975.

[22] J. V. Burke, A. S Lewis, and M. L. Overton. Approximating Subdifferentials by Random Sampling of Gradients. *Accepted to: SIAM Journal of Control*, October 2003.

[23] P. E. Caines and M. S. Shaikh. Optimality zone algorithms for hybrid systems: Efficient algorithms for optimal location and control computation. In J .Hespanha and A. Tiwari, editors, *Hybrid Systems: Computation and Control*, volume 3927 of *Lecture Notes in Computer Science*, pages 123 – 137. Springer, 2006.

[24] Y. Cao, S. Li, L. Petzold, and R. Serban. Adjoint Sensitivity Analysis for Differential-Algebraic Equations: The Adjoint DAE system and its numerical solution. *SIAM Journal of Scientific Computing*, 24(3):1076 – 1089, 2003.

[25] F. H. Clarke. *Optimization and Nonsmooth Analysis.* Number 5 in Classics in Applied Mathematics. SIAM, Philadelphia, 1990.

[26] E. A. Coddington and N. Levinson. *Theory of Ordinary Differential Equations.* McGraw Hill Co., Inc., New York, 1955.

[27] R. W. Cottle, J. S. Pang, and R. E. Stone. *The Linear Complementarity Problem.* Number 60 in Classics in Applied Mathematics. SIAM, Philadelphia, 2009.

[28] A. L. Dontchev and R. T. Rockafellar. Robinson's implicit function theorem and its extensions. *Mathematical Programming*, 117(1-2):129–147, 2009.

[29] Arne Drud. CONOPT: a GRG code for large sparse dynamic nonlinear optimization problems. *Mathematical Programming*, 31(2):153–191, 1985.

[30] Arne Drud. *CONOPT Solver Manual.* GAMS Development Corporation, Washington, DC, 2004.

[31] P. Dua, V. Dua, and E. N. Pistikopoulos. Optimal delivery of chemotherapeutic agents in cancer. *Computers & Chemical Engineering*, 32(1-2):99–107, JAN-FEB 2008.

[32] K. El-Rifai. *Robust adaptive control of switched systems.* PhD thesis, Massachussetts Institute of Technology, 2007.

[33] Y. M. Ermoliev and V. I. Norkin. On nonsmooth and discontinuous problems of stochastic systems optimization. *European Journal of Operational Research*, 101:230–244, 1997.

[34] Y. M. Ermoliev, V. I. Norkin, and R. J-B Wets. The Minimization of Semicontinuous Functions: Mollifier subgradients. *SIAM Journal of Control and Optimization*, 33(1):149–167, January 1995.

[35] F. Facchinei and J. S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems.* Springer, New York, 2003.

[36] 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.

[37] A. F. Filippov. *Differential Equations with Discontinuous Righthand Sides.* Springer, New York, 1988.

[38] S. Galán and P. I. Barton. Dynamic optimization of hybrid systems. *Computers & Chemical Engineering*, 22, Suppl:S183–S190, 1998.

[39] S. Galán, W. F. Feehery, and P. I. Barton. Parametric sensitivity functions for hybrid discrete/continuous systems. *Applied Numerical Mathematics*, 31:17–47, 1999.

[40] C. J. Goh and K. L. Teo. Control parametrization: A unified approach to optimal control problems with general constraints. *Automatica*, 24(1):3–18, 1988.

[41] K. Gökbayrak and C. Cassandras. Hybrid Controllers for Hierarchically Decomposed Systems. In B. Krogh and N. Lynch, editors, *Hybrid Systems: Computation and Control*, volume 1790 of *Lecture Notes in Computer Science*, pages 117–129. Springer Verlag, 2000.

[42] M. S. Gowda. Inverse and Implicit Function Theorems for H-Differentiable and Semismooth functions. *Optimization Method and Software*, 19(5):443–461, 2004.

[43] A. Griewank. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation.* Number 19 in Frontiers in Applied Mathematics. SIAM, Philadelphia, 2000.

[44] T. H. Gronwall. Note on derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, 20:292–296, 1919.

[45] S. Hedlund and A. Rantzer. Convex dynamic programming for hybrid systems. *IEEE Transactions on Automatic Control*, 47(9):1536–1540, September 2002.

[46] J. L. Hjersted and M. A. Henson. Optimization of fed-batch sacchoromyces cerevisia fermentation using dynamic flux balance models. *Biotechnology Progress*, 22:1239–1248, 2006.

[47] R. Horst and H. Tuy. *Global Optimization: Deterministic Approaches.* Springer, Berlin, 3. edition, 1996.

[48] C. Imbert. Support functions of the Clarke generalized Jacobian and of its plenary hull. *Nonlinear Analysis*, 49(8):1111–1125, 2002.

[49] E. M. Izhikevich. *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting.* The MIT Press, Cambridge, MA, 2007.

[50] M. A. Jenkins. Algorithm 493 Zeros of a Real Polynomial [C2]. *ACM Transactions on Mathematical Software*, 1(2):178–189, 1975.

[51] M. A. Jenkins and J. F. Traub. A three-stage algorithm for real polynomials using quadratic iteration. *SIAM Journal of Numerical Analysis*, 7(4):545–566, 1970.

[52] K. H. Johansson, M. Egerstedt, J. Lygeros, and S. Sastry. On the regularization of Zeno hybrid automata. *Systems & Control Letters*, 38:141–150, 1999.

[53] S. Kameswaran and L. T. Biegler. Convergence rates for direct transcription of optimal control problems using collocation at Radau points. *Computational Optimization and Applications*, 41:81–126, 2008.

[54] K. C. Kiwiel. *Methods of Descent for Nondifferentiable Optimization*, volume 1133 of *Lecture Notes in Mathematics.* Springer-Verlag, New York, 1985.

[55] C. K. Lee. *Global Optimization of Hybrid Systems.* PhD thesis, Massachusetts Institute of Technology, 2006.

[56] C. K. Lee and P. I. Barton. Global optimization of linear hybrid systems with varying transition times. *SIAM Journal on Control and Optimization*, 47(2):791–816, 2008.

[57] C. K. Lee, A. P. Singer, and P. I. Barton. Global optimization of linear hybrid systems with explicit transitions. *Systems & Control Letters*, 51(5):363–375, April 2004.

[58] J. M. Lee, E. P Gianchandani, and J. A. Papin. Flux balance analysis in the era of metabolomics. In *Briefings in Bioinformatics*, volume 7, pages 140–150. 2006.

[59] D. B. Leineweber, I. Bauer, H. G. Bock, and J. P. Schlöder. An efficient multiple shooting based reduced SQP strategy for large-scale dynamic process optimization. part 1: theoretical aspects. *Computers and Chemical Engineering*, 27:157–166, 2003.

[60] D. B. Leineweber, A. Schäfer, H. G. Bock, and J. P. Schlöder. An efficient multiple shooting based reduced SQP strategy for large-scale dynamic process optimization. part 1: Software aspects and Applications. *Computer and Chemical Engineering*, 27:167–174, 2003.

[61] C. Lemaréchal. *Optimization*, chapter Nondifferentiable Optimization. North-Holland, Amsterdam, 1989.

[62] Daniel Liberzon. *Switching in Systems and Control.* Birkhäuser, 2003.

[63] L. Lukšan and Vlček J. A bundle-Newton method for nonsmooth unconstrained minimization. *Mathematical Programming*, 83(1–3):373–391, 1998.

[64] L. Lukšan and Vlček J. Algorithm 811: NDA: Algorithms for nondifferentiable optimization. *ACM Transactions on Mathematical Software*, 27(2):193–213, June 2001.

[65] J. Lygeros, K. H. Johansson, S. N. Simić, J. Zhang, and S. S. Sastry. Dynamical properties of hybrid automata. *IEEE Transactions on Automatic Control*, 48(1):2–17, January 2003.

[66] M. M. Mäkelä. Survey of Bundle Methods for Nonsmooth Optimization. *Optimization Methods and Software*, 17(1):1–29, 2001.

[67] M. M. Mäkelä, M. Mietinnen, L. Lukšan, and Vlček J. Comparing Nonsmooth nonconvex bundle methods in solving hemivariational inequalities. *Journal of Global Optimization*, 14(2):117–135, 1999.

[68] M. M. Mäkelä and P. Neittaanmäki. *Nonsmooth Optimization: Analysis and Algorithms with Applications to Optimal Control.* World Scientific, Singapore, 1992.

[69] O. L. Mangasarian and T. H. Shiau. Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems. *SIAM Journal of Control and Optimization*, 25(3):583–595, 1987.

[70] R. Mifflin. An Algorithm for Constrained Optimization with Semismooth Functions. *Mathematics of Operations Research*, 2(2):191–207, 1977.

[71] R. Mifflin. Semismooth and Semiconvex Functions in Constrained Optimization. *SIAM Journal of Control and Optimization*, 15(6):959–972, 1977.

[72] R.E. Moore. A test for existence of solution to nonlinear systems. *SIAM Journal of Numerical Analysis*, 14:611–615, 1977.

[73] P. J. Mosterman. *Hybrid Dynamic Systems: A Hybrid Bond Graph Modeling Paradigm and Its Application in Diagnosis.* PhD thesis, Vanderbilt University, 1997.

[74] P. J. Mosterman. An overview of hybrid simulation phenomena and their support by simulation packages. In F. W. Vaandrager and J.H. van Schuppen, editors, *Hybrid Systems: Computation and Control*, volume 1569 of *Lecture Notes in Computer Science*, pages 165–177, Berlin, 1999. Springer-Verlag.

[75] P. J. Mosterman, F. Zhao, and G. Biswas. An ontology for transitions in physical systems. In *Proceedings of the National Conference on Artificial Intelligence (AAAI-98)*, pages 219–224, Madison, WI, 1998.

[76] J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables.* Academic Press, Inc., Boston, 1970.

[77] D. B. Özyurt and P. I. Barton. Cheap Second Order Directional Derivatives of Stiff ODE Embedded Functionals. *SIAM Journal of Scientific Computing*, 26(5):1725–1743, 2005.

[78] Z. Páles and V. Zeidan. Infinite dimensional generalized Jacobian: Properties and calculus rules. *Journal of Mathematical Analysis and Applications*, 344(1):55–75, 2008.

[79] J. S. Pang and D. Ralph. Piecewise Smoothness, Local Invertibility, and Parametric Analysis of Normal Maps. *Mathematics of Operations Research*, 21(2):401–426, 1996.

[80] J. S. Pang and D. E. Stewart. Differential variational inequalities. *Mathematical Programming*, 113(2):345–424, 2008.

[81] J. S. Pang and D. E. Stewart. Solution dependence on initial conditions in differential variational inequalities. *Mathematical Programming*, 116(1-2):429–460, 2009.

[82] J. C. Pannetta and J. Adam. A mathematical model of cycle-specific chemotherapy. *Mathematical and Computer Modelling*, 22(2):67–82, 1995.

[83] T. Park and P. I. Barton. State event location in differential-algebraic models. *ACM Trans. Model. Comput. Simul.*, 6(2):137–165, 1996.

[84] D. L. Pepyne and C. G. Cassandras. Optimal Control of Hybrid Systems in Manufacturing. *Proceedings of the IEEE*, 88(7):1108–1123, July 20010.

[85] L. Petzold, S. Li, Y. Cao, and R. Serban. Sensitivity analysis of differential-algebraic equations and partial differential equations. *Computers and Chemical Engineering*, 30:1553–1559, 2006.

[86] B. Piccoli. Necessary conditions for hybrid optimization. In *Proceedings of the 38th IEEE Conference on Decision and Control*, pages 410–415. IEEE, 1999.

[87] E. Polak, D.Q. Mayne, and Y. Wardi. On the extension of constrained optimization algorithms from differentiable to nondifferentiable case. *SIAM Journal of Control and Optimization*, 21(2):179–203, March 1983.

[88] L. Qi. Semismoothness and Decomposition of Maximal Normal Operators. *Journal of Mathematical Analysis and Applications*, 146(1):271–279, 1990.

[89] L. Qi. A nonsmooth version of Newton's method. *Mathematical Programming*, 58(3):353–367, 1993.

[90] A. U. Raghunathan, M. S. Diaz, and L. T. Biegler. An MPEC formulation for dynamic optimization of distillation operations. *Computers & Chemical Engineering*, 28(10):2037–2052, 2004.

[91] D. Ralph and S. Scholtes. Sensitivity analysis of composite piecewise smooth equations. *Mathematical Programming*, 76(3):593–612, 1997.

[92] R. T. Rockafellar and R. J-B. Wets. *Variational Analysis.* Number 317 in Grundlehren der mathematischen Wissenschaften. Springer, New York, 1998.

[93] E. N Rozenwasser. General sensitivity equations of discontinuous systems. *Automatic Remote Control*, pages 400–404, 1967.

[94] E. N. Rozenwasser and R. M. Yusupov. *Sensitivity of Automatic Control Systems.* CRC Press, Boca Raton, 1999.

[95] A. I. Ruban. Sensitivity Coefficients for Discontinuous Dynamic Systems. *Journal of Computer and Systems Sciences International*, 36(4):536–542, 1997.

[96] W. Rudin. *Principles of Mathematical Analysis.* McGraw-Hill Book Company, New York, 3rd edition, 1978.

[97] S. Sager, H. G. Bock, and G. Reinelt. Direct methods with maximal lower bound for mixed-integer optimal control problems. *Mathematical Programming*, 118(1), 2009.

[98] S. Scholtes. Introduction to piecewise differentiable equations, 1994. Habilitation Thesis, Institut für Statistik und Mathematische Wirtschaftstheorie, University of Karlsruhe.

[99] J. M. Schumacher. Complementarity systems in optimization. *Mathematical Programming*, 101(1):263 – 295, 295 2004.

[100] J. K. Scott, B. Chachuat, and P. I. Barton. Nonlinear convex and concave relaxations for the solution of parametric ODEs. submitted, 2009.

[101] M. S. Shaikh and P. E. Caines. On the Optimal Control of Hybrid Systems: Optimization of Trajectories, Switching Times and Location Schedules. In O. Maler and A. Pnueli, editors, *Hybrid Systems: Computation and Control*, volume 2623 of *Lecture Notes in Computer Science*, pages 466–481. Springer Verlag, 2003.

[102] O. Stursberg and S. Panek. Control of switched hybrid systems based on disjunctive formulations. In C. J. Tomlin and M. R. Greenstreet, editors, *Hybrid Systems:Computation and Control*, volume 2289 of *Lecture Notes in Computer Science*, pages 421–435, Berlin, 2002. Springer-Verlag.

[103] H. Sussman. A maximum principle for hybrid optimal control problems. In *Proceedings of the 38th IEEE Conference on Decision and Control*, pages 425–430. IEEE, 1999.

[104] L. Tavernini. Differential Automata and Their Discrete Simulators. *Nonlinear Analysis, Theory, Methods and Applications*, 11(6):665–683, 1987.

[105] K. L. Teo, C. J. Goh, and K. H. Wong. *A Unified Computational Approach to Optimal Control Problems.* John Wiley & Sons, Inc, 1991.

[106] J. Till, S. Engell, S. Panek, and O. Stursberg. Applied hybrid system optimization: An empirical investigation of complexity. *Control Engineering Practice*, 12:1291–1303, 2004.

[107] J. Tolsma and P. I. Barton. DAEPACK: an open modeling environment for legacy models. *Industrial & Engineering Chemistry Research*, 39(6):1826–1839, 2000. (http://yoric.mit.edu/daepack/daepack.html).

[108] J. E. Tolsma. DSL48SE manual (version 1.0). Technical report, Process Systems Engineering Laboratory, Department of Chemical Engineering, Massachusetts Institute of Technology, 2001. (http://yoric.mit.edu/daepack/download/Manuals_Pres/dsl48se.ps).

[109] J. E. Tolsma and P. I. Barton. Hidden discontinuities and parametric sensitivity analysis. *SIAM Journal on Scientific Computing*, 23(6):1861–1874, 2002.

[110] A. J. van der Schaft and J. M. Schumacher. Complementarity modeling of hybrid systems. *IEEE Transactions on Automatic Control*, 43(4):483 – 490, 1998.

[111] R. Vinter. *Optimal Control.* Birkhaüser, Boston, 2000.

[112] A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.

[113] J. Warga. Fat homeomorphisms and unbounded derivative containers. *Journal of Mathematical Analysis and Applications*, 81(2):545–560, 1981.

[114] X. Xu and P.J. Antsaklis. Results and Perspectives on Computational Methods for Optimal Control of Switched Systems. In O. Maler and A. Pnueli, editors, *Hybrid Systems: Computation and Control*, volume 2623 of *Lecture Notes in Computer Science*, pages 540–555. Springer, 2003.

[115] M. Yunt and P. I. Barton. Semismooth hybrid automata. In *2006 IEEE Conference on Computer-Aided Control System Design, Vols 1 and 2*, pages 336–341, New York, 2006. IEEE.

[116] J. Zhang, K. H. Johansson, M. Egerstedt, J. Lygeros, and S. Sastry. Zeno hybrid automata. *International Journal of Robust and Nonlinear Control*, 11:435–451, 2001.