Sensitivity Analysis for Nonsmooth Dynamic Systems

by

Kamil Ahmad Khan

B.S.E., Chemical Engineering, Princeton University (2009)
M.S. Chemical Engineering Practice, Massachusetts Institute of Technology (2012)

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

Doctor of Philosophy in Chemical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Februrary 2015

Author . . . . . .
Department of Chemical Engineering
December 16, 2014

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

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

Sensitivity Analysis for Nonsmooth Dynamic Systems

by

Kamil Ahmad Khan

Submitted to the Department of Chemical Engineering
on December 16, 2014, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Chemical Engineering

Abstract

Nonsmoothness in dynamic process models can hinder conventional methods for simulation, sensitivity analysis, and optimization, and can be introduced, for example, by transitions in flow regime or thermodynamic phase, or through discrete changes in the operating mode of a process. While dedicated numerical methods exist for nonsmooth problems, these methods require generalized derivative information that can be difficult to furnish. This thesis presents some of the first automatable methods for computing these generalized derivatives.

Firstly, Nesterov's lexicographic derivatives are shown to be elements of the plenary hull of Clarke's generalized Jacobian whenever they exist. Lexicographic derivatives thus provide useful local sensitivity information for use in numerical methods for nonsmooth problems. A vector forward mode of automatic differentiation is developed and implemented to evaluate lexicographic derivatives for finite compositions of simple lexicographically smooth functions, including the standard arithmetic operations, trigonometric functions, exp / log, piecewise differentiable functions such as the absolute-value function, and other nonsmooth functions such as the Euclidean norm. This method is accurate, automatable, and computationally inexpensive.

Next, given a parametric ordinary differential equation (ODE) with a lexicographically smooth right-hand side function, parametric lexicographic derivatives of a solution trajectory are described in terms of the unique solution of a certain auxiliary ODE. A numerical method is developed and implemented to solve this auxiliary ODE, when the right-hand side function for the original ODE is a composition of absolute-value functions and analytic functions. Computationally tractable sufficient conditions are also presented for differentiability of the original ODE solution with respect to system parameters.

Sufficient conditions are developed under which local inverse and implicit functions are lexicographically smooth. These conditions are combined with the results above to describe parametric lexicographic derivatives for certain hybrid discrete/continuous systems, including some systems whose discrete mode trajectories change when parameters are perturbed.

Lastly, to eliminate a particular source of nonsmoothness, a variant of McCormick's convex relaxation scheme is developed and implemented for use in global optimization methods. This variant produces twice-continuously differentiable convex underestimators for composite functions, while retaining the advantageous computational properties of McCormick's original scheme. Gradients are readily computed for these underestimators using automatic differentiation.

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

Acknowledgments

I could not have asked for more from an advisor than I received from Paul Barton. Beyond his guidance and support, Paul always met my concerns with honest, useful feedback, and my rough drafts with impressive scrutiny. He always gave me the latitude to pursue questions that particularly interested me; without this trust, much of the work in this thesis would not have been possible. Paul, thank you! My thesis committee, Professors Bill Green and Richard Braatz, have also contributed plenty of helpful advice over the years. I am also grateful to Professor Andreas Griewank for his support, and for many helpful discussions.

My PSE labmates at MIT have been great colleagues. I am particularly grateful to Joe, Kai, Mehmet, Stuart, Achim, and Harry for fielding my many questions, for our lively discussions, and for enhancing my understanding of both their own specialties and our broader research discipline.

Outside my professional life, I am grateful to my friends in Australia, in the US, and around the world. I'm glad that we have managed to keep in touch over the years, and I hope that we continue to do so. A special thank you to Jono, Millie, James, and Fiona for making the trek across the Pacific to Boston.

I am fortunate to have had many teachers throughout my life who believed in me, and who went well out of their way to support me on my path. As I have gotten older, I have come to appreciate their contributions to my life more and more. I cannot thank them enough.

Blair, my parents, Suroor, and the Hurley family have always been unwavering in their encouragement, support, and hospitality. Blair, no matter what life throws at us, I hope we can face it together.

The work in this thesis was funded by the MIT-BP Conversion Program, by Novartis Pharmaceuticals as part of the Novartis-MIT Center for Continuous Manufacturing, and by the National Science Foundation under Grant CBET-0933095.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis examines the local behavior of nonsmooth dynamic systems as underlying system parameters are varied, to provide useful sensitivity information to established methods for equation-solving and optimization. The contributions of this thesis include new theoretical results in nonsmooth analysis and the first numerical methods for nonsmooth dynamic sensitivity analysis. Much of the material appearing in this thesis has been published or submitted as the journal articles [54–56, 58–61] and the conference proceedings [53, 57]. The remainder of this chapter elaborates upon the motivation, goal, and contributions of this thesis, and summarizes established methods for approaching these goals.

1.1 Motivation

1.1.1 Nonsmoothness in chemical processes

Nonsmoothness in process systems and models can cause problems for simulation, sensitivity analysis, and optimization. When applied to nonsmooth problems, numerical methods developed for smooth functions can perform poorly, and their theoretical convergence results may no longer apply. Nonsmoothness can be introduced into models of chemical processes through various sources, some of which are listed in this section. As these sources suggest, nonsmoothness in a model of-

ten reflects a qualitative change in the behavior of the underlying system, as time or parameters are varied.

Firstly, transitions in thermodynamic phase or flow regime can require an underlying model to switch discretely between the phases or flow regimes of interest. Examples of transitions in flow regime include transitions between laminar flow and turbulent flow, or the onset of choked flow through a valve.

Discrete transitions in operating regime are a further source of nonsmoothness. Startup of a continuous process – modeled, for example, as in [6] – can involve individual process units being started up only when upstream process units are already sufficiently operational. Process shutdown is analogous. Safety mechanisms can also introduce discrete transitions: activating only when certain measured system variables leave a predetermined set of acceptable values. In processes with cyclic steady states, such as pressure-swing adsorption [105] and simulated moving bed processes [51], each process subunit typically cycles between discrete operating modes.

Embedded optimization problems provide another source of nonsmoothness in process models; the optimal solution values of such problems can be nondifferentiable functions of system parameters or state variables. Approaches to pinch analysis for heat integration [21], for example, represent the minimum heating and cooling demanded by a process in terms of various bivariate $\max\{\cdot,\cdot\}$ functions and a linear program that incorporates information concerning stream flow rates, heat capacities, and desired inlet and outlet temperatures. This approach is also used in the simulation of multi-stream heat exchangers [49]. In *dynamic flux balance analysis* models of bioreactors [36, 43, 44], a linear program that models quasi-steady state cellular metabolism is embedded in a dynamic model of a bioreactor.

The numerical methods applied to a problem may themselves introduce nonsmoothness. For example, established convex relaxation techniques [74, 76, 104] generate lower-bounding information for an objective function, for use in global optimization methods. As discussed in [76] and in Chapter 9, these techniques can introduce nondifferentiability even when the original objective function is smooth.

Methods for extending these relaxation techniques to dynamic systems [102, 103] inherit the original methods' nondifferentiability; these dynamic relaxation methods obtain their lower-bounding information from auxiliary dynamic systems with possibly nondifferentiable dependence on parameters.

1.1.2 Numerical methods for nonsmooth systems

For functions $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ that are locally Lipschitz continuous but not necessarily differentiable everywhere, several notions of *generalized derivatives* have been advanced as analogs of the classical Fréchet derivative. *Bundle methods* for local optimization and *semismooth Newton methods* for equation-solving use these generalized derivatives to construct local approximations of the objective function or residual function under consideration. These methods assume that, at any visited domain point of the objective function or residual function under consideration, both the function and an associated generalized derivative may be evaluated.

Semismooth Newton methods work as follows. Given an open set $X \subset \mathbb{R}^n$ and a locally Lipschitz continuous function $\mathbf{g} : X \to \mathbb{R}^n$ that is *semismooth* [92], consider the problem of determining $\mathbf{x}^* \in X$ for which

$$\mathbf{g}(\mathbf{x}^*) = \mathbf{0}. \tag{1.1}$$

Many nonsmooth functions encountered in practice are semismooth. In particular, all functions that are piecewise differentiable in the sense of Scholtes [97] are semismooth [23], compositions of semismooth functions are semismooth, and solutions of parametric ODEs with semismooth right-hand side functions are themselves semismooth with respect to the ODE parameters [88].

If the residual function $\mathbf{g}$ in (1.1) is twice-continuously differentiable, and if the derivative of $\mathbf{g}$ is nonsingular at some solution $\mathbf{x}^*$ of (1.1), then Newton's well-known method for equation-solving exhibits local Q-quadratic convergence to $\mathbf{x}^*$ [81]. The iterations of the basic Newton method take the following form:

$$\mathbf{x}_{(k+1)} \leftarrow \mathbf{x}_{(k)} + \mathbf{d}, \qquad \text{where} \qquad \mathbf{Jg}(\mathbf{x}_{(k)})\,\mathbf{d} = -\mathbf{g}(\mathbf{x}_{(k)}),$$

where $\mathbf{Jg}(\mathbf{y})$ denotes the (Fréchet) derivative of $\mathbf{g}$ at any $\mathbf{y} \in X$.

When $\mathbf{g}$ is not differentiable everywhere, the essence of a *semismooth Newton method* is to replace the derivative required by the classical Newton method with a *generalized derivative*. Thus, the prototypical Newton iteration above becomes

$$\mathbf{x}_{(k+1)} \leftarrow \mathbf{x}_{(k)} + \mathbf{d}, \qquad \text{where} \qquad \mathbf{Hd} = -\mathbf{g}(\mathbf{x}_{(k)}),$$

where $\mathbf{H}$ is chosen from a suitable set $V(\mathbf{x}_{(k)})$ of generalized derivative candidates. Intuitively, for this method to be useful, $V(\mathbf{x}_{(k)})$ must describe the local behavior of the function $\mathbf{g}$ near $\mathbf{x}_{(k)}$ meaningfully. To use this method, the ability to evaluate the function $\mathbf{g}$ and to compute an appropriate generalized derivative $\mathbf{H}$ are assumed. Several generalized derivative concepts have been formulated for use in nonsmooth numerical methods, including Clarke's generalized Jacobian [16] and its plenary hull [109], the B-subdifferential [92], Nesterov's lexicographic derivatives [79], and Mordukhovich's coderivatives [78]; the generalized derivatives that are pertinent to this thesis are summarized in Section 2.3.

As particular instances of the general semismooth Newton method above, Kojima and Shindo's Newton method [65, Algorithm EN] exhibits local Q-quadratic convergence when $\mathbf{g}$ is piecewise differentiable in the sense of Scholtes [97], $V(\mathbf{x}_{(k)})$ is chosen to be the B-subdifferential of $\mathbf{g}$ at $\mathbf{x}_{(k)}$, and the B-subdifferential of $\mathbf{g}$ at the solution $\mathbf{x}^*$ contains no singular matrices. Qi [91] shows that this method still exhibits local Q-superlinear convergence even if the assumption of piecewise differentiability is relaxed. This local Q-superlinear convergence is retained if Clarke's generalized Jacobian is used in place of the B-subdifferential [92]. A more recent *LP-Newton* method by Facchinei et al. [22] replaces the linear equation system at each iteration with a linear program, and weakens the invertibility requirements of the semismooth Newton method. This method can accommodate certain cases

in which a solution $\mathbf{x}^*$ is not isolated, and in which $\mathbf{x}^*$ is subject to constraints.

While the region of convergence for the classical Newton method may be enlarged significantly by *damping*, semismooth Newton methods cannot be globalized so easily. As shown by Ralph [93], a globalized Newton method for nonsmooth equation-solving would require solution of a *uniform first-order approximation* of the residual function at each iteration. While the first-order Taylor approximation provides such an approximation for a smooth residual function, any uniform first-order approximation of a nondifferentiable function must itself be nondifferentiable, and may therefore be nontrivial to solve.

Bundle methods [63, 67, 70, 71] for local optimization of semismooth functions maintain a bundle of objective function values and associated generalized derivatives computed at previously visited domain points. At each iteration, this bundle is used to construct a piecewise affine local approximation of the objective function; a quadratic program is then constructed and solved to determine a possible solution. If this solution does not satisfy certain necessary optimality conditions, then its corresponding objective function value and an associated generalized derivative are added to the bundle. Recent combinations [50] of bundle methods with cutting-plane methods exhibit global convergence for nonconvex problems.

1.2 Goal

Consider a nonsmooth dynamic process model that is represented as a system of parametric ordinary differential equations (ODEs), with state variables $\mathbf{x}$ and parameters $\mathbf{p}$:

$$\frac{d\mathbf{x}}{dt}(t, \mathbf{p}) = \mathbf{f}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})), \qquad \mathbf{x}(t_0, \mathbf{p}) = \mathbf{x}_0(\mathbf{p}).$$

Classical sensitivity analysis theory for ODEs is presented in [35], and is summarized in Section 2.2. According to this theory, if $\mathbf{f}$ and $\mathbf{x}_0$ are continuously differentiable, then so is any unique solution $\mathbf{x}$ of the above ODE system; moreover,

given any valid choice of parameters $\mathbf{p} := \mathbf{p}_0$, the partial derivative mapping $t \mapsto \frac{\partial \mathbf{x}}{\partial \mathbf{p}}(t, \mathbf{p}_0)$ evolves as the unique solution of a certain auxiliary linear ODE system.

This thesis, however, examines the case in which the functions $\mathbf{f}$ and $\mathbf{x}_0$ are locally Lipschitz continuous, but are not necessarily differentiable everywhere. Though a solution mapping $\mathbf{x}$ for this ODE system will remain continuously differentiable with respect to t in this case, the following example from [55] illustrates that nondifferentiability in $\mathbf{f}$ can lead to nondifferentiability in $\mathbf{x}$ with respect to the parameters $\mathbf{p}$.

Example 1.2.1. *Consider the following parametric ODE, with $c \in \mathbb{R}$ denoting a scalar parameter:*

$$\frac{dx}{dt}(t, c) = |x(t, c)|, \qquad x(0, c) = c.$$

By inspection, this ODE is uniquely solved by the mapping:

$$x : (t, c) \mapsto \begin{cases} c\,\mathrm{e}^{t}, & \text{if } c \geq 0, \\ c\,\mathrm{e}^{-t}, & \text{if } c < 0, \end{cases}$$

which is plotted in Figure 1-1. This expression for x shows that, for any fixed $t \neq 0$, the mapping $x(t, \cdot)$ is continuous but not differentiable at 0.

Given a nonsmooth parametric ODE system, the central goal of this thesis is the development and implementation of numerical methods for computing generalized derivatives for the solution of this ODE with respect to system parameters, for use in methods for optimization and equation-solving. The ODE right-hand side function is assumed to be a finite composition of known simple differentiable and nondifferentiable functions; beyond this, little *a priori* knowledge concerning the behavior of the dynamic system is assumed.

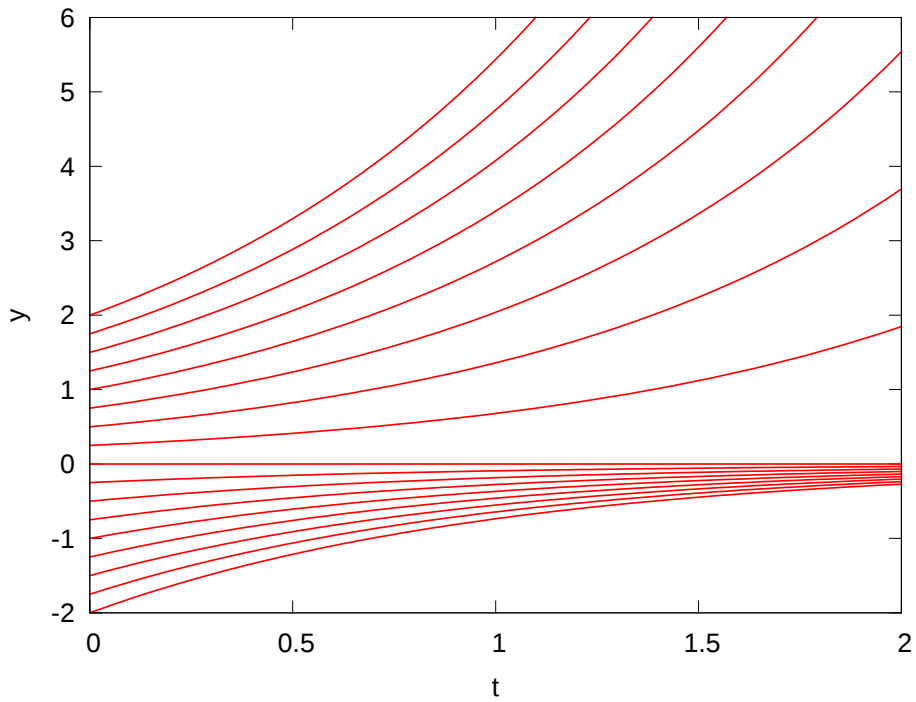

Figure 1-1: The ODE solution $y = x(t, c)$ described in Example 1.2.1, plotted against t for various values of $c \in [-2, 2]$.

1.3 Existing approaches

Since calculus rules [16, 78] for the well-known generalized derivatives hold only as inclusions rather than as equations, and since these inclusions may be strict [23], existing methods to evaluate generalized derivatives for known composite functions are limited. As summarized in [61], however, elements of the Clarke Jacobian have been computed analytically in certain special cases, for use in semismooth Newton methods. For example, the primal-dual active set methods of Hintermüller et al. [40] compute certain Clarke Jacobian elements for a nonsmooth equation system representing necessary optimality conditions for a quadratic program. This approach has been extended and applied successfully in certain large optimal control problems [39, 106] and data-fitting problems involving L^1-norms of functions [17]. Ulbrich [115] extends the underlying theory of semismooth Newton methods to general function spaces, to solve appropriate reformulations of certain variational inequalities: again computing the required Clarke Jacobian elements

analytically. Mordukhovich's coderivative has also been evaluated successfully for certain problems [7, 38]. As shown by Griewank [32], the classical *forward mode of automatic differentiation* continues to evaluate directional derivatives for a broad class of composite nondifferentiable functions. For semismooth functions, directional derivatives are linear combinations of the columns of certain generalized derivatives [92, Lemma 2.2].

Ralph's approach [93] for globalizing semismooth Newton methods, outlined in Section 1.1.2, has been implemented successfully for *mixed complementarity problems* in the PATH solver [19, 25]. A recent approach by Griewank [33] applies to a more general class of nonsmooth residual functions, and involves constructing uniform first-order approximations that are piecewise affine in the sense of Scholtes [97], using a method reminiscent of the forward mode of automatic differentiation.

As is well-known in convex analysis [42], nondifferentiable convex functions exhibit many useful calculus properties that their nonconvex counterparts lack. For any convex function on an open set, Clarke's generalized Jacobian is identical to the convex subdifferential; when the integrand in a parametric integral is convex with respect to parameters, the parametric subdifferential of the resulting integral is exactly the integral of the parametric subdifferential of the integrand [16]. A similar result holds for solutions of parametric ODE systems with convex right-hand side functions. Example 5.2.7 in Chapter 5, however, shows that this property of convex subdifferentials is not retained by Clarke's generalized Jacobian once nonconvexity is introduced. If a convex function of parameters is described as the solution of a parametric ODE with a nonconvex right-hand side function, as in the relaxation theory of [102], then the subdifferential of the ODE solution is not necessarily the solution of an auxiliary ODE in which the original nonconvex right-hand side function is replaced by its Clarke Jacobian. Moreover, these results do not naturally extend to compositions of well-behaved functions; the difference or composition of two convex functions, for example, is not necessarily convex.

As a key result in nonsmooth dynamic sensitivity analysis, Clarke [16, Theo-

rem 7.4.1] provides a sufficient condition for parametric differentiability of solutions of parametric ODEs, which was extended by Yunt [124] to certain hybrid discrete/continuous systems and differential-algebraic equation systems. Clarke's sufficient condition considers the times at which an ODE solution visits domain points at which the ODE right-hand side function is nondifferentiable. If the set of all such times has zero Lebesgue measure, then the ODE solution is in fact differentiable with respect to its parameters at that particular parameter value, and its corresponding parametric derivative solves a certain auxiliary linear ODE system. As shown in Chapter 6, this sufficient condition for differentiability becomes tractable to verify numerically for cases in which the ODE right-hand side is a finite composition of absolute-value functions and analytic functions. Established sensitivity analysis results for parametric hybrid discrete/continuous systems [30] also provide sufficient conditions for differentiability. These conditions require the sequence of visited discrete modes to be independent of the parameters, and the timing of each discrete event to be a well-defined implicit function of the parameters.

When Clarke's sufficient condition for differentiability is not satisfied by the solution of a parametric ODE, Pang and Stewart [88] provide the lone established result concerning parametric generalized derivatives of this solution. Pang and Stewart describe *linear Newton approximations* for such systems as the solutions of auxiliary linear ODE systems in which the original ODE right-hand side function is replaced by its parametric Clarke Jacobian. As shown in Example 5.1.1 in Chapter 5, however, linear Newton approximations do not necessarily share the desirable properties of the generalized derivatives studied in this thesis. In particular, linear Newton approximations for continuously differentiable functions or convex functions do not necessarily reduce to the derivative or the subdifferential, respectively; as a result, sufficient optimality conditions cannot be formulated in terms of linear Newton approximations.

As noted in [4], for example, it is possible to solve certain nonsmooth problems by instead solving a sequence of smooth problems which converges in some sense

to the problem of interest; this approach has no use for generalized derivatives, and is not pursued further in this thesis. By discretizing the independent variable in an ODE system, the solution of this system may be approximated as a large equation system, in which case dedicated sensitivity analysis results for dynamic systems would not be necessary. This approach, in essence, combines the original ODE system with a particular ODE solution method used to integrate it. However, with time discretized *a priori*, such an approach cannot take advantage of integration techniques such as adaptive time-stepping with error control, and may not be well-suited to stiff dynamic systems.

1.4 Contributions and thesis structure

The main contribution of this thesis is the development of the first numerical method for evaluating generalized derivatives for nonsmooth dynamic systems. To obtain this method, several theoretical results and incidental numerical methods were developed that are contributions in their own right. This section briefly summarizes the contents and contributions of each chapter of this thesis. As noted earlier, much of the material appearing in this thesis has been published or submitted as the journal articles [54–56, 58–61] and the conference proceedings [53, 57].

Chapter 2 summarizes the established mathematical concepts underlying the results and methods in the subsequent chapters. These concepts include basic notions of differentiability, classical results from the theory of ordinary differential equations, various formulations of *generalized derivatives* for nonsmooth functions, results concerning functions that are *piecewise differentiable* in the sense of Scholtes [97], and a description of the vector forward mode of automatic differentiation. Notational conventions used throughout this thesis are also described.

Chapter 3 develops new relationships between various generalized derivatives, and is collected from the articles [55, 61]. A new generalized derivative, the *LD-derivative*, is developed as a variant of Nesterov's lexicographic derivative [79] that satisfies a particularly simple variant of Nesterov's chain rule. Moreover, lexico-

graphic derivatives are readily computed from LD-derivatives. It is shown that lexicographic derivatives, whenever they exist, are elements of the plenary hull of Clarke's generalized Jacobian; this result generalizes a similar result by Nesterov [79] concerning scalar-valued functions. This result is strengthened for functions that are piecewise differentiable [97]; it is shown that all such functions are lexicographically smooth, with lexicographic derivatives that are always elements of the B-subdifferential.

In Chapter 4, the chain rule for LD-derivatives from Chapter 3 is exploited to yield a vector forward mode of automatic differentiation for lexicographic derivative evaluation. This chapter is reproduced from [61]. The developed method applies to finite compositions of lexicographically smooth *elemental functions*; these elemental functions can include smooth functions such as the standard arithmetic and trigonometric functions, piecewise differentiable functions such as the absolute-value function and the bivariate $\max\{\cdot,\cdot\}$ function, and other nonsmooth functions such as the Euclidean norm. This method is implemented in C++, and is computationally tractable, accurate, and automatable. This method is, essentially, an improved version of our first developed method [53, 54] for generalized derivative evaluation for vector-valued composite nonsmooth functions; this earlier method is reproduced in Appendix A for reference.

Chapter 5, reproduced from [55], considers parametric ordinary differential equation (ODE) systems, with right-hand side functions that are lexicographically smooth with respect to the differential variables. Lexicographic derivatives of a unique solution of such an ODE with respect to the ODE parameters are described in terms of the unique solution of a certain auxiliary ODE system, in an analogous manner to the development of classical sensitivity analysis [35, Ch. V] for ODEs with smooth right-hand side functions. Though the auxiliary ODE is guaranteed to have a unique solution, it does not necessarily satisfy the Carathéodory assumptions [26]. To our knowledge, this is the first description of a useful generalized derivative for a unique solution of a parametric nonsmooth ODE system. As an intermediate result, a result by Pang and Stewart [88] is generalized to describe

directional derivatives of the ODE solution with respect to the ODE parameters.

Chapter 6, reproduced from [59], considers the switching behavior of solutions of ODEs whose right-hand side functions are compositions of analytic functions and absolute-value functions. These ODE solutions are found to exhibit *non-Zenoness* [29, 48, 108], in that no absolute-value function in the ODE right-hand side function may switch between its two linear pieces infinitely often in any finite duration. This non-Zenoness persists even when a discontinuous control input is included, provided that this control input satisfies a certain *left/right-analyticity* property. The obtained non-Zenoness results are used to obtain a tractable formulation of Clarke's sufficient conditions [16, Theorem 7.4.1] for differentiability of a unique solution of a parametric ODE with a nondifferentiable right-hand side function. The obtained sufficient conditions can be tested during numerical integration of the ODE, and can, in certain cases, be verified to hold *a priori*.

Chapter 7, reproduced from [58], combines the main results of Chapters 5 and 6, to obtain a numerical method for evaluating parametric lexicographic derivatives for the solutions of parametric ODEs with right-hand side functions that are compositions of analytic functions and absolute-value functions. Though these lexicographic derivatives were described in terms of a non-Carathéodory ODE in Chapter 5, the non-Zenoness theory of Chapter 6 is exploited to reformulate this non-Carathéodory ODE as an equivalent hybrid discrete/continuous system. Theoretical properties of this hybrid system are obtained and exploited in the developed method. This is the first numerical method to evaluate generalized derivatives for general nonsmooth parametric ODEs.

Chapter 8, reproduced from [56], develops conditions under which local inverse and implicit functions are lexicographically smooth, and describes their LD-derivatives. These results are then combined with the results of Chapter 5 to present LD-derivatives for the hybrid discrete/continuous systems considered by Galán et al. [30]. Unlike the development in [30], however, the functions describing a solution's continuous evolution and discrete transitions are now permitted to be merely lexicographically smooth rather than continuously differentiable. With

this relaxation, even certain hybrid systems with varying discrete mode sequences can be shown to have solutions that are lexicographically smooth with respect to parameters. Moreover, the LD-derivatives of these solutions can be described.

In a departure from the chapters described above, Chapter 9 eliminates a particular source of nondifferentiability, and is reproduced from [60]. A variant of McCormick's scheme [74] for generating convex underestimators for composite functions is developed and implemented, in which the obtained convex relaxations are guaranteed to be twice-continuously differentiable, without sacrificing any of the useful properties of McCormick's original relaxation scheme. The modified relaxations can still be computed cheaply and accurately, and converge rapidly to the function they relax as the underlying parameter interval is reduced in width. Moreover, gradients can be obtained for these relaxations using automatic differentiation. These relaxations and their gradients are evaluated using a modification of the C++ library MC++ [15]. While McCormick's original relaxations are not guaranteed to be differentiable, the modified relaxations are amenable to treatment by numerical methods developed for twice-continuously differentiable functions.

Chapter 2

Mathematical background

This section describes relevant established mathematical concepts underlying the material in this thesis, and is largely reproduced from the background sections in the articles [54–56, 58, 59, 61]. This material covers concepts regarding ordinary differential equations, generalized derivatives for nonsmooth functions, Scholtes' piecewise differentiable functions [97], and automatic differentiation. Notational conventions used in this thesis are also described.

In addition to the material presented in this section, Chapter 8 presents further background information concerning *hybrid discrete/continuous systems*, and Chapter 9 presents further background information concerning McCormick's convex relaxation scheme [74] and the *generalized McCormick* theory of [100, 104].

2.1 Notation and basic concepts

Notational conventions used throughout this article are as follows; this section is largely reproduced from [61]. The vector space $\mathbb{R}^n$ is endowed with the usual Euclidean norm $\|\cdot\|$ and inner product $\langle\cdot,\cdot\rangle$, and the vector space $\mathbb{R}^{m\times n}$ of matrices is endowed with the corresponding induced norm. The *column space* of a matrix $\mathbf{M} \in \mathbb{R}^{n\times p}$ is defined as the set $\mathcal{R}(\mathbf{M}) := \{\mathbf{M}\mathbf{v} : \mathbf{v} \in \mathbb{R}^p\} \subset \mathbb{R}^n$. Elements of $\mathbb{R}$ and scalar-valued functions are denoted as lowercase letters (e.g. m), vectors in $\mathbb{R}^n$ and vector-valued functions are denoted as lowercase boldface letters (e.g. $\mathbf{m}$),

matrices in $\mathbb{R}^{n\times m}$ are denoted as uppercase boldface letters (e.g. $\mathbf{M}$), and sets are denoted as uppercase letters (e.g. M). For notational compactness, a well-defined vertical block matrix (or vector):

$$\begin{bmatrix}\mathbf{A}\\ \mathbf{B}\end{bmatrix}$$

will frequently be denoted as $(\mathbf{A}, \mathbf{B})$. The ith component of a vector $\mathbf{v}$ is denoted as v_i. The kth column of a matrix $\mathbf{M}$ is denoted with a parenthetical subscript as $\mathbf{m}_{(k)}$, whose i^{th} component is $m_{(k),i}$; a similar parenthetical subscript indicates a vector evaluated during the kth iteration of an algorithm, or the kth element of a sequence of vectors. Parenthetical superscripts (e.g. $\mathbf{f}^{(k)}$) are reserved for lexicographic differentiation, which is described in Section 2.3.3. $\mathbf{0}$ denotes a zero matrix or vector, and $\mathbf{I}$ denotes a square identity matrix. When the dimensions of $\mathbf{0}$ or $\mathbf{I}$ are unclear from the context, these will be noted in a subscript as, for example, $\mathbf{0}_m \in \mathbb{R}^m$, $\mathbf{0}_{m\times n} \in \mathbb{R}^{m\times n}$, or $\mathbf{I}_{n\times n} \in \mathbb{R}^{n\times n}$. The columns of $\mathbf{I}_{n\times n}$ are the *unit coordinate vectors*, and are denoted $\mathbf{e}_{(1)}, \ldots, \mathbf{e}_{(n)}$.

In the inductive proofs in this thesis, it will be convenient to refer to an *empty matrix* $\varnothing_{n\times 0}$ of real numbers, with n rows but no columns. In a further abuse of notation, the set $\{\varnothing_{n\times 0}\}$ will be denoted $\mathbb{R}^{n\times 0}$. No operations will be performed on $\varnothing_{n\times 0}$ beyond concatenation, which proceeds as expected:

$$\begin{bmatrix}\mathbf{A} & \varnothing_{n\times 0}\end{bmatrix} = \begin{bmatrix}\varnothing_{n\times 0} & \mathbf{A}\end{bmatrix} := \mathbf{A}, \qquad \forall \mathbf{A} \in \mathbb{R}^{n\times m}, \quad \forall m \in \mathbb{N}\cup\{0\}.$$

Given a collection of vectors $\mathbf{v}_{(1)}, \mathbf{v}_{(2)}, \ldots, \mathbf{v}_{(p)} \in \mathbb{R}^n$, $\begin{bmatrix}\mathbf{v}_{(1)} & \cdots & \mathbf{v}_{(j)}\end{bmatrix} \in \mathbb{R}^{n\times j}$ will denote $\varnothing_{n\times 0}$ when $j = 0$. Similarly, it will be convenient at times to refer to an *empty vector* $\varnothing_0 \in \mathbb{R}^0$ of real numbers, with no components.

If a function $\mathbf{f} : X \to Y$ satisfies some local property P at each $\mathbf{x} \in X$, then $\mathbf{f}$ will be said simply to satisfy P, without reference to any particular $\mathbf{x} \in X$.

The *convex hull*, *linear hull*, *closure*, and *interior* of a set $S \subset \mathbb{R}^n$ are denoted as $\operatorname{conv} S$, $\operatorname{span} S$, $\operatorname{cl}(S)$, and $\operatorname{int}(S)$, respectively. The *boundary* of $S \subset \mathbb{R}^n$ is the set $\operatorname{cl}(S)\backslash\operatorname{int}(S)$. If $S \subset \mathbb{R}^n$ is nonempty, then the *convex cone* generated by S is the set

$$\operatorname{cone} S := \left\{ \sum_{i=1}^{p} \lambda_i \mathbf{x}_{(i)} : p \in \mathbb{N}, \quad \mathbf{x}_{(i)} \in S,\ \lambda_i \geq 0,\ \forall i \in \{1, \ldots, p\} \right\} \subset \mathbb{R}^n.$$

If S is finite, then cone S is a *polyhedral cone*, and $|S|$ denotes the number of elements of S. A *conical subdivision* of $\mathbb{R}^n$ is a finite collection Λ of distinct polyhedral cones in $\mathbb{R}^n$ with nonempty interiors, such that

$$\bigcup_{C \in \Lambda} C = \mathbb{R}^n, \qquad \text{and} \qquad \operatorname{int}(C_1) \cap \operatorname{int}(C_2) = \emptyset,\ \forall C_1, C_2 \in \Lambda \text{ s.t. } C_1 \neq C_2.$$

2.1.1 Differentiability and directional differentiability

This section presents basic notions of differentiability of a function $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ at a domain point $\mathbf{x} \in X$.

Definition 2.1.1. *Consider an open set $X \subset \mathbb{R}^n$, some $\mathbf{x} \in X$, and a function $\mathbf{f} : X \to \mathbb{R}^m$. The following limit, if it exists, is the* directional derivative *of $\mathbf{f}$ at $\mathbf{x}$ in the direction $\mathbf{d} \in \mathbb{R}^n$:*

$$\mathbf{f}'(\mathbf{x}; \mathbf{d}) := \lim_{t \to 0^+} \frac{\mathbf{f}(\mathbf{x} + t\mathbf{d}) - \mathbf{f}(\mathbf{x})}{t}.$$

If $\mathbf{f}'(\mathbf{x}; \mathbf{d})$ exists in $\mathbb{R}^m$ for each $\mathbf{d} \in \mathbb{R}^n$, then $\mathbf{f}$ is directionally differentiable *at $\mathbf{x}$.*

As summarized by Scholtes [97], if $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ is directionally differentiable, then $\mathbf{f}'(\mathbf{x}; \cdot)$ is positively homogeneous for each $\mathbf{x} \in X$. If, in addition, $\mathbf{f}$ is locally Lipschitz continuous on its domain, then

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{\mathbf{f}(\mathbf{x} + \mathbf{h}) - (\mathbf{f}(\mathbf{x}) + \mathbf{f}'(\mathbf{x}; \mathbf{h}))}{\|\mathbf{h}\|} = \mathbf{0}, \qquad \forall \mathbf{x} \in X; \tag{2.1}$$

thus, the mapping $\mathbf{h} \mapsto \mathbf{f}(\mathbf{x}) + \mathbf{f}'(\mathbf{x}; \mathbf{h})$ approximates $\mathbf{f}$ well near $\mathbf{x}$. In this case, for any fixed $\mathbf{x} \in X$, the function $\mathbf{f}'(\mathbf{x}; \cdot)$ is Lipschitz continuous on $\mathbb{R}^n$. Functions that are both locally Lipschitz continuous and directionally differentiable satisfy the following useful chain rule.

Proposition 2.1.2 (Theorems 3.1.1 and 3.1.2 in [97]). *Consider open sets $X \subset \mathbb{R}^n$ and $Z \subset \mathbb{R}^p$, and functions $\mathbf{g} : Z \to X$ and $\mathbf{f} : X \to \mathbb{R}^m$ that are locally Lipschitz continuous and directionally differentiable at $\mathbf{z} \in Z$ and $\mathbf{g}(\mathbf{z})$, respectively. The composite function $\mathbf{f} \circ \mathbf{g}$ is then also locally Lipschitz continuous and directionally differentiable at $\mathbf{z}$, and satisfies*

$$[\mathbf{f} \circ \mathbf{g}]'(\mathbf{z}; \mathbf{d}) = \mathbf{f}'(\mathbf{g}(\mathbf{z}); \mathbf{g}'(\mathbf{z}; \mathbf{d})), \qquad \forall \mathbf{d} \in \mathbb{R}^p.$$

Definition 2.1.3. *Consider an open set $X \subset \mathbb{R}^n$, some $\mathbf{x} \in X$, and a function $\mathbf{f} : X \to \mathbb{R}^m$. The function $\mathbf{f}$ is* (Fréchet) differentiable *at $\mathbf{x} \in X$ if there exists a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ that satisfies:*

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

In this case, $\mathbf{A}$ is uniquely described by the above equation, and is called the (Fréchet) derivative *or* Jacobian matrix $\mathbf{J}\mathbf{f}(\mathbf{x}) \in \mathbb{R}^{m \times n}$. *If $m = 1$, in which case $\mathbf{f} \equiv f$ is scalar-valued, then the* gradient *of f at $\mathbf{x}$ is $\nabla f(\mathbf{x}) := (\mathbf{J}f(\mathbf{x}))^{\mathrm{T}} \in \mathbb{R}^n$.*

The function $\mathbf{f}$ is continuously differentiable ($\mathcal{C}^1$) *at $\mathbf{x}$ if there exists a neighborhood $N \subset X$ of $\mathbf{x}$ for which $\mathbf{f}$ is differentiable at each $\mathbf{y} \in N$, and for which the Jacobian mapping $\mathbf{y} \mapsto \mathbf{J}\mathbf{f}(\mathbf{y})$ is continuous at $\mathbf{x}$.*

Standard notation for *partial derivatives* is used; for example, given a mapping $(\mathbf{x}, \mathbf{z}) \mapsto \mathbf{g}(\mathbf{x}, \mathbf{z})$ that is differentiable at $(\bar{\mathbf{x}}, \bar{\mathbf{z}})$, the partial derivative $\frac{\partial \mathbf{g}}{\partial \mathbf{x}}(\bar{\mathbf{x}}, \bar{\mathbf{z}})$ denotes the derivative $\mathbf{J}[\mathbf{g}(\cdot, \bar{\mathbf{z}})](\bar{\mathbf{x}})$.

Any $\mathcal{C}^1$ function is also locally Lipschitz continuous. If $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\mathbf{x} \in X$, then $\mathbf{f}$ is also directionally differentiable at $\mathbf{x}$, with

$$\mathbf{f}'(\mathbf{x}; \mathbf{d}) = \mathbf{J}\mathbf{f}(\mathbf{x})\, \mathbf{d}, \qquad \forall \mathbf{d} \in \mathbb{R}^n.$$

As considered in [16, 82, 83], for example, a function $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ is *Gâteaux differentiable* at $\mathbf{x} \in X$ if the directional derivative mapping $\mathbf{d} \mapsto \mathbf{f}'(\mathbf{x}; \mathbf{d})$ is well-defined and linear on $\mathbb{R}^n$. If $\mathbf{f}$ is locally Lipschitz continuous, however, then Fréchet differentiability of $\mathbf{f}$ at $\mathbf{x}$ is equivalent to Gâteaux differentiability of

$\mathbf{f}$ at $\mathbf{x}$ [16]. This thesis is concerned primarily with locally Lipschitz continuous functions; there will be no need to consider any distinction between Gâteaux and Fréchet differentiability.

2.1.2 Analytic functions

The class of *analytic functions* is described at length in [66], and is defined below.

Definition 2.1.4 (adapted from Definition 1.1.5 in [66])**.** *Given an open set $T \subset \mathbb{R}$, a scalar-valued function f is* (real-)analytic ($\mathcal{C}^\omega$) *at $t^* \in T$ if there exists a neighborhood $N \subset T$ of t^* and constants $\{a_k\}_{k=0}^{\infty}$ in $\mathbb{R}$ for which the power series*

$$\sum_{k=0}^{\infty} a_k (t - t^*)^k$$

converges to $f(t)$ for each $t \in N$. A function $\mathbf{f} : T \to \mathbb{R}^m$ is analytic at t^* if each of its component functions $f_1, \ldots, f_m : T \to \mathbb{R}$ is analytic at t^*.

Observe that the basic arithmetic operations, trigonometric functions, power functions, and logarithmic functions are each $\mathcal{C}^\omega$ on the interiors of their respective domains. If a function f is $\mathcal{C}^\omega$ at t^*, then f is also $\mathcal{C}^\omega$ on some neighborhood of t^* [66, Corollary 1.2.4]. A well-defined composition of two $\mathcal{C}^\omega$ functions is itself $\mathcal{C}^\omega$ [66, Proposition 2.2.8]. The following elementary property of $\mathcal{C}^\omega$ functions will be exploited in Chapters 6 and 7.

Proposition 2.1.5 (adapted from Corollary 1.2.6 in [66])**.** *Consider an open set $T \subset \mathbb{R}$ and a $\mathcal{C}^\omega$ function $f : T \to \mathbb{R}$. If there exists a nonempty open set $U \subset T$ for which*

$$f(t) = 0, \qquad \forall t \in U,$$

then f is the zero mapping on T.

2.1.3 Set-valued mappings

As described in [3, 23], a *set-valued mapping* $F : Y \rightrightarrows Z$ is a function that maps each element of Y to a subset of Z. Suppose that $Y \subset \mathbb{R}^n$ is open and $Z = \mathbb{R}^m$. In this

case, f is *upper-semicontinuous* at $\mathbf{y} \in Y$ if, for each $\epsilon > 0$, there exists $\delta > 0$ such that whenever $\|\mathbf{z}\| < \delta$,

$$F(\mathbf{y} + \mathbf{z}) \subset F(\mathbf{y}) + \{\mathbf{v} \in \mathbb{R}^m : \|\mathbf{v}\| < \epsilon\}.$$

If F is upper-semicontinuous at $\mathbf{y} \in Y$, then given any convergent sequences $\{\mathbf{y}_{(i)}\}_{i\in\mathbb{N}}$ in Y and $\{\mathbf{z}_{(i)}\}_{i\in\mathbb{N}}$ in $\mathbb{R}^m$ such that $\lim_{i\to\infty} \mathbf{y}_{(i)} = \mathbf{y}$, $\lim_{i\to\infty} \mathbf{z}_{(i)} = \mathbf{z}$, and $\mathbf{z}_{(i)} \in F(\mathbf{y}_{(i)})$ for each $i \in \mathbb{N}$, it follows that $\mathbf{z} \in F(\mathbf{y})$.

2.2 Ordinary differential equations

This section summarizes relevant results concerning systems of ordinary differential equations (ODEs). Familiarity with ODEs is assumed; for further details, the reader is pointed to the texts [18, 26, 35].

Consider the following ODE system, whose right-hand side $\mathbf{g} : \mathbb{R} \times \mathbb{R}^{n_p} \times \mathbb{R}^n \to \mathbb{R}^n$ and initial condition depend directly on parameters $\mathbf{p}$.

$$\frac{d\mathbf{z}}{dt}(t, \mathbf{p}) = \mathbf{g}(t, \mathbf{p}, \mathbf{z}(t, \mathbf{p})), \qquad \mathbf{z}(t_0, \mathbf{p}) = \mathbf{z}_0(\mathbf{p}). \tag{2.2}$$

Throughout this thesis, the *solution* of such an ODE at $\mathbf{p} := \mathbf{p}_0$ refers to a solution in the Carathéodory sense, as summarized in [26]. Thus, a mapping $\tilde{\mathbf{z}}(\cdot, \mathbf{p}_0)$ solves the above ODE at $\mathbf{p} := \mathbf{p}_0$ on $[t_0, t_f]$ if and only if both $\tilde{\mathbf{z}}(t_0, \mathbf{p}_0) = \mathbf{z}_0(\mathbf{p}_0)$ and

$$\frac{d\tilde{\mathbf{z}}}{dt}(t, \mathbf{p}_0) = \mathbf{g}(t, \mathbf{p}_0, \tilde{\mathbf{z}}(t, \mathbf{p}_0))$$

for almost every $t \in [t_0, t_f]$ with respect to Lebesgue measure. Equivalently, $\tilde{\mathbf{z}}(\cdot, \mathbf{p}_0)$ solves the above ODE at $\mathbf{p} := \mathbf{p}_0$ on $[t_0, t_f]$ if and only if

$$\tilde{\mathbf{z}}(t) = \mathbf{z}_0(\mathbf{p}_0) + \int_{t_0}^{t} \mathbf{g}(s, \mathbf{p}_0, \tilde{\mathbf{z}}(s, \mathbf{p}_0))\, ds, \qquad \forall t \in [t_0, t_f].$$

Here, and throughout this thesis, the integral is understood to be a Lebesgue integral.

Conditions for the local existence and uniqueness of ODE solutions are presented in [18, 26, 35]. In particular, existence and uniqueness of a solution do not

require the ODE right-hand side $\mathbf{g}$ to be continuous with respect to its t argument. Neglecting the influence of the parameter $\mathbf{p}$, consider the simplified ODE:

$$\frac{d\mathbf{y}}{dt}(t) = \mathbf{h}(t, \mathbf{y}(t)), \qquad \mathbf{y}(t_0) = \mathbf{y}_0. \tag{2.3}$$

Suppose that $\mathbf{h}$ is defined on some open superset D of $(t_0, \mathbf{y}_0)$, and satisfies all of the following conditions, as presented in [26]:

- $\mathbf{h}(t, \cdot)$ is well-defined and continuous for almost all t,
- $\mathbf{h}(\cdot, \boldsymbol{\eta})$ is measurable for each $\boldsymbol{\eta}$, and
- there exists an integrable function $m_{\mathbf{h}}$ for which, for all $\boldsymbol{\eta}$, $\|\mathbf{h}(t, \boldsymbol{\eta})\| \leq m_{\mathbf{h}}(t)$.

Under these *Carathéodory existence conditions*, there exists a solution $\mathbf{y}$ of the ODE (2.3) on some neighborhood of t_0.

Next, suppose that there exists an integrable function $k_{\mathbf{h}}$ for which, for all t, $\boldsymbol{\eta}_A$, and $\boldsymbol{\eta}_B$ for which $(t, \boldsymbol{\eta}_A), (t, \boldsymbol{\eta}_B) \in D$,

$$\|\mathbf{h}(t, \boldsymbol{\eta}_A) - \mathbf{h}(t, \boldsymbol{\eta}_B)\| \leq k_{\mathbf{h}}(t)\, \|\boldsymbol{\eta}_A - \boldsymbol{\eta}_B\|.$$

Under this condition, any solution of (2.3) is unique [26]. When this condition is combined with the Carathéodory existence conditions, these conditions will together be referred to as the *Carathéodory existence and uniqueness* conditions.

In the results in this thesis concerning ODEs, it will often be assumed explicitly that an ODE solution exists on a given duration $[t_0, t_f]$. Under the Carathéodory existence and uniquenss assumptions, this solution is unique, and can be extended to yield a unique solution of the ODE on some open superset of $[t_0, t_f]$ [18, 35].

Returning now to the parametric ODE (2.2), suppose that the behavior of the state variables $\mathbf{z}$ with respect to parameters $\mathbf{p}$ is under investigation. Roughly, if the right-hand side function $\mathbf{g}$ is well-behaved with respect to its first and second arguments, then these arguments can be appended to $\mathbf{z}$ as extra ODE state variables: the first with a time-derivative of unity, and the second with a time-derivative of $\mathbf{0}_{n_p}$. Moreover, the influence of the initial-condition mapping $\mathbf{z}_0$ may

be handled *a posteriori* using an appropriate chain rule. With these modifications in mind, it suffices for our purposes to consider the simpler parametric ODE system:

$$\frac{d\mathbf{x}}{dt}(t,\mathbf{c}) = \mathbf{f}(\mathbf{x}(t,\mathbf{c})), \qquad \mathbf{x}(t_0,\mathbf{c}) = \mathbf{c}, \tag{2.4}$$

in which the parameter is the initial condition $\mathbf{c}$. In general, this modification simplifies the notation and presentation of relevant results in this thesis considerably. In the theoretical development in Chapter 5, however, it will be convenient to permit $\mathbf{g}$ to be discontinuous with respect to t, in which case any direct dependence of $\mathbf{g}$ with respect to t will not be ascribed to an augmented state variable instead. Several results in Chapter 6 will also consider situations in which $\mathbf{g}$ is discontinuous with respect to t.

2.2.1 Classical sensitivity analysis

Sensitivity analysis theory for parametric ODEs with $\mathcal{C}^1$ right-hand side functions is well-established, and is described, for example, by Hartman [35, Ch. V]. This classical theory serves as a useful analog of the theory developed in this thesis; any results concerning parametric generalized derivatives for ODEs with nondifferentiable right-hand side functions should, intuitively, reduce to the classical case when the right-hand side is $\mathcal{C}^1$.

As brief overview of this classical sensitivity theory, consider the following parametric ODE:

$$\frac{d\mathbf{x}}{dt}(t,\mathbf{c}) = \mathbf{f}(t,\mathbf{x}(t,\mathbf{c})), \qquad \mathbf{x}(t_0,\mathbf{c}) = \mathbf{c}, \tag{2.5}$$

with a $\mathcal{C}^1$ right-hand side function $\mathbf{f}$. Suppose that when $\mathbf{c} := \mathbf{c}_0$, there exists a unique ODE solution $\mathbf{x}(\cdot,\mathbf{c}_0)$ on $[t_0,t_f]$. Then, there also exists a unique solution $\mathbf{x}(\cdot,\mathbf{c})$ on $[t_0,t_f]$ for each $\mathbf{c}$ in some neighborhood of $\mathbf{c}_0$. Moreover, for each fixed $t \in [t_0,t_f]$, $\mathbf{x}(t,\cdot)$ is $\mathcal{C}^1$ at $\mathbf{c}_0$; the partial derivative mapping $t \mapsto \frac{\partial \mathbf{x}}{\partial \mathbf{c}}(t,\mathbf{c}_0)$ is the unique solution $\mathbf{A}$ on $[t_0,t_f]$ of the following linear ODE:

$$\frac{d\mathbf{A}}{dt}(t) = \frac{\partial \mathbf{f}}{\partial \mathbf{x}}(t, \mathbf{x}(t, \mathbf{c}))\, \mathbf{A}(t), \qquad \mathbf{A}(t_0) = \mathbf{I}.$$

This latter sensitivity ODE can be solved simultaneously with the original ODE in $\mathbf{x}$ to evaluate the partial derivative $\frac{\partial \mathbf{x}}{\partial \mathbf{c}}(t, \mathbf{c}_0)$. Established implicit integration methods [24, 73] perform this simultaneous integration accurately and efficiently, by exploiting an inherent redundancy and sparsity in the derivatives used in their corrector iterations.

The following result by Clarke [16] provides a sufficient condition for differentiability of the solution $\mathbf{x}$ of the ODE (2.5) with respect to the initial condition $\mathbf{c}$, even if the right-hand side function $\mathbf{f}$ is not differentiable everywhere.

Proposition 2.2.1 (adapted from Theorem 7.4.1 in [16])**.** *Suppose that the right-hand side function* $\mathbf{f}$ *of the ODE* (2.5) *is locally Lipschitz continuous, and that there exists a unique solution* $\mathbf{x}(\cdot, \mathbf{c}_0)$ *of* (2.5) *on* $[t_0, t_f]$*. Let* S *be the set on which* $\mathbf{f}$ *is differentiable. If the set*

$$\{t \in [t_0, t_f] : (t, \mathbf{x}(t, \mathbf{c}_0)) \notin S\}$$

has zero Lebesgue measure, then $\mathbf{x}(t^*, \cdot)$ *is differentiable at* $\mathbf{c}_0$ *for each* $t^* \in [t_0, t_f]$*. Moreover,* $\frac{\partial \mathbf{x}}{\partial \mathbf{c}}(t^*, c_0) = \mathbf{A}(t^*)$*, where* $\mathbf{A}$ *denotes the unique solution of the linear ODE:*

$$\frac{d\mathbf{A}}{dt}(t) = \begin{cases} \frac{\partial \mathbf{f}}{\partial \mathbf{x}}(t, \mathbf{x}(t, \mathbf{c}_0))\, \mathbf{A}(t), & \text{if } (t, \mathbf{x}(t, \mathbf{c}_0)) \in S, \\ \mathbf{0}, & \text{if } (t, \mathbf{x}(t, \mathbf{c}_0)) \notin S. \end{cases}$$

2.3 Generalized derivatives

Several concepts of generalized derivatives have been developed, as analogs of the Fréchet derivative for functions that are locally Lipschitz continuous but not everywhere differentiable. The pertinent generalized derivatives are summarized in this section. This section is largely reproduced from the articles [54, 55, 61].

2.3.1 Clarke's generalized Jacobian and the B-subdifferential

Clarke's generalized Jacobian [16] and the associated B-subdifferential [92] are used in established numerical methods for nonsmooth problems, and are defined as follows.

Definition 2.3.1 (from [16] and [91]). *Given an open set $X \subset \mathbb{R}^n$ and a locally Lipschitz continuous function $\mathbf{f} : X \to \mathbb{R}^m$, let $S \subset X$ be the set on which $\mathbf{f}$ is differentiable. The* B-subdifferential *of $\mathbf{f}$ at $\mathbf{x} \in X$ is defined as*

$$\partial_{\mathrm{B}}\mathbf{f}(\mathbf{x}) := \left\{ \mathbf{H} \in \mathbb{R}^{m\times n} : \mathbf{H} = \lim_{j\to\infty} \mathbf{J}\mathbf{f}(\mathbf{x}_{(j)}), \quad \mathbf{x} = \lim_{j\to\infty} \mathbf{x}_{(j)}, \quad \mathbf{x}_{(i)} \in S, \ \forall i \in \mathbb{N} \right\}.$$

Clarke's generalized Jacobian *of $\mathbf{f}$ at $\mathbf{x}$ is* $\partial\mathbf{f}(\mathbf{x}) := \operatorname{conv} \partial_{\mathrm{B}}\mathbf{f}(\mathbf{x})$.

In the above definition, for any $\mathbf{x} \in X$, the sets $\partial_{\mathrm{B}}\mathbf{f}(\mathbf{x})$ and $\partial\mathbf{f}(\mathbf{x})$ are necessarily nonempty and compact, and $\partial\mathbf{f}(\mathbf{x})$ is convex. If $\mathbf{f}$ is differentiable at $\mathbf{x}$, then $\mathbf{J}\mathbf{f}(\mathbf{x}) \in \partial\mathbf{f}(\mathbf{x})$. If $\mathbf{f}$ is $\mathcal{C}^1$ at $\mathbf{x}$, then $\{\mathbf{J}\mathbf{f}(\mathbf{x})\} = \partial_{\mathrm{B}}\mathbf{f}(\mathbf{x}) = \partial\mathbf{f}(\mathbf{x})$. As suggested by the notation of the Clarke Jacobian, if $\mathbf{f} \equiv f$ is both scalar-valued and convex, then $\partial f(\mathbf{x})$ is the set of transposed subgradients of f at $\mathbf{x}$.

The Clarke Jacobian satisfies classical calculus rules as inclusions instead of equations [16, Sections 2.3, 2.6, and 2.7], and exhibits several useful properties, including satisfaction of the following mean value theorem, inverse function theorem, and implicit function theorem [16]. Clarke's mean value theorem is reproduced below.

Proposition 2.3.2 (Proposition 2.6.5 in [16]). *Consider an open, convex set $X \subset \mathbb{R}^n$ and a locally Lipschitz continuous function $\mathbf{f} : X \to \mathbb{R}^m$. For each $\mathbf{x}, \boldsymbol{\xi} \in X$,*

$$\mathbf{f}(\mathbf{x}) - \mathbf{f}(\boldsymbol{\xi}) \in \operatorname{conv}\{\mathbf{H}(\mathbf{x} - \boldsymbol{\xi}) : \mathbf{H} \in \partial\mathbf{f}(\mathbf{z}), \ \mathbf{z} = \lambda\mathbf{x} + (1-\lambda)\boldsymbol{\xi}, \ \lambda \in [0,1]\}.$$

Useful optimality conditions can be expressed in terms of the Clarke Jacobian [16, Theorem 6.1.1]. Elements of the Clarke Jacobian are used in semismooth

Newton methods for equation solving [92], and in bundle methods for local optimization [67].

The set-valued mapping $\mathbf{x} \mapsto \partial \mathbf{f}(\mathbf{x})$ is upper-semicontinuous, as is the mapping $\mathbf{x} \mapsto \partial_{\mathrm{B}} \mathbf{f}(\mathbf{x})$ [23].

2.3.2 Plenary Jacobians

Sweetser [109] introduced the notion of *plenary sets* and *hulls*; combined with the Clarke Jacobian, these provide a generalized derivative that is a superset of the Clarke Jacobian, but does not sacrifice any of the latter's useful properties.

Definition 2.3.3 (from [109]). *A set $A \subset \mathbb{R}^{m \times n}$ is* plenary *if*

$$A = \{\mathbf{H} \in \mathbb{R}^{m \times n} : \forall \mathbf{d} \in \mathbb{R}^n,\ \exists \mathbf{A} \in A \ \textit{s.t.}\ \mathbf{Hd} = \mathbf{Ad}\}.$$

The plenary hull *of a set $S \subset \mathbb{R}^{m \times n}$ is the intersection of all plenary supersets of S in $\mathbb{R}^{m \times n}$, and is denoted* $\operatorname{plen} S \subset \mathbb{R}^{m \times n}$.

It is possible for $\operatorname{plen} S$ to be a strict superset of S, even if S is both convex and compact. The intersection of any collection of plenary sets is itself plenary [109]; thus, the plenary hull of any set of matrices is itself plenary. Given a set $S \subset \mathbb{R}^{m \times n}$ with either $n = 1$, $m = 1$, or both, S is itself plenary, and so $\operatorname{plen} S = S$.

Definition 2.3.4 (adapted from [109]). *Given an open set $X \subset \mathbb{R}^n$ and a function $\mathbf{f} : X \to \mathbb{R}^m$, the* plenary Jacobian $\partial_{\mathrm{P}} \mathbf{f}(\mathbf{x})$ *of $\mathbf{f}$ at $\mathbf{x} \in X$ is the plenary hull of the Clarke Jacobian $\partial \mathbf{f}(\mathbf{x})$. Equivalently,*

$$\partial_{\mathrm{P}} \mathbf{f}(\mathbf{x}) := \{\mathbf{H} \in \mathbb{R}^{m \times n} : \forall \mathbf{d} \in \mathbb{R}^n,\ \exists \mathbf{A} \in \partial \mathbf{f}(\mathbf{x}) \ \textit{s.t.}\ \mathbf{Hd} = \mathbf{Ad}\}.$$

The plenary Jacobian $\partial_{\mathrm{P}} \mathbf{f}(\mathbf{x})$ is convex, compact, and not empty [46]. The equivalence mentioned in the above definition follows immediately from (2.6) and [109, Lemma 3.1], since $\partial_{\mathrm{P}} \mathbf{f}(\mathbf{x})$ is compact and $\partial \mathbf{f}(\mathbf{x})$ is both convex and compact. This second characterization of the plenary Jacobian will be used in Chapter 3 to determine whether particular matrices are elements of $\partial_{\mathrm{P}} \mathbf{f}(\mathbf{x})$.

The plenary Jacobian has been investigated in [41, 46, 109, 123], and satisfies:

$$\partial \mathbf{f}(\mathbf{x}) \subset \partial_{\mathrm{P}} \mathbf{f}(\mathbf{x}) \subset \prod_{i=1}^{m} \partial f_i(\mathbf{x}), \tag{2.6}$$

where either or both of the above inclusions may be strict. (The rightmost set above denotes the set of matrices $\mathbf{M}$ whose i^{th} row is an element of $\partial f_i(\mathbf{x})$, for every $i \in \{1, \ldots, m\}$.) When $\min\{m, n\} = 1$, however, $\partial \mathbf{f}(\mathbf{x}) = \partial_{\mathrm{P}} \mathbf{f}(\mathbf{x})$. Since the objective functions in nonlinear programs (NLPs) are scalar-valued, it follows that *bundle methods* for finding local minima for nonsmooth NLPs [63, 67] are unaffected if the plenary Jacobian is used in place of Clarke's generalized Jacobian.

Combining Definition 2.3.4 with the inclusion $\partial \mathbf{f}(\mathbf{x}) \subset \partial_{\mathrm{P}} \mathbf{f}(\mathbf{x})$ yields:

$$\{\mathbf{H}\,\mathbf{e} \in \mathbb{R}^m : \mathbf{H} \in \partial_{\mathrm{P}} \mathbf{f}(\mathbf{x})\} = \{\mathbf{H}\,\mathbf{e} \in \mathbb{R}^m : \mathbf{H} \in \partial \mathbf{f}(\mathbf{x})\}, \qquad \forall \mathbf{e} \in \mathbb{R}^n. \tag{2.7}$$

The following proposition shows that if $m = n$, and if certain nonsingularity assumptions apply, then a similar relationship holds between images of inverses of elements of $\partial \mathbf{f}(\mathbf{x})$ and $\partial_{\mathrm{P}} \mathbf{f}(\mathbf{x})$. The condition that $\partial \mathbf{f}(\mathbf{x})$ does not contain any singular matrices is a key assumption in Clarke's inverse function theorem and implicit function theorem for locally Lipschitz continuous functions [16].

Proposition 2.3.5. *Given an open set $X \subset \mathbb{R}^n$ and a locally Lipschitz continuous function $\mathbf{f} : X \to \mathbb{R}^n$, suppose that for some $\mathbf{x} \in X$, $\partial \mathbf{f}(\mathbf{x})$ does not contain any singular matrices. Then $\partial_{\mathrm{P}} \mathbf{f}(\mathbf{x})$ does not contain any singular matrices either, and*

$$\{\mathbf{H}^{-1}\mathbf{e} \in \mathbb{R}^n : \mathbf{H} \in \partial_{\mathrm{P}} \mathbf{f}(\mathbf{x})\} = \{\mathbf{H}^{-1}\mathbf{e} \in \mathbb{R}^n : \mathbf{H} \in \partial \mathbf{f}(\mathbf{x})\} \qquad \forall \mathbf{e} \in \mathbb{R}^n.$$

Proof. Since $\partial \mathbf{f}(\mathbf{x})$ does not contain any singular matrices, [123, Proposition 3] implies that $\partial_{\mathrm{P}} \mathbf{f}(\mathbf{x})$ does not contain any singular matrices either. Since $\partial \mathbf{f}(\mathbf{x}) \subset \partial_{\mathrm{P}} \mathbf{f}(\mathbf{x})$, the inclusion $\{\mathbf{H}^{-1}\mathbf{e} \in \mathbb{R}^n : \mathbf{H} \in \partial_{\mathrm{P}} \mathbf{f}(\mathbf{x})\} \supset \{\mathbf{H}^{-1}\mathbf{e} \in \mathbb{R}^n : \mathbf{H} \in \partial \mathbf{f}(\mathbf{x})\}$ is trivial for each $\mathbf{e} \in \mathbb{R}^n$. To prove the reverse inclusion, choose any $\mathbf{e} \in \mathbb{R}^n$ and any $\mathbf{A} \in \partial_{\mathrm{P}} \mathbf{f}(\mathbf{x})$. This implies that $\mathbf{A}$ is nonsingular. By (2.7),

$$\mathbf{e} = \mathbf{A}(\mathbf{A}^{-1}\mathbf{e}) \in \{\mathbf{H}(\mathbf{A}^{-1}\mathbf{e}) \in \mathbb{R}^n : \mathbf{H} \in \partial_{\mathrm{P}}\mathbf{f}(\mathbf{x})\} = \{\mathbf{H}(\mathbf{A}^{-1}\mathbf{e}) \in \mathbb{R}^n : \mathbf{H} \in \partial\mathbf{f}(\mathbf{x})\}.$$

Thus, there exists $\mathbf{B} \in \partial\mathbf{f}(\mathbf{x})$ for which $\mathbf{e} = \mathbf{B}(\mathbf{A}^{-1}\mathbf{e})$. By the hypotheses of the proposition, $\mathbf{B}$ is nonsingular, and so $\mathbf{A}^{-1}\mathbf{e} = \mathbf{B}^{-1}\mathbf{e} \in \{\mathbf{H}^{-1}\,\mathbf{e} \in \mathbb{R}^n : \mathbf{H} \in \partial\mathbf{f}(\mathbf{x})\}$.

□

It follows that if the plenary Jacobian is used in place of Clarke's generalized Jacobian in a semismooth Newton method [92], then any sequence of iterates generated by the altered method can necessarily be generated using the original method. Thus, convergence results for the original method remain applicable when the plenary Jacobian is used in place of the Clarke Jacobian. Similarly, it follows from (2.7) that if the plenary Jacobian is used in place of the generalized Jacobian in Clarke's mean value theorem [16, Proposition 2.6.5], then the result is unaffected. Since $\partial_{\mathrm{P}}\mathbf{f}(\mathbf{x})$ contains a singular matrix if and only if $\partial\mathbf{f}(\mathbf{x})$ contains a singular matrix [123], Clarke's inverse function theorem [16, Theorem 7.1.1] for locally Lipschitz continuous functions is also unaffected if the generalized Jacobian is replaced with the plenary Jacobian.

As a set-valued mapping on X, $\partial_{\mathrm{P}}\mathbf{f}$ is upper-semicontinuous [123]. Thus, (2.7), [23, Definition 7.2.2], and [23, Definition 7.5.13] imply that $\partial_{\mathrm{P}}\mathbf{f}$ is a linear Newton approximation [23] of $\mathbf{f}$ at any $\mathbf{x} \in X$. In light of the previous paragraph, $\partial_{\mathrm{P}}\mathbf{f}$ is in some sense as good a linear Newton approximation of $\mathbf{f}$ as $\partial\mathbf{f}$. Similarly, it follows from (2.7) and Proposition 2.3.5 that the plenary Jacobian is a *Newton map* [64] satisfying the invertibility conditions in [64, Section 10.1] if and only if the Clarke Jacobian is as well.

2.3.3 Lexicographic derivatives

The theory of *lexicographic differentiation* was developed by Nesterov [79], and is summarized in this section. This thesis presents new theoretical results concerning lexicographic derivatives and new numerical methods for computing them.

Definition 2.3.6. *Given an open set $X \subset \mathbb{R}^n$ and a locally Lipschitz continuous function $\mathbf{f} : X \to \mathbb{R}^m$, $\mathbf{f}$ is* lexicographically smooth (L-smooth) *at $\mathbf{x} \in X$ if it is directionally differentiable at $\mathbf{x}$ and if, for any $p \in \mathbb{N}$ and $\mathbf{M} \in \mathbb{R}^{n \times p}$, the following* higher-order directional derivatives *are well-defined:*

$$
\begin{aligned}
&\mathbf{f}^{(0)}_{\mathbf{x},\mathbf{M}} : \mathbb{R}^n \to \mathbb{R}^m : \mathbf{h} \mapsto \mathbf{f}'(\mathbf{x};\mathbf{h}),\\
&\mathbf{f}^{(1)}_{\mathbf{x},\mathbf{M}} : \mathbb{R}^n \to \mathbb{R}^m : \mathbf{h} \mapsto [\mathbf{f}^{(0)}_{\mathbf{x},\mathbf{M}}]'(\mathbf{m}_{(1)};\mathbf{h}),\\
&\mathbf{f}^{(2)}_{\mathbf{x},\mathbf{M}} : \mathbb{R}^n \to \mathbb{R}^m : \mathbf{h} \mapsto [\mathbf{f}^{(1)}_{\mathbf{x},\mathbf{M}}]'(\mathbf{m}_{(2)};\mathbf{h}),\\
&\quad\vdots\\
&\mathbf{f}^{(p)}_{\mathbf{x},\mathbf{M}} : \mathbb{R}^n \to \mathbb{R}^m : \mathbf{h} \mapsto [\mathbf{f}^{(p-1)}_{\mathbf{x},\mathbf{M}}]'(\mathbf{m}_{(p)};\mathbf{h}).
\end{aligned}
$$

The following lemma provides basic properties of higher-order directional derivatives.

Lemma 2.3.7. *Given an open set $X \subset \mathbb{R}^n$, a function $\mathbf{f} : X \to \mathbb{R}^n$ that is L-smooth at $\mathbf{x} \in X$, some $p \in \mathbb{N}$, and some $\mathbf{M} = \begin{bmatrix}\mathbf{m}_{(1)} & \cdots & \mathbf{m}_{(p)}\end{bmatrix} \in \mathbb{R}^{n \times p}$, the following properties are satisfied:*

1. $\mathbf{f}^{(j)}_{\mathbf{x},\mathbf{M}}(\tau\mathbf{d}) = \tau\mathbf{f}^{(j)}_{\mathbf{x},\mathbf{M}}(\mathbf{d}), \quad \forall \tau \geq 0, \quad \forall \mathbf{d} \in \mathbb{R}^n, \quad \forall j \in \{0, 1, \ldots, p\},$
2. $\mathbf{f}^{(j)}_{\mathbf{x},\mathbf{M}}(\mathbf{d}) = \mathbf{f}^{(j+1)}_{\mathbf{x},\mathbf{M}}(\mathbf{d}) = \cdots = \mathbf{f}^{(p)}_{\mathbf{x},\mathbf{M}}(\mathbf{d}), \forall \mathbf{d} \in \operatorname{span}\{\mathbf{m}_{(1)}, \ldots, \mathbf{m}_{(j)}\}, \forall j \in \{1, \ldots, p\},$
3. $\mathbf{f}^{(j)}_{\mathbf{x},\mathbf{M}}$ *is linear on* $\operatorname{span}\{\mathbf{m}_{(1)}, \ldots, \mathbf{m}_{(j)}\}$ *for each* $j \in \{1, \ldots, p\}$,
4. $\mathbf{f}^{(j-1)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}_{(j)}) = \mathbf{f}^{(j)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}_{(j)}) = \cdots = \mathbf{f}^{(p)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}_{(j)}), \quad \forall j \in \{1, \ldots, p\},$
5. *With* $\tilde{\mathbf{M}} := \begin{bmatrix}\mathbf{m}_{(1)} & \cdots & \mathbf{m}_{(j)} & \mathbf{A}\end{bmatrix}$ *for any particular* $j \in \{0, 1, \ldots, p\}$, $q \in \mathbb{N} \cup \{0\}$, *and* $\mathbf{A} \in \mathbb{R}^{n \times q}$, $\mathbf{f}^{(j)}_{\mathbf{x},\mathbf{M}}(\mathbf{d}) = \mathbf{f}^{(j)}_{\mathbf{x},\tilde{\mathbf{M}}}(\mathbf{d})$ *for each* $\mathbf{d} \in \mathbb{R}^n$.

Proof. Properties 1, 2, and 3 are demonstrated in [79, Lemma 3]. To obtain the leftmost equation in Property 4, combining the definition of $\mathbf{f}^{(j)}_{\mathbf{x},\mathbf{M}}$ with the positive homogeneity of $\mathbf{f}^{(j-1)}_{\mathbf{x},\mathbf{M}}$ implied by Property 1 yields

$$\mathbf{f}^{(j)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}^{(j)}) = \lim_{\tau \to 0^+} \frac{(1+\tau)\mathbf{f}^{(j-1)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}^{(j)}) - \mathbf{f}^{(j-1)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}^{(j)})}{\tau} = \mathbf{f}^{(j-1)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}^{(j)}).$$

The remaining equations in Property 4 follow immediately from Property 2. Property 5 follows from the construction of the mappings $\mathbf{f}^{(j)}_{\mathbf{x},\mathbf{M}}$, noting that for each $\mathbf{d} \in \mathbb{R}^n$, $\mathbf{f}^{(j)}_{\mathbf{x},\mathbf{M}}(\mathbf{d})$ is independent of the rightmost $(p-j)$ columns of $\mathbf{M}$. □

Remark 2.3.8. *It follows from Property 5 of Lemma 2.3.7 that for any* $\mathbf{d} \in \mathbb{R}^n$, $\mathbf{f}^{(0)}_{\mathbf{x},\varnothing_{n\times 0}}(\mathbf{d})$ *is well-defined, and equals* $\mathbf{f}'(\mathbf{x};\mathbf{d})$.

The class of L-smooth functions is closed under composition, and includes all $\mathcal{C}^1$ functions and all convex functions, among others. If the columns of $\mathbf{M} \in \mathbb{R}^{n\times p}$ span $\mathbb{R}^n$, $\mathbf{f}^{(p)}_{\mathbf{x},then\mathbf{M}}$ is guaranteed to be linear [79], motivating the following definition.

Definition 2.3.9. *Consider an open set* $X \subset \mathbb{R}^n$ *and a function* $\mathbf{f} : X \to \mathbb{R}^m$ *that is L-smooth at* $\mathbf{x} \in X$. *For any nonsingular square matrix* $\mathbf{M} \in \mathbb{R}^{n\times n}$, *the* lexicographic derivative *of* $\mathbf{f}$ *at* $\mathbf{x}$ *in the directions* $\mathbf{M}$ *is* $\mathbf{J}_{\mathrm{L}}\mathbf{f}(\mathbf{x};\mathbf{M}) := \mathbf{J}\mathbf{f}^{(n)}_{\mathbf{x},\mathbf{M}}(\mathbf{0}) \in \mathbb{R}^{m\times n}$. *The* lexicographic subdifferential *of* $\mathbf{f}$ *at* $\mathbf{x}$ *is then*

$$\partial_{\mathrm{L}}\mathbf{f}(\mathbf{x}) := \{\mathbf{J}_{\mathrm{L}}\mathbf{f}(\mathbf{x};\mathbf{N}) : \mathbf{N} \in \mathbb{R}^{n\times n},\ \det \mathbf{N} \neq 0\} \subset \mathbb{R}^{m\times n}.$$

Observe that two functions with the same directional derivatives at $\mathbf{x} \in X$ will have the same lexicographic derivatives at $\mathbf{x}$. If $\mathbf{f}$ is differentiable at $\mathbf{x}$, then $\mathbf{f}$ is also L-smooth at $\mathbf{x}$; for all $\mathbf{M} \in \mathbb{R}^{n\times p}$ and $\mathbf{d} \in \mathbb{R}^n$,

$$\mathbf{f}^{(0)}_{\mathbf{x},\mathbf{M}}(\mathbf{d}) = \mathbf{f}^{(1)}_{\mathbf{x},\mathbf{M}}(\mathbf{d}) = \ldots = \mathbf{f}^{(p)}_{\mathbf{x},\mathbf{M}}(\mathbf{d}) = \mathbf{J}\mathbf{f}(\mathbf{x})\,\mathbf{d},$$

and so $\partial_{\mathrm{L}}\mathbf{f}(\mathbf{x}) = \{\mathbf{J}\mathbf{f}(\mathbf{x})\}$ in this case. Given a scalar-valued L-smooth function f, the inclusion $\partial_{\mathrm{L}} f(\mathbf{x}) \subset \partial f(\mathbf{x})$ was demonstrated in [79].

Unlike the Clarke Jacobian and the B-subdifferential, the lexicographic derivative satisfies sharp calculus rules. In an abuse of notation, for any $\mathbf{M} \in \mathbb{R}^{n\times p}$ with $p \in \mathbb{N}$, let $\tilde{\mathbf{J}}_{\mathrm{L}}\mathbf{f}(\mathbf{x};\mathbf{M}) \in \mathbb{R}^{m\times n}$ denote *any* particular matrix for which $\mathbf{f}^{(p)}_{\mathbf{x},\mathbf{M}}(\mathbf{d}) =$

$\tilde{\mathbf{J}}_{\mathrm{L}}\mathbf{f}(\mathbf{x};\mathbf{M})\,\mathbf{d}$ for each $\mathbf{d} \in \mathcal{R}(\mathbf{M})$. Property 3 of Lemma 2.3.7 shows that at least one such matrix exists.

According to Nesterov's chain rule for lexicographic derivatives [79, Theorem 5], if $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ are open, and if functions $\mathbf{g} : X \to Y$ and $\mathbf{f} : Y \to \mathbb{R}^q$ are lexicographically smooth, then the composition $\mathbf{f} \circ \mathbf{g}$ is also lexicographically smooth. Moreover, for any nonsingular matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ and any $\mathbf{x} \in X$,

$$\mathbf{J}_{\mathrm{L}}[\mathbf{f} \circ \mathbf{g}](\mathbf{x};\mathbf{M}) = \tilde{\mathbf{J}}_{\mathrm{L}}\mathbf{f}(\mathbf{g}(\mathbf{x});\mathbf{J}_{\mathrm{L}}\mathbf{g}(\mathbf{x};\mathbf{M})\,\mathbf{M})\;\mathbf{J}_{\mathrm{L}}\mathbf{g}(\mathbf{x};\mathbf{M}). \tag{2.8}$$

Observe that the matrix $\mathbf{J}_{\mathrm{L}}\mathbf{g}(\mathbf{x};\mathbf{M})\,\mathbf{M}$ in the above expression may be rectangular, and that its columns do not necessarily span $\mathbb{R}^m$. This chain rule offers a means of evaluating a lexicographic derivative $\mathbf{J}_{\mathrm{L}}\mathbf{f}(\mathbf{x};\mathbf{M})$ for a composite function $\mathbf{f}$, without explicitly constructing the higher-order directional derivatives $\mathbf{f}^{(j)}_{\mathbf{x},\mathbf{M}}$. In Chapter 3, this chain rule will be recast in a more tractable form.

Example 3.3.7 shows that, unlike the Clarke Jacobian and the B-subdifferential, the lexicographic subdifferential is not upper-semicontinuous as a set-valued mapping.

2.4 Piecewise differentiable functions

As formalized by Scholtes [97], piecewise differentiable functions represent a broad class of functions that need not be differentiable everywhere, but nevertheless satisfy useful properties. The material in this section is largely reproduced from [54, 56, 61]. Piecewise differentiable functions are defined as follows.

Definition 2.4.1 (from [97]). *Given an open set $X \subset \mathbb{R}^n$, some $\mathbf{x} \in X$, and a function $\mathbf{f} : X \to \mathbb{R}^m$, $\mathbf{f}$ is* piecewise differentiable *($\mathcal{PC}^1$) at $\mathbf{x}$ if there exists a neighborhood $N \subset X$ of $\mathbf{x}$ and a finite collection $\mathcal{F}_{\mathbf{f}}(\mathbf{x})$ of $\mathcal{C}^1$* selection functions *mapping N into $\mathbb{R}^m$, for which $\mathbf{f}$ is continuous on N, and*

$$\mathbf{f}(\mathbf{y}) \in \{\phi(\mathbf{y}) : \phi \in \mathcal{F}_{\mathbf{f}}(\mathbf{x})\}, \qquad \forall \mathbf{y} \in N.$$

If, in addition, each selection function $\phi \in \mathcal{F}_{\mathbf{f}}(\mathbf{x})$ is linear, then $\mathbf{f}$ is piecewise linear

($\mathcal{PL}$) at $\mathbf{x}$*. If each selection function* $\phi \in \mathcal{F}_{\mathbf{f}}(\mathbf{x})$ *is affine, then* $\mathbf{f}$ *is* piecewise affine *($\mathcal{PA}$) at* $\mathbf{x}$*.*

Proposition 2.4.2 (Proposition 4.1.1 in [97])**.** *Given an open set* $X \subset \mathbb{R}^n$*, a function* $\mathbf{f} : X \to \mathbb{R}^m$ *that is* $\mathcal{PC}^1$ *at* $\mathbf{x} \in X$*, and a finite collection* $\mathcal{F}_{\mathbf{f}}(\mathbf{x})$ *of* $\mathcal{C}^1$ *selection functions for* $\mathbf{f}$ *around* $\mathbf{x}$*, there exists a neighborhood* $N \subset X$ *of* $\mathbf{x}$ *and a collection* $\mathcal{E}_{\mathbf{f}}(\mathbf{x}) \subset \mathcal{F}_{\mathbf{f}}(\mathbf{x})$ *of* essentially active selection functions *for* $\mathbf{f}$ *around* $\mathbf{x}$*, for which both*

$$\mathbf{x} \in \mathrm{cl}(\mathrm{int}(\{\mathbf{y} \in N : \mathbf{f}(\mathbf{y}) = \phi(\mathbf{y})\}))\,, \qquad \forall \phi \in \mathcal{E}_{\mathbf{f}}(\mathbf{x}),$$

and

$$\mathbf{f}(\mathbf{y}) \in \{\phi(\mathbf{y}) : \phi \in \mathcal{E}_{\mathbf{f}}(\mathbf{x})\}, \qquad \forall \mathbf{y} \in N.$$

As described by Scholtes [97], $\mathcal{PC}^1$ functions exhibit several useful properties: they are locally Lipschitz continuous [97, Corollary 4.1.1] and directionally differentiable [97, Proposition 4.1.3] on their domains, and the class of $\mathcal{PC}^1$ functions is closed under composition. If $\mathbf{f}$ is $\mathcal{PC}^1$ at $\mathbf{x}$, then the directional derivative mapping $\mathbf{d} \mapsto \mathbf{f}'(\mathbf{x}; \mathbf{d})$ is $\mathcal{PL}$ on $\mathbb{R}^n$ [97, Proposition 4.1.3]. Moreover, the following proposition relates the generalized derivatives of a $\mathcal{PC}^1$ function to its essentially active selection functions.

Proposition 2.4.3. *Consider an open set* $X \subset \mathbb{R}^n$*, a function* $\mathbf{f} : X \to \mathbb{R}^m$ *that is* $\mathcal{PC}^1$ *at* $\mathbf{x} \in X$*, and any finite collection* $\mathcal{E}_{\mathbf{f}}(\mathbf{x})$ *of essentially active* $\mathcal{C}^1$ *selection functions for* $\mathbf{f}$ *around* $\mathbf{x}$*. Then:*

1. *for each* $\mathbf{d} \in \mathbb{R}^n$*,* $\mathbf{f}'(\mathbf{x}; \mathbf{d}) \in \{\mathbf{J}\phi(\mathbf{x})\,\mathbf{d} : \phi \in \mathcal{E}_{\mathbf{f}}(\mathbf{x})\}$*, and*
2. $\partial_{\mathrm{B}}\mathbf{f}(\mathbf{x}) = \{\mathbf{J}\phi(\mathbf{x}) : \phi \in \mathcal{E}_{\mathbf{f}}(\mathbf{x})\}$.

Proof. The property concerning $\mathbf{f}'(\mathbf{x}; \mathbf{d})$ is provided by [97, Proposition 4.1.3], and the property concerning $\partial_{\mathrm{B}}\mathbf{f}(\mathbf{x})$ follows from the proof of [97, Proposition 4.3.1]. □

Kojima and Shindo describe a Newton method [65, Algorithm EN] for solving equation systems with $\mathcal{PC}^1$ residual functions, using B-subdifferential elements at

each iteration. This method exhibits local Q-quadratic convergence, provided that its nonsingularity assumptions are met.

The B-subdifferential of a $\mathcal{PC}^1$ function $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ also satisfies the inclusion $\partial_{\mathrm{B}}[\mathbf{f}'(\mathbf{x};\cdot)](\mathbf{0}) \subset \partial_{\mathrm{B}}\mathbf{f}(\mathbf{x})$ for each $\mathbf{x} \in X$ [85, Lemma 2]. It will be shown in Chapter 3 that all $\mathcal{PC}^1$ functions are L-smooth, and that their lexicographic derivatives are always elements of their B-subdifferentials.

Useful characterizations of the inverses of invertible $\mathcal{PC}^1$ functions are provided in [94], and will be exploited in Chapter 8.

2.5 Factorable functions and automatic differentiation

The *vector forward mode of automatic differentiation (AD)* is a computationally efficient procedure for evaluating derivatives of finite compositions of simple *elemental functions*, as described in [34]. Such composite functions are formalized below.

Definition 2.5.1. *Given an open set $X \subset \mathbb{R}^n$, a function $\mathbf{f} : X \to \mathbb{R}^m$ is* factorable *if there exist:*

- *an* elemental library $\mathcal{L}$ *of functions, with each $\boldsymbol{\psi} \in \mathcal{L}$ mapping some open set $X_{\boldsymbol{\psi}} \subset \mathbb{R}^{n_{\boldsymbol{\psi}}}$ into $\mathbb{R}^{m_{\boldsymbol{\psi}}}$,*
- *a number $\ell \in \mathbb{N}$,*
- elemental functions $\boldsymbol{\psi}_{(j)} \in \mathcal{L}$ *for each $j \in \{1, \ldots, \ell\}$, and*
- *a binary relation $\prec$ such that $(i \prec j) \in \{\texttt{true}, \texttt{false}\}$ for each pair $(i, j) \in \{1, \ldots, \ell\}^2$ with $i < j$,*

such that $\mathbf{f}$ can be represented as follows. For any matrix (or vector) quantity $\mathbf{A}_{(j)}$ defined for each $j \in \{1, \ldots, \ell\}$ such that all $\mathbf{A}_{(j)}$s have the same number of columns, let $[\mathbf{A}_{(i)}]_{i \prec j}$ denote the matrix (or vector) obtained by stacking the elements of $\{\mathbf{A}_{(i)} : i \prec j\}$ vertically in order of increasing i. Then, for any $\mathbf{x} \in X$, $\mathbf{f}(\mathbf{x})$ can be evaluated according to the following factored representation *of $\mathbf{f}$:*

Set $\mathbf{v}_{(0)} \leftarrow \mathbf{x}$
for $j = 1$ ***to*** ℓ ***do***
  Set $\mathbf{u}_{(j)} \leftarrow [\mathbf{v}_{(i)}]_{i \prec j}$
  Set $\mathbf{v}_{(j)} \leftarrow \psi_{(j)}(\mathbf{u}_{(j)})$
end for
Set $\mathbf{f}(\mathbf{x}) \leftarrow \mathbf{v}_{(\ell)}$.

The following example illustrates the usefulness of including coordinate projection functions in the employed elemental library $\mathcal{L}$.

Example 2.5.2. *Consider the function* $f : \mathbb{R}^2 \to \mathbb{R} : (x_1, x_2) \mapsto \sin x_2$. *Defining the coordinate projection mapping* $\pi_2 : (x_1, x_2) \mapsto x_2$, *a factored representation for* f *is as follows:*

Set $\mathbf{v}_{(0)} \leftarrow (x_1, x_2)$
Set $\mathbf{u}_{(1)} \leftarrow \mathbf{v}_{(0)}$
Set $v_{(1)} \leftarrow \pi_2(\mathbf{u}_{(1)})$
Set $u_{(2)} \leftarrow v_{(1)}$
Set $v_{(2)} \leftarrow \sin u_{(2)}$
Set $f(x_1, x_2) \leftarrow v_{(2)}$.

The intermediate variables $\mathbf{u}_{(j)}$ and $\mathbf{v}_{(j)}$ in a factored representation may be thought of as functions of $\mathbf{x}$; for notational convenience, however, they will not be represented as $\mathbf{u}_{(j)}(\mathbf{x})$ or $\mathbf{v}_{(j)}(\mathbf{x})$. For each $\mathbf{x} \in X$, $\mathbf{u}_{(j)} \in X_{\psi_{(j)}}$ must hold. Factored representations are clearly not unique. Unless otherwise noted, any mentioned factorable function will be considered to have the generic factored representation given in Definition 2.5.1.

The following subclasses of factorable functions are amenable to certain varieties of AD. Another type of factorable function will be considered in Chapter 9.

Definition 2.5.3. *A factorable function* $\mathbf{f}$ *is* $\mathcal{C}^1$-factorable *if the elemental library* $\mathcal{L}$ *contains only* $\mathcal{C}^1$ *functions whose Jacobians can be computed.*

AD is conventionally applied to $\mathcal{C}^1$-factorable functions. A $\mathcal{C}^1$-factorable function is evidently $\mathcal{C}^1$ itself; Algorithm 1, adapted from [34], is the *vector forward AD mode* for evaluating its Jacobian, postmultiplied by a given matrix.

Algorithm 1 Computes $\mathbf{f}(\mathbf{x})$ and $\mathbf{J}\mathbf{f}(\mathbf{x})\,\mathbf{M}$ for a $\mathcal{C}^1$-factorable function $\mathbf{f}$

Require: $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ is $\mathcal{C}^1$-factorable, $\mathbf{x} \in X$, $\mathbf{M} \in \mathbb{R}^{n \times p}$
  Set $\mathbf{v}_{(0)} \leftarrow \mathbf{x}$
  Set $\dot{\mathbf{V}}_{(0)} \leftarrow \mathbf{M}$
  for $j = 1$ **to** ℓ **do**
    Set $\mathbf{u}_{(j)} \leftarrow [\mathbf{v}_{(i)}]_{i \prec j}$
    Set $\mathbf{v}_{(j)} \leftarrow \boldsymbol{\psi}_{(j)}(\mathbf{u}_{(j)})$
    Set $\dot{\mathbf{U}}_{(j)} \leftarrow [\dot{\mathbf{V}}_{(i)}]_{i \prec j}$
    Set $\dot{\mathbf{V}}_{(j)} \leftarrow \mathbf{J}\boldsymbol{\psi}_{(j)}(\mathbf{u}_{(j)})\,\dot{\mathbf{U}}_{(j)}$
  end for
  return $\mathbf{f}(\mathbf{x}) = \mathbf{v}_{(\ell)}$ and $\mathbf{J}\mathbf{f}(\mathbf{x})\,\mathbf{M} = \dot{\mathbf{V}}_{(\ell)}$

Definition 2.5.4. *A factorable function* $\mathbf{f}$ *is* abs-$\mathcal{C}^1$-factorable *if the elemental library* $\mathcal{L}$ *comprises both the absolute value function* $x \mapsto |x|$*, and various* $\mathcal{C}^1$ *functions whose Jacobians can be computed. If, in addition, each* $\mathcal{C}^1$ *function in* $\mathcal{L}$ *is also* $\mathcal{C}^\omega$*, then* $\mathbf{f}$ *is* abs-$\mathcal{C}^\omega$-factorable*, or simply* abs-factorable*.*

Abs-$\mathcal{C}^1$-factorable functions represent a broad class of nonsmooth functions encountered in practice. The class of abs-$\mathcal{C}^1$-factorable functions was considered previously in [32, 33, 53].

Definition 2.5.5. *A factorable function* $\mathbf{f}$ *is* $\mathcal{PC}^1$-factorable *if each function* $\boldsymbol{\psi} \in \mathcal{L}$ *is* $\mathcal{PC}^1$*, and for which the following information is known or computable for each* $\mathbf{x} \in X_{\boldsymbol{\psi}}$*:*

- *the value* $\boldsymbol{\psi}(\mathbf{x})$*,*
- *a finite* active normal set $H_{\boldsymbol{\psi}}(\mathbf{x}) \subset \mathbb{R}^{n_{\boldsymbol{\psi}}}$*, such that* $\boldsymbol{\psi}'(\mathbf{x}; \cdot)$ *is smooth except on the set* $\{\mathbf{d} \in \mathbb{R}^{n_{\boldsymbol{\psi}}} : \exists \mathbf{a} \in H_{\boldsymbol{\psi}}(\mathbf{x}) \text{ s.t. } \mathbf{a} \neq \mathbf{0} \text{ and } \langle \mathbf{a}, \mathbf{d} \rangle = 0\}$*. With* $r_{\max} := |H_{\boldsymbol{\psi}}(\mathbf{x})|$*, enumerate the elements of* $H_{\boldsymbol{\psi}}(\mathbf{x})$ *as* $\mathbf{a}_{\boldsymbol{\psi}}^{(1)}(\mathbf{x}), \ldots, \mathbf{a}_{\boldsymbol{\psi}}^{(r_{\max})}(\mathbf{x})$*,*
- *a function* $\zeta_{\boldsymbol{\psi}} : X_{\boldsymbol{\psi}} \to \{\texttt{true}, \texttt{false}\}$ *such that* $\zeta_{\boldsymbol{\psi}}(\mathbf{x}) = \texttt{false}$ *if and only if* $H_{\boldsymbol{\psi}}(\mathbf{x})$ *contains a nonzero vector, and*
- *a* branch-locked Jacobian *mapping* $\boldsymbol{\Gamma}_{\boldsymbol{\psi}}(\mathbf{x}; \cdot)$ *if* $\zeta_{\boldsymbol{\psi}}(\mathbf{x}) = \texttt{false}$*, such that for each* $\mathbf{s} \in \{-1, 1\}^{|H_{\boldsymbol{\psi}}(\mathbf{x})|}$*,*

$$\boldsymbol{\psi}'(\mathbf{x}; \mathbf{d}) = \boldsymbol{\Gamma}_{\boldsymbol{\psi}}(\mathbf{x}; \mathbf{s})\,\mathbf{d},$$

for all $\mathbf{d} \in \mathbb{R}^{n_\psi}$ *such that*

$$s_r \langle \mathbf{a}_\psi^{(r)}(\mathbf{x}), \mathbf{d} \rangle \geq 0, \qquad \forall r \in \{1, \ldots, |H_\psi(\mathbf{x})|\}.$$

As motivated by the above restrictions on $\mathbf{d}$*, components of the argument* $\mathbf{s}$ *are called* halfspace specifiers.

Such functions ψ *are called* elemental $\mathcal{PC}^1$ functions.

The existence of $H_\psi(\mathbf{x})$ and ζ_ψ for any $\mathcal{PC}^1$ function ψ is demonstrated in Appendix A; existence of a branch-locked Jacobian mapping follows from Lemmata A.1.11 and A.6.1. As shown in the examples in Section A.2 in Appendix A, the required information for elemental $\mathcal{PC}^1$ functions is readily furnished for many $\mathcal{PC}^1$ functions encountered in practice, including $\mathcal{C}^1$ functions with computable Jacobians, basic nonsmooth $\mathcal{PC}^1$ functions such as abs, $\min\{\cdot,\cdot\}$ and $\max\{\cdot,\cdot\}$, and more complicated $\mathcal{PC}^1$ functions such as $(x, y) \mapsto \sup_{z \in [x,y]} \sin z$ where $x < y$, which occur in interval arithmetic.

The first computationally tractable methods for evaluating Clarke Jacobian elements for a broad class of vector-valued nonsmooth functions were developed in [53, 54]; the article [54] is reproduced in Appendix A. Algorithm 12 in Appendix A evaluates an element of $\partial_{\mathrm{B}}\mathbf{f}(\mathbf{x}) \subset \partial\mathbf{f}(\mathbf{x})$ for any $\mathcal{PC}^1$-factorable function $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$, when for any particular nonsingular matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$, the vectors $\mathbf{q}_{(1)}, \ldots, \mathbf{q}_{(n)}$ in the algorithm are initialized as the columns $\mathbf{m}_{(1)}, \ldots, \mathbf{m}_{(n)}$ of $\mathbf{M}$. Denote the element of $\partial_{\mathrm{B}}\mathbf{f}(\mathbf{x})$ thus obtained as the *conical Jacobian* $\mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{x}; \mathbf{M})$, and define the *conical subdifferential* as

$$\partial_{\mathrm{C}}\mathbf{f}(\mathbf{x}) := \{\mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{x}; \mathbf{M}) : \mathbf{M} \in \mathbb{R}^{n \times n},\ \det \mathbf{M} \neq 0\} \subset \partial_{\mathrm{B}}\mathbf{f}(\mathbf{x}).$$

In the language of Appendix A, $\mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{x}; \mathbf{M})$ is the Jacobian of a conically active selection function of $\mathbf{f}$ at $\mathbf{x}$. Thus, in light of [97, Proposition 4.3.1], Example A.1.14 shows that the above inclusion may be strict.

Chapter 3

Relationships between generalized derivatives

This chapter develops new theoretical properties and relationships involving the generalized derivatives described in Section 2.3. Firstly, the *LD-derivative* is developed as an analog of the lexicographic derivative, and is shown to satisfy a particularly tractable extension of Nesterov's chain rule for lexicographic derivatives. Secondly, lexicographic derivatives are shown to be elements of the plenary Jacobian whenever they exist. This inclusion is tightened for $\mathcal{PC}^1$ functions: such functions are shown to be L-smooth, with lexicographic derivatives that are always B-subdifferential elements. The material in this chapter is reproduced from the articles [55, 61].

3.1 LD-derivatives and lexicographic derivatives

This section introduces the *LD-derivative*, which provides a more tractable extension of Nesterov's chain rule for lexicographic derivatives [79, Theorem 5]. Various properties relating LD-derivatives and lexicographic derivatives are developed. This section is reproduced from [61].

Definition 3.1.1. *Given an open set $X \subset \mathbb{R}^n$, a locally Lipschitz continuous function*

$\mathbf{f} : X \to \mathbb{R}^m$ *that is L-smooth at* $\mathbf{x} \in X$*, and a matrix* $\mathbf{M} := \begin{bmatrix} \mathbf{m}_{(1)} & \cdots & \mathbf{m}_{(p)} \end{bmatrix} \in \mathbb{R}^{n \times p}$*, the* lexicographic directional (LD-)derivative *of* $\mathbf{f}$ *at* $\mathbf{x}$ *in the directions* $\mathbf{M}$ *is*

$$\begin{aligned} \mathbf{f}'(\mathbf{x};\mathbf{M}) &:= \begin{bmatrix} \mathbf{f}^{(0)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}_{(1)}) & \mathbf{f}^{(1)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}_{(2)}) & \cdots & \mathbf{f}^{(p-1)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}_{(p)}) \end{bmatrix}, \\ &= \begin{bmatrix} \mathbf{f}^{(p)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}_{(1)}) & \cdots & \mathbf{f}^{(p)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}_{(p)}) \end{bmatrix}. \end{aligned} \tag{3.1}$$

The second equation in (3.1) follows from Property 4 in Lemma 2.3.7. Note that $\mathbf{f}'(\mathbf{x};\mathbf{M})$ is uniquely defined for all $\mathbf{M} \in \mathbb{R}^{n \times p}$ and all $p \in \mathbb{N}$, unlike the similar construction $g_p(\mathbf{f}, \mathbf{M}, \mathbf{x})$ described in [79], which is denoted as $\tilde{\mathbf{J}}_{\mathrm{L}}\mathbf{f}(\mathbf{x};\mathbf{M})$ in [55]. As its name and notation suggest, the LD-derivative is a generalization of the standard directional derivative; the two are equivalent when $\mathbf{M}$ has only one column.

It follows from the discussion in Section 2.3.3 that if $\mathbf{f}$ is differentiable at $\mathbf{x}$, then $\mathbf{f}'(\mathbf{x};\mathbf{M}) = \mathbf{J}\mathbf{f}(\mathbf{x})\,\mathbf{M}$. If $\mathbf{M}$ is square and nonsingular, then $\mathbf{f}'(\mathbf{x};\mathbf{M}) = \mathbf{J}_{\mathrm{L}}\mathbf{f}(\mathbf{x};\mathbf{M})\,\mathbf{M}$. A useful feature of the LD-derivative is that it simplifies the form and treatment of Nesterov's chain rule [79, Theorem 5], as follows.

Proposition 3.1.2. *Consider open sets* $X \subset \mathbb{R}^n$ *and* $Y \subset \mathbb{R}^m$*, and locally Lipschitz continuous functions* $\mathbf{g} : X \to Y$ *and* $\mathbf{f} : Y \to \mathbb{R}^q$ *that are L-smooth at* $\mathbf{x} \in X$ *and* $\mathbf{g}(\mathbf{x}) \in Y$*, respectively. The composition* $\mathbf{f} \circ \mathbf{g}$ *is L-smooth at* $\mathbf{x}$*. Moreover, given any matrix* $\mathbf{M} := \begin{bmatrix} \mathbf{m}_{(1)} & \cdots & \mathbf{m}_{(p)} \end{bmatrix} \in \mathbb{R}^{n \times p}$*, the following chain rule for LD-derivatives is satisfied:*

$$[\mathbf{f} \circ \mathbf{g}]'(\mathbf{x};\mathbf{M}) = \mathbf{f}'(\mathbf{g}(\mathbf{x}); \mathbf{g}'(\mathbf{x};\mathbf{M})). \tag{3.2}$$

Proof. Assume temporarily that $\mathbf{g}$ and $\mathbf{f}$ are L-smooth on neighborhoods of $\mathbf{x}$ and $\mathbf{g}(\mathbf{x})$, respectively. Under this additional assumption, [79, Theorem 1] shows that the composition $\mathbf{f} \circ \mathbf{g}$ is L-smooth at $\mathbf{x}$, and that:

$$[\mathbf{f} \circ \mathbf{g}]^{(p)}_{\mathbf{x},\mathbf{M}}(\mathbf{d}) = \mathbf{f}^{(p)}_{\mathbf{g}(\mathbf{x}),[\mathbf{g}^{(1)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}_{(1)}) \,\cdots\, \mathbf{g}^{(p)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}_{(p)})]}\left(\mathbf{g}^{(p)}_{\mathbf{x},\mathbf{M}}(\mathbf{d})\right), \qquad \forall \mathbf{d} \in \mathbb{R}^n. \tag{3.3}$$

Now, observe that the proof of [79, Theorem 1] makes use of [79, Lemma 1], and

that [97, Theorem 3.1.1] yields the same result as [79, Lemma 1] under milder assumptions. Thus, if [79, Lemma 1] is replaced with [97, Theorem 3.1.1] throughout the proof of [79, Theorem 1], then the above results are shown to hold without requiring the assumption that was invoked at the start of the current proof. Hence, this additional assumption can be removed.

Next, applying [79, Lemma 3] and the definition of the LD-derivative to (3.3) yields

$$[\mathbf{f} \circ \mathbf{g}]^{(p)}_{\mathbf{x},\mathbf{M}}(\mathbf{d}) = \mathbf{f}^{(p)}_{\mathbf{g}(\mathbf{x}),\mathbf{g}'(\mathbf{x};\mathbf{M})}\left(\mathbf{g}^{(p)}_{\mathbf{x},\mathbf{M}}(\mathbf{d})\right), \qquad \forall \mathbf{d} \in \mathbb{R}^n.$$

Thus, for each $k \in \{1, \ldots, p\}$, the definition of the LD-derivative and the above relationship yield

$$\begin{aligned}
[\mathbf{f} \circ \mathbf{g}]'(\mathbf{x};\mathbf{M})\, \mathbf{e}_{(k)} &= [\mathbf{f} \circ \mathbf{g}]^{(p)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}_{(k)}), \\
&= \mathbf{f}^{(p)}_{\mathbf{g}(\mathbf{x}),\mathbf{g}'(\mathbf{x};\mathbf{M})}(\mathbf{g}^{(p)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}_{(k)})), \\
&= \mathbf{f}'(\mathbf{g}(\mathbf{x});\mathbf{g}'(\mathbf{x};\mathbf{M}))\, \mathbf{e}_{(k)}.
\end{aligned}$$

Arranging the above equations for all $k \in \{1, \ldots, p\}$ as the columns of a single matrix equation yields (3.2). □

This chain rule for LD-derivatives resembles the chain rule for directional derivatives of functions that are directionally differentiable and locally Lipschitz continuous [97, Theorem 3.1.1], and does not demand $\mathbf{M}$ to be nonsingular or square. Note that Proposition 3.1.2 reduces to Nesterov's chain rule [79, Theorem 5] when $\mathbf{M}$ is chosen to be square and nonsingular, and reduces further to the classical chain rule when, in addition, $\mathbf{f}$ and $\mathbf{g}$ are both differentiable.

The following analogs of classical calculus rules are immediate consequences of Proposition 3.1.2, where $\mathbf{u}$ and $\mathbf{v}$ are L-smooth functions with appropriate domains and ranges, and where ψ is differentiable on its open domain:

- $[\mathbf{u} + \mathbf{v}]'(\mathbf{x};\mathbf{M}) = \mathbf{u}'(\mathbf{x};\mathbf{M}) + \mathbf{v}'(\mathbf{x};\mathbf{M})$: obtained by setting $\mathbf{f} : (\mathbf{x}, \mathbf{y}) \mapsto \mathbf{x} + \mathbf{y}$ and $\mathbf{g} : \mathbf{x} \mapsto (\mathbf{u}(\mathbf{x}), \mathbf{v}(\mathbf{x}))$,

- $[uv]'(\mathbf{x};\mathbf{M}) = v(\mathbf{x})\, u'(\mathbf{x};\mathbf{M}) + u(\mathbf{x})\, v'(\mathbf{x};\mathbf{M})$: obtained by setting $f : (x,y) \mapsto xy$ and $\mathbf{g} : \mathbf{x} \mapsto (u(\mathbf{x}), v(\mathbf{x}))$,
- $[\psi \circ \mathbf{u}]'(\mathbf{x};\mathbf{M}) = \mathbf{J}\psi(\mathbf{u}(\mathbf{x}))\, \mathbf{u}'(\mathbf{x};\mathbf{M})$,
- $[\mathbf{u} \circ \psi]'(\mathbf{x};\mathbf{M}) = \mathbf{u}'(\psi(\mathbf{x}); \mathbf{J}\psi(\mathbf{x})\, \mathbf{M})$.

Consider any L-smooth function $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$, some $\mathbf{x} \in X$, and some $\mathbf{M} \in \mathbb{R}^{n\times p}$, and define an affine transformation $\mathcal{T}_{\mathbf{M},\mathbf{x}} : \mathbb{R}^p \to \mathbb{R}^n : \mathbf{y} \mapsto \mathbf{x} + \mathbf{M}\mathbf{y}$. The mapping $\mathcal{T}_{\mathbf{M},\mathbf{x}}$ is evidently differentiable. It follows that $\mathcal{T}_{\mathbf{M},\mathbf{x}}(\mathbf{0}) = \mathbf{x}$, and $[\mathcal{T}_{\mathbf{M},\mathbf{x}}]'(\mathbf{0};\mathbf{I}) = \mathbf{J}\mathcal{T}_{\mathbf{M},\mathbf{x}}(\mathbf{0}) = \mathbf{M}$. Thus,

$$\mathbf{f}'(\mathbf{x};\mathbf{M}) = \mathbf{f}'(\mathcal{T}_{\mathbf{M},\mathbf{x}}(\mathbf{0}); [\mathcal{T}_{\mathbf{M},\mathbf{x}}]'(\mathbf{0};\mathbf{I})) = [\mathbf{f} \circ \mathcal{T}_{\mathbf{M},\mathbf{x}}]'(\mathbf{0};\mathbf{I}) = \mathbf{J}_{\mathrm{L}}[\mathbf{f} \circ \mathcal{T}_{\mathbf{M},\mathbf{x}}](\mathbf{0};\mathbf{I}). \quad (3.4)$$

In this way, $\mathbf{f}'(\mathbf{x};\mathbf{M})$ is expressed as the lexicographic derivative of a related function along the standard basis. Using this relationship, properties of lexicographic derivatives can be extended to describe LD-derivatives.

Combining the above observations, to obtain $\mathbf{J}_{\mathrm{L}}\mathbf{f}(\mathbf{x};\mathbf{M})$ for some nontrivial function $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ and some nonsingular matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$, it is in many cases easier to compute $\mathbf{f}'(\mathbf{x};\mathbf{M})$ first using the calculus rules for the LD-derivative, and to then solve the linear equation system $\mathbf{J}_{\mathrm{L}}\mathbf{f}(\mathbf{x};\mathbf{M})\, \mathbf{M} = \mathbf{f}'(\mathbf{x};\mathbf{M})$ for $\mathbf{J}_{\mathrm{L}}\mathbf{f}(\mathbf{x};\mathbf{M})$. When there is freedom to choose $\mathbf{M}$, setting $\mathbf{M}$ to $\mathbf{I}$ renders this linear equation system trivial.

Lemma 3.1.3 (from [56])**.** *Given an open set $X \subset \mathbb{R}^n$, suppose that a function $\mathbf{g} : X \to \mathbb{R}^m$ is L-smooth at $\mathbf{x} \in X$. For each $\mathbf{M} \in \mathbb{R}^{n\times p}$,*

$$\mathbf{g}'(\mathbf{x};\mathbf{M}) \in \{\mathbf{A}\mathbf{M} : \mathbf{A} \in \partial_{\mathrm{L}}\mathbf{g}(\mathbf{x})\}.$$

Proof. Choose a matrix $\mathbf{B} \in \mathbb{R}^{n\times q}$ for which the block matrix $\mathbf{N} := \begin{bmatrix}\mathbf{M} & \mathbf{B}\end{bmatrix} \in \mathbb{R}^{n\times(p+q)}$ has full row rank. The definition of the LD-derivative and Lemma 2.3.7 yield:

$$\mathbf{g}'(\mathbf{x};\mathbf{M}) = \begin{bmatrix}\mathbf{g}^{(p+q)}_{\mathbf{x},\mathbf{N}}(\mathbf{m}_{(1)}) & \cdots & \mathbf{g}^{(p+q)}_{\mathbf{x},\mathbf{N}}(\mathbf{m}_{(p)})\end{bmatrix}.$$

Moreover, by [79, Theorem 2 and Lemma 4], the function $\mathbf{g}_{\mathbf{x},\mathbf{N}}^{(p+q)}$ is linear, and has a derivative $\mathbf{A} \in \partial_{\mathrm{L}}\mathbf{g}(\mathbf{x})$. The required result follows immediately. $\square$

The notion of LD-derivatives suggests the following definition, which exists alongside the similar types of factorable functions defined in Chapter 2.

Definition 3.1.4. *A factorable function* $\mathbf{f}$ *is* L-factorable *if the elemental library* $\mathcal{L}$ *contains only lexicographically smooth functions whose LD-derivatives are known or computable.*

Any $\mathcal{C}^1$-factorable or abs-$\mathcal{C}^1$-factorable function is evidently L-factorable. The methods in Chapter 4 evaluate LD-derivatives efficiently for L-factorable functions. The following definition of a subclass of $\mathcal{PC}^1$ functions is adapted from [54], which is reproduced for reference as Appendix A. This definition be used in Chapters 3 and 4.

It will be shown in Section 4.2 that all $\mathcal{PC}^1$-factorable functions are also L-factorable. As noted by Griewank [33], given a function that is L-factorable but not $\mathcal{PC}^1$, it may be impossible to construct an approximation of this function that is both piecewise affine in the sense of Scholtes [97] and uniformly first-order in the sense of Ralph [93].

3.2 Lexicographic derivatives and plenary Jacobians

This section is reproduced from [55]. Consider an open set $X \subset \mathbb{R}^n$, some $\mathbf{x} \in X$, and a function $\mathbf{f} : X \to \mathbb{R}^m$ that is both locally Lipschitz continuous and directionally differentiable. The main results of this section are the inclusions $\partial_{\mathrm{B}}[\mathbf{f}'(\mathbf{x};\cdot)](\mathbf{0}) \subset \partial_{\mathrm{P}}\mathbf{f}(\mathbf{x})$ and $\partial_{\mathrm{L}}\mathbf{f}(\mathbf{x}) \subset \partial_{\mathrm{P}}\mathbf{f}(\mathbf{x})$, with the latter result assuming further that $\mathbf{f}$ is L-smooth at $\mathbf{x}$. It follows immediately that any numerical or analytical method for evaluating an element of $\partial_{\mathrm{B}}[\mathbf{f}'(\mathbf{x};\cdot)](\mathbf{0})$ or $\partial_{\mathrm{L}}\mathbf{f}(\mathbf{x})$ is also a method for evaluating an element of $\partial_{\mathrm{P}}\mathbf{f}(\mathbf{x})$.

Lemma 3.2.1. *Consider a function* $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ *that is positively homogeneous and locally Lipschitz continuous*[1]*. For any* $\mathbf{d} \in \mathbb{R}^n$, $\partial_{\mathrm{B}}\mathbf{f}(\mathbf{d}) \subset \partial_{\mathrm{B}}\mathbf{f}(\mathbf{0})$.

Proof. The result is trivial when $\mathbf{d} = \mathbf{0}$, so assume that $\mathbf{d} \neq \mathbf{0}$, and consider any particular $\mathbf{H} \in \partial_{\mathrm{B}}\mathbf{f}(\mathbf{d})$. By definition of the B-subdifferential, there exists a sequence $\{\mathbf{d}^{(i)}\}_{i\in\mathbb{N}}$ in $\mathbb{R}^n$ converging to $\mathbf{d}$, such that $\mathbf{f}$ is differentiable at each $\mathbf{d}^{(i)}$, and such that $\lim_{i\to\infty} \mathbf{J}\mathbf{f}(\mathbf{d}^{(i)}) = \mathbf{H}$. Since $\mathbf{d} \neq \mathbf{0}$ and $\lim_{i\to\infty} \mathbf{d}^{(i)} = \mathbf{d}$, it may be assumed without loss of generality that $\mathbf{d}^{(i)} \neq \mathbf{0}$ for all $i \in \mathbb{N}$. Making use of the positive homogeneity of $\mathbf{f}$, for each $i \in \mathbb{N}$, each $t > 0$, and each nonzero $\mathbf{h} \in \mathbb{R}^n$,

$$\frac{\mathbf{f}(t\mathbf{d}^{(i)} + \mathbf{h}) - \mathbf{f}(t\mathbf{d}^{(i)}) - \mathbf{J}\mathbf{f}(\mathbf{d}^{(i)})\,\mathbf{h}}{\|\mathbf{h}\|} = \frac{\mathbf{f}(\mathbf{d}^{(i)} + \frac{1}{t}\mathbf{h}) - \mathbf{f}(\mathbf{d}^{(i)}) - \mathbf{J}\mathbf{f}(\mathbf{d}^{(i)})\,(\frac{1}{t}\mathbf{h})}{\|\frac{1}{t}\mathbf{h}\|}$$

Noting that $\mathbf{f}$ is differentiable at $\mathbf{d}^{(i)}$ and taking the limit $\mathbf{h} \to \mathbf{0}$ yields:

$$\begin{aligned}\mathbf{0} &= \lim_{\mathbf{h}\to\mathbf{0}} \frac{\mathbf{f}(\mathbf{d}^{(i)} + \frac{1}{t}\mathbf{h}) - \mathbf{f}(\mathbf{d}^{(i)}) - \mathbf{J}\mathbf{f}(\mathbf{d}^{(i)})\,(\frac{1}{t}\mathbf{h})}{\|\frac{1}{t}\mathbf{h}\|}, \\ &= \lim_{\mathbf{h}\to\mathbf{0}} \frac{\mathbf{f}(t\mathbf{d}^{(i)} + \mathbf{h}) - \mathbf{f}(t\mathbf{d}^{(i)}) - \mathbf{J}\mathbf{f}(\mathbf{d}^{(i)})\,\mathbf{h}}{\|\mathbf{h}\|}.\end{aligned}$$

Thus, for each $i \in \mathbb{N}$ and each $t > 0$, $\mathbf{f}$ is differentiable at $(t\mathbf{d}^{(i)})$, with a derivative of $\mathbf{J}\mathbf{f}(t\mathbf{d}^{(i)}) = \mathbf{J}\mathbf{f}(\mathbf{d}^{(i)})$. Since $\lim_{i\to\infty} \mathbf{J}\mathbf{f}(\mathbf{d}^{(i)}) = \mathbf{H}$, it follows that

$$\mathbf{H} = \lim_{i\to\infty} \mathbf{J}\mathbf{f}\left(\frac{\mathbf{d}^{(i)}}{2^i\|\mathbf{d}^{(i)}\|}\right).$$

Noting that $\lim_{i\to\infty}\left(\frac{\mathbf{d}^{(i)}}{2^i\|\mathbf{d}^{(i)}\|}\right) = \mathbf{0}$, it follows that $\mathbf{H} \in \partial_{\mathrm{B}}\mathbf{f}(\mathbf{0})$. □

Lemma 3.2.2. *Consider an open set* $X \subset \mathbb{R}^n$*, some* $\mathbf{x} \in X$*, and a function* $\mathbf{f} : X \to \mathbb{R}^m$ *that is locally Lipschitz continuous and directionally differentiable. If* $\mathbf{f}'(\mathbf{x};\cdot)$ *is differentiable at some particular* $\mathbf{d} \in \mathbb{R}^n$*, then* $\mathbf{J}[\mathbf{f}'(\mathbf{x};\cdot)](\mathbf{d}) \in \partial_{\mathrm{P}}\mathbf{f}(\mathbf{x})$.

Proof. For notational simplicity, define $\mathbf{A} := \mathbf{J}[\mathbf{f}'(\mathbf{x};\cdot)](\mathbf{d})$. The differentiability of $\mathbf{f}'(\mathbf{x};\cdot)$ at $\mathbf{d}$ implies that

[1] Though irrelevant to this lemma, if $k_{\mathbf{f}}$ is a Lipschitz constant for $\mathbf{f}$ in a neighborhood of $\mathbf{0}$, then $k_{\mathbf{f}}$ is a global Lipschitz constant for $\mathbf{f}$ [97].

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\mathbf{f}'(\mathbf{x};\mathbf{d}+\mathbf{h}) - \mathbf{f}'(\mathbf{x};\mathbf{d}) - \mathbf{A}\mathbf{h}}{\|\mathbf{h}\|} = \mathbf{0}. \tag{3.5}$$

To prove the lemma, the cases in which $\mathbf{d} = \mathbf{0}$ and $\mathbf{d} \neq \mathbf{0}$ will be considered separately. If $\mathbf{d} = \mathbf{0}$, then applying (3.5) and the positive homogeneity of $\mathbf{f}'(\mathbf{x};\cdot)$ yields $\mathbf{f}'(\mathbf{x};\mathbf{0}) = \mathbf{0}$, and

$$\mathbf{0} = \lim_{t\to 0^+} \frac{\mathbf{f}'(\mathbf{x};t\mathbf{h}) - \mathbf{f}'(\mathbf{x};\mathbf{0}) - t\mathbf{A}\mathbf{h}}{t\|\mathbf{h}\|} = \frac{\mathbf{f}'(\mathbf{x};\mathbf{h}) - \mathbf{A}\mathbf{h}}{\|\mathbf{h}\|}, \qquad \forall \mathbf{h} \in \mathbb{R}^n \backslash \{\mathbf{0}\}.$$

Combining these statements, $\mathbf{f}'(\mathbf{x};\mathbf{h}) = \mathbf{A}\mathbf{h}$ for each $\mathbf{h} \in \mathbb{R}^n$. Hence, $\mathbf{f}$ is Gâteaux differentiable at $\mathbf{x}$, with a Gâteaux derivative of $\mathbf{A}$. Since Gâteaux and Fréchet differentiability are equivalent for locally Lipschitz continuous functions on $\mathbb{R}^n$ [16], it follows that $\mathbf{f}$ is Fréchet differentiable at $\mathbf{x}$, with $\mathbf{J}\mathbf{f}(\mathbf{x}) = \mathbf{A}$. Thus, $\mathbf{A} \in \partial\mathbf{f}(\mathbf{x}) \subset \partial_{\mathrm{P}}\mathbf{f}(\mathbf{x})$, as required.

Now consider the case in which $\mathbf{d} \neq \mathbf{0}$. In light of Definition 2.3.4, it suffices to show that for any particular $\mathbf{e} \in \mathbb{R}^n$, $\mathbf{A}\mathbf{e} = \mathbf{H}\mathbf{e}$ for some $\mathbf{H} \in \partial\mathbf{f}(\mathbf{x})$. This statement is trivial when $\mathbf{e} = \mathbf{0}$, so assume that $\mathbf{e} \neq \mathbf{0}$. It follows from (3.5) that for any $\epsilon > 0$, there exists some $\delta_\epsilon > 0$ such that whenever $|\tau| < \delta_\epsilon$,

$$\|\mathbf{f}'(\mathbf{x};\mathbf{d}+\tau\mathbf{e}) - \mathbf{f}'(\mathbf{x};\mathbf{d}) - \mathbf{A}(\tau\mathbf{e})\| < \epsilon\|\tau\mathbf{e}\|.$$

It will be assumed that $\delta_\epsilon < 1$ without loss of generality, since otherwise, setting $\delta_\epsilon \leftarrow \min\{\delta_\epsilon, \frac{1}{2}\}$ does not affect the validity of the above statement. Since $\mathbf{f}'(\mathbf{x};\cdot)$ is positively homogeneous, multiplying both sides of the above inequality by any $\alpha > 0$ and setting $\tau := \frac{1}{2}\delta_\epsilon$ yields:

$$\|\mathbf{f}'(\mathbf{x};\alpha(\mathbf{d}+\tfrac{1}{2}\delta_\epsilon\mathbf{e})) - \mathbf{f}'(\mathbf{x};\alpha\mathbf{d}) - \tfrac{1}{2}\alpha\delta_\epsilon\mathbf{A}\mathbf{e}\| < \tfrac{1}{2}\epsilon\alpha\delta_\epsilon\|\mathbf{e}\|, \qquad \forall \alpha > 0. \tag{3.6}$$

It follows from (2.1) that for any $\epsilon > 0$, there exists some $\bar{\delta}_\epsilon > 0$ such that whenever $\|\mathbf{v}\| \leq \bar{\delta}_\epsilon$, $\mathbf{f}$ is defined at $(\mathbf{x}+\mathbf{v})$, and

$$\|\mathbf{f}(\mathbf{x}+\mathbf{v}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x};\mathbf{v})\| < \epsilon\delta_\epsilon\|\mathbf{v}\|. \tag{3.7}$$

It will be assumed that $\lim_{\epsilon \to 0^+} \bar{\delta}_\epsilon = 0$ without loss of generality, since otherwise, setting $\bar{\delta}_\epsilon \leftarrow \min\{\bar{\delta}_\epsilon, \epsilon\}$ does not affect the validity of the above statement.

Now, choose any fixed $\epsilon > 0$, and set

$$\alpha_\epsilon := \frac{\bar{\delta}_\epsilon}{\|\mathbf{d}\| + \frac{1}{2}\delta_\epsilon \|\mathbf{e}\|} > 0.$$

The triangle inequality shows that for each $\tau \in [0, \frac{1}{2}\delta_\epsilon]$,

$$\alpha_\epsilon \|\mathbf{d} + \tau \mathbf{e}\| \leq \alpha_\epsilon (\|\mathbf{d}\| + \tau \|\mathbf{e}\|) \leq \alpha_\epsilon (\|\mathbf{d}\| + \tfrac{1}{2}\delta_\epsilon \|\mathbf{e}\|) = \bar{\delta}_\epsilon. \tag{3.8}$$

Thus, in (3.7), $\mathbf{v}$ may be set to $(\alpha_\epsilon(\mathbf{d} + \tau \mathbf{e}))$ for any $\tau \in [0, \frac{1}{2}\delta_\epsilon]$ to yield:

$$\|\mathbf{f}(\mathbf{x} + \alpha_\epsilon(\mathbf{d} + \tau \mathbf{e})) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x}; \alpha_\epsilon(\mathbf{d} + \tau \mathbf{e}))\| < \epsilon \delta_\epsilon \alpha_\epsilon \|\mathbf{d} + \tau \mathbf{e}\| \leq \epsilon \delta_\epsilon \bar{\delta}_\epsilon, \tag{3.9}$$

Setting τ to 0 and $\frac{1}{2}\delta_\epsilon$ in (3.9), respectively, yields:

$$\|\mathbf{f}(\mathbf{x}) + \mathbf{f}'(\mathbf{x}; \alpha_\epsilon \mathbf{d}) - \mathbf{f}(\mathbf{x} + \alpha_\epsilon \mathbf{d})\| < \epsilon \delta_\epsilon \bar{\delta}_\epsilon, \tag{3.10}$$

$$\|\mathbf{f}(\mathbf{x} + \alpha_\epsilon(\mathbf{d} + \tfrac{1}{2}\delta_\epsilon \mathbf{e})) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x}; \alpha_\epsilon(\mathbf{d} + \tfrac{1}{2}\delta_\epsilon \mathbf{e}))\| < \epsilon \delta_\epsilon \bar{\delta}_\epsilon. \tag{3.11}$$

Setting α to α_ϵ in (3.6), adding (3.10) and (3.11), and applying the triangle inequality yields:

$$\|\mathbf{f}(\mathbf{x} + \alpha_\epsilon(\mathbf{d} + \tfrac{1}{2}\delta_\epsilon \mathbf{e})) - \mathbf{f}(\mathbf{x} + \alpha_\epsilon \mathbf{d}) - \tfrac{1}{2}\alpha_\epsilon \delta_\epsilon \mathbf{A}\mathbf{e}\| < \epsilon \delta_\epsilon (\tfrac{1}{2}\alpha_\epsilon \|\mathbf{e}\| + 2\bar{\delta}_\epsilon). \tag{3.12}$$

Now, Clarke's mean value theorem for locally Lipschitz continuous functions [16, Proposition 2.6.5] implies that

$$\begin{aligned} &\mathbf{f}(\mathbf{x} + \alpha_\epsilon(\mathbf{d} + \tfrac{1}{2}\delta_\epsilon \mathbf{e})) - \mathbf{f}(\mathbf{x} + \alpha_\epsilon \mathbf{d}) \\ &\qquad \in \operatorname{conv}\{\tfrac{1}{2}\alpha_\epsilon \delta_\epsilon \mathbf{H}\mathbf{e} : \exists \tau \in [0, \tfrac{1}{2}\delta_\epsilon] \text{ s.t. } \mathbf{H} \in \partial \mathbf{f}(\mathbf{x} + \alpha_\epsilon(\mathbf{d} + \tau \mathbf{e}))\}. \end{aligned}$$

Substituting this result into (3.12) and applying the Carathéodory Theorem yields the existence of $\lambda_\epsilon^{(i)} \in [0, 1]$, $\tau_\epsilon^{(i)} \in [0, \frac{1}{2}\delta_\epsilon]$, and $\mathbf{H}_\epsilon^{(i)} \in \partial \mathbf{f}(\mathbf{x} + \alpha_\epsilon(\mathbf{d} + \tau_\epsilon^{(i)} \mathbf{e}))$ for each $i \in \{1, 2, \ldots, m+1\}$ such that:

$$1 = \sum_{i=1}^{m+1} \lambda_\epsilon^{(i)}, \quad \text{and} \quad \left\| \frac{1}{2} \sum_{i=1}^{m+1} \lambda_\epsilon^{(i)} \alpha_\epsilon \delta_\epsilon \mathbf{H}_\epsilon^{(i)} \mathbf{e} - \frac{1}{2} \alpha_\epsilon \delta_\epsilon \mathbf{A}\mathbf{e} \right\| < \epsilon \delta_\epsilon (\tfrac{1}{2} \alpha_\epsilon \|\mathbf{e}\| + 2\bar{\delta}_\epsilon). \tag{3.13}$$

Dividing both sides of the above inequality by $\frac{1}{2}\alpha_\epsilon \delta_\epsilon$, applying the definition of α_ϵ, and noting that $\delta_\epsilon < 1$ yields:

$$\left\| \sum_{i=1}^{m+1} \lambda_\epsilon^{(i)} \mathbf{H}_\epsilon^{(i)} \mathbf{e} - \mathbf{A}\mathbf{e} \right\| < \epsilon \left(\|\mathbf{e}\| + \frac{4\bar{\delta}_\epsilon}{\alpha_\epsilon} \right) < \epsilon (3\|\mathbf{e}\| + 4\|\mathbf{d}\|). \tag{3.14}$$

For each $\epsilon > 0$ and each $i \in \{1, \ldots, m+1\}$, $\lambda_\epsilon^{(i)}$ is an element of the compact set $[0, 1] \subset \mathbb{R}$, and $\tau_\epsilon^{(i)}$ is an element of the compact set $[0, \frac{1}{2}\delta_\epsilon]$. Moreover, if $k_{\mathbf{f}}$ denotes a Lipschitz constant for $\mathbf{f}$ on $\{\mathbf{y} \in X : \|\mathbf{y} - \mathbf{x}\| \leq \bar{\delta}_1\}$, then, noting that $\lim_{\epsilon \to 0^+} \bar{\delta}_\epsilon = 0$, it follows from (3.8) and [16, Proposition 2.6.2(d)] that for sufficiently small $\epsilon > 0$, $\mathbf{H}_\epsilon^{(i)}$ is an element of the compact set $\{\mathbf{H} \in \mathbb{R}^{m \times n} : \|\mathbf{H}\| \leq k_{\mathbf{f}}\}$ for each $i \in \{1, \ldots, m+1\}$.

Since any sequence in a compact set has a convergent subsequence, it follows that there exists a sequence $\{\epsilon_j\}_{j \in \mathbb{N}}$ such that each $\epsilon_j > 0$, $\lim_{j \to \infty} \epsilon_j = 0$, and the sequences $\{\lambda_{\epsilon_j}^{(i)}\}_{j \in \mathbb{N}}$, $\{\tau_{\epsilon_j}^{(i)}\}_{j \in \mathbb{N}}$, and $\{\mathbf{H}_{\epsilon_j}^{(i)}\}_{j \in \mathbb{N}}$ converge for each $i \in \{1, \ldots, m+1\}$, permitting the following definitions:

$$\bar{\lambda}^{(i)} := \lim_{j \to \infty} \lambda_{\epsilon_j}^{(i)}, \qquad \bar{\tau}^{(i)} := \lim_{j \to \infty} \tau_{\epsilon_j}^{(i)}, \qquad \text{and} \quad \bar{\mathbf{H}}^{(i)} := \lim_{j \to \infty} \mathbf{H}_{\epsilon_j}^{(i)}.$$

It follows from (3.13) and (3.14) that

$$1 = \sum_{i=1}^{m+1} \bar{\lambda}^{(i)}, \qquad \text{and} \qquad \left\| \sum_{i=1}^{m+1} \bar{\lambda}^{(i)} \bar{\mathbf{H}}^{(i)} \mathbf{e} - \mathbf{A}\mathbf{e} \right\| = 0. \tag{3.15}$$

Since each $\tau_{\epsilon_j}^{(i)} \in [0, \frac{1}{2}\delta_\epsilon]$, applying (3.8) with $\tau := \tau_{\epsilon_j}^{(i)}$ and $\epsilon := \epsilon_j$, taking the limit $j \to \infty$, and noting that $\lim_{\epsilon \to 0^+} \bar{\delta}_\epsilon = 0$,

$$0 \leq \limsup_{j \to \infty} \left\| \alpha_{\epsilon_j} (\mathbf{d} + \tau_{\epsilon_j}^{(i)} \mathbf{e}) \right\| \leq \lim_{j \to \infty} \bar{\delta}_{\epsilon_j} = 0.$$

Thus, for each $i \in \{1, \ldots, m+1\}$, $\lim_{j \to \infty} (\mathbf{x} + \alpha_{\epsilon_j} (\mathbf{d} + \tau_{\epsilon_j}^{(i)} \mathbf{e})) = \mathbf{x}$. Moreover, by construction,

$$\mathbf{H}_{\epsilon_j}^{(i)} \in \partial \mathbf{f}(\mathbf{x} + \alpha_{\epsilon_j}(\mathbf{d} + \tau_{\epsilon_j}^{(i)}\mathbf{e})), \qquad \forall i \in \{1, \ldots, m+1\}, \qquad \forall j \in \mathbb{N}.$$

The upper-semicontinuity of Clarke's generalized Jacobian then yields $\bar{\mathbf{H}}^{(i)} \in \partial \mathbf{f}(\mathbf{x})$. Since $\partial \mathbf{f}(\mathbf{x})$ is convex and $\sum_{i=1}^{m+1} \bar{\lambda}^{(i)} = 1$, it follows that $\bar{\mathbf{H}} := \sum_{i=1}^{m+1} \bar{\lambda}^{(i)} \bar{\mathbf{H}}^{(i)} \in \partial \mathbf{f}(\mathbf{x})$. Moreover, (3.15) shows that $\bar{\mathbf{H}}\mathbf{e} = \mathbf{A}\mathbf{e}$, as required. □

Theorem 3.2.3. *Given an open set $X \subset \mathbb{R}^n$ and a function $\mathbf{f} : X \to \mathbb{R}^m$ that is locally Lipschitz continuous and directionally differentiable, $\partial_{\mathrm{B}}[\mathbf{f}'(\mathbf{x}; \cdot)](\mathbf{0}) \subset \partial_{\mathrm{P}} \mathbf{f}(\mathbf{x})$ for each $\mathbf{x} \in X$.*

Proof. Consider any particular $\mathbf{x} \in X$ and any particular $\mathbf{H} \in \partial_{\mathrm{B}}[\mathbf{f}'(\mathbf{x}; \cdot)](\mathbf{0})$. By definition of the B-subdifferential, there exists a sequence $\{\mathbf{d}^{(i)}\}_{i \in \mathbb{N}}$ in $\mathbb{R}^n$ such that $\mathbf{f}'(\mathbf{x}; \cdot)$ is differentiable at each $\mathbf{d}^{(i)}$, and $\lim_{i \to \infty} \mathbf{J}[\mathbf{f}'(\mathbf{x}; \cdot)](\mathbf{d}^{(i)}) = \mathbf{H}$. By Lemma 3.2.2, $\mathbf{J}[\mathbf{f}'(\mathbf{x}; \cdot)](\mathbf{d}^{(i)}) \in \partial_{\mathrm{P}} \mathbf{f}(\mathbf{x})$ for each $i \in \mathbb{N}$. Since $\partial_{\mathrm{P}} \mathbf{f}(\mathbf{x})$ is a closed set [109], it follows that $\mathbf{H} \in \partial_{\mathrm{P}} \mathbf{f}(\mathbf{x})$. □

Corollary 3.2.4. *Given an open set $X \subset \mathbb{R}^n$ and a function $\mathbf{f} : X \to \mathbb{R}^m$ that is locally Lipschitz continuous and directionally differentiable,*

$$\partial_{\mathrm{B}}[\mathbf{f}'(\mathbf{x}; \cdot)](\mathbf{0}) \subset \partial[\mathbf{f}'(\mathbf{x}; \cdot)](\mathbf{0}) \subset \partial_{\mathrm{P}}[\mathbf{f}'(\mathbf{x}; \cdot)](\mathbf{0}) \subset \partial_{\mathrm{P}} \mathbf{f}(\mathbf{x}), \qquad \forall \mathbf{x} \in X.$$

Proof. Consider any particular $\mathbf{x} \in X$. The inclusions

$$\partial_{\mathrm{B}}[\mathbf{f}'(\mathbf{x}; \cdot)](\mathbf{0}) \subset \partial[\mathbf{f}'(\mathbf{x}; \cdot)](\mathbf{0}) \subset \partial_{\mathrm{P}}[\mathbf{f}'(\mathbf{x}; \cdot)](\mathbf{0})$$

follow immediately from the definitions of Clarke's generalized Jacobian and the plenary hull. Now, Theorem 3.2.3 yields the inclusion $\partial_{\mathrm{B}}[\mathbf{f}'(\mathbf{x}; \cdot)](\mathbf{0}) \subset \partial_{\mathrm{P}} \mathbf{f}(\mathbf{x})$. Since $\partial_{\mathrm{P}} \mathbf{f}(\mathbf{x})$ is convex [109], and since the convex hull of $\partial_{\mathrm{B}}[\mathbf{f}'(\mathbf{x}; \cdot)](\mathbf{0})$ is the intersection of all of its convex supersets in $\mathbb{R}^{m \times n}$, it follows that

$$\operatorname{conv} \partial_{\mathrm{B}}[\mathbf{f}'(\mathbf{x}; \cdot)](\mathbf{0}) = \partial[\mathbf{f}'(\mathbf{x}; \cdot)](\mathbf{0}) \subset \partial_{\mathrm{P}} \mathbf{f}(\mathbf{x}).$$

Since $\partial_{\mathrm{P}}\mathbf{f}(\mathbf{x})$ is plenary, and since the plenary hull of $\partial[\mathbf{f}'(\mathbf{x};\cdot)](\mathbf{0})$ is the intersection of all of its plenary supersets in $\mathbb{R}^{m\times n}$, it follows that $\partial_{\mathrm{P}}[\mathbf{f}'(\mathbf{x};\cdot)](\mathbf{0}) \subset \partial_{\mathrm{P}}\mathbf{f}(\mathbf{x})$. The required chain of inclusions is therefore established. □

Corollary 3.2.5. *Given an open set $X \subset \mathbb{R}^n$ and a function $\mathbf{f} : X \to \mathbb{R}^m$ that is L-smooth, $\partial_{\mathrm{L}}\mathbf{f}(\mathbf{x}) \subset \partial_{\mathrm{P}}\mathbf{f}(\mathbf{x})$ for each $\mathbf{x} \in X$.*

Proof. Consider any particular $\mathbf{x} \in X$ and any particular $\mathbf{H} \in \partial_{\mathrm{L}}\mathbf{f}(\mathbf{x})$. By definition of $\partial_{\mathrm{L}}\mathbf{f}(\mathbf{x})$, there exists some nonsingular matrix $\mathbf{M} := \begin{bmatrix}\mathbf{m}_{(1)} & \cdots & \mathbf{m}_{(n)}\end{bmatrix} \in \mathbb{R}^{n\times n}$ such that the following functions are well-defined:

$$\begin{aligned}
&\mathbf{f}^{(0)}_{\mathbf{x},\mathbf{M}} : \mathbb{R}^n \to \mathbb{R}^m : \mathbf{h} \mapsto \mathbf{f}'(\mathbf{x};\mathbf{h}),\\
&\mathbf{f}^{(j)}_{\mathbf{x},\mathbf{M}} : \mathbb{R}^n \to \mathbb{R}^m : \mathbf{h} \mapsto [\mathbf{f}^{(j-1)}_{\mathbf{x},\mathbf{M}}]'(\mathbf{m}_{(j)};\mathbf{h}), \qquad \forall j \in \{1,\ldots,n\},
\end{aligned}$$

and such that $\mathbf{f}^{(n)}_{\mathbf{x},\mathbf{M}}$ is linear (and therefore differentiable) on its domain, with a derivative of $\mathbf{J}\mathbf{f}^{(n)}_{\mathbf{x},\mathbf{M}}(\mathbf{0}) = \mathbf{H}$. As an intermediate result, it will be proved by induction on $k = n, (n-1), \ldots, 0$ that for each $k \in \{0, 1, \ldots, n\}$, $\mathbf{H} \in \partial_{\mathrm{P}}\mathbf{f}^{(k)}_{\mathbf{x},\mathbf{M}}(\mathbf{0})$. For the base case, the differentiability of $\mathbf{f}^{(n)}_{\mathbf{x},\mathbf{M}}$ implies that $\mathbf{H} = \mathbf{J}\mathbf{f}^{(n)}_{\mathbf{x},\mathbf{M}}(\mathbf{0}) \in \partial\mathbf{f}^{(n)}_{\mathbf{x},\mathbf{M}}(\mathbf{0}) \subset \partial_{\mathrm{P}}\mathbf{f}^{(n)}_{\mathbf{x},\mathbf{M}}(\mathbf{0})$.

For the inductive step, suppose that for some $k \in \{1, \ldots, n\}$, $\mathbf{H} \in \partial_{\mathrm{P}}\mathbf{f}^{(k)}_{\mathbf{x},\mathbf{M}}(\mathbf{0})$. It follows from the construction of $\mathbf{f}^{(k)}_{\mathbf{x},\mathbf{M}}$ that $\mathbf{f}^{(k-1)}_{\mathbf{x},\mathbf{M}}$ is directionally differentiable. Moreover, it follows from repeated application of [97, Theorem 3.1.2] that $\mathbf{f}^{(k-1)}_{\mathbf{x},\mathbf{M}}$ is Lipschitz continuous. Thus, noting that $\mathbf{f}^{(k)}_{\mathbf{x},\mathbf{M}} \equiv [\mathbf{f}^{(k-1)}_{\mathbf{x},\mathbf{M}}]'(\mathbf{m}_{(k)};\cdot)$, Corollary 3.2.4 implies that $\partial_{\mathrm{P}}\mathbf{f}^{(k)}_{\mathbf{x},\mathbf{M}}(\mathbf{0}) \subset \partial_{\mathrm{P}}\mathbf{f}^{(k-1)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}_{(k)})$. Applying the inductive assumption then yields:

$$\mathbf{H} \in \partial_{\mathrm{P}}\mathbf{f}^{(k-1)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}_{(k)}). \tag{3.16}$$

Now, Lemma 2.3.7 implies that $\mathbf{f}^{(k-1)}_{\mathbf{x},\mathbf{M}}$ is positively homogeneous, in which case Lemma 3.2.1 yields

$$\partial_{\mathrm{B}}\mathbf{f}^{(k-1)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}_{(k)}) \subset \partial_{\mathrm{B}}\mathbf{f}^{(k-1)}_{\mathbf{x},\mathbf{M}}(\mathbf{0}) \subset \partial_{\mathrm{P}}\mathbf{f}^{(k-1)}_{\mathbf{x},\mathbf{M}}(\mathbf{0}).$$

Since $\partial_{\mathrm{P}}\mathbf{f}^{(k-1)}_{\mathbf{x},\mathbf{M}}(\mathbf{0})$ is convex, it follows that $\partial\mathbf{f}^{(k-1)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}_{(k)}) \subset \partial_{\mathrm{P}}\mathbf{f}^{(k-1)}_{\mathbf{x},\mathbf{M}}(\mathbf{0})$. Similarly,

since $\partial_{\mathrm{P}}\mathbf{f}^{(k-1)}_{\mathbf{x},\mathbf{M}}(\mathbf{0})$ is plenary, it follows that $\partial_{\mathrm{P}}\mathbf{f}^{(k-1)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}_{(k)}) \subset \partial_{\mathrm{P}}\mathbf{f}^{(k-1)}_{\mathbf{x},\mathbf{M}}(\mathbf{0})$. Thus, (3.16) implies that $\mathbf{H} \in \partial_{\mathrm{P}}\mathbf{f}^{(k-1)}_{\mathbf{x},\mathbf{M}}(\mathbf{0})$, which completes the inductive step.

It follows from this inductive proof that $\mathbf{H} \in \partial_{\mathrm{P}}\mathbf{f}^{(0)}_{\mathbf{x},\mathbf{M}}(\mathbf{0}) = \partial_{\mathrm{P}}[\mathbf{f}'(\mathbf{x};\cdot)](\mathbf{0})$. A final application of Corollary 3.2.4 yields $\mathbf{H} \in \partial_{\mathrm{P}}\mathbf{f}(\mathbf{x})$. □

3.3 Specialization to $\mathcal{PC}^1$ functions

This section is reproduced from [61], and provides a stronger version of the result of the previous section for $\mathcal{PC}^1$ functions, using results from [54]. The main results of this section are that for a $\mathcal{PC}^1$ function $\mathbf{f}$ and a domain point $\mathbf{x}$, $\mathbf{f}$ is L-smooth, and $\partial_{\mathrm{L}}\mathbf{f}(\mathbf{x}) = \partial_{\mathrm{B}}[\mathbf{f}'(\mathbf{x};\cdot)](\mathbf{0}) \subset \partial_{\mathrm{B}}\mathbf{f}(\mathbf{x})$. These relationships will be exploited in Chapter 4 to yield improved methods for evaluating generalized derivatives for $\mathcal{PC}^1$-factorable functions.

Proposition 3.3.1. *Given an open set $X \subset \mathbb{R}^n$, any $\mathcal{PC}^1$ function $\mathbf{f} : X \to \mathbb{R}^m$ is L-smooth.*

Proof. Choose any such $\mathbf{f}$, some $\mathbf{x} \in X$, and some $\mathbf{M} := \begin{bmatrix} \mathbf{m}_{(1)} & \cdots & \mathbf{m}_{(p)} \end{bmatrix} \in \mathbb{R}^{n\times p}$ with $p \in \mathbb{N}$. Since $\mathbf{f}$ is $\mathcal{PC}^1$, it is also locally Lipschitz continuous. It will be shown by induction on $k \in \{0,\ldots,p\}$ that $\mathbf{f}^{(k)}_{\mathbf{x},\mathbf{M}}$ exists and is piecewise linear.

For the case in which $k = 0$, [97, Proposition 4.1.3] shows that $\mathbf{f}$ is directionally differentiable at $\mathbf{x}$, and that $\mathbf{f}'(\mathbf{x};\cdot)$ is piecewise linear. Hence, $\mathbf{f}^{(0)}_{\mathbf{x},\mathbf{M}} \equiv \mathbf{f}'(\mathbf{x};\cdot)$ exists and is piecewise linear.

Now consider any $k^* \in \{1,\ldots,p\}$, and suppose that the required result holds with $k := k^* - 1$. By the inductive assumption, $\mathbf{f}^{(k^*-1)}_{\mathbf{x},\mathbf{M}}$ exists and is piecewise linear. Thus, [97, Proposition 4.1.3] shows that $\mathbf{f}^{(k^*-1)}_{\mathbf{x},\mathbf{M}}$ is directionally differentiable, and that $\mathbf{f}^{(k^*)}_{\mathbf{x},\mathbf{M}} \equiv [\mathbf{f}^{(k^*-1)}_{\mathbf{x},\mathbf{M}}]'(\mathbf{m}_{(k^*)};\cdot)$ is piecewise linear. This completes the inductive argument. Since $\mathbf{x}$ and $\mathbf{M}$ were chosen arbitrarily, it follows that $\mathbf{f}$ is L-smooth. □

Lemma 3.3.2. *Given a polyhedral cone $C \subset \mathbb{R}^n$ with nonempty interior, there exist n linearly independent vectors $\mathbf{q}_{(1)},\ldots,\mathbf{q}_{(n)} \in \mathbb{R}^n$ such that* $\operatorname{cone}\{\mathbf{q}_{(1)},\ldots,\mathbf{q}_{(n)}\} \subset C$.

Proof. By the Farkas-Minkowski-Weyl Theorem (which is summarized as [97, Theorem 2.1.1]), there exists a matrix $\mathbf{M} = \begin{bmatrix} \mathbf{m}_{(1)} & \cdots & \mathbf{m}_{(p)} \end{bmatrix} \in \mathbb{R}^{n \times p}$ for which

$$C = \{\mathbf{M}\mathbf{v} : \mathbf{v} \in \mathbb{R}^p_+\} = \operatorname{cone}\{\mathbf{m}_{(1)}, \ldots, \mathbf{m}_{(p)}\}. \tag{3.17}$$

To obtain a contradiction, suppose that there do not exist n linearly independent vectors satisfying the statement of the lemma. By (3.17), each column of $\mathbf{M}$ is an element of C. It follows that the columns of $\mathbf{M}$ do not span $\mathbb{R}^n$, implying the existence of a vector $\mathbf{d} \in \mathbb{R}^n$ which is not in the column space of $\mathbf{M}$. Since C has nonempty interior, there exists $\mathbf{v} \in \mathbb{R}^p_+$ such that $\mathbf{M}\mathbf{v} \in \operatorname{int}(C)$. Hence, for sufficiently small $\epsilon > 0$, $(\mathbf{M}\mathbf{v} + \epsilon\mathbf{d}) \in C$, and so $(\mathbf{M}\mathbf{v} + \epsilon\mathbf{d}) = \mathbf{M}\mathbf{u}$ for some $\mathbf{u} \in \mathbb{R}^p_+$. It follows that $\mathbf{d} = \mathbf{M}\frac{1}{\epsilon}(\mathbf{u} - \mathbf{v})$, so $\mathbf{d}$ is in the column space of $\mathbf{M}$, which contradicts the definition of $\mathbf{d}$. □

Theorem 3.3.3. *Given an open set $X \subset \mathbb{R}^n$, a vector $\mathbf{x} \in X$, and a $\mathcal{PC}^1$-factorable function $\mathbf{f} : X \to \mathbb{R}^m$, $\partial_{\mathrm{C}}\mathbf{f}(\mathbf{x}) = \partial_{\mathrm{B}}[\mathbf{f}'(\mathbf{x}; \cdot)](\mathbf{0})$.*

Proof. To show that $\partial_{\mathrm{C}}\mathbf{f}(\mathbf{x}) \subset \partial_{\mathrm{B}}[\mathbf{f}'(\mathbf{x}; \cdot)](\mathbf{0})$, consider any $\mathbf{H} \in \partial_{\mathrm{C}}\mathbf{f}(\mathbf{x})$. By construction of $\partial_{\mathrm{C}}\mathbf{f}(\mathbf{x})$, there exists some nonsingular matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$ such that $\mathbf{H} = \mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{x}; \mathbf{M})$.

Since polyhedral cones are closed under nonnegative combinations of their elements, it follows that for any $\alpha > 0$, the operation $\mathbf{q}_{(k)} \leftarrow \mathbf{q}_{(k)} + \alpha\mathbf{q}_{(k^*)}$ transforms a polyhedral cone $\bar{C} := \operatorname{cone}\{\mathbf{q}_{(1)}, \ldots, \mathbf{q}_{(n)}\}$ into a subset of itself. Thus, Theorem A.3.2, Corollary A.6.3, and Lemma A.6.6 imply that there exists a polyhedral cone $C \subset \operatorname{cone}\{\mathbf{m}_{(1)}, \ldots, \mathbf{m}_{(n)}\} \subset \mathbb{R}^n$ with nonempty interior, for which

$$\mathbf{f}'(\mathbf{x}; \mathbf{d}) = \mathbf{H}\mathbf{d}, \qquad \forall \mathbf{d} \in C.$$

Since C is closed under multiplication by a positive scalar, it follows that if $t > 0$ and $\bar{\mathbf{d}} \in \operatorname{int}(C)$, then $t\bar{\mathbf{d}} \in \operatorname{int}(C)$ as well, and so $\mathbf{f}'(\mathbf{x}; \mathbf{d}) = \mathbf{H}\mathbf{d}$ for each $\mathbf{d}$ in a sufficiently small neighborhood of $t\bar{\mathbf{d}}$. Thus, for each $t > 0$, $\mathbf{f}'(\mathbf{x}; \cdot)$ is differentiable at $t\bar{\mathbf{d}}$, with $\mathbf{J}[\mathbf{f}'(\mathbf{x}; \cdot)](t\bar{\mathbf{d}}) = \mathbf{H}$. Taking the limit $t \to 0^+$ yields $\mathbf{H} \in \partial_{\mathrm{B}}[\mathbf{f}'(\mathbf{x}; \cdot)](\mathbf{0})$, as required.

To show that $\partial_{\mathrm{B}}[\mathbf{f}'(\mathbf{x};\cdot)](\mathbf{0}) \subset \partial_{\mathrm{C}}\mathbf{f}(\mathbf{x})$, consider any $\mathbf{H} \in \partial_{\mathrm{B}}[\mathbf{f}'(\mathbf{x};\cdot)](\mathbf{0})$. Since $\mathbf{f}$ is $\mathcal{PC}^1$, [97, Proposition 4.1.3] implies that $\mathbf{f}'(\mathbf{x};\cdot)$ is $\mathcal{PL}$. By [97, Proposition 2.2.3], there exists a conical subdivision Λ of $\mathbb{R}^n$ such that $\mathbf{f}'(\mathbf{x};\cdot)$ is linear on each particular cone $C \in \Lambda$, with some Jacobian $\mathbf{H}_C \in \mathbb{R}^{m\times n}$. Thus, for each $\mathbf{d} \in \mathbb{R}^n$, $\mathbf{f}'(\mathbf{x};\mathbf{d}) \in \{\mathbf{H}_C\,\mathbf{d} : C \in \Lambda\}$. It follows from the proof of [97, Proposition 4.3.1] that

$$\partial_{\mathrm{B}}[\mathbf{f}'(\mathbf{x};\cdot)](\mathbf{0}) \subset \{\mathbf{H}_C : C \in \Lambda\}.$$

Thus, $\mathbf{H} = \mathbf{H}_C$ for some $C \in \Lambda$. Since C has nonempty interior, Lemma 3.3.2 implies the existence of linearly independent vectors $\mathbf{q}_{(1)}, \ldots, \mathbf{q}_{(n)} \in C$. Thus,

$$\mathbf{f}'(\mathbf{x};\mathbf{d}) = \mathbf{H}\mathbf{d}, \qquad \forall \mathbf{d} \in \operatorname{cone}\{\mathbf{q}_{(1)}, \ldots, \mathbf{q}_{(n)}\}.$$

Noting that nonnegative combinations of elements of $Q := \operatorname{cone}\{\mathbf{q}_{(1)}, \ldots, \mathbf{q}_{(n)}\}$ are themselves elements of Q, it follows from Theorem A.3.2 and Lemma A.6.6 with $j := \ell$ that $\mathbf{H} = \mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{x}; \begin{bmatrix}\mathbf{q}_{(1)} & \cdots & \mathbf{q}_{(n)}\end{bmatrix}) \in \partial_{\mathrm{C}}\mathbf{f}(\mathbf{x})$. □

Theorem 3.3.4. *Given an open set $X \subset \mathbb{R}^n$, a vector $\mathbf{x} \in X$, a $\mathcal{PC}^1$-factorable function $\mathbf{f} : X \to \mathbb{R}^m$, and a nonsingular matrix $\mathbf{M} = \begin{bmatrix}\mathbf{m}_{(1)} & \cdots & \mathbf{m}_{(n)}\end{bmatrix} \in \mathbb{R}^{n\times n}$, $\mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{x};\mathbf{M}) = \mathbf{J}_{\mathrm{L}}\mathbf{f}(\mathbf{x};\mathbf{M})$.*

Proof. Since $\mathbf{f}$ is $\mathcal{PC}^1$, Proposition 3.3.1 implies that it is also L-smooth. According to Lemma A.6.2, Lemma A.6.6, and Theorem A.3.2, there exists $\ell_{k,j} \geq 0$ for each $k \in \{1, \ldots, n\}$ and each $j \in \{1, \ldots, k\}$ such that $\ell_{k,k} = 1$ for each k, and if

$$\mathbf{q}_{(k)} := \sum_{j=1}^{k} \ell_{k,j}\mathbf{m}_{(j)}, \qquad \forall k \in \{1, \ldots, n\},$$

then

$$\mathbf{f}'(\mathbf{x};\mathbf{d}) = \mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{x};\mathbf{M})\,\mathbf{d}, \qquad \forall \mathbf{d} \in \operatorname{cone}\{\mathbf{q}_{(1)}, \ldots, \mathbf{q}_{(n)}\}. \tag{3.18}$$

To obtain the required result, it will be proved by induction on $k \in \{0, 1, \ldots, n\}$ that

$$\mathbf{f}^{(k)}_{\mathbf{x},\mathbf{M}}(\mathbf{d}) = \mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{x};\mathbf{M})\,\mathbf{d}, \qquad \forall \mathbf{d} \in \operatorname{span}\{\mathbf{m}_{(1)},\ldots,\mathbf{m}_{(k)}\} + \operatorname{cone}\{\mathbf{q}_{(k+1)},\ldots,\mathbf{q}_{(n)}\}. \tag{3.19}$$

For notational consistency, when $k = 0$, $\operatorname{span}\{\mathbf{m}_{(1)},\ldots,\mathbf{m}_{(k)}\}$ denotes the set $\{\mathbf{0}\} \subset \mathbb{R}^n$. Likewise, when $k = n$, $\operatorname{cone}\{\mathbf{q}_{(k+1)},\ldots,\mathbf{q}_{(n)}\}$ denotes the set $\{\mathbf{0}\} \subset \mathbb{R}^n$.

Since $\mathbf{f}^{(0)}_{\mathbf{x},\mathbf{M}} \equiv \mathbf{f}'(\mathbf{x};\cdot)$, (3.18) yields the $k = 0$ case. For the inductive step, suppose that (3.19) holds with k equal to some $p \in \{0,1,\ldots,(n-1)\}$. Consider any particular $\bar{\mathbf{d}} \in \operatorname{span}\{\mathbf{m}_{(1)},\ldots,\mathbf{m}_{(p+1)}\} + \operatorname{cone}\{\mathbf{q}_{(p+2)},\ldots,\mathbf{q}_{(n)}\}$. By construction, $\bar{\mathbf{d}}$ can be decomposed as $\bar{\mathbf{d}} = \mathbf{u} + \mathbf{v} + \mathbf{w}$, where $\mathbf{u} \in \operatorname{span}\{\mathbf{m}_{(1)},\ldots,\mathbf{m}_{(p)}\}$, $\mathbf{v} := \alpha\mathbf{m}_{(p+1)}$ for some $\alpha \in \mathbb{R}$, and $\mathbf{w} \in \operatorname{cone}\{\mathbf{q}_{(p+2)},\ldots,\mathbf{q}_{(n)}\}$. (If $p = n-1$, then $\mathbf{w} = \mathbf{0}$.)

Now, applying [79, Equation 3.7] and the inductive assumption with $\mathbf{d} = \mathbf{u}$ yields:

$$\mathbf{f}^{(p+1)}_{\mathbf{x},\mathbf{M}}(\mathbf{u}) = \mathbf{f}^{(p)}_{\mathbf{x},\mathbf{M}}(\mathbf{u}) = \mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{x};\mathbf{M})\,\mathbf{u}. \tag{3.20}$$

Applying the definition of the directional derivative, the definition of $\mathbf{q}_{(p+1)}$, and the definition of $\mathbf{v}$ yields:

$$\begin{aligned}
&\mathbf{f}^{(p+1)}_{\mathbf{x},\mathbf{M}}(\mathbf{v}+\mathbf{w}) \\
&\quad = [\mathbf{f}^{(p)}_{\mathbf{x},\mathbf{M}}]'(\mathbf{m}_{(p+1)};\mathbf{v}+\mathbf{w}) \\
&\quad = \lim_{\tau\to 0^+} \frac{\mathbf{f}^{(p)}_{\mathbf{x},\mathbf{M}}((1+\tau\alpha)\mathbf{m}_{(p+1)} + \tau\mathbf{w}) - \mathbf{f}^{(p)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}_{(p+1)})}{\tau} \\
&\quad = \lim_{\tau\to 0^+} \frac{1}{\tau}\left[\mathbf{f}^{(p)}_{\mathbf{x},\mathbf{M}}\left((1+\tau\alpha)\mathbf{q}_{(p+1)} - (1+\tau\alpha)\sum_{j=1}^{p}\ell_{p+1,j}\mathbf{m}_{(j)} + \tau\mathbf{w}\right)\right. \\
&\qquad\qquad \left. - \mathbf{f}^{(p)}_{\mathbf{x},\mathbf{M}}\left(\mathbf{q}_{(p+1)} - \sum_{j=1}^{p}\ell_{p+1,j}\mathbf{m}_{(j)}\right)\right].
\end{aligned}$$

Applying [79, Equation 3.6] and collecting terms yields:

$$\mathbf{f}^{(p+1)}_{\mathbf{x},\mathbf{M}}(\mathbf{v}+\mathbf{w}) = \lim_{\tau\to 0^+} \frac{\mathbf{f}^{(p)}_{\mathbf{x},\mathbf{M}}((1+\tau\alpha)\mathbf{q}_{(p+1)}+\tau\mathbf{w}) - \tau\alpha\mathbf{f}^{(p)}_{\mathbf{x},\mathbf{M}}(\sum_{j=1}^{p}\ell_{p+1,j}\mathbf{m}_{(j)}) - \mathbf{f}^{(p)}_{\mathbf{x},\mathbf{M}}(\mathbf{q}_{(p+1)})}{\tau}. \tag{3.21}$$

Now, $(\sum_{j=1}^{p}\ell_{p+1,j}\mathbf{m}_{(j)}) \in \operatorname{span}\{\mathbf{m}_{(1)},\ldots,\mathbf{m}_{(p)}\}$, and $\mathbf{q}_{(p+1)} \in \operatorname{cone}\{\mathbf{q}_{(p+1)},\ldots,\mathbf{q}_{(n)}\}$. By construction, $\mathbf{w} \in \operatorname{cone}\{\mathbf{q}_{(p+2)},\ldots,\mathbf{q}_{(n)}\} \subset \operatorname{cone}\{\mathbf{q}_{(p+1)},\ldots,\mathbf{q}_{(n)}\}$. It follows that for sufficiently small $\tau > 0$, $(1+\tau\alpha) > 0$, and so $((1+\tau\alpha)\mathbf{q}_{(p+1)}+\tau\mathbf{w}) \in \operatorname{cone}\{\mathbf{q}_{(p+1)},\ldots,\mathbf{q}_{(n)}\}$. Thus, applying the inductive assumption to each term in the numerator of the right-hand side of (3.21) for sufficiently small $\tau > 0$ yields:

$$\begin{aligned}
&\mathbf{f}^{(p+1)}_{\mathbf{x},\mathbf{M}}(\mathbf{v}+\mathbf{w}) \\
&\quad= \lim_{\tau\to 0^+}\frac{1}{\tau}\Big[\mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{x};\mathbf{M})\,((1+\tau\alpha)\mathbf{q}_{(p+1)}+\tau\mathbf{w}) \\
&\qquad\qquad - \tau\alpha\mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{x};\mathbf{M})\left(\sum_{j=1}^{p}\ell_{p+1,j}\mathbf{m}_{(j)}\right) - \mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{x};\mathbf{M})\,\mathbf{q}_{(p+1)}\Big] \\
&\quad= \mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{x};\mathbf{M})\left(\alpha\mathbf{q}_{(p+1)} - \alpha\sum_{j=1}^{p}\ell_{p+1,j}\mathbf{m}_{(j)} + \mathbf{w}\right) \\
&\quad= \mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{x};\mathbf{M})\,(\alpha\mathbf{m}_{(p+1)}+\mathbf{w}) \\
&\quad= \mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{x};\mathbf{M})\,(\mathbf{v}+\mathbf{w}).
\end{aligned} \tag{3.22}$$

To complete the inductive step, since $\mathbf{u} \in \operatorname{span}\{\mathbf{m}_{(1)},\ldots,\mathbf{m}_{(p)}\}$, [79, Equation 3.6] may be applied to yield

$$\mathbf{f}^{(p+1)}_{\mathbf{x},\mathbf{M}}(\bar{\mathbf{d}}) = \mathbf{f}^{(p+1)}_{\mathbf{x},\mathbf{M}}(\mathbf{u}+\mathbf{v}+\mathbf{w}) = \mathbf{f}^{(p+1)}_{\mathbf{x},\mathbf{M}}(\mathbf{u}) + \mathbf{f}^{(p+1)}_{\mathbf{x},\mathbf{M}}(\mathbf{v}+\mathbf{w}).$$

Thus, (3.20) and (3.22) imply that

$$\mathbf{f}^{(p+1)}_{\mathbf{x},\mathbf{M}}(\bar{\mathbf{d}}) = \mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{x};\mathbf{M})\,\mathbf{u} + \mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{x};\mathbf{M})\,(\mathbf{v}+\mathbf{w}) = \mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{x};\mathbf{M})\,(\mathbf{u}+\mathbf{v}+\mathbf{w}) = \mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{x};\mathbf{M})\,\bar{\mathbf{d}}.$$

Since $\bar{\mathbf{d}}$ was chosen aribrarily from $\operatorname{span}\{\mathbf{m}_{(1)},\ldots,\mathbf{m}_{(p+1)}\} + \operatorname{cone}\{\mathbf{q}_{(p+2)},\ldots,\mathbf{q}_{(n)}\}$, (3.19) has been demonstrated for $k = p+1$, completing the inductive step.

Now, setting k to n in (3.19) yields the following, noting that the columns of $\mathbf{M}$ span $\mathbb{R}^n$:

$$\mathbf{f}^{(n)}_{\mathbf{x},\mathbf{M}}(\mathbf{d}) = \mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{x};\mathbf{M})\,\mathbf{d}, \qquad \forall \mathbf{d} \in \mathbb{R}^n.$$

Thus, $\mathbf{J}_{\mathrm{L}}\mathbf{f}(\mathbf{x};\mathbf{M}) = \mathbf{J}\mathbf{f}^{(n)}_{\mathbf{x},\mathbf{M}}(\mathbf{0}) = \mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{x};\mathbf{M})$, as required. □

Since LD-derivatives of a function do not depend on the existence or forms of the function's factored representations, the following corollary of the above theorem is immediate.

Corollary 3.3.5. $\mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{x};\mathbf{M})$ *and* $\partial_{\mathrm{C}}\mathbf{f}(\mathbf{x})$ *are independent of the particular factored representation of* $\mathbf{f}$ *used when carrying out Algorithm 12 in Appendix A.*

As the following corollary shows, the result from Corollary 3.2.5 that $\partial_{\mathrm{L}}\mathbf{f}(\mathbf{x}) \subset \partial_{\mathrm{P}}\mathbf{f}(\mathbf{x})$ for a L-smooth function $\mathbf{f}$ is strengthened when $\mathbf{f}$ is known to be $\mathcal{PC}^1$.

Corollary 3.3.6. *Given an open set* $X \subset \mathbb{R}^n$ *and a* $\mathcal{PC}^1$ *function* $\mathbf{f} : X \to \mathbb{R}^m$,

$$\partial_{\mathrm{L}}\mathbf{f}(\mathbf{x}) = \partial_{\mathrm{B}}[\mathbf{f}'(\mathbf{x};\cdot)](\mathbf{0}) \subset \partial_{\mathrm{B}}\mathbf{f}(\mathbf{x}) \subset \partial\mathbf{f}(\mathbf{x}), \qquad \forall \mathbf{x} \in X.$$

Proof. The function $\mathbf{f}$ may be considered to be an elemental $\mathcal{PC}^1$ function without loss of generality: while active normal sets and branch-locked Jacobians for $\mathbf{f}$ are not necessarily known or easily computed, their existence is sufficient for this corollary. In turn, $\mathbf{f}$ can be endowed with the following trivial factored representation, and may therefore be considered $\mathcal{PC}^1$-factorable:

Set $\mathbf{v}_{(0)} \leftarrow \mathbf{x}$
Set $\mathbf{u}_{(1)} \leftarrow \mathbf{v}_{(0)}$
Set $\mathbf{v}_{(1)} \leftarrow \mathbf{f}(\mathbf{u}_{(1)})$
Set $\mathbf{f}(\mathbf{x}) \leftarrow \mathbf{v}_{(1)}$.

Thus, for any $\mathbf{x} \in X$, Theorem 3.3.3 and Theorem A.3.2 imply that

$$\partial_{\mathrm{B}}[\mathbf{f}'(\mathbf{x};\cdot)](\mathbf{0}) = \partial_{\mathrm{C}}\mathbf{f}(\mathbf{x}) \subset \partial_{\mathrm{B}}\mathbf{f}(\mathbf{x}) \subset \partial\mathbf{f}(\mathbf{x}). \tag{3.23}$$

Applying Theorem 3.3.4 to $\mathbf{f}$ yields the required result, in combination with (3.23).

□

The following example uses the above corollary to show that even for $\mathcal{PC}^1$ functions, the lexicographic subdifferential is not necessarily upper semicontinuous as a set-valued mapping, in the sense of [3, Definition 1.4.1].

Example 3.3.7. *Consider the following $\mathcal{PC}^1$ function $f : \mathbb{R}^2 \to \mathbb{R}$, which was considered previously in [54, Example 2.14]:*

$$f : \mathbf{x} \mapsto \begin{cases} x_2 + x_1^2 & \text{if} \quad x_2 \leq -x_1^2, \\ 0 & \text{if} \quad -x_1^2 < x_2 < x_1^2, \\ x_2 - x_1^2 & \text{if} \quad x_1^2 \leq x_2. \end{cases}$$

The function f is plotted in Figure 3-1.

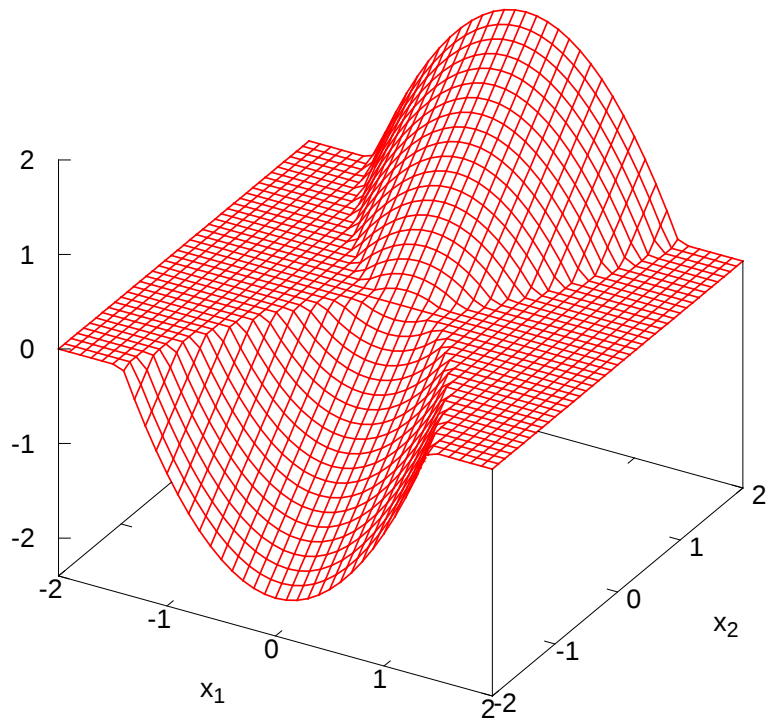

Figure 3-1: The function f described in Example 3.3.7.

For each $\epsilon > 0$, f is differentiable at $(\epsilon, 0)$, with $\mathbf{J}f(\epsilon, 0) = \begin{bmatrix} 0 & 0 \end{bmatrix}$. By inspection, the directional derivative of f at $\mathbf{0}$ is

$$f'(\mathbf{0};\mathbf{d}) = d_2, \qquad \forall \mathbf{d} \in \mathbb{R}^2,$$

and so $\partial_{\mathrm{B}}[f'(\mathbf{0};\cdot)](\mathbf{0}) = \left\{ \begin{bmatrix} 0 & 1 \end{bmatrix} \right\}$. *Thus, applying Corollary 3.3.6,*

$$\lim_{\substack{\mathbf{A} \in \partial_{\mathrm{L}} f(\epsilon,0) \\ \epsilon \to 0^+}} \mathbf{A} = \begin{bmatrix} 0 & 0 \end{bmatrix} \notin \left\{ \begin{bmatrix} 0 & 1 \end{bmatrix} \right\} = \partial_{\mathrm{L}} f(\mathbf{0}).$$

It follows that the set-valued mapping $\partial_{\mathrm{L}} f$ *is not upper semicontinuous at* $\mathbf{0}$.

Chapter 4

Evaluating lexicographic derivatives for factorable functions

4.1 Introduction

This chapter is reproduced from the article [61]. As discussed in Chapter 2, Clarke's generalized Jacobian [16] is a set-valued extension of the classical derivative to functions $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ that are locally Lipschitz continuous, but not necessarily differentiable everywhere. Several numerical methods for nonsmooth functions require Clarke Jacobian elements at each iteration, including semismooth Newton methods for equation-solving, and bundle methods for local optimization.

In Chapter 3, it was demonstrated that Nesterov's lexicographic derivatives [79] are elements of the plenary hull [109] of the Clarke Jacobian whenever they exist. Consequently, lexicographic derivatives are no less useful in nonsmooth numerical methods than Clarke Jacobian elements. Moreover, as also shown in Chapter 3, this result can be strengthened for functions that are piecewise differentiable in the sense of Scholtes [97]. Such a function has well-defined lexicographic derivatives, with these lexicographic derivatives being elements of the B-subdifferentials of both the original function and its directional derivative. In this case, it follows from [97, Proposition 4.3.1] that lexicographic derivatives provide the derivative

information required for local Q-quadratic convergence of Kojima and Shindo's Newton method [65, Algorithm EN], provided that this method's nonsingularity assumptions are satisfied.

Although the Clarke Jacobian satisfies calculus rules as inclusions rather than equations, recently developed methods [33, 54] evaluate Clarke Jacobian elements for finite compositions of simple *elemental functions* that are piecewise differentiable in the sense of Scholtes [97]. The method in [54] is reproduced for reference in Appendix A; this method requires applications of automatic differentiation (AD) at intermediate steps, in such a way that it cannot be implemented in a "tapeless" fashion like the standard forward AD mode for directional derivative evaluation [34]. Instead, this method requires storage of the composite function's computational graph, analogously to the reverse AD mode [34]. A system of linear equations must also be solved at the final step of the algorithm. The approach in [33] applies to a narrower class of functions: comprising compositions of simple smooth functions and the absolute value function. However, this method does not require either the remedial AD applications or the final linear solves of the method in Appendix A.

In this chapter, the methods described in the previous paragraph are unified and combined with techniques for lexicographic differentiation [79], to yield a new method for generalized derivative evaluation in the spirit of the vector forward AD mode. This method applies to a superset of the broad class of composite functions considered in Appendix A, while also retaining the computational benefits of the method in [33]. In particular, the composite functions under consideration need not be piecewise differentiable in the sense of Scholtes [97]. The method essentially proceeds by evaluating an LD-derivative: the variant of the lexicographic derivative introduced in Chapter 3 to simplify the treatment of the lexicographic derivative's calculus rules. Lexicographic derivatives are readily obtained from these LD-derivatives.

This chapter is structured as follows. Section 4.2 presents vector forward AD modes for generalized derivative evaluation and discusses their worst-case com-

putational performance, and Section 4.3 describes and applies a C++ implementation of these methods.

4.2 Generalized derivative evaluation

Corollary 3.3.6 shows that for a $\mathcal{PC}^1$ function, any lexicographic derivative is an element of both the B-subdifferential and the Clarke Jacobian. As shown in Section 3.1, lexicographic derivatives are readily obtained from LD-derivatives. Motivated by these results, this section presents numerical methods that evaluate LD-derivatives for L-factorable functions and $\mathcal{PC}^1$-factorable functions. These methods essentially combine the computational benefits of previous methods for Clarke Jacobian element evaluation [33, 54] and lexicographic derivative evaluation [79]. The computational tractability of these methods is established and discussed.

4.2.1 Method overview

Given an L-factorable function $\mathbf{f}$, applying Proposition 3.1.2 to the factored representation of $\mathbf{f}$ shows that Algorithm 2 evaluates $\mathbf{f}'(\mathbf{x};\mathbf{M})$. Algorithm 2 is, in essence, a generalization of Algorithm 1 in which LD-derivatives replace Jacobians throughout. In the special case in which each elemental function $\boldsymbol{\psi}_{(j)}$ is differentiable at $\mathbf{u}_{(j)}$,

$$[\boldsymbol{\psi}_{(j)}]'(\mathbf{u}_{(j)};\dot{\mathbf{U}}_{(j)}) = \mathbf{J}\boldsymbol{\psi}_{(j)}(\mathbf{u}_{(j)})\,\dot{\mathbf{U}}_{(j)},$$

and so Algorithm 2 reduces to Algorithm 1. If $\mathbf{f}$ is $\mathcal{PC}^1$ and $\mathbf{M} \in \mathbb{R}^{n\times n}$ is square and nonsingular, then Corollary 3.3.6 implies that Algorithm 2 computes $\mathbf{HM}$ for some matrix $\mathbf{H} \in \partial_{\mathrm{B}}[\mathbf{f}'(\mathbf{x};\cdot)](\mathbf{0}) \subset \partial_{\mathrm{B}}\mathbf{f}(\mathbf{x})$. Any $\mathcal{PC}^1$-factorable function is $\mathcal{PC}^1$ by construction; the following section shows that it is L-factorable as well, and discusses computation of the required LD-derivatives of elemental $\mathcal{PC}^1$ functions.

Unlike Algorithm 12 in Appendix A, this algorithm does not require remedial applications of the forward mode of AD. As a result, it can be implemented using

Algorithm 2 Computes $\mathbf{f}(\mathbf{x})$ and $\mathbf{f}'(\mathbf{x};\mathbf{M})$ for an L-factorable function $\mathbf{f}$

Require: $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ is L-factorable, $\mathbf{x} \in X$, $p \in \mathbb{N}$, $\mathbf{M} \in \mathbb{R}^{n\times p}$
  $\mathbf{v}_{(0)} \leftarrow \mathbf{x}$
  $\dot{\mathbf{V}}_{(0)} \leftarrow \mathbf{M}$
  for $j = 1$ **to** ℓ **do**
    $\mathbf{u}_{(j)} \leftarrow [\mathbf{v}_{(i)}]_{i\prec j}$
    $\mathbf{v}_{(j)} \leftarrow \boldsymbol{\psi}_{(j)}(\mathbf{u}_{(j)})$
    $\dot{\mathbf{U}}_{(j)} \leftarrow [\dot{\mathbf{V}}_{(i)}]_{i\prec j}$
    $\dot{\mathbf{V}}_{(j)} \leftarrow [\boldsymbol{\psi}_{(j)}]'(\mathbf{u}_{(j)};\dot{\mathbf{U}}_{(j)})$
  end for
  return $\mathbf{f}(\mathbf{x}) = \mathbf{v}_{(\ell)}$ and $\mathbf{f}'(\mathbf{x};\mathbf{M}) = \dot{\mathbf{V}}_{(\ell)}$

operator overloading in a similar tapeless fashion to the vector forward AD mode for smooth functions [34, Chapter 6]. Computational cost is correspondingly reduced, due to the relative computational expense of carrying out the forward AD mode.

A key difference between Algorithm 2 and the standard vector forward AD mode is that in Algorithm 2, the kth column $\dot{\mathbf{v}}_{(j),k}$ of $\dot{\mathbf{V}}_{(j)}$ does not represent the directional derivative $[\mathbf{v}_{(j)}]'(\mathbf{x};\mathbf{m}_{(k)})$. Instead, by inspection of the LD-derivative's definition, $\dot{\mathbf{v}}_{(j),k}$ denotes $[\mathbf{v}_{(j)}]^{(k-1)}_{\mathbf{x},\mathbf{M}}(\mathbf{m}_{(k)})$.

4.2.2 Evaluating LD-derivatives for elemental functions

To evaluate LD-derivatives for an L-factorable function $\mathbf{f}$ using Algorithm 2, LD-derivatives for the elemental functions $\boldsymbol{\psi}_{(j)}$ must be readily computable. Thus, this section presents closed-form expressions for LD-derivatives of abs, $\max\{\cdot,\cdot\}$, $\min\{\cdot,\cdot\}$, and $\|\cdot\|_2$, along with a general, computationally tractable method for computing LD-derivatives for elemental $\mathcal{PC}^1$ functions, when these are not known *a priori*.

Given a lexicographically smooth elemental function $\boldsymbol{\psi}$, some $\mathbf{x} \in X_{\boldsymbol{\psi}}$, and some $\mathbf{M} \in \mathbb{R}^{n\times p}$, the standard forward AD mode for directional derivative evaluation may be used to evaluate $\boldsymbol{\psi}^{(0)}_{\mathbf{x},\mathbf{M}}$, and then to successively evaluate $\boldsymbol{\psi}^{(k+1)}_{\mathbf{x},\mathbf{M}}$ from $\boldsymbol{\psi}^{(k)}_{\mathbf{x},\mathbf{M}}$ for each k. The computational cost of evaluating $\boldsymbol{\psi}^{(p-1)}_{\mathbf{x},\mathbf{M}}$ in this manner, however, scales worst-case exponentially with p. As a result, this is not an ideal

way to compute the LD-derivative $\psi(\mathbf{x};\mathbf{M})$ in general. When ψ is an elemental $\mathcal{PC}^1$ function, however, the following lemma shows that Algorithm 3 evaluates LD-derivatives for ψ, by exploiting (3.4) and Theorem 3.3.4. As discussed in the next section, this method is computationally tractable.

Algorithm 3 Computes $\psi'(\mathbf{x};\mathbf{M})$ for an elemental $\mathcal{PC}^1$ function ψ

Require: $\psi : X_\psi \subset \mathbb{R}^{n_\psi} \to \mathbb{R}^{m_\psi}$ is an elemental $\mathcal{PC}^1$ function, $\mathbf{x} \in X_\psi$, $p \in \mathbb{N}$, $\mathbf{M} \in \mathbb{R}^{n\times p}$
  if $\zeta_\psi(\mathbf{x}) = \texttt{true}$ **then**
    $\dot{\mathbf{V}} \leftarrow \mathbf{J}\psi(\mathbf{x})\,\mathbf{M}$
  else
    $\begin{bmatrix}\dot{\mathbf{w}}_{(1)} & \cdots & \dot{\mathbf{w}}_{(p)}\end{bmatrix} \leftarrow \mathbf{M}$
    for $r = 1$ **to** $|H_\psi(\mathbf{x})|$ **do**
      $s_r \leftarrow 1$
      $c^* \leftarrow 0$
      for $k = 1$ **to** p **do**
        Set $c \leftarrow \langle \mathbf{a}_\psi^{(r)}(\mathbf{x}), \dot{\mathbf{w}}_{(k)}\rangle \in \mathbb{R}$
        if $c \neq 0$ **then**
          if $c^* = 0$ **then**
            $c^* \leftarrow c$
            $s_r \leftarrow \operatorname{sign} c$
            $k^* \leftarrow k$
          else if $cc^* < 0$ **then**
            $\dot{\mathbf{w}}_{(k)} \leftarrow \dot{\mathbf{w}}_{(k)} - (\frac{c}{c^*})\dot{\mathbf{w}}_{(k^*)}$
          end if
        end if
      end for
    end for
    $\dot{\mathbf{V}} \leftarrow \mathbf{\Gamma}_\psi(\mathbf{x}; s_1, \ldots, s_{|H_\psi(\mathbf{x})|})\mathbf{M}$
  end if
  return $\psi'(\mathbf{x};\mathbf{M}) = \dot{\mathbf{V}}$

Lemma 4.2.1. *Given an open set $X \subset \mathbb{R}^n$, a vector $\mathbf{x} \in X$, an elemental $\mathcal{PC}^1$ function $\psi : X \to \mathbb{R}^m$, and a matrix $\mathbf{M} = \begin{bmatrix}\mathbf{m}_{(1)} & \cdots & \mathbf{m}_{(p)}\end{bmatrix} \in \mathbb{R}^{n\times p}$, Algorithm 3 computes $\psi'(\mathbf{x};\mathbf{M})$.*

Proof. Define the mapping $\mathcal{T}_{\mathbf{M},\mathbf{x}} : \mathbb{R}^p \to \mathbb{R}^n : \mathbf{y} \mapsto \mathbf{x} + \mathbf{M}\mathbf{y}$. If $\zeta_\psi(\mathbf{x}) = \texttt{true}$, then ψ is differentiable at $\mathbf{x}$. In this case, $\psi'(\mathbf{x};\mathbf{M}) = \mathbf{J}\psi(\mathbf{x})\,\mathbf{M}$, which is returned by Algorithm 3. Otherwise, if $\zeta_\psi(\mathbf{x}) = \texttt{false}$, then Theorem 3.3.4 implies that $\mathbf{J}_\mathrm{C}[\psi \circ \mathcal{T}_{\mathbf{M},\mathbf{x}}](\mathbf{0};\mathbf{I})$ is equal to $\mathbf{J}_\mathrm{L}[\psi \circ \mathcal{T}_{\mathbf{M},\mathbf{x}}](\mathbf{0};\mathbf{I})$, which is in turn equal to $\psi'(\mathbf{x};\mathbf{M})$ according to (3.4). In this case, it suffices to show that Algorithm 3 is equivalent

to the procedure for evaluating $\mathbf{J}_{\mathrm{C}}[\psi \circ \mathcal{T}_{\mathbf{M},\mathbf{x}}](\mathbf{0};\mathbf{I})$ according to [54, Algorithm 4.1], which is reproduced as Algorithm 12 in Appendix A.

To show that Algorithm 3 and the procedure for evaluating $\mathbf{J}_{\mathrm{C}}[\psi \circ \mathcal{T}_{\mathbf{M},\mathbf{x}}](\mathbf{0};\mathbf{I})$ are equivalent, consider the following factored representation of $\mathbf{f} \equiv \psi \circ \mathcal{T}_{\mathbf{M},\mathbf{x}}$:

$$
\begin{aligned}
&\mathbf{v}_{(0)} \leftarrow \mathbf{y} \\
&\mathbf{u}_{(1)} \leftarrow \mathbf{v}_{(0)} \\
&\mathbf{v}_{(1)} \leftarrow \mathcal{T}_{\mathbf{M},\mathbf{x}}(\mathbf{u}_{(1)}) \\
&\mathbf{u}_{(2)} \leftarrow \mathbf{v}_{(1)} \\
&\mathbf{v}_{(2)} \leftarrow \psi(\mathbf{u}_{(2)}) \\
&\mathbf{f}(\mathbf{y}) \leftarrow \mathbf{v}_{(2)},
\end{aligned}
$$

and suppose that Algorithm 12 in Appendix A is used to compute $\mathbf{J}_{\mathrm{C}}\mathbf{f}(\mathbf{0};\mathbf{I})$, noting that $\mathbf{f}$ is defined on some neighborhood of $\mathbf{0}$. Immediately before the outermost for-loop in the algorithm is reached, the vectors $\dot{\mathbf{u}}_{(2,1)}, \ldots, \dot{\mathbf{u}}_{(2,p)}$ will have been computed using the forward AD mode as

$$
\dot{\mathbf{u}}_{(2,k)} = \dot{\mathbf{v}}_{(1)}(\mathbf{0};\mathbf{e}_{(k)}) = [\mathcal{T}_{\mathbf{M},\mathbf{x}}]'(\mathbf{0};\mathbf{e}_{(k)}) = \mathbf{J}\mathcal{T}_{\mathbf{M},\mathbf{x}}(\mathbf{0})\,\mathbf{e}_{(k)} = \mathbf{M}\mathbf{e}_{(k)} = \mathbf{m}_{(k)}, \qquad \forall k \in \{1,\ldots,p\}.
$$

Hence, for each k, at this point in Algorithm 12, $\dot{\mathbf{u}}_{(2,k)}$ is equal to the initial value of $\dot{\mathbf{w}}_{(k)}$ in Algorithm 3.

Since $\mathcal{T}_{\mathbf{M},\mathbf{x}}$ is differentiable, no action is taken during the first iteration of the outermost for-loop of Algorithm 12, in which $j = 1$. Immediately after the second iteration, in which $j = 2$, the proof of Lemma A.6.6 and inspection of Algorithm 12 show that for each k, $\dot{\mathbf{u}}_{(2,k)}$ is now equal to the final value of $\dot{\mathbf{w}}_{(k)}$ in Algorithm 3. Moreover, since the same sequence of elementary column operations was applied to $\begin{bmatrix}\dot{\mathbf{u}}_{(2,1)} & \cdots & \dot{\mathbf{u}}_{(2,p)}\end{bmatrix}$ as to $\begin{bmatrix}\mathbf{q}_{(1)} & \cdots & \mathbf{q}_{(p)}\end{bmatrix}$, there exists some $\mathbf{A} \in \mathbb{R}^{p\times p}$ such that

$$
\begin{bmatrix}\dot{\mathbf{u}}_{(2,1)} & \cdots & \dot{\mathbf{u}}_{(2,p)}\end{bmatrix} = \mathbf{M}\mathbf{A}, \qquad \text{and} \qquad \begin{bmatrix}\mathbf{q}_{(1)} & \cdots & \mathbf{q}_{(p)}\end{bmatrix} = \mathbf{I}\mathbf{A}.
$$

Thus, $\mathbf{A} = \begin{bmatrix}\mathbf{q}_{(1)} & \cdots & \mathbf{q}_{(p)}\end{bmatrix}$, and

$$
\begin{bmatrix}\dot{\mathbf{u}}_{(2,1)} & \cdots & \dot{\mathbf{u}}_{(2,p)}\end{bmatrix} = \mathbf{M}\begin{bmatrix}\mathbf{q}_{(1)} & \cdots & \mathbf{q}_{(p)}\end{bmatrix}. \tag{4.1}
$$

Now, if c^* is assigned a nonzero value during the rth iteration of the middle for-loop in Algorithm 12, then the corresponding value of $\operatorname{sign} c^*$ is, by inspection, equal to the assigned value of s_r in the rth iteration of the outer for-loop in Algorithm 3. Moreover, as in the proof of Lemma A.6.6, once the second iteration of the outermost for-loop in Algorithm 12 is completed,

$$s_r \left\langle \mathbf{a}_{\psi}^{(r)}(\mathbf{x}), \dot{\mathbf{u}}_{(2,k)} \right\rangle \geq 0, \qquad \forall k \in \{1, \ldots, p\}.$$

The above relationship holds even if c^* remains at zero throughout the rth iteration of the middle for-loop in Algorithm 12, since for this to occur, it must be true that $\langle \mathbf{a}_{\psi}^{(r)}(\mathbf{x}), \dot{\mathbf{u}}_{(2,k)} \rangle = 0$ for each k.

Thus, as in the proof of Lemma A.6.6, there now exists a polyhedral cone

$$C^* := \bigcap_{r=1}^{|H_{\psi}(\mathbf{x})|} \left\{ \mathbf{z} \in \mathbb{R}^n : s_r \langle \mathbf{a}_{\psi}^{(r)}(\mathbf{x}), \mathbf{z} \rangle \geq 0 \right\},$$

which contains each $\dot{\mathbf{u}}_{(2,k)} = \dot{\mathbf{w}}_{(k)}$. The definition of $\mathbf{\Gamma}_{\psi}(\mathbf{x}; s_1, \ldots, s_{|H_{\psi}(\mathbf{x})|})$ implies that

$$\mathbf{\Gamma}_{\psi}(\mathbf{x}; s_1, \ldots, s_{|H_{\psi}(\mathbf{x})|}) \dot{\mathbf{u}}_{(2,k)} = \psi'(\mathbf{x}; \dot{\mathbf{u}}_{(2,k)}), \qquad \forall k \in \{1, \ldots, p\}. \tag{4.2}$$

Applying the chain rule for directional derivatives, for each $k \in \{1, \ldots, p\}$,

$$[\psi \circ \mathcal{T}_{\mathbf{M},\mathbf{x}}]'(\mathbf{0}; \mathbf{q}_{(k)}) = \psi'(\mathcal{T}_{\mathbf{M},\mathbf{x}}(\mathbf{0}); \mathbf{J}\mathcal{T}_{\mathbf{M},\mathbf{x}}(\mathbf{0})\, \mathbf{q}_{(k)}) = \psi'(\mathbf{x}; \mathbf{M}\mathbf{q}_{(k)}).$$

Applying (4.1) and (4.2), for each $k \in \{1, \ldots, p\}$,

$$\begin{aligned}
&[\psi \circ \mathcal{T}_{\mathbf{M},\mathbf{x}}]'(\mathbf{0}; \mathbf{q}_{(k)}) \\
&\quad = \psi'(\mathbf{x}; \dot{\mathbf{u}}_{(2,k)}) \\
&\quad = \mathbf{\Gamma}_{\psi}(\mathbf{x}; s_1, \ldots, s_{|H_{\psi}(\mathbf{x})|})\, \dot{\mathbf{u}}_{(2,k)} \\
&\quad = \mathbf{\Gamma}_{\psi}(\mathbf{x}; s_1, \ldots, s_{|H_{\psi}(\mathbf{x})|})\, \mathbf{M}\mathbf{q}_{(k)}.
\end{aligned}$$

Comparing these equations to the linear system solved at the end of Algorithm 12, and noting that $\begin{bmatrix} \mathbf{q}_{(1)} & \cdots & \mathbf{q}_{(p)} \end{bmatrix}$ is nonsingular by Lemma A.6.2, it follows that

$$\mathbf{J}_{\mathrm{C}}[\psi \circ \mathcal{T}_{\mathbf{M},\mathbf{x}}](\mathbf{0};\mathbf{I}) = \mathbf{\Gamma}_{\psi}(\mathbf{x}; s_1, \ldots, s_{|H_{\psi}(\mathbf{x})|})\, \mathbf{M},$$

which is the output of Algorithm 3. As discussed earlier, this result completes the proof. □

The following examples apply Algorithm 3 to evaluate LD-derivatives for simple nonsmooth functions which are commonly encountered in practice. These examples show that iterations of the for-loop in Algorithm 2 are readily executed when $\psi_{(j)}$ is a simple $\mathcal{PC}^1$ function.

Example 4.2.2. *When $\psi \equiv \mathrm{abs}\,(\cdot)$, Algorithm 3 reduces to the following. Since* $\mathrm{abs}\,(\cdot)$ *admits a scalar argument, $\dot{\mathbf{V}}$ and $\mathbf{M}$ are both row vectors in this case.*

```
if x ≠ 0 then
   V̇ ← (sign x)M
else
   s_1 ← 1
   for k = 1 to p do
      if m_(k) ≠ 0 then
         s_1 ← sign m_(k)
         Break out of for-loop
      end if
   end for
   V̇ ← s_1 M
end if
```

The above procedure may be further rearranged and simplified to yield the following variant of an expression in [79, Section 4]:

$$\begin{aligned}
\dot{\mathbf{V}} &\leftarrow \psi'(x;\mathbf{M}),\\
&= \begin{cases} (\operatorname{sign} x)\mathbf{M} & \textit{if } x \neq 0,\\ \mathbf{0} & \textit{if } \mathbf{M} = \mathbf{0},\\ (\operatorname{sign} m_{(k^*)})\mathbf{M} \quad \textit{with } k^* := \min\{k : m_{(k)} \neq 0\}, & \textit{if } x = 0 \textit{ and } \mathbf{M} \neq \mathbf{0}. \end{cases}
\end{aligned}$$

These procedures may be rephrased without reference to matrix operations, as follows:

```
s_1 ← sign x
for k = 1 to p do
   if s_1 = 0 then
      s_1 ← sign m_(k)
```

```
    end if
    v̇_(k) ← s_1 m_(k)
  end for
```

If $\mathbf{f}$ is abs -factorable, and if the abs $(\cdot)$ function is handled according to Example 4.2.2, then Algorithm 2 effectively reduces to the approach of [33, Section 6.2].

Example 4.2.3. *When $\psi \equiv \max\{\cdot,\cdot\}$, Algorithm 3 reduces to the following:*

```
  if x_1 > x_2 then
    V̇ ← [m_(1),1  ···  m_(p),1]
  else if x_1 < x_2 then
    V̇ ← [m_(1),2  ···  m_(p),2]
  else
    s_1 ← 1
    for k = 1 to p do
      if m_(k),1 ≠ m_(k),2 then
        s_1 ← sign (m_(k),1 − m_(k),2)
        Break out of for-loop
      end if
    end for
    if s_1 ≥ 0 then
      V̇ ← [m_(1),1  ···  m_(p),1]
    else
      V̇ ← [m_(1),2  ···  m_(p),2]
    end if
  end if
```

As in the $\psi \equiv \text{abs}\,(\cdot)$ case, this procedure can be further simplified to yield the following variant of an expression in [79, Section 4]:

$$\dot{\mathbf{V}} \leftarrow \psi'(\mathbf{x};\mathbf{M}) = \begin{cases} \begin{bmatrix}1 & 0\end{bmatrix}\mathbf{M} & \text{if } x_1 > x_2, \\ \begin{bmatrix}0 & 1\end{bmatrix}\mathbf{M} & \text{if } x_1 < x_2, \\ \begin{bmatrix}1 & 0\end{bmatrix}\mathbf{M} & \text{if } x_1 = x_2 \text{ and } \begin{bmatrix}1 & -1\end{bmatrix}\mathbf{M} = \mathbf{0}, \\ \begin{bmatrix}1 & 0\end{bmatrix}\mathbf{M} & \text{if } x_1 = x_2, \begin{bmatrix}1 & -1\end{bmatrix}\mathbf{M} \neq \mathbf{0}, \text{ and } m_{(k^*),1} > m_{(k^*),2}, \\ & \quad \text{with } k^* := \min\{k : m_{(k),1} \neq m_{(k),2}\}, \\ \begin{bmatrix}0 & 1\end{bmatrix}\mathbf{M} & \text{if } x_1 = x_2, \begin{bmatrix}1 & -1\end{bmatrix}\mathbf{M} \neq \mathbf{0}, \text{ and } m_{(k^*),1} < m_{(k^*),2}, \\ & \quad \text{with } k^* := \min\{k : m_{(k),1} \neq m_{(k),2}\}. \end{cases}$$

The procedure is identical when $\psi \equiv \min\{\cdot,\cdot\}$, except with the comparison operators $>$ and $<$ reversed in each instance.

The above procedures may be rephrased without reference to matrix operations, as follows:

$s_1 \leftarrow \operatorname{sign}(x_1 - x_2)$
for $k = 1$ ***to*** p ***do***
  if $s_1 = 0$ ***then***
    $s_1 \leftarrow \operatorname{sign}(m_{(k),1} - m_{(k),2})$
  end if
  if $s_1 \geq 0$ ***then***
    $\dot{v}_{(k)} \leftarrow m_{(k),1}$
  else
    $\dot{v}_{(k)} \leftarrow m_{(k),2}$
  end if
end for

Again, replacing the "$\geq$" operator with "$\leq$" yields an analogous procedure for $\min\{\cdot,\cdot\}$.

Any elemental $\mathcal{PC}^1$ function may be used in Algorithm 3, provided that the required information is available. If, however, at least one elemental function $\psi_{(j)}$ is lexicographically smooth but not $\mathcal{PC}^1$, then Corollary 3.3.6 may not apply. Nevertheless, a lexicographic derivative obtained from Algorithm 2 will still be a plenary Jacobian element due to Corollary 3.2.5. To this end, the following example presents LD-derivatives for the function $\|\cdot\|_2$, which is lexicographically smooth but not $\mathcal{PC}^1$.

Example 4.2.4. *LD-derivatives for $\psi \equiv \|\cdot\|_2$ on $\mathbb{R}^n$ are obtained as follows. With $\hat{\mathbf{q}}$ denoting the* unit vector $\frac{\mathbf{q}}{\|\mathbf{q}\|_2}$ *for any nonzero $\mathbf{q} \in \mathbb{R}^n$, it is readily verified that ψ is differentiable at each $\mathbf{x} \neq \mathbf{0}$, with*

$$\mathbf{J}\psi(\mathbf{x}) = \hat{\mathbf{x}}^{\mathrm{T}}, \qquad \forall \mathbf{x} \neq \mathbf{0}.$$

Since ψ is convex, it is lexicographically smooth; the directional derivative of ψ at $\mathbf{0}$ is

$$\psi'(\mathbf{0};\mathbf{d}) = \|\mathbf{d}\|_2 = \psi(\mathbf{d}) = \hat{\mathbf{d}}^{\mathrm{T}}\mathbf{d}, \qquad \forall \mathbf{d} \in \mathbb{R}^n.$$

Using the above results, it is readily shown that for any $\mathbf{M} := \begin{bmatrix} \mathbf{m}_{(1)} & \cdots & \mathbf{m}_{(p)} \end{bmatrix} \in \mathbb{R}^{n\times p}$ and any $k \in \{0, 1, \ldots, p\}$,

$$\psi^{(k)}_{\mathbf{0},\mathbf{M}}(\mathbf{d}) = \begin{cases} \psi(\mathbf{d}), & \textit{if } \mathbf{m}_{(j)} = \mathbf{0} \quad \forall j \in \{1,\ldots,k\} \\ (\hat{\mathbf{m}}_{(\ell)})^{\mathrm{T}}\mathbf{d}, \quad \textit{with } \ell := \min\{j : \mathbf{m}_{(j)} \neq \mathbf{0}\}, & \textit{otherwise.} \end{cases}$$

It follows that the LD-derivative of ψ is given by

$$\psi'(\mathbf{x};\mathbf{M}) = \begin{cases} \hat{\mathbf{x}}^{\mathrm{T}}\mathbf{M}, & \textit{if } \mathbf{x} \neq \mathbf{0}, \\ \mathbf{0}, & \textit{if } \mathbf{M} = \mathbf{0}, \\ (\hat{\mathbf{m}}_{(j^*)})^{\mathrm{T}}\mathbf{M} \quad \textit{with } j^* := \min\{j : \mathbf{m}_{(j)} \neq \mathbf{0}\}, & \textit{if } \mathbf{x} = \mathbf{0} \textit{ and } \mathbf{M} \neq \mathbf{0}. \end{cases}$$

4.2.3 Estimating computational complexity

The worst-case computational complexity of Algorithm 2 may be estimated as follows. Let the computational cost of evaluating a function $\mathbf{g}$ at some domain point $\mathbf{x}$ be $\mathcal{C}[\mathbf{g}(\mathbf{x})]$. Consider an L-factorable function $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$, whose factored representation consists of elemental functions from an elemental library $\mathcal{L}$. Comparing Algorithm 2 with the definition of a factored representation, for any particular $\mathbf{M} \in \mathbb{R}^{n\times p}$, it follows that

$$\begin{aligned} \mathcal{C}[\mathbf{f}'(\mathbf{x};\mathbf{M})] &\leq \left(\max_{j\in\{1,\ldots,\ell\}} \frac{\mathcal{C}\left[[\boldsymbol{\psi}_{(j)}]'(\mathbf{u}_{(j)};\dot{\mathbf{U}}_{(j)})\right]}{\mathcal{C}[\boldsymbol{\psi}_{(j)}(\mathbf{u}_{(j)})]} \right) \mathcal{C}[\mathbf{f}(\mathbf{x})] \\ &\leq \max\left\{ \frac{\mathcal{C}\left[\boldsymbol{\psi}'(\mathbf{z};\mathbf{N})\right]}{\mathcal{C}[\boldsymbol{\psi}(\mathbf{z})]} : \boldsymbol{\psi} \in \mathcal{L}, \mathbf{z} \in X_{\boldsymbol{\psi}}, \mathbf{N} \in \mathbb{R}^{n_{\boldsymbol{\psi}}\times p} \right\} \mathcal{C}[\mathbf{f}(\mathbf{x})]. \end{aligned} \tag{4.3}$$

The coefficient of $\mathcal{C}[\mathbf{f}(\mathbf{x})]$ in the above estimate is a library-dependent constant, and may itself be estimated as follows. Assume that each function in $\mathcal{L}$ is either a standard smooth function or operation such as $+$, $\times$, cos, or exp, a simple nonsmooth function as considered in Examples 4.2.2 to 4.2.4, or an elemental $\mathcal{PC}^1$ function. These cases will be considered separately.

If a function $\boldsymbol{\psi} \in \mathcal{L}$ is a standard smooth function or operation, then for any $\mathbf{z} \in X_{\boldsymbol{\psi}}$ and $\mathbf{N} \in \mathbb{R}^{n_{\boldsymbol{\psi}}\times p}$, it follows from the discussion in [34, Section 4.5] and the identity $\boldsymbol{\psi}'(\mathbf{z};\mathbf{N}) = \mathbf{J}\boldsymbol{\psi}(\mathbf{z})\,\mathbf{N}$ that

$$\mathcal{C}\left[\boldsymbol{\psi}'(\mathbf{z};\mathbf{N})\right] = \mathcal{C}\left[\mathbf{J}\boldsymbol{\psi}(\mathbf{z})\,\mathbf{N}\right] \leq (1+1.5p)\,\mathcal{C}[\boldsymbol{\psi}(\mathbf{z})].$$

If $\psi \equiv \text{abs}\,(\cdot)$, then the procedure at the end of Example 4.2.2 can be used to evaluate $\psi'(z;\mathbf{N})$. Assume for simplicity that if-statements, the sign $(\cdot)$ function, and products of the form sy for $s \in \{-1,0,1\}$ and $y \in \mathbb{R}$ are each of complexity $\mathcal{C}[\text{abs}\,(z)]$, since they each essentially require evaluation of one or two branching conditions, followed perhaps by a trivial negative operation. In this case, inspection of the final procedure in Example 4.2.2 shows that

$$\mathcal{C}\left[\psi'(z;\mathbf{N})\right] \leq (3p+1)\,\mathcal{C}[\text{abs}\,(z)] = (3p+1)\,\mathcal{C}[\psi(z)].$$

Assuming further that $\mathcal{C}[\max(y,z)] \approx \mathcal{C}[\text{abs}\,(z)]$, inspection of Example 4.2.3 shows that with $\psi \equiv \max(\cdot,\cdot)$ or $\min(\cdot,\cdot)$,

$$\mathcal{C}\left[\psi'((y,z);\mathbf{N})\right] \leq (3p+1)\,\mathcal{C}[\psi(y,z)].$$

With $\psi \equiv \|\cdot\|_2$ on $\mathbb{R}^{n_\psi}$, assume for simplicity that for any $\mathbf{y},\mathbf{z} \in \mathbb{R}^{n_\psi}$, $\mathcal{C}[\psi(\mathbf{z})] \geq \mathcal{C}[\mathbf{y}^{\mathrm{T}}\mathbf{z}]$, since $\psi(\mathbf{z}) = \sqrt{\mathbf{z}^{\mathrm{T}}\mathbf{z}}$. Since norms and inner products evidently dominate the computational complexity of the procedure in Example 4.2.4, it follows that for some $k' \in \{1,\ldots,p\}$,

$$\mathcal{C}\left[\psi'(\mathbf{z};\mathbf{N})\right] \lesssim \mathcal{C}\left[(\hat{\mathbf{n}}_{(k')})^{\mathrm{T}}\mathbf{N}\right] \leq \mathcal{C}[\hat{\mathbf{n}}_{(k')}] + \sum_{k=1}^{p}\mathcal{C}\left[\mathbf{y}^{\mathrm{T}}\mathbf{n}_{(k)}\right] \leq (p+1)\,\mathcal{C}[\psi(\mathbf{z})].$$

Lastly, assume that $\boldsymbol{\psi}$ is an elemental $\mathcal{PC}^1$ function, and that its LD-derivatives are evaluated using Algorithm 3. Elemental $\mathcal{PC}^1$ functions can be arbitrarily complicated; to restrict ourselves to elemental functions that are useful in practice, assume that

$$r_{\max} := \max\{|H_{\boldsymbol{\psi}}(\mathbf{y})| : \boldsymbol{\psi} \in \mathcal{L},\ \mathbf{y} \in X_{\boldsymbol{\psi}}\}$$

is well-defined and finite. Note that with $\psi \equiv \text{abs}\,(\cdot)$, $\max(\cdot,\cdot)$, or $\min(\cdot,\cdot)$, for example,

$$\max\{|H_{\boldsymbol{\psi}}(\mathbf{y})| : \mathbf{y} \in X_{\boldsymbol{\psi}}\} = 1.$$

In the spirit of complexity results for the standard vector forward AD mode [34, Section 4.5], assume that for some library-dependent constant $\gamma \in \mathbb{R}$,

$$\mathcal{C}\left[\mathbf{J}\boldsymbol{\psi}(\mathbf{z})\,\mathbf{N}\right] \leq \gamma p\,\mathcal{C}[\boldsymbol{\psi}(\mathbf{z})] \qquad \text{and} \qquad \mathcal{C}\left[\boldsymbol{\Gamma}_{\boldsymbol{\psi}}(\mathbf{z}; s_1, \ldots, s_{|H_{\boldsymbol{\psi}}(\mathbf{z})|})\,\mathbf{N}\right] \leq \gamma p\,\mathcal{C}[\boldsymbol{\psi}(\mathbf{z})],$$

whenever the left-hand quantities are defined. Inspection of Algorithm 3 shows that, retaining only the dominant terms,

$$\mathcal{C}\left[\boldsymbol{\psi}'(\mathbf{z}; \mathbf{N})\right] \lesssim r_{\max} p\left(\mathcal{C}[\langle \mathbf{a}, \dot{\mathbf{w}}\rangle] + \mathcal{C}[\dot{\mathbf{w}} - \alpha\dot{\mathbf{w}}^*]\right) + \gamma p\,\mathcal{C}[\boldsymbol{\psi}(\mathbf{z})]$$

Assuming further that $\mathcal{C}[\boldsymbol{\psi}(\mathbf{z})]$ exceeds both $\mathcal{C}[\langle \mathbf{a}, \dot{\mathbf{w}}\rangle]$ and $\mathcal{C}[\dot{\mathbf{w}} - \alpha\dot{\mathbf{w}}^*]$, as is likely the case for all but the simplest elemental $\mathcal{PC}^1$ functions $\boldsymbol{\psi}$, the above estimate simplifies to:

$$\mathcal{C}\left[\boldsymbol{\psi}'(\mathbf{z}; \mathbf{N})\right] \lesssim (2r_{\max} + \gamma) p\,\mathcal{C}[\boldsymbol{\psi}(\mathbf{z})].$$

It follows from the above discussion and from (4.3) that the computational complexity of using Algorithm 2 to evaluate $\mathbf{f}'(\mathbf{x}; \mathbf{M})$ satisfies the estimate:

$$\mathcal{C}\left[\mathbf{f}'(\mathbf{x}; \mathbf{M})\right] \lesssim \mathcal{O}(p)\ \mathcal{C}\left[\mathbf{f}(\mathbf{x})\right].$$

In particular, if the elemental library $\mathcal{L}$ contains only $\operatorname{abs}(\cdot)$, $\min(\cdot,\cdot)$, $\max(\cdot,\cdot)$, $\|\cdot\|_2$, and the standard smooth functions and operations, then

$$\mathcal{C}\left[\mathbf{f}'(\mathbf{x}; \mathbf{M})\right] \lesssim (3p + 1)\,\mathcal{C}\left[\mathbf{f}(\mathbf{x})\right].$$

Hence, Algorithm 2 is computationally tractable relative to the cost of a function evaluation. Like the standard vector forward AD mode [34], the cost of Algorithm 2 scales worst-case linearly with p.

4.3 Implementation and examples

This section describes an implementation of Algorithm 2 in C++, and presents results of applying this implementation to various example problems for illustration. An example of a nonsmooth equation system is presented in which if the generalized derivatives required by the semismooth Newton method are incorrectly computed using the standard vector forward AD mode for smooth functions, then the local convergence of the semismooth Newton method is lost.

4.3.1 Implementation

Algorithm 2 was implemented in C++, following a similar approach to the forward AD mode implementation described in [34, Section 6.1]. Analogously to the `adouble` class used in [34] to represent the function value/directional derivative pairs $(v_{(j)}, \dot{v}_{(j)})$, this implementation introduces an `ldouble` class to represent the function value/LD-derivative pairs $(v_{(j)}, \dot{\mathbf{V}}_{(j)})$ appearing in Algorithm 2, with the row vector $\dot{\mathbf{V}}_{(j)}$ represented as a `std::vector<double>`. For simplicity, all elemental functions considered are scalar-valued: these currently include the negative, the standard arithmetic operations $+$, $-$, $\times$, and $\div$, and the nonsmooth elemental functions $|\cdot|$, $\max\{\cdot,\cdot\}$, $\min\{\cdot,\cdot\}$, and $\|(\cdot,\cdot)\|_2 \equiv \sqrt{(\cdot)^2 + (\cdot)^2}$. Further elemental functions such as sin/cos/tan and exp/log can be included readily as well.

The implementation evaluates LD-derivatives as follows. The L-factorable function under consideration is entered as a template subroutine, expressed in terms of the elemental functions mentioned above, and written as though its inputs and outputs are `double` arrays. The various elemental functions are overloaded so that appropriate `ldouble` outputs are produced when `ldouble` inputs are given. To evaluate the LD-derivatives required by this overloading, the relationship $\mathbf{f}'(\mathbf{x}; \mathbf{M}) = \mathbf{J}\mathbf{f}(\mathbf{x})\,\mathbf{M}$ is applied for differentiable elemental functions, and the procedures in Examples 4.2.2, 4.2.3, and 4.2.4 are applied for the nonsmooth elemental functions considered.

To initialize the inputs $\mathbf{x}$ and $\mathbf{M}$ to Algorithm 2, an `ldouble` array is constructed,

whose k^{th} entry is $(x_k, \begin{bmatrix} m_{(1),k} & \cdots & m_{(p),k} \end{bmatrix})$. Thus, input components are effectively handled one at a time in the spirit of Example 2.5.2; introducing the coordinate projection function $\pi_k : \mathbf{x} \mapsto x_k$, it is readily verified that

$$[\pi_k]'(\mathbf{x}; \mathbf{M}) = \begin{bmatrix} m_{(1),k} & \cdots & m_{(p),k} \end{bmatrix}.$$

The outputs of Algorithm 2 are produced in a similar format to the inputs, with the LD-derivative of each coordinate projection of the output given separately. Note that for any lexicographically smooth function $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$,

$$\mathbf{f}'(\mathbf{x}; \mathbf{M}) = \begin{bmatrix} [f_1]'(\mathbf{x}; \mathbf{M}) \\ \vdots \\ [f_m]'(\mathbf{x}; \mathbf{M}) \end{bmatrix}.$$

and so an LD-derivative of a vector-valued function is readily assembled from LD-derivatives of its coordinate projections.

By inspection, several approaches to exploiting sparsity in the standard vector forward AD mode [34, Chapters 7 and 8] are also applicable to Algorithm 2, and could reduce the computational cost of LD-derivative evaluation significantly for large problems with sparse computational graphs. For simplicity, however, these approaches were not pursued further in the implementation described above.

4.3.2 Examples

To verify that the implementation of Algorithm 2 works for $\mathcal{PC}^1$-factorable functions, the generalized Jacobian elements computed for $\mathcal{PC}^1$-factorable functions in Examples A.5.1–A.5.5 were also computed using this implementation, setting the direction matrix $\mathbf{M}$ to $\mathbf{I}$ in each instance.

The following example demonstrates LD-derivative evaluation for a function that is L-factorable but not $\mathcal{PC}^1$-factorable, in order to solve a nonlinear complementarity problem using a semismooth Newton method. This example is small for illustration, and shows that, while compositions of nonsmooth functions traditionally make generalized derivative evaluation difficult [23], the methods in this chap-

ter are well-suited to such compositions. As described in Section 5.1, semismooth Newton methods have already been applied successfully to much larger problems in which the required generalized derivatives can be described analytically. Following a similar approach to this example, semismooth Newton methods may be combined with the methods in this chapter to locate points satisfying the Karush–Kuhn–Tucker optimality conditions for a constrained optimization problem, without requiring twice-differentiability of the objective function or constraints.

Example 4.3.1. *Similar to the formulation given in [23], a nonlinear complementarity problem (NCP) involves determining a vector* $\mathbf{x} \in \mathbb{R}^n$ *such that, for a given function* $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$*, the following conditions are satisfied simultaneously:*

$$\mathbf{x} \geq \mathbf{0}, \qquad \mathbf{f}(\mathbf{x}) \geq \mathbf{0}, \qquad \textit{and} \quad \mathbf{x}^{\mathrm{T}}\mathbf{f}(\mathbf{x}) = 0.$$

Such an NCP arises, for example, in formulating the Karush–Kuhn–Tucker optimality conditions for the constrained nonlinear program:

$$\min_{\mathbf{x}\in\mathbb{R}^n} h(\mathbf{x}) \qquad \textit{subject to} \quad \mathbf{x} \geq \mathbf{0},$$

where h *is differentiable, with* $\mathbf{f} := \nabla h$*.*

As discussed in [23], this NCP can be equivalently reformulated as the following nonsmooth equation system, using the Fischer-Burmeister function $(a, b) \mapsto \|(a, b)\|_2 - (a + b)$*:*

$$\mathbf{0} = \mathbf{g}(\mathbf{x}), \qquad \textit{where} \quad g_i : \mathbf{x} \mapsto \|(x_i, f_i(\mathbf{x}))\|_2 - (x_i + f_i(\mathbf{x})), \qquad \forall i \in \{1, \ldots, n\}. \tag{4.4}$$

As a variation of an example from [84], consider an instance of an NCP in which $n = 4$*, and* $\mathbf{f}$ *is defined as the mapping:*

$$\mathbf{f} : \mathbb{R}^4 \to \mathbb{R}^4 : \mathbf{x} \mapsto \begin{bmatrix} 3x_1^2 + 2x_1x_2 + 2x_2^2 + x_3 + 3x_4 + |x_3 - 2x_4 - 3| - 6 \\ 2x_1^2 + x_1 + x_2^2 + 10x_3 + 2x_4 - 2 \\ 3x_1^2 + x_1x_2 + 2x_2^2 + 2x_3 + 9x_4 + |x_3 - 2x_4 - 3| - 9 \\ x_1^2 + 3x_2^2 + 2x_3 + 3x_4 - 3 \end{bmatrix}. \tag{4.5}$$

Though the function $\mathbf{f}$ *is typically* $\mathcal{C}^1$ *in instances of NCP, this is not the case here. This nonsmoothness, however, does not present any problems when applying a semismooth Newton method to solve* (4.4)*, even though the redundancy of the nonsmooth terms in* $\mathbf{f}$ *yields nontrivial generalized Jacobians, as discussed in [23, Example 7.1.15].*

There are two sources of nonsmoothness in the resulting equation system (4.4)*: the 2-norms appearing in the definition of* $\mathbf{g}$*, and the absolute value functions appearing in the definition of* $\mathbf{f}$*. The kink of at least one* $\|\cdot\|_2$ *function in* (4.4) *is reachable: with* $\mathbf{x} = (1, 0, -0.1, 0)$*, for example,* $(x_2, f_2(\mathbf{x})) = (0, 0)$*. As a result, the residual function* $\mathbf{g}$ *in* (4.4) *is not* $\mathcal{PC}^1$ *everywhere, though it is L-factorable, and is readily verified using [75, Theorem 5] to be semismooth in the sense of [92].*

As discussed in Chapters 2 and 3, $\partial_{\mathrm{L}}\mathbf{g}(\mathbf{x}) \subset \partial_{\mathrm{P}}\mathbf{g}(\mathbf{x})$*, and an element of* $\partial_{\mathrm{P}}\mathbf{g}(\mathbf{x})$ *is as useful as an element of* $\partial\mathbf{g}(\mathbf{x})$ *in Qi and Sun's semismooth Newton method [92]. Hence, this semismooth Newton method was used to solve* (4.4) *in this instance, using the developed implementation of Algorithm 2 to evaluate* $\mathbf{g}'(\mathbf{x}_{(k)}; \mathbf{I}) = \mathbf{J}_{\mathrm{L}}\mathbf{g}(\mathbf{x}_{(k)}; \mathbf{I}) \in \partial_{\mathrm{L}}\mathbf{g}(\mathbf{x}_{(k)})$ *when computing the* $(k+1)^{th}$ *Newton step. The C++ library uBLAS included with Version 1.54.0 of Boost [117] was used to solve the linear equation system determining each Newton step. The method was determined to have converged to* $\mathbf{x}^*$ *if* $\|\mathbf{g}(\mathbf{x}^*)\|_2 < 10^{-6}$*.*

Results of the semismooth Newton method are presented for various initial guesses $\mathbf{x}_{(0)}$ *in Table 4.1; the method converged to a solution of the underlying NCP for each initial guess used. Convergence was observed to be at least Q-superlinear for each tested initial guess except* $(-5, 5, -5, 5)$*; for illustration, the progress of the method for an initial guess of* $(1.5, -1, 3.5, 0.25)$ *is shown in Table 4.2. Convergence from the initial guess* $(-5, 5, -5, 5)$ *to the solution* $(0, 0, 0, 1)$ *was observed to be Q-linear; to explain this behavior, the developed implementation of Algorithm 2 was used to show that*

$$\mathbf{J}_{\mathrm{L}}\mathbf{f}\left(\begin{bmatrix}0\\0\\0\\1\end{bmatrix}; \begin{bmatrix}0 & 1 & 0 & 0\\1 & 0 & 0 & 0\\0 & 0 & 1 & 0\\0 & 0 & 0 & 1\end{bmatrix}\right) = \begin{bmatrix}-1 & 0 & 0 & 0\\-1 & 0 & -10 & -2\\0 & 0 & -1 & 0\\0 & 0 & -2 & -3\end{bmatrix},$$

which was also verified by hand. This lexicographic derivative is singular; it follows from [123, Proposition 3(e)] and Corollary 3.2.5 that $\partial\mathbf{f}(0, 0, 0, 1)$ *contains a singular matrix, in*

Initial guess $\mathbf{x}_{(0)}$	Converged solution $\mathbf{x}^*$	Iterations until convergence
$(-5,-5,-5,-5)$	$(0, 0.717, 0.059, 0.447)$	7
$(-5,5,-5,5)$	$(0,0,0,1)$	18
$(5,-5,5,-5)$	$(0.612, 0, 0.75, 0.375)$	6
$(5,5,5,5)$	$(0, 0.717, 0.059, 0.447)$	8
$(5,-5,3,0)$	$(0, 0.717, 0.059, 0.447)$	8
$(5,-5,0,-1.5)$	$(0, 0.717, 0.059, 0.447)$	9
$(1,0,-0.1,0)$	$(0.612, 0, 0.75, 0.375)$	4
$(1.5,-1,3.5,0.25)$	$(1,0,3,0)$	5

Table 4.1: Results of using a semismooth Newton method to solve (4.4)–(4.5)

Iteration k	$\mathbf{x}_{(k)}$	$\|\mathbf{x}_{(k)} - \mathbf{x}^*\|_2$
0	$(1.50, -1.00, 3.50, 0.250)$	1.25
1	$(0.683, -0.004, 1.50, 0.001)$	1.53
2	$(0.997, 0.000, 2.56, 0.000)$	0.442
3	$(1.00, 0.000, 2.96, 0.000)$	3.9×10^{-2}
4	$(1.00, 0.000, 3.00, 0.000)$	2.6×10^{-4}
5	$(1.00, 0.000, 3.00, 0.000)$	1.2×10^{-8}

Table 4.2: Progress of a semismooth Newton method applied to (4.4)–(4.5) with an initial guess of $(1.5, -1, 3.5, 0.25)$

which case the assumptions of Qi and Sun's Q-superlinear convergence result [92, Proposition 3.1] are not satisfied at $(0,0,0,1)$*. Note that this approach does not represent a general method for verifying that a Clarke Jacobian element is singular: the existence of such an element does not imply the existence of a singular lexicographic derivative, and does not suggest which direction matrix to choose if it does exist.*

The solution discussed above, $\bar{\mathbf{x}} = (0,0,0,1)$*, is such that* $\bar{x}_2 = 0$ *and* $f_2(\bar{\mathbf{x}}) = 0$*. Thus, the residual function* $\mathbf{g}$ *is nonsmooth at* $\bar{\mathbf{x}}$*. The function* $\mathbf{f}$ *is nonsmooth at the initial guesses* $(5,-5,3,0)$*,* $(5,-5,0,-1.5)$*, and* $(1.5,-1,3.5,0.25)$*, and at the converged solution* $(1,0,3,0)$*. Moreover, the function* $\mathbf{g}$ *is nonsmooth at the initial guess* $(1,0,-0.1,0)$*, though* $\mathbf{f}$ *is not. Neither this nonsmoothness nor the 2-norms present in the definition of* $\mathbf{g}$ *posed an obstacle when evaluating LD-derivatives.*

The following example shows that even when the requirements for local quadratic convergence of Kojima and Shindo's nonsmooth Newton method [65, Theorem 1] are met, if the pseudo-Jacobian

$$\hat{\mathbf{J}}\mathbf{f}(\mathbf{x}) := \begin{bmatrix} \mathbf{f}'(\mathbf{x};\mathbf{e}_{(1)}) & \cdots & \mathbf{f}'(\mathbf{x};\mathbf{e}_{(n)}) \end{bmatrix}$$

is used in place of an element of the generalized Jacobian at each iteration, then the local convergence of this method is not necessarily preserved.

Example 4.3.2. *Consider the function:*

$$\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^2 : (x_1, x_2) \mapsto \begin{cases} \frac{1}{3}(x_1, x_2 - 2x_1) & \text{if } x_1 \leq 0 \text{ and } x_2 \geq 0, \\ \frac{1}{3}(2x_1, x_2) & \text{if } x_1 \geq 0 \text{ and } x_2 \leq 0, \\ \frac{1}{3}(2x_1 - 2x_2, x_2) & \text{if } x_1 \geq 2x_2 \text{ and } x_2 \geq 0, \\ \frac{1}{3}(x_1, 2x_2 - 2x_1) & \text{if } 2x_1 \leq x_2 \text{ and } x_2 \leq 0, \\ \frac{1}{3}(x_1, x_2) & \text{if } x_1 \leq 2x_2 \text{ and } x_1 \geq 0, \\ & \quad \text{or if } 2x_1 \geq x_2 \text{ and } x_1 \leq 0. \end{cases}$$

It can be verified that $\mathbf{f}$ *is continuous, and is therefore* $\mathcal{PL}$*, with linear selection functions whose Jacobians are each nonsingular. The equation system* $\mathbf{f}(\mathbf{x}^*) = \mathbf{0}$ *has the unique solution* $\mathbf{x}^* = \mathbf{0}$*. This equation system satisfies the hypotheses of [65, Theorem 1], so local Q-quadratic convergence of Kojima and Shindo's Newton method [65, Algorithm EN] to* $\mathbf{0}$ *is guaranteed. Indeed, Algorithm EN in [65] converges after one iteration given any initial guess other than* $\mathbf{0}$*.*

Suppose, instead, that the pseudo-Jacobian $\hat{\mathbf{J}}\mathbf{f}(\mathbf{x})$ *is used in place of a B-subdifferential element at each iteration, and suppose an initial guess of* $\mathbf{x}_{(0)} := (2a, a)$ *is chosen, for some* $a > 0$*. Since* $x_{(0),1} > 0$*,* $x_{(0),2} > 0$*, and* $x_{(0),1} = 2x_{(0),2}$*, it follows that* $\mathbf{f}(\mathbf{x}_{(0)}) = \mathbf{f}(2a, a) = (2a, a)$*, and*

$$\hat{\mathbf{J}}\mathbf{f}(\mathbf{x}_{(0)}) = \frac{1}{3}\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}.$$

Thus, the modified Newton method defines the next iterate as

$$\mathbf{x}_{(1)} := \mathbf{x}_{(0)} - \hat{\mathbf{J}}\mathbf{f}(\mathbf{x}_{(0)})^{-1}\mathbf{f}(\mathbf{x}_{(0)}) = \begin{bmatrix} 2a \\ a \end{bmatrix} - \begin{bmatrix} \frac{3}{2} & 0 \\ 0 & 3 \end{bmatrix}\begin{bmatrix} 2a \\ a \end{bmatrix} = \begin{bmatrix} -a \\ -2a \end{bmatrix}.$$

Since $x_{(1),1} < 0$*,* $x_{(1),2} < 0$*, and* $2x_{(1),1} = x_{(1),2}$*, it follows that* $\mathbf{f}(\mathbf{x}_{(1)}) = \mathbf{f}(-a, -2a) = (-a, -2a)$*, and*

$$\hat{\mathbf{J}}\mathbf{f}(\mathbf{x}_{(1)}) = \frac{1}{3}\begin{bmatrix}1 & 0\\ 0 & 2\end{bmatrix}.$$

Thus, the modified Newton method defines the next iterate as

$$\mathbf{x}_{(2)} := \mathbf{x}_{(1)} - \hat{\mathbf{J}}\mathbf{f}(\mathbf{x}_{(1)})^{-1}\mathbf{f}(\mathbf{x}_{(1)}) = \begin{bmatrix}-a\\ -2a\end{bmatrix} - \begin{bmatrix}3 & 0\\ 0 & \frac{3}{2}\end{bmatrix}\begin{bmatrix}-a\\ -2a\end{bmatrix} = \begin{bmatrix}2a\\ a\end{bmatrix} = \mathbf{x}_{(0)}.$$

It follows that $(2a, a) = \mathbf{x}_{(0)} = \mathbf{x}_{(2)} = \mathbf{x}_{(4)} = \ldots$, *and that* $(-a, -2a) = \mathbf{x}_{(1)} = \mathbf{x}_{(3)} = \mathbf{x}_{(5)} = \ldots$. *Hence, this modified Newton method does not converge given an initial guess of* $(2a, a)$ *for any* $a > 0$, *and therefore does not have a neighborhood of convergence around the solution* $\mathbf{x}^* = \mathbf{0}$. *This shows that the modified Newton method does not necessarily have a neighborhood of convergence around a solution, even when conditions for local quadratic convergence of Kojima and Shindo's nonsmooth Newton method are met.*

4.4 Conclusions

A vector forward AD mode has been developed for evaluating LD-derivatives for L-factorable functions, from which lexicographic derivatives are readily obtained. This method improves our earlier method in Appendix A, which was the first computationally tractable method for computation of generalized derivatives for a broad class of nonsmooth vector-valued functions. Moreover, the method in this chapter incorporates the computational advantages of the method in [33]; it can be implemented in a tapeless fashion using operator overloading.

Chapter 5

Lexicographic derivatives for solutions of nonsmooth ODEs

5.1 Introduction

This chapter reproduces the article [55] and the proceedings [57]. For any locally Lipschitz continuous mapping between finite-dimensional Euclidean spaces, *Clarke's generalized Jacobian* [16] is a set-valued mapping that provides useful local sensitivity information. Elements of Clarke's generalized Jacobian are used in *semismooth Newton methods* for equation-solving [65, 92], and in *bundle methods* for local optimization [63, 67, 70]. Chapter 4, Appendix A, and [33] present methods to evaluate generalized Jacobian elements for finite compositions of simple smooth and nonsmooth functions. However, there is currently no general method for determining generalized Jacobian elements for nonsmooth dynamic systems, which are defined in this chapter to be parametric Carathéodory ordinary differential equations (ODEs) with right-hand side functions that are not necessarily differentiable with respect to the dependent variables and parameters. These ODEs will be referred to as *nonsmooth parametric ODEs* throughout this chapter.

Classical results concerning parametric sensitivities of solutions of parametric ODEs require that the ODE right-hand side function has continuous partial deriva-

tives, and imply differentiability of a unique solution with respect to the parameters [35]. These results can be extended to certain hybrid discrete/continuous dynamic systems, in which any discontinuities or kinks in an otherwise differentiable solution are defined as the solutions of equation systems with residual functions that are both continuously differentiable and locally invertible [30]. Nevertheless, the following example shows that a solution of a nonsmooth parametric ODE system is not necessarily differentiable with respect to the parameters. In this case, classical sensitivity results for parametric ODEs do not apply.

Even if the solutions of nonsmooth parametric ODEs are known to be smooth or convex functions of the ODE parameters, there is no general method for evaluating their gradients or subgradients. Such applications arise in global optimization of systems with nonconvex parametric ODE solutions embedded, where convex underestimators of these nonconvex ODE solutions have been described as solutions of corresponding nonsmooth parametric ODEs [102].

As described in Chapter 2, Clarke [16, Theorem 7.4.1] presents the primary existing result describing generalized Jacobians of parametric ODE solutions, in which certain supersets of generalized Jacobians of the ODE solutions are constructed. Using properties of these supersets, sufficient conditions for the differentiability of the original ODE solution have been formulated [16, 125].

Pang and Stewart [88, Theorem 11 and Corollary 12] show that when a parametric ODE has a right-hand side function that is *semismooth* in the sense of Qi [92], the generalized Jacobian supersets described by Clarke are in fact *linear Newton approximations* about any domain point. As summarized in Section 7.5.1 of [23], a linear Newton approximation for a locally Lipschitz continuous function about a domain point is a set-valued mapping containing local sensitivity information. Throughout this chapter, all discussed linear Newton approximations are linear Newton approximations about *every* domain point simultaneously; any reference to a linear Newton approximation of a function at a domain point refers to the value of this linear Newton approximation when evaluated at that domain point. Yunt [124] extends Pang and Stewart's result to adjoint sensitivities, systems de-

scribed by index-1 differential-algebraic equations, and multi-stage systems with discontinuities in the right-hand side function occurring only at finitely many known values of the independent variable. However, the following example shows that linear Newton approximations are not guaranteed to satisfy certain properties that are satisfied by Clarke's generalized Jacobian. In particular, the linear Newton approximation of a continuously differentiable function at a domain point can include elements other than the derivative of the function at that point. Moreover, the linear Newton approximation of a convex scalar-valued function at a domain point can include elements that are not subgradients of the function at that point. Thirdly, given a convex scalar-valued function on an open set, the fact that the linear Newton approximation of the function at a domain point contains the origin is not a sufficient condition for a global minimum. Clarke's generalized Jacobian for a locally Lipschitz function, on the other hand, includes only the derivative whenever the function is continuously differentiable, and is identical to the convex subdifferential whenever the function is scalar-valued and convex [16]. In the latter case, the fact that the value of Clarke's generalized Jacobian at a domain point contains the origin is sufficient for a global minimum on an open set.

Example 5.1.1. *Consider the mappings* $f : \mathbb{R} \to \mathbb{R} : x \mapsto x$, $g : \mathbb{R} \to \mathbb{R} : x \mapsto \max\{x, 0\}$, *and* $h : \mathbb{R} \to \mathbb{R} : x \mapsto \min\{x, 0\}$. *Using [16, Theorem 2.5.1], the Clarke generalized Jacobians of g and h are evaluated as:*

$$\partial g(x) = \begin{cases} \{0\}, & \text{if } x < 0, \\ [0,1], & \text{if } x = 0, \\ \{1\}, & \text{if } x > 0, \end{cases} \qquad \partial h(x) = \begin{cases} \{1\}, & \text{if } x < 0, \\ [0,1], & \text{if } x = 0, \\ \{0\}, & \text{if } x > 0. \end{cases}$$

Now, g and h are each piecewise linear, and are therefore semismooth [23]. Since $f \equiv g + h$ *on* $\mathbb{R}$, *it follows from [23, Corollary 7.5.18] that the following set-valued mapping is a linear Newton approximation for f:*

$$\Gamma f : x \mapsto \partial g(x) + \partial h(x) = \begin{cases} \{1\}, & \text{if } x < 0, \\ [0,2], & \text{if } x = 0, \\ \{1\}, & \text{if } x > 0. \end{cases}$$

By inspection, f is convex and continuously differentiable on its domain, and has a deriva-

tive of $\mathbf{J}f(x) = 1$ *for each* $x \in \mathbb{R}$. *In addition,* f *does not have any local minima on* $\mathbb{R}$. *However, although* $\mathbf{J}f(0) \neq 0, 0 \in \Gamma f(0)$.

The plenary hull of Clarke's generalized Jacobian has been investigated in [41, 109, 123], and is referred to in this thesis as the *plenary Jacobian*. Though the plenary Jacobian is a superset of the generalized Jacobian, it satisfies several key nonsmooth analysis results in place of the generalized Jacobian. A benefit of the plenary Jacobian is that membership of the plenary Jacobian is easier to verify than membership of Clarke's generalized Jacobian. As argued in Chapter 2, the plenary Jacobian is in some sense as good a linear Newton approximation as the generalized Jacobian, and is just as useful in semismooth Newton methods and in bundle methods.

Sensitivities for unique solutions of a smooth parametric ODE system are traditionally expressed as the unique solutions of a corresponding linear ODE system obtained from the original system by application of the chain rule, as summarized in [35, Ch. V, Theorem 3.1]. In this spirit, the goal of this chapter is to present the first description of a plenary Jacobian element of the unique solution of a nonsmooth parametric ODE system as the unique solution of another ODE system. Nesterov's lexicographic derivatives [79] are used as a tool to construct this plenary Jacobian element.

The following section presents the main results of this chapter, in which directional derivatives and lexicographic derivatives for solutions of nonsmooth parametric ODEs are expressed as the unique solutions of corresponding ODE systems. Various implications of these results are discussed.

5.2 Generalized derivatives for solutions of parametric ODEs

This section extends a result by Pang and Stewart [88] to show that when the right-hand side function of a Carathéodory ODE [26] is directionally differentiable with

respect to the dependent variables, then directional derivatives of any ODE solution with respect to its initial condition can be expressed as the solution of another Carathéodory ODE. This result is in turn extended to show that if the right-hand side function of the original ODE is L-smooth with respect to the dependent variables, then lexicographic derivatives of the ODE solution can also be expressed as the unique solution of another ODE. This latter ODE may be decoupled into a sequence of Carathéodory ODEs, but does not necessarily satisfy the Carathéodory assumptions itself.

5.2.1 Propagating directional derivatives

The following theorem extends a result by Pang and Stewart [88, Theorem 7] concerning directional derivatives of ODE solutions to the case in which direct dependence of the ODE right-hand side function on the independent variable is measurable but not necessarily continuous. This theorem and the subsequent corollary show that these directional derivatives uniquely solve a corresponding ODE whose right-hand side function may be discontinuous in the independent variable, even if the right-hand side function of the original ODE was continuous. Hence, allowing for discontinuous dependence of the original right-hand side function on the independent variable is essential when using these results in inductive proofs to describe higher-order directional derivatives and lexicographic derivatives of the ODE solution.

Theorem 5.2.1. *Given an open, connected set $X \subset \mathbb{R}^n$ and real numbers $t_0 < t_f$, suppose that a function $\mathbf{f} : [t_0, t_f] \times X \to \mathbb{R}^n$ satisfies the following conditions:*

- *the mapping $\mathbf{f}(\cdot, \mathbf{c}) : [t_0, t_f] \to \mathbb{R}^n$ is measurable for each $\mathbf{c} \in X$,*
- *for each $t \in [t_0, t_f]$ except in a zero-measure subset $Z_{\mathbf{f}}$, the mapping $\mathbf{f}(t, \cdot) : X \to \mathbb{R}^n$ is continuous and directionally differentiable,*
- *with $\mathbf{x}(\cdot, \mathbf{c})$ denoting any solution of the parametric ODE system:*

$$\frac{d\mathbf{x}}{dt}(t,\mathbf{c}) = \mathbf{f}(t,\mathbf{x}(t,\mathbf{c})), \qquad \mathbf{x}(t_0,\mathbf{c}) = \mathbf{c}, \tag{5.1}$$

there exists a solution $\{\mathbf{x}(t,\mathbf{c}_0) : t \in [t_0,t_f]\} \subset X$ *for some* $\mathbf{c}_0 \in X$*,*

- *there exists an open set* $N \subset X$ *such that* $\{\mathbf{x}(t,\mathbf{c}_0) : t \in [t_0,t_f]\} \subset N$*, and such that there exist Lebesgue integrable functions* $k_{\mathbf{f}}, m_{\mathbf{f}} : [t_0,t_f] \to \mathbb{R}_+ \cup \{+\infty\}$ *for which*

$$\|\mathbf{f}(t,\mathbf{c})\| \le m_{\mathbf{f}}(t), \qquad \forall t \in [t_0,t_f], \quad \forall \mathbf{c} \in N,$$

and

$$\|\mathbf{f}(t,\mathbf{c}_1) - \mathbf{f}(t,\mathbf{c}_2)\| \le k_{\mathbf{f}}(t)\,\|\mathbf{c}_1 - \mathbf{c}_2\|, \qquad \forall t \in [t_0,t_f], \quad \forall \mathbf{c}_1,\mathbf{c}_2 \in N.$$

Then, for each $t \in [t_0,t_f]$*, the function* $\mathbf{x}_t \equiv \mathbf{x}(t,\cdot)$ *is well-defined and Lipschitz continuous on a neighborhood of* $\mathbf{c}_0$*, with a Lipschitz constant that is independent of* t*. Moreover,* $\mathbf{x}_t$ *is directionally differentiable at* $\mathbf{c}_0$ *for each* $t \in [t_0,t_f]$*, and for each* $\mathbf{d} \in \mathbb{R}^n$*, the mapping* $t \mapsto [\mathbf{x}_t]'(\mathbf{c}_0;\mathbf{d})$ *is the unique solution on* $[t_0,t_f]$ *of the ODE:*

$$\frac{d\mathbf{y}}{dt}(t) = [\hat{\mathbf{f}}_t]'(\mathbf{x}(t,\mathbf{c}_0);\mathbf{y}(t)), \qquad \mathbf{y}(t_0) = \mathbf{d}, \tag{5.2}$$

where $\hat{\mathbf{f}}_t : X \to \mathbb{R}^n$ *is defined in terms of* $\mathbf{f}$ *as follows, and is directionally differentiable for each* $t \in [t_0,t_f]$*:*

$$\hat{\mathbf{f}}_t(\mathbf{c}) = \begin{cases} \mathbf{f}(t,\mathbf{c}), & \text{if } t \in [t_0,t_f]\backslash Z_{\mathbf{f}}, \\ \mathbf{0}, & \text{if } t \in Z_{\mathbf{f}}. \end{cases}$$

Proof. By [26, Ch. 1, §1, Theorem 2], $\mathbf{x}(\cdot,\mathbf{c}_0)$ is the unique solution of (5.1) on $[t_0,t_f]$ with $\mathbf{c} = \mathbf{c}_0$. Consequently, by [26, Ch. 1, §1, Theorems 2 and 6], there exists a neighborhood $N_0 \subset N$ of $\mathbf{c}_0$ such that for each $\mathbf{c} \in N_0$, there exists a unique solution $\{\mathbf{x}(t,\mathbf{c}) : t \in [t_0,t_f]\} \subset N$ of (5.1).

To obtain the Lipschitz continuity of $\mathbf{x}(t,\cdot)$ near $\mathbf{c}_0$, choose any $t \in [t_0,t_f]$ and any $\mathbf{c}_1,\mathbf{c}_2 \in N_0$. Since the ODE solutions $\mathbf{x}(\cdot,\mathbf{c}_1)$ and $\mathbf{x}(\cdot,\mathbf{c}_2)$ exist on $[t_0,t_f]$, it

follows that

$$\begin{aligned}\|\mathbf{x}(t,\mathbf{c}_1)-\mathbf{x}(t,\mathbf{c}_2)\| &= \left\|\mathbf{c}_1+\int_{t_0}^{t}\mathbf{f}(s,\mathbf{x}(s,\mathbf{c}_1))\,\mathrm{d}s-\mathbf{c}_2-\int_{t_0}^{t}\mathbf{f}(s,\mathbf{x}(s,\mathbf{c}_2))\,\mathrm{d}s\right\|,\\ &\leq \|\mathbf{c}_1-\mathbf{c}_2\|+\int_{t_0}^{t}k_{\mathbf{f}}(s)\,\|\mathbf{x}(s,\mathbf{c}_1)-\mathbf{x}(s,\mathbf{c}_2)\|\,\mathrm{d}s.\end{aligned}$$

Let $k_{\mathbf{x}} := \exp\left(\int_{t_0}^{t_f} k_{\mathbf{f}}(s)\,\mathrm{d}s\right)$. Applying the version of Gronwall's Inequality described in Section 1 of [122], since the above inequality holds with any $\bar{t}\in[t_0,t]$ in place of t, it follows that

$$\|\mathbf{x}(t,\mathbf{c}_1)-\mathbf{x}(t,\mathbf{c}_2)\| \leq \|\mathbf{c}_1-\mathbf{c}_2\|\exp\left(\int_{t_0}^{t}k_{\mathbf{f}}(s)\,\mathrm{d}s\right)\leq k_{\mathbf{x}}\|\mathbf{c}_1-\mathbf{c}_2\|,\quad \forall \mathbf{c}_1,\mathbf{c}_2\in N_0.$$

This demonstrates the Lipschitz continuity of $\mathbf{x}(t,\cdot)$ near $\mathbf{c}_0$ for each $t\in[t_0,t_f]$, with a Lipschitz constant $k_{\mathbf{x}}$ that is independent of t.

By construction of $\hat{\mathbf{f}}_t$, $\hat{\mathbf{f}}_t$ is directionally differentiable on its domain for each $t\in[t_0,t_f]\backslash Z_{\mathbf{f}}$. $\hat{\mathbf{f}}_t$ is the zero function on $Z_{\mathbf{f}}$, and is therefore also directionally differentiable for each $t\in Z_{\mathbf{f}}$. Hence, $\hat{\mathbf{f}}_t$ is directionally differentiable for each $t\in[t_0,t_f]$. The mapping $\mathbf{g}:[t_0,t_f]\times\mathbb{R}^n\to\mathbb{R}^n:(t,\mathbf{v})\mapsto[\hat{\mathbf{f}}_t]'(\mathbf{x}(t,\mathbf{c}_0);\mathbf{v})$ is therefore well-defined, and is the right-hand side function of the ODE (5.2).

Now, choose any particular $\mathbf{v}\in\mathbb{R}^n$. Since $\mathbf{x}(\cdot,\mathbf{c}_0)$ is continuous on the compact set $[t_0,t_f]$, the set $\{\mathbf{x}(t,\mathbf{c}_0):t\in[t_0,t_f]\}\subset N$ is compact, and does not contain any points in the closed set $(\mathbb{R}^n\backslash N)$. Thus, there exists $\delta>0$ such that for any $\tau\in[0,\delta]$ and any $t\in[t_0,t_f]$, $(\mathbf{x}(t,\mathbf{c}_0)+\tau\mathbf{v})\in N$; this is trivial when $N=\mathbb{R}^n$, and follows from [26, Ch. 2, §5, Lemma 1] otherwise. Since $\mathbf{x}(\cdot,\mathbf{c}_0)$ is continuous on $[t_0,t_f]$, [26, Ch. 1, §1, Lemma 1] shows that the mapping $t\mapsto\mathbf{f}(t,\mathbf{x}(t,\mathbf{c}_0)+\tau\mathbf{v})$ is measurable on $[t_0,t_f]$ for each $\tau\in[0,\delta]$. Thus, the mapping $t\mapsto\hat{\mathbf{f}}_t(\mathbf{x}(t,\mathbf{c}_0)+\tau\mathbf{v})$ is also measurable on $[t_0,t_f]$ for each $\tau\in[0,\delta]$.

For each $\tau\in(0,\delta]$, the previous paragraph implies that the following mapping is well-defined and measurable:

$$\gamma_\tau : [t_0, t_f] \to \mathbb{R}^n : t \mapsto \frac{\hat{\mathbf{f}}_t(\mathbf{x}(t, \mathbf{c}_0) + \tau \mathbf{v}) - \hat{\mathbf{f}}_t(\mathbf{x}(t, \mathbf{c}_0))}{\tau}.$$

It follows from the directional differentiability of $\hat{\mathbf{f}}_t$ and the definition of $\mathbf{g}$ that for each $t \in [t_0, t_f]$, $\mathbf{g}(t, \mathbf{v}) = \lim_{\tau \to 0^+} \gamma_\tau(t)$. Noting that $\mathbf{v} \in \mathbb{R}^n$ was chosen arbitrarily, it follows that for each $\mathbf{v} \in \mathbb{R}^n$, the mapping $\mathbf{g}(\cdot, \mathbf{v})$ is the pointwise limit of a sequence of measurable functions, and is therefore measurable on $[t_0, t_f]$.

Now, define $Z_{k_\mathbf{f}} := \{t \in [t_0, t_f] : k_\mathbf{f}(t) = +\infty\}$. Since $k_\mathbf{f}$ is integrable on $[t_0, t_f]$, $Z_{k_\mathbf{f}}$ has zero measure. For each $t \in [t_0, t_f] \backslash (Z_\mathbf{f} \cup Z_{k_\mathbf{f}})$, the definition of $k_\mathbf{f}$ implies that $k_\mathbf{f}(t)$ is a finite Lipschitz constant for $\mathbf{f}_t$ near $\mathbf{x}(t, \mathbf{c}_0)$. Thus, [97, Theorem 3.1.2] implies that

$$\|\mathbf{g}(t, \mathbf{v}_1) - \mathbf{g}(t, \mathbf{v}_2)\| \leq k_\mathbf{f}(t) \, \|\mathbf{v}_1 - \mathbf{v}_2\|, \qquad \forall t \in [t_0, t_f] \backslash (Z_\mathbf{f} \cup Z_{k_\mathbf{f}}), \quad \forall \mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^n.$$

The above relationship still holds if $t \in Z_\mathbf{f}$, since $\mathbf{g}(t, \mathbf{v}) = \mathbf{0}$ for each $\mathbf{v} \in \mathbb{R}^n$ in this case. The relationship also holds if $t \in Z_{k_\mathbf{f}}$, since $k_\mathbf{f}(t) = +\infty$ in this case. Combining these cases,

$$\|\mathbf{g}(t, \mathbf{v}_1) - \mathbf{g}(t, \mathbf{v}_2)\| \leq k_\mathbf{f}(t) \, \|\mathbf{v}_1 - \mathbf{v}_2\|, \qquad \forall t \in [t_0, t_f], \quad \forall \mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^n. \tag{5.3}$$

Choose some $\mathbf{d} \in \mathbb{R}^n$, and let $m_\mathbf{y} := \|\mathbf{d}\| \exp\left(\int_{t_0}^{t_f} k_\mathbf{f}(s) \, \mathrm{d}s\right) + \|\mathbf{d}\| + 1$. Since $\mathbf{g}(t, \mathbf{0}) = \mathbf{0}$ for each $t \in [t_0, t_f]$, it follows that whenever $\|\mathbf{v}\| < m_\mathbf{y}$,

$$\|\mathbf{g}(t, \mathbf{v})\| = \|\mathbf{g}(t, \mathbf{v}) - \mathbf{g}(t, \mathbf{0})\| \leq k_\mathbf{f}(t) \, \|\mathbf{v}\| \leq k_\mathbf{f}(t) \, m_\mathbf{y}. \tag{5.4}$$

Thus, $\|\mathbf{g}\|$ is bounded above on $[t_0, t_f] \times \{\mathbf{v} \in \mathbb{R}^n : \|\mathbf{v}\| < m_\mathbf{y}\}$ by an integrable function of t. By Carathéodory's existence theorem [18, Ch. 2, Theorems 1.1 and 1.3], there exists a solution $\mathbf{y}$ of (5.2) on $[t_0, \bar{t}]$, where $\bar{t}$ is the least element of $]t_0, t_f]$ for which either $\bar{t} = t_f$, $\|\mathbf{y}(\bar{t})\| \geq m_\mathbf{y}$, or both. Now, for each $t \in [t_0, \bar{t}]$, (5.4) implies that

$$\|\mathbf{y}(t)\| = \left\| \mathbf{d} + \int_{t_0}^{t} \mathbf{g}(s, \mathbf{y}(s)) \, \mathrm{d}s \right\| \leq \|\mathbf{d}\| + \int_{t_0}^{t} k_\mathbf{f}(s) \, \|\mathbf{y}(s)\| \, \mathrm{d}s.$$

Thus, Gronwall's Inequality [122] implies that

$$\|\mathbf{y}(\bar{t})\| \leq \|\mathbf{d}\| \exp\left(\int_{t_0}^{\bar{t}} k_{\mathbf{f}}(s)\,\mathrm{d}s\right) \leq \|\mathbf{d}\| \exp\left(\int_{t_0}^{t_f} k_{\mathbf{f}}(s)\,\mathrm{d}s\right) < m_{\mathbf{y}}.$$

Comparing this inequality with the definition of $\bar{t}$, it follows that $\bar{t} = t_f$, and so there exists a solution $\mathbf{y}$ of (5.2) on $[t_0, t_f]$. Moreover, (5.3), (5.4), and [26, Ch. 1, §1, Theorem 2] show that this solution is unique.

The remainder of this proof proceeds similarly to the proof of [88, Theorem 7]. For sufficiently small $\bar{\tau} > 0$, $(\mathbf{c}_0 + \tau\mathbf{d}) \in N_0$. Thus, for each choice of $t \in [t_0, t_f]$ and $\tau \in (0, \bar{\tau}]$, let

$$\mathbf{e}_{\mathbf{x}}(t,\tau) := \frac{\mathbf{x}(t, \mathbf{c}_0 + \tau\mathbf{d}) - \mathbf{x}(t, \mathbf{c}_0)}{\tau} - \mathbf{y}(t),$$

and

$$\mathbf{e}_{\mathbf{f}}(t,\tau) := \frac{\mathbf{f}(t, \mathbf{x}(t, \mathbf{c}_0 + \tau\mathbf{d})) - \mathbf{f}(t, \mathbf{x}(t, \mathbf{c}_0))}{\tau} - \mathbf{g}\left(t, \frac{\mathbf{x}(t, \mathbf{c}_0 + \tau\mathbf{d}) - \mathbf{x}(t, \mathbf{c}_0)}{\tau}\right).$$

It follows from the established bounds that for each $t \in [t_0, t_f]$ and each $\tau \in (0, \bar{\tau}]$,

$$\begin{aligned}
\|\mathbf{e}_{\mathbf{f}}(t,\tau)\| &\leq \left\|\frac{\mathbf{f}(t, \mathbf{x}(t, \mathbf{c}_0 + \tau\mathbf{d})) - \mathbf{f}(t, \mathbf{x}(t, \mathbf{c}_0))}{\tau}\right\| + \left\|\mathbf{g}\left(t, \frac{\mathbf{x}(t, \mathbf{c}_0 + \tau\mathbf{d}) - \mathbf{x}(t, \mathbf{c}_0)}{\tau}\right)\right\|, \\
&\leq \frac{k_{\mathbf{f}}(t)}{\tau}\|\mathbf{x}(t, \mathbf{c}_0 + \tau\mathbf{d}) - \mathbf{x}(t, \mathbf{c}_0)\| + \frac{k_{\mathbf{f}}(t)}{\tau}\|\mathbf{x}(t, \mathbf{c}_0 + \tau\mathbf{d}) - \mathbf{x}(t, \mathbf{c}_0)\|, \\
&\leq 2k_{\mathbf{x}}k_{\mathbf{f}}(t)\|\mathbf{d}\|. \qquad (5.5)
\end{aligned}$$

Now, (5.3) and the definitions of $\mathbf{e}_{\mathbf{x}}$ and $\mathbf{e}_{\mathbf{f}}$ imply that for each $t \in [t_0, t_f]$ and $\tau \in (0, \bar{\tau}]$,

$$\begin{aligned}
\|\mathbf{e}_{\mathbf{x}}(t,\tau)\| &= \left\|\int_{t_0}^{t} \left(\frac{\mathbf{f}(s, \mathbf{x}(s, \mathbf{c}_0 + \tau\mathbf{d})) - \mathbf{f}(s, \mathbf{x}(s, \mathbf{c}_0))}{\tau} - \mathbf{g}(s, \mathbf{y}(s))\right)\mathrm{d}s\right\|, \\
&= \left\|\int_{t_0}^{t} \left(\mathbf{e}_{\mathbf{f}}(s,\tau) + \mathbf{g}\left(s, \frac{\mathbf{x}(s, \mathbf{c}_0 + \tau\mathbf{d}) - \mathbf{x}(s, \mathbf{c}_0)}{\tau}\right) - \mathbf{g}(s, \mathbf{y}(s))\right)\mathrm{d}s\right\|, \\
&\leq \int_{t_0}^{t} \left(\|\mathbf{e}_{\mathbf{f}}(s,\tau)\| + k_{\mathbf{f}}(s)\|\mathbf{e}_{\mathbf{x}}(s,\tau)\|\right)\mathrm{d}s.
\end{aligned}$$

Since $\|\mathbf{e}_{\mathbf{x}}(\cdot,\tau)\|$ is continuous, it is bounded on the compact set $[t_0, t_f]$. Hence, the mapping $t \mapsto k_{\mathbf{f}}(t)\|\mathbf{e}_{\mathbf{x}}(t,\tau)\|$ is integrable on $[t_0, t_f]$. This permits application of a variation [122, Theorem 2] of Gronwall's Inequality, which yields the following for any $t \in [t_0, t_f]$ and $\tau \in (0, \bar{\tau}]$:

$$0 \leq \|\mathbf{e}_{\mathbf{x}}(t,\tau)\| \leq \int_{t_0}^{t} \|\mathbf{e}_{\mathbf{f}}(s,\tau)\| \exp\left(\int_{s}^{t} k_{\mathbf{f}}(r)\,\mathrm{d}r\right)\mathrm{d}s \leq k_{\mathbf{x}} \int_{t_0}^{t} \|\mathbf{e}_{\mathbf{f}}(s,\tau)\|\,\mathrm{d}s. \quad (5.6)$$

Substituting (5.5) into (5.6) for each $t \in [t_0, t_f]$ and $\tau \in (0, \bar{\tau}]$ yields $\|\mathbf{e}_{\mathbf{x}}(t,\tau)\| \leq m_{\mathbf{e}_{\mathbf{x}}}$, where

$$m_{\mathbf{e}_{\mathbf{x}}} := 2\|\mathbf{d}\|(k_{\mathbf{x}})^2 \int_{t_0}^{t_f} k_{\mathbf{f}}(s)\,\mathrm{d}s.$$

Now, for each $t \in [t_0, t_f]\backslash(Z_{\mathbf{f}} \cup Z_{k_{\mathbf{f}}})$, the assumptions of the theorem imply that $\mathbf{f}(t,\cdot)$ is directionally differentiable and Lipschitz continuous on X, with a Lipschitz constant of $k_{\mathbf{f}}(t)$. Hence, (2.1) implies that for each $t \in [t_0, t_f]\backslash(Z_{\mathbf{f}} \cup Z_{k_{\mathbf{f}}})$, for each $\epsilon > 0$, there exists some $\delta_{t,\epsilon} > 0$ such that if $\|\mathbf{h}\| < \delta_{t,\epsilon}$,

$$\|\mathbf{f}(t, \mathbf{x}(t,\mathbf{c}_0) + \mathbf{h}) - \mathbf{f}(t, \mathbf{x}(t,\mathbf{c}_0)) - \mathbf{g}(t,\mathbf{h})\| \leq \epsilon\|\mathbf{h}\|.$$

Moreover, the Lipschitz continuity of $\mathbf{x}(t,\cdot)$ on N_0 for each $t \in [t_0, t_f]$ implies that for any $\epsilon > 0$, any $t \in [t_0, t_f]\backslash(Z_{\mathbf{f}} \cup Z_{k_{\mathbf{f}}})$, and any $\tau \in (0, \min\{\bar{\tau}, \frac{\delta_{t,\epsilon}}{k_{\mathbf{x}}\|\mathbf{d}\|+1}\})$,

$$\|\mathbf{x}(t, \mathbf{c}_0 + \tau\mathbf{d}) - \mathbf{x}(t,\mathbf{c}_0)\| \leq k_{\mathbf{x}}\tau\|\mathbf{d}\| < \delta_{t,\epsilon}.$$

Thus, if $t \in [t_0, t_f]\backslash(Z_{\mathbf{f}} \cup Z_{k_{\mathbf{f}}})$ and $0 < \tau < \min\{\bar{\tau}, \frac{\delta_{t,\epsilon}}{k_{\mathbf{x}}\|\mathbf{d}\|+1}\}$,

$$\begin{aligned}&\|\mathbf{f}(t, \mathbf{x}(t,\mathbf{c}_0 + \tau\mathbf{d})) - \mathbf{f}(t, \mathbf{x}(t,\mathbf{c}_0)) - \mathbf{g}(t, \mathbf{x}(t,\mathbf{c}_0 + \tau\mathbf{d}) - \mathbf{x}(t,\mathbf{c}_0))\| \\ &\quad \leq \epsilon\|\mathbf{x}(t,\mathbf{c}_0 + \tau\mathbf{d}) - \mathbf{x}(t,\mathbf{c}_0)\|.\end{aligned}$$

Noting that $\mathbf{g}(t,\cdot)$ is positively homogeneous and that $\tau > 0$, dividing both sides of the above inequality by τ yields the following, for each $t \in [t_0, t_f]\backslash(Z_{\mathbf{f}} \cup Z_{k_{\mathbf{f}}})$, each $\epsilon > 0$, and each $\tau \in (0, \min\{\bar{\tau}, \frac{\delta_{t,\epsilon}}{k_{\mathbf{x}}\|\mathbf{d}\|+1}\})$:

$$\|\mathbf{e_f}(t,\tau)\| \leq \epsilon \left\| \frac{\mathbf{x}(t,\mathbf{c}_0+\tau\mathbf{d}) - \mathbf{x}(t,\mathbf{c}_0)}{\tau} \right\| = \epsilon \|\mathbf{e_x}(t,\tau) + \mathbf{y}(t)\| < \epsilon(m_{\mathbf{e_x}} + m_{\mathbf{y}}).$$

Thus, $\lim_{\tau\to 0^+} \|\mathbf{e_f}(t,\tau)\| = 0$ for almost all $t \in [t_0, t_f]$. Using this limit and the bound (5.5), applying the dominated convergence theorem to (5.6) yields

$$\lim_{\tau\to 0^+} \|\mathbf{e_x}(t,\tau)\| = 0, \qquad \forall t \in [t_0, t_f].$$

Hence,

$$\lim_{\tau\to 0^+} \frac{\mathbf{x}(t,\mathbf{c}_0+\tau\mathbf{d}) - \mathbf{x}(t,\mathbf{c}_0)}{\tau} = \mathbf{y}(t), \qquad \forall t \in [t_0, t_f].$$

Noting that $\mathbf{d} \in \mathbb{R}^n$ was chosen arbitrarily, it follows that for each $\mathbf{d} \in \mathbb{R}^n$, the directional derivative $[\mathbf{x}_t]'(\mathbf{c}_0;\mathbf{d})$ exists and is finite for each $t \in [t_0, t_f]$, and so $\mathbf{x}_t$ is directionally differentiable at $\mathbf{c}_0$ for each $t \in [t_0, t_f]$. Moreover, the above equation shows that $t \mapsto [\mathbf{x}_t]'(\mathbf{c}_0;\mathbf{d})$ is the unique solution $\mathbf{y}$ of (5.2) on $[t_0, t_f]$. □

Corollary 5.2.2. *Under the assumptions of Theorem 5.2.1, and using the same notation as in the theorem, the mapping* $\mathbf{g} : [t_0, t_f] \times \mathbb{R}^n \to \mathbb{R}^n : (t, \mathbf{v}) \mapsto [\hat{\mathbf{f}}_t]'(\mathbf{x}(t,\mathbf{c}_0);\mathbf{v})$ *satisfies the following conditions:*

- *the mapping* $\mathbf{g}(\cdot,\mathbf{v}) : [t_0, t_f] \to \mathbb{R}^n$ *is measurable for each* $\mathbf{v} \in \mathbb{R}^n$*,*
- *for each* $t \in [t_0, t_f]$ *except in a zero-measure set* $Z_{\mathbf{g}}$*, the mapping* $\mathbf{g}(t,\cdot) : \mathbb{R}^n \to \mathbb{R}^n$ *is defined and continuous,*
- *for each* $\mathbf{d} \in \mathbb{R}^n$*, there exists a solution* $\{\mathbf{z}(t,\mathbf{d}) : t \in [t_0, t_f]\} \subset \mathbb{R}^n$ *of the parametric ODE system:*

$$\frac{d\mathbf{z}}{dt}(t,\mathbf{d}) = \mathbf{g}(t,\mathbf{z}(t,\mathbf{d})), \qquad \mathbf{z}(t_0,\mathbf{d}) = \mathbf{d}.$$

- *for each* $\mathbf{d} \in \mathbb{R}^n$*, there exists an open set* $N_{\mathbf{g}}(\mathbf{d}) \subset \mathbb{R}^n$ *such that*

$$\{\mathbf{z}(t,\mathbf{d}) : t \in [t_0, t_f]\} \subset N_{\mathbf{g}}(\mathbf{d}),$$

and such that there exist Lebesgue integrable functions $k_{\mathbf{g}}, m_{\mathbf{g}} : [t_0, t_f] \to \mathbb{R}_+ \cup \{+\infty\}$ *for which*

$$\|\mathbf{g}(t, \mathbf{v})\| \leq m_{\mathbf{g}}(t), \qquad \forall t \in [t_0, t_f], \quad \forall \mathbf{v} \in N_{\mathbf{g}}(\mathbf{d}),$$

and

$$\|\mathbf{g}(t, \mathbf{v}_1) - \mathbf{g}(t, \mathbf{v}_2)\| \leq k_{\mathbf{g}}(t)\, \|\mathbf{v}_1 - \mathbf{v}_2\|, \qquad \forall t \in [t_0, t_f], \quad \forall \mathbf{v}_1, \mathbf{v}_2 \in N_{\mathbf{g}}(\mathbf{d}).$$

If, in addition, the mapping $\mathbf{f}(t, \cdot) : X \to \mathbb{R}^n$ *is L-smooth for each* $t \in [t_0, t_f] \backslash Z_{\mathbf{f}}$*, then the mapping* $\mathbf{g}(t, \cdot) : \mathbb{R}^n \to \mathbb{R}^n$ *is L-smooth for each* $t \in [t_0, t_f]$*. In this case, the set* $Z_{\mathbf{g}}$ *described above may be set to* $\emptyset$*.*

Proof. The measurability of $\mathbf{g}(\cdot, \mathbf{v})$ and the existence and Lipschitz continuity of $\mathbf{g}(t, \cdot)$ except on some zero-measure set $Z_{\mathbf{g}} \subset (Z_{\mathbf{f}} \cup Z_{k_{\mathbf{f}}})$ were established in the proof of Theorem 5.2.1. For any $\mathbf{d} \in \mathbb{R}^n$, setting $\mathbf{z}(\cdot, \mathbf{d})$ to be the unique solution $\mathbf{y}$ of (5.2) establishes the existence of the trajectory $\{\mathbf{z}(t, \mathbf{d}) : t \in [t_0, t_f]\}$. The existence of a set $N_{\mathbf{g}}(\mathbf{d})$ and functions $k_{\mathbf{g}}$ and $m_{\mathbf{g}}$ satisfying the claimed properties follows from the proof of Theorem 5.2.1 as well, with the identifications $N_{\mathbf{g}}(\mathbf{d}) := \{\mathbf{v} \in \mathbb{R}^n : \|\mathbf{v}\| < m_{\mathbf{y}}\}$, $k_{\mathbf{g}} \equiv k_{\mathbf{f}}$ on $[t_0, t_f]$, and $m_{\mathbf{g}} : t \mapsto k_{\mathbf{f}}(t)\, m_{\mathbf{y}}$.

Now, suppose that the mapping $\mathbf{f}_t \equiv \mathbf{f}(t, \cdot) : X \to \mathbb{R}^n$ is L-smooth for each $t \in [t_0, t_f] \backslash Z_{\mathbf{f}}$. Choose any fixed $t \in [t_0, t_f] \backslash Z_{\mathbf{f}}$. The construction of $\mathbf{g}$ implies that $\mathbf{g}(t, \cdot) \equiv [\mathbf{f}_t]'(\mathbf{x}(t, \mathbf{c}_0); \cdot)$. Since $\mathbf{f}_t$ is L-smooth on X, it follows that $\mathbf{g}(t, \cdot)$ is L-smooth on $\mathbb{R}^n$. Now, choose any fixed $t \in Z_{\mathbf{f}}$. By construction of $\mathbf{g}$, $\mathbf{g}(t, \cdot)$ is the zero function, which is trivially L-smooth. Combining these cases, $\mathbf{g}(t, \cdot)$ is L-smooth on $\mathbb{R}^n$ for each $t \in [t_0, t_f]$. Since this demonstrates *a posteriori* that $\mathbf{g}(t, \cdot)$ is continuous on $\mathbb{R}^n$ for each $t \in [t_0, t_f]$, the set $Z_{\mathbf{g}}$ described in the statement of the corollary may be set to $\emptyset$. $\square$

5.2.2 Propagating lexicographic derivatives

The following corollary extends the results of the previous subsection to describe the higher-order directional derivatives of the solution of a nonsmooth parametric ODE. The subsequent theorem uses this result to express the LD-derivatives of the unique solution of an ODE with a L-smooth right-hand side function as the unique solution of another ODE. Some implications of this result are discussed. As discussed in Chapter 3, lexicographic derivatives are readily obtained from LD-derivatives in which the direction matrix is square and nonsingular.

Corollary 5.2.3. *Given an open, connected set $X \subset \mathbb{R}^n$ and real numbers $t_0 < t_f$, suppose that a function $\mathbf{f} : [t_0, t_f] \times X \to \mathbb{R}^n$ satisfies the conditions of Theorem 5.2.1, and suppose in addition that $\mathbf{f}(t, \cdot)$ is L-smooth on X for each $t \in [t_0, t_f] \backslash Z_{\mathbf{f}}$. Then, for each $t \in [t_0, t_f]$, with the function $\mathbf{x}_t \equiv \mathbf{x}(t, \cdot)$ defined as in the statement of Theorem 5.2.1, $\mathbf{x}_t$ is L-smooth at $\mathbf{c}_0$. Moreover, for each $p \in \mathbb{N}$, each $\mathbf{M} := \begin{bmatrix} \mathbf{m}_{(1)} & \cdots & \mathbf{m}_{(p)} \end{bmatrix} \in \mathbb{R}^{n \times p}$, each $j \in \{0, 1, \ldots, p\}$, and each $\mathbf{d} \in \mathbb{R}^n$, the mapping $t \mapsto [\mathbf{x}_t]^{(j)}_{\mathbf{c}_0, \mathbf{M}}(\mathbf{d})$ is the unique solution on $[t_0, t_f]$ of the ODE:*

$$\frac{d\mathbf{z}}{dt}(t) = \mathbf{h}_{(j)}(t, \mathbf{z}(t)), \qquad \mathbf{z}(t_0) = \mathbf{d}, \tag{5.7}$$

where the functions $\mathbf{h}_{(j)} : [t_0, t_f] \times \mathbb{R}^n \to \mathbb{R}^n$ are defined inductively as follows:

$$\begin{aligned} &\mathbf{h}_{(0)} : (t, \mathbf{v}) \mapsto [\hat{\mathbf{f}}_t]'(\mathbf{x}(t, \mathbf{c}_0); \mathbf{v}), \\ &\mathbf{h}_{(j)} : (t, \mathbf{v}) \mapsto [\mathbf{h}_{(j-1),t}]'([\mathbf{x}_t]^{(j-1)}_{\mathbf{c}_0, \mathbf{M}}(\mathbf{m}_{(j)}); \mathbf{v}), \qquad \forall j \in \{1, \ldots, p\}, \end{aligned}$$

where $\hat{\mathbf{f}}_t : X \to \mathbb{R}^n$ is defined for each $t \in [t_0, t_f]$ as in the statement of Theorem 5.2.1, and where $\mathbf{h}_{(j),t} \equiv \mathbf{h}_{(j)}(t, \cdot)$. Lastly, for each $t \in [t_0, t_f]$ and each $j \in \{0, 1, \ldots, p\}$, let

$$\mathbf{Y}(t, j, \mathbf{c}_0, \mathbf{M}) := \begin{bmatrix} [\mathbf{x}_t]^{(0)}_{\mathbf{c}_0, \mathbf{M}}(\mathbf{m}_{(1)}) & [\mathbf{x}_t]^{(1)}_{\mathbf{c}_0, \mathbf{M}}(\mathbf{m}_{(2)}) & \cdots & [\mathbf{x}_t]^{(j-1)}_{\mathbf{c}_0, \mathbf{M}}(\mathbf{m}_{(j)}) \end{bmatrix} \in \mathbb{R}^{n \times j}.$$

(Thus, $\mathbf{Y}(t, 0, \mathbf{c}_0, \mathbf{M}) = \varnothing_{n \times 0}$.) The functions $\mathbf{h}_{(j)}$ satisfy:

$$\mathbf{h}_{(j)}(t,\mathbf{v}) = [\hat{\mathbf{f}}_t]^{(j)}_{\mathbf{x}(t,\mathbf{c}_0),\mathbf{Y}(t,j,\mathbf{c}_0,\mathbf{M})}(\mathbf{v}), \qquad \forall (t,\mathbf{v}) \in [t_0,t_f] \times \mathbb{R}^n, \quad \forall j \in \{0,1,\ldots,p\}.$$

Proof. Corollary 5.2.2 shows that $\hat{\mathbf{f}}_t$ is L-smooth at $\mathbf{x}(t,\mathbf{c}_0)$ for each $t \in [t_0,t_f]$. Now, choose any fixed $p \in \mathbb{N}$ and $\mathbf{M} \in \mathbb{R}^{n\times p}$. It will be shown by induction on $j \in \{0,1,\ldots,p\}$ that for every such j and every $\mathbf{d} \in \mathbb{R}^n$, the mapping $t \mapsto [\mathbf{x}_t]^{(j)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{d})$ is the unique solution on $[t_0,t_f]$ of the ODE (5.7), that $\mathbf{h}_{(j)}(t,\cdot)$ is L-smooth for each $t \in [t_0,t_f]$, and that $\mathbf{h}_{(j)}$ satisfies the assumptions of Theorem 5.2.1 in place of $\mathbf{f}$, with $Z_{\mathbf{f}} = \emptyset$.

The case in which $j = 0$ follows immediately from Theorem 5.2.1 and Corollary 5.2.2. For the inductive step, suppose that for some $k \in \{1,\ldots,p\}$ and every $\mathbf{d} \in \mathbb{R}^n$, the mapping $t \mapsto [\mathbf{x}_t]^{(k-1)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{d})$ is the unique solution on $[t_0,t_f]$ of (5.7), and that $\mathbf{h}_{(k-1)}$ satisfies the assumptions of Theorem 5.2.1 in place of $\mathbf{f}$. The existence of kth-order directional derivatives of $\mathbf{x}_t$ is not assumed *a priori*. Applying Theorem 5.2.1 with $\mathbf{h}_{(k-1)}$ in place of $\mathbf{f}$, with $\mathbf{m}_{(k)}$ in place of $\mathbf{c}_0$, with the mapping $(t,\mathbf{d}) \mapsto [\mathbf{x}_t]^{(k-1)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{d})$ in place of $\mathbf{x}$, and with $Z_{\mathbf{f}} = \emptyset$, for each $\mathbf{d} \in \mathbb{R}^n$, the mapping $t \mapsto [(\mathbf{x}_t)^{(k-1)}_{\mathbf{c}_0,\mathbf{M}}]'(\mathbf{m}_{(k)};\mathbf{d})$ is the unique solution on $[t_0,t_f]$ of the ODE:

$$\frac{d\mathbf{z}}{dt}(t) = [\mathbf{h}_{(k-1),t}]'([\mathbf{x}_t]^{(k-1)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{m}_{(k)});\mathbf{z}(t)), \qquad \mathbf{z}(t_0) = \mathbf{d}.$$

Applying the definition of $\mathbf{h}_{(k)}$, it follows immediately that $t \mapsto [(\mathbf{x}_t)^{(k-1)}_{\mathbf{c}_0,\mathbf{M}}]'(\mathbf{m}_{(k)};\mathbf{d})$ is the unique solution on $[t_0,t_f]$ of (5.7) with $j := k$. Moreover, Theorem 5.2.1 shows that $[\mathbf{x}_t]^{(k-1)}_{\mathbf{c}_0,\mathbf{M}}$ is directionally differentiable at $\mathbf{m}_{(k)}$ for each $t \in [t_0,t_f]$, implying that $[\mathbf{x}_t]^{(k)}_{\mathbf{c}_0,\mathbf{M}} \equiv [(\mathbf{x}_t)^{(k-1)}_{\mathbf{c}_0,\mathbf{M}}]'(\mathbf{m}_{(k)};\cdot)$. Combining these remarks, for each $\mathbf{d} \in \mathbb{R}^n$, the mapping $t \mapsto [\mathbf{x}_t]^{(k)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{d})$ uniquely solves the ODE (5.7) with $j := k$. To complete the inductive step, Corollary 5.2.2 shows that $\mathbf{h}_{(k)}(t,\cdot)$ is L-smooth for each $t \in [t_0,t_f]$, and that $\mathbf{h}_{(k)}$ satisfies the assumptions of Theorem 5.2.1 in place of $\mathbf{f}$, with $Z_{\mathbf{f}} = \emptyset$.

Since p and $\mathbf{M}$ were arbitrary in the above inductive argument, this argument shows that $\mathbf{x}_t$ is L-smooth at $\mathbf{c}_0$ for each $t \in [t_0,t_f]$, as required.

Next, a simpler inductive proof shows that $\mathbf{h}_{(j)}(t,\cdot) \equiv [\hat{\mathbf{f}}_t]^{(j)}_{\mathbf{x}(t,\mathbf{c}_0),\mathbf{Y}(t,j,\mathbf{c}_0,\mathbf{M})}$ for

each $t \in [t_0, t_f]$ and each $j \in \{0, 1, \ldots, p\}$, as follows. For the base case, the definition of $\mathbf{h}_{(0)}$ implies that for each $t \in [t_0, t_f]$,

$$\mathbf{h}_{(0)}(t, \mathbf{v}) = [\hat{\mathbf{f}}_t]'(\mathbf{x}(t, \mathbf{c}_0); \mathbf{v}) = [\hat{\mathbf{f}}_t]^{(0)}_{\mathbf{x}(t,\mathbf{c}_0),\mathbf{Y}(t,0,\mathbf{c}_0,\mathbf{M})}(\mathbf{v}), \qquad \forall \mathbf{v} \in \mathbb{R}^n,$$

as required. For the inductive step, suppose that for some $k \in \{1, \ldots, p\}$,

$$\mathbf{h}_{(k-1),t} \equiv \mathbf{h}_{(k-1)}(t, \cdot) \equiv [\hat{\mathbf{f}}_t]^{(k-1)}_{\mathbf{x}(t,\mathbf{c}_0),\mathbf{Y}(t,k-1,\mathbf{c}_0,\mathbf{M})}, \qquad \forall t \in [t_0, t_f].$$

The constructive definition of $\mathbf{h}_{(k)}$, the inductive assumption, and the definitions of $\mathbf{Y}(t, k-1, \mathbf{c}_0, \mathbf{M})$ and $\mathbf{Y}(t, k, \mathbf{c}_0, \mathbf{M})$ imply that, for each $t \in [t_0, t_f]$,

$$\mathbf{h}_{(k)}(t, \cdot) \equiv [(\hat{\mathbf{f}}_t)^{(k-1)}_{\mathbf{x}(t,\mathbf{c}_0),\mathbf{Y}(t,k-1,\mathbf{c}_0,\mathbf{M})}]'([\mathbf{x}_t]^{(k-1)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{m}_{(k)}); \cdot) \equiv [\hat{\mathbf{f}}_t]^{(k)}_{\mathbf{x}(t,\mathbf{c}_0),\mathbf{Y}(t,k,\mathbf{c}_0,\mathbf{M})}.$$

This completes the inductive step. □

Using the notation of Corollary 5.2.3, if $\mathbf{e}_{(1)}, \ldots, \mathbf{e}_{(n)}$ denote the coordinate vectors in $\mathbb{R}^n$, then for any nonsingular $\mathbf{M} \in \mathbb{R}^{n \times n}$ and any $t \in [t_0, t_f]$,

$$\mathbf{J}_{\mathrm{L}}\mathbf{x}_t(\mathbf{c}_0; \mathbf{M}) = \left[[\mathbf{x}_t]^{(n)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{e}_{(1)}) \quad \cdots \quad [\mathbf{x}_t]^{(n)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{e}_{(n)}) \right].$$

Thus, Corollary 5.2.3 provides a method for evaluating lexicographic derivatives of $\mathbf{x}(t, \cdot)$. Without further assumptions, though, this method is computationally expensive in the worst case, as it involves construction and evaluation of the ODE right-hand side function $(t, \mathbf{v}) \mapsto \mathbf{h}_{(j)}(t, \mathbf{v}) = [\hat{\mathbf{f}}_t]^{(j)}_{\mathbf{x}(t,\mathbf{c}_0),\mathbf{Y}(t,j,\mathbf{c}_0,\mathbf{M})}(\mathbf{v})$ for each $j \in \{0, 1, \ldots, n\}$. If the forward mode of automatic differentiation is used to construct these mappings using the identity

$$[\hat{\mathbf{f}}_t]^{(j)}_{\mathbf{x}(t,\mathbf{c}_0),\mathbf{Y}(t,j,\mathbf{c}_0,\mathbf{M})} \equiv [(\hat{\mathbf{f}}_t)^{(j-1)}_{\mathbf{x}(t,\mathbf{c}_0),\mathbf{Y}(t,j-1,\mathbf{c}_0,\mathbf{M})}]'(\mathbf{m}_{(j)}; \cdot),$$

then the overall cost of this construction scales worst-case exponentially with j, relative to the cost of evaluating $\mathbf{f}$. To avoid this computational burden, the following theorem expresses LD-derivatives $[\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})$, and thus lexicographic derivatives $\mathbf{J}_{\mathrm{L}}\mathbf{x}_t(\mathbf{c}_0; \mathbf{M})$, in terms of the unique solution of an ODE, without requiring explicit construction of the intermediate directional derivatives $[\hat{\mathbf{f}}_t]^{(j)}_{\mathbf{x}(t,\mathbf{c}_0),\mathbf{Y}(t,j,\mathbf{c}_0,\mathbf{M})}$.

Theorem 5.2.4. *Given an open, connected set $X \subset \mathbb{R}^n$ and real numbers $t_0 < t_f$, suppose that a function $\mathbf{f} : [t_0, t_f] \times X \to \mathbb{R}^n$ satisfies the following conditions:*

- *the mapping $\mathbf{f}(\cdot, \mathbf{c}) : [t_0, t_f] \to \mathbb{R}^n$ is measurable for each $\mathbf{c} \in X$,*
- *for each $t \in [t_0, t_f]$ except in a zero-measure subset $Z_{\mathbf{f}}$, the mapping $\mathbf{f}(t, \cdot) : X \to \mathbb{R}^n$ is L-smooth,*
- *with $\mathbf{x}(\cdot, \mathbf{c})$ denoting any solution of the parametric ODE system:*

$$\frac{d\mathbf{x}}{dt}(t, \mathbf{c}) = \mathbf{f}(t, \mathbf{x}(t, \mathbf{c})), \qquad \mathbf{x}(t_0, \mathbf{c}) = \mathbf{c},$$

  there exists a solution $\{\mathbf{x}(t, \mathbf{c}_0) : t \in [t_0, t_f]\} \subset X$ for some $\mathbf{c}_0 \in X$,
- *there exists an open set $N \subset X$ such that $\{\mathbf{x}(t, \mathbf{c}_0) : t \in [t_0, t_f]\} \subset N$, and such that there exist Lebesgue integrable functions $k_{\mathbf{f}}, m_{\mathbf{f}} : [t_0, t_f] \to \mathbb{R}_+ \cup \{+\infty\}$ for which*

$$\|\mathbf{f}(t, \mathbf{c})\| \leq m_{\mathbf{f}}(t), \qquad \forall t \in [t_0, t_f], \quad \forall \mathbf{c} \in N,$$

  and

$$\|\mathbf{f}(t, \mathbf{c}_1) - \mathbf{f}(t, \mathbf{c}_2)\| \leq k_{\mathbf{f}}(t) \, \|\mathbf{c}_1 - \mathbf{c}_2\|, \qquad \forall t \in [t_0, t_f], \quad \forall \mathbf{c}_1, \mathbf{c}_2 \in N.$$

Then, for each $t \in [t_0, t_f]$, the function $\mathbf{x}_t \equiv \mathbf{x}(t, \cdot)$ is well-defined and Lipschitz continuous on a neighborhood of $\mathbf{c}_0$, with a Lipschitz constant that is independent of t. Moreover, $\mathbf{x}_t$ is L-smooth at $\mathbf{c}_0$; for any $p \in \mathbb{N}$ and any $\mathbf{M} \in \mathbb{R}^{n \times p}$, the LD-derivative mapping $t \mapsto [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})$ is the unique solution on $[t_0, t_f]$ of the following ODE:

$$\frac{d\mathbf{A}}{dt}(t) = [\hat{\mathbf{f}}_t]'(\mathbf{x}(t, \mathbf{c}_0); \mathbf{A}(t)), \qquad \mathbf{A}(t_0) = \mathbf{M}, \tag{5.8}$$

where $\hat{\mathbf{f}}_t : X \to \mathbb{R}^n$ is defined in terms of $\mathbf{f}$ as follows, and is L-smooth by construction for each $t \in [t_0, t_f]$:

$$\hat{\mathbf{f}}_t(\mathbf{c}) = \begin{cases} \mathbf{f}(t,\mathbf{c}), & \text{if } t \in [t_0, t_f] \backslash Z_{\mathbf{f}}, \\ \mathbf{0}, & \text{if } t \in Z_{\mathbf{f}}. \end{cases}$$

Proof. For each $t \in [t_0, t_f]$, the L-smoothness of $\mathbf{x}_t$ at $\mathbf{c}_0$ was established in Corollary 5.2.3. Moreover, it was established in the proof of Theorem 5.2.1 that $\mathbf{x}_t$ is Lipschitz continuous on a neighborhood of $\mathbf{c}_0$ for each $t \in [t_0, t_f]$, with a Lipschitz constant that is independent of t.

Now, consider any fixed $p \in \mathbb{N}$ and $\mathbf{M} := \begin{bmatrix} \mathbf{m}_{(1)} & \cdots & \mathbf{m}_{(p)} \end{bmatrix} \in \mathbb{R}^{n \times p}$. As an intermediate result, it will be shown by induction that for each $j \in \{1, \ldots, p\}$, the coupled ODE system:

$$\frac{d\mathbf{z}_{(i)}}{dt}(t) = [\hat{\mathbf{f}}_t]^{(i-1)}_{\mathbf{x}(t,\mathbf{c}_0),[\mathbf{z}_{(1)}(t)\ \ \mathbf{z}_{(2)}(t)\ \cdots\ \mathbf{z}_{(i-1)}(t)]}(\mathbf{z}_{(i)}(t)), \ \ \mathbf{z}_{(i)}(t_0) = \mathbf{m}_{(i)}, \ \ \forall i \in \{1, \ldots, j\} \tag{5.9}$$

has a unique solution on $[t_0, t_f]$, in which $\mathbf{z}_{(i)}(t) = [\mathbf{x}_t]^{(i-1)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{m}_{(i)})$ for each $t \in [t_0, t_f]$ and each $i \in \{1, \ldots, j\}$. (Note that the right-hand sides of the coupled ODEs above are all well-defined, since Corollary 5.2.2 established the L-smoothness of $\hat{\mathbf{f}}_t$ at $\mathbf{x}(t, \mathbf{c}_0)$ for each $t \in [t_0, t_f]$).

The case in which $j = 1$ follows immediately from Corollary 5.2.3. For the inductive step, suppose that for some $k \in \{2, 3, \ldots, p\}$, the coupled ODE system (5.9) has a unique solution on $[t_0, t_f]$ when $j := k - 1$, in which $\mathbf{z}_{(i)}(t) = [\mathbf{x}_t]^{(i-1)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{m}_{(i)})$ for each $t \in [t_0, t_f]$ and each $i \in \{1, \ldots, k-1\}$. Now, consider the case in which $j := k$. In this case, the ODEs in (5.9) with $i \in \{1, \ldots, k-1\}$ are unchanged from the case in which $j = k - 1$. Thus, by the inductive assumption, the ODEs in (5.9) with $i \in \{1, \ldots, k-1\}$ have unique solutions on $[t_0, t_f]$ in which $\mathbf{z}_{(i)}(t) = [\mathbf{x}_t]^{(i-1)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{m}_{(i)})$ for each $t \in [t_0, t_f]$. As a result, the ODE in (5.9) with $i = k$ becomes:

$$\frac{d\mathbf{z}_{(k)}}{dt}(t) = [\hat{\mathbf{f}}_t]^{(k-1)}_{\mathbf{x}(t,\mathbf{c}_0),\mathbf{Y}(t,k-1,\mathbf{c}_0,\mathbf{M})}(\mathbf{z}_{(k)}(t)), \qquad \mathbf{z}_{(k)}(t_0) = \mathbf{m}_{(k)}, \tag{5.10}$$

with $\mathbf{Y}(t, k-1, \mathbf{c}_0, \mathbf{M})$ defined as in the statement of Corollary 5.2.3. This corollary shows that (5.10) is uniquely solved on $[t_0, t_f]$ by the mapping $\mathbf{z}_{(k)} : t \mapsto [\mathbf{x}_t]^{(k-1)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{m}_{(k)})$. Combining this statement with the inductive assumption com-

pletes the inductive step.

Using this inductive result, the coupled ODE system:

$$\frac{d\mathbf{z}_{(i)}}{dt}(t) = [\hat{\mathbf{f}}_t]^{(i-1)}_{\mathbf{x}(t,\mathbf{c}_0),[\mathbf{z}_{(1)}(t)\ \ \mathbf{z}_{(2)}(t)\ \cdots\ \mathbf{z}_{(i-1)}(t)]}(\mathbf{z}_{(i)}(t)),\ \ \mathbf{z}_{(i)}(t_0) = \mathbf{m}_{(i)},\ \ \forall i \in \{1,\ldots,p\} \tag{5.11}$$

has a unique solution on $[t_0, t_f]$, in which $\mathbf{z}_{(i)}(t) = [\mathbf{x}_t]^{(i-1)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{m}_{(i)})$ for each $i \in \{1,\ldots,p\}$. Using the definition of the LD-derivative, it follows that for each $i \in \{1,\ldots,p\}$, each $t \in [t_0, t_f]$, and each choice of $\mathbf{v}_{(1)},\ldots,\mathbf{v}_{(p)} \in \mathbb{R}^n$,

$$[\hat{\mathbf{f}}_t]^{(i-1)}_{\mathbf{x}(t,\mathbf{c}_0),[\mathbf{v}_{(1)}\ \cdots\ \mathbf{v}_{(i-1)}]}(\mathbf{v}_{(i)}) = [\hat{\mathbf{f}}_t]'(\mathbf{x}(t,\mathbf{c}_0); \begin{bmatrix}\mathbf{v}_{(1)} & \cdots & \mathbf{v}_{(p)}\end{bmatrix})\,\mathbf{e}_{(i)}.$$

Thus, the following coupled ODE system is equivalent to (5.11):

$$\begin{cases} \dfrac{d\mathbf{z}_{(i)}}{dt}(t) &= [\hat{\mathbf{f}}_t]'(\mathbf{x}(t,\mathbf{c}_0); \begin{bmatrix}\mathbf{z}_{(1)}(t) & \cdots & \mathbf{z}_{(p)}(t)\end{bmatrix})\,\mathbf{e}_{(i)}, \qquad \forall i \in \{1,\ldots,p\}, \\ \mathbf{z}_{(i)}(t_0) &= \mathbf{m}_{(i)}, \end{cases} \tag{5.12}$$

and therefore has the same unique solution on $[t_0, t_f]$ as (5.11). Moreover, Property 4 in Lemma 2.3.7 and the definition of the LD-derivative imply that

$$[\mathbf{x}_t]^{(i-1)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{m}_{(i)}) = [\mathbf{x}_t]'(\mathbf{c}_0;\mathbf{M})\,\mathbf{e}_{(i)}, \qquad \forall t \in [t_0, t_f], \qquad \forall i \in \{1,\ldots,p\}.$$

Thus, the unique solution of (5.12) on $[t_0, t_f]$ satisfies $\mathbf{z}_{(i)}(t) = [\mathbf{x}_t]'(\mathbf{c}_0;\mathbf{M})\,\mathbf{e}_{(i)}$ for each $i \in \{1,\ldots,p\}$ and each $t \in [t_0, t_f]$. The coupled ODEs (5.12) may be written as the columns of a single ODE with the matrix-valued dependent variable $\mathbf{A} := \begin{bmatrix}\mathbf{z}_{(1)} & \cdots & \mathbf{z}_{(p)}\end{bmatrix}$ to yield the ODE (5.8), which therefore has the unique solution:

$$t \mapsto \begin{bmatrix}[\mathbf{x}_t]'(\mathbf{c}_0;\mathbf{M})\,\mathbf{e}_{(1)} & \cdots & [\mathbf{x}_t]'(\mathbf{c}_0;\mathbf{M})\,\mathbf{e}_{(p)}\end{bmatrix} = [\mathbf{x}_t]'(\mathbf{c}_0;\mathbf{M})$$

on $[t_0, t_f]$. □

Corollary 3.2.5 and Theorem 5.2.4 together show that plenary Jacobian elements can be obtained for solutions of a nonsmooth parametric ODE, provided

that lexicographic derivatives can be evaluated for the ODE right-hand side function, and provided that the unique solution of the ODE (5.8) can be determined or approximated numerically. This implies the following corollaries, which make use of *a priori* knowledge concerning the differentiability or convexity of the solution to a nonsmooth parametric ODE. These results do not require differentiability or convexity assumptions on the ODE right-hand side function.

Corollary 5.2.5. *Suppose that the hypotheses of Theorem 5.2.4 hold, and let* $\mathbf{x}_t \equiv \mathbf{x}(t, \cdot)$. *If* $\mathbf{x}_{t_f}$ *is known to be differentiable at* $\mathbf{c}_0$*, then the ODE:*

$$\frac{d\mathbf{A}}{dt}(t) = [\hat{\mathbf{f}}_t]'(\mathbf{x}(t, \mathbf{c}_0); \mathbf{A}(t)), \qquad \mathbf{A}(t_0) = \mathbf{I} \tag{5.13}$$

has a unique solution $\mathbf{A}$ *on* $[t_0, t_f]$*, which satisfies* $\mathbf{A}(t_f) = \mathbf{J}\mathbf{x}_{t_f}(\mathbf{c}_0)$.

Proof. By Theorem 5.2.4, the mapping $\mathbf{A} : t \mapsto \mathbf{J}_{\mathrm{L}}\mathbf{x}_t(\mathbf{c}_0; \mathbf{I})$ is the unique solution on $[t_0, t_f]$ of (5.13). Since $\mathbf{x}_{t_f}$ is differentiable at $\mathbf{c}_0$, it follows from [79] that

$$\mathbf{J}_{\mathrm{L}}\mathbf{x}_{t_f}(\mathbf{c}_0; \mathbf{I}) \in \partial_{\mathrm{L}}\mathbf{x}_{t_f}(\mathbf{c}_0) = \{\mathbf{J}\mathbf{x}_{t_f}(\mathbf{c}_0)\}.$$

Thus, $\mathbf{A}(t_f) = [\mathbf{x}_{t_f}]'(\mathbf{c}_0; \mathbf{I}) = \mathbf{J}_{\mathrm{L}}\mathbf{x}_{t_f}(\mathbf{c}_0; \mathbf{I}) = \mathbf{J}\mathbf{x}_{t_f}(\mathbf{c}_0)$. □

Now, for any function $\mathbf{g} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ that is piecewise differentiable in the sense of Scholtes [97], $\partial_{\mathrm{L}}\mathbf{g}(\mathbf{x}) \subset \partial\mathbf{g}(\mathbf{x})$ for each $\mathbf{x} \in X$ [61]. It follows that if the ODE right-hand side function $(t, \mathbf{c}) \mapsto \mathbf{f}(t, \mathbf{c})$ is piecewise differentiable with respect to $\mathbf{c}$ for almost all $t \in [t_0, t_f]$, then the solution to (5.8) is also an element of the linear Newton approximation to $\mathbf{x}(t, \cdot)$ at $\mathbf{c}_0$ described in [88, Corollary 12], right-multiplied by $\mathbf{M}$.

While the ODE (5.8) has a unique solution, the following example shows that its right-hand side function, $(t, \mathbf{A}) \mapsto [\hat{\mathbf{f}}_t]'(\mathbf{x}(t, \mathbf{c}_0); \mathbf{A})$, is not necessarily continuous with respect to $\mathbf{A}$ at almost every fixed $t \in [t_0, t_f]$. Thus, (5.8) is not necessarily a Carathéodory ODE. As the proof of Theorem 5.2.4 suggests, however, the columns of (5.8) can be decoupled to yield a sequence of Carathéodory ODEs, each with a unique solution.

Example 5.2.6. *Consider the following parametric ODE system with two differential variables:*

$$\frac{dx_1}{dt}(t,\mathbf{p}) = \frac{dx_2}{dt}(t,\mathbf{p}) = \max\{x_1(t,\mathbf{p}), x_2(t,\mathbf{p})\}, \qquad \mathbf{x}(0,\mathbf{p}) = \mathbf{p}.$$

This ODE system satisfies the Carathéodory existence and uniqueness conditions when $\mathbf{x}(t,\mathbf{p})$ *is restricted to any bounded neighborhood of* $\mathbf{p}$*; when* $\mathbf{p} = (0,0)$*, the unique solution is* $\mathbf{x}(t,\mathbf{0}) := (x_1(t,\mathbf{0}), x_2(t,\mathbf{0})) = \mathbf{0}$ *for each* $t \in \mathbb{R}$*. Now, with*

$$\mathbf{A} := \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \in \mathbb{R}^{2\times 2}, \qquad \textit{and} \quad \mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^2 : \mathbf{c} \mapsto (\max\{c_1, c_2\}, \max\{c_1, c_2\}),$$

it follows that $\mathbf{f}$ *is the composition of continuously differentiable functions and the function* $\mathbf{c} \mapsto \max\{c_1, c_2\}$*, and is therefore L-smooth. Since* $\mathbf{f}$ *is not an explicit function of* t*, it follows that* $\mathbf{f}$ *itself plays the role of* $\hat{\mathbf{f}}_t$ *in Theorems 5.2.1 and 5.2.4. By inspection, for any* $\mathbf{d} \in \mathbb{R}^2$ *and any* $t \in \mathbb{R}$*,*

$$\begin{aligned}
\mathbf{f}^{(0)}_{\mathbf{x}(t,\mathbf{0}),\mathbf{A}}(\mathbf{d}) &= \begin{cases} (d_1, d_1), & \textit{if } d_1 \geq d_2, \\ (d_2, d_2), & \textit{if } d_1 < d_2; \end{cases} \\
\mathbf{f}^{(1)}_{\mathbf{x}(t,\mathbf{0}),\mathbf{A}}(\mathbf{d}) &= [\mathbf{f}^{(0)}_{\mathbf{x}(t,\mathbf{0}),\mathbf{A}}]'((a_{11}, a_{21}); \mathbf{d}), \\
&= \begin{cases} (d_1, d_1), & \textit{if } a_{11} > a_{21}, \textit{ or if } a_{11} = a_{21} \textit{ and } d_1 \geq d_2, \\ (d_2, d_2), & \textit{if } a_{11} < a_{21}, \textit{ or if } a_{11} = a_{21} \textit{ and } d_1 < d_2. \end{cases}
\end{aligned}$$

Using Lemma 2.3.7, it follows that:

$$\begin{aligned}
\mathbf{f}'(\mathbf{x}(t,\mathbf{0}); \mathbf{A}) &= \begin{bmatrix} \mathbf{f}^{(2)}_{\mathbf{x}(t,\mathbf{0}),\mathbf{A}}(a_{11}, a_{21}) & \mathbf{f}^{(2)}_{\mathbf{x}(t,\mathbf{0}),\mathbf{A}}(a_{12}, a_{22}) \end{bmatrix}, \\
&= \begin{bmatrix} \mathbf{f}^{(0)}_{\mathbf{x}(t,\mathbf{0}),\mathbf{A}}(a_{11}, a_{21}) & \mathbf{f}^{(1)}_{\mathbf{x}(t,\mathbf{0}),\mathbf{A}}(a_{12}, a_{22}) \end{bmatrix}, \\
&= \begin{cases} \begin{bmatrix} a_{11} & a_{12} \\ a_{11} & a_{12} \end{bmatrix}, & \textit{if } a_{11} > a_{21}, \textit{ or if } a_{11} = a_{21} \textit{ and } a_{12} \geq a_{22}, \\ \begin{bmatrix} a_{21} & a_{22} \\ a_{21} & a_{22} \end{bmatrix}, & \textit{if } a_{11} < a_{21}, \textit{ or if } a_{11} = a_{21} \textit{ and } a_{12} < a_{22}. \end{cases}
\end{aligned}$$

It follows that for any $t \in \mathbb{R}$*, the mapping* $\mathbf{A} \mapsto \mathbf{f}'(\mathbf{x}(t,\mathbf{0}); \mathbf{A})$ *is discontinuous at any* $\mathbf{A} \in \mathbb{R}^{2\times 2}$ *for which both* $a_{11} = a_{21}$ *and* $a_{12} \neq a_{22}$*.*

The following example presents a straightforward application of Theorem 5.2.4, in which the relevant ODE systems can all be solved analytically.

Example 5.2.7. *Consider the function:*

$$\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^2 : \mathbf{y} \mapsto \begin{bmatrix} (1 - y_2)|y_1| \\ 1 \end{bmatrix},$$

and the following nonsmooth parametric ODE system with two differential variables, in which $\mathbf{c} := (c_1, c_2) \in \mathbb{R}^2$ *denotes a parameter:*

$$\frac{d\mathbf{x}}{dt}(t, \mathbf{c}) = \mathbf{f}(\mathbf{x}(t, \mathbf{c})), \qquad \mathbf{x}(0, \mathbf{c}) = \mathbf{c}.$$

It is readily verified that this ODE system is uniquely solved by the mapping:

$$\mathbf{x} : \mathbb{R} \times \mathbb{R}^2 \to \mathbb{R}^2 : (t, \mathbf{c}) \mapsto \begin{cases} \begin{bmatrix} c_1 \exp\left(-\frac{1}{2}t^2 + (1 - c_2)t\right) \\ c_2 + t \end{bmatrix}, & \text{if } c_1 \geq 0, \\ \begin{bmatrix} c_1 \exp\left(\frac{1}{2}t^2 + (c_2 - 1)t\right) \\ c_2 + t \end{bmatrix}, & \text{if } c_1 < 0. \end{cases} \tag{5.14}$$

Thus, $\mathbf{x}(t, \mathbf{0}) = (0, t)$ *for each* $t \in \mathbb{R}$*. The mapping* $t \mapsto x_1(t, (c_1, 0))$ *is plotted in Figure 5-1(a) for various values of* $c_1 \in [-2, 2]$*.*

B-subdifferentials of the parametric ODE solution can be evaluated analytically in this case, as follows. For each fixed $t \in \mathbb{R}$*, the mapping* $\mathbf{x}_t \equiv \mathbf{x}(t, \cdot)$ *is evidently differentiable at all domain points* $\mathbf{c}$ *for which* $c_1 \neq 0$*. The definition of the B-subdifferential can thus be used to show that, when* $\mathbf{c} = \mathbf{0}$*,*

$$\partial_{\mathrm{B}} \mathbf{x}_t(\mathbf{0}) = \left\{ \begin{bmatrix} \exp\left(-\frac{1}{2}t^2 + t\right) & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} \exp\left(\frac{1}{2}t^2 - t\right) & 0 \\ 0 & 1 \end{bmatrix} \right\}.$$

Thus, Definition 2.3.4 can be used to show that

$$\partial_{\mathrm{P}} \mathbf{x}_t(\mathbf{0}) = \partial \mathbf{x}_t(\mathbf{0}) = \operatorname{conv} \left\{ \begin{bmatrix} \exp\left(-\frac{1}{2}t^2 + t\right) & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} \exp\left(\frac{1}{2}t^2 - t\right) & 0 \\ 0 & 1 \end{bmatrix} \right\}.$$

Elements of the linear Newton approximation of $\mathbf{x}_t$ *described in [88, Corollary 12] can be evaluated as follows. The function* $\mathbf{f}$ *is evidently differentiable at all domain points* $\mathbf{y}$

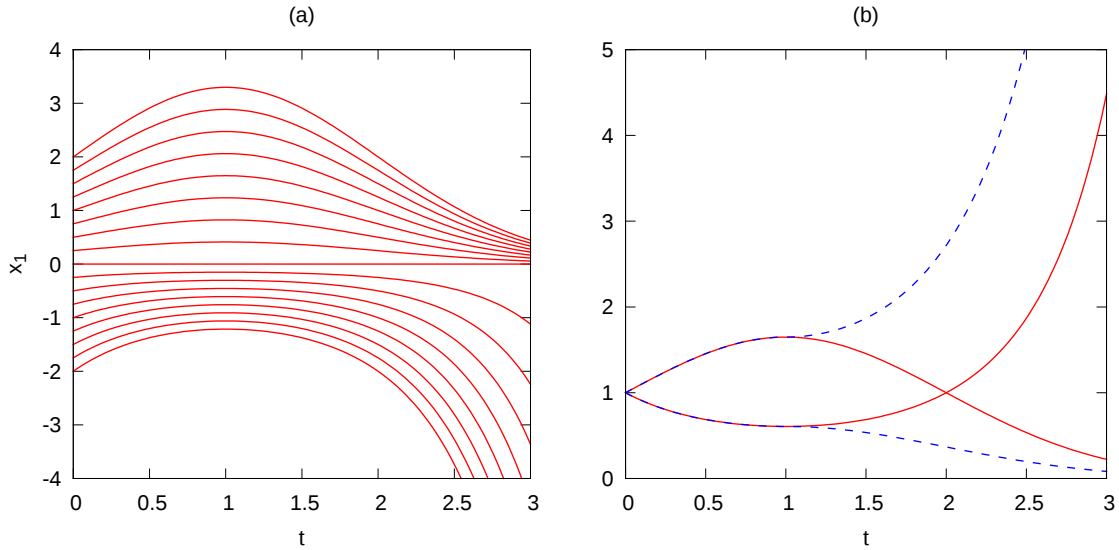

Figure 5-1: ODE solutions and sensitivities for Example 5.2.7: (a) $x_1(t,(c_1,0))$ vs. t for various values of $c_1 \in [-2,2]$, and (b) the (1,1)-entries of two elements of the linear Newton approximation $\Gamma\mathbf{x}_t(\mathbf{0})$ vs. t (dashed blue), and the set-valued (1,1)-entry of $\partial_{\mathrm{L}}\mathbf{x}_t(\mathbf{0})$ vs. t (solid red).

for which $y_1 \neq 0$. Thus, for each $t \in \mathbb{R}$, Clarke's generalized Jacobian of $\mathbf{f}$ is evaluated at $\mathbf{x}(t,\mathbf{0}) = (0,t)$ to be:

$$\partial\mathbf{f}(\mathbf{x}(t,\mathbf{0})) = \left\{\lambda\,(1-t)\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix} : \lambda \in [-1,1]\right\}, \qquad \forall t \in \mathbb{R}.$$

Now, define the mapping:

$$h : \mathbb{R}\times[-1,1] \to [-1,1] : (t,\mu) \mapsto \begin{cases} \mu, & \text{if } t \leq 1,\\ -\mu, & \text{if } t > 1.\end{cases}$$

The above results show that the linear Newton approximation of $\mathbf{x}_t$ at $\mathbf{0}$ described in [88, Corollary 12] includes the solutions of the following ODE for all $\mu \in [-1,1]$:

$$\frac{d\mathbf{A}}{dt}(t,\mu) = h(t,\mu)\,(1-t)\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}\mathbf{A}(t,\mu), \qquad \mathbf{A}(0,\mu) = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}.$$

This ODE is readily solved to yield:

$$\mathbf{A}(t,\mu) = \begin{cases} \begin{bmatrix} \exp(\mu(-\frac{1}{2}t^2+t)) & 0 \\ 0 & 1 \end{bmatrix}, & \text{if } t \leq 1, \\ \begin{bmatrix} \exp(\mu(\frac{1}{2}t^2-t+1)) & 0 \\ 0 & 1 \end{bmatrix}, & \text{if } t > 1. \end{cases}$$

Thus, for each $t > 1$, the linear Newton approximation $\Gamma\mathbf{x}_t(\mathbf{0})$ of $\mathbf{x}_t$ at $\mathbf{0}$ described in [88, Corollary 12] is such that

$$\operatorname{conv}\left\{ \begin{bmatrix} \exp(\frac{1}{2}t^2-t+1) & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} \exp(-\frac{1}{2}t^2+t-1) & 0 \\ 0 & 1 \end{bmatrix} \right\} \subset \Gamma\mathbf{x}_t(\mathbf{0}).$$

The $(1,1)$-entries of the linear Newton approximation elements on which the above convex hull is contructed are plotted in Figure 5-1(b).

Lexicographic derivatives of the parametric ODE solution can be evaluated using Theorem 5.2.4 as follows. Following a similar approach to Example 5.2.6, the following is obtained for each $\mathbf{A} \in \mathbb{R}^{2\times 2}$. Here, a_{ij} denotes the (i,j)–element of $\mathbf{A}$.

$$\mathbf{f}'(\mathbf{x}(t,\mathbf{0});\mathbf{A}) = \begin{cases} \begin{bmatrix} (1-t)a_{11} & (1-t)a_{12} \\ 0 & 0 \end{bmatrix}, & \text{if } a_{11} > 0, \text{ or if } a_{11} = 0 \text{ and } a_{12} \geq 0, \\ \begin{bmatrix} (t-1)a_{11} & (t-1)a_{12} \\ 0 & 0 \end{bmatrix}, & \text{if } a_{11} < 0, \text{ or if } a_{11} = 0 \text{ and } a_{12} < 0. \end{cases}$$

Thus, for any nonsingular $\mathbf{M} \in \mathbb{R}^{2\times 2}$, Theorem 5.2.4 shows that the mapping $t \mapsto [\mathbf{x}_t]'(\mathbf{0};\mathbf{M}) = \mathbf{J}_{\mathrm{L}}\mathbf{x}_t(\mathbf{c}_0;\mathbf{M})\,\mathbf{M}$ is the unique solution of the ODE:

$$\frac{d\mathbf{A}}{dt}(t) = \begin{cases} (1-t)\begin{bmatrix} a_{11}(t) & a_{12}(t) \\ 0 & 0 \end{bmatrix}, & \text{if } a_{11}(t) > 0, \text{ or if } a_{11}(t) = 0 \text{ and } a_{12}(t) \geq 0, \\ (t-1)\begin{bmatrix} a_{11}(t) & a_{12}(t) \\ 0 & 0 \end{bmatrix}, & \text{if } a_{11}(t) < 0, \text{ or if } a_{11}(t) = 0 \text{ and } a_{12}(t) < 0, \end{cases}$$
$$\mathbf{A}(0) = \mathbf{M}.$$

This ODE can be solved by inspection; post-multiplying the result by $\mathbf{M}^{-1}$ yields:

$$\mathbf{J}_{\mathrm{L}}\mathbf{x}_t(\mathbf{0};\mathbf{M}) = \begin{cases} \begin{bmatrix} \exp\left(-\frac{1}{2}t^2+t\right) & 0 \\ 0 & 1 \end{bmatrix}, & \textit{if } m_{11}>0, \textit{ or if } m_{11}=0 \textit{ and } m_{12}\geq 0, \\ \begin{bmatrix} \exp\left(\frac{1}{2}t^2-t\right) & 0 \\ 0 & 1 \end{bmatrix}, & \textit{if } m_{11}<0, \textit{ or if } m_{11}=0 \textit{ and } m_{12}<0, \end{cases}$$

and so

$$\partial_{\mathrm{L}}\mathbf{x}_t(\mathbf{0}) = \left\{ \begin{bmatrix} \exp\left(-\frac{1}{2}t^2+t\right) & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} \exp\left(\frac{1}{2}t^2-t\right) & 0 \\ 0 & 1 \end{bmatrix} \right\}.$$

The $(1,1)$*-entries of these lexicographic derivatives are plotted in Figure 5-1(b). This result is readily confirmed by lexicographic differentiation of* (5.14) *with respect to* $\mathbf{c}$ *at* $\mathbf{c}=\mathbf{0}$.

Collecting the above results, and noting that, for each $t>1$,

$$\begin{aligned} &\exp\left(-\tfrac{1}{2}t^2+t-1\right) < \min\{\exp\left(-\tfrac{1}{2}t^2+t\right), \exp\left(\tfrac{1}{2}t^2-t\right)\}, \\ \textit{and} \qquad &\exp\left(\tfrac{1}{2}t^2-t+1\right) > \max\{\exp\left(-\tfrac{1}{2}t^2+t\right), \exp\left(\tfrac{1}{2}t^2-t\right)\}, \end{aligned}$$

it follows that, for this example,

$$\partial_{\mathrm{L}}\mathbf{x}_t(\mathbf{0}) = \partial_{\mathrm{B}}\mathbf{x}_t(\mathbf{0}) \subset \partial\mathbf{x}_t(\mathbf{0}) = \partial_{\mathrm{P}}\mathbf{x}_t(\mathbf{0}) \subset \Gamma\mathbf{x}_t(\mathbf{0}), \qquad \forall t>1.$$

The rightmost inclusion above is strict. In particular, when $t=2$*, the evaluated generalized derivatives satisfy:*

$$\partial_{\mathrm{L}}\mathbf{x}_2(\mathbf{0}) = \partial_{\mathrm{P}}\mathbf{x}_2(\mathbf{0}) = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right\} \subset \left\{ \begin{bmatrix} \lambda & 0 \\ 0 & 1 \end{bmatrix} : \lambda \in [\tfrac{1}{\mathrm{e}}, \mathrm{e}] \right\} \subset \Gamma\mathbf{x}_2(\mathbf{0}).$$

Although $\mathbf{x}_2$ *is* strictly differentiable *at* $\mathbf{0}$ *in the sense of [16],* $\Gamma\mathbf{x}_2(\mathbf{0})$ *evidently contains elements other than* $\mathbf{J}\mathbf{x}_2(\mathbf{0})$.

The result of Theorem 5.2.4 is easily extended to cover ODEs whose initial conditions are nontrivial functions of parameters $\mathbf{p}\in\mathbb{R}^{n_p}$:

$$\frac{d\mathbf{x}}{dt}(t,\mathbf{p}) = \mathbf{f}(t,\mathbf{x}(t,\mathbf{p})), \qquad \mathbf{x}(t_0,\mathbf{p}) = \mathbf{f}_0(\mathbf{p}). \tag{5.15}$$

provided that $\mathbf{f}$ satisfies the hypotheses of Theorem 5.2.4 (with $\mathbf{f}_0(\mathbf{p}_0)$ in place of

$\mathbf{c}_0$ for some $\mathbf{p}_0 \in \mathbb{R}^{n_p}$), and provided that $\mathbf{f}_0 : \mathbb{R}^{n_p} \to \mathbb{R}^n$ is L-smooth at $\mathbf{p}_0$. Introducing the auxiliary parametrized ODE:

$$\frac{d\mathbf{z}}{dt}(t, \mathbf{c}) = \mathbf{f}(t, \mathbf{z}(t, \mathbf{c})), \qquad \mathbf{z}(t_0, \mathbf{c}) = \mathbf{c},$$

and defining $\mathbf{x}_t \equiv \mathbf{x}(t, \cdot)$ and $\mathbf{z}_t \equiv \mathbf{z}(t, \cdot)$, it follows that $\mathbf{x}_t \equiv \mathbf{z}_t \circ \mathbf{f}_0$ for each t. Now, for any nonsingular $\mathbf{M} \in \mathbb{R}^{n_p \times n_p}$, let $\mathbf{B} := [\mathbf{f}_0]'(\mathbf{p}_0; \mathbf{M})$. Applying the chain rule (2.8) and post-multiplying the result by $\mathbf{M}$ yields:

$$\mathbf{x}_t{}'(\mathbf{p}_0; \mathbf{M}) = [\mathbf{z}_t]'(\mathbf{f}_0(\mathbf{p}_0); \mathbf{B}). \tag{5.16}$$

Thus, $\mathbf{J}_{\mathrm{L}}\mathbf{x}_t(\mathbf{p}_0; \mathbf{M})$ can be evaluated by the following procedure:

Step 1: Evaluate $\mathbf{B}$.

Step 2: Use Theorem 5.2.4 to evaluate $[\mathbf{z}_t]'(\mathbf{f}_0(\mathbf{p}_0); \mathbf{B})$.

Step 3: Evaluate $\mathbf{J}_{\mathrm{L}}\mathbf{x}_t(\mathbf{p}_0; \mathbf{M})$ by solving the linear equation system (5.16).

Theorem 5.2.1 may be extended to cover (5.15) in a similar fashion.

This result may be extended in turn to parametric ODEs whose right-hand side functions depend explicitly on parameters $\mathbf{p} \in \mathbb{R}^{n_p}$:

$$\frac{d\mathbf{x}}{dt}(t, \mathbf{p}) = \mathbf{f}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})), \qquad \mathbf{x}(t_0, \mathbf{p}) = \mathbf{f}_0(\mathbf{p}). \tag{5.17}$$

Considering $\mathbf{p}$ as a constant dependent variable instead, the following ODE is constructed in terms of the augmented dependent variable $\mathbf{z} \equiv (\mathbf{p}, \mathbf{x})$, and is equivalent to (5.17):

$$\frac{d\mathbf{z}}{dt}(t, \mathbf{p}) = \mathbf{h}(t, \mathbf{z}(t, \mathbf{p})), \qquad \mathbf{z}(t_0, \mathbf{p}) = \mathbf{h}_0(\mathbf{p}),$$

where

$$\mathbf{h} : (t, (\mathbf{q}, \mathbf{c})) \mapsto \begin{bmatrix} \mathbf{0} \\ \mathbf{f}(t, \mathbf{q}, \mathbf{c}) \end{bmatrix}, \qquad \text{and} \qquad \mathbf{h}_0 : \mathbf{q} \mapsto \begin{bmatrix} \mathbf{q} \\ \mathbf{f}_0(\mathbf{q}) \end{bmatrix}.$$

Provided that $\mathbf{h}$ satisfies conditions analogous to the hypotheses of Theorem 5.2.4, the above ODE in $\mathbf{z}$ may be treated in the same manner as (5.15). In the special

case in which x_t is scalar-valued and convex on some neighborhood of $\mathbf{p}_0$, the discussion in Section 6.2 of [79] implies that $\mathbf{J}_{\mathrm{L}} x_t(\mathbf{p}_0; \mathbf{M})$ is a subgradient of x_t at $\mathbf{p}_0$. Hence, Theorem 5.2.4 describes certain subgradients of any convex solution of a nonsmooth parametric ODE system as the unique solutions of corresponding ODEs.

5.3 Sensitivities for optimal control

This section reproduces the conference proceedings [57], and combines the theory developed in this chapter with Nesterov's inclusion $\partial_{\mathrm{L}} f(\mathbf{x}) \subset \partial f(\mathbf{x})$ [79] for scalar-valued functions $f : X \subset \mathbb{R}^n \to \mathbb{R}$, to describe generalized derivatives for certain optimal control problems.

Given open sets $X \subset \mathbb{R}^n$ and $U \subset \mathbb{R}^{n_u}$, consider the following generic open-loop optimal control problem:

$$\inf_{\mathbf{u} \in \mathcal{U}} \phi(\mathbf{u}(t_f), \mathbf{x}(t_f, \mathbf{u})), \tag{5.18}$$

where $\mathcal{U} := L^1([t_0, t_f], U)$ is the class of Lebesgue-integrable functions mapping $[t_0, t_f]$ into U, where, for each $\mathbf{u} \in \mathcal{U}$, $t \mapsto \mathbf{x}(t, \mathbf{u})$ is an absolutely continuous solution of the following ordinary differential equation (ODE):

$$\begin{aligned} \dot{\mathbf{x}}(t, \mathbf{u}) &= \mathbf{f}(\mathbf{u}(t), \mathbf{x}(t, \mathbf{u})), \qquad \text{a.e.} \quad t \in [t_0, t_f] \\ \mathbf{x}(t_0, \mathbf{u}) &= \mathbf{x}_0 \in X, \end{aligned} \tag{5.19}$$

and where the functions $\phi : U \times X \to \mathbb{R}$ and $\mathbf{f} : U \times X \to \mathbb{R}^n$ are locally Lipschitz continuous and lexicographically smooth in the sense of Nesterov [79], but are not necessarily differentiable everywhere. Since $\mathbf{f}$ is locally Lipschitz continuous, it follows that for any fixed $\mathbf{u} \in \mathcal{U}$, any solution $t \mapsto \mathbf{x}(t, \mathbf{u})$ of the above ODE on $[t_0, t_f]$ is necessarily unique [26]. Applications of such problems include control of systems with discrete operating regimes, control of chemical processes with discrete transitions in thermodynamic phase or flow regime, and control of

bioreactors modelled using dynamic flux balance analysis [43].

Observe that any direct dependence of ϕ on t_f or of $\mathbf{f}$ on t may be handled in this framework by appending an extra state variable to $\mathbf{x}$ which holds the value of t. Integral terms may be similarly incorporated into the objective function of (5.18) by appending extra state variables to the ODE.

Standard optimal control approaches (summarized, for example, in [27, 62]) typically demand differentiability of the functions ϕ and $\mathbf{f}$ in (5.18) and (7.2); any nonsmoothness in these functions thus limits the applicability of these approaches. While there exist extensions [16, Ch. 5] of standard indirect methods to nonsmooth problems, these methods require full knowledge of the generalized derivatives of certain Hamiltonian functions, which can be nontrivial to furnish.

This section is concerned with direct methods for solving (5.18), in which the control $\mathbf{u}$ is discretized and represented by a finite collection of parameters. With this approximation, bundle methods [63, 67] for nonsmooth optimization can be used to solve (5.18) locally. These methods require evaluation of any single element of Clarke's generalized gradient [16] of the objective function at each iteration; describing such a generalized gradient element is the central goal of this work. Since the generalized gradient does not satisfy a sharp chain rule, lexicographic differentiation [79] will be employed to handle compositions of functions.

Thus, two representative discrete parametrizations of the control $\mathbf{u}$ will be considered. In the first parametrization, the control is represented as a finite linear combination of bounded, Lebesgue-measurable basis functions $\{\boldsymbol{\psi}_{(i)}\}_{i\in\mathbb{N}}$, with $\boldsymbol{\psi}_{(i)} : [t_0, t_f] \to \mathbb{R}^{n_u}$ for each $i \in \mathbb{N}$:

$$\mathbf{a} \in A \subset \mathbb{R}^{n_a}; \qquad \mathbf{u} : t \in [t_0, t_f] \mapsto \sum_{i=1}^{n_a} a_i \, \boldsymbol{\psi}_{(i)}(t), \tag{5.20}$$

where A is an open set, chosen so that $\sum_{i=1}^{n_a} a_i \boldsymbol{\psi}_{(i)}(t) \in U$ for each $t \in [t_0, t_f]$ and each $\mathbf{a} \in A$. For example, the functions $\boldsymbol{\psi}_{(i)}$ could be chosen as Lagrange polynomials or Legendre polynomials. In the second parametrization, the control is piecewise constant: constants $\{t_k\}_{k=1}^{n_s}$ are chosen in $[t_0, t_f]$ such that $t_0 < t_1 <$

$\ldots < t_{n_s} = t_f$, and the control $\mathbf{u}$ satisfies:

$$\mathbf{u}(t) = \mathbf{w}_{(i)}, \qquad \forall t \in [t_{i-1}, t_i), \quad \forall i \in \{1, \ldots, n_s\}, \tag{5.21}$$

and $\mathbf{u}(t_f) = \mathbf{w}_{(n_s)}$, for parameters $\mathbf{w}_{(1)}, \ldots, \mathbf{w}_{(n_s)} \in U$. This second parametrization may be expressed in the form of the first parametrization; however, it is less cumbersome to treat directly.

Under these parametrizations of $\mathbf{u}$, Corollaries 5.3.2 and 5.3.4 in this article describe generalized gradient elements of the objective function of (5.18) with respect to the parameters $\mathbf{a}$ and $\mathbf{w}_{(1)}, \ldots, \mathbf{w}_{(n_s)}$. This description builds on the results of this chapter concerning generalized derivatives of ODE solutions with respect to the ODE initial conditions, as well as results from Chapter 3 concerning the evaluation of generalized derivatives for compositions of known functions. The obtained generalized gradient elements are described in terms of the unique solutions of certain auxiliary ODEs.

In the special case in which the objective function of the considered parametrized version of (5.18) is convex, the described generalized gradient elements are subgradients in the sense of convex analysis. If the objective function is strictly differentiable at a considered domain point, then the described generalized element is the derivative of this function, even though $\mathbf{f}$ may not be differentiable. To our knowledge, this work represents the first description of generalized derivatives with these properties for a generic nonsmooth optimal control problem.

The remainder of this section extends the results of this chapter to describe elements of Clarke's generalized gradients for (5.18) under the control parametrizations (5.20) and (5.21).

Assume that LD-derivatives are computable for the lexicographically smooth functions ϕ and $\mathbf{f}$ describing the optimal control problem (5.18). This assumption is mild: a general procedure for evaluating LD-derivatives for finite compositions of known functions is presented in [61]. Under this assumption, the following results extend Theorem 4.2 of [55] to show that for each of the two considered discrete parametrizations of (5.18), elements of the generalized gradient of the objec-

tive function can be described in terms of the unique solutions of certain auxiliary ODEs.

5.3.1 Control with basis function expansion

With $\mathbf{u}$ parametrized according to (5.20), we may consider $\mathbf{x}(t,\cdot)$ to be a function of the parameter $\mathbf{a}$ instead of the function $\mathbf{u}$, yielding the following reformulation of the original optimal control problem (5.18):

$$\inf_{\mathbf{a}\in A} \phi\left(\bar{\mathbf{u}}(t_f,\mathbf{a}),\mathbf{x}(t_f,\mathbf{a})\right) \tag{5.22}$$

where

$$\bar{\mathbf{u}}(t,\mathbf{a}) := \sum_{i=1}^{n_a} a_i \boldsymbol{\psi}_{(i)}(t), \qquad \forall \mathbf{a}\in A,$$

and where we assume that, for each $\mathbf{a} \in A$, $\mathbf{x}(\cdot,\mathbf{a})$ solves the following ODE uniquely on $[t_0,t_f]$:

$$\dot{\mathbf{x}}(t,\mathbf{a}) = \mathbf{f}\left(\bar{\mathbf{u}}(t,\mathbf{a}),\mathbf{x}(t,\mathbf{a})\right), \qquad \mathbf{x}(t_0,\mathbf{a}) = \mathbf{x}_0. \tag{5.23}$$

Theorem 5.3.1. *At any $t \in [t_0,t_f]$, the mapping $\mathbf{x}_t \equiv \mathbf{x}(t,\cdot)$ described by* (5.23) *is lexicographically smooth at any particular $\hat{\mathbf{a}} \in A$. Moreover, defining a matrix:*

$$\boldsymbol{\Psi}(t) := \begin{bmatrix}\boldsymbol{\psi}_{(1)}(t) & \cdots & \boldsymbol{\psi}_{(n_a)}(t)\end{bmatrix} \in \mathbb{R}^{n_u\times n_a}, \ \forall t\in[t_0,t_f],$$

and choosing any $p\in\mathbb{N}$ and any matrix $\mathbf{M}\in\mathbb{R}^{n_a\times p}$, the mapping $t\mapsto [\mathbf{x}_t]'(\hat{\mathbf{a}};\mathbf{M})$ is the unique solution on $[t_0,t_f]$ of the ODE:

$$\dot{\mathbf{A}}(t) = \mathbf{f}'\left((\bar{\mathbf{u}}(t,\hat{\mathbf{a}}),\mathbf{x}(t,\hat{\mathbf{a}})); \begin{bmatrix}\boldsymbol{\Psi}(t)\,\mathbf{M}\\ \mathbf{A}(t)\end{bmatrix}\right), \quad \mathbf{A}(t_0) = \mathbf{0}_{n\times p}.$$

Proof. Define a mapping $\mathbf{g}:[t_0,t_f]\times X\times A\to\mathbb{R}^n$ such that:

$$\begin{aligned}\mathbf{g}(t,\boldsymbol{\xi},\boldsymbol{\alpha}) &:= \mathbf{f}(\bar{\mathbf{u}}(t,\boldsymbol{\alpha}),\boldsymbol{\xi}),\\ &\qquad \forall t\in[t_0,t_f], \quad \forall(\boldsymbol{\xi},\boldsymbol{\alpha})\in X\times A.\end{aligned}$$

For any vector $\mathbf{c} \in X \times A$ in a sufficiently small neighborhood of $(\mathbf{x}_0, \hat{\mathbf{a}})$, let $\mathbf{z}(\cdot, \mathbf{c})$ denote a solution on $[t_0, t_f]$ of the ODE:

$$\dot{\mathbf{z}}(t, \mathbf{c}) = \begin{bmatrix} \mathbf{g}(t, \mathbf{z}(t, \mathbf{c})) \\ \mathbf{0}_{n_a} \end{bmatrix}, \qquad \mathbf{z}(t_0, \mathbf{c}) = \mathbf{c}.$$

Since $\mathbf{f}$ is locally Lipschitz continuous, the right-hand side function of the above ODE is locally Lipschitz continuous as well, and thus $\mathbf{z}(\cdot, \mathbf{c})$ is unique. By construction of $\mathbf{g}$ and $\mathbf{z}$, observe that for each $t \in [t_0, t_f]$ and each $\mathbf{a}$ in a sufficiently small neighborhood of $\hat{\mathbf{a}} \in A$,

$$\mathbf{z}(t, \mathbf{x}_0, \mathbf{a}) = \begin{bmatrix} \mathbf{x}(t, \mathbf{a}) \\ \mathbf{a} \end{bmatrix}. \tag{5.24}$$

Applying [55, Theorem 4.2], for each $t \in [t_0, t_f]$, the mapping $\mathbf{z}_t \equiv \mathbf{z}(t, \cdot)$ is lexicographically smooth at $(\mathbf{x}_0, \hat{\mathbf{a}})$, and, for fixed $\mathbf{M} \in \mathbb{R}^{n_a \times p}$, the mapping

$$t \mapsto [\mathbf{z}_t]' \left((\mathbf{x}_0, \hat{\mathbf{a}}); \begin{bmatrix} \mathbf{0}_{n \times p} \\ \mathbf{M} \end{bmatrix} \right)$$

is the unique solution $(\mathbf{B}, \mathbf{C})$ on $[t_0, t_f]$ of the ODE system:

$$\begin{aligned} \dot{\mathbf{B}}(t) &= [\mathbf{g}_t]' \left(\mathbf{z}(t, \mathbf{x}_0, \hat{\mathbf{a}}); \begin{bmatrix} \mathbf{B}(t) \\ \mathbf{C}(t) \end{bmatrix} \right), \\ \dot{\mathbf{C}}(t) &= \mathbf{0}_{n_a \times p}, \\ \mathbf{B}(t_0) &= \mathbf{0}_{n \times p}, \qquad \mathbf{C}(t_0) = \mathbf{M}, \end{aligned}$$

where $\mathbf{g}_t \equiv \mathbf{g}(t, \cdot)$. By inspection, $\mathbf{C}(t) = \mathbf{M}$ for all t, and so $\mathbf{B}$ is the unique solution of the ODE:

$$\dot{\mathbf{B}}(t) = [\mathbf{g}_t]' \left(\mathbf{z}(t, \mathbf{x}_0, \hat{\mathbf{a}}); \begin{bmatrix} \mathbf{B}(t) \\ \mathbf{M} \end{bmatrix} \right), \qquad \mathbf{B}(t_0) = \mathbf{0}_{n \times p}. \tag{5.25}$$

Equation (5.24) shows that each component of $\mathbf{x}(t, \mathbf{a}) \in X$ is a component of $\mathbf{z}(t, \mathbf{x}_0, \mathbf{a})$, for each $t \in [t_0, t_f]$ and each $\mathbf{a}$ in some neighborhood of $\hat{\mathbf{a}}$. Thus, the above results and Proposition 3.1.2 show that $\mathbf{x}_t \equiv \mathbf{x}(t, \cdot)$ is lexicographically smooth at $\hat{\mathbf{a}}$, and that the mapping $t \mapsto [\mathbf{x}_t]'(\hat{\mathbf{a}}; \mathbf{M})$ is given by the function $\mathbf{B}$ described above. Now, for each $t \in [t_0, t_f]$, define a mapping $\gamma_t : X \times A \to U \times X$

such that:

$$\gamma_t(\boldsymbol{\xi}, \boldsymbol{\alpha}) \mapsto (\bar{\mathbf{u}}(t, \boldsymbol{\alpha}), \boldsymbol{\xi}) = \left(\sum_{i=1}^{n_a} \alpha_i \psi_{(i)}(t),\, \boldsymbol{\xi} \right), \qquad \forall (\boldsymbol{\xi}, \boldsymbol{\alpha}) \in X \times A.$$

Since γ_t is continuously differentiable for each t, Proposition 3.1.2 implies that for any matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{n \times p}$, any $\boldsymbol{\xi} \in X$, and any $\boldsymbol{\alpha} \in A$,

$$[\gamma_t]'\left((\boldsymbol{\xi}, \boldsymbol{\alpha}); \begin{bmatrix} \boldsymbol{\Gamma} \\ \mathbf{M} \end{bmatrix} \right) = \begin{bmatrix} \mathbf{0}_{n_u \times n} & \boldsymbol{\Psi}(t) \\ \mathbf{I}_{n \times n} & \mathbf{0}_{n \times n_a} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Gamma} \\ \mathbf{M} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Psi}(t)\, \mathbf{M} \\ \boldsymbol{\Gamma} \end{bmatrix}.$$

Thus, observing that $\mathbf{g}_t \equiv \mathbf{f} \circ \gamma_t$ by construction, Proposition 3.1.2 shows that the right-hand side functions of the ODE (6.8) and the ODE describing $\mathbf{A}$ in the statement of the theorem are identical. Hence, the ODE describing $\mathbf{A}$ has a unique solution $\mathbf{A} \equiv \mathbf{B}$ on $[t_0, t_f]$, and the result of the theorem follows immediately. □

The following corollary uses the above theorem to describe a generalized gradient element for the objective function of (5.22).

Corollary 5.3.2. *Denote the objective function of the parametrized optimal control problem* (5.22) *as*

$$J_1 : \mathbf{a} \mapsto \phi\left(\bar{\mathbf{u}}(t_f, \mathbf{a}), \mathbf{x}(t_f, \mathbf{a}) \right),$$

and define $\boldsymbol{\Psi}(t)$ for each t as in Theorem 5.3.1. For any nonsingular $\mathbf{M} \in \mathbb{R}^{n \times n}$, the unique vector $\mathbf{v}$ solving the following linear equation system is an element of the generalized gradient of J_1 at $\hat{\mathbf{a}} \in A$:

$$\mathbf{v}^{\mathrm{T}} \mathbf{M} = \phi'\left(\left(\bar{\mathbf{u}}(t_f, \hat{\mathbf{a}}), \mathbf{x}(t_f, \hat{\mathbf{a}}) \right); \begin{bmatrix} \boldsymbol{\Psi}(t_f)\, \mathbf{M} \\ [\mathbf{x}_{t_f}]'(\hat{\mathbf{a}}; \mathbf{M}) \end{bmatrix} \right),$$

in which the LD-derivative $[\mathbf{x}_{t_f}]'(\hat{\mathbf{a}}; \mathbf{M})$ is evaluated according to Theorem 5.3.1.

Proof. The nonsingularity of $\mathbf{M}$ implies the uniqueness of $\mathbf{v}$, and Theorem 5.3.1 evidently describes $[\mathbf{x}_{t_f}]'(\hat{\mathbf{a}}; \mathbf{M})$. To complete this proof, observe that, with the continuously differentiable function $\gamma_t : X \times A \to U \times X$ defined for each t as

in the proof of Theorem 5.3.1, Proposition 3.1.2 shows that J_1 is lexicographically smooth at $\hat{\mathbf{a}}$, with:

$$
\begin{aligned}
&[J_1]'(\hat{\mathbf{a}};\mathbf{M}) \\
&\quad = \phi'\left(\gamma_{t_f}(\mathbf{x}(t_f,\hat{\mathbf{a}}),\hat{\mathbf{a}});\, [\gamma_{t_f}]'\left((\mathbf{x}(t_f,\hat{\mathbf{a}}),\hat{\mathbf{a}});\, \begin{bmatrix} [\mathbf{x}_{t_f}]'(\hat{\mathbf{a}};\mathbf{M}) \\ \mathbf{M} \end{bmatrix}\right)\right), \\
&\quad = \phi'\left(\left(\bar{\mathbf{u}}(t_f,\hat{\mathbf{a}}),\mathbf{x}(t_f,\hat{\mathbf{a}})\right);\, \begin{bmatrix} \mathbf{\Psi}(t_f)\,\mathbf{M} \\ [\mathbf{x}_{t_f}]'(\hat{\mathbf{a}};\mathbf{M}) \end{bmatrix}\right).
\end{aligned}
$$

Moreover, since $\mathbf{M}$ is nonsingular and J_1 is scalar-valued, there exists an element $\hat{\mathbf{v}}$ of the generalized gradient of J_1 at $\hat{\mathbf{a}}$ for which $\hat{\mathbf{v}}^{\mathrm{T}}\mathbf{M} = [J_1]'(\hat{\mathbf{a}};\mathbf{M})$. Hence, $\mathbf{v} = \hat{\mathbf{v}}$ is an element of the generalized gradient of J_1 at $\hat{\mathbf{a}}$. □

5.3.2 Piecewise constant control

With $\mathbf{u}$ parametrized according to (5.20), define a vector $\mathbf{w} \in \mathbb{R}^{n_s n_u}$ as:

$$
\mathbf{w} := \begin{bmatrix} \mathbf{w}_{(1)} \\ \vdots \\ \mathbf{w}_{(n_s)} \end{bmatrix} \in U^{n_s} \subset \mathbb{R}^{n_s n_u},
$$

where each $\mathbf{w}_{(i)} \in U$. We may thus consider $\mathbf{x}(t,\cdot)$ to be a function of the parameter $\mathbf{w}$ instead of the function $\mathbf{u}$, yielding the following reformulation of the original optimal control problem (5.18):

$$
\inf_{\mathbf{w}\in U^{n_s}} \phi(\mathbf{w}_{(n_s)}, \mathbf{x}(t_f,\mathbf{w})) \tag{5.26}
$$

where we assume that, for each $\mathbf{w} \in U^{n_s}$, $\mathbf{x}(\cdot,\mathbf{w})$ is absolutely continuous on $[t_0, t_f]$, and solves the following ODE system uniquely:

$$
\begin{aligned}
\dot{\mathbf{x}}(t,\mathbf{w}) &= \mathbf{f}(\mathbf{w}_{(i)}, \mathbf{x}(t,\mathbf{w})), \\
&\qquad\qquad \forall t \in (t_{i-1}, t_i), \quad \forall i \in \{1,\ldots,n_s\}, \\
\mathbf{x}(t_0,\mathbf{w}) &= \mathbf{x}_0.
\end{aligned} \tag{5.27}
$$

Observe that for any fixed $\hat{\mathbf{w}} \in U^{n_s}$, the ODE solution $\mathbf{x}(\cdot,\hat{\mathbf{w}})$ is continuously

differentiable on the set $[t_0, t_f] \backslash \{t_1, \ldots, t_{n_s}\}$.

Theorem 5.3.3. *At any $t \in [t_0, t_f]$, the mapping $\mathbf{x}_t \equiv \mathbf{x}(t, \cdot)$ is lexicographically smooth at any particular $\hat{\mathbf{w}} \in U^{n_s}$. Moreover, consider any matrix*

$$\mathbf{M} := \begin{bmatrix} \mathbf{M}_{(1)} \\ \vdots \\ \mathbf{M}_{(n_s)} \end{bmatrix} \in \mathbb{R}^{n_s n_u \times p},$$

where $p \in \mathbb{N}$ and each $\mathbf{M}_{(i)} \in \mathbb{R}^{n_u \times p}$. The mapping $t \mapsto [\mathbf{x}_t]'(\hat{\mathbf{w}}; \mathbf{M})$ is then the unique, absolutely continuous solution on $[t_0, t_f]$ of the ODE:

$$\begin{aligned} \dot{\mathbf{A}}(t) &= \mathbf{f}' \left((\hat{\mathbf{w}}_{(i)}, \mathbf{x}(t, \hat{\mathbf{w}})); \begin{bmatrix} \mathbf{M}_{(i)} \\ \mathbf{A}(t) \end{bmatrix} \right), \\ &\qquad \forall t \in (t_{i-1}, t_i), \quad \forall i \in \{1, \ldots, n_s\}, \\ \mathbf{A}(t_0) &= \mathbf{0}_{n \times p}. \end{aligned}$$

Proof. Define $t_{-1} := t_0$ for notational convenience, in which case $[t_{-1}, t_0] = \{t_0\}$. It suffices to prove by induction on $i \in \{0, 1, \ldots, n_s\}$ that for each $t \in [t_{i-1}, t_i]$, $\mathbf{x}_t$ is lexicographically smooth, and that the mapping $\mathbf{A}$ described in the statement of the theorem is well-defined when restricted to $[t_{i-1}, t_i]$, and that the mapping $t \mapsto [\mathbf{x}_t]'(\hat{\mathbf{w}}; \mathbf{M})$ is identical to $\mathbf{A}$ on $[t_{i-1}, t_i]$.

As the base case of the induction, with $i := 0$, observe that $\mathbf{A}(t_0) = \mathbf{0}_{n \times p}$ by construction. Moreover, since $\mathbf{x}(t_0, \mathbf{w}) = \mathbf{x}_0$ for each $\mathbf{w} \in U^{n_s}$, $\mathbf{x}_{t_0}$ is constant, and is thus trivially lexicographically smooth, with

$$[\mathbf{x}_{t_0}]'(\hat{\mathbf{w}}; \mathbf{M}) = \mathbf{0}_{n \times p} = \mathbf{A}(t_0).$$

Since $[t_{-1}, t_0] = \{t_0\}$, the base case is thereby complete.

As the inductive step, suppose that the required statement holds for $i := j - 1$, for some $j \in \{1, \ldots, n_s\}$. Thus, $\mathbf{A}(t_{j-1})$ is well-defined, $\mathbf{x}_{t_{j-1}}$ is lexicographically smooth, and $\mathbf{A}(t_{j-1}) = [\mathbf{x}_{t_{j-1}}]'(\hat{\mathbf{w}}; \mathbf{M})$. Consider the case in which $i := j$.

Thus, consider the mapping:

$$\boldsymbol{\pi}_{(j)} : U^{n_s} \to U : \mathbf{w} \mapsto \mathbf{w}_{(j)}.$$

This mapping is evidently linear; its derivative is given by:

$$\mathbf{J}\boldsymbol{\pi}_{(j)}(\mathbf{w}) = \begin{bmatrix}\mathbf{P}_{(1)} & \cdots & \mathbf{P}_{(n_s)}\end{bmatrix},$$

where

$$\mathbf{P}_{(k)} := \begin{cases} \mathbf{I}_{n_u \times n_u}, & \text{if } k = j, \\ \mathbf{0}_{n_u \times n_u}, & \text{if } k \neq j, \end{cases} \qquad \forall k \in \{1, \ldots, n_s\}.$$

The ODE (5.27) shows that on (t_{j-1}, t_j), $\mathbf{x}(\cdot, \mathbf{w})$ evolves according to:

$$\dot{\mathbf{x}}(t, \mathbf{w}) = \mathbf{f}(\boldsymbol{\pi}_{(j)}(\mathbf{w}), \mathbf{x}(t, \mathbf{w})).$$

Now, define a mapping $\tilde{\mathbf{g}} : X \times U^{n_s} \to \mathbb{R}^n$ such that:

$$\tilde{\mathbf{g}}(\boldsymbol{\xi}, \boldsymbol{\omega}) = \mathbf{f}(\boldsymbol{\omega}_{(j)}, \boldsymbol{\xi}), \quad \forall (\boldsymbol{\xi}, \boldsymbol{\omega}) \in X \times U^{n_s}.$$

where $\boldsymbol{\omega}_{(j)} := \boldsymbol{\pi}_{(j)}(\boldsymbol{\omega})$ is defined analogously to $\mathbf{w}_{(j)}$. For any vector $\mathbf{d}$ in a sufficiently small neighborhood of $(\mathbf{x}_0, \hat{\mathbf{w}}) \in X \times U^{n_s}$, let $\boldsymbol{\zeta}(\cdot, \mathbf{d})$ denote a solution on $[t_{j-1}, t_j]$ of the ODE:

$$\dot{\boldsymbol{\zeta}}(t, \mathbf{d}) = \begin{bmatrix} \tilde{\mathbf{g}}(\boldsymbol{\zeta}(t, \mathbf{d})) \\ \mathbf{0}_{n_s n_u} \end{bmatrix}, \qquad \boldsymbol{\zeta}(t_{j-1}, \mathbf{d}) = \mathbf{d}.$$

Proceeding similarly to the proof of Theorem 5.3.1, since $\mathbf{f}$ is locally Lipschitz continuous, the right-hand side function of the above ODE is locally Lipschitz continuous as well, and thus $\boldsymbol{\zeta}(\cdot, \mathbf{d})$ is unique. By construction of $\tilde{\mathbf{g}}$ and $\boldsymbol{\zeta}$, observe that for each $t \in [t_{j-1}, t_j]$ and each $\mathbf{w}$ in a sufficiently small neighborhood of $\hat{\mathbf{w}} \in U^{n_s}$,

$$\boldsymbol{\zeta}(t, \mathbf{x}(t_{j-1}, \mathbf{w}), \mathbf{w}) = \begin{bmatrix} \mathbf{x}(t, \mathbf{w}) \\ \mathbf{w} \end{bmatrix}. \tag{5.28}$$

Applying [55, Theorem 4.2], for each $t \in [t_{j-1}, t_j]$, the mapping $\boldsymbol{\zeta}_t \equiv \boldsymbol{\zeta}(t, \cdot)$ is lexicographically smooth at $(\mathbf{x}(t_{j-1}, \hat{\mathbf{w}}), \hat{\mathbf{w}})$, and the LD-derivative mapping

$$t \mapsto [\boldsymbol{\zeta}_t]' \left((\mathbf{x}(t_{j-1}, \hat{\mathbf{w}}), \hat{\mathbf{w}}); \begin{bmatrix} [\mathbf{x}_{t_{j-1}}]'(\hat{\mathbf{w}}; \mathbf{M}) \\ \mathbf{M} \end{bmatrix} \right)$$

is the unique solution $(\tilde{\mathbf{B}}, \tilde{\mathbf{C}})$ on $[t_{j-1}, t_j]$ of the ODE system:

$$
\begin{aligned}
\dot{\tilde{\mathbf{B}}}(t) &= \tilde{\mathbf{g}}'\left(\zeta(t, \mathbf{x}(t_{j-1}, \hat{\mathbf{w}}), \hat{\mathbf{w}}); \begin{bmatrix} \tilde{\mathbf{B}}(t) \\ \tilde{\mathbf{C}}(t) \end{bmatrix}\right), \\
\dot{\tilde{\mathbf{C}}}(t) &= \mathbf{0}_{n_s n_u \times p}, \\
\tilde{\mathbf{B}}(t_{j-1}) &= [\mathbf{x}_{t_{j-1}}]'(\hat{\mathbf{w}}; \mathbf{M}), \qquad \tilde{\mathbf{C}}(t_{j-1}) = \mathbf{M}.
\end{aligned}
$$

By inspection, $\tilde{\mathbf{C}}(t) = \mathbf{M}$ for all t, and so $\tilde{\mathbf{B}}$ is the unique solution of the ODE:

$$
\begin{aligned}
\dot{\tilde{\mathbf{B}}}(t) &= \tilde{\mathbf{g}}'\left(\zeta(t, \mathbf{x}(t_{j-1}, \hat{\mathbf{w}}), \hat{\mathbf{w}}); \begin{bmatrix} \tilde{\mathbf{B}}(t) \\ \mathbf{M} \end{bmatrix}\right), \\
\tilde{\mathbf{B}}(t_{j-1}) &= [\mathbf{x}_{t_{j-1}}]'(\hat{\mathbf{w}}; \mathbf{M}).
\end{aligned} \tag{5.29}
$$

Equation (5.28) shows that each component of $\mathbf{x}(t, \hat{\mathbf{w}}) \in X$ is a component of $\zeta(t, (\mathbf{x}(t_{j-1}, \hat{\mathbf{w}}), \hat{\mathbf{w}}))$, for each $t \in [t_{j-1}, t_j]$. Thus, the above results and Proposition 3.1.2 show that $\mathbf{x}_t \equiv \mathbf{x}(t, \cdot)$ is lexicographically smooth at $\hat{\mathbf{w}}$, and that the mapping $t \mapsto [\mathbf{x}_t]'(\hat{\mathbf{w}}; \mathbf{M})$ is given on $[t_{j-1}, t_j]$ by the function $\tilde{\mathbf{B}}$ described above. Now, define a mapping $\boldsymbol{\eta} : X \times U^{n_s} \to U \times X$ such that:

$$
\boldsymbol{\eta}(\boldsymbol{\xi}, \boldsymbol{\omega}) \mapsto (\boldsymbol{\pi}_{(j)}(\boldsymbol{\omega}), \boldsymbol{\xi}), \quad \forall (\boldsymbol{\xi}, \boldsymbol{\omega}) \in X \times U^{n_s}.
$$

The function $\boldsymbol{\eta}$ is evidently continuously differentiable; Proposition 3.1.2 implies that for any matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{n \times p}$, any $\boldsymbol{\xi} \in X$, and any $\boldsymbol{\omega} \in U^{n_s}$,

$$
\boldsymbol{\eta}'\left((\boldsymbol{\xi}, \boldsymbol{\omega}); \begin{bmatrix} \boldsymbol{\Gamma} \\ \mathbf{M} \end{bmatrix}\right) = \begin{bmatrix} \mathbf{0}_{n_u \times n} & \mathbf{J}\boldsymbol{\pi}_{(j)}(\boldsymbol{\omega}) \\ \mathbf{I}_{n \times n} & \mathbf{0}_{n \times n_s n_u} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Gamma} \\ \mathbf{M} \end{bmatrix} = \begin{bmatrix} \mathbf{M}_{(j)} \\ \boldsymbol{\Gamma} \end{bmatrix}.
$$

Thus, observing that $\tilde{\mathbf{g}} \equiv \mathbf{f} \circ \boldsymbol{\eta}$ by construction, Proposition 3.1.2 shows that the right-hand side functions of the ODE (6.9) and the ODE describing $\mathbf{A}$ in the statement of the theorem are identical when restricted to $[t_{j-1}, t_j]$. Applying the inductive assumption, the ODE describing $\mathbf{A}$ is well-defined and has a unique solution $\mathbf{A} \equiv \tilde{\mathbf{B}}$ on $[t_{j-1}, t_j]$, thereby completing the inductive step. □

The following corollary uses the above theorem to describe a generalized gradient element for the objective function of (5.26).

Corollary 5.3.4. *Denote the objective function of the parametrized optimal control problem* (5.26) *as*

$$J_2 : \mathbf{w} \mapsto \phi\left(\mathbf{w}_{(n_s)}, \mathbf{x}(t_f, \mathbf{w})\right).$$

For any nonsingular matrix

$$\mathbf{M} := \begin{bmatrix} \mathbf{M}_{(1)} \\ \vdots \\ \mathbf{M}_{(n_s)} \end{bmatrix} \in \mathbb{R}^{n_s n_u \times n_s n_u},$$

with $\mathbf{M}_{(i)} \in \mathbb{R}^{n_u \times n_s n_u}$ *for each* i*, the unique vector* $\mathbf{v}$ *solving the following linear equation system is an element of the generalized gradient of* J_2 *at* $\hat{\mathbf{w}} \in U^{n_s}$*:*

$$\mathbf{v}^{\mathrm{T}}\mathbf{M} = \phi'\left(\left(\hat{\mathbf{w}}_{(n_s)}, \mathbf{x}(t_f, \hat{\mathbf{w}})\right); \begin{bmatrix} \mathbf{M}_{(n_s)} \\ [\mathbf{x}_{t_f}]'(\hat{\mathbf{w}}; \mathbf{M}) \end{bmatrix}\right),$$

in which the LD-derivative $[\mathbf{x}_{t_f}]'(\hat{\mathbf{w}}; \mathbf{M})$ *is evaluated according to Theorem 5.3.3.*

Proof. The nonsingularity of $\mathbf{M}$ implies the uniqueness of $\mathbf{v}$, and Theorem 5.3.3 evidently describes $[\mathbf{x}_{t_f}]'(\hat{\mathbf{w}}; \mathbf{M})$. Thus, with $\boldsymbol{\eta} : X \times U^{n_s} \to U \times X$ defined as in the proof of Theorem 5.3.3 when $j := n_s$, Proposition 3.1.2 shows that J_2 is lexicographically smooth at $\hat{\mathbf{w}}$, with:

$$\begin{aligned}
&[J_2]'(\hat{\mathbf{w}}; \mathbf{M}) \\
&\quad = \phi'\left(\boldsymbol{\eta}(\mathbf{x}(t_f, \hat{\mathbf{w}}), \hat{\mathbf{w}}); \boldsymbol{\eta}'\left((\mathbf{x}(t_f, \hat{\mathbf{w}}), \hat{\mathbf{w}}); \begin{bmatrix} [\mathbf{x}_{t_f}]'(\hat{\mathbf{w}}; \mathbf{M}) \\ \mathbf{M} \end{bmatrix}\right)\right), \\
&\quad = \phi'\left(\left(\hat{\mathbf{w}}_{(n_s)}, \mathbf{x}(t_f, \hat{\mathbf{w}})\right); \begin{bmatrix} \mathbf{M}_{(n_s)} \\ [\mathbf{x}_{t_f}]'(\hat{\mathbf{w}}; \mathbf{M}) \end{bmatrix}\right).
\end{aligned}$$

Since $\mathbf{M}$ is nonsingular and J_2 is scalar-valued, there exists an element $\hat{\mathbf{v}}$ of the generalized gradient of J_2 at $\hat{\mathbf{w}}$ for which $\hat{\mathbf{v}}^{\mathrm{T}}\mathbf{M} = [J_2]'(\hat{\mathbf{w}}; \mathbf{M})$. Hence, $\mathbf{v} = \hat{\mathbf{v}}$ is an element of the generalized gradient of J_2 at $\hat{\mathbf{w}}$. □

5.4 Conclusions

Theorems 5.2.1 and 5.2.4 describe directional derivatives and lexicographic derivatives for the unique solution of a parametric ODE system as the unique solutions of other ODEs. If the original ODE solution is known to be a scalar-valued convex function of the ODE parameters, then a subgradient is described, without requiring smoothness or convexity of the ODE right-hand side function. Similarly, if a differentiable function is the unique solution of a parametric ODE with a nonsmooth right-hand side, then its derivatives can be expressed as the solutions of corresponding ODE systems. To our knowledge, this chapter provides the first description of generalized derivatives of solutions of nonsmooth parametric ODEs that exhibit these properties.

Chapter 6

Switching behavior of solutions of nonsmooth ODEs

6.1 Introduction

This chapter is reproduced from the article [59], and focuses on solutions of the ordinary differential equation (ODE) system:

$$\frac{d\mathbf{x}}{dt}(t) = \mathbf{f}(t, \mathbf{x}(t)), \qquad \mathbf{x}(t_0) = \mathbf{c},$$

whose right-hand side function $\mathbf{f}$ is the finite composition of analytic functions and absolute-value functions. Such an *abs-factorable* function may be nonsmooth, but is necessarily locally Lipschitz continuous [32]; thus, any corresponding ODE solution $\mathbf{x}$ must be unique. Noting that every piecewise affine function is abs-factorable [97], this ODE formulation permits description of a broad range of nonsmooth dynamic systems. Applications of such systems include chemical process models with switches in thermodynamic phase and/or flow regime, auxiliary dynamic systems used to describe convex and concave relaxations of nonconvex dynamic systems [103], reachable-set methods employing ordinary differential equations with linear programs embedded [37], and *dynamic flux balance analysis* models [43], in which a nonsmooth model of cellular metabolism is embedded in the

dynamic model of a bioreactor.

Roughly, for a well-defined ODE system that switches between various smooth right-hand side functions, a solution trajectory which switches infinitely many times in a finite duration is said to exhibit *Zeno* behavior [29, 48]. In the systems considered in the present work, any switching behavior is entirely due to the absolute-value functions in the abs-factorable right-hand side function of the ODE. *Non-Zenoness* is a guarantee that Zeno behavior will not emerge. A general non-Zenoness result was demonstrated by Sussmann for ODEs with *extendably piecewise analytic* right-hand sides [108], and has been specialized to particular formulations of nonsmooth dynamic systems, such as *nonlinear complementarity systems* [86] and *piecewise affine dynamic systems* [111]. We formalize certain types of switching behavior for the abs-factorable ODEs under consideration, and obtain non-Zenoness results for these systems, even when a discontinuous control input is included. In this manner, we extend the main results of [110, 111] significantly. We also provide a pertinent restatement of our results for the special case of ODEs with linear programs embedded, as studied previously in [36, 44].

Clarke [16, Theorem 7.4.1] shows that for an ODE with a nonsmooth right-hand side function, nondifferentiability of the ODE solution $\mathbf{x}$ with respect to system parameters or initial conditions requires $\mathbf{f}$ to fail to be continuously differentiable at $(t, \mathbf{x}(t))$ for all t in some set of nonzero (Lebesgue) measure. We extend our non-Zenoness results to show that, when the control input is omitted, Clarke's necessary condition for nondifferentiability can only be satisfied when the ODE solution trajectory $\mathbf{x}$ exhibits a *valley-tracing mode*, in which the argument of some absolute-value function in $\mathbf{f}$ is identically zero for a nonzero duration. Moreover, necessary conditions that are readily verifiable during numerical integration are provided for the emergence of these valley-tracing modes. In certain cases, these conditions can be shown *a priori* not to be satisfied.

This chapter is structured as follows. Section 6.2 formalizes the classes of nonsmooth and discontinuous functions considered: *abs-factorable functions*, which are the right-hand side functions of the ODEs under consideration, and *left/right-*

analytic functions, which include the considered control inputs, and will be shown to include the solutions of the considered ODEs. Section 6.3 presents various non-Zenoness results for these ODEs, and characterizes the switching behavior of the absolute-value functions in the ODE right-hand sides. Section 6.4 extends the results of the previous section to posit necessary conditions for the emergence of valley-tracing modes. Examples are presented for illustration.

6.2 Left/right-analytic functions

This chapter exploits properties of two classes of nonsmooth systems: the *abs-factorable* functions described in Chapter 2, and the *left/right-analytic functions* described in this section, which are univariate functions that may be discontinuous, but are nevertheless piecewise analytic in a particular sense. The results in this work pertain to ODEs with abs-factorable right-hand side functions and left/right-analytic control inputs; as a key intermediate result, solutions of such ODEs will be shown to be left/right-analytic.

The signum function is defined as follows:

$$\text{sign} : \mathbb{R} \to \{-1, 0, +1\} : u \mapsto \begin{cases} +1, & \text{if } u > 0, \\ 0, & \text{if } u = 0, \\ -1, & \text{if } u < 0. \end{cases}$$

Definition 6.2.1. *Given an open set $T \subset \mathbb{R}$, a function $\mathbf{g} : T \to \mathbb{R}^n$ is* right-analytic *at $t \in T$ if there exist $a, b \in T$ such that $a < t < b$, and an analytic function $\tilde{\mathbf{g}} : (a, b) \to \mathbb{R}^n$ such that $\mathbf{g} \equiv \tilde{\mathbf{g}}$ on (t, b).*

Similarly, $\mathbf{h} : T \to \mathbb{R}^n$ is left-analytic *at t if there exist $a, b \in T$ such that $a < t < b$, and an analytic function $\tilde{\mathbf{h}} : (a, b) \to \mathbb{R}^n$ such that $\mathbf{h} \equiv \tilde{\mathbf{h}}$ on (a, t).*

A function that is both right-analytic and left-analytic at t is said to be left/right-analytic *at t.*

Remark 6.2.2. *A function that is left/right-analytic at each domain point is not necessarily analytic, or even continuous or differentiable. For example, a step function on $\mathbb{R}$ is*

evidently left/right-analytic. The class of continuous left/right-analytic functions is identical to the class of functions of single real variables which are defined on open sets and are extendably piecewise analytic *in the sense of Sussmann [108]. Thus, [108, Lemma 1] shows that any continuous left/right-analytic function is locally Lipschitz continuous.*

Remark 6.2.3. *Any left/right-analytic function* $\mathbf{h}$ *on an open set* $T \subset \mathbb{R}$ *is measurable: for any compact* $I \subset T$, $\mathbf{h}$ *may be expressed on* I *as a sum of finitely many pointwise products of analytic functions and indicator functions of measurable sets.*

Proof of Remark 2.5. Consider an open set $T \subset \mathbb{R}$ and a left/right-analytic function $\mathbf{x} : T \to \mathbb{R}^n$. Choose a compact set $K \subset T$; since K is chosen arbitrarily, it suffices to show that the restriction $\mathbf{x}|_K : K \to \mathbb{R}^n : t \mapsto \mathbf{x}(t)$ is measurable. Note that $\mathbf{x}|_K \equiv \mathbf{x}$ on K.

Since $\mathbf{x}$ is left/right-analytic, for each $t \in K$, there exists $\delta_t > 0$ such that $T_t := (t - \delta_t, t + \delta_t) \subset T$, and such that there exist analytic functions $\tilde{\mathbf{x}}^L_{(t)}, \tilde{\mathbf{x}}^R_{(t)} : T_t \to \mathbb{R}^n$, for which $\mathbf{x} \equiv \tilde{\mathbf{x}}^L_{(t)}$ on $T^L_t := (t - \delta_t, t)$, and $\mathbf{x} \equiv \tilde{\mathbf{x}}^R_{(t)}$ on $T^R_t := (t, t + \delta_t)$.

For any measurable set $S \subset \mathbb{R}$, let $\mathbb{I}_S$ denote the indicator function of S:

$$\mathbb{I}_S : \mathbb{R} \to \{0, 1\} : u \mapsto \begin{cases} 1, & \text{if } u \in S, \\ 0, & \text{if } u \notin S. \end{cases}$$

Since K is compact, and since the sets $\{T_t\}_{t \in K}$ comprise an open cover of K, there exists a finite subset $A \subset K$ such that

$$K \subset \bigcup_{t \in A} T_t = \bigcup_{t \in A} (T^L_t \cup \{t\} \cup T^R_t).$$

Enumerate the elements of A as $a_1 < a_2 < \ldots < a_m$. For each $i \in \{1, \ldots, m\}$, define sets S^L_i, S^C_i, S^R_i inductively as follows:

$$
\begin{aligned}
S_1^L &:= T_1^L, \\
S_1^C &:= \{a_1\}, \\
S_1^R &:= T_1^R, \\
S_i^L &:= T_i^L \setminus \left(\bigcup_{j=1}^{i-1} (S_j^L \cup S_j^C \cup S_j^R) \right), \qquad \forall i \in \{2, \ldots, m\}, \\
S_i^C &:= \{a_i\} \setminus \left(\bigcup_{j=1}^{i-1} (S_j^L \cup S_j^C \cup S_j^R) \right), \qquad \forall i \in \{2, \ldots, m\}, \\
S_i^R &:= T_i^R \setminus \left(\bigcup_{j=1}^{i-1} (S_j^L \cup S_j^C \cup S_j^R) \right), \qquad \forall i \in \{2, \ldots, m\}.
\end{aligned}
$$

By construction, the sets in the collection $\{S_i^L, S_i^C, S_i^R\}_{i=1}^m$ are measurable and disjoint, yet have the union $\bigcup_{t \in A} T_t \supset K$.

Combining the above results, it follows that for each $t \in K$,

$$
\mathbf{x}|_K(t) = \sum_{i=1}^m \left(\mathbb{I}_{S_i^L \cap K}(t)\, \tilde{\mathbf{x}}_{(a_i)}^L(t) + \mathbb{I}_{S_i^C \cap K}(t)\, \mathbf{x}(a_i) + \mathbb{I}_{S_i^R \cap K}(t)\, \tilde{\mathbf{x}}_{(a_i)}^R(t) \right),
$$

with $\tilde{\mathbf{x}}_{(a_i)}^L(t)$ and $\tilde{\mathbf{x}}_{(a_i)}^R(t)$ defined arbitrarily as 0 whenever $t \notin T_{a_i}$. (In this case, the indicator functions multiplying these terms would evaluate to 0 regardless.) It follows that $\mathbf{x}|_K$ is measurable. □

The following lemmata concern right-analytic functions, and are readily adapted to yield analogous results concerning left-analytic functions and left/right-analytic functions.

Lemma 6.2.4. *Given an open set $T \subset \mathbb{R}$, if a function $g : T \to \mathbb{R}$ is right-analytic at $t^* \in T$, then there exists $d \in T$ such that $d > t^*$ and exactly one of the following statements holds:*

- $g(t) > 0$ *for each* $t \in (t^*, d]$, *or*
- $g(t) < 0$ *for each* $t \in (t^*, d]$, *or*

- $g(t) = 0$ *for each* $t \in (t^*, d]$.

Proof. If g is right-analytic at t^*, then there exists $a, b \in T$ with $a < t^* < b$, and an analytic function $\tilde{g} : (a, b) \to \mathbb{R}$ such that $g \equiv \tilde{g}$ on (t^*, b). Thus, for some $\beta \in (t^*, b)$, there exists a sequence $\{\alpha_i\}_{i \in \mathbb{N}}$ in $\mathbb{R}$ such that

$$\tilde{g}(t) = \tilde{g}(t^*) + \sum_{i=1}^{\infty} \alpha_i (t - t^*)^i, \qquad \forall t \in [t^*, \beta],$$

with the above power series converging for each $t \in [t^*, \beta]$. The continuity of $\tilde{g}$ implies that the sets $\{t \in (a, b) : \tilde{g}(t) > 0\}$ and $\{t \in (a, b) : \tilde{g}(t) < 0\}$ are both open. If $\tilde{g}(t^*) \neq 0$, then, for sufficiently small $d \in (t^*, b)$, either $\tilde{g}(t) > 0$ for each $t \in [t^*, d]$ or $\tilde{g}(t) < 0$ for each $t \in [t^*, d]$. Since $g \equiv \tilde{g}$ on $(t^*, d]$, the required result then follows.

Thus, throughout the remainder of this proof, assume that $\tilde{g}(t^*) = 0$. If $\alpha_i = 0$ for each $i \in \mathbb{N}$, then $g(t) = \tilde{g}(t) = 0$ for each $t \in (t^*, \beta]$, as required. Otherwise, there exists $p := \min\{i \in \mathbb{N} : \alpha_i \neq 0\}$, allowing g to be expressed on $(t^*, \beta]$ as:

$$g(t) = \tilde{g}(t) = (t - t^*)^p \left(\alpha_p + \sum_{i=1}^{\infty} \alpha_{p+i} (t - t^*)^i \right), \qquad \forall t \in (t^*, \beta]. \tag{6.1}$$

Suppose that $\alpha_p > 0$; the case in which $\alpha_p < 0$ is analogous. Now,

$$\lim_{t \to (t^*)^+} \left(\alpha_p + \sum_{i=1}^{\infty} \alpha_{p+i} (t - t^*)^i \right) = \alpha_p > 0.$$

Thus, for sufficiently small $d \in (t^*, \beta]$,

$$\alpha_p + \sum_{i=1}^{\infty} \alpha_{p+i} (t - t^*)^i > 0, \qquad \forall t \in (t^*, d],$$

and so (6.1) implies that $g(t) > 0$ for each $t \in (t^*, d]$, as required. □

Lemma 6.2.5. *Given open sets* $T \subset \mathbb{R}$ *and* $Z \subset \mathbb{R}^n$*, a function* $\mathbf{g} : T \to Z$*, and a function* $\mathbf{h} : Z \to \mathbb{R}^m$*, if* $\mathbf{g}$ *is right-analytic at some* $t^* \in T$*, then the limit* $\mathbf{g}^* := \lim_{t \to (t^*)^+} \mathbf{g}(t)$ *exists in* $\mathbb{R}^n$*. If* $\mathbf{g}^* \in Z$*, and if* $\mathbf{h}$ *is analytic at* $\mathbf{g}^*$*, then the composite function* $\mathbf{h} \circ \mathbf{g}$ *is right-analytic at* t^**.*

Proof. The existence of $\mathbf{g}^* \in \mathbb{R}^n$ follows immediately from $\mathbf{g}$ being right-analytic at t^*. Suppose that $\mathbf{g}^* \in Z$, and, without loss of generality, assume that $m = 1$: this case is readily extended to cover the $m > 1$ case by considering $h_1 \circ \mathbf{g}, \ldots, h_m \circ \mathbf{g}$ separately. Under this assumption, $\mathbf{h}$ is a scalar-valued function $h : Z \to \mathbb{R}$.

Since $\mathbf{g}$ is right-analytic at $t^* \in T$, there exist $a, d \in T$ such that $a < t^* < d$, and analytic functions $\tilde{g}_1, \ldots, \tilde{g}_n : (a, d) \to \mathbb{R}$ for which

$$g_k(t) = \tilde{g}_k(t), \qquad \forall t \in (t^*, d), \quad \forall k \in \{1, \ldots, n\}.$$

Thus,

$$h \circ \mathbf{g}(t) = h(\tilde{g}_1(t), \ldots, \tilde{g}_n(t)), \qquad \forall t \in (t^*, d),$$

and $\mathbf{g}^* = (\tilde{g}_1(t^*), \ldots, \tilde{g}_n(t^*))$. Moreover, [66, Corollary 1.2.4 and Proposition 2.2.8] show that the mapping $t \mapsto h(\tilde{g}_1(t), \ldots, \tilde{g}_n(t))$ is well-defined and analytic on some neighborhood $N \subset T$ of t^*, as required. □

Lemma 6.2.6. *Consider open sets $T, U \subset \mathbb{R}$, some $t^* \in T$, a function $g : T \to U$, and a function $\mathbf{h} : U \to \mathbb{R}^n$. If g is right-analytic at t^*, then the limit $g^* := \lim_{t \to (t^*)^+} g(t)$ exists in $\mathbb{R}$. If $g^* \in U$, and if $\mathbf{h}$ is left/right-analytic at g^*, then the function $\mathbf{h} \circ g$ is right-analytic at t^*.*

Proof. The existence of $g^* \in \mathbb{R}$ is an immediate consequence of g being right-analytic at t^*. Suppose that $g^* \in U$. Since $\mathbf{h}$ is left/right-analytic at g^*, there exist $\delta > 0$ and analytic functions $\tilde{\mathbf{h}}_A : (g^* - \delta, g^* + \delta) \to \mathbb{R}^n$ and $\tilde{\mathbf{h}}_B : (g^* - \delta, g^* + \delta) \to \mathbb{R}^n$ such that, for each $s \in (g^*, g^* + \delta)$,

$$\mathbf{h}(s) = \tilde{\mathbf{h}}_A(s), \quad \forall s \in (g^* - \delta, g^*), \quad \text{and} \quad \mathbf{h}(s) = \tilde{\mathbf{h}}_B(s).$$

Since the function $\gamma \equiv g - g^*$ is right-analytic at t^*, Lemma 2.6 implies that for sufficiently small $d > t^*$, either $g(t) \geq g^*$ for all $t \in (t^*, d]$, or $g(t) \leq g^*$ for all $t \in (t^*, d]$, or both. Since g is right-analytic at t^*, there exist $t^L, t^U \in T$ with $t^L < t^* < t^U$ and an analytic function $\tilde{g} : (t^L, t^U) \to \mathbb{R}$ for which $g \equiv \tilde{g}$ on (t^*, t^U), and for which $t \in (t^*, t^U)$ implies $|g(t) - g^*| = |\tilde{g}(t) - \tilde{g}(t^*)| < \delta$. Set $\tilde{d} := \min(d, t^U)$.

The cases in which $g(t) \geq g^*$ for all $t \in (t^*, d]$, or $g(t) \leq g^*$ for all $t \in (t^*, d]$, will be considered separately. If $g(t) \geq g^*$ for all $t \in (t^*, d]$, then, combining the above results,

$$\mathbf{h}(g(t)) = \tilde{\mathbf{h}}_B(\tilde{g}(t)), \qquad \forall t \in (t^*, \tilde{d}).$$

The composition $\tilde{\mathbf{h}}_B \circ \tilde{g}$ is analytic at t^*, and so $\mathbf{h} \circ g$ is right-analytic at t^*, as required.

Similarly, if $g(t) \leq g^*$ for all $t \in (t^*, d]$, then

$$\mathbf{h}(g(t)) = \tilde{\mathbf{h}}_A(\tilde{g}(t)), \qquad \forall t \in (t^*, \tilde{d}).$$

The composition $\tilde{\mathbf{h}}_A \circ \tilde{g}$ is analytic at t^*, and so $\mathbf{h} \circ g$ is right-analytic at t^*, as required. □

Lemma 6.2.7. *Given open sets $T \subset \mathbb{R}$ and $Z \subset \mathbb{R}^n$, a function $\mathbf{z} : T \to Y$ for some closed set $Y \subset Z$, and an abs-factorable function $\mathbf{f} : Z \to \mathbb{R}^m$, if $\mathbf{z}$ is right-analytic at $t^* \in T$, then the composite function $\mathbf{f} \circ \mathbf{z}$ is also right-analytic at t^*, as are the functions $\mathbf{v}_{(j)} \circ \mathbf{z}$ for each $j \in \{0, 1 \ldots, \ell\}$ and $\mathbf{u}_{(j)} \circ \mathbf{z}$ for each $j \in \{1, \ldots, \ell\}$.*

Proof. Since Y is closed and $\mathbf{z}$ is right-analytic at t^*, $\lim_{t \to (t^*)^+} \mathbf{z}(t)$ exists and is an element of Z. Thus, using Lemmata 6.2.5 and 6.2.6, a simple strong inductive proof on $j = 0, 1, \ldots, \ell$ shows that $\mathbf{v}_{(j)} \circ \mathbf{z}$ is right-analytic for each $j \in \{0, 1 \ldots, \ell\}$, as is $\mathbf{u}_{(j)} \circ \mathbf{z}$ for each $j \in \{1, \ldots, \ell\}$. Since $\mathbf{f} \equiv \mathbf{v}_{(\ell)}$ on Z, the lemma is thereby demonstrated. □

6.3 Non-Zenoness for abs-factorable ODEs

This section establishes various notions of switching behavior and non-Zenoness for the ODE systems formalized in the following assumptions, the first of which includes a left/right-analytic control input that may be discontinuous. A pertinent restatement of the obtained non-Zenoness results is provided for ODE systems with linear programs embedded.

The results in this section depend heavily on the condition that all differentiable functions in the elemental library $\mathcal{L}$ are analytic. Example 1 in [72] shows that this condition cannot be relaxed in general while retaining non-Zenoness.

Assumption 6.3.1. *Consider open sets $\bar{T} \subset \mathbb{R}$, $X \subset \mathbb{R}^n$, and $W \subset \mathbb{R}^m$, a compact set $U \subset W$, some $\mathbf{c} \in X$, a left/right-analytic function $\mathbf{w} : \bar{T} \to U$, and an abs-factorable function $\mathbf{f} : \bar{T} \times X \times W \to \mathbb{R}^n$, and suppose that there exists a solution trajectory $\mathbf{x}$ (in the Carathéodory sense, as summarized in [26, Section 1]) of the following ODE on $[t_0, t_f] \subset \bar{T}$:*

$$\frac{d\mathbf{x}}{dt}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{w}(t)), \qquad \mathbf{x}(t_0) = \mathbf{c}. \tag{6.2}$$

The following assumption is the special case of the above assumption in which $m = 0$, and so the control input $\mathbf{w}$ is omitted. Thus, any result pertaining to the system in Assumption 6.3.1 also applies to the system in the following assumption.

Assumption 6.3.2. *Consider open sets $\bar{T} \subset \mathbb{R}$ and $X \subset \mathbb{R}^n$, some $\mathbf{c} \in X$, and an abs-factorable function $\mathbf{f} : \bar{T} \times X \to \mathbb{R}^n$, and suppose that there exists a solution trajectory $\mathbf{x}$ of the following ODE on $[t_0, t_f] \subset \bar{T}$:*

$$\frac{d\mathbf{x}}{dt}(t) = \mathbf{f}(t, \mathbf{x}(t)), \qquad \mathbf{x}(t_0) = \mathbf{c}. \tag{6.3}$$

If the control input $\mathbf{w}$ in Assumption 6.3.1 is abs-factorable, then Assumption 6.3.2 applies with $\mathbf{f}$ redefined as the mapping $(t, \mathbf{x}) \mapsto \mathbf{f}(t, \mathbf{x}, \mathbf{w}(t))$. Thus, results pertaining to the system in Assumption 6.3.2 are applicable in this case.

Under Assumption 6.3.1, since $\mathbf{x}$ is absolutely continuous, the image $\mathbf{x}([t_0, t_f])$ of $[t_0, t_f]$ under $\mathbf{x}$ is compact. Thus, since $\mathbf{f}$ is locally Lipschitz continuous, there exist open sets $Y \subset \bar{T} \times X$ and $V \subset W$ such that

$$[t_0, t_f] \times \mathbf{x}([t_0, t_f]) \times U \subset Y \times V,$$

and such that $\mathbf{f}$ is bounded and uniformly Lipschitz continuous (with Lipschitz constant $k_{\mathbf{f}}$) on $Y \times V$. Since $\mathbf{w}$ is bounded and measurable, it follows that the

mapping $\mathbf{f}_{\mathbf{w}} : \mathcal{Y} \to \mathbb{R}^n : (t, \mathbf{x}) \mapsto \mathbf{f}(t, \mathbf{x}, \mathbf{w}(t))$ is bounded and measurable, and is also uniformly Lipschitz continuous (also with Lipschitz constant $k_{\mathbf{f}}$) with respect to $\mathbf{x}$ for each fixed t. Consequently, the solution trajectory $\mathbf{x}$ described in Assumption 6.3.1 (or 6.3.2) is unique. Moreover, by [26, Section 1, Theorems 1 and 2], this solution may be extended to yield a unique solution $\mathbf{x}$ of (6.2) (or (6.3)) on $T := [t_\ell, t_u] \subset \bar{T}$, for some $t_\ell < t_0$ and some $t_u > t_f$. In the spirit of [33], for each $i \in \{1, \ldots, p_{\mathbf{f}}\}$ and each $t \in T$, define a *signature* $\sigma_i(t) := \operatorname{sign} u_{(\lambda_{\mathbf{f}}(i))}(t, \mathbf{x}(t), \mathbf{w}(t)) \in \{-1, 0, 1\}$; the $\mathbf{w}(t)$ argument will be understood to be omitted in the case of Assumption 6.3.2. Thus, for each $t \in T$ and $i \in \{1, \ldots, p_{\mathbf{f}}\}$,

$$\begin{aligned} v_{(\lambda_{\mathbf{f}}(i))}(t, \mathbf{x}(t), \mathbf{w}(t)) &= |u_{(\lambda_{\mathbf{f}}(i))}(t, \mathbf{x}(t), \mathbf{w}(t))| \\ &= \sigma_i(t)\, u_{(\lambda_{\mathbf{f}}(i))}(t, \mathbf{x}(t), \mathbf{w}(t)). \end{aligned} \tag{6.4}$$

Under Assumption 6.3.2, the following example illustrates that the ODE (6.3) can be expressed as a *hybrid automaton* in the sense of [72], with continuous evolution described by an analytic ODE right-hand side function in each discrete mode. The subsequent remark shows that (6.3) may also be represented as a *nonlinear complementarity system* (NCS) in the sense of [86, 99, 116]. Since the control $\mathbf{w}$ in (6.2) may be discontinuous, (6.2) is not necessarily representable as a NCS that is *strongly regular* in the sense of [86].

Example 6.3.3. *Consider the following ODE with two differential variables:*

$$\frac{d\mathbf{x}}{dt}(t) = \begin{bmatrix} |x_1(t)| \\ -|x_2(t) - x_1(t)| \end{bmatrix}, \qquad \mathbf{x}(0) = \mathbf{c}. \tag{6.5}$$

Note that for each $y \in \mathbb{R}$, $|y| \in \{qy : q \in \{-1, +1\}\}$. Thus, defining $Q := \{-1, +1\}^2$, the ODE can be reformulated as the following instance $(Q, \mathbb{R}^2, \mathrm{Init}, \mathbf{f}, I, E, G, R)$ of the generic hybrid automaton described in [72, Definition 2]:

- $\mathrm{Init} := \{(\mathbf{q}, \mathbf{x}) : q_1 x_1 \geq 0,\ q_2(x_2 - x_1) \geq 0, \mathbf{q} \in Q, \mathbf{x} \in \mathbb{R}^2\}$,
- $\mathbf{f}(\mathbf{q}, \mathbf{x}) := \begin{bmatrix} q_1 x_1 \\ -q_2(x_2 - x_1) \end{bmatrix}$, *for each* $\mathbf{q} \in Q$ *and* $\mathbf{x} \in \mathbb{R}^2$,

- $I(\mathbf{q}) := \{\mathbf{x} \in \mathbb{R}^2 : q_1 x_1 \geq 0,\ q_2(x_2 - x_1) \geq 0\}$, *for each* $\mathbf{q} \in Q$,

- $E := E_1 \cup E_2$, *where* $E_1 := \{((q_1, q_2), (-q_1, q_2)) : q_1, q_2 \in \{-1, +1\}\}$, *and* $E_2 := \{((q_1, q_2), (q_1, -q_2)) : q_1, q_2 \in \{-1, +1\}\}$,

- $G(e) := \{\mathbf{x} \in \mathbb{R}^2 : x_1 = 0\}$ *for each* $e \in E_1$, *and* $G(e) := \{\mathbf{x} \in \mathbb{R}^2 : x_2 - x_1 = 0\}$ *for each* $e \in E_2$,

- $R(e, \mathbf{x}) := \{\mathbf{x}\}$, *for each* $e \in E$ *and* $\mathbf{x} \in \mathbb{R}^2$.

Observe that $\mathbf{f}(\mathbf{q}, \cdot)$ *is linear – and therefore analytic – for each* $\mathbf{q} \in Q$. *This hybrid automaton is equivalent to the ODE* (6.5) *in the following sense. Given any solution* $\{\mathbf{x}(t) : t \in [0, t_f]\}$ *of the ODE* (6.5), *a valid execution of the hybrid automaton on* $[0, t_f]$ *is given by* $(\boldsymbol{\tau}, (q_1, q_2), \mathbf{x})$, *where* $q_1(t) := +1$ *if* $x_1(t) \geq 0$ *and* -1 *otherwise, where* $q_2(t) := +1$ *if* $x_2(t) - x_1(t) \geq 0$ *and* -1 *otherwise, and where each discontinuity in* q_1 *or* q_2 *coincides with the endpoint of an interval in* $\boldsymbol{\tau}$. *Conversely, given any execution* $(\boldsymbol{\tau}, \mathbf{q}, \mathbf{x})$ *of the hybrid automaton on* $[0, t_f]$, $\mathbf{x}$ *evidently solves the ODE* (6.5) *on* $[0, t_f]$.

Remark 6.3.4. *For any particular* $x \in \mathbb{R}$, *the statement* $z = |x|$ *is equivalent to the existence of* $y \in \mathbb{R}$ *for which*

$$z = 2y - x, \qquad \text{and} \qquad 0 \leq y \perp y - x \geq 0; \tag{6.6}$$

it is readily verified that these conditions are satisfied if and only if $y = \max\{x, 0\}$ *and* $z = |x|$.

When the domain of each $\psi_{(j)}$ *is* $\mathbb{R}^{n_j}$ *for appropriate* $n_j \in \mathbb{N}$, *the above relationship allows the ODE* (6.3) *to be expressed as an equivalent* nonlinear complementarity system *(NCS), as defined in [86, 87]. This can be shown as follows. For each fixed* $\mathbf{y} \in \mathbb{R}^{p_\mathbf{f}}$, *define an abs-factorable function* $\hat{\mathbf{f}}(\cdot, \cdot, \mathbf{y}) : \bar{T} \times X \to \mathbb{R}^n$ *as having the same factored representation as* $\mathbf{f}$, *except for the following changes. Let quantities relating to the factored representation of* $\hat{\mathbf{f}}(\cdot, \cdot, \mathbf{y})$ *be denoted with carets. For each* $i \in \{1, \ldots, p_\mathbf{f}\}$, *noting that* $\psi_{(\lambda_\mathbf{f}(i))}$ *is the mapping* $u \mapsto |u|$, *define* $\hat{\psi}_{(\lambda_\mathbf{f}(i))} : u \mapsto 2y_i - u$ *instead. Allowing now for variation in* $\mathbf{y}$, *the function* $\hat{\mathbf{f}}$ *is evidently analytic on its domain, since it is the composition*

of analytic functions. Thus, inspection of (6.6) *shows that the ODE* (6.3) *is equivalent to the NCS:*

$$\begin{aligned} \frac{d\mathbf{x}}{dt}(t) &= \hat{\mathbf{f}}(t, \mathbf{x}(t), \mathbf{y}(t)), \qquad \mathbf{x}(t_0) = \mathbf{c}, \\ \mathbf{0} \leq \mathbf{y}(t) &\perp \hat{\mathbf{h}}(t, \mathbf{x}(t), \mathbf{y}(t)) \geq \mathbf{0}, \end{aligned}$$

where the function $\hat{\mathbf{h}} : \bar{T} \times X \times \mathbb{R}^{p_{\mathbf{f}}} \to \mathbb{R}^{p_{\mathbf{f}}}$ *is defined so that for each* $i \in \{1, \ldots, p_{\mathbf{f}}\}$,

$$\hat{h}_i : (t, \mathbf{x}, \mathbf{y}) \mapsto y_i - \hat{u}_{(\lambda_{\mathbf{f}}(i))}(t, \mathbf{x}, \mathbf{y}).$$

Since each $\hat{u}_{(\lambda_{\mathbf{f}}(i))}$ *is the composition of analytic functions,* $\hat{\mathbf{h}}$ *is evidently analytic.*

This argument must be refined when the elemental library contains functions $\psi_{(j)}$ *whose domain is some strict subset of* $\mathbb{R}^{n_j}$*: in this case, certain values of* $\mathbf{y} \in \mathbb{R}^{p_{\mathbf{f}}}$ *may lead to domain violations in the function* $\hat{\mathbf{f}}$ *described above. Since the NCS representation of* (6.3) *is tangential to the results in this work, these refinements will not be pursued further here.*

6.3.1 Basic observations about switching behavior

The following definition formalizes intuitive notions of switching behavior in the trajectory $\mathbf{x}$, in which the arguments of the absolute-value functions in the ODE right-hand side change sign over time.

Definition 6.3.5. *Suppose that Assumption 6.3.1 (or 6.3.2) holds. The trajectory* $\mathbf{x}$ *has a* valley-crossing *at* $t^* \in [t_0, t_f]$ *if there exists* $i \in \{1, \ldots, p_{\mathbf{f}}\}$ *for which* $\limsup_{t \to t^*} \sigma_i(t) = +1$ *and* $\liminf_{t \to t^*} \sigma_i(t) = -1$. *The trajectory* $\mathbf{x}$ *has a* (valley)–crossing opportunity *at* $\bar{t} \in [t_0, t_f]$ *if* $\lim_{t \to \bar{t}} \sigma_i(t)$ *does not exist for some* $i \in \{1, \ldots, p_{\mathbf{f}}\}$. *The trajectory* $\mathbf{x}$ *has a* valley-tracing mode *on a nondegenerate interval* $[a, b] \subset [t_0, t_f]$ *if there exists* $i \in \{1, \ldots, p_{\mathbf{f}}\}$ *such that* $\sigma_i(t) = 0$ *for each* $t \in [a, b]$. *(An interval* $[a, b] \subset \mathbb{R}$ *is* nondegenerate *if* $a < b$*.)*

The "valleys" in the above definition refer to the shape of the absolute-value function's graph. Valley-tracing modes exhibit the *sliding motion* described by Fil-

ippov [26] when considering an ODE right-hand side that is smooth on particular subdomains. However, since the right-hand side functions considered here are locally Lipschitz continuous with respect to $\mathbf{x}$, the ODE solution $\mathbf{x}$ exists in the Carathéodory sense, and so Filippov's alternative notion of a solution [26] is not necessary.

Remark 6.3.6. *Valley-crossings can only occur at crossing opportunities. A discontinuity in some σ_i is not necessarily a crossing opportunity or a valley-crossing.*

Lemma 6.3.7. *Suppose that Assumption 6.3.2 holds. If $\mathbf{x}$ does not have a valley-crossing at $t^* \in [t_0, t_f]$, then $\mathbf{x}$ is analytic at t^*.*

Proof. Consider such a $t^* \in [t_0, t_f]$, and define $\mathbf{x}^* := \mathbf{x}(t^*)$. Since $t^* \in [t_0, t_f]$ is not a valley-crossing, it follows that for each $i \in \{1, \ldots, p_{\mathbf{f}}\}$, either $\limsup_{t \to t^*} \sigma_i(t) \leq 0$ or $\liminf_{t \to t^*} \sigma_i(t) \geq 0$. Thus, there exists a neighborhood $N \subset T$ of t^* such that either $u_{(\lambda_{\mathbf{f}}(i))}(t, \mathbf{x}(t)) \leq 0$ for each $t \in N$, or $u_{(\lambda_{\mathbf{f}}(i))}(t, \mathbf{x}(t)) \geq 0$ for each $t \in N$. Combining this statement with (6.4), there exists $\mathbf{s} \in \{-1, +1\}^{p_{\mathbf{f}}}$ such that

$$v_{(\lambda_{\mathbf{f}}(i))}(t, \mathbf{x}(t)) = s_i \, u_{(\lambda_{\mathbf{f}}(i))}(t, \mathbf{x}(t)), \quad \forall t \in N, \quad \forall i \in \{1, \ldots, p_{\mathbf{f}}\}.$$

Now, define an abs-factorable function $\tilde{\mathbf{f}}_{\mathbf{s}}$ as having the same factored representation as $\mathbf{f}$, but with $\psi_{(\lambda_{\mathbf{f}}(i))}$ replaced for each $i \in \{1, \ldots, p_{\mathbf{f}}\}$ by $\tilde{\psi}_{(\lambda_{\mathbf{f}}(i))} : u \mapsto s_i u$. Since each $\psi_{(j)}$ is locally Lipschitz continuous and defined on an open set, and since $\tilde{\mathbf{f}}_{\mathbf{s}}(t^*, \mathbf{x}^*) = \mathbf{f}(t^*, \mathbf{x}^*)$, it follows that $\tilde{\mathbf{f}}_{\mathbf{s}}$ is well-defined and locally Lipschitz continuous on some neighborhood of $(t^*, \mathbf{x}^*)$. Moreover, $\mathbf{x}$ evidently solves the ODE:

$$\frac{d\tilde{\mathbf{x}}}{dt}(t) = \tilde{\mathbf{f}}_{\mathbf{s}}(t, \tilde{\mathbf{x}}(t)), \qquad \tilde{\mathbf{x}}(t^*) = \mathbf{x}^* \tag{6.7}$$

on N. Noting that $\tilde{\mathbf{f}}_{\mathbf{s}}$ is analytic by construction, it follows from [35, Ch. II, Section 1] that $\mathbf{x}$ is analytic at t^*. □

Lemma 6.3.8. *Suppose that Assumption 6.3.1 holds. The set of crossing opportunities of $\mathbf{x}$ in $[t_0, t_f]$ is compact.*

Proof. Since $[t_0, t_f]$ is compact, it suffices to show that the set B of crossing opportunities in $[t_0, t_f]$ is closed. Suppose that $t^* \in [t_0, t_f]$ is a limit point of B; it will be shown that $t^* \in B$. By construction of t^*, there exists a sequence $\{t_j\}_{j\in\mathbb{N}}$ in B for which $|t_j - t^*| < 2^{-j}$ for each $j \in \mathbb{N}$.

By definition of a crossing opportunity, for each $j \in \mathbb{N}$, there exists $i \in \{1, \ldots, p_{\mathbf{f}}\}$ such that $\limsup_{t\to t_j} \sigma_i(t) \neq \liminf_{t\to t_j} \sigma_i(t)$. Since $\{1, \ldots, p_{\mathbf{f}}\}$ is a finite set, there exists $i^* \in \{1, \ldots, p_{\mathbf{f}}\}$ such that the set

$$I := \{j \in \mathbb{N} : \limsup_{t\to t_j} \sigma_{i^*}(t) \neq \liminf_{t\to t_j} \sigma_{i^*}(t)\}$$

is infinite. Since $\sigma_{i^*}(t) \in \{-1, 0, +1\}$ for each $t \in T$, it follows that

$$\limsup_{t\to t_j} \sigma_{i^*}(t) \leq +1$$

for each $j \in I$, and so $\liminf_{t\to t_j} \sigma_{i^*}(t) \in \{-1, 0\}$ for each $j \in I$. The cases in which the set $J := \{j \in I : \liminf_{t\to t_j} \sigma_{i^*}(t) = 0\}$ is infinite or finite will be considered separately.

First, suppose that J is infinite. Noting that t_j is a crossing opportunity for each $j \in \mathbb{N}$, it follows that

$$0 = \liminf_{t\to t_j} \sigma_{i^*}(t) < \limsup_{t\to t_j} \sigma_{i^*}(t), \qquad \forall j \in J,$$

and so $\limsup_{t\to t_j} \sigma_{i^*}(t) = +1$ for each $j \in J$. Thus, for each $j \in J$, there exist $r_j, s_j \in T$ such that $\sigma_{i^*}(r_j) = 0$, $\sigma_{i^*}(s_j) = +1$, $|r_j - t_j| < 2^{-j}$, and $|s_j - t_j| < 2^{-j}$. The triangle inequality yields the following for each $j \in J$:

$$\begin{aligned} |r_j - t^*| &\leq |r_j - t_j| + |t_j - t^*| < 2^{1-j}, \\ |s_j - t^*| &\leq |s_j - t_j| + |s_j - t^*| < 2^{1-j}. \end{aligned}$$

Thus, $\lim_{J\ni j\to\infty} s_j = \lim_{J\ni j\to\infty} r_j = t^*$, and so

$$\begin{aligned}\limsup_{t\to t^*}\sigma_{i^*}(t) &\geq \lim_{\substack{j\in J\\ j\to\infty}}\sigma_{i^*}(s_j)\\ &= +1 > 0 = \lim_{\substack{j\in J\\ j\to\infty}}\sigma_{i^*}(r_j) \geq \liminf_{t\to t^*}\sigma_{i^*}(t),\end{aligned}$$

which shows that t^* is a crossing opportunity.

Next, suppose, instead, that J is finite. The set

$$K := I\backslash J = \{j \in I : \liminf_{t\to t_j}\sigma_{i^*}(t) = -1\}$$

is then infinite. Since t_j is a crossing opportunity for each $j \in K$, it follows that $\limsup_{t\to t_j}\sigma_{i^*}(t) \geq 0$ for each $j \in K$. Thus, a similar argument to the previous case shows that

$$\limsup_{t\to t^*}\sigma_{i^*}(t) \geq 0 > -1 \geq \liminf_{t\to t^*}\sigma_{i^*}(t),$$

by considering appropriate sequences $\{s_j\}_{j\in K}$ and $\{r_j\}_{j\in K}$ which each converge to t^*. □

6.3.2 Establishing non-Zenoness

In this subsection, certain types of non-Zenoness are established for solutions of ODE systems satisfying Assumptions 6.3.1 and 6.3.2.

Firstly, under Assumption 6.3.1, the main non-Zenoness results of [110, 111] may be extended significantly, as the following theorem demonstrates. Observe that, by [97, Proposition 2.2.2], any function that is *piecewise affine* in the sense of Scholtes [97] is also abs-factorable.

Theorem 6.3.9. *Suppose that Assumption 6.3.1 holds. The trajectory* $\mathbf{x}$ *is absolutely continuous and left/right-analytic at each* $t^* \in [t_0, t_f]$.

Proof. Since $\mathbf{x}$ is a solution of a Carathéodory ODE, $\mathbf{x}$ is absolutely continuous on T. It will be shown that $\mathbf{x}$ is right-analytic at t^*; the proof that $\mathbf{x}$ is also left-analytic at t^* is analogous. Set $\mathbf{x}^* := \mathbf{x}(t^*)$ and $\mathbf{w}^* := \lim_{t\to(t^*)^+}\mathbf{w}(t)$. This limit exists, since $\mathbf{w}$ is

right-analytic. Moreover, there exists a neighborhood $N_w \subset \bar{T}$ of t^* and an analytic function $\tilde{\mathbf{w}} : N_w \to W$ such that $\mathbf{w} \equiv \tilde{\mathbf{w}}$ on (t^*, b) for some $b \in (t^*, t_u) \cap N_w$, and so $\tilde{\mathbf{w}}(t^*) = \mathbf{w}^*$.

This proof proceeds by induction on $p_{\mathbf{f}} \in \{0, 1, 2, \ldots\}$. For the base case, if $p_{\mathbf{f}} = 0$, then $\mathbf{f}$ is analytic. Thus, for some $d \in (t^*, b)$ and some $a \in (t_\ell, t^*) \cap N_w$, the function

$$\mathbf{g} : (a, d] \times X \to \mathbb{R}^n : (t, \mathbf{z}) \mapsto \mathbf{f}(t, \mathbf{z}, \tilde{\mathbf{w}}(t))$$

is well-defined and analytic, and $\mathbf{x}$ solves the following ODE on $[t^*, d]$:

$$\frac{d\boldsymbol{\xi}}{dt}(t) = \mathbf{g}(t, \boldsymbol{\xi}(t)), \qquad \boldsymbol{\xi}(t^*) = \mathbf{x}^*.$$

Moreover, since $\mathbf{g}$ is analytic, it is locally Lipschitz continuous, and so $\mathbf{x}$ is the unique solution on $[t^*, d]$ of the above ODE. From standard ODE existence theory, there exists a unique continuation $\tilde{\mathbf{x}}$ of this solution to $(\bar{a}, d]$ for some $\bar{a} \in (a, t^*)$. Since $\mathbf{g}$ is analytic, so is $\tilde{\mathbf{x}}$ [35]. Thus, since $\mathbf{x} \equiv \tilde{\mathbf{x}}$ on (t^*, d), and since $\tilde{\mathbf{x}}$ is defined and analytic on $(\bar{a}, d) \ni t^*$, $\mathbf{x}$ is right-analytic at t^*, as required.

Now, consider the case in which $p_{\mathbf{f}} = k \in \mathbb{N}$, and denote $\lambda_{\mathbf{f}}(k)$ as λ for simplicity. As the inductive assumption, suppose that the required result would hold if $\mathbf{f}$ were replaced by any abs-factorable function ϕ for which $p_\phi = k - 1$. To show that $\mathbf{x}$ is right-analytic at t^*, the following cases will be considered separately: $u_{(\lambda)}(t^*, \mathbf{x}^*, \mathbf{w}^*) > 0$, $u_{(\lambda)}(t^*, \mathbf{x}^*, \mathbf{w}^*) < 0$, and $u_{(\lambda)}(t^*, \mathbf{x}^*, \mathbf{w}^*) = 0$.

If $u_{(\lambda)}(t^*, \mathbf{x}^*, \mathbf{w}^*) > 0$, then the continuity of $\mathbf{x}$, $u_{(\lambda)}$, and $\tilde{\mathbf{w}}$ implies the existence of a neighborhood $N_a \subset N_w$ of t^* for which $u_{(\lambda)}(t, \mathbf{x}(t), \tilde{\mathbf{w}}(t)) > 0$ for each $t \in N_a$. Define an abs-factorable function $\mathbf{f}_A$ as having the same factored representation as $\mathbf{f}$, except with $\psi_{(\lambda)} : u \to |u|$ replaced by $\psi_{A,(\lambda)} : u \to u$. (Throughout this proof, the subscript A will denote quantities relating to $\mathbf{f}_A$ instead of $\mathbf{f}$.) Noting that $\mathbf{f}(t^*, \mathbf{x}^*, \mathbf{w}^*) = \mathbf{f}_A(t^*, \mathbf{x}^*, \mathbf{w}^*)$ by construction, and that each function in the factored representation of $\mathbf{f}_A$ is locally Lipschitz continuous and defined on an open set, it follows that $\mathbf{f}_A$ is well-defined and locally Lipschitz continuous on some neighborhood of $(t^*, \mathbf{x}^*, \mathbf{w}^*)$. Thus, there exists some sufficiently small $t_A \in$

$(t^*, b) \cap N_a$ for which the ODE:

$$\frac{d\mathbf{x}_A}{dt}(t) = \mathbf{f}_A(t, \mathbf{x}_A(t), \mathbf{w}(t)), \qquad \mathbf{x}_A(t^*) = \mathbf{x}^* \tag{6.8}$$

has a unique solution $\mathbf{x}_A$ on $[t^*, t_A]$. Since $p_{\mathbf{f}_A} = k - 1$ by construction, the inductive assumption shows that $\mathbf{x}_A$ is right-analytic at t^*. Moreover, for each $t \in (t^*, t_A]$, since $u_{(\lambda)}(t, \mathbf{x}(t), \mathbf{w}(t)) = u_{(\lambda)}(t, \mathbf{x}(t), \tilde{\mathbf{w}}(t)) > 0$, it follows that

$$\begin{aligned} v_{(\lambda)}(t, \mathbf{x}(t), \mathbf{w}(t)) &= |u_{(\lambda)}(t, \mathbf{x}(t), \mathbf{w}(t))| \\ &= u_{(\lambda)}(t, \mathbf{x}(t), \mathbf{w}(t)) = v_{A,(\lambda)}(t, \mathbf{x}(t), \mathbf{w}(t)). \end{aligned}$$

Thus, $\mathbf{f}(t, \mathbf{x}(t), \mathbf{w}(t)) = \mathbf{f}_A(t, \mathbf{x}(t), \mathbf{w}(t))$ for each $t \in (t^*, t_A]$, and so $\mathbf{x}$ solves the ODE (6.8) on $[t^*, t_A]$. The uniqueness of any solution of (6.8) yields $\mathbf{x} \equiv \mathbf{x}_A$ on $[t^*, t_A]$, which shows that $\mathbf{x}$ is right-analytic at t^*.

If $u_{(\lambda)}(t^*, \mathbf{x}^*, \mathbf{w}^*) < 0$, then define an abs-factorable function $\mathbf{f}_B$ as having the same factored representation as $\mathbf{f}$, except with $\psi_{B,(\lambda)} : u \to -u$. A similar argument to the previous case shows that $\mathbf{f}_B$ is well-defined on some neighborhood of $(t^*, \mathbf{x}^*, \mathbf{w}^*)$, and that, for sufficiently small $t_B \in (t^*, t_u)$, $\mathbf{x}$ uniquely solves the ODE:

$$\frac{d\mathbf{x}_B}{dt}(t) = \mathbf{f}_B(t, \mathbf{x}_B(t), \mathbf{w}(t)), \qquad \mathbf{x}_B(t^*) = \mathbf{x}^* \tag{6.9}$$

on $[t^*, t_B]$. By the inductive assumption, therefore, $\mathbf{x}$ is right-analytic at t^*.

If $u_{(\lambda)}(t^*, \mathbf{x}^*, \mathbf{w}^*) = 0$, then observe that, with $\mathbf{f}_A$ and $\mathbf{f}_B$ described as in the previous cases, both $\mathbf{f}_A$ and $\mathbf{f}_B$ are well-defined on some neighborhood N_{ab} of $(t^*, \mathbf{x}^*, \mathbf{w}^*)$, and that there exist unique solutions $\mathbf{x}_A$ and $\mathbf{x}_B$ of (6.8) and (6.9) on $[t^*, t_1]$ for some sufficiently small $t_1 \in (t^*, t_u)$. Noting that $u_{(\lambda)}$, $u_{A,(\lambda)}$, and $u_{B,(\lambda)}$ are equivalent mappings by construction, Lemma 6.2.7 and the inductive assumption show that the mappings $t \mapsto u_{(\lambda)}(t, \mathbf{x}_A(t), \mathbf{w}(t))$ and $t \mapsto u_{(\lambda)}(t, \mathbf{x}_B(t), \mathbf{w}(t))$ are both right-analytic at t^*. Thus, Lemma 6.2.4 shows that there exists $\beta \in (t^*, t_1) \cap (t^*, b)$ and $\sigma_A, \sigma_B \in \{-1, 0, +1\}$ for which

$$\operatorname{sign} u_{(\lambda)}(t, \mathbf{x}_D(t), \mathbf{w}(t)) = \sigma_D, \qquad \forall t \in (t^*, \beta], \quad \forall D \in \{A, B\}.$$

The following cases are exhaustive, and will be considered separately: $\sigma_A \geq 0$, $\sigma_B \leq 0$, and $(\sigma_A, \sigma_B) = (-1, +1)$.

If $\sigma_A \geq 0$, then $u_{(\lambda)}(t, \mathbf{x}_A(t), \mathbf{w}(t)) \geq 0$ for each $t \in (t^*, \beta]$. In this case, for each $t \in (t^*, \beta]$,

$$\begin{aligned} v_{(\lambda)}(t, \mathbf{x}_A(t), \tilde{\mathbf{w}}(t)) &= |u_{(\lambda)}(t, \mathbf{x}_A(t), \mathbf{w}(t))| \\ &= u_{(\lambda)}(t, \mathbf{x}_A(t), \mathbf{w}(t)) = v_{A,(\lambda)}(t, \mathbf{x}_A(t), \mathbf{w}(t)), \end{aligned}$$

and so $\mathbf{f}(t, \mathbf{x}_A(t), \tilde{\mathbf{w}}(t)) = \mathbf{f}_A(t, \mathbf{x}_A(t), \mathbf{w}(t))$ for each $t \in (t^*, \beta]$. Thus, $\mathbf{x}_A$ solves the ODE:

$$\frac{d\boldsymbol{\xi}}{dt}(t) = \mathbf{f}(t, \boldsymbol{\xi}(t), \tilde{\mathbf{w}}(t)), \qquad \boldsymbol{\xi}(t^*) = \mathbf{x}^*, \tag{6.10}$$

on $[t^*, \beta]$, as does $\mathbf{x}$, by inspection. Since $\mathbf{f}$ is locally Lipschitz continuous, uniqueness of any solution of (6.10) then implies that $\mathbf{x}_A \equiv \mathbf{x}$ on $[t^*, \beta]$. By the inductive assumption, $\mathbf{x}_A$ is right-analytic at t^*, implying that $\mathbf{x}$ is also right-analytic at t^*.

If $\sigma_B \leq 0$, then a similar argument shows that $\mathbf{x}_B \equiv \mathbf{x}$ on $[t^*, \beta]$, and that $\mathbf{x}$ is right-analytic at t^*.

The case in which $(\sigma_A, \sigma_B) = (-1, +1)$ does not occur. To show this, suppose, to obtain a contradiction, that both $\sigma_A = -1$ and $\sigma_B = +1$. Hence,

$$u_{(\lambda)}(t, \mathbf{x}_A(t), \tilde{\mathbf{w}}(t)) < 0 < u_{(\lambda)}(t, \mathbf{x}_B(t), \tilde{\mathbf{w}}(t)), \qquad \forall t \in (t^*, \beta],$$

which implies that $\mathbf{x}_A(t) \neq \mathbf{x}_B(t)$ for each $t \in (t^*, \beta]$. Define an abs-factorable function $\bar{\mathbf{f}}$ as having the same factored representation as $\mathbf{f}$, except with $\bar{\psi}_{(\lambda)} : u \to -|u|$. Let quantities relating to $\bar{\mathbf{f}}$ be denoted with overbars. Clearly $\bar{\mathbf{f}}(t^*, \mathbf{x}^*, \mathbf{w}^*) = \mathbf{f}(t^*, \mathbf{x}^*, \mathbf{w}^*)$, and so $\bar{\mathbf{f}}$ is well-defined and locally Lipschitz continuous on a neighborhood of $(t^*, \mathbf{x}^*, \mathbf{w}^*)$. Thus, for sufficiently small $\alpha \in (t^*, \beta)$, for each $t \in (t^*, \alpha]$,

$$
\begin{aligned}
\bar{v}_{(\lambda)}(t, \mathbf{x}_A(t), \tilde{\mathbf{w}}(t)) &= -|u_{(\lambda)}(t, \mathbf{x}_A(t), \mathbf{w}(t))| \\
&= u_{(\lambda)}(t, \mathbf{x}_A(t), \mathbf{w}(t)) = v_{A,(\lambda)}(t, \mathbf{x}_A(t), \mathbf{w}(t)); \\
\bar{v}_{(\lambda)}(t, \mathbf{x}_B(t), \tilde{\mathbf{w}}(t)) &= -|u_{(\lambda)}(t, \mathbf{x}_B(t), \mathbf{w}(t))| \\
&= -u_{(\lambda)}(t, \mathbf{x}_B(t), \mathbf{w}(t)) = v_{B,(\lambda)}(t, \mathbf{x}_B(t), \mathbf{w}(t)).
\end{aligned}
$$

Consequently, for each $t \in (t^*, \alpha]$, both $\bar{\mathbf{f}}(t, \mathbf{x}_A(t), \tilde{\mathbf{w}}(t)) = \mathbf{f}_A(t, \mathbf{x}_A(t), \mathbf{w}(t))$ and $\bar{\mathbf{f}}(t, \mathbf{x}_B(t), \tilde{\mathbf{w}}(t)) = \mathbf{f}_B(t, \mathbf{x}_B(t), \mathbf{w}(t))$. It follows that $\mathbf{x}_A$ and $\mathbf{x}_B$ both solve the ODE:

$$
\frac{d\bar{\mathbf{x}}}{dt}(t) = \bar{\mathbf{f}}(t, \bar{\mathbf{x}}(t), \tilde{\mathbf{w}}(t)), \qquad \bar{\mathbf{x}}(t^*) = \mathbf{x}^*
$$

on $[t^*, \alpha]$, contradicting the uniqueness of any solution of this ODE on $[t^*, \alpha]$, which follows from the local Lipschitz continuity of $\bar{\mathbf{f}}$. Therefore, it cannot be that both $\sigma_A = -1$ and $\sigma_B = +1$.

The above cases cover all possible situations, and thereby complete the inductive step. □

Next, the following corollary demonstrates non-Zenoness in the sense of [86], and motivates the subsequent definition.

Corollary 6.3.10. *Suppose that Assumption 6.3.1 holds. For each $t^* \in [t_0, t_f]$ and each $i \in \{1, \ldots, p_{\mathbf{f}}\}$, there exist $\sigma_i^L(t^*), \sigma_i^R(t^*) \in \{-1, 0, +1\}$ such that for sufficiently small $\delta > 0$,*

$$
\begin{aligned}
& \sigma_i(t) = \sigma_i^L(t^*), \qquad \forall t \in [t^* - \delta, t^*), \\
\textit{and} \qquad & \sigma_i(t) = \sigma_i^R(t^*), \qquad \forall t \in (t^*, t^* + \delta].
\end{aligned}
$$

Proof. By Lemma 6.2.7 and Theorem 6.3.9, the mapping $t \mapsto u_{(\lambda_{\mathbf{f}}(i))}(t, \mathbf{x}(t), \mathbf{w}(t))$ is left/right-analytic at any $t^* \in [t_0, t_f]$ for each $i \in \{1, \ldots, p_{\mathbf{f}}\}$. The required result then follows from Lemma 6.2.4 and the definition of each σ_i. □

Definition 6.3.11. *Suppose that Assumption 6.3.1 holds. For each $t^* \in [t_0, t_f]$ and each $i \in \{1, \ldots, p_{\mathbf{f}}\}$, define a* left-signature *$\sigma_i^L(t^*) \in \{-1, 0, +1\}$ and a* right-signature *$\sigma_i^R(t^*) \in \{-1, 0, +1\}$ as described in Corollary 6.3.10.*

Left/right-signatures play a similar role to the Lie derivatives considered in [72]. The concept of left/right-signatures permits a more intuitive characterization of valley-crossings and crossing opportunities, as described by the following two lemmata.

Lemma 6.3.12. *Suppose that Assumption 6.3.1 or 6.3.2 holds. There is a valley-crossing at $t^* \in [t_0, t_f]$ if and only if there exists some $i \in \{1, \ldots, p_{\mathbf{f}}\}$ such that $(\sigma_i^L(t^*), \sigma_i^R(t^*)) \in \{(-1, +1), (+1, -1)\}$. If Assumption 6.3.2 holds, then $\sigma_i(t^*) = 0$ at any valley-crossing t^* of $\mathbf{x}$.*

Proof. The first result follows immediately from Corollary 6.3.10 and the definition of a valley-crossing. If Assumption 6.3.2 holds, then the second result follows from the first result and the continuity of $t \mapsto u_{(\lambda_{\mathbf{f}}(i))}(t, \mathbf{x}(t))$ at t^*. □

Lemma 6.3.13. *Suppose that Assumption 6.3.1 or 6.3.2 holds. There is a crossing opportunity of $\mathbf{x}$ at $t^* \in [t_0, t_f]$ if and only if there exists some $i \in \{1, \ldots, p_{\mathbf{f}}\}$ such that*

$$(\sigma_i^L(t^*), \sigma_i^R(t^*)) \in \{(-1, 0), (-1, +1), (0, -1), \\ (0, +1), (+1, -1), (+1, 0)\}.$$

If Assumption 6.3.2 holds, then $\sigma_i(t^*) = 0$ at any crossing opportunity t^* of $\mathbf{x}$.

Proof. Comparing Corollary 6.3.10 and the definition of a crossing opportunity, there is a crossing opportunity at t^* if and only if $\sigma_i^L(t^*) \neq \sigma_i^R(t^*)$ for some $i \in \{1, \ldots, p_{\mathbf{f}}\}$, which is equivalent to the first required result. If Assumption 6.3.2 holds, then the second result follows from the first result and the continuity of $t \mapsto u_{(\lambda_{\mathbf{f}}(i))}(t, \mathbf{x}(t))$ at t^*. □

The following theorem essentially rules out the emergence of the Zeno automata illustrated in [48, Section 3.1] in the abs-factorable ODEs considered in this work.

Theorem 6.3.14. *Suppose that Assumption 6.3.1 holds. There are finitely many valley-crossings and crossing opportunities of $\mathbf{x}$ in $[t_0, t_f]$.*

Proof. Since any valley-crossing is also a crossing opportunity, it suffices to show that the set B of crossing opportunities in $[t_0, t_f]$ is finite. If, instead, B were infinite, then the compactness of $[t_0, t_f]$ would imply that B has a limit point, which would itself be an element of B due to Lemma 6.3.8. Hence, it suffices to show that each element of B is isolated.

If B is empty, then the theorem is trivially satisfied. Otherwise, choose any $t^* \in B$. Corollary 6.3.10 shows that for some $\delta > 0$, for each $i \in \{1, \ldots, p_{\mathbf{f}}\}$, σ_i is constant on $(t^*, t^* + \delta)$ and on $(t^* - \delta, t^*)$. This, in conjunction with Lemma 6.3.13, shows that the open intervals $(t^*, t^* + \delta)$ and $(t^* - \delta, t^*)$ do not contain any crossing opportunities. Hence, t^* is the only element of $B \cap (t^* - \delta, t^* + \delta)$, and is therefore isolated. □

Corollary 6.3.15. *Suppose that Assumption 6.3.1 holds. There exists $\epsilon > 0$ such that, for each $t^* \in [t_0, t_f]$, $\mathbf{x}$ has at most one crossing opportunity or valley-crossing on the set $[t^*, t^* + \epsilon] \cap [t_0, t_f]$.*

The remaining results in this subsection show that, under Assumption 6.3.2, the mappings $t \mapsto u_{(\lambda_{\mathbf{f}}(i))}(t, \mathbf{x}(t))$ are analytic between successive crossing opportunities, and each crossing opportunity for the trajectory $\mathbf{x}$ must also be a valley-crossing. Under Assumption 6.3.2, let B denote the set of crossing opportunities of $\mathbf{x}$ in $[t_0, t_f]$. In light of Theorem 6.3.14, the set $\bar{B} := B \cup \{t_0, t_f\}$ will be enumerated in the following two lemmata as $\{\tau_k\}_{k=0}^{q_{\mathbf{f}}}$, where $t_0 = \tau_0 < \tau_1 < \tau_2 < \ldots < \tau_{q_{\mathbf{f}}} = t_f$, for some $q_{\mathbf{f}} \in \mathbb{N}$.

Lemma 6.3.16. *Suppose that Assumption 6.3.2 holds. For each $k \in \{1, \ldots, q_{\mathbf{f}}\}$ and each $i \in \{1, \ldots, p_{\mathbf{f}}\}$, there exist $a, b \in T$ and an analytic function $\tilde{u}_{k,i} : (a, b) \to \mathbb{R}$ such that $a < \tau_{k-1} < \tau_k < b$, and $u_{(\lambda_{\mathbf{f}}(i))}(t, \mathbf{x}(t)) \equiv \tilde{u}_{k,i}(t)$ for each $t \in [\tau_{k-1}, \tau_k]$.*

Proof. Choose some particular $k \in \{1, \ldots, q_{\mathbf{f}}\}$ and $i \in \{1, \ldots, p_{\mathbf{f}}\}$. Since there are no crossing opportunities of $\mathbf{x}$ in (τ_{k-1}, τ_k), there cannot exist $t^*, t' \in [\tau_{k-1}, \tau_k]$ for which both $\sigma_i(t^*) = +1$ and $\sigma_i(t') = -1$. To show this, suppose, to obtain a contradiction, that such t^* and t' do exist. Noting that $u_{(\lambda_{\mathbf{f}}(i))}(t^*, \mathbf{x}(t^*)) > 0$, and that $u_{(\lambda_{\mathbf{f}}(i))}$ and $\mathbf{x}$ are continuous on their respective domains, there exists

a nondegenerate interval $[\alpha, \beta]$ that is a connected component of the closed set $P := \{t \in [\tau_{k-1}, \tau_k] : u_{(\lambda_{\mathbf{f}}(i))}(t, \mathbf{x}(t)) \geq 0\}$ containing t^*. Since $t' \notin P$, it follows that either $\alpha \neq \tau_{k-1}$ or $\beta \neq \tau_k$. Suppose that $\beta \neq \tau_k$; the case in which $\alpha \neq \tau_{k-1}$ is analogous. By definition of β, for each $\gamma \in (\beta, \tau_k]$, there must exist $t \in [\beta, \gamma]$ for which $u_{(\lambda_{\mathbf{f}}(i))}(t, \mathbf{x}(t)) < 0$. (Otherwise, $[\alpha, \gamma]$ would be a larger connected subset of P than $[\alpha, \beta]$.) Hence, $\sigma_i^L(\beta) \geq 0$, and $\sigma_i^R(\beta) = -1$, and so β is a crossing opportunity. This contradicts the established inequality $\tau_{k-1} < \beta < \tau_k$ and the construction of τ_{k-1} and τ_k.

It therefore follows that either $\sigma_i(t) \geq 0$ for all $t \in [\tau_{k-1}, \tau_k]$, or $\sigma_i(t) \leq 0$ for all $t \in [\tau_{k-1}, \tau_k]$. Noting that $i \in \{1, \ldots, p_{\mathbf{f}}\}$ was chosen arbitrarily, there exists $\mathbf{s} \in \{-1, +1\}^{p_{\mathbf{f}}}$ such that

$$v_{(\lambda_{\mathbf{f}}(i))}(t, \mathbf{x}(t)) = s_i \, u_{(\lambda_{\mathbf{f}}(i))}(t, \mathbf{x}(t)), \qquad \forall t \in [\tau_{k-1}, \tau_k], \quad \forall i \in \{1, \ldots, p_{\mathbf{f}}\}. \tag{6.11}$$

Thus, define an abs-factorable function $\bar{\mathbf{f}}$ as having the same factorable representation as $\mathbf{f}$, except with $\bar{\psi}_{(\lambda_{\mathbf{f}}(i))} : u \mapsto s_i u$ replacing $\psi_{(\lambda_{\mathbf{f}}(i))} : u \mapsto |u|$ for each $i \in \{1, \ldots, p_{\mathbf{f}}\}$. Since $\bar{\mathbf{f}}(t, \mathbf{x}(t)) = \mathbf{f}(t, \mathbf{x}(t))$ for each $t \in [\tau_{k-1}, \tau_k]$, and since each elemental function $\bar{\psi}_{(j)}$ in the factored representation of $\bar{\mathbf{f}}$ is locally Lipschitz continuous and defined on an open set, it follows that $\bar{\mathbf{f}}$ is well-defined and locally Lipschitz continuous on some open superset of $\{(t, \mathbf{x}(t)) : t \in [\tau_{k-1}, \tau_k]\}$. Thus, $\mathbf{x}$ is the unique solution on $[\tau_{k-1}, \tau_k]$ of the ODE:

$$\frac{d\bar{\mathbf{x}}}{dt}(t) = \bar{\mathbf{f}}(t, \bar{\mathbf{x}}(t)), \qquad \bar{\mathbf{x}}(\tau_{k-1}) = \mathbf{x}(\tau_{k-1}). \tag{6.12}$$

Noting that $\bar{\mathbf{f}}$ is analytic by construction, it follows that the above ODE has a unique analytic solution $\bar{\mathbf{x}}$ on some open superset of $[\tau_{k-1}, \tau_k]$, which implies that $\mathbf{x} \equiv \bar{\mathbf{x}}$ on $[\tau_{k-1}, \tau_k]$. Since $u_{\lambda_{\mathbf{f}}(i)}(t, \mathbf{x}(t)) = \bar{u}_{(\lambda_{\mathbf{f}}(i))}(t, \bar{\mathbf{x}}(t))$ for each $t \in [\tau_{k-1}, \tau_k]$ and each $i \in \{1, \ldots, p_{\mathbf{f}}\}$, since each $\bar{u}_{(\lambda_{\mathbf{f}}(i))}$ is analytic by construction, and since $t \mapsto \bar{u}_{(\lambda_{\mathbf{f}}(i))}(t, \bar{\mathbf{x}}(t))$ is defined on some open superset of $[\tau_{k-1}, \tau_k]$, the required result follows. $\square$

Lemma 6.3.17. *Suppose that Assumption 6.3.2 holds. For each $k \in \{1, \ldots, q_{\mathbf{f}}\}$ and each $i \in \{1, \ldots, p_{\mathbf{f}}\}$, the set*

$$S_{k,i} := \{t \in [\tau_{k-1}, \tau_k] : u_{(\lambda_{\mathbf{f}}(i))}(t, \mathbf{x}(t)) = 0\}$$

is either finite or equal to $[\tau_{k-1}, \tau_k]$.

Proof. Employing the analytic function $\tilde{u}_{k,i}$ provided by Lemma 6.3.16, observe that $S_{k,i} = \{t \in [\tau_{k-1}, \tau_k] : \tilde{u}_{k,i}(t) = 0\}$. The continuity of $\tilde{u}_{k,i}$ implies that $S_{k,i}$ is closed. If $S_{k,i}$ is finite, then the required result is trivially satisfied. Thus, suppose that $S_{k,i}$ is infinite. The compactness of $[\tau_{k-1}, \tau_k]$ implies that $S_{k,i}$ has a limit point, which must itself be an element of $S_{k,i}$ since $S_{k,i}$ is closed. In this case, [66, Corollary 1.2.7] implies that $\tilde{u}_{k,i} \equiv 0$, and so $S_{k,i} = [\tau_{k-1}, \tau_k]$. □

The following theorem demonstrates the converse of the first statement in Remark 6.3.6, under Assumption 6.3.2.

Theorem 6.3.18. *Suppose that Assumption 6.3.2 holds. Every crossing opportunity of $\mathbf{x}$ in $[t_0, t_f]$ is also a valley-crossing.*

Proof. To obtain a contradiction, suppose there exists a crossing opportunity $t^* \in [t_0, t_f]$ of $\mathbf{x}$ that is not also a valley-crossing, and set $\mathbf{x}^* := \mathbf{x}(t^*)$. Lemmata 6.3.12 and 6.3.13 imply that for some $i^* \in \{1, \ldots, p_{\mathbf{f}}\}$,

$$(\sigma^L_{i^*}(t^*), \sigma^R_{i^*}(t^*)) \in \{(-1, 0), (+1, 0), (0, -1), (0, +1)\}. \tag{6.13}$$

Since t^* is not a valley-crossing, Lemma 6.3.7 shows that $\mathbf{x}$ is analytic at t^*. Consider the analytic function $\tilde{\mathbf{f}}_{\mathbf{s}}$ described in the proof of Lemma 6.3.7, and let quantities related to the factorable representation of $\tilde{\mathbf{f}}_{\mathbf{s}}$ be denoted with tildes. It follows that $\mathbf{x}$ solves the ODE (6.7) on some neighborhood $N \subset T$ of t^*. Thus, $u_{(\lambda_{\mathbf{f}}(i^*))}(t, \mathbf{x}(t)) = \tilde{u}_{(\lambda_{\mathbf{f}}(i^*))}(t, \mathbf{x}(t))$ for each $t \in N$, and so $t \mapsto u_{(\lambda_{\mathbf{f}}(i^*))}(t, \mathbf{x}(t))$ is analytic at t^*. Equation (6.13) also shows that either $\sigma^L_{i^*}(t^*) = 0$ or $\sigma^R_{i^*}(t^*) = 0$; [66, Corollary 1.2.7] then implies that $u_{(\lambda_{\mathbf{f}}(i^*))}(t, \mathbf{x}(t)) = 0$ for all $t \in N$, which contradicts (6.13). □

Corollary 6.3.19. *Suppose that Assumption 6.3.2 holds. If there exist* $t^* \in [t_0, t_f]$ *and* $i \in \{1, \ldots, p_{\mathbf{f}}\}$ *such that*

$$(\sigma_i^L(t^*), \sigma_i^R(t^*)) \in \{(-1,0), (+1,0), (0,-1), (0,+1)\},$$

then there exists $j \in \{1, \ldots, p_{\mathbf{f}}\}$ *such that* $j \neq i$, $\sigma_j(t^*) = 0$, *and*

$$(\sigma_j^L(t^*), \sigma_j^R(t^*)) \in \{(-1,+1), (+1,-1)\}.$$

Proof. The corollary is a direct consequence of Theorem 6.3.18 and Lemmata 6.3.12 and 6.3.13. □

6.3.3 ODEs with linear programs embedded

This section extends Theorem 6.3.14 to the systems considered in [36, 44], in which an ODE right-hand side depends on the optimal costs of certain standard-form linear programs (LPs), expressed in terms of a varying right-hand side constraint vector. Such functions are formalized in the following definitions, which make use of results from LP sensitivity theory [10, Section 5.2].

Definition 6.3.20. *For any vectors* $\mathbf{c} \in \mathbb{R}^n$ *and* $\mathbf{b} \in \mathbb{R}^m$, *and any matrix* $\mathbf{A} \in \mathbb{R}^{m \times n}$, *let* $LP_{\mathbf{A},\mathbf{c}}(\mathbf{b})$ *denote the standard-form LP:*

$$\inf_{\mathbf{z} \in \mathbb{R}^n} \mathbf{c}^{\mathrm{T}}\mathbf{z}, \qquad \textit{subject to} \quad \mathbf{A}\mathbf{z} = \mathbf{b}, \quad \mathbf{z} \geq \mathbf{0}.$$

Definition 6.3.21. *Consider* $\mathbf{c} \in \mathbb{R}^n$ *and* $\mathbf{A} \in \mathbb{R}^{m \times n}$ *such that* $\mathbf{A}$ *has linearly independent rows, and the* dual-feasible *set* $DF_{\mathbf{A},\mathbf{c}} := \{\mathbf{p} \in \mathbb{R}^m : \mathbf{p}^{\mathrm{T}}\mathbf{A} \leq \mathbf{c}^{\mathrm{T}}\}$ *is nonempty. Define the* dual-extreme *set* $DE_{\mathbf{A},\mathbf{c}}$ *as the set of extreme points of* $DF_{\mathbf{A},\mathbf{c}}$. *Define the* dual-optimal and extreme *set* $DO_{\mathbf{A},\mathbf{c}}(\mathbf{b}) := \arg\max\{\mathbf{p}^{\mathrm{T}}\mathbf{b} : \mathbf{p} \in DE_{\mathbf{A},\mathbf{c}}\}$ *for each* $\mathbf{b} \in \mathbb{R}^m$, *and define the* LP-mapping $q_{\mathbf{A},\mathbf{c}} : \mathbb{R}^m \to \mathbb{R} : \mathbf{b} \mapsto \max\{\mathbf{p}^{\mathrm{T}}\mathbf{b} : \mathbf{p} \in DE_{\mathbf{A},\mathbf{c}}\}$.

Theorem 2.6 in [10] implies that $DE_{\mathbf{A},\mathbf{c}}$ is finite and nonempty, and that $q_{\mathbf{A},\mathbf{c}}$ is well-defined. Thus, $q_{\mathbf{A},\mathbf{c}}$ is convex and piecewise linear in the sense of Scholtes [97]. Strong duality for LPs implies that if $LP_{\mathbf{A},\mathbf{c}}(\mathbf{b})$ has a finite solution value, then $q_{\mathbf{A},\mathbf{c}}(\mathbf{b})$ is the solution value for $LP_{\mathbf{A},\mathbf{c}}(\mathbf{b})$.

Lemma 6.3.22. *Any LP-mapping $q_{\mathbf{A},\mathbf{c}} : \mathbb{R}^m \to \mathbb{R}$ is abs-factorable.*

Proof. For any $\mathbf{b} \in \mathbb{R}^m$,

$$\begin{aligned}
&q_{\mathbf{A},\mathbf{c}}(\mathbf{b}) \\
&= \max\{\mathbf{p}^{\mathrm{T}}\mathbf{b} : \mathbf{p} \in DE_{\mathbf{A},\mathbf{c}}\}, \\
&= \max\{\max\{\mathbf{p}^{\mathrm{T}}\mathbf{b}, \mathbf{q}^{\mathrm{T}}\mathbf{b}\} : \mathbf{p}, \mathbf{q} \in DE_{\mathbf{A},\mathbf{c}}\}, \\
&= \max\{\tfrac{1}{2}(\mathbf{p}+\mathbf{q})^{\mathrm{T}}\mathbf{b} + \tfrac{1}{2}|(\mathbf{p}-\mathbf{q})^{\mathrm{T}}\mathbf{b}| : \mathbf{p}, \mathbf{q} \in DE_{\mathbf{A},\mathbf{c}}\}. \qquad (6.14)
\end{aligned}$$

Since the set $DE_{\mathbf{A},\mathbf{c}}$ is finite, the outer 'max' operation in (6.14) can be represented as a finite composition of bivariate 'max' operations, which may in turn be expressed as a composition of linear functions and absolute-value functions. □

An explicit abs-factorable representation of an LP-mapping would be difficult to construct in practice, and would be intractable to use computationally. Nevertheless, the existence of such a representation shows that the non-Zenoness results developed in this work remain applicable if LP-mappings are added to the elemental library $\mathcal{L}$ used to define the abs-factorable ODE right-hand side $\mathbf{f}$. Let $\bar{\mathcal{L}}$ denote this augmented elemental library, and let an *$\bar{\mathcal{L}}$-factorable function* denote a factorable function defined in terms of the elemental library $\bar{\mathcal{L}}$. The following corollary establishes a pertinent restatement of the obtained non-Zenoness results for ODEs with $\bar{\mathcal{L}}$-factorable right-hand side functions.

Corollary 6.3.23. *Suppose that Assumption 6.3.1 holds, except with $\mathbf{f}$ being $\bar{\mathcal{L}}$-factorable instead of abs-factorable. Suppose there exists $j \in \{1, \ldots, \ell\}$ for which $\boldsymbol{\psi}_{(j)}$ is an LP-mapping $q_{\mathbf{A},\mathbf{c}}$. For each $t^* \in [t_0, t_f]$, there exists $\delta > 0$ such that the set-valued mapping $t \mapsto DO_{\mathbf{A},\mathbf{c}}(\mathbf{u}_{(j)}(t, \mathbf{x}(t), \mathbf{w}(t)))$ is constant on $(t^* - \delta, t^*)$ (with value $DO^L_{\mathbf{A},\mathbf{c}}(t^*) \subset DE_{\mathbf{A},\mathbf{c}}$) and on $(t^*, t^* + \delta)$ (with value $DO^R_{\mathbf{A},\mathbf{c}}(t^*) \subset DE_{\mathbf{A},\mathbf{c}}$). Moreover, there exists a finite set $Z_T \subset [t_0, t_f]$ such that if $t^* \notin Z_T$, then $DO^L_{\mathbf{A},\mathbf{c}}(t^*) = DO^R_{\mathbf{A},\mathbf{c}}(t^*)$.*

Proof. Since $\mathbf{f}$ is abs-factorable, Corollary 6.3.10, Lemma 6.3.13, and Theorem 6.3.14 apply. Thus, it suffices to show that for some particular factored representation of $\mathbf{f}$ using only the elemental library $\mathcal{L}$ used to define abs-factorable functions,

$DO_{\mathbf{A},\mathbf{c}}(\mathbf{u}_{(j)}(t,\mathbf{x}(t),\mathbf{w}(t)))$ is uniquely determined for each $t \in T$ by the quantities $\{\sigma_i(t)\}_{i=1}^{p_{\mathbf{f}}}$. Such a factored representation of $\mathbf{f}$ is readily generated by replacing each LP-mapping in the factored representation of $\mathbf{f}$ with the equivalent composition of linear functions and absolute-value functions considered in the proof of Lemma 6.3.22. Let overbars denote quantities relating to this new factored representation, and choose $\bar{j} \in \{1,\ldots,\bar{\ell}\}$ such that $\bar{\mathbf{u}}_{(\bar{j})} \equiv \mathbf{u}_{(j)}$. Observe that, for each $t \in T$, with $\mathbf{u} := \bar{\mathbf{u}}_{(\bar{j})}(t,\mathbf{x}(t),\mathbf{w}(t))$, $\mathbf{p} \in DO_{\mathbf{A},\mathbf{c}}(\mathbf{u})$ if and only if both $\operatorname{sign}(\mathbf{p}-\mathbf{q})^{\mathrm{T}}\mathbf{u} = 0$ for each $\mathbf{q} \in DO_{\mathbf{A},\mathbf{c}}(\mathbf{u})$, and $\operatorname{sign}(\mathbf{p}-\mathbf{q})^{\mathrm{T}}\mathbf{u} = +1$ for each $\mathbf{q} \in DE_{\mathbf{A},\mathbf{c}} \backslash DO_{\mathbf{A},\mathbf{c}}(\mathbf{u})$. Moreover, (6.14) shows that there exists $\bar{i} \in \{1,\ldots,\bar{\ell}\}$ such that $\bar{\psi}_{(\bar{i})} \equiv |\cdot|$ and $\bar{\sigma}_{\bar{i}}(t) = \operatorname{sign}(\mathbf{p}-\mathbf{q})^{\mathrm{T}}\mathbf{u}$. Thus, $DO_{\mathbf{A},\mathbf{c}}(\bar{\mathbf{u}}_{(\bar{j})}(t,\mathbf{x}(t),\mathbf{w}(t)))$, which is equal to $DO_{\mathbf{A},\mathbf{c}}(\mathbf{u}_{(j)}(t,\mathbf{x}(t),\mathbf{w}(t)))$ by construction, is uniquely determined for each $t \in T$ by the quantities $\{\bar{\sigma}_k(t)\}_{k=1}^{\bar{p}_{\mathbf{f}}}$, as required. □

6.4 Necessary conditions for valley-tracing modes

The main theorem of this section combines several of the results obtained in the previous section under Assumption 6.3.2, to describe necessary conditions for the emergence of valley-tracing modes (cf. Definition 6.3.5) in the solution trajectory $\mathbf{x}$ on $[t_0, t_f]$ of the ODE (6.3). Equivalently, failure of any of these conditions is sufficient to conclude that there are no valley-tracing modes of $\mathbf{x}$ on $[t_0, t_f]$. In practice, these conditions can be checked while solving (6.3) numerically. Moreover, the corollaries and examples throughout this section illustrate situations in which some of these conditions can be seen not to hold *a priori*, in which case there can be no valley-tracing modes. The theorem makes use of the classical directional derivative, which is defined as follows.

Definition 6.4.1. *Given an open set $Y \subset \mathbb{R}^n$ and some $\mathbf{y} \in Y$, a function $\mathbf{g} : Y \to \mathbb{R}^m$ is* directionally differentiable *at $\mathbf{y}$ if the following* directional derivative *exists and is finite for each $\mathbf{d} \in \mathbb{R}^n$:*

$$\mathbf{g}'(\mathbf{y};\mathbf{d}) := \lim_{t \to 0^+} \frac{\mathbf{g}(\mathbf{y}+t\mathbf{d}) - \mathbf{g}(\mathbf{y})}{t} \in \mathbb{R}^m.$$

As shown in [32], any abs-factorable function is directionally differentiable, and its directional derivatives may be evaluated numerically using a simple extension of the standard forward mode of automatic differentiation.

Theorem 6.4.2. *Suppose that Assumption 6.3.2 holds. For each $i \in \{1, \ldots, p_{\mathbf{f}}\}$, the set $Z_i := \{t \in [t_0, t_f] : \sigma_i(t) = 0\}$ is the union of finitely many points and intervals that are disjoint, compact, and nondegenerate. Any such interval $[a, b] \subset [t_0, t_f]$ satisfies all of the following conditions:*

1. $\sigma_i(a) = 0$ *and* $\sigma_i(b) = 0$,
2. $[u_{(\lambda_{\mathbf{f}}(i))}]'\left(\begin{bmatrix} a \\ \mathbf{x}(a) \end{bmatrix}; \begin{bmatrix} 1 \\ \mathbf{f}(a, \mathbf{x}(a)) \end{bmatrix}\right) = 0$,
3. $[u_{(\lambda_{\mathbf{f}}(i))}]'\left(\begin{bmatrix} b \\ \mathbf{x}(b) \end{bmatrix}; \begin{bmatrix} -1 \\ -\mathbf{f}(b, \mathbf{x}(b)) \end{bmatrix}\right) = 0$,
4. *either $a = t_0$, or there exists $i_A \in \{1, \ldots, p_{\mathbf{f}}\}$ such that each of the following conditions is satisfied:*
   - $i_A \neq i$,
   - $\sigma_{i_A}(a) = 0$,
   - $(\sigma^L_{i_A}(a), \sigma^R_{i_A}(a)) \in \{(-1, +1), (+1, -1)\}$,
5. *either $b = t_f$, or there exists $i_B \in \{1, \ldots, p_{\mathbf{f}}\}$ such that each of the following conditions is satisfied:*
   - $i_B \neq i$,
   - $\sigma_{i_B}(b) = 0$,
   - $(\sigma^L_{i_B}(b), \sigma^R_{i_B}(b)) \in \{(-1, +1), (+1, -1)\}$.

If there do not exist $i \in \{1, \ldots, p_{\mathbf{f}}\}$ and $a, b \in [t_0, t_f]$ satisfying $a < b$ and all of the above conditions, then each Z_i is finite, and $\mathbf{x}$ does not have any valley-tracing modes in $[t_0, t_f]$.

Proof. Using the notation of Lemma 6.3.17, $Z_i = \bigcup_{k=1}^{q_{\mathbf{f}}} S_{k,i}$ for each $i \in \{1, \ldots, p_{\mathbf{f}}\}$. Application of Lemma 6.3.17 then yields the first assertion of the theorem.

Next, suppose that $[a, b] \subset [t_0, t_f]$ is one of the disjoint, closed, nondegenerate intervals comprising Z_i for some $i \in \{1, \ldots, p_{\mathbf{f}}\}$. By definition of σ_i and Z_i, it follows that $t \mapsto u_{(\lambda_{\mathbf{f}}(i))}(t, \mathbf{x}(t))$ is the zero function when restricted to $[a, b]$. Condition 1 of the theorem follows immediately, as does the assertion that

$$[u_{(\lambda_{\mathbf{f}}(i))} \circ \mathbf{z}]'(a; 1) = 0 = [u_{(\lambda_{\mathbf{f}}(i))} \circ \mathbf{z}]'(b; -1),$$

where $\mathbf{z}$ is the mapping $t \mapsto (t, \mathbf{x}(t))$. Conditions 2 and 3 of the theorem then follow from [97, Theorem 3.1.1].

To show that Condition 4 holds, suppose that $a \neq t_0$. By construction of a, $\sigma_i^R(a) = 0$. Since $[a, b]$ is a connected component of Z_i, $\sigma_i^L(a)$ must be nonzero, otherwise Corollary 6.3.10 would imply that for some $\delta > 0$, $\sigma_i(t) = 0$ for all $t \in [a - \delta, a]$, and thus for all $t \in [a - \delta, b]$, contradicting the definition of a. It follows that $(\sigma_i^L(a), \sigma_i^R(a)) \in \{(-1, 0), (+1, 0)\}$; Condition 4 of the theorem then follows from Corollary 6.3.19. Condition 5 is demonstrated analogously. The remaining claim of the theorem is the contrapositive of the claims demonstrated above. □

During numerical integration, it is difficult to detect the start or end of a valley-tracing mode $[a, b]$ using the definition of a valley-tracing mode directly, since this requires verifying numerically that some $u_{(\lambda_{\mathbf{f}}(i))}(\cdot, \mathbf{x}(\cdot))$ is identically zero on an interval. However, Condition 4 of the above theorem shows that if $a \neq t_0$, then a coincides with a sign-change of a *discontinuity function* $t \mapsto u_{(\lambda_{\mathbf{f}}(i_A))}(t, \mathbf{x}(t))$ with $i_A \neq i$. This sign-change can be detected during integration using standard event-detection techniques [89]. An analogous situation occurs at b if $b \neq t_f$.

The following corollary shows that, under Assumption 6.3.2, Clarke's sufficient condition [16, Theorem 7.4.1] for differentiability of an ODE solution with respect to the initial condition is satisfied when there are no valley-tracing modes.

Corollary 6.4.3. *Suppose that Assumption 6.3.2 holds. Let $S_{\mathbf{f}} \subset \bar{T} \times X$ be the set on which $\mathbf{f}$ is not continuously differentiable. If there are no valley-tracing modes of $\mathbf{x}$ on*

$[t_0, t_f]$, *then the set* $S_t := \{t \in [t_0, t_f] : (t, \mathbf{x}(t)) \in S_{\mathbf{f}}\}$ *has zero Lebesgue measure.*

Proof. By Theorem 6.4.2, the absence of valley-tracing modes implies that $Z_i := \{t \in [t_0, t_f] : \sigma_i(t) = 0\}$ is finite for each $i \in \{1, \ldots, p_{\mathbf{f}}\}$. Thus, the set $\bar{Z} := \bigcup_{i=1}^{p_{\mathbf{f}}} Z_i$ has zero measure. It therefore suffices to show that $S_t \subset \bar{Z}$. For any particular $t^* \in [t_0, t_f] \backslash \bar{Z}$, $\sigma_i(t^*) \neq 0$ for each $i \in \{1, \ldots, p_{\mathbf{f}}\}$, and so $\psi_{(j)}$ is analytic at $\mathbf{u}_{(j)}(t^*, \mathbf{x}(t^*))$ for each $j \in \{1, \ldots, \ell\}$. This shows that $\mathbf{f}$ is analytic (and thus continuously differentiable) at $(t^*, \mathbf{x}(t^*))$, which implies that $t^* \notin S_t$. Since t^* was chosen arbitrarily, it follows that $S_t \subset \bar{Z}$, as required. □

The following two examples illustrate the results of Theorem 6.4.2 when applied to simple nonsmooth ODEs that exhibit valley-tracing modes and can be solved analytically. Together, these examples illustrate that there is no requirement in Conditions 4 and 5 of Theorem 6.4.2 that $i_A, i_B > i$, or that $i_A, i_B < i$.

Example 6.4.4. *Consider the following ODE, with a single differential variable:*

$$\frac{dx}{dt}(t) = \big|x(t) - 1 - |x(t) - 1|\big| + 2x(t) - 1, \qquad x(0) = 0.$$

It is readily confirmed that this ODE is solved by the mapping:

$$x : t \mapsto \begin{cases} t, & \text{if } t \leq 1, \\ \frac{1}{2}(1 + e^{2(t-1)}), & \text{if } t > 1. \end{cases}$$

Moreover, since the ODE right-hand side function is abs-factorable, and therefore locally Lipschitz continuous, this solution is unique. Two absolute-value functions appear in a direct factored representation of the ODE right-hand side function f*, with* $u_{(\lambda_f(1))}$ *and* $u_{(\lambda_f(2))}$ *equivalent to the arguments of the inner and outer absolute-value functions in the ODE right-hand side, respectively. Thus, for all* $(t, z) \in \mathbb{R}^2$*,*

$$u_{(\lambda_f(1))}(t, z) = z - 1 \quad \text{and} \quad u_{(\lambda_f(2))}(t, z) = z - 1 - |z - 1|.$$

The mappings $x, t \mapsto u_{\lambda_f(1)}(t, x(t))$*, and* $t \mapsto u_{\lambda_f(2)}(t, \mathbf{x}(t))$ *are plotted in Figure 6-1.*

Since $x(1) = 1$, $u_{(\lambda_f(1))}(1, x(1)) = u_{(\lambda_f(2))}(1, x(1)) = 0$. *For any* $t < 1$,

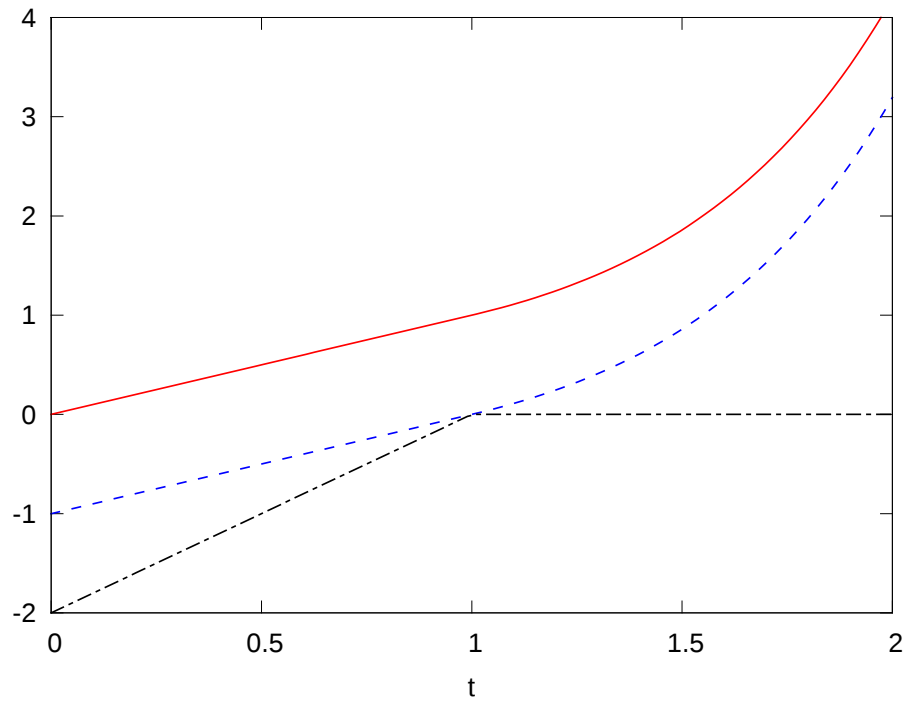

Figure 6-1: Plots of mappings described in Example 6.4.4: $x(t)$ vs. t (solid red), $u_{\lambda_f(1)}(t, x(t))$ vs. t (dashed blue), and $u_{\lambda_f(2)}(t, x(t))$ vs. t (dash-dotted black).

$$u_{\lambda_{\mathbf{f}}(1)}(t, x(t)) = x(t) - 1 = t - 1 < 0,$$
$$u_{\lambda_{\mathbf{f}}(2)}(t, x(t)) = x(t) - 1 + (x(t) - 1) = 2t - 2 < 0.$$

For any $t > 1$,

$$u_{\lambda_{\mathbf{f}}(1)}(t, x(t)) = x(t) - 1 = \tfrac{1}{2}(e^{2(t-1)} - 1) > 0,$$
$$u_{\lambda_{\mathbf{f}}(2)}(t, x(t)) = x(t) - 1 - (x(t) - 1) = 0.$$

Thus, $(\sigma_1^L(1), \sigma_1^R(1)) = (-1, +1)$, *and* $(\sigma_2^L(1), \sigma_2^R(1)) = (-1, 0)$. *Restricting* x *to the subdomain* $[0, 2] \subset \mathbb{R}$, *and defining* Z_1 *and* Z_2 *as in Theorem 6.4.2, note that* $Z_1 = \{1\}$ *and* $Z_2 = [1, 2]$. *There is a single valley-tracing mode of* x, *which occurs on* $[1, 2]$. *The conditions implied by Theorem 6.4.2 are evidently satisfied by this valley-tracing mode, with* $a := 1 \neq t_0$, $b := 2 = t_f$, $i := 2$, *and* $i_A := 1$. *Observe that* $i_A < i$.

Example 6.4.5. *Consider the following ODE, with a single differential variable:*

$$\frac{dx}{dt}(t) = \big|2x(t) - 2 + |x(t) - t|\big| + 3x(t) - t - 1, \qquad x(0) = 0.$$

It is readily confirmed that this ODE is solved by the mapping:

$$x : t \mapsto \begin{cases} t, & \text{if } t \leq 1, \\ \frac{1}{9}(3t + 5 + e^{6(t-1)}), & \text{if } t > 1. \end{cases}$$

Moreover, since the ODE right-hand side function is locally Lipschitz continuous, this solution is unique. As in the previous example, two absolute-value functions appear in any straightforward factored representation of the ODE right-hand side function f, with $u_{(\lambda_f(1))}$ and $u_{(\lambda_f(2))}$ equivalent to the arguments of the inner and outer absolute-value functions in the ODE right-hand side, respectively. Thus, for all $(t, z) \in \mathbb{R}^2$,

$$u_{(\lambda_f(1))}(t, z) = z - t \quad \text{and} \quad u_{(\lambda_f(2))}(t, z) = 2z - 2 - |z - t|.$$

The mappings x, $t \mapsto u_{\lambda_f(1)}(t, x(t))$, and $t \mapsto u_{\lambda_f(2)}(t, \mathbf{x}(t))$ are plotted in Figure 6-2.

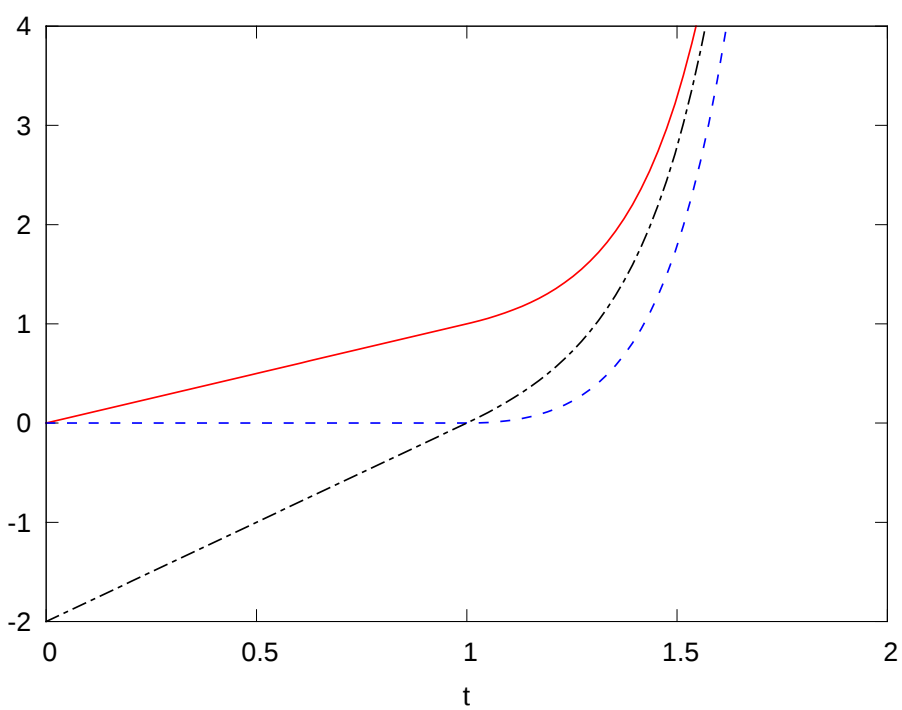

Figure 6-2: Plots of mappings described in Example 6.4.5: $x(t)$ vs. t (solid red), $u_{\lambda_f(1)}(t, x(t))$ vs. t (dashed blue), and $u_{\lambda_f(2)}(t, x(t))$ vs. t (dash-dotted black).

Since $x(1) = 1$, $u_{(\lambda_f(1))}(1, x(1)) = u_{(\lambda_f(2))}(1, x(1)) = 0$. For any $t < 1$,

$$u_{\lambda_f(1)}(t, x(t)) = x(t) - t = t - t = 0,$$
$$u_{\lambda_f(2)}(t, x(t)) = 2x(t) - 2 + |0| = 2t - 2 < 0.$$

For any $t > 1$,

$$\begin{aligned} u_{\lambda_f(1)}(t, x(t)) &= x(t) - t, \\ &= \tfrac{1}{9}(-6t + 5 + e^{6(t-1)}), \\ &> \tfrac{1}{9}(-6t + 5 + (1 + 6(t-1))) = 0, \\ u_{\lambda_f(2)}(t, x(t)) &= 2x(t) - 2 + (x(t) - t), \\ &= 3x(t) - t - 2 = \tfrac{1}{3}(-1 + e^{6(t-1)}) > 0. \end{aligned}$$

Thus, $(\sigma_1^L(1), \sigma_1^R(1)) = (0, +1)$, and $(\sigma_2^L(1), \sigma_2^R(1)) = (-1, +1)$. Restricting x to the subdomain $[0, 2] \subset \mathbb{R}$, and defining Z_1 and Z_2 as in Theorem 6.4.2, note that $Z_1 = [0, 1]$ and $Z_2 = \{1\}$. There is a single valley-tracing mode of x, which occurs on $[0, 1]$. The conditions implied by Theorem 6.4.2 are evidently satisfied by this valley-tracing mode, with $a := 0 = t_0$, $b := 1 \neq t_f$, $i := 1$, and $i_B := 2$. Observe that $i_B > i$.

The following corollaries and examples describe situations in which the indices $i_A, i_B \in \{1, \ldots, p_{\mathbf{f}}\}$ mentioned in Conditions 4 and 5 of Theorem 6.4.2 cannot be furnished. Hence, any valley-tracing mode of $\mathbf{x}$ must begin at t_0 and end at t_f.

Corollary 6.4.6. *Suppose that Assumption 6.3.2 holds. If $p_{\mathbf{f}} = 1$, then the set Z_1 is either finite or equal to $[t_0, t_f]$. If, in addition, $\sigma_1(t_0) \neq 0$ or $[u_{(\lambda_{\mathbf{f}}(1))}]'((t_0, \mathbf{c}); (1, \mathbf{f}(t_0, \mathbf{c}))) \neq 0$, then Z_1 is finite, so there are no valley-tracing modes of $\mathbf{x}$ on $[t_0, t_f]$.*

Example 6.4.7. *Given some fixed $\mathbf{c} \in \mathbb{R}^n$ and analytic functions $\mathbf{g} : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ and $h : \mathbb{R}^n \to \mathbb{R}$, suppose the ODE:*

$$\frac{d\mathbf{x}}{dt}(t) = \mathbf{g}(t, \mathbf{x}(t), |h(\mathbf{x}(t))|), \qquad \mathbf{x}(0) = \mathbf{c},$$

has a solution $\mathbf{x}$ on $[0, t_f]$. (Examples 1.2.1 and 5.2.7 have this structure.) If $h(\mathbf{c}) \neq 0$, or if

$$(\nabla h(\mathbf{c}))^{\mathrm{T}}\,\mathbf{g}(0,\mathbf{c},|h(\mathbf{c})|) \neq 0,$$

then there are no valley-tracing modes of $\mathbf{x}$ *on* $[0,t_f]$.

Example 6.4.8. *Given some fixed* $\mathbf{c} \in \mathbb{R}^n$ *and analytic functions* $\mathbf{g},\boldsymbol{\gamma} : \mathbb{R} \to \mathbb{R}^n$ *and* $h,\eta : \mathbb{R}^n \to \mathbb{R}$*, suppose the ODE:*

$$\begin{aligned}\frac{d\mathbf{x}}{dt}(t) &= \mathbf{g}(\max\{h(\mathbf{x}(t)),\eta(\mathbf{x}(t))\}) + \boldsymbol{\gamma}(\min\{h(\mathbf{x}(t)),\eta(\mathbf{x}(t))\}),\\ \mathbf{x}(0) &= \mathbf{c},\end{aligned}$$

has a solution $\mathbf{x}$ *on* $[0,t_f]$. *Defining* $\mathbf{f}$ *as the abs-factorable right-hand side function of the above ODE, it appears that* $p_{\mathbf{f}} = 2$. *However, noting that* $\max\{y,z\} = \frac{1}{2}(y+z) + \frac{1}{2}|y-z|$ *and* $\min\{y,z\} = \frac{1}{2}(y+z) - \frac{1}{2}|y-z|$*, it follows that* $\mathbf{f}$ *can be written as a composition of analytic functions and a single nonsmooth function,* $\mathbf{z} \mapsto |h(\mathbf{z}) - \eta(\mathbf{z})|$. *With* $\mathbf{f}$ *rewritten in this manner,* $p_{\mathbf{f}}$ *becomes unity, which permits application of Corollary 6.4.6.*

Thus, if $h(\mathbf{c}) \neq \eta(\mathbf{c})$*, or if*

$$(\nabla h(\mathbf{c}) - \nabla\eta(\mathbf{c}))^{\mathrm{T}}\,(\mathbf{g}(\max\{h(\mathbf{c}),\eta(\mathbf{c})\}) + \boldsymbol{\gamma}(\min\{h(\mathbf{c}),\eta(\mathbf{c})\})) \neq 0,$$

then there are no valley-tracing modes of $\mathbf{x}$ *on* $[0,t_f]$.

Corollary 6.4.9. *Suppose that Assumption 6.3.2 holds. If there is no* $t \in [t_0,t_f]$ *such that* $|\{i \in \{1,\ldots,p_{\mathbf{f}}\} : \sigma_i(t) = 0\}| \geq 2$*, then for each* $i \in \{1,\ldots,p_{\mathbf{f}}\}$*, the set* Z_i *is either finite or equal to* $[t_0,t_f]$.

Corollary 6.4.10. *Suppose that Assumption 6.3.2 holds. If there is no* $t \in [t_0,t_f]$ *such that* $|\{i \in \{1,\ldots,p_{\mathbf{f}}\} : \sigma_i(t) = 0\}| \geq 2$*, and if* $\sigma_{i^*}(t_0) \neq 0$ *or*

$$[u_{(\lambda_{\mathbf{f}}(i^*))}]'((t_0,\mathbf{c});(1,\mathbf{f}(t_0,\mathbf{c}))) \neq 0$$

for some $i^* \in \{1,\ldots,p_{\mathbf{f}}\}$*, then the set* Z_{i^*} *is finite.*

Example 6.4.11. *Given some fixed $c \in \mathbb{R}$ and analytic functions $g, h : \mathbb{R} \to \mathbb{R}$, suppose the ODE:*

$$\frac{dx}{dt}(t) = g(t)\,|x(t) - 1| + h(t)\,|x(t)|, \qquad x(0) = c$$

has a solution x on $[0, t_f]$. If $c \notin \{0, 1\}$, or if

$$g(0)\,|c - 1| \neq -h(0)\,|c|,$$

then there are no valley-tracing modes of x on $[0, t_f]$; the sets $\{t \in [0, t_f] : x(t) = 1\}$ and $\{t \in [0, t_f] : x(t) = 0\}$ are finite. If $c \in \{0, 1\}$, then $\{t \in [0, t_f] : x(t) = c\}$ is either finite or equal to $[t_0, t_f]$.

6.5 Conclusions

Non-Zenoness results and necessary conditions for valley-tracing mode emergence have been developed for the solutions of ODEs with abs-factorable right-hand side functions. The conditions for valley-tracing mode emergence are testable during numerical integration; if any of these fail, then there cannot be any valley-tracing modes.

Chapter 7

Evaluating lexicographic derivatives for ODE solutions

7.1 Introduction

This chapter is reproduced from the article [58]; it is concerned with the unique solution $\mathbf{x}(\cdot, \mathbf{p})$ on $[t_0, t_f]$ of the following parametric ordinary differential equation (ODE) system:

$$\frac{d\mathbf{x}}{dt}(t, \mathbf{p}) = \mathbf{f}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})), \qquad \mathbf{x}(t_0) = \mathbf{x}_0(\mathbf{p}),$$

which is presumed to be embedded either in an optimization problem:

$$\min_{\mathbf{p}} s(\mathbf{p}, \mathbf{x}(t_f, \mathbf{p})),$$

or in an equation-solving problem, in which a parameter $\mathbf{p} \in \mathbb{R}^m$ is sought for which

$$\mathbf{0} = \mathbf{r}(\mathbf{p}, \mathbf{x}(t_f, \mathbf{p})).$$

The former optimization problem represents a general class of dynamic optimization problems, while the latter equation-solving problem includes various formulations of boundary-value problems.

Numerical methods for solution of the above problems typically require derivative information for the objective function $\mathbf{p} \mapsto s(\mathbf{p}, \mathbf{x}(t_f, \mathbf{p}))$ or the residual function $\mathbf{p} \mapsto \mathbf{r}(\mathbf{p}, \mathbf{x}(t_f, \mathbf{p}))$. This derivative information is readily furnished when $\mathbf{f}$ and s or $\mathbf{r}$ are continuously differentiable; a classical sensitivity analysis result from ODE theory (summarized in [35]) states that $\mathbf{x}$ is also continuously differentiable in this case, and that the partial derivative mapping $t \mapsto \frac{\partial \mathbf{x}}{\partial \mathbf{p}}(t, \mathbf{p})$ solves a certain linear ODE system uniquely. This partial derivative can be combined with the classical chain rule to compute the desired sensitivity information. Moreover, adjoint sensitivity analysis [13] describes derivatives of the objective function of the above optimization problem without evaluating the partial derivative $\frac{\partial \mathbf{x}}{\partial \mathbf{p}}(t_f, \mathbf{p})$ at all.

These approaches are not applicable, however, when the function $\mathbf{f}$ is locally Lipschitz continuous but not differentiable everywhere. In this case, as shown in Example 1.2.1, $\mathbf{x}$ may also fail to be differentiable. In this chapter, $\mathbf{f}$ is assumed to be a finite composition of analytic functions and absolute value functions; without loss of generality, the same structure is imposed on $\mathbf{x}_0$ and s or $\mathbf{r}$. Though the classical Fréchet derivatives of $\mathbf{f}$ and $\mathbf{x}$ may not exist in this case, various notions of *generalized derivatives* [16, 78, 79, 109] are still well-defined, and can be used in dedicated numerical methods [22, 63, 92] for nonsmooth problems. These generalized derivatives are, however, more difficult to evaluate than their smooth counterparts, since they satisfy weakened versions of classical calculus rules [16, 23]. Nevertheless, Nesterov's lexicographic derivatives [79] satisfy an intuitive chain rule [61, 79], and lexicographic derivatives of the ODE solution $\mathbf{x}$ with respect to the parameter $\mathbf{p}$ were described in Chapter 5 in terms of the unique solution of a certain auxiliary ODE. Moreover, it was argued in Chapter 2 that if lexicographic derivatives are used in place of elements of Clarke's generalized Jacobian in nonsmooth numerical methods, then the convergence results of these methods are not weakened.

However, as shown in Chapter 5, the auxiliary ODE describing lexicographic derivatives of $\mathbf{x}$ is not a Carathéodory ODE, and thus cannot be solved directly using established methods for ODE integration. The goal of this chapter, therefore, is

to develop and present a tractable numerical method for computing lexicographic derivatives of $\mathbf{x}$ with respect to $\mathbf{p}$, so as to compute lexicographic derivatives of the residual function $\mathbf{p} \mapsto \mathbf{r}(\mathbf{p}, \mathbf{x}(t_f, \mathbf{p}))$ or the objective function $\mathbf{p} \mapsto s(\mathbf{p}, \mathbf{x}(t_f, \mathbf{p}))$ above. The developed numerical method exploits inherent structural properties of the auxiliary ODE, using the vector forward mode of automatic differentiation presented in Chapter 4, and the non-Zenoness theory of Chapter 6. This method represents the first tractable method for evaluating a plenary Jacobian element [55, 109] for the unique solution of a parametric ODE with a general nondifferentiable right-hand side function.

For notational simplicity, by considering the parameters $\mathbf{p}$ to be extra state variables that do not vary with t, the parameter vector $\mathbf{p}$ may be appended to the state variable vector $\mathbf{x}$ without loss of generality. The initial-condition generating function $\mathbf{x}_0$ may be considered *a posteriori* using the lexicographic derivative's chain rule, as considered in Chapter 3. These considerations yield the simpler ODE formulation:

$$\frac{d\mathbf{x}}{dt}(t, \mathbf{c}) = \mathbf{f}(t, \mathbf{x}(t, \mathbf{c})), \qquad \mathbf{x}(t_0) = \mathbf{c},$$

in which $\mathbf{x}$ depends on the ODE independent variable t and the initial condition $\mathbf{c}$. This description will be formalized in Section 7.3 below.

As a caveat, we note that the developed theory and method are unnecessary if there are no *valley-tracing modes* [59] in the ODE solution trajectory $\mathbf{x}$. In this case, the ODE solution $\mathbf{x}$ is differentiable with respect to the parameter $\mathbf{c}$; the parametric derivatives $\frac{\partial \mathbf{x}}{\partial \mathbf{c}}$ can then be evaluated using slight modifications of classical sensitivity analysis [16, 125]. Our method, nevertheless, computes the correct parametric derivatives even in this case.

This chapter is structured as follows. Section 7.2 summarizes relevant mathematical concepts which complement the material in Chapter 2, Section 7.3 formalizes the central problem formulation that was outlined above, Section 7.4 develops useful theoretical properties of the auxiliary ODE determining the sensitivities of $\mathbf{x}$, and Section 7.5 harnesses this theory to obtain a numerical method for evaluat-

ing lexicographic derivatives of an ODE solution with respect to the ODE initial conditions. An implementation of the method is described in Section 7.6, and is applied to various example problems for illustration.

7.2 Preliminaries

This section presents fundamental concepts and results which, together with the material presented in Chapter 2, underlie the results and methods in this chapter.

7.2.1 Left/right-analytic functions

The concept of *left/right-analyticity* was introduced in Chapter 6. As shown in Chapter 6, the unique solution of an ODE with an abs-factorable right-hand side function is left/right-analytic; roughly, this property prevents any absolute-value function in the ODE right-hand side from switching between its linear pieces infinitely often in any finite duration. This lack of pathological switching behavior is referred to as *non-Zenoness* [29, 48], and will be exploited heavily – if indirectly – throughout this chapter.

The following property of scalar-valued L/R-analytic functions extends Corollary 6.3.10, and will be used frequently in this chapter.

Lemma 7.2.1. *Given an open set $T \subset \mathbb{R}$, a finite set $S \subset \mathbb{R}$, and a function $g : T \to S$ that is L/R-analytic at $t^* \in T$, there exist $\gamma^L, \gamma^R \in S$ such that, for some sufficiently small $\delta > 0$,*

$$\begin{aligned} & g(t) = \gamma^L, \qquad \forall t \in [t^* - \delta, t^*), \\ \textit{and} \qquad & g(t) = \gamma^R, \qquad \forall t \in (t^*, t^* + \delta]. \end{aligned}$$

Moreover, for any compact set $U \subset T$, there exists a finite set $Z \subset U$ such that g is constant on some neighborhood of each $t \in U \backslash Z$.

Proof. Given an open set $W \subset T$, any $\mathcal{C}^\omega$ function $\tilde{g} : W \to S$ is continuous, and

must therefore be constant, since each element of S is isolated. The existence of γ^L and γ^R then follows from the definition of L/R-analyticity.

Now, choose any compact set $U \subset T$. The first result of the lemma implies that, for each $t^* \in U$, there exist $\gamma^L(t^*), \gamma^R(t^*) \in S$ and $\delta(t^*) > 0$ for which

$$\begin{aligned} & g(t) = \gamma^L(t^*), \qquad \forall t \in (t^* - \delta(t^*), t^*), \\ \text{and} \qquad & g(t) = \gamma^R(t^*), \qquad \forall t \in (t^*, t^* + \delta(t^*)). \end{aligned}$$

Moreover, since U is compact, there exists a finite set $Z \subset U$ for which

$$U \subset \bigcup_{t^* \in Z} (t^* - \delta(t^*), t^* + \delta(t^*)).$$

The above results show that g is constant on some neighborhood of each $t \in U \backslash Z$.

□

Definition 7.2.2. *Define the* signum function *as follows:*

$$\operatorname{sign} : \mathbb{R} \to \{-1, 0, +1\} : x \mapsto \begin{cases} +1, & \text{if } x > 0, \\ 0, & \text{if } x = 0, \\ -1, & \text{if } x < 0, \end{cases}$$

and define an unsignum function *as follows, for each $p \in \mathbb{N}$:*

$$\operatorname{unsign} : \mathbb{R}^p \to \{0, +1\} : \mathbf{x} \mapsto \prod_{k=1}^{p} (1 - |\operatorname{sign} x_p|) = \begin{cases} 0, & \text{if } \exists k \text{ s.t. } x_k \neq 0, \\ +1, & \text{if } x_k = 0 \quad \forall k. \end{cases}$$

Define $\operatorname{unsign} \varnothing_0 = +1$.

The signum function is evidently L/R-analytic, and the following result holds.

Lemma 7.2.3. *Given an open set $T \subset \mathbb{R}$ and an L/R-analytic function $\mathbf{v} : T \to \mathbb{R}^q$, the composite function $\phi \equiv \operatorname{unsign} \circ \mathbf{v}$ is also L/R-analytic on T.*

Proof. Choose any $t^* \in T$; it suffices to show that ϕ is L/R-analytic at t^*. This proof proceeds by strong induction on $q \in \mathbb{N}$.

First, suppose that $q = 1$, in which case $\mathbf{v}$ is a scalar-valued function v. Application of Lemma 6.2.6 with $\mathbf{g} := v$, $Z := \mathbb{R}$, and $\mathbf{h} := \operatorname{sign}(\cdot)$ shows that

the mapping $t \mapsto \operatorname{sign} v(t)$ is L/R-analytic at t^*. A second application of that lemma, with $\mathbf{g} : t \mapsto \operatorname{sign} v(t)$, $Z := (-2, 2)$, and $\mathbf{h} := \operatorname{abs}$, shows that the mapping $t \mapsto |\operatorname{sign} v(t)|$ is also L/R-analytic at t^*. Thus, Lemma 6.2.5 (with $\mathbf{g} : t \mapsto |\operatorname{sign} v(t)|$, $Z := (-2, 2)$, and $\mathbf{h} : z \mapsto 1 - z$) shows that $t \mapsto \operatorname{unsign} v(t)$ is L/R-analytic at t^*.

Next, suppose that the required result has been demonstrated for all $q \leq q^* \in \mathbb{N}$, and consider now the case in which $q := q^* + 1$. Thus, both of the functions $\phi_A : t \mapsto \operatorname{unsign} \mathbf{v}_{1:q^*}(t)$ and $\phi_B : t \mapsto \operatorname{unsign} v_{q^*+1}(t)$ are L/R-analytic at t^*. Observing that $\phi(t) = \phi_A(t)\, \phi_B(t)$ for each $t \in T$, applying Lemma 6.2.5 (with $\mathbf{g} : t \mapsto (\phi_A(t), \phi_B(t))$, $Z := (-2, 2)^2$, and $\mathbf{h} : \mathbf{z} \in \mathbb{R}^2 \mapsto z_1 z_2$) then shows that ϕ is also L/R-analytic at t^*. $\square$

The unsignum function is useful when describing the directional derivatives of the absolute-value function:

$$\operatorname{abs}'(x; d) = \begin{cases} (\operatorname{sign} x)\, d, & \text{if } x \neq 0, \\ |d|, & \text{if } x = 0, \end{cases} = (\operatorname{sign} x)\, d + (\operatorname{unsign} x)\, |d|, \qquad \forall x, d \in \mathbb{R}. \tag{7.1}$$

The *first-sign* function was introduced in [33], and is defined as follows.

Definition 7.2.4. *The* first-sign function *is defined as follows, for each $q \in \mathbb{N}$:*

$$\operatorname{fsign} : \mathbb{R}^q \to \{-1, 0, 1\} : \mathbf{v} \mapsto \begin{cases} \operatorname{sign} v_{k^*}, \text{ with } k^* := \min\{k : v_{(k)} \neq 0\}, & \text{if } \mathbf{v} \neq \mathbf{0}, \\ 0, & \text{if } \mathbf{v} = \mathbf{0}. \end{cases}$$

Define $\operatorname{fsign} \varnothing_0 = 0$.

The following two lemmata collect basic properties of the first-sign function.

Lemma 7.2.5. *The first-sign function satisfies the following identities, for each $q \in \mathbb{N}$, $\mathbf{z} \in \mathbb{R}^q$, and $k \in \{1, \ldots, q\}$:*

1. $\operatorname{fsign} \mathbf{z} = \operatorname{sign} z_1 + \sum_{k=1}^{q-1} (\operatorname{unsign} \mathbf{z}_{1:k})(\operatorname{sign} z_{k+1})$,

2. $\operatorname{fsign} \mathbf{z} = \operatorname{fsign} \mathbf{z}_{1:(q-1)} + (\operatorname{unsign} \mathbf{z}_{1:(q-1)})(\operatorname{sign} z_q)$,

3. $(\text{fsign}\,\mathbf{z})\,z_k = (\text{fsign}\,\mathbf{z}_{1:(k-1)})\,z_k + (\text{unsign}\,\mathbf{z}_{1:(k-1)})\,|z_k|$.

Proof. The first property follows immediately from the definitions of the signum, unsignum, and first-sign functions. The second property is trivial if $q = 1$, and follows immediately from the first property if $q > 1$.

The third property is trivial if $k = 1$ or $z_k = 0$, so assume that $k > 1$ and $z_k \neq 0$. Thus, $\text{fsign}\,\mathbf{z} = \text{fsign}\,\mathbf{z}_{1:k}$; the second property of the lemma then yields:

$$(\text{fsign}\,\mathbf{z})z_k = (\text{fsign}\,\mathbf{z}_{1:k})z_k = (\text{fsign}\,\mathbf{z}_{1:(k-1)})z_k + (\text{unsign}\,\mathbf{z}_{1:(k-1)})(\text{sign}\,z_k)z_k,$$

which is equivalent to the third claimed property. □

Lemma 7.2.6. *Given an open set $T \subset \mathbb{R}$, suppose that a function $\mathbf{v} : T \to \mathbb{R}^q$ is L/R-analytic on T. The composite function $\phi \equiv \text{fsign} \circ \mathbf{v}$ is also L/R-analytic on T.*

Proof. Choose any $t^* \in T$; it suffices to show that ϕ is L/R-analytic at t^*. This proof is analogous to the proof of Lemma 7.2.3, and proceeds by induction on $q \in \mathbb{N}$.

First, suppose that $q = 1$, in which case $\mathbf{v}$ is a scalar-valued function v. In this case, $\text{fsign}\,v(t) = \text{sign}\,v(t)$ for each $t \in T$, and Lemma 6.2.6 shows that the mapping $t \mapsto \text{sign}\,v(t)$ is L/R-analytic at t^*.

Next, suppose that the required result has been demonstrated for $q := q^* \in \mathbb{N}$, and consider now the case in which $q := q^* + 1$. Property 2 of Lemma 7.2.5 yields:

$$\phi(t) = \text{fsign}\,\mathbf{v}_{1:q}(t) + (\text{unsign}\,\mathbf{v}_{1:q}(t))(\text{sign}\,v_{q+1}(t)), \qquad \forall t \in T.$$

Again, Lemma 6.2.6 shows that the mapping $t \mapsto \text{sign}\,v_{q+1}(t)$ is L/R-analytic at t^*. This result, the inductive assumption, Lemma 7.2.3, and Lemma 6.2.5 thus show that ϕ is L/R-analytic at t^*, completing the inductive step. □

Definition 7.2.7. *Define a* first-nonzero locating *function as follows, for each $q \in \mathbb{N}$:*

$$\text{fnzero} : \mathbb{R}^q \to \{0, 1, \ldots, q\} :$$
$$\mathbf{z} \mapsto \sum_{k=1}^{q} \text{unsign}\,\mathbf{z}_{1:k} = \begin{cases} (\min\{k : z_k \neq 0\}) - 1, & \textit{if } \mathbf{z} \neq \mathbf{0}, \\ q, & \textit{if } \mathbf{z} = \mathbf{0}. \end{cases}$$

Lemma 7.2.8. *Given an open set $T \subset \mathbb{R}$, suppose that a function $\mathbf{v} : T \to \mathbb{R}^q$ is L/R-analytic. The composite function $\phi \equiv \text{fnzero} \circ \mathbf{v}$ is also L/R-analytic on T.*

Proof. The required result follows from Lemma 7.2.3 and Lemma 6.2.5. □

7.2.2 LD-derivatives for the absolute-value function

If a function $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\mathbf{x} \in X$, then $\mathbf{f}'(\mathbf{x}; \mathbf{M}) = \mathbf{J}\mathbf{f}(\mathbf{x})\,\mathbf{M}$ for each $\mathbf{M} \in \mathbb{R}^{n \times p}$. Adapting Example 4.2.2, LD-derivatives for the absolute-value function are given by:

$$\text{abs}'(x; \mathbf{M}) = (\text{fsign}\,(x, m_{(1)}, \ldots, m_{(p)}))\,\mathbf{M}, \qquad \forall \mathbf{M} \in \mathbb{R}^{1 \times p}, \quad \forall p \in \mathbb{N}.$$

By Lemma 7.2.5, the kth column of this row vector is:

$$\begin{aligned}
&\text{abs}_{x,\mathbf{M}}^{(k-1)}(m_{(k)}) \\
&\quad = (\text{fsign}\,(x, m_{(1)}, \ldots, m_{(p)}))\, m_{(k)}, \\
&\quad = (\text{fsign}\,(x, m_{(1)}, \ldots, m_{(k-1)}))\, m_{(k)} + (\text{unsign}\,(x, m_{(1)}, \ldots, m_{(k-1)}))\, |m_{(k)}|.
\end{aligned}$$

7.3 Problem formulation

The main results of this chapter concern the dynamic system formalized by the following assumption.

Assumption 7.3.1. *Consider open sets $X \subset \mathbb{R}^n$ and $\bar{T} \subset \mathbb{R}$, elements $t_0, t_f \in \bar{T}$ with $t_0 < t_f$, some $\mathbf{c}_0 \in X$, and an abs-factorable function $\mathbf{f} : \bar{T} \times X \to \mathbb{R}^n$ with at least one absolute-value function in its factored representation. Given the parametric ODE:*

$$\frac{d\mathbf{x}}{dt}(t, \mathbf{c}) = \mathbf{f}(t, \mathbf{x}(t, \mathbf{c})), \qquad \mathbf{x}(t_0, \mathbf{c}) = \mathbf{c}, \tag{7.2}$$

assume that, with $\mathbf{c} := \mathbf{c}_0$, there exists a unique solution $\mathbf{x}(\cdot, \mathbf{c}_0)$ of the above ODE on $[t_0, t_f]$.

In fact, since $\mathbf{f}$ is locally Lipschitz continuous, the particular assumption that the ODE solution $\mathbf{x}(\cdot, \mathbf{c}_0)$ is unique is implied by the remaining parts of Assumption 7.3.1. The inequality $t_0 < t_f$ is assumed without loss of generality; analogous results will hold if $t_f < t_0$ instead.

A classical result of ODE theory [35] shows that the unique ODE solution $\mathbf{x}(\cdot, \mathbf{c}_0)$ on $[t_0, t_f]$ described in Assumption 7.3.1 can be extended to yield a unique solution $\mathbf{x}(\cdot, \mathbf{c}_0)$ of (7.2) on some open set $T \subset \bar{T}$ for which $[t_0, t_f] \subset T$. Moreover, there exists a neighborhood $N \subset X$ of $\mathbf{c}_0$ such that, for any $\mathbf{c} \in N$, the ODE (7.2) has a unique solution $\mathbf{x}(\cdot, \mathbf{c})$ on $[t_0, t_f]$. The existence of T and N will prove to be useful in the following section.

Theorem 5.2.4 implies that $\mathbf{x}_t \equiv \mathbf{x}(t, \cdot)$ is L-smooth at $\mathbf{c}_0$ for each $t \in [t_0, t_f]$; moreover, for any $p \in \mathbb{N}$ and $\mathbf{M} \in \mathbb{R}^{n \times p}$, with $\mathbf{f}_t \equiv \mathbf{f}(t, \cdot)$, the LD-derivative mapping $t \mapsto [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})$ is the unique solution $\mathbf{A}$ on $[t_0, t_f]$ of the *sensitivity ODE*:

$$\frac{d\mathbf{A}}{dt}(t) = [\mathbf{f}_t]'(\mathbf{x}(t, \mathbf{c}_0); \mathbf{A}(t)) = \mathbf{f}'((t, \mathbf{x}(t, \mathbf{c}_0)); (\mathbf{0}_{1\times p}, \mathbf{A}(t))), \qquad \mathbf{A}(t_0) = \mathbf{M}. \quad (7.3)$$

As shown in Example 5.2.6, however, this ODE does not necessarily satisfy the Carathéodory assumptions (summarized in [26]), since the right-hand side of (7.3) may be discontinuous with respect to the $\mathbf{A}(t)$ term. As a result, this ODE is not amenable to established numerical integration methods. Thus, the goal of this chapter is to present a numerical method for computing the LD-derivative $[\mathbf{x}_{t_f}]'(\mathbf{c}_0; \mathbf{M})$ despite its non-Carathéodory nature.

For any $k \in \{1, \ldots, p\}$, Corollary 5.2.3 permits the kth column of the above ODE to be expressed in terms of the leftmost $(k-1)$ columns of $[\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})$ as follows. According to this corollary, the mapping $t \mapsto [\mathbf{x}_t]^{(k-1)}_{\mathbf{c}_0, \mathbf{M}}(\mathbf{m}_{(k)}) = [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})\, \mathbf{e}_{(k)}$ is the unique solution of the ODE:

$$
\begin{aligned}
\frac{d\mathbf{y}}{dt}(t) &= [\mathbf{f}_t]^{(k-1)}_{\mathbf{x}(t,\mathbf{c}_0),[\mathbf{x}_t]'(\mathbf{c}_0;\mathbf{M}_{(1:k-1)})}(\mathbf{y}(t)), \\
&= \mathbf{f}'\left(\begin{bmatrix} t \\ \mathbf{x}(t,\mathbf{c}_0) \end{bmatrix}; \begin{bmatrix} \mathbf{0}_{1\times(k-1)} & 0 \\ [\mathbf{x}_t]'(\mathbf{c}_0;\mathbf{M}_{(1:k-1)}) & \mathbf{y}(t) \end{bmatrix}\right) \mathbf{e}_{(k)}, \\
\mathbf{y}(t_0) &= \mathbf{m}_{(k)}.
\end{aligned}
\tag{7.4}
$$

7.4 Theoretical properties of the sensitivity system

This section develops theoretical properties of the sensitivity ODE (7.3), which will be exploited in later sections to develop a numerical method for solving the sensitivity ODE. The quantities introduced in this section depend on the initial condition $\mathbf{c}_0 \in X$, and are likely to vary nontrivially if $\mathbf{c}_0$ is changed. For notational simplicity, however, explicit references to $\mathbf{c}_0$ are generally omitted.

Algorithm 4 is a specialization of Algorithm 2 from Chapter 4, and computes the right-hand side function of the sensitivity ODE (7.3). This algorithm assumes for notational simplicity that the quantities described in the factored representation of $\mathbf{f}$ are available.

Algorithm 4 Computes $[\mathbf{f}_t]'(\mathbf{z};\mathbf{A}) = \mathbf{f}'((t,\mathbf{z});(\mathbf{0}_{1\times p},\mathbf{A}))$, with $\mathbf{f}$ described by Assumption 7.3.1

Require: $(t,\mathbf{z}) \in T \times X$, $\mathbf{A} \in \mathbb{R}^{n\times p}$
  $\dot{\mathbf{V}}_{(0)} \leftarrow (\mathbf{0}_{1\times p}, \mathbf{A})$
  for $j = 1$ **to** ℓ **do**
    $\dot{\mathbf{U}}_{(j)} \leftarrow [\dot{\mathbf{V}}_{(i)}]_{i\prec j}$
    if $j \in \Lambda_{\mathbf{f}}$ **then**
      $\dot{\mathbf{V}}_{(j)} \leftarrow (\mathrm{fsign}\,(u_{(j)}(t,\mathbf{z}), \dot{u}_{(j),1}, \ldots, \dot{u}_{(j),p}))\,\dot{\mathbf{U}}_{(j)}$
    else
      $\dot{\mathbf{V}}_{(j)} \leftarrow \mathbf{J}\psi_{(j)}(\mathbf{u}_{(j)}(t,\mathbf{z}))\,\dot{\mathbf{U}}_{(j)}$
    end if
  end for
  return $[\mathbf{f}_t]'(\mathbf{z};\mathbf{A}) = \dot{\mathbf{V}}_{(\ell)}$

7.4.1 Left/right-analyticity

This section shows that the solution $t \mapsto [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})$ of the sensitivity ODE (7.3) is L/R-analytic. In later sections, this L/R-analyticity will prove to be crucial for developing a numerical method to solve (7.3).

The following intermediate result is *not* subject to Assumption 7.3.1, and will be used in each step of an inductive proof in the subsequent proposition.

Lemma 7.4.1. *Given open sets $\bar{T}_{\mathbf{w}} \subset \mathbb{R}$, $Z \subset \mathbb{R}^n$, and $W \subset \mathbb{R}^m$, an interval $[t_0, t_f] \subset \bar{T}_{\mathbf{w}}$, and a compact set $U \subset W$, suppose that $\mathbf{w} : \bar{T}_{\mathbf{w}} \to U$ is L/R-analytic, $\mathbf{g} : Z \times W \to \mathbb{R}^n$ is abs-factorable, $\mathbf{z}_0 \in Z$, and the ODE:*

$$\frac{d\mathbf{z}}{dt}(t) = \mathbf{g}(\mathbf{z}(t), \mathbf{w}(t)), \qquad \mathbf{z}(t_0) = \mathbf{z}_0 \tag{7.5}$$

has a unique solution $\mathbf{z}$ on some open superset $T_{\mathbf{z}} \subset \bar{T}_{\mathbf{w}}$ of $[t_0, t_f]$. For each $\boldsymbol{\omega} \in W$, define $\mathbf{g}_{\boldsymbol{\omega}} : \boldsymbol{\zeta} \mapsto \mathbf{g}(\boldsymbol{\zeta}, \boldsymbol{\omega})$. Then, for each $\mathbf{d} \in \mathbb{R}^n$, there exists an open set $T_{\mathbf{y}} \subset \mathbb{R}$ for which:

- $[t_0, t_f] \subset T_{\mathbf{y}} \subset T_{\mathbf{z}}$,
- *the auxiliary ODE:*

$$\frac{d\mathbf{y}}{dt}(t) = [\mathbf{g}_{\mathbf{w}(t)}]'(\mathbf{z}(t); \mathbf{y}(t)) = \mathbf{g}'((\mathbf{z}(t), \mathbf{w}(t)); (\mathbf{y}(t), \mathbf{0}_m)), \qquad \mathbf{y}(t_0) = \mathbf{d} \tag{7.6}$$

  has a unique solution $\mathbf{y}$ on $T_{\mathbf{y}}$, and

- *this unique solution $\mathbf{y}$ is L/R-analytic on $T_{\mathbf{y}}$.*

Moreover, the ODE (7.6) *may be expressed in the form of* (7.5) *as follows. Defining $\tilde{Z} := \mathbb{R}^n$, there exist $r \in \mathbb{N}$, an open set $\tilde{T} \subset \mathbb{R}$, an open set $\tilde{W} \subset \mathbb{R}^r$, a compact set $\tilde{U} \subset \tilde{W}$, an L/R-analytic function $\tilde{\mathbf{w}} : \tilde{T} \to \tilde{U}$, and an abs-factorable function $\tilde{\mathbf{g}} : \tilde{Z} \times \tilde{W} \to \mathbb{R}^n$ such that $[t_0, t_f] \subset \tilde{T} \subset T_{\mathbf{y}}$, and*

$$\tilde{\mathbf{g}}(\boldsymbol{\eta}, \tilde{\mathbf{w}}(\tau)) = [\mathbf{g}_{\mathbf{w}(\tau)}]'(\mathbf{z}(\tau); \boldsymbol{\eta}), \qquad \forall \tau \in \tilde{T}, \boldsymbol{\eta} \in \tilde{Z}.$$

Proof. Define an interval $I_{\mathbf{z}} := [a_{\mathbf{z}}, b_{\mathbf{z}}] \subset T_{\mathbf{z}}$ for which $a_{\mathbf{z}} < t_0$ and $t_f < b_{\mathbf{z}}$. Theorem 6.3.9 shows that $\mathbf{z}$ is L/R-analytic on $I_{\mathbf{z}}$, and Theorem 5.2.1 shows that the ODE (7.6) does indeed have a unique solution $\mathbf{y}$ on $[t_0, t_f]$, which can be extended [35] to yield a unique solution $\mathbf{y}$ on some open set $\bar{T}_{\mathbf{y}} \subset \mathbb{R}$ for which $[t_0, t_f] \subset \bar{T}_{\mathbf{y}} \subset T_{\mathbf{z}}$.

Now, consider a particular factored representation of $\mathbf{g}$:

$\mathbf{v}_{(0)} \leftarrow (\boldsymbol{\zeta}, \boldsymbol{\omega}) \in Z \times W$
for $j = 1$ **to** ℓ **do**
  $\mathbf{u}_{(j)} \leftarrow [\mathbf{v}_{(i)}]_{i \prec j}$
  $\mathbf{v}_{(j)} \leftarrow \boldsymbol{\psi}_{(j)}(\mathbf{u}_{(j)})$
end for
$\mathbf{g}(\boldsymbol{\zeta}, \boldsymbol{\omega}) \leftarrow \mathbf{v}_{(\ell)}$,

and define $\Lambda_{\mathbf{g}} := \{j \in \{1, \ldots, \ell\} : \boldsymbol{\psi}_{(j)} \equiv \text{abs}\}$. Following the approaches of [32] and Chapter 6, a factored representation can be generated for the right-hand side $\gamma_A : (\boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{\omega}) \in \mathbb{R}^n \times Z \times W \mapsto [\mathbf{g}_{\boldsymbol{\omega}}]'(\boldsymbol{\zeta}; \boldsymbol{\eta})$ of the ODE (7.6), using (7.1) and the forward mode of automatic differentiation. This approach yields the following result:

$\mathbf{v}_{(0)} \leftarrow (\boldsymbol{\zeta}, \boldsymbol{\omega}) \in Z \times W$
$\dot{\mathbf{v}}_{(0)} \leftarrow (\boldsymbol{\eta}, \mathbf{0}_m) \in \mathbb{R}^n \times \mathbb{R}^m$
for $j = 1$ **to** ℓ **do**
  $\mathbf{u}_{(j)} \leftarrow [\mathbf{v}_{(i)}]_{i \prec j}$
  $\dot{\mathbf{u}}_{(j)} \leftarrow [\dot{\mathbf{v}}_{(i)}]_{i \prec j}$
  if $j \in \Lambda_{\mathbf{g}}$ **then**
    $v_{(j)} \leftarrow |u_{(j)}|$
    $\dot{v}_{(j)} \leftarrow (\operatorname{sign} u_{(j)})\, \dot{u}_{(j)} + (\operatorname{unsign} u_{(j)})\, |\dot{u}_{(j)}|$
  else
    $\mathbf{v}_{(j)} \leftarrow \boldsymbol{\psi}_{(j)}(\mathbf{u}_{(j)})$
    $\dot{\mathbf{v}}_{(j)} \leftarrow \mathbf{J}\boldsymbol{\psi}_{(j)}(\mathbf{u}_{(j)})\, \dot{\mathbf{u}}_{(j)}$
  end if
end for
$\gamma_A(\boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{\omega}) \leftarrow \dot{\mathbf{v}}_{(\ell)}$.

Next, the "$\operatorname{sign} u_{(j)}$" and "$\operatorname{unsign} u_{(j)}$" terms in the above representation will be represented as inputs instead. Observe that the factored representation for $\mathbf{g}$ defines each $\mathbf{u}_{(j)}$ as a function on $Z \times W$. Hence, consider a variation $\gamma_B : \mathbb{R}^n \times Z \times W \times (-2,2)^\ell \times (-2,2)^\ell \to \mathbb{R}^m$ of γ_A, so that $\gamma_B(\boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{\omega}, \boldsymbol{\sigma}, \bar{\boldsymbol{\sigma}})$ is defined according to the following factored representation:

$\mathbf{v}_{(0)} \leftarrow (\boldsymbol{\zeta}, \boldsymbol{\omega}) \in Z \times W$
$\dot{\mathbf{v}}_{(0)} \leftarrow (\boldsymbol{\eta}, \mathbf{0}_m) \in \mathbb{R}^n \times \mathbb{R}^m$
for $j = 1$ **to** ℓ **do**
  $\mathbf{u}_{(j)} \leftarrow [\mathbf{v}_{(i)}]_{i \prec j}$
  $\dot{\mathbf{u}}_{(j)} \leftarrow [\dot{\mathbf{v}}_{(i)}]_{i \prec j}$
  if $j \in \Lambda_{\mathbf{g}}$ **then**
    $v_{(j)} \leftarrow |u_{(j)}|$
    $\dot{v}_{(j)} \leftarrow \sigma_j \dot{u}_{(j)} + \bar{\sigma}_j |\dot{u}_{(j)}|$
  else
    $\mathbf{v}_{(j)} \leftarrow \boldsymbol{\psi}_{(j)}(\mathbf{u}_{(j)})$
    $\dot{\mathbf{v}}_{(j)} \leftarrow \mathbf{J}\boldsymbol{\psi}_{(j)}(\mathbf{u}_{(j)})\, \dot{\mathbf{u}}_{(j)}$
  end if
end for
$\gamma_B(\boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{\omega}, \boldsymbol{\sigma}, \bar{\boldsymbol{\sigma}}) \leftarrow \dot{\mathbf{v}}_{(\ell)}$.

Unlike γ_A, γ_B is abs-factorable.

Now, define mappings $\mathbf{s}, \bar{\mathbf{s}} : I_{\mathbf{z}} \to \{-1, 0, +1\}^{\ell}$, such that for each $j \in \{1, \ldots, \ell\}$ and $t \in I_{\mathbf{z}}$,

$$s_j(t) = \begin{cases} \operatorname{sign} u_{(j)}(\mathbf{z}(t), \mathbf{w}(t)) & \text{if } j \in \Lambda_{\mathbf{g}}, \\ 0 & \text{if } j \notin \Lambda_{\mathbf{g}}, \end{cases}$$
$$\text{and} \qquad \bar{s}_j(t) = \begin{cases} \operatorname{unsign} u_{(j)}(\mathbf{z}(t), \mathbf{w}(t)) & \text{if } j \in \Lambda_{\mathbf{g}}, \\ 0 & \text{if } j \notin \Lambda_{\mathbf{g}}. \end{cases}$$

Since $\mathbf{z}$ is continuous, the set $\{\mathbf{z}(t) : t \in I_{\mathbf{z}}\}$ is compact. Define an interval $\tilde{I}_{\mathbf{z}} := [\tilde{a}_{\mathbf{z}}, \tilde{b}_{\mathbf{z}}] \subset I_{\mathbf{z}}$ so that $a_{\mathbf{z}} < \tilde{a}_{\mathbf{z}} < t_0$ and $t_f < \tilde{b}_{\mathbf{z}} < b_{\mathbf{z}}$. Thus, since $\mathbf{z}$ and $\mathbf{w}$ are L/R-analytic on $I_{\mathbf{z}}$, Lemma 6.2.7 implies that the mapping $t \mapsto u_{(j)}(\mathbf{z}(t), \mathbf{w}(t))$ is L/R-analytic on $\tilde{I}_{\mathbf{z}}$ for each $j \in \Lambda_{\mathbf{g}}$; Lemma 6.2.6 and Lemma 7.2.3 then imply that both $\mathbf{s}$ and $\bar{\mathbf{s}}$ are L/R-analytic on $\tilde{I}_{\mathbf{z}}$. Now, observe that the ODE (7.6) is equivalent to the ODE:

$$\frac{d\mathbf{y}}{dt}(t) = \gamma_B(\mathbf{y}(t), \mathbf{z}(t), \mathbf{w}(t), \mathbf{s}(t), \bar{\mathbf{s}}(t)), \qquad \mathbf{y}(t_0) = \mathbf{d},$$

which has an abs-factorable right-hand side function γ_B. Moreover, the functions $\mathbf{z}$, $\mathbf{w}$, $\mathbf{s}$, and $\bar{\mathbf{s}}$ are each L/R-analytic on $\tilde{I}_{\mathbf{z}}$. Define an interval $\tilde{I}_{\mathbf{y}} := [\tilde{a}_{\mathbf{y}}, \tilde{b}_{\mathbf{y}}] \subset \tilde{I}_{\mathbf{z}}$ such that $\tilde{a}_{\mathbf{z}} < \tilde{a}_{\mathbf{y}} < t_0$ and $t_f < \tilde{b}_{\mathbf{y}} < \tilde{b}_{\mathbf{z}}$. Theorem 6.3.9 then implies that $\mathbf{y}$ is L/R-analytic on $\tilde{I}_{\mathbf{y}}$, and so $T_{\mathbf{y}} := (\tilde{a}_{\mathbf{y}}, \tilde{b}_{\mathbf{y}})$ satisfies the requirements of the lemma.

The remaining claim of the lemma is trivially satisfied by setting $r := n + m + 2\ell$, $\tilde{W} := Z \times W \times (-2,2)^\ell \times (-2,2)^\ell \subset \mathbb{R}^r$,

$$\tilde{U} := \{\mathbf{z}(t) \in Z : t \in \tilde{I}_{\mathbf{z}}\} \times U \times \{-1,0,+1\}^\ell \times \{-1,0,+1\}^\ell,$$

$\tilde{\mathbf{w}} \equiv (\mathbf{z}, \mathbf{w}, \mathbf{s}, \bar{\mathbf{s}})$, and $\tilde{\mathbf{g}} \equiv \gamma_B$. □

Proposition 7.4.2. *Suppose that Assumption 7.3.1 holds, and consider any* $\mathbf{M} \in \mathbb{R}^{n \times p}$. *The mapping* $t \mapsto [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})$ *is L/R-analytic on some open set* $\tilde{T} \subset \mathbb{R}$ *for which* $[t_0, t_f] \subset \tilde{T} \subset T$.

Proof. To obtain the required result, this proof demonstrates by induction that, for each $k \in \{1, \ldots, p\}$, the mapping $t \mapsto [\mathbf{x}_t]^{(k-1)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{m}_{(k)})$ is L/R-analytic on some open set $T_{(k)} \subset \mathbb{R}$ for which $[t_0, t_f] \subset T_{(k)} \subset T$. For the case in which $k = 1$, consider an interval $I_0 := [a_0, b_0] \subset T$ for which $a_0 < t_0$ and $t_f < b_0$. Theorem 5.2.1 implies that the mapping $t \mapsto [\mathbf{x}_t]^{(0)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{m}_{(1)}) = [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{m}_{(1)})$ solves the following ODE uniquely on I_0, with $\mathbf{f}_t \equiv \mathbf{f}(t, \cdot)$:

$$\frac{d\mathbf{y}}{dt}(t) = [\mathbf{f}_t]'(\mathbf{x}(t, \mathbf{c}_0); \mathbf{y}(t)), \qquad \mathbf{y}(t_0) = \mathbf{m}_{(1)}.$$

Applying Lemma 7.4.1 with $\mathbf{z} \equiv \mathbf{x}(\cdot, \mathbf{c}_0)$, $\bar{T}_{\mathbf{w}} := \bar{T} \cap (t_0 - 1, t_f + 1)$, $U := [t_0 - 1, t_f + 1]$, $W := \mathbb{R}$, $\mathbf{w} : t \mapsto t$, and $\mathbf{g} \equiv \mathbf{f}$ thus shows that there exists an open set $T_{(1)} \subset \mathbb{R}$ for which $[t_0, t_f] \subset T_{(1)} \subset (a_0, b_0) \cap \bar{T}_{\mathbf{w}} \cap T$, for which the mapping $t \mapsto [\mathbf{x}_t]^{(0)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{m}_{(1)})$ is L/R-analytic on $T_{(1)}$, and for which there exist $\tilde{r} \in \mathbb{N}$, an open set $\tilde{W} \subset \mathbb{R}^{\tilde{r}}$, a compact set $\tilde{U} \subset \tilde{W}$, an L/R-analytic function $\tilde{\mathbf{w}} : T_{(1)} \to \tilde{U}$, and an abs-factorable function $\tilde{\mathbf{g}} : \mathbb{R}^n \times \tilde{W} \to \mathbb{R}^n$ such that $\tilde{\mathbf{g}}(\boldsymbol{\eta}, \tilde{\mathbf{w}}(\tau)) = [\mathbf{f}_\tau]'(\mathbf{x}(\tau, \mathbf{c}_0); \boldsymbol{\eta})$ for each $(\tau, \boldsymbol{\eta}) \in T_{(1)} \times \mathbb{R}^n$.

As the inductive assumption, suppose that, with $k := q$ for some $q \in \{1, \ldots, p - 1\}$, the mapping $t \mapsto [\mathbf{x}_t]^{(q-1)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{m}_{(q)})$ is L/R-analytic on some open set $T_{(q)}$ for which $[t_0, t_f] \subset T_{(q)} \subset T$, and there exist $\hat{r} \in \mathbb{N}$, an open set $\hat{W} \subset \mathbb{R}^{\hat{r}}$, a compact set $\hat{U} \subset \hat{W}$, an L/R-analytic function $\hat{\mathbf{w}} : T_{(q)} \to \hat{U}$, and an abs-factorable function $\hat{\mathbf{g}} : \mathbb{R}^n \times \hat{W} \to \mathbb{R}^n$ such that $t \mapsto [\mathbf{x}_t]^{(q-1)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{m}_{(q)})$ is the unique solution $\mathbf{y}$ on $T_{(q)}$ of the ODE:

$$\frac{d\mathbf{y}}{dt}(t) = \hat{\mathbf{g}}(\mathbf{y}(t), \tilde{\mathbf{w}}(t)), \qquad \mathbf{y}(t_0) = \mathbf{m}_{(q)}.$$

With $\hat{\mathbf{g}}_{\boldsymbol{\omega}} \equiv \hat{\mathbf{g}}(\cdot, \boldsymbol{\omega})$, Theorem 5.2.1 implies the existence of some interval $I_{q+1} := [a_{q+1}, b_{q+1}] \subset T_{(q)}$ for which $a_{q+1} < t_0$ and $t_f < b_{q+1}$, and for which the mapping $t \mapsto [\mathbf{x}_t]^{(q)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{m}_{(q+1)}) = \left[[\mathbf{x}_t]^{(q-1)}_{\mathbf{c}_0,\mathbf{M}}\right]'(\mathbf{m}_{(q)}; \mathbf{m}_{(q+1)})$ is the unique solution $\mathbf{y}$ on I_{q+1} of the ODE:

$$\frac{d\mathbf{y}}{dt}(t) = [\hat{\mathbf{g}}_{\tilde{\mathbf{w}}(t)}]'([\mathbf{x}_t]^{(q-1)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{m}_{(q)}); \mathbf{y}(t)), \qquad \mathbf{y}(t_0) = \mathbf{m}_{(q+1)}.$$

Lemma 7.4.1 then implies that the inductive assumption holds for $k := q + 1$ as well, completing the inductive proof. The proposition is thereby demonstrated with $\tilde{T} := T_{(p)}$. □

Corollary 7.4.3. *Suppose that Assumption 7.3.1 holds, consider any* $\mathbf{M} \in \mathbb{R}^{n\times p}$*, and consider the set* $\tilde{T} \subset \mathbb{R}$ *described in Proposition 7.4.2. There exists an open set* $\hat{T} \subset \mathbb{R}$ *for which* $[t_0, t_f] \subset \hat{T} \subset \tilde{T}$*, and for which, for each* $j \in \{1, \ldots, \ell\}$*, the mappings* $t \mapsto \dot{\mathbf{U}}_{(j)}((t, \mathbf{x}(t, \mathbf{c}_0)); (\mathbf{0}_{1\times p}, [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})))$ *and* $t \mapsto \dot{\mathbf{V}}_{(j)}((t, \mathbf{x}(t, \mathbf{c}_0)); (\mathbf{0}_{1\times p}, [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})))$ *are L/R-analytic on* $\hat{T}$*.*

Proof. By Lemma 6.2.7 and Theorem 6.3.9, there exists an interval $I_{\mathbf{u}} := [a_{\mathbf{u}}, b_{\mathbf{u}}] \subset \tilde{T}$ for which $a_{\mathbf{u}} < t_0$ and $t_f < b_{\mathbf{u}}$, for which the mappings $t \mapsto \mathbf{x}(t, \mathbf{c}_0)$ and $t \mapsto \mathbf{u}_{(j)}(t, \mathbf{x}(t, \mathbf{c}_0))$ are L/R-analytic on $I_{\mathbf{u}}$ for each $j \in \{1, \ldots, \ell\}$. Noting that $\mathbf{x}(\cdot, \mathbf{c}_0)$ is continuous by construction, the set $\{\mathbf{x}(t, \mathbf{c}_0) : t \in I_{\mathbf{u}}\}$ is compact. Similarly, the set $\{\mathbf{u}_{(j)}(t, \mathbf{x}(t, \mathbf{c}_0)) : t \in I_{\mathbf{u}}\}$ is compact for each $j \in \{1, \ldots, \ell\}$.

Now, this proof proceeds by strong induction on $j \in \{0, \ldots, \ell\}$. For the $j := 0$ case, observe that

$$\dot{\mathbf{V}}_{(0)}((t, \mathbf{x}(t, \mathbf{c}_0)); (\mathbf{0}_{1\times p}, [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M}))) = (\mathbf{0}_{1\times p}, [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})) \in \mathbb{R}^{(n+1)\times p}, \qquad \forall t \in T.$$

Proposition 7.4.2 thus implies that the mapping

$$t \mapsto \dot{\mathbf{V}}_{(0)}((t, \mathbf{x}(t, \mathbf{c}_0)); (\mathbf{0}_{1\times p}, [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})))$$

is L/R-analytic on $\tilde{T}$.

Next, as the strong inductive assumption, choose any $j^* \in \{1, \ldots, \ell\}$, and suppose that, for each $j < j^*$, there exists an open set $T^*_{(j)} \subset \tilde{T}$ for which $[t_0, t_f] \subset T^*_{(j)}$, and for which the mapping $t \in T^*_{(j)} \mapsto \dot{\mathbf{V}}_{(j)}((t, \mathbf{x}(t, \mathbf{c}_0)); (\mathbf{0}_{1\times p}, [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})))$ is L/R-analytic. The cases in which $j^* \notin \Lambda_{\mathbf{f}}$ and $j^* \in \Lambda_{\mathbf{f}}$ will be considered separately.

If $j^* \notin \Lambda_{\mathbf{f}}$, then $\psi_{(j^*)}$ is $\mathcal{C}^\omega$ on its domain. Since the set $\{\mathbf{u}_{(j^*)}(t, \mathbf{x}(t, \mathbf{c}_0)) : t \in I_{\mathbf{u}}\}$ is a compact subset of the domain of $\psi_{(j^*)}$, Lemma 6.2.5 implies that the mapping $t \mapsto \mathbf{J}\psi_{(j^*)}(\mathbf{u}_{(j^*)}(t, \mathbf{x}(t, \mathbf{c}_0)))$ is L/R-analytic on $(a_{\mathbf{u}}, b_{\mathbf{u}})$. Defining an open superset $\tilde{T}^* := \bigcap_{\{j: j \prec j^*\}} T^*_{(j)}$ of $[t_0, t_f]$, the inductive assumption shows that the mapping $t \in \tilde{T}^* \mapsto \dot{\mathbf{U}}_{(j^*)}((t, \mathbf{x}(t, \mathbf{c}_0)); (\mathbf{0}_{1\times p}, [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})))$ is L/R-analytic. Thus, Algorithm 4 and Lemma 6.2.5 imply that, with $T^*_{(j^*)} := (a_{\mathbf{u}}, b_{\mathbf{u}}) \cap \tilde{T}^*$, the mapping $t \in T^*_{(j^*)} \mapsto \dot{\mathbf{V}}_{(j^*)}((t, \mathbf{x}(t, \mathbf{c}_0)); (\mathbf{0}_{1\times p}, [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})))$ is also L/R-analytic.

Otherwise, if $j^* \in \Lambda_{\mathbf{f}}$, then $\psi_{(j^*)} \equiv \text{abs}$, and there exists a single index $j < j^*$ for which $j \prec j^*$; the strong inductive assumption then shows that the mapping

$$\begin{aligned} t \in \tilde{T}^* \mapsto \dot{\mathbf{U}}_{(j^*)}((t, \mathbf{x}(t, \mathbf{c}_0)); (\mathbf{0}_{1\times p}, [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M}))), \\ = \dot{\mathbf{V}}_{(j)}((t, \mathbf{x}(t, \mathbf{c}_0)); (\mathbf{0}_{1\times p}, [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M}))) \end{aligned}$$

is L/R-analytic. Define $T^*_{(j^*)} := (a_{\mathbf{u}}, b_{\mathbf{u}}) \cap T^*_{(j)} \supset [t_0, t_f]$. The definition of $I_{\mathbf{u}}$, Algorithm 4, Lemma 7.2.6, the inductive assumption, and Lemma 6.2.5 thus imply that the mapping

$$t \in T^*_{(j^*)} \mapsto \dot{\mathbf{V}}_{(j^*)}((t, \mathbf{x}(t, \mathbf{c}_0)); (\mathbf{0}_{1\times p}, [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})))$$

is L/R-analytic.

The inductive argument is thus completed; the corollary is thereby demonstrated with $\hat{T} := \bigcap_{j=0}^{\ell} T^*_{(j)}$. □

7.4.2 Classical evolution between discrete valley crossings

This section uses the L/R-analyticity result of the previous section to show that there exist finitely many discrete events in the interval $[t_0, t_f]$, between which the

sensitivity ODE (7.3) evolves according to a classical linear sensitivity system corresponding to a certain auxiliary ODE. These discrete events will be referred to as *valley crossings*, and are generalizations of the valley crossings described previously in Chapter 6. Roughly, the theory developed in this section adds to the various non-Zenoness results obtained in Chapter 6.

Definition 7.4.4. *Suppose that Assumption 7.3.1 holds, and define the set $\hat{T} \subset \mathbb{R}$ as in Corollary 7.4.3. Consider any $t^* \in T \cap \hat{T}$ and $j \in \Lambda_{\mathbf{f}}$. For notational simplicity, in the remainder of this chapter, the superset $T \cap \hat{T}$ of $[t_0, t_f]$ will be denoted as T, the quantity $u_{(j)}(t^*, \mathbf{x}(t^*, \mathbf{c}_0))$ will frequently be denoted as "$u_{(j)}(t^*)$", and the quantity*

$$\dot{\mathbf{U}}_{(j)}((t^*, \mathbf{x}(t^*, \mathbf{c}_0)); (\mathbf{0}_{1\times p}, [\mathbf{x}_{t^*}]'(\mathbf{c}_0; \mathbf{M})))$$

will be denoted as "$\dot{\mathbf{U}}_{(j)}(t^*)$", whose k^{th} column is a scalar that will be denoted as "$\dot{u}_{(j),k}(t^*)$". Define:

$$\begin{aligned}
\sigma_{j,0}(t^*) &:= \operatorname{sign} u_{(j)}(t^*) \in \{-1, 0, +1\}, \\
\sigma_{j,k}(t^*) &:= \operatorname{sign} \dot{u}_{(j),k}(t^*) \in \{-1, 0, +1\}, \qquad \forall k \in \{1, \ldots, p\}, \\
\sigma_{j,p+1}(t^*) &:= 0, \\
\kappa_j(t^*) &:= \operatorname{fnzero}\left(u_{(j)}(t^*), \dot{u}_{(j),1}(t^*), \ldots, \dot{u}_{(j),p}(t^*)\right) \in \{0, 1, \ldots, p+1\}, \\
\text{and} \qquad \zeta_j(t^*) &:= \operatorname{fsign}\left(u_{(j)}(t^*), \dot{u}_{(j),1}(t^*), \ldots, \dot{u}_{(j),p}(t^*)\right) = \sigma_{j,\kappa_j(t^*)}(t^*) \in \{-1, 0, +1\}.
\end{aligned}$$

The quantities $\sigma_{j,k}(t^*)$, $\kappa_j(t^*)$, and $\zeta_j(t^*)$ will be referred to as signatures, valley-tracing depths, *and* critical signatures, *respectively.*

Remark 7.4.5. *Define a function $\phi : \{1, \ldots, p\} \times T \times \mathbb{R}^n$ according to Algorithm 5, which employs the constructions of the previous definition. For each $k \in \{1, \ldots, p\}$, comparison of Algorithms 4 and 5 shows that*

$$\begin{aligned}
\phi(k, t, \mathbf{a}) &= [\mathbf{f}_t]^{(k-1)}_{\mathbf{x}(t,\mathbf{c}_0), [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M}_{(1:k-1)})}(\mathbf{a}), \\
&= \mathbf{f}'\left(\begin{bmatrix} t \\ \mathbf{x}(t, \mathbf{c}_0) \end{bmatrix}; \begin{bmatrix} \mathbf{0}_{1\times(k-1)} & 0 \\ [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M}_{(1:k-1)}) & \mathbf{a} \end{bmatrix}\right) \mathbf{e}_{(k)}, \qquad \forall (t, \mathbf{a}) \in T \times \mathbb{R}^n.
\end{aligned}$$

Thus, $\phi(k,\cdot,\cdot)$ is the right-hand side function of the ODE (7.4); it follows immediately that the mapping $t \mapsto [\mathbf{x}_t]^{(k-1)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{m}_{(k)}) = [\mathbf{x}_t]'(\mathbf{c}_0;\mathbf{M})\,\mathbf{e}_{(k)}$ is the unique solution on $[t_0,t_f]$ of the ODE:

$$\frac{d\mathbf{y}}{dt}(t) = \phi(k,t,\mathbf{y}(t)), \qquad \mathbf{y}(t_0) = \mathbf{m}_{(k)}. \tag{7.7}$$

Algorithm 5 Computes $\phi(k,t,\mathbf{a})$, with $\mathbf{f}$ and $\mathbf{x}$ described by Assumption 7.3.1

Require: $k \in \{1,\ldots,p\}$, $t \in T$, $\mathbf{a} \in \mathbb{R}^n$
  $\bar{\mathbf{v}}_{(0)} \leftarrow (0,\mathbf{a})$
  for $j = 1$ **to** ℓ **do**
    $\bar{\mathbf{u}}_{(j)} \leftarrow [\bar{\mathbf{v}}_{(i)}]_{i\prec j}$
    if $j \in \Lambda_{\mathbf{f}}$ **and** $\kappa_j(t) < k$ **then**
      $\bar{v}_{(j)} \leftarrow \zeta_j(t)\,\bar{u}_{(j)}$
    else if $j \in \Lambda_{\mathbf{f}}$ **and** $\kappa_j(t) \geq k$ **then**
      $\bar{v}_{(j)} \leftarrow |\bar{u}_{(j)}|$
    else if $j \notin \Lambda_{\mathbf{f}}$ **then**
      $\bar{\mathbf{v}}_{(j)} \leftarrow \mathbf{J}\psi_{(j)}(\mathbf{u}_{(j)}(t,\mathbf{x}(t,\mathbf{c}_0)))\,\bar{\mathbf{u}}_{(j)}$
    end if
  end for
  return $\phi(k,t,\mathbf{a}) := \bar{\mathbf{v}}_{(\ell)}$

Corollary 7.4.3 and Lemmata 7.2.1, 7.2.6, and 7.2.8 together demonstrate the following lemma. Similar constructions were employed in Chapter 6.

Lemma 7.4.6. *Suppose that Assumption 7.3.1 holds. For each $t^* \in T$, $j \in \Lambda_{\mathbf{f}}$, and $k \in \{0,1,\ldots,p+1\}$, the mappings $\sigma_{j,k}$, κ_j, and ζ_j are L/R-analytic at t^*. Thus, there exist $\sigma^L_{j,k}(t^*), \sigma^R_{j,k}(t^*) \in \{-1,0,+1\}$ so that, for some sufficiently small $\delta > 0$,*

$$\begin{aligned} \sigma_{j,k}(t) &= \sigma^L_{j,k}(t^*), \qquad \forall t \in [t^*-\delta, t^*), \\ \textit{and} \qquad \sigma_{j,k}(t) &= \sigma^R_{j,k}(t^*), \qquad \forall t \in (t^*, t^*+\delta]. \end{aligned}$$

Similarly, there exist $\kappa^L_j(t^*), \kappa^R_j(t^*) \in \{0,1,\ldots,p+1\}$ so that, for some sufficiently small $\delta > 0$,

$$\begin{aligned} \kappa_j(t) &= \kappa^L_j(t^*), \qquad \forall t \in [t^*-\delta, t^*), \\ \textit{and} \qquad \kappa_j(t) &= \kappa^R_j(t^*), \qquad \forall t \in (t^*, t^*+\delta]. \end{aligned}$$

Lastly, there exist $\zeta_j^L(t^*), \zeta_j^R(t^*) \in \{-1, 0, +1\}$ *so that, for some sufficiently small* $\delta > 0$,

$$\begin{aligned} \zeta_j(t) &= \zeta_j^L(t^*), \qquad \forall t \in [t^* - \delta, t^*), \\ \textit{and} \qquad \zeta_j(t) &= \zeta_j^R(t^*), \qquad \forall t \in (t^*, t^* + \delta]. \end{aligned}$$

Definition 7.4.7. *Suppose that Assumption 7.3.1 holds. For each* $t^* \in T$, $j \in \Lambda_{\mathbf{f}}$, *and* $k \in \{0, 1, \ldots, p+1\}$, *define* $\sigma_{j,k}^L(t^*)$, $\sigma_{j,k}^R(t^*)$, $\kappa_j^L(t^*)$, $\kappa_j^R(t^*)$, $\zeta_j^L(t^*)$, *and* $\zeta_j^R(t^*)$ *as in the previous lemma. Collect these quantities over all* $j \in \Lambda_{\mathbf{f}}$ *as vectors* $\boldsymbol{\sigma}_{(k)}^L(t^*)$, $\boldsymbol{\sigma}_{(k)}^R(t^*)$, $\boldsymbol{\kappa}^L(t^*)$, $\boldsymbol{\kappa}^R(t^*)$, $\boldsymbol{\zeta}^L(t^*)$, *and* $\boldsymbol{\zeta}^R(t^*)$. *To preserve the notational convention that* z_i *denotes the* i^{th} *component of a vector* $\mathbf{z}$, *the aforementioned vectors will be considered to have dimension* ℓ, *but their components corresponding to any* $j \notin \Lambda_{\mathbf{f}}$ *will not be used.*

Lemma 7.4.8. *There exists a finite set* $Z \subset [t_0, t_f]$ *such that, for each* $t \in [t_0, t_f] \backslash Z$, $j \in \Lambda_{\mathbf{f}}$, *and* $k \in \{0, 1, \ldots, p+1\}$, *the following equations hold:* $\sigma_{j,k}(t) = \sigma_{j,k}^R(t)$, $\kappa_j(t) = \kappa_j^R(t)$, *and* $\zeta_j(t) = \zeta_j^R(t)$.

Proof. The required result follows immediately from Lemmata 7.2.1 and 7.4.6. □

Lemma 7.2.1 implies that the sets $\hat{V}$ and V in the following definition are both finite, and can therefore be enumerated as described. This definition modifies the definition of *valley crossings* in Chapter 6, and will be in effect throughout this chapter.

Definition 7.4.9. *Suppose that Assumption 7.3.1 holds. For any* $j \in \Lambda_{\mathbf{f}}$ *and* $k \in \{0, 1, \ldots, p\}$, $t^* \in T$ *is a* (j-)valley (k-)crossing *if both*

$$(\sigma_{j,k}^L(t^*), \sigma_{j,k}^R(t^*)) \in \{(-1, +1), (+1, -1)\},$$

and $\sigma_{j,q}^L(t^*) = \sigma_{j,q}^R(t^*) = 0$ *for each* $q \in \{0, 1, \ldots, (k-1)\}$. *Equivalently,* $t^* \in T$ *is a (j-)valley (k-)crossing if both*

$$\kappa_j^L(t^*) = \kappa_j^R(t^*) = k \qquad \textit{and} \qquad (\zeta_j^L(t^*), \zeta_j^R(t^*)) \in \{(-1, +1), (+1, -1)\}.$$

Denote the set of all valley crossings in $[t_0, t_f]$ *as* $\hat{V}$*, and set* $V := \hat{V} \cup \{t_0, t_f\}$*. Enumerate the elements of* V *as*

$$t_0 =: \omega_0 < \omega_1 < \ldots < \omega_\lambda := t_f.$$

For any $j \in \Lambda_{\mathbf{f}}$ *and* $k \in \{0, 1, \ldots, p\}$*,* $t^* \in T$ *is a* (j-)valley (k-)tracing switch *if*

$$(\sigma^L_{j,k}(t^*), \sigma^R_{j,k}(t^*)) \in \{(-1, 0), (+1, 0), (0, -1), (0, +1)\}.$$

As in Chapter 6, the "valleys" in the above definition refer to the shape of the absolute-value function's graph. Observe that the valley crossings described in Chapter 6 are all valley 0-crossings. The following lemma generalizes Corollary 6.3.19, by showing that each valley q^*-tracing switch is also a valley q-crossing for some $q \leq q^*$.

Lemma 7.4.10. *Suppose that Assumption 7.3.1 holds. If* $t^* \in T$ *is an* i^**-valley* q^**-tracing switch, with* $i^* \in \Lambda_{\mathbf{f}}$ *and* $q^* \in \{0, 1, \ldots, p\}$*, then there exist* $i \in \Lambda_{\mathbf{f}}$ *and* $q \in \{0, 1, \ldots, q^*\}$ *such that* t^* *is also an* i*-valley* q*-crossing.*

Proof. Since t^* is a valley q^*-tracing switch by assumption, there exists a least element k^* of $\{0, 1, \ldots, q^*\}$ for which t^* is a valley k^*-tracing switch. Let j^* be the least element of $\Lambda_{\mathbf{f}}$ for which t^* is a j^*-valley k^*-tracing switch.

If $k^* = 0$, then Corollary 6.3.19 yields the required result. Thus, assume that $k^* \geq 1$. To obtain a contradiction, suppose that t^* is not an i-valley q-crossing for any $i \in \Lambda_{\mathbf{f}}$ and $q \in \{0, 1, \ldots, k^*\}$. Thus, Definition 7.4.9 and the definition of k^* show that for each $j \in \Lambda_{\mathbf{f}}$, exactly one of the following cases holds:

1. both $\kappa^L_j(t^*) \geq k^*$ and $\kappa^R_j(t^*) \geq k^*$, and so $\sigma^L_{j,k} = \sigma^R_{j,k} = 0$ for each $k \in \{0, \ldots, k^* - 1\}$, or

2. both $\kappa^L_j(t^*) = \kappa^R_j(t^*) < k^*$ and $\zeta^L_j(t^*) = \zeta^R_j(t^*) \in \{-1, +1\}$.

In the first case, define a quantity $\bar{\zeta}_j := 0$; in the second case, define $\bar{\zeta}_j := \zeta^L_j(t^*) = \zeta^R_j(t^*) \in \{-1, +1\}$. Define $\bar{\delta}_j := \operatorname{unsign} \bar{\zeta}_j \in \{0, +1\}$. Thus, there exists a neighborhood $N \subset \hat{T}$ of t^* for which:

$$\text{fsign}\,(u_{(j)}(t), \dot{u}_{(j),1}(t), \ldots, \dot{u}_{(j),k^*-1}(t)) = \bar{\zeta}_j, \qquad \forall t \in N \backslash \{t^*\}, \quad \forall j \in \Lambda_{\mathbf{f}},$$

and

$$\text{unsign}\,(u_{(j)}(t), \dot{u}_{(j),1}(t), \ldots, \dot{u}_{(j),k^*-1}(t)) = \bar{\delta}_j, \qquad \forall t \in N \backslash \{t^*\}, \quad \forall j \in \Lambda_{\mathbf{f}}.$$

Define a function $\tilde{\mathbf{h}} : N \times \mathbb{R}^n \to \mathbb{R}^n$ so that $\tilde{\mathbf{h}}(t, \mathbf{a})$ is evaluated according to the following factored representation.

$\tilde{\mathbf{v}}_{(0)} \leftarrow (0, \mathbf{a}) \in \mathbb{R}^{n+1}$
for $j = 1$ **to** ℓ **do**
  $\tilde{\mathbf{u}}_{(j)} \leftarrow [\tilde{\mathbf{v}}_{(i)}]_{i \prec j}$
  if $j \in \Lambda_{\mathbf{f}}$ **then**
    $\tilde{v}_{(j)} \leftarrow \bar{\zeta}_j\, \tilde{u}_{(j)} + \bar{\delta}_j\, |\tilde{u}_{(j)}|$
  else
    $\tilde{\mathbf{v}}_{(j)} \leftarrow \mathbf{J}\psi_{(j)}(\mathbf{u}_{(j)}(t, \mathbf{x}(t, \mathbf{c}_0)))\, \tilde{\mathbf{u}}_{(j)}$
  end if
end for
$\tilde{\mathbf{h}}(t, \mathbf{a}) \leftarrow \tilde{\mathbf{v}}_{(\ell)}$

Observe that the mappings $t \mapsto \mathbf{u}_{(j)}(t, \mathbf{x}(t, \mathbf{c}_0))$ are L/R-analytic. This observation, Corollary 7.4.3, the definitions of $\bar{\zeta}_j$ and $\bar{\delta}_j$, and Lemmata 7.2.3 and 7.2.6 show that the ODE

$$\frac{d\mathbf{y}}{dt}(t) = \tilde{\mathbf{h}}(t, \mathbf{y}(t)), \qquad \mathbf{y}(t^*) = [\mathbf{x}_{t^*}]_{\mathbf{c}_0,\mathbf{M}}^{(k^*-1)}(\mathbf{m}_{(k^*)}) \tag{7.8}$$

is a Carathéodory ODE when restricted to $t \in N$. Comparing the definition of $\tilde{\mathbf{h}}$ with Algorithm 5, observe that

$$\phi(k^*, t, \mathbf{a}) = \tilde{\mathbf{h}}(t, \mathbf{a}), \qquad \forall t \in N \backslash \{t^*\}, \quad \forall \mathbf{a} \in \mathbb{R}^n.$$

Remark 7.4.5 then implies that the mapping $t \mapsto [\mathbf{x}_t]_{\mathbf{c}_0,\mathbf{M}}^{(k^*-1)}(\mathbf{m}_{(k^*)})$ solves the ODE (7.8) on N, and is thus the unique solution of this ODE on N, since $\tilde{\mathbf{h}}(t, \cdot)$ is evidently locally Lipschitz continuous for each fixed $t \in N$.

Since t^* is not a valley q-crossing for any $q \in \{0, 1, \ldots, k^*\}$, it follows that, for each $j \in \Lambda_{\mathbf{f}}$ for which $\bar{\delta}_j = +1$,

$$(\sigma^L_{j,k^*}(t^*), \sigma^R_{j,k^*}(t^*)) \in \{(-1,0),(+1,0),(-1,-1),(0,0),(+1,+1),(0,-1),(0,+1)\}.$$

Each of the above cases implies that, choosing N to be a smaller neighborhood of t^* if necessary, for each $j \in \Lambda_{\mathbf{f}}$ for which $\bar{\delta}_j = +1$, there exists $s_j \in \{-1,+1\}$ such that

$$\left|\tilde{u}_{(j)}\left(t, [\mathbf{x}_t]^{(k^*-1)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{m}_{(k^*)})\right)\right| = s_j\, \tilde{u}_{(j)}\left(t, [\mathbf{x}_t]^{(k^*-1)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{m}_{(k^*)})\right), \qquad \forall t \in N\backslash\{t^*\}.$$

Define $s_j := 0$ for each $j \in \Lambda_{\mathbf{f}}$ for which $\bar{\delta}_j = 0$. Thus, define a function $\hat{\mathbf{h}} : N \times \mathbb{R}^n \to \mathbb{R}^n$ so that $\hat{\mathbf{h}}(t,\mathbf{a})$ is evaluated according to the following factored representation.

$\hat{\mathbf{v}}_{(0)} \leftarrow (0,\mathbf{a}) \in \mathbb{R}^{n+1}$
for $j = 1$ **to** ℓ **do**
  $\hat{\mathbf{u}}_{(j)} \leftarrow [\hat{\mathbf{v}}_{(i)}]_{i \prec j}$
  if $j \in \Lambda_{\mathbf{f}}$ **then**
    $\hat{v}_{(j)} \leftarrow (\bar{\zeta}_j + s_j\bar{\delta}_j)\, \hat{u}_{(j)}$
  else
    $\hat{\mathbf{v}}_{(j)} \leftarrow \mathbf{J}\psi_{(j)}(\mathbf{u}_{(j)}(t,\mathbf{x}(t,\mathbf{c}_0)))\, \hat{\mathbf{u}}_{(j)}$
  end if
end for
$\hat{\mathbf{h}}(t,\mathbf{a}) \leftarrow \hat{\mathbf{v}}_{(\ell)}$

Observe that $\hat{\mathbf{h}}$ is $\mathcal{C}^\omega$, and therefore locally Lipschitz continuous, on its domain of definition. Moreover,

$$\tilde{\mathbf{h}}(t,\mathbf{a}) = \hat{\mathbf{h}}(t,\mathbf{a}), \qquad \forall t \in N\backslash\{t^*\}.$$

The above discussion implies that the ODE:

$$\frac{d\mathbf{y}}{dt}(t) = \hat{\mathbf{h}}(t,\mathbf{y}(t)), \qquad \mathbf{y}(t^*) = [\mathbf{x}_{t^*}]^{(k^*-1)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{m}_{(k^*)})$$

is a Carathéodory ODE for $t \in N$, and that the mapping $t \mapsto [\mathbf{x}_t]^{(k^*-1)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{m}_{(k^*)})$ also solves this ODE uniquely on N. Since $\hat{\mathbf{h}}$ is $\mathcal{C}^\omega$, the mapping $t \mapsto [\mathbf{x}_t]^{(k^*-1)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{m}_{(k^*)})$ is also $\mathcal{C}^\omega$ on N [35]. Thus, the mapping $t \mapsto \hat{u}_{(j^*)}(t, [\mathbf{x}_t]^{(k^*-1)}_{\mathbf{c}_0,\mathbf{M}}(\mathbf{m}_{(k^*)}))$ is also $\mathcal{C}^\omega$ on N. By construction, though,

$$\hat{u}_{(j^*)}\left(t, [\mathbf{x}_t]_{\mathbf{c}_0,\mathbf{M}}^{(k^*-1)}(\mathbf{m}_{(k^*)})\right) = \dot{u}_{(j^*),k^*}(t), \qquad \forall t \in N\backslash\{t^*\}. \tag{7.9}$$

Since t^* is a j^*-valley k^*-tracing switch, either $\sigma^L_{j^*,k^*}(t^*) = 0$ or $\sigma^R_{j^*,k^*}(t^*) = 0$. Thus, there exists $\delta^* > 0$ such that, for either each $t \in (t^* - \delta^*, t^*)$ or each $t \in (t^*, t^* + \delta^*)$,

$$\hat{u}_{(j^*)}\left(t, [\mathbf{x}_t]_{\mathbf{c}_0,\mathbf{M}}^{(k^*-1)}(\mathbf{m}_{(k^*)})\right) = 0.$$

The above analyticity result and [66, Corollary 1.2.6] then imply that

$$\hat{u}_{(j^*)}\left(t, [\mathbf{x}_t]_{\mathbf{c}_0,\mathbf{M}}^{(k^*-1)}(\mathbf{m}_{(k^*)})\right) = 0, \qquad \forall t \in N.$$

Equation (7.9) then implies that both $\sigma^L_{j^*,k^*}(t^*) = 0$ and $\sigma^R_{j^*,k^*}(t^*) = 0$, which contradicts the definitions of j^* and k^*.

Thus, there exist $i \in \Lambda_{\mathbf{f}}$ and $q \in \{0, 1, \ldots, k^*\}$ such that t^* is an i-valley q-crossing. Since $k^* \leq q^*$ by construction, the required result follows immediately. □

Corollary 7.4.11. *Suppose that Assumption 7.3.1 holds. If there exist $t^* \in T$ and $j \in \Lambda_{\mathbf{f}}$ such that $\kappa^L_j(t^*) \neq \kappa^R_j(t^*)$, then there exist $i \in \Lambda_{\mathbf{f}}\backslash\{j\}$ and*

$$q \in \left\{0, 1, \ldots, \min\left\{\kappa^L_j(t^*), \kappa^R_j(t^*)\right\}\right\}$$

such that t^* is an i-valley q-crossing.

Proof. Define $k^* := \min\{\kappa^L_j(t^*), \kappa^R_j(t^*)\} \in \{0, 1, \ldots, p\}$. Since $\kappa^L_j(t^*) \neq \kappa^R_j(t^*)$, the definition of κ_j implies that t^* is a j-valley k^*-tracing switch. Moreover, the definition of k^* shows that t^* is not a j-valley crossing. Lemma 7.4.10 then yields the required result. □

Corollary 7.4.12. *Suppose that Assumption 7.3.1 holds. For each $i^* \in \{1, \ldots, \lambda\}$, $j^* \in \Lambda_{\mathbf{f}}$, and $k^* \in \{0, 1, \ldots, p\}$, the mappings $t \mapsto \sigma^R_{j^*,k^*}(t)$, $t \mapsto \kappa^R_{j^*}(t)$, and $t \mapsto \zeta^R_{j^*}(t)$ are constant on $[\omega_{i^*-1}, \omega_{i^*})$.*

Moreover, if $i^* \in \{1, \ldots, \lambda - 1\}$, and ω_{i^*} is not a valley k-crossing for any $k \leq k^*$, then $\sigma^R_{j^*,k^*}(\omega_{i^*-1}) = \sigma^R_{j^*,k^*}(\omega_{i^*})$. If, in addition, $\kappa^R_{j^*}(\omega_{i^*-1}) \leq k^*$, then $\kappa^R_{j^*}(\omega_{i^*-1}) = \kappa^R_{j^*}(\omega_{i^*})$ and $\zeta^R_{j^*}(\omega_{i^*-1}) = \zeta^R_{j^*}(\omega_{i^*}) \in \{-1, +1\}$.

Proof. Lemma 7.4.10 and Corollary 7.4.11 imply that any discontinuity in the mappings $t \in [t_0, t_f] \mapsto \sigma^R_{j,k}(t)$ or $t \in [t_0, t_f] \mapsto \kappa^R_j(t)$ is also a valley crossing. Hence, for each $i \in \{1, \ldots, \lambda\}$, these mappings are constant on (ω_{i-1}, ω_i).

By construction, the mappings $t \in [t_0, t_f] \mapsto \sigma^R_{j,k}(t)$ and $t \in [t_0, t_f] \mapsto \kappa^R_j(t)$ are each right-continuous; the first required result follows immediately. Now, consider any $i^* \in \{1, \ldots, \lambda - 1\}$ for which ω_{i^*} is not a valley k-crossing for any $k \leq k^*$. The first result of this corollary shows that $\sigma^R_{j^*,k^*}(\omega_{i^*-1}) = \sigma^L_{j^*,k^*}(\omega_{i^*})$. Thus, Lemma 7.4.10 and the construction of i^* show that

$$\begin{aligned}
&(\sigma^R_{j^*,k^*}(\omega_{i^*-1}), \sigma^R_{j^*,k^*}(\omega_{i^*})) \\
&\qquad = (\sigma^L_{j^*,k^*}(\omega_{i^*}), \sigma^R_{j^*,k^*}(\omega_{i^*})), \\
&\qquad \notin \{(-1,+1), (+1,-1), (-1,0), (+1,0), (0,-1), (0,+1)\},
\end{aligned}$$

which implies that $\sigma^R_{j^*,k^*}(\omega_{i^*-1}) = \sigma^R_{j^*,k^*}(\omega_{i^*})$, as required. Suppose in addition that $\kappa^R_{j^*}(\omega_{i^*-1}) \leq k^*$. The first result of this corollary shows that $\kappa^R_{j^*}(\omega_{i^*-1}) = \kappa^L_{j^*}(\omega_{i^*})$ and $\zeta^R_{j^*}(\omega_{i^*-1}) = \zeta^L_{j^*}(\omega_{i^*})$. Corollary 7.4.11 and the construction of i^* then show that

$$\kappa^R_{j^*}(\omega_{i^*-1}) = \kappa^L_{j^*}(\omega_{i^*}) = \kappa^R_{j^*}(\omega_{i^*}) =: q^* \leq k^*,$$

as required. Thus, $\zeta^R_{j^*}(\omega_{i^*-1})$ and $\zeta^R_{j^*}(\omega_{i^*})$ are each nonzero. Noting that ω_{i^*} is not a valley q^*-crossing, the definitions of $\kappa^R_{j^*}$ and $\zeta^R_{j^*}$ imply that

$$\zeta^R_{j^*}(\omega_{i^*-1}) = \zeta^L_{j^*}(\omega_{i^*}) = \zeta^R_{j^*}(\omega_{i^*}),$$

as required. □

The following lemma shows that between any two successive valley crossings, the LD-derivative mapping $t \mapsto [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})$ evolves as the unique solution of a linear ODE.

Lemma 7.4.13. *Suppose that Assumption 7.3.1 holds. The LD-derivative mapping $t \mapsto [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})$ is the unique solution on $[t_0, t_f]$ of the Carathéodory ODE:*

$$\frac{d\mathbf{A}}{dt}(t) = \tilde{\mathbf{H}}(t, \mathbf{A}(t)), \qquad \mathbf{A}(t_0) = \mathbf{M}, \tag{7.10}$$

where the function $\tilde{\mathbf{H}} : [t_0, t_f] \times \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$ *is defined so that:*

$$\tilde{\mathbf{H}}(t, \mathbf{A}) := \begin{cases} \hat{\mathbf{H}}(\boldsymbol{\zeta}^R(\omega_{i-1}), t, \mathbf{A}), & \text{if } i \in \{1, \ldots, \lambda\} \text{ and } t \in [\omega_{i-1}, \omega_i), \\ \hat{\mathbf{H}}(\boldsymbol{\zeta}^R(\omega_{\lambda-1}), t_f, \mathbf{A}), & \text{if } t = t_f, \end{cases}$$

with the function $\hat{\mathbf{H}} : \{-1, 0, +1\}^\ell \times T \times \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$ *defined according to Algorithm 6.*

Algorithm 6 Evaluates $\hat{\mathbf{H}}(\hat{\boldsymbol{\zeta}}, t, \mathbf{A})$, with $\mathbf{f}$ and $\mathbf{x}$ described by Assumption 7.3.1

Require: $q \in \mathbb{N}$, $\hat{\boldsymbol{\zeta}} \in \{-1, 0, +1\}^\ell$, $t \in T$, and $\mathbf{A} \in \mathbb{R}^{n\times q}$
  $\hat{\mathbf{V}}_{(0)} \leftarrow (\mathbf{0}_{1\times q}, \mathbf{A})$
  for $j = 1$ **to** ℓ **do**
    $\hat{\mathbf{U}}_{(j)} \leftarrow [\hat{\mathbf{V}}_{(i)}]_{i \prec j}$
    if $j \in \Lambda_{\mathbf{f}}$ **then**
      $\hat{\mathbf{V}}_{(j)} \leftarrow \hat{\zeta}_j \, \hat{\mathbf{U}}_{(j)}$
    else
      $\hat{\mathbf{V}}_{(j)} \leftarrow \mathbf{J}\boldsymbol{\psi}_{(j)}(\mathbf{u}_{(j)}(t, \mathbf{x}(t, \mathbf{c}_0))) \, \hat{\mathbf{U}}_{(j)}$
    end if
  end for
  return $\hat{\mathbf{H}}(\hat{\boldsymbol{\zeta}}, t, \mathbf{A}) = \hat{\mathbf{V}}_{(\ell)}$

Proof. Consider the finite set $Z \subset [t_0, t_f]$ described by Lemma 7.4.8. Using the results of Lemma 7.4.8 and Corollary 7.4.12, inspection of Algorithms 4 and 6 shows that

$$\tilde{\mathbf{H}}(t, [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})) = [\mathbf{f}_t]'(\mathbf{x}(t, \mathbf{c}_0); [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})), \qquad \forall t \in [t_0, t_f] \backslash (Z \cup V).$$

Since both Z and V have zero Lebesgue measure, comparison of the ODEs (7.10) and (7.3) shows that the mapping $t \mapsto [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})$ solves (7.10) on $[t_0, t_f]$. Observing that the mapping $\hat{\mathbf{H}}(\hat{\boldsymbol{\zeta}}, t, \cdot)$ is linear, and thus Lipschitz continuous, for each fixed $\hat{\boldsymbol{\zeta}} \in \{-1, 0, +1\}^\ell$ and $t \in [t_0, t_f]$, it follows that $t \mapsto [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})$ is the unique solution of (7.10) on $[t_0, t_f]$. □

The following theorem extends the above results to show that, between any two successive valley crossings, the LD-derivative mapping $t \mapsto [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})$ evolves according to a classical ODE sensitivity system, as described, for example, by [35, Ch. V, Theorem 3.1].

Theorem 7.4.14. *Suppose that Assumption 7.3.1 holds, and consider any particular $i \in \{1, \ldots, \lambda\}$. Consider the following parametric ODE, with $\mathbf{p} \in \mathbb{R}^p$ denoting a parameter, and with a $\mathcal{C}^\omega$ function $\mathbf{h}(i, \cdot, \cdot) : [\omega_{i-1}, \omega_i] \times X \to \mathbb{R}^n$ defined according to Algorithm 7.*

$$\frac{d\boldsymbol{\xi}}{dt}(t, \mathbf{p}) = \mathbf{h}(i, t, \boldsymbol{\xi}(t, \mathbf{p})), \qquad \boldsymbol{\xi}(\omega_{i-1}, \mathbf{p}) = \mathbf{x}(\omega_{i-1}, \mathbf{c}_0) + [\mathbf{x}_{\omega_{i-1}}]'(\mathbf{c}_0; \mathbf{M})\, \mathbf{p}. \quad (7.11)$$

This ODE has a unique solution $\{\boldsymbol{\xi}(t, \mathbf{p}) : t \in [\omega_{i-1}, \omega_i]\}$ for each $\mathbf{p}$ in some neighborhood of $\mathbf{0} \in \mathbb{R}^p$. For each $t \in [\omega_{i-1}, \omega_i]$, the mapping $\boldsymbol{\xi}(t, \cdot)$ is differentiable at $\mathbf{0}$; the partial derivative mapping $t \mapsto \frac{\partial \boldsymbol{\xi}}{\partial \mathbf{p}}(t, \mathbf{0})$ is the unique solution on $[\omega_{i-1}, \omega_i]$ of the ODE:

$$\frac{d\mathbf{A}}{dt}(t) = \frac{\partial \mathbf{h}}{\partial \mathbf{z}}(i, t, \boldsymbol{\xi}(t, \mathbf{0}))\, \mathbf{A}(t), \qquad \mathbf{A}(\omega_{i-1}) = [\mathbf{x}_{\omega_{i-1}}]'(\mathbf{c}_0; \mathbf{M}). \quad (7.12)$$

Moreover, for each $t \in [\omega_{i-1}, \omega_i]$, $\mathbf{x}(t, \mathbf{c}_0) = \boldsymbol{\xi}(t, \mathbf{0})$ and $[\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M}) = \frac{\partial \boldsymbol{\xi}}{\partial \mathbf{p}}(t, \mathbf{0})$.

Algorithm 7 Evaluates $\mathbf{h}(i, t, \mathbf{z})$

Require: $i \in \{1, \ldots, \lambda\}$, $t \in [\omega_{i-1}, \omega_i]$, and $\mathbf{z} \in X$
  $\hat{\mathbf{v}}_{(0)} \leftarrow (t, \mathbf{z})$
  for $j = 1$ **to** ℓ **do**
    $\hat{\mathbf{u}}_{(j)} \leftarrow [\hat{\mathbf{v}}_{(i)}]_{i \prec j}$
    if $j \in \Lambda_{\mathbf{f}}$ **then**
      $\hat{v}_{(j)} \leftarrow \zeta_j^R(\omega_{i-1})\, \hat{u}_{(j)}$
    else
      $\hat{\mathbf{v}}_{(j)} \leftarrow \boldsymbol{\psi}_{(j)}(\hat{\mathbf{u}}_{(j)})$
    end if
  end for
  return $\mathbf{h}(i, t, \mathbf{z}) = \hat{\mathbf{v}}_{(\ell)}$

Proof. Consider any fixed $t \in [\omega_{i-1}, \omega_i] \backslash Z$ and $j \in \Lambda_{\mathbf{f}}$. Lemma 7.4.8 and Corollary 7.4.12 imply that $\zeta_j(t) = \zeta_j^R(\omega_{i-1})$. Now, if $u_{(j)}(t) = 0$, then $|u_{(j)}(t)| = 0 = \zeta_j(t)\, u_{(j)}(t)$. On the other hand, if $u_{(j)}(t) \neq 0$, then $\zeta_j(t) = \sigma_{j,0}(t) = \operatorname{sign} u_{(j)}(t) \in$

$\{-1, +1\}$, and so $|u_{(j)}(t)| = \zeta_j(t)\, u_{(j)}(t)$. Combining the above results, and permitting variation in t and j, it follows that

$$v_{(j)}(t) = |u_{(j)}(t)| = \zeta_j^R(\omega_{i-1})\, u_{(j)}(t), \qquad \forall t \in [\omega_{i-1}, \omega_i] \backslash Z, \quad \forall j \in \Lambda_{\mathbf{f}}.$$

Thus, noting that the set Z is finite, inspection of the ODE (7.11) shows that $\mathbf{x}(\cdot, \mathbf{c}_0)$ solves (7.11) with $\mathbf{p} := \mathbf{0}$ on $[\omega_{i-1}, \omega_i]$. By inspection of Algorithm 7, there exists an open set $T_i \supset [\omega_{i-1}, \omega_i]$ for which $\mathbf{h}(i, \cdot, \cdot)$ is in fact well-defined on $T_i \times X$, since $\mathbf{h}(i, \cdot, \cdot)$ is a finite composition of locally Lipschitz continuous functions on open sets. Observe that $\mathbf{h}(i, \cdot, \cdot)$ is $\mathcal{C}^\omega$ on $T_i \times X$, since Algorithm 7 expresses this function as a composition of $\mathcal{C}^\omega$ functions. Thus, since $\mathbf{h}(i, \cdot, \cdot)$ is locally Lipschitz continuous, it follows that $\mathbf{x}(\cdot, \mathbf{c}_0)$ is the unique solution of (7.11) with $\mathbf{p} := \mathbf{0}$ on $[\omega_{i-1}, \omega_i]$, as required. Standard ODE theory [35, Ch. V, Theorem 2.1] then implies that the ODE (7.11) has a unique solution on $[\omega_{i-1}, \omega_i]$ for each $\mathbf{p}$ in some neighborhood N of $\mathbf{0} \in \mathbb{R}^p$, as required.

[35, Ch. V, Theorem 3.1] then implies that, for each $t \in [\omega_{i-1}, \omega_i]$, the mapping $\boldsymbol{\xi}(t, \cdot)$ is differentiable at $\mathbf{0}$, and that the partial derivative mapping $t \mapsto \frac{\partial \boldsymbol{\xi}}{\partial \mathbf{p}}(t, \mathbf{0})$ solves the ODE (7.12) uniquely on $[\omega_{i-1}, \omega_i]$.

Lastly, for each $t \in [\omega_{i-1}, \omega_i]$ and each $\mathbf{A} \in \mathbb{R}^{n \times p}$, inspection of Algorithms 6 and 7 shows that if the vector forward mode of automatic differentiation [34] is applied to Algorithm 7 to compute the quantity

$$\frac{\partial \mathbf{h}}{\partial \mathbf{z}}(i, t, \boldsymbol{\xi}(t, \mathbf{0}))\, \mathbf{A} = \frac{\partial \mathbf{h}}{\partial \mathbf{z}}(i, t, \mathbf{x}(t, \mathbf{c}_0))\, \mathbf{A},$$

then the result is in fact $\hat{\mathbf{H}}(\boldsymbol{\zeta}^R(\omega_{i-1}), t, \mathbf{A})$. It follows immediately that the ODEs (7.10) and (7.12) have the same unique solution on $[\omega_{i-1}, \omega_i]$, which implies that $[\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M}) = \frac{\partial \boldsymbol{\xi}}{\partial \mathbf{p}}(t, \mathbf{0})$ for each $t \in [\omega_{i-1}, \omega_i]$. $\square$

The ODE (7.11) will be referred to as a *bank-locked ODE*, since each of the absolute-value function "valleys" in the right-hand side of the ODE (7.2) have been *locked* into one of its two linear "banks". Similarly, the linear ODE (7.12) will be referred to as a *bank-locked sensitivity ODE*. To compute $[\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})$, the ODEs (7.11)

and (7.12) can be solved simultaneously on $[\omega_{i-1}, \omega_i]$ for each $i \in \{1, \ldots, \lambda\}$, using any ODE solver with sensitivity analysis capabilities. Standard approaches [24, 73] for exploiting sparsity and redundancy in sensitivity systems remain applicable between each pair of successive valley crossings. This approach, however, requires computing the valley crossing ω_{i-1} and the critical signatures $\boldsymbol{\zeta}^R(\omega_{i-1})$ for each $i \in \{1, \ldots, \lambda\}$, since these quantities are not known *a priori*. The following sections discuss evaluation of these quantities.

7.4.3 Determining tracing depths and critical signatures

This section lays a theoretical foundation for determining the valley-tracing depths $\boldsymbol{\kappa}^R(\omega_i)$ and the critical signatures $\boldsymbol{\zeta}^R(\omega_i)$ for each $i \in \{0, 1, \ldots, \lambda - 1\}$. The results in this section are subject to the following remark.

Remark 7.4.15. *In the remainder of this section, Assumption 7.3.1 will be in effect, $i \in \{0, 1, \ldots, \lambda - 1\}$ will be fixed, and ω_i will be assumed to be known. Though λ is not known* a priori, *observe that $i < \lambda$ if and only if $\omega_i < t_f$. Thus, ω_{i+1} exists and is no greater than t_f; though ω_{i+1} is not known* a priori *either, its existence will be convenient when formulating the results in this section.*

For consistency, consider t_f to be a valley k-crossing for each $k \in \{0, 1, \ldots, p\}$. This consideration will simplify the presentation of results in this section; for example, the $i = \lambda - 1$ case will not need to be treated differently from the $i < \lambda - 1$ case.

Lemma 7.4.16. *Suppose that Remark 7.4.15 holds. If $t^* \in (\omega_i, t_f]$ is such that there are no valley 0-crossings in (ω_i, t^*), then*

$$\sigma^R_{j,0}(\omega_i) = \operatorname{sign} \int_{\omega_i}^{t^*} u_{(j)}(t, \mathbf{x}(t, \mathbf{c}_0))\, dt, \qquad \forall j \in \Lambda_{\mathbf{f}}. \tag{7.13}$$

Similarly, for any $q \in \{1, \ldots, p\}$, if $\hat{t} \in (\omega_i, t_f]$ is such that there are no valley k-crossings in $(\omega_i, \hat{t})$ for any $k \leq q$, then

$$\sigma^R_{j,q}(\omega_i) = \operatorname{sign} \int_{\omega_i}^{\hat{t}} \dot{u}_{(j),q}(t)\, dt, \qquad \forall j \in \Lambda_{\mathbf{f}}.$$

Proof. The required results follow from Corollary 7.4.12 and Lemma 7.4.8. Note that, for each $j \in \Lambda_{\mathbf{f}}$, the mapping $t \mapsto u_{(j)}(t, \mathbf{x}(t, \mathbf{c}_0))$ is continuous, and is therefore integrable on the compact set $[\omega_i, t^\star]$. Moreover, for each $j \in \Lambda_{\mathbf{f}}$ and $q \in \{1, \ldots, p\}$, $\dot{u}_{(j),q}$ is L/R-analytic, and is therefore measurable. Inspection of Algorithm 4 shows that $\dot{u}_{(j),q}$ is also bounded on $[\omega_i, \hat{t}]$, and is therefore integrable on $[\omega_i, \hat{t}]$. □

Observe that Theorem 5.2.4 provides necessary conditions for $\sigma^R_{j,0}(\omega_i)$ to be zero; these conditions can be exploited to avoid computing the integral (7.13) in certain situations. The following result is an immediate corollary of Theorem 5.2.4, noting that each mapping $t \mapsto u_{(j)}(t, \mathbf{x}(t, \mathbf{c}_0))$ is both L/R-analytic and continuous.

Corollary 7.4.17. *Choose any $j \in \Lambda_{\mathbf{f}}$. For any $\epsilon_\sigma \geq 0$, if $|u_{(j)}(\omega_i, \mathbf{x}(\omega_i, \mathbf{c}_0))| > \epsilon_\sigma$, then*

$$\zeta^R_j(\omega_i) = \sigma^R_{j,0}(\omega_i) = \operatorname{sign} u_{(j)}(\omega_i, \mathbf{x}(\omega_i, \mathbf{c}_0)) \in \{-1, +1\},$$

and $\kappa^R_j(\omega_i) = 0$. Similarly, if the directional derivative

$$d_j := [u_{(j)}]'((\omega_i, \mathbf{x}(\omega_i, \mathbf{c}_0)); (1, \mathbf{f}(\omega_i, \mathbf{x}(\omega_i, \mathbf{c}_0)))$$

satisfies $|d_j| > \epsilon_\sigma$, then

$$\zeta^R_j(\omega_i) = \sigma^R_{j,0}(\omega_i) = \operatorname{fsign}(u_{(j)}(\omega_i, \mathbf{x}(\omega_i, \mathbf{c}_0)), d_j) \in \{-1, +1\},$$

and $\kappa^R_j(\omega_i) = 0$.

For any $j \in \Lambda_{\mathbf{f}}$, the following quadrature ODE in w_j may be appended to the original ODE (7.2), to compute the integral (7.13):

$$\frac{dw_j}{dt}(t) = u_{(j)}(t, \mathbf{x}(t, \mathbf{c}_0)), \qquad w_j(\omega_i) = 0. \tag{7.14}$$

For each $j \in \Lambda_{\mathbf{f}}$, $\sigma^R_{j,0}(\omega_i)$ can be determined from Corollary 7.4.12, Lemma 7.4.16, or Corollary 7.4.17. Then, for each $j \in \Lambda_{\mathbf{f}}$, $\kappa^R_j(\omega_i) = 0$ if and only if $\sigma^R_{j,0}(\omega_i) \in \{-1, +1\}$, in which case $\zeta^R_{j,0}(\omega_i) \in \{-1, +1\}$. Thus, as a strong inductive assumption, suppose that, for some $q^\star \in \{1, \ldots, p\}$, the set

$$J(q^*) := \{j \in \Lambda_{\mathbf{f}} : q^\star \leq \kappa_j^R(\omega_i)\}$$

is known, as are $\kappa_j^R(\omega_i)$ and $\zeta_j^R(\omega_i)$ for each $j \in \Lambda_{\mathbf{f}} \backslash J(q^\star)$. Collect the known components of $\boldsymbol{\zeta}_j^R(\omega_i)$ in a vector $\bar{\boldsymbol{\zeta}} \in \{-1, 0, +1\}^\ell$, defined so that

$$\bar{\zeta}_j := \begin{cases} \zeta_j^R(\omega_i) \in \{-1, +1\}, & \text{if } j \in \Lambda_{\mathbf{f}} \backslash J(q^\star), \\ 0, & \text{if } j \in J(q^\star). \end{cases}$$

As with $\boldsymbol{\zeta}_j^R$, the components $\bar{\zeta}_j$ for which $j \notin \Lambda_{\mathbf{f}}$ will not be used. Assume that $J(q^\star)$ is nonempty; if $J(q^\star)$ is empty, then $\boldsymbol{\zeta}^R(\omega_i) = \bar{\boldsymbol{\zeta}}$, and all critical signatures have been determined already.

Under the assumptions in the previous paragraph, the results presented in the remainder of this section seek to identify all $j \in J(q^\star)$ for which $\kappa_j^R(\omega_i) = q^\star$, and to identify the corresponding values of $\zeta_j^R(\omega_i) = \sigma_j^R(\omega_i) \in \{-1, +1\}$.

Lemma 7.4.18. *Consider functions*

$$\mathbf{f}^\star : \{-1, 0, +1\}^\ell \times T \times \mathbb{R}^n \to \mathbb{R}^n, \qquad \textit{and} \qquad \mathbf{g}^\star : \{-1, 0, +1\}^\ell \times T \times \mathbb{R}^n \to \mathbb{R}^n$$

as described in Algorithm 8, and the following probe ODEs *in* $\mathbf{y}$ *and* $\mathbf{z}$*:*

$$\frac{d\mathbf{y}}{dt}(t) = \mathbf{f}^\star(\bar{\boldsymbol{\zeta}}, t, \mathbf{y}(t)), \qquad \mathbf{y}(\omega_i) = [\mathbf{x}_{\omega_i}]'(\mathbf{c}_0; \mathbf{M})\, \mathbf{e}_{(q^\star)}, \tag{7.15}$$

$$\frac{d\mathbf{z}}{dt}(t) = \dot{\mathbf{z}}(t) := \mathbf{g}^\star(\bar{\boldsymbol{\zeta}}, t, \mathbf{y}(t)), \qquad \mathbf{z}(\omega_i) = \mathbf{0} \in \mathbb{R}^\ell. \tag{7.16}$$

Let $\omega^\star$ be the least element of $(\omega_i, t_f]$ for which there exists $k < q^\star$ such that $\omega^\star$ is a valley k-crossing. (Here, t_f is considered to be a valley k-crossing for each $k \in \{0, 1, \ldots, p\}$.) The above ODEs have unique solutions $\mathbf{y}$ and $\mathbf{z}$ on some open superset $T^\star$ of $[\omega_i, \omega^\star]$. Moreover, $\mathbf{y}(t) = [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})\, \mathbf{e}_{(q^\star)}$ for each $t \in [\omega_i, \omega^\star]$. The function $\dot{\mathbf{z}}$ is L/R-analytic; for each $t^\star \in T^\star$ and $j \in \Lambda_{\mathbf{f}}$, there exist $\eta_j^L(t^\star), \eta_j^R(t^\star) \in \{-1, 0, +1\}$ for which, for sufficiently small $\delta > 0$

$$\begin{aligned} &\operatorname{sign} \dot{z}_j(t) = \eta_j^L(t^\star), \quad \forall t \in [t^\star - \delta, t^\star), \\ \textit{and} \qquad &\operatorname{sign} \dot{z}_j(t) = \eta_j^R(t^\star), \quad \forall t \in (t^\star, t^\star + \delta]. \end{aligned} \tag{7.17}$$

For each $t \in (\omega_i, \omega^\star)$ and $j \in \Lambda_{\mathbf{f}}$, $\eta_j^L(t) = \sigma_{j,q^\star}^L(t)$ and $\eta_j^R(t) = \sigma_{j,q^\star}^R(t)$.

Algorithm 8 Computes $\mathbf{f}^\star(\hat{\boldsymbol{\zeta}}, t, \mathbf{a}) \in \mathbb{R}^n$ and $\mathbf{g}^\star(\hat{\boldsymbol{\zeta}}, t, \mathbf{a}) \in \mathbb{R}^\ell$, with $\mathbf{f}$ and $\mathbf{x}$ described by Assumption 7.3.1

Require: $\hat{\boldsymbol{\zeta}} \in \{-1, 0, +1\}^\ell$, $t \in T$, $\mathbf{a} \in \mathbb{R}^n$
  $\mathbf{v}_{(0)}^\star \leftarrow (0, \mathbf{a})$
  $\mathbf{g} \leftarrow \mathbf{0} \in \mathbb{R}^\ell$
  for $j = 1$ **to** ℓ **do**
    $\mathbf{u}_{(j)}^\star \leftarrow [\mathbf{v}_{(i)}^\star]_{i \prec j}$
    if $j \in \Lambda_{\mathbf{f}}$ **and** $\hat{\zeta}_j \neq 0$ **then**
      $v_{(j)}^\star \leftarrow \hat{\zeta}_j \, u_{(j)}^\star$
      $g_j \leftarrow u_{(j)}^\star$
    else if $j \in \Lambda_{\mathbf{f}}$ **and** $\hat{\zeta}_j = 0$ **then**
      $v_{(j)}^\star \leftarrow |u_{(j)}^\star|$
      $g_j \leftarrow u_{(j)}^\star$
    else if $j \notin \Lambda_{\mathbf{f}}$ **then**
      $\mathbf{v}_{(j)}^\star \leftarrow \mathbf{J}\boldsymbol{\psi}_{(j)}(\mathbf{u}_{(j)}(t, \mathbf{x}(t, \mathbf{c}_0)))\, \mathbf{u}_{(j)}^\star$
    end if
  end for
  return $\mathbf{f}^\star(\hat{\boldsymbol{\zeta}}, t, \mathbf{a}) := \mathbf{v}_{(\ell)}^\star$ and $\mathbf{g}^\star(\hat{\boldsymbol{\zeta}}, t, \mathbf{a}) := \mathbf{g}$

Proof. With ϕ defined according to Algorithm 5, it follows from the definition of $\bar{\boldsymbol{\zeta}}$, Lemma 7.4.8, and Corollary 7.4.12 that

$$\phi(q^\star, t, \mathbf{a}) = \mathbf{f}^\star(\bar{\boldsymbol{\zeta}}, t, \mathbf{a}), \qquad \forall \mathbf{a} \in \mathbb{R}^n, \quad \forall t \in [\omega_i, \omega^\star] \backslash Z.$$

Remark 7.4.5 then implies that the mapping $t \mapsto [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})\, \mathbf{e}_{(q^\star)}$ solves the ODE (7.15) uniquely on $[\omega_i, \omega^\star]$. Standard ODE theory implies that this solution can be extended to yield a unique solution $\mathbf{y}$ of (7.15) on some open superset $T_{\mathbf{y}} \subset T$ of $[\omega_i, \omega^\star]$; moreover, $\mathbf{y}$ coincides with the mapping $t \mapsto [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})\, \mathbf{e}_{(q^\star)}$ on $[\omega_i, \omega^\star]$.

Next, observe that the right-hand side of the ODE (7.16) does not contain a "$\mathbf{z}(t)$" term, and is thus equivalent to a simple integral. Inspection of Algorithm 8

shows that the mapping $\mathbf{f}^\star(\bar{\boldsymbol{\zeta}}, \cdot, \cdot)$ is locally Lipschitz continuous; Gronwall's Inequality then implies that $\mathbf{y}$ is locally Lipschitz continuous on $T_{\mathbf{y}}$. This observation and Algorithm 8 imply that $t \mapsto \dot{\mathbf{z}}(t) = \mathbf{g}^\star(\bar{\boldsymbol{\zeta}}, t, \mathbf{y}(t))$ is locally Lipschitz continuous on $T_{\mathbf{y}}$. Thus, there exists an open set $T^\star \subset T_{\mathbf{y}}$ for which $[\omega_i, \omega^\star] \subset T^\star$, and for which the mapping $\dot{\mathbf{z}}$ is bounded and Lipschitz continuous on $T^\star$. Since $\mathbf{y}$ and $\mathbf{x}(\cdot, \mathbf{c}_0)$ are L/R-analytic, Lemmata 6.2.5 and 6.2.6 imply that $\dot{\mathbf{z}}$ is also L/R-analytic, and therefore measurable. It follows that, for each $t^\star \in T^\star$, the solution

$$\mathbf{z}(t^\star) = \int_{\omega_i}^{t^\star} \dot{\mathbf{z}}(t)\, dt$$

of the ODE (7.16) is unique and well-defined in $\mathbb{R}^n$. Moreover, Lemma 7.2.1 implies the existence of quantities $\eta_j^L(t^\star), \eta_j^R(t^\star) \in \{-1, 0, +1\}$ satisfying (7.17) for each $t^\star \in T^\star$ and $j \in \Lambda_{\mathbf{f}}$. Comparing Algorithms 4 and 8, it follows from the definition of $\bar{\boldsymbol{\zeta}}$, Lemma 7.4.8, and Corollary 7.4.12 that

$$\dot{z}_j(t) = \dot{u}_{(j),q^\star}(t), \qquad \forall t \in [\omega_i, \omega^\star] \backslash Z, \quad \forall j \in \Lambda_{\mathbf{f}}.$$

Moreover, since Z is finite, there exists $\delta > 0$ such that, for each $t^\star \in (\omega_i, \omega^\star)$, there is no element of Z in the set $[t^\star - \delta, t^\star) \cup (t^\star, t^\star + \delta]$. The final claim of the lemma follows immediately. □

The following theorem permits determination of $\zeta_j^R(\omega_i) \in \{-1, +1\}$ for each $j \in \Lambda_{\mathbf{f}}$ for which $\kappa_j^R(\omega_i) = q^\star$, without requiring the latter knowledge *a priori*.

Theorem 7.4.19. *Suppose that the conditions of Lemma 7.4.18 hold. Using the notation introduced in Lemma 7.4.18, for each $t^\star \in (\omega_i, \omega^\star)$, $t^\star$ is a valley $q^\star$-crossing if and only if there exists $j \in J(q^\star)$ for which $(\eta_j^L(t^\star), \eta_j^R(t^\star)) \in \{(-1, +1), (+1, -1)\}$.*

Next, choose any $\tau^\star \in (\omega_i, \omega^\star]$ for which there are no valley $q^\star$-crossings in $(\omega_i, \tau^\star)$. For each $j \in J(q^\star)$, each of the following conditional statements holds. In these statements, z_j refers to the jth component of the unique solution $\mathbf{z}$ of the ODE (7.16).

- *If $z_j(\tau^\star) = 0$, then $\sigma_{j,q^\star}^R(\omega_i) = 0$, and $\kappa_j^R(\omega_i) > q^\star$.*
- *If $z_j(\tau^\star) > 0$, then $\zeta_{j,q^\star}^R(\omega_i) = \sigma_{j,q^\star}^R(\omega_i) = +1$, and $\kappa_j^R(\omega_i) = q^\star$.*

- *If* $z_j(\tau^\star) < 0$, *then* $\zeta^R_{j,q^\star}(\omega_i) = \sigma^R_{j,q^\star}(\omega_i) = -1$, *and* $\kappa^R_j(\omega_i) = q^\star$.

Proof. Consider any $t^\star \in (\omega_i, \omega^\star)$. By construction of $\omega^\star$, there are no valley k-crossings in $(\omega_i, \omega^\star)$ for any $k < q^\star$. By construction of $J(q^\star)$, there are no $j^\star$-valley crossings in $(\omega_i, \omega^\star)$ for any $j^\star \in \Lambda_{\mathbf{f}} \backslash J(q^\star)$. Moreover, for each $j \in J(q^\star)$, Corollary 7.4.12 implies that $\sigma^L_{j,k}(t^\star) = \sigma^R_{j,k}(t^\star) = 0$ for each $k < q^\star$, and Lemma 7.4.18 implies that $(\eta^L_j(t^\star), \eta^R_j(t^\star)) = (\sigma^L_{j,q^\star}(t^\star), \sigma^R_{j,q^\star}(t^\star))$. The first claim of the theorem follows immediately.

The remaining claims of the theorem follow immediately from Lemmata 7.4.16 and 7.4.18, and the definitions of $J(q^\star)$, $\boldsymbol{\zeta}^R(\omega_i)$, and $\boldsymbol{\kappa}^R(\omega_i)$. □

7.4.4 Determining valley crossings

This section provides a theoretical foundation for determining the valley crossings ω_i for $i \in \{1, \ldots, \lambda\}$ during integration of the bank-locked ODE (7.11) and the bank-locked sensitivity ODE (7.12). Observe that λ is not known *a priori*; rather, $i = \lambda$ if and only if there are no valley crossings in the set (ω_{i-1}, t_f). By definition, $\omega_0 = t_0$. Thus, as an inductive assumption, suppose in the remainder of this section that Assumption 7.3.1 holds, and, for some fixed $i \in \mathbb{N}$, the quantities ω_{i-1}, $[\mathbf{x}_{\omega_{i-1}}]'(\mathbf{c}_0; \mathbf{M})$, $\boldsymbol{\kappa}^R(\omega_{i-1})$, and $\boldsymbol{\zeta}^R(\omega_{i-1})$ are known, with $\omega_{i-1} < t_f$.

Inspection of Algorithm 7 shows that $\mathbf{h}(i, \cdot, \cdot)$ is defined as a finite composition of functions that are locally Lipschitz continuous on open sets. Thus, there exists an open superset $\bar{T}_i \subset \mathbb{R}$ of $[\omega_{i-1}, \omega_i]$ for which Algorithm 7 describes a well-defined function $\mathbf{h}(i, \cdot, \cdot)$ on $\bar{T}_i \times \mathbb{R}^n$ rather than $[\omega_{i-1}, \omega_i] \times \mathbb{R}^n$. Using this extended domain, standard ODE extension theory implies that the unique solutions of the bank-locked ODEs (7.11) and (7.12) may be extended to yield unique solutions on some open set $T_i \subset \mathbb{R}$ for which $[\omega_{i-1}, \omega_i] \subset T_i \subset \bar{T}_i$. Note, however, that the extended solutions of these ODEs are not expected to correspond to $\mathbf{x}(t, \mathbf{c}_0)$ or its LD-derivatives for any $t \notin [\omega_{i-1}, \omega_i]$.

Inspection of Algorithm 7 shows that $\mathbf{h}(i, \cdot, \cdot)$ is $\mathcal{C}^\omega$ on its domain; it follows immediately that $\boldsymbol{\xi}(\cdot, \mathbf{0})$ is $\mathcal{C}^\omega$ on T_i. Thus, for each $j \in \Lambda_{\mathbf{f}}$, the function $\hat{u}_{(j)}(i, \cdot, \cdot)$

defined by Algorithm 7 is $\mathcal{C}^\omega$ as well. For any $t \in T_i$ and $j \in \Lambda_{\mathbf{f}}$, denote the quantity $\hat{u}_{(j)}(i, t, \boldsymbol{\xi}(t, \mathbf{0}))$ in Algorithm 7 as "$\hat{u}_{(j),0}(t)$". For each $k \in \{1, \ldots, p\}$, denote the scalar quantity

$$\frac{\partial \hat{u}_{(j)}}{\partial \mathbf{z}}(i, t, \boldsymbol{\xi}(t, \mathbf{0})) \; \frac{\partial \boldsymbol{\xi}}{\partial \mathbf{p}}(t, \mathbf{0}) \; \mathbf{e}_{(k)}$$

as "$\hat{u}_{(j),k}(t)$"; as discussed in the proof of Theorem 7.4.14, for each $k > 0$, $\hat{u}_{(j),k}(t)$ is also the kth column of the row vector

$$\hat{\mathbf{U}}_{(j)} \left(\boldsymbol{\zeta}^R(\omega_{i-1}), t, \tfrac{\partial \boldsymbol{\xi}}{\partial \mathbf{p}}(t, \mathbf{0}) \right)$$

described by Algorithm 6. Theorem 7.4.14, Lemmata 7.4.8 and 7.4.13, and Corollary 7.4.12 thus imply that

$$\hat{u}_{(j),0}(t) = u_{(j)}(t) \quad \text{and} \quad \hat{u}_{(j),k}(t) = \dot{u}_{(j),k}(t), \qquad \forall t \in [\omega_{i-1}, \omega_i] \backslash Z, \quad \forall k \in \{1, \ldots, p\}, \quad \forall j \in \Lambda_{\mathbf{f}}; \tag{7.18}$$

the continuity of $\hat{u}_{(j),0}$ and $u_{(j)}$ then yields

$$\hat{u}_{(j),0}(t) = u_{(j)}(t), \qquad \forall t \in [\omega_{i-1}, \omega_i], \quad \forall j \in \Lambda_{\mathbf{f}}. \tag{7.19}$$

Moreover, the above discussion implies that each $\hat{u}_{(j),k}$ is $\mathcal{C}^\omega$, and thus L/R-analytic, on T_i. The following two lemmata then follow from Lemma 7.2.1 and [66, Corollary 1.2.6], respectively.

Lemma 7.4.20. *For each $t^* \in T_i$, $j \in \Lambda_{\mathbf{f}}$, and $k \in \{0, 1, \ldots, p\}$, there exist quantities $s^L_{j,k}(t^*), s^R_{j,k}(t^*) \in \{-1, 0, +1\}$ so that, for some sufficiently small $\delta > 0$,*

$$s^L_{j,k}(t^*) = \operatorname{sign} \hat{u}_{(j),k}(t), \qquad \forall t \in [t^* - \delta, t^*),$$

and

$$s^R_{j,k}(t^*) = \operatorname{sign} \hat{u}_{(j),k}(t), \qquad \forall t \in (t^*, t^* + \delta].$$

Lemma 7.4.21. *For each $t^* \in T_i$, $j \in \Lambda_{\mathbf{f}}$, and $k \in \{0, 1, \ldots, p\}$,*

$$(s^L_{j,k}(t^*), s^R_{j,k}(t^*)) \notin \{(-1,0),(+1,0),(0,-1),(0,+1)\}.$$

The constructions in Lemma 7.4.20 motivate the following definition, which formalizes an analog of valley crossings for the bank-locked ODE and sensitivity ODE.

Definition 7.4.22. *For each $t^* \in T_i$, $j \in \Lambda_{\mathbf{f}}$, and $k \in \{0,1,\ldots,p\}$, define $s^L_{j,k}(t^*) \in \{-1,0,+1\}$ and $s^R_{j,k}(t^*) \in \{-1,0,+1\}$ as in Lemma 7.4.20. t^* is a* (j-)bank (k-)jump *if both $k = \kappa^R_j(\omega_{i-1})$ and*

$$(s^L_{j,k}(t^*), s^R_{j,k}(t^*)) \in \{(-1,+1),(+1,-1)\}.$$

Bank jumps can be found in T_i during numerical integration of the bank-locked ODE and sensitivity ODE, using standard event detection techniques [89]. Moreover, as the remaining results in this section demonstrate, certain bank jumps correspond to valley crossings in the original ODE (7.2) and sensitivity ODE (7.3), which permits determination of ω_i.

Lemma 7.4.23. *There are no bank jumps in (ω_{i-1}, ω_i).*

Proof. Choose any $t^* \in (\omega_{i-1}, \omega_i)$; it suffices to show that t^* is not a bank jump. Since the set Z described by Lemma 7.4.8 is finite, there exists some sufficiently small $\delta > 0$ for which $Z \cap ([t^* - \delta, t^*) \cup (t^*, t^* + \delta]) = \emptyset$. Thus, Lemma 7.4.20 and (7.18) imply that, for each $j \in \Lambda_{\mathbf{f}}$ and $k \in \{0,1,\ldots,p\}$, for sufficiently small $\delta > 0$, $s^L_{j,k}(t^*) = \sigma^L_{j,k}(t^*)$ and $s^R_{j,k}(t^*) = \sigma^R_{j,k}(t^*)$. Since there are no valley crossings in (ω_{i-1}, ω_i) by construction, the required result follows. □

Lemma 7.4.24. *Suppose that $\omega_i < t_f$, and let q^* be the least value of $k \in \{0,1,\ldots,p\}$ for which ω_i is a valley k-crossing. Then, ω_i is a bank q^*-jump, but is not a bank k-jump for any $k < q^*$.*

Proof. The second required result will be demonstrated first. Since the case in which $q^* = 0$ is trivial, suppose instead that $q^* \geq 1$; for an arbitrary choice of

$k^* \in \{0, 1, \ldots, q^* - 1\}$, it will be shown that ω_i is not a bank k^*-jump. Corollary 7.4.12 implies that $\sigma^R_{j,k}(\omega_i) = \sigma^R_{j,k}(\omega_{i-1})$ for each $j \in \Lambda_{\mathbf{f}}$ and $k \leq k^*$. For any particular $j \in \Lambda_{\mathbf{f}}$, Corollary 7.4.12 implies that if $\kappa^R_j(\omega_{i-1}) \leq k^*$, then $\zeta^R_j(\omega_i) = \zeta^R_j(\omega_{i-1})$. Corollary 7.4.11 implies that if $\kappa^R_j(\omega_{i-1}) > k^*$, then $\kappa^R_j(\omega_i) > k^*$ as well, in which case Lemma 7.4.8 implies that $\sigma_{j,k}(t) = \sigma^R_{j,k}(\omega_{i-1}) = 0$ for each $k \leq k^*$ and $t \in [\omega_{i-1}, \omega_{i+1}] \backslash Z$. Combining the above cases, it follows that

$$\zeta^R_j(\omega_i)\, u_{(j)}(t) = \zeta^R_j(\omega_{i-1})\, u_{(j)}(t), \qquad \forall t \in [\omega_{i-1}, \omega_{i+1}] \backslash Z, \quad \forall j \in \Lambda_{\mathbf{f}},$$

and

$$\zeta^R_j(\omega_i)\, \dot{u}_{(j),k}(t) = \zeta^R_j(\omega_{i-1})\, \dot{u}_{(j),k}(t), \qquad \forall t \in [\omega_{i-1}, \omega_{i+1}] \backslash Z, \quad \forall j \in \Lambda_{\mathbf{f}}, \quad \forall k \in \{1, \ldots, k^*\}.$$

With $\tilde{\mathbf{H}}$ defined as in Lemma 7.4.13, and with $\mathbf{h}$ defined as in Algorithm 7 (permitting, for the moment, the t argument of $\mathbf{h}$ to take values outside $[\omega_{i-1}, \omega_i]$), it follows that $\mathbf{h}(i, t, \mathbf{x}(t, \mathbf{c}_0))$ is well-defined at each $t \in [\omega_i, \omega_{i+1}] \backslash Z$, with

$$\mathbf{h}(i, t, \mathbf{x}(t, \mathbf{c}_0)) = \mathbf{h}(i+1, t, \mathbf{x}(t, \mathbf{c}_0)), \qquad \forall t \in [\omega_{i-1}, \omega_{i+1}] \backslash Z,$$

and

$$\tilde{\mathbf{H}}(t, [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M}_{(1:k^*)})) = \hat{\mathbf{H}}(\boldsymbol{\zeta}^R(\omega_{i-1}), t, [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M}_{(1:k^*)})), \qquad \forall t \in [\omega_{i-1}, \omega_{i+1}] \backslash Z.$$

So, by inspection, the mapping $t \mapsto \mathbf{x}(t, \mathbf{c}_0)$ solves the bank-locked ODE (7.11) on $[\omega_{i-1}, \omega_{i+1}]$, rather than just on $[\omega_{i-1}, \omega_i]$. Similarly, the mapping

$$t \mapsto [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M}_{(1:k^*)}) = [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})\, \mathbf{I}_{(1:k^*)}$$

solves the leftmost k^* columns of the bank-locked sensitivity ODE (7.12) on the interval $[\omega_{i-1}, \omega_{i+1}]$, rather than just on $[\omega_{i-1}, \omega_i]$. Thus, the proof of Theorem 7.4.14, Lemmata 7.4.8 and 7.4.13, and Corollary 7.4.12 imply that (7.18) remains applicable for each $t \in ([\omega_{i-1}, \omega_{i+1}] \cap T_i) \backslash Z$ and $k \leq k^*$, which implies that $s^L_{j,k}(\omega_i) = \sigma^L_{j,k}(\omega_i)$

and $s^R_{j,k}(\omega_i) = \sigma^R_{j,k}(\omega_i)$ for each $j \in \Lambda_{\mathbf{f}}$ and $k \leq k^*$.

Now suppose, to obtain a contradiction, that ω_i is a j-bank k^*-jump for some $j \in \Lambda_{\mathbf{f}}$. Thus, $k^* = \kappa^R_j(\omega_{i-1})$. Corollary 7.4.12 then implies that $\sigma^L_{j,k^*}(\omega_i) \in \{-1, +1\}$. Since ω_i is not a valley k^*-crossing, it follows that $(\sigma^L_{j,k^*}(\omega_i), \sigma^R_{j,k^*}(\omega_i) \in \{(-1,-1), (-1,0), (+1,+1), (+1,0)\}$. The final result in the previous paragraph then shows that $(s^L_{j,k^*}(\omega_i), s^R_{j,k^*}(\omega_i) \in \{(-1,-1), (-1,0), (+1,+1), (+1,0)\}$. This shows that ω_i is also not a bank k^*-jump, which is the required contradiction.

Next, to obtain a contradiction, suppose that ω_i is not a bank q^*-jump. The cases in which $q^* = 0$ and $q^* > 0$ will be considered separately. First, suppose that $q^* = 0$. Corollary 7.4.12 and (7.19) imply that $s^L_{j,0}(\omega_i) = \sigma^L_{j,0}(\omega_i) = \sigma^R_{j,0}(\omega_{i-1})$ for each $j \in \Lambda_{\mathbf{f}}$. Thus, if $s^L_{j,0}(\omega_i) \in \{-1, +1\}$ for some $j \in \Lambda_{\mathbf{f}}$, then $\kappa^R_j(\omega_{i-1}) = 0$. Using this result, since ω_i is not a bank 0-jump, Lemma 7.4.21 implies that $s^R_{j,0}(\omega_i) = s^L_{j,0}(\omega_i) = \sigma^R_{j,0}(\omega_{i-1}) \in \{-1, 0, +1\}$ for each $j \in \Lambda_{\mathbf{f}}$. The continuity of $\hat{u}_{(j),0}$ at ω_i then implies that, for sufficiently small $\delta > 0$,

$$|\hat{u}_{(j),0}(t)| = \sigma^R_{j,0}(\omega_{i-1})\, \hat{u}_{(j),0}(t) = \zeta^R_j(\omega_{i-1})\, \hat{u}_{(j),0}(t), \qquad \forall t \in (\omega_i - \delta, \omega_i + \delta), \quad \forall j \in \Lambda_{\mathbf{f}};$$

the rightmost equation above follows from (7.19), considering the cases in which $\sigma^R_{j,0}(\omega_{i-1})$ is either zero or nonzero separately. Inspecting the factored representation of $\mathbf{f}$, the above result, combined with (7.19), shows that the mapping $t \mapsto \boldsymbol{\xi}(t, \mathbf{0})$ solves the following ODE on $(\omega_i - \delta, \omega_i + \delta)$ uniquely:

$$\frac{d\mathbf{z}}{dt}(t) = \mathbf{f}(t, \mathbf{z}(t)), \qquad \mathbf{z}(\omega_i) = \boldsymbol{\xi}(\omega_i, \mathbf{0}) = \mathbf{x}(\omega_i, \mathbf{c}_0).$$

Thus, $\boldsymbol{\xi}(\cdot, \mathbf{0}) \equiv \mathbf{x}(\cdot, \mathbf{c}_0)$ on $(\omega_i - \delta, \omega_i + \delta)$, and so $s^L_{j,0}(\omega_i) = \sigma^L_{j,0}(\omega_i)$ and $s^R_{j,0}(\omega_i) = \sigma^R_{j,0}(\omega_i)$ for each $j \in \Lambda_{\mathbf{f}}$. Since there exists $j^* \in \Lambda_{\mathbf{f}}$ for which ω_i is a j^*-valley 0-crossing, it follows immediately that ω_i is a j^*-bank 0-jump, which is the required contradiction.

Next, suppose that $q^* > 0$. Using the final claim of the lemma, which was proven above, it follows that ω_i is not a bank k-jump for any $k \leq q^*$. Since ω_i

is not a bank 0-jump, a similar argument to the "$q^* = 0$" case shows that there exists a neighborhood $\hat{N} \subset T_i$ of ω_i on which $\boldsymbol{\xi}(\cdot, \mathbf{0}) \equiv \mathbf{x}(\cdot, \mathbf{c}_0)$, and so $s^L_{j,0}(\omega_i) = \sigma^L_{j,0}(\omega_i)$ and $s^R_{j,0}(\omega_i) = \sigma^R_{j,0}(\omega_i)$ for each $j \in \Lambda_{\mathbf{f}}$. Since ω_i is not a valley 0-crossing, Lemma 7.4.10 then implies that $s^L_{j,0}(\omega_i) = s^R_{j,0}(\omega_i)$ for each $j \in \Lambda_{\mathbf{f}}$.

Since ω_i is not a bank k-jump for any $k \leq q^*$, Corollary 7.4.12, Lemma 7.4.21, (7.18), and (7.19) imply that, for each $j \in \Lambda_{\mathbf{f}}$ for which $\bar{\kappa}_j := \kappa^R_j(\omega_{i-1}) \leq q^*$,

$$\zeta^R_j(\omega_{i-1}) = \sigma^R_{j,\bar{\kappa}_j}(\omega_{i-1}) = \sigma^L_{j,\bar{\kappa}_j}(\omega_i) = s^L_{j,\bar{\kappa}_j}(\omega_i) = s^R_{j,\bar{\kappa}_j}(\omega_i).$$

The rightmost equation above follows by contradiction; if, instead, $s^L_{j,\bar{\kappa}_j}(\omega_i) \neq s^R_{j,\bar{\kappa}_j}(\omega_i)$, then Lemma 7.4.21 implies that $(s^L_{j,\bar{\kappa}_j}(\omega_i), s^R_{j,\bar{\kappa}_j}(\omega_i)) \in \{(-1,+1),(+1,-1)\}$, in which case ω_i is a bank $\bar{\kappa}_j$-jump, which is false by assumption.

Moreover, for each $j \in \Lambda_{\mathbf{f}}$ and $k < \kappa^R_j(\omega_{i-1})$, (7.18) and (7.19) imply that

$$s^R_{j,k}(\omega_{i-1}) = \sigma^R_{j,k}(\omega_{i-1}) = 0;$$

the analyticity of $\hat{u}_{(j),k}$ on $T_{(i)}$ then implies that

$$\hat{u}_{(j),k}(t) = 0, \qquad \forall t \in T_i, \quad \forall j \in \Lambda_{\mathbf{f}}, \quad \forall k < \kappa^R_j(\omega_{i-1}).$$

Combining the above results and Lemma 7.2.1, and choosing $\hat{N}$ to be a smaller neighborhood of ω_i if necessary, there exists a finite set $\hat{Z} \subset \hat{N}$ for which:

$$\begin{aligned}\operatorname{fsign}\left(\hat{u}_{(j),0}(t), \ldots, \hat{u}_{(j),q^*}(t)\right) \hat{u}_{(j),k}(t) &= \zeta^R_j(\omega_{i-1})\, \hat{u}_{(j),k}(t), \\ &\qquad \forall t \in \hat{N} \backslash \hat{Z}, \quad \forall j \in \Lambda_{\mathbf{f}}, \quad \forall k \in \{1, \ldots, q^*\}.\end{aligned}$$

Comparing Algorithms 4 and 6 and combining the above results, it follows that, for each $t \in \hat{N} \backslash \hat{Z}$,

$$\hat{\mathbf{H}}\left(\boldsymbol{\zeta}^R(\omega_{i-1}), t, \tfrac{\partial \boldsymbol{\xi}}{\partial \mathbf{p}}(t, \mathbf{0})\, \mathbf{I}_{(1:q^*)}\right) = [\mathbf{f}_t]'(\mathbf{x}(t, \mathbf{c}_0); \tfrac{\partial \boldsymbol{\xi}}{\partial \mathbf{p}}(t, \mathbf{0})\, \mathbf{I}_{(1:q^*)}).$$

Moreover, it follows directly from the definition of the LD-derivative that

$$[\mathbf{x}_{\omega_i}]'(\mathbf{c}_0; \mathbf{M})\, \mathbf{I}_{(1:q^*)} = [\mathbf{x}_{\omega_i}]'(\mathbf{c}_0; \mathbf{M}_{(1:q^*)}).$$

Thus, the mapping $t \mapsto \frac{\partial \boldsymbol{\xi}}{\partial \mathbf{p}}(t, \mathbf{0})\, \mathbf{I}_{(1:q^*)}$ is the unique solution on $\hat{N}$ of the ODE:

$$\frac{d\mathbf{A}}{dt}(t) = [\mathbf{f}_t]'(\mathbf{x}(t, \mathbf{c}_0); \mathbf{A}(t)), \qquad \mathbf{A}(\omega_i) = \frac{\partial \boldsymbol{\xi}}{\partial \mathbf{p}}(\omega_i, \mathbf{0})\, \mathbf{I}_{(1:q^*)} = [\mathbf{x}_{\omega_i}]'(\mathbf{c}_0; \mathbf{M}_{(1:q^*)}).$$

Since the mapping $t \mapsto [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M}_{(1:q^*)})$ also solves this ODE, it follows that the mappings $t \mapsto \frac{\partial \boldsymbol{\xi}}{\partial \mathbf{p}}(t, \mathbf{0})\, \mathbf{I}_{(1:q^*)}$ and $t \mapsto [\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M}_{(1:q^*)})$ are identical on $\hat{N}$, and so $s^L_{j,k}(\omega_i) = \sigma^L_{j,k}(\omega_i)$ and $s^R_{j,k}(\omega_i) = \sigma^R_{j,k}(\omega_i)$ for each $j \in \Lambda_{\mathbf{f}}$ and $k \leq q^*$. Since there exists $j^* \in \Lambda_{\mathbf{f}}$ for which ω_i is a j^*-valley q^*-crossing, it follows immediately that ω_i is a j^*-bank q^*-jump, which is the required contradiction. □

The previous two lemmata yield the following theorem, which can be used to determine ω_i during integration of the bank-locked ODE and sensitivity ODE. Recall that the bank-locked ODE and sensitivity ODE have well-defined unique solutions on an open superset T_i of $[\omega_{i-1}, \omega_i]$, even though ω_i is not known *a priori*.

Theorem 7.4.25. *If there is no bank jump in the set (ω_{i-1}, t_f), then $\omega_i = t_f$. Otherwise, ω_i is the least bank jump in the set (ω_{i-1}, t_f). In the latter case, if k^* denotes the least value of $k \in \{0, 1, \ldots, p\}$ for which ω_i is a bank k-jump, then ω_i is a valley k^*-crossing, but is not a valley k-crossing for any $k < k^*$.*

Proof. If there is no bank jump in the set (ω_{i-1}, t_f), then the contrapositive of Lemma 7.4.24 implies that $\omega_i \geq t_f$. So, $\omega_i = t_f$, as required.

In the remainder of this proof, suppose that there exists a bank jump in the set (ω_{i-1}, t_f). In this case, Lemma 7.2.1 implies that there are finitely many bank jumps in the set $[\omega_{i-1}, t_f]$, and so there exists a least bank jump τ in (ω_{i-1}, t_f). Lemma 7.4.23 implies that $\omega_i \leq \tau < t_f$, and so Lemma 7.4.24 implies that ω_i is a bank jump. Thus, the definition of τ and the inequality $\omega_i \leq \tau$ together show that $\omega_i = \tau$, as required.

Choose $q^* \in \{0, 1, \ldots, p\}$ as in the statement of Lemma 7.4.24; Lemma 7.4.24 implies that ω_i is a bank q^*-jump, but is not a bank k-jump for any $k < q^*$. Thus, $q^* = k^*$; the definition of q^* then shows that ω_i is a valley k^*-crossing, but is not a valley k-crossing for any $k < q^* = k^*$. □

Remark 7.4.26. *Since the bank-locked sensitivity ODE is linear, its columns are uncoupled. Thus, for any $k \in \{1, \ldots, p\}$, locating a bank k-jump requires integrating only the bank-locked ODE and the kth column of the bank-locked sensitivity ODE.*

7.5 Numerical method

Algorithm 9 is a method that uses the main results of Section 7.4 to evaluate the LD-derivative $[\mathbf{x}_{t_f}]'(\mathbf{c}_0; \mathbf{M})$, assuming exact arithmetic and exact integration of ODEs. Algorithm 10 is a subroutine that is called in each iteration of the main loop of Algorithm 9. When this overall method is implemented in practice, numerical error will be introduced inevitably; in response to this error, further assumptions and modifications will be made below. This section describes Algorithm 9 and its implementation, and considers its advantages over alternative methods.

7.5.1 Method outline

This section presents a brief outline of Algorithm 9; a more detailed account is presented in the following section. Broadly, Algorithm 9 proceeds by first initializing $\omega_0 := t_0$, $\mathbf{x}(t_0, \mathbf{c}_0) := \mathbf{c}_0$, $[\mathbf{x}_{t_0}]'(\mathbf{c}_0; \mathbf{M}) := \mathbf{M}$, and a counter $i := 0$. The algorithm then uses the results of Section 7.4 to integrate a tractable reformulation of the sensitivity ODE (7.3), alternating between a *probe phase* and a *bank-locked integration phase*, until $\omega_\lambda = t_f$ is reached, and the required LD-derivative $[\mathbf{x}_{t_f}]'(\mathbf{c}_0; \mathbf{M})$ has been evaluated.

In the probe phase, the results of Section 7.4.3 are used to determine the quantities $\boldsymbol{\zeta}^R(\omega_i)$ and $\boldsymbol{\kappa}^R(\omega_i)$. After the probe phase, i is incremented by 1, and the bank-locked integration phase begins; in this phase, the results of Sections 7.4.2 and 7.4.4 are used to compute ω_i and the quantities $\mathbf{x}(\omega_i, \mathbf{c}_0)$ and $[\mathbf{x}_{\omega_i}]'(\mathbf{c}_0; \mathbf{M})$. If $\omega_i < t_f$, then the procedure returns to the beginning of the probe phase and continues.

Algorithm 9 Computes $[\mathbf{x}_{t_f}]'(\mathbf{c}_0;\mathbf{M})$, with $\mathbf{x}$ described by Assumption 7.3.1

Require: $p \in \mathbb{N}$, $\mathbf{M} \in \mathbb{R}^{n\times p}$
  $i \leftarrow 0$
  $\omega_0 \leftarrow t_0$
  $\mathbf{x}(t_0, \mathbf{c}_0) \leftarrow \mathbf{c}_0$
  $[\mathbf{x}_{t_0}]'(\mathbf{c}_0;\mathbf{M}) \leftarrow \mathbf{M}$
  $\bar{\boldsymbol{\zeta}} \leftarrow \mathbf{0} \in \{-1, 0, +1\}^{\ell}$
  $\bar{\boldsymbol{\kappa}} \leftarrow (p+1)\mathbf{1} \in \{0, 1, \ldots, p+1\}^{\ell}$
  $q^\star \leftarrow 0$
  $J^\star \leftarrow \Lambda_{\mathbf{f}}$
  loop
    Use Algorithm 10 to compute $\bar{\boldsymbol{\zeta}} \leftarrow \boldsymbol{\zeta}^R(\omega_i)$ and $\bar{\boldsymbol{\kappa}} \leftarrow \boldsymbol{\kappa}^R(\omega_i)$.
    $i \leftarrow i+1$
    $\mathbf{p} \leftarrow \mathbf{0} \in \mathbb{R}^p$
    Begin integrating the ODEs (7.11) and (7.12) simultaneously on $[\omega_{i-1}, t_f]$, detecting bank jumps.
    if a bank jump is detected in (ω_{i-1}, t_f) **then**
      Terminate integration at the least bank jump ω_i in (ω_{i-1}, t_f).
      $\mathbf{x}(\omega_i, \mathbf{c}_0) \leftarrow \boldsymbol{\xi}(\omega_i, \mathbf{0})$
      $[\mathbf{x}_{\omega_i}]'(\mathbf{c}_0;\mathbf{M}) \leftarrow \mathbf{A}(\omega_i)$
      Set $q^\star$ to be the least element k of $\{0, 1, \ldots, p\}$ for which ω_i is a bank k-jump.
    else
      return $[\mathbf{x}_{t_f}]'(\mathbf{c}_0;\mathbf{M}) = \mathbf{A}(t_f)$
    end if
    if $|\{j \in \Lambda_{\mathbf{f}} : \bar{\kappa}_j = q^\star\}| = 1$ **then**
      Find $j^\star \in \Lambda_{\mathbf{f}}$ for which $\bar{\kappa}_{j^\star} = q^\star$.
      $\bar{\zeta}_{j^\star} \leftarrow -\bar{\zeta}_{j^\star}$
      $J^\star \leftarrow \{j \in \Lambda_{\mathbf{f}} : \bar{\kappa}_j \geq q^\star\} \backslash \{j^\star\}$
    else
      $J^\star \leftarrow \{j \in \Lambda_{\mathbf{f}} : \bar{\kappa}_j \geq q^\star\}$
    end if
  end loop

Algorithm 10 Probe phase of Algorithm 9, which computes $\bar{\boldsymbol{\zeta}} \leftarrow \boldsymbol{\zeta}^R(\omega_i)$ and $\bar{\boldsymbol{\kappa}} \leftarrow \boldsymbol{\kappa}^R(\omega_i)$.

Require: $p, t_f, \mathbf{x}(\omega_i, \mathbf{c}_0), [\mathbf{x}_{\omega_i}]'(\mathbf{c}_0; \mathbf{M}), \bar{\boldsymbol{\zeta}} \in \{-1, 0, +1\}^\ell, \bar{\boldsymbol{\kappa}} \in \{0, 1, \ldots, p+1\}^\ell, q^\star \in \{0, 1, \ldots, p\}, J^\star \subset \Lambda_{\mathbf{f}}$ and $\omega_i \in [t_0, t_f)$ from Algorithm 9, and $\delta > 0$

```
if J* = ∅ then
    return
end if
t* ← min {ω_i + δ, t_f}
for all j ∈ J* do
    ζ̄_j ← 0, κ̄_j ← p + 1
end for
if q* = 0 then
    for all j ∈ J* do
        if a sufficient condition in Corollary 7.4.17 for κ_j^R(ω_i) to be 0 is met then
            Compute ζ̄_j ← σ_{j,0}^R(ω_i) ∈ {−1, +1} using Corollary 7.4.17.
            κ̄_j ← 0, J* ← J*\{j}
        end if
    end for
    if J* = ∅ then
        return
    end if
    Begin integrating the ODEs (7.2) and (7.14) for each j ∈ J* on [ω_i, t*], detecting valley 0-crossings.
    if a valley 0-crossing is detected in (ω_i, t*] then
        Terminate integration at the least valley 0-crossing τ* in (ω_i, t*]
        t* ← τ*
    end if
    for all j ∈ J* do
        if w_j(t*) ≠ 0 then
            ζ̄_j ← sign w_j(t*), κ̄_j ← 0, J* ← J*\{j}
        end if
    end for
    q* ← 1
else if q* ≥ 1 then
    p ← 0 ∈ ℝ^p
    Begin integrating the ODE (7.11) and the leftmost (q* − 1) columns of the ODE (7.12) simultaneously on [ω_i, t*],
    detecting bank k-jumps for all k < q*.
    if a bank k-jump is detected in (ω_i, t*] then
        Terminate integration at the least bank jump τ* detected in (ω_i, t*].
        t* ← τ*
    end if
end if
k* ← q*
while k* ≤ p and J* ≠ ∅ do
    Begin integrating the ODEs (7.2), (7.15), and (7.16) simultaneously on [ω_i, t*], detecting valley k*-crossings.
    if a valley k*-crossing is detected in (ω_i, t*) then
        Terminate integration at the least valley k*-crossing τ* in (ω_i, t*)
        t* ← τ*
    end if
    for all j ∈ J* do
        if z_j(t*) ≠ 0 then
            ζ̄_j ← sign z_j(t*), κ̄_j ← k*, J* ← J*\{j}
        end if
    end for
    k* ← k* + 1
end while
return
```

7.5.2 Method summary

A more detailed description of Algorithm 9 is presented in this section. Algorithm 9 sets a counter $i \leftarrow 0$, corresponding to the current index of ω_i. Integration of the ODEs (7.2) and (7.3) is initialized at $\omega_0 := t_0$. Vectors $\bar{\boldsymbol{\zeta}}$ and $\bar{\boldsymbol{\kappa}}$ are initialized; these vectors will hold the known values of $\zeta_j^R(\omega_i)$ and $\kappa_j^R(\omega_i)$ throughout the algorithm, and will otherwise default to 0 or $p+1$, respectively. A set $J^\star$ is also initialized to $\Lambda_{\mathbf{f}}$, and will play the role of $J(q^\star)$ in Section 7.4.3. A quantity $q^\star \in \{0, 1, \ldots, p\}$ is initialized to 0, and will play the role of its namesakes in Theorem 7.4.25 and Section 7.4.3. The following steps are then repeated until $\omega_\lambda := t_f$ is reached, at which point $[\mathbf{x}_{t_f}]'(\mathbf{c}_0; \mathbf{M})$ is returned.

Firstly, a *probe phase* described by Algorithm 10 is carried out, to determine and store the quantities $\boldsymbol{\zeta}^R(\omega_i)$ and $\boldsymbol{\kappa}^R(\omega_i)$ in the vectors $\bar{\boldsymbol{\zeta}}$ and $\bar{\boldsymbol{\kappa}}$, respectively. This phase takes an approach motivated by the results of Section 7.4.3. The ODEs described in this section are integrated on a duration $[\omega_i, \omega_i + \delta]$, where the parameter $\delta > 0$ is chosen to be small relative to $(t_f - t_0)$, but large enough that any appreciable deviation of an ODE state variable from its initial condition can be recognized. Ideally, δ should be short enough that there are no bank jumps or valley crossings of the corresponding ODE solutions on $(\omega_i, \omega_i + \delta]$; if any bank jumps or valley crossings are detected in this duration, then the algorithm effectively shortens δ via a proxy quantity $t^\star$.

In the probe phase, if $q^\star = 0$, then Lemma 7.4.16 and Corollary 7.4.17 are used to determine $\sigma_{j,0}^R(\omega_i)$ for each $j \in \Lambda_{\mathbf{f}}$. If $q^\star \geq 1$, then Corollary 7.4.12 implies that, for each $j \in \Lambda_{\mathbf{f}}$ for which $\kappa_j^R(\omega_{i-1}) < q^\star$, $\zeta_j^R(\omega_i) = \zeta_j^R(\omega_{i-1})$ and $\kappa_j^R(\omega_i) = \kappa_j^R(\omega_{i-1})$, so there is no need to update these quantities from their previously stored values. The bank-locked ODE and the leftmost $(q^\star - 1)$ columns of the bank-locked sensitivity ODE are used to check that there are no bank k-jumps on $(\omega_i, \omega_i + \delta)$ for any $k < q^\star$. Observe that, although $\zeta_j^R(\omega_i)$ is unknown at this point for each $j \in \Lambda_{\mathbf{f}}$ for which $\kappa_j^R(\omega_i) \geq q^\star$, these values are not necessary to carry out the required bank-locked integration, since the corresponding functions $\hat{u}_{(j),k}$ are iden-

tically zero. Now Corollary 7.4.12 also implies that, for each $j \in \Lambda_{\mathbf{f}}$ for which $\kappa_j^R(\omega_{i-1}) \geq q^\star$, the inequality $\kappa_j^R(\omega_i) \geq q^\star$ also holds. Thus, the strong inductive approach described in Section 7.4.3 is applied; Theorem 7.4.19 is used to compute $\sigma_{j,k}^R(\omega_i)$ values for $k := q^\star, q^\star + 1, \ldots, p$, terminating when $\boldsymbol{\zeta}^R(\omega_i)$ and $\boldsymbol{\kappa}^R(\omega_i)$ have been determined. If the jth components of these vectors have not been determined after the $k := p$ iteration, then it follows that $\sigma_{j,k}^R(\omega_i) = 0$ for each $k \in \{0, 1, \ldots, p\}$. In this case, $\zeta_j^R(\omega_i) = 0$ and $\kappa_j^R(\omega_i) = p + 1$, and so $\bar{\zeta}_j$ and $\bar{\kappa}_j$ need not be changed from their default values.

Once the probe phase is complete, the counter i is incremented by one, and a *bank-locked integration phase* begins. The results of Section 7.4.4 are then applied to solve the ODEs (7.2) and (7.3) on $[\omega_{i-1}, \omega_i]$, determining the unknown quantity ω_i in the process. To accomplish this, Theorem 7.4.25 is applied during simultaneous integration of the bank-locked ODE and sensitivity ODE, noting that the latter can be solved efficiently using established techniques for sensitivity analysis of smooth dynamic systems. If $\omega_i = t_f$, then the required LD-derivative can be returned; if not, then Theorem 7.4.25 is used to identify the least value of $q^\star$ for which ω_i is a valley $q^\star$-crossing. If there is only one $j^\star \in \Lambda_{\mathbf{f}}$ for which $\kappa_{j^\star}^R(\omega_{i-1}) = q^\star$, then this information is sufficient to conclude that $\kappa_{j^\star}^R(\omega_i) = \kappa_{j^\star}^R(\omega_{i-1})$ and $\zeta_{j^\star}^R(\omega_i) = -\zeta_{j^\star}^R(\omega_{i-1})$, since ω_i is certainly a $j^\star$-valley $q^\star$-crossing in this case. At this point, the method returns to the probe phase above.

7.5.3 Additional assumptions

The analysis in Section 7.4 concerning the structure of the sensitivity ODE (7.3) does not make any assumptions beyond the basic Assumption 7.3.1. To implement Algorithm 9 in practice, the following additional assumptions and considerations will be enforced to overcome the numerical error introduced by function evaluations and integration methods.

Firstly, to use the event detection algorithm of [89] to detect bank jumps and valley crossings, appropriate transversality conditions are assumed to hold at each

event; this assumption can be relaxed if a more rigorous yet more computationally expensive root-finding approach is used [47]. If these transversality conditions are not met, then event detection could yield false positives or false negatives; in Algorithm 10, this could lead to incorrect evaluations of $\boldsymbol{\zeta}^R(\omega_i)$ and $\boldsymbol{\kappa}^R(\omega_i)$, while in the main Algorithm 9, this could lead to valley crossings not being detected, or to spurious valley crossings being identified. If any of these errors are encountered, then the LD-derivatives be returned by the method would likely be incorrect.

Secondly, when using the integrals in Lemma 7.4.16 to determine $\sigma^R_{j,0}(\omega)$, or when using Theorem 7.4.19 to determine $\sigma^R_{j,q^\star}(\omega^\star)$, numerical error involved in computing the associated quadrature variables prevents checking that the integrals involved are exactly zero. Thus, if any such integral lies in a set $[-\epsilon, \epsilon]$ for a small tolerance $\epsilon > 0$, then this integral will be assumed to be zero for the purposes of Algorithm 10. If the integral is not, in fact, exactly zero, then the returned LD-derivatives would likely be incorrect. The chosen parameter ϵ should increase with both the parameter δ, and an estimate of the Lipschitz constant of the right-hand side of the bank-locked sensitivity ODE (7.12).

As a similar issue, observe that the sufficient conditions for $\kappa^R_j(\omega_i)$ to be zero provided by Corollary 7.4.17 are valid regardless of the value of $\epsilon_\sigma \geq 0$. Thus, ϵ_σ should be chosen to be a small positive tolerance: the smaller ϵ_σ is, the more likely the sufficient conditions of Corollary 7.4.17 are to hold. Nevertheless, ϵ_σ should be large enough that, if $u_{(j)}(\omega_i, \mathbf{x}(\omega_i, \mathbf{c}_0))$ or d_j would both be exactly zero in the absence of numerical error, then the numerical error introduced by computing these quantities would not lead to a sufficient condition in Corollary 7.4.17 from being erroneously satisfied.

Since there are a finite number of valley crossings in $[t_0, t_f]$, there exists $\hat{\delta} > 0$ for which, for each $i \in \{0, 1, \ldots, \lambda - 1\}$, the set $(\omega_i, \omega_i + \hat{\delta}]$ does not contain any valley crossings. If the parameter $\delta > 0$ in Algorithm 10 is assumed to be such a $\hat{\delta}$, then there is no need for detection of bank jumps or valley crossings in Algorithm 10. With these steps omitted, Algorithm 10 becomes the simpler Algorithm 11. However, as discussed earlier, $\delta > 0$ should also be chosen to be

large enough that any considered ODE solution that is not constant on $[\omega_i, \omega_i + \delta]$ can be reasonably expected to differ significantly from its value at ω_i over the interval $[\omega_i, \omega_i + \delta]$. Despite the discussion in this paragraph, detection of bank jumps in the main Algorithm 9 is a critical step of the method, and is required to determine the quantities ω_i.

Observe that the original ODE (7.2) has an abs-factorable right-hand side function, as do the probe ODEs (7.14)–(7.16). Thus, if an implicit integration method such as a backward differentiation formula (BDF) method is used to solve these ODEs, then each step of the integration method will require solving an equation system with a residual function that may not be differentiable everywhere. If these equation systems cannot be circumvented or solved with equation-solving methods developed for smooth problems, then dedicated equation-solving methods for nondifferentiable problems [91, 92] could be employed. Alternatively, specialized integration methods [33] have also been proposed for such nonsmooth ODEs, and could be employed here. Since the bank-locked ODE and bank-locked sensitivity ODE have smooth right-hand side functions, they are not affected by the issues in this paragraph.

7.5.4 Computational performance

The computational complexity of Algorithm 9 is evidently dominated by the complexity of the various ODE integration steps. The bank-locked integration steps of the main Algorithm 9 ultimately require solving the ODEs (7.11) and (7.12) on $[\omega_{i-1}, \omega_i]$ for each $i \in \{1, \ldots, \lambda\}$. If solved naïvely, the cost of formulating and solving the bank-locked sensitivity ODE (7.12) would be approximately p times the cost of solving the original ODE (7.2). Since (7.12) is a classical sensitivity system, however, it can be solved efficiently using the *staggered corrector* method [24]; the total computational cost of this bank-locked integration is therefore likely to be significantly less than in the naïve case.

The probe phase is carried out λ times in total: once at ω_i for each $i \in \{0, 1, \ldots, \lambda -$

Algorithm 11 Modified probe phase of Algorithm 9, as discussed in Section 7.5.3.

Require: p, t_f, $\mathbf{x}(\omega_i, \mathbf{c}_0)$, $[\mathbf{x}_{\omega_i}]'(\mathbf{c}_0; \mathbf{M})$, $\bar{\boldsymbol{\zeta}} \in \{-1, 0, +1\}^\ell$, $\bar{\boldsymbol{\kappa}} \in \{0, 1, \ldots, p+1\}^\ell$, $q^\star \in \{0, 1, \ldots, p\}$, $J^\star \subset \Lambda_{\mathbf{f}}$ and $\omega_i \in [t_0, t_f)$ from Algorithm 9, and $\delta > 0$

```
if J* = ∅ then
    return
end if
t* ← min {ω_i + δ, t_f}
for all j ∈ J* do
    ζ̄_j ← 0
    κ̄_j ← p + 1
end for
if q* = 0 then
    for all j ∈ J* do
        if a sufficient condition in Corollary 7.4.17 for κ_j^R(ω_i) to be 0 is met then
            Compute ζ̄_j ← σ_{j,0}^R(ω_i) ∈ {−1, +1} using Corollary 7.4.17.
            κ̄_j ← 0
            J* ← J*\{j}
        end if
    end for
    if J* = ∅ then
        return
    end if
    Integrate the ODEs (7.2) and (7.14) for each j ∈ J* on [ω_i, t*].
    for all j ∈ J* do
        if w_j(t*) ≠ 0 then
            ζ̄_j ← sign w_j(t*)
            κ̄_j ← 0
            J* ← J*\{j}
        end if
    end for
    q* ← 1
end if
k* ← q*
while k* ≤ p and J* ≠ ∅ do
    Integrate the ODEs (7.2), (7.15), and (7.16) simultaneously on [ω_i, t*].
    for all j ∈ J* do
        if z_j(t*) ≠ 0 then
            ζ̄_j ← sign z_j(t*)
            κ̄_j ← k*
            J* ← J*\{j}
        end if
    end for
    k* ← k* + 1
end while
return
```

1}. If Algorithm 11 is used in place of Algorithm 10, as discussed in Section 7.5.3, then the cost of each probe phase is dominated by the cost of solving the coupled ODEs (7.2) and (7.14) and the coupled ODEs (7.15) and (7.16) on a subset of $[\omega_i, \omega_i + \delta]$. By inspection of the ODEs involved, the cost of solving the former coupled pair is comparable to the cost of solving the latter coupled pair, so it suffices to consider only the latter in detail. Though we expect that large values of $\kappa_j^R(\omega_i)$ are unlikely, and that $\kappa_j^R(\omega_i)$ will often be 0 or 1 in practice, each probe phase could require p iterations of the `while`–loop in Algorithms 10 or 11 in the worst case. As a particular pathological case in which this worst case is reached, observe that, if the provided factored representation for $\mathbf{f}$ contains an absolute value $v_{(j)} := |u_{(j)}|$ whose argument is zero at each $\mathbf{x}(t, \mathbf{c})$ regardless of the value of $\mathbf{c}$, then $\dot{u}_{(j),k}(t)$ will be zero at each k and t, in which case $\kappa_j^R(t)$ will be $(p+1)$ at each t. Though this case could in principle be eliminated through preprocessing or manual reformulation of the provided representation of $\mathbf{f}$, the former approach would likely not be able to handle pathological formulations in the vein of

$$f : (t, x) \mapsto x^2 + |\sin^2(x) + \cos^2(x) - 1|,$$

without some capacity for symbolic manipulation.

Within the probe phase, evaluating the right-hand sides for the ODE pair (7.15) and (7.16) requires computing $\mathbf{x}(t, \mathbf{c}_0)$; to accomplish this, (7.15) and (7.16) can be solved simultaneously with the original ODE (7.2) on $[\omega_i, t^\star]$. The structural similarity of Algorithm 8 to the factored representation of $\mathbf{f}$ suggests that the staggered corrector method could be used again to reduce the computational cost of solving (7.15) and (7.16). Algorithm 10 has an `else if` block that Algorithm 11 lacks; this block demands solution of a reduced bank-locked sensitivity ODE on $[\omega_i, t^\star]$, and can be accomplished efficiently using the staggered corrector method. The detection of bank jumps and valley crossings in Algorithm 10 requires appending extra algebraic variables to the considered ODEs: one for each absolute-value function in the provided factored representation of $\mathbf{f}$.

7.5.5 Alternative methods

This section briefly describes why certain plausible alternative approaches to Algorithm 9 were not pursued.

Firstly, it is, of course, possible to attempt to integrate the sensitivity ODE (7.3) directly, using a standard numerical integration method. Since the ODE (7.3) has a right-hand side that may be discontinuous with respect to its $\mathbf{A}(t)$ term, however, there is no guarantee that any produced numerical solution will be meaningful [14]; the behavior of the integration method near any discontinuity of the right-hand side would be unpredictable.

Secondly, rather than solving the ODEs (7.2), (7.15), and (7.16) simultaneously in each iteration of the `while`–loop in Algorithm 10, the ODE (7.2) could be solved only once on $[\omega_i, t^*]$ at the start of this loop; the results of this integration could in principle be stored and fed into the right-hand sides of the ODEs (7.15) and (7.16) at each iteration of the `while`–loop. This approach would eliminate the need for redundant integration of (7.2), and would likely improve the computational efficiency of the overall procedure. However, this approach would also require careful storage and extraction of the LU-factors used in the corrector iterations of the integration method when solving (7.2), which would consequently tie the presented method to a particular nonstandard ODE solver.

Thirdly, according to Lemma 7.4.18, the probe ODE (7.15) in $\mathbf{y}^\star$ can be integrated to provide any particular column of the desired LD-derivative $[\mathbf{x}_t]'(\mathbf{c}_0; \mathbf{M})$. Moreover, valley crossings can be detected during integration of the ODE (7.16) according to Theorem 7.4.19. Combining these observations, if each integration of the probe ODEs (7.15) and (7.16) in Algorithm 10 were carried out on $[\omega_i, \omega_{i+1}]$ instead of $[\omega_i, t^*]$, with ω_{i+1} determined during integration of the probe ODEs, then certain columns of the required LD-derivatives could be computed in this manner instead. This method, however, would suffer even more from the computational burden of either integrating the original ODE (7.2) redundantly on $[\omega_i, \omega_{i+1}]$ during each probe integration, or storing the LU-factors corresponding to this integra-

tion of (7.2) on the interval $[\omega_i, \omega_{i+1}]$. Moreover, this method cannot exploit the computational efficiency of solving the bank-locked sensitivity ODE (7.12) to the same extent as the presented approach.

7.6 Implementation and examples

This section describes an implementation of the method described in Section 7.5, and demonstrates its application to example problems for illustration.

7.6.1 Implementation

Algorithm 9 was implemented in Fortran, with Algorithm 11 embedded as the probe phase. In this implementation, all integration is performed using the integrator DSL48SE [112, 113], which has capabilities for event detection and sparsity exploitation. To use DSL48SE, appropriate Fortran subroutines corresponding to the right-hand side functions of the quadrature ODE (7.14), the coupled probe ODEs (7.2), (7.15), and (7.16), and the bank-locked system (7.11) and (7.12) are generated automatically. This automatic code generation is performed in C++, using a modified version of a code generation tool developed for use in the reverse propagation of McCormick relaxations [120]. This modified code generation tool uses operator overloading in C++ to construct and store a factored representation the right-hand side function $\mathbf{f}$ of the ODE (7.2). Once this factored representation is stored, the tool constructs the required auxiliary right-hand side subroutines by stepping through this factored representation, and writing appropriate Fortran code corresponding to each elemental function encountered.

7.6.2 Examples

In this section, the developed implementation of Algorithms 9 and 11 is applied to various example problems. In each case, the tolerance parameters ϵ and ϵ_g were

Table 7.1: LD-derivative results for Example 7.6.1, with $\Lambda_f =: \{j^*\}$

t_0	t_f	c_0	$\mathbf{M}$	$\mathbf{x}(t_f, \mathbf{c}_0)$	$[x_{t_f}]'(c_0; \mathbf{M})$	$\bar{\zeta}_{j^*}(t_0)$	$\bar{\kappa}_{j^*}(t_0)$
0	2	1	$\begin{bmatrix}1 & 0 & 0\end{bmatrix}$	7.389	$\begin{bmatrix}7.389 & 0 & 0\end{bmatrix}$	$+1$	0
0	2	-1	$\begin{bmatrix}1 & 0 & 0\end{bmatrix}$	-0.1353	$\begin{bmatrix}0.1353 & 0 & 0\end{bmatrix}$	-1	0
0	2	0	$\begin{bmatrix}1 & 0 & 0\end{bmatrix}$	0	$\begin{bmatrix}7.389 & 0 & 0\end{bmatrix}$	$+1$	1
0	2	0	$\begin{bmatrix}-1 & 0 & 0\end{bmatrix}$	0	$\begin{bmatrix}-0.1353 & 0 & 0\end{bmatrix}$	-1	1
0	2	0	$\begin{bmatrix}0 & 2 & 0\end{bmatrix}$	0	$\begin{bmatrix}0 & 14.78 & 0\end{bmatrix}$	$+1$	2
0	2	0	$\begin{bmatrix}0 & 0 & 0\end{bmatrix}$	0	$\begin{bmatrix}0 & 0 & 0\end{bmatrix}$	0	4

chosen to be $\frac{10^{-3}}{t_f - t_0}$, and each required ODE integration was carried out with absolute and relative tolerances of 10^{-8}.

Example 7.6.1. *Consider an instance of the system described by Assumption 7.3.1, with* $n := 1$, $X := \mathbb{R}$, $t_0 := 0$, $t_f := 2$, *and*

$$f : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R} : (t, z) \mapsto |z|.$$

The developed implementation was applied to this system to evaluate $[x_{t_f}]'(c_0; \mathbf{M})$ *at various values of* $c_0 \in \mathbb{R}$ *and* $\mathbf{M} \in \mathbb{R}^{1\times 3}$*; the numerical results of this application are presented in Table 7.1. Observe that* Λ_f *contains a single element; in Table 7.1, this element is denoted* j^**. The obtained results were verified by direct computation in each case, since the considered ODE can be solved analytically.*

Example 7.6.2. *Consider another instance of the system described by Assumption 7.3.1, with* $n := 1$, $X := \mathbb{R}$, $t_0 := 0$, $t_f := 2$, *and*

$$f : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R} : (t, z) \mapsto (1 - t)|z|.$$

The developed implementation was applied to this system to evaluate $[x_{t_f}]'(c_0; \mathbf{M})$ *at various values of* $c_0 \in \mathbb{R}$ *and* $\mathbf{M} \in \mathbb{R}^{1\times 3}$*; the numerical results of this application are presented in Table 7.2. Again,* Λ_f *contains a single element; in Table 7.2, this element is denoted* j^**. The obtained results were verified by direct computation in each case, since the considered ODE system can be solved analytically.*

Table 7.2: LD-derivative results for Example 7.6.2, with $\Lambda_f =: \{j^*\}$

t_0	t_f	c_0	$\mathbf{M}$	$x(t_f, c_0)$	$[x_{t_f}]'(c_0; \mathbf{M})$	$\bar{\zeta}_{j^*}(t_0)$	$\bar{\kappa}_{j^*}(t_0)$
0	2	1	$\begin{bmatrix}1 & 0 & 0\end{bmatrix}$	1.000	$\begin{bmatrix}1.000 & 0 & 0\end{bmatrix}$	$+1$	0
0	2	-1	$\begin{bmatrix}1 & 0 & 0\end{bmatrix}$	-1.000	$\begin{bmatrix}1.000 & 0 & 0\end{bmatrix}$	-1	0
0	2	0	$\begin{bmatrix}1 & 0 & 0\end{bmatrix}$	0	$\begin{bmatrix}1.000 & 0 & 0\end{bmatrix}$	$+1$	1
0	2	0	$\begin{bmatrix}0 & -1 & 0\end{bmatrix}$	0	$\begin{bmatrix}0 & -1.000 & 0\end{bmatrix}$	-1	2
0	2	0	$\begin{bmatrix}0 & 0 & 0\end{bmatrix}$	0	$\begin{bmatrix}0 & 0 & 0\end{bmatrix}$	0	4

7.7 Conclusions

To our knowledge, the method presented in this chapter is the first numerical method for parametric generalized derivative evaluation for a broad class of parametric ordinary differential equations with right-hand side functions that are not differentiable everywhere. Thus, this method extends the scope of established methods for equation-solving and local optimization that are based on generalized derivatives. This method relies heavily on the theoretical foundation provided in Chapters 5 and 6.

Chapter 8

Lexicographic derivatives of hybrid systems

8.1 Introduction

This chapter is reproduced from the article [56]. *Hybrid discrete/continuous systems* [29, 30, 48, 72] represent dynamic systems exhibiting continuous evolution described by classical systems of differential equations, punctuated by well-defined discrete *events*. At these events, the underlying system of differential equations may be switched, and the state variables of the system may be permitted to jump. These discrete transitions permit intuitive modelling of discrete phenomena in the underlying system, including transitions between operating regimes of a process system, activation of safety mechanisms, and changes in thermodynamic phase. More abstract examples of hybrid systems emerge in [102], as auxiliary dynamic systems which provide convex underestimators for the state variables of an underlying continuous system, for use in global optimization methods [45].

Sensitivity analysis results have been obtained for certain hybrid systems [30], which are summarized in Section 8.2 below. In these systems, the functions describing the continuous evolution of the system and the handling of discrete events are all continuously differentiable, and each event time is described as a well-

defined continuously differentiable implicit function of the system parameters. The latter condition is an example of a *transversality condition*. Under these conditions, the state variables are themselves continuously differentiable functions of the parameters (except possibly at event times), with parametric derivatives that are described as the unique solution of an auxiliary hybrid system.

Chapters 3, 5, and 7 have shown that Nesterov's *lexicographic derivatives* [79] may be evaluated for the unique solutions of parametric ordinary differential equations (ODEs) with right-hand side functions that are locally Lipschitz continuous but not necessarily differentiable everywhere. Lexicographic derivatives have been shown in [79] and Chapter 2 to perform as well as elements of Clarke's generalized Jacobian [16] in established numerical methods for solving nonsmooth equation systems (e.g. [22, 92]) and local optimization problems (e.g. [63, 71]). Moreover, any lexicographic derivative of a differentiable function is in fact the classical (Fréchet) derivative, and any lexicographic derivative of a convex function is a subgradient in the sense of convex analysis [79]. To our knowledge, the lexicographic derivative is the only generalized derivative that has these properties and has been described for the solutions of parametric ODEs with nondifferentiable right-hand side functions.

Unlike the sensitivity analysis results for hybrid systems in [30], the theory of Chapter 5 does not permit the system state to jump, but does not require any transversality conditions to be satisfied when the system state visits points of nondifferentiability in the ODE right-hand side function. This chapter seeks to combine the benefits of both approaches, by considering a hybrid system as described in [30], but with all continuous differentiability requirements relaxed, so that the functions involved are lexicographically smooth in the sense of Nesterov [79], but need not be differentiable everywhere. Thus, discrete jumps in the system state are permitted, provided that transversality conditions are satisfied. However, nondifferentability in the functions determining the continuous evolution of the system, the event times, and the system state jumps is also permitted, relaxing the transversality conditions, and permitting more modelling flexibility. The hybrid systems

describing convex relaxations of ODEs in [102] are of this form, for example. Moreover, this flexibility permits sensitivity analysis for certain hybrid systems that exhibit pathological behavior, such as well-defined changes in the discrete mode sequence.

To handle the hybrid systems described in the previous paragraph, this chapter develops sufficient conditions under which a local inverse function or implicit function will be lexicographically smooth, and describes these functions' lexicographic derivatives. These lexicographic derivatives are readily computed when the functions involved are *piecewise differentiable* in the sense of Scholtes [97]. Combining this theory with the theory of Chapter 5, parametric lexicographic derivatives are described for the hybrid system in question, in terms of an auxiliary hybrid system. These lexicographic derivatives are then computed using the LD-derivative, which was defined in Chapter 3 as a variant of the lexicographic derivative that satisfies intuitive calculus rules.

This chapter is structured as follows. Section 8.2 summarizes established results concerning hybrid systems. Section 8.3 presents sufficient conditions under which a local inverse function or a local implicit function will be lexicographically smooth, and describes the corresponding lexicographic derivatives. Section 8.4 presents the main theorem of the chapter, in which lexicographic derivatives are presented for the hybrid systems described above, and Section 8.5 develops intermediate results that are used in the proof of this main theorem. Section 8.6 presents examples in which the developed theory is applied to various hybrid systems for illustration.

8.2 Classical sensitivity analysis for hybrid systems

This section summarizes relevant established results concerning sensitivity analysis for hybrid discrete/continuous systems. The sensitivity analysis results of Galán et al. [30] can be applied to a parametric hybrid system described by:

$$
\begin{aligned}
\mathbf{x}_{(1)}(\tau_{(1)}, \mathbf{p}) &= \boldsymbol{\xi}_{(1)}(\mathbf{p}), \\
\frac{d\mathbf{x}_{(i)}}{dt}(t, \mathbf{p}) &= \mathbf{f}_{(i)}(\mathbf{x}_{(i)}(t, \mathbf{p})), \qquad \forall t \in (\tau_{(i)}, \tau_{(i),f}], \qquad \forall i \in \{1, 2, \ldots\}, \\
0 &= g_{(i)}(\mathbf{x}_{(i)}(\tau_{(i+1)}(\mathbf{p}), \mathbf{p})), \qquad \forall i \in \{1, 2, \ldots\}, \\
\mathbf{x}_{(i+1)}(\tau_{(i+1)}(\mathbf{p}), \mathbf{p}) &= \boldsymbol{\theta}_{(i+1)}(\mathbf{x}_{(i)}(\tau_{(i+1)}(\mathbf{p}), \mathbf{p})), \qquad \forall i \in \{1, 2, \ldots\}.
\end{aligned}
$$

Here, for each $i \in \{1, 2, \ldots\}$, $\tau_{(i+1)}(\mathbf{p})$ denotes the least value of

$$
t^* \in (\tau_{(i)}(\mathbf{p}), \tau_{(i),f}(\mathbf{p}))
$$

for which

$$
0 = g_{(i)}(\mathbf{x}_{(i)}(t^*, \mathbf{p})).
$$

Note that direct dependence of the functions $\mathbf{f}_{(i)}$, $g_{(i)}$, or $\boldsymbol{\theta}_{(i)}$ on t or $\mathbf{p}$ may be included by considering t and $\mathbf{p}$ to be auxiliary state variables, and considering the evolution of an augmented state variable vector $\mathbf{z}_{(i)}(t, \mathbf{p}) := (t, \mathbf{p}, \mathbf{x}_{(i)}(t, \mathbf{p}))$.

The results in [30] assume that the functions $\boldsymbol{\xi}_{(1)}$, $\mathbf{f}_{(i)}$, $g_{(i)}$, and $\boldsymbol{\theta}_{(i)}$ are each $\mathcal{C}^1$; this assumption will be relaxed in Assumption 8.4.1 below. It is also assumed in [30] that a unique solution of the hybrid system exists on a duration $[\tau_{(1)}, \tau_f]$ for all $\mathbf{p}$ in a neighborhood of some $\bar{\mathbf{p}}$, and that the implicit functions $\tau_{(i)}$ are each $\mathcal{C}^1$ at $\bar{\mathbf{p}}$. This latter assumption is implied by the following *transversality conditions*:

$$
0 \neq (\nabla g_{(i)}(\mathbf{x}_{(i)}(\tau_{(i+1)}(\bar{\mathbf{p}}), \bar{\mathbf{p}})))^{\mathrm{T}} \, \mathbf{f}_{(i)}(\mathbf{x}_{(i)}(\tau_{(i+1)}(\bar{\mathbf{p}}), \bar{\mathbf{p}})), \qquad \forall i \in \{1, 2, \ldots\}.
$$

Under the assumptions above, the partial derivative $\frac{\partial \mathbf{x}_{(i)}}{\partial \mathbf{p}}(t, \bar{\mathbf{p}})$ exists whenever $\mathbf{x}_{(i)}(t, \bar{\mathbf{p}})$ is well-defined; this partial derivative is described as the unique solution of the following auxiliary hybrid system:

$$
\begin{aligned}
\frac{\partial \mathbf{x}_{(1)}}{\partial \mathbf{p}}(\tau_{(1)}, \bar{\mathbf{p}}) &= \mathbf{J}\boldsymbol{\xi}_{(1)}(\bar{\mathbf{p}}), \\
\frac{d}{dt}\left[\frac{\partial \mathbf{x}_{(i)}}{\partial \mathbf{p}}\right](t, \bar{\mathbf{p}}) &= \frac{\partial \mathbf{f}_{(i)}}{\partial \mathbf{x}}(\mathbf{x}_{(i)}(t, \bar{\mathbf{p}}))\, \frac{\partial \mathbf{x}_{(i)}}{\partial \mathbf{p}}(t, \bar{\mathbf{p}}), \\
&\qquad \forall t \in (\tau_{(i)}(\bar{\mathbf{p}}), \tau_{(i+1)}(\bar{\mathbf{p}})), \qquad \forall i \in \{1, 2, \ldots\}, \\
\mathbf{J}\tau_{(i+1)}(\bar{\mathbf{p}}) &= -\frac{(\nabla g_{(i)}(\mathbf{x}_{(i)}))^{\mathrm{T}}\, \frac{\partial \mathbf{x}_{(i)}}{\partial \mathbf{p}}}{(\nabla g_{(i)}(\mathbf{x}_{(i)}))^{\mathrm{T}}\, \mathbf{f}_{(i)}(\mathbf{x}_{(i)})}, \qquad \forall i \in \{1, 2, \ldots\}, \\
\frac{\partial \mathbf{x}_{(i+1)}}{\partial \mathbf{p}}(\tau_{(i+1)}(\bar{\mathbf{p}}), \bar{\mathbf{p}}) &= \Big(\mathbf{J}\boldsymbol{\theta}_{(i)}(\mathbf{x}_{(i)})\, \mathbf{f}_{(i)}(\mathbf{x}_{(i)}) - \mathbf{f}_{(i+1)}(\mathbf{x}_{(i+1)})\Big)\, \mathbf{J}\tau_{(i+1)}(\bar{\mathbf{p}}) \\
&\qquad + \mathbf{J}\boldsymbol{\theta}_{(i)}(\mathbf{x}_{(i)})\, \frac{\partial \mathbf{x}_{(i)}}{\partial \mathbf{p}}, \\
&\qquad \forall i \in \{1, 2, \ldots\}.
\end{aligned}
$$

The arguments of $\mathbf{x}_{(i)}$, $\mathbf{x}_{(i+1)}$, and $\frac{\partial \mathbf{x}_{(i)}}{\partial \mathbf{p}}$ in the bottom two equations above are $(\tau_{(i+1)}(\bar{\mathbf{p}}), \bar{\mathbf{p}})$ in each case, and are omitted for simplicity.

8.3 Lexicographic smoothness of inverse and implicit functions

In this section, sufficient conditions are provided for the L-smoothness of local inverse functions and local implicit functions, motivated by the classical inverse and implicit function theorems. Computationally tractable numerical methods are provided for computing the corresponding LD-derivatives, when the functions involved are *piecewise differentiable* in the sense of Scholtes [97].

Theorem 8.3.1. *Given an open set $Y \subset \mathbb{R}^n$ and some $\hat{\mathbf{y}} \in Y$, suppose that a function $\mathbf{f} : Y \to \mathbb{R}^n$ is a Lipschitz homeomorphism on some neighborhood $N \subset Y$ of $\hat{\mathbf{y}}$. Suppose, in addition, that $\mathbf{f}$ is L-smooth at $\hat{\mathbf{y}}$. Then, the corresponding local inverse function $\mathbf{f}^{-1}$ of $\mathbf{f}$ around $\hat{\mathbf{y}}$ is L-smooth at $\hat{\mathbf{z}} := \mathbf{f}(\hat{\mathbf{y}})$; for each $\mathbf{M} \in \mathbb{R}^{n \times p}$, $[\mathbf{f}^{-1}]'(\hat{\mathbf{z}}; \mathbf{M})$ is the unique solution $\mathbf{N} \in \mathbb{R}^{n \times p}$ of the equation system:*

$$
\mathbf{f}'(\hat{\mathbf{y}}; \mathbf{N}) = \mathbf{M}. \tag{8.1}
$$

Proof. Consider any particular matrix $\mathbf{M} := \begin{bmatrix} \mathbf{m}_{(1)} & \cdots & \mathbf{m}_{(p)} \end{bmatrix} \in \mathbb{R}^{n \times p}$. To demonstrate the L-smoothness of $\mathbf{f}^{-1}$ at $\hat{\mathbf{z}}$, [97, Theorem 3.2.3] will be used in an inductive argument to show that, for each $k \in \{0, 1, \ldots, p\}$, there exists a matrix $\mathbf{N}_{(k)} \in \mathbb{R}^{n \times k}$ such that $\mathbf{f}^{(k)}_{\hat{\mathbf{y}}, \mathbf{N}_{(k)}}$ is a Lipschitz homeomorphism on $\mathbb{R}^m$, and such that $[\mathbf{f}^{-1}]^{(k)}_{\hat{\mathbf{z}}, \mathbf{M}}$ exists and is equivalent to $[\mathbf{f}^{(k)}_{\hat{\mathbf{y}}, \mathbf{N}_{(k)}}]^{-1}$ on $\mathbb{R}^n$.

For the case in which $k = 0$, note that $\mathbf{f}$ is directionally differentiable at $\hat{\mathbf{y}}$. Thus, [97, Theorem 3.2.3] implies that $\mathbf{f}'(\hat{\mathbf{y}}; \cdot) \equiv \mathbf{f}^{(0)}_{\hat{\mathbf{y}}, \varnothing_{n \times 0}}$ is a Lipschitz homeomorphism on $\mathbb{R}^n$, and that $\mathbf{f}^{-1}$ is Lipschitz continuous and directionally differentiable at $\hat{\mathbf{z}}$, with

$$[\mathbf{f}^{-1}]'(\hat{\mathbf{z}}; \cdot) \equiv \left[\mathbf{f}^{-1}\right]^{(0)}_{\hat{\mathbf{z}}, \mathbf{M}} \equiv \left[\mathbf{f}^{(0)}_{\hat{\mathbf{y}}, \varnothing_{n \times 0}}\right]^{-1}.$$

For the inductive step, suppose that, for some $k \in \{1, \ldots, p\}$, there exists a matrix $\mathbf{N}_{(k-1)} \in \mathbb{R}^{n \times (k-1)}$ such that $\mathbf{f}^{(k-1)}_{\hat{\mathbf{y}}, \mathbf{N}_{(k-1)}}$ is a Lipschitz homeomorphism on $\mathbb{R}^n$, and such that

$$\left[\mathbf{f}^{(k-1)}_{\hat{\mathbf{y}}, \mathbf{N}_{(k-1)}}\right]^{-1} \equiv \left[\mathbf{f}^{-1}\right]^{(k-1)}_{\hat{\mathbf{z}}, \mathbf{M}}. \tag{8.2}$$

Since $\mathbf{f}^{(k-1)}_{\hat{\mathbf{y}}, \mathbf{N}_{(k-1)}}$ is a homeomorphism, there exists a vector $\mathbf{n}_{(k)} \in \mathbb{R}^n$ such that

$$\mathbf{m}_{(k)} = \mathbf{f}^{(k-1)}_{\hat{\mathbf{y}}, \mathbf{N}_{(k-1)}}(\mathbf{n}_{(k)}).$$

Since $\mathbf{f}$ is L-smooth at $\hat{\mathbf{y}}$, $\mathbf{f}^{(k-1)}_{\hat{\mathbf{y}}, \mathbf{N}_{(k-1)}}$ is directionally differentiable. Thus, by the inductive assumption and [97, Theorem 3.2.3], with $\mathbf{N}_{(k)} := \begin{bmatrix} \mathbf{N}_{(k-1)} & \mathbf{n}_{(k)} \end{bmatrix} \in \mathbb{R}^{n \times k}$, it follows that the mapping

$$\left[\mathbf{f}^{(k-1)}_{\hat{\mathbf{y}}, \mathbf{N}_{(k-1)}}\right]'(\mathbf{n}_{(k)}; \cdot) \equiv \left[\mathbf{f}^{(k-1)}_{\hat{\mathbf{y}}, \mathbf{N}_{(k)}}\right]'(\mathbf{n}_{(k)}; \cdot) \equiv \mathbf{f}^{(k)}_{\hat{\mathbf{y}}, \mathbf{N}_{(k)}}$$

is a Lipschitz homeomorphism. Moreover, by [97, Theorem 3.2.3] and (8.2), the inverse mapping $\left[\mathbf{f}^{(k-1)}_{\hat{\mathbf{y}}, \mathbf{N}_{(k)}}\right]^{-1}$ is directionally differentiable at $\mathbf{m}_{(k)}$, with

$$
\begin{aligned}
\left[\mathbf{f}^{(k)}_{\hat{\mathbf{y}},\mathbf{N}_{(k)}}\right]^{-1} &\equiv \left[\left[\mathbf{f}^{(k-1)}_{\hat{\mathbf{y}},\mathbf{N}_{(k)}}\right]^{-1}\right]'(\mathbf{m}_{(k)};\cdot), \\
&\equiv \left[\left[\mathbf{f}^{(k-1)}_{\hat{\mathbf{y}},\mathbf{N}_{(k-1)}}\right]^{-1}\right]'(\mathbf{m}_{(k)};\cdot), \\
&\equiv \left[\left[\mathbf{f}^{-1}\right]^{(k-1)}_{\hat{\mathbf{z}},\mathbf{M}}\right]'(\mathbf{m}_{(k)};\cdot).
\end{aligned}
$$

It follows that $\left[\mathbf{f}^{-1}\right]^{(k)}_{\hat{\mathbf{z}},\mathbf{M}}$ exists and is equivalent to $\left[\mathbf{f}^{(k)}_{\hat{\mathbf{y}},\mathbf{N}_{(k)}}\right]^{-1}$, which completes the inductive argument. Since $\mathbf{M}$ was chosen arbitrarily, the L-smoothness of $\mathbf{f}^{-1}$ at $\hat{\mathbf{z}}$ is thereby demonstrated.

By definition, the identity $\mathbf{f}(\mathbf{f}^{-1}(\mathbf{z})) = \mathbf{z}$ holds for all $\mathbf{z}$ in $\mathbf{f}(N)$, which is open. Applying the chain rule for LD-derivatives to this identity, for any particular $\mathbf{M} := \begin{bmatrix}\mathbf{m}_{(1)} & \cdots & \mathbf{m}_{(p)}\end{bmatrix} \in \mathbb{R}^{n\times p}$,

$$
\mathbf{f}'\left(\hat{\mathbf{y}}; \left[\mathbf{f}^{-1}\right]'(\hat{\mathbf{z}};\mathbf{M})\right) = \mathbf{M}.
$$

Thus, $\left[\mathbf{f}^{-1}\right]'(\hat{\mathbf{z}};\mathbf{M})$ is a solution $\mathbf{N}$ of (8.1). To show that (8.1) has no more than one solution, let $\mathbf{N} := \begin{bmatrix}\mathbf{n}_{(1)} & \cdots & \mathbf{n}_{(p)}\end{bmatrix} \in \mathbb{R}^{n\times p}$ be a solution of (8.1). Writing the columns of (8.1) separately yields

$$
\mathbf{f}^{(k-1)}_{\hat{\mathbf{y}},\mathbf{M}}(\mathbf{n}_{(k)}) = \mathbf{m}_{(k)}, \qquad \forall k \in \{1,\ldots,p\}. \tag{8.3}
$$

As the earlier inductive argument shows, $\mathbf{f}^{(k-1)}_{\hat{\mathbf{y}},\mathbf{M}}$ is a Lipschitz homeomorphism for each $k \in \{1,\ldots,p\}$. Each equation in (8.3) therefore has a unique solution $\mathbf{n}_{(k)}$, which specifies $\mathbf{N}$ uniquely. □

With $\mathbf{f}$ and $\hat{\mathbf{y}}$ satisfying the conditions of Theorem 8.3.1, let $\partial\mathbf{f}(\hat{\mathbf{y}})$ denote Clarke's generalized Jacobian [16] of $\mathbf{f}$ at $\mathbf{y}$. Since $\mathbf{f}$ is Lipschitz continuous on some neighborhood of $\hat{\mathbf{y}}$, if $\partial\mathbf{f}(\hat{\mathbf{y}})$ contains no singular matrices, then [16, Theorem 7.1.1] demonstrates that $\mathbf{f}$ is a Lipschitz homeomorphism on some neighborhood of $\hat{\mathbf{y}}$.

Theorem 8.3.2. *Given open sets $Y \subset \mathbb{R}^m$ and $Z \subset \mathbb{R}^n$, and a function $\mathbf{h} : Y \times Z \to \mathbb{R}^m$, suppose that $\mathbf{h}(\hat{\mathbf{y}},\hat{\mathbf{z}}) = \mathbf{0}_m$ for some $\hat{\mathbf{y}} \in Y$ and $\hat{\mathbf{z}} \in Z$, and that $\mathbf{h}$ is L-smooth at $(\hat{\mathbf{y}},\hat{\mathbf{z}})$. Suppose, in addition, that the auxiliary mapping $\mathbf{g} : Y \times Z \to \mathbb{R}^m \times Z : (\mathbf{y},\mathbf{z}) \mapsto$*

$(\mathbf{h}(\mathbf{y},\mathbf{z}),\mathbf{z})$ *is a Lipschitz homeomorphism on some open set* $N_{\mathbf{y}} \times N_{\mathbf{z}}$ *that contains* $(\hat{\mathbf{y}},\hat{\mathbf{z}})$*. Then, there exists a locally Lipschitz continuous function* $\boldsymbol{\eta} : N_{\mathbf{z}} \subset \mathbb{R}^n \to \mathbb{R}^m$ *such that* $\boldsymbol{\eta}(\hat{\mathbf{z}}) = \hat{\mathbf{y}}$ *and* $\mathbf{h}(\boldsymbol{\eta}(\mathbf{z}),\mathbf{z}) = \mathbf{0}_m$ *for all* $\mathbf{z} \in N_{\mathbf{z}}$*. Moreover,* $\boldsymbol{\eta}$ *is L-smooth at* $\hat{\mathbf{z}}$*; for each* $\mathbf{M} \in \mathbb{R}^{n\times p}$*,* $\boldsymbol{\eta}'(\hat{\mathbf{z}};\mathbf{M})$ *is the unique solution* $\mathbf{N} \in \mathbb{R}^{m\times p}$ *of the equation system:*

$$\mathbf{h}'((\hat{\mathbf{y}},\hat{\mathbf{z}});(\mathbf{N},\mathbf{M})) = \mathbf{0}_{m\times p}. \tag{8.4}$$

Proof. The existence of $\boldsymbol{\eta}$ follows from Clarke's implicit function theorem [16, Section 7.1]; the inverse of $\mathbf{g}$ on $\mathbf{g}(N_{\mathbf{y}} \times N_{\mathbf{z}})$ is such that $\mathbf{g}^{-1}(\mathbf{0}_m,\mathbf{z}) = (\boldsymbol{\eta}(\mathbf{z}),\mathbf{z})$ for each $\mathbf{z} \in N_{\mathbf{z}}$. By Theorem 8.3.1, $\mathbf{g}^{-1}$ is L-smooth at $(\mathbf{0}_m,\hat{\mathbf{z}})$, and so $\mathbf{g}^{-1}(\mathbf{0}_m,\cdot)$ and $\boldsymbol{\eta}$ are L-smooth at $\hat{\mathbf{z}}$.

Now, Theorem 8.3.1 implies that $[\mathbf{g}^{-1}]'\big((\mathbf{0}_m,\hat{\mathbf{z}});(\mathbf{0}_{m\times p},\mathbf{M})\big)$ is the unique solution $\mathbf{W} := (\mathbf{N},\mathbf{P})$ of the equation system:

$$(\mathbf{0}_{m\times p},\mathbf{M}) = \mathbf{g}'((\hat{\mathbf{y}},\hat{\mathbf{z}});\mathbf{W}).$$

Applying the definition of $\mathbf{g}$, this equation system is equivalent to

$$\begin{bmatrix}\mathbf{0}_{m\times p}\\ \mathbf{M}\end{bmatrix} = \begin{bmatrix}\mathbf{h}'((\hat{\mathbf{y}},\hat{\mathbf{z}});(\mathbf{N},\mathbf{P}))\\ \mathbf{P}\end{bmatrix},$$

and so $\mathbf{P} = \mathbf{M}$. Hence, this equation system is in turn equivalent to (8.4), which therefore has a unique solution $\mathbf{N}$. Now, lexicographic differentiation of the identity $\mathbf{0}_m = \mathbf{h}(\boldsymbol{\eta}(\mathbf{z}),\mathbf{z})$ with respect to $\mathbf{z}$ at $\mathbf{z} = \hat{\mathbf{z}}$ yields

$$\mathbf{0}_{m\times p} = \mathbf{h}'\left((\hat{\mathbf{y}},\hat{\mathbf{z}});\begin{bmatrix}\boldsymbol{\eta}'(\hat{\mathbf{z}};\mathbf{M})\\ \mathbf{M}\end{bmatrix}\right),$$

and so $\boldsymbol{\eta}'(\hat{\mathbf{z}};\mathbf{M})$ is the unique matrix $\mathbf{N}$ which satisfies (8.4). □

As earlier, with $\mathbf{h}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ satisfying the conditions of Theorem 8.3.2, let $\partial\mathbf{h}$ denote Clarke's generalized Jacobian [16] of $\mathbf{h}$. Since $\mathbf{h}$ is Lipschitz continuous on some neighborhood of $(\hat{\mathbf{y}},\hat{\mathbf{z}})$, if the set

$$\left\{\begin{bmatrix}\mathbf{I}_{m\times m} & \mathbf{0}_{m\times n}\end{bmatrix}\mathbf{A} : \mathbf{A} \in \partial\mathbf{h}(\hat{\mathbf{y}},\hat{\mathbf{z}})\right\}$$

contains no singular matrices, then [16, Corollary of Theorem 7.1.1] demonstrates

that the auxiliary function $\mathbf{g}$ described in the statement of the theorem is a Lipschitz homeomorphism on some neighborhood of $(\hat{\mathbf{y}}, \hat{\mathbf{z}})$.

The equation systems (8.1) and (8.4) have residual functions that may be discontinuous with respect to their $\mathbf{N}$ terms. Thus, even though these equation systems have unique solutions, these solutions may be difficult to identify in practice. However, as in the proof of Theorem 8.3.1, if the first $(k-1)$ columns of the solution $\mathbf{N}$ of (8.1) are known, then, with $\mathbf{N}_{(k-1)} := \begin{bmatrix} \mathbf{n}_{(1)} & \cdots & \mathbf{n}_{(k-1)} \end{bmatrix}$, $\mathbf{n}_{(k)}$ can be determined as the unique solution $\mathbf{n}$ of the following equation system. This equation system has a residual function that is Lipschitz continuous, but may not be differentiable everywhere. Here, $\mathbf{e}_{(k)}$ denotes the kth unit coordinate vector in $\mathbb{R}^k$.

$$\mathbf{0}_n = \mathbf{f}^{(k-1)}_{\hat{\mathbf{y}},\mathbf{N}_{(k-1)}}(\mathbf{n}) - \mathbf{m}_{(k)} = \left(\mathbf{f}'(\hat{\mathbf{y}}; \begin{bmatrix} \mathbf{N}_{(k-1)} & \mathbf{n} \end{bmatrix})\right) \mathbf{e}_{(k)} - \mathbf{m}_{(k)}.$$

Thus, the equation system (8.1) may be solved one column at a time, using a numerical method for nonsmooth equation solving. The equation system (8.4) may be approached similarly.

Moreover, when the relevant inverse or implicit functions are described in terms of functions $\mathbf{f}$ or $\mathbf{h}$ that are piecewise differentiable in the sense of Scholtes [97], the above theorems suggest an alternative class of tractable methods for computing the corresponding LD-derivatives. Scholtes' definition is as follows.

Definition 8.3.3 (from [97]). *Given an open set $X \subset \mathbb{R}^n$, a function $\mathbf{g} : X \to \mathbb{R}^m$ is* piecewise differentiable ($\mathcal{PC}^1$) *at $\mathbf{x} \in X$ if there exist a neighborhood $N \subset X$ of $\mathbf{x}$ and a finite collection $\mathcal{F}_{\mathbf{g}}(\mathbf{x})$ of $\mathcal{C}^1$* selection functions *mapping N into $\mathbb{R}^m$, for which $\mathbf{g}$ is continuous on N, and*

$$\mathbf{g}(\mathbf{y}) \in \{\gamma(\mathbf{y}) : \gamma \in \mathcal{F}_{\mathbf{g}}(\mathbf{x})\}, \qquad \forall \mathbf{y} \in N.$$

Let $\partial_{\mathrm{B}}\mathbf{g}$ denote the *B-subdifferential* [92] of $\mathbf{g}$. As in Chapter 3, any $\mathcal{PC}^1$ function $\mathbf{g}$ is L-smooth, and satisfies $\partial_{\mathrm{L}}\mathbf{g}(\mathbf{x}) = \partial_{\mathrm{B}}[\mathbf{g}'(\mathbf{x}; \cdot)](\mathbf{0}_n) \subset \partial_{\mathrm{B}}\mathbf{g}(\mathbf{x})$ for each domain point $\mathbf{x}$. The following examples describe approaches to furnishing a finite collection of selection functions, thus motivating the subsequent methods for solving

the equation systems (8.1) and (8.4) when the functions $\mathbf{f}$ or $\mathbf{h}$ are $\mathcal{PC}^1$. Note that sufficient conditions for a $\mathcal{PC}^1$ function to be a local Lipschitz homeomorphism have been presented in [94, 97].

Example 8.3.4. *Consider a composition* $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ *of simple* $\mathcal{PC}^1$ *functions, such as continuously differentiable functions, the absolute-value function, and multivariate* max *and* min *functions. For any* $\mathbf{x} \in X$*, a finite collection* $\mathcal{F}_{\mathbf{f}}(\mathbf{x})$ *may be furnished by considering the compositions of selection functions of the* $\mathcal{PC}^1$ *functions used to define* $\mathbf{f}$ *that are* active *[97]. For example, consider the* $\mathcal{PC}^1$ *function:*

$$\mathbf{f} : \mathbf{y} \in \mathbb{R}^2 \mapsto |y_1 - y_2 + 1| + |y_1|.$$

Since each absolute-value function has a $y \mapsto -y$ *branch and a* $y \mapsto y$ *branch, the following collection of selection functions for* $\mathbf{f}$ *around* $\mathbf{x} := (1, 0) \in \mathbb{R}^2$ *is readily furnished.*

$$\mathcal{F}_{\mathbf{f}}(\mathbf{x}) := \{\mathbf{y} \mapsto s_1(y_1 - y_2 + 1) + s_2 y_1 : s_1, s_2 \in \{-1, +1\}\}.$$

It is possible, however, to furnish an even smaller collection of selection functions for $\mathbf{f}$*. In a sufficiently small neighborhood of* $\mathbf{x}$*, observe that the second absolute-value function in the definition of* $\mathbf{f}$ *is restricted to its* $y \mapsto y$ *branch, while the first absolute-value function could be described by either its* $y \mapsto y$ *branch or its* $y \mapsto -y$ *branch. Thus, the mappings* $\mathbf{y} \mapsto (y_1 - y_2 + 1) + y_1$ *and* $\mathbf{y} \mapsto -(y_1 - y_2 + 1) + y_1$ *together comprise a smaller collection* $\mathcal{F}_{\mathbf{f}}(\mathbf{x})$ *for* $\mathbf{f}$ *around* $\mathbf{x}$*.*

Consider a finite set $Z \subset \mathbb{N}$*, a finite collection* $\{g_{(i)} : i \in Z\}$ *of continuously differentiable mappings from* $\mathbb{R}^n$ *into* $\mathbb{R}$*, and the function*

$$g : \mathbf{y} \in \mathbb{R}^n \mapsto \max\{g_{(i)}(\mathbf{y}) : i \in Z\}.$$

The function g *is evidently* $\mathcal{PC}^1$*, and* $\{g_{(i)} : i \in Z\}$ *is clearly a finite collection of selection functions for* $g_{(i)}$ *about any* $\mathbf{y} \in \mathbb{R}^n$*. Again, though, it may not be necessary to retain each function in this collection. Following a similar approach to the treatment of* $\mathbf{f}$ *above, it is readily verified that, given any particular* $\mathbf{z} \in \mathbb{R}^n$*, the following is a collection of selection functions for* g *around* $\mathbf{z}$*:*

$$\{g_{(i)} : i \in Z,\ g_{(i)}(\mathbf{z}) = g(\mathbf{z})\}.$$

Observe that each condition $g_{(i)}(\mathbf{z}) = g(\mathbf{z})$ would be checked incidentally during evaluation of $g(\mathbf{z})$.

Proposition 8.3.5. *Suppose that the conditions of Theorem 8.3.1 hold, that the function $\mathbf{f}$ is $\mathcal{PC}^1$ at $\hat{\mathbf{y}}$, and that a finite collection $\mathcal{F}_{\mathbf{f}}(\hat{\mathbf{y}})$ of $\mathcal{C}^1$ selection functions for $\mathbf{f}$ around $\hat{\mathbf{y}}$ is known. For any $\mathbf{M} \in \mathbb{R}^{n\times p}$, there exists $\phi \in \mathcal{F}_{\mathbf{f}}(\hat{\mathbf{y}})$ for which $\mathbf{J}\phi(\hat{\mathbf{y}})$ is nonsingular and $[\mathbf{f}^{-1}]'(\hat{\mathbf{z}};\mathbf{M}) = (\mathbf{J}\phi(\hat{\mathbf{y}}))^{-1}\mathbf{M}$. Thus, the following method solves the equation system* (8.1) *for $\mathbf{N} = [\mathbf{f}^{-1}]'(\hat{\mathbf{z}};\mathbf{M})$:*

```
for all φ ∈ F_f(ŷ) do
  if Jφ(ŷ) is nonsingular then
    Solve the linear equation system Jφ(ŷ) A = M for A ∈ R^{n×p}
    if f'(ŷ; A) = M then
      return N := A
    end if
  end if
end for
```

Proof. This proof depends on the following established results. [97, Proposition 4.2.1] implies that $\mathbf{f}^{-1}$ is $\mathcal{PC}^1$ at $\hat{\mathbf{z}}$; Corollary 3.2.4 then yields $\partial_{\mathrm{L}}[\mathbf{f}^{-1}](\hat{\mathbf{z}}) \subset \partial_{\mathrm{B}}[\mathbf{f}^{-1}](\hat{\mathbf{z}})$. There exists a subset $\mathcal{E}_{\mathbf{f}}(\hat{\mathbf{y}}) \subset \mathcal{F}_{\mathbf{f}}(\hat{\mathbf{y}})$ of $\mathcal{C}^1$ selection functions for $\mathbf{f}$ about $\hat{\mathbf{y}}$ that are *essentially active* at $\hat{\mathbf{y}}$ in the sense of Scholtes [97]. Next, [97, Proposition 4.2.1] and its proof imply that $\mathbf{J}\phi(\hat{\mathbf{y}})$ is nonsingular for each $\phi \in \mathcal{E}_{\mathbf{f}}(\hat{\mathbf{y}})$. The classical inverse function theorem then implies that, for each $\phi \in \mathcal{E}_{\mathbf{f}}(\hat{\mathbf{y}})$, ϕ is invertible in a neighborhood of $\hat{\mathbf{y}}$, with an inverse ϕ^{-1} that is $\mathcal{C}^1$ at $\hat{\mathbf{z}}$, and satisfies

$$\mathbf{J}[\phi^{-1}](\hat{\mathbf{z}}) = (\mathbf{J}\phi(\hat{\mathbf{y}}))^{-1}.$$

Lastly, the proofs of [97, Proposition 4.2.1 and Proposition 4.3.1] imply that

$$\partial_{\mathrm{B}}[\mathbf{f}^{-1}](\hat{\mathbf{z}}) \subset \{\mathbf{J}[\phi^{-1}](\hat{\mathbf{z}}) : \phi \in \mathcal{E}_{\mathbf{f}}(\hat{\mathbf{y}})\}.$$

Combining the above results with Lemma 3.1.3 and Theorem 8.3.1, there exists $\phi \in \mathcal{E}_{\mathbf{f}}(\hat{\mathbf{y}}) \subset \mathcal{F}_{\mathbf{f}}(\hat{\mathbf{y}})$ such that $\mathbf{J}\phi(\hat{\mathbf{y}})$ is nonsingular, and satisfies

$$\mathbf{f}'(\hat{\mathbf{y}}; (\mathbf{J}\phi(\hat{\mathbf{y}}))^{-1}\mathbf{M}) = \mathbf{M}.$$

Thus, $(\mathbf{J}\phi(\hat{\mathbf{y}}))^{-1}\mathbf{M}$ is the unique solution $\mathbf{N}$ of (8.1) that was described by Theorem 8.3.1. The remaining required results follow immediately. □

Proposition 8.3.6. *Suppose that the conditions of Theorem 8.3.2 hold, that the function* $\mathbf{h}$ *is* $\mathcal{PC}^1$ *at* $(\hat{\mathbf{y}}, \hat{\mathbf{z}})$*, and that a finite collection* $\mathcal{F}_{\mathbf{h}}(\hat{\mathbf{y}}, \hat{\mathbf{z}})$ *of* $\mathcal{C}^1$ *selection functions for* $\mathbf{h}$ *around* $(\hat{\mathbf{y}}, \hat{\mathbf{z}})$ *is known. For any* $\mathbf{M} \in \mathbb{R}^{n\times p}$*, there exists* $\psi \in \mathcal{F}_{\mathbf{h}}(\hat{\mathbf{y}}, \hat{\mathbf{z}})$ *for which* $\frac{\partial \psi}{\partial \mathbf{y}}(\hat{\mathbf{y}}, \hat{\mathbf{z}})$ *is nonsingular and*

$$\boldsymbol{\eta}'(\hat{\mathbf{z}}; \mathbf{M}) = -\left(\frac{\partial \psi}{\partial \mathbf{y}}(\hat{\mathbf{y}}, \hat{\mathbf{z}})\right)^{-1} \frac{\partial \psi}{\partial \mathbf{z}}(\hat{\mathbf{y}}, \hat{\mathbf{z}})\, \mathbf{M}.$$

Thus, the following method solves the equation system (8.4) *for* $\mathbf{N} = \boldsymbol{\eta}'(\hat{\mathbf{z}}; \mathbf{M})$*:*

for all $\psi \in \mathcal{F}_{\mathbf{h}}(\hat{\mathbf{y}}, \hat{\mathbf{z}})$ ***do***
  if $\frac{\partial \psi}{\partial \mathbf{y}}(\hat{\mathbf{y}}, \hat{\mathbf{z}})$ *is nonsingular* ***then***
    Solve the linear equation system $\frac{\partial \psi}{\partial \mathbf{y}}(\hat{\mathbf{y}}, \hat{\mathbf{z}})\, \mathbf{A} = \frac{\partial \psi}{\partial \mathbf{z}}(\hat{\mathbf{y}}, \hat{\mathbf{z}})\, \mathbf{M}$ *for* $\mathbf{A} \in \mathbb{R}^{m\times p}$
    if $\mathbf{h}'\left((\hat{\mathbf{y}}, \hat{\mathbf{z}}); (\mathbf{A}, \mathbf{M})\right) = \mathbf{0}_{m\times p}$ ***then***
      return $\mathbf{N} := \mathbf{A}$
    end if
  end if
end for

Proof. This proof proceeds similarly to the proof of Proposition 8.3.5. Define auxiliary linear mappings $\boldsymbol{\pi} : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m : (\mathbf{y}, \mathbf{z}) \mapsto \mathbf{y}$ and $\boldsymbol{\theta} : \mathbb{R}^n \to \mathbb{R}^{m+n} : \mathbf{z} \mapsto (\mathbf{0}_m, \mathbf{z})$. Since $\mathbf{h}$ is $\mathcal{PC}^1$ at $(\hat{\mathbf{y}}, \hat{\mathbf{z}})$, and since any well-defined finite composition of $\mathcal{PC}^1$ functions is itself $\mathcal{PC}^1$ [97], the auxiliary mapping $\mathbf{g}$ described in Theorem 8.3.2 is also $\mathcal{PC}^1$ at $(\hat{\mathbf{y}}, \hat{\mathbf{z}})$. It follows from the proof of Theorem 8.3.2 that $\mathbf{g}$ is invertible on some neighborhood of $(\hat{\mathbf{y}}, \hat{\mathbf{z}})$, with an inverse $\mathbf{g}^{-1}$ that is L-smooth at $(\mathbf{0}_m, \hat{\mathbf{z}}) = \boldsymbol{\theta}(\hat{\mathbf{z}})$. Moreover, [97, Proposition 4.2.1] implies that $\mathbf{g}^{-1}$ is $\mathcal{PC}^1$ at $\boldsymbol{\theta}(\hat{\mathbf{z}})$.

Now, there exists a subset $\mathcal{E}_{\mathbf{h}}(\hat{\mathbf{y}}, \hat{\mathbf{z}}) \subset \mathcal{F}_{\mathbf{h}}(\hat{\mathbf{y}}, \hat{\mathbf{z}})$ of $\mathcal{C}^1$ selection functions for $\mathbf{h}$ about $(\hat{\mathbf{y}}, \hat{\mathbf{z}})$ that are essentially active at $(\hat{\mathbf{y}}, \hat{\mathbf{z}})$ in the sense of Scholtes [97]. For each $\psi \in \mathcal{E}_{\mathbf{h}}(\hat{\mathbf{y}}, \hat{\mathbf{z}})$, define a $\mathcal{C}^1$ mapping $\gamma_\psi : (\mathbf{y}, \mathbf{z}) \mapsto (\psi(\mathbf{y}, \mathbf{z}), \mathbf{z})$ on some neighborhood of $(\hat{\mathbf{y}}, \hat{\mathbf{z}})$. Thus,

$$\mathbf{J}\gamma_\psi(\hat{\mathbf{y}},\hat{\mathbf{z}}) = \begin{bmatrix} \frac{\partial\psi}{\partial\mathbf{y}}(\hat{\mathbf{y}},\hat{\mathbf{z}}) & \frac{\partial\psi}{\partial\mathbf{z}}(\hat{\mathbf{y}},\hat{\mathbf{z}}) \\ \mathbf{0}_{n\times m} & \mathbf{I}_{n\times n} \end{bmatrix} \in \mathbb{R}^{(m+n)\times(m+n)}. \tag{8.5}$$

The definition of $\mathbf{g}$ implies that a collection of essentially active $\mathcal{C}^1$ selection functions for $\mathbf{g}$ about $(\hat{\mathbf{y}},\hat{\mathbf{z}})$ can be expressed as:

$$\mathcal{E}_{\mathbf{g}}(\hat{\mathbf{y}},\hat{\mathbf{z}}) := \{\gamma_\psi : \psi \in \mathcal{E}_{\mathbf{h}}(\hat{\mathbf{y}},\hat{\mathbf{z}})\}.$$

Proposition 8.3.5 (with $\mathbf{f} := \mathbf{g}$) shows that $\mathbf{J}\gamma_\psi(\hat{\mathbf{y}},\hat{\mathbf{z}})$ is nonsingular for each $\psi \in \mathcal{E}_{\mathbf{h}}(\hat{\mathbf{y}},\hat{\mathbf{z}})$; (8.5) then implies that $\frac{\partial\psi}{\partial\mathbf{y}}(\hat{\mathbf{y}},\hat{\mathbf{z}})$ is nonsingular for each $\psi \in \mathcal{E}_{\mathbf{h}}(\hat{\mathbf{y}},\hat{\mathbf{z}})$, and that

$$(\mathbf{J}\gamma_\psi(\hat{\mathbf{y}},\hat{\mathbf{z}}))^{-1} = \begin{bmatrix} \left(\frac{\partial\psi}{\partial\mathbf{y}}(\hat{\mathbf{y}},\hat{\mathbf{z}})\right)^{-1} & -\left(\frac{\partial\psi}{\partial\mathbf{y}}(\hat{\mathbf{y}},\hat{\mathbf{z}})\right)^{-1}\frac{\partial\psi}{\partial\mathbf{z}}(\hat{\mathbf{y}},\hat{\mathbf{z}}) \\ \mathbf{0}_{n\times m} & \mathbf{I}_{n\times n} \end{bmatrix} \in \mathbb{R}^{(m+n)\times(m+n)}.$$

Moreover, as in the proof of Proposition 8.3.5,

$$\begin{aligned} &\partial_{\mathrm{L}}[\mathbf{g}^{-1}](\boldsymbol{\theta}(\hat{\mathbf{z}})) \\ &\quad \subset \partial_{\mathrm{B}}[\mathbf{g}^{-1}](\boldsymbol{\theta}(\hat{\mathbf{z}})) \\ &\quad \subset \{\mathbf{J}[\gamma^{-1}](\boldsymbol{\theta}(\hat{\mathbf{z}})) : \gamma \in \mathcal{E}_{\mathbf{g}}(\hat{\mathbf{y}},\hat{\mathbf{z}})\} \\ &\quad = \{(\mathbf{J}\gamma(\hat{\mathbf{y}},\hat{\mathbf{z}}))^{-1} : \gamma \in \mathcal{E}_{\mathbf{g}}(\hat{\mathbf{y}},\hat{\mathbf{z}})\} \\ &\quad = \{(\mathbf{J}\gamma_\psi(\hat{\mathbf{y}},\hat{\mathbf{z}}))^{-1} : \psi \in \mathcal{E}_{\mathbf{h}}(\hat{\mathbf{y}},\hat{\mathbf{z}})\}. \end{aligned}$$

Now, inspection of the proof of Theorem 8.3.2 shows that $\boldsymbol{\eta} \equiv \boldsymbol{\pi} \circ \mathbf{g}^{-1} \circ \boldsymbol{\theta}$ on some neighborhood of $\hat{\mathbf{z}}$. Thus, the above results, Lemma 3.1.3, and the chain rule for LD-derivatives yield:

$$
\begin{aligned}
&\boldsymbol{\eta}'(\hat{\mathbf{z}};\mathbf{M}) \\
&\quad = \mathbf{J}\boldsymbol{\pi}(\mathbf{g}^{-1}(\boldsymbol{\theta}(\hat{\mathbf{z}}))) \; [\mathbf{g}^{-1}]'(\boldsymbol{\theta}(\hat{\mathbf{z}});\mathbf{J}\boldsymbol{\theta}(\hat{\mathbf{z}})\,\mathbf{M}) \\
&\quad = \begin{bmatrix}\mathbf{I}_{m\times m} & \mathbf{0}_{m\times n}\end{bmatrix} \; [\mathbf{g}^{-1}]' \left(\boldsymbol{\theta}(\hat{\mathbf{z}});(\mathbf{0}_{m\times p},\mathbf{M})\right) \\
&\quad \in \left\{ \begin{bmatrix}\mathbf{I}_{m\times m} & \mathbf{0}_{m\times n}\end{bmatrix} (\mathbf{J}\boldsymbol{\gamma}_{\boldsymbol{\psi}}(\hat{\mathbf{y}},\hat{\mathbf{z}}))^{-1} \begin{bmatrix}\mathbf{0}_{m\times p} \\ \mathbf{M}\end{bmatrix} : \boldsymbol{\psi} \in \mathcal{E}_{\mathbf{h}}(\hat{\mathbf{y}},\hat{\mathbf{z}}) \right\} \\
&\quad = \left\{ \begin{bmatrix}\mathbf{I}_{m\times m} & \mathbf{0}_{m\times n}\end{bmatrix} \begin{bmatrix} \left(\frac{\partial\boldsymbol{\psi}}{\partial\mathbf{y}}(\hat{\mathbf{y}},\hat{\mathbf{z}})\right)^{-1} & -\left(\frac{\partial\boldsymbol{\psi}}{\partial\mathbf{y}}(\hat{\mathbf{y}},\hat{\mathbf{z}})\right)^{-1} \frac{\partial\boldsymbol{\psi}}{\partial\mathbf{z}}(\hat{\mathbf{y}},\hat{\mathbf{z}}) \\ \mathbf{0}_{n\times m} & \mathbf{I}_{n\times n} \end{bmatrix} \begin{bmatrix}\mathbf{0}_{m\times p} \\ \mathbf{M}\end{bmatrix} : \right. \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \left. \boldsymbol{\psi} \in \mathcal{E}_{\mathbf{h}}(\hat{\mathbf{y}},\hat{\mathbf{z}}) \right\} \\
&\quad = \left\{ -\left(\frac{\partial\boldsymbol{\psi}}{\partial\mathbf{y}}(\hat{\mathbf{y}},\hat{\mathbf{z}})\right)^{-1} \frac{\partial\boldsymbol{\psi}}{\partial\mathbf{z}}(\hat{\mathbf{y}},\hat{\mathbf{z}})\,\mathbf{M} : \boldsymbol{\psi} \in \mathcal{E}_{\mathbf{h}}(\hat{\mathbf{y}},\hat{\mathbf{z}}) \right\}.
\end{aligned}
$$

The remaining required results follow immediately. □

Observe that computational complexity of the method in Proposition 8.3.5 scales worst-case linearly with the number of selection functions in the provided collection $\mathcal{F}_{\mathbf{f}}(\hat{\mathbf{y}})$, and the cost of solving each required linear equation system scales as $\mathcal{O}(n^3 p)$. Similarly, the computational complexity of the method in Proposition 8.3.6 scales worst-case linearly with the number of selection functions in the provided collection $\mathcal{F}_{\mathbf{h}}(\hat{\mathbf{y}},\hat{\mathbf{z}})$, and the cost of solving each required linear system scales as $\mathcal{O}(m^3 p)$.

When the functions $\mathbf{f}$ and $\mathbf{h}$ are represented as finite compositions of simple L-smooth functions, the derivatives and partial derivatives required by the above methods can typically be computed using standard automatic differentiation techniques [34], and the required LD-derivatives can be computed using the variant of the vector forward mode of automatic differentiation developed in Chapter 4.

8.4 LD-derivatives for hybrid systems

This section describes a hybrid discrete/continuous system that is based on the presentation by Galán et al. [30], but is generalized to permit the functions involved to be L-smooth instead of $\mathcal{C}^1$. Theorem 8.4.2 below is the main theorem of

this chapter, and describes the LD-derivatives of this hybrid system as the unique solution of an auxiliary hybrid system. The proof of this theorem depends on intermediate results that are developed in Section 8.5.

The following assumption formalizes the hybrid system for which lexicographic derivatives are desired. As in Section 8.2, any direct dependence of the functions $\mathbf{f}_{(i)}$, $g_{(i)}$, and $\boldsymbol{\theta}_{(i)}$ on t or $\mathbf{p}$ is neglected without loss of generality; any such direct dependence may be included by appending auxiliary state variables to $\mathbf{x}$ that contain the values of t and $\mathbf{p}$. Discontinuities of $\mathbf{f}_{(i)}$ with respect to t may also be handled in this framework by appending corresponding functions $g_{(j)}$ and $\boldsymbol{\theta}_{(j)}$ to represent every such discontinuity.

Assumption 8.4.1. *For some $n_m \in \mathbb{N}$, for each $i \in \{1, 2, \ldots, n_m\}$, let $X_{(i)} \subset \mathbb{R}^{n_{(i)}}$ be an open set, and let a function $\mathbf{f}_{(i)} : X_{(i)} \to \mathbb{R}^{n_{(i)}}$ be bounded, Lipschitz continuous, and L-smooth. Let functions $g_{(i)} : X_{(i)} \to \mathbb{R}$ for each $i \in \{1, \ldots, n_m\}$ and $\boldsymbol{\theta}_{(i)} : X_{(i-1)} \to X_{(i)}$ for each $i \in \{2, 3, \ldots, n_m\}$ be Lipschitz continuous and L-smooth.*

For some open set $P \subset \mathbb{R}^{n_p}$ and some particular $\bar{\mathbf{p}} \in P$, let functions $\tau_{(1)} : P \to \mathbb{R}$ and $\boldsymbol{\xi}_{(1)} : P \to X_{(1)}$ be L-smooth at $\bar{\mathbf{p}}$. Suppose $\bar{\tau}_f \in \mathbb{R}$ is such that $\bar{\tau}_f > \tau_{(1)}(\mathbf{p})$ for each $\mathbf{p} \in P$.

Consider the following parametric hybrid discrete/continuous system, defined for each $\mathbf{p} \in P$:

$$\begin{aligned}
\mathbf{x}_{(1)}(\tau_{(1)}(\mathbf{p}), \mathbf{p}) &= \boldsymbol{\xi}_{(1)}(\mathbf{p}), \\
\frac{d\mathbf{x}_{(i)}}{dt}(t, \mathbf{p}) &= \mathbf{f}_{(i)}(\mathbf{x}_{(i)}(t, \mathbf{p})), && \forall i \in \{1, \ldots, n_m\}, && (8.6) \\
0 &= g_{(i)}(\mathbf{x}_{(i)}(\tau_{(i+1)}(\mathbf{p}), \mathbf{p})), && \forall i \in \{1, \ldots, n_m - 1\}, && (8.7) \\
\mathbf{x}_{(i+1)}(\tau_{(i+1)}(\mathbf{p}), \mathbf{p}) &= \boldsymbol{\theta}_{(i+1)}(\mathbf{x}_{(i)}(\tau_{(i+1)}(\mathbf{p}), \mathbf{p})), && \forall i \in \{1, , \ldots, n_m - 1\}. && (8.8)
\end{aligned}$$

Here, for each $i \in \{1, \ldots, n_m - 1\}$, $\tau_{(i+1)}(\mathbf{p})$ denotes the least value of $t^* \in (\tau_{(i)}(\mathbf{p}), \bar{\tau}_f)$ such that

$$0 = g_{(i)}(\mathbf{x}_{(i)}(t^*, \mathbf{p})),$$

and $\tau_{(n_m+1)}(\mathbf{p}) := \bar{\tau}_f$. For each $\mathbf{p} \in P$ in a particular neighborhood of some $\bar{\mathbf{p}} \in P$,

for each $i \in \{1, \ldots, n_m\}$*, suppose that there exists a unique solution* $\{\mathbf{x}_{(i)}(t, \mathbf{p}) : t \in [\tau_{(i)}(\mathbf{p}), \tau_{f,i}(\mathbf{p})]\} \subset X_{(i)}$ *of* (8.6) *for some* $\tau_{f,i}(\mathbf{p}) > \tau_{(i+1)}(\mathbf{p})$*.*

Suppose that for each $i \in \{1, \ldots, n_m - 1\}$*, the composition* $g_{(i)} \circ \mathbf{x}_{(i)}$ *satisfies the assumptions of Theorem 8.3.2 at* $(\tau_{(i+1)}(\bar{\mathbf{p}}), \bar{\mathbf{p}})$*. For simplicity, define* $\tau^*_{(i)} := \tau_{(i)}(\bar{\mathbf{p}})$ *for each* $i \in \{1, \ldots, n_m + 1\}$*, and* $\mathbf{x}^*_{(i)} := \mathbf{x}_{(i)}(\tau^*_{(i+1)}, \bar{\mathbf{p}})$ *for each* $i \in \{1, \ldots, n_m - 1\}$*.*

Galan et al. [30] permit each $\mathbf{x}_{(i+1)}(\tau_{(i+1)}(\mathbf{p}), \mathbf{p})$ to be described as an implicit function of $\mathbf{x}_{(i)}(\tau_{(i+1)}(\mathbf{p}), \mathbf{p})$, instead of being specified explicitly by the function $\boldsymbol{\theta}_{(i)}$ in (8.8). The results in this chapter are compatible with this approach; an implicit version of (8.8) can be handled using Theorem 8.3.2. For simplicity, we do not pursue this further.

Moreover, the results in this chapter remain valid if, in Assumption 8.4.1, the functions $\mathbf{f}_{(i)}$, $g_{(i)}$, and $\boldsymbol{\theta}_{(i)}$ are not uniformly Lipschitz continuous. In fact, the local Lipschitz continuity implied by L-smoothness yields uniform Lipschitz continuity on some open superset of the domain points visited by the solution of the hybrid system with $\mathbf{p} := \bar{\mathbf{p}}$. Explicit consideration of these sets, however, would obscure the arguments underlying our developed results, and so we retain Lipschitz continuity in Assumption 8.4.1 for simplicity.

The following theorem is the main theorem of this chapter, and describes LD-derivatives of the hybrid system described by Assumption 8.4.1. The proof of this theorem depends on various intermediate results that are presented in the following section. Various implications of this theorem are described at the end of the current section.

Theorem 8.4.2. *Suppose that Assumption 8.4.1 holds. Then, for each* $j \in \{1, \ldots, n_m\}$ *and each* $\tilde{t} \in (\tau_{(j)}(\bar{\mathbf{p}}), \tau_{(j+1)}(\bar{\mathbf{p}})]$*,* $\mathbf{x}_{(j)}(\tilde{t}, \cdot)$ *is L-smooth at* $\bar{\mathbf{p}}$*; for any* $\mathbf{M} \in \mathbb{R}^{n_p \times p}$*,* $[\mathbf{x}_{(j),\tilde{t}}]'(\bar{\mathbf{p}}; \mathbf{M})$ *is the matrix* $\mathbf{A}_{(j)}(\tilde{t}) \in \mathbb{R}^{n_{(j)} \times p}$ *defined uniquely by the following hybrid discrete/continuous system:*

$$\mathbf{A}_{(1)}(\tau_{(1)}(\bar{\mathbf{p}})) = [\boldsymbol{\xi}_{(1)}]'(\bar{\mathbf{p}};\mathbf{M}) - \mathbf{f}_{(1)}(\boldsymbol{\xi}_{(1)}(\bar{\mathbf{p}}))\,[\tau_{(1)}]'(\bar{\mathbf{p}};\mathbf{M}), \tag{8.9}$$

$$\frac{d\mathbf{A}_{(i)}}{dt}(t) = [\mathbf{f}_{(i)}]'(\mathbf{x}_{(i)}(t,\bar{\mathbf{p}});\mathbf{A}_{(i)}(t)), \tag{8.10}$$

$$\forall t \in (\tau^*_{(i)}, \tau^*_{(i+1)}], \qquad \forall i \in \{1,\ldots,n_m\},$$

$$\mathbf{0}_{1\times p} = [g_{(i)}]'\Big(\mathbf{x}^*_{(i)}; \mathbf{f}_{(i)}(\mathbf{x}^*_{(i)})\,[\tau_{(i+1)}]'(\bar{\mathbf{p}};\mathbf{M}) + \mathbf{A}_{(i)}(\tau^*_{(i+1)})\Big), \tag{8.11}$$

$$\forall i \in \{1,\ldots,n_m-1\},$$

$$\begin{aligned}\mathbf{A}_{(i+1)}(\tau^*_{(i+1)}) = {}& [\boldsymbol{\theta}_{(i+1)}]'\Big(\mathbf{x}^*_{(i)}; \mathbf{A}_{(i)}(\tau^*_{(i+1)}) + \mathbf{f}_{(i)}(\mathbf{x}^*_{(i)})\,[\tau_{(i+1)}]'(\bar{\mathbf{p}};\mathbf{M})\Big) \\ & - \mathbf{f}_{(i+1)}\left(\mathbf{x}_{(i+1)}(\tau^*_{(i+1)},\bar{\mathbf{p}})\right)\,[\tau_{(i+1)}]'(\bar{\mathbf{p}};\mathbf{M}),\end{aligned} \tag{8.12}$$

$$\forall i \in \{1,\ldots,n_m-1\},$$

where each $[\tau_{(i+1)}]'(\bar{\mathbf{p}};\mathbf{M})$ *is defined implicitly as the unique solution of* (8.11). *If, for some* $i \in \{1,\ldots,n_m-1\}$, $g_{(i)}$ *is differentiable at* $\mathbf{x}^*_{(i)}$, *then*

$$[\tau_{(i+1)}]'(\bar{\mathbf{p}};\mathbf{M}) = -\frac{(\nabla g_{(i)}(\mathbf{x}^*_{(i)}))^{\mathrm{T}}\,\mathbf{A}_{(i)}(\tau^*_{(i+1)})}{(\nabla g_{(i)}(\mathbf{x}^*_{(i)}))^{\mathrm{T}}\,\mathbf{f}_{(i)}(\mathbf{x}^*_{(i)})}; \tag{8.13}$$

if $\boldsymbol{\theta}_{(i+1)}$ *is differentiable at* $\mathbf{x}^*_{(i)}$, *then*

$$\begin{aligned}&\mathbf{A}_{(i+1)}(\tau^*_{(i+1)}) - \mathbf{A}_{(i)}(\tau^*_{(i+1)}) \\ &\quad = \Big(\mathbf{J}\boldsymbol{\theta}_{(i+1)}(\mathbf{x}^*_{(i)}) - \mathbf{I}\Big)\,\mathbf{A}_{(i)}(\tau^*_{(i+1)}) \\ &\qquad - \Big(\mathbf{f}_{(i+1)}\left(\mathbf{x}_{(i+1)}(\tau^*_{(i+1)},\bar{\mathbf{p}})\right) - \mathbf{J}\boldsymbol{\theta}_{(i+1)}(\mathbf{x}^*_{(i)})\,\mathbf{f}_{(i)}(\mathbf{x}^*_{(i)})\Big)\,[\tau_{(i+1)}]'(\bar{\mathbf{p}};\mathbf{M}).\end{aligned} \tag{8.14}$$

In particular, if $\mathbf{f}_{(i+1)}(\mathbf{x}_{(i+1)}(\tau^*_{(i+1)},\bar{\mathbf{p}})) = \mathbf{J}\boldsymbol{\theta}_{(i+1)}(\mathbf{x}^*_{(i)})\,\mathbf{f}_{(i)}(\mathbf{x}^*_{(i)})$, *then there is no need to evaluate* $[\tau_{(i+1)}]'(\bar{\mathbf{p}};\mathbf{M})$. *If, in addition,* $\boldsymbol{\theta}_{(i+1)}$ *is the identity mapping, then*

$$\mathbf{A}_{(i+1)}(\tau^*_{(i+1)}) = \mathbf{A}_{(i)}(\tau^*_{(i+1)}).$$

Proof. For each $j \in \{2,\ldots,n_m\}$, define the mapping $\boldsymbol{\xi}_{(j)} : \mathbf{p} \mapsto \mathbf{x}_{(j)}(\tau_{(j)}(\mathbf{p}),\mathbf{p})$, which is analogous to $\boldsymbol{\xi}_{(1)}$. In an inductive argument, it will be shown that for each $j \in \{1,\ldots,n_m\}$:

(A) $\tau_{(j)}$ is L-smooth at $\bar{\mathbf{p}}$, and, if $j > 1$, $[\tau_{(j)}]'(\bar{\mathbf{p}};\mathbf{M})$ is determined uniquely by (8.11),

(B) $\boldsymbol{\xi}_{(j)}$ is L-smooth at $\bar{\mathbf{p}}$, and, if $j > 1$,

$$[\boldsymbol{\xi}_{(j)}]'(\bar{\mathbf{p}};\mathbf{M}) = [\boldsymbol{\theta}_{(j)}]'\Big(\mathbf{x}^*_{(j-1)};\mathbf{A}_{(j-1)}(\tau^*_{(j)}) + \mathbf{f}_{(j-1)}(\mathbf{x}^*_{(j-1)})\,[\tau_{(j)}]'(\bar{\mathbf{p}};\mathbf{M})\Big)\,,$$

(C) for each $t \in (\tau^*_{(j)}, \tau^*_{(j+1)}]$, $\mathbf{x}_{(j),t} \equiv \mathbf{x}_{(j)}(t,\cdot)$ is L-smooth at $\bar{\mathbf{p}}$, and $[\mathbf{x}_{(j),t}]'(\bar{\mathbf{p}};\mathbf{M}) = \mathbf{A}_{(j)}(t)$, where $\mathbf{A}_{(j)}(t)$ is defined uniquely by (8.9)–(8.12).

Corollary 8.5.5 and the assumptions of the theorem yield (A), (B), and (C) when $j := 1$. For the inductive step, choose some particular $k \in \{2,\ldots,n_m\}$, and suppose that (A), (B), and (C) hold for $j := k-1$.

By the inductive assumption, the hypotheses of Corollary 8.5.7 and Lemma 8.5.6 are satisfied with $\tau := \tau_{(k-1)}$, $\boldsymbol{\xi} := \boldsymbol{\xi}_{(k-1)}$, $\mathbf{x} := \mathbf{x}_{(k-1)}$, $\mathbf{f} := \mathbf{f}_{(k-1)}$, and $t_e := \tau_{(k)}$. Thus, Corollary 8.5.7 yields (A) with $j := k$, and Lemma 8.5.6 yields the L-smoothness of $\mathbf{x}_{(k-1)}$ in some neighborhood of $(\tau_{(k)}(\bar{\mathbf{p}}), \bar{\mathbf{p}})$.

Defining the mapping $\boldsymbol{\gamma}_{(k)} : \mathbf{p} \mapsto (\tau_{(k)}(\mathbf{p}), \mathbf{p})$, it follows that $\boldsymbol{\gamma}_{(k)}$ is L-smooth near $\bar{\mathbf{p}}$, and so the composite mapping $[\boldsymbol{\theta}_{(k)} \circ \mathbf{x}_{(k-1)} \circ \boldsymbol{\gamma}_{(k)}]$ is L-smooth at $\bar{\mathbf{p}}$. By definition, for each $\mathbf{p}$ near $\bar{\mathbf{p}}$,

$$\boldsymbol{\xi}_{(k)}(\mathbf{p}) = \mathbf{x}_{(k)}(\boldsymbol{\gamma}_{(k)}(\mathbf{p})) = \boldsymbol{\theta}_{(k)}(\mathbf{x}_{(k-1)}(\boldsymbol{\gamma}_{(k)}(\mathbf{p}))).$$

Applying the chain rule for LD-derivatives, it follows that $\boldsymbol{\xi}_{(k)}$ is L-smooth at $\bar{\mathbf{p}}$, with

$$\begin{aligned}[\boldsymbol{\xi}_{(k)}]'(\bar{\mathbf{p}};\mathbf{M}) &= [\boldsymbol{\theta}_{(k)}]'\Big(\mathbf{x}^*_{(k-1)};[\mathbf{x}_{(k-1)}]'\Big(\boldsymbol{\gamma}_{(k)}(\bar{\mathbf{p}});[\boldsymbol{\gamma}_{(k)}]'(\bar{\mathbf{p}};\mathbf{M})\Big)\Big)\\ &= [\boldsymbol{\theta}_{(k)}]'\left(\mathbf{x}^*_{(k-1)};[\mathbf{x}_{(k-1)}]'\left(\boldsymbol{\gamma}_{(k)}(\bar{\mathbf{p}});\begin{bmatrix}[\tau_{(k)}]'(\bar{\mathbf{p}};\mathbf{M})\\ \mathbf{M}\end{bmatrix}\right)\right).\end{aligned}$$

By applying Lemma 8.5.6 as above to evaluate the LD-derivative of $\mathbf{x}_{(k-1)}$ in the above expression, (B) is obtained with $j := k$.

Since (A) and (B) have been demonstrated when $j := k$, Corollary 8.5.5 yields (C) when $j := k$. This completes the inductive assumption, and thereby demonstrates the claims of the theorem regarding (8.9)–(8.12).

If $g_{(i)}$ is differentiable for some particular $i \in \{1, \ldots, n_m - 1\}$, then (8.11) becomes

$$\mathbf{0}_{1\times p} = (\nabla g_{(i)}(\mathbf{x}^*_{(i)}))^{\mathrm{T}} \left(\mathbf{f}_{(i)}(\mathbf{x}^*_{(i)})\, [\tau_{(i+1)}]'(\bar{\mathbf{p}}; \mathbf{M}) + \mathbf{A}_{(i)}(\tau^*_{(i+1)})\right).$$

As established by the inductive proof above, this equation has a unique solution $[\tau_{(i+1)}]'(\bar{\mathbf{p}}; \mathbf{M})$, and so a similar argument to the proof of Corollary 8.5.7 demonstrates (8.13). Moreover, if $\boldsymbol{\theta}_{(i+1)}$ is differentiable, then (8.12) becomes

$$\begin{aligned}\mathbf{A}_{(i+1)}(\tau^*_{(i+1)}) = \mathbf{J}\boldsymbol{\theta}_{(i+1)}(\mathbf{x}^*_{(i)}) &\left(\mathbf{A}_{(i)}(\tau^*_{(i+1)}) + \mathbf{f}_{(i)}(\mathbf{x}^*_{(i)})\, [\tau_{(i+1)}]'(\bar{\mathbf{p}}; \mathbf{M})\right) \\ &- \mathbf{f}_{(i+1)}\left(\mathbf{x}_{(i+1)}(\tau^*_{(i+1)}, \bar{\mathbf{p}})\right)\, [\tau_{(i+1)}]'(\bar{\mathbf{p}}; \mathbf{M}),\end{aligned}$$

which can be rearranged to yield (8.14). The remaining claims of the theorem follow immediately. □

Observe that the auxiliary hybrid system (8.9)–(8.12) described by Theorem 8.4.2 reduces to the classical hybrid sensitivities described in Section 8.2 when the functions $\mathbf{f}_{(i)}$, $g_{(i)}$, and $\boldsymbol{\theta}_{(i)}$ are each $\mathcal{C}^1$. Observe also that if $\mathbf{M}$ has a single column, then the auxiliary hybrid system (8.9)–(8.12) describes directional derivatives of the state variables $\mathbf{x}$ with respect to the parameters $\mathbf{p}$.

Theorem 8.4.2 does not apply to all hybrid systems; Assumption 8.4.1 requires that the *discrete modes* of the hybrid system, enumerated by the index i, are necessarily visited in the order $i := 1, 2, 3, \ldots$. However, although Assumption 8.4.1 nominally requires $\tau_{(i+1)}(\mathbf{p})$ to be strictly greater than $\tau_{(i)}(\mathbf{p})$ for each i and each $\mathbf{p}$ near $\bar{\mathbf{p}}$, the proofs of Theorem 8.4.2 and Corollary 8.5.7 in fact suggest that $\tau_{(i+1)}(\mathbf{p}) = \tau_{(i)}(\mathbf{p})$ is permissible, provided that there exists a neighborhood N of $\bar{\mathbf{p}}$ such that $\tau_{(i+1)}(\mathbf{p}) \geq \tau_{(i)}(\mathbf{p})$ for each $\mathbf{p} \in N$, and such that $\tau_{(i+1)}$ is still a well-defined L-smooth implicit function near $\bar{\mathbf{p}}$. This observation permits handling of certain changes in the discrete *mode sequence* visited by the solution trajectory, in which small changes in parameters $\mathbf{p}$ can lead to the discrete index i being updated in an order other than $i := 1, 2, \ldots$. This possibility is illustrated in Example 8.6.3.

As an incidental corollary of Theorem 8.4.2, observe that LD-derivatives of each

event time $\tau_{(i+1)}$ at $\bar{\mathbf{p}}$ are described by (8.11) (or by (8.13), if applicable).

We conclude this section by noting that it would be difficult, if not impossible, to verify the *transversality conditions* in Assumption 8.4.1 that the composite functions $g_{(i)} \circ \mathbf{x}_{(i)}$ each satisfy the conditions of Theorem 8.3.2 at $(\tau_{(i+1)}(\bar{\mathbf{p}}), \bar{\mathbf{p}})$. By contrast, in the simpler hybrid system described in Section 8.2, the corresponding transversality conditions can be verified during the evaluation of the derivative $\mathbf{J}\tau_{(i+1)}(\bar{\mathbf{p}})$. Unlike the development in [30], however, the transversality conditions in Assumption 8.4.1 apply only to the discrete transitions described by the functions $g_{(i)}$ and $\boldsymbol{\theta}_{(i+1)}$, and not to any nondifferentiabilities in the functions $\mathbf{f}_{(i)}$. If the latter were incorporated directly into the theory of [30], then appropriate transversality conditions would need to be applied.

8.5 Intermediate results

This section presents intermediate results that were used in the proof of Theorem 8.4.2 above. Roughly, these intermediate results permit LD-derivatives to be propagated over each discrete mode i of the hybrid system described in Assumption 8.4.1: detecting the events $\tau_{(i)}(\bar{\mathbf{p}})$, carrying out the transitions described by $\boldsymbol{\theta}_{(i)}$, and resuming continuous evolution of the system following these transitions. These results are developed without considering the hybrid system in Assumption 8.4.1 explicitly; rather, it suffices in this section to consider simpler ODEs with a right-hand side function $\mathbf{f}$ that is described by the following assumption.

Assumption 8.5.1. *Let $X \subset \mathbb{R}^n$ be an open set, and consider a function $\mathbf{f} : X \to \mathbb{R}^n$ which is bounded, Lipschitz continuous, and directionally differentiable. Suppose that $m_{\mathbf{f}} > 0$ is a bound for $\mathbf{f}$ on X, and that $k_{\mathbf{f}} > 0$ is a Lipschitz constant for $\mathbf{f}$ on X.*

Remark 8.5.2. *Under Assumption 8.5.1, it follows from [26, §1, Theorem 2] that any solution of an ODE with $\mathbf{f}$ as its right-hand side function is unique. Thus, the assumed uniqueness of ODE solutions in the following results is for clarity only, and does not contribute at all to the hypotheses of these results.*

The following lemma provides a variant of Theorem 5.2.4 that considers the dependence of an ODE dependent variable $\mathbf{x}$ on the initial value τ of the independent variable t.

Lemma 8.5.3. *Suppose that Assumption 8.5.1 holds. With $\mathbf{x}(\cdot, \tau, \boldsymbol{\xi})$ denoting any solution of the parametric ODE system:*

$$\frac{d\mathbf{x}}{dt}(t, \tau, \boldsymbol{\xi}) = \mathbf{f}(\mathbf{x}(t, \tau, \boldsymbol{\xi})), \qquad \mathbf{x}(\tau, \tau, \boldsymbol{\xi}) = \boldsymbol{\xi}, \tag{8.15}$$

suppose that there exists a unique solution $\{\mathbf{x}(t, \bar{\tau}, \bar{\boldsymbol{\xi}}) : t \in [\bar{\tau}, \bar{\tau}_f]\} \subset X$ for some $\bar{\tau}, \bar{\tau}_f \in \mathbb{R}$ with $\bar{\tau} < \bar{\tau}_f$ and some $\bar{\boldsymbol{\xi}} \in X$. Under these assumptions, for each $t \in [\bar{\tau}, \bar{\tau}_f]$, the mapping $\mathbf{x}_t \equiv \mathbf{x}(t, \cdot, \cdot)$ is well-defined and Lipschitz continuous on a neighborhood of $(\bar{\tau}, \bar{\boldsymbol{\xi}})$, with a Lipschitz constant that is independent of t. Moreover, $\mathbf{x}_t$ is directionally differentiable at $(\bar{\tau}, \bar{\boldsymbol{\xi}})$; for any $\alpha \in \mathbb{R}$ and $\mathbf{d} \in \mathbb{R}^n$, the mapping $t \mapsto [\mathbf{x}_t]'((\bar{\tau}, \bar{\boldsymbol{\xi}}); (\alpha, \mathbf{d}))$ is the unique solution $\mathbf{y}$ on $[\bar{\tau}, \bar{\tau}_f]$ of the ODE:

$$\frac{d\mathbf{y}}{dt}(t) = \mathbf{f}'(\mathbf{x}(t, \bar{\tau}, \bar{\boldsymbol{\xi}}); \mathbf{y}(t)), \qquad \mathbf{y}(\bar{\tau}) = \mathbf{d} - \alpha \mathbf{f}(\bar{\boldsymbol{\xi}}). \tag{8.16}$$

If, in addition, $\mathbf{f}$ is L-smooth on X, then $\mathbf{x}_t$ is also L-smooth at $(\bar{\tau}, \bar{\boldsymbol{\xi}})$; for any $\mathbf{v} \in \mathbb{R}^p$ and any $\mathbf{M} \in \mathbb{R}^{n \times p}$, the mapping $t \mapsto [\mathbf{x}_t]'\big((\bar{\tau}, \bar{\boldsymbol{\xi}}); (\mathbf{v}^{\mathrm{T}}, \mathbf{M})\big)$ is the unique solution $\mathbf{A}$ on $[\bar{\tau}, \bar{\tau}_f]$ of the ODE:

$$\frac{d\mathbf{A}}{dt}(t) = \mathbf{f}'(\mathbf{x}(t, \bar{\tau}, \bar{\boldsymbol{\xi}}); \mathbf{A}(t)), \qquad \mathbf{A}(\bar{\tau}) = \mathbf{M} - \mathbf{f}(\bar{\boldsymbol{\xi}})\, \mathbf{v}^{\mathrm{T}}. \tag{8.17}$$

Proof. By [18, Chapter 1, Theorem 4.1], the solution $\{\mathbf{x}(t, \bar{\tau}, \bar{\boldsymbol{\xi}}) : t \in [\bar{\tau}, \bar{\tau}_f]\} \subset X$ may be continued to some interval $[a, b] \subset \mathbb{R}$ such that $a < \bar{\tau}$ and $b > \bar{\tau}_f$, while remaining in X. Thus, [18, Chapter 1, Theorem 7.1] shows that there exists some neighborhood $N_\tau \subset \mathbb{R}$ of 0 and $\bar{N} \subset X$ of $\bar{\boldsymbol{\xi}}$ such that for each $(\tau, \boldsymbol{\xi}) \in (\{\bar{\tau}\} + N_\tau) \times \bar{N}$, there exists a unique solution $\{\mathbf{x}(t, \tau, \boldsymbol{\xi}) : t \in [\tau, \tau_f]\} \subset X$ of (8.15), for any $\tau_f \in (\bar{\tau}_f + N_\tau)$. Hence, $\mathbf{x}(t, \cdot, \cdot)$ is well-defined on $(\{\bar{\tau}\} + N_\tau) \times \bar{N}$ for each $t \in [\bar{\tau}, \bar{\tau}_f]$.

Suppose that $\mathbf{f}$ is directionally differentiable, but not necessarily L-smooth. Consider the following auxiliary ODE system, with a parameter $\mathbf{c} \in \mathbb{R}^n$:

$$\frac{d\mathbf{z}}{dt}(t,\mathbf{c}) = \mathbf{f}(\mathbf{z}(t,\mathbf{c})), \qquad \mathbf{z}(\bar{\tau},\mathbf{c}) = \mathbf{c}. \tag{8.18}$$

Unlike (8.15), (8.18) has a fixed initial independent variable, and is therefore amenable to the treatment of Chapter 5. Since the mapping $t \mapsto \mathbf{x}(t,\bar{\tau},\bar{\boldsymbol{\xi}})$ evidently solves (8.18) on $[\bar{\tau},\bar{\tau}_f]$ when $\mathbf{c} := \bar{\boldsymbol{\xi}}$, (8.18) satisfies the hypotheses of Theorem 5.2.1. Thus, according to that theorem, for each $t \in [\bar{\tau},\bar{\tau}_f]$, the mapping $\mathbf{z}_t \equiv \mathbf{z}(t,\cdot)$ is well-defined and Lipschitz continuous on a neighborhood $N_{\mathbf{z}}$ of $\bar{\boldsymbol{\xi}}$, with a Lipschitz constant $k_{\mathbf{z}}$ that is independent of t.

To show that $\mathbf{x}_t$ is Lipschitz continuous on $(\{\bar{\tau}\} + N_\tau) \times (\bar{N} \cap N_{\mathbf{z}})$ for any particular $t \in [\bar{\tau},\bar{\tau}_f]$, with a Lipschitz constant that is independent of t, consider any $(\tau_1,\boldsymbol{\xi}_1),(\tau_2,\boldsymbol{\xi}_2) \in (\{\bar{\tau}\} + N_\tau) \times (\bar{N} \cap N_{\mathbf{z}})$. It follows that

$$\begin{aligned}
&\|\mathbf{x}_t(\tau_1,\boldsymbol{\xi}_1) - \mathbf{x}_t(\tau_2,\boldsymbol{\xi}_2)\| \\
&\quad \le \|\mathbf{x}_t(\tau_1,\boldsymbol{\xi}_1) - \mathbf{x}_t(\tau_1,\boldsymbol{\xi}_2)\| + \|\mathbf{x}_t(\tau_1,\boldsymbol{\xi}_2) - \mathbf{x}_t(\tau_2,\boldsymbol{\xi}_2)\| \\
&\quad = \|\mathbf{x}_{t+\bar{\tau}-\tau_1}(\bar{\tau},\boldsymbol{\xi}_1) - \mathbf{x}_{t+\bar{\tau}-\tau_1}(\bar{\tau},\boldsymbol{\xi}_2)\| + \|\mathbf{x}_{t+\bar{\tau}-\tau_1}(\bar{\tau},\boldsymbol{\xi}_2) - \mathbf{x}_{t+\bar{\tau}-\tau_2}(\bar{\tau},\boldsymbol{\xi}_2)\| \\
&\quad \le \|\mathbf{z}_{t+\bar{\tau}-\tau_1}(\boldsymbol{\xi}_1) - \mathbf{z}_{t+\bar{\tau}-\tau_1}(\boldsymbol{\xi}_2)\| + \left\| \int_{t+\bar{\tau}-\tau_2}^{t+\bar{\tau}-\tau_1} \mathbf{f}(\mathbf{x}(s,\bar{\tau},\boldsymbol{\xi}_2))\, ds \right\| \\
&\quad \le k_{\mathbf{z}}\|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\| + m_{\mathbf{f}}|\tau_1 - \tau_2| \\
&\quad \le (k_{\mathbf{z}} + m_{\mathbf{f}})\|(\tau_1,\boldsymbol{\xi}_1) - (\tau_2,\boldsymbol{\xi}_2)\|_1.
\end{aligned}$$

Next, defining the mapping

$$\bar{\mathbf{x}} : \mathbb{R} \times X \to \mathbb{R}^n : (\tau,\boldsymbol{\xi}) \mapsto \boldsymbol{\xi} - (\tau - \bar{\tau})\mathbf{f}(\bar{\boldsymbol{\xi}}),$$

this proof proceeds by showing that for each $t \in [\bar{\tau},\bar{\tau}_f]$, $\mathbf{z}_t \circ \bar{\mathbf{x}}$ is a good enough first-order approximation of $\mathbf{x}_t$ near $(\bar{\tau},\bar{\boldsymbol{\xi}})$ to share the same directional derivatives and lexicographic derivatives at $(\bar{\tau},\bar{\boldsymbol{\xi}})$.

Now, choose any $\alpha \in \mathbb{R}$ and $\mathbf{d} \in \mathbb{R}^n$. By [26, §1, Theorems 2 and 6], noting that $\bar{\mathbf{x}}$ is Lipschitz continuous on its domain, there exists $\bar{s} > 0$ such that for each $s \in (0,\bar{s})$ and each $t \in [\bar{\tau},\bar{\tau}_f]$, $\mathbf{x}(t,\bar{\tau} + s\alpha,\bar{\boldsymbol{\xi}} + s\mathbf{d})$ and $\mathbf{z}(t,\bar{\mathbf{x}}(\bar{\tau} + s\alpha,\bar{\boldsymbol{\xi}} + s\mathbf{d}))$ are each well-defined.

For each $t \in [\bar{\tau},\bar{\tau}_f]$ and each $s \in (0,\bar{s})$, let

$$\mathbf{e}_{\mathbf{z}}(s,t) := (\mathbf{z}(t,\bar{\mathbf{x}}(\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d})) - \mathbf{z}(t,\bar{\mathbf{x}}(\bar{\tau},\bar{\boldsymbol{\xi}}))) - (\mathbf{x}(t,\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d}) - \mathbf{x}(t,\bar{\tau},\bar{\boldsymbol{\xi}}))\,.$$

By Theorem 5.2.1, $\mathbf{z}_t$ is directionally differentiable at $\bar{\mathbf{x}}(\bar{\tau},\bar{\boldsymbol{\xi}}) = \bar{\boldsymbol{\xi}}$. Applying the chain rule for directional derivatives,

$$[\mathbf{z}_t \circ \bar{\mathbf{x}}]'((\bar{\tau},\bar{\boldsymbol{\xi}});(\alpha,\mathbf{d})) = [\mathbf{z}_t]'(\bar{\mathbf{x}}(\bar{\tau},\bar{\boldsymbol{\xi}});\bar{\mathbf{x}}'((\bar{\tau},\bar{\boldsymbol{\xi}});(\alpha,\mathbf{d}))) = [\mathbf{z}_t]'(\bar{\boldsymbol{\xi}};\mathbf{d}-\alpha\mathbf{f}(\bar{\boldsymbol{\xi}})).$$

Noting that $\mathbf{x}(t,\bar{\tau},\bar{\boldsymbol{\xi}}) = \mathbf{z}(t,\bar{\boldsymbol{\xi}})$ for each $t \in [\bar{\tau},\bar{\tau}_f]$, it follows from Theorem 5.2.1 that the mapping $t \mapsto [\mathbf{z}_t \circ \bar{\mathbf{x}}]'((\bar{\tau},\bar{\boldsymbol{\xi}});(\alpha,\mathbf{d}))$ is the unique solution $\mathbf{y}$ of (8.16) on $[\bar{\tau},\bar{\tau}_f]$. Thus, to show that the mapping $t \mapsto [\mathbf{x}_t]'((\bar{\tau},\bar{\boldsymbol{\xi}});(\alpha,\mathbf{d}))$ exists and solves (8.16) uniquely on $[\bar{\tau},\bar{\tau}_f]$, it suffices to show that

$$\lim_{s\to 0^+} \frac{\mathbf{e}_{\mathbf{z}}(s,t)}{s} = \mathbf{0}_n, \qquad \forall t \in [\bar{\tau},\bar{\tau}_f].$$

Noting again that $\mathbf{x}(t,\bar{\tau},\bar{\boldsymbol{\xi}}) = \mathbf{z}(t,\bar{\mathbf{x}}(\bar{\tau},\bar{\boldsymbol{\xi}}))$ for each $t \in [\bar{\tau},\bar{\tau}_f]$ yields

$$\mathbf{e}_{\mathbf{z}}(s,t) = \mathbf{z}(t,\bar{\mathbf{x}}(\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d})) - \mathbf{x}(t,\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d}).$$

Fixing $s \in (0,\bar{s})$ and $i \in \{1,\ldots,n\}$, note that

$$\bar{\mathbf{x}}(\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d}) = \bar{\boldsymbol{\xi}}+s\mathbf{d}-s\alpha\mathbf{f}(\bar{\boldsymbol{\xi}}).$$

Assume that $\alpha \geq 0$; the case in which $\alpha < 0$ is analogous. Since $\mathbf{f}$ is continuous, the mean-value theorem [95, Theorem 5.19] yields the existence of $\theta_s \in [0,s\alpha]$ such that

$$\begin{aligned}\left\| \int_{\bar{\tau}}^{\bar{\tau}+s\alpha} (\mathbf{f}(\mathbf{z}(t,\bar{\mathbf{x}}(\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d}))) - \mathbf{f}(\bar{\boldsymbol{\xi}}))\,dt \right\| \\ \leq s\alpha \|\mathbf{f}(\mathbf{z}(\bar{\tau}+\theta_s,\bar{\mathbf{x}}(\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d}))) - \mathbf{f}(\bar{\boldsymbol{\xi}})\|.\end{aligned}$$

Using the obtained results, it follows that

$$
\begin{aligned}
&\|\mathbf{e_z}(s,\bar{\tau}+s\alpha)\| \\
&\quad = \|\mathbf{z}(\bar{\tau}+s\alpha,\bar{\mathbf{x}}(\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d}))-\mathbf{x}(\bar{\tau}+s\alpha,\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d})\| \\
&\quad = \left\|\left(\bar{\mathbf{x}}(\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d})+\int_{\bar{\tau}}^{\bar{\tau}+s\alpha}\mathbf{f}(\mathbf{z}(t,\bar{\mathbf{x}}(\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d})))\,dt\right)-(\bar{\boldsymbol{\xi}}+s\mathbf{d})\right\| \\
&\quad = \left\|\int_{\bar{\tau}}^{\bar{\tau}+s\alpha}\mathbf{f}(\mathbf{z}(t,\bar{\mathbf{x}}(\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d})))\,dt-s\alpha\mathbf{f}(\bar{\boldsymbol{\xi}})\right\| \\
&\quad \le s\alpha\|\mathbf{f}(\mathbf{z}(\bar{\tau}+\theta_s,\bar{\mathbf{x}}(\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d})))-\mathbf{f}(\bar{\boldsymbol{\xi}})\| \\
&\quad \le s\alpha k_{\mathbf{f}}\,\|\mathbf{z}(\bar{\tau}+\theta_s,\bar{\boldsymbol{\xi}}+s\mathbf{d}-s\alpha\mathbf{f}(\bar{\boldsymbol{\xi}}))-\bar{\boldsymbol{\xi}}\| \\
&\quad = s\alpha k_{\mathbf{f}}\,\|\mathbf{z}(\bar{\tau}+\theta_s,\bar{\boldsymbol{\xi}}+s\mathbf{d}-s\alpha\mathbf{f}(\bar{\boldsymbol{\xi}}))-\mathbf{z}(\bar{\tau},\bar{\boldsymbol{\xi}})\| \\
&\quad \le s\alpha k_{\mathbf{f}}\,(\|\mathbf{z}(\bar{\tau}+\theta_s,\bar{\boldsymbol{\xi}}+s\mathbf{d}-s\alpha\mathbf{f}(\bar{\boldsymbol{\xi}}))-\mathbf{z}(\bar{\tau},\bar{\boldsymbol{\xi}}+s\mathbf{d}-s\alpha\mathbf{f}(\bar{\boldsymbol{\xi}}))\| \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + \|\mathbf{z}(\bar{\tau},\bar{\boldsymbol{\xi}}+s\mathbf{d}-s\alpha\mathbf{f}(\bar{\boldsymbol{\xi}}))-\mathbf{z}(\bar{\tau},\bar{\boldsymbol{\xi}})\|) \\
&\quad \le s\alpha k_{\mathbf{f}}\left(\int_{\bar{\tau}}^{\bar{\tau}+\theta_s}\|\mathbf{f}(\mathbf{z}(t,\bar{\boldsymbol{\xi}}+s\mathbf{d}-s\alpha\mathbf{f}(\bar{\boldsymbol{\xi}})))\|\,dt+k_{\mathbf{z}}\,\|s\mathbf{d}-s\alpha\mathbf{f}(\bar{\boldsymbol{\xi}})\|\right) \\
&\quad \le s\alpha k_{\mathbf{f}}\,(\theta_s m_{\mathbf{f}}+sk_{\mathbf{z}}(\|\mathbf{d}\|+\alpha m_{\mathbf{f}})) \\
&\quad \le s^2\alpha k_{\mathbf{f}}\,(\alpha m_{\mathbf{f}}+k_{\mathbf{z}}(\|\mathbf{d}\|+\alpha m_{\mathbf{f}}))\,.
\end{aligned}
$$

So, with $\delta_{\mathbf{z}} := \alpha k_{\mathbf{f}}\,(\alpha m_{\mathbf{f}}+k_{\mathbf{z}}(\|\mathbf{d}\|+\alpha m_{\mathbf{f}}))$, $\|\mathbf{e_z}(s,\bar{\tau}+s\alpha)\| \le s^2\delta_{\mathbf{z}}$. For each $t \in [\bar{\tau},\bar{\tau}_f]$,

$$
\begin{aligned}
\|\mathbf{e_z}(s,t)\| &= \Bigg\|\mathbf{z}(\bar{\tau}+s\alpha,\bar{\mathbf{x}}(\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d}))+\int_{\bar{\tau}+s\alpha}^{t}\mathbf{f}(\mathbf{z}(r,\bar{\mathbf{x}}(\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d})))\,dr \\
&\qquad -\mathbf{x}(\bar{\tau}+s\alpha,\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d})-\int_{\bar{\tau}+s\alpha}^{t}\mathbf{f}(\mathbf{x}(r,\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d}))\,dr\Bigg\| \\
&\le \|\mathbf{z}(\bar{\tau}+s\alpha,\bar{\mathbf{x}}(\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d}))-\mathbf{x}(\bar{\tau}+s\alpha,\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d})\| \\
&\qquad +\int_{\bar{\tau}+s\alpha}^{t}\|\mathbf{f}(\mathbf{z}(r,\bar{\mathbf{x}}(\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d})))-\mathbf{f}(\mathbf{x}(r,\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d}))\|\,dr \\
&\le \|\mathbf{e_z}(s,\bar{\tau}+s\alpha)\| \\
&\qquad + k_{\mathbf{f}}\int_{\bar{\tau}+s\alpha}^{t}\|\mathbf{z}(r,\bar{\mathbf{x}}(\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d}))-\mathbf{x}(r,\bar{\tau}+s\alpha,\bar{\boldsymbol{\xi}}+s\mathbf{d})\|\,dr \\
&\le s^2\delta_{\mathbf{z}}+k_{\mathbf{f}}\int_{\bar{\tau}+s\alpha}^{t}\|\mathbf{e_z}(s,r)\|\,dr.
\end{aligned}
$$

Applying the version of Gronwall's inequality described in [122, Section 1], for each $t \in [\bar{\tau},\bar{\tau}_f]$,

$$
\|\mathbf{e_z}(s,t)\| \le s^2\delta_z\exp(k_{\mathbf{f}}(t-\bar{\tau}-s\alpha)).
$$

Noting that s is positive, dividing both sides of the above inequality by s and taking the limit $s \to 0^+$ yields

$$0 \leq \lim_{s \to 0^+} \frac{\|\mathbf{e}_{\mathbf{z}}(s,t)\|}{s} \leq 0 \exp(k_{\mathbf{f}}(t - \bar{\tau})) = 0,$$

as required. As discussed earlier, this implies that $\mathbf{x}_t$ is directionally differentiable at $(\bar{\tau}, \bar{\boldsymbol{\xi}})$ for each $t \in [\bar{\tau}, \bar{\tau}_f]$, and that the mapping $t \mapsto [\mathbf{x}_t]'((\bar{\tau}, \bar{\boldsymbol{\xi}}); (\alpha, \mathbf{d}))$ is the unique solution $\mathbf{y}$ of (8.16) on $[\bar{\tau}, \bar{\tau}_f]$.

In the remainder of this proof, suppose that $\mathbf{f}$ is L-smooth, and consider any fixed $t \in [\bar{\tau}, \bar{\tau}_f]$. To show that $\mathbf{x}_t$ is L-smooth, consider any $\mathbf{v} := (v_1, \ldots, v_p) \in \mathbb{R}^p$ and any $\mathbf{M} := \begin{bmatrix} \mathbf{m}_{(1)} & \cdots & \mathbf{m}_{(p)} \end{bmatrix} \in \mathbb{R}^{n \times p}$. Define $\bar{\gamma} := (\bar{\tau}, \bar{\boldsymbol{\xi}}) \in \mathbb{R}^{n+1}$ and $\mathbf{W} := (\mathbf{v}^{\mathrm{T}}, \mathbf{M}) \in \mathbb{R}^{(n+1) \times p}$. The directional differentiability results established above show that with $k := 0$,

$$[\mathbf{x}_t]^{(k)}_{\bar{\gamma},\mathbf{W}} \equiv [\mathbf{z}_t \circ \bar{\mathbf{x}}]^{(k)}_{\bar{\gamma},\mathbf{W}}.$$

The following simple inductive argument shows that the above equivalence holds for each $k \in \{0, 1, \ldots, p\}$. Suppose that for some $k \in \{1, \ldots, p\}$, $[\mathbf{x}_t]^{(k-1)}_{\bar{\gamma},\mathbf{W}}$ exists, and is equivalent to $[\mathbf{z}_t \circ \bar{\mathbf{x}}]^{(k-1)}_{\bar{\gamma},\mathbf{W}}$. By Theorem 5.2.4, $\mathbf{z}_t$ is L-smooth at $\bar{\mathbf{x}}(\bar{\gamma}) = \bar{\boldsymbol{\xi}}$. Since $\bar{\mathbf{x}}$ is linear, it is also L-smooth, and so the composition $\mathbf{z}_t \circ \bar{\mathbf{x}}$ is L-smooth at $\bar{\gamma}$ as well. Hence, $[\mathbf{z}_t \circ \bar{\mathbf{x}}]^{(k)}_{\bar{\gamma},\mathbf{W}}$ exists, and the inductive assumption yields

$$[\mathbf{z}_t \circ \bar{\mathbf{x}}]^{(k)}_{\bar{\gamma},\mathbf{W}} := \left[[\mathbf{z}_t \circ \bar{\mathbf{x}}]^{(k-1)}_{\bar{\gamma},\mathbf{W}} \right]' \left(\begin{bmatrix} v_k \\ \mathbf{m}_{(k)} \end{bmatrix}; \cdot \right) \equiv \left[[\mathbf{x}_t]^{(k-1)}_{\bar{\gamma},\mathbf{W}} \right]' \left(\begin{bmatrix} v_k \\ \mathbf{m}_{(k)} \end{bmatrix}; \cdot \right).$$

Thus, $[\mathbf{x}_t]^{(k)}_{\bar{\gamma},\mathbf{W}}$ exists and is equivalent to $[\mathbf{z}_t \circ \bar{\mathbf{x}}]^{(k)}_{\bar{\gamma},\mathbf{W}}$, which completes the inductive argument. The arbitrariness of $\mathbf{v}$ and $\mathbf{M}$ shows that $\mathbf{x}_t$ is L-smooth at $\bar{\gamma}$, with

$$[\mathbf{x}_t]'(\bar{\gamma}; \mathbf{W}) = [\mathbf{z}_t \circ \bar{\mathbf{x}}]'(\bar{\gamma}; \mathbf{W}).$$

Since the chain rule for LD-derivatives implies that

$$[\mathbf{z}_t \circ \bar{\mathbf{x}}]'(\bar{\gamma}; \mathbf{W}) = [\mathbf{z}_t]'(\bar{\mathbf{x}}(\bar{\gamma}); \bar{\mathbf{x}}'(\bar{\gamma}; \mathbf{W})) = [\mathbf{z}_t]'(\bar{\boldsymbol{\xi}}; \mathbf{M} - \mathbf{f}(\bar{\boldsymbol{\xi}})\, \mathbf{v}^{\mathrm{T}}),$$

a second application of Theorem 5.2.4 shows that the mapping $t \mapsto [\mathbf{x}_t]'(\bar{\gamma}; \mathbf{W})$ is

the unique solution $\mathbf{A}$ of (8.17). $\square$

The following assumption extends Assumption 8.5.1 to describe an ODE with a parameter-dependent initial condition and initial independent variable. The subsequent corollary describes directional derivatives for the corresponding ODE solution with respect to the parameter.

Assumption 8.5.4. *Suppose that Assumption 8.5.1 holds, and let $P \subset \mathbb{R}^{n_p}$ be open. Suppose that functions $\tau : P \to \mathbb{R}$ and $\boldsymbol{\xi} : P \to X$ are locally Lipschitz continuous. With $\mathbf{x}(\cdot, \mathbf{p})$ denoting any solution of the parametric ODE system:*

$$\frac{d\mathbf{x}}{dt}(t, \mathbf{p}) = \mathbf{f}(\mathbf{x}(t, \mathbf{p})), \qquad \mathbf{x}(\tau(\mathbf{p}), \mathbf{p}) = \boldsymbol{\xi}(\mathbf{p}),$$

suppose that there exists a unique solution $\{\mathbf{x}(t, \mathbf{p}) : t \in [\tau(\mathbf{p}), \bar{\tau}_f]\} \subset X$ for some $\bar{\tau}_f \in \mathbb{R}$ and each $\mathbf{p} \in P$ in a particular neighborhood of some $\bar{\mathbf{p}} \in P$. Define $\bar{\tau} := \tau(\bar{\mathbf{p}})$ and $\bar{\boldsymbol{\xi}} := \boldsymbol{\xi}(\bar{\mathbf{p}})$. Suppose that τ and $\boldsymbol{\xi}$ are each directionally differentiable at $\bar{\mathbf{p}}$.

Corollary 8.5.5. *Suppose that Assumption 8.5.4 holds. For each $t \in [\bar{\tau}, \bar{\tau}_f]$, the mapping $\mathbf{x}_t \equiv \mathbf{x}(t, \cdot)$ is well-defined and Lipschitz continuous on a neighborhood of $\bar{\mathbf{p}}$, with a Lipschitz constant that is independent of t. Moreover, $\mathbf{x}_t$ is directionally differentiable at $\bar{\mathbf{p}}$ for each $t \in [\bar{\tau}, \bar{\tau}_f]$; for any $\mathbf{d} \in \mathbb{R}^{n_p}$, the mapping $t \mapsto [\mathbf{x}_t]'(\bar{\mathbf{p}}; \mathbf{d})$ is the unique solution $\mathbf{y}$ on $[\bar{\tau}, \bar{\tau}_f]$ of the ODE:*

$$\frac{d\mathbf{y}}{dt}(t) = \mathbf{f}'(\mathbf{x}(t, \bar{\mathbf{p}}); \mathbf{y}(t)), \qquad \mathbf{y}(\bar{\tau}) = \boldsymbol{\xi}'(\bar{\mathbf{p}}; \mathbf{d}) - \mathbf{f}(\bar{\boldsymbol{\xi}})\, \tau'(\bar{\mathbf{p}}; \mathbf{d}).$$

If, in addition, τ and $\boldsymbol{\xi}$ are L-smooth at $\bar{\mathbf{p}}$, and $\mathbf{f}$ is L-smooth on X, then the mapping $\mathbf{x}_t \equiv \mathbf{x}(t, \cdot)$ is L-smooth at $\bar{\mathbf{p}}$ for each $t \in [\bar{\tau}, \bar{\tau}_f]$; for any $\mathbf{M} \in \mathbb{R}^{n_p \times p}$, the mapping $t \mapsto [\mathbf{x}_t]'(\bar{\mathbf{p}}; \mathbf{M})$ is the unique solution $\mathbf{A}$ on $[\bar{\tau}, \bar{\tau}_f]$ of the ODE:

$$\frac{d\mathbf{A}}{dt}(t) = \mathbf{f}'(\mathbf{x}(t, \bar{\mathbf{p}}); \mathbf{A}(t)), \qquad \mathbf{A}(\bar{\tau}) = \boldsymbol{\xi}'(\bar{\mathbf{p}}; \mathbf{M}) - \mathbf{f}(\bar{\boldsymbol{\xi}})\, \tau'(\bar{\mathbf{p}}; \mathbf{M}).$$

Proof. Choose any $\mathbf{d} \in \mathbb{R}^{n_p}$, consider the following auxiliary ODE system, with parameters $a \in \mathbb{R}$ and $\mathbf{c} \in \mathbb{R}^n$:

$$\frac{d\mathbf{z}}{dt}(t, a, \mathbf{c}) = \mathbf{f}(\mathbf{z}(t, a, \mathbf{c})), \qquad \mathbf{z}(a, a, \mathbf{c}) = \mathbf{c},$$

and define the mappings $\mathbf{z}_t \equiv \mathbf{z}(t,\cdot,\cdot)$ and $\boldsymbol{\gamma} : \mathbf{p} \mapsto (\tau(\mathbf{p}), \boldsymbol{\xi}(\mathbf{p}))$. By Lemma 8.5.3, for each $t \in [\bar{\tau}, \bar{\tau}_f]$, $\mathbf{z}_t$ is well-defined and Lipschitz continuous on a neighborhood of $\boldsymbol{\gamma}(\bar{\mathbf{p}})$, with a Lipschitz constant that is independent of t. Furthermore, $\mathbf{z}_t$ is directionally differentiable at $\boldsymbol{\gamma}(\bar{\mathbf{p}})$. Since $\mathbf{x}_t \equiv \mathbf{z}_t \circ \boldsymbol{\gamma}$, and since $\boldsymbol{\gamma}$ is locally Lipschitz continuous, it follows that $\mathbf{x}_t$ is well-defined and Lipschitz continuous on a neighborhood of $\bar{\mathbf{p}}$, with a Lipschitz constant that is independent of t. Moreover, using the chain rule for directional derivatives, it follows that $\mathbf{x}_t$ is directionally differentiable at $\bar{\mathbf{p}}$, with

$$[\mathbf{x}_t]'(\bar{\mathbf{p}};\mathbf{d}) = [\mathbf{z}_t]'\big(\boldsymbol{\gamma}(\bar{\mathbf{p}});\boldsymbol{\gamma}'(\bar{\mathbf{p}};\mathbf{d})\big) = [\mathbf{z}_t]'\left((\bar{\tau},\bar{\boldsymbol{\xi}});\begin{bmatrix}\tau'(\bar{\mathbf{p}};\mathbf{d})\\ \boldsymbol{\xi}'(\bar{\mathbf{p}};\mathbf{d})\end{bmatrix}\right).$$

Applying Lemma 8.5.3 to the right-hand side of the above equation yields the required ODE expression for $t \mapsto [\mathbf{x}_t]'(\bar{\mathbf{p}};\mathbf{d})$.

The case in which τ, $\boldsymbol{\xi}$, and $\mathbf{f}$ are L-smooth is analogous, with an arbitrary $\mathbf{M} \in \mathbb{R}^{n_p \times p}$ replacing $\mathbf{d}$ in the above argument, and with Lemma 8.5.3 now used to demonstrate the L-smoothness of $\mathbf{z}_t$ at $\boldsymbol{\gamma}(\bar{\mathbf{p}})$. □

The following lemma effectively decouples LD-derivatives of an ODE solution $\mathbf{x}$ into a component reflecting dependence of $\mathbf{x}$ on t, and a component reflecting dependence of $\mathbf{x}$ on its initial condition. The latter component can be evaluated using Corollary 8.5.5.

Lemma 8.5.6. *Suppose that Assumption 8.5.4 holds, and let* $\mathbf{x}_t \equiv \mathbf{x}(t,\cdot)$. *For any* $t \in [\bar{\tau}, \bar{\tau}_f]$, $\mathbf{x}$ *is well-defined and Lipschitz continuous on a neighborhood of* $(t,\bar{\mathbf{p}})$ *and is directionally differentiable at* $(t,\bar{\mathbf{p}})$*; for any* $\alpha \in \mathbb{R}$ *and* $\mathbf{d} \in \mathbb{R}^{n_p}$,

$$\mathbf{x}'((t,\bar{\mathbf{p}});(\alpha,\mathbf{d})) = \alpha\mathbf{f}(\mathbf{x}(t,\bar{\mathbf{p}})) + [\mathbf{x}_t]'(\bar{\mathbf{p}};\mathbf{d}). \tag{8.19}$$

If, in addition, τ *and* $\boldsymbol{\xi}$ *are L-smooth at* $\bar{\mathbf{p}}$*, and* $\mathbf{f}$ *is L-smooth on* X*, then* $\mathbf{x}$ *is L-smooth at* $(t,\bar{\mathbf{p}})$ *for each* $t \in [\bar{\tau}, \bar{\tau}_f]$*; for any* $\mathbf{v} \in \mathbb{R}^p$ *and* $\mathbf{M} \in \mathbb{R}^{n_p \times p}$,

$$\mathbf{x}'\Big((t,\bar{\mathbf{p}});(\mathbf{v}^{\mathrm{T}},\mathbf{M})\Big) = \mathbf{f}(\mathbf{x}(t,\bar{\mathbf{p}}))\,\mathbf{v}^{\mathrm{T}} + [\mathbf{x}_t]'(\bar{\mathbf{p}};\mathbf{M}).$$

Proof. With γ defined as in the proof of Corollary 8.5.5, using the local Lipschitz continuity of γ on P, [18, Theorem 7.1] yields the existence of some $a < \bar{\tau}$, some $b > \bar{\tau}_f$, and some neighborhood $\bar{N}$ of $\bar{\mathbf{p}}$ such that $\mathbf{x}$ is well-defined on $(a, b) \times \bar{N}$. Choosing $\hat{\tau}, \hat{\tau}_f$ such that $a < \hat{\tau} < \bar{\tau}$ and $b > \hat{\tau}_f > \bar{\tau}_f$, Corollary 8.5.5 implies that $\mathbf{x}_t$ is Lipschitz continuous on some neighborhood $\hat{N}$ of $\bar{\mathbf{p}}$ for any fixed $t \in [\hat{\tau}, \hat{\tau}_f]$, with a Lipschitz constant $\bar{k}_{\mathbf{x}}$ that is independent of t. Thus, for any $(t_1, \mathbf{p}_1), (t_2, \mathbf{p}_2) \in [\hat{\tau}, \hat{\tau}_f] \times (\bar{N} \cap \hat{N})$,

$$\begin{aligned}
&\|\mathbf{x}(t_1, \mathbf{p}_1) - \mathbf{x}(t_2, \mathbf{p}_2)\| \\
&\quad \leq \|\mathbf{x}_{t_1}(\mathbf{p}_1) - \mathbf{x}_{t_1}(\mathbf{p}_2)\| + \|\mathbf{x}(t_1, \mathbf{p}_2) - \mathbf{x}(t_2, \mathbf{p}_2)\| \\
&\quad \leq \bar{k}_{\mathbf{x}} \|\mathbf{p}_1 - \mathbf{p}_2\| + \left\| \int_{t_2}^{t_1} \mathbf{f}(\mathbf{x}(s, \mathbf{p}_2))\, ds \right\| \\
&\quad \leq \bar{k}_{\mathbf{x}} \|\mathbf{p}_1 - \mathbf{p}_2\| + m_{\mathbf{f}} |t_1 - t_2| \\
&\quad \leq (\bar{k}_{\mathbf{x}} + m_{\mathbf{f}}) \|(t_1, \mathbf{p}_1) - (t_2, \mathbf{p}_2)\|_1,
\end{aligned}$$

and so $\mathbf{x}$ is Lipschitz continuous on $[\hat{\tau}, \hat{\tau}_f] \times (\bar{N} \cap \hat{N})$.

Now, choose any fixed $t \in [\bar{\tau}, \bar{\tau}_f]$. For sufficiently small $s > 0$, by the mean-value theorem, there exists $\theta_{s,i} \in [0, s\alpha]$ such that

$$\int_t^{t+s\alpha} f_i(\mathbf{x}(r, \bar{\mathbf{p}} + s\mathbf{d}))\, dr = s\alpha f_i(\mathbf{x}(t + \theta_{s,i}, \bar{\mathbf{p}} + s\mathbf{d})).$$

Since $f_i \circ \mathbf{x}$ is continuous, and since $\lim_{s \to 0^+} \theta_{s,i} = 0$, it follows that

$$\lim_{s \to 0^+} \frac{\int_t^{t+s\alpha} f_i(\mathbf{x}(r, \bar{\mathbf{p}} + s\mathbf{d}))\, dr}{s} = \alpha \lim_{s \to 0^+} f_i(\mathbf{x}(t + \theta_{s,i}, \bar{\mathbf{p}} + s\mathbf{d})) = \alpha f_i(\mathbf{x}(t, \bar{\mathbf{p}})).$$

Concatenating the above limits for all i,

$$\lim_{s \to 0^+} \frac{\int_t^{t+s\alpha} \mathbf{f}(\mathbf{x}(r, \bar{\mathbf{p}} + s\mathbf{d}))\, dr}{s} = \alpha \mathbf{f}(\mathbf{x}(t, \bar{\mathbf{p}})). \tag{8.20}$$

Now, for sufficiently small $s > 0$,

$$\frac{\mathbf{x}(t+s\alpha,\bar{\mathbf{p}}+s\mathbf{d})-\mathbf{x}(t,\bar{\mathbf{p}})}{s} = \frac{\mathbf{x}(t+s\alpha,\bar{\mathbf{p}}+s\mathbf{d})-\mathbf{x}(t,\bar{\mathbf{p}}+s\mathbf{d})}{s} + \frac{\mathbf{x}(t,\bar{\mathbf{p}}+s\mathbf{d})-\mathbf{x}(t,\bar{\mathbf{p}})}{s}$$
$$= \frac{\int_t^{t+s\alpha}\mathbf{f}(\mathbf{x}(r,\bar{\mathbf{p}}+s\mathbf{d}))\,dr}{s} + \frac{\mathbf{x}_t(\bar{\mathbf{p}}+s\mathbf{d})-\mathbf{x}_t(\bar{\mathbf{p}})}{s}.$$

Since Corollary 8.5.5 implies the directional differentiability of $\mathbf{x}_t$ at $\bar{\mathbf{p}}$, it follows from (8.20) and Corollary 8.5.5 that

$$\lim_{s\to 0^+}\frac{\mathbf{x}(t+s\alpha,\bar{\mathbf{p}}+s\mathbf{d})-\mathbf{x}(t,\bar{\mathbf{p}})}{s} = \alpha\mathbf{f}(\mathbf{x}(t,\bar{\mathbf{p}})) + [\mathbf{x}_t]'(\bar{\mathbf{p}};\mathbf{d}).$$

This in turn yields the directional differentiability of $\mathbf{x}$ at $(t,\bar{\mathbf{p}})$, with a directional derivative described by (8.19).

Suppose now that τ and $\boldsymbol{\xi}$ are L-smooth at $\bar{\mathbf{p}}$, and that $\mathbf{f}$ is L-smooth on X. Consider any $\mathbf{v} := (v_1,\ldots,v_p) \in \mathbb{R}^p$ and any $\mathbf{M} := \begin{bmatrix}\mathbf{m}_{(1)} & \cdots & \mathbf{m}_{(p)}\end{bmatrix}$. Define $\bar{\gamma} := (\tau(\bar{\mathbf{p}}),\boldsymbol{\xi}(\bar{\mathbf{p}})) \in \mathbb{R}^{n+1}$ and $\mathbf{W} := (\mathbf{v}^{\mathrm{T}},\mathbf{M}) \in \mathbb{R}^{(n_p+1)\times p}$. The following simple inductive argument shows that for each $k \in \{0,1,\ldots,p\}$, $\mathbf{x}^{(k)}_{\bar{\gamma},\mathbf{W}}$ exists, and that for each $\alpha \in \mathbb{R}$ and each $\mathbf{d} \in \mathbb{R}^{n_p}$,

$$\mathbf{x}^{(k)}_{\bar{\gamma},\mathbf{W}}(\alpha,\mathbf{d}) = \alpha\mathbf{f}(\mathbf{x}(t,\bar{\mathbf{p}})) + [\mathbf{x}_t]^{(k)}_{\bar{\mathbf{p}},\mathbf{M}}(\mathbf{d}). \tag{8.21}$$

The case in which $k = 0$ follows immediately from (8.19). For the inductive step, suppose the above statement is true for some particular $k \in \{0,1,\ldots,(p-1)\}$. Then, for any $s > 0$, any $\alpha \in \mathbb{R}$, and any $\mathbf{d} \in \mathbb{R}^{n_p}$,

$$\frac{\mathbf{x}^{(k)}_{\bar{\gamma},\mathbf{W}}(v_{k+1}+s\alpha,\mathbf{m}_{(k+1)}+s\mathbf{d})-\mathbf{x}^{(k)}_{\bar{\gamma},\mathbf{W}}(v_{k+1},\mathbf{m}_{(k+1)})}{s}$$
$$= \left(\frac{(v_{k+1}+s\alpha)-v_{k+1}}{s}\right)\mathbf{f}(\mathbf{x}(t,\bar{\mathbf{p}})) + \frac{[\mathbf{x}_t]^{(k)}_{\bar{\mathbf{p}},\mathbf{M}}(\mathbf{m}_{(k+1)}+s\mathbf{d})-[\mathbf{x}_t]^{(k)}_{\bar{\mathbf{p}},\mathbf{M}}(\mathbf{m}_{(k+1)})}{s}$$
$$= \alpha\mathbf{f}(\mathbf{x}(t,\bar{\mathbf{p}})) + \frac{[\mathbf{x}_t]^{(k)}_{\bar{\mathbf{p}},\mathbf{M}}(\mathbf{m}_{(k+1)}+s\mathbf{d})-[\mathbf{x}_t]^{(k)}_{\bar{\mathbf{p}},\mathbf{M}}(\mathbf{m}_{(k+1)})}{s}.$$

By Corollary 8.5.5, $\mathbf{x}_t$ is L-smooth at $\bar{\mathbf{p}}$. Thus,

$$\lim_{s\to 0^+} \frac{\mathbf{x}^{(k)}_{\bar{\gamma},\mathbf{W}}(v_{k+1}+s\alpha, \mathbf{m}_{(k+1)}+s\mathbf{d}) - \mathbf{x}^{(k)}_{\bar{\gamma},\mathbf{W}}(v_{k+1}, \mathbf{m}_{(k+1)})}{s}$$
$$= \alpha \mathbf{f}(\mathbf{x}(t,\bar{\mathbf{p}})) + [\mathbf{x}_t]^{(k+1)}_{\bar{\mathbf{p}},\mathbf{M}}(\mathbf{d}),$$

yielding the existence of $\mathbf{x}^{(k+1)}_{\bar{\gamma},\mathbf{W}}$ and its equivalence to the mapping

$$(\alpha, \mathbf{d}) \mapsto \alpha \mathbf{f}(\mathbf{x}(t,\bar{\mathbf{p}})) + [\mathbf{x}_t]^{(k+1)}_{\bar{\mathbf{p}},\mathbf{M}}(\mathbf{d}).$$

This completes the inductive assumption. Since $\mathbf{v}$ and $\mathbf{M}$ were chosen arbitrarily, this yields the L-smoothness of $\mathbf{x}$ at $(t, \bar{\mathbf{p}})$.

To complete the proof, (8.21) implies that with $\mathbf{v}$ and $\mathbf{M}$ chosen as in the inductive argument above, and with $\bar{\gamma}$ and $\mathbf{W}$ defined as above,

$$\begin{aligned}
&\mathbf{x}'((t,\bar{\mathbf{p}});\mathbf{W}) \\
&\quad = \left[\mathbf{x}^{(0)}_{\bar{\gamma},\mathbf{W}}(v_1, \mathbf{m}_{(1)}) \quad \cdots \quad \mathbf{x}^{(p-1)}_{\bar{\gamma},\mathbf{W}}(v_p, \mathbf{m}_{(p)})\right] \\
&\quad = \left[\left(v_1 \mathbf{f}(\mathbf{x}(t,\bar{\mathbf{p}})) + [\mathbf{x}_t]^{(0)}_{\bar{\mathbf{p}},\mathbf{M}}(\mathbf{m}_{(1)})\right) \quad \cdots \quad \left(v_p \mathbf{f}(\mathbf{x}(t,\bar{\mathbf{p}})) + [\mathbf{x}_t]^{(p-1)}_{\bar{\mathbf{p}},\mathbf{M}}(\mathbf{m}_{(p)})\right)\right] \\
&\quad = \mathbf{f}(\mathbf{x}(t,\bar{\mathbf{p}}))\,\mathbf{v}^{\mathrm{T}} + \left[[\mathbf{x}_t]^{(0)}_{\bar{\mathbf{p}},\mathbf{M}}(\mathbf{m}_{(1)}) \quad \cdots \quad [\mathbf{x}_t]^{(p-1)}_{\bar{\mathbf{p}},\mathbf{M}}(\mathbf{m}_{(p)})\right] \\
&\quad = \mathbf{f}(\mathbf{x}(t,\bar{\mathbf{p}}))\,\mathbf{v}^{\mathrm{T}} + [\mathbf{x}_t]'(\bar{\mathbf{p}};\mathbf{M}),
\end{aligned}$$

as required. □

The following corollary describes parametric LD-derivatives of an *event*, defined as the earliest value of the independent variable t for which a *transition condition* is satisfied. This result is, essentially, an analog of [30, Equation 50] that permits the functions involved to be L-smooth rather than $\mathcal{C}^1$.

Corollary 8.5.7. *Suppose that Assumption 8.5.4 holds, suppose a function $g : X \to \mathbb{R}$ is L-smooth, and suppose that a function t_e is implicitly defined on some open subset of P containing $\bar{\mathbf{p}}$ such that for each applicable $\mathbf{p} \in P$, $t_e(\mathbf{p})$ is the least element t^* of $(\tau(\mathbf{p}), \bar{\tau}_f]$ for which*

$$0 = g(\mathbf{x}(t^*, \mathbf{p})).$$

Suppose that τ and $\boldsymbol{\xi}$ are L-smooth at $\bar{\mathbf{p}}$, and $\mathbf{f}$ is L-smooth on X. Let $\tau^* := t_e(\bar{\mathbf{p}})$ and $\mathbf{x}^* := \mathbf{x}(\tau^*, \bar{\mathbf{p}})$, and suppose the composite mapping $g \circ \mathbf{x}$ satisfies the conditions of Theorem 8.3.2 at $(\tau^*, \bar{\mathbf{p}})$. Then t_e is L-smooth at $\bar{\mathbf{p}}$; for any $\mathbf{M} \in \mathbb{R}^{n_p \times p}$, $[t_e]'(\bar{\mathbf{p}}; \mathbf{M})$ is the unique solution $\mathbf{v}^{\mathrm{T}} \in \mathbb{R}^{1 \times p}$ of:

$$\mathbf{0}_{1\times p} = g'\Big(\mathbf{x}^*; \mathbf{f}(\mathbf{x}^*)\,\mathbf{v}^{\mathrm{T}} + [\mathbf{x}_{\tau^*}]'(\bar{\mathbf{p}}; \mathbf{M})\Big)\,. \tag{8.22}$$

If, in addition, g is differentiable at $\mathbf{x}^*$, then

$$[t_e]'(\bar{\mathbf{p}}; \mathbf{M}) = -\frac{(\nabla g(\mathbf{x}^*))^{\mathrm{T}}\,[\mathbf{x}_{\tau^*}]'(\bar{\mathbf{p}}; \mathbf{M})}{(\nabla g(\mathbf{x}^*))^{\mathrm{T}}\,\mathbf{f}(\mathbf{x}^*)}. \tag{8.23}$$

Proof. Choose any $\mathbf{M} \in \mathbb{R}^{n_p \times p}$. By Lemma 8.5.6, $\mathbf{x}$ is L-smooth in some neighborhood of $(\tau^*, \bar{\mathbf{p}})$, and so the mapping $g \circ \mathbf{x}$ is as well. By Theorem 8.3.2, the implicit function t_e is L-smooth in some neighborhood of $\bar{\mathbf{p}}$, and $[t_e]'(\bar{\mathbf{p}}; \mathbf{M})$ is the unique solution $\mathbf{v}^{\mathrm{T}}$ of

$$\mathbf{0}_{1\times p} = [g \circ \mathbf{x}]'\Big((\tau^*, \bar{\mathbf{p}}); (\mathbf{v}^{\mathrm{T}}, \mathbf{M})\Big)\,.$$

Applying the chain rule for LD-derivatives, the above equation is equivalent to:

$$\mathbf{0}_{1\times p} = g'\Big(\mathbf{x}^*; \mathbf{x}'\Big((\tau^*, \bar{\mathbf{p}}); (\mathbf{v}^{\mathrm{T}}, \mathbf{M})\Big)\Big)\,.$$

Lemma 8.5.6 implies that this equation is in turn equivalent to (8.22), which therefore has the unique solution $\mathbf{v}^{\mathrm{T}} = [t_e]'(\bar{\mathbf{p}}; \mathbf{M})$.

If g is differentiable at $\mathbf{x}^*$, then (8.22) is equivalent to

$$\begin{aligned}\mathbf{0}_{1\times p} &= (\nabla g(\mathbf{x}^*))^{\mathrm{T}}\Big(\mathbf{f}(\mathbf{x}^*)\,\mathbf{v}^{\mathrm{T}} + [\mathbf{x}_{\tau^*}]'(\bar{\mathbf{p}}; \mathbf{M})\Big)\,,\\ &= (\nabla g(\mathbf{x}^*))^{\mathrm{T}}\,\mathbf{f}(\mathbf{x}^*)\,\mathbf{v}^{\mathrm{T}} + (\nabla g(\mathbf{x}^*))^{\mathrm{T}}\,[\mathbf{x}_{\tau^*}]'(\bar{\mathbf{p}}; \mathbf{M}).\end{aligned} \tag{8.24}$$

Suppose, to obtain a contradiction, that $(\nabla g(\mathbf{x}^*))^{\mathrm{T}}\,\mathbf{f}(\mathbf{x}^*) = 0$. In this case, the right-hand side of (8.24) is independent of $\mathbf{v}$, and so (8.24) has either no solutions or infinitely many solutions $\mathbf{v}^{\mathrm{T}}$. However, since g is differentiable at $\mathbf{x}^*$, (8.24) is equivalent to (8.22), which has a unique solution $\mathbf{v}^{\mathrm{T}} = [t_e]'(\bar{\mathbf{p}}; \mathbf{M})$. This yields a contradiction, and thereby shows that $(\nabla g(\mathbf{x}^*))^{\mathrm{T}}\,\mathbf{f}(\mathbf{x}^*) \neq 0$ under the assumptions

of this corollary. Noting that $(\nabla g(\mathbf{x}^*))^{\mathrm{T}}\,\mathbf{f}(\mathbf{x}^*)$ is a scalar, (8.24) is readily solved for $\mathbf{v}^{\mathrm{T}} = [t_e]'(\bar{\mathbf{p}};\mathbf{M})$ to yield (8.23). $\square$

Although the above corollary nominally requires $t_e(\mathbf{p})$ to be strictly greater than $\tau(\mathbf{p})$ for each $\mathbf{p} \in P$, the proof of the corollary implies that the result remains valid if $t_e(\bar{\mathbf{p}}) = \bar{\tau}$, provided that $t_e(\mathbf{p}) \geq \tau(\mathbf{p})$ for each $\mathbf{p} \in P$, and provided that t_e is still a well-defined L-smooth implicit function near $\bar{\mathbf{p}}$.

8.6 Examples

The examples in this section illustrate hybrid systems in which, for various reasons, classical sensitivity analysis approaches [30] cannot be used to compute parametric derivatives for the state variables, or even to confirm that these derivatives exist. Theorem 8.4.2, however, permits evaluation of parametric LD-derivatives for the system state variables in each case. In each of these examples, the requirement in Assumption 8.4.1 that each composition $g_{(i)} \circ \mathbf{x}_{(i)}$ satisfies the conditions of Theorem 8.3.2 at $(\tau^*_{(i+1)}, \bar{\mathbf{p}})$ can be verified to hold by direct computation.

Example 8.6.1. *This example applies Theorem 8.4.2 to a hybrid system with a nondifferentiable discontinuity function. Consider an instance of the parametric hybrid discrete/continuous system described in Assumption 8.4.1, with $n_m := 2$ discrete modes, $n_{(i)} := 2$ state variables for each mode $i \in \{1,2\}$, and $n_p := 2$ parameters, and with the following function definitions:*

$$
\begin{aligned}
&\boldsymbol{\xi}_{(1)} : \mathbf{p} \in \mathbb{R}^2 \mapsto \mathbf{p}, \\
&\tau_{(1)} : \mathbf{p} \in \mathbb{R}^2 \mapsto 0, \\
&\mathbf{f}_{(i)} : \mathbf{z} \in \mathbb{R}^2 \mapsto \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \mathbf{z} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad \forall i \in \{1,2\}, \\
&g_{(1)} : \mathbf{z} \in \mathbb{R}^2 \mapsto z_1 - e^2 + |z_2 - 2|, \\
&\boldsymbol{\theta}_{(2)} : \mathbf{z} \in \mathbb{R}^2 \mapsto \tfrac{1}{2}\mathbf{z}.
\end{aligned}
$$

Observe that $\mathbf{p}$ is simply the initial condition for the hybrid system; consider the particular

initial condition $\bar{\mathbf{p}} := (1,0)$. *Set* $\bar{\tau}_f := 3$, *and choose* $P \subset \mathbb{R}^2$ *to be a sufficiently small neighborhood of* $\bar{\mathbf{p}}$ *to meet the existence and uniqueness conditions of Assumption 8.4.1. By inspection, this hybrid system has the following solution for* $\mathbf{p} := \bar{\mathbf{p}}$*:*

$$\begin{aligned}
\mathbf{x}_{(1)}(t,(1,0)) &= (e^t, t), && \forall t \in [0,3],\\
\tau^*_{(2)} &:= \tau_{(2)}(1,0) = 2,\\
\mathbf{x}^*_{(1)} &:= \mathbf{x}_{(1)}(2,(1,0)) = (e^2, 2),\\
\mathbf{x}_{(2)}(t,(1,0)) &= (\tfrac{1}{2}e^{t+2}, t-1), && \forall t \geq 2.
\end{aligned}$$

Observe that $g_{(1)}$ *is nondifferentiable at* $\mathbf{x}^\star_{(1)}$*; it follows that established sensitivity theory [30] for hybrid systems is not applicable to this system at* $\mathbf{p} := (1,0)$. *Nevertheless, the approach of this chapter applies. As in [33], define a* first-sign *function* fsign *as follows:*

$$\operatorname{fsign} : \mathbb{R}^p \to \{-1,0,+1\} : \mathbf{z} \mapsto \begin{cases} 0, & \text{if } \mathbf{z} = \mathbf{0},\\ \operatorname{sign} z_{k^*}, \text{ with } k^* := \min\{k : z_k \neq 0\}, & \text{if } \mathbf{z} \neq \mathbf{0}. \end{cases}$$

The following LD-derivatives are then readily computed, with the absolute-value function in $g_{(1)}$ *handled as in Chapter 4. Given any matrix* $\mathbf{Q}$, *let* $q_{i,j}$ *denote the* (i,j)*-entry of* $\mathbf{Q}$. *The following expressions hold for each* $\mathbf{z} \in \mathbb{R}^2$ *and* $\mathbf{N} \in \mathbb{R}^{2\times p}$.

$$\begin{aligned}
[f_{(1)}]'(\mathbf{z};\mathbf{N}) &= [f_{(2)}]'(\mathbf{z};\mathbf{N}) = \begin{bmatrix} 1 & 0\\ 0 & 0 \end{bmatrix}\mathbf{N},\\
[g_{(1)}]'(\mathbf{x}^*_{(1)};\mathbf{N}) &= \begin{cases} \begin{bmatrix} 1 & -1 \end{bmatrix}\mathbf{N}, & \text{if } \operatorname{fsign}(n_{2,1},\ldots,n_{2,p}) \leq 0,\\ \begin{bmatrix} 1 & 1 \end{bmatrix}\mathbf{N}, & \text{if } \operatorname{fsign}(n_{2,1},\ldots,n_{2,p}) > 0, \end{cases}\\
[\boldsymbol{\theta}_{(2)}]'(\mathbf{z};\mathbf{N}) &= \tfrac{1}{2}\mathbf{N}.
\end{aligned} \tag{8.25}$$

Thus, Theorem 8.4.2 implies that, for $j = 1$ *and* $\tilde{t} \in (0,2]$, *and for* $j = 2$ *and* $\tilde{t} \in (2,3]$, *for any* $\mathbf{M} \in \mathbb{R}^{2\times p}$, *the LD-derivative* $[\mathbf{x}_{(j),\tilde{t}}]'((1,0);\mathbf{M})$ *is the matrix* $\mathbf{A}_{(j)}(\tilde{t})$ *described as follows. Here,* $\mathbf{v} \in \mathbb{R}^p$ *is the unique solution of the equation system* (8.26) *below.*

$$
\begin{aligned}
\mathbf{A}_{(1)}(t) &= \begin{bmatrix} e^t & 0 \\ 0 & 1 \end{bmatrix} \mathbf{M}, && \forall t \in [0,2], \\
\mathbf{0}_{1\times p} &= [g_{(1)}]' \left(\begin{bmatrix} e^2 \\ 2 \end{bmatrix}; \begin{bmatrix} e^2 \\ 1 \end{bmatrix} \mathbf{v}^{\mathrm{T}} + \begin{bmatrix} e^2 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{M} \right), && (8.26) \\
\mathbf{A}_{(2)}(2) &= \frac{1}{2} \left(\begin{bmatrix} e^2 & 0 \\ 0 & 2 \end{bmatrix} \mathbf{M} - \begin{bmatrix} e^2 \\ 1 \end{bmatrix} \mathbf{v}^{\mathrm{T}} \right), \\
\mathbf{A}_{(2)}(t) &= \begin{bmatrix} e^{t-2} & 0 \\ 0 & 1 \end{bmatrix} \mathbf{A}_{(2)}(2), && \forall t > 2.
\end{aligned}
$$

The implicitly defined quantity $\mathbf{v}$ *can be evaluated analytically in this case. Define an intermediate matrix quantity*

$$
\mathbf{W} := \begin{bmatrix} e^2 \\ 1 \end{bmatrix} \mathbf{v}^{\mathrm{T}} + \begin{bmatrix} e^2 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{M} \in \mathbb{R}^{2\times p},
$$

and define $\sigma^* := \mathrm{fsign}(w_{2,1}, \ldots, w_{2,p})$. *Equations* (8.25) *and* (8.26) *imply that, if* $\sigma^* \leq 0$, *then* $\begin{bmatrix} 1 & -1 \end{bmatrix} \mathbf{W} = \mathbf{0}_{1\times p}$. *Otherwise, if* $\sigma^* > 0$, *then* $\begin{bmatrix} 1 & 1 \end{bmatrix} \mathbf{W} = \mathbf{0}_{1\times p}$. *Since these cases are exhaustive, the definition of* $\mathbf{W}$ *implies that*

$$
\mathbf{v}^{\mathrm{T}} = \begin{cases} \frac{1}{e^2-1} \begin{bmatrix} -e^2 & 1 \end{bmatrix} \mathbf{M}, & \text{if } \sigma^* \leq 0, \\ \frac{-1}{e^2+1} \begin{bmatrix} e^2 & 1 \end{bmatrix} \mathbf{M}, & \text{if } \sigma^* > 0. \end{cases} \tag{8.27}
$$

Substituting this relationship into the definition of $\mathbf{W}$ *yields:*

$$
w_{2,k} \in \left\{ \frac{e^2(m_{2,k} - m_{1,k})}{e^2 - 1}, \frac{e^2(m_{2,k} - m_{1,k})}{e^2 + 1} \right\}, \qquad \forall k \in \{1, \ldots, p\}.
$$

Noting that both coefficients of $(m_{2,k} - m_{1,k})$ *in the above expression are positive, it follows immediately that* $\sigma^* = \mathrm{fsign}\left((\begin{bmatrix} -1 & 1 \end{bmatrix} \mathbf{M})^{\mathrm{T}} \right)$; $\mathbf{v}$ *is then obtained by substituting this expression into* (8.27).

Example 8.6.2. *This example applies Theorem 8.4.2 to a hybrid system that visits non-differentiable domain points of its ODE right-hand side function for a nonzero duration. Consider an instance of the parametric hybrid discrete/continuous system described in Assumption 8.4.1, with* $n_m := 2$ *discrete modes,* $n_{(i)} := 2$ *state variables for each mode* $i \in \{1, 2\}$, *and* $n_p := 2$ *parameters, and with the following function definitions:*

$$
\begin{aligned}
\boldsymbol{\xi}_{(1)} &: \mathbf{p} \in \mathbb{R}^2 \mapsto \mathbf{p}, \\
\tau_{(1)} &: \mathbf{p} \in \mathbb{R}^2 \mapsto 0, \\
\mathbf{f}_{(1)} &: \mathbf{z} \in \mathbb{R}^2 \mapsto (|z_1|, 1), \\
g_{(1)} &: \mathbf{z} \in \mathbb{R}^2 \mapsto z_1 + z_2 - 2, \\
\boldsymbol{\theta}_{(2)} &: \mathbf{z} \in \mathbb{R}^2 \mapsto \mathbf{z}, \\
\mathbf{f}_{(2)} &: \mathbf{z} \in \mathbb{R}^2 \mapsto (z_1, 1).
\end{aligned}
$$

Again, $\mathbf{p}$ *is simply the initial condition for the hybrid system; consider the particular initial condition* $\bar{\mathbf{p}} := (0,0)$*. Set* $\bar{\tau}_f := 3$*, and choose* $P \subset \mathbb{R}^2$ *to be a sufficiently small neighborhood of* $\bar{\mathbf{p}}$ *to meet the existence and uniqueness conditions of Assumption 8.4.1. By inspection, this hybrid system has the following solution for* $\mathbf{p} := \bar{\mathbf{p}}$*:*

$$
\begin{aligned}
\mathbf{x}_{(1)}(t,(0,0)) &= (0,t), && \forall t \in [0,3], \\
\tau^*_{(2)} &:= \tau_{(2)}(0,0) = 2, \\
\mathbf{x}^*_{(1)} &:= \mathbf{x}_{(1)}(2,(0,0)) = (0,2), \\
\mathbf{x}_{(2)}(t,(0,0)) &= (0,t), && \forall t \geq 2.
\end{aligned}
$$

Indeed, for any choice of $a \in [0,2)$*, the hybrid system has the following solution for* $\mathbf{p} := (ae^{a-2}, 0)$*:*

$$
\begin{aligned}
\mathbf{x}_{(1)}(t,(ae^{a-2},0)) &= (ae^{t+a-2},t), && \forall t \in [0,3-a], \\
\tau_{(2)}(ae^{a-2},0) &= 2-a, \\
\mathbf{x}_{(2)}(t,(ae^{a-2},0)) &= (ae^{t+a-2},t), && \forall t \geq 2-a,
\end{aligned}
$$

and has the following solution for $\mathbf{p} := (-ae^{a+2}, 0)$*:*

$$
\begin{aligned}
\mathbf{x}_{(1)}(t,(-ae^{a+2},0)) &= (-ae^{a+2-t},t), && \forall t \in [0,a+3], \\
\tau_{(2)}(-ae^{a+2},0) &= a+2, \\
\mathbf{x}_{(2)}(t,(-ae^{a+2},0)) &= (-ae^{t-a-2},t), && \forall t \geq a+2.
\end{aligned}
$$

The above solution trajectories are illustrated in Figure 8-1 for various values of $a \in [0,2)$*.*

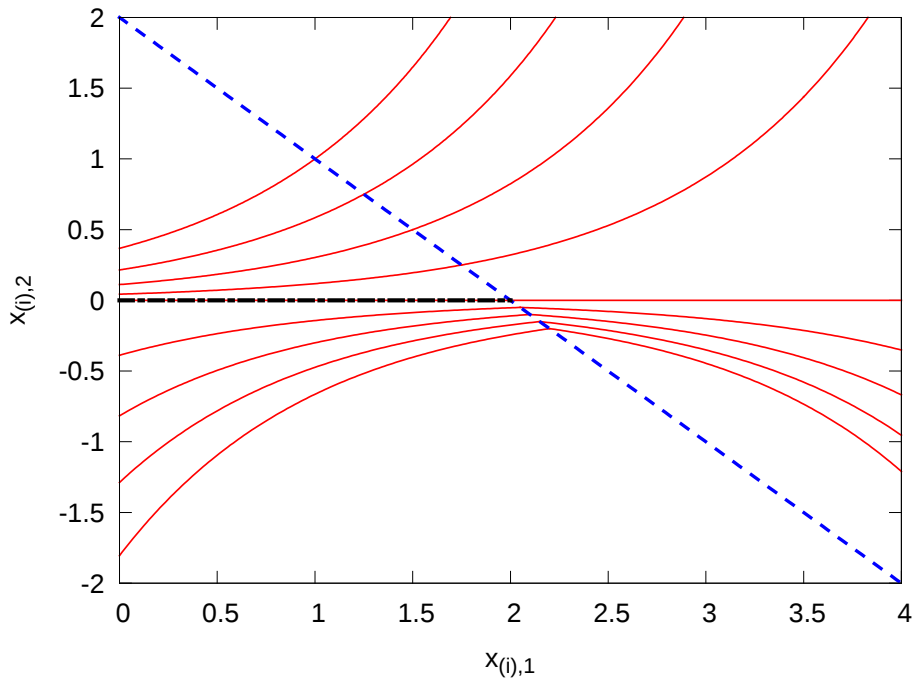

Figure 8-1: Solution trajectories (solid red) for the hybrid system considered in Example 8.6.2, for various choices of $a \in [0,2)$ and $\mathbf{p} \in \{(ae^{a-2},0),(-ae^{a+2},0)\}$, the set $\{\mathbf{z} \in \mathbb{R}^2 : g_{(1)}(\mathbf{z}) = 0\}$ on which a discrete event occurs (dashed blue), and a subset of $\mathbb{R}^2$ on which $\mathbf{f}_{(1)}$ is nondifferentiable (dash-dotted black).

For each $t \in [0,2]$, $\mathbf{f}_{(1)}$ is nondifferentiable at $\mathbf{x}_{(1)}(t,(0,0))$; the sensitivity theory of [30] is therefore not applicable in this case. Observe that $\mathbf{f}_{(1)}(\mathbf{x}^*_{(1)}) = (0,1)$, that $\boldsymbol{\theta}_{(2)}$ is the identity mapping, and that

$$\mathbf{f}_{(2)}(\mathbf{x}_{(2)}(2,(0,0))) = \mathbf{J}\boldsymbol{\theta}_{(2)}(\mathbf{x}^*_{(1)})\,\mathbf{f}_{(1)}(\mathbf{x}^*_{(1)});$$

the final claim of Theorem 8.4.2 is therefore applicable with $i := 1$. Defining the first-sign function as in Example 8.6.1, and handling the absolute-value function as in Chapter 4, Equation (8.10) *in Theorem 8.4.2 becomes the following ODE, when $i := 1$. Here, "$a_{1,k}(t)$" refers to the $(1,k)$-element of $\mathbf{A}_{(1)}(t)$.*

$$\frac{d\mathbf{A}_{(1)}}{dt}(t) = \begin{bmatrix} \mathrm{fsign}(x_{(1),1}(t,\bar{\mathbf{p}}), a_{1,1}(t), \ldots, a_{1,p}(t)) & 0 \\ 0 & 0 \end{bmatrix} \mathbf{A}_{(1)}(t).$$

Thus, the auxiliary hybrid system in $\mathbf{A}_{(j)}$ presented in Theorem 8.4.2 is readily solved by inspection. For $j = 1$ and $\tilde{t} \in (0,2]$, and for $j = 2$ and $\tilde{t} \in (2,3]$, for any $\mathbf{M} \in \mathbb{R}^{2\times p}$,

the LD-derivative $[\mathbf{x}_{(j),\tilde{t}}]'(\bar{\mathbf{p}};\mathbf{M})$ *is the matrix* $\mathbf{A}_{(j)}(\tilde{t})$ *described as follows.*

$$
\begin{aligned}
\sigma &:= \operatorname{fsign}\left(\left(\begin{bmatrix}1 & 0\end{bmatrix}\mathbf{M}\right)^{\mathrm{T}}\right), \\
\mathbf{A}_{(1)}(t) &= \begin{bmatrix} e^{\sigma t} & 0 \\ 0 & 1 \end{bmatrix}\mathbf{M}, && \forall t \in [0,2], \\
\mathbf{A}_{(2)}(2) &= \mathbf{A}_{(1)}(2) = \begin{bmatrix} e^{2\sigma} & 0 \\ 0 & 1 \end{bmatrix}\mathbf{M}, \\
\mathbf{A}_{(2)}(t) &= \begin{bmatrix} e^{t-2} & 0 \\ 0 & 1 \end{bmatrix}\mathbf{A}_{(2)}(2) = \begin{bmatrix} e^{t+2\sigma-2} & 0 \\ 0 & 1 \end{bmatrix}\mathbf{M}, && \forall t > 2.
\end{aligned}
$$

If $\mathbf{x}_{(1)}(t,\cdot)$ *were differentiable at* $\bar{\mathbf{p}}$ *for some fixed* $t \in (0,2]$*, then the expression for* $\mathbf{A}_{(1)}(t)$ *above would vary linearly with* $\mathbf{M}$*. However, the intermediate quantity* σ *varies nonlinearly with* $\mathbf{M}$*; by choosing* $\mathbf{M}$ *appropriately, the parameter* σ *can be made to take any particular value in the set* $\{-1,0,+1\}$*. Inspection of the above expression for* $\mathbf{A}_{(1)}$ *then shows that* $\mathbf{x}_{(1)}(t,\cdot)$ *is nondifferentiable at* $\bar{\mathbf{p}}$ *for each* $t \in (0,2]$*. A similar argument shows that* $\mathbf{x}_{(2)}(t,\cdot)$ *is nondifferentiable at* $\bar{\mathbf{p}}$ *for each* $t \in [2,3]$*.*

Example 8.6.3. *This example examines a hybrid system in which small perturbations in the system parameters can change the discrete mode sequence visited by the solution trajectory. This hybrid system is reformulated so as to satisfy Assumption 8.4.1, whereupon parametric LD-derivatives are obtained for its solution trajectory using Theorem 8.4.2.*

Consider functions:

$$
\begin{aligned}
\mathbf{f}_I &: \mathbf{z} \in \mathbb{R}^2 \mapsto \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\mathbf{z} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \\
\mathbf{f}_A &: \mathbf{z} \in \mathbb{R}^2 \mapsto 2\mathbf{f}_I(\mathbf{z}), \\
\mathbf{f}_B &: \mathbf{z} \in \mathbb{R}^2 \mapsto \tfrac{1}{2}\mathbf{f}_I(\mathbf{z}), \\
\mathbf{f}_F &: \mathbf{z} \in \mathbb{R}^2 \mapsto \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}\mathbf{z},
\end{aligned}
$$

and

$$
g_{\pm} : \mathbf{z} \in \mathbb{R}^2 \mapsto z_1 - e^2 \pm (z_2 - 2),
$$

and consider the following hybrid discrete/continuous system with four discrete modes,

indexed by I, A, B, F, and with a parameter $\mathbf{p} \in \mathbb{R}^2$ *chosen from a sufficiently small neighborhood of* $\bar{\mathbf{p}} := (1, 0)$. *The system's state variables are initialized in mode I with* $\mathbf{x}_I(0, \mathbf{p}) = \mathbf{p}$. *When in any mode* $J \in \{I, A, B, F\}$, $\mathbf{x}(\cdot, \mathbf{p}) \equiv \mathbf{x}_J(\cdot, \mathbf{p})$ *evolves according to the ODE:*

$$\frac{d\mathbf{x}_J}{dt}(t, \mathbf{p}) = \mathbf{f}_J(\mathbf{x}_J(t, \mathbf{p})).$$

The system does not satisfy Assumption 8.4.1 directly. At each transition between discrete modes, there is no jump in the system's state variables. The discrete mode is changed from mode I at the least value of $\tau_I > 0$ *for which*

$$g_+(\mathbf{x}_I(\tau_I, \mathbf{p})) = 0 \qquad OR \qquad g_-(\mathbf{x}_I(\tau_I, \mathbf{p})) = 0.$$

If $g_+(\mathbf{x}_I(\tau_I, \mathbf{p})) = 0$, *then the discrete mode changes from I to A at* τ_I; *otherwise, the discrete mode changes from I to B at* τ_I.

Once the system is in discrete mode A, the mode is changed from A to F at the least value of $\tau_A \geq \tau_I$ *for which*

$$g_-(\mathbf{x}_A(\tau_A, \mathbf{p})) = 0.$$

Once the system is in discrete mode B, the mode is changed from B to F at the least value of $\tau_B \geq \tau_I$ *for which*

$$g_+(\mathbf{x}_B(\tau_B, \mathbf{p})) = 0.$$

Once the system enters mode F, there are no further changes to the discrete mode. The event times τ_I, τ_A, *and* τ_B *depend on* $\mathbf{p}$ *whenever they exist, and will thus be denoted as functions of* $\mathbf{p}$. *The discrete mode structure of this hybrid system is illustrated in Figure 8-2(a). This hybrid system is, essentially, an ODE with a discontinuous right-hand side function, as considered in [26].*

At $\mathbf{p} := \bar{\mathbf{p}} = (1, 0)$, *it is readily verified that the hybrid system above has the following unique solution trajectory:*

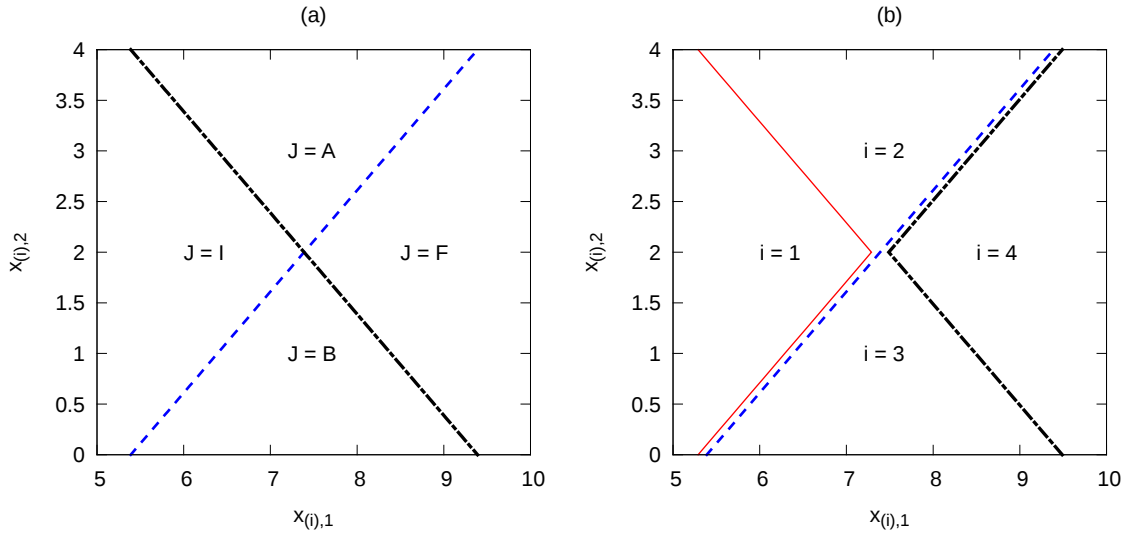

Figure 8-2: Discrete modes for the hybrid system considered in Example 8.6.3: (a) the relationship between the continuous state $\mathbf{x}_J(t, \mathbf{p})$ and discrete mode $J \in \{I, A, B, F\}$ for the original formulation, assuming that $\mathbf{p}$ is sufficiently close to $(1, 0)$, the set $\{\mathbf{z} \in \mathbb{R}^2 : g_+(\mathbf{z}) = 0\}$ (dash-dotted black), and the set $\{\mathbf{z} \in \mathbb{R}^2 : g_-(\mathbf{z}) = 0\}$ (dashed blue), and (b) the relationship between the continuous state $\mathbf{x}_{(i)}(t, \mathbf{p})$ and discrete mode $i \in \{1, 2, 3, 4\}$ for the modified formulation, the set $\{\mathbf{z} \in \mathbb{R}^2 : g_{(1)}(\mathbf{z}) = 0\}$ (solid red), the set $\{\mathbf{z} \in \mathbb{R}^2 : g_{(2)}(\mathbf{z}) = 0\}$ (dashed blue), and the set $\{\mathbf{z} \in \mathbb{R}^2 : g_{(3)}(\mathbf{z}) = 0\}$ (dash-dotted black).

$$
\begin{aligned}
\mathbf{x}_I(t, (1,0)) &= (e^t, t), && \forall t \in [0, 2], \\
\tau_I(1,0) &= 2, \\
\mathbf{x}_A(2, (1,0)) &= \mathbf{x}_I(2, (1,0)) = (e^2, 2), \\
\tau_A(1,0) &= 2, \\
\mathbf{x}_F(2, (1,0)) &= \mathbf{x}_A(2, (1,0)) = (e^2, 2), \\
\mathbf{x}_F(t, (1,0)) &= (e^{2t-2}, 2e^{4-2t}), && \forall t > 2.
\end{aligned}
$$

Observe that this trajectory visits the discrete modes in the order $I \to A \to F$, and that both $g_+(\mathbf{x}_I(\tau_I(1,0), (1,0)) = 0$ and $g_-(\mathbf{x}_I(\tau_I(1,0), (1,0)) = 0$. It is readily verified that, for all sufficiently small $\epsilon > 0$, the solution trajectory with $\mathbf{p} := \bar{\mathbf{p}} + (\epsilon, 0) = (1 + \epsilon, 0)$ instead visits the discrete modes in the order $I \to B \to F$. Since small changes in parameters may change the sequence of visited discrete modes, conventional sensitivity analysis theory [30] cannot describe sensitivities of the solution trajectory for this particu-

lar system at $\mathbf{p} := \bar{\mathbf{p}}$.

However, as noted in Section 8.4, if Assumption 8.4.1 is relaxed to permit $\tau_{(i+1)}(\mathbf{p}) = \tau_{(i)}(\mathbf{p})$*, then this assumption's admittance of nondifferentiable discontinuity functions* $g_{(i)}$ *can be exploited to provide the following alternative formulation for the above hybrid system. Unlike the original formulation, this reformulation exhibits a discrete mode sequence that is independent of* $\mathbf{p}$*, provided that* $\mathbf{p}$ *is chosen from some sufficiently small neighborhood of* $\bar{\mathbf{p}}$.

$$
\begin{aligned}
&\boldsymbol{\xi}_{(1)} : \mathbf{p} \in \mathbb{R}^2 \mapsto \mathbf{p}, \\
&\tau_{(1)} : \mathbf{p} \in \mathbb{R}^2 \mapsto 0, \\
&\mathbf{f}_{(1)} : \mathbf{z} \in \mathbb{R}^2 \mapsto \mathbf{f}_I(\mathbf{z}), \\
&g_{(1)} : \mathbf{z} \in \mathbb{R}^2 \mapsto z_1 - e^2 + |z_2 - 2|, \\
&\boldsymbol{\theta}_{(i)} : \mathbf{z} \in \mathbb{R}^2 \mapsto \mathbf{z}, && \forall i \in \{2,3,4\}, \\
&\mathbf{f}_{(2)} : \mathbf{z} \in \mathbb{R}^2 \mapsto \mathbf{f}_A(\mathbf{z}), \\
&g_{(2)} : \mathbf{z} \in \mathbb{R}^2 \mapsto z_1 - e^2 - (z_2 - 2), \\
&\mathbf{f}_{(3)} : \mathbf{z} \in \mathbb{R}^2 \mapsto \mathbf{f}_B(\mathbf{z}), \\
&g_{(3)} : \mathbf{z} \in \mathbb{R}^2 \mapsto z_1 - e^2 - |z_2 - 2|, \\
&\mathbf{f}_{(4)} : \mathbf{z} \in \mathbb{R}^2 \mapsto \mathbf{f}_F(\mathbf{z}).
\end{aligned}
$$

The above reformulation may be handled using Theorem 8.4.2; the discrete mode structure for this reformulation is illustrated in Figure 8-2(b). In this figure, the level sets $\{\mathbf{z} \in \mathbb{R}^2 : g_{(i)}(\mathbf{z}) = 0\}$ *have been translated by small amounts to illustrate the invariance of the mode sequence* $i : 1 \to 2 \to 3 \to 4$*; in fact* $\{\mathbf{z} \in \mathbb{R}^2 : g_{(2)}(\mathbf{z}) = 0\} \subset \{\mathbf{z} \in \mathbb{R}^2 : g_{(1)}(\mathbf{z}) = 0\} \cup \{\mathbf{z} \in \mathbb{R}^2 : g_{(3)}(\mathbf{z}) = 0\}$.

Observe that this reformulation has the following solution trajectory at $\bar{\mathbf{p}}$*, which is analogous to the trajectory obtained for the original formulation above.*

$$
\begin{aligned}
\mathbf{x}_{(1)}(t,(1,0)) &= (e^t, t), && \forall t \in [0,3],\\
\tau^*_{(2)} &:= \tau_{(2)}(1,0) = 2,\\
\mathbf{x}^*_{(1)} &:= \mathbf{x}_{(1)}(2,(1,0)) = (e^2, 2),\\
\mathbf{x}_{(2)}(2,(1,0)) &= (e^2, 2),\\
\tau^*_{(3)} &:= \tau_{(3)}(1,0) = 2,\\
\mathbf{x}^*_{(2)} &:= \mathbf{x}_{(2)}(2,(1,0)) = (e^2, 2),\\
\mathbf{x}_{(3)}(2,(1,0)) &= (e^2, 2),\\
\tau^*_{(4)} &:= \tau_{(4)}(1,0) = 2,\\
\mathbf{x}^*_{(3)} &:= \mathbf{x}_{(3)}(2,(1,0)) = (e^2, 2)\\
\mathbf{x}_{(4)}(t,(1,0)) &= (e^{2t-2}, 2e^{4-2t}), && \forall t \geq 2.
\end{aligned}
$$

According to Theorem 8.4.2, for each $t \in (0,2]$, $\mathbf{x}_{(1)}(t,\cdot)$ is L-smooth at $\bar{\mathbf{p}}$. Similarly, for each $t > 2$, $\mathbf{x}_{(4)}(t,\cdot)$ is L-smooth at $\bar{\mathbf{p}}$. Since L-smoothness implies local Lipschitz continuity, the state variables of this hybrid system are thereby shown to be locally Lipschitz continuous at $\bar{\mathbf{p}}$ for each fixed $t \neq 2$, even in the original formulation.

Defining the first-sign function as in Example 8.6.1, the following expressions hold for each $\mathbf{z} \in \mathbb{R}^2$ and $\mathbf{N} \in \mathbb{R}^{2\times p}$:

$$
\begin{aligned}
[f_{(1)}]'(\mathbf{z};\mathbf{N}) &= [f_{(4)}]'(\mathbf{z};\mathbf{N}) = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \mathbf{N},\\
[f_{(2)}]'(\mathbf{z};\mathbf{N}) &= \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \mathbf{N},\\
[f_{(3)}]'(\mathbf{z};\mathbf{N}) &= \begin{bmatrix} 0.5 & 0 \\ 0 & 0 \end{bmatrix} \mathbf{N},\\
[g_{(1)}]'(\mathbf{x}^*_{(1)};\mathbf{N}) &= \begin{cases} \begin{bmatrix} 1 & -1 \end{bmatrix} \mathbf{N}, & \textit{if } \operatorname{fsign}(n_{2,1},\ldots,n_{2,p}) \leq 0,\\ \begin{bmatrix} 1 & 1 \end{bmatrix} \mathbf{N}, & \textit{if } \operatorname{fsign}(n_{2,1},\ldots,n_{2,p}) > 0, \end{cases}\\
[g_{(2)}]'(\mathbf{x}^*_{(2)};\mathbf{N}) &= (\nabla g_{(2)}(\mathbf{x}^*_{(2)}))^{\mathrm{T}}\mathbf{N} = \begin{bmatrix} 1 & -1 \end{bmatrix} \mathbf{N},\\
[g_{(3)}]'(\mathbf{x}^*_{(3)};\mathbf{N}) &= \begin{cases} \begin{bmatrix} 1 & -1 \end{bmatrix} \mathbf{N}, & \textit{if } \operatorname{fsign}(n_{2,1},\ldots,n_{2,p}) \geq 0,\\ \begin{bmatrix} 1 & 1 \end{bmatrix} \mathbf{N}, & \textit{if } \operatorname{fsign}(n_{2,1},\ldots,n_{2,p}) < 0, \end{cases}\\
[\boldsymbol{\theta}_{(i)}]'(\mathbf{z};\mathbf{N}) &= \mathbf{N}, && \forall i \in \{2,3,4\}.
\end{aligned}
\tag{8.28}
$$

Thus, Theorem 8.4.2 implies that, for $j = 1$ and $\tilde{t} \in (0,2]$, and for $j = 4$ and $\tilde{t} > 2$, for

any $\mathbf{M} \in \mathbb{R}^{2\times p}$*, the LD-derivative* $[\mathbf{x}_{(j),\tilde{t}}]'(\bar{\mathbf{p}};\mathbf{M})$ *is the matrix* $\mathbf{A}_{(j)(\tilde{t})}$ *described as follows. Here,* $\mathbf{v}$*,* $\mathbf{u}$*, and* $\mathbf{w}$ *are the unique solutions of* (8.29)*,* (8.30)*, and* (8.31) *below. Though the matrices* $\mathbf{A}_{(2)}(2)$ *and* $\mathbf{A}_{(3)}(2)$ *below must be computed in order to compute* $\mathbf{A}_{(4)}(t)$*, they do not represent LD-derivatives for the hybrid system.*

$$
\begin{aligned}
\mathbf{A}_{(1)}(t) &= \begin{bmatrix} e^t & 0 \\ 0 & 1 \end{bmatrix} \mathbf{M}, && \forall t \in (0,2], \\
\mathbf{0}_{1\times p} &= [g_{(1)}]' \left(\begin{bmatrix} e^2 \\ 2 \end{bmatrix} ; \begin{bmatrix} e^2 \\ 1 \end{bmatrix} \mathbf{v}^{\mathrm{T}} + \begin{bmatrix} e^2 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{M} \right), && (8.29) \\
\mathbf{A}_{(2)}(2) &= \mathbf{A}_{(1)}(2) + \left(\begin{bmatrix} e^2 \\ 1 \end{bmatrix} - \begin{bmatrix} 2e^2 \\ 2 \end{bmatrix} \right) \mathbf{v}^{\mathrm{T}} = \begin{bmatrix} e^2 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{M} - \begin{bmatrix} e^2 \\ 1 \end{bmatrix} \mathbf{v}^{\mathrm{T}}, \\
\mathbf{u}^{\mathrm{T}} &= -\frac{\begin{bmatrix} 1 & -1 \end{bmatrix} \mathbf{A}_{(2)}(2)}{\begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} 2e^2 \\ 2 \end{bmatrix}}, && (8.30) \\
\mathbf{A}_{(3)}(2) &= \mathbf{A}_{(2)}(2) - \left(\begin{bmatrix} \frac{1}{2}e^2 \\ \frac{1}{2} \end{bmatrix} - \begin{bmatrix} 2e^2 \\ 2 \end{bmatrix} \right) \mathbf{u}^{\mathrm{T}}, \\
\mathbf{0}_{1\times p} &= [g_{(3)}]' \left(\begin{bmatrix} e^2 \\ 2 \end{bmatrix} ; \begin{bmatrix} \frac{1}{2}e^2 \\ \frac{1}{2} \end{bmatrix} \mathbf{w}^{\mathrm{T}} + \mathbf{A}_{(3)}(2) \right), && (8.31) \\
\mathbf{A}_{(4)}(2) &= \mathbf{A}_{(3)}(2) + \left(\begin{bmatrix} \frac{1}{2}e^2 \\ \frac{1}{2} \end{bmatrix} - \begin{bmatrix} e^2 \\ 1 \end{bmatrix} \right) \mathbf{w}^{\mathrm{T}}. \\
\mathbf{A}_{(4)}(t) &= \begin{bmatrix} e^{2t-4} & 0 \\ 0 & e^{4-2t} \end{bmatrix} \mathbf{A}_{(4)}(2), && \forall t > 2.
\end{aligned}
$$

As in Example 8.6.1, construct $\sigma^* := \mathrm{fsign}((\begin{bmatrix} -1 & 1 \end{bmatrix} \mathbf{M})^{\mathrm{T}}) \in \{-1, 0, +1\}$*. With this construction,* (8.27) *remains applicable; thus,*

$$
\mathbf{v}^{\mathrm{T}} = \begin{cases} \frac{1}{e^2-1} \begin{bmatrix} -e^2 & 1 \end{bmatrix} \mathbf{M}, & \text{if } \sigma^* \leq 0, \\ \frac{-1}{e^2+1} \begin{bmatrix} e^2 & 1 \end{bmatrix} \mathbf{M}, & \text{if } \sigma^* > 0. \end{cases}
$$

At this point, determination of the intermediate vector quantity $\mathbf{w}$ *is the only obstacle to evaluating* $\mathbf{A}_{(4)}(t)$ *by direct computation using the above expressions. Following a similar approach to the determination of* $\mathbf{v}$ *in Example 8.6.1, the equation system* (8.31) *may be solved for* $\mathbf{w}$ *analytically, as follows. For notational convenience, define* $\mathbf{B} := \mathbf{A}_{(3)}(2) \in \mathbb{R}^{2\times p}$*,*

$$\mathbf{Q} := \frac{1}{2}\begin{bmatrix} e^2 \\ 1 \end{bmatrix} \mathbf{w}^{\mathrm{T}} + \mathbf{B} \in \mathbb{R}^{2\times p},$$

and $\zeta^* := \mathrm{fsign}(q_{2,1}, \ldots, q_{2,p})$. *Equations* (8.28) *and* (8.31) *imply that, if* $\zeta^* \geq 0$, *then* $\begin{bmatrix} 1 & -1 \end{bmatrix} \mathbf{Q} = \mathbf{0}_{1\times p}$. *Otherwise, if* $\zeta^* < 0$, *then* $\begin{bmatrix} 1 & 1 \end{bmatrix} \mathbf{Q} = \mathbf{0}_{1\times p}$. *Since these cases are exhaustive, the definitions of* $\mathbf{Q}$ *and* $\mathbf{B}$ *imply that*

$$\mathbf{w}^{\mathrm{T}} = \begin{cases} \frac{-2}{e^2-1}\begin{bmatrix} 1 & -1 \end{bmatrix} \mathbf{A}_{(3)}(2), & \text{if } \zeta^* \geq 0, \\ \frac{-2}{e^2+1}\begin{bmatrix} 1 & 1 \end{bmatrix} \mathbf{A}_{(3)}(2), & \text{if } \zeta^* < 0. \end{cases} \tag{8.32}$$

Substituting this description into the definition of $\mathbf{Q}$ *yields:*

$$q_{2,k} \in \left\{ \tfrac{1}{e^2-1}(e^2 b_{2,k} - b_{1,k}), \tfrac{1}{e^2+1}(e^2 b_{2,k} - b_{1,k}) \right\}, \qquad \forall k \in \{1, \ldots, p\}.$$

Noting that both coefficients of $(e^2 b_{2,k} - b_{1,k})$ *in the above expression are positive, it follows immediately that*

$$\zeta^* = \mathrm{fsign}\left(\left(\begin{bmatrix} -1 & e^2 \end{bmatrix} \mathbf{B} \right)^{\mathrm{T}} \right) = \mathrm{fsign}\left(\left(\begin{bmatrix} -1 & e^2 \end{bmatrix} \mathbf{A}_{(3)}(2) \right)^{\mathrm{T}} \right);$$

$\mathbf{w}$ *is then obtained by substituting this expression for* ζ^* *into* (8.32). *With* $\mathbf{w}$ *thus determined,* $\mathbf{A}_{(4)}(t)$ *can be evaluated for any* $t \geq 2$ *using the expressions above.*

In this example, there were no jumps in state variables. Suppose now that the original formulation is modified to permit jumps in state variables at its discrete events, with these jumps governed by a function $\boldsymbol{\theta}_+$ *at the* $I \to A$ *and* $B \to F$ *transitions, and by a function* $\boldsymbol{\theta}_-$ *at the* $I \to B$ *and* $A \to F$ *transitions. To reflect the behavior of this modified system, the transition functions in the reformulated system would, by inspection of Figure 8-2, need to satisfy the following conditions:*

$$\boldsymbol{\theta}_{(2)} \equiv \boldsymbol{\theta}_+, \qquad \boldsymbol{\theta}_{(4)} \equiv \boldsymbol{\theta}_+, \qquad \text{and} \qquad \boldsymbol{\theta}_{(3)} \circ \boldsymbol{\theta}_{(2)} \equiv \boldsymbol{\theta}_{(4)} \circ \boldsymbol{\theta}_{(3)} \equiv \boldsymbol{\theta}_-.$$

Thus, there must exist a function $\boldsymbol{\theta}_{(3)}$ *such that* $\boldsymbol{\theta}_{(3)} \circ \boldsymbol{\theta}_+ \equiv \boldsymbol{\theta}_+ \circ \boldsymbol{\theta}_{(3)} \equiv \boldsymbol{\theta}_-$. *If this condition cannot be satisfied, then the approach of this example will not be applicable. Observe that this condition is trivially satisfied by the identity transformation when the state variables in the original system do not jump; in this case,* $\boldsymbol{\theta}_+$ *and* $\boldsymbol{\theta}_-$ *are both the*

identity transformation.

8.7 Conclusions

Sufficient conditions have been presented for L-smoothness of local inverse functions and local implicit functions, and the corresponding LD-derivatives have been described as the unique solutions of the equation systems (8.1) and (8.4). In the special case that these functions are described in terms of piecewise differentiable functions, numerical methods have been provided for efficient evaluation of these LD-derivatives.

Using the above results, Theorem 8.4.2 provides parametric LD-derivatives for the broad class of hybrid discrete/continuous systems described by Assumption 8.4.1, in which nonsmoothness may be present in any or all of the functions determining the continuous evolution, discrete event timing, and state variable jumps of the hybrid system. This assumption requires the discrete mode sequence visited by the solution trajectory to be invariant under sufficiently small perturbations of the parameters. Nevertheless, Example 8.6.3 illustrates the possibility of reformulating certain hybrid systems violating this assumption, so that Assumption 8.4.1 and Theorem 8.4.2 apply to the equivalent reformulated system. Through this reformulation, the original formulation is demonstrated *a posteriori* to have state variables that are locally Lipschitz continuous with respect to the system parameters at each fixed time that is not a discrete event. This approach is applicable certain ODE systems with discontinuous right-hand side functions, in the form considered by Filippov [26].

Chapter 9

Twice-continuously differentiable convex relaxations of factorable functions

9.1 Introduction

As a departure from the earlier chapters in this thesis, this chapter is concerned with eliminating one particular source of nonsmoothness in optimization problems. Specifically, a variant of McCormick's scheme [74] for the generation of convex underestimators is developed. While the original scheme may produce nondifferentiable relaxations even when the relaxed function is smooth, this variant produces twice-continuously differentiable relaxations, without sacrificing any of the useful computational properties of McCormick's scheme. This chapter is reproduced from [60].

Branch-and-bound methods for global optimization [45] require the ability to evaluate a lower bound on a nonconvex function on particular classes of subdomains. This lower-bounding information may be generated using a relaxation scheme by McCormick [74], which evaluates convex underestimators of a nonconvex objective function on interval subdomains. McCormick's relaxation method

assumes that the objective function can be expressed as a finite, known composition of simple functions and arithmetic operations. Subgradients may be computed for these underestimators using dedicated variants [5, 76] of automatic differentiation [34]. Using this information, a lower bound on a nonconvex objective function on an interval may be supplied by minimizing the corresponding convex McCormick underestimator using a local optimization solver. Other methods for global optimization, such as nonconvex outer approximation [52] and nonconvex generalized Benders decomposition [68], also require the construction and minimization of convex underestimators.

McCormick's relaxation method has several useful properties: accurate evaluation of a convex underestimator and a corresponding subgradient is computationally inexpensive and automatable; the C++ library MC++ [15, 76] uses operator overloading to compute these quantities for well-defined user-supplied compositions of the basic arithmetic operations and functions such as sin / cos and exp / log. Moreover, as the width of the interval on which a McCormick relaxation is constructed is reduced to zero, the relaxation approaches the objective function sufficiently rapidly [11] to mitigate a phenomenon called the *cluster effect* [20, 119], in which a branch-and-bound method will branch many times on intervals that either contain or are near a global minimum. By extending McCormick's method in an intuitive manner, *generalized McCormick relaxations* [100, 104] have been developed to handle compositions of functions in a more systematic manner, and to handle various extensions of McCormick's theory to implicit functions [101, 107, 120].

However, as the following example shows, McCormick's relaxations can be nondifferentiable.

Example 9.1.1. *Let a function* $\mathrm{mid} : \mathbb{R}^3 \to \mathbb{R}$ *map to the median of its three scalar arguments, consider the smooth composite function* $g : \mathbb{R} \to \mathbb{R} : z \mapsto \exp(z^3)$*, and set* $z^* := -1 + \sqrt{3}$*. As shown in [76, Example 2.1], the function:*

$$g^{\mathrm{cv}} : [-1,1] \to \mathbb{R} : z \mapsto \exp(\mathrm{mid}(z^3 + 3z^2 - 3, z^3 - 3z^2 + 3, -1)),$$
$$= \begin{cases} \exp(-1), & \textit{if } z \leq z^*, \\ \exp(z^3 + 3z^2 - 3), & \textit{if } z > z^* \end{cases}$$

can be generated from g according to McCormick's rule [76, Section 3] for generating convex relaxations of a composite function, when αBB relaxations [1] of the inner function $z \mapsto z^3$ are employed. Indeed, g^{cv} is convex on $[-1,1]$, and $g^{\mathrm{cv}}(z) \leq g(z)$ for each $z \in [-1,1]$. However, even though g^{cv} satisfies McCormick's proposed sufficient condition for differentiability of a convex relaxation [74, p. 151], it is in fact nondifferentiable at z^*.

Several factors can introduce failure of continuous or twice-continuous differentiability of McCormick's relaxations. Firstly, as illustrated by the above example, the median function used in defining McCormick's composition rule is itself nondifferentiable. Secondly, any nondifferentiability in supplied relaxations of composed functions can propagate to yield nondifferentiability in constructed relaxations of composite functions. (Whether the composed functions are themselves smooth is irrelevant.) Thirdly, as presented in [76, Proposition 2.6], McCormick's rule for generating relaxations of products introduces nondifferentiability, due to its use of bivariate max and min functions. A relaxation scheme preserving continuous or twice-continuous differentiability would be desirable for a number of reasons, which are summarized in the following two paragraphs.

In general, nondifferentiable convex relaxations must be solved using dedicated numerical methods for nondifferentiable optimization problems such as bundle methods [63, 67], which lack the strong convergence rate results of their smooth counterparts. Continuously differentiable convex relaxations may be minimized using gradient-based algorithms for local optimization, which typically exhibit Q-linear convergence. Twice-continuously differentiable relaxations can be minimized by Newton's method, which exhibits Q-quadratic convergence under certain invertibility assumptions on the Hessian matrix. Computation of the required Hessian or Hessian-vector products can be avoided by using a secant-based quasi-Newton method instead, which exhibits Q-superlinear convergence under

the assumptions of Newton's method.

Furthermore, a method for generating continuously differentiable relaxations would yield theoretical and numerical benefits when used in established methods for generating convex and concave relaxations of solutions of parametric ordinary differential equations (ODEs). If continuously differentiable relaxations are available for the right-hand side function of such an ODE, then the relaxation-generating ODE described in [103] would have a continuously differentiable right-hand side function. Theoretical hurdles concerning evaluation of subgradients for these relaxations would thus be overcome, since the ODE solution relaxations would now be differentiable with respect to the ODE parameters. Moreover, the corresponding parametric derivatives may be computed according to classical ODE theory [35]. Similarly, incorporation of continuously differentiable relaxations of an ODE right-hand side function into the relaxation method of [102] would yield ODEs whose parametric sensitivities are decribed by the hybrid system sensitivity results of [30].

Thus, the goal of this chapter is to present a variant of McCormick's relaxation scheme which produces continuously or twice-continuously differentiable relaxations, while retaining the various theoretical and computational benefits of McCormick's original method. To achieve this, variants of McCormick's product rule are introduced in Definition 9.3.19, in which the original product rule is further relaxed in a particular manner. An additional assumption (Assumption 9.2.21) is imposed on user-supplied relaxations of composed *intrinsic functions*, so as to enforce differentiability in McCormick's composition rule. This assumption is readily satisfied for standard arithmetic operations and functions. Under these modifications, the aforementioned sources of nonsmoothness in McCormick's relaxation scheme are circumvented. For broader applicability, the relaxation theory developed in this chapter is presented in the framework of generalized McCormick relaxations [100]. To construct twice-continuously differentiable relaxations rather than once-continuously differentiable relaxations according to the methods in this chapter, more stringent (yet readily satisfied) assumptions are required on the sup-

plied relaxations of univariate intrinsic functions, and the employed product rule must be relaxed further. Gradients of the developed relaxations can be evaluated efficiently using the standard forward or reverse modes of automatic differentiation [34].

The product rule variants developed in this chapter make use of certain smoothing approximations. Smooth approximations of simple nonsmooth functions have previously been considered [8], particularly in the context of complementarity problems [23, 28, 90]. The smoothing approach taken in this chapter is similar in spirit, but is modified so as to accommodate our requirement that the posited convex/concave relaxations are well-defined, are indeed convex or concave, are valid bounds on the underlying function, and are either once- or twice-continuously differentiable, as desired.

Observe that the αBB relaxation scheme [1] represents an alternative to McCormick's scheme, and shares several of the features of McCormick's method outlined above. Moreover, αBB relaxations of twice-continuously differentiable functions are themselves twice-continuously differentiable. This chapter instead focuses on variations of McCormick's method, due to the ability of McCormick's theory to handle more general compositions of functions, and due to its extensions to relaxations of implicit functions, and to relaxations of solutions of differential-algebraic equations.

This chapter is structured as follows. Section 9.2 summarizes established definitions and properties concerning differentiability on intervals, interval analysis, and McCormick's relaxation technique. Section 9.3 develops the smoothing constructions used in the remainder of the chapter. Section 9.4 develops variants of McCormick's relaxation technique, and presents the main theorem of the chapter, in which these variants are asserted to have desirable properties. The proof of this theorem is spread over the next three sections: Section 9.5 shows that the proposed McCormick relaxation variants are indeed valid relaxations, Section 9.6 demonstrates continuous or twice-continuous differentiability of these relaxations and demonstrates a technique for propagating their gradients, and Section 9.7 shows

that as the underlying parameter interval is reduced in size, the relaxations converge to the original function sufficiently rapidly to mitigate clustering in a branch-and-bound scheme for global optimization. Section 9.8 describes a C++ implementation of the methods in this chapter, and presents examples of its application for illustration.

9.2 Background

This section summarizes relevant, established definitions and properties concerning differentiability on closed sets, interval analysis, McCormick's convex/concave relaxation scheme, and convergence orders of relaxation schemes.

9.2.1 Differentiability on open and closed sets

Let $\|\cdot\|$ denote the Euclidean norm on $\mathbb{R}^n$. Given an open set $X \subset \mathbb{R}^n$, a function $\mathbf{f} : X \to \mathbb{R}^m$ is *(Fréchet) differentiable* at $\mathbf{x} \in X$ if there exists a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ for which

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

In this case, the above equation defines $\mathbf{A}$ uniquely, and $\mathbf{A}$ is called the *Jacobian matrix* $\mathbf{J}\mathbf{f}(\mathbf{x})$ of $\mathbf{f}$ at $\mathbf{x}$. If $m = 1$, in which case $\mathbf{f} \equiv f$ is scalar-valued, then the *gradient* of f at $\mathbf{x}$ is the column vector $\nabla f(\mathbf{x}) := (\mathbf{J}f(\mathbf{x}))^{\mathrm{T}} \in \mathbb{R}^m$.

Given an open set $X \subset \mathbb{R}^n$, a function $\mathbf{f} : X \to \mathbb{R}^n$ is *continuously differentiable* ($\mathcal{C}^1$) on X if it is differentiable on X, and the Jacobian mapping $\mathbf{x} \mapsto \mathbf{J}\mathbf{f}(\mathbf{x})$ is continuous on X. Equivalently, $\mathbf{f}$ is $\mathcal{C}^1$ on X if its first-order partial derivatives each exist on X and are continuous. If $m = 1$, in which case $\mathbf{f} \equiv f$ is scalar-valued, then f is *twice-continuously differentiable* ($\mathcal{C}^2$) on X if f is $\mathcal{C}^1$ on X and there exists a continuous *Hessian* mapping $\mathbf{x} \to \nabla^2 f(\mathbf{x})$ for which

$$\mathbf{0} = \lim_{\mathbf{h}\to\mathbf{0}} \frac{f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^{\mathrm{T}}\,\mathbf{h} - \frac{1}{2}\mathbf{h}^{\mathrm{T}}\,\nabla^2 f(\mathbf{x})\,\mathbf{h}}{\|\mathbf{h}\|^2}, \qquad \forall \mathbf{x} \in X.$$

Equivalently, f is $\mathcal{C}^2$ on X if its second-order partial derivatives each exist on X and are continuous. A vector-valued function $\mathbf{f}$ is $\mathcal{C}^2$ if each of its component functions is $\mathcal{C}^2$.

By specializing a classical result by Whitney [121], differentiability on closed sets such as intervals can be defined in a manner that is consistent with the classical chain rule of differentiation, as follows.

Definition 9.2.1. *Given a closed set $B \subset \mathbb{R}^n$ and some $i \in \{1, 2\}$, a function $\mathbf{f} : B \to \mathbb{R}^m$ is $\mathcal{C}^i$ on B if there exists an open set $X \subset \mathbb{R}^n$ such that $B \subset X$, and a function $\hat{\mathbf{f}} : X \to \mathbb{R}^m$ such that $\hat{\mathbf{f}}(\mathbf{x}) = \mathbf{f}(\mathbf{x})$ for each $\mathbf{x} \in B$, and such that $\hat{\mathbf{f}}$ is $\mathcal{C}^i$ (in the classical sense) on X. Given any point $\mathbf{x}$ in the boundary of B, define $\mathbf{J}\mathbf{f}(\mathbf{x}) := \mathbf{J}\hat{\mathbf{f}}(\mathbf{x})$. If $m = 1$, in which case $\mathbf{f} \equiv f$ is scalar-valued, then define $\nabla f(\mathbf{x}) := \mathbf{J}f(\mathbf{x})^{\mathrm{T}}$.*

Remark 9.2.2. *When $\mathbf{x}$ lies in the boundary of B, it is possible that $\mathbf{J}\mathbf{f}(\mathbf{x})$ is not uniquely specified by the above definition, since $\hat{\mathbf{f}}$ might not be specified uniquely. For example, if B comprises a single point $\{\mathbf{x}_0\} \subset \mathbb{R}^n$, then $\hat{\mathbf{f}}$ may be chosen to be any $\mathcal{C}^i$ function for which $\hat{\mathbf{f}}(\mathbf{x}_0) = \mathbf{f}(\mathbf{x}_0)$.*

Despite the possible nonuniqueness implied by the previous remark, the following propositions show that the classical chain rule continues to hold.

Proposition 9.2.3. *Consider B, i, and $\mathbf{f}$ as in Definition 9.2.1, and any point $\mathbf{x}$ in the boundary of B. If there exists any sequence $\{\mathbf{x}_{(k)}\}_{k \in \mathbb{N}} \to \mathbf{x}$ in $B \backslash \{\mathbf{x}\}$, then any Jacobian $\mathbf{J}\mathbf{f}(\mathbf{x})$ satisfies*

$$\mathbf{0} = \lim_{\substack{\mathbf{h} \to \mathbf{0} \\ (\mathbf{x}+\mathbf{h}) \in B}} \frac{\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{J}\mathbf{f}(\mathbf{x})\,\mathbf{h}}{\|\mathbf{h}\|}.$$

Proof. This proposition is an immediate corollary of Theorem 1 in [121]. □

Corollary 9.2.4. *Given a closed convex set $B \subset \mathbb{R}^n$ and a convex $\mathcal{C}^1$ function $f : B \to \mathbb{R}$, for each $\mathbf{x} \in B$, $\nabla f(\mathbf{x})$ is a* subgradient *of f at $\mathbf{x}$ in that*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^{\mathrm{T}} (\mathbf{y} - \mathbf{x}), \qquad \forall \mathbf{y} \in B.$$

Proposition 9.2.5. *Consider nonempty sets $B \subset \mathbb{R}^n$ and $D \subset \mathbb{R}^m$ such that B is either closed, open, or both, and such that D is either closed, open, or both. For any fixed $i \in \{1, 2\}$, given $\mathcal{C}^i$ functions $\mathbf{g} : B \to D$ and $\mathbf{f} : D \to \mathbb{R}^p$, the composite function $\mathbf{h} \equiv \mathbf{f} \circ \mathbf{g} : B \to \mathbb{R}^p$ is well-defined and $\mathcal{C}^i$ on B.*

Moreover, for each $\mathbf{x} \in B$, $\mathbf{Jh}(\mathbf{x}) = \mathbf{Jf}(\mathbf{g}(\mathbf{x}))\,\mathbf{Jg}(\mathbf{x})$. (If B is closed and $\mathbf{x}$ lies in the boundary of B, then this construction of $\mathbf{Jh}(\mathbf{x})$ satisfies Definition 9.2.1 and Proposition 9.2.3 for some valid choice of $\hat{\mathbf{h}}$.)

Proof. This proposition is an immediate corollary of Theorem 1 in [121]. □

9.2.2 Interval analysis

This section presents a brief overview of relevant definitions and concepts from interval analysis; for further details, the reader is directed to introductory sources [2, 77, 80].

An *interval* $\boldsymbol{x} \equiv [\underline{x}, \overline{x}]$ is a nonempty compact set $\{z \in \mathbb{R} : \underline{x} \leq z \leq \overline{x}\} \subset \mathbb{R}$; the set of all such intervals is denoted $\mathbb{IR}$. Intervals and vectors of intervals are denoted in this chapter as boldfaced, italicized, lowercase letters (e.g., $\boldsymbol{y}$), whereas vectors in $\mathbb{R}^n$ are denoted as boldfaced, romanized, lowercase letters (e.g., $\mathbf{y}$). Given a set $B \subset \mathbb{R}^n$, the set of intervals (or vectors of intervals) that are subsets of B will be denoted as $\mathbb{I}B$. If B is nonempty, then $\mathbb{I}B$ is necessarily nonempty. An interval vector $\boldsymbol{y} \in \mathbb{IR}^n$ will be represented equivalently as $[\underline{\mathbf{y}}, \overline{\mathbf{y}}]$, where $\underline{\mathbf{y}} := (\underline{y}_1, \ldots, \underline{y}_n) \in \mathbb{R}^n$ and $\overline{\mathbf{y}} := (\overline{y}_1, \ldots, \overline{y}_n) \in \mathbb{R}^n$.

An interval $\boldsymbol{x} \in \mathbb{IR}$ has a *width* of $\operatorname{wid} \boldsymbol{x} := \overline{x} - \underline{x}$, and an interval vector $\boldsymbol{y} \equiv (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) \in \mathbb{IR}^n$ has a width of $\operatorname{wid} \boldsymbol{y} := \max_{k \in \{1,\ldots,n\}} \operatorname{wid} \boldsymbol{y}_k$. An interval or interval vector with zero width is *degenerate*, and is *nondegenerate* otherwise.

Definition 9.2.6 (from [2])**.** *For each $c \in \mathbb{R}$, define scalar-interval multiplication so that for each $\boldsymbol{x} \in \mathbb{IR}$,*

$$c\boldsymbol{x} := \begin{cases} [c\underline{x}, c\overline{x}], & \text{if } c \geq 0, \\ [c\overline{x}, c\underline{x}], & \text{if } c < 0. \end{cases}$$

Setting $c \leftarrow -1$ corresponds to a negative *operation. Define interval operations $+, -, \times : \mathbb{IR} \times \mathbb{IR} \to \mathbb{IR}$ such that*

$$
\begin{aligned}
+(\boldsymbol{x}, \boldsymbol{y}) &\equiv \boldsymbol{x} + \boldsymbol{y} := [\underline{x} + \underline{y}, \overline{x} + \overline{y}], && \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{IR}. \\
-(\boldsymbol{x}, \boldsymbol{y}) &\equiv \boldsymbol{x} - \boldsymbol{y} := [\underline{x} - \overline{y}, \overline{x} - \underline{y}], && \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{IR}. \\
\times(\boldsymbol{x}, \boldsymbol{y}) &\equiv \boldsymbol{x}\boldsymbol{y} := [\min\{\underline{x}\underline{y}, \underline{x}\overline{y}, \overline{x}\underline{y}, \overline{x}\overline{y}\}, \max\{\underline{x}\underline{y}, \underline{x}\overline{y}, \overline{x}\underline{y}, \overline{x}\overline{y}\}], && \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{IR}.
\end{aligned}
$$

It is readily verified that for any interval operation $\circ \in \{+, -, \times\}$, $\boldsymbol{x} \circ \boldsymbol{y} = \{v \circ w : v \in \boldsymbol{x}, w \in \boldsymbol{y}\}$ for any intervals $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{IR}$, and $[v, v] \circ [w, w] = [v \circ w, v \circ w]$ for any $v, w \in \mathbb{R}$.

Lemma 9.2.7. *Consider an interval $\boldsymbol{x} \in \mathbb{IR}$ and scalars $a, b \in \mathbb{R}$ for which $a \leq b$. If $\underline{x} \geq 0$, then $\underline{(a\boldsymbol{x})} \leq \underline{(b\boldsymbol{x})}$. If $\overline{x} \leq 0$, then $\underline{(a\boldsymbol{x})} \geq \underline{(b\boldsymbol{x})}$. Similarly, if $\underline{x} \geq 0$, then $\overline{(a\boldsymbol{x})} \leq \overline{(b\boldsymbol{x})}$. If $\overline{x} \leq 0$, then $\overline{(a\boldsymbol{x})} \geq \overline{(b\boldsymbol{x})}$.*

Proof. For any $c \in \mathbb{R}$,

$$
\underline{(c\boldsymbol{x})} = \begin{cases} c\underline{x}, & \text{if } c \geq 0, \\ c\overline{x}, & \text{if } c < 0 \end{cases} = \overline{x}\min\{c, 0\} + \underline{x}\max\{c, 0\}. \tag{9.1}
$$

If $\underline{x} \geq 0$, then each term in the final expression above is evidently increasing with respect to c, yielding the first required inequality. If, instead, $\overline{x} \leq 0$, then each term in the final expression is decreasing with respect to c, which yields the second required inequality. Next, for any $c \in \mathbb{R}$,

$$
\overline{(c\boldsymbol{x})} = \begin{cases} c\overline{x}, & \text{if } c \geq 0, \\ c\underline{x}, & \text{if } c < 0 \end{cases} = \underline{x}\min\{c, 0\} + \overline{x}\max\{c, 0\}. \tag{9.2}
$$

Using this result, a similar argument to the previous case yields the remaining inequalities. □

Definition 9.2.8 (from [77])**.** *Consider a nonempty set $B \subset \mathbb{R}^n$. An interval-valued function $\boldsymbol{f} : \mathbb{I}B \to \mathbb{IR}^m$ is* inclusion monotonic *if $\boldsymbol{f}(\boldsymbol{x}) \subset \boldsymbol{f}(\boldsymbol{y})$ for any pair $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{I}B$ for which $\boldsymbol{x} \subset \boldsymbol{y}$.*

Given a function $\mathbf{g} : B \to \mathbb{R}^m$, an interval-valued function $\tilde{\boldsymbol{g}} : \mathbb{I}B \to \mathbb{IR}^m$ is an interval extension *of $\mathbf{g}$ if $\tilde{\boldsymbol{g}}([\mathbf{x}, \mathbf{x}]) = [\mathbf{g}(\mathbf{x}), \mathbf{g}(\mathbf{x})]$ for each $\mathbf{x} \in B$.*

The following result from [77] motivates the above definition.

Theorem 9.2.9 (Theorem 3.1 in [77])**.** *Consider a function* $\mathbf{g} : B \subset \mathbb{R}^n \to \mathbb{R}^m$*. If a function* $\tilde{\boldsymbol{g}} : \mathbb{I}B \to \mathbb{I}\mathbb{R}^m$ *is inclusion monotonic and is an interval extension of* $\mathbf{g}$*, then* $\mathbf{g}(\boldsymbol{x}) := \{\mathbf{g}(\mathbf{z}) : \mathbf{z} \in \boldsymbol{x}\} \subset \tilde{\boldsymbol{g}}(\boldsymbol{x})$ *for all* $\boldsymbol{x} \in \mathbb{I}B$*.*

Definition 9.2.10 (adapted from [77])**.** *Consider a nonempty set* $B \subset \mathbb{R}^n$*. An interval function* $\boldsymbol{f} : \mathbb{I}B \to \mathbb{I}\mathbb{R}^m$ *is* locally Lipschitz continuous *if for each* $\boldsymbol{q} \in \mathbb{I}B$*, there exists* $k \geq 0$ *for which*

$$\operatorname{wid}(\boldsymbol{f}(\boldsymbol{x})) \leq k \operatorname{wid} \boldsymbol{x}, \qquad \forall \boldsymbol{x} \in \mathbb{I}\boldsymbol{q}.$$

A locally Lipschitz continuous, inclusion-monotonic interval extension of a function f will be called a *tight interval extension* of f.

Definition 9.2.11. *Given a set* $D \subset \mathbb{R}^n$*, define the* interval hull $\square D$ *of* D *as the intersection of all intervals in* $\mathbb{I}\mathbb{R}^n$ *that are supersets of* D*. Given a function* $\mathbf{f} : B \subset \mathbb{R}^n \to \mathbb{R}^m$*, define the interval hull* $\square\mathbf{f} : \mathbb{I}B \to \mathbb{I}\mathbb{R}^m$ *so that* $\square\mathbf{f}(\boldsymbol{x}) = \square\{\mathbf{y} \in \mathbb{R}^m : \exists \mathbf{z} \in \boldsymbol{x} \text{ s.t. } \mathbf{y} = \mathbf{f}(\mathbf{z})\}$*.*

Definition 9.2.12. *Given an open set* $B \subset \mathbb{R}$*, a function* $u : B \to \mathbb{R}$ *is* univariate intrinsic *if there exists a known tight interval extension* $\tilde{\boldsymbol{u}} : \mathbb{I}B \to \mathbb{I}\mathbb{R}$ *of* u*, and if, with* $\bar{B} := \{(\boldsymbol{x}, z) \in \mathbb{I}B \times B : z \in \boldsymbol{x}\}$*, there exist known functions* $u^{\mathrm{cv}}, u^{\mathrm{cc}} : \bar{B} \to \mathbb{R}$ *and* $\zeta_u^{\min}, \zeta_u^{\max} : \mathbb{I}B \to \mathbb{R}$ *such that:*

- *For each* $\boldsymbol{x} \in \mathbb{I}B$*,* $u^{\mathrm{cv}}(\boldsymbol{x}, \cdot)$ *is convex on* $\boldsymbol{x}$*,* $u^{\mathrm{cc}}(\boldsymbol{x}, \cdot)$ *is concave on* $\boldsymbol{x}$*, and* $u^{\mathrm{cv}}(\boldsymbol{x}, z) \leq u(z) \leq u^{\mathrm{cc}}(\boldsymbol{x}, z)$ *for each* $z \in \boldsymbol{x}$*.*

- *For each* $\boldsymbol{x} \in \mathbb{I}B$*,* $\zeta_u^{\min}(\boldsymbol{x}) \in \arg\min\{u^{\mathrm{cv}}(\boldsymbol{x}, z) : z \in \boldsymbol{x}\}$ *and* $\zeta_u^{\max}(\boldsymbol{x}) \in \arg\max\{u^{\mathrm{cc}}(\boldsymbol{x}, z) : z \in \boldsymbol{x}\}$*.*

- *For any* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{I}B$ *with* $\boldsymbol{x} \subset \boldsymbol{y}$*, and for any* $z \in \boldsymbol{x}$*,* $u^{\mathrm{cv}}(\boldsymbol{y}, z) \leq u^{\mathrm{cv}}(\boldsymbol{x}, z)$ *and* $u^{\mathrm{cc}}(\boldsymbol{y}, z) \geq u^{\mathrm{cc}}(\boldsymbol{x}, z)$*.*

- *For each* $z \in B$*,* $u^{\mathrm{cv}}([z, z], z) = u^{\mathrm{cc}}([z, z], z) = u(z)$*.*

For any $z \in \boldsymbol{x} \in \mathbb{I}B$, define

$$
\begin{aligned}
& u_I^{\mathrm{cv}}(\boldsymbol{x}, z) := u^{\mathrm{cv}}(\max\{z, \zeta_u^{\min}(\boldsymbol{x})\}), \\
& u_D^{\mathrm{cv}}(\boldsymbol{x}, z) := u^{\mathrm{cv}}(\min\{z, \zeta_u^{\min}(\boldsymbol{x})\}), \\
& u_I^{\mathrm{cc}}(\boldsymbol{x}, z) := u^{\mathrm{cc}}(\min\{z, \zeta_u^{\max}(\boldsymbol{x})\}), \\
\textit{and} \qquad & u_D^{\mathrm{cc}}(\boldsymbol{x}, z) := u^{\mathrm{cc}}(\max\{z, \zeta_u^{\max}(\boldsymbol{x})\}).
\end{aligned}
$$

The interval hull of a locally Lipschitz continuous function is clearly a tight interval extension of the function. The interval operations in Definition 9.2.6 are interval hulls of the corresponding operations on real numbers. Tight interval extensions are provided for a number of univariate intrinsic functions in Table 9.1; these interval extensions are also interval hulls.

Appropriate constructions of the functions u^{cv} and u^{cc} are also provided for these univariate intrinsic functions in Table 9.2. By inspection, these particular constructions satisfy the properties:

$$
\min_{z \in \boldsymbol{x}} u^{\mathrm{cv}}(\boldsymbol{x}, z) = \min_{z \in \boldsymbol{x}} u(z), \qquad \text{and} \qquad \max_{z \in \boldsymbol{x}} u^{\mathrm{cc}}(\boldsymbol{x}, z) = \max_{z \in \boldsymbol{x}} u(z);
$$

in general, a weaker version of these properties will be required in Assumption 9.2.21 below.

Definition 9.2.13. *Given a nonempty set $B \subset \mathbb{R}^n$, a function $\mathbf{f} : B \to \mathbb{R}^m$ is* MC-factorable *if each of the following conditions is satisfied:*

- $\mathbf{f}$ *can be expressed on B as a finite composition (in some order) of addition, multiplication, and univariate intrinsic functions with known tight interval extensions, and*
- *a well-defined* natural interval extension $\tilde{\boldsymbol{f}} : \mathbb{I}B \to \mathbb{I}\mathbb{R}^n$ *of* $\mathbf{f}$ *can be constructed by replacing each addition/multipication/univariate intrinsic function by its corresponding tight interval extension, without introducing any domain violations.*

The natural interval extension of an MC-factorable function is a tight interval extension of the function [77, Section 3.3].

Table 9.1: Tight interval extensions for various univariate intrinsic functions u.

B	$u(z)$ for $z \in B$	$\tilde{\boldsymbol{u}}(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathbb{I}B$
$\mathbb{R}$	cz for fixed $c \in \mathbb{R}$	$c\boldsymbol{x}$
$\mathbb{R}$	$\exp z$	$[\exp \underline{x}, \exp \overline{x}]$
$(0, +\infty)$	$\ln z$	$[\ln \underline{x}, \ln \overline{x}]$
$\mathbb{R}$	z^{2k} for fixed $k \in \mathbb{N}$	$[(\mathrm{mid}(0, \underline{x}, \overline{x}))^{2k}, \max\{\underline{x}^{2k}, \overline{x}^{2k}\}]$
$\mathbb{R}$	z^{2k+1} for fixed $k \in \mathbb{N}$	$[\underline{x}^{2k+1}, \overline{x}^{2k+1}]$
$(0, +\infty)$	$\sqrt{z}$	$[\sqrt{\underline{x}}, \sqrt{\overline{x}}]$
$\mathbb{R}$	$\lvert z \rvert$	$[\lvert\mathrm{mid}(0, \underline{x}, \overline{x})\rvert, \max\{\lvert\underline{x}\rvert, \lvert\overline{x}\rvert\}]$
$(0, +\infty)$	$\frac{1}{z^k}$ for fixed $k \in \mathbb{N}$	$\left[\frac{1}{\overline{z}^k}, \frac{1}{\underline{z}^k}\right]$
$(-\infty, 0)$	$\frac{1}{z^{2k}}$ for fixed $k \in \mathbb{N}$	$\left[\frac{1}{\underline{z}^{2k}}, \frac{1}{\overline{z}^{2k}}\right]$
$(-\infty, 0)$	$\frac{1}{z^{2k-1}}$ for fixed $k \in \mathbb{N}$	$\left[\frac{1}{\overline{z}^{2k-1}}, \frac{1}{\underline{z}^{2k-1}}\right]$

9.2.3 McCormick objects and relaxations

This section presents and extends definitions and properties from the PhD thesis [100], which essentially reframe the classical development of McCormick's relaxation technique [74] in terms of the abstract objects containing bounding and relaxing information that are propagated by MC++ [15] in order to carry out McCormick's scheme in practice.

Definition 9.2.14. *The set of* McCormick objects *of n variables is defined as* $\mathbb{MR}^n := \{(\boldsymbol{z}^{\mathrm{B}}, \boldsymbol{z}^{\mathrm{C}}) \in \mathbb{IR}^n \times \mathbb{IR}^n : \boldsymbol{z}^{\mathrm{B}} \cap \boldsymbol{z}^{\mathrm{C}} \neq \emptyset\}$. *For any* $\mathcal{X} \in \mathbb{MR}^n$, $\mathcal{X}$ *will be represented equivalently as*

$$\mathcal{X} \equiv (\boldsymbol{x}^{\mathrm{B}}, \boldsymbol{x}^{\mathrm{C}}) \equiv ([\underline{\mathbf{x}}^{\mathrm{B}}, \overline{\mathbf{x}}^{\mathrm{B}}], [\underline{\mathbf{x}}^{\mathrm{C}}, \overline{\mathbf{x}}^{\mathrm{C}}]).$$

Given $\mathcal{X}, \mathcal{Y} \in \mathbb{MR}^n$, $\mathcal{X} \subset \mathcal{Y}$ *if and only if both* $\boldsymbol{x}^{\mathrm{B}} \subset \boldsymbol{y}^{\mathrm{B}}$ *and* $\boldsymbol{x}^{\mathrm{C}} \subset \boldsymbol{y}^{\mathrm{C}}$. *The set of* proper McCormick objects *of n variables is* $\mathbb{MR}^n_{\mathrm{prop}} := \{(\boldsymbol{z}^{\mathrm{B}}, \boldsymbol{z}^{\mathrm{C}}) \in \mathbb{MR}^n : \boldsymbol{z}^{\mathrm{C}} \subset \boldsymbol{z}^{\mathrm{B}}\}$. *Given a set* $B \subset \mathbb{R}^n$, *define* $\mathbb{M}B := \{\mathcal{X} \in \mathbb{MR}^n : \boldsymbol{x}^{\mathrm{B}} \in \mathbb{I}B\}$, *and* $\mathbb{M}B_{\mathrm{prop}} := \{\mathcal{X} \in \mathbb{MR}^n : \boldsymbol{x}^{\mathrm{C}} \subset \boldsymbol{x}^{\mathrm{B}} \in \mathbb{I}B\} \subset \mathbb{MR}^n_{\mathrm{prop}}$.

Definition 9.2.15. *Given a function* $\mathbf{f} : B \subset \mathbb{R}^n \to \mathbb{R}^m$, *a mapping*

$$\mathcal{F} : \mathbb{M}B \ (\textit{or } \mathbb{M}B_{\text{prop}}) \to \mathbb{M}\mathbb{R}^m$$

is a McCormick extension *of* $\mathbf{f}$ *if* $\mathcal{F}([\mathbf{x},\mathbf{x}],[\mathbf{x},\mathbf{x}]) = ([\mathbf{f}(\mathbf{x}),\mathbf{f}(\mathbf{x})],[\mathbf{f}(\mathbf{x}),\mathbf{f}(\mathbf{x})])$ *for each* $\mathbf{x} \in B$.

Definition 9.2.16. *Given a set* $B \subset \mathbb{R}^n$*, a function* $\mathcal{F} : \mathbb{M}B \ (\textit{or } \mathbb{M}B_{\text{prop}}) \to \mathbb{M}\mathbb{R}^m$ *is* inclusion monotonic *if* $\mathcal{F}(\mathcal{X}) \subset \mathcal{F}(\mathcal{Y})$ *for each pair* $\mathcal{X},\mathcal{Y} \in \mathbb{M}B$ *(or* $\mathbb{M}B_{\text{prop}}$*) such that* $\mathcal{X} \subset \mathcal{Y}$.

Definition 9.2.17. *A pair* $\mathcal{X},\mathcal{Y} \in \mathbb{M}\mathbb{R}^n$ *is* coherent *if* $\boldsymbol{x}^{\text{B}} = \boldsymbol{y}^{\text{B}}$*. Given coherent* $\mathcal{X},\mathcal{Y} \in \mathbb{M}\mathbb{R}^n$*, for each* $\lambda \in [0,1]$*, define:*

$$\mathcal{C}\text{onv}(\lambda,\mathcal{X},\mathcal{Y}) := (\boldsymbol{x}^{\text{B}}, \lambda\boldsymbol{x}^{\text{C}} + (1-\lambda)\boldsymbol{y}^{\text{C}}).$$

Given a set $B \subset \mathbb{R}^n$*, a function* $\mathcal{F} : \mathbb{M}B \ (\textit{or } \mathbb{M}B_{\text{prop}}) \to \mathbb{M}\mathbb{R}^m$ *is* coherent *if* $\mathcal{F}(\mathcal{X})$ *is coherent to* $\mathcal{F}(\mathcal{Y})$ *for every coherent* $\mathcal{X},\mathcal{Y} \in \mathbb{M}B$ *(or* $\mathbb{M}B_{\text{prop}}$*). A function* $\mathcal{F} : \mathbb{M}B \ (\textit{or } \mathbb{M}B_{\text{prop}}) \to \mathbb{M}\mathbb{R}^m$ *is* coherently concave *if it is coherent, and, for every* $\mathcal{X},\mathcal{Y} \in \mathbb{M}B$ *(or* $\mathbb{M}B_{\text{prop}}$*),*

$$\mathcal{F}(\mathcal{C}\text{onv}(\lambda,\mathcal{X},\mathcal{Y})) \supset \mathcal{C}\text{onv}(\lambda,\mathcal{F}(\mathcal{X}),\mathcal{F}(\mathcal{Y})), \qquad \forall\lambda \in [0,1].$$

The following definition is stricter than in [100], and combines the above properties.

Definition 9.2.18. *Given a function* $\mathbf{f} : B \subset \mathbb{R}^n \to \mathbb{R}^m$*, a mapping*

$$\mathcal{F} : \mathbb{M}B \ (\textit{or } \mathbb{M}B_{\text{prop}}) \to \mathbb{M}\mathbb{R}^m$$

is a relaxation function *for* $\mathbf{f}$ *if it is coherently concave, inclusion monotonic, and a McCormick extension of* $\mathbf{f}$.

The following two propositions demonstrate the utility of relaxation functions: they are closed under composition, and effectively define convex underestimators and concave overestimators of the underlying functions they relax.

Proposition 9.2.19 (Lemmas 2.4.15 and 2.4.17 in [100]). *Consider functions* $\mathbf{f} : B \subset \mathbb{R}^n \to D \subset \mathbb{R}^m$ *and* $\mathbf{g} : D \to \mathbb{R}^k$*, a relaxation function* $\mathcal{F} : \mathbb{M}B$ *(or* $\mathbb{M}B_{\text{prop}}$*)* $\to \mathbb{M}\mathbb{R}^m$ *for* $\mathbf{f}$*, and a relaxation function* $\mathcal{G} : \mathbb{M}D$ *(or* $\mathbb{M}D_{\text{prop}}$*)* $\to \mathbb{M}\mathbb{R}^k$ *for* $\mathbf{g}$*. Define* $B_0 := \{\mathcal{X} \in \mathbb{M}B$ *(or* $\mathbb{M}B_{\text{prop}}$*)* $: \mathcal{F}(\mathcal{X}) \in \mathbb{M}D\}$*. If there are no domain violations in constructing the composition* $\mathcal{G} \circ \mathcal{F} : B_0 \to \mathbb{M}\mathbb{R}^k$*, then* $\mathcal{G} \circ \mathcal{F}$ *is a relaxation function for* $\mathbf{g} \circ \mathbf{f} : B \to \mathbb{R}^k$*.*

Proposition 9.2.20 (Lemma 2.4.11 in [100]). *Given a function* $f : B \subset \mathbb{R}^n \to \mathbb{R}$*, a relaxation function* $\mathcal{F} : \mathbb{M}B$ *(or* $\mathbb{M}B_{\text{prop}}$*)* $\to \mathbb{M}\mathbb{R}$ *for* f *on* B*, and some* $\boldsymbol{x} \in \mathbb{I}B$*, define functions* $\phi_{f,\boldsymbol{x}}, \psi_{f,\boldsymbol{x}} : \boldsymbol{x} \to \mathbb{R}$ *such that:*

$$\phi_{f,\boldsymbol{x}}(\mathbf{z}) = \underline{f}^{\mathrm{C}}(\boldsymbol{x}, [\mathbf{z}, \mathbf{z}]), \qquad \textit{and} \qquad \psi_{f,\boldsymbol{x}}(\mathbf{z}) = \overline{f}^{\mathrm{C}}(\boldsymbol{x}, [\mathbf{z}, \mathbf{z}]), \qquad \forall \mathbf{z} \in \boldsymbol{x}.$$

Then $\phi_{f,\boldsymbol{x}}$ *is convex on* $\boldsymbol{x}$*,* $\psi_{f,\boldsymbol{x}}$ *is concave on* $\boldsymbol{x}$*, and* $\phi_{f,\boldsymbol{x}}(\mathbf{z}) \leq f(\mathbf{z}) \leq \psi_{f,\boldsymbol{x}}(\mathbf{z})$ *for each* $\mathbf{z} \in \boldsymbol{x}$*.*

The goal of this chapter is to obtain $\mathcal{C}^i$ relaxations of an MC-factorable function, for some $i \in \{1, 2\}$. Achieving this will require appending the following non-standard assumption to Definition 9.2.12 for each employed univariate intrinsic function. This assumption will be invoked explicitly whenever it is required.

Assumption 9.2.21. *For particular* $i \in \{1, 2\}$*, given a univariate intrinsic function* $u : B \subset \mathbb{R} \to \mathbb{R}$*, assume for each* $\boldsymbol{x} \in \mathbb{I}B$ *that the functions* $u^{\mathrm{cv}}(\boldsymbol{x}, \cdot)$ *and* $u^{\mathrm{cc}}(\boldsymbol{x}, \cdot)$ *are each* $\mathcal{C}^i$ *on* $\boldsymbol{x}$*, and that* $\underline{\tilde{u}}(\boldsymbol{x}) \leq u^{\mathrm{cv}}(\boldsymbol{x}, \zeta_u^{\min}(\boldsymbol{x}))$ *and* $\overline{\tilde{u}}(\boldsymbol{x}) \geq u^{\mathrm{cc}}(\boldsymbol{x}, \zeta_u^{\max}(\boldsymbol{x}))$*. If* $i = 2$*, assume additionally that:*

- *if* $\zeta_u^{\min}(\boldsymbol{x}) \in \operatorname{int}(\boldsymbol{x})$*, then the second derivative of* $u^{\mathrm{cv}}(\boldsymbol{x}, \cdot)$ *is zero at* $\zeta_u^{\min}(\boldsymbol{x})$*, and*
- *if* $\zeta_u^{\max}(\boldsymbol{x}) \in \operatorname{int}(\boldsymbol{x})$*, then the second derivative of* $u^{\mathrm{cc}}(\boldsymbol{x}, \cdot)$ *is zero at* $\zeta_u^{\max}(\boldsymbol{x})$*.*

Observe that the above assumption does not require u itself to be $\mathcal{C}^i$. Indeed, the $\mathcal{C}^i$ relaxations obtained in this chapter will remain valid even when nondifferentiable univariate intrinsic functions are employed. However, [11] shows that nondifferentiable intrinsic functions cannot satisfy Assumption 9.2.38 below, which

will be required in this chapter to ensure sufficiently rapid convergence of the obtained relaxations to the original function as the width of $\boldsymbol{x}$ approaches zero.

Remark 9.2.22. *The following two lemmata show that if $i = 1$ and a function $u : B \to \mathbb{R}$ is $\mathcal{C}^1$, then Assumption 9.2.21 and the conditions of Definition 9.2.12 are satisfied when $u^{\mathrm{cv}}(\boldsymbol{x}, \cdot)$ and $u^{\mathrm{cc}}(\boldsymbol{x}, \cdot)$ are chosen to be the convex and concave envelopes of u on $\boldsymbol{x}$, respectively, and when $\zeta_u^{\min}(\boldsymbol{x})$ and $\zeta_u^{\max}(\boldsymbol{x})$ are chosen according to Definition 9.2.12.*

Lemma 9.2.23. *Consider an interval $\boldsymbol{x} \in \mathbb{IR}$, a Lipschitz continuous function $u : \boldsymbol{x} \to \mathbb{R}$, and the convex envelope $u^{\mathrm{cv}} : \boldsymbol{x} \to \mathbb{R}$ of u on $\boldsymbol{x}$. Then, $u^{\mathrm{cv}}(\underline{x}) = u(\underline{x})$ and $u^{\mathrm{cv}}(\overline{x}) = u(\overline{x})$. Moreover, u^{cv} is Lipschitz continuous on $\boldsymbol{x}$, with the same Lipschitz constant as u.*

Proof. The required result is trivial if $\underline{x} = \overline{x}$, so assume that $\underline{x} < \overline{x}$. Let k_u denote a Lipschitz constant for u on $\boldsymbol{x}$. Applying the definition of the convex envelope,

$$u(y) \geq u^{\mathrm{cv}}(y) \geq u(\underline{x}) - k_u(y - \underline{x}), \qquad \forall y \in \boldsymbol{x}; \tag{9.3}$$

the first inequality above is due to u dominating u^{cv}, and the second inequality is due to u^{cv} dominating each convex underestimator of u on $\boldsymbol{x}$. Setting y to $\underline{x}$ in the above inequality chain yields $u^{\mathrm{cv}}(\underline{x}) = u(\underline{x})$.

A similar argument yields:

$$u(y) \geq u^{\mathrm{cv}}(y) \geq u(\overline{x}) + k_u(y - \overline{x}), \qquad \forall y \in \boldsymbol{x}; \tag{9.4}$$

setting y to $\overline{x}$ yields $u^{\mathrm{cv}}(\overline{x}) = u(\overline{x})$.

Thus, (9.3) and (9.4) become:

$$\begin{aligned} u^{\mathrm{cv}}(y) - u^{\mathrm{cv}}(\underline{x}) &\geq -k_u(y - \underline{x}), \qquad \forall y \in \boldsymbol{x}, \\ u^{\mathrm{cv}}(y) - u^{\mathrm{cv}}(\overline{x}) &\geq k_u(y - \overline{x}), \qquad \forall y \in \boldsymbol{x}. \end{aligned}$$

Defining D_+u^{cv} and D_-u^{cv} as the right-derivative and left-derivative of u^{cv} described in [42, Part I, Theorem 4.1.1], it follows from [42, Part I, Proposition 4.1.3] that $D_+u^{\mathrm{cv}}(\underline{x})$ and $D_-u^{\mathrm{cv}}(\overline{x})$ both exist, are finite, and satisfy

$$D_+u^{\mathrm{cv}}(\underline{x}) \geq -k_u, \qquad \text{and} \qquad D_-u^{\mathrm{cv}}(\overline{x}) \leq k_u.$$

Thus, u^{cv} is continuous at $\underline{x}$ and $\overline{x}$. Moreover, [42, Part I, Theorem 4.2.1] implies that for each $y \in \mathrm{int}(\boldsymbol{x})$, each subgradient of u^{cv} at y is an element of $[-k_u, k_u]$. This result, combined with the mean-value theorem [42, Part I, Theorem 4.2.4], shows that u^{cv} is Lipschitz continuous on $\boldsymbol{x}$, with a Lipschitz constant of k_u. □

Lemma 9.2.24. *Consider an interval $\boldsymbol{x} \in \mathbb{IR}$, and a $\mathcal{C}^1$ function $u : \boldsymbol{x} \to \mathbb{R}$. The convex envelope $u^{\mathrm{cv}} : \boldsymbol{x} \to \mathbb{R}$ of u on $\boldsymbol{x}$ is also $\mathcal{C}^1$ on $\boldsymbol{x}$.*

Proof. The required result is trivial if $\underline{x} = \overline{x}$, so assume that $\underline{x} < \overline{x}$. Theorem 3.2 in [31] implies that u^{cv} is $\mathcal{C}^1$ on $(\underline{x}, \overline{x}) = \mathrm{int}(\boldsymbol{x})$; it remains to be shown that u^{cv} is also $\mathcal{C}^1$ at $\underline{x}$ and $\overline{x}$. Noting that u is Lipschitz continuous on $\boldsymbol{x}$, construct the right-derivative D_+u^{cv} and the left-derivative D_-u^{cv} as in the proof of Lemma 9.2.23. As in the proof of Lemma 9.2.23, $D_+u^{\mathrm{cv}}(\underline{x})$ and $D_-u^{\mathrm{cv}}(\overline{x})$ each exist and are finite. Define the following function, which extends the domain of u^{cv} to $\mathbb{R}$:

$$\psi : \mathbb{R} \to \mathbb{R} : y \mapsto \begin{cases} u^{\mathrm{cv}}(\underline{x}) + (D_+u^{\mathrm{cv}}(\underline{x}))(y - \underline{x}), & \text{if } y < \underline{x}, \\ u^{\mathrm{cv}}(y), & \text{if } y \in \boldsymbol{x}, \\ u^{\mathrm{cv}}(\overline{x}) + (D_-u^{\mathrm{cv}}(\overline{x}))(y - \overline{x}), & \text{if } \overline{x} < y. \end{cases}$$

The function ψ is evidently continuous, and is $\mathcal{C}^1$ at each $y \in \mathbb{R}\backslash\{\underline{x}, \overline{x}\}$. Applying the definitions of D_+u^{cv} and D_-u^{cv}, it follows that ψ is differentiable at $\underline{x}$ and $\overline{x}$ as well; thus,

$$\nabla\psi(y) = \begin{cases} D_+u^{\mathrm{cv}}(\underline{x}), & \text{if } y \leq \underline{x}, \\ \nabla u^{\mathrm{cv}}(y), & \text{if } y \in \mathrm{int}(\boldsymbol{x}), \\ D_-u^{\mathrm{cv}}(\overline{x}), & \text{if } \overline{x} \leq y. \end{cases}$$

This equation, together with [42, Part I, Theorem 4.2.1(iii)], shows that ψ is $\mathcal{C}^1$ even at $\underline{x}$ and $\overline{x}$, and is therefore $\mathcal{C}^1$ on $\mathbb{R}$. Hence, u^{cv} is $\mathcal{C}^1$ on $\boldsymbol{x}$. □

Remark 9.2.25. *Consider a univariate function u that is $\mathcal{C}^2$, is either convex or concave, and is either monotonically increasing or monontonically decreasing on its domain. Moreover, note that the concave envelope of a univariate convex function on an interval is a secant, as is the convex envelope of a univariate concave function on an interval. Thus, even if $i = 2$, Assumption 9.2.21 and the conditions of Definition 9.2.12 are satisfied when $u^{\mathrm{cv}}(\boldsymbol{x}, \cdot)$ and $u^{\mathrm{cc}}(\boldsymbol{x}, \cdot)$ are chosen to be the convex and concave envelopes of u on $\boldsymbol{x}$, respectively, and when $\zeta_u^{\min}(\boldsymbol{x})$ and $\zeta_u^{\max}(\boldsymbol{x})$ are chosen according to Definition 9.2.12.*

Remark 9.2.26. *The condition in Assumption 9.2.21 that both* $\underline{\tilde{u}}(\boldsymbol{x}) \leq u^{\mathrm{cv}}(\boldsymbol{x}, \zeta_u^{\min}(\boldsymbol{x}))$ *and* $\overline{\tilde{u}}(\boldsymbol{x}) \geq u^{\mathrm{cv}}(\boldsymbol{x}, \zeta_u^{\max}(\boldsymbol{x}))$ *can be imposed without loss of generality, as detailed in Remark 9.3.13 below.*

Example 9.2.27. *Suppose the function* $u : \mathbb{R} \to \mathbb{R} : x \mapsto x^2$ *is considered as a univariate intrinsic function. Since* u *is convex, it is its own convex envelope on any subinterval of* $\mathbb{R}$*. In line with Remark 9.2.22, if* $i = 1$*, setting* $u^{\mathrm{cv}}(\boldsymbol{x}, \cdot) \equiv u$ *for each* $\boldsymbol{x} \in \mathbb{IR}$ *is therefore consistent with Definition 9.2.12 and Assumption 9.2.21. However, on an interval* $\boldsymbol{y}$ *with* $\underline{y} < 0 < \overline{y}$*,* $\zeta_u^{\min}(\boldsymbol{y}) = 0 \in \operatorname{int}(\boldsymbol{y})$*, but* $\nabla^2 u(0) = 2$*, so setting* $u^{\mathrm{cv}}(\boldsymbol{y}, \cdot) \equiv u$ *is inconsistent with Assumption 9.2.21 when* $i = 2$*. Nevertheless, it is readily shown that the following choice of* u^{cv} *is consistent with Definition 9.2.12 and Assumption 9.2.21 when* $i = 2$*:*

$$u^{\mathrm{cv}}(\boldsymbol{y}, z) := \begin{cases} z^2 & \text{if } 0 \notin (\underline{y}, \overline{y}), \\ \frac{z^3}{(\overline{y})} & \text{if } \underline{y} < 0 < \overline{y} \text{ and } 0 \leq z, \\ \frac{z^3}{(\underline{y})} & \text{if } \underline{y} < 0 < \overline{y} \text{ and } z < 0. \end{cases}$$

Observe that setting $u^{\mathrm{cc}}(\boldsymbol{x}, \cdot)$ *to be the affine concave envelope of* u *on* $\boldsymbol{x}$ *is consistent with Assumption 9.2.21, since, in this case, either* $\underline{x}$ *or* $\overline{x}$ *will be a valid choice of* $\zeta_u^{\max}(\boldsymbol{x})$*.*

Example 9.2.28. *Suppose the function* $u : \mathbb{R} \to \mathbb{R} : x \mapsto |x|$ *is considered as a univariate intrinsic function. In the spirit of the previous example, it is readily confirmed that the following choice of* u^{cv} *is consistent with Definition 9.2.12 and Assumption 9.2.21 for each* $i \in \{1, 2\}$*:*

$$u^{\mathrm{cv}}(\boldsymbol{y}, x) := \begin{cases} |x| & \text{if } 0 \notin (\underline{y}, \overline{y}), \\ \frac{x^{2+i}}{\overline{y}^{1+i}} & \text{if } \underline{y} < 0 < \overline{y} \text{ and } 0 \leq x, \\ \left| \frac{x^{2+i}}{\underline{y}^{1+i}} \right| & \text{if } \underline{y} < 0 < \overline{y} \text{ and } x < 0. \end{cases}$$

As in the previous example, observe that setting $u^{\mathrm{cc}}(\boldsymbol{x}, \cdot)$ *to be the affine concave envelope of* u *on* $\boldsymbol{x}$ *is consistent with Assumption 9.2.21.*

Example 9.2.29. *For some fixed* $k \in \mathbb{N}$*, suppose the function* $u : \mathbb{R} \to \mathbb{R} : x \mapsto x^{2k+1}$ *is considered as a univariate intrinsic function. In line with Remark 9.2.22, if* $i = 1$*,*

setting $u^{\mathrm{cv}}(\boldsymbol{x},\cdot)$ and $u^{\mathrm{cc}}(\boldsymbol{x},\cdot)$ to be the convex/concave envelopes of u described in [69] is consistent with Definition 9.2.12 and Assumption 9.2.21. If $i = 2$, then it is readily verified that the following choices of u^{cv} and u^{cc} are consistent with Definition 9.2.12 and Assumption 9.2.21:

$$u^{\mathrm{cv}}(\boldsymbol{x},z) := \begin{cases} \underline{x}^{2k+1} + (\overline{x}^{2k+1} - \underline{x}^{2k+1})\left(\frac{z-\underline{x}}{\overline{x}-\underline{x}}\right), & \text{if } \overline{x} \le 0, \\ \underline{x}^{2k+1}\left(\frac{\overline{x}-z}{\overline{x}-\underline{x}}\right) + (\max\{0,z\})^{2k+1}, & \text{if } \underline{x} < 0 < \overline{x}, \\ z^{2k+1}, & \text{if } 0 \le \underline{x}, \end{cases}$$

$$u^{\mathrm{cc}}(\boldsymbol{x},z) := \begin{cases} z^{2k+1}, & \text{if } \overline{x} \le 0, \\ \overline{x}^{2k+1}\left(\frac{z-\underline{x}}{\overline{x}-\underline{x}}\right) + (\min\{0,z\})^{2k+1}, & \text{if } \underline{x} < 0 < \overline{x}, \\ \underline{x}^{2k+1} + (\overline{x}^{2k+1} - \underline{x}^{2k+1})\left(\frac{z-\underline{x}}{\overline{x}-\underline{x}}\right), & \text{if } 0 \le \underline{x}. \end{cases}$$

The functions $u^{\mathrm{cv}}(\boldsymbol{x},\cdot)$ and $u^{\mathrm{cc}}(\boldsymbol{x},\cdot)$ described above are evidently strictly increasing on $\boldsymbol{x}$ for each $\boldsymbol{x} \in \mathbb{IR}$. Thus, setting $\zeta_u^{\min}(\boldsymbol{x}) := \underline{x}$ and $\zeta_u^{\max}(\boldsymbol{x}) := \overline{x}$ is consistent with Definition 9.2.12.

Any univariate function $u : B \subset \mathbb{R} \to \mathbb{R}$ on an open set can be considered to be a univariate intrinsic function, provided that the functions $\tilde{\boldsymbol{u}}$, u^{cv}, and u^{cc} are known or can be constructed. Table 9.2 presents functions $u^{\mathrm{cv}}, u^{\mathrm{cc}}$ which satisfy the conditions of Definition 9.2.12 and Assumption 9.2.21 for the univariate intrinsic functions u considered in Table 9.1.

Within this framework, McCormick's relaxations [74] can be restated as the convex/concave relaxations implied by Propositions 9.2.19 and 9.2.20 for an MC-factorable function, when the following relaxation functions are applied in place of each addition/multiplication/univariate intrinsic operation. As demonstrated in [100], these McCormick operations are indeed relaxation functions of the corresponding operations on real numbers.

Definition 9.2.30. *Define an addition operation $+ : \mathbb{MR}^2 \to \mathbb{MR}$ so that, for each $\mathcal{X}, \mathcal{Y} \in \mathbb{MR}$,*

$$+(\mathcal{X},\mathcal{Y}) \equiv \mathcal{X} + \mathcal{Y} := (\boldsymbol{x}^{\mathrm{B}} + \boldsymbol{y}^{\mathrm{B}}, \boldsymbol{x}^{\mathrm{C}} + \boldsymbol{y}^{\mathrm{C}}),$$

Table 9.2: Functions $u^{\mathrm{cv}}, u^{\mathrm{cc}}$ that satisfy the conditions of Definition 9.2.12 and Assumption 9.2.21 for various univariate intrinsic functions u.

B	$u(z)$ for $z \in B$	$u^{\mathrm{cv}}(\boldsymbol{x}, z)$ for $\boldsymbol{x} \in \mathbb{I}B, z \in B$	$u^{\mathrm{cc}}(\boldsymbol{x}, z)$ for $\boldsymbol{x} \in \mathbb{I}B, z \in B$
$\mathbb{R}$	cz for fixed $c \in \mathbb{R}$	cz	cz
$\mathbb{R}$	$\exp z$	$\exp z$	$\exp \underline{x} + (\exp \overline{x} - \exp \underline{x}) \left(\frac{z-\underline{x}}{\overline{x}-\underline{x}}\right)$
$(0, +\infty)$	$\ln z$	$\ln \underline{x} + (\ln \overline{x} - \ln \underline{x}) \left(\frac{z-\underline{x}}{\overline{x}-\underline{x}}\right)$	$\ln z$
$\mathbb{R}$	z^2	See Example 9.2.27	$\underline{x}^2 + (\overline{x}^2 - \underline{x}^2) \left(\frac{z-\underline{x}}{\overline{x}-\underline{x}}\right)$
$\mathbb{R}$	z^{2k+2} for fixed $k \in \mathbb{N}$	z^{2k+2}	$\underline{x}^{2k+2} + (\overline{x}^{2k+2} - \underline{x}^{2k+2}) \left(\frac{z-\underline{x}}{\overline{x}-\underline{x}}\right)$
$\mathbb{R}$	z^{2k+1} for fixed $k \in \mathbb{N}$	See Example 9.2.29	See Example 9.2.29
$(0, +\infty)$	$\sqrt{z}$	$\sqrt{\underline{x}} + (\sqrt{\overline{x}} - \sqrt{\underline{x}}) \left(\frac{z-\underline{x}}{\overline{x}-\underline{x}}\right)$	$\sqrt{z}$
$\mathbb{R}$	$\lvert z \rvert$	See Example 9.2.28	$\lvert \underline{x} \rvert + (\lvert \overline{x} \rvert - \lvert \underline{x} \rvert) \left(\frac{z-\underline{x}}{\overline{x}-\underline{x}}\right)$
$(0, +\infty)$	$\frac{1}{z^k}$ for fixed $k \in \mathbb{N}$	$\frac{1}{z^k}$	$\frac{1}{\underline{x}^k} + (\frac{1}{\overline{x}^k} - \frac{1}{\underline{x}^k}) \left(\frac{z-\underline{x}}{\overline{x}-\underline{x}}\right)$
$(-\infty, 0)$	$\frac{1}{z^{2k}}$ for fixed $k \in \mathbb{N}$	$\frac{1}{z^{2k}}$	$\frac{1}{\underline{x}^{2k}} + (\frac{1}{\overline{x}^{2k}} - \frac{1}{\underline{x}^{2k}}) \left(\frac{z-\underline{x}}{\overline{x}-\underline{x}}\right)$
$(-\infty, 0)$	$\frac{1}{z^{2k-1}}$ for fixed $k \in \mathbb{N}$	$\frac{1}{\underline{x}^{2k-1}} + (\frac{1}{\overline{x}^{2k-1}} - \frac{1}{\underline{x}^{2k-1}}) \left(\frac{z-\underline{x}}{\overline{x}-\underline{x}}\right)$	$\frac{1}{z^{2k-1}}$

and define $+ : \mathbb{MR}^2_{\text{prop}} \to \mathbb{MR}$ *as the restriction of* $+ : \mathbb{MR}^2 \to \mathbb{MR}$ *to the domain* $\mathbb{MR}^2_{\text{prop}}$.

The following definition of a McCormick multiplication operation is adapted from McCormick's original presentation [74] and [76], and will be replaced in this chapter by Definition 9.3.19 further below. The following operation will be denoted by the symbol "$\bullet$"; the usual notation for multiplication will be reserved for Definition 9.3.19. Note that multiplication of a scalar and a McCormick object was previously treated as a univariate intrinsic function in Tables 9.1 and 9.2; the following definition instead concerns multiplication of two McCormick objects.

Definition 9.2.31. *Define a* classical McCormick multiplication *operation* $\bullet : \mathbb{MR}^2 \to \mathbb{MR}$ *so that, for each* $\mathcal{X}, \mathcal{Y} \in \mathbb{MR}$,

$$\bullet(\mathcal{X}, \mathcal{Y}) \equiv \mathcal{X} \bullet \mathcal{Y} := (\boldsymbol{x}^{\mathrm{B}} \boldsymbol{y}^{\mathrm{B}}, \boldsymbol{z}),$$

where $\boldsymbol{z} \equiv [\underline{z}, \overline{z}] \in \mathbb{IR}$ *is defined in terms of the intermediate quantities* $\boldsymbol{v} := \boldsymbol{x}^{\mathrm{B}} \cap \boldsymbol{x}^{\mathrm{C}}$ *and* $\boldsymbol{w} := \boldsymbol{y}^{\mathrm{B}} \cap \boldsymbol{y}^{\mathrm{C}}$ *as follows:*

$$\begin{aligned}
\underline{z} &:= \max\left(\underline{(\underline{y}^{\mathrm{B}}\boldsymbol{v})} + \underline{(\underline{x}^{\mathrm{B}}\boldsymbol{w})} - \underline{x}^{\mathrm{B}}\underline{y}^{\mathrm{B}},\ \underline{(\overline{y}^{\mathrm{B}}\boldsymbol{v})} + \underline{(\overline{x}^{\mathrm{B}}\boldsymbol{w})} - \overline{x}^{\mathrm{B}}\overline{y}^{\mathrm{B}}\right), \\
\overline{z} &:= \min\left(\overline{(\underline{y}^{\mathrm{B}}\boldsymbol{v})} + \overline{(\overline{x}^{\mathrm{B}}\boldsymbol{w})} - \overline{x}^{\mathrm{B}}\underline{y}^{\mathrm{B}},\ \overline{(\overline{y}^{\mathrm{B}}\boldsymbol{v})} + \overline{(\underline{x}^{\mathrm{B}}\boldsymbol{w})} - \underline{x}^{\mathrm{B}}\overline{y}^{\mathrm{B}}\right).
\end{aligned}$$

Definition 9.2.32. *Define a function* $\mathrm{mid} : \mathbb{R}^3 \to \mathbb{R}$ *as mapping to the median of its three scalar arguments. Given a univariate intrinsic function* $u : B \subset \mathbb{R} \to \mathbb{R}$ *that satisfies Assumption 9.2.21, define* $\mathcal{U} : \mathbb{M}B \to \mathbb{MR}$ *so that for each* $\mathcal{X} \in \mathbb{M}B$, $\mathcal{U}(\mathcal{X}) := (\tilde{\boldsymbol{u}}(\boldsymbol{x}^{\mathrm{B}}), \boldsymbol{z})$, *where*

$$\boldsymbol{z} := [u^{\mathrm{cv}}(\boldsymbol{x}^{\mathrm{B}}, \mathrm{mid}(\zeta_u^{\min}(\boldsymbol{x}^{\mathrm{B}}), \underline{x}^{\mathrm{C}}, \overline{x}^{\mathrm{C}})), u^{\mathrm{cc}}(\boldsymbol{x}^{\mathrm{B}}, \mathrm{mid}(\zeta_u^{\max}(\boldsymbol{x}^{\mathrm{B}}), \underline{x}^{\mathrm{C}}, \overline{x}^{\mathrm{C}}))].$$

The above definitions suggest the construction of an analog of a natural interval extension for an MC-factorable function, using McCormick objects instead of

intervals. This notion is formalized in the following definition, which is motivated by the subsequent theorem.

Definition 9.2.33 (adapted from [100]). *Given an MC-factorable function* $\mathbf{f} : B \subset \mathbb{R}^n \to \mathbb{R}^m$*, a* natural McCormick extension $\mathcal{F} : \mathbb{M}B \to \mathbb{M}\mathbb{R}^m$ *of* $\mathbf{f}$ *is defined by replacing each addition operation, multiplication operation, and univariate intrinsic function in the construction of* $\mathbf{f}$ *with its McCormick counterpart described by Definitions 9.2.30–9.2.32, provided that there are no domain violations in the introduced McCormick arithmetic.*

Theorem 9.2.34 (Theorem 2.4.32 in [100]). *Given an MC-factorable function* $\mathbf{f} : B \subset \mathbb{R}^n \to \mathbb{R}^m$ *with a well-defined natural McCormick extension* $\mathcal{F}$*,* $\mathcal{F}$ *is a relaxation function for* $\mathbf{f}$*.*

The classical McCormick relaxations of an MC-factorable function are the convex/concave relaxations implied by the above theorem and by Proposition 9.2.20. These relaxations may be nonsmooth; the central goal of this chapter is to develop $\mathcal{C}^1$ and $\mathcal{C}^2$ variants of these relaxations.

9.2.4 Convergence order

Intuitively, to be useful, a scheme for constructing convex and concave relaxations of a scalar-valued function on an interval should converge rapidly to the underlying function as the width of interval is reduced to zero. Appopriate notions of convergence were formalized in [11], and were extended to McCormick objects in the PhD thesis [96]. This section summarizes the definitions and properties that are relevant to the results in this chapter.

Definition 9.2.35 (adapted from [11]). *Given a continuous function* $f : B \subset \mathbb{R}^n \to \mathbb{R}$*, functions* $\{f^{\mathrm{cv}}(\boldsymbol{x}, \cdot), f^{\mathrm{cc}}(\boldsymbol{x}, \cdot) : \boldsymbol{x} \to \mathbb{R}\}_{\boldsymbol{x} \in \mathbb{I}B}$ *comprise a* scheme of estimators *for* f *if, for each* $\boldsymbol{x} \in \mathbb{I}B$*,* $f^{\mathrm{cv}}(\boldsymbol{x}, \cdot)$ *is convex on* $\boldsymbol{x}$*,* $f^{\mathrm{cc}}(\boldsymbol{x}, \cdot)$ *is concave on* $\boldsymbol{x}$*, and*

$$f^{\mathrm{cv}}(\boldsymbol{x}, \mathbf{z}) \leq f(\mathbf{z}) \leq f^{\mathrm{cc}}(\boldsymbol{x}, \mathbf{z}), \qquad \forall \mathbf{z} \in \boldsymbol{x}.$$

Such a scheme is pointwise convergent *of order t_0 if for each $\boldsymbol{q} \in \mathbb{I}B$, there exists $a_0 > 0$ such that*

$$\begin{aligned} & \sup_{\mathbf{z}\in\boldsymbol{x}}(f(\mathbf{z}) - f^{\mathrm{cv}}(\boldsymbol{x},\mathbf{z})) \leq a_0(\operatorname{wid}\boldsymbol{x})^{t_0}, && \forall \boldsymbol{x} \in \mathbb{I}\boldsymbol{q}, \\ \textit{and} \quad & \sup_{\mathbf{z}\in\boldsymbol{x}}(f^{\mathrm{cc}}(\boldsymbol{x},\mathbf{z}) - f(\mathbf{z})) \leq a_0(\operatorname{wid}\boldsymbol{x})^{t_0}, && \forall \boldsymbol{x} \in \mathbb{I}\boldsymbol{q}. \end{aligned}$$

The following example motivates the incorporation of the interval $\boldsymbol{q}$ into this definition.

Example 9.2.36. *Consider a function $f : \mathbb{R} \to \mathbb{R}$ and a scheme of estimators*

$$\{f^{\mathrm{cv}}(\boldsymbol{x},\cdot), f^{\mathrm{cc}}(\boldsymbol{x},\cdot)\}_{\boldsymbol{x}\in\mathbb{I}\mathbb{R}}$$

for f, for which, for each $\boldsymbol{x} \in \mathbb{I}\mathbb{R}$,

$$\sup_{z\in\boldsymbol{x}}(f(z) - f^{\mathrm{cv}}(\boldsymbol{x},z)) = \sup_{z\in\boldsymbol{x}}(f^{\mathrm{cc}}(\boldsymbol{x},z) - f(z)) = \begin{cases} (\operatorname{wid}\boldsymbol{x})^2 & \textit{if } \operatorname{wid}\boldsymbol{x} \leq 1, \\ (\operatorname{wid}\boldsymbol{x})^3 & \textit{if } \operatorname{wid}\boldsymbol{x} > 1. \end{cases}$$

According to the above definition, this scheme is pointwise convergent of order 2. In the original definition [11], however, this scheme is not pointwise convergent of order 2, since, for each $a_0 > 0$, there exists a sufficiently large interval $\boldsymbol{x} \in \mathbb{I}\mathbb{R}$ for which

$$a_0(\operatorname{wid}\boldsymbol{x})^2 < (\operatorname{wid}\boldsymbol{x})^3.$$

In fact, according to the definition in [11], this scheme is not pointwise convergent of any order. Since applications of pointwise convergence in [11] are only concerned with sufficiently small intervals, the interval $\boldsymbol{q}$ was added to the definition above so that the constants a_0 and t_0 need not apply to arbitrarily large intervals in $\mathbb{I}B$.

By Theorem 2 in [11], if f is nonaffine and twice-continuously differentiable, then there does not exist any scheme of estimators for f with pointwise convergence of order greater than 2. Given an MC-factorable function expressed as a composition of twice-continuously differentiable functions, the classical McCormick relaxations of this function are pointwise convergent of order 2 [11], as are the

αBB relaxations [1, 11]. A scheme of estimators with second-order pointwise convergence is typically necessary to mitigate clustering when carrying out a branch-and-bound method for global optimization [20]. Certain problems with nondifferentiable objective functions, however, are not subject to this requirement [118].

Consider an MC-factorable function f that is a composition only of locally Lipschitz continuous functions. Given a natural interval extension $\tilde{\boldsymbol{f}}$ of f, the constant mappings $\{\mathbf{z} \mapsto \underline{\tilde{f}}(\boldsymbol{x}), \mathbf{z} \mapsto \overline{\tilde{f}}(\boldsymbol{x})\}_{\boldsymbol{x} \in \mathbb{I}B}$ comprise a scheme of estimators for f that is pointwise convergent of order 1 [98].

The following definition formalizes a notion of width of a McCormick object, and a corresponding notion of convergence of a function of McCormick objects, as the width of the argument tends to zero.

Definition 9.2.37 (adapted from [96]). *A McCormick object $\mathcal{X} \in \mathbb{M}\mathbb{R}$ has a* width *of*

$$\operatorname{wid}_{\mathcal{M}} \mathcal{X} := \operatorname{wid}\left(\boldsymbol{x}^{\mathrm{B}} \cap \boldsymbol{x}^{\mathrm{C}}\right) = \min\{\overline{x}^{\mathrm{C}}, \overline{x}^{\mathrm{B}}\} - \max\{\underline{x}^{\mathrm{C}}, \underline{x}^{\mathrm{B}}\}.$$

A vector $\mathcal{Y} \in \mathbb{M}\mathbb{R}^n$ of McCormick objects has a width of

$$\operatorname{wid}_{\mathcal{M}} \mathcal{Y} \equiv \operatorname{wid}_{\mathcal{M}} (\mathcal{Y}_1, \ldots, \mathcal{Y}_n) := \max_{k \in \{1,\ldots,n\}} \operatorname{wid}_{\mathcal{M}} \mathcal{Y}_k.$$

A function $\mathcal{F} : \mathbb{M}B$ (or $\mathbb{M}B_{\text{prop}}$) $\subset \mathbb{M}\mathbb{R}^n \to \mathbb{M}\mathbb{R}^m$ has (t_1, t_2)-convergence on $\mathbb{M}B$ (or $\mathbb{M}B_{\text{prop}}$) if for each $\boldsymbol{q} \in \mathbb{I}B$, there exist $a_1, a_2 > 0$ such that

$$\operatorname{wid}_{\mathcal{M}} (\mathcal{F}(\mathcal{X})) \leq a_1 (\operatorname{wid}_{\mathcal{M}} \mathcal{X})^{t_1} + a_2 (\operatorname{wid} \boldsymbol{x}^{\mathrm{B}})^{t_2}, \qquad \forall \mathcal{X} \in \mathbb{M}\boldsymbol{q} \ (\text{or } \mathbb{M}\boldsymbol{q}_{\text{prop}}).$$

Again, the interval $\boldsymbol{q}$ has been added to this definition to prevent the fixed constants a_1, a_2 from having to be applicable to every choice of $\boldsymbol{x}^{\mathrm{B}} \in \mathbb{I}B$.

As described in Section 3.2 of [96], given a (t_1, t_2)-convergent relaxation function $\mathcal{F}$ for a function f, the corresponding convex/concave relaxations of f described by Proposition 9.2.20 exhibit pointwise convergence of order t_2. Moreover, as described in Section 3.9.7 of [96], a well-defined composition of $(1, 2)$-convergent McCormick-valued functions is itself $(1, 2)$-convergent. This notion

motivates the following assumption, which will be appended frequently to Definition 9.2.12.

Assumption 9.2.38. *Given a univariate intrinsic function $u : B \subset \mathbb{R} \to \mathbb{R}$, assume that the functions $\{u^{\mathrm{cv}}(\boldsymbol{x}, \cdot), u^{\mathrm{cc}}(\boldsymbol{x}, \cdot)\}_{\boldsymbol{x} \in \mathbb{I}B}$ comprise a scheme of estimators for u on B that is pointwise convergent of order 2.*

The above assumption is satisfied by the functions $u^{\mathrm{cv}}, u^{\mathrm{cc}}$ described in Table 9.2, except when u is the absolute value function $z \mapsto |z|$: this is demonstrated in [11] for each u other than $z \mapsto z^2$, which is considered in the following lemma.

Lemma 9.2.39. *Consider the relaxation scheme $\{u^{\mathrm{cv}}(\boldsymbol{x}, \cdot), u^{\mathrm{cc}}(\boldsymbol{x}, \cdot)\}_{\boldsymbol{x} \in \mathbb{IR}}$ for $u : z \mapsto z^2$ on $\mathbb{R}$ described in Example 9.2.27. This scheme satisfies Assumption 9.2.38.*

Proof. By [11, Theorem 10], the concave relaxations of u described in Example 9.2.27 are pointwise convergent of order 2, so it remains to consider only the convex relaxations of u.

Now, if $\boldsymbol{x} \in \mathbb{IR}$ but $0 \notin \operatorname{int}(\boldsymbol{x})$, then $u(z) - u^{\mathrm{cv}}(\boldsymbol{x}, z) = 0$ for all $z \in \boldsymbol{x}$. If, instead, $\boldsymbol{x} \in \mathbb{IR}$ and $0 \in \operatorname{int}(\boldsymbol{x})$, then

$$\begin{aligned}
\sup_{z \in \boldsymbol{x}} (u(z) - u^{\mathrm{cv}}(\boldsymbol{x}, z)) &= \max \left\{ \sup_{z \in [\underline{x}, 0]} (u(z) - u^{\mathrm{cv}}(\boldsymbol{x}, z)), \sup_{z \in [0, \overline{x}]} (u(z) - u^{\mathrm{cv}}(\boldsymbol{x}, z)) \right\}, \\
&= \max \left\{ \sup_{z \in [\underline{x}, 0]} \left(z^2 - \tfrac{z^3}{(\underline{x})} \right), \sup_{z \in [0, \overline{x}]} \left(z^2 - \tfrac{z^3}{(\overline{x})} \right) \right\}, \\
&= \max \left\{ \tfrac{4}{27} \underline{x}^2, \tfrac{4}{27} \overline{x}^2 \right\}, \\
&\leq \tfrac{4}{27} (\operatorname{wid} \boldsymbol{x})^2.
\end{aligned}$$

Combining the above cases,

$$\sup_{z \in \boldsymbol{x}} (u(z) - u^{\mathrm{cv}}(\boldsymbol{x}, z)) \leq \tfrac{4}{27} (\operatorname{wid} \boldsymbol{x})^2, \qquad \forall \boldsymbol{x} \in \mathbb{IR},$$

as required. □

Lemma 9.2.40. *For fixed $k \in \mathbb{N}$, consider the relaxation scheme $\{u^{\mathrm{cv}}(\boldsymbol{x}, \cdot), u^{\mathrm{cc}}(\boldsymbol{x}, \cdot)\}_{\boldsymbol{x} \in \mathbb{IR}}$*

for $u : z \mapsto z^{2k+1}$ on $\mathbb{R}$ described in Example 9.2.29. This scheme satisfies Assumption 9.2.38.

Proof. It will be shown that the convex relaxations $u^{\mathrm{cv}}(\boldsymbol{x}, \cdot)$ of u are pointwise convergent of order 2; a similar argument applies to the concave relaxations $u^{\mathrm{cc}}(\boldsymbol{x}, \cdot)$.

Consider any fixed interval $\boldsymbol{q} \in \mathbb{IR}$, and any $\boldsymbol{x} \in \mathbb{I}\boldsymbol{q}$. If $i = 1$ or $0 \notin \boldsymbol{x}$, then $u^{\mathrm{cv}}(\boldsymbol{x}, \cdot)$ is the convex envelope of u on $\boldsymbol{x}$, which, by [11, Theorem 10], is pointwise convergent of order 2 with respect to $\boldsymbol{x}$.

If $i = 2$ and $0 \in \boldsymbol{x}$, then, noting that $u^{\mathrm{cv}}(\boldsymbol{x}, \cdot)$ is increasing, we obtain:

$$\begin{aligned}
\sup_{z \in \boldsymbol{x}} (u(z) - u^{\mathrm{cv}}(\boldsymbol{x}, z)) &\leq \sup_{z \in \boldsymbol{x}} (u(z) - u^{\mathrm{cv}}(\boldsymbol{x}, \underline{x})) \\
&= \sup_{z \in \boldsymbol{x}} (z^{2k+1} - \underline{x}^{2k+1}) \\
&\leq (\overline{x} - \underline{x})^{2k+1} + (\overline{x} - \underline{x})^{2k+1} \\
&\leq 2(\operatorname{wid} \boldsymbol{q})^{2k-1} (\overline{x} - \underline{x})^2.
\end{aligned}$$

The above results together show that $u^{\mathrm{cv}}(\boldsymbol{x}, \cdot)$ is pointwise convergent of order 2 to u with respect to $\boldsymbol{x}$, as required. □

9.3 Smoothing constructions

This section establishes basic properties of certain $\mathcal{C}^1$ and $\mathcal{C}^2$ relaxations of simple nonsmooth functions such as $z \mapsto \max\{z, 0\}$ and $(x, y) \mapsto \max\{x, y\}$, and uses these to construct variants of McCormick's multiplication rule. These rules will be shown in subsequent sections to have various desirable properties.

9.3.1 Relaxing simple nonsmooth functions

Definition 9.3.1. *Define functions $\mu_1, \mu_2 : \mathbb{R} \to \mathbb{R}$ as follows:*

$$\mu_1 : y \mapsto \begin{cases} 0 & \text{if } y \leq 0, \\ \frac{1}{4} y^2 & \text{if } 0 < y < 2, \\ y - 1 & \text{if } 2 \leq y, \end{cases} \qquad \mu_2 : y \mapsto \begin{cases} 0 & \text{if } y \leq 0, \\ \frac{1}{16} y^3 (4 - y) & \text{if } 0 < y < 2, \\ y - 1 & \text{if } 2 \leq y. \end{cases}$$

Observe that μ_1 is a member of the family of functions considered in [23, Example 11.8.11(c)]. In this chapter, in the spirit of [8], [23, Section 11.8] and [9, Section 1.10], μ_1 and μ_2 essentially serve as analogs of the mapping $y \mapsto \max\{y, 0\}$ which exhibit several useful properties. Ultimately, μ_1 will be used to construct $\mathcal{C}^1$ analogs of McCormick relaxations, and μ_2 will be used to construct $\mathcal{C}^2$ relaxations. By inspection, μ_1 and μ_2 are each $\mathcal{C}^1$, with

$$\nabla\mu_1 : y \mapsto \begin{cases} 0 & \text{if } y \leq 0, \\ \frac{1}{2}y & \text{if } 0 < y < 2, \\ 1 & \text{if } 2 \leq y, \end{cases} \qquad \nabla\mu_2 : y \mapsto \begin{cases} 0 & \text{if } y \leq 0, \\ \frac{1}{4}y^2(3-y) & \text{if } 0 < y < 2, \\ 1 & \text{if } 2 \leq y. \end{cases} \tag{9.5}$$

The above expressions show that $\mu_1(y)$, $\mu_2(y)$, $\nabla\mu_1(y)$, and $\nabla\mu_2(y)$ are each nonnegative for each $y \in \mathbb{R}$, noting that $3 - y > 0$ when $0 < y < 2$. Thus, μ_1 and μ_2 are increasing on $\mathbb{R}$. Moreover, μ_2 is $\mathcal{C}^2$, with

$$\nabla^2\mu_2 : y \mapsto \begin{cases} 0 & \text{if } y \leq 0, \\ \frac{3}{4}y(2-y) & \text{if } 0 < y < 2, \\ 0 & \text{if } 2 \leq y. \end{cases} \tag{9.6}$$

Lemma 9.3.2. *For each $i \in \{1, 2\}$ and each $y \in \mathbb{R}$, $\max\{y-1, 0\} \leq \mu_i(y) \leq \max\{y, 0\}$.*

Proof. The cases in which $y \leq 0$, $0 < y \leq 1$, $1 < y < 2$, and $2 \leq y$ will be considered separately.

If $y \leq 0$, then, for each $i \in \{1, 2\}$,

$$\max\{y-1, 0\} = 0 = \mu_i(y) = \max\{y, 0\}.$$

If $0 < y \leq 1$, then, noting that $3 \leq 4 - y < 4$, the following inequality chains are satisfied:

$$\begin{aligned} &\max\{y-1, 0\} = 0 < \tfrac{1}{4}y^2 = \mu_1(y) = y \cdot \tfrac{y}{4} < y = \max\{y, 0\}, \\ &\max\{y-1, 0\} = 0 < \tfrac{1}{16}y^3(4-y) = \mu_2(y) \leq y(\tfrac{1}{16} \cdot 1^2 \cdot 4) < y = \max\{y, 0\}. \end{aligned}$$

If $1 < y < 2$, then, noting that $4 - y^2 > 0$ and $4 - y < 3$, the following inequality chains are satisfied:

$\max\{y-1,0\} = y-1 = \frac{1}{4}y^2 - \frac{1}{4}(y-2)^2 \leq \frac{1}{4}y^2 = \mu_1(y) = y \cdot \frac{y}{4} < y = \max\{y,0\}$,
$\max\{y-1,0\} = y-1 = \frac{1}{16}y^3(4-y) - \frac{1}{16}(y-2)^2(4-y^2) \leq \frac{1}{16}y^3(4-y) = \mu_2(y)$,
$\mu_2(y) = y(\frac{1}{16}y^2(4-y)) \leq y(\frac{1}{16} \cdot 2^2 \cdot 3) < y = \max\{y,0\}$.

If $2 \leq y$, then, for each $i \in \{1,2\}$,

$$\max\{y-1,0\} = y-1 = \mu_i(y) < y = \max\{y,0\}.$$

Thus, the required inequalities have been demonstrated for each $i \in \{1,2\}$ and each $y \in \mathbb{R}$. □

Lemma 9.3.3. *The functions μ_1 and μ_2 are convex on $\mathbb{R}$.*

Proof. Inspection of (9.5) and (9.6) shows that $\nabla\mu_1$ is increasing on $\mathbb{R}$, and that $\nabla^2\mu_2(y)$ is nonnegative for each $y \in \mathbb{R}$. The convexity of μ_1 and μ_2 follows immediately. □

Definition 9.3.4. *For each $i \in \{1,2\}$, define functions $\gamma_i, \sigma_i : \mathbb{R} \times \mathbb{R} \times [0,+\infty) \to \mathbb{R}$ as follows:*

$$\gamma_i : (z,a,p) \mapsto \begin{cases} \max\{z,a\} & \text{if } p = 0, \\ a + p\,\mu_i(\frac{z-a}{p}) & \text{if } p > 0, \end{cases}$$
$$\sigma_i : (z,b,p) \mapsto \begin{cases} \min\{z,b\} & \text{if } p = 0, \\ b - p\,\mu_i(\frac{b-z}{p}) & \text{if } p > 0, \end{cases}$$

and define functions $\nu_i, \lambda_i : \mathbb{R} \times \mathbb{R} \times [0,+\infty) \to \mathbb{R}$ as follows:

$$\nu_i : (x,y,p) \mapsto \tfrac{1}{2}(\gamma_i(x,y,p) + \gamma_i(y,x,p)),$$
$$\lambda_i : (x,y,p) \mapsto \tfrac{1}{2}(\sigma_i(x,y,p) + \sigma_i(y,x,p)).$$

Useful properties of γ_i, σ_i, ν_i, and λ_i will be established below. Intuitively, throughout this chapter, $\gamma_i(z,a,p)$ plays a similar role to $\max\{z,a\}$ for fixed a, $\sigma_i(z,b,p)$ is analogous to $\min\{z,b\}$ for fixed b, $\nu_i(x,y,p)$ is analogous to $\max\{x,y\}$

for varying x and y, and $\lambda_i(x, y, p)$ is analogous to $\min\{x, y\}$ for varying x and y. Roughly, the parameter p quantifies the extent to which γ_i and σ_i are relaxed to yield a differentiable underestimator of $\max\{\cdot, a\}$ and a differentiable overestimator of $\min\{\cdot, b\}$, as formalized in the following lemma.

Lemma 9.3.5. *Consider any fixed $i \in \{1, 2\}$, $a, b \in \mathbb{R}$, and $p \geq 0$. The mapping $\gamma_i(\cdot, a, p)$ is convex and increasing. Moreover,*

$$a \leq \max\{z - p, a\} \leq \gamma_i(z, a, p) \leq \max\{z, a\}, \qquad \forall z \in \mathbb{R}.$$

Similarly, the mapping $\sigma_i(\cdot, b, p)$ is concave and increasing, with

$$\min\{z, b\} \leq \sigma_i(z, b, p) \leq \min\{z + p, b\} \leq b, \qquad \forall z \in \mathbb{R}.$$

If $p > 0$, then $\gamma_i(\cdot, a, p)$ and $\sigma_i(\cdot, b, p)$ are both $\mathcal{C}^i$.

Proof. If $p = 0$, then $\gamma_i(z, a, p) = \max\{z, a\}$ and $\sigma_i(z, b, p) = \min\{z, b\}$ for each $z \in \mathbb{R}$, from which the required results follow immediately.

If $p > 0$, then, for any $a \in \mathbb{R}$, $\gamma_i(\cdot, a, p)$ is a translated and dilated version of μ_i. Hence, the required results concerning γ_i follow immediately from Lemmas 9.3.2 and 9.3.3, and the fact that μ_i is $\mathcal{C}^i$. The required results concerning σ_i are demonstrated similarly. □

Proposition 9.3.6. *Consider any fixed $i \in \{1, 2\}$, $a, b \in \mathbb{R}$ and $p > 0$. Gradients of the mappings $z \mapsto \gamma_i(z, a, p)$ and $z \mapsto \sigma_i(z, b, p)$ at some $z_0 \in \mathbb{R}$ may be computed using* (9.5) *as follows.*

$$\frac{\partial \gamma_i}{\partial z}(z_0, a, p) = \nabla \mu_i\left(\tfrac{z_0 - a}{p}\right), \qquad \frac{\partial \sigma_i}{\partial z}(z_0, b, p) = \nabla \mu_i\left(\tfrac{b - z_0}{p}\right).$$

Lemma 9.3.7. *Given triples $(z_1, a_1, p_1), (z_2, a_2, p_2) \in \mathbb{R} \times \mathbb{R} \times [0, +\infty)$, suppose that $z_1 \leq z_2$, $a_1 \leq a_2$, and $p_1 \geq p_2$. Then, for each $i \in \{1, 2\}$, $\gamma_i(z_1, a_1, p_1) \leq \gamma_i(z_2, a_2, p_2)$.*

Similarly, given triples $(z_1, b_1, p_1), (z_2, b_2, p_2) \in \mathbb{R} \times \mathbb{R} \times [0, +\infty)$, suppose that $z_1 \geq z_2$, $b_1 \geq b_2$, and $p_1 \geq p_2$. Then, for each $i \in \{1, 2\}$, $\sigma_i(z_1, b_1, p_1) \geq \sigma_i(z_2, b_2, p_2)$.

Proof. The result concerning γ_i will be shown; the result concerning σ_i is analogous. Observe that

$$\begin{aligned}\gamma_i(z_2, a_2, p_2) - \gamma_i(z_1, a_1, p_1) = (\gamma_i(z_2, a_2, p_2) - \gamma_i(z_1, a_1, p_2)) \\ + (\gamma_i(z_1, a_1, p_2) - \gamma_i(z_1, a_1, p_1));\end{aligned}$$

it suffices to show that each parenthetical term in the right-hand side of the above equation is nonnegative. If $p_2 = 0$, then

$$\gamma_i(z_2, a_2, p_2) - \gamma_i(z_1, a_1, p_2) = \max\{z_2, a_2\} - \max\{z_1, a_1\} \geq 0.$$

Otherwise, if $p_2 > 0$, then

$$\begin{aligned}&\gamma_i(z_2, a_2, p_2) - \gamma_i(z_1, a_1, p_2) \\ &\quad = (\gamma_i(z_2, a_2, p_2) - \gamma_i(z_2, a_1, p_2)) + (\gamma_i(z_2, a_1, p_2) - \gamma_i(z_1, a_1, p_2)), \\ &\quad = \int_{a_1}^{a_2} \frac{\partial \gamma_i}{\partial a}(z_2, s, p_2)\, ds + p_2 \left(\mu_i(\tfrac{z_2 - a_1}{p_2}) - \mu_i(\tfrac{z_1 - a_1}{p_2})\right), \\ &\quad = \int_{a_1}^{a_2} \left(1 - \nabla \mu_i(\tfrac{z_2 - s}{p_2})\right)\, ds + p_2 \left(\mu_i(\tfrac{z_2 - a_1}{p_2}) - \mu_i(\tfrac{z_1 - a_1}{p_2})\right), \\ &\quad \geq 0;\end{aligned}$$

to obtain the final inequality, note that (9.5) shows that $\nabla \mu_i(x) \leq 1$ for each x, and so the integrand in the integral above is nonnegative. The non-integral term is also nonnegative, since $p_2 > 0$, $z_2 \geq z_1$, and μ_i is increasing.

It remains to be shown that $\gamma_i(z_1, a_1, p_2) \geq \gamma_i(z_1, a_1, p_1)$. This is trivial if $p_2 = p_1$. Otherwise, either $p_1 > p_2 = 0$ or $p_1 > p_2 > 0$; these two cases will be considered separately. If $p_1 > p_2 = 0$, then Lemma 9.3.5 yields:

$$\gamma_i(z_1, a_1, p_2) = \max\{z_1, a_1\} \geq \gamma_i(z_1, a_1, p_1),$$

as required. If $p_1 > p_2 > 0$, then $\frac{z_1 - a_1}{p_2} \geq \frac{z_1 - a_1}{p_1}$, and so, since μ_i is increasing,

$$\gamma_i(z_1, a_1, p_2) - \gamma_i(z_1, a_1, p_1) = \mu_i(\tfrac{z_1 - a_1}{p_2}) - \mu_i(\tfrac{z_1 - a_1}{p_1}) \geq 0,$$

as required. □

Lemma 9.3.8. *For any fixed $p > 0$ and $i \in \{1,2\}$, the mappings $(x,y) \mapsto \nu_i(x,y,p)$ and $(x,y) \mapsto \lambda_i(x,y,p)$ are each $\mathcal{C}^i$ on $\mathbb{R}^2$.*

Proof. With $p > 0$, for each $x, y \in \mathbb{R}$,

$$\begin{aligned} \nu_i(x,y,p) &= \frac{1}{2}\left(x + y + p(\mu_i(\tfrac{x-y}{p}) + \mu_i(\tfrac{y-x}{p}))\right), \\ \text{and} \qquad \lambda_i(x,y,p) &= \frac{1}{2}\left(x + y - p(\mu_i(\tfrac{x-y}{p}) + \mu_i(\tfrac{y-x}{p}))\right). \end{aligned}$$

Noting that μ_i is $\mathcal{C}^i$ then yields the required result. □

Proposition 9.3.9. *Consider any fixed $i \in \{1,2\}$ and $p > 0$. Partial derivatives of the mappings $(x,y) \mapsto \nu_i(x,y,p)$ and $(x,y) \mapsto \lambda_i(x,y,p)$ at some $x_0, y_0 \in \mathbb{R}$ may be computed using* (9.5) *as follows.*

$$\begin{aligned} \frac{\partial \nu_i}{\partial x}(x_0,y_0,p) &= \frac{\partial \lambda_i}{\partial y}(x_0,y_0,p) = \frac{1}{2}\left(1 + \nabla\mu_i(\tfrac{x_0-y_0}{p}) - \nabla\mu_i(\tfrac{y_0-x_0}{p})\right), \\ \frac{\partial \nu_i}{\partial y}(x_0,y_0,p) &= \frac{\partial \lambda_i}{\partial x}(x_0,y_0,p) = \frac{1}{2}\left(1 - \nabla\mu_i(\tfrac{x_0-y_0}{p}) + \nabla\mu_i(\tfrac{y_0-x_0}{p})\right). \end{aligned}$$

Proof. This result follows immediately from the previous lemma. □

Lemma 9.3.10. *Given triples $(x_1,y_1,p_1), (x_2,y_2,p_2) \in \mathbb{R} \times \mathbb{R} \times [0,+\infty)$, suppose that $x_1 \le x_2$, $y_1 \le y_2$, and $p_1 \ge p_2$. Then, for each $i \in \{1,2\}$, $\nu_i(x_1,y_1,p_1) \le \nu_i(x_2,y_2,p_2)$.*

Similarly, given triples $(x_3,y_3,p_3), (x_4,y_4,p_4) \in \mathbb{R} \times \mathbb{R} \times [0,+\infty)$, suppose that $x_3 \ge x_4$, $y_3 \ge y_4$, and $p_3 \ge p_4$. Then, for each $i \in \{1,2\}$, $\lambda_i(x_3,y_3,p_3) \ge \lambda_i(x_4,y_4,p_4)$.

Proof. The required result follows immediately from Lemma 9.3.7 and the definitions of ν_i and λ_i. □

Lemma 9.3.11. *Given $p \ge 0$ and $i \in \{1,2\}$, the mapping $(x,y) \mapsto \nu_i(x,y,p)$ is convex on $\mathbb{R}^2$, and the mapping $(x,y) \mapsto \lambda_i(x,y,p)$ is concave on $\mathbb{R}^2$. Moreover, for all $x, y \in \mathbb{R}$,*

$$\begin{aligned} &\tfrac{1}{2}(x+y) \le \tfrac{1}{2}(\max\{x-p,y\} + \max\{x,y-p\}) \le \nu_i(x,y,p) \le \max\{x,y\}, \\ \textit{and} \qquad &\min\{x,y\} \le \lambda_i(x,y,p) \le \tfrac{1}{2}(\min\{x+p,y\} + \min\{x,y+p\}) \le \tfrac{1}{2}(x+y). \end{aligned}$$

Proof. It will be shown that the mapping $(x, y) \mapsto \nu_i(x, y, p)$ is convex; a similar argument shows that $(x, y) \mapsto \lambda_i(x, y, p)$ is concave. Choose any $(x_A, y_A), (x_B, y_B) \in \mathbb{R}^2$ and any $\ell \in (0, 1)$, and define $\tilde{x} := \ell x_A + (1 - \ell) x_B$ and $\tilde{y} := \ell y_A + (1 - \ell) y_B$. The cases in which $p = 0$ and $p > 0$ will be considered separately.

If $p = 0$, then $\nu_i(x_j, y_j, p) = \max\{x_j, y_j\}$ for each $j \in \{A, B\}$; the convexity of the bivariate max function then implies that $(x, y) \mapsto \nu_i(x, y, p)$ is convex.

If $p > 0$, then, noting that $\tilde{x} - \tilde{y} = \ell(x_A - y_A) + (1 - \ell)(x_B - y_B)$, the convexity of μ_i implies that

$$\ell\, \mu_i(\tfrac{x_A - y_A}{p}) + (1 - \ell)\, \mu_i(\tfrac{x_B - y_B}{p}) \geq \mu_i(\tfrac{\tilde{x} - \tilde{y}}{p});$$

multiplying both sides of the above inequality by p and adding $\tilde{y}$ yields

$$\ell\, \gamma_i(x_A, y_A, p) + (1 - \ell)\, \gamma_i(x_B, y_B, p) \geq \gamma_i(\tilde{x}, \tilde{y}, p). \tag{9.7}$$

Since x_A, x_B, y_A, y_B were chosen arbitrarily, interchanging x_j with y_j for each $j \in \{A, B\}$ in the above argument yields

$$\ell\, \gamma_i(y_A, x_A, p) + (1 - \ell)\, \gamma_i(y_B, x_B, p) \geq \gamma_i(\tilde{y}, \tilde{x}, p);$$

adding this inequality to (9.7) and multiplying the result by $\frac{1}{2}$ yields

$$\ell\, \nu_i(x_A, y_A, p) + (1 - \ell)\, \nu_i(x_B, y_B, p) \geq \nu_i(\tilde{x}, \tilde{y}, p),$$

which shows that $(x, y) \mapsto \nu_i(x, y, p)$ is convex, as required.

The remaining claims of the lemma follow immediately from Lemma 9.3.5 and the definitions of ν_i and λ_i. □

Definition 9.3.12. *Define a function $p : \mathbb{IR} \to [0, +\infty)$ such that for some constant $a_p > 0$, $p(\boldsymbol{x}) := a_p(\operatorname{wid} \boldsymbol{x})^2$ for each $\boldsymbol{x} \in \mathbb{IR}$. Denote $p(\boldsymbol{x})$ as $p_{\boldsymbol{x}}$.*

This particular quadratic expression for p is irrelevant to the results developed in Sections 9.5 and 9.6 below; the results in these sections remain valid if p is redefined so that $p(\boldsymbol{x}) := \pi(\operatorname{wid} \boldsymbol{x})$, where $\pi : [0, +\infty) \to [0, +\infty)$ is any particular strictly-increasing function for which $\pi(0) = 0$. Defining $\pi : z \mapsto a_p z^2$, however,

yields the convergence results obtained in Section 9.7. The particular choice of the constant a_p does not affect the theoretical results developed in this chapter; appropriate choices of a_p will be discussed in Section 9.8.1 from a numerical standpoint.

Remark 9.3.13. *As claimed earlier, the condition in Assumption 9.2.21 that both $\underline{\tilde{u}}(\boldsymbol{x}) \leq u^{\mathrm{cv}}(\boldsymbol{x}, \zeta_u^{\min}(\boldsymbol{x}))$ and $\overline{\tilde{u}}(\boldsymbol{x}) \geq u^{\mathrm{cc}}(\boldsymbol{x}, \zeta_u^{\max}(\boldsymbol{x}))$ can be imposed without loss of generality. If this condition either fails or is not known to be true, then, for each $\boldsymbol{x} \in \mathbb{I}B$, $u^{\mathrm{cv}}(\boldsymbol{x}, \cdot)$ can be replaced with the mapping $z \mapsto \gamma_i(u^{\mathrm{cv}}(\boldsymbol{x}, z), \underline{\tilde{u}}(\boldsymbol{x}), p_{\boldsymbol{x}})$, and $u^{\mathrm{cc}}(\boldsymbol{x}, \cdot)$ can be replaced with the mapping $z \mapsto \sigma_i(u^{\mathrm{cc}}(\boldsymbol{x}, z), \overline{\tilde{u}}(\boldsymbol{x}), p_{\boldsymbol{x}})$; these replacements now satisfy the condition. The established properties of γ_i and σ_i ensure that the other conditions required of u^{cv} and u^{cc} by Definition 9.2.12 and Assumption 9.2.21 continue to hold under these replacements.*

9.3.2 Relaxing intersections of bounds and relaxations

Roughly, for each $i \in \{1, 2\}$, the $\mathcal{S}\mathrm{qu}_i$ and $\boldsymbol{belt}_i$ operations introduced in this section are $\mathcal{C}^i$ relaxations of the "Cut" and "Enc" operations presented in Definitions 2.4.3 and 2.4.5 of [100], and serve analogous roles. It will be shown in this section that $\mathcal{S}\mathrm{qu}_i$ is a relaxation function of the identity function on $\mathbb{R}$. Further in this chapter, Lemma 9.7.1 will show that $\mathcal{S}\mathrm{qu}_i$ is $(1, 2)$-convergent. Intuitively, $\mathcal{S}\mathrm{qu}_i$ also inherits the $\mathcal{C}^i$ nature of γ_i and σ_i; this property will be exploited in Section 9.6.

Definition 9.3.14. *For each $\mathcal{X} \in \mathbb{M}\mathbb{R}$ and each $i \in \{1, 2\}$, define a* belt operation *$\boldsymbol{belt}_i(\mathcal{X}) \in \mathbb{I}\mathbb{R}$ as follows:*

$$\boldsymbol{belt}_i(\mathcal{X}) := \begin{cases} [x, x] & \text{if } \underline{x}^{\mathrm{B}} = \overline{x}^{\mathrm{B}} =: x, \\ [\gamma_i(\underline{x}^{\mathrm{C}}, \underline{x}^{\mathrm{B}}, p_{\boldsymbol{x}^{\mathrm{B}}}), \sigma_i(\overline{x}^{\mathrm{C}}, \overline{x}^{\mathrm{B}}, p_{\boldsymbol{x}^{\mathrm{B}}})] & \text{if } \underline{x}^{\mathrm{B}} < \overline{x}^{\mathrm{B}}. \end{cases}$$

Define a squashing operation *$\mathcal{S}\mathrm{qu}_i(\mathcal{X}) := (\boldsymbol{x}^{\mathrm{B}}, \boldsymbol{belt}_i(\mathcal{X})) \in \mathbb{I}\mathbb{R}^2$. Given a vector $\mathcal{Y} \in \mathbb{M}\mathbb{R}^n$, define*

$$\mathcal{S}\mathrm{qu}_i(\mathcal{Y}) := \begin{bmatrix} \mathcal{S}\mathrm{qu}_i(\mathcal{Y}_1) \\ \vdots \\ \mathcal{S}\mathrm{qu}_i(\mathcal{Y}_n) \end{bmatrix} \in (\mathbb{I}\mathbb{R}^2)^n.$$

For any $\mathcal{X} \in \mathbb{MR}$, Lemma 9.3.5 implies that $\boldsymbol{x}^{\mathrm{B}} \cap \boldsymbol{x}^{\mathrm{C}} \subset \boldsymbol{belt}_i(\mathcal{X}) \subset \boldsymbol{x}^{\mathrm{B}}$, which in turn yields $\mathcal{S}\mathrm{qu}_i(\mathcal{X}) \in \mathbb{MR}_{\mathrm{prop}}$. Thus, $\mathcal{S}\mathrm{qu}_i(\mathcal{Y}) \in \mathbb{MR}^n_{\mathrm{prop}}$ for any $\mathcal{Y} \in \mathbb{MR}^n$. Furthermore, observe that $\boldsymbol{belt}_i(([x,x],[x,x])) = [x,x]$ for each $x \in \mathbb{R}$.

Lemma 9.3.15. *For each $i \in \{1,2\}$, for each coherent pair $\mathcal{X}, \mathcal{Y} \in \mathbb{MR}$ and each $\ell \in [0,1]$,*

$$\boldsymbol{belt}_i(\mathcal{C}\mathrm{onv}(\ell, \mathcal{X}, \mathcal{Y})) \supset \ell\, \boldsymbol{belt}_i(\mathcal{X}) + (1-\ell)\, \boldsymbol{belt}_i(\mathcal{Y}).$$

Moreover, $\mathcal{S}\mathrm{qu}_i$ is coherently concave for each $i \in \{1,2\}$.

Proof. Since $\mathcal{X}$ and $\mathcal{Y}$ are coherent, define $\boldsymbol{z} := \boldsymbol{x}^{\mathrm{B}} = \boldsymbol{y}^{\mathrm{B}}$. The following pair of inequalities is obtained from the convexity of $\gamma_i(\cdot, \underline{z}^{\mathrm{B}}, p_{\boldsymbol{z}})$ and the concavity of $\sigma_i(\cdot, \overline{z}^{\mathrm{B}}, p_{\boldsymbol{z}})$:

$$\begin{aligned}
\gamma_i(\ell \underline{x}^{\mathrm{C}} + (1-\ell)\underline{y}^{\mathrm{C}}, \underline{z}^{\mathrm{B}}, p_{\boldsymbol{z}}) &\leq \ell\, \gamma_i(\underline{x}^{\mathrm{C}}, \underline{z}^{\mathrm{B}}, p_{\boldsymbol{z}}) + (1-\ell)\, \gamma_i(\underline{y}^{\mathrm{C}}, \underline{z}^{\mathrm{B}}, p_{\boldsymbol{z}}), \\
\sigma_i(\ell \overline{x}^{\mathrm{C}} + (1-\ell)\overline{y}^{\mathrm{C}}, \overline{z}^{\mathrm{B}}, p_{\boldsymbol{z}}) &\geq \ell\, \sigma_i(\overline{x}^{\mathrm{C}}, \overline{z}^{\mathrm{B}}, p_{\boldsymbol{z}}) + (1-\ell)\, \sigma_i(\overline{y}^{\mathrm{C}}, \overline{z}^{\mathrm{B}}, p_{\boldsymbol{z}}),
\end{aligned}$$

which are equivalent to the required inclusion. Moreover, since $\mathcal{X}$, $\mathcal{Y}$, and ℓ were chosen arbitrarily, it follows immediately that $\mathcal{S}\mathrm{qu}_i$ is coherently concave. □

Lemma 9.3.16. *For each $i \in \{1,2\}$, $\boldsymbol{belt}_i$ and $\mathcal{S}\mathrm{qu}_i$ are inclusion monotonic.*

Proof. Consider any $\mathcal{X}, \mathcal{Y} \in \mathbb{MR}$ for which $\mathcal{X} \subset \mathcal{Y}$. If $\underline{x}^{\mathrm{B}} = \overline{x}^{\mathrm{B}} =: x$, then $\boldsymbol{x}^{\mathrm{B}} \cap \boldsymbol{x}^{\mathrm{C}} \neq \emptyset$ implies $x \in \boldsymbol{x}^{\mathrm{C}}$. Thus, $\mathcal{X} \subset \mathcal{Y}$ implies

$$\boldsymbol{belt}_i(\mathcal{X}) = [x,x] = \boldsymbol{x}^{\mathrm{B}} = \boldsymbol{x}^{\mathrm{B}} \cap \boldsymbol{x}^{\mathrm{C}} \subset \boldsymbol{y}^{\mathrm{B}} \cap \boldsymbol{y}^{\mathrm{C}} \subset \boldsymbol{belt}_i(\mathcal{Y}),$$

as required. If $\underline{x}^{\mathrm{B}} < \overline{x}^{\mathrm{B}}$, then since $\underline{x}^{\mathrm{C}} \geq \underline{y}^{\mathrm{C}}$, $\underline{x}^{\mathrm{B}} \geq \underline{y}^{\mathrm{B}}$, and $p_{\boldsymbol{x}^{\mathrm{B}}} \leq p_{\boldsymbol{y}^{\mathrm{B}}}$, Lemma 9.3.7 implies that

$$\gamma_i(\underline{x}^{\mathrm{C}}, \underline{x}^{\mathrm{B}}, p_{\boldsymbol{x}^{\mathrm{B}}}) \geq \gamma_i(\underline{y}^{\mathrm{C}}, \underline{y}^{\mathrm{B}}, p_{\boldsymbol{y}^{\mathrm{B}}}).$$

A similar argument shows that

$$\sigma_i(\overline{x}^{\mathrm{C}}, \overline{x}^{\mathrm{B}}, p_{\boldsymbol{x}^{\mathrm{B}}}) \leq \sigma_i(\overline{y}^{\mathrm{C}}, \overline{y}^{\mathrm{B}}, p_{\boldsymbol{y}^{\mathrm{B}}}),$$

and so $\boldsymbol{belt}_i(\mathcal{X}) \subset \boldsymbol{belt}_i(\mathcal{Y})$. The inclusion $\mathcal{S}\mathrm{qu}_i(\mathcal{X}) \subset \mathcal{S}\mathrm{qu}_i(\mathcal{Y})$ follows immediately. □

Lemma 9.3.17. *For each fixed $i \in \{1,2\}$ and $\boldsymbol{x}^{\mathrm{B}} := [\underline{x}^{\mathrm{B}}, \overline{x}^{\mathrm{B}}] \in \mathbb{IR}$, consider the interval-valued mapping $\boldsymbol{y} : (\underline{\xi}, \overline{\xi}) \mapsto \boldsymbol{belt}_i((\boldsymbol{x}^{\mathrm{B}}, [\underline{\xi}, \overline{\xi}]))$. The mappings $\underline{y}$ and $\overline{y}$ are both $\mathcal{C}^i$ on $\mathbb{R}^2$.*

Proof. If $\underline{x}^{\mathrm{B}} = \overline{x}^{\mathrm{B}}$, then the mapping $\boldsymbol{belt}_i((\boldsymbol{x}^{\mathrm{B}}, \cdot))$ is constant, and is therefore $\mathcal{C}^i$. Otherwise, if $\underline{x}^{\mathrm{B}} < \overline{x}^{\mathrm{B}}$, then the required result follows immediately from Lemma 9.3.5 and Definition 9.3.12. □

Remark 9.3.18. *It follows from the above definitions and lemmata that, for each $i \in \{1,2\}$, $\mathcal{S}\mathrm{qu}_i$ is a relaxation function for the identity mapping $\mathbf{x} \in \mathbb{R}^n \mapsto \mathbf{x}$.*

9.3.3 Relaxing multiplication

Throughout this section, let the value of $i \in \{1,2\}$ be fixed. Setting $i = 1$ will yield $\mathcal{C}^1$ relaxations; setting $i = 2$ will yield $\mathcal{C}^2$ relaxations, but will place stricter requirements on the univariate intrinsic functions considered, as formalized in Assumption 9.2.21.

The following definition replaces Definition 9.2.31; it will be shown in this chapter that this replacement weakens McCormick's classical multiplication operation to yield an alternative that is $\mathcal{C}^i$, while maintaining $(1,2)$-convergence. This modified multiplication operation depends on i, but this dependence will not be reflected in its "$\mathcal{X}\mathcal{Y}$" notation.

Definition 9.3.19. *Define a multiplication operation $\times_i : \mathbb{MR}^2_{\mathrm{prop}} \to \mathbb{MR}$ so that, for each $\mathcal{X}, \mathcal{Y} \in \mathbb{MR}$,*

$$\times_i(\mathcal{X}, \mathcal{Y}) \equiv \mathcal{X}\mathcal{Y} := \mathcal{S}\mathrm{qu}_i((\boldsymbol{x}^{\mathrm{B}}\boldsymbol{y}^{\mathrm{B}}, \boldsymbol{z})),$$

where $\boldsymbol{z} \equiv [\underline{z}, \overline{z}] \in \mathbb{IR}$ is defined so that:

$$\underline{z} := \nu_i \left(\underline{(\underline{y}^{\mathrm{B}}\boldsymbol{x}^{\mathrm{C}})} + \underline{(\underline{x}^{\mathrm{B}}\boldsymbol{y}^{\mathrm{C}})} - \underline{x}^{\mathrm{B}}\underline{y}^{\mathrm{B}},\ \underline{(\overline{y}^{\mathrm{B}}\boldsymbol{x}^{\mathrm{C}})} + \underline{(\overline{x}^{\mathrm{B}}\boldsymbol{y}^{\mathrm{C}})} - \overline{x}^{\mathrm{B}}\overline{y}^{\mathrm{B}},\ p_{\boldsymbol{x}^{\mathrm{B}}\boldsymbol{y}^{\mathrm{B}}} \right),$$
$$\overline{z} := \lambda_i \left(\overline{(\underline{y}^{\mathrm{B}}\boldsymbol{x}^{\mathrm{C}})} + \overline{(\overline{x}^{\mathrm{B}}\boldsymbol{y}^{\mathrm{C}})} - \overline{x}^{\mathrm{B}}\underline{y}^{\mathrm{B}},\ \overline{(\overline{y}^{\mathrm{B}}\boldsymbol{x}^{\mathrm{C}})} + \overline{(\underline{x}^{\mathrm{B}}\boldsymbol{y}^{\mathrm{C}})} - \underline{x}^{\mathrm{B}}\overline{y}^{\mathrm{B}},\ p_{\boldsymbol{x}^{\mathrm{B}}\boldsymbol{y}^{\mathrm{B}}} \right).$$

9.3.4 Restrictions to proper McCormick objects

The following result shows that the codomains of $+ : \mathbb{MR}^2_{\mathrm{prop}} \to \mathbb{MR}$ (cf. Definition 9.2.30) and $\times_i : \mathbb{MR}^2_{\mathrm{prop}} \to \mathbb{MR}$ may be restricted to $\mathbb{MR}_{\mathrm{prop}}$ without loss of generality.

Proposition 9.3.20. *Consider any $\mathcal{X}, \mathcal{Y} \in \mathbb{MR}_{\mathrm{prop}}$, and define $\mathcal{S} := \mathcal{X} + \mathcal{Y}$ and $\mathcal{P} := \mathcal{X}\mathcal{Y}$ for some $i \in \{1, 2\}$. Then, $\mathcal{S}, \mathcal{P} \in \mathbb{MR}_{\mathrm{prop}}$.*

Proof. Firstly, to show that $\mathcal{S} \in \mathbb{MR}_{\mathrm{prop}}$, observe that

$$\boldsymbol{s}^{\mathrm{C}} = [\underline{s}^{\mathrm{C}}, \overline{s}^{\mathrm{C}}] = [\underline{x}^{\mathrm{C}}, \overline{x}^{\mathrm{C}}] + [\underline{y}^{\mathrm{C}}, \overline{y}^{\mathrm{C}}] \subset [\underline{x}^{\mathrm{B}}, \overline{x}^{\mathrm{B}}] + [\underline{y}^{\mathrm{B}}, \overline{y}^{\mathrm{B}}] = \boldsymbol{x}^{\mathrm{B}} + \boldsymbol{y}^{\mathrm{B}} = \boldsymbol{s}^{\mathrm{B}}.$$

Secondly, $\mathcal{P} = \mathcal{S}\mathrm{qu}_i((\boldsymbol{p}^{\mathrm{B}}, \boldsymbol{z}))$, with $\boldsymbol{z} \in \mathbb{IR}$ given as in Definition 9.3.19. Define $\boldsymbol{v} := \boldsymbol{x}^{\mathrm{B}} \cap \boldsymbol{x}^{\mathrm{C}}$ and $\boldsymbol{w} := \boldsymbol{y}^{\mathrm{B}} \cap \boldsymbol{y}^{\mathrm{C}}$. Since $\mathcal{X}, \mathcal{Y} \in \mathbb{MR}_{\mathrm{prop}}$, it follows that $\boldsymbol{v} = \boldsymbol{x}^{\mathrm{C}}$ and $\boldsymbol{w} = \boldsymbol{y}^{\mathrm{C}}$. Now, making use of Lemma 9.3.11, it follows that

$$\begin{aligned} &\underline{z} \leq \max \left(\underline{(\underline{y}^{\mathrm{B}}\boldsymbol{v})} + \underline{(\underline{x}^{\mathrm{B}}\boldsymbol{w})} - \underline{x}^{\mathrm{B}}\underline{y}^{\mathrm{B}},\ \underline{(\overline{y}^{\mathrm{B}}\boldsymbol{v})} + \underline{(\overline{x}^{\mathrm{B}}\boldsymbol{w})} - \overline{x}^{\mathrm{B}}\overline{y}^{\mathrm{B}} \right) \\ \text{and} \quad &\overline{z} \geq \min \left(\overline{(\underline{y}^{\mathrm{B}}\boldsymbol{v})} + \overline{(\overline{x}^{\mathrm{B}}\boldsymbol{w})} - \overline{x}^{\mathrm{B}}\underline{y}^{\mathrm{B}},\ \overline{(\overline{y}^{\mathrm{B}}\boldsymbol{v})} + \overline{(\underline{x}^{\mathrm{B}}\boldsymbol{w})} - \underline{x}^{\mathrm{B}}\overline{y}^{\mathrm{B}} \right). \end{aligned}$$

Defining

$$\begin{aligned} &\underline{q} := \max \left(\underline{(\underline{y}^{\mathrm{B}}\boldsymbol{v})} + \underline{(\underline{x}^{\mathrm{B}}\boldsymbol{w})} - \underline{x}^{\mathrm{B}}\underline{y}^{\mathrm{B}},\ \underline{(\overline{y}^{\mathrm{B}}\boldsymbol{v})} + \underline{(\overline{x}^{\mathrm{B}}\boldsymbol{w})} - \overline{x}^{\mathrm{B}}\overline{y}^{\mathrm{B}} \right), \\ \text{and} \quad &\overline{q} := \min \left(\overline{(\underline{y}^{\mathrm{B}}\boldsymbol{v})} + \overline{(\overline{x}^{\mathrm{B}}\boldsymbol{w})} - \overline{x}^{\mathrm{B}}\underline{y}^{\mathrm{B}},\ \overline{(\overline{y}^{\mathrm{B}}\boldsymbol{v})} + \overline{(\underline{x}^{\mathrm{B}}\boldsymbol{w})} - \underline{x}^{\mathrm{B}}\overline{y}^{\mathrm{B}} \right), \end{aligned}$$

it is argued on [100, Page 69] that $[\underline{q}, \overline{q}] \cap \boldsymbol{p}^{\mathrm{B}} \neq \emptyset$. Thus,

$$\boldsymbol{z} \cap \boldsymbol{p}^{\mathrm{B}} \supset [\underline{q}, \overline{q}] \cap \boldsymbol{p}^{\mathrm{B}} \neq \emptyset.$$

This shows that $(\boldsymbol{p}^{\mathrm{B}}, \boldsymbol{z}) \in \mathbb{MR}$, which implies that $\mathcal{P} = \mathcal{S}\mathrm{qu}_i((\boldsymbol{p}^{\mathrm{B}}, \boldsymbol{z})) \in \mathbb{MR}_{\mathrm{prop}}$. $\square$

The following result considers univariate intrinsic functions in the same manner as the above result, and shows that the codomains of their McCormick analogs may be restricted to $\mathbb{MR}_{\mathrm{prop}}$ without loss of generality.

Proposition 9.3.21. *Consider a univariate intrinsic function $u : B \subset \mathbb{R} \to \mathbb{R}$ that satisfies Assumption 9.2.21. With $\mathcal{U}$ described by Definition 9.2.32, $\mathcal{U}(\mathcal{X}) \in \mathbb{MR}_{\mathrm{prop}}$ for each $\mathcal{X} \in \mathbb{M}B_{\mathrm{prop}}$.*

Proof. By construction, $\zeta_u^{\min}(\boldsymbol{x}^{\mathrm{B}}) \in \boldsymbol{x}^{\mathrm{B}}$ and $\zeta_u^{\max}(\boldsymbol{x}^{\mathrm{B}}) \in \boldsymbol{x}^{\mathrm{B}}$. Since $\mathcal{X} \in \mathbb{M}B_{\mathrm{prop}}$, $\boldsymbol{x}^{\mathrm{C}} \subset \boldsymbol{x}^{\mathrm{B}}$; it follows that $\mathrm{mid}(\zeta_u^{\min}(\boldsymbol{x}^{\mathrm{B}}), \underline{x}^{\mathrm{C}}, \overline{x}^{\mathrm{C}}) \in \boldsymbol{x}^{\mathrm{B}}$ and $\mathrm{mid}(\zeta_u^{\max}(\boldsymbol{x}^{\mathrm{B}}), \underline{x}^{\mathrm{C}}, \overline{x}^{\mathrm{C}}) \in \boldsymbol{x}^{\mathrm{B}}$. It therefore follows from the bounds on u^{cv} and u^{cc} in Assumption 9.2.21 that $[\underline{u}^{\mathrm{C}}(\mathcal{X}), \overline{u}^{\mathrm{C}}(\mathcal{X})] \subset \tilde{\boldsymbol{u}}(\boldsymbol{x}^{\mathrm{B}})$, which implies that $\mathcal{U}(\mathcal{X}) \in \mathbb{MR}_{\mathrm{prop}}$. $\square$

9.4 Main theorem

The following definition is a variation of Definition 9.2.33. Observe that all univariate intrinsic functions listed in Table 9.2 satisfy Assumption 9.2.21 for each $i \in \{1, 2\}$, and that all of these functions except for the absolute-value function satisfy Assumption 9.2.38. The result following this definition is the main theorem of this chapter.

Definition 9.4.1. *Given some $i^* \in \{1, 2\}$ and an MC-factorable function $\mathbf{f} : B \subset \mathbb{R}^n \to \mathbb{R}^m$ whose composed univariate intrinsic functions each satisfy Assumption 9.2.21 with $i := i^*$, a* natural $\mathcal{C}^{i^*}$ McCormick extension $\mathcal{F} : \mathbb{M}B_{\mathrm{prop}} \to \mathbb{MR}^m$ *of $\mathbf{f}$ is defined by replacing each addition operation in the description of f with its McCormick counterpart described in Definition 9.2.30, each multiplication operation with its counterpart in Definition 9.3.19 with $i := i^*$, and each univariate intrinsic function with its counterpart in Definition 9.2.32.*

Define an unconstrained $\mathcal{C}^{i^*}$ McCormick extension *of $\mathbf{f}$ as $\mathcal{F}_{\mathrm{unc}} := \mathcal{F} \circ \mathcal{S}\mathrm{qu}_{i^*} : \mathbb{M}B \to \mathbb{MR}^m$.*

Theorem 9.4.2. *Given some $i^* \in \{1,2\}$ and an MC-factorable function $\mathbf{f} : B \subset \mathbb{R}^n \to \mathbb{R}^m$ whose composed univariate intrinsic functions each satisfy Assumption 9.2.21 with $i := i^*$, there are no domain violations in the construction of a natural $\mathcal{C}^{i^*}$ McCormick extension $\mathcal{F} : \mathbb{M}B_{\text{prop}} \to \mathbb{M}\mathbb{R}^m$ of $\mathbf{f}$ on B. The function $\mathcal{F}$ is a relaxation function for $\mathbf{f}$ on B. Additionally, if each univariate intrinsic function describing $\mathbf{f}$ satisfies Assumption 9.2.38, then $\mathcal{F}$ is $(1,2)$-convergent.*

Moreover, if $m = 1$, in which case $\mathbf{f} \equiv f$ is scalar-valued, then the functions $\phi_{\boldsymbol{x}}, \psi_{\boldsymbol{x}}$ defined by Proposition 9.2.20 in terms of $\mathcal{F}$ for each $\boldsymbol{x} \in \mathbb{I}B$ are each $\mathcal{C}^{i^*}$ on $\boldsymbol{x}$.

Proof. Since $\mathbf{f}$ is MC-factorable, it has a well-defined natural interval extension. Thus, Proposition 9.3.20, Proposition 9.3.21, and Assumption 9.2.21 imply that there are no domain violations in the construction of $\mathcal{F}$. The remaining claims of the theorem are proved separately as Theorems 9.5.1, 9.6.1, and 9.7.3 below. □

Roughly, an unconstrained $\mathcal{C}^{i^*}$ McCormick extension of a function $f : B \subset \mathbb{R}^n \to \mathbb{R}$ yields convex/concave relaxations of f that are weaker than those described by a natural $\mathcal{C}^{i^*}$ McCormick extension, yet are well-defined on all of $\mathbb{R}^n$ rather than particular interval subsets, and satisfy the following corollary. As a result, natural $\mathcal{C}^{i^*}$ McCormick extensions are preferable to unconstrained $\mathcal{C}^{i^*}$ McCormick extensions in general, since the former generate tighter relaxations. Unconstrained $\mathcal{C}^{i^*}$ McCormick extensions are preferable in two particular situations: firstly, if the problem $\min_{\mathbf{z} \in \boldsymbol{x}} \phi_{\boldsymbol{x}}(\mathbf{z})$ is solved using a constrained convex optimization method which visits infeasible points, and secondly, if generalized McCormick relaxations [104] are employed in a manner that permits inputs $\mathcal{X} \equiv (\boldsymbol{x}^{\text{B}}, \boldsymbol{x}^{\text{C}})$ for which $\boldsymbol{x}^{\text{C}} \nsubseteq \boldsymbol{x}^{\text{B}}$.

Corollary 9.4.3. *Given some $i^* \in \{1,2\}$ and an MC-factorable function $\mathbf{f} : B \subset \mathbb{R}^n \to \mathbb{R}^m$ whose composed univariate intrinsic functions each satisfy Assumption 9.2.21 with $i := i^*$, an unconstrained $\mathcal{C}^{i^*}$ McCormick extension $\mathcal{F}_{\text{unc}} : \mathbb{M}B \to \mathbb{M}\mathbb{R}^m$ of $\mathbf{f}$ is a relaxation function for $\mathbf{f}$. Additionally, if each univariate intrinsic function describing $\mathbf{f}$ satisfies Assumption 9.2.38, then $\mathcal{F}_{\text{unc}}$ is $(1,2)$-convergent.*

Moreover, if $m = 1$, in which case $\mathbf{f} \equiv f$ is scalar-valued, then the functions $\phi_{\boldsymbol{x}}, \psi_{\boldsymbol{x}}$ defined by Proposition 9.2.20 in terms of $\mathcal{F}_{\text{unc}}$ for each $\boldsymbol{x} \in \mathbb{I}B$ are each $\mathcal{C}^{i^*}$ on $\mathbb{R}^n$. For each $\boldsymbol{x} \in \mathbb{I}B$, $\phi_{\boldsymbol{x}}$ is convex on $\mathbb{R}^n$, and $\psi_{\boldsymbol{x}}$ is concave on $\mathbb{R}^n$.

9.5 Elemental relaxation functions

The following theorem shows that Definitions 9.2.30, 9.3.19, and 9.2.32 provide relaxation functions of addition, multiplication, and univariate intrinsic functions; the remainder of this section is concerned with proving this theorem.

Theorem 9.5.1. *The functions $+ : \mathbb{M}\mathbb{R}^2_{\text{prop}} \to \mathbb{M}\mathbb{R}_{\text{prop}}$, $\times_i : \mathbb{M}\mathbb{R}^2_{\text{prop}} \to \mathbb{M}\mathbb{R}_{\text{prop}}$, and $\mathcal{U} : \mathbb{M}B \to \mathbb{M}\mathbb{R}$ described in Definitions 9.2.30, 9.3.19, and 9.2.32 are relaxation functions for $+ : \mathbb{R}^2 \to \mathbb{R}$, $\times : \mathbb{R}^2 \to \mathbb{R}$, and $u : B \to \mathbb{R}$, respectively.*

Proof. This theorem combines Lemmata 9.5.2–9.5.6 below. □

Lemma 9.5.2. *The function $+ : \mathbb{M}\mathbb{R}^2_{\text{prop}} \to \mathbb{M}\mathbb{R}_{\text{prop}}$ is coherently concave, inclusion monotonic, and a McCormick extension of $+ : \mathbb{R}^2 \to \mathbb{R}$.*

Proof. This result follows from Theorem 2.4.20 in [100], noting that for any choice of $\mathcal{X}, \mathcal{Y} \in \mathbb{M}\mathbb{R}_{\text{prop}}$, $\boldsymbol{x}^{\text{B}} \cap \boldsymbol{x}^{\text{C}} = \boldsymbol{x}^{\text{C}}$ and $\boldsymbol{y}^{\text{B}} \cap \boldsymbol{y}^{\text{C}} = \boldsymbol{y}^{\text{C}}$. □

Lemma 9.5.3. *The function $\times_i : \mathbb{M}\mathbb{R}^2_{\text{prop}} \to \mathbb{M}\mathbb{R}_{\text{prop}}$ is coherently concave.*

Proof. Consider a coherent pair $(\mathcal{X}_1, \mathcal{Y}_1), (\mathcal{X}_2, \mathcal{Y}_2) \in \mathbb{M}\mathbb{R}^2_{\text{prop}}$, and a scalar $\ell \in [0, 1]$. Since this pair is coherent, define $\boldsymbol{x}^{\text{B}} := \boldsymbol{x}^{\text{B}}_1 = \boldsymbol{x}^{\text{B}}_2$ and $\boldsymbol{y}^{\text{B}} := \boldsymbol{y}^{\text{B}}_1 = \boldsymbol{y}^{\text{B}}_2$. Define $\mathcal{Q}_1 := \mathcal{X}_1\mathcal{Y}_1$ and $\mathcal{Q}_2 := \mathcal{X}_2\mathcal{Y}_2$. Using the definition of the squashing operation, it follows that $\boldsymbol{q}^{\text{B}}_1 = \boldsymbol{q}^{\text{B}}_2 = \boldsymbol{x}^{\text{B}}\boldsymbol{y}^{\text{B}} =: \boldsymbol{q}^{\text{B}}$, and so $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are coherent. Define $\mathcal{X}_0 := \mathcal{C}\text{onv}(\ell, \mathcal{X}_1, \mathcal{X}_2)$, and define $\mathcal{Y}_0$ and $\mathcal{Q}_0$ analogously. To obtain the required result, it suffices to show that $\mathcal{X}_0\mathcal{Y}_0 \supset \mathcal{Q}_0$.

If $\underline{q}^{\text{B}} = \overline{q}^{\text{B}} =: q$, then the definition of the squashing operation implies that $\mathcal{X}_0\mathcal{Y}_0 = [q, q] = \mathcal{Q}_0$, as required. Thus, it will be assumed throughout the rest of this proof that $\underline{q}^{\text{B}} < \overline{q}^{\text{B}}$.

For each $j \in \{0, 1, 2\}$, define $\boldsymbol{z}_j \equiv [\underline{z}_j, \overline{z}_j] \in \mathbb{I}\mathbb{R}$ such that

$$\underline{z}_j := \nu_i \left(\underline{(\underline{y}^{\mathrm{B}} \boldsymbol{x}_j^{\mathrm{C}})} + \underline{(\underline{x}^{\mathrm{B}} \boldsymbol{y}_j^{\mathrm{C}})} - \underline{x}^{\mathrm{B}} \underline{y}^{\mathrm{B}},\ \underline{(\overline{y}^{\mathrm{B}} \boldsymbol{x}_j^{\mathrm{C}})} + \underline{(\overline{x}^{\mathrm{B}} \boldsymbol{y}_j^{\mathrm{C}})} - \overline{x}^{\mathrm{B}} \overline{y}^{\mathrm{B}},\ p_{\boldsymbol{q}^{\mathrm{B}}} \right),$$

and $$\overline{z}_j := \lambda_i \left(\overline{(\underline{y}^{\mathrm{B}} \boldsymbol{x}_j^{\mathrm{C}})} + \overline{(\overline{x}^{\mathrm{B}} \boldsymbol{y}_j^{\mathrm{C}})} - \overline{x}^{\mathrm{B}} \underline{y}^{\mathrm{B}},\ \overline{(\overline{y}^{\mathrm{B}} \boldsymbol{x}_j^{\mathrm{C}})} + \overline{(\underline{x}^{\mathrm{B}} \boldsymbol{y}_j^{\mathrm{C}})} - \underline{x}^{\mathrm{B}} \overline{y}^{\mathrm{B}},\ p_{\boldsymbol{q}^{\mathrm{B}}} \right).$$

Since $\boldsymbol{q}_1^{\mathrm{B}} = \boldsymbol{q}_2^{\mathrm{B}} = \boldsymbol{x}^{\mathrm{B}} \boldsymbol{y}^{\mathrm{B}} = \boldsymbol{q}^{\mathrm{B}}$, the required inclusion, $\mathcal{X}_0 \mathcal{Y}_0 \supset \mathcal{C}\mathrm{onv}(\ell, \mathcal{Q}_1, \mathcal{Q}_2)$, is equivalent to the inclusion:

$$\mathcal{S}\mathrm{qu}_i((\boldsymbol{q}^{\mathrm{B}}, \boldsymbol{z}_0)) \supset \mathcal{C}\mathrm{onv}(\ell, \mathcal{S}\mathrm{qu}_i((\boldsymbol{q}^{\mathrm{B}}, \boldsymbol{z}_1)), \mathcal{S}\mathrm{qu}_i((\boldsymbol{q}^{\mathrm{B}}, \boldsymbol{z}_2))),$$

which is in turn equivalent to the inclusion:

$$\boldsymbol{belt}_i((\boldsymbol{q}^{\mathrm{B}}, \boldsymbol{z}_0)) \supset \ell\, \boldsymbol{belt}_i((\boldsymbol{q}^{\mathrm{B}}, \boldsymbol{z}_1)) + (1-\ell)\, \boldsymbol{belt}_i((\boldsymbol{q}^{\mathrm{B}}, \boldsymbol{z}_2)).$$

Thus, due to Lemma 9.3.15, it suffices to demonstrate the following inclusion:

$$\boldsymbol{belt}_i((\boldsymbol{q}^{\mathrm{B}}, \boldsymbol{z}_0)) \supset \boldsymbol{belt}_i(\mathcal{C}\mathrm{onv}(\ell, (\boldsymbol{q}^{\mathrm{B}}, \boldsymbol{z}_1), (\boldsymbol{q}^{\mathrm{B}}, \boldsymbol{z}_2))),$$

which can be rewritten as:

$$\boldsymbol{belt}_i((\boldsymbol{q}^{\mathrm{B}}, \boldsymbol{z}_0)) \supset \boldsymbol{belt}_i((\boldsymbol{q}^{\mathrm{B}}, [\ell \underline{z}_1 + (1-\ell)\underline{z}_2, \ell \overline{z}_1 + (1-\ell)\overline{z}_2)]).$$

Since $\boldsymbol{belt}_i$ is inclusion monotonic, it thus suffices to demonstrate the inequalities:

$$\underline{z}_0 \leq \ell \underline{z}_1 + (1-\ell)\underline{z}_2, \qquad \text{and} \qquad \overline{z}_0 \geq \ell \overline{z}_1 + (1-\ell)\overline{z}_2.$$

The first of these inequalities will be demonstrated here; the second can be shown to hold by an analogous argument. For each $j \in \{0, 1, 2\}$, define:

$$\alpha_j := \underline{(\underline{y}^{\mathrm{B}} \boldsymbol{x}_j^{\mathrm{C}})} + \underline{(\underline{x}^{\mathrm{B}} \boldsymbol{y}_j^{\mathrm{C}})} - \underline{x}^{\mathrm{B}} \underline{y}^{\mathrm{B}}, \qquad \text{and} \qquad \beta_j := \underline{(\overline{y}^{\mathrm{B}} \boldsymbol{x}_j^{\mathrm{C}})} + \underline{(\overline{x}^{\mathrm{B}} \boldsymbol{y}_j^{\mathrm{C}})} - \overline{x}^{\mathrm{B}} \overline{y}^{\mathrm{B}}.$$

Now, for each $j \in \{0, 1, 2\}$,

$$\underline{(\underline{y}^{\mathrm{B}} \boldsymbol{x}_j^{\mathrm{C}})} = \begin{cases} \underline{y}^{\mathrm{B}} \underline{x}_j^{\mathrm{C}} & \text{if } \underline{y}^{\mathrm{B}} \geq 0, \\ \underline{y}^{\mathrm{B}} \overline{x}_j^{\mathrm{C}} & \text{if } \underline{y}^{\mathrm{B}} < 0. \end{cases}$$

Moreover, by definition of the $\mathcal{C}$onv operation,

$$\underline{x}_0^{\mathrm{C}} = \ell \underline{x}_1^{\mathrm{C}} + (1-\ell)\underline{x}_2^{\mathrm{C}}, \qquad \text{and} \qquad \overline{x}_0^{\mathrm{C}} = \ell \overline{x}_1^{\mathrm{C}} + (1-\ell)\overline{x}_2^{\mathrm{C}}.$$

Combining the above results, it follows that

$$\underline{(\underline{y}^{\mathrm{B}}\boldsymbol{x}_0^{\mathrm{C}})} = \ell \underline{(\underline{y}^{\mathrm{B}}\boldsymbol{x}_1^{\mathrm{C}})} + (1-\ell)\underline{(\underline{y}^{\mathrm{B}}\boldsymbol{x}_2^{\mathrm{C}})};$$

an analogous argument shows that

$$\underline{(\underline{x}^{\mathrm{B}}\boldsymbol{y}_0^{\mathrm{C}})} = \ell \underline{(\underline{x}^{\mathrm{B}}\boldsymbol{y}_1^{\mathrm{C}})} + (1-\ell)\underline{(\underline{x}^{\mathrm{B}}\boldsymbol{y}_2^{\mathrm{C}})}.$$

Adding these two equations and subtracting the constant term $\underline{x}^{\mathrm{B}}\underline{y}^{\mathrm{B}}$, it follows that

$$\alpha_0 = \ell\alpha_1 + (1-\ell)\alpha_2;$$

an analogous argument shows that

$$\beta_0 = \ell\beta_1 + (1-\ell)\beta_2.$$

Thus,

$$\nu_i(\alpha_0, \beta_0, p_{\boldsymbol{q}^{\mathrm{B}}}) = \nu_i(\ell\alpha_1 + (1-\ell)\alpha_2, \ell\beta_1 + (1-\ell)\beta_2, p_{\boldsymbol{q}^{\mathrm{B}}}),$$

which, by Lemma 9.3.11, implies that

$$\nu_i(\alpha_0, \beta_0, p_{\boldsymbol{q}^{\mathrm{B}}}) \leq \ell\, \nu_i(\alpha_1, \beta_1, p_{\boldsymbol{q}^{\mathrm{B}}}) + (1-\ell)\, \nu_i(\alpha_2, \beta_2, p_{\boldsymbol{q}^{\mathrm{B}}}).$$

Comparing this inequality with the definitions of α_j, β_j, and $\underline{z}_j$ for each $j \in \{0,1,2\}$, it follows immediately that

$$\underline{z}_0 \leq \ell\underline{z}_1 + (1-\ell)\underline{z}_2,$$

as required. □

Lemma 9.5.4. *The function* $\times_i : \mathbb{MR}^2_{\mathrm{prop}} \to \mathbb{MR}_{\mathrm{prop}}$ *is inclusion monotonic.*

Proof. Consider any $\mathcal{X}_1, \mathcal{X}_2, \mathcal{Y}_1, \mathcal{Y}_2 \in \mathbb{MR}_{\mathrm{prop}}$ such that $\mathcal{X}_2 \subset \mathcal{X}_1$ and $\mathcal{Y}_2 \subset \mathcal{Y}_1$. It will be shown that $\mathcal{X}_2\mathcal{Y}_2 \subset \mathcal{X}_1\mathcal{Y}_1$. Since $\times : \mathbb{IR}^2 \to \mathbb{IR}$ is inclusion monotonic, it follows that $\boldsymbol{x}_2^{\mathrm{B}}\boldsymbol{y}_2^{\mathrm{B}} \subset \boldsymbol{x}_1^{\mathrm{B}}\boldsymbol{y}_1^{\mathrm{B}}$, and so $p_{\boldsymbol{x}_2^{\mathrm{B}}\boldsymbol{y}_2^{\mathrm{B}}} \leq p_{\boldsymbol{x}_1^{\mathrm{B}}\boldsymbol{y}_1^{\mathrm{B}}}$. By construction $\boldsymbol{x}_2^{\mathrm{C}} \subset \boldsymbol{x}_1^{\mathrm{C}}$ and

$\boldsymbol{y}_2^{\mathrm{C}} \subset \boldsymbol{y}_1^{\mathrm{C}}$. Define $\boldsymbol{z}_1 \in \mathbb{IR}$ as in Definition 9.3.19 for the product $\mathcal{X}_1\mathcal{Y}_1$, and define $\boldsymbol{z}_2 \in \mathbb{IR}$ analogously for the product $\mathcal{X}_2\mathcal{Y}_2$.

Due to Lemma 9.3.16 and the inclusion $\boldsymbol{x}_2^{\mathrm{B}}\boldsymbol{y}_2^{\mathrm{B}} \subset \boldsymbol{x}_1^{\mathrm{B}}\boldsymbol{y}_1^{\mathrm{B}}$, it suffices to show that $\boldsymbol{z}_2 \subset \boldsymbol{z}_1$. It will be shown that $\underline{z}_2 \geq \underline{z}_1$; an analogous argument shows that $\overline{z}_2 \leq \overline{z}_1$. In turn, due to Lemma 9.3.10 and the inequality $p_{\boldsymbol{x}_2^{\mathrm{B}}\boldsymbol{y}_2^{\mathrm{B}}} \leq p_{\boldsymbol{x}_1^{\mathrm{B}}\boldsymbol{y}_1^{\mathrm{B}}}$, it suffices to demonstrate the inequalities:

$$\begin{aligned} &\underline{(\underline{y}_1^{\mathrm{B}}\boldsymbol{x}_1^{\mathrm{C}})} + \underline{(\underline{x}_1^{\mathrm{B}}\boldsymbol{y}_1^{\mathrm{C}})} - \underline{x}_1^{\mathrm{B}}\underline{y}_1^{\mathrm{B}} \leq \underline{(\underline{y}_2^{\mathrm{B}}\boldsymbol{x}_2^{\mathrm{C}})} + \underline{(\underline{x}_2^{\mathrm{B}}\boldsymbol{y}_2^{\mathrm{C}})} - \underline{x}_2^{\mathrm{B}}\underline{y}_2^{\mathrm{B}}, \\ \text{and} \quad &\underline{(\overline{y}_1^{\mathrm{B}}\boldsymbol{x}_1^{\mathrm{C}})} + \underline{(\overline{x}_1^{\mathrm{B}}\boldsymbol{y}_1^{\mathrm{C}})} - \overline{x}_1^{\mathrm{B}}\overline{y}_1^{\mathrm{B}} \leq \underline{(\overline{y}_2^{\mathrm{B}}\boldsymbol{x}_2^{\mathrm{C}})} + \underline{(\overline{x}_2^{\mathrm{B}}\boldsymbol{y}_2^{\mathrm{C}})} - \overline{x}_2^{\mathrm{B}}\overline{y}_2^{\mathrm{B}}. \end{aligned}$$

Noting that $\boldsymbol{x}_j^{\mathrm{C}} = \boldsymbol{x}_j^{\mathrm{B}} \cap \boldsymbol{x}_j^{\mathrm{C}}$ and $\boldsymbol{y}_j^{\mathrm{C}} = \boldsymbol{y}_j^{\mathrm{B}} \cap \boldsymbol{y}_j^{\mathrm{C}}$ for each $j \in \{1,2\}$ by construction, the proof of [100, Theorem 2.4.23] demonstrates the above inequalities. □

Lemma 9.5.5. *The function* $\times_i : \mathbb{MR}^2_{\mathrm{prop}} \to \mathbb{MR}_{\mathrm{prop}}$ *is a McCormick extension of* $\times : \mathbb{R}^2 \to \mathbb{R}$.

Proof. Choose $x, y \in \mathbb{R}$, and consider the McCormick objects $\mathcal{X}_0 := ([x,x],[x,x]) \in \mathbb{MR}_{\mathrm{prop}}$ and $\mathcal{Y}_0 := ([y,y],[y,y]) \in \mathbb{MR}_{\mathrm{prop}}$. Observe that $[x,x][y,y] = [xy,xy]$, and that $p_{[x,x]} = p_{[y,y]} = p_{[x,x][y,y]} = 0$. Thus, according to Definition 9.3.19 and the established properties of γ_i and σ_i, it follows that:

$$\mathcal{X}_0\mathcal{Y}_0 = \mathcal{S}\mathrm{qu}_i(([x,x][y,y],\boldsymbol{z})),$$

where $\boldsymbol{z} \equiv [\underline{z},\overline{z}]$ is defined as follows:

$$\begin{aligned} \underline{z} &:= \nu_i\left(\underline{(y[x,x])} + \underline{(x[y,y])} - xy,\ \underline{(y[x,x])} + \underline{(x[y,y])} - xy,\ 0\right), \\ &= \nu_i(xy,xy,0) = \gamma_i(xy,xy,0) = xy, \\ \text{and} \quad \overline{z} &:= \lambda_i\left(\overline{(y[x,x])} + \overline{(x[y,y])} - xy,\ \overline{(y[x,x])} + \overline{(x[y,y])} - xy,\ 0\right), \\ &= \lambda_i(xy,xy,0) = \sigma_i(xy,xy,0) = xy. \end{aligned}$$

Thus, $\mathcal{X}_0\mathcal{Y}_0 = \mathcal{S}\mathrm{qu}_i(([xy,xy],[xy,xy])) = ([xy,xy],[xy,xy])$. □

Lemma 9.5.6 (Theorems 2.4.27, 2.4.29, and 2.4.30 in [100]). *For any univariate intrinsic function $u : B \subset \mathbb{R} \to \mathbb{R}$, the function $\mathcal{U} : \mathbb{M}B \to \mathbb{M}\mathbb{R}$ is coherently concave, inclusion monotonic, and a McCormick extension of u.*

9.6 Continuous and twice-continuous differentiability

The results in this section show that for any natural or unconstrained $\mathcal{C}^i$ McCormick extension of an MC-factorable function, the convex/concave relaxations suggested by Proposition 9.2.20 are indeed $\mathcal{C}^i$ on their interval domains, and their gradients may be evaluated using the standard forward or reverse modes of automatic differentiation [34].

In particular, Lemmata 9.6.5 and 9.6.6 effectively correct McCormick's proposed sufficient condition for differentiability of relaxations of composite functions [74, p. 151], by applying Assumption 9.2.21.

Theorem 9.6.1. *Consider any $i^* \in \{1, 2\}$. For fixed intervals $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{I}\mathbb{R}$, the mappings*

$$\begin{aligned}
&(\underline{y}_1, \overline{y}_1, \underline{y}_2, \overline{y}_2) \mapsto \underline{+}^{\mathrm{C}}\left((\boldsymbol{x}_1, [\underline{y}_1, \overline{y}_1]), (\boldsymbol{x}_2, [\underline{y}_2, \overline{y}_2])\right), \\
&(\underline{y}_1, \overline{y}_1, \underline{y}_2, \overline{y}_2) \mapsto \overline{+}^{\mathrm{C}}\left((\boldsymbol{x}_1, [\underline{y}_1, \overline{y}_1]), (\boldsymbol{x}_2, [\underline{y}_2, \overline{y}_2])\right), \\
&(\underline{y}_1, \overline{y}_1, \underline{y}_2, \overline{y}_2) \mapsto \underline{\times}_i^{\mathrm{C}}\left((\boldsymbol{x}_1, [\underline{y}_1, \overline{y}_1]), (\boldsymbol{x}_2, [\underline{y}_2, \overline{y}_2])\right), \\
\text{and}\quad &(\underline{y}_1, \overline{y}_1, \underline{y}_2, \overline{y}_2) \mapsto \overline{\times}_i^{\mathrm{C}}\left((\boldsymbol{x}_1, [\underline{y}_1, \overline{y}_1]), (\boldsymbol{x}_2, [\underline{y}_2, \overline{y}_2])\right)
\end{aligned}$$

described in Definitions 9.2.30 and 9.3.19 are each $\mathcal{C}^{i^*}$ on $\{(\underline{y}_1, \overline{y}_1, \underline{y}_2, \overline{y}_2) \in \mathbb{R}^4 : \underline{y}_1 \leq \overline{y}_1, \underline{y}_2 \leq \overline{y}_2, [\underline{y}_1, \overline{y}_1] \in \boldsymbol{x}_1, [\underline{y}_2, \overline{y}_2] \in \boldsymbol{x}_2\}$.

Next, consider a univariate intrinsic function $u : B \subset \mathbb{R} \to \mathbb{R}$ that satisfies Assumption 9.2.21 with $i := i^*$, and choose any fixed interval $\boldsymbol{x} \in \mathbb{I}B$. The mappings

$$(\underline{y}, \overline{y}) \mapsto \underline{u}^{\mathrm{C}}((\boldsymbol{x}, [\underline{y}, \overline{y}])) \qquad \text{and} \qquad (\underline{y}, \overline{y}) \mapsto \overline{u}^{\mathrm{C}}((\boldsymbol{x}, [\underline{y}, \overline{y}])),$$

described in Definition 9.2.32, are each $\mathcal{C}^{i^*}$ on $\{(\underline{y}, \overline{y}) \in \mathbb{R}^2 : \underline{y} \leq \overline{y}, [\underline{y}, \overline{y}] \subset \boldsymbol{x}\}$.

Proof. This theorem collects the results of Lemmata 9.6.2–9.6.6 below. □

Lemma 9.6.2. *For fixed intervals $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{IR}$, the mappings*

$$
\begin{aligned}
& (\underline{y}_1, \overline{y}_1, \underline{y}_2, \overline{y}_2) \mapsto \underline{+}^{\mathrm{C}} \left((\boldsymbol{x}_1, [\underline{y}_1, \overline{y}_1]), (\boldsymbol{x}_2, [\underline{y}_2, \overline{y}_2]) \right), \\
\textit{and} \qquad & (\underline{y}_1, \overline{y}_1, \underline{y}_2, \overline{y}_2) \mapsto \overline{+}^{\mathrm{C}} \left((\boldsymbol{x}_1, [\underline{y}_1, \overline{y}_1]), (\boldsymbol{x}_2, [\underline{y}_2, \overline{y}_2]) \right)
\end{aligned}
$$

are each $\mathcal{C}^2$ on $\{(\underline{y}_1, \overline{y}_1, \underline{y}_2, \overline{y}_2) \in \mathbb{R}^4 : \underline{y}_1 \leq \overline{y}_1, \underline{y}_2 \leq \overline{y}_2, [\underline{y}_1, \overline{y}_1] \in \boldsymbol{x}_1, [\underline{y}_2, \overline{y}_2] \in \boldsymbol{x}_2\}$.

Proof. The definition of $+ : \mathbb{MR}^2_{\text{prop}} \mapsto \mathbb{MR}_{\text{prop}}$ implies that the mappings in question are linear, and are therefore $\mathcal{C}^2$. □

Lemma 9.6.3. *Suppose that scalars $a, b, c, d \in \mathbb{R}$ are such that $ab = ad = cb = cd$. At least one of the following conditions must hold:*

- *both $a = c$ and $b = d$ hold simultaneously,*
- *$a = c = 0$,*
- *$b = d = 0$.*

Proof. Suppose that the first condition does not hold; it will be shown that either the second or third condition must hold in this case. Thus, suppose that either $a \neq c$ or $b \neq d$. If $a \neq c$, then the equations $(a - c)b = 0 = (a - c)d$ imply that $b = d = 0$, as required. Otherwise, if $b \neq d$, then the equations $a(b - d) = 0 = c(b - d)$ imply that $a = c = 0$, as required. □

Lemma 9.6.4. *For each $i \in \{1, 2\}$, given fixed intervals $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{IR}$, the mappings*

$$
\begin{aligned}
& (\underline{y}_1, \overline{y}_1, \underline{y}_2, \overline{y}_2) \mapsto \underline{\times}_i^{\mathrm{C}} \left((\boldsymbol{x}_1, [\underline{y}_1, \overline{y}_1]), (\boldsymbol{x}_2, [\underline{y}_2, \overline{y}_2]) \right), \\
\textit{and} \qquad & (\underline{y}_1, \overline{y}_1, \underline{y}_2, \overline{y}_2) \mapsto \overline{\times}_i^{\mathrm{C}} \left((\boldsymbol{x}_1, [\underline{y}_1, \overline{y}_1]), (\boldsymbol{x}_2, [\underline{y}_2, \overline{y}_2]) \right)
\end{aligned}
$$

are each $\mathcal{C}^i$ on $\{(\underline{y}_1, \overline{y}_1, \underline{y}_2, \overline{y}_2) \in \mathbb{R}^4 : \underline{y}_1 \leq \overline{y}_1, \underline{y}_2 \leq \overline{y}_2, [\underline{y}_1, \overline{y}_1] \in \boldsymbol{x}_1, [\underline{y}_2, \overline{y}_2] \in \boldsymbol{x}_2\}$.

Proof. The cases in which $\operatorname{wid}(\boldsymbol{x}_1 \boldsymbol{x}_2) > 0$ and $\operatorname{wid}(\boldsymbol{x}_1 \boldsymbol{x}_2) = 0$ will be considered separately.

Firstly, suppose that $\operatorname{wid}(\boldsymbol{x}_1\boldsymbol{x}_2) > 0$. For any $c \in \mathbb{R}$, (9.1) and (9.2) imply that the mappings $(v,w) \mapsto \underline{(c[v,w])}$ and $(v,w) \mapsto \overline{(c[v,w])}$ are both linear on $\{(v,w) \in \mathbb{R}^2 : v \leq w\}$, and are therefore $\mathcal{C}^i$. This observation, together with Lemma 9.3.8, Lemma 9.3.17, and Definition 9.3.19, implies that the required result holds.

Secondly, suppose that $\operatorname{wid}(\boldsymbol{x}_1\boldsymbol{x}_2) = 0$, in which case $\underline{x}_1\underline{x}_2 = \underline{x}_1\overline{x}_2 = \overline{x}_1\underline{x}_2 = \overline{x}_1\overline{x}_2$. Applying Lemma 9.6.3, it suffices to consider separately the cases in which $\underline{x}_1 = \overline{x}_1 = 0$, $\underline{x}_2 = \overline{x}_2 = 0$, and both $\underline{x}_1 = \overline{x}_1$ and $\underline{x}_2 = \overline{x}_2$.

If $\underline{x}_1 = \overline{x}_1 = 0$, then $\boldsymbol{x}_1\boldsymbol{x}_2 = [0,0]$, in which case the outer squashing operation in Definition 9.3.19 implies that $(\boldsymbol{x}_1, [\underline{y}_1, \overline{y}_1]) \times (\boldsymbol{x}_2, [\underline{y}_2, \overline{y}_2]) = ([0,0],[0,0])$. Thus, each of the two mappings in the statement of the lemma is the zero mapping, which is trivially $\mathcal{C}^i$. The case in which $\underline{x}_2 = \overline{x}_2 = 0$ is analogous.

Lastly, if both $\underline{x}_1 = \overline{x}_1 =: x_1$ and $\underline{x}_2 = \overline{x}_2 =: x_2$, then $\boldsymbol{x}_1\boldsymbol{x}_2 = [x_1x_2, x_1x_2]$, in which case the outer squashing operation in Definition 9.3.19 implies that

$$(\boldsymbol{x}_1, [\underline{y}_1, \overline{y}_1]) \times (\boldsymbol{x}_2, [\underline{y}_2, \overline{y}_2]) = ([x_1x_2, x_1x_2], [x_1x_2, x_1x_2]).$$

Thus, each of the two mappings in the statement of the lemma is a constant mapping, which, again, is trivially $\mathcal{C}^i$. □

The following two lemmata essentially show that McCormick's proposed sufficient condition for differentiable relaxations of composite functions [74, p. 151] becomes valid when Assumption 9.2.21 is applied.

Lemma 9.6.5. *Consider a univariate intrinsic function $u : B \subset \mathbb{R} \to \mathbb{R}$ that satisfies Assumption 9.2.21. For any intervals $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{I}B$ for which $\boldsymbol{y} \subset \boldsymbol{x}$,*

$$\begin{aligned} & u^{\mathrm{cv}}(\boldsymbol{x}, \operatorname{mid}(\zeta_u^{\min}(\boldsymbol{x}), \underline{y}, \overline{y})) = u_I^{\mathrm{cv}}(\boldsymbol{x}, \underline{y}) + u_D^{\mathrm{cv}}(\boldsymbol{x}, \overline{y}) - u^{\mathrm{cv}}(\boldsymbol{x}, \zeta_u^{\min}(\boldsymbol{x})), \\ \textit{and} \quad & u^{\mathrm{cc}}(\boldsymbol{x}, \operatorname{mid}(\zeta_u^{\max}(\boldsymbol{x}), \underline{y}, \overline{y})) = u_I^{\mathrm{cc}}(\boldsymbol{x}, \underline{y}) + u_D^{\mathrm{cc}}(\boldsymbol{x}, \overline{y}) - u^{\mathrm{cc}}(\boldsymbol{x}, \zeta_u^{\max}(\boldsymbol{x})). \end{aligned}$$

Proof. The first required equation will be shown to hold; the second can be demonstrated analogously. By construction,

$$
\begin{aligned}
u_I^{\mathrm{cv}}(\boldsymbol{x},\underline{y}) + u_D^{\mathrm{cv}}(\boldsymbol{x},\overline{y}) - u^{\mathrm{cv}}(\boldsymbol{x},\zeta_u^{\min}(\boldsymbol{x})) \\
= u^{\mathrm{cv}}(\boldsymbol{x},\max\{\underline{y},\zeta_u^{\min}(\boldsymbol{x})\}) + u^{\mathrm{cv}}(\boldsymbol{x},\min\{\overline{y},\zeta_u^{\min}(\boldsymbol{x})\}) \\
- u^{\mathrm{cv}}(\boldsymbol{x},\zeta_u^{\min}(\boldsymbol{x})).
\end{aligned} \tag{9.8}
$$

Since $\underline{y} \leq \overline{y}$, at least one of the following three cases must apply: $\zeta_u^{\min}(\boldsymbol{x}) \leq \underline{y} \leq \overline{y}$, $\underline{y} \leq \zeta_u^{\min}(\boldsymbol{x}) \leq \overline{y}$, or $\underline{y} \leq \overline{y} \leq \zeta_u^{\min}(\boldsymbol{x})$. These cases will be considered separately.

If $\zeta_u^{\min}(\boldsymbol{x}) \leq \underline{y} \leq \overline{y}$, then $\underline{y} = \mathrm{mid}(\zeta_u^{\min}(\boldsymbol{x}),\underline{y},\overline{y})$, and (9.8) becomes

$$
\begin{aligned}
&u_I^{\mathrm{cv}}(\boldsymbol{x},\underline{y}) + u_D^{\mathrm{cv}}(\boldsymbol{x},\overline{y}) - u^{\mathrm{cv}}(\boldsymbol{x},\zeta_u^{\min}(\boldsymbol{x})) \\
&\quad = u^{\mathrm{cv}}(\boldsymbol{x},\underline{y}) + u^{\mathrm{cv}}(\boldsymbol{x},\zeta_u^{\min}(\boldsymbol{x})) - u^{\mathrm{cv}}(\boldsymbol{x},\zeta_u^{\min}(\boldsymbol{x})) \\
&\quad = u^{\mathrm{cv}}(\boldsymbol{x},\underline{y}).
\end{aligned}
$$

If $\underline{y} \leq \zeta_u^{\min}(\boldsymbol{x}) \leq \overline{y}$, then $\zeta_u^{\min}(\boldsymbol{x}) = \mathrm{mid}(\zeta_u^{\min}(\boldsymbol{x}),\underline{y},\overline{y})$, and (9.8) becomes

$$
\begin{aligned}
&u_I^{\mathrm{cv}}(\boldsymbol{x},\underline{y}) + u_D^{\mathrm{cv}}(\boldsymbol{x},\overline{y}) - u^{\mathrm{cv}}(\boldsymbol{x},\zeta_u^{\min}(\boldsymbol{x})) \\
&\quad = u^{\mathrm{cv}}(\boldsymbol{x},\zeta_u^{\min}(\boldsymbol{x})) + u^{\mathrm{cv}}(\boldsymbol{x},\zeta_u^{\min}(\boldsymbol{x})) - u^{\mathrm{cv}}(\boldsymbol{x},\zeta_u^{\min}(\boldsymbol{x})) = u^{\mathrm{cv}}(\boldsymbol{x},\zeta_u^{\min}(\boldsymbol{x})).
\end{aligned}
$$

If $\underline{y} \leq \overline{y} \leq \zeta_u^{\min}(\boldsymbol{x})$, then $\overline{y} = \mathrm{mid}(\zeta_u^{\min}(\boldsymbol{x}),\underline{y},\overline{y})$, and (9.8) becomes

$$
\begin{aligned}
&u_I^{\mathrm{cv}}(\boldsymbol{x},\underline{y}) + u_D^{\mathrm{cv}}(\boldsymbol{x},\overline{y}) - u^{\mathrm{cv}}(\boldsymbol{x},\zeta_u^{\min}(\boldsymbol{x})) \\
&\quad = u^{\mathrm{cv}}(\boldsymbol{x},\zeta_u^{\min}(\boldsymbol{x})) + u^{\mathrm{cv}}(\boldsymbol{x},\overline{y}) - u^{\mathrm{cv}}(\boldsymbol{x},\zeta_u^{\min}(\boldsymbol{x})) \\
&\quad = u^{\mathrm{cv}}(\boldsymbol{x},\overline{y}).
\end{aligned}
$$

In each case, the required result is satisfied. □

Lemma 9.6.6. *Consider a univariate intrinsic function $u : B \subset \mathbb{R} \to \mathbb{R}$ that satisfies Assumption 9.2.21, and an interval $\mathbf{x} \in \mathbb{I}B$. The functions $u_I^{\mathrm{cv}}(\boldsymbol{x},\cdot)$, $u_D^{\mathrm{cv}}(\boldsymbol{x},\cdot)$, $u_I^{\mathrm{cc}}(\boldsymbol{x},\cdot)$, and $u_D^{\mathrm{cc}}(\boldsymbol{x},\cdot)$ are each $\mathcal{C}^i$ on $\mathbf{x}$.*

Proof. It will be shown that $u_I^{\mathrm{cv}}(\boldsymbol{x},\cdot)$ and $u_D^{\mathrm{cv}}(\boldsymbol{x},\cdot)$ are $\mathcal{C}^i$; the remaining results can be demonstrated analogously. The cases in which $\underline{x} < \zeta_u^{\min}(\boldsymbol{x}) < \overline{x}$, $\zeta_u^{\min}(\boldsymbol{x}) = \underline{x}$, or $\zeta_u^{\min}(\boldsymbol{x}) = \overline{x}$ will be considered separately.

Suppose first that $\underline{x} < \zeta_u^{\min}(\boldsymbol{x}) < \overline{x}$. Since the mapping $\phi := u^{\mathrm{cv}}(\boldsymbol{x}, \cdot)$ is $\mathcal{C}^1$ on $\boldsymbol{x}$, regardless of the value of $i \in \{1, 2\}$, it follows that $\nabla\phi(\zeta_u^{\min}(\boldsymbol{x})) = 0$. Using this result, it is readily verified that $\phi_I := u_I^{\mathrm{cv}}(\boldsymbol{x}, \cdot)$ and $\phi_D := u_D^{\mathrm{cv}}(\boldsymbol{x}, \cdot)$ are $\mathcal{C}^1$ on $\boldsymbol{x}$, with

$$\begin{aligned} \nabla\phi_I(z) &= \begin{cases} 0 & \text{if } z \leq \zeta_u^{\min}(\boldsymbol{x}), \\ \nabla\phi(z) & \text{if } z > \zeta_u^{\min}(\boldsymbol{x}), \end{cases} \\ \text{and} \qquad \nabla\phi_D(z) &= \begin{cases} \nabla\phi(z) & \text{if } z < \zeta_u^{\min}(\boldsymbol{x}), \\ 0 & \text{if } z \geq \zeta_u^{\min}(\boldsymbol{x}). \end{cases} \end{aligned} \tag{9.9}$$

Furthermore, if $i = 2$, then Assumption 9.2.21 implies that ϕ is $\mathcal{C}^2$ on $\boldsymbol{x}$, and that $\nabla^2\phi(\zeta_u^{\min}(\boldsymbol{x})) = 0$. Using this result, it is readily verified that ϕ_D and ϕ_I are $\mathcal{C}^2$ on $\boldsymbol{x}$, with

$$\begin{aligned} \nabla^2\phi_I(z) &= \begin{cases} 0 & \text{if } z \leq \zeta_u^{\min}(\boldsymbol{x}), \\ \nabla^2\phi(z) & \text{if } z > \zeta_u^{\min}(\boldsymbol{x}), \end{cases} \\ \text{and} \qquad \nabla^2\phi_D(z) &= \begin{cases} \nabla^2\phi(z) & \text{if } z < \zeta_u^{\min}(\boldsymbol{x}), \\ 0 & \text{if } z \geq \zeta_u^{\min}(\boldsymbol{x}). \end{cases} \end{aligned}$$

Next, suppose that either $\zeta_u^{\min}(\boldsymbol{x}) = \underline{x}$ or $\zeta_u^{\min}(\boldsymbol{x}) = \overline{x}$. In these cases, the functions ϕ_I and ϕ_D are each equivalent on $\boldsymbol{x}$ to either ϕ or to the constant mapping $\phi^* : z \mapsto \phi(\zeta_u^{\min}(\boldsymbol{x}))$, and are therefore $\mathcal{C}^i$ on $\boldsymbol{x}$. □

9.6.1 Gradient propagation

Using the obtained differentiability results, the standard forward or reverse modes of automatic differentiation [34] can be used to evaluate derivatives of the convex/concave relaxations obtained for natural or unconstrained $\mathcal{C}^i$ McCormick extensions, provided that gradients can be evaluated for the composed addition, multiplication, and univariate intrinsic operations. The obtained gradients are clearly subgradients of the corresponding relaxations.

To evaluate derivatives for $\mathcal{C}^i$ McCormick extensions, addition and univariate intrinsic composition can be treated exactly as in Proposition 2.9 and Theorem 3.2 in [76], with all subgradients mentioned in these results replaced by the corre-

sponding gradients. For multiplication, repeated application of the chain rule to Definition 9.3.19 yields the following, which makes use of the partial derivatives of ν_i and λ_i provided by Proposition 9.3.9.

Theorem 9.6.7. *Consider functions $f, g : D \subset \mathbb{R}^n \to \mathbb{R}$, and relaxation functions $\mathcal{F}, \mathcal{G} : \mathbb{M}D$ (or $\mathbb{M}D_{\text{prop}}$) $\to \mathbb{M}\mathbb{R}$ for f and g on D, such that the mappings $\mathcal{X} \mapsto \boldsymbol{f}^{\text{B}}(\mathcal{X})$ and $\mathcal{X} \mapsto \boldsymbol{g}^{\text{B}}(\mathcal{X})$ are each independent of their $\boldsymbol{x}^{\text{C}}$ argument. Consider the product function $h : D \to \mathbb{R} : \mathbf{z} \mapsto f(\mathbf{z})\, g(\mathbf{z})$, and the corresponding product relaxation function $\mathcal{H} : \mathbb{M}D$ (or $\mathbb{M}D_{\text{prop}}$) $\to \mathbb{M}\mathbb{R} : \mathcal{X} \mapsto \mathcal{F}(\mathcal{X})\, \mathcal{G}(\mathcal{X})$. As in Proposition 9.2.20, for some fixed $\boldsymbol{y} \in \mathbb{I}D$, construct the convex/concave relaxations $\phi_{h,\boldsymbol{y}} : \mathbf{z} \mapsto \underline{h}^{\text{C}}((\boldsymbol{y}, [\mathbf{z}, \mathbf{z}]))$ and $\psi_{h,\boldsymbol{y}} : \mathbf{z} \mapsto \overline{h}^{\text{C}}((\boldsymbol{y}, [\mathbf{z}, \mathbf{z}]))$ of h on $\boldsymbol{y}$, and construct the analogous relaxations $\phi_{f,\boldsymbol{y}}/\psi_{f,\boldsymbol{y}}$ of f and $\phi_{g,\boldsymbol{y}}/\psi_{g,\boldsymbol{y}}$ of g. Gradients of $\phi_{h,\boldsymbol{y}}$ and $\psi_{h,\boldsymbol{y}}$ at some particular $\mathbf{x} \in \boldsymbol{y}$ may be computed as follows, with $\mathcal{Y} := (\boldsymbol{y}, \mathbf{x}) \in \mathbb{M}D$ (or $\mathbb{M}D_{\text{prop}}$). For notational simplicity, the $\mathcal{Y}$ arguments of $\boldsymbol{f}^{\text{B}}(\mathcal{Y}) \equiv [\underline{f}^{\text{B}}(\mathcal{Y}), \overline{f}^{\text{B}}(\mathcal{Y})]$, $\boldsymbol{g}^{\text{B}}(\mathcal{Y}) \equiv [\underline{g}^{\text{B}}(\mathcal{Y}), \overline{g}^{\text{B}}(\mathcal{Y})]$, and $\boldsymbol{h}^{\text{B}}(\mathcal{Y}) \equiv [\underline{h}^{\text{B}}(\mathcal{Y}), \overline{h}^{\text{B}}(\mathcal{Y})]$ will be omitted.*

If $\overline{h}^{\text{B}} = \underline{h}^{\text{B}}$, then $\nabla\phi_{h,\boldsymbol{y}}(\mathbf{x}) = \nabla\psi_{h,\boldsymbol{y}}(\mathbf{x}) = \mathbf{0}$. Otherwise, if $\overline{h}^{\text{B}} > \underline{h}^{\text{B}}$, then define intermediate scalar quantities:

$$
\begin{aligned}
n_1(\mathbf{x}) &:= \underline{(\underline{g}^{\text{B}}\, \boldsymbol{f}^{\text{C}}(\mathcal{Y}))} + \underline{(\underline{f}^{\text{B}}\, \boldsymbol{g}^{\text{C}}(\mathcal{Y}))} - \underline{f}^{\text{B}}\underline{g}^{\text{B}}, \\
n_2(\mathbf{x}) &:= \underline{(\overline{g}^{\text{B}}\, \boldsymbol{f}^{\text{C}}(\mathcal{Y}))} + \underline{(\overline{f}^{\text{B}}\, \boldsymbol{g}^{\text{C}}(\mathcal{Y}))} - \overline{f}^{\text{B}}\overline{g}^{\text{B}}, \\
n_3(\mathbf{x}) &:= \overline{(\underline{g}^{\text{B}}\, \boldsymbol{f}^{\text{C}}(\mathcal{Y}))} + \overline{(\overline{f}^{\text{B}}\, \boldsymbol{g}^{\text{C}}(\mathcal{Y}))} - \overline{f}^{\text{B}}\underline{g}^{\text{B}}, \\
n_4(\mathbf{x}) &:= \overline{(\overline{g}^{\text{B}}\, \boldsymbol{f}^{\text{C}}(\mathcal{Y}))} + \overline{(\underline{f}^{\text{B}}\, \boldsymbol{g}^{\text{C}}(\mathcal{Y}))} - \underline{f}^{\text{B}}\overline{g}^{\text{B}},
\end{aligned}
$$

If $\overline{f}^{\text{B}} = \underline{f}^{\text{B}}$, then define intermediate scalar quantities $b_1(\mathbf{x}) = b_2(\mathbf{x}) = b_3(\mathbf{x}) = b_4(\mathbf{x}) := 0$. Otherwise, if $\overline{f}^{\text{B}} > \underline{f}^{\text{B}}$, then define:

$$
b_1(\mathbf{x}) := \begin{cases} \underline{g}^{\text{B}}\, \nabla\phi_{f,\boldsymbol{y}}(\mathbf{x}), & \text{if } \underline{g}^{\text{B}} \geq 0, \\ \underline{g}^{\text{B}}\, \nabla\psi_{f,\boldsymbol{y}}(\mathbf{x}), & \text{if } \underline{g}^{\text{B}} < 0, \end{cases} \qquad
b_2(\mathbf{x}) := \begin{cases} \overline{g}^{\text{B}}\, \nabla\phi_{f,\boldsymbol{y}}(\mathbf{x}), & \text{if } \overline{g}^{\text{B}} \geq 0, \\ \overline{g}^{\text{B}}\, \nabla\psi_{f,\boldsymbol{y}}(\mathbf{x}), & \text{if } \overline{g}^{\text{B}} < 0, \end{cases}
$$

$$
b_3(\mathbf{x}) := \begin{cases} \underline{g}^{\text{B}}\, \nabla\psi_{f,\boldsymbol{y}}(\mathbf{x}), & \text{if } \underline{g}^{\text{B}} \geq 0, \\ \underline{g}^{\text{B}}\, \nabla\phi_{f,\boldsymbol{y}}(\mathbf{x}), & \text{if } \underline{g}^{\text{B}} < 0, \end{cases} \qquad
b_4(\mathbf{x}) := \begin{cases} \overline{g}^{\text{B}}\, \nabla\psi_{f,\boldsymbol{y}}(\mathbf{x}), & \text{if } \overline{g}^{\text{B}} \geq 0, \\ \overline{g}^{\text{B}}\, \nabla\phi_{f,\boldsymbol{y}}(\mathbf{x}), & \text{if } \overline{g}^{\text{B}} < 0. \end{cases}
$$

If $\overline{g}^{\mathrm{B}} = \underline{g}^{\mathrm{B}}$*, then define intermediate scalar quantities* $b_5(\mathbf{x}) = b_6(\mathbf{x}) = b_7(\mathbf{x}) = b_8(\mathbf{x}) := 0$*. Otherwise, if* $\overline{g}^{\mathrm{B}} > \underline{g}^{\mathrm{B}}$*, then define:*

$$b_5(\mathbf{x}) := \begin{cases} \underline{f}^{\mathrm{B}} \nabla \phi_{g,\boldsymbol{y}}(\mathbf{x}), & \text{if } \underline{f}^{\mathrm{B}} \geq 0, \\ \underline{f}^{\mathrm{B}} \nabla \psi_{g,\boldsymbol{y}}(\mathbf{x}), & \text{if } \underline{f}^{\mathrm{B}} < 0, \end{cases} \qquad b_6(\mathbf{x}) := \begin{cases} \overline{f}^{\mathrm{B}} \nabla \phi_{g,\boldsymbol{y}}(\mathbf{x}), & \text{if } \overline{f}^{\mathrm{B}} \geq 0, \\ \overline{f}^{\mathrm{B}} \nabla \psi_{g,\boldsymbol{y}}(\mathbf{x}), & \text{if } \overline{f}^{\mathrm{B}} < 0, \end{cases}$$

$$b_7(\mathbf{x}) := \begin{cases} \overline{f}^{\mathrm{B}} \nabla \psi_{g,\boldsymbol{y}}(\mathbf{x}), & \text{if } \overline{f}^{\mathrm{B}} \geq 0, \\ \overline{f}^{\mathrm{B}} \nabla \phi_{g,\boldsymbol{y}}(\mathbf{x}), & \text{if } \overline{f}^{\mathrm{B}} < 0, \end{cases} \qquad b_8(\mathbf{x}) := \begin{cases} \underline{f}^{\mathrm{B}} \nabla \psi_{g,\boldsymbol{y}}(\mathbf{x}), & \text{if } \underline{f}^{\mathrm{B}} \geq 0, \\ \underline{f}^{\mathrm{B}} \nabla \phi_{g,\boldsymbol{y}}(\mathbf{x}), & \text{if } \underline{f}^{\mathrm{B}} < 0. \end{cases}$$

Next, define the following intermediate scalar quantities:

$$\begin{aligned}
a_1(\mathbf{x}) &:= \frac{\partial \nu_i}{\partial x}(n_1(\mathbf{x}), n_2(\mathbf{x}), p_{\boldsymbol{h}^{\mathrm{B}}})\, b_1(\mathbf{x}) + \frac{\partial \nu_i}{\partial y}(n_1(\mathbf{x}), n_2(\mathbf{x}), p_{\boldsymbol{h}^{\mathrm{B}}})\, b_2(\mathbf{x}), \\
a_2(\mathbf{x}) &:= \frac{\partial \nu_i}{\partial x}(n_1(\mathbf{x}), n_2(\mathbf{x}), p_{\boldsymbol{h}^{\mathrm{B}}})\, b_5(\mathbf{x}) + \frac{\partial \nu_i}{\partial y}(n_1(\mathbf{x}), n_2(\mathbf{x}), p_{\boldsymbol{h}^{\mathrm{B}}})\, b_6(\mathbf{x}), \\
a_3(\mathbf{x}) &:= \frac{\partial \lambda_i}{\partial x}(n_3(\mathbf{x}), n_4(\mathbf{x}), p_{\boldsymbol{h}^{\mathrm{B}}})\, b_3(\mathbf{x}) + \frac{\partial \lambda_i}{\partial y}(n_3(\mathbf{x}), n_4(\mathbf{x}), p_{\boldsymbol{h}^{\mathrm{B}}})\, b_4(\mathbf{x}), \\
a_4(\mathbf{x}) &:= \frac{\partial \lambda_i}{\partial x}(n_3(\mathbf{x}), n_4(\mathbf{x}), p_{\boldsymbol{h}^{\mathrm{B}}})\, b_7(\mathbf{x}) + \frac{\partial \lambda_i}{\partial y}(n_3(\mathbf{x}), n_4(\mathbf{x}), p_{\boldsymbol{h}^{\mathrm{B}}})\, b_8(\mathbf{x}).
\end{aligned}$$

Then,

$$\begin{aligned}
\nabla \phi_{h,\boldsymbol{y}}(\mathbf{x}) &= \frac{\partial \gamma_i}{\partial z}(\underline{h}^{\mathrm{C}}(\mathcal{Y}), \underline{h}^{\mathrm{B}}, p_{\boldsymbol{h}^{\mathrm{B}}})\ (a_1(\mathbf{x}) + a_2(\mathbf{x}))\,, \\
\nabla \psi_{h,\boldsymbol{y}}(\mathbf{x}) &= \frac{\partial \sigma_i}{\partial z}(\overline{h}^{\mathrm{C}}(\mathcal{Y}), \overline{h}^{\mathrm{B}}, p_{\boldsymbol{h}^{\mathrm{B}}})\ (a_3(\mathbf{x}) + a_4(\mathbf{x}))\,.
\end{aligned}$$

Proof. This result follows immediately from Definition 9.3.19 and the chain rule in Proposition 9.2.5. Observe that, in light of Remark 9.2.2, if a composed function is defined only at a single point, then its derivative at this point may be set to $\mathbf{0}$ without affecting the validity of this chain rule. □

When constructing unconstrained $\mathcal{C}^i$ McCormick relaxations, the following gradient propagation result can be used to handle the initial squashing operation.

Proposition 9.6.8. *For fixed* $\boldsymbol{y} \in \mathbb{IR}$ *and* $i \in \{1, 2\}$*, consider the functions* $\underline{s}^{\mathrm{C}}_{\boldsymbol{y}}, \overline{s}^{\mathrm{C}}_{\boldsymbol{y}} : \mathbb{R}^2 \to \mathbb{R}$ *defined so that* $\mathcal{S}\mathrm{qu}_i((\boldsymbol{y}, \boldsymbol{z})) = (\boldsymbol{y}, [\underline{s}^{\mathrm{C}}_{\boldsymbol{y}}(\underline{z}, \overline{z}), \overline{s}^{\mathrm{C}}_{\boldsymbol{y}}(\underline{z}, \overline{z})])$ *for each* $\boldsymbol{z} \in \mathbb{IR}$*. Then*

$$\nabla \underline{s}^{\mathrm{C}}_{\boldsymbol{y}}(\underline{z}, \overline{z}) = \begin{bmatrix} \frac{\partial \gamma_i}{\partial z}(\underline{z}, \underline{y}, p_{\boldsymbol{y}}) & 0 \end{bmatrix}, \qquad \textit{and} \qquad \nabla \overline{s}^{\mathrm{C}}_{\boldsymbol{y}}(\underline{z}, \overline{z}) = \begin{bmatrix} 0 & \frac{\partial \sigma_i}{\partial z}(\overline{z}, \overline{y}, p_{\boldsymbol{y}}) \end{bmatrix}.$$

Proof. This result follows immediately from the definition of the squashing operation. □

9.7 Convergence order

This section shows that both natural and unconstrained $\mathcal{C}^i$ McCormick extensions are $(1,2)$-convergent, provided that each employed univariate intrinsic function satisfies Assumptions 9.2.21 and 9.2.38. Thus, convex/concave relaxations based on these McCormick extensions exhibit second-order pointwise convergence. Each univariate function in Table 9.2 satisfies Assumption 9.2.38 except the absolute-value function; the nonsmoothness of the absolute-value function prevents second-order pointwise convergence from being achievable [11, Example 5].

Lemma 9.7.1. *The squashing operation is* $(1,2)$*-convergent for each fixed* $i \in \{1,2\}$.

Proof. Choose any $\mathcal{X} \in \mathbb{MR}$. If $\operatorname{wid} \boldsymbol{x}^{\mathrm{B}} = 0$, then

$$\operatorname{wid}_{\mathcal{M}} (\mathcal{S}\mathrm{qu}_i(\mathcal{X})) = 0 = \operatorname{wid}_{\mathcal{M}} \mathcal{X} + 2a_p(\operatorname{wid} \boldsymbol{x}^{\mathrm{B}})^2.$$

If $\operatorname{wid} \boldsymbol{x}^{\mathrm{B}} > 0$, then, using Lemma 9.3.5, and noting that $\mathcal{S}\mathrm{qu}_i(\mathcal{X}) \in \mathbb{MR}_{\mathrm{prop}}$, it follows that:

$$\begin{aligned}
\operatorname{wid}_{\mathcal{M}} (\mathcal{S}\mathrm{qu}_i(\mathcal{X})) &= \operatorname{wid} (\boldsymbol{belt}_i(\mathcal{X})) \\
&= \sigma_i(\overline{x}^{\mathrm{C}}, \overline{x}^{\mathrm{B}}, p_{\boldsymbol{x}^{\mathrm{B}}}) - \gamma_i(\underline{x}^{\mathrm{C}}, \underline{x}^{\mathrm{B}}, p_{\boldsymbol{x}^{\mathrm{B}}}) \\
&\leq \min\{\overline{x}^{\mathrm{C}} + p_{\boldsymbol{x}^{\mathrm{B}}}, \overline{x}^{\mathrm{B}}\} - \max\{\underline{x}^{\mathrm{C}} - p_{\boldsymbol{x}^{\mathrm{B}}}, \underline{x}^{\mathrm{B}}\} \\
&\leq \min\{\overline{x}^{\mathrm{C}} + p_{\boldsymbol{x}^{\mathrm{B}}}, \overline{x}^{\mathrm{B}} + p_{\boldsymbol{x}^{\mathrm{B}}}\} - \max\{\underline{x}^{\mathrm{C}} - p_{\boldsymbol{x}^{\mathrm{B}}}, \underline{x}^{\mathrm{B}} - p_{\boldsymbol{x}^{\mathrm{B}}}\} \\
&= \min\{\overline{x}^{\mathrm{C}}, \overline{x}^{\mathrm{B}}\} - \max\{\underline{x}^{\mathrm{C}}, \underline{x}^{\mathrm{B}}\} + 2p_{\boldsymbol{x}^{\mathrm{B}}} \\
&= \operatorname{wid}_{\mathcal{M}} \mathcal{X} + 2a_p(\operatorname{wid} \boldsymbol{x}^{\mathrm{B}})^2.
\end{aligned} \tag{9.10}$$

Noting that $\mathcal{X}$ was chosen arbitrarily, the required result follows. □

Lemma 9.7.2. *The multiplication operation described in Definition 9.3.19 is* $(1,2)$*-convergent for each fixed* $i \in \{1,2\}$.

Proof. Choose any $\mathbf{q} \equiv (\boldsymbol{q}_1, \boldsymbol{q}_2) \in \mathbb{IR}^2$, and any $\mathcal{X} \in (\mathbb{M}\boldsymbol{q}_1)_{\text{prop}}, \mathcal{Y} \in (\mathbb{M}\boldsymbol{q}_2)_{\text{prop}}$, in which case $\operatorname{wid}_{\mathcal{M}} \mathcal{X} \leq \operatorname{wid} \boldsymbol{x}^{\text{B}} \leq \operatorname{wid} \boldsymbol{q}_1$, and $\operatorname{wid}_{\mathcal{M}} \mathcal{Y} \leq \operatorname{wid} \boldsymbol{y}^{\text{B}} \leq \operatorname{wid} \boldsymbol{q}_1$. Construct the interval $\boldsymbol{z} \in \mathbb{IR}$ described in Definition 9.3.19. Define $\mathcal{Z} := (\boldsymbol{x}^{\text{B}}\boldsymbol{y}^{\text{B}}, \boldsymbol{z}) \in \mathbb{MR}_{\text{prop}}$ and $p := p_{\boldsymbol{x}^{\text{B}}\boldsymbol{y}^{\text{B}}}$ for notational convenience.

Applying Lemma 3.9.19 in [96], and noting that $\boldsymbol{x}_1^{\text{B}} = \boldsymbol{x}^{\text{B}}$ and $\boldsymbol{y}_1^{\text{B}} = \boldsymbol{y}^{\text{B}}$ by construction, there exist $a_1, a_2 > 0$ (which may depend on $\boldsymbol{q}_1$, but are independent of $\mathcal{X}$ and $\mathcal{Y}$) for which

$$\operatorname{wid}_{\mathcal{M}} (\mathcal{X} \bullet \mathcal{Y}) \leq a_1 \operatorname{wid}_{\mathcal{M}} (\mathcal{X}, \mathcal{Y}) + a_2 (\operatorname{wid} (\boldsymbol{x}^{\text{B}}, \boldsymbol{y}^{\text{B}}))^2.$$

(Recall that the symbol "$\bullet$" refers to the classical McCormick product described in Definition 9.2.31.) Define the following intermediate quantities:

$$\begin{aligned} n_1 &:= \underline{(\underline{y}^{\text{B}}\boldsymbol{x}^{\text{C}})} + \underline{(\underline{x}^{\text{B}}\boldsymbol{y}^{\text{C}})} - \underline{x}^{\text{B}}\underline{y}^{\text{B}}, & n_2 &:= \underline{(\overline{y}^{\text{B}}\boldsymbol{x}^{\text{C}})} + \underline{(\overline{x}^{\text{B}}\boldsymbol{y}^{\text{C}})} - \overline{x}^{\text{B}}\overline{y}^{\text{B}}, \\ n_3 &:= \overline{(\underline{y}^{\text{B}}\boldsymbol{x}^{\text{C}})} + \overline{(\overline{x}^{\text{B}}\boldsymbol{y}^{\text{C}})} - \overline{x}^{\text{B}}\underline{y}^{\text{B}}, & n_4 &:= \overline{(\overline{y}^{\text{B}}\boldsymbol{x}^{\text{C}})} + \overline{(\underline{x}^{\text{B}}\boldsymbol{y}^{\text{C}})} - \underline{x}^{\text{B}}\overline{y}^{\text{B}}. \end{aligned}$$

Using Lemma 9.3.11,

$$\begin{aligned} \operatorname{wid}_{\mathcal{M}} \mathcal{Z} &\leq \operatorname{wid} \boldsymbol{z} \\ &= \lambda_i(n_3, n_4, p) - \nu_i(n_1, n_2, p) \\ &\leq \tfrac{1}{2} (\min\{n_3 + p, n_4\} + \min\{n_3, n_4 + p\}) \\ &\qquad\qquad - \tfrac{1}{2} (\max\{n_1 - p, n_2\} + \min\{n_1, n_2 - p\}) \\ &\leq \min\{n_3 + p, n_4 + p\} - \max\{n_1 - p, n_2 - p\} \\ &= \min\{n_3, n_4\} - \max\{n_1, n_2\} + 2p \\ &= \operatorname{wid}_{\mathcal{M}} (\mathcal{X} \bullet \mathcal{Y}) + 2p. \end{aligned}$$

Define the absolute value of any interval $\boldsymbol{a} \in \mathbb{IR}$ as $|\boldsymbol{a}| := \max\{|\underline{a}|, |\overline{a}|\} \geq 0$. Using [77, Equation 4.3],

$$\operatorname{wid} (\boldsymbol{x}^{\text{B}}\boldsymbol{y}^{\text{B}}) \leq |\boldsymbol{x}^{\text{B}}| \operatorname{wid} \boldsymbol{y}^{\text{B}} + |\boldsymbol{y}^{\text{B}}| \operatorname{wid} \boldsymbol{x}^{\text{B}} \leq |\boldsymbol{q}_1| \operatorname{wid} (\boldsymbol{x}^{\text{B}}, \boldsymbol{y}^{\text{B}}).$$

Thus,

$$p \leq a_p(|\boldsymbol{q}_1| \operatorname{wid}(\boldsymbol{x}^{\mathrm{B}}, \boldsymbol{y}^{\mathrm{B}}))^2.$$

Combining the above results, Lemma 9.7.1, and (9.10),

$$\begin{aligned}
\operatorname{wid}_{\mathcal{M}}(\mathcal{X}\mathcal{Y}) &= \operatorname{wid}_{\mathcal{M}}(\mathcal{S}\mathrm{qu}_i(\mathcal{Z})) \\
&\leq \operatorname{wid}_{\mathcal{M}} \mathcal{Z} + 2p \\
&\leq \operatorname{wid}_{\mathcal{M}}(\mathcal{X} \bullet \mathcal{Y}) + 4p \\
&\leq a_1 \operatorname{wid}_{\mathcal{M}}(\mathcal{X}, \mathcal{Y}) + a_2(\operatorname{wid}(\boldsymbol{x}^{\mathrm{B}}, \boldsymbol{y}^{\mathrm{B}}))^2 + 4p \\
&\leq a_1 \operatorname{wid}_{\mathcal{M}}(\mathcal{X}, \mathcal{Y}) + (4a_p|\boldsymbol{q}_1|^2 + a_2)(\operatorname{wid}(\boldsymbol{x}^{\mathrm{B}}, \boldsymbol{y}^{\mathrm{B}}))^2,
\end{aligned}$$

which yields the required result, since a_1, a_2, a_p, and $|\boldsymbol{q}_1|$ are each independent of $\mathcal{X}$ and $\mathcal{Y}$. □

Theorem 9.7.3. *Given some $i^* \in \{1, 2\}$ and an MC-factorable function $\mathbf{f} : B \subset \mathbb{R}^n \to \mathbb{R}^m$ whose composed univariate intrinsic functions satisfy Assumptions 9.2.21 and 9.2.38 with $i := i^*$, any natural $\mathcal{C}^{i^*}$ McCormick extension $\mathcal{F} : \mathbb{M}B_{\text{prop}} \to \mathbb{M}\mathbb{R}^m$ of $\mathbf{f}$ is $(1, 2)$-convergent. Any unconstrained $\mathcal{C}^{i^*}$ McCormick extension $\mathcal{F}_{\text{unc}} : \mathbb{M}B \to \mathbb{M}\mathbb{R}^m$ of $\mathbf{f}$ is also $(1, 2)$-convergent.*

Proof. As discussed in [96, Section 3.9.7], the composition of $(1, 2)$-convergent functions is itself $(1, 2)$-convergent. The addition operation $+ : \mathbb{M}\mathbb{R}^2_{\text{prop}} \to \mathbb{M}\mathbb{R}_{\text{prop}}$ is $(1, 2)$-convergent [96, Lemma 3.9.17], as is any univariate intrinsic function which satisfies Assumption 9.2.38 [96, Lemma 3.9.23]. Lemmata 9.7.1 and 9.7.2 show that the squashing operation and the multiplication operation described in Definition 9.3.19 are each $(1, 2)$-convergent as well. Combining these results, $\mathcal{F}$ and $\mathcal{F}_{\text{unc}}$ are each $(1, 2)$-convergent. □

9.8 Implementation and examples

This section first discusses how to choose the parameter a_p in Definition 9.3.12 in accordance with numerical considerations. A C++ implementation of the re-

laxation theory in this chapter is then described, and is subsequently applied to various example problems for illustration.

9.8.1 Choosing the parameter a_p

When constructing a $\mathcal{C}^i$ McCormick extension of a function f, the parameter a_p in Definition 9.3.12 is only used if either f is described in terms of at least one product function, an unconstrained $\mathcal{C}^i$ McCormick extension is desired, or the constructions described in Remark 9.3.13 for pathological univariate intrinsic functions are required. If none of these circumstances apply, then there is no need to choose a_p.

Although the established $(1,2)$-convergence of $\mathcal{C}^i$ McCormick extensions is independent of a_p, larger values of a_p ultimately yield weaker relaxations $\phi_{f,\boldsymbol{x}}/\psi_{f,\boldsymbol{x}}$ when $\operatorname{wid}\boldsymbol{x}$ is large, making fathoming by value dominance less likely at the early stages of a branch-and-bound procedure for nonconvex optimization. On the other hand, smaller values of a_p yield relaxations that are theoretically $\mathcal{C}^i$, yet may differ (with respect to the L^2-norm) only marginally from a nondifferentiable function when $\operatorname{wid}\boldsymbol{x}$ is reduced.

Moreover, observe that in the results established in this chapter, there is no need for the same value of a_p to be used each time the function $p : \mathbb{IR} \to [0, +\infty)$ is invoked during construction of a particular $\mathcal{C}^i$ McCormick extension of a function. This notion provides a degree of freedom which can be exploited to ensure that the values of a_p employed are neither too great or too small, in accordance with the previous paragraph.

Now, it follows from Lemma 9.3.5 that for any $\mathcal{X} \equiv (\boldsymbol{x}^{\mathrm{B}}, \boldsymbol{x}^{\mathrm{C}}) \in \mathbb{MR}_{\mathrm{prop}}$ and each $i \in \{1,2\}$,

$$0 \leq \frac{\operatorname{wid}(\boldsymbol{belt}_i(\mathcal{X}))}{\operatorname{wid}\boldsymbol{x}^{\mathrm{B}}} - \frac{\operatorname{wid}\boldsymbol{x}^{\mathrm{C}}}{\operatorname{wid}\boldsymbol{x}^{\mathrm{B}}} \leq \max\left\{\frac{2p_{\boldsymbol{x}^{\mathrm{B}}}}{\operatorname{wid}\boldsymbol{x}^{\mathrm{B}}}, 1\right\} = \max\{2a_p \operatorname{wid}\boldsymbol{x}^{\mathrm{B}}, 1\}.$$

This sequence of inequalities suggests that the belt operation increases the ratio $\frac{\operatorname{wid}\boldsymbol{x}^{\mathrm{C}}}{\operatorname{wid}\boldsymbol{x}^{\mathrm{B}}}$ by at most $(2a_p \operatorname{wid}\boldsymbol{x}^{\mathrm{B}})$. Note that if $\frac{\operatorname{wid}\boldsymbol{x}^{\mathrm{C}}}{\operatorname{wid}\boldsymbol{x}^{\mathrm{B}}} = 1$, then, intuitively, the re-

laxation information contained in $\boldsymbol{x}^{\mathrm{C}}$ is simply returning the interval bounds $\boldsymbol{x}^{\mathrm{B}}$. If $\frac{\operatorname{wid}(\boldsymbol{belt}_i(\mathcal{X}))}{\operatorname{wid}\boldsymbol{x}^{\mathrm{B}}} \approx \frac{\operatorname{wid}\boldsymbol{x}^{\mathrm{C}}}{\operatorname{wid}\boldsymbol{x}^{\mathrm{B}}}$, then there is little numerical difference between the $\mathcal{C}^i$ McCormick relaxations and the classical McCormick relaxations.

In light of the above discussion, suppose that during execution of a branch-and-bound procedure, when any interval subdomain $\boldsymbol{x}$ is visited, then the $\mathcal{C}^i$ McCormick extension of a function demands evaluation of $p_{\boldsymbol{y}^{\mathrm{B}}(\boldsymbol{x})}$, where the interval-valued function $\boldsymbol{y}^{\mathrm{B}}$ is defined by the natural interval extension of the MC-factorable objective function. Due to inclusion monotonicity of natural interval extensions, $\operatorname{wid}(\boldsymbol{y}^{\mathrm{B}}(\boldsymbol{x}))$ decreases as $\operatorname{wid}\boldsymbol{x}$ decreases. Now, if $\boldsymbol{x}_0$ denotes the interval domain considered at the root node of the branch-and-bound procedure, the above discussion suggests setting

$$a_p \leftarrow \frac{b_p}{2\operatorname{wid}\boldsymbol{y}^{\mathrm{B}}(\boldsymbol{x}_0)} \tag{9.11}$$

for some constant b_p in the range $[0.01, 0.2]$. With this choice, the $\mathcal{C}^i$ McCormick extensions are not relaxed too much relative to the corresponding original natural McCormick extensions, and yet $(2a_p \operatorname{wid}\boldsymbol{y}^{\mathrm{B}}(\boldsymbol{x}))$ remains significantly greater than 0 (relative to a computer's typical numerical precision) even after several successive branches in the branch-and-bound procedure.

Lastly, note that $(2a_p \operatorname{wid}\boldsymbol{y}^{\mathrm{B}}(\boldsymbol{x})) \to 0^+$ in the limit $(\operatorname{wid}\boldsymbol{x}) \to 0^+$. If the quantity $(2a_p \operatorname{wid}\boldsymbol{y}^{\mathrm{B}}(\boldsymbol{x}))$ falls below some small tolerance $\epsilon > 0$, then affine relaxations defined either by the subgradients of the classical natural McCormick extensions or the gradients of $\mathcal{C}^i$ McCormick extensions may be preferable to the McCormick extensions themselves.

9.8.2 Implementation

A C++ implementation of $\mathcal{C}^i$ McCormick extension evaluation was developed by modifying version 1.0 of the header library MC++ [15] to carry out the methods in this chapter. This new implementation describes McCormick objects using a template class `mc::smoothMcC<T>`, which is a modified version of the class

mc::McCormick<T> defined by MC++. As in MC++, the templated argument T refers to the interval objects used by an employed interval arithmetic library. The specific modifications used to construct the class mc::smoothMcC<T> from the class mc::McCormick<T> from MC++ are as follows.

Firstly, static member variables _MCbp and _MCi were added to the class, so as to hold the values of the parameters b_p and $i \in \{1, 2\}$, respectively. These parameters can be set and retrieved using static member functions setBp, getBp, setI, and getI. Static member functions MCp, MCmu, dMCmu, MCgamma, MCsigma, MCnu, ddxMCnu, ddyMCnu, MClambda, ddxMClambda, and ddyMClambda were also included, to evaluate the functions p, μ_i, $\nabla\mu_i$, γ_i, σ_i, ν_i, $\frac{\partial\nu_i}{\partial x}$, $\frac{\partial\nu_i}{\partial y}$, λ_i, $\frac{\partial\lambda_i}{\partial x}$, and $\frac{\partial\lambda_i}{\partial y}$, respectively. The execution of MCp is detailed in the next paragraph. In the following description, let mcX denote an arbitrary mc::smoothMcC<T> objects representing a McCormick object $\mathcal{X}$. Member functions squash and p were added to the class, so that mcX.squash() replaces its calling member $\mathcal{X}$ with $\mathcal{S}\mathrm{qu}_i(\mathcal{X})$, and so that mcX.p() invokes MCp to return the value $p_{\boldsymbol{x}^{\mathrm{B}}}$. Using these constructions, McCormick-McCormick multiplication (via a friend function operator*(const& mc::smoothMcC<T>, const& mc::smoothMcC<T>)) was implemented according to Definition 9.3.19, with gradients propagated according to Theorem 9.6.7. The relaxations described in Examples 9.2.27, 9.2.28 and 9.2.29 were implemented by modifying the overloaded operations fabs and pow appropriately, along with a squaring function sqr that was implemented in MC++.

To implement evaluation of p via MCp according to the discussion in Section 9.8.1, a static member enum variable _apMode was added to the mc::smoothMcC<T> class, to describe whether the parameters a_p should be evaluated as if the root node in a branch-and-bound process is being visited, or whether a child node is being visited instead. If _apMode=SET_AP, which can be forced using a static void member function beginStoringAp, then each time p is evaluated, the parameter a_p is evaluated in the root-node mode described in Section 9.8.1, and the value of a_p is pushed onto the end of a static member std::vector<double> named _apList. To handle child nodes in a branch-and-bound process, when values of a_p have already been

stored in _apList, a static void member function beginRetrievingAp sets _apMode to GET_AP. In this mode, each time p is evaluated, the appropriate value of a_p is retrieved from _apList; the appropriate component of _apList to be retrieved is tracked using a static member std::vector<double>::const_iterator variable named _apListIterator.

Ultimately, given a user-supplied template subroutine f that is written as if its inputs and outputs are doubles or double arrays, the implementation described above permits natural $\mathcal{C}^i$ McCormick extensions of f to be evaluated using operator overloading, along with directional derivatives that are evaluated using the forward mode of automatic differentiation. To obtain unconstrained $\mathcal{C}^i$ McCormick extensions instead, the squash operation should first be applied to each mc::smoothMcC<T> input to f. The univariate intrinsic functions and operations described in Table 9.2 are all supported in this implementation.

9.8.3 Complexity analysis

Roughly, denote the computational cost of evaluating an MC-factorable function $f : X \subset \mathbb{R}^n \to \mathbb{R}$ using its factored representation as $\mathcal{C}ost(f)$. Observe that, when constructing the convex or concave relaxation suggested by a natural $\mathcal{C}^2$ McCormick extension for f, each addition, multiplication, and univariate intrinsic function in the factored representation of f is replaced with its $\mathcal{C}^2$ McCormick counterpart. Thus, there exists $\gamma_c > 0$ for which the computational cost of evaluating a $\mathcal{C}^2$ convex or concave relaxation of f is no greater than $\gamma_c \mathcal{C}ost(f)$. The parameter γ_c is independent of f, but depends on the library of univariate intrinsic functions considered.

Similarly, using standard complexity results for automatic differentiation [34], it follows that there exist similar library-dependent constants $\gamma_a, \gamma_t > 0$, satisfying the following claim. If the reverse mode of automatic differentiation is used to evaluate a subgradient of such a relaxation, then the cost of doing so is bounded above by $\gamma_a \mathcal{C}ost(f)$; if the forward mode is used instead, then the cost of evaluat-

ing this subgradient is bounded above by $n\gamma_t \mathcal{C}ost(f)$, where n denotes the domain dimension of f.

9.8.4 Examples

In this section, the implementation of $\mathcal{C}^2$ McCormick relaxation described in Section A.5.1 is applied to various example problems for illustration.

Example 9.8.1. *To illustrate the modified multiplication rule provided by Definition 9.3.19, consider the function $f : \mathbb{R}^2 \to \mathbb{R} : (x, y) \mapsto y(x^2 - 1)$, which is plotted in Figure 9-1(a). The function f is (real-)analytic but nonconvex on $\mathbf{z} := [-4, 4]^2 \subset \mathbb{R}^2$.*

Using MC++ [15], the classical McCormick convex relaxation of f was constructed on $\mathbf{z}$, and is plotted in Figure 9-1(b). This relaxation is not differentiable everywhere; this nondifferentiability is introduced via McCormick's rule for relaxing the product of terms whose signs change on the interval of interest. A natural $\mathcal{C}^2$ McCormick relaxation of f on $\mathbf{z}$ was constructed using the implementation described in Sections 9.8.1 and A.5.1, with $b_p := 0.2$; this relaxation is plotted in Figure 9-1(c). Observe that this relaxation is visibly differentiable (and is, in fact, $\mathcal{C}^2$), but is otherwise qualitatively similar to the classical McCormick relaxation. The classical McCormick relaxation dominates its $\mathcal{C}^2$ counterpart on $\mathbf{z}$.

For comparison, the αBB relaxation of f on $\mathbf{z}$ with a nonuniform diagonal shift matrix that minimizes maximum separation distance [1] was computed directly to be:

$$f^\alpha : (x, y) \mapsto f(x, y) + 8(x^2 - 16) + 4(y^2 - 16),$$

and is plotted in Figure 9-1(d). The obtained αBB relaxation is analytic, and has a minimum at $(x^*, y^*) := (0, 0.125)$. Observe that $f^\alpha(x^*, y^*) = -192.0625$, which is less than the lower bound $\underline{\tilde{f}}(\mathbf{z}) = -60$ provided by the natural interval extension of f on $\mathbf{z}$. This interval lower bound coincides with $\min_{(x,y)\in\mathbf{z}} f(x, y)$, and is dominated on $\mathbf{z}$ by both the constructed classical McCormick relaxation and the constructed $\mathcal{C}^2$ McCormick relaxation.

Example 9.8.2. *To illustrate the handling of the absolute-value function according to Example 9.2.28, consider the function*

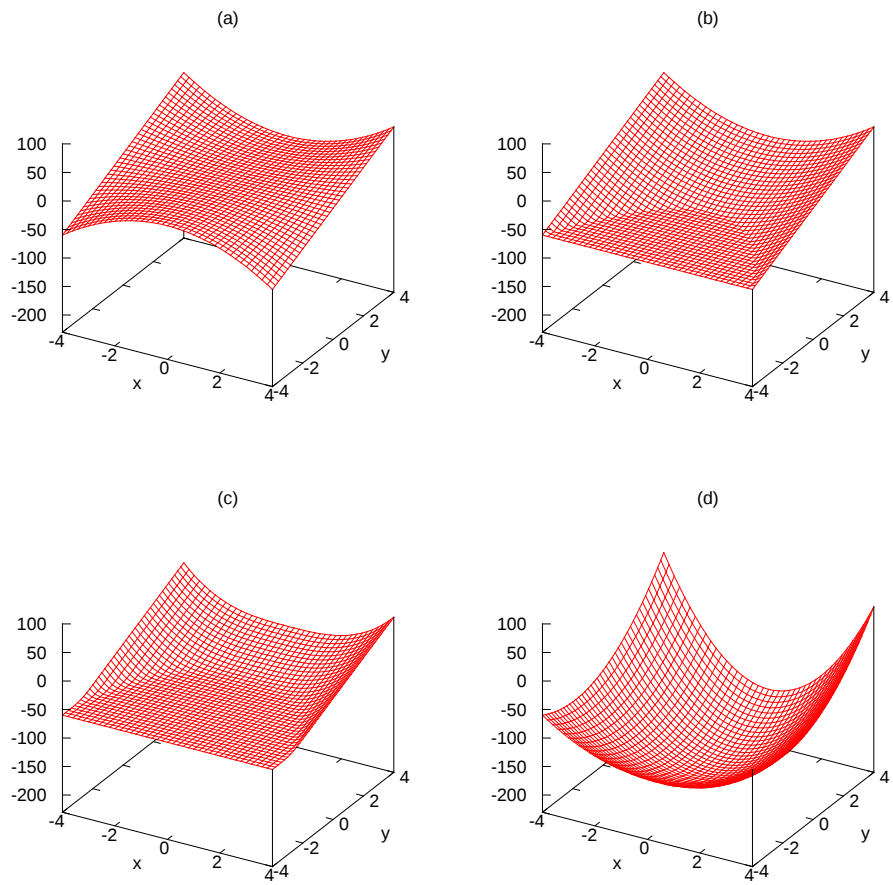

Figure 9-1: The function $f : (x, y) \mapsto y(x^2 - 1)$ and its convex relaxations on $[-4, 4]^2$: (a) the function f, (b) the classical McCormick relaxation of f, (c) a $\mathcal{C}^2$ McCormick relaxation of f, and (d) the αBB relaxation of f that minimizes maximum separation distance.

$$g : \mathbb{R}^2 \to \mathbb{R} : (x, y) \mapsto |x + 1| + |x - 1| - |x + y - 1| - |x - y + 1|, \tag{9.12}$$

which is plotted in Figure 9-2(a). The function g is piecewise affine, and is nonconvex on $\mathbf{z} := [-2, 2]^2 \subset \mathbb{R}^2$.

As in the previous example, the classical McCormick convex relaxation of g on $\mathbf{z}$ was constructed using MC++, and is plotted in Figure 9-2(b); this relaxation is readily verified to be piecewise affine. The $\mathcal{C}^2$ McCormick relaxation of g on $\mathbf{z}$ was evaluated using the implementation described in Section A.5.1, and is plotted in Figure 9-2(c). Since there

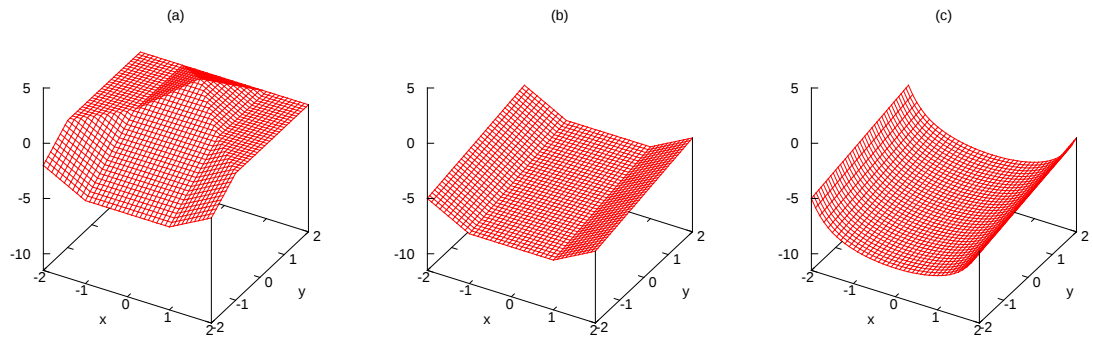

Figure 9-2: The function g described in (9.12) and its convex relaxations on $[-2,2]^2$: (a) the function g, (b) the classical McCormick relaxation of g, and (c) a $\mathcal{C}^2$ McCormick relaxation of g.

does not exist a scheme of estimators satisfying Assumption 9.2.38 for the absolute-value function, the generated McCormick and $\mathcal{C}^2$ McCormick relaxations are not guaranteed to be pointwise convergent of order 2.

Example 9.8.3. *To illustrate the handling of the squaring function $z \mapsto z^2$ according to Example 9.2.27, consider the function*

$$h : \mathbb{R}^2 \to \mathbb{R} : (x, y) \mapsto (xy - 1)^2, \tag{9.13}$$

which is plotted in Figure 9-3(a). The function h is analytic and nonconvex on $\mathbf{z} := [-2,2]^2 \subset \mathbb{R}^2$.

The classical McCormick convex relaxation h^{cv} of h on $\mathbf{z}$ was evaluated using MC++, along with a subgradient at each point. This relaxation h^{cv} is plotted in Figure 9-3(b); x- and y-components of the evaluated subgradients of h^{cv} are plotted in Figures 9-3(c) and 9-3(d), respectively. As a function of (x, y), the evaluated subgradient is evidently not differentiable everywhere; it follows that h^{cv} is not twice-differentiable, let alone $\mathcal{C}^2$. This example illustrates that, even though the squaring function is convex, considering the squaring function as its own convex relaxation can yield failures of twice-continuous differentiability. This observation motivates Assumption 9.2.21 and Example 9.2.27.

A natural $\mathcal{C}^2$ McCormick relaxation $\tilde{h}^{cv}$ of h on $\mathbf{z}$ was constructed using the imple-

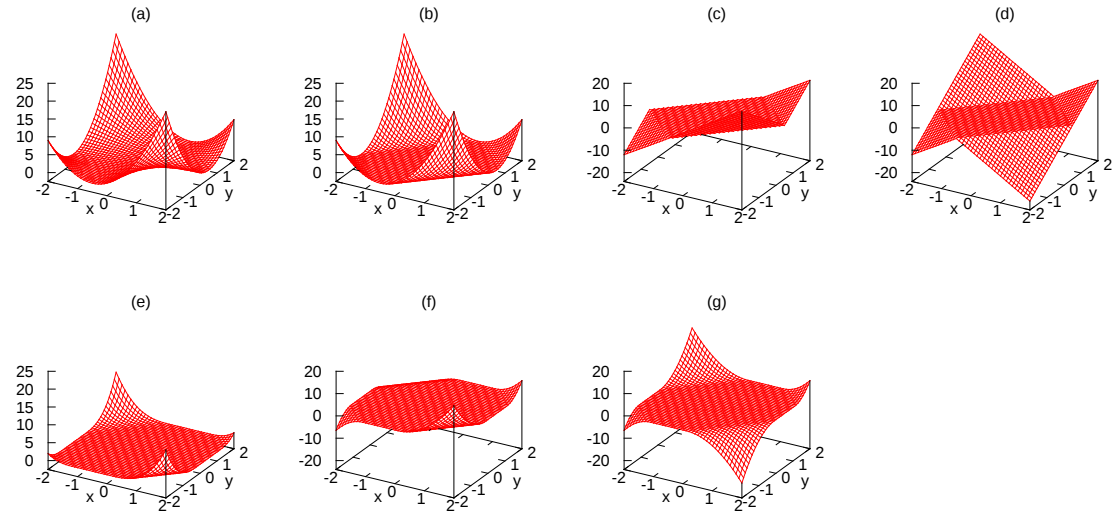

Figure 9-3: The function h described in (9.13) and its convex relaxations and associated subgradients on $[-2,2]^2$: (a) the function h, (b) the classical McCormick relaxation h^{cv} of h, (c) the x-component of some subgradient of h^{cv}, (d) the y-component of some subgradient of h^{cv}, (e) a $\mathcal{C}^2$ McCormick relaxation $\tilde{h}^{\text{cv}}$ of h, (f) the partial derivative $\frac{\partial \tilde{h}^{\text{cv}}}{\partial x}$, and (g) the partial derivative $\frac{\partial \tilde{h}^{\text{cv}}}{\partial y}$.

mentation described in Sections 9.8.1 and A.5.1, with $b_p := 0.2$; this relaxation is plotted in Figure 9-1(e). Gradients of $\tilde{h}^{cv}$ were also evaluated using the described implementation; the partial derivatives $\frac{\partial \tilde{h}^{cv}}{\partial x}$ and $\frac{\partial \tilde{h}^{cv}}{\partial y}$ are plotted in Figures 9-3(f) and 9-3(g), respectively. These partial derivatives appear to be differentiable, and are indeed $\mathcal{C}^1$.

Example 9.8.4. *This example illustrates the second-order pointwise convergence of the $\mathcal{C}^2$ McCormick relaxations presented in this chapter. As in [11, Example 7], consider the function*

$$f : \mathbb{R}_+ \to \mathbb{R} : x \mapsto (z - z^2)(\log z + e^{-z})$$

on intervals of the form $[0.5 - \epsilon, 0.5 + \epsilon]$ for $\epsilon \in (0, 0.2)$. The function f is plotted in Figure 9-4, together with a series of $\mathcal{C}^2$ relaxations $\psi_{\boldsymbol{x}(\epsilon)}$ of f constructed using the implementation described in Sections 9.8.1 and A.5.1, on intervals $\boldsymbol{x} \in \{[0.5 - \epsilon, 0.5 + \epsilon] : \epsilon = 0.4(2^k), k \in \{1, \ldots, 20\}\}$, with the parameters in (9.11) set to $\boldsymbol{x}_0 := [0.3, 0.7]$ and $b_p := 0.2$.

For the considered values of ϵ, Figure 9-4(b) plots $\sup_{x \in \boldsymbol{x}(\epsilon)} (f(x) - \psi_{\boldsymbol{x}(\epsilon)}(x))$ against

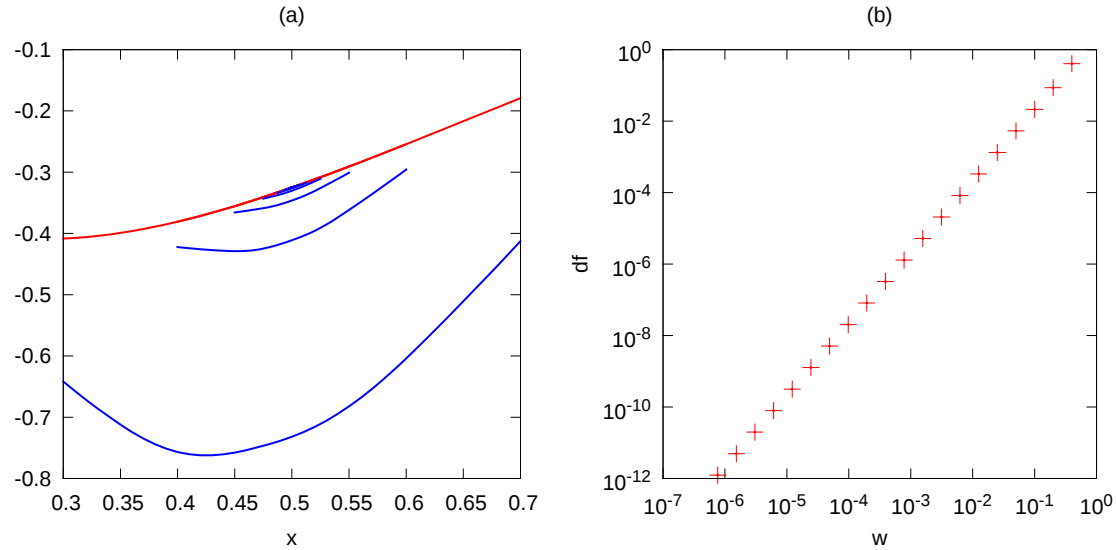

Figure 9-4: (a) The function f described in Example 9.8.4 (red) and its $\mathcal{C}^2$ convex relaxations $\psi_{\boldsymbol{x}(\epsilon)}$ of f on intervals $\boldsymbol{x}(\epsilon) := [0.5-\epsilon, 0.5+\epsilon]$ for $\epsilon \in \{0.4(2^{-k}) : k \in \mathbb{N}\}$ (blue), and (b) a plot of $df := \sup_{x\in\boldsymbol{x}(\epsilon)}(f(x) - \psi_{\boldsymbol{x}(\epsilon)}(x))$ vs. $w := \operatorname{wid}\boldsymbol{x}(\epsilon) = 2\epsilon$.

$\operatorname{wid}\boldsymbol{x}(\epsilon)$ *on a logarithmic scale; the slope of this plot suggests second-order pointwise convergence of the convex relaxation* $\psi_{\boldsymbol{x}}(\epsilon)$ *to* f *as* $\epsilon \to 0^+$.

9.9 Conclusions

A variant of McCormick's relaxation scheme has been presented, which produces $\mathcal{C}^2$ convex and concave relaxations of a provided MC-factorable function, while retaining the computational benefits of McCormick's method. Gradients are readily evaluated for the provided relaxations using standard automatic differentiation methods. As an avenue for possible future work, we expect that the methods in this chapter are compatible with an established scheme for reverse propagaion of McCormick relaxations [120], and could yield a scheme for constructing $\mathcal{C}^2$ relaxations for implicit functions.

As an open problem, observe that the methods in this chapter do not extend immediately to the multivariate relaxations described by Tsoukalas and Mitsos [114]. Such an extension would be desirable, since the multivariate product relaxations

are tighter than the classical McCormick product relaxation described in Definition 9.2.31.

Chapter 10

Conclusions

In this thesis, numerical methods have been developed and implemented to evaluate Nesterov's lexicographic derivatives for composite L-smooth functions, and for the unique solutions of parametric ODE systems with L-smooth right-hand side functions. The methods presented in Appendix A and the conference proceedings [53] are the first tractable methods for computing generalized derivatives for a broad class of vector-valued composite nonsmooth functions, and the method presented in Chapter 7 is the first method for computing a useful generalized derivative for a broad class of nonsmooth dynamic systems. These methods broaden the scope of equation-solving problems and optimization problems that may be approached using semismooth Newton methods, bundle methods, or their variants.

This thesis has also presented several new theoretical results in nonsmooth sensitivity analysis. The LD-derivative was introduced as a tool to facilitate evaluation of lexicographic derivatives. In Chapter 3, lexicographic derivatives of any L-smooth function were shown to be plenary Jacobian elements, and are also B-subdifferential elements when the underlying function is piecewise differentiable in the sense of Scholtes [97]. Chapter 5 presented the first theoretical description of a useful generalized derivative for a parametric ODE system in terms of an auxiliary ODE system, thus extending classical sensitivity results for smooth ODE systems to the nonsmooth case. Chapter 8 develops lexicographic deriva-

tives for local inverse functions and implicit functions that are described in terms of L-smooth functions, and exploits these to describe lexicographic derivatives for certain pathological hybrid discrete/continuous systems that cannot be treated by classical sensitivity theory [30].

Numerical tools were also developed to mitigate the impact of nonsmoothness on certain problems. McCormick's classic method [74] for computing convex relaxations of composite functions was weakened in Chapter 9 to yield a variant that computes twice-continuous differentiable relaxations, while preserving the various computational advantages of McCormick's original method. In Chapter 6, Clarke's sufficient condition [16, Theorem 7.4.1] for parametric differentiability of the solution of a parametric nonsmooth ODE was shown to take a numerically tractable form when the ODE right-hand side is a finite composition of analytic functions and absolute-value functions.

10.1 Avenues for future work

The work in this thesis suggests several theoretical and numerical avenues for future work. As an open theoretical question, it is unknown if lexicographic derivatives are always B-subdifferential elements, as was shown in Chapter 3 for the special case of piecewise differentiable functions. Even if they are not, these two generalized derivatives may nevertheless behave similarly when used in numerical methods for equation-solving or optimization, thus mirroring the relationship between the Clarke Jacobian and its plenary hull. As another open theoretical question, it is currently unknown whether the solution of a parametric ODE with a piecewise differentiable right-hand side function is itself piecewise differentiable with respect to the ODE parameters. If true, then these ODE solutions would be subject to the results in Section 3.3, and could be treated using Kojima and Shindo's Newton method [65], which exhibits local Q-quadratic convergence if its invertibility requirements are met.

The numerical method for dynamic LD-derivative computation in Chapter 7

depends strongly on the ODE right-hand side function being a known composition of analytic functions and absolute value functions. It would be useful to extend this numerical method to include dynamic systems with linear programs embedded, without having to express these linear programs in an abs-factored form; numerical methods for ODE integration have already been extended in such a manner [36]. It may be possible to accommodate ODE right-hand side functions that are L-smooth but not piecewise differentiable: perhaps including the Euclidean norm.

Taken together, the results of Chapters 5, 7, and 8 suggest that it may be possible to develop a method for computing parametric LD-derivatives for unique solutions of certain parametric index-1 differential-algebraic equation systems; such a method would, inevitably, be a generalization of the method in Chapter 7.

Appendix A

Previous methods for Clarke Jacobian element evaluation

For reference, this appendix reproduces most of the article [54]. The numerical methods developed in this article and the related conference proceedings [53] were the first tractable, accurate methods for evaluation of a generalized derivative for a broad class of nonsmooth vector-valued functions. Unlike the methods in Chapter 4, the methods in this appendix require storage of the computational graph of the function under consideration; they are not *tapeless*. The theory underlying these methods is used to obtain certain results in Chapters 2 and 3.

Note that the *elemental $\mathcal{PC}^1$ functions* defined in this appendix are different from the elemental $\mathcal{PC}^1$ functions considered in Chapter 2.

A.1 Mathematical background

This section presents key theoretical results from polyhedral theory, nonsmooth analysis, and the theory of piecewise differentiable functions. These results will be useful in formulating and validating the methods developed in this work. Apart from the development of *hyperplane normal sets* in Sections A.1.1 and A.1.3, this section echoes the background presented in [53].

General notational conventions used in this work are as follows. The Euclidean

metric spaces considered are equipped with the Euclidean norm $\|\cdot\|$. If a function $f : X \to Y$ satisfies a local property P at every $x \in X$, then f is said to satisfy P, without reference to any particular $x \in X$.

For any column vector $\mathbf{x} \in \mathbb{R}^n$, if $\mathbf{e}_{(1)}, \ldots, \mathbf{e}_{(n)}$ are the unit coordinate vectors in $\mathbb{R}^n$, then x_i denotes the inner product $\langle \mathbf{e}_{(i)}, \mathbf{x} \rangle \in \mathbb{R}$ for each $i \in \{1, \ldots, n\}$. An equivalent representation of $\mathbf{x}$ is then $(x_1, \ldots, x_n)$.

Given an open set $X \subset \mathbb{R}^n$, a function $\mathbf{f} : X \to \mathbb{R}^m$, some $\mathbf{x} \in X$, and some $\mathbf{d} \in \mathbb{R}^n$, if the one-sided limit

$$\lim_{t \to 0^+} \frac{\mathbf{f}(\mathbf{x} + t\mathbf{d}) - \mathbf{f}(\mathbf{x})}{t}$$

exists, then this limit is the *directional derivative* of $\mathbf{f}$ at $\mathbf{x}$ in the direction $\mathbf{d}$, and is denoted by $\mathbf{f}'(\mathbf{x}; \mathbf{d})$. If this limit exists and is finite for all $\mathbf{d} \in \mathbb{R}^n$, then $\mathbf{f}$ is *directionally differentiable* at $\mathbf{x}$.

Given an open set $X \subset \mathbb{R}^n$, a function $\mathbf{f} : X \to \mathbb{R}^m$ is (Fréchet)-differentiable at $\mathbf{x} \in X$ if there exists a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ for which

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

In this case, $\mathbf{A}$ is the unique *Jacobian matrix* of $\mathbf{f}$ at $\mathbf{x}$, and is denoted by $\mathbf{J}\mathbf{f}(\mathbf{x})$. If $\mathbf{f}$ is differentiable at $\mathbf{x}$, then it is also directionally differentiable at $\mathbf{x}$, with the directional derivative:

$$\mathbf{f}'(\mathbf{x}; \mathbf{d}) = \mathbf{J}\mathbf{f}(\mathbf{x})\, \mathbf{d}, \qquad \forall \mathbf{d} \in \mathbb{R}^n. \tag{A.1}$$

Given a function $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ and some $\mathbf{x} \in X$, $\mathbf{f}$ is *continuously differentiable* ($\mathcal{C}^1$) at $\mathbf{x}$ if there exists an open set $N \subset X$ such that $\mathbf{x} \in N$, $\mathbf{f}$ is differentiable at each $\mathbf{y} \in N$, and $\mathbf{J}\mathbf{f}$ is continuous at $\mathbf{x}$.

A.1.1 Polyhedral theory

Given a set $S \subset \mathbb{R}^n$, the *interior*, *closure*, and *convex hull* of S are denoted by $\operatorname{int}(S)$, $\operatorname{cl}(S)$, and $\operatorname{conv} S$, respectively. If S is nonempty, then the *convex cone* generated

by S is the set of nonnegative combinations of elements of S, and is denoted by $\operatorname{cone} S$. The number of elements in a finite set S is denoted by $|S|$. If S is finite and nonempty, then $\operatorname{cone} S$ is a *polyhedral cone*, and is closed and convex. Any polyhedral cone in $\mathbb{R}^n$ can be represented equivalently as the set $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \leq \mathbf{0}\}$ for some real-valued matrix $\mathbf{A}$ of appropriate dimensions [126].

A *partition* of a discrete set S is a collection of mutually disjoint sets whose union is S. A partition of a connected set $\hat{S}$ is a collection of sets whose union is $\hat{S}$, but whose interiors are mutually disjoint. A *conical subdivision* of $\mathbb{R}^n$ is a partition of $\mathbb{R}^n$ comprising finitely many polyhedral cones with nonempty interior.

Rather than deal with a particular conical subdivision Λ explicitly, it is more convenient to work with the halfspaces of $\mathbb{R}^n$ whose intersections describe the polyhedral cones in Λ. This motivates the following lemmas and definition, which use similar notation to [126, Chapter 7].

Lemma A.1.1. *Given a conical subdivision Λ of $\mathbb{R}^n$, there exists a finite subset $\mathcal{H} := \{\mathbf{a}_{(1)}, \ldots, \mathbf{a}_{(p)}\} \subset \mathbb{R}^n$ such that for any cone $\sigma \in \Lambda$, there exists a vector $\mathbf{s} \in \{-1, 0, 1\}^p$ such that $\sigma = \bigcap_{r=1}^{p} \{\mathbf{x} \in \mathbb{R}^n : s_r \langle \mathbf{a}_{(r)}, \mathbf{x} \rangle \leq 0\}$.*

Proof. For each $\sigma \in \Lambda$, there exists a matrix $\mathbf{A}_\sigma$ of appropriate dimensions for which $\sigma = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}_\sigma \mathbf{x} \leq \mathbf{0}\}$. Thus, if $\mathcal{H}$ is defined to be

$$\mathcal{H} = \bigcup_{\sigma \in \Lambda} \{\mathbf{a} \in \mathbb{R}^n : \mathbf{a}^{\mathrm{T}} \text{ is a row of } \mathbf{A}_\sigma\},$$

then $\mathcal{H}$ is finite, and so its elements may be enumerated as $\mathbf{a}_{(1)}, \ldots, \mathbf{a}_{(p)}$ for some $p \in \mathbb{N}$.

Now, given any $\bar{\sigma} \in \Lambda$, each row of $\mathbf{A}_{\bar{\sigma}}$ is a transposed element of $\mathcal{H}$. As a result, if a vector $\mathbf{s} \in \{-1, 0, 1\}^p$ is constructed so that for each $r \in \{1, \ldots, p\}$,

$$s_r = \begin{cases} 1 & \text{if } \mathbf{a}_{(r)}^{\mathrm{T}} \text{ is a row of } \mathbf{A}_{\bar{\sigma}}, \\ 0 & \text{otherwise,} \end{cases}$$

then $\bar{\sigma} = \bigcap_{r=1}^{p} \{\mathbf{x} \in \mathbb{R}^n : s_r \langle \mathbf{a}_{(r)}, \mathbf{x} \rangle \leq 0\}$, as required. □

Definition A.1.2. *Given a conical subdivision Λ of $\mathbb{R}^n$, a* hyperplane normal set *corresponding to Λ is a set $\mathcal{H} \subset \mathbb{R}^n$ satisfying the description in the statement of Lemma A.1.1.*

Remark A.1.3. *For a given conical subdivision Λ of $\mathbb{R}^n$, the proof of Lemma A.1.1 shows that the lemma remains true with the additional restriction that each $\mathbf{s} \in \{0,1\}^p$. Nevertheless, permitting $\mathbf{s}$ to be chosen from the set $\{-1,0,1\}^p$ can yield hyperplane normal sets with fewer elements, such as those discussed in Section A.2.1.*

A.1.2 Nonsmooth analysis

A directionally differentiable function need not be smooth, as the following example shows.

Example A.1.4. *For the absolute value function* $\mathrm{abs} : \mathbb{R} \to \mathbb{R} : x \mapsto |x|$, *it is readily verified that for each $x, d \in \mathbb{R}$,*

$$\mathrm{abs}'(x;d) = \begin{cases} d & \text{if } x > 0, \text{ or if } x = 0 \text{ and } d \geq 0, \\ -d & \text{if } x < 0, \text{ or if } x = 0 \text{ and } d < 0. \end{cases} \tag{A.2}$$

Definition A.1.5. *Given an open set $X \subset \mathbb{R}^n$, some $\mathbf{x} \in X$, and a locally Lipschitz continuous function $\mathbf{f} : X \to \mathbb{R}^m$, let $S \subset X$ be the set on which $\mathbf{f}$ is not differentiable. The* Bouligand (B-)subdifferential $\partial_{\mathrm{B}}\mathbf{f}(\mathbf{x})$ *of $\mathbf{f}$ at $\mathbf{x}$ is then defined as*

$$\partial_{\mathrm{B}}\mathbf{f}(\mathbf{x}) = \Big\{ \mathbf{H} \in \mathbb{R}^{m\times n} : \mathbf{H} = \lim_{i\to\infty} \mathbf{J}\mathbf{f}(\mathbf{x}_{(i)}) \\ \text{for some sequence } \{\mathbf{x}_{(i)}\}_{i\in\mathbb{N}} \text{ in } X\backslash S \text{ such that } \lim_{i\to\infty} \mathbf{x}_{(i)} = \mathbf{x} \Big\}.$$

The (Clarke) generalized Jacobian $\partial\mathbf{f}(\mathbf{x})$ *of $\mathbf{f}$ at $\mathbf{x}$ is the convex hull of $\partial_{\mathrm{B}}\mathbf{f}(\mathbf{x})$ [16]. Both $\partial_{\mathrm{B}}\mathbf{f}(\mathbf{x})$ and $\partial\mathbf{f}(\mathbf{x})$ exist, are unique, and are nonempty. If $\mathbf{f}$ is differentiable at $\mathbf{x}$, then $\mathbf{J}\mathbf{f}(\mathbf{x}) \in \partial\mathbf{f}(\mathbf{x})$. If $\mathbf{f}$ is $\mathcal{C}^1$ at $\mathbf{x}$, then $\partial_{\mathrm{B}}\mathbf{f}(\mathbf{x}) = \partial\mathbf{f}(\mathbf{x}) = \{\mathbf{J}\mathbf{f}(\mathbf{x})\}$.*

Computing generalized Jacobian elements for composite functions is a nontrivial task, since the generalized Jacobian satisfies calculus rules as inclusions instead of equations [16].

A.1.3 Piecewise differentiable functions

As defined rigorously in Definition A.1.6, piecewise differentiable functions include a broad range of nonsmooth functions, yet preserve many useful properties of $\mathcal{C}^1$ functions. Unless otherwise noted, the definitions and properties presented in this subsection are as stated and proven in [97].

Definition A.1.6. *Given an open set $X \subset \mathbb{R}^n$, a function $\mathbf{f} : X \to \mathbb{R}^m$ is* piecewise differentiable *($\mathcal{PC}^1$) at $\mathbf{x} \in X$ if there exists an open neighborhood $N \subset X$ of $\mathbf{x}$ such that $\mathbf{f}$ is continuous on N, and such that there exists a finite collection $\mathcal{F}_{\mathbf{f}}(\mathbf{x})$ of $\mathcal{C}^1$ functions which map N into $\mathbb{R}^m$ and satisfy*

$$\mathbf{f}(\mathbf{y}) \in \{\mathbf{f}^*(\mathbf{y}) : \mathbf{f}^* \in \mathcal{F}_{\mathbf{f}}(\mathbf{x})\}, \qquad \forall \mathbf{y} \in N. \tag{A.3}$$

The functions $\mathbf{f}^* \in \mathcal{F}_{\mathbf{f}}(\mathbf{x})$ are called selection functions *for $\mathbf{f}$ around $\mathbf{x}$, and a collection $\mathcal{F}_{\mathbf{f}}(\mathbf{x})$ of selection functions satisfying* (A.3) *is called a* sufficient collection of selection functions *for $\mathbf{f}$ around $\mathbf{x}$.*

If there exists a sufficient collection of selection functions for $\mathbf{f}$ which are each linear (i.e. affine and homogeneous), then $\mathbf{f}$ is piecewise linear *($\mathcal{PL}$).*

Remark A.1.7. *Any $\mathcal{C}^1$ or $\mathcal{PL}$ function is trivially $\mathcal{PC}^1$. The* abs *function mentioned in Example A.1.4 is $\mathcal{PL}$, since the functions $y \mapsto y$ and $y \mapsto -y$ are a sufficient collection of selection functions for* abs *around any domain point.*

Lemma A.1.8. *Any $\mathcal{PC}^1$ function $\mathbf{f} : X \to \mathbb{R}^m$ on an open set $X \subset \mathbb{R}^n$ exhibits the following properties [97, Corollary 4.1.1, Proposition 4.1.3, and Theorem 3.1.1]:*

1. *$\mathbf{f}$ is locally Lipschitz continuous.*

2. *$\mathbf{f}$ is directionally differentiable, and $\mathbf{f}'(\mathbf{x}; \cdot)$ is $\mathcal{PL}$ for any fixed $\mathbf{x} \in X$.*

3. *Given an open set $Y \subset \mathbb{R}^m$ containing the range of $\mathbf{f}$, and a $\mathcal{PC}^1$ function $\mathbf{g} : Y \to \mathbb{R}^\ell$, the composite function $\mathbf{h} : X \to \mathbb{R}^\ell : \mathbf{x} \mapsto \mathbf{g} \circ \mathbf{f}(\mathbf{x})$ is also $\mathcal{PC}^1$. Moreover, the directional derivative of $\mathbf{h}$ satisfies the chain rule:*

$$\mathbf{h}'(\mathbf{x}; \mathbf{d}) = \mathbf{g}'(\mathbf{f}(\mathbf{x}); \mathbf{f}'(\mathbf{x}; \mathbf{d})), \qquad \forall \mathbf{x} \in X, \quad \forall \mathbf{d} \in \mathbb{R}^n. \tag{A.4}$$

Definition A.1.9. *Given a sufficient collection $\mathcal{F}_{\mathbf{f}}(\mathbf{x})$ of selection functions for a $\mathcal{PC}^1$ function $\mathbf{f} : X \to \mathbb{R}^m$ at $\mathbf{x} \in X$, a selection function $\mathbf{f}^* \in \mathcal{F}_{\mathbf{f}}(\mathbf{x})$ is* essentially active *for $\mathbf{f}$ at $\mathbf{x}$ if $\mathbf{x} \in \mathrm{cl}(\mathrm{int}(\{\mathbf{y} \in X : \mathbf{f}(\mathbf{y}) = \mathbf{f}^*(\mathbf{y})\}))$.*

Lemma A.1.10. *Given an open set $X \subset \mathbb{R}^n$, a $\mathcal{PC}^1$ function $\mathbf{f} : X \to \mathbb{R}^m$, and some $\mathbf{x} \in X$, $\mathbf{f}$ exhibits the following properties involving essentially active selection functions [97, Propositions 4.1.1, 4.1.3, and A.4.1]:*

1. *There exists a sufficient collection $\mathcal{E}_{\mathbf{f}}(\mathbf{x})$ of selection functions for $\mathbf{f}$ at $\mathbf{x}$ which are each essentially active at $\mathbf{x}$. Thus, any selection function not in $\mathcal{E}_{\mathbf{f}}(\mathbf{x})$ may be discarded without loss of generality.*

2. *For any $\mathbf{d} \in \mathbb{R}^n$, the directional derivative of $\mathbf{f}$ at $\mathbf{x}$ in the direction $\mathbf{d}$ satisfies:*

$$\mathbf{f}'(\mathbf{x}; \mathbf{d}) \in \{\mathbf{J}\mathbf{f}^*(\mathbf{x})\, \mathbf{d} : \mathbf{f}^* \in \mathcal{E}_{\mathbf{f}}(\mathbf{x})\}.$$

3. *The B-subdifferential of $\mathbf{f}$ at $\mathbf{x}$ satisfies:*

$$\partial_{\mathrm{B}}\mathbf{f}(\mathbf{x}) = \{\mathbf{J}\mathbf{f}^*(\mathbf{x}) : \mathbf{f}^* \in \mathcal{E}_{\mathbf{f}}(\mathbf{x})\} \subset \partial\mathbf{f}(\mathbf{x}).$$

As defined in the subsequent lemma and definition, *conically active selection functions* are introduced in this work to describe the essentially active selection functions for a $\mathcal{PC}^1$ function $\mathbf{f}$ that are necessary to define the directional derivatives of $\mathbf{f}$.

Lemma A.1.11. *Given an open set $X \subset \mathbb{R}^n$, a $\mathcal{PC}^1$ function $\mathbf{f} : X \to \mathbb{R}^m$, and a vector $\mathbf{x} \in X$, there exists a conical subdivision $\Lambda_{\mathbf{f}}(\mathbf{x})$ of $\mathbb{R}^n$ such that for each polyhedral cone $\sigma \in \Lambda_{\mathbf{f}}(\mathbf{x})$, there is an essentially active selection function $\mathbf{f}_\sigma \in \mathcal{E}_{\mathbf{f}}(\mathbf{x})$ for which*

$$\mathbf{f}'(\mathbf{x}; \mathbf{d}) = \mathbf{J}\mathbf{f}_\sigma(\mathbf{x})\, \mathbf{d}, \qquad \forall \mathbf{d} \in \sigma. \tag{A.5}$$

Proof. The result follows immediately from Property 2 in Lemma A.1.8, Property 2 in Lemma A.1.10, and [97, Proposition 2.2.3]. □

Definition A.1.12. *A conical subdivision $\Lambda_{\mathbf{f}}(\mathbf{x})$ as described in Lemma A.1.11 is called an* active conical subdivision *for* $\mathbf{f}$ *at* $\mathbf{x}$*. Each cone $\sigma \in \Lambda_{\mathbf{f}}(\mathbf{x})$ is an* active cone *for* $\mathbf{f}$ *at* $\mathbf{x}$*. For each active cone σ, an essentially active selection function $\mathbf{f}_\sigma$ satisfying* (A.5) *is called a* conically active selection function *for* $\mathbf{f}$ *at* $\mathbf{x}$ *corresponding to σ. A hyperplane normal set $H_{\mathbf{f}}(\mathbf{x})$ corresponding to $\Lambda_{\mathbf{f}}(\mathbf{x})$ is called an* active normal set *for* $\mathbf{f}$ *at* $\mathbf{x}$*.*

The following example shows that active conical subdivisions and active normal sets for $\mathbf{f}$ at $\mathbf{x}$ need not be unique.

Example A.1.13. *Consider an arbitrary $\mathcal{C}^1$ function $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$. Any conical subdivision Λ of $\mathbb{R}^n$ is an active conical subdivision for $\mathbf{f}$ at any particular $\mathbf{x} \in X$, since* (A.5) *is satisfied for each $\sigma \in \Lambda$ with $\mathbf{f}_\sigma := \mathbf{f}$. In this case, any hyperplane normal set corresponding to Λ is an active normal set for $\mathbf{f}$ at $\mathbf{x}$.*

As the following example shows, an essentially active selection function is not necessarily also conically active.

Example A.1.14. *Consider the $\mathcal{PC}^1$ function $f : \mathbb{R}^2 \to \mathbb{R}$ defined as follows:*

$$f(x,y) = \begin{cases} y - x^2 & \text{if } y > x^2, \\ y + x^2 & \text{if } y < -x^2, \\ 0 & \text{otherwise} \end{cases} \qquad \forall (x,y) \in \mathbb{R}^2.$$

Then the function $g : \mathbb{R}^2 \to \mathbb{R} : (x,y) \mapsto 0$ is an essentially active function for f at $\mathbf{0}$, since

$$\mathbf{0} \in \operatorname{cl}\Big(\operatorname{int}(\{(x,y) \in \mathbb{R}^2 : f(x,y) = 0\})\Big) = \{(x,y) \in \mathbb{R}^2 : |y| \leq x^2\}.$$

However, $\{\mathbf{d} \in \mathbb{R}^2 : f'(\mathbf{0};\mathbf{d}) = g'(\mathbf{0};\mathbf{d})\} = \{\mathbf{d} \in \mathbb{R}^2 : d_2 = 0\}$, which has an empty interior. Since all of the polyhedral cones in any conical subdivision have nonempty interior, it follows that g is not a conically active selection function for f at $\mathbf{0}$, regardless of the particular active conical subdivision employed.

A.2 PC^1-factorable functions

This section introduces the broad subclass of $\mathcal{PC}^1$ functions to which the methods developed in this work can be applied. A generalization of a result in [32] is developed to show that the forward mode of automatic differentiation produces directional derivatives for this class of functions.

A.2.1 Elemental PC^1 functions

As formalized in the following definition, the class of elemental $\mathcal{PC}^1$ functions is intuitively the class of simple, known $\mathcal{PC}^1$ functions, and includes abs, min, and max. The methods developed in this work apply to finite compositions of these elemental $\mathcal{PC}^1$ functions.

Definition A.2.1. *Given an open set $X \subset \mathbb{R}^n$, a $\mathcal{PC}^1$ function $\mathbf{f} : X \to \mathbb{R}^m$ is an* elemental $\mathcal{PC}^1$ function *if the following information is known:*

- *analytical directional derivatives for $\mathbf{f}$,*
- *an active normal set $H_{\mathbf{f}}(\mathbf{x})$ for $\mathbf{f}$ at each $\mathbf{x} \in X$, with its elements enumerated arbitrarily as $\{\mathbf{a}_{\mathbf{f}}^{(r)}(\mathbf{x})\}_{r=1}^{|H_{\mathbf{f}}(\mathbf{x})|}$, and*
- *a Boolean function $\zeta_{\mathbf{f}} : X \to \{\texttt{true}, \texttt{false}\}$, for which $\zeta_{\mathbf{f}}(\mathbf{x}) = \texttt{false}$ if and only if $H_{\mathbf{f}}(\mathbf{x})$ contains an element other than the zero vector.*

The remainder of this subsection presents examples of elemental $\mathcal{PC}^1$ functions. Further examples are given in Examples A.2.9 and A.2.10. Though each active normal set $H_{\mathbf{f}}(\mathbf{x})$ could be constructed as in the proof of Lemma A.1.1, it will instead be advantageous to choose each $H_{\mathbf{f}}(\mathbf{x})$ to contain as few elements as possible while remaining easy to compute. Note that knowledge of $H_{\mathbf{f}}(\mathbf{x})$ for each $\mathbf{x} \in X$ is sufficient to define $\zeta_{\mathbf{f}}$.

Remark A.2.2. *If a $\mathcal{PC}^1$ function $\mathbf{f} : X \to \mathbb{R}^m$ is $\mathcal{C}^1$ at some $\mathbf{x} \in X$, and if $\mathbf{Jf}(\mathbf{x})$ is known, then the directional derivatives of $\mathbf{f}$ at $\mathbf{x}$ can be computed using* (A.1)*. Moreover, $\{\mathbf{0}\}$ is trivially an active normal set for $\mathbf{f}$ at $\mathbf{x}$.*

Remark A.2.3. *If a $\mathcal{PC}^1$ function* $\mathbf{f}$ *on $X \subset \mathbb{R}$ is not $\mathcal{C}^1$ at $x \in X$, then $\{1\}$ is an active normal set for* $\mathbf{f}$ *at x. This is because polyhedral cones are closed under multiplication by nonnegative scalars, and so any polyhedral cone in $\mathbb{R}$ must equal either $\{0\}$, $\mathbb{R}$, $\{d \in \mathbb{R} : d \leq 0\}$, or $\{d \in \mathbb{R} : -d \leq 0\}$.*

In light of the above remarks, all $\mathcal{C}^1$ functions with known Jacobians are elemental $\mathcal{PC}^1$ functions. The following examples show that abs, min, and max are nonsmooth elemental $\mathcal{PC}^1$ functions. Further examples of elemental $\mathcal{PC}^1$ functions are given in Section A.2.2.

Example A.2.4. *The absolute value function* abs *is an elemental $\mathcal{PC}^1$ function, since its directional derivatives are given in Example A.1.4, and since for each $x \in \mathbb{R}$, the set*

$$H_{\text{abs}}(x) = \begin{cases} \{1\} & \text{if } x = 0, \\ \{0\} & \text{otherwise} \end{cases}$$

is an active normal set for abs *at x.*

Example A.2.5. *The* max *and* min *functions on $\mathbb{R}^n$ are elemental $\mathcal{PC}^1$ functions, as the following argument demonstrates. It follows from the definition of the directional derivative that for any $\mathbf{x}, \mathbf{d} \in \mathbb{R}^n$, if $f(\mathbf{x}) := \max(x_1, \ldots, x_n)$, then*

$$f'(\mathbf{x}; \mathbf{d}) = \max\{d_i : x_i = \max(x_1, \ldots, x_n)\}.$$

Directional derivatives for min *are analogous.*

Noting that the max *function on $\mathbb{R}^2$ is locally linear except on the hyperplane $S := \{\mathbf{x} \in \mathbb{R}^2 : x_1 = x_2\}$, which is perpendicular to $(1, -1)$, it is readily verified that for any given $\mathbf{x} \in \mathbb{R}^2$,*

$$H_{\max}(\mathbf{x}) = \begin{cases} \{(1, -1)\} & \text{if } x_1 = x_2, \\ \{\mathbf{0}\} & \text{otherwise} \end{cases}$$

is an active normal set for max *at* $\mathbf{x}$.

By a similar argument, for any $n \geq 2$, an active normal set for $\max : \mathbb{R}^n \to \mathbb{R}$ *at any given $\mathbf{x} \in \mathbb{R}^n$ is*

$$H_{\max}(\mathbf{x}) = \{\mathbf{e}_{(i)} - \mathbf{e}_{(j)} : (x_i = x_j = \max(x_1, \ldots, x_n)) \wedge (i < j)\} \cup \{\mathbf{0}\}.$$

An active normal set for the min *function on* $\mathbb{R}^n$ *at* $\mathbf{x}$ *is analogous.*

A.2.2 Composing elemental PC^1 functions

The class of $\mathcal{PC}^1$-factorable functions is defined as follows, and is analogous to the class of composite smooth functions to which automatic differentiation is conventionally applied [34]. Intuitively, $\mathcal{PC}^1$-factorable functions include all well-defined finite compositions of elemental $\mathcal{PC}^1$ functions.

Definition A.2.6. *Given an open set* $X \subset \mathbb{R}^n$*, a* $\mathcal{PC}^1$-factorable function $\mathbf{f} : X \to \mathbb{R}^m$ *is a function for which the following exist and are known:*

- *an* intermediate function number $\ell \in \mathbb{N}$,
- *a Boolean* dependence operator $\prec$*, such that* $(i \prec j) \in \{\texttt{true}, \texttt{false}\}$ *for each* $j \in \{1, 2, \ldots, \ell\}$ *and each* $i \in \{0, 1, \ldots, j-1\}$*, and*
- *an elemental* $\mathcal{PC}^1$ *function* $\psi_{(j)} : X_{(j)} \subset \mathbb{R}^{n_j} \to Y_{(j)} \subset \mathbb{R}^{m_j}$ *for each* $j \in \{1, \ldots, \ell\}$*, where* $\prod_{\{i : i \prec j\}} Y_{(i)} \subset X_{(j)}$*, and where* $m_\ell = m$,

and where for any $\mathbf{x} \in X$, $\mathbf{f}(\mathbf{x})$ *can be evaluated by the following procedure:*

> *Set* $\mathbf{v}_{(0)} \leftarrow \mathbf{x}$
> ***for*** $j = 1$ ***to*** ℓ ***do***
> *Set* $\mathbf{u}_{(j)} \in X_{(j)}$ *to be a column vector consisting of all* $\mathbf{v}_{(i)}$*s for which* $i \prec j$*, stacked in order of increasing* i*.*
> *Set* $\mathbf{v}_{(j)} \leftarrow \psi_{(j)}(\mathbf{u}_{(j)})$
> ***end for***
> *Set* $\mathbf{f}(\mathbf{x}) \leftarrow \mathbf{v}_{(\ell)}$

The above procedure defines $\mathbf{f}$ *completely, and is called a* $\mathcal{PC}^1$-factored representation *of* $\mathbf{f}$.

The functions considered in Examples A.5.1 to A.5.5 are all $\mathcal{PC}^1$-factorable functions. $\mathcal{PC}^1$-factored representations are constructed in Examples A.5.2 and A.5.3.

Given a $\mathcal{PC}^1$-factored representation of a function, if matrices (or vectors) $\mathbf{A}_{(i)}$ are defined for each $i \in \{0, \ldots, \ell\}$ so that each has the same number of columns, then $[\mathbf{A}_{(i)}]_{i \prec j}$ denotes the matrix (or vector) constructed by stacking the elements of $\{\mathbf{A}_{(i)} : i \prec j\}$ vertically in order of increasing i.

Remark A.2.7. *The class of $\mathcal{PC}^1$-factorable functions is evidently closed under composition. Moreover, Property 3 in Lemma A.1.8 implies that each $\mathcal{PC}^1$-factorable function is itself $\mathcal{PC}^1$.*

In practice, the elemental $\mathcal{PC}^1$ functions employed would be chosen from an implementation-dependent *library*. Such a library would typically contain the standard elemental $\mathcal{C}^1$ functions used in automatic differentiation, the abs function, and the max and min functions on $\mathbb{R}^2$. Nevertheless, depending on the particular application, it may be convenient to add further elemental $\mathcal{PC}^1$ functions to the library. In principle, this could include any $\mathcal{PC}^1$ function whose directional derivatives and active normal sets have been computed.

The following example demonstrates that many $\mathcal{PC}^1$ functions described by `if..then..else` statements can be represented as $\mathcal{PC}^1$-factorable functions without adding further elemental $\mathcal{PC}^1$ functions to the library. When this approach is inconvenient or impossible, the subsequent example demonstrates how to construct active normal sets for a broad subclass of `if..then..else`-type functions.

Example A.2.8. *Given $\mathcal{PC}^1$-factorable functions $\mathbf{f}_A, \mathbf{f}_B : X \subset \mathbb{R}^n \to \mathbb{R}^m$ and $g : X \to \mathbb{R}$, suppose that $\mathbf{f}_A(\mathbf{x}) = \mathbf{f}_B(\mathbf{x})$ whenever $g(\mathbf{x}) = 0$. Then the function*

$$\mathbf{f} : X \to \mathbb{R}^m : \mathbf{x} \mapsto \begin{cases} \mathbf{f}_A(\mathbf{x}) & \text{if } g(\mathbf{x}) \geq 0, \\ \mathbf{f}_B(\mathbf{x}) & \text{if } g(\mathbf{x}) < 0 \end{cases}$$

is $\mathcal{PC}^1$. Moreover, if there exists a $\mathcal{PC}^1$-factorable function $\mathbf{h} : X \to \mathbb{R}^m$ such that

$$g(\mathbf{x})\mathbf{h}(\mathbf{x}) = \mathbf{f}_A(\mathbf{x}) - \mathbf{f}_B(\mathbf{x}), \quad \forall \mathbf{x} \in X, \tag{A.6}$$

then

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}_B(\mathbf{x}) + \max(g(\mathbf{x}), 0)\, \mathbf{h}(\mathbf{x}), \quad \forall \mathbf{x} \in X,$$

and so $\mathbf{f}$ is $\mathcal{PC}^1$-factorable.

Example A.2.9. *Given elemental $\mathcal{PC}^1$ functions $\mathbf{f}_A, \mathbf{f}_B : X \subset \mathbb{R}^n \to \mathbb{R}^m$, suppose that for some $\mathbf{a} \in \mathbb{R}^n$ and some $c \in \mathbb{R}$, $\mathbf{f}_A(\mathbf{x}) = \mathbf{f}_B(\mathbf{x})$ whenever $\langle \mathbf{a}, \mathbf{x} \rangle = c$. Then the function*

$$\mathbf{f} : X \to \mathbb{R}^m : \mathbf{x} \mapsto \begin{cases} \mathbf{f}_A(\mathbf{x}) & \textit{if } \langle \mathbf{a}, \mathbf{x} \rangle \geq c, \\ \mathbf{f}_B(\mathbf{x}) & \textit{if } \langle \mathbf{a}, \mathbf{x} \rangle < c \end{cases}$$

is $\mathcal{PC}^1$. Using the definition of the directional derivative, it is readily verified that for any $\mathbf{x} \in X$ and any $\mathbf{d} \in \mathbb{R}^n$,

$$\mathbf{f}'(\mathbf{x}; \mathbf{d}) = \begin{cases} \mathbf{f}_A{}'(\mathbf{x}; \mathbf{d}) & \textit{if } \langle \mathbf{a}, \mathbf{x} \rangle > c, \textit{ or if } \langle \mathbf{a}, \mathbf{x} \rangle = c \textit{ and } \langle \mathbf{a}, \mathbf{d} \rangle \geq 0, \\ \mathbf{f}_B{}'(\mathbf{x}; \mathbf{d}) & \textit{if } \langle \mathbf{a}, \mathbf{x} \rangle < c, \textit{ or if } \langle \mathbf{a}, \mathbf{x} \rangle = c \textit{ and } \langle \mathbf{a}, \mathbf{d} \rangle < 0. \end{cases} \tag{A.7}$$

An active normal set for $\mathbf{f}$ at any $\mathbf{x} \in X$ is then

$$H_{\mathbf{f}}(\mathbf{x}) = \begin{cases} H_{\mathbf{f}_A}(\mathbf{x}) & \textit{if } \langle \mathbf{a}, \mathbf{x} \rangle > c, \\ H_{\mathbf{f}_B}(\mathbf{x}) & \textit{if } \langle \mathbf{a}, \mathbf{x} \rangle < c, \\ H_{\mathbf{f}_A}(\mathbf{x}) \cup H_{\mathbf{f}_B}(\mathbf{x}) \cup \{\mathbf{a}\} & \textit{if } \langle \mathbf{a}, \mathbf{x} \rangle = c. \end{cases}$$

In the following example, the approach of Example A.2.9 is used to construct active normal sets for a particular $\mathcal{PC}^1$ function.

Example A.2.10. *For any fixed $\epsilon \in (0, \frac{\pi}{2})$, consider the sets $S := (-\epsilon, 2\pi + \epsilon) \subset \mathbb{R}$ and $P := \{(x, y) \in S^2 : x < y\}$. Then P is open. Consider the function:*

$$f : P \to \mathbb{R} : (x, y) \mapsto \sup_{z \in [x,y]} \sin z,$$

which can be expressed as follows for any $(x, y) \in P$:

$$f(x, y) = \begin{cases} 1 & \textit{if } y \geq \frac{\pi}{2} \textit{ and } x < \frac{\pi}{2}, \\ \sin x & \textit{if } x \geq \frac{\pi}{2} \textit{ and } x + y < \frac{3\pi}{2}, \\ \sin y & \textit{otherwise.} \end{cases}$$

This representation demonstrates that f is $\mathcal{PC}^1$. There is no $(x, y) \in P$ which satisfies more than one of the statements $x = \frac{\pi}{2}$, $y = \frac{\pi}{2}$, and $x + y = \frac{3\pi}{2}$ simultaneously. It follows that since f is only nondifferentiable on the three hyperplanes described by these statements, f can be represented as three nested `if..then..else`*–type functions of the form described in Example A.2.9.*

Alternatively, in the spirit of Example A.2.9, f can be represented as an elemental $\mathcal{PC}^1$ function directly. Directional derivatives for f can be evaluated analogously to (A.7)*, and an active normal set for f at any $(x, y) \in P$ is as follows:*

$$H_f(x,y) = \begin{cases} \{\mathbf{e}_{(2)}\} & \text{if } y = \frac{\pi}{2}, \\ \{\mathbf{e}_{(1)}\} & \text{if } x = \frac{\pi}{2}, \\ \{(1,1)\} & \text{if } x + y = \frac{3\pi}{2}, \\ \{\mathbf{0}\} & \text{otherwise.} \end{cases}$$

A.2.3 Automatic differentiation

The forward mode of automatic differentiation (AD) is a fully automatable technique for computing directional derivatives of composite $\mathcal{C}^1$ functions [34]. The following definition of the forward mode of AD for $\mathcal{PC}^1$-factorable functions is analogous to the smooth case.

Definition A.2.11. *Given an open set $X \subset \mathbb{R}^n$, a $\mathcal{PC}^1$-factorable function $\mathbf{f}$ as described in Definition A.2.6, some $\mathbf{x} \in X$ and a direction vector $\mathbf{d} \in \mathbb{R}^n$, the* forward mode of AD for $\mathcal{PC}^1$-factorable functions *generates a vector $\dot{\mathbf{f}}(\mathbf{x};\mathbf{d}) \in \mathbb{R}^m$ according to the following procedure:*

Set $\dot{\mathbf{v}}_{(0)} \leftarrow \mathbf{d}$
for** $j = 1$ **to** ℓ **do
  Set $\dot{\mathbf{u}}_{(j)} \leftarrow [\dot{\mathbf{v}}_{(i)}]_{i \prec j}$, and set $\dot{\mathbf{v}}_{(j)} \leftarrow \psi_{(j)}{}'(\mathbf{u}_{(j)};\dot{\mathbf{u}}_{(j)})$
end for
Set $\dot{\mathbf{f}}(\mathbf{x};\mathbf{d}) \leftarrow \dot{\mathbf{v}}_{(\ell)}$

Remark A.2.12. *Given a $\mathcal{PC}^1$-factorable function $\mathbf{f}$, the intermediate variables $\mathbf{v}_{(j)}$, $\mathbf{u}_{(j)}$, $\dot{\mathbf{v}}_{(j)}$, and $\dot{\mathbf{u}}_{(j)}$ described in Definitions A.2.6 and A.2.11 are uniquely specified for each $\mathbf{x} \in X$ and each $\mathbf{d} \in \mathbb{R}^n$. Hence, there exist mappings $\mathbf{v}_{(j)} : X \to Y_{(j)}$, $\mathbf{u}_{(j)} : X \to X_{(j)}$, $\dot{\mathbf{v}}_{(j)} : X \times \mathbb{R}^n \to \mathbb{R}^{n_j}$, and $\dot{\mathbf{u}}_{(j)} : X \times \mathbb{R}^n \to \mathbb{R}^{m_j}$ which produce the values of these intermediate variables for each $\mathbf{x} \in X$ and each $\mathbf{d} \in \mathbb{R}^n$.*

Generalizing a similar result in [32], the following lemma shows that the forward mode of AD produces directional derivatives for $\mathcal{PC}^1$-factorable functions.

Lemma A.2.13. *Given a $\mathcal{PC}^1$-factored representation of a $\mathcal{PC}^1$-factorable function $\mathbf{f} : X \to \mathbb{R}^n$, the vectors $\dot{\mathbf{v}}_{(j)}(\mathbf{x};\mathbf{d})$ and $\dot{\mathbf{u}}_{(j)}(\mathbf{x};\mathbf{d})$ generated by the forward mode of AD for each $j \in \{1,\ldots,\ell\}$ are, respectively, the directional derivatives $\mathbf{v}_{(j)}{}'(\mathbf{x};\mathbf{d})$ and $\mathbf{u}_{(j)}{}'(\mathbf{x};\mathbf{d})$. In particular, $\dot{\mathbf{f}}(\mathbf{x};\mathbf{d})$ is the directional derivative $\mathbf{f}'(\mathbf{x};\mathbf{d})$.*

Proof. Consider any fixed $\mathbf{x} \in X$ and $\mathbf{d} \in \mathbb{R}^n$. By Property 3 in Lemma A.1.8, the mappings $\mathbf{v}_{(j)}$ and $\mathbf{u}_{(j)}$ are $\mathcal{PC}^1$, and are therefore directionally differentiable. The lemma is proved by strong induction on $j \in \{0, 1, \ldots, \ell\}$ as follows.

Base case: Since $\mathbf{v}_{(0)}(\mathbf{y}) = \mathbf{y}$ for each $\mathbf{y} \in X$, it follows that $\mathbf{v}_{(0)}{}'(\mathbf{x}; \mathbf{d}) = \mathbf{d} = \dot{\mathbf{v}}_{(0)}(\mathbf{x}; \mathbf{d})$.

Strong inductive step: Suppose that for some $j \in \{0, 1, \ldots, \ell - 1\}$, for each $i \leq j$, $\mathbf{v}_{(i)}{}'(\mathbf{x}; \mathbf{d}) = \dot{\mathbf{v}}_{(i)}(\mathbf{x}; \mathbf{d})$. Since $\mathbf{u}_{(j+1)}(\mathbf{y}) = [\mathbf{v}_{(i)}(\mathbf{y})]_{i \prec j+1}$ for each $\mathbf{y} \in X$, the definition of the directional derivative can be applied to yield $\mathbf{u}_{(j+1)}{}'(\mathbf{x}; \mathbf{d}) = [\mathbf{v}_{(i)}{}'(\mathbf{x}; \mathbf{d})]_{i \prec j+1}$. The strong inductive assumption then implies that

$$\mathbf{u}_{(j+1)}{}'(\mathbf{x}; \mathbf{d}) = [\dot{\mathbf{v}}_{(i)}(\mathbf{x}; \mathbf{d})]_{i \prec j+1} = \dot{\mathbf{u}}_{(j+1)}(\mathbf{x}; \mathbf{d}).$$

Combining this result with (A.4) yields:

$$\begin{aligned}
\mathbf{v}_{(j+1)}{}'(\mathbf{x}; \mathbf{d}) &= \psi_{(j+1)}{}'(\mathbf{u}_{(j+1)}(\mathbf{x}); \mathbf{u}_{(j+1)}{}'(\mathbf{x}; \mathbf{d})) \\
&= \psi_{(j+1)}{}'(\mathbf{u}_{(j+1)}(\mathbf{x}); \dot{\mathbf{u}}_{(j+1)}(\mathbf{x}; \mathbf{d})) \\
&= \dot{\mathbf{v}}_{(j+1)}(\mathbf{x}; \mathbf{d}),
\end{aligned}$$

which completes the strong induction. Since $\mathbf{f} \equiv \mathbf{v}_{(\ell)}$ by construction, it follows that $\dot{\mathbf{f}}(\mathbf{x}; \mathbf{d}) = \mathbf{f}'(\mathbf{x}; \mathbf{d})$. □

A.3 Generalized Jacobian element evaluation

As discussed in [53], if any essentially active selection function in $\mathcal{E}_{\mathbf{f}}(\mathbf{x})$ is known *a priori* for a $\mathcal{PC}^1$-factorable function $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ at $\mathbf{x} \in X$, then an element of $\partial_{\mathrm{B}}\mathbf{f}(\mathbf{x}) \subset \partial\mathbf{f}(\mathbf{x})$ can be evaluated using Property 3 in Lemma A.1.10. However, if little is known about $\mathbf{f}$ *a priori* beyond a $\mathcal{PC}^1$-factored representation, then essentially active selection functions for $\mathbf{f}$ at $\mathbf{x}$ can be difficult to obtain. In particular, composing essentially active selection functions of the elemental $\mathcal{PC}^1$ functions describing $\mathbf{f}$ does not necessarily yield an essentially active selection function of $\mathbf{f}$ [53]. Similarly, active conical subdivisions for $\mathbf{f}$ are nontrivial to construct.

Nevertheless, the results in Section A.2.3 show that directional derivatives are computable for $\mathcal{PC}^1$-factorable functions, and Lemma A.1.10 shows that these directional derivatives provide information regarding generalized Jacobian elements. Therefore, in this section, numerical methods are developed to evaluate a generalized Jacobian element for a $\mathcal{PC}^1$-factorable function $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ at any $\mathbf{x} \in X$, by determining the Jacobian of a conically active selection function of $\mathbf{f}$ at $\mathbf{x}$ according to Lemma A.1.11. Section A.3.1 covers the special case in which $n = 1$, and Section A.3.2 covers the general case.

A.3.1 PC^1-factorable functions of a single variable

As the following theorem demonstrates, generalized Jacobian elements are readily obtained for $\mathcal{PC}^1$-factorable functions of a single variable using the forward AD mode.

Theorem A.3.1. *Given a $\mathcal{PC}^1$-factorable function $\mathbf{f} : X \subset \mathbb{R} \to \mathbb{R}^m$,*

$$\partial \mathbf{f}(x) = \operatorname{conv}\{\mathbf{f}'(x;1), -\mathbf{f}'(x;-1)\}, \qquad \forall x \in X. \tag{A.8}$$

In particular, a single application of the forward mode of AD to $\mathbf{f}$ is sufficient to evaluate an element of $\partial \mathbf{f}(x)$.

Proof. The result is an immediate consequence of Remark A.2.3 and Lemma A.1.11. □

In a sense, the above result depends on the fact that any neighborhood of $x \in \mathbb{R}$ only extends in finitely many directions away from x: the positive direction and the negative direction. This property clearly does not extend to higher-dimensional Euclidean spaces. Moreover, [16, Example 2.5.2] shows that when $n > 1$, obtaining a generalized Jacobian element for a $\mathcal{PC}^1$-factorable function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ is not a simple matter of concatenating elements of *partial generalized Jacobians* $\partial_i \mathbf{f}(\mathbf{x})$ obtained using Theorem A.3.1.

A.3.2 PC^1-factorable functions of multiple variables

In this section, it is shown that if Algorithm 12 is applied to a $\mathcal{PC}^1$-factorable function $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ at $\mathbf{x} \in X$, then an element of $\partial\mathbf{f}(\mathbf{x})$ is returned. The proof of this result depends on several intermediate results which are stated and proved in the appendix of this article. The performance of Algorithm 12 is discussed in Section A.4, along with methods for improving the efficiency of the algorithm.

An outline of Algorithm 12 is as follows. The algorithm evaluates directional derivatives of the given function $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ along the n unit coordinate directions using the forward AD mode. These basis directions are then perturbed until for each elemental $\mathcal{PC}^1$ function in the $\mathcal{PC}^1$-factored representation of $\mathbf{f}$, the directions in which the directional derivatives of these elemental $\mathcal{PC}^1$ functions are evaluated all lie in the same active cone of the elemental $\mathcal{PC}^1$ function. Each perturbation of the basis directions involves addition of a positive scalar multiple of one basis direction to another, so as to maintain the linear independence of the basis. The generalized Jacobian element returned by the algorithm is the Jacobian of the composition of the corresponding conically active selection functions of each elemental function. Though the algorithm contains three nested `for`–loops, if the only permitted elemental $\mathcal{PC}^1$ functions are $\mathcal{C}^1$ functions, abs, min on $\mathbb{R}^2$, and max on $\mathbb{R}^2$, then the constructions in Section A.2.1 imply that at most one iteration of the middle `for`–loop is carried out during each iteration of the outermost `for`–loop.

Theorem A.3.2. *Given an open set $X \subset \mathbb{R}^n$, a $\mathcal{PC}^1$-factorable function $\mathbf{f} : X \to \mathbb{R}^m$, and a vector $\mathbf{x} \in X$, suppose that a matrix $\mathbf{B} \in \mathbb{R}^{m\times n}$ is constructed according to Algorithm 12. $\mathbf{B}$ is then well-defined, and is an element of both $\partial_{\mathrm{B}}\mathbf{f}(\mathbf{x})$ and $\partial\mathbf{f}(\mathbf{x})$.*

Proof. By Lemma A.6.2, the matrix $\begin{bmatrix}\mathbf{q}_{(1)} & \cdots & \mathbf{q}_{(n)}\end{bmatrix}$ is invertible when Line 23 of the algorithm is reached. Consequently, the linear equation system in Line 23 has a unique solution, and so $\mathbf{B}$ is well-defined.

Since $\mathbf{f} \equiv \mathbf{v}_{(\ell)}$ on X, Lemma A.6.6 implies that when Line 23 is reached, there exists some $\mathbf{f}^* \in \mathcal{E}_{\mathbf{f}}(\mathbf{x})$ such that $\mathbf{f}'(\mathbf{x};\mathbf{q}_{(k)}) = \mathbf{J}\mathbf{f}^*(\mathbf{x})\,\mathbf{q}_{(k)}$ for each $k \in \{1,\ldots,n\}$. It follows that

Algorithm 12 Computes an element of $\partial \mathbf{f}(\mathbf{x})$ for a $\mathcal{PC}^1$-factorable function $\mathbf{f}$

Require: $\mathbf{f} : X \to \mathbb{R}^m$ is $\mathcal{PC}^1$-factorable, $\mathbf{x} \in X$

Set $\mathbf{q}_{(k)} \leftarrow \mathbf{e}_{(k)} \in \mathbb{R}^n$ for each $k \in \{1, \ldots, n\}$

Use the $\mathcal{PC}^1$-factored representation of $\mathbf{f}$ to evaluate the intermediate variable $\mathbf{v}_{(j)}(\mathbf{x})$ for each $j \in \{1, \ldots, \ell\}$

For each $k \in \{1, \ldots, n\}$, use the forward mode of AD to evaluate $\mathbf{f}'(\mathbf{x}; \mathbf{q}_{(k)})$, and set $\dot{\mathbf{u}}_{(j,k)} \leftarrow \dot{\mathbf{u}}_{(j)}(\mathbf{x}; \mathbf{q}_{(k)})$ for each $j \in \{1, \ldots, \ell\}$

for $j = 1$ **to** ℓ **do**

  if $\zeta_{\psi_{(j)}}\left(\mathbf{u}_{(j)}(\mathbf{x})\right) = \texttt{false}$ **then**

    for $r = 1$ **to** $|H_{\psi_{(j)}}\left(\mathbf{u}_{(j)}(\mathbf{x})\right)|$ **do**

      Set $c^* \leftarrow 0$

      for $k = 1$ **to** n **do**

        Set $c \leftarrow \langle \mathbf{a}^{(r)}_{\psi_{(j)}}(\mathbf{u}_{(j)}(\mathbf{x})), \dot{\mathbf{u}}_{(j,k)} \rangle \in \mathbb{R}$

        if $c \neq 0$ **then**

          if $c^* = 0$ **then**

            Set $c^* \leftarrow c$ and $k^* \leftarrow k$

          else if $cc^* < 0$ **then**

            Set $\alpha \leftarrow -\frac{c}{c^*} \in \mathbb{R}_+$

            Set $\mathbf{q}_{(k)} \leftarrow \mathbf{q}_{(k)} + \alpha \mathbf{q}_{(k^*)}$

            Use the forward mode of AD to evaluate $\mathbf{f}'(\mathbf{x}; \mathbf{q}_{(k)})$, and set $\dot{\mathbf{u}}_{(i,k)} \leftarrow \dot{\mathbf{u}}_{(i)}(\mathbf{x}; \mathbf{q}_{(k)})$ for each $i \in \{1, \ldots, \ell\}$

          end if

        end if

      end for

    end for

  end if

end for

Solve the following linear system for $\mathbf{B} \in \mathbb{R}^{m \times n}$:

$$\mathbf{B} \begin{bmatrix} \mathbf{q}_{(1)} & \cdots & \mathbf{q}_{(n)} \end{bmatrix} = \begin{bmatrix} \mathbf{f}'(\mathbf{x}; \mathbf{q}_{(1)}) & \cdots & \mathbf{f}'(\mathbf{x}; \mathbf{q}_{(n)}) \end{bmatrix}$$

return $\mathbf{B}$

$$\mathbf{J}\mathbf{f}^*(\mathbf{x}) \begin{bmatrix} \mathbf{q}_{(1)} & \cdots & \mathbf{q}_{(n)} \end{bmatrix} = \begin{bmatrix} \mathbf{f}'(\mathbf{x};\mathbf{q}_{(1)}) & \cdots & \mathbf{f}'(\mathbf{x};\mathbf{q}_{(n)}) \end{bmatrix}.$$

Since $\begin{bmatrix} \mathbf{q}_{(1)} & \cdots & \mathbf{q}_{(n)} \end{bmatrix}$ is invertible, $\mathbf{J}\mathbf{f}^*(\mathbf{x})$ is the matrix $\mathbf{B}$ that is returned by Algorithm 12. Property 3 in Lemma A.1.10 then yields $\mathbf{B} = \mathbf{J}\mathbf{f}^*(\mathbf{x}) \in \partial_{\mathrm{B}}\mathbf{f}(\mathbf{x}) \subset \partial\mathbf{f}(\mathbf{x})$. □

Corollary A.3.3. *Given a $\mathcal{PC}^1$-factorable function $\mathbf{f} : X \to \mathbb{R}^m$, Algorithm 12 produces $\mathbf{J}\mathbf{f}(\mathbf{x})$ if $\mathbf{f}$ is $\mathcal{C}^1$ at $\mathbf{x}$, and produces a subgradient of $\mathbf{f}$ at $\mathbf{x}$ if $\mathbf{f}$ is convex.*

Proof. If a $\mathcal{PC}^1$ function $\mathbf{f} : X \to \mathbb{R}^m$ is $\mathcal{C}^1$ at $\mathbf{x} \in X$, then $\partial\mathbf{f}(\mathbf{x}) = \{\mathbf{J}\mathbf{f}(\mathbf{x})\}$, which yields the first result. The subdifferential of a locally Lipschitz continuous and convex function is identical to its generalized Jacobian [16], which yields the second result. □

Note that for the results of Corollary A.3.3 to hold, the elemental $\mathcal{PC}^1$ functions used to represent $\mathbf{f}$ need not be $\mathcal{C}^1$ or satisfy rules for convex composite functions such as in [12, Sections 3.2.4 and 3.6.2].

The following corollary shows that when Algorithm 12 is applied to a function of a single variable, the algorithm produces a result that is consistent with Theorem A.3.1. In this case, it is nevertheless more computationally efficient to apply Theorem A.3.1 than to apply Algorithm 12.

Corollary A.3.4. *Given a $\mathcal{PC}^1$-factorable function $\mathbf{f} : X \subset \mathbb{R} \to \mathbb{R}^m$ and some $x \in X$, Algorithm 12 produces $\mathbf{f}'(x;1) \in \partial\mathbf{f}(x)$.*

Proof. During execution of the algorithm, $\mathbf{q}_{(1)}$ is never altered after its initial assignment to $\mathbf{e}_{(1)}$. Since $X \subset \mathbb{R}$ in this case, $\mathbf{e}_{(1)} = 1$, and so the final linear system in the algorithm reduces to $\mathbf{B} = \mathbf{f}'(x;1)$. The algorithm therefore returns $\mathbf{f}'(x;1)$, which is an element of $\partial\mathbf{f}(x)$ according to Theorem A.3.1. □

Theorem 2 in our previous work [53] follows as a corollary to Theorem A.3.2 above, since with $H_{\text{abs}}(x)$ given for each $x \in \mathbb{R}$ as in Example A.2.4, it follows that $\zeta_{\text{abs}}(x) = \texttt{false}$ if and only if $x = 0$, in which case $H_{\text{abs}}(x) = \{1\}$. Thus, the method in [53, Theorem 2] is the special case of Algorithm 12 in which each elemental $\mathcal{PC}^1$ function $\psi_{(j)}$ is chosen from the class $\mathcal{C}^1 \cup \{\text{abs}\}$.

A.3.3 Modifications to Algorithm 12

Algorithm 13 is a variant of Algorithm 12 which is intended to improve computational performance. As discussed in Section A.4, if nonsmooth elemental $\mathcal{PC}^1$ functions other than abs, min, and max are employed, then Algorithm 13 demands fewer applications of the forward AD mode than Algorithm 12 in the worst case.

The following theorem demonstrates that, given the same input, Algorithms 12 and 13 produce the same result. The proof of this theorem describes the motivation behind the modifications to Algorithm 12 which produce Algorithm 13. The performance of Algorithm 13 is discussed in Section A.4.

Theorem A.3.5. *Given a $\mathcal{PC}^1$-factorable function $\mathbf{f} : X \to \mathbb{R}^m$ and a vector $\mathbf{x} \in X$, Algorithms 12 and 13 produce the same matrix $\mathbf{B} \in \partial_{\mathrm{B}}\mathbf{f}(\mathbf{x}) \subset \partial\mathbf{f}(\mathbf{x})$ when applied to $\mathbf{f}$ at $\mathbf{x}$.*

Proof. Algorithm 13 is a modified version of Algorithm 12, and so it suffices to show that each modification made to Algorithm 12 does not alter the values of the basis vectors $\mathbf{q}_{(k)}$ at the end of each iteration of the outermost `for`–loop.

Firstly, for each $j \in \{0, 1, \ldots, \ell - 1\}$, once the jth iteration of the outermost `for`–loop in Algorithm 12 has been performed, the values of $\dot{\mathbf{u}}_{(i,1)}, \ldots, \dot{\mathbf{u}}_{(i,n)}$ are never used in the remainder of the algorithm unless $i > j$ and $\zeta_{\psi_{(i)}}\left(\mathbf{u}_{(i)}(\mathbf{x})\right) = \texttt{false}$. Except in this case, there is no need to store each $\dot{\mathbf{u}}_{(i,k)}$.

Secondly, the values of $\dot{\mathbf{u}}_{(i,k)}$ are not used during the jth iteration of the outermost `for`–loop if $i \neq j$. Moreover, setting $j \leftarrow j^*$ in (A.15) shows that Line 16 of Algorithm 13 produces the same change in $\dot{\mathbf{u}}_{(j,k)}$ as Line 16 of Algorithm 12. Hence, rather than carrying out the forward AD mode in the innermost `for`–loop of Algorithm 12, it suffices to update $\dot{\mathbf{u}}_{(j,k)}$ using (A.15) in the innermost loop, and only carry out the forward AD mode at the end of each iteration of the outermost `for`–loop.

Lastly, it follows from the above discussion that when the ℓth iteration of the outermost `for`-loop is reached, the remainder of the algorithm does not make use of $\dot{\mathbf{u}}_{(i,k)}$ whenever $i \neq \ell$. Since (A.4) implies that

Algorithm 13 Computes an element of $\partial \mathbf{f}(\mathbf{x})$ for a $\mathcal{PC}^1$-factorable function $\mathbf{f}$

Require: $\mathbf{f} : X \to \mathbb{R}^m$ is $\mathcal{PC}^1$-factorable, $\mathbf{x} \in X$

Set $\mathbf{q}_{(k)} \leftarrow \mathbf{e}_{(k)} \in \mathbb{R}^n$ for each $k \in \{1, \ldots, n\}$

Use the $\mathcal{PC}^1$-factored representation of $\mathbf{f}$ to evaluate the intermediate variable $\mathbf{v}_{(j)}(\mathbf{x})$ for each $j \in \{1, \ldots, \ell\}$

For each $k \in \{1, \ldots, n\}$, use the forward mode of AD to evaluate $\mathbf{f}'(\mathbf{x}; \mathbf{q}_{(k)})$, and set $\dot{\mathbf{u}}_{(j,k)} \leftarrow \dot{\mathbf{u}}_{(j)}(\mathbf{x}; \mathbf{q}_{(k)})$ for each $j \in \{1, \ldots, \ell\}$ such that $\zeta_{\psi_{(j)}}\left(\mathbf{u}_{(j)}(\mathbf{x})\right) = \texttt{false}$

for $j = 1$ **to** ℓ **do**

  if $\zeta_{\psi_{(j)}}\left(\mathbf{u}_{(j)}(\mathbf{x})\right) = \texttt{false}$ **then**

    Set a Boolean variable $\texttt{qUpdated}(k) \leftarrow \texttt{false}$ for each $k \in \{1, \ldots, n\}$

    for $r = 1$ **to** $|H_{\psi_{(j)}}\left(\mathbf{u}_{(j)}(\mathbf{x})\right)|$ **do**

      Set $c^* \leftarrow 0$

      for $k = 1$ **to** n **do**

        Set $c \leftarrow \langle \mathbf{a}^{(r)}_{\psi_{(j)}}(\mathbf{u}_{(j)}(\mathbf{x})), \dot{\mathbf{u}}_{(j,k)} \rangle \in \mathbb{R}$

        if $c \neq 0$ **then**

          if $c^* = 0$ **then**

            Set $c^* \leftarrow c$ and $k^* \leftarrow k$

          else if $cc^* < 0$ **then**

            Set $\alpha \leftarrow -\frac{c}{c^*} \in \mathbb{R}_+$

            Set $\mathbf{q}_{(k)} \leftarrow \mathbf{q}_{(k)} + \alpha \mathbf{q}_{(k^*)}$ and $\dot{\mathbf{u}}_{(j,k)} \leftarrow \dot{\mathbf{u}}_{(j,k)} + \alpha \dot{\mathbf{u}}_{(j,k^*)}$

            Set $\texttt{qUpdated}(k) \leftarrow \texttt{true}$

          end if

        end if

      end for

    end for

    if $j = \ell$ **then**

      Evaluate $\mathbf{f}'(\mathbf{x}; \mathbf{q}_{(k)}) = \psi_{(\ell)}{}'(\mathbf{u}_{(\ell)}(\mathbf{x}); \dot{\mathbf{u}}_{(\ell,k)})$ for each $k \in \{1, \ldots, n\}$ such that $\texttt{qUpdated}(k) = \texttt{true}$

    else

      for all $k \in \{1, \ldots, n\}$ such that $\texttt{qUpdated}(k) = \texttt{true}$ **do**

        Use the forward mode of AD to evaluate $\mathbf{f}'(\mathbf{x}; \mathbf{q}_{(k)})$, and set $\dot{\mathbf{u}}_{(i,k)} \leftarrow \dot{\mathbf{u}}_{(i)}(\mathbf{x}; \mathbf{q}_{(k)})$ for each $i \in \{j+1, \ldots, \ell\}$ such that $\zeta_{\psi_{(i)}}\left(\mathbf{u}_{(i)}(\mathbf{x})\right) = \texttt{false}$

      end for

    end if

  end if

end for

Solve the following linear system for $\mathbf{B} \in \mathbb{R}^{m \times n}$:

$$\mathbf{B}\begin{bmatrix}\mathbf{q}_{(1)} & \cdots & \mathbf{q}_{(n)}\end{bmatrix} = \begin{bmatrix}\mathbf{f}'(\mathbf{x}; \mathbf{q}_{(1)}) & \cdots & \mathbf{f}'(\mathbf{x}; \mathbf{q}_{(n)})\end{bmatrix}$$

return $\mathbf{B}$

$$\mathbf{f}'(\mathbf{x};\mathbf{q}_{(k)}) = \psi_{(\ell)}{}'(\mathbf{u}_{(\ell)}(\mathbf{x});\dot{\mathbf{u}}_{(\ell,k)})$$

as well, there is no need to carry out the forward AD mode once the ℓth iteration of the outermost loop has been reached. $\square$

A.4 Computational performance

In this section, a worst-case complexity analysis is applied to Algorithms 12 and 13, to demonstrate the computational tractability of the algorithms relative to the cost of a function evaluation. Since semismooth Newton methods for solving nonsmooth equations demand evaluation of a generalized Jacobian element during each iteration [23], computationally tractable generalized Jacobian element evaluation is necessary for these methods to be practical.

Useful potential variations of these algorithms are also discussed.

A.4.1 Complexity analysis

Given a $\mathcal{PC}^1$-factorable function $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ and some $\mathbf{x} \in X$, the computational cost of applying Algorithms 13 and 12 to $\mathbf{f}$ at $\mathbf{x}$ is evidently dominated by the costs of applying the forward mode of AD, solving the final linear system in each algorithm, computing inner products, and carrying out AXPY ("Alpha X Plus Y") operations of the form $\mathbf{y} \leftarrow \mathbf{y} + \alpha\mathbf{x}$ with $\alpha \in \mathbb{R}$.

To this end, a worst-case complexity analysis will be conducted for these algorithms by describing the worst-case computational effort involved in each of the above operations in terms of the following parameters:

- the domain dimension n and the range dimension m of $\mathbf{f}$,
- $\rho := \max\left\{\left|H_{\psi_{(j)}}\left(\mathbf{u}_{(j)}(\mathbf{x})\right)\right| : j \in \{1,\ldots,\ell\}\right\}$,
- $\gamma := \left|\left\{j \in \{1,\ldots,\ell\} : \psi_{(j)} \text{ is not } \mathcal{C}^1 \text{ at } \mathbf{u}_{(j)}(\mathbf{x})\right\}\right|$, and

- the maximum domain dimension ν of the nonsmooth elemental $\mathcal{PC}^1$ functions $\boldsymbol{\psi}_{(j)}$.

Note that γ is no greater than $\left|\{j \in \{1, \ldots, \ell\} : \boldsymbol{\psi}_{(j)} \text{ is nonsmooth}\}\right|$, which is computed easily from a $\mathcal{PC}^1$-factored representation of $\mathbf{f}$, and which is in turn no greater than ℓ.

The employed library $\mathcal{L}$ of elemental $\mathcal{PC}^1$ functions places an upper bound on ρ, with

$$\rho \leq \sup\{|H_{\boldsymbol{\psi}^*}(\mathbf{y})| : \boldsymbol{\psi}^* \in \mathcal{L},\ \mathbf{y} \in \operatorname{dom} \boldsymbol{\psi}^*\}.$$

In particular, if $\mathcal{L}$ consists only of $\mathcal{C}^1$ functions, max and min on $\mathbb{R}^2$, and abs, then Examples A.2.4 and A.2.5 imply that $\rho \leq 1$. The employed library similarly places an upper bound on ν; when min and max on $\mathbb{R}^2$ and abs are the only nonsmooth elemental $\mathcal{PC}^1$ functions permitted, then $\nu = 2$.

Let $\mathbf{Q}_f$ denote the coefficient matrix of the final linear systems in Algorithms 13 and 12. Lemma A.6.2 shows that $\mathbf{Q}_f$ is guaranteed to be unit upper triangular. Now, $\mathbf{Q}_f$ is constructed from the identity matrix in each algorithm by applying operations in which a scalar multiple of the $(k^*)^{\text{th}}$ column is added to another column successively. By inspection, k^* can take no more than $\gamma\rho$ distinct values in either algorithm. It follows that no more than $\gamma\rho$ of the rows of $\mathbf{Q}_f$ can contain nonzero off-diagonal elements, and so since $\mathbf{Q}_f$ is unit upper triangular, the final linear system in each algorithm requires only $\mathcal{O}(nm\rho\gamma)$ FLOPs to solve.

With this observation, Algorithm 12 requires no more than:

- $n + (n-1)\rho\gamma$ applications of the forward AD mode to $\mathbf{f}$,
- $\mathcal{O}(nm\rho\gamma)$ FLOPs to solve the final linear system,
- $n\rho\gamma$ inner products of vectors of dimension no greater than ν, and
- $(n-1)\rho\gamma$ AXPY operations on vectors of dimension n.

Similarly, Algorithm 13 requires no more than:

- $(n\gamma + n - \gamma)$ applications of the forward AD mode to $\mathbf{f}$,
- $\mathcal{O}(nm\rho\gamma)$ FLOPs to solve the final linear system,
- $n\rho\gamma$ inner products of vectors of dimension no greater than ν, and
- $(n-1)\rho\gamma$ AXPY operations on vectors of dimension no greater than $(n+\nu)$.

The computational cost of applying the forward mode of AD is typically a small constant multiple of the cost of a function evaluation [34], where the value of the constant depends on the library of elemental functions employed. As a result, Algorithms 12 and 13 are both computationally tractable relative to the cost of a function evaluation.

Remark A.4.1. *Since it is in some sense improbable for an inner product of two vectors to be zero, it is expected that throughout most executions of either algorithm, k^* will only take the value 1 whenever it is assigned a value. This would simplify the structure of $\mathbf{Q}_f$, but is not considered in the worst-case analyses above.*

The worst-case number of forward AD mode applications in Algorithm 13 is evidently independent of ρ. Thus, if ρ is significantly greater than unity, if $n > 1$, and if $\gamma > 0$, then Algorithm 13 demands fewer applications of the forward AD mode than Algorithm 12 in the worst case. As discussed above, however, $\rho > 1$ only if nonsmooth elemental $\mathcal{PC}^1$ functions other than abs, min on $\mathbb{R}^2$, and max on $\mathbb{R}^2$ are employed.

In the special case where $\zeta_{\psi_{(j)}}\left(\mathbf{u}_{(j)}(\mathbf{x})\right) = \texttt{true}$ for each $j \in \{1, 2, \ldots, \ell\}$, Algorithms 12 and 13 each require n applications of the forward AD mode to $\mathbf{f}$, no evaluation of inner products, and no AXPY operations. Since the basis vectors $\mathbf{q}_{(k)}$ are never altered in this case, the coefficient matrix in the final linear system in each algorithm is the identity matrix, and so the linear system does not require any operations to solve. The computational cost of either algorithm is therefore comparable to the cost of evaluating the Jacobian of an analogous smooth function using the forward mode of AD.

A.4.2 Further potential modifications

In this subsection, further potential modifications to the algorithms are discussed that would either change the generalized Jacobian element produced, or would improve the computational performance of the algorithms.

Since $\mathbf{q}_{(1)}$ is not altered during either algorithm, these methods are guaranteed to produce the Jacobian of a conically active selection function for $\mathbf{f}$ at $\mathbf{x}$ whose active cone contains $\mathbf{e}_{(1)}$. In fact, the methods can be altered to produce the Jacobian of some conically active selection function whose active cone overlaps with any given set cone $\mathbf{s}_{(1)}, \ldots, \mathbf{s}_{(n)} \subset \mathbb{R}^n$ with nonempty interior. This can be accomplished by setting $\mathbf{q}_{(k)} \leftarrow \mathbf{s}_{(k)}$ initially instead of setting $\mathbf{q}_{(k)} \leftarrow \mathbf{e}_{(k)}$. If the matrix $\begin{bmatrix} \mathbf{s}_{(1)} & \cdots & \mathbf{s}_{(n)} \end{bmatrix}$ is not (unit) upper triangular, however, then the coefficient matrix in the final linear system will likely not be (unit) upper triangular either.

It is evident from the statements of the algorithms that not all of the evaluated quantities are actually used, and that some are not used after certain points in the procedures. If available memory is limited, or if n or ℓ is particularly large, then overwriting could be used to reduce the memory footprint of the algorithms. Moreover, at the cost of introducing additional overhead, unused intermediate quantities need not be evaluated in the first place. Rather than carrying out the full forward AD mode, it may be advantageous to carry out each step of the forward mode only when the corresponding intermediate variable is required. To further reduce the work involved in carrying out the forward AD mode, Lemma A.6.6 implies that during the $(j^*)^{\text{th}}$ iteration of the outermost `for`–loop in either algorithm, for each $j < j^*$, the intermediate values $\dot{\mathbf{v}}_{(j)}(\mathbf{x}; \mathbf{q}_{(k)})$ can be updated using AXPY operations analogous to (A.15) instead of using the forward AD mode explicitly. If this idea is applied to Algorithm 13, then these AXPY updates should be carried out whenever $\mathbf{q}_{(k)}$ is updated, rather than when the forward AD mode is carried out.

As described in [34, Section 4.5], the *vector forward mode* of AD effectively carries out the forward AD mode in several directions simultaneously, with less compu-

tational burden than sequentially carrying out the forward AD mode in each direction. A similar approach could reduce the computational cost of Algorithm 13, since this algorithm typically demands evaluation of directional derivatives in several directions in one step.

Lastly, given a $\mathcal{PC}^1$-factorable function $\mathbf{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$, when n and/or m are large, but each output $f_1, \ldots, f_m$ depends on few of the input variables $x_1, \ldots, x_n$, then this sparsity of the computational graph of $\mathbf{f}$ can be exploited in order to reduce the computational cost of performing AD [34, Chapter 7]. This option is not explored in the present work, but would likely be useful for large-scale practical problems.

A.5 Implementation and examples

In this section, an implementation of Algorithm 12 is discussed, and the methods developed in this paper are applied to several examples for illustration.

A.5.1 Implementation in C++

An implementation of Algorithm 12 was developed in C++. The implementation requires a $\mathcal{PC}^1$-factorable function to be entered as a template function which is written as though all input, output, and intermediate variables are of double precision type. `If`–statements and `while`–loops are not permitted; `for`–loops are permitted only if the number of iterations performed is independent of the values of the input variables. The standard arithmetic and trigonometric functions are permitted, along with abs, min, and max. Branching functions of the form of Example A.2.9 are supported in the special case where $n = 1$ and $\mathbf{a} = 1$. Further elemental $\mathcal{PC}^1$ functions can be added to the library, provided that their directional derivatives and active normal sets can be computed.

Given a point $\mathbf{x}$ in the function's domain, fully-automated computation of a generalized Jacobian element of the function at $\mathbf{x}$ proceeds as follows. Operator

overloading is used to construct a $\mathcal{PC}^1$-factored representation of the function, which is stored in the form of several arrays, in the spirit of the reverse AD mode implementation discussed in [34, Chapter 6]. Using the notation of Sections A.2.2 and A.2.3, for each $j \in \{1, 2, \ldots, \ell\}$, the identities of the elemental $\mathcal{PC}^1$ functions $\psi_{(j)}$ are stored in an *operation trace* array, and the indices $\{i : i \prec j\}$ are stored in an *index trace* array. The values $\mathbf{v}_{(j)}(\mathbf{x})$ are computed using the operation and index traces, and are stored in a *value trace* array. The basis vectors $\mathbf{q}_{(1)}, \ldots, \mathbf{q}_{(n)}$ are stored, and are used in conjunction with the various traces to compute the values $\dot{\mathbf{v}}_{(j)}(\mathbf{x}; \mathbf{q}_{(k)})$ using the forward AD mode. These $\dot{\mathbf{v}}_{(j)}(\mathbf{x}; \mathbf{q}_{(k)})$ values are stored in a *tangent trace* array. Noting that each $\dot{\mathbf{u}}_{(j)}(\mathbf{x}; \mathbf{q}_{(k)})$ can be constructed using the tangent and index traces, the various traces are then used to carry out Algorithm 12 as written, with elements of the traces overwritten as necessary.

Linear algebra was performed using the Boost library uBLAS [117].

A.5.2 Examples

In this subsection, key features of the developed methods are illustrated through various example problems. Wherever generalized Jacobian elements were computed in Examples A.5.1 to A.5.4, this was accomplished both by hand and by using the implementation described in Section A.5.1. The generalized Jacobian elements obtained using these two approaches agreed in every case.

In the following three examples, Algorithm 12 is used to obtain generalized Jacobian elements for various $\mathcal{PC}^1$ functions.

Example A.5.1. *Consider the functions $f : \mathbb{R} \to \mathbb{R} : x \mapsto |x| - |x|$ and $g : \mathbb{R} \to \mathbb{R} : x \mapsto \max(x, 0) + \min(x, 0)$, which have been used in [23] and [32] to illustrate the lack of sharp calculus rules for the generalized Jacobian.*

Suppose that elements of $\partial f(0)$ and $\partial g(0)$ are desired. Since f and g are each functions of single variables, Corollary A.3.4 implies that performing Algorithm 12 on f and g at 0 in their above representations yields $f'(0; 1) = 0 \in \partial f(0)$ and $g'(0; 1) = 1 \in \partial g(0)$, respectively.

Table A.1: $\mathcal{PC}^1$-factored representation of f in Example A.5.2

j	Algebraic expression for $\mathbf{v}_{(j)}$	$\mathbf{v}_{(j)}(\mathbf{0})$	$\zeta_{\psi_{(j)}}\left(\mathbf{u}_{(j)}(\mathbf{0})\right)$
0	$\mathbf{v}_{(0)} = \mathbf{x}$	$\mathbf{0}$	–
1	$v_{(1)} = v_{(0),1}$	0	`true`
2	$v_{(2)} = -v_{(0),2}$	0	`true`
3	$v_{(3)} = \min(v_{(1)}, v_{(2)})$	0	`false`
4	$v_{(4)} = v_{(0),2} - v_{(0),1}$	0	`true`
5	$v_{(5)} = \max(v_{(3)}, v_{(4)})$	0	`false`

In this case, analytical generalized Jacobians for f and g are trivial to compute, since f is the zero mapping on $\mathbb{R}$ and g is the identity mapping on $\mathbb{R}$. As a result, for each $x \in \mathbb{R}$, $\partial f(x) = \{0\}$ and $\partial g(x) = \{1\}$, which is consistent with the result obtained from Algorithm 12.

Example A.5.2. *Consider the following function from [16, Example 2.5.2], which was used as an example in our previous work:*

$$f : \mathbb{R}^2 \to \mathbb{R} : (x, y) \mapsto \max(\min(x, -y), y - x).$$

Suppose that an element of $\partial f(\mathbf{0})$ is desired. A $\mathcal{PC}^1$-factored representation of f is given in Table A.1. Unlike in [53], there is no longer any need to express max *and* min *in terms of the absolute value function.*

To carry out Line 2 of Algorithm 12, a function evaluation was carried out to determine the values of $f(\mathbf{0}) = v_{(5)}(\mathbf{0})$, all intermediate variables $\mathbf{v}_{(j)}(\mathbf{0})$, and the values of $\zeta_{\psi_{(j)}}\left(\mathbf{u}_{(j)}(\mathbf{0})\right)$ for each $j \in \{1, \ldots, 5\}$. These are shown in the rightmost two columns of Table A.1. Note that $\zeta_{\psi_{(j)}}\left(\mathbf{u}_{(j)}(\mathbf{0})\right) =$ `false` only for $j \in \{3, 5\}$, and that the $\mathcal{PC}^1$-factored representation of f implies that $\mathbf{u}_{(3)} \equiv (v_{(1)}, v_{(2)})$ and $\mathbf{u}_{(5)} \equiv (v_{(3)}, v_{(4)})$. Thus, $\mathbf{u}_{(3)}(\mathbf{0}) = \mathbf{u}_{(5)}(\mathbf{0}) = (0, 0)$.

To carry out Line 3 of the algorithm, the forward mode of AD was applied to f at $\mathbf{0}$ in the directions $\mathbf{q}_{(1)} = \mathbf{e}_{(1)} = (1, 0)$ and $\mathbf{q}_{(2)} = \mathbf{e}_{(2)} = (0, 1)$. The results are shown in Table A.2, along with algebraic instructions for carrying out the forward mode of AD. The directional derivatives of max *and* min *were evaluated as in Example A.2.5.*

Continuing with the procedure, key iterations of the innermost `for`*–loop in Algo-*

Table A.2: Intermediate quantities used to evaluate $\partial f(\mathbf{0})$ in Example A.5.2

j	Algebraic expression for $\dot{\mathbf{v}}_{(j)}$	$\dot{\mathbf{v}}_{(j)}(\mathbf{0};(1,0))$	$\dot{\mathbf{v}}_{(j)}(\mathbf{0};(0,1))$	$\dot{\mathbf{v}}_{(j)}(\mathbf{0};(2,1))$
0	$\dot{\mathbf{v}}_{(0)} = \mathbf{d}$	$(1,0)$	$(0,1)$	$(2,1)$
1	$\dot{v}_{(1)} = \dot{v}_{(0),1}$	1	0	2
2	$\dot{v}_{(2)} = -\dot{v}_{(0),2}$	0	-1	-1
3	$\dot{v}_{(3)} = \min'((v_{(1)}, v_{(2)}); (\dot{v}_{(1)}, \dot{v}_{(2)}))$	0	-1	-1
4	$\dot{v}_{(4)} = \dot{v}_{(0),2} - \dot{v}_{(0),1}$	-1	1	-1
5	$\dot{v}_{(5)} = \max'((v_{(3)}, v_{(4)}); (\dot{v}_{(3)}, \dot{v}_{(4)}))$	0	1	-1

rithm 12 were as follows:

- *At* $(j,r,k) = (3,1,1)$, $\dot{\mathbf{u}}_{(j,k)} = \dot{\mathbf{u}}_{(3)}(\mathbf{0};(1,0)) = (1,0)$. *Thus,*

$$c \leftarrow \langle \mathbf{a}, \dot{\mathbf{u}}_{(j,k)} \rangle = \langle (1,-1), (1,0) \rangle = 1 \neq 0,$$

  so c^* *was set to* 1, *and* k^* *was set to* 1.

- *At* $(j,r,k) = (3,1,2)$, $\dot{\mathbf{u}}_{(j,k)} = \dot{\mathbf{u}}_{(3)}(\mathbf{0};(0,1)) = (0,-1)$. *Thus,*

$$c \leftarrow \langle \mathbf{a}, \dot{\mathbf{u}}_{(j,k)} \rangle = \langle (1,-1), (0,-1) \rangle = 1.$$

  Since $cc^* = 1 \geq 0$, $\mathbf{q}_{(2)}$ *was left unchanged.*

- *At* $(j,r,k) = (5,1,1)$, $\dot{\mathbf{u}}_{(j,k)} = \dot{\mathbf{u}}_{(5)}(\mathbf{0};(1,0)) = (0,-1)$. *Thus,*

$$c \leftarrow \langle \mathbf{a}, \dot{\mathbf{u}}_{(j,k)} \rangle = \langle (1,-1), (0,-1) \rangle = 1 \neq 0,$$

  so c^* *was set to* 1, *and* k^* *was set to* 1.

- *At* $(j,r,k) = (5,1,2)$, $\dot{\mathbf{u}}_{(j,k)} = \dot{\mathbf{u}}_{(5)}(\mathbf{0};(0,1)) = (-1,1)$. *Thus,*

$$c \leftarrow \langle \mathbf{a}, \dot{\mathbf{u}}_{(j,k)} \rangle = \langle (1,-1), (-1,1) \rangle = -2.$$

  Since $cc^* = -2 < 0$, $\mathbf{q}_{(2)}$ *was updated according to:*

$$\mathbf{q}_{(2)} \leftarrow \mathbf{q}_{(2)} + \left(\frac{-(-2)}{1}\right) \mathbf{q}_{(1)} = (0,1) + 2(1,0) = (2,1).$$

$f'(\mathbf{0};(2,1))$ was then evaluated using the forward mode of AD, and is shown in the rightmost column of Table A.2.

In this case, the single basis vector update was the same as the single update performed in [53, Example 3]. It is nevertheless possible for different $\mathcal{PC}^1$-factored representations of the same function to yield different sequences of basis vector updates.

Thus, exactly as in [53, Example 3], the matrix $\mathbf{B}$ was then defined so as to solve the linear system:

$$\mathbf{B}\begin{bmatrix}1 & 2\\ 0 & 1\end{bmatrix} = \begin{bmatrix}0 & -1\end{bmatrix}.$$

It follows that $\mathbf{B} = \begin{bmatrix}0 & -1\end{bmatrix}$, which is an element of $\partial f(\mathbf{0})$ according to Theorem A.3.2.

This example is simple enough that its generalized Jacobians can be evaluated analytically, as follows. Note that f can be expressed in terms of $\mathcal{C}^1$ selection functions as:

$$f:\mathbb{R}^2\to\mathbb{R}:(x,y)\mapsto\begin{cases} x & \text{if } y\le -x \text{ and } y\le 2x,\\ -y & \text{if } -x\le y\le \frac{1}{2}x,\\ y-x & \text{otherwise.}\end{cases} \tag{A.9}$$

It is readily verified that the three selection functions above are each essentially active for f at $\mathbf{0}$. Property 3 in Lemma A.1.10 then yields $\partial f(\mathbf{0}) = \operatorname{conv}\{[1\ 0],[0\ -1],[-1\ 1]\}\ni\mathbf{B}$, which confirms the result obtained from Algorithm 12.

This example additionally illustrates a phenomenon which is absent in the smooth case. Equation (A.9) demonstrates that f is $\mathcal{PL}$, and is therefore positively homogeneous. It follows that for each $\mathbf{d}\in\mathbb{R}^2$ and each $\epsilon>0$,

$$\Delta_\epsilon f(\mathbf{0};\mathbf{d}) := \frac{f(\epsilon\mathbf{d})-f(\mathbf{0})}{\epsilon} = \lim_{t\to 0^+}\frac{f(\mathbf{0}+t\mathbf{d})-f(\mathbf{0})}{t} = f'(\mathbf{0};\mathbf{d}),$$

and so forward finite differencing provides exact directional derivatives for f at $\mathbf{0}$. If the nonsmoothness of f were ignored, and if forward finite differencing were used to approximate the (actually nonexistent) Jacobian of f at $\mathbf{0}$, then this approach would yield the following for some $\epsilon>0$:

Table A.3: $\mathcal{PC}^1$-factored representation of f in Example A.5.3

j	Algebraic expression for $\mathbf{v}_{(j)}$	$\mathbf{v}_{(j)}(\mathbf{0})$	$\zeta_{\psi_{(j)}}\left(\mathbf{u}_{(j)}(\mathbf{0})\right)$
0	$\mathbf{v}_{(0)} = \mathbf{x}$	$\mathbf{0}$	–
1	$v_{(1)} = v_{(0),1} - v_{(0),2}$	0	`true`
2	$v_{(2)} = \lvert v_{(1)} \rvert$	0	`false`
3	$v_{(3)} = 1 + v_{(2)}$	1	`true`
4	$v_{(4)} = v_{(1)} v_{(3)}$	0	`true`

Table A.4: Intermediate quantities used to evaluate $\partial f(\mathbf{0})$ in Example A.5.3

j	Algebraic expression for $\dot{\mathbf{v}}_{(j)}$	$\dot{\mathbf{v}}_{(j)}(\mathbf{0};(1,0))$	$\dot{\mathbf{v}}_{(j)}(\mathbf{0};(0,1))$	$\dot{\mathbf{v}}_{(j)}(\mathbf{0};(1,1))$
0	$\dot{\mathbf{v}}_{(0)} = \mathbf{d}$	$(1,0)$	$(0,1)$	$(1,1)$
1	$\dot{v}_{(1)} = \dot{v}_{(0),1} - \dot{v}_{(0),2}$	1	-1	0
2	$\dot{v}_{(2)} = \mathrm{abs}'(v_{(1)}; \dot{v}_{(1)})$	1	1	0
3	$\dot{v}_{(3)} = \dot{v}_{(2)}$	1	1	0
4	$\dot{v}_{(4)} = v_{(1)}\dot{v}_{(3)} + v_{(3)}\dot{v}_{(1)}$	1	-1	0

$$\begin{aligned}\mathbf{J}f(\mathbf{0}) &\stackrel{?}{=} \begin{bmatrix}\Delta_\epsilon f\,(\mathbf{0};(1,0)) & \Delta_\epsilon f\,(\mathbf{0};(0,1))\end{bmatrix} \\ &= \begin{bmatrix} f'(\mathbf{0};(1,0)) & f'(\mathbf{0};(0,1))\end{bmatrix} = \begin{bmatrix}0 & 1\end{bmatrix} \notin \partial f(\mathbf{0}).\end{aligned}$$

Thus, in the limit $\epsilon \to 0^+$, the above approximation does not tend to an element of $\partial f(\mathbf{0})$. Centered finite differencing similarly fails to provide an element of $\partial f(\mathbf{0})$, since

$$\lim_{\epsilon \to 0^+} \frac{1}{2\epsilon}\begin{bmatrix}(f(\epsilon,0) - f(-\epsilon,0)) & (f(0,\epsilon) - f(0,-\epsilon))\end{bmatrix} = \tfrac{1}{2}\begin{bmatrix}1 & 1\end{bmatrix} \notin \partial f(\mathbf{0}).$$

It follows that if the active cones of a nonsmooth function are not known a priori, *then finite differencing does not necessarily approximate elements of the generalized Jacobian.*

Example A.5.3. *Consider the $\mathcal{PC}^1$ function $f : \mathbb{R}^2 \to \mathbb{R} : (x,y) \mapsto (1 + |x-y|)(x-y)$. A $\mathcal{PC}^1$-factored representation of f is given in Table A.3, and Table A.4 shows the result of applying the forward AD mode to f at $\mathbf{0}$ in certain directions.*

Though f is expressed in terms of the nonsmooth absolute value function, it is in fact $\mathcal{C}^1$ on its domain. If this is known a priori, *then $\mathbf{J}f(\mathbf{0})$ may be evaluated using two applications of the forward mode of AD, since* (A.1) *implies that*

$$\mathbf{J}f(\mathbf{0}) = \mathbf{J}f(\mathbf{0})\,\mathbf{I} = \begin{bmatrix} f'(\mathbf{0};(1,0)) & f'(\mathbf{0};(0,1)) \end{bmatrix} = \begin{bmatrix} 1 & -1 \end{bmatrix}. \tag{A.10}$$

Suppose it is not known a priori *that f is $\mathcal{C}^1$, and that an element of $\partial f(\mathbf{0})$ is desired. If Algorithm 12 is applied, then only one basis vector update is carried out. This update occurs during the iteration of the innermost* `for`*–loop at which $(j,r,k) = (2,1,2)$, and is as follows:*

$$\mathbf{q}_{(2)} \leftarrow \mathbf{q}_{(2)} + \left(\frac{-(-1)}{1}\right)\mathbf{q}_{(1)} = (0,1) + (1,0) = (1,1).$$

$f'(\mathbf{0};(1,1))$ was therefore evaluated using the forward mode of AD, and is shown in Table A.4.

Algorithm 12 therefore returns a matrix $\mathbf{B}$ for which

$$\mathbf{B}\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

Thus $\mathbf{B} = \begin{bmatrix} 1 & -1 \end{bmatrix}$, which agrees with (A.10)*. Algorithm 12 therefore returns the Jacobian of f in this case, which is consistent with Corollary A.3.3.*

In the following example, a semismooth Newton method is used to determine a root of a $\mathcal{PC}^1$ function. In this case, if the initial guess is chosen from a certain set of nonzero (Lebesgue) measure, then the Newton method visits domain points at which the function is nondifferentiable. Thus, even though Rademacher's Theorem implies that $\mathcal{PC}^1$ functions can only be nondifferentiable on a set of measure zero, these points of nondifferentiability are still reachable in practice.

Example A.5.4. *Consider the following function:*

$$\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^2 :$$
$$(x,y) \mapsto \begin{cases} (x,y) & \text{if } x \leq -4, \\ (\max[\frac{1}{2}x - 2, -xy - \frac{1}{2}x - 4y - 6], y - \frac{1}{2}x - 2) & \text{if } -4 \leq x \leq -2, \\ (\max[x-1, 2x-2y-1], y-1) & \text{if } -2 \leq x \leq 2, \\ (\max[1, 3-2y], y-1) & \text{if } 2 \leq x. \end{cases}$$

It is readily verified that $\mathbf{f}$ is continuous on $\mathbb{R}^2$, and is therefore $\mathcal{PC}^1$. Suppose that a root of $\mathbf{f}$ is desired, and so a semismooth Newton method [92] is applied. Now, Lemma A.1.10 and

Theorem A.3.2 imply that the generalized Jacobian elements computed by Algorithm 12 are each Jacobians of essentially active selection functions. When only these generalized Jacobian elements are used, the semismooth Newton method reduces to the nonsmooth Newton method described in [65]. Suppose that the initial guess $\mathbf{x}^{(0)} \in \mathbb{R}^2$ *satisfies* $x_1^{(0)} < -4$.

Since $\partial\mathbf{f}(\mathbf{x}^{(0)}) = \{\mathbf{J}\mathbf{f}(\mathbf{x}^{(0)})\} = \{\mathbf{I}\}$, *the result of the first Newton step is*

$$\mathbf{x}^{(1)} \leftarrow \mathbf{x}^{(0)} - \mathbf{I}^{-1}\mathbf{f}(\mathbf{x}^{(0)}) = \mathbf{x}^{(0)} - \mathbf{x}^{(0)} = \mathbf{0}.$$

Now,

$$\partial\mathbf{f}(\mathbf{x}^{(1)}) = \partial\mathbf{f}(\mathbf{0}) = \operatorname{conv}\left\{\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}, \begin{bmatrix}2 & -2\\ 0 & 1\end{bmatrix}\right\} = \left\{\begin{bmatrix}1+\lambda & -2\lambda\\ 0 & 1\end{bmatrix} : \lambda \in [0,1]\right\}$$

Noting that for each $\lambda \in [0,1]$,

$$\begin{bmatrix}1+\lambda & -2\lambda\\ 0 & 1\end{bmatrix}^{-1} = \frac{1}{1+\lambda}\begin{bmatrix}1 & 2\lambda\\ 0 & 1+\lambda\end{bmatrix},$$

the second Newton step is as follows, for some $\lambda \in [0,1]$*:*

$$\mathbf{x}^{(2)} \leftarrow \mathbf{0} - \frac{1}{1+\lambda}\begin{bmatrix}1 & 2\lambda\\ 0 & 1+\lambda\end{bmatrix}\begin{bmatrix}-1\\ -1\end{bmatrix} = (2 - \tfrac{1}{1+\lambda}, 1) \in \operatorname{conv}\left\{(1,1), (\tfrac{3}{2}, 1)\right\}.$$

(In this case, the C++ implementation of Algorithm 12 produced the result corresponding to $\lambda = 1$.*) Now,* $\mathbf{f}(\mathbf{y}) = \mathbf{y} - (1,1)$ *for each* $\mathbf{y}$ *near* $\mathbf{x}^{(2)}$. *Thus* $\partial\mathbf{f}(\mathbf{x}^{(2)}) = \{\mathbf{I}\}$, *and so the third Newton step is:*

$$\mathbf{x}^{(3)} \leftarrow \mathbf{x}^{(2)} - \mathbf{I}(\mathbf{x}^{(2)} - (1,1)) = (1,1).$$

Since $\mathbf{f}(1,1) = \mathbf{0}$, *the semismooth Newton method is successful, and terminates.*

For illustration, suppose that, instead of evaluating an element of $\partial\mathbf{f}(\mathbf{0})$, *the method for evaluating Jacobians of smooth functions were applied naïvely. In this case, the second Newton step in the above procedure becomes:*

$$\mathbf{x}^{(2)} \leftarrow \mathbf{0} - \begin{bmatrix}\mathbf{f}'(\mathbf{0};(1,0)) & \mathbf{f}'(\mathbf{0};(0,1))\end{bmatrix}^{-1} \begin{bmatrix}-1\\-1\end{bmatrix} = -\begin{bmatrix}1 & -2\\0 & 1\end{bmatrix}^{-1} \begin{bmatrix}-1\\-1\end{bmatrix} = (3,1).$$

Now,

$$\partial\mathbf{f}(3,1) = \left\{\begin{bmatrix}0 & -2\\0 & 1\end{bmatrix}\right\},$$

and so each element of $\partial\mathbf{f}(3,1)$ *is singular. The semismooth Newton method therefore cannot proceed after the incorrect second step, and terminates without finding a solution.*

In the following example, the C++ implementation of Algorithm 12 is applied to a problem in pinch analysis in [21]. This problem is sufficiently complicated that an analytical generalized Jacobian is nontrivial to obtain.

Example A.5.5. *Algorithm 14, adapted from [21], is a procedure for conducting a pinch analysis of a chemical process at the design stage. This procedure computes the minimum heating* Q^{H} *and cooling* Q^{C} *demanded by the process, in terms of the following process parameters.*

Algorithm 14 Returns the minimum heating Q^{H} and cooling Q^{C} required by a process

for all $j \in \mathcal{H}$ **do**
  Set $T_j^{\mathrm{in}} \leftarrow T_j^{\mathrm{in}} - \Delta T$
  Set $T_j^{\mathrm{out}} \leftarrow T_j^{\mathrm{out}} - \Delta T$
end for
Set $Q^{\mathrm{H}} \leftarrow 0$
for all $i \in \mathcal{H} \cup \mathcal{C}$ **do**
  Set $z^{\mathrm{P}} \leftarrow 0$
  for all $j \in \mathcal{H} \cup \mathcal{C}$ **do**
    Set $z^{\mathrm{P}} \leftarrow^{\mathrm{P}} + f_j c_j^{\mathrm{P}}[\max(0, T_j^{\mathrm{out}} - T_i^{\mathrm{in}}) - \max(0, T_j^{\mathrm{in}} - T_i^{\mathrm{in}})]$
  end for
  Set $Q^{\mathrm{H}} \leftarrow \max(Q^{\mathrm{H}}, z^{\mathrm{P}})$
end for
Set $\Omega \leftarrow 0$
for all $j \in \mathcal{I}$ **do**
  Set $\Omega \leftarrow \Omega + f_j c_j^{\mathrm{P}}(T_j^{\mathrm{in}} - T_j^{\mathrm{out}})$
end for
Set $Q^{\mathrm{C}} \leftarrow Q^{\mathrm{H}} + \Omega$
return $Q^{\mathrm{H}}, Q^{\mathrm{C}}$

Suppose that the process contains N process streams, which are indexed by $\mathcal{I} := \{1, 2, \ldots, N\}$. For each $j \in \mathcal{I}$, f_j [kg/s] denotes the material flow rate of stream j, c_j^{P} [kJ/(kg·°C)] denotes the specific heat capacity of this material, and T_j^{in} and T_j^{out} [°C] denote the desired inlet and outlet stream temperatures. $\mathcal{H} := \{j \in \mathcal{I} : T_j^{\mathrm{in}} > T_j^{\mathrm{out}}\}$ denotes the indices of "hot" streams which require cooling, and $\mathcal{C} := \{j \in \mathcal{I} : T_j^{\mathrm{in}} < T_j^{\mathrm{out}}\}$ denotes the indices of "cold" streams which require heating. ΔT [°C] denotes the minimum temperature approach permitted in a heat integration scheme for the process. It is assumed that a hot utility is available at a temperature T^{H}, and that a cold utility is available at a temperature T^{C}, where

$$T^{\mathrm{H}} > \max\{T_j^{\mathrm{out}} + \Delta T : j \in \mathcal{C}\} \quad \textit{and} \quad T^{\mathrm{C}} < \min\{T_j^{\mathrm{out}} - \Delta T : j \in \mathcal{H}\}.$$

Suppose that all process parameters mentioned above are held constant, except for the inlet and outlet stream temperatures $\mathbf{T} := (T_1^{\mathrm{in}}, T_1^{\mathrm{out}}, T_2^{\mathrm{in}}, \ldots, T_N^{\mathrm{out}})$. Under the assumption that there is no $j \in \mathcal{I}$ for which $T_j^{\mathrm{in}} = T_j^{\mathrm{out}}$, sufficiently small perturbations of $\mathbf{T}$ will not alter $\mathcal{H}$. In this case, unrolling the `for`*–loops in Algorithm 14 shows that this algorithm is essentially a $\mathcal{PC}^1$-factored representation of $\boldsymbol{\chi} := (Q^{\mathrm{H}}, Q^{\mathrm{C}})$, expressed as a function of $\mathbf{T}$. By inspection, this representation involves $N(2N+1)$ computations of* max *functions on $\mathbb{R}^2$, and it is unclear at the outset whether nondifferentiable domain points of the* max *function in Line 11 of the algorithm are reached.*

Consider a variation of Example 1 in [21, Appendix B], in which a chemical process has four process streams, with $f_j c_j^{\mathrm{P}}$-values given by the second column in Table A.5. Let $\mathbf{T}_0 \in \mathbb{R}^8$ denote the vector $(T_1^{\mathrm{in}}, T_1^{\mathrm{out}}, T_2^{\mathrm{in}}, \ldots, T_4^{\mathrm{out}})$, with the values of each T_j^{in} and T_j^{out} given by the rightmost two columns of Table A.5. Suppose that an element of $\partial\boldsymbol{\chi}(\mathbf{T}_0)$ is required, assuming a minimum temperature approach of $\Delta T = 10$°C. To obtain this $\partial\boldsymbol{\chi}(\mathbf{T}_0)$ element, the C++ implementation of Algorithm 12 was applied to the $\mathcal{PC}^1$-factored representation presented in Algorithm 14. To verify that Algorithm 14 was entered correctly, the results in Example 1 in [21, Appendix B] were replicated, and several smaller toy examples were tested in the same way.

Given these parameter values, the minimum heating and cooling demanded by the pro-

Table A.5: Stream parameters used in Example A.5.5

Stream index j	$f_j c_j^{\mathrm{P}}$ [kW/°C]	T_j^{in} [°C]	T_j^{out} [°C]
1	8.79	160	93
2	10.55	170	126
3	7.62	60	160
4	6.08	116	260

cess were found to be $Q^{\mathrm{H}} = 639.5$ kW *and* $Q^{\mathrm{C}} = 55.11$ kW. *Application of Algorithm 12 produced:*

$$\begin{bmatrix} -3.15 & 0 & -10.55 & 0 & 0 & 7.62 & 0 & 6.08 \\ 5.64 & -8.79 & 0 & -10.55 & 7.62 & 0 & 6.08 & 0 \end{bmatrix} \in \partial\boldsymbol{\chi}(\mathbf{T}_0),$$

where each element has units [kW/°C]. *During this application of Algorithm 12, the following two basis vector updates were performed:*

$$\mathbf{q}_{(6)} \leftarrow (0,0,1,0,0,1,0,0) \quad \textit{and} \quad \mathbf{q}_{(7)} \leftarrow (0,0,0,1,0,0,1,0).$$

A.6 Intermediate results

This appendix provides several intermediate results which are used in the proof of Theorem A.3.2.

Lemma A.6.1. *Consider a conical subdivision* Λ *of* $\mathbb{R}^n$ *and a corresponding hyperplane normal set* $\mathcal{H} = \{\mathbf{a}_{(1)}, \ldots, \mathbf{a}_{(p)}\}$. *For any* $\bar{\mathbf{s}} \in \{-1,1\}^p$, *let* $\bar{\sigma} = \bigcap_{r=1}^{p} \{\mathbf{x} \in \mathbb{R}^n : \bar{s}_r \langle \mathbf{a}_{(r)}, \mathbf{x} \rangle \leq 0\}$. *There exists a cone* $\sigma \in \Lambda$ *for which* $\bar{\sigma} \subset \sigma$.

Proof. For each $r \in \{1, \ldots, p\}$ and each $\mathbf{d} \in \bar{\sigma}$, $\bar{s}_r \langle \mathbf{a}_{(r)}, \mathbf{d} \rangle \leq 0$. It follows that for each $r \in \{1, \ldots, p\}$, either $\bar{s}_r \langle \mathbf{a}_{(r)}, \mathbf{d} \rangle = 0 = \langle \mathbf{a}_{(r)}, \mathbf{d} \rangle$ for all $\mathbf{d} \in \bar{\sigma}$, or there exists some $\mathbf{d} \in \bar{\sigma}$ such that $\bar{s}_r \langle \mathbf{a}_{(r)}, \mathbf{d} \rangle < 0$. Thus, $\{1, \ldots, p\}$ can be partitioned into sets $\mathcal{I}$ and $\mathcal{J}$, where

$$\mathcal{I} := \{r \in \{1, \ldots, p\} : \langle \mathbf{a}_{(r)}, \mathbf{d} \rangle = 0, \quad \forall \mathbf{d} \in \bar{\sigma}\}, \text{ and}$$
$$\mathcal{J} := \{r \in \{1, \ldots, p\} : \exists \mathbf{d} \in \bar{\sigma} \text{ such that } \bar{s}_r \langle \mathbf{a}_{(r)}, \mathbf{d} \rangle < 0\}.$$

The cases in which $\mathcal{J}$ is either empty or nonempty will be considered separately.

If $\mathcal{J}$ is empty, then for each $r \in \mathcal{I} = \{1, \ldots, p\}$ and each $\mathbf{d} \in \bar{\sigma}$, for any $\mathbf{s} \in \{-1, 0, 1\}^p$, $s_r \langle \mathbf{a}_{(r)}, \mathbf{d} \rangle = 0 \leq 0$, and so $\mathbf{d} \in \bigcap_{r=1}^{p} \{\mathbf{x} \in \mathbb{R}^n : s_r \langle \mathbf{a}_{(r)}, \mathbf{x} \rangle \leq 0\}$. Using the definition of $\mathcal{H}$, it follows that $\mathbf{d} \in \sigma$ for every $\sigma \in \Lambda$ and every $\mathbf{d} \in \bar{\sigma}$, which implies that $\bar{\sigma} \subset \sigma$ for *every* $\sigma \in \Lambda$.

If $\mathcal{J}$ is nonempty, then for each $r \in \mathcal{J}$, choose $\mathbf{d}_{(r)} \in \bar{\sigma}$ such that $\bar{s}_r \langle \mathbf{a}_{(r)}, \mathbf{d}_{(r)} \rangle < 0$. Consider the vector $\mathbf{d}^* := \sum_{r \in \mathcal{J}} \mathbf{d}_{(r)}$. Since $\bar{\sigma}$ is a convex cone by inspection, $\mathbf{d}^* \in \bar{\sigma}$. Moreover, for each $r \in \mathcal{J}$,

$$\bar{s}_r \langle \mathbf{a}_{(r)}, \mathbf{d}^* \rangle = \sum_{\rho \in \mathcal{J}} \bar{s}_r \langle \mathbf{a}_{(r)}, \mathbf{d}_{(\rho)} \rangle = \bar{s}_r \langle \mathbf{a}_{(r)}, \mathbf{d}_{(r)} \rangle + \sum_{\rho \in \mathcal{J} \setminus \{r\}} \bar{s}_r \langle \mathbf{a}_{(r)}, \mathbf{d}_{(\rho)} \rangle < 0, \quad \text{(A.11)}$$

where the strict inequality above follows from the construction of $\mathbf{d}_{(r)}$ and the definition of $\bar{\sigma}$. Since Λ is a partition of $\mathbb{R}^n$, there exists some $\sigma \in \Lambda$ such that $\mathbf{d}^* \in \sigma$, and some $\mathbf{s} \in \{-1, 0, 1\}^p$ such that $\sigma = \bigcap_{r=1}^{p} \{\mathbf{x} \in \mathbb{R}^n : s_r \langle \mathbf{a}_{(r)}, \mathbf{x} \rangle \leq 0\}$. To complete this proof, it will be shown that $\bar{\sigma} \subset \sigma$. It suffices to show that for each $r \in \{1, \ldots, p\}$,

$$\bar{\sigma} \subset \{\mathbf{x} \in \mathbb{R}^n : s_r \langle \mathbf{a}_{(r)}, \mathbf{x} \rangle \leq 0\}. \quad \text{(A.12)}$$

The cases in which $r \in \mathcal{I}$ and $r \in \mathcal{J}$ will be considered separately.

If $r \in \mathcal{I}$, then the definition of $\mathcal{I}$ implies that, regardless of the value of s_r,

$$\bar{\sigma} \subset \{\mathbf{x} \in \mathbb{R}^n : \langle \mathbf{a}_{(r)}, \mathbf{x} \rangle = 0\} \subset \{\mathbf{x} \in \mathbb{R}^n : s_r \langle \mathbf{a}_{(r)}, \mathbf{x} \rangle \leq 0\}.$$

If $r \in \mathcal{J}$, then suppose, to obtain a contradiction, that $s_r = -\bar{s}_r$. Since $\mathbf{d}^* \in \sigma$, it follows that $0 \geq s_r \langle \mathbf{a}_{(r)}, \mathbf{d}^* \rangle = -\bar{s}_r \langle \mathbf{a}_{(r)}, \mathbf{d}^* \rangle$, which contradicts (A.11). Thus $s_r = -\bar{s}_r$ cannot be true, and so either $s_r = 0$ or $s_r = \bar{s}_r$. If $s_r = 0$, then $\{\mathbf{x} \in \mathbb{R}^n : s_r \langle \mathbf{a}_{(r)}, \mathbf{x} \rangle \leq 0\} = \mathbb{R}^n$, and so (A.12) is trivial. If $s_r = \bar{s}_r$, then the definition of $\bar{\sigma}$

yields

$$\bar{\sigma} \subset \{\mathbf{x} \in \mathbb{R}^n : \bar{s}_r \langle \mathbf{a}_{(r)}, \mathbf{x} \rangle \leq 0\} = \{\mathbf{x} \in \mathbb{R}^n : s_r \langle \mathbf{a}_{(r)}, \mathbf{x} \rangle \leq 0\}.$$

Since (A.12) has now been demonstrated for each $r \in \{1, \ldots, p\}$, it follows that $\bar{\sigma} \subset \sigma$. □

Lemma A.6.2. *At every point in Algorithm 12, the matrix* $\mathbf{Q} \equiv \begin{bmatrix} \mathbf{q}_{(1)} & \cdots & \mathbf{q}_{(n)} \end{bmatrix}$ *is unit upper triangular, and is therefore nonsingular.*

Proof. Noting that $\mathbf{Q} = \mathbf{I}$ initially, and that $\mathbf{Q}$ is only altered during Line 15 of the algorithm, it suffices to show that execution of Line 15 preserves the unit upper triangularity of $\mathbf{Q}$.

In Line 15, a scalar multiple of the $(k^*)^{\text{th}}$ column of $\mathbf{Q}$ is added to the kth column, where $k > k^*$. Moreover, if $\mathbf{Q}$ is unit upper triangular immediately before Line 15 is carried out, then the $(k^*)^{\text{th}}$ column of $\mathbf{Q}$ contains only zeroes below its $(k^*)^{\text{th}}$ entry. The unit upper triangularity of $\mathbf{Q}$ is therefore unaffected by execution of Line 15. □

Corollary A.6.3. *At every point in Algorithm 12, the set* $\sigma_{\mathbf{q}} := \operatorname{cone}\{\mathbf{q}_{(1)}, \ldots, \mathbf{q}_{(n)}\}$ *has nonempty interior.*

Proof. By Lemma A.6.2, $\mathbf{Q}$ is nonsingular throughout Algorithm 12, and so its columns comprise a basis of $\mathbb{R}^n$. Consider the vector $\mathbf{q}^* := \sum_{k=1}^n \mathbf{q}_{(k)} \in \sigma_{\mathbf{q}}$. For any $\mathbf{y} \in \mathbb{R}^n$,

$$\mathbf{q}^* + \mathbf{y} = \mathbf{q}^* + \mathbf{Q}(\mathbf{Q}^{-1}\mathbf{y}) = \sum_{k=1}^{n} \left(1 + (\mathbf{Q}^{-1}\mathbf{y})_k\right) \mathbf{q}_{(k)}. \tag{A.13}$$

Let $\|\cdot\|$ denote the Euclidean norm on $\mathbb{R}^n$. For sufficiently small $\delta > 0$, any $\mathbf{y} \in \mathbb{R}^n$ with $\|\mathbf{y}\| < \delta$ satisfies $\|\mathbf{Q}^{-1}\mathbf{y}\| < 1$. In this case, $(\mathbf{Q}^{-1}\mathbf{y})_k > -1$ for each $k \in \{1, \ldots, n\}$. It follows from (A.13) that $(\mathbf{q}^* + \mathbf{y}) \in \sigma_{\mathbf{q}}$ whenever $\|\mathbf{y}\| < \delta$, and so $\mathbf{q}^* \in \operatorname{int}(\sigma_{\mathbf{q}})$. □

Lemma A.6.4. *Given a set* $S \subset \mathbb{R}^n$ *with nonempty interior and a conical subdivision* Λ *of* $\mathbb{R}^n$*, there exists a polyhedral cone* $\sigma \in \Lambda$ *for which* $\operatorname{int}(S) \cap \operatorname{int}(\sigma)$ *is nonempty.*

Proof. Consider some $\mathbf{x} \in \text{int}(S)$. There exists a neighborhood $N \subset \mathbb{R}^n$ of $\mathbf{x}$ such that $N \subset S$. Since Λ is a partition of $\mathbb{R}^n$, there exists some $\sigma \in \Lambda$ such that $\mathbf{x} \in \sigma$. By definition of Λ, there exists some $\mathbf{y} \in \text{int}(\sigma)$. Since σ is convex, $\mathbf{x} \in \sigma$, and $\mathbf{y} \in \text{int}(\sigma)$, it follows that $\mathbf{z}(\lambda) := (\lambda\mathbf{y} + (1-\lambda)\mathbf{x}) \in \text{int}(\sigma)$ for each $\lambda \in (0,1)$. Moreover, $\mathbf{z}(\lambda) \in N \subset \text{int}(S)$ when $\lambda > 0$ is sufficiently small, in which case $\mathbf{z}(\lambda) \in \text{int}(S) \cap \text{int}(\sigma)$. □

Lemma A.6.5. *Given a finite set of vectors $\mathbf{s}_{(1)}, \ldots, \mathbf{s}_{(p)} \in \mathbb{R}^n$ and a vector $\mathbf{a} \in \mathbb{R}^n$, suppose that the following procedure is carried out:*

```
Set c* ← 0
for k = 1 to p do
   Set c ← ⟨a, s_(k)⟩ ∈ ℝ
   if c ≠ 0 then
      if c* = 0 then
         Set c* ← c and k* ← k
      else if cc* < 0 then
         Set s_(k) ← s_(k) − (c/c*) s_(k*)
      end if
   end if
end for
```

At the end of this procedure, there is some $s \in \{-1, 1\}$ such that for each $k \in \{1, \ldots, p\}$, $s\langle \mathbf{a}, \mathbf{s}_{(k)} \rangle \leq 0$. Moreover, this relationship is invariant under further operations of the form $\mathbf{s}_{(k_1)} \leftarrow \mathbf{s}_{(k_1)} + \beta \mathbf{s}_{(k_2)}$ with $k_1, k_2 \in \{1, \ldots, p\}$ and $\beta > 0$.

Proof. If k^* is never assigned a value during the procedure, then $\langle \mathbf{a}, \mathbf{s}_{(k)} \rangle = 0$ for each $k \in \{1, \ldots, p\}$, and so s may be arbitrarily set to 1. Otherwise, suppose that k^* is assigned by the procedure to some $q \in \{1, \ldots, p\}$. By construction of k^*, there is some $s^* \in \{-1, 1\}$ such that $s^*\langle \mathbf{a}, \mathbf{s}_{(q)} \rangle < 0$.

Now, for each $k \neq q$ in $\{1, \ldots, p\}$, if $k < q$, then $\mathbf{s}_{(k)}$ remains unchanged during the procedure and $s^*\langle \mathbf{a}, \mathbf{s}_{(k)} \rangle = 0$.

If $k > q$ and if $s^*\langle \mathbf{a}, \mathbf{s}_{(k)} \rangle \leq 0$ initially, then $\langle \mathbf{a}, \mathbf{s}_{(k)} \rangle \langle \mathbf{a}, \mathbf{s}_{(q)} \rangle \geq 0$, and so $\mathbf{s}_{(k)}$ remains unchanged during the procedure. Otherwise, if $s^*\langle \mathbf{a}, \mathbf{s}_{(k)} \rangle > 0$ initially, then $\langle \mathbf{a}, \mathbf{s}_{(k)} \rangle \langle \mathbf{a}, \mathbf{s}_{(q)} \rangle < 0$ initially, and so $\mathbf{s}_{(k)}$ is updated by the procedure to $\mathbf{s}_{(k)_{new}} := \mathbf{s}_{(k)} - \left(\frac{\langle \mathbf{a}, \mathbf{s}_{(k)} \rangle}{\langle \mathbf{a}, \mathbf{s}_{(q)} \rangle} \right) \mathbf{s}_{(q)}$. Thus,

$$\langle \mathbf{a}, \mathbf{s}_{(k)_{new}} \rangle = \langle \mathbf{a}, \mathbf{s}_{(k)} \rangle - \left(\frac{\langle \mathbf{a}, \mathbf{s}_{(k)} \rangle}{\langle \mathbf{a}, \mathbf{s}_{(q)} \rangle} \right) \langle \mathbf{a}, \mathbf{s}_{(q)} \rangle = 0,$$

and so $s^* \langle \mathbf{a}, \mathbf{s}_{(k)} \rangle = 0$ at the end of the procedure.

Combining the above cases, it follows that $s^* \langle \mathbf{a}, \mathbf{s}_{(k)} \rangle \leq 0$ for each $k \in \{1, \ldots, p\}$ at the end of the procedure. Setting s to s^* then yields the first result of the lemma.

To obtain the remaining result of the lemma, it suffices to note that the set $\{\mathbf{y} \in \mathbb{R}^n : s\langle \mathbf{a}, \mathbf{y} \rangle \leq 0\}$ is a convex cone, and is therefore closed under nonnegative combinations of its elements. □

Lemma A.6.6. *For each $j \in \{0, 1, \ldots, \ell\}$, at each point in Algorithm 12 after the jth iteration of the outermost* `for`*–loop, the basis vectors $\mathbf{q}_{(1)}, \ldots, \mathbf{q}_{(n)}$ satisfy:*

$$\begin{aligned} &\forall i \in \{0, 1, \ldots, j\}, \exists \mathbf{v}^*_{(i)} \in \mathcal{E}_{\mathbf{v}_{(i)}}(\mathbf{x}) \text{ such that} \\ &\qquad \forall \mathbf{d} \in \operatorname{cone}\{\mathbf{q}_{(1)}, \ldots, \mathbf{q}_{(n)}\}, \quad \mathbf{v}_{(i)}{}'(\mathbf{x}; \mathbf{d}) = \mathbf{J}\mathbf{v}^*_{(i)}(\mathbf{x})\, \mathbf{d}, \end{aligned} \tag{A.14}$$

where $\mathcal{E}_{\mathbf{v}_{(i)}}(\mathbf{x})$ denotes some sufficient collection of essentially active selection functions for $\mathbf{v}_{(i)}$ at $\mathbf{x}$.

Proof. This proof proceeds by induction on $j \in \{0, 1, \ldots, \ell\}$. For notational simplicity, let Q denote the set $\{\mathbf{q}_{(1)}, \ldots, \mathbf{q}_{(n)}\}$.

Base case: Since $\mathbf{v}_{(0)}(\mathbf{y}) = \mathbf{y}$ for all $\mathbf{y} \in X$, it follows that $\{\mathbf{v}_{(0)}\}$ is a sufficient collection of essentially active selection functions for $\mathbf{v}_{(0)}$ at $\mathbf{x}$, and that $\mathbf{v}_{(0)}{}'(\mathbf{x}; \mathbf{d}) = \mathbf{d} = \mathbf{J}\mathbf{v}_{(0)}(\mathbf{x})\, \mathbf{d}$ for each $\mathbf{d} \in \mathbb{R}^n$.

Inductive step: Suppose that for some particular $j^* \in \{1, 2, \ldots, \ell\}$, (A.14) holds with $j = j^* - 1$ once the $(j^* - 1)^{\text{th}}$ iteration of the outermost `for`–loop has been completed.

Let Λ denote a particular active conical subdivision of $\psi_{(j^*)}$ at $\mathbf{u}_{(j^*)}(\mathbf{x})$ to which $H_{\psi_{(j^*)}}\left(\mathbf{u}_{(j^*)}(\mathbf{x})\right)$ corresponds. As an intermediate result, it will be shown that immediately after the $(j^*)^{\text{th}}$ iteration of the outermost `for`–loop, there exists some $\sigma^* \in \Lambda$ such that $\dot{\mathbf{u}}_{(j^*,k)} := \dot{\mathbf{u}}_{(j^*)}(\mathbf{x}; \mathbf{q}_{(k)}) \in \sigma^*$ for each $k \in \{1, \ldots, n\}$. If

$$\zeta_{\psi_{(j^*)}}\left(\mathbf{u}_{(j^*)}(\mathbf{x})\right) = \texttt{true},$$

then $\mathbb{R}^{n_{j^*}} \in \Lambda$, and so the result is trivial. Thus, it will be assumed that $\zeta_{\psi_{(j^*)}}\left(\mathbf{u}_{(j^*)}(\mathbf{x})\right) =$ `false`. It follows from Lemma A.2.13 and the inductive assumption that immediately after the $(j^*)^{\text{th}}$ iteration of the outermost `for`–loop,

$$\dot{\mathbf{u}}_{(j^*)}(\mathbf{x};\mathbf{d}) = [\dot{\mathbf{v}}_{(i)}(\mathbf{x};\mathbf{d})]_{i \prec j^*} = [\mathbf{J}\mathbf{v}^*_{(i)}(\mathbf{x})]_{i \prec j^*}\mathbf{d}, \quad \forall \mathbf{d} \in \operatorname{cone} Q.$$

Hence, $\dot{\mathbf{u}}_{(j^*)}(\mathbf{x};\cdot)$ is linear on $\operatorname{cone} Q$, and so Lines 15 and 16 of the algorithm produce the following change in $\dot{\mathbf{u}}_{(j^*,k)}$:

$$\dot{\mathbf{u}}_{(j^*,k)} \leftarrow \dot{\mathbf{u}}_{(j^*,k)} + \alpha\dot{\mathbf{u}}_{(j^*,k^*)}. \tag{A.15}$$

Lemmas A.6.1 and A.6.5 then imply that immediately after the $(j^*)^{\text{th}}$ iteration of the outermost `for`–loop, $\dot{\mathbf{u}}_{(j^*,1)}, \ldots, \dot{\mathbf{u}}_{(j^*,n)}$ all lie in a single cone $\sigma^* \in \Lambda$. This completes the proof of the intermediate result.

With $\sigma^* \in \Lambda$ chosen as in the statement of the intermediate result, let ψ^* denote the conically active selection function for $\psi_{(j^*)}$ at $\mathbf{u}_{(j^*)}(\mathbf{x})$ corresponding to σ^*. Immediately after the $(j^*)^{\text{th}}$ iteration of the outermost `for`–loop, the inductive assumption and Property 3 in Lemma A.1.8 imply that for each $\mathbf{d} \in \operatorname{cone} Q$,

$$\mathbf{v}_{(j^*)}{}'(\mathbf{x};\mathbf{d}) = \psi_{(j^*)}{}'(\mathbf{u}_{(j^*)}(\mathbf{x});\dot{\mathbf{u}}_{(j^*)}(\mathbf{x};\mathbf{d})) = \mathbf{J}\psi^*(\mathbf{u}_{(j^*)}(\mathbf{x}))[\mathbf{J}\mathbf{v}^*_{(i)}(\mathbf{x})]_{i \prec j^*}\mathbf{d}. \tag{A.16}$$

This relationship continues to hold for the remainder of the algorithm, since $\alpha := -\frac{c}{c^*} > 0$ whenever $cc^* < 0$, and so Line 15 transforms $\operatorname{cone} Q$ into a subset of itself. Thus, $\mathbf{v}_{(j^*)}{}'(\mathbf{x};\cdot)$ is linear on $\operatorname{cone} Q$ at each point in the algorithm after the $(j^*)^{\text{th}}$ iteration of the outermost `for`–loop.

Consider any active conical subdivision $\Lambda_{\mathbf{v}_{(j^*)}}(\mathbf{x})$ of $\mathbf{v}_{(j^*)}$ at $\mathbf{x}$. At each point in the algorithm after the $(j^*)^{\text{th}}$ iteration of the outermost `for`–loop, Corollary A.6.3 and Lemma A.6.4 imply that there is some $\sigma \in \Lambda_{\mathbf{v}_{(j^*)}}(\mathbf{x})$ such that $N := \operatorname{int}(\sigma) \cap \operatorname{int}(\operatorname{cone} Q)$ is nonempty. Moreover, by Lemma A.1.11, there is some $\mathbf{v}^*_{(j^*)} \in \mathcal{E}_{\mathbf{v}_{(j^*)}}(\mathbf{x})$ such that $\mathbf{v}_{(j^*)}{}'(\mathbf{x};\mathbf{d}) = \mathbf{J}\mathbf{v}^*_{(j^*)}(\mathbf{x})\,\mathbf{d}$ for each $\mathbf{d} \in \sigma$. Thus, by construction of N, (A.16) implies that

$$\mathbf{J}\boldsymbol{\psi}^*(\mathbf{u}_{(j^*)}(\mathbf{x}))[\mathbf{J}\mathbf{v}^*_{(i)}(\mathbf{x})]_{i\prec j^*}\mathbf{d} = \mathbf{J}\mathbf{v}^*_{(j^*)}(\mathbf{x})\,\mathbf{d}, \qquad \forall \mathbf{d} \in N. \tag{A.17}$$

Since N is nonempty, there exists some $\mathbf{d}^* \in N$. Since N is open, for some sufficiently small $\epsilon > 0$, $(\mathbf{d}^* + \epsilon\mathbf{e}_{(k)}) \in N$ for each $k \in \{1, \ldots, n\}$. Thus, (A.17) is satisfied when $\mathbf{d} = \mathbf{d}^*$, and when $\mathbf{d} = \mathbf{d}^* + \epsilon\mathbf{e}_{(k)}$ for any $k \in \{1, \ldots, n\}$. Since both sides of (A.17) are linear in $\mathbf{d}$, it follows that

$$\mathbf{J}\boldsymbol{\psi}^*(\mathbf{u}_{(j^*)}(\mathbf{x}))[\mathbf{J}\mathbf{v}^*_{(i)}(\mathbf{x})]_{i\prec j^*}\mathbf{e}_{(k)} = \mathbf{J}\mathbf{v}^*_{(j^*)}(\mathbf{x})\,\mathbf{e}_{(k)}, \qquad \forall k \in \{1, \ldots, n\},$$

and so

$$\mathbf{J}\boldsymbol{\psi}^*(\mathbf{u}_{(j^*)}(\mathbf{x}))[\mathbf{J}\mathbf{v}^*_{(i)}(\mathbf{x})]_{i\prec j^*} = \mathbf{J}\mathbf{v}^*_{(j^*)}(\mathbf{x}).$$

Substituting this result into (A.16) completes the inductive step. □

Bibliography

[1] C. S. Adjiman, S. Dallwig, C. A. Floudas, and A. Neumaier. A global optimization method, αBB, for general twice-differentiable constrained NLPs – I. Theoretical advances. *Computers Chem. Engng*, 22:1137–1158, 1998.

[2] Götz Alefeld and Günter Mayer. Interval analysis: theory and applications. *J. Comput. Appl. Math.*, 121:421–464, 2000.

[3] Jean-Pierre Aubin and Hélène Frankowska. *Set-Valued Analysis*. Modern Birkhäuser Classics. Birkhäuser Boston, Boston, MA, 2009.

[4] B. T. Baumrucker, J. G. Renfro, and L. T. Biegler. MPEC problem formulations and solution strategies with chemical engineering applications. *Comput. Chem. Eng.*, 32:2903–2913, 2008.

[5] Markus Beckers, Viktor Mosenkis, Michael Maier, and Uwe Naumann. Adjoint mode computation of subgradients for McCormick relaxations. Technical Report AIB-2011-10, RWTH Aachen, 2011.

[6] Brahim Benyahia, Richard Lakerveld, and Paul I. Barton. A plant-wide dynamic model of a continuous pharmaceutical process. *Ind. Eng. Chem. Res.*, 51:15393–15412, 2012.

[7] P. Beremlijski, J. Haslinger, M. Kočvara, R. Kučera, and J. V. Outrata. Shape optimization in three-dimensional contact problems with Coulomb friction. *SIAM J. Optim.*, 20(1):416–444, 2009.

[8] Dimitri P. Bertsekas. Nondifferentiable optimization via approximation. In M.L. Balinski and Philip Wolfe, editors, *Mathematical Programming Study 3*, pages 1–25. North-Holland Publishing Company, Amsterdam, 1975.

[9] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, second edition, 1999.

[10] Dimitris Bertsimas and John N. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific and Dynamic Ideas, LLC, Belmont, MA, 1997.

[11] A. Bompadre and A. Mitsos. Convergence rate of McCormick relaxations. *J. Glob. Optim.*, 52:1–28, 2012.

[12] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[13] Yang Cao, Shengtai Li, Linda Petzold, and Radu Serban. Adjoint sensitivity analysis for differential-algebraic equations: The adjoint DAE system and its numerical solution. *SIAM J. Sci. Comput.*, 24:1076–1089, 2003.

[14] Francois Edouard Cellier. *Combined continuous/discrete system simulation by use of digital computers: Techniques and tools*. PhD thesis, Swiss Federal Institute of Technology in Zurich, 1979.

[15] Benoit Chachuat. MC++: A toolkit for bounding factorable functions, v1.0, 2014. Retrieved online on July 2, 2014, from `https://projects.coin-or.org/MCpp`.

[16] Frank H. Clarke. *Optimization and Nonsmooth Analysis*. SIAM, Philadelphia, PA, 1990.

[17] Christian Clason, Bangti Jin, and Karl Kunisch. A semismooth Newton method for L^1 data fitting with automatic choice of regularization parameters and noise calibration. *SIAM J. Imaging Sci.*, 3(2):199–231, 2010.

[18] Earl A. Coddington and Norman Levinson. *Theory of Ordinary Differential Equations*. McGraw Hill Co., Inc., New York, NY, 1955.

[19] Steven P. Dirkse and Michael C. Ferris. The PATH solver: a non-monotone stabilization scheme for mixed complementarity problems. *Optim. Method. Softw.*, 5:123–156, 1995.

[20] Kaisheng Du and R. Baker Kearfott. The cluster problem in multivariate global optimization. *J. Global Optim.*, 5:253–265, 1994.

[21] Marco A. Duran and Ignacio E. Grossmann. Simultaneous optimization and heat integration of chemical processes. *AIChE Journal*, 32(1):123–138, 1986.

[22] Francisco Facchinei, Andreas Fischer, and Markus Herrich. An LP-Newton method: nonsmooth equations, KKT systems, and nonisolated solutions. *Math. Program., Ser. A*, 146:1–36, 2014.

[23] Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer-Verlag New York, Inc., New York, NY, 2003.

[24] William F. Feehery, John E. Tolsma, and Paul I. Barton. Efficient sensitivity analysis of large-scale differential-algebraic systems. *Appl. Numer. Meth.*, 25:41–54, 1997.

[25] Michael C. Ferris and Todd S. Munson. Interfaces to PATH 3.0: design, implementation and usage. *Comput. Optim. Appl.*, 12:207–227, 1999.

[26] Alekseĭ F. Filippov. *Differential Equations with Discontinuous Righthand Sides*. Kluwer Academic Publishers, Dordrecht, 1988.

[27] Bernard Friedland. *Control System Design: An Introduction to State-Space Methods*. McGraw-Hill, New York, 1986.

[28] Steven A. Gabriel and Jorge J. Moré. Smoothing of mixed complementarity problems. Preprint MCS-P541-0995, Argonne National Laboratory, 1995.

[29] Santos Galán and Paul I. Barton. Dynamic optimization of hybrid systems. *Comput. Chem. Eng.*, 22 (Suppl.):S183–S190, 1998.

[30] Santos Galán, William F. Feehery, and Paul I. Barton. Parametric sensitivity functions for hybrid discrete/continuous systems. *Appl. Numer. Math.*, 31:17–47, 1999.

[31] A. Griewank and P. J. Rabier. On the smoothness of convex envelopes. *T. Am. Math. Soc.*, 322:691–709, 1990.

[32] Andreas Griewank. Automatic directional differentiation of nonsmooth composite functions. In *Recent Developments in Optimization, French-German Conference on Optimization*, Dijon, 1994.

[33] Andreas Griewank. On stable piecewise linearization and generalized algorithmic differentiation. *Optim. Method. Softw.*, 28(6):1139–1178, 2013.

[34] Andreas Griewank and Andrea Walther. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Other Titles in Applied Mathematics. SIAM, Philadelphia, PA, 2nd edition, 2008.

[35] Philip Hartman. *Ordinary Differential Equations*. SIAM, Philadelphia, PA, second edition, 2002.

[36] Stuart M. Harwood, Kai Höffner, and Paul I. Barton. Solution of ordinary differential equations with a linear program embedded: the right-hand side case. *Submitted*, 2013.

[37] Stuart M. Harwood, Joseph K. Scott, and Paul I. Barton. Bounds on Reachable Sets Using Ordinary Differential Equations with Linear Programs Embedded. In Sophie Tarbouriech and Miroslav Krstic, editors, *Nonlinear Control Systems*, volume 9, pages 62–67, 2013.

[38] Jaroslav Haslinger, Jiří V. Outrata, and Róbert Pathó. Shape optimization in 2D contact problems with given friction and a solution-dependent coefficient of friction. *Set-Valued Anal.*, 20:31–59, 2012.

[39] M. Hintermüller and M. Hinze. A SQP-semismooth Newton-type algorithm applied to control of the instationary Navier-Stokes system subject to control constraints. *SIAM J. Optim.*, 16(4):1177–1200, 2006.

[40] M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM J. Optim.*, 13(3):865–888, 2003.

[41] J.-B. Hiriart-Urruty. Characterizations of the plenary hull of the generalized Jacobian matrix. *Math. Program. Stud.*, 17:1–12, 1982.

[42] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms I*. Springer-Verlag, Berlin, 2003.

[43] J. L. Hjersted and M. A. Henson. Steady-state and dynamic flux balance analysis of ethanol production by *saccharomyces cerevisiae*. *IET Syst. Biol.*, 3:167–179, 2009.

[44] Kai Höffner, Stuart M. Harwood, and Paul I. Barton. A reliable simulator for dynamic flux balance analysis. *Biotechnol. Bioeng.*, 110:792–802, 2013.

[45] Reiner Horst and Hoang Tuy. *Global Optimization: Deterministic Approaches*. Springer-Verlag, Berlin, second edition, 1993.

[46] C. Imbert. Support functions of the Clarke generalized Jacobian and of its plenary hull. *Nonlinear Anal.-Theor.*, 49:1111–1125, 2002.

[47] M. A. Jenkins and J. F. Traub. A three-stage algorithm for real polynomials using quadratic iteration. *SIAM J. Numer. Anal.*, 7:545–566, 1970.

[48] Karl Henrik Johansson, Magnus Egerstedt, John Lygeros, and Shakar Sastry. On the regularization of Zeno hybrid automata. *Syst. Control Lett.*, 38:141–150, 1999.

[49] Ravindra S. Kamath, Lorenz T. Biegler, and Ignacio E. Grossmann. Modeling multistream heat exchangers with and without phase changes for simultaneous optimization and heat integration. *AIChE Journal*, 58:190–204, 2012.

[50] Napsu Karmitsa, Mario Tanaka Filho, and José Herskovits. Globally convergent cutting plane method for nonconvex nonsmooth minimization. *J. Optim. Theory Appl.*, 148:528–549, 2011.

[51] Yoshiaki Kawajiri and Lorenz T. Biegler. Nonlinear programming superstructure for optimal dynamic operations of simulated moving bed processes. *Ind. Eng. Chem. Res.*, 45:8503–8513, 2006.

[52] Padmanaban Kesavan, Russell J. Allgor, Edward P. Gatzke, and Paul I. Barton. Outer approximation algorithms for separable nonconvex mixed-integer nonlinear programs. *Math. Program., Ser. A*, 100:517–535, 2004.

[53] Kamil A. Khan and Paul I. Barton. Evaluating an element of the Clarke generalized Jacobian of a piecewise differentiable function. In Shaun Forth, Paul Hovland, Eric Phipps, Jean Utke, and Andrea Walther, editors, *Recent Advances in Algorithmic Differentiation*, pages 115–125. Springer-Verlag, Berlin, 2012.

[54] Kamil A. Khan and Paul I. Barton. Evaluating an element of the Clarke generalized Jacobian of a composite piecewise differentiable function. *ACM T. Math. Software*, 39(4):23:1–23:28, 2013.

[55] Kamil A. Khan and Paul I. Barton. Generalized derivatives for solutions of parametric ordinary differential equations with non-differentiable right-hand sides. *J. Optimiz. Theory App.*, 163:355–386, 2014.

[56] Kamil A. Khan and Paul I. Barton. Generalized derivatives of hybrid systems. *In preparation*, 2014.

[57] Kamil A. Khan and Paul I. Barton. Generalized gradient elements for nonsmooth optimal control problems. In *Proceedings of the 53rd IEEE Conference on Decision and Control*, Los Angeles, 2014.

[58] Kamil A. Khan and Paul I. Barton. A numerical method for evaluating generalized derivatives for nonsmooth parametric ordinary differential equations. *In preparation*, 2014.

[59] Kamil A. Khan and Paul I. Barton. Switching behavior of solutions of ordinary differential equations with nonsmooth right-hand sides. *Under review*, 2014.

[60] Kamil A. Khan and Paul I. Barton. A twice-continuously differentiable version of McCormick's relaxations. *In preparation*, 2014.

[61] Kamil A. Khan and Paul I. Barton. A vector forward mode of automatic differentiation for generalized derivative evaluation. *Under review*, 2014.

[62] Donald E. Kirk. *Optimal Control Theory: An Introduction*. Dover Publications, Mineola, NY, 2004.

[63] Krzysztof C. Kiwiel. *Methods of Descent for Nondifferentiable Optimization*. Lecture Notes in Mathematics. Springer-Verlag, Berlin, 1985.

[64] Diethard Klatte and Bernd Kummer. *Nonsmooth Equations in Optimization*. Nonconvex Optimization and Its Applications. Springer, Dordrecht, 2002.

[65] Masakazu Kojima and Susumu Shindo. Extension of Newton and quasi-Newton methods to systems of PC^1 equations. *J. Oper. Res. Soc. Jpn*, 29(4):352–375, 1986.

[66] Steven G. Krantz and Harold R. Parks. *A Primer of Real Analytic Functions*. Birkhäuser Advanced Texts. Birkhäuser, Boston, MA, second edition, 2002.

[67] Claude Lemaréchal, Jean-Jacques Strodiot, and André Bihain. On a bundle algorithm for nonsmooth optimization. In Olvi L. Mangasarian, Robert R. Meyer, and Stephen M. Robinson, editors, *Nonlinear Programming 4*, New York, NY, 1981. Academic Press.

[68] Xiang Li, Asgeir Tomasgard, and Paul I. Barton. Nonconvex generalized Benders decomposition for stochastic separable mixed-integer nonlinear programs. *J. Optim. Theory Appl.*, 151:425–454, 2011.

[69] Leo Liberti and Constantinos C. Pantelides. Convex envelopes of monomials of odd degree. *J. Glob. Optim.*, 25:157–168, 2003.

[70] Ladislav Lukšan and Jan Vlček. A bundle-Newton method for nonsmooth unconstrained minimization. *Math. Program.*, 83:373–391, 1998.

[71] Ladislav Lukšan and Jan Vlček. Algorithm 811: NDA: Algorithms for nondifferentiable optimization. *ACM T. Math. Software*, 27(2):193–213, 2001.

[72] John Lygeros, Karl Henrik Johansson, Shankar Sastry, and Magnus Egerstedt. On the existence of executions of hybrid automata. *IEEE Conf. Decision Control*, pages 2249–2254, 1999.

[73] Timothy Maly and Linda R. Petzold. Numerical methods and software for sensitivity analysis of differential-algebraic systems. *Appl. Numer. Meth.*, 20:57–79, 1996.

[74] G. P. McCormick. Computability of global solutions to factorable nonconvex programs: Part I - Convex underestimating problems. *Math. Program.*, 10:147–175, 1976.

[75] Robert Mifflin. Semismooth and semiconvex functions in constrained optimization. *SIAM J. Control Optim.*, 15(6):959–972, 1977.

[76] Alexander Mitsos, Benoit Chachuat, and Paul I. Barton. McCormick-based relaxations of algorithms. *SIAM J. Optim.*, 20:573–601, 2009.

[77] Ramon E. Moore. *Methods and Applications of Interval Analysis*. SIAM, Philadelphia, 1979.

[78] Boris S. Mordukhovich. *Variational analysis and generalized differentiation I: Basic theory*. Springer, Berlin, 2006.

[79] Yurii Nesterov. Lexicographic differentiation of nonsmooth functions. *Math. Program. B*, 104:669–700, 2005.

[80] Arnold Neumaier. *Interval methods for systems of equations*. Cambridge University Press, Cambridge, 1990.

[81] J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonliner equations in several variables*. Academic Press, Inc., San Diego, CA, 1970.

[82] Zsolt Páles and Vera Zeidan. Infinite dimensional Clarke generalized Jacobian. *J. Convex Anal.*, 14:433–454, 2007.

[83] Zsolt Páles and Vera Zeidan. Infinite dimensional generalized Jacobian: Properties and calculus rules. *J. Math. Anal. Appl.*, 344:55–75, 2008.

[84] Jong-Shi Pang and Steven A. Gabriel. NE/SQP: A robust algorithm for the nonlinear complementarity problem. *Math. Program.*, 60:295–337, 1993.

[85] Jong-Shi Pang and Daniel Ralph. Piecewise smoothness, local invertibility, and parametric analysis of normal maps. *Math. Oper. Res.*, 21:401–426, 1996.

[86] Jong-Shi Pang and Jinglai Shen. Strongly regular differential variational systems. *IEEE T. Automat. Contr.*, 52:242–255, 2007.

[87] Jong-Shi Pang and David E. Stewart. Differential variational inequalities. *Math. Program. A*, 113:345–424, 2008.

[88] Jong-Shi Pang and David E. Stewart. Solution dependence on initial conditions in differential variational inequalities. *Math. Program. B*, 116:429–460, 2009.

[89] Taeshin Park and Paul I. Barton. State event location in differential-algebraic models. *ACM T. Model. Comput. S.*, 6:137–165, 1996.

[90] L. Qi and D. Sun. Smoothing functions and smoothing Newton method for complementarity and variational inequality problems. *J. Optim. Theory App.*, 113:121–147, 2002.

[91] Liqun Qi. Convergence analysis of some algorithms for solving nonsmooth equations. *Math. Oper. Res.*, 18(1):227–244, 1993.

[92] Liqun Qi and Jie Sun. A nonsmooth version of Newton's method. *Math. Program.*, 58:353–367, 1993.

[93] Daniel Ralph. Global convergence of damped Newton's method for nonsmooth equations via the path search. *Math. Oper. Res.*, 19(2):352–389, 1994.

[94] Daniel Ralph and Stefan Scholtes. Sensitivity analysis of composite piecewise smooth equations. *Math. Program.*, 76:593–612, 1997.

[95] Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, Inc., New York, NY, third edition, 1976.

[96] Spencer D. Schaber. *Tools for dynamic model development*. PhD thesis, Massachusetts Institute of Technology, 2014.

[97] Stefan Scholtes. *Introduction to piecewise differentiable equations*. SpringerBriefs in Optimization. Springer, New York, NY, 2012.

[98] Daniel Scholz. Theoretical rate of convergence for interval inclusion functions. *J. Glob. Optim.*, 53:749–767, 2012.

[99] J. M. Schumacher. Complementarity systems in optimization. *Math. Program. B*, 101:263–295, 2004.

[100] Joseph K. Scott. *Reachability analysis and deterministic global optimization of differential-algebraic systems*. PhD thesis, Massachusetts Institute of Technology, 2012.

[101] Joseph K. Scott and Paul I. Barton. Convex and concave relaxations for the parametric solutions of semi-explicit index-one differential-algebraic equations. *J. Optim. Theory App.*, 156:617–649, 2013.

[102] Joseph K. Scott and Paul I. Barton. Improved relaxations for the parametric solutions of ODEs using differential inequalities. *J. Global Optim.*, 125:27–60, 2013.

[103] Joseph K. Scott, Benoit Chachuat, and Paul I. Barton. Nonlinear convex and concave relaxations for the solutions for parametric ODEs. *Optim. Control Appl. Meth.*, 34:145–163, 2013.

[104] Joseph K. Scott, Matthew D. Stuber, and Paul I. Barton. Generalized McCormick relaxations. *J. Glob. Optim.*, 51:569–606, 2011.

[105] Oliver J. Smith, IV and Arthur W. Westerberg. Acceleration of cyclic steady state convergence for pressure swing adsorption models. *Ind. Eng. Chem. Res.*, 31:1569–1573, 1992.

[106] Georg Stadler. Elliptic optimal control problems with L^1 control cost and applications for the placement of control devices. *Comput. Optim. Appl.*, 44(2):159–181, 2009.

[107] Matthew D. Stuber, Joseph K. Scott, and Paul I. Barton. Convex and concave relaxations of implicit functions. *Optim. Method Softw.*, In press, 2014.

[108] Héctor J. Sussmann. Bounds on the number of switchings for trajectories of piecewise analytic vector fields. *J. Differ. Equations*, 43:399–418, 1982.

[109] T. H. Sweetser, III. A minimal set-valued strong derivative for vector-valued Lipschitz functions. *J. Optimiz. Theory App.*, 23:549–562, 1977.

[110] Le Quang Thuan. Non-Zenoness of piecewise affine dynamical systems and affine complementarity systems with inputs. *Control Theory Tech.*, 12:35–47, 2014.

[111] Le Quang Thuan and M. Kanat Camlibel. Continuous piecewise affine dynamical systems do not exhibit Zeno behavior. *IEEE T. Automat. Contr.*, 56:1932–1936, 2011.

[112] John E. Tolsma. *DSL48SE Manual (Version 1.0): Hybrid Discrete/Continuous Numerical Integration and Parametric Sensitivity Analysis*, 2001.

[113] John E. Tolsma and Paul I. Barton. DAEPACK: An open modeling environment for legacy models. *Ind. Eng. Chem. Res.*, 39:1826–1839, 2000.

[114] A. Tsoukalas and A. Mitsos. Multivariate McCormick relaxations. *J. Glob. Optim.*, 59:633–662, 2014.

[115] Michael Ulbrich. *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*. MOS-SIAM Series on Optimization. SIAM, Philadelphia, PA, 2011.

[116] A. J. van der Schaft and J. M. Schumacher. The complementary-slackness class of hybrid systems. *Math. Control Signals Systems*, 9:266–301, 1996.

[117] Joerg Walter, Mathias Koch, Gunter Winkler, and David Bellot. BOOST basic linear algebra library, 2010. Retrieved online on October 30, 2013, from `http://www.boost.org/doc/libs/1_54_0/libs/numeric/ublas/doc/index.htm`.

[118] Achim Wechsung. *Global optimization in reduced space*. PhD thesis, Massachusetts Institute of Technology, 2014.

[119] Achim Wechsung, Spencer D. Schaber, and Paul I. Barton. The cluster problem revisited. *J. Glob. Optim.*, 58:429–438, 2014.

[120] Achim Wechsung, Joseph K. Scott, Harry A. J. Watson, and Paul I. Barton. Reverse propagation of McCormick relaxations. *Submitted*, 2014.

[121] Hassler Whitney. Analytic extensions of differentiable functions defined in closed sets. *Trans. Amer. Math. Soc.*, 36:63–89, 1934.

[122] D. Willett and J. S. W. Wong. On the discrete analogues of some generalizations of Gronwall's inequality. *Monatsh. Math.*, 69:362–367, 1965.

[123] H. Xu. Set-valued approximations and Newton's methods. *Math. Program.*, 84:401–420, 1999.

[124] Mehmet Yunt. *Nonsmooth Dynamic Optimization of Systems with Varying Structure*. PhD thesis, Massachusetts Institute of Technology, 2011.

[125] Mehmet Yunt, Kamil A. Khan, and Paul I. Barton. Parametric sensitivity analysis of dynamic systems using nonsmooth analysis: I. Theory. *In preparation*, 2014.

[126] Günter M. Ziegler. *Lectures on Polytopes*. Springer-Verlag, New York, NY, 1998.