Winter 12-17-2019

Theory of Lexicographic Differentiation in the Banach Space Setting

Jaeho Choi
University of Maine, jaeho.choi@maine.edu

THEORY OF LEXICOGRAPHIC DIFFERENTIATION IN THE BANACH SPACE SETTING

By

Jaeho Choi

B.A. Williams College, 2017

A THESIS

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Arts in Mathematics

(in The Department of Mathematics and Statistics)

The Graduate School

The University of Maine

December 2019

Advisory Committee:

Peter Stechlinski, Assistant Professor of Mathematics, Advisor

Andrew Knightly, Professor of Mathematics

David Bradley, Associate Professor of Mathematics

THEORY OF LEXICOGRAPHIC DIFFERENTIATION IN THE BANACH SPACE SETTING

By Jaeho Choi

Thesis Advisor: Professor Peter Stechlinski

An Abstract of the Thesis Presented
in Partial Fulfillment of the Requirements for the
Degree of Master of Arts in Mathematics
(in The Department of Mathematics and Statistics)
December 2019

Derivative information is useful for many problems found in science and engineering that require equation solving or optimization. Driven by its utility and mathematical curiosity, researchers over the years have developed a variety of generalized derivatives. In this thesis, we will first take a look at Clarke's generalized derivative for locally Lipschitz continuous functions between Euclidean spaces, which roughly is the smallest convex set containing all nearby derivatives of a domain point of interest. Clarke's generalized derivative in this setting possesses a strong theoretical and numerical toolkit, which is analogous to that of the classical derivative. It includes nonsmooth versions of the chain rule, the mean value theorem, and the implicit function theorem, as well as nonsmooth equation-solving and optimization methods. However, it is generally difficult to obtain elements of Clarke's generalized derivative in the Euclidean space setting. To address this issue, we use lexicographic differentiation by Nesterov and lexicographic directional differentiation by Khan and Barton. They are generalized derivatives theories for a subclass of locally Lipschitz continuous functions, called the class of lexicographically smooth functions, which help to find elements of Clarke's generalized derivative in the Euclidean space setting systematically. Lexicographic derivatives are either elements of Clarke's generalized derivative in the Euclidean space setting or at least indistinguishable

from them as far as numerical tools are concerned. We outline a process by which we can find a lexicographic derivative once a lexicographic directional derivative is known. Lastly, we present lexicographic differentiation theory for a subclass of locally Lipschitz continuous functions mapping between Banach spaces that have Schauder bases, called, unsurprisingly, the class of lexicographically smooth functions. We provide a proof for Nesterov's result that, as in the Euclidean space setting, lexicographic derivatives in this setting satisfy a sharp calculus rule.

DEDICATION

To my family

ACKNOWLEDGEMENTS

I would like to thank all my family, friends, and mentors whose support has made this thesis possible. In particular, I would like to express my gratitude to my advisor, Professor Peter Stechlinski, for his guidance, encouragement, and patience.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1
INTRODUCTION

Differentiation is a key concept in mathematics that finds a wide range of applications in fundamental problems in science and engineering such as equation-solving and optimization. One classic example of equation-solving is Newton's method, an algorithm that uses derivative information of a function to find its roots. Given a function $f : X \subset \mathbb{R} \to \mathbb{R}$, where X is an open subset of $\mathbb{R}$, and an initial guess x_0 for a root $x \in X$ of f, Newton's method provides a recursive formula for subsequent approximations to x, i.e.,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n \in \{0, 1, 2, \ldots\}.$$

If x_0 is "sufficiently" close to x and there exists a closed neighborhood of x containing x_0 in which f is twice continuously differentiable and f' does not attain zero, then Newton's method guarantees that the sequence $\{x_n\}$ of approximations converges to x, at a rate such that

$$|x_{n+1} - x| \leq M \, |x_n - x|^2$$

for some $M \geq 0$. Simply put, Newton's method works provided that x_0 is "sufficiently" close to x and f is "well-behaved" around x. For example, consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^3$. For any nonzero initial guess $x_0 \in \mathbb{R}$, the subsequent approximations to the root 0 are given recursively by

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{(x_n)^3}{3\,(x_n)^2} = \frac{2}{3} x_n, \quad n \in \{0, 1, 2, \ldots\}.$$

Hence,

$$x_n = \left(\frac{2}{3}\right)^n x_0.$$

Therefore, the sequence $\{x_n\}$ of approximations converges to 0. In this example, any nonzero initial guess x_0 is "sufficiently" close to 0. However, this is not the case in general.

Consider the function $g : \mathbb{R} \to \mathbb{R}$ defined by $g(x) = x^{\frac{1}{3}}$. For any nonzero initial guess $x_0 \in \mathbb{R}$, the subsequent approximations to the root 0 are given recursively by

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{(x_n)^{\frac{1}{3}}}{\frac{1}{3}(x_n)^{-\frac{2}{3}}} = -2x_n, \quad n \in \{0, 1, 2, \dots\}.$$

Hence,

$$x_n = (-2)^n x_0.$$

Therefore, no matter how close x_0 is to 0, the sequence $\{x_n\}$ of approximations does not converge to 0. The complete failure of Newton's method in this example is because the function g is not “well-behaved" around 0, as the derivative of g does not exist at 0.

There are functions for which Newton's method fails. Take for example the function $h : \mathbb{R} \to \mathbb{R}$ defined by

$$h(x) = \begin{cases} |x| & \text{if } |x| < 1, \\ x^2 & \text{if } |x| \geq 1. \end{cases}$$

Suppose that we want to use Newton's method to find the root 0 of h with the initial guess $x_0 = 1$. Newton's method cannot compute x_1 since the derivative of h is not defined at x_0. This example necessitates a generalization of Newton's method to a larger class of functions that contains “nonsmooth" functions such as h. To that end, we must first generalize the notion of a derivative for a function. Over the years, researchers have extended the classical derivative to nonsmooth functions in a number of ways, motivated either by mathematical curiosity or by problems arising in science and engineering. These generalized notions of the classical derivative differ from one another in many regards, such as the class of functions for which they apply and the kind and purpose of derivative information that they provide. In general, the broader the class of functions for which a theory of generalized differentiation applies, the less useful the derivative information. In this thesis, we will focus on Clarke's theory of generalized differentiation [3], which is applicable to the class of locally Lipschitz continuous functions. Roughly speaking, the Clarke generalized derivative at a point in the domain of a locally Lipschitz continuous function is the smallest convex set containing “nearby derivatives." For example, let us recall the nonsmooth function h, introduced above. Since $h'(x) = 1$ for $x \in (-1, 1)$ and $h'(x) = 2x$ for $x \in (1, \infty)$, the Clarke generalized derivative of h at $x = 1$ is the convex hull of the set $\{1, 2(1)\} = \{1, 2\}$, i.e., $[1, 2]$, and is denoted $\partial h(1)$. Clarke generalized derivatives are just what we need to generalize Newton's method for locally Lipschitz continuous functions. Suppose that we want to find the root 0 of h with the initial guess $x_0 = 1$ using a generalized Newton's method in which the “derivative" of h at x_0 is set to be any number in $\partial h(1)$, i.e., $[1, 2]$. Then the first approximation is given by

$$x_1 = x_0 - \frac{h(x_0)}{a}, \quad a \in [1, 2] = \partial h(x_0).$$

Since $x_1 \in [0, \frac{1}{2}]$ and $h'(x) = 1$ for all $x \in (-1, 1)$, the Clarke generalized derivative of h at x_1, denoted $\partial h(x_1)$, is the smallest convex set containing 1, i.e., $\{1\}$. Then the second approximation is given by

$$x_2 = x_1 - \frac{h(x_1)}{b}, \quad b \in \{1\} = \partial h(x_1),$$

i.e., $x_2 = 0$. Hence, regardless of the choice of the number $a \in \partial h(1)$, the sequence $\{x_n\}$ of approximations does converge to the root, which demonstrates the power of Clarke's theory of generalized differentiation in equation-solving.

This thesis will concentrate on Clarke's theory of generalized differentiation for two main reasons. First of all, it is designed for the class of locally Lipschitz continuous functions, which contains many functions commonly encountered in problems arising in science and engineering. In addition, Clarke's theory of generalized differentiation possesses a strong theoretical toolkit that contains powerful theorems, such as a mean value theorem and inverse and implicit function theorems [3], and a strong numerical toolkit that includes techniques useful in nonsmooth equation-solving and optimization methods, such as the generalized Newton's method described above [10]. Unfortunately, it is quite challenging to compute Clarke generalized derivative elements in general for a number of reasons. First of all, while classical calculus rules are equality-based, the calculus rules for Clarke generalized derivatives are inclusion-based. For example, consider the functions $A : \mathbb{R} \to \mathbb{R}$ defined by $A(x) = \max(0, x)$ and $B : \mathbb{R} \to \mathbb{R}$ defined by $B(x) = \min(0, x)$. Then the sum of the two functions $A + B : \mathbb{R} \to \mathbb{R}$ is the identity map, i.e., $[A + B](x) = x$. Being the smallest convex sets containing “nearby derivatives," the Clarke generalized derivatives of A, B, and their sum at the origin are

$$\partial A(0) = [0, 1]$$

$$\partial B(0) = [0, 1]$$

$$\partial [A + B](0) = \{1\}$$

respectively. Hence, $\partial [A + B](0) \subset \partial A(0) + \partial B(0)$, where the inclusion is proper. Therefore, it is not always possible to obtain an element of the Clarke generalized derivative of $A + B$ by adding together elements of the Clarke generalized derivative of A and B. Secondly, it is generally not possible to obtain a Clarke generalized derivative element by “stitching" together directional derivatives in coordinate directions, which are more straightforward to compute. Lastly, in case of vector-valued functions, it is generally not possible to obtain a Clarke generalized derivative element by calculating the corresponding Clarke generalized derivative elements of the component functions.

Recently, two theories have been developed for a subclass of locally Lipschitz continuous functions to address these challenges: lexicographic differentiation by Nesterov [9] and lexicographic directional differentiation by Khan and Barton [6]. This subclass consisting of lexicographically smooth (L-smooth) functions, is a broad class of functions containing commonly encountered functions in problems that require numerical treatment, such as convex functions, piecewise differentiable functions [11], and any compositions of L-smooth functions [9, 6]. In the context of many numerical methods, lexicographic derivatives are indistinguishable from Clarke generalized derivative elements, but can be computed in an efficient, accurate, and automated way, unlike Clarke generalized derivative elements. This advantage of lexicographic derivatives comes from the fact that lexicographic directional derivatives are “well-behaved" in that they can be calculated component-wise and satisfy equality-based calculus rules. Lexicographic differentiation and lexicographic directional differentiation are both theories with powerful theoretical toolkits containing tools such as implicit and inverse function theorems. Since their development, lexicographic differentiation and lexicographic directional differentiation have been successfully applied to finite-dimensional problems arising in nonsmooth optimization [13] and nonsmooth dynamic systems [5, 12] (see [1] for more details).

For some subclasses of L-smooth functions, lexicographic derivatives are indistinguishable not only from Clarke generalized derivative elements, but also from other generalized derivative elements. For example, if f is piecewise differentiable [11], then the lexicographic derivative of f is an element of the Bouligand subdifferential [3] and an element of the Mordukhovich subdifferential [7]. Just as Clarke's generalized derivative is developed for locally Lipschitz continuous functions, a number of generalized derivatives, such as Dini's derivative, are developed for discontinuous functions. However, this thesis will not expand on such generalized derivatives, as they are not suitable for the equation-solving and optimization problems of our interest.

In Chapter 2, we introduce the classical theory of differentiation and present Newton's method as an example of a numerical tool for which the classical derivative is useful. In Chapter 3, we present Clarke's theory of differentiation as a generalized theory of differentiation for the class of locally Lipschitz continuous functions mapping between Euclidean spaces. We highlight the important subclass consisting of piecewise differentiable functions, for which Clarke's theory of differentiation is particularly simple to formulate. We present the theoretical toolkit of Clarke's generalized derivative object, which mirrors that of the classical derivative. Moreover, we highlight a Newton's method for the class of locally Lipschitz continuous functions to illustrate a numerical tool to which Clarke's theory of differentiation applies.

Despite the utility of Clarke's theory, it is difficult to obtain elements of Clarke's generalized derivative object, even for functions that are constructed from functions whose

Clarke's derivative objects are known, since they obey inclusion-based calculus rules. In Chapter 4, we will address this issue by presenting Nesterov's theory of lexicographic differentiation and Khan and Barton's theory of lexicographic directional differentiation. These generalized differentiation theories for a subclass of locally Lipschitz continuous functions, called the class of L-smooth functions, provide a systematic way to find elements of Clarke's derivative object. First, we introduce these theories for the class of L-smooth functions mapping between Euclidean spaces. We highlight that lexicographic derivatives are at best elements of Clarke's derivative object and at least indistinguishable from elements of Clarke's derivative object in nonsmooth numerical tools. We note that unlike Clarke's generalized derivative object, L-derivatives satisfy an equality-based chain rule. To make it easier to find L-derivatives, we present lexicographic directional derivatives for L-smooth functions. Unlike elements of Clarke's generalized derivative object, it is straightforward to find LD-derivatives of elementary functions and LD-derivatives of functions constructed from functions whose LD-derivatives are known. An L-derivative is easy to obtain once an LD-derivative is known. Next, we introduce Nesterov's lexicographic differentiation for a subclass of locally Lipschitz continuous functions mapping between Banach spaces with Schauder bases, called, as expected, the class of L-smooth functions. We state L-derivatives for such functions and provide a proof for Nesterov's result that these L-derivatives obey an equality-based chain rule, just like the L-derivatives in the Euclidean space setting. For illustration, we present examples of L-smooth functions, find their L-derivatives, and show that they indeed obey the sharp chain rule. In Chapter 5, we give conclusions and discuss future work.

The contributions made in this thesis include elaboration of existing proofs wherever deemed appropriate, a proof for Theorem 4.3.5, and provision of new illustrative examples in Section 4.3.5.

CHAPTER 2
PRELIMINARIES AND BACKGROUND

In this chapter, we review the classical theory of differentiation.

2.1 Notation

In this chapter, we use X and Y to denote open subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively.

2.2 Vector Spaces and Linear Maps

Definition 2.2.1. *Let $\mathbb{F}$ be a field and V a set equipped with two operations: addition and scalar multiplication. Addition is a function $+ : V \times V \to V$ mapping $(\boldsymbol{v}, \boldsymbol{w})$ to $\boldsymbol{v} + \boldsymbol{w}$. Scalar multiplication is a function $\cdot : \mathbb{F} \times V \to V$ mapping $(\alpha, \boldsymbol{v})$ to $\alpha\boldsymbol{v}$. If V satisfies the following axioms, then V is called a vector space over $\mathbb{F}$:*

For any vectors $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in V$ and any scalars $a, b \in \mathbb{F}$,

1. $\boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w}) = (\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w}$.

2. $\boldsymbol{u} + \boldsymbol{v} = \boldsymbol{v} + \boldsymbol{u}$.

3. *There exists an element $\boldsymbol{0} \in V$, called a zero vector, such that $\boldsymbol{v} + \boldsymbol{0} = \boldsymbol{v}$ for all $\boldsymbol{v} \in V$.*

4. *For every $\boldsymbol{v} \in V$, there exists an element $-\boldsymbol{v} \in V$, called an additive inverse of $\boldsymbol{v}$, such that $\boldsymbol{v} + (-\boldsymbol{v}) = \boldsymbol{0}$.*

5. $a(b\boldsymbol{v}) = (ab)\boldsymbol{v}$.

6. $1\boldsymbol{v} = \boldsymbol{v}$*, where 1 denotes the multiplicative identity in $\mathbb{F}$.*

7. $a(\boldsymbol{u} + \boldsymbol{v}) = a\boldsymbol{u} + a\boldsymbol{v}$.

8. $(a + b)\boldsymbol{v} = a\boldsymbol{v} + b\boldsymbol{v}$.

Elements of V are called vectors and elements of $\mathbb{F}$ are called scalars.

Definition 2.2.2. *A normed vector space is a vector space V over a field $\mathbb{F}$ $(=\mathbb{R}$ or $\mathbb{C})$, equipped with a function $\|\cdot\| : \boldsymbol{V} \to \mathbb{R}$, called a norm on V, that satisfies the following properties:*

1. $\|\boldsymbol{x}\| \geq 0$ *for any* $\boldsymbol{x} \in V$, *and* $\|\boldsymbol{x}\| = 0$ *if and only if* $\boldsymbol{x} = \mathbf{0}$.

2. $\|a\boldsymbol{x}\| = |a| \, \|\boldsymbol{x}\|$ *for any* $a \in \mathbb{F}$ *and any* $\boldsymbol{x} \in V$.

3. $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ *for any* $\boldsymbol{x}, \boldsymbol{y} \in V$.

Definition 2.2.3. *A normed vector space V over a field $\mathbb{F}$ $(=\mathbb{R}$ or $\mathbb{C})$ is complete if every Cauchy sequence of elements of V converges to an element of V with respect to the norm. A complete normed vector space is called a Banach space.*

Remark 2.2.4. *In this thesis, we will always view $\mathbb{R}^n$ as a Banach space with the Euclidean norm.*

Definition 2.2.5. *Let V be a Banach space over a field $\mathbb{F}$ $(=\mathbb{R}$ or $\mathbb{C})$. A Schauder basis is a sequence $\{\boldsymbol{b}_n\}$ of elements of V such that for every element $\boldsymbol{v} \in V$ there exists a unique sequence $\{a_n\}$ of scalars in $\mathbb{F}$ for which*

$$\boldsymbol{v} = \sum_{n=0}^{\infty} a_n \boldsymbol{b}_n,$$

where the convergence is with respect to the norm.

Definition 2.2.6. *Let V and W be vector spaces over a field $\mathbb{F}$. A function $\boldsymbol{f} : V \to W$ is a linear map if for any vectors $\boldsymbol{u}, \boldsymbol{v} \in V$ and any scalar $c \in \mathbb{F}$,*

$$\boldsymbol{f}(\boldsymbol{u} + \boldsymbol{v}) = \boldsymbol{f}(\boldsymbol{u}) + \boldsymbol{f}(\boldsymbol{v})$$

$$\boldsymbol{f}(c\boldsymbol{u}) = c\boldsymbol{f}(\boldsymbol{u}).$$

If $V = W$, then a linear map is called a linear operator.

Definition 2.2.7. *Let V and W be normed vector spaces over a field $\mathbb{F}$ $(=\mathbb{R}$ or $\mathbb{C})$, where $\|\cdot\|_V$ is the norm on V and $\|\cdot\|_W$ is the norm on W. A linear map $\boldsymbol{f}: V \to W$ is bounded if there exists some $M \geq 0$ such that for all $\boldsymbol{v} \in V$,*

$$\|\boldsymbol{f}(\boldsymbol{v})\|_W \leq M \|\boldsymbol{v}\|_V .$$

Remark 2.2.8. *Let V and W be normed vector spaces over a field $\mathbb{F}$ $(=\mathbb{R}$ or $\mathbb{C})$, where $\|\cdot\|_V$ is the norm on V and $\|\cdot\|_W$ is the norm on W. The set of bounded linear maps from V to W equipped with the two operations*

$$(\boldsymbol{A}+\boldsymbol{B})(\boldsymbol{v}) = \boldsymbol{A}(\boldsymbol{v}) + \boldsymbol{B}(\boldsymbol{v}),\ \boldsymbol{v} \in V$$

$$(\alpha \boldsymbol{A})(\boldsymbol{v}) = \alpha \boldsymbol{A}(\boldsymbol{v}),\ \boldsymbol{v} \in V$$

for any bounded linear maps $\boldsymbol{A}$ and $\boldsymbol{B}$ and any $\alpha \in \mathbb{F}(=\mathbb{R}$ or $\mathbb{C})$ forms a vector space denoted by $\mathcal{C}(V,W)$. This notation is justified by the fact that a linear map $\boldsymbol{A}: V \to W$ is bounded if and only if it is continuous with respect to topologies induced from the norms $\|\cdot\|_V$ and $\|\cdot\|_W$ of V and W, respectively. The vector space $\mathcal{C}(V,W)$ equipped with the norm $\|\cdot\| : \mathcal{C}(V,W) \to \mathbb{R}$ defined by

$$\|\boldsymbol{A}\| = \inf\{M \geq 0 : \|\boldsymbol{A}\boldsymbol{v}\|_W \leq M \|\boldsymbol{v}\|_V \ \text{for all } \boldsymbol{v} \in V\}$$

forms a normed vector space. The norm $\|\cdot\|$ is called the operator norm.

2.3 Theory of Differentiation

We introduce differentiation notions that are commonly encountered in the Euclidean space setting.

Definition 2.3.1. *Let $f : X \subset \mathbb{R}^n \to \mathbb{R}$ be a function. The partial derivative of f at $\boldsymbol{a} = (a_1 \dots a_n)^T \in X$ with respect to the i-th variable x_i is*

$$\frac{\partial f}{\partial x_i}(\boldsymbol{a}) := \lim_{h \to 0} \frac{f(a_1, \dots, a_{i-1}, a_i + h, a_{i+1}, \dots, a_n) - f(a_1, \dots, a_i, \dots, a_n)}{h}$$

if the limit exists.

Definition 2.3.2. *Let $\boldsymbol{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ be a function that maps $\boldsymbol{x} = (x_1 \dots x_n)^T \in X$ to $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}) \dots f_m(\boldsymbol{x}))^T \in \mathbb{R}^m$. If the partial derivative $\frac{\partial f_i}{\partial x_j}$ of f_i with respect to the j-th variable x_j exists at $\boldsymbol{a} \in X$ for all $i \in \{1, \dots, m\}$ and all $j \in \{1, \dots, n\}$, then the Jacobian matrix of $\boldsymbol{f}$ at $\boldsymbol{a}$ is the matrix*

$$\boldsymbol{J}\boldsymbol{f}(\boldsymbol{a}) := \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\boldsymbol{a}) & \cdots & \frac{\partial f_1}{\partial x_n}(\boldsymbol{a}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\boldsymbol{a}) & \cdots & \frac{\partial f_m}{\partial x_n}(\boldsymbol{a}) \end{pmatrix}.$$

Definition 2.3.3. *A function $\boldsymbol{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ is called (Fréchet-) differentiable at $\boldsymbol{x} \in X$ if there exists a bounded linear map $\boldsymbol{A}_{\boldsymbol{x}} \in \mathcal{C}(\mathbb{R}^n, \mathbb{R}^m)$ such that*

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

Remark 2.3.4. *If a function $\boldsymbol{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\boldsymbol{x} \in X$, then the bounded linear map $\boldsymbol{A}_{\boldsymbol{x}} \in \mathcal{C}(\mathbb{R}^n, \mathbb{R}^m)$ is unique, all of the partial derivatives of $\boldsymbol{f}$ at $\boldsymbol{x}$ exist, and $\boldsymbol{A}_{\boldsymbol{x}}$ is the Jacobian matrix $\boldsymbol{J}\boldsymbol{f}(\boldsymbol{x})$ taken as an element of $\mathcal{C}(\mathbb{R}^n, \mathbb{R}^m)$. We call $\boldsymbol{A}_{\boldsymbol{x}}$ the (Fréchet) derivative of $\boldsymbol{f}$ at $\boldsymbol{x}$. In general, the existence of all of the partial derivatives of a function $\boldsymbol{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ at a point $\boldsymbol{x} \in X$ does not imply that $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}$. For example, consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by*

$$f(x\ y)^T = \begin{cases} x & \text{if } y \neq x^2, \\ 0 & \text{if } y = x^2. \end{cases}$$

Then f is not differentiable at $(0\ 0)^T$, but all of the partial derivatives of f at $(0\ 0)^T$ exist. However, if all of the partial derivatives of a function $\boldsymbol{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ exist in a neighborhood of a point $\boldsymbol{x} \in X$, and are continuous at $\boldsymbol{x}$, then $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}$. If $\boldsymbol{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at all points in X, then we say that $\boldsymbol{f}$ is differentiable on X.

Definition 2.3.5. *A function $\boldsymbol{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ is continuously differentiable ($\mathcal{C}^1$) on X if it is differentiable on X and the derivative $\boldsymbol{J}\boldsymbol{f} : X \to \mathcal{C}(\mathbb{R}^n, \mathbb{R}^m)$ of $\boldsymbol{f}$ is continuous on X.*

Example 2.3.6. *Let $f : \mathbb{R} \to \mathbb{R}$ be the function defined by $f(x) = x^2$. Then f is differentiable on $\mathbb{R}$ since at any $x \in \mathbb{R}$ we have a bounded linear map $A_x : \mathbb{R} \to \mathbb{R}$ defined by $A_x(h) = 2xh$ such that*

$$\lim_{h\to 0} \frac{|f(x+h) - f(x) - A_x h|}{|h|} = \lim_{h\to 0} \frac{|(x+h)^2 - x^2 - 2xh|}{|h|} = 0.$$

Hence, $A_x = \boldsymbol{J}f(x) = 2x$ is the derivative of f at x, as expected. In fact, f is $\mathcal{C}^1$ since the derivative of f mapping x to $2x$ is continuous on $\mathbb{R}$.

Definition 2.3.7. *A function $\boldsymbol{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ is directionally differentiable at $\boldsymbol{x} \in X$ in the direction of $\boldsymbol{h} \in \mathbb{R}^n$ if*

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

exists. In this case, $\boldsymbol{f}'(\boldsymbol{x}; \boldsymbol{h})$ is called the directional derivative of $\boldsymbol{f}$ at $\boldsymbol{x}$ in the direction of $\boldsymbol{h}$. If $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}$ in all directions in $\mathbb{R}^n$, then $\boldsymbol{f}$ is called directionally differentiable at $\boldsymbol{x}$. If $\boldsymbol{f}$ is directionally differentiable at all points in X, then $\boldsymbol{f}$ is called directionally differentiable on X.

Proposition 2.3.8. *Let $\boldsymbol{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ be a $\mathcal{C}^1$ function. For any $\boldsymbol{x} \in X$ and $\boldsymbol{h} \in \mathbb{R}^n$, $\boldsymbol{f}$ is directionally differentiable at $\boldsymbol{x}$ in the direction of $\boldsymbol{h}$. Also,*

$$\boldsymbol{f}'(\boldsymbol{x}; \boldsymbol{h}) = \boldsymbol{J}\boldsymbol{f}(\boldsymbol{x})\boldsymbol{h}.$$

Example 2.3.9. *Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a function defined by $f(x_1\ x_2)^T = x_1^2 + x_2^2$. Then f is $\mathcal{C}^1$. Note that*

$$\boldsymbol{J}f(\boldsymbol{x})\boldsymbol{d} = (2x_1\ 2x_2)\begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = 2x_1 d_1 + 2x_2 d_2.$$

Then

$$\begin{aligned} f'(\boldsymbol{x}; \boldsymbol{d}) &= \lim_{\alpha\downarrow 0} \frac{f(\boldsymbol{x} + \alpha\boldsymbol{d}) - f(\boldsymbol{x})}{\alpha} \\ &= \lim_{\alpha\downarrow 0} \frac{(x_1 + \alpha d_1)^2 + (x_2 + \alpha d_2)^2 - x_1^2 - x_2^2}{\alpha} \\ &= 2x_1 d_1 + 2x_2 d_2 \\ &= \boldsymbol{J}f(\boldsymbol{x})\boldsymbol{d}, \end{aligned}$$

as expected.

2.3.1 The Theoretical Toolkit of Differentiation

Here, we present the classical theory of differentiation.

Theorem 2.3.10 (The classical mean value theorem)**.** *Let $[a, b] \subset \mathbb{R}$ be a compact interval and $f : [a, b] \to \mathbb{R}$ a function continuous on $[a, b]$ and differentiable on (a, b). Then there exists $c \in (a, b)$ such that*

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Theorem 2.3.11. *If $\boldsymbol{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ and $\boldsymbol{g} : X \to \mathbb{R}^m$ are functions differentiable at $\boldsymbol{x} \in X$, then for any $a, b \in \mathbb{R}$, the function $a\boldsymbol{f} + b\boldsymbol{g} : X \to \mathbb{R}^m$ defined by $[a\boldsymbol{f} + b\boldsymbol{f}](\boldsymbol{y}) = a\boldsymbol{f}(\boldsymbol{y}) + b\boldsymbol{g}(\boldsymbol{y})$ is differentiable at $\boldsymbol{x}$. Moreover,*

$$\boldsymbol{J}[a\boldsymbol{f} + b\boldsymbol{g}](\boldsymbol{x}) = a\boldsymbol{J}\boldsymbol{f}(\boldsymbol{x}) + b\boldsymbol{J}\boldsymbol{g}(\boldsymbol{x}).$$

Theorem 2.3.12 (The classical chain rule)**.** *If $\boldsymbol{f} : X \subset \mathbb{R}^n \to Y \subset \mathbb{R}^m$ is a function differentiable at $\boldsymbol{x} \in X$ and $\boldsymbol{g} : Y \to \mathbb{R}^k$ is a function differentiable at $\boldsymbol{f}(\boldsymbol{x}) \in Y$, then the composition $\boldsymbol{g} \circ \boldsymbol{f} : X \to \mathbb{R}^k$ is differentiable at $\boldsymbol{x}$. Moreover, its derivative at $\boldsymbol{x}$ is given by*

$$\boldsymbol{J}[\boldsymbol{g} \circ \boldsymbol{f}](\boldsymbol{x}) = \boldsymbol{J}\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}))\,\boldsymbol{J}\boldsymbol{f}(\boldsymbol{x}).$$

Let $\boldsymbol{f} : X \times Y \subset \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be a function. Consider the following equation:

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{0}.$$

An implicit function theorem provides conditions under which it is possible to solve this equation for $\boldsymbol{y}$ as a function of $\boldsymbol{x}$ in the neighborhood of a known solution $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})$. There are a number of implicit function theorems; the stronger the assumptions, the stronger the conclusions. Below, we give a standard implicit function theorem.

Theorem 2.3.13 (The classical implicit function theorem)**.** *Let $\boldsymbol{f} : X \times Y \subset \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be a $\mathcal{C}^1$ function and suppose there exist $\bar{x} \in X$ and $\bar{y} \in Y$*

such that $\boldsymbol{f}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}) = \boldsymbol{0}$. *If*

$$\begin{pmatrix} \frac{\partial f_1}{\partial y_1}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}) & \cdots & \frac{\partial f_1}{\partial y_m}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial y_1}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}) & \cdots & \frac{\partial f_m}{\partial y_m}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}) \end{pmatrix}$$

is invertible, then there exist neighborhoods $U \subset X$ *and* $V \subset Y$ *of* $\bar{\boldsymbol{x}}$ *and* $\bar{\boldsymbol{y}}$, *respectively, on which the equation*

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{0}$$

uniquely defines $\boldsymbol{y}$ *as a function of* $\boldsymbol{x}$. *That is, there is a function* $\boldsymbol{p} : U \to V$ *such that*

1. $\boldsymbol{f}(\boldsymbol{u}, \boldsymbol{p}(\boldsymbol{u})) = \boldsymbol{0}$ *for all* $\boldsymbol{u} \in U$.

2. *For each* $\boldsymbol{u} \in U$, $\boldsymbol{p}(\boldsymbol{u})$ *is the unique solution to* $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{0}$ *lying in* V. *In particular,* $\boldsymbol{p}(\bar{\boldsymbol{x}}) = \bar{\boldsymbol{y}}$.

3. $\boldsymbol{p}$ *is* $\mathcal{C}^1$ *on* U. *Moreover, for any* $\boldsymbol{u} \in U$,

$$\begin{pmatrix} \frac{\partial p_1}{\partial u_1}(\boldsymbol{u}) & \cdots & \frac{\partial p_1}{\partial u_n}(\boldsymbol{u}) \\ \vdots & \ddots & \vdots \\ \frac{\partial p_m}{\partial u_1}(\boldsymbol{u}) & \cdots & \frac{\partial p_m}{\partial u_n}(\boldsymbol{u}) \end{pmatrix} = -\begin{pmatrix} \frac{\partial f_1}{\partial y_1}(\boldsymbol{u}) & \cdots & \frac{\partial f_1}{\partial y_m}(\boldsymbol{u}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial y_1}(\boldsymbol{u}) & \cdots & \frac{\partial f_m}{\partial y_m}(\boldsymbol{u}) \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\boldsymbol{u}) & \cdots & \frac{\partial f_1}{\partial x_n}(\boldsymbol{u}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\boldsymbol{u}) & \cdots & \frac{\partial f_m}{\partial x_n}(\boldsymbol{u}) \end{pmatrix}.$$

2.3.2 Newton's Method

In addition to a useful theoretical toolkit, the theory of differentiation possesses a strong numerical toolkit. In this section, we demonstrate how the derivative is used in equation-solving methods. In particular, we will use Newton's method to solve the system of equations

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$$

where $\boldsymbol{f} : X \subset \mathbb{R}^n \to \mathbb{R}^n$ is a $\mathcal{C}^1$ function. Let $\boldsymbol{x}_0$ be an initial guess for a root of $\boldsymbol{f}$. If $\boldsymbol{x}_{k-1}$ is a current approximation for the root, then Newton's method gives a formula for the subsequent approximation $\boldsymbol{x}_k$, i.e.,

$$\boldsymbol{x}_k = \boldsymbol{x}_{k-1} - [\boldsymbol{J}\boldsymbol{f}(\boldsymbol{x}_{k-1})]^{-1}\,\boldsymbol{f}(\boldsymbol{x}_{k-1}).$$

For sufficiently simple systems, we can often compute the inverse of the Jacobian matrix directly with no difficulty. For more complicated systems, it is usually more efficient to solve the following equivalent system of linear equations for $\boldsymbol{x}_k - \boldsymbol{x}_{k-1}$:

$$\boldsymbol{J}\boldsymbol{f}(\boldsymbol{x}_{k-1})(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}) = -\boldsymbol{f}(\boldsymbol{x}_{k-1}).$$

Example 2.3.14. *Suppose that we are given the following system of equations:*

$$\begin{pmatrix} x_1 + x_2 \\ {x_1}^2 - x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Geometrically, solutions to this system of equations correspond to the intersections of the curves $y = -x$ and $y = x^2$ on the Cartesian plane. In this example, we will use Newton's method to solve this system of equations. Letting $\boldsymbol{x} = (x_1\ x_2)^T$, we can rewrite this system of equations as

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0},$$

where $\boldsymbol{f}(\boldsymbol{x}) = (x_1 + x_2\ {x_1}^2 - x_2)^T$. Then, the Jacobian matrix of $\boldsymbol{f}$ at $\boldsymbol{x}$ is

$$\boldsymbol{J}\boldsymbol{f}(\boldsymbol{x}) = \begin{pmatrix} 1 & 1 \\ 2x_1 & -1 \end{pmatrix}$$

and its inverse is

$$[\boldsymbol{J}\boldsymbol{f}(\boldsymbol{x})]^{-1} = \frac{1}{2x_1 + 1}\begin{pmatrix} 1 & 1 \\ 2x_1 & -1 \end{pmatrix}.$$

$\boldsymbol{x}_0$	$(-2\ 3)^T$
$\boldsymbol{x}_1$	$(-\frac{4}{3}\ \frac{4}{3})^T$
$\boldsymbol{x}_2$	$(-\frac{16}{15}\ \frac{16}{15})^T$
$\boldsymbol{x}_3$	$(-\frac{256}{255}\ \frac{256}{255})^T$
$\boldsymbol{x}_4$	$(-\frac{65536}{65535}\ \frac{65536}{65535})^T$

Table 2.1. First few approximations to the root $(-1\ 1)^T$

The above table displays the first few approximations to the left intersection point $(-1\ 1)^T$ *of the two curves provided that we choose* $\boldsymbol{x}_0 = (-2\ 3)^T$ *as the initial guess for it. As expected, these approximations are approaching* $(-1\ 1)^T$.

CHAPTER 3
GENERALIZED DIFFERENTIATION THEORY

In this chapter, we present a generalized theory of differentiation for a class of functions called locally Lipschitz continuous functions.

3.1 Notation

In this chapter, we use X and Y to denote open subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively.

3.2 Nonsmooth Functions

3.2.1 Locally Lipschitz Continuous Functions

Definition 3.2.1. *A function $\boldsymbol{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous if there exists $K \geq 0$ such that for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in X,*

$$|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})| \leq K \, |\boldsymbol{x} - \boldsymbol{y}| \, .$$

Any such constant K is called a Lipschitz constant. A function $\boldsymbol{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ is locally Lipschitz continuous if for any $\boldsymbol{x} \in X$, there exists a neighborhood $U_{\boldsymbol{x}}$ of $\boldsymbol{x}$ such that $\boldsymbol{f}$ restricted to $U_{\boldsymbol{x}}$ is Lipschitz continuous.

Remark 3.2.2. *A Lipschitz continuous function is locally Lipschitz continuous. However, the converse is not true. For example, the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$ is locally Lipschitz continuous, but not Lipschitz continuous. Moreover, compositions of locally Lipschitz continuous functions are locally Lipschitz continuous.*

A $\mathcal{C}^1$ function is locally Lipschitz continuous. Not every differentiable function is locally Lipschitz continuous. For example, consider the function $h : \mathbb{R} \to \mathbb{R}$ defined by

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

Then h is differentiable, with the derivative

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

However, h is not locally Lipschitz continuous, due to its behavior at 0*. Conversely, not every locally Lipschitz continuous function is differentiable. For example, the absolute value function* $f : \mathbb{R} \to \mathbb{R}$ *defined by* $f(x) = |x|$ *is locally Lipschitz continuous, but not differentiable.*

Example 3.2.3. *The square root function* $f : \mathbb{R} \to \mathbb{R}$ *defined by* $f(x) = \sqrt{|x|}$ *is not locally Lipschitz continuous.*

It is convenient to develop a generalized theory of differentiation for locally Lipschitz continuous functions due to Rademacher's theorem, because it allows us to work with nondifferentiability in a controlled manner. The following theorem is from [4].

Theorem 3.2.4 (Rademacher's Theorem)**.** *If* $\boldsymbol{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ *is locally Lipschitz continuous, then* $\boldsymbol{f}$ *is differentiable almost everywhere on* X*.*

Example 3.2.5. *For the absolute value function* $f : \mathbb{R} \to \mathbb{R}$ *defined by* $f(x) = |x|$*, the set of points of nondifferentiability in the domain is* $\{0\}$*, which has Lebesgue measure* 0*.*

Example 3.2.6. *For the function* $g : \mathbb{R}^2 \to \mathbb{R}$ *defined by* $g(x_1\ x_2)^T = \min(x_1, x_2)$*, the set of points of nondifferentiability in the domain is* $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 = x_2\}$*, which has Lebesgue measure* 0*.*

3.2.2 Piecewise Differentiable Functions

There is an important subclass of locally Lipschitz continuous functions containing the class of $\mathcal{C}^1$ functions, which we introduce below. The material here comes from [11].

Definition 3.2.7. *A function* $\boldsymbol{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ *is piecewise differentiable (*$\mathcal{PC}^1$*) at* $\boldsymbol{x} \in X$ *if there exists a neighborhood* $N_{\boldsymbol{x}} \subset X$ *of* $\boldsymbol{x}$ *and a finite collection*

$\mathcal{F}_{\boldsymbol{f}}(\boldsymbol{x}) = \{\boldsymbol{f}_{(1)}, \ldots, \boldsymbol{f}_{(k)}\}$ of $\mathcal{C}^1$ functions mapping $N_{\boldsymbol{x}}$ into $\mathbb{R}^m$ such that $\boldsymbol{f}$ is continuous on $N_{\boldsymbol{x}}$ and for all $\boldsymbol{y} \in N_{\boldsymbol{x}}$, $\boldsymbol{f}(\boldsymbol{y}) \in \{\boldsymbol{f}_{(i)}(\boldsymbol{y}) : i \in \{1, \ldots, k\}\}$. If $\boldsymbol{f}$ is $\mathcal{PC}^1$ at all points of X, then $\boldsymbol{f}$ is called $\mathcal{PC}^1$ on X. Given the set of functions $\mathcal{F}_{\boldsymbol{f}}(\boldsymbol{x})$, there is an associated set

$$I^{ess}_{\boldsymbol{f}}(\boldsymbol{x}) := \left\{i \in \{1, \ldots, k\} : \boldsymbol{x} \in cl\left\{int\left\{\left\{\boldsymbol{y} \in N_{\boldsymbol{x}} : \boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{f}_{(i)}(\boldsymbol{y})\right\}\right\}\right\}\right\},$$

called the set of essentially active indices of $\boldsymbol{f}$ at $\boldsymbol{x}$ with respect to $\mathcal{F}_{\boldsymbol{f}}(\boldsymbol{x})$. Corresponding to a set of essentially active indices is a set of essentially active functions

$$\mathcal{E}_{\boldsymbol{f}}(\boldsymbol{x}) := \left\{\boldsymbol{f}_{(i)} : i \in I^{ess}_{\boldsymbol{f}}(\boldsymbol{x})\right\}.$$

Remark 3.2.8. *A composition of $\mathcal{PC}^1$ functions is $\mathcal{PC}^1$. A $\mathcal{PC}^1$ function is directionally differentiable.*

Example 3.2.9. *Let $g : \mathbb{R}^2 \to \mathbb{R}$ be the function defined by $g(x_1\ x_2)^T = \max(0, \min(x_1, x_2))$. Then g is $\mathcal{PC}^1$, with $\mathcal{F}_g\left((0\ 0)^T\right) = \left\{g_{(1)}, g_{(2)}, g_{(3)}\right\}$, where*

$$g_{(1)}(x_1\ x_2)^T = 0$$
$$g_{(2)}(x_1\ x_2)^T = x_1$$
$$g_{(3)}(x_1\ x_2)^T = x_2.$$

We will show that a directional derivative of g can be obtained from an appropriate selection function. Since

$$g'\left((1\ 1)^T; (-1\ 0)^T\right) = \lim_{\alpha \to 0^+} \frac{g\left((1\ 1)^T + \alpha(-1\ 0)^T\right) - g(1\ 1)^T}{\alpha} = -1$$

and

$$g'_{(2)}((1\ 1)^T; (-1\ 0)^T) = \boldsymbol{J}g_{(2)}(1\ 1)^T \begin{pmatrix} -1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} -1 \\ 0 \end{pmatrix} = -1,$$

we see that $g'((1\ 1)^T; (-1\ 0)^T) = g'_{(2)}((1\ 1)^T; (-1\ 0)^T)$.

Example 3.2.10. *The absolute value function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = |x|$ is $\mathcal{PC}^1$ on $\mathbb{R}$.*

Example 3.2.11. *For any $n \in \mathbb{N}$, the minimum function* $\min : \mathbb{R}^n \to \mathbb{R}$ *defined by* $\min(\boldsymbol{x}) = \min\{x_1, \ldots, x_n\}$ *where* $\boldsymbol{x} = (x_1 \ldots x_n)^T$ *and the maximum function* $\max : \mathbb{R}^n \to \mathbb{R}$ *defined by* $\max(\boldsymbol{x}) = \max\{x_1, \ldots, x_n\}$ *where* $\boldsymbol{x} = (x_1 \ldots x_n)^T$ *are* $\mathcal{PC}^1$ *on* $\mathbb{R}^n$.

Example 3.2.12. *For any $n \in \{2, 3, 4, \ldots\}$, the Euclidean norm function* $|\cdot| : \mathbb{R}^n \to \mathbb{R}$ *mapping* $\boldsymbol{x}$ *to* $|\boldsymbol{x}|$ *is not* $\mathcal{PC}^1$ *at* $\mathbf{0}$.

3.3 Clarke's Theory of Differentiation

In this section, definitions and results come from [3].

3.3.1 Bouligand Subdifferentials and Clarke Jacobians

Definition 3.3.1. *Let* $\boldsymbol{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ *be a locally Lipschitz continuous function. If* $Z_{\boldsymbol{f}} \subset X$ *is the set of points of nondifferentiability in the domain of* $\boldsymbol{f}$*, then the Bouligand (B-) subdifferential of* $\boldsymbol{f}$ *at* $\boldsymbol{x} \in X$ *is*

$$\partial_B \boldsymbol{f}(\boldsymbol{x}) := \left\{ \boldsymbol{H} \in \mathbb{R}^{m \times n} : \boldsymbol{H} = \lim_{j \to \infty} \boldsymbol{J}\boldsymbol{f}(\boldsymbol{x}_{(j)}) \text{ for some sequence } \left\{\boldsymbol{x}_{(j)}\right\} \text{ in } X \setminus Z_{\boldsymbol{f}} \text{ converging to } \boldsymbol{x} \right\}.$$

Definition 3.3.2. *Let* $\boldsymbol{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ *be a locally Lipschitz continuous function. If* $Z_{\boldsymbol{f}} \subset X$ *is the set of points of nondifferentiability in the domain of* $\boldsymbol{f}$*, then the Clarke Jacobian of* $\boldsymbol{f}$ *at* $\boldsymbol{x} \in X$ *is*

$$\partial \boldsymbol{f}(\boldsymbol{x}) := \mathit{conv}\ \partial_B \boldsymbol{f}(\boldsymbol{x}).$$

In words, the Clarke Jacobian of $\boldsymbol{f}$ at $\boldsymbol{x} \in X$ is the convex hull of the B-subdifferential of $\boldsymbol{f}$ at $\boldsymbol{x}$.

Proposition 3.3.3. *A Clarke Jacobian is nonempty, compact, and convex.*

Example 3.3.4. *For the absolute value function* $f : \mathbb{R} \to \mathbb{R}$ *defined by* $f(x) = |x|$*,*

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

and

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

In the above example, $\boldsymbol{f}$ is $\mathcal{PC}^1$. In general, the Clarke Jacobian is straightforward to compute for $\mathcal{PC}^1$ functions.

Proposition 3.3.5. *Let* $\boldsymbol{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ *be a* $\mathcal{PC}^1$ *function on* X*, and* $\boldsymbol{x} \in X$*. Then*

$$\partial \boldsymbol{f}(\boldsymbol{x}) = conv \ \left\{ \boldsymbol{J}\boldsymbol{f}_{(i)}(\boldsymbol{x}) : i \in I_{\boldsymbol{f}}^{ess}(\boldsymbol{x}) \right\}.$$

Example 3.3.6. *Recall the function* g *in Example 3.2.9. We have*

$$\partial_B g(0 \ 0)^T = \{[1 \ 0], [0 \ 1], [0 \ 0]\}$$

and

$$\partial g(0 \ 0)^T = \{[\lambda_1 \ \lambda_2] : \lambda_1, \lambda_2 \geq 0, \lambda_1 + \lambda_2 \leq 1\}.$$

3.3.2 The Theoretical Toolkit of Clarke Jacobians

Just like classical differentiaton, Clarke's theory of differentiation possesses a useful theoretical toolkit.

Theorem 3.3.7 (The chain rule for Clarke Jacobians)**.** *If* $\boldsymbol{f} : X \subset \mathbb{R}^n \to Y \subset \mathbb{R}^m$ *and* $\boldsymbol{h} : Y \to \mathbb{R}^r$ *are locally Lipschitz continuous functions, then* $\boldsymbol{h} \circ \boldsymbol{f}$ *is locally Lipschitz continuous and the Clarke Jacobian of* $\boldsymbol{h} \circ \boldsymbol{f}$ *satisfies*

$$\partial \left[\boldsymbol{h} \circ \boldsymbol{f}\right](\boldsymbol{x}) \subset conv \ \{\boldsymbol{H}\boldsymbol{F} : \boldsymbol{H} \in \partial \boldsymbol{h}(\boldsymbol{f}(\boldsymbol{x})), \boldsymbol{F} \in \partial \boldsymbol{f}(\boldsymbol{x})\}.$$

Remark 3.3.8. *Unlike the classical chain rule, which is an equality-based rule, the Clarke Jacobian chain rule is an inclusion-based rule.*

Example 3.3.9. *Let* $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x_1\ x_2)^T = \min(x_1, x_2)$, *and* $h : \mathbb{R} \to \mathbb{R}$, $h(y) = \max(0, y)$, *both of which are locally Lipschitz continuous functions. Then* $[h \circ f](x_1\ x_2)^T = \max(0, \min(x_1, x_2))$. *As seen above,*

$$\partial_B [h \circ f](0\ 0)^T = \{[1\ 0], [0\ 1], [0\ 0]\}$$

and

$$\partial [h \circ f](0\ 0)^T = \{[\lambda_1\ \lambda_2] : \lambda_1, \lambda_2 \geq 0, \lambda_1 + \lambda_2 \leq 1\}.$$

On the other hand,

$$\begin{aligned}
&\mathit{conv}\ \left\{H\boldsymbol{F} : H \in \partial h\left(f(0\ 0)^T\right), \boldsymbol{F} \in \partial f(0\ 0)^T\right\} \\
&= \mathit{conv}\ \{H\boldsymbol{F} : H \in [0,1], \boldsymbol{F} \in \{[\lambda\ 1-\lambda] : 0 \leq \lambda \leq 1\}\} \\
&= \mathit{conv}\ \{H[\lambda\ 1-\lambda] : 0 \leq H \leq 1, 0 \leq \lambda \leq 1\} \\
&= \{[\lambda_1\ \lambda_2] : \lambda_1, \lambda_2 \geq 0, \lambda_1 + \lambda_2 \leq 1\} \\
&= \partial [h \circ f](0\ 0)^T.
\end{aligned}$$

In the above example, the chain rule relation holds with equality. In general, however, it is not the case.

Example 3.3.10. *Let* $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \min(0, x)$, *and* $h : \mathbb{R} \to \mathbb{R}$, $h(y) = \max(0, y)$, *both of which are locally Lipschitz continuous functions. Then* $[h \circ f](x) = \max(0, \min(0, x)) = 0$, *i.e., the composition is identically zero. Note that*

$$\begin{aligned}
\partial_B [h \circ f](0) &= \{0\} \\
\partial [h \circ f](0) &= \mathit{conv}\ \{0\} = \{0\}.
\end{aligned}$$

On the other hand,

$$\mathit{conv}\ \{HF : H \in \partial h(f(0)), F \in \partial f(0)\} = \mathit{conv}\ \{HF : H \in [0,1], F \in [0,1]\} = [0,1].$$

In the above example, the chain rule relation is a proper inclusion.

The Clarke Jacobian possesses a theoretical toolkit that mirrors that of the classical derivative, which includes a mean value theorem and an implicit function theorem.

Theorem 3.3.11 (The mean value theorem for Clarke Jacobians)**.** *Let $[a,b] \subset \mathbb{R}$ be a compact interval and $f : [a,b] \to \mathbb{R}$ a function Lipschitz continuous on $[a,b]$. Then there exists $c \in (a,b)$ such that*

$$\frac{f(b) - f(a)}{b - a} \in \partial f(c).$$

Theorem 3.3.12 (The implicit function theorem for Clarke Jacobians)**.** *If $\boldsymbol{g} : X \subset \mathbb{R}^n \times Y \subset \mathbb{R}^m \to \mathbb{R}^m$ is a locally Lipschitz continuous function such that $\boldsymbol{g}(\boldsymbol{x}_0, \boldsymbol{y}_0) = \boldsymbol{0}$ for some $\boldsymbol{x}_0 \in X$ and $\boldsymbol{y}_0 \in Y$, and det$(\boldsymbol{Y}) \neq 0$ for all*

$$\boldsymbol{Y} \in \left\{ \boldsymbol{Y} \in \mathbb{R}^{n \times n} : [\boldsymbol{X}\ \boldsymbol{Y}] \in \partial \boldsymbol{g}(\boldsymbol{x}_0, \boldsymbol{y}_0) \right\},$$

then there exists a Lipschitz continuous function $\boldsymbol{r} : \mathbb{R}^n \to \mathbb{R}^m$ such that $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{r}(\boldsymbol{x})) = \boldsymbol{0}$ in a neighborhood of $\boldsymbol{x}_0$.

3.3.3 Nonsmooth Newton's Method

In addition to a useful theoretical toolkit, Clarke's theory of differentiation possesses a strong numerical toolkit that includes Newton's methods and local optimization methods. In this section, we introduce a Newton's method for the class of locally Lipschitz continuous functions. Consider a locally Lipschitz continuous function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$. Let $\boldsymbol{x}_{(0)}$ be an initial guess for a root of $\boldsymbol{f}$. We can obtain subsequent approximations to the root using the recursive formula

$$\boldsymbol{x}_{(k)} = \boldsymbol{x}_{(k-1)} - \boldsymbol{H}^{-1} \boldsymbol{f}\left(\boldsymbol{x}_{(k-1)}\right), \text{ where } \boldsymbol{H} \in \partial \boldsymbol{f}\left(\boldsymbol{x}_{(k-1)}\right), \quad k \in \mathbb{N}.$$

Example 3.3.13. *Suppose that we are given the equation*

$$f(x) = 0$$

to solve, where

$$f(x) = \begin{cases} |x| & \text{if } |x| < 1 \\ x^2 & \text{if } |x| \geq 1. \end{cases}$$

To find a root of this equation, we can apply the locally Lipschitz continuous version of Newton's method, since f is locally Lipschitz continuous. Let $x_0 = 1$ be an initial guess for the root 0*. Then*

$$x_1 = x_0 - \frac{1}{a} f(x_0) = 1 - \frac{1}{a}, \text{ where } a \in \partial f(x_0) = \text{conv } \{1, 2\} = [1, 2].$$

To obtain x_2, we may choose any $a \in \partial f(x_0) = [1, 2]$. For any $a \in [1, 2]$,

$$x_2 = x_1 - \frac{1}{b} f(x_1) = \left(1 - \frac{1}{a}\right) - \frac{1}{b}\left(1 - \frac{1}{a}\right), \text{ where } b \in \partial f(x_1) = \text{conv } \{1\} = \{1\},$$

i.e., $x_2 = 0$, the root of the equation.

It is worth noting that the classical version of Newton's method would have failed to give us x_1 since f is not differentiable at $x_0 = 1$.

As we have just seen with Newton's method, Clarke Jacobian elements are essential in implementing numerical tools for the class of locally Lipschitz continuous functions. However, the fact that calculus rules for Clarke Jacobians are inclusion-based makes it difficult to obtain Clarke Jacobian elements even for a function constructed from functions whose Clarke Jacobians are known.

Example 3.3.14. *If $\boldsymbol{g} : X \subset \mathbb{R}^n \to \mathbb{R}^m$ and $\boldsymbol{h} : X \to \mathbb{R}^m$ are locally Lipschitz continuous functions, then $\boldsymbol{g} + \boldsymbol{h} : X \to \mathbb{R}^m$ is locally Lipschitz continuous and*

$$\partial[\boldsymbol{g} + \boldsymbol{h}](\mathbf{0}) \subset \partial \boldsymbol{g}(\mathbf{0}) + \partial \boldsymbol{h}(\mathbf{0}),$$

where the inclusion may be proper. Thus, adding a Clarke Jacobian element of $\boldsymbol{g}$ and a Clarke Jacobian element of $\boldsymbol{h}$ will not necessarily yield a Clarke Jacobian element of $\boldsymbol{g} + \boldsymbol{h}$. For instance, consider $g : \mathbb{R} \to \mathbb{R}$, $g(x) = \max(0, x)$, and $h : \mathbb{R} \to \mathbb{R}$, $h(x) = \min(0, x)$, both of which are locally Lipschitz continuous with $\partial g(0) = [0, 1]$ and $\partial h(0) = [0, 1]$. By the above *rule,* $g+h:\mathbb{R}\to\mathbb{R}$*,* $[g+h](x)=x$*, is locally Lipschitz continuous with* $\partial[g+h](0)=\{1\}$*. However, although* $0\in\partial g(0)\cup\partial h(0)=[0,1]$*, we have* $0+0\notin\partial f(0)=\{1\}$*.*

Example 3.3.15. *In general, we cannot compute Clarke Jacobian elements for a vector-valued function* $\boldsymbol{f}=(f_1\ \dots\ f_m)^T:\mathbb{R}^n\to\mathbb{R}^m$ *component-wise due to the following rule:*

$$\partial\boldsymbol{f}(\boldsymbol{x})\subset\partial f_1(\boldsymbol{x})\times\cdots\times\partial f_m(\boldsymbol{x})=\left\{\begin{pmatrix}\boldsymbol{F}_1\\ \vdots\\ \boldsymbol{F}_m\end{pmatrix}:\boldsymbol{F}_i\in\partial f_i(\boldsymbol{x})\right\},$$

where the inclusion may be proper. For instance, consider the function $\boldsymbol{f}:\mathbb{R}^2\to\mathbb{R}^2$ *defined by*

$$\boldsymbol{f}(x_1\ x_2)^T=\begin{pmatrix}x_1+|x_2|\\ x_1-|x_2|\end{pmatrix}.$$

Then

$$\partial\boldsymbol{f}(0\ 0)^T=\left\{\begin{pmatrix}1 & 2\lambda-1\\ 1 & 1-2\lambda\end{pmatrix}:0\le\lambda\le 1\right\}.$$

However,

$$\partial\boldsymbol{f}(0\ 0)^T\subset\partial f_1(0\ 0)^T\times\partial f_2(0\ 0)^T=\left\{\begin{pmatrix}1 & 2\lambda_1-1\\ 1 & 2\lambda_2-1\end{pmatrix}:(\lambda_1\ \lambda_2)^T\in[0,1]^2\right\},$$

where the inclusion is proper.

Example 3.3.16. *In general, it is not possible to compute a B-subdifferential element or Clarke Jacobian element of a function by calculating its directional derivatives in coordinate directions and then "stitching" them together. For instance, let* $f:\mathbb{R}^2\to\mathbb{R}$ *be a function defined by* $f(x_1\ x_2)^T=|x_1-x_2|$*. Note that* f *is* $\mathcal{PC}^1$*, with* $\mathcal{F}_f((0\ 0)^T)=\{f_{(1)},f_{(2)}\}$*, where*

$$f_{(1)}(x_1\ x_2)^T=x_1-x_2$$
$$f_{(2)}(x_1\ x_2)^T=-x_1+x_2.$$

Then

$$f'((0\ 0)^T;(1\ 0)^T) = \left[\boldsymbol{J}f_{(1)}(0\ 0)^T\right](1\ 0)^T = (1\ \ -1)(1\ 0)^T = 1$$

$$f'((0\ 0)^T;(0\ 1)^T) = \left[\boldsymbol{J}f_{(2)}(0\ 0)^T\right](0\ 1)^T = (-1\ 1)(0\ 1)^T = 1.$$

Moreover,

$$\partial_B f(0\ 0)^T = \{(1\ \ -1),(-1\ 1)\}$$

$$\partial f(0\ 0)^T = \{(1-\lambda\ \ -1+\lambda) : 0 \leq \lambda \leq 2\}.$$

However,

$$\big(f'((0\ 0)^T;(1\ 0)^T)\ f'((0\ 0)^T;(0\ 1)^T)\big) = (1\ 1) \notin \partial f(0\ 0)^T.$$

CHAPTER 4
THEORY OF LEXICOGRAPHIC DIFFERENTIATION

Recently, lexicographic differentiation introduced by Nesterov [9] and lexicographic directional differentiation by Khan and Barton [6] have provided a way to find Clarke Jacobian elements systematically. In this chapter, we will present both theories, which are developed for a subclass of locally Lipschitz functions called lexicographically (L-) smooth functions. We will fill in some details for lexicographic differentiation in the Banach space setting that had been omitted by Nesterov and provide new illustrative examples.

4.1 Preliminaries

4.1.1 Notation

In this section, let E_1, E_2, and E_3 be real normed vector spaces and let $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_3$ denote the norms of E_1, E_2, and E_3, respectively.

4.1.2 Locally Lipschitz Continuous and Directionally Differentiable Functions

Definition 4.1.1. *A function $\boldsymbol{f} : E_1 \to E_2$ is Lipschitz continuous if there exists $K \geq 0$ such that for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in E_1,*

$$\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})\|_2 \leq K \|\boldsymbol{x} - \boldsymbol{y}\|_1 .$$

Any such constant K is called a Lipschitz constant. A function $\boldsymbol{f} : E_1 \to E_2$ is locally Lipschitz continuous if for any $\boldsymbol{x} \in E_1$, there exists a neighborhood $U_{\boldsymbol{x}}$ of $\boldsymbol{x}$ such that $\boldsymbol{f}$ restricted to $U_{\boldsymbol{x}}$ is Lipschitz continuous.

Definition 4.1.2. *A function $\boldsymbol{f} : E_1 \to E_2$ is differentiable at $\boldsymbol{x} \in E_1$ in the direction of $\boldsymbol{h} \in E_1$ if*

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

exists. In that case, $\boldsymbol{f}'(\boldsymbol{x};\boldsymbol{h})$ *is called the directional derivative of* $\boldsymbol{f}$ *at* $\boldsymbol{x}$ *in the direction* $\boldsymbol{h}$*. If* $\boldsymbol{f}$ *is differentiable at* $\boldsymbol{x}$ *in all directions in* E_1*, then* $\boldsymbol{f}$ *is called directionally differentiable at* $\boldsymbol{x}$*. If* $\boldsymbol{f}$ *is directionally differentiable at all points in* E_1*, then* $\boldsymbol{f}$ *is called directionally differentiable on* E_1*.*

If $\boldsymbol{f} : E_1 \to E_2$ is locally Lipschitz continuous and directionally differentiable on E_1, we write

$$\boldsymbol{f} \in \mathcal{D}(E_1, E_2).$$

Definition 4.1.3. *Let* $\boldsymbol{f} \in \mathcal{D}(E_1, E_2)$ *and* $\boldsymbol{x} \in E_1$*. The function* $H_{\boldsymbol{x}}[\boldsymbol{f}] : E_1 \to E_2$ *defined by*

$$H_{\boldsymbol{x}}[\boldsymbol{f}](\boldsymbol{h}) = \boldsymbol{f}'(\boldsymbol{x};\boldsymbol{h}), \quad \boldsymbol{h} \in E_1$$

is called the homogenization of $\boldsymbol{f}$ *at* $\boldsymbol{x}$*.*

Proposition 4.1.4. *Let* $\boldsymbol{f}, \boldsymbol{g} \in \mathcal{D}(E_1, E_2)$ *and* $\boldsymbol{x} \in E_1$*. Then for any* $\alpha, \beta \in \mathbb{R}$*,*

$$H_{\boldsymbol{x}}[\alpha \boldsymbol{f} + \beta \boldsymbol{g}] = \alpha H_{\boldsymbol{x}}[\boldsymbol{f}] + \beta H_{\boldsymbol{x}}[\boldsymbol{g}].$$

Proof. Let $\boldsymbol{f}, \boldsymbol{g} \in \mathcal{D}(E_1, E_2)$, $\boldsymbol{x} \in E_1$, and $\alpha, \beta \in \mathbb{R}$. First, we check that $\alpha \boldsymbol{f} + \beta \boldsymbol{g} \in \mathcal{D}(E_1, E_2)$ to ensure that the function $H_{\boldsymbol{x}}[\alpha \boldsymbol{f} + \beta \boldsymbol{g}] : E_1 \to E_2$ is defined. Let $\boldsymbol{z} \in E_1$. Since $\boldsymbol{f}, \boldsymbol{g} \in \mathcal{D}(E_1, E_2)$, there exist neighborhoods $U_{\boldsymbol{z}}$ and $V_{\boldsymbol{z}}$ of $\boldsymbol{z}$ on which $\boldsymbol{f}$ and $\boldsymbol{g}$ are Lipschitz continuous, respectively. Then for all $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ in the neighborhood $U_{\boldsymbol{z}} \cap V_{\boldsymbol{z}}$ of $\boldsymbol{z}$,

$$\begin{aligned}
\|(\alpha \boldsymbol{f} + \beta \boldsymbol{g})(\boldsymbol{w}_1) - (\alpha \boldsymbol{f} + \beta \boldsymbol{g})(\boldsymbol{w}_2)\|_2 &\leq \alpha \|\boldsymbol{f}(\boldsymbol{w}_1) - \boldsymbol{f}(\boldsymbol{w}_2)\|_2 + \beta \|\boldsymbol{g}(\boldsymbol{w}_1) - \boldsymbol{g}(\boldsymbol{w}_2)\|_2 \\
&\leq \alpha K_1 \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_1 + \beta K_2 \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_1 \\
&\leq K_3 \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_1,
\end{aligned}$$

where $K_3 = \max\{\alpha K_1, \beta K_2\}$. Therefore, $\alpha \boldsymbol{f} + \beta \boldsymbol{g}$ is locally Lipschitz continuous. Let $\boldsymbol{h} \in E_1$. Since

$$\begin{aligned}
\lim_{t \to 0^+} \frac{(\alpha \boldsymbol{f} + \beta \boldsymbol{g})(\boldsymbol{x} + t\boldsymbol{h}) - (\alpha \boldsymbol{f} + \beta \boldsymbol{g})(\boldsymbol{x})}{t} &= \lim_{t \to 0^+} \left[\alpha \frac{\boldsymbol{f}(\boldsymbol{x} + t\boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{x})}{t} + \beta \frac{\boldsymbol{g}(\boldsymbol{x} + t\boldsymbol{h}) - \boldsymbol{g}(\boldsymbol{x})}{t} \right] \\
&= \alpha \boldsymbol{f}'(\boldsymbol{x};\boldsymbol{h}) + \beta \boldsymbol{g}'(\boldsymbol{x};\boldsymbol{h}),
\end{aligned}$$

$\alpha \boldsymbol{f} + \beta \boldsymbol{g}$ is directionally differentiable at $\boldsymbol{x}$. Therefore, $\alpha \boldsymbol{f} + \beta \boldsymbol{g} \in \mathcal{D}(E_1, E_2)$, as needed. The proposition follows immediately from above and the definition of homogenization. $\square$

Proposition 4.1.5. *Let $\boldsymbol{f} \in \mathcal{D}(E_1, E_2)$ and $\boldsymbol{x} \in E_1$. Then $H_{\boldsymbol{x}}[\boldsymbol{f}]$ is Lipschitz continuous.*

Proof. Let $\boldsymbol{f} \in \mathcal{D}(E_1, E_2)$ and $\boldsymbol{x} \in E_1$. Since $\boldsymbol{f}$ is locally Lipschitz continuous, there exists a neighborhood $U_{\boldsymbol{x}}$ of $\boldsymbol{x}$ on which $\boldsymbol{f}$ is Lipschitz continuous. Then, for all $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ in E_1,

$$\begin{aligned} \|\boldsymbol{f}'(\boldsymbol{x};\boldsymbol{h}_1) - \boldsymbol{f}'(\boldsymbol{x};\boldsymbol{h}_2)\|_2 &= \left\| \lim_{t\to 0^+} \frac{\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{h}_1) - \boldsymbol{f}(\boldsymbol{x})}{t} - \lim_{t\to 0^+} \frac{\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{h}_2) - \boldsymbol{f}(\boldsymbol{x})}{t} \right\|_2 \\ &= \left\| \lim_{t\to 0^+} \frac{\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{h}_1) - \boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{h}_2)}{t} \right\|_2 \\ &= \lim_{t\to 0^+} \frac{\|\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{h}_1) - \boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{h}_2)\|_2}{t} \\ &\le \lim_{t\to 0^+} \frac{K\,\|(\boldsymbol{x}+t\boldsymbol{h}_1) - (\boldsymbol{x}+t\boldsymbol{h}_2)\|_1}{t} \\ &= K\,\|\boldsymbol{h}_1 - \boldsymbol{h}_2\|_1 . \end{aligned}$$

Therefore, $H_{\boldsymbol{x}}[\boldsymbol{f}]$ is Lipschitz continuous. $\square$

Proposition 4.1.6 (The chain rule for directional derivatives)**.** *If $\boldsymbol{f} \in \mathcal{D}(E_1, E_2)$ and $\boldsymbol{F} \in \mathcal{D}(E_2, E_3)$, then $\boldsymbol{F} \circ \boldsymbol{f} \in \mathcal{D}(E_1, E_3)$. Moreover,*

$$H_{\boldsymbol{x}}[\boldsymbol{F} \circ \boldsymbol{f}](\cdot) = H_{\boldsymbol{f}(\boldsymbol{x})}[\boldsymbol{F}](H_{\boldsymbol{x}}[\boldsymbol{f}](\cdot)).$$

Proof. Let $\boldsymbol{f} \in \mathcal{D}(E_1, E_2)$ and $\boldsymbol{F} \in \mathcal{D}(E_2, E_3)$. First, we show that $\boldsymbol{F} \circ \boldsymbol{f} \in \mathcal{D}(E_1, E_3)$. Let $\boldsymbol{z} \in E_1$. Since $\boldsymbol{f} \in \mathcal{D}(E_1, E_2)$ and $\boldsymbol{F} \in \mathcal{D}(E_2, E_3)$, there exist a neighborhood $U_{\boldsymbol{z}}$ of $\boldsymbol{z}$ on which $\boldsymbol{f}$ is Lipschitz continuous and a neighborhood $V_{\boldsymbol{f}(\boldsymbol{z})}$ of $\boldsymbol{f}(\boldsymbol{z})$ on which $\boldsymbol{F}$ is Lipschitz continuous. Clearly, $\boldsymbol{z} \in U_{\boldsymbol{z}} \cap \boldsymbol{f}^{-1}\left(V_{\boldsymbol{f}(\boldsymbol{z})}\right)$. Note that $\boldsymbol{f}$ is continuous since it is locally Lipschitz continuous. Since the preimage of every open set under a continuous function is open, $\boldsymbol{f}^{-1}\left(V_{\boldsymbol{f}(\boldsymbol{z})}\right)$ is open. Hence, $U_{\boldsymbol{z}} \cap \boldsymbol{f}^{-1}\left(V_{\boldsymbol{f}(\boldsymbol{z})}\right)$ is a neighborhood of $\boldsymbol{z}$. For all $\boldsymbol{x}$ and $\boldsymbol{y}$ in $U_{\boldsymbol{z}} \cap \boldsymbol{f}^{-1}\left(V_{\boldsymbol{f}(\boldsymbol{z})}\right)$,

$$\|\boldsymbol{F}(\boldsymbol{f}(\boldsymbol{x})) - \boldsymbol{F}(\boldsymbol{f}(\boldsymbol{y}))\|_3 \le K_1\,\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})\|_2 \le K_1 K_2\,\|\boldsymbol{x} - \boldsymbol{y}\|_1 .$$

Therefore, $\boldsymbol{F} \circ \boldsymbol{f}$ is locally Lipschitz continuous. Let $\boldsymbol{x} \in E_1$ and $\boldsymbol{h} \in E_1$. To show that $\boldsymbol{F} \circ \boldsymbol{f}$ is directionally differentiable on E_1, it is enough to show that

$$(\boldsymbol{F} \circ \boldsymbol{f})'(\boldsymbol{x}; \boldsymbol{h}) = \boldsymbol{F}'(\boldsymbol{f}(\boldsymbol{x}); \boldsymbol{f}'(\boldsymbol{x}; \boldsymbol{h})).$$

Since $\boldsymbol{f}$ and $\boldsymbol{F}$ are directionally differentiable on E_1 and E_2, respectively,

$$\boldsymbol{f}(\boldsymbol{x} + \alpha \boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}'(\boldsymbol{x}; \boldsymbol{h})\alpha + \boldsymbol{\epsilon}_{\boldsymbol{x},\boldsymbol{h}}(\alpha)\alpha,$$

$$\boldsymbol{F}(\boldsymbol{f}(\boldsymbol{x}) + \beta \boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{f}(\boldsymbol{x})) = \boldsymbol{F}'(\boldsymbol{f}(\boldsymbol{x}); \boldsymbol{h})\beta + \boldsymbol{\eta}_{\boldsymbol{x},\boldsymbol{h}}(\beta)\beta,$$

where $\boldsymbol{\epsilon}_{\boldsymbol{x},\boldsymbol{h}} : \mathbb{R}^+ \to E_2$ and $\boldsymbol{\eta}_{\boldsymbol{x},\boldsymbol{h}} : \mathbb{R}^+ \to E_3$ are functions such that

$$\lim_{\alpha \downarrow 0} \boldsymbol{\epsilon}_{\boldsymbol{x},\boldsymbol{h}}(\alpha) = \boldsymbol{0},$$

$$\lim_{\beta \downarrow 0} \boldsymbol{\eta}_{\boldsymbol{x},\boldsymbol{h}}(\beta) = \boldsymbol{0}.$$

Hence,

$$\begin{aligned}\boldsymbol{F}(\boldsymbol{f}(\boldsymbol{x} + \alpha \boldsymbol{h})) - \boldsymbol{F}(\boldsymbol{f}(\boldsymbol{x})) &= \boldsymbol{F}(\boldsymbol{f}(\boldsymbol{x}) + \alpha \boldsymbol{k}_{\boldsymbol{x},\boldsymbol{h}}(\alpha)) - \boldsymbol{F}(\boldsymbol{f}(\boldsymbol{x})) \\ &= \boldsymbol{F}'(\boldsymbol{f}(\boldsymbol{x}); \boldsymbol{k}_{\boldsymbol{x},\boldsymbol{h}}(\alpha))\alpha + \boldsymbol{\eta}_{\boldsymbol{x},\boldsymbol{k}_{\boldsymbol{x},\boldsymbol{h}}(\alpha)}(\alpha)\alpha,\end{aligned}$$

where $\boldsymbol{k}_{\boldsymbol{x},\boldsymbol{h}} : \mathbb{R}^+ \to E_2$ is a function given by

$$\boldsymbol{k}_{\boldsymbol{x},\boldsymbol{h}}(\cdot) = \boldsymbol{f}'(\boldsymbol{x}; \boldsymbol{h}) + \boldsymbol{\epsilon}_{\boldsymbol{x},\boldsymbol{h}}(\cdot).$$

$H_{\boldsymbol{f}(\boldsymbol{x})}[\boldsymbol{F}]$ is continuous since it is Lipschitz continuous by Proposition 4.1.5. Hence,

$$\begin{aligned}\lim_{\alpha \downarrow 0} \frac{\boldsymbol{F}(\boldsymbol{f}(\boldsymbol{x} + \alpha \boldsymbol{h})) - \boldsymbol{F}(\boldsymbol{f}(\boldsymbol{x}))}{\alpha} &= \lim_{\alpha \downarrow 0} \Big(\boldsymbol{F}'(\boldsymbol{f}(\boldsymbol{x}); \boldsymbol{k}_{\boldsymbol{x},\boldsymbol{h}}(\alpha)) + \boldsymbol{\eta}_{\boldsymbol{x},\boldsymbol{k}_{\boldsymbol{x},\boldsymbol{h}}(\alpha)}(\alpha) \Big) \\ &= \lim_{\alpha \downarrow 0} \boldsymbol{F}'\left(\boldsymbol{f}(\boldsymbol{x}); \boldsymbol{k}_{\boldsymbol{x},\boldsymbol{h}}(\alpha)\right) \\ &= \boldsymbol{F}'(\boldsymbol{f}(\boldsymbol{x}); \lim_{\alpha \downarrow 0} \boldsymbol{k}_{\boldsymbol{x},\boldsymbol{h}}(\alpha)) \\ &= \boldsymbol{F}'(\boldsymbol{f}(\boldsymbol{x}); \boldsymbol{f}'(\boldsymbol{x}; \boldsymbol{h})),\end{aligned}$$

as needed. The proposition follows immediately from above and the definition of homogenization. □

Example 4.1.7. *Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined by $f(x_1\ x_2)^T = \min(x_1, x_2)$ and $h : \mathbb{R} \to \mathbb{R}$ the function defined by $h(y) = \max(0, y)$. Then*

$$h'(y; d) = \lim_{\alpha \downarrow 0} \frac{h(y + \alpha d) - h(y)}{\alpha}$$
$$= \begin{cases} 0 & \text{if } y < 0, \text{ or } y = 0 \text{ and } d \leq 0, \\ d & \text{if } y > 0, \text{ or } y = 0 \text{ and } d > 0, \end{cases}$$

and

$$f'((x_1\ x_2)^T; (d_1\ d_2)^T) = \lim_{\alpha \downarrow 0} \frac{f\left((x_1\ x_2)^T + \alpha(d_1\ d_2)^T\right) - f(x_1\ x_2)^T}{\alpha}$$
$$= \begin{cases} d_1 & \text{if } x_1 < x_2, \text{ or } x_1 = x_2 \text{ and } d_1 \leq d_2, \\ d_2 & \text{if } x_1 > x_2, \text{ or } x_1 = x_2 \text{ and } d_1 > d_2. \end{cases}$$

Hence, $h'(0; d) = \max(0, d)$ and $f'((0\ 0)^T; (d_1\ d_2)^T) = \min(d_1, d_2)$. On the other hand,

$$[h \circ f]'\left((0\ 0)^T; (d_1\ d_2)^T\right) = \lim_{\alpha \downarrow 0} \frac{g\left((0\ 0)^T + \alpha(d_1\ d_2)^T\right) - g(0\ 0)^T}{\alpha}$$
$$= \begin{cases} d_1 & \text{if } 0 \leq d_1 \leq d_2, \\ d_2 & \text{if } d_1 > d_2 > 0, \\ 0 & \text{otherwise}, \end{cases}$$
$$= \max(0, \min(d_1, d_2)).$$

Thus, $[h \circ f]'\left((0\ 0)^T; (d_1\ d_2)^T\right) = h'\left(f(0\ 0)^T; f'\left((0\ 0)^T; (d_1\ d_2)^T\right)\right)$, as expected.

Definition 4.1.8. *A function $\boldsymbol{f} : E_1 \to E_2$ is called homogeneous if for all $\alpha \in \mathbb{R}$ and all $\boldsymbol{x} \in E_1$,*

$$\boldsymbol{f}(\alpha \boldsymbol{x}) = \alpha \boldsymbol{f}(\boldsymbol{x}).$$

Remark 4.1.9. *Let a function $\boldsymbol{f} : E_1 \to E_2$ be directionally differentiable on E_1. If $\boldsymbol{x} \in E_1$, then the directional derivative of $\boldsymbol{f}$ at $\boldsymbol{x}$ is homogeneous since*

$$\boldsymbol{f}'(\boldsymbol{x}; \boldsymbol{0}) = \boldsymbol{0}$$

and if $\boldsymbol{h} \in E_1$ *and* $\alpha \neq 0$, *then*

$$\begin{aligned}
\boldsymbol{f}'(\boldsymbol{x}; \alpha \boldsymbol{h}) &= \lim_{t \to 0^+} \frac{\boldsymbol{f}(\boldsymbol{x} + t\alpha \boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{x})}{t} \\
&= \lim_{t \to 0^+} \alpha \frac{\boldsymbol{f}(\boldsymbol{x} + \alpha t \boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{x})}{\alpha t} \\
&= \alpha \lim_{t \to 0^+} \frac{\boldsymbol{f}(\boldsymbol{x} + t \boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{x})}{t} \\
&= \alpha \boldsymbol{f}'(\boldsymbol{x}; \boldsymbol{h}).
\end{aligned}$$

Proposition 4.1.10. *Let* $\boldsymbol{f} \in \mathcal{D}(E_1, E_2)$ *be a homogeneous function. Then for any* $\boldsymbol{x} \in E_1$,

$$H_{\boldsymbol{x}}[H_{\boldsymbol{x}}[\boldsymbol{f}]] = H_{\boldsymbol{x}}[\boldsymbol{f}].$$

Moreover,

$$H_{\tau \boldsymbol{x}}[\boldsymbol{f}] = H_{\boldsymbol{x}}[\boldsymbol{f}] \text{ for any } \tau > 0,$$

$$H_{\boldsymbol{x}}[\boldsymbol{f}](\alpha \boldsymbol{x}) = \alpha \boldsymbol{f}(\boldsymbol{x}) \text{ for any } \alpha \in \mathbb{R}.$$

Finally, for any $\boldsymbol{x}, \boldsymbol{y} \in E_1$ *and any* $\alpha \in \mathbb{R}$,

$$H_{\boldsymbol{x}}[\boldsymbol{f}](\boldsymbol{y} + \alpha \boldsymbol{x}) = H_{\boldsymbol{x}}[\boldsymbol{f}](\boldsymbol{y}) + \alpha \boldsymbol{f}(\boldsymbol{x}).$$

Proof. Let $\boldsymbol{f} \in \mathcal{D}(E_1, E_2)$ be a homogeneous function, $\boldsymbol{x} \in E_1$, and $\alpha \in \mathbb{R}$. Since $\boldsymbol{f}$ is homogeneous,

$$\begin{aligned}
H_{\boldsymbol{x}}[\boldsymbol{f}](\alpha \boldsymbol{x}) &= \lim_{t \to 0^+} \frac{\boldsymbol{f}(\boldsymbol{x} + t\alpha \boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})}{t} \\
&= \lim_{t \to 0^+} \frac{\boldsymbol{f}((1 + t\alpha)\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})}{t} \\
&= \lim_{t \to 0^+} \frac{(1 + t\alpha)\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})}{t} \\
&= \alpha \boldsymbol{f}(\boldsymbol{x}).
\end{aligned}$$

For all $\boldsymbol{h} \in E_1$,

$$\begin{aligned}
H_{\boldsymbol{x}}[H_{\boldsymbol{x}}[\boldsymbol{f}]](\boldsymbol{h}) &= \lim_{t \to 0^+} \frac{H_{\boldsymbol{x}}[\boldsymbol{f}](\boldsymbol{x} + t\boldsymbol{h}) - H_{\boldsymbol{x}}[\boldsymbol{f}](\boldsymbol{x})}{t} \\
&= \lim_{t \to 0^+} \frac{1}{t} \left(\lim_{s \to 0^+} \frac{\boldsymbol{f}\left(\boldsymbol{x} + s\left(\boldsymbol{x} + t\boldsymbol{h}\right)\right) - \boldsymbol{f}(\boldsymbol{x} + s\boldsymbol{x})}{s} \right) \\
&= \lim_{t \to 0^+} \frac{1}{t} \left(\lim_{s \to 0^+} \frac{\boldsymbol{f}\left(\boldsymbol{x} + s(1+s)^{-1} t\boldsymbol{h}\right) - \boldsymbol{f}(\boldsymbol{x})}{s(1+s)^{-1}} \right) \\
&= \lim_{t \to 0^+} \frac{\boldsymbol{f}'(\boldsymbol{x}; t\boldsymbol{h})}{t} \\
&= \boldsymbol{f}'(\boldsymbol{x}; \boldsymbol{h}) \\
&= H_{\boldsymbol{x}}[\boldsymbol{f}](\boldsymbol{h}).
\end{aligned}$$

For $\tau > 0$ and $\boldsymbol{h} \in E_1$,

$$\begin{aligned}
H_{\tau\boldsymbol{x}}[\boldsymbol{f}](\boldsymbol{h}) &= \boldsymbol{f}'(\tau\boldsymbol{x}; \boldsymbol{h}) \\
&= \lim_{t \to 0^+} \frac{\boldsymbol{f}(\tau\boldsymbol{x} + t\boldsymbol{h}) - \boldsymbol{f}(\tau\boldsymbol{x})}{t} \\
&= \lim_{t \to 0^+} \tau \frac{\boldsymbol{f}(\boldsymbol{x} + \frac{t}{\tau}\boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{x})}{t} \\
&= \lim_{t \to 0^+} \frac{\boldsymbol{f}(\boldsymbol{x} + \frac{t}{\tau}\boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{x})}{\frac{t}{\tau}} \\
&= \boldsymbol{f}'(\boldsymbol{x}; \boldsymbol{h}).
\end{aligned}$$

Finally, for $\boldsymbol{x}, \boldsymbol{y} \in E_1$ and $\alpha \in \mathbb{R}$,

$$\begin{aligned}
H_{\boldsymbol{x}}[\boldsymbol{f}](\boldsymbol{y}) &= \lim_{t \to 0^+} \frac{\boldsymbol{f}(\boldsymbol{x} + t\boldsymbol{y}) - \boldsymbol{f}(\boldsymbol{x})}{t} \\
&= \lim_{t \to 0^+} \frac{\boldsymbol{f}((1 - t\alpha)\boldsymbol{x} + t(\boldsymbol{y} + \alpha\boldsymbol{x})) - \boldsymbol{f}(\boldsymbol{x})}{t} \\
&= \lim_{t \to 0^+} \frac{1}{t} \left((1 - t\alpha)\boldsymbol{f}\left(\boldsymbol{x} + \frac{t}{1 - t\alpha}(\boldsymbol{y} + \alpha\boldsymbol{x})\right) - \boldsymbol{f}(\boldsymbol{x}) \right) \\
&= \lim_{t \to 0^+} \frac{1}{t} \left((1 - t\alpha)\boldsymbol{f}\left(\boldsymbol{x} + \frac{t}{1 - t\alpha}(\boldsymbol{y} + \alpha\boldsymbol{x})\right) - (1 - t\alpha)\boldsymbol{f}(\boldsymbol{x}) - t\alpha\boldsymbol{f}(\boldsymbol{x}) \right) \\
&= \lim_{t \to 0^+} \frac{1}{t} \left((1 - t\alpha)\boldsymbol{f}\left(\boldsymbol{x} + \frac{t}{1 - t\alpha}(\boldsymbol{y} + \alpha\boldsymbol{x})\right) - (1 - t\alpha)\boldsymbol{f}(\boldsymbol{x}) \right) - \alpha\boldsymbol{f}(\boldsymbol{x}) \\
&= \lim_{t \to 0^+} \frac{\boldsymbol{f}\left(\boldsymbol{x} + \frac{t}{1-t\alpha}(\boldsymbol{y} + \alpha\boldsymbol{x})\right) - \boldsymbol{f}(\boldsymbol{x})}{\frac{t}{1-t\alpha}} - \alpha\boldsymbol{f}(\boldsymbol{x}) \\
&= \boldsymbol{f}'(\boldsymbol{x}; \boldsymbol{y} + \alpha\boldsymbol{x}) - \alpha\boldsymbol{f}(\boldsymbol{x}) \\
&= H_{\boldsymbol{x}}[\boldsymbol{f}](\boldsymbol{y} + \alpha\boldsymbol{x}) - \alpha\boldsymbol{f}(\boldsymbol{x}).
\end{aligned}$$

□

4.1.3 Lexicographically Smooth Functions

Nesterov's lexicographic differentiation [9] is developed for the class of lexicographically (L-) smooth functions. The definitions and results here are drawn from that work. The proofs here, due to Nesterov, have been expanded for ease of comprehension.

Definition 4.1.11. *Let* $\boldsymbol{f} \in \mathcal{D}(E_1, E_2)$*,* $\boldsymbol{x} \in E_1$*, and* $\boldsymbol{U} = \{\boldsymbol{u}_k\}_{k=1}^{\infty}$*, a sequence of vectors, called directions, in* E_1*. The sequence of the recursively defined functions*

$$\boldsymbol{f}^{(0)}_{\boldsymbol{x},\boldsymbol{U}} = H_{\boldsymbol{x}}[\boldsymbol{f}],$$
$$\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}} = H_{\boldsymbol{u}_k}[\boldsymbol{f}^{(k-1)}_{\boldsymbol{x},\boldsymbol{U}}], \quad k \in \mathbb{N},$$

is called the homogenization sequence of $\boldsymbol{f}$ *generated by* $\boldsymbol{x}$ *and* $\boldsymbol{U}$*, if it exists.*

Definition 4.1.12. *A function* $\boldsymbol{f} \in \mathcal{D}(E_1, E_2)$ *is lexicographically smooth on* E_1*, or L-smooth on* E_1 *for short, if its homogenization sequence exists for any* $\boldsymbol{x} \in E_1$ *and any sequence* $\boldsymbol{U}$ *of directions in* E_1*. If* $\boldsymbol{f}$ *is L-smooth on* E_1*, we write* $\boldsymbol{f} \in \mathcal{L}(E_1, E_2)$*.*

Remark 4.1.13. *The class of L-smooth functions contains convex functions, differentiable functions, and compositions of L-smooth functions. In the case of* $E_1 = \mathbb{R}^n$ *and* $E_2 = \mathbb{R}^m$*, the class of L-smooth functions also contains* $\mathcal{PC}^1$ *functions.*

Proposition 4.1.14. *Let* $\boldsymbol{f} \in \mathcal{L}(E_1, E_2)$*,* $\boldsymbol{x} \in E_1$*, and* $\boldsymbol{U} = \{\boldsymbol{u}_k\}_{k=1}^{\infty}$*, a sequence of directions in* E_1*. Then* $\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}$ *is Lipschitz continuous for each* $k \in \{0, 1, 2, \dots\}$*.*

Proof. By Proposition 4.1.5, $\boldsymbol{f}^{(0)}_{\boldsymbol{x},\boldsymbol{U}}$ is Lipschitz continuous. Suppose that $\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}$ is Lipchitz continuous. Then, for all $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ in E_1,

$$\begin{aligned}
&\left\|\boldsymbol{f}^{(k+1)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h}_1) - \boldsymbol{f}^{(k+1)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h}_2)\right\|_2 \\
&= \left\| \lim_{t\to 0^+} \frac{\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_{k+1} + t\boldsymbol{h}_1) - \boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_{k+1})}{t} - \lim_{t\to 0^+} \frac{\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_{k+1} + t\boldsymbol{h}_2) - \boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_{k+1})}{t} \right\|_2 \\
&= \left\| \lim_{t\to 0^+} \frac{\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_{k+1} + t\boldsymbol{h}_1) - \boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_{k+1} + t\boldsymbol{h}_2)}{t} \right\|_2 \\
&= \lim_{t\to 0^+} \frac{\left\|\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_{k+1} + t\boldsymbol{h}_1) - \boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_{k+1} + t\boldsymbol{h}_2)\right\|_2}{t} \\
&\leq \lim_{t\to 0^+} \frac{K\left\|(\boldsymbol{u}_{k+1} + t\boldsymbol{h}_1) - (\boldsymbol{u}_{k+1} + t\boldsymbol{h}_2)\right\|_1}{t} \\
&= K\left\|\boldsymbol{h}_1 - \boldsymbol{h}_2\right\|_1.
\end{aligned}$$

Therefore, the proposition follows from the principle of induction. □

Proposition 4.1.15. *Let $\boldsymbol{f}, \boldsymbol{g} \in \mathcal{L}(E_1, E_2)$. Then, for any $\alpha, \beta \in \mathbb{R}$,*

$$\alpha\boldsymbol{f} + \beta\boldsymbol{g} \in \mathcal{L}(E_1, E_2).$$

Proof. Let $\boldsymbol{f}, \boldsymbol{g} \in \mathcal{L}(E_1, E_2)$. Hence, $\boldsymbol{f}, \boldsymbol{g} \in \mathcal{D}(E_1, E_2)$. Then, $\alpha\boldsymbol{f} + \beta\boldsymbol{g} \in \mathcal{D}(E_1, E_2)$ by Proposition 4.1.4. Let $\boldsymbol{x} \in E_1$ and $\boldsymbol{U} = \{\boldsymbol{u}_k\}_{k=1}^{\infty}$ a sequence of directions in E_1. It suffices to show that the homogenization sequence of $\alpha\boldsymbol{f} + \beta\boldsymbol{g}$ generated by $\boldsymbol{x}$ and $\boldsymbol{U}$ exists. We claim that

$$(\alpha\boldsymbol{f} + \beta\boldsymbol{g})^{(k)}_{\boldsymbol{x},\boldsymbol{U}} = \alpha\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}} + \beta\boldsymbol{g}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}, \quad k \in \{0, 1, 2, \dots\}.$$

Since

$$(\boldsymbol{\alpha}\boldsymbol{f} + \boldsymbol{\beta}\boldsymbol{g})^{(0)}_{\boldsymbol{x},\boldsymbol{U}} = H_{\boldsymbol{x}}[\alpha\boldsymbol{f} + \beta\boldsymbol{g}] = \alpha H_{\boldsymbol{x}}[\boldsymbol{f}] + \beta H_{\boldsymbol{x}}[\boldsymbol{g}]$$

by Proposition 4.1.4, the statement holds when $k = 0$. Suppose that the k^{th} statement is true. Then, for all $\boldsymbol{h} \in E_1$,

$$
\begin{aligned}
&(\alpha \boldsymbol{f} + \beta \boldsymbol{g})^{(k+1)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h}) \\
&= ((\alpha \boldsymbol{f} + \beta \boldsymbol{g})^{(k)}_{\boldsymbol{x},\boldsymbol{U}})'(\boldsymbol{u}_{k+1}; \boldsymbol{h}) \\
&= \lim_{t \to 0^+} \frac{(\alpha \boldsymbol{f} + \beta \boldsymbol{g})^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_{k+1} + t\boldsymbol{h}) - (\alpha \boldsymbol{f} + \beta \boldsymbol{g})^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_{k+1})}{t} \\
&= \lim_{t \to 0^+} \left[\alpha \frac{\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_{k+1} + t\boldsymbol{h}) - \boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_{k+1})}{t} + \beta \frac{\boldsymbol{g}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_{k+1} + t\boldsymbol{h}) - \boldsymbol{g}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_{k+1})}{t} \right] \\
&= \alpha \boldsymbol{f}^{(k+1)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h}) + \beta \boldsymbol{g}^{(k+1)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h}).
\end{aligned}
$$

Hence, the claim follows from the principle of induction. Thus, the homogenization sequence of $\alpha \boldsymbol{f} + \beta \boldsymbol{g}$ exists for any $\boldsymbol{x} \in E_1$ and any sequence $\boldsymbol{U} = \{\boldsymbol{u_k}\}_{k=1}^{\infty}$ of directions in E_1. Therefore, $\alpha \boldsymbol{f} + \beta \boldsymbol{g} \in \mathcal{L}(E_1, E_2)$. □

The homogenization sequence functions satisfy an equality-based chain rule.

Theorem 4.1.16. *If $\boldsymbol{f} \in \mathcal{L}(E_1, E_2)$ and $\boldsymbol{F} \in \mathcal{L}(E_2, E_3)$, then $\boldsymbol{F} \circ \boldsymbol{f} \in \mathcal{L}(E_1, E_3)$. Moreover, for any $\boldsymbol{x} \in E_1$ and any sequence $\boldsymbol{U} = \{\boldsymbol{u}_k\}_{k=1}^{\infty}$ of directions in E_1,*

$$(\boldsymbol{F} \circ \boldsymbol{f})^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\cdot) = \boldsymbol{F}^{(k)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}} \left(\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\cdot) \right), \quad k \in \{0, 1, 2, \ldots\},$$

where the sequence $\boldsymbol{V} = \{\boldsymbol{v}_k\}_{k=1}^{\infty}$ of directions in E_2 is given by

$$\boldsymbol{v}_k = \boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_k), \quad k \in \mathbb{N}.$$

Proof. Let $\boldsymbol{f} \in \mathcal{L}(E_1, E_2)$ and $\boldsymbol{F} \in \mathcal{L}(E_2, E_3)$. Then $\boldsymbol{f} \in \mathcal{D}(E_1, E_2)$ and $\boldsymbol{F} \in \mathcal{D}(E_2, E_3)$. Thus, $\boldsymbol{F} \circ \boldsymbol{f} \in \mathcal{D}(E_1, E_3)$ by Proposition 4.1.6. We claim that if $\boldsymbol{x} \in E_1$ and $\boldsymbol{U} = \{\boldsymbol{u}_k\}_{k=1}^{\infty}$, a sequence of directions in E_1, then

$$(\boldsymbol{F} \circ \boldsymbol{f})^{(j)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h}) = \boldsymbol{F}^{(j)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}} \left(\boldsymbol{f}^{(j)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h}) \right), \quad j \in \{0, 1, 2, \ldots\}$$

for all $\boldsymbol{h} \in E_1$, where the sequence $\boldsymbol{V} = \{\boldsymbol{v}_k\}_{k=1}^{\infty}$ of directions in E_2 is given by

$$\boldsymbol{v}_k = \boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_k), \quad k \in \mathbb{N}.$$

By Proposition 4.1.6,

$$(\boldsymbol{F}\circ\boldsymbol{f})'(\boldsymbol{x};\boldsymbol{h}) = \boldsymbol{F}'\left(\boldsymbol{f}(\boldsymbol{x});\boldsymbol{f}'(\boldsymbol{x};\boldsymbol{h})\right) \text{ for any } \boldsymbol{h}\in E_1,$$

so

$$(\boldsymbol{F}\circ\boldsymbol{f})^{(0)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h}) = \boldsymbol{F}^{(0)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}}\left(\boldsymbol{f}^{(0)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h})\right) \text{ for any } \boldsymbol{h}\in E_1,$$

which confirms that the statement is true when $j=0$. By Propositions 4.1.6 and 4.1.10,

$$\begin{aligned}
(\boldsymbol{F}\circ\boldsymbol{f})^{(1)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h}) &= H_{\boldsymbol{u}_1}\left[(\boldsymbol{F}\circ\boldsymbol{f})^{(0)}_{\boldsymbol{x},\boldsymbol{U}}\right](\boldsymbol{h}) \\
&= H_{\boldsymbol{u}_1}\left[\boldsymbol{F}^{(0)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}}\circ\boldsymbol{f}^{(0)}_{\boldsymbol{x},\boldsymbol{U}}\right](\boldsymbol{h}) \\
&= H_{\boldsymbol{f}^{(0)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_1)}\left[\boldsymbol{F}^{(0)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}}\right]\left(H_{\boldsymbol{u}_1}\left[\boldsymbol{f}^{(0)}_{\boldsymbol{x},\boldsymbol{U}}\right](\boldsymbol{h})\right) \\
&= H_{\boldsymbol{f}^{(0)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_1)}\left[\boldsymbol{F}^{(0)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}}\right]\left(\boldsymbol{f}^{(1)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h})\right) \\
&= \left(\boldsymbol{F}^{(0)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}}\right)'\left(\boldsymbol{f}^{(0)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_1);\boldsymbol{f}^{(1)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h})\right) \\
&= \left(\boldsymbol{F}^{(0)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}}\right)'\left(\boldsymbol{f}^{(1)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_1);\boldsymbol{f}^{(1)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h})\right) \\
&= \left(\boldsymbol{F}^{(0)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}}\right)'\left(\boldsymbol{v}_1;\boldsymbol{f}^{(1)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h})\right) \\
&= \boldsymbol{F}^{(1)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}}\left(\boldsymbol{f}^{(1)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h})\right).
\end{aligned}$$

Hence, the statement is true when $j=1$. Now, suppose that

$$(\boldsymbol{F}\circ\boldsymbol{f})^{(j)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h}) = \boldsymbol{F}^{(j)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}}\circ\boldsymbol{f}^{(j)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h})$$

for all $\boldsymbol{h}\in E_1$, where the sequence $\boldsymbol{V}=\{\boldsymbol{v}_k\}_{k=1}^{j}$ of directions in E_2 is given by

$$\boldsymbol{v}_k = \boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_k), \quad k\in\{1,\cdots,j\}.$$

By Propositions 4.1.6 and 4.1.10,

$$\begin{aligned}
(\boldsymbol{F}\circ\boldsymbol{f})^{(k+1)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h}) &= H_{\boldsymbol{u}_{k+1}}\left[(\boldsymbol{F}\circ\boldsymbol{f})^{(k)}_{\boldsymbol{x},\boldsymbol{U}}\right](\boldsymbol{h}) \\
&= H_{\boldsymbol{u}_{k+1}}\left[\boldsymbol{F}^{(k)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}}\circ\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}\right](\boldsymbol{h}) \\
&= H_{\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_{k+1})}\left[\boldsymbol{F}^{(k)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}}\right]\left(H_{\boldsymbol{u}_{k+1}}\left[\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}\right](\boldsymbol{h})\right) \\
&= H_{\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_{k+1})}\left[\boldsymbol{F}^{(k)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}}\right]\left(\boldsymbol{f}^{(k+1)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h})\right) \\
&= \left(\boldsymbol{F}^{(k)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}}\right)'\left(\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_{k+1});\boldsymbol{f}^{(k+1)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h})\right) \\
&= \left(\boldsymbol{F}^{(k)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}}\right)'\left(\boldsymbol{f}^{(k+1)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_{k+1});\boldsymbol{f}^{(k+1)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h})\right) \\
&= \left(\boldsymbol{F}^{(k)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}}\right)'\left(\boldsymbol{v}_{k+1};\boldsymbol{f}^{(k+1)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h})\right) \\
&= \boldsymbol{F}^{(k+1)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}}\left(\boldsymbol{f}^{(k+1)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h})\right).
\end{aligned}$$

Hence, the theorem follows from the principle of induction. □

Definition 4.1.17. *Let $\boldsymbol{U}$ be a sequence of directions in E_1. For any $k\in\mathbb{N}$, we denote the span of the first k directions of $\boldsymbol{U}$ by $L_k(\boldsymbol{U})$. Let $L_0(\boldsymbol{U}) := \{0\}$.*

Proposition 4.1.18. *Let $\boldsymbol{f}\in\mathcal{L}(E_1,E_2)$, $\boldsymbol{x}\in E_1$, and $\boldsymbol{U}=\{\boldsymbol{u}_k\}_{k=1}^{\infty}$, a sequence of directions in E_1. Then the members of the homogenization sequence of $\boldsymbol{f}$ generated by $\boldsymbol{x}$ and $\boldsymbol{U}$ satisfy the following:*

$$\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\tau\boldsymbol{h}) = \tau\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h}),\ \text{where } \boldsymbol{h}\in E_1 \text{ and } \tau\geq 0,$$

$$\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h}+\alpha\boldsymbol{d}) = \boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h})+\alpha\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{d}),\ \text{where } \boldsymbol{h}\in E_1,\ \boldsymbol{d}\in L_k(\boldsymbol{U}),\ \text{and } \alpha\in\mathbb{R},$$

$$\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h}) = \boldsymbol{f}^{(k-1)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h}),\ \text{where } \boldsymbol{h}\in L_{k-1}(\boldsymbol{U}).$$

Proof. Let $\boldsymbol{f}\in\mathcal{L}(E_1,E_2)$, $\boldsymbol{x}\in E_1$, and $\boldsymbol{U}=\{\boldsymbol{u}_k\}_{k=1}^{\infty}$, a sequence of directions in E_1. The first statement in the proposition follows from the Remark on Definition 4.1.8. We show that the second statement is true using the principle of induction. It is clear that the statement holds when $k=0$. Suppose that the k^{th} statement is true. Let $\boldsymbol{h}\in E_1$,

$\boldsymbol{d} \in L_{k+1}(\boldsymbol{U})$, and $\alpha \in \mathbb{R}$. Then $\boldsymbol{d} = \boldsymbol{d}_k + \beta \boldsymbol{u}_{k+1}$ for some $\beta \in \mathbb{R}$, where $\boldsymbol{d}_k \in L_k(\boldsymbol{U})$. Then

$$\begin{aligned}
H_{\boldsymbol{u}_{k+1}}\left[\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}\right](\boldsymbol{h}+\alpha\boldsymbol{d}_k) &= \left(\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}\right)'(\boldsymbol{u}_{k+1};\boldsymbol{h}+\alpha\boldsymbol{d}_k) \\
&= \lim_{t\to 0^+} \frac{\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}\left(\boldsymbol{u}_{k+1}+t(\boldsymbol{h}+\alpha\boldsymbol{d}_k)\right) - \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{u}_{k+1})}{t} \\
&= \lim_{t\to 0^+} \frac{\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{u}_{k+1}+t\boldsymbol{h}) + t\alpha\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{d}_k) - \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{u}_{k+1})}{t} \\
&= \lim_{t\to 0^+} \frac{\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{u}_{k+1}+t\boldsymbol{h}) - \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{u}_{k+1})}{t} + \alpha\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{d}_k) \\
&= \left(\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}\right)'(\boldsymbol{u}_{k+1};\boldsymbol{h}) + \alpha\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{d}_k) \\
&= H_{\boldsymbol{u}_{k+1}}\left[\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}\right](\boldsymbol{h}) + \alpha\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{d}_k).
\end{aligned}$$

Note that when $\boldsymbol{h} = \boldsymbol{0}$ and $\alpha = 1$,

$$\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k+1)}(\boldsymbol{d}_k) = \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{d}_k).$$

Then, by Proposition 4.1.10,

$$\begin{aligned}
\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k+1)}(\boldsymbol{h}+\alpha\boldsymbol{d}) &= \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k+1)}(\boldsymbol{h}+\alpha\boldsymbol{d}_k+\alpha\beta\boldsymbol{u}_{k+1}) \\
&= \left(\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}\right)'(\boldsymbol{u}_{k+1};\boldsymbol{h}+\alpha\boldsymbol{d}_k+\alpha\beta\boldsymbol{u}_{k+1}) \\
&= H_{\boldsymbol{u}_{k+1}}\left[\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}\right](\boldsymbol{h}+\alpha\boldsymbol{d}_k+\alpha\beta\boldsymbol{u}_{k+1}) \\
&= H_{\boldsymbol{u}_{k+1}}\left[\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}\right](\boldsymbol{h}+\alpha\boldsymbol{d}_k) + \alpha\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\beta\boldsymbol{u}_{k+1}) \\
&= H_{\boldsymbol{u}_{k+1}}\left[\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}\right](\boldsymbol{h}) + \alpha\left(\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{d}_k) + \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\beta\boldsymbol{u}_{k+1})\right) \\
&= \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k+1)}(\boldsymbol{h}) + \alpha\left(\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{k+1}(\boldsymbol{d}_k) + \beta\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{u}_{k+1})\right) \\
&= \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k+1)}(\boldsymbol{h}) + \alpha\left(H_{\boldsymbol{u}_{k+1}}\left[\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}\right](\boldsymbol{d}_k) + \beta\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{u}_{k+1})\right) \\
&= \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k+1)}(\boldsymbol{h}) + \alpha\left(H_{\boldsymbol{u}_{k+1}}\left[\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}\right](\boldsymbol{d}_k+\beta\boldsymbol{u}_{k+1})\right) \\
&= \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k+1)}(\boldsymbol{h}) + \alpha\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k+1)}(\boldsymbol{d}_k+\beta\boldsymbol{u}_{k+1}) \\
&= \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k+1)}(\boldsymbol{h}) + \alpha\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k+1)}(\boldsymbol{d}).
\end{aligned}$$

Therefore, the second statement follows from the principle of induction. We show that the third statement is true using the principle of induction. The statement holds trivially when

$k = 0$. Suppose that the k^{th} statement is true. Let $\boldsymbol{h} \in L_k(\boldsymbol{U})$. By the second statement,

$$\begin{aligned}
\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k+1)}(\boldsymbol{h}) &= \lim_{t \to 0^+} \frac{\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{u}_{k+1} + t\boldsymbol{h}) - \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{u}_{k+1})}{t} \\
&= \lim_{t \to 0^+} \frac{\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{u}_{k+1}) + t\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{h}) - \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{u}_{k+1})}{t} \\
&= \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{h}).
\end{aligned}$$

Therefore, the third statement follows from the principle of induction. □

This result allows us to establish linearity of the kth homogenization sequence function restricted to the span of the first k direction vectors for any $k \in \mathbb{N}$.

Corollary 4.1.19. *Let $\boldsymbol{f} \in \mathcal{L}(E_1, E_2)$, $\boldsymbol{x} \in E_1$, and $\boldsymbol{U} = \{\boldsymbol{u}_k\}_{k=1}^{\infty}$, a sequence of directions in E_1. Let $k \in \mathbb{N}$. Then, for any $\boldsymbol{h} = \sum_{i=1}^{k} \alpha_i \boldsymbol{u}_i \in L_k(\boldsymbol{U})$,*

$$\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{h}) = \sum_{i=1}^{k} \alpha_i \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{u}_i).$$

Proof. Let $\boldsymbol{f} \in \mathcal{L}(E_1, E_2)$, $\boldsymbol{x} \in E_1$, and $\boldsymbol{U} = \{\boldsymbol{u}_k\}_{k=1}^{\infty}$, a sequence of directions in E_1. We show the corollary using the principle of induction. Let $\boldsymbol{h} = \alpha_1 \boldsymbol{u}_1 \in L_1(\boldsymbol{U})$. Since

$$\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(1)}(\alpha_1 \boldsymbol{u}_1) = \alpha_1 \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(1)}(\boldsymbol{u}_1),$$

the statement holds when $k = 1$. Suppose that the k^{th} statement is true. Let $\boldsymbol{h} \in L_{k+1}(\boldsymbol{U})$. Then $\boldsymbol{h} = \boldsymbol{h}_k + \alpha_{k+1}\boldsymbol{u}_{k+1}$ for some $\alpha_{k+1} \in \mathbb{R}$, where $\boldsymbol{h}_k = \sum_{i=1}^{k} \alpha_i \boldsymbol{u}_i \in L_k(\boldsymbol{U})$. Then, by Proposition 4.1.18,

$$\begin{aligned}
\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k+1)}(\boldsymbol{h}) &= \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k+1)}(\boldsymbol{h}_k + \alpha_{k+1}\boldsymbol{u}_{k+1}) \\
&= \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k+1)}(\boldsymbol{h}_k) + \alpha_{k+1}\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k+1)}(\boldsymbol{u}_{k+1}) \\
&= \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{h}_k) + \alpha_{k+1}\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k+1)}(\boldsymbol{u}_{k+1}) \\
&= \sum_{i=1}^{k} \alpha_i \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{u}_i) + \alpha_{k+1}\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k+1)}(\boldsymbol{u}_{k+1}) \\
&= \sum_{i=1}^{k} \alpha_i \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k+1)}(\boldsymbol{u}_i) + \alpha_{k+1}\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k+1)}(\boldsymbol{u}_{k+1}) \\
&= \sum_{i=1}^{k+1} \alpha_i \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k+1)}(\boldsymbol{u}_i).
\end{aligned}$$

The corollary follows from the principle of induction. □

The properties of homogenization sequences outlined in this section will be useful in the upcoming sections on lexicographic differentiation.

4.2 Lexicographic Differentiation in the Euclidean Space Setting

4.2.1 Notation

In this section, we restrict our discussion to finite-dimensional real normed vector spaces. That is, let E_1, E_2, and E_3 be finite-dimensional real normed vector spaces and let $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_3$ denote the norms of E_1, E_2, and E_3, respectively.

4.2.2 Preliminaries

Definition 4.2.1. *Let $\boldsymbol{f} \in \mathcal{L}(E_1, E_2)$, $\boldsymbol{x} \in E_1$, and $\boldsymbol{U} = \{\boldsymbol{u}_k\}_{k=1}^m$, an ordered set of directions in E_1. For each $k \in \{1, \ldots, m\}$, the function $\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}$ restricted to $L_k(\boldsymbol{U})$ is called the L_k derivative of $\boldsymbol{f}$ at $\boldsymbol{x}$ along $\boldsymbol{U}$.*

Remark 4.2.2. *By Corollary 4.1.19, the function $\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}$ restricted to $L_k(\boldsymbol{U})$ is linear. Hence, the L_k derivative of $\boldsymbol{f}$ at $\boldsymbol{x}$ along $\boldsymbol{U}$ has a matrix representation given a basis for E_1 and a basis for E_2.*

Theorem 4.2.3. *Let $\boldsymbol{f} \in \mathcal{L}(E_1, E_2)$, $\boldsymbol{x} \in E_1$, and $\boldsymbol{U} = \{\boldsymbol{u}_k\}_{k=1}^m$, an ordered set of directions in E_1 that span E_1. Then there exists $k_0 \in \{0, 1, \ldots, m\}$ such that the function $\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}$ is linear for any $k \geq k_0$. For any $k \geq k_0$,*

$$\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}\Big|_{L_k(\boldsymbol{U})} = \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k_0)}\Big|_{L_{k_0}(\boldsymbol{U})}.$$

Proof. Let $\boldsymbol{f} \in \mathcal{L}(E_1, E_2)$, $\boldsymbol{x} \in E_1$, and $\boldsymbol{U} = \{\boldsymbol{u}_k\}_{k=1}^m$, an ordered set of directions in E_1 that span E_1. Since $L_m(\boldsymbol{U}) = E_1$, there exists $k_0 \in \{0, 1, \ldots, m\}$ such that $L_{k_0}(\boldsymbol{U}) = E_1$. Since

$$\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k_0)} = \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k_0)}\Big|_{L_{k_0}(\boldsymbol{U})},$$

the function $\boldsymbol{f}^{(k_0)}_{\boldsymbol{x},\boldsymbol{U}} : E_1 \to E_2$ is linear by Corollary 4.1.19. Therefore, by Proposition 4.1.18,

$$\left.\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}\right|_{L_k(\boldsymbol{U})} = \left.\boldsymbol{f}^{(k_0)}_{\boldsymbol{x},\boldsymbol{U}}\right|_{L_{k_0}(\boldsymbol{U})} \quad \text{for any } k \geq k_0,$$

and the function $\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}$ is linear for any $k \geq k_0$. □

Remark 4.2.4. *The well-ordering of $\mathbb{N}$ guarantees that without loss of generality, we can set k_0 in Theorem 4.2.3 to be the least such number. Then, k_0 is called the degree of nondifferentiability of $\boldsymbol{f}$ at $\boldsymbol{x}$ along $\boldsymbol{U}$.*

4.2.3 The Lexicographic Derivative

Definition 4.2.5. *Let $\boldsymbol{f} \in \mathcal{L}(E_1, E_2)$, $\boldsymbol{x} \in E_1$, and $\boldsymbol{U} = \{\boldsymbol{u}_k\}_{k=1}^m$, an ordered set of directions in E_1 that span E_1. For any k greater than or equal to the degree of nondifferentiability of $\boldsymbol{f}$ at $\boldsymbol{x}$ along $\boldsymbol{U}$, the function $\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}$ restricted to $L_k(\boldsymbol{U})$ is called the lexicographic derivative (or L-derivative for short) of $\boldsymbol{f}$ at $\boldsymbol{x}$ along $\boldsymbol{U}$.*

Equivalently, we have the following, more straightforward, definition of the L-derivative.

Definition 4.2.6. *Let $\boldsymbol{f} \in \mathcal{L}(E_1, E_2)$, $\boldsymbol{x} \in E_1$, and $\boldsymbol{U} = \{\boldsymbol{u}_k\}_{k=1}^m$, an ordered set of directions in E_1 that span E_1. The function $\boldsymbol{f}^{(m)}_{\boldsymbol{x},\boldsymbol{U}} : E_1 \to E_2$ is called the lexicographic derivative (or L-derivative for short) of $\boldsymbol{f}$ at $\boldsymbol{x}$ along $\boldsymbol{U}$ and is denoted by $\boldsymbol{J}_L\boldsymbol{f}(\boldsymbol{x};\boldsymbol{U})$.*

Definition 4.2.7. *The lexicographic subdifferential (or L subdifferential for short) of $\boldsymbol{f} \in \mathcal{L}(E_1, E_2)$ at $\boldsymbol{x} \in E_1$ is defined to be*

$$\partial_L \boldsymbol{f}(\boldsymbol{x}) := \{\boldsymbol{J}_L\boldsymbol{f}(\boldsymbol{x};\boldsymbol{U}) : \boldsymbol{U} \text{ spans } E_1\}.$$

Example 4.2.8. *Let $g : \mathbb{R}^2 \to \mathbb{R}$ be a function defined by $g(x_1\ x_2)^T = \max(0, \min(x_1, x_2))$ and $\boldsymbol{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ a direction matrix. That is, the sequence of directions in $\mathbb{R}^2$ is*

$$\boldsymbol{U} = \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\}.$$

We have $g \in \mathcal{L}(\mathbb{R}^2, \mathbb{R})$ *and*

$$\begin{aligned}
g^{(0)}_{(0\ 0)^T,\boldsymbol{I}}(d_1\ d_2)^T &= g'((0\ 0)^T; (d_1\ d_2)^T) = \max\left(0, \min\left(d_1, d_2\right)\right) \\
g^{(1)}_{(0\ 0)^T,\boldsymbol{I}}(d_1\ d_2)^T &= \left[g^{(0)}_{(0\ 0)^T,\boldsymbol{I}}\right]'\left(\begin{pmatrix}1\\0\end{pmatrix}; (d_1\ d_2)^T\right) = \max\left(0, d_2\right) \\
g^{(2)}_{(0\ 0)^T,\boldsymbol{I}}(d_1\ d_2)^T &= \left[g^{(1)}_{(0\ 0)^T,\boldsymbol{I}}\right]'\left(\begin{pmatrix}0\\1\end{pmatrix}; (d_1\ d_2)^T\right) = d_2.
\end{aligned}$$

As expected, $g^{(2)}_{(0\ 0)^T,\boldsymbol{I}}$ *is linear. Moreover, the L derivative of* g *at* $(0\ 0)^T$ *along* $\boldsymbol{I}$ *is*

$$\boldsymbol{J}_L g\left((0\ 0)^T; \boldsymbol{I}\right) = \boldsymbol{J} g^{(2)}_{(0\ 0)^T,\boldsymbol{I}} = (0\ 1).$$

It is worth noting that

$$\boldsymbol{J}_L g\left((0\ 0)^T; \boldsymbol{I}\right) \in \partial_B g(0\ 0)^T = \left\{(1\ 0), (0\ 1), (0\ 0)\right\}.$$

4.2.4 The Theoretical Toolkit of L-Derivatives

L-derivatives possess an equality-based chain rule, as stated below.

Theorem 4.2.9. *Let* $\boldsymbol{f} \in \mathcal{L}(E_1, E_2)$ *and* $\boldsymbol{F} \in \mathcal{L}(E_2, E_3)$*. Then* $\boldsymbol{F} \circ \boldsymbol{f} \in \mathcal{L}(E_1, E_3)$*. Moreover, for any* $\boldsymbol{x} \in E_1$ *and any basis* $\boldsymbol{U} = \{\boldsymbol{u}_k\}_{k=1}^m$ *for* E_1*,*

$$(\boldsymbol{F} \circ \boldsymbol{f})^{(m)}_{\boldsymbol{x},\boldsymbol{U}} = \boldsymbol{F}^{(m)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}} \circ \boldsymbol{f}^{(m)}_{\boldsymbol{x},\boldsymbol{U}},$$

where the sequence $\boldsymbol{V} = \{\boldsymbol{v}_k\}_{k=1}^m$ *of directions in* $\boldsymbol{E}_2$ *is given by*

$$\boldsymbol{v}_k = \boldsymbol{f}^{(m)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_k), \quad k \in \{1, \ldots, m\}.$$

Proof. The theorem follows immediately from Theorem 4.1.16 and Proposition 4.1.18. □

In Example 4.2.8, we observed that the L-derivative of g at $(0\ 0)^T$ along $\boldsymbol{I}$ is an element of the B-subdifferential of g at $(0\ 0)^T$. In the following proposition, we present a generalized derivatives landscape showing how different generalized derivatives are related to one another, which combines results from [5, 6, 9].

Proposition 4.2.10. *Let X be an open subset of $\mathbb{R}^n$.*

(i) If $f : X \to \mathbb{R}$ is an L-smooth function, then $\partial_L \boldsymbol{f}(\boldsymbol{x}) \subset \partial \boldsymbol{f}(\boldsymbol{x})$ for any $\boldsymbol{x} \in X$.

(ii) If $\boldsymbol{f} : X \to \mathbb{R}^m$ is a $\mathcal{PC}^1$ function, then $\partial_L \boldsymbol{f}(\boldsymbol{x}) \subset \partial_B \boldsymbol{f}(\boldsymbol{x}) \subset \partial \boldsymbol{f}(\boldsymbol{x})$ for any $\boldsymbol{x} \in X$.

(iii) If $\boldsymbol{f} : X \to \mathbb{R}^m$ is a $\mathcal{C}^1$ function, then $\partial_L \boldsymbol{f}(\boldsymbol{x}) = \partial_B \boldsymbol{f}(\boldsymbol{x}) = \partial \boldsymbol{f}(\boldsymbol{x}) = \{\boldsymbol{J}\boldsymbol{f}(\boldsymbol{x})\}$ for any $\boldsymbol{x} \in X$.

(iv) If $\boldsymbol{f} : X \to \mathbb{R}^m$ is an L-smooth function, then for any $\boldsymbol{x} \in X$ and any $\boldsymbol{d} \in \mathbb{R}^n$,

$$\{\boldsymbol{A}\boldsymbol{d} : \boldsymbol{A} \in \partial_L \boldsymbol{f}(\boldsymbol{x})\} \subset \{\boldsymbol{A}\boldsymbol{d} : \boldsymbol{A} \in \partial \boldsymbol{f}(\boldsymbol{x})\}.$$

In the above proposition, the last item states that for an L-smooth function the plenary hull of the L-subdifferential is contained in the plenary hull of the Clarke Jacobian. One implication of this statement is that L-derivatives can be used in place of Clarke Jacobian elements in implementing a numerical tool associated with Clarke Jacobians, such as a Newton's method, since it is concerned with matrix-vector products.

4.2.5 The Lexicographic Directional Derivative

Lexicographic directional differentiation by Khan and Barton [6] provides a systematic way to find L-derivatives, allowing for implementation of numerical tools requiring Clarke Jacobians.

Definition 4.2.11. *Let $\boldsymbol{f} : E_1 \to E_2$ be an L-smooth function. The lexicographic directional derivative (or LD-derivative for short) of $\boldsymbol{f}$ at $\boldsymbol{x} \in E_1$ for a matrix $\boldsymbol{M} = \left[\boldsymbol{m}_{(1)} \cdots \boldsymbol{m}_{(k)}\right]$ of directions in E_1 is defined as*

$$\boldsymbol{f}'(\boldsymbol{x}; \boldsymbol{M}) := \left[\boldsymbol{f}^{(0)}_{\boldsymbol{x},\boldsymbol{M}}(\boldsymbol{m}_{(1)}) \;\; \cdots \;\; \boldsymbol{f}^{(k-1)}_{\boldsymbol{x},\boldsymbol{M}}(\boldsymbol{m}_{(k)})\right].$$

The LD-derivative is a generalization of the directional derivative for the class of L-smooth functions.

Proposition 4.2.12. *Let X be an open subset of $\mathbb{R}^n$.*

(i) If $\boldsymbol{f} : X \to \mathbb{R}^m$ is an L-smooth function and $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ is nonsingular, then

$$\boldsymbol{f}'(\boldsymbol{x}; \boldsymbol{M}) = \boldsymbol{J}_L\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{M})\boldsymbol{M}.$$

(ii) If $\boldsymbol{f} : X \to \mathbb{R}^m$ is an L-smooth function that is differentiable at $\boldsymbol{x} \in X$, then

$$\boldsymbol{f}'(\boldsymbol{x}; \boldsymbol{M}) = \boldsymbol{J}\boldsymbol{f}(\boldsymbol{x})\boldsymbol{M}.$$

The name "LD-derivative" comes from the fact that LD-derivatives for elementary functions, such as the max and min functions, the absolute value function, and the 2-norm function, are naturally expressed using lexicographic ordering on $\mathbb{R}^n$.

Definition 4.2.13. *The binary relation $\prec$ on $\mathbb{R}^n$ is defined such that if $\boldsymbol{x} = (x_1 \dots x_n)^T, \boldsymbol{y} = (y_1 \dots y_n)^T \in \mathbb{R}^n$, then*

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \prec \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

if there exists $k \in \{1, \dots, n\}$ such that $x_k < y_k$ and $x_i = y_i$ for all $i \in \{1, \dots, k-1\}$. The relations $\preceq$, $\succ$, and $\succeq$ are defined similarly.

Remark 4.2.14. *The binary relations in Definition 4.2.13 are total orders on $\mathbb{R}^n$.*

Example 4.2.15. *In $\mathbb{R}^3$, we have*

$$\begin{pmatrix} 2 \\ 6 \\ 9 \end{pmatrix} \prec \begin{pmatrix} 2 \\ 7 \\ 1 \end{pmatrix}.$$

Example 4.2.16. *LD-derivatives of elementary functions are naturally expressed using lexicographic ordering. For* $\min : \mathbb{R}^2 \to \mathbb{R}$, $\min(x_1\ x_2)^T = \min(x_1, x_2)$, *and* $\max : \mathbb{R}^2 \to \mathbb{R}$, $\max(x_1\ x_2)^T = \max(x_1, x_2)$,

$$\min{}'\left(\begin{pmatrix} x \\ y \end{pmatrix}; \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}\right) = \begin{cases} [m_{11}\ m_{12}] & \text{if } \begin{pmatrix} x \\ m_{11} \\ m_{12} \end{pmatrix} \prec \begin{pmatrix} y \\ m_{21} \\ m_{22} \end{pmatrix} \\ [m_{21}\ m_{22}] & \text{if } \begin{pmatrix} x \\ m_{11} \\ m_{12} \end{pmatrix} \succeq \begin{pmatrix} y \\ m_{21} \\ m_{22} \end{pmatrix} \end{cases}$$

$$\max{}'\left(\begin{pmatrix} x \\ y \end{pmatrix}; \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}\right) = \begin{cases} [m_{21}\ m_{22}] & \text{if } \begin{pmatrix} x \\ m_{11} \\ m_{12} \end{pmatrix} \prec \begin{pmatrix} y \\ m_{21} \\ m_{22} \end{pmatrix} \\ [m_{11}\ m_{12}] & \text{if } \begin{pmatrix} x \\ m_{11} \\ m_{12} \end{pmatrix} \succeq \begin{pmatrix} y \\ m_{21} \\ m_{22} \end{pmatrix} \end{cases}.$$

4.2.6 The Theoretical Toolkit of LD-Derivatives

Recall that although Clarke Jacobian elements are critical in implementing numerical tools for the class of locally Lipschitz continuous functions, the fact that calculus rules for Clarke Jacobians are inclusion-based makes it difficult to obtain Clarke Jacobian elements even for a function constructed from functions whose Clarke Jacobians are known. Unlike Clarke Jacobians, however, LD-derivatives possess equality-based calculus rules [6].

Proposition 4.2.17 (The chain rule for the LD-derivative)**.** *Let X and Y be open subsets of $\mathbb{R}^n$. Let $\boldsymbol{f} : X \to \mathbb{R}^m$ and $\boldsymbol{g} : Y \to \mathbb{R}^m$ be L-smooth functions. Then the composition*

$\boldsymbol{f} \circ \boldsymbol{g}$ *is L-smooth and*

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

Proposition 4.2.18. *Let X be an open subset of $\mathbb{R}^n$. Let $\boldsymbol{u} = (u_1 \dots u_n)^T : X \to \mathbb{R}^m$ and $\boldsymbol{v} = (v_1 \dots v_n)^T : X \to \mathbb{R}^m$ be L-smooth functions and $\boldsymbol{M} = \left[\boldsymbol{m}_{(1)} \cdots \boldsymbol{m}_{(k)}\right]$ a row of directions in X. For any $\boldsymbol{x} \in X$,*

(i) $\boldsymbol{u}'(\boldsymbol{x}; \boldsymbol{M}) = (u_1'(\boldsymbol{x}; \boldsymbol{M}) \ u_2'(\boldsymbol{x}; \boldsymbol{M}) \dots u_m'(\boldsymbol{x}; \boldsymbol{M}))^T$

(ii) $[\boldsymbol{u} + \boldsymbol{v}]' (\boldsymbol{x}; \boldsymbol{M}) = \boldsymbol{u}'(\boldsymbol{x}; \boldsymbol{M}) + \boldsymbol{v}'(\boldsymbol{x}; \boldsymbol{M})$

(iii) If $n = 1$, then

$$[uv]' (\boldsymbol{x}; \boldsymbol{M}) = u'(\boldsymbol{x}; \boldsymbol{M})v(\boldsymbol{x}) + u(\boldsymbol{x})v'(\boldsymbol{x}; \boldsymbol{M}).$$

These equality-based calculus rules allow us to find an L-derivative, i.e., a Clarke Jacobian element, in a systematic way. The process is outlined below for an L-smooth function $\boldsymbol{f} : X \subset \mathbb{R}^n \to \mathbb{R}^m$. Let $\boldsymbol{x} \in X$.

(i) Choose a nonsingular directions matrix $\boldsymbol{M} \in \mathbb{R}^{n \times n}$.

(ii) Find the LD-derivative $\boldsymbol{f}'(\boldsymbol{x}; \boldsymbol{M})$ using either the definition or the equality-based calculus rules.

(iii) Solve the linear equation system

$$\boldsymbol{f}'(\boldsymbol{x}; \boldsymbol{M}) = \boldsymbol{J}_L \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{M}) \boldsymbol{M}$$

for the L-derivative $\boldsymbol{J}_L \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{M})$, which is unique since $\boldsymbol{M}$ is nonsingular.

Example 4.2.19. *Consider the function $g : \mathbb{R}^2 \to \mathbb{R}$ defined by $g(x_1 \ x_2)^T = \max(0, \min(x_1, x_2))$. Then*

$$g^{(0)}_{(0\ 0)^T, \boldsymbol{I}}(d_1 \ d_2)^T = \max(0, \min(d_1, d_2))$$
$$g^{(1)}_{(0\ 0)^T, \boldsymbol{I}}(d_1 \ d_2)^T = \max(0, d_2).$$

Hence,

$$g'\left((0\ 0)^T; \boldsymbol{I}\right) = \left[g^{(0)}_{(0\ 0)^T,\boldsymbol{I}}(1\ 0)^T\ g^{(1)}_{(0\ 0)^T,\boldsymbol{I}}(0\ 1)^T\right] = [0\ 1].$$

Therefore,

$$\boldsymbol{J}_L g\left((0\ 0)^T; \boldsymbol{I}\right) = g'\left((0\ 0)^T; \boldsymbol{I}\right)\boldsymbol{I}^{-1} = [0\ 1].$$

4.3 Lexicographic Differentiation in the Banach Space Setting

Motivated by the utility of lexicographic differentiation in obtaining Clarke Jacobian elements via the LD-derivative, we now focus our attention on the Banach space setting. In this section, we restrict our discussion to Banach spaces that have Schauder bases. The definitions and results here are from [9]. The proofs here are due to Nesterov except for the one for Theorem 4.3.5. His proofs have been expanded for ease of comprehension.

4.3.1 Notation

Let E_1, E_2, and E_3 be Banach spaces and let $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_3$ denote the norms of E_1, E_2, and E_3, respectively.

4.3.2 Preliminaries

Definition 4.3.1. *Let $\boldsymbol{f} \in \mathcal{L}(E_1, E_2)$, $\boldsymbol{x} \in E_1$, and $\boldsymbol{U} = \{\boldsymbol{u}_k\}_{k=1}^{\infty}$, a sequence of directions in E_1. For each $k \in \mathbb{N}$, the function $\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}$ restricted to $L_k(\boldsymbol{U})$ is called the L_k derivative of $\boldsymbol{f}$ at $\boldsymbol{x}$ along $\boldsymbol{U}$.*

Proposition 4.3.2. *Let $\boldsymbol{f} \in \mathcal{L}(E_1, E_2)$, $\boldsymbol{x} \in E_1$, and $\boldsymbol{U} = \{\boldsymbol{u}_k\}_{k=1}^{\infty}$, a sequence of directions in E_1. For each $k \in \mathbb{N}$, the function $\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}$ restricted to $L_k(\boldsymbol{U})$ is a bounded linear operator.*

Proof. Let $\boldsymbol{f} \in \mathcal{L}(E_1, E_2)$, $\boldsymbol{x} \in E_1$, and $\boldsymbol{U} = \{\boldsymbol{u}_k\}_{k=1}^{\infty}$, a sequence of directions in E_1. By Corollary 4.1.19, the function $\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}$ restricted to $L_k(\boldsymbol{U})$ is linear for each $k \in \mathbb{N}$. We show that the function $\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}$ restricted to $L_k(\boldsymbol{U})$ is bounded for each $k \in \mathbb{N}$. Let $\boldsymbol{h} \in L_1(\boldsymbol{U})$.

Since

$$
\begin{aligned}
\left\|\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(1)}(\boldsymbol{h})\right\|_2 &= \left\|\lim_{t\downarrow\infty} \frac{\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(0)}(\boldsymbol{x}+t\boldsymbol{h}) - \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(0)}(\boldsymbol{x})}{t}\right\|_2 \\
&= \lim_{t\downarrow\infty} \left\|\frac{\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(0)}(\boldsymbol{x}+t\boldsymbol{h}) - \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(0)}(\boldsymbol{x})}{t}\right\|_2 \\
&= \lim_{t\downarrow\infty} \frac{\left\|\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(0)}(\boldsymbol{x}+t\boldsymbol{h}) - \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(0)}(\boldsymbol{x})\right\|_2}{t} \\
&\leq \lim_{t\downarrow\infty} \frac{K\left\|(\boldsymbol{x}+t\boldsymbol{h}) - \boldsymbol{x}\right\|_1}{t} \\
&= K\left\|\boldsymbol{h}\right\|_1
\end{aligned}
$$

for some $K \geq 0$, the function $\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(1)}$ is bounded. Suppose that the function $\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}$ restricted to $L_k(\boldsymbol{U})$ is bounded. Let $\boldsymbol{h} \in L_{k+1}(\boldsymbol{U})$. Then

$$
\begin{aligned}
\left\|\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k+1)}(\boldsymbol{h})\right\|_2 &= \left\|\lim_{t\downarrow\infty} \frac{\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{x}+t\boldsymbol{h}) - \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{x})}{t}\right\|_2 \\
&= \lim_{t\downarrow\infty} \left\|\frac{\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{x}+t\boldsymbol{h}) - \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{x})}{t}\right\|_2 \\
&= \lim_{t\downarrow\infty} \frac{\left\|\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{x}+t\boldsymbol{h}) - \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{x})\right\|_2}{t} \\
&\leq \lim_{t\downarrow\infty} \frac{K\left\|(\boldsymbol{x}+t\boldsymbol{h}) - \boldsymbol{x}\right\|_1}{t} \\
&= K\left\|\boldsymbol{h}\right\|_1
\end{aligned}
$$

for some $K \geq 0$. Hence, the function $\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k+1)}$ restricted to $L_{k+1}(\boldsymbol{U})$ is bounded. The theorem follows from the principle of induction. □

4.3.3 The Lexicographic Derivative

Theorem 4.3.3. *Let* $\boldsymbol{f} \in \mathcal{L}(E_1, E_2)$, $\boldsymbol{x} \in E_1$, *and* $\boldsymbol{U} = \{\boldsymbol{u}_k\}_{k=1}^{\infty}$, *a basis for* E_1. *Then there exists a unique bounded linear operator* $\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(\infty)} : E_1 \to E_2$ *such that*

$$\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(\infty)}(\boldsymbol{h}) = \lim_{k\to\infty} \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k)}(\boldsymbol{h}), \quad \boldsymbol{h} \in E_1.$$

Proof. Let $\boldsymbol{f} \in \mathcal{L}(E_1, E_2)$, $\boldsymbol{x} \in E_1$, and $\boldsymbol{U} = \{\boldsymbol{u}_k\}_{k=1}^{\infty}$, a basis for E_1. If

$$\boldsymbol{h} = \sum_{k=1}^{\infty} h_k \boldsymbol{u}_k \in E_1,$$

then let $\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(\infty)} : E_1 \to E_2$ be a function defined by

$$\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(\infty)}(\boldsymbol{h}) = \sum_{k=1}^{\infty} h_k \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k-1)}(\boldsymbol{u}_k).$$

We first show that $\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(\infty)}$ exists. For each $m \in \mathbb{N}$, let

$$\begin{aligned}
\boldsymbol{h}_m &= \sum_{k=1}^{m} h_k \boldsymbol{u}_k, \\
\boldsymbol{y}_m &= \sum_{k=1}^{m} h_k \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k-1)}(\boldsymbol{u}_k).
\end{aligned}$$

By Proposition 4.1.14, each member of the homogenization sequence of $\boldsymbol{f}$ generated by $\boldsymbol{x}$ and $\boldsymbol{U}$ is Lipschitz continuous with a Lipschitz constant $K > 0$. Let $\epsilon > 0$. Since $\{\boldsymbol{h}_m\}_{m=1}^{\infty}$ is a Cauchy sequence, there exists $N \in \mathbb{N}$ such that for any $m_1, m_2 \geq N$,

$$\|h_{m_1} - h_{m_2}\|_1 < \frac{\epsilon}{K}.$$

Then, for $m_1 > m_2 \geq N$,

$$\begin{aligned}
\left\|\boldsymbol{y}_{m_1} - \boldsymbol{y}_{m_2}\right\|_2 &= \left\| \sum_{k=m_2+1}^{m_1} h_k \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k-1)}(\boldsymbol{u}_k) \right\|_2 \\
&= \left\| \sum_{k=m_2+1}^{m_1} h_k \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(m_1)}(\boldsymbol{u}_k) \right\|_2 \\
&= \left\| \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(m_1)} \left(\sum_{k=m_2+1}^{m_1} h_k \boldsymbol{u}_k \right) \right\|_2 \\
&= \left\| \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(m_1)} \left(\boldsymbol{h}_{m_1} - \boldsymbol{h}_{m_2} \right) \right\|_2 \\
&\leq K \left\| \boldsymbol{h}_{m_1} - \boldsymbol{h}_{m_2} \right\|_1 \\
&< \epsilon.
\end{aligned}$$

Hence, $\{\boldsymbol{y}_m\}_{m=1}^{\infty}$ is Cauchy and therefore convergent, as needed. The function $\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(\infty)} : E_1 \to E_2$ is linear since for

$$\boldsymbol{a} = \sum_{k=1}^{\infty} a_k \boldsymbol{u}_k \in E_1,$$
$$\boldsymbol{b} = \sum_{k=1}^{\infty} b_k \boldsymbol{u}_k \in E_1,$$

and $\alpha, \beta \in \mathbb{R}$,

$$\begin{aligned}
\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(\infty)}(\alpha \boldsymbol{a} + \beta \boldsymbol{b}) &= \sum_{k=1}^{\infty} (\alpha a_k + \beta b_k) \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k-1)}(\boldsymbol{u}_k) \\
&= \lim_{m\to\infty} \sum_{k=1}^{m} (\alpha a_k + \beta b_k) \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k-1)}(\boldsymbol{u}_k) \\
&= \lim_{m\to\infty} \left(\alpha \sum_{k=1}^{m} a_k \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k-1)}(\boldsymbol{u}_k) + \beta \sum_{k=1}^{m} b_k \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k-1)}(\boldsymbol{u}_k) \right) \\
&= \alpha \sum_{k=1}^{\infty} a_k \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k-1)}(\boldsymbol{u}_k) + \beta \sum_{k=1}^{\infty} b_k \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k-1)}(\boldsymbol{u}_k) \\
&= \alpha \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(\infty)}(\boldsymbol{a}) + \beta \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(\infty)}(\boldsymbol{b}).
\end{aligned}$$

It is bounded since if

$$\boldsymbol{h} = \sum_{k=1}^{\infty} h_k \boldsymbol{u}_k \in E_1,$$

then

$$\begin{aligned}
\left\| \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(\infty)}(\boldsymbol{h}) \right\|_2 &= \left\| \lim_{m\to\infty} \sum_{k=1}^{m} h_k \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k-1)}(\boldsymbol{u}_k) \right\|_2 \\
&= \lim_{m\to\infty} \left\| \sum_{k=1}^{m} h_k \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k-1)}(\boldsymbol{u}_k) \right\|_2 \\
&= \lim_{m\to\infty} \left\| \sum_{k=1}^{m} h_k \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(m)}(\boldsymbol{u}_k) \right\|_2 \\
&= \lim_{m\to\infty} \left\| \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(m)} \left(\sum_{k=1}^{m} h_k \boldsymbol{u}_k \right) \right\|_2 \\
&= \lim_{m\to\infty} \left\| \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(m)} (\boldsymbol{h}_m) \right\|_2 \\
&\leq \lim_{m\to\infty} M \left\| \boldsymbol{h}_m \right\|_1 \\
&= M \left\| \boldsymbol{h} \right\|_1 .
\end{aligned}$$

Note that

$$\begin{aligned}
\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(\infty)}(\boldsymbol{h}) &= \lim_{m\to\infty} \sum_{k=1}^{m} h_k \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(k-1)}(\boldsymbol{u}_k) \\
&= \lim_{m\to\infty} \sum_{k=1}^{m} h_k \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(m)}(\boldsymbol{u}_k) \\
&= \lim_{m\to\infty} \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(m)}\left(\sum_{k=1}^{m} h_k \boldsymbol{u}_k\right) \\
&= \lim_{m\to\infty} \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(m)}(\boldsymbol{h}_m).
\end{aligned}$$

Hence,

$$\begin{aligned}
\left\| \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(\infty)}(\boldsymbol{h}) - \lim_{m\to\infty} \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(m)}(\boldsymbol{h}) \right\|_2 &= \left\| \lim_{m\to\infty} \left(\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(m)}(\boldsymbol{h}_m) - \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(m)}(\boldsymbol{h}) \right) \right\|_2 \\
&= \lim_{m\to\infty} \left\| \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(m)}(\boldsymbol{h}_m) - \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(m)}(\boldsymbol{h}) \right\|_2 \\
&\leq K \lim_{m\to\infty} \|\boldsymbol{h}_m - \boldsymbol{h}\|_2 \\
&= 0.
\end{aligned}$$

Therefore,

$$\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(\infty)}(\boldsymbol{h}) = \lim_{m\to\infty} \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(m)}(\boldsymbol{h}).$$

Note that E_2 is Hausdorff since it is a metric space with respect to the norm $\|\cdot\|_2$. This limit is unique since every convergent sequence has a unique limit in a Hausdorff space.

□

Remark 4.3.4. *The function* $\boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(\infty)} : E_1 \to E_2$ *is called the lexicographic derivative (or L derivative for short) of* $\boldsymbol{f}$ *at* $\boldsymbol{x}$ *for* $\boldsymbol{U}$*.*

4.3.4 The Theoretical Toolkit of L-Derivatives

Theorem 4.3.5. *Let* $\boldsymbol{f} \in \mathcal{L}(E_1, E_2)$ *and* $\boldsymbol{F} \in \mathcal{L}(E_2, E_3)$*. Then* $\boldsymbol{F} \circ \boldsymbol{f} \in \mathcal{L}(E_1, E_3)$*. Moreover, for any* $\boldsymbol{x} \in E_1$ *and* $\boldsymbol{U} = \{\boldsymbol{u}_k\}_{k=1}^{\infty}$*, a basis for* E_1*,*

$$(\boldsymbol{F} \circ \boldsymbol{f})_{\boldsymbol{x},\boldsymbol{U}}^{(\infty)} = \boldsymbol{F}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}}^{(\infty)} \circ \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(\infty)},$$

where the sequence $\boldsymbol{V} = \{\boldsymbol{v}_k\}_{k=1}^{\infty}$ *of directions in* E_2 *is given by*

$$\boldsymbol{v}_k = \boldsymbol{f}_{\boldsymbol{x},\boldsymbol{U}}^{(\infty)}(\boldsymbol{u}_k), \quad k \in \mathbb{N}.$$

Proof. Let $\boldsymbol{f} \in \mathcal{L}(E_1, E_2)$ and $\boldsymbol{F} \in \mathcal{L}(E_2, E_3)$. Then $\boldsymbol{f} \in \mathcal{D}(E_1, E_2)$ and $\boldsymbol{F} \in \mathcal{D}(E_2, E_3)$. Thus, $\boldsymbol{F} \circ \boldsymbol{f} \in \mathcal{D}(E_1, E_3)$ by Proposition 4.1.6. Let $\boldsymbol{x} \in E_1$ and $\boldsymbol{U} = \{\boldsymbol{u}_k\}_{k=1}^{\infty}$, a basis for E_1. By Theorem 4.1.16,

$$(\boldsymbol{F} \circ \boldsymbol{f})^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\cdot) = \boldsymbol{F}^{(k)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}}\left(\boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\cdot)\right), \quad k \in \{0, 1, 2, \dots\},$$

where the sequence $\boldsymbol{V} = \{\boldsymbol{v}_k\}_{k=1}^{m}$ of directions in E_2 is given by

$$\boldsymbol{v}_k = \boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_k), \quad k \in \mathbb{N}.$$

For each $k \in \mathbb{N}$,

$$\boldsymbol{v}_k = \boldsymbol{f}^{(k)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_k) = \lim_{m\to\infty} \boldsymbol{f}^{(m)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_k) = \boldsymbol{f}^{(\infty)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_k).$$

Moreover, if $\boldsymbol{h} \in E_1$,

$$\begin{aligned}
&\left\| (\boldsymbol{F} \circ \boldsymbol{f})^{(\infty)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h}) - \left(\boldsymbol{F}^{(\infty)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}} \circ \boldsymbol{f}^{(\infty)}_{\boldsymbol{x},\boldsymbol{U}}\right)(\boldsymbol{h}) \right\|_3 \\
&= \left\| \lim_{m\to\infty} (\boldsymbol{F} \circ \boldsymbol{f})^{(m)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h}) - \lim_{m\to\infty} \boldsymbol{F}^{(m)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}}\left(\boldsymbol{f}^{(\infty)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h})\right) \right\|_3 \\
&= \lim_{m\to\infty} \left\| \boldsymbol{F}^{(m)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}}\left(\boldsymbol{f}^{(m)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h})\right) - \boldsymbol{F}^{(m)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}}\left(\boldsymbol{f}^{(\infty)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h})\right) \right\|_3 \\
&\leq K \lim_{m\to\infty} \left\| \boldsymbol{f}^{(m)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h}) - \boldsymbol{f}^{(\infty)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h}) \right\|_2 \\
&= 0.
\end{aligned}$$

Hence,

$$(\boldsymbol{F} \circ \boldsymbol{f})^{(\infty)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{h}) = \left(\boldsymbol{F}^{(\infty)}_{\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{V}} \circ \boldsymbol{f}^{(\infty)}_{\boldsymbol{x},\boldsymbol{U}}\right)(\boldsymbol{h}), \quad \boldsymbol{h} \in E_1,$$

where the sequence $\boldsymbol{V} = \{\boldsymbol{v}_k\}_{k=1}^{\infty}$ of directions in E_2 is given by

$$\boldsymbol{v}_k = \boldsymbol{f}^{(\infty)}_{\boldsymbol{x},\boldsymbol{U}}(\boldsymbol{u}_k), \quad k \in \mathbb{N},$$

as needed. □

4.3.5 Examples

In this section, we provide examples to illustrate lexicographic differentiation in the Banach space setting. Consider the collection $\mathcal{C}$ of all real sequences. Let

$\{x_k\}_{k=1}^\infty, \{y_k\}_{k=1}^\infty \in \mathcal{C}$ and $\alpha \in \mathbb{R}$. Define vector addition by

$$\{x_k\}_{k=1}^\infty + \{y_k\}_{k=1}^\infty = \{x_n + y_n\}_{k=1}^\infty$$

and scalar multiplication by

$$\alpha\{x_k\}_{k=1}^\infty = \{\alpha x_k\}_{k=1}^\infty.$$

Then $\mathcal{C}$ equipped with the vector addition and the scalar multiplication is a vector space. Consider the collection $\ell^2(\mathbb{R})$ of real sequences $\{x_k\}_{k=1}^\infty$ satisfying

$$\sum_{k=1}^\infty |x_k|^2 < \infty.$$

Then $\ell^2(\mathbb{R})$ is a subspace of $\mathcal{C}$. Moreover,

$$\|\{x_k\}_{k=1}^\infty\|_2 = \left(\sum_{k=1}^\infty |x_k|^2\right)^{1/2}, \quad \{x_k\}_{k=1}^\infty \in \ell^2(\mathbb{R})$$

is a norm for $\ell^2(\mathbb{R})$. In fact, $\ell^2(\mathbb{R})$ is a complete metric space with respect to this norm and therefore is a complete normed vector space.

Example 4.3.6. *Let $\boldsymbol{f} : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ be a function given by*

$$\boldsymbol{f}(\{x_k\}_{k=1}^\infty) = \{\max(0, x_k)\}_{k=1}^\infty.$$

The function $\boldsymbol{f}$ exists since

$$\sum_{k=1}^\infty |\max(0, x_k)|^2 \le \sum_{k=1}^\infty |x_k|^2 < \infty.$$

It is Lipschitz continuous since for any $\{x_k\}_{k=1}^\infty$ and $\{y_k\}_{k=1}^\infty$ in $\ell^2(\mathbb{R})$,

$$\begin{aligned}\|\boldsymbol{f}(\{x_k\}_{k=1}^\infty) - \boldsymbol{f}(\{y_k\}_{k=1}^\infty)\|_2 &= \left(\sum_{k=1}^\infty |\max(0, x_k) - \max(0, y_k)|^2\right)^{1/2} \\ &= \lim_{m\to\infty} \left(\sum_{k=1}^m |\max(0, x_k) - \max(0, y_k)|^2\right)^{1/2} \\ &\le \lim_{m\to\infty} \left(\sum_{k=1}^m |x_k - y_k|^2\right)^{1/2} \\ &= \|\{x_k\}_{k=1}^\infty - \{y_k\}_{k=1}^\infty\|_2,\end{aligned}$$

and is directionally differentiable on $\ell^2(\mathbb{R})$ *since for any* $\{x_k\}_{k=1}^{\infty}$ *and* $\{h_k\}_{k=1}^{\infty}$ *in* $\ell^2(\mathbb{R})$,

$$\boldsymbol{f}'(\{x_k\}_{k=1}^{\infty};\{h_k\}_{k=1}^{\infty}) = \{y_k\}_{k=1}^{\infty},$$

where for each $k \in \mathbb{N}$,

$$y_k = \begin{cases} h_i & \text{if } x_k > 0, \\ 0 & \text{if } x_k < 0, \\ h_k & \text{if } x_k = 0 \text{ and } h_i > 0, \\ 0 & \text{if } x_i = 0 \text{ and } h_i \leq 0. \end{cases}$$

In fact, $\boldsymbol{f} \in \mathcal{L}(\ell^2(\mathbb{R}), \ell^2(\mathbb{R}))$ *since for any* $\{x_k\}_{k=1}^{\infty}, \{h_k\}_{k=1}^{\infty} \in \ell^2(\mathbb{R})$ *and any sequence* $\boldsymbol{U} = \{\{u_{kj}\}_{j=1}^{\infty}\}_{k=1}^{\infty}$ *of directions in* $\ell^2(\mathbb{R})$,

$$\boldsymbol{f}^{(n)}_{\{x_k\}_{k=1}^{\infty},\boldsymbol{U}}(\{h_k\}_{k=1}^{\infty}) = \{y_{nk}\}_{k=1}^{\infty}, \quad n \in \mathbb{N},$$

where for each $k \in \mathbb{N}$,

$$y_{nk} = \begin{cases} h_k & \text{if } \begin{pmatrix} x_k \\ u_{1k} \\ u_{2k} \\ \vdots \\ u_{nk} \\ d_k \end{pmatrix} \succ \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \\ 0 & \text{if } \begin{pmatrix} x_k \\ u_{1k} \\ u_{2k} \\ \vdots \\ u_{nk} \\ d_k \end{pmatrix} \preceq \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}. \end{cases}$$

Example 4.3.7. *Let $\boldsymbol{g} : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ be a function given by*

$$\boldsymbol{g}(\{x_k\}_{k=1}^{\infty}) = \{|x_k|\}_{k=1}^{\infty}.$$

The function $\boldsymbol{g}$ exists since

$$\sum_{k=1}^{\infty} |(|x_k|)|^2 = \sum_{k=1}^{\infty} |x_k|^2 < \infty.$$

It is Lipschitz continuous since for any $\{x_k\}_{k=1}^{\infty}$ and $\{y_k\}_{k=1}^{\infty}$ in $\ell^2(\mathbb{R})$,

$$\begin{aligned}
\|\boldsymbol{g}(\{x_k\}_{k=1}^{\infty}) - \boldsymbol{g}(\{y_k\}_{k=1}^{\infty})\|_2 &= \left(\sum_{k=1}^{\infty} ||x_k| - |y_k||^2\right)^{1/2} \\
&= \lim_{m\to\infty} \left(\sum_{k=1}^{m} ||x_k| - |y_k||^2\right)^{1/2} \\
&\leq \lim_{m\to\infty} \left(\sum_{k=1}^{m} |x_k - y_k|^2\right)^{1/2} \\
&= \|\{x_k\}_{k=1}^{\infty} - \{y_k\}_{k=1}^{\infty}\|_2
\end{aligned}$$

by the reverse triangle inequality, and is directionally differentiable on $\ell^2(\mathbb{R})$ since for any $\{x_k\}_{k=1}^{\infty}$ and $\{h_k\}_{k=1}^{\infty}$ in $\ell^2(\mathbb{R})$,

$$\boldsymbol{g}'(\{x_k\}_{k=1}^{\infty}; \{h_k\}_{k=1}^{\infty}) = \{y_k\}_{k=1}^{\infty},$$

where for each $k \in \mathbb{N}$,

$$y_k = \begin{cases} h_k & \text{if } x_k > 0, \\ -h_k & \text{if } x_k < 0, \\ h_k & \text{if } x_k = 0 \text{ and } d_k > 0, \\ -h_k & \text{if } x_k = 0 \text{ and } d_k \leq 0. \end{cases}$$

In fact, $\boldsymbol{g} \in \mathcal{L}(\ell^2(\mathbb{R}), \ell^2(\mathbb{R}))$ since for any $\{x_k\}_{k=1}^{\infty}, \{h_k\}_{k=1}^{\infty} \in \ell^2(\mathbb{R})$ and any sequence $\boldsymbol{U} = \{\{u_{kj}\}_{j=1}^{\infty}\}_{k=1}^{\infty}$ of directions in $\ell^2(\mathbb{R})$,

$$\boldsymbol{g}^{(n)}_{\{x_k\}_{k=1}^{\infty}, \boldsymbol{U}}(\{h_k\}_{k=1}^{\infty}) = \{y_{nk}\}_{k=1}^{\infty}, \quad n \in \mathbb{N},$$

where for each $k \in \mathbb{N}$,

$$y_{nk} = \begin{cases} h_k & \text{if } \begin{pmatrix} x_k \\ u_{1k} \\ u_{2k} \\ \vdots \\ u_{nk} \\ d_k \end{pmatrix} \succ \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \\ -h_k & \text{if } \begin{pmatrix} x_k \\ u_{1k} \\ u_{2k} \\ \vdots \\ u_{nk} \\ d_k \end{pmatrix} \preceq \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}. \end{cases}$$

We check that Theorem 4.3.5 holds for the functions $\boldsymbol{f}$ and $\boldsymbol{g}$ and their composition $\phi = \boldsymbol{f} \circ \boldsymbol{g}$ with $\boldsymbol{U} = \{\{u_{kj}\}_{j=1}^{\infty}\}_{k=1}^{\infty}$, a standard basis for $\ell^2(\mathbb{R})$, i.e., for each $k \in \mathbb{N}$,

$$u_{kj} = \begin{cases} 0 & \text{if } j \neq k, \\ 1 & \text{if } j = k. \end{cases}$$

First of all, $\boldsymbol{\phi} = \boldsymbol{g}$ since if $\{x_k\}_{k=1}^{\infty} \in \ell^2(\mathbb{R})$, then

$$\boldsymbol{\phi}\left(\{x_k\}_{k=1}^{\infty}\right) = \boldsymbol{f}\left(\boldsymbol{g}(\{x_k\}_{k=1}^{\infty})\right) = \boldsymbol{f}(\{|x_k|\}_{k=1}^{\infty}) = \{\max\left(0, |x_k|\right)\}_{k=1}^{\infty} = \{|x_k|\}_{k=1}^{\infty} = \boldsymbol{g}\left(\{x_k\}_{k=1}^{\infty}\right).$$

Thus, $\boldsymbol{\phi} \in \mathcal{L}(\ell^2(\mathbb{R}), \ell^2(\mathbb{R}))$. Let $\{x_k\}_{k=1}^{\infty} \in \ell^2(\mathbb{R})$. The sequence $\boldsymbol{V} = \{\{v_{kj}\}_{j=1}^{\infty}\}_{k=1}^{\infty}$ of directions in $\ell^2(\mathbb{R})$ is given by

$$\{v_{kj}\}_{j=1}^{\infty} = \boldsymbol{g}^{(k)}_{\{x_j\}_{j=1}^{\infty}, \boldsymbol{U}}(\{u_{kj}\}_{j=1}^{\infty}), \quad k \in \mathbb{N},$$

i.e., for each $j \in \mathbb{N}$,

$$v_{kj} = \begin{cases} 1 & \text{if } j = k \text{ and } x_j \geq 0 \\ -1 & \text{if } j = k \text{ and } x_j < 0 \\ 0 & \text{if } j \neq k \end{cases}, \quad k \in \mathbb{N}.$$

Let $\{h_k\}_{k=1}^{\infty} \in \ell^2(\mathbb{R})$. Then

$$\boldsymbol{f}^{(0)}_{\boldsymbol{g}\left(\{x_j\}_{j=1}^{\infty}\right),\boldsymbol{V}}(\{h_j\}_{j=1}^{\infty}) = \{z_k\}_{k=1}^{\infty},$$

where for each $k \in \mathbb{N}$,

$$z_k = \begin{cases} h_k & \text{if } \boldsymbol{g}(\{x_j\}_{j=1}^{\infty})_k > 0, \\ 0 & \text{if } \boldsymbol{g}(\{x_j\}_{j=1}^{\infty})_k > 0, \\ h_k & \text{if } \boldsymbol{g}(\{x_j\}_{j=1}^{\infty})_k = 0 \text{ and } h_k > 0, \\ 0 & \text{if } \boldsymbol{g}(\{x_j\}_{j=1}^{\infty})_k = 0 \text{ and } h_k \leq 0. \end{cases}$$

Hence,

$$\boldsymbol{f}^{(n)}_{\boldsymbol{g}\left(\{x_j\}_{j=1}^{\infty}\right),\boldsymbol{V}}(\{h_j\}_{j=1}^{\infty}) = \{z_{nk}\}_{k=1}^{\infty}, \quad n \in \mathbb{N},$$

where for each $k \in \mathbb{N}$,

$$z_{nk} = \begin{cases} h_k & \text{if } \boldsymbol{g}\left(\{x_j\}_{j=1}^{\infty}\right)_k > 0, \\ 0 & \text{if } \boldsymbol{g}\left(\{x_j\}_{j=1}^{\infty}\right)_k < 0, \\ h_k & \text{if } \boldsymbol{g}\left(\{x_j\}_{j=1}^{\infty}\right)_k = 0,\ k \in \{1, \ldots, n\}, \text{ and } x_k \geq 0, \\ 0 & \text{if } \boldsymbol{g}\left(\{x_j\}_{j=1}^{\infty}\right)_k = 0,\ k \in \{1, \ldots, n\}, \text{ and } x_k < 0, \\ h_k & \text{if } \boldsymbol{g}\left(\{x_j\}_{j=1}^{\infty}\right)_k = 0,\ k \notin \{1, \ldots, n\}, \text{ and } h_k > 0, \\ 0 & \text{if } \boldsymbol{g}\left(\{x_j\}_{j=1}^{\infty}\right)_k = 0,\ k \notin \{1, \ldots, n\}, \text{ and } h_k \leq 0. \end{cases}$$

Thus, the L derivative of $\boldsymbol{f}$ at $\boldsymbol{g}\left(\{x_k\}_{k=1}^{\infty}\right)$ for $\boldsymbol{V}$ is identity, i.e.,

$$\boldsymbol{f}^{(\infty)}_{\boldsymbol{g}\left(\{x_k\}_{k=1}^{\infty}\right),\boldsymbol{V}} = \mathbf{1}.$$

Therefore,

$$\boldsymbol{\phi}^{(\infty)}_{\{x_k\}_{k=1}^{\infty},\boldsymbol{U}} = \mathbf{1} \circ \boldsymbol{\phi}^{(\infty)}_{\{x_k\}_{k=1}^{\infty},\boldsymbol{U}} = \boldsymbol{f}^{(\infty)}_{\boldsymbol{g}\left(\{x_k\}_{k=1}^{\infty}\right),\boldsymbol{V}} \circ \boldsymbol{g}^{(\infty)}_{\{x_k\}_{k=1}^{\infty},\boldsymbol{U}},$$

as expected.

4.4 A Remark on the Domains of Functions

In this chapter, we have presented results for functions whose domains are real normed vector spaces or Banach spaces. The same results can be shown to hold for functions whose domains are open subsets of such spaces.

CHAPTER 5
CONCLUSION AND FUTURE WORK

In this thesis, we have presented a generalized theory of differentiation for the class of locally Lipschitz continuous functions. First, we studied the classical theory of differentiation in the Euclidean space setting. We explored the theoretical toolkit for the classical derivative. We also noted the utility of the numerical toolkit for the classical derivative, which includes Newton's method for solving systems of nonlinear equations. Next, we studied Clarke's theory of differentiation, which generalizes the classical notion of derivative to the class of locally Lipschitz continuous functions. In particular, we studied Clarke's derivative objects for a select subclass of locally Lipschitz continuous functions called $\mathcal{PC}^1$ functions, because they demonstrate Clarke's theory of differentiation clearly. We also explored both the theoretical toolkit and the numerical toolkit for Clarke's derivative objects, which preserve the powerful theorems and numerical techniques for the classical derivative. We observed that although Clarke's derivative objects are useful, they are difficult to obtain, even for functions constructed from functions whose Clarke's derivative objects are known, because they obey inclusion-based calculus rules. To address this issue, we introduced Nesterov's lexicographic differentiation and Khan and Barton's lexicographic directional differentiation, i.e., generalized derivatives theories for a subclass of locally Lipschitz continuous functions called L-smooth functions, which help to find elements of Clarke's derivative object systematically. We highlighted that L-derivatives are at best elements of Clarke's derivative object and at least indistinguishable from elements of Clarke's derivative objects as far as nonsmooth numerical tools are concerned. We also noted that unlike Clarke's derivative object, L-derivatives obey an equality-based chain rule. Next, we introduced LD-derivatives for L-smooth functions, whose name, as well as the term "L-derivative," comes from the fact that LD-derivatives of elementary L-smooth functions are naturally expressed with lexicographic ordering. We observed that

LD-derivatives of elementary L-smooth functions can be obtained readily from the definition. Unlike elements of Clarke's derivative object, it is less difficult to obtain LD-derivatives of functions that are constructed from functions whose LD-derivatives are known, thanks to the equality-based chain rule obeyed by LD-derivatives. A proposition relating an LD-derivative to an L-derivative makes it straightforward to calculate an L-derivative once an LD-derivative is known.

The definition of Clarke's derivative object can be generalized to the class of locally Lipschitz continuous functions mapping a Banach space into $\mathbb{R}$ (see page 27 of [3]). Ultimately, the aim of our work is to find a systematic way to access elements of Clarke's generalized derivative object in this setting. To that end, we presented the class of L-smooth functions, a subclass of locally Lipschitz continuous functions mapping between Banach spaces that have Schauder bases. Next, we presented L-derivatives for such functions and showed that these L-derivatives obey an equality-based chain rule, just like the L-derivatives defined earlier. Lastly, we gave examples of L-smooth functions, found their L-derivatives, and demonstrated that they indeed satisfy the chain rule.

In terms of future work, it remains to be shown that L-derivatives for L-smooth functions mapping between Banach spaces that have Schauder bases are at best elements of Clarke's generalized derivative object or at least indistinguishable from elements of Clarke's generalized derivative object as far as nonsmooth numerical tools are concerned. Next, it is desirable to define LD-derivatives for L-smooth functions mapping between Banach spaces that have Schauder bases and to find LD-derivatives of elementary functions for illustration. We also want to show that LD-derivatives satisfy an equality-based chain rule in this setting. Moreover, we want to find a relation between an L-derivative and an LD-derivative that enables us to obtain an L-derivative once an LD-derivative is known.

Once we have a theory of lexicographic directional differentiation for the class of L-smooth functions mapping between Banach spaces that have Schauder bases, we want to apply this machinery to real-world problems arising in science and engineering, including those involving variational inequalities, elliptic equations, and optimal control [2, 3, 8, 14].

REFERENCES

[1] Paul I. Barton, Kamil A. Khan, Peter Stechlinski, and Harry A. J. Watson. Computationally relevant generalized derivatives: theory, evaluation and applications. *Optimization Methods and Software*, 33(4-6):1030–1072, 2018.

[2] Constantin Christof, Christian Clason, Christian Meyer, and Walther Stephan. Optimal control of a non-smooth semilinear elliptic equation. *Mathematical Control and Related Fields*, 8(1):247–276, 2018.

[3] Frank H. Clarke. *Optimization and Nonsmooth Analysis*. SIAM, Philadelphia, PA, 1990.

[4] Juha Heinonen. Lectures on Lipschitz analysis, August 2004.

[5] Kamil A. Khan and Paul I. Barton. Generalized derivatives for solutions of parametric ordinary differential equations with non-differentiable right-hand sides. *Journal of Optimization Theory and Applications*, 163:355–386, 2014.

[6] Kamil A. Khan and Paul I. Barton. A vector forward mode of automatic differentiation for generalized derivative evaluation. *Optimization Methods and Software*, 30(6):1185–1212, 2015.

[7] Boris S. Mordukhovich. *Variational Analysis and Generalized Differentiation I: Basic Theory*. Springer, Berlin, 2006.

[8] Boris S. Mordukhovich. *Variational Analysis and Generalized Differentiation II: Applications*. Springer, Berlin, 2006.

[9] Yu Nesterov. Lexicographic differentiation of nonsmooth functions. *Mathematical Programming*, 104(2-3):669–700, 2005.

[10] Liqun Qi and Jie Sun. A nonsmooth version of Newton's method. *Mathematical Programming*, 58(1-3):353–367, 1993.

[11] Stefan Scholtes. *Introduction to Piecewise Differentiable Equations*. Springer, New York, NY, 2012.

[12] Peter G. Stechlinski and Paul I. Barton. Generalized derivatives of differential-algebraic equations. *Journal of Optimization Theory and Applications*, 171:1–26, 2016.

[13] Peter G. Stechlinski, Kamil A. Khan, and Paul I. Barton. Generalized sensitivity analysis of nonlinear programs. *SIAM Journal on Optimization*, 28(1):272–301, 2018.

[14] Michael Ulbrich. Semismooth Newton methods for operator equations in function spaces. *SIAM Journal on Optimization*, 13(3):805–841, 2003.

BIOGRAPHY OF THE AUTHOR

Jaeho Choi was born in Seoul, South Korea. He graduated from Thomas S. Wootton High School in Rockville, Maryland, and from Williams College in Williamstown, Massachusetts, in 2017 with a Bachelor of Arts degree in Mathematics, Physics, and Economics. He has co-authored a paper titled *On the* L^p *regularity of solutions to the generalized Hunter-Saxton system*, which has been published in *Journal of Discrete Continuous Dynamical Systems Series B*, and a paper titled *The Boltzmann distribution and the quantum-classical correspondence*, which has been published in *Journal of Physics A: Mathematical and Theoretical*. Jaeho Choi is a candidate for the Master of Arts in Mathematics degree in The Department of Mathematics and Statistics from the University of Maine in December 2019.