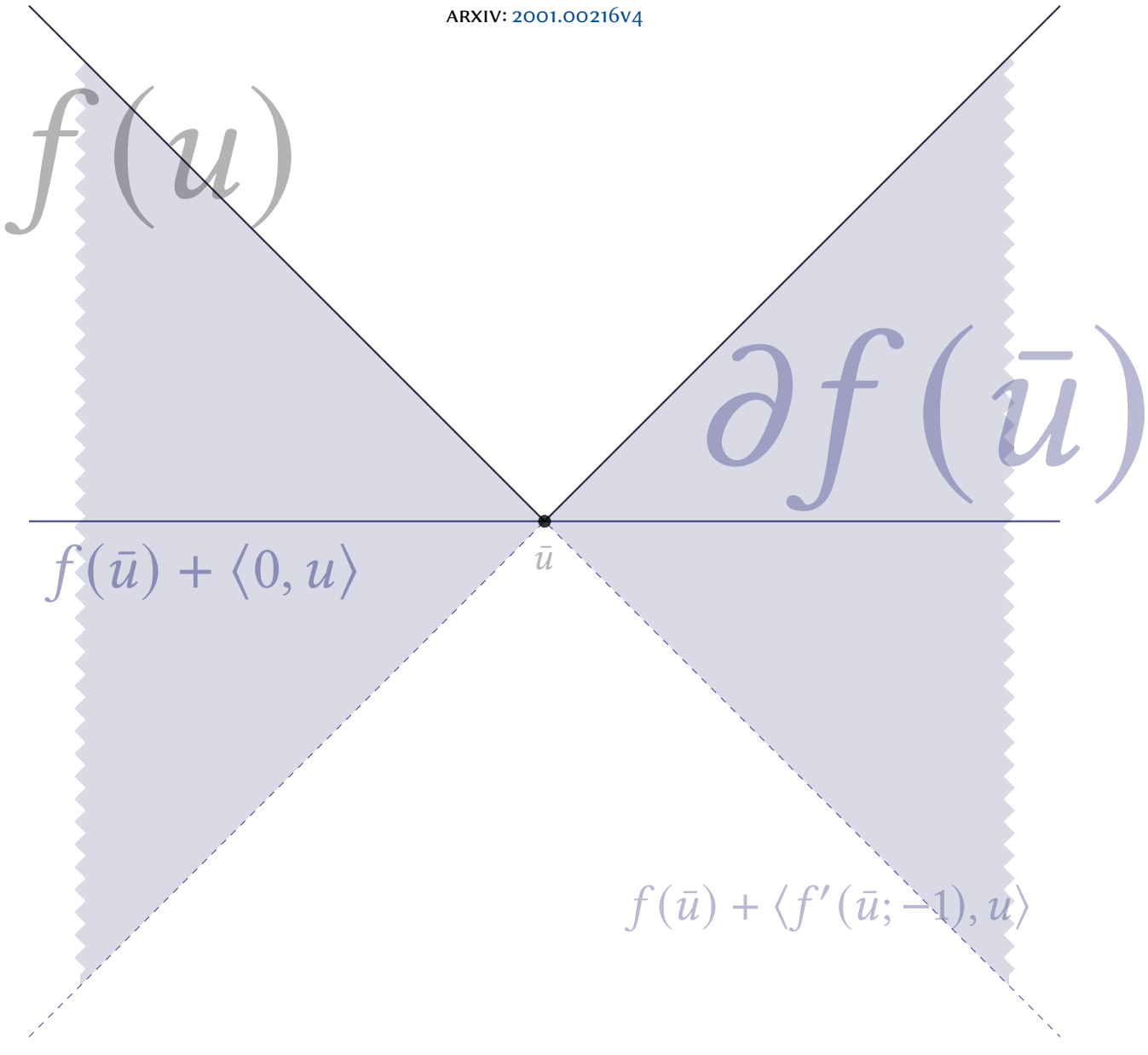
INTRODUCTION TO NONSMOOTH ANALYSIS AND OPTIMIZATION

Christian Clason
c.clason@uni-graz.at
https://homepage.uni-graz.at/c.clason
ORCID: 0000-0002-9948-8426

Tuomo Valkonen
tuomov@iki.fi
https://tuomov.iki.fi
ORCID: 0000-0001-6683-3572

April 15, 2024

ARXIV: 2001.00216v4



Contents

III NONCONVEX ANALYSIS

IV SET-VALUED ANALYSIS

V APPLICATIONS

PREFACE

One of the major applications of classical analysis is *optimization* or the search for minima (or maxima) of a given function; this search may be motivated by the function directly representing an outcome of which lower values are desirable (say, the total cost of an economic production plan) or by the minimizing property indirectly being an essential characterization of a point of interest (say, a physical state as the minimizer of an energy functional, or the solution of an inverse or imaging problem as the minimizer of a regularization functional). In particular, analytical concepts are crucial in every stage of the treatment of optimization problems: continuity properties for showing existence of solutions (that the minimal value is actually attained in a feasible point), first derivatives for intrinsic characterizations of solutions (via *Fermat principles* or *optimality conditions*) and for the numerical solution via *steepest descent* or *gradient methods*, and second derivatives for the numerical solution via *Newton methods* and for deriving stability results, e.g, with respect to computational errors (via implicit function theorems).

However, there are many practically relevant functions that are *not* differentiable, such as the absolute value or maximum function. The goal of nonsmooth analysis is therefore to find generalized derivative concepts that on the one hand allow the above sketched approach for such functions and on the other hand admit a sufficiently rich calculus to give *explicit* derivatives for a sufficiently large class of functions. In this book, we specifically aim at treating problems of the form

$$\min_{x \in C} \frac{1}{p}\|S(x) - z\|_Y^p + \frac{\alpha}{q}\|x\|_X^q \tag{P}$$

for a closed convex *constraint* or *feasible* set $C \subset X$, a (possibly nonlinear but differentiable) operator $S : X \to Y$, $\alpha \geq 0$ and $p, q \in [1, \infty)$ (in particular, $p = 1$ and/or $q = 1$). Such problems are ubiquitous in inverse problems, imaging, and optimal control of differential equations. Hence, we consider optimization in *infinite-dimensional* function spaces; i.e., we are looking for functions as minimizers. The main benefit (beyond the frequently cleaner notation) is that the developed algorithms become *discretization independent*: they can be applied to any (reasonable) finite-dimensional approximation, and the details – in particular, the fineness – of the approximation do not influence the convergence behavior of the algorithms. A special role will be played throughout the book by integral functionals and superposition operators that act pointwise on functions, since these allow transferring the often more explicit finite-dimensional calculus to the infinite-dimensional setting.

Nonsmooth analysis and optimization in finite dimensions has a long history; we refer here only to the classical textbooks [Boyd and Vandenberghe, 2004; Hiriart-Urruty and Lemaréchal, 1993a,b; Mäkelä and Neittaanmäki, 1992; Rockafellar and Wets, 1998; Ruszczyǹski, 2006] as well as the recent [Bagirov et al., 2014; Beck, 2017; Cui and Pang, 2021; Nesterov, 2018; Royset and Wets, 2021]. There also exists a large body of literature on specific nonsmooth optimization problems, in particular ones involving variational inequalities and equilibrium constraints; see, e.g., [Facchinei and Pang, 2003a,b; Outrata et al., 1998]. In contrast, the infinite-dimensional setting is still being actively developed, with monographs and textbooks focusing on either theory [Barbu and Precupanu, 2012; Clarke, 2013, 1990; Dontchev, 2021; Ioffe, 2017; Mordukhovich, 2006, 2018; Penot, 2013; Schirotzek, 2007; Zălinescu, 2002] or algorithms [Ito and Kunisch, 2008; Ulbrich, 2011]. Two exceptions are [Bauschke and Combettes, 2017] and [Peypouquet, 2015], the former containing an impressively comprehensive and integrated treatment of convex analysis and proximal point methods in Hilbert spaces, and the latter giving an equally impressively concise introduction to these topics in normed vector spaces. As this book neared completion, [Bauschke and Moursi, 2023] was published, which serves as a very gentle introduction to convex optimization and first-order methods in Hilbert spaces as treated in [Bauschke and Combettes, 2017]. The aim of this book is thus to draw together results scattered throughout the literature in order to give a unified presentation of theory – both convex and nonconvex – and algorithms – both first- and second-order – in Banach spaces that is suitable for an advanced class on mathematical optimization. In order to do this, we focus on optimization of nonsmooth functionals rather than nonsmooth constraints; in particular, we do not treat optimization with complementarity or equilibrium constraints, which still see significant active development in infinite dimensions. We also restrict the treatment to the two classes of

i) convex functions and

ii) locally Lipschitz continuous functions,

which together cover a wide spectrum of applications. In particular, the first class will lead us to generalized gradient methods, while the second class are the basis for generalized Newton methods. These methods are chosen since they have become increasingly popular in recent years and fit particularly well within the integrated approach of this book. On the other hand, this focus leads us to omit other, more classical, methods and in particular bundle methods, which have very recently seen developments in Hilbert spaces. Here, too, we can only refer to the research literature as well as to the classical books cited above for finite-dimensional treatments. Regarding generalized derivatives of set-valued mappings required for the mentioned stability results, we similarly do not aim for a (possibly fuzzy) general theory and instead restrict ourselves to situations where a regularity condition (one out of the veritable zoo of conditions found in the literature) holds that allows deriving exact results that still apply to problems of the form (P). The general theory can be found in, e.g., [Aubin and Frankowska, 1990; Mordukhovich, 2006, 2018; Rockafellar and Wets, 1998] (to which this book is, among other things, intended as a gentle introduction).

The book is intended for students and researchers with a solid background in analysis and linear algebra and an interest in the mathematical foundations of nonsmooth optimization. Since we deal with infinite-dimensional spaces, some knowledge of functional analysis is assumed, but the necessary background will be summarized in Chapter 1. Similarly, Chapter 2 collects needed fundamental results from the calculus of variations, including the direct method for existence of minimizers and the related notion of lower semicontinuity as well as differential calculus in Banach spaces, where the results on pointwise superposition operators on Lebesgue spaces require elementary (Lebesgue) measure and integration theory. Basic familiarity with classical nonlinear optimization is helpful but not necessary.

In Part II we then start our study of *convex* optimization problems. After introducing convex functionals and their basic properties in Chapter 3, we define our first generalized derivative in Chapter 4: the *convex subdifferential*, which is no longer a single unique derivative but consists of a *set* of equally admissible subderivatives. Nevertheless, we obtain a useful corresponding Fermat principle as well as calculus rules. A particularly useful calculus rule in convex optimization is *Fenchel duality*, which assigns to any optimization problem a *dual problem* that can help treating the original *primal* problem; this is the content of Chapter 5. We change our viewpoint in Chapter 6 slightly to study the subdifferential as a set-valued *monotone operator*, which leads us to the corresponding *resolvent* or *proximal point mapping*, which will later become the basis of all algorithms. The following Chapter 7 discusses the relation between convexity and smoothness of primal and dual problem and introduces the *Moreau–Yosida regularization*, which has better properties in both regards that can be used to accelerate the convergence of algorithms. We turn to these in Chapter 8, where we start by deriving a number of popular first-order methods including *forward-backward splitting* and *primal-dual proximal splitting* (also known as the *Chambolle–Pock method*). Their convergence under rather general assumptions is then shown in Chapter 9. If additional convexity properties hold, we can even show convergence rates for the iterates using a general *testing approach*; this is carried out in Chapter 10. Otherwise we either have to restrict ourselves to more abstract criticality measures as in Chapter 11 or modify the algorithms to include *over-relaxation* or *inertia* as in Chapter 12. One philosophy we here wish to pass to the reader is that the development of optimization methods consists, firstly, in suitable *reformulation* of the problem; secondly, in the *preconditioning* of the raw optimality conditions; and, thirdly, in *testing* with appropriate operators whether this yields fast convergence.

We leave the convex world in Part III. For locally Lipschitz continuous functions, we introduce the *Clarke subdifferential* in Chapter 13 and derive calculus rules. Not only is this useful for obtaining a Fermat principle for problems of the form (P), it is also the basis for defining a further generalized derivative that can be used in place of the Hessian in a generalized Newton method. This *Newton derivative* and the corresponding *semismooth Newton method* is studied in Chapter 14. We also derive and analyze a variant of the primal-dual proximal splitting method suitable for (P) in Chapter 15. We end this part with a short outlook Chapters 16 and 17 to further subdifferential concepts that can lead to sharper

optimality conditions but in general admit a weaker calculus; we will look at some of these in detail in the next part.

To derive stability properties of minimization problems, we need to study the sensitivity of subdifferentials to perturbations and hence generalized derivative concepts for set-valued mappings; this is the goal of Part IV. The construction of the generalized derivatives is geometric, based on *tangent* and *normal cones* introduced in Chapter 18. From these, we obtain *Fréchet* and *limiting (co)derivatives* in Chapter 20 and derive calculus rules for them in Chapters 22 to 25. In particular, we show how to lift the (more extensive) finite-dimensional theory to the special case of pointwise-defined sets and mappings operators on Lebesgue spaces in Chapters 19 and 21. We then address second-order conditions for nonsmooth nonconvex optimization problems in Chapter 26. In Chapter 27, we use these derivatives to characterize Lipschitz-like properties of set-valued mappings, which then are used to obtain the desired stability properties in Chapter 28. We also show in Chapter 29 that these regularity properties imply faster convergence of first-order methods.

Finally, Part V illustrates how these results apply to concrete optimization problems arising in inverse problems and mathematical imaging (Chapters 30 to 32) and in optimal control (Chapters 33 and 34), where we freely admit that the selection of examples is subjective and driven by the authors' interests. These chapters are accompanied by Julia implementations [Clason and Valkonen, 2023] of the discussed algorithms, which can be used to recreate the presented numerical results.

This book can serve as a textbook for several different classes:

(i) an introductory course on convex optimization based on Chapters 3 to 10 (excluding Section 3.3 and results on superposition operators) and adding Chapters 11, 12, and 15 as time permits;

(ii) an intermediate course on nonsmooth optimization based on Chapters 3 to 9 (including Section 3.3 and results on superposition operators) together with Chapters 13, 14, 16, and 17;

(iii) an intermediate course on nonsmooth analysis based on Chapters 3 to 6 together with Chapter 13 and Chapters 16 to 20, adding Chapters 22 to 21 as time permits;

(iv) an advanced course on set-valued analysis based on Chapters 16 to 29.

This book is based in part on such graduate lectures given by the first author in 2014 (in slightly different form) and 2016–2017 at the University of Duisburg-Essen and by the second author at the University of Cambridge in 2015 and Escuela Politécnica Nacional in Quito in 2020. Shorter seminars were also delivered at the University of Jyväskylä and the Escuela Politécnica Nacional in 2017. Part IV of the book was also used in a course on variational analysis at the EPN in 2019. Parts of the book were also taught by both authors at the Winter School "Modern Methods in Nonsmooth Optimization" organized

by Christian Kanzow and Daniel Wachsmuth at the University of Würzburg in February 2018, for which the notes were further adapted and extended. As such, much (but not all) of this material is classical. In particular, Chapters 3 to 7 as well as Chapter 13 are based on [Attouch et al., 2014; Barbu and Precupanu, 2012; Bauschke and Combettes, 2017; Brokate, 2014; Clarke, 2013; Schirotzek, 2007], Chapter 14 is based on [Ito and Kunisch, 2008; Schiela, 2008; Ulbrich, 2002], Chapter 16 is extracted from [Mordukhovich, 2006], and Chapters 18 to 25 are adapted from [Mordukhovich, 2006; Rockafellar and Wets, 1998]. Parts of Chapter 17 are adapted from [Ioffe, 2017], and other parts are original work. On the other hand, Chapters 8 to 12 as well as Chapters 15, 21, and 29 are adapted from [Clason et al., 2019; Valkonen, 2020b, 2021c], [Clason and Valkonen, 2017b], and [Valkonen, 2021c], respectively.

Finally, we would like to thank Sebastian Angerhausen, Andreas Habring, Fernando Jimenez Torres, Heikki von Koch, Anton Schiela, Ensio Suonperä, Diego Vargas Jaramillo, Bjørn Jensen, Daniel Wachsmuth, and in particular Gerd Wachsmuth for carefully reading parts of the manuscript, finding mistakes and bits that could be expressed more clearly, and making helpful suggestions. All remaining errors are of course our own.

Essen/Graz and Quito/Helsinki, March 2024

Part I

BACKGROUND

1 FUNCTIONAL ANALYSIS

Functional analysis is the study of infinite-dimensional vector spaces and of the operators acting between them, and has since its foundations in the beginning of the 20th century grown into the *lingua franca* of modern applied mathematics. In this chapter we collect the basic concepts and results (and, more importantly, fix notations) from linear functional analysis that will be used throughout the rest of the book. For details and proofs, the reader is referred to the standard literature, e.g., [Alt, 2016; Brezis, 2010; Rynne and Youngson, 2008], or to [Clason, 2020a].

1.1 NORMED VECTOR SPACES

In the following, X will denote a real vector space. A mapping $\|\cdot\| : X \to \mathbb{R}^+ := [0, \infty)$ is called a *norm* (on X), if for all $x \in X$ there holds

(i) $\|\lambda x\| = |\lambda|\|x\|$ for all $\lambda \in \mathbb{R}$;

(ii) $\|x + y\| \leq \|x\| + \|y\|$ for all $y \in X$;

(iii) $\|x\| = 0$ if and only if $x = 0 \in X$.

Example 1.1. (i) The following mappings define norms on $X = \mathbb{R}^N$:

$$\|x\|_p = \left(\sum_{i=1}^{N} |x_i|^p \right)^{1/p}, \qquad 1 \leq p < \infty,$$
$$\|x\|_\infty = \max_{i=1,\dots,N} |x_i|.$$

(ii) The following mappings define norms on $X = \ell^p$ (the space of real-valued sequences for which these terms are finite):

$$\|x\|_p = \left(\sum_{i=1}^{\infty} |x_i|^p \right)^{1/p}, \qquad 1 \leq p < \infty,$$
$$\|x\|_\infty = \sup_{i=1,\dots,\infty} |x_i|.$$

(iii) The following mappings define norms on $X = L^p(\Omega)$ (the space of real-valued measurable functions on the domain $\Omega \subset \mathbb{R}^d$ for which these terms are finite):

$$\|u\|_{L^p} = \left(\int_\Omega |u(x)|^p \right)^{1/p}, \qquad 1 \le p < \infty,$$
$$\|u\|_{L^\infty} = \operatorname*{ess\,sup}_{x \in \Omega} |u(x)|,$$

where ess sup stands for the essential supremum; for details on these definitions, see, e.g., [Alt, 2016].

(iv) The following mapping defines a norm on $X = C(\overline{\Omega})$ (the space of continuous functions on $\overline{\Omega}$):

$$\|u\|_C = \sup_{x \in \overline{\Omega}} |u(x)|.$$

An analogous norm is defined on $X = C_0(\Omega)$ (the space of continuous functions on Ω with compact support), if the supremum is taken only over the space of continuous functions on Ω with compact support), if the supremum is taken only over $x \in \Omega$.

If $\|\cdot\|$ is a norm on X, the tuple $(X, \|\cdot\|)$ is called a *normed vector space*, and one frequently denotes this by writing $\|\cdot\|_X$. If the norm is canonical (as in Example 1.1 (ii)–(iv)), it is often omitted, and one speaks simply of "the normed vector space X".

Two norms $\|\cdot\|_1$, $\|\cdot\|_2$ are called *equivalent* on X, if there are constants $c_1, c_2 > 0$ such that

$$c_1\|x\|_2 \le \|x\|_1 \le c_2\|x\|_2 \qquad \text{for all } x \in X.$$

If X is finite-dimensional, all norms on X are equivalent. However, the corresponding constants c_1 and c_2 may depend on the dimension N of X; avoiding such dimension-dependent constants is one of the main reasons to consider optimization in infinite-dimensional spaces.

If $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ are normed vector spaces with $X \subset Y$, we call X *continuously embedded* in Y, denoted by $X \hookrightarrow Y$, if there exists a $C > 0$ with

$$\|x\|_Y \le C\|x\|_X \qquad \text{for all } x \in X.$$

For example, if $\Omega \subset \mathbb{R}^d$ is a bounded domain, $L^q(\Omega) \hookrightarrow L^p(\Omega)$ for every $1 \le p \le q \le \infty$.

A norm directly induces a notion of convergence, the so-called *strong convergence*. A sequence $\{x_n\}_{n\in\mathbb{N}} \subset X$ *converges* (*strongly* in X) to a $x \in X$, denoted by $x_n \to x$, if

$$\lim_{n\to\infty} \|x_n - x\|_X = 0.$$

A set $U \subset X$ is called

- *closed*, if for every convergent sequence $\{x_n\}_{n\in\mathbb{N}} \subset U$ the limit $x \in X$ is an element of U as well;
- *compact*, if every sequence $\{x_n\}_{n\in\mathbb{N}} \subset U$ contains a convergent subsequence $\{x_{n_k}\}_{k\in\mathbb{N}}$ with limit $x \in U$.

A mapping $F : X \to Y$ is *continuous* if and only if $x_n \to x$ implies $F(x_n) \to F(x)$. If $x_n \to x$ and $F(x_n) \to y$ imply that $F(x) = y$ (i.e., graph $F \subset X \times Y$ is a closed set), we say that F has *closed graph*.

Further we define for later use for $x \in X$ and $r > 0$

- the *open ball* $\mathbb{O}(x, r) := \{z \in X \mid \|x - z\|_X < r\}$ and
- the *closed ball* $\mathbb{B}(x, r) := \{z \in X \mid \|x - z\|_X \leq r\}$.

The closed ball around $0 \in X$ with radius 1 is also referred to as the *unit ball* $\mathbb{B}_X$. A set $U \subset X$ is called

- *open*, if for all $x \in U$ there exists an $r > 0$ with $\mathbb{O}(x, r) \subset U$ (i.e., all $x \in U$ are *interior points* of U);
- *bounded*, if it is contained in $\mathbb{B}(0, r)$ for a $r > 0$;
- *convex*, if for any $x, y \in U$ and $\lambda \in [0, 1]$ also $\lambda x + (1 - \lambda) y \in U$.

In normed vector spaces it always holds that the complement of an open set is closed and vice versa (i.e., the closed sets in the sense of topology are exactly the (sequentially) closed set as defined above). The definition of a norm directly implies that both open and closed balls are convex.

For arbitrary U, we denote by $\operatorname{cl} U$ the *closure* of U, defined as the smallest closed set that contains U (which coincides with the set of all limit points of convergent sequences in U); we write $\operatorname{int} U$ for the *interior* of U, which is the largest open set contained in U; and we write $\operatorname{bd} U := \operatorname{cl} U \setminus \operatorname{int} U$ for the *boundary* of U. Finally, we write $\operatorname{co} U$ for the *convex hull* of U, defined as the smallest convex set that contains U.

A normed vector space X is called *complete* if every Cauchy sequence in X is convergent; in this case, X is called a *Banach space*. All spaces in Example 1.1 are Banach spaces. Convex subsets of Banach spaces have the following useful property which derives from the Baire Theorem.

Lemma 1.2. *Let X be a Banach space and $U \subset X$ be closed and convex. Then*

$$\operatorname{int} U = \{x \in U \mid \text{for all } h \in X \text{ there is a } \delta > 0 \text{ with } x + th \in U \text{ for all } t \in [0, \delta]\}.$$

The set on the right-hand side is called *algebraic interior* or *core*. For this reason, Lemma 1.2 is sometimes referred to as the "core-int Lemma". Note that the inclusion "$\subset$" always holds in normed vector spaces due to the definition of interior points via open balls.

We now consider mappings between normed vector spaces. In the following, let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed vector spaces, $U \subset X$, and $F : U \to Y$ be a mapping. We denote by

- $\ker F := \{x \in U \mid F(x) = 0\}$ the *kernel* or *null space* of F;
- $\operatorname{ran} F := \{F(x) \in Y \mid x \in U\}$ the *range* of F;
- $\operatorname{graph} F := \{(x, y) \in X \times Y \mid y = F(x)\}$ the *graph* of F.

We call $F : U \to Y$

- *continuous* at $x \in U$, if for all $\varepsilon > 0$ there exists a $\delta > 0$ with
$$\|F(x) - F(z)\|_Y \leq \varepsilon \qquad \text{for all } z \in \mathbb{O}(x, \delta) \cap U;$$
- *Lipschitz continuous*, if there exists an $L > 0$ (called *Lipschitz constant*) with
$$\|F(x_1) - F(x_2)\|_Y \leq L\|x_1 - x_2\|_X \qquad \text{for all } x_1, x_2 \in U.$$
- *locally Lipschitz continuous at* $x \in U$, if there exists a $\delta > 0$ and a $L = L(x, \delta) > 0$ with
$$\|F(x) - F(\tilde{x})\|_Y \leq L\|x - \tilde{x}\|_X \qquad \text{for all } \tilde{x} \in \mathbb{O}(x, \delta) \cap U;$$
- *locally Lipschitz continuous near* $x \in U$, if there exists a $\delta > 0$ and a $L = L(x, \delta) > 0$ with
$$\|F(x_1) - F(x_2)\|_Y \leq L\|x_1 - x_2\|_X \qquad \text{for all } x_1, x_2 \in \mathbb{O}(x, \delta) \cap U.$$
We will refer to the $\mathbb{O}(x, \delta)$ as the *Lipschitz neighborhood* of x (for F). If F is locally Lipschitz continuous near every $x \in U$, we call F *locally Lipschitz continuous on* U.

If $T : X \to Y$ is linear, continuity is equivalent to the existence of a constant $C > 0$ with

$$\|Tx\|_Y \leq C\|x\|_X \qquad \text{for all } x \in X.$$

For this reason, continuous linear mappings are called *bounded*; one speaks of a bounded linear *operator*. The space $\mathbb{L}(X; Y)$ of bounded linear operators is itself a normed vector space if endowed with the *operator norm*

$$\|T\|_{\mathbb{L}(X;Y)} = \sup_{x \in X \setminus \{0\}} \frac{\|Tx\|_Y}{\|x\|_X} = \sup_{\|x\|_X = 1} \|Tx\|_Y = \sup_{\|x\|_X \leq 1} \|Tx\|_Y$$

(which is equal to the smallest possible constant C in the definition of continuity). If $(Y, \|\cdot\|_Y)$ is a Banach space, then so is $(\mathbb{L}(X;Y), \|\cdot\|_{\mathbb{L}(X;Y)})$.

Finally, if $T \in \mathbb{L}(X;Y)$ is bijective, the inverse $T^{-1} : Y \to X$ is continuous if and only if there exists a $c > 0$ with

$$c\|x\|_X \leq \|Tx\|_Y \qquad \text{for all } x \in X.$$

In this case, $\|T^{-1}\|_{\mathbb{L}(Y;X)} = c^{-1}$ for the largest possible choice of c.

1.2 DUAL SPACES, SEPARATION, AND WEAK CONVERGENCE

Of particular importance to us is the special case $\mathbb{L}(X;Y)$ for $Y = \mathbb{R}$, the space of *bounded linear functionals* on X. In this case, $X^* := \mathbb{L}(X;\mathbb{R})$ is called the *dual space* (or just *dual*) of X. For $x^* \in X^*$ and $x \in X$, we set

$$\langle x^*, x\rangle_X := x^*(x) \in \mathbb{R}.$$

This *duality pairing* indicates that we can also interpret it as x acting on x^*, which will become important later. The definition of the operator norm immediately implies that

$$\langle x^*, x\rangle_X \leq \|x^*\|_{X^*}\|x\|_X \qquad \text{for all } x \in X, x^* \in X^*. \tag{1.1}$$

In many cases, the dual of a Banach space can be identified with another known Banach space.

Example 1.3. (i) $(\mathbb{R}^N, \|\cdot\|_p)^* \cong (\mathbb{R}^N, \|\cdot\|_q)$ with $p^{-1} + q^{-1} = 1$, where we set $0^{-1} = \infty$ and $\infty^{-1} = 0$. The duality pairing is given by

$$\langle x^*, x\rangle_p = \sum_{i=1}^{N} x_i^* x_i.$$

(ii) $(\ell^p)^* \cong (\ell^q)$ for $1 < p < \infty$. The duality pairing is given by

$$\langle x^*, x\rangle_p = \sum_{i=1}^{\infty} x_i^* x_i.$$

Furthermore, $(\ell^1)^* = \ell^\infty$, but $(\ell^\infty)^*$ is not a sequence space.

(iii) Analogously, $L^p(\Omega)^* \cong L^q(\Omega)$ with $p^{-1} + q^{-1} = 1$ for $1 < p < \infty$. The duality pairing is given by

$$\langle u^*, u\rangle_p = \int_\Omega u^*(x)u(x)\,dx.$$

Furthermore, $L^1(\Omega)^* \cong L^\infty(\Omega)$, but $L^\infty(\Omega)^*$ is not a function space.

(iv) $C_0(\Omega)^* \cong \mathcal{M}(\Omega)$, the space of *Radon measure*; it contains among others the Lebesgue measure as well as Dirac measures δ_x for $x \in \Omega$, defined via $\delta_x(u) = u(x)$ for $u \in C_0(\Omega)$. The duality pairing is given by

$$\langle u^*, u \rangle_C = \int_\Omega u(x)\, du^*.$$

A central result on dual spaces is the Hahn–Banach Theorem, which comes in both an algebraic and a geometric version.

Theorem 1.4 (Hahn–Banach, algebraic). *Let X be a normed vector space and $x \in X \setminus \{0\}$. Then there exists a $x^* \in X^*$ with*

$$\|x^*\|_{X^*} = 1 \qquad \text{and} \qquad \langle x^*, x \rangle_X = \|x\|_X.$$

Theorem 1.5 (Hahn–Banach, geometric). *Let X be a normed vector space and $A, B \subset X$ be convex, nonempty, and disjoint.*

(i) If A is open, there exists an $x^* \in X^*$ and a $\lambda \in \mathbb{R}$ with

$$\langle x^*, x_1 \rangle_X < \lambda \leq \langle x^*, x_2 \rangle_X \qquad \text{for all } x_1 \in A, x_2 \in B.$$

(ii) If A is closed and B is compact, there exists an $x^* \in X^*$ and a $\lambda \in \mathbb{R}$ with

$$\langle x^*, x_1 \rangle_X \leq \lambda < \langle x^*, x_2 \rangle_X \qquad \text{for all } x_1 \in A, x_2 \in B.$$

Particularly the geometric version – also referred to as *separation theorems* – is of crucial importance in convex analysis. We will also require their following variant, which is known as *Eidelheit Theorem*.

Corollary 1.6. *Let X be a normed vector space and $A, B \subset X$ be convex and nonempty. If the interior* $\operatorname{int} A$ *of A is nonempty and disjoint with B, there exists an $x^* \in X^* \setminus \{0\}$ and a $\lambda \in \mathbb{R}$ with*

$$\langle x^*, x_1 \rangle_X \leq \lambda \leq \langle x^*, x_2 \rangle_X \qquad \text{for all } x_1 \in A, x_2 \in B.$$

Proof. Theorem 1.5 (i) yields the existence of x^* and λ satisfying the claim for all $x_1 \in \operatorname{int} A$; this inequality is even strict, which also implies $x^* \neq 0$. It thus remains to show that the first inequality also holds for the remaining $x_1 \in A \setminus \operatorname{int} A$. Since $\operatorname{int} A$ is nonempty, there exists an $x_0 \in \operatorname{int} A$, i.e., there is an $r > 0$ with $\mathbb{O}(x_0, r) \subset A$. The convexity of A then implies that $t\tilde{x} + (1-t)x_1 \in A$ for all $\tilde{x} \in \mathbb{O}(x_0, r)$ and $t \in [0, 1]$. Hence,

$$t\mathbb{O}(x_0, r) + (1-t)x = \mathbb{O}(tx_0 + (1-t)x_1, tr) \subset A,$$

and in particular $z(t) := tx_0 + (1-t)x_1 \in \operatorname{int} A$ for all $t \in (0,1)$.

We can thus find a sequence $\{z_n\}_{n\in\mathbb{N}} \subset \operatorname{int} A$ (e.g., $z_n = z(n^{-1})$) with $z_n \to x_1$. Due to the continuity of $x^* \in X^* = \mathbb{L}(X;\mathbb{R})$ we can thus pass to the limit $n \to \infty$ and obtain

$$\langle x^*, x_1\rangle_X = \lim_{n\to\infty} \langle x^*, z_n\rangle_X \leq \lambda. \qquad \square$$

This can be used to characterize a normed vector space by its dual. For example, a direct consequence of Theorem 1.4 is that the norm on a Banach space can be expressed as an operator norm.

Corollary 1.7. *Let X be a Banach space. Then for all $x \in X$,*

$$\|x\|_X = \sup_{\|x^*\|_{X^*}\leq 1} |\langle x^*, x\rangle_X|,$$

and the supremum is attained.

A vector $x \in X$ can therefore be considered as a linear and, by (1.1), bounded functional on X^*, i.e., as an element of the *bidual* $X^{**} := (X^*)^*$. The embedding $X \hookrightarrow X^{**}$ is realized by the *canonical injection*

$$J : X \to X^{**}, \qquad \langle Jx, x^*\rangle_{X^*} := \langle x^*, x\rangle_X \quad \text{for all } x^* \in X^*. \tag{1.2}$$

Clearly, J is linear; Theorem 1.4 furthermore implies that $\|Jx\|_{X^{**}} = \|x\|_X$. If the canonical injection is surjective and we can thus identify X^{**} with X, the space X is called *reflexive*. All finite-dimensional spaces are reflexive, as are Example 1.1 (ii) and (iii) for $1 < p < \infty$; however, ℓ^1, ℓ^∞ as well as $L^1(\Omega), L^\infty(\Omega)$ and $C(\overline{\Omega})$ are not reflexive. In general, a normed vector space is reflexive if and only if its dual space is reflexive.

The following consequence of the separation Theorem 1.5 will be of crucial importance in Part IV. For a set $A \subset X$, we define the *polar cone*

$$A^\circ := \{x^* \in X^* \mid \langle x^*, x\rangle_X \leq 0 \text{ for all } x \in A\},$$

cf. Figure 1.1. Similarly, we define for $B \subset X^*$ the *prepolar cone*

$$B_\circ := \{x \in X \mid \langle x^*, x\rangle_X \leq 0 \text{ for all } x^* \in B\}.$$

The *bipolar cone* of $A \subset X$ is then defined as

$$A^{\circ\circ} := (A^\circ)_\circ \subset X.$$

(If X is reflexive, $A^{\circ\circ} = (A^\circ)^\circ$.) For the following statement about polar cones, recall that a set $C \subset X$ is called a *cone* if $x \in C$ and $\lambda > 0$ implies that $\lambda x \in C$ (such that (pre-, bi-)polar cones are indeed cones).

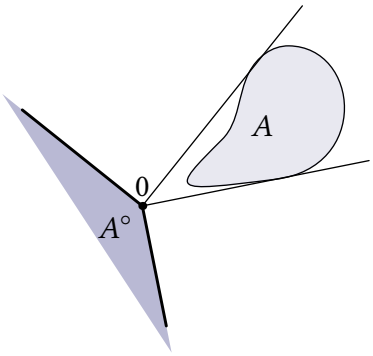


Figure 1.1: The polar cone A° is the normal cone at zero to the smallest cone containing A.

Theorem 1.8 (bipolar theorem). *Let X be a normed vector space and $A \subset X$. Then*

(i) A° is closed and convex;

(ii) $A \subset A^{\circ\circ}$;

(iii) if $A \subset B$, then $B^\circ \subset A^\circ$;

(iv) if C is a closed and convex cone with $0 \in C$, then $C = C^{\circ\circ}$.

Proof. *(i):* This follows directly from the definition and the continuity of the duality pairing.

(ii): Let $x \in A$ be arbitrary. Then by definition of the polar cone, every $x^* \in A^\circ$ satisfies

$$\langle x^*, x\rangle_X \leq 0,$$

i.e., $x \in (A^\circ)_\circ = A^{\circ\circ}$.

(iii): This is immediate from the definition.

(iv): By (ii), we only need to prove $C^{\circ\circ} \subset C$ which we do by contradiction. Assume therefore that there exists $x \in C^{\circ\circ} \setminus \{0\}$ with $x \notin C$. Applying Theorem 1.5 (ii) to the nonempty (due to (ii)) closed, and convex set $C^{\circ\circ}$ and the disjoint compact convex set $\{x\}$, we obtain $x^* \in X^* \setminus \{0\}$ and $\lambda \in \mathbb{R}$ such that

$$\langle x^*, \tilde{x}\rangle_X \leq \lambda < \langle x^*, x\rangle_X \quad \text{for all } \tilde{x} \in C. \tag{1.3}$$

Since C is a cone, the first inequality must also hold for $t\tilde{x} \in C$ for every $t > 0$. This implies that

$$\langle x^*, \tilde{x}\rangle_X \leq t^{-1}\lambda \to 0 \quad \text{for } t \to \infty,$$

i.e., $\langle x^*, \tilde{x}\rangle_X \leq 0$ for all $\tilde{x} \in C$ must hold, i.e., $x^* \in C^\circ$. On the other hand, if $\lambda < 0$, we obtain by the same argument that

$$\langle x^*, \tilde{x}\rangle_X \leq t^{-1}\lambda \to -\infty \quad \text{for } t \to 0,$$

which cannot hold. Hence, we can take $\lambda = 0$ in (1.3). Together, we obtain from $x \in C^{\circ\circ}$ the contradiction

$$0 < \langle x^*, x \rangle_X \leq 0. \qquad \square$$

The duality pairing induces further notions of convergence.

(i) A sequence $\{x_n\}_{n\in\mathbb{N}} \subset X$ *converges weakly* (in X) to $x \in X$, denoted by $x_n \rightharpoonup x$, if

$$\langle x^*, x_n \rangle_X \to \langle x^*, x \rangle_X \qquad \text{for all } x^* \in X^*.$$

(ii) A sequence $\{x_n^*\}_{n\in\mathbb{N}} \subset X^*$ *converges weakly-** (in X^*) to $x^* \in X^*$, denoted by $x_n^* \overset{*}{\rightharpoonup} x^*$, if

$$\langle x_n^*, x \rangle_X \to \langle x^*, x \rangle_X \qquad \text{for all } x \in X.$$

Weak convergence generalizes the concept of componentwise convergence in $\mathbb{R}^N$, which – as can be seen from the proof of the Heine–Borel Theorem – is the appropriate concept in the context of compactness. Strong convergence in X implies weak convergence by continuity of the duality pairing; in the same way, strong convergence in X^* implies weak-$*$ convergence. If X is reflexive, weak and weak-$*$ convergence (both in $X = X^{**}$) coincide. In finite-dimensional spaces, all convergence notions coincide.

Weakly convergent sequences are always bounded; if X is a Banach space, so are weakly-$*$ convergent sequences. If $x_n \to x$ and $x_n^* \overset{*}{\rightharpoonup} x^*$ or $x_n \rightharpoonup x$ and $x_n^* \to x^*$, then $\langle x_n^*, x_n \rangle_X \to \langle x^*, x \rangle_X$. However, the duality pairing of weak(-$*$) convergent sequences does not converge in general.

As for strong convergence, one defines weak(-$*$) continuity and closedness of mappings as well as weak(-$*$) closedness and compactness of sets. The last property is of fundamental importance in optimization; its characterization is therefore a central result of this chapter.

Theorem 1.9 (Eberlein–Šmulyan). *If X is a normed vector space, then $\mathbb{B}_X$ is weakly compact if and only if X is reflexive.*

Hence in a reflexive space, all bounded sequences contain a *weakly* (but in general not strongly) convergent subsequence. Note that weak closedness is a *stronger* claim than closedness, since the property has to hold for more sequences. For convex sets, however, both concepts coincide.

Lemma 1.10. *Let X be a normed vector space and $U \subset X$ be convex. Then U is weakly closed if and only if U is closed.*

Proof. Weakly closed sets are always closed since a convergent sequence is also weakly convergent. Let now $U \subset X$ be convex closed and nonempty (otherwise nothing has to be shown) and consider a sequence $\{x_n\}_{n\in\mathbb{N}} \subset U$ with $x_n \rightharpoonup x \in X$. Assume that $x \in X \setminus U$. Then the sets U and $\{x\}$ satisfy the premise of Theorem 1.5 (ii); we thus find an $x^* \in X^*$ and a $\lambda \in \mathbb{R}$ with

$$\langle x^*, x_n\rangle_X \leq \lambda < \langle x^*, x\rangle_X \quad \text{for all } n \in \mathbb{N}.$$

Passing to the limit $n \to \infty$ in the first inequality yields the contradiction

$$\langle x^*, x\rangle_X < \langle x^*, x\rangle_X. \qquad \square$$

If X is not reflexive (e.g., $X = L^\infty(\Omega)$), we have to turn to weak-$*$ convergence.

Theorem 1.11 (Banach–Alaoglu). *If X is a separable normed vector space (i.e., contains a countable dense subset), then $\mathbb{B}_{X^*}$ is weakly-$*$ compact.*

By the Weierstraß Approximation Theorem, both $C(\overline{\Omega})$ and $L^p(\Omega)$ for $1 \leq p < \infty$ are separable; also, ℓ^p is separable for $1 \leq p < \infty$. Hence, bounded and weakly-$*$ closed balls in ℓ^∞, $L^\infty(\Omega)$, and $\mathcal{M}(\Omega)$ are weakly-$*$ compact. However, these spaces themselves are not separable.

We also have the following straightforward improvement of Theorem 1.8 (i).

Lemma 1.12. *Let X be a separable normed vector space and $A \subset X$. Then A° is weakly-$*$ closed and convex.*

Note, however, that arbitrary closed convex sets in nonreflexive spaces do *not* have to be weakly-$*$ closed.

Finally, we will also need the following "weak-$*$" separation theorem, whose proof is analogous to the proof of Theorem 1.5 (using the fact that the linear weakly-$*$ continuous functionals are exactly those of the form $x^* \mapsto \langle x^*, x\rangle_X$ for some $x \in X$); see also [Rudin, 2021, Theorem 3.4(b)].

Theorem 1.13. *Let X be a normed vector space and $A \subset X^*$ be a nonempty, convex, and weakly-$*$ closed subset and $x^* \in X^* \setminus A$. Then there exist an $x \in X$ and a $\lambda \in \mathbb{R}$ with*

$$\langle z^*, x\rangle_X \leq \lambda < \langle x^*, x\rangle_X \qquad \text{for all } z^* \in A.$$

Since a normed vector space is characterized by its dual, this is also the case for linear operators acting on this space. For any $T \in \mathbb{L}(X;Y)$, the *adjoint operator* $T^* \in \mathbb{L}(Y^*;X^*)$ is defined via

$$\langle T^* y^*, x\rangle_X = \langle y^*, Tx\rangle_Y \qquad \text{for all } x \in X,\ y^* \in Y^*.$$

It always holds that $\|T^*\|_{\mathbb{L}(Y^*;X^*)} = \|T\|_{\mathbb{L}(X;Y)}$. Furthermore, the continuity of T implies that T^* is weakly-$*$ continuous (and T weakly continuous).

1.3 HILBERT SPACES

Especially strong duality properties hold in Hilbert spaces. A mapping $(\cdot \mid \cdot) : X \times X \to \mathbb{R}$ on a vector space X over $\mathbb{R}$ is called *inner product*, if

(i) $(\alpha x + \beta y \mid z) = \alpha(x \mid z) + \beta(y \mid z)$ for all $x, y, z \in X$ and $\alpha, \beta \in \mathbb{R}$;

(ii) $(x \mid y) = (y \mid x)$ for all $x, y \in X$;

(iii) $(x \mid x) \geq 0$ for all $x \in X$ with equality if and only if $x = 0$.

An inner product induces a norm

$$\|x\|_X := \sqrt{(x \mid x)_X},$$

which satisfies the *Cauchy–Schwarz inequality*

$$(x \mid y)_X \leq \|x\|_X \|y\|_X.$$

If X is complete with respect to the induced norm (i.e., if $(X, \|\cdot\|_X)$ is a Banach space), then X is called a *Hilbert space*; if the inner product is canonical, it is frequently omitted, and the Hilbert space is simply denoted by X. The spaces in Example 1.3 (i)–(iii) for $p = 2 (= q)$ are all Hilbert spaces, where the inner product coincides with the duality pairing and induces the canonical norm.

Directly from the definition of the induced norm we obtain the *binomial expansion*

$$\|x + y\|_X^2 = \|x\|_X^2 + 2(x \mid y)_X + \|y\|_X^2, \tag{1.4}$$

which in turn can be used to verify the *three-point identity*

$$(x - y \mid x - z)_X = \frac{1}{2}\|x - y\|_X^2 - \frac{1}{2}\|y - z\|_X^2 + \frac{1}{2}\|x - z\|_X^2 \quad \text{for all } x, y, z \in X. \tag{1.5}$$

(This can be seen as a generalization of the classical Pythagorean theorem in plane geometry.)

The relevant point in our context is that the dual of a Hilbert space X can be identified with X itself.

Theorem 1.14 (Fréchet–Riesz). *Let X be a Hilbert space. Then for each $x^* \in X^*$ there exists a unique $z_{x^*} \in X$ with $\|x^*\|_{X^*} = \|z_{x^*}\|_X$ and*

$$\langle x^*, x \rangle_X = (x \mid z_{x^*})_X \qquad \text{for all } x \in X.$$

The element z_{x^*} is called *Riesz representation* of x^*. The (linear) mapping $J_X : X^* \to X$, $x^* \mapsto z_{x^*}$, is called *Riesz isomorphism*, and can be used to show that every Hilbert space is reflexive.

Theorem 1.14 allows to use the inner product instead of the duality pairing in Hilbert spaces. For example, a sequence $\{x_n\}_{n\in\mathbb{N}} \subset X$ converges weakly to $x \in X$ if and only if

$$(x_n \,|\, z)_X \to (x \,|\, z)_X \qquad \text{for all } z \in X.$$

This implies that if $x_n \rightharpoonup x$ and in addition $\|x_n\|_X \to \|x\|_X$ (in which case we say that x_n *strictly converges* to x),

$$\|x_n - x\|_X^2 = \|x_n\|_X^2 - 2(x_n \,|\, x)_X + \|x\|_X^2 \to 0, \tag{1.6}$$

i.e., $x_n \to x$. A normed vector space in which strict convergence implies strong convergence is said to have the *Radon–Riesz property*.

Similar statements hold for linear operators on Hilbert spaces. For a linear operator $T \in \mathbb{L}(X;Y)$ between Hilbert spaces X and Y, the *Hilbert space adjoint operator* $T^\star \in \mathbb{L}(Y;X)$ is defined via

$$(T^\star y \,|\, x)_X = (Tx \,|\, y)_Y \qquad \text{for all } x \in X,\, y \in Y.$$

If $T^\star = T$, the operator T is called *self-adjoint*. A self-adjoint operator is called *positive definite*, if there exists a $c > 0$ such that

$$(Tx \,|\, x)_X \geq c\|x\|_X^2 \quad \text{for all } x \in X.$$

In this case, T has a bounded inverse T^{-1} with $\|T^{-1}\|_{\mathbb{L}(X;X)} \leq c^{-1}$. We will also use the notation $S \geq T$ for two operators $S, T : X \to X$ if

$$(Sx \,|\, x)_X \geq (Tx \,|\, x)_X \quad \text{for all } x \in X.$$

Hence T is positive definite if and only if $T \geq c\mathrm{Id}$ for some $c > 0$; if $T \geq 0$, we say that T is merely *positive semi-definite*.

The Hilbert space adjoint is related to the (Banach space) adjoint via $T^\star = J_X T^* J_Y^{-1}$. If the context is obvious, we will not distinguish the two in notation. Similarly, we will also – by a moderate abuse of notation – use angled brackets to denote inner products in Hilbert spaces except where we need to refer to both at the same time (which will rarely be the case, and the danger of confusing inner products with elements of a product space is much greater).

2 CALCULUS OF VARIATIONS

We first consider the question of the existence of solutions to optimization problems of the form

$$\min_{x \in U} F(x)$$

for a (nonlinear) functional $F : U \to \mathbb{R}$ and a subset U of a Banach space X. Answering such questions is one of the goals of the *calculus of variations*.

Note that we don't require F to be defined on all of X; this is important for example when F involves the solution of a nonlinear partial differential equation which may only exist if x is sufficiently small. For the purposes of existence of a minimizer, however, we do not need to distinguish whether U represents such a domain of definition or an additional constraint in the optimization problem. In both cases, we can get rid of the constraint by extending F to all of X with the value ∞ by setting

$$\overline{F} : X \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}, \qquad \overline{F}(x) = \begin{cases} F(x) & \text{if } x \in U, \\ \infty & \text{if } x \in X \setminus U. \end{cases}$$

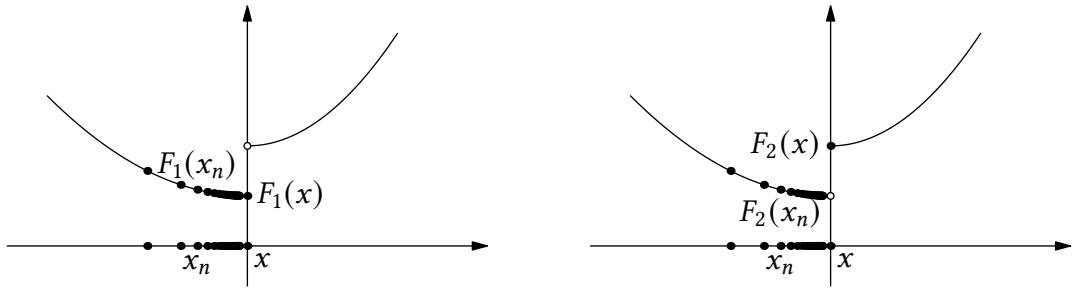
We extend the usual arithmetic on $\mathbb{R}$ to $\overline{\mathbb{R}}$ by letting $t < \infty$ and $t + \infty = \infty$ for all $t \in \mathbb{R}$; subtraction and multiplication of negative numbers with ∞ and in particular $F(x) = -\infty$ is not allowed, however. Thus if there is any $x \in U$ at all, a minimizer $\bar{x}$ of $\overline{F}$ necessarily must lie in U and coincide with a minimizer of F over U.

2.1 THE DIRECT METHOD

Our goal now is to find conditions under which a functional $F : X \to \overline{\mathbb{R}}$ attains a (real-valued) minimum over X. First, there clearly must exist a point with finite value. We call the set on which F is finite the *effective domain*

$$\operatorname{dom} F := \{x \in X \mid F(x) < \infty\}.$$

If $\operatorname{dom} F \neq \emptyset$, the functional F is called *proper*.



(a) F_1 is lower semicontinuous at x

(b) F_2 is *not* lower semicontinuous at x

Figure 2.1: Illustration of lower semicontinuity: two functions $F_1, F_2 : \mathbb{R} \to \mathbb{R}$ and a sequence $\{x_n\}_{n\in\mathbb{N}}$ realizing their (identical) limes inferior.

Next, we require a form of continuity to prevent the function from "jumping over" possible minima. We call F *lower semicontinuous* in $x \in X$ if

$$F(x) \leq \liminf_{n\to\infty} F(x_n) \qquad \text{for every } \{x_n\}_{n\in\mathbb{N}} \subset X \text{ with } x_n \to x,$$

see Figure 2.1, where F_2 is an example for a function that is not lower semicontinuous and does not attain a minimum. Analogously, we define *weakly(-∗) lower semicontinuous* functionals via weakly(-∗) convergent sequences.

Finally, we need to prevent the function from having a "minimum at infinity". Here we use the following property: We call F *coercive* if for every sequence $\{x_n\}_{n\in\mathbb{N}} \subset X$ with $\|x_n\|_X \to \infty$ we also have $F(x_n) \to \infty$.

We now have everything at hand to prove the central existence result in the calculus of variations. The strategy for its proof is known as the *direct method*.[1]

Theorem 2.1. *Let X be a reflexive Banach space and $F : X \to \overline{\mathbb{R}}$ be proper, coercive, and weakly lower semicontinuous. Then the minimization problem*

$$\min_{x\in X} F(x)$$

has a solution $\bar{x} \in \operatorname{dom} F$.

Proof. The proof can be separated into three steps.

[1] This strategy is applied so often in the literature that one usually just writes "Existence of a minimizer follows from the direct method." or even just "Existence follows from standard arguments." The basic idea goes back to Hilbert; the version based on lower semicontinuity which we use here is due to Leonida Tonelli (1885–1946), who through it had a lasting influence on the modern calculus of variations.

(i) *Pick a minimizing sequence.*

Since F is proper, there exists an $M := \inf_{x\in X} F(x) < \infty$ (although $M = -\infty$ is not excluded so far). We can thus find a sequence $\{y_n\}_{n\in\mathbb{N}} \subset \operatorname{ran} F \setminus \{\infty\} \subset \mathbb{R}$ with $y_n \to M$, i.e., there exists a sequence $\{x_n\}_{n\in\mathbb{N}} \subset X$ with

$$F(x_n) \to M = \inf_{x\in X} F(x).$$

Such a sequence is called *minimizing sequence*. Note that from the convergence of $\{F(x_n)\}_{n\in\mathbb{N}}$ we cannot conclude the convergence of $\{x_n\}_{n\in\mathbb{N}}$ (yet).

(ii) *Show that the minimizing sequence contains a weakly convergent subsequence.*

Assume to the contrary that $\{x_n\}_{n\in\mathbb{N}}$ is unbounded, i.e., that $\|x_n\|_X \to \infty$ for $n \to \infty$. The coercivity of F then implies that $F(x_n) \to \infty$ as well, in contradiction to $F(x_n) \to M < \infty$ by definition of the minimizing sequence. Hence, the sequence is bounded, i.e., there is an $M > 0$ with $\|x_n\|_X \le M$ for all $n \in \mathbb{N}$. In particular, $\{x_n\}_{n\in\mathbb{N}} \subset \mathbb{B}(0, M)$. The Eberlein–Šmulyan Theorem 1.9 therefore implies the existence of a weakly converging subsequence $\{x_{n_k}\}_{k\in\mathbb{N}}$ with limit $\bar{x} \in X$. (This limit is a candidate for the minimizer.)

(iii) *Show that its limit is a minimizer.*

From the definition of the minimizing sequence, we also have $F(x_{n_k}) \to M$ for $k \to \infty$. Together with the weak lower semicontinuity of F and the definition of the infimum we thus obtain

$$\inf_{x\in X} F(x) \le F(\bar{x}) \le \liminf_{k\to\infty} F(x_{n_k}) = M = \inf_{x\in X} F(x) < \infty.$$

This implies that $\bar{x} \in \operatorname{dom} F$ and that $\inf_{x\in X} F(x) = F(\bar{x}) > -\infty$. Hence, the infimum is attained in $\bar{x}$ which is therefore the desired minimizer. □

Remark 2.2. If X is not reflexive but the dual of a separable Banach space, we can argue analogously for weakly-$*$ lower semicontinuous functionals using the Banach–Alaoglu Theorem 1.11

Note how the topology on X used in the proof is restricted in step (ii) and (iii): Step (ii) profits from a coarse topology (in which more sequences are convergent), while step (iii) profits from a fine topology (the fewer sequences are convergent, the easier it is to satisfy the lim inf conditions). Since in the cases of interest to us no more than boundedness of a minimizing sequence can be expected, we cannot use a finer than the weak topology. We thus have to ask whether a sufficiently large class of (interesting) functionals are weakly lower semicontinuous.

A first example is the class of bounded linear functionals: For any $x^* \in X^*$, the functional

$$F : X \to \mathbb{R}, \qquad x \mapsto \langle x^*, x\rangle_X,$$

is weakly continuous by definition of weak convergence and hence *a fortiori* weakly lower semicontinuous. Another advantage of (weak) lower semicontinuity is that it is preserved under certain operations.

Lemma 2.3. *Let X and Y be Banach spaces and $F : X \to \overline{\mathbb{R}}$ be weakly(-$*$) lower semicontinuous. Then the following functionals are weakly(-$*$) lower semicontinuous as well:*

(i) αF *for all* $\alpha \geq 0$*;*

(ii) $F + G$ *for* $G : X \to \overline{\mathbb{R}}$ *weakly(-$*$) lower semicontinuous;*

(iii) $\varphi \circ F$ *for* $\varphi : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ *lower semicontinuous and monotonically increasing.*

(iv) $F \circ \Phi$ *for* $\Phi : Y \to X$ *weakly(-$*$) continuous, i.e.,* $y_n \rightharpoonup^{(*)} y$ *implies* $\Phi(y_n) \rightharpoonup^{(*)} \Phi(y)$*;*

(v) $x \mapsto \sup_{i \in I} F_i(x)$ *with* $F_i : X \to \overline{\mathbb{R}}$ *weakly(-$*$) lower semicontinuous for all* $i \in I$ *and an arbitrary set* I*.*

Note that (v) does *not* hold for continuous functions.

Proof. We only show the claim for the case of weak lower semicontinuity; the statements for weak-$*$ lower semicontinuity follow by the same arguments.

Statements (i) and (ii) follow directly from the properties of the limes inferior.

For statement (iii), it first follows from the monotonicity of φ and the weak lower semicontinuity of F that $x_n \rightharpoonup x$ implies

$$\varphi(F(x)) \leq \varphi(\liminf_{n\to\infty} F(x_n)).$$

It remains to show that the right-hand side can be bounded by $\liminf_{n\to\infty} \varphi(F(x_n))$. For that purpose, we consider the subsequence $\{\varphi(F(x_{n_k})\}_{k\in\mathbb{N}}$ which realizes the lim inf, i.e., for which $\liminf_{n\to\infty} \varphi(F(x_n)) = \lim_{k\to\infty} \varphi(F(x_{n_k}))$. By passing to a further subsequence which we index by k' for simplicity, we can also obtain that $\liminf_{k\to\infty} F(x_{n_k}) = \lim_{k'\to\infty} F(x_{n_{k'}})$. Since the lim inf restricted to a subsequence can never be smaller than that of the full sequence, the monotonicity of φ together with its weak lower semicontinuity now implies that

$$\varphi(\liminf_{n\to\infty} F(x_n)) \leq \varphi(\lim_{k'\to\infty} F(x_{n_{k'}})) \leq \liminf_{k'\to\infty} \varphi(F(x_{n_{k'}})) = \liminf_{n\to\infty} \varphi(F(x_n)),$$

where we have used in the last step that a subsequence of a convergent sequence has the same limit (which coincides with the lim inf).

Statement (iv) follows directly from the weak continuity of Φ, as $y_n \rightharpoonup y$ implies that $x_n := \Phi(y_n) \rightharpoonup \Phi(y) =: x$, and the lower semicontinuity of F yields

$$F(\Phi(y)) \leq \liminf_{n\to\infty} F(\Phi(y_n)).$$

Finally, let $\{x_n\}_{n\in\mathbb{N}}$ be a weakly converging sequence with limit $x \in X$. Then the definition of the supremum implies that

$$F_j(x) \le \liminf_{n\to\infty} F_j(x_n) \le \liminf_{n\to\infty} \sup_{i\in I} F_i(x_n) \qquad \text{for all } j \in I.$$

Taking the supremum over all $j \in I$ on both sides yields statement (v). □

Corollary 2.4. *If X is a Banach space, then the norm $\|\cdot\|_X$ is proper, coercive, and weakly lower semicontinuous. Similarly, the dual norm $\|\cdot\|_{X^*}$ is proper, coercive, and weakly-$*$ lower semicontinuous.*

Proof. Coercivity and $\operatorname{dom}\|\cdot\|_X = X$ follow directly from the definition. Weak lower semicontinuity follows from Lemma 2.3 (v) and Corollary 1.7 since

$$\|x\|_X = \sup_{\|x^*\|_{X^*}\le 1} |\langle x^*, x\rangle_X|.$$

The claim for $\|\cdot\|_{X^*}$ follows analogously using the definition of the operator norm in place of Corollary 1.7. □

Another frequently occurring functional is the *indicator function*[2] of a set $U \subset X$, defined as

$$\delta_U(x) = \begin{cases} 0 & x \in U, \\ \infty & x \in X \setminus U. \end{cases}$$

The purpose of this definition is of course to write the minimization of a functional $F : X \to \mathbb{R}$ (i.e., defined on all of X) under the additional constraint $x \in U$ to the minimization of $\overline{F} := F + \delta_U$ over X. The following result is therefore important for showing the existence of such a constrained minimizer.

Lemma 2.5. *Let X be a Banach space and $U \subset X$. Then $\delta_U : X \to \overline{\mathbb{R}}$ is*

(i) proper if U is nonempty;

(ii) weakly lower semicontinuous if U is convex and closed;

(iii) coercive if U is bounded.

Proof. Statement (i) is clear. For (ii), consider a weakly converging sequence $\{x_n\}_{n\in\mathbb{N}} \subset X$ with limit $x \in X$. If $x \in U$, then $\delta_U \ge 0$ immediately yields

$$\delta_U(x) = 0 \le \liminf_{n\to\infty} \delta_U(x_n).$$

[2] not to be confused with the *characteristic function* $\mathbb{1}_U$ with $\mathbb{1}_U(x) = 1$ for $x \in U$ and 0 else

Let now $x \notin U$. Since U is convex and closed and hence by Lemma 1.10 also weakly closed, there must be a $N \in \mathbb{N}$ with $x_n \notin U$ for all $n \geq N$ (otherwise we could – by passing to a subsequence if necessary – construct a sequence with $x_n \rightharpoonup x \in U$, in contradiction to the assumption). Thus, $\delta_U(x_n) = \infty$ for all $n \geq N$, and therefore

$$\delta_U(x) = \infty = \liminf_{n\to\infty} \delta_U(x_n).$$

For (iii), let U be bounded, i.e., there exist an $M > 0$ with $U \subset \mathbb{B}(0, M)$. If $\|x_n\|_X \to \infty$, then there exists an $N \in \mathbb{N}$ with $\|x_n\|_X > M$ for all $n \geq N$, and thus $x_n \notin \mathbb{B}(0, M) \supset U$ for all $n \geq N$. Hence, $\delta_U(x_n) \to \infty$ as well. □

2.2 DIFFERENTIAL CALCULUS IN NORMED VECTOR SPACES

To characterize minimizers of functionals on infinite-dimensional spaces using the Fermat principle, we transfer the classical derivative concepts to normed vector spaces.

Let X and Y be normed vector spaces, $F : X \to Y$ be a mapping, and $x, h \in X$ be given.

- If the one-sided limit

  $$F'(x; h) := \lim_{t \searrow 0} \frac{F(x + th) - F(x)}{t} \in Y$$

  (where $t \searrow 0$ denotes the limit for arbitrary positive decreasing null sequences) exists, it is called the *directional derivative* of F in x in direction h.

- If $F'(x; h)$ exists for all $h \in X$ and

  $$DF(x) : X \to Y, \qquad h \mapsto F'(x; h)$$

  defines a bounded linear operator, we call F *Gâteaux differentiable* (at x) and $DF(x) \in \mathbb{L}(X; Y)$ its *Gâteaux derivative*.

- If additionally

  $$\lim_{\|h\|_X \to 0} \frac{\|F(x + h) - F(x) - DF(x)h\|_Y}{\|h\|_X} = 0,$$

  then F is called *Fréchet differentiable* (in x) and $F'(x) := DF(x) \in \mathbb{L}(X; Y)$ its *Fréchet derivative*.

- If additionally the mapping $F' : X \to \mathbb{L}(X; Y)$ is (Lipschitz) continuous, we call F *(Lipschitz) continuously differentiable*.

The difference between Gâteaux and Fréchet differentiable lies in the approximation error of F near x by $F(x) + DF(x)h$: While it only has to be bounded in $\|h\|_X$ – i.e., linear in $\|h\|_X$ – for a Gâteaux differentiable function, it has to be superlinear in $\|h\|_X$ if F is Fréchet differentiable. (For a *fixed* direction h, this is of course also the case for Gâteaux differentiable functions; Fréchet differentiability thus additionally requires a uniformity in h.) We also point out that continuous differentiability always entails Fréchet differentiability.

Remark 2.6. Sometimes a weaker notion than continuous differentiability is used. A mapping $F : X \to Y$ is called *strictly differentiable* in x if

$$\lim_{\substack{y \to x \\ \|h\|_X \to 0}} \frac{\|F(y+h) - F(y) - F'(x)h\|_Y}{\|h\|_X} = 0. \tag{2.1}$$

The benefit of this definition over that of continuous differentiability is that the limit process is now in the function F rather than the derivative F'; strict differentiability can therefore hold if every neighborhood of x contains points where F is not differentiable. However, if F is differentiable everywhere in a neighborhood of x, then F is strictly differentiable if and only if F' is continuous; see [Dontchev and Rockafellar, 2014, Proposition 1D.7]. Although many results of Chapters 13 to 25 actually hold under the weaker assumption of strict differentiability, we will therefore work only with the more standard notion of continuous differentiability.

If F is Gâteaux differentiable, the Gâteaux derivative can be computed via

$$DF(x)h = \left(\tfrac{d}{dt} F(x + th)\right)\Big|_{t=0}.$$

Bounded linear operators $F \in \mathbb{L}(X; Y)$ are obviously Fréchet differentiable with derivative $F'(x) = F \in \mathbb{L}(X; Y)$ for all $x \in X$. Further derivatives can be obtained through the usual calculus, whose proof in normed vector spaces is exactly as in $\mathbb{R}^N$. As an example, we prove a chain rule.

Theorem 2.7. *Let X, Y, and Z be normed vector spaces, and let $F : X \to Y$ be Fréchet differentiable at $x \in X$ and $G : Y \to Z$ be Fréchet differentiable at $y := F(x) \in Y$. Then $G \circ F$ is Fréchet differentiable at x and*

$$(G \circ F)'(x) = G'(F(x))F'(x).$$

Proof. For $h \in X$ with $x + h \in \operatorname{dom} F$ we have

$$(G \circ F)(x+h) - (G \circ F)(x) = G(F(x+h)) - G(F(x)) = G(y+g) - G(y)$$

with $g := F(x+h) - F(x)$. The Fréchet differentiability of G thus implies that

$$\|(G \circ F)(x+h) - (G \circ F)(x) - G'(y)g\|_Z = r_1(\|g\|_Y)$$

with $r_1(t)/t \to 0$ for $t \to 0$. The Fréchet differentiability of F further implies

$$\|g - F'(x)h\|_Y = r_2(\|h\|_X)$$

with $r_2(t)/t \to 0$ for $t \to 0$. In particular,

$$\|g\|_Y \le \|F'(x)h\|_Y + r_2(\|h\|_X). \tag{2.2}$$

Hence, with $c := \|G'(F(x))\|_{\mathbb{L}(Y;Z)}$ we have

$$\|(G \circ F)(x+h) - (G \circ F)(x) - G'(F(x))F'(x)h\|_Z \le r_1(\|g\|_Y) + c\, r_2(\|h\|_X).$$

If $\|h\|_X \to 0$, we obtain from (2.2) and $F'(x) \in \mathbb{L}(X;Y)$ that $\|g\|_Y \to 0$ as well, and the claim follows. □

A similar rule for Gâteaux derivatives does not hold, however.

Of special importance in Part IV will be the following inverse function theorem, whose proof can be found, e.g., in [Renardy and Rogers, 2004, Theorem 10.4].

Theorem 2.8 (inverse function theorem). *Let $F : X \to Y$ be a continuously differentiable mapping between the Banach spaces X and Y and $x \in X$. If $F'(x) : X \to Y$ is bijective, then there exists an open set $V \subset Y$ with $F(x) \in V$ such that $F^{-1} : V \to X$ exists and is continuously differentiable.*

Of particular relevance in optimization is of course the special case $F : X \to \mathbb{R}$, where $DF(x) \in \mathbb{L}(X;\mathbb{R}) = X^*$ (if the Gâteaux derivative exists). Following the usual notation from Section 1.2, we will then write $F'(x;h) = \langle DF(x), h\rangle_X$ for the directional derivative in direction $h \in X$. Our first result is the classical Fermat principle characterizing minimizers of a differentiable functions.

Theorem 2.9 (Fermat principle). *Let $F : X \to \mathbb{R}$ be Gâteaux differentiable and $\bar{x} \in X$ be a local minimizer of F. Then $DF(\bar{x}) = 0$, i.e.,*

$$\langle DF(\bar{x}), h\rangle_X = 0 \qquad \text{for all } h \in X.$$

Proof. Let $h \in X$ be arbitrary. Since $\bar{x}$ is a local minimizer, the core–int Lemma 1.2 implies that there exists an $\varepsilon > 0$ such that $F(\bar{x}) \le F(\bar{x} + th)$ for all $t \in (0, \varepsilon)$, i.e.,

$$0 \le \frac{F(\bar{x}+th) - F(\bar{x})}{t} \to F'(\bar{x};h) = \langle DF(\bar{x}), h\rangle_X \quad \text{for } t \to 0, \tag{2.3}$$

where we have used the Gâteaux differentiability and hence directional differentiability of F. Since the right-hand side is linear in h, the same argument for $-h$ yields $\langle DF(\bar{x}), h\rangle_X \le 0$ and therefore the claim. □

We will also need the following version of the mean value theorem.

Theorem 2.10. *Let $F : X \to \mathbb{R}$ be Fréchet differentiable. Then for all $x, h \in X$,*

$$F(x+h) - F(x) = \int_0^1 \langle F'(x+th), h\rangle_X \, dt.$$

Proof. Consider the scalar function

$$f : [0,1] \to \mathbb{R}, \qquad t \mapsto F(x+th).$$

From Theorem 2.7 we obtain that f (as a composition of mappings on normed vector spaces) is differentiable with

$$f'(t) = \langle F'(x+th), h\rangle_X,$$

and the fundamental theorem of calculus in $\mathbb{R}$ yields that

$$F(x+h) - F(x) = f(1) - f(0) = \int_0^1 f'(t)\, dt = \int_0^1 \langle F'(x+th), h\rangle_X \, dt. \qquad \square$$

As in classical analysis, this result is useful for relating local and pointwise properties of smooth functions. A typical example is the following lemma.

Lemma 2.11. *Let $F : X \to Y$ be continuously Fréchet differentiable in a neighborhood U of $x \in X$. Then F is locally Lipschitz continuous near $x \in U$.*

Proof. Since $F' : U \to \mathbb{L}(X;Y)$ is continuous in U, there exists a $\delta > 0$ with $\|F'(z) - F'(x)\|_{\mathbb{L}(X;Y)} \le 1$ and hence $\|F'(z)\|_{\mathbb{L}(X;Y)} \le 1 + \|F'(x)\|_{\mathbb{L}(X;Y)}$ for all $z \in \mathbb{B}(x,\delta) \subset U$. For any $x_1, x_2 \in \mathbb{B}(x,\delta)$ we also have $x_2 + t(x_1 - x_2) \in \mathbb{B}(x,\delta)$ for all $t \in [0,1]$ (since balls in normed vector spaces are convex), and hence Theorem 2.10 implies that

$$\begin{aligned} \|F(x_1) - F(x_2)\|_Y &\le \int_0^1 \|F'(x_2 + t(x_1 - x_2))\|_{\mathbb{L}(X;Y)} \|x_1 - x_2\|_X \, dt \\ &\le (1 + \|F'(x)\|_{\mathbb{L}(X;Y)}) \|x_1 - x_2\|_X, \end{aligned}$$

and thus local Lipschitz continuity near x with constant $L = \frac{1}{2}(1 + \|F'(x)\|_{\mathbb{L}(X;Y)})$. $\square$

Note that since the Gâteaux derivative of $F : X \to \mathbb{R}$ is an element of X^*, it cannot be added to elements in X (as required for, e.g., a steepest descent method). However, in Hilbert spaces (and in particular in $\mathbb{R}^N$), we can use the Fréchet–Riesz Theorem 1.14 to identify $DF(x) \in X^*$ with an element $\nabla F(x) \in X$, called the *gradient* of F at x, in a canonical way via

$$\langle DF(x), h\rangle_X = (\nabla F(x) \mid h)_X \qquad \text{for all } h \in X.$$

We illustrate this with a simple example.

Example 2.12. Let $F(x) = \frac{1}{2}\|x\|_X^2 = \frac{1}{2}(x \mid x)_X$. Then we have for all $x, h \in X$ that

$$F'(x;h) = \lim_{t \searrow 0} \frac{\frac{1}{2}(x+th \mid x+th)_X - \frac{1}{2}(x \mid x)_X}{t} = (x \mid h)_X = \langle DF(x), h\rangle_X,$$

since the inner product is linear in h for fixed x. Hence, the squared norm is Gâteaux differentiable at every $x \in X$ with derivative $DF(x) = h \mapsto (x \mid h)_X \in X^*$; it is even Fréchet differentiable since

$$\lim_{\|h\|_X \to 0} \frac{\left|\frac{1}{2}\|x+h\|_X^2 - \frac{1}{2}\|x\|_X^2 - (x,h)_X\right|}{\|h\|_X} = \lim_{\|h\|_X \to 0} \frac{1}{2}\|h\|_X = 0.$$

The gradient $\nabla F(x) \in X$ by definition is given by

$$(\nabla F(x) \mid h)_X = \langle DF(x), h\rangle_X = (x \mid h)_X \qquad \text{for all } h \in X,$$

i.e., $\nabla F(x) = x$.

The following example demonstrates how the gradient (in contrast to the derivative) depends on the inner product on X – which may be different from the inner product inducing the squared norm.

Example 2.13. Let $M \in \mathbb{L}(X;X)$ be self-adjoint and positive definite (and thus continuously invertible). Then $(x \mid y)_Z := (Mx \mid y)_X$ also defines an inner product on the vector space X and induces an (equivalent) norm $\|x\|_Z := (x \mid x)_Z^{1/2}$ on X. Hence $(X, (\cdot \mid \cdot)_Z)$ is a Hilbert space as well, which we will denote by Z. Consider now the functional $\tilde{F} : Z \to \mathbb{R}$ with $\tilde{F}(x) := \frac{1}{2}\|x\|_X^2$ (which is well-defined since $\|\cdot\|_X$ is also an equivalent norm on Z). Then, the derivative $D\tilde{F}(x) \in Z^*$ is still given by $\langle D\tilde{F}(x), h\rangle_Z = (x \mid h)_X$ for all $h \in Z$ (or, equivalently, for all $h \in X$ since we defined Z via the same vector space). However, $\nabla\tilde{F}(x) \in Z$ is now characterized by

$$(x \mid h)_X = \langle D\tilde{F}(x), h\rangle_Z = (\nabla\tilde{F}(x) \mid h)_Z = (M\nabla\tilde{F}(x) \mid h)_X \qquad \text{for all } h \in Z,$$

i.e., $\nabla\tilde{F}(x) = M^{-1}x \neq \nabla F(x)$.

(The situation is even more delicate if M is only positive definite on a subspace, as in the case of $X = L^2(\Omega)$ and $Z = H^1(\Omega)$.)

2.3 SUPERPOSITION OPERATORS

A special class of operators on function spaces arise from pointwise application of a real-valued function, e.g., $u(x) \mapsto \sin(u(x))$. We thus consider for $f : \Omega \times \mathbb{R} \to \mathbb{R}$ with $\Omega \subset \mathbb{R}^d$

open and bounded as well as $p, q \in [1, \infty]$ the corresponding *superposition* or *Nemytskii operator*

$$F : L^p(\Omega) \to L^q(\Omega), \qquad [F(u)](x) = f(x, u(x)) \quad \text{for almost every } x \in \Omega. \tag{2.4}$$

For this operator to be well-defined requires certain restrictions on f. We call $f : \Omega \times \mathbb{R} \to \mathbb{R}$ a *Carathéodory function* if

(i) for all $z \in \mathbb{R}$, the mapping $x \mapsto f(x, z)$ is measurable;

(ii) for almost every $x \in \Omega$, the mapping $z \mapsto f(x, z)$ is continuous.

We additionally require the following growth condition: For given $p, q \in [1, \infty)$ there exist $a \in L^q(\Omega)$ and $b \in L^\infty(\Omega)$ with

$$|f(x, z)| \leq a(x) + b(x)|z|^{p/q}. \tag{2.5}$$

Under these conditions, F is well-defined and even continuous.

Theorem 2.14. *If the Carathéodory function $f : \Omega \times \mathbb{R} \to \mathbb{R}$ satisfies the growth condition (2.5) for $p, q \in [1, \infty)$, then the superposition operator $F : L^p(\Omega) \to L^q(\Omega)$ defined via (2.4) is continuous.*

Proof. We sketch the essential steps; a complete proof can be found in, e.g., [Appell and Zabrejko, 1990, Theorems 3.1, 3.7]. First, one shows for given $u \in L^p(\Omega)$ the measurability of $F(u)$ using the Carathéodory properties. It then follows from (2.5) and the triangle inequality that

$$\|F(u)\|_{L^q} \leq \|a\|_{L^q} + \|b\|_{L^\infty} \||u|^{p/q}\|_{L^q} = \|a\|_{L^q} + \|b\|_{L^\infty} \|u\|_{L^p}^{p/q} < \infty,$$

i.e., $F(u) \in L^q(\Omega)$.

To show continuity, we consider a sequence $\{u_n\}_{n \in \mathbb{N}} \subset L^p(\Omega)$ with $u_n \to u \in L^p(\Omega)$. Then there exists a subsequence, again denoted by $\{u_n\}_{n \in \mathbb{N}}$, that converges pointwise almost everywhere in Ω, as well as a $v \in L^p(\Omega)$ with $|u_n(x)| \leq |v(x)| + |u_1(x)| =: g(x)$ for all $n \in \mathbb{N}$ and almost every $x \in \Omega$ (see, e.g., [Alt, 2016, Lemma 3.22 as well as (3-14) in the proof of Theorem 3.17]). The continuity of $z \mapsto f(x, z)$ then implies $F(u_n) \to F(u)$ pointwise almost everywhere as well as

$$|[F(u_n)](x)| \leq a(x) + b(x)|u_n(x)|^{p/q} \leq a(x) + b(x)|g(x)|^{p/q} \quad \text{for almost every } x \in \Omega.$$

Since $g \in L^p(\Omega)$, the right-hand side defines a function in $L^q(\Omega)$, and we can apply Lebesgue's dominated convergence theorem to deduce that $F(u_n) \to F(u)$ in $L^q(\Omega)$. As this argument can be applied to any subsequence, the whole sequence must converge to $F(u)$, which yields the claimed continuity. □

In fact, the growth condition (2.5) is also necessary for continuity; see [Appell and Zabrejko, 1990, Theorem 3.2]. In addition, it is straightforward to show that for $p = q = \infty$, the growth condition (2.5) (with $p/q := 0$ in this case) implies that F is even locally Lipschitz continuous.

Similarly, one would like to show that differentiability of f implies differentiability of the corresponding superposition operator F, ideally with "pointwise" derivative $[F'(u)h](x) = f'(u(x))h(x)$ (compare Example 1.3 (iii)). However, this does not hold in general; for example, the superposition operator defined by $f(x, z) = \sin(z)$ is *not* differentiable at $u = 0$ for $1 \leq p = q < \infty$. The reason is that for a Fréchet differentiable superposition operator $F : L^p(\Omega) \to L^q(\Omega)$ and a direction $h \in L^p(\Omega)$, the pointwise(!) product $F'(u)h$ has to be in $L^q(\Omega)$. This leads to additional conditions on the superposition operator F' defined by f', which is known as *two-norm discrepancy*.

Theorem 2.15. *Let $f : \Omega \times \mathbb{R} \to \mathbb{R}$ be a Carathéodory function that satisfies the growth condition (2.5) for $1 \leq q < p < \infty$. If the partial derivative f'_z is a Carathéodory function as well and satisfies (2.5) for $p' = p - q$, the superposition operator $F : L^p(\Omega) \to L^q(\Omega)$ is continuously Fréchet differentiable, and its derivative in $u \in L^p(\Omega)$ in direction $h \in L^p(\Omega)$ is given by*

$$[F'(u)h](x) = f'_z(x, u(x))h(x) \qquad \text{for almost every } x \in \Omega.$$

Proof. Theorem 2.14 yields that for $r := \frac{pq}{p-q}$ (i.e., $\frac{r}{p} = \frac{q}{p'}$), the superposition operator

$$G : L^p(\Omega) \to L^r(\Omega), \qquad [G(u)](x) = f'_z(x, u(x)) \quad \text{for almost every } x \in \Omega,$$

is well-defined and continuous. The Hölder inequality further implies that for any $u \in L^p(\Omega)$,

$$\|G(u)h\|_{L^q} \leq \|G(u)\|_{L^r} \|h\|_{L^p} \qquad \text{for all } h \in L^p(\Omega), \tag{2.6}$$

i.e., the pointwise multiplication $h \mapsto G(u)h$ defines a bounded linear operator $DF(u) : L^p(\Omega) \to L^q(\Omega)$.

Let now $h \in L^p(\Omega)$ be arbitrary. Since $z \mapsto f(x, z)$ is continuously differentiable by assumption, the classical mean value theorem together with the properties of the integral (in particular, monotonicity, Jensen's inequality on $[0, 1]$, and Fubini's theorem) and (2.6)

implies that

$$\begin{aligned}
&\|F(u+h) - F(u) - DF(u)h\|_{L^q} \\
&\qquad = \left(\int_\Omega |f(x, u(x)+h(x)) - f(x, u(x)) - f_z'(x, u(x))h(x)|^q \, dx \right)^{\frac{1}{q}} \\
&\qquad = \left(\int_\Omega \left| \int_0^1 f_z'(x, u(x)+th(x))h(x) \, dt - f_z'(x, u(x))h(x) \right|^q \, dx \right)^{\frac{1}{q}} \\
&\qquad \leq \left(\int_0^1 \int_\Omega \left| \left(f_z'(x, u(x)+th(x)) - f_z'(x, u(x)) \right) h(x) \right|^q \, dx \, dt \right)^{\frac{1}{q}} \\
&\qquad = \int_0^1 \|(G(u+th) - G(u))h\|_{L^q} \, dt \\
&\qquad \leq \int_0^1 \|G(u+th) - G(u)\|_{L^r} \, dt \; \|h\|_{L^p}.
\end{aligned}$$

Due to the continuity of $G : L^p(\Omega) \to L^r(\Omega)$, the integrand tends to zero uniformly in $[0,1]$ for $\|h\|_{L^p} \to 0$, and hence F is by definition Fréchet differentiable with derivative $F'(u) = DF(u)$ (whose continuity we have already shown). □

In fact, this result is sharp: except for the case $p = q = \infty$, no superposition operator is differentiable from $L^p(\Omega)$ to $L^p(\Omega)$ (unless it is affine-linear); see, e.g., [Appell and Zabrejko, 1990, Theorem 3.12].

2.4 VARIATIONAL PRINCIPLES

As the example $f(t) = 1/t$ on $\{t \in \mathbb{R} : t \geq 1\}$ shows, the coercivity requirement in Theorem 2.1 is necessary to obtain minimizers even if the functional is bounded from below. However, sometimes one does not need an exact minimizer and is satisfied with "almost minimizers". *Variational principles* state that such almost minimizers can be obtained as minimizers of a perturbed functional and even give a precise relation between the size of the perturbation needed in terms of the desired distance from the infimum.

The most well-known variational principle is *Ekeland's variational principle*, which holds in general complete metric spaces but which we here state in Banach spaces for the sake of notation. In the statement of the following theorem, note that we do not assume the functional to be *weakly* lower semicontinuous.

Theorem 2.16 (Ekeland's variational principle). *Let X be a Banach space and $F : X \to \overline{\mathbb{R}}$ be proper, lower semicontinuous, and bounded from below. Let $\varepsilon > 0$ and $z_\varepsilon \in X$ be such that*

$$F(z_\varepsilon) < \inf_{x \in X} F(x) + \varepsilon.$$

Then for any $\lambda > 0$, there exists an $x_\lambda \in X$ with

(i) $\|x_\lambda - z_\varepsilon\|_X \leq \lambda$,

(ii) $F(x_\lambda) + \frac{\varepsilon}{\lambda}\|x_\lambda - z_\varepsilon\|_X \leq F(z_\varepsilon)$,

(iii) $F(x_\lambda) < F(x) + \frac{\varepsilon}{\lambda}\|x - x_\lambda\|_X$ *for all* $x \in X \setminus \{x_\lambda\}$.

Proof. The proof proceeds similarly to that of Theorem 2.1: We construct an "almost minimizing" sequence, show that it converges, and verify that the limit has the desired properties. Here we proceed inductively. First, set $x_0 \coloneqq z_\varepsilon$. For given x_n, define now

$$S_n \coloneqq \left\{ x \in X \;\middle|\; F(x) + \frac{\varepsilon}{\lambda}\|x - x_n\|_X \leq F(x_n) \right\}.$$

Since $x_n \in S_n$, this set is nonempty. We can thus choose $x_{n+1} \in S_n$ such that

$$F(x_{n+1}) \leq \frac{1}{2}F(x_n) + \frac{1}{2}\inf_{x \in S_n} F(x), \tag{2.7}$$

which is possible because either the right-hand side equals $F(x_n)$ (in which case we choose $x_{n+1} = x_n$) or is strictly greater, in which case there must exist such an x_{n+1} by the properties of the infimum. By construction, the sequence $\{F(x_n)\}_{n\in\mathbb{N}}$ is thus decreasing as well as bounded from below and therefore convergent. Using the triangle inequality, the fact that $x_{n+1} \in S_n$, and the telescoping sum, we also obtain that for any $m \geq n \in \mathbb{N}$,

$$\frac{\varepsilon}{\lambda}\|x_n - x_m\|_X \leq \sum_{j=n}^{m-1} \frac{\varepsilon}{\lambda}\|x_j - x_{j+1}\|_X \leq F(x_n) - F(x_m). \tag{2.8}$$

Hence, $\{x_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence since $\{F(x_n)\}_{n\in\mathbb{N}}$ is one and hence converges to some $x_\lambda \in X$ since X is complete.

We now show that this limit has the claimed properties. We begin with (ii), for which we use the fact that both F and the norm in X are lower semicontinuous and hence obtain from (2.8) by taking $m \to \infty$ that

$$\frac{\varepsilon}{\lambda}\|x_n - x_\lambda\|_X + F(x_\lambda) \leq \limsup_{m\to\infty} \frac{\varepsilon}{\lambda}\|x_n - x_m\|_X + F(x_m) \leq F(x_n) \quad \text{for any } n \geq 0. \tag{2.9}$$

Choosing in particular $n = 0$ such that $x_0 = z_\varepsilon$ yields (ii).

Furthermore, by definition of z_ε, this implies that

$$\frac{\varepsilon}{\lambda}\|z_\varepsilon - x_\lambda\|_X \leq F(z_\varepsilon) - F(x_\lambda) \leq F(z_\varepsilon) - \inf_{x\in X} F(x) < \varepsilon$$

and hence (i).

Assume now that (iii) does not hold, i.e., that there exists an $x \in X \setminus \{x_\lambda\}$ such that

$$F(x) \leq F(x_\lambda) - \frac{\varepsilon}{\lambda}\|x - x_\lambda\|_X < F(x_\lambda). \tag{2.10}$$

Estimating $F(x_\lambda)$ using (2.9) and then using the productive zero together with the triangle inequality, we obtain from the first inequality that for all $n \in \mathbb{N}$,

$$F(x) \leq F(x_n) - \frac{\varepsilon}{\lambda}\|x_n - x_\lambda\|_X - \frac{\varepsilon}{\lambda}\|x - x_\lambda\|_X \leq F(x_n) - \frac{\varepsilon}{\lambda}\|x_n - x\|_X.$$

Hence, $x \in S_n$ for all $n \in \mathbb{N}$. From (2.7), we then deduce that

$$2F(x_{n+1}) - F(x_n) \leq F(x) \quad \text{for all } n \in \mathbb{N}.$$

The convergence of $\{F(x_n)\}_{n\in\mathbb{N}}$ together with (2.10) and the lower semicontinuity of F thus yields the contradiction

$$\lim_{n\to\infty} F(x_n) \leq F(x) < F(x_\lambda) \leq \lim_{n\to\infty} F(x_n). \qquad \square$$

Ekeland's variational principle has the disadvantage that even for differentiable F, the perturbed function that is minimized by x_λ is inherently nonsmooth. This is different for *smooth variational principles* such as the following one due to Borwein and Preiss [Borwein and Preiss, 1987].

Theorem 2.17 (Borwein–Preiss variational principle). *Let X be a Banach space and $F : X \to \overline{\mathbb{R}}$ be proper, lower semicontinuous, and bounded from below. Let $\varepsilon > 0$ and $z_\varepsilon \in X$ be such that*

$$F(z_\varepsilon) < \inf_{x\in X} F(x) + \varepsilon.$$

Then for any $\lambda > 0$ and $p \geq 1$, there exists

- *a sequence $\{x_n\}_{n\in\mathbb{N}_0} \subset X$ with $x_0 = z_\varepsilon$ converging strongly to some $x_\lambda \in X$ and*
- *a sequence $\{\mu_n\}_{n\in\mathbb{N}_0} \subset (0,\infty)$ with $\sum_{n=0}^{\infty} \mu_n = 1$*

such that

(i) $\|x_\lambda - x_n\|_X \leq \lambda$ *for all* $n \in \mathbb{N} \cup \{0\}$,

(ii) $F(x_\lambda) + \frac{\varepsilon}{\lambda^p} \sum_{n=0}^{\infty} \mu_n \|x_\lambda - x_n\|_X^p \leq F(z_\varepsilon)$,

(iii) $F(x_\lambda) + \frac{\varepsilon}{\lambda^p} \sum_{n=0}^{\infty} \mu_n \|x_\lambda - x_n\|_X^p \leq F(x) + \frac{\varepsilon}{\lambda^p} \sum_{n=0}^{\infty} \mu_n \|x - x_n\|_X^p$ *for all* $x \in X$.

Proof. We proceed similarly to the proof of Theorem 2.16 by induction. First, we chose constants $\gamma, \eta, \mu, \theta > 0$ such that

- $F(z_\varepsilon) - \inf_{x\in X} F(x) < \eta < \gamma < \varepsilon$,
- $\mu < 1 - \frac{\gamma}{\varepsilon}$,
- $\theta < \mu\left(1 - \left(\frac{\eta}{\gamma}\right)^{1/p}\right)^p$.

Let now $x_0 := z_\varepsilon$ and $F_0 := F$ and set $\delta := (1-\mu)\frac{\varepsilon}{\lambda^p} > 0$. We then define

$$F_1(x) := F_0(x) + \delta\mu\|x - x_0\|_X^p \qquad \text{for all } x \in X.$$

By construction, we then have

$$\inf_{x\in X} F_1(x) \le F_1(x_0) = F_0(x_0),$$

and thus we can find, by the same argument as for (2.7), an $x_1 \in X$ with

$$F_1(x_1) \le \theta F_0(x_0) + (1-\theta)\inf_{x\in X} F_1(x).$$

Continuing in this manner, we obtain sequences $\{x_n\}_{n\in\mathbb{N}}$ and $\{F_n\}_{n\in\mathbb{N}}$ with

$$F_{n+1}(x) = F_n(x) + \delta\mu^n\|x - x_n\|_X^p \tag{2.11}$$

and

$$F_{n+1}(x_{n+1}) \le \theta F_n(x_n) + (1-\theta)\inf_{x\in X} F(x). \tag{2.12}$$

Set now $s_n := \inf_{x\in X} F_n(x)$ and $a_n := F_n(x_n)$. Then (2.11) implies that $\{s_n\}_{n\ge 0}$ is monotonically increasing, while (2.12) implies that $\{a_n\}_{n\ge 0}$ is monotonically decreasing. We thus have

$$s_n \le s_{n+1} \le a_{n+1} \le \theta a_n + (1-\theta)s_{n+1} \le a_n, \tag{2.13}$$

which can be rearranged to show for all $n \ge 0$ that

$$a_{n+1} - s_{n+1} \le \theta a_n + (1-\theta)s_{n+1} - s_{n+1} = \theta(a_n - s_{n+1}) \le \theta(a_n - s_n) \le \theta^n(a_0 - s_0). \tag{2.14}$$

This together with the monotonicity of the two sequences and the boundedness of F from below shows that $\lim_{n\to\infty} a_n = \lim_{n\to\infty} s_n \in \mathbb{R}$. We now use (2.11) in (2.13) to obtain that

$$a_n \ge a_{n+1} = F_n(x_n) + \delta\mu^n\|x_{n+1} - x_n\|_X^p \ge s_n + \delta\mu^n\|x_{n+1} - x_n\|_X^p,$$

which together with (2.14) and the choice of η yields

$$\delta\mu^n\|x_{n+1} - x_n\|_X^p \le a_n - s_n \le \theta^n(a_0 - s_0) < \eta\theta^n.$$

The choice of θ and μ now ensure that $0 < \frac{\theta}{\mu} < 1$, which implies that

$$\begin{aligned}\|x_m - x_n\|_X &\le \sum_{k=n}^{m-n-1} \|x_{k+1} - x_k\|_X \le \left(\frac{\eta}{\delta}\right)^{1/p} \sum_{k=n}^{m-n-1} \left(\frac{\theta}{\mu}\right)^{k/p} \\ &\le \left(\frac{\eta}{\delta}\right)^{1/p} \left(\frac{\theta}{\mu}\right)^{n/p} \left(1 - \left(\frac{\theta}{\mu}\right)^{1/p}\right)^{-1} \qquad \text{for all } m, n \ge 0\end{aligned} \tag{2.15}$$

using the partial geometric series

$$\sum_{k=n}^{m-n-1} \alpha^k = \sum_{k=0}^{m-n-1} \alpha^k - \sum_{k=0}^{n-1} \alpha^k = \frac{1-\alpha^{m-n}}{1-\alpha} - \frac{1-\alpha^n}{1-\alpha} < \frac{\alpha^n}{1-\alpha}$$

valid for any $\alpha \in (0,1)$. Hence $\{x_n\}_n \in \mathbb{N}$ is a Cauchy sequence which therefore converges to some $x_\lambda \in X$. Setting $\mu_n := \mu^n(1-\mu) > 0$, we also have $\sum_{n=0}^{\infty} \mu_n = 1$ by the choice of $\mu < 1$. Furthermore, the definition of μ_n and δ implies for all $x \in X$ that

$$F(x) + \frac{\varepsilon}{\lambda^p} \sum_{k=0}^{\infty} \mu_k \|x - x_k\|_X^p = \lim_{n\to\infty} F(x) + \sum_{k=0}^{n} \delta\mu^k \|x - x_k\|_X^p = \lim_{n\to\infty} F_n(x). \tag{2.16}$$

It remains to verify the claims on x_λ. First, (2.15) together with the choice of θ and δ implies for all $n, m \geq 0$ that

$$\|x_m - x_n\|_X \leq \left(\frac{\eta}{\delta}\right)^{1/p} \left(\frac{\eta}{\gamma}\right)^{-1/p} = \left(\frac{\gamma}{\delta}\right)^{1/p} < \left(\frac{\varepsilon}{\delta}\right)^{1/p} (1-\mu)^{1/p} = \lambda.$$

Letting $m \to \infty$ for fixed $n \in \mathbb{N} \cup \{0\}$ now shows (i).

Second, by (2.11) and the definition of δ, we have

$$F(x_n) + \frac{\varepsilon}{\lambda^p} \sum_{k=0}^{\infty} \mu_k \|x_n - x_k\|_X^p = F_n(x_n) + \frac{\varepsilon}{\lambda^p} \sum_{k=n+1}^{\infty} \mu_k \|x_n - x_k\|_X^p \leq a_n + \varepsilon \sum_{k=n+1}^{\infty} \mu_k,$$

where the inequality follows from (i). The lower semicontinuity of F and of the norm thus yield

$$F(x_\lambda) + \frac{\varepsilon}{\lambda^p} \sum_{k=0}^{\infty} \mu_k \|x_\lambda - x_k\|_X^p \leq \lim_{n\to\infty} a_n \leq a_0 = F(z_\varepsilon) \tag{2.17}$$

since $\{a_n\}_{n\geq 0}$ is monotonically decreasing. This shows (ii).

Finally, (2.16) and the definition of s_n imply for all $x \in X$ that

$$F(x) + \frac{\varepsilon}{\lambda^p} \sum_{k=0}^{\infty} \mu_k \|x - x_k\|_X^p = \lim_{n\to\infty} F_n(x) \geq \lim_{n\to\infty} s_n = \lim_{n\to\infty} a_n,$$

which together with (2.17) yields (iii). □

The Borwein–Preiss variational principle therefore guarantees a smooth perturbation if, e.g., X is a Hilbert space and $p = 2$. Further smooth variational principles that allow for more general smooth perturbations such as the *Deville–Godefroy–Zizler variational principle* can be found in, e.g., [Borwein and Zhu, 2005; Schirotzek, 2007].

Part II

CONVEX ANALYSIS

3 CONVEX FUNCTIONS

Now that we know from the direct method of the calculus of variations when a functional $F : X \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ admits a minimizer $\bar{x} \in X$, our next goal is to characterize such minimizers using optimality conditions, i.e., without comparing its function value to that at every other point. If F is differentiable at $\bar{X}$, the classical optimality condition is by Fermat's principle, $F'(\bar{x}) = 0$, and we can use calculus rules to evaluate this derivative in order to make this condition as explicit as possible. We wish to extend this as far as possible to nonsmooth F, i.e., not classically (Fréchet or Gâteaux) differentiable, is nonsmooth. Clearly, *not* being differentiable is not much to work with, so we have to assume other properties instead. One possibility is to replace the *local* property of differentiability with the *global* property of convexity. As we will see in this and the following chapters, this property will allow us to recover a satisfying calculus for a class of relevant nonsmooth functionals. We begin in this chapter with deriving several fundamental properties of convex functions relevant for optimization, while the corresponding Fermat principle and calculus rules are the topic of the next Chapter 4.

Throughout this and the following chapters, X will be a normed vector space unless noted otherwise.

3.1 DEFINITION AND BASIC PROPERTIES

A functional $F : X \to \overline{\mathbb{R}}$ is called *convex* if for all $x, y \in X$ and $\lambda \in [0, 1]$, it holds that

$$F(\lambda x + (1 - \lambda) y) \leq \lambda F(x) + (1 - \lambda) F(y) \tag{3.1}$$

(where the function value ∞ is allowed on both sides). If for all $x, y \in \operatorname{dom} F$ with $x \neq y$ and all $\lambda \in (0, 1)$ we even have

$$F(\lambda x + (1 - \lambda) y) < \lambda F(x) + (1 - \lambda) F(y),$$

we call F *strictly convex*.

As illustrated in Figure 3.1, an alternative characterization of the convexity of a functional $F : X \to \overline{\mathbb{R}}$ is based on its *epigraph*

$$\operatorname{epi} F := \{(x, t) \in X \times \mathbb{R} \mid F(x) \leq t\}.$$

(a) a convex set C

(b) the (convex) epigraph of a convex function

Figure 3.1: Illustration of a convex set and of the characterization of a convex function in terms of the convexity of its epigraph: all line segments between two points of corresponding set are completely contained in that set.

Lemma 3.1. *Let $F : X \to \overline{\mathbb{R}}$. Then* epi F *is*

(i) nonempty if and only if F is proper;

(ii) convex if and only if F is convex;

(iii) (weakly) closed if and only if F is (weakly) lower semicontinuous.[1]

Proof. Statement (i) follows directly from the definition: F is proper if and only if there exists an $x \in X$ and a $t \in \mathbb{R}$ with $F(x) \leq t < \infty$, i.e., $(x, t) \in \operatorname{epi} F$.

For (ii), let F be convex and $(x, r), (y, s) \in \operatorname{epi} F$ be given. For any $\lambda \in [0, 1]$, the definition (3.1) then implies that

$$F(\lambda x + (1-\lambda)y) \leq \lambda F(x) + (1-\lambda)F(y) \leq \lambda r + (1-\lambda)s,$$

i.e., that

$$\lambda(x, r) + (1-\lambda)(y, s) = (\lambda x + (1-\lambda)y, \lambda r + (1-\lambda)s) \in \operatorname{epi} F,$$

and hence epi F is convex. Let conversely epi F be convex and $x, y \in X$ be arbitrary, where we can assume that $F(x) < \infty$ and $F(y) < \infty$ (otherwise (3.1) is trivially satisfied). We clearly have $(x, F(x)), (y, F(y)) \in \operatorname{epi} F$. The convexity of epi F then implies for all $\lambda \in [0, 1]$ that

$$(\lambda x + (1-\lambda)y, \lambda F(x) + (1-\lambda)F(y)) = \lambda(x, F(x)) + (1-\lambda)(y, F(y)) \in \operatorname{epi} F,$$

and hence by definition of epi F that (3.1) holds.

Finally, we show (iii): Let first F be lower semicontinuous, and let $\{(x_n, t_n)\}_{n\in\mathbb{N}} \subset \operatorname{epi} F$ be an arbitrary sequence with $(x_n, t_n) \to (x, t) \in X \times \mathbb{R}$. Then we have that

$$F(x) \leq \liminf_{n\to\infty} F(x_n) \leq \limsup_{n\to\infty} t_n = t,$$

[1] For that reason, some authors use the term *closed* also to refer to lower semicontinuous functionals. We will stick with the latter, much less ambiguous, term throughout the following.

i.e., $(x,t) \in \operatorname{epi} F$. Let conversely $\operatorname{epi} F$ be closed and assume that F is proper (otherwise the claim holds trivially) and not lower semicontinuous. Then there exists a sequence $\{x_n\}_{n\in\mathbb{N}} \subset X$ with $x_n \to x \in X$ and

$$F(x) > \liminf_{n\to\infty} F(x_n) =: M \in [-\infty, \infty).$$

We now distinguish two cases.

a) $x \in \operatorname{dom} F$: In this case, we can select a subsequence, again denoted by $\{x_n\}_{n\in\mathbb{N}}$, such that there exists an $\varepsilon > 0$ with $F(x_n) \le F(x) - \varepsilon$ and thus $(x_n, F(x)-\varepsilon) \in \operatorname{epi} F$ for all $n \in \mathbb{N}$. From $x_n \to x$ and the closedness of $\operatorname{epi} F$, we deduce that $(x, F(x)-\varepsilon) \in \operatorname{epi} F$ and hence $F(x) \le F(x) - \varepsilon$, contradicting $\varepsilon > 0$.

b) $x \notin \operatorname{dom} F$: In this case, we can argue similarly using $F(x_n) \le M + \varepsilon$ for $M > -\infty$ or $F(x_n) \le \varepsilon$ for $M = -\infty$ to obtain a contradiction with $F(x) = \infty$.

The equivalence of weak lower semicontinuity and weak closedness follows in exactly the same way. □

Note that $(x,t) \in \operatorname{epi} F$ implies that $x \in \operatorname{dom} F$; hence the effective domain of a proper, convex, and lower semicontinuous functional is always nonempty, convex, and closed as well. Also, together with Lemma 1.10 we immediately obtain

Corollary 3.2. *Let $F : X \to \overline{\mathbb{R}}$ be convex. Then F is weakly lower semicontinuous if and only F is lower semicontinuous.*

Also useful for the study of a functional $F : X \to \overline{\mathbb{R}}$ are the corresponding *sublevel sets*

$$\operatorname{sub}_t F := \{x \in X \mid F(x) \le t\}, \qquad t \in \mathbb{R},$$

for which one shows as in Lemma 3.1 the following properties.

Lemma 3.3. *Let $F : X \to \overline{\mathbb{R}}$.*

(i) If F is convex, $\operatorname{sub}_t F$ is convex for all $t \in \mathbb{R}$ (but the converse does not hold).

(ii) F is (weakly) lower semicontinuous if and only if $\operatorname{sub}_t F$ is (weakly) closed for all $t \in \mathbb{R}$.

Directly from the definition we obtain the convexity of

(i) *continuous affine functionals* of the form $x \mapsto \langle x^*, x\rangle_X - \alpha$ for fixed $x^* \in X^*$ and $\alpha \in \mathbb{R}$;

(ii) the norm $\|\cdot\|_X$ in a normed vector space X;

(iii) the indicator function δ_C for a convex set C.

If X is a Hilbert space, $F(x) = \|x\|_X^2$ is even strictly convex: For $x, y \in X$ with $x \neq y$ and any $\lambda \in (0,1)$,

$$\begin{aligned}
\|\lambda x + (1-\lambda)y\|_X^2 &= (\lambda x + (1-\lambda)y \mid \lambda x + (1-\lambda)y)_X \\
&= \lambda^2 (x \mid x)_X + 2\lambda(1-\lambda)(x \mid y)_X + (1-\lambda)^2 (y \mid y)_X \\
&= \lambda\Big(\lambda(x \mid x)_X - (1-\lambda)(x-y \mid x)_X + (1-\lambda)(y \mid y)_X\Big) \\
&\quad + (1-\lambda)\Big(\lambda(x \mid x)_X + \lambda(x-y \mid y)_X + (1-\lambda)(y \mid y)_X\Big) \\
&= (\lambda + (1-\lambda))\Big(\lambda(x \mid x)_X + (1-\lambda)(y \mid y)_X\Big) - \lambda(1-\lambda)(x-y \mid x-y)_X \\
&= \lambda\|x\|_X^2 + (1-\lambda)\|y\|_X^2 - \lambda(1-\lambda)\|x-y\|_X^2 \\
&< \lambda\|x\|_X^2 + (1-\lambda)\|y\|_X^2.
\end{aligned}$$

Further examples can be constructed as in Lemma 2.3 through the following operations.

Lemma 3.4. *Let X and Y be normed vector spaces and let $F : X \to \overline{\mathbb{R}}$ be convex. Then the following functionals are convex as well:*

(i) αF *for all* $\alpha \geq 0$*;*

(ii) $F + G$ *for* $G : X \to \overline{\mathbb{R}}$ *convex (strictly if* F or G *is strictly convex);*

(iii) $\varphi \circ F$ *for* $\varphi : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ *convex and increasing;*

(iv) $F \circ K$ *for* $K : Y \to X$ *linear;*

(v) $x \mapsto \sup_{i \in I} F_i(x)$ *with* $F_i : X \to \overline{\mathbb{R}}$ *convex for an arbitrary set* I*.*

Lemma 3.4 (v) in particular implies that the pointwise supremum of continuous affine functionals is always convex. In fact, any convex functional can be written in this way. To show this, we define for a proper functional $F : X \to \overline{\mathbb{R}}$ the *convex envelope*

$$F^\Gamma(x) := \sup\{a(x) \mid a \text{ continuous affine with } a(\tilde{x}) \leq F(\tilde{x}) \text{ for all } \tilde{x} \in X\}.$$

Note that $F^\Gamma : X \to [-\infty, \infty]$ without further assumptions of F.

Lemma 3.5. *Let $F : X \to \overline{\mathbb{R}}$ be proper. Then F is convex and lower semicontinuous if and only if $F = F^\Gamma$.*

Proof. Since affine functionals are convex, Lemma 3.4 (v) and Lemma 2.3 (v) imply that $F = F^\Gamma$ is always convex and lower semicontinuous.

Let now $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. It is clear from the definition of F^Γ as a pointwise supremum that $F^\Gamma \leq F$ always holds. Assume therefore that $F^\Gamma < F$. Then there exists an $x_0 \in X$ and a $\lambda \in \mathbb{R}$ with

$$F^\Gamma(x_0) < \lambda < F(x_0).$$

We now use the Hahn–Banach separation theorem to construct a continuous affine functional a with $a \leq F$ but $a(x_0) > \lambda > F^\Gamma(x_0)$, which would contradict the definition of F^Γ. Since F is proper, convex, and lower semicontinuous, epi F is nonempty, convex, and closed by Lemma 3.1. Furthermore, $\{(x_0, \lambda)\}$ is compact and, as $\lambda < F(x_0)$, disjoint with epi F. Theorem 1.5 (ii) hence yields a $z^* \in (X \times \mathbb{R})^*$ and an $\alpha \in \mathbb{R}$ with

$$\langle z^*, (x,t)\rangle_{X\times\mathbb{R}} \leq \alpha < \langle z^*, (x_0,\lambda)\rangle_{X\times\mathbb{R}} \qquad \text{for all } (x,t) \in \operatorname{epi} F.$$

We now define an $x^* \in X^*$ via $\langle x^*, x\rangle_X = \langle z^*, (x,0)\rangle_{X\times\mathbb{R}}$ for all $x \in X$ and set $s := \langle z^*, (0,1)\rangle_{X\times\mathbb{R}} \in \mathbb{R}$. Then $\langle z^*, (x,t)\rangle_{X\times\mathbb{R}} = \langle x^*, x\rangle_X + st$ and hence

$$\langle x^*, x\rangle_X + st \leq \alpha < \langle x^*, x_0\rangle_X + s\lambda \qquad \text{for all } (x,t) \in \operatorname{epi} F. \tag{3.2}$$

Now for $(x,t) \in \operatorname{epi} F$ we also have $(x,t') \in \operatorname{epi} F$ for all $t' > t$, and the first inequality in (3.2) implies that for all sufficiently large $t' > 0$,

$$s \leq \frac{\alpha - \langle x^*, x\rangle_X}{t'} \to 0 \qquad \text{for } t' \to \infty.$$

Hence $s \leq 0$. We continue with a case distinction.

(i) $s < 0$: We set

$$a : X \to \mathbb{R}, \qquad x \mapsto \frac{\alpha - \langle x^*, x\rangle_X}{s},$$

which is continuous affine. Furthermore, using the "productive zero" (i.e., adding and subtracting the same term) in the first inequality in (3.2) for $(x, F(x)) \in \operatorname{epi} F$ implies (noting $s < 0$!) that

$$a(x) = \tfrac{1}{s}\left(\alpha - \langle x^*, x\rangle_X - sF(x)\right) + F(x) \leq F(x).$$

(For $x \notin \operatorname{dom} F$ this holds trivially.) But the second inequality in (3.2) implies that

$$a(x_0) = \tfrac{1}{s}\left(\alpha - \langle x^*, x_0\rangle_X\right) > \lambda.$$

(ii) $s = 0$: Then $\langle x^*, x\rangle_X \leq \alpha < \langle x^*, x_0\rangle_X$ for all $x \in \operatorname{dom} F$, which can only hold for $x_0 \notin \operatorname{dom} F$. But F is proper, and hence we can find a $y_0 \in \operatorname{dom} F$, for which we can construct as in case (i) by separating epi F and (y_0, μ) for sufficiently small μ a continuous affine functional $a_0 : X \to \mathbb{R}$ with $a_0 \leq F$ pointwise. For $\rho > 0$ we now set

$$a_\rho : X \to \mathbb{R}, \qquad x \mapsto a_0(x) + \rho\left(\langle x^*, x\rangle_X - \alpha\right),$$

which is continuous affine as well. Since $\langle x^*, x\rangle_X \leq \alpha$, we also have that $a_\rho(x) \leq a_0(x) \leq F(x)$ for all $x \in \operatorname{dom} F$ and any $\rho > 0$. But due to $\langle x^*, x_0\rangle_X > \alpha$, we can choose $\rho > 0$ with $a_\rho(x_0) > \lambda$.

In both cases, the definition of F^Γ as a supremum implies that $F^\Gamma(x_0) > \lambda$ as well, contradicting the assumption $F^\Gamma(x_0) < \lambda$. □

Remark 3.6. Using the weak-∗ Hahn–Banach Theorem 1.13 in place of Theorem 1.5, the same proof shows that a proper functional $F : X^* \to \overline{\mathbb{R}}$ is convex and weakly-∗ lower semicontinuous if and only if $F = F_\Gamma$ for

$$F_\Gamma(x^*) := \sup\{\langle x^*, x\rangle_X + \alpha \mid x \in X, \alpha \in \mathbb{R}, \langle \tilde{x}^*, x\rangle_X + \alpha \le F(\tilde{x}^*) \text{ for all } \tilde{x}^* \in X^*\}.$$

(Note that a convex and weakly-∗ lower semicontinuous functional need not be lower semicontinuous, since convex and closed sets need not be weakly-∗ closed.)

A particularly useful class of convex functionals in the calculus of variations arises from integral functionals with convex integrands defined through superposition operators.

Lemma 3.7. *Let $f : \mathbb{R} \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. If $\Omega \subset \mathbb{R}^d$ is bounded and $1 \le p \le \infty$, this also holds for*

$$F : L^p(\Omega) \to \overline{\mathbb{R}}, \qquad u \mapsto \begin{cases} \int_\Omega f(u(x))\,dx & \text{if } f \circ u \in L^1(\Omega), \\ \infty & \text{else.} \end{cases}$$

Proof. First, Lemma 3.5 implies that there exist $a, \alpha \in \mathbb{R}$ such that

$$f(t) \ge at - \alpha \qquad \text{for all } t \in \mathbb{R}. \tag{3.3}$$

Since Ω is bounded and hence $L^p(\Omega) \subset L^1(\Omega)$ for any $p \ge 1$, this implies that

$$F(u) \ge \int_\Omega au(x) - \alpha\,dx \in \mathbb{R} \qquad \text{for any } u \in L^p(\Omega).$$

In particular, $F(u) > -\infty$ for all $u \in L^p(\Omega)$. Since f is proper, there is a $t_0 \in \operatorname{dom} f$. Hence (using again that Ω is bounded) the constant function $u_0 \equiv t_0 \in \operatorname{dom} F$ satisfies $F(u_0) < \infty$. This shows that F is proper.

To show convexity, we take $u, v \in \operatorname{dom} F$ (since otherwise (3.1) is trivially satisfied) and $\lambda \in [0,1]$ arbitrary. The convexity of f now implies that

$$f(\lambda u(x) + (1-\lambda)v(x)) \le \lambda f(u(x)) + (1-\lambda) f(v(x)) \quad \text{for almost every } x \in \Omega.$$

Since $u, v \in \operatorname{dom} F$ and $L^1(\Omega)$ is a vector space, $\lambda f(u(x)) + (1-\lambda) f(v(x)) \in L^1(\Omega)$ as well. Similarly, the left-hand side is bounded from below by $a(\lambda u(x) + (1-\lambda)v(x)) - \alpha \in L^1(\Omega)$ by (3.3). We can thus integrate the inequality over Ω to obtain the convexity of F.

To show lower semicontinuity, we use Lemma 3.1. Let $\{(u_n, t_n)\}_{n\in\mathbb{N}} \subset \operatorname{epi} F$ with $u_n \to u$ in $L^p(\Omega)$ and $t_n \to t$ in $\mathbb{R}$. Then there exists a subsequence $\{u_{n_k}\}_{k\in\mathbb{N}}$ with $u_{n_k}(x) \to u(x)$

almost everywhere. Hence, the lower semicontinuity of f together with Fatou's Lemma implies that

$$\begin{aligned}\int_\Omega f(u(x)) - (au(x) - \alpha)\,dx &\le \int_\Omega \liminf_{k\to\infty} (f(u_{n_k}(x)) - (au_{n_k}(x) - \alpha))\,dx \\ &\le \liminf_{k\to\infty} \int_\Omega f(u_{n_k}(x)) - (au_{n_k}(x) - \alpha)\,dx \\ &= \liminf_{k\to\infty} \int_\Omega f(u_{n_k}(x))\,dx - \int_\Omega au(x) - \alpha\,dx\end{aligned}$$

as the integrands are nonnegative due to (3.3). Since $(u_{n_k}, t_{n_k}) \in \operatorname{epi} F$, this yields

$$F(u) = \int_\Omega f(u(x))\,dx \le \liminf_{k\to\infty} \int_\Omega f(u_{n_k}(x))\,dx = \liminf_{k\to\infty} F(u_{n_k}) \le \lim_{k\to\infty} t_{n_k} = t,$$

i.e., $(u, t) \in \operatorname{epi} F$. Hence $\operatorname{epi} F$ is closed, and the lower semicontinuity of F follows from Lemma 3.1 (iii). □

3.2 EXISTENCE OF MINIMIZERS

After all this preparation, we can quickly prove the main result on existence of solutions to convex minimization problems.

Theorem 3.8. *Let X be a reflexive Banach space and let*

(i) $U \subset X$ be nonempty, convex, and closed;

(ii) $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous with $\operatorname{dom} F \cap U \neq \emptyset$;

(iii) U be bounded or F be coercive.

Then the problem

$$\min_{x\in U} F(x)$$

admits a solution $\bar{x} \in U \cap \operatorname{dom} F$. If F is strictly convex, the solution is unique.

Proof. We consider the extended functional $\overline{F} = F + \delta_U : X \to \overline{\mathbb{R}}$. Assumption (i) together with Lemma 2.5 implies that δ_U is proper, convex, and weakly lower semicontinuous. From (i) we obtain an $x_0 \in U$ with $\overline{F}(x_0) < \infty$, and hence $\overline{F}$ is proper, convex, and (by Corollary 3.2) weakly lower semicontinuous. Finally, due to (iii), $\overline{F}$ is coercive since for bounded U, we can use that $F > -\infty$, and for coercive F, we can use that $\delta_U \ge 0$. Hence we can apply Theorem 2.1 to obtain the existence of a minimizer $\bar{x} \in \operatorname{dom} \overline{F} = U \cap \operatorname{dom} F$ of $\overline{F}$ with

$$F(\bar{x}) = \overline{F}(\bar{x}) \le \overline{F}(x) = F(x) \qquad \text{for all } x \in U,$$

i.e., $\bar{x}$ is the claimed solution.

Let now F be strictly convex, and let $\bar{x}$ and $\bar{x}' \in U$ be two different minimizers, i.e., $F(\bar{x}) = F(\bar{x}') = \min_{x \in U} F(x)$ and $\bar{x} \neq \bar{x}'$. Then by the convexity of U we have for all $\lambda \in (0,1)$ that

$$x_\lambda := \lambda \bar{x} + (1-\lambda)\bar{x}' \in U,$$

while the strict convexity of F implies that

$$F(x_\lambda) < \lambda F(\bar{x}) + (1-\lambda) F(\bar{x}') = F(\bar{x}).$$

But this is a contradiction to $F(\bar{x}) \leq F(x)$ for all $x \in U$. □

Note that for a sum of two convex functionals to be coercive, it is in general not sufficient that only one of them is. Functionals for which this is the case – such as the indicator function of a bounded set – are called *supercoercive*; another example which will be helpful later is the squared norm.

Lemma 3.9. *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $x_0 \in X$ be given. Then the functional*

$$J : X \to \overline{\mathbb{R}}, \qquad x \mapsto F(x) + \frac{1}{2}\|x - x_0\|_X^2$$

is coercive.

Proof. Since F is proper, convex, and lower semicontinuous, it follows from Lemma 3.5 that F is bounded from below by a continuous affine functional, i.e., there exists an $x^* \in X^*$ and an $\alpha \in \mathbb{R}$ with $F(x) \geq \langle x^*, x \rangle_X - \alpha$ for all $x \in X$. Together with the reverse triangle inequality and (1.1), we obtain that

$$\begin{aligned} J(x) &\geq \langle x^*, x \rangle_X - \alpha + \tfrac{1}{2}\left(\|x\|_X - \|x_0\|_X\right)^2 \\ &\geq -\|x^*\|_{X^*}\|x\|_X - \alpha + \tfrac{1}{2}\|x\|_X^2 - \|x\|_X\|x_0\|_X \\ &= \|x\|_X \left(\tfrac{1}{2}\|x\|_X - \|x^*\|_{X^*} - \|x_0\|_X\right) - \alpha. \end{aligned}$$

Since x^* and x_0 are fixed, the term in parentheses is positive for $\|x\|_X$ sufficiently large, and hence $J(x) \to \infty$ for $\|x\|_X \to \infty$ as claimed. □

3.3 CONTINUITY PROPERTIES

To close this chapter, we show the following remarkable result: *Any (locally) bounded convex functional is (locally) continuous.* Besides being of use in later chapters, this result illustrates the beauty of convex analysis: an algebraic but global property (convexity) connects two topological but local properties (neighborhood and continuity). Here we consider of course the strong topology in a normed vector space.

Lemma 3.10. *Let X be a normed vector space, $F : X \to \overline{\mathbb{R}}$ be convex, and $x \in X$. If there is a $\rho > 0$ such that F is bounded from above on $\mathbb{O}(x, \rho)$, then F is locally Lipschitz continuous near x.*

Proof. By assumption, there exists an $M \in \mathbb{R}$ with $F(y) \leq M$ for all $y \in \mathbb{O}(x, \rho)$. We first show that F is locally bounded from below as well. Let $y \in \mathbb{O}(x, \rho)$ be arbitrary. Since $\|x - y\|_X < \rho$, we also have that $z := 2x - y = x - (y - x) \in \mathbb{O}(x, \rho)$, and the convexity of F implies that $F(x) = F\left(\frac{1}{2}y + \frac{1}{2}z\right) \leq \frac{1}{2}F(y) + \frac{1}{2}F(z)$ and hence that

$$-F(y) \leq F(z) - 2F(x) \leq M - 2F(x) =: m,$$

i.e., $-m \leq F(y) \leq M$ for all $y \in \mathbb{O}(x, \rho)$.

We now show that this implies Lipschitz continuity on $\mathbb{O}(x, \frac{\rho}{2})$. Let $y_1, y_2 \in \mathbb{O}(x, \frac{\rho}{2})$ with $y_1 \neq y_2$ and set

$$z := y_1 + \frac{\rho}{2}\frac{y_1 - y_2}{\|y_1 - y_2\|_X} \in \mathbb{O}(x, \rho),$$

which holds because $\|z - x\|_X \leq \|y_1 - x\|_X + \frac{\rho}{2} < \rho$. By construction, we thus have that

$$y_1 = \lambda z + (1 - \lambda)y_2 \quad \text{for} \quad \lambda := \frac{\|y_1 - y_2\|_X}{\|y_1 - y_2\|_X + \frac{\rho}{2}} \in (0, 1),$$

and the convexity of F now implies that $F(y_1) \leq \lambda F(z) + (1 - \lambda)F(y_2)$. Together with the definition of λ as well as $F(z) \leq M$ and $-F(y_2) \leq m = M - 2F(x)$, this yields the estimate

$$\begin{aligned} F(y_1) - F(y_2) \leq \lambda(F(z) - F(y_2)) &\leq \lambda(2M - 2F(x)) \\ &= \frac{2(M - F(x))}{\|y_1 - y_2\|_X + \frac{\rho}{2}}\|y_1 - y_2\|_X \\ &\leq \frac{2(M - F(x))}{\frac{\rho}{2}}\|y_1 - y_2\|_X. \end{aligned}$$

Exchanging the roles of y_1 and y_2, we obtain that

$$|F(y_1) - F(y_2)| \leq \frac{4}{\rho}(M - F(x))\|y_1 - y_2\|_X \quad \text{for all } y_1, y_2 \in \mathbb{O}\left(x, \frac{\rho}{2}\right)$$

and hence the local Lipschitz continuity with constant $L(x, \rho/2) := \frac{4}{\rho}(M - F(x))$. □

This result can be extended by showing that convex functions are bounded everywhere in the interior (again a topological concept!) of their effective domain. As an intermediary step, we first consider the scalar case.[2]

[2] With a bit more effort, one can show that the claim holds for $F : \mathbb{R}^N \to \overline{\mathbb{R}}$ with arbitrary $N \in \mathbb{N}$; see, e.g., [Schirotzek, 2007, Corollary 1.4.2].

Lemma 3.11. *If $f : \mathbb{R} \to \overline{\mathbb{R}}$ is convex, then f is locally bounded from above on* $\operatorname{int}(\operatorname{dom} f)$.

Proof. Let $x \in (\operatorname{dom} f)^o$, i.e., there exist $a, b \in \mathbb{R}$ with $x \in (a, b) \subset \operatorname{dom} f$; by possibly shrinking the interval we can even assume that $[a, b] \subset \operatorname{dom} f$. Let now $z \in (a, b)$. Since intervals are convex, there exists a $\lambda \in (0, 1)$ with $z = \lambda a + (1 - \lambda)b$. By convexity, we thus have

$$f(z) \leq \lambda f(a) + (1 - \lambda) f(b) \leq \max\{|f(a)|, |f(b)|\} < \infty.$$

Hence f is locally bounded from above in x. □

The proof of the general case requires further assumptions on X and F.

Theorem 3.12. *Let X be a Banach space. If $F : X \to \overline{\mathbb{R}}$ is convex and lower semicontinuous, then F is locally bounded from above on* $\operatorname{int}(\operatorname{dom} F)$.

Proof. We first show the claim for the case $x = 0 \in \operatorname{int}(\operatorname{dom} F)$, which implies in particular that $M := |F(0)|$ is finite. Consider now for arbitrary $h \in X$ the mapping

$$f : \mathbb{R} \to \overline{\mathbb{R}}, \qquad t \mapsto F(th).$$

It is straightforward to verify that f is convex and satisfies $0 \in \operatorname{int}(\operatorname{dom} f)$. By Lemmas 3.10 and 3.11, f is thus locally Lipschitz continuous near 0; hence in particular $|f(t) - f(0)| \leq Lt \leq 1$ for sufficiently small $t > 0$. The reverse triangle inequality therefore yields a $\delta > 0$ with

$$F(0 + th) \leq |F(0 + th)| = |f(t)| \leq |f(0)| + 1 = M + 1 \qquad \text{for all } t \in [0, \delta].$$

Hence 0 lies in the algebraic interior of the sublevel set $\operatorname{sub}_{M+1} F$, which is convex and closed (since we assumed F to be lower semicontinuous) by Lemma 3.3. The core–int Lemma 1.2 thus yields that $0 \in \operatorname{int}(\operatorname{sub}_{M+1} F)$, i.e., there exists a $\rho > 0$ with $F(z) \leq M + 1$ for all $z \in \mathbb{O}(0, \rho)$.

For the general case $x \in \operatorname{int}(\operatorname{dom} F)$, consider

$$\tilde{F} : X \to \overline{\mathbb{R}}, \qquad y \mapsto F(y + x).$$

Again, it is straightforward to verify convexity and lower semicontinuity of $\tilde{F}$ and that $0 \in \operatorname{int}(\operatorname{dom} \tilde{F})$. It follows from what we've shown so far that $\tilde{F}$ is locally bounded from above on $\mathbb{O}(0, \rho)$, which immediately implies that F is locally bounded from above on $\mathbb{O}(x, \rho)$. □

Together with Lemma 3.10, we thus obtain the desired result.

Theorem 3.13. *Let X be a Banach space. If $F : X \to \overline{\mathbb{R}}$ is convex and lower semicontinuous, then F is locally Lipschitz continuous on* $\operatorname{int}(\operatorname{dom} F)$.

We shall have several more occasions to observe the unreasonably nice behavior of convex lower semicontinuous functions on the interior of their effective domain.

4 CONVEX SUBDIFFERENTIALS

For convex functionals, we can use the general properties from the previous chapter to obtain explicit optimality conditions. We do this by first deriving a Fermat principle in terms of a generalized derivative that can be used to characterize global minimizers of nonsmooth functionals. The remainder of the chapter is then devoted to the explicit characterization of this generalized derivative specifically for convex lower semicontinuous functionals; first directly for elementary examples, then for more complicated functions by deriving calculus rules like a sum and a chain rule.

4.1 DEFINITION AND BASIC PROPERTIES

The motivation for our notion of generalized derivative is geometric: The classical derivative $f'(t)$ of a scalar function $f : \mathbb{R} \to \mathbb{R}$ at t can be interpreted as the slope of the tangent at $f(t)$. If the function is not differentiable, the tangent – if it exists at all – need no longer be unique. The idea is then to define as the generalized derivative the *set of all* tangent slopes. Correspondingly, we define in a normed vector space X the *(convex) subdifferential* of $F : X \to \overline{\mathbb{R}}$ at $x \in \operatorname{dom} F$ as

$$\partial F(x) := \{x^* \in X^* \mid \langle x^*, \tilde{x} - x\rangle_X \leq F(\tilde{x}) - F(x) \quad \text{for all } \tilde{x} \in X\}\,. \tag{4.1}$$

(Note that $\tilde{x} \notin \operatorname{dom} F$ is allowed since in this case the inequality is trivially satisfied.) For $x \notin \operatorname{dom} F$, we set $\partial F(x) = \emptyset$. An element $x^* \in \partial F(x)$ is called a *subderivative*. (Following the terminology for classical derivatives, we reserve the more common term *subgradient* for its Riesz representation $z_{x^*} \in X$ when X is a Hilbert space.)

The following example shows that the subdifferential can also be empty for $x \in \operatorname{dom} F$, even if F is convex.

Example 4.1. We take $X = \mathbb{R}$ (and hence $X^* \cong X = \mathbb{R}$) and consider

$$F(x) = \begin{cases} -\sqrt{x} & \text{if } x \geq 0, \\ \infty & \text{if } x < 0. \end{cases}$$

Since (3.1) is trivially satisfied if x or y is negative, we can assume $x, y \geq 0$ so that

we are allowed to take the square of both sides of (3.1). A straightforward algebraic manipulation then shows that this is equivalent to $\lambda(\lambda-1)(\sqrt{x}-\sqrt{y})^2 \geq 0$, which holds for any $x, y \geq 0$ and $\lambda \in [0,1]$. Hence F is convex.

However, for $x = 0$, any $x^* \in \partial F(0)$ by definition must satisfy

$$x^* \cdot \tilde{x} \leq -\sqrt{\tilde{x}} \qquad \text{for all } \tilde{x} \in \mathbb{R}.$$

Taking now $\tilde{x} > 0$ arbitrary, we can divide by it on both sides and let $\tilde{x} \to 0$ to obtain

$$x^* \leq -\left(\sqrt{\tilde{x}}\right)^{-1} \to -\infty.$$

This is impossible for $x^* \in \mathbb{R} \cong X^*$. Hence, $\partial F(0)$ is empty.

In fact, it will become clear that the nonexistence of tangents is much more problematic than the nonuniqueness. However, we will later show that for proper, convex, and lower semicontinuous functionals, $\partial F(x)$ is nonempty (and bounded) for all $x \in \operatorname{int}(\operatorname{dom} F)$; see Theorem 13.17. Furthermore, it follows directly from the definition that for all $x \in X$, the set $\partial F(x)$ is convex and weakly-$*$ closed.

The definition immediately yields a Fermat principle.

Theorem 4.2 (Fermat principle). *Let $F : X \to \overline{\mathbb{R}}$ and $\bar{x} \in \operatorname{dom} F$. Then the following statements are equivalent:*

(i) $0 \in \partial F(\bar{x})$;

(ii) $F(\bar{x}) = \min_{x \in X} F(x)$.

Proof. This is a direct consequence of the definitions: $0 \in \partial F(\bar{x})$ if and only if

$$0 = \langle 0, \tilde{x} - \bar{x} \rangle_X \leq F(\tilde{x}) - F(\bar{x}) \qquad \text{for all } \tilde{x} \in X,$$

i.e., $F(\bar{x}) \leq F(\tilde{x})$ for all $\tilde{x} \in X$.[1] □

This matches the geometrical intuition: If $X = \mathbb{R} \cong X^*$, the affine function $\tilde{F}(\tilde{x}) := F(x) + x^*(\tilde{x} - x)$ with $x^* \in \partial F(x)$ describes a tangent at $(x, F(x))$ with slope x^*; the condition $x^* = 0 \in \partial F(\tilde{x})$ thus means that F has a horizontal tangent in $\bar{x}$. (Conversely, the

[1] Note that convexity of F is not required for Theorem 4.2. The condition $0 \in \partial F(\bar{x})$ therefore characterizes the global(!) minimizers of *any* function F. However, nonconvex functionals can also have local minimizers, for which the subdifferential inclusion is not satisfied. In fact, (convex) subdifferentials of nonconvex functionals are usually empty. (And conversely, one can show that $\partial F(x) \neq \emptyset$ for all $x \in \operatorname{dom} F$ implies that F is convex.) This leads to problems in particular for the proof of calculus rules, for which we will indeed have to assume convexity.

function from Example 4.1 only has a vertical tangent in $x = 0$, which corresponds to an infinite slope that is not an element of any vector space.)

Not surprisingly, the convex subdifferential behaves more nicely for convex functions. The key property is an alternative characterization using directional derivatives, which exist (at least in the extended real-valued sense) for any convex function.

Lemma 4.3. *Let $F : X \to \overline{\mathbb{R}}$ be convex and let $x \in \operatorname{dom} F$ and $h \in X$ be given. Then:*

(i) the function

$$\varphi : (0, \infty) \to \overline{\mathbb{R}}, \qquad t \mapsto \frac{F(x+th) - F(x)}{t},$$

is increasing;

(ii) there exists a limit $F'(x;h) = \lim_{t \searrow 0} \varphi(t) \in [-\infty, \infty]$, which satisfies

$$F'(x;h) \leq F(x+h) - F(x);$$

(iii) if $x \in \operatorname{int}(\operatorname{dom} F)$, the limit $F'(x;h)$ is finite.

Proof. (i): Inserting the definition and sorting terms shows that for all $0 < s \leq t$, the condition $\varphi(s) \leq \varphi(t)$ is equivalent to

$$F(x+sh) \leq \frac{s}{t} F(x+th) + \left(1 - \frac{s}{t}\right) F(x),$$

which follows from the convexity of F since $x + sh = \frac{s}{t}(x+th) + (1 - \frac{s}{t})x$.

(ii): The claim immediately follows from (i) since

$$F'(x;h) = \lim_{t \searrow 0} \varphi(t) = \inf_{t > 0} \varphi(t) \leq \varphi(1) = F(x+h) - F(x) \in \overline{\mathbb{R}}.$$

(iii): Since $\operatorname{int}(\operatorname{dom} F)$ is contained in the algebraic interior of $\operatorname{dom} F$, there exists an $\varepsilon > 0$ such that $x + th \in \operatorname{dom} F$ for all $t \in (-\varepsilon, \varepsilon)$. Proceeding as in (i), we obtain that $\varphi(s) \leq \varphi(t)$ for all $s < t < 0$ as well. From $x = \frac{1}{2}(x+th) + \frac{1}{2}(x-th)$ for $t > 0$, we also obtain that

$$\varphi(-t) = \frac{F(x-th) - F(x)}{-t} \leq \frac{F(x+th) - F(x)}{t} = \varphi(t)$$

and hence that φ is increasing on all $\mathbb{R} \setminus \{0\}$. As in (ii), the choice of ε now implies that

$$-\infty < \varphi(-\varepsilon) \leq F'(x;h) \leq \varphi(\varepsilon) < \infty. \qquad \square$$

Lemma 4.4. *Let $F : X \to \overline{\mathbb{R}}$ be convex and $x \in \operatorname{dom} F$. Then*

$$\partial F(x) = \{x^* \in X^* \mid \langle x^*, h \rangle_X \leq F'(x;h) \text{ for all } h \in X\}.$$

Proof. Since any $\tilde{x} \in X$ can be written as $\tilde{x} = x + h$ for some $h \in X$ and vice versa, it suffices to show that for any $x^* \in X^*$, the following statements are equivalent:

(i) $\langle x^*, h\rangle_X \leq F'(x;h)$ for all $h \in X$;

(ii) $\langle x^*, h\rangle_X \leq F(x+h) - F(x)$ for all $h \in X$.

If $x^* \in X^*$ satisfies $\langle x^*, h\rangle_X \leq F'(x;h)$ for all $h \in X$, we immediately obtain from Lemma 4.3 (ii) that

$$\langle x^*, h\rangle_X \leq F'(x;h) \leq F(x+h) - F(x) \qquad \text{for all } h \in X.$$

Setting $\tilde{x} = x + h \in X$ then yields $x^* \in \partial F(x)$.

Conversely, if $\langle x^*, h\rangle \leq F(x+h) - F(x)$ holds for all $h \in X$, it also holds for th for all $h \in X$ and $t > 0$. Dividing by t and passing to the limit (which exists by Lemma 4.3 (ii)) then yields that

$$\langle x^*, h\rangle_X \leq \lim_{t \searrow 0} \frac{F(x+th) - F(x)}{t} = F'(x;h). \qquad \square$$

4.2 FUNDAMENTAL EXAMPLES

We now look at some examples. First, the construction from the directional derivative indicates that the subdifferential is indeed a generalization of the Gâteaux derivative.

Theorem 4.5. *Let $F : X \to \overline{\mathbb{R}}$ be convex. If F is Gâteaux differentiable at x, then $\partial F(x) = \{DF(x)\}$.*

Proof. By definition of the Gâteaux derivative, we have that

$$\langle DF(x), h\rangle_X = DF(x)h = F'(x;h) \quad \text{for all } h \in X.$$

Lemma 4.4 now immediately yields $DF(x) \in \partial F(x)$. Conversely, $x^* \in \partial F(x)$ again by Lemma 4.4 implies that

$$\langle x^*, h\rangle_X \leq F'(x;h) = \langle DF(x), h\rangle_X \quad \text{for all } h \in X.$$

Taking the supremum over all h with $\|h\|_X \leq 1$ now yields that $\|x^* - DF(x)\|_{X^*} \leq 0$, i.e., $x^* = DF(x)$. $\square$

The converse holds as well: If $x \in \operatorname{int}(\operatorname{dom} F)$ and $\partial F(x)$ is a singleton, then F is Gâteaux differentiable; see Theorem 13.18.

Of course, we also want to compute subdifferentials of functionals that are not differentiable. The canonical example is the norm $\|\cdot\|_X$ on a normed vector space, which even for $X = \mathbb{R}$ is not differentiable at $x = 0$.

Theorem 4.6. *For any $x \in X$,*

$$\partial(\|\cdot\|_X)(x) = \begin{cases} \{x^* \in X^* \mid \langle x^*, x\rangle_X = \|x\|_X \text{ and } \|x^*\|_{X^*} = 1\} & \text{if } x \neq 0, \\ \mathbb{B}_{X^*} & \text{if } x = 0. \end{cases}$$

Proof. For $x = 0$, we have $x^* \in \partial(\|\cdot\|_X)(x)$ by definition if and only if

$$\langle x^*, \tilde{x}\rangle_X \leq \|\tilde{x}\|_X \qquad \text{for all } \tilde{x} \in X \setminus \{0\}$$

(since the inequality is trivial for $\tilde{x} = 0$), which by the definition of the operator norm holds if and only if $\|x^*\|_{X^*} \leq 1$.

Let now $x \neq 0$ and consider $x^* \in \partial(\|\cdot\|_X)(x)$. Inserting first $\tilde{x} = 0$ and then $\tilde{x} = 2x$ into the definition (4.1) yields the sequence of inequalities

$$\|x\|_X \leq \langle x^*, x\rangle_X = \langle x^*, 2x - x\rangle \leq \|2x\|_X - \|x\|_X = \|x\|_X,$$

which imply that $\langle x^*, x\rangle_X = \|x\|_X$. Similarly, we have for all $\tilde{x} \in X$ that

$$\langle x^*, \tilde{x}\rangle_X = \langle x^*, (\tilde{x} + x) - x\rangle_X \leq \|\tilde{x} + x\|_X - \|x\|_X \leq \|\tilde{x}\|_X,$$

As in the case $x = 0$, this implies that $\|x^*\|_{X^*} \leq 1$. For $\tilde{x} = x/\|x\|_X$ we further have that

$$\langle x^*, \tilde{x}\rangle_X = \|x\|_X^{-1}\langle x^*, x\rangle_X = \|x\|_X^{-1}\|x\|_X = 1.$$

Hence, $\|x^*\|_{X^*} = 1$ is in fact attained.

Conversely, let $x^* \in X^*$ with $\langle x^*, x\rangle_X = \|x\|_X$ and $\|x^*\|_{X^*} = 1$. Then we obtain for all $\tilde{x} \in X$ from (1.1) the relation

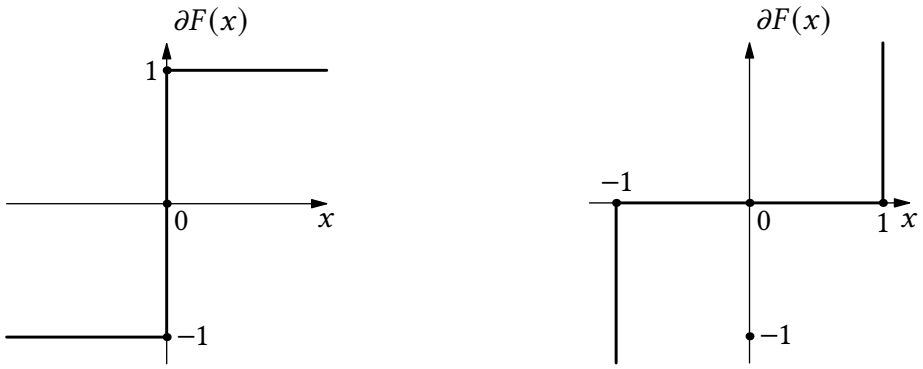
$$\langle x^*, \tilde{x} - x\rangle_X = \langle x^*, \tilde{x}\rangle_X - \langle x^*, x\rangle_X \leq \|\tilde{x}\|_X - \|x\|_X,$$

and hence $x^* \in \partial(\|\cdot\|_X)(x)$ by definition. □

Example 4.7. In particular, we obtain for $X = \mathbb{R}$ the subdifferential of the absolute value function as[2]

$$\partial(|\cdot|)(t) = \operatorname{sign}(t) := \begin{cases} \{1\} & \text{if } t > 0, \\ \{-1\} & \text{if } t < 0, \\ [-1, 1] & \text{if } t = 0, \end{cases} \tag{4.2}$$

cf. Figure 4.1a.



(a) $F(x) = |x|$ (b) $F(x) = \delta_{[-1,1]}(x)$

Figure 4.1: Illustration of graph ∂F for two different functions $F : \mathbb{R} \to \overline{\mathbb{R}}$.

We can also give a more explicit characterization of the subdifferential of the indicator functional of a set $C \subset X$.

Lemma 4.8. *For any $C \subset X$,*

$$\partial\delta_C(x) = \{x^* \in X^* \mid \langle x^*, \tilde{x} - x\rangle_X \leq 0 \quad \text{for all } \tilde{x} \in C\}\,.$$

Proof. For any $x \in C = \operatorname{dom}\delta_C$, we have that

$$\begin{aligned} x^* \in \partial\delta_C(x) &\Leftrightarrow \langle x^*, \tilde{x} - x\rangle_X \leq \delta_C(\tilde{x}) \quad \text{for all } \tilde{x} \in X \\ &\Leftrightarrow \langle x^*, \tilde{x} - x\rangle_X \leq 0 \qquad \text{for all } \tilde{x} \in C, \end{aligned}$$

since the first inequality is trivially satisfied for all $\tilde{x} \notin C$. □

The set $N_C(x) := \partial\delta_C(x)$ is also called the (convex) *normal cone* to C at x (which may be empty if C is not convex). We illustrate such sets in Figure 4.2. Depending on the set C, this can be made even more explicit.

Example 4.9. Let $X = \mathbb{R}$ and $C = [-1, 1]$, and let $t \in C$. Then we have $x^* \in \partial\delta_{[-1,1]}(t)$ if and only if $x^*(\tilde{t} - t) \leq 0$ for all $\tilde{t} \in [-1, 1]$. We proceed by distinguishing three cases.

Case 1: $t = 1$. Then $\tilde{t} - t \in [-2, 0]$, and hence the product is nonpositive if and only if $x^* \geq 0$.

Case 2: $t = -1$. Then $\tilde{t} - t \in [0, 2]$, and hence the product is nonpositive if and only if $x^* \leq 0$.

Case 3: $t \in (-1, 1)$. Then $\tilde{t} - t$ can be positive as well as negative, and hence only

[2]Note that this set-valued definition of $\operatorname{sign}(t)$ differs from the usual (single-valued) one, in particular for $t = 0$; to make this distinction clear, one often refers to (4.2) as the *sign in the sense of convex analysis*. Throughout this book, we will always use the sign in this sense.

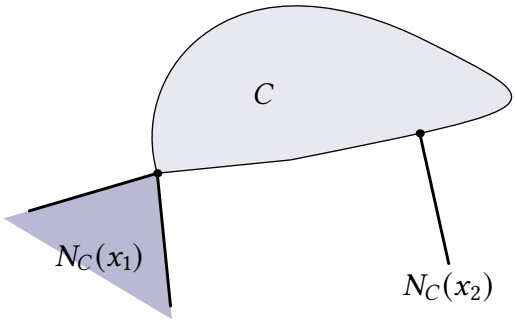


Figure 4.2: Normal cones of a convex set C at two points x_1 and x_2.

$x^* = 0$ is possible.

We thus obtain that

$$\partial\delta_{[-1,1]}(t) = \begin{cases} [0,\infty) & \text{if } t = 1, \\ (-\infty, 0] & \text{if } t = -1, \\ \{0\} & \text{if } t \in (-1,1), \\ \emptyset & \text{if } t \in \mathbb{R} \setminus [-1,1], \end{cases} \tag{4.3}$$

cf. Figure 4.1b. Readers familiar with (non)linear optimization will recognize these as the *complementarity conditions* for Lagrange multipliers corresponding to the inequalities $-1 \le t \le 1$.

Conversely, subdifferentials of convex functionals can be obtained from normal cones to corresponding epigraphs (which are convex sets by Lemma 3.1). This relation will be the basis for defining further subdifferentials for more general classes of mappings in Part IV. We illustrate this result for the absolute value function of Example 4.7 in Figure 4.3.

Lemma 4.10. *Let $F : X \to \overline{\mathbb{R}}$ be convex and $x \in \operatorname{dom} F$. Then $x^* \in \partial F(x)$ if and only if $(x^*, -1) \in N_{\operatorname{epi} F}(x, F(x))$.*

Proof. By definition of the normal cone, $(x^*, -1) \in N_{\operatorname{epi} F}(x, F(x))$ is equivalent to

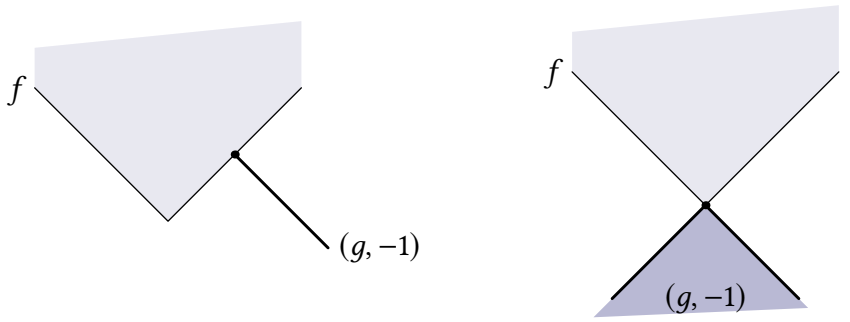
$$\langle x^*, \tilde{x} - x\rangle_X - (\tilde{t} - F(x)) \le 0 \quad \text{for all } (\tilde{x}, \tilde{t}) \in \operatorname{epi} F, \tag{4.4}$$

i.e., for all $\tilde{x} \in X$ and $\tilde{t} \ge F(\tilde{x})$. Taking $\tilde{t} = F(\tilde{x})$ and rearranging, this yields that $x^* \in \partial F(x)$.

Conversely, from $x^* \in \partial F(x)$ we immediately obtain that

$$\langle x^*, \tilde{x} - x\rangle_X \le F(\tilde{x}) - F(x) \le \tilde{t} - F(x) \quad \text{for all } \tilde{x} \in X, \tilde{t} \ge F(\tilde{x}),$$

i.e., (4.4) and thus $(x^*, -1) \in \operatorname{epi} F$. □



(a) $\partial f(x) = \{\operatorname{sign} x\}$ at $x \neq 0$ (b) $\partial f(x) = [-1, 1]$ at $x = 0$

Figure 4.3: Subdifferentials of $f(x) = |x|$ in terms of the normal cone of the epigraph.

The following result furnishes a crucial link between finite- and infinite-dimensional convex optimization. We again assume (as we will from now on) that $\Omega \subset \mathbb{R}^d$ is open and bounded.

Theorem 4.11. *Let $f : \mathbb{R} \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and let $F : L^p(\Omega) \to \overline{\mathbb{R}}$ with $1 \leq p < \infty$ be as in Lemma 3.7. Then we have for all $u \in \operatorname{dom} F$ with $q := \frac{p}{p-1}$ that*

$$\partial F(u) = \{u^* \in L^q(\Omega) \mid u^*(x) \in \partial f(u(x)) \quad \text{for almost every } x \in \Omega\}.$$

Proof. Let $u, \tilde{u} \in \operatorname{dom} F$, i.e., $f \circ u, f \circ \tilde{u} \in L^1(\Omega)$ (otherwise there is nothing to show), and let $u^* \in L^q(\Omega)$ be arbitrary. If $u^*(x) \in \partial f(u(x))$ almost everywhere, we can integrate over all $x \in \Omega$ to obtain

$$F(\tilde{u}) - F(u) = \int_\Omega f(\tilde{u}(x)) - f(u(x))\,dx \geq \int_\Omega u^*(x)(\tilde{u}(x) - u(x))\,dx = \langle u^*, \tilde{u} - u\rangle_{L^p},$$

i.e., $u^* \in \partial F(u)$.

Conversely, let $u^* \in \partial F(u)$. Then by definition it holds that

$$\int_\Omega u^*(x)(\tilde{u}(x) - u(x))\,dx \leq \int_\Omega f(\tilde{u}(x)) - f(u(x))\,dx \quad \text{for all } \tilde{u} \in L^p(\Omega).$$

Let now $t \in \mathbb{R}$ be arbitrary and let $A \subset \Omega$ be an arbitrary measurable set. Setting

$$\tilde{u}(x) := \begin{cases} t & \text{if } x \in A, \\ u(x) & \text{if } x \notin A, \end{cases}$$

the above inequality implies due to $\tilde{u} \in L^p(\Omega)$ that

$$\int_A u^*(x)(t - u(x))\,dx \leq \int_A f(t) - f(u(x))\,dx.$$

Since A was arbitrary, it must hold that

$$u^*(x)(t - u(x)) \leq f(t) - f(u(x)) \qquad \text{for almost every } x \in \Omega.$$

Furthermore, since $t \in \mathbb{R}$ was arbitrary, we obtain that $u^*(x) \in \partial f(u(x))$ for almost every $x \in \Omega$. □

Remark 4.12. A similar representation representation can be shown for vector-valued and spatially-dependent integrands $f : \Omega \times \mathbb{R} \to \mathbb{R}^m$ under stronger assumptions; see, e.g., [Rockafellar, 1976a, Corollary 3F].

A similar proof shows that for $F : \mathbb{R}^N \to \overline{\mathbb{R}}$ with $F(x) = \sum_{i=1}^N f_i(x_i)$ and $f_i : \mathbb{R} \to \overline{\mathbb{R}}$ convex, we have for any $x \in \operatorname{dom} F$ that

$$\partial F(x) = \left\{ x^* \in \mathbb{R}^N \;\middle|\; x_i^* \in \partial f_i(x_i), \quad 1 \leq i \leq N \right\}.$$

Together with the above examples, this yields componentwise expressions for the subdifferential of the norm $\|\cdot\|_1$ as well as of the indicator functional of the unit ball with respect to the supremum norm in $\mathbb{R}^N$.

4.3 CALCULUS RULES

As for classical derivatives, one rarely obtains subdifferentials from the fundamental definition but rather by applying calculus rules. It stands to reason that these are more difficult to derive the weaker the derivative concept is (i.e., the more functionals are differentiable in that sense). For convex subdifferentials, the following two rules still follow directly from the definition.

Lemma 4.13. *Let $F : X \to \overline{\mathbb{R}}$ be convex and $x \in \operatorname{dom} F$. Then,*

(i) $\partial(\lambda F)(x) = \lambda(\partial F(x)) := \{\lambda x^* \mid x^* \in \partial F(x)\}$ *for* $\lambda \geq 0$;

(ii) $\partial F(\cdot + x_0)(x) = \partial F(x + x_0)$ *for* $x_0 \in X$ *with* $x + x_0 \in \operatorname{dom} F$.

Already the sum rule is considerably more delicate.

Theorem 4.14 (sum rule). *Let X be a Banach space, $F, G : X \to \overline{\mathbb{R}}$ be convex and lower semicontinuous, and $x \in \operatorname{dom} F \cap \operatorname{dom} G$. Then*

$$\partial F(x) + \partial G(x) \subset \partial(F + G)(x),$$

with equality if there exists an $x_0 \in \operatorname{int}(\operatorname{dom} F) \cap \operatorname{dom} G$.

Proof. The inclusion follows directly from adding the definitions of the two subdifferentials. Let therefore $x \in \operatorname{dom} F \cap \operatorname{dom} G$ and $x^* \in \partial(F+G)(x)$, i.e., satisfying

$$\langle x^*, \tilde{x} - x\rangle_X \leq (F(\tilde{x}) + G(\tilde{x})) - (F(x) + G(x)) \quad \text{for all } \tilde{x} \in X. \tag{4.5}$$

Our goal is now to use (as in the proof of Lemma 3.5) the characterization of convex functionals via their epigraph together with the Hahn–Banach separation theorem to construct a bounded linear functional $y^* \in \partial G(x) \subset X^*$ with $x^* - y^* \in \partial F(x)$, i.e.,

$$\begin{aligned} F(\tilde{x}) - F(x) - \langle x^*, \tilde{x} - x\rangle_X &\geq \langle y^*, x - \tilde{x}\rangle_X \quad \text{for all } \tilde{x} \in \operatorname{dom} F, \\ G(x) - G(\tilde{x}) &\leq \langle y^*, x - \tilde{x}\rangle_X \quad \text{for all } \tilde{x} \in \operatorname{dom} G. \end{aligned}$$

For that purpose, we define the sets

$$\begin{aligned} C_1 &:= \{(\tilde{x}, t - (F(x) - \langle x^*, x\rangle_X)) \mid \tilde{x} \in \operatorname{dom} F,\ t \geq F(\tilde{x}) - \langle x^*, \tilde{x}\rangle_X\}, \\ C_2 &:= \{(\tilde{x}, G(x) - t) \mid \tilde{x} \in \operatorname{dom} G,\ t \geq G(\tilde{x})\}, \end{aligned}$$

i.e.,

$$C_1 = \operatorname{epi}(F - x^*) - (0, F(x) - \langle x^*, x\rangle_X), \qquad C_2 = -(\operatorname{epi} G - (0, G(x))),$$

cf. Figure 4.4. To apply Corollary 1.6 to these sets, we have to verify its conditions.

(i) Since $x \in \operatorname{dom} F \cap \operatorname{dom} G$, both C_1 and C_2 are nonempty. Furthermore, since F and G are convex, it is straightforward (if tedious) to verify from the definition that C_1 and C_2 are convex.

(ii) The critical point is of course the nonemptiness of $\operatorname{int} C_1$, for which we argue as follows. Since $x_0 \in \operatorname{int}(\operatorname{dom} F)$, we know from Theorem 3.12 that F is bounded in an open neighborhood $U \subset \operatorname{int}(\operatorname{dom} F)$ of x_0. We can thus find an open interval $I \subset \mathbb{R}$ such that $U \times I \subset C_1$. Since $U \times I$ is open by the definition of the product topology on $X \times \mathbb{R}$, any (x_0, α) with $\alpha \in I$ is an interior point of C_1.

(iii) It remains to show that $\operatorname{int} C_1 \cap C_2 = \emptyset$. Assume there exists a $(\tilde{x}, \alpha) \in \operatorname{int} C_1 \cap C_2$. But then the definitions of these sets and of the product topology imply that

$$F(\tilde{x}) - F(x) - \langle x^*, \tilde{x} - x\rangle_X < \alpha \leq G(x) - G(\tilde{x}),$$

contradicting (4.5). Hence $\operatorname{int} C_1$ and C_2 are disjoint.

We can thus apply Corollary 1.6 to obtain a pair $(z^*, s) \in (X^* \times \mathbb{R}) \setminus \{(0,0)\} \cong (X \times \mathbb{R})^* \setminus \{(0,0)\}$ and a $\lambda \in \mathbb{R}$ with

$$\langle z^*, \tilde{x}\rangle_X + s(t - (F(x) - \langle x^*, x\rangle_X)) \leq \lambda, \quad \tilde{x} \in \operatorname{dom} F, t \geq F(\tilde{x}) - \langle x^*, \tilde{x}\rangle_X, \tag{4.6a}$$

$$\langle z^*, \tilde{x}\rangle_X + s(G(x) - t) \geq \lambda, \quad \tilde{x} \in \operatorname{dom} G, t \geq G(\tilde{x}). \tag{4.6b}$$

We now show that $s < 0$. If $s = 0$, we can insert $\tilde{x} = x_0 \in \operatorname{dom} F \cap \operatorname{dom} G$ to obtain the contradiction

$$\langle z^*, x_0\rangle_X < \lambda \leq \langle z^*, x_0\rangle_X,$$

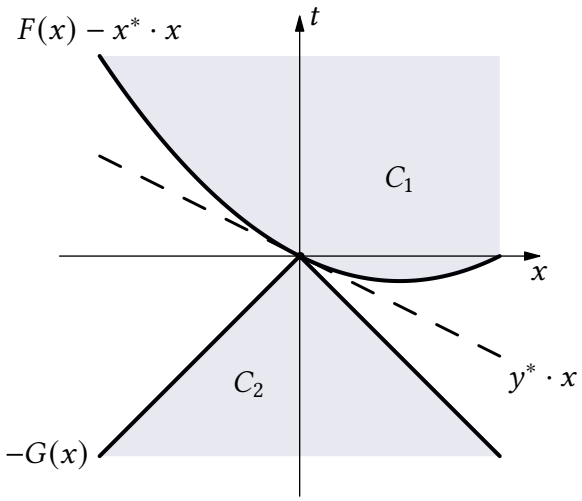


Figure 4.4: Illustration of the proof of Theorem 4.14 for $F(x) = \frac{1}{2}|x|^2$, $G(x) = |x|$, and $x^* = \frac{1}{2} \in \partial(F+G)(0)$. The dashed line is the separating hyperplane $\{(x,t) \mid z^* \cdot x + st = \lambda\}$, i.e., $\lambda = 0$, $z^* = -1$, $s = -2$ and hence $y^* = \frac{1}{2} \in \partial G(0)$.

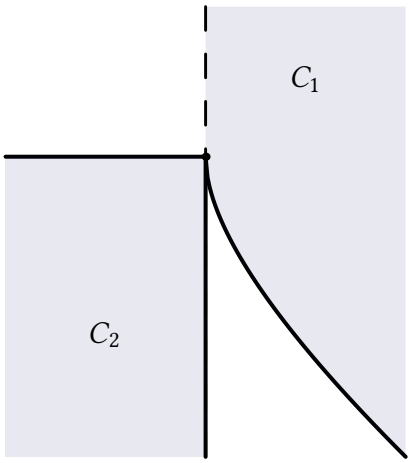


Figure 4.5: Illustration of the situation in Example 4.15. Here the dashed separating hyperplane corresponds to the vertical line $\{(x,t) \mid x = 0\}$ (i.e., $z^* = 1$ and $s = 0$), and hence $y^* \notin \mathbb{R}$.

which follows since (x_0, α) for α large enough is an interior point of C_1 and hence can be *strictly* separated from C_2 by Theorem 1.5 (i). If $s > 0$, choosing $t > F(x) - \langle x^*, x\rangle_X$ makes the term in parentheses in (4.6a) strictly positive, and taking $t \to \infty$ with fixed $\tilde{x}$ leads to a contradiction to the boundedness by λ.

Hence $s < 0$, and (4.6a) with $t = F(\tilde{x}) - \langle x^*, \tilde{x}\rangle_X$ and (4.6b) with $t = G(\tilde{x})$ imply that

$$\begin{aligned} F(\tilde{x}) - F(x) - \langle x^*, \tilde{x} - x\rangle_X &\geq s^{-1}(\lambda - \langle z^*, \tilde{x}\rangle_X) \quad \text{for all } \tilde{x} \in \operatorname{dom} F, \\ G(x) - G(\tilde{x}) &\leq s^{-1}(\lambda - \langle z^*, \tilde{x}\rangle_X) \quad \text{for all } \tilde{x} \in \operatorname{dom} G. \end{aligned}$$

Taking $\tilde{x} = x \in \operatorname{dom} F \cap \operatorname{dom} G$ in both inequalities immediately yields that $\lambda = \langle z^*, x\rangle_X$. Hence, $y^* = s^{-1}z^* \in X^*$ is the desired functional with $(x^* - y^*) \in \partial F(x)$ and $y^* \in \partial G(x)$, i.e., $x^* \in \partial F(x) + \partial G(x)$. □

The following example demonstrates that the inclusion is strict in general (although naturally the situation in infinite-dimensional vector spaces is nowhere near as obvious).

Example 4.15. We take again $X = \mathbb{R}$ and $F : X \to \overline{\mathbb{R}}$ from Example 4.1, i.e.,

$$F(x) = \begin{cases} -\sqrt{x} & \text{if } x \geq 0, \\ \infty & \text{if } x < 0, \end{cases}$$

as well as $G(x) = \delta_{(-\infty,0]}(x)$. Both F and G are convex, and $0 \in \operatorname{dom} F \cap \operatorname{dom} G$. In fact, $(F+G)(x) = \delta_{\{0\}}(x)$ and hence it is straightforward to verify that $\partial(F+G)(0) = \mathbb{R}$.

On the other hand, we know from Example 4.1 and the argument leading to (4.3) that

$$\partial F(0) = \emptyset, \qquad \partial G(0) = [0, \infty),$$

and hence that

$$\partial F(0) + \partial G(0) = \emptyset \subsetneq \mathbb{R} = \partial(F+G)(0).$$

(As F only admits a vertical tangent as $x = 0$, this example corresponds to the situation where $s = 0$ in (4.6a), cf. Figure 4.5.)

Remark 4.16. There exist alternative conditions that guarantee that the sum rule holds with equality. For example, if X is a Banach space and F and G are in addition lower semicontinuous, this holds under the *Attouch–Brézis condition* that

$$\bigcup_{\lambda \geq 0} \lambda \left(\operatorname{dom} F - \operatorname{dom} G\right) =: Z \text{ is a closed subspace of } X,$$

see [Attouch and Brezis, 1986]. (Note that this condition is not satisfied in Example 4.15 either, since in this case $Z = -\operatorname{dom} G = [0, \infty)$ which is closed but not a subspace.)

It is not difficult to see that the condition $x_0 \in \operatorname{int}(\operatorname{dom} F) \cap \operatorname{dom} G$ in the statement of Lemma 4.13 implies the Attouch–Brézis condition. In fact, the latter allows us to generalize the condition to $x_0 \in \operatorname{ri}(\operatorname{dom} F) \cap \operatorname{dom} G$ where $\operatorname{ri} A$ for a set A denotes the *relative interior*: the interior of A with respect to the smallest closed affine set that contains A. As an example, $\operatorname{ri}\{c\} = \{c\}$ for a point $c \in X$.

By induction, we obtain from this sum rules for an arbitrary (finite) number of functionals (where x_0 has to be an interior point of all but one effective domain). A chain rule for linear operators can be proved similarly.

Theorem 4.17 (chain rule). *Let X, Y be Banach spaces, $K \in \mathbb{L}(X; Y)$, $F : Y \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $x \in \operatorname{dom}(F \circ K)$. Then,*

$$\partial(F \circ K)(x) \supset K^* \partial F(Kx) := \{K^* y^* \mid y^* \in \partial F(Kx)\}$$

with equality if there exists an $x_0 \in X$ with $Kx_0 \in \operatorname{int}(\operatorname{dom} F)$.

Proof. The inclusion is again a direct consequence of the definition: If $y^* \in \partial F(Kx) \subset Y^*$, we in particular have for all $\tilde{y} = K\tilde{x} \in Y$ with $\tilde{x} \in X$ that

$$F(K\tilde{x}) - F(Kx) \geq \langle y^*, K\tilde{x} - Kx \rangle_Y = \langle K^* y^*, \tilde{x} - x \rangle_X,$$

i.e., $x^* := K^* y^* \in \partial(F \circ K) \subset X^*$.

To show the claimed equality under the additional assumption, let $x \in \operatorname{dom}(F \circ K)$ and $x^* \in \partial(F \circ K)(x)$, i.e.,

$$F(Kx) + \langle x^*, \tilde{x} - x\rangle_X \leq F(K\tilde{x}) \quad \text{for all } \tilde{x} \in X.$$

We now construct a $y^* \in \partial F(Kx)$ with $x^* = K^* y^*$ by applying the sum rule to[3]

$$H : X \times Y \to \overline{\mathbb{R}}, \qquad (x, y) \mapsto F(y) + \delta_{\operatorname{graph} K}(x, y).$$

Since K is linear and continuous, $\operatorname{graph} K$ is convex and closed, and hence $\delta_{\operatorname{graph} K}$ is convex and lower semicontinuous. Furthermore, $Kx \in \operatorname{dom} F$ by assumption and thus $(x, Kx) \in \operatorname{dom} H$.

We begin by showing that $x^* \in \partial(F \circ K)(x)$ if and only if $(x^*, 0) \in \partial H(x, Kx)$. First, let $(x^*, 0) \in \partial H(x, Kx)$. Then we have for all $\tilde{x} \in X$, $\tilde{y} \in Y$ that

$$\langle x^*, \tilde{x} - x\rangle_X + \langle 0, \tilde{y} - Kx\rangle_Y \leq F(\tilde{y}) - F(Kx) + \delta_{\operatorname{graph} K}(\tilde{x}, \tilde{y}) - \delta_{\operatorname{graph} K}(x, Kx).$$

In particular, this holds for all $\tilde{y} \in \operatorname{ran}(K) = \{K\tilde{x} \mid \tilde{x} \in X\}$. By $\delta_{\operatorname{graph} K}(\tilde{x}, K\tilde{x}) = 0$ we thus obtain that

$$\langle x^*, \tilde{x} - x\rangle_X \leq F(K\tilde{x}) - F(Kx) \quad \text{for all } \tilde{x} \in X,$$

i.e., $x^* \in \partial(F \circ K)(x)$. Conversely, let $x^* \in \partial(F \circ K)(x)$. Since $\delta_{\operatorname{graph} K}(x, Kx) = 0$ and $\delta_{\operatorname{graph} K}(\tilde{x}, \tilde{y}) \geq 0$, it then follows for all $\tilde{x} \in X$ and $\tilde{y} \in Y$ that

$$\begin{aligned}\langle x^*, \tilde{x} - x\rangle_X + \langle 0, \tilde{y} - Kx\rangle_Y &= \langle x^*, \tilde{x} - x\rangle_X \\ &\leq F(K\tilde{x}) - F(Kx) + \delta_{\operatorname{graph} K}(\tilde{x}, \tilde{y}) - \delta_{\operatorname{graph} K}(x, Kx) \\ &= F(\tilde{y}) - F(Kx) + \delta_{\operatorname{graph} K}(\tilde{x}, \tilde{y}) - \delta_{\operatorname{graph} K}(x, Kx),\end{aligned}$$

where we have used that the last equality holds trivially as $\infty = \infty$ for $\tilde{y} \neq K\tilde{x}$. Hence, $(x^*, 0) \in \partial H(x, Kx)$.

We now consider $\tilde{F} : X \times Y \to \overline{\mathbb{R}}$, $(x, y) \mapsto F(y)$, and $(x_0, Kx_0) \in \operatorname{graph} K = \operatorname{dom} \delta_{\operatorname{graph} K}$. Since $Kx_0 \in \operatorname{int}(\operatorname{dom} F) \subset Y$ by assumption, $(x_0, Kx_0) \in \operatorname{int}(\operatorname{dom} \tilde{F}) = X \times \operatorname{int}(\operatorname{dom} F) \subset X \times Y$ as well. We can thus apply Theorem 4.14 to obtain

$$(x^*, 0) \in \partial H(x, Kx) = \partial \tilde{F}(x, Kx) + \partial \delta_{\operatorname{graph} K}(x, Kx),$$

i.e., $(x^*, 0) = (x_1^*, y_1^*) + (x_2^*, y_2^*)$ for some $(x_1^*, y_1^*) \in \partial \tilde{F}(x, Kx)$ and $(x_2^*, y_2^*) \in \partial \delta_{\operatorname{graph} K}(x, Kx)$.

Finally, we "collapse" these subdifferentials back to the individual spaces to obtain the desired characterization. First, we have $(x_1^*, y_1^*) \in \partial \tilde{F}(x, Kx)$ if and only if

$$\langle x_1^*, \tilde{x} - x\rangle_X + \langle y_1^*, \tilde{y} - Kx\rangle_Y \leq F(\tilde{y}) - F(Kx) \quad \text{for all } \tilde{x} \in X, \tilde{y} \in Y.$$

[3] This technique of "lifting" a problem to a product space in order to separate operators is also useful in many other contexts.

Fixing in turn $\tilde{x} = x$ and $\tilde{y} = Kx$ implies that $y_1^* \in \partial F(Kx)$ and $x_1^* = 0$, respectively. Second, $(x_2^*, y_2^*) \in \partial\delta_{\operatorname{graph} K}(x, Kx)$ if and only if

$$\langle x_2^*, \tilde{x} - x\rangle_X + \langle y_2^*, \tilde{y} - Kx\rangle_Y \leq 0 \quad \text{for all } (\tilde{x}, \tilde{y}) \in \operatorname{graph} K,$$

i.e., for all $\tilde{x} \in X$ and $\tilde{y} = K\tilde{x}$. Therefore,

$$\langle x_2^* + K^* y_2^*, \tilde{x} - x\rangle_X \leq 0 \quad \text{for all } \tilde{x} \in X$$

and hence $x_2^* = -K^* y_2^* \in X^*$. Together we obtain

$$(x^*, 0) = (0, y_1^*) + (-K^* y_2^*, y_2^*),$$

which implies that $y_1^* = -y_2^*$ and thus that $x^* = -K^* y_2^* = K^* y_1^*$ with $y_1^* \in \partial F(Kx)$ as claimed. □

The condition for equality in particular holds if K is surjective and $\operatorname{dom} F$ has nonempty interior. Again, the inequality can be strict.

Example 4.18. Here we take $X = Y = \mathbb{R}$ and again $F : X \to \overline{\mathbb{R}}$ from Examples 4.1 and 4.15 as well as

$$K : \mathbb{R} \to \mathbb{R}, \qquad Kx = 0.$$

Clearly, $(F \circ K)(x) = 0$ for all $x \in \mathbb{R}$ and hence $\partial(F \circ K)(x) = \{0\}$ by Theorem 4.5. On the other hand, $\partial F(0) = \emptyset$ by Example 4.1 and hence

$$K^* \partial F(Kx) = K^* \partial F(0) = \emptyset \subsetneq \{0\}.$$

(Note the problem: K^* is far from surjective, and $\operatorname{ran} K \cap \operatorname{int}(\operatorname{dom} F) = \emptyset$.)

We can also obtain a chain rule when the *inner* mapping is nondifferentiable.

Theorem 4.19. *Let $F : X \to \mathbb{R}$ be convex and $\varphi : \mathbb{R} \to \mathbb{R}$ be convex, increasing, and differentiable. Then $\varphi \circ F$ is convex, and for all $x \in X$,*

$$\partial(\varphi \circ F)(x) = \varphi'(F(x))\partial F(x) = \{\varphi'(F(x))x^* \mid x^* \in \partial F(x)\}.$$

Proof. First, the convexity of $\varphi \circ F$ follows from Lemma 3.4 (iii). To calculate the subdifferential, we fix $x \in X$ and observe from Theorem 3.13 that φ is Lipschitz continuous with

some constant L near $F(x) \in \operatorname{int}(\operatorname{dom}\varphi) = \mathbb{R}$. Thus, for any $h \in X$,

$$\begin{aligned}(\varphi \circ F)'(x;h) &= \lim_{t \searrow 0} \frac{(\varphi \circ F)(x+th) - (\varphi \circ F)(x)}{t} \\ &= \lim_{t \searrow 0} \frac{\varphi(F(x+th)) - \varphi(F(x) + tF'(x;h))}{t} \\ &\quad + \lim_{t \searrow 0} \frac{\varphi(F(x) + tF'(x;h)) - \varphi(F(x))}{t} \\ &\leq \lim_{t \searrow 0} L \left| \frac{F(x+th) - F(x)}{t} - F'(x;h) \right| + \varphi'(F(x);F'(x;h)) \\ &= \varphi'(F(x);F'(x;h)),\end{aligned}$$

where we have used the directional differentiability of F in $x \in \operatorname{int}(\operatorname{dom} F) = X$ in the last step. Similarly, we prove the opposite inequality using $\varphi(t_1) - \varphi(t_2) \geq -L|t_1 - t_2|$ for all t_1, t_2 sufficiently close to $F(x)$. Hence

$$(\varphi \circ F)'(x;h) = \varphi'(F(x);F'(x;h)) = \varphi'(F(x))F'(x;h)$$

by the differentiability of φ.

Now Lemma 4.4 yields that

$$\partial(\varphi \circ F)(x) = \{z^* \in X^* \mid \langle z^*, h\rangle_X \leq \varphi'(F(x))F'(x;h) \text{ for all } h \in X\}.$$

Since $\varphi : \mathbb{R} \to \mathbb{R}$ is monotone and differentiable, $\varphi'(F(x)) \geq 0$. Hence if $\varphi'(F(x)) > 0$, we can set $x^* := \varphi'(F(x))^{-1}z^* \in X^*$; otherwise $z^* = 0$ is the only element of $\partial(\varphi \circ F(x))$. In either case, we can write

$$\partial(\varphi \circ F)(x) = \{\varphi'(F(x))x^* \mid \langle x^*, h\rangle_X \leq F'(x;h) \text{ for all } h \in X\}$$

so that the claim follows by Lemma 4.4. □

Remark 4.20. The differentiability assumption on φ in Theorem 4.19 is not necessary, but the proof is otherwise much more involved and demands the support functional machinery of Section 13.3. See also [Hiriart-Urruty and Lemaréchal, 2001, Section D.4.3] for a version with set-valued F in finite dimensions.

The Fermat principle together with the sum rule yields the following characterization of minimizers of convex functionals under convex constraints.

Corollary 4.21. *Let $U \subset X$ be nonempty, convex, and closed, and let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. If there exists an $x_0 \in \operatorname{int} U \cap \operatorname{dom} F$, then $\bar{x} \in U$ solves*

$$\min_{x \in U} F(x)$$

if and only if $0 \in \partial F(\bar{x}) + N_U(\bar{x})$ *or, in other words, if there exists an* $x^* \in X^*$ *with*

$$\begin{cases} x^* \in \partial F(\bar{x}), \\ \langle x^*, \tilde{x} - \bar{x} \rangle_X \geq 0 \quad \text{for all } \tilde{x} \in U. \end{cases} \tag{4.7}$$

Proof. Due to the assumptions on F and U, we can apply Theorem 4.2 to $J := F + \delta_U$. Furthermore, since $x_0 \in \operatorname{int} U = \operatorname{int}(\operatorname{dom} \delta_U)$, we can also apply Theorem 4.14. Hence F has a minimum in $\bar{x}$ if and only if

$$0 \in \partial J(\bar{x}) = \partial F(\bar{x}) + \partial \delta_U(\bar{x}).$$

Together with the characterization of subdifferentials of indicator functionals as normal cones, this yields (4.7). □

If $F : X \to \mathbb{R}$ is Gâteaux differentiable (and hence finite-valued), (4.7) coincide with the classical *Karush–Kuhn–Tucker conditions*; the existence of an interior point $x_0 \in \operatorname{int} U$ is related to a *Slater condition* in nonlinear optimization needed to show existence of the Lagrange multiplier x^* for inequality constraints.

5 FENCHEL DUALITY

One of the main tools in convex optimization is *duality*: Any convex optimization problem can be related to a *dual problem*, and the joint study of both problems yields additional information about the solution. Our main objective in this chapter, the *Fenchel–Rockafellar duality theorem*, will be our main tool for deriving explicit optimality conditions as well as numerical algorithms for convex minimization problems that can be expressed as the sum of (simple) functionals.

5.1 FENCHEL CONJUGATES

Let X be a normed vector space and $F : X \to \overline{\mathbb{R}}$ be proper but not necessarily convex. We then define the *Fenchel conjugate* (or *convex conjugate*) of F as

$$F^* : X^* \to \overline{\mathbb{R}}, \qquad F^*(x^*) = \sup_{x \in X} \{\langle x^*, x\rangle_X - F(x)\}.$$

(Since $\operatorname{dom} F = \emptyset$ is excluded, we have that $F^*(x^*) > -\infty$ for all $x^* \in X^*$, and hence the definition is meaningful.) An alternative interpretation is that $F^*(x^*)$ is the (negative of the) affine part of the tangent to F (in the point x at which the supremum is attained) with slope x^*, see Figure 5.1. Lemma 3.4 (v) and Lemma 2.3 (v) immediately imply that F^* is always convex and weakly-$*$ lower semicontinuous (as long as F is indeed proper). If F is bounded from below by an affine functional (which is always the case if F is proper, convex, and lower semicontinuous by Lemma 3.5), then F^* is proper as well. Finally, the definition directly yields the *Fenchel–Young inequality*

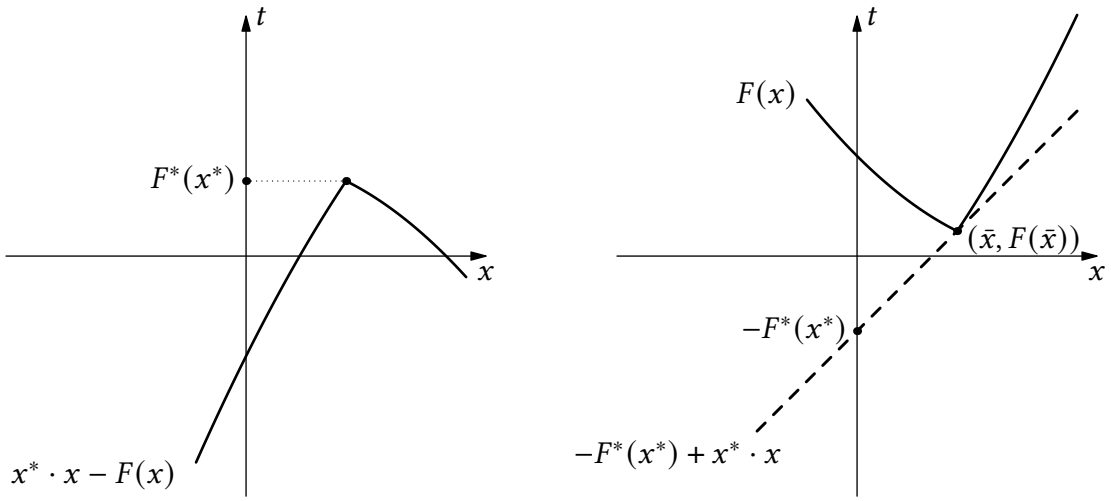
$$\langle x^*, x\rangle_X \le F(x) + F^*(x^*) \qquad \text{for all } x \in X, x^* \in X^*. \tag{5.1}$$

If X is not reflexive, we can similarly define for (weakly-$*$ lower semicontinuous) $F : X^* \to \mathbb{R}$ the *Fenchel preconjugate*

$$F_* : X \to \overline{\mathbb{R}}, \qquad F_*(x) = \sup_{x^* \in X^*} \{\langle x^*, x\rangle_X - F(x^*)\}.$$

The point of this convention is that even in nonreflexive spaces, the *biconjugate*

$$F^{**} : X \to \overline{\mathbb{R}}, \qquad F^{**}(x) = (F^*)_*(x)$$



(a) $F^*(x^*)$ as maximizer of $x^* \cdot x - F(x)$

(b) Alternative interpretation: $-F^*(x^*)$ as offset for tangent to F with given slope x^*. Note that in this case, $x^* \in \partial F(\bar{x})$ and $-F^*(x^*) + x^* \cdot \bar{x} = F(\bar{x})$.

Figure 5.1: Geometrical illustration of the Fenchel conjugate.

is again defined on X (rather than $X^{**} \supset X$). For reflexive spaces, of course, we have $F^{**} = (F^*)^*$. Intuitively, F^{**} is the convex envelope of F, which by Lemma 3.5 coincides with F itself if F is convex.

Theorem 5.1 (Fenchel–Moreau–Rockafellar). *Let $F : X \to \overline{\mathbb{R}}$ be proper. Then,*

(i) $F^{**} \leq F$;

(ii) $F^{**} = F^{\Gamma}$;

(iii) $F^{**} = F$ *if and only if* F *is convex and lower semicontinuous.*

Proof. For (i), we take the supremum over all $x^* \in X^*$ in the Fenchel–Young inequality (5.1) and obtain that

$$F(x) \geq \sup_{x^* \in X^*} \{\langle x^*, x \rangle_X - F^*(x^*)\} = F^{**}(x).$$

For (ii), we first note that F^{**} is convex and lower semicontinuous by definition as a Fenchel conjugate as well as proper by (i). Hence, Lemma 3.5 yields that

$$F^{**}(x) = (F^{**})^{\Gamma}(x) = \sup \{a(x) \mid a : X \to \mathbb{R} \text{ continuous affine with } a \leq F^{**}\}.$$

We now show that we can replace F^{**} with F on the right-hand side. For this, let $a(x) = \langle x^*, x \rangle_X - \alpha$ with arbitrary $x^* \in X^*$ and $\alpha \in \mathbb{R}$. If $a \leq F^{**}$, then (i) implies that $a \leq F$.

Conversely, if $a \leq F$, we have that $\langle x^*, x\rangle_X - F(x) \leq \alpha$ for all $x \in X$, and taking the supremum over all $x \in X$ yields that $\alpha \geq F^*(x^*)$. By definition of F^{**}, we thus obtain that

$$a(x) = \langle x^*, x\rangle_X - \alpha \leq \langle x^*, x\rangle_X - F^*(x^*) \leq F^{**}(x) \quad \text{for all } x \in X,$$

i.e., $a \leq F^{**}$.

Statement (iii) now directly follows from (ii) and Lemma 3.5. □

Remark 5.2. Continuing from Remark 3.6, we can adapt the proof of Theorem 5.1 to proper functionals $F : X^* \to \overline{\mathbb{R}}$ to show that $F = (F_*)^*$ if and only if F is convex and weakly-$*$ lower semicontinuous.

We again consider some relevant examples.

Example 5.3.

(i) Let $\mathbb{B}_X$ be the unit ball in the normed vector space X and take $F = \delta_{\mathbb{B}_X}$. Then we have for any $x^* \in X^*$ that

$$(\delta_{\mathbb{B}_X})^*(x^*) = \sup_{x\in X}\left\{\langle x^*, x\rangle_X - \delta_{\mathbb{B}_X}(x)\right\} = \sup_{\|x\|_X \leq 1}\{\langle x^*, x\rangle_X\} = \|x^*\|_{X^*}.$$

Similarly, one shows using the definition of the Fenchel preconjugate and Corollary 1.7 that $(\delta_{\mathbb{B}_{X^*}})_*(x) = \|x\|_X$.

(ii) Let X be a normed vector space and take $F(x) = \|x\|_X$. We now distinguish two cases for a given $x^* \in X^*$.

Case 1: $\|x^*\|_{X^*} \leq 1$. Then it follows from (1.1) that $\langle x^*, x\rangle_X - \|x\|_X \leq 0$ for all $x \in X$. Furthermore, $\langle x^*, 0\rangle = 0 = \|0\|_X$, which implies that

$$F^*(x^*) = \sup_{x\in X}\{\langle x^*, x\rangle_X - \|x\|_X\} = 0.$$

Case 2: $\|x^*\|_{X^*} > 1$. Then by definition of the dual norm, there exists an $x_0 \in X$ with $\langle x^*, x_0\rangle_X > \|x_0\|_X$. Hence, taking $t \to \infty$ in

$$0 < t(\langle x^*, x_0\rangle_X - \|x_0\|_X) = \langle x^*, tx_0\rangle_X - \|tx_0\|_X \leq F^*(x^*)$$

yields $F^*(x^*) = \infty$.

Together we obtain that $F^* = \delta_{\mathbb{B}_{X^*}}$. As above, a similar argument shows that $(\|\cdot\|_{X^*})_* = \delta_{\mathbb{B}_X}$.

We can generalize Example 5.3 (ii) to powers of norms.

Lemma 5.4. *Let X be a normed vector space and $F(x) := \frac{1}{p}\|x\|_X^p$ for $p \in (1,\infty)$. Then $F^*(x^*) = \frac{1}{q}\|x^*\|_{X^*}^q$ for $q := \frac{p}{p-1}$.*

Proof. We first consider the scalar function $\varphi(t) := \frac{1}{p}|t|^p$ and compute the Fenchel conjugate $\varphi^*(s)$ for $s \in \mathbb{R}$. By the choice of p and q, we then can write $\frac{1}{q} = 1 - \frac{1}{p}$ as well as $|s|^q = \operatorname{sign}(s)s|s|^{1/(p-1)} = |\operatorname{sign}(s)|s|^{1/(p-1)}|^p$ for any $s \in \mathbb{R}$ and therefore obtain

$$\frac{1}{q}|s|^q = \left(\operatorname{sign}(s)|s|^{1/(p-1)}\right)s - \frac{1}{p}\left|\operatorname{sign}(s)|s|^{1/(p-1)}\right|^p \le \sup_{t\in\mathbb{R}}\left\{ts - \frac{1}{p}|t|^p\right\} \le \frac{1}{q}|s|^q,$$

where we have used the classical Young inequality $ts \le \frac{1}{q}|t|^p + \frac{1}{q}|s|^q$ in the last step. This shows that $\varphi^*(s) = \frac{1}{q}|s|^q$.[1]

We now write using the definition of the norm in X^* that

$$\begin{aligned} F^*(x^*) &= \sup_{x\in X}\left\{\langle x^*, x\rangle_X - \frac{1}{p}\|x\|_X^p\right\} = \sup_{t\ge 0}\left\{\sup_{x\in\mathbb{B}_X}\left\{\langle x^*, tx\rangle_X - \frac{1}{p}\|tx\|_X^p\right\}\right\} \\ &= \sup_{t\ge 0}\left\{t\|x^*\|_{X^*} - \frac{1}{p}|t|^p\right\} = \frac{1}{q}\,|\|x^*\|_{X^*}|^q \end{aligned}$$

since φ is even and the supremum over all $t \in \mathbb{R}$ is thus attained for $t \ge 0$. □

As for convex subdifferentials, Fenchel conjugates of integral functionals can be computed pointwise.

Theorem 5.5. *Let $f : \mathbb{R} \to \overline{\mathbb{R}}$ be measurable, proper and lower semicontinuous, and let $F : L^p(\Omega) \to \overline{\mathbb{R}}$ with $1 \le p < \infty$ be defined as in Lemma 3.7. Then we have for $q = \frac{p}{p-1}$ that*

$$F^* : L^q(\Omega) \to \overline{\mathbb{R}}, \qquad F^*(u^*) = \int_\Omega f^*(u^*(x))\,dx.$$

Proof. We argue similarly as in the proof of Theorem 4.11, with some changes that are needed since measurability of $f \circ u$ does not immediately imply that of $f^* \circ u^*$. Let $u^* \in L^q(\Omega)$ be arbitrary and consider for all $x \in \Omega$ the functions

$$\varphi(x) := \sup_{t\in\mathbb{R}}\{tu^*(x) - f(t)\} = f^*(u^*(x)),$$

as well as for $n \in \mathbb{N}$

$$\varphi_n(x) := \sup_{|t|\le n}\{tu^*(x) - f(t)\} \le f^*(u^*(x)).$$

[1] Which is how the Fenchel–Young inequality got its name.

By a measurable selection theorem ([Ekeland and Témam, 1999, Theorem VIII.1.2]), the pointwise supremum in the definition of φ_n is attained at some t_x^* for almost every $x \in \Omega$ and defines a measurable mapping $x \mapsto u_n(x) := t_x^*$ with $\|u_n\|_{L^\infty} \leq n$. This also implies that $\varphi_n = u_n \cdot u^* - f \circ u_n$ is measurable. Furthermore, by assumption there exists a $t_0 \in \operatorname{dom} f$, and hence $u_0 := t_0 u^*(x) - f(t_0)$ is measurable and satisfies $u_0 \leq \varphi_n(x)$ for all $n \geq |t_0|$. Finally, by construction, $\varphi_n(x)$ is monotonically increasing and converges to $\varphi(x)$ for all $x \in \Omega$. The sequence $\{\varphi_n - u_0\}_{n\in\mathbb{N}}$ of functions is thus measurable and nonnegative, and the monotone convergence theorem yields that

$$\int_\Omega \varphi(x) - u_0(x)\,dx = \int_\Omega \sup_{n\in\mathbb{N}} \varphi_n(x) - u_0(x)\,dx = \sup_{n\in\mathbb{N}} \int_\Omega \varphi_n(x) - u_0(x)\,dx.$$

Hence the pointwise limit $\varphi = f^* \circ u^*$ is measurable as well.

The measurable selection theorem also yields that

$$\begin{aligned}\int_\Omega f^*(u^*(x))\,dx &= \sup_{n\in\mathbb{N}} \int_\Omega \sup_{|t|\leq n} \{tu^*(x) - f(t)\}\;dx \\ &= \sup_{n\in\mathbb{N}} \int_\Omega u^*(x)u_n(x) - f(u_n(x))\,dx \\ &\leq \sup_{u\in L^p(\Omega)} \int_\Omega u^*(x)u(x) - f(u(x))\,dx = F^*(u^*),\end{aligned}$$

since $u_n \in L^\infty(\Omega) \subset L^p(\Omega)$ for all $n \in \mathbb{N}$.

For the converse inequality, we can now proceed as in the proof of Theorem 4.11. For any $u \in L^p(\Omega)$ and $u^* \in L^q(\Omega)$, we have by the Fenchel–Young inequality (5.1) applied to f and f^* that

$$f(u(x)) + f^*(u^*(x)) \geq u^*(x)u(x) \quad \text{for almost every } x \in \Omega.$$

Since both sides are measurable, this implies that

$$\int_\Omega f^*(u^*(x))\,dx \geq \int_\Omega u^*(x)u(x) - f(u(x))\,dx,$$

and taking the supremum over all $u \in L^p(\Omega)$ yields the claim. □

Remark 5.6. A similar representation representation can be shown for vector-valued and spatially-dependent integrands $f : \Omega \times \mathbb{R} \to \mathbb{R}^m$ under stronger assumptions; see, e.g., [Rockafellar, 1976a, Corollary 3C].

Fenchel conjugates satisfy a number of useful calculus rules, which follow directly from the properties of the supremum.

Lemma 5.7. *Let $F : X \to \overline{\mathbb{R}}$ be proper. Then,*

(i) $(\alpha F)^* = \alpha F^* \circ (\alpha^{-1}\mathrm{Id})$ *for any* $\alpha > 0$;

(ii) $(F(\cdot + x_0) + \langle x_0^*, \cdot\rangle_X)^* = F^*(\cdot - x_0^*) - \langle \cdot - x_0^*, x_0\rangle_X$ *for all* $x_0 \in X$, $x_0^* \in X^*$;

(iii) $(F \circ K)^* = F^* \circ K^{-*}$ *for continuously invertible* $K \in \mathbb{L}(Y;X)$ *and* $K^{-*} := (K^{-1})^*$.

Proof. *(i):* For any $\alpha > 0$, we have that

$$(\alpha F)^*(x^*) = \sup_{x\in X}\left\{\alpha\langle \alpha^{-1}x^*, x\rangle_X - \alpha F(x)\right\} = \alpha \sup_{x\in X}\left\{\langle \alpha^{-1}x^*, x\rangle_X - F(x)\right\} = \alpha F^*(\alpha^{-1}x^*).$$

(ii): Since $\{x + x_0 \mid x \in X\} = X$, we have that

$$\begin{aligned}(F(\cdot + x_0) + \langle x_0^*, \cdot\rangle_X)^*(x^*) &= \sup_{x\in X}\{\langle x^*, x\rangle_X - F(x + x_0)\} - \langle x_0^*, x\rangle_X\\ &= \sup_{x\in X}\left\{\langle x^* - x_0^*, x + x_0\rangle_X - F(x + x_0)\right\} - \langle x^* - x_0^*, x_0\rangle_X\\ &= \sup_{\tilde{x}=x+x_0, x\in X}\left\{\langle x^* - x_0^*, \tilde{x}\rangle_X - F(\tilde{x})\right\} - \langle x^* - x_0^*, x_0\rangle_X\\ &= F^*(x^* - x_0^*) - \langle x^* - x_0^*, x_0\rangle_X.\end{aligned}$$

(iii): Since $X = \operatorname{ran} K$, we have that

$$\begin{aligned}(F \circ K)^*(y^*) &= \sup_{y\in Y}\left\{\langle y^*, K^{-1}Ky\rangle_Y - F(Ky)\right\}\\ &= \sup_{x=Ky, y\in Y}\{\langle K^{-*}y^*, x\rangle_X - F(x)\} = F^*(K^{-*}y^*).\end{aligned}$$

□

There are some obvious similarities between the definitions of the Fenchel conjugate and of the subdifferential, which yield the following very useful property that plays the role of a "convex inverse function theorem". (See also Figure 5.1b and compare Figures 4.1a and 4.1b.)

Lemma 5.8 (Fenchel–Young). *Let* $F : X \to \overline{\mathbb{R}}$ *be proper, convex, and lower semicontinuous. Then the following statements are equivalent for any* $x \in X$ *and* $x^* \in X^*$*:*

(i) $\langle x^*, x\rangle_X = F(x) + F^*(x^*)$;

(ii) $x^* \in \partial F(x)$;

(iii) $x \in \partial F^*(x^*)$.

Proof. If (i) holds, the definition of F^* as a supremum immediately implies that

$$\langle x^*, x\rangle_X - F(x) = F^*(x^*) \geq \langle x^*, \tilde{x}\rangle_X - F(\tilde{x}) \qquad \text{for all } \tilde{x} \in X, \tag{5.2}$$

which again by definition is equivalent to (ii). Conversely, taking the supremum over all $\tilde{x} \in X$ in (5.2) yields

$$\langle x^*, x\rangle_X \geq F(x) + F^*(x^*),$$

which together with the Fenchel–Young inequality (5.1) leads to (i).

Similarly, (i) in combination with Theorem 5.1 implies that

$$\langle x^*, x\rangle_X - F^*(x^*) = F(x) = F^{**}(x) \geq \langle \tilde{x}^*, x\rangle - F^*(\tilde{x}^*) \qquad \text{for all } \tilde{x}^* \in X^*,$$

yielding as above the equivalence of (i) and (iii). □

Remark 5.9. If F is not convex, the above proof shows that we still have the equivalence (i) $\Leftrightarrow$ (ii). Furthermore since always $F^{**} \leq F$ by Theorem 5.1 (i), it still holds that (i) $\Rightarrow$ (iii). However, we can only conclude from (iii) that (i) and (ii) hold for $F^{**} \neq F$ in place of F. Applying Lemma 5.8 to nonconvex functionals therefore inevitably introduces a *convexification* (by replacing the nonconvex F with its convex envelope F^{**}).

Remark 5.10. Recall that $\partial F^*(x^*) \subset X^{**}$. Therefore, if X is not reflexive, $x \in \partial F^*(x^*)$ in (iii) has to be understood via the canonical injection $J : X \hookrightarrow X^{**}$ as $Jx \in \partial F^*(x^*)$, i.e., as

$$\langle Jx, \tilde{x}^* - x^*\rangle_{X^*} = \langle \tilde{x}^* - x^*, x\rangle_X \leq F^*(\tilde{x}^*) - F^*(x^*) \quad \text{for all } \tilde{x}^* \in X.$$

Using (iii) to conclude equality in (i) or, equivalently, the subdifferential inclusion (ii) therefore requires the additional condition that $x \in X \hookrightarrow X^{**}$. Conversely, if (i) or (ii) hold, (iii) also guarantees that the subderivative x is an element of $\partial F^*(x^*) \cap X$, which is a stronger claim (see [gerw, 2022] for a counterexample).

Similar statements apply to (weakly-$*$ lower semicontinuous) $F : X^* \to \overline{\mathbb{R}}$ and $F_* : X \to \overline{\mathbb{R}}$.

5.2 DUALITY OF OPTIMIZATION PROBLEMS

Lemma 5.8 can be used to replace the subdifferential of a (complicated) norm with that of a (simpler) conjugate indicator functional (or vice versa). For example, given a problem of the form

$$\inf_{x \in X} F(x) + G(Kx) \tag{5.3}$$

for $F : X \to \overline{\mathbb{R}}$ and $G : Y \to \overline{\mathbb{R}}$ proper, convex, and lower semicontinuous, and $K \in \mathbb{L}(X; Y)$, we can use Theorem 5.1 to replace G with the definition of G^{**} and obtain the *saddle-point problem*

$$\inf_{x \in X} \sup_{y^* \in Y^*} F(x) + \langle y^*, Kx\rangle_Y - G^*(y^*). \tag{5.4}$$

If(!) we were now able to exchange inf and sup, we could write (with $\inf F = -\sup(-F)$)

$$\begin{aligned}\inf_{x\in X}\sup_{y^*\in Y^*} F(x) + \langle y^*, Kx\rangle_Y - G^*(y^*) &= \sup_{y^*\in Y^*}\inf_{x\in X} F(x) + \langle y^*, Kx\rangle_Y - G^*(y^*)\\ &= \sup_{y^*\in Y^*} -\left\{\sup_{x\in X} -F(x) + \langle -K^*y^*, x\rangle_X\right\} - G^*(y^*).\end{aligned}$$

From the definition of F^*, we thus obtain the *dual problem*

$$\sup_{y^*\in Y^*} -G^*(y^*) - F^*(-K^*y^*). \tag{5.5}$$

As a side effect, we have shifted the operator K from G to F^* without having to invert it.

The following theorem uses in an elegant way the Fermat principle, the sum and chain rules, and the Fenchel–Young equality to derive sufficient conditions for the exchangeability.

Theorem 5.11 (Fenchel–Rockafellar). *Let X and Y be Banach spaces, $F : X \to \overline{\mathbb{R}}$ and $G : Y \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $K \in \mathbb{L}(X;Y)$. Assume furthermore that*

(i) the primal problem (5.3) *admits a solution $\bar{x} \in X$;*

(ii) there exists an $x_0 \in \operatorname{dom}(G\circ K) \cap \operatorname{dom} F$ with $Kx_0 \in \operatorname{int}(\operatorname{dom} G)$.

Then the dual problem (5.5) admits a solution $\bar{y}^* \in Y^*$ and

$$\min_{x\in X} F(x) + G(Kx) = \max_{y^*\in Y^*} -G^*(y^*) - F^*(-K^*y^*). \tag{5.6}$$

Furthermore, $\bar{x}$ and $\bar{y}^*$ are solutions to (5.3) and (5.5), respectively, if and only if

$$\begin{cases} \bar{y}^* \in \partial G(K\bar{x}),\\ -K^*\bar{y}^* \in \partial F(\bar{x}).\end{cases} \tag{5.7}$$

Proof. Let first $\bar{x} \in X$ be a solution to (5.3). By assumption (ii), Theorems 4.14 and 4.17 are applicable; Theorem 4.2 thus implies that

$$0 \in \partial(F + G\circ K)(\bar{x}) = K^*\partial G(K\bar{x}) + \partial F(\bar{x})$$

and thus the existence of a $\bar{y}^* \in \partial G(K\bar{x})$ with $-K^*\bar{y}^* \in \partial F(\bar{x})$, i.e., satisfying (5.7).

Conversely, let (5.7) hold for $\bar{x} \in X$ and $\bar{y}^* \in Y^*$. Then again by Theorems 4.2, 4.14, and 4.17, $\bar{x}$ is a solution to (5.3). Furthermore, (5.7) together with Lemma 5.8 imply equality in the Fenchel–Young inequalities for F and G, i.e.,

$$\begin{cases} \langle \bar{y}^*, K\bar{x}\rangle_Y = G(K\bar{x}) + G^*(\bar{y}^*),\\ \langle -K^*\bar{y}^*, \bar{x}\rangle_X = F(\bar{x}) + F^*(-K^*\bar{y}^*).\end{cases} \tag{5.8}$$

Adding both equations and rearranging now yields

$$F(\bar{x}) + G(K\bar{x}) = -F^*(-K^*\bar{y}^*) - G^*(\bar{y}^*). \tag{5.9}$$

It remains to show that $\bar{y}^*$ is a solution to (5.5). For this purpose, we introduce

$$L : X \times Y^* \to \overline{\mathbb{R}}, \qquad L(x, y^*) = F(x) + \langle y^*, Kx\rangle_Y - G^*(y^*). \tag{5.10}$$

For all $\tilde{x} \in X$ and $\tilde{y}^* \in Y^*$, we always have that

$$\sup_{y^* \in Y^*} L(\tilde{x}, y^*) \geq L(\tilde{x}, \tilde{y}^*) \geq \inf_{x \in X} L(x, \tilde{y}^*), \tag{5.11}$$

and hence (taking the infimum over all $\tilde{x}$ in the first and the supremum over all $\tilde{y}^*$ in the second inequality) that

$$\inf_{x \in X} \sup_{y^* \in Y^*} L(x, y^*) \geq \sup_{y^* \in Y^*} \inf_{x \in X} L(x, y^*). \tag{5.12}$$

We thus obtain that

$$\begin{aligned} F(\bar{x}) + G(K\bar{x}) &= \inf_{x \in X} \sup_{y^* \in Y^*} F(x) + \langle y^*, Kx\rangle_Y - G^*(y^*) \\ &\geq \sup_{y^* \in Y^*} \inf_{x \in X} F(x) + \langle y^*, Kx\rangle_Y - G^*(y^*) \\ &= \sup_{y^* \in Y^*} -G^*(y^*) - F^*(-K^*y^*) \end{aligned} \tag{5.13}$$

(i.e., *weak duality* holds merely under assumption (i)). Combining this with (5.9) yields that

$$-G^*(\bar{y}^*) - F^*(-K^*\bar{y}^*) = F(\bar{x}) + G(K\bar{x}) \geq \sup_{y^* \in Y^*} -G^*(y^*) - F^*(-K^*y^*),$$

i.e., $\bar{y}^*$ is a solution to (5.5), which in particular shows the claimed existence of a solution.

Since all solutions to (5.5) have by definition the same (maximal) functional value, (5.9) also implies (5.6).

Finally, if $\bar{x} \in X$ and $\bar{y}^* \in Y^*$ are solutions to (5.3) and (5.5), respectively, the just derived strong duality (5.6) conversely implies that (5.9) holds. Together with the productive zero, we obtain from this that

$$0 = \left[G(K\bar{x}) + G^*(\bar{y}^*) - \langle \bar{y}^*, K\bar{x}\rangle_X\right] + \left[F(\bar{x}) + F^*(-K^*\bar{y}^*) - \langle -K^*\bar{y}^*, \bar{x}\rangle_Y\right].$$

Since both brackets have to be nonnegative due to the Fenchel–Young inequality, they each have to be zero. We therefore deduce that (5.8) holds, and hence Lemma 5.8 implies (5.7). □

Remark 5.12. If X is the dual of a separable Banach space X_*, it is possible to derive a similar duality result with the (weakly-$*$ lower semicontinuous) preconjugate $F_* : X_* \to \overline{\mathbb{R}}$ in place of $F^* : X^* \to \overline{\mathbb{R}}$ under the additional assumption that $\operatorname{ran} K^* \subset X_* \subsetneq X^*$ (using Remark 5.10 in (5.8)). If X_* is a "nicer" space than X^* (e.g., for $X = \mathcal{M}(\Omega)$, the space of bounded Radon measures on a domain Ω with $X_* = C_0(\Omega)$, the space of continuous functions with compact support), the *predual problem*

$$\sup_{y^* \in Y^*} -G^*(y^*) - F_*(-K^* y^*)$$

may be easier to treat than the dual problem (5.5). This is the basis of the "preduality trick" used in, e.g., [Clason and Kunisch, 2011; Hintermüller and Kunisch, 2004].

Remark 5.13. The condition (ii) was only used to guarantee equality in the sum and chain rules Theorems 4.14 and 4.17 applied to $F + G \circ K$. Since these rules hold under the weaker condition of Remark 4.16 (recall that the chain rule was proved by reduction to the sum rule), Theorem 5.11 and Corollary 5.14 hold under this weaker condition as well.

The relations (5.7) are referred to as *Fenchel extremality conditions*; we can use Lemma 5.8 to generate further, equivalent, optimality conditions by inverting one or the other subdifferential inclusion. We will later exploit this to derive implementable algorithms for solving optimization problems of the form (5.3). Furthermore, Theorem 5.11 characterizes the subderivative $\bar{y}^*$ produced by the sum and chain rules as solution to a convex minimization problem, which may be useful. For example, if either F^* or G^* is strongly convex, this subderivative will be unique, which has beneficial consequences for the stability and the convergence of algorithms for the computation of solutions to (5.7).

For their analysis, it will sometimes be more convenient to apply the consequences of Theorem 5.11 in the form of the saddle-point problem (5.4). For a general mapping $L : X \times Y^* \to \overline{\mathbb{R}}$, we call $(\tilde{x}, \tilde{y}^*)$ a *saddle point* of L if

$$\sup_{y^* \in Y^*} L(\tilde{x}, y^*) \leq L(\tilde{x}, \tilde{y}^*) \leq \inf_{x \in X} L(x, \tilde{y}^*). \tag{5.14}$$

(Note that the opposite inequality (5.11) always holds.)

Corollary 5.14. *Assume that the conditions of Theorem 5.11 hold. Then there exists a saddle point $(\bar{x}, \bar{y}^*) \in X \times Y^*$ to*

$$L(x, y^*) := F(x) + \langle y^*, Kx \rangle_Y - G^*(y^*).$$

Furthermore, for any $(x, y^*) \in X \times Y^*$,

$$\begin{aligned} F(\bar{x}) + \langle y^*, K\bar{x} \rangle_Y - G^*(y^*) &\leq F(\bar{x}) + \langle \bar{y}^*, K\bar{x} \rangle_Y - G^*(\bar{y}^*) \\ &\leq F(x) + \langle \bar{y}^*, Kx \rangle_Y - G^*(\bar{y}^*). \end{aligned} \tag{5.15}$$

Proof. Both statements follow from the fact that under the assumption, the inequality in (5.13) and hence in (5.14) holds as an equality. □

With the notation $u = (x, y)$, let us define the *duality gap*

$$\bar{\mathcal{G}}(u) \coloneqq F(x) + G(Kx) + G^*(y^*) + F^*(-K^*y^*). \tag{5.16}$$

By Theorem 5.11, we have $\bar{\mathcal{G}} \geq 0$ and $\bar{\mathcal{G}}(\bar{u}) = 0$ if and only if $\bar{u}$ is a saddle point.

On the other hand, for any saddle point $\bar{u} = (\bar{x}, \bar{y}^*)$ of a Lagrangian $L : X \times Y^* \to \overline{\mathbb{R}}$, we can also define the *Lagrangian duality gap*

$$\mathcal{G}_L(u; \bar{u}) \coloneqq L(x, \bar{y}^*) - L(\bar{x}, y^*).$$

For L defined in (5.10), we always have by the definition of the convex conjugate that

$$0 \leq \mathcal{G}_L(u; \bar{u}) \leq \bar{\mathcal{G}}(u).$$

However, $\mathcal{G}_L(u; \bar{u}) = 0$ does not necessarily imply that u is a saddle point. (This is the case if L is strictly convex in x or strictly concave in y, i.e., if either F or G^* is strictly convex.) Nevertheless, as we will see in later chapters, the *Lagrangian duality gap* can generally be shown to converge for iterates produced by optimization algorithms, while this is more difficult for the conventional duality gap.

6 MONOTONE OPERATORS AND PROXIMAL POINTS

Any minimizer $\bar{x} \in X$ of a convex functional $F : X \to \overline{\mathbb{R}}$ satisfies by Theorem 4.2 the Fermat principle $0 \in \partial F(\bar{x})$. To use this to characterize $\bar{x}$, and, later, to derive implementable algorithms for its iterative computation, we now study the mapping $x \mapsto \partial F(x)$ in more detail.

6.1 BASIC PROPERTIES OF SET-VALUED MAPPINGS

We start with some basic concepts. For two normed vector spaces X and Y we consider a *set-valued mapping* $A : X \to \mathcal{P}(Y)$, also denoted by $A : X \rightrightarrows Y$, and define

- its *domain of definition* $\operatorname{dom} A = \{x \in X \mid A(x) \neq \emptyset\}$;
- its *range* $\operatorname{ran} A = \bigcup_{x \in X} A(x)$;
- its *graph* $\operatorname{graph} A = \{(x, y) \in X \times Y \mid y \in A(x)\}$;
- its *inverse* $A^{-1} : Y \rightrightarrows X$ via $A^{-1}(y) = \{x \in X \mid y \in A(x)\}$ for all $y \in Y$.

(Note that $A^{-1}(y) = \emptyset$ is allowed by the definition; hence for set-valued mappings, the inverse and always exists.) Similarly, we will say that $A : X \rightrightarrows Y$ is *surjective* if $\operatorname{ran} A = Y$.

For $A, B : X \rightrightarrows Y$, $C : Y \rightrightarrows Z$, and $\lambda \in \mathbb{R}$ we further define

- $\lambda A : X \rightrightarrows Y$ via $(\lambda A)(x) = \{\lambda y \mid y \in A(x)\}$;
- $A + B : X \rightrightarrows Y$ via $(A + B)(x) = \{y + z \mid y \in A(x), z \in B(x)\}$;
- $C \circ A : X \rightrightarrows Z$ via $(C \circ A)(x) = \{z \mid \text{there is } y \in A(x) \text{ with } z \in C(y)\}$.

Of particular importance not only in the following but also in Part IV is the continuity of set-valued mappings. We first introduce notions of convergence of sets. So let $\{X_n\}_{n \in \mathbb{N}}$ be a sequence of subsets of X. We define

(i) the *outer limit* as the set

$$\limsup_{n \to \infty} X_n := \left\{ x \in X \;\middle|\; \text{there exists } \{n_k\}_{k \in \mathbb{N}} \text{ with } x_{n_k} \in X_{n_k} \text{ and } \lim_{k \to \infty} x_{n_k} = x \right\},$$

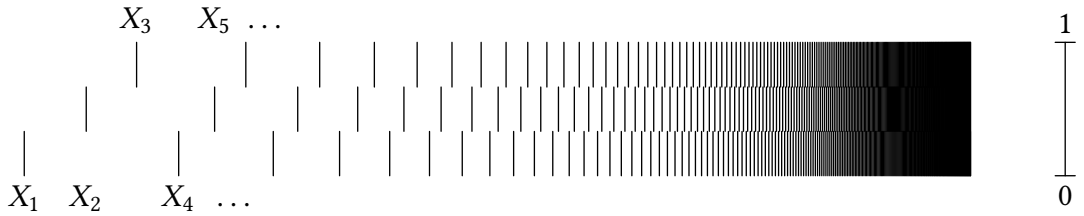


Figure 6.1: Illustration of Example 6.1 with $\limsup_{n\to\infty} X_n = [0,1]$ while $\liminf_{n\to\infty} X_n = \emptyset$.

(ii) the *inner limit* as the set

$$\liminf_{n\to\infty} X_n := \left\{ x \in X \;\middle|\; \text{there exist } x_n \in X_n \text{ with } \lim_{n\to\infty} x_n = x \right\}.$$

Correspondingly, we define the *weak outer limit* and the *weak inner limit*, denoted by $\text{w-lim sup}_{n\to\infty} X_n$ and $\text{w-lim inf}_{n\to\infty} X_n$, respectively, using weakly converging (sub)sequences. Similarly, for a dual space X^*, we define the *weak-$*$ outer limit* $\text{w-}*\text{-lim sup}_{n\to\infty} X_n^*$ and the *weak-$*$ inner limit* $\text{w-}*\text{-lim inf}_{n\to\infty} X_n^*$.

The outer limit consists of all points approximable through *some* subsequence of the sets X_n, while the inner limit has to be approximable through *every* subsequence. The vast difference between inner and outer limits is illustrated by the following extreme example.

Example 6.1. Let $X = \mathbb{R}$ and $\{X_n\}_{n\in\mathbb{N}}$, $X_n \subset [0,1]$, be given as

$$X_n := \begin{cases} [0, \frac{1}{3}) & \text{if } n = 3k-2 \text{ for some } k \in \mathbb{N}, \\ [\frac{1}{3}, \frac{2}{3}) & \text{if } n = 3k-1 \text{ for some } k \in \mathbb{N}, \\ [\frac{2}{3}, 1] & \text{if } n = 3k \text{ for some } k \in \mathbb{N}, \end{cases}$$

see Figure 6.1. Then,

$$\limsup_{n\to\infty} X_n = [0,1],$$

since for any $x \in [0,1]$, we can find a subsequence of $\{X_n\}_{n\in\mathbb{N}}$ (by selecting subsequences with, e.g., $n = 3k-2$ for $k \in \mathbb{N}$ if $x < \frac{1}{3}$) that contain x. On the other hand,

$$\liminf_{n\to\infty} X_n = \emptyset,$$

since for any $x \in [0,1]$, there will be a subsequence of X_n (again, selecting only subsequences with, e.g., $n = 3k$ for $k \in \mathbb{N}$ if $x < \frac{1}{3}$) that will not contain points arbitrarily close to x.

Lemma 6.2. *Let $\{X_n\}_{n\in\mathbb{N}}$, $X_n \subset X$. Then $\limsup_{n\to\infty} X_n$ and $\liminf_{n\to\infty} X_n$ are (possibly empty) closed sets.*

Proof. Let $X_\infty := \limsup_{n\to\infty} X_n$. If X_∞ is empty, there is nothing to prove. So suppose, $\{x_k\}_{k\in\mathbb{N}} \subset X_\infty$ converges to some $\bar{y} \in X$. Then by the definition of X_∞ as an outer limit, there exist infinite subsets $N_k \subset \mathbb{N}$ and subsequences $x_{k,n} \in X_n$ for $n \in N_k$ with $\lim_{N_k \ni n\to\infty} x_{k,n} = x_k$. We can find for each $k \in \mathbb{N}$ an index $n_k \in N_k$ such that $\|x_k - x_{k,n_k}\|_X \leq 1/n$. Thus $\|\bar{y} - x_{k,n_k}\|_X \leq \|\bar{y} - x_k\|_X + 1/k$. Letting $k \to \infty$ we see that $X_{n_k} \ni x_{k,n_k} \to \bar{y}$. Thus $\bar{y} \in X_\infty$, that is, X_∞ is (strongly) closed.

Let then $X_\infty := \liminf_{n\to\infty} X_n$. If X_∞ is empty, there is nothing to prove. So suppose $\{x_k\}_{k\in\mathbb{N}} \subset X_\infty$ converges to some $\bar{y} \in X$. Then for each $n \in \mathbb{N}$ there exist $x_{k,n} \in X_n$ with $\lim_{n\to\infty} x_{k,n} = x_k$. We can consequently find for each $k \in \mathbb{N}$ an index $n_k \in \mathbb{N}$ such that $\|x_k - x_{k,n}\|_X < 1/k$ for $n \geq n_k$. Thus for every $n \in \mathbb{N}$ we can find $k_n \in \mathbb{N}$ such that $\|x_{k_n} - x_{k_n,n}\|_X \leq 1/k_n$ with $k_n \to \infty$ as $n \to \infty$. Since this implies $\|\bar{y} - x_{k_n,n}\|_X \leq \|\bar{y} - x_{k_n}\|_X + 1/k_n$, letting $n \to \infty$ we see that $X_n \ni x_{k_n,n} \to \bar{y}$. Thus $\bar{y} \in X_\infty$, that is, X_∞ is (strongly) closed. □

With these definitions, we can define limits and continuity of set-valued mappings. Specifically, for $A : X \rightrightarrows Y$, and a subset $C \subset X$, we define the inner and outer limits (relative to C, if $C \neq X$) as

$$\limsup_{C\ni\tilde{x}\to x} A(\tilde{x}) := \bigcup_{C\ni x_n\to x} \limsup_{n\to\infty} A(x_n),$$

and

$$\liminf_{C\ni\tilde{x}\to x} A(\tilde{x}) := \bigcap_{C\ni x_n\to x} \liminf_{n\to\infty} A(x_n).$$

If $C = X$, we drop C from the notations. Analogously, we define weak-to-strong, strong-to-weak, and weak-to-weak limits by replacing $x_n \to x$ by $x_n \rightharpoonup x$ and/or the outer/inner limit by the weak outer/inner limit.

Corollary 6.3. *Let $A : X \rightrightarrows Y$ and $x \in X$. Then $\limsup_{\tilde{x}\to x} A(\tilde{x})$ and $\liminf_{\tilde{x}\to x} A(\tilde{x})$ are (possibly empty) closed sets.*

Proof. The proof of the closedness of the outer limit is analogous to Lemma 6.2, while the proof of the closedness of the inner limit is a consequence of Lemma 6.2 and of the fact that the intersections of closed sets are closed. □

Let then $A : X \rightrightarrows Y$ be a set-valued mapping. We say that

(i) A is *outer semicontinuous* at x if $\limsup_{C\ni\tilde{x}\to x} A(\tilde{x}) \subset A(x)$ with $C = X$.

(ii) A is *inner semicontinuous* at x if $\liminf_{C\ni\tilde{x}\to x} A(\tilde{x}) \supset A(x)$ with $C = X$.

(iii) The map A is outer/inner semicontinuous if it is outer/inner semicontinuous at all $x \in X$.

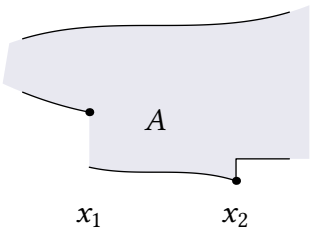


Figure 6.2: Illustration of outer and inner semicontinuity. The black line indicates the bounds on the boundary of graph F that belong to the graph. The set-valued mapping A is not outer semicontinuous at x_1, because $A(x_1)$ does not include all limits from the right. It is outer semicontinuous at the "discontinuous" point x_2, as $A(x_2)$ includes all limits from both sides. The mapping A is not inner semicontinuous at x_2, because at this point, $A(x)$ cannot be approximated from both sides. It is inner semicontinuous at every other point x, including x_1, as at this points $A(x)$ can be approximated from both sides.

(iv) *continuous* (at x) if it is both inner and outer semicontinuous (at x).

(v) We say that these properties are "relative C" when we restrict $\tilde{x} \in C$ for some $C \subset X$.

These concepts are illustrated in Figure 6.2.

Just like lower semicontinuity of functionals, the outer semicontinuity of set-valued mappings can be interpreted as a closedness property and will be crucial. The following lemma is stated for strong-to-strong outer semicontinuity, but corresponding statements hold (with identical proof) for weak-to-strong, strong-to-weak, and weak-to-weak outer semicontinuity as well.

Lemma 6.4. *A set-valued mapping $A : X \rightrightarrows Y$ is outer semicontinuous if and only if* $\operatorname{graph} A \subset X \times Y$ *is closed, i.e., $x_n \to x$ and $A(x_n) \ni y_n \to y$ imply that $y \in A(x)$.*

Proof. Let $x_n \to x$ and $y_n \in A(x_n)$, and suppose also $y_n \to y$. Then if $\operatorname{graph} A$ is closed, $(x, y) \in \operatorname{graph} A$ and hence $y \in A(x)$. Since this holds for arbitrary sequences $\{x_n\}_{n\in\mathbb{N}}$, A is outer semicontinuous.

If, on the other hand, A is outer semicontinuous, and $(x_n, y_n) \in \operatorname{graph} A$ converge to $(x, y) \in X \times Y$, then $y \in A(x)$ and hence $(x, y) \in \operatorname{graph} A$. Since this holds for arbitrary sequences $\{(x_n, y_n)\}_{n\in\mathbb{N}}$, $\operatorname{graph} A$ is closed. □

6.2 MONOTONE OPERATORS

For the codomain $Y = X^*$ (as in the case of $x \mapsto \partial F(x)$), additional properties become important. A set-valued mapping $A : X \rightrightarrows X^*$ is called *monotone* if

$$\langle x_1^* - x_2^*, x_1 - x_2 \rangle_X \geq 0 \quad \text{for all} \quad (x_1, x_1^*), (x_2, x_2^*) \in \operatorname{graph} A. \tag{6.1}$$

Straight from the definition, we obtain the monotonicity of the following mappings.

Example 6.5. (i) If $A : X \rightrightarrows X^*$ is monotone and $\lambda \geq 0$, then λA is monotone as well.

(ii) If $A, B : X \rightrightarrows X^*$ are monotone, then $A + B$ is monotone as well.

(iii) If $F : X \to \overline{\mathbb{R}}$ is proper, then $\partial F : X \rightrightarrows X^*$, $x \mapsto \partial F(x)$, is monotone since for any $x_1, x_2 \in X$ with $x_1^* \in \partial F(x_1)$ and $x_2^* \in \partial F(x_2)$, we have by definition that

$$\begin{aligned} \langle x_1^*, \tilde{x} - x_1 \rangle_X &\leq F(\tilde{x}) - F(x_1) \qquad \text{for all} \quad \tilde{x} \in X, \\ \langle x_2^*, \tilde{x} - x_2 \rangle_X &\leq F(\tilde{x}) - F(x_2) \qquad \text{for all} \quad \tilde{x} \in X. \end{aligned}$$

Adding the first inequality for $\tilde{x} = x_2$ and the second for $\tilde{x} = x_1$ and rearranging the result yields (6.1).

(Example 6.5 (iii) generalizes the well-known fact that if $f : \mathbb{R} \to \mathbb{R}$ is convex and differentiable, its derivative f' is monotonically increasing.)

In fact, we will need the following, stronger, property, which guarantees that A is outer semicontinuous: A monotone operator $A : X \rightrightarrows X^*$ is called *maximally monotone* if there does not exist another monotone operator $\tilde{A} : X \rightrightarrows X^*$ such that $\operatorname{graph} A \subsetneq \operatorname{graph} \tilde{A}$. In other words, A is maximal monotone if for any $x \in X$ and $x^* \in X^*$ the condition

$$\langle x^* - \tilde{x}^*, x - \tilde{x} \rangle_X \geq 0 \qquad \text{for all } (\tilde{x}, \tilde{x}^*) \in \operatorname{graph} A \tag{6.2}$$

implies that $x^* \in A(x)$. (In other words, (6.2) holds if *and only if* $(x, x^*) \in \operatorname{graph} A$.) For fixed $x \in X$ and $x^* \in X^*$, the condition claims that if A is monotone, then so is the extension

$$\tilde{A} : X \rightrightarrows X^*, \qquad \tilde{x} \mapsto \begin{cases} A(x) \cup \{x^*\} & \text{if } \tilde{x} = x, \\ A(\tilde{x}) & \text{if } \tilde{x} \neq x. \end{cases}$$

For A to be maximally monotone means that this is not a true extension, i.e., $\tilde{A} = A$.

Example 6.6. The operator

$$A : \mathbb{R} \rightrightarrows \mathbb{R}, \qquad t \mapsto \begin{cases} \{1\} & \text{if } t > 0, \\ \{0\} & \text{if } t = 0, \\ \{-1\} & \text{if } t < 0, \end{cases}$$

is monotone but not maximally monotone, since A is a proper subset of the monotone operator defined by $\tilde{A}(t) = \operatorname{sign}(t) = \partial(|\cdot|)(t)$ from Example 4.7.

Several useful properties follow directly from the definition.

Lemma 6.7. *If $A : X \rightrightarrows X^*$ is maximally monotone, then so is λA for all $\lambda > 0$.*

Proof. Let $x \in X$ and $x^* \in X^*$, and assume that

$$0 \leq \langle x^* - \tilde{x}^*, x - \tilde{x}\rangle_X = \lambda\langle \lambda^{-1}x^* - \lambda^{-1}\tilde{x}^*, x - \tilde{x}\rangle_X \quad \text{for all } (\tilde{x}, \tilde{x}^*) \in \operatorname{graph} \lambda A.$$

Since $\tilde{x}^* \in \lambda A(\tilde{x})$ if and only if $\lambda^{-1}\tilde{x}^* \in A(\tilde{x})$ and A is maximally monotone, this implies that $\lambda^{-1}x^* \in A(x)$, i.e., $x^* \in (\lambda A)(x)$. Hence, λA is maximally monotone. □

Lemma 6.8. *If $A : X \rightrightarrows X^*$ is maximally monotone, then $A(x)$ is convex and closed for all $x \in X$.*

Proof. Closedness follows from Lemma 6.10. Assume then that $A(x)$ is not convex, i.e., $x^*_\lambda := \lambda x^* + (1-\lambda)\tilde{x}^* \notin A(x)$ for some $x^*, \tilde{x}^* \in A(x)$ and $\lambda \in (0,1)$. We then show that A is not maximal. To see this, we define $\tilde{A}$ via

$$\tilde{A}(y) := \begin{cases} A(y) & y \neq x, \\ A(x) \cup \{x^*_\lambda\}, & y = x, \end{cases}$$

and show that $\tilde{A}$ is monotone. By the definition of $\tilde{A}$, it suffices to show for all $y \in X$ and $y^* \in A(y)$ that

$$\langle x^*_\lambda - y^*, x - y\rangle_X \geq 0.$$

But this follows directly from the definition of x^*_λ and the monotonicity of A. □

Lemma 6.9. *Let X be a reflexive Banach space. If $A : X \rightrightarrows X^*$ is maximally monotone, then so is $A^{-1} : X^* \rightrightarrows X^{**} \simeq X$.*

Proof. First, recall that the inverse $A^{-1} : X^* \rightrightarrows X$ always exists as a set-valued mapping and can be identified with a set-valued mapping from X^* to X^{**} with the aid of the canonical injection $J : X \to X^{**}$ from (1.2), i.e.,

$$A^{-1}(x^*) := \{Jx \in X^{**} \mid x^* \in A(x)\} \qquad \text{for all } x^* \in X^*$$

From this and the definition (1.2), it is clear that A^{-1} is monotone if and only if A is.

Let now $x^* \in X^*$ and $x^{**} \in X^{**}$ be given, and assume that

$$\langle x^{**} - \tilde{x}^{**}, x^* - \tilde{x}^*\rangle_{X^*} \geq 0 \qquad \text{for all } (\tilde{x}^*, \tilde{x}^{**}) \in \operatorname{graph} A^{-1}. \tag{6.3}$$

Since X is reflexive, J is surjective such that there exists an $x \in X$ with $x^{**} = Jx$. Similarly, we can write $\tilde{x}^{**} = J\tilde{x}$ for some $\tilde{x} \in X$ with $\tilde{x}^* \in A(\tilde{x})$. By definition of the duality pairing, (6.3) is thus equivalent to

$$\langle x^* - \tilde{x}^*, x - \tilde{x} \rangle_X \geq 0$$

for all $\tilde{x} \in X$ and $\tilde{x}^* \in A(\tilde{x})$. But since A is maximally monotone, this implies that $x^* \in A(x)$ and hence $x^{**} = Jx \in A^{-1}(x)$. □

We now come to the outer semicontinuity.

Lemma 6.10. *Let $A : X \rightrightarrows X^*$ be maximally monotone. Then A is both weak-to-strong and strong-to-weak-$*$ outer semicontinuous.*

Proof. Let $x \in X$ and $x^* \in X^*$ and consider sequences $\{x_n\}_{n\in\mathbb{N}} \subset X$ with $x_n \rightharpoonup x$ and $\{x_n^*\}_{n\in\mathbb{N}} \subset X^*$ with $x_n^* \in A(x_n)$ and $x_n^* \to x^*$ (or $x_n \to x$ and $x_n^* \overset{*}{\rightharpoonup} x^*$). For arbitrary $\tilde{x} \in X$ and $\tilde{x}^* \in A(\tilde{x})$, the monotonicity of A implies that

$$0 \leq \langle x_n^* - \tilde{x}^*, x_n - \tilde{x} \rangle_X \to \langle x^* - \tilde{x}^*, x - \tilde{x} \rangle_X$$

since the duality pairing of strongly and weakly (or weakly-$*$ and strongly) converging sequences is convergent. Since A is maximally monotone, we obtain that $x^* \in A(x)$ and hence A is weak-to-strong (or strong-to-weakly-$*$) outer semicontinuous by Lemma 6.4. □

Since the pairing of weakly and weakly-$*$ convergent sequences does not converge in general, weak-to-weak-$*$ outer semicontinuity requires additional assumptions on the two sequences. Although we will not need to make use of it, the following notion can prove useful in other contexts. We call a set-valued mapping $A : X \rightrightarrows X^*$ *BCP outer semicontinuous* (for *Brezis–Crandall–Pazy*), if for any sequences $\{x_n\}_{n\in\mathbb{N}} \subset X$ and $\{x_n^*\}_{n\in\mathbb{N}} \subset X^*$ with

(i) $x_n \rightharpoonup x$ and $A(x_n) \ni x_n^* \overset{*}{\rightharpoonup} x^*$,

(ii) $\limsup_{n\to\infty} \langle x_n^* - x^*, x_n - x \rangle_X \leq 0$,

we have $x^* \in A(x)$. The following result from [Brezis et al., 1970, Lemma 1.2] (hence the name) shows that maximally monotone operators are BCP outer semicontinuous.

Lemma 6.11. *Let X be a Banach space and let $A : X \rightrightarrows X^*$ be maximally monotone. Then A is BCP outer semicontinuous.*

Proof. First, the monotonicity of A and assumption (ii) imply that

$$0 \leq \liminf_{n\to\infty} \langle x_n^* - x^*, x_n - x \rangle_X \leq \limsup_{n\to\infty} \langle x_n^* - x^*, x_n - x \rangle_X \leq 0. \tag{6.4}$$

Furthermore, from assumption (i) and the fact that X is a Banach space, it follows that $\{x_n\}_{n\in\mathbb{N}}$ and $\{x_n^*\}_{n\in\mathbb{N}}$ and hence also $\{\langle x_n^*, x_n\rangle_X\}_{n\in\mathbb{N}}$ are bounded. Thus there exists a subsequence such that $\langle x_{n_k}^*, x_{n_k}\rangle_X \to L$ for some $L \in \mathbb{R}$. Passing to the limit, and using (6.4), we obtain that

$$\begin{aligned} 0 &= \lim_{k\to\infty} \langle x_{n_k}^* - x^*, x_{n_k} - x\rangle_X \\ &= \lim_{k\to\infty} \langle x_{n_k}^*, x_{n_k}\rangle_X - \lim_{k\to\infty} \langle x_{n_k}^*, x\rangle_X - \lim_{k\to\infty} \langle x^*, x_{n_k}\rangle_X + \langle x^*, x\rangle_X \\ &= L - \langle x^*, x\rangle_X. \end{aligned}$$

Since the limit does not depend on the subsequence, we have that $\langle x_n^*, x_n\rangle_X \to \langle x^*, x\rangle_X$.

Let now $\tilde{x} \in X$ and $\tilde{x}^* \in A(\tilde{x})$ be arbitrary. Using again the monotonicity of A and assumption (i) together with the first claim yields

$$\begin{aligned} 0 &\le \liminf_{n\to\infty} \langle x_n^* - \tilde{x}^*, x_n - \tilde{x}\rangle_X \\ &\le \lim_{n\to\infty} \langle x_n^*, x_n\rangle_X - \lim_{n\to\infty} \langle x_n^*, \tilde{x}\rangle_X - \lim_{n\to\infty} \langle \tilde{x}^*, x_n\rangle_X + \langle \tilde{x}^*, \tilde{x}\rangle_X \\ &= \langle x^* - \tilde{x}^*, x - \tilde{x}\rangle_X \end{aligned}$$

and hence that $x^* \in A(x)$ by the maximal monotonicity of A. □

The usefulness of BCP outer semicontinuity arises from the fact that it also implies weak-to-strong outer semicontinuity under slightly weaker conditions on A.

Lemma 6.12. *Suppose $A : X \rightrightarrows X^*$ is monotone (but not necessarily maximally monotone) and BCP outer semicontinuous. Then A is also weak-to-strong outer semicontinuous.*

Proof. Let $x_n \rightharpoonup x$ and $x_n^* \to x^*$ with $x_n^* \in A(x_n)$ for all $n \in \mathbb{N}$. This implies that $x_n^* \overset{*}{\rightharpoonup} x^*$ as well and that $\{x_n\}_{n\in\mathbb{N}}$ is bounded. We thus have for some $C > 0$ that

$$\limsup_{n\to\infty} \langle x_n^* - x^*, x_n - x\rangle_X \le C \limsup_{n\to\infty} \|x_n^* - x^*\|_{X^*} = 0.$$

Hence, condition (ii) is satisfied, and the BCP outer semicontinuity yields $x^* \in A(x)$. □

We now show that convex subdifferentials are maximally monotone. Although this result (known as *Rockafellar's Theorem*, see [Rockafellar, 1970]) holds in arbitrary Banach spaces, the proof (adapted here from [Simons, 2009]) greatly simplifies in reflexive Banach spaces.

Theorem 6.13. *Let X be a reflexive Banach space and $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. Then $\partial F : X \rightrightarrows X^*$ is maximally monotone.*

Proof. First, we already know from Example 6.5 (iii) that ∂F is monotone. Let now $x \in X$ and $x^* \in X^*$ be given such that

$$\langle x^* - \tilde{x}^*, x - \tilde{x}\rangle_X \geq 0 \qquad \text{for all } \tilde{x} \in X, \tilde{x}^* \in \partial F(\tilde{x}). \tag{6.5}$$

We consider

$$J : X \to \overline{\mathbb{R}}, \qquad z \mapsto F(z+x) - \langle x^*, z\rangle_X + \frac{1}{2}\|z\|_X^2,$$

which is proper, convex and lower semicontinuous by the assumptions on F. Furthermore, J is coercive by Lemma 3.9. Theorem 3.8 thus yields a $\bar{z} \in X$ with $J(\bar{z}) = \min_{z\in X} J(z)$. By Theorems 4.2, 4.5, and 4.14 and Lemma 4.13 (ii) then

$$0 \in \partial F(\bar{z}+x) - \{x^*\} + \partial j(\bar{z}), \tag{6.6}$$

where we have introduced $j(z) := \frac{1}{2}\|z\|_X^2$. In other words, there exists a $z^* \in \partial j(\bar{z})$ such that $x^* - z^* \in \partial F(\bar{z}+x)$. Combining Lemma 5.4 for $p = q = 2$ and Lemma 5.8, we furthermore have that $z^* \in \partial j(\bar{z})$ if and only if

$$\langle z^*, \bar{z}\rangle_X = \frac{1}{2}\|\bar{z}\|_X^2 + \frac{1}{2}\|z^*\|_{X^*}^2. \tag{6.7}$$

Applying now (6.5) for $\tilde{x} = \bar{z} + x$ and $\tilde{x}^* = x^* - z^* \in \partial F(\tilde{x})$, we obtain using (6.7) that

$$0 \leq \langle x^* - x^* + z^*, x - \bar{z} - x\rangle_X = -\langle z^*, \bar{z}\rangle_X = -\frac{1}{2}\|z^*\|_{X^*}^2 - \frac{1}{2}\|\bar{z}\|_X^2,$$

implying that both $\bar{z} = 0$ and $z^* = 0$. Hence by (6.6) we conclude that $x^* \in \partial F(x)$, which shows that ∂F is maximally monotone. □

The argument in the preceding proof can be modified to give a characterization of maximal monotonicity for general monotone operators; this is known as *Minty's Theorem* and is a central result in the theory of monotone operators. We again make use of the *duality mapping* $\partial j : X \rightrightarrows X^*$ for $j(x) = \frac{1}{2}\|x\|_X^2$, noting for later use that if X is a Hilbert space (and we identify X^* with X), then $\partial j = \mathrm{Id}$.

Theorem 6.14 (Minty). *Let X be a reflexive Banach space and $A : X \rightrightarrows X^*$ be monotone with* $\operatorname{graph} A \neq \emptyset$. *If A is maximally monotone, then $\partial j + A$ is surjective.*

Proof. We proceed similarly as in the proof of Theorem 6.13 by constructing a functional F_A which plays the same role for A as F does for ∂F. Specifically, we define for a maximally monotone operator $A : X \rightrightarrows X^*$ with $\operatorname{graph} A \neq \emptyset$ the *Fitzpatrick functional*

$$F_A : X \times X^* \to (-\infty, \infty], \qquad (x, x^*) \mapsto \sup_{(z,z^*)\in \operatorname{graph} A} \left(\langle x^*, z\rangle_X + \langle z^*, x\rangle_X - \langle z^*, z\rangle_X\right), \tag{6.8}$$

which can be written equivalently as

$$F_A(x, x^*) = \langle x^*, x\rangle_X - \inf_{(z,z^*)\in \operatorname{graph} A} \langle x^* - z^*, x - z\rangle_X. \tag{6.9}$$

Each characterization implies useful properties.

(i) By maximal monotonicity of A, we have by definition that $\langle x^* - z^*, x - z\rangle_X \geq 0$ for all $(z, z^*) \in \operatorname{graph} A$ if and only if $(x, x^*) \in \operatorname{graph} A$. In particular, for all $(x, x^*) \notin \operatorname{graph} A$ there exists $(z, z^*) \in \operatorname{graph} A$ with $\langle x^* - z^*, x - z\rangle_X < 0$, and therefore $\inf_{(z,z^*)\in\operatorname{graph} A}\langle x^* - z^*, x - z\rangle_X < 0$ for all $(x, x^*) \notin \operatorname{graph} A$. Furthermore, for $(x, x^*) \in \operatorname{graph} A$ the infimum is attained in $(z, z^*) = (x, x^*)$). Hence (6.9) implies that $F_A(x, x^*) \geq \langle x^*, x\rangle_X$ with equality for $(x, x^*) \in \operatorname{graph} A$. Since $\operatorname{graph} A \neq \emptyset$, this shows that F_A is proper.

(ii) On the other hand, the definition (6.8) yields that

$$F_A = (G_A)^* \qquad \text{for} \qquad G_A(z^*, z) = \langle z^*, z\rangle_X + \delta_{\operatorname{graph} A^{-1}}(z^*, z)$$

(since $(z, z^*) \in \operatorname{graph} A$ if and only if $(z^*, z) \in \operatorname{graph} A^{-1}$). Furthermore, since $\operatorname{graph} A \neq \emptyset$ was assumed, F_A is the Fenchel conjugate of a proper functional and therefore convex and lower semicontinuous.

As a first step, we show that $0 \in \operatorname{ran}(\partial j + A)$. We set $\Xi := X \times X^*$ as well as $\xi := (x, x^*) \in \Xi$ and consider the functional

$$J_A : \Xi \to \overline{\mathbb{R}}, \qquad \xi \mapsto F_A(\xi) + \frac{1}{2}\|\xi\|_\Xi^2.$$

We first note that property (i) implies for all $\xi \in \Xi$ that

$$\begin{aligned} J_A(\xi) = F_A(\xi) + \frac{1}{2}\|\xi\|_\Xi^2 &= F_A(x, x^*) + \frac{1}{2}\|x\|_X^2 + \frac{1}{2}\|x^*\|_{X^*}^2 \\ &\geq \langle x^*, x\rangle_X + \frac{1}{2}\|x\|_X^2 + \frac{1}{2}\|x^*\|_{X^*}^2 \\ &\geq 0, \end{aligned} \tag{6.10}$$

where the last inequality follows from the Fenchel–Young inequality for j applied to $(x, -x^*)$. Furthermore, J_A is proper, convex, lower semicontinuous, and (by Lemma 3.9) coercive. Theorem 3.8 thus yields a $\bar\xi := (\bar x, \bar x^*) \in \Xi$ with $J_A(\bar\xi) = \min_{\xi\in\Xi} J_A(\xi)$, which by Theorems 4.2, 4.5, and 4.14 satisfies that

$$0 \in \partial J_A(\bar\xi) = \partial\left(\frac{1}{2}\|\bar\xi\|_X^2\right) + \partial F_A(\bar\xi),$$

i.e., there exists a $\bar\xi^* = (\bar w^*, \bar w) \in \Xi^* \simeq X^* \times X$ (since X is reflexive) such that $\bar\xi^* \in \partial F_A(\bar\xi)$ and $-\bar\xi^* \in \partial(\frac{1}{2}\|\bar\xi\|_X^2)$.

By definition of the subdifferential, we thus have for all $\xi \in \Xi$ that

$$F_A(\xi) \geq F_A(\bar\xi) + \langle\bar\xi^*, \xi - \bar\xi\rangle_\Xi = J_A(\bar\xi) + \frac{1}{2}\|\bar\xi^*\|_{\Xi^*}^2 + \langle\bar\xi^*, \xi\rangle_\Xi \geq \frac{1}{2}\|\bar\xi^*\|_\Xi^2 + \langle\bar\xi^*, \xi\rangle_\Xi,$$

where the second step uses again the Fenchel–Young inequality holding with equality for $(\bar\xi, -\bar\xi^*)$, and the last step follows from (6.10). Property (i) then implies for all $(x, x^*) \in \operatorname{graph} A$ that

$$\langle x^*, x\rangle_X = F_A(x, x^*) \geq \frac{1}{2}\|\bar w^*\|_{X^*}^2 + \frac{1}{2}\|\bar w\|_X^2 + \langle\bar w^*, x\rangle_X + \langle x^*, \bar w\rangle_X.$$

Adding $\langle \bar{w}^*, \bar{w} \rangle_X$ on both sides and rearranging yields

$$\langle x^* - \bar{w}^*, x - \bar{w} \rangle_X \geq \langle \bar{w}^*, \bar{w} \rangle_X + \frac{1}{2}\|\bar{w}^*\|_{X^*}^2 + \frac{1}{2}\|\bar{w}\|_X^2 \geq 0, \tag{6.11}$$

again by the Fenchel–Young inequality. The maximal monotonicity of A thus yields that $\bar{w}^* \in A(\bar{w})$, i.e., $(\bar{w}, \bar{w}^*) \in \operatorname{graph} A$. Inserting this for (x, x^*) in (6.11) then shows that

$$\langle \bar{w}^*, \bar{w} \rangle_X + \frac{1}{2}\|\bar{w}^*\|_{X^*}^2 + \frac{1}{2}\|\bar{w}\|_X^2 = 0.$$

Hence the Fenchel–Young inequality for ∂j holds with equality at $(\bar{w}, -\bar{w}^*)$, implying $-\bar{w}^* \in \partial j(\bar{w})$. Together, we obtain that $0 = -\bar{w}^* + \bar{w}^* \in (\partial j + A)(\bar{w})$.

Finally, let $z^* \in X^*$ be arbitrary and set $B : X \rightrightarrows X^*, x \mapsto \{-z^*\} + A(x)$. Using the definition, it is straightforward to verify that B is maximally monotone with graph $B \neq \emptyset$ as well. As we have just shown, there now exists a $\bar{x}^* \in X^*$ with $0 \in (\partial j + B)(\bar{x}^*) = \{\bar{x}^*\} + \{-z^*\} + A(\bar{x}^*)$, i.e., $z^* \in (\partial j + A)(\bar{x}^*)$. Hence $\partial j + A$ is surjective. □

6.3 RESOLVENTS AND PROXIMAL POINTS

The proof of Theorem 6.13 is based on associating to any $x^* \in \partial F(x)$ an element $\bar{z} \in X$ as the minimizer of a suitable functional. If X is a Hilbert space, this functional is even strictly convex and hence the minimizer $\bar{z}$ is unique. This property can be exploited to define a new *single-valued* mapping that is more useful for algorithms than the set-valued subdifferential mapping. For this purpose, we restrict the discussion in the remainder of this chapter to Hilbert spaces (but see Remark 6.29 below). This allows identifying X^* with X; in particular, we will from now on identify the set $\partial F(x) \subset X^*$ of subderivatives with the corresponding set in X of subgradients (i.e., their Riesz representations). By the same token, we will also use the same notation for inner products as for duality pairings to avoid the danger of confusing pairs of elements $(x, x^*) \in \operatorname{graph} \partial F$ with their inner product.

We can then define for a maximally monotone operator $A : X \rightrightarrows X$ with $\operatorname{graph} A \neq \emptyset$ the *resolvent*

$$\mathcal{R}_A : X \rightrightarrows X, \qquad \mathcal{R}_A(x) = (\mathrm{Id} + A)^{-1}x,$$

as well as for a proper, convex, and lower semicontinuous functional $F : X \to \overline{\mathbb{R}}$ the *proximal point mapping*

$$\operatorname{prox}_F : X \to X, \qquad \operatorname{prox}_F(x) = \operatorname*{arg\,min}_{z \in X} \frac{1}{2}\|z - x\|_X^2 + F(z). \tag{6.12}$$

Since a similar argument as in the proof of Theorem 6.13 shows that $w \in \mathcal{R}_{\partial F}(x)$ is equivalent to the necessary and sufficient conditions for the *proximal point* w to be a minimizer of the strictly convex functional in (6.12), we have that

$$\operatorname{prox}_F = (\mathrm{Id} + \partial F)^{-1} = \mathcal{R}_{\partial F}. \tag{6.13}$$

Resolvents of monotone and, in particular, maximal monotone operators have useful properties.

Lemma 6.15. *If $A : X \rightrightarrows X$ is monotone, $\mathcal{R}_A$ is* firmly nonexpansive, *i.e.,*

$$\|z_1 - z_2\|_X^2 \leq \langle x_1 - x_2, z_1 - z_2\rangle_X \qquad \text{for all } (x_1, z_1), (x_2, z_2) \in \operatorname{graph} \mathcal{R}_A \tag{6.14}$$

or equivalently,

$$\|z_1 - z_2\|_X^2 + \|(x_1 - z_1) - (x_2 - z_2)\|_X^2 \leq \|x_1 - x_2\|_X^2 \tag{6.15}$$
$$\text{for all } (x_1, z_1), (x_2, z_2) \in \operatorname{graph} \mathcal{R}_A.$$

Proof. Let $x_1, x_2 \in \operatorname{dom} \mathcal{R}_A$ as well as $z_1 \in \mathcal{R}_A(x_1)$ and $z_2 \in \mathcal{R}_A(x_2)$. By definition of the resolvent, this implies that $x_1 - z_1 \in A(z_1)$ and $x_2 - z_2 \in A(z_2)$. By the monotonicity of A, we thus have

$$0 \leq \langle (x_1 - z_1) - (x_2 - z_2), z_1 - z_2\rangle_X,$$

which after rearranging yields (6.14). The equivalence of (6.14) and (6.15) is straightforward to verify using binomial expansion. □

Corollary 6.16. *Let $A : X \rightrightarrows X$ be maximally monotone with* $\operatorname{graph} A \neq \emptyset$. *Then $\mathcal{R}_A : X \to X$ is single-valued and Lipschitz continuous with constant $L = 1$.*

Proof. Under the stated assumptions, $\mathrm{Id} + A$ is surjective by Theorem 6.14, which implies that $\mathcal{R}_A(x) \neq \emptyset$ for all $x \in X$, i.e., $\operatorname{dom} \mathcal{R}_A = X$. Let now $x \in X$ and $z_1, z_2 \in \mathcal{R}_A(x)$. Since A is monotone, $\mathcal{R}_A$ is nonexpansive by Lemma 6.15, which yields both single-valuedness of $\mathcal{R}_A$ (by taking $x_1 = x_2 = x$ implies $z_1 = z_2$) and its Lipschitz continuity (by applying the Cauchy–Schwarz inequality). □

In particular, by Theorem 6.13, this holds for the proximal point mapping $\operatorname{prox}_F : X \to X$ of a proper, convex, and lower semicontinuous functional $F : X \to \mathbb{R}$.

Remark 6.17. Conversely, it can be shown that every nonexpansive mapping $T : X \to X$ that satisfies $T(x) \in \partial G(x)$ for all $x \in X$ for some proper, convex, and lower semicontinuous functional $G : X \to \overline{\mathbb{R}}$ is the proximal mapping of some proper, convex, and lower semicontinuous functional $F : X \to \overline{\mathbb{R}}$; see [Gribonval and Nikolova, 2020; Moreau, 1965].

Lipschitz continuous mappings with constant $L = 1$ are also called *nonexpansive*. A related concept that is sometimes used is the following. A mapping $T : X \to X$ is called *α-averaged* for some $\alpha \in (0, 1)$, if $T = (1 - \alpha)\mathrm{Id} + \alpha J$ for some nonexpansive $J : X \to X$. We then have the following relation.

Lemma 6.18. *Let $T : X \to X$. Then T is firmly nonexpansive if and only if T is $(1/2)$-averaged.*

Proof. Suppose T is $(1/2)$-averaged. Then $T = \frac{1}{2}(\mathrm{Id} + J)$ for some nonexpansive J. We compute

$$\begin{aligned}\|T(x) - T(y)\|_X^2 &= \frac{1}{4}\left(\|J(x) - J(y)\|_X^2 + 2\langle J(x) - J(y), x - y\rangle_X + \|x - y\|_X^2\right)\\ &\leq \frac{1}{2}\left(\langle J(x) - J(y), x - y\rangle_X + \|x - y\|_X^2\right)\\ &= \langle T(x) - T(y), x - y\rangle_X.\end{aligned}$$

Thus T is firmly nonexpansive.

Suppose then that T is firmly nonexpansive. If we show that $J := 2T - \mathrm{Id}$ is nonexpansive, it follows that T is $(1/2)$-averaged. This is established by the simple calculations

$$\begin{aligned}\|J(x) - J(y)\|_X^2 &= 4\|T(x) - T(y)\|_X^2 - 4\langle T(x) - T(y), x - y\rangle_X + \|x - y\|_X^2\\ &\leq \|x - y\|_X^2.\end{aligned}$$

This completes the proof. □

Like maximally monotone operators, α-averaged operators always have outer semicontinuity properties. To show this, we will use that in Hilbert spaces, the converse of Minty's Theorem 6.14 holds (with the duality mapping $\partial j = \mathrm{Id}$).

Lemma 6.19. *Let $A : X \rightrightarrows X$ be monotone. If $\mathrm{Id} + A$ is surjective, then A is maximally monotone.*

Proof. Consider $x \in X$ and $x^* \in X$ with

$$\langle x^* - \tilde{x}^*, x - \tilde{x}\rangle_X \geq 0 \qquad \text{for all } (\tilde{x}, \tilde{x}^*) \in \operatorname{graph} A. \tag{6.16}$$

If $\mathrm{Id} + A$ is surjective, then for $x + x^* \in X$ there exist a $z \in X$ and a $z^* \in A(z)$ with

$$x + x^* = z + z^* \in (\mathrm{Id} + A)z. \tag{6.17}$$

Inserting $(\tilde{x}, \tilde{x}^*) = (z, z^*)$ into (6.16) then yields that

$$0 \leq \langle x^* - z^*, x - z\rangle_X = \langle z - x, x - z\rangle_X = -\|x - z\|_X^2 \leq 0,$$

i.e., $x = z$. From (6.17) we further obtain $x^* = z^* \in A(z) = A(x)$, and hence A is maximally monotone. □

Lemma 6.20. *Let $T : X \to X$ be α-averaged. Then T is weak-to-strong and strong-to-weakly-$*$ outer semicontinuous, and the set of fixed points $\bar{x} = T(\bar{x})$ of T is convex and closed.*

Proof. Let $T = (1-\alpha)\mathrm{Id} + \alpha J$ for some nonexpansive operator $J : X \to X$. Then clearly $x \in X$ is a fixed point of T if and only if x is a fixed point of J. It thus suffices to show the claim for the fixed-point set $\{\bar{x} \mid \bar{x} = J(\bar{x})\} = (\mathrm{Id} - J)^{-1}(0)$ of a nonexpansive operator J. By Lemmas 6.8 to 6.10, we thus only need to show that $\mathrm{Id} - J$ is maximally monotone.

First, $\mathrm{Id} - J$ is clearly monotone. Moreover, $2\mathrm{Id} - J = \mathrm{Id} + (\mathrm{Id} - J)$ is surjective since otherwise $2x - J(x) = 2y - J(y)$ for $x \neq y$, which together with the assumed nonexpansivity would lead to the contradiction $0 \neq 2\|x - y\| \leq \|x - y\|$. Lemma 6.19 then shows that $\mathrm{Id} - J$ is maximally monotone, and the claim follows. □

The following useful result allows characterizing minimizers of convex functionals as proximal points.

Lemma 6.21. *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $x, x^* \in X$. Then for any $\gamma > 0$,*

$$x^* \in \partial F(x) \quad \Leftrightarrow \quad x = \mathrm{prox}_{\gamma F}(x + \gamma x^*).$$

Proof. Multiplying both sides of the subdifferential inclusion by $\gamma > 0$ and adding x yields that

$$\begin{aligned} x^* \in \partial F(x) &\Leftrightarrow x + \gamma x^* \in (\mathrm{Id} + \gamma \partial F)(x) \\ &\Leftrightarrow x \in (\mathrm{Id} + \gamma \partial F)^{-1}(x + \gamma x^*) \\ &\Leftrightarrow x = \mathrm{prox}_{\gamma F}(x + \gamma x^*), \end{aligned}$$

where in the last step we have used that $\gamma \partial F = \partial(\gamma F)$ by Lemma 4.13 (i) and hence that $\mathrm{prox}_{\gamma F} = \mathcal{R}_{\partial(\gamma F)} = \mathcal{R}_{\gamma \partial F}$. □

By applying Lemma 6.21 to the Fermat principle $0 \in \partial F(\bar{x})$, we obtain the following fixed-point characterization of minimizers of F.

Corollary 6.22. *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex and lower semicontinuous, and $\gamma > 0$ be arbitrary. Then $\bar{x} \in \mathrm{dom}\, F$ is a minimizer of F if and only if*

$$\bar{x} = \mathrm{prox}_{\gamma F}(\bar{x}).$$

This simple result should not be underestimated: It allows replacing (explicit) set inclusions in optimality conditions by equivalent (implicit) Lipschitz continuous equations, which (as we will show in following chapters) can be solved by fixed-point iteration or Newton-type methods.

We can also derive a generalization of the orthogonal decomposition of vector spaces.

Theorem 6.23 (Moreau decomposition). *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. Then we have for all $x \in X$ that*

$$x = \operatorname{prox}_F(x) + \operatorname{prox}_{F^*}(x).$$

Proof. Setting $w = \operatorname{prox}_F(x)$, Lemmas 5.8 and 6.21 for $\gamma = 1$ imply that

$$\begin{aligned} w = \operatorname{prox}_F(x) = \operatorname{prox}_F(w + (x - w)) &\Leftrightarrow x - w \in \partial F(w) \\ &\Leftrightarrow w \in \partial F^*(x - w) \\ &\Leftrightarrow x - w = \operatorname{prox}_{F^*}((x - w) + w) = \operatorname{prox}_{F^*}(x). \end{aligned}$$ □

The following calculus rules will prove useful.

Lemma 6.24. *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. Then,*

(i) for $\lambda \neq 0$ and $z \in X$ we have with $H(x) := F(\lambda x + z)$ that

$$\operatorname{prox}_H(x) = \lambda^{-1}(\operatorname{prox}_{\lambda^2 F}(\lambda x + z) - z);$$

(ii) for $\gamma > 0$ we have that

$$\operatorname{prox}_{\gamma F^*}(x) = x - \gamma \operatorname{prox}_{\gamma^{-1} F}(\gamma^{-1} x);$$

(iii) for proper, convex, lower semicontinuous $G : Y \to \overline{\mathbb{R}}$ and $\gamma > 0$ we have with $H(x, y) := F(x) + G(y)$ that

$$\operatorname{prox}_{\gamma H}(x, y) = \begin{pmatrix} \operatorname{prox}_{\gamma F}(x) \\ \operatorname{prox}_{\gamma G}(y) \end{pmatrix}.$$

Proof. *(i):* By definition,

$$\operatorname{prox}_H(x) = \operatorname*{arg\,min}_{w \in X} \frac{1}{2}\|w - x\|_X^2 + F(\lambda w + z) =: \bar{w}.$$

Now note that since X is a vector space,

$$\min_{w \in X} \frac{1}{2}\|w - x\|_X^2 + F(\lambda w + z) = \min_{v \in X} \frac{1}{2}\|\lambda^{-1}(v - z) - x\|_X^2 + F(v),$$

and the respective minimizers $\bar{w}$ and $\bar{v}$ are related by $\bar{v} = \lambda \bar{w} + z$. The claim then follows from

$$\begin{aligned} \bar{v} &= \operatorname*{arg\,min}_{v \in X} \frac{1}{2}\|\lambda^{-1}(v - z) - x\|_X^2 + F(v) \\ &= \operatorname*{arg\,min}_{v \in X} \frac{1}{2\lambda^2}\|v - (\lambda x + z)\|_X^2 + F(v) \\ &= \operatorname*{arg\,min}_{v \in X} \frac{1}{2}\|v - (\lambda x + z)\|_X^2 + \lambda^2 F(v) \\ &= \operatorname{prox}_{\lambda^2 F}(\lambda x + z). \end{aligned}$$

(ii): Theorem 6.23, Lemma 5.7 (i), and (i) for $\lambda = \gamma^{-1}$ and $z = 0$ together imply that

$$\begin{aligned}\operatorname{prox}_{\gamma F}(x) &= x - \operatorname{prox}_{(\gamma F)^*}(x)\\ &= x - \operatorname{prox}_{\gamma F^* \circ (\gamma^{-1}\mathrm{Id})}(x)\\ &= x - \gamma \operatorname{prox}_{\gamma(\gamma^{-2}F^*)}(\gamma^{-1}x).\end{aligned}$$

Applying this to F^* and using that $F^{**} = F$ by Theorem 5.1 (iii) now yields the claim.

(iii): By definition of the norm on the product space $X \times Y$, we have that

$$\begin{aligned}\operatorname{prox}_{\gamma H}(x, y) &= \operatorname*{arg\,min}_{(u,v)\in X\times Y} \frac{1}{2}\|(u, v) - (x, y)\|^2_{X\times Y} + \gamma H(u, v)\\ &= \operatorname*{arg\,min}_{u\in X, v\in Y} \left(\frac{1}{2}\|u - x\|^2_X + \gamma F(u)\right) + \left(\frac{1}{2}\|v - y\|^2_Y + \gamma G(v)\right).\end{aligned}$$

Since there are no mixed terms in u and v, the two terms in parentheses can be minimized separately. Hence, $\operatorname{prox}_{\gamma H}(x, y) = (\bar{u}, \bar{v})$ for

$$\begin{aligned}\bar{u} &= \operatorname*{arg\,min}_{u\in X} \frac{1}{2}\|u - x\|^2_X + \gamma F(u) = \operatorname{prox}_{\gamma F(x)},\\ \bar{v} &= \operatorname*{arg\,min}_{v\in Y} \frac{1}{2}\|v - y\|^2_Y + \gamma G(v) = \operatorname{prox}_{\gamma G(x)}.\end{aligned}$$

□

Computing proximal points is difficult in general since evaluating prox_F by its definition entails minimizing F. In some cases, however, it is possible to give an explicit formula for prox_F.

Example 6.25. We first consider scalar functions $f : \mathbb{R} \to \overline{\mathbb{R}}$.

(i) $f(t) = \frac{1}{2}|t|^2$. Since f is differentiable, we can set the derivative of $\frac{1}{2}(s - t)^2 + \frac{\gamma}{2}s^2$ to zero and solve for s to obtain $\operatorname{prox}_{\gamma f}(t) = (1 + \gamma)^{-1}t$.

(ii) $f(t) = |t|$. By Example 4.7, we have that $\partial f(t) = \operatorname{sign}(t)$; hence $s := \operatorname{prox}_{\gamma f}(t) = (\mathrm{Id} + \gamma \operatorname{sign})^{-1}(t)$ if and only if $t \in \{s\} + \gamma \operatorname{sign}(s)$. Let t be given and assume this holds for some $\bar{s}$. We now proceed by case distinction.

Case 1: $\bar{s} > 0$. This implies that $t = \bar{s} + \gamma$, i.e., $\bar{s} = t - \gamma$, and hence that $t > \gamma$.

Case 2: $\bar{s} < 0$. This implies that $t = \bar{s} - \gamma$, i.e., $\bar{s} = t + \gamma$, and hence that $t < -\gamma$.

Case 3: $\bar{s} = 0$. This implies that $t \in \gamma[-1, 1] = [-\gamma, \gamma]$.

Since this yields a complete and disjoint case distinction for t, we can conclude

that

$$\operatorname{prox}_{\gamma f}(t) = \begin{cases} t-\gamma & \text{if } t > \gamma, \\ 0 & \text{if } t \in [-\gamma, \gamma], \\ t+\gamma & \text{if } t < -\gamma. \end{cases}$$

This mapping is also known as the *soft-shrinkage* or]soft-thresholding operator.

(iii) $f(t) = \delta_{[-1,1]}(t)$. We can proceed here in the same way as in (ii), but for the sake of variety we instead use Lemma 6.24 (ii) to compute the proximal point mapping from that of $f^*(t) = |t|$ (see Example 5.3 (ii)) via

$$\begin{aligned} \operatorname{prox}_{\gamma f}(t) &= t - \gamma \operatorname{prox}_{\gamma^{-1} f^*}(\gamma^{-1} t) \\ &= \begin{cases} t - \gamma(\gamma^{-1} t - \gamma^{-1}) & \text{if } \gamma^{-1} t > \gamma^{-1}, \\ t - 0 & \text{if } \gamma^{-1} t \in [-\gamma^{-1}, \gamma^{-1}], \\ t - \gamma(\gamma^{-1} t + \gamma^{-1}) & \text{if } \gamma^{-1} t < -\gamma^{-1} \end{cases} \\ &= \begin{cases} 1 & \text{if } t > 1, \\ t & \text{if } t \in [-1, 1], \\ -1 & \text{if } t < -1. \end{cases} \end{aligned}$$

For every $\gamma > 0$, the proximal point of t is thus its projection onto $[-1, 1]$.

Example 6.26. We can generalize Example 6.25 to $X = \mathbb{R}^N$ (endowed with the Euclidean inner product) by applying Lemma 6.24 (iii) N times. We thus obtain componentwise

(i) for $F(x) = \frac{1}{2}\|x\|_2^2 = \sum_{i=1}^N \frac{1}{2} x_i^2$ that

$$[\operatorname{prox}_{\gamma F}(x)]_i = \left(\frac{1}{1+\gamma}\right) x_i, \quad 1 \le i \le N;$$

(ii) for $F(x) = \|x\|_1 = \sum_{i=1}^N |x_i|$ that

$$[\operatorname{prox}_{\gamma F}(x)]_i = (|x_i| - \gamma)^+ \operatorname{sign}(x_i), \quad 1 \le i \le N;$$

(iii) for $F(x) = \delta_{\mathbb{B}_\infty}(x) = \sum_{i=1}^N \delta_{[-1,1]}(x_i)$ that

$$[\operatorname{prox}_{\gamma F}(x)]_i = x_i - (x_i - 1)^+ - (x_i + 1)^- = \frac{x_i}{\max\{1, |x_i|\}}, \qquad 1 \le i \le N.$$

Here we have used the convenient notation $(t)^+ := \max\{t, 0\}$ and $(t)^- := \min\{t, 0\}$.

Many more examples of projection operators and proximal mappings can be found in

[Cegielski, 2012], [Parikh and Boyd, 2014, § 6.5], [Beck, 2017], as well as at https://www.proximity-operator.net.

Since the subdifferential of convex integral functionals can be evaluated pointwise by Theorem 4.11, the same holds for the definition (6.13) of the proximal point mapping.

Corollary 6.27. *Let $f : \mathbb{R} \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $F : L^2(\Omega) \to \overline{\mathbb{R}}$ be defined by superposition as in Lemma 3.7. Then we have for all $\gamma > 0$ and $u \in L^2(\Omega)$ that*

$$[\operatorname{prox}_{\gamma F}(u)](x) = \operatorname{prox}_{\gamma f}(u(x)) \qquad \text{for almost every } x \in \Omega.$$

Example 6.28. Let X be a Hilbert space. Similarly to Example 6.25 one can show that

(i) for $F = \frac{1}{2}\|\cdot\|_X^2 = \frac{1}{2}\langle\cdot,\cdot\rangle_X$, that

$$\operatorname{prox}_{\gamma F}(x) = \left(\frac{1}{1+\gamma}\right) x;$$

(ii) for $F = \|\cdot\|_X$, using a case distinction as in Theorem 4.6, that

$$\operatorname{prox}_{\gamma F}(x) = \left(1 - \frac{\gamma}{\|x\|_X}\right)^+ x;$$

(iii) for $F = \delta_C$ with $C \subset X$ nonempty, convex, and closed, that by definition

$$\operatorname{prox}_{\gamma F}(x) = \operatorname{proj}_C(x) := \operatorname*{arg\,min}_{z \in C} \|z - x\|_X$$

the *metric projection* of x onto C; the proximal point mapping thus generalizes the concept projection onto convex sets. Explicit or at least constructive formulas for the projection onto different classes of sets can be found in [Cegielski, 2012, Chapter 4.1].

Remark 6.29. The results of this section can be extended to (reflexive) Banach spaces if the identity is replaced by the duality mapping $\partial j : X \rightrightarrows X^*$ for $j(x) = \frac{1}{2}\|x\|_X^2$. If the norm is differentiable (which is the case if the unit ball of X^* is *strictly* convex as for, e.g., $X = L^p(\Omega)$ with $p \in (1, \infty)$), the duality mapping is in fact single-valued [Cioranescu, 1990, Theorem 2.16], and hence the corresponding resolvent $(\partial j + A)^{-1}$ is well-defined. However, the proximal mapping need no longer be Lipschitz continuous, although the definition can be modified to obtain uniform continuity; see [Bačák and Kohlenbach, 2018]. Similarly, the Moreau decomposition (Theorem 6.23) needs to be modified appropriately; see [Combettes and Reyes, 2013]. The main difficulty from our point of view, however, lies in the evaluation of the proximal mapping, which then rarely admits a closed form even for simple functionals.

7 SMOOTHNESS AND CONVEXITY

Before we turn to algorithms for the solution of nonsmooth optimization problems, we derive consequences of convexity for *differentiable* functionals that will be useful in proving convergence of splitting methods for functionals involving a smooth component. In particular, we will show that Lipschitz continuous differentiability is linked via Fenchel duality to strong convexity.

7.1 SMOOTHNESS

We now derive useful consequences of Lipschitz differentiability and their relation to convexity. Recall from Theorem 4.5 that for $F : X \to \overline{\mathbb{R}}$ convex and Gâteaux differentiable, $\partial F(x) = \{DF(x)\}$ (which can be identified with $\{\nabla F(x)\} \subset X$ in Hilbert spaces).

Lemma 7.1. *Let X be a Banach space and let $F : X \to \mathbb{R}$ be Gâteaux differentiable. Consider the properties:*

(i) The property

$$F(y) \leq F(x) + \langle DF(y), y - x\rangle_X - \frac{1}{2L}\|DF(x) - DF(y)\|_{X^*}^2 \quad \text{for all } x, y \in X. \tag{7.1}$$

(ii) The co-coercivity *of DF with factor L^{-1}:*

$$L^{-1}\|DF(x) - DF(y)\|_{X^*}^2 \leq \langle DF(x) - DF(y), x - y\rangle_X \quad \text{for all } x, y \in X. \tag{7.2}$$

(iii) Lipschitz continuity of DF with factor L:

$$\|DF(x) - DF(y)\|_{X^*} \leq L\|x - y\|_X \quad \text{for all } x, y \in X. \tag{7.3}$$

(iv) The property

$$\langle DF(x + h) - DF(x), h\rangle_X \leq L\|h\|_X^2 \quad \text{for all } x, h \in X. \tag{7.4}$$

(v) The smoothness *(also known as* descent lemma*) of F with factor L:*

$$F(x+h) \le F(x) + \langle DF(x), h\rangle_X + \frac{L}{2}\|h\|_X^2 \quad \text{for all } x, h \in X. \tag{7.5}$$

(vi) The uniform smoothness *of F with factor L:*

$$F(\lambda x + (1-\lambda)y) + \lambda(1-\lambda)\frac{L}{2}\|x-y\|_X^2 \ge \lambda F(x) + (1-\lambda)F(y) \quad \text{for all } x, y \in X, \lambda \in [0,1]. \tag{7.6}$$

Then (i) ⇒ (ii) ⇒ (iii) ⇒ (iv) ⇔ (v) ⇔ (vi). If F is convex and X is reflexive, then all the properties are equivalent.

Proof. (i) ⇒ (ii): Summing the estimate (7.1) with the same estimate with x and y exchanged, we obtain (7.2).

(ii) ⇒ (iii): This follows immediately from (1.1).

(iii) ⇒ (iv): Taking $y = x + h$ and multiplying (7.3) by $\|h\|_X$, the property follows again from (1.1).

(iv) ⇒ (v): Using the mean value Theorem 2.10 and (7.4), we obtain

$$\begin{aligned} F(x+h) - F(x) - \langle DF(x), h\rangle_X &= \int_0^1 \langle DF(x+th), h\rangle_X \, dt - \langle DF(x), h\rangle_X \\ &= \int_0^1 \langle DF(x+th) - DF(x), h\rangle_X \, dt \\ &\le \int_0^1 t \, dt \cdot L\|h\|_X^2 = \frac{L}{2}\|h\|_X^2. \end{aligned}$$

(v) ⇒ (iv): This follows by adding together (7.5) and the same inequality with $x + h$ in place of x.

(v) ⇒ (vi): Set $x_\lambda := \lambda x + (1-\lambda)y$. Multiplying (7.5) first for $x = x_\lambda$ and $h = x - x_\lambda = (1-\lambda)(x-y)$ with λ and then for $x = x_\lambda$ and $h = y - x_\lambda = \lambda(y-x)$ with $1-\lambda$ and adding the results yields (7.6).

(vi) ⇒ (v): This follows by dividing (7.6) by $\lambda > 0$ and taking the limit $\lambda \to 0$.

(v) ⇒ (i) when F is convex and X is reflexive: Since F is convex, we have from Theorem 4.5 that

$$\langle DF(y), (x+h) - y\rangle_X \le F(x+h) - F(y).$$

Combining this with (7.5) yields

$$F(y) \le F(x) + \langle DF(x), h\rangle_X - \langle DF(y), (x+h) - y\rangle_X + \frac{L}{2}\|h\|_X^2 \tag{7.7}$$
$$= F(x) + \langle DF(y), y - x\rangle_X + \langle DF(x) - DF(y), h\rangle_X + \frac{L}{2}\|h\|_X^2.$$

Let $z^* := -L^{-1}(DF(x) - DF(y))$. Since X is reflexive, the algebraic Hahn–Banach Theorem 1.4 yields (after multiplication by $\|z^*\|_{X^*}$) an $h \in X$ such that

$$\|h\|_X = \|z^*\|_{X^*} \quad \text{and} \quad \langle z^*, h\rangle_X = \|z^*\|_{X^*}^2.$$

Consequently, continuing from (7.7),

$$\begin{aligned} F(y) &\le F(x) + \langle DF(y), y - x\rangle_X - L\langle z^*, h\rangle_X + \frac{L}{2}\|h\|_X^2 \\ &= F(x) + \langle DF(y), y - x\rangle_X - \frac{L}{2}\|z^*\|_{X^*}^2 \\ &= F(x) + \langle DF(y), y - x\rangle_X - \frac{1}{2L}\|DF(x) - DF(y)\|_{X^*}^2. \end{aligned}$$

This proves (7.1). □

The next "smoothness three-point corollary" will be valuable for the study of splitting methods that involve a smooth component function.

Corollary 7.2. *Let X be a reflexive Banach space and let $F : X \to \mathbb{R}$ be convex and Gâteaux differentiable. Then the following are equivalent:*

(i) F has L^{-1}-co-coercive derivative (or any of the equivalent properties of Lemma 7.1).

(ii) The three-point smoothness

$$\langle DF(z), x - \widehat{x}\rangle_X \ge F(x) - F(\widehat{x}) - \frac{L}{2}\|x - z\|_X^2 \quad \text{for all } \widehat{x}, z, x \in X, \tag{7.8}$$

(iii) The three-point monotonicity

$$\langle DF(z) - DF(\widehat{x}), x - \widehat{x}\rangle_X \ge -\frac{L}{4}\|x - z\|_X^2 \quad \text{for all } \widehat{x}, z, x \in X. \tag{7.9}$$

Proof. If ∇F is L^{-1}-co-coercive, using Lemma 7.1, we have the L-smoothness

$$F(z) - F(x) \ge \langle DF(z), z - x\rangle_X - \frac{L}{2}\|x - z\|_X^2.$$

By convexity $F(\widehat{x}) - F(z) \ge \langle DF(z), \widehat{x} - z\rangle_X$. Summing up, we obtain (7.8).

Regarding (7.9), by assumption we have the co-coercivity

$$\langle DF(z) - DF(\widehat{x}), z - \widehat{x}\rangle_X \geq L^{-1}\|DF(z) - DF(\widehat{x})\|_{X^*}^2.$$

Thus, using (1.1) and Young's inequality in the form $ab \leq \frac{1}{2\alpha}a^2 + \frac{\alpha}{2}b^2$ for $a, b \in \mathbb{R}$ and $\alpha > 0$, we obtain

$$\begin{aligned}\langle DF(z) - DF(\widehat{x}), x - \widehat{x}\rangle_X &= \langle DF(z) - DF(\widehat{x}), z - \widehat{x}\rangle_X + \langle DF(z) - DF(\widehat{x}), x - z\rangle_X \\ &\geq L^{-1}\|DF(z) - DF(\widehat{x})\|_{X^*}^2 - \|DF(z) - DF(\widehat{x})\|_{X^*}\|x - z\|_X \\ &\geq -\frac{L}{4}\|x - z\|_X^2.\end{aligned}$$

This is (7.9)

For the reverse implications, we assume that (7.9) holds and set $z^* := -2L^{-1}(DF(z) - DF(\widehat{x}))$. By the assumed reflexivity, we can again apply the algebraic Hahn–Banach Theorem 1.4 to obtain an $h \in X$ such that

$$\|h\|_X = \|z^*\|_{X^*} \quad \text{and} \quad \langle z^*, h\rangle_X = \|z^*\|_{X^*}^2.$$

With $x = z + h$, (7.9) gives

$$\begin{aligned}\langle DF(z) - DF(\widehat{x}), z - \widehat{x}\rangle_X &\geq -\langle DF(z) - DF(\widehat{x}), h\rangle_X - \frac{L}{4}\|h\|_X^2 \\ &= \frac{L}{2}\langle z^*, h\rangle_X - \frac{L}{4}\|z^*\|_{X^*}^2 \\ &= \frac{L}{4}\|z^*\|_{X^*}^2 = \frac{1}{L}\|DF(z) - DF(\widehat{x})\|_{X^*}^2.\end{aligned}$$

This is the L^{-1}-co-coercivity (7.2). The remaining equivalences follow from Lemma 7.1. □

7.2 STRONG CONVEXITY

The central notion in this chapter (and later for obtaining higher convergence rates for first-order algorithms) is the following "quantitative" version of convexity. We say that $F : X \to \overline{\mathbb{R}}$ is *strongly convex* with the factor $\gamma > 0$ if for all $x, y \in X$ and $\lambda \in [0, 1]$,

$$F(\lambda x + (1 - \lambda)y) + \lambda(1 - \lambda)\frac{\gamma}{2}\|x - y\|_X^2 \leq \lambda F(x) + (1 - \lambda)F(y). \tag{7.10}$$

Obviously, strong convexity implies strict convexity, so strongly convex functions have a unique minimizer. If X is a Hilbert space, it is straightforward if tedious to verify by expanding the squared norm that (7.10) is equivalent to $F - \frac{\gamma}{2}\|\cdot\|_X^2$ being convex.

We have the following important duality result that was first shown in [Azé and Penot, 1995].

Theorem 7.3. *Let $F : X \to \overline{\mathbb{R}}$ be proper and convex. Then the following are true:*

(i) If F is strongly convex with factor γ, then F^* is uniformly smooth with factor γ^{-1}.

(ii) If F is uniformly smooth with factor L, then F^* is strongly convex with factor L^{-1}.

(iii) If F is lower semicontinuous, then F is uniformly smooth with factor L if and only if F^* is strongly convex with factor L^{-1}.

Proof. *(i):* Let $x^*, y^* \in X^*$ and $\alpha_x, \alpha_y \in \mathbb{R}$ with $\alpha_x < F^*(x^*)$ and $\alpha_y < F^*(y^*)$. From the definition of the Fenchel conjugate, there exist $x, y \in X$ such that

$$\alpha_x < \langle x^*, x\rangle_X - F(x), \qquad \alpha_y < \langle y^*, y\rangle_X - F(y).$$

Multiplying the first inequality with $\lambda \in [0,1]$, the second with $(1-\lambda)$, and using the Fenchel–Young inequality (5.1) in the form

$$0 \le F(x_\lambda) + F^*(x_\lambda^*) - \langle x_\lambda^*, x_\lambda\rangle_X$$

for $x_\lambda^* := \lambda x^* + (1-\lambda)y^*$ and $x_\lambda := \lambda x + (1-\lambda)y$ then yields

$$\begin{aligned}
\lambda\alpha_x + (1-\lambda)\alpha_y &\le F(x_\lambda) + F^*(x_\lambda^*) - \lambda F(x) - (1-\lambda)F(y) + \lambda(1-\lambda)\langle x^* - y^*, x - y\rangle_X \\
&\le F^*(x_\lambda^*) + \lambda(1-\lambda)\left(\langle x^* - y^*, x - y\rangle_X - \frac{\gamma}{2}\|x-y\|_X^2\right) \\
&\le F^*(x_\lambda^*) + \lambda(1-\lambda)\sup_{z\in X}\left\{\langle x^* - y^*, z\rangle_X - \frac{\gamma}{2}\|z\|_X^2\right\} \\
&= F^*(x_\lambda^*) + \lambda(1-\lambda)\frac{1}{2\gamma}\|x^* - y^*\|_{X^*}^2,
\end{aligned}$$

where we have used the definition (7.10) of strong convexity in the second inequality and Lemma 5.4 together with Lemma 5.7 (i) in the final equality. Letting now $\alpha_x \to F^*(x^*)$ and $\alpha_y \to F^*(y^*)$, we obtain (7.6) for F^* with $L := \gamma^{-1}$.

(ii): Let $x^*, y^* \in X^*$ and $\lambda \in [0,1]$. Set again $x_\lambda^* := \lambda x^* + (1-\lambda)y^*$. Then we obtain from the definition of the Fenchel conjugate and (7.6) that for any $x, y \in X$,

$$\begin{aligned}
\lambda F^*(x^*) + (1-\lambda)F^*(y^*) &\ge \lambda\left[\langle x^*, x + (1-\lambda)y\rangle_X - F(x + (1-\lambda)y)\right] \\
&\quad + (1-\lambda)\left[\langle y^*, x - \lambda y\rangle_X - F(x - \lambda y)\right] \\
&\ge \lambda\langle x^*, x + (1-\lambda)y\rangle_X + (1-\lambda)\langle y^*, x - \lambda y\rangle_X \\
&\quad - F(x) - \lambda(1-\lambda)\frac{L}{2}\|y\|_X^2 \\
&= \langle x_\lambda^*, x\rangle_X - F(x) + \lambda(1-\lambda)\left(\langle y^* - x^*, y\rangle_X - \frac{L}{2}\|y\|_X^2\right).
\end{aligned}$$

Taking now the supremum over all $x, y \in X$ and using again Lemma 5.4 together with Lemma 5.7 (i), we obtain the strong convexity (7.10) with $\gamma := L^{-1}$.

(iii): One direction of the claim is clear from (ii). For the other direction, if F^* is strongly convex with factor L^{-1}, then its preconjugate $(F^*)_*$ is uniformly smooth with factor L by a proof completely analogous to (i). Then we use Theorem 5.1 to see that $F = F^{**} := (F^*)_*$ under the lower semicontinuity assumption. □

Just as convexity of F implies monotonicity of ∂F, strong convexity has the following consequences.

Lemma 7.4. *Let X be a Banach space and $F : X \to \overline{\mathbb{R}}$. Consider the properties:*

(i) F is strongly convex with factor $\gamma > 0$.

(ii) F is strongly subdifferentiable *with factor γ:*

$$F(y) - F(x) \geq \langle x^*, y - x\rangle_X + \frac{\gamma}{2}\|y - x\|_X^2 \quad \text{for all } x, y \in X;\ x^* \in \partial F(x). \tag{7.11}$$

(iii) ∂F is strongly monotone *with factor γ:*

$$\langle y^* - x^*, y - x\rangle_X \geq \gamma\|y - x\|_X^2 \quad \text{for all } x, y \in X;\ x^* \in \partial F(x),\ y^* \in \partial F(y). \tag{7.12}$$

Then (i) ⇒ (ii) ⇒ (iii). If X is reflexive and F is proper, convex, and lower semicontinuous, then also (iii) ⇒ (i).

Proof. *(i) ⇒ (ii):* Let $x, y \in X$ and $\lambda \in (0, 1)$ be arbitrary. Dividing (7.10) by λ and rearranging yields

$$\frac{F(y + \lambda(x - y)) - F(y)}{\lambda} \leq F(x) - F(y) - (1 - \lambda)\frac{\gamma}{2}\|x - y\|_X^2.$$

Since strongly convex functions are also convex, we can apply Lemma 4.3 (ii) to pass to the limit $\lambda \to 0$ on both sides to obtain

$$F'(y, x - y) \leq F(x) - F(y) - \frac{\gamma}{2}\|x - y\|_X^2.$$

Using Lemma 4.4 for $h = x - y$, we thus obtain that for any $y^* \in \partial F(y)$,

$$\langle y^*, x - y\rangle_X \leq F'(y, x - y) \leq F(x) - F(y) - \frac{\gamma}{2}\|x - y\|_X^2.$$

Exchanging the roles of x and y and rearranging yields (7.11).

(ii) ⇒ (iii): Adding (7.11) with the same inequality with x and y exchanged immediately yields (7.12).

(iii) ⇒ (i): Suppose first that ∂F is surjective. Then $\operatorname{dom} \partial F^* = X^*$. Using the duality between ∂F and ∂F^* in Lemma 5.8, we rewrite (7.12) as

$$\langle y^* - x^*, y - x\rangle_X \geq \gamma\|y - x\|_X^2 \quad \text{for all } x^*, y^* \in X^*;\ x \in \partial F^*(x^*),\ y \in \partial F^*(y^*). \tag{7.13}$$

Taking $y = x$, this implies that $x^* = y^*$, i.e., $\partial F^*(x^*)$ is a singleton for all $x^* \in X^*$. Here we use that $\operatorname{dom} \partial F^* = X^*$ to avoid the possibility that $\partial F^*(x^*) = \emptyset$. By Theorem 4.5 it follows that F^* is Gâteaux differentiable. Thus (7.13) describes the co-coercivity (7.2) of DF^* with factor γ. By Lemma 7.1 it follows that F^* is uniformly smooth with factor γ^{-1}. Consequently, by Theorem 7.3 F is strongly convex with factor γ.

If ∂F is not surjective, we replace F by $F + \varepsilon j$ for the duality mapping $j(x) := \frac{1}{2}\|x\|_X^2$ and some $\varepsilon > 0$. By Theorem 6.13 and Minty's Theorem 6.14 now $\partial(F + \varepsilon j)$ is surjective. It also remains strongly monotone with factor γ as ∂j is monotone. Now, by the above reasoning, $F + \varepsilon j$ is strongly convex with factor γ. Since $\varepsilon > 0$ was arbitrary, we deduce from the defining (7.10) that F is strongly convex with factor γ. □

Note that the factor γ enters into the strong monotonicity (7.12) directly rather than as $\frac{\gamma}{2}$ as in the strong subdifferentiability (7.11) (and strong convexity).

We can also derive a stronger, quantitative, version of the fact that for convex functions, points that satisfy the Fermat principle are minimizers.

Lemma 7.5. *Let X be a Banach space and let $F : X \to \overline{\mathbb{R}}$ be strongly convex with factor $\gamma > 0$. Assume that F admits a minimum $M := \min_{x \in X} F(x)$. Then the* Polyak–Łojasewicz inequality *holds:*

$$F(x) - M \leq \frac{1}{2\gamma}\|x^*\|_{X^*}^2 \quad \text{for all } x \in X,\ x^* \in \partial F(x). \tag{7.14}$$

Proof. Let $x \in X$ and $x^* \in \partial F(x)$ be arbitrary. Then from Lemma 7.4 (ii) we have that

$$-F(x) + \langle x^*, x - y\rangle_X - \frac{\gamma}{2}\|x - y\|_X^2 \geq -F(y).$$

Taking the supremum over all $y \in X$, noting that this is equivalent to taking the supremum over all $x - y \in X$, and inserting the Fenchel conjugate of the squared norm from Lemma 5.4 together with Lemma 5.7 (i), we obtain

$$-F(x) + \frac{1}{2\gamma}\|x^*\|_{X^*}^2 \geq \sup_{y \in X} -F(y) = -\min_{y \in X} F(y)$$

and hence, after rearranging, (7.14). □

Comparing the consequences of strong convexity in Lemma 7.4 and those of uniform smoothness in Lemma 7.1, we can already see a certain duality between them: While the former give lower bounds, the latter give upper bounds and vice versa. A simple example is the following

Corollary 7.6. *If $F : X \to \mathbb{R}$ is strongly convex with factor γ and uniformly smooth with factor L, then*

$$\gamma\|x-y\|_X^2 \le \langle DF(x)-DF(y), x-y\rangle_X \le L\|x-y\|_X^2 \quad \text{for all } x, y \in X. \tag{7.15}$$

Proof. The first inequality follows from Lemma 7.4 (iii), while the second follows from (1.1) together with Lemma 7.1 (iii). □

The estimates of Corollary 7.2 can be improved if F is in addition strongly convex.

Corollary 7.7. *Let X be a Banach space and let $F : X \to \mathbb{R}$ be strongly convex with factor $\gamma > 0$ as well as Lipschitz differentiable with constant $L > 0$. Then for any $\alpha > 0$,*

$$\langle DF(z), x-\widehat{x}\rangle_X \ge F(x) - F(\widehat{x}) + \frac{\gamma-\alpha L}{2}\|x-\widehat{x}\|_X^2 - \frac{L}{2\alpha}\|x-z\|_X^2 \quad \text{for all } \widehat{x}, z, x \in X, \tag{7.16}$$

as well as

$$\langle DF(z)-DF(\widehat{x}), x-\widehat{x}\rangle_X \ge (\gamma-\alpha L)\|x-\widehat{x}\|_X^2 - \frac{L}{4\alpha}\|x-z\|_X^2 \quad \text{for all } \widehat{x}, z, x \in X. \tag{7.17}$$

Proof. Using the strong subdifferentiability from Lemma 7.4 (ii), the Lipschitz continuity of DF, (1.1), and Young's inequality, we obtain

$$\begin{aligned}
\langle DF(z), x-\widehat{x}\rangle_X &= \langle DF(x), x-\widehat{x}\rangle_X + \langle DF(z)-DF(x), x-\widehat{x}\rangle_X \\
&\ge F(x) - F(\widehat{x}) + \frac{\gamma}{2}\|x-\widehat{x}\|_X^2 - \frac{\alpha L}{2}\|x-\widehat{x}\|_X^2 - \frac{1}{2\alpha L}\|DF(z)-DF(x)\|_{X^*}^2 \\
&\ge F(x) - F(\widehat{x}) + \frac{\gamma}{2}\|x-\widehat{x}\|_X^2 - \frac{\alpha L}{2}\|x-\widehat{x}\|_X^2 - \frac{L}{2\alpha}\|x-z\|_X^2.
\end{aligned}$$

For (7.17), we can use the strong monotonicity of DF from Lemma 7.4 (iii) to estimate analogously

$$\begin{aligned}
\langle DF(z)-DF(\widehat{x}), x-\widehat{x}\rangle_X &= \langle DF(x)-DF(\widehat{x}), x-\widehat{x}\rangle_X + \langle DF(z)-DF(x), x-\widehat{x}\rangle_X \\
&\ge \gamma\|x-\widehat{x}\|_X^2 - \alpha L\|x-\widehat{x}\|_X^2 - \frac{L}{4\alpha}\|x-z\|_X^2.
\end{aligned}$$

□

7.3 MOREAU–YOSIDA REGULARIZATION

We now look at another way to reformulate optimality conditions using proximal point mappings. Although these are no longer equivalent reformulations, they will serve as a link to the Newton-type methods which will be introduced in Chapter 14.

We again assume that X is a Hilbert space and identify X^* with X via the Riesz isomorphism. Let $A : X \rightrightarrows X$ be a maximally monotone operator with graph $A \neq \emptyset$ and $\gamma > 0$. Then we define the *Yosida approximation* of A as

$$A_\gamma := \frac{1}{\gamma}\left(\mathrm{Id} - \mathcal{R}_{\gamma A}\right).$$

In particular, the Yosida approximation of the subdifferential of a proper, convex, and lower semicontinuous functional $F : X \to \overline{\mathbb{R}}$ is given by

$$(\partial F)_\gamma := \frac{1}{\gamma}\left(\mathrm{Id} - \mathrm{prox}_{\gamma F}\right), \tag{7.18}$$

which by Corollary 6.16 and Theorem 6.13 is always Lipschitz continuous with constant $L = \gamma^{-1}$.

An alternative point of view is the following. For a proper, convex, and lower semicontinuous functional $F : X \to \overline{\mathbb{R}}$ and $\gamma > 0$, we define the *Moreau envelope*[1]

$$F_\gamma : X \to \mathbb{R}, \qquad x \mapsto \inf_{z \in X} \frac{1}{2\gamma}\|z - x\|_X^2 + F(z), \tag{7.19}$$

see Figure 7.1. Comparing this with the definition (6.12) of the proximal point mapping of F, we see that

$$F_\gamma(x) = \frac{1}{2\gamma}\|\mathrm{prox}_{\gamma F}(x) - x\|_X^2 + F(\mathrm{prox}_{\gamma F}(x)). \tag{7.20}$$

(Note that multiplying a functional by $\gamma > 0$ does not change its minimizers.) Hence F_γ is indeed well-defined on X and single-valued. Furthermore, we can deduce from (7.20) that F_γ is convex as well.

Lemma 7.8. *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $\gamma > 0$. Then F_γ is convex.*

Proof. We first show that for any convex $G : X \to \overline{\mathbb{R}}$, the mapping

$$H : X \times X \to \overline{\mathbb{R}}, \qquad (x, z) \mapsto F(z) + G(z - x)$$

is convex as well. Indeed, for any $(x_1, z_1), (x_2, z_2) \in X \times X$ and $\lambda \in [0, 1]$, the convexity of F and G implies that

$$\begin{aligned} H(\lambda(x_1, z_1) + (1-\lambda)(x_2, z_2)) &= F\left(\lambda z_1 + (1-\lambda)z_2\right) + G\left(\lambda(z_1 - x_1) + (1-\lambda)(z_2 - x_2)\right) \\ &\leq \lambda\left(F(z_1) + G(z_1 - x_1)\right) + (1-\lambda)\left(F(z_2) + G(z_2 - x_2)\right) \\ &= \lambda H(x_1, z_1) + (1-\lambda)H(x_2, z_2). \end{aligned}$$

[1] not to be confused with the *convex* envelope F^Γ!

Let now $x_1, x_2 \in X$ and $\lambda \in [0,1]$. Since $F_\gamma(x) = \inf_{z\in X} H(x,z)$ for $G(y) := \frac{1}{2\gamma}\|y\|_X^2$, there exist two minimizing sequences $\{z_n^1\}_{n\in\mathbb{N}}, \{z_n^2\}_{n\in\mathbb{N}} \subset X$ with

$$H(x_1, z_n^1) \to F_\gamma(x_1), \qquad H(x_2, z_n^2) \to F_\gamma(x_2).$$

From the properties of the infimum together with the convexity of H, we thus obtain for all $n \in \mathbb{N}$ that

$$\begin{aligned} F_\gamma(\lambda x_1 + (1-\lambda)x_2) &\le H(\lambda(x_1, z_n^1) + (1-\lambda)(x_2, z_n^2)) \\ &\le \lambda H(x_1, z_n^1) + (1-\lambda)H(x_2, z_n^2), \end{aligned}$$

and passing to the limit $n \to \infty$ yields the desired convexity. □

We will also show later that Moreau–Yosida regularization preserves (global!) Lipschitz continuity.

The next theorem links the two concepts of Moreau envelope and of Yosida approximation and hence justifies the term *Moreau–Yosida regularization.*

Theorem 7.9. *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $\gamma > 0$. Then F_γ is Fréchet differentiable with*

$$\nabla(F_\gamma) = (\partial F)_\gamma.$$

Proof. Let $x, y \in X$ be arbitrary and set $x^* = \operatorname{prox}_{\gamma F}(x)$ and $y^* = \operatorname{prox}_{\gamma F}(y)$. We first show that

$$\frac{1}{\gamma}\langle y^* - x^*, x - x^*\rangle_X \le F(y^*) - F(x^*). \tag{7.21}$$

(Note that for proper F, the definition of proximal points as minimizers necessarily implies that $x^*, y^* \in \operatorname{dom} F$.) To this purpose, consider for $t \in (0,1)$ the point $x_t^* := ty^* + (1-t)x^*$. Using the minimizing property of the proximal point x^* together with the convexity of F and completing the square, we obtain that

$$\begin{aligned} F(x^*) &\le F(x_t^*) + \frac{1}{2\gamma}\|x_t^* - x\|_X^2 - \frac{1}{2\gamma}\|x^* - x\|_X^2 \\ &\le tF(y^*) + (1-t)F(x^*) - \frac{t}{\gamma}\langle x - x^*, y^* - x^*\rangle_X + \frac{t^2}{2\gamma}\|x^* - y^*\|_X^2. \end{aligned}$$

Rearranging the terms, dividing by $t > 0$ and passing to the limit $t \to 0$ then yields (7.21). Combining this with (7.20) implies that

$$\begin{aligned} F_\gamma(y) - F_\gamma(x) &= F(y^*) - F(x^*) + \frac{1}{2\gamma}\left(\|y - y^*\|_X^2 - \|x - x^*\|_X^2\right) \\ &\ge \frac{1}{2\gamma}\left(2\langle y^* - x^*, x - x^*\rangle_X + \|y - y^*\|_X^2 - \|x - x^*\|_X^2\right) \\ &= \frac{1}{2\gamma}\left(2\langle y - x, x - x^*\rangle_X + \|y - y^* - x + x^*\|_X^2\right) \\ &\ge \frac{1}{\gamma}\langle y - x, x - x^*\rangle_X. \end{aligned}$$

By exchanging the roles of x^* and y^* in (7.21), we obtain that

$$F_\gamma(y) - F_\gamma(x) \leq \frac{1}{\gamma}\langle y - x, y - y^*\rangle_X.$$

Together, these two inequalities yield that

$$\begin{aligned} 0 &\leq F_\gamma(y) - F_\gamma(x) - \frac{1}{\gamma}\langle y - x, x - x^*\rangle_X \\ &\leq \frac{1}{\gamma}\langle y - x, (y - y^*) - (x - x^*)\rangle_X \\ &\leq \frac{1}{\gamma}\left(\|y - x\|_X^2 - \|y^* - x^*\|_X^2\right) \\ &\leq \frac{1}{\gamma}\|y - x\|_X^2, \end{aligned}$$

where the next-to-last inequality follows from the firm nonexpansivity of proximal point mappings (Lemma 6.15).

If we now set $y = x + h$ for arbitrary $h \in X$, we obtain that

$$0 \leq \frac{F_\gamma(x + h) - F_\gamma(x) - \langle \gamma^{-1}(x - x^*), h\rangle_X}{\|h\|_X} \leq \frac{1}{\gamma}\|h\|_X \to 0 \qquad \text{for } h \to 0,$$

i.e., F_γ is Fréchet differentiable with gradient $\frac{1}{\gamma}(x - x^*) = (\partial F)_\gamma(x)$. □

Since F_γ is convex by Lemma 7.8, this result together with Theorem 4.5 yields the catchy relation $\partial(F_\gamma) = (\partial F)_\gamma$.

Example 7.10. We consider again $X = \mathbb{R}^N$.

(i) For $F(x) = \frac{1}{2}\|x\|_2^2$, Example 6.26 (ii) yields $\operatorname{prox}_{\gamma F}(x) = \frac{1}{1+\gamma}x$. Inserting this into the definition of the Yosida approximation and the Moreau envelope and simplifying yields that

$$(\partial F)_\gamma(x) = \frac{1}{\gamma}\left(x - \frac{1}{1+\gamma}\right)x = \frac{1}{1+\gamma}x$$

and

$$F_\gamma(x) = \frac{1}{2\gamma}\|\tfrac{1}{1+\gamma}x\|_2^2 + \frac{1}{2}\|\tfrac{1}{1+\gamma}x\|_2^2 = \frac{1}{2(1+\gamma)}\|x\|_2^2.$$

(Unsurpisingly, the Moreau envelope of a quadratic function remains quadratic and is simply scaled.)

(ii) For $F(x) = \|x\|_1$, we have from Example 6.26 (ii) that the proximal point mapping is given by the componentwise soft-shrinkage operator. Inserting this into the

definition yields that

$$\big[(\partial\|\cdot\|_1)_\gamma(x)\big]_i = \begin{cases} \frac{1}{\gamma}(x_i - (x_i - \gamma)) = 1 & \text{if } x_i > \gamma, \\ \frac{1}{\gamma}x_i & \text{if } x_i \in [-\gamma, \gamma], \\ \frac{1}{\gamma}(x_i - (x_i + \gamma)) = -1 & \text{if } x_i < -\gamma. \end{cases}$$

Comparing this to the corresponding subdifferential (4.2), we see that the set-valued case in the point $x_i = 0$ has been replaced by a linear function on a small interval.

Similarly, inserting the definition of the proximal point into (7.20) shows that

$$F_\gamma(x) = \sum_{i=1}^N f_\gamma(x_i) \text{ for } f_\gamma(t) = \begin{cases} \frac{1}{2\gamma}|t - (t-\gamma)|^2 + |t-\gamma| = t - \frac{\gamma}{2} & \text{if } t > \gamma, \\ \frac{1}{2\gamma}|t|^2 & \text{if } t \in [-\gamma, \gamma], \\ \frac{1}{2\gamma}|t - (t+\gamma)|^2 + |t+\gamma| = -t - \frac{\gamma}{2} & \text{if } t < -\gamma. \end{cases}$$

For small values, the absolute value is thus replaced by a quadratic function (which removes the nondifferentiability at 0). This modification is well-known under the name *Huber norm*; see Figure 7.1a.

(iii) For $F(x) = \delta_{\mathbb{B}_\infty}(x)$, we have from Example 6.26 (iii) that the proximal mapping is given by the componentwise projection onto $[-1, 1]$ and hence that

$$\big[(\partial\delta_{\mathbb{B}_\infty})_\gamma(x)\big]_i = \frac{1}{\gamma}\Big(x_i - \big(x_i - (x_i-1)^+ - (x_i+1)^-\big)\Big) = \frac{1}{\gamma}(x_i-1)^+ + \frac{1}{\gamma}(x_i+1)^-.$$

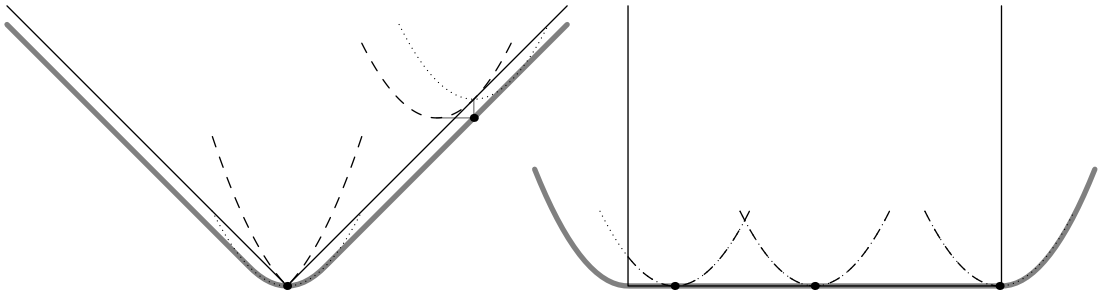
Similarly, inserting this and using that $\operatorname{prox}_{\gamma F}(x) \in \mathbb{B}_\infty$ and $\langle (x+1)^-, (x-1)^+\rangle_X = 0$ yields that

$$(\delta_{\mathbb{B}_\infty})_\gamma(x) = \frac{1}{2\gamma}\|(x-1)^+\|_2^2 + \frac{1}{2\gamma}\|(x+1)^-\|_2^2,$$

which corresponds to the classical penalty functional for the inequality constraints $x - 1 \le 0$ and $x + 1 \ge 0$ in nonlinear optimization; see Figure 7.1b.

Using Corollary 6.27, analogous characterizations can be derived for the squared L^2-norm, the L^1-norm, and the indicator functional of the L^∞-ball on $L^2(\Omega)$.

By Theorem 7.9, F_γ is Fréchet differentiable with Lipschitz continuous gradient with factor γ^{-1}. From Theorem 7.3, we thus know that F_γ^* is strongly convex with factor γ, which in Hilbert spaces is equivalent to $F_\gamma^* - \frac{\gamma}{2}\|\cdot\|_X^2$ being convex. In fact, this can be made even more explicit.



(a) $f(t) = |t|$ (b) $f(t) = \delta_{[-1,1]}(t)$

Figure 7.1: Illustration of the Moreau–Yosida regularization (thick solid line) of F (thin solid line). The dotted line indicates the quadratic function $z \mapsto \frac{1}{2\gamma}\|x - z\|_X^2$, while the dashed line is $z \mapsto F(z) + \frac{1}{2\gamma}\|x - z\|_X^2$. The dots and the horizontal and vertical lines (nontrivial only in the second point of (a)) emanating from the dots indicate the pair $(x, F_\gamma(x))$ and how it relates to the minimization of the shifted quadratic functional. (In (b) the two lines are overlaid within $[-1, 1]$, as only the domain of definition of the two functions is different.)

Theorem 7.11. *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. Then we have for all $\gamma > 0$ that*

$$(F_\gamma)^* = F^* + \frac{\gamma}{2}\|\cdot\|_X^2.$$

Proof. We obtain directly from the definition of the Fenchel conjugate in Hilbert spaces and of the Moreau envelope that

$$\begin{aligned}
(F_\gamma)^*(x^*) &= \sup_{x\in X}\left\{\langle x^*, x\rangle_X - \inf_{z\in X}\left[\tfrac{1}{2\gamma}\|x - z\|_X^2 + F(z)\right]\right\} \\
&= \sup_{x\in X}\left\{\langle x^*, x\rangle_X + \sup_{z\in X}\left\{-\tfrac{1}{2\gamma}\|x - z\|_X^2 - F(z)\right\}\right\} \\
&= \sup_{z\in X}\left\{\langle x^*, z\rangle_X - F(z) + \sup_{x\in X}\left\{\langle x^*, x - z\rangle_X - \tfrac{1}{2\gamma}\|x - z\|_X^2\right\}\right\} \\
&= F^*(x^*) + \left(\tfrac{1}{2\gamma}\|\cdot\|_X^2\right)^*(x^*),
\end{aligned}$$

since for any given $z \in X$, the inner supremum is always taken over the full space X. The claim now follows from Lemma 5.4 with $p = 2$ (using again the fact that we have identified X^* with X) and Lemma 5.7 (i). □

This immediately yields a Moreau decomposition of the envelope; cf. Lemma 6.24 (ii).

Corollary 7.12. *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. Then for all $x \in X$ and $\gamma > 0$,*

$$\frac{1}{2\gamma}\|x\|_X^2 = F_\gamma(x) + (F^*)_{\gamma^{-1}}(\gamma^{-1}x).$$

Proof. By definition of the Moreau envelope, we have that

$$\begin{aligned} F_\gamma(x) &= \inf_{z\in X} F(z) + \frac{1}{2\gamma}\|x-z\|_X^2 \\ &= \inf_{z\in X} F(z) + \frac{1}{2\gamma}\|x\|_X^2 - \frac{1}{\gamma}(x \mid z)_X + \frac{1}{2\gamma}\|z\|_X^2 \\ &= \frac{1}{2\gamma}\|x\|_X^2 - \sup_{z\in X}\left\{(\gamma^{-1}x \mid z)_X - F(z) - \frac{1}{2\gamma}\|z\|_X^2\right\} \\ &= \frac{1}{2\gamma}\|x\|_X^2 - \left(F + \frac{1}{2\gamma}\|\cdot\|_X^2\right)^*(\gamma^{-1}x). \end{aligned}$$

The claim now follows since F (by assumption) and F_γ (by Lemma 7.8 and Theorem 7.9) are convex and lower semicontinous, and hence Theorem 5.1 (iii) together with Theorem 7.11 implies

$$\left(F + \frac{1}{2\gamma}\|\cdot\|_X^2\right)^* = \left(F^{**} + \frac{1}{2\gamma}\|\cdot\|_X^2\right)^* = \left((F^*)_{\gamma^{-1}}\right)^{**} = (F^*)_{\gamma^{-1}}. \qquad \square$$

Taking the derivative of this identity and using Theorem 7.9 together with (7.18), we again obtain Lemma 6.24 (ii).

With the help of Theorem 7.11, we can also show the converse of Theorem 7.9: every smooth function can be obtained through Moreau–Yosida regularization.

Corollary 7.13. *Let $F : X \to \overline{\mathbb{R}}$ be convex and L-smooth. Then for all $x \in X$,*

$$F(x) = (G^*)_{L^{-1}}(x) \quad \text{and} \quad \nabla F(x) = \operatorname{prox}_{LG}(Lx)$$

for

$$G : X \to \overline{\mathbb{R}}, \qquad G(x) = F^*(x) - \frac{1}{2L}\|x\|_X^2.$$

Proof. Since F is convex and L-smooth and X is a Hilbert space, Lemma 7.1 and Theorem 7.3 yields that F^* is strongly convex with factor L^{-1} and thus that G is convex. Furthermore, as a Fenchel conjugate of a proper convex functional, F^* and thus G is proper and lower semicontinuous. Theorems 5.1 and 7.11 now imply that for all $x \in X$,

$$(G^*)_{L^{-1}}(x) = (G^*)^{**}_{L^{-1}}(x) = \left(G + \frac{1}{2L}\|\cdot\|_X^2\right)^*(x) = F^{**}(x) = F(x).$$

Furthermore, by Lemma 4.13 and Theorems 4.5 and 4.14, we have that

$$\partial G(z) = \partial F^*(z) - \{L^{-1}z\} \qquad \text{for all } z \in X.$$

By the definition of the proximal mapping, this is equivalent to $z = \operatorname{prox}_{LG} Lx$ for any $x \in \partial F^*(z)$. But by Lemma 5.8, $x \in \partial F^*(z)$ holds if and only if $z \in \partial F(x) = \{\nabla F(x)\}$, and combining these two yields the first expression for the gradient. $\square$

Let us briefly consider the relevance of the previous results to optimization.

Approximation by smooth mappings For a convex functional $F : X \to \overline{\mathbb{R}}$, every minimizer $\bar{x} \in X$ satisfies the Fermat principle $0 \in \partial F(\bar{x})$, which we can write equivalently as $\bar{x} \in \partial F^*(0)$. If we now replace ∂F^* with its Yosida approximation $(\partial F^*)_\gamma$, we obtain the regularized optimality condition

$$x_\gamma = (\partial F^*)_\gamma(0) = -\frac{1}{\gamma}\operatorname{prox}_{\gamma F^*}(0).$$

This is now an *explicit* and even Lipschitz continuous relation. Although x_γ is no longer a minimizer of F, the convexity of F_γ implies that $x_\gamma \in (\partial F^*)_\gamma(0) = \partial(F_\gamma^*)(0)$ is equivalent to

$$0 \in \partial(F_\gamma^*)^*(x_\gamma) = \partial\left(F^{**} + \tfrac{\gamma}{2}\|\cdot\|_X^2\right)(x_\gamma) = \partial\left(F + \tfrac{\gamma}{2}\|\cdot\|_X^2\right)(x_\gamma),$$

i.e., x_γ is the (unique due to the strict convexity of the squared norm) minimizer of the functional $F + \frac{\gamma}{2}\|\cdot\|_X^2$. Hence, the regularization of ∂F^* has not made the original problem smooth but merely (more) strongly convex. The equivalence can also be used to show (similarly to the proof of Theorem 2.1) that $x_\gamma \rightharpoonup \bar{x}$ for $\gamma \to 0$. In practice, this straightforward approach fails due to the difficulty of computing F^* and $\operatorname{prox}_{F^*}$ and is therefore usually combined with one of the splitting techniques that will be introduced in the next chapter.

Conversion between gradients and proximal mappings According to Corollary 7.13, solving $\min_x F(x)$ for an L-smooth function F is equivalent to solving

$$\min_{x,\tilde{x}\in X} G^*(x) + \frac{1}{2L}\|x - \tilde{x}\|_X^2.$$

Observe that G^* may be nonsmooth. Suppose we apply an algorithm for the latter that makes use of the proximal mapping of G^* (such as the splitting methods that will be discussed in the following chapters). Then using the Moreau decomposition of Lemma 6.24 (ii) with Corollary 7.13, we see that

$$\operatorname{prox}_{L^{-1}G^*}(x) = x - L^{-1}\nabla F(x).$$

Therefore, this can still be done purely in terms of the gradient evaluations of F.

Remark 7.14. Continuing from Remark 6.29, Moreau–Yosida regularization can also be defined in reflexive Banach spaces; we refer to [Brezis et al., 1970] for details. Again, the main issue is the practical evaluation of F_γ and $(\partial F)_\gamma$ if the duality mapping is no longer the identity.

8 PROXIMAL POINT AND SPLITTING METHODS

We now turn to the development of algorithms for computing minimizers of functionals $J : X \to \overline{\mathbb{R}}$ of the form

$$J(x) \coloneqq F(x) + G(x)$$

for $F, G : X \to \overline{\mathbb{R}}$ convex but not necessarily differentiable. One of the main difficulties compared to the differentiable setting is that the naive equivalent to steepest descent, the iteration

$$x^{k+1} \in x^k - \tau_k \partial J(x^k),$$

does not work since even in finite dimensions, arbitrary subgradients need not be descent directions – this can only be guaranteed for the subgradient of minimal norm; see, e.g., [Ruszczyński, 2006, Example 7.1, Lemma 2.77]. Furthermore, the minimal norm subgradient of J cannot be computed easily from those of F and G. We thus follow a different approach and look for a root $\widehat{x}$ of the set-valued mapping $x \mapsto \partial J(x)$ (which coincides with the minimizer $\bar{x}$ of J if J is convex). In this chapter, we only derive methods, postponing proofs of convergence, in various different senses, to Chapters 9 to 11. For the reasons mentioned in the beginning of Section 6.3, we will assume in this and the following chapters that X (as well as all further occurring spaces) is a Hilbert space so that we can identify $X^* \cong X$.

8.1 PROXIMAL POINT METHOD

We have seen in Corollary 6.22 that a root $\widehat{x}$ of $\partial J : X \rightrightarrows X$ can be characterized as a fixed point of $\operatorname{prox}_{\tau J}$ for any $\tau > 0$. This suggests a fixed-point iteration: Choose $x^0 \in X$ and for an appropriate sequence $\{\tau_k\}_{k \in \mathbb{N}}$ of step sizes set

$$x^{k+1} = \operatorname{prox}_{\tau_k J}(x^k). \tag{8.1}$$

This iteration naturally generalizes to finding a root $\widehat{x} \in A^{-1}(0)$ of a set-valued (usually monotone) operator $A : X \rightrightarrows X$ as

$$x^{k+1} = \mathcal{R}_{\tau_k A}(x^k). \tag{8.2}$$

This is the *proximal point method*, which is the basic building block for all methods in this chapter. Using the definition of the resolvent, this can also be written in implicit form as

$$0 \in \tau_k A(x^{k+1}) + (x^{k+1} - x^k), \tag{8.3}$$

which will be useful for the analysis of the method.

If A is maximal monotone (in particular if $A = \partial J$), Lemma 6.15 shows that the iteration map $x \mapsto \mathcal{R}_{\tau_k A}(x)$ is firmly nonexpansive. Mere (nonfirm) nonexpansivity already implies that

$$\|x^{k+1} - \widehat{x}\|_X = \|\mathcal{R}_{\tau_k A}(x^k) - \widehat{x}\|_X \leq \|x^k - \widehat{x}\|_X.$$

In other words, the method does not escape from a fixed point. Either a more refined analysis based on firm nonexpansivity of the iteration map or a more direct analysis based on the maximal monotonicity of A can be used to further show that the iterates $\{x^k\}_{k\in\mathbb{N}}$ indeed converge to a fixed point $\widehat{x}$ for an initial iterate x^0. The latter will be the topic of Chapter 9.

A practical issue is the steps (8.1) of the basic proximal point method are typically just as difficult as the original problem, so the method is not feasible for problems that demand an iterative method for their solution in the first place. However, the proximal step does form an important building block of several more practical *splitting* methods for problems of the form $J = F + G$, which we derive in the following by additional clever manipulations.

Remark 8.1. The proximal point algorithm can be traced back to Krasnosel′skiĭ [Krasnosel′skiĭ, 1955] and Mann [Mann, 1953] (as a special case of the *Krasnosel′skiĭ–Mann iteration*); it was also studied in [Martinet, 1970]. The formulation considered here was proposed in [Rockafellar, 1976b].

8.2 EXPLICIT SPLITTING: FORWARD-BACKWARD SPLITTING

As we have noted, the proximal point method is not feasible for most functionals of the form $J(x) = F(x) + G(x)$, since the evaluation of prox_J is not significantly easier than solving the original minimization problem – even if prox_F and prox_G have a closed-form expression. (Such functionals are called *prox-simple*). We thus proceed differently: instead of applying the proximal point reformulation directly to $0 \in \partial J(\widehat{x})$, we first apply the subdifferential sum rule (Theorem 4.14) to deduce the existence of $\widehat{p} \in X$ with

$$\begin{cases} \widehat{p} \in \partial F(\widehat{x}), \\ -\widehat{p} \in \partial G(\widehat{x}). \end{cases} \tag{8.4}$$

We can now replace one or both of these subdifferential inclusions by a proximal point reformulation that only involves F or G.

Explicit splitting methods – also known as *forward-backward splitting* – are based on applying Lemma 6.21 only to, e.g., the second inclusion in (8.4) to obtain

$$\begin{cases} \widehat{p} \in \partial F(\widehat{x}), \\ \widehat{x} = \operatorname{prox}_{\tau G}(\widehat{x} - \tau\widehat{p}). \end{cases} \tag{8.5}$$

The corresponding fixed-point iteration then consists in

(i) choosing $p^k \in \partial F(x^k)$ (with minimal norm);

(ii) setting $x^{k+1} = \operatorname{prox}_{\tau_k G}(x^k - \tau_k p^k)$.

Again, computing a subgradient with minimal norm can be complicated in general. It is, however, easy if F is additionally differentiable since in this case $\partial F(x) = \{\nabla F(x)\}$ by Theorem 4.5. This leads to the *proximal gradient* or *forward-backward splitting method*

$$x^{k+1} = \operatorname{prox}_{\tau_k G}(x^k - \tau_k \nabla F(x^k)). \tag{8.6}$$

(The special case $G = \delta_C$ – i.e., $\operatorname{prox}_{\tau G}(x) = \operatorname{proj}_C(x)$ – is also known as the *projected gradient method*). Similarly to the proximal point method, this method can be written in implicit form as

$$0 \in \tau_k[\partial G(x^{k+1}) + \nabla F(x^k)] + (x^{k+1} - x^k). \tag{8.7}$$

Based on this, we will see in Chapter 9 that the iterates $\{x^k\}$ converge weakly if $\tau_k L < 2$ for L the Lipschitz factor of ∇F. The need to know L is one drawback of the explicit splitting method. This can to some extend be circumvented by performing a line search, i.e., testing for various choices of τ_k until a sufficient decrease in function values is achieved. We will discuss such strategies later on in Section 12.3. Another highly successful variant of explicit splitting applies *inertia* to the iterates for faster convergence; this we will discuss in Section 12.2 after developing tools for the study of convergence rates.

Remark 8.2. Forward-backward splitting for finding the root of the sum of two monotone operators was already proposed in [Lions and Mercier, 1979]. It has become especially popular under the name *iterative soft-thresholding* (ISTA) in the context of *sparse regression* (i.e., regularization of linear inverse problems with ℓ^1 penalties), see, e.g., [Chambolle et al., 1998; Daubechies et al., 2004; Wright et al., 2009].

8.3 IMPLICIT SPLITTING: DOUGLAS–RACHFORD SPLITTING

Even with a line search, the restriction on the step sizes τ_k in explicit splitting remain unsatisfactory. Such restrictions are not needed in implicit splitting methods. (Compare the properties of explicit vs. implicit Euler methods for differential equations.) Here, the proximal point formulation is applied to both subdifferential inclusions in (8.4), which yields the optimality conditions

$$\begin{cases} \widehat{x} = \operatorname{prox}_{\tau F}(\widehat{x} + \tau\widehat{p}), \\ \widehat{x} = \operatorname{prox}_{\tau G}(\widehat{x} - \tau\widehat{p}). \end{cases}$$

To eliminate $\widehat{p}$ from these equations, we set $\widehat{z} := \widehat{x} + \tau\widehat{p}$ and $\widehat{w} := \widehat{x} - \tau\widehat{p} = 2\widehat{x} - \widehat{z}$. It remains to derive a recursion for $\widehat{z}$, which we obtain from the productive zero $\widehat{z} = \widehat{z} + (\widehat{x} - \widehat{x})$.

Further replacing some copies of $\widehat{x}$ by a new variable $\widehat{y}$ leads to the overall fixed point system

$$\begin{cases} \widehat{x} = \operatorname{prox}_{\tau F}(\widehat{z}), \\ \widehat{y} = \operatorname{prox}_{\tau G}(2\widehat{x} - \widehat{z}), \\ \widehat{z} = \widehat{z} + \widehat{x} - \widehat{y}. \end{cases}$$

The corresponding fixed-point iteration leads to the *Douglas–Rachford splitting* (DRS) method

$$\begin{cases} x^{k+1} = \operatorname{prox}_{\tau F}(z^k), \\ y^{k+1} = \operatorname{prox}_{\tau G}(2x^{k+1} - z^k), \\ z^{k+1} = z^k + y^{k+1} - x^{k+1}. \end{cases} \tag{8.8}$$

Of course, the algorithm and its derivation generalize to arbitrary monotone operators $A, B : X \rightrightarrows X$:

$$\begin{cases} x^{k+1} = \mathcal{R}_{\tau B}(z^k), \\ y^{k+1} = \mathcal{R}_{\tau A}(2x^{k+1} - z^k), \\ z^{k+1} = z^k + y^{k+1} - x^{k+1}. \end{cases} \tag{8.9}$$

We can also write the DRS method in more implicit form. Indeed, inverting the resolvents in (8.9) and using the last update to change variables in the first two yields

$$\begin{cases} 0 \in \tau B(x^{k+1}) + y^{k+1} - z^{k+1}, \\ 0 \in \tau A(y^{k+1}) + z^{k+1} - x^{k+1}, \\ 0 = x^{k+1} - y^{k+1} + (z^{k+1} - z^k). \end{cases}$$

Therefore, with $u := (x, y, z) \in X^3$, and the operators[1]

$$H(x, y, z) := \begin{pmatrix} \tau B(x) + y - z \\ \tau A(y) + z - x \\ x - y \end{pmatrix} \quad \text{and} \quad M := \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \mathrm{Id} \end{pmatrix}, \tag{8.10}$$

we can write the DRS method as the *preconditioned proximal point method*

$$0 \in H(u^{k+1}) + M(u^{k+1} - u^k). \tag{8.11}$$

Indeed, the basic proximal point in implicit form (8.3) is just (8.11) with the *preconditioner* $M = \tau^{-1}\mathrm{Id}$. It is furthermore straightforward to verify that $0 \in H(\widehat{u})$ is equivalent to $0 \in A(\widehat{x}) + B(\widehat{x})$.

[1] Here and in the following, we identify $x \in X$ with the singleton set $\{x\} \subset X$ whenever there is no danger of confusion.

The formulation (8.11) will in the following chapter form the basis for proving the convergence of the method. Recalling the discussion on convergence in Section 8.1, it seems beneficial for H to be maximally monotone, as then (although this is not immediate from Lemma 6.15) it is reasonable to expect the nonexpansivity of the iterates with respect to the semi-norm $u \mapsto \|u\|_M := \sqrt{\langle Mu, u\rangle}$ on X^3 induced by the self-adjoint operator M , i.e., that

$$\|x^{k+1} - \widehat{x}\|_M \leq \|x^k - \widehat{x}\|_M.$$

While it is straightforward to verify that H is monotone if A and B are, the question of maximal monotonicity is more involved and will be addressed in Chapter 9. There, we will also show that the expected nonexpansivity holds in a slightly stronger sense and that this will yield the convergence of the method.

Remark 8.3. The Douglas–Rachford splitting was first introduced in [Douglas and Rachford, 1956]; the relationship to the proximal point method was discovered in [Eckstein and Bertsekas, 1992]. The DRS is the *unique* 2-operator splitting method that needs to propagate only one variable from each iteration to the next one, z^{k+1} [Ryu, 2019]. An extension of the DRS with a forward step with respect to a third operator is studied [Davis and Yin, 2017]. It is also possible to devise acceleration schemes under strong monotonicity [see, e.g., Bredies and Sun, 2016].

8.4 PRIMAL-DUAL PROXIMAL SPLITTING

We now consider problems of the form

$$\min_{x \in X} F(x) + G(Kx) \tag{8.12}$$

for $F : X \to \overline{\mathbb{R}}$ and $G : Y \to \overline{\mathbb{R}}$ proper, convex, and lower semicontinuous, and $K \in \mathbb{L}(X; Y)$. Applying Theorem 5.11 and Lemma 5.8 to such a problem yields the Fenchel extremality conditions

$$\begin{cases} -K^* \bar{y} \in \partial F(\bar{x}), \\ \quad \bar{y} \in \partial G(K\bar{x}), \end{cases} \quad \Leftrightarrow \quad \begin{cases} -K^* \bar{y} \in \partial F(\bar{x}), \\ \quad K\bar{x} \in \partial G^*(\bar{y}). \end{cases} \tag{8.13}$$

With the general notation $u := (x, y)$, this can be written as $0 \in H(\widehat{u})$ for

$$H(u) := \begin{pmatrix} \partial F(x) + K^* y \\ \partial G^*(y) - Kx \end{pmatrix}, \tag{8.14}$$

It is again not difficult to see that H is monotone. This suggests that we might be able to apply the proximal point method to find a root of H. In practice we however need to work a little bit more, as the resolvent of H can rarely be given an explicit, easily solvable form. If, however, the resolvents of G^* and F can individually be computed explicitly, it makes sense to try to "decouple" the primal and dual variables. This is what we will do.

To do so, we reformulate for arbitrary $\sigma, \tau > 0$ the extremality conditions (8.13) using Lemma 6.21 as

$$\begin{cases} \widehat{x} = \operatorname{prox}_{\tau F}(\widehat{x} - \tau K^* \widehat{y}), \\ \widehat{y} = \operatorname{prox}_{\sigma G^*}(\widehat{y} + \sigma K \widehat{x}). \end{cases}$$

This suggests the fixed-point iterations

$$\begin{cases} x^{k+1} = \operatorname{prox}_{\tau F}(x^k - \tau K^* y^k), \\ y^{k+1} = \operatorname{prox}_{\sigma G^*}(y^k + \sigma K x^{k+1}). \end{cases} \tag{8.15}$$

In the first equation, we now use $\operatorname{prox}_{\tau F} = (\mathrm{Id} + \tau \partial F)^{-1}$ to obtain that

$$\begin{aligned} x^{k+1} = \operatorname{prox}_{\tau F}(x^k - \tau K^* y^k) &\Leftrightarrow x^k - \tau K^* y^k \in x^{k+1} + \tau \partial F(x^{k+1}) \\ &\Leftrightarrow 0 \in \tau^{-1}(x^{k+1} - x^k) - K^*(y^{k+1} - y^k) \\ &\qquad + [\partial F(x^{k+1}) + K^* y^{k+1}]. \end{aligned} \tag{8.16}$$

Similarly, the second equation of (8.15) gives

$$\begin{aligned} y^{k+1} = \operatorname{prox}_{\sigma G^*}(y^k + \sigma K x^{k+1}) &\Leftrightarrow \sigma^{-1} y^k \in \sigma^{-1} y^{k+1} + \partial G^*(y^{k+1}) - K x^{k+1} \\ &\Leftrightarrow 0 \in \sigma^{-1}(y^{k+1} - y^k) + [\partial G^*(y^{k+1}) - K x^{k+1}]. \end{aligned} \tag{8.17}$$

With the help of (8.16), (8.17), and the operator

$$\tilde{M} := \begin{pmatrix} \tau^{-1}\mathrm{Id} & -K^* \\ 0 & \sigma^{-1}\mathrm{Id} \end{pmatrix},$$

we can then rearrange (8.15) as the preconditioned proximal point method (8.11). Furthermore, provided the step lengths are such that $M = \tilde{M}$ is invertible, this can be written

$$0 \in H(u^{k+1}) + M(u^{k+1} - u^k) \Leftrightarrow u^{k+1} = \mathcal{R}_{M^{-1}H} u^k. \tag{8.18}$$

However, considering the remarks on convergence in the previous sections, there is a problem: M is not self-adjoint and therefore does not induce a (semi-)norm on $X \times Y$. We therefore change our algorithm and take

$$M := \begin{pmatrix} \tau^{-1}\mathrm{Id} & -K^* \\ -K & \sigma^{-1}\mathrm{Id} \end{pmatrix}. \tag{8.19}$$

Correspondingly, replacing (8.17) by

$$\begin{aligned} y^{k+1} = \operatorname{prox}_{\sigma G^*}(y^k + \sigma K(2x^{k+1} - x^k)) &\Leftrightarrow \sigma^{-1} y^k - K x^k \in \sigma^{-1} y^{k+1} + \partial G^*(y^{k+1}) - 2K x^{k+1} \\ &\Leftrightarrow 0 \in \sigma^{-1}(y^{k+1} - y^k) - K(x^{k+1} - x^k) \\ &\qquad + [\partial G^*(y^{k+1}) - K x^{k+1}], \end{aligned}$$

we then obtain from (8.18) the *Primal-Dual Proximal Splitting* (PDPS) method

$$\begin{cases} x^{k+1} = \operatorname{prox}_{\tau F}(x^k - \tau K^* y^k), \\ \bar{x}^{k+1} = 2x^{k+1} - x^k, \\ y^{k+1} = \operatorname{prox}_{\sigma G^*}(y^k + \sigma K \bar{x}^{k+1}). \end{cases} \tag{8.20}$$

The middle over-relaxation step is a consequence of our choice of the bottom-left corner of M defined in (8.19). This itself was forced to have its current form through the self-adjointness requirement on M and the choice of the top-right corner of M. As mentioned above, the role of the latter is to "decouple" the primal update from the dual update by shifting $K^* y^{k+1}$ within H to $K^* y^k$ so that the primal iterate x^{k+1} can be computed without knowing y^{k+1}. (Alternatively, we could zero out the off-diagonal of M and still have a self-adjoint operator, but then we would generally not be able to compute x^{k+1} independent of y^{k+1}.)

In the following chapters, we will demonstrate that the PDPS method converges if the step sizes are chosen to ensure $\sigma\tau\|K\|^2_{\mathbb{L}(X;Y)} < 1$, and that in fact it has particularly good convergence properties. Note that although the iteration (8.20) is implicit in F and G, it is still explicit in K; it is therefore not surprising that step size restrictions based on K remain. Applying, for example, the PDPS method with $\tilde{G}(x) := G(Kx)$ (i.e., applying only the sum rule but not the chain rule) would lead to a fully implicit method. This would, however, require computing K^{-1} in the primal proximal step involving $\operatorname{prox}_{\sigma\tilde{G}^*}$. It is precisely the point of the primal-dual proximal splitting to avoid having to invert K, which is often prohibitively expensive if not impossible (e.g., if K does not have closed range as in many inverse problems).

Remark 8.4. The primal-dual proximal splitting was first introduced in [Pock et al., 2009] for specific image segmentation problems, and later more generally in [Chambolle and Pock, 2011]. For this reason, it is frequently referred to as the *Chambolle–Pock method*. The relation to proximal point methods was first pointed out in [He and Yuan, 2012]. In [Esser et al., 2010] it was classified as the *Primal-Dual Hybrid Gradient method, Modified* or PDHGM after the method (8.15), which is called the PDHG. The latter is due to [Zhu and Chan, 2008].

Banach space generalizations of the PDPS method, based on a so-called *Bregman divergence* in place of $u \mapsto \frac{1}{2}\|u\|^2$, were introduced in [Hohage and Homann, 2014]. We will discuss Bregman divergences in further detail in Section 11.1.

The PDPS method has been also generalized to different types of nonconvex problems in [Möllenhoff et al., 2015; Valkonen, 2014]. Stochastic generalizations are considered in [Chambolle et al., 2018; Valkonen, 2019].

8.5 PRIMAL-DUAL EXPLICIT SPLITTING

The PDPS method is useful for dealing with the sum of functionals where one summand includes a linear operator. However, if this is the case for *both* operators, i.e.,

$$\min_{x\in X} F(Ax) + G(Kx)$$

for $F : Z \to \overline{\mathbb{R}}$, $G : Y \to \overline{\mathbb{R}}$, $A \in \mathbb{L}(X;Z)$ and $K \in \mathbb{L}(X;Y)$, we again have the problem of dealing with a complicated proximal mapping. One workaround is the following "lifting trick": we introduce

$$\tilde{F}(x) := 0, \quad \tilde{G}(y,z) := G(y) + F(z) \quad \text{and} \quad \tilde{K}x := (Kx, Ax), \tag{8.21}$$

and then apply the PDPS method to the reformulated problem $\min_x \tilde{F}(x) + \tilde{G}(\tilde{K}x)$. According to Lemma 6.24 (iii), the dual step of the PDPS method will then split into separate proximal steps with respect to G^* and F^*, while the proximal map in the primal step will be trivial. However, an additional dual variable will have been introduced through the introduction of z above, which can be costly.

An alternative approach is the following. Analogously to (8.15), but only using Lemma 6.21 on the second relation of (8.13) together with the chain rule (Theorem 4.17), we can reformulate the latter as

$$\begin{cases} \widehat{x} \in \widehat{x} - \tau[\partial A^* F(A\widehat{x}) + K^*\widehat{y}], \\ \widehat{y} = \operatorname{prox}_{\sigma G^*}(\widehat{y} + \sigma K\widehat{x}). \end{cases} \tag{8.22}$$

(For $K = \mathrm{Id}$, we can alternatively obtain (8.22) from the derivation of explicit splitting by using Moreau's identity, Theorem 6.23, in the second relation of (8.5).)

If F is Gâteaux differentiable (and taking $A = \mathrm{Id}$ for the sake of presentation), inserting the first relation in the second relation, (8.22) can be further rewritten as

$$\begin{cases} \widehat{x} = \widehat{x} - \tau[\nabla F(\widehat{x}) + K^*\widehat{y}], \\ \widehat{y} = \operatorname{prox}_{\sigma G^*}(\widehat{y} + \sigma K\widehat{x} - \sigma\tau K[\nabla F(\widehat{x}) + K^*\widehat{y}]). \end{cases}$$

Reordering the lines and fixing $\tau = \sigma = 1$, the corresponding fixed-point iteration leads to the *primal-dual explicit splitting* (PDES) method

$$\begin{cases} y^{k+1} = \operatorname{prox}_{G^*}((\mathrm{Id} - KK^*)y^k + K(x^k - \nabla F(x^k))), \\ x^{k+1} = x^k - \nabla F(x^k) - K^* y^{k+1}. \end{cases} \tag{8.23}$$

Again, we can write (8.23) in more implicit form as

$$\begin{cases} 0 \in \partial G^*(y^{k+1}) - K(x^k - \nabla F(x^k) - K^* y^k) + (y^{k+1} - y^k), \\ 0 = \nabla F(x^k) + K^* y^{k+1} + (x^{k+1} - x^k). \end{cases}$$

Inserting the second relation in the first, this is

$$\begin{cases} 0 \in \partial G^*(y^{k+1}) - Kx^{k+1} + (\mathrm{Id} - KK^*)(y^{k+1} - y^k), \\ 0 = \nabla F(x^k) + K^* y^{k+1} + (x^{k+1} - x^k). \end{cases}$$

If we now introduce the preconditioning operator

$$M := \begin{pmatrix} \mathrm{Id} & 0 \\ 0 & \mathrm{Id} - KK^* \end{pmatrix}, \tag{8.24}$$

then in terms of the monotone operator H introduced in (8.14) for the PDPS method and $u = (x, y)$, the PDES method (8.23) can be written in implicit form as

$$0 \in H(u^{k+1}) + \begin{pmatrix} \nabla F(x^k) - \nabla F(x^{k+1}) \\ 0 \end{pmatrix} + M(u^{k+1} - u^k). \tag{8.25}$$

The middle term switches the step with respect to F to be explicit. Note that (8.7) could have also been written with a similar middle term; we can therefore think of the PDES method as a *preconditioned explicit splitting* method.

The preconditioning operator M is self-adjoint as well as positive semi-definite if $\|K\|_{\mathbb{L}(X;Y)} \leq 1$. It does not have the off-diagonal decoupling terms that the preconditioner for the PDPS method has. Instead, through the special structure of the problem the term $\mathrm{Id} - KK^*$ decouple y^{k+1} from x^{k+1}, allowing y^{k+1} be computed first.

We will in Section 9.4 see that the iterates of the PDES method converge weakly when ∇F is Lipschitz with factor strictly less than 2.

Remark 8.5. The primal-dual explicit splitting was introduced in [Loris and Verhoeven, 2011] as *Generalized Iterative Soft Thresholding* (GIST) for $F(x) = \frac{1}{2}\|b - x\|^2$. The general case has later been called the *primal-dual fixed point method* (PDFP) in [Chen et al., 2013] and the *proximal alternating predictor-corrector* (PAPC) in [Drori et al., 2015].

8.6 AUGMENTED LAGRANGIAN AND ADMM

Let $F : X \to \overline{\mathbb{R}}$ and $G : Z \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous. Also let $A \in \mathbb{L}(X; Y)$, and $B \in \mathbb{L}(Z; Y)$, and consider for some $c \in Y$ the problem

$$\min_{x,z} \; F(x) + G(z) \quad \text{s.t.} \quad Ax + Bz = c. \tag{8.26}$$

A traditional way to handle this kind of constraint problems is by means of the augmented Lagrangian. We start by introducing the *Lagrangian*

$$\mathcal{L}(x, z; \lambda) := F(x) + G(z) + \langle Ax + Bz - c, \lambda \rangle_Y.$$

Then (8.26) has the same solutions as the saddle-point problem

$$\min_{x\in X, z\in Z}\max_{\lambda\in Y}\ \mathcal{L}(x,z;\lambda). \tag{8.27}$$

We may then "augment" the Lagrangian by a squared penalty on the violation of the constraint, hence obtaining the equivalent problem

$$\min_{x\in X, z\in Z}\max_{\lambda\in Y}\ \mathcal{L}_\tau(x,z;\lambda) := F(x)+G(z)+\langle Ax+Bz-c,\lambda\rangle_Y+\frac{\tau}{2}\|Ax+Bz-c\|_Y^2, \tag{8.28}$$

where $\mathcal{L}_\tau$ is the *augmented Lagrangian*.

A classical approach for the solution of (8.28) is by alternatingly solving for one variable, keeping the others fixed. If we take a proximal step for the dual variable or *Lagrange multiplier* λ, this yields the *Alternating Directions Method of Multipliers* (ADMM)

$$\begin{cases} x^{k+1} := \operatorname{arg\,min}_{x\in X}\ \mathcal{L}_\tau(x,z^k;\lambda^k),\\ z^{k+1} := \operatorname{arg\,min}_{z\in Z}\ \mathcal{L}_\tau(x^{k+1},z;\lambda^k),\\ \lambda^{k+1} := \operatorname{arg\,max}_{\lambda\in Y}\ \mathcal{L}_\tau(x^{k+1},z^{k+1};\lambda)-\dfrac{1}{2\tau}\|\lambda-\lambda^k\|_Y^2. \end{cases} \tag{8.29}$$

This can be rewritten as

$$\begin{cases} x^{k+1}\in (A^*A+\tau^{-1}\partial F)^{-1}(A^*(c-Bz^k-\tau^{-1}\lambda^k)),\\ z^{k+1}\in (B^*B+\tau^{-1}\partial G)^{-1}(B^*(c-Ax^{k+1}-\tau^{-1}\lambda^k)),\\ \lambda^{k+1} := \lambda^k+\tau(Ax^{k+1}+Bz^{k+1}-c). \end{cases} \tag{8.30}$$

As can be observed, the ADMM requires inverting relatively complicated set-valued operators in place of simple proximal point operations. This is why the basic ADMM is seldom practically implementable without the application of a further optimization method to solve the x and z updates.

In the literature, there have been various remedies to the nonimplementability of the ADMM. In particular, one can modify the ADMM iterations by adding to (8.29) additional proximal terms. Introducing for some $Q_x\in\mathbb{L}(X;X)$ and $Q_z\in\mathbb{L}(Z;Z)$ the weighted norms $\|x\|_{Q_x} := \sqrt{\langle Q_x x,x\rangle_X}$ and $\|z\|_{Q_z} := \sqrt{\langle Q_z z,z\rangle_Z}$, this leads to the iteration

$$\begin{cases} x^{k+1} := \operatorname{arg\,min}_{x\in X}\ \mathcal{L}_\tau(x,z^k;\lambda^k)+\dfrac{1}{2}\|x-x^k\|_{Q_x}^2,\\ z^{k+1} := \operatorname{arg\,min}_{z\in Z}\ \mathcal{L}_\tau(x^{k+1},z;\lambda^k)+\dfrac{1}{2}\|z-z^k\|_{Q_z}^2,\\ \lambda^{k+1} := \operatorname{arg\,max}_{\lambda\in Y}\ \mathcal{L}_\tau(x^{k+1},z^{k+1};\lambda)-\dfrac{1}{2\tau}\|\lambda-\lambda^k\|_Y^2. \end{cases} \tag{8.31}$$

If we specifically take $Q_x := \sigma^{-1}\mathrm{Id} - \tau A^*A$ and $Q_z := \theta^{-1}\mathrm{Id} - \tau B^*B$ for some $\sigma, \theta > 0$ with $\theta\tau\|A\| < 1$ and $\sigma\tau\|B\| < 1$, then we can expand

$$\begin{aligned}\mathcal{L}_\tau(x,z;\lambda) + \frac{1}{2}\|x - x^k\|^2_{Q_x} &= F(x) + G(z) + \langle Ax + Bz - c, \lambda\rangle_Y \\ &\quad + \tau\langle x, A^*(Bz - c)\rangle_X + \frac{\tau}{2}\|Bz - c\|^2_Y \\ &\quad + \frac{1}{2\sigma}\|x - x^k\|^2_X + \tau\langle x^{k+1}, A^*Ax^k\rangle_X - \frac{\tau}{2}\|Ax^k\|^2_Y,\end{aligned}$$

which has the "partial" subdifferential ∂_x with respect to x (keeping z, λ fixed)

$$\partial_x\mathcal{L}_\tau(x,z;\lambda) = \partial F(x) + A^*\lambda + \tau A^*(Bz - c) + \sigma^{-1}(x - x^k) + \tau A^*Ax^k.$$

Similarly computing the partial subdifferential ∂_z with respect to z, (8.31) can thus be written as the *preconditioned ADMM*

$$\begin{cases} x^{k+1} := \mathrm{prox}_{\sigma F}((\mathrm{Id} - \sigma\tau)A^*Ax^k + \sigma A^*(\tau(c - Bz^k) - \lambda^k)), \\ z^{k+1} := \mathrm{prox}_{\theta G}((\mathrm{Id} - \theta\tau)B^*Bz^k + \theta B^*(\tau(c - Ax^{k+1}) - \lambda^k)), \\ \lambda^{k+1} := \lambda^k + \tau(Ax^{k+1} + Bz^{k+1} - c). \end{cases} \tag{8.32}$$

We will see in the next section that this method is just the PDPS method with the primal and dual variables exchanged.

Remark 8.6. The ADMM was introduced in [Arrow et al., 1958; Gabay, 1983] as an alternating approach to the classical Augmented Lagrangian method. The preconditioned ADMM is due to [Zhang et al., 2011].

8.7 CONNECTIONS

In Section 8.5 we have seen the importance and interplay of problem formulation and algorithm choice for problems with a specific structure. We will now see that many of the algorithms we have presented are actually equivalent when applied to differing formulations of the problem. Hence, if one algorithm is efficient on one formulation of the problem, another algorithm may work equally well on a different formulation.

We start by considering the ADMM problem (8.26), which we can reformulate as

$$\min_{x,z}\ F(x) + G(z) + \delta_{\{c\}}(Ax + Bz).$$

Applying the PDPS method (8.20) to this formulation yields the algorithm

$$\begin{cases} x^{k+1} := \operatorname{prox}_{\tau F}(x^k - \tau A^* \lambda^k), \\ z^{k+1} := \operatorname{prox}_{\tau G}(z^k - \tau B^* \lambda^k), \\ \bar{x}^{k+1} := 2x^{k+1} - x^k, \\ \bar{z}^{k+1} := 2z^{k+1} - z^k, \\ \lambda^{k+1} := \lambda^k + \sigma(A\bar{x}^{k+1} + B\bar{z}^{k+1} - c). \end{cases} \tag{8.33}$$

Note that both the ADMM (8.30) and the preconditioned ADMM (8.32) have a very similar form to this iteration. We will now demonstrate that if $A = \mathrm{Id}$ and so $X = Y$, i.e., if we want to solve the (primal) problem

$$\min_{z \in Z} F(c - Bz) + G(z), \tag{8.34}$$

then the ADMM is equivalent to the PDPS method (8.20) applied to the (dual) problem

$$\min_{y \in Y} \left[G^*(B^* y) - \langle c, y \rangle_Y\right] + F^*(y), \tag{8.35}$$

where the dual step will be performed with respect to F^*.

To make the exact way the PDPS method is applied in each instance clearer, and to highlight the primal-dual nature of the PDPS method, it will be more convenient to write the problem to which the PDPS method is applied in *saddle-point form*. Specifically, minding (5.4) together with the discussion following Theorem 5.11, the problem $\min_x F(x) + G(Kx)$ can be written as the saddle-point problem

$$\min_{x \in X} \max_{y \in Y} F(x) + \langle Kx, y \rangle_Y - G^*(y).$$

This formulation also shows the dual variable directly in the problem formulation. Applied to (8.35), we then obtain the problem

$$\min_{y \in Y} \max_{x \in X} \left[G^*(B^* y) - \langle c, y \rangle_Y\right] + \langle x, y \rangle_Y - F(x). \tag{8.36}$$

Our claim is that the PDPS method applied to this saddle-point formulation is equivalent to the ADMM in case of $A = \mathrm{Id}$. The iterates of the two algorithms will be different, as the variables solved for will be different aside from the shared x. However, all the variables will be related by affine transformations.

We will also demonstrate that the preconditioned ADMM is equivalent to the PDPS method when $B = \mathrm{Id}$. In fact, we will demonstrate a chain of relationships from ADMM or preconditioned ADMM (primal problem) via the PDPS (saddle-point problem) method to the DRS method (dual problem); the equivalence between the ADMM and the DRS method even holds generally.

To demonstrate the idea, we start with $A = B = \mathrm{Id}$. Then (8.30) reads

$$\begin{cases} x^{k+1} := \operatorname{prox}_{\tau^{-1}F}(c - z^k - \tau^{-1}\lambda^k), \\ z^{k+1} := \operatorname{prox}_{\tau^{-1}G}(c - x^{k+1} - \tau^{-1}\lambda^k), \\ \lambda^{k+1} := \lambda^k + \tau(x^{k+1} + z^{k+1} - c). \end{cases} \tag{8.37}$$

Using the third step for the previous iteration to obtain an expression for z^k, we can rewrite the first step as

$$x^{k+1} := \operatorname{prox}_{\tau^{-1}F}(x^k - \tau^{-1}(2\lambda^k - \lambda^{k-1})).$$

If we use Lemma 6.24 (ii), the second step reads

$$z^{k+1} := (c - x^{k+1} - \tau^{-1}\lambda^k) - \tau^{-1}\operatorname{prox}_{\tau G^*}(\tau(c - x^{k+1}) - \lambda^k).$$

Minding the third step of (8.37), this yields $\lambda^{k+1} = -\operatorname{prox}_{\tau G^*}(\tau(c - x^{k+1}) - \lambda^k)$. Replacing λ^{k+1} by $y^{k+1} := -\lambda^{k+1}$, moving c into the proximal part, and reordering the steps such that x^{k+1} becomes x^k, transforms (8.37) into

$$\begin{cases} y^{k+1} := \operatorname{prox}_{\tau(G^* - \langle c, \cdot \rangle)}(y^k - \tau x^k), \\ x^{k+1} := \operatorname{prox}_{\tau^{-1}F}(x^k + \tau^{-1}(2y^{k+1} - y^k)). \end{cases} \tag{8.38}$$

This is the PDPS method applied to (8.36) with $B = \mathrm{Id}$. However, the step lengths τ and $\sigma = \tau^{-1}$ do not satisfy $\tau\sigma\|K\|^2 < 1$, which would be needed to deduce convergence of the ADMM from that of the PDPS method. But we will see in Chapter 11 that these step lengths at least lead to convergence of a certain "Lagrangian duality gap", and for the ADMM we can in general only prove such gap estimates.

To show the relation of ADMM to implicit splitting, we further use Lemma 6.24 (ii) in the second step of (8.38) to obtain

$$x^{k+1} = \tau^{-1}(2y^{k+1} - y^k) + x^k - \tau^{-1}\operatorname{prox}_{\tau F^*}(2y^{k+1} - y^k + \tau x^k).$$

Introducing $w^{k+1} := y^{k+1} - \tau x^{k+1}$ and changing variables, we thus transform (8.38) into

$$\begin{cases} y^{k+1} := \operatorname{prox}_{\tau(G^* - \langle c, \cdot \rangle)}(w^k), \\ w^{k+1} := w^k - y^{k+1} + \operatorname{prox}_{\tau F^*}(2y^{k+1} - w^k). \end{cases} \tag{8.39}$$

But this is the DRS method (8.8) applied to

$$\min_{x \in X} F^*(x) + [G^*(x) - \langle c, x \rangle_X].$$

Recall now from Lemma 5.4 that $[G(c - \cdot)]^* = G^*(-\cdot) + \langle c, \cdot \rangle_Y$. Theorem 5.11 thus shows that this is the dual problem of (8.34), so we can at least deduce from Corollary 9.11 the convergence of y^k to a solution of the dual problem.

We can make the correspondence more general with the help of the following generalization of Moreau's identity (Theorem 6.23).

Lemma 8.7. *Let $S = G \circ K$ for convex, proper, and lower semicontinuous $G : Y \to \overline{\mathbb{R}}$ and $K \in \mathbb{L}(X;Y)$. If there exists an $x_0 \in \operatorname{dom} S$ such that $Kx_0 \in \operatorname{int}(\operatorname{dom} G)$, then for all $x \in X$ and $\gamma > 0$,*

$$x = \operatorname{prox}_{\gamma S}(x) + \gamma K^*(KK^* + \gamma^{-1}\partial G^*)^{-1}(\gamma^{-1}Kx).$$

In particular,

$$\operatorname{prox}_{S^*}(x) = K^*(KK^* + \partial G^*)^{-1}(Kx).$$

Proof. By Theorem 5.11, $w = \operatorname{prox}_{\gamma S}(x)$ if and only if for some $y^* \in Y^*$ holds

$$\begin{cases} -K^*y^* \in w - x, \\ \quad y^* \in \gamma \partial G(Kw). \end{cases}$$

In other words, by Lemma 5.8,

$$\begin{cases} -K^*y^* = w - x, \\ \quad Kw \in \partial G^*(\gamma^{-1}y^*). \end{cases}$$

Applying K to the first relation, inserting the second, and multiplying by γ^{-1} yields

$$KK^*\gamma^{-1}y^* + \gamma^{-1}\partial G^*(\gamma^{-1}x^*) = \gamma^{-1}Ky,$$

i.e., $\gamma^{-1}x^* \in (KK^* + \gamma^{-1}\partial G^*)^{-1}(\gamma^{-1}Kx)$. Combined with $-K^*y^* = w - x$, this yields the first claim. The second claim then follows from Theorem 7.11 together with the first claim for $\gamma = 1$. □

Theorem 8.8. *Let $F : X \to \overline{\mathbb{R}}$ and $G : Z \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous. Also let $A \in \mathbb{L}(X;Y)$, and $B \in \mathbb{L}(Z;Y)$, and $c \in Y$. Assume the existence of a point $(x_0, z_0) \in \operatorname{dom} F \times \operatorname{dom} G$ with $Ax_0 + Bz_0 = c$. Then the iterates of the following algorithms can be transformed to one another with affine transformations and (to obtain the ADMM) the addition of elements of $\ker A$ and $\ker B$:*

(i) The ADMM applied to the (primal) problem

$$\min_{x \in X, z \in Z} F(x) + G(z) \quad \text{s.t.} \quad Ax + Bz = c. \tag{8.40}$$

(ii) The DRS method applied to the (dual) problem

$$\min_{y \in Y} F^*(A^*y) + [G^*(B^*y) - \langle c, y\rangle_Y]. \tag{8.41}$$

(iii) If $A = \mathrm{Id}$, $X = Y$, and $\sigma = \tau^{-1}$, the PDPS method applied to the (saddle-point) problem

$$\min_{y \in Y} \max_{x \in X} [G^*(B^*y) - \langle c, y\rangle_Y] + \langle x, y\rangle_Y - F(x). \tag{8.42}$$

Proof. We first show that the ADMM updates for (8.40) can be transformed, via affine transformations alone, to the DRS updates for (8.41), and the PDPS updates for (8.42). Observe that the assumption on the existence of (x_0, z_0) ensures that the infimum in (8.40) is finite. Thus, multiplying the first and second updates of (8.30) by A and B, and changing variables x^{k+1} and z^{k+1} to $\tilde{x}^{k+1} := Ax^{k+1}$ and $\tilde{z}^{k+1} := Bz^{k+1}$, we obtain

$$\begin{cases} \tilde{x}^{k+1} \in A(A^*A + \tau^{-1}\partial F)^{-1}(A^*(c - \tilde{z}^k - \tau^{-1}\lambda^k)), \\ \tilde{z}^{k+1} \in B(B^*B + \tau^{-1}\partial G)^{-1}(B^*(c - \tilde{x}^{k+1} - \tau^{-1}\lambda^k)), \\ \lambda^{k+1} := \lambda^k + \tau(Ax^{k+1} + Bz^{k+1} - c). \end{cases} \tag{8.43}$$

Using Lemma 8.7 with $y^k := -\lambda^k$ and $-y^{k+1} = -y^k + \tau(\tilde{x}^{k+1} + z^{k+1} - c)$, we transform this as above to

$$\begin{cases} y^{k+1} \in \operatorname{prox}_{\tau G^* \circ B^*}(\tau(c - \tilde{x}^k) + y^k), \\ \tilde{x}^{k+1} \in A(A^*A + \tau^{-1}\partial F)^{-1}(A^*(2y^{k+1} - y^k + \tilde{x}^k)). \end{cases} \tag{8.44}$$

If $A = \mathrm{Id}$, this is the PDPS method for (8.42) with the iterate equivalence $\tilde{x}^{k+1} = x^{k+1}$. We continue with Lemma 8.7 and $w^{k+1} := y^{k+1} - \tau\tilde{x}^{k+1}$ to transform (8.44) further into

$$\begin{cases} y^{k+1} := \operatorname{prox}_{\tau(G^* \circ B^* - \langle c, \cdot \rangle)}(w^k), \\ w^{k+1} := w^k - y^{k+1} + \operatorname{prox}_{\tau F^* \circ A^*}(2y^{k+1} - w^k). \end{cases} \tag{8.45}$$

This is the DRS method for (8.41).

In the other direction, it is clear from the derivation above that the DRS (8.45) generates (8.44) and, via affine transformations, its iterates. Likewise (8.44) and be transformed back into (8.44) by reversing the steps. The passage from the iterates of (8.43) back to the iterates of the ADMM (8.30) cannot be achieved with affine transformations alone, unless A and B are injective. However, when there exist $\tilde{x}^{k+1}$ and $\tilde{z}^{k+1}$ solving (8.43), there must exist some x^{k+1} and z^{k+1} with $\tilde{x}^{k+1} := Ax^{k+1}$ and $\tilde{z}^{k+1} := Bz^{k+1}$ that satisfy (8.30).

We still need to establish the claimed duality relationship between the problems (8.40), (8.41), and (8.42). To pass from (8.40) to (8.41), we would like to apply Theorem 5.11 to $\tilde{F}(x, z) := F(x) + G(z)$, $\tilde{G}(y) := \delta_{\{y=c\}}(y)$, and $\tilde{K} := (A, B)$. However, $\operatorname{dom} G = \{c\}$ has empty interior, so condition (ii) of the theorem does not hold. Recalling Remarks 4.16 and 5.13, we can however replace the interior with the relative interior $\operatorname{ri} \operatorname{dom} \tilde{G} = \{c\}$. Thus the condition reduces to the existence of $y_0 \in \operatorname{dom} \tilde{F}$ with $Ky_0 = c$, which is satisfied by $y_0 = (x_0, z_0)$.

Finally, the relationship to (8.42) when $A = \mathrm{Id}$ is immediate from (8.41) and the definition of the conjugate function F^*. The existence of a saddle point follows from the proof of Theorem 5.11. □

The methods in the proof of Theorem 8.8 are rarely computationally feasible or efficient unless $A = B = \mathrm{Id}$, due to the difficult proximal mappings for compositions of functionals with operators or the set-valued operator inversions required. On the other hand, the PDPS method (8.33) only requires that we can compute the proximal mappings of G and F. This demonstrates the importance of problem formulation.

Similar connections hold for the preconditioned ADMM (8.32). With the help of the third step of (8.32), the first step can be rewritten

$$x^{k+1} := \operatorname{prox}_{\sigma F}(x^k - \sigma A^*(2\lambda^k - \lambda^{k-1})).$$

If $\theta\tau = 1$ and $B = \mathrm{Id}$, the second step reads

$$z^{k+1} := \operatorname{prox}_{\tau^{-1}G}((c - Ax^{k+1}) - \tau^{-1}\lambda^k).$$

We transform this with Lemma 6.24 (ii) into

$$z^{k+1} = (c - Ax^{k+1}) - \tau^{-1}\lambda^k - \tau^{-1}\operatorname{prox}_{\tau G^*}(\tau(c - Ax^{k+1}) - \lambda^k).$$

Using the third step of (8.32), this is equivalent to

$$-\lambda^{k+1} = \operatorname{prox}_{\tau G^*}(\tau(c - Ax^{k+1}) - \lambda^k).$$

Introducing $y^{k+1} := -\lambda^{k+1}$ and changing the order of the first and second step, we therefore transform (8.32) into the PDPS method

$$\begin{cases} y^{k+1} := \operatorname{prox}_{\tau G^*}(y^k - \tau A x^k), \\ x^{k+1} := \operatorname{prox}_{\sigma F}(x^k + \sigma A^*(2y^{k+1} - y^k)). \end{cases} \tag{8.46}$$

We therefore have obtained the following result.

Theorem 8.9. *Let $F : X \to \overline{\mathbb{R}}$ and $G : Y \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous. Also let $A \in \mathbb{L}(X; Y)$ and $c \in Y$. Assume the existence of a point $(x_0, z_0) \in \operatorname{dom} F \times \operatorname{dom} G$ with $Ax_0 + z_0 = c$. Take $\theta = \tau^{-1}$. Then subject to affine transformations to obtain iterates not explicitly generated in each case, the following are equivalent:*

(i) The preconditioned ADMM (8.32) applied to the (primal) problem

$$\min_{x \in X, z \in Y} F(x) + G(z) \quad s.t. \quad Ax + z = c.$$

(ii) The PDPS method applied to the (saddle point) problem

$$\min_{y \in Y} \max_{x \in X} [G^*(y) - \langle c, y\rangle_Y] + \langle Ax, y\rangle_Y - F(x).$$

(iii) *If* $A = \mathrm{Id}$, $X = Y$, *and* $\sigma = \tau^{-1}$, *the Douglas–Rachford splitting method applied to the (dual) problem*

$$\min_{y \in X} \; F^*(y) + [G^*(y) - \langle c, y \rangle_Y].$$

Proof. We have already proved the equivalence of the preconditioned ADMM and the PDPS method. For equivalence to the DRS method, we observe that under the additional assumptions of this theorem, (8.46) reduces to (8.38). □

9 SPLITTING METHODS: WEAK CONVERGENCE

Now that we have in the previous chapter derived several iterative procedures through the manipulation of fixed-point equations, we have to show that they indeed converge to a fixed point (which by construction is then the solution of an optimization problem, making these procedures optimization algorithms). We start with weak convergence, as this is the most that can generally be expected.

The classical approach to proving weak convergence is by introducing suitable contractive (or at least firmly nonexpansive) operators related to the algorithm and then applying classical fixed-point theorems (see Remark 9.5 below). We will instead introduce a very direct approach that will then extend in the following chapters to be also capable of proving convergence rates. The three main ingredients of all convergence proofs will be

(i) The three-point identity (1.5), which we recall here as

$$\langle x-y, x-z\rangle_X = \frac{1}{2}\|x-y\|_X^2 - \frac{1}{2}\|y-z\|_X^2 + \frac{1}{2}\|x-z\|_X^2 \quad \text{for all } x, y, z \in X. \tag{9.1}$$

(ii) The monotonicity of the operator H whose roots we seek to find (which in the simplest case equals ∂F for the functional F we want to minimize).

(iii) The nonnegativity of the preconditioning operators M defining the implicit forms of the algorithms we presented in Chapter 8.

In the later chapters, stronger versions of the last two ingredients will be required to obtain convergence rates and the convergence of function value differences $F(x^{k+1}) - F(\widehat{x})$ or of more general gap functionals.

9.1 OPIAL'S LEMMA AND FEJÉR MONOTONICITY

The next lemma forms the basis of all our weak convergence proofs. It is a generalized subsequence argument, showing that if all weak limit points of a sequence lie in a set and if the sequence does not diverge (in the strong sense) away from this set, the full sequence converges weakly. We recall that $\bar{x} \in X$ is a weak(-$*$) limit point of the sequence $\{x^k\}_{k\in\mathbb{N}}$, if there exists a subsequence such that $x^{k_\ell} \rightharpoonup \bar{x}$ weakly(-$*$) in X.

Lemma 9.1 (Opial). *Let X be a Hilbert space and $\hat{X} \subset X$ be a nonempty subset. If the sequence $\{x^k\}_{k\in\mathbb{N}} \subset X$ satisfies*

(i) $\|x^{k+1} - \bar{x}\|_X \leq \|x^k - \bar{x}\|_X$ for all $\bar{x} \in \hat{X}$ and $k \in \mathbb{N}$;

(ii) all weak limit points of $\{x^k\}_{k\in\mathbb{N}}$ belong to $\hat{X}$;

then $x^k \rightharpoonup \widehat{x}$ in X for some $\widehat{x} \in \hat{X}$.

Proof. First, the assumption (i) implies that the sequence $\{x^k\}_{k\in\mathbb{N}}$ is bounded and hence by Theorem 1.9 contains a weakly convergent subsequence. Let now $\bar{x}$ and $\hat{x}$ be weak limit points. The assumption (i) then implies that both $\{\|x^k - \bar{x}\|_X\}_{k\in\mathbb{N}}$ and $\{\|x^k - \hat{x}\|_X\}_{k\in\mathbb{N}}$ are decreasing and bounded from below and therefore convergent. This yields that

$$\langle x^k, \bar{x} - \hat{x}\rangle_X = \frac{1}{2}\left(\|x^k - \hat{x}\|_X^2 - \|x^k - \bar{x}\|_X^2 + \|\bar{x}\|_X^2 - \|\hat{x}\|_X^2\right) \to c \in \mathbb{R}.$$

Since $\bar{x}$ is a weak accumulation point, there exists a subsequence $\{x^{k_n}\}_{n\in\mathbb{N}}$ with $x^{k_n} \rightharpoonup \bar{x}$; similarly, there exists a subsequence $\{x^{k_m}\}_{m\in\mathbb{N}}$ with $x^{k_m} \rightharpoonup \hat{x}$. Hence,

$$\langle \bar{x}, \bar{x} - \hat{x}\rangle_X = \lim_{n\to\infty} \langle x^{k_n}, \bar{x} - \hat{x}\rangle_X = c = \lim_{m\to\infty} \langle x^{k_m}, \bar{x} - \hat{x}\rangle_X = \langle \hat{x}, \bar{x} - \hat{x}\rangle_X,$$

and therefore

$$0 = \langle \bar{x} - \hat{x}, \bar{x} - \hat{x}\rangle_X = \|\bar{x} - \hat{x}\|_X^2,$$

i.e., $\bar{x} = \hat{x}$. Every convergent subsequence thus has the same weak limit (which lies in $\hat{X}$ by assumption (ii)). The claim now follows from a standard subsequence–subsequence argument: Assume to the contrary that there exists a subsequence of $\{x^k\}_{k\in\mathbb{N}}$ that does not converge to $\hat{x}$. Then we can apply the above argument to obtain a further subsequence converging to $\hat{x}$, which is a contradiction to the fact that any subsequence of a convergent sequences converges to the same limit. □

A sequence satisfying the condition (i) is called *Fejér monotone* (with respect to $\hat{X}$); this is a crucial property of iterates generated by any fixed-point algorithm.

Remark 9.2. Lemma 9.1 first appeared in the proof of [Opial, 1967, Theorem 1]. (There $\hat{X}$ is assumed to be closed and convex, but we do not require this since Condition (ii) is already sufficient to show the claim.)

The concept of Fejér monotone sequences first appears in [Fejér, 1922], where it was observed that for every point outside the convex hull of a subset of the Euclidean plane, it is always possible to construct a point that is closer to each point in the subset than the original point (and that this property in fact characterizes the convex hull). The term *Fejér monotone* itself appears in [Motzkin and Schoenberg, 1954], where this construction is used to show convergence of an iterative scheme for the projection onto a convex polytope.

9.2 THE FUNDAMENTAL METHODS: PROXIMAL POINT AND EXPLICIT SPLITTING

Using Opial's Lemma 9.1, we can fairly directly show weak convergence of the proximal point and forward-backward splitting methods.

PROXIMAL POINT METHOD

We recall our most fundamental nonsmooth optimization algorithm, the proximal point method. For later use, we treat the general version of (8.1) for an arbitrary set-valued operator $H : X \rightrightarrows X$, i.e.,

$$x^{k+1} = \mathcal{R}_{\tau_k H}(x^k). \tag{9.2}$$

We will need the next lemma to allow a very general choice of the step lengths $\{\tau_k\}_{k\in\mathbb{N}}$. (If we assume $\tau_k \geq \varepsilon > 0$, in particular if we keep $\tau_k \equiv \tau$ constant, it will not be needed.) For the statement, note that by the definition of the resolvent, (9.2) is equivalent to $\tau_k^{-1}(x^k - x^{k+1}) \in H(x^{k+1})$.

Lemma 9.3. *Let $\{\tau_k\}_{k\in\mathbb{N}} \subset (0,\infty)$ with $\sum_{k=0}^{\infty} \tau_k^2 = \infty$, and let $H : X \rightrightarrows X$ be monotone. Suppose $\{x^k\}_{k\in\mathbb{N}}$ and $w^{k+1} := -\tau_k^{-1}(x^{k+1} - x^k)$ satisfies*

(i) $0 \neq w^{k+1} \in H(x^{k+1})$ *and*

(ii) $\displaystyle\sum_{k=0}^{\infty} \tau_k^2 \|w^k\|_X^2 < \infty.$

Then $\|w^k\|_X \to 0$.

Proof. Since $w^k \in H(x^k)$ and H is monotone, we have from the definition of w^k that

$$0 \leq \langle w^{k+1} - w^k, x^{k+1} - x^k\rangle_X = \tau_k \langle w^k - w^{k+1}, w^{k+1}\rangle_X \leq \tau_k \|w^{k+1}\|_X (\|w^k\|_X - \|w^{k+1}\|_X).$$

Thus the nonnegative sequence $\{\|w^k\|_X\}_{k\in\mathbb{N}}$ is decreasing and hence converges to some $M \geq 0$. Since $\sum_{k=0}^{\infty} \tau_k^2 = \infty$, the second assumption implies that $\liminf_{k\to\infty} \|w^k\|_X = 0$. Since the full sequence converges, $M = 0$, i.e., $\|w^k\|_X \to 0$ as claimed. □

This shows that the "generalized residual" w^k in the inclusion $w^k \in H(x^k)$ converges (strongly) to zero. As usual, this does not (yet) imply that $\{x^k\}_{k\in\mathbb{N}}$ itself converges; but if it does, we expect the limit to be a root of H. This is what we prove next, using the three fundamental ingredients we introduced in the beginning of the chapter.

Theorem 9.4. *Let $H : X \rightrightarrows X$ be monotone and weak-to-strong outer semicontinuous with $H^{-1}(0) \neq 0$. Furthermore, let $\{\tau_k\}_{k\in\mathbb{N}} \subset (0,\infty)$ with $\sum_{k=0}^{\infty} \tau_k^2 = \infty$. If $\{x^k\}_{k\in\mathbb{N}} \subset X$ is given by the iteration (9.2) for any initial iterate $x^0 \in X$, then $x^k \rightharpoonup \widehat{x}$ for some root $\widehat{x} \in H^{-1}(0)$.*

Proof. We recall that the proximal point iteration can be written in implicit form as

$$0 \in \tau_k H(x^{k+1}) + (x^{k+1} - x^k). \tag{9.3}$$

We "test" (9.3) by the application of $\langle \cdot, x^{k+1} - \widehat{x}\rangle_X$ for an arbitrary $\widehat{x} \in H^{-1}(0)$. Thus we obtain

$$0 \in \langle \tau_k H(x^{k+1}) + (x^{k+1} - x^k), x^{k+1} - \widehat{x}\rangle_X, \tag{9.4}$$

where the right-hand side should be understood as the set of all possible inner products involving elements of $H(x^{k+1})$. By the monotonicity of H, since $0 \in H(\widehat{x})$, we have

$$\langle H(x^{k+1}), x^{k+1} - \widehat{x}\rangle_X \geq 0,$$

which again should be understood to hold for any $w \in H(x^{k+1})$. (We will frequently make use of this notation and the one from (9.4) throughout this and the following chapters to keep the presentation concise.) Thus (9.4) yields

$$\langle x^{k+1} - x^k, x^{k+1} - \widehat{x}\rangle_X \leq 0.$$

Applying now the three-point identity (9.1) for $x = x^{k+1}$, $y = x^k$, and $z = \widehat{x}$, yields

$$\frac{1}{2}\|x^{k+1} - \widehat{x}\|_X^2 + \frac{1}{2}\|x^{k+1} - x^k\|_X^2 \leq \frac{1}{2}\|x^k - \widehat{x}\|_X^2. \tag{9.5}$$

This shows the Fejér monotonicity of $\{x^k\}_{k\in\mathbb{N}}$ with respect to $\hat{X} = H^{-1}(0)$.

Furthermore, summing (9.5) over $k = 0, \dots, N-1$ gives

$$\frac{1}{2}\|x^N - \widehat{x}\|_X^2 + \sum_{k=0}^{N-1} \frac{1}{2}\|x^{k+1} - x^k\|_X^2 \leq \frac{1}{2}\|x^0 - \widehat{x}\|_X^2 =: C_0. \tag{9.6}$$

Writing $w^{k+1} := -\tau_k^{-1}(x^{k+1} - x^k)$, the implicit iteration (9.3) shows that $w^{k+1} \in H(x^{k+1})$. From (9.6) we also deduce that

$$\sum_{k=0}^{N-1} \tau_k^2 \|w^{k+1}\|_X^2 \leq 2C_0.$$

If $\tau_k \geq \varepsilon > 0$, letting $N \to \infty$ shows $\|w^{k+1}\|_X \to 0$. Otherwise, we can use Lemma 9.3 to establish the same.

Let finally $\bar{x}$ be any weak limit point of $\{x^k\}_{k\in\mathbb{N}}$, that is $x^{k_i} \rightharpoonup \bar{x}$ for a subsequence $\{k_i\}_{i\in\mathbb{N}} \subset \mathbb{N}$. Recall that $w^{k_i} \in H(x^{k_i})$. The weak-to-strong outer semicontinuity of H now immediately yields $0 \in H(\bar{x})$. We then finish by applying Opial's Lemma 9.1 for the set $\hat{X} = H^{-1}(0)$. □

Note that the conditions of Theorem 9.4 are in particularly satisfied if H is either maximally monotone (Lemma 6.10) or monotone and BCP outer semicontinuous (Lemma 6.12). In particular, applying Theorem 9.4 to $H = \partial J$ yields the convergence of the proximal point method (8.1) for any proper, convex, and lower semicontinuous functional $J : X \to \overline{\mathbb{R}}$.

Remark 9.5. A conventional way of proving the convergence of the proximal point method is with Browder's fixed-point theorem [Browder, 1965], which shows the existence of fixed points of firmly nonexpansive or, more generally, α-averaged mappings. (We have already shown in Lemma 6.15 the firm nonexpansivity of the proximal map.) On the other hand, to prove Browder's fixed-point theorem itself, we can use similar arguments as Theorem 9.4, see Theorem 9.22 below.

EXPLICIT SPLITTING

The convergence of the forward-backward splitting method

$$x^{k+1} = \operatorname{prox}_{\tau_k G}(x^k - \tau_k \nabla F(x^k)) \tag{9.7}$$

can be shown analogously. To do so, we need to assume the Lipschitz continuity of the gradient of F (since we are not using a proximal point mapping for F which is always firmly nonexpansive and hence Lipschitz continuous).

Theorem 9.6. *Let $F : X \to \mathbb{R}$ and $G : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. Suppose $(\partial(F+G))^{-1}(0) \neq \emptyset$, i.e., that $J := F + G$ has a minimizer. Furthermore, let F be Gâteaux differentiable with L-Lipschitz gradient. If $0 < \tau_{\min} \leq \tau_k \leq \tau_{\max} < 2L^{-1}$, then for any initial iterate $x^0 \in X$ the sequence generated by (9.7) converges weakly to a root $\widehat{x} \in (\partial(F+G))^{-1}(0)$.*

Proof. We again start by writing (9.7) in implicit form as

$$0 \in \tau_k[\partial G(x^{k+1}) + \nabla F(x^k)] + (x^{k+1} - x^k). \tag{9.8}$$

By the monotonicity of ∂G and the three-point monotonicity (7.9) of F from Corollary 7.2, we first deduce for any $\widehat{x} \in \hat{X} := (\partial(F+G))^{-1}(0)$ that

$$\langle \partial G(x^{k+1}) + \nabla F(x^k), x^{k+1} - \widehat{x} \rangle_X \geq -\frac{L}{4}\|x^{k+1} - x^k\|_X^2.$$

Thus, again testing (9.8) with $\langle \cdot, x^{k+1} - \widehat{x} \rangle_X$ yields

$$\langle x^{k+1} - x^k, x^{k+1} - \widehat{x} \rangle_X \leq \frac{L\tau_k}{4}\|x^{k+1} - x^k\|_X^2.$$

The three-point identity (9.1) now implies that

$$\frac{1}{2}\|x^{k+1} - \widehat{x}\|_X^2 + \frac{1 - \tau_k L/2}{2}\|x^{k+1} - x^k\|_X^2 \leq \frac{1}{2}\|x^k - \widehat{x}\|_X^2. \tag{9.9}$$

The assumption $2 > \tau_k L$ then establishes the Fejér monotonicity of $\{x^k\}_{k\in\mathbb{N}}$ with respect to $\hat{X}$. Let now $\bar{x}$ be a weak limit point of $\{x^k\}_{k\in\mathbb{N}}$, i.e., $x^{k_i} \rightharpoonup \bar{x}$ for a subsequence $\{k_i\}_{i\in\mathbb{N}} \subset \mathbb{N}$. Since (9.9) implies $x^{k+1} - x^k \to 0$ by the assumption on the step lengths, we have $\nabla F(x^{k_i+1}) - \nabla F(x^{k_i}) \to 0$ by the Lipschitz continuity of ∇F. Consequently, using again the subdifferential sum rule Theorem 4.14, $\partial(G+F)(x^{k_i+1}) \ni w^{k_i+1}+\nabla F(x^{k_i+1})-\nabla F(x^{k_i}) \to 0$. By the weak-to-strong outer semicontinuity of $\partial(G+F)$ from Lemma 6.10 and Theorem 6.13, it follows that $0 \in \partial(G+F)(\bar{x})$. We finish by applying Opial's Lemma 9.1 with $\hat{X} = (\partial(F+G))^{-1}(0)$. □

Remark 9.7. The L-Lipschitz requirement on ∇F can in some cases be restrictive, or upper estimates of L difficult to obtain. In the latter case, line search can be used, as we will discuss in Section 12.3. However, a slight modification of the forward-backward iteration can avoid the requirement entirely. Indeed, [Malitsky and Tam, 2020] prove the convergence of the *forward-reflected-backward* iteration $x^{k+1} := \operatorname{prox}_{\tau G}(x^k - 2\tau\nabla F(x^k) + \tau\nabla F(x^{k-1}))$, merely requiring F to be $\tilde{L}$-Lipschitz, and the fixed step length $0 < \tau < 1/(2\tilde{L})$.

9.3 PRECONDITIONED PROXIMAL POINT METHODS: DRS AND PDPS

We now extend the analysis of the previous section to the *preconditioned proximal point method* (8.11), which we recall can be written in implicit form as

$$0 \in H(x^{k+1}) + M(x^{k+1} - x^k) \tag{9.10}$$

for some preconditioning operator $M \in \mathbb{L}(X;X)$ and includes the Douglas–Rachford splitting (DRS) and the primal-dual proximal splitting (PDPS) methods as special cases. To deal with M, we need to improve Theorem 9.6 slightly. First, we introduce the preconditioned norm $\|x\|_M := \sqrt{\langle Mx, x\rangle}$, which satisfies the *preconditioned three-point identity*

$$\langle M(x-y), x-z\rangle = \frac{1}{2}\|x-y\|_M^2 - \frac{1}{2}\|y-z\|_M^2 + \frac{1}{2}\|x-z\|_M^2 \quad \text{for all } x, y, z \in X. \tag{9.11}$$

The boundedness assumption in the statement of the next theorem holds in particular for $M = \mathrm{Id}$ and H maximally monotone by Corollary 6.16.

Theorem 9.8. *Suppose $H : X \rightrightarrows X$ is monotone and weak-to-strong outer semicontinuous with $H^{-1}(0) \neq \emptyset$, that $M \in \mathbb{L}(X;X)$ is self-adjoint and positive semi-definite, and that either M has a bounded inverse, or $(H+M)^{-1} \circ M^{1/2}$ is bounded on bounded sets. Let the initial iterate $x^0 \in X$ be arbitrary, and assume that (9.10) has a unique solution x^{k+1} for all $k \in \mathbb{N}$. Then the iterates $\{x^k\}_{k\in\mathbb{N}}$ of (9.10) are bounded and satisfy $0 \in \limsup_{k\to\infty} H(x^k)$ and $M^{1/2}(x^k - \widehat{x}) \rightharpoonup 0$ for some $\widehat{x} \in H^{-1}(0)$.*

Proof. Let $\widehat{x} \in H^{-1}(0)$ be arbitrary. By the monotonicity of H, we then have as before

$$\langle H(x^{k+1}), x^{k+1} - \widehat{x} \rangle_X \geq 0,$$

which together with (9.10) yields

$$\langle M(x^{k+1} - x^k), x^{k+1} - \widehat{x} \rangle_X \leq 0. \tag{9.12}$$

Applying the preconditioned three-point identity (9.11) for $x = x^{k+1}$, $y = x^k$, and $z = \widehat{x}$ in (9.12) shows that

$$\frac{1}{2}\|x^{k+1} - \widehat{x}\|_M^2 + \frac{1}{2}\|x^{k+1} - x^k\|_M^2 \leq \frac{1}{2}\|x^k - \widehat{x}\|_M^2, \tag{9.13}$$

and summing (9.13) over $k = 0, \ldots, N-1$ yields

$$\frac{1}{2}\|x^N - \widehat{x}\|_M^2 + \sum_{k=0}^{N-1} \frac{1}{2}\|x^{k+1} - x^k\|_M^2 \leq \frac{1}{2}\|x^0 - \widehat{x}\|_M^2. \tag{9.14}$$

Let now $z^k := M^{1/2}x^k$. Our objective is then to show $z^k \rightharpoonup \bar{z}$ for some $\bar{z} \in \hat{Z} := M^{1/2}H^{-1}(0)$, which we do by using Opial's Lemma 9.1. From (9.13), we obtain the necessary Fejér monotonicity of $\{z^k\}_{k\in\mathbb{N}}$ with respect to the set $\hat{Z}$. It remains to verify that $\hat{Z}$ contains all weak limit points of $\{z^k\}_{k\in\mathbb{N}}$.

Let therefore $\bar{z}$ be such a limit point, i.e., $z^{k_i} \rightharpoonup \bar{z}$ for a subsequence $\{k_i\}_{i\in\mathbb{N}}$. We want to show that $\bar{z} = M^{1/2}\bar{x}$ for a weak limit point $\bar{x}$ of $\{x^k\}_{k\in\mathbb{N}}$. We proceed by first showing in two cases the boundedness of $\{x^k\}_{k\in\mathbb{N}}$:

(i) If M has a bounded inverse, then $M \geq \theta I$ for some $\theta > 0$, and thus the sequence $\{x^k\}_{k\in\mathbb{N}}$ is bounded by (9.14).

(ii) Otherwise, $(H + M)^{-1} \circ M^{1/2}$ is bounded on bounded sets. Now (9.14) only gives boundedness of $\{z^k\}_{k\in\mathbb{N}}$. However, $x^{k+1} \in (H + M)^{-1}(Mx^k) = (H + M)^{-1}(M^{1/2}z^k)$, and $\{z^k\}_{k\in\mathbb{N}}$ is bounded by (9.14), so we obtain the boundedness of $\{x^k\}_{k\in\mathbb{N}}$.

Thus there exists a further subsequence of $\{x^{k_i}\}_{i\in\mathbb{N}}$, weakly converging to some $\bar{x} \in X$. Since $z^k = M^{1/2}x^k$, it follows that $\bar{z} = M^{1/2}\bar{x}$. To show that $\bar{z} \in \hat{Z}$, if therefore suffices to show that the weak limit points of $\{x^k\}_{k\in\mathbb{N}}$ belong to $H^{-1}(0)$.

Let thus $\bar{x}$ be any weak limit point of $\{x^k\}_{k\in\mathbb{N}}$, i.e., $x^{k_i} \rightharpoonup \bar{x}$ for some subsequence $\{k_i\}_{k\in\mathbb{N}} \subset \mathbb{N}$. From (9.14), we obtain first that $M^{1/2}(x^{k+1} - x^k) \to 0$ and hence that $w^{k+1} := -M(x^{k+1} - x^k) \to 0$. From (8.18), we also know that $w^{k+1} \in H(x^{k+1})$. It follows that $0 = \lim_{k\to\infty} w^{k+1} \in \limsup_{k\to\infty} H(x^{k+1})$. The weak-to-strong outer semicontinuity now immediately yields $0 \in H(\bar{x})$. Hence, $\hat{Z}$ contains all weak limit points of $\{z^k\}_{k\in\mathbb{N}}$.

The claim now follows from Lemma 9.1. □

In the following, we verify that the DRS and PDPS methods satisfy the assumptions of this theorem.

DOUGLAS–RACHFORD SPLITTING

Recall that the DRS method (8.8), i.e.,

$$\begin{cases} x^{k+1} = \operatorname{prox}_{\tau F}(z^k), \\ y^{k+1} = \operatorname{prox}_{\tau G}(2x^{k+1} - z^k), \\ z^{k+1} = z^k + y^{k+1} - x^{k+1}, \end{cases} \tag{9.15}$$

can be written as the preconditioned proximal point method (8.11) in terms of $u = (x, y, z) \in U := X^3$ and the operators

$$H(x, y, z) := \begin{pmatrix} \tau B(x) + y - z \\ \tau A(y) + z - x \\ x - y \end{pmatrix} \quad \text{and} \quad M := \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I \end{pmatrix} \tag{9.16}$$

for $B = \partial F$ and $A = \partial G$. We are now interested in the properties of H in terms of those of A and B. For this, we can make use of the generic structure of H, which will reappear several times in the following.

Lemma 9.9. *If $A : X \rightrightarrows X$ is maximally monotone and $\Xi \in \mathbb{L}(X; X)$ is skew-adjoint (i.e., $\Xi^* = -\Xi$), then $H := A + \Xi$ is maximally monotone. In particular, any skew-adjoint operator Ξ is maximally monotone.*

Proof. Let $x, z^* \in X$ be given such that

$$\langle z^* - \tilde{z}^*, x - \tilde{x} \rangle_X \geq 0 \quad \text{for all } \tilde{x} \in X, \tilde{z}^* \in H(\tilde{x}).$$

Recalling (6.2), we need to show that $z^* \in H(x)$. By the definition of H, for any $\tilde{z}^* \in H(\tilde{x})$ there exists a $\tilde{x}^* \in A(\tilde{x})$ with $\tilde{z}^* = \tilde{x}^* + \Xi\tilde{x}$. On the other hand, setting $x^* := z^* - \Xi x$, we have $z^* = x^* + \Xi x$. We are thus done if we can show that $x^* \in A(x)$. But using the skew-adjointness of H and the symmetry of the inner product, we can write

$$\begin{aligned} 0 &\leq \langle z^* - \tilde{z}^*, x - \tilde{x} \rangle_X \\ &= \langle x^* - \tilde{x}^*, x - \tilde{x} \rangle_X + \langle \Xi(x - \tilde{x}), x - \tilde{x} \rangle_X \\ &= \langle x^* - \tilde{x}^*, x - \tilde{x} \rangle_X + \frac{1}{2}\langle \Xi(x - \tilde{x}), x - \tilde{x} \rangle_X - \frac{1}{2}\langle x - \tilde{x}, \Xi(x - \tilde{x}) \rangle_X \\ &= \langle x^* - \tilde{x}^*, x - \tilde{x} \rangle_X, \end{aligned} \tag{9.17}$$

and $x^* \in A(x)$ follows from the maximal monotonicity of A.

To prove the final claim about skew-adjoint operators being maximally monotone, we take $A = \{0\} = \partial S$ for the constant functional $S \equiv 0$, which is maximally monotone by Theorem 6.13. □

Corollary 9.10. *Let A and B be maximally monotone. Then the operator H defined in (9.16) is maximally monotone.*

Proof. Let

$$\tilde{A}(u) := \begin{pmatrix} \tau B(x) \\ \tau A(y) \\ 0 \end{pmatrix} \quad \text{and} \quad \Xi := \begin{pmatrix} 0 & \mathrm{Id} & -\mathrm{Id} \\ -\mathrm{Id} & 0 & \mathrm{Id} \\ \mathrm{Id} & -\mathrm{Id} & 0 \end{pmatrix}.$$

From the definition of the inner product on the product space X^3 together with Lemma 6.7, we have that $\tilde{A}$ is maximally monotone, while Ξ is clearly skew-adjoint. The claim now follows from Lemma 9.9. □

We can now show convergence of the DRS method.

Corollary 9.11. *Let $A, B : X \rightrightarrows X$ be maximally monotone, and suppose $(A+B)^{-1}(0) \neq \emptyset$. Pick a step length $\tau > 0$ and an initial iterate $z^0 \in X$. Then the iterates $\{(x^k, y^k, z^k)\}_{k\in\mathbb{N}}$ of the DRS method (9.15) converge weakly to $(\widehat{x}, \widehat{y}, \widehat{z}) \in H^{-1}(0)$ satisfying $\widehat{x} = \widehat{y} \in (A+B)^{-1}(0)$. Moreover, $x^k - y^k \to 0$.*

Proof. Since A and B are maximally monotone, Corollary 6.16 shows that the DRS iteration is always solvable for u^{k+1}. Regarding convergence, we start by proving that the sequence $\{u^k = (x^k, y^k, z^k)\}_{k\in\mathbb{N}}$ is bounded, $z^k \rightharpoonup \widehat{z}$ for some $\widehat{z}$, and $0 \in \limsup_{k\to\infty} H(u^k)$. Note that the latter implies as claimed that $x^k - y^k \to 0$ strongly. We do this using Theorem 9.8 whose conditions we have to verify. By Corollary 9.10, H is maximally monotone and hence weak-to-strong outer semicontinuous by Lemma 6.10. Since M is noninvertible, we also have to verify that $(H+M)^{-1} \circ M^{1/2}$ is bounded on bounded sets. But since $u^{k+1} \in (H+M)^{-1}(Mu^k) = (H+M)^{-1}(M^{1/2}u^k)$ is an equivalent formulation of the iteration (9.15), this follows from the Lipschitz continuity of the resolvent (Corollary 6.16). Hence, we can apply Theorem 9.8 to deduce $0 \in \limsup_{k\to\infty} H(u^k)$ as well as $M^{1/2}(u^k - \widehat{u}) \rightharpoonup 0$ for some $\widehat{u} = (\widehat{x}, \widehat{y}, \widehat{z})$ with $0 \in H(\widehat{u})$. By the definition of M, this gives $z^k \rightharpoonup \widehat{z}$. Moreover, the third line in the definition of H implies that $\widehat{x} = \widehat{y}$. Adding the first two lines in the same definition, we then obtain $0 \in A(\widehat{x}) + B(\widehat{x})$.

It remains to show weak convergence of the other variables. Since $\{u^k\}_{k\in\mathbb{N}}$ is bounded, it contains a subsequence converging weakly to some $\tilde{u} = (\tilde{x}, \tilde{y}, \tilde{z})$ which satisfies $0 \in H(\tilde{x}, \tilde{y}, \tilde{z})$ such that $\tilde{x} = \tilde{y}$. Since $z^k \rightharpoonup \widehat{z}$, we have $\tilde{z} = \widehat{z}$. The first relation of the inclusion then can be rearranged to $\tilde{y} = \tilde{x} = \mathcal{R}_{\tau B}(\tilde{z}) = \mathcal{R}_{\tau B}(\widehat{z})$ by the single-valuedness of the resolvent (Corollary 6.16). The limit is thus independent of the subsequence, and hence a subsequence–subsequence argument shows that the full sequence converges. □

In particular, this convergence result applies to the special case of $B = \partial F$ and $A = \partial G$ for proper, convex, lower semicontinuous $F, G : X \to \overline{\mathbb{R}}$. However, the fixed point provided by the DRS method is related to a solution of the problem $\min_{x\in X} F(x) + G(x)$ only if the subdifferential sum rule (Theorem 4.14) holds with equality.

PRIMAL-DUAL PROXIMAL SPLITTING

To study the PDPS method, we recall from (8.14) and (8.19) the operators

$$H(u) := \begin{pmatrix} \partial F(x) + K^* y \\ \partial G^*(y) - Kx \end{pmatrix}, \quad \text{and} \quad M := \begin{pmatrix} \tau^{-1}\mathrm{Id} & -K^* \\ -K & \sigma^{-1}\mathrm{Id} \end{pmatrix} \tag{9.18}$$

for $u = (x, y) \in X \times Y =: U$. With these we have already shown in Section 8.4 that the PDPS method

$$\begin{cases} x^{k+1} = \operatorname{prox}_{\tau F}(x^k - \tau K^* y^k), \\ \bar{x}^{k+1} = 2x^{k+1} - x^k, \\ y^{k+1} = \operatorname{prox}_{\sigma G^*}(y^k + \sigma K \bar{x}^{k+1}). \end{cases} \tag{9.19}$$

has the form (9.10) of the preconditioned proximal point method. To show convergence, we first have to establish some basic properties of both H and M.

Lemma 9.12. *The operator $M : U \to U$ defined in (8.19) is bounded and self-adjoint. If $\sigma\tau\|K\|^2_{\mathbb{L}(X;Y)} < 1$, then M is positive definite.*

Proof. The definition of M directly implies boundedness (since $K \in \mathbb{L}(X;Y)$ is bounded) and self-adjointness. Let now $u = (x, y) \in U$ be given. Then

$$\begin{aligned} \langle Mu, u\rangle_U &= \langle \tau^{-1}x - K^* y, x\rangle_X + \langle \sigma^{-1}y - Kx, y\rangle_Y \\ &= \tau^{-1}\|x\|_X^2 - 2\langle x, K^* y\rangle_X + \sigma^{-1}\|y\|_Y^2 \\ &\geq \tau^{-1}\|x\|_X^2 - 2\|K\|_{\mathbb{L}(X;Y)}\|x\|_X\|y\|_Y + \sigma^{-1}\|y\|_Y^2 \\ &\geq \tau^{-1}\|x\|_X^2 - \|K\|_{\mathbb{L}(X;Y)}\sqrt{\sigma\tau}(\tau^{-1}\|x\|_X^2 + \sigma^{-1}\|y\|_Y^2) + \sigma^{-1}\|y\|_Y^2 \\ &= (1 - \|K\|_{\mathbb{L}(X;Y)}\sqrt{\sigma\tau})(\tau^{-1}\|x\|_X^2 + \sigma^{-1}\|y\|_Y^2) \\ &\geq C(\|x\|_X^2 + \|y\|_Y^2) \end{aligned} \tag{9.20}$$

for $C := (1 - \|K\|_{\mathbb{L}(X;Y)}\sqrt{\sigma\tau})\min\{\tau^{-1}, \sigma^{-1}\} > 0$. Hence, $\langle Mu, u\rangle_U \geq C\|u\|_U^2$ for all $u \in U$, and therefore M is positive definite. □

Lemma 9.13. *The operator $H : U \rightrightarrows U$ defined in (9.18) is maximally monotone.*

Proof. Let $A(u) := \begin{pmatrix} \partial F(x) \\ \partial G^*(y) \end{pmatrix}$ and $\Xi := \begin{pmatrix} 0 & K^* \\ -K & 0 \end{pmatrix}$. Then Ξ is skew-adjoint, and A is maximally monotone by the definition of the inner product on $U = X \times Y$ and Theorem 6.13. The claim now follows from Lemma 9.9. □

With this, we can deduce the convergence of the PDPS method.

Corollary 9.14. *Let the convex, proper, and lower semicontinuous functions $F : X \to \overline{\mathbb{R}}$, $G : Y \to \overline{\mathbb{R}}$, and the linear operator $K \in \mathbb{L}(X;Y)$ satisfy the assumptions of Theorem 5.11. If, moreover, $\sigma\tau\|K\|^2_{\mathbb{L}(X;Y)} < 1$, then the sequence $\{u^k := (x^k, y^k)\}_{k\in\mathbb{N}}$ generated by the PDPS method (9.19) for any initial iterate $u^0 \in X \times Y$ converges weakly in U to a pair $\widehat{u} := (\widehat{x}, \widehat{y}) \in H^{-1}(0)$, i.e., satisfying (8.13).*

Proof. By Lemma 9.12, M is self-adjoint and positive definite and thus has a bounded inverse. Minding Lemma 9.13, we can therefore apply Theorem 9.8 to show that $(u^k - \widehat{u}) \rightharpoonup 0$ for some $\widehat{u} \in H^{-1}(0)$ with respect to the inner product $\langle M\bullet, \bullet\rangle_U$. Since M is has a bounded inverse, this implies that

$$\langle u^k, Mw\rangle_U = \langle Mu^k, w\rangle_U \to \langle M\widehat{u}, w\rangle_U = \langle \widehat{u}, Mw\rangle_U \quad \text{for all } w \in U$$

and hence $u^k \rightharpoonup \widehat{u}$ in U since $\operatorname{ran} M = U$ due to the invertibility of M. □

Remark 9.15. Through a general approach to degenerately preconditioned proximal point methods, i.e., singular M, [Bredies et al., 2022] prove the weak convergence of PDPS in the degenerate case $\tau\sigma\|K\|^2 = 1$. Like our Corollary 9.11, their approach also readily establishes the convergence of $\{(x^k, y^k, z^k)\}_{k\in\mathbb{N}}$ for the DRS; classical proofs only show the convergence of $\{z^k\}_{k\in\mathbb{N}}$.

9.4 PRECONDITIONED EXPLICIT SPLITTING METHODS: PDES AND MORE

Let $A, B : X \rightrightarrows X$ be monotone operators and consider the iterative scheme

$$0 \in A(x^{k+1}) + B(x^k) + M(x^{k+1} - x^k), \tag{9.21}$$

which is implicit in A but explicit in B. We obviously intend to use this method to find some $\widehat{x} \in (A + B)^{-1}(0)$.

As we have seen, the proximal point, PDPS, and DRS methods are all of the form (9.21) with $B = 0$. The basic explicit splitting method is also of this form with $A = \partial G$, $B = \nabla F$, and $M = \tau^{-1}\mathrm{Id}$. It is moreover not difficult to see from (8.25) that the primal-dual explicit splitting (PDES) method is also of the form (9.21) with nonzero B. So to prove the convergence of this algorithm, we want to improve Theorem 9.6 to be able to deal with the preconditioning operator M and the general monotone operators A and B in place of subdifferentials and gradients.

To proceed, we need a suitable notion of smoothness for B to be able to deal with the explicit step. In Theorem 9.6 we only used the Lipschitz continuity of ∇F in two places: first, to establish the three-point monotonicity using Corollary 7.2, and second, at the end of the proof for a continuity argument. To simplify dealing with B that may only act on a subspace, as in the case of the primal-dual explicit splitting in Section 8.5, we now make this three-point monotonicity with respect to an operator Λ our main assumption.

Specifically, we say that $B : X \rightrightarrows X$ is *three-point monotone* at $\widehat{x} \in X$ with respect to $\Lambda \in \mathbb{L}(X;X)$ if

$$\langle B(z) - B(\widehat{x}), x - \widehat{x}\rangle \geq -\frac{1}{4}\|z - x\|_\Lambda^2 \quad \text{for all } x, z \in X. \tag{9.22}$$

If this holds for every $\widehat{x}$, we say that B is *three-point monotone with respect to* Λ. From Corollary 7.2, it is clear that if ∇F is Lipschitz continuous with constant L, then $B = \nabla F$ is three-point monotone with respect to $\Lambda = L\,\mathrm{Id}$.

We again start with a lemma exploiting the structural properties of the saddle-point operator to show a "shifted outer semicontinuity".

Lemma 9.16. *Let $H = A + B : X \rightrightarrows X$ be weak-to-strong outer semicontinuous with B single-valued and Lipschitz continuous. If $w^{k+1} \in A(x^{k+1}) + B(z^k)$ for $k \in \mathbb{N}$ with $w^k \to \bar{w}$ and $x^{k+1} - z^k \to 0$ strongly in X and $x^k \rightharpoonup \bar{x}$ weakly in X, then $\bar{w} \in H(\bar{x})$.*

Proof. We have $w^{k+1} \in A(x^{k+1}) + B(z^k)$ so that

$$\tilde{w}^{k+1} := w^{k+1} - B(z^k) + B(x^{k+1}) \in H(x^{k+1}).$$

Since $w^{k+1} \to \bar{w}$ and $x^{k+1} - z^k \to 0$ and B is Lipschitz continuous, we have $\tilde{w}^{k+1} \to \bar{w}$ as well. The weak-to-strong outer semicontinuity of H then immediately yields $\bar{w} \in H(\bar{x})$. □

Theorem 9.17. *Let $H = A + B$ with $H^{-1}(0) \neq \emptyset$ for $A, B : X \rightrightarrows X$ with A monotone and B single-valued Lipschitz continuous and three-point monotone with respect to some $\Lambda \in \mathbb{L}(X;X)$. Furthermore, let $M \in \mathbb{L}(X;X)$ be self-adjoint, positive definite, with a bounded inverse, and satisfy $(2 - \varepsilon)M \geq \Lambda$ for some $\varepsilon > 0$. Suppose H is weak-to-strong outer semicontinuous. Let the starting point $x^0 \in X$ be arbitrary and assume that (9.21) has a unique solution x^{k+1} for every $k \in \mathbb{N}$. Then the iterates $\{x^k\}_{k\in\mathbb{N}}$ of (9.21) satisfy $M^{1/2}(x^k - \widehat{x}) \rightharpoonup 0$ for some $\widehat{x} \in H^{-1}(0)$.*

Proof. The proof follows along the same lines as that of Theorem 9.8 with minor modifications. First, since $0 \in H(\widehat{x})$, the monotonicity of A and the three-point monotonicity (9.22) of B yields

$$\langle A(x^{k+1}) + B(x^k), x^{k+1} - \widehat{x}\rangle \geq -\frac{1}{4}\|x^{k+1} - x^k\|_\Lambda^2,$$

which together with (9.21) leads to

$$\langle M(x^{k+1} - x^k), x^{k+1} - \widehat{x}\rangle \leq \frac{1}{4}\|x^{k+1} - x^k\|_\Lambda^2.$$

From the preconditioned three-point identity (9.11) we then obtain

$$\frac{1}{2}\|x^{k+1} - \widehat{x}\|_M^2 + \frac{1}{2}\|x^{k+1} - x^k\|_{M-\Lambda/2}^2 \leq \frac{1}{2}\|x^k - \widehat{x}\|_M^2. \tag{9.23}$$

Our assumption that $(2-\varepsilon)M \geq \Lambda$ implies that $M - \Lambda/2 \geq \varepsilon M/2$. By definition, we can therefore bound the second norm on the left-hand side from below to obtain (9.13) with an additional constant depending on ε. We may thus proceed as in the proof of Theorem 9.8 to establish $w^{k+1} := -M(x^{k+1} - x^k) \to 0$. We now have $w^{k+1} \in A(x^{k+1}) + B(x^k)$ and therefore use Lemma 9.16 with $z^k = x^k$ and $\bar{w} = 0$ to establish $0 \in H(\bar{x})$. The rest of the proof again proceeds as for Theorem 9.8 with the application of Opial's Lemma 9.1. □

We again apply this result to show the convergence of specific splitting methods containing an explicit step.

PRIMAL-DUAL EXPLICIT SPLITTING

We now return to algorithms for problems of the form

$$\min_{x \in X} F(x) + G(Kx)$$

for Gâteaux differentiable F and linear K. Recall from (8.23) the primal-dual explicit splitting (PDES) method

$$\begin{cases} y^{k+1} = \operatorname{prox}_{G^*}((\mathrm{Id} - KK^*)y^k + K(x^k - \nabla F(x^k))), \\ x^{k+1} = x^k - \nabla F(x^k) - K^* y^{k+1}, \end{cases} \tag{9.24}$$

which can be written in implicit form as

$$0 \in H(u^{k+1}) + \begin{pmatrix} \nabla F(x^k) - \nabla F(x^{k+1}) \\ 0 \end{pmatrix} + M(u^{k+1} - u^k) \tag{9.25}$$

with

$$H(u) := \begin{pmatrix} \partial F(x) + K^* y \\ \partial G^*(y) - Kx \end{pmatrix} \quad \text{and} \quad M := \begin{pmatrix} \mathrm{Id} & 0 \\ 0 & \mathrm{Id} - KK^* \end{pmatrix}, \tag{9.26}$$

for $u = (x, y) \in X \times Y =: U$.

Corollary 9.18. *Let $F : X \to \mathbb{R}$ and $G : Y \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $K \in \mathbb{L}(X; Y)$. Suppose F is Gâteaux differentiable with L-Lipschitz gradient for $L < 2$, that $\|K\|_{\mathbb{L}(X;Y)} < 1$, and that the assumptions of Theorem 5.11 are satisfied. Then for any initial iterate $u^0 \in X \times Y$ the iterates $\{u^k = (x^k, y^k)\}_{k \in \mathbb{N}}$ of the (8.23) converge weakly to some $\widehat{u} \in H^{-1}(0)$ with H given by (8.14).*

Proof. We recall that Theorem 5.11 guarantees that $H^{-1}(0) \neq \emptyset$. To apply Theorem 9.17, we write $H = A + B$ for

$$A(u) := \begin{pmatrix} 0 \\ \partial G^*(y) \end{pmatrix} + \Xi u, \quad B(u) := \begin{pmatrix} \nabla F(x) \\ 0 \end{pmatrix}, \quad \Xi := \begin{pmatrix} 0 & K^* \\ -K & 0 \end{pmatrix}.$$

We first note that M as given in (9.26) is self-adjoint and positive definite under our assumption $\|K\|_{\mathbb{L}(X;Y)} < 1$. By Corollary 7.2, the three-point monotonicity (9.22) holds for $\Lambda := \left(\begin{smallmatrix} L & 0 \\ 0 & 0 \end{smallmatrix}\right)$. Since $L < 2$, there furthermore exists an $\varepsilon > 0$ sufficiently small such that $(2-\varepsilon)M \geq \Lambda$. Finally, Lemma 9.13 shows that H is maximally monotone and hence weak-to-strong outer semicontinuous by Lemma 6.10. The claim now follows from Theorem 9.17. □

Remark 9.19. It is possible to improve the result to $\|K\| \leq 1$ if we increase the complexity of Theorem 9.17 slightly to allow for $M \geq 0$. However, in this case it is only possible to show the convergence of the partial iterates $\{x^k\}_{k\in\mathbb{N}}$.

PRIMAL-DUAL PROXIMAL SPLITTING WITH AN ADDITIONAL FORWARD STEP

Using a similar switching term as in the implicit formulation (9.25) of the PDES method, it is possible to incorporate additional forward steps in the PDPS method. For $F = F_0 + E$ with F_0, E convex and E Gâteaux differentiable, we therefore consider

$$\min_{x\in X} F_0(x) + E(x) + G(Kx). \tag{9.27}$$

With $u = (x, y)$ and following Section 8.4, any minimizer $\widehat{x} \in X$ satisfies $0 \in H(\widehat{u})$ for

$$H(u) := \begin{pmatrix} \partial F(x) + \nabla E(x) + K^* y \\ \partial G^*(y) - Kx \end{pmatrix}. \tag{9.28}$$

Similarly, following the arguments in Section 8.4, we can show that the iteration

$$\begin{cases} x^{k+1} = \operatorname{prox}_{\tau F_0}(x^k - \tau\nabla E(x^k) - \tau K^* y^k), \\ \bar{x}^{k+1} = 2x^{k+1} - x^k, \\ y^{k+1} = \operatorname{prox}_{\sigma G^*}(y^k + \sigma K\bar{x}^{k+1}), \end{cases} \tag{9.29}$$

is equivalent to the implicit formulation

$$0 \in \begin{pmatrix} \partial F_0(x^{k+1}) + \nabla E(x^k) + K^* y^{k+1} \\ \partial G(y^{k+1}) - Kx^{k+1} \end{pmatrix} + M(u^{k+1} - u^k)$$

with the preconditioner M defined as in (9.18). The convergence can thus be shown as for the PDES method.

Corollary 9.20. *Let $E : X \to \mathbb{R}$, $F_0 : X \to \overline{\mathbb{R}}$, and $G : Y \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $K \in \mathbb{L}(X;Y)$. Suppose E is Gâteaux differentiable with an L-Lipschitz gradient, and that the assumptions of Theorem 5.11 are satisfied with $F := F_0 + E$. Assume, moreover, that $\tau, \sigma > 0$ satisfy*

$$1 > \|K\|^2_{\mathbb{L}(X;Y)}\tau\sigma + \tau\frac{L}{2}. \tag{9.30}$$

Then for any initial iterate $u^0 \in X \times Y$ the iterates $\{u^k\}_{k\in\mathbb{N}}$ of (9.29) converge weakly to some $\widehat{u} \in H^{-1}(0)$ for H given by (9.28).

Proof. As before, Theorem 5.11 guarantees that $H^{-1}(0) \neq \emptyset$. We apply Theorem 9.17 to

$$A(u) := \begin{pmatrix} \partial F_0(x) \\ \partial G^*(y) \end{pmatrix} + \Xi u, \quad B(u) := \begin{pmatrix} \nabla E(x) \\ 0 \end{pmatrix}, \quad \Xi := \begin{pmatrix} 0 & K^* \\ -K & 0 \end{pmatrix},$$

and M given by (9.18). By Corollary 7.2, the three-point monotonicity (9.22) holds with $\Lambda := \left(\begin{smallmatrix} L & 0 \\ 0 & 0 \end{smallmatrix}\right)$. We have already shown in Lemma 9.12 that M is self-adjoint and positive definite. Furthermore, from (9.20) in the proof of Lemma 9.12, we have

$$\langle Mu, u \rangle \geq (1 - \|K\|_{\mathbb{L}(X;Y)} \sqrt{\sigma\tau})(\tau^{-1}\|x\|_X^2 + \sigma^{-1}\|y\|_Y^2)$$

Thus (9.30) implies that M is positive definite. Arguing similarly to (9.20), we also estimate

$$\langle Mu, u \rangle_U \geq \tau^{-1}\|x\|_X^2 - 2\|K\|_{\mathbb{L}(X;Y)}\|x\|_X\|y\|_Y + \sigma^{-1}\|y\|_Y^2 \geq (1 - \|K\|_{\mathbb{L}(X;Y)}^2 \sigma\tau)\tau^{-1}\|x\|_X^2.$$

By the strict inequality in (9.30), we thus deduce $(2 - \varepsilon)M \geq \Lambda$ for some $\varepsilon > 0$.

Now by Lemma 9.13, H is again maximally monotone and therefore weak-to-strong outer semicontinuous by Lemma 6.10, and the claim follows from Theorem 9.17. □

Remark 9.21. The forward step was introduced to the basic PDPS method in [Condat, 2013; Vũ, 2013], see also [Chambolle and Pock, 2015]. These papers also introduced an additional over-relaxation step that we will discuss in Chapter 12.

9.5 FIXED-POINT THEOREMS

Based on our generic approach, we now prove the classical *Browder fixed-point theorem*, which can itself be used to prove the convergence of optimization methods and other fixed-point iterations (see Remark 9.5). We recall from Lemma 6.18 that firmly nonexpansive maps are $(1/2)$-averaged, so the result applies by Lemma 6.15 to the resolvents of maximally monotone maps in particular – hence proving the convergence of the proximal point method.

Theorem 9.22 (Browder Fixed-point Theorem). *On a Hilbert space X, suppose $T : X \to X$ is α-averaged for some $\alpha \in (0, 1)$ and has a fixed point $\widehat{x} = T(\widehat{x})$. Let $x^{k+1} := T(x^k)$. Then $x^k \rightharpoonup \bar{x}$ weakly in X for some fixed point $\bar{x}$ of T.*

Proof. Finding a fixed point of T is equivalent to finding a root of $H(x) := T(x) - x$. Similarly, we can rewrite the fixed-point iteration as solving for x^{k+1} the inclusion

$$0 = x^k - T(x^k) + (x^{k+1} - x^k). \tag{9.31}$$

Proceeding as in the previous sections, we test this by the application of $\langle \cdot, x^{k+1} - \widehat{x}\rangle_X$. After application of the three-point identity (9.1), we then obtain

$$\frac{1}{2}\|x^{k+1} - \widehat{x}\|_X^2 + \frac{1}{2}\|x^{k+1} - x^k\|_X^2 + \langle x^k - T(x^k), x^{k+1} - \widehat{x}\rangle_X \leq \frac{1}{2}\|x^k - \widehat{x}\|_X^2. \tag{9.32}$$

Since $x^{k+1} = T(x^k)$, $\widehat{x}$ is a fixed point of T, and by assumption $T = (1-\alpha)\mathrm{Id} + \alpha J$ for some nonexpansive operator $J : X \to X$, we have

$$\begin{aligned}
\langle x^k - T(x^k), x^{k+1} - \widehat{x}\rangle_X &= \langle x^k - \widehat{x} - (T(x^k) - T(\widehat{x})), T(x^k) - T(\widehat{x})\rangle_X \\
&= \alpha\langle x^k - \widehat{x} - (J(x^k) - J(\widehat{x})), (1-\alpha)(x^k - \widehat{x}) + \alpha(J(x^k) - J(\widehat{x}))\rangle_X \\
&= (\alpha - \alpha^2)\|x^k - \widehat{x}\|_X^2 - \alpha^2\|J(x^k) - J(\widehat{x})\|_X^2 \\
&\quad + (2\alpha^2 - \alpha)\langle x^k - \widehat{x}, J(x^k) - J(\widehat{x})\rangle_X
\end{aligned}$$

as well as

$$\begin{aligned}
\frac{1}{2}\|x^{k+1} - x^k\|_X^2 = \frac{1}{2}\|T(x^k) - x^k\|_X^2 = \frac{\alpha^2}{2}\|J(x^k) - x^k\|_X^2 = \frac{\alpha^2}{2}\|J(x^k) - J(\widehat{x}) - (x^k - \widehat{x})\|_X^2 \\
= \frac{\alpha^2}{2}\|x^k - \widehat{x}\|_X^2 + \frac{\alpha^2}{2}\|J(x^k) - J(\widehat{x})\|_X^2 - \alpha^2\langle x^k - \widehat{x}, J(x^k) - J(\widehat{x})\rangle_X.
\end{aligned}$$

Thus, for any $\delta > 0$,

$$\begin{aligned}
\frac{1-\delta}{2}\|x^{k+1} - x^k\|_X^2 + \langle x^k - T(x^k), x^{k+1} - \widehat{x}\rangle_X &= ((1+\delta)\alpha^2 - \alpha)\langle x^k - \widehat{x}, J(x^k) - J(\widehat{x})\rangle_X \\
&\quad + \frac{2\alpha - (1+\delta)\alpha^2}{2}\|x^k - \widehat{x}\|_X^2 - \frac{(1+\delta)\alpha^2}{2}\|J(x^k) - J(\widehat{x})\|_X^2.
\end{aligned}$$

Taking $\delta = \frac{1}{\alpha} - 1$, we have $\delta > 0$ and $\alpha = (1+\delta)\alpha^2$. Thus the factor in front of the inner product term is positive, and hence we obtain by the nonexpansivity of J

$$\frac{1-\delta}{2}\|x^{k+1} - x^k\|_X^2 + \langle x^k - T(x^k), x^{k+1} - \widehat{x}\rangle_X = \frac{\alpha}{2}\|x^k - \widehat{x}\|_X^2 - \frac{\alpha}{2}\|J(x^k) - J(\widehat{x})\|_X^2 \geq 0.$$

From (9.32), it now follows that

$$\frac{1}{2}\|x^{k+1} - \widehat{x}\|_X^2 + \frac{\delta}{2}\|x^{k+1} - x^k\|_X^2 \leq \frac{1}{2}\|x^k - \widehat{x}\|_X^2.$$

As before, this implies Fejér monotonicity of $\{x^k\}_{k\in\mathbb{N}}$ and that $\|x^{k+1} - x^k\|_X \to 0$. The latter implies $\|T(x^k) - x^k\|_X \to 0$ via (9.31). Let $\bar{x}$ be any weak limit point of $\{x^k\}_{k\in\mathbb{N}}$. Denote by $N \subset \mathbb{N}$ be the indices of the corresponding subsequence. We show that $\bar{x}$ is a fixed point of T. Since by Lemma 6.20 the set of fixed points is convex and closed, the claim then follows from Opial's Lemma 9.1.

To show that $\bar{x}$ is a fixed point of T, first, we expand

$$\frac{1}{2}\|x^k - T(\bar{x})\|_X^2 = \frac{1}{2}\|x^k - \bar{x}\|_X^2 + \frac{1}{2}\|\bar{x} - T(\bar{x})\|_X^2 + \langle x^k - \bar{x}, \bar{x} - T(\bar{x})\rangle_X.$$

Since $x^k \rightharpoonup \bar{x}$, this gives

$$\limsup_{N \ni k \to \infty} \frac{1}{2}\|x^k - T(\bar{x})\|_X^2 \geq \limsup_{N \ni k \to \infty} \frac{1}{2}\|x^k - \bar{x}\|_X^2 + \|\bar{x} - T(\bar{x})\|_X^2.$$

On the other hand, by the nonexpansivity of T and $T(x^k) - x^k \to 0$, we have

$$\limsup_{N \ni k \to \infty} \|x^k - T(\bar{x})\|_X \leq \limsup_{N \ni k \to \infty} \Big(\|T(x^k) - T(\bar{x})\|_X + \|x^k - T(x^k)\|_X \Big) \leq \limsup_{N \ni k \to \infty} \|x^k - \bar{x}\|_X.$$

Together this two inequalities show, as desired, that $\|T(\bar{x}) - \bar{x}\| = 0$. □

Remark 9.23. Theorem 9.22 in its modern form (stated for firmly nonexpansive or more generally α-averaged maps) can be first found in [Browder, 1967]. However, similar results for what are now called *Krasnosel′skiĭ–Mann iterations* – which are closely related to α-averaged maps – were stated in more limited settings in [Krasnosel′skiĭ, 1955; Mann, 1953; Opial, 1967; Petryshyn, 1966; Schaefer, 1957]. Our overall approach in this book, based on [Valkonen, 2020b], is an "implicit" counterpart to the more classical fixed point theorems. Instead of considering explicit iterations $x^{k+1} := T(x^k)$, the theory is based on x^{k+1} defined implicitly through equations $0 = H(x^{k+1}) + (x^{k+1} - x^k)$.

10 SPLITTING METHODS: RATES OF CONVERGENCE

As we have seen, minimizers of convex problems in a Hilbert space X can generally be characterized by the inclusion

$$0 \in H(\widehat{x})$$

for the unknown $\widehat{x} \in X$ and a suitable monotone operator $H : X \rightrightarrows X$. This inclusion in turn can be solved using a (preconditioned) proximal point iteration that converges weakly under suitable assumptions. In the present chapter, we want to improve this analysis to obtain *convergence rates*, i.e., estimates of the distance $\|x^k - \widehat{x}\|_X$ of iterates to $\widehat{x}$ in terms of the iteration number k. Our general approach will be to consider this distance multiplied by an iteration-dependent *testing parameter* φ_k (or, for structured algorithms, consider the norm relative to a *testing operator*) and to show by roughly the same arguments as in Chapter 9 that this product stays bounded: $\varphi_k \|x^k - \widehat{x}\|_X \leq C$. If we can then show that this testing parameter grows at a certain rate, the distance must decay at the reciprocal rate. Consequently, we can now avoid the complications of dealing with weak convergence; in fact, this chapter will consist of simple algebraic manipulations. However, for this to work we need to assume additional properties of H, namely strong monotonicity. Recall from Lemma 7.4 that H is called strongly monotone with factor $\gamma > 0$ if

$$\langle H(\tilde{x}) - H(x), \tilde{x} - x \rangle_X \geq \gamma \|\tilde{x} - x\|_X^2 \quad (\tilde{x}, x \in X), \tag{10.1}$$

where, in a slight abuse of notation, the left-hand side is understood to stand for any choice of elements from $H(\tilde{x})$ and $H(x)$.

Before we turn to the actual estimates, we first define various notions of convergence rates. Consider a function $r : \mathbb{N} \to [0, \infty)$ (e.g., $r(k) = \|x^k - \widehat{x}\|_X$ or $r(k) = G(x^k) - G(\widehat{x})$ for $\widehat{x}$ a minimizer of G).

(i) We say that $r(k)$ *converges (to zero as $k \to \infty$) at the rate $O(f(k))$* if $r(k) \leq C f(k)$ for some constant $C > 0$ for all $k \in \mathbb{N}$ and a decreasing function $f : \mathbb{N} \to [0, \infty)$ with $\lim_{k\to\infty} f(k) = 0$ (e.g., $f(k) = 1/k$ or $f(k) = 1/k^2$).

(ii) Analogously, we say that a function $R : \mathbb{N} \to [0, \infty)$ *grows at the rate $\Omega(F(k))$* if $R(k) \geq cF(k)$ for all $k \in \mathbb{N}$ for some constant $c > 0$ and an increasing function $F : \mathbb{N} \to [0, \infty)$ with $\lim_{k\to\infty} F(k) = \infty$.

Clearly $r = 1/R$ converges to zero at the rate $f = 1/F$ if and only if R grows at the rate F. The most common cases are $F(k) = k$ or $F(k) = k^2$.

We can alternatively characterize *orders* of convergence via

$$\mu := \lim_{k\to\infty} \frac{r(k+1)}{r(k)}.$$

(i) If $\mu = 1$, we say that $r(k)$ converges (to zero as $k \to \infty$) *sublinearly*.

(ii) If $\mu \in (0, 1)$, then this convergence is *linear*. This is equivalent to a convergence at the rate $O(\tilde{\mu}^k)$ for any $\tilde{\mu} \in (\mu, 1)$.

(iii) If $\mu = 0$, then the convergence is *superlinear*.

Different rates of superlinear convergence can also be studied. We say that $r(k)$ converges (to zero as $k \to \infty$) *superlinearly with order* $q > 1$ if

$$\lim_{k\to\infty} \frac{r(k+1)}{r(k)^q} < \infty.$$

The most common case is $q = 2$, which is also known as *quadratic convergence*. (This is not to be confused with the – much slower – convergence at the rate $O(1/k^2)$; similarly, linear convergence is different from – and much faster than – convergence at the rate $O(1/k)$.)

10.1 THE FUNDAMENTAL METHODS

Before going into this abstract operator-based theory, we demonstrate the general concept of testing by studying the fundamental methods, the proximal point and explicit splitting methods. These are purely primal methods with a single step length parameter, which simplifies the testing approach since we only need a single testing parameter. (It should be pointed out that the proofs in this section can be carried out – and in fact shortened – without introducing testing parameters at all. Nevertheless, we follow this approach since it provides a blueprint for the proofs for the structured primal-dual methods where these are required.)

PROXIMAL POINT METHOD

We start with the basic proximal point method for solving $0 \in H(\widehat{x})$ for a monotone operator $H : X \rightrightarrows X$, which we recall can be written in implicit form as

$$0 \in \tau_k H(x^{k+1}) + (x^{k+1} - x^k). \tag{10.2}$$

Theorem 10.1 (proximal point method iterate rates). *Suppose $H : X \rightrightarrows X$ is strongly monotone with $H^{-1}(0) \neq \emptyset$. Let $x^{k+1} := \mathcal{R}_{\tau_k H}(x^k)$ for some $\{\tau_k\}_{k\in\mathbb{N}} \subset (0,\infty)$ and $x^0 \in X$ be arbitrary. Then the following hold for the iterates $\{x^k\}_{k\in\mathbb{N}}$ and the unique point $\widehat{x} \in H^{-1}(0)$:*

(i) If $\tau_k \equiv \tau$ is constant, then $\|x^k - \widehat{x}\|_X \to 0$ linearly.

(ii) If $\tau_k \nearrow \infty$, then $\|x^k - \widehat{x}\|_X \to 0$ superlinearly.

Proof. Let $\widehat{x} \in H^{-1}(0)$; by assumption, such a point exists and is unique due to the assumed strong monotonicity of H (since inserting any two roots $\hat{x}, \tilde{x} \in X$ of H in (10.1) yields $\|\hat{x} - \tilde{x}\|_X \leq 0$). For each iteration $k \in \mathbb{N}$, pick a *testing parameter* $\varphi_k > 0$ and apply the *test* $\varphi_k \langle \cdot, x^{k+1} - \widehat{x}\rangle_X$ to (10.2) to obtain (using the same notation from Theorem 9.4)

$$0 \in \varphi_k \tau_k \langle H(x^{k+1}), x^{k+1} - \widehat{x}\rangle_X + \varphi_k \langle x^{k+1} - x^k, x^{k+1} - \widehat{x}\rangle_X. \tag{10.3}$$

By the strong monotonicity of H, and the fact that $0 \in H(\widehat{x})$, for some $\gamma > 0$,

$$\langle H(x^{k+1}), x^{k+1} - \widehat{x}\rangle_X \geq \gamma \|x^{k+1} - \widehat{x}\|_X^2$$

Multiplying this inequality with $\varphi_k \tau_k$ and using (10.3), we obtain

$$\varphi_k \tau_k \gamma \|x^{k+1} - \widehat{x}\|_X^2 + \varphi_k \langle x^{k+1} - x^k, x^{k+1} - \widehat{x}\rangle_X \leq 0.$$

An application of the three-point identity (9.1) then yields

$$\frac{\varphi_k(1 + 2\tau_k\gamma)}{2} \|x^{k+1} - \widehat{x}\|_X^2 + \frac{\varphi_k}{2} \|x^{k+1} - x^k\|_X^2 \leq \frac{\varphi_k}{2} \|x^k - \widehat{x}\|_X^2. \tag{10.4}$$

Let us now force on the testing parameters the recursion

$$\varphi_0 = 1, \qquad \varphi_{k+1} = \varphi_k(1 + 2\tau_k\gamma). \tag{10.5}$$

Then (10.4) yields

$$\frac{\varphi_{k+1}}{2} \|x^{k+1} - \widehat{x}\|_X^2 + \frac{\varphi_k}{2} \|x^{k+1} - x^k\|_X^2 \leq \frac{\varphi_k}{2} \|x^k - \widehat{x}\|_X^2. \tag{10.6}$$

We now distinguish the two cases for the step sizes τ_k.

(i) Summing (10.6) for $k = 0, \ldots, N-1$ gives

$$\frac{\varphi_N}{2} \|x^N - \widehat{x}\|_X^2 + \sum_{k=0}^{N-1} \frac{\varphi_k}{2} \|x^{k+1} - x^k\|_X^2 \leq \frac{\varphi_0}{2} \|x^0 - \widehat{x}\|_X^2.$$

In particular, $\varphi_0 = 1$ implies that

$$\|x^N - \widehat{x}\|_X^2 \leq \varphi_N^{-1} \|x^0 - \widehat{x}\|_X^2.$$

Since $\tau_k \equiv \tau$, (10.5) implies that $\varphi_N = (1 + 2\tau\gamma)^N$. Setting $\tilde{\mu} := (1 + 2\tau\gamma)^{-1/2} < 1$ now gives convergence at the rate $O(\tilde{\mu}^{-N})$ and therefore the claimed linear rate.

(ii) From (10.6) combined with (10.5) follows directly that

$$\frac{\|x^{k+1} - \widehat{x}\|_X^2}{\|x^k - \widehat{x}\|_X^2} \leq \frac{\varphi_k}{\varphi_{k+1}} = (1 + 2\tau_k\gamma)^{-1} \to 0$$

since $\tau_k \to \infty$, which implies the claimed superlinear convergence of $\|x^k - \widehat{x}\|_X$. (A similar argument can be used to directly show linear convergence for constant step sizes.) □

EXPLICIT SPLITTING

We now return to problems of the form

$$\min_{x \in X} F(x) + G(x) \tag{10.7}$$

for Gâteaux differentiable F, and study the convergence rates of the explicit (or forward–backward) splitting method

$$x^{k+1} := \operatorname{prox}_{\tau G}(x^k - \tau \nabla F(x^k)), \tag{10.8}$$

which we recall can be written in implicit form as

$$0 \in \tau[\partial G(x^{k+1}) + \nabla F(x^k)] + (x^{k+1} - x^k). \tag{10.9}$$

Theorem 10.2 (explicit splitting iterate rates). *Let $F : X \to \mathbb{R}$ and $G : X \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous. Suppose further that F is Gâteaux differentiable, ∇F is Lipschitz continuous with constant $L > 0$, and G is γ-strongly convex for some $\gamma > 0$. If $[\partial(F+G)]^{-1}(0) \neq \emptyset$ and the step length parameter $\tau > 0$ satisfies $\tau L \leq 2$, then for any initial iterate $x^0 \in X$ the iterates $\{x^k\}_{k\in\mathbb{N}}$ generated by the explicit splitting method (10.8) converge linearly to the unique minimizer of (10.7).*

Proof. Let $\widehat{x} \in [\partial(F+G)]^{-1}(0)$; by assumption, such a point exists and is unique due to the strong and therefore strict convexity of G. As in the proof of Theorem 10.1, for each iteration $k \in \mathbb{N}$, pick a *testing parameter* $\varphi_k > 0$ and apply the *test* $\varphi_k\langle \cdot, x^{k+1} - \widehat{x}\rangle$ to (10.9) to obtain

$$0 \in \varphi_k\tau\langle \partial G(x^{k+1}) + \nabla F(x^k), x^{k+1} - \widehat{x}\rangle_X + \varphi_k\langle x^{k+1} - x^k, x^{k+1} - \widehat{x}\rangle_X. \tag{10.10}$$

Since G is strongly convex, it follows from (10.9) and Lemma 7.4 (iii) that

$$\langle \partial G(x^{k+1}) - \partial G(\widehat{x}), x^{k+1} - \widehat{x}\rangle_X \geq \gamma\|x^{k+1} - \widehat{x}\|_X^2.$$

Similarly, since ∇F is Lipschitz continuous, it follows from Corollary 7.2 that

$$\langle \nabla F(x^k) - \nabla F(\widehat{x}), x^{k+1} - \widehat{x}\rangle_X \geq -\frac{L}{4}\|x^{k+1} - x^k\|_X^2.$$

Combining the last two inequalities with $0 \in \partial G(\widehat{x}) + \nabla F(\widehat{x})$, we obtain

$$\langle \partial G(x^{k+1}) + \nabla F(x^k), x^{k+1} - \widehat{x} \rangle_X \geq \gamma \|x^{k+1} - \widehat{x}\|_X^2 - \frac{L}{4}\|x^{k+1} - x^k\|_X^2. \tag{10.11}$$

Inserting this into (10.10) and using the three-point identity, as in the proof of Theorem 10.1, we now obtain

$$\frac{\varphi_k(1+2\tau\gamma)}{2}\|x^{k+1} - \widehat{x}\|_X^2 + \frac{\varphi_k(1-\tau L/2)}{2}\|x^{k+1} - x^k\|_X^2 \leq \frac{\varphi_k}{2}\|x^k - \widehat{x}\|_X^2. \tag{10.12}$$

Since $1 - \tau L/2 \geq 0$, summing over $k = 0, \dots, N-1$, we arrive at

$$\frac{\varphi_N}{2}\|x^N - \widehat{x}\|_X^2 \leq \frac{\varphi_0}{2}\|x^0 - \widehat{x}\|_X^2.$$

As in Theorem 10.1, the definition of φ_k shows that $\|x^k - \widehat{x}\|_X^2 \longrightarrow 0$ linearly. □

Observe that it is not possible to obtain superlinear convergence in this case since the assumption $\tau_k \leq 2L^{-1}$ forces the step lengths to remain bounded.

10.2 STRUCTURED ALGORITHMS AND ACCELERATION

We now to extend the analysis above to the structured case where $H = A + B$, since we have already seen that most common first-order algorithm can be written as calculating in each step the next iterate x^{k+1} from a specific instance of the general preconditioned implicit–explicit splitting method

$$0 \in A(x^{k+1}) + B(x^k) + M(x^{k+1} - x^k). \tag{10.13}$$

In the proofs of convergence of the proximal point and explicit splitting methods (e.g., in Theorems 10.1 and 10.2 as well as in Chapter 9), we had the step length τ_k in front of H or $\nabla F + \partial G$. On the other hand, in Section 9.3 on structured algorithms, we incorporated the step length parameters into the preconditioning operator M. To transfer the testing approach from these fundamental methods to the structured methods, we will now split them out from M and move them in front of H as well by introducing a *step length operator* W_{k+1}. We will also allow the preconditioner M_{k+1} to vary by iteration; as we will see below, this is required for accelerated versions of the PDPS method. Correspondingly, we consider the scheme

$$0 \in W_{k+1}[A(x^{k+1}) + B(x^k)] + M_{k+1}(x^{k+1} - x^k). \tag{10.14}$$

Since we now have a step length operator instead of a single scalar step length, we will also have to consider instead of a scalar testing parameter an iteration-dependent *testing operator* $Z_{k+1} \in \mathbb{L}(X;X)$. The rough idea is that $Z_{k+1}M$ – or, as needed for accelerated

algorithms, $Z_{k+1}M_{k+1}$ – will form a "local norm" that measures the rate of convergence in a nonuniform way; and rather than testing the (scalar) three-point identity (10.4), we will build the testing already into the initial strong monotonicity inequality. We therefore require an operator-level version of strong monotonicity, which we introduce next.

Let $A : X \rightrightarrows X$ and let $Z, \Gamma \in \mathbb{L}(X;X)$ be such that $Z\Gamma$ is positive semi-definite. Then we say that A is Γ-*strongly monotone* at $\widehat{x} \in X$ with respect to Z if

$$\langle A(x) - A(\widehat{x}), x - \widehat{x}\rangle_Z \geq \|x - \widehat{x}\|^2_{Z\Gamma} \quad (x \in X). \tag{10.15}$$

If this holds for all $\widehat{x} \in X$, we say that A is Γ-strongly monotone with respect to Z.

It is clear that strongly monotone operators with parameter $\gamma > 0$ are $\gamma \cdot \mathrm{Id}$-strongly monotone with respect to $Z = \mathrm{Id}$. More generally, operators with a separable block-structure, $A(x) = (A_1(x_1), \ldots, A_n(x_n))$ for $x = (x_1, \ldots, x_n)$ satisfy the property, as as illustrated in more detail in the next example for the two-block case.

Example 10.3. Let $A(x) = (A_1(x_1), A_2(x_2))$ for $x = (x_1, x_2) \in X_1 \times X_2$ and the monotone operators $A_1 : X_1 \rightrightarrows X_1$ and $A_2 : X_2 \rightrightarrows X_2$. Suppose A_1 and A_2 are, respectively γ_1- and γ_2-(strongly) monotone for $\gamma_1, \gamma_2 \geq 0$. Then (10.15) holds for any $\varphi_1, \varphi_2 > 0$ for

$$\Gamma := \begin{pmatrix} \gamma_1 \mathrm{Id} & 0 \\ 0 & \gamma_2 \mathrm{Id} \end{pmatrix} \quad \text{and} \quad Z := \begin{pmatrix} \varphi_1 \mathrm{Id} & 0 \\ 0 & \varphi_2 \mathrm{Id} \end{pmatrix}$$

We do not impose $Z\Gamma$ to be self-adjoint in (10.15), although we use the norm notation. Forgoing with self-adjointness allows Γ to have skew-adjoint parts $\Xi = -\Xi^*$, cf. Lemma 9.9. Indeed, for the operator H for the PDPS method from (8.14), we can for $Z = \mathrm{Id}$ always choose $\Gamma = \left(\begin{smallmatrix} 0 & K^* \\ -K & 0 \end{smallmatrix}\right)$ skew-adjoint. With either F or G^* strongly convex, Γ will also have corresponding components as in Example 10.3.

Let further $B : X \rightrightarrows X$ and let $Z, \Lambda \in \mathbb{L}(X;X)$ be such that $Z\Lambda$ is positive semi-definite. Then we say that B is *three-point monotone* at $\widehat{x} \in X$ with respect to Z and Λ if

$$\langle B(z) - B(\widehat{x}), x - \widehat{x}\rangle_Z \geq -\frac{1}{4}\|x - z\|^2_{Z\Lambda} \quad (x, z \in X). \tag{10.16}$$

If this holds for all $\widehat{x} \in X$, we say that B is three-point monotone with respect to Z and Λ.

Example 10.4. Let $B(x) = (\nabla E_1(x_1), \nabla E_2(x_2))$ for $x = (x_1, x_2) \in X_1 \times X_2$ and the respectively L_1- and L_2-smooth convex functions $E_1 : X_1 \to \mathbb{R}$ and $E_2 : X_2 \to \mathbb{R}$. Then a referral to Corollary 7.2 shows (10.16) to hold for any $\varphi_1, \varphi_2 > 0$ for

$$\Lambda := \begin{pmatrix} L_1 \mathrm{Id} & 0 \\ 0 & L_2 \mathrm{Id} \end{pmatrix} \quad \text{and} \quad Z := \begin{pmatrix} \varphi_1 \mathrm{Id} & 0 \\ 0 & \varphi_2 \mathrm{Id} \end{pmatrix}$$

More generally, we can take $B(x) = (B_1(x_1), B_2(x_2))$ for $B_1 : X_1 \to X_1$ and $B_2 : X_2 \to X_2$ three-point monotone as defined in (7.9).

Clearly Example 10.4 as Example 10.3 generalizes to a large number of blocks, and both to operators acting separably on more general direct sums of orthogonal subspaces.

We are now ready to forge our hammer for producing convergence rates for structured algorithms. In the following, for any $M, N \in \mathbb{L}(X;X)$, we write $M \succeq N$ to mean that $M - N$ is positive semi-definite: $\|x\|_M^2 \geq \|x\|_N^2$ for all $x \in X$.

Theorem 10.5. *Let $A, B : X \rightrightarrows X$ and $H := A+B$. For each $k \in \mathbb{N}$, let further $Z_{k+1}, W_{k+1}, M_{k+1} \in \mathbb{L}(X;X)$ be such that $Z_{k+1}M_{k+1}$ is self-adjoint and positive semi-definite. Assume that there exists a $\widehat{x} \in H^{-1}(0)$. For each $k \in \mathbb{N}$, suppose for some $\Gamma, \Lambda \in \mathbb{L}(X;X)$ that A is Γ-strongly monotone at $\widehat{x}$ with respect to $Z_{k+1}W_{k+1}$ and that B is three-point monotone at $\widehat{x}$ with respect to $Z_{k+1}W_{k+1}$ and Λ. Let the initial iterate $x^0 \in X$ be arbitrary, and suppose $\{x^{k+1}\}_{k\in\mathbb{N}}$ are generated by (10.14). If for every $k \in \mathbb{N}$ both*

$$Z_{k+1}(M_{k+1} + 2W_{k+1}\Gamma) \succeq Z_{k+2}M_{k+2} \quad \text{and} \tag{10.17}$$

$$Z_{k+1}M_{k+1} \succeq Z_{k+1}W_{k+1}\Lambda/2. \tag{10.18}$$

hold, then

$$\frac{1}{2}\|x^N - \widehat{x}\|^2_{Z_{N+1}M_{N+1}} \leq \frac{1}{2}\|x^0 - \widehat{x}\|^2_{Z_1M_1}. \tag{10.19}$$

Proof. For brevity, denote $\widetilde{H}_{k+1}(x^{k+1}) := W_{k+1}[A(x^{k+1})+B(x^k)]$. First, from (10.15) and (10.16) we have that

$$\langle \widetilde{H}_{k+1}(x^{k+1}), x^{k+1} - \widehat{x}\rangle_{Z_{k+1}} \geq \|x^{k+1} - \widehat{x}\|^2_{Z_{k+1}W_{k+1}\Gamma} - \frac{1}{4}\|x^k - x^{k+1}\|^2_{Z_{k+1}W_{k+1}\Lambda}. \tag{10.20}$$

Multiplying (10.14) with Z_{k+1} and rearranging, we obtain

$$-Z_{k+1}M_{k+1}(x^{k+1} - x^k) \in Z_{k+1}\widetilde{H}_{k+1}(x^{k+1}).$$

Inserting this into (10.20) and applying the preconditioned three-point formula (9.11) for $M = Z_{k+1}M_{k+1}$ yields

$$\frac{1}{2}\|x^{k+1} - \widehat{x}\|^2_{Z_{k+1}(M_{k+1}+2W_{k+1}\Gamma)} + \frac{1}{2}\|x^{k+1} - x^k\|^2_{Z_{k+1}(M_{k+1}-W_{k+1}\Lambda/2)} \leq \frac{1}{2}\|x^k - \widehat{x}\|^2_{Z_{k+1}M_{k+1}}.$$

Using (10.17) and (10.18), this implies that

$$\frac{1}{2}\|x^{k+1} - \widehat{x}\|^2_{Z_{k+2}M_{k+2}} \leq \frac{1}{2}\|x^k - \widehat{x}\|^2_{Z_{k+1}M_{k+1}}. \tag{10.21}$$

Summing over $k = 0, \ldots, N-1$ now yields the claim. □

The inequality (10.21) is a quantitative or *variable metric* version of the Fejér monotonicity of Lemma 9.1 (i) with respect to $\hat{X} = \{\widehat{x}\}$.

If Theorem 10.5 is applicable, we immediately obtain the convergence rate result.

Corollary 10.6 (convergence with a rate). *If* (10.19) *holds and* $Z_{N+1}M_{N+1} \geq \mu(N)I$ *for some* $\mu : \mathbb{N} \to \mathbb{R}$, *then* $\|x^N - \widehat{u}\|^2 \to 0$ *at the rate* $O(1/\mu(N))$.

PRIMAL-DUAL PROXIMAL SPLITTING METHODS

We now apply this operator-testing technique to primal-dual splitting methods for the solution of

$$\min_{x \in X} F_0(x) + E(x) + G(Kx) \tag{10.22}$$

with $F_0 : X \to \overline{\mathbb{R}}$, $E : X \to \mathbb{R}$, and $G : Y \to \overline{\mathbb{R}}$ convex, proper, and lower semicontinuous and $K \in \mathbb{L}(X;Y)$. We will also write $F := F_0 + E$. The methods include in particular the PDPS method with a forward step (9.29). Now allowing varying step lengths and an over-relaxation parameter ω_k, this can be written

$$\begin{cases} x^{k+1} := (I + \tau_k \partial F_0)^{-1}(x^k - \tau_k K^* y^k - \tau_k \nabla E(x^k)), \\ \bar{x}^{k+1} := \omega_k(x^{k+1} - x^k) + x^{k+1}, \\ y^{k+1} := (I + \sigma_{k+1}\partial G^*)^{-1}(y^k + \sigma_{k+1} K \bar{x}^{k+1}). \end{cases} \tag{10.23}$$

For the basic version of the algorithm with $\omega_k = 1$, $\tau_k \equiv \tau_0 > 0$, and $\sigma_k \equiv \sigma_0 > 0$, we have seen in Corollary 9.20 that the iterates converge weakly if the step length parameters satisfy

$$L\tau_0/2 + \tau_0\sigma_0\|K\|^2_{\mathbb{L}(X;Y)} < 1, \tag{10.24}$$

where L is the Lipschitz constant of ∇E. We will now show that under strong convexity of F_0, we can choose these parameters to *accelerate* the algorithm to yield convergence at a rate $O(1/N^2)$. If both F_0 and G^* are strongly convex, we can even obtain linear convergence. Throughout $\widehat{u} = (\widehat{x}, \widehat{y})$ denotes a root of

$$H(u) := \begin{pmatrix} \partial F_0(x) + \nabla E(x) + K^* y \\ \partial G^*(y) - Kx \end{pmatrix},$$

which we assume exists. From Theorem 5.11, this is the case if an interior point condition is satisfied for $G \circ K$ and (10.22) admits a solution.

We will also require the following technical lemma in place of the simpler growth argument for the choice (10.5).

Lemma 10.7. *Pick $\varphi_0 > 0$ arbitrarily, and define iteratively $\varphi_{k+1} := \varphi_k\big(1 + 2\gamma\varphi_k^{-1/2}\big)$ for some $\gamma > 0$. Then there exists a constant $c > 0$ such that $\varphi_k \geq \big(ck + \varphi_0^{1/2}\big)^2$ for all $k \in \mathbb{N}$.*

Proof. Replacing φ_k by $\varphi_k' := \gamma^{-2}\varphi_k$, we may assume without loss of generality that $\gamma = 1$. We claim that $\varphi_k^{1/2} \geq ck + \varphi_0^{1/2}$ for some $c > 0$. We proceed by induction. The case $k = 0$ is clear. If the claim holds for $k = 0, \dots, N-1$, we can unroll the recursion to obtain the estimate

$$\varphi_N - \varphi_0 = \sum_{k=0}^{N-1} 2\varphi_k^{1/2} \geq 2\sum_{k=0}^{N-1} ck + 2\varphi_0^{1/2}N = cN(N-1) + 2\varphi_0^{1/2}N = cN^2 + (2\varphi_0^{1/2} - c)N.$$

Expanding $(cN + \varphi_0^{1/2})^2 = c^2N^2 + 2c\varphi_0^{1/2}N + \varphi_0$, we see that the claim for φ_N holds if $c \geq c^2$ and $2\varphi_0^{1/2} - c \geq 2c\varphi_0^{1/2}$. Taking the latter with equality and solving for c yields $c = 2\varphi_0^{1/2}/(1 + 2\varphi_0^{1/2}) < 1$ and hence also the former. Since this choice of c does not depend on N, the claim follows. □

Theorem 10.8 (accelerated and linearly convergent PDPS). *Let $F_0 : X \to \overline{\mathbb{R}}$, $E : X \to \mathbb{R}$ and $G : Y \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous with ∇E Lipschitz continuous with constant $L > 0$. Also let $K \in \mathbb{L}(X; Y)$, and suppose the assumptions of Theorem 5.11 are satisfied with $F := F_0 + E$. Pick initial step lengths $\tau_0, \sigma_0 > 0$ subject to (10.24). For any initial iterate $u^0 \in X \times Y$, suppose $\{u^{k+1} = (x^{k+1}, y^{k+1})\}_{k\in\mathbb{N}}$ are generated by (10.23).*

(i) If F_0 is strongly convex with factor $\gamma > 0$, and we take

$$\omega_k := 1/\sqrt{1 + 2\gamma\tau_k}, \quad \tau_{k+1} := \tau_k\omega_k, \quad \text{and} \quad \sigma_{k+1} := \sigma_k/\omega_k, \tag{10.25}$$

then $\|x^N - \widehat{x}\|_X^2 \to 0$ at the rate $O(1/N^2)$.

(ii) If both F_0 and G^* are strongly convex with factor $\gamma > 0$ and $\rho > 0$, respectively, and we take

$$\omega_k := 1/(1 + 2\theta), \quad \theta := \min\{\rho\sigma_0, \gamma\tau_0\}, \quad \tau_k := \tau_0 \quad \text{and} \quad \sigma_k := \sigma_0, \tag{10.26}$$

then $\|x^N - \widehat{x}\|_X^2 + \|y^N - \widehat{y}\|_Y^2 \to 0$ linearly.

Proof. Recalling Corollary 9.20, we write (10.23) in the form (10.14) by taking

$$A(u) := \begin{pmatrix} \partial F_0(x) \\ \partial G^*(y) \end{pmatrix} + \Xi u, \quad B(u) := \begin{pmatrix} \nabla E(x) \\ 0 \end{pmatrix}, \quad \Xi := \begin{pmatrix} 0 & K^* \\ -K & 0 \end{pmatrix},$$
$$W_{k+1} := \begin{pmatrix} \tau_k\mathrm{Id} & 0 \\ 0 & \sigma_{k+1}\mathrm{Id} \end{pmatrix}, \quad \text{and} \quad M_{k+1} := \begin{pmatrix} \mathrm{Id} & -\tau_k K^* \\ -\omega_k\sigma_{k+1}K & \mathrm{Id} \end{pmatrix}.$$

As before, Theorem 5.11 guarantees that $H^{-1}(0) \neq \emptyset$. For some primal and dual testing parameters $\varphi_k, \psi_{k+1} > 0$, we also take as our testing operator

$$Z_{k+1} := \begin{pmatrix} \varphi_k \mathrm{Id} & 0 \\ 0 & \psi_{k+1}\mathrm{Id} \end{pmatrix}. \tag{10.27}$$

By Examples 10.3 and 10.4, A is then Γ-strongly monotone with respect to $Z_{k+1}W_{k+1}$ and B is three-point monotone with respect to $Z_{k+1}W_{k+1}$ and Λ for

$$\Gamma := \Xi + \begin{pmatrix} \gamma \mathrm{Id} & 0 \\ 0 & \rho \mathrm{Id} \end{pmatrix}, \quad \text{and} \quad \Lambda := \begin{pmatrix} L\,\mathrm{Id} & 0 \\ 0 & 0 \end{pmatrix},$$

where $\rho = 0$ if G^* is not strongly convex.

We will apply Theorem 10.5. Taking $\omega_k := \sigma_{k+1}^{-1}\psi_{k+1}^{-1}\varphi_k\tau_k$, we expand

$$Z_{k+1}M_{k+1} = \begin{pmatrix} \varphi_k \mathrm{Id} & -\varphi_k\tau_k K^* \\ -\varphi_k\tau_k K & \psi_{k+1}\mathrm{Id} \end{pmatrix}. \tag{10.28}$$

Thus $Z_{k+1}M_{k+1}$ is self-adjoint as required. We still need to show that it is nonnegative and indeed grows at a rate that gives our claims. We also need to verify (10.17) and (10.18), which expand as

$$\begin{pmatrix} (\varphi_k(1+2\gamma\tau_k) - \varphi_{k+1})\mathrm{Id} & (\varphi_k\tau_k + \varphi_{k+1}\tau_{k+1})K^* \\ (\varphi_{k+1}\tau_{k+1} - 2\psi_{k+1}\sigma_{k+1} - \varphi_k\tau_k)K & (\psi_{k+1}(1+2\rho\sigma_{k+1}) - \psi_{k+2})\mathrm{Id} \end{pmatrix} \geq 0, \quad \text{and} \tag{10.29}$$

$$\begin{pmatrix} \varphi_k(1-\tau_k L/2)\mathrm{Id} & -\varphi_k\tau_k K^* \\ -\varphi_k\tau_k K & \psi_{k+1}\mathrm{Id} \end{pmatrix} \geq 0. \tag{10.30}$$

We now proceed backward by deriving the step length rules as sufficient conditions for these two inequalities. First, clearly (10.29) holds if for all $k \in \mathbb{N}$ we can guarantee that

$$\varphi_{k+1} \leq \varphi_k(1+2\gamma\tau_k), \quad \psi_{k+1} \leq \psi_k(1+2\rho\sigma_k), \quad \text{and} \quad \varphi_k\tau_k = \psi_k\sigma_k. \tag{10.31}$$

We deal with (10.30) and the lower bounds on $Z_{k+1}M_{k+1}$ in one go. By Young's inequality, we have for any $\delta \in (0,1)$ that

$$2\varphi_k\tau_k\langle Kx, y\rangle \leq (1-\delta)\varphi_k\|x\|^2 + \varphi_k\tau_k^2(1-\delta)^{-1}\|K^*y\|^2 \quad (x \in X,\ y \in Y),$$

hence recalling (10.28) also

$$Z_{k+1}M_{k+1} \geq \begin{pmatrix} \delta\varphi_k\mathrm{Id} & 0 \\ 0 & \psi_{k+1}\mathrm{Id} - \varphi_k\tau_k^2(1-\delta)^{-1}KK^* \end{pmatrix}. \tag{10.32}$$

Similarly, for the operator from (10.30), we have

$$\begin{pmatrix} \varphi_k(1-\tau_k L/2)\mathrm{Id} & -\varphi_k\tau_k K^* \\ -\varphi_k\tau_k K & \psi_{k+1}\mathrm{Id} \end{pmatrix} \geq \begin{pmatrix} \varphi_k(\delta - \tau_k L/2)\mathrm{Id} & 0 \\ 0 & \psi_{k+1}\mathrm{Id} - \varphi_k\tau_k^2(1-\delta)^{-1}K^*K \end{pmatrix}.$$

The condition (10.30) is therefore satisfied and $Z_{k+1}M_{k+1} \geq \varepsilon Z_{k+1}$ for some $\varepsilon > 0$ if (10.31) holds and both

$$\delta\varphi_k \geq \varepsilon\varphi_k + \varphi_k\tau_k L/2 \quad \text{and} \quad \psi_{k+1} \geq \varepsilon\psi_{k+1} + \varphi_k\tau_k^2(1-\delta)^{-1}\|K\|^2. \tag{10.33}$$

By (10.31), $\psi_{k+1} \geq \psi_k$, so using also using the last part of (10.31), we see (10.33) to hold if

$$\delta - \varepsilon \geq \tau_k L/2 \quad \text{and} \quad (1-\delta)(1-\varepsilon) \geq \tau_k\sigma_k\|K\|^2. \tag{10.34}$$

If we choose τ_k and σ_k such that their product stays constant (i.e., $\tau_k\sigma_k = \sigma_0\tau_0$), then the second equality holds for $\delta = 1 - \sigma_0\tau_0\|K\|^2/(1-\varepsilon)$, which has to be positive. Inserting this into the first part of (10.34), we see that it to hold if $1 \geq \sigma_0\tau_0\|K\|^2/(1-\varepsilon) + \varepsilon + \tau_k L/2$. This holds for some $\varepsilon > 0$ due to the assumed (10.24), i.e., $\tau_k L/2 + \sigma_0\tau_0\|K\|^2 < 1$. Since $\{\tau_k\}_{k\in\mathbb{N}}$ is nonincreasing, we see that (10.34) and hence (10.30) is satisfied when the initialization condition (10.24) holds.

To apply Theorem 10.5, all that remains is to verify (10.31) and that $\tau_k\sigma_k = \tau_0\sigma_0$. To obtain convergence rates, we need to further study the rate of increase of φ_k and ψ_{k+1}, which we recall that we wish to make as high as possible.

(i) If $\gamma > 0$ and $\rho = 0$, the best possible choice allowed by (10.31) is $\psi_k \equiv \psi_0$ and $\varphi_{k+1} = \varphi_k(1 + 2\gamma\tau_k)$ with $\sigma_k = \varphi_k\tau_k/\psi_0$. Together with the condition $\tau_k\sigma_k = \sigma_0\tau_0$, this forces $\sigma_0\tau_0 = \varphi_k\tau_k^2/\psi_0$. If we take $\psi_0 = 1/(\sigma_0\tau_0)$, we thus need $\tau_k = \varphi_k^{-1/2}$. Since $\sigma_{k+1} = \sigma_0\tau_0/\tau_{k+1} = 1/(\psi_0\tau_{k+1})$, we obtain the relations

$$\omega_k = \frac{\varphi_k\tau_k}{\sigma_{k+1}\psi_{k+1}} = \frac{\varphi_k^{1/2}}{\varphi_{k+1}^{1/2}} = \frac{1}{\sqrt{1+2\gamma\tau_k}},$$

which are satisfied for the choices of ω_k, τ_{k+1}, and σ_{k+1} in (10.25).

We now use Theorem 10.5 and Corollary 10.6 and (10.32) to obtain

$$\frac{\delta\varphi_N}{2}\|x^N - \widehat{x}\|_X^2 \leq \frac{1}{2}\|u^N - \widehat{u}\|_{Z_{N+1}M_{N+1}}^2 \leq C_0 := \frac{1}{2}\|u^0 - \widehat{u}\|_{Z_1M_1}^2.$$

Although this does not tell us anything about the convergence of the dual iterates $\{y^N\}_{N\in\mathbb{N}}$ as $\psi_N \equiv \psi$ stays constant, Lemma 10.7 shows that the primal test φ_N grows at the rate $\Omega(N^2)$ Hence we obtain the claimed convergence of the primal iterates at the rate $O(1/N^2)$.

(ii) If $\gamma > 0$ and $\rho > 0$ and we take $\tau_k \equiv \tau_0$ and $\sigma_k \equiv \sigma_0$, the last condition of (10.31) forces $\psi_k = \varphi_k\tau_0/\sigma_0$. Inserting this into the second condition yields $\varphi_{k+1} \leq \varphi_k(1 + 2\rho\sigma_0)$. Together with the first condition, we therefore at best can take $\varphi_{k+1} = \varphi_k(1+2\theta)$ for $\theta := \min\{\rho\sigma_0, \gamma\tau_0\}$. Reversing the roles of ψ and φ, we see that we can at best take $\psi_{k+1} = \psi_k(1+2\theta)$. This leads to the relations

$$\omega_k = \frac{\varphi_k\tau_0}{\sigma_0\psi_{k+1}} = \frac{\varphi_k}{\varphi_{k+1}} = \frac{1}{1+2\theta},$$

which are again satisfied by the respective choices in (10.26).

We finish the proof with Theorem 10.5 and Corollary 10.6, observing now from (10.32) that $Z_N M_N \geq C(1+2\theta)^N \mathrm{Id}$ for some $C > 0$. □

Note that if $\gamma = 0$ and $\rho = 0$, (10.31) forces $\varphi_k \equiv \varphi_0$ as well as $\psi_k \equiv \psi_0$. If we take $\varphi_k \equiv 1$, then we also have to take $\tau_k = \sigma_k \psi_0$. We can use this to define ψ_0 if we also fix $\tau_k \equiv \tau_0$ and $\sigma_k \equiv \sigma_0$. This also forces $\omega_k \equiv 1$. We thus again arrive at (10.31) as well as $\tau_k \sigma_k = \sigma_0 \tau_0$. However, we cannot obtain from this convergence rates for the iterates, merely boundedness and hence weak convergence as in Section 9.4.

11 SPLITTING METHODS: GAPS AND ERGODIC RESULTS

We continue with the testing framework introduced in Chapter 10 for proving rates of convergence of iterates of optimization methods. This generally required strong convexity, which is not always available. In this chapter, we use the testing idea to derive convergence rates of objective function values and other, more general, *gap functionals* that indicate algorithm convergence more indirectly than iterate convergence. This can be useful in cases where we can only obtain weak convergence of iterates, but can obtain rates of convergence of such a gap functional. Nevertheless, this gap convergence often will only be *ergodic*, i.e., the estimates only apply to a weighted sum of the history of iterates instead of the most recent iterate. In fact, we will first derive ergodic estimates for all algorithms. If we can additionally show that the algorithm is monotonic with respect to this gap, we can improve the ergodic estimate to the nonergodic ones as in the previous chapters.

11.1 GAP FUNCTIONALS

We recall that one of the three fundamental ingredients in the convergence proofs of Chapter 9 was the monotonicity of H (with one of the points fixed to a root $\widehat{x}$). We now modify this requirement to be able to prove estimates on the convergence of function values when $H = \partial F$ for some proper, convex, and lower semicontinuous $F : X \to \overline{\mathbb{R}}$. In this case, by the definition of the convex subdifferential,

$$\langle \partial F(x^{k+1}), x^{k+1} - \bar{x} \rangle_X \geq F(x^{k+1}) - F(\bar{x}) \quad (\bar{x} \in X). \tag{11.1}$$

On the other hand, for an L-smooth functional $G : X \to \mathbb{R}$, we can use the three-point estimates of Corollary 7.2 to obtain

$$\langle \nabla G(x^k), x^{k+1} - \bar{x} \rangle_X \geq G(x^{k+1}) - G(\bar{x}) - \frac{1}{2L}\|x^{k+1} - x^k\|_X^2 \quad (\bar{x} \in X). \tag{11.2}$$

These two inequalities are enough to obtain function value estimates for the more general case $H = \partial F + \nabla G$ including a forward step with respect to G. We will produce such estimates in Section 11.2.

GENERIC GAP FUNCTIONALS

More generally, when H does not directly arise from subdifferentials or gradients but has a more complicated structure, we introduce several *gap functionals*. We identified in Chapter 9 that for some lifted functionals $\tilde{F}$ and $\tilde{G}$ and a skew-adjoint operator $\Xi = -\Xi^*$, the unaccelerated PDPS, PDES, and DRS consist in taking $H = \partial\tilde{F} + \nabla\tilde{G} + \Xi$ and iterating

$$0 \in \partial\tilde{G}(x^{k+1}) + \nabla\tilde{F}(x^k) + \Xi x^{k+1} + M(x^{k+1} - x^k), \tag{11.3}$$

where the skew-adjoint operator Ξ does not arise as a subdifferential of any function. Working with this requires extra effort, especially when we later study accelerated methods.

Note that by the skew-adjointness of Ξ, we have $\langle \Xi\widehat{x}, \widehat{x}\rangle_X = 0$. Using this and the estimates (11.1) and (11.2) on $\tilde{F}$ and $\tilde{G}$, we obtain for the basic unaccelerated scheme (11.3) the estimate

$$\langle \partial\tilde{G}(x^{k+1}) + \nabla\tilde{F}(x^k) + \Xi x^{k+1}, x^{k+1} - \widehat{x}\rangle_X \geq \tilde{\mathcal{G}}(x; \widehat{x}) - \frac{1}{2L}\|x^{k+1} - x^k\|_X^2$$

with the *generic gap functional*

$$\tilde{\mathcal{G}}(x; \bar{x}) := (\tilde{G} + \tilde{F})(x) - (\tilde{G} + \tilde{F})(\bar{x}) + \langle \Xi\bar{x}, x\rangle_X. \tag{11.4}$$

In the next lemma, we collect some elementary properties of this functional. Note that $\tilde{\mathcal{G}}(x, z) = 0$ is possible even for $x \neq z$.

Lemma 11.1. *Let $H := \partial\tilde{F} + \nabla\tilde{G} + \Xi$, where $\Xi \in \mathbb{L}(X; X)$ is skew-adjoint and $\tilde{G} : X \to \overline{\mathbb{R}}$ and $\tilde{F} : X \to \mathbb{R}$ are convex, proper, and lower semicontinuous. If $\widehat{x} \in H^{-1}(0)$, then $\tilde{\mathcal{G}}(\,\cdot\,; \widehat{x}) \geq 0$ and $\tilde{\mathcal{G}}(\widehat{x}; \widehat{x}) = 0$.*

Proof. We first note that $\widehat{x} \in H^{-1}(0)$ is equivalent to $-\Xi\widehat{x} \in \partial(\tilde{F} + \tilde{G})(\widehat{x})$. Hence using the definition of the convex subdifferential and the fact that $\langle \Xi\widehat{x}, \widehat{x}\rangle_X = 0$ due to the skew-adjointness of Ξ, we deduce for arbitrary $x \in X$ that

$$(\tilde{F} + \tilde{G})(x) - (\tilde{F} + \tilde{G})(\widehat{x}) \geq \langle -\Xi\widehat{x}, x - \widehat{x}\rangle_X = \langle -\Xi\widehat{x}, x\rangle_X,$$

i.e., $\tilde{\mathcal{G}}(x, \widehat{x}) \geq 0$. The fact that $\tilde{\mathcal{G}}(\widehat{x}, \widehat{x}) = 0$ follows immediately from the skew-adjointness of Ξ. □

The function value estimates (11.1) and (11.2) – unlike simple monotonicity-based nonnegativity estimates – do not depend on $\bar{x}$ being a root of H. Therefore, taking any *bounded* set $B \subset X$ such that $H^{-1}(0) \cap B \neq \emptyset$, we see that the *partial gap*

$$\tilde{\mathcal{G}}(x; B) := \sup_{\bar{x} \in B} \tilde{\mathcal{G}}(x; \bar{x})$$

also satisfies $\tilde{\mathcal{G}}(\,\cdot\,; B) \geq 0$.

THE LAGRANGIAN DUALITY GAP

Let us now return to the problem

$$\min_{x\in X} F(x) + G(Kx), \tag{11.5}$$

where we split $F = F_0 + E$ assuming E to have a Lipschitz-continuous gradient. With the notation $u = (x, y)$, we recall that Theorem 5.11 guarantees the existence of a primal-dual solution $\widehat{u}$ whenever its conditions are satisfied. This, we further recall, can be written as $0 \in H(\widehat{u})$ for

$$H(u) := \begin{pmatrix} \partial F(x) + K^* y \\ \partial G^*(y) - Kx \end{pmatrix}. \tag{11.6a}$$

As we have already seen in, e.g., Theorem 10.8, we can express this choice of H in the present framework with

$$\tilde{F}(u) := F_0(x) + G(y), \quad \tilde{G}(u) := E(x), \quad \text{and} \quad \Xi := \begin{pmatrix} 0 & K^* \\ -K & 0 \end{pmatrix}. \tag{11.6b}$$

With this, the generic gap functional $\tilde{\mathcal{G}}$ from (11.4) becomes the *Lagrangian duality gap*

$$\mathcal{G}_L(u;\bar{u}) := \big(F(x) + \langle \bar{y}, Kx\rangle - G^*(\bar{y})\big) - \big(F(\bar{x}) + \langle y, K\bar{x}\rangle - G^*(y)\big) \le \bar{\mathcal{G}}(u), \tag{11.7}$$

where

$$\bar{\mathcal{G}}(u) := F(x) + G(Kx) + F^*(-K\bar{y}) + G^*(\bar{y})$$

is the real *duality gap*, cf. (5.16). As Corollary 5.14 shows, when its conditions are satisfied and $\bar{u} = \widehat{u} \in H^{-1}(0)$, the Lagrangian duality gap is nonnegative.

Since (11.1) and (11.2) do not depend on $\bar{x}$ being a root of H, convergence results for the Lagrangian duality gap can sometimes be improved slightly by taking any bounded set $B \subset X \times Y$ such that $B \cap H^{-1}(0) \neq \emptyset$ and defining the *partial duality gap*

$$\mathcal{G}(u;B) := \sup_{\bar{u}\in B} \mathcal{G}_L(u;\bar{u}). \tag{11.8}$$

This satisfies $0 \le \mathcal{G}(u;B) \le \bar{\mathcal{G}}(u)$. Moreover, by the definition of F^* and $G^{**} = G$, we have $\mathcal{G}(u; X \times Y) = \bar{\mathcal{G}}(u)$, which explains both the importance of partial duality gaps and the term "partial gap".

BREGMAN DIVERGENCES AND GAP FUNCTIONALS

Although we will not need this in the following, we briefly discuss a possible extension to Banach spaces. Let $J : X \to \overline{\mathbb{R}}$ be convex on a Banach space X. Then for $x \in \operatorname{dom} J$ and $p \in \partial J(x)$, one can define the asymmetric *Bregman divergence* (or *distance*)

$$B_J^p(z, x) := J(z) - J(x) - \langle p, z - x\rangle_X, \quad (x \in X).$$

Due to the definition of the convex subdifferential, this is nonnegative. It is also possible to symmetrize the distance by considering $\tilde{B}_J(x,z) \coloneqq B_J^q(x,z) + B_J^p(z,x)$ with $q \in \partial J(z)$ and $z \in \operatorname{dom} J$, but even the symmetrized divergence is not generally a true distance as it can happen that $B_J(x,z) = 0$ even if $x \neq z$.

The Bregman divergence satisfies a three-point identity for any $\bar{x} \in \operatorname{dom} J$: We have

$$B_J^p(\bar{x},x) - B_J^p(\bar{x},z) + B_J^q(x,z) = [J(\bar{x}) - J(x) - \langle p, \bar{x} - x\rangle_X] - [J(\widehat{x}) - J(z) - \langle q, \widehat{x} - z\rangle_X] \\ + [J(x) - J(z) - \langle q, x - z\rangle_X],$$

which immediately gives the three-point identity

$$\langle p - q, x - \bar{x}\rangle_X = B_J^p(\bar{x},x) - B_J^q(\bar{x},z) + B_J^q(x,z) \quad (\bar{x},x,z \in X,\ p \in \partial J(z),\ q \in \partial J(x)). \tag{11.9}$$

If X is a Hilbert space, we can take $J(x) = \frac{1}{2}\|x\|^2$ to obtain $B_J^{x-z}(z,x) = \tilde{B}_J(z,x) = \frac{1}{2}\|z-x\|_X^2$. Therefore this three-point identity generalizes the classical three-point identity (9.1) in Hilbert spaces. This could be used to generalize our convergence proofs to Banach spaces to treat methods of the general form

$$0 \in H(x^{k+1}) + \partial_1 B_J^{q^k}(x^{k+1}, x^k),$$

where ∂_1 denotes taking a subdifferential with respect to the first variable. To see how (11.9) applies, observe that

$$\partial_1 B_J^{q^k}(x^{k+1}, x^k) = \partial J(x^{k+1}) - q^k = \{p^{k+1} - q^k \mid q^{k+1} \in \partial J(x^{k+1})\}.$$

This would, however, not provide convergence in norm but with respect to B_J. For a general approach to primal-dual methods based on Bregman divergences, see [Valkonen, 2021a].

Returning to our generic gap functional $\tilde{\mathcal{G}}$ defined in (11.4), we have already observed in the proof of Lemma 11.1 that $-\Xi\widehat{x} \in \partial(\tilde{F} + \tilde{G})(\widehat{x})$. Since due to the skew-adjointness of Ξ we also have $\langle \Xi\widehat{x}, x\rangle_X = \langle \Xi\widehat{x}, x - \widehat{x}\rangle_X$ for a solution $\widehat{x} \in H^{-1}(0)$, this means that

$$\tilde{\mathcal{G}}(x,\widehat{x}) = B_{\tilde{G}+\tilde{F}}^{-\Xi\widehat{x}}(x,\widehat{x}).$$

In other words, the gap based at a solution $\widehat{x} \in H^{-1}(0)$ is also a Bregman divergence. In general, as we have already remarked, it can be zero for $x \neq \widehat{x}$.

11.2 CONVERGENCE OF FUNCTION VALUES

We start with the fundamental algorithms: the proximal point method and explicit splitting. In the following, we write $G_{\min} \coloneqq \min_{x\in X} G(x)$ whenever the minimum exists.

Theorem 11.2 (proximal point method ergodic function value). *Let G be proper, lower semicontinuous, and (strongly) convex with factor $\gamma \geq 0$. Suppose $[\partial G]^{-1}(0) \neq \emptyset$. Pick an arbitrary $x^0 \in X$. Let $\varphi_{k+1} := \varphi_k(1+\gamma\tau_k)$, and $\varphi_0 := 1$. For the iterates $x^{k+1} := \operatorname{prox}_{\tau_k G}(x^k)$ of the proximal point method, define the* ergodic sequence

$$\tilde{x}^N := \frac{1}{\zeta_N}\sum_{k=0}^{N-1}\tau_k\varphi_k x^{k+1} \quad \text{for} \quad \zeta_N := \sum_{k=0}^{N-1}\tau_k\varphi_k \quad (N \geq 1). \tag{11.10}$$

(i) If $\tau_k \equiv \tau > 0$ and G is not strongly convex ($\gamma = 0$), then $G(\tilde{x}^N) \to G_{\min}$ at the rate $O(1/N)$.

(ii) If $\tau_k \equiv \tau > 0$ and G is strongly convex ($\gamma > 0$), then $G(\tilde{x}^N) \to G_{\min}$ linearly.

(iii) If $\tau_k \nearrow \infty$ and G is strongly convex, then $G(\tilde{x}^N) \to G_{\min}$ superlinearly.

Proof. Let the root $\widehat{x} \in [\partial G]^{-1}(0)$ be arbitrary; by assumption at least one exists. Then $G_{\min} = G(\widehat{x})$ by Theorem 4.2. We recall that the proximal point iteration for minimizing G can be written as

$$0 \in \tau_k \partial G(x^{k+1}) + (x^{k+1} - x^k). \tag{11.11}$$

As in the proof of Theorem 9.4, we test (11.11) by the application of $\varphi_k\langle \cdot, x^{k+1} - \widehat{x}\rangle_X$ for some testing parameter $\varphi_k > 0$ to obtain

$$0 \in \varphi_k\tau_k\langle \partial G(x^{k+1}), x^{k+1} - \widehat{x}\rangle_X + \varphi_k\langle x^{k+1} - x^k, x^{k+1} - \widehat{x}\rangle_X. \tag{11.12}$$

The next step will differ from the proof of Theorem 9.4, as we want a value estimate. Indeed, by the subdifferential characterization of strong convexity, Lemma 7.4 (ii),

$$\langle \partial G(x^{k+1}), x^{k+1} - \widehat{x}\rangle_X \geq G(x^{k+1}) - G(\widehat{x}) + \frac{\gamma}{2}\|x^{k+1} - \widehat{x}\|_X^2.$$

Using this and the three-point-identity (9.1) in (11.12), we obtain similarly to the proof of Theorem 10.1 the estimate

$$\frac{\varphi_k(1+\tau_k\gamma)}{2}\|x^{k+1} - \widehat{x}\|_X^2 + \varphi_k\tau_k[G(x^{k+1}) - G(\widehat{x})] + \frac{\varphi_k}{2}\|x^{k+1} - x^k\|_X^2 \leq \frac{\varphi_k}{2}\|x^k - \widehat{x}\|_X^2. \tag{11.13}$$

We now impose the recursion

$$\varphi_k(1+\tau_k\gamma) = \varphi_{k+1}. \tag{11.14}$$

(Observe the factor-of-two difference compared to (10.5).) Thus

$$\frac{\varphi_{k+1}}{2}\|x^{k+1} - \widehat{x}\|_X^2 + \varphi_k\tau_k[G(x^{k+1}) - G(\widehat{x})] + \frac{\varphi_k}{2}\|x^{k+1} - x^k\|_X^2 \leq \frac{\varphi_k}{2}\|x^k - \widehat{x}\|_X^2. \tag{11.15}$$

Summing over $k = 0, \dots, N-1$ then yields

$$\frac{\varphi_N}{2}\|x^N - \widehat{x}\|_X^2 + \sum_{k=0}^{N-1} \varphi_k \tau_k [G(x^{k+1}) - G(\widehat{x})] + \sum_{k=0}^{N-1} \frac{\varphi_k}{2}\|x^{k+1} - x^k\|_X^2 \leq \frac{\varphi_0}{2}\|x^0 - \widehat{x}\|_X^2 =: C_0. \tag{11.16}$$

Using Jensen's inequality, it follows for the ergodic sequence defined in (11.10) that

$$\zeta_N [G(\tilde{x}^N) - G(\widehat{x})] \leq C_0.$$

If $\varphi_k \equiv \varphi_0$ and $\gamma = 0$, we therefore have that $\zeta_N = N\varphi_0\tau$ and thus obtain $O(1/N)$ convergence of function values for the ergodic variable $\tilde{x}^N$.

If $\varphi_k \equiv \varphi_0$ and $\gamma > 0$, we deduce from (11.14) that $\zeta_N = \sum_{k=0}^{N-1}(1+\gamma\tau_k)^k \tau_k \varphi_0$. This grows exponentially and hence we obtain the claimed linear convergence.

Finally, if $\tau_k \to \infty$, we would similarly to Theorem 10.1 (ii) obtain superlinear convergence if $\zeta_N/\zeta_{N+1} \to 0$ were to hold. To show this, we can write

$$\frac{\zeta_N}{\zeta_{N+1}} = \frac{\sum_{k=0}^{N-1} \varphi_k \tau_k}{\sum_{k=0}^{N} \varphi_k \tau_k} = \frac{\sum_{k=0}^{N-1} \frac{\varphi_k \tau_k}{\varphi_N \tau_N}}{1 + \sum_{k=0}^{N-1} \frac{\varphi_k \tau_k}{\varphi_N \tau_N}}$$

So it suffices to show that $c_N := \sum_{k=k_0}^{N-1} \frac{\varphi_k \tau_k}{\varphi_N \tau_N} \to 0$ as $N \to \infty$. This we obtain by estimating

$$\begin{aligned} c_N &= \sum_{k=0}^{N-1} \frac{\tau_k/\tau_N}{\prod_{j=k}^{N-1}(1+\gamma\tau_j)} \leq \sum_{k=0}^{N-1} \frac{(1+\gamma\tau_k)/(1+\gamma\tau_N)}{\prod_{j=k}^{N-1}(1+\gamma\tau_j)} \\ &= \sum_{k=0}^{N-1} \frac{1}{\prod_{j=k+1}^{N}(1+\gamma\tau_j)} \leq \sum_{k=0}^{N-1} (1+\gamma\tau_{k+1})^{-(N-k)}. \end{aligned}$$

In the first and last step we have used that $\{\tau_k\}_{k\in\mathbb{N}}$ is increasing. Now we pick $a > 1$ and find $k_0 \in \mathbb{N}$ such that $1 + \gamma\tau_k \geq a$ for $k \geq k_0$. Then for $N > k_0$,

$$c_N \leq \sum_{k=0}^{k_0-1} (1+\gamma\tau_{k+1})^{-(N-k)} + \sum_{k=k_0}^{N-1} a^{-(N-k)} = \sum_{k=0}^{k_0-1} (1+\gamma\tau_{k+1})^{-(N-k)} + \sum_{j=1}^{N-k_0} a^{-j}.$$

The first term goes to zero as $N \to \infty$ while the second term, as a geometric series, converges to $a^{-1}/(1-a^{-1})$. We therefore deduce that $\lim_{N\to\infty} c_N \leq a^{-1}/(1-a^{-1})$. Letting $a \to \infty$, we see that $c_N \searrow 0$. □

It is possible to improve the result to be nonergodic by showing that the proximal point method is in fact monotonic.

Corollary 11.3 (proximal point method function value). *The proximal point method is monotonic, i.e., $G(x^{k+1}) \leq G(x^k)$ for all $k \in \mathbb{N}$. Therefore the convergence rates of Theorem 11.2 also hold for $G(x^N) \to G_{\min}$.*

Proof. We know from (11.11) that

$$0 \le \tau_k^{-1}\|x^{k+1} - x^k\|_X^2 = \langle \partial G(x^{k+1}), x^k - x^{k+1}\rangle_X \le G(x^k) - G(x^{k+1}).$$

This proves monotonicity. Now (11.16) gives

$$\zeta_N[G(x^N) - G(\widehat{x})] \le C_0.$$

Now we proceed using the growth estimates for ζ_N in the proof of Theorem 11.2. □

These results can be extended to the explicit splitting method,

$$x^{k+1} := \operatorname{prox}_{\tau G}(x^k - \tau \nabla F(x^k)),$$

in a straightforward manner. In the next theorem, observe in comparison to Theorem 10.2 that $\tau L \le 1$ instead of $\tau L \le 2$. This kind of factor-of-two stricter step length or Lipschitz factor bound is a general feature of function value estimates of methods involving an explicit step, as well as of the gap estimates in the following sections. It stems from the corresponding difference between the value estimate (7.8) and the non-value estimate (7.9) in Corollary 7.2.

Theorem 11.4 (explicit splitting function value). *Let $J := F + G$ where $G : X \to \overline{\mathbb{R}}$ and $F : X \to \mathbb{R}$ are convex, proper, and lower semicontinuous, with F moreover L-smooth. Suppose $[\partial J]^{-1}(0) \neq \emptyset$. If $\tau L \le 1$, the explicit splitting method satisfies both $J(\tilde{x}^N) \to J_{\min}$ at the rate $O(1/N)$. If G is strongly convex, then this convergence is linear.*

Proof. With $\tau_k := \tau$, as usual, we write the method as

$$0 \in \tau_k[\partial G(x^{k+1}) + \nabla F(x^k)] + (x^{k+1} - x^k). \tag{11.17}$$

We then take arbitrary $\widehat{x} \in [\partial(F+G)]^{-1}(0)$ and use the three-point smoothness of F proved in Corollary 7.2, and the subdifferential characterization of strong convexity of G, Lemma 7.4 (ii), to obtain

$$\langle \partial G(x^{k+1}) + \nabla F(x^k), x^{k+1} - \widehat{x}\rangle_X \ge J(x^{k+1}) - J(\widehat{x}) + \frac{\gamma}{2}\|x^{k+1} - \widehat{x}\|_X^2 - \frac{L}{4}\|x^{k+1} - x^k\|_X^2.$$

As in the proof of Theorem 11.2, after testing (11.17) by the application of $\varphi_k\langle \cdot, x^{k+1} - \widehat{x}\rangle_X$, we now obtain

$$\frac{\varphi_{k+1}}{2}\|x^{k+1} - \widehat{x}\|_X^2 + \varphi_k\tau_k[J(x^{k+1}) - J(\widehat{x})] + \frac{\varphi_k(1 - \tau_k L)}{2}\|x^{k+1} - x^k\|_X^2 \le \frac{\varphi_k}{2}\|x^k - \widehat{x}\|_X^2. \tag{11.18}$$

Since $\tau_k L \le 1$, we may proceed as in Theorem 11.2 to prove the ergodic convergences. □

Again, we can show nonergodic convergence due to the monotonicity of the iteration.

Corollary 11.5. *The convergence rates of Theorem 11.4 also hold for* $J(x^N) \to J_{\min}$.

Proof. We obtain from (11.17) and the smoothness of F (see (7.5)) that

$$\tau_k^{-1}\|x^{k+1} - x^k\|_X^2 = \langle \partial G(x^{k+1}) + \nabla F(x^k), x^k - x^{k+1}\rangle_X \le J(x^k) - J(x^{k+1}) + \frac{L}{2}\|x^{k+1} - x^k\|_X^2.$$

Since $L\tau_k \le 1 < 2$, we obtain monotonicity. The rest now follows as in Theorem 11.2 and Corollary 11.3. □

Remark 11.6. Based on Corollary 7.7, any strong convexity of F can also be used to obtain linear convergence by adapting the steps of the proof of Theorem 11.4.

11.3 ERGODIC GAP ESTIMATES

We now study the convergence of gap functionals for general unaccelerated schemes of the form (11.3). Since $\tilde{G}$ may in general not have the same factor L of smoothness on all subspaces, we introduce the condition (11.19) of the next result. It is simply a version of the standard result of Corollary 7.2 that allows a block-separable structure through the operator Λ in place of the factor L; compare Example 10.4.

Theorem 11.7. *Let* $H := \partial\tilde{F} + \nabla\tilde{G} + \Xi$, *where* $\Xi \in \mathbb{L}(X;X)$ *is skew-adjoint and* $\tilde{G} : X \to \overline{\mathbb{R}}$ *and* $\tilde{F} : X \to \mathbb{R}$ *are convex, proper, and lower semicontinuous. Suppose* $\tilde{F}$ *satisfies for some* $\Lambda \in \mathbb{L}(X;X)$ *the three-point smoothness condition*

$$\langle \nabla\tilde{F}(z), x - \bar{x}\rangle_X \ge F(x) - F(\bar{x}) - \frac{1}{2}\|z - x\|_\Lambda^2 \quad (\bar{x}, x, z \in X). \tag{11.19}$$

Also let $M \in \mathbb{L}(X;X)$ *be positive semi-definite and self-adjoint. Pick* $x^0 \in X$, *and let the sequence* $\{x^{k+1}\}_{k\in\mathbb{N}}$ *be generated through the iterative solution of* (11.3). *Then for every* $\bar{x} \in X$,

$$\frac{1}{2}\|x^N - \bar{x}\|_{ZM}^2 + \sum_{k=0}^{N-1}\left(\tilde{\mathcal{G}}(x^{k+1};\bar{x}) + \frac{1}{2}\|x^{k+1} - \bar{x}\|_{M-\Lambda}^2\right) \le \frac{1}{2}\|x^1 - \bar{x}\|_{ZM}^2. \tag{11.20}$$

Proof. Observe that (11.19) implies

$$\langle \nabla\tilde{F}(z), x - \bar{x}\rangle \ge \tilde{F}(x) - \tilde{F}(\bar{x}) - \frac{1}{2}\|z - x\|_\Lambda^2 \quad (x, z \in X). \tag{11.21}$$

Likewise, by the convexity of $\tilde{G}$ we have

$$\langle \partial\tilde{G}(x), x - \bar{x}\rangle \ge \tilde{G}(x) - \tilde{G}(\bar{x}) \quad (x \in X). \tag{11.22}$$

Using (11.21) and (11.22), we obtain

$$\begin{aligned}(11.23)\qquad &\langle \partial\tilde{G}(x^{k+1}) + \nabla\tilde{F}(x^k) + \Xi x^{k+1}, x^{k+1} - \bar{x}\rangle \\ &\qquad \geq (\tilde{G}+\tilde{F})(x^{k+1}) - (\tilde{G}+\tilde{F})(\bar{x}) + \langle \Xi x^{k+1}, x^{k+1} - \bar{x}\rangle_X - \frac{1}{2}\|z - x\|_\Lambda^2 \\ &\qquad = \tilde{\mathcal{G}}(x^{k+1};\bar{x}) - \frac{1}{2}\|z - x\|_\Lambda^2.\end{aligned}$$

In the final step we have also referred to the definition of $\tilde{\mathcal{G}}$ in (11.4) and the skew-adjointness of Ξ.

From here on, our arguments are already standard: We test (11.3) through the application of $\langle \bullet, x^{k+1} - \bar{x}\rangle$, obtaining

$$0 \in \langle \partial\tilde{G}(x^{k+1}) + \nabla\tilde{F}(x^k) + \Xi x^{k+1} + M(x^{k+1} - x^k), x^{k+1} - \bar{x}\rangle.$$

Then we insert (11.23), which gives

$$\frac{1}{2}\|x^{k+1} - \bar{x}\|_M^2 + \tilde{\mathcal{G}}(x^{k+1};\bar{x}) + \frac{1}{2}\|x^{k+1} - x^k\|_{M-\Lambda}^2 \leq \frac{1}{2}\|x^k - \bar{x}\|_M^2.$$

Summing over $k = 0, \dots, N-1$ yields (11.20). □

In particular, we obtain the following corollary that shows that $\tilde{\mathcal{G}}(\tilde{x}^N;\widehat{x}) \to \tilde{\mathcal{G}}(\widehat{x};\widehat{x}) = 0$ at the rate $O(1/N)$ for any $\widehat{x} \in H^{-1}(0)$. Even further, taking any *bounded* set $B \subset X$ such that $H^{-1}(0) \cap B \neq \emptyset$, we see that also the *partial gap* $\tilde{\mathcal{G}}(\tilde{x}^N;B) \to \tilde{\mathcal{G}}(\widehat{x};B) = 0$.

Corollary 11.8. *In Theorem 11.7, suppose in addition that $M \geq \Lambda$ and define the ergodic sequence*

$$\tilde{x}^N := \frac{1}{N}\sum_{k=0}^{N-1} x^{k+1}.$$

Then

$$\tilde{\mathcal{G}}(\tilde{x}^N;\bar{x}) \leq \frac{1}{2N}\|x^1 - \bar{x}\|_M^2.$$

Proof. This follows immediately from using $M \geq \Lambda$ to eliminate the term $\frac{1}{2}\|x^{k+1} - \bar{x}\|_{M-\Lambda}^2$ from (11.20) and then using Jensen's inequality on the gap. □

Due to the presence of Ξ, we cannot in general prove monotonicity of the abstract proximal point method and thus get rid of the ergodicity of the estimates.

IMPLICIT SPLITTING

We now consider the solution of

$$\min_{x\in X} F(x) + G(x).$$

Setting $B = \partial F$ and $A = \partial G$, (9.16), the Douglas–Rachford or implicit splitting method can be written in the general form (11.3) with $u = (x, y, z)$,

$$\tilde{G}(u) := \tau G(y) + \tau F(x), \qquad \tilde{F} \equiv 0,$$
$$\Xi := \begin{pmatrix} 0 & \mathrm{Id} & -\mathrm{Id} \\ -\mathrm{Id} & 0 & \mathrm{Id} \\ \mathrm{Id} & -\mathrm{Id} & 0 \end{pmatrix}, \quad \text{and} \qquad M := \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I \end{pmatrix}.$$

Moreover,

$$H(u) := \partial \tilde{G}(u) + \Xi u. \tag{11.24}$$

We then have the following ergodic estimate for

$$\mathcal{G}_{\mathrm{DRS}}(u;\widehat{u}) = [G(y) + F(x)] - [G(\widehat{x}) + F(\widehat{x})] + \langle \widehat{x} - \widehat{z}, x - y\rangle \geq 0.$$

Theorem 11.9. *Let $F : X \to \overline{\mathbb{R}}$ and $G : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. Let $\widehat{u} \in H^{-1}(0)$ for H given by (11.24). Then for any initial iterate $u^0 = (x^0, y^0, z^0) \in X^3$, the iterates $\{u^k\}_{k\in N}$ of the implicit splitting method (8.8) satisfy*

$$\mathcal{G}_{DRS}(\tilde{u}^N;\widehat{u}) \leq \frac{1}{2N\tau}\|u^1 - \widehat{u}\|_M^2, \quad \textit{where} \quad \tilde{u}^N := \frac{1}{N}\sum_{k=0}^{N-1} u^{k+1}.$$

Proof. Clearly M is self-adjoint and positive semi-definite, and $M \geq \Lambda := 0$. The rest is clear from Corollary 11.8 by moving τ from $\tilde{\mathcal{G}}$ on the right-hand side, and using that $\widehat{x} = \widehat{y}$. □

Clearly, following the discussion in Section 11.1, we can define a partial version of $\mathcal{G}_{\mathrm{DRS}}$ and obtain its convergence from Theorem 11.9.

PRIMAL-DUAL EXPLICIT SPLITTING

We recall that the PDES method (8.23) for (11.5) corresponds to (11.5) with the choice $F_0 = 0$ and $E = F$, while the preconditioning operator is given by

$$M := \begin{pmatrix} \mathrm{Id} & 0 \\ 0 & \mathrm{Id} - KK^* \end{pmatrix}$$

With this, we obtain the following estimate for the Lagrangian duality gap defined in (11.7).

Theorem 11.10. *Let $F : X \to \mathbb{R}$ and $G : Y \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $K \in \mathbb{L}(X;Y)$. Suppose F is be Gâteaux differentiable with L-Lipschitz gradient for $L \leq 1$, and that $\|K\|_{\mathbb{L}(X;Y)} \leq 1$. Then for any initial iterate $u^0 \in X \times Y$, the iterates $\{u^k = (x^k, y^k)\}_{k\in\mathbb{N}}$ of (8.23) satisfy for all $\bar{u} = (\bar{x}, \bar{y}) \in X \times Y$ the ergodic gap estimate*

$$\mathcal{G}(\tilde{u}^N; \bar{u}) \leq \frac{1}{2N}\|u^1 - \bar{u}\|_M^2, \quad \text{where} \quad \tilde{u}^N := \frac{1}{N}\sum_{k=0}^{N-1} u^{k+1}.$$

In particular, if $B \subset X$ is bounded and $B \cap H^{-1}(0) \neq \emptyset$, the partial duality gap $\mathcal{G}(u^N, B) \to 0$ at the rate $O(1/N)$.

Proof. We use Corollary 11.8. Using the assumed bound $\|K\|_{\mathbb{L}(X;Y)} \leq 1$, clearly M is self-adjoint and positive semi-definite. By Corollary 7.2, the three-point smoothness condition (11.19) holds with $\Lambda := \left(\begin{smallmatrix} L & 0 \\ 0 & 0 \end{smallmatrix}\right)$, where L is the Lipschitz factor of ∇F. Since $\|K\|_{\mathbb{L}(X;Y)} \leq 1$ and $L \leq 1$, we also verify $M \geq \Lambda$. The rest now follows from Corollary 11.8 as well as the nonnegativity of the partial duality gap (11.8). □

PRIMAL-DUAL PROXIMAL SPLITTING

We continue with the problem (11.5) and the corresponding structure (11.6) for H. We recall from Corollaries 9.14 and 9.20 that for the unaccelerated PDPS we take the preconditioning operator as

$$M := \begin{pmatrix} \tau^{-1}\mathrm{Id} & -K^* \\ -K & \sigma^{-1}\mathrm{Id} \end{pmatrix} \tag{11.25}$$

for some primal and dual step length parameters $\tau, \sigma > 0$. We now obtain the following result for the Lagrangian duality gap defined in (11.7).

Theorem 11.11. *Let $F_0 : X \to \overline{\mathbb{R}}$, $E : X \to \mathbb{R}$, and $G : Y \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $K \in \mathbb{L}(X;Y)$. Suppose E is Gâteaux differentiable with L-Lipschitz gradient. Take $\sigma, \tau > 0$ satisfying*

$$L\tau + \tau\sigma\|K\|^2 < 1.$$

Then for any initial iterate $u^0 \in X \times Y$ the iterates $\{u^k = (x^k, y^k)\}_{k\in\mathbb{N}}$ of the PDPS method (9.29) satisfy for any $\bar{u} = (\bar{x}, \bar{y}) \in X \times Y$ the ergodic gap estimate

$$\mathcal{G}(\tilde{u}^N; \bar{u}) \leq \frac{1}{2N\tau}\|u^1 - \bar{u}\|_M^2, \quad \text{where} \quad \tilde{u}^N := \frac{1}{N}\sum_{k=0}^{N-1} u^{k+1}.$$

In particular, if $B \subset X$ is bounded and $B \cap H^{-1}(0) \neq \emptyset$, the partial duality gap $\mathcal{G}(u^N, B) \to 0$ at the rate $O(1/N)$.

Proof. We use Corollary 11.8. By Corollary 7.2, the three-point smoothness condition (11.19) holds with $\Lambda := \left(\begin{smallmatrix} L & 0 \\ 0 & 0 \end{smallmatrix}\right)$, where L is the Lipschitz factor of ∇E. In Corollary 9.20 we have already proved that ZM is self-adjoint and positive semi-definite. Similarly to the proof of the corollary, we verify that the condition $L\tau + \tau\sigma\|K\|^2 < 1$ guarantees $M \geq \Lambda$. (The only difference to the conditions in that result is the standard gap estimate factor-of-two difference in the term containing L.) The rest is clear from Corollary 11.8 as well as the nonnegativity of the partial duality gap (11.8). □

11.4 THE TESTING APPROACH IN ITS GENERAL FORM

We now want to produce gap estimates for accelerated methods. As we have seen in Section 10.1, as an extension of (11.3) these iteratively solve

$$0 \in W_{k+1}[\partial\tilde{G}(x^{k+1}) + \nabla\tilde{F}(x^k) + \Xi x^{k+1}] + M_{k+1}(x^{k+1} - x^k) \tag{11.26}$$

for iteration-dependent step length and preconditioning operators $W_{k+1} \in \mathbb{L}(X;X)$ and $M_{k+1} \in \mathbb{L}(X;X)$. We also introduced testing operators $Z_{k+1} \in \mathbb{L}(X;X)$ such that $Z_{k+1}M_{k+1}$ is self-adjoint and positive semi-definite.

Unless $Z_{k+1}W_{k+1}$ is a scalar multiple of the identity, we will not be able to extract in a straightforward way any of the gap functionals of Section 11.1 out of (11.26). Indeed, it is not clear how to provide a completely general approach to gap functionals of accelerated or otherwise complex algorithms. We will specifically see the difficulties when performing gap realignment for the accelerated PDPS in Section 11.5 and when developing very specific gap functionals for the ADMM in Section 11.6.

Towards brevity in the following sections, we however do some general preparatory work. Observe that the method (11.26) can be written more abstractly as

$$0 \in \widetilde{H}_{k+1}(x^{k+1}) + M_{k+1}(x^{k+1} - x^k) \tag{11.27}$$

for some iteration-dependent set-valued function $\widetilde{H}_{k+1} : X \rightrightarrows X$. The estimate (11.28) in the next theorem is in essence a quantitative or variable-metric version of the three-point smoothness and strong convexity estimate (7.16). The proof of the following result is already standard, where the abstract value $\mathcal{V}_{k+1}(\widehat{x})$ models a suitable gap functional for iterate x^{k+1}.

Theorem 11.12. *On a Hilbert space X, let $\widetilde{H}_{k+1} : X \rightrightarrows X$, and $M_{k+1}, Z_{k+1} \in \mathbb{L}(X;X)$ for $k \in \mathbb{N}$. Suppose (11.27) is solvable for the iterates $\{x^k\}_{k\in\mathbb{N}}$. If $Z_{k+1}M_{k+1}$ is self-adjoint and*

$$\begin{aligned}\langle \widetilde{H}_{k+1}(x^{k+1}), x^{k+1} - \widehat{x}\rangle_{Z_{k+1}} \geq \mathcal{V}_{k+1}(\widehat{x}) &+ \frac{1}{2}\|x^{k+1} - \widehat{x}\|^2_{Z_{k+2}M_{k+2} - Z_{k+1}M_{k+1}} \\ &- \frac{1}{2}\|x^{k+1} - x^k\|^2_{Z_{k+1}M_{k+1}}\end{aligned} \tag{11.28}$$

for all $k \in \mathbb{N}$ *and some* $\widehat{x} \in X$ *and* $\mathcal{V}_{k+1}(\widehat{x}) \in \mathbb{R}$, *then both*

$$\frac{1}{2}\|x^{k+1} - \widehat{x}\|^2_{Z_{k+2}M_{k+2}} + \mathcal{V}_{k+1}(\widehat{x}) \le \frac{1}{2}\|x^k - \widehat{x}\|^2_{Z_{k+1}M_{k+1}} \quad (k \in \mathbb{N}) \tag{11.29}$$

and

$$\frac{1}{2}\|x^N - \widehat{x}\|^2_{Z_{N+1}M_{N+1}} + \sum_{k=0}^{N-1} \mathcal{V}_{k+1}(\widehat{x}) \le \frac{1}{2}\|x^0 - \widehat{x}\|^2_{Z_1M_1} \quad (N \ge 1). \tag{11.30}$$

Proof. Inserting (11.27) into (11.28), we obtain

$$\begin{aligned} -\langle x^{k+1} - x^k, x^{k+1} - \widehat{x}\rangle_{Z_{k+1}M_{k+1}} \ge \frac{1}{2}\|x^{k+1} - \widehat{x}\|^2_{Z_{k+2}M_{k+2} - Z_{k+1}M_{k+1}} \\ - \frac{1}{2}\|x^{k+1} - x^k\|^2_{Z_{k+1}M_{k+1}} + \mathcal{V}_{k+1}(\widehat{x}). \end{aligned} \tag{11.31}$$

We recall for general self-adjoint M the three-point formula (9.1), i.e.,

$$\langle x^{k+1} - x^k, x^{k+1} - \widehat{x}\rangle_M = \frac{1}{2}\|x^{k+1} - x^k\|^2_M - \frac{1}{2}\|x^k - \widehat{x}\|^2_M + \frac{1}{2}\|x^{k+1} - \widehat{x}\|^2_M.$$

Using this with $M = Z_{k+1}M_{k+1}$, we rewrite (11.31) as (11.29). Summing (11.29) over $k = 0, \dots, N-1$, we obtain (11.30). □

11.5 ERGODIC GAPS FOR ACCELERATED PRIMAL-DUAL METHODS

To derive ergodic gap estimates for the accelerated primal-dual proximal splitting of Theorem 10.8, we need to perform significant additional work due to the fact that $\eta_k := \varphi_k\tau_k \ne \psi_{k+1}\sigma_{k+1}$. The overall idea of the proof remains the same, but we need to pay special attention to the blockwise structure of the problem and to do some realignment of the blocks to get the same factor η_k in front of both G and F.

DUALITY GAP REALIGNMENT

We continue with the problem (11.5) and the setup (11.6). Working with the general scheme (11.27), we write

$$\widetilde{H}_{k+1}(u) := W_{k+1}(\partial\tilde{G}(u^{k+1}) + \nabla\tilde{F}(u^k) + \Xi) \tag{11.32a}$$

taking as in Theorem 10.8 the testing and step length operators

$$W_{k+1} := \begin{pmatrix} \tau_k\mathrm{Id} & 0 \\ 0 & \sigma_{k+1}\mathrm{Id} \end{pmatrix} \quad \text{and} \quad Z_{k+1} := \begin{pmatrix} \varphi_k\mathrm{Id} & 0 \\ 0 & \psi_{k+1}\mathrm{Id} \end{pmatrix} \tag{11.32b}$$

for some step length and testing parameters $\tau_k, \sigma_{k+1}, \varphi_k, \sigma_{k+1} > 0$. Throughout this section we also take

$$\Gamma := \begin{pmatrix} \gamma \cdot \mathrm{Id} & 0 \\ 0 & \rho \cdot \mathrm{Id} \end{pmatrix} \quad \text{and} \quad \Lambda := \begin{pmatrix} L \cdot \mathrm{Id} & 0 \\ 0 & 0 \end{pmatrix}. \tag{11.32c}$$

For the moment, we do not yet need to know the specific structure of M_{k+1}; hence the following estimates apply not only to the PDPS method but also to the PDES method and its potential accelerated variants.

Lemma 11.13. *Let us be given $K \in \mathbb{L}(X;Y)$, $F = F_0 + E$ with $F_0 : X \to \overline{\mathbb{R}}$, $E : X \to \mathbb{R}$, and $G^* : Y \to \overline{\mathbb{R}}$ convex, proper, and lower semicontinuous on Hilbert spaces X and Y. Suppose F_0 and G^* are (strongly) convex for some $\gamma, \rho \geq 0$, and E has L-Lipschitz continuous gradient. With the setup of (11.6) and (11.32), for any $u, \widehat{u} \in X \times Y$ and any $k \in \mathbb{N}$ we have*

$$\langle \widetilde{H}_{k+1}(u), u - \widehat{u} \rangle_{Z_{k+1}} \geq \mathcal{G}_{k+1}(u; \widehat{u}) + \frac{1}{2}\|u - \widehat{u}\|^2_{Z_{k+1}W_{k+1}(2\Xi+\Gamma)} - \frac{1}{4}\|u - u^k\|^2_{Z_{k+1}W_{k+1}\Lambda}$$

for

$$\begin{aligned} \mathcal{G}_{k+1}(u;\widehat{u}) := {} & \varphi_k\tau_k(F(x) - F(\widehat{x})) + \psi_{k+1}\sigma_{k+1}(G^*(y) - G^*(\widehat{y})) \\ & + \langle (\varphi_k\tau_k K^*)\widehat{y}, x\rangle_X - \langle (\psi_{k+1}\sigma_{k+1}K)\widehat{x}, y\rangle_Y - \langle (K\varphi_k\tau_k - \psi_{k+1}\sigma_{k+1}K)\widehat{x}, \widehat{y}\rangle_Y. \end{aligned}$$

Proof. Expanding $\widetilde{H}_{k+1}$, we have

$$\begin{aligned} \langle \widetilde{H}_{k+1}(u), u - \widehat{u} \rangle_{Z_{k+1}} = {} & \varphi_k\tau_k\langle \partial F_0(x), x - \widehat{x}\rangle_X \\ & + \varphi_k\tau_k\langle \nabla E(x^k), x - \widehat{x}\rangle_X \\ & + \psi_{k+1}\sigma_{k+1}\langle \partial G^*(y), y - \widehat{y}\rangle_Y \\ & + \langle (\varphi_k\tau_k K^*)y, x - \widehat{x}\rangle_X - \langle (\psi_{k+1}\sigma_{k+1}K)x, y - \widehat{y}\rangle_Y. \end{aligned}$$

Observe that

$$\begin{aligned} & \langle (\varphi_k\tau_k K^*)y, x - \widehat{x}\rangle_X - \langle (\psi_{k+1}\sigma_{k+1}K)x, y - \widehat{y}\rangle_Y \\ & \quad = \langle (K\varphi_k\tau_k - \psi_{k+1}\sigma_{k+1}K)(x - \widehat{x}), y - \widehat{y}\rangle_Y \\ & \qquad + \langle (\varphi_k\tau_k K^*)\widehat{y}, x - \widehat{x}\rangle_X - \langle (\psi_{k+1}\sigma_{k+1}K)\widehat{x}, y - \widehat{y}\rangle_Y \\ & \quad = \frac{1}{2}\|u - \widehat{u}\|^2_{2Z_{k+1}W_{k+1}\Xi} - \langle (K\varphi_k\tau_k - \psi_{k+1}\sigma_{k+1}K)\widehat{x}, \widehat{y}\rangle_Y \\ & \qquad + \langle (\varphi_k\tau_k K^*)\widehat{y}, x\rangle_X - \langle (\psi_{k+1}\sigma_{k+1}K)\widehat{x}, y\rangle_Y. \end{aligned}$$

Therefore

$$\begin{aligned} \langle \widetilde{H}_{k+1}(u), u - \widehat{u} \rangle_{Z_{k+1}} = {} & \varphi_k\tau_k\langle \partial F_0(x), x - \widehat{x}\rangle_X \\ & + \varphi_k\tau_k\langle \nabla E(x^k), x - \widehat{x}\rangle_X \\ & + \psi_{k+1}\sigma_{k+1}\langle \partial G^*(y), y - \widehat{y}\rangle_Y \\ & + \frac{1}{2}\|u - \widehat{u}\|^2_{2Z_{k+1}W_{k+1}\Xi} - \langle (K\varphi_k\tau_k - \psi_{k+1}\sigma_{k+1}K)\widehat{x}, \widehat{y}\rangle_Y \\ & + \langle (\varphi_k\tau_k K^*)\widehat{y}, x\rangle_X - \langle (\psi_{k+1}\sigma_{k+1}K)\widehat{x}, y\rangle_Y. \end{aligned} \tag{11.33}$$

Due to the smoothness three-point corollaries, specifically (7.8), we have

$$\langle \nabla E(x^k), x - \widehat{x} \rangle_X \geq E(x) - E(\widehat{x}) - \frac{L}{2}\|x - x^k\|_X^2. \tag{11.34a}$$

Also, by the (strong) convexity of F_0, we have

$$\langle \partial F_0(x), x - \widehat{x} \rangle_X \geq F_0(x) - F_0(\widehat{x}) + \frac{\gamma}{2}\|x - \widehat{x}\|_X^2, \tag{11.34b}$$

as well as by the (strong) convexity of G^*

$$\langle \partial G^*(y), y - \widehat{y} \rangle_Y \geq G^*(y) - G^*(\widehat{y}) + \frac{\rho}{2}\|y - \widehat{y}\|_Y^2. \tag{11.34c}$$

Applying these estimates in (11.33), and using the structure (11.32b) and (11.32c) of the involved operators, we obtain the claim. □

If $\varphi_k \tau_k = \psi_{k+1}\sigma_{k+1}$, clearly $\mathcal{G}_{k+1}(u^{k+1}; \widehat{u}) \geq \varphi_k \tau_k \mathcal{G}(u^{k+1})$. This is the case in the unaccelerated case already considered in Theorems 11.10 and 11.11. Some specific stochastic accelerated algorithms also satisfy this [see Valkonen, 2019]. Applying the techniques of Section 11.3, we could then use Jensen's inequality to estimate $\sum_{k=0}^{n-1} \mathcal{G}_{k+1}(u^{k+1}; \widehat{u}) \geq \sum_{k=0}^{N-1} \varphi_k \tau_k \mathcal{G}(u^{k+1})$ further from below to obtain a gap on suitable ergodic sequences. However, in our primary accelerated algorithm of interest, the PDPS method, instead $\varphi_k \tau_k = \psi_k \sigma_k$. We will therefore have to do some rearrangements.

Lemma 11.14. *Let $K \in \mathbb{L}(X;Y)$, $F = F_0 + E$ with $F_0 : X \to \overline{\mathbb{R}}$, $E : X \to \mathbb{R}$, and $G^* : Y \to \overline{\mathbb{R}}$ convex, proper, and lower semicontinuous on Hilbert spaces X and Y. Suppose F_0 and G^* are (strongly) convex for some $\gamma, \rho \geq 0$, and E has L-Lipschitz gradient. With the setup of (11.6) and (11.32), suppose $\varphi_k \tau_k = \psi_k \sigma_k$. If $\widehat{u} \in H^{-1}(0)$, then*

$$\begin{aligned}\langle \widetilde{H}_{k+1}(u^{k+1}), u^{k+1} - \widehat{u} \rangle_{Z_{k+1}} \geq \mathcal{G}_{*,k+1}(x^{k+1}, y^k; \widehat{u}) &+ \frac{1}{2}\|u^{k+1} - \widehat{u}\|^2_{Z_{k+1}W_{k+1}(2\Xi+\Gamma)} \\ &- \frac{1}{2}\|u^{k+1} - u^k\|^2_{Z_{k+1}W_{k+1}\Lambda} \quad (N \geq 2)\end{aligned} \tag{11.35}$$

for some $\mathcal{G}_{*,k+1}(x^{k+1}, y^k; \widehat{u})$ satisfying with $\mathcal{G}$ given by (11.7) the estimate

$$\sum_{k=0}^{N-1} \mathcal{G}_{*,k+1}(x^{k+1}, y^k; \widehat{u}) \geq \sum_{k=1}^{N-1} \varphi_k \tau_k \mathcal{G}(x^{k+1}, y^k; \widehat{u}). \tag{11.36}$$

Proof. First, note that (11.35) holds for

$$\begin{aligned}\mathcal{G}_{*,k+1}(x^{k+1}, y^k; \widehat{u}) := \inf_{w^{k+1} \in \widetilde{H}_{k+1}(u^{k+1})} &\langle w^{k+1}, u^{k+1} - \widehat{u} \rangle_{Z_{k+1}} \\ &- \frac{1}{2}\|u^{k+1} - \widehat{u}\|^2_{Z_{k+1}W_{k+1}(2\Xi+\Gamma)} + \frac{1}{2}\|u^{k+1} - u^k\|^2_{Z_{k+1}W_{k+1}\Lambda}.\end{aligned}$$

It remains to prove the estimate (11.36) for this choice.

With $N \geq 1$, let us define the set

$$\begin{aligned} S_N &:= \sum_{k=0}^{N-1} \Big(\langle \widetilde{H}_{k+1}(u^{k+1}), u^{k+1} - \widehat{u} \rangle_{Z_{k+1}} - \frac{1}{2} \|u^{k+1} - \widehat{u}\|^2_{Z_{k+1} W_{k+1}(2\Xi + \Gamma)} + \frac{1}{2} \|u^{k+1} - u^k\|^2_{Z_{k+1} W_{k+1} \Lambda} \Big) \\ &= \sum_{k=0}^{N-1} \Bigg(\varphi_k \tau_k \Big(\langle \partial F_0(x^{k+1}) + \nabla E(x^k), x^{k+1} - \widehat{x} \rangle_X - \frac{\gamma}{2} \|x^{k+1} - \widehat{x}\|^2_X + \frac{L}{2} \|x^{k+1} - x^k\|^2_X \Big) \\ &\qquad + \psi_{k+1} \sigma_{k+1} \Big(\langle \partial G^*(y^{k+1}), y^{k+1} - \widehat{y} \rangle_Y - \frac{\rho}{2} \|y^{k+1} - \widehat{y}\|^2_Y \Big) \Bigg). \end{aligned}$$

Observe that in the second expression, $Z_{k+1} W_{k+1} \Xi$ has canceled the corresponding component of $\widetilde{H}_{k+1}$. Then it is enough to prove that $S_N \geq \sum_{k=1}^{N-1} \varphi_k \tau_k \mathcal{G}(x^{k+1}, y^k; \widehat{u})$. To do this, we need to shift y^{k+1} to y^k. With $N \geq 2$, we therefore rearrange terms to obtain

$$S_N = A_N + B_N$$

for

$$\begin{aligned} A_N &= \varphi_0 \tau_0 \Big(\langle \partial F_0(x^1) + \nabla E(x^0), x^1 - \widehat{x} \rangle_X - \frac{\gamma}{2} \|x^1 - \widehat{x}\|^2_X + \frac{L}{2} \|x^1 - x^0\|^2_X \Big) \\ &\quad + \psi_N \sigma_N \Big(\langle \partial G^*(y^N), y^N - \widehat{y} \rangle_Y - \frac{\rho}{2} \|y^N - \widehat{y}\|^2_Y \Big) \\ &\quad - \langle (K \varphi_0 \tau_0 - \psi_N \sigma_N K) \widehat{x}, \widehat{y} \rangle_Y + \langle (\varphi_0 \tau_0 K^*) \widehat{y}, x^1 \rangle_X - \langle (\psi_N \sigma_N K) \widehat{x}, y^N \rangle_Y \end{aligned}$$

and

$$\begin{aligned} B_N &:= \sum_{k=1}^{N-1} \Bigg(\varphi_k \tau_k \Big(\langle \partial F_0(x^{k+1}) + \nabla E(x^k), x^{k+1} - \widehat{x} \rangle_X - \frac{\gamma}{2} \|x^{k+1} - \widehat{x}\|^2_X + \frac{L}{2} \|x^{k+1} - x^k\|^2_X \Big) \\ &\qquad + \psi_k \sigma_k \Big(\langle \partial G^*(y^k), y^k - \widehat{y} \rangle_Y - \frac{\rho}{2} \|y^{k+1} - \widehat{y}\|^2_Y \Big) \\ &\qquad + \langle (\varphi_k \tau_k K^*) \widehat{y}, x^{k+1} \rangle_X - \langle (\psi_k \sigma_k K) \widehat{x}, y^k \rangle_Y \Bigg) \end{aligned}$$

Observe that we only sum over $k = 1, \ldots, N-1$ instead of $k = 0, \ldots, N-1$.

We can now use (11.34) and our assumption $\varphi_k \tau_k = \psi_k \sigma_k$ to estimate

$$B_N \geq \sum_{k=1}^{N-1} \varphi_k \tau_k \mathcal{G}(x^{k+1}, y^k). \tag{11.37}$$

By Corollary 7.2, E satisfies the three-point monotonicity estimate (7.9); in particular,

$$\langle \nabla E(x^0) - \nabla E(\widehat{x}), x^1 - \widehat{x} \rangle_X \geq -\frac{L}{2} \|x^1 - x^0\|^2_X.$$

Since $K^*\widehat{x} \in \partial G^*(\widehat{y})$, and $-K\widehat{y} \in \partial F_0(\widehat{x}) + \nabla E(\widehat{x})$, and ∂F_0 and ∂G are strongly monotone, we also obtain

$$\langle \partial F_0(x^1) + \nabla E(\widehat{x}) + K^*\widehat{y}, x^1 - \widehat{x}\rangle_X - \frac{\gamma}{2}\|x^1 - \widehat{x}\|_X^2 \geq 0 \quad \text{and}$$
$$\langle \partial G^*(y^N) - K\widehat{x}, y^N - \widehat{y}\rangle_Y - \frac{\rho}{2}\|y^N - \widehat{y}\|_Y^2 \geq 0.$$

Rearranging and using these estimates we obtain

$$\begin{aligned} A_N = \varphi_0\tau_0\Big(\langle \partial F_0(x^1) + \nabla E(x^0) + K^*\widehat{y}, x^1 - \widehat{x}\rangle_X - \frac{\gamma}{2}\|x^1 - \widehat{x}\|_X^2 + \frac{L}{2}\|x^1 - x^0\|_X^2\Big) \\ + \psi_N\sigma_N\Big(\langle \partial G^*(y^N) - K\widehat{x}, y^N - \widehat{y}\rangle_Y - \frac{\gamma}{2}\|y^N - \widehat{y}\|_Y^2\Big) \geq 0. \end{aligned} \tag{11.38}$$

The estimates (11.37) and (11.38) finally give $S_N \geq \sum_{k=1}^{N-1}\varphi_k\tau_k\mathcal{G}(x^{k+1}, y^k; \widehat{u})$ as we set out to prove. □

In the proof of Lemma 11.14, we required $\widehat{u} \in H^{-1}(0)$ to show that $A_N \geq 0$. Therefore, as the estimate (11.35) will not hold for an arbitrary base point $\bar{u}$ in place $\widehat{u}$, we will not be able to obtain for accelerated methods the convergence of the *partial duality gap* (11.8) that converges for unaccelerated methods.

The next theorem is our main result regarding ergodic gaps for general accelerated methods. As γ and ρ feature as acceleration parameters in algorithms, the conditions of this theorem imply that gap estimates require slower acceleration.

Theorem 11.15. *Let $K \in \mathbb{L}(X;Y)$, $F = F_0 + E$ with $F_0 : X \to \overline{\mathbb{R}}$, $E : X \to \mathbb{R}$, and $G^* : Y \to \overline{\mathbb{R}}$ convex, proper, and lower semicontinuous on Hilbert spaces X and Y. Suppose F_0 and G^* are (strongly) convex for some $\gamma, \rho \geq 0$, and E has L-Lipschitz gradient. Assume the setup (11.6) and (11.32). For each $k \in \mathbb{N}$, also take $M_{k+1} \in \mathbb{L}(X \times Y; X \times Y)$ such that $Z_{k+1}M_{k+1}$ is self-adjoint. Pick an initial iterate $u^0 \in X \times Y$ and suppose $\{u^{k+1} = (x^{k+1}, y^{k+1})\}_{k\in\mathbb{N}}$ are generated by (11.27). Let $\widehat{u} = (\widehat{x}, \widehat{y}) \in H^{-1}(0)$. If $\varphi_k\tau_k = \psi_k\sigma_k$, and*

$$\frac{1}{2}\|u^{k+1} - u^k\|^2_{Z_{k+1}(M_{k+1} - W_{k+1}\Lambda)} + \frac{1}{2}\|u^{k+1} - \widehat{u}\|^2_{Z_{k+1}(M_{k+1} + W_{k+1}(2\Xi+\Gamma)) - Z_{k+2}M_{k+2}} \geq 0, \tag{11.39}$$

then

$$\frac{1}{2}\|u^N - \widehat{u}\|^2_{Z_{N+1}M_{N+1}} + \zeta_{*,N}\mathcal{G}(\tilde{x}_{*,N}, \tilde{y}_{*,N}; \widehat{u}) \leq \|u^0 - \widehat{u}\|^2_{Z_1M_1} \quad (N \geq 2) \tag{11.40}$$

for $\mathcal{G}$ given by (11.7) and the ergodic sequences

$$\tilde{x}_{*,N} := \zeta_{*,N}^{-1}\sum_{k=1}^{N-1}\tau_k\varphi_k x^{k+1} \quad \text{and} \quad \tilde{y}_{*,N} := \zeta_{*,N}^{-1}\sum_{k=1}^{N-1}\sigma_k\psi_k y^k \quad \text{for} \quad \zeta_{*,N} := \sum_{k=1}^{N-1}\eta_k.$$

Proof. Using (11.35) in (11.39), we obtain (11.28) for $\mathcal{V}_{k+1}(\widehat{u}) := \mathcal{G}_{*,k+1}(x^{k+1}, y^k; \widehat{u})$. By Jensen's inequality,

$$\sum_{k=0}^{N-1} \mathcal{G}_{*,k+1}(x^{k+1}, y^k; \widehat{u}) \geq \zeta_{*,N} \mathcal{G}(\tilde{x}_{*,N}, \tilde{y}_{*,N}; \widehat{u}).$$

We therefore obtain (11.40) from (11.30) in Theorem 11.12. □

ACCELERATED PRIMAL-DUAL PROXIMAL SPLITTING

We now obtain gap estimates for the accelerated PDPS method. Observe the factor-of-two differences in the definitions of ω_k and in the initial conditions for the step lengths in the following theorem compared to Theorem 10.8. Because strong convexity with factor γ implies strong convexity with the factor $\gamma/2$, the conditions and step length rules of this theorem imply the iterate convergence results of Corollary 9.20 and Theorem 10.8 as well.

Theorem 11.16 (gap estimates for PDPS). *Let $F_0 : X \to \overline{\mathbb{R}}$, $E : X \to \mathbb{R}$ and $G : Y \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous on Hilbert spaces X and Y with ∇E L-Lipschitz. Also let $K \in \mathbb{L}(X;Y)$ and let $\widehat{u} = (\widehat{x}, \widehat{y})$ be a primal-dual solution to the problem (11.5). Pick initial step lengths $\tau_0, \sigma_0 > 0$ subject to $L\tau_0 + \tau_0\sigma_0\|K\|^2_{\mathbb{L}(X;Y)} < 1$. For any initial iterate $u^0 \in X \times Y$, suppose $\{u^{k+1}\}_{k\in\mathbb{N}}$ are generated by the (accelerated) PDPS method (10.23). Let the Lagrangian duality gap functional $\mathcal{G}$ be given by (11.7), and the ergodic iterates $\tilde{x}_{*,N}$ and $\tilde{y}_{*,N}$ by (11.15).*

(i) If we take $\tau_k \equiv \tau_0$ and $\sigma_k \equiv \sigma_0$, then the ergodic gap $\mathcal{G}(\tilde{x}_{*,N}, \tilde{y}_{*,N}; \widehat{u}) \to 0$ at the rate $O(1/N)$.

(ii) If F_0 is strongly convex with factor $\gamma > 0$, and we take

$$\omega_k := 1/\sqrt{1+\gamma\tau_k}, \quad \tau_{k+1} := \tau_k\omega_k, \quad \text{and} \quad \sigma_{k+1} := \sigma_k/\omega_k,$$

then $\mathcal{G}(\tilde{x}_{*,N}, \tilde{y}_{*,N}; \widehat{u}) \to 0$ at the rate $O(1/N^2)$

(iii) If both F_0 and G^* are strongly convex with respective factors $\gamma > 0$ and $\rho > 0$, and we take

$$\omega_k := 1/\sqrt{1+\theta}, \quad \theta := \min\{\rho\sigma_0, \gamma\tau_0\}, \quad \tau_k := \tau_0 \quad \text{and} \quad \sigma_k := \sigma_0,$$

then $\mathcal{G}(\tilde{x}_{*,N}, \tilde{y}_{*,N}; \widehat{u}) \to 0$ linearly.

Proof. We use Theorem 11.15 in place of Theorem 10.5 in the proof of Theorem 10.8. We recall that the latter consists of showing $Z_{k+1}M_{k+1}$ to be self-adjoint and (10.17) and $M \geq \Lambda$ to hold, i.e.,

$$Z_{k+1}(M_{k+1} + 2W_{k+1}\Gamma) \geq Z_{k+2}M_{k+2}, \quad \text{and} \quad Z_{k+1}(M_{k+1} - W_{k+1}\Lambda/2) \geq 0,$$

Now, to prove (11.39), we instead prove the self-adjointness as well as

$$Z_{k+1}(M_{k+1} + W_{k+1}\Gamma) \geq Z_{k+2}M_{k+2}, \quad \text{and} \qquad Z_{k+1}(M_{k+1} - W_{k+1}\Lambda) \geq 0.$$

These all follows from the proof of Theorem 10.8 with the factor-of-two differences in the formulas for ω_k and the initialization condition apparent from the statements of these two theorems. The proof of Theorem 10.8 also verifies that $\varphi_k\tau_k = \psi_k\sigma_k$.

All the conditions Theorem 11.15 are therefore satisfied, so (11.40) holds; in particular, $\zeta_{*,N}\mathcal{G}(\tilde{x}_{*,N}, \tilde{y}_{*,N}; \widehat{u}) \leq C_0 := \|u^0 - \widehat{u}\|^2_{Z_1M_1}$ for all $N \geq 2$. It remains to study the convergence rate of the gap from this estimate. We have $\zeta_{*,N} = \sum_{k=1}^{N-1} \varphi_k^{1/2}$. In the unaccelerated case ($\gamma = 0$), we get $\zeta_{*,N} = N\varphi_0^{1/2}$. This gives the claimed $O(1/N)$ rate. In the accelerated case, φ_k is of the order $\Omega(k^2)$ by the proof of Theorem 10.8. Therefore also $\zeta_{*,N}$ is of the order $\Theta(N^2)$, so we get the claimed $O(1/N^2)$ convergence. In the linear convergence case, likewise, φ_k is exponential. Therefore so is $\zeta_{*,N}$. □

Remark 11.17 (spatially adaptive and stochastic methods). Recalling the block-separability Example 10.3, consider the spaces $X = X_1 \times \cdots \times X_m$ and $Y = Y_1 \times \cdots \times Y_n$. Suppose $F(x) = \sum_{j=1}^m F_j(x_j)$ and $G^*(y) = \sum_{\ell=1}^n G_\ell^*(y_\ell)$ for $x = (x_1, \ldots, x_m) \in X$ and $y = (y_1, \ldots, y_n) \in Y$. Take $Z_{k+1} := \left(\begin{smallmatrix} \Phi_k & 0 \\ 0 & \Psi_{k+1} \end{smallmatrix}\right)$ as well as $W_{k+1} := \left(\begin{smallmatrix} T_k & 0 \\ 0 & \Sigma_{k+1} \end{smallmatrix}\right)$ for $T_k := \sum_{j=1}^n \tau_{k,j}P_j$, and similar expressions for Φ_k, Σ_{k+1}, and Σ_{k+1}, where $P_jx := x_j$ projects into X_j. Instead of $\varphi_k\tau_k = \psi_k\sigma_k$ that we required in (10.8), imposing $\mathbb{E}[\Phi_kT_k] = \eta_kI$ and $\mathbb{E}[\Psi_k\Sigma_k] = \eta_kI$ for some scalar η_k, we may then start following through the proof of Theorem 10.8 to derive stochastic block-coordinate methods that randomly update only some of the blocks on each iteration, as well as methods that adapt the blockwise step lengths to the spatial or blockwise structure of the problem. With somewhat more effort, we can also follow through the proofs of the present Section 11.5. Specifically, if we replace our ergodic sequences by

$$\tilde{x}_{*,N} := \zeta_{*,N}^{-1}\mathbb{E}\left[\sum_{k=1}^{N-1} T_k^*\Phi_k^*x^{k+1}\right] \quad \text{and} \quad \tilde{y}_{*,N} := \zeta_{*,N}^{-1}\mathbb{E}\left[\sum_{k=1}^{N-1} \Sigma_k^*\Psi_k^*y^k\right] \quad \text{for} \quad \zeta_{*,N} := \sum_{k=1}^{N-1} \eta_k,$$

we then obtain in place of (11.40) the estimate

$$\mathbb{E}\left[\frac{1}{2}\|u^N - \widehat{u}\|^2_{Z_{N+1}M_{N+1}}\right] + \zeta_{*,N}\mathcal{G}(\tilde{x}_{*,N}, \tilde{y}_{*,N}) + \sum_{k=0}^{N-1} \mathbb{E}\left[\mathcal{V}_{k+1}(\widehat{u})\right] \leq \|u^0 - \widehat{u}\|^2_{Z_1M_1}.$$

If instead $\mathbb{E}[\Phi_kT_k] = \eta_kI$, and $\mathbb{E}[\Psi_{k+1}\Sigma_{k+1}] = \eta_kI$, we get the result for the ergodic sequences

$$\tilde{x}_N := \zeta_N^{-1}\mathbb{E}\left[\sum_{k=0}^{N-1} T_k^*\Phi_k^*x^{k+1}\right] \quad \text{and} \quad \tilde{y}_N := \zeta_N^{-1}\mathbb{E}\left[\sum_{k=0}^{N-1} \Sigma_{k+1}^*\Psi_{k+1}^*y^{k+1}\right] \quad \text{where} \quad \zeta_N := \sum_{k=0}^{N-1} \eta_k.$$

In either case, if we do not or cannot, due to lack of strong convexity of some of the F_ℓ, accelerate all of the blockwise step lengths $\tau_{k+1,j}$ with the same factor $\gamma = \gamma_j$, it will generally be the case that $\mathbb{E}\left[\mathcal{V}_{k+1}(\widehat{u})\right] < 0$. This quantity will have such an order of magnitude that we get mixed $O(1/N^2) + O(1/N)$ convergence rates. We refer to [Valkonen, 2019] for details on such spatially adaptive and stochastic primal-dual methods, and [Wright, 2015] for an introduction to the idea of stochastic coordinate descent.

11.6 CONVERGENCE OF THE ADMM

Let $G : X \to \overline{\mathbb{R}}$, $F : Z \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous, $A \in \mathbb{L}(X;Y)$, $B \in \mathbb{L}(Z;Y)$, and $c \in Y$. Recall the problem

$$\min_{x,z} J(x,z) := G(x) + F(z) + \delta_C(x,z), \tag{11.41}$$

where

$$C := \{(x,z) \in X \times Z \mid Ax + Bz = c\}.$$

We now show an ergodic convergence result for the ADMM applied to this problem, which we recall from (8.30) to read

$$\begin{cases} x^{k+1} \in (A^*A + \tau^{-1}\partial F)^{-1}(A^*(c - Bz^k - \tau^{-1}\lambda^k)), \\ z^{k+1} \in (B^*B + \tau^{-1}\partial G)^{-1}(B^*(c - Ax^{k+1} - \tau^{-1}\lambda^k)), \\ \lambda^{k+1} := \lambda^k + \tau(Ax^{k+1} + Bz^{k+1} - c). \end{cases} \tag{11.42}$$

The general structure of the convergence proof is very similar to all the other algorithms we have studied. However, now the forward-step component does not arise as a gradient $\nabla\tilde{E}$ but is a special nonself-adjoint preconditioner $\tilde{M}_{i+1}$. Moreover, in the first stage of the proof we obtain a convergence estimate for a duality gap that we then refine at the end of the proof to separate function value and constraint satisfaction estimates.

Theorem 11.18. *Let $G : X \to \overline{\mathbb{R}}$ and $F : Z \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous, $A \in \mathbb{L}(X;Y)$, $B \in \mathbb{L}(Z;Y)$, and $c \in Y$. Let J be defined as in (11.41), which we assume to admit a solution $(\widehat{x},\widehat{z}) \in X \times Z$. For arbitrary initial iterates (x^0, y^0, z^0), let $\{(x^{k+1}, z^{k+1}, \lambda^{k+1})\}_{k\in\mathbb{N}} \subset X \times Z \times Y$ be generated by the ADMM (11.42) for (11.41). Define the ergodic sequences $\tilde{x}^N := \frac{1}{N}\sum_{k=0}^{N-1} x^{k+1}$ and $\tilde{z}^N := \frac{1}{N}\sum_{k=0}^{N-1} z^{k+1}$. Then both $(G+F)(\tilde{x}^N, \tilde{z}^N) \to \min_{x\in X} J(x)$ and $\|A\tilde{x}^N + B\tilde{z}^N - c\|_Y \to 0$ at the rate $O(1/N)$.*

Proof. We consider the augmented problem

$$\min_{(x,z)\in X\times Z} J_\tau(x,z) := G(x) + F(z) + \delta_C(x,z) + \frac{\tau}{2}\|Ax + Bz - c\|_Y^2,$$

which has the same solutions as (11.41). As the normal cone to the constraint set C at any point $(x,z) \in C$ is given by $N_C(x,z) = \{(A^*\lambda, B^*\lambda) \mid \lambda \in Y\}$, setting $u = (x,z,\lambda)$ and

$$H(u) := \begin{pmatrix} \partial G(x) + A^*\lambda + \tau A^*(Ax + Bz - c) \\ \partial F(z) + B^*\lambda + \tau B^*(Ax + Bz - c) \\ -(Ax + Bz - c) \end{pmatrix},$$

the optimality conditions for this problem can be written as $0 \in H(u)$. In particular, there exists $\widehat{\lambda} \in Y$ such that $(\widehat{x},\widehat{z},\widehat{\lambda}) \in H^{-1}(0)$. However, we will not be needing this, and take $\widehat{\lambda}$ arbitrary.

We could rewrite the algorithm (11.42) as (11.27) with

$$\widetilde{H}_{k+1}(u) = H(u) \quad \text{and} \quad M_{k+1} = \begin{pmatrix} 0 & -\tau A^*B & -A^* \\ 0 & 0 & -B^* \\ 0 & 0 & \tau^{-1}I \end{pmatrix}.$$

However, M_{k+1} is nonsymmetric, and any symmetrizing Z_{k+1} would make $Z_{k+1}\widetilde{H}_{k+1}$ difficult to analyze. We therefore take instead

$$\widetilde{H}_{k+1}(u) := H(u) + \tilde{M}_{k+1}(u - u^k) \quad \text{with} \quad \tilde{M}_{k+1} := \begin{pmatrix} 0 & -\tau A^*B & -A^* \\ 0 & -\tau B^*B & -B^* \\ 0 & 0 & 0 \end{pmatrix},$$

as well as

$$M_{k+1} := \begin{pmatrix} 0 & 0 & 0 \\ 0 & \tau B^*B & 0 \\ 0 & 0 & \tau^{-1}I \end{pmatrix}, \quad \text{and} \quad Z_{k+1} := I.$$

Clearly $Z_{k+1}M_{k+1}$ is self-adjoint.

Let us set

$$\Gamma := \tau \begin{pmatrix} A^*A & A^*B & 0 \\ B^*A & B^*B & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \Xi := \begin{pmatrix} 0 & 0 & A^* \\ 0 & 0 & B^* \\ -A & -B & 0 \end{pmatrix}.$$

Using the fact that $A\widehat{x} + B\widehat{x} = c$, observe that we can split $H = \partial\tilde{F} + \Xi$, where

$$\begin{aligned} \tilde{F}(u) &:= G(x) + F(z) + \frac{\tau}{2}\|Ax + Bz - c\|_Y^2 + \langle c, \lambda\rangle_Y \\ &= G(x) + F(z) + \frac{1}{2}\|u - \widehat{u}\|_\Gamma^2 + \langle c, \lambda\rangle_Y. \end{aligned}$$

It follows

$$\begin{aligned} \langle H(u^{k+1}), u^{k+1} - \widehat{u}\rangle_{Z_{k+1}} &\geq \tilde{F}(u^{k+1}) - \tilde{F}(\widehat{u}) + \frac{1}{2}\|u^{k+1} - \widehat{u}\|_\Gamma^2 + \langle \widehat{u}, u^{k+1}\rangle_\Xi \\ &= [F(x^{k+1}) + G(z^{k+1})] - [F(\widehat{x}) + G(\widehat{x})] + \langle c, \lambda^{k+1} - \widehat{\lambda}\rangle_Y \\ &\quad + \|u^{k+1} - \widehat{u}\|_\Gamma^2 + \langle \widehat{u}, u^{k+1}\rangle_\Xi. \end{aligned}$$

Again using $A\widehat{x} + B\widehat{x} = c$, we expand

$$\begin{aligned} \langle \widehat{u}, u^{k+1}\rangle_\Xi &= \langle \widehat{\lambda}, Ax^{k+1} + Bz^{k+1}\rangle_Y - \langle A\widehat{x} + B\widehat{z}, \lambda^{k+1}\rangle_Y \\ &= \langle \widehat{\lambda}, Ax^{k+1} + Bz^{k+1} - c\rangle_Y - \langle c, \lambda^{k+1} - \widehat{\lambda}\rangle_Y. \end{aligned}$$

Thus

$$\begin{aligned} \langle H(u^{k+1}), u^{k+1} - \widehat{u}\rangle_{Z_{k+1}} &\geq [F(x^{k+1}) + G(z^{k+1})] - [F(\widehat{x}) + G(\widehat{x})] \\ &\quad + \|u^{k+1} - \widehat{u}\|_\Gamma^2 + \langle \widehat{\lambda}, Ax^{k+1} + Bz^{k+1} - c\rangle_Y \\ &= \bar{F}(u^{k+1}; \widehat{\lambda}) - \bar{F}(\widehat{u}; \widehat{\lambda}) + \|u^{k+1} - \widehat{u}\|_\Gamma^2 \end{aligned} \tag{11.43}$$

for

$$\bar{F}(u;\widehat{\lambda}) := F(x) + G(z) + \langle \widehat{\lambda}, Ax + Bz - c\rangle_Y. \tag{11.44}$$

On the other hand,

$$\begin{aligned}\langle u^{k+1} - u^k, u^{k+1} - \widehat{u}\rangle_{Z_{k+1}\tilde{M}_{k+1}} &= \langle -\tau B(z^{k+1} - z^k) - (\lambda^{k+1} - \lambda^k), A(x^{k+1} - \widehat{x})\rangle_Y \\ &\quad + \langle -\tau B(z^{k+1} - z^k) - (\lambda^{k+1} - \lambda^k), B(z^{k+1} - \widehat{z})\rangle_Y \\ &= \langle -\tau B(z^{k+1} - z^k) - (\lambda^{k+1} - \lambda^k), A(x^{k+1} - \widehat{x}) + B(z^{k+1} - \widehat{z})\rangle_Y.\end{aligned}$$

From (11.42) we recall

$$\lambda^{k+1} - \lambda^k = \tau(Ax^{k+1} + Bz^{k+1} - c) = \tau[A(x^{k+1} - \widehat{x}) + B(z^{k+1} - \widehat{z})].$$

Hence

$$\begin{aligned}\langle u^{k+1} - u^k, u^{k+1} - \widehat{u}\rangle_{Z_{k+1}\tilde{M}_{k+1}} &= -\|u^{k+1} - \widehat{u}\|_\Gamma^2 - \langle B(z^{k+1} - z^k), \lambda^{k+1} - \lambda^k\rangle_Y \\ &\geq -\|u^{k+1} - \widehat{u}\|_\Gamma^2 - \frac{1}{2}\|u^{k+1} - u^k\|_{Z_{i+1}M_{i+1}}^2.\end{aligned} \tag{11.45}$$

Combining (11.43) and (11.45) it follows that

$$\langle \widetilde{H}_{k+1}(u^{k+1}), u^{k+1} - \widehat{u}\rangle_{Z_{k+1}} \geq \bar{F}(u^{k+1};\widehat{\lambda}) - \bar{F}(\widehat{u};\widehat{\lambda}) - \frac{1}{2}\|u^{k+1} - u^k\|_{Z_{i+1}M_{i+1}}^2.$$

By Theorem 11.12 now

$$\frac{1}{2}\|u^N - \widehat{u}\|_{Z_{N+1}M_{N+1}}^2 + \sum_{k=0}^{N-1}\left(\bar{F}(u^{k+1};\widehat{\lambda}) - \bar{F}(\widehat{u};\widehat{\lambda})\right) \leq \frac{1}{2}\|u^0 - \widehat{u}\|_{Z_1M_1}^2 \quad (N \geq 1).$$

Writing $\tilde{u}^N = (\tilde{x}^N, \tilde{y}^N, \tilde{\lambda}^N) := \frac{1}{N}\sum_{k=0}^{N-1} u^{k+1}$, Jensen's inequality now shows that

$$\bar{F}(\tilde{u}^N;\widehat{\lambda}) - \bar{F}(\widehat{u};\widehat{\lambda}) \leq \frac{1}{2N}\|u^0 - \widehat{u}\|_{Z_1M_1}^2 \quad (N \geq 1). \tag{11.46}$$

Since $A\widehat{x} + B\widehat{z} = c$, observe that $\bar{F}(\,\cdot\,;\widehat{\lambda}) - \bar{F}(\widehat{u};\widehat{\lambda})$ is the Lagrangian duality gap (11.7) for the minmax formulation (8.27) of (11.41), hence nonnegative when $\widehat{u} \in H^{-1}(0)$. So (11.46) shows the convergence of the duality gap. However, we can improve the result somewhat since $\widehat{\lambda}$ was taken as arbitrary. Expanding $\bar{F}$ using (11.44) and taking the supremum over $\widehat{\lambda} \in \mathbb{B}(0,\kappa)$ in (11.46), we thus obtain for any $\kappa > 0$ the estimate

$$\begin{aligned}0 \leq [F(\tilde{x}^N) + G(\tilde{z}^N)] - [F(\widehat{x}) + G(\widehat{x})] + \kappa\|A\tilde{x}^N + B\tilde{z}^N - c\|_Y \\ = \sup_{\widehat{\lambda}\in\mathbb{B}(0,\kappa)}\left(F(\tilde{u}^N;\widehat{\lambda}) - \bar{F}(\widehat{u};\widehat{\lambda})\right) \leq \sup_{\widehat{\lambda}\in\mathbb{B}(0,\kappa)}\frac{1}{2N}\|u^0 - \widehat{u}\|_{Z_1M_1}^2.\end{aligned}$$

This gives the claim. □

12 META-ALGORITHMS

In this chapter, we consider several *"meta-algorithms"* for accelerating minimization algorithms such as the ones derived in the previous chapters. These include *inertia* and *over-relaxation*, as well as *line searches*. These schemes differ from the strong convexity based acceleration of Chapter 9 in that no additional assumptions are made on F and G. Rather, through the use of an additional extrapolated or interpolated point, the first two schemes attempt to obtain a second-order approximation of the function. Line search, on the other hand, can be used to find optimal parameters or to estimate unknown parameters. Throughout the chapter, we base our work in the abstract algorithm (11.27), i.e.,

$$0 \in \widetilde{H}_{k+1}(x^{k+1}) + M_{k+1}(x^{k+1} - x^k), \tag{12.1}$$

where the iteration-dependent set-valued operator $\widetilde{H}_{k+1} : X \rightrightarrows X$ in suitable sense approximates a (monotone) operator $H : X \rightrightarrows X$, whose root we intend to find, and $M_{k+1} \in \mathbb{L}(X; X)$ is a linear preconditioner.

12.1 OVER-RELAXATION

We start with *over-relaxation*. Essentially, this amounts to taking (12.1) and replacing x^k in the preconditioner by an over-relaxed point z^k defined for some parameters $\lambda_k > 0$ through the recurrence

$$z^{k+1} := \lambda_k^{-1} x^{k+1} + (1 - \lambda_k^{-1}) z^k. \tag{12.2}$$

We thus seek to solve

$$0 \in \widetilde{H}_{k+1}(x^{k+1}) + M_{k+1}(x^{k+1} - z^k). \tag{12.3}$$

Since $z^{k+1} - z^k = \lambda_k^{-1}(x^{k+1} - z^k)$, we can write (12.1) as

$$0 \in \widetilde{H}_{k+1}(x^{k+1}) + \lambda_k M_{k+1}(z^{k+1} - z^k). \tag{12.4}$$

We can therefore lift the overall algorithm into the form (12.1) as

$$0 \in \hat{H}_{k+1}(q^{k+1}) + \hat{M}_{k+1}(q^{k+1} - q^k) \tag{12.5}$$

by taking $q := (x, z)$ with

$$\hat{H}_{k+1}(q) := \begin{pmatrix} \widetilde{H}_{k+1}(x) \\ \lambda_k^{-1}(z-x) \end{pmatrix} \quad \text{and} \quad \hat{M}_{k+1} := \begin{pmatrix} 0 & \lambda_k M_{k+1} \\ 0 & (I-\lambda_k^{-1})I \end{pmatrix}. \tag{12.6}$$

To be able to use our previous estimate on $\langle \widetilde{H}_{k+1}(x^{k+1}), x^{k+1} - \widehat{x} \rangle_{Z_{k+1}}$, we would like to test with

$$\hat{Z}_{k+1} := \begin{pmatrix} \lambda_k Z_{k+1} & 0 \\ 0 & 0 \end{pmatrix}.$$

Unfortunately, $Z_{k+1}M_{k+1}$ is not self-adjoint, so Theorem 11.12 does not apply. However, observing from (12.2) that

$$z^{k+1} - x^{k+1} = (1-\lambda_k)(z^{k+1} - z^k), \tag{12.7}$$

we are able to proceed along the same lines of proof.

Theorem 12.1. *On a Hilbert space X, let $\widetilde{H}_{k+1} : X \rightrightarrows X$, and $M_{k+1}, Z_{k+1} \in \mathbb{L}(X;X)$ for $k \in \mathbb{N}$. Suppose (12.3) is solvable for the iterates $\{x^k\}_{k\in\mathbb{N}}$. If $Z_{k+1}M_{k+1}$ is self-adjoint,*

$$\lambda_k^2 Z_{k+1}M_{k+1} \geq \lambda_{k+1}^2 Z_{k+2}M_{k+2}, \tag{12.8}$$

and

$$\langle \widetilde{H}_{k+1}(x^{k+1}), x^{k+1} - \widehat{x} \rangle_{Z_{k+1}} \geq \mathcal{V}_{k+1}(\widehat{x}) - \frac{1}{2}\|x^{k+1} - z^k\|^2_{Z_{k+1}Q_{k+1}} \tag{12.9}$$

for some $Q_{k+1} \in \mathbb{L}(X;X)$, for all $k \in \mathbb{N}$ and some $\widehat{x} \in X$ and $\mathcal{V}_{k+1}(\widehat{x}) \in \mathbb{R}$, then

$$\frac{\lambda_{k+1}^2}{2}\|z^{k+1} - \widehat{x}\|^2_{Z_{k+2}M_{k+2}} + \lambda_k \mathcal{V}_{k+1}(\widehat{x}) + \frac{\lambda_k}{2}\|z^{k+1} - z^k\|^2_{\lambda_k(2\lambda_k - 1)Z_{k+1}M_{k+1} - Z_{k+1}Q_{k+1}} \leq \frac{\lambda_k^2}{2}\|z^k - \widehat{x}\|^2_{Z_{k+1}M_{k+1}} \quad (k \in \mathbb{N}). \tag{12.10}$$

Proof. Taking $\widehat{q} := (\widehat{x}, \widehat{x})$, we apply $\langle \cdot, q^{k+1} - \widehat{q} \rangle_{\hat{Z}_{k+1}}$ to (12.3). Thus

$$0 \in \langle \hat{H}_{k+1}(q^{k+1}) + \hat{M}_{k+1}(q^{k+1} - q^k), q^{k+1} - \widehat{q} \rangle_{\hat{Z}_{k+1}}.$$

Observe that

$$\hat{Z}_{k+1}\hat{M}_{k+1} = \begin{pmatrix} 0 & \lambda_k^2 Z_{k+1}M_{k+1} \\ 0 & 0 \end{pmatrix}.$$

Thus

$$0 \in \langle \widetilde{H}_{k+1}(x^{k+1}), x^{k+1} - \widehat{x} \rangle_{\lambda_k Z_{k+1}} + \lambda_k^2 \langle z^{k+1} - z^k, x^{k+1} - \widehat{x} \rangle_{Z_{k+1}M_{k+1}}.$$

Using (12.7) we then get

$$0 \in \langle \widetilde{H}_{k+1}(x^{k+1}), x^{k+1} - \widehat{x} \rangle_{\lambda_k Z_{k+1}} - \lambda_k^2(1-\lambda_k)\|z^{k+1} - z^k\|^2_{Z_{k+1}M_{k+1}} + \lambda_k^2 \langle z^{k+1} - z^k, z^{k+1} - \widehat{x} \rangle_{Z_{k+1}M_{k+1}}.$$

Using the three-point-identity (9.1), we rearrange this into

$$0 \in \langle \widetilde{H}_{k+1}(x^{k+1}), x^{k+1} - \widehat{x} \rangle_{\lambda_k Z_{k+1}} + \frac{\lambda_k^2 - 2\lambda_k^2(1-\lambda_k)}{2} \|z^{k+1} - z^k\|^2_{Z_{k+1}M_{k+1}} + \frac{\lambda_k^2}{2}\|z^{k+1} - \widehat{x}\|^2_{Z_{k+1}M_{k+1}} - \frac{\lambda_k^2}{2}\|z^k - \widehat{x}\|^2_{Z_{k+1}M_{k+1}}.$$

Observe that $\lambda_k^2 - 2\lambda_k^2(1-\lambda_k) = \lambda_k^2(2\lambda_k - 1)$. Using (12.2), (12.9), and (12.8), this gives (12.10). □

Clearly we should try to ensure $\lambda_k(2\lambda_k - 1)Z_{k+1}M_{k+1} \geq Z_{k+1}Q_{k+1}$. If $Z_{k+1}M_{k+1} = Z_0M_0$ is constant and $Q_{k+1} = 0$, this holds if $\{\lambda_k\}_{k\in N}$ is nonincreasing and satisfies $\lambda_k \geq 1/2$. Therefore, we cannot get any convergence rates from the iterates in this case. It is, however, possible to obtain convergence of a gap, and it would be possible to obtain weak convergence.

The next result is a variant of Corollary 11.8 for over-relaxed methods.

Corollary 12.2. *Let $H := \partial\tilde{F} + \nabla\tilde{G} + \Xi$, where $\Xi \in \mathbb{L}(X;X)$ is skew-adjoint, and $\tilde{G} : X \to \overline{\mathbb{R}}$ and $\tilde{F} : X \to \mathbb{R}$ convex, proper, and lower semicontinuous. Suppose $\tilde{F}$ satisfies for some $\Lambda \in \mathbb{L}(X;X)$ the three-point smoothness condition (11.19). Also let $M \in \mathbb{L}(X;X)$ be positive semi-definite and self-adjoint. Pick $x^0 = z^0 \in X$, and define the sequence $\{(x^{k+1}, z^{k+1})\}_{k\in\mathbb{N}}$ through*

$$\begin{cases} 0 \in [\partial\tilde{G}(x^{k+1}) + \partial\tilde{F}(z^k) + \Xi x^{k+1}] + M(x^{k+1} - z^k), \\ z^{k+1} := \lambda_k^{-1}x^{k+1} - (\lambda_k^{-1} - 1)z^k. \end{cases} \tag{12.11}$$

Suppose $\{\lambda_k\}_{k\in\mathbb{N}}$ is nonincreasing and

$$\lambda_k(2\lambda_k - 1)M \geq \Lambda \quad (k \in \mathbb{N}). \tag{12.12}$$

Then for every $\widehat{x} \in H^{-1}(0)$ and the gap functional $\tilde{\mathcal{G}}$ defined in (11.4),

$$\tilde{\mathcal{G}}(\tilde{x}^N; \widehat{x}) \leq \frac{\lambda_0^2}{2\sum_{k=0}^{N-1}\lambda_k}\|z^0 - \widehat{x}\|_M^2, \quad \text{where} \quad \tilde{x}^N := \frac{1}{\sum_{k=0}^{N-1}\lambda_k}\sum_{k=0}^{N-1}\lambda_k x^{k+1}. \tag{12.13}$$

Proof. The method (12.11) is (12.3) with $\tilde{H}_{k+1}(x) := \partial\tilde{G}(x) + \nabla\tilde{F}(z^k) + \Xi x$ as well as $M_{k+1} \equiv M$ and $Z_{k+1} \equiv \mathrm{Id}$. Using (11.19) for $\tilde{F}$, the convexity of $\tilde{G}$, and the assumption $ZW = \eta\mathrm{Id}$, we obtain as in the proof of (11.7) the estimate

$$\langle \widetilde{H}_{k+1}(x^{k+1}), x^{k+1} - \widehat{x} \rangle \geq \tilde{\mathcal{G}}(x^{k+1}; \widehat{x}) - \frac{1}{2}\|z^k - x^{k+1}\|_\Lambda^2$$

This provides (12.9) while (12.12) and the constant choice of the testing and preconditioning operators guarantee that $\lambda_k(2\lambda_k - 1)Z_{k+1}M_{k+1} \geq Z_{k+1}Q_{k+1}$ for $Q_{k+1} \equiv \Lambda$. By Theorem 12.1, we now obtain

$$\frac{\lambda_{k+1}^2}{2}\|z^{k+1} - \widehat{x}\|_M^2 + \lambda_k \tilde{\mathcal{G}}(x^{k+1}; \widehat{x}) \leq \frac{\lambda_k^2}{2}\|z^k - \widehat{x}\|_M^2. \tag{12.14}$$

Summing over $k = 0, \dots, N-1$ and an application of Jensen's inequality finishes the proof. □

OVER-RELAXED PROXIMAL POINT METHOD

We apply the above results to the *over-relaxed proximal point method*

$$\begin{cases} x^{k+1} := \operatorname{prox}_{\tau G}(z^k), \\ z^{k+1} := \lambda_k^{-1} x^{k+1} - (\lambda_k^{-1} - 1) z^k. \end{cases} \tag{12.15}$$

Theorem 12.3. *Let $G : X \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous with $[\partial G]^{-1}(0) \neq \emptyset$. Pick an initial iterate $x^0 = z^0 \in X$. If $\{\lambda_k\}_{k\in\mathbb{N}} \geq 1/2$ is nonincreasing, the ergodic sequence $\{\tilde{x}^N\}_{N\in\mathbb{N}}$ defined in (12.13) and generated from the iterates $\{x^k\}_{k\in\mathbb{N}}$ of the over-relaxed proximal point method (12.15) satisfies $G(\tilde{x}^N) \to G_{\min} := \min_{x\in X} G(x)$ at the rate $O(1/N)$.*

Proof. We apply Corollary 12.2 with $\tilde{G} = G$, $\tilde{F} = 0$, $M = \tau^{-1}\mathrm{Id}$. Clearly $\tilde{F}$ satisfies (11.19) with $\Lambda = 0$. Then (12.12) holds if $2\lambda_k \geq 1$, that is to say $\lambda_k \geq 1/2$. For $\widehat{x} \in \arg\min G$, we have $\tilde{\mathcal{G}}(x; \widehat{x}) = G(x) - G(\widehat{x}) = G(x) - G_{\min}$. Therefore Corollary 12.2 gives

$$G(\tilde{x}^N) \leq G_{\min} + \frac{\lambda_0^2}{2\tau \sum_{k=0}^{N-1} \lambda_k}\|z^0 - \widehat{x}\|_X^2 \tag{12.16}$$

Since $\sum_{k=0}^{N-1} \lambda_k \geq N/2$ by the lower bound on λ_k, we get the claimed $O(1/N)$ convergence rate of the function values for the ergodic sequence. □

OVER-RELAXED EXPLICIT SPLITTING

For a smooth function F, the *over-relaxed explicit splitting method* iterates

$$\begin{cases} x^{k+1} := \operatorname{prox}_{\tau G}(z^k - \tau \nabla F(z^k)), \\ z^{k+1} := \lambda_k^{-1} x^{k+1} - (\lambda_k^{-1} - 1) z^k. \end{cases} \tag{12.17}$$

Theorem 12.4. *Let $J := G + F$ for $G : X \to \overline{\mathbb{R}}$ and $F : X \to \mathbb{R}$ be convex, proper, and lower semicontinuous with ∇F L-Lipschitz. Suppose $[\partial J]^{-1}(0) \neq \emptyset$. Pick an initial iterate $x^0 = z^0 \in X$. If $\{\lambda_k\}_{k\in\mathbb{N}}$ is nonincreasing and satisfies*

$$\lambda_k \geq \frac{1}{4}(1 + \sqrt{1 + 8L\tau}), \tag{12.18}$$

then the ergodic sequence $\{\tilde{x}^N\}_{N\in\mathbb{N}}$ defined in (12.13) and generated from the iterates $\{x^k\}_{k\in\mathbb{N}}$ of the over-relaxed explicit splitting method (12.17) satisfies $J(\tilde{x}^N) \to J_{\min} := \min_{x\in X} J(x)$ at the rate $O(1/N)$.

Proof. We apply Corollary 12.2 with $\tilde{G} = G$, $\tilde{F} = F$, and $M = \tau^{-1}\mathrm{Id}$. By Corollary 7.2, $\tilde{F}$ satisfies the three-point smoothness condition (11.19) with $\Lambda = L\,\mathrm{Id}$. The condition (12.12) consequently holds if $\lambda_k(2\lambda_k - 1) > L\tau$, which holds under the assumption (12.18). The rest follows as in the proof of Theorem 12.3. □

OVER-RELAXED PDPS

With $F = F_0 + E : X \to \overline{\mathbb{R}}$, $G^* : Y \to \overline{\mathbb{R}}$, and $K \in \mathbb{L}(X;Y)$, take $H : X \times Y \rightrightarrows X \times Y$ as well as $\tilde{F}, \tilde{G}$, and Ξ as in (11.6), and the preconditioner M as in (11.25) for fixed step length parameters $\tau, \sigma > 0$. Writing $z^k = (\xi^k, v^k)$, and, as usual $u^k = (x^k, y^k)$, the method (12.4) then becomes the *over-relaxed primal-dual proximal splitting (PDPS) method* with a forward step, also known as the *Vũ–Condat method*:

$$\begin{cases} x^{k+1} := (I + \tau\partial F_0)^{-1}(\xi^k - \tau K^* y^k - \tau\nabla E(\xi^k)), \\ \bar{x}^{k+1} := (x^{k+1} - \xi^k) + x^{k+1}, \\ y^{k+1} := (I + \sigma\partial G^*)^{-1}(v^k + \sigma K\bar{x}^{k+1}), \\ \xi^{k+1} := \lambda_k^{-1}x^{k+1} - (\lambda_k^{-1} - 1)\xi^k, \\ v^{k+1} := \lambda_k^{-1}y^{k+1} - (\lambda_k^{-1} - 1)v^k. \end{cases} \tag{12.19}$$

For the statement of the next result, we recall that for the primal-dual saddle-point operator H from (11.6), the generic gap functional $\tilde{\mathcal{G}}$ becomes the primal-dual gap $\mathcal{G}$ given in (11.7).

Theorem 12.5. *Suppose $F_0 : X \to \overline{\mathbb{R}}$, $E : X \to \mathbb{R}$ and $G : Y \to \overline{\mathbb{R}}$ are convex, proper, and lower semicontinuous on Hilbert spaces X and Y with ∇E L-Lipschitz. Let also $K \in \mathbb{L}(X;Y)$. With $F = F_0 + E$, suppose the assumptions of Theorem 5.11 are satisfied. Pick an initial iterate $u^0 = z^0 \in X \times Y$. If the sequence $\{\lambda_k\}_{k\in\mathbb{N}}$ is nonincreasing and satisfies*

$$\lambda_k \geq \frac{1}{4}(1 + \sqrt{1 + 8L\tau/(1 - \tau\sigma\|K\|^2)}) \quad \text{and} \quad \tau\sigma\|K\|^2 \leq 1, \tag{12.20}$$

then the ergodic sequence $\{\tilde{u}^N = (\tilde{x}^N, \tilde{y}^N)\}_{N\in\mathbb{N}}$ defined as in (12.13) and generated from the iterates $\{u^k = (x^k, y^k)\}_{k\in\mathbb{N}}$ of the over-relaxed PDPS method (12.19) satisfies $\mathcal{G}(\tilde{x}_{*,N}, \tilde{y}_{*,N}) \to 0$ at the rate $O(1/N)$.

Proof. We recall that $H^{-1}(0) \neq \emptyset$ under the assumptions of Theorem 5.11. Clearly M is self-adjoint. The condition (12.12) can with (10.32) be reduced to

$$\begin{pmatrix} \lambda_k(2\lambda_k - 1)\delta\tau \mathrm{Id} & 0 \\ 0 & \sigma^{-1}I - \tau(1-\delta)^{-1}KK^* \end{pmatrix} \geq \begin{pmatrix} L & 0 \\ 0 & 0 \end{pmatrix}$$

for some $\delta \in (0,1)$. As in (10.34) in the proof of Theorem 10.8, these conditions reduce to

$$\lambda_k(2\lambda_k - 1)\delta \geq \tau L \quad \text{and} \quad 1 - \delta \geq \tau\sigma\|K\|^2. \tag{12.21}$$

The first inequality holds if $\lambda_k \geq \frac{1}{4}(1 + \sqrt{1 + 8L\tau\delta^{-1}})$. Solving the second inequality as an equality for δ yields the condition

$$\lambda_k \geq \frac{1}{4}(1 + \sqrt{1 + 8L\tau[(1 - \tau\sigma\|K\|^2)]^{-1}}),$$

i.e., (12.20). Now we obtain the gap convergence from Corollary 12.2. □

Remark 12.6. The method (12.19) is due to [Condat, 2013; Vũ, 2013]. The convergence of the ergodic gap was observed in [Chambolle and Pock, 2015].

12.2 INERTIA

Our next *inertial* meta-algorithm will likewise not yield convergence of the main iterates, but through a special arrangement of variables combined with intricate unrolling arguments, is able to do away with the word *ergodic* in the gap estimates. In essence, the meta-algorithm replaces the previous iterate x^k in the linear preconditioner of (12.1) by an inertial point

$$\bar{x}^k := (1 + \alpha_k)x^k - \alpha_k x^{k-1} \quad \text{for} \quad \alpha_k := \lambda_k(\lambda_{k-1}^{-1} - 1) \tag{12.22}$$

for some *inertial parameter* sequence $\{\lambda_k\}_{k\in\mathbb{N}}$. We thus solve

$$0 \in \widetilde{H}_{k+1}(x^{k+1}) + M_{k+1}(x^{k+1} - \bar{x}^k). \tag{12.23}$$

We can relate this to over-relaxation as follows: we simply replace z^k in the definition (12.2) of z^{k+1} by x^k, i.e., we take

$$z^{k+1} := \lambda_k^{-1}x^{k+1} - (\lambda_k^{-1} - 1)x^k. \tag{12.24}$$

Since

$$\begin{aligned} \lambda_k(z^{k+1} - z^k) &= x^{k+1} - (1 - \lambda_k)x^k - \lambda_k[\lambda_{k-1}^{-1}x^k - (\lambda_{k-1}^{-1} - 1)x^{k-1}] \\ &= x^{k+1} - [1 - \lambda_k + \lambda_k\lambda_{k-1}^{-1}]x^k + \lambda_k(\lambda_{k-1}^{-1} - 1)x^{k-1} \\ &= x^{k+1} - [(1 + \alpha_k)x^k - \alpha_k x^{k-1}] = x^{k+1} - \bar{x}^k, \end{aligned} \tag{12.25}$$

we obtain the method (12.4), with the differing update (12.24) of z^{k+1}. Again we can also lift the overall algorithm into the form (12.1), specifically (12.5), by taking $q := (x, z)$ with

$$\hat{H}_{k+1}(q) := \begin{pmatrix} \widetilde{H}_{k+1}(x) \\ z - x \end{pmatrix}, \quad \text{and} \quad \hat{M}_{k+1} := \begin{pmatrix} 0 & \lambda_k M_{k+1} \\ (I - \lambda_k^{-1})I & 0 \end{pmatrix}.$$

Now comes the trick with inertial methods: Unlike with over-relaxed methods, where we wanted to avoid having to estimate $\langle \widetilde{H}_{k+1}(x^{k+1}), z^{k+1} - \widehat{z} \rangle_{Z_{k+1}}$, with inertial methods we are brave enough to do this. Indeed, our specific choice (12.24) makes this possible, as we shall see below. We therefore test with

$$\hat{Z}_{k+1} := \begin{pmatrix} 0 & 0 \\ \lambda_k Z_{k+1} & 0 \end{pmatrix}$$

to obtain a self-adjoint and positive semi-definite

$$\hat{Z}_{k+1} M_{k+1} = \begin{pmatrix} 0 & 0 \\ 0 & \lambda_k^2 Z_{k+1} M_{k+1} \end{pmatrix}. \tag{12.26}$$

Therefore Theorem 11.12 applies, and we obtain the following:

Theorem 12.7. *Let X be a Hilbert space, $\widetilde{H}_{k+1} : X \rightrightarrows X$, and $M_{k+1}, Z_{k+1} \in \mathbb{L}(X; X)$ for $k \in \mathbb{N}$. Suppose (12.23) is solvable for the iterates $\{x^k\}_{k\in\mathbb{N}}$ and inertial parameters $\{\lambda_k\}_{k\in\mathbb{N}} \subset (0, \infty)$. If $Z_{k+1}M_{k+1}$ is self-adjoint, and*

$$\begin{aligned} \lambda_k \langle \widetilde{H}_{k+1}(x^{k+1}), z^{k+1} - \widehat{z} \rangle_{Z_{k+1}} \geq \mathcal{V}_{k+1}(\widehat{x}) + \frac{1}{2}\|z^{k+1} - \widehat{z}\|^2_{\lambda_{k+1}^2 Z_{k+2} M_{k+2} - \lambda_k^2 Z_{k+1} M_{k+1}} \\ - \frac{\lambda_k^2}{2}\|z^{k+1} - z^k\|^2_{Z_{k+1}M_{k+1}} \end{aligned} \tag{12.27}$$

for all $k \in \mathbb{N}$ and some $\widehat{x} \in X$ and $\mathcal{V}_{k+1}(\widehat{x}) \in \mathbb{R}$, then

$$\frac{\lambda_N^2}{2}\|z^N - \widehat{x}\|^2_{Z_{N+1}M_{N+1}} + \sum_{k=0}^{N-1} \mathcal{V}_{k+1}(\widehat{x}) \leq \frac{\lambda_0^2}{2}\|z^0 - \widehat{x}\|^2_{Z_1 M_1} \quad (N \geq 1).$$

Proof. This follows directly from Theorem 11.12 and the expansion (12.26). □

We now provide examples of how to apply this result to the proximal point method and explicit splitting. As we recall, in these algorithms we take $Z_{k+1} = \varphi_k I$ and $W_{k+1} = \tau_k I$. To proceed, we will need a few further general-purpose technical lemmas. The first one is the fundamental lemma for inertia, which provides inertial function value unrolling.

Lemma 12.8. *Let $G : X \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous. Suppose $\lambda_k \in [0,1]$ and $\varphi_k, \tau_k > 0$ for $k \in \mathbb{N}$ with*

$$\varphi_{k+1}\tau_{k+1}(1-\lambda_{k+1}) \le \varphi_k\tau_k \quad (k \ge 0). \tag{12.28}$$

Assume $q^{k+1} \in \partial G(x^{k+1})$ for $k = 0, \dots, N-1$, and $0 \in \partial G(\widehat{x})$. Then

$$\begin{aligned} s_{G,N} &:= \sum_{k=0}^{N-1} \varphi_k\tau_k\lambda_k\langle q^{k+1}, z^{k+1} - \widehat{x}\rangle_X \\ &\ge \varphi_{N-1}\tau_{N-1}(G(x^N) - G(\widehat{x})) - \varphi_0\tau_0(1-\lambda_0)(G(x^0) - G(\widehat{x})). \end{aligned} \tag{12.29}$$

Proof. Using (12.24), observe that

$$\begin{aligned} \lambda_k(z^{k+1} - \widehat{x}) &= \lambda_k[\lambda_{k+1}^{-1}x^{k+1} - (\lambda_k^{-1} - 1)x^k - \widehat{x}] \\ &= \lambda_k(x^{k+1} - \widehat{x}) + (1-\lambda_k)(x^{k+1} - x^k). \end{aligned} \tag{12.30}$$

Recalling from (12.30) that $\lambda_k(z^{k+1} - \widehat{x}) = \lambda_k(x^{k+1} - \widehat{x}) + (1-\lambda_k)(x^{k+1} - x^k)$ and using the convexity of G, we can estimate

$$\begin{aligned} s_{G,N} &= \sum_{k=0}^{N-1} \varphi_k\tau_k\Big[\lambda_k\langle q^{k+1}, x^{k+1} - \widehat{x}\rangle_X + (1-\lambda_k)\langle q^{k+1}, x^{k+1} - x^k\rangle_X\Big] \\ &\ge \sum_{k=0}^{N-1} \varphi_k\tau_k\Big[\lambda_k(G(x^{k+1}) - G(\widehat{x})) + (1-\lambda_k)(G(x^{k+1}) - G(x^k))\Big] \\ &= \sum_{k=0}^{N-1} \Big[\varphi_k\tau_k(G(x^{k+1}) - G(\widehat{x})) - \varphi_k\tau_k(1-\lambda_k)(G(x^k) - G(\widehat{x}))\Big]. \end{aligned} \tag{12.31}$$

Since $G(x^k) \ge G(\widehat{x})$, the recurrence inequality (12.28) together with a telescoping argument now gives

$$s_{G,N} \ge \varphi_{N-1}\tau_{N-1}(G(x^N) - G(\widehat{x})) - \varphi_0\tau_0(1-\lambda_0)(G(x^0) - G(\widehat{x})).$$

This is the claim. □

Lemma 12.9. *Suppose $\lambda_0 = 1$ and $\lambda_k^{-2} = \lambda_{k+1}^{-2} - \lambda_{k+1}^{-1}$ for $k = 0, \dots, N-1$. Then*

$$\lambda_{k+1} = \frac{2}{1 + \sqrt{1 + 4\lambda_k^{-2}}} \qquad (k = 0, \dots, N-1) \tag{12.32}$$

and $\lambda_N^{-1} \ge (N+1)$.

Proof. First, the recurrence (12.32) is a simple solution of the assumed quadratic equation. We show the lower bound by total induction on N. Assume that $\lambda_k^{-1} \geq (k+1)$ for all $k = 0, \dots, N-1$. Rearranging the original update as

$$\lambda_{k+1}^{-2} - \lambda_{k+1}^{-1} = \lambda_k^{-2} - \lambda_k^{-1} + \lambda_k^{-1},$$

summing over $k = 0, \dots, N-1$, and telescoping yields

$$\lambda_N^{-2} - \lambda_N^{-1} = \sum_{k=0}^{N-1} \lambda_k^{-1}.$$

From the induction assumption, we thus obtain $\lambda_N^{-2} - \lambda_N^{-1} \geq (N+2)(N+1)$. Solving this quadratic inequality as an equality then shows that $\lambda_N^{-1} \geq (1 + \sqrt{1 + 4(N+2)(N+1)}) \geq (N+1)$, which completes the proof. □

INERTIAL PROXIMAL POINT METHOD

Let $H = \partial G$ and $\tilde{H}_{k+1} = \tau \partial G$ for a convex, proper, lower semicontinuous function G. Take $\tau > 0$ and λ_{k+1} by (12.32) for $\lambda_0 = 1$. Then (12.23) becomes the *inertial proximal point method*

$$\begin{cases} x^{k+1} := \operatorname{prox}_{\tau G}(\bar{x}^k), \\ \alpha_{k+1} := \lambda_{k+1}(\lambda_k^{-1} - 1), \\ \bar{x}^{k+1} := (1 + \alpha_{k+1})x^{k+1} - \alpha_{k+1}x^k. \end{cases} \tag{12.33}$$

Note that x^0 is never needed, as $\alpha_1 = 0$. The real initial iterate, which can be freely chosen, is $\bar{x}^0$.

Theorem 12.10. *Let $G : X \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous. Suppose $[\partial G]^{-1}(0) \neq \emptyset$. Take $\tau > 0$ and $\lambda_0 = 1$, and pick an initial iterate $\bar{x}^0 \in X$. Then the inertial proximal point method (12.33) satisfies $G(x^N) \to G_{\min}$ at the rate $O(1/N^2)$.*

Proof. If we take $\tau_k = \tau$ as stated and $\varphi_k = \lambda_k^{-2}$, then (12.9) verifies (12.28). Since now $\lambda_{k+1}^2 \varphi_{k+1} = \lambda_k^2 \varphi_k$, (12.27) holds if

$$\lambda_k \varphi_k \tau_k \langle \partial G(x^{k+1}), z^{k+1} - \widehat{z} \rangle_X \geq \mathcal{V}_{k+1}(\widehat{x}) - \frac{\lambda_k^2 \varphi_k}{2} \|z^{k+1} - z^k\|_X^2 \tag{12.34}$$

for some $\mathcal{V}_{k+1}(\widehat{x}) \in \mathbb{R}$. This is verified by Lemma 12.8 for some $\mathcal{V}_{k+1}(\widehat{x})$ such that

$$\sum_{k=0}^{N-1} \mathcal{V}_{k+1}(\widehat{x}) \geq \varphi_{N-1}\tau_{N-1}(G(x^N) - G(\widehat{x})) - \varphi_0\tau_0(1 - \lambda_0)(G(x^0) - G(\widehat{x})).$$

Since $\lambda_0 = 1$, Theorem 12.7 gives the estimate

$$\frac{\varphi_N \lambda_N^2}{2} \|x^N - \widehat{x}\|_X^2 + \varphi_{N-1}\tau_{N-1}(G(x^N) - G(\widehat{x})) \leq \frac{\varphi_0 \lambda_0^2}{2} \|x^0 - \widehat{x}\|_X^2.$$

By Lemma 12.9 now $\varphi_{N-1}\tau_{N-1} = \lambda_{N-1}^{-2}\tau \geq \tau N^2$. Therefore we obtain the claimed convergence rate. □

INERTIAL EXPLICIT SPLITTING

Let $H = \partial G + \nabla F$ and $\tilde{H}_{k+1}(x) = \tau(\partial G(x) + \nabla F(\bar{x}^k))$ for convex, proper, lower semicontinuous functions G and F with F smooth. Take $\tau > 0$ and λ_{k+1} by (12.32) for $\lambda_0 = 1$. Then (12.23) becomes the *inertial explicit splitting method*

$$\begin{cases} x^{k+1} := \operatorname{prox}_{\tau G}(\bar{x}^k - \tau \nabla F(\bar{x}^k)), \\ \alpha_{k+1} := \lambda_{k+1}(\lambda_k^{-1} - 1), \\ \bar{x}^{k+1} := (1 + \alpha_{k+1}) x^{k+1} - \alpha_{k+1} x^k. \end{cases} \tag{12.35}$$

Again, x^0 is never needed, as $\alpha_1 = 0$. The freely pickable initial iterate is $\bar{x}^0$.

To prove the convergence of this method, we need to incorporate the forward step into Lemma 12.8.

Lemma 12.11. *Let $J := F + G$ for $G : X \to \overline{\mathbb{R}}$ and $F : X \to \mathbb{R}$ be convex, proper, and lower semicontinuous. Suppose F has L-Lipschitz gradient and that $\lambda_k \in [0, 1]$ and $\varphi_k, \tau_k > 0$ satisfy the recurrence inequality (12.28) for $k \in \mathbb{N}$. Assume $w^{k+1} \in \partial G(x^{k+1})$ for all $k = 0, \dots, N-1$, and that $0 \in \partial J(\widehat{x})$. Then*

$$\begin{aligned} s_N := \sum_{k=0}^{N-1} \left(\varphi_k \tau_k \lambda_k \langle w^{k+1} + \nabla F(\bar{x}^k), z^{k+1} - \widehat{x} \rangle_X + \frac{\varphi_k \tau_k \lambda_k^2 L}{2} \|z^{k+1} - z^k\|_X^2 \right) \\ \geq \varphi_{N-1}\tau_{N-1}(J(x^N) - J(\widehat{x})) - \varphi_0 \tau_0 (1 - \lambda_0)(J(x^0) - J(\widehat{x})). \end{aligned} \tag{12.36}$$

Proof. We recall from (12.25) that $\frac{\lambda_k^2}{2}\|z^{k+1} - z^k\|_X^2 = \frac{1}{2}\|x^{k+1} - \bar{x}^k\|_X^2$. We therefore estimate

using Corollary 7.2 that

(12.37)
$$\begin{aligned} s_{F,N} &:= \sum_{k=0}^{N-1} \left(\varphi_k \tau_k \lambda_k \langle \nabla F(\bar{x}^k), z^{k+1} - \widehat{x} \rangle_X + \frac{\varphi_k \tau_k \lambda_k^2 L}{2} \|z^{k+1} - z^k\|_X^2 \right) \\ &= \sum_{k=0}^{N-1} \varphi_k \tau_k \left[\lambda_k \langle \nabla F(\bar{x}^k), x^{k+1} - \widehat{x} \rangle_X + (1-\lambda_k) \langle \nabla F(\bar{x}^k), x^{k+1} - x^k \rangle_X + \frac{L}{2} \|x^{k+1} - \bar{x}^k\|_X^2 \right] \\ &\geq \sum_{k=0}^{N-1} \varphi_k \tau_k \left[\lambda_k (F(x^{k+1}) - F(\widehat{x})) + (1-\lambda_k)(F(x^{k+1}) - F(x^k)) \right] \\ &= \sum_{k=0}^{N-1} \left[\varphi_k \tau_k (F(x^{k+1}) - F(\widehat{x})) - \varphi_k \tau_k (1-\lambda_k)(F(x^k) - F(\widehat{x})) \right]. \end{aligned}$$

Summing with the estimate (12.31) for G, we deduce

$$s_N \geq \sum_{k=0}^{N-1} \left[\varphi_k \tau_k ((F+G)(x^{k+1}) - (F+G)(\widehat{x})) - \varphi_k \tau_k (1-\lambda_k)((F+G)(x^k) - (F+G)(\widehat{x})) \right].$$

Since $(F+G)(x^k) \geq (F+G)(\widehat{x})$, the recurrence inequality (12.28) together with a telescoping argument now gives the claim. □

Theorem 12.12. *Let $J := G + F$ for $G : X \to \overline{\mathbb{R}}$ and $F : X \to \mathbb{R}$ be convex, proper, and lower semicontinuous with ∇F Lipschitz. Suppose $[\partial J]^{-1}(0) \neq \emptyset$. Take $\tau > 0$ with $\tau L \leq 1$, and $\lambda_0 = 1$, and pick an initial iterate $\bar{x}^0 \in X$. Then the inertial explicit splitting (12.35) satisfies $J(x^N) \to \min_{x \in X} J(x)$ at the rate $O(1/N^2)$.*

Proof. The proof follows that of Theorem 12.10: in place of (12.34) we reduce (12.27) to the condition

$$\lambda_k \varphi_k \tau_k \langle \partial G(x^{k+1}) + \nabla F(\bar{x}^k), z^{k+1} - \widehat{z} \rangle_X \geq \mathcal{V}_{k+1}(\widehat{x}) - \frac{\lambda_k^2 \varphi_k}{2} \|z^{k+1} - z^k\|_X^2.$$

This is verified for some $\mathcal{V}_{k+1}(\widehat{x})$ such that

$$\sum_{k=0}^{N-1} \mathcal{V}_{k+1}(\widehat{x}) \geq \varphi_{N-1} \tau_{N-1} (J(x^N) - J(\widehat{x})) - \varphi_0 \tau_0 (1-\lambda_0)(J(x^0) - J(\widehat{x}))$$

by using Lemma 12.11 and the bound $\tau L \leq 1$ in place of Lemma 12.8. □

Remark 12.13 (accelerated gradient methods, FISTA). The inertial scheme was first introduced by [Nesterov, 1983] for the basic gradient descent method for smooth functions. The extension to explicit splitting is due to [Beck and Teboulle, 2009a], which proposed a *fast iterative shrinkage-thresholding algorithm* (FISTA) for the specific problem of minimizing a least-squares term plus a weighted ℓ^1 norm. (Note that in most treatments of FISTA, λ_k^{-1} is written as t_k.) We refer to [Beck, 2017; Nesterov, 2004] for a further discussion of these algorithms and more general accelerated gradient methods based on combinations of a history of iterates.

Remark 12.14 (PDPS, Douglas–Rachford, and correctors). The above unrolling arguments cannot be directly applied to PDPS, Douglas–Rachford splitting, and other methods based on (12.1) with non-maximally monotone H. Following [Chambolle and Pock, 2015], one can apply inertia to the PDPS method with the restricted choice $\alpha_k \in (0, 1/3)$. This prevents the use of the FISTA rule (12.32) and only yields $O(1/N)$ convergence of an ergodic gap. Based on alternative argumentation, when one of the functions is quadratic, [Patrinos et al., 2014] managed to employ the FISTA rule and obtain $O(1/N^2)$ rates for inertial Douglas–Rachford splitting. Moreover, [Valkonen, 2020a] observed that by introducing a *corrector* for the non-subdifferential component of H, in essence Ξ_{k+1}, the gap unrolling arguments can be performed. This approach also allows combining inertial acceleration with strong monotonicity based acceleration.

12.3 LINE SEARCH

Let us return to the basic results on weak convergence (Theorem 9.6), strong convergence with rates (Theorem 10.2), and function value convergence (Theorem 11.4) of the explicit splitting method. These results depend on the three-point inequalities of Corollary 7.2 (or, for faster rates under strong convexity, Corollary 7.7), specifically either the non-value estimate

$$\langle \nabla F(x^k) - \nabla F(\widehat{x}), x^{k+1} - \widehat{x} \rangle_X \geq -\frac{L}{4} \|x^{k+1} - x^k\|_X^2 \tag{12.38}$$

or the value estimate

$$\langle \nabla F(x^k), x^{k+1} - \widehat{x} \rangle_X \geq F(x^{k+1}) - F(\widehat{x}) - \frac{L}{2} \|x^{k+1} - x^k\|_X^2. \tag{12.39}$$

Recall that for weak convergence of iterates, we required the step length parameters $\{\tau_k\}_{k\in\mathbb{N}}$ to satisfy on each iteration the bound $\tau_k L < 2$. Under a strong convexity assumption, the bound $\tau_k L \leq 2$ was sufficient for strong convergence of iterates. Function value convergence was finally shown under the bound $\tau_k L \leq 1$. All cases thus hold for $\tau_k L \leq 1$, which we assume in the following for simplicity.

In this section, we address the following question: What if we do not know the Lipschitz factor L? A basic idea is to take L large enough. But what is large enough? Finding such a large enough L is the same as taking τ_k small enough and $L = 1/\tau_k$. This leads us to the following rough *line search* rule: for some $\tau > 0$ and line search parameter $\theta \in (0, 1)$, start with $\tau_k := \tau$, and iterate $\tau_k \mapsto \theta\tau_k$ until (12.39) (or (12.38)) is satisfied with $L = 1/\tau_k$. Note that on each update of τ_k, we need to recalculate $x^{k+1} := \operatorname{prox}_{\tau_k G}(x^k - \tau_k \nabla F(x^k)))$.

Performing this line search still appears to depend on knowing $\widehat{x}$ through (12.39). However, going back to the proof of Corollary 7.2, we see that what is really needed is to satisfy the smoothness (or descent) inequality (7.5) which was used to derive (12.39). We are therefore lead to the following practical line search method to guarantee the inequality

$$\langle \nabla F(x^k), x^{k+1} - x^k \rangle_X \geq F(x^{k+1}) - F(x^k) - \frac{1}{2\tau_k} \|x^{k+1} - x^k\|_X^2 \tag{12.40}$$

on every iteration:

0. Pick $\theta \in (0,1)$, $\tau > 0$, $\lambda_0 := 1$, $x^0 \in X$; set $k = 0$.
1. Set $\tau_k = \tau$.
2. Calculate $x^{k+1} := \operatorname{prox}_{\tau_k G}(x^k - \tau_k \nabla F(x^k))$.
3. If (12.40) does not hold, update $\tau_k := \theta \tau_k$, and go back to step 2.
4. Set $k := k+1$, and continue from step 1.

Theorem 12.15 (explicit splitting line search). *Let $J := F + G$ where $G : X \to \overline{\mathbb{R}}$ and $F : X \to \mathbb{R}$ are convex, proper, and lower semicontinuous, with ∇F moreover Lipschitz. Suppose $[\partial J]^{-1}(0) \neq \emptyset$. Then the above line search method satisfies $J(x^N) \to \min_{x \in X} J(x)$ at the rate $O(1/N)$. If G is strongly convex, then this convergence is linear.*

Proof. Since ∇F is $\tilde{L}$-smooth for some unknown $\tilde{L} > 0$, eventually the line search procedure satisfies $1/\tau_k \geq \tilde{L}$. Hence (12.40) is satisfied, and $\tau_k \geq \varepsilon > 0$ for some $\varepsilon > 0$. We can therefore follow through the proof of Theorem 11.4 with $L = 1/\tau_k$. □

We can also combine the line search method with the inertial explicit splitting (12.35). If in place of (12.40) we seek to satisfy

$$\langle \nabla F(\bar{x}^k), x^{k+1} - x^k \rangle_X \geq F(x^{k+1}) - F(x^k) - \frac{1}{2\tau_k} \|x^{k+1} - \bar{x}^k\|_X^2, \tag{12.41}$$

then also

$$\langle \nabla F(\bar{x}^k), x^{k+1} - \widehat{x} \rangle_X \geq F(x^{k+1}) - F(\widehat{x}) - \frac{1}{2\tau_k} \|x^{k+1} - \bar{x}^k\|_X^2.$$

This allows the inequality of (12.37) to be shown.

We are therefore lead to the following practical backtracking inertial explicit splitting:

0. Pick $\theta \in (0,1)$, $\tau > 0$, $\lambda_0 := 1$, $\bar{x}^0 = x^0 \in X$; set $k = 0$.
1. Set $\tau_k = \tau$.
2. Calculate $x^{k+1} := \operatorname{prox}_{\tau_k G}(\bar{x}^k - \tau_k \nabla F(\bar{x}^k)))$.
3. If (12.41) does not hold, update $\tau_k := \theta \tau_k$, and go back to step 2.
4. Set $\bar{x}^{k+1} := (1 + \alpha_{k+1}) x^{k+1} - \alpha_{k+1} x^k$ for $\alpha_{k+1} := \lambda_{k+1}(\lambda_k^{-1} - 1)$.
5. Set $k := k+1$, and continue from step 1.

The proof of the following is immediate:

Theorem 12.16. *Let $J \coloneqq G + F$ for $G : X \to \overline{\mathbb{R}}$ and $F : X \to \mathbb{R}$ be convex, proper, and lower semicontinuous with ∇F Lipschitz. Suppose $[\partial J]^{-1}(0) \neq \emptyset$. Take $\tau > 0$ and $\lambda_0 = 1$, and pick an initial iterate $\bar{x}^0 \in X$. Then the above backtracking inertial explicit splitting satisfies $J(x^N) \to \min_{x \in X} J(x)$ at the rate $O(1/N^2)$.*

The reader may now work out how to use line search to satisfy the nonnegativity of the metric $Z_{k+1}M_{k+1}$ in the PDPS method when $\|K\|$ is not known, or how to satisfy the condition $L\tau_0 + \tau_0\sigma_0\|K\|^2 < 1$ when the Lipschitz factor L of the forward step component E is not known.

Remark 12.17 (adaptive inertial parameters, quasi-Newton methods, and primal-dual proximal line searches). Regarding our statement in the beginning of the chapter about inertia methods attempting to construct a second-order approximation of the function, [Ochs and Pock, 2019] show that an adaptive inertial explicit splitting, performing an optimal line search on λ_k instead of τ_k, is equivalent to a proximal quasi-Newton method. Such a method is a further development of variants [see Beck and Teboulle, 2009b] of the method that attempt to restore the monotonicity of explicit splitting that is lost by inertia. Indeed, if $J(\bar{x}^{k+1}) \leq J(\bar{x}^k)$ does not hold for $\lambda_k < 1$, we can revert to $\lambda_k = 1$ to ensure descent as the step reduces to basic explicit splitting, which we know to be monotone by Theorem 11.4. Finally, a line search for the PDPS method is studied in [Malitsky and Pock, 2018].

Part III

NONCONVEX ANALYSIS

13 CLARKE SUBDIFFERENTIALS

We now turn to a concept of generalized derivatives that covers, among others, both Fréchet derivatives and convex subdifferentials. Again, we start with the general class of functionals that admit such a derivative. It is clear that we need to require some continuity properties, since otherwise there would be no relation between functional values at neighboring points and thus no hope of characterizing optimality through pointwise properties. In Part II, we used lower semicontinuity for this purpose, which together with convexity yielded the required properties. In this part, we want to drop the latter, global, assumption; in turn we need to strengthen the local continuity assumption. We thus consider now locally Lipschitz continuous functionals. Recall that $F : X \to \mathbb{R}$ is locally Lipschitz continuous near $x \in X$ if there exist a $\delta > 0$ and an $L > 0$ (which in the following will always denote the local Lipschitz constant of F) such that

$$|F(x_1) - F(x_2)| \leq L\|x_1 - x_2\|_X \qquad \text{for all } x_1, x_2 \in \mathbb{O}(x, \delta).$$

We will refer to the $\mathbb{O}(x, \delta)$ from the definition as the *Lipschitz neighborhood* of x. Note that for this we have to require that F is (locally) finite-valued (but see Remark 13.27 below). Throughout this chapter, we will assume that X is a Banach space unless stated otherwise.

13.1 DEFINITION AND BASIC PROPERTIES

We proceed as for the convex subdifferential and first define for $F : X \to \mathbb{R}$ the *generalized directional derivative* in $x \in X$ in direction $h \in X$ as

$$F^\circ(x; h) := \limsup_{\substack{y \to x \\ t \searrow 0}} \frac{F(y + th) - F(y)}{t}. \tag{13.1}$$

Note the difference to the classical directional derivative: We no longer require the existence of a limit but merely of accumulation points. We will need the following properties.

Lemma 13.1. *Let $F : X \to \mathbb{R}$ be locally Lipschitz continuous near $x \in X$ with the factor L. Then the mapping $h \mapsto F^\circ(x; h)$ is*

(i) Lipschitz continuous with constant L and satisfies $|F^\circ(x; h)| \leq L\|h\|_X < \infty$;

(ii) subadditive, *i.e.*, $F^\circ(x;h+g) \le F^\circ(x;h) + F^\circ(x;g)$ *for all* $h,g \in X$;

(iii) positively homogeneous, *i.e.*, $F^\circ(x;\alpha h) = (\alpha F)^\circ(x;h)$ *for all* $\alpha > 0$ *and* $h \in X$;

(iv) reflective, *i.e.*, $F^\circ(x;-h) = (-F)^\circ(x;h)$ *for all* $h \in X$.

Proof. *(i):* Let $h, g \in X$ be arbitrary. The local Lipschitz continuity of F implies that

$$F(y+th) - F(y) \le F(y+tg) - F(y) + tL\|h-g\|_X$$

for all y sufficiently close to x and t sufficiently small. Dividing by $t > 0$ and taking the lim sup then yields that

$$F^\circ(x;h) \le F^\circ(x;g) + L\|h-g\|_X.$$

Exchanging the roles of h and g shows the Lipschitz continuity of $F^\circ(x;\cdot)$, which also yields the claimed boundedness since $F^\circ(x;g) = 0$ for $g = 0$ from the definition.

(ii): Since $t \searrow 0$ and $g \in X$ is fixed, $y \to x$ if and only if $y + tg \to x$. The definition of the lim sup and the productive zero thus immediately yield

$$\begin{aligned} F^\circ(x;h+g) &= \limsup_{\substack{y\to x\\ t\searrow 0}} \frac{F(y+th+tg)-F(y)}{t} \\ &\le \limsup_{\substack{y\to x\\ t\searrow 0}} \frac{F(y+th+tg)-F(y+tg)}{t} + \limsup_{\substack{y\to x\\ t\searrow 0}} \frac{F(y+tg)-F(y)}{t} \\ &= F^\circ(x;h) + F^\circ(x;g). \end{aligned}$$

(iii): The claim is clear for $\alpha = 0$. For $\alpha > 0$, we obain again from the definition that

$$\begin{aligned} F^\circ(x;\alpha h) &= \limsup_{\substack{y\to x\\ t\searrow 0}} \frac{F(y+t(\alpha h))-F(y)}{t} \\ &= \limsup_{\substack{y\to x\\ \alpha t\searrow 0}} \alpha\frac{F(y+(\alpha t)h)-F(y)}{\alpha t} = (\alpha F)^\circ(x;h). \end{aligned}$$

(iv): Similarly, since $t \searrow 0$ and $h \in X$ is fixed, $y \to x$ if and only if $w := y - th \to x$. We thus have that

$$\begin{aligned} F^\circ(x;-h) &= \limsup_{\substack{y\to x\\ t\searrow 0}} \frac{F(y-th)-F(y)}{t} \\ &= \limsup_{\substack{w\to x\\ t\searrow 0}} \frac{-F(w+th)-(-F(w))}{t} = (-F)^\circ(x;h). \end{aligned}$$

□

In particular, Lemma 13.1 (i)–(iii) imply that the mapping $h \mapsto F^\circ(x;h)$ is proper, convex, and lower semicontinuous.

We now define for a locally Lipschitz continuous functional $F : X \to \mathbb{R}$ the *Clarke subdifferential* in $x \in X$ as

$$\partial_C F(x) := \{x^* \in X^* \mid \langle x^*, h\rangle_X \leq F^\circ(x;h) \text{ for all } h \in X\}. \tag{13.2}$$

The definition together with Lemma 13.1 (i) directly implies the following properties.

Lemma 13.2. *Let $F : X \to \mathbb{R}$ be locally Lipschitz continuous and $x \in X$. Then $\partial_C F(x)$ is convex, weakly-$*$ closed, and bounded. Specifically, if F is Lipschitz near x with constant L, then $\partial_C F(x) \subset \mathbb{B}(0, L)$.*

Furthermore, we have the following useful continuity property.

Lemma 13.3. *Let $F : X \to \mathbb{R}$. Then $\partial_C F(x)$ is strong-to-weak-$*$ outer semicontinuous, i.e., if $x_n \to x$ and if $\partial_C F(x_n) \ni x_n^* \overset{*}{\rightharpoonup} x^*$, then $x^* \in \partial_C F(x)$.*

Proof. Let $h \in X$ be arbitrary. By assumption, we then have that $\langle x_n^*, h\rangle_X \leq F^\circ(x_n;h)$ for all $n \in \mathbb{N}$. The weak-$*$ convergence of $\{x_n^*\}_{n\in\mathbb{N}}$ then implies that

$$\langle x^*, h\rangle_X = \lim_{n\to\infty} \langle x_n^*, h\rangle_X \leq \limsup_{n\to\infty} F^\circ(x_n;h).$$

Hence we are finished if we can show that $\limsup_{n\to\infty} F^\circ(x_n;h) \leq F^\circ(x;h)$ (since then $x^* \in \partial_C F(x)$ by definition).

For this, we use that by definition of $F^\circ(x_n;h)$, there exist sequences $\{y_{n,m}\}_{m\in\mathbb{N}}$ and $\{t_{n,m}\}_{m\in\mathbb{N}}$ with $y_{n,m} \to x_n$ and $t_{n,m} \searrow 0$ for $m \to \infty$ realizing the lim sup for each x_n. Hence, for all $n \in \mathbb{N}$ we can find a $y_n := y_{n,m(n)}$ and a $t_n := t_{n,m(n)}$ such that $\|y_n - x_n\|_X + t_n < n^{-1}$ (and hence in particular $y_n \to x$ and $t_n \searrow 0$) as well as

$$F^\circ(x_n;h) - \tfrac{1}{n} \leq \frac{F(y_n + t_n h) - F(y_n)}{t_n}$$

for n sufficiently large. Taking the lim sup for $n \to \infty$ on both sides yields the desired inequality. □

Again, the construction immediately yields a Fermat principle.[1]

[1] Similarly to Theorem 4.2, we do not need to require Lipschitz continuity of F – the Fermat principle for the Clarke subdifferential characterizes (among others) *any* local minimizer. However, if we want to use this principle to verify that a given $\bar{x} \in X$ is indeed a (candidate for) a minimizer, we need a suitable characterization of the subdifferential – and this is only possible for (certain) locally Lipschitz continuous functionals.

Theorem 13.4 (Fermat principle). *If $F : X \to \mathbb{R}$ has a local minimum in $\bar{x}$, then $0 \in \partial_C F(\bar{x})$.*

Proof. If $\bar{x} \in X$ is a local minimizer of F, then $F(\bar{x}) \leq F(\bar{x} + th)$ for all $h \in X$ and $t > 0$ sufficiently small (since the topological interior is always included in the algebraic interior). But this implies that

$$\langle 0, h \rangle_X = 0 \leq \liminf_{t \searrow 0} \frac{F(\bar{x} + th) - F(\bar{x})}{t} \leq \limsup_{t \searrow 0} \frac{F(\bar{x} + th) - F(\bar{x})}{t} \leq F^\circ(x; h)$$

and hence $0 \in \partial_C F(\bar{x})$ by definition. □

Note that F is not assumed to be convex, and hence the condition is in general not sufficient (consider, e.g., $f(t) = -|t|$).

13.2 FUNDAMENTAL EXAMPLES

Next, we show that the Clarke subdifferential is indeed a generalization of the derivative concepts we've studied so far.

Theorem 13.5. *Let $F : X \to \mathbb{R}$ be continuously Fréchet differentiable in a neighborhood U of $x \in X$. Then $\partial_C F(x) = \{F'(x)\}$.*

Proof. First, we note that F is locally Lipschitz continuous near x by Lemma 2.11. We now show that $F^\circ(x; h) = F'(x)h \ (= F'(x; h))$ for all $h \in X$. Take again sequences $\{y_n\}_{n \in \mathbb{N}}$ and $\{t_n\}_{n \in \mathbb{N}}$ with $y_n \to x$ and $t_n \searrow 0$ realizing the lim sup in (13.1). Applying the mean value Theorem 2.10 and using the continuity of F' yields for any $h \in X$ that

$$\begin{aligned} F^\circ(x; h) &= \lim_{n \to \infty} \frac{F(y_n + t_n h) - F(y_n)}{t_n} \\ &= \lim_{n \to \infty} \int_0^1 \frac{1}{t_n} \langle F'(y_n + s(t_n h)), t_n h \rangle_X \, ds \\ &= \langle F'(x), h \rangle_X \end{aligned}$$

since the integrand converges uniformly in $s \in [0, 1]$ to $\langle F'(x), h \rangle_X$. Hence by definition, $x^* \in \partial_C F(x)$ if and only if $\langle x^*, h \rangle_X \leq \langle F'(x), h \rangle_X$ for all $h \in X$, which is only possible for $x^* = F'(x)$. □

The following example shows that Theorem 13.5 does *not* hold if F is merely Fréchet differentiable.

Example 13.6. Let $F : \mathbb{R} \to \mathbb{R}$, $F(x) = x^2 \sin(x^{-1})$. Then it is straightforward (if tedious) to show that F is differentiable on $\mathbb{R}$ with

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

In particular, F is not continuously differentiable at $x = 0$. But a similar limit argument shows that for all $h \in \mathbb{R}$,

$$F^\circ(0; h) = |h|$$

and hence that

$$\partial_C F(0) = [-1, 1] \supsetneq \{0\} = \{F'(0)\}.$$

(The first equality also follows more directly from Theorem 13.26 below.)

As the example suggests, we always have the following weaker relation.

Lemma 13.7. *Let $F : X \to \mathbb{R}$ be Lipschitz continuous and Gâteaux differentiable in a neighborhood U of $x \in X$. Then $DF(x) \in \partial_C F(x)$.*

Proof. Let $h \in X$ be arbitrary. First, note that we always have that

$$F'(x; h) = \lim_{t \searrow 0} \frac{F(x + th) - F(x)}{t} \leq \limsup_{\substack{y \to x \\ t \searrow 0}} \frac{F(y + th) - F(y)}{t} = F^\circ(x; h). \tag{13.3}$$

Since F is Gâteaux differentiable, it follows that

$$\langle DF(x), h \rangle_X = F'(x; h) \leq F^\circ(x; h) \qquad \text{for all } h \in X,$$

and thus $DF(x) \in \partial_C F(x)$ by definition. □

Similarly, the Clarke subdifferential reduces to the convex subdifferential in some situations.

Theorem 13.8. *Let $F : X \to \overline{\mathbb{R}}$ be convex and lower semicontinuous. Then $\partial_C F(x) = \partial F(x)$ for all $x \in \operatorname{int}(\operatorname{dom} F)$.*

Proof. By Theorem 3.13, F is locally Lipschitz continuous near $x \in \operatorname{int}(\operatorname{dom} F)$. We now show that $F^\circ(x; h) = F'(x; h)$ for all $h \in X$, which together with Lemma 4.4 yields the claim. By (13.3), we always have that $F'(x; h) \leq F^\circ(x; h)$. To show the reverse inequality,

let $\delta > 0$ be arbitrary. Since the difference quotient of convex functionals is increasing by Lemma 4.3 (i), we obtain that

$$\begin{aligned} F^\circ(x;h) &= \lim_{\varepsilon \searrow 0} \sup_{y \in \mathbb{B}(x,\delta\varepsilon)} \sup_{0<t<\varepsilon} \frac{F(y+th) - F(y)}{t} \\ &\leq \lim_{\varepsilon \searrow 0} \sup_{y \in \mathbb{B}(x,\delta\varepsilon)} \frac{F(y+\varepsilon h) - F(y)}{\varepsilon} \\ &\leq \lim_{\varepsilon \searrow 0} \frac{F(x+\varepsilon h) - F(x)}{\varepsilon} + 2L\delta \\ &= F'(x;h) + 2L\delta, \end{aligned}$$

where the last inequality follows by adding two productive zeros and using the local Lipschitz continuity in x. Since $\delta > 0$ was arbitrary, this implies that $F^\circ(x;h) \leq F'(x;h)$, and the claim follows. □

A locally Lipschitz continuous functional $F : X \to \mathbb{R}$ with $F^\circ(x;h) = F'(x;h)$ for all $h \in X$ is called *regular* in $x \in X$. We have just shown that every continuously differentiable and every convex and lower semicontinuous functional is regular; intuitively, a function is thus regular at any points in which it is either differentiable or has at most a "convex kink".

Finally, similarly to Theorem 4.11 one can show the following pointwise characterization of the Clarke subdifferential of integral functionals with Lipschitz continuous integrands. We again assume that $\Omega \subset \mathbb{R}^d$ is open and bounded.

Theorem 13.9. *Let $f : \mathbb{R} \to \mathbb{R}$ be Lipschitz continuous and $F : L^p(\Omega) \to \overline{\mathbb{R}}$ with $1 \leq p < \infty$ as in Lemma 3.7. Then we have for all $u \in L^p(\Omega)$ with $q = \frac{p}{p-1}$ (where $q = \infty$ for $p = 1$) that*

$$\partial_C F(u) \subset \{u^* \in L^q(\Omega) \mid u^*(x) \in \partial_C f(u(x)) \text{ for almost every } x \in \Omega\}.$$

If f is regular at $u(x)$ for almost every $x \in \Omega$, then F is regular at u, and equality holds.

Proof. First, by the properties of the Lebesgue integral and the Lipschitz continuity of f, we have for any $u, v \in L^p(\Omega)$ that

$$|F(u) - F(v)| \leq \int_\Omega |f(u(x)) - f(v(x))|\,dx \leq L \int_\Omega |u(x) - v(x)|\,dx \leq LC_p \|u - v\|_{L^p},$$

where L is the Lipschitz constant of f and C_p the constant from the continuous embedding $L^p(\Omega) \hookrightarrow L^1(\Omega)$ for $1 \leq p \leq \infty$. Hence $F : L^p(\Omega) \to \mathbb{R}$ is Lipschitz continuous and therefore finite-valued as well.

Let now $\xi \in \partial_C F(u) \subset L^p(\Omega)^*$ be given and $h \in L^p(\Omega)$ be arbitrary. By definition, we thus have

$$\begin{aligned} \langle \xi, h \rangle_{L^p} \le F^\circ(u;h) &= \limsup_{\substack{v \to u \\ t \searrow 0}} \frac{F(v+th) - F(v)}{t} \\ &\le \int_\Omega \limsup_{\substack{v \to u \\ t \searrow 0}} \frac{f(v(x) + th(x)) - f(v(x))}{t}\, dx \\ &\le \int_\Omega \limsup_{\substack{v_x \to u(x) \\ t_x \searrow 0}} \frac{f(v_x + t_x h(x)) - f(v_x)}{t_x}\, dx \\ &= \int_\Omega f^\circ(u(x); h(x))\, dx, \end{aligned} \tag{13.4}$$

where we were able to use the Reverse Fatou Lemma to exchange the lim sup with the integral in the first inequality since the integrand is bounded from above by the integrable function $L|h|$ due to Lemma 13.1 (i); the second inequality follows by bounding for almost every $x \in \Omega$ the (pointwise) limit over the sequences realizing the lim sup in the second line by the lim sup over all admissible sequences.

In order to interpret (13.4) pointwise, we use that Lemma 13.1 (i) together with the (global) Lipschitz continuity of f implies that the function $x \mapsto f^\circ(u(x);t)$ is integrable for any $t \in \mathbb{R}$. We can thus argue exactly as in the proof of Theorem 4.11: Let $t \in \mathbb{R}$ be arbitrary and $A \subset \Omega$ be an arbitrary measurable subset. Setting

$$h(x) = \begin{cases} t & \text{if } x \in A, \\ 0 & \text{if } x \notin A, \end{cases}$$

(so that $h \in L^\infty(\Omega) \subset L^p(\Omega)$) and using $f^\circ(u(x);0) = 0$, we obtain from (13.4) together with the representation of $\xi \in L^p(\Omega)^*$ via some $u^* \in L^q(\Omega)$ that

$$\int_A u^*(x) t\, dx = \langle \xi, h \rangle_{L^p} \le \int_\Omega f^\circ(u(x); h(x))\, dx = \int_A f^\circ(u(x); t)\, dx.$$

Since A was arbitrary, this implies that

$$u^*(x) t \le f^\circ(u(x);t) \qquad \text{for almost every } x \in \Omega.$$

Since $t \in \mathbb{R}$ was arbitrary, we obtain $u^*(x) \in \partial_C f(u(x))$ almost everywhere.

It remains to show the remaining assertions when f is regular. In this case, it follows from (13.4) that for any $h \in L^p(\Omega)$,

$$\begin{aligned} F^\circ(u;h) &\le \int_\Omega f^\circ(u(x); h(x))\, dx = \int_\Omega f'(u(x); h(x))\, dx \\ &\le \lim_{t \searrow 0} \frac{F(u+th) - F(u)}{t} = F'(u;h) \le F^\circ(u;h), \end{aligned} \tag{13.5}$$

where the second inequality is obtained by applying Fatou's Lemma, this time appealing to the integrable lower bound $-L|h(x)|$. This shows that $F'(u;h) = F^\circ(u;h)$ and hence that F is regular. We further obtain for any $u^* \in L^q(\Omega)$ with $u^*(x) \in \partial_C f(u(x))$ almost everywhere and any $h \in L^p(\Omega)$, that

$$\langle u^*, h\rangle_{L^p} = \int_\Omega u^*(x)h(x)\,dx \le \int_\Omega f^\circ(u(x);h(x))\,dx \le F^\circ(u,h),$$

where we have used (13.5) in the last inequality. Since $h \in L^p(\Omega)$ was arbitrary, this implies that $u^* \in \partial_C F(u)$. □

Under additional assumptions similar to those of Theorem 2.14 and with more technical arguments, this result can be extended to spatially varying integrands $f : \Omega \times \mathbb{R} \to \mathbb{R}$; see, e.g., [Clarke, 1990, Theorem 2.7.5].

13.3 CALCULUS RULES

We now turn to calculus rules. The first one follows directly from the definition.

Theorem 13.10. *Let $F : X \to \mathbb{R}$ be locally Lipschitz continuous near $x \in X$ and $\alpha \in \mathbb{R}$. Then,*

$$\partial_C(\alpha F)(x) = \alpha\partial_C(F)(x).$$

Proof. First, αF is clearly locally Lipschitz continuous near x for any $\alpha \in \mathbb{R}$. If $\alpha = 0$, both sides of the claimed equality are zero (which is easiest seen from Theorem 13.5). If $\alpha > 0$, we have that $(\alpha F)^\circ(x;h) = \alpha F^\circ(x;h)$ for all $h \in X$ from the definition. Hence,

$$\begin{aligned}
\alpha\partial_C F(x) &= \{\alpha x^* \in X^* \mid \langle x^*, h\rangle_X \le F^\circ(x;h) \quad \text{for all } h \in X\} \\
&= \{\alpha x^* \in X^* \mid \langle \alpha x^*, h\rangle_X \le \alpha F^\circ(x;h) \quad \text{for all } h \in X\} \\
&= \{y^* \in X^* \mid \langle y^*, h\rangle_X \le (\alpha F)^\circ(x;h) \quad \text{for all } h \in X\} \\
&= \partial_C(\alpha F)(x).
\end{aligned}$$

To conclude the proof, it suffices to show the claim for $\alpha = -1$. For that, we use Lemma 13.1 (iv) to obtain that

$$\begin{aligned}
\partial_C(-F)(x) &= \{x^* \in X^* \mid \langle x^*, h\rangle_X \le (-F)^\circ(x;h) \quad \text{for all } h \in X\} \\
&= \{x^* \in X^* \mid \langle -x^*, -h\rangle_X \le F^\circ(x;-h) \quad \text{for all } h \in X\} \\
&= \{-y^* \in X^* \mid \langle y^*, g\rangle_X \le F^\circ(x;g) \quad \text{for all } g \in X\} \\
&= -\partial_C F(x).
\end{aligned}$$

□

Corollary 13.11. *Let $F : X \to \mathbb{R}$ be locally Lipschitz continuous near $\bar{x} \in X$. If F has a local maximum in $\bar{x}$, then $0 \in \partial_C F(\bar{x})$.*

Proof. If $\bar{x}$ is a local maximizer of F, it is a local minimizer of $-F$. Hence, Theorems 13.4 and 13.10 imply that

$$0 \in \partial_C(-F)(\bar{x}) = -\partial_C F(\bar{x}),$$

i.e., $0 = -0 \in \partial_C F(\bar{x})$. □

SUPPORT FUNCTIONALS

The remaining rules are significantly more involved. As in the previous proofs, a key step is to relate different sets of the form (13.2), which we will do with the help of the following lemmas due to [Hörmander, 1955].

Lemma 13.12. *Let $S : X \to \mathbb{R}$ be positively homogeneous, subadditive, and lower semicontinuous, and let*

$$A = \{x^* \in X^* \mid \langle x^*, x\rangle_X \leq S(x) \quad \text{for all } x \in X\}.$$

Then

$$S(x) = \sup_{x^* \in A} \langle x^*, x\rangle_X \qquad \text{for all } x \in X. \tag{13.6}$$

Proof. By definition of A, the inequality $\langle x^*, x\rangle_X - S(x) \leq 0$ holds for all $x \in X$ if and only if $x^* \in A$. Thus a case distinction as in Example 5.3 (ii) using the positive homogeneity of S (which in particular implies that $S(0) = 0$) shows that

$$S^*(x^*) = \sup_{x \in X} \langle x^*, x\rangle_X - S(x) = \begin{cases} 0 & x^* \in A, \\ \infty & x^* \notin A, \end{cases}$$

i.e., $S^* = \delta_A$. Furthermore, by assumption S is also subadditive and hence convex as well as lower semicontinuous; it is also proper. Theorem 5.1 thus yields

$$S(x) = S^{**}(x) = (\delta_A)_*(x) = \sup_{x^* \in A} \langle x^*, x\rangle_X. \tag{13.7}$$ □

The right-hand side of (13.6) is called the *support functional* of $A \subset X^*$; see, e.g., [Hiriart-Urruty and Lemaréchal, 2001] for their use in convex analysis (in finite dimensions). Note that (13.7) implies that any set of the form A is nonempty since the supremum over the empty set is $-\infty$ and S was assumed to be real-valued.

Lemma 13.13. *Let $A, B \subset X^*$ be nonempty, convex, and weakly-$*$ closed. Then $A \subset B$ if and only if*

$$\sup_{x^* \in A} \langle x^*, x\rangle_X \leq \sup_{x^* \in B} \langle x^*, x\rangle_X \qquad \text{for all } x \in X. \tag{13.8}$$

Proof. If $A \subset B$, then the right-hand side of (13.8) is obviously not less than the left-hand side. Conversely, assume that there exists an $x^* \in A$ with $x^* \notin B$. By the assumptions on A and B, we then obtain from Theorem 1.13 an $x \in X$ and a $\lambda \in \mathbb{R}$ with

$$\langle z^*, x\rangle_X \leq \lambda < \langle x^*, x\rangle_X \qquad \text{for all } z^* \in B.$$

Taking the supremum over all $z^* \in B$ and estimating the right-hand side by the supremum over all $x^* \in A$ then yields that

$$\sup_{z^* \in B} \langle z^*, x\rangle_X < \sup_{x^* \in A} \langle x^*, x\rangle_X.$$

Hence (13.8) is violated, and the claim follows by contraposition. □

Corollary 13.14. *Let $A, B \subset X^*$ be nonempty, convex, and weakly-$*$ closed. Then $A = B$ if and only if*

$$\sup_{x^* \in A} \langle x^*, x\rangle_X = \sup_{x^* \in B} \langle x^*, x\rangle_X \qquad \text{for all } x \in X. \tag{13.9}$$

Proof. Again, the claim is obvious if $A = B$. Conversely, if (13.9) holds, then in particular (13.8) holds, and we obtain from Lemma 13.13 that $A \subset B$. Exchanging the roles of A and B now yields the claim. □

Since generalized directional derivatives are real-valued, Lemma 13.12 together with Lemma 13.1 directly yields the following useful representation.

Corollary 13.15. *Let $F : X \to \mathbb{R}$ be locally Lipschitz continuous and $x \in X$. Then*

$$F^\circ(x; h) = \sup_{x^* \in \partial_C F(x)} \langle x^*, h\rangle_X \quad \text{for all } h \in X.$$

In particular, $\partial_C F(x)$ is nonempty.

For example, this implies a converse result to Theorem 13.5.

Corollary 13.16. *Let $F : X \to \mathbb{R}$ be locally Lipschitz continuous near x. If $\partial_C F(x) = \{x^*\}$ for some $x^* \in X^*$, then F is Gâteaux differentiable in x with $DF(x) = x^*$.*

Proof. Under the assumption, it follows from Corollary 13.15 that

$$F^\circ(x; h) = \sup_{\tilde{x}^* \in \partial F_C(x)} \langle \tilde{x}^*, h\rangle_X = \langle x^*, h\rangle_X$$

for all $h \in X$. In particular, $F^\circ(x;h)$ is linear (and not just reflective) in h. It thus follows from Lemma 13.1 (iv) that for any $h \in X$,

$$\begin{aligned}\liminf_{\substack{y\to x\\ t\searrow 0}} \frac{F(y+th)-F(y)}{t} &= -\limsup_{\substack{y\to x\\ t\searrow 0}} \frac{-F(y+th)-(-F(y))}{t}\\ &= -(-F)^\circ(x;h) = -F^\circ(x;-h) = F^\circ(x,h)\\ &= \limsup_{\substack{y\to x\\ t\searrow 0}} \frac{F(y+th)-F(y)}{t}.\end{aligned}$$

Hence the lim sup is a proper limit, and thus $F^\circ(x;h) = F'(x;h)$; i.e., F is regular in x. This shows that $F'(x;h)$ is linear and bounded in h, and hence x^* is by definition the Gâteaux derivative. □

It is not hard to verify from the definition and the Lipschitz continuity of F that in this case, x^* is in fact a Fréchet derivative.

We can also use this to show the promised nonemptiness of the convex subdifferential.

Theorem 13.17. *Let X be a Banach space and let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $x \in \operatorname{int}(\operatorname{dom} F)$. Then $\partial F(x)$ is nonempty, convex, weakly-$*$ closed, and bounded.*

Proof. Since $x \in (\operatorname{dom} F)^o$, Theorem 13.8 shows that $\partial F(x) = \partial_C F(x) \neq \emptyset$ by Corollary 13.15. The remaining properties follow similarly from Lemma 13.2. □

By a similar argument, we now obtain the promised converse of Theorem 4.5; we combine both statements here for the sake of reference.

Theorem 13.18. *Let X be a Banach space and let $F : X \to \overline{\mathbb{R}}$ be convex. If F is Gâteaux differentiable at x, then $\partial F(x) = \{DF(x)\}$. Conversely, if $x \in \operatorname{int}(\operatorname{dom} F)$ and $\partial F(x) = \{x^*\}$ is a singleton, then F is Gâteaux differentiable at x with $DF(x) = x^*$.*

Proof. The first claim was already shown in Theorem 4.5, while the second follows from Corollary 13.16 together with Theorem 13.8. □

As another consequence, we can show that Moreau–Yosida regularization defined in Section 7.3 preserves (global!) Lipschitz continuity.

Lemma 13.19. *Let X be a Hilbert space and let $F : X \to \mathbb{R}$ be Lipschitz continuous with constant L. Then F_γ is Lipschitz continuous with constant L as well. If F is in addition convex, then $F - \frac{\gamma L^2}{2} \leq F_\gamma \leq F$.*

Proof. Let $x, z \in X$. We expand

$$F_\gamma(x) - F_\gamma(z) = \sup_{y_z \in X} \inf_{y_x \in X} \left(F(y_x) - F(y_z) + \frac{1}{2\gamma}\|y_x - x\|_X^2 - \frac{1}{2\gamma}\|y_z - z\|_X^2 \right).$$

Taking $y_x = y_z + x - z$, we estimate

$$F_\gamma(x) - F_\gamma(z) \leq \sup_{y_z \in X} \left(F(y_z + x - z) - F(y_z) \right) \leq L\|x - z\|_X.$$

Exchanging x and y, we obtain the first claim.

For the second claim, we first observe that by assumption $\operatorname{dom} F = X$. Hence by Theorem 13.17 and Lemma 13.2, for every $x \in X$, there exists some $x^* \in \partial F(x)$ with $\|x^*\|_{X^*} \leq L$. Thus, using Lemma 4.4, for any $x^* \in \partial F(x)$,

$$F_\gamma(x) = \inf_{y \in X} F(z) + \frac{1}{2\gamma}\|x - y\|_X^2 \geq F(x) + \langle x^*, z - x \rangle_X + \frac{1}{2\gamma}\|x - z\|_X^2.$$

The Cauchy–Schwarz and generalized Young's inequality then yield $F_\gamma(x) \geq F(x) - \frac{\gamma}{2}\|x^*\|_{X^*}^2 \geq F(x) - \frac{\gamma}{2}L^2$. The second inequality follows by estimating the infimum in (7.19) by $z = x$. □

SUM RULE

With the aid of these results on support functionals, we can now show a sum rule.

Theorem 13.20. *Let $F, G : X \to \mathbb{R}$ be locally Lipschitz continuous near $x \in X$. Then,*

$$\partial_C(F + G)(x) \subset \partial_C F(x) + \partial_C G(x).$$

If F and G are regular at x, then $F + G$ is regular at x and equality holds.

Proof. It is clear that $F + G$ is locally Lipschitz continuous near x. Furthermore, from the properties of the lim sup we always have for all $h \in X$ that

$$(F + G)^\circ(x; h) \leq F^\circ(x; h) + G^\circ(x; h). \tag{13.10}$$

If F and G are regular at x, the calculus of limits yields that

$$F^\circ(x; h) + G^\circ(x; h) = F'(x; h) + G'(x; h) = (F + G)'(x; h) \leq (F + G)^\circ(x; h),$$

which implies that $(F + G)^\circ(x; h) = (F + G)'(x; h)$, i.e., $F + G$ is regular.

By the definition of the Clarke subdifferential, it follows from (13.10)

$$\partial_C(F + G)(x) \subset \{x^* \in X^* \mid \langle x^*, h \rangle_X \leq F^\circ(x; h) + G^\circ(x; h) \text{ for all } h \in X\} =: A$$

(with equality if F and G are regular); it thus remains to show that $A = \partial_C F(x) + \partial_C G(x)$. For this, we use that $\partial_C F(x)$ and $\partial_C G(x)$ are convex and weakly-$*$ closed by Lemma 13.2 and nonempty by Corollary 13.15, and hence so is their sum since both sets are bounded. Furthermore, as shown in Lemma 13.1, generalized directional derivatives and hence their sums are real-valued, positively homogeneous, convex, and lower semicontinuous. We thus obtain from Lemma 13.12 for all $h \in X$ that

$$\begin{aligned}\sup_{x^* \in \partial_C F(x) + \partial_C G(x)} \langle x^*, h\rangle_X &= \sup_{x_1^* \in \partial_C F(x)} \langle x_1^*, h\rangle_X + \sup_{x_2^* \in \partial_C G(x)} \langle x_2^*, h\rangle_X \\ &= F^\circ(x;h) + G^\circ(x;h) = \sup_{x^* \in A} \langle x^*, h\rangle_X.\end{aligned}$$

The claimed equality of A (which is nonempty, convex, and weakly-$*$ closed as well) and the sum of the subdifferentials now follows from Corollary 13.14. □

Note the differences to the convex sum rule: The generic inclusion is now in the other direction; furthermore, *both* functionals have to be regular, and in exactly the point where the sum rule is applied. By induction, one obtains from this sum rule for an arbitrary number of functionals (which all have to be regular).

CHAIN RULE

To prove a chain rule, we need the following "nonsmooth" mean value theorem due to [Lebourg, 1975, 1979].

Theorem 13.21. *Let $F : X \to \mathbb{R}$ be locally Lipschitz continuous near $x \in X$ and $\tilde{x}$ be in the Lipschitz neighborhood of x. Then there exists a $\lambda \in (0,1)$ and an $x^* \in \partial_C F(x + \lambda(\tilde{x} - x))$ such that*

$$F(\tilde{x}) - F(x) = \langle x^*, \tilde{x} - x\rangle_X.$$

Proof. Define $\psi, \varphi : [0,1] \to \mathbb{R}$ as

$$\psi(\lambda) := F(x + \lambda(\tilde{x} - x)), \qquad \varphi(\lambda) := \psi(\lambda) + \lambda(F(x) - F(\tilde{x})).$$

By the assumptions on F and $\tilde{x}$, both ψ and φ are Lipschitz continuous. In addition, $\varphi(0) = F(x) = \varphi(1)$, and hence φ has a local minimum or maximum in an interior point $\bar{\lambda} \in (0,1)$. From the Fermat principle Theorem 13.4 or Corollary 13.11, respectively, together with the sum rule from Theorem 13.20 and the characterization of the subdifferential of the second term from Theorem 13.5, we thus obtain that

$$0 \in \partial_C \varphi(\bar{\lambda}) \subset \partial_C \psi(\bar{\lambda}) + \{F(x) - F(\tilde{x})\}.$$

Hence we are finished if we can show for $x_{\bar{\lambda}} := x + \bar{\lambda}(\tilde{x} - x)$ that

$$\partial_C \psi(\bar{\lambda}) \subset \left\{ \langle x^*, \tilde{x} - x\rangle_X \;\middle|\; x^* \in \partial_C F(x_{\bar{\lambda}}) \right\} =: A. \tag{13.11}$$

For this purpose, consider for arbitrary $s \in \mathbb{R}$ the generalized directional derivative

$$\begin{aligned}\psi^\circ(\bar\lambda; s) &= \limsup_{\substack{\lambda \to \bar\lambda \\ t \searrow 0}} \frac{\psi(\lambda + ts) - \psi(\lambda)}{t} \\ &= \limsup_{\substack{\lambda \to \bar\lambda \\ t \searrow 0}} \frac{F(x + (\lambda + ts)(\tilde x - x)) - F(x + \lambda(\tilde x - x))}{t} \\ &\le \limsup_{\substack{z \to x_{\bar\lambda} \\ t \searrow 0}} \frac{F(z + ts(\tilde x - x)) - F(z)}{t} = F^\circ(x_{\bar\lambda}; s(\tilde x - x)),\end{aligned}$$

where the inequality follows from considering arbitrary sequences $z \to x_{\bar\lambda}$ (instead of special sequences of the form $z_n = x + \lambda_n(\tilde x - x)$) in the last lim sup. Again, the definition of the Clarke subdifferential thus implies that

$$\partial_C \psi(\bar\lambda) \subset \left\{ t^* \in \mathbb{R} \;\middle|\; t^* s \le F^\circ(x_{\bar\lambda}; s(\tilde x - x)) \text{ for all } s \in \mathbb{R} \right\} =: B. \tag{13.12}$$

It remains to show that the sets A and B from (13.11) and (13.12) coincide. But this follows again from Lemma 13.12 and Corollary 13.14, since for all $s \in \mathbb{R}$ we have that

$$\sup_{t^* \in A} t^* s = \sup_{x^* \in \partial_C F(x_{\bar\lambda})} \langle x^*, s(\tilde x - x)\rangle_X = F^\circ(x_{\bar\lambda}; s(\tilde x - x)) = \sup_{t^* \in B} t^* s. \qquad \square$$

We also need the following generalization of the argument in Theorem 13.5.

Lemma 13.22. *Let X, Y be Banach spaces and $F : X \to Y$ be continuously Fréchet differentiable at $x \in X$. Let $\{x_n\}_{n\in\mathbb{N}} \subset X$ be a sequence with $x_n \to x$ and $\{t_n\}_{n\in\mathbb{N}} \subset (0, \infty)$ be a sequence with $t_n \searrow 0$. Then for any $h \in X$,*

$$\lim_{n\to\infty} \frac{F(x_n + t_n h) - F(x_n)}{t_n} = F'(x)h.$$

Proof. Let $h \in X$ be arbitrary. By the Hahn–Banach extension Theorem 1.4, for every $n \in \mathbb{N}$ there exists a $y_n^* \in Y^*$ with $\|y_n^*\|_{Y^*} = 1$ and

$$\|t_n^{-1}(F(x_n + t_n h) - F(x_n)) - F'(x)h\|_Y = \langle y_n^*, t_n^{-1}(F(x_n + t_n h) - F(x_n)) - F'(x)h\rangle_Y.$$

Applying now the classical mean value theorem to the scalar functions

$$f_n : [0,1] \to \mathbb{R}, \qquad f_n(s) = \langle y_n^*, F(x_n + s t_n h)\rangle_Y,$$

we obtain similarly to the proof of Theorem 2.10 for all $n \in \mathbb{N}$ that

$$\begin{aligned}\|t_n^{-1}(F(x_n + t_n h) - F(x_n)) - F'(x)h\|_Y &= t_n^{-1} \int_0^1 \langle y_n^*, F'(x_n + s t_n h) t_n h\rangle_Y \, ds - \langle y_n^*, F'(x)h\rangle_Y \\ &= \int_0^1 \langle y_n^*, [F'(x_n + s t_n h) - F'(x)]h\rangle_Y \, ds \\ &\le \int_0^1 \|F'(x_n + s t_n h) - F'(x))\|_{\mathbb{L}(X;Y)} \, ds \, \|h\|_X,\end{aligned}$$

where we have used (1.1) together with $\|y_n^*\|_{Y^*} = 1$ in the last step. Since F' is continuous by assumption, the integrand goes to zero as $n \to \infty$ uniformly in $s \in [0,1]$, and the claim follows. □

We now come to the chain rule, which in contrast to the convex case does not require the inner mapping to be linear; this is one of the main advantages of the Clarke subdifferential in the context of nonsmooth optimization.

Theorem 13.23. *Let Y be a separable Banach space, $F : X \to Y$ be continuously Fréchet differentiable at $x \in X$, and $G : Y \to \mathbb{R}$ be locally Lipschitz continuous near $F(x)$. Then,*

$$\partial_C(G \circ F)(x) \subset F'(x)^* \partial_C G(F(x)) := \{F'(x)^* y^* \mid y^* \in \partial_C G(F(x))\}.$$

If G is regular at $F(x)$, then $G \circ F$ is regular at x, and equality holds.

Proof. The local Lipschitz continuity of $G \circ F$ follows from that of G and F (which in turn follows from Lemma 2.11). For the claimed inclusion (respectively, equality), we argue as before using the support calculus. First we show that for every $h \in X$ there exists a $y^* \in \partial_C G(F(x))$ with

$$(G \circ F)^\circ(x; h) = \langle y^*, F'(x)h \rangle_Y. \tag{13.13}$$

To this end, consider for given $h \in X$ sequences $\{x_n\}_{n\in\mathbb{N}} \subset X$ and $\{t_n\}_{n\in\mathbb{N}} \subset (0,\infty)$ with $x_n \to x$, $t_n \searrow 0$, and

$$(G \circ F)^\circ(x; h) = \lim_{n\to\infty} \frac{G(F(x_n + t_n h)) - G(F(x_n))}{t_n}.$$

Let us now write $U_{F(x)}$ for the neighborhood of $F(x)$ where G is Lipschitz with factor L. By continuity of F, we can then find $n_0 \in \mathbb{N}$ such that $F(x_n), F(x_n + t_n h) \in U_{F(x)}$ for all $n \geq n_0$. Theorem 13.21 thus yields for all $n \geq n_0$ a $y_n^* \in \partial_C G(y_n)$ with $y_n := F(x_n) + \lambda_n(F(x_n + t_n h) - F(x_n))$ for some $\lambda_n \in (0,1)$ such that

$$\frac{G(F(x_n + t_n h)) - G(F(x_n))}{t_n} = \langle y_n^*, q_n \rangle_Y \quad \text{with} \quad q_n := \frac{F(x_n + t_n h) - F(x_n)}{t_n} \tag{13.14}$$

Since $\lambda_n \in (0,1)$ is uniformly bounded, we also have that $y_n \to F(x)$ for $n \to \infty$. Hence y_n is in the Lipschitz neighborhood of $F(x)$ for $n \in \mathbb{N}$ large enough, and Lemma 13.2 yields that $y_n^* \in \partial_C G(y_n) \subset \mathbb{B}(0, L)$ for $n \in \mathbb{N}$ sufficiently large. This implies that $\{y_n^*\}_{n\in\mathbb{N}} \subset Y^*$ is bounded, and the Banach–Alaoglu Theorem 1.11 yields a weakly-$*$ convergent subsequence with limit $y^* \in \partial_C G(F(x))$ by Lemma 13.3. Finally, since F is continuously Fréchet differentiable, $q_n \to F'(x)h$ strongly in Y by Lemma 13.22. Hence, $\langle y_n^*, q_n \rangle_Y \to \langle y^*, F'(x)h \rangle$ as the duality pairing of weakly-$*$ and strongly converging sequences. Passing to the limit in (13.14) therefore yields (13.13) (first along the subsequence chosen above; by convergence

of the left-hand side of (13.14) and the uniqueness of limits then for the full sequence as well). By definition of the Clarke subdifferential, we thus have for $y^* \in \partial_C G(F(x))$ that

$$(G \circ F)^\circ(x;h) = \langle y^*, F'(x)h\rangle_Y \leq G^\circ(F(x);F'(x)h). \tag{13.15}$$

If G is now regular at x, we have that $G^\circ(F(x);F'(x)h) = G'(F(x);F'(x)h)$ and hence by the local Lipschitz continuity of G and the Fréchet differentiability of F that

$$\begin{aligned}
&G^\circ(F(x);F'(x)h)\\
&\qquad = \lim_{t\searrow 0} \frac{G(F(x)+tF'(x)h) - G(F(x))}{t}\\
&\qquad = \lim_{t\searrow 0} \frac{G(F(x)+tF'(x)h) - G(F(x+th)) + G(F(x+th)) - G(F(x))}{t}\\
&\qquad \leq \lim_{t\searrow 0} \left(L\|h\|_X \frac{\|F(x)+F'(x)th - F(x+th)\|_Y}{\|th\|_X} + \frac{G(F(x+th)) - G(F(x))}{t} \right)\\
&\qquad = (G\circ F)'(x;h) \leq (G\circ F)^\circ(x;h).
\end{aligned}$$

Together with (13.15), this implies that $(G \circ F)'(x;h) = (G \circ F)^\circ(x;h)$ (i.e., $G \circ F$ is regular at x) and that

$$(G \circ F)^\circ(x;h) = G^\circ(F(x);F'(x)h). \tag{13.16}$$

As before, Lemma 13.12 now implies for all $h \in X$ that

$$\sup_{x^* \in F'(x)^* \partial_C G(F(x))} \langle x^*, h\rangle_X = \sup_{y^* \in \partial_C G(F(x))} \langle y^*, F'(x)h\rangle_Y = G^\circ(F(x);F'(x)h)$$

and hence by Lemma 13.13 that

$$F'(x)^* \partial_C G(F(x)) = \{x^* \in X^* \mid \langle x^*, h\rangle_X \leq G^\circ(F(x);F'(x)h) \text{ for all } h \in X\}.$$

Combined with (13.15) or (13.16) and the definition of the Clarke subdifferential in (13.2), this now yields the claimed inclusion or equality, respectively, for the Clarke subdifferential of the composition. □

Again, the generic inclusion is the reverse of the one in the convex chain rule. Note that equality in the chain rule also holds if $-G$ is regular, since we can then apply Theorem 13.23 to $-G \circ F$ and use that $\partial_C(-G)(F(x)) = -\partial_C G(F(x))$ by Theorem 13.10. Furthermore, if G is not regular but $F'(x)$ is surjective, a similar proof shows that equality (but not the regularity of $G \circ F$) holds in the chain rule; see [Clarke, 2013, Theorem 10.19].

Example 13.24. As a simple example, we consider

$$F : \mathbb{R}^2 \to \mathbb{R}, \qquad (x_1, x_2) \mapsto |x_1 x_2|,$$

which is not convex. To compute the Clarke subdifferential, we write $F = g \circ T$ for

$$g : \mathbb{R} \to \mathbb{R}, \quad t \mapsto |t|, \qquad T : \mathbb{R}^2 \to \mathbb{R}, \quad (x_1, x_2) \mapsto x_1 x_2,$$

where g is finite-valued, convex, and Lipschitz continuous, and hence regular at any $t \in \mathbb{R}$, and T is continuously differentiable for all $x \in \mathbb{R}^2$ with Fréchet derivative

$$T'(x) : \mathbb{R}^2 \to \mathbb{R}, \qquad T'(x)h := x_2 h_1 + x_1 h_2.$$

Its adjoint is easily verified to be given by

$$T'(x)^* : \mathbb{R} \to \mathbb{R}^2, \qquad T'(x)^* t := \left(\begin{smallmatrix} x_2 t \\ x_1 t \end{smallmatrix}\right).$$

Hence, Theorem 13.23 together with Theorem 13.8 yields that F is regular at any $x \in \mathbb{R}^2$ and that

$$\partial_C F(x) = T'(x)^* \partial g(T(x)) = \begin{pmatrix} x_2 \\ x_1 \end{pmatrix} \operatorname{sign}(x_1 x_2),$$

for the set-valued sign function from Example 4.7.

13.4 CHARACTERIZATION IN FINITE DIMENSIONS

A more explicit characterization of the Clarke subdifferential is possible in finite-dimensional spaces. The basis is the following theorem, which only holds in $\mathbb{R}^N$; a proof can be found in, e.g., [DiBenedetto, 2002, Theorem 23.2] or [Heinonen, 2005, Theorem 3.1].

Theorem 13.25 (Rademacher). *Let $U \subset \mathbb{R}^N$ be open and $F : U \to \mathbb{R}$ be Lipschitz continuous. Then F is Fréchet differentiable at almost every $x \in U$.*

This result allows replacing the lim sup in the definition of the Clarke subdifferential (now considered as a subset of $\mathbb{R}^N$, i.e., identifying the dual of $\mathbb{R}^N$ with $\mathbb{R}^N$ itself) with a proper limit.

Theorem 13.26. *Let $F : \mathbb{R}^N \to \mathbb{R}$ be locally Lipschitz continuous near $x \in \mathbb{R}^N$. Then F is Fréchet differentiable on $\mathbb{R}^N \setminus E_F$ for a set $E_F \subset \mathbb{R}^N$ of Lebesgue measure 0 and*

$$\partial_C F(x) = \operatorname{co} \left\{ \lim_{n\to\infty} \nabla F(x_n) \;\middle|\; x_n \to x,\ x_n \notin E_F \right\}, \tag{13.17}$$

where $\operatorname{co} A$ denotes the convex hull of $A \subset \mathbb{R}^N$.

Proof. We first note that the Rademacher Theorem ensures that such a set E_F exists and has Lebesgue measure 0. Hence there indeed exist sequences $\{x_n\}_{n\in\mathbb{N}} \in \mathbb{R}^N \setminus E_F$ with $x_n \to x$.

Furthermore, the local Lipschitz continuity of F yields that for any x_n in the Lipschitz neighborhood of x and any $h \in \mathbb{R}^N$, we have that

$$|\langle \nabla F(x_n), h\rangle| = \left|\lim_{t \searrow 0} \frac{F(x_n + th) - F(x_n)}{t}\right| \leq L\|h\|$$

and hence that $\|\nabla F(x_n)\| \leq L$ for all $n \in \mathbb{N}$ large enough. This implies that $\{\nabla F(x_n)\}_{n\in\mathbb{N}} \subset \mathbb{R}^N$ is bounded and thus contains a convergent subsequence. The set on the right-hand side of (13.17) is therefore nonempty.

Let now $\{x_n\}_{n\in\mathbb{N}} \subset \mathbb{R}^N \setminus E_F$ be an arbitrary sequence with $x_n \to x$ and $\{\nabla F(x_n)\}_{n\in\mathbb{N}} \to x^*$ for some $x^* \in \mathbb{R}^N$. Since F is differentiable at every $x_n \notin E_F$ by definition, Lemma 13.7 yields that $\nabla F(x_n) \in \partial_C F(x_n)$, and hence $x^* \in \partial_C F(x)$ by Lemma 13.3. The convexity of $\partial_C F(x)$ from Lemma 13.2 now implies that any convex combination of such limits x^* is contained in $\partial_C F(x)$, which shows the inclusion "$\supset$" in (13.17).

For the other inclusion, we first show for all $h \in \mathbb{R}^N$ and $\varepsilon > 0$ that

$$F^\circ(x; h) - \varepsilon \leq \limsup_{E_F \not\ni y \to x} \langle \nabla F(y), h\rangle =: M(h). \tag{13.18}$$

Indeed, by definition of $M(h)$ and of the lim sup, for every $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\langle \nabla F(y), h\rangle \leq M(h) + \varepsilon \qquad \text{for all } y \in \mathbb{O}(x, \delta) \setminus E_F.$$

Here, $\delta > 0$ can be chosen sufficiently small for F to be Lipschitz continuous on $\mathbb{O}(x, \delta)$. In particular, $E_F \cap \mathbb{O}(x, \delta)$ is a set of zero measure. Hence, F is differentiable at $y + th$ for almost all $y \in \mathbb{O}(x, \frac{\delta}{2})$ and almost all $t \in (0, \frac{\delta}{2\|h\|})$ by Fubini's Theorem. The classical mean value theorem therefore yields for all such y and t that

$$F(y + th) - F(y) = \int_0^t \langle \nabla F(y + sh), h\rangle \, ds \leq t(M(h) + \varepsilon) \tag{13.19}$$

since $y + sh \in \mathbb{O}(x, \delta)$ for all $s \in (0, t)$ by the choice of t. The continuity of F implies that the full inequality (13.19) even holds for *all* $y \in \mathbb{O}(x, \frac{\delta}{2})$ and *all* $t \in (0, \frac{\delta}{2\|h\|})$. Dividing by $t > 0$ and taking the lim sup over all $y \to x$ and $t \searrow 0$ now yields (13.18).

Since $\varepsilon > 0$ was arbitrary, this implies that $F^\circ(x; h) \leq M(h)$ for all $h \in \mathbb{R}^N$ and hence that

$$\partial_C F(x) \subset \left\{x^* \in \mathbb{R}^N \;\middle|\; \langle x^*, h\rangle \leq M(h) \quad \text{for all } h \in \mathbb{R}^N\right\} =: B.$$

We are thus finished if we can show that B is equal to the set on the right-hand side of (13.17), which we denote by $\operatorname{co} A$. For this, we once again appeal to Corollary 13.14. First, we note that the definition of the convex hull implies for all $h \in \mathbb{R}^N$ that

$$\sup_{x^* \in \operatorname{co} A} \langle x^*, h\rangle = \sup_{\substack{x_i^* \in A \\ \sum_i t_i = 1, t_i \geq 0}} \sum_i t_i \langle x_i^*, h\rangle = \sup_{\sum_i t_i = 1, t_i \geq 0} \sum_i t_i \sup_{x_i^* \in A} \langle x_i^*, h\rangle = \sup_{x^* \in A} \langle x^*, h\rangle$$

since the sum is maximal if and only if each summand is maximal. Next we have that

$$M(h) = \limsup_{E_F \not\ni y \to x} \langle \nabla F(y), h \rangle = \sup_{E_F \not\ni x_n \to x} \langle \lim_{n\to\infty} \nabla F(x_n), h \rangle = \sup_{x^* \in A} \langle x^*, h \rangle.$$

Finally, one can show as in Lemma 13.1 that the mapping $h \mapsto M(h)$ is positively homogeneous, subadditive, and lower semicontinuous. From Lemma 13.12, we thus have that

$$\sup_{x^* \in B} \langle x^*, h \rangle = M(h) = \sup_{x^* \in A} \langle x^*, h \rangle = \sup_{x^* \in \operatorname{co} A} \langle x^*, h \rangle.$$

Since both sets are clearly convex and closed as well as nonempty (which we've already argued for $\operatorname{co} A$ and which follows from (13.18) for B), (13.9) yields $B = \operatorname{co} A$ and thus the claim. □

Remark 13.27. It is possible to extend the Clarke subdifferential defined here to extended-real valued functions using an equivalent, more geometrical, construction involving generalized normal cones to epigraphs; see [Clarke, 1990, Definition 2.4.10]. We will follow this approach when studying the more general subdifferentials for set-valued functionals in Chapters 18 and 20.

14 SEMISMOOTH NEWTON METHODS

The proximal point and splitting methods in Chapter 8 are generalizations of gradient methods and in general have at most linear convergence. In this chapter, we will therefore consider second-order methods, specifically a generalization of Newton's method which admits (locally) superlinear convergence.

14.1 CONVERGENCE OF GENERALIZED NEWTON METHODS

As a motivation, we first consider the most general form of a Newton-type method. Let X and Y be normed vector spaces and $F : X \to Y$ be given and suppose we are looking for an $\bar{x} \in X$ with $F(\bar{x}) = 0$. A Newton-type method to find such an $\bar{x}$ then consists of repeating the following steps:

1. choose an invertible $M_k := M(x^k) \in \mathbb{L}(X; Y)$;
2. solve the *Newton step* $M_k s^k = -F(x^k)$;
3. update $x^{k+1} = x^k + s^k$.

We can now ask under which conditions this method converges to $\bar{x}$, and in particular, when the convergence is *superlinear*, i.e.,

$$\lim_{k\to\infty} \frac{\|x^{k+1} - \bar{x}\|_X}{\|x^k - \bar{x}\|_X} = 0. \tag{14.1}$$

(Recall the discussion in the beginning of Chapter 10.) For this purpose, we set $e^k := x^k - \bar{x}$ and use the Newton step together with the fact that $F(\bar{x}) = 0$ to obtain that

$$\begin{aligned} \|x^{k+1} - \bar{x}\|_X &= \|x^k - M(x^k)^{-1}F(x^k) - \bar{x}\|_X \\ &= \|M(x^k)^{-1}\left[F(x^k) - F(\bar{x}) - M(x^k)(x^k - \bar{x})\right]\|_X \\ &= \|M(\bar{x} + e^k)^{-1}\left[F(\bar{x} + e^k) - F(\bar{x}) - M(\bar{x} + e^k)e^k\right]\|_X \\ &\leq \|M(\bar{x} + e^k)^{-1}\|_{\mathbb{L}(Y;X)}\|F(\bar{x} + e^k) - F(\bar{x}) - M(\bar{x} + e^k)e^k\|_Y. \end{aligned} \tag{14.2}$$

Hence, (14.1) holds under

(i) a *regularity condition*: there exists a $C > 0$ with

$$\|M(x^k)^{-1}\|_{\mathbb{L}(Y;X)} \leq C \qquad \text{for all } k \in \mathbb{N};$$

(ii) an *approximation condition*:

$$\lim_{k\to\infty} \frac{\|F(\bar{x}+e^k) - F(\bar{x}) - M(\bar{x}+e^k)e^k\|_Y}{\|e^k\|_X} = 0.$$

This motivates the following definition: We call $F : X \to Y$ *Newton differentiable* in $x \in X$ with *Newton derivative* $D_N F(x)$ if there exists a neighborhood $U \subset X$ of x and a mapping $D_N F : U \to \mathbb{L}(X;Y)$ such that

$$\lim_{\|h\|_X\to 0} \frac{\|F(x+h) - F(x) - D_N F(x+h)h\|_Y}{\|h\|_X} = 0. \tag{14.3}$$

Note the differences to the Fréchet derivative: First, the Newton derivative is evaluated in $x + h$ instead of x. More importantly, we have not required *any* connection between $D_N F$ with F, while the only possible candidate for the Fréchet derivative was the Gâteaux derivative (which itself was linked to F via the directional derivative). A function thus can only be Newton differentiable (or not) with respect to a concrete choice of $D_N F$. In particular, Newton derivatives are not unique.

If F is Newton differentiable with Newton derivative $D_N F$, we can set $M(x^k) = D_N F(x^k)$ and obtain the *semismooth Newton method*

$$x^{k+1} := x^k - D_N F(x^k)^{-1} F(x^k). \tag{14.4}$$

Its local superlinear convergence follows directly from the construction.

Theorem 14.1. *Let X, Y be normed vector spaces and let $F : X \to Y$ be Newton differentiable near $\bar{x} \in X$ with $F(\bar{x}) = 0$ with Newton derivative $D_N F(\bar{x})$. Assume further that there exist $\delta > 0$ and $C > 0$ with $\|D_N F(x)^{-1}\|_{\mathbb{L}(Y;X)} \leq C$ for all $x \in \mathbb{O}(\bar{x}, \delta)$. Then the semismooth Newton method (14.4) converges superlinearly to $\bar{x}$ for all x^0 sufficiently close to $\bar{x}$.*

Proof. The proof is virtually identical to that for the classical Newton method. We have already shown that for any $x^0 \in \mathbb{O}(\bar{x}, \delta)$,

$$\|e^1\|_X \leq C\|F(\bar{x}+e^0) - F(\bar{x}) - D_N F(\bar{x}+e^0)e^0\|_Y. \tag{14.5}$$

Let now $\varepsilon \in (0,1)$ be arbitrary. The Newton differentiability of F then implies that there exists a $\rho > 0$ such that

$$\|F(\bar{x}+h) - F(\bar{x}) - D_N F(\bar{x}+h)h\|_Y \leq \frac{\varepsilon}{C}\|h\|_X \qquad \text{for all } \|h\|_X \leq \rho. \tag{14.6}$$

Hence, if we choose x^0 such that $\|\bar{x} - x^0\|_X \leq \min\{\delta, \rho\}$, the estimate (14.5) implies that $\|\bar{x}-x^1\|_X \leq \varepsilon\|\bar{x}-x^0\|_X$. By induction, we obtain from this that $\|\bar{x}-x^k\|_X \leq \varepsilon^k\|\bar{x}-x^0\|_X \to 0$. Since $\varepsilon \in (0,1)$ was arbitrary, we can take in each step k a different $\varepsilon_k \to 0$ to obtain that $\|x^{k+1} - \bar{x}\|_X \leq \varepsilon_k\|x^k - \bar{x}\|_X$ and hence that the convergence is superlinear. □

Sometimes, the Newton derivatives $D_N F(x)$ are poorly conditioned, or the region of convergence impractically small. In that case, it may help to *dampen* the method to

$$x^{k+1} := x^k - [D_N F(x^k) + \theta \mathrm{Id}]^{-1} F(x^k),$$

for some $\theta > 0$. As shown in the next theorem, this method still converges, but only linearly. For this scheme, we would take $M(x) = D_N F(x) + \theta \mathrm{Id}$ in the theorem. As we will learn in Chapter 30, it is also possible to modify $D_N F(x)$ only on a subspace.

Theorem 14.2. *Let X, Y be normed vector spaces and let $F : X \to Y$ be Newton differentiable near $\bar{x} \in X$ with $F(\bar{x}) = 0$ with Newton derivative $D_N F(\bar{x})$. Also assume to be given $M(x) \in \mathbb{L}(X;Y)$ that satisfy $\|M(x) - D_N F(x)\|_{\mathbb{L}(X;Y)} \leq \theta$ and $\|M(x)^{-1}\|_{\mathbb{L}(Y;X)} \leq C$ for all $x \in \mathbb{O}(\bar{x}, \delta)$ for some $\theta, \delta > 0$ and $0 < C < \theta^{-1}$. Then $x^{k+1} := x^k - M(x^k)^{-1} F(x^k)$ converge linearly to $\bar{x}$ for all x^0 sufficiently close to $\bar{x}$.*

Proof. Following (14.2), we have

$$\|e^1\|_X \leq C \|F(\bar{x} + e^0) - F(\bar{x}) - M(\bar{x} + e^0) e^0\|_Y. \tag{14.7}$$

Let $\varepsilon > 0$. Using the Newton differentiability of F, following (14.6), we deduce the existence of $\rho > 0$ such that whenever $\|h\|_X \leq \rho$, we have

$$\begin{aligned} \|F(\bar{x} + h) - F(\bar{x}) - M(\bar{x} + h)h\|_Y &\leq \|F(\bar{x} + h) - F(\bar{x}) - D_N F(\bar{x} + h)h\|_Y \\ &\quad + \|[D_N F(\bar{x} + h) - M(\bar{x} + h)]h\|_Y \\ &\leq \left(\frac{\varepsilon}{C} + \theta\right) \|h\|_X. \end{aligned}$$

Hence, if we choose x^0 such that $\|\bar{x} - x^0\|_X \leq \min\{\delta, \rho\}$, the estimate (14.7) implies that $\|\bar{x} - x^1\|_X \leq (C\theta + \varepsilon)\|\bar{x} - x^0\|_X$. Since $0 < C\theta < 1$, taking $\varepsilon > 0$ small enough, we have $\beta := C\theta + \varepsilon \in (0, 1)$. By induction, we obtain from this that $\|\bar{x} - x^k\|_X \leq \beta^k \|\bar{x} - x^0\|_X \to 0$. This shows the linear convergence. □

14.2 NEWTON DERIVATIVES

The remainder of this chapter is dedicated to the construction of Newton derivatives that satisfy the approximation condition (although it should be pointed out that the verification of the regularity condition is usually the much more involved step in practice, which is usually very specific to the concrete problem). We begin with the obvious connection with the Fréchet derivative.

Theorem 14.3. *Let X, Y be normed vector spaces. If $F : X \to Y$ is continuously differentiable at $x \in X$, then F is also Newton differentiable at x with Newton derivative $D_N F(x) = F'(x)$.*

Proof. We have for arbitrary $h \in X$ that

$$\begin{aligned}\|F(x+h) - F(x) - F'(x+h)h\|_Y \leq{}& \|F(x+h) - F(x) - F'(x)h\|_Y \\ &+ \|F'(x) - F'(x+h)\|_{\mathbb{L}(X;Y)}\|h\|_X,\end{aligned}$$

where the first summand is $o(\|h\|_X)$ by definition of the Fréchet derivative and the second by the continuity of F'. □

Calculus rules can be shown similarly to those for Fréchet derivatives. For the sum rule this is immediate; here we prove a chain rule by way of example.

Theorem 14.4. *Let X, Y, and Z be normed vector spaces, and let $F : X \to Y$ be Newton differentiable at $x \in X$ with Newton derivative $D_N F(x)$ and $G : Y \to Z$ be Newton differentiable at $y := F(x) \in Y$ with Newton derivative $D_N G(y)$. If $D_N F$ and $D_N G$ are uniformly bounded in a neighborhood of x and y, respectively, then $G \circ F$ is also Newton differentiable at x with Newton derivative*

$$D_N(G \circ F)(x) = D_N G(F(x)) \circ D_N F(x).$$

Proof. We proceed as in the proof of Theorem 2.7. For $h \in X$ and $g := F(x+h) - F(x)$ we have that

$$(G \circ F)(x+h) - (G \circ F)(x) = G(y+g) - G(y).$$

The Newton differentiability of G then implies that

$$\|(G \circ F)(x+h) - (G \circ F)(x) - D_N G(y+g)g\|_Z \leq r_1(\|g\|_Y)$$

with $r_1(t)/t \to 0$ for $t \to 0$. The Newton differentiability of F further implies that

$$\|g - D_N F(x+h)h\|_Y \leq r_2(\|h\|_X)$$

with $r_2(t)/t \to 0$ for $t \to 0$. In particular,

$$\|g\|_Y \leq \|D_N F(x+h)\|_{\mathbb{L}(X;Y)}\|h\|_Y + r_2(\|h\|_X).$$

The uniform boundedness of $D_N F$ now implies that $\|g\|_Y \to 0$ for $\|h\|_X \to 0$. Hence, using that $y + g = F(x+h)$, we obtain

$$\begin{aligned}&\|(G \circ F)(x+h) - (G \circ F)(x) - D_N G(F(x+h))D_N F(x+h)h\|_Z \\ &\qquad\leq \|G(y+g) - G(y) - D_N G(y+g)g\|_Z \\ &\qquad\quad + \|D_N G(y+g)\,[g - D_N F(x+h)h]\,\|_Z \\ &\qquad\leq r_1(\|g\|_Y) + \|D_N G(y+g)\|_{\mathbb{L}(Y;Z)} r_2(\|h\|_X),\end{aligned}$$

and the claim thus follows from the uniform boundedness of $D_N G$. □

Finally, it follows directly from the definition of the product norm and Newton differentiability that Newton derivatives of vector-valued functions can be computed componentwise.

Theorem 14.5. *Let X, Y_i be normed vector spaces and let $F_i : X \to Y_i$ be Newton differentiable with Newton derivative $D_N F_i$ for $1 \leq i \leq m$. Then*

$$F : X \to (Y_1 \times \cdots \times Y_m), \qquad x \mapsto (F_1(x), \ldots, F_m(x))^T,$$

is also Newton differentiable with Newton derivative

$$D_N F(x) = (D_N F_1(x), \ldots, D_N F_m(x))^T.$$

Since the definition of a Newton derivative is not constructive, allowing different choices, the question remains how to obtain a candidate for which the approximation condition in the definition can be verified. For two classes of functions, such an explicit construction is known.

LOCALLY LIPSCHITZ CONTINUOUS FUNCTIONS ON $\mathbb{R}^N$

If $F : \mathbb{R}^N \to \mathbb{R}$ is locally Lipschitz continuous, candidates can be taken from the Clarke subdifferential, which has an explicit characterization by Theorem 13.26. Under some additional assumptions, each candidate is indeed a Newton derivative.

A function $F : \mathbb{R}^N \to \mathbb{R}$ is called *piecewise (continuously) differentiable* or *PC^1 function*, if

(i) F is continuous on $\mathbb{R}^N$;

(ii) for all $x \in \mathbb{R}^N$ there exists an open neighborhood $U_x \subset \mathbb{R}^N$ of x and a finite set $\{F_i : U_x \to \mathbb{R}\}_{i \in I_x}$ of continuously differentiable functions with

$$F(\tilde{x}) \in \{F_i(\tilde{x})\}_{i \in I_x} \qquad \text{for all } \tilde{x} \in U_x.$$

In this case, we call F a *continuous selection* of the F_i in U_x. The set

$$I_a(x) := \{i \in I_x \mid F(x) = F_i(x)\}$$

is called the *active index set* at x. Since the F_i are continuous, we have that $F(\tilde{x}) \neq F_j(\tilde{x})$ for all $j \notin I_a(x)$ and $\tilde{x}$ sufficiently close to x. Hence, indices that are only active on sets of zero measure do not have to be considered in the following. We thus define the *essentially active index set*

$$I_e(x) := \{i \in I_x \mid x \in \operatorname{cl}(\operatorname{int}\{\tilde{x} \in U_x \mid F(\tilde{x}) = F_i(\tilde{x})\})\} \subset I_a(x).$$

An example of an active but not essentially active index set is the following.

Example 14.6. Consider the function

$$f : \mathbb{R} \to \mathbb{R}, \qquad t \mapsto \max\{0, t, t/2\},$$

i.e., $f_1(t) = 0$, $f_2(t) = t$, and $f_3(t) = t/2$. Then $I_a(0) = \{1, 2, 3\}$ but $I_e(0) = \{1, 2\}$, since f_3 is active only at $t = 0$ and hence $\operatorname{int}\{t \in \mathbb{R} \mid f(t) = f_3(t)\} = \emptyset = \operatorname{cl}\emptyset$.

Since any C^1 function $F_i : U_x \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant $L_i := \sup_{\tilde{x}\in U_x} |\nabla F(\tilde{x})|$ by Lemma 2.11, PC^1 functions are always locally Lipschitz continuous; see [Scholtes, 2012, Corollary 4.1.1].

Theorem 14.7. *Let $F : \mathbb{R}^N \to \mathbb{R}$ be piecewise differentiable. Then F is locally Lipschitz continuous on $\mathbb{R}^N$ with local constant $L(x) = \max_{i\in I_a(x)} L_i$.*

This yields the following explicit characterization of the Clarke subdifferential of a PC^1 function.

Theorem 14.8. *Let $F : \mathbb{R}^N \to \mathbb{R}$ be piecewise differentiable and $x \in \mathbb{R}^N$. Then*

$$\partial_C F(x) = \operatorname{co}\{\nabla F_i(x) \mid i \in I_e(x)\}.$$

Proof. Let $x \in \mathbb{R}^N$ be arbitrary. By Theorem 13.26 it suffices to show that

$$\left\{ \lim_{n\to\infty} \nabla F(x_n) \;\middle|\; x_n \to x,\ x_n \notin E_F \right\} = \{\nabla F_i(x) \mid i \in I_e(x)\},$$

where E_F is the set of Lebesgue measure 0 where F is not differentiable from Rademacher's Theorem. For this, let $\{x_n\}_{n\in\mathbb{N}} \subset \mathbb{R}^N$ be a sequence with $x_n \to x$, F is differentiable at x_n for all $n \in \mathbb{N}$, and $\nabla F(x_n) \to x^* \in \mathbb{R}^N$. Since F is differentiable at x_n, it must hold that $F(\tilde{x}) = F_{i_n}(\tilde{x})$ for some $i_n \in I_a(x)$ and all $\tilde{x}$ sufficiently close to x_n, which implies that $\nabla F(x_n) = \nabla F_{i_n}(x_n)$. For sufficiently large $n \in \mathbb{N}$, we can further assume that $i_n \in I_e(x)$ (if necessary, by adding x_n with $i_n \notin I_e(x)$ to E_F, which does not increase its Lebesgue measure). If we now consider subsequences $\{x_{n_k}\}_{k\in\mathbb{N}}$ with constant index $i_{n_k} =: i \in I_e(x)$ (which exist since $I_e(x)$ is finite), we obtain using the continuity of ∇F_i that

$$x^* = \lim_{k\to\infty} \nabla F(x_{n_k}) = \lim_{k\to\infty} \nabla F_i(x_{n_k}) \in \{\nabla F_i(x) \mid i \in I_e(x)\}.$$

Conversely, for every $\nabla F_i(x)$ with $i \in I_e(x)$ there exists by definition of the essentially active indices a sequence $\{x_n\}_{n\in\mathbb{N}}$ with $x_n \to x$ and $F = F_i$ in a sufficiently small neighborhood of each x_n for n large enough. The continuous differentiability of the F_i thus implies that $\nabla F(x_n) = \nabla F_i(x_n)$ for all $n \in \mathbb{N}$ large enough and hence that

$$\nabla F_i(x) = \lim_{n\to\infty} \nabla F_i(x_n) = \lim_{n\to\infty} \nabla F(x_n). \qquad \square$$

From this, we obtain the Newton differentiability of PC^1 functions.

Theorem 14.9. *Let $F : \mathbb{R}^N \to \mathbb{R}$ be piecewise differentiable. Then F is Newton differentiable for all $x \in \mathbb{R}^N$, and every $D_N F(x) \in \partial_C F(x)$ is a Newton derivative.*

Proof. Let $x \in \mathbb{R}^N$ be arbitrary and $h \in X$ with $x + h \in U_x$. By Theorem 14.8, every $D_N F(x+h) \in \partial_C F(x+h)$ is of the form

$$D_N F(x+h) = \sum_{i \in I_e(x+h)} \lambda_i \nabla F_i(x+h) \qquad \text{for} \quad \sum_{i \in I_e(x+h)} \lambda_i = 1,\ \lambda_i \geq 0.$$

Since F is continuous, we have for all $h \in \mathbb{R}^N$ sufficiently small that $I_e(x+h) \subset I_a(x+h) \subset I_a(x)$, where the second inclusion follows from the fact that by continuity, $F(x) \neq F_i(x)$ implies that $F(x+h) \neq F_i(x+h)$. Hence, $F(x+h) = F_i(x+h)$ and $F(x) = F_i(x)$ for all $i \in I_e(x+h)$. Theorem 14.3 then yields that

$$|F(x+h) - F(x) - D_N F(x+h)h| \leq \sum_{i \in I_e(x+h)} \lambda_i |F_i(x+h) - F_i(x) - \nabla F_i(x+h)h| = o(\|h\|),$$

since all F_i are continuously differentiable by assumption. □

A natural application of the above are proximal point mappings of convex and lower semicontinuous functionals.

Example 14.10.

(i) We first consider the proximal mapping for the indicator function $\delta_A : \mathbb{R}^N \to \overline{\mathbb{R}}$ of the set $A := \{x \in \mathbb{R}^N \mid x_i \in [a,b]\}$ for some $a < b \in \mathbb{R}$. Analogously to (iii), the corresponding proximal mapping is the componentwise projection

$$[\mathrm{proj}_A(x)]_i = \mathrm{proj}_{[a,b]} x_i = \begin{cases} a & \text{if } x_i < a, \\ x_i & \text{if } x_i \in [a,b], \\ b & \text{if } x_i > b, \end{cases}$$

which is clearly piecewise differentiable. Theorem 14.8 thus yields (also componentwise) that

$$\partial_C [\mathrm{proj}_A(x)]_i = \begin{cases} \{1\} & \text{if } x_i \in (a,b), \\ \{0\} & \text{if } x_i \notin [a,b], \\ [0,1] & \text{if } x_i \in \{a,b\}. \end{cases}$$

By Theorems 14.5 and 14.9, a possible Newton derivative is therefore given by

$$[D_N \mathrm{proj}_A(x)h]_i = [\mathbb{1}_{[a,b]}(x) \odot h]_i := \begin{cases} h_i & \text{if } x_i \in [a,b], \\ 0 & \text{if } x_i \notin [a,b], \end{cases}$$

where the choice of which case to include $x_i \in \{a, b\}$ in is arbitrary. (The componentwise product $[x \odot y]_i := x_i y_i$ on $\mathbb{R}^N$ is also known as the *Hadamard product*.)

(ii) Consider now the proximal mapping for $G : \mathbb{R}^N \to \mathbb{R}$, $G(x) := \|x\|_1$, whose proximal mapping for arbitrary $\gamma > 0$ is given by Example 6.26 (ii) componentwise as

$$[\operatorname{prox}_{\gamma G}(x)]_i = \begin{cases} x_i - \gamma & \text{if } x_i > \gamma, \\ 0 & \text{if } x_i \in [-\gamma, \gamma], \\ x_i + \gamma & \text{if } x_i < -\gamma. \end{cases}$$

Again, this is clearly piecewise differentiable, and Theorem 14.8 thus yields (also componentwise) that

$$\partial_C[(\operatorname{prox}_{\gamma G})(x)]_i = \begin{cases} \{1\} & \text{if } |x_i| > \gamma, \\ \{0\} & \text{if } |x_i| < \gamma, \\ [0,1] & \text{if } |x_i| = \gamma. \end{cases}$$

By Theorems 14.5 and 14.9, a possible Newton derivative is therefore given by

$$[D_N \operatorname{prox}_{\gamma G}(x)h]_i = [\mathbb{1}_{\{|t| \geq \gamma\}}(x) \odot h]_i := \begin{cases} h_i & \text{if } |x_i| \geq \gamma, \\ 0 & \text{if } |x_i| < \gamma, \end{cases}$$

where again we could have taken the value th_i for any $t \in [0,1]$ for $|x_i| = \gamma$.

SUPERPOSITION OPERATORS ON $L^p(\Omega)$

Rademacher's Theorem does not hold in infinite-dimensional function spaces, and hence the Clarke subdifferential no longer yields an algorithmically useful candidate for a Newton derivative in general. One exception is the class of superposition operators defined by scalar Newton differentiable functions, for which the Newton derivative can be evaluated pointwise as well.

We thus consider as in Section 2.3 for an open and bounded domain $\Omega \subset \mathbb{R}^N$, a Carathéodory function $f : \Omega \times \mathbb{R} \to \mathbb{R}$ (i.e., $(x, z) \mapsto f(x, z)$ is measurable in x and continuous in z), and $1 \leq p, q \leq \infty$ the corresponding superposition operator

$$F : L^p(\Omega) \to L^q(\Omega), \qquad [F(u)](x) = f(x, u(x)) \quad \text{for almost every } x \in \Omega.$$

The goal is now to similarly obtain a Newton derivative $D_N F$ for F as a superposition operator defined by the Newton derivative $D_N f(x, z)$ of $z \mapsto f(x, z)$. Here, the assumption that $D_N f$ is also a Carathéodory function is too restrictive, since we want to allow discontinuous derivatives as well (see Example 14.10). Luckily, for our purpose, a weaker

property is sufficient: A function is called *Baire–Carathéodory function* if it can be written as a pointwise limit of Carathéodory functions, i.e., if

$$f(x,z) = \lim_{n\to\infty} f_n(x,z) \qquad \text{for almost every } x \in \Omega \text{ and all } z \in \mathbb{R},$$

where f_n is a Carathéodory function for all $n \in \mathbb{N}$; see [Appell and Zabrejko, 1990, Lemma 1.4].

Under certain growth conditions on f and $D_N f$,[1] we can transfer the Newton differentiability of f to F, but we again have to take a *two-norm discrepancy* into account.

Theorem 14.11. *Let $f : \Omega \times \mathbb{R} \to \mathbb{R}$ be a Carathéodory function. Furthermore, assume that*

(i) $z \mapsto f(x,z)$ is uniformly Lipschitz continuous for almost every $x \in \Omega$ and $x \mapsto f(x,0)$ is bounded;

(ii) $z \mapsto f(x,z)$ is Newton differentiable with Newton derivative $z \mapsto D_N f(x,z)$ for almost every $x \in \Omega$;

(iii) $D_N f$ is a Baire–Carathéodory function and uniformly bounded.

Then for any $1 \le q < p < \infty$, the corresponding superposition operator $F : L^p(\Omega) \to L^q(\Omega)$ is Newton differentiable with Newton derivative

$$D_N F : L^p(\Omega) \to \mathbb{L}(L^p(\Omega); L^q(\Omega)), \qquad [D_N F(u)h](x) = D_N f(x, u(x))h(x)$$

for almost every $x \in \Omega$ and all $h \in L^p(\Omega)$.

Proof. First, the uniform Lipschitz continuity together with the reverse triangle inequality yields that

$$|f(x,z)| \le |f(x,0)| + L|z| \le C + L|z|^{q/q} \quad \text{for almost every } x \in \Omega \text{ and all } z \in \mathbb{R},$$

and hence the growth condition (2.5) is satisfied for all $1 \le q \le \infty$. Due to the continuous embedding $L^p(\Omega) \hookrightarrow L^q(\Omega)$ for all $1 \le q \le p \le \infty$, the superposition operator $F : L^p(\Omega) \to L^q(\Omega)$ is therefore well-defined and continuous by Theorem 2.14.

For any measurable $u : \Omega \to \mathbb{R}$, we have that $x \mapsto D_N f(x, u(x))$ is by assumption (iii) the pointwise limit of measurable functions and hence itself measurable. Furthermore, its uniform boundedness in particular implies the growth condition (2.5) for $p' := p$ and $q' := p - q > 0$. As in the proof of Theorem 2.15, we deduce that the corresponding superposition operator $D_N F : L^p(\Omega) \to L^s(\Omega)$ is well-defined and continuous for $s := \frac{pq}{p-q}$, and that for any $u \in L^p(\Omega)$, the mapping $h \mapsto D_N F(u) \cdot h$ defines a bounded linear operator $D_N F(u) : L^p(\Omega) \to L^q(\Omega)$. (This time, we do not distinguish in notation between the linear operator and the function defining this operator by pointwise multiplication.)

[1] which can be significantly relaxed; see [Schiela, 2008, Proposition A.1]

To show that $D_N F(u)$ is a Newton derivative for F in $u \in L^p(\Omega)$, we consider the pointwise residual

$$r : \Omega \times \mathbb{R} \to \mathbb{R}, \qquad r(x,z) := \begin{cases} \frac{|f(x,z)-f(x,u(x))-D_N f(x,z)(z-u(x))|}{|z-u(x)|} & \text{if } z \neq u(x), \\ 0 & \text{if } z = u(x). \end{cases}$$

Since f is a Carathéodory function and $D_N f$ is a Baire–Carathéodory function, the function $x \mapsto r(x, \tilde{u}(x)) =: [R(\tilde{u})](x)$ is measurable for any measurable $\tilde{u} : \Omega \to \mathbb{R}$ (since sums, products, and quotients of measurable functions are again measurable). Furthermore, for $\tilde{u} \in L^p(\Omega)$, the uniform Lipschitz continuity of f and the uniform boundedness of $D_N f$ imply that for almost every $x \in \Omega$ with $\tilde{u}(x) \neq u(x)$,

$$|[R(\tilde{u})](x)| = \frac{|f(x,\tilde{u}(x)) - f(x,u(x)) - D_N f(x,\tilde{u}(x))(\tilde{u}(x)-u(x))|}{|\tilde{u}(x)-u(x)|} \leq L + C \tag{14.8}$$

and thus that $R(\tilde{u}) \in L^\infty(\Omega)$. Hence, the superposition operator $R : L^p(\Omega) \to L^s(\Omega)$ is well-defined.

Let now $\{u_n\}_{n\in\mathbb{N}} \subset L^p(\Omega)$ be a sequence with $u_n \to u \in L^p(\Omega)$. Then there exists a subsequence, again denoted by $\{u_n\}_{n\in\mathbb{N}}$, with $u_n(x) \to u(x)$ for almost every $x \in \Omega$. Since $z \mapsto f(x,z)$ is Newton differentiable almost everywhere, we have by definition that $r(x, u_n(x)) \to 0$ for almost every $x \in \Omega$. Together with the boundedness from (14.8), Lebesgue's dominated convergence theorem therefore yields that $R(u_n) \to 0$ in $L^s(\Omega)$ (and hence along the full sequence since the limit is unique).[2] For any $\tilde{u} \in L^p(\Omega)$, the Hölder inequality with $\frac{1}{p} + \frac{1}{s} = \frac{1}{q}$ thus yields that

$$\|F(\tilde{u}) - F(u) - D_N F(\tilde{u})(\tilde{u}-u)\|_{L^q} = \|R(\tilde{u})(\tilde{u}-u)\|_{L^q} \leq \|R(\tilde{u})\|_{L^s} \|\tilde{u}-u\|_{L^p}.$$

If we now set $\tilde{u} := u + h$ for $h \in L^p(\Omega)$ with $\|h\|_{L^p} \to 0$, we have that $\|R(u+h)\|_{L^s} \to 0$ and hence by definition the Newton differentiability of F in u with Newton derivative $h \mapsto D_N F(u)h$ as claimed. □

Example 14.12.

(i) Consider

$$A := \left\{ u \in L^2(\Omega) \;\middle|\; a \leq u(x) \leq b \quad \text{for almost every } x \in \Omega \right\}$$

and $\operatorname{proj}_A : L^p(\Omega) \to L^2(\Omega)$ for $p > 2$, which by (iii) can be written as a superposition operator of the corresponding Lipschitz continuous scalar proximal point operator, whose Newton derivative is given in Example 14.10 (i). Since this derivative is clearly bounded (by 1) and the pointwise limit of continuous functions,

[2] This is the step that fails for $F : L^\infty(\Omega) \to L^\infty(\Omega)$, since pointwise convergence and boundedness together do not imply uniform convergence almost everywhere.

Theorem 14.11 yields the pointwise almost everywhere Newton derivative

$$[D_N \mathrm{proj}_A(u)h](x) = [\mathbb{1}_{[a,b]}(u)h](x) := \begin{cases} h(x) & \text{if } u(x) \in [a,b], \\ 0 & \text{if } u(x) \notin [a,b]. \end{cases}$$

(ii) Consider now

$$G : L^2(\Omega) \to \mathbb{R}, \qquad G(u) = \|u\|_{L^1} = \int_\Omega |u(x)|\,dx$$

and $\mathrm{prox}_{\gamma G} : L^p(\Omega) \to L^2(\Omega)$ for $p > 2$ and $\gamma > 0$, which by (ii) can be written as a superposition operator of the corresponding Lipschitz continuous scalar proximal point operator, whose Newton derivative is given in Example 14.10 (ii). Since this derivative is clearly bounded (by 1) and the pointwise limit of continuous functions, Theorem 14.11 yields the pointwise almost everywhere Newton derivative

$$[D_N \mathrm{prox}_{\gamma G}(u)h](x) = [\mathbb{1}_{\{|t|\geq\gamma\}}(u)h](x) := \begin{cases} h(x) & \text{if } |u(x)| \geq \gamma, \\ 0 & \text{if } |u(x)| < \gamma. \end{cases}$$

For $p = q \in [1,\infty]$, however, the claim is false in general, as can be shown by counterexamples.

Example 14.13. We take

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(z) = \max\{0,z\} := \begin{cases} 0 & \text{if } z \leq 0, \\ z & \text{if } z \geq 0. \end{cases}$$

This is a piecewise differentiable function, and hence by Theorem 14.9 we can for any $\delta \in [0,1]$ take as Newton derivative

$$D_N f(z)h = \begin{cases} 0 & \text{if } z < 0, \\ \delta h & \text{if } z = 0, \\ h & \text{if } z > 0. \end{cases}$$

We now consider the corresponding superposition operators $F : L^p(\Omega) \to L^p(\Omega)$ and $D_N F(u) \in \mathbb{L}(L^p(\Omega); L^p(\Omega))$ for any $p \in [1,\infty)$ and show that the approximation condition (14.3) is violated for $\Omega = (-1,1)$, $u(x) = -|x|$, and

$$h_n(x) = \begin{cases} \frac{1}{n} & \text{if } |x| < \frac{1}{n}, \\ 0 & \text{if } |x| \geq \frac{1}{n}. \end{cases}$$

First, it is straightforward to compute $\|h_n\|_{L^p}^p = \frac{2}{n^{p+1}}$. Then since $[F(u)](x) = \max\{0, -|x|\} = 0$ almost everywhere, we have that

$$[F(u+h_n) - F(u) - D_N F(u+h_n)h_n](x) = \begin{cases} -|x| & \text{if } |x| < \frac{1}{n}, \\ 0 & \text{if } |x| > \frac{1}{n}, \\ -\frac{\delta}{n} & \text{if } |x| = \frac{1}{n}, \end{cases}$$

and thus

$$\|F(u+h_n) - F(u) - D_N F(u+h_n)h_n\|_{L^p}^p = \int_{-\frac{1}{n}}^{\frac{1}{n}} |x|^p \, dx = \frac{2}{p+1}\left(\frac{1}{n}\right)^{p+1}.$$

This implies that

$$\lim_{n\to\infty} \frac{\|F(u+h_n) - F(u) - D_N F(u+h_n)h_n\|_{L^p}}{\|h_n\|_{L^p}} = \left(\frac{1}{p+1}\right)^{\frac{1}{p}} \neq 0$$

and hence that F is not Newton differentiable from $L^p(\Omega)$ to $L^p(\Omega)$ for any $p < \infty$.

For the case $p = q = \infty$, we take $\Omega = (0,1)$, $u(x) = x$, and

$$h_n(x) = \begin{cases} nx - 1 & \text{if } x \leq \frac{1}{n}, \\ 0 & \text{if } x \geq \frac{1}{n}, \end{cases}$$

such that $\|h_n\|_{L^\infty} = 1$ for all $n \in \mathbb{N}$. We also have that $x + h_n = (1+n)x - 1 \leq 0$ for $x \leq \frac{1}{n+1} \leq \frac{1}{n}$ and hence that

$$[F(u+h_n) - F(u) - D_N F(u+h_n)h_n](x) = \begin{cases} (1+n)x - 1 & \text{if } x \leq \frac{1}{n+1}, \\ 0 & \text{if } x \geq \frac{1}{n+1}, \end{cases}$$

since either $h_n = 0$ or $F(u+h_n) = F(u) + D_N F(u)h_n$ in the second case. Now,

$$\sup_{x \in (0, \frac{1}{n+1}]} |(1+n)x - 1| = 1 \qquad \text{for all } n \in \mathbb{N},$$

which implies that

$$\lim_{n\to\infty} \frac{\|F(u+h_n) - F(u) - D_N F(u+h_n)h_n\|_{L^p}}{\|h_n\|_{L^p}} = 1 \neq 0$$

and hence that F is not Newton differentiable from $L^\infty(\Omega)$ to $L^\infty(\Omega)$ either.

Remark 14.14. Semismoothness was introduced in [Mifflin, 1977] for Lipschitz-continuous functionals $F : \mathbb{R}^N \to \mathbb{R}$ as a condition relating Clarke subderivatives and directional derivatives near a point. This definition was extended to functions $F : \mathbb{R}^N \to \mathbb{R}^M$ in [Qi, 1993; Qi and Sun, 1993] and shown to imply a uniform version of the approximation condition (14.3) for all elements of the Clarke subdifferential and hence superlinear convergence of the semismooth Newton method in finite dimensions. A semismooth Newton method specifically for PC^1 functions was already considered in [Kojima and Shindo, 1986]. In normed vector spaces, [Kummer, 1988] was the first to study an abstract class of Newton methods for nonsmooth equations based on the condition (14.3), unifying the previous results; see [Klatte and Kummer, 2002]. In all these works, the analysis was based on semismoothness as a property relating $F : X \to Y$ to a set-valued mapping $G : X \rightrightarrows L(X, Y)$, whose elements (uniformly) satisfy (14.3). In contrast, [Chen et al., 2000; Kummer, 2000] considered – as we do in this book – single-valued Newton derivatives (named *Newton maps* in the former and *slanting functions* in the latter) in Banach spaces. This approach was later followed in [Hintermüller et al., 2002; Ito and Kunisch, 2008] to show that for a specific choice of Newton derivative, the classical *primal-dual active set method* for solving quadratic optimization problems under linear inequality constraints can be interpreted as a semismooth Newton method. In parallel, [Ulbrich, 2002, 2011] showed that superposition operators defined by semismooth functions (in the sense of [Qi and Sun, 1993]) are semismooth (in the sense of [Kummer, 1988]) between the right spaces. A similar result for single-valued Newton derivatives was shown in [Schiela, 2008] using a proof that is much closer to the one for the classical differentiability of superposition operators; compare Theorems 2.15 and 14.11. It should, however, be mentioned that not all calculus results for semismooth functions are available in the single-valued setting; for example, the implicit function theorem from [Kruse, 2018] requires set-valued Newton derivatives, since the selection of the Newton derivative of the implicit function need not correspond to the selection of the given mapping. Finally, we remark that the notion of semismoothness and semismooth Newton methods were very recently extended to set-valued mappings in [Gferer and Outrata, 2021].

15 NONLINEAR PRIMAL-DUAL PROXIMAL SPLITTING

In this chapter, our goal is to extend the primal-dual proximal splitting (PDPS) method to *nonlinear* operators $K \in C^1(X;Y)$, i.e., to problems of the form

$$\min_X F(x) + G(K(x)), \tag{15.1}$$

where we still assume $F : X \to \overline{\mathbb{R}}$ and $G : Y \to \overline{\mathbb{R}}$ to be convex, proper, and lower semicontinuous on the Hilbert spaces X and Y. For simplicity, we will only consider linear convergence under a strong convexity assumption and refer to the literature for weak convergence and acceleration under partial strong convexity (see Remark 15.12 below). As in earlier chapters, we use the same notation for the inner product as for the duality pairing in Hilbert spaces to distinguish them better from pairs of elements.

We recall the three-point program for convergence proofs of first-order methods from Chapter 9, which remains fundamentally the same in the nonlinear setting. However, we need to make some of the concepts local. Thus the three main ingredients of our convergence proofs will be the following.

(i) The three-point identity (1.5).

(ii) The *local monotonicity* of the operator H whose roots correspond to the (primal-dual) critical points of (15.1). We fix one of the points in the definition of monotonicity in Section 6.2 to a root $\widehat{x}$ of H, and only vary the other point in a neighborhood of $\widehat{x}$. This is essentially a nonsmooth variant of the standard second-order sufficient (or local quadratic growth) condition $\nabla^2 F(x) \succ 0$ (i.e., positive definiteness of the Hessian) for minimizing a smooth function $F : \mathbb{R}^N \to \mathbb{R}$.

(iii) The nonnegativity of the preconditioning operators M_{k+1} defining the implicit form of the algorithm. These will now in general depend on the current iterate, and thus we can only show the nonnegativity in a neighborhood of suitable $\widehat{x}$.

15.1 NONCONVEX EXPLICIT SPLITTING

To motivate our more specific assumptions on K, we start by showing that forward-backward splitting can be applied to a nonconvex function for the forward step. We

thus consider for the problem

$$\min_{x\in X} G(x) + F(x), \tag{15.2}$$

with F smooth but possibly nonconvex, the algorithm

$$x^{k+1} = \operatorname{prox}_{\tau G}(x^k - \tau\nabla F(x^k)). \tag{15.3}$$

To show convergence of this algorithm, we extend the non-value three-point smoothness inequalities of Corollaries 7.2 and 7.7 from convex smooth functions to C^2 functions. (It is also possible to obtain corresponding value inequalities.)

Lemma 15.1. *Suppose $F \in C^2(X)$. Let $z, \widehat{x} \in X$, and suppose for some $L > 0$ and $\gamma \geq 0$ for all $\zeta \in \mathbb{B}(\widehat{x}, \|z - \widehat{x}\|_X)$ that $\gamma \cdot \mathrm{Id} \leq \nabla^2 F(\zeta) \leq L \cdot \mathrm{Id}$. Then for any $\beta \in (0, 2]$ and $x \in X$ we have*

$$\langle \nabla F(z) - \nabla F(\widehat{x}), x - \widehat{x}\rangle_X \geq \frac{\gamma(2-\beta)}{2}\|x - \widehat{x}\|_X^2 - \frac{L}{2\beta}\|x - z\|_X^2. \tag{15.4}$$

Proof. By the one-dimensional mean value theorem applied to $t \mapsto \langle \nabla F(\widehat{x}+t(z-\widehat{x})), x-\widehat{x}\rangle_X$, we obtain for $\zeta = \widehat{x} + s(z - \widehat{x})$ for some $s \in [0, 1]$ that

$$\langle \nabla F(z) - \nabla F(\widehat{x}), x - \widehat{x}\rangle_X = \langle \nabla^2 F(\zeta)(z - \widehat{x}), x - \widehat{x}\rangle_X.$$

Therefore, for any $\beta > 0$,

$$\begin{aligned}\langle \nabla F(z) - \nabla F(\widehat{x}), x - \widehat{x}\rangle_X &= \|x - \widehat{x}\|^2_{\nabla^2 F(\zeta)} + \langle \nabla^2 F(\zeta)(z - x), x - \widehat{x}\rangle_X \\ &\geq \frac{2-\beta}{2}\|x - \widehat{x}\|^2_{\nabla^2 F(\zeta)} - \frac{1}{2\beta}\|x - z\|^2_{\nabla^2 F(\zeta)}.\end{aligned} \tag{15.5}$$

By the definition of γ and L, we obtain (15.4). □

The following result is almost a carbon copy of Theorems 9.6 and 10.2 for convex smooth F. However, since our present problem is nonconvex, we can only expect local convergence to a critical point of $J := F + G$.

Theorem 15.2. *Let $F \in C^2(X)$ and let $G : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. Given an initial iterate x^0 and a critical point $\widehat{x} \in [\partial G + \nabla F]^{-1}(0)$ of $J := F + G$, let $\mathcal{X} := \mathbb{B}(\widehat{x}, \|x^0 - \widehat{x}\|)$, and suppose for some $L > 0$ and $\gamma \geq 0$ that*

$$\gamma \cdot \mathrm{Id} \leq \nabla^2 F(\zeta) \leq L \cdot \mathrm{Id} \quad (\zeta \in \mathcal{X}). \tag{15.6}$$

Take $0 < \tau < 2L^{-1}$.

(i) If $\gamma > 0$, then the sequence $\{x^k\}_{k\in\mathbb{N}}$ generated by (15.3) converges linearly to $\widehat{x}$.

(ii) If $\gamma = 0$, then the sequence $\{x^k\}_{k\in\mathbb{N}}$ converges weakly to a critical point of J.

Note that if G is locally finite-valued, then by Theorem 13.20 our definition of a critical point in this theorem means $\widehat{x} \in [\partial_C J]^{-1}(0)$.

Proof. As usual, we write (15.3) as

$$0 \in \tau[\partial G(x^{k+1}) + \nabla F(x^k)] + (x^{k+1} - x^k). \tag{15.7}$$

Suppose $x^k \in \mathcal{X}$ and let $\beta \in (L\tau, 2)$ be arbitrary (which is possible since $\tau L < 2$). By the monotonicity of ∂G and the local three-point monotonicity (15.4) of F implied by Lemma 15.1, we obtain

$$\langle \partial G(x^{k+1}) + \nabla F(x^k), x^{k+1} - \widehat{x}\rangle_X \geq \frac{\gamma(2-\beta)}{2}\|x^{k+1} - \widehat{x}\|_X^2 - \frac{L}{2\beta}\|x^{k+1} - x^k\|_X^2. \tag{15.8}$$

Observe that if we had $x^{k+1} = x^k$ (or $F = 0$), this would show the local quadratic growth of $F + G$ at $\widehat{x}$. Since, in general, $x^{k+1} \neq x^k$, we however need to compensate for taking the forward step with respect to F.

Testing (15.7) by the application of $\varphi_k\langle \cdot, x^{k+1} - \widehat{x}\rangle_X$ for some testing parameter $\varphi_k > 0$ and afterwards applying (15.8) yields

$$\frac{\varphi_k\gamma\tau(2-\beta)}{2}\|x^{k+1} - \widehat{x}\|_X^2 - \frac{\varphi_k L\tau}{2\beta}\|x^{k+1} - x^k\|_X^2 + \varphi_k\langle x^{k+1} - x^k, x^{k+1} - \widehat{x}\rangle_X \leq 0.$$

Taking

$$\varphi_{k+1} := \varphi_k(1 + \gamma\tau(2-\beta)) \quad \text{with} \quad \varphi_0 > 0, \tag{15.9}$$

the three-point formula (9.1) yields

$$\frac{\varphi_{k+1}}{2}\|x^{k+1} - \widehat{x}\|_X^2 + \frac{\varphi_k(1 - \tau L/\beta)}{2}\|x^{k+1} - x^k\|_X^2 \leq \frac{\varphi_k}{2}\|x^k - \widehat{x}\|_X^2. \tag{15.10}$$

Since $\beta \in (L\tau, 2)$ and $x^k \in \mathcal{X}$, this implies that $x^{k+1} \in \mathcal{X}$. By induction, we thus obtain that $\{x^k\}_{k\in\mathbb{N}} \subset \mathcal{X}$ under our assumption $x^0 \in \mathcal{X}$.

If $\gamma > 0$, the recursion (15.9) together with $\beta < 2$ shows that φ_k grows exponentially. Using that $\tau L/\beta \leq 1$ and telescoping (15.10) then shows the claimed linear convergence.

Let us then consider weak convergence. With $\gamma = 0$ and $\beta < 2$, the recursion (15.9) reduces to $\varphi_{k+1} \equiv \varphi_0 > 0$. Since $\tau L \leq \beta$, the estimate (15.10) yields Fejér monotonicity of the iterates $\{x^k\}_{k\in\mathbb{N}}$. Moreover, we establish for $w^{k+1} := -\tau^{-1}(x^{k+1} - x^k)$ that $\|w^{k+1}\|_X \to 0$ and $w^{k+1} \in \partial G(x^{k+1}) + \nabla F(x^k)$ for all $k \in \mathbb{N}$. Let $\bar{x}$ be any weak limit point of $\{x^k\}_{k\in\mathbb{N}}$, i.e., there exists a subsequence $\{x^{k_n}\}_{n\in\mathbb{N}}$ with $x^{k_n} \rightharpoonup \bar{x} \in \mathcal{X}$. Then also $x^{k_n+1} \rightharpoonup \bar{x} \in \mathcal{X}$. Since ∇F is by (15.6) Lipschitz continuous in $\mathcal{X}$, we have $\nabla F(x^{k_n+1}) - \nabla F(x^{k_n}) \to 0$. Consequently, $\partial G(x^{k_n+1}) + \nabla F(x^{k_n+1}) \ni w^{k_n+1} + \nabla F(x^{k_n+1}) - \nabla F(x^{k_n}) \to 0$. By the outer semicontinuity of $\partial G + \nabla F$, it follows that $0 \in \partial G(\bar{x}) + \nabla F(\bar{x})$ and therefore $\bar{x} \in (\partial F + \nabla G)^{-1}(0) \subset \mathcal{X}$. The claim thus follows by applying Opial's Lemma 9.1. □

15.2 NONCONVEX PRIMAL-DUAL SPLITTING: ALGORITHM AND ASSUMPTIONS

As mentioned above, we consider the problem (15.1) with $F : X \to \overline{\mathbb{R}}$ and $G : Y \to \overline{\mathbb{R}}$ convex, proper, and lower semicontinuous, and $K \in C^1(X; Y)$. We will soon state more precise assumptions on K. When either the null space of $[\nabla K(x)]^*$ is trivial or $\operatorname{dom} G = X$, we can apply the chain rule Theorem 13.23 for Clarke subdifferentials as well as the equivalences of Theorems 13.5 and 13.8 for convex and differentiable functions, respectively, to rewrite as in Section 8.4 the critical point conditions for this problem as $0 \in H(\widehat{u})$ for the set-valued operator $H : X \times Y \rightrightarrows X \times Y$ defined for $u = (x, y) \in X \times Y$ as

$$H(u) := \begin{pmatrix} \partial F(x) + [\nabla K(x)]^* y \\ \partial G^*(y) - K(x) \end{pmatrix}. \tag{15.11}$$

Throughout the rest of this chapter, we write $\widehat{u} = (\widehat{x}, \widehat{y}) \in H^{-1}(0)$ for an arbitrary root of H that we assume to exist.

In analogy to the basic PDPS method, the basic unaccelerated NL-PDPS method then iterates

$$\begin{cases} x^{k+1} := (I + \tau \partial F)^{-1}(x^k - \tau [\nabla K(x^k)]^* y^k), \\ \bar{x}^{k+1} := (1 + \omega) x^{k+1} - \omega x^k, \\ y^{k+1} := (I + \sigma \partial G^*)^{-1}(y^k + \sigma K(\bar{x}^{k+1})) \end{cases} \tag{15.12}$$

for some acceleration parameter $\tilde{\gamma}_{G^*} \geq 0$ (later to be fixed to be less than the factor of strong convexity of G^*), and where we set the over-relaxation parameter

$$\omega = \frac{1}{1 + 2\tilde{\gamma}_{G^*}\sigma}. \tag{15.13}$$

We can write this algorithm in the general form of Theorem 11.12 as follows. For each iteration $k \in \mathbb{N}$ with some primal and dual testing parameters $\varphi_k, \psi_{k+1} > 0$, we define the step length and testing operators

$$W := \begin{pmatrix} \tau \mathrm{Id} & 0 \\ 0 & \sigma \mathrm{Id} \end{pmatrix} \quad \text{and} \quad Z_{k+1} := \begin{pmatrix} \varphi_k \mathrm{Id} & 0 \\ 0 & \psi_{k+1} \mathrm{Id} \end{pmatrix}.$$

We also define the linear preconditioner M_{k+1} and the step length weighted partial linearization $\widetilde{H}_{k+1}$ of H by

$$M_{k+1} := \begin{pmatrix} \mathrm{Id} & -\tau [\nabla K(x^k)]^* \\ -\omega \sigma \nabla K(x^k) & \mathrm{Id} \end{pmatrix}, \quad \text{and} \tag{15.14}$$

$$\widetilde{H}_{k+1}(u) := W \begin{pmatrix} \partial F(x) + [\nabla K(x^k)]^* y \\ \partial G^*(y) - K(\bar{x}^{k+1}) - \nabla K(x^k)(x - \bar{x}^{k+1}). \end{pmatrix} \tag{15.15}$$

Observe that $\widetilde{H}_{k+1}(u)$ simplifies to $WH(u)$ for linear K. Then (15.12) becomes

$$0 \in \widetilde{H}_{k+1}(u^{k+1}) + M_{k+1}(u^{k+1} - u^k). \tag{15.16}$$

We will need K to be locally Lipschitz differentiable.

Assumption 15.3 (locally Lipschitz ∇K). The operator $K : X \to Y$ is Fréchet differentiable, and for some $L \geq 0$ and a neighborhood $\mathcal{X}_K$ of $\widehat{x}$,

$$\|\nabla K(x) - \nabla K(z)\|_{\mathbb{L}(X,Y)} \leq L\|x - z\|_X \quad (x, z \in \mathcal{X}_K). \tag{15.17}$$

We also require a three-point assumption on K. This assumption combines a second-order growth condition with a three-point smoothness estimate. Note that the factor γ_K can be negative; if it is, it will need to be offset by sufficient strong convexity of F.

Assumption 15.4 (three-point condition on K). For a neighborhood $\mathcal{X}_K$ of $\widehat{x}$, and some $\gamma_K \in \mathbb{R}$ and $L, \theta \geq 0$, we require

$$\begin{aligned} &\langle [\nabla K(z) - \nabla K(\widehat{x})]^* \widehat{y}, x - \widehat{x} \rangle_X \\ &\quad \geq \gamma_K \|x - \widehat{x}\|_X^2 + \theta \|K(\widehat{x}) - K(x) - \nabla K(x)(\widehat{x} - x)\|_Y - \frac{\lambda}{2}\|x - z\|_X^2 \quad (x, z \in \mathcal{X}_K). \end{aligned} \tag{15.18}$$

We observe the following special cases of Assumption 15.4:

(a) For linear K, the assumption trivially holds for any $\gamma_K \leq 0$, $\theta \geq 0$ and $\lambda = 0$.

(b) Let $G^* = \delta_{\{1\}}$, so that $K : X \to \mathbb{R}$ and the problem (15.1) reduces to (15.2) with K in place of F. Minding that in this case $\widehat{y} = 1$, Lemma 15.1 with $\beta = 1$ proves Assumption 15.4 for $\lambda = L$, any $\theta \geq 0$ and $\gamma_K \leq \gamma$ with $\gamma, L \geq 0$ satisfying $\gamma \cdot \mathrm{Id} \leq \nabla^2 K(\zeta) \leq L \cdot \mathrm{Id}$ or all $\zeta \in \mathcal{X}_K$.

In more general settings, the verification of Assumption 15.4 can demand some effort. We refer to [Clason et al., 2019] for examples and to [Clason et al., 2020] for further generalizations.

15.3 NONCONVEX PRIMAL-DUAL SPLITTING: CONVERGENCE PROOF

For simplicity of treatment, and to demonstrate the main ideas without excessive technicalities, we only show linear convergence under strong convexity of both F and G^*.

We will base our proof on Theorem 11.12 and thus have to verify its assumptions. Most of the work is in verifying the inequality (11.28), which we do in several steps. First, we ensure that the operator $Z_{k+1}M_{k+1}$ giving rise to the local metric is self-adjoint. Then we

show that $Z_{k+2}M_{k+2}$ and the update $Z_{k+1}(M_{k+1}+\Xi_{k+1})$ actually performed by the algorithm yield identical norms, where Ξ_{k+1} represents some off-diagonal components from the algorithm as well as any strong convexity provided by F and G^*. Finally, we estimate the local monotonicity of $\widetilde{H}_{k+1}$.

We write $\gamma_F, \gamma_{G^*} \geq 0$ for the factors of (strong) convexity of F and G^*, and recall the factor $\gamma_K \in \mathbb{R}$ from Assumption 15.4 Then for some "acceleration parameters" $\tilde{\gamma}_F, \tilde{\gamma}_{G^*} \geq 0$ and $\kappa \in [0,1)$, we require that

$$\gamma_F + \gamma_K \geq \tilde{\gamma}_F \geq 0, \qquad \gamma_{G^*} \geq \tilde{\gamma}_{G^*} \geq 0, \tag{15.19a}$$

$$\eta_k \coloneqq \varphi_k \tau = \psi_k \sigma, \qquad 1 - \kappa \leq \tau\sigma \|\nabla K(x^k)\|^2, \tag{15.19b}$$

$$\varphi_{k+1} = \varphi_k(1 + 2\tilde{\gamma}_F \tau), \quad \text{and} \qquad \psi_{k+1} = \psi_k(1 + 2\tilde{\gamma}_{G^*}\sigma) \quad (k \in \mathbb{N}). \tag{15.19c}$$

With this, ω defined in (15.13) satisfies

$$\omega = \psi_{k+1}^{-1}\psi_k = \eta_{k+1}^{-1}\eta_k \quad (k \in \mathbb{N}). \tag{15.20}$$

The next lemma adapts Lemma 9.12.

Lemma 15.5. *Fix $k \in \mathbb{N}$ and suppose (15.19) holds. Then $Z_{k+1}M_{k+1}$ is self-adjoint and satisfies*

$$Z_{k+1}M_{k+1} \geq \begin{pmatrix} \delta\varphi_k \cdot \mathrm{Id} & 0 \\ 0 & (\kappa - \delta)(1-\delta)^{-1}\psi_{k+1} \cdot \mathrm{Id} \end{pmatrix} \quad \text{for any } \delta \in [0, \kappa].$$

Proof. From (15.19) and (15.20) we have $\varphi_k \tau = \psi_{k+1}\omega\sigma = \eta_k$. By (15.14) then

$$Z_{k+1}M_{k+1} = \begin{pmatrix} \varphi_k \cdot \mathrm{Id} & -\eta_k[\nabla K(x^k)]^* \\ -\eta_k \nabla K(x^k) & \psi_{k+1} \cdot \mathrm{Id} \end{pmatrix}. \tag{15.21}$$

This shows that $Z_{k+1}M_{k+1}$ is self-adjoint. Furthermore, since Young's inequality followed by (15.19) and (15.20) shows that

$$\begin{aligned} 2\eta_k \langle \nabla K(x^k)\tilde{x}, \tilde{y}\rangle &\leq (1-\delta)\eta_k\tau^{-1}\|\tilde{x}\|^2 + \frac{\eta_k\tau}{1-\delta}\|\nabla K(x^k)^*\tilde{y}\|^2 \\ &= (1-\delta)\varphi_k\|\tilde{x}\|^2 + \psi_{k+1}\omega\frac{\tau\sigma}{1-\delta}\|\nabla K(x^k)^*\tilde{y}\|^2 \quad (\tilde{x} \in X,\ \tilde{y} \in Y), \end{aligned}$$

we obtain from (15.21) that

$$Z_{k+1}M_{k+1} \geq \begin{pmatrix} \delta\varphi_k \mathrm{Id} & 0 \\ 0 & \psi_{k+1}\left(\mathrm{Id} - \omega\frac{\tau\sigma}{1-\delta}\nabla K(x^k)[\nabla K(x^k)]^*\right) \end{pmatrix}. \tag{15.22}$$

The claimed estimate then follows from the assumptions (15.19). □

Our next step is to simplify the operator $Z_{k+1}M_{k+1} - Z_{k+2}M_{k+2}$ occuring in the inequality (11.28) we are trying to prove.

Lemma 15.6. *Fix $k \in \mathbb{N}$, and suppose (15.19) holds. Then $\frac{1}{2}\|\cdot\|^2_{Z_{k+1}(M_{k+1}+\Xi_{k+1})-Z_{k+2}M_{k+2}} = 0$ for*

$$\Xi_{k+1} \coloneqq \begin{pmatrix} 2\tilde{\gamma}_F \tau \mathrm{Id} & 2\tau[\nabla K(x^k)]^* \\ -2\sigma \nabla K(x^{k+1}) & 2\tilde{\gamma}_{G^*}\sigma \mathrm{Id} \end{pmatrix}. \tag{15.23}$$

Proof. Using (15.19) and (15.21) can write

$$Z_{k+1}(M_{k+1} + \Xi_{k+1}) - Z_{k+2}M_{k+2} = D_{k+1}$$

for the skew-symmetric operator

$$D_{k+1} \coloneqq \begin{pmatrix} 0 & [\eta_{k+1}\nabla K(x^{k+1}) + \eta_k \nabla K(x^k)]^* \\ -[\eta_{k+1}\nabla K(x^{k+1}) + \eta_k \nabla K(x^k)] & 0 \end{pmatrix}.$$

This yields the claim. □

For our convergence claim, we need to assume that the dual variables stay bounded within the "nonlinear range" of K. To this end, we introduce the (possibly empty) subspace Y_{L} of Y in which K acts linearly, i.e.,

$$Y_{\mathrm{L}} \coloneqq \{y \in Y \mid \text{the mapping } x \mapsto \langle y, K(x)\rangle \text{ is linear}\} \quad \text{and} \quad Y_{\mathrm{NL}} \coloneqq Y_{\mathrm{L}}^{\perp}.$$

We then denote by P_{NL} the orthogonal projection to Y_{NL}. We also write

$$\mathbb{B}_{\mathrm{NL}}(\widehat{y}, r) \coloneqq \{y \in Y \mid \|y - \widehat{y}\|_{P_{\mathrm{NL}}} \leq r\}$$

for the closed cylinder in Y of the radius r with axis orthogonal to Y_{NL}.

With $\mathcal{X}_K$ given by Assumption 15.3, we now define for some radius $\rho_y > 0$ the neighborhood

$$\mathcal{U}(\rho_y) \coloneqq \mathcal{X}_K \times \mathbb{B}_{\mathrm{NL}}(\widehat{y}, \rho_y). \tag{15.24}$$

We will require that the iterates $\{u^k\}_{k\in\mathbb{N}}$ of (15.12) stay within this neighborhood for some fixed $\rho_y > 0$.

The next lemma provides the necessary three-point inequality to estimate the linearizations performed within $\widetilde{H}_{k+1}$.

Lemma 15.7. *For a fixed $k \in \mathbb{N}$, suppose $\bar{x}^{k+1} \in \mathcal{X}_K$, and let $\rho_y \geq 0$ be such that $u^k, u^{k+1} \in \mathcal{U}(\rho_y)$. Suppose K satisfies Assumptions 15.3 and 15.4 with $\omega\theta \geq \rho_y$. If (15.19) holds, then*

$$\langle \widetilde{H}_{k+1}(u^{k+1}), u^{k+1} - \widehat{u}\rangle_{Z_{k+1}} \geq \frac{1}{2}\|u^{k+1} - \widehat{u}\|^2_{Z_{k+1}\Xi_{k+1}} - \frac{\eta_k[\lambda + 3L\rho_y]}{2}\|x^{k+1} - x^k\|_X^2.$$

Proof. From (15.11), (15.15), (15.19), and (15.23), we calculate

$$
\begin{aligned}
(15.25)\qquad D &\coloneqq \langle \widetilde{H}_{k+1}(u^{k+1}), u^{k+1} - \widehat{u}\rangle_{Z_{k+1}} - \frac{1}{2}\|u^{k+1} - \widehat{u}\|^2_{Z_{k+1}\Xi_{k+1}} \\
&= \langle H(u^{k+1}), u^{k+1} - \widehat{u}\rangle_{Z_{k+1}W} - \eta_k\tilde{\gamma}_F\|x^{k+1} - \widehat{x}\|_X^2 - \eta_{k+1}\tilde{\gamma}_{G^*}\|y^{k+1} - \widehat{y}\|_Y^2 \\
&\quad + \eta_k\langle[\nabla K(x^k) - \nabla K(x^{k+1})](x^{k+1} - \widehat{x}), y^{k+1}\rangle_Y \\
&\quad + \eta_{k+1}\langle K(x^{k+1}) - K(\bar{x}^{k+1}) - \nabla K(x^k)(x^{k+1} - \bar{x}^{k+1}), y^{k+1} - \widehat{y}\rangle_Y \\
&\quad + \langle(\eta_{k+1}\nabla K(x^{k+1}) - \eta_k\nabla K(x^k))(x^{k+1} - \widehat{x}), y^{k+1} - \widehat{y}\rangle_Y.
\end{aligned}
$$

Here the first of the terms involving K comes from the first lines of $\widetilde{H}_{k+1}$ and H, the second of the terms from the second line, and the third from Ξ_{k+1}. Since $0 \in H(\widehat{u})$, we have $q_F \coloneqq -[\nabla K(\widehat{x})]^*\widehat{y} \in \partial F(\widehat{x})$ and $q_{G^*} \coloneqq K(\widehat{x}) \in \partial G^*(\widehat{y})$. Using (15.19), we can therefore expand

$$
\begin{aligned}
\langle H(u^{k+1}), u^{k+1} - \widehat{u}\rangle_{Z_{k+1}W} &= \eta_k\langle\partial F(x^{k+1}) - q_F, x^{k+1} - \widehat{x}\rangle_X \\
&\quad + \eta_{k+1}\langle\partial G^*(y^{k+1}) - q_{G^*}, y^{k+1} - \widehat{y}\rangle_Y \\
&\quad + \eta_k\langle[\nabla K(x^{k+1})]^* y^{k+1} - [\nabla K(\widehat{x})]^*\widehat{y}, x^{k+1} - \widehat{x}\rangle_X \\
&\quad + \eta_{k+1}\langle K(\widehat{x}) - K(x^{k+1}), y^{k+1} - \widehat{y}\rangle_Y.
\end{aligned}
$$

Using the γ_F-strong monotonicity of ∂F and the γ_{G^*}-strong monotonicity of ∂G^*, and rearranging terms, we obtain

$$
\begin{aligned}
\langle H(u^{k+1}), u^{k+1} - \widehat{u}\rangle_{Z_{k+1}W} &\geq \eta_k\gamma_F\|x^{k+1} - \widehat{x}\|_X^2 + \eta_{k+1}\gamma_{G^*}\|y^{k+1} - \widehat{y}\|_Y^2 \\
&\quad + \eta_k\langle\nabla K(x^{k+1})(x^{k+1} - \widehat{x}), y^{k+1}\rangle_Y \\
&\quad - \eta_k\langle\nabla K(\widehat{x})(x^{k+1} - \widehat{x}), \widehat{y}\rangle_Y + \eta_{k+1}\langle K(\widehat{x}) - K(x^{k+1}), y^{k+1} - \widehat{y}\rangle_Y.
\end{aligned}
$$

Combining this estimate with (15.25) and rearranging terms, we obtain

$$
\begin{aligned}
D &\geq \eta_k(\gamma_F - \tilde{\gamma}_F)\|x^{k+1} - \widehat{x}\|_X^2 + \eta_{k+1}(\gamma_{G^*} - \tilde{\gamma}_{G^*})\|y^{k+1} - \widehat{y}\|_Y^2 \\
&\quad - \eta_k\langle\nabla K(\widehat{x})(x^{k+1} - \widehat{x}), \widehat{y}\rangle_Y + \eta_k\langle\nabla K(x^k)(x^{k+1} - \widehat{x}), y^{k+1}\rangle_Y \\
&\quad + \eta_{k+1}\langle K(\widehat{x}) - K(\bar{x}^{k+1}) - \nabla K(x^k)(x^{k+1} - \bar{x}^{k+1}), y^{k+1} - \widehat{y}\rangle_Y \\
&\quad + \langle(\eta_{k+1}\nabla K(x^{k+1}) - \eta_k\nabla K(x^k))(x^{k+1} - \widehat{x}), y^{k+1} - \widehat{y}\rangle_Y.
\end{aligned}
$$

Further rearrangements and $\gamma_F + \gamma_K \geq \tilde{\gamma}_F$ and $\gamma_{G^*} \geq \tilde{\gamma}_{G^*}$ give

$$
\begin{aligned}
(15.26)\qquad D &\geq -\eta_k\gamma_K\|x^{k+1} - \widehat{x}\|_X^2 + \eta_k\langle[\nabla K(x^k) - \nabla K(\widehat{x})](x^{k+1} - \widehat{x}), \widehat{y}\rangle_Y \\
&\quad + \eta_{k+1}\langle K(\widehat{x}) - K(x^{k+1}) - \nabla K(x^{k+1})(\widehat{x} - x^{k+1}), y^{k+1} - \widehat{y}\rangle_Y \\
&\quad + \eta_{k+1}\langle K(x^{k+1}) - K(\bar{x}^{k+1}) + \nabla K(x^{k+1})(\bar{x}^{k+1} - x^{k+1}), y^{k+1} - \widehat{y}\rangle_Y \\
&\quad + \eta_{k+1}\langle(\nabla K(x^k) - \nabla K(x^{k+1}))(\bar{x}^{k+1} - x^{k+1}), y^{k+1} - \widehat{y}\rangle_Y.
\end{aligned}
$$

Using Assumption 15.3 and the mean value theorem in the form

$$K(x') = K(x) + \nabla K(x)(x' - x) + \int_0^1 (\nabla K(x + s(x' - x)) - \nabla K(x))(x' - x)ds,$$

we obtain for any $x, x' \in \mathcal{X}_K$ and $y \in Y$ the inequality

$$\langle K(x') - K(x) - \nabla K(x)(x' - x), y\rangle_Y \leq (L/2)\|x - x'\|_X^2 \|y\|_{P_{\mathrm{NL}}}. \tag{15.27}$$

Applying Assumption 15.3, the inequality (15.27), and $\bar{x}^{k+1} - x^{k+1} = \omega(x^{k+1} - x^k)$ to the last two terms of (15.26), we obtain

$$\langle K(x^{k+1}) - K(\bar{x}^{k+1}) + \nabla K(x^{k+1})(\bar{x}^{k+1} - x^{k+1}), y^{k+1} - \widehat{y}\rangle_Y \geq -\frac{L\omega^2}{2}\|x^{k+1} - x^k\|^2 \|y^{k+1} - \widehat{y}\|_{P_{\mathrm{NL}}}$$

and

$$\langle (\nabla K(x^k) - \nabla K(x^{k+1}))(\bar{x}^{k+1} - x^{k+1}), y^{k+1} - \widehat{y}\rangle_Y \geq -L\omega\|x^{k+1} - x^k\|_X^2 \|y^{k+1} - \widehat{y}\|_{P_{\mathrm{NL}}}.$$

These estimates together with (15.19) and $u^{k+1} \in \mathcal{U}(\rho_y)$ now imply that $D \geq \eta_k D_{k+1}^K$ for

$$\begin{aligned} D_{k+1}^K := \langle [\nabla K(x^k) - \nabla K(\widehat{x})](x^{k+1} - \widehat{x}), \widehat{y}\rangle_Y - \gamma_K \|x^{k+1} - \widehat{x}\|_X^2 - L(1 + \omega/2)\rho_y \|x^{k+1} - x^k\|_X^2 \\ - \omega^{-1}\|y^{k+1} - \widehat{y}\|_{P_{\mathrm{NL}}} \|K(\widehat{x}) - K(x^{k+1}) - \nabla K(x^{k+1})(\widehat{x} - x^{k+1})\|_Y. \end{aligned}$$

Finally, we use Assumption 15.4 and Young's inequality to estimate

$$\begin{aligned} D_{k+1}^K \geq (\theta - \omega^{-1}\|y^{k+1} - \widehat{y}\|_{P_{\mathrm{NL}}})\|K(\widehat{x}) - K(x^{k+1}) - \nabla K(x^{k+1})(\widehat{x} - x^{k+1})\|_Y \\ - \frac{\lambda + 3L\rho_y}{2}\|x^{k+1} - x^k\|_X^2. \end{aligned}$$

Now observe that $\theta - \omega^{-1}\|y^{k+1} - \widehat{y}\|_{P_{\mathrm{NL}}} \geq \theta - \omega^{-1}\rho_y \geq 0$. Combining with the estimate $D \geq \eta_k D_{k+1}^K$, we therefore obtain our claim. □

We now have all the necessary tools at hand to prove the main estimate (11.28) needed for the application of Theorem 11.12.

Theorem 15.8. *Let $F : X \to \overline{\mathbb{R}}$ and $G : Y \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous. Suppose $K : X \to Y$ satisfies Assumptions 15.3 and 15.4. Fix $k \in \mathbb{N}$, and also suppose $\bar{x}^{k+1} \in \mathcal{X}_K$ and that $u^k, u^{k+1} \in \mathcal{U}(\rho_y)$ for some $\rho_y \geq 0$ Suppose (15.19) holds for some $\kappa \in [0, 1)$ and*

$$\tau < \frac{\kappa}{\lambda + 3L\rho_y} \tag{15.28}$$

as well as $\omega\theta \geq \rho_y$. Then

$$\frac{1}{2}\|u^{k+1} - \widehat{u}\|_{Z_{k+2}M_{k+2}}^2 \leq \frac{1}{2}\|u^k - \widehat{u}\|_{Z_{k+1}M_{k+1}}^2 \quad (k \in \mathbb{N}). \tag{15.29}$$

Proof. We show that (15.30) holds with $\mathcal{V}_{k+1} \equiv 0$, i.e., that

$$\langle \widetilde{H}_{k+1}(u^{k+1}), u^{k+1} - \widehat{u} \rangle_{Z_{k+1}} \geq \frac{1}{2}\|u^{k+1} - \widehat{u}\|^2_{Z_{k+2}M_{k+2} - Z_{k+1}M_{k+1}} - \frac{1}{2}\|u^{k+1} - u^k\|^2_{Z_{k+1}M_{k+1}}. \tag{15.30}$$

The claim then follows from Theorem 11.12 and Lemma 15.5, the latter of which provides the necessary self-adjointness of $Z_{k+1}M_{k+1}$.

Let thus $\delta \in (0, \kappa)$ be arbitrary, and define

$$S_{k+1} := \begin{pmatrix} (\delta\varphi_k - \eta_k[\lambda + 3L\rho_y])\mathrm{Id} & 0 \\ 0 & \psi_{k+1}\left(\mathrm{Id} - \omega\frac{\tau\sigma}{1-\delta}\nabla K(x^k)[\nabla K(x^k)]^*\right) \end{pmatrix}.$$

Using (15.22) and (15.21) and, in the second and third step, Lemmas 15.6 and 15.7, we estimate

$$\begin{aligned} \frac{1}{2}\|u^{k+1} - u^k\|^2_{S_{k+1} - Z_{k+1}M_{k+1}} &\leq -\frac{\eta_k[\lambda + 3L\rho_y]}{2}\|x^{k+1} - x^k\|^2_X \\ &\leq \langle \widetilde{H}_{k+1}(u^{k+1}), u^{k+1} - \widehat{u} \rangle_{Z_{k+1}} - \frac{1}{2}\|u^{k+1} - \widehat{u}\|^2_{Z_{k+1}\Xi_{k+1}} \\ &= \langle \widetilde{H}_{k+1}(u^{k+1}), u^{k+1} - \widehat{u} \rangle_{Z_{k+1}} - \frac{1}{2}\|u^{k+1} - \widehat{u}\|^2_{Z_{k+2}M_{k+2} - Z_{k+1}M_{k+1}}. \end{aligned}$$

Then (15.30) holds if $S_{k+1} \geq 0$. This readily follows from (15.28) with $\delta \in (0, \kappa)$ for the primal variable and (15.19) for the dual variable. □

Theorem 15.9. *Let $F : X \to \overline{\mathbb{R}}$ and $G : Y \to \overline{\mathbb{R}}$ be strongly convex, proper, and lower semicontinuous. Suppose $K : X \to Y$ satisfies Assumptions 15.3 and 15.4. Let $R_K > 0$ be such that $\sup_{x \in \mathcal{X}_K} \|\nabla K(x)\| \leq R_K$. Pick $0 < \tau < 1/(\lambda + 3L\rho_y)$ for a given $\rho_y \geq 0$, and take $\sigma = \tau\tilde{\gamma}_F/\tilde{\gamma}_{G^*}$ for some $\tilde{\gamma}_F \in (0, \gamma_F + \gamma_K]$ and $\tilde{\gamma}_{G^*} \in (0, \gamma_G]$ such that $\omega\theta \geq \rho_y$. Let the iterates $\{(u^k, \bar{x}^{k+1})\}_{k\in\mathbb{N}}$ be generated by the NL-PDPS method (15.12). If $\bar{x}^{k+1} \in \mathcal{X}_K$ and $u^k \in \mathcal{U}(\rho_y)$ for all $k \in \mathbb{N}$ and some $\widehat{u} \in H^{-1}(0)$ for H given in (15.11), then $u^k \to \widehat{u}$ linearly.*

Proof. Take $\varphi_{k+1} := \varphi_k(1 + 2\tilde{\gamma}_F\tau)$ and $\psi_{k+1} := \psi_k(1 + 2\tilde{\gamma}_{G^*}\sigma)$ for $\varphi_0 = 1$ and $\psi_1 := \tau/\sigma$. Then $\varphi_k\tau = \psi_{k+1}\sigma$ if and only if $1 + 2\tilde{\gamma}_F\tau = 1 + 2\tilde{\gamma}_{G^*}\sigma$, i.e., for $\sigma = \tau\tilde{\gamma}_F/\tilde{\gamma}_{G^*}$ as stated. Consequently (15.19) is satisfied and the testing parameters φ_k and ψ_{k+1} grow exponentially. Clearly (15.28) holds for some $\kappa \in [0, 1)$. Combining (15.29) from Theorem 15.8 with Lemma 15.5 now shows the claimed linear convergence. □

Besides step length bounds and structural properties of the problem, Theorem 15.9 still requires us to ensure that the iterates stay close enough to the critical point $\widehat{x}$. This can be done if we initialize close enough to a critical point. As the proof is very technical, we merely state the following result.

Theorem 15.10 ([Clason et al., 2019, Proposition 4.8]). *Under the assumptions of Theorem 15.9, for any $\rho_y > 0$ there exists an $\varepsilon > 0$ such that $\{u^k\}_{k\in\mathbb{N}} \subset \mathcal{U}(\rho_y)$ for all initial iterates $u^0 = (x^0, y^0)$ satisfying*

$$\sqrt{2\delta^{-1}(\|x^0 - \widehat{x}\|^2 + \tau\sigma^{-1}\|y^0 - \widehat{y}\|^2)} \leq \varepsilon. \tag{15.31}$$

Remark 15.11 (weaker assumptions, weaker convergence). We have only demonstrated linear convergence of the method under the strong convexity of both F and G^*. However, under similarly lesser assumptions as for the basic PDPS method familiar from Part II, both an accelerated $O(1/N^2)$ rate and weak convergence can be proved. We refer to [Clason et al., 2019] for details, noting that Opial's Lemma 9.1 extends straightforwardly to the quantitative Fejér monotonicity (10.21) that is the basis of our proofs here. We also note that our linear convergence result differs from that in [Clason et al., 2019] by taking the over-relaxation parameter $\omega = 1$ in (15.12) instead of $\omega = 1/(1 + 2\tilde{\gamma}_F\tau) < 1$; compare Theorem 10.8.

Remark 15.12 (historical development of the NL-PDPS). The NL-PDPS method was first introduced in [Valkonen, 2014] in finite dimensions with applications to inverse problems in magnetic resonance imaging. The method was later extended in [Clason and Valkonen, 2017a] to infinite dimensions and applied to PDE-constrained optimization problems. In these works, only (weak) convergence of the iterates is shown, based on the metric regularity of the operator H. We discuss metric regularity later in Chapters 27 and 28. Convergence rates were then first shown in [Clason et al., 2019]. In that paper, alternative forms of the three-point condition Assumption 15.4 on K are also discussed.

Similarly to how we showed in Section 8.7 that the preconditioned ADMM is equivalent to the PDPS method, it is possible to derive a preconditioned nonlinear ADMM that is equivalent to the NL-PDPS method; such algorithms are considered in [Benning et al., 2016]. The NL-PDPS method has been extended in [Clason et al., 2020] by replacing $\langle K(x), y\rangle_Y$ by a general saddle term $K(x, y)$, which can be applied to nonconvex optimization problems such as ℓ^0-TV denoising or elliptic Nash equilibrium problems. Block-adapted and stochastic variants in the spirit of Remark 11.17 can be found in [Mazurenko et al., 2020]. Finally, a simplified approach using the Bregman divergences of Section 11.1 is presented in [Valkonen, 2021a].

16 LIMITING SUBDIFFERENTIALS

While the Clarke subdifferential is a suitable concept for nonsmooth but convex or nonconvex but smooth functionals, it has severe drawbacks for nonsmooth *and* nonconvex functionals: As shown in Corollary 13.11, its Fermat principle cannot distinguish minimizers from maximizers. The reason is that the Clarke subdifferential is always convex, which is a direct consequence of its construction (13.2) via polarity with respect to (generalized) directional derivatives. To obtain sharper results for such functionals, it is therefore necessary to construct *nonconvex* subdifferentials directly via a *dual* limiting process. On the other hand, deriving calculus rules for the previous subdifferentials crucially exploited their convexity by applying Hahn–Banach separation theorems, and calculus rules for nonconvex subdifferentials are thus significantly more difficult to obtain. As in Chapter 13, we will assume throughout this chapter that X is a Banach space unless stated otherwise.

16.1 BOULIGAND SUBDIFFERENTIALS

The first definition is motivated by Theorem 13.26: We *define* a subdifferential as a suitable limit of classical derivatives (without convexification). For $F : X \to \overline{\mathbb{R}}$, we first define the *set of Gâteaux points*

$$G_F := \{x \in X \mid F \text{ is Gâteaux differentiable at } x\} \subset \operatorname{dom} F$$

and then the *Bouligand subdifferential* of F at x as

$$\partial_B F(x) := \{x^* \in X^* \mid DF(x_n) \overset{*}{\rightharpoonup} x^* \text{ for some } G_F \ni x_n \to x\}\,. \tag{16.1}$$

For $F : \mathbb{R}^N \to \mathbb{R}$ locally Lipschitz, it follows from Theorem 13.26 that $\partial_C F(x) = \operatorname{co} \partial_B F(x)$. However, unless X is finite-dimensional, it is not clear a priori that the Bouligand subdifferential is nonempty even for $x \in \operatorname{dom} F$.[1] Furthermore, the subdifferential does not admit a satisfactory calculus; not even a Fermat principle holds.

[1] Although in special cases it is possible to give a full characterization in Hilbert spaces; see, e.g., [Christof et al., 2018].

Example 16.1. Let $F : \mathbb{R} \to \mathbb{R}$, $F(x) := |x|$. Then F is differentiable at every $x \neq 0$ with $F'(x) = \operatorname{sign}(x)$. Correspondingly,

$$0 \notin \{-1, 1\} = \partial_B F(0).$$

To make this approach work therefore requires a more delicate limiting process. The remainder of this chapter is devoted to one such approach, where we only give an overview and state important results following [Mordukhovich, 2006]. The full theory is based on a geometric construction similar to Lemma 4.10 making use of tangent and normal cones (corresponding to generalized directional derivatives and subgradients, respectively) that also allows for differentiation of set-valued mappings. We will develop this theory in Chapters 18 to 21. For an alternative, more axiomatic, approach to generalized derivatives of nonconvex functionals, we refer to [Ioffe, 2017; Penot, 2013].

16.2 FRÉCHET SUBDIFFERENTIALS

We begin with the following limiting construction, which combines the characterizations of both the Fréchet derivative and the convex subdifferential. Let X be a Banach space and $F : X \to \overline{\mathbb{R}}$. The *Fréchet subdifferential* (or *regular subdifferential* or *presubdifferential*) of F at x is then defined as[2]

$$\partial_F F(x) := \left\{ x^* \in X^* \;\middle|\; \liminf_{y \to x} \frac{F(y) - F(x) - \langle x^*, y - x \rangle_X}{\|y - x\|_X} \geq 0 \right\}. \tag{16.2}$$

Note how this "localizes" the definition of the convex subdifferential around the point of interest: the numerator does not need to be nonnegative for all y; it suffices if this holds for any y sufficiently close to x. By a similar argument as for Theorem 4.2, we thus obtain a Fermat principle for *local* minimizers.

Theorem 16.2. *Let $F : X \to \overline{\mathbb{R}}$ be proper and $\bar{x} \in \operatorname{dom} F$ be a local minimizer. Then $0 \in \partial_F F(\bar{x})$.*

Proof. Let $\bar{x} \in \operatorname{dom} F$ be a local minimizer. Then there exists an $\varepsilon > 0$ such that $F(\bar{x}) \leq F(y)$ for all $y \in \mathbb{O}(\bar{x}, \varepsilon)$, which is equivalent to

$$\frac{F(y) - F(\bar{x}) - \langle 0, y - \bar{x} \rangle_X}{\|y - \bar{x}\|_X} \geq 0 \quad \text{for all } y \in \mathbb{O}(\bar{x}, \varepsilon).$$

Now for any strongly convergent sequence $y_n \to \bar{x}$, we have that $y_n \in \mathbb{O}(\bar{x}, \varepsilon)$ for n large enough. Taking the lim inf in the above inequality thus yields $0 \in \partial_F F(\bar{x})$. □

[2] The equivalence of (16.2) with the usual definition based on corresponding normal cones follows from, e.g., [Mordukhovich, 2006, Theorem 1.86].

For convex functionals, of course, the numerator is always nonnegative by definition, and the Fréchet subdifferential reduces to the convex subdifferential.

Theorem 16.3. *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous and $x \in \operatorname{dom} F$. Then $\partial_F F(x) = \partial F(x)$.*

Proof. By definition of the convex subdifferential, any $x^* \in \partial F(x)$ satisfies

$$F(y) - F(x) - \langle x^*, y - x\rangle_X \geq 0 \quad \text{for all } y \in X.$$

Dividing by $\|x-y\|_X > 0$ for $y \neq x$ and taking the lim inf as $y \to x$ thus yields $x^* \in \partial_F F(x)$.

Conversely, let $x^* \in \partial_F F(x)$ and $h \in X \setminus \{0\}$ be arbitrary. Then for an $\delta > 0$, there exists an $\varepsilon > 0$ such that

$$\frac{F(x+th) - F(x) - \langle x^*, th\rangle_X}{t\|h\|_X} \geq -\delta \quad \text{for all } t \in (0, \varepsilon).$$

Multiplying by $\|h\|_X > 0$ and letting $t \to 0$, we obtain from Lemma 4.3 that

$$\langle x^*, h\rangle_X \leq \frac{F(x+th) - F(x)}{t} + \delta \to F'(x; h) + \delta. \tag{16.3}$$

Since $\delta > 0$ was arbitrary, this implies by Lemma 4.4 that $x^* \in \partial F(x)$. □

Similarly, for Fréchet differentiable functionals, the limit in (16.2) is zero for all sequences.

Theorem 16.4. *Let $F : X \to \mathbb{R}$ be Fréchet differentiable at $x \in X$. Then $\partial_F F(x) = \{F'(x)\}$.*

Proof. The definition of the Fréchet derivative immediately yields

$$\lim_{y \to x} \frac{F(y) - F(x) - \langle F'(x), y - x\rangle_X}{\|x - y\|_X} = \lim_{\|h\|_X \to 0} \frac{F(x+h) - F(x) - F'(x)h}{\|h\|_X} = 0$$

and hence $F'(x) \in \partial_F F(x)$.

Conversely, let $x^* \in \partial_F F(x)$ and let again $h \in X \setminus \{0\}$ be arbitrary. As in the proof of Theorem 16.3, we then obtain that

$$\langle x^*, h\rangle_X \leq F'(x; h) = \langle F'(x), h\rangle_X. \tag{16.4}$$

Applying the same argument to $-h$ then yields $\langle x^*, h\rangle_X = \langle F'(x), h\rangle_X$ for all $h \in X$, i.e., $x^* = F'(x)$. □

For nonsmooth and nonconvex functionals, the Fréchet subdifferential can be strictly smaller than the Clarke subdifferential.

Example 16.5. Consider $F : \mathbb{R} \to \mathbb{R}$, $F(x) := -|x|$. For any $x \neq 0$, it follows from Theorem 16.4 that $\partial_F F(x) = \{-\operatorname{sign}(x)\}$. But for $x = 0$ and arbitrary $x^* \in \mathbb{R}$, we have that

$$\liminf_{y \to 0} \frac{F(y) - F(0) - \langle x^*, y - 0 \rangle}{|y - 0|} = \liminf_{y \to 0} (-1 - x^* \cdot \operatorname{sign}(y)) = -1 - |x^*| < 0$$

and hence that

$$\partial_F F(0) = \emptyset \subsetneq [-1, 1] = \partial_C F(0).$$

Note that $0 \in \operatorname{dom} F$ in this example. Although the Fréchet subdifferential does not pick up a maximizer in contrast to the Clarke subdifferential, the fact that $\partial_F F(x)$ can be empty even for $x \in \operatorname{dom} F$ is a problem when trying to derive calculus rules that hold with equality. In fact, as Example 16.5 shows, the Fréchet subdifferential fails to be outer semicontinuous, which is also not desirable. This leads to the next and final definition.

16.3 MORDUKHOVICH SUBDIFFERENTIALS

Let X be a reflexive Banach space and $F : X \to \overline{\mathbb{R}}$. The *Mordukhovich subdifferential* (or *basic subdifferential* or *limiting subdifferential*) of F at $x \in \operatorname{dom} F$ is then defined as the strong-to-weak* outer closure of $\partial_F F(x)$, i.e.,[3]

$$\begin{aligned} \partial_M F(x) &:= \text{w-}*\text{-}\limsup_{y \to x} \partial_F F(y) \\ &= \left\{ x^* \in X^* \;\middle|\; x_n^* \overset{*}{\rightharpoonup} x^* \text{ for some } x_n^* \in \partial_F F(x_n) \text{ with } x_n \to x \right\}, \end{aligned} \tag{16.5}$$

which can be seen as a generalization of the definition (16.1) of the Bouligand subdifferential. Note that in contrast to (16.1), this definition includes the constant sequence $x_n^* \equiv x^*$ even at nondifferentiable points, which makes this a more useful concept in general. This also implies that $\partial_F F(x) \subset \partial_M F(x)$ for any F, and Theorem 16.2 immediately yields a Fermat principle.

Corollary 16.6. *Let $F : X \to \overline{\mathbb{R}}$ be proper and $\bar{x} \in \operatorname{dom} F$ be a local minimizer. Then $0 \in \partial_M F(\bar{x})$.*

As for the Fréchet subdifferential, maximizers do not satisfy the Fermat principle.

Example 16.7. Consider again $F : \mathbb{R} \to \mathbb{R}$, $F(x) := -|x|$. Using Example 16.5, we directly

[3] The equivalence of this definition with the original geometric definition – which holds in *reflexive* Banach spaces – follows from [Mordukhovich, 2006, Theorem 2.34].

obtain from (16.5) that $\partial_M F(0) = \{-1, 1\} = \partial_B F(0)$.

Since the convex subdifferential is strong-to-weak* outer semicontinuous, the Mordukhovich subdifferential reduces to the convex subdifferential as well.

Theorem 16.8. *Let X be a reflexive Banach space, $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $x \in \operatorname{dom} F$. Then $\partial_M F(x) = \partial F(x)$.*

Proof. From Theorem 16.3, it follows that $\partial F(x) = \partial_F F(x) \subset \partial_M F(x)$. Let therefore $x^* \in \partial_M F(x)$ be arbitrary. Then by definition there exists a sequence $\{x_n^*\}_{n\in\mathbb{N}} \subset X^*$ with $x_n^* \overset{*}{\rightharpoonup} x^*$ and $x_n^* \in \partial_F F(x_n) = \partial F(x_n)$ for $x_n \to x$. From Theorem 6.13 and Lemma 6.10, it then follows that $x^* \in \partial F(x)$ as well. □

A similar result holds for *continuously* differentiable functionals.

Theorem 16.9. *Let X be a reflexive Banach space and $F : X \to \overline{\mathbb{R}}$ be continuously differentiable at $x \in X$. Then $\partial_M F(x) = \{F'(x)\}$.*

Proof. From Theorem 16.3, it follows that $\{F'(x)\} = \partial_F F(x) \subset \partial_M F(x)$. Let therefore $x^* \in \partial_M F(x)$ be arbitrary. Then by definition there exists a sequence $\{x_n^*\}_{n\in\mathbb{N}} \subset X^*$ with $x_n^* \overset{*}{\rightharpoonup} x^*$ and $x_n^* \in \partial_F F(x_n) = \{F'(x_n)\}$ for $x_n \to x$. The continuity of F' then immediately implies that $F'(x_n) \to F'(x)$, and since strong limits are also weak-$*$ limits, we obtain $x^* = F'(x)$. □

The same function as in Example 13.6 shows that this equality does not hold if F is merely Fréchet differentiable.

We also have the following relation to Clarke subdifferentials, which should be compared to Theorem 13.26. We will give a proof in a more restricted setting in Chapter 20, cf. Corollary 20.21.

Theorem 16.10 ([Mordukhovich, 2006, Theorem 3.57]). *Let X be a reflexive Banach space and $F : X \to \mathbb{R}$ be locally Lipschitz continuous around $x \in X$. Then $\partial_C F(x) = \operatorname{cl}^* \operatorname{co} \partial_M F(x)$, where $\operatorname{cl}^* A$ stands for the weak-$*$ closure of the set $A \subset X^*$.*[4]

The following example illustrates that the Mordukhovich subdifferential can be nonconvex.

[4] Of course, in reflexive Banach spaces the weak-$*$ closure coincides with the weak closure. The statement holds more general in so-called *Asplund spaces* which include some nonreflexive Banach spaces.

Example 16.11. Let $F : \mathbb{R}^2 \to \mathbb{R}$, $F(x_1, x_2) = |x_1| - |x_2|$. Since F is continuously differentiable for any (x_1, x_2) where $x_1, x_2 \neq 0$ with

$$\nabla F(x_1, x_2) \in \{(1, 1), (-1, 1), (1, -1), (-1, -1)\},$$

we obtain from (16.2) that

$$\partial_F F(x_1, x_2) = \begin{cases} \{(1, -1)\} & \text{if } x_1 > 0, x_2 > 0, \\ \{(-1, -1)\} & \text{if } x_1 < 0, x_2 > 0, \\ \{(-1, 1)\} & \text{if } x_1 < 0, x_2 < 0, \\ \{(1, 1)\} & \text{if } x_1 > 0, x_2 < 0, \\ \{(t, -1) \mid t \in [-1, 1]\} & \text{if } x_1 = 0, x_2 > 0, \\ \{(t, 1) \mid t \in [-1, 1]\} & \text{if } x_1 = 0, x_2 < 0, \\ \emptyset & \text{if } x_2 = 0. \end{cases}$$

In particular, $\partial_F F(0, 0) = \emptyset$. However, from (16.5) it follows that

$$\partial_M F(0, 0) = \{(t, -1) \mid t \in [-1, 1]\} \cup \{(t, 1) \mid t \in [-1, 1]\}.$$

In particular, $0 \notin \partial_M F(0, 0)$. On the other hand, Theorem 16.10 then yields that

$$\partial_C F(0, 0) = \{(t, s) \mid t, s \in [-1, 1]\} = [-1, 1]^2 \tag{16.6}$$

and hence $0 \in \partial_C F(0, 0)$. (Note that F attains neither a minimum nor a maximum on $\mathbb{R}^2$, while $(0, 0)$ is a nonsmooth saddle-point.)

In contrast to the Bouligand subdifferential, the Mordukhovich subdifferential admits a satisfying calculus, although the assumptions are understandably more restrictive than in the convex setting. The first rule follows as always straight from the definition.

Theorem 16.12. *Let X be a reflexive Banach space and $F : X \to \overline{\mathbb{R}}$. Then for any $\lambda \geq 0$ and $x \in X$,*

$$\partial_M(\lambda F)(x) = \lambda \partial_M F(x).$$

Full calculus in infinite-dimensional spaces holds only for a rather small class of mappings.

Theorem 16.13 ([Mordukhovich, 2006, Proposition 1.107]). *Let X be a reflexive Banach space, $F : X \to \mathbb{R}$ be continuously differentiable, and $G : X \to \overline{\mathbb{R}}$ be arbitrary. Then for any $x \in \operatorname{dom} G$,*

$$\partial_M(F + G)(x) = \{F'(x)\} + \partial_M G(x).$$

While the previous two theorems also hold for the Fréchet subdifferential (the latter even for merely Fréchet differentiable F), the following chain rule is only valid for the Mordukhovich subdifferential. Compared to Theorem 13.23, it also allows for the outer functional to be extended-real valued.

Theorem 16.14 ([Mordukhovich, 2006, Proposition 1.112]). *Let X be a reflexive Banach space, $F : X \to Y$ be continuously differentiable, and $G : Y \to \overline{\mathbb{R}}$ be arbitrary. Then for any $x \in X$ with $F(x) \in \operatorname{dom} G$ and $F'(x) : X \to Y$ surjective,*

$$\partial_M (G \circ F)(x) = F'(x)^* \partial_M G(F(x)).$$

More general calculus rules require X to be a reflexive Banach[5] space as well as additional, nontrivial, assumptions on F and G; see, e.g., [Mordukhovich, 2006, Theorem 3.36 and Theorem 3.41].

We will illustrate how to prove the above calculus results and more in Section 20.4 and Chapter 25, after studying the differentiation of set-valued mappings.

[5] or Asplund

17 ε-SUBDIFFERENTIALS AND APPROXIMATE FERMAT PRINCIPLES

We now study an approximate variant of the Fréchet subdifferential of Section 16.2 as well as related approximate Fermat principles; these will be needed in Chapter 18 to study limiting tangent and normal cones.

17.1 ε-SUBDIFFERENTIALS

Just like the ε-minimizers in Section 2.4, it can be useful to consider "relaxed" ε-subdifferentials. In particular, it is possible to derive *exact* calculus rules for these relaxed subdifferentials, which can lead to tighter results than inclusions for the corresponding exact subdifferentials (in particular, for the Fréchet subdifferential). We will make use of this in Chapter 27.

Similarly to the Fréchet subdifferential (16.2), we thus define for $F : X \to \overline{\mathbb{R}}$ the *ε(-Fréchet)-subdifferential* by

$$\partial_\varepsilon F(x) := \left\{ x^* \in X^* \;\middle|\; \liminf_{y \to x} \frac{F(y) - F(x) - \langle x^*, y - x \rangle_X}{\|y - x\|_X} \geq -\varepsilon \right\}, \tag{17.1}$$

where $\partial_0 F = \partial_F F$. The following lemma provides further insight into the ε-subdifferential.

Lemma 17.1. *Let $F : X \to \overline{\mathbb{R}}$ on a Banach space X, and $\varepsilon \geq 0$. Then the following are equivalent:*

(i) $x^* \in \partial_\varepsilon F(x)$;

(ii) $x^* \in \partial_F [F + \varepsilon \|\cdot - x\|_X](x)$;

(iii) $0 \in \partial_F [F + \varepsilon \|\cdot - x\|_X - \langle x^*, \cdot - x \rangle](x)$.

Proof. Each of the alternatives is by (17.1) and (16.2) equivalent to

$$\liminf_{y \to x} \frac{\varepsilon \|y - x\|_X + F(y) - F(x) - \langle x^*, y - x \rangle_X}{\|y - x\|_X} \geq 0. \qquad \square$$

We have the following "fuzzy" ε-sum rule.

Lemma 17.2. *Let X be a Banach space, $G : X \to \overline{\mathbb{R}}$, and $F : X \to \mathbb{R}$ be convex with $\partial F(x) \subset \mathbb{B}(\bar{x}^*, \varepsilon)$ for some $\varepsilon \geq 0$ and $\bar{x}^* \in X^*$. Then for all $\delta \geq 0$,*

$$\partial_\delta G(x) + \partial F(x) \subset \partial_\delta [G + F](x) \subset \partial_{\varepsilon+\delta} G(x) + \{\bar{x}^*\}.$$

In particular, if $\bar{x}^* \in \partial F(x)$, then

$$\partial_\delta G(x) + \partial F(x) \subset \partial_\delta [G + F](x) \subset \partial_{\varepsilon+\delta} G(x) + \partial F(x).$$

Proof. We start with the first inclusion. Let $\tilde{x}^* \in \partial F(x)$ and $x^* \in \partial_\delta G(x)$. Then the definitions (4.1) and (17.1), respectively, imply that

$$\begin{aligned}\liminf_{y \to x} \frac{G(y) - G(x) + F(y) - F(x) - \langle x^* + \tilde{x}^*, y - x\rangle_X}{\|y - x\|_X} \\ \geq \liminf_{y \to x} \frac{G(y) - G(x) - \langle x^*, y - x\rangle_X}{\|y - x\|_X} \geq -\delta,\end{aligned}$$

i.e., $x^* + \tilde{x}^* \in \partial_\delta [G + F](x)$.

To prove the second inclusion, let $x^* \in \partial_\delta [G + F](x)$ and $h \in X$ with $\|h\|_X = 1$. Then (17.1) implies that for all $t_n \searrow 0$ and $h_n \to h$,

$$\liminf_{n \to \infty} \frac{F(x + t_n h_n) - F(x) + G(x + t_n h_n) - G(x) - t_n \langle x^*, h_n\rangle_X}{t_n} \geq -\delta. \tag{17.2}$$

Since F is directionally differentiable by Lemma 4.3 and locally Lipschitz around $x \in \operatorname{int}(\operatorname{dom} F) = X$ by Theorem 3.13 with Lipschitz constant $L > 0$, we have

$$\lim_{n \to \infty} \frac{F(x + t_n h_n) - F(x)}{t_n} \leq \lim_{n \to \infty} \left(\frac{F(x + t_n h) - F(x)}{t_n} + L\|h_n - h\|_X \right) = F'(x; h).$$

Let now $\rho > 0$ be arbitrary. Then by Lemma 4.4, Theorem 13.8, , and Corollary 13.15 there exists an $x^*_{h,\rho} \in \partial F(x)$ such that $F'(x; h) \leq \langle x^*_{h,\rho}, h\rangle_X + \rho$. Therefore

$$\begin{aligned}\lim_{n \to \infty} \frac{F(x + t_n h_n) - F(x) - t_n \langle \bar{x}^*, h_n\rangle_X}{t_n} &\leq F'(x; h) - \langle \bar{x}^*, h\rangle_X \\ &\leq \langle x^*_{h,\rho} - \bar{x}^*, h\rangle_X + \rho \\ &\leq \varepsilon + \rho,\end{aligned}$$

where we have used that $\partial F(x) \subset \mathbb{B}(\bar{x}^*, \varepsilon)$ and $\|h\|_X = 1$ in the last inequality. Since $\rho > 0$ was arbitrary, the characterization (17.2) now implies

$$\liminf_{n \to \infty} \frac{G(x + t_n h_n) - G(x) - t_n \langle x^* - \bar{x}^*, h_n\rangle_X}{t_n} \geq -(\delta + \varepsilon).$$

Since $y_n := x + t_n h_n \to x$ was arbitrary, this proves $x^* - \bar{x}^* \in \partial_{\varepsilon+\delta} G(x)$, i.e., $\partial_\delta [G + F](x) \subset \partial_{\varepsilon+\delta} G(x) + \{\bar{x}^*\}$. □

The following is now immediate from Theorem 4.5, since we are allowed to take $\varepsilon = 0$ if $\partial F(x)$ is a singleton.

Corollary 17.3. *Let X be a Banach space, $G : X \to \overline{\mathbb{R}}$, and $F : X \to \mathbb{R}$ be convex and Gâteaux differentiable at $x \in X$. Then for every $\delta \geq 0$,*

$$\partial_\delta[G + F](x) = \partial_\delta G(x) + \{DF(x)\}.$$

In particular,

$$\partial_F[G + F](x) = \partial_F G(x) + \{DF(x)\}.$$

17.2 SMOOTH SPACES

For the remaining results in this chapter, we need additional assumptions on the normed vector space X. In particular, we need to assume that the norm is Gâteaux differentiable on $X \setminus \{0\}$; we call such spaces *Gâteaux smooth.*

Recalling from Chapter 7 the duality between differentiability and convexity, it is not surprising that this property can be related to the convexity of the dual norm. Here we need the following property: a normed vector space X is called *locally uniformly convex* if for any $x \in X$ with $\|x\|_X = 1$ and all $\varepsilon \in (0, 2]$ there exists a $\delta(\varepsilon, x) > 0$ such that

$$\|\tfrac{1}{2}(x + y)\|_X \leq 1 - \delta(\varepsilon, x) \quad \text{for all } y \in X \text{ with } \|y\|_X = 1 \text{ and } \|x - y\|_X \geq \varepsilon. \tag{17.3}$$

Lemma 17.4. *Let X be a Banach space and X^* be locally uniformly convex. Then X is Gâteaux smooth.*

Proof. Let $x \in X \setminus \{0\}$ be given. Since norms are convex, it suffices by Theorem 13.18 to show that $\partial\|\cdot\|_X(x)$ is a singleton. Let therefore $x_1^*, x_2^* \in \partial\|\cdot\|_X(x)$, i.e., satisfying by Theorem 4.6

$$\|x_1^*\|_{X^*} = \|x_2^*\|_{X^*} = 1, \qquad \langle x_1^*, x\rangle_X = \langle x_2^*, x\rangle_X = \|x\|_X.$$

This implies that

$$2 = \frac{1}{\|x\|_X}\left(\langle x_1^*, x\rangle_X + \langle x_2^*, x\rangle_X\right) = \langle x_1^* + x_2^*, \tfrac{x}{\|x\|_X}\rangle_X \leq \|x_1^* + x_2^*\|_{X^*}$$

by (1.1) and hence that $\|\frac{1}{2}(x_1^* + x_2^*)\|_{X^*} \geq 1$. Since X^* is locally uniformly convex, this is only possible if $x_1^* = x_2^*$, as otherwise we could choose for $\varepsilon := \|x_1^* - x_2^*\|_{X^*} \in (0, 2]$ a $\delta(\varepsilon, x) > 0$ such that $\|\frac{1}{2}(x_1^* + x_2^*)\|_{X^*} \leq 1 - \delta(\varepsilon, \delta) < 1$. □

Remark 17.5. In fact, if X is additionally reflexive, the norm is even continuously (Fréchet) differentiable; see [Schirotzek, 2007, Proposition 4.7.10]. We will not need this stronger property, however. In addition, locally uniformly convex spaces always have the Radon–Riesz property; see [Schirotzek, 2007, Lemma 4.7.9].

Example 17.6. The following spaces are locally uniformly convex:

(i) X a Hilbert space. This follows from the *parallelogram identity*

$$\|\tfrac{1}{2}(x+y)\|_X^2 = \frac{1}{2}\|x\|_X^2 + \frac{1}{2}\|y\|_X^2 - \frac{1}{4}\|x-y\|_X^2 \qquad \text{for all } x, y \in X,$$

which in fact characterizes precisely those norms that are induced by an inner product. This identity immediately yields for all $\varepsilon > 0$ and all $x, y \in X$ satisfying $\|x-y\|_X \geq \varepsilon$ that

$$\|\tfrac{1}{2}(x+y)\|_X^2 \leq 1 - \frac{\varepsilon^2}{4} \leq \left(1 - \frac{\varepsilon^2}{8}\right)^2,$$

which in particular verifies (17.3) with $\delta := \frac{\varepsilon^2}{8}$.

(ii) $X = L^p(\Omega)$ for $p \in (2, \infty)$. This follows from the algebraic inequality

$$|a+b|^p + |a-b|^p \leq 2^{p-1}(|a|^p + |b|^p) \qquad \text{for all } a, b \in \mathbb{R},$$

see [Cioranescu, 1990, Lemma II.4.1]. This implies that

$$\|\tfrac{1}{2}(u+v)\|_{L^p(\Omega)}^p \leq \frac{1}{2}\|u\|_{L^p(\Omega)}^p + \frac{1}{2}\|v\|_{L^p(\Omega)}^p - \frac{1}{2^p}\|u-v\|_{L^p(\Omega)}^p \quad \text{for all } u, v \in L^p(\Omega).$$

We can now argue exactly as in case (i).

(iii) $X = L^p(\Omega)$ for $p \in (1, 2)$. This follows from the algebraic inequality

$$|a+b|^p + |a-b|^p \leq 2(|a|^p + |b|^p)^{p/(p-1)} \qquad \text{for all } a, b \in \mathbb{R},$$

see [Cioranescu, 1990, Lemma II.4.1], implying a similar inequality for the $L^p(\Omega)$ norms from which the claim follows as for (i) and (ii).

Hence every Hilbert space (by identifying X with X^*) and every $L^p(\Omega)$ for $p \in (1, \infty)$ (identifying $L^p(\Omega)$ with $L^q(\Omega)$, $q = \frac{p}{p-1} \in (1, \infty)$) is Gâteaux smooth.

In fact, the celebrated Lindenstrauss and Trojanski *renorming theorems* show that every reflexive Banach space admits an equivalent norm such that the space (with that norm) becomes locally uniformly convex; see [Cioranescu, 1990, Theorem III.2.10]. (Of course, even though that means that the dual space of the *renormed space* is Gâteaux smooth, this does not imply anything about the differentiability of the original norm, as the obvious example of $\mathbb{R}^N$ endowed with the 1- or the ∞-norm shows.) For many more details on smooth and uniformly convex spaces, see [Cioranescu, 1990; Fabian et al., 2001; Schirotzek, 2007].

Note that even in Gâteaux smooth spaces, the norm will not be differentiable at $x = 0$. But this can be addressed by considering $\|x\|_X^p$ for $p > 1$; for later use, we state this for $p = 2$.

Lemma 17.7. *Let X be a Gâteaux smooth Banach space and $F(x) = \|x\|_X^2$. Then F is Gâteaux differentiable at any $x \in X$ with*

$$DF(x) = 2\|x\|_X x^* \qquad \text{for any } x^* \in X^* \text{ with } \|x^*\|_{X^*} = 1 \text{ and } \langle x^*, x\rangle_X = \|x\|_X.$$

Proof. Since norms are convex, we can apply Theorems 4.6 and 4.19 to obtain that

$$\partial F(x) = \{2\|x\|_X x^* \mid x^* \in X^* \text{ with } \|x^*\|_{X^*} = 1 \text{ and } \langle x^*, x\rangle_X = \|x\|_X\} \quad (x \in X).$$

At any $x \neq 0$, this set is a singleton by Theorem 4.5 and the assumption that X is Gâteaux smooth. Clearly also $\partial F(0) = \{0\}$, and hence the claim follows from Theorem 13.18. □

Remark 17.8 (Asplund spaces). *Asplund spaces* are, by (one equivalent) definition, those Banach spaces where every continuous, convex, real-valued function is Fréchet-differentiable on a dense set. (This is a limited version of Rademacher's Theorem 13.25 in $\mathbb{R}^N$.) We refer to [Yost, 1993] for an introduction to Asplund spaces. Importantly, reflexive Banach spaces are Asplund.

The norm of an Asplund space is thus differentiable on a dense set D. It was shown in [Ekeland and Lebourg, 1976] that perturbed optimization problems on Asplund spaces have solutions on a dense set of perturbation parameters and that the objective function is differentiable at such a solution. If we worked in the following sections with perturbed optimization problems and applied such an existence result instead of the Ekeland or the Borwein–Preiss variational principles (Theorem 2.16 or Theorem 2.17, respectively), we would be able to extend the following results to Asplund spaces.

17.3 FUZZY FERMAT PRINCIPLES

The following result generalizes the Fermat principle of Theorem 16.2 to sums of two functions in a "fuzzy" fashion. We will use it to show a fuzzy containment formula for ε-subdifferentials. Its generalizations to more than two functions can also be used to derive more advanced fuzzy sum rules than 17.2. Our focus is, however, on exact calculus, so we will not be developing such generalizations.

Lemma 17.9 (fuzzy Fermat principle). *Let X be a Gâteaux smooth Banach space and $F, G : X \to \overline{\mathbb{R}}$. If $F + G$ attains a local minimum at a point $\bar{x} \in X$ where F is lower semicontinuous and G is locally Lipschitz, then for any $\delta, \mu > 0$ we have*

$$0 \in \bigcup_{x,y \in \mathbb{B}(\bar{x},\delta)} (\partial_F F(x) + \partial_F G(y)) + \mu \mathbb{B}_{X^*}.$$

Proof. Let $\rho, \alpha > 0$ be arbitrary. The idea is to separate the two nonsmooth functions F and G, and hence be able to use the exact sum rule of Corollary 17.3, by locally relaxing the problem $\min_{x\in X}(F+G)$ to

$$\inf_{x,y\in X} J_\alpha(x,y) := F(x) + G(y) + \alpha\|x-y\|_X^2 + \|x-\bar{x}\|_X^2 + \delta_{\mathbb{B}(\bar{x},\rho)^2}(x,y).$$

We take $\rho > 0$ small enough that $\bar{x}$ minimizes $F+G$ within $\mathbb{B}(\bar{x},\rho)$, and both $F \geq F(\bar{x}) - 1$ and $G \geq G(\bar{y}) - 1$ on $\mathbb{B}(\bar{x},\rho)$. The first requirement is possible by the assumption of $F+G$ attaining its local minimum at $\bar{x}$, while the latter follows from the lower semicontinuity of F and the local Lipschitz continuity of G. In the following, we denote by L the Lipschitz factor of G on $\mathbb{B}(\bar{x},\rho)$. It follows that $J_\alpha(x,y) \geq F(\bar{x}) + G(\bar{x}) - 2$ for all $(x,y) \in \mathbb{B}(\bar{x},\rho)^2 = \operatorname{dom} J_\alpha$, and hence J_α is bounded from below.

We study the approximate solutions of the relaxed problem in several steps.

Step 1: constrained infimal values converge to $J(\bar{x},\bar{x})$. Let $x_\alpha, y_\alpha \in \mathbb{B}(\bar{x},\rho)$ be such that

$$J_\alpha(x_\alpha, y_\alpha) < j_\alpha + \alpha^{-1} \quad \text{where} \quad j_\alpha := \inf_{x,y\in X} J_\alpha(x,y). \tag{17.4}$$

We show that

$$J_\alpha(\bar{x},\bar{x}) < j_\alpha + \varepsilon_\alpha \quad \text{for} \quad \varepsilon_\alpha := L\sqrt{\frac{\alpha^{-1}+2}{\alpha}} + \alpha^{-1}.$$

To start with, we have

$$\begin{aligned} F(\bar{x}) + G(\bar{x}) + \alpha^{-1} &= J(\bar{x},\bar{x}) + \alpha^{-1} \\ &\geq j_\alpha + \alpha^{-1} \\ &> J_\alpha(x_\alpha, y_\alpha) \\ &= F(x_\alpha) + G(y_\alpha) + \alpha\|x_\alpha - y_\alpha\|_X^2 + \|x_\alpha - \bar{x}\|_X^2 \\ &\geq F(\bar{x}) + G(\bar{x}) + \alpha\|x_\alpha - y_\alpha\|_X^2 + \|x_\alpha - \bar{x}\|_X^2 - 2. \end{aligned}$$

This implies that $\|x_\alpha - y_\alpha\|_X < \sqrt{\frac{\alpha^{-1}+2}{\alpha}}$. Since $\bar{x}$ minimizes $F+G$ within $\mathbb{B}(\bar{x},\rho)$, we obtain the bound (17.4) through

$$\begin{aligned} J_\alpha(\bar{x},\bar{x}) &= F(\bar{x}) + G(\bar{x}) \\ &\leq F(x_\alpha) + G(x_\alpha) \\ &\leq F(x_\alpha) + G(y_\alpha) + L\|x_\alpha - y_\alpha\|_X. \\ &\leq J(x_\alpha, y_\alpha) + L\|x_\alpha - y_\alpha\|_X. \\ &< j_\alpha + \varepsilon_\alpha. \end{aligned}$$

Step 2: exact unconstrained minimizers exist for a perturbed problem. By (17.4), we can apply the Borwein–Preiss variational principle (Theorem 2.17) for any $\lambda, \alpha > 0$, small enough $\rho > 0$ (all to be fixed later), and $p = 2$ to obtain a sequence $\{\mu_n\}_{n\geq 0}$ of nonnegative weights summing to 1 and a sequence $\{(x_n, y_n)\}_{n\geq 0} \subset X \in X$ with $(x_0, y_0) = (\bar{x},\bar{x})$ converging strongly to some $(\widehat{x}_\alpha, \widehat{y}_\alpha) \in X \times X$ (endowed with the euclidean product norm) such that

(i) $\|x_n - \widehat{x}_\alpha\|_X^2 + \|y_n - \widehat{y}_\alpha\|_X^2 \le \lambda^2$ for all $n \ge 0$ (in particular, $\|\bar{x} - \widehat{x}_\alpha\|_X \le \lambda$);

(ii) the function

$$H_\alpha(x, y) := J_\alpha(x, y) + \frac{\varepsilon_\alpha}{\lambda^2} \sum_{n=0}^{\infty} \mu_n \left(\|x - x_n\|^2 + \|y - y_n\|^2 \right)$$

attains its global minimum at $(\widehat{x}_\alpha, \widehat{y}_\alpha)$.

Note that since J_α includes the constraint $(x, y) \in \mathbb{B}(\bar{x}, \rho)^2$, we have $(\widehat{x}_\alpha, \widehat{y}_\alpha) \in \mathbb{B}(\bar{x}, \rho)^2$. In fact, by taking $\lambda \in (0, \rho)$, it follows from (i) and the convergence $(x_n, y_n) \to (\widehat{x}_\alpha, \widehat{y}_\alpha)$ that the minimizer $(\widehat{x}_\alpha, \widehat{y}_\alpha) \in \mathbb{B}(\bar{x}, \lambda)^2 \subset \operatorname{int} \mathbb{B}(\bar{x}, \rho)^2$ is unconstrained.

Step 3: the perturbed minimizers satisfy the claim for large α and small λ. Setting $\Psi_y(x) := \|x - y\|_X^2$, it follows from Lemma 17.7 that Ψ_y is Gâteaux differentiable for any $y \in X$ with $D\Psi_y(x) \in 2\|x - y\|_X \mathbb{B}_{X^*}$. Furthermore, since $(\widehat{x}_\alpha, \widehat{y}_\alpha) \in \operatorname{int} \mathbb{B}(\bar{x}, \rho)^2$, we have $\partial \delta_{\mathbb{B}(\bar{x},\rho)^2}(\widehat{x}_\alpha, \widehat{y}_\alpha) = (0, 0)$. Hence the only nonsmooth component of H_α at $(\widehat{x}_\alpha, \widehat{y}_\alpha)$ is $(x, y) \mapsto F(x) + G(y)$. We can thus apply Theorem 16.2 and Corollary 17.3 to obtain

$$0 \in \partial_F H_\alpha(\widehat{x}_\alpha, \widehat{y}_\alpha) = \begin{pmatrix} \partial_F F(\bar{x}) + \alpha D\Psi_{\widehat{y}_\alpha}(\widehat{x}_\alpha) + D\Psi_{\bar{x}}(\widehat{x}_\alpha) + \frac{\varepsilon_\alpha}{\lambda^2} \sum_{n=0}^{\infty} \mu_n D\Psi_{x_n}(\widehat{x}_\alpha) \\ \partial_F G(\bar{x}) + \alpha D\Psi_{\widehat{x}_\alpha}(\widehat{y}_\alpha) + \frac{\varepsilon_\alpha}{\lambda^2} \sum_{n=0}^{\infty} \mu_n D\Psi_{y_n}(\widehat{y}_\alpha) \end{pmatrix}.$$

By (i) and $\widehat{x}_\alpha, \widehat{y}_\alpha \in \mathbb{B}(\bar{x}, \lambda)$ we have $\|\widehat{x}_\alpha - x_n\|_X, \|\widehat{y}_\alpha - y_n\|_X \le \lambda$ for all $n \ge 0$. In addition, $\sum_{n=0}^{\infty} \mu_n = 1$, and thus $\frac{\varepsilon_\alpha}{\lambda^2} \sum_{n=0}^{\infty} \mu_n D\Psi_{x_n}(\widehat{x}_\alpha) \in \frac{2\varepsilon_\alpha}{\lambda} \mathbb{B}_{X^*}$ and likewise for $D\Psi_{y_n}$ (so that in fact we were justified in differentiating the series term-wise). By (i) also $\|\widehat{x}_\alpha - \bar{x}\|_X \le \lambda$, so that $D\Psi_{\bar{x}}(\widehat{x}_\alpha) \in 2\lambda \mathbb{B}_{X^*}$. Finally, since $-x^* \in \partial \|\cdot\|_X(-x)$ for any $x^* \in \partial \|\cdot\|_X(x)$ and any $x \in X$, we have $D\Psi_y(x) = -D\Psi_x(y)$ for all $x, y \in X$. We thus have

$$\begin{cases} -\alpha D\Psi_{\widehat{y}_\alpha}(\widehat{x}_\alpha) \in \partial_F F(\widehat{x}_\alpha) + \left(2\lambda + \dfrac{2\varepsilon_\alpha}{\lambda}\right) \mathbb{B}_{X^*}, \\ \alpha D\Psi_{\widehat{y}_\alpha}(\widehat{x}_\alpha) \in \partial_F G(\widehat{y}_\alpha) + \dfrac{2\varepsilon_\alpha}{\lambda} \mathbb{B}_{X^*}, \end{cases}$$

which implies that

$$0 \in \partial_F F(\widehat{x}_\alpha) + \partial_F G(\widehat{y}_\alpha) + \left(2\lambda + \frac{4\varepsilon_\alpha}{\lambda}\right) \mathbb{B}_{X^*}.$$

Since $(\widehat{x}_\alpha, \widehat{y}_\alpha) \in \mathbb{B}(\bar{x}, \lambda)^2$, the claim now follows by taking $\lambda \in (0, \rho)$ small enough and then $\alpha > 0$ large (and thus ε_α small) enough. □

Remark 17.10 (fuzzy Fermat principles and trustworthy subdifferentials). Lemma 17.9 is due to [Fabian, 1988]. Such fuzzy Fermat principles are studied in more detail from the point of view of *fuzzy variational principles* in [Ioffe, 2017]. Specifically, the claim of Lemma 17.9 has to hold for an arbitrary subdifferential operator ∂_* for it to be called *trustworthy*, whereas the opposite inclusion $\partial_* G(x) + \partial_* F(x) \subset \partial_*[G + F](x)$ is required for the subdifferential to be called *elementary*.

Remark 17.11 (notes on the proof of Lemma 17.9). Note how we had to apply the Borwein–Preiss variational principle instead of Ekeland's to obtain a differentiable convex perturbation and thus to be able to apply the sum rule Corollary 17.3. In contrast, the proof in [Ioffe, 2017] is based on the Deville–Godefroy–Zizler variational principle, which makes no convexity assumption on the perturbation function and hence requires the stronger property of Fréchet smoothness (i.e., Fréchet instead of Gâteaux differentiability of the norm outside the origin).

Finally, with an additional argument showing $J_\alpha(\widehat{x}_\alpha, \widehat{y}_\alpha) \leq j_\alpha + \beta_\alpha$ for a suitable β_α, it would be possible to further constrain $|F(x) - F(\bar{x})| \leq \delta$ in the claim of Lemma 17.9, as is done in [Ioffe, 2017, Theorem 4.30].

Corollary 17.12. *Let X be a Gâteaux smooth Banach space, let $F : X \to \overline{\mathbb{R}}$ be lower semicontinuous near $\bar{x} \in X$, and $\varepsilon > 0$. Then for any $\delta > 0$ and $\varepsilon' > \varepsilon$ we have*

$$\partial_\varepsilon F(\bar{x}) \subset \bigcup_{z \in \mathbb{B}(\bar{x},\delta)} \partial_F F(z) + \varepsilon' \mathbb{B}_{X^*}.$$

Proof. We may assume that $\bar{x} \in \operatorname{dom} F$, in particular that there exists some $x^* \in \partial_\varepsilon F(\bar{x})$, i.e., such that

$$\liminf_{\bar{x} \neq y \to \bar{x}} \frac{F(y) - F(\bar{x}) - \langle x^*, y - \bar{x} \rangle_X}{\|y - \bar{x}\|_X} \geq -\varepsilon.$$

Taking any $\varepsilon' > \varepsilon$ and defining

$$\bar{F}(x) := F(x) - \langle x^*, x - \bar{x} \rangle_X \quad \text{and} \quad \bar{G}(x) := \varepsilon' \|x - \bar{x}\|_X,$$

we obtain as in Lemma 17.1 that

$$\liminf_{\bar{x} \neq y \to \bar{x}} \frac{(\bar{G} + \bar{F})(y) - (\bar{G} + \bar{F})(\bar{x})}{\|y - \bar{x}\|_X} \geq (\varepsilon' - \varepsilon).$$

Thus $\bar{F} + \bar{G}$ achieves its local minimum at $\bar{x}$. The function $\bar{G}$ is convex and Lipschitz while $\bar{F}$ lower semicontinuous. Hence Lemma 17.9 implies for any $\delta > 0$ and $\mu' > 0$ that

$$0 \in \bigcup_{z,y \in \mathbb{B}(\bar{x},\delta)} \big(\partial_F \bar{F}(y) + \partial_F \bar{G}(z)\big) + \mu' \mathbb{B}_X.$$

Since $\partial_F \bar{F}(y) = \partial_F F(y) - \{x^*\}$ (by Corollary 17.3 or directly from the definition) and $\partial_F \bar{G}(z) = \partial \bar{G}(z) \subset \varepsilon' \mathbb{B}_X$, we obtain

$$x^* \in \bigcup_{z \in \mathbb{B}(x,\delta)} \partial_F \bar{F}(z) + (\mu' + \varepsilon') \mathbb{B}_X.$$

Since $\mu' > 0$ and $\varepsilon' > \varepsilon$ were arbitrary, the claim follows. □

17.4 APPROXIMATE FERMAT PRINCIPLES AND PROJECTIONS

We now introduce an *approximate Fermat principle*, which can be invoked when we *do not know* whether a minimizer exists; in particular, when F fails to be *weakly* lower semicontinuous so that Theorem 2.1 is not applicable.

Theorem 17.13. *Let X be a Banach space and $F : X \to \overline{\mathbb{R}}$ be proper, lower semicontinuous, and bounded from below. Then for every $\varepsilon, \delta > 0$ there exists an $\bar{x}_\varepsilon \in X$ such that*

(i) $F(\bar{x}_\varepsilon) \leq \inf_{x \in X} F(x) + \varepsilon$;

(ii) $F(\bar{x}_\varepsilon) < F(x) + \delta \|x - \bar{x}_\varepsilon\|_X$ *for all* $x \neq \bar{x}_\varepsilon$;

(iii) $0 \in \partial_\delta F(\bar{x}_\varepsilon)$.

Proof. Since F is bounded from below, $\inf_{x \in X} F(x) > -\infty$. We can thus take a minimizing sequence $\{x_n\}_{n \in \mathbb{N}}$ with $F(x_n) \searrow \inf_{x \in X} F(x)$ and find a $n(\varepsilon) \in \mathbb{N}$ such that $x_\varepsilon := x_{n(\varepsilon)}$ satisfies (i). Ekeland's variational principle Theorem 2.16 thus yields for $\lambda := \varepsilon/\delta$ an $\bar{x}_\varepsilon := \bar{x}_{\varepsilon,\lambda}$ such that $\|\bar{x}_\varepsilon - x_\varepsilon\|_X \leq \lambda$,

$$F(\bar{x}_\varepsilon) \leq F(\bar{x}_\varepsilon) + \frac{\varepsilon}{\lambda}\|\bar{x}_\varepsilon - x_\varepsilon\|_X \leq F(x_\varepsilon),$$

as well as

$$F(\bar{x}_\varepsilon) < F(x) + \frac{\varepsilon}{\lambda}\|\bar{x}_\varepsilon - x\|_X \quad (x \neq \bar{x}_\varepsilon).$$

Thus (i) as well as (ii) hold. The latter implies for all $x \neq \bar{x}_\varepsilon$ that

$$\frac{F(x) - F(\bar{x}_\varepsilon) - \langle 0, x - \bar{x}_\varepsilon \rangle_X}{\|x - \bar{x}_\varepsilon\|_X} \geq -\delta,$$

i.e., $0 \in \partial_\delta F(\bar{x}_\varepsilon)$ by definition. □

As an example for possible applications of approximate Fermat principles, we use it to prove the following result on projections and approximate projections onto a *nonconvex* set $C \subset X$. For nonconvex sets, even the exact projection need no longer be unique; furthermore, for the reasons discussed before Theorem 17.13, the set of projections $P_C(x)$ may be empty when $C \neq \emptyset$ is closed but not *weakly* closed. We recall that by Lemma 1.10, convex closed sets are weakly closed, as are, of course, finite-dimensional closed sets. However, more generally, weak closedness can be elusive. Hence we will need to perform *approximate projections* in Part IV. It is not surprising that this requires additional assumptions on the containing space to make up for this.

Theorem 17.14. *Let X be a Gâteaux smooth Banach space and let $C \subset X$ be nonempty and closed. Define the (possibly multi-valued) projection*

$$P_C : X \rightrightarrows X, \qquad P_C(x) := \operatorname*{arg\,min}_{\tilde{x} \in C} \|\tilde{x} - x\|_X$$

and the corresponding distance function

$$d_C : X \to \mathbb{R}, \qquad d_C(x) := \inf_{\tilde{x} \in C} \|\tilde{x} - x\|_X.$$

Then the following hold:

(i) For any $\bar{x} \in P_C(x)$, there exists an $\bar{x}^* \in \partial_F \delta_C(\bar{x})$ such that

$$\langle \bar{x}^*, x - \bar{x} \rangle_X = \|x - \bar{x}\|_X, \qquad \|\bar{x}^*\|_{X^*} \leq 1. \tag{17.5}$$

(ii) For any $\varepsilon > 0$, there exists an approximate projection $\bar{x}_\varepsilon \in C$ satisfying

$$\|\bar{x}_\varepsilon - x\|_X \leq d_C(x) + \varepsilon$$

as well as (17.5) for some $\bar{x}^* \in \partial_\varepsilon \delta_C(\bar{x}_\varepsilon)$.

(iii) If X is a Hilbert space, then $x - \bar{x} \in \partial_\varepsilon \delta_C(\bar{x})$ for all $\varepsilon \geq 0$.

Proof. (i): Let $x \notin C$, since otherwise $\bar{x}^* := 0 \in \partial_F(\bar{x})$ for $\bar{x} = x \in C$ by the definition of the Fréchet subdifferential. Set $F(\tilde{x}) := \|\tilde{x} - x\|_X$ and assume that $\bar{x} \in P_C(x)$. The Fermat principle Theorem 16.2 then yields that $0 \in \partial_F[\delta_C + F](\bar{x})$. Since $x \notin C$ and $\bar{x} \in C$, by assumption F is differentiable at $\bar{x}$. Thus Theorem 4.5 shows that $\partial F(\bar{x}) = \{DF(\bar{x})\}$ is a singleton. The sum rule of Corollary 17.3 then yields that $\bar{x}^* := -DF(\bar{x}) \in \partial_F \delta_C(\bar{x})$. The claim of (17.5) now follows from Theorem 4.6.

(ii): Compared to (i), we merely invoke the approximate Fermat principle of Theorem 17.13 in place of Theorem 16.2, which establishes the existence of $\bar{x}_\varepsilon \in C$ satisfying $\|\bar{x}_\varepsilon - x\|_X \leq d_C(x)$ and $0 \in \partial_\varepsilon[\delta_C + F](\bar{x})$. The sum rule of Lemma 17.2 then shows that $\bar{x}^* := -DF(\bar{x}) \in \partial_\varepsilon \delta_C(\bar{x})$.

(iii): In a Hilbert space, we can identify $-DF(\bar{x})$ with the corresponding gradient $-\nabla F(\bar{x}) = (x - \bar{x})/\|x - \bar{x}\|_X \in X$ for $\bar{x} \neq 0$ (otherwise $-\nabla F(\bar{x}) = 0 = x - \bar{x}$). Since $\partial_\varepsilon \delta_C(\bar{x})$ is a cone, this implies that $x - \bar{x} \in \partial_F \delta_C(\bar{x})$ as well. □

In the next chapters, we will see that $\partial_F \delta_C(\bar{x})$ coincides with a suitable normal cone to C at $\bar{x}$. In other words, $\bar{x}^*$ is a normal vector to the set C. In Hilbert spaces, this normal vector can be identified with the (normalized) vector pointing from $\bar{x}$ to x.

Part IV

SET-VALUED ANALYSIS

18 TANGENT AND NORMAL CONES

We now start our study of stability properties of the solutions to nonsmooth optimization problems. As we have characterized the latter via subdifferential inclusions, we need to study the sensitivity of such relations to perturbations. As in the smooth case, this can be done through derivatives of these conditions with respect to relevant parameters; however, these conditions are expressed as inclusions instead of simple equations. Hence we require notions of derivatives for set-valued mappings.

To motivate how we will develop differential calculus for set-valued mappings, recall from Lemma 4.10 how the subdifferential of a convex function F can be defined in terms of the normal cone to the epigraph of F. This idea forms the basis of differentiating general set-valued mappings $H : X \rightrightarrows Y$, where instead of taking the normal cone at $(x, F(x))$ to $\operatorname{epi} F$, we do this at any point (x, y) of $\operatorname{graph} H := \{(x, y) \in X \times Y \mid y \in H(x)\}$. Since we are generally not in the nice convex setting – even for a convex function F, the set $\operatorname{graph} \partial F$ is not convex unless F is linear – there are some complications which result in having to deal with various nonequivalent definitions. In this chapter, we introduce the relevant graphical notions of tangent and normal cones. In Chapter 19, we develop specific expressions for these conses to sets in $L^p(\Omega)$ defined as pointwise via finite-dimensional sets. In the following Chapters 20 to 25, we then define and further develop notions of differentiation of set-valued mappings based on these cones.

18.1 DEFINITIONS AND EXAMPLES

THE FUNDAMENTAL CONES

Our first type of tangent cone is defined using roughly the same limiting process on difference quotients as basic directional derivatives. Let X be a Banach space. We define the *tangent cone* (or *Bouligand* or *contingent cone*) of the set $C \subset X$ at $x \in X$ as

$$T_C(x) := \limsup_{\tau \searrow 0} \frac{C - x}{\tau} = \left\{ \Delta x \in X \;\middle|\; \Delta x = \lim_{k \to \infty} \frac{x_k - x}{\tau_k} \text{ for some } C \ni x_k \to x,\ \tau_k \searrow 0 \right\}, \tag{18.1}$$

i.e., the tangent cone is the outer limit (in the sense of Section 6.1) of the "blown up" sets $(C - x)/\tau$ as $\tau \searrow 0$.

The tangent cone is closely related to the *Fréchet normal cone*, which is based on the same limiting process as the Fréchet subdifferential in Chapter 16:

$$\widehat{N}_C(x) := \left\{ x^* \in X^* \;\middle|\; \limsup_{C \ni \tilde{x} \to x} \frac{\langle x^*, \tilde{x} - x \rangle_X}{\|\tilde{x} - x\|_X} \leq 0 \right\}. \tag{18.2}$$

LIMITING CONES IN FINITE DIMENSIONS

One difficulty with the Fréchet normal cone is that it is not outer semicontinuous. By taking their outer limit (in the sense of set-valued mappings), we obtain the less "irregular" (*basic* or *limiting* or *Mordukhovich*) *normal cone*. This definition is somewhat more involved in infinite dimensions, so we first consider $C \subset \mathbb{R}^N$ at $x \in \mathbb{R}^N$. In this case, the limiting normal cone is defined as

$$\begin{aligned} N_C(x) &:= \limsup_{C \ni \tilde{x} \to x} \widehat{N}_C(\tilde{x}) \\ &= \left\{ x^* \in \mathbb{R}^N \;\middle|\; x^* = \lim_{k \to \infty} x_k^* \text{ for some } x_k^* \in \widehat{N}_C(x_k),\ C \ni x_k \to x \right\}. \end{aligned} \tag{18.3}$$

Despite N_C being obtained by the outer semicontinuous regularization of $\widehat{N}_C$, the *latter* is sometimes in the literature called the *regular normal cone*. We stick to the convention of calling $\widehat{N}_C$ the *Fréchet normal cone* and N_C the *limiting normal cone*.

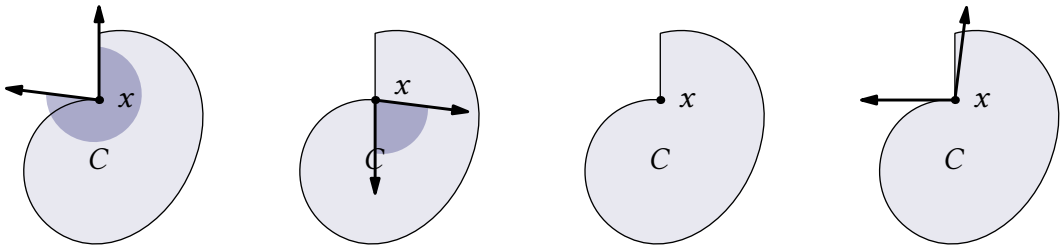
The limiting variant of the tangent cone is the *Clarke tangent cone* (also known as the *regular tangent cone*), defined for a set $C \subset \mathbb{R}^N$ at $x \in \mathbb{R}^N$ as the inner limit

$$\begin{aligned} \widehat{T}_C(x) &:= \liminf_{\substack{C \ni \tilde{x} \to x, \\ \tau \searrow 0}} \frac{C - \tilde{x}}{\tau} \\ &= \left\{ \Delta x \in \mathbb{R}^N \;\middle|\; \begin{array}{r} \text{for all } \tau_k \searrow 0,\ C \ni x_k \to x \text{ there exists } C \ni \tilde{x}_k \to x \\ \text{with } (\tilde{x}_k - x_k)/\tau_k \to \Delta x \end{array} \right\}. \end{aligned} \tag{18.4}$$

We will later in Corollary 18.20 see that for a closed set $C \subset \mathbb{R}^N$, we in fact have that $\widehat{T}_C(x) = \liminf_{C \ni \tilde{x} \to x} T_C(\tilde{x})$.

The following example as well as Figure 18.1 illustrate the different cones.

Example 18.1. We compute the different tangent and normal cones at all points $x \in C$ for different $C \subset \mathbb{R}^2$.



(a) tangent cone $T_C(x)$ (b) Clarke tangent cone $\widehat{T}_C(x)$ (c) Fréchet normal cone $\widehat{N}_C(x) = \{0\}$ (d) limiting normal cone $N_C(x)$

Figure 18.1: Illustration of the different normal and tangent cones at a nonregular point of a set C. The dot indicates the base point x. The thick arrows and dark filled-in areas indicate the directions included in the cones.

(i) $C = \mathbb{B}(0,1)$: Clearly, if $x \in \operatorname{int} C$, then

$$N_C(x) = \widehat{N}_C(x) = \{0\},$$
$$T_C(x) = \widehat{T}_C(x) = \mathbb{R}^2.$$

For any $x \in \operatorname{bd} C$, on the other hand,

$$N_C(x) = \widehat{N}_C(x) = [0,\infty)x := \{tx \mid t \geq 0\},$$
$$T_C(x) = \widehat{T}_C(x) = \{z \mid \langle z, x\rangle \leq 0\}.$$

(ii) $C = [0,1]^2$: For $x \in \operatorname{int} C$, we again have that $N_C(x) = \widehat{N}_C(x) = \{0\}$ and $T_C(x) = \widehat{T}_C(x) = \mathbb{R}^2$; similarly, for $x \in \operatorname{bd} C \setminus \{(0,0),(0,1),(1,0),(1,1)\}$ (i.e., x is not one of the corners of C), again $N_C(x) = \widehat{N}_C(x) = [0,\infty)x$ and $T_C(x) = \widehat{T}_C(x) = \{z \mid \langle z, x\rangle = 0\}$. Of the corners, we concentrate on $x = (1,1)$, the others being analogous. Then

$$N_C(x) = \widehat{N}_C(x) = \{(\Delta x, \Delta y) \mid \Delta x, \Delta y \geq 0\},$$
$$T_C(x) = \widehat{T}_C(x) = \{(\Delta x, \Delta y) \mid \Delta x, \Delta y \leq 0\}.$$

(iii) $C = [0,1]^2 \setminus [\frac{1}{2},1]^2$: Here as well $N_C(x) = \widehat{N}_C(x) = \{0\}$ and $T_C(x) = \widehat{T}_C(x) = \mathbb{R}^2$ for $x \in \operatorname{int} C$. Other points on $\operatorname{bd} C$ are computed analogously to similar corners and edges of the square $[0,1]^2$, but we have to be careful with the "interior corner" $x = (\frac{1}{2},\frac{1}{2})$. Here, similarly to Figure 18.1c, we see that $\widehat{N}_C(x) = \{0\}$. However, as a lim sup,

$$N_C(x) = (0,1)[0,\infty) \cup (1,0)[0,\infty).$$

For the tangent cones, we then get

$$T_C(x) = \{(\Delta x, \Delta y) \mid \Delta x \le 0 \text{ or } \Delta y \le 0\},$$

while, as a lim inf,

$$\widehat{T}_C(x) = T_C(x) \cup (1,0)\mathbb{R} \cup (0,1)\mathbb{R}.$$

LIMITING CONES IN INFINITE DIMENSIONS

Let now X be again a Banach space. Although the fundamental cones – the (basic) tangent cone and the Fréchet normal cone – were defined based on strongly convergent sequences, in infinite-dimensional spaces weak modes of convergence better replicate various relationships between the different cones. We thus call an element $\Delta x \in X$ *weakly tangent* to C at x if

$$\Delta x = \operatorname*{w\text{-}lim}_{k\to\infty} \frac{x_k - x}{\tau_k} \quad \text{for some} \quad C \ni x_k \to x,\ \tau_k \searrow 0, \tag{18.5}$$

where the w-lim of course stands for $\tau_k^{-1}(x_k - x) \rightharpoonup \Delta x$. We denote by the *weak tangent cone* (or *weak contingent cone*) $T_C^w(x) \subset X$ the set of all such Δx. Using the notion of outer limits of set-valued mappings from Chapter 6, we can also write

$$T_C^w(x) = \operatorname*{w\text{-}limsup}_{\tau\searrow 0} \frac{C - x}{\tau}. \tag{18.6}$$

Likewise, the limiting normal cone $N_C(x)$ to $C \subset X$ in a general infinite-dimensional Banach space X is based on weak-$*$ limits. Moreover, several proofs will be easier if we slightly relax the definition. Therefore, given $\varepsilon \ge 0$ we first introduce the *ε-normal cone* of $x^* \in X^*$ satisfying

$$\widehat{N}_C^\varepsilon(x) := \left\{ x^* \in X^* \;\middle|\; \limsup_{C \ni \tilde{x} \to x} \frac{\langle x^*, \tilde{x} - x\rangle_X}{\|\tilde{x} - x\|_X} \le \varepsilon \right\}. \tag{18.7}$$

The *Fréchet normal cone* is then simply $\widehat{N}_C(x) := \widehat{N}_C^0(x)$.

Now, the (*basic* or *limiting* or *Mordukhovich*) *normal cone* is defined as

$$N_C(x) := \operatorname*{w\text{-}*\text{-}limsup}_{\tilde{x}\to x,\ \varepsilon\searrow 0} \widehat{N}_C^\varepsilon(\tilde{x}). \tag{18.8}$$

In other words, $x^* \in N_C(x)$ if and only if there exist $C \ni x_k \to x$, $\varepsilon_k \searrow 0$ and $x_k^* \in N_C^{\varepsilon_k}(x_k)$ such that $x_k^* \overset{*}{\rightharpoonup} x^*$.

In Gâteaux smooth Banach spaces, we can fix $\varepsilon \equiv 0$ in (18.8). Thus such spaces can be treated similarly to the finite-dimensional case in (18.3).

Theorem 18.2. *Let X be a Gâteaux smooth Banach space, $C \subset X$, and $x \in X$. Then*

$$N_C(x) = \text{w-}*\text{-}\limsup_{\tilde{x}\to x} \widehat{N}_C(\tilde{x}). \tag{18.9}$$

Proof. Denote by K the set on the right hand side of (18.9). Then by the definition (18.8), clearly $N_C(x) \supset K$. To show $N_C(x) \subset K$, let $x^* \in N_C(x)$. Then (18.8) yields $x_k \to x$, $\varepsilon_k \searrow 0$, and $x_k^* \overset{*}{\rightharpoonup} x_k$ with $x_k^* \in \widehat{N}_C^{\varepsilon_k}(x_k)$. We need to show that there exist some $\tilde{x}_k \to x$ and $\tilde{x}_k^* \overset{*}{\rightharpoonup} x^*$ with $\tilde{x}_k^* \in \widehat{N}_C(\tilde{x}_k)$. Indeed, since $\widehat{N}_C^{\varepsilon} = \partial_\varepsilon \delta_C$, by Corollary 17.12 applied to $F = \delta_C$, we have for any sequence $\delta_k \searrow 0$ that

$$x_k^* \in N_C^{\varepsilon_k}(x_k) \subset \bigcup_{\tilde{x}\in\mathbb{B}(x_k,\delta_k)} N_C(\tilde{x}) + \delta_k \mathbb{B}_{X^*} \qquad (k \in \mathbb{N}).$$

In particular, there exist $\tilde{x}_k \in \mathbb{B}(x_k, \delta_k)$ and $\tilde{x}_k^* \in N_C(\tilde{x}_k) \cap \mathbb{B}(x_k^*, \delta_k)$, which implies that $\tilde{x}_k \to x$ and $\tilde{x}_k^* \overset{*}{\rightharpoonup} x^*$ as desired. □

Remark 18.3. Theorem 18.2 can be extended to Asplund spaces – in particular to reflexive Banach spaces. The equivalence of (18.9) and (18.8) can, in fact, be used as a definition of an Asplund space. For details we refer to [Mordukhovich, 2006, Theorem 2.35].

Finally, the *Clarke tangent cone* is defined as in finite dimensions as

$$\begin{aligned} \widehat{T}_C(x) &:= \liminf_{\substack{C\ni\tilde{x}\to x,\\ \tau\searrow 0}} \frac{C-\tilde{x}}{\tau} \\ &= \left\{ \Delta x \in X \;\middle|\; \begin{array}{r} \text{for all } \tau_k \searrow 0,\ C \ni x_k \to x \text{ there exists } C \ni \tilde{x}_k \to x \\ \text{with } (\tilde{x}_k - x_k)/\tau_k \to \Delta x \end{array} \right\}. \end{aligned} \tag{18.10}$$

In infinite-dimensional spaces, however, we in general only have the inclusion $\widehat{T}_C(x) \subset \liminf_{C\ni\tilde{x}\to x} T_C(\tilde{x})$; see Corollary 18.20.

Remark 18.4 (a much too brief history of various cones). The (Bouligand) tangent cone was already introduced for smooth sets by Peano in 1908 [Peano, 1908]; the term *contingent cone* is due to Bouligand [Bouligand, 1930]. The Clarke tangent cone (also called *circatangent cone*) was introduced in [Clarke, 1973, 1975]; see also [Clarke, 1990]. The limiting normal cone can be found in [Mordukhovich, 1976], who stressed the need of defining (nonconvex) normal cones directly rather than as (necessarily convex) polars of tangent cones. The history of the Fréchet normal cone is harder to trace, but it has appeared in the literature as the polar of the tangent cone. We will see that in finite dimensions, $\widehat{N}_C(x) = T_C(x)^\circ$. In infinite dimensions, $T_C(x)^\circ$ is sometimes called the *Dini normal cone* and is in general not equal to the Fréchet normal cone.

We do not attempt to do full justice to the muddier parts of the historical development here, and rather refer to the accounts in [Bigolin and Golo, 2014; Dolecki and Greco, 2011] as well as [Rockafellar and Wets, 1998, Commentary to Ch. 6] and [Mordukhovich, 2018, Commentary to Ch. 1]. Various further cones are also discussed in [Aubin and Frankowska, 1990].

18.2 BASIC RELATIONSHIPS AND PROPERTIES

As seen in Example 18.1, the limiting normal cone $N_C(x)$ can be larger than the Fréchet normal cone $\widehat{N}_C(x)$; conversely, the Clarke tangent cone $\widehat{T}_C(x)$ is *smaller* than the tangent cone $T_C(x)$; see Figure 18.1. These inclusions hold in general.

Theorem 18.5. *Let $C \subset X$ and $x \in X$. Then*

(i) $\widehat{T}_C(x) \subset T_C(x) \subset T_C^w(x)$;

(ii) $\widehat{N}_C(x) \subset N_C(x)$.

Proof. If we fix the base point $\tilde{x}$ as x in the definition (18.10) of $\widehat{T}_C(x)$, the tangent inclusion $\widehat{T}_C(x) \subset T_C(x)$ is clear from the definition (18.1) of $T_C(x)$ as an outer limit and of $\widehat{T}_C(x)$ as an inner limit. The inclusion $T_C(x) \subset T_C^w(x)$ is likewise clear from the definition of $T^C(x)$ as a strong outer limit and of $T_C^w(x)$ is the corresponding weak outer limit.

The normal inclusion $\widehat{N}_C(x) \subset N_C(x)$ follows from the definition (18.8) of $N_C(x)$ as the outer limit of $\widehat{N}_C^\varepsilon(\tilde{x})$ as $\tilde{x} \to x$ and $\varepsilon \searrow 0$. (In finite dimensions, we can fix $\varepsilon = 0$ in this argument or refer to the equivalence of definitions shon in Theorem 18.2.) □

For a closed and convex set C, however, both the Fréchet and limiting normal cones coincide with the convex normal cone defined in Lemma 4.8 (which we here denote by $\partial\delta_C(x)$ to avoid confusion).

Lemma 18.6. *Let $C \subset X$ be nonempty, closed, and convex. Then for all $x \in X$,*

(i) $\widehat{N}_C(x) = \partial\delta_C(x)$;

(ii) if X is Gâteaux smooth (in particular, finite-dimensional), $N_C(x) = \partial\delta_C(x)$.

Proof. If $x \notin C$, it follows from their definitions that all three cones are empty. We can thus assume that $x \in C$.

(i): If $x^* \in \partial\delta_C(x)$, we have by definition that

$$\langle x^*, y - x\rangle_X \le 0 \quad \text{for all } y \in C.$$

Taking in particular $y = \tilde{x}$ and passing to the limit $\tilde{x} \to x$ thus implies that $x^* \in \widehat{N}_C(x)$.

Conversely, let $x^* \in \widehat{N}_C(x)$ and let $y \in C$ be arbitrary. Since C is convex, this implies that $x_t := x + t(y - x) \in C$ for any $t \in (0, 1)$ as well. We also have that $x_t \to x$ for $t \to 0$. From (18.2), it then follows by inserting the definition of x_t and dividing by $t > 0$ that

$$0 \ge \lim_{t\to 0} \frac{\langle x^*, x_t - x\rangle_X}{\|x_t - x\|_X} = \frac{\langle x^*, y - x\rangle_X}{\|y - x\|_X}.$$

and hence, since $y \in C$ was arbitrary, that $x^* \in \partial\delta_C(x)$.

(ii): By Lemmas 2.5 and 6.10 and Theorem 6.13, $\partial\delta_C$ is strong-to-weak-$*$ outer semicontinuous, which by Theorem 18.5 and the $\varepsilon \equiv 0$ characterization of Theorem 18.2 implies that

$$N_C(x) = \text{w-}*\text{-}\limsup_{\tilde{x}\to x} \widehat{N}_C(\tilde{x}) = \text{w-}*\text{-}\limsup_{\tilde{x}\to x} \partial\delta_C(\tilde{x}) \subset \partial\delta_C(x) = \widehat{N}_C(x) \subset N_C(x).$$

Hence $\widehat{N}_C(x) = N_C(x)$. □

Note that convexity was only used for the second inclusion, and hence $\partial\delta_C(x) \subset N_C(x)$ always holds. In general, comparing (18.2) with (16.2), we have the following relation.

Corollary 18.7. *Let $C \subset X$ and $x \in X$. Then $\widehat{N}_C(x) = \partial_F\delta_C(x)$.*

The next theorem lists some of the most basic properties of the various tangent and normal cones.

Theorem 18.8. *Let $C \subset X$ and $x \in X$. Then*

(i) $T_C(x)$, $\widehat{T}_C(x)$, $\widehat{N}_C(x)$, and $N_C(x)$ are cones;

(ii) $T_C(x)$, $\widehat{T}_C(x)$, and $\widehat{N}^\varepsilon_C(x)$ for every $\varepsilon \geq 0$ are closed;

(iii) $\widehat{T}_C(x)$ and $\widehat{N}^\varepsilon_C(x)$ for every $\varepsilon \geq 0$ are convex;

(iv) if X is finite-dimensional, then $N_C(x)$ is closed.

Proof. We argue the different properties for each type of cone in turn.

The Fréchet (ε-)normal cone: It is clear from the definition of $\widehat{N}_C(x)$ that it is a cone, i.e., that $x^* \in \widehat{N}_C(x)$ implies that $\lambda x^* \in \widehat{N}_C(x)$ for all $\lambda > 0$.

Let now $\varepsilon \geq 0$ be arbitrary. Let $x^*_k \in \widehat{N}^\varepsilon_C(x)$ converge to some $x^* \in X^*$. Also suppose $C \ni x_\ell \to x$. Then for any $\ell, k \in \mathbb{N}$, we have by the Cauchy–Schwarz inequality that

$$\frac{\langle x^*, x_\ell - x\rangle_X}{\|x_\ell - x\|_X} \leq \frac{\langle x^*_k, x_\ell - x\rangle_X}{\|x_\ell - x\|_X} + \|x^*_k - x^*\|_X$$

and thus that

$$\limsup_{\ell\to\infty} \frac{\langle x^*, x_\ell - x\rangle_X}{\|x_\ell - x\|_X} \leq \varepsilon + \|x^*_k - x^*\|_X.$$

Since $k \in \mathbb{N}$ was arbitrary and $x^*_k \to x^*$, we see that $x^* \in \widehat{N}^\varepsilon_C(x)$ and may conclude that $\widehat{N}^\varepsilon_C(x)$ is closed.

To show convexity, take $x_1^*, x_2^* \in \widehat{N}_C^\varepsilon(x)$ and let $x^* := \lambda x_1^* + (1-\lambda)x_2^*$ for some $\lambda \in (0,1)$. We then have

$$\frac{\langle x^*, x_\ell - x\rangle_X}{\|x_\ell - x\|_X} = \lambda \frac{\langle x_1^*, x_\ell - x\rangle_X}{\|x_\ell - x\|_X} + (1-\lambda)\frac{\langle x_2^*, x_\ell - x\rangle_X}{\|x_\ell - x\|_X}.$$

Taking the limit $x_\ell \to x$ now yields $x^* \in \widehat{N}_C^\varepsilon(x)$ and hence the convexity.

The limiting normal cone: If X is finite-dimensional, the set $N_C(x)$ is a closed cone as the strong outer limit of the (closed) cones $\widehat{N}_C(x_\ell)$ as $x_\ell \to x$; see Lemma 6.2.

The tangent cone: By Lemma 6.2, $T_C(x)$ is closed as the outer limit of the sets $C_\tau := (C-x)/\tau$ as $\tau \searrow 0$. To see that it is a cone, suppose $\Delta x \in T_C(x)$. Then there exist by definition $\tau_k \searrow 0$ and $C \ni x_k \to x$ such that $(x_k - x)/\tau_k \to \Delta x$. Now, for any $\lambda > 0$, taking $\tilde{\tau}_k := \lambda^{-1}\tau_k$, we have $(x_k - x)/\tilde{\tau}_k \to \lambda \Delta x$. Hence $\lambda \Delta x \in T_C(x)$.

The Clarke tangent cone: Finally, $\widehat{T}_C(x)$ is a closed set through its definition as an inner limit, cf. Corollary 6.3, as well as a cone by analogous arguments as for $T_C(x)$. To see that it is convex, take $\Delta x^1, \Delta x^2 \in \widehat{T}_C(x)$. Since $\widehat{T}_C(x)$ is a cone, we only need to show that $\Delta x := \Delta x^1 + \Delta x^2 \in \widehat{T}_C(x)$. By the definition of $\widehat{T}_C(x)$ as an inner limit, we therefore have to show that for any sequence $\tau_k \searrow 0$ and any "base point sequence" $C \ni x_k \to x$, there exist $\tilde{x}_k \in C$ such that $(\tilde{x}_k - x_k)/\tau_k \to \Delta x$. We do this by using the varying base point in the definition of $\widehat{T}_C(x)$ to "bridge" between the sequences generating Δx_1 and Δx_2; see Figure 18.2. First, since $\Delta x^1 \in \widehat{T}_C(x)$, by the very same definition of $\widehat{T}_C(x)$ as an inner limit, we can find for the base point sequence $\{x_k\}_{k\in\mathbb{N}}$ points $C \ni x_k^1 \to X$ with $(x_k^1 - x_k)/\tau_k \to \Delta x^1$. Continuing in the same way, since $\Delta x^2 \in \widehat{T}_C(x)$, we can now find with $\{x_k^1\}_{k\in\mathbb{N}}$ as the base point sequence points $x_k^2 \in C$ such that $(x_k^2 - x_k^1)/\tau_k \to \Delta x^2$. It follows

$$\frac{x_k^2 - x_k}{\tau_k} = \frac{x_k^2 - x_k^1}{\tau_k} + \frac{x_k^1 - x_k}{\tau_k} \to \Delta x^1 + \Delta x^2 = \Delta x.$$

Thus $\{\tilde{x}_k\}_{k\in\mathbb{N}} = \{x_k^2\}_{k\in\mathbb{N}}$ is the sequence we are looking for, showing that $\Delta x \in \widehat{T}_C(x)$ and hence that the Clarke tangent cone is convex. □

One might expect $T_C^w(x)$ to be weakly closed and $N_C(x)$ to be weak-$*$ closed. However, this is not necessarily the case, since weak and weak-$*$ inner and outer limits need not be closed in the respective topologies. Consequently, N_C may also not be (strong-to-weak-$*$) outer semicontinuous at a point x, as this would imply $N_C(x)$ to be weak-$*$ closed and hence closed. However, in finite dimensions we do have outer semicontinuity.

Corollary 18.9. *If X is finite-dimensional, then the mapping $x \mapsto N_C(x)$ is outer semicontinuous.*

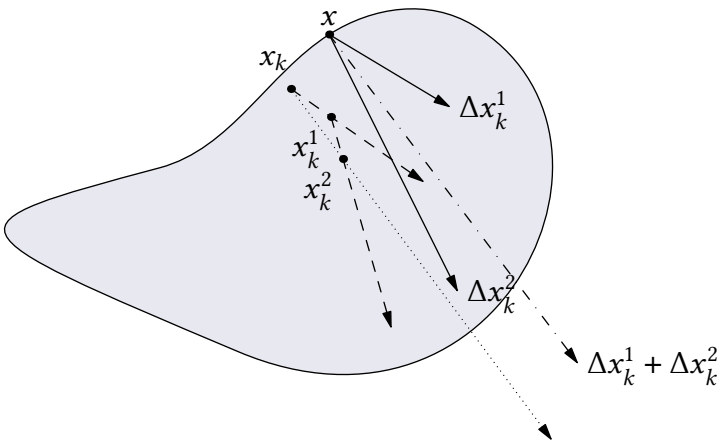


Figure 18.2: Illustration of the "bridging" argument in the proof of Theorem 18.8. As x_k converges to x, the dashed arrows converge to the solid arrows, while the dotted arrow converges to the dash-dotted one, which depicts the point $\Delta x_k^1 + \Delta x_k^2$ that we are trying to prove to be in $\widehat{T}_C(x)$.

Proof. Let $C \ni x_k \to x$ and $x_k^* \in N_C(x_k)$ with $x_k^* \to x^*$. Then for $\delta_k \searrow 0$, the definition (18.3) provides $\tilde{x}_k \in C$ and $\tilde{x}_k^* \in \widehat{N}_C(\tilde{x}_k)$ with $\|\tilde{x}_k^* - x_k^*\| \le \delta_k$ and $\|\tilde{x}_k - x_k\| \le \delta_k$. It follows that $C \ni \tilde{x}_k \to x$ and $\tilde{x}_k^* \to x^*$ with $\tilde{x}_k^* \in \widehat{N}_C(\tilde{x}_k)$. Thus by definition, $x^* \in N_C(x)$, and hence N_C is outer semicontinuous. □

18.3 POLARITY AND LIMITING RELATIONSHIPS

The tangent and normal cones satisfy various polarity relationships. To state these, recall from Section 1.2 for a general set $C \subset X$ the definition of the *polar cone*

$$C^\circ = \{x^* \in X^* \mid \langle x^*, x\rangle_X \le 0 \text{ for all } x \in C\}$$

as well as of the *bipolar cone* $C^{\circ\circ} = (C^\circ)_\circ \subset X$.

THE FUNDAMENTAL CONES

The relations in the following result will be crucial.

Lemma 18.10. *Let X be a Banach space, $C \subset X$, and $x \in X$. Then*

(i) $\widehat{N}_C(x) \subset T_C^w(x)^\circ \subset T_C(x)^\circ$;

(ii) if X is reflexive, then $\widehat{N}_C(x) = T_C^w(x)^\circ$;

(iii) if X is finite-dimensional, then $\widehat{N}_C(x) = T_C(x)^\circ$.

Proof. (i): We take $\Delta x \in T_C^w(x)$ and $x^* \in \widehat{N}_C(x)$. Then there exist $\tau_k \searrow 0$ and $C \ni x_k \to x$ such that $(x_k - x)/\tau_k \rightharpoonup \Delta x$ weakly in X. Thus

$$\langle x^*, \Delta x\rangle_X = \limsup_{k\to\infty} \frac{\langle x^*, x_k - x\rangle_X}{\tau_k} = \limsup_{k\to\infty} \frac{\langle x^*, x_k - x\rangle_X}{\|x_k - x\|_X} \cdot \frac{\|x_k - x\|_X}{\tau_k}.$$

Since $x^* \in \widehat{N}_C(x)$ and $C \ni x_k \to x$, we have by definition that $\limsup_{k\to\infty} \langle x^*, x_k - x\rangle_X / \|x_k - x\|_X \le 0$. Moreover, $(x_k - x)/\tau_k \rightharpoonup \Delta x$ implies that $\|x_k - x\|_X/\tau_k$ is bounded. Passing to the limit, it therefore follows that $\langle x^*, \Delta x\rangle_X \le 0$. Since this holds for every $\Delta x \in T_C^w(x)$, we see that $x^* \in T_C^w(x)^\circ$. This shows that $\widehat{N}_C(x) \subset T_C^w(x)^\circ$. Since $T_C(x) \subset T_C^w(x)$ by Theorem 18.5, $T_C^w(x)^\circ \subset T_C(x)^\circ$ follows from Theorem 1.8.

(ii): Due to (i), we only need to show "$\supset$". Let $x^* \notin \widehat{N}_C(x)$. Then, by definition, there exist $C \ni x_k \to x$ with

$$\lim_{k\to\infty} \langle x^*, \Delta x_k\rangle > 0 \quad \text{for} \quad \Delta x_k := \frac{x_k - x}{\|x_k - x\|_X}. \tag{18.11}$$

We now use the reflexivity of X and the Eberlein–Šmulyan Theorem 1.9 to pass to a subsequence, unrelabelled, such that that $\Delta x_k \rightharpoonup \Delta x$ for some $\Delta x \in X$ that by definition satisfies $\Delta x \in T_C^w(x)$. However, passing to the limit in (18.11) now shows that $\langle x^*, \Delta x\rangle_X > 0$ and hence that $x^* \notin T_C^w(x)^\circ$.

(iii): This is immediate from (ii) since $T_C(x) = T_C^w(x)$ in finite-dimensional spaces. □

THE LIMITING CONES: PRELIMINARY LEMMAS

For a polarity relationship between the basic normal cone and the Clarke tangent cone, we need to work significantly harder. We start here with some preliminary lemmas shared between the finite-dimensional and infinite-dimensional setting, and then treat the two in that order.

Lemma 18.11. *Let X be a reflexive Banach space, $C \subset X$, and $x \in X$. Then*

$$\widehat{T}_C(x) \subset \liminf_{C\ni\tilde{x}\to x} T_C^w(\tilde{x}). \tag{18.12}$$

If $X = \mathbb{R}^N$, then

$$\widehat{T}_C(x) \subset \liminf_{C\ni\tilde{x}\to x} T_C(\tilde{x}).$$

Proof. The case $X = \mathbb{R}^N$ trivially follows from (18.12). To prove (18.12), denote by K the set on its right-hand side. If $\Delta x \notin K$, then there exist $\varepsilon > 0$ and a sequence $C \ni x_k \to x$ such that

$$\inf_{\Delta x_k \in T_C^w(x_k)} \|\Delta x_k - \Delta x\|_X \geq 3\varepsilon. \tag{18.13}$$

Fix $k \in \mathbb{N}$ and suppose that for some $\tau_\ell \searrow 0$ and $\tilde{x}_\ell \in C$,

$$\left\| \tfrac{\tilde{x}_\ell - x_k}{\tau_\ell} - \Delta x \right\|_X \leq 2\varepsilon \qquad (\ell \in \mathbb{N}). \tag{18.14}$$

Using the reflexivity of X and the Eberlein–Šmulyan Theorem 1.9, we then find a further, unrelabelled, subsequence of $\{(\tilde{x}_\ell, \tau_\ell)\}_{\ell\in\mathbb{N}}$ such that $(\tilde{x}_\ell - x_k)/\tau_k\ell \rightharpoonup \Delta x_k$ as $\ell \to \infty$ for some $\Delta x_k \in T_C^w(x_k)$ with $\|\Delta x_k - \Delta x\|_X \leq 2\varepsilon$, in contradiction to (18.13). We thus have

$$\lim_{\tau\searrow 0} \inf_{\tilde{x}\in C} \left\| \tfrac{\tilde{x}-x_k}{\tau} - \Delta x \right\|_X \geq 2\varepsilon.$$

Since this holds for all $k \in \mathbb{N}$, we can find $\tau_k > 0$ with $\tau_k \searrow 0$ satisfying the inequality

$$\liminf_{k\to\infty} \inf_{\tilde{x}\in C} \left\| \tfrac{\tilde{x}-x_k}{\tau_k} - \Delta x \right\|_X \geq \varepsilon$$

implying that $\Delta x \notin \widehat{T}_C(x)$. Therefore (18.12) holds. □

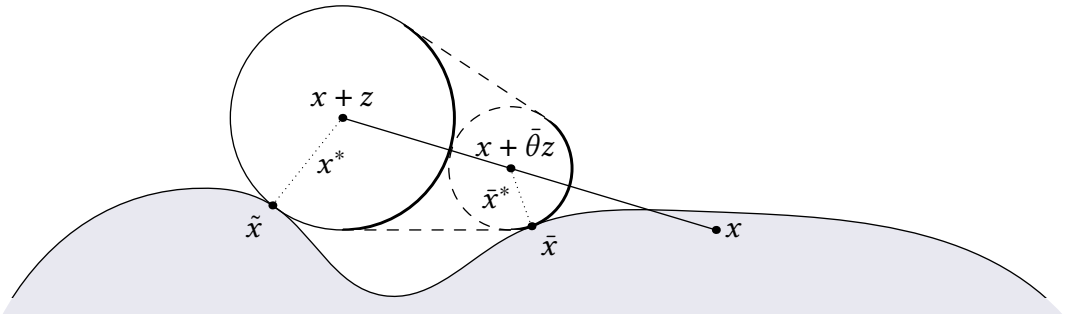
Lemma 18.12. *Let X be a reflexive and Gâteaux smooth (or finite-dimensional) Banach space, $C \subset X$, and $x \in X$. Then*

$$\widehat{T}_C(x) \subset N_C(x)^\circ.$$

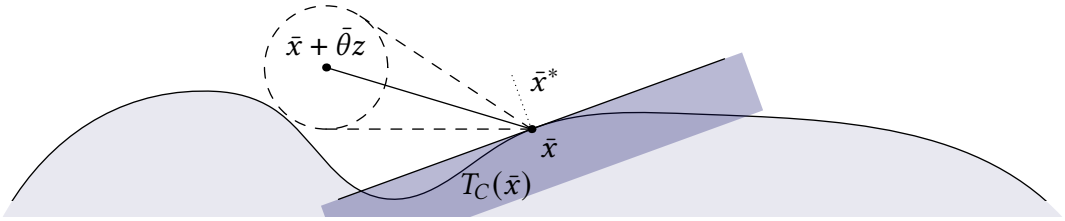
Proof. Take $x^* \in N_C(x)$ and $\Delta x \in \widehat{T}_C(x)$. This gives by Theorem 18.2 (or (18.3) if X is finite-dimensional) sequences $x_k \to x$ and $x_k^* \overset{*}{\rightharpoonup} x^*$ with $x_k^* \in \widehat{N}_C(x_k)$. By Lemma 18.11, we can find for each $k \in \mathbb{N}$ a $\Delta x_k \in T_C^w(x_k)$ such that $\Delta x_k \to \Delta x$. Since $\widehat{N}_C(x_k) = T_C^w(x_k)^\circ$ by Lemma 18.10 (ii) when X is reflexive, we have $\langle x_k^*, \Delta x_k\rangle_X \leq 0$. Combining all these observations, we obtain

$$\begin{aligned}\langle x^*, \Delta x\rangle_X &= \lim_{k\to\infty} \left(\langle x_k^*, \Delta x_k\rangle_X + \langle x^* - x_k^*, \Delta x\rangle_X + \langle x_k^*, \Delta x - \Delta x_k\rangle_X\right)\\ &= \lim_{k\to\infty} \langle x_k^*, \Delta x_k\rangle_X \leq 0.\end{aligned}$$

Since $x^* \in N_C(x)$ was arbitrary, we deduce that $\Delta x \in N_C(x)^\circ$ and hence the claim. □



(a) By assumption, the interior of the ball around $x+z$ of radius ε does not intersect C (shaded). In this example, the point $\tilde x \in C$ intersects the boundary; however, it is not on the leading edge (thick lines) where the normal vector x^* would satisfy $\langle z, x^*\rangle \geq \varepsilon$. Reducing $\theta < 1$ produces an intersecting point $\bar x$ on the leading edge.



(b) The "ice cream cone" emanating from $\bar x$ along the line $[\bar x, \bar x+\bar\theta z]$ with a ball of radius $\bar\varepsilon\bar\theta$ does not intersect C (light shading). From this it follows that the tangent cone $T_C(\bar x)$ (incomplete dark shading) is at a distance $\bar\varepsilon$ from z.

Figure 18.3: Geometric illustration of the construction in the proof of Lemma 18.13.

THE LIMITING CONES IN FINITE DIMENSIONS

We now start our development of polarity relationships between the limiting cones, as well as limiting relationships between the tangent and Clarke tangent cones. Our main tool will be the following "ice cream cone lemma", for which it is important that we endow $\mathbb{R}^N$ with the Euclidean norm.

Lemma 18.13. *Let $C \subset \mathbb{R}^N$ be closed and let $x \in C$. Let $z \in \mathbb{R}^N \setminus \{0\}$ and $\varepsilon > 0$ be such that*

$$\operatorname{int} \mathbb{B}(x+z, \varepsilon) \cap C = \emptyset. \tag{18.15}$$

Then for any $\bar\varepsilon \in (0, \varepsilon)$, there exists an $\bar x \in C$ such that there exist

(i) $\bar\theta \in (0,1]$ satisfying $\|(\bar x + \bar\theta z) - (x+z)\| \leq \bar\varepsilon$ and $\inf_{\Delta x \in T_C(\bar x)} \|\Delta x - z\| \geq \bar\varepsilon$;

(ii) $\bar x^* \in \widehat N_C(\bar x)$ satisfying $\langle \bar x^*, z\rangle \geq \bar\varepsilon$ and $\|\bar x^*\| \leq 1$.

Proof. We define the increasing real function $\varphi(t) := \sqrt{1+t^2}$ and $F, G : \mathbb{R}^N \times \mathbb{R} \to \overline{\mathbb{R}}$ by

$$F(\tilde x, \theta) := \varphi(\bar\varepsilon)\theta + \varphi(\|(\tilde x + \theta z) - (x+z)\|) \quad \text{and} \quad G(\tilde x, \theta) := \delta_C(\tilde x) + \delta_{[0,\infty)}(\theta).$$

Then $F+G$ is proper, coercive, and lower semicontinuous and hence admits a minimizer $(\bar{x},\bar{\theta}) \in C\times[0,\infty)$ by Theorem 2.1. (We illustrate the idea of such a minimizer geometrically in Figure 18.3.) Let $\bar{y} := (\bar{x}+\bar{\theta}z) - (x+z)$.

(i): We first prove $\bar{\theta} \in (0,1]$. Suppose $\bar{\theta} = 0$. Since $\bar{x} \in C$ we obtain using (18.15) that

$$[F+G](\bar{x},0) = \varphi(\|\bar{x}-(x+z)\|) \geq \varphi(\varepsilon) > \varphi(\bar{\varepsilon}) = [F+G](x,1).$$

This is a contradiction to $(\bar{x},0)$ being a minimizer. Thus $\bar{\theta} \neq 0$. Likewise,

$$\varphi(\bar{\varepsilon})\bar{\theta} + \varphi(\|\bar{y}\|) = [F+G](\bar{x},\bar{\theta}) \leq [F+G](x,1) = \varphi(\bar{\varepsilon}),$$

where both terms on the left-hand side are nonnegative. Hence $\bar{\theta} \leq 1$. By the monotonicity of φ, this also verifies the claim $\|\bar{y}\| \leq \bar{\varepsilon}$.

We still need to prove the claim on the tangent cone. Since $(\bar{x},\bar{\theta})$ is a minimizer of $F+G$, for any $\tilde{\theta} \geq 0$ and $\tilde{x} \in C$ we have

$$\varphi(\bar{\varepsilon})\bar{\theta} + \varphi(\|\bar{y}\|) \leq [F+G](\bar{x},\bar{\theta}) \leq [F+G](\tilde{x},\tilde{\theta}) = \varphi(\bar{\varepsilon})\tilde{\theta} + \varphi(\|y\|).$$

Letting $y := (\tilde{x}+\tilde{\theta}z) - (x+z)$ and using first this inequality and then the convexity of φ with $\varphi'(t) = t/\varphi(t) \leq 1$ for all $t \geq 0$ yields

$$\begin{aligned}\varphi(\bar{\varepsilon})(\bar{\theta}-\tilde{\theta}) &\leq \varphi(\|\bar{y}\|) - \varphi(\|y\|)\\ &\leq \varphi'(\|\bar{y}\|)\,(\|\bar{y}\| - \|y\|)\\ &\leq \|\bar{y}-y\| = \|\tilde{x}-\bar{x}-(\bar{\theta}-\tilde{\theta})z\|.\end{aligned}$$

Dividing by $\tau = \bar{\theta}-\tilde{\theta}$ for $\tilde{\theta} \in [0,\bar{\theta})$, we obtain that $\bar{\varepsilon} \leq \varphi(\bar{\varepsilon}) \leq \left\|\frac{\tilde{x}-\bar{x}}{\tau} - z\right\|$. Taking the infimum over $\tilde{x} \in C$ and $\tau \in (0,\bar{\theta}]$ thus yields $\inf_{\Delta x \in T_C(\bar{x})} \|\Delta x - z\| \geq \bar{\varepsilon}$.

(ii): By Lemma 3.4 (iv), F is convex. Furthermore, $\operatorname{int}(\operatorname{dom} F) = \mathbb{R}^{N+1}$ so that F is Lipschitz near $(\bar{x},\bar{\theta})$ by Theorem 3.13. Using Theorems 4.6, 4.17, and 4.19 with $K(x,\theta) := x+\theta z$, it follows that

$$\partial F(\bar{x},\bar{\theta}) = \left\{ \begin{pmatrix} \varphi'(\|\bar{y}\|)\,y^* \\ \varphi(\bar{\varepsilon}) + \varphi'(\|\bar{y}\|)\langle z, y^*\rangle \end{pmatrix} \;\middle|\; \begin{array}{ll} \langle y^*, \bar{y}\rangle = \|\bar{y}\|, \|y^*\| = 1 & \text{if } \bar{y} \neq 0 \\ \|y^*\| \leq 1 & \text{if } \bar{y} = 0 \end{array} \right\}. \tag{18.16}$$

Since $\mathbb{R}^N$ endowed with the euclidean norm is a Hilbert space, $x \mapsto \|x\|^2$ is Gâteaux differentiable by Example 17.6 (i) and Lemma 17.7. Hence $\partial F(\bar{x},\bar{\theta})$ is a singleton, and therefore F is Gâteaux differentiable at $(\bar{x},\bar{\theta})$ due to Lemma 13.7 and Theorem 13.8. We can thus apply the Fermat principle (Theorem 16.2) and the Fréchet sum rule (Corollary 17.3) to deduce $0 \in \partial_F F(\bar{x},\bar{\theta}) + \partial_F G(\bar{x},\bar{\theta})$. Since $\bar{\theta} > 0$, we have $\partial_F G(\bar{x},\bar{\theta}) = \widehat{N}_C(\bar{x}) \times \{0\}$ by Corollary 18.7, which implies that

$$-\varphi'(\|\bar{y}\|)\,y^* \in \widehat{N}_C(\bar{x}) \quad \text{and} \quad \varphi(\bar{\varepsilon}) + \varphi'(\|\bar{y}\|)\langle z, y^*\rangle = 0. \tag{18.17}$$

Since $\varphi(\bar{\varepsilon}) > 0$, the second equation in (18.17) yields $\varphi'(\|\bar{y}\|) \neq 0$ as well. As $\varphi'(t) \in (0,1)$ and $\varphi(t) > t$ for all $t > 0$, we can set $x^* := -\varphi'(\|\bar{y}\|)\,y^*$ to obtain $x^* \in \widehat{N}_C(\bar{x})$ with $\|x^*\| \leq 1$ and $\langle z, x^*\rangle = \frac{\varphi(\bar{\varepsilon})}{\varphi'(\|\bar{y}\|)} \geq \varphi(\bar{\varepsilon}) \geq \bar{\varepsilon}$. □

The following consequence of the ice cream cone lemma will be useful for several polarity relations. We call a set C *closed near* $x \in C$, if there exists a $\delta > 0$ such that $C \cap \mathbb{B}(x, \delta)$ is closed.

Lemma 18.14. *Let $C \subset \mathbb{R}^N$ be closed near x. If $z \notin \widehat{T}_C(x)$, then there exist $\tilde{\varepsilon} > 0$ and a sequence $C \ni \tilde{x}_k \to x$ such that for all $k \in \mathbb{N}$,*

(i) $\inf_{\Delta\tilde{x}_k \in T_C(\tilde{x}_k)} \|\Delta\tilde{x}_k - z\| \geq \tilde{\varepsilon}$;

(ii) there exists $\tilde{x}_k^* \in \widehat{N}_C(\tilde{x}_k)$ with $\|\tilde{x}_k^*\| \leq 1$ and $\langle \tilde{x}_k^*, z\rangle \geq \tilde{\varepsilon}$.

Proof. First, $z \notin \widehat{T}_C(x)$ implies by (18.10) the existence of $\varepsilon > 0$, $C \ni x_k \to x$, and $\tau_k \searrow 0$ such that

$$\inf_{\tilde{x} \in C} \left\| \tfrac{\tilde{x} - x_k}{\tau_k} - z \right\| \geq \varepsilon \qquad (k \in \mathbb{N}),$$

implying that

$$\operatorname{int} \mathbb{B}(x_k + \tau_k z, \tau_k \varepsilon) \cap C = \emptyset.$$

By taking τ_k small enough – i.e., $k \in \mathbb{N}$ large enough – we may without loss of generality assume that C is closed. For any $\tilde{\varepsilon} \in (0, \varepsilon)$ and every $k \in \mathbb{N}$, Lemma 18.13 now yields $\tilde{x}_k \in C$ and $\tilde{\theta}_k \in (0, 1]$ satisfying

(i') $\|(\tilde{x}_k + \tilde{\theta}_k \tau_k z) - (x + \tau_k z)\| \leq \tilde{\varepsilon}\tau_k$ and $\inf_{\Delta\tilde{x}_k \in T_C(\tilde{x}_k)} \|\Delta\tilde{x}_k - \tau_k z\| \geq \tilde{\varepsilon}\tau_k$;

(ii') there exists an $\tilde{x}_k^* \in \widehat{N}_C(\tilde{x}_k)$ such that $\langle \tilde{x}_k^*, \tau_k z\rangle \geq \tau_k \tilde{\varepsilon}$ and $\|\tilde{x}_k^*\| \leq 1$.

We readily obtain (i) from (i') and (ii) from (ii') Since (i') also shows that $\tilde{x}_k \to x$ as $\tau_k \searrow 0$, this finishes the proof. □

We can now show the converse inclusion of Lemma 18.12 when the set is closed near x.

Theorem 18.15. *If $C \subset \mathbb{R}^N$ is closed near x, then*

$$\widehat{T}_C(x) = N_C(x)^\circ.$$

Proof. By Lemma 18.12, we only need to prove $\widehat{T}_C(x) \supset N_C(x)^\circ$. We argue by contraposition. Let $z \notin \widehat{T}_C(x)$. Then Lemma 18.14 yields a sequence $\{x_k^*\}_{k\in\mathbb{N}} \subset \mathbb{R}^N$ such that $x_k^* \in \widehat{N}_C(x_k)$ for $C \ni x_k \to x$ and $\langle x_k^*, z\rangle \geq \varepsilon > 0$ as well as $\|x_k^*\| \leq 1$. Since $\{x_k^*\}_{k\in\mathbb{N}}$ is bounded, we can extract a subsequence that converges to some $x^* \in \mathbb{R}^N$. By definition of the limiting normal cone, $x^* \in N_C(x)$. Moreover, $\langle x^*, z\rangle \geq \varepsilon > 0$. This provides, as required, that $z \notin N_C(x)^\circ$. □

THE LIMITING CONES IN INFINITE DIMENSIONS

We now repeat the arguments above in infinite dimensions, however, we need extra care and extra assumptions. Besides reflexivity (to obtain weak-$*$ compactness from Eberlein–Šmulyan Theorem 1.9) and Gâteaux smoothness (to obtain differentiability of the norm), we need to use the approximate Fermat principle of Theorem 17.13 since exact projections to general sets C may not exist; compare Theorem 17.14. This introduces ε-normal cones into the proof. The geometric ideas of the proof, however, are the same as illustrated in Figure 18.3.

Lemma 18.16. *Let X be a Banach space, $C \subset X$ be closed, and $x \in C$. Let $z \in X \setminus \{0\}$ and $\varepsilon > 0$ be such that*

$$\operatorname{int} \mathbb{B}(x+z, \varepsilon) \cap C = \emptyset. \tag{18.18}$$

Then for any $\bar\varepsilon \in (0, \varepsilon)$ and $\rho > 0$, there exists $\bar x \in C$ such that there exist

(i) $\bar\theta \in (0,1]$ such that $\|(\bar x + \bar\theta z) - (x+z)\|_X \le \bar\varepsilon$ and $\inf_{\Delta x \in T_C(\bar x)} \|\Delta x - z\|_X \ge \bar\varepsilon$;

(ii) if X is Gâteaux smooth, $\bar x^* \in \widehat N_C^\rho(\bar x)$ such that $\langle \bar x^*, z\rangle_X \ge \bar\varepsilon$ and $\|\bar x^*\|_{X^*} \le 1$.

Proof. We define the convex and increasing real function $\varphi(t) := \sqrt{1+t^2}$ and pick arbitrary

$$\tilde\varepsilon \in (\bar\varepsilon, \varepsilon), \quad 0 < \rho < \frac{\varphi(\tilde\varepsilon) - \bar\varepsilon}{2 + \bar\varepsilon}, \quad \text{and} \quad 0 < \delta < \varphi(\varepsilon) - \varphi(\tilde\varepsilon). \tag{18.19}$$

The upper bound on ρ is without loss of generality for (ii) because $\widehat N_C^\rho(\bar x) \subset \widehat N_C^{\rho'}(\bar x)$ for $\rho' \ge \rho$. Then we define $F, G : X \times \mathbb{R} \to \overline{\mathbb{R}}$ by

$$F(\tilde x, \theta) := \varphi(\tilde\varepsilon)\theta + \varphi(\|(\tilde x + \theta z) - (x+z)\|_X) \quad \text{and} \quad G(\tilde x, \theta) := \delta_C(\tilde x) + \delta_{[0,\infty)}(\theta).$$

The function $F + G$ is proper and coercive, hence $\inf(F+G) > -\infty$. However, it may not admit a minimizer. Nevertheless, the approximate Fermat principle of Theorem 17.13 produces an approximate minimizer $(\bar x, \bar\theta) \in C \times [0,\infty)$ with

(a) $[F+G](\bar x, \bar\theta) \le \inf[F+G] + \delta$,

(b) $[F+G](\bar x, \bar\theta) < [F+G](\tilde x, \theta) + \rho\|\tilde x - \bar x\|_X + \rho|\theta - \bar\theta|$ for all $(\tilde x, \theta) \ne (\bar x, \bar\theta)$, and

(c) $0 \in \partial_\rho [F+G](\bar x, \bar\theta)$.

Let again $\bar y := (\bar x + \bar\theta z) - (x+z)$.

(i): We first prove $\bar\theta \in (0, \frac{\varphi(\varepsilon)}{\varphi(\tilde\varepsilon)}]$, which will in particular imply that $\bar\theta \in (0, 1+\varepsilon)$. Suppose $\bar\theta = 0$. Since $\bar x \in C$, using (18.18) and the convexity of φ, we obtain

$$[F+G](\bar x, 0) - \delta = \varphi(\|\bar x - (x+z)\|_X) - \delta \ge \varphi(\varepsilon) - \delta > \varphi(\tilde\varepsilon) = [F+G](x, 1)$$

in contradiction to (a). Thus $\bar\theta \neq 0$. Likewise,

$$\varphi(\tilde\varepsilon)\bar\theta + \varphi(\|\bar y\|_X) = [F+G](\bar x, \bar\theta) \leq [F+G](x,1) + \delta = \varphi(\tilde\varepsilon) + \delta < \varphi(\varepsilon).$$

where both terms on the left-hand side are nonnegative. Hence $\bar\theta \leq \frac{\varphi(\varepsilon)}{\varphi(\tilde\varepsilon)}$. By monotonicity of φ, this also verifies the claim $\|\bar y\|_X \leq \varepsilon$.

We still need to prove the claim on the tangent cone. Letting $y := (\tilde x + \theta z) - (x + z)$, we rearrange (b) as

$$\varphi(\tilde\varepsilon)(\bar\theta - \theta) - \rho|\theta - \bar\theta| \leq \varphi(\|\bar y\|_X) - \varphi(\|y\|_X) + \rho\|\tilde x - \bar x\|_X \tag{18.20}$$

Using the convexity of φ, we also have

$$\varphi(\|\bar y\|_X) - \varphi(\|y\|_X) \leq \frac{1}{\varphi(\|\bar y\|_X)}\,(\|\bar y\|_X - \|y\|_X) \leq \|\bar y - y\|_X = \|\tilde x - \bar x - (\bar\theta - \theta)z\|_X$$

Further estimating $\|\tilde x - \bar x\|_X \leq \|\tilde x - \bar x - (\bar\theta - \theta)z\|_X + |\bar\theta - \theta|$, (18.20) now yields

$$[\varphi(\tilde\varepsilon) - 2\rho](\bar\theta - \theta) \leq (1+\rho)\|\tilde x - \bar x - (\bar\theta - \theta)z\|_X \qquad (\theta \in [0, \bar\theta),\ \tilde x \in C).$$

Dividing by $(1+\rho)(\bar\theta - \theta)$ and using (18.19) (for the first inequality), we obtain that

$$\bar\varepsilon \leq \frac{\varphi(\tilde\varepsilon) - 2\rho}{1+\rho} \leq \inf_{\tilde x \in C,\, \theta \in [0,\bar\theta)} \left\| \frac{\tilde x - \bar x}{\bar\theta - \theta} - z \right\|_X .$$

This shows $\inf_{\Delta x \in T_C(\bar x)} \|\Delta x - z\|_X \geq \bar\varepsilon$.

(ii): By Lemma 3.4 (iv), F is convex. Furthermore $\operatorname{int}(\operatorname{dom} F) = X \times \mathbb{R}$, and hence F is Lipschitz near $(\bar x, \bar\theta)$ by Theorem 3.13. Using Theorems 4.6, 4.17, and 4.19 with $K(x,\theta) := x + \theta z$, it follows that

$$\partial F(\bar x, \bar\theta) = \left\{ \begin{pmatrix} \varphi'(\|\bar y\|_X)\, y^* \\ \varphi(\tilde\varepsilon) + \varphi'(\|\bar y\|_X)\langle z, y^*\rangle_X \end{pmatrix} \;\middle|\; \begin{array}{ll} \langle y^*, \bar y\rangle_X = \|\bar y\|_X, \|y^*\|_{X^*} = 1 & \text{if } \bar y \neq 0 \\ \|y^*\|_{X^*} \leq 1 & \text{if } \bar y = 0 \end{array} \right\}. \tag{18.21}$$

Again, $\partial F(\bar x, \bar\theta)$ is a singleton by Lemma 17.7 and the assumption that X is Gâteaux smooth.

We can thus apply the ε-sum rule (Lemma 17.2) in (c) to deduce $0 \in \partial F(\bar x, \bar\theta) + \partial_\rho G(\bar x, \bar\theta)$. Since $\bar\theta > 0$, we have $\partial_\rho G(\bar x, \bar\theta) = \widehat N_C^\rho(\bar x) \times \{0\}$, which implies that

$$-\varphi'(\|\bar y\|_X)\, y^* \in \widehat N_C^\rho(\bar x) \quad \text{and} \quad \varphi(\tilde\varepsilon) + \varphi'(\|\bar y\|_X)\langle z, y^*\rangle_X = 0. \tag{18.22}$$

Since $\varphi(\tilde\varepsilon) > 0$, the second equation in (18.22) yields $\varphi'(\|\bar y\|_X) \neq 0$ as well. As $\varphi'(t) \in (0,1)$ and $\varphi(t) > t$ for all $t > 0$, we can set $x^* := -\varphi'(\|\bar y\|_X)\, y^*$ to obtain $x^* \in \widehat N_C^\rho(\bar x)$ with $\|x^*\|_{X^*} \leq 1$ and $\langle z, x^*\rangle_X = \frac{\varphi(\tilde\varepsilon)}{\varphi'(\|\bar y\|_X)} \geq \varphi(\tilde\varepsilon) \geq \bar\varepsilon$. □

Remark 18.17. If X is in addition reflexive, we can use the Eberlein–Šmulyan Theorem 1.9 to pass to the limit as $\rho \searrow 0$ in Lemma 18.16 and produce $\bar x^* \in \widehat N_C(\bar x)$ satisfying the other claims of the lemma.

Lemma 18.18. *Let X be a Banach space and $C \subset X$ be closed near $x \in C$. If $z \notin \widehat{T}_C(x)$, then there exist $\tilde{\varepsilon} > 0$ and a sequence $C \ni \tilde{x}_k \to x$ such that for all $k \in \mathbb{N}$,*

(i) $\inf_{\Delta\tilde{x}_k \in T_C(\tilde{x}_k)} \|\Delta\tilde{x}_k - z\|_X \geq \tilde{\varepsilon}$;

(ii) if X is Gâteaux smooth, there exists $\tilde{x}_k^* \in \widehat{N}_C(\tilde{x}_k)$ with $\|\tilde{x}_k^*\|_{X^*} \leq 1$ and $\langle \tilde{x}_k^*, z\rangle_{X^*} \geq \tilde{\varepsilon}$.

Proof. The assumption $z \notin \widehat{T}_C(x)$ implies by (18.10) the existence of $\varepsilon > 0$, $C \ni x_k \to x$, and $\tau_k \searrow 0$ such that

$$\inf_{\tilde{x} \in C} \left\| \tfrac{\tilde{x} - x_k}{\tau_k} - z \right\|_X \geq \varepsilon \qquad (k \in \mathbb{N}).$$

This implies that

$$\operatorname{int} \mathbb{B}(x_k + \tau_k z, \tau_k \varepsilon) \cap C = \emptyset.$$

Since the argument is local, by taking τ_k small enough – i.e., $k \in \mathbb{N}$ large enough – we may without loss of generality assume that C is closed. For any $\tilde{\varepsilon} \in (0, \varepsilon)$ and every $k \in \mathbb{N}$, Lemma 18.16 now produces $\tilde{x}_k \in C$ and $\tilde{\theta}_k \in (0, 1]$ satisfying

(i') $\|(\tilde{x}_k + \tilde{\theta}_k \tau_k z) - (x + \tau_k z)\|_X \leq \tilde{\varepsilon}\tau_k$ and $\inf_{\Delta\tilde{x}_k \in T_C(\tilde{x}_k)} \|\Delta\tilde{x}_k - \tau_k z\|_X \geq \tilde{\varepsilon}\tau_k$;

(ii') if X is Gâteaux smooth, there exists $\tilde{x}_k^* \in \widehat{N}_C^{\tilde{\varepsilon}\tau_k}(\tilde{x}_k)$ such that $\langle x_k^*, \tau_k z\rangle_X \geq \tau_k \tilde{\varepsilon}$ and $\|\tilde{x}_k^*\|_{X^*} \leq 1$.

We readily obtain (i) from (i') and (ii) from (ii'). Since (i') also shows that $\tilde{x}_k \to x$ as $\tau_k \searrow 0$, this finishes the proof. □

Theorem 18.19. *Let X be a reflexive and Gâteaux smooth Banach space and let $C \subset X$ be closed near $x \in C$. Then*

$$\widehat{T}_C(x) = N_C(x)^\circ.$$

Proof. By Lemma 18.12, we only need to prove $\widehat{T}_C(x) \supset N_C(x)^\circ$. Let $z \notin \widehat{T}_C(x)$. Then Lemma 18.18 yields a sequence $\{\tilde{x}_k^*\}_{k\in\mathbb{N}} \subset \mathbb{B}_{X^*}$ such that $\tilde{x}_k^* \in \widehat{N}_C(x_k)$ and $\langle \tilde{x}_k^*, z\rangle_X \geq \varepsilon$. Since X is reflexive, X^* is reflexive as well, and so we can apply Theorem 1.9 to extract a subsequence of $\{\tilde{x}_k^*\}_{k\in\mathbb{N}}$ that converges weakly and thus, again by reflexivity, also weakly-$*$ to some $x^* \in N_C(x)$ (by definition of the limiting normal cone) with $\langle x^*, z\rangle_X \geq \varepsilon > 0$. □

THE CLARKE TANGENT CONE

We can now show the promised alternative characterization of the Clarke tangent cone $\widehat{T}_C(x)$ as the inner limit of tangent cones.

Corollary 18.20. *Let X be a reflexive Banach space and let $C \subset X$ be closed near $x \in X$. Then*

$$\liminf_{C \ni \tilde{x} \to x} T_C(\tilde{x}) \subset \widehat{T}_C(x) \subset \liminf_{C \ni \tilde{x} \to x} T_C^w(\tilde{x}). \tag{18.23}$$

In particular, if X is finite-dimensional, then

$$\widehat{T}_C(x) = \liminf_{C \ni \tilde{x} \to x} T_C(\tilde{x}).$$

Proof. We have already proved the second inclusion of (18.23) in Lemma 18.11. For the first inclusion, suppose $z \notin \widehat{T}_C(x)$. Then Lemma 18.18 yields an $\tilde{\varepsilon} > 0$ and a sequence $C \ni \tilde{x}_k \to x$ such that $\inf_{\Delta\tilde{x}_k \in T_C(\tilde{x}_k)} \|\Delta\tilde{x}_k - z\|_X \geq \tilde{\varepsilon}$ for all k. This shows that $z \notin \liminf_{C \ni \tilde{x} \to x} T_C(\tilde{x})$. □

Remark 18.21. Lemma 18.18 and thus the first inclusion of (18.23) do not actually require the reflexivity of X. In contrast , Lemma 18.11 and thus the second inclusion of (18.23) do not require the local closedness assumption. Besides X being reflexive, it holds more generally if X has the Radon–Riesz property and is Fréchet smooth; see [Mordukhovich, 2006, Theorem 1.9] and compare Remark 17.5.

18.4 REGULARITY

It stands to reason that without any assumptions on the set $C \subset X$ such as convexity, there is little hope of obtaining precise characterizations or exact transformation rules for the various cones. Similarly, precise characterizations or exact calculus rules for the derivatives of set-valued mappings – which, respectively, we will derive from the former – require strong assumptions on these mappings. This is especially true of the limiting cones. As befitting the introductory character of this textbook, we will therefore only develop calculus for the derivatives based on the limiting cones when they are equal the corresponding basic cones. This will allow deriving exact results that are nevertheless applicable to the situations we have been focusing on in the previous parts, such as problems of the form (P). These conditions can be compared to constraint qualifications in nonlinear optimization that guarantee that the tangent cone coincides with the linearization cone. However, "fuzzy" results are available under more general assumptions, for which we refer to the monographs [Aubin and Frankowska, 1990; Mordukhovich, 2006, 2018; Rockafellar and Wets, 1998].

Specifically, we say that $C \subset X$ is *tangentially regular* at $x \in C$ if $T_C(x) = \widehat{T}_C(x)$, and *normally regular* at x if $N_C(x) = \widehat{N}_C(x)$. We call C *regular* at x if C is both normally and tangentially regular.

Example 18.22. With $C \subset \mathbb{R}^2$ as in Example 18.1, we see that $C = \mathbb{B}(0,1)$ and $C = [0,1]^2$ are regular at every $x \in C$, while $C = [0,1]^2 \setminus [\frac{1}{2},1]^2$ is regular everywhere except at $x = (\frac{1}{2}, \frac{1}{2})$.

In finite dimensions, the two concepts of regularity are equivalent and have various characterizations. By Lemma 18.6, these hold in particular for closed convex sets.

Theorem 18.23. *Let $C \subset \mathbb{R}^N$ be closed near x. Then the following conditions are equivalent:*

(i) C is normally regular at x;

(ii) C is tangentially regular at x;

(iii) $\widehat{N}_C$ is outer semicontinuous at x;

(iv) T_C is inner semicontinuous at x (relative to C).

In particular, if any of these hold, C is regular at x.

Proof. (i) $\Leftrightarrow$ (ii): If (i) holds, then by Theorems 1.8, 18.5, and 18.15 and Lemma 18.10

$$T_C(x) \subset T_C(x)^{\circ\circ} = \widehat{N}_C(x)^\circ = N_C(x)^\circ = \widehat{T}_C(x) \subset T_C(x),$$

which shows (ii). The other direction is completely analogous, exchanging the roles of "N" and "T" to obtain

$$N_C(x) \subset N_C(x)^{\circ\circ} = \widehat{T}_C(x)^\circ = T_C(x)^\circ = \widehat{N}_C(x) \subset N_C(x).$$

(i) $\Leftrightarrow$ (iii): If (i) holds, then the outer semicontinuity of N_C (Corollary 18.9) and the inclusion $\widehat{N}_C(\tilde{x}) \subset N_C(\tilde{x})$ from Theorem 18.5 show that $\limsup_{\tilde{x}\to x} \widehat{N}_C(\tilde{x}) \subset \widehat{N}_C(x)$, i.e., the outer semicontinuity of $\widehat{N}_C$. Conversely, the outer semicontinuity of $\widehat{N}_C$ and the definition $N_C(x) = \limsup_{\tilde{x}\to x} \widehat{N}_C(\tilde{x})$ show that $N_C(x) \subset \widehat{N}_C(x)$. Combined with the inclusion $\widehat{N}_C(\tilde{x}) \subset N_C(\tilde{x})$ from Theorem 18.5, we obtain (i).

(ii) $\Leftrightarrow$ (iv): To show that (iv) implies (ii), recall from Corollary 18.20 that

$$\widehat{T}_C(x) = \liminf_{C \ni \tilde{x} \to x} T_C(\tilde{x}). \tag{18.24}$$

By the assumed inner semicontinuity and the definition of the inner limit, we thus obtain that $T_C(x) = \liminf_{C\ni\tilde{x}\to x} T_C(\tilde{x}) = \widehat{T}_C(x)$. For the other direction, we simply use $\widehat{T}_C(x) = T_C(x)$ in (18.24). □

Combining the previous result with Lemma 18.10 and Theorem 18.15, we deduce the following.

Corollary 18.24. *If $C \subset \mathbb{R}^N$ is regular at x and closed near x, then both $T_C(x)$ and $N_C(x)$ are convex. Furthermore,*

(i) $N_C(x) = T_C(x)^\circ$;

(ii) $T_C(x) = N_C(x)^\circ$.

In infinite dimensions, our main equivalent characterization of normal regularity is the following. (We do not have a similar characterization of tangential regularity.)

Theorem 18.25. *Let X be a reflexive and Gâteaux smooth Banach space. Then $C \subset X$ is normally regular at $x \in C$ if and only if $\widehat{T}_C(x) = \widehat{N}_C(x)^\circ$.*

Proof. Suppose first that $\widehat{T}_C(x) = \widehat{N}_C(x)^\circ$. Since $\widehat{T}_C(x) \subset N_C(x)^\circ$ by Lemma 18.12, we have $\widehat{N}_C(x)^\circ \subset N_C(x)^\circ$. Furthermore, Theorem 18.5 (ii) yields $\widehat{N}_C(x) \subset N_C(x)$ and thus $\widehat{N}_C(x)^\circ \supset N_C(x)^\circ$ by Theorem 1.8. It follows that $\widehat{N}_C(x)^\circ = N_C(x)^\circ$. We now recall from Theorem 18.8 that $\widehat{N}_C(x)$ is closed and convex. Hence $\bar{x}^* \in N_C(x) \setminus \widehat{N}_C(x)$ implies by Theorem 1.13 that there exists $\bar{x} \in X$ and $\lambda \in \mathbb{R}$ such that

$$\langle x^*, \bar{x} \rangle_X \leq \lambda < \langle \bar{x}^*, \bar{x} \rangle_X \qquad (x^* \in \widehat{N}_C(x)).$$

Since $\widehat{N}_C(x)$ is a cone, this is only possible for $\lambda \geq 0$. Thus the first inequality shows that $\bar{x} \in \widehat{N}_C(x)^\circ$ and the second that $\bar{x} \notin N_C(x)^\circ$. This is in contradiction to $\widehat{N}_C(x)^\circ = N_C(x)^\circ$. Hence $N_C(x) = \widehat{N}_C(x)$, i.e., C is normally regular at x.

Conversely, if C is normally regular at x, we obtain using Lemma 18.12 that

$$\widehat{T}_C(x) \subset N_C(x)^\circ = \widehat{N}_C(x)^\circ.$$

By Lemma 18.10 (i), Theorem 18.5 (i), and Theorem 1.8 using the fact that $\widehat{T}_C(x)$ is a closed convex cone by Theorem 18.8, we also have

$$\widehat{N}_C(x)^\circ \supset T_C(x)^{\circ\circ} \supset \widehat{T}_C(x)^{\circ\circ} = \widehat{T}_C(x).$$

Therefore $\widehat{T}_C(x) = \widehat{N}_C(x)^\circ$ as claimed. □

In sufficiently regular spaces, normal regularity implies tangential regularity of closed sets.

Lemma 18.26. *Let X be a reflexive and Gâteaux smooth Banach space and let $C \subset X$ be closed near $x \in C$. If C is normally regular at x, then C is tangentially regular at x.*

Proof. Arguing as in the proof of Theorem 18.23 (i) ⇔ (ii), by Theorems 1.8, 18.5, and 18.19 and Lemma 18.10 we have

$$T_C(x) \subset T_C^w(x) \subset T_C^w(x)^{\circ\circ} = \widehat{N}_C(x)^\circ = N_C(x)^\circ = \widehat{T}_C(x) \subset T_C(x).$$

This shows that $T_C(x) = \widehat{T}_C(x)$. □

From Lemmas 18.6 and 18.26, we immediately obtain the following regularity result.

Corollary 18.27. *Let X be a Gâteaux smooth Banach space and let $C \subset X$ be nonempty, closed, and convex. Then C is normally regular at every $x \in C$. If X is additionally reflexive, then C is also tangentially regular at every $x \in C$.*

19 TANGENT AND NORMAL CONES OF POINTWISE-DEFINED SETS

As we have seen in Chapter 18, the relationships between the different tangent and normal cones are less complete in infinite-dimensional spaces than in finite-dimensional ones. In this chapter, however, we show that certain *pointwise-defined* sets on $L^p(\Omega)$ for $p \in (1, \infty)$ largely satisfy the finite-dimensional relations. We will use these results in Chapter 21 to derive expressions for generalized derivatives of pointwise-defined set-valued mappings, in particular for subdifferentials of integral functionals. As mentioned in Section 18.4, these relations are less satisfying for the limiting cones than for the basic cones. To treat the limiting cones, we will therefore assume the regularity of the underlying pointwise sets. For the basic cones, we also require an assumption, which however is weaker than (tangential) regularity.

19.1 DERIVABILITY

We start with the fundamental regularity assumption. Let X be a Banach space and $C \subset X$. We then say that a tangent vector $\Delta x \in T_C(x)$ at $x \in C$ is *derivable* if there exists an $\varepsilon > 0$ and a curve $\xi : [0, \varepsilon] \to C$ that generates Δx at 0, i.e.,

$$\xi(0) = x \quad \text{and} \quad \Delta x = \lim_{\tau \searrow 0} \frac{\xi(\tau) - \xi(0)}{\tau} = \xi'(0). \tag{19.1}$$

Note that we do not make any assumptions on the differentiability or continuity of ξ except at $\tau = 0$. We say that C is *geometrically derivable* at $x \in C$ if every $\Delta x \in T_C(x)$ is derivable.

As the next lemma shows, the point of this definition is that derivable tangent vectors are characterized by a full limit instead of just an inner limit; this additional property will allow us to construct tangent vectors in $L^p(\Omega)$ from pointwise tangent vectors, similarly to how Clarke regularity was used to obtain equality in the pointwise characterization of Clarke subdifferentials of integral functionals in Theorem 13.9.

Lemma 19.1. *Let $C \subset X$ and $x \in C$. Then the set $T_C^0(x)$ of derivable tangent vectors is given by*

$$T_C^0(x) = \liminf_{\tau \searrow 0} \frac{C - x}{\tau}. \tag{19.2}$$

Proof. We first recall that by definition of the inner limit, Δx is an element of the set on the right-hand side if for every sequence $\tau_k \searrow 0$ there exist $x_k \in C$ such that $(x_k - x)/\tau_k \to \Delta x$. For a derivable tangent vector $\Delta x \in T_C^0(x)$ and any $\tau_k \searrow 0$, we can simply take $x_k = \xi(\tau_k)$. For the converse inclusion, let Δx be an element of the right-hand side set. Let now $\tau_k \searrow 0$ be given and take $x_k \in C$ realizing the inner limit. Since $\tau_k \searrow 0$ was arbitrary, setting $\xi(\tau_k) := x_k$ for all $k \in \mathbb{N}$ defines a curve $\xi : [0, \varepsilon] \to C$ for some $\varepsilon > 0$, and hence $\Delta x \in T_C^0(x)$. □

By taking $\tilde{x} \equiv x$ constant in (18.10) and comparing with (19.2), we immediately obtain that all Clarke tangent vectors are derivable.

Corollary 19.2. *Let $C \subset X$ and $x \in C$. Then every $\Delta x \in \widehat{T}_C(x)$ is derivable.*

Clearly, if C is tangentially regular at x, then also every tangent vector is derivable.

Corollary 19.3. *If $C \subset X$ is tangentially regular at $x \in C$, then every $\Delta x \in T_C(x)$ is derivable.*

However, a set can be geometrically derivable without being tangentially regular.

Example 19.4. Let $C := ([0, \infty) \times \{0\}) \cup (\{0\} \times [0, \infty)) \subset \mathbb{R}^2$. Then we obtain directly from the definition of the tangent cone that

$$T_C(x_1, x_2) = \begin{cases} C, & \text{if } (x_1, x_2) = (0, 0), \\ \mathbb{R} \times \{0\}, & \text{if } x_1 = 0, x_2 > 0, \\ \{0\} \times \mathbb{R}, & \text{if } x_1 > 0, x_2 = 0, \\ \emptyset, & \text{otherwise.} \end{cases}$$

However, it follows from Corollary 18.20 that $\widehat{T}_C(0, 0) = \{(0, 0)\}$. Thus C is not tangentially regular at $(0, 0)$.

On the other hand, for any $\Delta x = (t_1, 0) \in T_C(0, 0)$, $t_1 \in \mathbb{R}$, setting $\xi(s) := (st_1, 0)$ yields $\xi(0) = (0, 0)$ and $\xi'(0) = (t_1, 0) = \Delta x$. Hence Δx is derivable. Similarly, setting $\xi(s) := (0, st_2)$ shows that $\Delta x = (0, t_2) \in T_C(0, 0)$ is derivable for every $t_2 \in \mathbb{R}$. Thus C is geometrically derivable at $(0, 0)$.

19.2 TANGENT AND NORMAL CONES

As the goal is to define derivatives of set-valued mappings $F : X \rightrightarrows Y$ via tangent cones to their epigraphs $\operatorname{epi} F \subset X \times Y$, we need to consider product spaces of p-integrable functions (with possibly different p). Let therefore $\Omega \subset \mathbb{R}^d$ be an open and bounded domain. For $\vec{p} := (p_1, \dots, p_m) \in (1, \infty)^m$, we then define

$$L^{\vec{p}}(\Omega) := L^{p_1}(\Omega) \times \cdots \times L^{p_m}(\Omega),$$

endowed with the canonical euclidean product norm, i.e.,

$$\|u\|_{L^{\vec{p}}} := \sqrt{\sum_{k=1}^{m} \|u_k\|_{L^{p_k}}^2} \quad (u = (u_1, \dots, u_m) \in L^{\vec{p}}).$$

We will need the case $m = 2$ in Chapter 21; on first reading of the present chapter, we recommend picturing $m = 1$, i.e., $L^{\vec{p}}(\Omega) = L^p(\Omega)$ for some $p \in (1, \infty)$. We further denote by p^* the conjugate exponent of $p \in (1, \infty)$, defined as satisfying $1/p + 1/p^* = 1$, and write $\vec{p}^* := (p_1^*, \dots, p_m^*)$ so that $L^{\vec{p}}(\Omega)^* \cong L^{\vec{p}^*}(\Omega)$. Note that $L^{\vec{p}}(\Omega)$ is reflexive and Gâteaux smooth as the product of reflexive and Gâteaux smooth spaces; cf. Example 17.6. Finally, we will write $\mathcal{L}(\Omega)$ for the d-dimensional Lebesgue measure of Ω and recall the characteristic function $\mathbb{1}_U$ of a set $U \subset L^{\vec{p}}(\Omega)$, which satisfies $\mathbb{1}_U(u) = (1, \dots, 1) \in \mathbb{R}^m$ if $u \in U$ and $\mathbb{1}_U(u) = 0 \in \mathbb{R}^m$ otherwise.

We then call a set $U \subset L^{\vec{p}}(\Omega)$ for $\vec{p} \in (1, \infty)^m$ *pointwise defined* if

$$U := \left\{ u \in L^{\vec{p}}(\Omega) \mid u(x) \in C(x) \text{ for a.e. } x \in \Omega \right\}$$

for a Borel-measurable mapping $C : \Omega \rightrightarrows \mathbb{R}^m$ with $C(x) \subset \mathbb{R}^m$. We say that U is *pointwise derivable* if $C(x)$ is geometrically derivable at every $\xi \in C(x)$ for almost every $x \in \Omega$.

THE FUNDAMENTAL CONES

We now derive pointwise characterizations of the fundamental cones to pointwise defined sets, starting with the tangent cone.

Theorem 19.5. *Let $U \subset L^{\vec{p}}(\Omega)$ be pointwise derivable. Then for every $u \in U$,*

$$T_U(u) = \left\{ \Delta u \in L^{\vec{p}}(\Omega) \;\middle|\; \Delta u(x) \in T_{C(x)}(u(x)) \text{ for a.e. } x \in \Omega \right\}. \tag{19.3}$$

Proof. The inclusion "$\subset$" follows from (18.1) and the fact that a sequence convergent in $L^{\vec{p}}(\Omega)$ for $\vec{p} \in (1, \infty)$ converges, after possibly passing to a subsequence, pointwise almost everywhere.

For the converse inclusion, we take for almost every $x \in \Omega$ a tangent vector $\Delta u(x) \in T_{C(x)}(u(x))$ at $u(x) \in C(x)$. We only need to consider the case $\Delta u \in L^{\vec{p}}(\Omega)$. By geometric derivability, we may find for almost every $x \in \Omega$ an $\varepsilon(x) > 0$ and a curve $\xi(\cdot, x) : [0, \varepsilon(x)] \to C(x)$ such that $\xi(0, x) = u(x)$ and $\xi'_+(0, x) = \Delta u(x)$. In particular, for any given $\rho > 0$, we may find $\varepsilon_\rho(x) \in (0, \varepsilon(x)]$ such that

$$\frac{|\xi(t, x) - \xi(0, x) - \Delta u(x)t|_2}{t} \le \rho \qquad (t \in (0, \varepsilon_\rho(x)], \text{ a.e. } x \in \Omega). \tag{19.4}$$

For $t > 0$, let us set

$$E_{\rho,t} := \{x \in \Omega \mid t \le \varepsilon_\rho(x)\}$$

and define

$$\tilde{u}^{\rho,t}(x) := \begin{cases} \xi(t, x) & \text{if } x \in E_{\rho,t}, \\ u(x) & \text{if } x \in \Omega \setminus E_{\rho,t}. \end{cases}$$

Writing $\xi = (\xi_1, \dots, \xi_m)$ and $\Delta u = (\Delta u_1, \dots, \Delta u_m)$, we have from (19.4) that

$$\frac{|\xi_j(t, x) - \xi_j(0, x) - \Delta u_j(x)t|}{t} \le \rho \quad (j = 1, \dots, m,\ t \in (0, \varepsilon_\rho(x)] \text{ for a.e. } x \in \Omega). \tag{19.5}$$

Therefore, using the elementary inequality $(a + b)^2 \le 2a^2 + 2b^2$, we obtain

$$\begin{aligned} \|\tilde{u}^{\rho,t} - u\|_{L^{\vec{p}}}^2 &= \sum_{j=1}^m \|[\tilde{u}_j^{\rho,t} - u]_j\|_{L^{p_j}}^2 \\ &\le \sum_{j=1}^m \left(\int_\Omega t^{p_j} (\rho + |\Delta u_j(x)|)^{p_j} \, dx \right)^{2/p_j} \\ &\le \sum_{j=1}^m \left(t\rho \mathcal{L}(\Omega)^{1/p_j} + t\|\Delta u_j\|_{L^{p_j}} \right)^2 \\ &\le 2t^2 \sum_{j=1}^m \left(\rho \mathcal{L}(\Omega)^{1/p_j} \right)^2 + 2t^2 \|\Delta u\|_{L^{\vec{p}}}^2. \end{aligned} \tag{19.6}$$

Similarly, (19.5) and the same elementary inequality together with Minkowski's inequality in the form $(a^p + b^p)^{1/p} \le |a| + |b|$ yield

$$\begin{aligned} \frac{\|\tilde{u}^{\rho,t} - u - t\Delta u\|_{L^{\vec{p}}}^2}{t^2} &= \sum_{j=1}^m \frac{1}{t^2} \left(\int_{E_{\rho,t}} |\xi_j(t, x) - \xi_j(0, x) - t\Delta u_j(x)|^{p_j} \, dx \right. \\ &\qquad \left. + \int_{\Omega \setminus E_{\rho,t}} |\Delta u_j(x) t|^{p_j} \, dx \right)^{2/p_j} \\ &\le \sum_{j=1}^m \left(\rho^{p_j} \mathcal{L}(\Omega) + \|\Delta u \mathbb{1}_{\Omega \setminus E_{\rho,t}}\|_{L^{\vec{p}}}^{p_j} \right)^{2/p_j} \\ &\le 2 \sum_{j=1}^m \left(\rho \mathcal{L}(\Omega)^{1/p_j} \right)^2 + 2\|\Delta u \mathbb{1}_{\Omega \setminus E_{\rho,t}}\|_{L^{\vec{p}}}^2. \end{aligned} \tag{19.7}$$

Now for each $k \in \mathbb{N}$, we can find $t_k \searrow 0$ such that $\|\Delta u \mathbb{1}_{\Omega \setminus E_{1/k,t_k}}\|_{L^{\vec{p}}} \le 1/k$. This follows from Lebesgue's dominated convergence theorem and the fact that $\mathcal{L}(\Omega \setminus E_{\rho,t}) \to 0$ as $t \to 0$. The estimates (19.6) and (19.7) with $\rho = 1/k$ and $t = t_k$ thus show for $u_k := \tilde{u}^{1/k,t_k}$ that $u_k \to u$ and $(u_k - u)/t_k \to \Delta u$, i.e., $\Delta u \in T_U(u)$. □

We next consider the Fréchet normal cone.

Theorem 19.6. *Let $U \subset L^{\vec{p}}(\Omega)$ be pointwise derivable. Then for every $u \in U$,*

$$\widehat{N}_U(u) = \left\{u^* \in L^{\vec{p}^*}(\Omega) \;\middle|\; u^*(x) \in \widehat{N}_{C(x)}(u(x)) \text{ for a.e. } x \in \Omega\right\}. \tag{19.8}$$

Proof. Recalling the definition of $\widehat{N}_U(u)$ from (18.7), we need to find all $u^* \in L^{\vec{p}^*}(\Omega)$ satisfying for every given sequence $U \ni u_k \to u$

$$0 \ge \limsup_{k\to\infty} \frac{\langle u^*, u_k - u\rangle_{L^{\vec{p}}}}{\|u_k - u\|_{L^{\vec{p}}}} =: \limsup_{k\to\infty} L_k. \tag{19.9}$$

Let $\varepsilon > 0$ be arbitrary and set $v_k := u - u_k$ as well as

$$Z_k^1 := \{x \in \Omega \mid |v_k(x)|_2 \le \varepsilon^{-1}\|v_k\|_{L^{\vec{p}}}\} \qquad (k \in \mathbb{N}). \tag{19.10a}$$

Furthermore, let $Z^2 \subset \Omega$ be such that

$$u^* \text{ is bounded on } Z^2, \tag{19.10b}$$

$$\mathcal{L}(Z_k^1 \setminus Z^2) \le \varepsilon \qquad (k \in \mathbb{N}). \tag{19.10c}$$

Using Hölder's inequality, (19.10a), and (19.10c), we then estimate for $k = 1, \dots, m$

$$\begin{aligned} L_k &= \frac{\int_{\Omega\setminus(Z_k^1\cap Z^2)} \langle u^*(x), v_k(x)\rangle_2 \, dx}{\|v_k\|_{L^{\vec{p}}}} + \frac{\int_{Z_k^1\cap Z^2} \langle u^*(x), v_k(x)\rangle_2 \, dx}{\|v_k\|_{L^{\vec{p}}}} \\ &\le \frac{\|\mathbb{1}_{\Omega\setminus(Z_k^1\cap Z^2)} u^*\|_{L^{\vec{p}^*}} \|v_k\|_{L^{\vec{p}}}}{\|v_k\|_{L^{\vec{p}}}} + \int_{Z_k^1\cap Z^2} \frac{\langle u^*(x), v_k(x)\rangle_2}{|v_k(x)|_2} \cdot \frac{|v_k(x)|_2}{\|v_k\|_{L^{\vec{p}}}} \, dx \\ &\le \|\mathbb{1}_{\Omega\setminus(Z_k^1\cap Z^2)} u^*\|_{L^{\vec{p}^*}} + \varepsilon^{-1} \int_{Z^2} \max\left\{0, \frac{\langle u^*(x), v_k(x)\rangle_2}{|v_k(x)|_2}\right\} dx. \end{aligned}$$

If now for almost every $x \in \Omega$ we have that $u^*(x) \in \widehat{N}_{C(x)}(u(x))$, then also $\langle u^*(x), v_k(x)\rangle_2 \le 0$ for almost every $x \in \Omega$. It follows using (19.10b) and the reverse Fatou inequality in the previous estimate that

$$\limsup_{k\to\infty} L_k \le \limsup_{k\to\infty} \|\mathbb{1}_{\Omega\setminus(Z_k^1\cap Z^2)} u^*\|_{L^{\vec{p}^*}}. \tag{19.11}$$

Since $|v_k(x)|_2 \ge \varepsilon^{-1}\|v_k\|_{L^{\vec{p}}}$ for $x \in \Omega \setminus Z_k^q$, we have

$$\|v_k\|_{L^{\vec{p}}} \ge \|\mathbb{1}_{\Omega\setminus Z_k^1} v_k\|_{L^{\vec{p}}} \ge (\varepsilon^{-p}\mathcal{L}(\Omega \setminus Z_k^1))^{1/p}\|v_k\|_{L^{\vec{p}}}.$$

Hence $\mathcal{L}(\Omega \setminus Z_k^1) \le \varepsilon^p$ and $\mathcal{L}(\Omega \setminus (Z_k^1 \cap Z_2)) \le \mathcal{L}(\Omega \setminus Z_k^1) + \mathcal{L}(\Omega \setminus Z_2) \le C\varepsilon$ for some constant $C > 0$ and small enough $\varepsilon > 0$. It therefore follows from Egorov's theorem that $\mathbb{1}_{\Omega\setminus(Z_k^1\cap Z^2)}u^*$ converge to 0 in measure as $k \to \infty$. Since $u^* \in L^{\vec{p}^*}(\Omega)$ and $\mathbb{1}_{\Omega\setminus(Z_k^1\cap Z^2)}u^* \le u^*$, it follows from Vitali's convergence theorem (see, e.g., [Fonseca and Leoni, 2007, Proposition 2.27]) that $\limsup_{k\to\infty} \|\mathbb{1}_{\Omega\setminus(Z_k^1\cap Z^2)}u^*\|_{L^{\vec{p}^*}} = 0$. Since $\varepsilon > 0$ was arbitrary, we deduce from (19.11) that (19.9) holds and, consequently,

$$\widehat{N}_U(u) \supset \{u^* \in L^{\vec{p}^*}(\Omega) \mid u^*(x) \in \widehat{N}_{C(x)}(u(x)) \text{ for a.e. } x \in \Omega\}.$$

This proves one direction of (19.8), which therefore holds even without geometric derivability.

For the converse inclusion, let $u^* \in \widehat{N}_U(u)$. We have to show that $u^*(x) \in \widehat{N}_{C(x)}(u(x))$ for almost every $x \in \Omega$, which we do by contradiction. Assume therefore that the pointwise inclusion does not hold. By the polarity relationship $\widehat{N}_{C(x)}(u(x)) = T_{C(x)}(u(x))^\circ$ from Lemma 18.10, we can find $\delta > 0$ and a Borel set $E \subset \Omega$ of finite positive Lebesgue measure such that for each $x \in E$, there exists $w(x) \in T_{C(x)}(u(x))$ with $|w(x)|_2 = 1$ and $\langle u^*(x), w(x)\rangle_2 \ge \delta$. We may without loss of generality assume that $C(x)$ is geometrically derivable at $w(x)$ for every $x \in E$, i.e., for each $x \in E$ there exists a curve $\xi(\cdot, x) : [0, \varepsilon(x)] \to C(x)$ such that $\xi'_+(0, x) = w(x)$ and $\xi(0, x) = u(x)$. Let now $c \in (0, \delta)$ be arbitrary. By replacing E by a subset of positive measure, we may by Egorov's theorem assume the existence of $\varepsilon > 0$ such that

$$|\xi(t, x) - \xi(0, x) - w(x)t|_2 \le ct \quad (t \in [0, \varepsilon],\ x \in E). \tag{19.12}$$

Let us define

$$\tilde{u}^t(x) := \begin{cases} \xi(t, x) & \text{if } x \in E, \\ u(x) & \text{if } x \in \Omega \setminus E. \end{cases}$$

Setting $v^t := \tilde{u}^t - u$, we have $v^t(x) = \xi(t, x) - \xi(0, x)$ for $x \in E$ and $v^t(x) = 0$ for $x \in \Omega \setminus E$. Therefore, writing $v^t = (v_1^t, \dots, v_m^t)$, $w = (w_1, \dots, w_m)$, and $\xi = (\xi_1, \dots \xi_m)$, we obtain using (19.12) for $t \in (0, \varepsilon]$ and some $c' > 0$ that

$$\begin{aligned} \|v^t\|_{L^{\vec{p}}}^2 &= \sum_{j=1}^m \left(\int_E |\xi_j(t, x) - \xi_j(0, x)|^{p_j}\, dx \right)^{2/p_j} \\ &\le \sum_{j=1}^m \left(\int_E (|w_j(x)|t + ct)^{p_j}\, dx \right)^{2/p_j} \le c't^2. \end{aligned}$$

Likewise,

$$\langle u^*(x), v^t(x)\rangle_2 \ge \langle u^*(x), w(x)\rangle_2 - |u^*(x)|_2 \cdot |\xi(t, x) - \xi(0, x) - wt|_2 \ge \delta t - ct.$$

It follows that

$$\limsup_{t\searrow 0} \int_E \frac{\langle u^*(x), v^t(x)\rangle_2}{\|v^t\|_{L^{\vec{p}}}}\, dx \ge \limsup_{t\searrow 0} \frac{\mathcal{L}(E)(\delta t - ct)}{c't} = \frac{\mathcal{L}(E)(\delta - c)}{c'} > 0.$$

Taking $u_k := \tilde{u}^{1/k}$ for $k \in \mathbb{N}$, we obtain $\lim_{k\to\infty} L_k > 0$ and therefore $u^* \notin \widehat{N}_U(u)$. By contraposition, this shows that $u^*(x) \in \widehat{N}_{C(x)}(u(x))$ for almost every $x \in \Omega$. □

We can now derive a similar polarity relationships as to the finite-dimensional one in Lemma 18.10.

Corollary 19.7. *Let $U \subset L^{\vec{p}}(\Omega)$ be pointwise derivable and $u \in U$. Then $\widehat{N}_U(u) = T_U(u)^\circ$.*

Proof. By Theorems 19.5 and 19.6 and Lemma 18.10, we have

$$\begin{aligned}
u^* \in \widehat{N}_U(u) &\Leftrightarrow u^*(x) \in \widehat{N}_{C(x)}(u(x)) && \text{(a.e. } x \in \Omega\text{)} \\
&\Leftrightarrow \langle u^*(x), \Delta u(x)\rangle_2 \le 0 && \text{(a.e. } x \in \Omega \text{ when } \Delta u(x) \in T_{C(x)}(u(x))\text{)} \\
&\Rightarrow \langle u^*, \Delta u\rangle_{L^{\vec{p}}} \le 0 && \text{(when } \Delta u \in T_U(u)\text{)} \\
&\Leftrightarrow u^* \in T_U(u)^\circ.
\end{aligned} \tag{19.13}$$

Hence $\widehat{N}_U(u) \subset T_U(u)^\circ$.

For the converse inclusion, we need to improve the implication in (19.13) to an equivalence. We argue by contradiction. Assume that $u^* \in T_U(u)^\circ$ and that there exists some $\Delta\bar{u} \in T_U(u)$ and a subset $E \subset \Omega$ with $\mathcal{L}(\Omega \setminus E) > 0$ and

$$\langle u^*(x), \Delta\bar{u}(x)\rangle_2 > 0 \qquad (x \in E).$$

Taking $\bar{u}^*(x) := (1 + t\mathbb{1}_E(x))u^*(x)$, we obtain for sufficient large t that $\langle \bar{u}^*, \Delta\bar{u}\rangle_{L^{\vec{p}}} > 0$. This contradicts that $u^* \in T_U(u)^\circ$. Hence $\widehat{N}_U(u) \supset T_U(u)^\circ$. □

THE LIMITING CONES

For the limiting cones, we in general only have an inclusion of the pointwise cones.

Theorem 19.8. *Let $U \subset L^{\vec{p}}(\Omega)$ be pointwise derivable. Then for every $u \in U$,*

$$\widehat{T}_U(u) \supset \left\{\Delta u \in L^{\vec{p}}(\Omega) \;\middle|\; \Delta u(x) \in \widehat{T}_{C(x)}(u(x)) \text{ for a.e. } x \in \Omega\right\}.$$

Proof. Let $\Delta u \in L^{\vec{p}}\Omega)$ with $\Delta u(x) \in \widehat{T}_{C(x)}(u(x))$ for almost every $x \in \Omega$ and let $u_k \to u$ in $L^{\vec{p}}(\Omega)$. In particular, we then have $u_k(x) \to u(x)$ for almost every $x \in \Omega$. Furthermore, by the inner limit characterization of $\widehat{T}_{C(x)}(u(x))$ in Corollary 18.20, there exist $\Delta\tilde{u}_k(x) \in T_{C(x)}(u_k(x))$ with $\Delta\tilde{u}_k(x) \to \Delta u(x)$. Egorov's theorem, then yields for all $\ell \ge 1$ a Borel-measurable set $E_\ell \subset \Omega$ such that $\mathcal{L}(\Omega \setminus E_\ell) < 1/\ell$ and $\Delta\tilde{u}_k \to \Delta u$ uniformly on E_ℓ. Since $T_{C(x)}(u_k(x))$ is a cone, we have $0 \in T_{C(x)}(u_k(x))$. It follows that

$$T_{C(x)}(u_k(x)) \ni \Delta u_{\ell,k}(x) := \mathbb{1}_{E_\ell}(x)\Delta\tilde{u}_k(x).$$

In particular, (19.3) shows that $\Delta u_{\ell,k} \in T_U(u_k)$ with $\Delta u_{\ell,k} \to \Delta u_\ell := \Delta u \mathbb{1}_{E_\ell}$ in $L^{\vec{p}}(\Omega)$ as $k \to \infty$. By Vitali's convergence theorem (compare the proof of Theorem 19.6), $\Delta u \mathbb{1}_{E_\ell} \to \Delta u$ in $L^{\vec{p}}(\Omega)$ as $\ell \to \infty$. Therefore, we may extract a diagonal subsequence $\{\Delta \tilde{u}_k := \Delta u_{\ell_k,k}\}_{k\geq 1}$ of $\{\Delta u_{\ell,k}\}_{k,\ell\geq 1}$ such that $\Delta \tilde{u}_k \to \Delta u$. Since $u_k \to u$ was arbitrary and $\Delta \tilde{u}_k \in T_U(u_k)$, we deduce that $\Delta u \in \widehat{T}_U(u)$. □

Theorem 19.9. *Let $U \subset L^{\vec{p}}(\Omega)$ be pointwise derivable. Then for every $u \in U$,*

$$N_U(u) \supset \big\{u^* \in L^{\vec{p}^*}(\Omega) \;\big|\; u^*(x) \in N_{C(x)}(u(x)) \text{ for a.e. } x \in \Omega\big\}.$$

Proof. Let $u^* \in L^{\vec{p}^*}(\Omega)$ with $u^*(x) \in N_{C(x)}(u(x))$ for almost every $x \in \Omega$. Then by definition, for almost all $x \in \Omega$ there exist $C(x) \ni \tilde{u}_k(x) \to u(x)$ as well as $\widehat{N}_{C(x)}(u(x)) \ni \tilde{u}_k^*(x) \to u^*(x)$. By Egorov's theorem, for every $\ell \geq 1$ there exists a Borel-measurable set $E_\ell \subset \Omega$ such that $\mathcal{L}(\Omega \setminus E_\ell) < 1/\ell$ and $\tilde{u}_k^* \to u^*$ as well as $\tilde{u}_k \to u$ uniformly on E_ℓ. We set $u_{\ell,k} := \mathbb{1}_{E_\ell} u_k + (1 - \mathbb{1}_{E_\ell})u$ and $u_{\ell,k}^* := \mathbb{1}_{E_\delta} \tilde{u}_k^*$. Then $u_{\ell,k}^*(x) \in \widehat{N}_{C(x)}(u_{\ell,k}(x))$ for almost every $x \in \Omega$. By Vitali's convergence theorem (compare the proof of Theorem 19.6), both $u_{\ell,k} \to u$ in $L^{\vec{p}}(\Omega)$ and $u_{\ell,k}^* \to u_\ell^*$ in $L^{\vec{p}^*}(\Omega)$ for $u_\ell^* := \mathbb{1}_{E_\ell} u^*$. Since $u_\ell^* \to u^*$ in $L^{\vec{p}^*}(\Omega)$, we can extract a diagonal subsequence of $\{(u_{\ell,k}, u_{\ell,k}^*)\}_{\ell,k\geq 1}$ to deduce that $u^* \in N_U(u)$. □

If the pointwise sets $C(x)$ are regular, we have the following polarity between the cones to the pointwise-defined set U.

Lemma 19.10. *Let $U \subset L^{\vec{p}}(\Omega)$ be pointwise derivable and $u \in U$. If $C(x)$ is regular at $u(x)$ and closed near $u(x)$ for almost every $x \in \Omega$, then $T_U(u) = \widehat{N}_U(u)^\circ$.*

Proof. By the regularity of $C(x)$ at $u(x)$ for almost every $x \in \Omega$ and Theorem 19.6, we have

$$\widehat{N}_U(u) = \big\{u^* \in L^{\vec{p}^*}(\Omega) \;\big|\; u^*(x) \in N_{C(x)}(u(x)) \text{ for a.e. } x \in \Omega\big\}.$$

By Theorem 18.15, $N_{C(x)}(u(x))^\circ = \widehat{T}_{C(x)}(u(x))$ for almost every $x \in \Omega$. Arguing as in the proof of Corollary 19.7, we thus obtain

$$\widehat{N}_U(u)^\circ = \big\{\Delta u \in L^{\vec{p}}(\Omega) \;\big|\; \Delta u(x) \in \widehat{T}_{C(x)}(u(x)) \text{ for a.e. } x \in \Omega\big\}.$$

The regularity of $C(x)$ also implies that $\widehat{T}_{C(x)}(u(x)) = T_{C(x)}(u(x))$ for almost every $x \in \Omega$. The claims now follow from Theorem 19.5. □

We can use this result to transfer the regularity of $C(x)$ to U.

Lemma 19.11. *Let $U \subset L^{\vec{p}}(\Omega)$ be pointwise derivable and $u \in U$. If $C(x)$ is regular at $u(x)$ and closed near $u(x)$ for almost every $x \in \Omega$, then U is regular at u and*

$$T_U^w(u) = T_U(u) = \widehat{T}_U(u).$$

Proof. Since $L^{\vec{p}}(\Omega)$ is reflexive, we have $\widehat{N}_U(u) = T_U^w(u)^\circ$ by Lemma 18.10 (ii). This fact together with Lemma 19.10 and Theorems 1.8 and 18.5 shows that

$$T_U^w(u) \subset T_U^w(u)^{\circ\circ} = \widehat{N}_U(u)^\circ = T_U(u) \subset T_U^w(u).$$

Furthermore, by the regularity and closedness assumptions, we obtain from Theorems 19.5 and 19.8 that $T_U(u) = \widehat{T}_U(u)$, which also implies tangential regularity.

Since $L^{\vec{p}}(\Omega)$ for $\vec{p} \in (1,\infty)^m$ is reflexive and Gâteaux smooth, normal regularity follows from Theorem 18.25 together with Lemma 19.10. □

From this, we obtain pointwise expressions with equality. For the Clarke tangent cone, we only require local closedness of the underlying sets.

Theorem 19.12. *Let $U \subset L^{\vec{p}}(\Omega)$ be pointwise derivable. If $C(x)$ is closed near $u(x)$ for almost every $x \in \Omega$ for every $u \in U$, then*

$$\widehat{T}_U(u) = \big\{\Delta u \in L^{\vec{p}}(\Omega) \;\big|\; \Delta u(x) \in \widehat{T}_{C(x)}(u(x)) \text{ for a.e. } x \in \Omega\big\}.$$

Proof. The inclusion "$\supset$" was already shown in Theorem 19.8. To prove the converse inclusion when $C(x)$ is closed near $u(x)$ for almost every $x \in \Omega$, we only need to observe from Lemma 18.12 and Theorem 19.9 and

$$\begin{aligned}\widehat{T}_C(u) \subset N_C(u)^\circ &\subset \big\{u^* \in L^{\vec{p}^*}(\Omega) \;\big|\; u^*(x) \in N_{C(x)}(u(x)) \text{ for a.e. } x \in \Omega\big\}^\circ \\ &= \big\{\Delta u \in L^{\vec{p}}(\Omega) \;\big|\; \Delta u(x) \in \widehat{T}_{C(x)}(u(x)) \text{ for a.e. } x \in \Omega\big\},\end{aligned}$$

where the last equality again follows from Theorem 18.15 together with an argument as in the proof of Corollary 19.7. □

For the limiting normal cone, however, we *do* require regularity.

Theorem 19.13. *Let $U \subset L^{\vec{p}}(\Omega)$ be pointwise derivable. If $C(x)$ is regular at $u(x)$ and closed near $u(x)$ for almost every $x \in \Omega$, then for every $u \in U$,*

$$N_U(u) = \big\{u^* \in L^{\vec{p}^*}(\Omega) \;\big|\; u^*(x) \in N_{C(x)}(u(x)) \text{ for a.e. } x \in \Omega\big\}.$$

Proof. The inclusion "$\supset$" was already shown in Theorem 19.9. The converse inclusion for regular and closed $C(x)$ follows from Lemma 19.11 and Theorem 19.6. □

Remark 19.14. Theorems 19.5 and 19.6 on the fundamental cones are based on [Clason and Valkonen, 2017b]. Without regularity, the characterization of the limiting normal cone of a pointwise-defined set is much more delicate. A full characterization was given in [Mehlitz and Wachsmuth, 2018, 2019], which showed that even for a closed nonconvex set, the limiting normal cone contains the convex hull of the strong limiting normal cone (where the limit is taken with respect to strong convergence instead of weak-$*$ convergence) and is dense in the Dini normal cone $\widehat{T}_C^\circ(x)$ – in the words of the authors, it may be "unpleasantly large". This is due to an inherent convexifying effect of integration with respect to the Lebesgue measure.

A characterization of specific pointwise-defined sets in Sobolev spaces was derived in [Harder and Wachsmuth, 2018], with similar conclusions.

20 DERIVATIVES AND CODERIVATIVES OF SET-VALUED MAPPINGS

We are now ready to differentiate set-valued mappings; as already discussed, these generalized derivatives are based on the tangent and normal cones of the previous Chapter 18. To account for the changed focus, we will slightly switch notation and use in this and the following chapters of Part IV uppercase letters for set-valued mappings and lowercase letters for scalar-valued functionals such that, e.g., $F(x) = \partial f(x)$. We focus in this chapter on examples, basic properties, and relationships between the various derivative concepts. In the following Chapters 22 to 25, we then develop calculus rules for each of the different derivatives and coderivatives.

20.1 DEFINITIONS

To motivate the following definitions, it is instructive to recall the geometric intuition behind the classical derivative of a scalar function f as limit of a difference quotient: given an (infinitesimal) change Δx of the argument x, it gives the corresponding (infinitesimal) change Δy of the value $y = f(x)$ required to stay on the graph of f. In other words, $(\Delta x, \Delta y)$ is a tangent vector to graph f. For a proper set-valued mapping F, however, it is also possible to remain on the graph of F by varying y without changing x; it thus also makes sense to ask the "dual" question of, given a change Δy in image space, what change Δx in domain space is required to stay inside the graph of F. In geometric terms, the answer is given by Δx such that $(\Delta x, -\Delta y)$ is a normal vector to graph F. (Note that normal vectors point *away* from a set, while we are trying to correct by moving *towards* it. Recall also that $(f'(x), -1)$ is normal to epi f for a smooth function f; see Figure 20.1 and compare Lemma 4.10 as well as Section 20.4 below.) In Banach spaces, of course, normal vectors are subsets of the dual space.

We thus distinguish

(i) *graphical derivatives*, which generalize classical derivatives and are based on tangent cones;

(ii) *coderivatives*, which generalize adjoint derivatives and are based on normal cones.

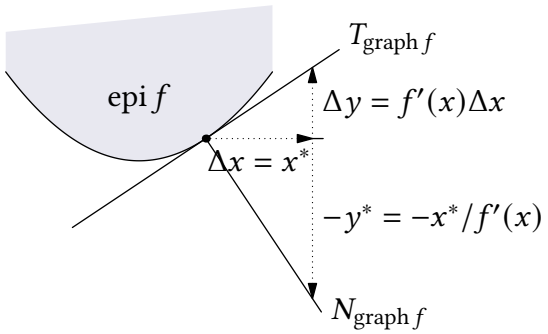


Figure 20.1: Illustration why the coderivatives negate y^* in comparison to the normal cone.

In each case, we can use either basic or limiting cones, leading to four different definitions.

Specifically, let X, Y be Banach spaces and $F : X \rightrightarrows Y$. Then we define

(i) the *graphical derivative* of F at $x \in X$ for $y \in Y$ as

$$DF(x|y) : X \rightrightarrows Y, \qquad DF(x|y)(\Delta x) := \left\{\Delta y \in Y \;\middle|\; (\Delta x, \Delta y) \in T_{\operatorname{graph} F}(x, y)\right\};$$

(ii) the *Clarke graphical derivative* of F at $x \in X$ for $y \in Y$ as

$$\widehat{D}F(x|y) : X \rightrightarrows Y, \qquad \widehat{D}F(x|y)(\Delta x) := \left\{\Delta y \in Y \;\middle|\; (\Delta x, \Delta y) \in \widehat{T}_{\operatorname{graph} F}(x, y)\right\};$$

(iii) the *Fréchet coderivative* of F at $x \in X$ for $y \in Y$ as

$$\widehat{D}^*F(x|y) : Y^* \rightrightarrows X^*, \qquad \widehat{D}^*F(x|y)(y^*) := \left\{x^* \in X^* \;\middle|\; (x^*, -y^*) \in \widehat{N}_{\operatorname{graph} F}(x, y)\right\};$$

(iv) the (*basic* or *limiting* or *Mordukhovich*) *coderivative* of F at $x \in X$ for $y \in Y$ as

$$D^*F(x|y) : Y^* \rightrightarrows X^*, \qquad D^*F(x|y)(y^*) := \left\{x^* \in X^* \;\middle|\; (x^*, -y^*) \in N_{\operatorname{graph} F}(x, y)\right\}.$$

Observe how the coderivatives operate from Y^* to X^*, while the derivatives operate from X to Y. It is crucial that these are defined directly via (possibly nonconvex) normal cones rather than via polarity from the corresponding graphical derivatives to avoid convexification. This will allow for sharper results involving these coderivatives.

We illustrate these definitions with the simplest example of a single-valued linear operator.

Example 20.1 (single-valued linear operators). Let $F(x) := \{Ax\}$ for $A \in \mathbb{L}(X; Y)$ and $u = (x, Ax) \in \operatorname{graph} F$. Note that $\operatorname{graph} F$ is a linear subspace of $X \times Y$. Since $\operatorname{graph} F$ is regular by Corollary 18.27, both of the tangent cones are given by

$$T_{\operatorname{graph} F}(u) = \widehat{T}_{\operatorname{graph} F}(u) = \operatorname{graph} F = \{(\Delta x, A\Delta x) \in X \times Y \mid \Delta x \in X\},$$

while the normal cones are given by

$$\begin{aligned} N_{\operatorname{graph} F}(u) = \widehat{N}_{\operatorname{graph} F}(u) &= \{u^* \in X^* \times Y^* \mid u^* \perp \operatorname{graph} F\} \\ &= \{(x^*, y^*) \in X^* \times Y^* \mid \langle x^*, \Delta x\rangle_X + \langle y^*, A\Delta x\rangle_Y = 0 \text{ for all } \Delta x \in X\} \\ &= \{(A^* y^*, -y^*) \in X^* \times Y^* \mid y^* \in Y^*\}. \end{aligned}$$

This immediately yields the graphical derivatives

$$DF(x|Ax)(\Delta x) = \widehat{D}F(x|Ax)(\Delta x) = \{A\Delta x\}$$

as well as the coderivatives

$$D^*F(x|y)(y^*) = \widehat{D}^*F(x|y)(y^*) = \{A^* y^*\}.$$

Using (18.1), we can also write the graphical derivative as

$$DF(x|y)(\Delta x) = \limsup_{t \searrow 0,\, \Delta\tilde{x} \to \Delta x} \frac{F(x + t\Delta\tilde{x}) - y}{t}, \tag{20.1}$$

since

$$(\Delta x, \Delta y) \in \limsup_{\tau \searrow 0} \frac{\operatorname{graph} F - (x, y)}{\tau}$$

if and only if there exist $\tau_k \searrow 0$ and x_k such that

$$\Delta x = \lim_{k\to\infty} \frac{x_k - x}{\tau_k} \quad \text{and} \quad \Delta y \in \limsup_{k\to\infty} \frac{F(x_k) - y}{\tau_k}. \tag{20.2}$$

The former forces $x_k = x - \tau_k \Delta x_k$ for $\Delta x_k \to \Delta x$, so the latter gives (20.1).

In infinite-dimensional spaces, we also have to distinguish the *weak graphical derivative* $D^w F(x|y)$ and the *ε-coderivative* $\widehat{D}^*_\varepsilon F(x|y)$, both constructed analogously from the weak tangent cone $T^w_{\operatorname{graph} F}(x, y)$ and the ε-normal cone $\widehat{N}^\varepsilon_{\operatorname{graph} F}(x, y)$, respectively. However, we will not be working directly with these and instead switch to the setting of the corresponding cones when they would be needed.

Remark 20.2 (a much too brief history of various (co)derivatives). As for the various tangent and normal cones, the (more recent) development of derivatives and coderivatives of set-valued mappings is convoluted, and we do not attempt to give a full account, instead referring to the commentaries to [Rockafellar and Wets, 1998, Chapter 8], [Mordukhovich, 2006, Chapter 1.4.12], and [Mordukhovich, 2018, Chapter 1].

The graphical derivative goes back to Aubin [Aubin, 1981], who also introduced the Clarke graphical derivative (under the name *circatangent derivative*) in [Aubin, 1984]. Coderivatives based on normal cones were mainly treated there for mappings whose graphs are convex, for which these cones can be defined as polars of the appropriate tangent cones. Graphical derivatives were further studied in

[Thibault, 1983]. In parallel, Mordukhovich introduced the (nonconvex) limiting coderivative via his limiting normal cone in [Morduhovič, 1980], again stressing the need for a genuinely nonconvex direct construction. The term *coderivative* was coined by Ioffe, who was the first to study these mappings systematically in [Ioffe, 1984].

20.2 BASIC PROPERTIES

We now translate various results of Chapter 18 on tangent and normal cones to the setting of graphical derivatives and coderivatives. From Theorem 18.5, we immediately obtain

Corollary 20.3. *For $F : X \rightrightarrows Y$, $x \in X$, and $y \in Y$, we have the inclusions*

(i) $\widehat{D}F(x|y)(\Delta x) \subset DF(x|y)(\Delta x) \subset D^w F(x|y)(\Delta x)$ *for all* $\Delta x \in X$;

(ii) $\widehat{D}^*F(x|y)(y^*) \subset D^*F(x|y)(y^*)$ *for all* $y^* \in Y^*$.

Similarly, we obtain from Theorem 18.8 the following outer semicontinuity and convexity properties.

Corollary 20.4. *For $F : X \rightrightarrows Y$, $x \in X$, and $y \in Y$,*

(i) $DF(x|y)$, $\widehat{D}F(x|y)$, *and* $\widehat{D}^*F(x|y)$ *are closed;*

(ii) if X and Y are finite-dimensional, then $D^*F(x|y)$ is closed;

(iii) $\widehat{D}F(x|y)$ *and* $\widehat{D}^*F(x|y)$ *are convex.*

Graphical derivatives and coderivatives behave completely symmetrically with respect to inversion of a set-valued mapping (which we recall is always possible in the sense of preimages).

Lemma 20.5. *Let $F : X \rightrightarrows Y$, $x \in X$, and $y \in Y$. Then*

$$\begin{aligned}
\Delta y \in DF(x|y)(\Delta x) &\quad\Leftrightarrow\quad \Delta x \in DF^{-1}(y|x)(\Delta y),\\
\Delta y \in \widehat{D}F(x|y)(\Delta x) &\quad\Leftrightarrow\quad \Delta x \in \widehat{D}F^{-1}(y|x)(\Delta y),\\
x^* \in \widehat{D}^*F(x|y)(y^*) &\quad\Leftrightarrow\quad -y^* \in \widehat{D}^*F^{-1}(y|x)(-x^*),\\
x^* \in D^*F(x|y)(y^*) &\quad\Leftrightarrow\quad -y^* \in D^*F^{-1}(y|x)(-x^*).
\end{aligned}$$

Proof. We have

$$\begin{aligned}
\Delta y \in DF(x|y)(\Delta x) &\Leftrightarrow (\Delta x, \Delta y) \in T_{\operatorname{graph} F}(x, y)\\
&\Leftrightarrow (\Delta y, \Delta x) \in T_{\operatorname{graph} F^{-1}}(y, x)\\
&\Leftrightarrow \Delta x \in DF^{-1}(y|x)(\Delta y).
\end{aligned}$$

The proof for the regular derivative and the coderivatives is completely analogous. □

ADJOINTS OF SET-VALUED MAPPINGS

From the various relations between normal and tangent cones, we obtain corresponding relations between these derivatives. To state these relationships, we need to introduce the upper and lower adjoints of set-valued mappings. Let $H : X \rightrightarrows Y$ be a set-valued mapping. Then the *upper adjoint* of H is defined as

$$H^{\circ+}(y^*) := \{x^* \mid \langle x^*, x\rangle_X \le \langle y^*, y\rangle_Y \text{ for all } y \in H(x),\ x \in X\},$$

and the *lower adjoint* of H as

$$H^{\circ-}(y^*) := \{x^* \mid \langle x^*, x\rangle_X \ge \langle y^*, y\rangle_Y \text{ for all } y \in H(x),\ x \in X\}.$$

As the next example shows, these notions generalize the definition of the adjoint of a linear operator.

Example 20.6 (upper and lower adjoints of linear mappings). Let $H(x) := \{Ax\}$ for $A \in \mathbb{L}(X;Y)$. Then

$$\begin{aligned} H^{\circ+}(y^*) &= \{x^* \in X^* \mid \langle x^*, x\rangle_X \le \langle y^*, y\rangle_Y \text{ for all } y = Ax,\ x \in X\} \\ &= \{x^* \in X^* \mid \langle x^*, x\rangle_X \le \langle y^*, Ax\rangle_Y \text{ for all } x \in X\} \\ &= \{x^* \in X^* \mid \langle x^* - A^*y^*, x\rangle_X \le 0 \text{ for all } x \in X\} \\ &= \{A^*y^*\}. \end{aligned}$$

Similarly, $H^{\circ-}(y^*) = \{A^*y^*\}$.

For solution mappings of linear equations, we have the following adjoints.

Example 20.7 (upper and lower adjoints of solution maps to linear equations). Let $H(x) := \{y \mid Ay = x\}$ for $A \in \mathbb{L}(X;Y)$. Then

$$H^{\circ+}(y^*) = \{x^* \mid \langle x^*, x\rangle \le \langle y^*, y\rangle \text{ for all } Ay = x,\ x \in X\}$$

If $y^* \notin \operatorname{ran} A^*$, then $\operatorname{ran} A^* \perp \ker A \neq \emptyset$, so for every $x^* \in X^*$ and $x \in X$ we can choose $y \in Y$ such that the above condition is not satisfied. Therefore $H^{\circ+}(y) = \emptyset$. Otherwise, if $y^* = A^*\tilde{x}^*$, we continue to calculate

$$H^{\circ+}(y^*) = \{x^* \in X^* \mid \langle x^*, x\rangle_X \le \langle \tilde{x}^*, x\rangle_X \text{ for all } x \in X\} = \{\tilde{x}^*\}.$$

Therefore

$$H^{\circ+}(y^*) = \{x^* \in X \mid A^*x^* = y^*\}.$$

A similar argument shows that $H^{\circ-}(y^*) = H^{\circ+}(y^*)$.

These examples and Example 20.1 suggest the adjoint relationships of the next corollary. Note that in infinite-dimensional spaces, we only have a relationship between the limiting derivatives, i.e., between the Clarke graphical derivative and the limiting coderivative.

Corollary 20.8. *Let X, Y be Banach spaces and $F : X \rightrightarrows Y$.*

(i) If X and Y are finite-dimensional, then

$$\widehat{D}^* F(x|y) = DF(x|y)^{\circ+}.$$

(ii) If X and Y are reflexive and Gâteaux smooth (in particular, if they are finite-dimensional), and graph F *is closed near (x, y), then*

$$\widehat{D}F(x|y) = D^* F(x|y)^{\circ-}.$$

Proof. (i): Identifying X^* with X and Y^* with Y in finite dimension, we have by definition that

$$DF(x|y)(\Delta x) = \left\{\Delta y \in Y \;\middle|\; (\Delta x, \Delta y) \in T_{\operatorname{graph} F}(x, y)\right\}$$

and

$$\widehat{D}^* F(x|y)(\Delta y) = \left\{\Delta x \in X \;\middle|\; (\Delta x, -\Delta y) \in \widehat{N}_{\operatorname{graph} F}(x, y)\right\}.$$

Using Lemma 18.10 (iii), we then see that

$$\begin{aligned} x^* \in DF(x|y)^{\circ+}(y^*) &\Leftrightarrow \langle x^*, \Delta x\rangle_X \le \langle y^*, \Delta y\rangle_Y \text{ for } \Delta y \in DF(x|y)(\Delta x) \\ &\Leftrightarrow \langle x^*, \Delta x\rangle_X + \langle -y^*, \Delta y\rangle_Y \le 0 \text{ for } (\Delta x, \Delta y) \in T_{\operatorname{graph} F}(x, y) \\ &\Leftrightarrow (x^*, -y^*) \in T_{\operatorname{graph} F}(x, y)^\circ = \widehat{N}_{\operatorname{graph} F}(x, y) \\ &\Leftrightarrow x^* \in \widehat{D}^* F(x|y)(y^*). \end{aligned}$$

This proves the claim.

(ii): We proceed analogously to (i) using Theorem 18.19 (or Theorem 18.15 if X and Y are finite-dimensional):

$$\begin{aligned} \Delta y \in D^* F(x|y)^{\circ-}(\Delta x) &\Leftrightarrow \langle y^*, \Delta y\rangle_Y \ge \langle x^*, \Delta x\rangle_X \text{ for } x^* \in D^* F(x|y)(y^*) \\ &\Leftrightarrow \langle x^*, \Delta x\rangle_X + \langle -y^*, \Delta y\rangle_Y \le 0 \text{ for } (x^*, -y^*) \in N_{\operatorname{graph} F}(x, y) \\ &\Leftrightarrow (\Delta x, \Delta y) \in N_{\operatorname{graph} F}(x, y)^\circ = \widehat{T}_{\operatorname{graph} F}(x, y) \\ &\Leftrightarrow \Delta y \in \widehat{D}F(x|y)(\Delta x). \end{aligned}$$

□

LIMITING CHARACTERIZATIONS IN FINITE DIMENSIONS

In finite dimensions, we can characterize the limiting coderivative and the Clarke derivative directly as inner and outer limits, respectively.

Corollary 20.9. *Let X and Y be finite-dimensional and $F : X \rightrightarrows Y$. Then for all $(x, y) \in X \times Y$ and all $y^* \in Y$,*

$$D^*F(x|y)(y^*) = \left\{ x^* \in X \;\middle|\; \begin{array}{r} \text{there exists } \operatorname{graph} F \ni (\tilde{x}, \tilde{y}) \to (x, y) \\ \text{and } (\tilde{x}^*, \tilde{y}^*) \to (x^*, y^*) \\ \text{with } \tilde{x}^* \in \widehat{D}^*F(\tilde{x}|\tilde{y})(\tilde{y}^*) \end{array} \right\}. \tag{20.3}$$

If $\operatorname{graph} F$ *is closed near* (x, y)*, then for all* $\Delta x \in \mathbb{R}^N$

$$\widehat{D}F(x|y)(\Delta x) = \left\{ \Delta y \in Y \;\middle|\; \begin{array}{r} \text{for all } \operatorname{graph} F \ni (\tilde{x}, \tilde{y}) \to (x, y) \\ \text{there exists } (\Delta\tilde{x}, \Delta\tilde{y}) \to (\Delta x, \Delta y) \\ \text{with } \Delta\tilde{y} \in DF(\tilde{x}|\tilde{y})(\Delta\tilde{x}) \end{array} \right\}. \tag{20.4}$$

Proof. The characterization (20.3) of the limiting coderivative is a direct application of the definition of the limiting normal cone (18.3) as an outer limit of the Fréchet normal. The characterization (20.4) of the Clarke graphical derivative follows from the characterization of Corollary 18.20 of the Clarke tangent cone as an inner limit of (basic) tangent cones. □

REGULARITY

Based on the regularity concepts of sets from Section 18.4, we can define concepts of regularity of set-valued mappings. We say that F at $(x, y) \in \operatorname{graph} F$ (or at x for $y \in F(x)$) is

(i) *T-regular* if $DF(x|y) = \widehat{D}F(x|y)$ (i.e., if $\operatorname{graph} F$ has tangential regularity);

(ii) *N-regular*, if $D^*F(x|y) = \widehat{D}^*F(x|y)$ (i.e., if $\operatorname{graph} F$ has normal regularity).

If F is both T- and N-regular at (x, y), we say that F is *graphically regular*.

From Theorem 18.25, we immediately obtain the following characterization of N-regularity.

Corollary 20.10. *Let X, Y be reflexive and Gâteaux smooth Banach spaces, $F : X \rightrightarrows Y$, and let $(x, y) \in \operatorname{graph} F$ with $\operatorname{graph} F$ closed near (x, y). Then F is N-regular at (x, y) if and only if $\widehat{D}F(x|y) = [\widehat{D}^*F(x|y)]^{\circ-}$.*

Writing out various alternatives of Theorem 18.23 for set-valued mappings, we obtain full equivalence of the notions and alternative characterizations in finite dimensions.

Corollary 20.11. *Let X, Y be finite-dimensional and $F : X \rightrightarrows Y$. If* graph F *is closed near (x, y), then the following conditions are equivalent:*

(i) F is N-regular at x for y, i.e., $D^*F(x|y) = \widehat{D}^*F(x|y)$;

(ii) F is T-regular at x for y, i.e., $DF(x|y) = \widehat{D}F(x|y)$;

(iii) $$\widehat{D}^*F(x|y)(y^*) \supset \left\{ x^* \in X \;\middle|\; \begin{array}{r} \text{there exists } \operatorname{graph} F \ni (\tilde{x}, \tilde{y}) \to (x, y) \\ \text{and } (\tilde{x}^*, \tilde{y}^*) \to (x^*, y^*) \\ \text{with } \tilde{x}^* \in \widehat{D}^*F(\tilde{x}|\tilde{y})(\tilde{y}^*) \end{array} \right\};$$

(iv) $$DF(x|y)(\Delta x) \subset \left\{ \Delta y \in Y \;\middle|\; \begin{array}{r} \text{for all } (\tilde{x}, \tilde{y}) \to (x, y) \\ \text{there exists } \operatorname{graph} F \ni (\Delta\tilde{x}, \Delta\tilde{y}) \to (\Delta x, \Delta y) \\ \text{with } \Delta\tilde{y} \in DF(\tilde{x}|\tilde{y})(\Delta\tilde{x}) \end{array} \right\}.$$

In particular, if any of these hold, F is graphically regular at x for y.

20.3 EXAMPLES

As the following examples demonstrate, the graphical derivatives and coderivatives generalize classical (sub)differentials.

SINGLE-VALUED MAPPINGS AND THEIR INVERSES

For the Clarke graphical derivative and the limiting coderivatives (which are obtained as inner or outer limits), we have to require – just as for the Clarke subdifferential in Theorem 13.5 – slightly more than just Fréchet differentiability.

Theorem 20.12. *Let X, Y be Banach spaces and let $F : X \to Y$ be single-valued and Fréchet-differentiable at $x \in X$. Then*

$$DF(x|y)(\Delta x) = \begin{cases} \{F'(x)\Delta x\} & \text{if } y = F(x), \\ \emptyset & \text{otherwise,} \end{cases}$$

and

$$\widehat{D}^*F(x|y)(y^*) = \begin{cases} \{F'(x)^*y^*\} & \text{if } y = F(x), \\ \emptyset & \text{otherwise.} \end{cases}$$

If F is continuously Fréchet-differentiable at x, then F is graphically regular at x for $F(x)$, and hence the corresponding expressions also hold for $\widehat{D}F(x|y)$ and $D^*F(x|y)$.

Proof. The graphical derivative: We have $(\Delta x, \Delta y) \in T_{\operatorname{graph} F}(x, y)$ if and only if for some $x_k \to x$, $y_k := F(x_k)$, and $\tau_k \searrow 0$ there holds

$$\Delta x = \lim_{k\to\infty} \frac{x_k - x}{\tau_k} =: \lim_{k\to\infty} \Delta x_k \tag{20.5a}$$

and

$$\Delta y = \lim_{k\to\infty} \frac{y_k - y}{\tau_k} = \lim_{k\to\infty} \frac{F(x + \tau_k \Delta x_k) - F(x)}{\tau_k}. \tag{20.5b}$$

If $\Delta x_k = 0$ for all sufficiently large $k \in \mathbb{N}$, clearly both $\Delta x = 0$ and $\Delta y = 0$. This satisfies the claimed expression. So we may assume that $\Delta x_k \neq 0$ for all $k \in \mathbb{N}$. In this case, (20.5b) holds if and only if

$$\lim_{k\to\infty} \frac{F(x + h_k) - F(x) - \tau_k \Delta y_k}{\|h_k\|_X} = 0$$

for $h_k := \tau_k \Delta x_k$ and any $\Delta y_k \to \Delta y$. Since F is Fréchet differentiable, this clearly holds with

$$\Delta y_k := \tau_k^{-1} F'(x) h_k = F'(x) \Delta x_k \to F'(x) \Delta x =: \Delta y.$$

This shows that $DF(x|y)(\Delta x) = \{F'(x)\Delta x\}$.

The Clarke graphical derivative: To calculate $\widehat{D}F(x|y)$, we have to find all Δx and Δy such that for every $\tau_k \searrow 0$ and $(\tilde{x}_k, \tilde{y}_k) \to (x, y)$ with $\tilde{y}_k = F(\tilde{x}_k)$, there exists $x_k \to x$ with

$$\Delta x = \lim_{k\to\infty} \frac{x_k - \tilde{x}_k}{\tau_k} \qquad \text{and} \qquad \Delta y = \lim_{k\to\infty} \frac{F(x_k) - F(\tilde{x}_k)}{\tau_k}.$$

Setting $x_k = \tilde{x}_k + \tau_k \Delta x_k$ with $\Delta x_k \to \Delta x$, the second condition becomes

$$\Delta y = \lim_{k\to\infty} \frac{F(\tilde{x}_k + \tau_k \Delta x_k) - F(\tilde{x}_k)}{\tau_k}.$$

Taking $\tilde{x}_k = x$, arguing as for DF shows that $\Delta y = F'(x)\Delta x$ is the only candidate. It just remains to show that any choice of $\tilde{x}_k$ gives the same limit, i.e., that

$$\lim_{k\to\infty} \frac{F(\tilde{x}_k + \tau_k \Delta x_k) - F(\tilde{x}_k) - \tau_k F'(x)\Delta x}{\tau_k} = 0.$$

But this follows from the assumed continuous differentiability using Lemma 13.22. Thus for $y = F(x)$,

$$\widehat{D}F(x|y)(\Delta x) = \{F'(x)\Delta x\} = DF(x|y)(\Delta x).$$

This shows that F is T-regular at x for y.

The Fréchet coderivative: The claim follows from proving that

$$\widehat{D}^*_\varepsilon F(x|y)(y^*) = \begin{cases} \mathbb{B}(F'(x)^* y^*, \varepsilon) & \text{if } y = F(x), \\ \emptyset & \text{otherwise,} \end{cases} \tag{20.6}$$

To show this, we note that $x^* \in \widehat{D}^*_\varepsilon F(x|y)(y^*)$ if and only if for every sequence $x_k \to x$ with $F(x_k) \to F(x)$,

$$\limsup_{k\to\infty} \frac{\langle x^*, x_k - x\rangle_X - \langle y^*, F(x_k) - F(x)\rangle_Y}{\sqrt{\|x_k - x\|_X^2 + \|F(x_k) - F(x)\|_Y^2}} \le \varepsilon.$$

Dividing both numerator and denominator by $\|x_k - x\|_X > 0$, we obtain the equivalent condition that

$$\limsup_{k\to\infty} q_k \le \varepsilon \qquad \text{for} \qquad q_k := \frac{\langle x^*, x_k - x\rangle_X - \langle y^*, F(x_k) - F(x)\rangle_Y}{\|x_k - x\|_X}.$$

If we take $x^* \in \mathbb{B}(F'(x)^* y^*, \varepsilon)$, this condition is verified by the Fréchet differentiability of F at x. Conversely, to show that this implies $x^* \in \mathbb{B}(F'(x)^* y^*, \varepsilon)$, we take $x_k := x + \tau_k h$ for some $\tau_k \searrow 0$ and $h \in X$ with $\|h\|_X = 1$. Then again by the Fréchet differentiability of F,

$$\varepsilon \ge \lim_{k\to\infty} q_k = \langle x^*, h\rangle - \langle y^*, F'(x)h\rangle.$$

Since $h \in \mathbb{B}_X$ was arbitrary, this shows that $x^* \in \mathbb{B}(F'(x)^* y^*, \varepsilon)$.

The limiting coderivative: By the definition (18.8), the formula (20.6) for ε-coderivatives, and the continuous differentiability, we have

$$\begin{aligned}
N_{\operatorname{graph} F}(x, F(x)) &= \text{w-}*\text{-}\limsup_{\tilde{x}\to x,\, \varepsilon\searrow 0} \widehat{N}^\varepsilon_{\operatorname{graph} F}(\tilde{x}, F(\tilde{x})) \\
&= \text{w-}*\text{-}\limsup_{\tilde{x}\to x\, \varepsilon\searrow 0} \{(y^*, F'(\tilde{x})^* y^* + z^*) \in Y^* \times X^* \mid y^* \in Y^*,\, z^* \in \mathbb{B}(0,\varepsilon)\} \\
&= \text{w-}*\text{-}\limsup_{\tilde{x}\to x} \{(y^*, F'(\tilde{x})^* y^*) \in Y^* \times X^* \mid y^* \in Y^*\} \\
&= \{(y^*, F'(x)^* y^*) \in Y^* \times X^* \mid y^* \in Y^*\}.
\end{aligned}$$

This shows the claimed formula for the limiting coderivative and hence N- and therefore graphical regularity. □

Remark 20.13. In finite dimensional spaces, it would be possible to more concisely prove the expression for $\widehat{D}F(x|y)$ using Corollary 18.20. Likewise, we could use the polarity relationships of Corollary 20.8 to obtain the expression for $\widehat{D}^*F(x|y)$. These approaches will, however, not be possible in more general spaces.

Combining Theorem 20.12 with Lemma 20.5 allows us to compute the graphical derivatives and coderivatives of inverses of single-valued functions.

Corollary 20.14. *Let X, Y be Banach spaces and let $F : X \to Y$ be single-valued and Fréchet-differentiable at $x \in X$. Then*

$$DF^{-1}(y|x)(\Delta y) = \begin{cases} \{\Delta x \in X \mid F'(x)\Delta x = \Delta y\} & \text{if } y = F(x), \\ \emptyset & \text{otherwise,} \end{cases}$$

and

$$\widehat{D}^* F^{-1}(y|x)(x^*) = \begin{cases} \{y^* \in Y^* \mid F'(x)^* y^* = x^*\} & \text{if } y = F(x), \\ \emptyset & \text{otherwise.} \end{cases}$$

If F is continuously Fréchet-differentiable at x, then F^{-1} is graphically at $y = F(x)$ for x, and hence the corresponding expressions also hold for $\widehat{D}F^{-1}(y|x)$ and $D^*F^{-1}(y|x)$.

It is important that Theorem 20.12 concerns the strong graphical derivatives DF instead of the weak graphical derivative $D^w F$. Indeed, as the next counter-example demonstrates, $D^w F$ is more of a theoretical tool (with the important property in reflexive spaces that $\widehat{D}^* F(x|y) = D^w F(x|y)^{\circ+}$ by Lemma 18.10 (ii)) which does not enjoy a rich calculus consistent with conventional notions. In the following chapters, we will therefore not develop calculus rules for the weak graphical derivative.

Example 20.15 (counter-example to single-valued weak graphical derivatives). Let $f \in C^1(\mathbb{R})$, $\Omega \subset \mathbb{R}^d$ be open, and

$$F : L^2(\Omega) \to \mathbb{R}, \qquad F(u) = \int_0^1 f(u(x))\,dx.$$

Then by the above,

$$DF(u|F(u))(\Delta u) = \left\{ \int_0^1 f'(u(x))\Delta u(x)\,dx \right\}.$$

In particular, $DF(u|F(u))(0) = \{0\}$.

However, choosing, e.g., $f(t) = \sqrt{1+t^2}$, $\Omega = (0,1)$, and $u_k(x) := \operatorname{sign}\sin(2^k \pi x)$, we have $u_k \rightharpoonup 0$ in $L^2(\Omega)$ but $|u_k(x)| = 1$ for a.e. $x \in [0,1]$. Take now $\tilde{u}_k := \alpha \tau_k u_k$ for any given $\tau_k \searrow 0$ and $\alpha > 0$. Then $\tilde{u}_k \rightharpoonup 0$ as well, while

$$F(\tilde{u}_k) - F(0) = \sqrt{1+\alpha^2\tau_k^2} - 1 \to 0.$$

Moreover, $(\tilde{u}_k - 0)/\tau_k = \alpha u_k \rightharpoonup 0$ and $\lim_{k\to\infty} \left(\sqrt{1+\alpha^2\tau_k^2} - 1\right)/\tau_k = \alpha^2$. As $\alpha > 0$ was arbitrary, we deduce that $D^w F(u|F(u))(0) \supset [0,\infty)$.

DERIVATIVES AND CODERIVATIVES OF SUBDIFFERENTIALS

We now apply these notions to set-valued mappings arising as subdifferentials of convex functionals. First, we directly obtain from Theorem 20.12 an expression for the squared norm in Hilbert spaces.

Corollary 20.16. *Let X be a Hilbert space and $f(x) = \frac{1}{2}\|x\|_X^2$ for $x \in X$. Then*

$$\widehat{D}[\partial f](x|y)(\Delta x) = D[\partial f](x|y)(\Delta x) = \begin{cases} \{\Delta x\} & \text{if } y = x, \\ \emptyset & \text{otherwise,} \end{cases}$$

and

$$D^*[\partial f](x|y)(y^*) = \widehat{D}^*[\partial f](x|y)(y^*) = \begin{cases} \{y^*\} & \text{if } y = x, \\ \emptyset & \text{otherwise.} \end{cases}$$

In particular, ∂f is graphically regular at every $x \in X$.

Of course, we are more interested in subdifferentials of nonsmooth functionals. We first study the indicator functional of an interval; see Figure 20.2.

Theorem 20.17. *Let $f(x) := \delta_{[-1,1]}(x)$ for $x \in \mathbb{R}$. Then*

$$D[\partial f](x|y)(\Delta x) = \begin{cases} \mathbb{R} & \text{if } |x| = 1,\ y \in (0,\infty)x,\ \Delta x = 0, \\ [0,\infty)x & \text{if } |x| = 1,\ y = 0,\ \Delta x = 0, \\ \{0\} & \text{if } |x| = 1,\ y = 0,\ x\Delta x < 0, \\ \{0\} & \text{if } |x| < 1,\ y = 0, \\ \emptyset & \text{otherwise,} \end{cases} \tag{20.7}$$

$$\widehat{D}^*[\partial f](x|y)(y^*) = \begin{cases} \mathbb{R}, & \text{if } |x| = 1,\ y \in (0,\infty)x,\ y^* = 0 \\ [0,\infty)x & \text{if } |x| = 1,\ y = 0,\ xy^* \geq 0, \\ \{0\} & \text{if } |x| < 1,\ y = 0, \\ \emptyset & \text{otherwise,} \end{cases} \tag{20.8}$$

$$\widehat{D}[\partial f](x|y)(\Delta x) = \begin{cases} \mathbb{R} & \text{if } |x| = 1,\ y \in (0,\infty)x,\ \Delta x = 0, \\ \{0\} & \text{if } |x| = 1,\ y = 0,\ \Delta x = 0, \\ \{0\} & |x| < 1,\ y = 0, \\ \emptyset & \text{if otherwise,} \end{cases} \tag{20.9}$$

and

$$D^*[\partial f](x|y)(y^*) = \begin{cases} \mathbb{R} & \text{if } |x| = 1,\ y \in [0,\infty)x,\ y^* = 0 \\ [0,\infty)x & \text{if } |x| = 1,\ y = 0,\ xy^* > 0, \\ \{0\} & \text{if } |x| = 1,\ y = 0,\ xy^* < 0, \\ \{0\} & \text{if } |x| < 1,\ y = 0, \\ \emptyset & \text{otherwise.} \end{cases} \tag{20.10}$$

In particular, ∂f is graphically regular at x for $y \in \partial f(x)$ if and only if $|x| < 1$ or $y \neq 0$.

Proof. We first of all recall from Example 4.9 that graph ∂f is closed with

$$\partial f(x) = \begin{cases} [0,\infty)x & \text{if } |x| = 1, \\ \{0\} & \text{if } |x| < 1, \\ \emptyset & \text{otherwise.} \end{cases} \tag{20.11}$$

We now verify (20.7). If $y \in \partial f(x)$ and $\Delta y \in D[\partial f](x|y)(\Delta x)$, there exist by (20.1) sequences $t_k \searrow 0$, $x_k \to x$, and $y_k \in \partial f(x + t_k \Delta x_k)$ such that

$$\Delta x = \lim_{k\to\infty} \frac{x_k - x}{t_k} \qquad \text{and} \qquad \Delta y = \lim_{k\to\infty} \frac{y_k - y}{t_k}. \tag{20.12}$$

We proceed by case distinction.

(i) $|x| = 1$, $\Delta x = 0$, and $y \in (0,\infty)x$: Then choosing $x_k \equiv x$, any $\Delta y \in \mathbb{R}$ and k large enough, we can take $y_k = y + t_k \Delta y \in [0,\infty)x = \partial f(x)$. This yields the first case of (20.7).

(ii) $|x| = 1$, $\Delta x = 0$, but $y = 0$: In this case, choosing $x_k \equiv x$, we can take any $y_k \in \partial f(x + t_k \Delta x_k) = \partial f(x) = [0,\infty)x$. Picking any $\Delta y \in [0,\infty)x$ and setting $y_k := y + t_k \Delta y$, we deduce that $\Delta y \in D[\partial f](x|y)(\Delta x)$. Thus "$\supset$" holds in the second case of (20.7). Since $\Delta y \in -(0,\infty)x$ is clearly not obtainable with $y_k \in [0,\infty)x$, also "$\subset$" holds.

(iii) $|x| = 1$ and $\Delta x = 0$, but $y \in -(0,\infty)x$: Then we have $y_k \in [0,\infty)x$ for k large enough since in this case either $x_k = x$ or $x_k \in (-1,1)$. Thus $|y - y_k| \geq |y| > 0$, so the second limit in (20.12) cannot exist. Therefore the coderivative is empty, which is covered by the last case of (20.7).

(iv) $|x| = 1$ and $x\Delta x > 0$: Then the first limit in (20.12) requires that $x_k \notin \operatorname{dom} \partial f$, and hence $\partial f(x_k) = \emptyset$ for k large enough. This is again covered by the last case of (20.7).

(v) $|x| = 1$ and $x\Delta x < 0$ (the case $x\Delta x = 0$ being covered by (i)–(iii)): Since $\Delta x \neq 0$ has a different sign from x, it follows from the first limit in (20.12) that $x_k \in (-1,1)$ for k large enough. Consequently, $\partial f(x_k) = \{0\}$, i.e., $y_k = 0$. The limit (20.12) in this case only exists if $y = 0$, in which case also $\Delta y = 0$. This is covered by the third case of (20.7), while $y \neq 0$ is covered by the last case.

(vi) $|x| < 1$: Then $y = 0$ and necessarily $y_k = 0$ for k large enough. Therefore also $\Delta y = 0$, which yields the fourth case in (20.7).

(vii) $|x| > 1$: Then $\partial f(x) = \emptyset$ and therefore the coderivative is empty as well, yielding again the final case (20.7).

The expression for $\widehat{D}^*[\partial f](x|y)$ can be verified using Corollary 20.8 (i). It can also be seen graphically from Figure 20.2.

By the inner and outer limit characterizations of Corollary 20.9, we now obtain the expressions for the Clarke graphical derivative $\widehat{D}[\partial f](x|y)$ and the limiting coderivative

$D^*[\partial f](x|y)$. Since graph ∂f is locally contained in an affine subspace outside of the "corner cases" $(x, y) \in \{(1, 0), (-1, 0)\}$, only the latter need special inspection. For the Clarke graphical derivative, we need to write Δy as the limit of $\Delta y_k \in D[\partial f](x_k, y_k)(\Delta x_k)$ for some $\Delta x_k \to \Delta x$ and *all* graph $\partial f \ni (x_k, y_k) \to (x, y)$. Consider for example $(x, y) = (-1, 0)$. Trying both $(x_k, y_k) = (-1+1/k, 0)$ and $(x_k, y_k) = (-1, -1/k)$, we see that this is only possible for $(\Delta x, \Delta y) = (\Delta x_k, \Delta y_k) = (0, 0)$. This yields the second case of (20.9). Conversely, for the limiting coderivative, it suffices to find *one* such sequence from the Fréchet coderivative. Choosing for $(x, y) = (-1, 0)$ again $(x_k, y_k) = (-1 + 1/k, 0)$ and $(x_k, y_k) = (-1, -1/k)$ as well as the constant sequence $(x_k, y_k) = (-1, 0)$ yields the second, third, and first case of (20.16), respectively.

Finally, in finite dimensions the mapping ∂f is graphically regular if and only if $D[\partial f](x|y) = \widehat{D}[\partial f](x|y)$ by Corollary 20.11, which is the case exactly when $|x| < 1$ or $y \neq 0$ as claimed. □

In nonlinear optimization with inequality constraints, the case where ∂f is graphically regular corresponds precisely to the case of *strict complementarity* of the minimizer $\bar{x}$ and the Lagrange multiplier $\bar{y}$ for the constraint $x \in [-1, 1]$.

We next study the different derivatives and graphical regularity of the subdifferential of the absolute value function; see Figure 20.3.

Theorem 20.18. *Let $f(x) := |x|$ for $x \in \mathbb{R}$. Then*

$$D[\partial f](x|y)(\Delta x) = \begin{cases} \{0\} & \text{if } x \neq 0,\ y = \operatorname{sign} x, \\ \{0\} & \text{if } x = 0,\ \Delta x \neq 0,\ y = \operatorname{sign} \Delta x, \\ (-\infty, 0]\, y & \text{if } x = 0,\ \Delta x = 0,\ |y| = 1, \\ \mathbb{R} & \text{if } x = 0,\ \Delta x = 0,\ |y| < 1, \\ \emptyset & \text{if otherwise,} \end{cases} \tag{20.13}$$

$$\widehat{D}^*[\partial f](x|y)(y^*) = \begin{cases} \{0\} & \text{if } x \neq 0,\ y = \operatorname{sign} x, \\ (-\infty, 0]\, y & \text{if } x = 0,\ yy^* \leq 0,\ |y| = 1, \\ \mathbb{R} & \text{if } x = 0,\ y^* = 0,\ |y| < 1, \\ \emptyset & \text{otherwise,} \end{cases} \tag{20.14}$$

$$\widehat{D}[\partial f](x|y)(\Delta x) = \begin{cases} \{0\} & \text{if } x \neq 0,\ y = \operatorname{sign} x, \\ \{0\} & \text{if } x = 0,\ \Delta x = 0,\ |y| = 1, \\ \mathbb{R} & \text{if } x = 0,\ \Delta x = 0,\ |y| < 1, \\ \emptyset & \text{otherwise,} \end{cases} \tag{20.15}$$

and

$$D^*[\partial f](x|y)(y^*) = \begin{cases} \{0\} & \text{if } x \neq 0,\ y = \operatorname{sign} x, \\ \{0\} & \text{if } x = 0,\ yy^* > 0,\ |y| = 1, \\ (-\infty, 0]\, y & \text{if } x = 0,\ yy^* < 0,\ |y| = 1, \\ \mathbb{R} & \text{if } x = 0,\ y^* = 0,\ |y| \leq 1, \\ \emptyset & \text{otherwise.} \end{cases} \tag{20.16}$$

In particular, ∂f is graphically regular if and only if $x \neq 0$ or $|y| < 1$.

Proof. To start with proving (20.13), we recall from Example 4.7 that

$$\partial f(x) = \operatorname{sign}(x) = \begin{cases} \{1\} & \text{if } x > 0 \\ \{-1\} & \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0. \end{cases} \tag{20.17}$$

To calculate the graphical derivative, we use that if $y \in \partial f(x)$ and $\Delta y \in D[\partial f](x|y)(\Delta x)$, there exist by (20.1) sequences $t_k \searrow 0$, $x_k \to x$, and $y_k \in \partial f(x + t_k \Delta x_k)$ such that

$$\Delta x = \lim_{k\to\infty} \frac{x_k - x}{t_k} \qquad \text{and} \qquad \Delta y = \lim_{k\to\infty} \frac{y_k - y}{t_k}. \tag{20.18}$$

We proceed by case distinction:

(i) $x \neq 0$ and $y \neq \operatorname{sign} x$: Then $y \notin \partial f(x)$ and therefore $D[\partial f](x|y) = \emptyset$, which is covered by the last case of (20.13).

(ii) $x \neq 0$ and $y = \operatorname{sign} x$: Then for any $x_k \to x$, we have that $\partial f(x_k) = \partial f(x) = \{\operatorname{sign} x\}$ for k large enough. Therefore, for any $\Delta x \in \mathbb{R}$ we have that $\Delta y = 0$, which is the first case of (20.13).

(iii) $x = 0$ and $\Delta x \neq 0$: Then $x_k \neq 0$ and $y_k = \operatorname{sign} x_k = \operatorname{sign} \Delta x$. Therefore the limits in (20.18) will only exist if $|y| = 1$, which holds from $y = \operatorname{sign} \Delta x$. Thus $\Delta y = 0$, i.e., we obtain the second case of (20.13).

(iv) $x = 0$ and $\Delta x = 0$: Then taking $x_k \equiv x$, we can choose $y_k \in [-1, -1]$ arbitrarily. If $|y| = 1$, then $(y - y_k) \operatorname{sign} y \leq 0$, so (20.18) shows that $\Delta y \operatorname{sign} y \leq 0$, which is the third case of (20.13). If $|y| < 1$, we may obtain any $\Delta y \in \mathbb{R}$ by the limit in (20.18). This is the fourth case of (20.13).

The expression for $\widehat{D}^*[\partial f](x|y)$ can be verified using Corollary 20.8 (i). It can also be seen graphically from Figure 20.3.

By the inner and outer limit characterizations of Corollary 20.9, we now obtain the expressions for the Clarke graphical derivative $\widehat{D}[\partial f](x|y)$ and the limiting coderivative $D^*[\partial f](x|y)$. Since graph ∂f is locally contained in an affine subspace outside of the "corner cases" $(x, y) \in \{(0, 1), (0, -1)\}$, only the latter need special inspection. For the Clarke

graphical derivative, we need to write Δy as the limit of $\Delta y_k \in D[\partial f](x_k, y_k)(\Delta x_k)$ for some $\Delta x_k \to \Delta x$ and *all* graph $\partial f \ni (x_k, y_k) \to (x, y)$. Consider for example $(x, y) = (0, -1)$. Trying both $(x_k, y_k) = (0, -1+1/k)$ and $(x_k, y_k) = (-1/k, -1)$, we see that this is only possible for $(\Delta x, \Delta y) = (\Delta x_k, \Delta y_k) = (0, 0)$. This yields the third case of (20.15). Conversely, for the limiting coderivative, it suffices to find *one* such sequence from the Fréchet coderivative. Choosing for $(x, y) = (0, -1)$ again $(x_k, y_k) = (0, -1 + 1/k)$ and $(x_k, y_k) = (-1/k, 1)$ as well as the constant sequence $(x_k, y_k) = (-1, 0)$ yields the fourth, second, and third case of (20.16), respectively.

Finally, in finite dimensions the mapping ∂f is graphically regular if and only if $D[\partial f](x|y) = \widehat{D}[\partial f](x|y)$ by Corollary 20.11, which is the case exactly when $x \neq 0$ or $|y| < 1$ as claimed. □

20.4 RELATION TO SUBDIFFERENTIALS

All of the subdifferentials that we have studied in Part III can be constructed from the corresponding normal cones to the epigraph of a functional $J : X \to \overline{\mathbb{R}}$ as in the convex case; see Lemma 4.10. For the Fréchet and limiting subdifferentials, it is easy to see the relationships

$$\partial_F J(x) = \{x^* \in X^* \mid (x^*, -1) \in \widehat{N}_{\operatorname{epi} J}(x, J(x))\}, \tag{20.19}$$

$$\partial_M J(x) = \{x^* \in X^* \mid (x^*, -1) \in N_{\operatorname{epi} J}(x, J(x))\}, \tag{20.20}$$

from the corresponding definitions. For the Clarke subdifferential, however, we have to work a bit harder.

First, we define for $A \subset X$ and $x \in X$ the *Clarke normal cone*

$$N_A^C(x) := \widehat{T}_A(x)^\circ. \tag{20.21}$$

We can now extend the definition of the Clarke subdifferential to arbitrary functionals $J : X \to \overline{\mathbb{R}}$ on Gâteaux smooth Banach spaces via the Clarke normal cone to their epigraph.

Lemma 20.19. *Let X be a reflexive and Gâteaux smooth Banach space and let $J : X \to \mathbb{R}$ be locally Lipschitz continuous around $x \in X$. Then*

$$\partial_C J(x) = \{x^* \in X^* \mid (x^*, -1) \in N^C_{\operatorname{epi} J}(x, J(x))\}.$$

Proof. The Clarke tangent cone to epi J by definition is

$$\widehat{T}_{\operatorname{epi} J}(x, J(x)) = \left\{ (\Delta x, \Delta t) \in X \times \mathbb{R} \;\middle|\; \begin{array}{r} \text{for all } \tau_k \searrow 0,\ x_k \to x,\ J(x_k) \le t_k \to J(x) \\ \text{there exist } \tilde{x}_k \in X \text{ and } \tilde{t}_k \ge J(\tilde{x}_k) \\ \text{with } (\tilde{x}_k - x_k)/\tau_k \to \Delta x \text{ and } (\tilde{t}_k - t_k)/\tau_k \to \Delta t \end{array} \right\}.$$

If $(\Delta x, \Delta t) \in \widehat{T}_{\operatorname{epi} J}(x, J(x))$, then replacing $\tilde{t}_k$ by $\tilde{t}_k + \tau_k(\Delta s - \Delta t) \geq J(\tilde{x}_k)$ shows that also $(\Delta x, \Delta s) \in \widehat{T}_{\operatorname{epi} J}(x, J(x))$ for all $\Delta s \geq \Delta t$. Thus we may make the minimal choices $\tilde{t}_k = J(\tilde{x}_k)$ and $t_k = J(x_k)$ to see that

$$\widehat{T}_{\operatorname{epi} J}(x, J(x)) = \left\{ (\Delta x, \Delta t) \in X \times \mathbb{R} \;\middle|\; \begin{array}{r} \text{for all } \tau_k \searrow 0,\ x_k \to x \text{ there exist } \tilde{x}_k \in X \\ \text{with } (\tilde{x}_k - x_k)/\tau_k \to \Delta x \\ \text{and } \limsup_{k\to\infty} (J(\tilde{x}_k) - J(x_k))/\tau_k \leq \Delta t \end{array} \right\}.$$

Since J is locally Lipschitz continuous, it suffices to take $\tilde{x}_k = x_k + \tau_k \Delta x$ to obtain

$$\widehat{T}_{\operatorname{epi} J}(x, J(x)) = \{(\Delta x, \Delta t) \in X \times \mathbb{R} \mid x \in X,\ \Delta t \geq J^\circ(x; \Delta x)\} = \operatorname{epi}[J^\circ(x; \cdot)].$$

Hence $(x^*, -1) \in N^C_{\operatorname{epi} J}(x, J(x)) = \widehat{T}_{\operatorname{epi} J}(x, J(x))^\circ$ if and only if $\langle x^*, \Delta x\rangle_X \leq J^\circ(x; \Delta x)$ for all $x \in X$, which by definition is equivalent to $x^* \in \partial_C J(x)$. □

We furthermore have the following relationship between the Clarke and limiting normal cones.

Corollary 20.20. *Let X be a reflexive and Gâteaux smooth Banach space and $A \subset X$ be closed near $x \in A$. Then*

$$N^C_A(x) = N_A(x)^{\circ\circ} = \operatorname{cl} \operatorname{co}^* N_A(x),$$

where $\operatorname{cl} \operatorname{co}^*$ *denotes the weak-$*$ closed convex hull.*

Proof. First, $N_A(x) \neq \emptyset$ since $x \in A$. Furthermore, $\operatorname{cl} \operatorname{co}^* N_A(x)$ is the smallest weak-$*$-closed and convex set that contains $N_A(x)$, and therefore Theorem 1.8 and Lemma 1.10 imply $N_A(x)^{\circ\circ} = \operatorname{cl} \operatorname{co}^* N_A(x)^{\circ\circ} = \operatorname{cl} \operatorname{co}^* N_A(x)$. The relationship $N^C_A(x) = N_A(x)^{\circ\circ}$ is an immediate consequence of Theorem 18.19. □

Assuming that X is Gâteaux smooth, we now have everything at hand to give a proof of Theorem 16.10, which characterizes the Clarke subdifferential as the weak-$*$ closed convex hull of the limiting subdifferential.

Corollary 20.21. *Let X be a reflexive and Gâteaux smooth Banach space and $J : X \to \mathbb{R}$ be locally Lipschitz continuous around $x \in X$. Then $\partial_C J(x) = \operatorname{cl}^* \operatorname{co} \partial_M J(x)$.*

Proof. Together, Lemma 20.19 and Corollary 20.20 and (20.20) directly yield

$$\begin{aligned} \partial_C J(x) &= \{x^* \in X^* \mid (x^*, -1) \in N^C_{\operatorname{epi} J}(x, J(x))\} \\ &= \{x^* \in X^* \mid (x^*, -1) \in \operatorname{cl}^* \operatorname{co} N_{\operatorname{epi} J}(x, J(x))\} \\ &= \operatorname{cl}^* \operatorname{co}\{x^* \in X^* \mid (x^*, -1) \in N_{\operatorname{epi} J}(x, J(x))\} \\ &= \operatorname{cl}^* \operatorname{co} \partial_M J(x). \end{aligned}$$

□

(The Gâteaux smoothness of X can be relaxed to X being an Asplund space following Remark 17.8.)

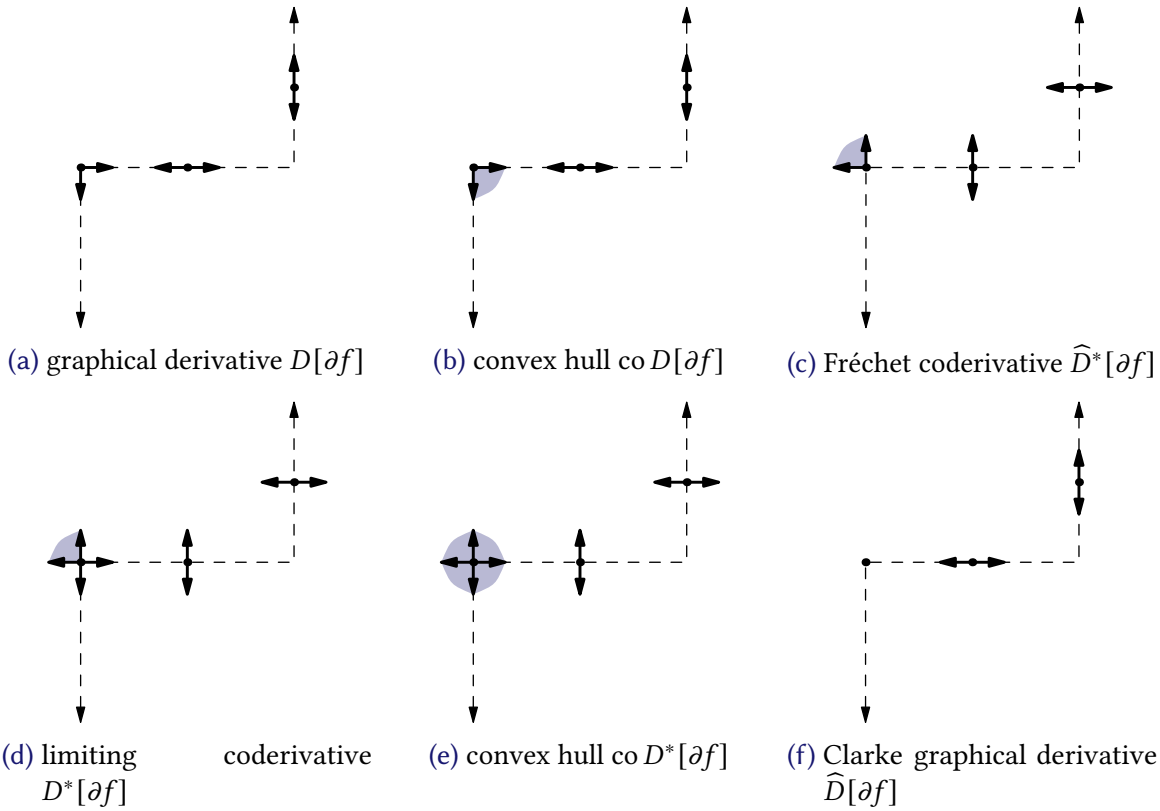
From the corresponding definitions, it also follows that

$$\partial_F J(x) = \widehat{D}^*[\operatorname{epif} J](x|J(x))(1),$$
$$\partial_M J(x) = D^*[\operatorname{epif} J](x|J(x))(1),$$

where the *epigraphical function*

$$\operatorname{epif} J(x) := \{t \in \mathbb{R} \mid t \geq J(x)\}$$

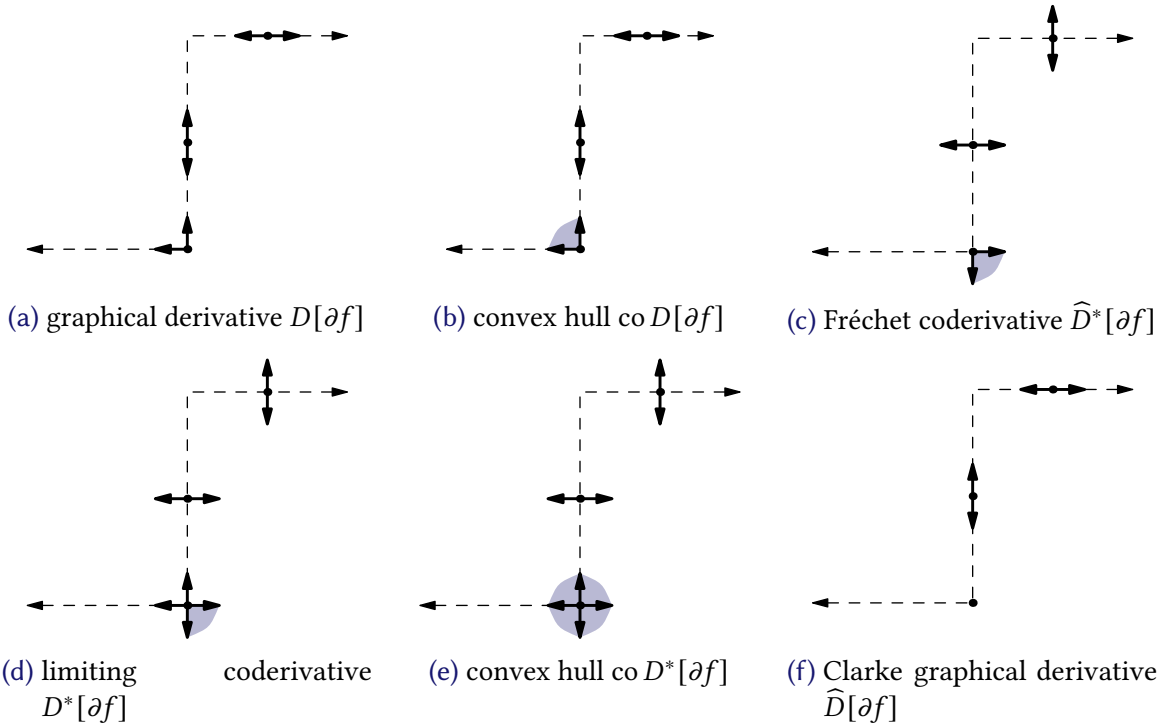
satisfies $\operatorname{graph}[\operatorname{epif} J] = \operatorname{epi} J$. Thus the results of the following Chapters 23 and 25 can be used to derive the missing calculus rules for the Fréchet and limiting subdifferentials. In particular, Theorem 25.14 will provide the missing proof of the sum rule (Theorem 16.13) for the limiting subdifferential.



(a) graphical derivative $D[\partial f]$ (b) convex hull co $D[\partial f]$ (c) Fréchet coderivative $\widehat{D}^*[\partial f]$

(d) limiting coderivative $D^*[\partial f]$ (e) convex hull co $D^*[\partial f]$ (f) Clarke graphical derivative $\widehat{D}[\partial f]$

Figure 20.2: Illustration of the different graphical derivatives and coderivatives of ∂f for $f = \delta_{[-1,1]}$. The dashed line is graph ∂f. The dots indicate the base points (x, y) where $D[\partial f](x|y)$ is calculated, and the thick arrows and filled-in areas the directions of $(\Delta x, \Delta y)$ (resp. $(\Delta x, -\Delta y)$ for the coderivatives) relative to the base point. Observe that there is no graphical regularity at $(x, y) \in \{(-1, 0), (1, 0)\}$. Everywhere else, ∂f is graphically regular. Observe also that cones in the last figures of each row are polar to the cones in the first and the second figures on the same row.



(a) graphical derivative $D[\partial f]$ (b) convex hull co $D[\partial f]$ (c) Fréchet coderivative $\widehat{D}^*[\partial f]$

(d) limiting coderivative $D^*[\partial f]$ (e) convex hull co $D^*[\partial f]$ (f) Clarke graphical derivative $\widehat{D}[\partial f]$

Figure 20.3: Illustration of the different graphical derivatives and coderivatives of ∂f for $f = |\cdot|$. The dashed line is graph ∂f. The dots indicate the base points (x, y) where $D[\partial f](x|y)$ is calculated, and the thick arrows and filled-in areas the directions of $(\Delta x, \Delta y)$ (resp. $(\Delta x, -\Delta y)$ for the coderivatives) relative to the base point. Observe that there is no graphical regularity at $(x, y) \in \{(0, -1), (0, 1)\}$. Everywhere else, ∂f is graphically regular. Observe that cones in the last figures of each row are polar to the cones in the first and the second figures on the same row.

21 DERIVATIVES AND CODERIVATIVES OF POINTWISE-DEFINED MAPPINGS

Just as for tangent and normal cones, the relationships between the basic and limiting derivatives and coderivatives are less complete in infinite-dimensional spaces than in finite-dimensional ones. In this chapter, we apply the results of Chapter 19 to derive pointwise characterizations analogous to Theorem 4.11 for the basic derivatives of pointwise-defined set-valued mappings, which (only) in the case of graphically regularity transfer to their limiting variants.

21.1 PROTO-DIFFERENTIABILITY

For our superposition formulas, we need some regularity from the finite-dimensional mappings. The appropriate notion is that of *proto-differentiability*, which corresponds to the *geometric derivability* of the underlying tangent cone.

Let X, Y be Banach spaces. We say that a set-valued mapping $F : X \rightrightarrows Y$ is *proto-differentiable* at $x \in X$ for $y \in F(x)$ if

$$\text{(21.1a)} \quad \text{for every } \Delta y \in DF(x|y)(\Delta x) \quad \text{and} \quad \tau_k \searrow 0,$$

$$\text{(21.1b)} \quad \text{there exist } x_k \in X \text{ with } \frac{x_k - x}{\tau_k} \to \Delta x \quad \text{and} \quad y_k \in F(x_k) \text{ with } \frac{y_k - y}{\tau_k} \to \Delta y.$$

In other words, in addition to the basic limit (20.1) defining $DF(x|y)$, a corresponding inner limit holds in the graph space.

By application of Lemma 19.1 and Corollary 19.3, we immediately obtain the following equivalent characterization.

Corollary 21.1. *Let X, Y be Banach spaces and $F : X \rightrightarrows Y$. Then F is proto-differentiable at every $x \in X$ for every $y \in F(x)$ if and only if* graph F *is geometrically derivable at (x, y). In particular, if F is graphically regular at (x, y), then F is proto-differentiable at x for y.*

Clearly, differentiable single-valued mappings are proto-differentiable. Another large class are maximally monotone set-valued mappings on Hilbert spaces.

Lemma 21.2. *Let X be a Hilbert space and let $A : X \rightrightarrows X$ be maximally monotone. Then A is proto-differentiable at any $x \in \operatorname{dom} A$ for any $x^* \in A(x)$.*

Proof. Let $\Delta x^* \in D[A](x|x^*)(\Delta x)$. By definition, there then exist $\tau_k \searrow 0$ and $(x_k, x_k^*) \in \operatorname{graph} A$ such that $(x_k - x)/\tau_k \to \Delta x$ and $(x_k^* - x^*)/\tau_k \to \Delta x^*$. To show that A is proto-differentiable, we will construct for an arbitrary sequence $\tilde{\tau}_k \searrow 0$ sequences $(\tilde{x}_k, \tilde{x}_k^*) \in \operatorname{graph} A$ such that $(\tilde{x}_k - x)/\tilde{\tau}_k \to \Delta x$ and $(\tilde{x}_k^* - x^*)/\tilde{\tau}_k \to \Delta x^*$. We will do so using resolvents. Similarly to Lemma 6.21, we have that

$$\begin{aligned} x^* \in A(x) \quad &\Leftrightarrow \quad x \in A^{-1}(x^*) \quad \Leftrightarrow \quad x^* + x \in \{x^*\} + A^{-1}(x^*) \\ &\Leftrightarrow \quad x^* \in \mathcal{R}_{A^{-1}}(x^* + x). \end{aligned}$$

Since A is maximally monotone and X is reflexive, A^{-1} is maximally monotone by Lemma 6.9 as well, and thus the resolvent $\mathcal{R}_{A^{-1}}$ is single-valued by Corollary 6.16. We therefore take

$$\begin{aligned} \tilde{x}_k &:= x + \frac{\tilde{\tau}_k}{\tau_k}(x_k - x) + \frac{\tilde{\tau}_k}{\tau_k} x_k^* + \left(1 - \frac{\tilde{\tau}_k}{\tau_k}\right) x^* - \tilde{x}_k^* \quad \text{and} \\ \tilde{x}_k^* &:= \mathcal{R}_{A^{-1}}\left(x + \frac{\tilde{\tau}_k}{\tau_k}(x_k - x) + \frac{\tilde{\tau}_k}{\tau_k} x_k^* + \left(1 - \frac{\tilde{\tau}_k}{\tau_k}\right) x^* - \tilde{\tau}_k(\Delta x + \Delta x^*)\right) + \tilde{\tau}_k \Delta x^* \\ &= \mathcal{R}_{A^{-1}}(\tilde{x}_k^* + \tilde{x}_k - \tilde{\tau}_k(\Delta x + \Delta x^*)) + \tilde{\tau}_k \Delta x^*. \end{aligned}$$

Since resolvents of maximally monotone operators are 1-Lipschitz by Lemma 6.15, we have

$$\begin{aligned} \lim_{k\to\infty} \frac{\|\tilde{x}_k^* - x^* - \tilde{\tau}_k \Delta x^*\|_X}{\tilde{\tau}_k} &= \lim_{k\to\infty} \frac{\|\mathcal{R}_{A^{-1}}(\tilde{x}_k^* + \tilde{x}_k - \tilde{\tau}_k(\Delta x + \Delta x^*)) - \mathcal{R}_{A^{-1}}(x^* + x)\|_X}{\tilde{\tau}_k} \\ &\le \lim_{k\to\infty} \frac{\|(\tilde{x}_k^* + \tilde{x}_k - \tilde{\tau}_k(\Delta x + \Delta x^*)) - (x^* + x)\|_X}{\tilde{\tau}_k} \\ &= \lim_{k\to\infty} \frac{\|(x_k - x - \tau_k \Delta x) + (x_k^* - x^* - \tau_k \Delta x^*)\|_X}{\tau_k} = 0. \end{aligned}$$

Likewise, by inserting the definition of $\tilde{x}_k$ and using the triangle inequality, we obtain

$$\begin{aligned} \lim_{k\to\infty} \frac{\|\tilde{x}_k - x - \tilde{\tau}_k \Delta x\|_X}{\tilde{\tau}_k} &\le \lim_{k\to\infty} \frac{\|(x_k - x - \tau_k \Delta x) + (x_k^* - x^* - \tau_k \Delta x^*)\|_X}{\tau_k} \\ &\quad + \lim_{k\to\infty} \frac{\|\tilde{x}_k^* - x^* - \tilde{\tau}_k \Delta x^*\|_X}{\tilde{\tau}_k} \\ &= 0. \end{aligned}$$

This shows the claimed proto-differentiability. □

Since subdifferentials of convex and lower semicontinuous functionals on reflexive Banach spaces are maximally monotone by Theorem 6.13, we immediately obtain the following.

Corollary 21.3. *Let X be a Hilbert space and let $J : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. Then ∂J is proto-differentiable at any $x \in \operatorname{dom} J$ for any $x^* \in \partial J(x)$.*

This corollary combined with Theorems 20.17 and 20.18 shows that proto-differentiability is a strictly weaker property than graphical regularity.

21.2 GRAPHICAL DERIVATIVES AND CODERIVATIVES

As a corollary of the tangent and normal cone representations from Theorems 19.5 and 19.6, we obtain explicit characterizations of the graphical derivative and the Fréchet coderivative of a class of pointwise-defined set-valued mappings. In the following, let $\Omega \subset \mathbb{R}^d$ be an open and bounded domain and write again p^* for the conjugate exponent of $p \in (1, \infty)$ satisfying $1/p + 1/p^* = 1$.

Theorem 21.4. *Let $F : L^p(\Omega) \rightrightarrows L^q(\Omega)$ for $p, q \in (1, \infty)$ have the form*

$$F(u) = \{w \in L^q(\Omega) \mid w(x) \in f(u(x)) \text{ for a.e. } x \in \Omega\}$$

for some pointwise almost everywhere proto-differentiable mapping $f : \mathbb{R} \rightrightarrows \mathbb{R}$. Then for every $w^* \in L^{q^*}(\Omega)$ and $\Delta u \in L^p(\Omega)$,

$$\widehat{D}^* F(u|w)(w^*) = \left\{ u^* \in L^{p^*}(\Omega) \;\middle|\; \begin{array}{r} u^*(x) \in \widehat{D}^* f(u(x)|w(x))(w^*(x)) \\ \text{for a.e. } x \in \Omega \end{array} \right\}, \tag{21.2a}$$

$$DF(u|w)(\Delta u) = \left\{ \Delta w \in L^q(\Omega) \;\middle|\; \begin{array}{r} \Delta w(x) \in Df(u(x)|w(x))(\Delta u(x)) \\ \text{for a.e. } x \in \Omega \end{array} \right\}. \tag{21.2b}$$

Moreover, if f is graphically regular at $u(x)$ for $w(x)$ for almost every $x \in \Omega$, then F is graphically regular at u for w and

$$\begin{aligned} \widehat{D} F(u|w) &= D^w F(u|w) = DF(u|w), \\ D^* F(u|w) &= \widehat{D}^* F(u|w). \end{aligned}$$

Proof. First, graph f is geometrically derivable by Corollary 21.1 due to the assumed proto-differentiability of f. We further have

$$\operatorname{graph} F = \left\{ (u, w) \in L^p(\Omega) \times L^q(\Omega) \;\middle|\; (u(x), w(x)) \in \operatorname{graph} f \text{ for a.e. } x \in \Omega \right\}.$$

Now (21.2b) and (21.2a) follow from Theorems 19.5 and 19.6, respectively, for $C : x \mapsto \operatorname{graph} f$ and $U = \operatorname{graph} F$ together with definitions of the graphical derivative in terms of the tangent cone the Fréchet coderivative in terms of the Fréchet normal cone. The remaining claims under graphical regularity follow similarly from Lemma 19.11. □

The above result directly applies to second derivatives of integral functionals.

Corollary 21.5. *Let $J : L^p(\Omega) \to \overline{\mathbb{R}}$ for $p \in (1, \infty)$ be given by*

$$J(u) = \int_\Omega j(u(x))\,dx$$

for some proper, convex, and lower semicontinuous integrand $j : \mathbb{R} \to (-\infty, \infty]$. Then

$$\widehat{D}^*[\partial J](u|u^*)(\Delta u) = \left\{ \Delta u^* \in L^{p^*}(\Omega) \;\middle|\; \begin{array}{r} \Delta u^*(x) \in \widehat{D}^*[\partial j](u(x)|u^*(x))(\Delta u(x)) \\ \text{for a.e. } x \in \Omega \end{array} \right\},$$
$$D[\partial J](u|u^*)(\Delta u) = \left\{ \Delta u^* \in L^{p^*}(\Omega) \;\middle|\; \begin{array}{r} \Delta u^*(x) \in D[\partial j](u(x)|u^*(x))(\Delta u(x)) \\ \text{for a.e. } x \in \Omega \end{array} \right\}.$$

Moreover, if ∂j is graphically regular at $u(x)$ for $u^*(x)$ for almost every $x \in \Omega$, then ∂J is graphically regular at u for u^* and

$$\widehat{D}[\partial J](u|u^*) = D^w[\partial J](u|u^*) = D[\partial J](u|u^*),$$
$$D^*[\partial J](u|u^*) = \widehat{D}^*[\partial J](u|u^*).$$

Proof. By Corollary 21.3, ∂j is proto-differentiable. Since

$$\partial J(u) = \left\{ u^* \in L^{p^*}(\Omega) \mid u^*(x) \in \partial j(u(x)) \text{ for a.e. } x \in \Omega \right\}$$

by Theorem 4.11 and therefore

$$\operatorname{graph}[\partial J] = \left\{ (u, u^*) \in L^p(\Omega) \times L^{p^*}(\Omega) \mid u^*(x) \in \partial j(u(x)) \text{ for a.e. } x \in \Omega \right\},$$

the remaining claims follow from Theorem 21.4 with $F = \partial J$, $f = \partial j$, and $q = p^*$. □

Remark 21.6. The case of vector-valued and spatially-varying set-valued mappings and convex integrands can be found in [Clason and Valkonen, 2017b].

We illustrate this result with the usual examples. To keep the presentation simple, we focus on the case $p^* = p = 2$ such that $L^2(\Omega)$ is a Hilbert space and we can identify $X \cong X^*$.

First, we immediately obtain from Corollary 20.16 together with Corollary 21.5

Corollary 21.7. *Let $J : L^2(\Omega) \to \mathbb{R}$ be given by*

$$J(u) \coloneqq \int_\Omega \frac{1}{2}|u(x)|^2\,dx.$$

Then for $u^* = u$ and all $\Delta u \in L^2(\Omega)$, we have

$$\widehat{D}[\partial J](u|u^*)(\Delta u) = D^w[\partial J](u|u^*)(\Delta u) = D[\partial J](u|u^*)(\Delta u) = \Delta u,$$
$$D^*[\partial J](u|u^*)(\Delta u) = \widehat{D}^*[\partial J](u|u^*)(\Delta u) = \Delta u.$$

If $u^* \neq u$, all the derivatives and coderivatives are empty.

From Theorem 20.17, we also obtain expressions for the basic derivatives of indicator functionals for pointwise constraints. For the limiting derivatives, we only obtain expressions at points where graphical regularity (corresponding to strict complementarity) holds; cf. Remark 19.14.

Corollary 21.8. *Let $J : L^2(\Omega) \to \overline{\mathbb{R}}$ be given by*

$$J(u) := \int_\Omega \delta_{[-1,1]}(u(x))\,dx.$$

Let $u \in \operatorname{dom} J$ and $u^* \in \partial J(u)$. Then $\Delta u^* \in D[\partial J](u|u^*)(\Delta u) \subset L^2(\Omega)$ if and only if for almost every $x \in \Omega$,

$$\Delta u^*(x) \in \begin{cases} \mathbb{R} & \text{if } |u(x)| = 1,\ u^*(x) \in (0,\infty)u(x),\ \Delta u(x) = 0, \\ [0,\infty)u(x) & \text{if } |u(x)| = 1,\ u^*(x) = 0,\ \Delta u(x) = 0, \\ \{0\} & \text{if } |u(x)| = 1,\ u^*(x) = 0,\ u(x)\Delta u(x) < 0, \\ \{0\} & \text{if } |u(x)| < 1,\ u^*(x) = 0, \\ \emptyset & \text{otherwise.} \end{cases}$$

Similarly, $\Delta u \in D[\partial J](u|u^*)(\Delta u^*) \subset L^2(\Omega)$ if and only if for almost every $x \in \Omega$,

$$\Delta u(x) \in \begin{cases} \mathbb{R}, & \text{if } |u(x)| = 1,\ u^*(x) \in (0,\infty)u(x), \Delta u^*(x) = 0, \\ [0,\infty)u(x) & \text{if } |u(x)| = 1,\ u^*(x) = 0,\ u(x)\Delta u^*(x) \geq 0, \\ \{0\} & \text{if } |u(x)| < 1,\ u^*(x) = 0, \\ \emptyset & \text{otherwise.} \end{cases}$$

If either $|u(x)| < 1$ or $u^*(x) \neq 0$, then $\Delta u^* \in \widehat{D}[\partial J](u|u^*)(\Delta u) = D^*[\partial J](u|u^*)(\Delta u)$ if and only if for almost every $x \in \Omega$,

$$\Delta u^*(x) \in \begin{cases} \mathbb{R} & \text{if } |u(x)| = 1,\ u^*(x) \in (0,\infty)u(x),\ \Delta u(x) = 0, \\ \{0\} & \text{if } |u(x)| < 1,\ \Delta u(x) \in \mathbb{R}, \\ \emptyset & \text{otherwise.} \end{cases}$$

A similar characterization holds for the basic derivatives of the L^1 norm (as a functional on $L^2(\Omega)$).

Corollary 21.9. *Let $J : L^2(\Omega) \to \mathbb{R}$ be given by*

$$J(u) := \int_\Omega |u(x)|\,dx.$$

Let $u \in \operatorname{dom} J$ and $u^* \in \partial J(u)$. Then $\Delta u^* \in D[\partial J](u|u^*)(\Delta u) \subset L^2(\Omega)$ if and only if for almost every $x \in \Omega$,

$$\Delta u^*(x) \in \begin{cases} \mathbb{R} & \text{if } |u(x)| = 1,\ u^*(x) \in (0,\infty)u(x),\ \Delta u(x) = 0, \\ [0,\infty)u(x) & \text{if } |u(x)| = 1,\ u^*(x) = 0,\ \Delta u(x) = 0, \\ \{0\} & \text{if } |u(x)| = 1,\ u^*(x) = 0,\ u(x)\Delta u(x) < 0, \\ \{0\} & \text{if } |u(x)| < 1,\ u^*(x) = 0, \\ \emptyset & \text{otherwise,} \end{cases}$$

Similarly, $\Delta u \in D[\partial J](u|u^*)(\Delta u^*) \subset L^2(\Omega)$ if and only if for almost every $x \in \Omega$,

$$\Delta u(x) \in \begin{cases} \mathbb{R}, & \text{if } |u(x)| = 1,\ u^*(x) \in (0,\infty)u(x), \Delta u^*(x) = 0, \\ [0,\infty)u(x) & \text{if } |u(x)| = 1,\ u^*(x) = 0,\ u(x)\Delta u^*(x) \geq 0, \\ \{0\} & \text{if } |u(x)| < 1,\ u^*(x) = 0, \\ \emptyset & \text{otherwise,} \end{cases}$$

If either $u(x) \neq 0$ or $|u^*(x)| < 1$, then $\Delta u^* \in \widehat{D}[\partial J](u|u^*)(\Delta u) = D^*[\partial J](u|u^*)(\Delta u)$ if and only if for almost every $x \in \Omega$,

$$\Delta u^*(x) \in \begin{cases} \{0\} & \text{if } u(x) \neq 0,\ u^*(x) = \operatorname{sign} u(x),\ \Delta u(x) \in \mathbb{R}, \\ \mathbb{R} & \text{if } u(x) = 0,\ |u^*(x)| < 1,\ \Delta u(x) = 0, \\ \emptyset & \text{otherwise.} \end{cases}$$

Obtaining similar characterizations for derivatives of the Clarke subdifferential of integral functions with nonsmooth nonconvex integrands requires verifying proto-differentiability of the pointwise subdifferential mapping, which is challenging since the Clarke subdifferential in general does not have the nice properties of the convex subdifferential as a set-valued mapping. For problems of the form (P) in the introduction, it is therefore simpler to first apply the calculus rules from the following chapters (assuming they are applicable) and to then use the above results for the derivatives of the convex or smooth component mappings.

22 CALCULUS FOR THE GRAPHICAL DERIVATIVE

We now turn to calculus such as sum and product rules. We concentrate on the situation where at least one of the mappings involved is classically differentiable, which allows exact results and is already useful in practice. For a much fuller picture of infinite-dimensional calculus in high generality, the reader is referred to [Mordukhovich, 2006]. For further finite-dimensional calculus we refer to [Mordukhovich, 2018; Rockafellar and Wets, 1998].

The rules we develop for the various (co)derivatives are in each case based on linear transformation formulas of the underlying cones as well as on a fundamental composition lemma. These fundamental lemmas, however, require further regularity assumptions that are satisfied in particular by (continuously) Fréchet differentiable single-valued mappings and their inverses. For the sake of presentation, we treat each derivative in its own chapter, starting with the relevant regularity concept, then proving the fundamental lemmas, and finally deriving the calculus rules. We start with the (basic) graphical derivative.

22.1 SEMI-DIFFERENTIABILITY

Let X, Y be Banach spaces and $F : X \rightrightarrows Y$. We say that F is *semi-differentiable* at $x \in X$ for $y \in F(x)$ if

$$\text{(22.1a)} \qquad \text{for every } \Delta y \in DF(x|y)(\Delta x) \quad \text{and} \quad x_k \to x,\ \tau_k \searrow 0 \quad \text{with} \quad \frac{x_k - x}{\tau_k} \to \Delta x$$

$$\text{(22.1b)} \qquad \text{there exist } y_k \in F(x_k) \quad \text{with} \quad \frac{y_k - y}{\tau_k} \to \Delta y.$$

In other words, $DF(x|y)$ is a full limit.

Lemma 22.1. *A mapping $F : X \rightrightarrows Y$ is semi-differentiable at $x \in X$ for $y \in Y$ if and only if*

$$\text{(22.2)} \qquad DF(x|y)(\Delta x) = \lim_{\tau \searrow 0,\, \Delta\tilde{x} \to \Delta x} \frac{F(x + \tau\Delta\tilde{x}) - y}{\tau} \qquad (\Delta x \in X).$$

Proof. First, note that (20.1) shows that $DF(x|y)(\Delta x)$ is the outer limit corresponding to (22.2). Similarly, by (22.1), F is semidifferentiable if $DF(x|y)$ is the corresponding inner limit. (For any sequence $\tau_k \searrow 0$, we can relate x_k in (22.1) and $\Delta\tilde{x} =: \Delta x_k$ in (22.2) via

$\Delta x_k = (x_k - x)/\tau_k$.) Hence, F is semidifferentiable if and only if the outer limit in (20.1) is a full limit. □

Compared to the definition of proto-differentiability in Section 21.1, we now require that Δy can be written as the limit of a difference quotient taken from $F(x_k)$ for *any* sequence $\{x_k\}$ similarly realizing Δx (while for proto-differentiability, this only has to be possible for *one* such sequence). Hence, semi-differentiability is a stronger property than proto-differentiability with the former implying the latter.

Example 22.2 (proto-differentiable but not semi-differentiable). Let $F : \mathbb{R} \rightrightarrows \mathbb{R}$ have graph $F = \mathbb{Q} \times \{0\}$. Then F is proto-differentiable at any $x \in \mathbb{Q}$ by the density of $\mathbb{Q}$ in $\mathbb{R}$. However, F is not semi-differentiable, as we can take $x_k \notin \mathbb{Q}$ in (22.1).

To characterize the semi-differentiability of the inverses of single-valued mappings, we require the next lemma. We say that $A \in \mathbb{L}(Y;X)$ has a *right-inverse* $A^\dagger \in \mathbb{L}(X;Y)$ if $AA^\dagger = \mathrm{Id}$. Then $A^* \in \mathbb{L}(Y^*;X^*)$ has the *left-inverse* $A^{\dagger *} \in \mathbb{L}(X^*;Y^*)$, i.e., $A^{\dagger *}A^* = \mathrm{Id}$.

Lemma 22.3. *On Banach spaces X and Y, suppose $F : X \to Y$ is continuously differentiable at x and $F'(x) \in \mathbb{L}(X;Y)$ has a right-inverse $F'(x)^\dagger \in \mathbb{L}(Y;X)$. For $P := \mathrm{Id} - F'(x)^\dagger F'(x)$, define*

$$\bar{F} : X \to Y \times \ker F'(x), \qquad \bar{F}(\tilde{x}) := (F(\tilde{x}), P\tilde{x}) \quad \text{for all } \tilde{x} \in X.$$

Then $\bar{F}$ is bijective in a neighborhood U of $\bar{F}(x)$ with a continuously differentiable inverse satisfying $\bar{F}^{-1}(\tilde{w}) \in F^{-1}(\tilde{y})$ for all $\tilde{w} = (\tilde{y}, \tilde{q}) \in U$ as well as

$$(\bar{F}^{-1})'(\bar{F}(x))(\Delta y, \Delta q) = F'(x)^\dagger \Delta y + \Delta q \quad \text{for all } (\Delta y, \Delta x) \in Y \times \ker F'(x). \tag{22.3}$$

Proof. Let $A := F'(x)$ and $A^\dagger := F'(x)^\dagger$. Then $P = \mathrm{Id} - A^\dagger A$ is a projection into $\ker A = \ker F'(x)$, in particular, $AP = 0$. We further define

$$M : Y \times \ker A \to X, \qquad M(\tilde{y}, \tilde{q}) := A^\dagger \tilde{y} + \tilde{q}, \quad \text{for all } \tilde{y} \in Y \text{ and } \tilde{q} \in \ker A.$$

Then for all $\Delta x \in X$,

$$M\bar{F}'(x)\Delta x = A^\dagger A \Delta x + P\Delta x = \Delta x.$$

Thus M is a left-inverse of $\bar{F}'(x)$, and consequently $\ker \bar{F}'(x) = \{0\}$. Since $\bar{F}(x)'\Delta x = (A\Delta x, P\Delta x)$ for all $\Delta x \in X$, similarly, for all $(\tilde{y}, \tilde{q}) \in Y \times \ker A$, we have

$$\bar{F}'(x)M(\tilde{y}, \tilde{q}) = (AA^\dagger \tilde{y} + A\tilde{q}, PA^\dagger \tilde{y} + P\tilde{q}) = (AA^\dagger \tilde{y}, P\tilde{q}) = (\tilde{y}, \tilde{q}),$$

which shows that M is also the right-inverse of $\bar{F}'(x)$ on $Y \times \ker F'(x)$. Hence $\bar{F}'(x)$ is bijective, $(\bar{F}^{-1})'(\bar{F}(x)) = M$, and the construction of M establishes (22.3).

By the inverse function Theorem 2.8, a continuosly differentiable $\bar{F}^{-1}$ exists in a neighborhood U of $w = (y, q) := \bar{F}(x)$ in $Y \times \ker A$ with $(\bar{F}^{-1})'(w) = M$ and $\bar{F}^{-1}(w) = x$. By construction, $\bar{F}^{-1}(\tilde{w}) \in F^{-1}(\tilde{y})$ for $\tilde{w} = (\tilde{y}, \tilde{q}) \in U$. □

Now, we have the following characterizations for the semi-differentiability of single-valued mappings and their inverses.

Lemma 22.4. *Let X, Y be Banach spaces and $F : X \to Y$.*

(i) If F is Fréchet differentiable at x, then F is semi-differentiable at x for $y = F(x)$.

(ii) If F is continuously differentiable at x and $F'(x) \in \mathbb{L}(X;Y)$ has a right-inverse $F'(x)^\dagger \in \mathbb{L}(Y;X)$, then $F^{-1} : Y \rightrightarrows X$ is semi-differentiable at $y = F(x)$ for x.

Proof. (i): This follows directly from the definition of semi-differentiability and the Fréchet derivative.

(ii): By Corollary 20.14, $DF^{-1}(y|x)(\Delta y) = \{\Delta x \in X \mid F'(x)\Delta x = \Delta y\}$ for $y = F(x)$. Hence (22.1) for F^{-1} requires showing that for all $\tau_k \searrow 0$ and $y_k \in Y$ with $(y_k - y)/\tau_k \to F'(x)\Delta x$, there exist x_k with $y_k = F(x_k)$ and $(x_k - x)/\tau_k \to \Delta x$. Let $\bar{F}$ be given by Lemma 22.3. Since $\bar{F}$ is invertible in a neighborhood of $\bar{F}(x) = (y, q) =: w$ for $q := Px \in \ker F'(x)$, let us take $x_k := \bar{F}^{-1}(y_k, q + \tau_k \Delta q)$ for $\Delta q := P\Delta x$. Then, by construction, $\bar{F}(x_k) = (F(x_k), Px_k) = (y_k, q + \tau_k \Delta q)$, Moreover

$$\lim_{k\to\infty} \frac{x_k - x}{\tau_k} = \lim_{k\to\infty} \frac{\bar{F}^{-1}(y_k, q + \tau_k \Delta q) - \bar{F}^{-1}(w)}{\tau_k} = (\bar{F}^{-1})'(w)(\Delta y, \Delta q).$$

By (22.3),

$$(\bar{F}^{-1})'(w)(\Delta y, \Delta q) = F'(x)^\dagger \Delta y + \Delta q = F'(x)^\dagger F'(x)\Delta x + (\mathrm{Id} - F'(x)^\dagger F'(x))\Delta x = \Delta x.$$

This finishes the proof. □

Remark 22.5. In Lemma 22.4 (ii), if X is finite-dimensional, it suffices to assume that F is continuously differentiable with $\ker F'(x)^* = \{0\}$. In this case we can take $F'(x)^{\dagger *} := A^*(AA^*)^{-1}$ for $A := F'(x)$.

22.2 CONE TRANSFORMATION FORMULAS

At their heart, calculus rules for (co)derivatives of set-valued mappings derive from corresponding transformation formulas for the underlying cones. To formulate these, let $C \subset Y$ and $R \in \mathbb{L}(Y;X)$. Take $x \in RC := \{Ry \mid y \in C\}$. We then say that there exists a *family of continuous inverse selections*

$$\{R_y^{-1} : U_y \to C \mid y \in C,\ Ry = x\}$$

of R to C at $x \in RC$ if for each $y \in C$ with $Ry = x$ there exists a neighborhood $U_y \subset RC$ of $x = Ry$ and a map $R_y^{-1} : U_y \to C$ continuous at x with $R_y^{-1}(x) = y$ and $RR_y^{-1}(\tilde{x}) = \tilde{x}$ for every $\tilde{x} \in U_x$ a neighborhood of x.

Example 22.6. Let $G : \mathbb{R}^{N-1} \to \mathbb{R}$ be continuous at x, and set $C := \operatorname{epi} G$ as well as $R(\tilde{x}, \tilde{t}) := \tilde{x}$. Then by the classical inverse function Theorem 2.8,

$$\{R^{-1}_{(x,t)}(\tilde{x}) := (\tilde{x}, t - G(x) + G(\tilde{x})) \mid t \geq G(x)\}$$

is a family of continuous inverse selections to C at x. If G is Fréchet differentiable at x, then so is $R^{-1}_{(t,x)}$.

Lemma 22.7. *Let X, Y be Banach spaces and assume there exists a family of continuous inverse selections $\{R_y^{-1} : U_y \to C \mid y \in C,\ Ry = x\}$ of $R \in \mathbb{L}(Y;X)$ to $C \subset Y$ at $x \in X$. If each R_y^{-1} is Fréchet differentiable at x, then*

$$T_{RC}(x) = \bigcup_{y:Ry=x} RT_C(y).$$

Proof. We first prove "$\supset$". Suppose $\Delta y \in T_C(y)$ for some $y \in \operatorname{cl} C$ with $Ry = x$. Then $\Delta y = \lim_{k\to\infty}(y_k - y)/\tau_k$ for some $y_k \in C$ and $\tau_k \searrow 0$. Consequently, since R is bounded, $R(y_k - y)/\tau_k \to R\Delta y$. But $Ry \in \operatorname{cl} RC$, so $R\Delta y \in T_{RC}(x)$. On the other hand, if $y \notin \operatorname{cl} C$, then $T_C(y) = \emptyset$ and thus there is nothing to show. Hence "$\supset$" holds.

To establish "$\subset$", we first of all note that $T_{RC}(x) = \emptyset$ if $x \notin \operatorname{cl} RC$. So suppose $x \in \operatorname{cl} RC$ and $\Delta x \in T_{RC}(x)$. Then $x = Ry$ for some $y \in \operatorname{cl} C$. Since $0 \in T_C(y)$, we can concentrate on $\Delta x \neq 0$. Then $\Delta x = \lim_{k\to\infty}(x_k - x)/\tau_k$ for some $x_k \in RC$ and $\tau_k \searrow 0$. We have $x_k = Ry_k$ for $y_k := R_y^{-1}(x_k)$. If we can show that $(y_k - y)/\tau_k \to \Delta y$ for some $\Delta y \in Y$, then $\Delta y \in T_C(y)$ and $\Delta x = R\Delta y$. Since R_y^{-1} is Fréchet differentiable at x, letting $h_k := x_k - x$ and using that $(h_k - \tau_k \Delta x)/\tau_k = (x_k - x)/\tau_k - \Delta x \to 0$ and $\|h_k\|_X/\tau_k \to \|\Delta x\|_X$, indeed

$$\begin{aligned}\lim_{k\to\infty}\left(\frac{y_k - y}{\tau_k} - (R_y^{-1})'(x)\Delta x\right) &= \lim_{k\to\infty}\frac{R_y^{-1}(x_k) - R_y^{-1}(x) - \tau_k (R_y^{-1})'(x)\Delta x}{\tau_k} \\ &= \lim_{k\to\infty}\frac{R_y^{-1}(x + h_k) - R_y^{-1}(x) - (R_y^{-1})'(x)h_k}{\tau_k} = 0.\end{aligned}$$

Thus $\Delta y = (R_y^{-1})'(x)\Delta x$, which proves "$\subset$". □

Remark 22.8 (qualification conditions in finite dimensions). If X and Y are finite-dimensional, we could replace the existence of the family of $\{R_y^{-1}\}$ of continuous selections in Lemma 22.7 by the more conventional *qualification condition*

$$\bigcup_{y:Ry=x} T_C(y) \cap \ker R = \{0\}.$$

We do not employ such a condition, as the extension to Banach spaces would have to be based not on $T_C(y)$ but on the weak tangent cone $T_C^w(y)$ that is difficult to compute explicitly.

We base all our calculus rules on the previous linear transformation lemma and the following composition lemma for the tangent cone T_C.

Lemma 22.9 (fundamental lemma on compositions). *Let X, Y, Z be Banach spaces and*

$$C := \{(x, y, z) \mid y \in F(x),\ z \in G(y)\}$$

for $F : X \rightrightarrows Y$, and $G : Y \rightrightarrows Z$. If $(x, y, z) \in C$ and either

(i) G is semi-differentiable at y for z, or

(ii) F^{-1} is semi-differentiable at y for x,

then

$$T_C(x, y, z) = \{(\Delta x, \Delta y, \Delta z) \mid \Delta y \in DF(x|y)(\Delta x),\ \Delta z \in DG(y|z)(\Delta y)\}. \tag{22.4}$$

Proof. We only consider the case (i); the case (ii) is shown analogously. By definition, we have $(\Delta x, \Delta y, \Delta z) \in T_C(x, y, z)$ if and only if for some $(x_k, y_k, z_k) \in C$ and $\tau_k \searrow 0$,

$$\Delta x = \lim_{k\to\infty} \frac{x_k - x}{\tau_k}, \qquad \Delta y = \lim_{k\to\infty} \frac{y_k - y}{\tau_k}, \qquad \Delta z = \lim_{k\to\infty} \frac{z_k - z}{\tau_k}.$$

On the other hand, we have $\Delta y \in DF(x|y)(\Delta x)$ if and only if the first two limits hold for some $(x_k, y_k) \in \operatorname{graph} F$ and $\tau_k \searrow 0$. Likewise, we have $\Delta z \in DG(y|z)(\Delta y)$ if and only if the last two limits hold for some $(y_k, z_k) \in \operatorname{graph} G$. This immediately yields "$\subset$".

To prove "$\supset$", take $\tau_k > 0$ and $(x_k, y_k) \in \operatorname{graph} F$ such that the first two limits hold. By the semi-differentiability of G at y for z, for any $\Delta z \in DG(y|z)(\Delta y)$ we can find $z_k \in G(y_k)$ such that $(z_k - z)/\tau_k \to \Delta z$. This shows the remaining limit. □

If one of the two mappings is single-valued, we can use Lemma 22.4 for verifying its semi-differentiability and Theorem 20.12 for the expression of its graphical derivative to obtain from Lemma 22.9 the following two special cases.

Corollary 22.10 (fundamental lemma on compositions: single-valued outer mapping). *Let X, Y, Z be Banach spaces and*

$$C := \{(x, y, G(y)) \mid y \in F(x)\}$$

for $F : X \rightrightarrows Y$ and $G : Y \to Z$. If $(x, y, z) \in C$ and G is Fréchet differentiable at y, then

$$T_C(x, y, z) = \{(\Delta x, \Delta y, G'(y)\Delta y) \mid \Delta y \in DF(x|y)(\Delta x)\}.$$

Corollary 22.11 (fundamental lemma on compositions: single-valued inner mapping). *Let X, Y, Z be Banach spaces and*

$$C := \{(x, y, z) \mid y = F(x),\ z \in G(y)\}$$

for $F : X \rightrightarrows Y$ and $G : Y \to Z$. If $(x, y, z) \in C$, F is continuously Fréchet differentiable at x and $F'(x)$ has a right-inverse $F'(x)^\dagger \in \mathbb{L}(Y; X)$, then

$$T_C(x, y, z) = \{(\Delta x, \Delta y, \Delta z) \mid \Delta y = F'(x)\Delta x,\ \Delta z \in DG(y|z)(\Delta y)\}.$$

22.3 CALCULUS RULES

Combining now the previous results, we quickly obtain various calculus rules. We begin as usual with a sum rule.

Theorem 22.12 (addition of a single-valued differentiable mapping). *Let X, Y be Banach spaces, let $G : X \to Y$ be Fréchet differentiable, and $F : X \rightrightarrows Y$. Then for any $x \in X$ and $y \in H(x) := F(x) + G(x)$,*

$$DH(x|y)(\Delta x) = DF(x|y - G(x))(\Delta x) + G'(x)\Delta x \qquad (\Delta x \in X).$$

Proof. We have graph $H = RC$ for

$$C := \{(u, \tilde{x}, G(\tilde{x})) \mid \tilde{x} \in X,\ u \in F(\tilde{x})\} \quad \text{and} \quad R(u, \tilde{x}, v) := (\tilde{x}, u + v). \tag{22.5}$$

We now use Lemma 22.7 to calculate T_{RC}. Accordingly, for all $(u, \tilde{x}, G(\tilde{x})) \in C$ such that $R(u, \tilde{x}, G(\tilde{x})) = (x, y)$ – i.e., only for $\tilde{x} = x$ and $u = y - G(x)$ – we define the inverse selection

$$R^{-1}_{(u,x,G(x))} : RC \to C, \qquad R^{-1}_{(u,x,G(x))}(\tilde{x}, \tilde{y}) := (\tilde{y} - G(\tilde{x}), \tilde{x}, G(\tilde{x})), \tag{22.6}$$

Then $R^{-1}_{(u,x,G(x))}(x, u + G(x)) = (u, x, G(x))$ and $R^{-1}_{(u,x,G(x))}(\tilde{x}, \tilde{y}) \in C$ for every $(\tilde{x}, \tilde{y}) \in RC$. Furthermore, $R^{-1}_{(u,x,G(x))}$ is continuous and Fréchet differentiable at (x, z).

Lemma 22.7 now yields

$$T_{\operatorname{graph} H}(x, y) = \{(\Delta x, \Delta u + \Delta v) \mid (\Delta u, \Delta x, \Delta v) \in T_C(y - G(x), x, G(x))\}.$$

Moreover, C given in (22.5) coincides with the C defined in Corollary 22.10 with F^{-1} in place of F. Thus, using the corollary and inserting the expression from Lemma 20.5 for DF^{-1} into the result, it follows

$$T_C(u, x, v) = \{(\Delta u, \Delta x, G'(x)\Delta x) \mid \Delta u \in DF(x|u)(\Delta x)\}.$$

Thus

$$\begin{aligned} DH(x|y)(\Delta x) &= \{\Delta u + \Delta v \mid (\Delta u, \Delta x, \Delta v) \in T_C(y - G(x), x, G(x))\} \\ &= \{\Delta u + G'(x)\Delta x \mid \Delta u \in DF(x|y - G(x))(\Delta x)\}, \end{aligned}$$

which yields the claim. □

We now turn to chain rules, beginning with the case that the outer mapping is single-valued.

Theorem 22.13 (outer composition with a single-valued differentiable mapping). *Let X, Y be Banach spaces, $F : X \rightrightarrows Y$, and $G : Y \to Z$. Let $x \in X$ and $z \in H(x) := G(F(x))$ be given. If G is Fréchet differentiable at every $y \in F(x)$, left-invertible on* ran G *near z, and the left-inverse G^{-1} is Fréchet differentiable at z, then*

$$DH(x|z)(\Delta x) = \bigcup_{y:G(y)=z} G'(y)DF(x|y)(\Delta x) \qquad (\Delta x \in X).$$

Proof. Observing that graph $H = RC$ for

$$C := \{(\tilde{x}, \tilde{y}, G(\tilde{y})) \mid \tilde{y} \in F(\tilde{x})\} \quad \text{and} \quad R(\tilde{x}, \tilde{y}, \tilde{z}) := (\tilde{x}, \tilde{z}), \tag{22.7}$$

we again use Lemma 22.7 to calculate T_{RC}. Accordingly, we define for $y \in G^{-1}(z) \cap F(x)$ the family of inverse selections

$$R^{-1}_{(x,y,z)} : RC \to C, \qquad R^{-1}_{(x,y,z)}(\tilde{x}, \tilde{z}) := (\tilde{x}, G^{-1}(\tilde{z}), \tilde{z}). \tag{22.8}$$

Clearly, $R^{-1}_{(x,y,z)}(x, z) = (x, y, z)$. Furthermore, G is by assumption invertible on its range near $z = G(y)$. Hence $G^{-1}(\tilde{z}) \in F(\tilde{x})$, and thus in fact $R^{-1}_{(x,y,z)}(\tilde{x}, \tilde{z}) \in C$ for all $(\tilde{x}, \tilde{z}) \in RC$. Moreover, since G^{-1} has the same properties at z, $R^{-1}_{(x,y,z)}$ is at (x, z) continuous, Fréchet differentiable, and locally Lipschitz with a factor independent of y.

We are therefore justified in applying Lemma 22.7, which yields

$$T_{\operatorname{graph} H}(x, z) = \bigcup_{y:G(y)=z} \{(\Delta x, \Delta z) \mid (\Delta x, \Delta y, \Delta z) \in T_C(x, y, z)\}.$$

Using Corollary 22.10, we then obtain

$$\begin{aligned} DH(x|z)(\Delta x) &= \bigcup_{y:G(y)=z} \{\Delta z \mid (\Delta x, \Delta y, \Delta z) \in T_C(x, y, z)\} \\ &= \bigcup_{y:G(y)=z} \{G'(y)\Delta y \mid \Delta y \in DF(x|y)(\Delta x)\}. \end{aligned}$$

After further simplification, we arrive at the claimed expression. □

In particular, this result holds if G is Fréchet differentiable and $G'(y)$ is bijective, since in this case the inverse function Theorem 2.8 guarantees the local existence and differentiability of G^{-1}.

Another useful special case is when the mapping G is linear.

Corollary 22.14 (outer composition with a linear operator). *Let X, Y, Z be Banach spaces, $A \in \mathbb{L}(Y; Z)$, and $F : X \rightrightarrows Y$. If A has a bounded left-inverse $A^\dagger$, then for any $x \in X$ and $z \in H(x) := AF(x)$,*

$$DH(x|z)(\Delta x) = ADF(x|y)(\Delta x) \qquad (\Delta x \in X)$$

for the unique $y \in Y$ such that $Ay = z$.

Proof. We apply Theorem 22.13 to $G(y) := Ay$, which is clearly continuously differentiable at every $y \in F(x)$. Since A has a bounded left-inverse $A^\dagger$, $G^{-1}(y) = A^\dagger y$ is an inverse of G on $G(y) = \operatorname{ran} A$, which is also clearly differentiable. Moreover, $\{y \mid G(y) = z\}$ is a singleton, which removes the intersections and unions from the expressions provided by Theorem 22.13. □

The assumption of left-invertibility is in particular satisfied if Y and Z are Hilbert spaces and A is injective and has closed range, since in this case we can take $A^\dagger = (A^*A)^{-1}A^*y$ (the *Moore–Penrose pseudoinverse* of A) and $A^{\dagger *} = (A^\dagger)^*$.

We next consider chain rules where the inner mapping is single-valued.

Theorem 22.15 (inner composition with a single-valued differentiable mapping). *Let X, Y, Z be Banach spaces, $F : X \to Y$ and $G : Y \rightrightarrows Z$. Let $x \in X$ and $z \in H(x) := G(F(x))$. If F is continuously Fréchet differentiable near x and $F'(x)$ has a right-inverse $F'(x)^\dagger \in \mathbb{L}(Y;X)$, then*

$$DH(x|z)(\Delta x) = DG(F(x)|z)(F'(x)\Delta x) \qquad (\Delta x \in X).$$

Proof. Observing that $\operatorname{graph} H = RC$ for

$$C := \{(\tilde{x}, \tilde{y}, \tilde{z}) \mid \tilde{y} = F(\tilde{x}),\ \tilde{z} \in G(\tilde{y})\} \quad \text{and} \quad R(\tilde{x}, \tilde{y}, \tilde{z}) := (\tilde{x}, \tilde{z}), \tag{22.9}$$

we again use Lemma 22.7 to compute T_{RC}. Accordingly, we define a family of inverse selections for all $(\tilde{x}, \tilde{y}, \tilde{z}) \in C$ such that $R(\tilde{x}, \tilde{y}, \tilde{z}) = (x, z)$. The latter only holds for $(\tilde{x}, \tilde{y}, \tilde{z}) = (x, F(x), z)$, and hence we only need

$$R^{-1}_{(x,F(x),z)} : RC \to C, \qquad R^{-1}_{(x,F(x),z)}(\tilde{x}, \tilde{z}) := (\tilde{x}, F(\tilde{x}), \tilde{z}).$$

Clearly $R^{-1}_{(x,F(x),z)}(x, z) = (x, F(x), z)$ and $R^{-1}_{(x,F(x),z)}(\tilde{x}, \tilde{z}) \in C$ for $(\tilde{x}, \tilde{z}) \in RC$. Moreover, $R^{-1}_{(x,F(x),z)}$ is continuous and Fréchet differentiable at (x, z).

Thus Lemma 22.7 yields

$$T_{\operatorname{graph} H}(x, z) = \{(\Delta x, \Delta z) \mid (\Delta x, \Delta y, \Delta z) \in T_C(x, F(x), z)\}.$$

On the other hand, due to the continuous differentiability of F and the right-invertibility of $F'(x)$, we can apply Corollary 22.11 to obtain

$$T_C(x, y, z) = \{(\Delta x, \Delta y, \Delta z) \mid \Delta y = F'(x)\Delta x,\ \Delta z \in DG(y|z)(\Delta y)\}.$$

Thus

$$\begin{aligned} DH(x|z)(\Delta x) &= \{\Delta z \mid (\Delta x, \Delta y, \Delta z) \in T_C(x, F(x), z)\} \\ &= \{\Delta z \mid \Delta y = F'(x)\Delta x,\ \Delta z \in DG(F(x)|z)(\Delta y)\}, \end{aligned}$$

which yields the claim. □

Again, we can specialize this result to the case where the single-valued mapping is linear.

Corollary 22.16 (inner composition with a linear operator). *Let X, Y, Z be Banach spaces, $A \in \mathbb{L}(X;Y)$, and $G : Y \rightrightarrows Z$. Let $H := G \circ A$ for $A \in \mathbb{L}(X;Y)$ and $G : Y \rightrightarrows Z$ on Banach spaces X, Y, and Z. If A has a right-inverse $A^\dagger \in \mathbb{L}(Y;X)$, then for all $x \in X$ and $z \in H(x) := G(Ax)$,*

$$DH(x|z)(\Delta x) = DG(Ax|z)(A\Delta x) \qquad (\Delta x \in X).$$

We wish to apply these results to further differentiate the chain rules from Theorems 4.17 and 13.23. For the former, this is straight-forward based on the two corollaries so far obtained.

Corollary 22.17 (second derivative chain rule for convex subdifferential). *Let X, Y be Banach spaces, let $f : Y \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $A \in \mathbb{L}(X;Y)$ be such that A has a right-inverse $A^\dagger \in \mathbb{L}(Y;X)$, and $\operatorname{ran} A \cap \operatorname{int} \operatorname{dom} f \neq \emptyset$. Let $h := f \circ A$. Then for any $x \in X$ and $x^* \in \partial h(x) = A^* \partial f(Ax)$,*

$$D[\partial h](x|x^*)(\Delta x) = A^* D[\partial f](Ax|y^*)(A\Delta x) \qquad (\Delta x \in X)$$

for the unique $y^* \in Y^*$ satisfying $A^* y^* = x^*$.

Proof. The expression for $\partial h(x)$ follows from Theorem 4.17, to which we apply Corollary 22.16 as well as Corollary 22.14 with A^* in place of A, recalling that a right-inverse $A^\dagger$ of A produces a left-inverse $A^{\dagger *}$ of A^*. □

To further differentiate the result of applying a chain rule such as Theorem 13.23, we also need a product rule for a single-valued mapping G and a set-valued mapping F. In principle, this could be obtained as a composition of $x \mapsto (x_1, x_2)$, $(x_1, x_2) \mapsto \{G(x_1)\} \times F(x_2)$, and $(y_1, y_2) \mapsto y_1 y_2$; however, the last one of these mappings does not possess the left-inverse required by Theorem 22.13. We therefore take another route, which starts with the following lemma.

Lemma 22.18. *Let X and Y be Banach spaces, and $F : X \rightrightarrows Y$. Define $\bar{F} : X \rightrightarrows X \times Y$ by $\bar{F}(x) := \{x\} \times F(x)$. Then, for all $x, \Delta x \in X$ and $y \in F(x)$, we have*

$$D\bar{F}(x|x, y)(\Delta x) = \{\Delta x\} \times DF(x|y)(\Delta x).$$

Proof. We have

$$\operatorname{graph} \bar{F} = R_0 \operatorname{graph} F \quad \text{for} \quad R_0(\tilde{x}, \tilde{y}) := (\tilde{x}, (\tilde{x}, \tilde{y})).$$

Let now $y \in F(x)$. Clearly $R_{0,v}^{-1}(\tilde{x},(\tilde{x},\tilde{y})) := (\tilde{x},\tilde{y})$, $R_{0,v}^{-1} : R_0 \operatorname{graph} F \to \operatorname{graph} F$ is a Fréchet differentiable inverse selection of R_0 at $(x,(x,y)) \in R_0 \operatorname{graph} F$ for the unique $v = (x,y) \in \operatorname{graph} F$ with $R_0 v = (x,(x,y))$. Therefore, by Lemma 22.7, we have

$$T_{R_0 \operatorname{graph} F}(x,(x,y)) = \{(\Delta x,(\Delta x,\Delta y)) \mid (\Delta x,\Delta y) \in T_{\operatorname{graph} F}(x,y)\},$$

which establishes the claim. □

Theorem 22.19 (product rule). *Let X, Y, Z be Banach spaces, let $G : X \to \mathbb{L}(Y;Z)$ be Fréchet differentiable, and $F : X \rightrightarrows Y$. If $G(\tilde{x}) \in \mathbb{L}(Y;Z)$ has a left-inverse $G(\tilde{x})^\dagger \in \mathbb{L}(Z;Y)$ for $\tilde{x}$ near $x \in X$ and the mapping $\tilde{x} \mapsto G(\tilde{x})^\dagger$ is Fréchet differentiable at x, then for all $z \in H(x) := G(x)F(x) := \bigcup_{y \in F(x)} G(x)y$,*

$$DH(x|z)(\Delta x) = [G'(x)\Delta x]y + G(x)DF(x|y)\Delta x \quad (z \in H(x),\ \Delta x \in X)$$

for the unique $y \in F(x)$ satisfying $G(x)y = z$.

Proof. Let $\bar{F}$ be as in Lemma 22.18. Then $\operatorname{graph} H = R_1 \operatorname{graph}(\bar{G} \circ \bar{F})$ for

$$\bar{G}(\tilde{x},\tilde{y}) = (\tilde{x}, G(\tilde{x})\tilde{y}) \quad \text{and} \quad R_1(\tilde{x}_1,\tilde{x}_2,\tilde{z}) := (\tilde{x}_1,\tilde{z}),$$

where

$$\begin{aligned}\operatorname{graph}(\bar{G} \circ \bar{F}) &= \{(x,x,G(x)y) \mid (x,(x,y)) \in \operatorname{graph} \bar{F}\} \\ &= \{(x,x,G(x)y) \mid x \in X,\ y \in F(x)\}.\end{aligned}$$

We now wish to apply Theorem 22.13 on $\bar{G} \circ \bar{F}$. First, $\bar{G}$ is single-valued and differentiable. Since $G(\tilde{x})$ is assumed left-invertible for $\tilde{x}$ near x, the mapping $Q : (\tilde{x},\tilde{z}) \mapsto (\tilde{x}, G(\tilde{x})^\dagger \tilde{z})$ is a left-inverse of $\bar{G}$, which is Fréchet differentiable at (x,z) since $\tilde{x} \to G(\tilde{x})^\dagger$ is Fréchet differentiable at x. Finally, we also have

$$\bar{G}'(x,y)(\Delta x,\Delta y) = (\Delta x, [G'(x)\Delta x]y + G(x)\Delta y).$$

Thus Theorem 22.13 and Lemma 22.18 yield

$$\begin{aligned}D[\bar{G} \circ \bar{F}](x|x,z)(\Delta x) &= \bigcup_{y:\bar{G}(x,y)=(x,z)} \bar{G}'(x,y)D\bar{F}(x|x,y)(\Delta x) \\ &= \bigcup_{y:G(x)y=z} \bar{G}'(x,y)(\Delta x, DF(x|y)\Delta x) \\ &= \bigcup_{y:G(x)y=z} \{\Delta x\} \times ([G'(x)\Delta x]y + G(x)DF(x|y)\Delta x).\end{aligned}$$

It follows that

$$T_{\operatorname{graph}(\bar{G}\circ\bar{F})}(x,x,z) = \bigcup_{y:G(x)y=z} \{(\Delta x,\Delta x,\Delta z) \mid \Delta z \in ([G'(x)\Delta x]y + G(x)DF(x|y)\Delta x)\}.$$

Observe then that $R_{1,w}^{-1}(\tilde{x}_1, \tilde{z}) := (\tilde{x}_1, (\tilde{x}_1, \tilde{z}))$, $R_{1,w}^{-1} : R_1 \operatorname{graph}(\bar{G} \circ \bar{F}) \to \operatorname{graph}(\bar{G} \circ \bar{F})$ is a Fréchet differentiable inverse selection of R_1 at $(x, G(x)y) \in R_0 \operatorname{graph}(\bar{G} \circ \bar{F})$ for the unique $w = (x, (x, G(x)y)) \in \operatorname{graph}(\bar{G} \circ \bar{F})$ with $R_1 w = (x, G(x)y)$. Therefore, another application of Lemma 22.7 yields

$$T_{\operatorname{graph} H}(x, z) = \bigcup_{y : G(x)y = z} \{(\Delta x, \Delta z) \mid \Delta z \in ([G'(x)\Delta x]y + G(x)DF(x|y)\Delta x)\}.$$

Since the y is unique by our invertibility assumptions on $G(x)$ and exists due to $z \in H(x)$, we obtain the claim. □

Corollary 22.20 (second derivative chain rule for Clarke subdifferential). *Let X, Y be Banach spaces, let $f : Y \to R$ be locally Lipschitz continuous, and let $S : X \to Y$ be twice continuously differentiable. Set $h : X \to Y$, $h(x) := f(S(x))$. If there exists a neighborhood U of $x \in X$ such that*

(i) f is Clarke regular at $S(\tilde{x})$ for all $\tilde{x} \in U$;

(ii) $S'(\tilde{x})$ has a right-inverse $S'(\tilde{x})^\dagger \in \mathbb{L}(Y; X)$ for all $\tilde{x} \in U$;

(iii) the mapping $\tilde{x} \mapsto S'(\tilde{x})^{\dagger *}$ is Fréchet differentiable at x;

then for all $x^* \in \partial_C h(x) = S'(x)^* \partial_C f(S(x))$,

$$D[\partial_C h](x|x^*)(\Delta x) = (S''(x)\Delta x)^* y^* + S'(x)^* D[\partial_C f](S(x)|y^*)(S'(x)\Delta x) \quad (\Delta x \in X)$$

for the unique $y^* \in \partial_C f(S(x))$ such that $S'(x)^* y^* = x^*$.

Proof. The expression for $\partial_C h(x)$ follows from Theorem 13.23. Let now $\tilde{S} : X \to \mathbb{L}(Y^*; X^*)$, $\tilde{S}(\tilde{x}) := S'(\tilde{x})^*$. Then $\tilde{S}$ is Fréchet differentiable at x, and has the left-inverse $S'(\tilde{x})^{*\dagger}$ for all $\tilde{x} \in U$. Together with assumption (iii) this allows us to apply Theorem 22.19 to obtain

$$D[\partial_C h](x|x^*)(\Delta x) = (\tilde{S}'(x)\Delta x)y^* + S'(x)^* D[(\partial_C f) \circ S](x|x^*)(\Delta x) \quad (\Delta \tilde{x} \in X).$$

Furthermore, since $S'(x)$ has a bounded right-inverse, we can apply Theorem 22.15 to obtain for all $x \in U$ and all $x^* \in \partial_C f(S(x))$

$$D[(\partial_C f) \circ S](x|x^*)(\Delta x) = D[\partial_C f](S(x)|y^*)(S'(x)\Delta x) \qquad (\Delta x \in X)$$

for the unique $y^* \in \partial_C f(S(x))$ such that $S'(x)^* y^* = x^*$. Finally, since the adjoint mapping $A \mapsto A^*$ is linear and an isometry, it is straightforward to verify using the definition that $\tilde{S}'(x)\Delta x = (S''(x)\Delta x)^*$, which yields the claim. □

23 CALCULUS FOR THE FRÉCHET CODERIVATIVE

We continue with calculus rules for the Fréchet coderivative. As in Chapter 22, we start with the relevant regularity concept, then prove the fundamental lemmas, and finally derive the calculus rules.

23.1 SEMI-CODIFFERENTIABILITY

Let X, Y be Banach spaces. We say that F is *semi-codifferentiable* at $x \in X$ for $y \in F(x)$ if for each $y^* \in Y^*$ there exists some $x^* \in \widehat{D}^*F(x|y)(y^*)$ satisfying

$$\lim_{\operatorname{graph} F \ni (x_k, y_k) \to (x,y)} \frac{\langle x^*, x_k - x\rangle_X - \langle y^*, y_k - y\rangle_Y}{\|(x_k - x, y_k - y)\|_{X\times Y}} = 0. \tag{23.1}$$

Recalling (18.7), this is equivalent to requiring that $-x^* \in \widehat{D}^*F(x|y)(-y^*)$ as well. For single-valued mappings and their inverses we have the following characterization.

Lemma 23.1. *Let X, Y be Banach spaces and let $F : X \to Y$ be single-valued. If F is Fréchet differentiable at $x \in X$, then*

(i) F is semi-codifferentiable at x for $y = F(x)$.

If, moreover, $F'(x) \in \mathbb{L}(X;Y)$ has a left-inverse $F'(x)^\dagger \in \mathbb{L}(Y;X)$, then

(ii) F^{-1} is semi-codifferentiable at $y = F(x)$ for x.

Proof. Recalling from Theorem 20.12 that $\widehat{D}^*F(x|y)(y^*) = \{F'(x)^*y^*\}$ when $y = F(x)$, the claim (i) follows immediately from the observation above that semi-codifferentiability is equivalent to the existence for all y^* of $x^* \in \widehat{D}^*F(x|y)(y^*)$ such that $-x^* \in \widehat{D}^*F(x|y)(-y^*)$ as well.

As for (ii), recalling the inverse relationships of Lemma 20.5 and again using Theorem 20.12, we have $\widehat{D}^*F^{-1}(y|x)(x^*) = \{y^* \mid x^* = F'(x)^*y^*\}$. Moreover, we recall that for a left-inverse $F'(x)^\dagger$ of $F'(x)$, the operator $F'(x)^{\dagger *}$ is a right-inverse of $F'(x)^*$. Thus, for any x^*, we have $y^* := F'(x)^{\dagger *}x^* \in \widehat{D}^*F^{-1}(y|x)(x^*)$, and, by linearity, $-y^* \in \widehat{D}^*F^{-1}(y|x)(-x^*)$. Hence F^{-1} is semi-codifferentiable at y for x. □

23.2 CONE TRANSFORMATION FORMULAS

In the following, we consider more general ε-normal cones for $\varepsilon \geq 0$ as these results will be needed later in Chapter 25 for proving the corresponding expressions for the limiting normal cone. We refer to Section 22.2 for the definition of a family of continuous inverse selections.

Lemma 23.2. *Let X, Y be Banach spaces, and assume there exists a family of continuous inverse selections $\{R_y^{-1} : U_y \to C \mid y \in C,\ Ry = x\}$ of $R \in \mathbb{L}(Y;X)$ to $C \subset Y$ at $x \in X$. If each R_y^{-1} for all $y \in C$ with $Ry = x$ is locally Lipschitz at x with the factor L_x, then for all $\varepsilon \geq 0$,*

$$\widehat{N}_{RC}^{\varepsilon/\|R\|_{\mathbb{L}(Y;X)}}(x) \subset \bigcap_{y \in C : Ry = x} \{x^* \in X^* \mid R^*x^* \in \widehat{N}_C^{\varepsilon}(y)\} \subset \widehat{N}_{RC}^{\varepsilon L_x}(x). \tag{23.2}$$

In particular,

$$\widehat{N}_{RC}(x) = \bigcap_{y \in C : Ry = x} \{x^* \in X^* \mid R^*x^* \in \widehat{N}_C(y)\}.$$

Proof. By (18.7), $x^* \in \widehat{N}_{RC}^{\tilde{\varepsilon}}(x)$ for a given $\tilde{\varepsilon} > 0$ if and only if

$$\limsup_{RC \ni Ry_k \to Ry} \frac{\langle R^*x^*, y_k - y\rangle_Y}{\|R(y_k - y)\|_X} \leq \tilde{\varepsilon} \tag{23.3}$$

for *all* $y \in C$ such that $Ry = x$. The specific choice of y is inconsequential; if the expression holds for one, it holds for all. Since the expression inside the limit is invariant under perturbations $\tilde{y} \in \ker R$, the limit can be further be equivalently be written in terms of $C \ni y_k \to y$. Thus we see that $x^* \in \widehat{N}_{RC}^{\tilde{\varepsilon}}(x)$ if and only if for every $\varepsilon' > \tilde{\varepsilon}$ and all y with $Ry = x$, there exists a $\delta > 0$ such that

$$\langle R^*x^*, y_k - y\rangle_Y \leq \varepsilon' \|R(y_k - y)\|_X \quad (y_k \in \mathbb{B}(y, \delta) \cap C). \tag{23.4}$$

Similarly, $R^*x^* \in \widehat{N}_C^{\varepsilon}(y)$ if and only if

$$\limsup_{C \ni y_k \to y} \frac{\langle R^*x^*, y_k - y\rangle_Y}{\|y_k - y\|_Y} \leq \varepsilon. \tag{23.5}$$

Now if $x^* \in \widehat{N}_{RC}^{\tilde{\varepsilon}}(x)$, then using (23.4) and estimating $\|R(y_k - y)\|_X \leq \|R\|_{\mathbb{L}(X;Y)}\|y_k - y\|_Y$ yields (23.5) for $\varepsilon = \varepsilon'\|R\|_{\mathbb{L}(X;Y)}$ for all $\varepsilon' > \tilde{\varepsilon}$. Hence the first inclusion in (23.2) holds by taking $\tilde{\varepsilon} = \varepsilon/\|R\|_{\mathbb{L}(X;Y)}$ and letting $\varepsilon' \searrow \tilde{\varepsilon}$

For the second inclusion, we first observe that equivalently to (23.3), $x^* \in \widehat{N}_{RC}^{\tilde{\varepsilon}}(x)$ if and only if

$$\limsup_{RC \ni x_k \to x} \frac{\langle R^*x^*, R_y^{-1}(x_k) - R_y^{-1}(x)\rangle_X}{\|x_k - x\|_X} \leq \tilde{\varepsilon} \tag{23.6}$$

for all $y \in C$ with $Ry = x$. Let $y \in C$ with $Ry = x$ be arbitrary, and $R^*x^* \in \widehat{N}_C^\varepsilon(y)$. Then (23.5) holds for $y_k = R_y^{-1}(x_k)$ for any $U_y \ni x_k \to x$. By the local Lipschitz assumption, we have $L_x \geq \limsup_{k\to\infty} \|R_y^{-1}(x_k) - R_y^{-1}(x)\|_Y / \|x_k - x\|_X$. Applying these choices and estimates in (23.5), we obtain (23.6) for $\tilde{\varepsilon} = \varepsilon L_x$. Since $y \in C$ with $Ry = x$ was arbitrary, we conclude that $x^* \in \widehat{N}_{RC}^{\tilde{\varepsilon}}(x)$, which yields the second inclusion in (23.2). □

Remark 23.3 (polarity and qualification condition in finite dimensions). In finite dimensions, Lemma 23.2 for $\varepsilon = 0$ could also be proved with the help of the polarity relationships $\widehat{N}_{RC}(x) = T_{RC}(x)^\circ$ and $\widehat{N}_C(y) = T_C(y)^\circ$ from Lemma 18.10. Furthermore, the existence of a family of continuous selections could be replaced by a qualification condition as in Remark 22.8.

We are now ready to prove the fundamental composition lemma, this time for the Fréchet normal cone. We say that $G : Y \rightrightarrows Z$ is *inner Lipschitz at y for z* if for some $L, \delta > 0$ and all $\tilde{y} \in B(\delta, y)$ we have

$$\inf_{\tilde{z}\in G(\tilde{y})} \|\tilde{z} - z\| \leq L\|\tilde{y} - y\|.$$

For single-valued mappings, this property obviously reduces to the Lipschitz-continuity at y. We return to futher, different, Lipschitz-like properties of set-valued mappings in Chapter 27.

Lemma 23.4 (fundamental lemma on compositions). *Let X, Y, Z be Banach spaces and*

$$C := \{(x, y, z) \mid y \in F(x),\ z \in G(y)\}$$

for $F : X \rightrightarrows Y$, and $G : Y \rightrightarrows Z$. Let $(x, y, z) \in C$.

(i) If G is semi-codifferentiable and inner Lipschitz at y for z with factor $L > 0$, then

$$K_\varepsilon \subset \widehat{N}_C^\varepsilon(x, y, z) \subset K_{L\varepsilon}$$

for all $\varepsilon \geq 0$ and

$$K_\varepsilon := \left\{(x^*, y^*, z^*) \;\middle|\; \begin{array}{l} x^* \in \widehat{D}_\varepsilon^* F(x|y)(-\tilde{y}^* - y^*), \\ \tilde{y}^* \in \widehat{D}^* G(y|z)(z^*),\ -\tilde{y}^* \in \widehat{D}^* G(y|z)(-z^*) \end{array}\right\}.$$

(ii) If F^{-1} is semi-codifferentiable and inner Lipschitz at y for x with factor $\ell > 0$, then

$$Q_\varepsilon \subset \widehat{N}_C^\varepsilon(x, y, z) \subset Q_{\ell\varepsilon}$$

for all $\varepsilon \geq 0$ and

$$Q_\varepsilon := \left\{(x^*, y^*, z^*) \;\middle|\; \begin{array}{l} x^* \in \widehat{D}^* F(x|y)(-\tilde{y}^* - y^*),\ -x^* \in \widehat{D}^* F(x|y)(\tilde{y}^* + y^*), \\ -\tilde{y}^* \in \widehat{D}_\varepsilon^* G(y|z)(-z^*) \end{array}\right\}.$$

Proof. We recall that $(x^*, y^*, z^*) \in \widehat{N}^\varepsilon_C(x, y, z)$ if and only if

$$\limsup_{C \ni (x_k, y_k, z_k) \to (x,y,z)} \frac{\langle x^*, x_k - x\rangle_X + \langle y^*, y_k - y\rangle_Y + \langle z^*, z_k - z\rangle_Z}{\|(x_k, y_k, z_k) - (x, y, z)\|_{X\times Y\times Z}} \le \varepsilon. \tag{23.7}$$

In case (i), the semi-codifferentiability of G implies that for some $\tilde{y}^* \in \widehat{D}^*G(y|z)(z^*)$ we have $-\tilde{y}^* \in \widehat{D}^*G(y|z)(-z^*)$ or equivalently

$$\lim_{\operatorname{graph} G \ni (y_k, z_k) \to (y,z)} \frac{\langle \tilde{y}^*, y_k - y\rangle_Y - \langle z^*, z_k - z\rangle_Z}{\|(y_k, z_k) - (y, z)\|_{Y\times Z}} = 0.$$

Thus (23.7) holds if and only if for some $\tilde{y}^* \in \widehat{D}^*G(y|z)(z^*)$ with $-\tilde{y}^* \in \widehat{D}^*G(y|z)(-z^*)$ we have

$$\limsup_{C \ni (x_k, y_k, z_k) \to (x,y,z)} \frac{\langle x^*, x_k - x\rangle_X + \langle \tilde{y}^* + y^*, y_k - y\rangle_Y}{\|(x_k, y_k, z_k) - (x, y, z)\|_{X\times Y\times Z}} \le \varepsilon. \tag{23.8}$$

But this follows from $x^* \in \widehat{D}^*_\varepsilon F(x|y)(-\tilde{y}^* - y^*)$, which yields the first inclusion in (i). For the second inclusion, taking large enough k, we use the inner Lipschitz assumption to choose $z_k \in G(y_k)$ such that $\|z_k - z\| \le (L + 1/k)\|y_k - y\|$. Then (23.8) implies

$$\limsup_{\operatorname{graph} F \ni (x_k, y_k) \to (x,y,z)} \frac{\langle x^*, x_k - x\rangle_X + \langle \tilde{y}^* + y^*, y_k - y\rangle_Y}{\|(x_k, y_k) - (x, y)\|_{X\times Y}} \le L\varepsilon,$$

which is to say $x^* \in \widehat{D}^*_{L\varepsilon} F(x|y)(-\tilde{y}^* - y^*)$.

In case (ii), the semi-codifferentiability of F^{-1} implies that there exists $\tilde{y}^* \in -y^* + \widehat{D}^*F^{-1}(y|x)(-x^*)$, i.e., satisfying $x^* \in \widehat{D}^*F(x|y)(-\tilde{y}^* - y^*)$, such that also $-x^* \in \widehat{D}^*F(x|y)(\tilde{y}^* + y^*)$. This is again equivalently written as

$$\lim_{\operatorname{graph} G \ni (y_k, x_k) \to (y,x)} \frac{-\langle \tilde{y}^* + y^*, y_k - y\rangle_Y - \langle x^*, x_k - x\rangle_X}{\|(y_k, x_k) - (y, x)\|_{Y\times X}} = 0.$$

Thus (23.7) holds if and only if for some $\tilde{y}^*$ we have both $x^* \in \widehat{D}^*F(x|y)(-\tilde{y}^* - y^*)$ and $-x^* \in \widehat{D}^*F(x|y)(\tilde{y}^* + y^*)$, as well as

$$\limsup_{C \ni (x_k, y_k, z_k) \to (x,y,z)} \frac{\langle z^*, z_k - z\rangle_X - \langle \tilde{y}^*, y_k - y\rangle_Y}{\|(x_k, y_k, z_k) - (x, y, z)\|_{X\times Y\times Z}} \le \varepsilon.$$

But this follows from $-\tilde{y}^* \in \widehat{D}^*_\varepsilon G(y|z)(-z^*)$, which yields the first inclusion in (ii). For the second inclusion, as in case (i), we use the inner Lipschitz assumption. □

For the remaining results, we fix $\varepsilon = 0$. If one of the two mappings is single-valued, Lemma 23.4 yields the following two special cases.

Corollary 23.5 (fundamental lemma on compositions: single-valued outer mapping). *Let X, Y, Z be Banach spaces and*

$$C := \{(x, y, G(y)) \mid y \in F(x)\}$$

for $F : X \rightrightarrows Y$ and $G : Y \to Z$. If $(x, y, z) \in C$ and G is Fréchet differentiable at y, then

$$\widehat{N}_C(x, y, z) = \{(x^*, y^*, z^*) \mid x^* \in \widehat{D}^*F(x|y)(-[G'(y)]^*z^* - y^*),\ y^* \in Y^*\}.$$

Proof. We first use Lemma 23.1 (i) to show the semi-codifferentiability of G at y for z. The assumed Fréchet differentiability at y implies that G is Lipschitz and hence inner Lipschitz at y for $z = G(y)$. Thus we may apply Lemma 23.4 (i) to get an expression for $\widehat{N}_C(x, y, z)$. We finish by inserting therein the expression given by Theorem 20.12 for $\widehat{D}^*G(y|z)(z^*)$. □

The corresponding result for a single-valued inner mapping is not quite as straightforward, unless we assume full (left and right) invertibility of $F'(x)^{-1}$. We first do so, and then relax the assumption to mere right invertibility.

Corollary 23.6 (initial lemma on compositions: single-valued inner mapping). *Let X, Y, Z be Banach spaces and*

$$C := \{(x, y, z) \mid y = F(x),\ z \in G(y)\}$$

for $F : X \to Y$ and $G : Y \rightrightarrows Z$. If $(x, y, z) \in C$, F is continuously Fréchet differentiable at x and $F'(x)$ has an inverse $F'(x)^{-1} \in \mathbb{L}(Y; X)$, then

$$\widehat{N}_C(x, y, z) = \{(F'(x)^*(-\tilde{y}^* - y^*), y^*, z^*) \mid -\tilde{y}^* \in \widehat{D}^*G(y|z)(-z^*);\ \tilde{y}^*, y^* \in Y^*\}.$$

Proof. Similarly to the previous proof, we apply Theorem 20.12 and Lemma 23.1 (ii) to F to prove its semi-codifferentiability and then use Lemma 23.4 (ii). To prove that F^{-1} is inner Lipschitz at $y = F(x)$ for x, we apply the Inverse Function Theorem 2.8, which shows that F^{-1} exists and is continuously differentiable. Then Lemma 2.11 shows that F^{-1} is locally Lipschitz at y, which implies that F^{-1} is inner Lipschitz, as required. □

Lemma 23.7 (fundamental lemma on compositions: single-valued inner mapping). *Let X, Y, Z be Banach spaces and*

$$C := \{(x, y, z) \mid y = F(x),\ z \in G(y)\}$$

for $F : X \to Y$ and $G : Y \rightrightarrows Z$. If $(x, y, z) \in C$, the mapping F is continuously Fréchet differentiable at x, and $F'(x)$ has a right inverse $F'(x)^\dagger \in \mathbb{L}(Y; X)$, then

$$\widehat{N}_C(x, y, z) = \{(F'(x)^*(-\tilde{y}^* - y^*), y^*, z^*) \mid -\tilde{y}^* \in \widehat{D}^*G(y|z)(-z^*);\ \tilde{y}^*, y^* \in Y^*\}.$$

Proof. Let $\bar{F} : X \to Y \times \ker F'(x)$, $\bar{F}(x) := (F(x), Px)$ for $P := \mathrm{Id} - F'(x)^\dagger F'(x)$. Also let $\bar{G} : Y \times \ker F'(x) \rightrightarrows Z$ be defined by $G(y,q) := G(y)$. Then, by Lemma 22.3, $\bar{F}$ is invertible, and by either the proof of the lemma or by the Inverse Function Theorem 2.8, $\bar{F}'(x)$ has an inverse $\bar{F}'(x)^{-1} \in \mathbb{L}(Y \times \ker F'(x); X)$. Directly from the definition, we deduce that for every $z^* \in Z^*$,

$$\widehat{D}^*\bar{G}(y|z)(z^*) = \{(z^*, q^*) \in Z^* \times [\ker F'(x)]^* \mid z^* \in \widehat{D}^*\bar{G}(y|z)(z^*),\ q^* = 0\}.$$

We apply Corollary 23.6 to $\bar{F}$ and $\bar{G}$ to obtain for

$$\bar{C} := \{(x, (y,q), z) \mid (y,q) = \bar{F}(x),\ z \in \bar{G}(y,q)\}$$

the normal expression

$$\begin{aligned}
&\widehat{N}_{\bar{C}}(x,(y,q),z) \\
&\quad = \left\{ (\bar{F}'(x)^*(-\tilde{y}^* - y^*, -\tilde{q}^* - q^*), (y^*, q^*), z^*) \;\middle|\; \begin{array}{l} -(\tilde{y}^*, \tilde{q}^*) \in \widehat{D}^*\bar{G}(y|z)(-z^*); \\ \tilde{y}^*, y^* \in Y^*;\ \tilde{q}^*, q^* \in [\ker F'(x)]^* \end{array} \right\} \\
&\quad = \left\{ (F'(x)^*(-\tilde{y}^* - y^*) + P^*(-\tilde{q}^* - q^*), (y^*, q^*), z^*) \;\middle|\; \begin{array}{l} -\tilde{y}^* \in \widehat{D}^*G(y|z)(-z^*); \\ \tilde{y}^*, y^* \in Y^*; \\ q^* \in [\ker F'(x)]^*;\ \tilde{q}^* = 0 \end{array} \right\}.
\end{aligned}$$

Now we write $C = R\bar{C}$ for $R(\tilde{x}, (\tilde{y}, \tilde{q}), \tilde{z}) := (\tilde{x}, \tilde{y}, \tilde{z})$, and observe that $R^{-1}_{(x,(y,q),z)}(\tilde{x}, \tilde{y}, \tilde{z}) := (\tilde{x}, (\tilde{y}, P\tilde{x}), \tilde{z})$ for $q = Px$ is a Lipschitz inverse selection of R at $(x, (y,q), z)$ to $\bar{C}$. Therefore Lemma 23.2 establishes

$$\begin{aligned}
\widehat{N}_C(x,y,z) &= \bigcap_{(\tilde{x},(\tilde{y},\tilde{q}),\tilde{z}) \in \bar{C}, R(\tilde{x},(\tilde{y},\tilde{q}),\tilde{z}) = (x,y,z)} \{(x^*, y^*, z^*) \mid (x^*, (y^*, 0), z^*) \in \widehat{N}_{\bar{C}}(\tilde{x}, (\tilde{y}, \tilde{q}), \tilde{z})\} \\
&= \{(F'(x)^*(-\tilde{y}^* - y^*), y^*, z^*) \mid -\tilde{y}^* \in \widehat{D}^*G(y|z)(-z^*),\ y^* \in Y^*\},
\end{aligned}$$

as claimed. □

23.3 CALCULUS RULES

Using the above lemmas, we again obtain calculus rules. The proofs are similar to those in Section 22.3, and we only note the differences.

Theorem 23.8 (addition of a single-valued differentiable mapping). *Let X, Y be Banach spaces, $G : X \to Y$ be continuously Fréchet differentiable, and $F : X \rightrightarrows Y$. Then for any $x \in X$ and $y \in H(x) := F(x) + G(x)$,*

$$\widehat{D}^*H(x|y)(y^*) = \widehat{D}^*F(x|y - G(x))(y^*) + [G'(x)]^* y^* \qquad (y^* \in Y^*).$$

Proof. We have graph $H = RC$ for C and R given by (22.5) in the proof of Theorem 22.12. Since G is continuously Fréchet differentiable, it is locally Lipschitz by Lemma 2.11. As shown in Theorem 22.12, the set of $v \in C$ with $Rv = (x, y)$ is a singleton. Consequently the map R_v given by (22.6) for the unique v is locally Lipschitz with a factor $L_{(x,y)}$. We may therefore apply Lemma 23.2 in place of Lemma 22.7 in the proof of Theorem 22.12 to obtain

$$\widehat{N}_{\operatorname{graph} H}(x, y) = \{(x^*, y^*) \mid (y^*, x^*, y^*) \in \widehat{N}_C(y - G(x), x, G(x))\}.$$

Moreover, C given in (22.5) coincides with the C defined in Corollary 23.5 with F^{-1} in place of F. Inserting the expression from Lemma 20.5 for $\widehat{D}^*F^{-1}$ into the result of the corollary, it follows that

$$\begin{aligned}\widehat{N}_C(u, x, v) &= \{(u^*, x^*, v^*) \in Y^* \times X^* \times Y^* \mid u^* \in \widehat{D}^*F^{-1}(u|x)(-[G'(y)]^*v^* - x^*)\} \\ &= \{(u^*, x^*, v^*) \in Y^* \times X^* \times Y^* \mid [G'(x)]^*v^* + x^* \in \widehat{D}^*F(x|u)(-u^*)\}.\end{aligned}$$

Thus

$$\begin{aligned}\widehat{D}^*H(x|y)(y^*) &= \{x^* \mid (-y^*, x^*, -y^*) \in \widehat{N}_C(y - G(x), x, G(x))\} \\ &= \{x^* \mid -[G'(x)]^*y^* + x^* \in \widehat{D}^*F(x|y - G(x))(y^*)\},\end{aligned}$$

which yields the claim. □

Theorem 23.9 (outer composition with a single-valued differentiable mapping). *Let X, Y be Banach spaces, $F : X \rightrightarrows Y$, and $G : Y \to Z$. Let $x \in X$ and $z \in H(x) := G(F(X))$ be given. If G is Fréchet differentiable at every $y \in F(x)$, left-invertible on* $\operatorname{ran} G$ *near z, and the inverse G^{-1} is continuously Fréchet differentiable in a neighborhood z, then*

$$\widehat{D}^*H(x|z)(z^*) = \bigcap_{y:G(y)=z} \widehat{D}^*F(x|y)([G'(y)]^*z^*) \qquad (z^* \in Z^*).$$

Proof. We have graph $H = RC$ for R and C as given by (22.7) in the proof of Theorem 22.13. Using the assumed continuous Fréchet differentiability of G^{-1} at z, Lemma 2.11 establishes that G^{-1} is Lipschitz at z. Consequently, so are the selections $R^{-1}_{(x,y,z)} : RC \to C$ constructed in (22.8) Applying Lemma 23.2 in place of Lemma 22.7 then yields

$$\widehat{N}_{\operatorname{graph} H}(x, z) = \bigcap_{y:G(y)=z} \{(x^*, z^*) \mid (x^*, 0, z^*) \in \widehat{N}_C(x, y, z)\}.$$

Corollary 23.5 then shows that

$$\begin{aligned}\widehat{D}^*H(x|z)(z^*) &= \bigcap_{y:G(y)=z} \{x^* \mid (x^*, 0, -z^*) \in \widehat{N}_C(x, y, z)\} \\ &= \bigcap_{y:G(y)=z} \{x^* \mid x^* \in \widehat{D}^*F(x|y)([G'(y)]^*z^*)\}.\end{aligned}$$

After further simplification, we arrive at the claimed expression. □

Corollary 23.10 (outer composition with a linear operator). *Let X, Y, Z be Banach spaces, $A \in \mathbb{L}(Y;Z)$, and $F : X \rightrightarrows Y$. If A has a bounded left-inverse $A^\dagger$, then for any $x \in X$ and $z \in H(x) := AF(x)$,*

$$\widehat{D}^*H(x|z)(z^*) = \widehat{D}^*F(x|y)(A^*z^*) \qquad (z^* \in Z^*)$$

for the unique $y \in Y$ such that $Ay = z$.

Proof. We only need to verify that $G(y) := Az$ satisfies the assumptions of Theorem 23.9, which can be done exactly as in the proof of Corollary 22.14. □

Theorem 23.11 (inner composition with a single-valued mapping). *Let X, Y, Z be Banach spaces, $F : X \to Y$ and $G : Y \rightrightarrows Z$. Let $x \in X$ and $z \in H(x) := G(F(x))$. If F is continuously Fréchet differentiable in a neighborhood of x and $F'(x)$ has a right-inverse $F'(x)^\dagger \in \mathbb{L}(Y;X)$, then*

$$\widehat{D}^*H(x|z)(z^*) = [F'(x)]^*\widehat{D}^*G(F(x)|z)(z^*) \qquad (z^* \in Z^*).$$

Proof. We have graph $H = RC$ for C and R as given by (22.9) in the proof of Theorem 22.15. Similarly to the proof of continuity and Fréchet differentiability therein but now using the continuous Fréchet differentiability assumption on F and Lemma 2.11, we observe that R_v^{-1} is locally Lipschitz at (x,z) for the unique $v = (x, F(x), z) \in C$ with $Rv = (x,z)$. We are therefore justified in applying Lemma 23.2 in place of Theorem 22.15. It yields

$$\widehat{N}_{\operatorname{graph} H}(x,z) = \{(x^*, z^*) \mid (x^*, 0, z^*) \in \widehat{N}_C(x, F(x), z)\}.$$

On the other hand, since F is continuously Fréchet differentiable, Lemma 23.7 implies that

$$\widehat{N}_C(x,y,z) = \{(F'(x)^*(-\tilde{y}^* - y^*), y^*, z^*) \mid -\tilde{y}^* \in \widehat{D}^*G(y|z)(-z^*),\ y^* \in Y^*\}.$$

Thus

$$\begin{aligned}\widehat{D}^*H(x|z)(z^*) &= \{x^* \mid (x^*, 0, -z^*) \in \widehat{N}_C(x, F(x), z)\}\\ &= \{F'(x)^*\tilde{y}^* \mid \tilde{y}^* \in \widehat{D}^*G(y|z)(z^*)\},\end{aligned}$$

which yields the claim. □

Corollary 23.12 (inner composition with a linear operator). *Let X, Y, Z be Banach spaces, $A \in \mathbb{L}(X;Y)$, and $G : Y \rightrightarrows Z$. Let $H := G \circ A$ for $A \in \mathbb{L}(X;Y)$ and $G : Y \rightrightarrows Z$ on Banach spaces X, Y, and Z. If A has a right-inverse $A^\dagger \in \mathbb{L}(Y;X)$, then for all $x \in X$ and $z \in H(x) := G(Ax)$,*

$$\widehat{D}^*H(x|z)(z^*) = A^*\widehat{D}^*G(Ax|z)(z^*) \qquad (z^* \in Z^*).$$

We again apply this to the chain rule from Theorem 4.17. Compare the following expression with that from Corollary 22.17, noting that $\partial f : X \rightrightarrows X^*$ in Banach spaces such that $\widehat{D}^*[\partial f](x|x^*) : X^{**} \to X^*$.

Corollary 23.13 (second derivative chain rule for convex subdifferential). *Let X, Y be Banach spaces, $f : Y \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $A \in \mathbb{L}(X;Y)$ be such that A has a right-inverse $A^\dagger \in \mathbb{L}(Y;X)$, and $\operatorname{ran} A \cap \operatorname{int} \operatorname{dom} f \neq \emptyset$. Let $h := f \circ A$. Then for any $x \in X$ and $x^* \in \partial h(x) = A^* \partial f(Ax)$,*

$$\widehat{D}^*[\partial h](x|x^*)(x^{**}) = A^* \widehat{D}^*[\partial f](Ax|y^*)(A^{**}x^{**}) \qquad (x^{**} \in X^{**})$$

for the unique $y^* \in Y^*$ satisfying $A^* y^* = x^*$.

Proof. The expression for $\partial h(x)$ follows from Theorem 4.17, to which we apply Corollary 23.12 as well as Corollary 23.10 with A^* in place of A, recalling that a right-inverse $A^\dagger$ for A produces the left-inverse $A^{\dagger *}$ for A^*. □

Hence if X is reflexive, the expression for the coderivative is identical to that for the graphical derivative.

For the corresponding result for the Clarke subdifferential, we again need a product rule. We start with the following lemma.

Lemma 23.14. *Let X, Y be Banach spaces and $F : X \rightrightarrows Y$. Define $\bar{F} : X \rightrightarrows X \times Y$ by $\bar{F}(x) := \{x\} \times F(x)$. Then, for all $x \in X$, $y \in F(x)$, $x^* \in X^*$, and $y^* \in Y^*$, we have*

$$\widehat{D}^* \bar{F}(x|x, y)(x^*, y^*) = \widehat{D}^* F(x|y)(y^*) + x^*.$$

Proof. The proof is analogous to Lemma 22.18 for the graphical derivative. We have

$$\operatorname{graph} \bar{F} = R_0 \operatorname{graph} F \quad \text{for} \quad R_0(\tilde{x}, \tilde{y}) := (\tilde{x}, (\tilde{x}, \tilde{y})).$$

Clearly $R_{0,v}^{-1}(\tilde{x}, (\tilde{x}, \tilde{y})) := (\tilde{x}, \tilde{y})$, $R_{0,v}^{-1} : R_0 \operatorname{graph} F \to \operatorname{graph} F$, is a Fréchet differentiable and Lipschitz inverse selection of R_0 at $(x, (x, y)) \in R_0 \operatorname{graph} F$ for the unique $v = (x, y) \in \operatorname{graph} F$ with $R_0 v = (x, (x, y))$. Therefore, by Lemma 23.2, we have

$$\widehat{N}_{R_0 \operatorname{graph} F}(x, (x, y)) = \{(x_0^*, (-x^*, -y^*)) \mid (x_0^* - x^*, -y^*) \in \widehat{N}_{\operatorname{graph} F}(x, y)\},$$

which establishes the claim. □

Theorem 23.15 (product rule). *Let X, Y, Z be Banach spaces, $G : X \to \mathbb{L}(Y;Z)$ be Fréchet differentiable, and $F : X \rightrightarrows Y$. If $G(\tilde{x}) \in \mathbb{L}(Y;Z)$ has a left-inverse $G(\tilde{x})^\dagger \in \mathbb{L}(Z;Y)$ for $\tilde{x}$ near $x \in X$ and the mapping $\tilde{x} \mapsto G(\tilde{x})^\dagger$ is continuously Fréchet differentiable in a neighborhood of x, then for all $z \in H(x) := G(x)F(x) := \bigcup_{y \in F(x)} G(x)y$,*

$$\widehat{D}^* H(x|z)(z^*) = \widehat{D}^* F(x|y)(G(x)^* z^*) + ([G'(x) \cdot]y)^* z^* \qquad (z^* \in Z^*)$$

for the unique $y \in F(x)$ satisfying that $G(x)y = z$.

Proof. The proof is analogous to Theorem 22.19 for the graphical derivative. We again have $\operatorname{graph} H = R_1 \operatorname{graph}(\bar{G} \circ \bar{F})$ for $\bar{F}$ as in Lemma 23.14,

$$\bar{G}(\tilde{x}, \tilde{y}) = (\tilde{x}, G(\tilde{x})\tilde{y}), \quad \text{and} \quad R_1(\tilde{x}_1, \tilde{x}_2, \tilde{z}) := (\tilde{x}_1, \tilde{z}).$$

We also have

$$\bar{G}'(x, y)^*(x_0^*, z^*) = (x_0^* + ([G'(x)\cdot]y)^* z^*, G(x)^* z^*).$$

We now apply Theorem 23.9, whose remaining assumptions are verified exactly as those of Theorem 22.13 in Theorem 22.19, only now using the *continuous* Fréchet differentiability of $\tilde{x} \mapsto G(\tilde{x})^\dagger$. This combined with Lemma 23.14 yields

$$\begin{aligned}
\widehat{D}^*[\bar{G} \circ \bar{F}](x|x, z)(x_0^*, z^*) &= \bigcap_{y:\bar{G}(x,y)=(x,z)} \widehat{D}^*\bar{F}(x|x, y)(\bar{G}'(x, y)^*(x_0^*, z^*)) \\
&= \bigcap_{y:G(x)y=z} \widehat{D}^*\bar{F}(x|x, y)(x_0^* + ([G'(x)\cdot]y)^* z^*, G(x)^* z^*) \\
&= \bigcap_{y:G(x)y=z} x_0^* + \widehat{D}^*F(x|y)(G(x)^* z^*) + ([G'(x)\cdot]y)^* z^*.
\end{aligned}$$

It follows that

$$\widehat{N}_{\operatorname{graph}(\bar{G}\circ\bar{F})}(x, x, z) = \bigcap_{y:G(x)y=z} \left\{ (x^*, -x_0^*, -z^*) \;\middle|\; \begin{array}{rl} x^* - x_0^* & \in \widehat{D}^*F(x|y)(G(x)^* z^*) \\ & \quad + ([G'(x)\cdot]y)^* z^* \end{array} \right\}.$$

Write $w := (x, (x, z)) \in \operatorname{graph}(\bar{G} \circ \bar{F})$. Since the inverse selection $R_{1,w}^{-1}$ constructed in Theorem 22.19 is linear, it is Lipschitz. As w is the unique point in $\operatorname{graph}(\bar{G} \circ \bar{F})$ with $R_1 w = (x, z)$, the entire family of inverse selections corresponding to $(x, z) \in R_1 \operatorname{graph}(\bar{G}\circ\bar{F})$ has a uniform Lipschitz factor. Therefore, another application of Lemma 23.2 yields

$$\widehat{N}_{\operatorname{graph} H}(x, z) = \bigcap_{y:G(x)y=z} \{(x^*, -z^*) \mid x^* \in \widehat{D}^*F(x|y)(G(x)^* z^*) + ([G'(x)\cdot]y)^* z^*\}.$$

Since the y is unique by our invertibility assumptions on $G(x)$ and exists due to $z \in H(x)$, we obtain the claim. □

Corollary 23.16 (second derivative chain rule for Clarke subdifferential). *Let X, Y be Banach spaces, let $f : Y \to R$ be locally Lipschitz continuous, and let $S : X \to Y$ be twice continuously differentiable. Set $h : X \to Y$, $h(x) := f(S(x))$. If there exists a neighborhood U of $x \in X$ such that*

(i) f is Clarke regular at $S(\tilde{x})$ for all $\tilde{x} \in X$;

(ii) $S'(\tilde{x})$ has a right-inverse $S'(\tilde{x})^\dagger \in \mathbb{L}(Y; X)$ for all $\tilde{x} \in U$;

(iii) the mapping $\tilde{x} \mapsto S'(\tilde{x})^{\dagger *}$ is continuously Fréchet differentiable at x;

then for all $x^* \in \partial_C h(x) = S'(x)^* \partial_C f(S(x))$,

$$\widehat{D}^*[\partial_C h](x|x^*)(x^{**}) = \hat{S}(x)^* x^{**} + S'(x)^* \widehat{D}^*[\partial_C f](S(x)|y^*)(S'(x)^{**} x^{**}) \quad (x^{**} \in X^{**})$$

for the linear operator $\hat{S} : X \to \mathbb{L}(X; X^*)$, $\hat{S}(x)\Delta x := (S''(x)\Delta x)^* y^*$ *and the unique* $y^* \in \partial_C f(S(x))$ *such that* $S'(x)^* y^* = x^*$.

Proof. The expression for $\partial_C h(x)$ follows from Theorem 13.23. Let now $\tilde{S} : X \to \mathbb{L}(Y^*; X^*)$, $\tilde{S}(x) := S'(\tilde{x})^*$. Then $\tilde{S}$ is Fréchet differentiable in U as well, which together with assumption (iii) allows us to apply Theorem 23.15 to obtain

$$\widehat{D}^*[\partial_C h](x|x^*)(x^{**}) = (\tilde{S}'(x)y^*)^* x^{**} + \widehat{D}^*[(\partial_C f) \circ S](x|y^*)(S'(x)^{**} x^{**}) \quad (x^{**} \in X^{**}).$$

Furthermore, since $S'(x)$ has a bounded right-inverse, we can apply Theorem 23.11 to obtain for all $x^* \in \partial_C f(S(x))$ that

$$\widehat{D}^*[(\partial_C f) \circ S](x|y^*)(y^{**}) = S'(x)^* \widehat{D}^*[\partial_C f](S(x)|y^*)(y^{**}) \qquad (y^{**} \in Y^{**})$$

for the unique $y^* \in \partial_C f(S(x))$ such that $S'(x)^* y^* = x^*$. The claim now follows again from the fact that $\tilde{S}'(x)\Delta x = (S''(x)\Delta x)^*$. □

Note that $\hat{S}(x)\Delta x := (S''(x)\Delta x)^* y^*$ also occurs in the corresponding Corollary 22.20 and recall from Examples 20.1 and 20.6 and Theorem 20.12 that coderivatives for differentiable single-valued mappings amount to taking adjoints of their Fréchet derivative.

24 CALCULUS FOR THE CLARKE GRAPHICAL DERIVATIVE

We now turn to the limiting (co)derivatives. Compared to the basic (co)derivatives, calculus rules for these are much more challenging and require even more assumptions. In this chapter, we consider the Clarke graphical derivative, where in addition to strict differentiability we will for the sake of simplicity assume T-regularity of the set-valued mapping (so that the Clarke graphical derivative coincides with the graphical derivative) and show that this regularity is preserved under addition and composition with a single-valued mapping.

24.1 STRICT DIFFERENTIABILITY

The following concept generalizes the notion of strict differentiability for single-valued mappings (see Remark 2.6) to set-valued mappings. Let X, Y be Banach spaces. We say that $F : X \rightrightarrows Y$ is *strictly differentiable* at $x \in X$ for $y \in F(x)$ if $\operatorname{graph} F$ is closed near (x, y) and

$$\text{(24.1a)}\qquad \text{for every } \Delta y \in \widehat{D}F(x|y)(\Delta x),\quad \tau_k \searrow 0,\quad \tilde{x}_k \to x \quad \text{with} \quad \frac{x_k - \tilde{x}_k}{\tau_k} \to \Delta x,$$
$$\text{and} \quad \tilde{y}_k \in F(\tilde{x}_k) \quad \text{with} \quad \tilde{y}_k \to y,$$

$$\text{(24.1b)}\qquad \text{there exist } y_k \in F(x_k) \quad \text{with} \quad \frac{y_k - \tilde{y}_k}{\tau_k} \to \Delta y.$$

Compared to semi-differentiability, strict differentiability requires that the limits realizing the various directions are interchangeable with limits of the base points; in other words, that the graphical derivative is itself an inner limit, i.e.,

$$\text{(24.2)}\qquad \widehat{D}F(x|y)(\Delta x) = \liminf_{\substack{\tau \searrow 0,\ \Delta\tilde{x} \to \Delta x \\ \operatorname{graph} F \ni (\tilde{x}, \tilde{y}) \to (x, y)}} \frac{F(\tilde{x} + \tau\Delta\tilde{x}) - \tilde{y}}{\tau} \qquad (\Delta x \in X).$$

Lemma 24.1. *If X and Y are finite-dimensional, then $F : X \rightrightarrows Y$ is strictly differentiable at $x \in X$ for $y \in Y$ if and only if*

$$\text{(24.3)}\qquad \widehat{D}F(x|y)(\Delta x) = \liminf_{\substack{\operatorname{graph} F \ni (\tilde{x}, \tilde{y}) \to (x, y), \\ \Delta\tilde{x} \to \Delta x,\ DF(\tilde{x}|\tilde{y})(\Delta\tilde{x}) \neq \emptyset}} DF(\tilde{x}|\tilde{y})(\Delta\tilde{x}) \qquad (\Delta x \in X).$$

Proof. We need to show that graph $\widehat{D}F(x,y) = K$ for

$$K := \left\{ (\Delta x, \Delta y) \;\middle|\; \Delta y \in \liminf_{\substack{\operatorname{graph} F \ni (\tilde{x},\tilde{y}) \to (x,y),\\ \Delta\tilde{x} \to \Delta x,\, DF(\tilde{x}|\tilde{y})(\Delta\tilde{x}) \neq \emptyset}} DF(\tilde{x}|\tilde{y})(\Delta\tilde{x}) \right\}.$$

We first show that $\operatorname{graph} \widehat{D}F(x,y) \subset K$. If $(\Delta x, \Delta y) \notin K$, then there exist $\operatorname{graph} F \ni (\tilde{x}_k, \tilde{y}_k) \to (x,y)$ and $\Delta x_k \to \Delta x$ with $DF(\tilde{x}_k|\tilde{y}_k)(\Delta x_k) \neq \emptyset$ such that for some $\varepsilon > 0$ and an infinite subset $N \subset \mathbb{N}$,

$$\inf_{\Delta y_k \in DF(\tilde{x}_k|\tilde{y}_k)(\Delta x_k)} \|\Delta y_k - \Delta y\| \geq 2\varepsilon \qquad (k \in N).$$

By the characterization (20.1) of $DF(\tilde{x}_k|\tilde{y}_k)$, this implies the existence of $\tau_k \searrow 0$ such that

$$\limsup_{k\to\infty} \inf_{y_k \in F(x_k + \tau_k \Delta x_k)} \left\| \frac{y_k - \tilde{y}_k}{\tau_k} - \Delta y \right\| \geq \varepsilon.$$

Thus $(\Delta x, \Delta y) \notin \operatorname{graph} \widehat{D}F(x,y)$ and hence $\operatorname{graph} \widehat{D}F(x,y) \subset K$.

Rewriting then

$$K = \left\{ (\Delta x, \Delta y) \in X \times Y \;\middle|\; \begin{array}{r} (\tilde{x}, \tilde{y}, \Delta\tilde{x}) \to (x, y, \Delta x) \Rightarrow \exists\, \Delta\tilde{y} \to \Delta y \\ \text{with } \Delta\tilde{y} \in DF(\tilde{x}|\tilde{y})(\Delta\tilde{x}) \end{array} \right\},$$

the characterization (20.4) of $\widehat{D}F(x,y)$ provides the opposite inclusion $\operatorname{graph} \widehat{D}F(x,y) \subset K$. Therefore (24.3) holds. □

In particular, single-valued continuously differentiable mappings and their inverses are strictly differentiable.

Lemma 24.2. *Let X, Y be Banach spaces and let $F : X \to Y$ be single-valued.*

(i) If F is continuously differentiable at $x \in X$, then F is strictly differentiable at x for $y = F(x)$.

(ii) If F is continuously differentiable near $x \in X$ and $F'(x)$ has a right-inverse $F'(x)^\dagger \in \mathbb{L}(Y;X)$, then F^{-1} is strictly differentiable at $y = F(x)$ for x.

Proof. The proof is analogous to Lemma 22.4, since the inverse function Theorem 2.8 establishes the continuous differentiability of $\bar{F}^{-1}$ and hence strict differentiability. □

Remark 24.3. As in Remark 22.5, if X is finite-dimensional, it suffices in Lemma 24.2 (ii) to assume that F is continuously differentiable with $\ker F'(x)^* = \{0\}$.

24.2 CONE TRANSFORMATION FORMULAS

The main aim in the following lemmas is to show that tangential regularity is preserved under certain transformations. We do this by proceeding as in Section 22.2 to derive explicit expressions for the transformed cones and then comparing them with the corresponding expressions obtained there for the graphical derivative.

Lemma 24.4. *Let X, Y be Banach spaces and assume there exists a family of continuous inverse selections $\{R_y^{-1} : U_y \to C \mid y \in C,\ Ry = x\}$ of $R \in \mathbb{L}(Y;X)$ to $C \subset Y$ at $x \in X$. If each R_y^{-1} is Fréchet differentiable at x and C is tangentially regular at all $y \in C$ with $Ry = x$, then RC is tangentially regular at x and*

$$\widehat{T}_{RC}(x) = \bigcup_{y:Ry=x} R\widehat{T}_C(y).$$

Proof. We first prove "$\subset$". Suppose $\Delta y \in \widehat{T}_C(y)$ for some $y \in Y$ with $Ry = x$. Then for any $C \ni \tilde{y}_k \to y$ there exist $y_k \in C$ and $\tau_k \searrow 0$ such that $\Delta y = \lim_{k\to\infty}(y_k - \tilde{y}_k)/\tau_k$. Consequently, since R is bounded, $R(y_k - \tilde{y}_k)/\tau_k \to R\Delta y$. To show that $R\Delta y \in \widehat{T}_{RC}(x)$, let $RC \ni \bar{x}_k \to x$ be given. Take now $\tilde{y}_k = R_y^{-1}(\bar{x}_k)$, which satisfies $\tilde{y}_k \to y = R_y^{-1}(x)$ due to $\bar{x}_k \to x$. Then $(Ry_k - \bar{x}_k)/\tau_k = R(y_k - \tilde{y}_k)/\tau_k \to R\Delta y$, which shows "$\supset$".

To prove "$\subset$", suppose that $\Delta x \in \widehat{T}_{RC}(x)$ and hence $\Delta x \in T_{RC}(x)$ by Theorem 18.5. By Lemma 22.7, $\Delta x = R\Delta y$ for some $y \in Y$ with $Ry = x$ and $\Delta y \in T_C(y) = \widehat{T}_C(y)$ by the assumed tangential regularity of C at y. This shows "$\subset$".

Comparing now the expression for $\widehat{T}_{RC}(y) = T_C(y)$ with the expression for $T_{RC}(x)$ provided by Lemma 22.7 and using the tangential regularity of C shows the claimed tangential regularity of RC. □

Remark 24.5 (regularity assumptions). The assumption in Lemma 24.4 that C is tangentially regular is not needed if $\ker R = \{0\}$ or, more generally, if R is a continuously differentiable mapping with $\ker \nabla R(y) = \{0\}$; see [Mordukhovich, 1994, Corollary 5.4].

Lemma 24.6 (fundamental lemma on compositions). *Let X, Y, Z be Banach spaces and*

$$C := \{(x, y, z) \mid y \in F(x),\ z \in G(y)\}$$

for $F : X \rightrightarrows Y$, and $G : Y \rightrightarrows Z$. If $(x, y, z) \in C$ and either

(a) G is inner semicontinuous, strictly differentiable, and T-regular at y for z, or

(b) F^{-1} is inner semicontinuous, strictly differentiable, and T-regular at y for x,

then

$$\widehat{T}_C(x,y,z) = \{(\Delta x, \Delta y, \Delta z) \mid \Delta y \in \widehat{D}F(x|y)(\Delta x),\ \Delta z \in \widehat{D}G(y|z)(\Delta y)\}. \tag{24.4}$$

Moreover, if F is T-regular at x for y and G is T-regular at y for z, then C is tangentially regular at (x,y,z).

Proof. We only consider the case (a) as the case (b) is again proved similarly. The proof is analogous to Lemma 22.9, using in this case the strict differentiability of G in place of semi-differentiability. First, we observe that $(\Delta x, \Delta y, \Delta z) \in \widehat{T}_C(x,y,z)$ if and only if for all $\tau_k \searrow 0$ and $C \ni (\tilde{x}_k, \tilde{y}_k, \tilde{z}_k) \to (x,y,z)$, there exist $(x_k, y_k, z_k) \in C$ such that

$$\Delta x = \lim_{k\to\infty} \frac{x_k - \tilde{x}_k}{\tau_k}, \qquad \Delta y = \lim_{k\to\infty} \frac{y_k - \tilde{y}_k}{\tau_k}, \qquad \Delta z = \lim_{k\to\infty} \frac{z_k - \tilde{z}_k}{\tau_k}.$$

Suppose $(\Delta x, \Delta y, \Delta z) \in \widehat{T}_C(x,y,z)$. Taking $(\tilde{x}_k, \tilde{y}_k, \tilde{z}_k) = (x,y,z)$, it is immediate that $\Delta y \in DF(x|y)(\Delta x)$ and $\Delta z \in DG(y|z)(\Delta y)$. By the T-regularity of G, it follows that $\Delta z \in \widehat{D}G(y|z)(\Delta y)$. Now take any graph $F \ni (\tilde{x}_k, \tilde{y}_k) \to (x,y)$. By the assumption that G is inner semicontinuous, there exists some $G(\tilde{y}_k) \ni \tilde{z}_k \to z$. Thus, by the above characterization of $(\Delta x, \Delta y, \Delta z) \in \widehat{T}_C(x,y,z)$, there exist $(x_k, y_k) \in \operatorname{graph} F$ such that $(x_k - \tilde{x}_k)/\tau_k \to \Delta x$ and $(y_k - \tilde{y}_k)/\tau_k \to \Delta y$. That is, $(\Delta x, \Delta y) \in \widehat{T}_{\operatorname{graph} F}(x,y)$. This shows "$\subset$" in (24.4).

To prove "$\supset$", suppose $\Delta y \in \widehat{D}F(x|y)(\Delta x)$ and $\Delta z \in \widehat{D}G(y|z)(\Delta y)$ and take $\tau_k \searrow 0$ and $C \ni (\tilde{x}_k, \tilde{y}_k, \tilde{z}_k) \to (x,y,z)$. By definition of $\widehat{D}F(x|y)$, there then exist $(x_k, y_k) \in \operatorname{graph} F$ such that the first two limits hold. By the strict differentiability of G at y for z, we can also find $z_k \in G(y_k)$ such that $(z_k - \tilde{z}_k)/\tau_k \to \Delta z$. This shows the remaining limit.

Finally, the tangential regularity of C follows from the assumed T-regularities of F and G by comparing (24.4) with the corresponding expression (22.4). □

If one of the two mappings is single-valued, we can use Lemma 24.2 for verifying its semi-differentiability and Theorem 20.12 for regularity and the expression of its graphical derivative to obtain from Lemma 24.6 the following two special cases.

Corollary 24.7 (fundamental lemma on compositions: single-valued outer mapping). *Let X, Y, Z be Banach spaces and*

$$C := \{(x, y, G(y)) \mid y \in F(x)\}$$

for $F : X \rightrightarrows Y$ and $G : Y \to Z$. If $(x,y,z) \in C$ and G is continuously differentiable at y, then

$$\widehat{T}_C(x,y,z) = \{(\Delta x, \Delta y, G'(y)\Delta y) \mid \Delta y \in \widehat{D}F(x|y)(\Delta x)\}.$$

Moreover, if F is T-regular at (x,y), then C is tangentially-regular at $(x,y,G(y))$.

Corollary 24.8 (fundamental lemma on compositions: single-valued inner mapping). *Let X, Y, Z be Banach spaces and*

$$C := \{(x, y, z) \mid y = F(x),\ z \in G(y)\}$$

for $F : X \rightrightarrows Y$ and $G : Y \to Z$. If $(x, y, z) \in C$, F is continuously Fréchet differentiable at x, and $F'(x)$ has a right-inverse $F'(x)^\dagger \in \mathbb{L}(Y; X)$, then

$$\widehat{T}_C(x, y, z) = \{(\Delta x, \Delta y, \Delta z) \mid \Delta y = F'(x)\Delta x,\ \Delta z \in \widehat{D}G(y|z)(\Delta y)\}.$$

Moreover, if G is T-regular at (y, z), then C is tangentially regular at (x, y, z).

24.3 CALCULUS RULES

Using these lemmas, we again obtain calculus rules under the assumption that the involved set-valued mapping is regular.

Theorem 24.9 (addition of a single-valued differentiable mapping). *Let X, Y be Banach spaces, let $G : X \to Y$ be Fréchet differentiable, and $F : X \rightrightarrows Y$. If G is continuously Fréchet differentiable at $x \in X$ and F is T-regular at $(x, y - G(x))$ for $y \in H(x) := F(x) + G(x)$, then H is T-regular at (x, y) and*

$$\widehat{D}H(x|y)(\Delta x) = \widehat{D}F(x|y - G(x))(\Delta x) + G'(x)\Delta x \qquad (\Delta x \in X).$$

Proof. We construct H from C and R as in Theorem 22.12. Due to the assumptions (noting that continuous differentiability implies strict differentiability), C and RC are tangentially regular by Lemmas 24.4 and 24.6, respectively. We now obtain the claimed expression from Theorem 22.12. □

Theorem 24.10 (outer composition with a single-valued differentiable mapping). *Let X, Y, Z be Banach spaces, $F : X \rightrightarrows Y$, and $G : Y \to Z$. Let $x \in X$ and $z \in H(x) := G(F(X))$ be given. If G is continuously Fréchet differentiable at each $y \in F(x)$, invertible on $\operatorname{ran} G$ near z with Fréchet differentiable inverse at z, and F is T-regular at (x, y), then H is T-regular at (x, z) and*

$$\widehat{D}H(x|z)(\Delta x) = \bigcup_{y:G(y)=z} G'(y)\widehat{D}F(x|y)(\Delta x) \qquad (\Delta x \in X).$$

Proof. We construct H from C and R as in Theorem 22.13. Due to the assumptions, C and RC are tangentially regular by Corollary 24.7 and Lemma 24.4, respectively. We now obtain the claimed expression from Theorem 22.13. □

The special case for a linear operator follows from this exactly as in the proof of Corollary 22.14.

Corollary 24.11 (outer composition with a linear operator). *Let X, Y, Z be Banach spaces, $A \in \mathbb{L}(Y;Z)$, and $F : X \rightrightarrows Y$. If A has a bounded left-inverse $A^\dagger$ and F is T-regular at (x, y) for $x \in X$ and the unique $y \in Y$ with $Ay = z$, then for any $x \in X$ and $z \in H(x) := AF(x)$, then H is T-regular at (x, z) and*

$$\widehat{D}H(x|z)(\Delta x) = A\widehat{D}F(x|y)(\Delta x) \qquad (\Delta x \in X).$$

Theorem 24.12 (inner composition with a single-valued differentiable mapping). *Let X, Y, Z be Banach spaces, $F : X \to Y$ and $G : Y \rightrightarrows Z$. Let $x \in X$ and $z \in H(x) := G(F(x))$. If F is continuously Fréchet differentiable near x such that $F'(x)$ has a right-inverse $F'(x)^\dagger \in \mathbb{L}(Y;X)$ and G is T-regular at $(F(x), z)$, then H is T-regular at (x, z) and*

$$\widehat{D}H(x|z)(\Delta x) = \widehat{D}G(F(x)|z)(F'(x)\Delta x) \qquad (\Delta x \in X).$$

Proof. We construct H from C and R as in Theorem 22.15. Due to the assumptions, C and RC are tangentially regular by Corollary 24.8 and Lemma 24.4, respectively. We now obtain the claimed expression from Theorem 22.15. □

Corollary 24.13 (inner composition with a linear operator). *Let X, Y, Z be Banach spaces, $A \in \mathbb{L}(X;Y)$, and $G : Y \rightrightarrows Z$. Let $H := G \circ A$ for $A \in \mathbb{L}(X;Y)$ and $G : Y \rightrightarrows Z$ on Banach spaces X, Y, and Z. If A has a right-inverse $A^\dagger \in \mathbb{L}(Y;X)$ and G is T-regular at (Ax, z) for $x \in X$ and $z \in H(x) := G(Ax)$, then H is T-regular at (x, z) and*

$$\widehat{D}H(x|z)(\Delta x) = \widehat{D}G(Ax|z)(A\Delta x) \qquad (\Delta x \in X).$$

As in Section 22.3, we can apply these results to chain rules for subdifferentials, this time only at points where these subdifferentials are T-regular.

Corollary 24.14 (second derivative chain rule for convex subdifferential). *Let X, Y be Banach spaces, let $f : Y \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $A \in \mathbb{L}(X;Y)$ be such that A has a right-inverse $A^\dagger \in \mathbb{L}(Y;X)$, and $\operatorname{ran} A \cap \operatorname{int} \operatorname{dom} f \neq \emptyset$. Let $h := f \circ A$. If ∂f is T-regular at Ax, $x \in X$, for $y^* \in \partial f(Ax)$, then ∂h is T-regular at x for $x^* = A^* y^*$ and*

$$\widehat{D}[\partial h](x|x^*)(\Delta x) = A^* \widehat{D}[\partial f](Ax|y^*)(A\Delta x) \qquad (\Delta x \in X).$$

Theorem 24.15 (product rule). *Let X, Y, Z be Banach spaces, let $G : X \to \mathbb{L}(Y;Z)$ be Fréchet differentiable, and $F : X \rightrightarrows Y$. Assume that $G(\tilde{x}) \in \mathbb{L}(Y;Z)$ has a left-inverse $G(\tilde{x})^\dagger \in \mathbb{L}(Z;Y)$ for $\tilde{x}$ near $x \in X$ and that the mapping $\tilde{x} \mapsto G(\tilde{x})^\dagger$ is Fréchet differentiable at x. Let $x \in X$ and $z \in H(x) := G(x)F(x) := \bigcup_{y \in F(x)} G(x)y$. If F is T-regular at x for the unique $y \in F(x)$ satisfying $G(x)y = z$ and G is continuously differentiable at y, then H is T-regular at x for z and*

$$\widehat{D}H(x|z)(\Delta x) = [G'(x)\Delta x]\, y + G(x)\widehat{D}F(x|y)\Delta x \qquad (\Delta x \in X).$$

Proof. We construct H from R_1 and $\operatorname{graph}(\bar{G} \circ \bar{F})$ as in Theorem 22.19. Due to the assumptions, $\bar{G}$ and $\bar{F}$ are T-regular, and hence H is tangentially regular by Theorem 24.10 and Lemma 24.4. We now obtain the claimed expression from Theorem 22.19. □

Corollary 24.16 (second derivative chain rule for Clarke subdifferential). *Let X, Y be Banach spaces, let $f : Y \to R$ be locally Lipschitz continuous, and let $S : X \to Y$ be twice continuously differentiable. Set $h : X \to Y$, $h(x) := f(S(x))$. If there exists a neighborhood U of $x \in X$ such that*

(i) f is Clarke regular at $S(\tilde{x})$ for all $\tilde{x} \in X$;

(ii) $S'(\tilde{x})$ has a right-inverse $S'(\tilde{x})^\dagger \in \mathbb{L}(Y; X)$ for all $\tilde{x} \in U$;

(iii) the mapping $\tilde{x} \mapsto S'(\tilde{x})^{\dagger *}$ is Fréchet differentiable at x;

and $\partial_C f$ is T-regular at $S(x)$ for $y^* \in \partial_C f(S(x))$, then $\partial_C h$ is T-regular at x for $x^* = S'(x)^* y^*$ and

$$\widehat{D}[\partial_C h](x|x^*)(\Delta x) = (S''(x)\Delta x)^* y^* + S'(x)^* \widehat{D}[\partial_C f](S(x)|y^*)(S'(x)\Delta x) \quad (\Delta x \in X).$$

25 CALCULUS FOR THE LIMITING CODERIVATIVE

The limiting coderivative is the most challenging of all the graphical and coderivatives, and developing exact calculus rules for it requires the most assumptions. In particular, we will here assume a stronger variant of the assumptions of Chapter 23 for the Fréchet coderivative that also implies N-regularity of the set-valued mapping so that we can exploit the stronger properties of the Fréchet coderivative. To prove the fundamental composition lemmas, we will also need to introduce the concept of *partial sequential normal compactness* that will be used to prevent certain unit-length coderivatives from converging weakly-$*$ to zero. This concept will also be needed in Chapter 27.

25.1 STRICT CODIFFERENTIABILITY

Let X, Y be Banach spaces. We say that F is *strictly codifferentiable* at $x \in X$ for $y \in F(x)$ if

$$D^*F(x|y)(y^*) = \left\{ x^* \in X^* \;\middle|\; \begin{array}{l} \forall \operatorname{graph} F \ni (x_k, y_k) \to (x, y),\ \varepsilon_k \searrow 0 : \\ \exists (x_k^*, y_k^*) \xrightarrow{*} (x^*, y^*) \text{ with } x_k^* \in \widehat{D}^*_{\varepsilon_k} F(x_k|y_k)(y_k^*) \end{array} \right\}, \tag{25.1}$$

i.e., if (18.8) is a full weak-$*$-limit. From Theorem 20.12 and Corollary 20.14, it is clear that single-valued continuously differentiable mappings and their inverses are strictly codifferentiable.

Lemma 25.1. *Let X, Y be Banach spaces, $F : X \to Y$, $x \in X$, and $y = F(x)$.*

(i) If F is continuously differentiable at x, then F is strictly codifferentiable at x for y.

(ii) If F is continuously differentiably near x, then F^{-1} is strictly codifferentiable at y for x.

The next lemma and counterexample demonstrate that strict codifferentiability is a strictly stronger assumption than N-regularity.

Lemma 25.2. *Let X, Y be Banach spaces and let $F : X \rightrightarrows Y$ be strictly codifferentiable at x for y. Then F is N-regular at x for y.*

Proof. By Theorem 18.5, strict codifferentiability, and the definition of the inner limit, respectively,

$$\begin{aligned}\widehat{N}_{\operatorname{graph} F}(x,y) &\subset N_{\operatorname{graph} F}(x,y)\\ &= \limsup_{\operatorname{graph} F \ni (\tilde{x},\tilde{y}) \to (x,y),\, \varepsilon \searrow 0} \widehat{N}^{\varepsilon}_{\operatorname{graph} F}(\tilde{x},\tilde{y})\\ &\subset \widehat{N}_{\operatorname{graph} F}(x,y).\end{aligned}$$

Therefore $N_{\operatorname{graph} F}(x,y) = \widehat{N}_{\operatorname{graph} F}(x,y)$, i.e., $\operatorname{graph} F$ is normally regular at (x,y). □

Example 25.3 (graphical regularity does not imply strict codifferentiability). Consider $F(x) := [|x|,\infty)$, $x \in \mathbb{R}$. Then $\operatorname{graph} F = \operatorname{epi} |\cdot|$ is a convex set and therefore graphically regular at all points and

$$N_{\operatorname{graph} F}(x,|x|) = \begin{cases} (\operatorname{sign} x, -1)[0,\infty) & \text{if } x \neq 0,\\ \operatorname{graph} F^\circ = \{(x^*,y^*) \mid -y^* \geq |x^*|\} & \text{if } x = 0.\end{cases}$$

Hence $N_{\operatorname{graph} F}$ is not continuous and therefore, a fortiori, F is not strictly codifferentiable at $(0,0)$.

25.2 PARTIAL SEQUENTIAL NORMAL COMPACTNESS

One central difficulty in working with infinite-dimensional spaces is the need to distinguish weak-$*$ convergence and strong convergence. In particular, we need to prevent certain sequences whose norm is bounded away from zero from weak-$*$ converging to zero. As we cannot guarantee this in general, we need to add this as an assumption. In our specific setting, this is the *partial sequential normal compactness* (*PSNC*) of $G : Y \rightrightarrows Z$ at y for z, which holds if

$$\begin{aligned}(25.2)\quad \varepsilon_k \searrow 0,\ (y_k,z_k) \to (y,z),\ y_k^* \xrightarrow{*} 0,\ \|z_k^*\|_{Z^*} \to 0,\ \text{and } y_k^* \in \widehat{D}^*_{\varepsilon_k} G(y_k|z_k)(z_k^*)\\ \Rightarrow \|y_k^*\|_{Y^*} \to 0.\end{aligned}$$

Obviously, if Y^* finite-dimensional, then every mapping $G : Y \rightrightarrows Z$ is PSNC. To prove the PSNC property of single-valued mappings and their inverses, we will need an estimate of ε-coderivatives.

Lemma 25.4. *Let X, Y be Banach spaces and let $F : X \to Y$ be continuously differentiable at $x \in X$. Then for any $\varepsilon > 0$, $L := \|F'(x)\|_{\mathbb{L}(X;Y)}$, and $y = F(x)$,*

$$\widehat{D}^*_\varepsilon F(x|y)(y^*) \subset \mathbb{B}(F'(x)^* y^*, (L+1)\varepsilon) \qquad (y^* \in Y^*).$$

Proof. By definition, $x^* \in \widehat{D}^* F(x|y)(y^*)$ if and only if for every sequence $x_k \to x$,

$$\limsup_{k\to\infty} \frac{\langle x^*, x_k - x\rangle_X - \langle y^*, F(x_k) - F(x)\rangle_Y}{\sqrt{\|x_k - x\|_X^2 + \|F(x_k) - F(x)\|_Y^2}} \le \varepsilon. \tag{25.3}$$

Let $\ell > L$. Then by the continuous differentiability and therefore local Lipschitz continuity of F at x, we have $\|F(x_k) - F(x)\|_Y \le \ell \|x_k - x\|_X$ for large enough k and therefore

$$\limsup_{k\to\infty} \frac{\langle x^*, x_k - x\rangle_X - \langle y^*, F(x_k) - F(x)\rangle_Y}{\|x_k - x\|_X} \le \varepsilon(\ell + 1).$$

Furthermore, the Fréchet differentiability of F implies that

$$\limsup_{k\to\infty} \frac{\langle F'(x)^* y^*, x_k - x\rangle_X - \langle y^*, F(x_k) - F(x)\rangle_Y}{\|x_k - x\|_X} = 0$$

and hence that

$$\limsup_{k\to\infty} \frac{\langle x^* - F'(x)^* y^*, x_k - x\rangle_X}{\|x_k - x\|_X} \le \varepsilon(\ell + 1).$$

Since $x_k \to x$ was arbitrary, this implies $\|x^* - F'(x)^* y^*\|_{X^*} \le \varepsilon(\ell + 1)$, and since $\ell > L$ was arbitrary, the claim follows. □

Lemma 25.5. *Let Y, Z be Banach spaces and $G : Y \to Z$. If either*

(a) G is continuously differentiable near $y \in Y$ or

(b) Y^* is finite-dimensional,

then G is PSNC at y for $z = G(y)$.

Proof. The finite-dimensional case (b) is clear from the definition (25.2) of the PSNC property.

For case (a), we have from Lemma 25.4 that $\widehat{D}^*_{\varepsilon_k} G(y_k|z_k)(z_k^*) \subset \mathbb{B}(G'(y_k)^* z_k^*, \ell\varepsilon_k)$ for any $\ell > \|G'(y_k)\|_{\mathbb{L}(Y;Z)}$. By the continuous differentiability of G, this will hold for $\ell > \|G'(y)\|_{\mathbb{L}(Y;Z)}$ and any $k \in \mathbb{N}$ large enough. Thus there exist $d_k^* \in \mathbb{B}(0, \ell\varepsilon_k)$ such that

$$y_k^* = G'(y_k)^* z_k^* + d_k^* = G'(y)^* z_k^* + [G'(y_k) - G'(y)]^* z_k^* + d_k^* \to 0$$

since $d_k^* \to 0$ (due to $\varepsilon_k \searrow 0$), $\|z_k^*\|_{Z^*} \to 0$, $y_k \to y$, and G is continuously differentiable near y. □

Lemma 25.6. *Let Y, Z be Banach spaces and $G : Y \to Z$. If either*

(a) G is continuously differentiable near $y \in Y$ and $G'(y) \in \mathbb{L}(Y;Z)$ has a right-inverse $G'(y)^\dagger \in \mathbb{L}(Z;Y)$, or

(b) Z^* is finite-dimensional,

then G^{-1} is PSNC at $z = G(y)$ for y.

Proof. The finite-dimensional case (b) is clear from the definition (25.2) of the PSNC property.

For case (a), we have from the definition of $\widehat{D}^*_\varepsilon F$ via $\widehat{N}^\varepsilon_{\operatorname{graph} F}$ that $\Delta z^*_k \in \widehat{D}^*_\varepsilon G^{-1}(z_k|y_k)(\Delta y^*_k)$ if and only if $\Delta y^*_k \in \widehat{D}^*_\varepsilon G(y_k|z_k)(\Delta z^*_k)$. We thus have to show that

$$\varepsilon_k \searrow 0,\ (y_k, z_k) \to (y, z),\ z^*_k \xrightarrow{*} 0,\ \|y^*_k\|_{Y^*} \to 0,\ \text{and } y^*_k \in \widehat{D}^*_{\varepsilon_k} G(y_k|z_k)(z^*_k) \Rightarrow \|z^*_k\|_{Z^*} \to 0.$$

From Lemma 25.4, it follows that $\widehat{D}^*_{\varepsilon_k} G(y_k|z_k)(z^*_k) \subset \mathbb{B}(G'(y_k)^* z^*_k, \ell\varepsilon_k))$ for any $\ell > \|G'(y_k)\|_{\mathbb{L}(Y;Z)}$. As in Lemma 25.5, we now deduce that $y^*_k = G'(y_k)^* z^*_k + d^*_k$ for some $d^*_k \in \mathbb{B}(0, \ell\varepsilon_k)$. Since $y^*_k - d^*_k \to 0$, we also have $G'(y_k)^* z^*_k \to 0$ and thus $G'(y)^* z^*_k + [G'(y_k) - G'(y)]^* z^*_k \to 0$. Since $\{z^*_k\}_{k\in\mathbb{N}}$ is bounded by the continuous differentiability of G and $y_k \to y$, we obtain $G'(y)^* z^*_k \to 0$. Since $G'(y)$ is assumed to have a right-inverse, $G'(y)^*$ has a left-inverse, Hence this implies $z^*_k \to 0$ as required. □

We will use PSNC to obtain the following partial compactness property for the limiting coderivative, for which we need to assume reflexivity (or finite-dimensionality) of Y.

Lemma 25.7. *Let Y, Z be Banach spaces and $G : Y \rightrightarrows Z$. Let $y \in Y$ and $z \in G(y)$ be given. Assume $y^* \in D^*G(y|z)(0)$ implies $y^* = 0$ and either*

(a) Y is finite-dimensional or

(b) Y is reflexive and G is PSNC at y for z.

If

$$(y_k, z_k) \to (y, z), \quad z^*_k \xrightarrow{*} z^*, \quad \tilde{\varepsilon}_k \searrow 0, \quad \text{and} \quad \bar{y}^*_k \in \widehat{D}^*_{\tilde{\varepsilon}_k} G(y_k|z_k)(z^*_k),$$

then there exists a subsequence such that $\bar{y}^*_k \xrightarrow{*} \bar{y}^* \in D^*G(y|z)(z^*)$.

Proof. We first show that $\{\bar{y}^*_k\}_{k\in\mathbb{N}}$ is bounded. We argue by contradiction and suppose that $\{\bar{y}^*_k\}_{k\in\mathbb{N}}$ is unbounded. We may then assume that $\|\bar{y}^*_k\|_{Y^*} \to \infty$ by switching to an (unrelabelled) subsequence. Since $\widehat{D}^*_{\tilde{\varepsilon}_k} G(y_k|z_k)$ is formed from a cone, we also have

$$B_{Y^*} \ni \bar{y}^*_k / \|\bar{y}^*_k\|_{Y^*} \in \widehat{D}^*_{\tilde{\varepsilon}_k} G(y_k|z_k)(z^*_k / \|\bar{y}^*_k\|_{Y^*}).$$

Observe that $\|z^*_k / \|\bar{y}^*_k\|_{Y^*}\|_{Z^*} \to 0$ because $\{z^*_k\}_{k\in\mathbb{N}}$ is bounded. Since Y is reflexive, we can use the Eberlein–Šmulyan Theorem 1.9 to extract a subsequence such that $\bar{y}^*_k / \|\bar{y}^*_k\|_{Y^*} \xrightarrow{*} \bar{y}^*$ for some $\bar{y}^* \in D^*G(y|z)(0)$. If Y is finite-dimensional, clearly $\bar{y}^* \neq 0$. Otherwise we need to

use the assumed PSNC property. If $\bar{y}^* = 0$, then (25.2) implies that $1 = \|\bar{y}_k^*/\|\bar{y}_k^*\|_{Y^*}\|_{Y^*} \to 0$, which is a contradiction. Therefore $\bar{y}^* \neq 0$. However, we have assumed $\bar{y}^* \in D^*G(y|z)(0)$ to imply $\bar{y}^* = 0$, so we obtain a contradiction.

Therefore $\{\bar{y}_k^*\}_{k\in\mathbb{N}}$ is bounded, so we may again use the Eberlein–Šmulyan Theorem 1.9 to extract a subsequence converging to some $\bar{y}^* \in Y$. By the definition of the limiting coderivative, this implies $\bar{y}^* \in D^*G(y|z)(z^*)$ and hence the claim. □

Remark 25.8. The PSNC property, its stronger variant *sequential normal compactness* (*SNC*), and their implications are studied in significant detail in [Mordukhovich, 2006].

25.3 CONE TRANSFORMATION FORMULAS

As in Section 24.2, we now show that normal regularity is preserved under certain transformations by deriving explicit expressions for the transformed cones and then comparing them with the corresponding expressions of the Fréchet coderivative.

Lemma 25.9. *Let X, Y be Banach spaces and assume there exists a family of continuous inverse selections $\{R_y^{-1} : U_y \to C \mid y \in C,\ Ry = x\}$ of $R \in \mathbb{L}(Y;X)$ to $C \subset Y$ at $x \in X$. If each R_y^{-1} for all $y \in C$ with $Ry = x$ is locally Lipschitz at x with the factor L_x, and C is normally regular at all such y, then RC is normally regular at x and*

$$N_{RC}(x) = \bigcap_{y\in C:Ry=x} \{x^* \in X^* \mid R^*x^* \in N_C(y)\}.$$

Proof. We first prove "$\subset$". Let $x^* \in N_{RC}(x)$. By definition, this holds if and only if there exist $\varepsilon_k \searrow 0$ as well as $x_k^* \xrightarrow{*} x^*$ and $x_k \to x$ with $x_k^* \in \widehat{N}_{RC}^{\varepsilon_k}(x_k)$. Let $y \in Y$ be such that $Ry = x$. Defining $y_k := R_y^{-1}x_k$, we have $Ry_k = x_k$ and $C \ni y_k \to y$. Thus Lemma 23.2 yields $R^*x_k^* \in \widehat{N}_C^{\varepsilon_k L}(y_k)$. By definition of the limiting coderivative, this implies that $R^*x^* \in N_C(y)$. Since this holds for all $y \in Y$ with $Ry = x$, we obtain "$\subset$".

For "$\supset$", Let $x^* \in X^*$ be such that $R^*x^* \in N_C(y)$ for all $y \in Y$ with $Ry = x$. Then the assumption of regularity of C at y implies that $R^*x^* \in \widehat{N}_C(y)$. Hence taking $y_k = y$, $x_k^* = x^*$, and $x_k = x$, we deduce from Lemma 23.2 that $x_k^* \in \widehat{N}_{RC}(x_k)$. Again by definition, this implies that $x^* \in N_{RC}(x)$.

Finally, the normal regularity of RC at x is clear from writing $N_C(y) = \widehat{N}_C(y)$ and comparing our expression for $N_{RC}(x)$ to the expression for $\widehat{N}_{RC}(x)$ provided by Lemma 23.2. □

Remark 25.10 (regularity assumptions). Again, the assumption in Lemma 24.4 that C is normally regular is not needed if $\ker R = \{0\}$ or, more generally, if R is a continuously differentiable mapping with $\ker \nabla R(y) = \{0\}$.

For the fundamental lemma for the limiting coderivative, we need to assume reflexivity of Y in order to apply the PSNC via Lemma 25.7.

Lemma 25.11 (fundamental lemma on compositions). *Let X, Y, Z be Banach spaces with Y reflexive and*

$$C := \{(x, y, z) \mid y \in F(x),\ z \in G(y)\}$$

for $F : X \rightrightarrows Y$, and $G : Y \rightrightarrows Z$. Let $(x, y, z) \in C$.

(i) If G is strictly codifferentiable and PSNC at y for z, semi-codifferentiable near $(y, z) \in \operatorname{graph} G$, and $y^* \in D^*G(y|z)(0)$ implies $y^* = 0$, then

$$N_C(x, y, z) = \{(x^*, y^*, z^*) \mid x^* \in D^*F(x|y)(-\tilde{y}^* - y^*),\ \tilde{y}^* \in D^*G(y|z)(z^*)\}.$$

(ii) If F^{-1} is strictly codifferentiable and PSNC at y for x, semi-codifferentiable near $(y, x) \in \operatorname{graph} F^{-1}$, and $y^* \in D^*F^{-1}(y|x)(0)$ implies $y^* = 0$, then

$$N_C(x, y, z) = \{(x^*, y^*, z^*) \mid x^* \in D^*F(x|y)(-\tilde{y}^* - y^*),\ -\tilde{y}^* \in D^*G(y|z)(-z^*)\}.$$

Moreover, if F is N-regular at x for y and G is N-regular at y for z, then C is normally regular at (x, y, z).

Proof. We only consider the case (i); the case (ii) is shown analogously. To show the inclusion "$\subset$", let $(x^*, y^*, z^*) \in N_C(x, y, z)$, which by definition holds if and only if there exist $\varepsilon_k \searrow 0$ as well as $(x_k^*, y_k^*, z_k^*) \xrightarrow{*} (x^*, y^*, z^*)$ and $C \ni (x_k, y_k, z_k) \to (x, y, z)$ with $(x_k^*, y_k^*, z_k^*) \in \widehat{N}_C^{\varepsilon_k}(x_k, y_k, z_k)$. Since by assumption G is semi-codifferentiable at $(y_k, z_k) \in \operatorname{graph} G$ for $k \in \mathbb{N}$ sufficiently large, we can apply Lemma 23.4 (i) to obtain a $\tilde{y}_k^* \in \widehat{D}^*G(y_k|z_k)(z_k^*)$ such that

$$x_k^* \in \widehat{D}_{\varepsilon_k}^* F(x_k|y_k)(-\tilde{y}_k^* - y_k^*). \tag{25.4}$$

Since $z_k^* \xrightarrow{*} z^*$, $(y_k, z_k) \to (y, z)$, and $\varepsilon_k \searrow 0$, we deduce from Lemma 25.7 that $\tilde{y}_k^* \xrightarrow{*} \tilde{y}^*$ (for a subsequence) for some $\tilde{y}^* \in D^*G(y|z)(z^*)$. Since also $x_k^* \xrightarrow{*} x^*$ and $y_k^* \xrightarrow{*} y^*$, by (25.4) and the definition of the limiting coderivative, this implies that $x^* \in D^*F(x|y)(-\tilde{y}^* - y^*)$.

To show "$\supset$", let $x^* \in D^*F(x|y)(-\tilde{y}^* - y^*)$ and $\tilde{y}^* \in D^*G(y|z)(z^*)$. We can then by the definition of $D^*F(x|y)$ find $\varepsilon_k \searrow 0$ as well as $(x_k, y_k) \to (x, y)$ and $(x_k^*, \bar{y}_k^*) \xrightarrow{*} (x^*, \tilde{y}^* + y^*)$ with $x_k^* \in \widehat{D}_{\varepsilon_k}^* F(x_k|y_k)(-\bar{y}_k^*)$. Since G is strictly codifferentiable at y for z, taking any $z_k \to z$, we can now find $z_k^* \xrightarrow{*} z^*$ and $\tilde{y}_k^* \xrightarrow{*} \tilde{y}^*$ with $\tilde{y}_k^* \in \widehat{D}_{\varepsilon_k}^* G(y_k|z_k)(z_k^*)$. Letting $y_k^* := \bar{y}_k^* - \tilde{y}_k^*$, this implies that $y_k^* \xrightarrow{*} y^*$ and that $x_k^* \in \widehat{D}_{\varepsilon_k}^* F(x_k|y_k)(-\tilde{y}_k^* - y_k^*)$. By Lemma 23.4 (i), it follows that $(x_k^*, y_k^*, z_k^*) \in \widehat{N}_C^{\varepsilon_k}(x_k, y_k, z_k)$. The claim now follows again from the definition of $N_C(x, y, z)$ as the corresponding outer limit.

Finally, the normal regularity of C follows from the N-regularity of F and G (via Lemma 25.2) by comparing Lemma 23.4 with Lemma 23.4 (i) for $\varepsilon = 0$. □

If one of the two mappings is single-valued, we can use Lemma 25.1 for verifying its semi-differentiability and Theorem 20.12 for the expression of its graphical derivative to obtain from Lemma 25.11 the following two special cases.

Corollary 25.12 (fundamental lemma on compositions: single-valued outer mapping). *Let X, Y, Z be Banach spaces with Y reflexive and*

$$C := \{(x, y, G(y)) \mid y \in F(x)\}$$

for $F : X \rightrightarrows Y$ and $G : Y \to Z$. If $(x, y, z) \in C$ and G is continuously differentiable near y, then

$$N_C(x, y, z) = \{(x^*, y^*, z^*) \mid x^* \in D^*F(x|y)(-[G'(y)]^*z^* - y^*),\ y^* \in Y^*\}.$$

Moreover, if F is N-regular at (x, y), then C is normally regular at $(x, y, G(y))$.

Proof. We apply Lemma 25.11, where the strict and semi-codifferentiability requirements on G are verified by Lemmas 23.1 and 25.1; the PSNC requirement follows from Lemma 25.5; and the requirement of $y^* \in D^*G(y|z)(0)$ implying $y^* = 0$ follows from the expression of Theorem 20.12 for $D^*G(y|z)(0)$. The claimed normal regularity of C for N-regular F follows from the N-regularity of G established by Theorem 20.12. □

Corollary 25.13 (fundamental lemma on compositions: single-valued inner mapping). *Let X, Y, Z be Banach spaces with Y reflexive and*

$$C := \{(x, y, z) \mid y = F(x),\ z \in G(y)\}$$

for $F : X \rightrightarrows Y$ and $G : Y \to Z$. If $(x, y, z) \in C$, F is continuously differentiable near x, and either

(a) $F'(x)$ has a right-inverse $F'(x)^\dagger \in \mathbb{L}(Y; X)$ or

(b) Y^* is finite-dimensional,

then

$$N_C(x, y, z) = \{(F'(x)^*(-\tilde{y}^* - y^*), y^*, z^*) \mid -\tilde{y}^* \in D^*G(y|z)(-z^*),\ y^* \in Y^*\}.$$

Moreover, if G is N-regular at (y, z), then C is normally regular at (x, y, z).

Proof. We apply Lemma 25.11, where the strict and semi-codifferentiability requirements on F^{-1} are verified by Lemmas 23.1 and 25.1; the PSNC requirement follows from Lemma 25.6; and the requirement of $y^* \in D^*F^{-1}(y|x)(0)$ implying $y^* = 0$ follows from the expression of Corollary 20.14 for $D^*F^{-1}(y|x)(0)$. The claimed normal regularity of C for N-regular G follows from the N-regularity of F established by Theorem 20.12. □

25.4 CALCULUS RULES

Using these lemmas, we obtain again calculus rules.

Theorem 25.14 (addition of a single-valued differentiable mapping). *Let X, Y be Banach spaces with X reflexive, let $G : X \to Y$ be Fréchet differentiable, and $F : X \rightrightarrows Y$. If G is continuously Fréchet differentiable at $x \in X$ and F is N-regular at $(x, y - G(x))$ for $y \in H(x) := F(x) + G(x)$, then H is N-regular at (x, y) and*

$$D^*H(x|y)(y^*) = D^*F(x|y - G(x))(y^*) + [G'(x)]^* y^* \qquad (y^* \in Y^*).$$

Proof. We follow Theorem 23.8 to construct H from C and R, and prove properties of the inverse selections of R. Due to the assumptions (noting that continuous differentiability implies strict differentiability), C and RC are normally regular by Lemmas 25.9 and 25.11, respectively. We now obtain the claimed expression from Theorem 23.8. □

Theorem 25.15 (outer composition with a single-valued differentiable mapping). *Let X, Y, Z be Banach spaces with Y reflexive, $F : X \rightrightarrows Y$, and $G : Y \to Z$. Let $x \in X$ and $z \in H(x) := G(F(X))$ be given. If G is continuously Fréchet differentiable at each $y \in F(x)$, invertible on* ran G *near z with Fréchet differentiable inverse at z, and F is N-regular at (x, y), then H is N-regular at (x, z) and*

$$D^*H(x|z)(z^*) = \bigcap_{y:G(y)=z} D^*F(x|y)([G'(y)]^* z^*) \qquad (z^* \in Z^*).$$

Proof. We follow the proof of Theorem 23.9 to construct H from C and R, and prove properties of the inverse selections of R. Due to the assumptions, C and RC are normally regular by Corollary 25.12 and Lemma 25.9, respectively. We now obtain the claimed expression from Theorem 23.9. □

Corollary 25.16 (outer composition with a linear operator). *Let X, Y, Z be Banach spaces with Y reflexive, $A \in \mathbb{L}(Y; Z)$, and $F : X \rightrightarrows Y$. If A has a bounded left-inverse $A^\dagger$ and F is N-regular at (x, y) for $x \in X$ and the unique $y \in Y$ with $Ay = z$, then for any $x \in X$ and $z \in H(x) := AF(x)$, then H is N-regular at (x, z) and*

$$D^*H(x|z)(z^*) = D^*F(x|y)(A^* z^*) \qquad (z^* \in Z^*).$$

Theorem 25.17 (inner composition with a single-valued differentiable mapping). *Let X, Y, Z be Banach spaces with Y reflexive, $F : X \to Y$ and $G : Y \rightrightarrows Z$. Let $x \in X$ and $z \in H(x) := G(F(x))$. If F is continuously Fréchet differentiable near x such that $F'(x)$ has a left-inverse $F'(x)^\dagger \in \mathbb{L}(Y; X)$ and G is T-regular at $(F(x), z)$, then H is T-regular at (x, z) and*

$$D^*H(x|z)(z^*) = [F'(x)]^* D^*G(F(x)|z)(z^*) \qquad (z^* \in Z^*).$$

Proof. We follow Theorem 23.11 to construct H from C and R, and prove properties of the inverse selections of R. Due to the assumptions, C and RC are normally regular by Corollary 25.13 and Lemma 25.9, respectively. We now obtain the claimed expression from Theorem 23.11. □

Corollary 25.18 (inner composition with a linear operator). *Let X, Y, Z be Banach spaces with Y reflexive, $A \in \mathbb{L}(X;Y)$, and $G : Y \rightrightarrows Z$. Let $H := G \circ A$ for $A \in \mathbb{L}(X;Y)$ and $G : Y \rightrightarrows Z$ on Banach spaces X, Y, and Z. If A has a right-inverse $A^\dagger \in \mathbb{L}(Y;X)$ and G is N-regular at (Ax, z) for $x \in X$ and $z \in H(x) := G(Ax)$, then H is N-regular at (x, z) and*

$$D^*H(x|z)(z^*) = A^*D^*G(Ax|z)(z^*) \qquad (z^* \in Z^*).$$

To apply these results for chain rules of subdifferentials, we now need to assume that *both* spaces are reflexive in addition to N-regularity.

Corollary 25.19 (second derivative chain rule for convex subdifferential). *Let X, Y be reflexive Banach spaces, let $f : Y \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, $A \in \mathbb{L}(X;Y)$ be such that A has a right-inverse $A^\dagger \in \mathbb{L}(Y;X)$, and $\operatorname{ran} A \cap \operatorname{int} \operatorname{dom} f \neq \emptyset$. Let $h := f \circ A$. If ∂f is N-regular at Ax, $x \in X$, for $y^* \in \partial f(Ax)$, then ∂h is N-regular at x for $x^* = A^*y^*$ and*

$$D^*[\partial h](x|x^*)(\Delta x) = A^*D^*[\partial f](Ax|y^*)(A\Delta x) \qquad (\Delta x \in X).$$

Theorem 25.20 (product rule). *Let X, Y, Z be Banach spaces with X, Y reflexive, let $G : X \to \mathbb{L}(Y;Z)$ be Fréchet differentiable, and $F : X \rightrightarrows Y$. Assume that $G(\tilde{x}) \in \mathbb{L}(Y;Z)$ has a left-inverse $G(\tilde{x})^\dagger \in \mathbb{L}(Z;Y)$ for $\tilde{x}$ near $x \in X$ and that the mapping $\tilde{x} \mapsto G(\tilde{x})^\dagger$ is Fréchet differentiable at x. Let $x \in X$ and $z \in H(x) := G(x)F(x) := \bigcup_{y \in F(x)} G(x)y$. If F is N-regular at x for the unique $y \in F(x)$ satisfying $G(x)y = z$ and G is continuously differentiable at y, then H is N-regular at x for z and*

$$D^*H(x|z)(z^*) = D^*F(x|y)(G(x)^*z^*) + ([G'(x)\cdot]y)^*z^* \qquad (z^* \in Z^*).$$

Proof. We follow Theorem 23.15 to construct H from R_1 and $\operatorname{graph}(\bar{G} \circ \bar{F})$, and prove properties of the inverse selections of R. Due to the assumptions, $\bar{G}$ and $\bar{F}$ are T-regular, and hence H is tangentially regular by Theorem 25.15 and Lemma 25.9. We now obtain the claimed expression from Theorem 23.15. □

Corollary 25.21 (second derivative chain rule for Clarke subdifferential). *Let X, Y be reflexive Banach spaces, let $f : Y \to R$ be locally Lipschitz continuous, and let $S : X \to Y$ be twice continuously differentiable. Set $h : X \to Y$, $h(x) := f(S(x))$. If there exists a neighborhood U of $x \in X$ such that*

(i) f is Clarke regular at $S(\tilde{x})$ for all $\tilde{x} \in X$;

(ii) $S'(\tilde{x})$ *has a right-inverse* $S'(\tilde{x})^\dagger \in \mathbb{L}(Y;X)$ *for all* $\tilde{x} \in U$*;*

(iii) the mapping $\tilde{x} \mapsto S'(\tilde{x})^{\dagger *}$ *is Fréchet differentiable at* x*;*

and $\partial_C f$ *is N-regular at* $S(x)$ *for* $y^* \in \partial_C f(S(x))$*, then* $\partial_C h$ *is N-regular at* x *for* $x^* = S'(x)^* y^*$ *and*

$$\widehat{D}^*[\partial_C h](x|x^*)(x^{**}) = \hat{S}(x)^* x^{**} + S'(x)^* \widehat{D}^*[\partial_C f](S(x)|y^*)(S'(x)^{**} x^{**}) \quad (x^{**} \in X^{**}).$$

Remark 25.22. Even in finite dimensions, calculus rules for the sum $F+G$ of arbitrary set-valued mappings $F, G : \mathbb{R}^N \rightrightarrows \mathbb{R}^M$ or the composition $F \circ H$ for $H : \mathbb{R}^N \rightrightarrows \mathbb{R}^N$ are much more limited, and in general only yield inclusions of the form

$$D^*[F+G](x|y)(y^*) \subset \bigcup_{\substack{y=y_1+y_2,\\ y_1 \in F(x),\\ y_2 \in G(x)}} D^*F(x|y_1)(y^*) + D^*G(x|y_2)(y^*),$$

and

$$D^*[F \circ H](x|y)(y^*) \subset \bigcup_{z \in H(x) \cap F^{-1}(y)} D^*H(x|z) \circ D^*F(z|y)(y^*).$$

We refer to [Mordukhovich, 2018; Rockafellar and Wets, 1998] for these and other results.

26 SECOND-ORDER OPTIMALITY CONDITIONS

We now illustrate the use of set-valued derivatives for optimization problems by showing how these can be used to derive second-order (sufficient and necessary) optimality conditions for non-smooth problems. Again, we do not aim for the most general or sharpest possible results and focus instead on problems having the form (P) involving the composition of a nonsmooth convex functional with a smooth nonlinear operator. As in the previous chapters, we will also assume a regularity conditions that allows for cleaner results.

26.1 SECOND-ORDER DERIVATIVES

Let X be a Banach space and $f : X \to \overline{\mathbb{R}}$. In this chapter, we set

$$\partial_C f(x) := \left\{ x^* \in X^* \;\middle|\; (x^*, -1) \in N^C_{\operatorname{epi} f}(x, f(x)) \right\},$$

where $N^C_A := \widehat{T}^\circ_A$ is the Clarke normal cone. By Lemma 20.19, this coincides with the classical Clarke subdifferential if $f : X \to \mathbb{R}$ is locally Lipschitz continuous.

As in the smooth case, second-order conditions are based on a local quadratic model built from curvature information at a point. Since in the nonsmooth case, second derivatives, i.e., graphical derivatives of the subdifferential, are no longer unique, we need to consider the entire set of them when building this curvature information. We therefore need to distinguish a *lower curvature model* at $x \in X$ for $x^* \in \partial_C f(x)$ in direction $\Delta x \in X$

$$Q_f(\Delta x; x|x^*) := \inf_{\Delta x^* \in D[\partial_C f](x|x^*)(\Delta x)} \langle \Delta x^*, \Delta x \rangle_X$$

as well as an *upper curvature model*

$$Q^f(\Delta x; x|x^*) := \sup_{\Delta x^* \in D[\partial_C f](x|x^*)(\Delta x)} \langle \Delta x^*, \Delta x \rangle_X.$$

It turns out that even for $\Delta x \neq 0$, we need to consider the *stationary upper model*

$$Q^f_0(\Delta x; x|x^*) := \sup_{\Delta x^* \in D[\partial_C f](x|x^*)(0)} \langle \Delta x^*, \Delta x \rangle_X,$$

which we use to define the *extended upper model*

$$\begin{aligned}\hat{Q}^f(\Delta x; x|x^*) &:= \max\left\{Q^f(\Delta x; x|x^*), Q_0^f(\Delta x; x|x^*)\right\} \\ &= \sup_{\Delta x^* \in D[\partial_C f](x|x^*)(\Delta x) \cup D[\partial_C f](x|x^*)(0)} \langle \Delta x^*, \Delta x\rangle_X.\end{aligned}$$

For smooth functionals, these models coincide with the usual Hessian.

Theorem 26.1. *Let X be a Banach space and let $f : X \to \mathbb{R}$ be twice continuously differentiable. Then for every $x, \Delta x \in X$,*

$$Q_f(\Delta x; x|f'(x)) = Q^f(\Delta x; x|f'(x)) = \langle f''(x)\Delta x, \Delta x\rangle_X$$

and

$$\hat{Q}^f(\Delta x; x|f'(x)) = \max\left\{0, \langle f''(x)\Delta x, \Delta x\rangle_X\right\}.$$

Proof. Since $\partial_C f(x) = \{f'(x)\}$ by Theorem 13.5, it follows from Theorem 20.12 that

$$D[\partial_C f(x)](x|x^*)(\Delta x) = \langle f''(x)\Delta x, \Delta x\rangle_X$$

and in particular $D[\partial_C f(x)](x|x^*)(0) = 0$, which immediately yields the claim. □

We illustrate the nonsmooth case with the usual examples of the indicator functional of the unit ball and the norm on $\mathbb{R}$.

Lemma 26.2. *Let $f(x) = \delta_{[-1,1]}(x)$, $x \in \mathbb{R}$. Then for every $x^* \in \partial f(x)$ and $\Delta x \in \mathbb{R}$,*

$$Q_f(\Delta x; x|x^*) = \begin{cases} \infty & \text{if } |x| = 1,\ x^* = 0,\ x\Delta x > 0, \\ \infty & \text{if } |x| = 1,\ x^* \in (0,\infty)x,\ \Delta x \neq 0, \\ 0, & \text{otherwise,} \end{cases}$$

$$Q^f(\Delta x; x|x^*) = \begin{cases} -\infty & \text{if } |x| = 1,\ x^* = 0,\ x\Delta x > 0, \\ -\infty & \text{if } |x| = 1,\ x^* \in (0,\infty)x,\ \Delta x \neq 0, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$\hat{Q}^f(\Delta x; x|x^*) = Q_0^f(\Delta x; x|x^*) = \begin{cases} \infty & \text{if } |x| = 1,\ x^* \in (0,\infty)x, \\ \infty & \text{if } |x| = 1,\ x^* = 0,\ x\Delta x > 0, \\ 0 & \text{if } |x| = 1,\ x^* = 0,\ x\Delta x \leq 0, \\ 0 & \text{if } |x| < 1. \end{cases}$$

Proof. The claims follow directly from the expression (20.7) in Theorem 20.17 with $\sup \emptyset = -\infty$ and $\inf \emptyset = \infty$. □

Lemma 26.3. *Let $f(x) = |x|$, $x \in \mathbb{R}$. Then for every $x^* \in \partial f(x)$ and $\Delta x \in \mathbb{R}$,*

$$Q_f(\Delta x; x|x^*) = \begin{cases} \infty & \text{if } x = 0,\ \Delta x \neq 0,\ \operatorname{sign} \Delta x \neq x^*, \\ 0 & \text{otherwise,} \end{cases}$$

$$Q^f(\Delta x; x|x^*) = \begin{cases} -\infty & \text{if } x = 0,\ \Delta x \neq 0,\ \operatorname{sign} \Delta x \neq x^*, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\hat{Q}^f(\Delta x; x|x^*) = Q_0^f(\Delta x; x|x^*) = \begin{cases} 0 & \text{if } x \neq 0,\ x^* = \operatorname{sign} x, \\ 0 & \text{if } x = 0,\ |x^*| = 1,\ x^* \Delta x \geq 0, \\ \infty & \text{if } x = 0,\ |x^*| = 1,\ x^* \Delta x < 0, \\ \infty & \text{if } x = 0,\ |x^*| < 1. \end{cases}$$

Proof. The claims follow directly from the expression (20.13) in Theorem 20.18 with $\sup \emptyset = -\infty$ and $\inf \emptyset = \infty$. □

These results can be lifted to the corresponding integral functionals on $L^p(\Omega)$ using the results of Chapter 21. Similarly, we obtain calculus rules for the curvature functionals from the corresponding results in Chapter 22.

Theorem 26.4 (sum rule). *Let X be a Banach space, let $f : X \to \mathbb{R}$ be locally Lipschitz continuous, and let $g : X \to \mathbb{R}$ be twice continuously differentiable. Set $j(x) := f(x) + g(x)$. Then for every $x \in X$ and $x^* \in \partial_C f(x)$,*

$$Q_j(\Delta x; x|x^* + g'(x)) = Q_f(\Delta x; x|x^*) + \langle g''(x)\Delta x, \Delta x\rangle_X \qquad (\Delta x \in X),$$
$$Q^j(\Delta x; x|x^* + g'(x)) = Q^f(\Delta x; x|x^*) + \langle g''(x)\Delta x, \Delta x\rangle_X \qquad (\Delta x \in X).$$

Proof. We only show the expression for the upper model, the lower model being analogous. First, by Theorem 13.20, we have $\partial_C j(x) = \{x^* + g'(x) \mid x^* \in \partial_C f(x)\}$. The sum rule Theorem 22.12 for the graphical derivative together with Theorem 20.12 then yields

$$D[\partial_C j](x|x^* + g'(x))(\Delta x) = D[\partial_C f](x|x^*)(\Delta x) + g''(x)\Delta x$$

and therefore

$$\begin{aligned} Q^j(\Delta x; x|x^* + g'(x)) &= \sup_{\Delta x^* \in D[\partial_C j](x|x^* + g'(x))(\Delta x)} \langle \Delta x^*, \Delta x\rangle_X \\ &= \sup_{\Delta x^* \in D[\partial_C f](x|x^*)(\Delta x)} \langle \Delta x^*, \Delta x\rangle_X + \langle g''(x)\Delta x, \Delta x\rangle_X. \end{aligned}$$ □

Theorem 26.5 (chain rule). *Let X, Y be Banach spaces, let $f : Y \to \mathbb{R}$ be convex, and let $S : X \to Y$ be twice continuously differentiable. Set $j(x) := f(S(x))$. If there exists a neighborhood U of $x \in X$ such that*

(i) f *is Clarke regular at* $S(\tilde{x})$ *for all* $\tilde{x} \in X$*;*

(ii) $S'(\tilde{x})^*$ *has a bounded left-inverse* $S'(\tilde{x})^{*\dagger} \in \mathbb{L}(X^*; Y^*)$ *for all* $\tilde{x} \in U$*;*

(iii) the mapping $\tilde{x} \mapsto S'(\tilde{x})^{\dagger *}$ *is Fréchet differentiable at* x*;*

then for all $x^* \in \partial_C h(x) = S'(x)^* \partial_C f(S(x))$*,*

$$Q_j(\Delta x; x|x^*) = \langle y^*, [S''(x)\Delta x]\Delta x\rangle_Y + Q^f(S'(x)\Delta x; S(x)|y^*) \qquad (\Delta x \in X),$$
$$Q^j(\Delta x; x|x^*) = \langle y^*, [S''(x)\Delta x]\Delta x\rangle_Y + Q^f(S'(x)\Delta x; S(x)|y^*) \qquad (\Delta x \in X),$$

for the unique $y^* \in \partial_C f(S(x))$ *such that* $S'(x)^* y^* = x^*$*.*

Proof. We again only consider the upper model Q^j, the lower model being analogous. Due to our assumptions, we can apply Corollary 24.16 to obtain

$$D[\partial_C(f \circ S)](x|x^*)(\Delta x) = [S''(x)^* \Delta x] y^* + S'(x)^* D[\partial f](S(x)|y^*)(S'(x)\Delta x),$$

where $S'' : X \to [X \to \mathbb{L}(Y^*; X^*)]$. Thus every $\Delta x^* \in D[\partial_C(f \circ S)](x|x^*)(\Delta x)$ can be written for some $\Delta y^* \in D[\partial f](S(x)|y^*)(S'(x)\Delta x)$ as $\Delta x^* = [S''(x)\Delta x]^* y^* + S'(x)^* \Delta y^*$. Inserting this into the definition of Q^j yields

$$\begin{aligned} Q^j(\Delta x; x|x^*) &= \sup_{\Delta y^* \in D[\partial f](S(x)|y^*)(S'(x)\Delta x)} \langle [S''(x)\Delta x]^* y^* + S'(x)^* \Delta y^*, \Delta x\rangle_X \\ &= \langle y^*, [S''(x)\Delta x]\Delta x\rangle_Y + \sup_{\Delta y^* \in D[\partial f](S(x)|y^*)(S'(x)\Delta x)} \langle \Delta y^*, S'(x)\Delta x\rangle_Y. \end{aligned}$$ □

26.2 SUBCONVEXITY

We say that $f : X \to \overline{\mathbb{R}}$ is *subconvex near* $\bar{x}$ for $\bar{x}^* \in \partial_C f(x)$ if for all $\rho > 0$, there exists $\varepsilon > 0$ such that

$$f(\tilde{x}) - f(x) \geq \langle x^*, \tilde{x} - x\rangle_X - \frac{\rho}{2}\|\tilde{x} - x\|_X^2 \qquad (x, \tilde{x} \in \mathbb{B}(\bar{x}, \varepsilon);\ x^* \in \partial_C f(x) \cap \mathbb{B}(\bar{x}^*, \varepsilon)). \tag{26.1}$$

We say that f is *subconvex at* $\bar{x}$ for $\bar{x}^*$ if this holds with $\tilde{x} = \bar{x}$ fixed. It is clear that convex functions are subconvex near any point for any subderivative. By extension, scalar functions such as $t \mapsto |t|^q$ for $q \in (0, 1)$ that are locally minorized by $\tilde{x} \mapsto f(\bar{x}) + \langle x^*, \tilde{x} - x\rangle_X$ at points of nonsmoothness are also subconvex.

The sum of two subconvex functions for which the subdifferential sum rule holds is clearly also subconvex. The next result shows that smooth functions simply need to have a non-negative Hessian at the point $\bar{x}$ to be subconvex. This is in contrast to the everywhere non-negative Hessian of convex functions.

Lemma 26.6. *Let X be a Banach space and let $f : X \to \mathbb{R}$ be twice continuously differentiable. If $\langle f''(\bar{x})\Delta x, \Delta x\rangle_X \geq 0$ for all $\Delta x \in X$, then f is subconvex near $\bar{x} \in X$ for $f'(\bar{x})$.*

Proof. Fix $\rho > 0$. We apply Theorem 2.10 first to f to obtain for every $x, h \in X$ that

$$f(x+h) - f(x) = \int_0^1 \langle f'(x+th), h\rangle_X \, dt.$$

Similarly, the same theorem applied to $t \mapsto \langle f'(x+th), h\rangle$ for any $x, h \in X$ yields

$$\langle f'(x+th), h\rangle_X - \langle f'(x), h\rangle_X = \int_0^1 \langle f''(x+sth)h, h\rangle_X \, ds.$$

Combined these two expansions yield

$$f(x+h) - f(x) = \langle f'(x), h\rangle_X + \int_0^1 \int_0^1 \langle f''(x+sth)h, h\rangle_X \, ds\, dt. \tag{26.2}$$

Since $\langle f''(\bar{x})h, h\rangle_X \geq 0$, we have

$$\langle f''(x+q)h, h\rangle_X \geq \langle [f''(x+q) - f''(\bar{x})]h, h\rangle_X \qquad (x, q, h \in X).$$

Therefore, by the continuity of f'', for any $\rho > 0$ we can find $\varepsilon > 0$ such that

$$\langle f''(x+q)h, h\rangle_X \geq -\frac{\rho}{2}\|h\|_X^2 \qquad (q \in \mathbb{B}(0,\varepsilon),\ x \in \mathbb{B}(\bar{x},\varepsilon),\ h \in X).$$

Taking $q = sth$, this and (26.2) shows that

$$f(x+h) - f(x) \geq \langle f'(x), h\rangle_X - \frac{\rho}{2}\|h\|_X^2.$$

The claim now follows by taking $h = \tilde{x} - x$. □

Remark 26.7. Subconvexity, which to our knowledge has not previously been treated in the literature, is a stronger condition than the prox-regularity introduced in [Poliquin and Rockafellar, 1996]. The latter requires (26.1) to hold merely for a fixed $\rho > 0$. The definition in [Rockafellar and Wets, 1998] is slightly broader and implies the earlier one. Their definition is itself a modification of the *primal-lower-nice* functions of [Thibault and Zagrodny, 1995]. Our notion of subconvexity is also related to those of *subsmooth* sets and *submonotone* operators introduced in [Aussel et al., 2005]. An alternative concept for functions, *subsmoothness* and *lower-C^k*, has been introduced in [Rockafellar, 1981].

26.3 SUFFICIENT AND NECESSARY CONDITIONS

We start with sufficient conditions, which are based on the upper model.

Theorem 26.8. *Let X be a Banach space and $f : X \to \overline{\mathbb{R}}$. If for $\bar{x} \in X$,*

(i) f is subconvex near $\bar{x}$ *for $\bar{x}^* = 0$;*

(ii) $0 \in \partial_C f(\bar{x})$;

(iii) there exists a $\mu > 0$ such that

$$\hat{Q}^f(\Delta x; \bar{x}|0) \geq \mu \|\Delta x\|_X^2 \qquad (\Delta x \in X);$$

then $\bar{x}$ is a strict local minimizer of f.

Proof. Let $\bar{x}^* := 0$ and $\Delta x \in X$. By the assumed subconvexity, for every $\rho > 0$ there exists $\varepsilon_\rho > 0$ such that for $x \in \mathbb{B}(\bar{x}, \varepsilon_\rho/2)$ and $x^* \in \partial_C f(x) \cap \mathbb{B}(\bar{x}^*, \varepsilon_\rho)$, we have for every $t > 0$ with $t\|\Delta x\|_X < \frac{1}{2}\varepsilon_\rho$ that

$$\frac{f(x + t\Delta x) - f(x) - t\langle \bar{x}^*, \Delta x\rangle_X}{t^2} \geq \frac{\langle x^* - \bar{x}^*, \Delta x\rangle_X}{t} - \frac{\rho}{2}\|\Delta x\|_X^2.$$

Since $\rho > 0$ was arbitrary, we thus obtain for every $\Delta\tilde{x} \in X$ and $\Delta x^* \in D[\partial f](x|x^*)(\Delta\tilde{x})$ that

$$\begin{aligned} A(\Delta x, \Delta\tilde{x}, \Delta x^*) &:= \liminf_{\substack{t\searrow 0,\,(x-\bar{x})/t\to\Delta\tilde{x}\\ (x^*-\bar{x}^*)/t\to\Delta x^*,\,x^*\in\partial_C f(x)}} \frac{f(x + t\Delta x) - f(x) - t\langle \bar{x}^*, \Delta x\rangle_X}{t^2} \\ &\geq \liminf_{\substack{t\searrow 0,\,(x-\bar{x})/t\to\Delta\tilde{x}\\ (x^*-\bar{x}^*)/t\to\Delta x^*,\,x^*\in\partial_C f(x)}} \frac{\langle x^* - \bar{x}^*, \Delta x\rangle_X}{t} = \langle \Delta x^*, \Delta x\rangle_X. \end{aligned}$$

This implies that

$$\sup_{\Delta x^* \in D[\partial_C f](\bar{x}|\bar{x}^*)(\Delta\tilde{x})} A(\Delta x, \Delta\tilde{x}, \Delta x^*) \geq \sup_{\Delta x^* \in D[\partial_C f](\bar{x}|\bar{x}^*)(\Delta\tilde{x})} \langle \Delta x^*, \Delta x\rangle_X =: B(\Delta x, \Delta\tilde{x}).$$

Since $\bar{x}^* = 0$, we can fix $x = \bar{x} + t\Delta x$ and $\Delta\tilde{x} = \Delta x$ in the lim inf above and use (iii) to obtain

$$\liminf_{t\searrow 0} \frac{f(\bar{x} + 2t\Delta x) - f(\bar{x} + t\Delta x)}{t^2} \geq B(\Delta x, \Delta x). \tag{26.3}$$

Similarly, fixing $x = \bar{x}$ and $\Delta\tilde{x} = 0$ yields

$$\liminf_{t\searrow 0} \frac{f(\bar{x} + t\Delta x) - f(\bar{x})}{t^2} \geq B(\Delta x, 0) \geq 0, \tag{26.4}$$

where the final inequality follows from the definition of B by taking $\Delta x^* = 0$ (which is possible since $\bar{x}^* \in \partial_C f(\bar{x})$). We now make a case distinction.

(I) $B(\Delta x, 0) \geq \mu \|\Delta x\|_X^2$. In this case, the lim inf is strictly positive for $\Delta x \neq 0$ and hence $f(\bar{x} + t\Delta x) > f(\bar{x})$ for all $t > 0$ sufficiently small.

(II) $B(\Delta x, 0) < \mu \|\Delta x\|_X^2$. In this case, it follows from (iii) that

$$\mu \|\Delta x\|_X^2 \leq \hat{Q}^f(\Delta x; \bar{x}|0) = \max\{B(\Delta x, \Delta x), B(\Delta x, 0)\}$$

and hence that $B(\Delta x, \Delta x) = \hat{Q}^f(\Delta x; \bar{x}|0) \geq \mu \|\Delta x\|_X^2$. Summing (26.3) and (26.4) then yields

$$\liminf_{t \searrow 0} \frac{f(\bar{x} + 2t\Delta x) - f(\bar{x})}{t^2} \geq B(\Delta x, \Delta x) \geq \mu \|\Delta x\|_X^2,$$

which again implies for $\Delta x \neq 0$ that $f(\bar{x} + t\Delta x) > 0 f(\bar{x})$ for all $t > 0$ sufficiently small.

Since $\Delta x \in X$ was arbitrary, $\bar{x}$ is by definition a strict local minimizer of f. □

Remark 26.9. The use of the stationarity curvature model Q_0^f in the second-order condition is required since the upper curvature model may not provide any information about the growth of f at $\bar{x}$ in certain directions. However, since $D[\partial_C f](\bar{x}|\bar{x}^*)(0)$ is a cone, if it contains *any* element Δx^* such that $\langle \Delta x^*, \Delta x \rangle_X > 0$, then $B(\Delta x, 0) = Q_0^f(\Delta x; \bar{x}|\bar{x}^*) = \infty$, ensuring that the condition (iii) holds in the direction Δx for any $\mu > 0$. For example, if $f(x) = |x|$, then Lemma 26.3 shows that $Q^f(\Delta x; 0|0) = 0$ for $\Delta x \neq 0$, which indeed does not provide any information about the growth of f at 0. Conversely, $Q_0^f(\Delta x; 0|0) = \infty$ for any $\Delta x \neq 0$, so the growth is more rapid than Q^f can measure.

Combining Theorem 26.8 with Theorem 26.1, we obtain the classical sufficient second-order condition. (Recall that in infinite-dimensional spaces, positive definiteness and coercivity are no longer equivalent, and the latter, stronger, property is usually required.)

Corollary 26.10. *Let X be a Banach space and let $f : X \to \mathbb{R}$ be twice continuously differentiable. If for $\bar{x} \in X$,*

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

(ii) there exists a $\mu > 0$ such that

$$\langle f''(\bar{x})\Delta x, \Delta x \rangle_X \geq \mu \|\Delta x\|_X^2 \qquad (\Delta x \in X);$$

then $\bar{x}$ is a local minimizer of f.

Proof. To apply Theorem 26.8, it suffices to note that $\partial_C f(x) = \{f'(x)\}$ by Theorem 13.5 and that the second-order condition ensures subconvexity of f at $\bar{x}$ for $\bar{x}^* = 0$ by Lemma 26.6. □

For nonsmooth functionals, we merely illustrate the sufficient second-order condition with a simple but nontrivial scalar example.

Corollary 26.11. *Let $X = \mathbb{R}$ and $j := f + g$ for $g : \mathbb{R} \to \mathbb{R}$ twice continuously differentiable and $f(x) = |x|$. Then the sufficient condition of Theorem 26.8 holds at $\bar{x} \in \mathbb{R}$ if and only if one of the following cases holds:*

(a) $\bar{x} = 0$ and $|g'(\bar{x})| < 1$;

(b) $\bar{x} = 0$, $|g'(\bar{x})| = 1$, and $g''(\bar{x}) > 0$; or

(c) $\bar{x} \neq 0$, $g'(\bar{x}) = -\operatorname{sign} \bar{x}$, and $g''(\bar{x}) > 0$.

Proof. We apply Theorem 26.8, for which we need to verify its conditions. First, note that (ii) is equivalent to $0 = x^* + g'(\bar{x})$ for some $x^* \in \partial f(\bar{x}) = \operatorname{sign} \bar{x}$ by Theorem 13.20 and Example 4.7.

We now verify the subconvexity of j near $\bar{x}$ for $\bar{x}^* = 0$. Expanding the definition (26.1), this requires

$$|\tilde{x}| - |x| + g(\tilde{x}) - g(x) \geq \langle x^* + g'(x), \tilde{x} - x \rangle - \frac{\rho}{2}\|\tilde{x} - x\|^2 \tag{26.5}$$
$$(x, \tilde{x} \in \mathbb{B}(\bar{x}, \varepsilon);\ x^* \in \partial_C |\cdot|(x) \cap \mathbb{B}(\bar{x}^* - g'(x), \varepsilon)).$$

In cases (b) and (c), we can apply Lemma 26.6 to deduce the subconvexity of g and therefore of $j = f + g$ since f is convex. For case (a), we have $\bar{x} = 0$ with $|g'(\bar{x})| < 1$. Since g' is continuous, we consequently have $\bar{x}^* - g'(x) = -g'(x) \in (-1, 1)$ when $|x - \bar{x}| = |x|$ is small enough. Since $\partial f(x) \in \{-1, 1\}$ for $x \neq 0$, it follows that $\partial_C |\cdot|(x) \cap \mathbb{B}(\bar{x}^* - g'(x), \varepsilon) = \emptyset$ for $x \in \mathbb{B}(\bar{x}, \varepsilon) \setminus \{\bar{x}\}$ for small enough $\varepsilon > 0$. Therefore, for small enough $\varepsilon > 0$, the condition (26.5) reduces to

$$|\tilde{x}| + g(\tilde{x}) - g(0) \geq \langle x^* + g'(0), \tilde{x} \rangle - \frac{\rho}{2}|\tilde{x}|^2 \quad (\tilde{x} \in [-\varepsilon, \varepsilon],\ |x^*| \leq 1,\ |x^* + g'(0)| \leq \varepsilon). \tag{26.6}$$

Furthermore, $|g'(0)| < 1$ implies that for every $\rho > 0$ and $c > 0$, we can find an $\varepsilon > 0$ sufficiently small that

$$(1 - \varepsilon - |g'(0)|)|\tilde{x}| \geq \frac{c - \rho}{2}|\tilde{x}|^2 \qquad (\tilde{x} \in [-\varepsilon, \varepsilon]).$$

Since $g : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable, we can apply a Taylor expansion in $\bar{x} = 0$ to obtain for some $c > 0$ and $|\tilde{x}|$ sufficiently small that

$$g(0) \leq g(\tilde{x}) + \langle g'(0), -\tilde{x} \rangle + \frac{c}{2}|\tilde{x}|^2.$$

Adding this to the previous inequality, we obtain for sufficiently small $\varepsilon > 0$ and $x^* \in [-1, 1]$ satisfying $|x^* + g'(0)| \le \varepsilon$ that

$$\begin{aligned} |\tilde{x}| + g(\tilde{x}) - g(0) &\ge (|g'(0)| + \varepsilon)|\tilde{x}| + \langle g'(0), \tilde{x}\rangle - \frac{\rho}{2}|\tilde{x}|^2 \\ &\ge \langle x^* + g'(0), \tilde{x}\rangle - \frac{\rho}{2}|\tilde{x}|^2 \end{aligned}$$

for every $|\tilde{x}| \le \varepsilon$, which is (26.6). Hence $j = f + g$ is subconvex near $\bar{x} = 0$ for $0 = x^* + g'(0)$.

To verify (iii), we compute the upper curvature model. Let $\Delta x \in X$. Then by Theorems 26.1 and 26.4,

$$\begin{aligned} Q^j(\Delta x; x|x^* + g'(x)) &= Q^f(\Delta x; x|x^*) + \langle g''(x)\Delta x, \Delta x\rangle, \\ Q_0^j(\Delta x; x|x^* + g'(x)) &= Q^f(\Delta x; x|x^*), \end{aligned}$$

where Q^f is given by Lemma 26.3. It follows that

$$Q^j(\Delta x; x|x^* + g'(x)) = \begin{cases} -\infty & \text{if } x = 0,\ \Delta x \ne 0,\ \operatorname{sign}\Delta x \ne x^*, \\ \langle g''(x)\Delta x, \Delta x\rangle & \text{otherwise,} \end{cases}$$

and

$$Q_0^j(\Delta x; x|x^* + g'(x)) = \begin{cases} 0 & \text{if } x \ne 0, x^* = \operatorname{sign} x, \\ 0 & \text{if } x = 0,\ |x^*| = 1,\ x^*\Delta x \ge 0, \\ \infty & \text{if } x = 0,\ |x^*| = 1,\ x^*\Delta x < 0, \\ \infty & \text{if } x = 0, |x^*| < 1. \end{cases}$$

Thus

$$\hat{Q}^j(\Delta x; x|x^* + g'(x)) = \begin{cases} \max\{0, \langle g''(x)\Delta x, \Delta x\rangle\} & \text{if } x \ne 0,\ x^* = \operatorname{sign} x, \\ \max\{0, \langle g''(x)\Delta x, \Delta x\rangle\} & \text{if } x = 0,\ |x^*| = 1,\ x^*\Delta x \ge 0, \\ \infty & \text{if } x = 0,\ |x^*| = 1,\ x^*\Delta x < 0, \\ \infty & \text{if } x = 0,\ |x^*| < 1. \end{cases}$$

The condition (iii) is thus equivalent to

$$\max\{0, \langle g''(\bar{x})\Delta x, \Delta x\rangle\} \ge \mu\|\Delta x\|^2 \quad \text{when} \quad \begin{cases} \bar{x} \ne 0 \text{ or} \\ \bar{x} = 0,\ |g'(\bar{x})| = 1, \text{ and } g'(\bar{x})\Delta x < 0. \end{cases}$$

The left inequality can only hold for arbitrary $\Delta x \in \mathbb{R}$ if $\mu = g''(\bar{x}) > 0$. Hence (ii) and (iii) hold if and only if one of the cases (a)–(c) holds. □

Note that case (a) corresponds to the case of strict complementarity or graphical regularity of ∂f in Theorem 20.18. Conversely, cases (b), and (c) imply that g and therefore j is locally convex, recalling from Theorem 4.2 that for convex functionals, the first-order optimality conditions are necessary *and* sufficient.

Now we formulate our necessary condition, which is based on the *lower* curvature model.

Theorem 26.12. *Let X be a Banach space and $f : X \to \overline{\mathbb{R}}$. If $\bar{x} \in X$ is a local minimizer of f and f is locally Lipschitz continuous and subconvex at $\bar{x}$ for $0 \in X^*$, then*

$$Q_f(\Delta x; \bar{x}|0) \geq 0 \qquad (\Delta x \in X).$$

Proof. We have from Theorem 13.4 that $\bar{x}^* := 0 \in \partial_C f(\bar{x})$. By the assumed subconvexity, for every $\rho > 0$ there exists $\varepsilon > 0$ such that for $x \in \mathbb{B}(\bar{x}, \varepsilon/2)$ and $x_t^* \in \partial_C f(\bar{x} + t\Delta\tilde{x}) \cap \mathbb{B}(\bar{x}^*, \varepsilon)$, we have for every $t > 0$ with $t\|\Delta x\|_X < \varepsilon/2$ that

$$\frac{f(\bar{x} + t\Delta\tilde{x}) - f(\bar{x}) - t\langle \bar{x}^*, \Delta\tilde{x}\rangle_X}{t^2} \leq \frac{\langle x_t^* - \bar{x}^*, \Delta\tilde{x}\rangle_X}{t} + \frac{\rho}{2}\|\Delta\tilde{x}\|_X^2.$$

For every $\Delta x^* \in D[\partial_C f](\bar{x}|\bar{x}^*)(\Delta x)$, by definition there exist $\Delta\tilde{x} \to \Delta x$ and, for small enough $t > 0$, $x_t^* \in \partial_C f(x + t\Delta\tilde{x}) \cap \mathbb{B}(\bar{x}^*, \varepsilon)$ such that $(x_t^* - \bar{x}^*)/t \to \Delta x^* \in X^*$. Since $\rho > 0$ was arbitrary and $\bar{x}^* = 0$, it follows that

$$\begin{aligned}
\liminf_{\substack{\Delta\tilde{x}\to\Delta x \\ t\searrow 0}} \frac{f(\bar{x} + t\Delta\tilde{x}) - f(\bar{x})}{t^2} &\leq \liminf_{\substack{\Delta\tilde{x}\to\Delta x \\ t\searrow 0}} \left(\frac{\langle x_t^* - \bar{x}^*, \Delta x\rangle_X}{t} + \frac{\langle x_t^* - \bar{x}^*, \Delta\tilde{x} - \Delta x\rangle_X}{t} \right) \\
&= \liminf_{t\searrow 0} \frac{\langle x_t^* - \bar{x}^*, \Delta x\rangle_X}{t} \\
&\leq \inf_{\Delta x^* \in D[\partial_C f](\bar{x}|\bar{x}^*)(\Delta x)} \langle \Delta x^*, \Delta x\rangle_X \\
&= Q_f(\Delta x; \bar{x}|\bar{x}^*) = Q_f(\Delta x; \bar{x}|0).
\end{aligned}$$

Since $\bar{x}$ is a local minimizer, we have $f(\bar{x}) \leq f(\bar{x} + t\Delta\tilde{x})$ for $t > 0$ sufficiently small and $\Delta\tilde{x}$ sufficiently close to Δx. Rearranging and passing to the limit thus yields the claimed nonnegativity of $Q_f(\Delta x; \bar{x}|0)$. □

Remark 26.13. Compared to the sufficient condition of Theorem 26.8, the necessary condition does not involve a *stationary lower model*

$$Q_{f,0}(\Delta x; \bar{x}|0) := \inf_{\Delta x^* \in D[\partial_C f](\bar{x}|0)(0)} \langle \Delta x^*, \Delta x\rangle_X.$$

In fact, $Q_{f,0}(\Delta x; \bar{x}|0) \geq 0$ is *not* a necessary optimality condition: let $f(x) = |x|$, $x \in \mathbb{R}$, and $\bar{x} = 0$. Then by Theorem 20.18, $D[\partial f](0|0)(0) = \mathbb{R}$ and hence $Q_{f,0}(\Delta x; 0|0) = -\infty$ for all $\Delta x \neq 0$.

For smooth functions, we recover the usual second-order necessary condition from Theorem 26.1.

Corollary 26.14. *Let X be a Banach space and let $f : X \to \mathbb{R}$ be twice continuously differentiable. If $\bar{x} \in X$ is a local minimizer of f, then*

$$\langle f''(\bar{x})\Delta x, \Delta x\rangle_X \geq 0 \qquad (\Delta x \in X).$$

We again illustrate the nonsmooth case with a scalar example.

Corollary 26.15. *Let $X = \mathbb{R}$ and $j := f + g$ for $g : \mathbb{R} \to \mathbb{R}$ twice continuously differentiable and $f(x) = |x|$. Then the necessary condition of Theorem 26.12 holds at $\bar{x}$ if and only if $g''(\bar{x}) \geq 0$.*

Proof. We apply Theorem 26.12, for which we need to verify its conditions. Both f and g are locally Lipschitz continuous by Theorem 3.13 and Lemma 2.11, respectively, and hence so is j. We have already verified the subconvexity of j in Corollary 26.11.

By Theorems 13.4 and 13.20 and Example 4.7, we again have $0 = x^* + g'(\bar{x})$ for some $x^* \in \partial f(\bar{x}) = \operatorname{sign} \bar{x}$. It remains to compute the lower curvature model. Let $\Delta x \in X$. By Theorems 26.1 and 26.4,

$$Q_j(\Delta x; x|x^* + g'(x)) = Q_f(\Delta x; x|x^*) + \langle g''(x)\Delta x, \Delta x\rangle,$$

where Q_f is given by Lemma 26.3. It follows that

$$Q_j(\Delta x; x|x^* + g'(x)) = \begin{cases} \infty & \text{if } x = 0,\ \Delta x \neq 0,\ \operatorname{sign} \Delta x \neq x^*, \\ \langle g''(x)\Delta x, \Delta x\rangle & \text{otherwise.} \end{cases}$$

Hence the condition $Q_j(\Delta x; \bar{x}|0) \geq 0$ for all $\Delta x \in X$ reduces to $g''(\bar{x}) \geq 0$. □

Remark 26.16. Second-order optimality conditions can also be based on *epigraphical derivatives*, which were introduced in [Rockafellar, 1985, 1988]; we refer to [Rockafellar and Wets, 1998] for a detailed discussion. A related approach based on second-order directional curvature functionals was used in [Christof and Wachsmuth, 2018] for deriving necessary and sufficient second-order optimality conditions for smooth optimization problems subject to nonsmooth and possibly nonconvex constraints.

27 LIPSCHITZ-LIKE PROPERTIES

A related issue to second-order conditions is that of *stability* of the solution to optimization problems under perturbation. To motivate the following, let $f : X \to \overline{\mathbb{R}}$ and suppose we wish to find $\bar{x} \in X$ such that $0 \in \partial f(\bar{x})$ for a suitable subdifferential. Suppose further that we are given some $\tilde{x} \in X$ with $w \in \partial f(\tilde{x})$ with $\|w\|_{X^*} \leq \varepsilon$ – say, from one of the algorithms in Chapter 8. A natural question is then for an error estimate $\|\bar{x} - \tilde{x}\|_X$ in terms of ε. Clearly, if ∂f has a single-valued and Lipschitz continuous inverse, this is the case since then

$$\|\bar{x} - \tilde{x}\|_X = \|(\partial f)^{-1}(0) - (\partial f)^{-1}(w)\|_X \leq L\|w\|_{X^*}.$$

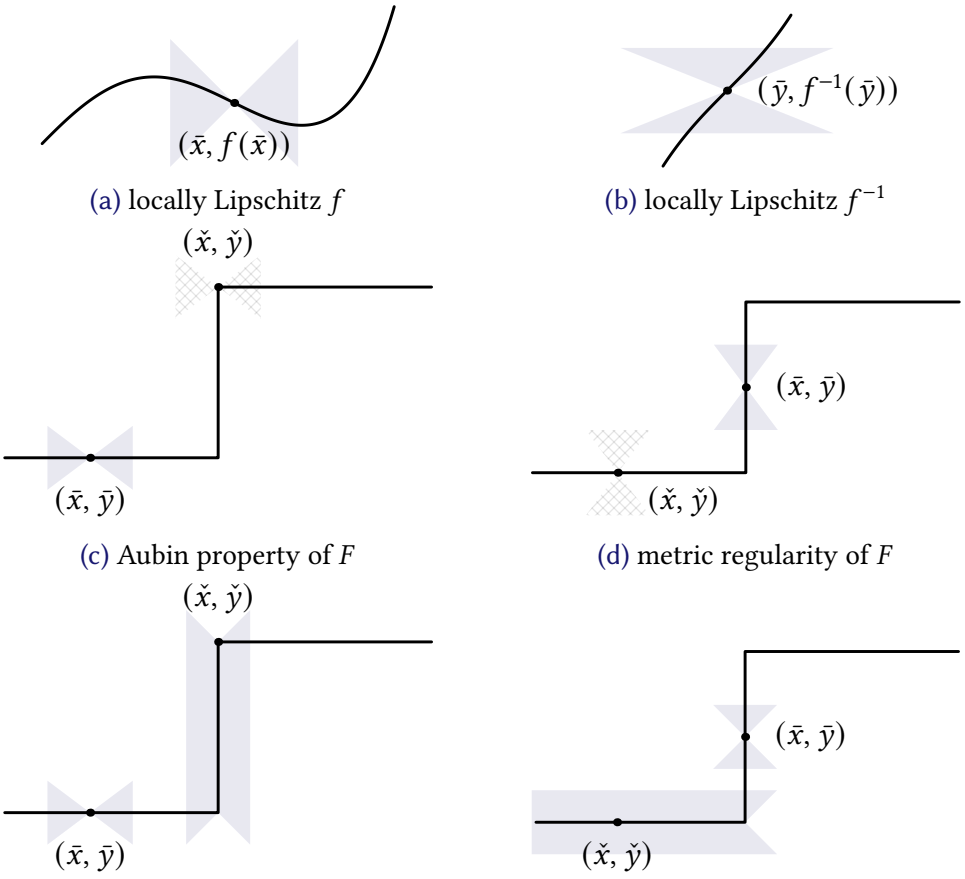
Of course, the situation is much more complicated in the set-valued case. To treat this, we first have to define suitable notions of Lipschitz-like behavior of set-valued mappings, which we then characterize using coderivatives (generalizing the characterization of the Lipschitz constant of a differentiable single-valued mapping through the norm of its derivative). We return to the question of stability of minimizers in the more general context of perturbations of parametrized solution mappings in Chapter 28.

27.1 LIPSCHITZ-LIKE PROPERTIES OF SET-VALUED MAPPINGS

To set up the definition of Lipschitz-like properties for set-valued mappings, it is helpful to recall from Section 1.1 for single-valued functions the distinction between (point-based) local Lipschitz continuity *at* a point and (neighborhood-based) local Lipschitz continuity *near* a point. (Figure 27.2b below shows a function that is locally Lipschitz at but not near the give point.) Similarly, we will have to distinguish for set-valued mappings the corresponding notions of *Aubin property* (which is point-based) and *calmness* (which is neighborhood-based). If these properties hold for the inverse of a mapping, we will call the mapping itself *metrically regular* and *metrically subregular*, respectively. These four properties are illustrated in Figure 27.1.

Recall also from Lemma 17.4 the definition of the distance of a point $x \in X$ to a set $A \subset X$, which we here write for the sake of convenience as

$$\operatorname{dist}(A, x) := \operatorname{dist}(x, A) := d_A(x) = \inf_{\tilde{x} \in A} \|x - \tilde{x}\|_X.$$



(a) locally Lipschitz f (b) locally Lipschitz f^{-1}

(c) Aubin property of F (d) metric regularity of F

(e) calmness of F (f) metric subregularity of F

Figure 27.1: Illustration of Lipschitz-like properties using cones. The thick lines are the graph of the function; if this graph is locally contained in a filled cone, the property holds, while a cross-hatched cone indicates that the property is violated.

We then say that $F : X \rightrightarrows Y$ has the *Aubin* or *pseudo-Lipschitz property at $\bar{x}$ for $\bar{y}$* if graph F is closed near $(\bar{x}, \bar{y})$ and there exist $\delta, \kappa > 0$ such that

$$\operatorname{dist}(y, F(x)) \leq \kappa \operatorname{dist}(F^{-1}(y), x) \quad (x \in \mathbb{B}(\bar{x}, \delta),\ y \in \mathbb{B}(\bar{y}, \delta)). \tag{27.1}$$

We call the infimum of all $\kappa > 0$ for which (27.1) holds for some $\delta > 0$ the *graphical modulus* of F at $\bar{x}$ for $\bar{y}$, written $\operatorname{lip} F(\bar{x}|\bar{y})$.

When we are interested in the stability of the optimality condition $0 \in F(\bar{x})$, it is typically more beneficial to study the Aubin property of the inverse F^{-1}. This is called the *metric regularity* of F at a point $(\bar{x}, \bar{y}) \in \operatorname{graph} F$, which holds if there exist $\kappa, \delta > 0$ such that

$$\operatorname{dist}(x, F^{-1}(y)) \leq \kappa \operatorname{dist}(y, F(x)) \quad (x \in \mathbb{B}(\bar{x}, \delta),\ y \in \mathbb{B}(\bar{y}, \delta)). \tag{27.2}$$

We call the infimum of all $\kappa > 0$ for which (27.2) holds for some $\delta > 0$ the *modulus of metric regularity* of F at $\bar{x}$ for $\bar{y}$, written $\operatorname{reg} F(\bar{x}|\bar{y})$.

The metric regularity and Aubin property are too strong to be satisfied in many applications. A weaker notion is provided by *(metric) subregularity* at $(\bar{x}, \bar{y}) \in \operatorname{graph} F$, which holds if there exist $\kappa, \delta > 0$ such that

$$\operatorname{dist}(x, F^{-1}(\bar{y})) \leq \kappa \operatorname{dist}(\bar{y}, F(x)) \quad (x \in \mathbb{B}(\bar{x}, \delta)). \tag{27.3}$$

Compared to metric regularity, this allows much more leeway for F by fixing $y = \bar{y} \in F(\bar{x})$ (while still allowing x to vary). We call the infimum of all $\kappa > 0$ for which (27.3) holds for some $\delta > 0$ for the *modulus of (metric) subregularity* of F at $\bar{x}$ for $\bar{y}$, written $\operatorname{subreg} F(\bar{x}|\bar{y})$.

The counterpart of metric subregularity that relaxes the Aubin property is known as *calmness*. We say that $F : X \rightrightarrows Y$ is calm at $\bar{x}$ for $\bar{y}$ if there exist $\kappa, \delta > 0$ such that

$$\operatorname{dist}(y, F(\bar{x})) \leq \kappa \operatorname{dist}(\bar{x}, F^{-1}(y)) \quad (y \in \mathbb{B}(\bar{y}, \delta)). \tag{27.4}$$

We call the infimum of all $\kappa > 0$ for which (27.4) holds for some $\delta > 0$ the *modulus of calmness* of F at $\bar{x}$ for $\bar{y}$, written $\operatorname{calm} F(\bar{x}|\bar{y})$. Clearly the Aubin property implies calmness, while metric regularity implies metric subregularity.

Unfortunately, the direct calculation of the different moduli is often infeasible in practice. Much of the rest of this chapter concentrates on calculating the graphical modulus and the modulus of metric regularity in special cases. We will consider metric subregularity (as well as a related, weaker, notion of strong submonotonicity) in the following Section 29.1.

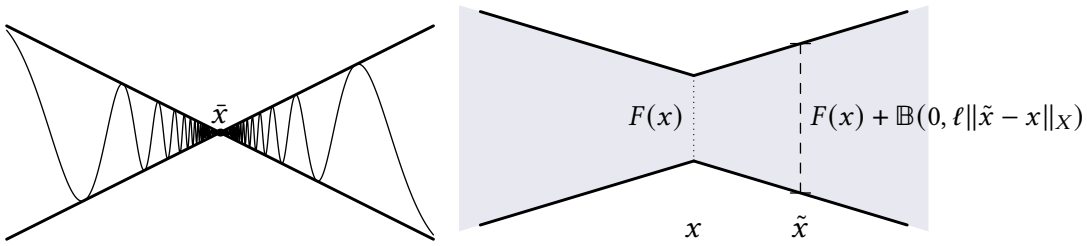
Remark 27.1. The Aubin property is due to [Aubin, 1984], whereas metric subregularity is due to [Ioffe, 1979], first given the modern name in [Dontchev and Rockafellar, 2004]. Calmness was introduced in [Robinson, 1981] as the *upper Lipschitz property*. Metric regularity is equivalent to *openness at a linear rate* near $(\widehat{u}, \widehat{w})$ and holds for smooth maps by the classical Lyusternik–Graves theorem. We refer in particular to [Dontchev and Rockafellar, 2014; Ioffe, 2017] for further information on these and other related properties.

In particular, related to metric subregularity is the stronger concept of *strong metric subregularity*, which was introduced in [Rockafellar, 1989] and requires the existence of $\kappa, \delta > 0$ such that

$$\|x - \bar{x}\|_X \leq \kappa \operatorname{dist}(\bar{y}, F(x)) \quad (x \in \mathbb{B}(\bar{x}, \delta)),$$

i.e., a bound on the norm distance to $\bar{x}$ rather than the closest preimage of $\bar{y}$. Its many properties are studied in [Cibulka et al., 2018], which also introduced q-exponent versions. Particularly worth noting is that strong metric subregularity is invariant with respect to perturbations by smooth functions, while metric subregularity is not.

Weaker and "partial" concepts of regularity have also been considered in the literature. Of particular note is the directional metric subregularity of [Gfrerer, 2013]. The idea here is to study necessary optimality conditions by requiring metric regularity or subregularity only along critical directions instead of all directions. In [Valkonen, 2021c], by contrast, the norms in the definition of subregularity are made operator-relative to study the partial subregularity on subspaces; compare the testing of algorithms for structured problems in Section 10.2.



(a) oscillating single-valued function (b) graph of $x \mapsto F(x) + \mathbb{B}(0, \ell\|\tilde{x} - x\|_X)$

Figure 27.2: The oscillating example in (a) illustrates a function f that is locally Lipschitz (or calm) *at* $\bar{x}$, but not locally Lipschitz (or does not have the Aubin property) *near* the same point: the graph of the function stays in the cone formed by the thick lines and based at $(\bar{x}, f(\bar{x})) \in \operatorname{graph} f$. If, however, we move the cone locally along the graph, even increasing its width, the graph will not be contained the cone. In (b) we illustrate the "fat cone" structure $\operatorname{graph}(\tilde{x} \mapsto F(x) + \mathbb{B}(0, \ell\|\tilde{x} - x\|_X))$ appearing on the right-hand-side in Theorem 27.2 (i), and varying with the second base point x around $\bar{x}$. This is to be contrasted with the leaner cone $\operatorname{graph}(\tilde{x} \mapsto f(\bar{x}) + \mathbb{B}(0, \ell\|\tilde{x} - \bar{x}\|_X))$ bounding the function in (a).

We now provide alternative characterizations of the Aubin property and of calmness. These extend to metric regularity and subregularity, respectively, by application to the inverse.

The right-hand-side of the set-inclusion characterization (i) in the next theorem forms a "fat cone" that we illustrate in Figure 27.2b. It should locally at each base point x around $\bar{x}$ bound F for the Aubin property to be satisfied. Based on the formulation (i), we illustrate in Figure 27.3 the satisfaction and dissatisfaction of the Aubin property. The other two new characterizations show that we do not need to restrict $\tilde{x}$ to a tiny neighborhood of $\bar{x}$ in neither (i) nor the original characterization (27.1).

Theorem 27.2. *Let X, Y be Banach spaces and $F : X \rightrightarrows Y$. Then the following are equivalent for $\bar{x} \in X$ and $\bar{y} \in F(\bar{x})$:*

(i) There exists $\kappa, \delta > 0$ such that

$$F(\tilde{x}) \cap \mathbb{B}(\bar{y}, \delta) \subset F(x) + \mathbb{B}(0, \kappa\|\tilde{x} - x\|_X) \quad (\tilde{x}, x \in \mathbb{B}(\bar{x}, \delta)).$$

(ii) There exists $\kappa, \delta > 0$ such that

$$F(\tilde{x}) \cap \mathbb{B}(\bar{y}, \delta) \subset F(x) + \mathbb{B}(0, \kappa\|\tilde{x} - x\|_X) \quad (x \in \mathbb{B}(\bar{x}, \delta);\ \tilde{x} \in X).$$

(iii) The Aubin property (27.1).

(iv) There exists $\kappa, \delta > 0$ such that

$$\operatorname{dist}(y, F(x)) \leq \kappa \operatorname{dist}(F^{-1}(y) \cap \mathbb{B}(\bar{x}, \delta), x) \quad (x \in \mathbb{B}(\bar{x}, \delta),\ y \in \mathbb{B}(\bar{y}, \delta)).$$

The infimum of $\kappa > 0$ for which each of these characterizations holds is equal to the graphical modulus $\operatorname{lip} F(\bar{x}|\bar{y})$. *(The radius of validity $\delta > 0$ for any given $\kappa > 0$ may be distinct in each of the characterizations, however.)*

Proof. *(i)* $\Leftrightarrow$ *(ii):* Clearly (ii) implies (i) with the same $\kappa, \delta > 0$. To show the implication in the other direction, we start by applying (i) with $\tilde{x} = \bar{x}$, which yields

$$F(\bar{x}) \cap \mathbb{B}(\bar{y}, \delta) \subset F(x) + \mathbb{B}(0, \kappa\|\bar{x} - x\|_X) \quad (x \in \mathbb{B}(\bar{x}, \delta)).$$

Taking $x \in \mathbb{B}(\bar{x}, \delta')$ for some $\delta' \in (0, \delta]$, we thus deduce that

$$\bar{y} \in F(x) + \mathbb{B}(0, \kappa\|\bar{x} - x\|_X) \subset F(x) + \mathbb{B}(0, \kappa\delta').$$

In particular, for any $\varepsilon' > 0$, we have

$$\mathbb{B}(\bar{y}, \varepsilon') \subset F(x) + \mathbb{B}(0, \kappa\delta' + \varepsilon'). \tag{27.5}$$

For $\tilde{x} \in \mathbb{B}(\bar{x}, \delta)$, (ii) is immediate from (i), so we may concentrate on $\tilde{x} \in X \setminus \mathbb{B}(\bar{x}, \delta)$. Then

$$\|\tilde{x} - x\|_X \geq \|\tilde{x} - \bar{x}\|_X - \|\bar{x} - x\|_X \geq \delta - \delta'.$$

If we pick $\varepsilon', \delta' > 0$ such that $\kappa\delta' + \varepsilon' \leq \kappa(\delta - \delta')$, it follows

$$\kappa\delta' + \varepsilon' \leq \kappa\|\tilde{x} - x\|_X.$$

Thus (27.5) gives, as illustrated in Figure 27.4,

$$F(\tilde{x}) \cap \mathbb{B}(\bar{y}, \varepsilon') \subset \mathbb{B}(\bar{y}, \varepsilon') \subset F(x) + \mathbb{B}(0, \kappa\delta' + \varepsilon') \subset F(x) + \kappa\mathbb{B}(0, \|\tilde{x} - x\|_X),$$

which is (ii).

(ii) $\Leftrightarrow$ *(iii):* We expand (ii) as

$$\{\tilde{y}\} \cap \mathbb{B}(\bar{y}, \delta) \subset F(x) + \mathbb{B}(0, \kappa\|\tilde{x} - x\|_X) \quad (\tilde{y} \in F(\tilde{x}); x \in \mathbb{B}(\bar{x}, \delta);\ \tilde{x} \in X).$$

By rearranging and taking the infimum over all $y \in F(x)$, this yields

$$\inf_{y \in F(x)} \|\tilde{y} - y\|_Y \leq \kappa\|\tilde{x} - x\|_X \quad (\tilde{y} \in F(\tilde{x}) \cap \mathbb{B}(\bar{y}, \delta);\ x \in \mathbb{B}(\bar{x}, \delta);\ \tilde{x} \in X).$$

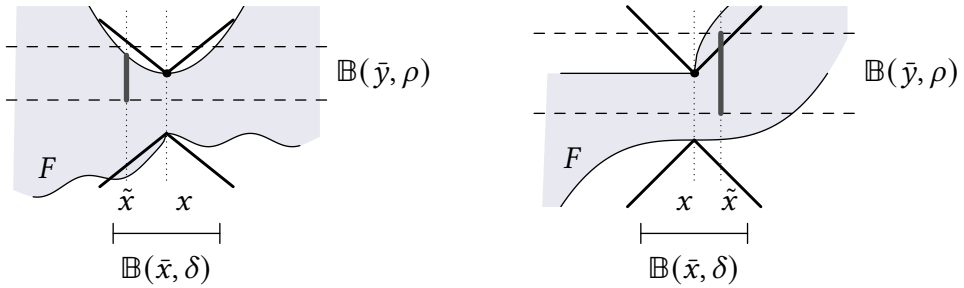
This may further be rewritten as

$$\inf_{y \in F(x)} \|\tilde{y} - y\|_Y \leq \inf_{\tilde{x} \in F^{-1}(\tilde{y})} \kappa\|\tilde{x} - x\|_X \quad (x \in \mathbb{B}(\bar{x}, \delta);\ \tilde{y} \cap \mathbb{B}(\bar{y}, \delta)).$$

Thus (iii) is equivalent to (i).

(iii) $\Rightarrow$ *(iv):* This is immediate from the definition of dist, which yields

$$\operatorname{dist}(F^{-1}(y), x) \leq \operatorname{dist}(F^{-1}(y) \cap \mathbb{B}(\bar{x}, \delta), x).$$



(a) property is satisfied

(b) property is not satisfied

Figure 27.3: Illustration of satisfaction and dissatisfaction of the Aubin property for $x = \bar{x}$. The dashed lines indicate $\mathbb{B}(\bar{y}, \rho)$, and the dot marks $(\bar{x}, \bar{y})$, while the dark gray thick lines indicate $F(\tilde{x}) \cap \mathbb{B}(\bar{y}, \rho)$. It should remain within the bounds of the black thick lines indicating "fat" cone $F(x) + \mathbb{B}(0, \kappa\|\tilde{x} - x\|_X)$. The violation of the bounds at the bottom in (a) does not matter, because we are only interested in the area between the dashed lines.

(iv) $\Rightarrow$ *(i):* We express (iv) as

$$\inf_{\tilde{y} \in F(x)} \|y - \tilde{y}\|_Y \leq \kappa \|\tilde{x} - x\|_X \quad (x \in \mathbb{B}(\bar{x}, \delta),\ y \in \mathbb{B}(\bar{y}, \delta),\ \tilde{x} \in F^{-1}(y) \cap \mathbb{B}(\bar{x}, \delta)).$$

This can be rearranged to imply that

$$\{y\} \subset F(x) + \mathbb{B}(0, \kappa\|\tilde{x} - x\|_X) \quad (x \in \mathbb{B}(\bar{x}, \delta),\ y \in \mathbb{B}(\bar{y}, \delta) \cap F(\tilde{x}),\ \tilde{x} \in \mathbb{B}(\bar{x}, \delta)),$$

which can be further rewritten as

$$F(\tilde{x}) \cap \mathbb{B}(\bar{y}, \delta) \subset F(x) + \mathbb{B}(0, \kappa\|\tilde{x} - x\|_X) \quad (x, \tilde{x} \in \mathbb{B}(\bar{x}, \delta)),$$

yielding (i). □

We have similar characterizations of calmness. The proof is analogous, simply fixing $x = \bar{x}$.

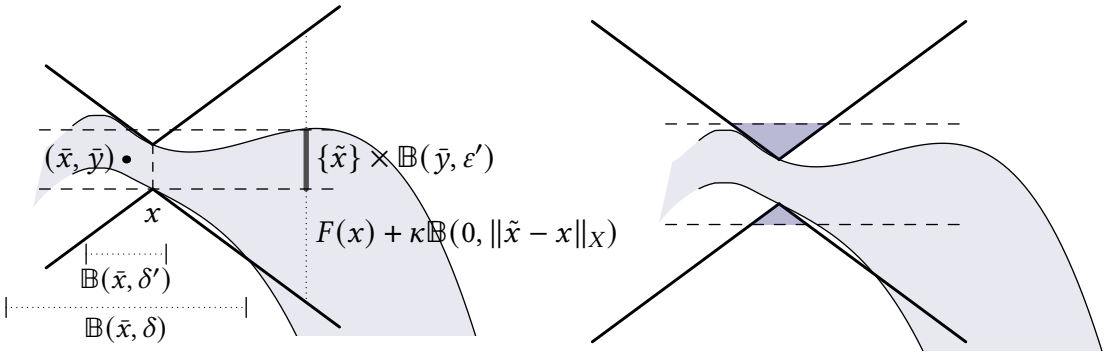
Corollary 27.3. *Let X, Y be Banach spaces and $F : X \rightrightarrows Y$. Then the following are equivalent for $\bar{x} \in X$ and $\bar{y} \in F(\bar{x})$:*

(i) There exists $\kappa, \delta > 0$ such that

$$F(\tilde{x}) \cap \mathbb{B}(\bar{y}, \delta) \subset F(\bar{x}) + \mathbb{B}(0, \kappa\|\tilde{x} - \bar{x}\|_X) \quad (\tilde{x} \in \mathbb{B}(\bar{x}, \delta)).$$

(ii) There exists $\kappa, \delta > 0$ such that

$$F(\tilde{x}) \cap \mathbb{B}(\bar{y}, \delta) \subset F(\bar{x}) + \mathbb{B}(0, \kappa\|\tilde{x} - \bar{x}\|_X) \quad (\tilde{x} \in X).$$



(a) illustration of technique (b) critical areas

Figure 27.4: (a) Illustration of the technique in Theorem 27.2 to prove the equivalence of the two set inclusion formulations of the Aubin property. For $\tilde{x}$ outside the ball $\mathbb{B}(\bar{x}, \delta)$, the set $\mathbb{B}(\bar{y}, \varepsilon')$ indicated by the thick dark gray line, is completely contained in the fat-cone structure $F(x) + \kappa\mathbb{B}(0, \|\tilde{x} - x\|_X)$ of Figure 27.2b, indicated by the thick black and dotted lines. Closer to x, within $\mathbb{B}(\bar{x}, \delta)$, this is not the case, although $F(\tilde{x}) \cap \mathbb{B}(\bar{y}, \varepsilon')$ itself is still contained in the structure. (b) highlights in darker color the areas that are critical for the Aubin property to hold.

(iii) Calmness (27.4).

(iv) There exists $\kappa, \delta > 0$ *such that*

$$\operatorname{dist}(y, F(\bar{x})) \leq \kappa \operatorname{dist}(F^{-1}(y) \cap \mathbb{B}(\bar{x}, \delta), \bar{x}) \quad (y \in \mathbb{B}(\bar{y}, \delta)).$$

The infimum of $\kappa > 0$ *for which each of these characterizations holds is equal to the modulus of calmness* $\operatorname{calm} F(\bar{x}|\bar{y})$. *(The radius of validity* $\delta > 0$ *for any given* $\kappa > 0$ *may be distinct in each of the characterizations, however.)*

27.2 NEIGHBORHOOD-BASED CODERIVATIVE CRITERIA

Our goal is now to relate the Aubin property to "outer norms" of limiting coderivatives, just as the Lipschitz property of differentiable single-valued functions can be related to norms of their derivatives. Before embarking on this in the next section, as a preparatory step we relate in this section the Aubin property to neighborhood-based criteria on Fréchet coderivatives. To this end, we define for a set-valued mapping $F : X \rightrightarrows Y$, $(\bar{x}, \bar{y}) \in \operatorname{graph} F$, and $\delta, \varepsilon > 0$

$$\kappa_\delta^\varepsilon(\bar{x}|\bar{y}) := \sup \left\{ \|x^*\|_{X^*} \;\middle|\; \begin{array}{l} x^* \in \widehat{D}_\varepsilon^* F(x|y)(y^*),\ \|y^*\|_{Y^*} \leq 1, \\ x \in \mathbb{B}(\bar{x}, \delta),\ y \in F(x) \cap \mathbb{B}(\bar{y}, \delta) \end{array} \right\}, \tag{27.6}$$

which measures locally the opening of the cones $\widehat{N}^{\varepsilon}_{\operatorname{graph} F}(x|y)$ around $(\bar{x}, \bar{y})$; for smooth functions and $\varepsilon = 0$, it coincides with the local supremum of $\|DF(x)\|_{\mathbb{L}(X;Y)}$ around $(\bar{x}, F(\bar{x}))$ (cf. Theorem 20.12). The next lemma bounds these openings in terms of the graphical modulus.

Lemma 27.4. *Let X, Y be Banach spaces and $F : X \rightrightarrows Y$. If* graph F *is closed near $(\bar{x}, \bar{y})$, then*

$$\inf_{\delta>0} \kappa^{\delta}_{\delta}(\bar{x}|\bar{y}) \le \inf_{\delta>0} \kappa^{0}_{\delta}(\bar{x}|\bar{y}) \le \operatorname{lip} F(\bar{x}|\bar{y}).$$

Proof. Since $\widehat{D}^{*}_{\varepsilon}F(x|y)(y^*) \subset \widehat{D}^{*}(x|y)(y^*)$, we always have $\kappa^{\delta}_{\delta}(\bar{x}|\bar{y}) \le \kappa^{0}_{\delta}(\bar{x}|\bar{y})$. It hence suffices to prove for any choice of $\varepsilon(\delta) \in [0, \delta]$ that

$$\kappa := \inf_{\delta>0} \kappa^{\delta}_{\varepsilon(\delta)}(\bar{x}|\bar{y}) \le \operatorname{lip} F(\bar{x}|\bar{y}).$$

We may assume that $\operatorname{lip} F(\bar{x}|\bar{y}) < \infty$, since otherwise there is nothing to prove. This implies in particular that the Aubin property holds, so the definition (27.1) yields for any $\kappa' > \operatorname{lip} F(x|y)$ a $\delta' > 0$ such that

$$\inf_{\tilde{y}\in F(\tilde{x})} \|\tilde{y} - y\|_Y \le \kappa' \|\tilde{x} - x\|_X, \quad (y \in F(x) \cap \mathbb{B}(\bar{y}, \delta'),\ \tilde{x} \in \mathbb{B}(\bar{x}, \delta')). \tag{27.7}$$

Pick $\tilde{\kappa} \in (0, \kappa)$ and $\delta \in (0, \delta')$. By the definition of $\kappa^{\varepsilon(\delta)}_{\delta}(\bar{x}|\bar{y})$, there exist $x \in \mathbb{B}(\bar{x}, \delta)$, $y \in F(x) \cap \mathbb{B}(\bar{y}, \delta)$, and $(x^*, -y^*) \in \widehat{N}^{\varepsilon(\delta)}_{\operatorname{graph} F}(x, y)$ such that $\|x^*\|_{X^*} \ge \tilde{\kappa}$ and $\|y^*\|_{Y^*} \le 1$. Theorem 1.4 then yields a $\Delta x \in X$ such that

$$\langle x^*, \Delta x\rangle_X = \|x^*\|_{X^*} \quad \text{and} \quad \|\Delta x\|_X = 1. \tag{27.8}$$

Let $\tau_k \searrow 0$ with $\tau_k \le \delta$ and set $x_k := x + \tau_k \Delta x$. Then taking $\tilde{x} = x_k$ in (27.7), we can take $y_k \in F(x_k)$ such that

$$\liminf_{k\to\infty} \tau_k^{-1} \|y_k - y\|_Y \le \kappa' \|\Delta x\|_X. \tag{27.9}$$

In particular, after passing to a subsequence if necessary, we may assume that $y_k \to y$ strongly in Y. Using (27.8), $\|x^*\|_{X^*} \ge \tilde{\kappa}$, and $\|y^*\|_{Y^*} \le 1$, this leads to

$$\begin{aligned} \limsup_{k\to\infty} \tau_k^{-1} \left(\langle x^*, x_k - x\rangle_X - \langle y^*, y_k - y\rangle_Y\right) & \\ = \limsup_{k\to\infty} \left(\langle x^*, \Delta x\rangle_X - \tau_k^{-1}\langle y^*, y_k - y\rangle_Y\right) & \\ \ge (\|x^*\|_{X^*} - \kappa')\|\Delta x\|_X \ge \tilde{\kappa} - \kappa'. & \end{aligned} \tag{27.10}$$

By (27.9) (for the chosen subsequence) and the construction of x_k, we have

$$\limsup_{k\to\infty} \tau_k^{-1} \|(x_k, y_k) - (x, y)\|_{X\times Y} \le (1 + \kappa')\|\Delta x\|_X = 1 + \kappa'. \tag{27.11}$$

Since $(x^*, -y^*) \in \widehat{N}^{\varepsilon(\delta)}_{\operatorname{graph} F}(x, y)$, the defining equation (18.7) of $\widehat{N}^{\varepsilon(\delta)}_{\operatorname{graph} F}(x, y)$ we have

$$\limsup_{k\to\infty} \frac{\langle x^*, x_k - x\rangle_X - \langle y^*, y_k - y\rangle_Y}{\|(x_k, y_k) - (x, y)\|_{X\times Y}} \leq \varepsilon(\delta). \tag{27.12}$$

Therefore, (27.10), (27.11), and (27.12) together yield

$$(1 + \kappa')\varepsilon(\delta) \geq \tilde{\kappa} - \kappa'.$$

Taking the infimum over $\delta > 0$, it follows that $\tilde{\kappa} \geq \kappa'$. Since $\kappa' > \operatorname{lip} F(\bar{x}|\bar{y})$ and $\tilde{\kappa} < \kappa$ were arbitrary, we obtain $\kappa \leq \operatorname{lip} F(\bar{x}|\bar{y})$ as desired. □

For the next theorem, recall the definition of Gâteaux smooth spaces from Section 17.2.

Theorem 27.5. *Let X, Y be Gâteaux smooth Banach spaces and let $F : X \rightrightarrows Y$ be such that* $\operatorname{graph} F$ *is closed near $(\bar{x}, \bar{y}) \in X \times Y$. Then F has the Aubin property at $\bar{x}$ for $\bar{y}$ if and only if $\kappa^\delta_\delta(\bar{x}|\bar{y}) < \infty$ or $\kappa^0_\delta(\bar{x}|\bar{y}) < \infty$ for some $\delta > 0$. Furthermore, in this case*

$$\inf_{\delta>0} \kappa^\delta_\delta(\bar{x}|\bar{y}) = \inf_{\delta>0} \kappa^0_\delta(\bar{x}|\bar{y}) = \operatorname{lip} F(\bar{x}|\bar{y}). \tag{27.13}$$

Proof. By Lemma 27.4, it suffices to show that

$$\kappa := \inf_{\delta>0} \kappa^\delta_\delta(\bar{x}|\bar{y}) \geq \operatorname{lip} F(\bar{x}|\bar{y}).$$

We may assume that $\operatorname{lip} F(\bar{x}|\bar{y}) > 0$ as otherwise there is nothing to show. Our plan is now to take arbitrary $0 < \tilde{\kappa} < \operatorname{lip} F(\bar{x}|\bar{y})$ and show that $\kappa \geq \tilde{\kappa}$. This implies $\kappa \geq \operatorname{lip} F(\bar{x}|\bar{y})$ as desired.

To show that $\kappa \geq \tilde{\kappa}$, it suffices to show that $\kappa^\delta_\delta(\bar{x}|\bar{y}) \geq \tilde{\kappa}$ for all $\delta > 0$. To do this, for a parameter $t \searrow 0$, we take $\varepsilon_t \searrow 0$ and $(x_t, y_t) \to (\bar{x}, \bar{y})$ as $t \searrow 0$ as well as ε_t-normals $(x^*_t, -y^*_t) \in \widehat{N}^{\varepsilon_t}_{\operatorname{graph} F}(x_t, y_t)$ that satisfy $\liminf_{t\to 0} \|x^*_t\|_{X^*} \geq \tilde{\kappa}$ and $\limsup_{t\to 0} \|y^*_t\|_{Y^*} \leq 1$. By the definition of $\kappa^\delta_\delta(\bar{x}|\bar{y})$ in (27.6), taking for each $\delta > 0$ the index $t > 0$ such that $\max\{\varepsilon_t, \|x_t - \bar{x}\|_X, \|y_t - \bar{y}\|_Y\} \leq \delta$, this shows as claimed that $\kappa^\delta_\delta(\bar{x}|\bar{y}) \geq \tilde{\kappa}$. The rough idea is to construct the ε_t-normals by projecting points not in $\operatorname{graph} F$ back onto this set. There are, however, some technical difficulties along our way. We divide the construction into three steps.

Step 1: setting up the projection problem. Let $0 < \tilde{\kappa} < \operatorname{lip} F(\bar{x}|\bar{y})$. Since then the Aubin property does not hold for $\tilde{\kappa}$, by the characterization of Theorem 27.2 (iv) there exist

$$\tilde{y}_t \in F(\tilde{x}_t) \cap \mathbb{B}(\bar{y}, t) \quad \text{and} \quad \tilde{x}_t, x_t \in \mathbb{B}(\bar{x}, t) \quad \text{for all } t > 0 \tag{27.14}$$

such that

$$\inf_{y_t \in F(x_t)} \|y_t - \tilde{y}_t\|_Y > \tilde{\kappa}\|x_t - \tilde{x}_t\|_X. \tag{27.15}$$

Since $\inf_{y_t \in F(\tilde{x}_t)} \|y_t - \tilde{y}_t\|_Y = 0$, this implies that $x_t \neq \tilde{x}_t$ and $(x_t, \tilde{y}_t) \notin \operatorname{graph} F$. We want to locally project $(x_t, \tilde{y}_t)$ back onto $\operatorname{graph} F$. However, the nondifferentiability of the distance function $\|\cdot - \tilde{y}_t\|_Y$ at $\tilde{y}_t$ would cause difficulties, so – similarly to the proof of Lemma 18.13 – we modify the projection by composing the norm with the "smoothing function"

$$\varphi_\mu(r) := \sqrt{\mu^2 + r^2} - \mu. \tag{27.16}$$

By Theorems 4.5, 4.6, and 4.19 and the assumed differentiability of $\|\cdot\|_Y$ away from the origin, $\varphi_\mu(\|\cdot\|_Y)$ is convex and has a single-valued subdifferential mapping with elements of norm less than one. Hence this smoothed distance function is Gâteaux differentiable by Lemma 13.7. Due to (27.16), for every $t > 0$ and $\mu_t > 0$, we further have

$$\|y - \tilde{y}_t\|_Y - \mu_t \leq \varphi_{\mu_t}(\|y - \tilde{y}_t\|_Y) \leq \|y - \tilde{y}_t\|_Y \quad (y \in Y). \tag{27.17}$$

To locally project $(x_t, \tilde{y}_t)$ onto $\operatorname{graph} F$, we thus seek to minimize the function

$$\psi_t(x, y) := \delta_{C_t}(x, y) + \tilde{\kappa}\|x - x_t\|_X + \varphi_{\mu_t}(\|y - \tilde{y}_t\|_Y) \tag{27.18}$$

for

$$C_t := [\mathbb{B}(\bar{x}, t + 2\tilde{\kappa}) \times \mathbb{B}(\bar{y}, t + 2\tilde{\kappa})] \cap \operatorname{graph} F.$$

Clearly, ψ_t is bounded from below by $-\mu_t$ as well as coercive since C_t is bounded. If t is is small enough, then C_t is closed by the local closedness of $\operatorname{graph} F$. Therefore ψ_t is lower semicontinuous (but not weakly lower semicontinuous since $\operatorname{graph} F$ need not be convex).

Step 2: finding approximate minimizers. We would like to find a minimizer of ψ_t, but the lack of weak lower semicontinuity prevents the use of Tonelli's direct method of Theorem 2.1. We therefore use Ekeland's variational principle (Theorem 2.16) to find an approximate minimizer. Towards this end, choose for every $t > 0$

$$\mu_t := t^{-1/2}\tilde{\kappa}\|\tilde{x}_t - x_t\|_X^2 \leq \tilde{\kappa}t^{1/2}\|\tilde{x}_t - x_t\|_X \quad \text{and} \quad \lambda_t := \|\tilde{x}_t - x_t\|_X + t^{1/2} \leq t + t^{1/2}, \tag{27.19}$$

where the inequalities hold due to (27.14). Then

$$\psi_t(\tilde{x}_t, \tilde{y}_t) = \tilde{\kappa}\|\tilde{x}_t - x_t\|_X \leq (\tilde{\kappa}\|\tilde{x}_t - x_t\|_X + \mu_t) + \inf \psi_t. \tag{27.20}$$

Therefore, applying Theorem 2.16 for $\lambda = \lambda_t$ and

$$\varepsilon = \tilde{\kappa}\|\tilde{x}_t - x_t\|_X + \mu_t = \tilde{\kappa}\|\tilde{x}_t - x_t\|_X t^{-1/2}\lambda_t = \frac{\mu_t \lambda_t}{\|\tilde{x}_t - x_t\|_X},$$

we obtain for each $t > 0$ a strict minimizer $(\bar{x}_t, \bar{y}_t)$ of

$$\tilde{\psi}_t(x, y) := \psi_t(x, y) + \frac{\mu_t}{\|\tilde{x}_t - x_t\|_X}(\|x - \bar{x}_t\|_X + \|y - \bar{y}_t\|_Y) \tag{27.21a}$$

with

$$\psi_t(\bar{x}_t, \bar{y}_t) + \frac{\mu_t}{\|\tilde{x}_t - x_t\|_X}(\|\tilde{x}_t - \bar{x}_t\|_X + \|\tilde{y}_t - \bar{y}_t\|_Y) \leq \psi_t(\tilde{x}_t, \tilde{y}_t) = \kappa\|\tilde{x}_t - x_t\|_X \tag{27.21b}$$

and

$$\|\bar{x}_t - \tilde{x}_t\|_X + \|\bar{y}_t - \tilde{y}_t\|_Y \le \lambda_t. \tag{27.21c}$$

We claim that $\bar{x}_t \neq x_t$, which we show by contradiction. Assume therefore that $\bar{x}_t = x_t$. Then $\bar{y}_t \in F(x_t)$, and (27.17) yields

$$\psi_t(x_t, \bar{y}_t) = \varphi_{\mu_t}(\|\bar{y}_t - \tilde{y}_t\|_Y) \ge \|\bar{y}_t - \tilde{y}_t\|_Y - \mu_t.$$

Thus by (27.20) and (27.21b),

$$\|\bar{y}_t - \tilde{y}_t\|_Y \le \|\bar{y}_t - \tilde{y}_t\|_Y - \mu_t + \frac{\mu_t}{\|\tilde{x}_t - x_t\|_X}\left(\|\tilde{x}_t - x_t\|_X + \|\tilde{y}_t - \bar{y}_t\|_Y\right) \le \tilde{\kappa}\|\tilde{x}_t - x_t\|_X.$$

But this contradicts (27.15) as $\bar{y}_t \in F(x_t)$.

Step 3: constructing ε-normals. We are now ready to construct the desired ε-normals. We write

$$\tilde{\psi}_t(x_t, y_t) = \delta_{C_t}(x, y) + F(x, y) \tag{27.22}$$

for the convex and Lipschitz continuous function

$$F(x, y) := \tilde{\kappa}\|x - x_t\|_X + \varphi_{\mu_t}(\|y - \tilde{y}_t\|_Y) + \frac{\mu_t}{\|\tilde{x}_t - x_t\|_X}\left(\|x - \tilde{x}_t\|_X + \|y - \tilde{y}_t\|_Y\right).$$

Since we assume X to be Gâteaux smooth, $x \mapsto \tilde{\kappa}\|x - x_t\|_Y$ is Gâteaux differentiable at $\tilde{x}_t \neq x_t$. Furthermore, $y \mapsto \varphi_{\mu_t}(\|y - \tilde{y}_t\|_Y)$ is by construction Gâteaux differentiable for all y. By (27.19), we have $\frac{\mu_t}{\|\tilde{x}_t - x_t\|_X} \le t^{1/2}\tilde{\kappa}$. Since $x_t \neq \bar{x}_t$, Theorems 4.6, 4.14, and 4.19 now yield

$$\partial F(\bar{x}_t, \bar{y}_t) \subset \mathbb{B}((-x_t^*, y_t^*), t^{1/2}\tilde{\kappa}) \quad \text{for} \quad \begin{cases} -x_t^* = \tilde{\kappa} D[\|\cdot - x_t\|_X](\bar{x}_t), \\ y_t^* = D[\varphi'_{\mu_t}(\|\cdot - \tilde{y}_t\|_Y)](\bar{y}_t). \end{cases} \tag{27.23}$$

Since $\bar{x}_t \neq x_t$, we have $\|x_t^*\|_{X^*} = \tilde{\kappa}$ by Theorem 4.6. Moreover, $\|y_t^*\|_{Y^*} \le 1$ as observed in Step 1. Theorem 16.2 further yields $0 \in \partial_F \tilde{\psi}_t(\bar{x}_t, \bar{y}_t)$.

Due to (27.22) and (27.23), Lemma 17.2 now shows that

$$(x_t^*, -y_t^*) \in \widehat{N}_{C_t}^{\varepsilon_t}(\bar{x}_t, \bar{y}_t), \quad \text{i.e.,} \quad x_t^* \in \widehat{D}_{\varepsilon_t}^* F(\bar{x}_t|\bar{y}_t)(y_t^*) \quad \text{for} \quad \varepsilon_t := t^{1/2}\tilde{\kappa}.$$

We illustrate this construction in Figure 27.5. Since $\lambda_t \le t + t^{1/2}$ by (27.19), it follows from (27.21c) that $\|\bar{x}_t - \bar{x}\|_X, \|\bar{y}_t - \bar{y}\|_Y \le 2t + t^{1/2}$ and hence that $(\bar{x}_t, \bar{y}_t) \to (\bar{x}, \bar{y})$ as $t \searrow 0$. We also have both $\liminf_{t\searrow 0} \|x_t^*\|_{X^*} \ge \tilde{\kappa}$ and $\limsup_{t\searrow 0} \|y_t^*\|_{Y^*} \le 1$. Thus we have constructed the desired sequence of ε_t-normals. □

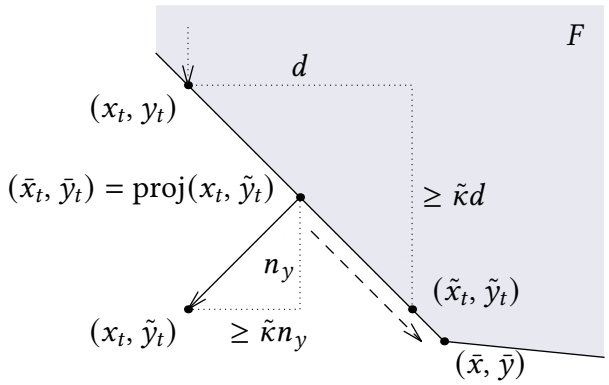


Figure 27.5: The construction in the final part of the proof of Theorem 27.5. The dotted arrow indicates how y_t minimizes the distance to $\tilde{y}_t$ within $F(x_t)$, which ensures that $\|y_t - \tilde{y}_t\|_Y \geq \tilde{\kappa} d$ for $d := \|x_t - \tilde{x}_t\|_X$. The point $(x_t, \tilde{y}_t)$ is outside graph F; when projected back as $(\bar{x}_t, \bar{y}_t)$, the normal vector to graph F indicated by the solid arrow has x-component larger than the y-component n_y by the factor $\tilde{\kappa}$. The dashed arrow indicates the convergence of the other points to $(\bar{x}, \bar{y})$ as $t \searrow 0$.

Remark 27.6. Our proof of Theorem 27.5 differs from those in [Mordukhovich, 2006, 2018] by the specific construction of the point $(\tilde{y}_t, x_t) \notin \operatorname{graph} F$ and the use of the smoothed distance $\varphi_{\varepsilon_t}(\|\cdot\|_X)$. In contrast, the earlier proofs first translate the Aubin property (or metric regularity) into a *covering* or *linear openness* property to construct the point outside graph F that is to be projected back onto this set. In finite dimensions, [Mordukhovich, 2018] develops calculus for the limiting subdifferential of Section 16.3 to avoid the lack of calculus for the Fréchet subdifferential; we instead apply the fuzzy calculus of Lemma 17.2 to the smoothed distance function $\varphi_{\varepsilon_t}(\|\cdot\|_X)$. A further alternative in finite dimensions involves the proximal subdifferentials used in [Rockafellar and Wets, 1998]. In infinite dimensions, [Mordukhovich, 2006] develops advanced extremal principles to work with the Fréchet subdifferential.

Remark 27.7 (relaxation of Gâteaux smoothness). The assumption that Y (or, with somewhat more work, X) is Gâteaux smooth in Theorem 27.5 may be replaced with the assumption of the existence of a family $\{\theta_\mu : Y \to \mathbb{R}\}_{\mu>0}$ of Gâteaux differentiable norm approximations satisfying

$$\|y\|_Y - \mu \leq \theta_\mu(y) \leq \|y\|_Y \quad (y \in Y).$$

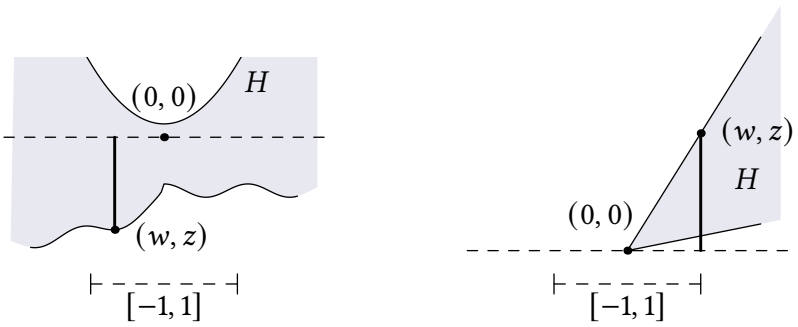
Then (27.17) holds with $\theta_{\mu_t}(y - \tilde{y}_t)$ in place of $\varphi_{\mu_t}(\|y - \tilde{y}_t\|_Y)$. For example, with φ_μ as in (27.16), in $L^p(\Omega)$ we can set

$$\theta_\mu(y) := \|\varphi_\mu(|y(\xi)|)\|_{L^p(\Omega)} \quad (y \in L^1(\Omega)).$$

With somewhat more effort, the Gâteaux smoothness of X can be similarly relaxed.

27.3 POINT-BASED CODERIVATIVE CRITERIA

We will now convert the neighborhood-based criterion of Lemma 27.4 and Theorem 27.5 into a simpler point-based criterion. For the statement, we need to introduce a new smaller



(a) general set-valued mapping H (b) mapping whose graph H is a cone

Figure 27.6: Points (w, z) achieving the supremum in the expression of the outer norm $|H|^+$.

coderivative of $F : X \rightrightarrows Y$ at x for y, the *mixed (limiting) coderivative* $D^*_M F(x|y) : Y^* \rightrightarrows X^*$,

$$D^*_M F(x|y)(y^*) := \text{w-}*\text{-}\limsup_{\substack{(\tilde{x},\tilde{y})\to(x,y)\\ \tilde{y}^*\to y^*,\, \varepsilon\searrow 0}} \widehat{D}^*_\varepsilon F(\tilde{x}|\tilde{y})(\tilde{y}^*), \tag{27.24}$$

which differs from the "normal" coderivative

$$D^* F(x|y)(y^*) = \text{w-}*\text{-}\limsup_{\substack{(\tilde{x},\tilde{y})\to(x,y)\\ \tilde{y}^*\xrightarrow{*} y^*,\, \varepsilon\searrow 0}} \widehat{D}^*_\varepsilon F(\tilde{x}|\tilde{y})(\tilde{y}^*), \tag{27.25}$$

by the use of weak-$*$ convergence in X^* and strong convergence in Y^* instead of weak-$*$ convergence in both. (The mixed coderivative is not obtained directly from any of the usual normal cones, although one can naturally define corresponding mixed normal cones on product spaces.)

We further define for any $H : W \rightrightarrows Z$ the *outer norm*

$$|H|^+ := \sup\{\|z\|_Z \mid z \in H(w),\ \|w\|_W \le 1\}.$$

We illustrate the outer norm by two examples in Figure 27.6. We are mainly interested in the outer norms of coderivatives, in particular of

$$|D^*_M F(\bar{x}|\bar{y})|^+ = \sup\{\|\bar{x}^*\|_{X^*} \mid \bar{x}^* \in D^*_M F(\bar{x}|\bar{y})(\bar{y}^*),\ \|\bar{y}^*\|_{Y^*} \le 1\}. \tag{27.26}$$

Recalling Theorem 18.5, we have

$$\widehat{D}^* F(x|y)(y^*) \subset D^*_M F(x|y)(y^*) \subset D^* F(x|y)(y^*), \tag{27.27}$$

so the outer norms satisfy

$$|D^*_M F(\bar{x}|\bar{y})|^+ \le |D^* F(\bar{x}|\bar{y})|^+.$$

We say that F is *coderivatively normal* at $\bar{x}$ for $\bar{y}$ if $|D^*_M F(\bar{x}|\bar{y})|^+ = |D^* F(\bar{x}|\bar{y})|^+$. Of course, if Y is finite-dimensional, then $D^*_M F(x|y) = D^* F(x|y)$ and thus F is always coderivatively normal. Note that $|D^* F(\bar{x}|\bar{y})|^+$ can be directly related to the neighborhood-based κ^δ_δ defined in (27.6). In particular, it measures the opening of the cone $N_{\operatorname{graph} F}(\bar{x}, \bar{y})$; compare Figure 27.6b.

As the central result of this chapter, we now use this connection to derive a characterization of the Aubin property and the graphical modulus (and hence also of metric regularity and the modulus of metric regularity) through the outer norm of the mixed limiting coderivative. This *Mordukhovich criterion* generalizes the classical relation between the Lipschitz constant of a C^1 function and the norm of its derivative.

Lemma 27.8 (Mordukhovich criterion in general Banach spaces). *Let X, Y be Banach spaces and let $F : X \rightrightarrows Y$ be such that* graph F *is closed near $(\bar{x}, \bar{y}) \in X \times Y$. If F has the Aubin property at $\bar{x}$ for $\bar{y}$, then*

$$D^*_M F(\bar{x}|\bar{y})(0) = \{0\} \tag{27.28}$$

and

$$|D^*_M F(\bar{x}|\bar{y})|^+ \leq \operatorname{lip} F(\bar{x}|\bar{y}). \tag{27.29}$$

Proof. As the first step, we show that the Aubin property implies (27.29) and hence that $\kappa := |D^*_M F(\bar{x}|\bar{y})|^+ < \infty$. Let $\rho > 0$. By the definition of $D^*_M F(\bar{x}|\bar{y})$ in (27.24), there then exist $\delta \in (0, \rho)$, $x \in \mathbb{B}(\bar{x}, \rho)$, and $y \in F(x) \cap \mathbb{B}(\bar{y}, \rho)$ as well as $y^* \in Y^*$ and $x^* \in \widehat{D}^*_\delta F(x|y)(y^*)$ such that $\|y^*\|_{Y^*} \leq 1 + \rho$ and $\|x^*\|_{X^*} \geq \kappa(1-\rho)^2$. (The upper bound on $\|y^*\|_{Y^*}$ is why we need the *mixed* coderivative, since $\|\cdot\|_{Y^*}$ is continuous only in the strong topology. For the lower bound on $\|x^*\|_{X^*}$, in contrast, the weak-$*$ lower semicontinuity of $\|\cdot\|_{X^*}$ is sufficient.) Since $\widehat{D}^*_\delta F(x|y)$ is formed from a cone, we may divide x^* and y^* by $1+\rho$ and thus assume that $\|y^*\|_{Y^*} \leq 1$ and $\|x^*\|_{X^*} \geq \kappa(1-\rho)$. Consequently

$$\kappa(1-\rho) \leq \kappa^\delta_\delta(\bar{x}|\bar{y}) = \sup \left\{ \|x^*\|_{X^*} \,\middle|\, \begin{array}{l} x^* \in \widehat{D}^*_\delta F(x|y)(y^*),\ \|y^*\|_{Y^*} \leq 1, \\ x \in \mathbb{B}(\bar{x}, \delta),\ y \in F(x) \cap \mathbb{B}(\bar{y}, \delta) \end{array} \right\}.$$

Taking the infimum over $\delta > 0$ and letting $\rho \searrow 0$ thus shows

$$\kappa \leq \inf_{\delta > 0} \kappa^\delta_\delta(\bar{x}|\bar{y}).$$

It now follows from Lemma 27.4 that $\kappa \leq \operatorname{lip} F(\bar{x}|\bar{y})$, which yields (27.29).

As the second step, we prove that the Aubin property implies (27.28). We argue by contraposition. First, note that since graph $D^*_M F(x|y)$ is a cone, $0 \in D^*_M F(x|y)(0)$. Hence if (27.28) does not hold, there exists $x^* \in X^* \setminus \{0\}$ such that

$$x^*[0, \infty) \subset D^*_M F(x|y)(0).$$

By (27.26) and the first step, this implies that $\infty = \kappa \leq \operatorname{lip} F(\bar{x}|\bar{y})$ and hence that the Aubin property of F at $\bar{x}$ for $\bar{y}$ is violated. □

Applied to F^{-1}, we obtain a corresponding result for metric regularity.

Corollary 27.9 (Mordukhovich criterion for metric regularity in general Banach spaces). *Let X, Y be Banach spaces and let $F : X \rightrightarrows Y$ be such that* graph F *is closed near $(\bar{x}, \bar{y}) \in X \times Y$. If F is metrically regular at $(\bar{x}, \bar{y})$, then*

$$0 \in D^*_M F(\bar{x}|\bar{y})(y^*) \Rightarrow y^* = 0 \tag{27.30}$$

and

$$|D^*_M F^{-1}(\bar{y}|\bar{x})|^+ \leq \operatorname{reg} F(\bar{x}|\bar{y}). \tag{27.31}$$

Proof. We apply Lemma 27.8 to F^{-1}, observing that (27.28) applied to F^{-1} is (27.30). □

Under stronger assumptions on the spaces and the set-valued mapping, we obtain equivalence. For the following theorem, recall the definition of partial sequential normal compactness (PSNC) from Section 25.2.

Theorem 27.10 (Mordukhovich criterion in smooth Banach spaces). *Let X, Y be Gâteaux smooth Banach spaces with X reflexive and let $F : X \rightrightarrows Y$ be such that* graph F *is closed near $(\bar{x}, \bar{y}) \in X \times Y$. If F is PSNC at $\bar{x}$ for $\bar{y}$, then the following are equivalent:*

(i) the Aubin property of F at $\bar{x}$ for $\bar{y}$;

(ii) the implication (27.28);

(iii) $|D^*_M F(\bar{x}|\bar{y})|^+ < \infty$.

Proof. Due to Lemma 27.8, it suffices to show that (iii) ⇒ (ii) ⇒ (i). We start with the second implication. Since X and Y are Gâteaux smooth, Theorem 27.5 yields

$$\operatorname{lip} F(\bar{x}|\bar{y}) = \tilde{\kappa} := \inf_{\delta>0} \sup \left\{ \|x^*\|_{X^*} \,\middle|\, \begin{array}{l} x^* \in \widehat{D}^*_\delta F(x|y)(y^*),\ \|y^*\|_{Y^*} \leq 1, \\ x \in \mathbb{B}(\bar{x}, \delta),\ y \in F(x) \cap \mathbb{B}(\bar{y}, \delta) \end{array} \right\} \tag{27.32}$$

and that the Aubin property holds if $\tilde{\kappa} < \infty$. We now argue by contradiction. Assume that the Aubin property does not hold. Then $\tilde{\kappa} = \infty$ and hence we can find $(x_k, y_k) \to (\bar{x}, \bar{y})$, $\varepsilon_k \searrow 0$, and $x^*_k \in \widehat{D}^*_{\varepsilon_k} F(x_k|y_k)(y^*_k)$ with $\|y^*_k\|_{Y^*} \leq 1$ and $\|x^*_k\|_{X^*} \to \infty$. In particular, $y^*_k / \|x^*_k\|_{X^*} \to 0$. Since X is reflexive, we can apply the Eberlein–Šmulyan Theorem 1.9 to extract a subsequence (not relabelled) such that $x^*_k / \|x^*_k\|_{X^*} \overset{*}{\rightharpoonup} x^*$ for some $x^* \in X^*$. Since graph $\widehat{D}^*_{\varepsilon_k} F(x_k|y_k)$ is a cone, we also have

$$x^*_k / \|x^*_k\|_{X^*} \in \widehat{D}^*_{\varepsilon_k} F(x_k|y_k)(y^*_k / \|x^*_k\|_{X^*}).$$

By the definition (27.24) of the mixed coderivative, we deduce that $x^* \in D^*_M F(\bar{x}|\bar{y})(0)$. We now make a case distinction: If $x^* \neq 0$, then this contradicts the qualification condition (27.28). On the other hand, if $x^* = 0$, the PSNC of F at $\bar{x}$ for $\bar{y}$, implies that $1 = \|x^*_k/\|x^*_k\|_{X^*}\|_{X^*} \to 0$, which is also a contradiction. Therefore (27.28) implies the Aubin property.

It remains to show that (iii) $\Rightarrow$ (ii). First, since graph $D^*_M F(\bar{x}|\bar{y})$ is a cone, $D^*_M F(\bar{x}|\bar{y})(0)$ is a cone as well. Hence by (27.26), $|D^*_M F(\bar{x}|\bar{y})|^+ < \infty$ implies that $D^*_M F(\bar{x}|\bar{y})(0) = \{0\}$, which is (27.28). □

Again, applying Theorem 27.10 to F^{-1} yields a characterization of metric regularity.

Corollary 27.11 (Mordukhovich criterion for metric regularity in smooth Banach spaces). *Let X, Y be Gâteaux smooth Banach spaces with X reflexive and let $F : X \rightrightarrows Y$ be such that* graph F *is closed near $(\bar{x}, \bar{y}) \in X \times Y$. If F^{-1} is PSNC at $\bar{y}$ for $\bar{x}$, then the following are equivalent:*

(i) the metric regularity of F at $(\bar{x}, \bar{y})$;

(ii) the implication (27.30);

(iii) $|D^*_M F^{-1}(\bar{y}|\bar{x})|^+ < \infty$.

Remark 27.12 (separable and Asplund spaces). The reflexivity of X (resp. Y) was used to obtain the weak-$*$ compactness of the unit ball in X^* via the Eberlein–Šmulyan Theorem 1.9 applied to X^*. Alternatively, this can be obtained by assuming separability of X and using the Banach–Alaoglu Theorem 1.11. More generally, dual spaces of Asplund spaces have weak-$*$-compact unit balls; we refer to [Mordukhovich, 2006] for the full theory in Asplund spaces.

In finite dimensions, we have a full characterization of the graphical modulus via the outer norm of the limiting coderivative (which here coincides with the mixed coderivative).

Corollary 27.13 (Mordukhovich criterion for the graphical modulus in finite dimensions). *Let X, Y be finite-dimensional Gâteaux smooth Banach spaces and let $F : X \rightrightarrows Y$ be such that* graph F *is closed near $(\bar{x}, \bar{y}) \in X \times Y$. Then*

$$\operatorname{lip} F(\bar{x}|\bar{y}) = |D^* F(\bar{x}|\bar{y})|^+.$$

Proof. Due to Lemma 27.8, we only have to show that

$$\operatorname{lip} F(\bar{x}|\bar{y}) \leq |D^* F(\bar{x}|\bar{y})|^+. \tag{27.33}$$

As in the proof of Theorem 27.10, the smoothness of X and Y allows applying Theorem 27.5 to obtain that $\operatorname{lip} F(\bar{x}|\bar{y}) = \tilde{\kappa}$ given by (27.32). It therefore suffices to show that $\tilde{\kappa} \leq |D^* F(\bar{x}|\bar{y})|^+$. Let $\kappa' < \tilde{\kappa}$ be arbitrary. By (27.32), we can then find $(x_k, y_k) \to (\bar{x}, \bar{y})$ and

(a) property is satisfied (b) property is not satisfied

Figure 27.7: Illustration of $|D^*F(x|y)|^+ = \sup\{\|x^*\|_{X^*} \mid (x^*, -y^*) \in N_{\operatorname{graph} F}(x,y), \|y^*\|_{Y^*} \leq 1\}$, where the arrows denote the directions contained in the normal cone. In (a), $-y^* \in [0,\infty)$ but $x^* = 0$, hence $|D^*F(x|y)|^+ = 0$ and the Aubin property is satisfied. In (b), we can take for $y^* = 0$ any $x^* \in (-\infty, 0]$, hence $|D^*F(x|y)|^+ = \infty$ and the Aubin property is not satisfied.

$\varepsilon_k \searrow 0$ as well as $x_k^* \in \widehat{D}^*_{\varepsilon_k} F(x_k|y_k)(y_k^*)$ with $\|y_k^*\|_{Y^*} \leq 1$, and $\tilde{\kappa} \geq \|x_k^*\| \geq \kappa'$. Since X and Y are finite-dimensional, we can apply the Heine–Borel Theorem to extract *strongly* converging subsequences (not relabelled) such that $x_k^* \to x^*$ with $\|x^*\|_{X^*} \geq \kappa'$ and $y_k^* \to y^*$ with $\|y^*\|_{Y^*} \leq 1$. Since strongly converging sequences also converge weakly-$*$, the expression (27.25) for the normal coderivative implies that $x^* \in D^*F(\bar{x}|\bar{y})(y^*)$ and that $|D^*F(\bar{x}|\bar{y})|^+ \geq \|x^*\|_{X^*} \geq \kappa'$. Since $\kappa' < \tilde{\kappa}$ was arbitrary, we obtain (27.33). □

We illustrate in Figure 27.7 how the outer norm of the coderivative relates to the Aubin property.

Corollary 27.14 (Mordukhovich criterion for the modulus of metric regularity in finite dimensions). *Let X, Y be finite-dimensional Gâteaux smooth Banach spaces and let $F : X \rightrightarrows Y$ be such that* graph F *is closed near $(\bar{x}, \bar{y}) \in X \times Y$. Then*

$$\operatorname{reg} F(\bar{x}|\bar{y}) = |D^*F(\bar{x}|\bar{y})^{-1}|^+.$$

Proof. By Lemma 20.5, we have

$$\begin{aligned} |D^*F^{-1}(\bar{y}|\bar{x})|^+ &= \sup\{\|y^*\|_{Y^*} \mid -y^* \in D^*F^{-1}(\bar{y}|\bar{x})(-x^*),\ \|x^*\|_{X^*} \leq 1\} \\ &= \sup\{\|y^*\|_{Y^*} \mid x^* \in D^*F(\bar{x}|\bar{y})(y^*),\ \|x^*\|_{X^*} \leq 1\} \\ &= |[D^*F(\bar{x}|\bar{y})]^{-1}|^+. \end{aligned}$$

The claim now follows by applying Corollary 27.13 to F^{-1} together with Corollary 27.9. □

Remark 27.15. Derivative-based characterizations of calmness and metric subregularity are significantly more involved than those of the Aubin property and metric regularity discussed above. We refer to [Gfrerer, 2011; Gfrerer and Outrata, 2016; Henrion et al., 2002; Zheng and Ng, 2010] to a few characterizations in special cases.

To close this section, we relate the Mordukhovich criterion to the classical inverse function theorem (Theorem 2.8).

Corollary 27.16 (inverse function theorem). *Let X, Y be reflexive and Gâteaux smooth Banach spaces and let $F : X \to Y$ be continuously differentiable around $\bar{x} \in X$. If $F'(\bar{x})^* \in \mathbb{L}(Y^*; X^*)$ has a left-inverse $F'(\bar{x})^{*\dagger} \in \mathbb{L}(X^*; Y^*)$, then there exist $\kappa > 0$ and $\delta > 0$ such that for all $y \in \mathbb{B}(F(\bar{x}), \delta)$ there exists a single-valued selection $J(y) \in F^{-1}(y)$ with*

$$\|\bar{x} - J(y)\|_X \leq \kappa \|F(\bar{x}) - y\|_Y.$$

Proof. Let $\bar{y} := F(\bar{x})$. By Theorem 20.12 and the reflexivity of X and Y,

$$D^*F(\bar{x}|\bar{y}) = \widehat{D}^*F(\bar{x}|\bar{y}) = \{F'(\bar{x})^*\}. \tag{27.34}$$

We have both $D^*F^{-1}(\bar{y}|\bar{x}) = [D^*F(\bar{x}|\bar{y})]^{-1}$ and $\widehat{D}^*F^{-1}(\bar{y}|\bar{x}) = [\widehat{D}^*F(\bar{x}|\bar{y})]^{-1}$ by Lemma 20.5. Due to (27.27), this then implies that $D^*_M F^{-1}(\bar{y}|\bar{x}) \subset D^*F^{-1}(\bar{y}|\bar{x}) = [D^*F(\bar{x}|\bar{y})]^{-1}$. The existence of a left-inverse implies that $F'(\bar{x})^*$ is injective, which together with (27.34) yields (27.30).

By the continuity of F, graph F^{-1} is closed near $(\bar{y}, \bar{x})$. By Lemma 25.6, F^{-1} is PSNC at $\bar{y}$ for $\bar{x}$. Consequently, Corollary 27.14 shows that F is metrically regular at $\bar{x}$ for $\bar{y}$. By the definition (27.2) of metrical regularity, there thus exists for any $\tilde{\kappa} > \operatorname{reg} F(\bar{x}|\bar{y})$ a $\delta > 0$ such that

$$\inf_{\tilde{x} \in F^{-1}(y)} \|x - \tilde{x}\|_X \leq \tilde{\kappa} \|F(x) - y\|_Y \quad (x \in \mathbb{B}(\bar{x}, \delta), y \in \mathbb{B}(\bar{y}, \delta)).$$

Taking in particular $x = \bar{x}$ yields

$$\inf_{\tilde{x} \in F^{-1}(y)} \|\bar{x} - \tilde{x}\|_X \leq \tilde{\kappa} \|F(\bar{x}) - y\|_Y \quad (y \in \mathbb{B}(F(\bar{x}), \delta)).$$

Although the infimum might not be attained, this implies that we can take arbitrary $\kappa > \tilde{\kappa}$ to obtain for any $y \in \mathbb{B}(F(\bar{x}), \delta)$ the existence of some $J(y) := \tilde{x} \in F^{-1}(y)$ satisfying $\|\bar{x} - \tilde{x}\|_X \leq \kappa \|F(\bar{x}) - y\|_Y$, which is the claim. □

28 STABILITY WITH RESPECT TO PERTURBATIONS

We now apply the Lipschitz-like properties of Chapter 27 to study the stability of optimization problems under perturbations. As a motivating problem, we recall the introductory problem (P) and consider the mapping

$$j(x; y, \alpha) := \frac{1}{2}\|Ax - y\|_Y^2 + \alpha g(x).$$

Assuming that a minimizer $\bar{x} = x(y, \alpha)$ of $x \mapsto j(x, y, \alpha)$ exists, we can ask further questions about *stability*, i.e., the dependence of $\bar{x}$ on y and α, in particular whether $\bar{x}$ depends (Lipschitz-)continuously on these parameters. This is of particular relevance in *inverse problems*, which study the solution of ill-posed operator equations $Ax = y$ via families of approximate *well-posed* problems. The central question of *regularization theory* is whether $x(\tilde{y}, \alpha)$ converges to a solution $\hat{x}$ of the operator equation $A\hat{x} = \hat{y}$ as $\hat{y} \to y$ and $\alpha \to 0$.

We study the question of stability in Section 28.1. After deriving in Section 28.2 a convenient characterization of the metric subregularity of convex subdifferentials, we prove the convergence of minimizers in the sense of regularization theory.

28.1 STABILITY WITH RESPECT TO PERTURBATIONS

Let X, P be Banach spaces and $f : X \times P \to \overline{\mathbb{R}}$. We then consider for some parameter $\bar{p} \in P$ the *parametric optimization problem*

$$\min_{x \in X} f(x; \bar{p})$$

and study how a minimizer (or critical point) $\bar{x} \in X$ behaves under perturbations of $\bar{p}$. For this purpose, we introduce the set-valued *solution mapping* (or, if $x \mapsto f(x; p)$ is not convex, *critical point mapping*)

$$S : P \rightrightarrows X, \quad S(p) := \{x \in X \mid 0 \in \partial f(x; p)\}, \tag{28.1}$$

where ∂ is a suitable (convex or Clarke) subdifferential with respect to x for fixed p. We apply the concepts from Section 27.1 to this problem. Specifically, if S has the Aubin property at $\bar{p}$ for $\bar{x}$, then we can take $y = \bar{y}$ in (27.1) to obtain

$$\inf_{x \in S(p)} \|\bar{x} - x\|_X \leq \kappa \|p - \bar{p}\|_P \quad (p \in \mathbb{B}(\bar{p}, \delta))$$

for some $\delta, \kappa > 0$. In other words, the Aubin property of the solution mapping S at $\bar{p}$ for $\bar{x}$ implies the local Lipschitz stability of solutions $x = S(p)$ under perturbations p around the parameter $\bar{p}$. This of course begs the question when a solution mapping has the Aubin property.

We start with a simple special case. Returning to the motivation at the beginning of this chapter, $w \in \partial f(\tilde{x})$ is of course equivalent to $0 \in \partial f(\tilde{x}) - \{w\} = \partial(f - \langle w, \cdot \rangle_X)(\tilde{x})$ since continuous linear mappings are differentiable. Such a perturbation of f is called a *tilt perturbation*, with $w \in X^*$ called *tilt parameter*.

To make this more precise, let $g : X \to \mathbb{R}$ be locally Lipschitz. For a tilt parameter $p \in X^*$, we then define

$$f(x; p) = g(x) - \langle p, x \rangle_X \tag{28.2}$$

and refer to the stability of minimizers (or critical points) of f with respect to p as *tilt stability*. By Theorems 13.4 and 13.20, the solution mapping for f is

$$S(p) = \{x \in X \mid p \in \partial_C g(x)\} = (\partial g_C)^{-1}(p),$$

which thus has the Aubin property – and f is tilt-stable – if and only if $\partial_C g$ is metrically regular at $\bar{x}$ for 0, i.e., by (27.2) that there exist $\kappa, \delta > 0$ such that

$$\operatorname{dist}(x, (\partial_C g)^{-1}(x^*)) \leq \kappa \operatorname{dist}(\partial_C g(x), x^*) \quad (x^* \in \mathbb{B}(0, \delta);\ x \in \mathbb{B}(\bar{x}, \delta)). \tag{28.3}$$

We illustrate this with two examples. The first concerns data stability of least squares fitting, which in Hilbert spaces can be formulated as tilt stability.

Example 28.1 (data stability of least squares fitting). Let X, Y be Hilbert spaces and $g(x) = \frac{1}{2}\|Ax - y\|_Y^2$ for some $A \in \mathbb{L}(X; Y)$ and $y \in Y$. Taking $p = A^*\Delta y$ for some $\Delta y \in Y$, we can write this in the form of (28.2) via

$$f(x; p) = g(x) - \langle A^*\Delta y, x \rangle_X = \frac{1}{2}\|Ax - (y + \Delta y)\|_Y^2 - \frac{1}{2}\|\Delta y\|_Y^2.$$

Data stability thus follows from the metric regularity of ∂g at a minimizer $\bar{x}$ of the convex functional g. We have $\partial_C g(x) = \{A^*(Ax - y)\}$, so

$$(\partial g)^{-1}(x^*) = \{\tilde{x} \in X \mid A^*A\tilde{x} = A^*y + x^*\}.$$

Therefore (28.3) is equivalent to

$$\inf_{\tilde{x}\in X}\{\|\tilde{x}-x\|_X \mid A^*A\tilde{x} = A^*y + x^*\} \le \kappa\|A^*Ax - (A^*y+x^*)\|_X \qquad (x^* \in \mathbb{B}(0,\delta);\ x \in \mathbb{B}(\bar{x},\delta)).$$

If A^*A has a bounded inverse $(A^*A)^{-1} \in \mathbb{L}(X;X)$, then we can take $\kappa = \|(A^*A)^{-1}\|_{\mathbb{L}(X;X)}$ for any $\delta > 0$. On the other hand, if A^*A is not surjective, then there cannot be metric regularity (simply take an appropriate choice of $x^* \notin \operatorname{ran} A^*A$).

For a genuinely nonsmooth example, we consider the (academic) problem of minimizing the (non-squared) norm on a Hilbert space.

Example 28.2 (tilt stability of least norm fitting). Let X be a Hilbert space and $g(x) = \|x-z\|_X$ for some $z \in X$. To show tilt stability, we have to verify (28.3) for some $\kappa, \delta > 0$. For $x \neq z$, we have $\partial g(x) = \{(x-z)/\|x-z\|_X\}$, and for $x = z$, we have $\partial g(x) = \mathbb{B}(0,1)$. Thus (28.3) reads

$$\operatorname{dist}(x, (\partial g)^{-1}(x^*)) \le \kappa \begin{cases} \left\| \frac{x-z}{\|x-z\|_X} - x^* \right\|_X & \text{if } x \neq z, \\ \operatorname{dist}(x^*, \mathbb{B}(0,1)) & \text{if } x = z, \end{cases} \tag{28.4}$$

for all $x^* \in \mathbb{B}(0,\delta)$ and $x \in \mathbb{B}(\bar{x},\delta)$ where

$$\operatorname{dist}(x, (\partial g)^{-1}(x^*)) = \begin{cases} \operatorname{dist}(x-z, x^*[0,\infty)) & \text{if } \|x^*\|_X = 1, \\ \|x-z\|_X & \text{if } \|x^*\|_X < 1, \\ \infty & \text{if } \|x^*\|_X > 1. \end{cases}$$

As the inequality cannot hold if $\|x^*\|_X > 1$, we take $\delta \in (0,1]$ to ensure that this does not happen. If $x = z$, then (28.4) trivially holds for any $\kappa > 0$, both sides being zero. For $x^* \in \mathbb{B}(0,\delta)$ and $x \in \mathbb{B}(\bar{x},\delta) \setminus \{z\}$, the inequality (28.4) reads

$$\kappa \left\| \frac{x-z}{\|x-z\|_X} - x^* \right\|_X \ge \begin{cases} \operatorname{dist}(x-z, x^*[0,\infty)) & \text{if } \|x^*\|_X = 1, \\ \|x-z\|_X & \text{if } \|x^*\|_X < 1. \end{cases}$$

Choosing $x^* = \lambda(x-z)/\|x-z\|_X$, and letting $\lambda \nearrow 1$, we see that the inequality cannot hold unless $\delta \in (0,1)$ (which prevents $\lambda \nearrow 1$). Thus, taking the infimum of the left-hand side over $\|x^*\|_X \le \delta < 1$ and the supremum of the right-hand side over $x \in \mathbb{B}(\bar{x},\delta)$, the inequality holds if $\kappa(1-\delta) \ge \delta$. This can be satisfied for any $\kappa > 0$ for sufficiently small $\delta \in (0,1)$.

Since $x^* \in X$ is comparable to the tilt parameter $p \in X$, this says that we can only stably "tilt" g by an amount $\|p\|_X < 1$. If we tilt with $\|p\|_X > 1$, the tilted function has no minimizer, while for $\|p\|_X = 1$, every $x = z + tp$ for $t \ge 0$ is a minimizer.

We now return to the general solution mapping (28.1). The following proposition applied to $F(x,p) := \partial f(x;p)$ provides a general tool for our analysis.

Theorem 28.3. *Let P, X, and Y be reflexive and Gâteaux smooth Banach spaces. For $F : X \times P \to Y$, let*

$$S(p) := \{x \in X \mid 0 \in F(x,p)\}.$$

Then S has the Aubin property at $\bar{p}$ for $\bar{x} \in S(\bar{p})$ if

$$(0,p^*) \in D^*_N F(\bar{x},\bar{p}|0)(y^*) \Rightarrow y^* = 0,\ p^* = 0 \tag{28.5}$$

and

$$Q(y,p) := \{x \in X \mid y \in F(x,p)\}.$$

is PSNC at $(\bar{y},\bar{p})$ for $\bar{x}$.

Proof. We have $S(p) = Q(0,p)$. Hence if we can show that Q has the Aubin property at $(0,\bar{p})$ for $\bar{x}$, this will imply the Aubin property of S at $\bar{p}$ for $\bar{x}$ by simple restriction of the free variables in Theorem 27.2 (i) to the subspace $\{0\} \times P$.

We do this by applying Theorem 27.10 to Q, which holds if we can show that

$$D^*_M Q(0,\bar{p}|\bar{x})(0) = \{0\}.$$

By (27.27), a sufficient assumption for this is that

$$D^* Q(0,\bar{p}|\bar{x})(0) = \{0\},$$

which can equivalently be expressed as

$$(y^*,p^*,0) \in N_{\operatorname{graph} Q}(0,\bar{p},\bar{x}) \Rightarrow y^* = 0,\ p^* = 0. \tag{28.6}$$

Now

$$\operatorname{graph} Q = \{(y,p,x) \mid y \in F(x,p)\} = \pi \operatorname{graph} F$$

for the permutation $\pi(x,p,y) := (y,p,x)$ (which applied to a set should be understood as applied to every element of that set). We thus also have

$$N_{\operatorname{graph} Q}(y,p,x) = \pi N_{\operatorname{graph} F}(\pi(y,p,x)).$$

In particular, (28.6) becomes

$$(0,p^*,y^*) \in N_{\operatorname{graph} F}(\bar{x},\bar{p},0) \Rightarrow y^* = 0,\ p^* = 0.$$

But this is equivalent to (28.5). □

Remark 28.4. Theorem 28.3 is related to the classical implicit function theorem. If F is graphically regular at $(\bar{x}, \bar{p}, 0)$, it is also possible derive explicit characterizations of DS such as

$$DS(\bar{p}|\bar{x})(\Delta p) = \{\Delta x \in X \mid DF(\bar{x}, \bar{p}|0)(\Delta x, \Delta p) \ni 0\}.$$

For details in finite dimensions, we refer to [Rockafellar and Wets, 1998, Theorem 9.56 & Proposition 8.41].

We close this section by illustrating the requirements of Theorem 28.3 for the stability of specific problems of the form (P) with respect to the penalty parameter α. (Naturally, these can be relaxed or made further explicit in more concrete situations.) We consider for $\alpha > 0$ and $h, g : X \to \overline{\mathbb{R}}$ the problem

$$\min_{x \in X} h(x) + \alpha g(x).$$

For this problem, we define the Clarke-critical point mapping

$$S(\alpha) := \{x \in X \mid 0 \in \partial_C(h + \alpha g)(x)\}. \tag{28.7}$$

When the problem is convex, this coincides with the solution mapping. Subject to a non-degeneracy condition, the next theorem yields a stability estimate for convex g and smooth h.

Theorem 28.5. *Let X be a finite-dimensional and Gâteaux smooth Banach space and let $h : X \to \mathbb{R}$ be twice continuously differentiable and $g : X \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous. Suppose*

$$0 \in h''(\bar{x})^* y + \bar{\alpha} D^*[\partial g](\bar{x}| - \bar{\alpha}^{-1} h'(\bar{x}))(y) \implies y = 0. \tag{28.8}$$

Then S has the Aubin property at $\bar{\alpha}$ for any $\bar{x} \in S(\bar{\alpha})$.

Proof. By Theorems 13.4, 13.5, and 13.20, we can expand

$$S(\alpha) = \{x \in X \mid 0 \in F(x; \alpha)\} \quad \text{for} \quad F(x; \alpha) := h'(x) + \alpha \partial g(x).$$

To apply Theorem 28.3 to prove the Aubin property, we need to verify its assumptions. First, by Theorems 25.14 and 25.20, we have

$$D^* F(\bar{x}; \bar{\alpha}|0)(y) = \begin{pmatrix} h''(\bar{x})^* y + \bar{\alpha} D^*[\partial g](\bar{x}| - \bar{\alpha}^{-1} h'(\bar{x}))(y) \\ -\langle h'(\bar{x}), y \rangle_X \end{pmatrix}.$$

Thus (28.5) holds by (28.8). Furthermore, since $X^* \times \mathbb{R}$ is finite-dimensional, the PSNC holds at every (y, α) with $y \in F(\bar{x}, \bar{\alpha})$ and $\alpha > 0$ by Lemma 25.5. Hence Theorem 28.3 is indeed applicable and implies that S has the Aubin property at $\bar{\alpha}$. □

Corollary 28.6. *Under the assumptions of Theorem 28.5,*

$$\inf_{x\in S(\alpha)} \|\bar{x} - x\|_X \le \kappa|\bar{\alpha} - \alpha|$$

for some $\kappa > 0$ and all α sufficiently close to $\bar{\alpha}$.

Proof. The claim follows directly from the definition (27.1) of the Aubin property for S given by (28.7) in $y = \bar{x} \in S(\bar{\alpha})$, which yields

$$\inf_{x\in S(\alpha)} \|\bar{x} - x\|_X = \operatorname{dist}(\bar{x}, S(\alpha)) \le \kappa \operatorname{dist}(S^{-1}(\bar{x}), \alpha) = \kappa|\bar{\alpha} - \alpha|. \qquad \square$$

28.2 METRIC SUBREGULARITY OF CONVEX SUBDIFFERENTIALS

We recall from (27.3) that a set-valued mapping $H : X \rightrightarrows X^*$ is metrically subregular at $\widehat{x} \in X$ for $\widehat{w} \in X^*$ if there exist $\delta > 0$ and $\kappa > 0$ such that

$$\operatorname{dist}(x, H^{-1}(\widehat{w})) \le \kappa \operatorname{dist}(\widehat{w}, H(x)) \quad (x \in \mathbb{B}(\widehat{x}, \delta)).$$

We also recall that the infimum of all $\kappa > 0$ for which this inequality holds for some $\delta > 0$ is denoted by $\operatorname{subreg} H(\widehat{x}|\widehat{w})$, the modulus of (metric) subregularity of H at $\widehat{x}$ for $\widehat{w}$. In the following, we will also make use of the *squared* distance of $x \in X$ to a set $A \subset X$,

$$\operatorname{dist}^2(x, A) := \inf_{\tilde{x}\in A} \|x - \tilde{x}\|_X^2.$$

We then have the following characterization of metric subregularity of convex functionals.

Theorem 28.7. *Let $g : X \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous and let $\widehat{x} \in X$ with $0 \in \partial g(\widehat{x})$. If there exist $\gamma > 0$ and $\delta > 0$ such that*

$$g(x) \ge g(\widehat{x}) + \gamma \operatorname{dist}^2(x, [\partial g]^{-1}(0)) \quad (x \in \mathbb{B}_X(\widehat{x}, \delta)), \tag{28.9}$$

then ∂g is metrically subregular at $\widehat{x}$ for 0 with $\kappa = \gamma^{-1}$ and the same δ.

Conversely, if ∂g is metrically subregular at $\widehat{x}$ for 0 with some $\kappa, \delta > 0$, then (28.9) holds for any $\gamma \in (0, 1/(4\kappa))$.

Proof. Let first (28.9) hold for $\gamma, \delta > 0$. We need to show that

$$\gamma \operatorname{dist}(x, [\partial g]^{-1}(0)) \le \operatorname{dist}(0, \partial g(x)) \quad (x \in \mathbb{B}(\widehat{x}, \delta)). \tag{28.10}$$

To that end, let $x \in \mathbb{B}(\widehat{x}, \delta)$. Clearly, if $\partial g(x) = \emptyset$, there is nothing to prove. So assume that there exists an $x^* \in \partial g(x)$. Then $x \in \operatorname{dom} g$, so that (28.9) shows that $\operatorname{dist}^2(x, [\partial g]^{-1}(0)) <$

∞. Consequently $[\partial g]^{-1}(0) \neq \emptyset$. For each $\varepsilon > 0$, by the definition of the set-distance, we can therefore find $x_\varepsilon \in [\partial g]^{-1}(0)$ such that

$$\|x - x_\varepsilon\|_X \leq \operatorname{dist}(x, [\partial g]^{-1}(0)) + \varepsilon. \tag{28.11}$$

By the definition of the convex subdifferential and $\widehat{x}, x_\varepsilon \in \arg\min g$, we have

$$\langle x^*, x - x_\varepsilon \rangle_X \geq g(x) - g(x_\varepsilon) = g(x) - g(\widehat{x}).$$

Combined with (28.9) and (28.11), this yields

$$\begin{aligned} \gamma \operatorname{dist}^2(x, [\partial g]^{-1}(0)) &\leq \langle x^*, x - x_\varepsilon \rangle_X \\ &\leq \|x^*\|_{X^*} \|x - x_\varepsilon\|_X \leq \|x^*\|_{X^*} (\operatorname{dist}(x, [\partial g]^{-1}(0)) + \varepsilon). \end{aligned}$$

Since $\varepsilon > 0$ was arbitrary and $\|x^*\|_{X^*} \leq \operatorname{dist}(0, \partial g(x))$, we obtain (28.10).

Conversely, let ∂g be metrically subregular at $\widehat{x}$ for 0 for some parameters $\kappa, \delta > 0$. Take any $\gamma \in (0, 1/(4\kappa))$. We argue by contradiction. Assume that (28.9) does not hold. Then we can find some $\tilde{x} \in \mathbb{B}(\widehat{x}, 2\delta/3)$ such that

$$g(\tilde{x}) < g(\widehat{x}) + \gamma \operatorname{dist}^2(\tilde{x}, [\partial g]^{-1}(0)). \tag{28.12}$$

However, $\widehat{x}$ is a minimizer of g, so necessarily $\gamma \operatorname{dist}^2(\tilde{x}, [\partial g]^{-1}(0)) > 0$. By Ekeland's variational principle (Theorem 2.16), we can thus find $y \in X$ satisfying

$$\|y - \tilde{x}\|_X \leq \frac{1}{2} \operatorname{dist}(\tilde{x}, [\partial g]^{-1}(0)) \tag{28.13}$$

and for all $x \in X$ that

$$g(x) \geq g(y) - \frac{\gamma \operatorname{dist}^2(\tilde{x}, [\partial g]^{-1}(0))}{\frac{1}{2} \operatorname{dist}(\tilde{x}, [\partial g]^{-1}(0))} \|x - y\|_X = g(y) - 2\gamma \operatorname{dist}(\tilde{x}, [\partial g]^{-1}(0)) \|x - y\|_X.$$

In follows that y minimizes $g + 2\gamma \operatorname{dist}(\tilde{x}, [\partial g]^{-1}(0)) \|\cdot - y\|_X$, which by Theorems 4.2, 4.6, and 4.14 is equivalent to $0 \in \partial g(y) + 2\gamma \operatorname{dist}(\tilde{x}, [\partial g]^{-1}(0)) \mathbb{B}_{X^*}$. Hence we can find some $y^* \in \partial g(y)$ satisfying $\|y^*\|_{X^*} \leq 2\gamma \operatorname{dist}(\tilde{x}, [\partial g]^{-1}(0))$. Using (28.13), we now obtain

$$\begin{aligned} 2\kappa \operatorname{dist}(0, \partial g(y)) &< (2\gamma)^{-1} \operatorname{dist}(0, \partial g(y)) \\ &\leq (2\gamma)^{-1} \|y^*\|_{X^*} \leq \operatorname{dist}(\tilde{x}, [\partial g]^{-1}(0)) \\ &= 2 \operatorname{dist}(\tilde{x}, [\partial g]^{-1}(0)) - \operatorname{dist}(\tilde{x}, [\partial g]^{-1}(0)) \\ &\leq 2\|y - \tilde{x}\|_X + 2 \operatorname{dist}(y, [\partial g]^{-1}(0)) - \operatorname{dist}(\tilde{x}, [\partial g]^{-1}(0)) \\ &\leq 2 \operatorname{dist}(y, [\partial g]^{-1}(0)). \end{aligned}$$

By (28.13) and our choice of $\tilde{x} \in \mathbb{B}(\widehat{x}, 2\delta/3)$,

$$\|y - \widehat{x}\|_X \leq \|y - \tilde{x}\|_X + \|\tilde{x} - \widehat{x}\|_X \leq \frac{3}{2} \|\tilde{x} - \widehat{x}\|_X \leq \delta.$$

Therefore $y \in \mathbb{B}(\widehat{x}, \delta)$ violates the assumed metric subregularity (28.10) with the factor $\tilde{\gamma}$, and hence (28.9) holds. □

Applying Theorem 28.7 to $x \mapsto g(x) + \langle \widehat{x}^*, x \rangle_X$ now yields the following characterization due to [Aragón Artacho and Geoffroy, 2014].

Corollary 28.8. *Let $g : X \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous and let $\widehat{x} \in X$ and $\widehat{x}^* \in \partial g(\widehat{x})$. If there exist $\gamma > 0$ and $\delta > 0$ such that*

$$g(x) \geq g(\widehat{x}) + \langle \widehat{x}^*, x - \widehat{x} \rangle_X + \gamma \operatorname{dist}^2(x, [\partial g]^{-1}(\widehat{x}^*)) \quad (x \in \mathbb{B}_X(\widehat{x}, \delta)), \tag{28.14}$$

then ∂g is metrically subregular at $\widehat{x}$ for $\widehat{x}^*$ with $\kappa = \gamma^{-1}$ and the same δ.

Conversely, if ∂g is metrically subregular at $\widehat{x}$ for $\widehat{x}^*$ with some $\kappa, \delta > 0$, then (28.9) holds for any $\gamma \in (0, 1/(4\kappa))$.

If we denote by $\hat{\gamma}(\widehat{x}|\widehat{x}^*)$ the supremum of $\gamma > 0$ for which (28.14) holds for some $\delta > 0$, then we obtain the following estimate involving the modulus of subregularity.

Corollary 28.9. *Let $g : X \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous and let $\widehat{x} \in X$ and $\widehat{x}^* \in \partial g(\widehat{x})$. Then*

$$\operatorname{subreg} \partial g(\widehat{x}|\widehat{x}^*) \leq \hat{\gamma}(\widehat{x}|\widehat{x}^*)^{-1} \leq 4 \operatorname{subreg} \partial g(\widehat{x}|\widehat{x}^*).$$

Remark 28.10 (strong metric subregularity). As in Remark 27.1, we can also characterize *strong* metric subregularity using a strong notion of local subdifferentiability. In the setting of Corollary 28.8, it was shown in [Aragón Artacho and Geoffroy, 2014] that strong metric subregularity of ∂g at $\widehat{x}$ for $\widehat{x}^*$ is equivalent to

$$g(x) \geq g(\widehat{x}) + \langle \widehat{x}^*, x - \widehat{x} \rangle_X + \gamma \|x - \widehat{x}\|_X^2 \quad (x \in \mathbb{B}_X(\widehat{x}, \delta)), \tag{28.15}$$

i.e., a local form of strong subdifferentiability. Compared to the characterization of metric subregularity in (28.14), intuitively the strong version does not "squeeze" $[\partial g]^{-1}(\widehat{x}^*)$ into a single point.

Strong metric subregularity may almost trivially be used in the convergence proofs of Part II and Chapter 15 as a relaxation of strong convexity; compare [Clason et al., 2020]. Also observe that (28.15) can be expressed in terms of the *Bregman divergence* (see Section 11.1) as

$$B_g^{\widehat{x}^*}(x, \widehat{x}) \geq \gamma \|x - \widehat{x}\|_X^2 \quad (x \in \mathbb{B}_X(\widehat{x}, \delta)),$$

i.e., that $B_g^{\widehat{x}^*}$ is *elliptic* at $\widehat{x}$ in the sense of [Valkonen, 2021a]. In optimization methods based on preconditioning by Bregman divergences instead of the linear preconditioner M as discussed in Section 11.1, this generalizes the positive definiteness requirement on M.

28.3 TIKHONOV-TYPE REGULARIZATION OF INVERSE PROBLEMS

Let now the data y^δ depend on a *noise level* $\delta > 0$, and consider for a corresponding parameter $\alpha_\delta > 0$ the problem

$$\min_{x\in X} \frac{1}{2}\|Ax - y^\delta\|_Y^2 + \alpha_\delta g(x), \tag{28.16}$$

where $A \in \mathbb{L}(X;Y)$ between a Banach space X and a Hilbert space Y. This problem is called a *Tikhonov-type regularization* of the inverse problem $Ax = y^\delta$. If $g(x) = \frac{1}{2}\|\cdot\|_X^2$ with X a Hilbert space, we talk simply of *Tikhonov regularization*.

We assume for some true data $\hat{y}$ that

$$\|y - \hat{y}\|_Y \le \delta. \tag{28.17}$$

Suppose there exists a solution $\hat{x}$ to the problem

$$\min_{x\in C} g(x) \quad \text{where} \quad C := \{x \in X \mid Ax = \hat{y}\}. \tag{28.18}$$

Denote by $\hat{X}$ the set of solutions to (28.18). In inverse problems, the question whether solutions x_δ to the Tikhonov-type problem (28.16) converge to some $\hat{x} \in \hat{X}$ is a topic of *regularization theory*. The condition (28.19) of the next lemma is known as a *source condition* in that context.

Lemma 28.11. *Suppose $A \in \mathbb{L}(X;Y)$ and that $g : X \to \overline{\mathbb{R}}$ is convex, proper, and lower semicontinuous with* $\operatorname{int}\operatorname{dom} g \cap C \neq \emptyset$. *We have $\hat{x} \in \hat{X}$ if and only if there exists $\hat{w} \in Y$ such that*

$$A\hat{x} = \hat{y} \quad \text{and} \quad -A^*\hat{w} \in \partial g(\hat{x}), \tag{28.19}$$

Proof. The condition $\operatorname{int}\operatorname{dom} g \cap C \neq \emptyset$ guarantees that the sum rule Theorem 4.14 holds as an equality for $\delta_C + g$. Writing $\delta_C(x) = \delta_{\{\hat{y}\}}(Ax)$, and using the chain rule (4.17) and the fact that

$$\delta_{\{\hat{y}\}}(y) = \begin{cases} Y & y = \hat{y}, \\ \emptyset, & \text{otherwise} \end{cases}$$

we therefore obtain

$$\partial[\delta_C + g](x) = A^*Y + \partial g(x) \quad \text{whenever} \quad Ax = \hat{y}.$$

Thus $0 \in \partial[\delta_C + g](x)$ whenever (28.19) holds. Now the Fermat principle of Theorem 4.2 establishes the claim. □

The next result characterizes convergence. For brevity we write

$$f_\delta(x) := \frac{1}{2}\|Ax - y^\delta\|_Y^2.$$

The condition (28.20) in the next theorem is satisfied in particular if x_δ is an e_δ-minimizer of $f_\delta + \alpha_\delta g$. Observe that the right-hand side of (28.20) does not depend on the choice of $\hat{x} \in \hat{X}$. We directly assume the characterization (28.14) of metric subregularity to be able to use an optimal modulus γ for which the characterization holds.

Theorem 28.12. *Let $A \in \mathbb{L}(X;Y)$ and $g : X \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous with $\operatorname{int}\operatorname{dom} g \cap C \neq \emptyset$. Suppose (28.17) holds, $\hat{x} \in X$ satisfies (28.19), and that for every $\delta > 0$, for some $e_\delta > 0$ there exists $x_\delta \in X$ satisfying*

$$[f_\delta + \alpha_\delta g](x_\delta) \le [f_\delta + \alpha_\delta g](\hat{x}) + e_\delta. \tag{28.20}$$

Suppose for some $\tilde{\delta} > 0$ and $\tilde{X} \subset \hat{X}$ that for all $\delta \in (0, \tilde{\delta})$ and $\tilde{x} \in \tilde{X}$, the subdifferential mapping $\partial[f_\delta + \alpha_\delta g]$ satisfies (28.9) at $\tilde{x}$ for $f_\delta'(\tilde{x}) - \alpha_\delta A^* \hat{w}$ in the neighborhood $U_{\tilde{x}}$ (independent of δ) with the factor $\gamma > 0$ (independent of both δ and $\tilde{x}$) with respect to the norm

$$\|x\|_\delta := \sqrt{\|Ax\|_Y^2 + \alpha_\delta \|x\|_X^2} \quad (x \in X).$$

Assume further for some $\rho > 0$ that

$$\bigcup_{\tilde{x} \in \tilde{X}} U_{\tilde{x}} \supset U_\rho := \{x \in X \mid \|A(x - \hat{x})\| \le \rho,\ R(x) \le R(\hat{x}) + \rho\}. \tag{28.21}$$

Then there exists $\bar{\delta} > 0$ such that

$$\operatorname{dist}^2(x_\delta, \hat{X}) \le \frac{e_\delta}{\gamma \alpha_\delta} + \frac{\delta^2}{2\gamma^2 \alpha_\delta} + \frac{\alpha_\delta}{2\gamma^2}\|\hat{w}\|_Y^2 \quad (\delta \in (0, \bar{\delta})).$$

Proof. Since $A\hat{x} = \hat{y}$, using Young's inequality, (28.20), and (28.17), we have

$$\begin{aligned}
\frac{1}{2}\|A(x_\delta - \hat{x})\|_Y^2 + 2\alpha_\delta g(x_\delta) &\le \|Ax_\delta - y^\delta\|_Y^2 + 2\alpha_\delta g(x_\delta) + \|y^\delta - \hat{y}\|_Y^2 \\
&\le 2e_\delta + 2\|y^\delta - \hat{y}\|_Y^2 + 2\alpha_\delta g(\hat{x}) \\
&\le 2(e_\delta + \delta^2 + \alpha_\delta g(\hat{x})).
\end{aligned}$$

Thus both

$$\|A(x_\delta - \hat{x})\|_Y^2 \le 4(e_\delta + \delta + \alpha_\delta g(\hat{x})) \quad \text{and} \quad g(x_\delta) \le g(\hat{x}) + \frac{e_\delta + \delta^2}{\alpha_\delta}.$$

This implies the existence of $\bar{\delta} \in (0, \tilde{\delta}]$ such that $x_\delta \in U_\rho$ for $\delta \in (0, \bar{\delta})$. Consequently (28.21) establishes for every such δ an element $\hat{x}_\delta \in \hat{X}$ such that $x_\delta \in U_{\hat{x}_\delta}$. By $f_\delta + \alpha_\delta g$ satisfying (28.14) at $\hat{x}_\delta$ for $f_\delta'(\tilde{x}) - \alpha_\delta A^* \hat{w}$ for such δ, therefore

$$[f_\delta + \alpha_\delta g](x_\delta) - [f_\delta + \alpha_\delta g](\hat{x}_\delta) \ge \langle f_\delta'(\hat{x}_\delta) - \alpha_\delta A^* \hat{w}, x_\delta - \hat{x}_\delta \rangle_X + \gamma \operatorname{dist}_\delta^2(x_\delta, \hat{X}), \tag{28.22}$$

where $\operatorname{dist}_\delta$ denotes the distance-to-set function with respect to $\|\cdot\|_\delta$.

We next expand

$$f_\delta'(\hat{x}_\delta) - \alpha_\delta A^* \hat{w} = A^*(A\hat{x}_\delta - y^\delta - \alpha_\delta \hat{w}) = A^*(\hat{y} - y^\delta - \alpha_\delta \hat{w}).$$

Hence (28.19) and (28.22) establish

$$\begin{aligned} e_\delta &\geq \langle f_\delta'(\hat{x}_\delta) - \alpha_\delta A^* \hat{w}, x_\delta - \hat{x}\rangle_X + \gamma \operatorname{dist}_\delta^2(x_\delta, \hat{X}) \\ &= \langle \hat{y} - y^\delta - \alpha_\delta \hat{w}, A(x_\delta - \hat{x}_\delta)\rangle_Y + \gamma \inf_{\bar{x}\in\hat{X}} \left(\|A(x_\delta - \bar{x})\|_Y^2 + \alpha_\delta \|x_\delta - \bar{x}\|_X^2 \right). \end{aligned}$$

Since $A\bar{x} = A\hat{x}_\delta$ due to $\hat{X} \subset C$, distributing the inf over the entire right-hand side and using Young's inequality establishes

$$\begin{aligned} e_\delta &\geq \inf_{\bar{x}\in\hat{X}} \left(\langle \hat{y} - y^\delta - \alpha_\delta \hat{w}, A(x_\delta - \bar{x})\rangle_Y + \gamma \|A(x_\delta - \bar{x})\|_Y^2 + \gamma\alpha_\delta \|x_\delta - \bar{x}\|_X^2 \right) \\ &\geq \inf_{\bar{x}\in\hat{X}} \left(-\frac{1}{4\gamma} \|\hat{y} - y^\delta - \alpha_\delta \hat{w}\|_Y^2 + \gamma\alpha_\delta \|x_\delta - \bar{x}\|_X^2 \right). \end{aligned}$$

Thus, again using Young's inequality and (28.17), we obtain

$$\operatorname{dist}^2(x_\delta, \hat{X}) \leq \frac{e_\delta}{\gamma\alpha_\delta} + \frac{1}{4\gamma^2\alpha_\delta} \|\hat{y} - y^\delta - \alpha_\delta \hat{w}\|_Y^2 \leq \frac{e_\delta}{\gamma\alpha_\delta} + \frac{\delta^2}{2\gamma^2\alpha_\delta} + \frac{\alpha_\delta}{2\gamma^2} \|\hat{w}\|_Y^2.$$

This is the claim. □

Immediately we obtain the following characterization of convergence of regularized solutions.

Corollary 28.13. *Under the assumptions of Theorem 28.12, if*

$$\lim_{\delta \searrow 0} \left(\alpha_\delta, \frac{\delta^2}{\alpha_\delta}, \frac{e_\delta}{\alpha_\delta} \right) = 0,$$

then

$$\lim_{\delta \searrow 0} \operatorname{dist}(x_\delta, \hat{X}) = 0.$$

Remark 28.14. For an introduction to inverse problems, we refer to [Clason, 2020b; Hanke, 2017; Mueller and Siltanen, 2012]; a classical treatise on regularization theory is [Engl et al., 1996] with Banach spaces and other advanced aspects covered in [Ito and Jin, 2014; Kaltenbacher et al., 2008; Schuster et al., 2012]; see also Chapters 30 to 32 and the remarks therein. Our specialized account is based on [Valkonen, 2021b], which also shows that using *strong* metric subregularity in Theorem 28.12 in place of metric subregularity yields convergence to a specific $\hat{x} \in \hat{X}$ instead of the set $\hat{X}$. Those results also relax the requirement $\bigcup_{\tilde{x}\in\tilde{X}} U_{\tilde{x}} \supset U_\rho$ through assumptions of weak(-∗) closedness and openness.

29 SPLITTING METHODS: FASTER CONVERGENCE FROM REGULARITY

As we have seen in Chapter 10, proximal point and splitting methods can be accelerated if at least one of the involved functionals is strongly convex. However, this can be a too strong requirement, and we will show in this chapter how faster convergence (even without acceleration) can be shown under the weaker requirements of metric subregularity or strong submonotonicity. We begin in Section 29.1 by introducing the latter notion before illustrating in Section 29.2 the effect of the two properties on splitting methods by showing local linear convergence of forward-backward splitting.

29.1 SUBMONOTONICITY OF CONVEX SUBDIFFERENTIALS

Throughout this section, let X be a Banach space and $G : X \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous. Our goal is now to give conditions for metric subregularity and strong submonotonicity of $\partial G : X \rightrightarrows X^*$ at a critical point $\widehat{x} \in X$ with $0 \in \partial G(\widehat{x})$.

Recall the characterization of metric subregularity of a convex subdifferential shown in Section 28.2. As a weaker alternative to that result, we now relax the strong monotonicity assumption of Chapter 10 more directly. We say that a set-valued mapping $H : X \rightrightarrows X^*$ is (γ, θ)-*strongly submonotone at* $\widehat{x}$ *for* $\widehat{x}^* \in H(\widehat{x})$ with $\theta \geq \gamma > 0$ if there exists $\delta > 0$ such that for all $x \in \mathbb{B}_X(\widehat{x}, \delta)$ and $x^* \in H(x) \cap \mathbb{B}_{X^*}(\widehat{x}^*, \delta)$,

$$\inf_{\bar{x} \in H^{-1}(\widehat{x}^*)} \left(\langle x^* - \widehat{x}^*, x - \bar{x} \rangle_X + (\theta - \gamma) \|x - \bar{x}\|_X^2 \right) \geq \theta \operatorname{dist}^2(x, H^{-1}(\widehat{x}^*)). \tag{29.1}$$

If this only holds for $\theta \geq \gamma = 0$, then we call H *submonotone* at $\widehat{x}$ for $\widehat{x}^*$.

Clearly, (strong) monotonicity (see Lemma 7.4) implies (strong) submonotonicity at any $\widehat{x} \in X$ and $\widehat{x}^* \in H(\widehat{x})$. However, subdifferentials of convex functionals need not be strongly monotone. The next theorem shows that local second-order growth away from the set of minimizers implies strong submonotonicity of such subdifferentials at any minimizer $\widehat{x}$ for $\widehat{x}^* = 0$, which is the monotonicity-based analogue of the characterization of metric subregularity in Theorem 28.7.

Theorem 29.1. *Let $G : X \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous and let $\widehat{x} \in X$ with $0 \in \partial G(\widehat{x})$. If there exists $\delta > 0$ such that*

$$G(x) \geq G(\widehat{x}) + \gamma \operatorname{dist}^2(x, [\partial G]^{-1}(0)) \quad (x \in \mathbb{B}_X(\widehat{x}, \delta)), \tag{29.2}$$

then ∂G is (γ, θ)-strongly submonotone at $\widehat{x}$ for 0 for any $\theta \geq \gamma$.

Proof. Since $\theta \geq \gamma$, (29.2) is equivalent to

$$\inf_{\bar{x} \in [\partial G]^{-1}(0)} \left(G(x) - G(\widehat{x}) + (\theta - \gamma)\|x - \bar{x}\|_X^2\right) \geq \theta \operatorname{dist}^2(x, [\partial G]^{-1}(0)) \tag{29.3}$$

for all $x \in \mathbb{B}_X(\widehat{x}, \delta)$. By the definition of the convex subdifferential, we have for all $\bar{x} \in [\partial G]^{-1}(0)$ and $\widehat{x}^* = 0$ that

$$\langle x^* - \widehat{x}^*, x - \bar{x}\rangle_X \geq G(x) - G(\bar{x}) = G(x) - G(\widehat{x}).$$

Inserting this into (29.3) yields the definition (29.1) of strong submonotonicity for $H = \partial G$. □

Together with Theorem 28.7, this shows that for convex subdifferentials, metric subregularity implies strong submonotonicity, which is thus a weaker property.

We conclude this section by showing that the subdifferentials of the indicator functional of the finite-dimensional unit ball and of the absolute value function are both subregular and strongly submonotone. Note that neither of these subdifferentials is strongly monotone in the conventional sense. Here we restrict ourselves to showing (γ, γ)-strong submonotonicity for some $(\widehat{x}, \widehat{x}^*) \in \operatorname{graph} \partial G$, i.e., that there exists $\delta > 0$ such that

$$\langle x^* - \widehat{x}^*, x - \widehat{x}\rangle_X \geq \gamma \operatorname{dist}^2(x, [\partial G]^{-1}(\widehat{x}^*)) \quad (x \in \mathbb{B}(\widehat{x}, \delta),\ x^* \in \partial G(x)). \tag{29.4}$$

Lemma 29.2. *Let $G := \delta_{\mathbb{B}(0,\alpha)}$ on $(\mathbb{R}^N, \|\cdot\|_2)$ and $(\widehat{x}, \widehat{x}^*) \in \operatorname{graph} \partial G$. Then ∂G is*

(i) metrically subregular at $\widehat{x}$ for $\widehat{x}^*$ for any $\delta \in (0, \alpha]$ and

$$\kappa \geq \begin{cases} 2\alpha/\|\widehat{x}^*\|_2 & \text{if } \widehat{x}^* \neq 0, \\ 0 & \text{if } \widehat{x}^* = 0; \end{cases}$$

(ii) (γ, γ)-strongly submonotone at $\widehat{x}$ for $\widehat{x}^*$ for any $\delta > 0$ and

$$\gamma \leq \begin{cases} \|\widehat{x}^*\|_2/(2\alpha) & \text{if } \widehat{x}^* \neq 0, \\ \infty & \text{if } \widehat{x}^* = 0. \end{cases}$$

Proof. We first verify (28.14) for $\delta = \alpha$ and $\gamma = \kappa^{-1}$ as stated. To that end, let $x \in \mathbb{B}(0,\alpha)$. If $\widehat{x}^* = 0$, then (28.14) trivially holds by the subdifferentiability of G and $\operatorname{dist}^2(x, [\partial G]^{-1}(\widehat{x}^*)) = \operatorname{dist}^2(x, \mathbb{B}(0,\alpha)) = 0$. Let therefore $\widehat{x}^* \neq 0$. Then $[\partial G]^{-1}(\widehat{x}^*) = \{\widehat{x}\}$ as well as $\|\widehat{x}\|_2 = \alpha$ and $\widehat{x}^* = \beta\widehat{x}$ for $\beta = \|\widehat{x}^*\|_2/\|\widehat{x}\|_2$. Since $\gamma \leq \|\widehat{x}^*\|_2/(2\alpha)$, we have $\beta \geq 2\gamma$. Then $\|x\|_2 \leq \alpha$ yields

$$\begin{aligned}
\gamma \operatorname{dist}^2(x, [\partial G]^{-1}(\widehat{x}^*)) &= \gamma\|x - \widehat{x}\|_2^2 \\
&\leq \beta\langle \widehat{x}, \widehat{x} - x\rangle_2 - \frac{\beta}{2}\|\widehat{x}\|_2^2 + \frac{\beta}{2}\|x\|_2^2 \\
&\leq \beta\langle \widehat{x}, \widehat{x} - x\rangle_2 \\
&= \langle \widehat{x}^*, \widehat{x} - x\rangle_2 \\
&\leq \langle \widehat{x}^*, \widehat{x} - x\rangle_2 + G(x) - G(\widehat{x}).
\end{aligned}$$

Since $\operatorname{dom} G = \mathbb{B}(0,\alpha)$, this shows that (28.14) holds for any $\delta > 0$.

Corollary 28.8 now yields (i). Adding

$$G(\widehat{x}) - G(x) \geq \langle x^*, \widehat{x} - x\rangle_2 \quad (x^* \in \partial G(x))$$

to (28.14), we also obtain (29.4) and thus (ii). □

Lemma 29.3. *Let $G := |\cdot|$ on $\mathbb{R}$ and $(\widehat{x}, \widehat{x}^*) \in \operatorname{graph} \partial G$. Then ∂G is*

(i) metrically subregular at $\widehat{x}$ for $\widehat{x}^*$ for any $\kappa > 0$ and

$$\delta \leq \begin{cases} 2\kappa & \text{if } \widehat{x}^* \in \{1, -1\}, \\ \kappa & \text{if } |\widehat{x}^*| < 1; \end{cases}$$

(ii) (γ,γ)-strongly submonotone at $\widehat{x}$ for $\widehat{x}^*$ for any $\gamma > 0$ and

$$\delta \leq \begin{cases} 2\gamma^{-1} & \text{if } \widehat{x}^* \in \{1, -1\}, \\ \gamma^{-1} & \text{if } |\widehat{x}^*| < 1. \end{cases}$$

Proof. We first verify (28.14) for any $\delta > 0$ and $\gamma = \kappa^{-1}$ as stated. Suppose first that $\widehat{x}^* = 1$ so that $\widehat{x} \in [\partial G]^{-1}(\widehat{x}^*) = [0, \infty)$. This implies that $\widehat{x} = |\widehat{x}|$, and hence (28.14) becomes

$$|x| \geq x + \gamma \inf_{\bar{x} \geq 0} (x - \bar{x})^2 \quad (|x - \widehat{x}| \leq \delta).$$

If $x \geq 0$, this trivially holds by taking $\bar{x} = x$. If $x \leq 0$, the right-hand side is minimized by $\bar{x} = 0$, and thus the inequality holds for $x \geq -2\gamma^{-1}$. Since $\widehat{x} \geq 0$, this is guaranteed by our bound on δ. The case $\widehat{x}^* = -1$ is analogous.

If $|\widehat{x}^*| < 1$, then $\widehat{x} \in [\partial G]^{-1}(\widehat{x}^*) = \{0\}$, and hence (28.14) becomes

$$|x| \geq \gamma |x|^2 \quad (|x| \leq \delta).$$

This again holds by our choice of δ.

Corollary 28.8 now yields (i). Adding

$$G(\widehat{x}) - G(x) \geq \langle x^*, \widehat{x} - x \rangle \quad (x^* \in \partial G(x))$$

to (28.14), we also obtain (29.4) and thus (ii). □

Remark 29.4. If we allow in the definition of subregularity or submonotonicity an arbitrary neighborhood of $\widehat{x}$ instead of a ball, then Lemma 29.3 holds in a much larger neighborhood.

29.2 LOCAL LINEAR CONVERGENCE OF EXPLICIT SPLITTING

Returning to the notation used in Chapters 8 to 12, we now assume throughout that X is a Hilbert space, $F, G : X \to \overline{\mathbb{R}}$ are convex, proper, and lower semicontinuous, and that F is Fréchet differentiable and has a Lipschitz continuous gradient ∇F with Lipschitz constant $L \geq 0$. Let further an initial iterate $x^0 \in X$ and a step size $\tau > 0$ be given and let the sequence $\{x^k\}_{k\in\mathbb{N}}$ be generated by the forward-backward splitting method (or basic proximal point method if $F = 0$), i.e., by solving for x^{k+1} in

$$0 \in \tau[\partial G(x^{k+1}) + \nabla F(x^k)] + (x^{k+1} - x^k). \tag{29.5}$$

We also write $H := \partial G + \nabla F : X \rightrightarrows X$. Finally, it is worth recalling the approach of Chapter 10 for encoding convergence rate into "testing" parameters $\varphi_k > 0$.

We start our analysis by adapting the proofs of Theorems 10.2 and 11.4 to employ the squared distance function $x \mapsto \operatorname{dist}^2(x; \widehat{X})$ to the entire solution set $\widehat{X} = H^{-1}(0)$ in place of the squared distance function $x \mapsto \|x - \widehat{x}\|_X^2$ to a fixed $\widehat{x} \in H^{-1}(0)$.

Lemma 29.5. *Let $\widehat{X} \subset X$. If for all $k \in \mathbb{N}$ and $w^{k+1} := -\nabla F(x^k) - \tau^{-1}(x^{k+1} - x^k) \in \partial G(x^{k+1})$,*

$$\inf_{\bar{x}\in\widehat{X}} \left(\frac{\varphi_k}{2} \|x^{k+1} - \bar{x}\|_X^2 + \varphi_k \tau \langle w^{k+1} + \nabla F(x^k), x^{k+1} - \bar{x} \rangle_X \right) \geq \frac{\varphi_{k+1}}{2} \operatorname{dist}^2(x^{k+1}, \widehat{X}) - \frac{\varphi_k}{2} \|x^{k+1} - x^k\|_X^2, \tag{29.6}$$

then

$$\frac{\varphi_N}{2} \operatorname{dist}^2(x^N, \widehat{X}) \leq \frac{\varphi_1}{2} \operatorname{dist}^2(x^0, \widehat{X}) \quad (N \geq 1). \tag{29.7}$$

Proof. Inserting (29.5) into (29.6) yields

$$\inf_{\bar{x}\in H^{-1}(0)} \varphi_k \left(\frac{1}{2}\|x^{k+1}-x^k\|_X^2 + \frac{1}{2}\|x^{k+1}-\bar{x}\|_X^2 - \langle x^{k+1}-x^k, x^{k+1}-\bar{x}\rangle_X \right) \geq \frac{\varphi_{k+1}}{2}\operatorname{dist}^2(x^{k+1}; H^{-1}(0)). \tag{29.8}$$

Using the three-point formula (9.1), we can then rewrite (29.8) as

$$\frac{\varphi_k}{2}\operatorname{dist}^2(x^k; H^{-1}(0)) \geq \frac{\varphi_{k+1}}{2}\operatorname{dist}^2(x^{k+1}; H^{-1}(0)).$$

The claim now follows by a telescoping sum over $k = 0, \dots, N-1$. □

RATES FROM ERROR BOUNDS AND METRIC SUBREGULARITY

Our first approach for the satisfaction of (29.6) is based on *error bounds*, which we will prove using metric subregularity. The essence of error bounds is to prove for some $\theta > 0$ that

$$\|x^{k+1}-x^k\|_X \geq \theta\|x^{k+1}-\widehat{x}\|_X.$$

We slightly weaken this condition, and assume the bound to be relative to the entire solution set, i.e.,

$$\|x^{k+1}-x^k\|_X^2 \geq \theta \operatorname{dist}^2(x^{k+1}; H^{-1}(0)). \tag{29.9}$$

This bound holds under metric subregularity. We first need the following technical lemma on the iteration (29.5).

Lemma 29.6. *If* $0 < \tau L < 1$, *then*

$$\frac{1}{2}\|x^{k+1}-x^k\|_X^2 \geq \frac{\tau^2}{4(1+L^2\tau^2)}\operatorname{dist}^2(0, \partial G(x^{k+1}) + \nabla F(x^{k+1})).$$

Proof. Since $-(x^{k+1}-x^k) \in \tau[\partial G(x^{k+1}) + \nabla F(x^k)]$ by (29.5), we have

$$\frac{1}{2}\|x^{k+1}-x^k\|_X^2 = \frac{1}{2}\operatorname{dist}^2(0, \{-(x^{k+1}-x^k)\}) \geq \frac{1}{2}\operatorname{dist}^2(0, \tau[\partial G(x^{k+1}) + \nabla F(x^k)]). \tag{29.10}$$

The generalized Young's inequality for any $\alpha \in (0,1)$ then yields

$$\begin{aligned}
&\frac{1}{2}\operatorname{dist}^2(0, \tau[\partial G(x^{k+1}) + \nabla F(x^k)]) \\
&\qquad = \frac{\tau^2}{2}\operatorname{dist}^2(\nabla F(x^{k+1}) - \nabla F(x^k), \partial G(x^{k+1}) + \nabla F(x^{k+1})) \\
&\qquad = \inf_{q \in \partial G(x^{k+1})} \frac{\tau^2}{2}\|(\nabla F(x^{k+1}) - \nabla F(x^k)) - (q + \nabla F(x^{k+1}))\|_X^2 \\
&\qquad \geq \frac{\tau^2(1-\alpha^{-1})}{2}\|\nabla F(x^{k+1}) - \nabla F(x^k)\|_X^2 + \inf_{q \in \partial G(x^{k+1})} \frac{\tau^2(1-\alpha)}{2}\|q + \nabla F(x^{k+1})\|_X^2 \\
&\qquad \geq \frac{\tau^2(1-\alpha^{-1})L^2}{2}\|x^{k+1} - x^k\|_X^2 + \frac{\tau^2(1-\alpha)}{2}\operatorname{dist}^2(0, \partial G(x^{k+1}) + \nabla F(x^{k+1})),
\end{aligned}$$

where we have used in the last step that $1 - \alpha^{-1} < 0$ and that ∇F is Lipschitz continuous. Combining this estimate with (29.10), we obtain that

$$\frac{1 - \tau^2(1-\alpha^{-1})L^2}{2}\|x^{k+1} - x^k\|_X^2 \geq \frac{\tau^2(1-\alpha)}{2}\operatorname{dist}^2(0, \partial G(x^{k+1}) + \nabla F(x^{k+1})).$$

Rearranging and using that $1 > \tau^2(1-\alpha^{-1})L^2$ by assumption then yields

$$\frac{1}{2}\|x^{k+1} - x^k\|_X^2 \geq \frac{\theta}{2}\operatorname{dist}^2(0, \partial G(x^{k+1}) + \nabla F(x^{k+1})).$$

for

$$\theta := \frac{\tau^2(1-\alpha)}{1 - \tau^2(1-\alpha^{-1})L^2},$$

which for $\alpha = 1/2$ yields the claim. □

Metric subregularity then immediately yields the error bound (29.9).

Lemma 29.7. *Let H be metrically subregular at $\widehat{x}$ for $\widehat{w} = 0$ for $\kappa > 0$ and $\delta > 0$. If $0 < \tau L \leq 2$ and $x^{k+1} \in \mathbb{B}(\widehat{x}, \delta)$, then (29.9) holds with $\theta = \frac{\tau^2}{2\kappa^2(1+L^2\tau^2)}$.*

Proof. Combining Lemma 29.6 and the definition of metric subregularity yields

$$\frac{1}{2}\|x^{k+1} - x^k\|_X^2 \geq \frac{\tau^2}{4(1+L^2\tau^2)}\operatorname{dist}^2(0, H(x^{k+1})) \geq \frac{\tau^2}{4\kappa^2(1+L^2\tau^2)}\operatorname{dist}^2(x^{k+1}, H^{-1}(0)).$$ □

From this lemma, we now obtain *local* linear convergence of the forward-backward splitting method when H is metrically subregular at a solution.

Theorem 29.8. *Let H be metrically subregular at $\widehat{x} \in H^{-1}(0)$ for $\widehat{w} = 0$ for $\kappa > 0$ and $\delta > 0$. If $0 < \tau L \leq 2$ and $x^k \in \mathbb{B}(\widehat{x}, \delta)$ for all $k \in \mathbb{N}$, then (29.7) holds for $\varphi_{k+1} := \varphi_k(1+\theta)$ and $\varphi_0 = 1$ with $\theta = \frac{\tau^2}{2\kappa^2(1+L^2\tau^2)}$. In particular, $\operatorname{dist}^2(x^N; H^{-1}(0)) \to 0$ at a linear rate.*

Proof. Let $\bar{x} \in H^{-1}(0)$ and $w^{k+1} \in \partial G(x^{k+1})$ as in Lemma 29.5. From (10.11) in the proof of Theorem 10.2 and using $\tau L \leq 2$, we obtain

$$\langle w^{k+1} + \nabla F(x^k), x^{k+1} - \bar{x}\rangle_X \geq -\frac{L}{4}\|x^{k+1} - x^k\|_X^2 \geq -\frac{1}{2\tau}\|x^{k+1} - x^k\|_X^2.$$

Lemma 29.7 now yields the error bound (29.9) and hence for all $\bar{x} \in H^{-1}(0)$ that

$$\frac{\varphi_k}{2}\|x^{k+1} - x^k\|_X^2 + \frac{\varphi_{k+1} - \varphi_k\theta}{2}\|x^{k+1} - \bar{x}\|_X \geq \frac{\varphi_{k+1}}{2}\operatorname{dist}^2(x^{k+1}; H^{-1}(0)).$$

Summing these two estimates yields

$$\frac{\varphi_k}{2}\|x^{k+1} - x^k\|_X^2 + \frac{\varphi_k}{2}\|x^{k+1} - \bar{x}\|_X^2 + \varphi_k\tau\langle w^{k+1} + \nabla F(x^k), x^{k+1} - \bar{x}\rangle_X \geq \frac{\varphi_{k+1}}{2}\operatorname{dist}^2(x^{k+1}, H^{-1}(0)).$$

Taking the infimum over $\bar{x} \in H^{-1}(0)$, we obtain (29.6) for $\widehat{X} = H^{-1}(0)$. The claim now follows from Lemma 29.5 and the exponential growth of φ_k. □

The convergence is local due to the requirement $x^{k+1} \in \mathbb{B}(\widehat{x}, \delta)$ for applying subregularity. In finite dimensions, the weak convergence result of Theorem 9.6 of course guarantees that the iterates enter and remain in this neighborhood after a finite number of steps.

RATES FROM STRONG SUBMONOTONICITY

If H is instead strongly submonotone, we can (locally) ensure (29.6) directly.

Theorem 29.9. *Let H be $(\gamma/2, \theta/2)$-strongly submonotone at $\widehat{x} \in H^{-1}(0)$ for $\widehat{w} = 0$ for $\delta > 0$. If $\gamma > \theta + L^2\tau$ and $x^0 \in \mathbb{B}(\widehat{x}, \varepsilon)$ for some $\varepsilon > 0$ sufficiently small, then (29.7) holds for $\varphi_{k+1} := \varphi_k(1 + (\gamma - L^2\tau)\tau)$ and $\varphi_0 = 1$. In particular, $\operatorname{dist}^2(x^N; H^{-1}(0)) \to 0$ at a linear rate.*

Proof. Let $w^{k+1} := -\tau^{-1}(x^{k+1} - x^k) - \nabla F(x^k) \in \partial G(x^{k+1})$ by (29.5). By (9.9) in the proof of Theorem 9.6, if $x^0 \in \mathbb{B}(\widehat{x}, \varepsilon)$ for $\varepsilon > 0$ small enough, then $\|x^{k+1} - x^k\|_X \leq \delta/(L + \tau^{-1})$ for all $k \in \mathbb{N}$ such that the Lipschitz continuity of ∇F yields

$$\|\nabla F(x^{k+1}) - \nabla F(x^k) - \tau^{-1}(x^{k+1} - x^k)\|_X \leq \delta.$$

Thus $w^{k+1} \in \partial G(x^{k+1}) \cap \mathbb{B}(-\nabla F(x^{k+1}), \delta)$ and $x^{k+1} \in \mathbb{B}(\widehat{x}, \delta)$ for all $k \in \mathbb{N}$. Now, for all $\bar{x} \in H^{-1}(0)$, the strong submonotonicity of H at $\widehat{x}$ for 0 implies that

$$\varphi_k \tau \langle w^{k+1} + \nabla F(x^{k+1}), x^{k+1} - \bar{x} \rangle_X + \frac{(\theta - \gamma)\varphi_k \tau}{2} \|x^{k+1} - \bar{x}\|_X^2 \geq \frac{\theta \varphi_k \tau}{2} \operatorname{dist}^2(x^{k+1}; H^{-1}(0))$$

for all $k \in \mathbb{N}$. Cauchy's inequality and the Lipschitz continuity of ∇F then yields

$$\varphi_k \tau \langle \nabla F(x^k) - \nabla F(x^{k+1}), x^{k+1} - \bar{x} \rangle_X \geq -\frac{\varphi_k}{2} \|x^{k+1} - x^k\|_X^2 - \frac{\varphi_k \tau^2 L^2}{2} \|x^{k+1} - \bar{x}\|_X^2.$$

We now sum the last two inequalities to obtain

$$\begin{aligned} &\frac{\varphi_k}{2} \|x^{k+1} - x^k\|_X^2 + \varphi_k \tau \langle w^{k+1} + \nabla F(x^k), x^{k+1} - \bar{x} \rangle_X \\ &\qquad \geq \frac{\theta \varphi_k \tau}{2} \operatorname{dist}^2(x^{k+1}; H^{-1}(0)) + \frac{(\gamma - \theta - L^2 \tau)\varphi_k \tau}{2} \|x^{k+1} - \bar{x}\|_X^2. \end{aligned}$$

Using that $\theta - \gamma + L^2 \tau < 0$ and taking the infimum over all $\bar{x} \in H^{-1}(0)$ then yields

$$\begin{aligned} &\inf_{\bar{x} \in H^{-1}(0)} \left(\frac{\varphi_k}{2} \|x^{k+1} - \bar{x}\|_X^2 + \varphi_k \tau \langle w^{k+1} + \nabla F(x^k), x^{k+1} - \bar{x} \rangle_X \right) \\ &\qquad \geq \frac{(\gamma - L^2 \tau)\varphi_k \tau + \varphi_k}{2} \operatorname{dist}^2(x^{k+1}; H^{-1}(0)) - \frac{\varphi_k}{2} \|x^{k+1} - x^k\|_X^2. \end{aligned}$$

Since $\gamma - L^2 \tau > 0$ and $\varphi_{k+1} = \varphi_k(1 + (\gamma - L^2 \tau)\tau)$, this shows (29.6) with $\widehat{X} = H^{-1}(0)$. The claim now follows from Lemma 29.5 and the exponential growth of φ_k. □

Remark 29.10. Similarly to Theorem 10.1 (ii), if $F \equiv 0$ we can let $\tau \nearrow \infty$ to obtain local superlinear convergence of the proximal point method under strong submonotonicity of ∂G at the solution.

Remark 29.11 (local linear convergence). Local linear convergence was first derived from error bounds in [Luo and Tseng, 1992] for matrix splitting problems and was studied for other methods, including the ADMM and the proximal point method among others, in [Aspelmeier et al., 2016; Han and Yuan, 2013; Leventhal, 2009; Li and Mordukhovich, 2012]. An alternative approach to the proximal point method was taken in [Aragón Artacho and Gaydu, 2012] based on Lyusternik–Graves-style estimates, while [Adly et al., 2015] presented an approach based on metric regularity to Newton's method for variational inclusions. Furthermore, [Zhou and So, 2017] proposed a unified approach to error bounds for generic smooth constrained problems. Finally, [Liu et al., 2018; Valkonen, 2021c] introduced *partial* or subspace versions of error bounds and showed the fast convergence of only some variables of structured algorithms such as the ADMM or PDPS. The relationships between error bounds and metric subregularity is studied in more detail in [Dontchev and Rockafellar, 2014; Gfrerer, 2011; Ioffe, 2017; Kruger, 2015; Ngai and Théra, 2008]. Submonotonicity was introduced in [Valkonen, 2021c].

Part V

APPLICATIONS

30 SPARSE REGULARIZATION

In this and the following chapters, we illustrate the application of the results and methods of the previous parts to selected nonsmooth optimization problems.

We first study the application of the optimization theory and methods that we have developed to the solution of some *inverse problems*, including imaging problems, which we treat in finite dimensions to avoid technical difficulties unrelated to nonsmooth optimization. In a nutshell, inverse problems consist in trying to obtain quantities of interest that are not directly accessible by combining measured (incomplete, noisy) data with a mathematical model linking the desired quantity to the predicted measurements. Such problems are usually *ill-posed* in the sense that a solution may not exist, may not be unique, or may not be stable with respect to perturbations of the data. Hence one needs to apply *regularization* to obtain a stable approximation. For an introduction to the regularization of inverse problems, we refer the reader to the seminal work [Engl et al., 1996] as well as to the more recent [Clason, 2020b; Ito and Jin, 2014]. One particular approach is *Tikhonov regularization*, which consists in solving an optimization problem that involves the sum of (a) a *data term* that matches the model prediction against available data and of (b) a *regularization term* that attempts to promote expected and desirable features in the reconstruction (and is typically required to obtain well-posedness of the regularized problem). An increasingly popular class of regularization terms promotes "sparsity" of the solution in the sense that it can explain the data with a minimal number of features; as we will see, such terms require nonsmooth optimization. This class (and nonsmooth optimization in general) is particularly relevant in the context of *mathematical image processing*, where the quantity of interest is an image rather than an abstract physical parameter; see, e.g., [Bredies and Lorenz, 2018; Scherzer et al., 2009].

In this chapter we start with perhaps the simplest nonsmooth regularization of an inverse problem: ℓ^1-regularized data-fitting, sometimes known as the *Lasso problem*. The starting point is linear regression, but we wish to explain the data "in simple terms" only through its most important features. We then move on to signal recovery applications in the next Chapters 31 and 32.

30.1 PROBLEM DESCRIPTION

Let $b_i \in \mathbb{R}$ be a single measurement of an unknown signal $x \in \mathbb{R}^M$ through the filter $a_i \in \mathbb{R}^M$. Without the presence of noise, $b_i = a_i^T x$ for the $i = 1, \dots, N$ measurements. In statistical contexts, b_i is known as a dependent variable and a_i as a data vector. Since each b_i and a_i may be noisy, and the system

$$a_i^T x = b_i, \quad (i = 1, \dots, N),$$

may be over- or under-determined, direct solution of x from this system is not in general well-posed. Basic linear regression instead seeks the least squares solution x through solution of the optimization problem

$$\min_{x \in \mathbb{R}^M} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} (b_i - a_i^T x)^2. \tag{30.1}$$

To explain the data $\{(a_i, b_i)\}$ through its most important features, we want x to be sparse, i.e., to have many zero elements, and few nonzero elements. For example, a_i might be the attributes (genre, length, etc.) of a film, and b_i its rating. A sparse vector x would then contain only the most relevant attributes for the rating and their relative weighting. To perform such *sparse regression*, let us add to the data fitting term of (30.1) the regularization term $g(x) = \lambda \|x\|_1$. Then we obtain the so-called *Lasso problem*

$$\min_{x \in \mathbb{R}^M} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} (b_i - a_i^T x)^2 + \lambda \|x\|_1 \tag{30.2}$$

The hope is that to explain the data, the ℓ^1-norm regularization term will cause the minimizer to select more relevant features from the data, ignoring irrelevant ones.

In the following, we write (30.2) more succinctly as

$$\min_{x \in \mathbb{R}^M} J(x) \quad \text{for} \quad J(x) := F(x) + G(x), \tag{30.3}$$

where

$$A := (a_1, \dots, a_N)^T \in \mathbb{R}^{N \times M}, \quad F(x) := \frac{1}{2} \|Ax - b\|_2^2, \quad \text{and} \quad G(y) := \lambda \|x\|_1.$$

30.2 OPTIMALITY CONDITIONS

Our first result characterizes the solutions of (30.3).

Theorem 30.1. *The vector $\widehat{x} \in \mathbb{R}^M$ is a solution to (30.2) if and only if there exists a $\widehat{p} \in \mathbb{R}^M$ such that*

$$-A^*(A\widehat{x} - b) = \lambda \widehat{p} \qquad \text{and} \qquad \widehat{p}_i \in \begin{cases} \{1\} & \text{if } \widehat{x}_i > 0, \\ \{-1\} & \text{if } \widehat{x}_i < 0, \\ [-1, 1] & \text{if } \widehat{x}_i = 0. \end{cases} \tag{30.4}$$

Proof. Since A is linear and F and G are convex, J is convex as well. Therefore the convex Fermat principle of Theorem 4.2 is an equivalent characterization of solutions to (30.2) as those $\widehat{x}$ satisfying $0 \in \partial J(\widehat{x})$. Since both F and G have full domain and are proper and lower semicontinuous, we may further use the subdifferential sum rule of Theorem 4.14 to deduce for all $x \in \mathbb{R}^M$ that $\partial J(x) = \partial F(x) + \partial G(x)$. Since F is differentiable, using Theorem 4.5 we therefore characterize the solutions as those points $\widehat{x}$ satisfying

$$-\nabla F(\widehat{x}) \in \partial G(\widehat{x}). \tag{30.5}$$

Since F is smooth, expanding $\nabla F(\widehat{x}) = A^*(A\widehat{x} - b)$ and using Example 4.7 to calculate $\partial G(\widehat{x})$ componentwise yields (30.4). □

Note the complementarity between the primal variable $\widehat{x}$ and the dual variable $\widehat{p}$, which yields the desired sparsity: a component $\widehat{x}_i$ is zero if the corresponding scaled and "back-propagated" residual $\widehat{p}_i$ is smaller than 1 in magnitude. However, $\widehat{x}_i$ can be zero even if $|\widehat{p}_i| = 1$; if this case can be excluded, we say that *strict complementarity* holds, i.e.,

$$\text{either } \widehat{x}_i \neq 0 \text{ or } |\widehat{p}_i| < 1, \quad (i = 1, \dots, M). \tag{30.6}$$

Thus strict complementarity avoids, whenever $\widehat{x}_i = 0$, the boundary cases $|\widehat{p}_i| = 1$ that happen when $\widehat{x} \neq 0$.

30.3 ALGORITHMS

The starting point for deriving implementable algorithms for the solution of (30.2) is the following reformulation of the optimality conditions using the proximal point mapping.

Lemma 30.2. *The vector $\widehat{x} \in \mathbb{R}^M$ is a solution to (30.2) if and only if*

$$\widehat{x} = \operatorname{prox}_{\tau G}(\widehat{x} - \tau A^*(A\widehat{x} - b)). \tag{30.7}$$

Proof. Applying Lemma 6.21 to G, we may rewrite (30.5) for any $\tau > 0$ as

$$\widehat{x} = \operatorname{prox}_{\tau G}(\widehat{x} - \tau \nabla F(\widehat{x})),$$

which after inserting $\nabla F(\widehat{x}) = A^*(A\widehat{x} - b)$ yields (30.7). □

FORWARD-BACKWARD SPLITTING

The forward-backward or explicit splitting method of (8.6) is our first iterative method for solving (30.2). As we did in the general setting in Chapter 8, the method can be directly developed from the proximal-form optimality conditions (30.7). First, using Example 6.25 (ii) we write the proximal map of G in terms of the soft-thresholding operator as

$$\operatorname{prox}_{\tau G}(x) = (\operatorname{soft}_{\lambda\tau}(x_1), \dots, \operatorname{soft}_{\lambda\tau}(x_M)) \quad \text{for} \quad \operatorname{soft}_\theta(t) := \begin{cases} t-\theta & \text{if } t > \theta, \\ 0 & \text{if } t \in [-\theta, \theta], \\ t+\theta & \text{if } t < -\theta. \end{cases}$$

Then (8.6) becomes

$$\begin{aligned} x^{k+1} &:= \operatorname{prox}_{\tau G}(x^k - \tau \nabla F(x^k)) \\ &= \operatorname{soft}_{\lambda\tau}((\mathrm{Id} - \tau A^* A)x^k + \tau A^* b). \end{aligned} \tag{30.8}$$

Under mild conditions, the iterates converge.

Theorem 30.3. *Suppose $\tau\|A\|^2 < 2$. Then for any starting point $x^0 \in \mathbb{R}^M$, the iterates $\{x^k\}_{k\in\mathbb{N}}$ generated by (30.8) converge to a solution $\widehat{x}$ of (30.3).*

Proof. The Lipschitz factor of $\nabla F(x) = A^*(Ax - b)$ is $\|A\|^2$. Therefore, the claim follows from Theorem 9.6. □

Convergence of function values can be similarly deduced from Theorem 11.4.

Theorem 30.3 provides no convergence rates as, indeed, no rates for iterates are in general known for forward-backward splitting without some sort of stronger growth assumptions. However, Theorem 30.10 in Section 30.4 below will show that $\partial[F+G]$ is metrically regular at $\widehat{x}$ for 0, provided that the strict complementarity condition (30.6) holds. Since this implies metric subregularity, Theorem 29.8 can be used to demonstrate the local linear convergence of (30.8) near a strictly complementary solution.

We can also apply the inertial FISTA method of (12.35) to (30.3). Based on the basic explicit splitting (30.8), this method becomes

$$\begin{cases} x^{k+1} = \operatorname{soft}_{\lambda\tau}((\mathrm{Id} - \tau A^* A)\bar{x}^k + \tau A^* b), \\ \alpha_{k+1} := \lambda_{k+1}(\lambda_k^{-1} - 1), \\ \bar{x}^{k+1} := (1+\alpha_{k+1})x^{k+1} - \alpha_{k+1}x^k. \end{cases}$$

The initial inertial parameter $\lambda_0 = 1$, while $\bar{x}^0 \in \mathbb{R}^M$ can be chosen freely. Since $\alpha_1 = 0$, x^0 is never used. Regarding convergence, Theorem 12.12 readily gives the following result.

Theorem 30.4. *Suppose $\tau\|A\|^2 \leq 1$. Then for any starting point $x^0 \in \mathbb{R}^M$, the iterates $\{x^k\}_{k\in\mathbb{N}}$ generated by (30.3) satisfy $J(x^k) \to \min J$ at the rate $O(1/k^2)$.*

In fact, under strict complementarity, zeroes are identified in a finite number of steps.

Theorem 30.5. *Assume the solution $\widehat{x} \in \mathbb{R}^M$ to (30.2) is unique and satisfies strict complementarity. Then there exists $K \in \mathbb{N}$ such that the iterates $\{x^k\}_{k\in\mathbb{N}}$ generated by (30.3) satisfy $x_i^k = \widehat{x}_i$ for all $k \geq K$ and $i = 1, \dots, N$ with $\widehat{x}_i = 0$.*

Proof. Indeed, forward-backward splitting for $\min_x(F + G)$ and a step length $\tau > 0$ by definition satisfies

$$0 \in \partial G(x^{k+1}) + \nabla F(x^k) + \tau(x^{k+1} - x^k). \tag{30.9}$$

Following the proof of Theorem 9.6, we have $\|x^{k+1} - x^k\| \to 0$ and $x^{k+1} \to \widehat{x}$ provided the solution $\widehat{x}$ is unique. It follows that $\nabla F(x^k) \to \nabla F(\widehat{x})$. Furthermore, strict complementarity yields $-[\nabla F(\widehat{x})]_i \in (-\lambda, \lambda)$ for all i with $\widehat{x}_i = 0$, and hence for those same i it holds that $-[\nabla F(x^k)]_i \in (-\lambda, \lambda)$ for all $k \geq K$ for some $K \in \mathbb{N}$. By (30.9) and $\|x^{k+1} - x^k\| \to 0$, it is then necessary that $[\partial G(x^{k+1})]_i$ contains a point in $(-\lambda, \lambda)$. This is only possible if $x_i^{k+1} = 0$ after a finite number of steps. □

Remark 30.6 (unconditional linear convergence and activity identification). It is shown in [Bolte et al., 2017] through error bounds that forward-backward splitting for the Lasso problem converges linearly without any assumptions. Error bounds can also be proved more generally based on piecewise polynomial properties derived in [Li, 2013]. These are used in [Garrigos et al., 2020] to obtain error bounds in separable Hilbert spaces. This follows earlier works such as [Bredies and Lorenz, 2008] with stricter assumptions. In the former also a "finite identification property" is studied following earlier efforts in [Lewis, 2002; Liang et al., 2014], among others; it can be shown that the forward-backward splitting and other methods converge in a finite number of steps to a smooth submanifold. As verified by elementary analysis in Theorem 30.5, in the case of the Lasso problem, forward-backward splitting identifies the strictly complementary zeroes in a finite number of steps.

SEMISMOOTH NEWTON METHOD

By Theorem 30.1, we know that minimizers $\bar{x}$ of (30.3) satisfy

$$\bar{x} - \operatorname{prox}_{\gamma G}(\bar{x} - \gamma\nabla F(\bar{x})) = 0$$

for any $\gamma > 0$. We therefore look for a root of

$$H(x) \coloneqq x - \operatorname{prox}_{\gamma G}(x - \gamma\nabla F(x)).$$

If we can produce an invertible and well-conditioned Newton derivative $D_N H(x)$ for all x in a sufficiently large neighborhood of $\bar{x}$, this can be done with the semismooth Newton method (14.4), i.e., solving s^k from $D_N H(x^k)s_k = -H(x^k)$ and updating $x^{k+1} = x^k + s^k$.

In fact, let $T(x) := x - \gamma\nabla F(x)$ and consider the composition $\mathrm{prox}_{\gamma G} \circ T$. We may use Theorem 14.3 to obtain $D_N T(x) = \mathrm{Id} - \gamma\nabla^2 F(x)$, and Example 14.10 (ii) to obtain $D_N \mathrm{prox}_{\gamma G}$. Both $D_N \mathrm{prox}_{\gamma G}$ and $D_N T$ are locally uniformly bounded (obviously from the characterization and the continuous differentiability, respectively). Thus, we are justified in using the chain rule from Theorem 14.4 on the composition to calculate

$$\begin{aligned} D_N H(x) &= \mathrm{Id} - D_N \mathrm{prox}_{\gamma G}(T(x)) \circ D_N T(x) \\ &= \mathrm{Id} - \mathbb{1}_{\mathcal{A}(x)}[\mathrm{Id} - \gamma\nabla^2 F(x)] \\ &= \mathbb{1}_{\mathcal{I}(x)} + \gamma\mathbb{1}_{\mathcal{A}(x)}\nabla^2 F(x), \end{aligned}$$

where we have defined the *inactive* and *active sets*, respectively, as

$$\mathcal{I}(x) := \{i \in \{1,\dots,N\} \mid |x_i - \gamma[\nabla F(x)]_i| < \gamma\}, \qquad \mathcal{A}(x) := \{1,\dots,M\} \setminus \mathcal{I}(x). \tag{30.10}$$

The matrix $D_N H(x)$ may in general not be invertible, or may be poorly conditioned *on the active components*, as we will soon see in more detail. For some $\theta > 0$, we therefore replace it with the *active-dampened* matrix

$$M(x) := \mathbb{1}_{\mathcal{I}(x)} + \gamma\mathbb{1}_{\mathcal{A}(x)}\nabla^2 F(x) + \theta\mathbb{1}_{\mathcal{A}(x)}. \tag{30.11}$$

Write $P_{\mathcal{I}(x)}$ and $P_{\mathcal{A}(x)}$ for the projections to the inactive and active components, so that $\mathbb{1}_{\mathcal{I}(x)} = P^*_{\mathcal{I}(x)}P_{\mathcal{I}(x)}$, and likewise for the active components. Thus the active-dampened semismooth Newton step s^k is determined by

$$\left(\mathbb{1}_{\mathcal{I}(x^k)} + \gamma\mathbb{1}_{\mathcal{A}(x^k)}\nabla^2 F(x^k) + \theta\mathbb{1}_{\mathcal{A}(x^k)}\right) s^k = -x^k + \mathrm{prox}_{\gamma G}(x^k - \gamma\nabla F(x^k)). \tag{30.12}$$

Since the proximal point mapping of G is the soft shrinkage operator, we have using the definition of the inactive set that

$$P_{\mathcal{I}(x^k)}\mathrm{prox}_{\gamma G}(x^k - \gamma\nabla F(x^k)) = 0.$$

Hence, multiplying (30.12) from the left by $P_{\mathcal{I}(x^k)}$, we deduce that $P_{\mathcal{I}(x^k)}s^k = -P_{\mathcal{I}(x^k)}x^k$. Thus $s_i^k = -x_i^k$ for the inactive components $i \in \mathcal{I}(x^k)$. It follows that $x_i^{k+1} = 0$ for $i \in \mathcal{I}(x^k)$. On the other, writing $s^k = \mathbb{1}_{\mathcal{A}(x^k)}s^k + \mathbb{1}_{\mathcal{I}(x^k)}s^k = P^*_{\mathcal{A}(x^k)}P_{\mathcal{A}(x^k)}s^k - \mathbb{1}_{\mathcal{I}(x^k)}x^k$ and multiplying (30.12) from the left by $P_{\mathcal{A}(x^k)}$ yields

$$\begin{aligned} &[\gamma P_{\mathcal{A}(x^k)}\nabla^2 F(x^k)P^*_{\mathcal{A}(x^k)} + \theta\mathrm{Id}]P_{\mathcal{A}(x^k)}s^k \\ &\qquad = P_{\mathcal{A}(x^k)}(-x^k + \mathrm{prox}_{\gamma G}(x^k - \gamma\nabla F(x^k)) + \gamma\nabla^2 F(x^k)\mathbb{1}_{\mathcal{I}(x^k)}x^k). \end{aligned} \tag{30.13}$$

Since $P_{\mathcal{A}(x^k)}\nabla^2 F(x^k)P^*_{\mathcal{A}(x^k)} + \theta\mathrm{Id}$ is positive definite, we can solve this for $P_{\mathcal{A}(x^k)}s^k$. Altogether, therefore, the semismooth Newton method for (30.3) becomes

(i) form the inactive and active sets $\mathcal{I}(x^k)$ and $\mathcal{A}(x^k)$ following (30.10);

(ii) solve $P_{\mathcal{A}(x^k)} s^k$ from (30.13);

(iii) update $x^{k+1} := \mathbb{1}_{\mathcal{A}(x^k)}(x^k + s^k)$.

This coincides with an *active set strategy* similar to those used for solving quadratic subproblems in sequential programming methods with inequality constraints; cf. [Ito and Kunisch, 2008, Chapter 8.4].

For convergence, we need to assume that $P_{\mathcal{A}(\bar{x})} \nabla^2 F(\bar{x}) P^*_{\mathcal{A}(\bar{x})}$ is invertible. Practically this means that there are more measurements than attributes that describe the measurements. Although superlinear convergence has superficially no stricter conditions than linear convergence, the convergence radius can in practice be smaller, and hence convergence may not hold for an arbitrary initial iterate x^0.

To improve readability of the next theorem proving these properties, we recall the following "operator Young's inequality".

Lemma 30.7. *On Hilbert spaces X and Y, let $A \in \mathbb{L}(X;Y)$ and $B \in \mathbb{L}(X;Y)$. Then for any $\beta > 0$, we have*

$$2A^*B \preceq \beta A^*A + \beta^{-1} B^*B,$$

where $A \preceq B$ means that $B - A$ is positive semi-definite.

Proof. Take any $x \in X$. Then using the Cauchy–Schwarz and Young's inequality yields

$$2\langle x, A^*Bx\rangle_X = 2\langle Ax, Bx\rangle_Y \le \beta \|Ax\|_Y^2 + \beta^{-1}\|Bx\|_Y^2 = \langle x, (\beta A^*A + \beta^{-1}B^*B)x\rangle_X.$$

Since this holds for all $x \in X$, this means that $(\beta A^*A + \beta^{-1}B^*B) - 2A^*B$ is positive semi-definite. □

Theorem 30.8. *Let $\bar{x}$ be a (unique) minimizer of (30.3). Let $\gamma, \theta > 0$ satisfy $2(1-\theta^2) > \gamma\theta\|A\|^2$, and suppose that $P_{\mathcal{A}(\bar{x})} A^* A P^*_{\mathcal{A}(\bar{x})}$ is positive definite. If x^0 is sufficiently close to $\bar{x}$, then the sequence $\{x^{k+1}\}_{k\in\mathbb{N}}$ generated by iterating (i)–(iii) above converges linearly to $\bar{x}$. If $\theta = 0$, and $\gamma > 0$ is arbitrary, the convergence is superlinear.*

Proof. We first consider linear convergence. Let $M(x)$ be given by (30.11). Then

$$\|M(x) - D_N H(x)\|_{\mathbb{L}(\mathbb{R}^N;\mathbb{R}^N)} = \|\theta \mathbb{1}_{\mathcal{A}(x)}\|_{\mathbb{L}(\mathbb{R}^N;\mathbb{R}^N)} \le \theta,$$

so the corresponding assumption of Theorem 14.2 (applied to H in place of F) holds. To apply the theorem, we still need to prove $\|M(x)^{-1}\|_{\mathbb{L}(\mathbb{R}^N;\mathbb{R}^N)} \le C$ for all $x \in U$ for some

neighborhood U of $\bar{x}$ and some $C > 0$ with $C\theta < 1$. That is to say, $M(x)^*M(x) \geq C^{-2}\,\mathrm{Id}$. We expand

$$\begin{aligned} M(x)^*M(x) &= \mathbb{1}_{\mathcal{I}(x)} + \gamma^2 A^*A\mathbb{1}_{\mathcal{A}(x)}A^*A + \theta\gamma\mathbb{1}_{\mathcal{A}(x)}A^*A + \theta\gamma A^*A\mathbb{1}_{\mathcal{A}(x)} + \theta^2\mathbb{1}_{\mathcal{A}(x)} \\ &= \mathbb{1}_{\mathcal{I}(x)} + \gamma^2 A^*A\mathbb{1}_{\mathcal{A}(x)}A^*A + 2\theta\gamma\mathbb{1}_{\mathcal{A}(x)}A^*A\mathbb{1}_{\mathcal{A}(x)} + \theta^2\mathbb{1}_{\mathcal{A}(x)} \\ &\quad + \theta\gamma\mathbb{1}_{\mathcal{A}(x)}A^*A\mathbb{1}_{\mathcal{I}(x)} + \theta\gamma\mathbb{1}_{\mathcal{I}(x)}A^*A\mathbb{1}_{\mathcal{A}(x)}. \end{aligned} \tag{30.14}$$

Eliminating the second term by positive semi-definiteness, and applying Lemma 30.7 to the last two terms yields for any $\beta > 0$ that

$$M(x)^*M(x) \geq \mathbb{1}_{\mathcal{I}(x)} + (2-\beta)\theta\gamma\mathbb{1}_{\mathcal{A}(x)}A^*A\mathbb{1}_{\mathcal{A}(x)} + \theta^2\mathbb{1}_{\mathcal{A}(x)} - \beta^{-1}\theta\gamma\mathbb{1}_{\mathcal{I}(x)}A^*A\mathbb{1}_{\mathcal{I}(x)}. \tag{30.15}$$

By assumption, $P_{\mathcal{A}(x)}A^*AP^*_{\mathcal{A}(x)}$ is positive definite and $\mathcal{A}(x) = \mathcal{A}(\bar{x})$ for all x in some open neighborhood U of $\bar{x}$. Therefore $\mathbb{1}_{\mathcal{A}(x)}A^*A\mathbb{1}_{\mathcal{A}(x)} \geq \varepsilon\mathbb{1}_{\mathcal{A}(x)}$ for some $\varepsilon > 0$ and all $x \in U$. Consequently, it follows from (30.15) that

$$M(x)^*M(x) \geq (1 - \beta^{-1}\gamma\theta\|A\|^2)\mathbb{1}_{\mathcal{I}(x)} + ((2-\beta)\theta\gamma\varepsilon + \theta^2)\mathbb{1}_{\mathcal{A}(x)} \quad \text{for all} \quad x \in U.$$

We have $M(x)^*M(x) \geq C^{-2}\,\mathrm{Id}$ for some $C > 0$ with $C\theta < 1$ if both factors in this expression are strictly greater than θ^2. For the second factor, this follows from taking *any* $\beta \in (0,2)$. Minding our assumption $2(1-\theta^2) > \gamma\theta\|A\|^2$, also the first factor is greater than θ^2 for *some* $\beta \in (0,2)$. The linear convergence claim now follows from Theorem 14.2.

To show superlinear convergence when $\theta = 0$, we will use Theorem 14.1, which requires us to show that $\|D_N H(x)^{-1}\|_{\mathbb{L}(\mathbb{R}^n;\mathbb{R}^n)} \leq C$ for some $C > 0$. Since now $M = D_N H$, this amounts to showing $C^{-2}\,\mathrm{Id} \leq M(x)^*M(x)$. We expand

$$\begin{aligned} A^*A\mathbb{1}_{\mathcal{A}(x)}A^*A &= (\mathbb{1}_{\mathcal{A}(x)}A^*A\mathbb{1}_{\mathcal{A}(x)})^2 + \mathbb{1}_{\mathcal{A}(x)}A^*A\mathbb{1}_{\mathcal{A}(x)}A^*A\mathbb{1}_{\mathcal{I}(x)} \\ &\quad + \mathbb{1}_{\mathcal{I}(x)}A^*A\mathbb{1}_{\mathcal{A}(x)}A^*A\mathbb{1}_{\mathcal{A}(x)} + \mathbb{1}_{\mathcal{I}(x)}A^*A\mathbb{1}_{\mathcal{A}(x)}A^*A\mathbb{1}_{\mathcal{I}(x)}. \end{aligned}$$

We then apply Lemma 30.7 to the middle terms and follow with $\mathbb{1}_{\mathcal{A}(x)}A^*A\mathbb{1}_{\mathcal{A}(x)} \geq \varepsilon\mathbb{1}_{\mathcal{A}(x)}$ to obtain for any $\mu > 0$ the bound

$$\begin{aligned} \gamma^2 A^*A\mathbb{1}_{\mathcal{A}(x)}A^*A &\geq (1-\mu)(\mathbb{1}_{\mathcal{A}(x)}A^*A\mathbb{1}_{\mathcal{A}(x)})^2 - (\mu^{-1}-1)\mathbb{1}_{\mathcal{I}(x)}A^*A\mathbb{1}_{\mathcal{A}(x)}A^*A\mathbb{1}_{\mathcal{I}(x)} \\ &\geq (1-\mu)\varepsilon^2\mathbb{1}_{\mathcal{A}(x)} - (1-\mu^{-1})\|A\|^4\mathbb{1}_{\mathcal{I}(x)}. \end{aligned}$$

Inserting this lower bound into the expansion $M(x)^*M(x) = \mathbb{1}_{\mathcal{I}(x)} + \gamma^2 A^*A\mathbb{1}_{\mathcal{A}(x)}A^*A$ from (30.14), we deduce for some $\mu \in (0,1)$ the existence of $C > 0$ such that $C^{-2}\,\mathrm{Id} \leq M(x)^*M(x)$. Superlinear convergence now follows from Theorem 14.1. □

Remark 30.9. The superlinear convergence of semismooth Newton methods for (30.3) was proved by [Griesse and Lorenz, 2008].

NUMERICAL ILLUSTRATION

To give a practical perspective on the above algorithms, we illustrate their performance on a simple numerical example. We take $x \in \mathbb{R}^{1024}$ and $A \in \mathbb{L}(\mathbb{R}^{1024}; \mathbb{R}^{128})$ as convolution with a Gaussian kernel (standard deviation $\sigma = 7$ on the domain $[0, 1024]$) followed by subsampling. To generate the data b, we apply A to the true solution depicted in Figure 30.1, and apply normally distributed noise of variance 0.03. As regularization parameter, we take $\lambda = 0.008$. For all algorithms, we use the initial iterate $x^0 = 0$. For the first-order methods we take the step length $\tau = 0.9/L^2$, where L is an estimate of $\|A\|$. For the SSN method, we take the proximal parameter $\gamma = 100/L^2$. Since the basic SSN method ($\theta = 0$) does not exhibit convergence, we use the active-dampened variant ($\theta > 0$). Further details on the experimental setup can be found in the accompanying code [Clason and Valkonen, 2023].

We show the data and the reconstructions in Figure 30.1 and algorithm performance in Figures 30.2 and 30.3. As predicted by the theory, the inertial acceleration of FISTA (30.3) makes it faster than the unaccelerated forward-backward splitting method (30.8). Since the SSN method has to be dampened and hence converges only linearly in this ill-posed setting, it is clearly outperformed by FISTA.

30.4 STABILITY UNDER PERTURBATIONS

We now study stability of solutions to the ℓ^1-regularized least problem (30.3). We add a further perturbation parameter $p \in \mathbb{R}^N$ to J, setting

$$J(x; p) := \frac{1}{2}\|Ax - b - p\|_2^2 + \lambda\|x\|_1,$$

so that $J(x) := J(x; 0)$. Then

$$\partial_x J(x; p) = A^*(Ax - b - p) + \lambda\partial\|\cdot\|_1(x),$$

so that for perturbed data the solution mapping is given by

$$S(p) := \{x \in \mathbb{R}^M \mid 0 \in \partial_x J(x; p)\} = \{x \in \mathbb{R}^M \mid A^*p \in \partial J(x)\} = [(\partial J)^{-1} \circ A^*](p).$$

The next result shows that the Lasso problem is data-stable at the solution $\widehat{x}$ for data b if (for simplicity) the solution is strictly complementary, and the matrix A^*A is invertible on the subspace corresponding to the active (i.e., explaining) features. This is the same condition as in the convergence Theorem 30.8 for the SSN method. Indeed, due to optimality of $\widehat{x}$ and the strict complementarity of $\widehat{x}$ and $\widehat{p}$, $\mathcal{I}$ is the same set as $\mathcal{I}(\widehat{x})$ defined for the SSN method in (30.10). This can seen by using the proximal characterization (8.5) of the optimality conditions, and the zero-projection properties of the soft-thresholding operator, Example 6.25 (ii).

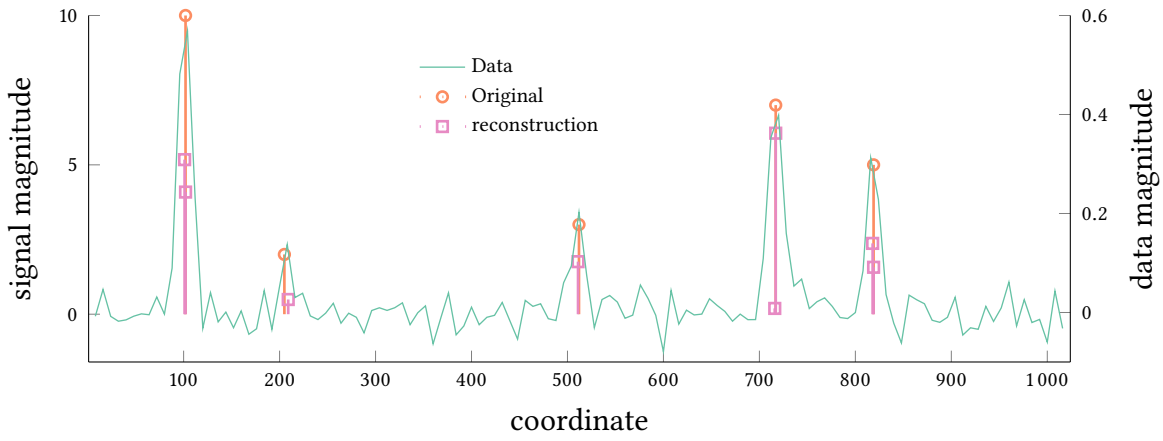


Figure 30.1: Sparse reconstruction data and result.

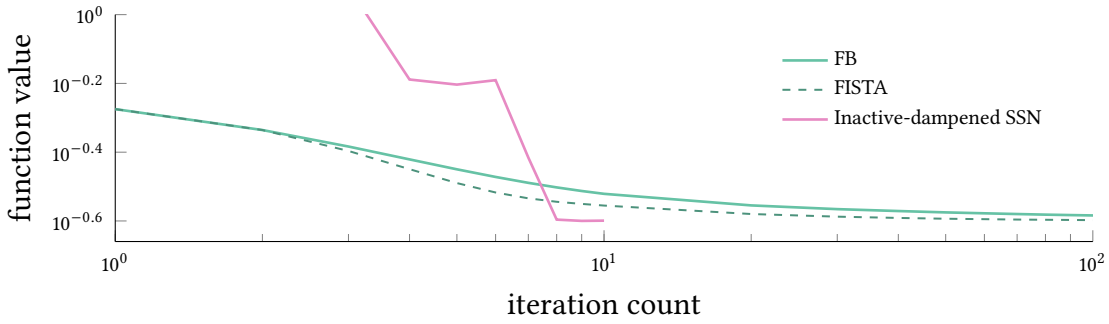


Figure 30.2: Sparse reconstruction algorithm performance: iterations versus function value.

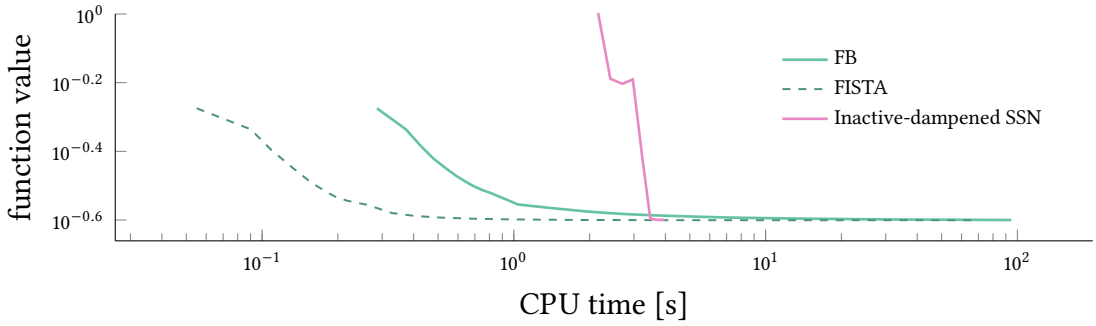


Figure 30.3: Sparse reconstruction algorithm performance: time (in seconds) versus function value.

Theorem 30.10. *For* (30.3), *suppose* $0 \in \partial J(\widehat{x})$ *and that* $\widehat{x}$ *and* $\widehat{p} := -\lambda^{-1}A^*(A\widehat{x}-b) \in \partial\|\cdot\|_1(\widehat{x})$ *satisfy the strict complementarity condition* (30.6). *Let*

$$\mathcal{I} := \{i \in \{1,\dots,M\} \mid \widehat{x}_i = 0\}$$

be the set of inactive indices, and set

$$V = \left\{x \in \mathbb{R}^M \;\middle|\; x_i = 0 \text{ for } i \in \mathcal{I}\right\}$$

Denote by P_V the orthogonal projection onto V. Then S has the Aubin property at 0 for $\widehat{x}$ if $P_V A^* A P_V^*$ is nonsingular on V.

Proof. Since the solution mapping S has the Aubin property at 0 for $\widehat{x}$ if ∂J is metrically regular at $\widehat{x}$ for 0, we use Corollary 27.14 to verify the latter. To do so, we need an expression for $D^*[\partial J]$. For simplicity of notation, we write $g := \|\cdot\|_1$, so that $J(x) = \frac{1}{2}\|Ax-b\|_2^2 + \lambda g(x)$. Then Theorem 4.14 and Example 4.7 give

$$\partial J(x) = A^T(Ax - b) + \lambda \partial g(x) \quad \text{for} \quad \partial g(\widehat{x}) = \prod_{i=1}^{M} \begin{cases} \operatorname{sign} \widehat{x}_i & \widehat{x}_i \neq 0, \\ [-1,1] & \widehat{x}_i = 0. \end{cases}$$

To calculate $D^*[\partial J](\widehat{x}|0)$, we need ∂g to be graphically regular at $\widehat{x}$ for $\widehat{p}$. By Theorem 20.18, this is equivalent to the strict complementarity assumed in (30.6).

Since the first part of $\partial J(x)$ is single-valued and linear, using Theorems 25.14 and 25.15 together with the assumption (30.6), we obtain for any $p^* \in \mathbb{R}^M$ that

$$D^*[\partial J](\widehat{x}|0)(p^*) = A^*Ap^* + \lambda D^*[\partial g](\widehat{x}|\widehat{p})(p^*). \tag{30.16}$$

In Theorem 20.18, for the strictly complementary cases (30.6), we have already calculated that

$$D^*[\partial g](\widehat{x}|\widehat{p})(p^*) = \prod_{i=1}^{M} \begin{cases} \{0\} & \text{if } \widehat{x}_i \neq 0,\ \widehat{p}_i = \operatorname{sign} \widehat{x}_i, \\ \mathbb{R} & \text{if } \widehat{x}_i = 0,\ [p^*]_i = 0,\ |\widehat{p}_i| < 1, \\ \emptyset & \text{otherwise.} \end{cases}$$

From (30.16) we now obtain

$$D^*[\partial J](\widehat{x}|0)(p^*) = \begin{cases} A^*Ap^* + V^\perp, & p^* \in V, \\ \emptyset, & p^* \notin V. \end{cases}$$

Note how λ disappears from the expression, as V and $V^\perp$ are subspaces and thus invariant under multiplication by λ. We then calculate

$$\begin{aligned} |D^*[\partial J](\widehat{x}|0)^{-1}|^+ &= \sup\{\|p^*\| \mid \exists \Delta p^* \in D^*[\partial g](\widehat{x}|0)(p^*),\ \|p^*\| \leq 1\} \\ &= \sup\{\|p^*\| \mid \Delta x \in V,\ z \in V^\perp\ \|A^*Ap^* + z\| \leq 1\} \\ &= \sup\{\|p^*\| \mid \Delta x \in V,\ \|P_V A^* A P_V^* p^*\| \leq 1\}. \end{aligned}$$

Thus Corollary 27.14 shows that ∂J is metrically regular at $\widehat{x}$ for 0. □

We can also prove sensitivity with respect to the regularization parameter under the exact same conditions as in the previous theorem.

Theorem 30.11. *Suppose that the conditions of Theorem 30.10 hold and that $P_V A^* A P_V^*$ is nonsingular on V. Let*

$$Z(\tilde{\lambda}) \coloneqq \{x \in \mathbb{R}^M \mid 0 \in \partial \tilde{J}(x; \tilde{\lambda})\} \quad \text{for} \quad \tilde{J}(x; \tilde{\lambda}) \coloneqq \frac{1}{2}\|Ax - b\|_2^2 + \tilde{\lambda}\|x\|_1.$$

Then Z has the Aubin property at λ for any $x \in Z(\lambda)$.

Proof. In Theorem 28.5, take $g(x) = \|x\|_1$ and $h(x) = \frac{1}{2}\|Ax - b\|_2^2$. If we verify (28.8), i.e.,

$$0 \in A^* A y + \lambda D^*[\partial g](\bar{x}| - \lambda^{-1} A^*(A\bar{x} - b))(y) \implies y = 0,$$

then Theorem 28.5 establishes that Z has the Aubin property at λ. The strict complementarity condition (30.6) implies that either $\bar{x}_i \neq 0$ or $|[\lambda^{-1} A^*(A\bar{x} - b)]_i| < 1$ for all components $i = 1, \dots, M$. Therefore, Theorem 20.18 shows that

$$D^*[\partial g](\bar{x}| - \lambda^{-1} A^*(A\bar{x} - b))(y) = V^\perp \neq \emptyset$$

if and only if $y \in V$. Consequently, (28.8) becomes

$$0 \in A^* A y + V^\perp \text{ and } y \in V \implies y = 0.$$

But this follows from the assumption that $P_V A^* A P_V^*$ is invertible on V. □

Remark 30.12 (regularization theory). A proof of convergence of solutions to the sparse regularization problems (30.3) in the sense of Section 28.3 as $\lambda \searrow 0$ can be found in [Valkonen, 2021b].

31 ℓ^1 FITTING

Nonsmooth norms are not only useful as regularization terms. In (30.2), the use of the sum of squares as a data fitting term was justified by statistical arguments: For Gaussian noise, its minimizer coincides with the *mean* of the signal, which is the *maximum likelihood estimator* under this assumption. However, that connection is lost for non-Gaussian noise, in particular if the data contains *outliers* (rare deviations of much larger magnitude than a normal distribution would predict). A particular such error model is *impulsive noise*, which is characterized by containing *only* outliers. Such errors are relevant in digital signal and image processing, where they can arise from malfunctioning pixels in camera sensors, faulty memory locations in hardware, or transmission in noisy channels. One particular model is *random-valued impulsive noise*, which corresponds to additive errors of the form

$$\eta(x) = \begin{cases} \xi & \text{with probability } r, \\ 0 & \text{with probability } 1 - r, \end{cases}$$

where $r \in [0, 1]$ is the fraction of faulty channels and the normally distributed random variable ξ with mean 0 and variance $\sigma > 0$ is the (independent) noise on each affected channel. A more extreme model is *salt-and-pepper noise*, where the data in each affected channel is replaced by either 0 or 1 (modeling, e.g., pixels in CCD sensors that are either defective or saturated by cosmic noise).

Statistically, a more robust estimator in the presence of outliers is the *median*, which minimizes – instead of the sum of squares – the sum of absolute values; see [Gelman et al., 2013; Huber, 2009]. This leads to replacing in (30.3) the squared ℓ^2 norm by the (nonsquared) ℓ^1 norm. (Another motivation for this is that at least for impulsive noise, the model output should match the data everywhere except for the outliers – i.e., that the residual data mismatch is sparse.) Due to their relevance in signal and image processing, such problems have attracted increasing interest in the last decade; here we only mention [Clason et al., 2010; Kärkkäinen et al., 2005; Yang et al., 2009] as a sample of relevant work. To avoid additional complexity, we here consider again the regression problem from Chapter 30, where we now assume as in [Clason et al., 2010; Kärkkäinen et al., 2005] that the noise is sparse but the solution is smooth.

31.1 PROBLEM DESCRIPTION

We consider for $A \in \mathbb{R}^{N\times M}$ and $b \in \mathbb{R}^N$ as in (30.3) the *ℓ^1-fitting problem*

$$\min_{x\in\mathbb{R}^M} \lambda\|Ax - b\|_1 + \frac{1}{2}\|x\|_2^2, \tag{31.1}$$

where $\lambda > 0$ is a (inverse) regularization parameter related to the noise level. (The benefit of writing the problem in this form instead of using $\alpha := \lambda^{-1}$ as in Chapter 30 will become apparent in the following.) The regularization term is smooth to indicate no sparsity requirements on the reconstructed signal, merely the desire for small values.

31.2 OPTIMALITY CONDITIONS

Using the same approach as in Chapter 30, we obtain optimality conditions for (31.1). We write the problem in the canonical form

$$\min_{x\in\mathbb{R}^M} J(x) \quad \text{where} \quad J(x) := F(Ax) + G(x)$$

by taking

$$F(y) := \lambda\|y - b\|_1 \quad \text{and} \quad G(x) := \frac{1}{2}\|x\|_2^2.$$

We then have the following explicit optimality conditions.

Theorem 31.1. *A vector $\widehat{x} \in \mathbb{R}^M$ is a solution to (31.1) if and only if there exists $\widehat{y} \in \mathbb{R}^N$ such that*

$$-\widehat{x} = A^*\widehat{y} \qquad \text{and} \qquad \widehat{y}_i \in \begin{cases} \{\lambda\} & \text{if } [A\widehat{x} - b]_i > 0, \\ \{-\lambda\} & \text{if } [A\widehat{x} - b]_i < 0, \\ [-\lambda, \lambda] & \text{if } [A\widehat{x} - b]_i = 0. \end{cases} \tag{31.2}$$

Proof. Since F and G are convex and A is linear, also J is convex. Therefore the convex Fermat principle of Theorem 4.2 characterizes the solution of (31.1) as those $\widehat{x}$ satisfying $0 \in \partial J(\widehat{x})$. Since both F and G have full domains and are proper and lower semicontinuous, we may further use the subdifferential sum rule of Theorem 4.14 and the chain rule of Theorem 4.17 to calculate for all $x \in \mathbb{R}^m$ that $\partial J(x) = A^*\partial F(Ax) + \partial G(x)$. Since G is differentiable, using Theorem 4.5 we therefore characterize the solutions as those points $\widehat{x}$ satisfying

$$-\widehat{x} \in A^*\partial F(A\widehat{x}). \tag{31.3}$$

Using Example 4.7 to calculate $\partial F(A\widehat{x})$ componentwise, we obtain (31.2). □

Based on the Fenchel–Rockafellar Theorem 5.11, we may alternatively study optimality conditions for the dual problem

$$\min_{y\in\mathbb{R}^N} Q(y) := F^*(y) + G^*(-A^*y).$$

We know from Lemma 5.4 that $G^*(x) = \frac{1}{2}\|x\|_2^2$. By Example 5.3 (ii) and Lemma 5.7 we also calculate that

$$F^*(y) = \delta_{\lambda\mathbb{B}_\infty}(y) + \langle b, y\rangle. \tag{31.4}$$

Therefore, the dual problem is given by

$$\min_{y\in\mathbb{R}^N} \delta_{\lambda\mathbb{B}_\infty}(y) + \langle b, y\rangle + \frac{1}{2}\|A^*y\|_2^2. \tag{31.5}$$

For this problem, we can also derive explicit optimality conditions.

Theorem 31.2. *A vector $\widehat{y} \in \mathbb{R}^N$ is a solution to (31.5) of (31.1) if and only if there exists a $\widehat{p} \in \mathbb{R}^N$ such that*

$$-AA^*\widehat{y} = \widehat{p} \qquad \text{and} \qquad [\widehat{p} - b]_i \in \begin{cases} [0,\infty) & \text{if } \widehat{y}_i = \lambda, \\ 0 & \text{if } \widehat{y}_i \in (-\lambda,\lambda), \\ (-\infty, 0] & \text{if } \widehat{y}_i = -\lambda, \\ \emptyset & \text{otherwise.} \end{cases} \tag{31.6}$$

Proof. Again, the Fermat principle characterizes the solutions via $0 \in \partial Q(\widehat{y})$. Since G^* has a full domain and both F^* and G^* are proper and lower semicontinuous, we may further use the subdifferential sum rule of Theorem 4.14 and the chain rule of Theorem 4.17 to calculate for all $y \in \mathbb{R}^N$ that $\partial Q(y) = -A\partial G^*(-A^*y) + \partial F^*(x)$. By the differentiability of G^*, again any dual solution $\hat{y}$ is therefore characterized by

$$-AA^*\widehat{y} = A\nabla G^*(-A^*\widehat{y}) \in \partial F^*(\widehat{y}). \tag{31.7}$$

Using Example 4.9 the expression of the subdifferential of the indicator function of an interval, we obtain (31.6). □

We can also characterize the primal and dual solutions through a primal-dual system.

Theorem 31.3. *The solutions $\widehat{x} \in \mathbb{R}^M$ and $\widehat{y} \in \mathbb{R}^N$ to the primal problem (31.1) and the dual problem (31.5) are simultaneously characterized by (31.2) or, equivalently,*

$$-A^*\widehat{y} = \widehat{x} \qquad \text{and} \qquad [A\widehat{x} - b]_i \in \begin{cases} [0,\infty) & \text{if } \widehat{y}_i = \lambda, \\ 0 & \text{if } \widehat{y} \in (-\lambda,\lambda), \\ (-\infty, 0] & \text{if } \widehat{y} = -\lambda, \\ \emptyset & \text{otherwise,} \end{cases} \tag{31.8}$$

Proof. According to Theorem 5.11, the primal and dual solutions are characterized by

$$\widehat{y} \in \partial F(A\widehat{x}) \quad \text{and} \quad -A^*\widehat{y} \in \partial G(\widehat{x}). \tag{31.9}$$

This expands as (31.2) where $\widehat{y}$ is indeed the dual variable. By the Fenchel–Young Lemma 5.8, the conditions (31.9) equivalently be written

$$A\widehat{x} \in \partial F^*(\widehat{y}) \quad \text{and} \quad -A^*\widehat{y} \in \partial G(\widehat{x}).$$

Similarly to the proof of Theorem 31.2, this condition becomes (31.8). □

One may note that the primal-dual condition (31.8) implies the dual condition (31.6) with $\widehat{p} = A\widehat{x}$.

31.3 ALGORITHMS

Once more, the starting point for implementable algorithms is the proximal point reformulation of the optimality conditions, this time for the dual problem.

Lemma 31.4. *A vector $\widehat{y} \in \mathbb{R}^N$ is a solution to (31.5) of (31.1) if and only if*

$$\widehat{y} = \operatorname{proj}_{\lambda\mathbb{B}_\infty}(\widehat{y} - \tau[AA^*\widehat{y} + b]). \tag{31.10}$$

Proof. Recalling Lemma 6.21, we may also rewrite (31.7) in terms of the proximal operator of F^*, for any $\tau > 0$ (we just multiply (31.3) by τ) as

$$\widehat{y} = \operatorname{prox}_{\tau F^*}(\widehat{y} + \tau A\nabla G^*(-A^*\widehat{y})),$$

Using the expression for F^* in (31.4) and the definition of the conjugate, we have for any y that

$$\operatorname{prox}_{\tau F^*}(y) = \operatorname{prox}_{\tau\delta_{\lambda\mathbb{B}_\infty}}(y - \tau b) = \operatorname{proj}_{\lambda\mathbb{B}_\infty}(y - \tau b).$$

Hence we obtain (31.10). □

DUAL FORWARD-BACKWARD SPLITTING

Following Section 8.2, we obtain from (31.10) the forward-backward splitting method

$$y^{k+1} := \operatorname{proj}_{\lambda\mathbb{B}_\infty}(y^k - \tau[AA^*y^k + b]). \tag{31.11}$$

As an immediate consequence of Theorem 9.6, the iterates of (31.11) converge subject to a bound on the step length parameter $\tau > 0$.

Theorem 31.5. *Suppose $\tau\|A\|^2 < 2$. Then for any starting point $y^0 \in \mathbb{R}^N$, the iterates $\{y^k\}_{k\in\mathbb{N}}$ generated by (31.11) converge to a solution $\widehat{y}$ of the dual problem (31.5).*

By Theorem 31.3, the primal and dual solutions $\widehat{x}$ and $\widehat{y}$ to (31.1) satisfy $\widehat{x} = -A^*\widehat{y}$. We can therefore recover a primal approximate solution $x^k = -A^*y^k$ from a dual approximate solution y^k.

Convergence of function values can be obtained in a similar fashion from Corollary 11.5 under the stricter condition $\tau\|A\|^2 \leq 1$. Under this condition, we also obtain from Theorem 12.12 the $O(1/k^2)$ convergence of the inertial variant

$$\begin{cases} y^{k+1} := \operatorname{proj}_{\lambda\mathbb{B}_\infty}(y^k - \tau[AA^*\bar{y}^k + b]). \\ \alpha_{k+1} := \lambda_{k+1}(\lambda_k^{-1} - 1), \\ \bar{y}^{k+1} := (1+\alpha_{k+1})y^{k+1} - \alpha_{k+1}y^k. \end{cases} \tag{31.12}$$

Here the initial inertial parameter is initialized with $\lambda_0 = 1$, while $\bar{y}^0 \in \mathbb{R}^N$ can be chosen freely. Since $\alpha_1 = 0$, the initial iterate y^0 is in fact never used.

PRIMAL-DUAL PROXIMAL SPLITTING

We can also apply the PDPS method (8.20) to (31.1) by taking

$$F(x) = \frac{1}{2}\|x\|_2^2, \qquad G(z) = \frac{1}{2}\|z - b\|_1, \qquad K = A,$$

in the canonical problem (8.12). Using Example 5.3 and Lemmas 5.4 and 5.7 we see that $G^*(y) = \delta_{\lambda\mathbb{B}_\infty}(y) + \langle y, b\rangle$. Consequently, it is not difficult to verify that we then have

$$\operatorname{prox}_{\sigma G^*}(y) = \operatorname{proj}_{\lambda\mathbb{B}_\infty}(y - \sigma b).$$

The projection reduces to a simple componentwise "clamping" of values in the range $[-\lambda, \lambda]$. The PDPS method of (8.20) then becomes

$$\begin{cases} x^{k+1} := \dfrac{1}{1+\tau}(x^k - \tau A^* y^k), \\ \bar{x}^{k+1} := 2x^{k+1} - x^k, \\ y^{k+1} := \operatorname{proj}_{\lambda\mathbb{B}_\infty}(y^k + \sigma(A\bar{x}^{k+1} - b)). \end{cases} \tag{31.13}$$

The method converges subject to a simple step length condition.

Theorem 31.6. *Suppose $\tau\sigma\|A\|^2 < 1$. Then for any starting point $(x^0, y^0) \in \mathbb{R}^M \times \mathbb{R}^N$, the iterates $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ generated by (31.13) converge to solutions $\widehat{x}$ and $\widehat{y}$ of (31.1) and (31.5), respectively.*

Since F is strongly convex with factor $\gamma = 1$, we can also apply the accelerated method of (10.23), updating the step length parameter according to (10.25) in

$$\begin{cases} \omega_k := 1/\sqrt{1+2\tau_k}, \quad \tau_{k+1} := \tau_k \omega_k, \quad \sigma_{k+1} := \sigma_k/\omega_k, \\ x^{k+1} := \dfrac{1}{1+\tau_k}(x^k - \tau_k A^* y^k), \\ \bar{x}^{k+1} := (1+\omega_k)x^{k+1} - \omega_k x^k, \\ y^{k+1} := \operatorname{proj}_{\lambda \mathbb{B}_\infty}(y^k + \sigma_{k+1}(A\bar{x}^{k+1} - b)). \end{cases} \tag{31.14}$$

Theorem 10.8 immediately yields its convergence.

Theorem 31.7. *Suppose $\tau_0 \sigma_0 \|A\|^2 < 1$. Then for any starting point $(x^0, y^0) \in \mathbb{R}^M \times \mathbb{R}^N$, the primal iterates $\{x^k\}_{k\in\mathbb{N}}$ generated by (31.14) converge to a minimizer $\widehat{x}$ of (31.1) at the rate $O(1/k^2)$.*

Convergence of the Lagrangian duality gap can be obtained from Theorem 11.11 or, in the accelerated case, Theorem 11.16, under the same conditions as for iterate convergence.

SEMISMOOTH NEWTON METHOD

Similar to sparse regularization from Chapter 30, we apply a semismooth Newton method to the proximal point reformulation (31.10) by looking for a root $\widehat{y}$ of

$$H(y) := y - \operatorname{proj}_{\lambda \mathbb{B}_\infty}(y - \tau(AA^* y + b))$$

with arbitrary $\tau > 0$. From Example 14.10 (i) and the chain rule Theorem 14.4, a Newton derivative in direction h is given componentwise by

$$\begin{aligned} [D_N H(y)h]_i &= \left[h - \mathbb{1}_{[-\lambda,\lambda]}(y - \tau(AA^* y + b)) \odot (h - \tau AA^* h)\right]_i \\ &= \begin{cases} h_i & \text{if } |y_i - \tau[AA^* y + b]_i| > \lambda, \\ \tau[AA^* h]_i & \text{if } |y_i - \tau[AA^* y + b]_i| \le \lambda. \end{cases} \end{aligned}$$

Here we recall the notation $[x \odot y]_i := x_i y_i$ for the componentwise or Hadamard product on $\mathbb{R}^N$. We can write this concisely as

$$D_N H(y) = \mathbb{1}_{\mathcal{A}(y^k)} + \tau \mathbb{1}_{\mathcal{I}(y^k)} AA^*$$

for the active and inactive sets

$$\mathcal{A}(y^k) := \{i \in \{1, \dots, N\} \mid |y_i - \tau[AA^* y + b]_i| > \lambda\}, \quad \text{and} \tag{31.15a}$$

$$\mathcal{I}(y^k) := \{1, \dots, N\} \setminus \mathcal{A}(y^k). \tag{31.15b}$$

Thus the semismooth Newton algorithm is $y^{k+1} := y^k + s^k$, where we solve for s^k in

$$(\mathbb{1}_{\mathcal{A}(y^k)} + \tau \mathbb{1}_{\mathcal{I}(y^k)} AA^*)s^k = -H(y^k).$$

Proceeding as for (30.12), we deduce that $s_i^k = -H(y^k)_i = [\mathrm{proj}_{\lambda \mathbb{B}_\infty}(y^k - \tau(AA^* y^k + b)) - y^k]_i$ for $i \in \mathcal{A}(y^k)$, hence

$$y_i^{k+1} = [\mathrm{proj}_{\lambda \mathbb{B}_\infty}(y^k - \tau(AA^* y^k + b))]_i \quad \text{for} \quad i \in \mathcal{A}(y^k). \tag{31.16}$$

For $i \in \mathcal{I}(y^k)$, we have $[\mathrm{proj}_{\lambda \mathbb{B}_\infty}(y^k - \tau(AA^* y^k + b))]_i = [y^k - \tau(AA^* y^k + b)]_i$. Hence, by introducing the projection $P_{\mathcal{I}(y^k)}$ to the inactive set and writing

$$s^k = P^*_{\mathcal{I}(y^k)} P_{\mathcal{I}(y^k)} s^k + \mathbb{1}_{\mathcal{A}(y^k)} s^k = P^*_{\mathcal{I}(y^k)} P_{\mathcal{I}(y^k)} s^k - \mathbb{1}_{\mathcal{A}(y^k)} H(y^k)$$

we deduce as after (30.12) that the inactive components $P_{\mathcal{I}(y^k)} s^k$ are characterized by

$$\tau P_{\mathcal{I}(y^k)} AA^* P^*_{\mathcal{I}(y^k)} [P_{\mathcal{I}(y^k)} s^k] = -P_{\mathcal{I}(y^k)} (\mathrm{Id} - \tau AA^* \mathbb{1}_{\mathcal{A}(y^k)}) H(y^k). \tag{31.17}$$

Altogether, therefore, the semismooth Newton method for the dual problem (31.5) iterates

(i) form the active and inactive sets $\mathcal{A}(y^k)$ and $\mathcal{I}(y^k)$ following (31.15);

(ii) update y_i^{k+1} for $i \in \mathcal{A}(y^k)$ by (31.16);

(iii) solve $P_{\mathcal{I}(y^k)} s^k$ from (31.17);

(iv) update $y_i^{k+1} := y_i^k + s_i^k$ for $i \in \mathcal{I}(y^k)$.

From the dual iterate y^k, an approximation of the corresponding primal solution can again be recovered via $x^k := -A^* y^k$.

Completely analogously to the proof of superlinear convergence in Theorem 30.8, Theorem 14.1 establishes the following covergence result.

Theorem 31.8. *Let $\bar{y}$ be a (unique) minimizer of the dual problem (31.5) to (31.1) and $\gamma > 0$. Suppose that $P_{\mathcal{I}(\bar{y})} AA^* P^*_{\mathcal{I}(\bar{y})}$ is positive definite. If y^0 is sufficiently close to $\bar{y}$, then $\{y^{k+1}\}_{k \in \mathbb{N}}$ generated by iterating (i)–(iv) above converge superlinearly to $\bar{y}$.*

NUMERICAL ILLUSTRATION

Again we illustrate the performance of the aforementioned algorithms on a simple numerical example. We take $x \in \mathbb{R}^{1024}$ and $A \in \mathbb{L}(\mathbb{R}^{1024}; \mathbb{R}^{128})$ as convolution with a Gaussian kernel (standard deviation $\sigma = 7$ on the domain $[0, 1024]$) followed by subsampling. To generate the data b, we apply A to the true signal depicted in Figure 30.1, and follow with salt-and-pepper noise of magnitude 1.8. As the inverse regularization parameter, we take

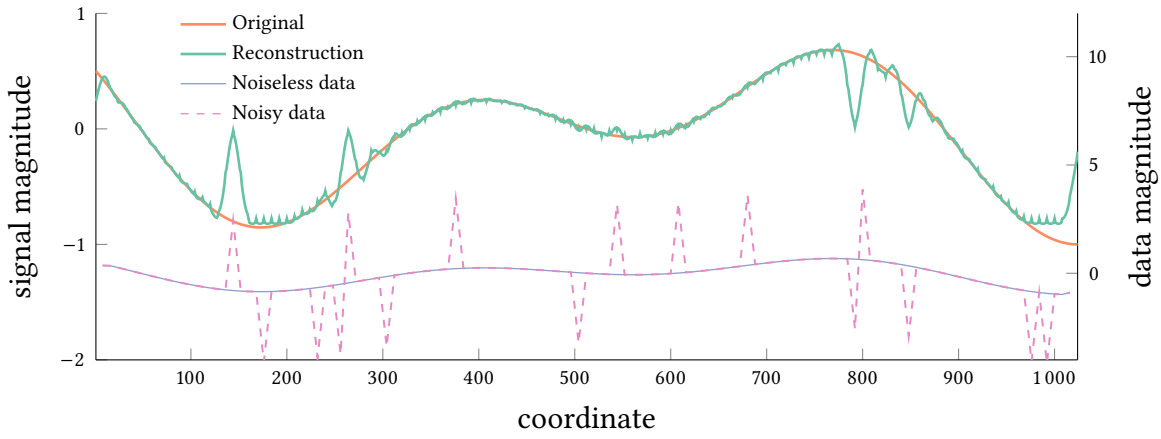


Figure 31.1: ℓ^1 fitting data and result.

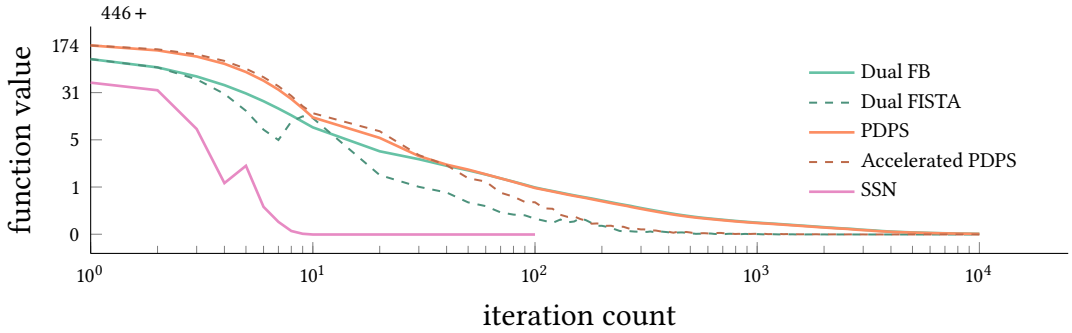


Figure 31.2: ℓ^1 fitting reconstruction algorithm performance: iteration vs. function value.

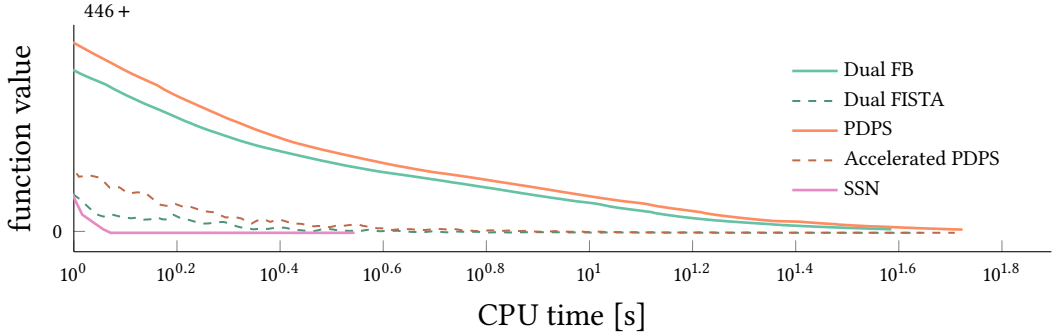


Figure 31.3: ℓ^1 fitting reconstruction algorithm performance: time (in seconds) vs. function value.

$\lambda = 6.5$. For all algorithms, we use the initial iterate $x^0 = 0$. For the forward-backward type methods, we take the step length $\tau = 0.9/L^2$, where L is again an estimate of $\|A\|$. For the PDPS method, we take the step lengths $\tau = 0.5/L$ and $\sigma = 1.9/L$. For the SSN method, we take the proximal parameter $\gamma = 9/L^2$. The precise experimental details can be found in the accompanying code [Clason and Valkonen, 2023].

We illustrate the data and the reconstruction in Figure 31.1 and the convergence behavior

in Figures 31.2 and 31.3. The latter clearly show that acceleration improves performances of the first-order methods, but the superlinearly convergent SSN method requires significantly fewer iterations than any first-order method. In fact, even though each iteration of the former is much more expensive, the total time to reach the objective is still smaller. On the other hand, first-order methods are much faster in reducing the objective value in the beginning and therefore may be the method of choice if high accuracy is not desired.

32 TOTAL VARIATION REGULARIZATION

We now turn to *mathematical image processing*, where the unknown to be reconstructed from data is a digital image. The most basic mathematical image processing task is *denoising*, i.e., removing the noise in an image (for example, a photograph taken in low light conditions), which corresponds to taking the forward operator as the identity. More advanced image processing tasks include *inpainting*, *deblurring*, and *superresolution*. These correspond to filling in missing parts of an image, reducing blur caused by defocussed lenses or motion, and recovering additional detail, and involve more complicated linear forward operators. For an introduction to mathematical image processing, we refer to [Bredies and Lorenz, 2018; Scherzer et al., 2009]. In true *inverse imaging problems*, the given data is not itself an image but related to it via some mathematical model describing the physical measurements; examples are magnetic resonance imaging (MRI), involving the Fourier transform [Nishimura, 1996], or positron emission tomography (PET) and computed X-ray tomography (CT), both involving the Radon transform [Natterer, 2001]. More challenging imaging modalities such as electrical impedance tomography (EIT) and more advanced MRI techniques require the forward operator A to be nonlinear. We do not treat such operators here, but point towards the primal-dual method of Chapter 15 as one possible solution technique. Alternative Gauss–Newton type methods are introduced by [Jauhiainen et al., 2020].

The salient point here is the particular structure of images, which requires an adapted regularization term. The key observation here is that images contain sharp edges (representing jumps in intensity) separating mostly smooth areas. Mathematically, this can be related to requiring sparsity of the *gradient* of the image, rather than the image itself; the corresponding sparse regularization of the gradient is called the *total variation regularization*, which was introduced for denoising by [Rudin et al., 1992] and has become very popular for other (inverse) imaging tasks such as the ones mentioned above.

Treating such problems in an infinite-dimensional function space framework is very challenging and requires the unknown to be considered in the space $\mathrm{BV}(\Omega)$ of *functions of bounded variation* on a domain $\Omega \subset \mathbb{R}^2$, which are characterized by their distributional gradient being a Radon measure $y \in \mathcal{M}(\Omega;\mathbb{R}^2)$. This is a nonreflexive Banach space with a complicated structure; see [Ambrosio et al., 2000; Attouch et al., 2014] for the rich functional analysis and geometric measure theory in this space. As our primary focus

here is on algorithms that require a Hilbert space structure, we will treat this problem in a finite-dimensional discretized setting.

32.1 PROBLEM DESCRIPTION

We consider the problem

$$\min_x \ \frac{1}{2}\|Ax - b\|_2^2 + \alpha\|Dx\|_{1,2}, \tag{32.1}$$

where $x \in \mathbb{R}^M$ for $M = n_1 n_2$ is a vectorization of the two-dimensional image, consisting of an $n_1 \times n_2$ grid of components called *pixels*; $A \in \mathbb{R}^{N\times M}$ for some N is the linear forward operator; and $D \in \mathbb{R}^{2M\times M}$ is a discretization of the image gradient to be specified below. We index $x \in \mathbb{R}^M$ using two coordinates $i \in \{1, \dots, n_1\}$ and $j \in \{1, \dots, n_2\}$, identifying x_{ij} with $x_{\iota(i,j)}$ for a suitable linear index ι, for example $\iota(i, j) = i + n_1(j - 1)$. Likewise we index variables $y \in \mathbb{R}^{2M}$ with $k \in \{1, 2\}$ along with i and j, identifying y_{kij} with $y_{\iota_2(k,i,j)}$ for a suitable linear index ι_2, for example $\iota_2(k, i, j) = k + 2(\iota(i, j) - 1)$. We also write $y_{\cdot ij} := (y_{1ij}, y_{2ij}) \in \mathbb{R}^2$. When necessary for clarity, we insert commas between the indices. As a discretized derivative, we take forward differences with Neumann boundary conditions, which with the above notation corresponds to setting

$$[Du]_{1ij} = \begin{cases} u_{i+1,j} - u_{i,j}, & 1 \le i < n_1,\ 1 \le j \le n_2, \\ 0, & i = n_1,\ 1 \le j \le n_2, \end{cases}$$

$$[Du]_{2ij} = \begin{cases} u_{i,j+1} - u_{i,j}, & 1 \le i \le n_1,\ 1 \le j < n_2, \\ 0, & 1 \le i \le n_2,\ j = n_2. \end{cases}$$

It remains to discuss the vector-sparsity penalty

$$\|y\|_{1,2} := \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \|y_{\cdot ij}\|_2 = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \sqrt{y_{1ij}^2 + y_{2ij}^2} \qquad (y \in \mathbb{R}^{2M}).$$

First, it is straightforward to verify that

$$(\mathbb{R}^{2M}, \|\cdot\|_{1,2})^* = (\mathbb{R}^{2M}, \|\cdot\|_{\infty,2}),$$

where

$$\|y\|_{\infty,2} := \max_{\substack{i=1,\dots,n_1 \\ j=1,\dots,n_2}} \|y_{\cdot ij}\|_2,$$

using that

$$\langle y^*, y\rangle_{1,2} := \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{2} y^*_{kij} y_{kij} \le \|y^*\|_{1,2}\|y\|_{\infty,2}.$$

This allows us to compute various objects by applying the convex analysis of Part II pixel-wise, i.e., separately for each pair of pixel coordinates (i, j). First, by applying Theorem 4.6, we obtain an explicit expression for the subdifferential.

Lemma 32.1. *Let $y, y^* \in \mathbb{R}^{2M}$. Then $y^* \in \partial\|\cdot\|_{1,2}(y)$ if and only if*

$$y^*_{\cdot ij} \in \begin{cases} \left\{\frac{y_{\cdot ij}}{\|y_{\cdot ij}\|_2}\right\} & \text{if } y_{\cdot ij} \neq 0, \\ \mathbb{B}_2 & \text{if } y_{\cdot ij} = 0, \end{cases} \tag{32.2}$$

where $\mathbb{B}_2$ is the Euclidean unit ball in $\mathbb{R}^2$.

By Example 5.3 (ii), the Fenchel conjugate of a norm is given by the indicator functional of the dual unit ball, which in this case is

$$\mathbb{B}_{\infty,2} := \{y \in \mathbb{R}^{2M} \mid \|y\|_{\infty,2} \leq 1\} = \{y \in \mathbb{R}^{2M} \mid y_{\cdot ij} \in \mathbb{B}_2 \text{ for each } i, j\}.$$

By Lemma 5.7 (i), we thus have

$$(\alpha\|\cdot\|_{1,2})^* = \delta_{\alpha\mathbb{B}_{\infty,2}}.$$

A case distinction similar to Example 4.9 then yields the following characterization of the subdifferential.

Lemma 32.2. *Let $y, y^* \in \mathbb{R}^{2M}$ and $\alpha > 0$. Then $y^* \in \partial\delta_{\alpha\mathbb{B}_{\infty,2}}(y)$ if and only if*

$$y^*_{\cdot ij} \in \begin{cases} [0,\infty)\, y_{\cdot ij} & \text{if } \|y_{\cdot ij}\|_2 = \alpha, \\ 0 & \text{if } \|y_{\cdot ij}\|_2 < \alpha, \\ \emptyset & \text{otherwise.} \end{cases} \tag{32.3}$$

Finally, similar to Corollary 6.27 (iii) we can show that the corresponding proximal point mapping for $\gamma > 0$ is given pixelwise by

$$[\operatorname{proj}_{\alpha\mathbb{B}_{\infty,2}}(y)]_{\cdot ij} = \operatorname{proj}_{\alpha\mathbb{B}_2}(y_{\cdot ij}) = y_{\cdot ij} \begin{cases} \frac{\alpha}{\|y_{\cdot ij}\|_2} & \text{if } \|y_{\cdot ij}\|_2 > \alpha, \\ 1 & \text{if } \|y_{\cdot ij}\|_2 \leq \alpha. \end{cases} \tag{32.4}$$

32.2 OPTIMALITY CONDITIONS

Our derivation of optimality conditions for (32.1) follows that for sparse regularization in Chapter 30. Setting

$$F(x) := \frac{1}{2}\|Ax - b\|_2^2 \quad \text{and} \quad G(y) := \alpha\|y\|_{1,2},$$

we can write (32.1) in canonical form as

$$\min_{x \in \mathbb{R}^M} J(x) \quad \text{where} \quad J(x) := F(x) + G(Dx). \tag{32.5}$$

The following result characterizes the solutions of this convex problem.

Theorem 32.3. *Let $\widehat{x} \in \mathbb{R}^M$ be a solution to (32.1). Then there exists a $\widehat{y} \in \mathbb{R}^{2M}$ such that*

$$-A^*(A\widehat{x} - b) = D^*\widehat{y} \qquad \text{and} \qquad \widehat{y}_{\cdot ij} \in \begin{cases} \alpha \left\{ \frac{[D\widehat{x}]_{\cdot ij}}{\|[D\widehat{x}]_{\cdot ij}\|_2} \right\} & \text{if } [D\widehat{x}]_{\cdot ij} \neq 0, \\ \alpha \mathbb{B}_2 & \text{if } [D\widehat{x}]_{\cdot ij} = 0. \end{cases} \tag{32.6}$$

Proof. Since F and G are convex, and D is linear, J is convex as well. Therefore the Fermat principle of Theorem 4.2 characterizes the solution of (32.1) as those $\widehat{x}$ satisfying $0 \in \partial J(\widehat{x})$. Since both F and G have full domains and are proper and lower semicontinuous, we may further use the subdifferential sum rule of Theorem 4.14 and the chain rule of Theorem 4.17 to calculate for all $x \in \mathbb{R}^M$ that $\partial J(x) = \partial F(x) + D^* \partial G(Dx)$. Since F is differentiable, we can use Theorem 4.5 to characterize the solutions as those points $\widehat{x}$ satisfying

$$-\nabla F(\widehat{x}) \in D^* \partial G(D\widehat{x}). \tag{32.7}$$

Together with (32.2), this yields (32.6). □

The expression for $\widehat{y}$ in (32.6) is difficult to work with in practice, in particular for deriving algorithms. With the help of the Fenchel–Rockafellar Theorem 5.11, we may alternatively study optimality conditions for the dual problem

$$\min_{y \in \mathbb{R}^{2M}} Q(y) \coloneqq F^*(-D^* y) + G^*(y), \tag{32.8}$$

where $G^* = \delta_{\alpha \mathbb{B}_{\infty,2}}$. If $A = \mathrm{Id}$, we also obtain a simple expression for F^*, which yields the following result.

Theorem 32.4. *For $A = \mathrm{Id}$, the solutions $\widehat{y} \in \mathbb{R}^{2M}$ to the dual problem (32.8) of (32.1) are characterized by*

$$-D(D^*\widehat{y} - b) = \widehat{p} \qquad \text{and} \qquad \widehat{p}_{\cdot ij} \in \begin{cases} [0, \infty) \widehat{y}_{\cdot ij} & \text{if } \|\widehat{y}_{\cdot ij}\|_2 = \alpha, \\ \{0\} & \text{if } \|\widehat{y}_{\cdot ij}\|_2 < \alpha, \\ \emptyset & \text{otherwise.} \end{cases} \tag{32.9}$$

Proof. Again we can apply the Fermat principle. We calculate using Lemma 5.4 and 5.7 (ii) for $K = \mathrm{Id}$ that $F^*(y) = \frac{1}{2}\|y\|_2^2 + \langle b, y \rangle$. Since F^* has a full domain and both G^* and F^* are proper and lower semicontinuous, we may further use the subdifferential sum rule of Theorem 4.14 and the chain rule of Theorem 4.17 to calculate for all $y \in \mathbb{R}^{2M}$ that

$$\partial Q(y) = -D \partial F^*(-D^* y) + \partial G^*(y).$$

By the differentiability of F^*, the dual solutions $\widehat{y}$ are therefore characterized by

$$D \nabla F^*(-D^* \widehat{y}) \in \partial G^*(\widehat{y}). \tag{32.10}$$

Together with (32.3), this yields (32.9). □

The Fenchel–Rockafellar Theorem 5.11 also gives a primal-dual characterization of optimality. In contrast to the primal result Theorem 32.3 and the dual result Theorem 32.4, it has simple expressions for all variables even for $A \neq \mathrm{Id}$.

Theorem 32.5. *The solutions $\widehat{x} \in \mathbb{R}^M$ and $\widehat{y} \in \mathbb{R}^{2M}$ to the primal problem (32.1) and the dual problem (32.8) are simultaneously characterized by (32.6) or, equivalently,*

$$-D^*\widehat{y} = A^*(A\widehat{x} - b) \qquad \text{and} \qquad [D\widehat{x}]_{\cdot ij} \in \begin{cases} [0,\infty)\widehat{y}_{\cdot ij} & \text{if } \|\widehat{y}_{\cdot ij}\|_2 = \alpha, \\ \{0\} & \text{if } \|\widehat{y}_{\cdot ij}\|_2 < \alpha, \\ \emptyset & \text{otherwise.} \end{cases} \tag{32.11}$$

Proof. According to Theorem 5.11, the primal and dual solutions are characterized by

$$\widehat{y} \in \partial G(D\widehat{x}) \quad \text{and} \quad -D^*\widehat{y} = \nabla F(\widehat{x}). \tag{32.12}$$

This is simply (32.6), where $\widehat{y}$ is indeed the dual variable. By the Fenchel–Young Lemma 5.8, the conditions (32.12) can equivalently be written

$$D\widehat{x} \in \partial G^*(\widehat{y}) \quad \text{and} \quad -D^*\widehat{y} = \nabla F(\widehat{x}).$$

Together with (32.3) for an expression of ∂G^*, this yields (32.11). □

32.3 ALGORITHMS

Following the approach established in the previous chapters, we now derive some algorithms for (32.1) based on either the dual optimality conditions (32.9) or the primal-dual optimality conditions (32.11). We start with the former and the corresponding forward-backward type methods. We then move onto primal-dual splitting methods. As the discretized gradient D has a nontrivial kernel, semismooth Newton methods cannot be applied directly without dampening as in Chapter 30, which would negate the performance advantage over splitting methods. We will therefore focus here on splitting methods, but refer to [Hintermüller and Stadler, 2006] for a modified semismooth Newton method that retains superlinear convergence.

DUAL FORWARD-BACKWARD SPLITTING FOR DENOISING

For $A = \mathrm{Id}$, we can directly apply the forward-backward splitting method (8.6) to the dual problem (32.8) by rewriting the optimality condition (32.10) using Lemma 6.21 for any $\tau > 0$ as

$$0 = \operatorname{proj}_{\alpha\mathbb{B}_{\infty,2}}(\widehat{y} - \tau D(D^*\widehat{y} - b)).$$

We recall from the proof of Theorem 32.4 with $A = \mathrm{Id}$ that $F^*(y) = \frac{1}{2}\|y\|_2^2 + \langle b, y\rangle$. Therefore, we obtain the iteration

$$y^{k+1} := \mathrm{prox}_{\tau G^*}(y^k + \tau D^* \nabla F^*(-Dy^k)) = \mathrm{proj}_{\alpha \mathbb{B}_{\infty,2}}(y^k - \tau D(D^* y^k - b)), \tag{32.13}$$

where the projection operator is given by (32.4).

By (32.12), the primal and dual solutions $\widehat{x}$ and $\widehat{y}$ satisfy $-D^*\widehat{y} \in \nabla F(\widehat{x}) = \{\widehat{x} - b\}$, which allows us to recover a primal solution from a dual solution $\widehat{y}$ via $\widehat{x} = b - D^*\widehat{y}$.

Again, the method converges subject to a simple step length bound.

Theorem 32.6. *Suppose $\tau\|D\|^2 < 2$. Then for any starting point $y^0 \in \mathbb{R}^{2M}$, the iterates $\{y^k\}_{k\in\mathbb{N}}$ generated by (32.13) converge to a solution $\widehat{y}$ of the dual problem (32.8).*

Proof. The Lipschitz factor of $\nabla[F^* \circ D]$ is $\|D\|^2$, and hence the claim follows from Theorem 9.6. □

Convergence of function values for the dual objective (31.5) can be obtained in a similar fashion from Corollary 11.5 under the stricter condition $\tau\|D\|^2 \leq 1$.

PRIMAL-DUAL PROXIMAL SPLITTING FOR UNITARY-SIMPLE FORWARD OPERATORS

Dual forward-backward splitting requires that we are able to compute ∇F^*, which can be difficult and numerically expensive for general $A \neq \mathrm{Id}$; compare Lemma 5.7 (iii). Furthermore, F^* may not even be a smooth function when A is not invertible. Similarly, the primal-dual proximal splitting (8.20) for (32.5) is given by

$$\begin{cases} x^{k+1} := \mathrm{prox}_{\tau F}(x^k - \tau D^* y^k), \\ \bar{x}^{k+1} = 2x^{k+1} - x^k, \\ y^{k+1} := \mathrm{proj}_{\alpha \mathbb{B}_{\infty,2}}(y^k + \sigma D \bar{x}^{k+1}), \end{cases} \tag{32.14}$$

which still requires computing the proximal mapping of F, which can in general be difficult.

However, suppose that $A = SU$ for an unitary operator U and S such that $\tau S^*S + \mathrm{Id}$ has a simple inverse. For example, U can be the Fourier transform and S can be a sub-sampling operator, in which case $\tau S^*S + \mathrm{Id}$ is diagonal; such type of problems appear in magnetic resonance imaging. In this case, we can write $x = \mathrm{prox}_{\tau F}(z)$ as

$$0 = \tau U^* S^*(SUx - b) + x - z.$$

Multiplying by U, yields

$$\tau S^* b + Uz = (\tau S^* S + \mathrm{Id})Ux.$$

By assumption, we can solve this for

$$x = U^*(\tau S^* S + \mathrm{Id})^{-1}(\tau S^* b + Uz).$$

This shows that

$$\mathrm{prox}_{\tau F}(z) = U^*(\tau S^* S + \mathrm{Id})^{-1}(S^* b + Uz).$$

In this case, (32.14) is practical to implement. In particular, for $U = S = \mathrm{Id}$, i.e., for image denoising, we have

$$\mathrm{prox}_{\tau F}(z) = \frac{1}{1+\tau}(b + z). \tag{32.15}$$

From Corollary 9.14, we directly obtain the following convergence result which even holds for general A.

Theorem 32.7. *Suppose $\tau\sigma\|D\|^2 < 1$. Then for any starting point $(x^0, y^0) \in \mathbb{R}^M \times \mathbb{R}^{2M}$, the iterates $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ generated by (32.14) converge to solutions $\widehat{x}$ and $\widehat{y}$ of (32.1) and (32.8).*

If F is γ-strongly convex (in particular if $U = S = \mathrm{Id}$, where $\gamma = 1$), we can apply the accelerated variant (10.23) to obtain the iteration

$$\begin{cases} \omega_k := 1/\sqrt{1 + 2\gamma\tau_k}, \quad \tau_{k+1} := \tau_k \omega_k, \quad \sigma_{k+1} := \sigma_k/\omega_k, \\ x^{k+1} := \mathrm{prox}_{\tau_k F}(x^k - \tau_k D^* y^k), \\ \bar{x}^{k+1} = (1 + \omega_k)x^{k+1} - \omega_k x^k, \\ y^{k+1} := \mathrm{proj}_{\alpha\mathbb{B}_{\infty,2}}(y^k + \sigma_{k+1} D\bar{x}^{k+1}). \end{cases} \tag{32.16}$$

From Theorem 10.8, we then obtain convergence at the faster rate $O(1/k^2)$.

Theorem 32.8. *Suppose $\tau_0\sigma_0\|D\|^2 < 1$ and that F is γ-strongly convex. Then for any starting point $(x^0, y^0) \in \mathbb{R}^M \times \mathbb{R}^{2M}$, the primal iterates $\{x^k\}_{k\in\mathbb{N}}$ generated by (32.16) converge to a minimizer $\widehat{x}$ of (32.1) at the rate $O(1/k^2)$.*

Convergence of the Lagrangian duality gap can be obtained from Theorem 11.11 or in the strongly convex case from Theorem 11.16.

PRIMAL-DUAL PROXIMAL SPLITTING FOR GENERAL FORWARD OPERATORS

If A is a more complex operator, $\mathrm{prox}_{\frac{\lambda}{2}\|A\cdot - b\|_2^2}$ in general cannot be computed efficiently. To overcome this, we will split the problem in two different ways. First, as we observed in Section 8.5, we can equivalently write (32.1) as

$$\min_{x\in\mathbb{R}^N} \tilde{G}(x) + \tilde{F}(Kx) \tag{32.17}$$

for

$$\tilde{G} \equiv 0, \quad \tilde{F}(y,z) := \alpha\|y\|_1 + \frac{1}{2}\|z-b\|_2^2, \quad \text{and} \quad Kx := (Dx, Ax).$$

We write for brevity $F_0(z) := \frac{1}{2}\|z-b\|_2^2$. By Lemma 6.24 (i) and Example 6.26 – or from (32.15) – for any $\gamma > 0$, we have

$$\operatorname{prox}_{\gamma F_0}(z) = \operatorname{prox}_{\gamma\frac{1}{2}\|\cdot\|_2^2}(z-b) + b = \frac{1}{1+\gamma}(z-b) + b = \frac{1}{1+\gamma}(z+\gamma b)$$

Hence by Lemma 6.24 (ii),

$$\begin{aligned}\operatorname{prox}_{\sigma F_0^*}(z) &= z - \sigma \operatorname{prox}_{\sigma^{-1}\tilde{F}}(\sigma^{-1}z)\\ &= z - \sigma\frac{1}{1+\sigma^{-1}}(\sigma^{-1}z + \sigma^{-1}b)\\ &= \frac{1}{1+\sigma}(z-\sigma b).\end{aligned}$$

By Lemma 6.24 (iii) we thus obtain

$$\operatorname{prox}_{\sigma F_0^*}(y,z) = (\operatorname{proj}_{\alpha\mathbb{B}_{\infty,2}}(y), \operatorname{prox}_{\sigma F_0^*}(z)) = (\operatorname{proj}_{\alpha\mathbb{B}_{\infty,2}}(y), \tfrac{1}{1+\sigma}(z-\sigma b)).$$

Therefore the primal-dual proximal splitting method (8.20) for (32.17) is given by

$$\begin{cases} x^{k+1} := x^k - \tau[D^* y^k + A^* z^k], \\ \bar{x}^{k+1} = 2x^{k+1} - x^k, \\ y^{k+1} := \operatorname{proj}_{\alpha\mathbb{B}_{\infty,2}}(y^k + \sigma D\bar{x}^{k+1}), \\ z^{k+1} := \dfrac{1}{1+\sigma}(z^k + \sigma[A\bar{x}^{k+1} - b]). \end{cases} \tag{32.18}$$

As before, we can apply the general convergence result from Corollary 9.14 to show that the iterates converge to a solution of the problem (32.1).

Theorem 32.9. *Suppose $\tau\sigma(\|D\|^2 + \|A\|^2) < 1$. For any starting point $(x^0, y^0, z^0) \in \mathbb{R}^{M+2M+N}$, let the iterates $\{(x^k, y^k, z^k)\}_{k\in\mathbb{N}}$ be generated by (32.18). Then the primal iterates $\{x^k\}_{k\in\mathbb{N}}$ converge to a minimizer of (32.1).*

The convergence of a Lagrangian duality gap corresponding to the formulation (32.17) can be obtained from Theorem 11.11.

PRIMAL-DUAL PROXIMAL SPLITTING WITH A FORWARD STEP

The dualization trick of the expanded PDPS method does not require the data term F to be differentiable; we could have derived (32.18) for an $F(x) = F_0(Ax)$ for an arbitrary convex,

possibly nonsmooth F_0. It does, however, require introducing the additional variable z, which may come at the cost of performance. This can be avoided for smooth F by using the variant of the PDPS method with a forward step introduced in (9.29). To apply it, we write (32.1) as

$$\min_{x\in X} F_0(x) + E(x) + G(Kx)$$

for

$$F_0 \equiv 0, \quad E(x) = \frac{1}{2}\|Ax - b\|_2^2, \quad G(x) = \|\cdot\|_{2,1}, \quad \text{and} \quad K = D.$$

Thus G and K are as in (32.14); however, the primal update becomes

$$x^{k+1} := \operatorname{prox}_{\tau F_0}(x^k - \tau[\nabla E(x^k) + D^* y^k])$$

We thus obtain from (9.29) the algorithm

$$\begin{cases} x^{k+1} := x^k - \tau[A^*(Ax^k - b) + D^* y^k], \\ \bar{x}^{k+1} := 2x^{k+1} - x^k, \\ y^{k+1} := \operatorname{proj}_{\alpha\mathbb{B}_{\infty,2}}(y^k + \sigma D\bar{x}^{k+1}). \end{cases} \tag{32.19}$$

We have the following convergence result.

Theorem 32.10. *Suppose* $1 > \|D\|^2\tau\sigma + \frac{\tau}{2}\|A\|^2$. *Then for any starting point* $(x^0, y^0) \in \mathbb{R}^{M+2M}$, *the iterates* $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ *generated by* (32.19) *converge a solution* $(\widehat{x}, \widehat{y})$ *of the primal-dual optimality conditions* (32.11). *In particular, the primal iterates* $\{x^k\}_{k\in\mathbb{N}}$ *converge to a minimizer of* (32.1).

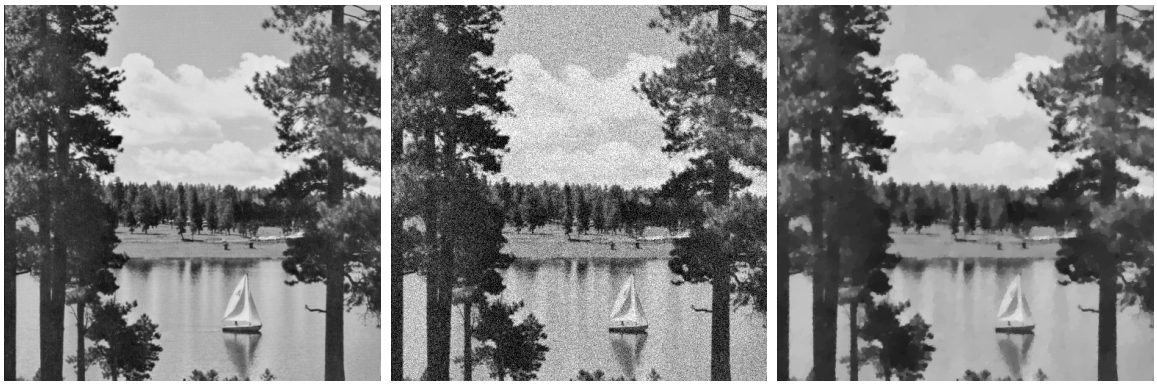
Proof. Since ∇E is Lipschitz with constant $L = \|A\|^2$, the claim is a direct consequence of Corollary 9.20. □

Again, convergence of the Lagrangian duality gap can be obtained from Theorem 11.11. We can also apply acceleration similarly to (32.16), for which convergence rates can be obtained from Theorems 10.8 and 11.16.

PRIMAL-DUAL EXPLICIT SPLITTING

Just like the PDPS method with a forward step, the PDES method of (8.23) avoids the need to introduce an additional variable. To apply the latter, we write the problem (32.1) in the form $\min_x F(x) + G(Kx)$ for $F(x) = \frac{1}{2}\|Ax - b\|_2^2$ and $G(y) = \alpha\|y\|_{1,2}$. However, since the convergence result from Corollary 9.18 has the restriction $\|K\| < 1$, we rescale by taking $G = \alpha\lambda\|\cdot\|_{1,2}$ and $K = \lambda^{-1}D$ for some $\lambda > \|D\|$. Then the PDES method (8.23) becomes

$$\begin{cases} y^{k+1} = \operatorname{proj}_{\lambda\alpha\mathbb{B}_{\infty,2}}((\mathrm{Id} - \lambda^{-2}DD^*)y^k + K(x^k - A^*(Ax^k - b))), \\ x^{k+1} = x^k - A^*(Ax^k - b) - \lambda^{-1}D^* y^{k+1}. \end{cases} \tag{32.20}$$



(a) original (b) noisy (c) reconstruction

Figure 32.1: TV denoising data and result.

We have the following convergence result.

Corollary 32.11. *For any initial iterate* $(x^0, y^0) \in \mathbb{R}^{M\times 2M}$*, the sequence* $\{x^k, \lambda^{-1}y^k\}_{k\in\mathbb{N}}$ *constructed by* (32.20) *converges to a solution of the primal-dual optimality conditions* (32.11)*. In particular, the primal iterates* $\{x^k\}_{k\in\mathbb{N}}$ *converge to a minimizer of* (32.1)*.*

Proof. Since ∇F is Lipschitz with constant $L = \|A\|^2$ and $\|K\| < 1$, Corollary 9.18 immediately yields the convergence of $\{x^k, \lambda^{-1}y^k\}_{k\in\mathbb{N}}$ to some $(\bar{x}, \bar{y})$ satisfying $-K^*\bar{y} = \nabla F(\bar{x})$ and $K\bar{x} \in \partial G^*(\bar{y})$, i.e., $\bar{y} \in \partial G(K\bar{x})$. Since $\|\cdot\|_{2,1}$ is positively homogeneous,

$$\partial G(K\bar{x}) = \alpha\lambda\partial\|\cdot\|_{2,1}(\lambda^{-1}D\bar{x}) = \alpha\partial\|\cdot\|_{2,1}(D\bar{x}).$$

Inserting the definition of $K = \lambda^{-1}D$ and dividing by $\lambda > 0$, respectively, we thus obtain that as $-D^*(\lambda^{-1}\bar{y}) = A^*(A\bar{x} - b)$ and $(\lambda^{-1}\bar{y}) \in \alpha\partial\|\cdot\|_{2,1}(D\bar{x})$. Hence $(\widehat{x}, \widehat{y}) := (\bar{x}, \lambda^{-1}\bar{y})$ satisfies (32.11). □

Convergence of the Lagrangian duality gap can be obtained from Theorem 11.10.

NUMERICAL ILLUSTRATION

We illustrate the performance of the various variants of the forward-backward splitting, PDPS, and PDES methods on total variation denoising and superresolution.

We start with denoising. We include in our experiments the dual forward-backward splitting (32.13), the PDPS method (32.14), the forward PDPS method (32.19), and their accelerated variants. We use as b the noisy image shown in Figure 32.1b, which was obtained from the original ("ground-truth") image in Figure 32.1a by applying normally-distributed noise with mean 0 and standard deviation 0.1. As the regularization parameter, we take $\alpha = 0.1$;

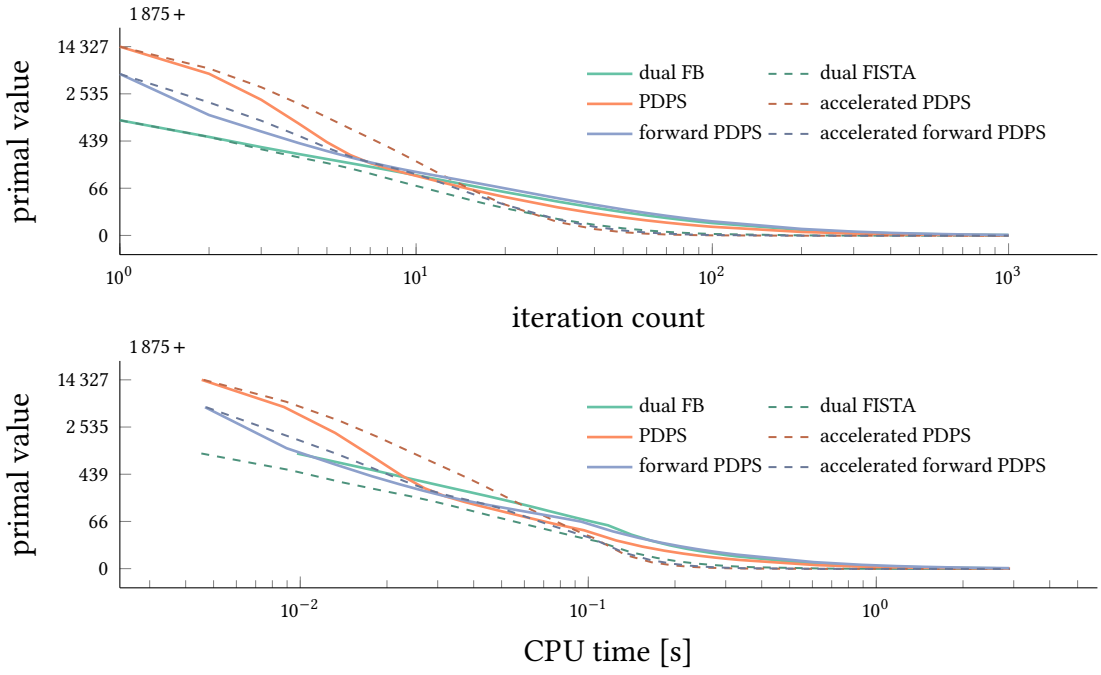


Figure 32.2: TV denoising algorithm performance: primal function value.

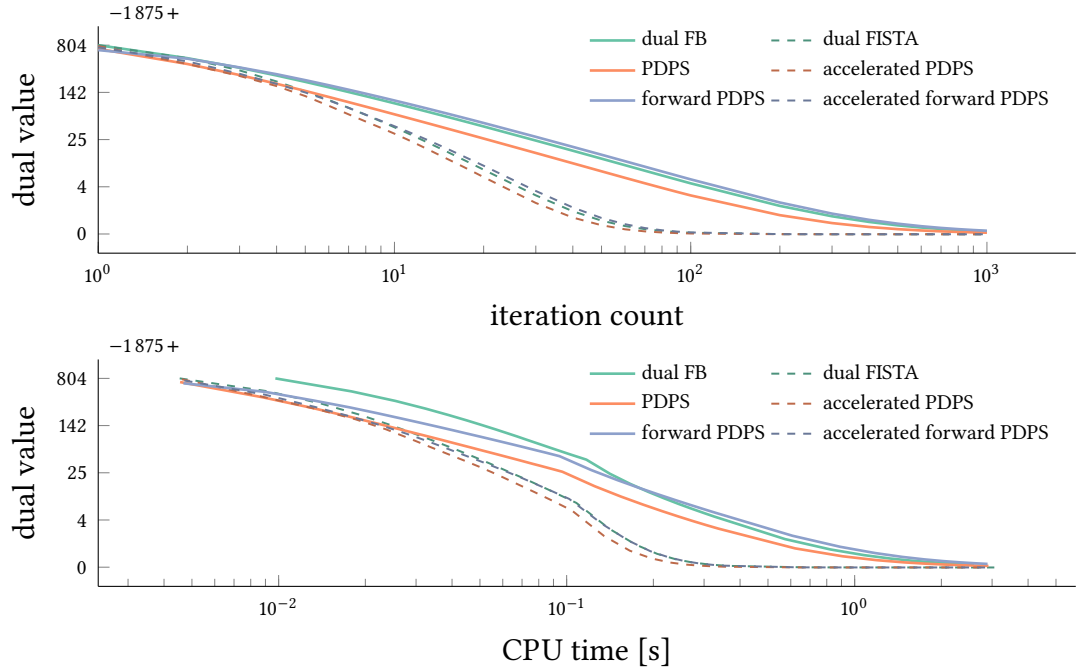


Figure 32.3: TV denoising algorithm performance: dual function value.

the corresponding denoised image is shown in Figure 32.1c. For forward-backward splitting and its inertial variant, we take $\tau = 0.99/M^2$, where M is an estimate of $\|D\|$. For the basic PDPS method and its accelerated variant we take $\tau = 1.99/M$ and $\sigma = 0.5/M$ to satisfy $\tau\sigma M^2 < 1$. For the forward PDPS method and its accelerated variant we take $\tau = 0.35 \cdot 2/L$ and $\sigma = 0.95(1 - \tau L/2)/(\tau M^2)$ to satisfy (9.30), where $L = 1$ is the Lipschitz factor of ∇F. Further experimental details can be found in the accompanying code [Clason and Valkonen, 2023].

We plot the convergence behavior in Figure 32.2 with respect to the primal functional value (32.1). For the primal-dual methods, we use the iterates x^k to directly calculate the primal function values. For the forward-backward methods, which do not directly generate primal variables, we use the first part of the optimality conditions (32.11) (with $A = \mathrm{Id}$) to generate x^k from y^k. Although initially the accelerated variants seem to be slower, they eventually outperform the unaccelerated variants, in line with their better *asymptotic* convergence rates. The same phenomenon can be observed in relation to the different base algorithms: Asymptotically, all algorithms converge at the same rate even though in the beginning, the simpler dual forward-backward splitting outperforms the forward PDPS method which outperforms the PDPS method.

The picture is clearer when considering convergence of the dual function values, which can be directly calculated from all iterates, and for which Theorem 32.6 ensures convergence for dual forward-backward splitting. Note that since all algorithms involve a dual projection step, the dual iterates are feasible, so the dual functional reduces to the strongly convex $F^*(y) = \frac{1}{2}\|y\|_2^2 + \langle b, y\rangle$. Here, Figure 32.3 shows the expected behavior of the algorithms, with the PDPS method outperforming the dual forward-backward splitting method and the forward PDPS method (albeit at the same asymptotic rate), and the accelerated variants clearly outperforming the base algorithms (at a higher asymptotic rate).

For the superresolution demonstration, we consider the forward PDPS method (32.19), the expanded PDPS method (32.18), and the PDES method (32.20). In this experiment, the operator $A \in \mathbb{L}(\mathbb{R}^{512^2}; \mathbb{R}^{64^2})$ performs convolution with a Gaussian kernel (standard deviation $\sigma = 5$ on the domain $\Omega = [0, 512]^2$) followed by subsampling by factor of 8. We illustrate the data and the reconstruction in Figure 32.4. The low-resolution data is obtained from the original image by applying A and adding normally-distributed noise of mean 0 and standard deviation 0.001. As the regularization parameter, we take $\alpha = 0.0001$. For the forward PDPS method, we take $\tau = 0.95 \cdot 2/L$ and $\sigma = 0.95(1 - \tau L/2)/(\tau M^2)$ to satisfy (9.30), where L is an upper estimate of the Lipschitz factor of ∇F, i.e., of $\|A\|^2$. For the expanded PDPS method, we take $\tau = 1.9/\sqrt{M^2 + L}$ and $\sigma = 0.5/\sqrt{M^2 + L}$ to satisfy $\tau\sigma(\|D\|^2 + \|A\|^2) < 1$ via $\tau\sigma(M^2 + L) < 1$. The PDES method has no step length parameters.

We illustrate the convergence behavior in Figure 32.5. As we can see, the expanded variant of the PDPS method is slower than the other two algorithms that do not introduce additional variables. Moreover, the accelerated algorithms eventually outperform all the unaccelerated variants. The PDPS method with forward step is somewhat faster than the PDES method.

(a) original (b) low-resolution data (c) reconstruction

Figure 32.4: TV superresolution data and result.

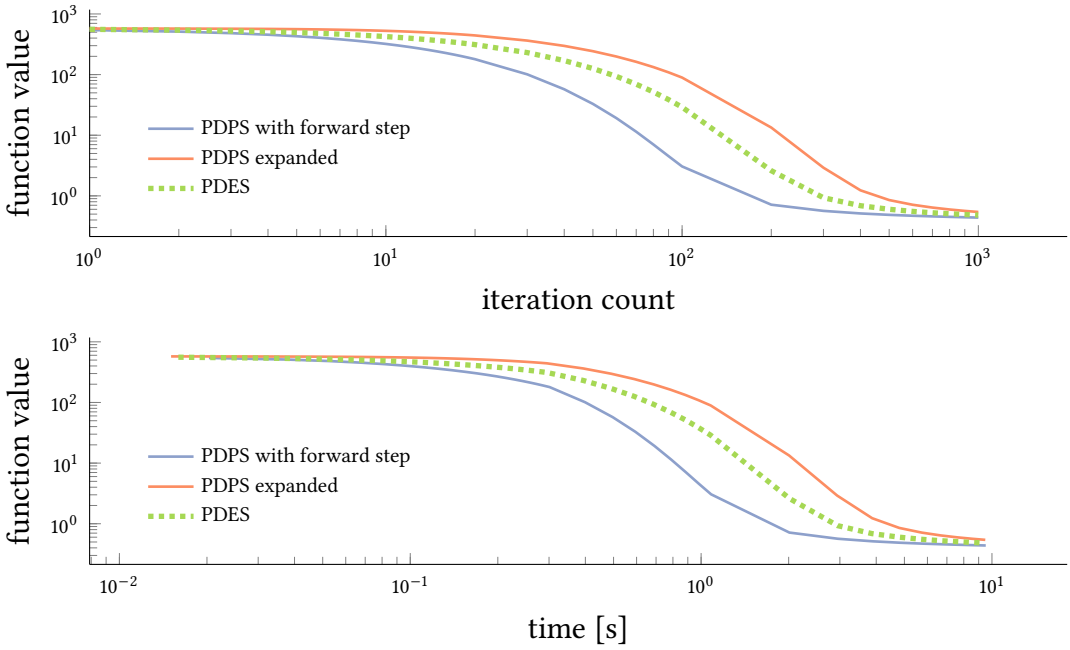


Figure 32.5: TV superresolution algorithm performance.

33 OPTIMAL CONTROL WITH CONSTRAINTS

We now illustrate the applications of the theory of Parts II and III in infinite-dimensional spaces, in particular function spaces. Specifically, we consider *optimal control problems*, where the solution of a (partial) differential equation – the *state* – is sought to be brought as close as possible to a desired state by adjusting a relevant *control*. Typically this control is the right-hand side, boundary conditions, or coefficients of the differential equation. Optimal control problems occur in a wide variety of applications such as autonomous vehicles, process engineering, and optimal design; they are also closely related to inverse problems for partial differential equations. Typically, this involves minimizing a weighted sum of a *tracking term* involving the state and a *control cost* involving the control; these are linked through the differential equation as an equality constraint, and hence this is also known as *PDE-constrained optimization*. Using the implicit function theorem, one can use this constraint to define a *control-to-state mapping*; much of optimal control theory is concerned with analyzing the properties (in particular regarding differentiability) of this mapping, especially for (systems of) time-dependent and/or nonlinear equations or controls appearing as the coefficients. On these and other issues, we refer the reader to the seminal monograph [Lions, 1971], to the standard textbook [Tröltzsch, 2010], as well as to [De los Reyes, 2015; Hinze et al., 2009] in particular regarding applications and numerical methods.

Here we focus on dealing with optimal control problems where either the tracking term or the control costs are nonsmooth, which allows imposing additional structure on the optimal state or control. Correspondingly, such problems have received increasing attention in recent years. To avoid unnecessary technical difficulties, we restrict ourselves to the simplest possible partial differential equation: the Poisson equation with homogeneous boundary conditions and the control appearing as a right-hand side. We briefly introduce the required notation and refer to, e.g., [Tröltzsch, 2010] for details and proofs of the claimed properties. Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary. We then introduce for $k \in \mathbb{N}$ and $1 < p < \infty$ the *Sobolev space*

$$W^{k,p}(\Omega) := \left\{ v \in L^p(\Omega) \;\middle|\; D^\alpha v \in L^p(\Omega) \text{ for all } |\alpha| \leq k \right\},$$

where $D^\alpha v$ is the *weak derivative* of v of order $|\alpha|$. These are Banach spaces with the natural norm; for $p = 2$, $H^k(\Omega) := W^{k,p}(\Omega)$ is a Hilbert space. Under the assumptions on the

domain Ω, we have the continuous embeddings

$$W^{k,p}(\Omega) \hookrightarrow L^q(\Omega) \qquad \text{for } 1 \le q \le \frac{dp}{d-kp} \quad (:= \infty \text{ if } kp \ge d),$$
$$W^{k,p}(\Omega) \hookrightarrow C(\overline{\Omega}) \qquad \text{for } kp > d;$$

see, e.g., [Tröltzsch, 2010, Theorem 7.1]. Furthermore, the embedding $W^{k,p}(\Omega) \hookrightarrow L^p(\Omega)$ is compact for every $k \in \mathbb{N}$ and $1 < p < \infty$; see, e.g., [Tröltzsch, 2010, Theorem 7.4]. In particular, weakly convergent sequences in $W^{k,p}(\Omega)$ for $k \ge 2$ converge strongly in $L^p(\Omega)$. Finally, we denote by $W_0^{k,p}(\Omega)$ the closure of $C_0^\infty(\overline{\Omega})$ with respect to the $W^{k,p}$-norm, whose elements have vanishing trace on the boundary of Ω.

We now consider for given $u \in L^2(\Omega)$ the *weak formulation* of the Poisson equation $-\Delta y = u$ with homogeneous boundary condition, i.e., we look for $y \in H_0^1(\Omega)$ satisfying

$$\int_\Omega \nabla y(x) \cdot \nabla v(x)\, dx = \int_\Omega u(x) v(x)\, dx \qquad \text{for all } v \in H_0^1(\Omega). \tag{33.1}$$

Under the assumptions on Ω, this equation admits a unique solution $y \in H_0^1(\Omega)$ which depends continuously on u; see, e.g., [Tröltzsch, 2010, Theorem 2.4]. This allows defining a linear bounded control-to-state mapping $S : L^2(\Omega) \to L^2(\Omega)$ (which is even compact since the range $\operatorname{ran} S \subset H_0^1(\Omega)$ embeds compactly into $L^p(\Omega)$ for any $1 < p < \infty$). If $d \le 3$ and $\Omega \subset \mathbb{R}^d$ is convex, we even have $y \in H^2(\Omega) \hookrightarrow C(\overline{\Omega})$; see [Grisvard, 2011, Theorem 3.2.1.2].

We will also need the adjoint $S^* : L^2(\Omega) \to L^2(\Omega)$ of S. Using either the implicit function theorem or formal Lagrange multiplier calculus, we can characterize $p := S^* h \in L^2(\Omega)$ for given $h \in L^2(\Omega)$ as the unique solution to the *adjoint equation*

$$\int_\Omega \nabla w(x) \cdot \nabla p(x)\, dx = \int_\Omega w(x) h(x)\, dx \qquad \text{for all } w \in H_0^1(\Omega); \tag{33.2}$$

see, e.g., [Tröltzsch, 2010, Lemma 2.24, Chapter 2.10] or [Hinze et al., 2009, Chapter 1.6]. This implies that $\operatorname{ran} S^* \subset H_0^1(\Omega) \hookrightarrow L^p(\Omega)$ for any $1 < p < \infty$ as well.

33.1 CONTROL CONSTRAINTS

We start with the simplest nonsmooth optimal control problems: quadratic control problems with pointwise constraints on the control or state. Although these problems can be treated by well-known standard methods of constrained smooth optimization (cf., e.g., [Tröltzsch, 2010, Chapters 2 and 6.2], respectively), they serve well to illustrate the application of the abstract results of Part II.

PROBLEM DESCRIPTION

Let $y^d \in L^2(\Omega)$ be a desired state, $\alpha > 0$, and $a > b$ be given. We then consider the "mother problem"

$$\min_{u \in L^2(\Omega), y \in H_0^1(\Omega)} \frac{1}{2}\|y - y^d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2}\|u\|_{L^2(\Omega}^2$$
$$\text{subject to (33.1)} \qquad \text{and} \qquad a \le u(x) \le b \quad \text{for almost every } x \in \Omega.$$

Introducing the *admissible set*

$$U_{\text{ad}} := \left\{ u \in L^2(\Omega) \,\middle|\, a \le u(x) \le b \quad \text{for almost every } x \in \Omega \right\}$$

and using the control-to-state-mapping $S : L^2(\Omega) \to L^2(\Omega)$, $u \mapsto y$ solving (33.1), introduced above, we can write this problem in *reduced form* as

$$\min_{u \in U_{\text{ad}}} \frac{1}{2}\|Su - y^d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2}\|u\|_{L^2(\Omega)}^2. \tag{33.3}$$

To apply the general theory of the previous parts, we write this as $\min_{u \in L^2(\Omega)} J(u)$ for $J = F + G$ with

$$F(u) := \frac{1}{2}\|Su - y^d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2}\|u\|_{L^2(\Omega)}^2,$$
$$G(u) := \delta_{U_{\text{ad}}}(u).$$

EXISTENCE

Since S is linear and bounded (and hence weakly continuous) and the norm is weakly lower semicontinuous by Corollary 2.4 and convex, it follows from Lemmas 2.3 and 3.4 that F is weakly lower semicontinuous and convex; it is even strictly convex due to the control costs. Furthermore, $\operatorname{dom} F = L^2(\Omega)$ since S is well-defined on this space. Similarly, it can be shown that $U_{\text{ad}} \subset L^2(\Omega)$ is nonempty, closed, convex, and bounded and thus G is proper, lower semicontinuous, convex, and coercive by Lemma 2.5. We thus immediately obtain from Theorem 3.8 the existence of a unique optimal control $\bar{u} \in U_{\text{ad}}$ as well as a corresponding optimal state $\bar{y} := S\bar{u} \in H_0^1(\Omega)$.

OPTIMALITY CONDITIONS

To derive optimality conditions, we apply the Fermat principle as well as the calculus rules from Chapter 4. Although $\operatorname{dom} G = U_{\text{ad}} \subset L^2(\Omega)$ does *not* contain any interior points, we have $\operatorname{dom} F = L^2(\Omega)$ and hence we can still apply the sum rule from Theorem 4.14. In fact, since the squared norm in the Hilbert space $L^2(\Omega)$ (which we always identify with its dual

via the Fréchet–Riesz Theorem 1.14) is Fréchet differentiable, we obtain using the chain rule from Theorem 2.7 that

$$\nabla F(u) = S^*(Su - y^d) + \alpha u.$$

Using Theorems 4.5 and 4.14 and Lemma 4.8 and introducing the adjoint state $\bar{p} \in H_0^1(\Omega)$, we thus arrive at the primal-dual optimality conditions[1]

$$\begin{cases} \bar{p} = S^*(S\bar{u} - y^d), \\ (\bar{p} + \alpha\bar{u} \mid u - \bar{u})_{L^2(\Omega)} \geq 0 \quad \text{for all } u \in U_{\mathrm{ad}}, \end{cases} \tag{33.4}$$

where the second relation is often called a *variational inequality* for the optimal control; cf. [Tröltzsch, 2010, Theorem 2.25]. This relation, which is the explicit form of $-\bar{p} - \alpha\bar{u} \in \partial\delta_{U_{\mathrm{ad}}}(\bar{u})$, can by Lemma 6.21 and Example 6.28 (iii) be written equivalently for any $\gamma > 0$ as

$$\bar{u} = \operatorname{prox}_{\gamma\delta_{U_{\mathrm{ad}}}}(\bar{u} + \gamma(-\bar{p} - \alpha\bar{u}))\,. \tag{33.5}$$

Using the special choice $\gamma = \alpha^{-1}$ in the first expression as well as the pointwise characterization of proximal mappings on $L^2(\Omega)$ from Corollary 6.27 together with Example 6.25 (iii), we obtain the well-known *projection formula*

$$\bar{u}(x) = \operatorname{proj}_{[a,b]}\left(-\frac{1}{\alpha}\bar{p}(x)\right) = \begin{cases} a & \text{if } -\frac{1}{\alpha}\bar{p}(x) < a, \\ -\frac{1}{\alpha}p(x) & \text{if } -\frac{1}{\alpha}\bar{p}(x) \in [a,b], \\ b & \text{if } -\frac{1}{\alpha}\bar{p}(x) > b; \end{cases} \tag{33.6}$$

cf. [Tröltzsch, 2010, Theorem 2.28].

Remark 33.1. The relation (33.6) could also have been obtained by recognizing that $G(u)+\frac{\alpha}{2}\|u\|^2_{L^2(\Omega)} = (G^*_\alpha)^*$ by Theorem 7.11, where G^*_α is the Moreau envelope of G^*. We therefore obtain via Theorem 7.9

$$\begin{cases} \bar{p} = S^*(S\bar{u} - y^d), \\ \bar{u} = (\partial G^*)_\alpha(-\bar{p}), \end{cases}$$

where $(\partial G^*)_\alpha$ is the Yosida approximation of ∂G^*. Using its definition (7.18) together with Lemma 6.24 (ii), it is straightforward to verify that the second relation is in fact equivalent to (33.6).

[1] If the control-to-state mapping S is nonlinear but continuously differentiable, we can proceed in exactly the same fashion by using Theorems 13.5, 13.8, and 13.20 instead to arrive at (33.4) with $S'(\bar{u})^*$ in place of S^*.

EXPLICIT SPLITTING METHODS

Since F and G are proper, convex, and lower semicontinuous, and F is Fréchet differentiable with Lipschitz continuous gradient (since $\nabla F(u)$ is affine, it is globally Lipschitz with constant $L := \|S^*S + \alpha\mathrm{Id}\|_{\mathbb{L}(L^2(\Omega);L^2(\Omega))} = \|S\|^2_{\mathbb{L}(L^2(\Omega);L^2(\Omega))} + \alpha$), the optimal control $\bar{u}$ can be computed using the explicit splitting method (9.7). In our specific instance, this becomes the *projected gradient method*: Choose $u^0 \in L^2(\Omega)$ (e.g., $u^0 = 0$) and $\tau < 2L^{-1}$ and compute for $k = 0, \dots$

$$
\begin{cases}
y^{k+1} = Su^k & \text{by solving (33.1)},\\
p^{k+1} = S^*(y^{k+1} - y^d) & \text{by solving (33.2) for } h = y^{k+1} - y^d,\\
u^{k+1} = \operatorname{proj}_{[a,b]}\left((1-\tau\alpha)u^k - \tau p^{k+1}\right) & \text{almost everywhere.}
\end{cases} \tag{33.7}
$$

By Theorem 9.6, we then have $u^k \rightharpoonup \bar{u}$ in $L^2(\Omega)$. (Since G is not strongly convex, we do not obtain any rates.)

We can also apply the acceleration strategies from Chapter 12. Specifically, the *inertial projected gradient method* for $z^0 = u^0 \in L^2(\Omega)$, $\tau > 0$, and $\lambda_0 = 1$ consists in computing for $k = 0, \dots$

$$
\begin{cases}
y^{k+1} = Sz^k & \text{by solving (33.1)},\\
p^{k+1} = S^*(y^{k+1} - y^d) & \text{by solving (33.2) for } h = y^{k+1} - y^d,\\
u^{k+1} = \operatorname{proj}_{[a,b]}\left((1-\tau\alpha)u^k - \tau p^{k+1}\right) & \text{almost everywhere,}\\
\lambda_{k+1} = 2\left(1 + \sqrt{1 + 4\lambda_k^{-2}}\right), & \beta_{k+1} = \lambda_{k+1}(\lambda_k^{-1} - 1),\\
z^{k+1} = (1+\beta_{k+1})u^{k+1} - \beta_{k+1}u^k. &
\end{cases} \tag{33.8}
$$

By Theorem 12.12, we obtain the convergence of the function values $J(\tilde{u}^k) \to J(\bar{u})$ at the rate $O(1/k^2)$ as $k \to \infty$ (for the *nonergodic* sequence).

Similarly, we could also derive the *over-relaxed projected gradient method* from (12.17); however, since this method does not show any benefit over the projected gradient method for this problem, this is left as an exercise to the reader.

Instead, we will consider an alternative splitting. Since the maximal step length is constrained by the Lipschitz constant of F, it is beneficial to include as many parts of the functional as possible in the proximal point mapping. We thus turn to the splitting

$$
\begin{aligned}
F(u) &:= \frac{1}{2}\|Su - y^d\|^2_{L^2(\Omega)},\\
G_\alpha(u) &:= \delta_{U_{\mathrm{ad}}}(u) + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)}.
\end{aligned}
$$

To compute $\operatorname{prox}_{\gamma G_\alpha}$, we first observe that completing the square yields the scalar equality

$$\frac{1}{2\gamma}(z-t)^2 + \frac{\alpha}{2}z^2 = \frac{1+\alpha\gamma}{\gamma}\left(z - \frac{1}{1+\alpha\gamma}t\right)^2 + \frac{\gamma}{1+\alpha\gamma}t^2.$$

By ignoring the constant term, we hence have pointwise almost everywhere that for all $\gamma > 0$ and $v \in L^2(\Omega)$,

$$\begin{aligned}[\operatorname{prox}_{\gamma G_\alpha}(v)](x) &= \operatorname*{arg\,min}_{z\in[a,b]} \frac{1}{2\gamma}(z - v(x))^2 + \frac{\alpha}{2}z^2 \\ &= \operatorname*{arg\,min}_{z\in[a,b]} \frac{1+\alpha\gamma}{\gamma}\left(z - \frac{1}{1+\alpha\gamma}v(x)\right)^2 \\ &= \operatorname{proj}_{[a,b]}\left(\frac{1}{1+\alpha\gamma}v(x)\right).\end{aligned}$$

In place of (33.5), we thus have the equivalent optimality conditions

$$\bar{u} = \operatorname{prox}_{\gamma\delta_{U_{\mathrm{ad}}}}\left(\frac{1}{1+\alpha\gamma}(\bar{u} - \gamma\bar{p})\right). \tag{33.9}$$

From this, we obtain the corresponding (inertial) explicit splitting method for G_α by replacing the update for u^{k+1} in (33.7) (or (33.8) by

$$u^{k+1} = \operatorname{proj}_{[a,b]}\left(\frac{1}{1+\tau\alpha}\left(u^k - \tau p^{k+1}\right)\right) \qquad \text{almost everywhere,}$$

where τ now is only constrained by the smaller Lipschitz constant $L = \|S\|^2_{\mathbb{L}(L^2(\Omega);L^2(\Omega))}$, allowing larger steps.

In addition, since G_α is now strongly convex, we even get from Theorem 10.2 strong convergence of u^k at a linear rate.

SEMISMOOTH NEWTON METHOD

Using again the specific choice $\gamma = \alpha^{-1}$ and the definition of the adjoint state $\bar{p}$, we can write (33.5) as the nonsmooth equation $H(\bar{u}) = 0$ for

$$H : L^2(\Omega) \to L^2(\Omega), \qquad H(u) = u - \operatorname{proj}_{U_{\mathrm{ad}}}\left(-\frac{1}{\alpha}S^*(Su - y^d)\right). \tag{33.10}$$

Since $\operatorname{ran} S^* \subset H^1_0(\Omega) \hookrightarrow L^p(\Omega)$ for any $p > 2$ and $y^d \in L^2(\Omega)$, it follows from Example 14.12 (i) together with the chain rule Theorem 14.4 (since both $D_N\operatorname{proj}_{U_{\mathrm{ad}}}$ and S^*S are

clearly uniformly bounded) that H is Newton differentiable with a Newton derivative whose application to any $\delta u \in L^2(\Omega)$ is given pointwise almost everywhere by

$$[D_N H(u)\delta u](x) = \delta u(x) + \frac{1}{\alpha}\mathbb{1}_{[a,b]}\left(-\frac{1}{\alpha}[S^*Su](x)\right)[S^*S\delta u](x), \tag{33.11}$$

where $\mathbb{1}_{[a,b]}(t) = 1$ for $t \in [a,b]$ and 0 else. Clearly, $D_N H(u)$ is self-adjoint. Furthermore, since

$$\begin{aligned}(D_N H(u)\delta u \mid \delta u)_{L^2(\Omega)} &= \|\delta u\|^2_{L^2(\Omega)} + \frac{1}{\alpha}\int_{\{x\in\Omega \mid -\alpha^{-1}[S^*Su](x)\in[a,b]\}} |[S\delta u](x)|^2\,dx \\ &\geq \|\delta u\|^2_{L^2(\Omega)}\end{aligned}$$

for any $u \in L^2(\Omega)$, $D_N H(u)$ is uniformly positive definite and hence invertible. Theorem 14.1 thus guarantees that for any $u^0 \in L^2(\Omega)$, the semismooth Newton iteration

$$u^{k+1} = u^k - D_N H(u^k)^{-1} H(u^k) \tag{33.12}$$

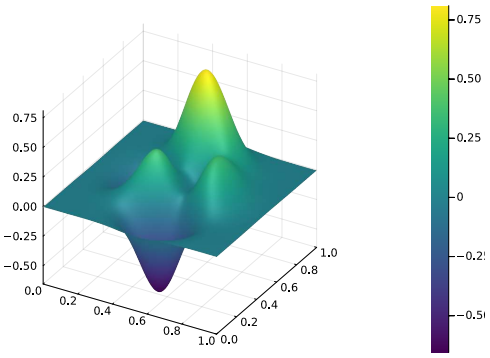
is locally superlinearly convergent. The properties of $D_N H(u)$ also imply that the Newton step (33.12) can be solved efficiently using a matrix-free conjugate gradient (CG) method (where for each CG iteration, one needs to solve two partial differential equations to apply S and S^*, followed by setting the result to zero almost everywhere where $u^k(x) \notin [a,b]$).

We indicate the performance of the projected gradient method, the explicit splitting method with G_α, its inertial variant (FISTA), and the semismooth Newton (SSN) method for the control constraints problem for the the control constraintsare $[a,b] = [-1,1]$, the control cost parameter $\alpha = 0.005$, and the target

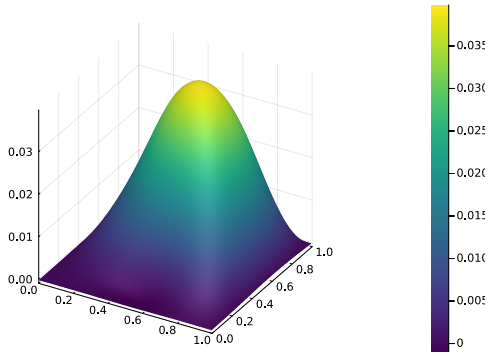
$$\begin{aligned}y^d(x_1,x_2) = \frac{3}{10}(4-6x_1)^2 e^{-(6x_1-3)^2-(6x_2-2)^2} \\ -\left(\frac{1}{5}(6x_1-3) - (6x_1-3)^3 - (6x_2-3)^5\right)e^{-(6x_1-3)^2-(6x_2-3)^2} \\ -\frac{1}{30}e^{-(6x_1-2)^2-(6x_2-3)^2};\end{aligned} \tag{33.13}$$

see Figure 33.1, which also shows the corresponding computed optimal control and state. Here and in the following, variables are discretized to a $N \times N$ grid for $N = 256$. For the splitting methods we take $\tau = 0.9/L^2$, where L an estimate of $\|S\|$. More details can again be found in the accompanying code [Clason and Valkonen, 2023].

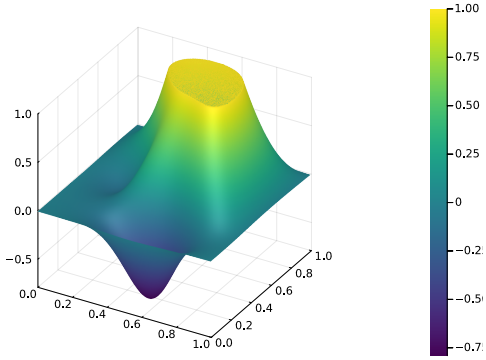
As function values are not meaningful in this problem (since they can be infinite for infeasible controls), we compare the residual norm $\|H(u^k)\|_{\mathbb{L}(L^2(\Omega);L^2(\Omega)}$ for H given by (33.10); these are shown in Figure 33.2. FISTA turns out to be the slowest algorithm; but this is not surprising since it only has a $O(1/k^2)$ rate of convergence, while for this strongly convex problem, the explicit splitting method has linear convergence, and the



(a) target y^d



(b) optimal state $\bar{y}$



(c) optimal control $\bar{u}$

Figure 33.1: Control constraints: target and optimal control and state.

SSN method has superlinear convergence. The projected gradient method shows linear convergence as well, albeit with a smaller constant than the explicit splitting method with G_α. Theoretically, indeed, the iterates of both methods converge linearly due to the strongly convex regularization term $\frac{\alpha}{2}\|u\|^2$. For the explicit splitting method with G_α, this is a direct consequence of Theorem 10.2. For the projected gradient method, where the strongly convex term is in F, we would need to adapt the proof to use Corollary 7.7 in place of Corollary 7.2. That the projected gradient method is slightly slower than the explicit splitting method with G_α can be attributed to the fact that when the regularization term is included in F, the Lipschitz factor L of F is higher, and consequently the step length parameter τ smaller. When the proximal step can be easily calculated, it is often more efficient to do more in the proximal step and less in the gradient step, as the former does not constrain the step length parameter. Indeed, taking iteration-dependent step length parameters $\tau_k \nearrow \infty$, the plain proximal point method converges by Theorem 10.1 superlinearly for strongly convex objectives. Of course, its steps can be very expensive.

As in the ℓ^1 fitting example, even though each SSN iteration involves solving a large indefinite system, the total computational time for getting the residual to machine precision is still lower than for the first-order splitting methods. Conversely, the (non-accelerated)

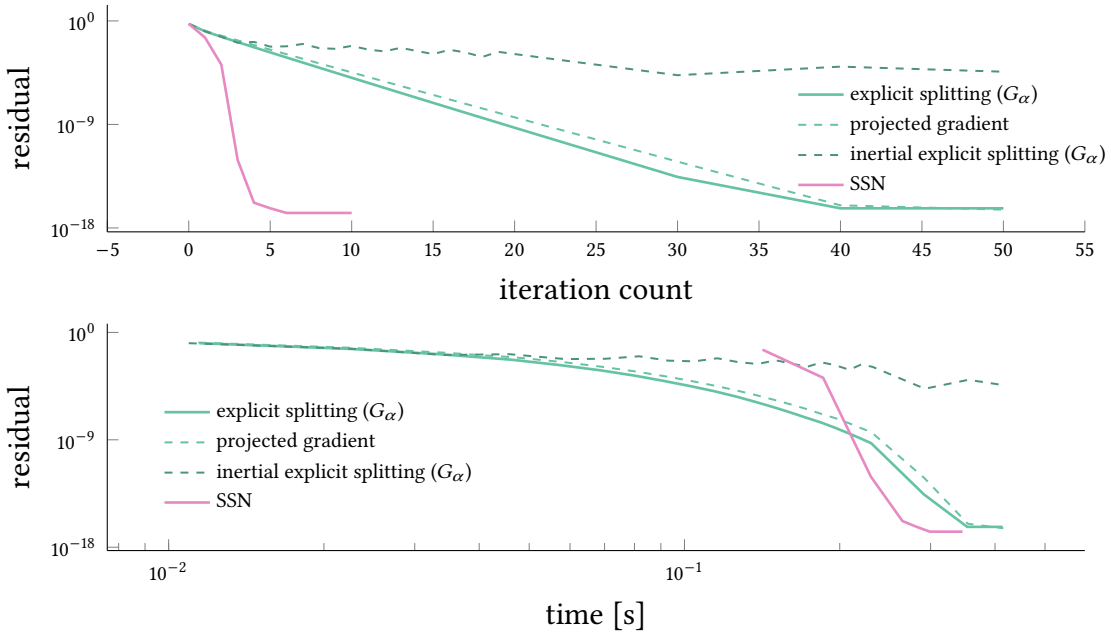


Figure 33.2: Algorithm performance for the control constraints example. We plot the residual $\|H(u^k)\|$ for H given by (33.10).

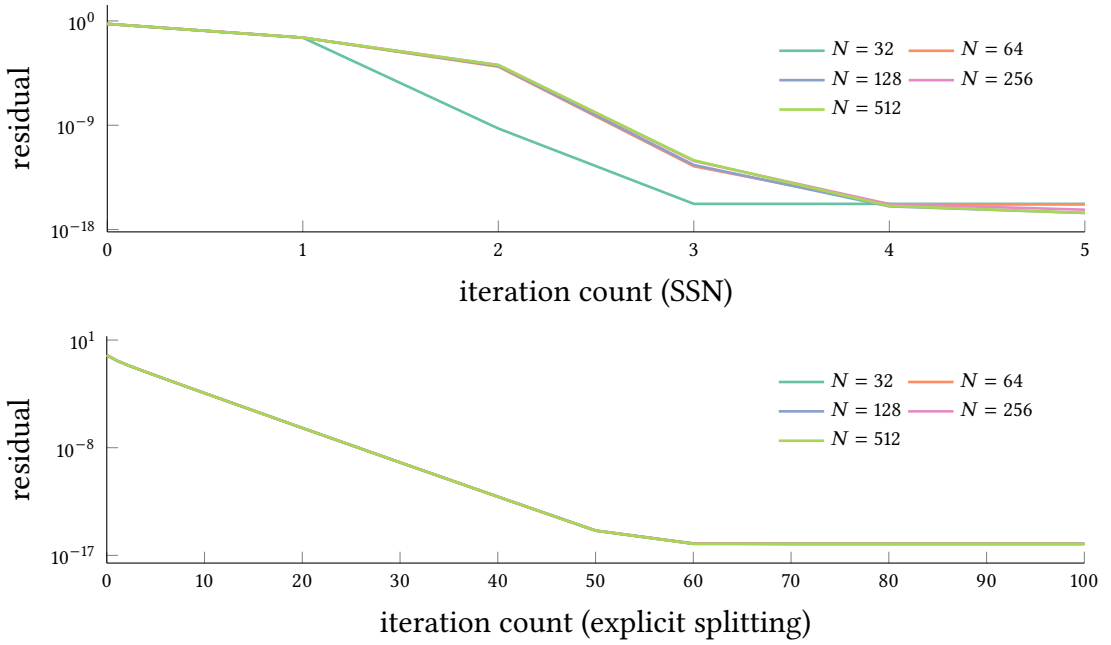


Figure 33.3: Control constraints: SSN performance versus dimension N. We plot the norm of residual $H(u^k)$.

splitting methods are faster in achieving a higher tolerance of about 10^{-6} and hence again may be preferable if high accuracy is not required.

Finally, as the convergence of these methods was shown on the infinite-dimensional level, it can be expected that the number of iterations required to solve optimality conditions for discretizations of the problem is independent of the fineness of the discretization. This beneficial property is referred to as *mesh independence*; see, e.g., [Hintermüller and Ulbrich, 2004] for its proof for a semismooth Newton method. We numerically indicate the dimension independence of both the SSN and explicit splitting methods in Figure 33.3.

33.2 STATE CONSTRAINTS

PROBLEM DESCRIPTION

There are also occasions when one wishes to put pointwise bounds on the state, for example when looking for optimal heat sources to achieve on average a comfortable temperature in a room without risking hot spots of a face-melting temperature. Staying in the current setting otherwise, we thus want to solve for a given upper bound $y_{\max} > 0$ (for simplicity) the *state-constrained optimal control problem*

$$\min_{u\in L^2(\Omega), y\in H_0^1(\Omega)} \frac{1}{2}\|y - y^d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2}\|u\|_{L^2(\Omega)}^2$$
$$\text{subject to (33.1)} \qquad \text{and} \qquad y(x) \le y_{\max} \quad \text{for almost every } x \in \Omega.$$

This has a similar structure as the control-constrained problem, and we will follow the same general approach. However, this is more delicate here, since we now have to apply the chain rule for the subdifferential of the indicator functional, which requires a nonempty interior of the corresponding set – which does not hold in $L^2(\Omega)$, and the dual space of $L^\infty(\Omega)$ is very difficult to characterize. We thus instead assume that $\Omega \subset \mathbb{R}^d$ is convex for $d \le 3$ so that the solutions to the state equation are continuous and we can impose the state constraints *everywhere*. We then define the admissible set

$$Y_{\text{ad}} := \left\{ w \in C(\overline{\Omega}) \;\middle|\; w(x) \le y_{\max} \text{ for all } x \in \overline{\Omega} \right\}$$

as well as

$$\tilde{S} : L^2(\Omega) \to C(\overline{\Omega}), \qquad u \mapsto y \text{ solving (33.1)},$$

which is well-defined and continuous under our assumptions on Ω. The problem in reduced form is then

$$\min_{u\in L^2(\Omega)} \frac{1}{2}\|Su - y^d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2}\|u\|_{L^2(\Omega)}^2 + \delta_{Y_{\text{ad}}}(\tilde{S}u), \tag{33.14}$$

which has the general form $J = F + \tilde{G}$ with $F : L^2(\Omega) \to \mathbb{R}$ as above and $\tilde{G} = \delta_{Y_{\text{ad}}} \circ \tilde{S} : L^2(\Omega) \to C(\overline{\Omega})$.

EXISTENCE

Since $\tilde{S}$ is continuous, Y_{ad} clearly is nonempty, convex, and closed, and F is coercive on $L^2(\Omega)$ due to the control costs while $\tilde{G}$ is nonnegative, we immediately obtain the existence of an optimal control $\bar{u} \in L^2(\Omega)$ and an admissible optimal state $\bar{y} \in Y_{\mathrm{ad}} \cap H_0^1(\Omega)$ by Theorem 2.1. Since F is strictly convex, this control is again unique.

OPTIMALITY CONDITIONS

Setting $G = \delta_{Y_{\mathrm{ad}}} : C(\overline{\Omega}) \to \overline{\mathbb{R}}$, the problem (33.14) has the form $\min_u F(u) + G(\tilde{S}u)$. Since we are working with continuous functions here and the state equation is linear, we have for $u_0 = 0 \in L^2(\Omega)$ that $y_0 := \tilde{S}u_0 = 0 < y_{\max}$ and hence that $y_0 \in \operatorname{int} Y_{\mathrm{ad}}$. We can thus apply the Fenchel–Rockafellar Theorem 5.11 to obtain the primal-dual optimality condition

$$\begin{cases} \bar{\mu} \in \partial\delta_{Y_{\mathrm{ad}}}(\tilde{S}\bar{u}), \\ -\tilde{S}^*\bar{\mu} = S^*(S\bar{u} - y^d) + \alpha\bar{u}, \end{cases} \tag{33.15}$$

where we have again used the fact that F is Fréchet differentiable with the given gradient. Since $\bar{\mu} \in \partial\delta_{Y_{\mathrm{ad}}} \subset C(\overline{\Omega})^* \cong \mathcal{M}(\Omega)$ is a Radon measure (cf. Example 1.3 (iv)), a more explicit, "pointwise", interpretation analogous to (33.4) and (33.5) is more involved and involves results from measure theory.

First, by Lemma 4.8, $\bar{\mu} \in \mathcal{M}(\Omega)$ and $\bar{y} \in C(\overline{\Omega})$ satisfy

$$\int_\Omega (\tilde{y}(x) - \bar{y}(x))\, d\bar{\mu}(x) \le 0 \qquad \text{for all } \tilde{y} \le y_{\max}.$$

By a pointwise argument similar to Example 4.9, it follows that

$$\bar{\mu} \ge 0 \qquad \text{and} \qquad \int_\Omega (\bar{y}(x) - y_{\max})\, d\bar{\mu}(x) = 0, \tag{33.16}$$

i.e., that $\bar{\mu}$ is a nonnegative Radon measure whose support is contained in the active set $\{x \in \Omega \mid \bar{y}(x) = y_{\max}\}$.

Second, using the continuous (and dense) embedding of $W^{1,p}(\Omega) \hookrightarrow C(\overline{\Omega})$ for p sufficiently large, it is possible to show that any $\mu \in \mathcal{M}(\Omega)$ satisfies $S^*\mu \in W^{1,q}(\Omega)$ for some sufficiently small $q > 1$ and can therefore be characterized as the unique solution $\tilde{p} = S^*\mu$ to

$$\int_\Omega \nabla w(x) \cdot \nabla \tilde{p}(x)\, dx = \int_\Omega w(x)\, d\mu(x) \qquad \text{for all } w \in W_0^{1,p}(\Omega), \tag{33.17}$$

see, e.g., [Clason and Schiela, 2017; Meyer et al., 2011] and the references therein.

Combining (33.1), (33.17) added to (33.2), and (33.16), we obtain from (33.15) the (suitably interpreted) necessary and sufficient optimality conditions

$$\text{(33.18)}\qquad \begin{cases} \alpha\bar{u} + \bar{p} = 0, \\ \quad -\Delta\bar{y} = \bar{u}, \\ \quad -\Delta\bar{p} = \bar{y} - y^d + \bar{\mu}, \\ \qquad \bar{y} \le y_{\max}, \qquad \bar{\mu} \ge 0, \qquad \displaystyle\int_\Omega (\bar{y}(x) - y_{\max})\, d\bar{\mu}(x) = 0, \end{cases}$$

compare [Tröltzsch, 2010, Theorem 6.5]. (The last line corresponds again to the classical complementarity conditions from nonlinear optimization.)

SEMISMOOTH NEWTON METHOD

Since (33.18) cannot fully be expressed pointwise, a numerical solution is difficult. We thus instead apply the Moreau–Yosida regularization from Section 7.3 to G, which entails replacing $\partial G : C(\overline{\Omega}) \rightrightarrows \mathcal{M}(\Omega)$ in (33.15) by its Yosida approximation $(\partial G)_\gamma : L^2(\Omega) \to L^2(\Omega)$ for $\gamma > 0$ (and, as a consequence, $\tilde{S}$ by S). Following the computation in Example 7.10 (iii) and using Corollary 6.27, we obtain the pointwise almost everywhere expression

$$[H_\gamma(y)](x) := [(\partial G)_\gamma(y)](x) = \frac{1}{\gamma}(y(x) - y_{\max})^+ := \frac{1}{\gamma}\max\{0, y(x) - y_{\max}\}$$

and hence the regularized optimality conditions for $(u_\gamma, y_\gamma, p_\gamma)$

$$\text{(33.19)}\qquad \begin{cases} \alpha u_\gamma + p_\gamma = 0, \\ \quad -\Delta y_\gamma = u_\gamma, \\ \quad -\Delta p_\gamma = y_\gamma - y^d + \dfrac{1}{\gamma}(y_\gamma - y_{\max})^+, \end{cases}$$

where we have used the single-valued regularized relation $\mu_\gamma = H_\gamma(y_\gamma)$ to eliminate μ_γ in the last line. By Theorem 7.9 and the computation in Example 7.10 (iii), $u_\gamma \in L^2(\Omega)$ is the (unique) minimizer of

$$\text{(33.20)}\qquad \min_{u \in L^2(\Omega)} \frac{1}{2}\|Su - y^d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)} + \frac{1}{2\gamma}\|(Su - y_{\max})^+\|^2_{L^2(\Omega)},$$

which guarantees the existence of a (unique) solution $(u_\gamma, y_\gamma, p_\gamma) \in L^2(\Omega) \times H^1_0(\Omega) \times H^1_0(\Omega)$. Of course, we cannot expect $y_\gamma \in Y_{\mathrm{ad}}$ in general; but a lower semicontinuity argument as in Theorem 2.1 shows that $u_\gamma \to \bar{u}$ and $y_\gamma \to \bar{y} \in Y_{\mathrm{ad}}$ strongly in $L^2(\Omega)$ as $\gamma \to 0$; see [De los Reyes, 2015, Theorem 6.5]

To apply a semismooth Newton method, we first eliminate $u_\gamma = -\frac{1}{\alpha}p_\gamma$ from the first relation of (33.19) in the second relation to obtain the *reduced optimality system*

$$\begin{cases} -\Delta y_\gamma + \dfrac{1}{\alpha} p_\gamma = 0, \\ -\Delta p_\gamma - y_\gamma - \dfrac{1}{\gamma}(y_\gamma - y_{\max})^+ + y^d = 0, \end{cases} \tag{33.21}$$

which is a nonsmooth system of equations for (p_γ, y_γ). Since $y_\gamma \in H^1_0(\Omega) \hookrightarrow L^r(\Omega)$ for some $r > 2$, the superposition operator $H_\gamma : y \mapsto \frac{1}{\gamma}(y - y_{\max})^+$ is semismooth from $L^r(\Omega) \to L^2(\Omega)$ by Example 14.12 (i) with Newton derivative given pointwise almost everywhere by

$$[D_N H_\gamma(y)](x) = \frac{1}{\gamma} \mathbb{1}_{(y_{\max},\infty)}(y(x)).$$

A semismooth Newton step thus consists in solving for $(\delta p, \delta y) \in H^1_0(\Omega)$ in

$$\begin{pmatrix} \frac{1}{\alpha}\mathrm{Id} & -\Delta \\ -\Delta & -\mathrm{Id} - \frac{1}{\gamma}\mathbb{1}_{(y_{\max},\infty)}(y^k) \end{pmatrix} \begin{pmatrix} \delta p \\ \delta y \end{pmatrix} = - \begin{pmatrix} -\Delta y^k + \frac{1}{\alpha} p^k \\ -\Delta p^k - y^k - \frac{1}{\gamma}(y^k - y_{\max})^+ + y^d \end{pmatrix} \tag{33.22}$$

and then setting $p^{k+1} := p^k + \delta p$, $y^{k+1} := y^k + \delta y$. The block operator on the left-hand side of (33.22) is a self-adjoint block operator that can be shown to be boundedly invertible (by using the fact that $-\Delta$ is a self-adjoint and positive definite operator) for any $y \in H^1_0(\Omega)$. Hence this semismooth Newton method converges locally superlinearly according to Theorem 14.1.

Since the PDE constraint is linear, we can further rewrite the Newton step to avoid applying differential operators when evaluating the right-hand side. Using $\delta p = p^{k+1} - p^k$ and similarly for δy, and writing $(y - y_{\max})^+ = \mathbb{1}_{(y_{\max},\infty)}(y)(y - y_{\max})$, we can rearrange the Newton step as

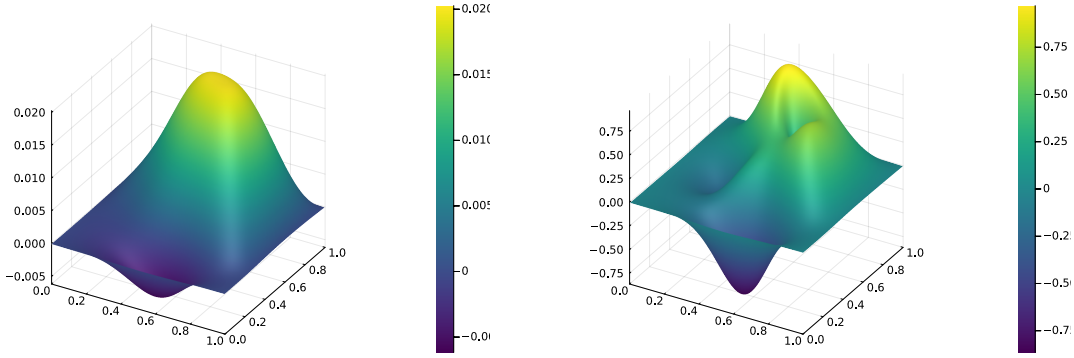
$$\begin{pmatrix} \frac{1}{\alpha}\mathrm{Id} & -\Delta \\ -\Delta & -\mathrm{Id} - \frac{1}{\gamma}\mathbb{1}_{(y_{\max},\infty)}(y^k) \end{pmatrix} \begin{pmatrix} p^{k+1} \\ y^{k+1} \end{pmatrix} = \begin{pmatrix} 0 \\ -y^d - \frac{1}{\gamma}\mathbb{1}_{(y_{\max},\infty)}(y^k) y_{\max} \end{pmatrix}. \tag{33.23}$$

This is closely related to the *primal-dual active set method* for quadratic optimization problems with box constraints; see [Hintermüller et al., 2002; Ito and Kunisch, 2008]. Furthermore, if

$$\mathbb{1}_{(y_{\max},\infty)}(y^{k+1}) = \mathbb{1}_{(y_{\max},\infty)}(y^k)$$

almost everywhere, it is straightforward to verify that (33.23) coincides with the reduced optimality conditions (33.21), which implies that $u^{k+1} := -\frac{1}{\alpha}p^{k+1} = -\Delta y^{k+1}$ is the desired optimal control. (This *finite termination property* of semismooth Newton methods for quadratic optimization problems is one reason for its efficiency for such problems.)

In practice, the radius of convergence for the semismooth Newton method applied to such a Moreau–Yosida regularization shrinks with $\gamma \to 0$. A possible way of dealing with this



(a) optimal state y_γ (b) optimal control u_γ

Figure 33.4: State constraints: optimal control and state for $\gamma = 10^{-4}$.

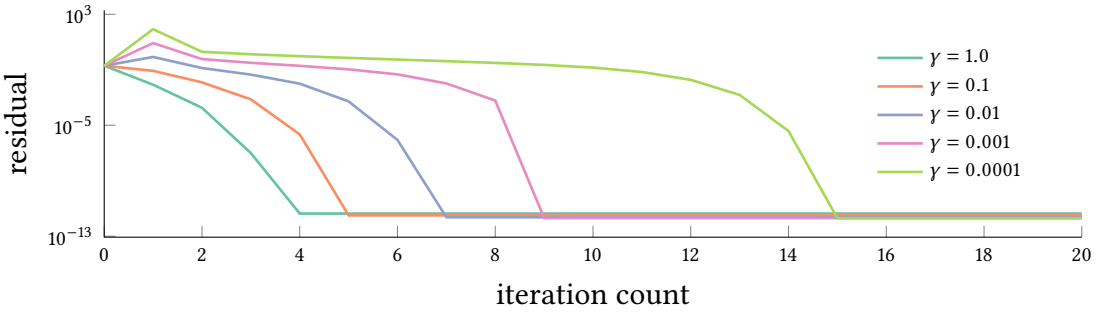


Figure 33.5: State constraints: SSN performance versus Moreau–Yosida parameter γ. We plot the norm of residual, which corresponds to the left hand side of (33.21) instantiated at (p^k, y^k).

is the following *continuation strategy*: Starting with a sufficiently large value of γ, solve a sequence of problems with decreasing γ (e.g., $\gamma^k = \gamma^0/2^k$), taking the solution of the previous problem as the starting point for the next (which is hopefully close enough to the solution to lie within the convergence region; otherwise the continuation has to be terminated or the reduction strategy for γ adapted).

Our target y^d, dimension $N = 256$, and control cost parameter $\alpha = 0.005$ are exactly the same as for control constraints in the previous section. We take $y_{\max} = 0.02$; all other details are specified in [Clason and Valkonen, 2023]. The optimal state, control, and adjoint state for $\gamma = 10^{-4}$ are exemplarily shown in 33.4. We illustrate in Figure 33.5 the dependence of the convergence speed of the SSN method on the Moreau–Yosida parameter γ.

34 DISCRETE-VALUED OPTIMAL CONTROL

The final example illustrates the application to a challenging class of *mixed-integer PDE-constrained optimization problems*, where the desired controls are functions that should only take values from a specified discrete set. Such problems arise in, e.g., topology optimization, material parameter identification with a priori information, and joint image reconstruction and segmentation. The purpose of this example is to demonstrate how nonsmooth optimization can be used to impose strong, non-trivial, structural properties on the solution.

Specifically, for a given set of values $u_1 < u_2 < \cdots < u_m \in \mathbb{R}$, we consider the *admissible set*

$$U_{\mathrm{ad}} := \left\{ u \in L^2(\Omega) \;\middle|\; u(x) \in \{u_1, \ldots, u_m\} \quad \text{for almost every } x \in \Omega \right\}.$$

This set is nonconvex and not weakly closed, which makes the standard theory inapplicable. The usual approach of replacing U_{ad} with its closed convex hull

$$\overline{\mathrm{co}}\, U_{\mathrm{ad}} = \left\{ u \in L^2(\Omega) \;\middle|\; u(x) \in [u_1, u_m] \quad \text{for almost every } x \in \Omega \right\}$$

however is insufficient as it loses information about the interior values $u_2, \ldots, u_{m-1}$. We therefore proceed differently by first adding a *pointwise* quadratic penalty that promotes discrete values of lower magnitude (assuming that lower magnitude is preferable, all other things being equal), i.e., we consider instead of $\delta_{\{u_1,\ldots,u_m\}}$ the weighted indicator function

$$\hat{g}(t) := \frac{1}{2}|t|^2 + \delta_{\{u_1,\ldots,u_m\}}(t)$$

whose convex envelope is readily seen by graphical arguments to be

$$g(t) = \begin{cases} \frac{1}{2}\left((u_i + u_{i+1})t - u_i u_{i+1}\right) & \text{if } t \in [u_i, u_{i+1}], \quad 1 \le i < m, \\ \infty & \text{else;} \end{cases} \tag{34.1}$$

see Figure 34.1. (This will be rigorously verified in Remark 34.2 below.)

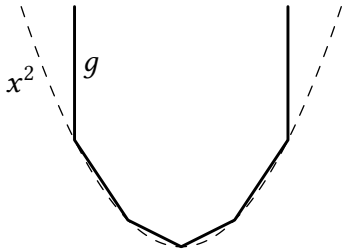


Figure 34.1: Plot of g given by (34.1) for $u_1, \dots, u_5 = -1, -0.5, 0, 0.5, 1$. The graph of $x \mapsto x^2$ is also drawn with a dashed line.

34.1 PROBLEM DESCRIPTION

We now consider for given $u_1 < \cdots < u_m$ and $y^d \in L^2(\Omega)$ the model *discrete-valued control problem*

$$\min_{u \in L^2(\Omega)} \frac{1}{2}\|Su - y^d\|^2_{L^2(\Omega)} + \alpha G(u), \tag{34.2}$$

where $S : L^2(\Omega) \to L^2(\Omega)$ is again the control-to-state mapping for (33.1) introduced at the beginning of Chapter 33, $\alpha > 0$, and

$$G : L^2(\Omega) \to \overline{\mathbb{R}}, \qquad G(u) := \int_\Omega g(u(x))\,dx.$$

Since g, given by (34.1), is proper, convex, and lower semicontinuous, so is G by Lemma 3.7. We thus again obtain from Theorem 3.8 the existence of an optimal control $\bar{u} \in L^2(\Omega)$ as well as a corresponding optimal state $\bar{y} := S\bar{u} \in H^1_0(\Omega)$. (Since G is convex but not strictly convex, we cannot directly conclude uniqueness; however, since F is strictly convex, the optimal state $\bar{y} = S\bar{u}$ must be unique, which then yields uniqueness of $\bar{u}$ by the continuous invertibility of S.)

Remark 34.1. In the special case that $m = 3$ and $u_1 = -M \ll u_2 = 0 \ll u_3 = M$, the convex penalty (34.1) simplifies to

$$g(v) = \begin{cases} \frac{M}{2}|t| & \text{if } |t| \le M, \\ \infty & \text{else.} \end{cases}$$

In other words – after rescaling $\alpha \mapsto \frac{2}{M}\alpha$ – (34.2) becomes the *sparse control problem*

$$\min_{u \in L^2(\Omega)} \frac{1}{2}\|Su - y^d\|^2_{L^2(\Omega)} + \alpha\|u\|_{L^1} + \delta_{\{[-M,M]\}}(u),$$

which seeks to find a (bounded) optimal control that is zero on as large a part of the domain Ω as possible; see [Stadler, 2009; Vossen and Maurer, 2006]. Hence all results in this chapter can be specialized to this problem as well.

However, in the absence of the control constraints $-M \le u(x) \le M$ almost everywhere (or additional $L^2(\Omega)$ regularization), the problem is no longer coercive in $L^1(\Omega)$, and an optimal control must be sought in the space $\mathcal{M}(\Omega)$ of Radon measures [Bidaut, 1975]. In this case, it is still possible to exploit similar arguments using the "preduality" of $C_0(\Omega)$ and $\mathcal{M}(\Omega)$; see Remark 5.12 and [Casas et al., 2012; Clason and Kunisch, 2014; Clason and Schiela, 2017].

34.2 OPTIMALITY CONDITIONS

Again we can derive optimality conditions from the Fermat principle together with calculus rules. As in Chapter 33, Theorems 4.5 and 4.14 yield the primal-dual optimality conditions

$$\begin{cases} -\bar{p} = S^*(S\bar{u} - y^d), \\ \bar{p} \in \partial\left(\alpha G(\bar{u})\right), \end{cases}$$

for the adjoint state $\bar{p} \in H_0^1(\Omega)$. The last relation implies by Lemma 4.13 (i) that $\bar{p} = \alpha\bar{q}$ for some $\bar{q} \in \partial G(\bar{u})$, i.e., $\frac{1}{\alpha}\bar{p} \in \partial G(\bar{u})$. Further applying the "convex inverse function" Lemma 5.8 leads to the equivalent optimality conditions

$$\begin{cases} -\bar{p} = S^*(S\bar{u} - y^d), \\ \bar{u} \in \partial G^*\left(\dfrac{1}{\alpha}\bar{p}\right). \end{cases} \tag{34.3}$$

To derive from this system some information on the structure of optimal controls, we need to obtain an explicit representation for ∂G^*, which we can do pointwise via Theorems 4.11 and 5.5.

We first compute the subdifferential $\partial g(v)$ at a point $v \in [u_1, u_m]$. To that end, we write

$$g(t) = g_1(t) + \delta_{[u_1,u_m]}$$

for the real-valued extension

$$g_1 : \mathbb{R} \to \mathbb{R}, \qquad g_1(t) = \begin{cases} u_1 t - \frac{1}{2}u_1^2 & \text{if } t \le u_1, \\ \frac{1}{2}\left((u_i + u_{i+1})t - u_i u_{i+1}\right) & \text{if } t \in [u_i, u_{i+1}], \quad 1 \le i < m, \\ u_m t - \frac{1}{2}u_m^2 & \text{if } t \ge u_m. \end{cases}$$

This is a convex PC^1 function, hence by Theorems 13.8 and 14.8 we have that

$$\partial g_1(t) = \begin{cases} \{u_1\} & \text{if } t < u_1, \\ [u_1, \frac{1}{2}(u_1 + u_2)] & \text{if } t = u_1, \\ \{\frac{1}{2}(u_i + u_{i+1})\} & \text{if } t \in (u_i, u_{i+1}), \quad 1 \le i < m, \\ \left[\frac{1}{2}(u_{i-1} + u_i), \frac{1}{2}(u_i + u_{i+1})\right] & \text{if } t = u_i, \quad 1 \le i < m, \\ [\frac{1}{2}(u_{m-1} + u_m), u_m] & \text{if } t = u_m, \\ \{u_m\} & \text{if } t > u_m. \end{cases}$$

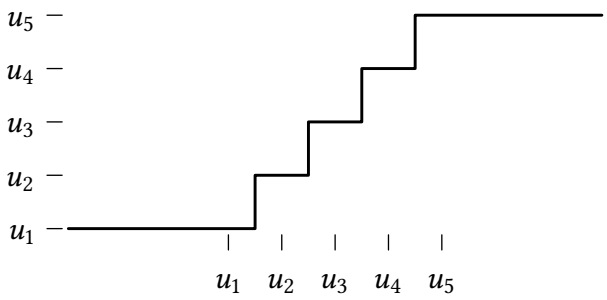


Figure 34.2: Plot of ∂g^* given by (34.5) for $u_1, \dots, u_5 = -1, -0.5, 0, 0.5, 1$.

Furthermore, since g_1 is continuous in any $t \in \mathbb{R}$, we can apply the sum rule Theorem 4.14 together with the characterization of the subdifferential of the indicator function as a normal cone analogous to Example 4.9 to obtain

$$\partial g(t) = \begin{cases} (-\infty, \frac{1}{2}(u_1 + u_2)] & \text{if } t = u_1, \\ \{\frac{1}{2}(u_i + u_{i+1})\} & \text{if } t \in (u_i, u_{i+1}), \quad 1 \le i < m, \\ \left[\frac{1}{2}(u_{i-1} + u_i), \frac{1}{2}(u_i + u_{i+1})\right] & \text{if } t = u_i, \quad 1 \le i < m, \\ [\frac{1}{2}(u_{m-1} + u_m), \infty) & \text{if } t = u_m, \\ \emptyset & \text{else.} \end{cases} \tag{34.4}$$

We can now simply appeal to Lemma 5.8 (keeping in mind that subdifferentials are always closed) to obtain

$$\partial g^*(q) \in \begin{cases} \{u_1\} & \text{if } q \in \left(-\infty, \frac{1}{2}(u_1 + u_2)\right), \\ [u_i, u_{i+1}] & \text{if } q = \frac{1}{2}(u_i + u_{i+1}), \quad 1 \le i < m, \\ \{u_i\} & \text{if } q \in \left(\frac{1}{2}(u_{i-1} + u_i), \frac{1}{2}(u_i + u_{i+1})\right), \quad 1 < i < m, \\ \{u_d\} & \text{if } q \in \left(\frac{1}{2}(u_{m-1} + u_m), \infty\right), \\ \emptyset & \text{else.} \end{cases} \tag{34.5}$$

We illustrate ∂g^* in Figure 34.2.

Applying (34.5) and Theorems 4.11 and 5.5 in (34.12), we now obtain the explicit primal-dual optimality conditions

$$\begin{cases} -\bar{p} = S^*(S\bar{u} - y^d), \\ \bar{u}(x) \in \begin{cases} \{u_i\} & \text{if } \bar{p}(x) \in Q_i, \\ [u_i, u_{i+1}] & \text{if } \bar{p}(x) \in Q_{i,i+1}, \end{cases} \end{cases} \tag{34.6}$$

for the sets

$$\begin{aligned} Q_i &= \left\{ q \;\middle|\; \tfrac{\alpha}{2}(u_{i-1} + u_i) < q < \tfrac{\alpha}{2}(u_i + u_{i+1}) \right\}, \quad 1 \le i \le m, \\ Q_{i,i+i} &= \left\{ q \;\middle|\; q = \tfrac{\alpha}{2}(u_i + u_{i+1}) \right\}, \quad 1 \le i < m, \end{aligned}$$

where we have set $u_0 = -\infty$ and $u_{m+1} = \infty$ to avoid the need for further case distinctions. This immediately implies that even after convex relaxation, the optimal control will take on almost everywhere one of the prescribed discrete values except where the adjoint state happens to attain one of the critical values $\frac{\alpha}{2}(u_i + u_{i+1})$, $i = 1, \dots, m$. If this attainment can be excluded – as in our case, where $\bar{p}$ is harmonic as the solution of a Poisson equation and thus cannot be constant on a set of positive measure unless it vanishes everywhere – the relaxed control will still be admissible for the original nonconvex problem and thus locally optimal for the (weighted) discrete problem. We also see the effect of α on the control: the larger α, the more likely that $\bar{p}(x) \in Q_i$ corresponding to an u_i of lower magnitude.

Remark 34.2. We point out that it was not necessary to derive the explicit form of the conjugate itself in order to obtain explicit primal-dual optimality conditions. Nevertheless, this can be useful for verifying that g is indeed the convex envelope of $\hat{g}$.

First, we have by definition that

$$\hat{g}^*(q) \coloneqq \sup_{t \in \{u_1, \dots, u_m\}} q \cdot t - \frac{1}{2}|t|^2 = u_i q - \frac{1}{2}|u_i|^2$$

for some $1 \leq i \leq m$. Since the u_i are assumed to be ordered by increasing magnitude, it therefore suffices to check for given $q \in \mathbb{R}$ whether

$$u_i q - \frac{1}{2}|u_i|^2 \leq u_{i+1} q - \frac{1}{2}|u_{i+1}|^2$$

or, equivalently, whether

$$q(u_{i+1} - u_i) \leq \frac{1}{2}(u_{i+1}^2 - u_i^2).$$

Since by assumption $u_{i+1} - u_i > 0$, this in turn is equivalent to

$$q \leq \frac{1}{2}(u_{i+1} + u_i).$$

Hence

$$\hat{g}^*(q) = \begin{cases} q u_1 - \frac{1}{2}u_1^2 & \text{if } q \leq \frac{1}{2}(u_1 + u_2), \\ q u_i - \frac{1}{2}u_i^2 & \text{if } \frac{1}{2}(u_{i-1} + u_i) \leq q \leq \frac{1}{2}(u_i + u_{i+1}), 1 < i < m, \\ q u_m - \frac{1}{2}u_m^2 & \text{if } \frac{1}{2}(u_m + u_{m-1}) \leq q. \end{cases}$$

A similar – albeit more tedious – calculation using the piecewise differentiability of g shows that

$$g^*(q) = \hat{g}^*(q).$$

By Theorem 5.1 and the convexity of g, we thus have

$$\hat{g}^\Gamma = \hat{g}^{**} = (\hat{g}^*)^* = (g^*)^* = g.$$

34.3 ALGORITHMS

34.3.1 PROXIMAL GRADIENT METHODS

As in Section 33.1, we can compute a solution to (34.2) via an explicit splitting method, for which we only need an explicit characterization of the proximal point mapping $\operatorname{prox}_{\gamma(\alpha G)}$. By Corollary 6.27, this is given pointwise almost everywhere by the proximal point mapping for αg, which we can derive analogously to Example 6.25 (ii). For the sake of presentation, we fix $\alpha = 1$ for now.

By the definition of the proximal point mapping, $w = \operatorname{prox}_{\gamma g}(t) = (\mathrm{Id} + \gamma \partial g)^{-1}(t)$ holds for any $t \in \mathbb{R}$ if and only if $t \in \{w\} + \gamma \partial g(w)$. Using (34.4), we thus distinguish the following cases for w:

(i) $w = u_1$: In this case,

$$t \in \{w\} + \gamma \left(-\infty, \tfrac{1}{2}(u_1 + u_2)\right] = \left(-\infty, (1 + \tfrac{\gamma}{2})u_1 + \tfrac{\gamma}{2}u_2\right].$$

(ii) $w \in (u_i, u_{i+1})$ for $1 \leq i < m$: In this case,

$$t \in \{w\} + \gamma\{\tfrac{1}{2}(u_i + u_{i+1})\},$$

which first can be solved for w to yield

$$w = t - \tfrac{\gamma}{2}(u_i + u_{i+1});$$

inserting this into $w \in (u_i, u_{i+1})$ and simplifying then gives

$$t \in \left((1 + \tfrac{\gamma}{2})u_i + \tfrac{\gamma}{2}u_{i+1}, \tfrac{\gamma}{2}u_i + (1 + \tfrac{\gamma}{2})u_{i+1}\right).$$

(iii) $w = u_i$, $1 < i < m$: Proceeding as in the first case, we obtain

$$t \in \left[\tfrac{\gamma}{2}u_{i-1} + (1 + \tfrac{\gamma}{2})u_i, (1 + \tfrac{\gamma}{2})u_i + \tfrac{\gamma}{2}u_{i+1}\right].$$

(iv) $w = u_m$: Similarly, this implies that

$$t \in \left[\tfrac{\gamma}{2}u_{m-1} + (1 + \tfrac{\gamma}{2})u_m, \infty\right).$$

Since this is a complete and disjoint case distinction for $t \in \mathbb{R}$, we obtain that

$$\operatorname{prox}_{\gamma g}(t) = \begin{cases} u_i & \text{if } t \in \left[\left(1 + \frac{\gamma}{2}\right) u_i + \frac{\gamma}{2}u_{i-1}, \left(1 + \frac{\gamma}{2}\right) u_i + \frac{\gamma}{2}u_{i+1}\right], \\ t - \frac{\gamma}{2}(u_i + u_{i-1}) & \text{if } t \in \left(\left(1 + \frac{\gamma}{2}\right) u_{i-1} + \frac{\gamma}{2}u_i, \left(1 + \frac{\gamma}{2}\right) u_i + \frac{\gamma}{2}u_{i-1}\right), \end{cases} \tag{34.7}$$

again with the convention that $u_0 = -\infty$ and $u_{m+1} = \infty$. The proximal point mapping therefore has the form of a *generalized shrinkage operator*. We illustrate this mapping in Figure 34.3.

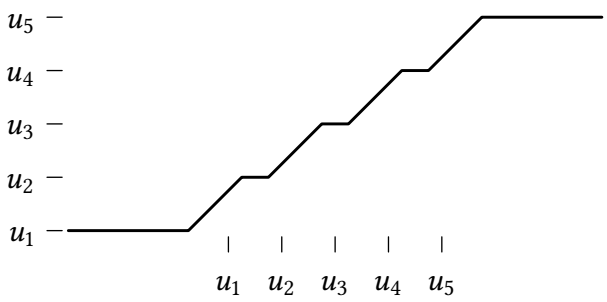


Figure 34.3: Plot of proxγg given by (34.7) for $u_1, \dots, u_5 = -1, -0.5, 0, 0.5, 1$ and $\gamma = 0.5$.

Remark 34.3. In the special case of sparse control ($m = 3$ and $u_1 = -M \ll u_2 = 0 \ll u_3 = M$), the proximal point mapping reduces to a projection of the well-known *soft-shrinkage operator* from Example 6.25.

Choosing $\tau < 2L^{-1}$ for $L = \|S\|^2_{\mathbb{L}(L^2(\Omega);L^2(\Omega))}$ and $u^0 = u_i$ for some $1 \le i \le m$, we can thus apply the *proximal gradient method*

$$\text{(34.8)} \qquad \begin{cases} y^{k+1} = Su^k & \text{by solving (33.1)}, \\ p^{k+1} = S^*(y^d - y^{k+1}) & \text{by solving (33.2) for } h = y^d - y^{k+1}, \\ u^{k+1}(x) = \operatorname{prox}_{(\tau\alpha)g}\left(u^k(x) + \tau p^{k+1}(x)\right) & \text{almost everywhere.} \end{cases}$$

By Theorem 9.6, we then have $u^k \rightharpoonup \bar{u}$ in $L^2(\Omega)$. (Since G is not strongly convex, we do not obtain any rates.)

Similarly, we can apply the acceleration strategies from Chapter 12: The *over-relaxed proximal gradient method* for $z^0 = u^0 \in L^2(\Omega)$, $\tau > 0$, and $\lambda = \frac{1}{4}(1 + \sqrt{1 + 8L\tau})$ consists in computing for $k = 0, \dots$

$$\text{(34.9)} \qquad \begin{cases} y^{k+1} = Sz^k & \text{by solving (33.1)}, \\ p^{k+1} = S^*(y^d - y^{k+1}) & \text{by solving (33.2) for } h = y^d - y^{k+1}, \\ u^{k+1}(x) = \operatorname{prox}_{(\tau\alpha)g}\left(u^k(x) + \tau p^{k+1}(x)\right) & \text{almost everywhere,} \\ z^{k+1} = \lambda^{-1}u^{k+1} - (\lambda^{-1} - 1)z^k. \end{cases}$$

By Theorem 12.4, we obtain the convergence of the function values $J(\tilde{u}^N) \to J(\bar{u})$ at the rate $O(1/N)$ as $N \to \infty$ for the ergodic sequence $u^N := \frac{1}{N}\sum_{k=0}^{N} u^{k+1}$.

The *inertial proximal gradient method* for $z^0 = u^0 \in L^2(\Omega)$, $\tau > 0$, and $\lambda_0 = 1$ consists in

computing for $k = 0, \dots$

$$
\begin{cases}
y^{k+1} = Sz^k & \text{by solving (33.1)},\\
p^{k+1} = S^*(y^d - y^{k+1}) & \text{by solving (33.2) for } h = y^d - y^{k+1},\\
u^{k+1}(x) = \operatorname{prox}_{(\tau\alpha)g}\left(u^k(x) + \tau p^{k+1}(x)\right) & \text{almost everywhere},\\
z^{k+1} = (1+\beta_{k+1})u^{k+1} - \beta_{k+1}u^k. &
\end{cases} \tag{34.10}
$$

By Theorem 12.12, we obtain the convergence of the function values $J(\tilde{u}^k) \to J(\bar{u})$ at the rate $O(1/k^2)$ as $k \to \infty$ (for the nonergodic sequence).

Note that in all these algorithms, the number m of desired values only enters (linearly!) through the case distinction in (34.7) for the proximal point mapping. In particular, the cost of each step – which in practice is dominated by computing the solutions y^{k+1} and p^{k+1} of the state and adjoint equation, respectively – is only mildly affected by m. The convex relaxation thus avoids the combinatorial complexity of classical (e.g., branch-and-bound) approaches to mixed-integer optimization.

34.3.2 SEMISMOOTH NEWTON METHOD

The starting point for applying a semismooth Newton method is again the Moreau–Yosida regularization of (34.3), i.e., replacing the set-valued subdifferential ∂G^* by its single-valued Yosida approximation

$$(\partial G^*)_\gamma = \frac{1}{\gamma}\left(\mathrm{Id} - \operatorname{prox}_{\gamma G^*}\right)$$

for some $\gamma > 0$. Again, we can exploit Corollary 6.27 for carrying out the computation pointwise. By Lemma 6.24 (ii), we have that

$$
\begin{aligned}
\operatorname{prox}_{\gamma g^*}(t) &= t - \gamma \operatorname{prox}_{\gamma^{-1}g}(\tfrac{1}{\gamma}t)\\
&= \begin{cases} t - \gamma u_i & \text{in case (i)},\\ t - \gamma\left(\frac{1}{\gamma}t - \frac{1}{2\gamma}(u_i + u_{i+1})\right) = \frac{1}{2}(u_i + u_{i+1}) & \text{in case (ii)}, \end{cases}
\end{aligned}
$$

where case (i) corresponds to

$$\tfrac{1}{\gamma}t \in \left[1 + \tfrac{1}{2\gamma}u_i + \tfrac{1}{2\gamma}u_{i-1}, (1+\tfrac{1}{2\gamma})u_i + \tfrac{1}{2\gamma}u_{i+1}\right],$$

i.e.,

$$t \in \left[\gamma u_i + \frac{1}{2}(u_{i-1} + u_i), \gamma u_i + \frac{1}{2}(u_i + u_{i+1})\right];$$

and case (ii) corresponds to

$$\tfrac{1}{\gamma}t \in \left(1 + \tfrac{1}{2\gamma}u_{i-1} + \tfrac{1}{2\gamma}u_i, (1+\tfrac{1}{2\gamma})u_i + \tfrac{1}{2\gamma}u_{i-1}\right),$$

i.e.,

$$t \in \left(\gamma u_{i-1} + \frac{1}{2}(u_{i-1} + u_i), \gamma u_i + \frac{1}{2}(u_{i-1} + u_i)\right);$$

again with the convention $u_0 = -\infty$ and $u_{m+1} = \infty$. Hence

$$H_\gamma(p) \coloneqq (\partial G^*)_\gamma \left(\tfrac{1}{\alpha} p\right)$$

is given pointwise almost everywhere by

$$[H_\gamma(p)](x) = h_\gamma(p(x)) \coloneqq \begin{cases} u_i & \text{if } p(x) \in Q_i^\gamma, \\ \frac{1}{\alpha\gamma}\left(p(x) - \frac{\alpha}{2}(u_{i-1} + u_i)\right) & \text{if } p(x) \in Q_{i,i+1}^\gamma, \end{cases} \tag{34.11}$$

for

$$Q_i^\gamma \coloneqq \left[\alpha\gamma u_i + \frac{\alpha}{2}(u_{i-1} + u_i), \alpha\gamma u_i + \frac{\alpha}{2}(u_i + u_{i+1})\right],$$
$$Q_{i,i+1}^\gamma \coloneqq \left(\alpha\gamma u_i + \frac{\alpha}{2}(u_i + u_{i+1}), \alpha\gamma u_{i+1} + \frac{\alpha}{2}(u_i + u_{i+1})\right).$$

We illustrate H_γ in Figure 34.4. Replacing $\partial G^*(\frac{1}{\alpha}\cdot)$ by H_γ in (34.3) leads to the regularized optimality conditions

$$\begin{cases} -p_\gamma = S^*(Su_\gamma - y^d), \\ u_\gamma = H_\gamma(p_\gamma). \end{cases} \tag{34.12}$$

Comparing this system with the expansion (34.6) of (34.3), we see that the general structure – in particular, the fact that $u_\gamma(x) = [H_\gamma(p_\gamma)](x) \in \{u_1, \dots, u_m\}$ in the first case – is conserved; the main difference is that the set-valued second case at a point has been replaced by an affine function (with slope $\frac{1}{\gamma}$) in an interval, for which the case distinctions have been adjusted to make room. (This relates to the fact that by Theorem 7.11, the Moreau–Yosida regularization (34.12) is equivalent to replacing G in (34.2) by $G + \frac{\gamma}{2}\|\cdot\|_{L^2}^2$, i.e., the regularized problem still has the original nonsmooth structure and has merely been made *strongly* convex.) Comparing (34.11) and (34.6), it is straightforward to verify that a solution satisfying $u_\gamma(x) \in \{u_1, \dots, u_m\}$ for almost every $x \in \Omega$ also satisfies the unregularized optimality conditions (34.6) and is therefore optimal for (34.2) as well; in this sense, the Moreau–Yosida regularization is an *exact (dual) penalization*.

We now derive the semismooth Newton iteration for solving (34.12). First, it is again advantageous to reformulate the system using the definition of S and S^* as well as the second equation to

$$\begin{cases} -\Delta p_\gamma + y_\gamma - y^d = 0, \\ -\Delta y_\gamma - H_\gamma(p_\gamma) = 0, \end{cases} \tag{34.13}$$

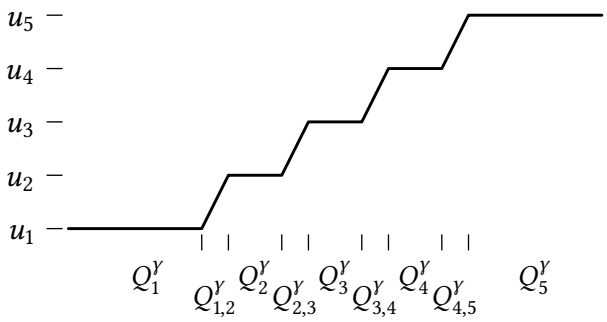


Figure 34.4: Plot of h_γ given by (34.11) for $u_1, \dots, u_5 = -1, -0.5, 0, 0.5, 1$ and $\gamma = 0.5$ and $\alpha = 1$.

cf. (33.21), which we can consider as a nonlinear equation $T(y,p) = 0$ for $T : H_0^1(\Omega) \times H_0^1(\Omega) \to H_0^1(\Omega)^* \times H_0^1(\Omega)^*$. (The corresponding optimal control can be recovered from its solution via $u_\gamma = H_\gamma(p_\gamma)$, which is a simple pointwise evaluation.)

To obtain a Newton derivative $D_N T(y,p)$, we clearly only need to compute one for H_γ, which we again do pointwise. First, it is straightforward to verify that h_γ is continuous and piecewise linear, so that by Theorems 14.8 and 14.9 we have that

$$D_N h_\gamma(t) := \begin{cases} \frac{1}{\alpha\gamma} & \text{if } t \in Q_{i,i+1}^\gamma, \\ 0 & \text{else,} \end{cases} \tag{34.14}$$

is a Newton derivative for h_γ at t. Clearly, this function is uniformly bounded by $\frac{1}{\alpha\gamma}$. For fixed $\gamma > 0$, the intervals $Q_{i,i+1}^\gamma$ are also separated, and hence $D_N h_\gamma$ is a Baire–Caratheodory function. Since $p_\gamma \in H_0^1(\Omega) \hookrightarrow L^r(\Omega)$ for some $r > 2$, it thus follows from Theorem 14.11 that a Newton derivative of H_γ at p in direction $\delta p \in L^r(\Omega)$ is given pointwise almost everywhere by

$$[D_N H_\gamma(p)\delta p](x) = \begin{cases} \frac{1}{\alpha\gamma}\delta p(x) & \text{if } p(x) \in Q_{i,i+1}^\gamma, \\ 0 & \text{else.} \end{cases}$$

Setting $Q^\gamma := \bigcup_{i=1}^m Q_{i,i+1}^\gamma$, we thus obtain as a Newton derivative for T at $(y,p) \in H_0^1(\Omega) \times H_0^1(\Omega)$

$$D_N T(y,p) = \begin{pmatrix} \mathrm{Id} & -\Delta \\ -\Delta & -\frac{1}{\alpha\gamma}\mathbb{1}_{Q^\gamma}(p) \end{pmatrix}.$$

This is a self-adjoint operator that can be shown to be uniformly (with respect to p) boundedly invertible; see [Clason and Kunisch, 2014, Proposition 4.3]. Hence by Theorem 14.1, the following semismooth Newton method converges locally superlinearly to a solution to (34.13): Given $(p^k, y^k) \in H_0^1(\Omega) \times H_0^1(\Omega)$,

(i) solve for $(\delta p, \delta y) \in H_0^1(\Omega) \times H_0^1(\Omega)$ the coupled linear system

$$\begin{aligned} -\Delta\delta p + \delta y &= y^d - y^k + \Delta p^k, \\ -\Delta\delta y - \frac{1}{\alpha\gamma}\mathbb{1}_{Q^\gamma}(p^k)\delta p &= H_\gamma(p^k) + \Delta y^k, \end{aligned}$$

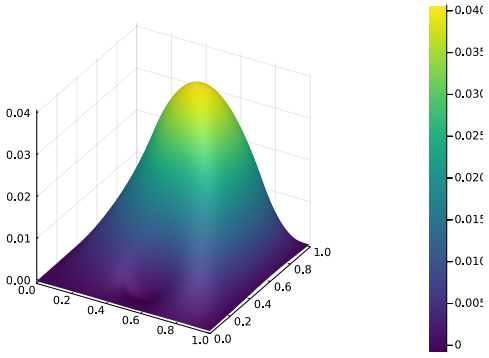
(ii) set

$$y^{k+1} = y^k + \delta y, \qquad p^{k+1} = p^k + \delta p.$$

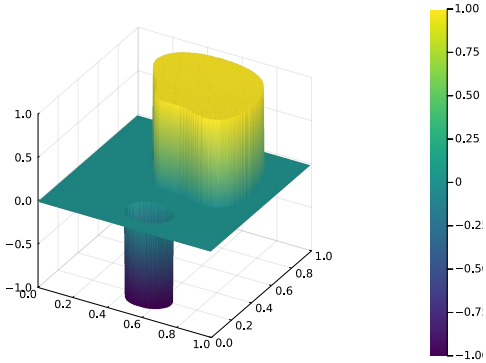
Using the linearity of the state equation and comparing (34.11) with (34.14), this can again be reformulated as a linear system for (p^{k+1}, y^{k+1}). Similarly to the proximal gradient methods, the number m of desired states only enters linearly via the case distinction in Q^γ. In particular, the computation of the Newton step itself is independent of the value of m, hence avoiding combinatorial complexity. As in Section 33.2, this will in practice be embedded in a continuation strategy for $\gamma \to 0$.

We indicate the dependence of solutions to the model discrete-valued control problem (34.2) on the set of allowed values $u_1, \dots, u_m$, by taking $m = 3, 10, 20$ equally spaced controls on $[a, b] = [-1, 1]$. Other than this restriction on the values of the control u, the experimental setup is the same as for control constraints in Section 33.1. For the first-order methods, we take the step length parameter $\tau = 0.9/L^2$, where L is an estimate of $\|S\|$. For the SSN method, the Moreau–Yosida regularization parameter is set to $\gamma = 10^{-6}$. The corresponding optimal controls u_γ for this value of γ are verified to only take on admissible values almost everywhere and thus are also optimal for (34.2); see Figure 34.5. Note also how the discrete-valued controls better approximate the solution shown in Figure 33.1 as m increases. Regarding performance, the SSN method converges to near machine precision within 15 iterations. The forward-backward splitting methods are significantly slower both in number of iterations and actual runtime; see Figure 34.6. In fact, for the shown parameters, all three methods will yield a control taking only allowed values only after 5000 iterations, at which the residual norm drops to machine precision similarly to the SSN method.

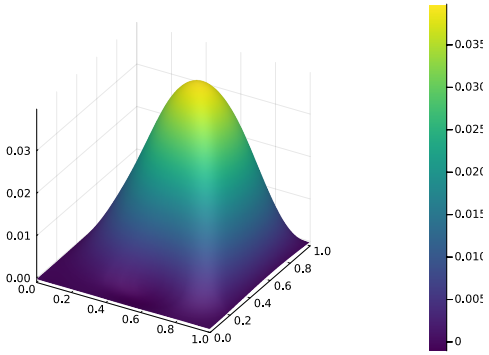
Remark 34.4. The convex relaxation described in this chapter was first proposed in [Clason and Kunisch, 2014] (corresponding to the formal limit $\beta \to \infty$ there) and later applied to topology optimization [Clason and Kunisch, 2016; Clason et al., 2021a] and parameter identification [Clason and Do, 2018] problems. Vector-valued problems were considered in [Clason et al., 2021b] and [Clason et al., 2016], the latter treating the related problem of "switching controls", where at most one of a pair (u, v) of distributed controls should be active at any point, i.e., $u(x)v(x) = 0$ should hold pointwise almost everywhere. The presentation here is condensed from [Clason and Do, 2018; Clason and Kunisch, 2016; Clason et al., 2016, 2021a].



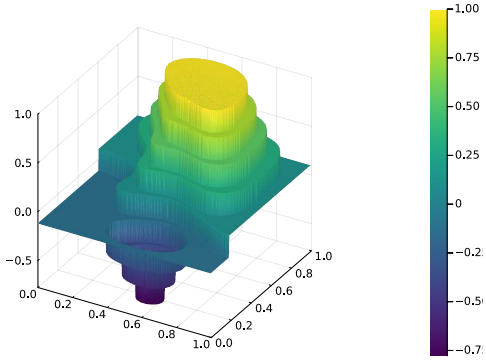
(a) optimal state y, $m = 3$



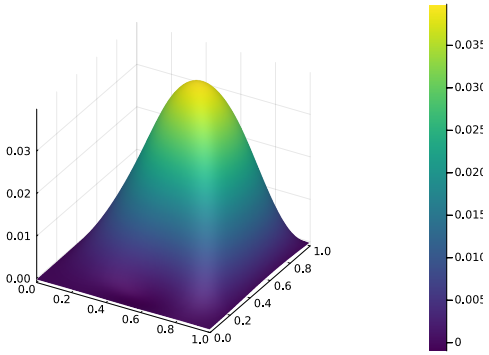
(b) optimal control u, $m = 3$



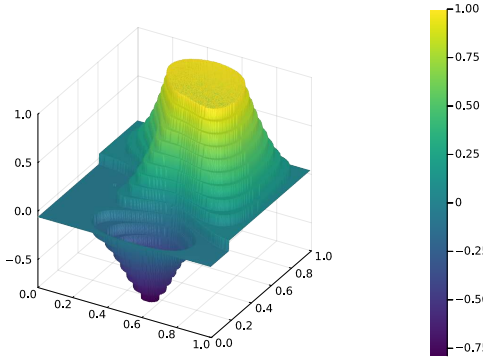
(c) optimal state y, $m = 10$



(d) optimal control u, $m = 10$



(e) optimal state y, $m = 20$



(f) optimal control u, $m = 20$

Figure 34.5: Discrete control example control and state for $\gamma = 10^{-6}$ and $m = 3, 10, 20$. The target y^d is the same as for control constraints in Figure 33.1 and state constraints in Figure 33.4.

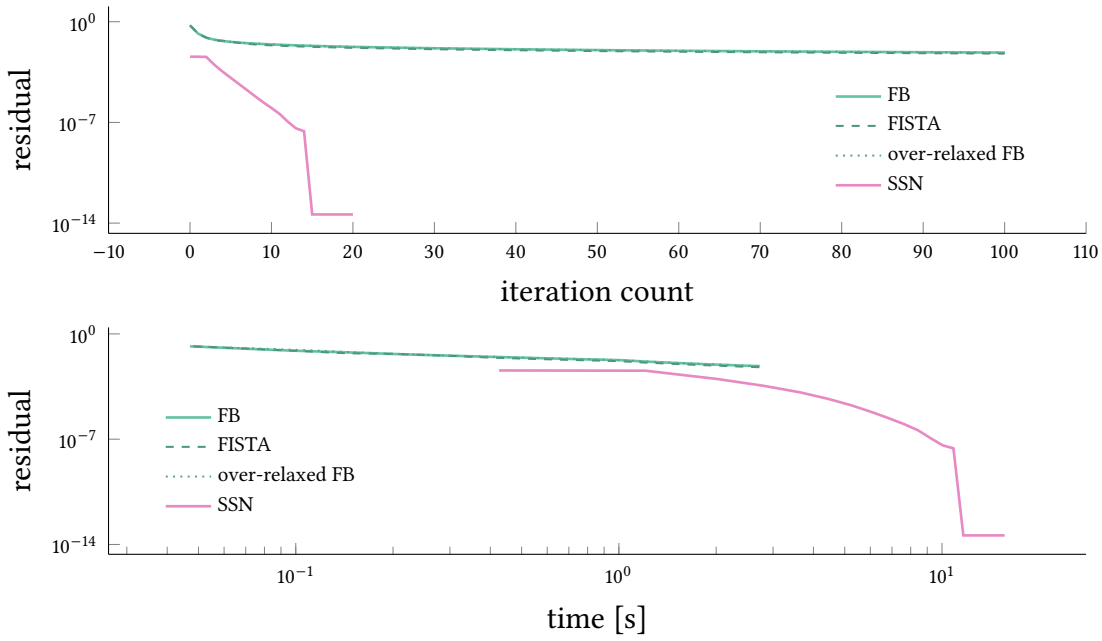


Figure 34.6: Algorithm performance for the discrete control problem, $m = 10$, $\gamma = 10^{-6}$. For the SSN method, we plot the residual $\|H_\gamma(p^k)\|_2$, while for the first-order methods, with $\gamma = 0$, the residual is similarly given by the violation of (34.6).

Bibliography

S. Adly, R. Cibulka, and H. V. Ngai. Newton's method for solving inclusions using set-valued approximations. *SIAM Journal on Optimization*, 25(1):159–184, 2015. doi:10.1137/130926730.

H. W. Alt. *Linear Functional Analysis*. Universitext. Springer, 2016. doi:10.1007/978-1-4471-7280-2.

L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, New York, 2000. doi:10.1007/978-3-0348-8974-2_2.

J. Appell and P. P. Zabrejko. *Nonlinear Superposition Operators*. Cambridge University Press, 1990. doi:10.1017/cbo9780511897450.

F. J. Aragón Artacho and M. Gaydu. A Lyusternik–Graves theorem for the proximal point method. *Computational Optimization and Applications*, 52(3):785–803, 2012. doi:10.1007/s10589-011-9439-6.

F. J. Aragón Artacho and M. H. Geoffroy. Metric subregularity of the convex subdifferential in Banach spaces. *Journal of Nonlinear and Convex Analysis*, 15(1):35–47, 2014.

K. J. Arrow, L. Hurwicz, and H. Uzawa. *Studies in Linear and Non-Linear Programming*. Stanford University Press, 1958.

T. Aspelmeier, C. Charitha, and D. R. Luke. Local linear convergence of the ADMM/Douglas–Rachford algorithms without strong convexity and application to statistical imaging. *SIAM Journal on Imaging Sciences*, 9(2):842–868, 2016. doi:10.1137/15m103580x.

H. Attouch and H. Brezis. Duality for the sum of convex functions in general Banach spaces. In *Aspects of Mathematics and its Applications*, volume 34 of *North-Holland Math. Library*, pages 125–133. North-Holland, Amsterdam, 1986. doi:10.1016/s0924-6509(09)70252-1.

H. Attouch, G. Buttazzo, and G. Michaille. *Variational Analysis in Sobolev and BV Spaces*, volume 6 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics, 2 edition, 2014. doi:10.1137/1.9781611973488.

J. Aubin and H. Frankowska. *Set-Valued Analysis*. Birkhäuser Basel, 1990. doi:10.1007/978-0-8176-4848-0.

J.-P. Aubin. Contingent derivatives of set-valued maps and existence of solutions to nonlinear inclusions and differential inclusions. In *Mathematical Analysis and Applications, Part A*, volume 7 of *Adv. in Math. Suppl. Stud.*, pages 159–229. Academic Press, New York-London, 1981.

J.-P. Aubin. Lipschitz behavior of solutions to convex minimization problems. *Mathematics of Operations Research*, 9(1):87–111, 1984. doi:10.1287/moor.9.1.87.

D. Aussel, A. Daniilidis, and L. Thibault. Subsmooth sets: functional characterizations and related concepts. *Transactions of the American Mathematical Society*, 357(4):1275–1301, 2005. doi:10.1090/s0002-9947-04-03718-3.

D. Azé and J.-P. Penot. Uniformly convex and uniformly smooth convex functions. *Annales de la Faculté des sciences de Toulouse: Mathématiques*, 4(4):705–730, 1995. http://eudml.org/doc/73364.

A. Bagirov, N. Karmitsa, and M. M. Mäkelä. *Introduction to Nonsmooth Optimization*. Springer, Cham, 2014. doi:10.1007/978-3-319-08114-4. Theory, practice and software.

V. Barbu and T. Precupanu. *Convexity and Optimization in Banach Spaces*. Springer Monographs in Mathematics. Springer, Dordrecht, 4 edition, 2012. doi:10.1007/978-94-007-2247-7.

H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, 2 edition, 2017. doi:10.1007/978-3-319-48311-5.

H. H. Bauschke and W. M. Moursi. *An Introduction to Convexity, Optimization, and Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2023. doi:10.1137/1.9781611977806.

M. Bačák and U. Kohlenbach. On proximal mappings with Young functions in uniformly convex Banach spaces. *Journal of Convex Analysis*, 25(4):1291–1318, 2018.

A. Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, 2017. doi:10.1137/1.9781611974997.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009a. doi:10.1137/080716542.

A. Beck and M. Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing*, 18(11):2419–2434, 2009b. doi:10.1109/tip.2009.2028250.

M. Benning, F. Knoll, C.-B. Schönlieb, and T. Valkonen. Preconditioned ADMM with nonlinear operator constraint. In L. Bociu, J.-A. Désidéri, and A. Habbal, editors, *System Modeling and Optimization: 27th IFIP TC 7 Conference, CSMO 2015, Sophia Antipolis, France, June 29–July 3, 2015, Revised Selected Papers*, pages 117–126. Springer International Publishing, 2016. doi:10.1007/978-3-319-55795-3_10. http://tuomov.iki.fi/m/nonlinearADMM.pdf.

M.-F. Bidaut. Un problème de contrôle optimal à fonction coût en norme L^1. *C. R. Acad. Sci. Paris Sér. A-B*, 281(9):A273–A276, 1975.

F. Bigolin and S. N. Golo. A historical account on characterizations of C^1-manifolds in Euclidean spaces by tangent cones. *Journal of Mathematical Analysis and Applications*, 412(1):63–76, 2014. doi:10.1016/j.jmaa.2013.10.035. arXiv:1003.1332.

J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2): 471–507, 2017. doi:10.1007/s10107-016-1091-6.

J. M. Borwein and D. Preiss. A smooth variational principle with applications to subdifferentiability and to differentiability of convex functions. *Transactions of the American Mathematical Society*, 303:517–527, 1987. doi:10.2307/2000681.

J. M. Borwein and Q. J. Zhu. *Techniques of Variational Analysis*, volume 20 of *CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC*. Springer-Verlag, New York, 2005. doi:10.1007/0-387-28271-8.

G. Bouligand. Sur quelques points de méthodologie géométrique. *Rev. Gén. des Sciences*, 41: 39–43, 1930.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004. doi:10.1017/cbo9780511804441.

K. Bredies and D. Lorenz. *Mathematical Image Processing*. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, Cham, 2018. doi:10.1007/978-3-030-01458-2.

K. Bredies and D. A. Lorenz. Linear convergence of iterative soft-thresholding. *Journal of Fourier Analysis and Applications*, 14(5):813–837, 2008. doi:10.1007/s00041-008-9041-1.

K. Bredies and H. Sun. Accelerated Douglas–Rachford methods for the solution of convex-concave saddle-point problems, 2016. Preprint.

K. Bredies, E. Chenchene, D. A. Lorenz, and E. Naldi. Degenerate preconditioned proximal point algorithms. *SIAM Journal on Optimization*, 32(3):2376–2401, 2022. doi:10.1137/21m1448112.

H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, 2010. doi:10.1007/978-0-387-70914-7.

H. Brezis, M. G. Crandall, and A. Pazy. Perturbations of nonlinear maximal monotone sets in Banach space. *Communications on Pure and Applied Mathematics*, 23:123–144, 1970. doi:10.1002/cpa.3160230107.

M. Brokate. Konvexe analysis und evolutionsprobleme. Zentrum Mathematik, TU München, 2014. http://www-m6.ma.tum.de/~brokate/cev_ss14.pdf.

F. E. Browder. Nonexpansive nonlinear operators in a banach space. *Proceedings of the National Academy of Sciences of the United States of America*, 54(4):1041, 1965. doi:10.1073/pnas.54.4.1041.

F. E. Browder. Convergence theorems for sequences of nonlinear operators in Banach spaces. *Mathematische Zeitschrift*, 100(3):201–225, 1967. doi:10.1007/bf01109805.

E. Casas, C. Clason, and K. Kunisch. Approximation of elliptic control problems in measure spaces with sparse solutions. *SIAM Journal on Control and Optimization*, 50(4):1735–1752, 2012. doi:10.1137/110843216.

A. Cegielski. *Iterative methods for fixed point problems in Hilbert spaces*, volume 2057 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2012. doi:10.1007/978-3-642-30901-4.

A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011. doi:10.1007/s10851-010-0251-1.

A. Chambolle and T. Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, pages 1–35, 2015. doi:10.1007/s10107-015-0957-3.

A. Chambolle, R. A. DeVore, N.-y. Lee, and B. J. Lucier. Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage. *IEEE Transactions on Image Processing*, 7(3):319–335, 1998. doi:10.1109/83.661182.

A. Chambolle, M. Ehrhardt, P. Richtárik, and C. Schönlieb. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *SIAM Journal on Optimization*, 28(4):2783–2808, 2018. doi:10.1137/17m1134834.

P. Chen, J. Huang, and X. Zhang. A primal-dual fixed point algorithm for convex separable minimization with applications to image restoration. *Inverse Problems*, 29(2):025011, 2013. doi:10.1088/0266-5611/29/2/025011.

X. Chen, Z. Nashed, and L. Qi. Smoothing methods and semismooth methods for nondifferentiable operator equations. *SIAM Journal on Numerical Analysis*, 38(4):1200–1216, 2000. doi:10.1137/s0036142999356719.

C. Christof and G. Wachsmuth. No-gap second-order conditions via a directional curvature functional. *SIAM Journal on Optimization*, 28(3):2097–2130, 2018. doi:10.1137/17m1140418.

C. Christof, C. Clason, C. Meyer, and S. Walter. Optimal control of a non-smooth semilinear elliptic equation. *Mathematical Control and Related Fields*, 8(1):247–276, 2018. doi:10.3934/mcrf.2018011.

R. Cibulka, A. Dontchev, and A. Kruger. Strong metric subregularity of mappings in variational analysis and optimization. *Journal of Mathematical Analysis and Applications*, 457(2):1247–1282, 2018. doi:10.1016/j.jmaa.2016.11.045. Special Issue on Convex Analysis and Optimization: New Trends in Theory and Applications.

I. Cioranescu. *Geometry of Banach Spaces, Duality Mappings and Nonlinear Problems*, volume 62 of *Mathematics and Its Applications*. Springer, 1990. doi:10.1007/978-94-009-2121-4.

F. Clarke. *Functional Analysis, Calculus of Variations and Optimal Control*. Springer, 2013. doi:10.1007/978-1-4471-4820-3.

F. H. Clarke. Necessary conditions for nonsmooth problems in optimal control and the calculus of variations, 1973.

F. H. Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975. doi:10.1090/s0002-9947-1975-0367131-6.

F. H. Clarke. *Optimization and Nonsmooth Analysis*, volume 5 of *Classics Appl. Math.* Society for Industrial and Applied Mathematics, 1990. doi:10.1137/1.9781611971309.

C. Clason. *Introduction to Functional Analysis*. Compact Textbooks in Mathematics. Springer International Publishing, Basel, 2020a. doi:10.1007/978-3-030-52784-6.

C. Clason. Regularization of Inverse Problems, 2020b. arXiv:2001.00617.

C. Clason and T. B. T. Do. Convex regularization of discrete-valued inverse problems. In B. Hofmann, A. Leitão, and J. Zubelli, editors, *New Trends in Parameter Identification for Mathematical Models*, Trends in Mathematics, pages 31–51. Springer, 2018. doi:10.1007/978-3-319-70824-9_2. arXiv:1707.01041.

C. Clason and K. Kunisch. A duality-based approach to elliptic control problems in non-reflexive Banach spaces. *ESAIM: Control, Optimisation and Calculus of Variations*, 17(1): 243–266, 2011. doi:10.1051/cocv/2010003.

C. Clason and K. Kunisch. Multi-bang control of elliptic systems. *Annales de l'Institut Henri Poincaré (C) Analyse Non Linéaire*, 31(6):1109–1130, 2014. doi:10.1016/j.anihpc.2013.08.005.

C. Clason and K. Kunisch. A convex analysis approach to multi-material topology optimization. *ESAIM: Mathematical Modelling and Numerical Analysis*, 50(6):1917–1936, 2016. doi:10.1051/m2an/2016012. arXiv:1702.07525.

C. Clason and A. Schiela. Optimal control of elliptic equations with positive measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 23(1):217–240, 2017. doi:10.1051/cocv/2015046. arXiv:1702.07528.

C. Clason and T. Valkonen. Primal-dual extragradient methods for nonlinear nonsmooth PDE-constrained optimization. *SIAM Journal on Optimization*, 27(3):1314–1339, 2017a. doi:10.1137/16m1080859.

C. Clason and T. Valkonen. Stability of saddle points via explicit coderivatives of pointwise subdifferentials. *Set-Valued and Variational Analysis*, 25:69–112, 2017b. doi:10.1007/s11228-016-0366-7. http://tuomov.iki.fi/m/pdex2_stability.pdf.

C. Clason and T. Valkonen. Introduction to nonsmooth analysis and optimization: example algorithm implementations. Software on Zenodo, 2023. To upload!

C. Clason, B. Jin, and K. Kunisch. A semismooth Newton method for L^1 data fitting with automatic choice of regularization parameters and noise calibration. *SIAM Journal on Imaging Sciences*, 3(2):199–231, 2010. doi:10.1137/090758003.

C. Clason, K. Ito, and K. Kunisch. A convex analysis approach to optimal controls with switching structure for partial differential equations. *ESAIM: Control, Optimisation and Calculus of Variations*, 22(2):581–609, 2016. doi:10.1051/cocv/2015017. arXiv:1702.07540.

C. Clason, S. Mazurenko, and T. Valkonen. Acceleration and global convergence of a first-order primal–dual method for nonconvex problems. *SIAM Journal on Optimization*, 29:933–963, 2019. doi:10.1137/18m1170194.

C. Clason, S. Mazurenko, and T. Valkonen. Primal–dual proximal splitting and generalized conjugation in non-smooth non-convex optimization. *Applied Mathematics & Optimization*, 84(2):1239–1284, 2020. doi:10.1007/s00245-020-09676-1. http://tuomov.iki.fi/m/nlpdhgm_general.pdf.

C. Clason, K. Kunisch, and P. Trautmann. Optimal control of the principal coefficient in a scalar wave equation. *Applied Mathematics & Optimization*, 84(3):2889–2921, 2021a. doi:10.1007/s00245-020-09733-9. arXiv:1912.08672.

C. Clason, C. Tameling, and B. Wirth. Convex relaxation of discrete vector-valued optimization problems. *SIAM Review*, 63(4):783–821, 2021b. doi:10.1137/21m1426237. arXiv:2108.10077.

P. L. Combettes and N. N. Reyes. Moreau's decomposition in Banach spaces. *Mathematical Programming*, 139(1):103–114, 2013. doi:10.1007/s10107-013-0663-y.

L. Condat. A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 158(2):460–479, 2013. doi:10.1007/s10957-012-0245-9.

Y. Cui and J.-S. Pang. *Modern Nonconvex Nondifferentiable Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2021. doi:10.1137/1.9781611976748.

I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004. doi:10.1002/cpa.20042.

D. Davis and W. Yin. A three-operator splitting scheme and its optimization applications. *Set-Valued and Variational Analysis*, 25(4):829–858, 2017. ISSN 1877-0541. doi:10.1007/s11228-017-0421-z.

J. C. De los Reyes. *Numerical PDE-Constrained Optimization*. Springer, 2015. doi:10.1007/978-3-319-13395-9.

E. DiBenedetto. *Real Analysis*. Birkhäuser Boston, Inc., Boston, MA, 2002. doi:10.1007/978-1-4612-0117-5.

S. Dolecki and G. H. Greco. Tangency vis-à-vis differentiability by Peano, Severi and Guareschi. *Journal of Convex Analysis*, 18(2):301–339, 2011. arXiv:1003.1332.

A. L. Dontchev. *Lectures on variational analysis*, volume 205 of *Appl. Math. Sci.* Cham: Springer, 2021. doi:10.1007/978-3-030-79911-3.

A. L. Dontchev and R. T. Rockafellar. Regularity and conditioning of solution mappings in variational analysis. *Set-Valued and Variational Analysis*, 12(1-2):79–109, 2004. doi:10.1023/b:svan.0000023394.19482.30.

A. L. Dontchev and R. T. Rockafellar. *Implicit Functions and Solution Mappings*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2 edition, 2014. doi:10.1007/978-1-4939-1037-3.

J. Douglas, Jim and J. Rachford, H. H. On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the American Mathematical Society*, 82(2):421–439, 1956. doi:10.2307/1993056.

Y. Drori, S. Sabach, and M. Teboulle. A simple algorithm for a class of nonsmooth convex–concave saddle-point problems. *Operations Research Letters*, 43(2):209–214, 2015. doi:10.1016/j.orl.2015.02.001.

J. Eckstein and D. P. Bertsekas. On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318, 1992. doi:10.1007/bf01581204.

I. Ekeland and G. Lebourg. Generic Fréchet-differentiability and perturbed optimization problems in Banach spaces. *Transactions of the American Mathematical Society*, 224(2):193–216, 1976. doi:10.1090/s0002-9947-1976-0431253-2.

I. Ekeland and R. Témam. *Convex Analysis and Variational Problems*, volume 28 of *Classics Appl. Math.* Society for Industrial and Applied Mathematics, 1999. doi:10.1137/1.9781611971088.

H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*, volume 375 of *Mathematics and its Applications*. Springer Netherlands, 1996. doi:10.1007/978-94-009-1740-8.

E. Esser, X. Zhang, and T. F. Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3(4):1015–1046, 2010. doi:10.1137/09076934x.

M. Fabian. On classes of subdifferentiability spaces of Ioffe. *Nonlinear Analysis: Theory, Methods & Applications*, 12(1):63–74, 1988. doi:10.1016/0362-546x(88)90013-2.

M. Fabian, P. Habala, P. Hájek, V. M. Santalucía, J. Pelant, and V. Zizler. *Differentiability of norms*, pages 241–284. Springer New York, New York, NY, 2001. doi:10.1007/978-1-4757-3480-5_8.

F. Facchinei and J.-S. Pang. *Finite-dimensional Variational Inequalities and Complementarity Problems. Vol. I.* Springer Series in Operations Research. Springer-Verlag, New York, 2003a. doi:10.1007/b97543.

F. Facchinei and J.-S. Pang. *Finite-dimensional Variational Inequalities and Complementarity Problems. Vol. II.* Springer Series in Operations Research. Springer-Verlag, New York, 2003b. doi:10.1007/b97544.

L. Fejér. Über die Lage der Nullstellen von Polynomen, die aus Minimumforderungen gewisser Art entspringen. *Mathematische Annalen*, 85(1):41–48, 1922. doi:10.1007/bf01449600.

I. Fonseca and G. Leoni. *Modern Methods in the Calculus of Variations: L^p Spaces*. Springer, 2007. doi:10.1007/978-0-387-69006-3.

D. Gabay. Applications of the method of multipliers to variational inequalities. In M. Fortin and R. Glowinski, editors, *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*, volume 15 of *Studies in Mathematics and its Applications*, pages 299–331. North-Holland, 1983.

G. Garrigos, L. Rosasco, and S. Villa. Thresholding gradient methods in Hilbert spaces: support identification and linear convergence. *ESAIM: Control, Optimisation and Calculus of Variations*, 26:28, 2020. doi:10.1051/cocv/2019011.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, 3 edition, 2013. doi:10.1201/9780429258480. http://www.stat.columbia.edu/~gelman/book/.

gerw. Subgradient in a predual under weak* continuity, 2022. https://mathoverflow.net/q/414752 (visited 2022-01-26).

H. Gferer and J. Outrata. On a semismooth* Newton method for solving generalized equations. *SIAM Journal on Optimization*, 31(1):489–517, 2021. doi:10.1137/19m1257408.

H. Gfrerer. First order and second order characterizations of metric subregularity and calmness of constraint set mappings. *SIAM Journal on Optimization*, 21(4):1439–1474, 2011. doi:10.1137/100813415.

H. Gfrerer. On directional metric regularity, subregularity and optimality conditions for nonsmooth mathematical programs. *Set-Valued and Variational Analysis*, 21(2):151–176, 2013. doi:10.1007/s11228-012-0220-5.

H. Gfrerer and J. Outrata. On Lipschitzian properties of implicit multifunctions. *SIAM Journal on Optimization*, 26(4):2160–2189, 2016. doi:10.1137/15m1052299.

R. Gribonval and M. Nikolova. A characterization of proximity operators. *Journal of Mathematical Imaging and Vision*, 62:773–789, 2020. doi:10.1007/s10851-020-00951-y.

R. Griesse and D. A. Lorenz. A semismooth newton method for Tikhonov functionals with sparsity constraints. *Inverse Problems*, 24(3):035007, 2008. ISSN 1361–6420. doi:10.1088/0266-5611/24/3/035007.

P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. Society for Industrial and Applied Mathematics, 2011. doi:10.1137/1.9781611972030.

D. Han and X. Yuan. Local linear convergence of the alternating direction method of multipliers for quadratic programs. *SIAM Journal on Numerical Analysis*, 51(6):3446–3457, 2013. doi:10.1137/120886753.

M. Hanke. *A Taste of Inverse Problems: Basic Theory and Examples*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics, 2017. ISBN 9781611974942. doi:10.1137/1.9781611974942.

F. Harder and G. Wachsmuth. The limiting normal cone of a complementarity set in Sobolev spaces. *Optimization*, 67(5):1579–1603, 2018. doi:10.1080/02331934.2018.1484467.

B. He and X. Yuan. Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective. *SIAM Journal on Imaging Sciences*, 5(1):119–149, 2012. doi:10.1137/100814494.

J. Heinonen. Lectures on Lipschitz analysis, 2005. http://www.math.jyu.fi/research/reports/rep100.pdf.

R. Henrion, A. Jourani, and J. Outrata. On the calmness of a class of multifunctions. *SIAM Journal on Optimization*, 13(2):603–618, 2002. doi:10.1137/s1052623401395553.

M. Hintermüller and K. Kunisch. Total bounded variation regularization as a bilaterally constrained optimization problem. *SIAM Journal on Applied Mathematics*, 64(4):1311–1333, 2004. doi:10.1137/s0036139903422784.

M. Hintermüller and G. Stadler. An infeasible primal-dual algorithm for total bounded variation-based inf-convolution-type image restoration. 28(1):1–23 (electronic), 2006. doi:10.1137/040613263.

M. Hintermüller and M. Ulbrich. A mesh-independence result for semismooth Newton methods. *Math. Program.*, 101(1, Ser. B):151–184, 2004. doi:10.1007/s10107-004-0540-9.

M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM Journal on Optimization*, 13(3):865–888 (2003), 2002. doi:10.1137/s1052623401383558.

M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*. Springer, New York, 2009. doi:10.1007/978-1-4020-8839-1.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I*, volume 305 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, Berlin, 1993a. doi:10.1007/978-3-662-02796-7.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms II*, volume 306 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, Berlin, 1993b. doi:10.1007/978-3-662-06409-2.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer-Verlag, 2001. doi:10.1007/978-3-642-56468-0.

T. Hohage and C. Homann. A generalization of the Chambolle–Pock algorithm to Banach spaces with applications to inverse problems, 2014. Preprint.

P. J. Huber. *Robust Statistics*. John Wiley & Sons Inc., New York, 2 edition, 2009. doi:10.1002/9780470434697.

L. Hörmander. Sur la fonction d'appui des ensembles convexes dans un espace localement convexe. *Arkiv för Matematik*, 3:181–186, 1955. doi:10.1007/bf02589354.

A. D. Ioffe. Regular points of Lipschitz functions. *Transactions of the American Mathematical Society*, 251:61–69, 1979. doi:10.1090/s0002-9947-1979-0531969-6.

A. D. Ioffe. Approximate subdifferentials and applications. I. The finite-dimensional theory. *Transactions of the American Mathematical Society*, 281(1):389–416, 1984. doi:10.2307/1999541.

A. D. Ioffe. *Variational Analysis of Regular Mappings: Theory and Applications*. Springer Monographs in Mathematics. Springer International Publishing, 2017. doi:10.1007/978-3-319-64277-2.

K. Ito and B. Jin. *Inverse Problems*, volume 22 of *Series on Applied Mathematics*. World Scientific, Singapore, 2014. doi:10.1142/9120.

K. Ito and K. Kunisch. *Lagrange Multiplier Approach to Variational Problems and Applications*, volume 15 of *Advances in Design and Control*. Society for Industrial and Applied Mathematics, 2008. doi:10.1137/1.9780898718614.

J. Jauhiainen, P. Kuusela, A. Seppänen, and T. Valkonen. Relaxed Gauss–Newton methods with applications to electrical impedance tomography. *SIAM Journal on Imaging Sciences*, 13(3):1415–1445, 2020. doi:10.1137/20m1321711. arXiv:2002.08044.

B. Kaltenbacher, A. Neubauer, and O. Scherzer. *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*. Number 6 in Radon Series on Computational and Applied Mathematics. De Gruyter, 2008. ISBN 9783110208276.

D. Klatte and B. Kummer. *Nonsmooth Equations in Optimization*, volume 60 of *Nonconvex Optimization and its Applications*. Kluwer Academic Publishers, Dordrecht, 2002. doi:10.1007/b130810. Regularity, calculus, methods and applications.

M. Kojima and S. Shindo. Extension of Newton and quasi-Newton methods to systems of PC^1 equations. *Journal of the Operations Research Society of Japan*, 29(4):352–375, 1986. doi:10.15807/jorsj.29.352.

M. A. Krasnosel′skiĭ. Two remarks on the method of successive approximations. *Uspekhi Matematicheskikh Nauk*, 10(1(63)):123–127, 1955.

A. Y. Kruger. Error bounds and metric subregularity. *Optimization*, 64(1):49–79, 2015. doi:10.1080/02331934.2014.938074.

F. Kruse. Semismooth implicit functions. *Journal of Convex Analysis*, 28(2):595–622, 2018. https://imsc.uni-graz.at/mannel/SIF_Kruse.pdf.

B. Kummer. Newton's method for non-differentiable functions. *Mathematical Research*, 45: 114–125, 1988.

B. Kummer. Generalized Newton and NCP-methods: convergence, regularity, actions. *Discussiones Mathematicae. Differential Inclusions, Control and Optimization*, 20(2):209–244, 2000. doi:10.7151/dmdico.1013.

T. Kärkkäinen, K. Kunisch, and K. Majava. Denoising of smooth images using L^1-fitting. *Computing*, 74(4):353–376, 2005. doi:10.1007/s00607-004-0097-8.

G. Lebourg. Valeur moyenne pour gradient généralisé. *C. R. Acad. Sci. Paris Sér. A-B*, 281 (19):A795–A797, 1975.

G. Lebourg. Generic differentiability of Lipschitzian functions. *Transactions of the American Mathematical Society*, 256:125–144, 1979. doi:10.2307/1998104.

D. Leventhal. Metric subregularity and the proximal point method. *Journal of Mathematical Analysis and Applications*, 360(2):681–688, 2009. doi:10.1016/j.jmaa.2009.07.012.

A. S. Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13 (3):702–725, 2002. doi:10.1137/s1052623401387623.

G. Li. Global error bounds for piecewise convex polynomials. *Mathematical Programming*, 137(1):37–64, 2013. doi:10.1007/s10107-011-0481-z.

G. Li and B. S. Mordukhovich. Hölder metric subregularity with applications to proximal point method. *SIAM Journal on Optimization*, 22(4):1655–1684, 2012. doi:10.1137/120864660.

J. Liang, J. Fadili, and G. Peyré. Local linear convergence of forward–backward under partial smoothness. *Advances in Neural Information Processing Systems*, 27:1970–1978, 2014. http://papers.nips.cc/paper/5260-local-linear-convergence-of-forward-backward-under-partial-smoothness.pdf.

J.-L. Lions. *Optimal Control of Systems Governed by Partial Differential Equations*, volume 170 of *Die Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, New York-Berlin, 1971.

P. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979. doi:10.1137/0716071.

Y. Liu, X. Yuan, S. Zeng, and J. Zhang. Partial error bound conditions and the linear convergence rate of the alternating direction method of multipliers. *SIAM Journal on Numerical Analysis*, 56(4):2095–2123, 2018. doi:10.1137/17m1144623.

I. Loris and C. Verhoeven. On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty. *Inverse Problems*, 27(12):125007, 2011. doi:10.1088/0266-5611/27/12/125007.

Z.-Q. Luo and P. Tseng. Error bound and convergence analysis of matrix splitting algorithms for the affine variational inequality problem. *SIAM Journal on Optimization*, 2(1):43–54, 1992. doi:10.1137/0802004.

M. M. Mäkelä and P. Neittaanmäki. *Nonsmooth Optimization*. World Scientific Publishing Co., Inc., River Edge, NJ, 1992. doi:10.1142/1493. Analysis and algorithms with applications to optimal control.

Y. Malitsky and T. Pock. A first-order primal-dual algorithm with linesearch. *SIAM Journal on Optimization*, 28(1):411–432, 2018. doi:10.1137/16m1092015.

Y. Malitsky and M. K. Tam. A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM Journal on Optimization*, 30(2):1451–1472, 2020. doi:10.1137/18m1207260.

W. R. Mann. Mean value methods in iteration. *Proceedings of the American Mathematical Society*, 4:506–510, 1953. doi:10.2307/2032162.

B. Martinet. Régularisation d'inéquations variationnelles par approximations successives. *Rev. Française Informat. Recherche Opérationnelle*, 4(Sér. R-3):154–158, 1970.

S. Mazurenko, J. Jauhiainen, and T. Valkonen. Primal-dual block-proximal splitting for a class of non-convex problems. *Electronic Transactions on Numerical Analysis*, 52:509–552, 2020. doi:10.1553/etna_vol52s509. arXiv:1911.06284.

P. Mehlitz and G. Wachsmuth. The limiting normal cone to pointwise defined sets in Lebesgue spaces. *Set-Valued and Variational Analysis*, 26(3):449–467, 2018. doi:10.1007/s11228-016-0393-4.

P. Mehlitz and G. Wachsmuth. The weak sequential closure of decomposable sets in Lebesgue spaces and its application to variational geometry. *Set-Valued and Variational Analysis*, 27(1):265–294, 2019. doi:10.1007/s11228-017-0464-1.

C. Meyer, L. Panizzi, and A. Schiela. Uniqueness criteria for solutions of the adjoint equation in state-constrained optimal control. *Numerical Functional Analysis and Optimization*, 32(9):983–1007, 2011. doi:10.1080/01630563.2011.587074.

R. Mifflin. Semismooth and semiconvex functions in constrained optimization. *SIAM Journal on Control and Optimization*, 15(6):959–972, 1977. doi:10.1137/0315061.

B. v. Morduhovič. Metric approximations and necessary conditions for optimality for general classes of nonsmooth extremal problems. *Doklady Akademii Nauk SSSR*, 254(5): 1072–1076, 1980.

B. S. Mordukhovich. Maximum principle in the problem of time optimal response with nonsmooth constraints. *Journal of Mathematical Analysis and Applications*, 40(6):960–969, 1976. doi:10.1016/0021-8928(76)90136-2.

B. S. Mordukhovich. Generalized differential calculus for nonsmooth and set-valued mappings. *Journal of Mathematical Analysis and Applications*, 183(1):250–288, 1994. doi:10.1006/jmaa.1994.1144.

B. S. Mordukhovich. *Variational Analysis and Generalized Differentiation I*, volume 330 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2006. doi:10.1007/3-540-31247-1.

B. S. Mordukhovich. *Variational Analysis and Applications*. Springer Monographs in Mathematics. Springer International Publishing, 2018. doi:10.1007/978-3-319-92775-6.

J.-J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965. doi:10.24033/bsmf.1625.

T. S. Motzkin and I. J. Schoenberg. The relaxation method for linear inequalities. *Canadian Journal of Mathematics*, 6:393–404, 1954. doi:10.4153/cjm-1954-038-x.

J. L. Mueller and S. Siltanen. *Linear and Nonlinear Inverse Problems with Practical Applications*. Society for Industrial and Applied Mathematics, 2012. doi:10.1137/1.9781611972344.

T. Möllenhoff, E. Strekalovskiy, M. Moeller, and D. Cremers. The primal-dual hybrid gradient method for semiconvex splittings. *SIAM Journal on Imaging Sciences*, 8(2):827–857, 2015. doi:10.1137/140976601.

F. Natterer. *The Mathematics of Computerized Tomography*. Society for Industrial and Applied Mathematics, 2001. doi:10.1137/1.9780898719284.

Y. Nesterov. *Introductory Lectures on Convex Optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004. doi:10.1007/978-1-4419-8853-9.

Y. Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, Cham, 2018. doi:10.1007/978-3-319-91578-4.

Y. E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Doklady Akademii Nauk SSSR*, 27(2):372–376, 1983.

H. V. Ngai and M. Théra. Error bounds in metric spaces and application to the perturbation stability of metric regularity. *SIAM Journal on Optimization*, 19(1):1–20, 2008. doi:10.1137/060675721.

D. Nishimura. *Principles of Magnetic Resonance Imaging*. Stanford University, 1996.

P. Ochs and T. Pock. Adaptive FISTA for non-convex optimization. *SIAM Journal on Optimization*, 29(4):2482–2503, 2019. doi:10.1137/17m1156678.

Z. Opial. Weak convergence of the sequence of successive approximations for nonexpansive mappings. *Bulletin of the American Mathematical Society*, 73(4):591–597, 1967. doi:10.1090/s0002-9904-1967-11761-0.

J. Outrata, M. Kočvara, and J. Zowe. *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints*, volume 28 of *Nonconvex Optimization and its Applications*. Kluwer Academic Publishers, Dordrecht, 1998. doi:10.1007/978-1-4757-2825-5. Theory, applications and numerical results.

N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3): 123–231, 2014. doi:10.1561/2400000003.

P. Patrinos, L. Stella, and A. Bemporad. Douglas–Rachford splitting: complexity estimates and accelerated variants. In *53rd IEEE Conference on Decision and Control*, pages 4234–4239, 2014. doi:10.1109/cdc.2014.7040049.

G. Peano. *Formulario Mathematico*. Fratelli Boca, Torino, 1908.

J.-P. Penot. *Calculus Without Derivatives*, volume 266 of *Graduate Texts in Mathematics*. Springer, New York, 2013. doi:10.1007/978-1-4614-4538-8.

W. V. Petryshyn. Construction of fixed points of demicompact mappings in Hilbert space. *Journal of Mathematical Analysis and Applications*, 14(2):276–284, 1966. doi:10.1016/0022-247x(66)90027-8.

J. Peypouquet. *Convex Optimization in Normed Spaces*. SpringerBriefs in Optimization. Springer, Cham, 2015. doi:10.1007/978-3-319-13710-0.

T. Pock, D. Cremers, H. Bischof, and A. Chambolle. An algorithm for minimizing the Mumford–Shah functional. In *12th IEEE Conference on Computer Vision*, pages 1133–1140, 2009. doi:10.1109/iccv.2009.5459348.

R. Poliquin and R. T. Rockafellar. Prox-regular functions in variational analysis. *Transactions of the American Mathematical Society*, 348(5):1805–1838, 1996. doi:10.1090/s0002-9947-96-01544-9.

L. Qi. Convergence analysis of some algorithms for solving nonsmooth equations. *Mathematics of Operations Research*, 18(1):227–244, 1993. doi:10.1287/moor.18.1.227.

L. Qi and J. Sun. A nonsmooth version of Newton's method. *Mathematical Programming*, 58(3, Ser. A):353–367, 1993. doi:10.1007/bf01581275.

M. Renardy and R. C. Rogers. *An Introduction to Partial Differential Equations*, volume 13 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 2 edition, 2004. doi:10.1007/b97427.

S. M. Robinson. Some continuity properties of polyhedral multifunctions. In H. König, B. Korte, and K. Ritter, editors, *Mathematical Programming at Oberwolfach*, pages 206–214. Springer Berlin Heidelberg, Berlin, Heidelberg, 1981. doi:10.1007/bfb0120929.

R. T. Rockafellar. On the maximal monotonicity of subdifferential mappings. *Pacific Journal of Mathematics*, 33:209–216, 1970. doi:10.2140/pjm.1970.33.209.

R. T. Rockafellar. Integral functionals, normal integrands and measurable selections. In *Nonlinear Operators and the Calculus of Variations (Summer School, Univ. Libre Bruxelles, Brussels, 1975)*, volume 543 of *Lecture Notes in Math.*, pages 157–207. Springer, 1976a. doi:10.1007/bfb0079944.

R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976b. doi:10.1137/0314056.

R. T. Rockafellar. Favorable classes of Lipschitz continuous functions in subgradient optimization. In *Progress in Nondifferentiable Optimization*, page 125–143, 1981. http://pure.iiasa.ac.at/id/eprint/1760/.

R. T. Rockafellar. Maximal monotone relations and the second derivatives of nonsmooth functions. *Annales de l'Institut Henri Poincare (C) Non Linear Analysis*, 2(3):167–184, 1985. doi:10.1016/s0294-1449(16)30401-2.

R. T. Rockafellar. First-and second-order epi-differentiability in nonlinear programming. *Transactions of the American Mathematical Society*, 307(1):75–108, 1988. doi:10.1090/s0002-9947-1988-0936806-9.

R. T. Rockafellar. Proto-differentiability of set-valued mappings and its applications in optimization. *Annales de l'Institut Henri Poincare (C) Non Linear Analysis*, 6(S6):449–482, 1989. ISSN 0294-1449. doi:10.1016/s0294-1449(17)30034-3.

R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer, 1998. doi:10.1007/978-3-642-02431-3.

J. O. Royset and R. J.-B. Wets. *An Optimization Primer*. Springer, 2021. doi:10.1007/978-3-030-76275-9.

L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992. doi:10.1016/0167-2789(92)90242-f.

W. Rudin. *Analysis*. De Gruyter, 2 edition, 2021. doi:10.1515/9783110750430.

A. Ruszczyński. *Nonlinear Optimization*. Princeton University Press, 2006. doi:10.2307/j.ctvcm4hcj.

B. P. Rynne and M. A. Youngson. *Linear Functional Analysis*. Springer Undergraduate Mathematics Series. Springer, London, 2 edition, 2008. doi:10.1007/978-1-84800-005-6.

E. K. Ryu. Uniqueness of drs as the 2 operator resolvent-splitting and impossibility of 3 operator resolvent-splitting. *Mathematical Programming*, 182(1–2):233––273, 2019. ISSN 1436-4646. doi:10.1007/s10107-019-01403-1.

H. Schaefer. Über die methode sukzessiver approximationen. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 59:131–140, 1957. http://eudml.org/doc/146424.

O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen. *Variational Methods in Imaging*. Springer, New York, 2009. doi:10.1007/978-0-387-69277-7.

A. Schiela. A simplified approach to semismooth newton methods in function space. *SIAM Journal on Optimization*, 19(3):1417–1432, 2008. doi:10.1137/060674375.

W. Schirotzek. *Nonsmooth Analysis*. Universitext. Springer, 2007. doi:10.1007/978-3-540-71333-3.

S. Scholtes. *Introduction to Piecewise Differentiable Equations*. Springer Briefs in Optimization. Springer, New York, 2012. doi:10.1007/978-1-4614-4340-7.

T. Schuster, B. Kaltenbacher, B. Hofmann, and K. Kazimierski. *Regularization Methods in Banach Spaces*. Radon Series on Computational and Applied Mathematics. De Gruyter, 2012. ISBN 9783110255720.

S. Simons. A new proof of the maximal monotonicity of subdifferentials. *Journal of Convex Analysis*, 16(1):165–168, 2009.

G. Stadler. Elliptic optimal control problems with L^1-control cost and applications for the placement of control devices. 44(2):159–181, 2009. doi:10.1007/s10589-007-9150-9.

L. Thibault. Tangent cones and quasi-interiorly tangent cones to multifunctions. *Transactions of the American Mathematical Society*, 277(2):601–621, 1983. doi:10.2307/1999227.

L. Thibault and D. Zagrodny. Integration of subdifferentials of lower semicontinuous functions on Banach spaces. *Journal of Mathematical Analysis and Applications*, 189(1): 33–58, 1995. doi:10.1006/jmaa.1995.1003.

F. Tröltzsch. *Optimal Control of Partial Differential Equations*, volume 112 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2010. doi:10.1090/gsm/112. Translated from the 2005 German original by Jürgen Sprekels.

M. Ulbrich. Semismooth Newton methods for operator equations in function spaces. *SIAM Journal on Optimization*, 13(3):805–841, 2002. doi:10.1137/s1052623400371569.

M. Ulbrich. *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, volume 11 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics, 2011. doi:10.1137/1.9781611970692.

T. Valkonen. A primal-dual hybrid gradient method for nonlinear operators with applications to MRI. *Inverse Problems*, 30(5):055012, 2014. doi:10.1088/0266-5611/30/5/055012.

T. Valkonen. Block-proximal methods with spatially adapted acceleration. *Electronic Transactions on Numerical Analysis*, 51:15–49, 2019. doi:10.1553/etna_vol51s15. http://tuomov.iki.fi/m/blockcp.pdf.

T. Valkonen. Inertial, corrected, primal-dual proximal splitting. *SIAM Journal on Optimization*, 30(2):1391–1420, 2020a. doi:10.1137/18m1182851. arXiv:1804.08736.

T. Valkonen. Testing and non-linear preconditioning of the proximal point method. *Applied Mathematics & Optimization*, 82(2):591–636, 2020b. doi:10.1007/s00245-018-9541-6. arXiv:1703.05705. http://tuomov.iki.fi/m/proxtest.pdf.

T. Valkonen. First-order primal-dual methods for nonsmooth nonconvex optimisation. In K. Chen, C.-B. Schönlieb, X.-C. Tai, and L. Younes, editors, *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging*. Springer, Cham, 2021a. ISBN 978-3-030-03009-4. doi:10.1007/978-3-030-03009-4_93-1. arXiv:1910.00115.

T. Valkonen. Regularisation, optimisation, subregularity. *Inverse Problems*, 37(4):045010, 2021b. doi:10.1088/1361-6420/abe4aa. arXiv:2011.07575.

T. Valkonen. Preconditioned proximal point methods and notions of partial subregularity. *Journal of Convex Analysis*, 28(1):251–278, 2021c. arXiv:1711.05123. http://tuomov.iki.fi/m/proxtest.pdf.

G. Vossen and H. Maurer. On L^1-minimization in optimal control and applications to robotics. 27(6):301–321, 2006. doi:10.1002/oca.781.

B. C. Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013. doi:10.1007/s10444-011-9254-8.

S. J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015. doi:10.1007/s10107-015-0892-3.

S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009. doi:10.1109/tsp.2009.2016892.

J. Yang, Y. Zhang, and W. Yin. An efficient TVL1 algorithm for deblurring multichannel images corrupted by impulsive noise. *SIAM Journal on Scientific Computation*, 31(4):2842–2865, 2009. doi:10.1137/080732894.

D. Yost. Asplund spaces for beginners. *Acta Universitatis Carolinae. Mathematica et Physica*, 34(2):159–177, 1993. https://dml.cz/dmlcz/702006.

X. Zhang, M. Burger, and S. Osher. A unified primal-dual algorithm framework based on Bregman iteration. *Journal of Scientific Computing*, 46(1):20–46, 2011. doi:10.1007/s10915-010-9408-8.

X. Y. Zheng and K. F. Ng. Metric subregularity and calmness for nonconvex generalized equations in Banach spaces. *SIAM Journal on Optimization*, 20(5):2119–2136, 2010. doi:10.1137/090772174.

Z. Zhou and A. M.-C. So. A unified approach to error bounds for structured convex optimization problems. *Mathematical Programming*, 165(2):689–728, 2017. doi:10.1007/s10107-016-1100-9.

M. Zhu and T. Chan. An efficient primal-dual hybrid gradient algorithm for total variation image restoration. CAM Report 08-34, UCLA, 2008. ftp://ftp.math.ucla.edu/pub/camreport/cam08-34.pdf.

C. Zălinescu. *Convex Analysis in General Vector Spaces*. World Scientific Publishing Co., Inc., River Edge, NJ, 2002. doi:10.1142/9789812777096.

Index