

NONSMOOTH ANALYSIS AND OPTIMIZATION

LECTURE NOTES

Christian Clason

August 15, 2017

christian.clason@uni-due.de
<https://udue.de/clason>

CONTENTS

I BACKGROUND

- 1 FUNCTIONAL ANALYSIS 4
 - 1.1 Normed vector spaces 4
 - 1.2 Strong and weak convergence 6
 - 1.3 Hilbert spaces 12
- 2 CALCULUS OF VARIATIONS 13
 - 2.1 The direct method 13
 - 2.2 Differential calculus in Banach spaces 17
 - 2.3 Superposition operators 20

II CONVEX ANALYSIS

- 3 CONVEX FUNCTIONS 24
- 4 CONVEX SUBDIFFERENTIALS 33
- 5 FENCHEL DUALITY 42
- 6 MONOTONE OPERATORS AND PROXIMAL POINTS 49
 - 6.1 Monotone operators 49
 - 6.2 Resolvents and proximal points 53
 - 6.3 Moreau–Yosida regularization 59
- 7 PROXIMAL POINT AND SPLITTING METHODS 64
 - 7.1 Proximal point method 64
 - 7.2 Explicit splitting 66
 - 7.3 Implicit splitting 70
 - 7.4 Primal-dual extragradient method 71

III LIPSCHITZ ANALYSIS

8 CLARKE SUBDIFFERENTIALS 76

9 SEMISMOOTH NEWTON METHODS 89

INTRODUCTION

Optimization is concerned with finding solutions to problems of the form

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

for a function $F : X \rightarrow \mathbb{R}$ and a set $U \subset X$. Specifically, one considers the following questions:

1. Does this problem admit a solution, i.e., is there an $\bar{x} \in U$ such that

$$F(\bar{x}) \leq F(x) \quad \text{for all } x \in U?$$

2. Is there an intrinsic characterization of \bar{x} , i.e., one not requiring comparison with all other $x \in U$?
3. How can this \bar{x} be computed (efficiently)?

For $U \subset \mathbb{R}^n$, these questions can be answered roughly as follows.

1. If U is compact and F is continuous, the Weierstraß Theorem yields that F attains its minimum at a point $\bar{x} \in U$ (as well as its maximum).
2. If F is differentiable, the *Fermat principle*

$$0 = F'(\bar{x})$$

holds.

3. If F is continuously differentiable and U is open, one can apply the *steepest descent* or gradient method to compute an \bar{x} satisfying the Fermat principle: Choosing a starting point x^0 and setting

$$x^{k+1} = x^k - t_k F'(x^k), \quad k = 0, \dots,$$

for suitable step sizes t_k , we have that $x^k \rightarrow \bar{x}$ for $k \rightarrow \infty$.

If F is even twice continuously differentiable, one can apply Newton's method to the Fermat principle: Choosing a suitable starting point x^0 and setting

$$x^{k+1} = x^k - F''(x^k)^{-1} F'(x^k), \quad k = 0, \dots,$$

we have that $x^k \rightarrow \bar{x}$ for $k \rightarrow \infty$.

However, there are many practically relevant functions that are *not* differentiable, such as the absolute value or maximum function. The aim 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. Here we concentrate on the two classes of

- i) convex functions,
- 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. To fix ideas, we aim at treating problems of the form

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

for a convex set $C \subset X$, a (possibly nonlinear but differentiable) operator $F : X \rightarrow Y$, $\alpha \geq 0$ 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 algorithm.

Since we deal with infinite-dimensional spaces, some knowledge of functional analysis is assumed, but the necessary background will be summarized in [Chapter 1](#). The results on pointwise operators on Lebesgue spaces also require elementary (Lebesgue) measure and integration theory. Basic familiarity with classical nonlinear optimization is helpful but not necessary.

These notes are based on graduate lectures given 2014 (in slightly different form) and 2016–2017 at the University of Duisburg-Essen. As such, no claim is made of originality (beyond possibly the selection – and, more importantly, omission – of material). Rather, like a magpie, I have collected the shiniest results and proofs I could find, mainly from [[Brokate 2014](#); [Schirotzek 2007](#); [Attouch et al. 2006](#); [Bauschke & Combettes 2017](#); [Clarke 2013](#); [Ulbrich 2002](#); [Schiela 2008](#)]. All mistakes, of course, are entirely my own.

Part I
BACKGROUND

1 FUNCTIONAL ANALYSIS

In this chapter we collect the basic concepts and results (and, more importantly, fix notations) from linear functional analysis that will be used in the following. For details and proofs, the reader is referred to the standard literature, e.g., [Alt 2016; Brezis 2010].

1.1 NORMED VECTOR SPACES

In the following, X will denote a vector space over the field \mathbb{K} , where we restrict ourselves for the sake of simplicity to the case $\mathbb{K} = \mathbb{R}$. A mapping $\|\cdot\| : X \rightarrow \mathbb{R}^+ := [0, \infty)$ is called a *norm* (on X), if for all $x \in X$ there holds

- (i) $\|\lambda x\| = |\lambda| \|x\|$ for all $\lambda \in \mathbb{K}$,
- (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 $X = \mathbb{R}^N$:

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

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

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

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

$$\|u\|_p = \left(\int_{\Omega} |u(x)|^p \right)^{1/p} \quad 1 \leq p < \infty,$$

$$\|u\|_{\infty} = \operatorname{ess\,sup}_{x \in \Omega} |u(x)|.$$

(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

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

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

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

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 $F : U \rightarrow Y$ be a mapping. We denote by

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

We call $F : U \rightarrow Y$

- *continuous* in $x \in U$, if for all $\varepsilon > 0$ there exists a $\delta > 0$ with

$$\|F(x) - F(z)\|_Y \leq \varepsilon \quad \text{for all } z \in U \text{ with } \|x - z\|_X \leq \delta;$$

- *Lipschitz continuous*, if there exists an $L > 0$ (called *Lipschitz constant*) with

$$\|F(x_1) - F(x_2)\|_Y \leq L\|x_1 - x_2\|_X \quad \text{for all } x_1, x_2 \in U.$$

- *locally Lipschitz continuous* in $x \in U$, if there exists a $\delta > 0$ and a $L = L(x, \delta) > 0$ with

$$\|F(x) - F(z)\|_Y \leq L\|x - z\|_X \quad \text{for all } z \in U \text{ with } \|x - z\|_X \leq \delta.$$

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

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

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

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

(which is equal to the smallest possible constant C in the definition of continuity). If $T \in L(X, Y)$ is bijective, the inverse $T^{-1} : Y \rightarrow X$ is continuous if and only if there exists a $c > 0$ with

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

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

1.2 STRONG AND WEAK CONVERGENCE

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 $x_n \rightarrow x$, if

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

A subset $U \subset X$ is called

- *closed*, if for every convergent sequence $\{x_n\}_{n \in \mathbb{N}} \subset U$ the limit $x \in U$ 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 $F : X \rightarrow Y$ is continuous if and only if $x_n \rightarrow x$ implies $F(x_n) \rightarrow F(x)$, and *closed*, if $x_n \rightarrow x$ and $F(x_n) \rightarrow y$ imply $F(x) = y$ (i.e., $\text{graph } F \subset X \times Y$ is a closed set).

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

- the *open ball* $O_r(x) := \{z \in X : \|x - z\|_X < r\}$ and
- the *closed ball* $K_r(x) := \{z \in X : \|x - z\|_X \leq r\}$.

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

- *open*, if for all $x \in U$ there exists an $r > 0$ with $O_r(x) \subset U$ (i.e., all $x \in U$ are *interior points* of U , which together form the *interior* U^o);
- *bounded*, if it is contained in $K_r(0)$ 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.

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. If Y is a Banach space, so is $L(X, Y)$ if endowed with the operator norm. 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*

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

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.

Of particular importance to us is the special case $L(X, Y)$ for $Y = \mathbb{R}$, the space of *bounded linear functionals* on X . In this case, $X^* := L(X, \mathbb{R})$ 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

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

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, $L^p(\Omega)^* \cong L^q(\Omega)$ 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. For any $x \in X$ there exists a $x^* \in X^*$ with*

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

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*

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

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

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

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 A° of A is nonempty and disjoint with B , there exists an $x^* \in X^* \setminus \{0\}$ and a $\lambda \in \mathbb{R}$ with*

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

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

$$tO_r(x_0) + (1-t)x = O_{tr}(tx_0 + (1-t)x) \subset A,$$

and in particular $x(t) := tx_0 + (1-t)x \in A^\circ$ for all $t \in (0, 1)$.

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

$$\langle x^*, x \rangle_X = \lim_{n \rightarrow \infty} \langle x^*, x_n \rangle_X \leq \lambda. \quad \square$$

In a certain way, a normed vector space is thus characterized by its dual. A direct consequence of **Theorem 1.4** is that the norm on a Banach space can be expressed in the manner of 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 $X \subset X^{**}$ is realized by the *canonical injection*

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

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 $1 < p < \infty$ but not ℓ^1, ℓ^∞ as well as $L^1(\Omega), L^\infty(\Omega)$ and $C(\overline{\Omega})$.

The duality pairing induces further notions of convergence: the *weak convergence* on X as well as the *weak-* convergence* on X^* .

(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

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

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

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

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 implies weak convergence by continuity of the duality pairing; in the same way, convergence with respect to the operator norm (also called *pointwise convergence*) implies weak-* convergence. If X is reflexive, weak and weak-* convergence (both in $X = X^{**}$!) coincide. In finite-dimensional spaces, all convergence notions coincide.

If $x_n \rightharpoonup x$ and $x_n^* \rightharpoonup^* x^*$ or $x_n \rightharpoonup x$ and $x_n^* \rightarrow x^*$, then $\langle x_n^*, x_n \rangle_X \rightarrow \langle x^*, x \rangle_X$. 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 1.8 (Eberlein–Šmuljan). *If X is a normed vector space, B_X is weakly compact if and only if X is reflexive.*

Hence in a reflexive space, all bounded and weakly closed sets are weakly compact. 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 1.9. *A convex set $U \subset X$ is closed if and only if it is weakly 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 $n \rightarrow \infty$ in the first inequality yields the contradiction

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

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

Theorem 1.10 (Banach–Alaoglu). *If X is a separable normed vector space (i.e., contains a countable dense subset), 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.

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 1991](#), Theorem 3.4(b)].

Theorem 1.11. *Let $A \subset X^*$ be a non-empty, 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*

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

Note, however, that closed convex sets in non-reflexive spaces do *not* have to be weakly-* closed.

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

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

It always holds that $\|T^*\|_{L(Y^*, X^*)} = \|T\|_{L(X, Y)}$. 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 $(\cdot, \cdot) : X \times X \rightarrow \mathbb{R}$ on a vector space X over \mathbb{R} is called *inner product*, if

- (i) $(\alpha x + \beta y, z) = \alpha(x, z) + \beta(y, z)$ for all $x, y, z \in X$ and $\alpha, \beta \in \mathbb{R}$;
- (ii) $(x, y) = (y, x)$ for all $x, y \in X$;
- (iii) $(x, x) \geq 0$ for all $x \in X$ with equality if and only if $x = 0$.

A Banach space together with an inner product $(X, (\cdot, \cdot)_X)$ is called a *Hilbert space*; if the inner product is canonical, it is frequently omitted, and the Hilbert space is simply denoted by X . An inner product induces a norm

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

which satisfies the *Cauchy–Schwarz inequality*:

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

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.

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

Theorem 1.12 (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*

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

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

[Theorem 1.12](#) 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

$$(x_n, z)_X \rightarrow (x, z)_X \quad \text{for all } z \in X.$$

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

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

If $T^* = T$, the operator T is called *self-adjoint*. Both definitions of adjoints are related via $T^* = J_X T^* J_Y^{-1}$. If the context is obvious, we will not distinguish the two in notation.

2 CALCULUS OF VARIATIONS

We first consider the question about the existence of minimizers of a (nonlinear) functional $F : U \rightarrow \mathbb{R}$ for a subset U of a Banach space X . Answering such questions is one of the goals of the *calculus of variations*.

2.1 THE DIRECT METHOD

It is helpful to include the constraint $x \in U$ into the functional by extending F to all of X with the value ∞ . We thus consider

$$\bar{F} : X \rightarrow \bar{\mathbb{R}} := \mathbb{R} \cup \{\infty\}, \quad \bar{F}(x) = \begin{cases} F(x) & x \in U, \\ \infty & x \in X \setminus U. \end{cases}$$

We use the usual arithmetic on $\bar{\mathbb{R}}$, i.e., $t < \infty$ 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} necessarily must lie in U .

We thus consider from now on functionals $F : X \rightarrow \bar{\mathbb{R}}$. The set on which F is finite is called the *effective domain*

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

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

We now generalize the Weierstraß Theorem (every real-valued continuous function on a compact set attains its minimum and maximum) to Banach spaces and in particular to functions of the form \bar{F} . Since we are only interested in minimizers, we only require a “one-sided” continuity: We call F *lower semi-continuous* in $x \in X$ if

$$F(x) \leq \liminf_{n \rightarrow \infty} F(x_n) \quad \text{for every } \{x_n\}_{n \in \mathbb{N}} \subset X \text{ with } x_n \rightarrow x.$$

Analogously, we define *weakly(-*) lower semi-continuous* functionals via weakly(-*) convergent sequences. Finally, F is called *coercive* if for every sequence $\{x_n\}_{n \in \mathbb{N}} \subset X$ with $\|x_n\|_X \rightarrow \infty$ we also have $F(x_n) \rightarrow \infty$.

We now have all concepts in 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 $F : X \rightarrow \overline{\mathbb{R}}$ be proper, coercive, and weakly lower semi-continuous. Then the minimization problem*

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

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

Proof. The proof can be separated into four steps.

(i) *Show that F is bounded from below.*

Assume that F is not bounded from below. Then there exists a sequence $\{x_n\}_{n \in \mathbb{N}} \subset X$ with $F(x_n) \rightarrow -\infty$. The coercivity of F then implies that this sequence is bounded (otherwise we would have $F(x_n) \rightarrow \infty$), i.e., there is an $M > 0$ with $\|x_n\|_X \leq M$ for all $n \in \mathbb{N}$. In particular, $\{x_n\}_{n \in \mathbb{N}} \subset K_M(0)$. The [Eberlein–Šmulyan Theorem 1.8](#) therefore implies the existence of a weakly converging subsequence $\{x_{n_k}\}_{k \in \mathbb{N}}$, whose limit $\tilde{x} \in X$ satisfies by weak lower semi-continuity

$$F(\tilde{x}) \leq \liminf_{k \rightarrow \infty} F(x_{n_k}) = -\infty.$$

But this is a contradiction to $\text{ran } F \subset \overline{\mathbb{R}}$.

(ii) *Show that there exists a minimizing sequence.*

Since F is proper and bounded from below, there must exist an $M := \inf_{x \in X} F(x) \in \mathbb{R}$. Thus, by the definition of the infimum, there exists a sequence $\{y_n\}_{n \in \mathbb{N}} \subset \text{ran } F \setminus \{\infty\} \subset \mathbb{R}$ with $y_n \rightarrow M$, i.e., there exists a sequence $\{x_n\}_{n \in \mathbb{N}} \subset X$ with

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

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).

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

The coercivity of F again implies that the minimizing sequence is bounded and thus by [Theorem 1.8](#) contains a weakly convergent subsequence $\{x_{n_k}\}_{k \in \mathbb{N}}$ with limit $\bar{x} \in X$. (This limit is a candidate for the minimizer.)

¹This 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 semi-continuity which we use here is due to [Leonida Tonelli](#) (1885–1946), who had a lasting influence on the modern calculus of variations through it.

(iv) Show that this limit is a minimizer.

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

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

Hence the infimum is attained in $\bar{x} \in \text{dom } F$, which is therefore the desired minimizer. \square

If X is not reflexive but the dual of a separable Banach space, we can argue analogously using the [Banach–Alaoglu Theorem 1.10](#)

Note how the topology on X used in the proof is restricted in step (iii) and (iv): Step (iii) profits from a coarse topology (in which more sequences are convergent), while step (iv) profits from a fine topology (the fewer sequences are convergent, the easier it is to satisfy the \liminf 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 semi-continuous.

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

$$F : X \rightarrow \overline{\mathbb{R}}, \quad x \mapsto \langle x^*, x \rangle_X,$$

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

Lemma 2.2. *Let X and Y be Banach spaces and $F : X \rightarrow \overline{\mathbb{R}}$ be weakly lower semi-continuous. Then, the following functionals are weakly lower semi-continuous as well:*

- (i) αF for all $\alpha \geq 0$;
- (ii) $F + G$ for $G : X \rightarrow \overline{\mathbb{R}}$ weakly lower semi-continuous;
- (iii) $\varphi \circ F$ for $\varphi : \overline{\mathbb{R}} \rightarrow \overline{\mathbb{R}}$ lower semi-continuous and strictly increasing.
- (iv) $F \circ \Phi$ for $\Phi : Y \rightarrow X$ 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 $F_i : X \rightarrow \overline{\mathbb{R}}$ weakly lower semi-continuous for an arbitrary set I .

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

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

Statement (iii) follows from the strict monotonicity and weak lower semi-continuity of φ since $x_n \rightharpoonup x$ implies

$$\varphi(F(x)) \leq \varphi(\liminf_{n \in \mathbb{N}} F(x_n)) \leq \liminf_{n \in \mathbb{N}} \varphi(F(x_n)).$$

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

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

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

$$F_j(x) \leq \liminf_{n \rightarrow \infty} F_j(x_n) \leq \liminf_{n \rightarrow \infty} \sup_{i \in I} F_i(x_n) \quad \text{for all } j \in I.$$

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

Corollary 2.3. *If X is a Banach space, the norm $\|\cdot\|_X$ is proper, coercive, and weakly lower semi-continuous.*

Proof. Coercivity and $\text{dom } \|\cdot\|_X = X$ follow directly from the definition. Weak lower semi-continuity follows from [Lemma 2.2](#) (v) and [Corollary 1.7](#) since

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

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 reduce the minimization of a functional $F : X \rightarrow \mathbb{R}$ over U to the minimization of $\bar{F} := F + \delta_U$ over X . The following result is therefore important for showing the existence of a minimizer.

Lemma 2.4. *Let X be a Banach space and $U \subset X$. Then, δ_U is*

- (i) *proper if U is non-empty;*
- (ii) *weakly lower semi-continuous if U is convex and closed;*
- (iii) *coercive if U is bounded.*

²not to be confused with the *characteristic function* χ_U with $\chi_U(x) = 1$ for $x \in U$ and 0 else

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 $\delta_U \geq 0$ immediately yields

$$\delta_U(x) = 0 \leq \liminf_{n \rightarrow \infty} \delta_U(x_n).$$

Let now $x \notin U$. Since U is convex and closed and hence by [Lemma 1.9](#) 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 $x_n \rightarrow x \in U$, in contradiction to the assumption). Thus, $\delta_U(x_n) = \infty$ for all $n \geq N$, and therefore

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

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

2.2 DIFFERENTIAL CALCULUS IN BANACH SPACES

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

Let X and Y be Banach spaces, $F : X \rightarrow Y$ be a mapping, and $x, h \in X$ be given.

- If the one-sided limit

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

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

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

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

- If additionally

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

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

- If the mapping $x \mapsto F'(x)$ 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 of course also the case for Gâteaux differentiable functions; Fréchet differentiability thus additionally requires a uniformity in h .)

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

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

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

Theorem 2.5. *Let X, Y , and Z be Banach spaces, and let $F : X \rightarrow Y$ be Fréchet differentiable in $x \in X$ and $G : Y \rightarrow Z$ be Fréchet differentiable in $y := F(x) \in Y$. Then, $G \circ F$ is Fréchet differentiable in x and*

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

Proof. For $h \in X$ with $x + h \in \text{dom } F$ 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 $r_1(t)/t \rightarrow 0$ for $t \rightarrow 0$. The Fréchet differentiability of F further implies

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

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

$$(2.1) \quad \|g\|_Y \leq \|F'(x)h\|_Y + r_2(\|h\|_X).$$

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

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

If $\|h\|_X \rightarrow 0$, we obtain from (2.1) and $F'(x) \in L(X, Y)$ that $\|g\|_Y \rightarrow 0$ as well, and the claim follows. \square

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

We will also need the following variant of the mean value theorem. Let $[a, b] \subset \mathbb{R}$ be a bounded interval and $f : [a, b] \rightarrow X$ be continuous. Then the Bochner integral $\int_a^b f(t) dt \in X$ is well-defined and by definition satisfies

$$(2.2) \quad \left\langle x^*, \int_a^b f(t) dt \right\rangle_X = \int_a^b \langle x^*, f(t) \rangle_X dt \quad \text{for all } x^* \in X^*,$$

as well as

$$(2.3) \quad \left\| \int_a^b f(t) dt \right\|_X \leq \int_a^b \|f(t)\|_X dt,$$

see, e.g., [Yosida 1995, Corollary v.1].

Theorem 2.6. *Let $F : U \rightarrow Y$ be Fréchet differentiable, and let $y \in U$ and $h \in Y$ be given with $y + th \in U$ for all $t \in [0, 1]$. Then*

$$F(y + h) - F(y) = \int_0^1 F'(y + th)h dt.$$

Proof. Consider for arbitrary $y^* \in Y^*$ the function

$$f : [0, 1] \rightarrow \mathbb{R}, \quad t \mapsto \langle y^*, F(y + th) \rangle_Y.$$

From [Theorem 2.5](#) we obtain that f (as a composition of mappings on Banach spaces) is differentiable with

$$f'(t) = \langle y^*, F'(y + th)h \rangle_Y,$$

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

$$\langle y^*, F(y + h) - F(y) \rangle_Y = f(1) - f(0) = \int_0^1 f'(t) dt = \left\langle y^*, \int_0^1 F'(y + th)h dt \right\rangle_Y,$$

where the last equality follows from (2.2). Since $y^* \in Y^*$ was arbitrary, the claim follows from this together with [Corollary 1.7](#). \square

We now turn to the characterization of minimizers of a differentiable function $F : X \rightarrow \mathbb{R}$.³

Theorem 2.7 (Fermat principle). *Let $F : X \rightarrow \mathbb{R}$ be Gâteaux differentiable and $\bar{x} \in X$ be a minimizer of F . Then $DF(\bar{x}) = 0$, i.e.,*

$$DF(\bar{x})h = F'(x; h) = 0 \quad \text{for all } h \in X.$$

³The *indirect method* of the calculus of variations uses this to show existence of minimizers as well, e.g., as the solution of a partial differential equation.

Proof. If \bar{x} is a minimizer of F , then for all $h \in X$ and sufficiently small $\varepsilon > 0$, the function $f : (-\varepsilon, \varepsilon) \rightarrow \mathbb{R}, t \mapsto F(\bar{x} + th)$, must have a minimum in $t = 0$. Since F is Gâteaux differentiable, the derivative $f'(t)$ in $t = 0$ exists and hence must satisfy

$$0 = f'(0) = \lim_{t \rightarrow 0^+} \frac{f(t) - f(0)}{t} = F'(x; h). \quad \square$$

Note that the Gâteaux derivative of a functional $F : X \rightarrow \mathbb{R}$ is an element of the *dual space* $X^* = L(X, \mathbb{R})$ and thus cannot be added to elements in X . However, in Hilbert spaces (and in particular in \mathbb{R}^n), we can use the [Fréchet–Riesz Theorem 1.12](#) to identify $DF(x) \in X^*$ with an element $\nabla F(x) \in X$, called *gradient* of F , in a canonical way via

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

As an example, let us consider the functional $F(x) = \frac{1}{2}\|x\|_X^2$, where the norm is induced by the inner product. Then we have for all $x, h \in X$ that

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

since the inner product is linear in h for fixed x . Hence, the squared norm is Gâteaux differentiable in x with derivative $DF(x) = (x, \cdot)_X \in X^*$ and gradient $\nabla F(x) = x \in X$; it is even Fréchet differentiable since

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

If the same mapping is now considered on a smaller Hilbert space $X' \hookrightarrow X$ (e.g., $X = L^2(\Omega)$ and $X' = H^1(\Omega)$), then the derivative $DF(x) \in (X')^*$ is still given by $DF(x)h = (x, h)_X$ (now only for all $h \in X'$), but the gradient $\nabla F \in X'$ is now characterized by

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

Different inner products thus lead to different gradients.

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 $f : \Omega \times \mathbb{R} \rightarrow \mathbb{R}$ with $\Omega \subset \mathbb{R}^n$ open and bounded as well as $p, q \in [1, \infty]$ the corresponding *superposition* or *Nemytskii operator*

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

For this operator to be well-defined requires certain restrictions on f . We call f 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

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

Under these conditions, F is even continuous.

Theorem 2.8. *If the Carathéodory function $f : \Omega \times \mathbb{R} \rightarrow \mathbb{R}$ satisfies the growth condition (2.5) for $p, q \in [1, \infty)$, then the superposition operator $F : L^p(\Omega) \rightarrow L^q(\Omega)$ defined via (2.4) is continuous.*

Proof. We sketch the essential steps; a complete proof can be found in, e.g., [Appell & Zabreiko 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 $u_n \rightarrow u \in L^p(\Omega)$. 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 $F(u_n) \rightarrow F(u)$ 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 is in $L^q(\Omega)$, and we can apply Lebesgue's dominated convergence theorem to deduce that $F(u_n) \rightarrow F(u)$ 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. \square

In fact, the growth condition (2.5) is also necessary for continuity; see [Appell & Zabreiko 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 $[F'(u)h](x) = f'(u(x))h(x)$. However, this does not hold in general; for example, the superposition operator

defined by $f(x, z) = \sin(z)$ is *not* differentiable in $u = 0$ for $1 \leq p = q < \infty$. The reason is that for a Fréchet differentiable superposition operator $F : L^p(\Omega) \rightarrow L^q(\Omega)$ and a direction $h \in L^p(\Omega)$, the pointwise(!) product has to satisfy $F'(u)h \in L^q(\Omega)$. This leads to additional conditions on the superposition operator F' defined by f' , which is known as *two norm discrepancy*.

Theorem 2.9. *Let $f : \Omega \times \mathbb{R} \rightarrow \mathbb{R}$ 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 $F : L^p(\Omega) \rightarrow L^q(\Omega)$ is continuously Fréchet differentiable, and its derivative in $u \in L^p(\Omega)$ in direction $h \in L^p(\Omega)$ is given by*

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

Proof. Theorem 2.8 yields that for $r := \frac{pq}{p-q}$ (i.e., $\frac{r}{p} = \frac{p'}{q}$), the superposition operator

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

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

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

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

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 (2.3) and (2.6) implies that

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

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

Part II

CONVEX ANALYSIS

3 CONVEX FUNCTIONS

The classical derivative concepts from the previous chapter are not sufficient for our purposes, since many interesting functionals are not differentiable in this sense; also, they cannot handle functionals with values in $\overline{\mathbb{R}}$. We therefore need a derivative concept that is more general than Gâteaux and Fréchet derivatives and still allows a Fermat principle and a rich calculus.

We first consider a general class of functionals that admit such a generalized derivative. A proper functional $F : X \rightarrow \overline{\mathbb{R}}$ is called *convex*, if for all $x, y \in X$ and $\lambda \in [0, 1]$, it holds that

$$(3.1) \quad F(\lambda x + (1 - \lambda)y) \leq \lambda F(x) + (1 - \lambda)F(y)$$

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

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

we call F *strictly convex*.

An alternative characterization of the convexity of a functional $F : X \rightarrow \overline{\mathbb{R}}$ is based on its *epigraph*

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

Lemma 3.1. *Let $F : X \rightarrow \overline{\mathbb{R}}$. Then $\text{epi } F$ 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 semi-continuous.*

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., $(x, t) \in \text{epi } F$.

For (ii), let F be convex and $(x, r), (y, s) \in \text{epi } F$ 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

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

and hence $\text{epi } F$ is convex. Let conversely $\text{epi } F$ 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 $(x, F(x)), (y, F(y)) \in \text{epi } F$. The convexity of $\text{epi } F$ then implies for all $\lambda \in [0, 1]$ that

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

and hence by definition of $\text{epi } F$ that (3.1) holds.

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

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

i.e., $(x, t) \in \text{epi } F$. Let conversely $\text{epi } F$ be closed and assume that F is not lower semi-continuous. Then there exists a sequence $\{x_n\}_{n \in \mathbb{N}} \subset X$ with $x_n \rightarrow x \in X$ and

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

We now distinguish two cases.

- a) $x \in \text{dom } F$: 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 $F(x_n) \leq F(x) - \varepsilon$ and thus $(x_n, F(x) - \varepsilon) \in \text{epi } F$ for all $n \in \mathbb{N}$. From $x_n \rightarrow x$ and the closedness of $\text{epi } F$, we deduce that $(x, F(x) - \varepsilon) \in \text{epi } F$ and hence $F(x) \leq F(x) - \varepsilon$, contradicting $\varepsilon > 0$.
- b) $x \notin \text{dom } F$: In this case, we can argue similarly using $F(x_n) \leq M + \varepsilon$ for $M > -\infty$ or $F(x_n) \leq \varepsilon$ for $M = -\infty$ to obtain a contradiction with $F(x) = \infty$.

The equivalence of weak lower semi-continuity and weak closedness follows in exactly the same way. \square

Note that $(x, t) \in \text{epi } F$ implies that $x \in \text{dom } F$; hence the effective domain of a proper, convex, and lower semi-continuous functional is always nonempty, convex, and closed as well. Also, together with Lemma 1.9 we immediately obtain

Corollary 3.2. *Let $F : X \rightarrow \overline{\mathbb{R}}$ be convex. Then, F is weakly lower semi-continuous if and only if F is lower semi-continuous.*

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

$$F_\alpha := \{x \in X : F(x) \leq \alpha\}, \quad \alpha \in \mathbb{R},$$

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

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

- (i) If F is convex, F_α is convex for all $\alpha \in \mathbb{R}$, but the converse does not hold.
- (ii) F is (weakly) lower semi-continuous if and only if F_α is (weakly) closed for all $\alpha \in \mathbb{R}$.

Directly from the definition we obtain the convexity of

- (i) 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 $t \in (0, 1)$,

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

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

Lemma 3.4. Let $f : \mathbb{R} \rightarrow \overline{\mathbb{R}}$ be proper, convex, and lower semi-continuous. If $\Omega \subset \mathbb{R}^n$ is bounded and $1 \leq p \leq \infty$, this also holds for

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

Proof. Since f is proper, there is a $t_0 \in \text{dom } f$. Hence, the constant function $u \equiv t_0 \in \text{dom } F$ as $f(u) \equiv f(t_0) \in L^\infty(\Omega) \subset L^1(\Omega)$.

For $u, v \in \text{dom } F$ (otherwise (3.1) is trivially satisfied) and $\lambda \in [0, 1]$, the convexity of f implies that

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

Since for $f, g \in L^1(\Omega)$ and $\alpha, \beta \in \mathbb{R}$ we also have $\alpha f + \beta g \in L^1(\Omega)$, it follows that $\lambda u + (1 - \lambda)v \in \text{dom } F$, and integration of the inequality over Ω yields the convexity of F .

Consider now $\{u_n\}_{n \in \mathbb{N}}$ with $u_n \rightarrow u$ in $L^p(\Omega)$. Then there exists a subsequence $\{u_{n_k}\}_{k \in \mathbb{N}}$ with $u_{n_k}(x) \rightarrow u(x)$ almost everywhere. Hence, the lower semi-continuity of f together with Fatou's Lemma implies that

$$F(u) = \int_{\Omega} f(u(x)) dx \leq \int_{\Omega} \liminf_{k \rightarrow \infty} f(u_{n_k}(x)) dx \leq \liminf_{k \rightarrow \infty} \int_{\Omega} f(u_{n_k}(x)) dx = \liminf_{k \rightarrow \infty} F(u_{n_k}).$$

As this argument can be applied to every further subsequence, the claim must hold for the whole sequence. \square

Further examples can be constructed as in [Lemma 2.2](#) through the following operations.

Lemma 3.5. *Let X and Y be normed vector spaces and let $F : X \rightarrow \overline{\mathbb{R}}$ be convex. Then the following functionals are convex as well:*

- (i) αF for all $\alpha \geq 0$;
- (ii) $F + G$ for $G : X \rightarrow \overline{\mathbb{R}}$ convex (if F or G are strictly convex, so is $F + G$);
- (iii) $\varphi \circ F$ for $\varphi : \overline{\mathbb{R}} \rightarrow \overline{\mathbb{R}}$ convex and increasing;
- (iv) $F \circ A$ for $A : Y \rightarrow X$ linear;
- (v) $x \mapsto \sup_{i \in I} F_i(x)$ with $F_i : X \rightarrow \overline{\mathbb{R}}$ convex for an arbitrary set I .

[Lemma 3.5](#) (v) in particular implies that the pointwise supremum of 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 $F : X \rightarrow \overline{\mathbb{R}}$ the *convex hull*

$$F^{\Gamma} : X \rightarrow \mathbb{R}, \quad x \mapsto \sup \{a(x) : a \text{ affine with } a(x) \leq F(x) \text{ for all } x \in X\}.$$

Lemma 3.6. *Let $F : X \rightarrow \overline{\mathbb{R}}$ be proper. Then F is convex and lower semi-continuous if and only if $F = F^{\Gamma}$.*

Proof. Since affine functionals are convex and continuous, [Lemma 3.5](#) (v) and [Lemma 2.2](#) (v) imply that $F = F^{\Gamma}$ is always continuous and lower semi-continuous.

Let now $F : X \rightarrow \overline{\mathbb{R}}$ be proper, convex, and lower semi-continuous. It is obvious from the definition of F^{Γ} as a supremum that $F^{\Gamma} \leq F$ always holds pointwise. Assume 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 an affine functional $a \in X^*$ 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 semi-continuous, $\text{epi } F$ is nonempty, convex, and closed

by Lemma 3.1. Furthermore, $\{(x_0, \lambda)\}$ is compact and, as $\lambda < F(x_0)$, disjoint with $\text{epi } F$. Theorem 1.5 (ii) hence yields a $z^* \in (X \times \mathbb{R})^*$ and an $\alpha \in \mathbb{R}$ with

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

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

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

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

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

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

(i) $s < 0$: We set

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

which is affine and continuous. Furthermore, $(x, F(x)) \in \text{epi } F$ for any $x \in \text{dom } F$, and using the productive zero in the first inequality in (3.2) implies (noting $s < 0$!) that

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

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

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

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

$$a_\rho : X \rightarrow \mathbb{R}, \quad x \mapsto a_0(x) + \rho(\langle x^*, x \rangle_X - \alpha),$$

which is a_ρ affine and continuous 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 $x \in \text{dom } F$ and arbitrary $\rho > 0$. 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 $F^\Gamma(x_0) < \lambda$. \square

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

Theorem 3.7. *Let X be a reflexive Banach space and let*

- (i) $U \subset X$ be nonempty, convex, and closed;
- (ii) $F : U \rightarrow \overline{\mathbb{R}}$ be proper, convex, and lower semi-continuous with $\text{dom } F \cap U \neq \emptyset$;
- (iii) U be bounded or F be coercive.

Then the problem

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

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

Proof. We consider the extended functional $\bar{F} = F + \delta_U : X \rightarrow \overline{\mathbb{R}}$. Assumption (i) together with Lemma 2.2 implies that δ_U is proper, convex, and weakly lower semi-continuous. From (ii) we obtain an $x_0 \in U$ with $\bar{F}(x_0) < \infty$, and hence \bar{F} is proper, convex, and weakly lower semi-continuous. Finally, \bar{F} is coercive since for bounded U , we can use that $F > -\infty$, and for coercive F , we can use that $\delta_U \geq 0$. Hence we can apply Theorem 2.1 to obtain the existence of a minimizer $\bar{x} \in \text{dom } \bar{F} = U \cap \text{dom } F$ of \bar{F} with

$$F(\bar{x}) = \bar{F}(\bar{x}) \leq \bar{F}(x) = F(x) \quad \text{for all } x \in U,$$

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 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 3.8. *Let $F : X \rightarrow \overline{\mathbb{R}}$ be proper, convex, and lower semi-continuous, and $x_0 \in X$ be given. Then the functional*

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

is coercive.

Proof. Since F is proper, convex, and lower semi-continuous, it follows from [Lemma 3.6](#) that F is bounded from below by an 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

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

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

To close this chapter, we show the following remarkable result: *Any real-valued convex functional is continuous.* (An extended real-valued proper functional must necessarily be discontinuous in some point.) This is remarkable because continuity is a topological property, while convexity is a purely algebraic property and doesn't imply any topology. So it is not even clear in *which* topology the functional is supposed to be continuous! This apparent contradiction is resolved by considering the following, sharper, result from which the claim will follow: A convex function that is bounded from above *in a neighborhood* of a point is continuous in this point (and even locally Lipschitz continuous). But “neighborhood” is a topological concept, and the function will be continuous in the topology that is defined by the choice of neighborhood. Here we will consider the strong topology in a normed vector space.

Theorem 3.9. *Let X be a normed vector space, $F : X \rightarrow \overline{\mathbb{R}}$ be convex, and $x \in X$. If there is a $\rho > 0$ such that F is bounded from above on $O_\rho(x)$, then F is locally Lipschitz continuous in x .*

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

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

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

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 $-F(y_1) \leq m = M - 2F(x)$, this yields the estimate

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

Exchanging the roles of y_1 and y_2 , we obtain that

$$|F(y_1) - F(y_2)| \leq \frac{2(M - F(x))}{\rho/2} \|y_1 - y_2\|_X \quad \text{for all } y_1, y_2 \in O_{\frac{\rho}{2}}(x)$$

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

We first deduce from this the desired result in the scalar case.

Corollary 3.10. *Let $f : \mathbb{R} \rightarrow \overline{\mathbb{R}}$ be convex. Then, f is locally Lipschitz continuous on $(\text{dom } f)^\circ$.*

Proof. Let $x \in (\text{dom } f)^\circ$, i.e., there exist $a, b \in \mathbb{R}$ with $x \in [a, b] \subset \text{dom } f$. 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 , and the claim follows from [Theorem 3.9](#). \square

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

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

Theorem 3.11. *Let X be a Banach space $F : X \rightarrow \overline{\mathbb{R}}$ be convex and lower semi-continuous. Then, F is locally Lipschitz continuous on $(\text{dom } F)^\circ$.*

Proof. We first show the claim for the case $x = 0 \in (\text{dom } F)^\circ$, which implies in particular that $M := |F(0)| < \infty$. Consider now for arbitrary $h \in X$ the mapping

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

It is straightforward to verify that f is convex and lower semi-continuous as well and satisfies $0 \in (\text{dom } f)^\circ$. By [Corollary 3.10](#), f is thus locally Lipschitz continuous in 0, i.e., $|f(t) - f(0)| \leq Lt \leq 1$ for sufficiently small $t > 0$. The reverse triangle inequality therefore yields a $\delta > 0$ with

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

Hence, 0 lies in the algebraic interior of the sublevel set F_{M+1} , which is convex and closed by Lemma 3.3. The core-int Lemma 1.2 thus yields that $0 \in (F_{M+1})^o$, i.e., there exists a $\rho > 0$ with $F(z) \leq M + 1$ for all $z \in O_\rho(0)$. This implies that F is locally bounded from above in 0 and hence locally Lipschitz continuous by Theorem 3.9.

For the general case $x \in (\text{dom } F)^o$, consider

$$\tilde{F} : X \rightarrow \overline{\mathbb{R}}, \quad y \mapsto F(y - x).$$

Again, it is straightforward to verify convexity and lower semi-continuity of \tilde{F} and that $0 \in (\text{dom } \tilde{F})^o$. It follows from the above that \tilde{F} is locally Lipschitz continuous in a neighborhood $O_\rho(0)$, which implies that

$$|F(y_1) - F(y_2)| = |\tilde{F}(y_1 + x) - \tilde{F}(y_2 + x)| \leq L\|y_1 - y_2\|_X \quad \text{for all } y_1, y_2 \in O_\rho(x)$$

and hence the local Lipschitz continuity of F . □

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

4 CONVEX SUBDIFFERENTIALS

We now turn to the characterization of minimizers of convex functionals via a Fermat principle. A first candidate for the required notion of derivative is the directional derivative since it exists (at least in the extended real-valued sense) for any convex function.

Lemma 4.1. *Let $F : X \rightarrow \overline{\mathbb{R}}$ be convex and let $x \in \text{dom } F$ and $h \in X$ be given. Then:*

(i) *the function*

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

is increasing;

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

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

(iii) *if $x \in (\text{dom } F)^0$, the limit $F'(x; h)$ is finite.*

Proof. *Ad (i):* Inserting the definition and sorting terms shows that for all $0 < s < t$, 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 $x + sh = (1 - \frac{s}{t})x + \frac{s}{t}(x + th)$.

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

$$F'(x; h) = \lim_{t \rightarrow 0^+} \varphi(t) = \inf_{t > 0} \varphi(t) \leq \varphi(1) = F(x + h) - F(x).$$

Ad (iii): Since $(\text{dom } F)^0$ is contained in the algebraic interior of $\text{dom } F$, there exists an $\varepsilon > 0$ such that $x + th \in \text{dom } F$ 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

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

Unfortunately, this concept can't yet be what we are looking for, since the convex function $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(t) = |t|$ has a minimum in $t = 0$, but $f'(0; h) = |h| > 0$ for $h \in \mathbb{R} \setminus \{0\}$. We thus don't have $f'(0; h) = 0$ for some $h \neq 0$ – but we at least have $0 \leq f'(0; h)$ for all $h \in \mathbb{R}$. It is this condition that we now generalize to normed vector spaces. For this purpose, consider for convex $F : X \rightarrow \overline{\mathbb{R}}$ and any $x \in \text{dom } F$ the set

$$(4.1) \quad \{x^* \in X^* : \langle x^*, h \rangle_X \leq F'(x; h) \quad \text{for all } h \in X\}.$$

With the help of [Lemma 4.1](#), this set (which can be empty!) can also be expressed without directional derivatives.

Lemma 4.2. *Let $F : X \rightarrow \overline{\mathbb{R}}$ be convex and $x \in \text{dom } F$. For any $x^* \in X^*$, the following statements are equivalent:*

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

Proof. If (i) holds, we immediately obtain from [Lemma 4.1](#) (ii) that

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

Conversely, if (ii) 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 then yields that

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

If we introduce $\tilde{x} = x + h \in X$, the second statement leads to our desired derivative concept: For $F : X \rightarrow \overline{\mathbb{R}}$ and $x \in \text{dom } F$, we define the (*convex*) *subdifferential* as

$$(4.2) \quad \partial F(x) := \{x^* \in X^* : \langle x^*, \tilde{x} - x \rangle_X \leq F(\tilde{x}) - F(x) \quad \text{for all } \tilde{x} \in X\}.$$

(Note that $\tilde{x} \notin \text{dom } F$ is allowed since in this case the inequality is trivially satisfied.) For $x \notin \text{dom } F$, we set $\partial F(x) = \emptyset$.¹ It follows directly from the definition that $\partial F(x)$ is convex and weakly-* closed. An element $\xi \in \partial F(x)$ is called a *subgradient*.

Theorem 4.3. *Let $F : X \rightarrow \overline{\mathbb{R}}$ and $\bar{x} \in \text{dom } F$. Then the following statements are equivalent:*

- (i) $0 \in \partial F(\bar{x})$;
- (ii) $F(\bar{x}) = \min_{x \in X} F(x)$.

¹We will later show that $\partial F(x)$ is nonempty and bounded for all $x \in (\text{dom } F)^\circ$; see [Corollary 8.9](#).

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

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

i.e., $F(\bar{x}) \leq F(\tilde{x})$ for all $\tilde{x} \in X$.² □

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

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 4.4. *Let $F : X \rightarrow \overline{\mathbb{R}}$ be convex and Gâteaux differentiable in 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 4.2 with $\tilde{x} := x + h$ now immediately yields $DF(x) \in \partial F(x)$.

Conversely, the definition of $\xi \in \partial F(x)$ with $h := \tilde{x} - x \in X$ implies that

$$\langle \xi, h \rangle_X \leq F'(x; h) = \langle DF(x), h \rangle_X.$$

Since $\tilde{x} \in X$ was arbitrary, this has to hold for all $h \in X$. Taking the supremum over all h with $\|h\|_X \leq 1$ now yields that $\|\xi - DF(x)\|_{X^*} \leq 0$, i.e., $\xi = DF(x)$. □

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

Theorem 4.5. *For any $x \in X$,*

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

²Note that convexity of F is not required for **Theorem 4.3**! 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 $x \in \text{dom } F$ 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.

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

$$\langle \xi, \tilde{x} \rangle_X \leq \|\tilde{x}\|_X \quad \text{for all } \tilde{x} \in X \setminus \{0\}$$

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

Let now $x \neq 0$ and consider $\xi \in \partial(\|\cdot\|_X)(x)$. Successively inserting $\tilde{x} = 0$ and $\tilde{x} = 2x$ in the definition (4.2) yields

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

i.e., $\langle \xi, x \rangle_X = \|x\|_X$. Similarly, we have for all $\tilde{x} \in X$ that

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

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

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

Hence, $\|\xi\|_{X^*} = 1$ 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. \square

In particular, we obtain for $X = \mathbb{R}$ the subdifferential of the absolute value function as

$$(4.3) \quad \partial(|\cdot|)(t) = \text{sign}(t) := \begin{cases} \{1\} & \text{if } t > 0, \\ \{-1\} & \text{if } t < 0, \\ [-1, 1] & \text{if } t = 0. \end{cases}$$

We can also give a more explicit characterization of the subdifferential of the indicator functional of a convex set $C \subset X$: For any $x \in C = \text{dom } \delta_C$, we have that

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

since the first inequality is trivially satisfied for all $\tilde{x} \notin C$. The set $\partial\delta_C(x)$ is also called the *normal cone* to C at x . Depending on the set C , this can be made even more explicit. Let $X = \mathbb{R}$ and $C = [-1, 1]$, and let $t \in C$. Then we have $\xi \in \partial\delta_{[-1,1]}(t)$ if and only if $\xi(\tilde{t} - t) \leq 0$ 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 positive if and only if $\xi \geq 0$.

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

Case 3: $t \in (-1, 1)$. Then $\tilde{t} - t$ can be positive as well as negative, and hence only $\xi = 0$ is possible.

We thus obtain that

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

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

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

Lemma 4.6. *Let $f : \mathbb{R} \rightarrow \overline{\mathbb{R}}$ be proper, convex, and lower semi-continuous, and let $F : L^p(\Omega) \rightarrow \overline{\mathbb{R}}$ with $1 \leq p < \infty$ be as in Lemma 3.4. Then we have for all $u \in \text{dom } F$ with $q := \frac{p}{p-1}$ that*

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

Proof. Let $u \in \text{dom } F$, i.e., $f \circ u \in L^1(\Omega)$, and let $u^* \in L^p(\Omega)$ be arbitrary. If $u^* \in L^q(\Omega)$ satisfies $u^*(x) \in \partial f(u(x))$ almost everywhere, we can insert $\tilde{u}(x)$ into the definition and 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

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

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

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

$$\partial F(x) = \left\{ \xi \in \mathbb{R}^N : \xi_i \in \partial f_i(x_i), \quad 1 \leq i \leq N \right\}.$$

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 .

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 4.7. *Let $F : X \rightarrow \overline{\mathbb{R}}$ be convex and $x \in \text{dom } F$. Then,*

- (i) $\partial(\lambda F)(x) = \lambda(\partial F(x)) := \{\lambda\xi : \xi \in \partial F(x)\}$ for $\lambda > 0$;
- (ii) $\partial F(\cdot + x_0)(x) = \partial F(x + x_0)$ for $x_0 \in X$ with $x + x_0 \in \text{dom } F$.

The sum rule is already considerably more delicate.

Theorem 4.8 (sum rule). *Let $F, G : X \rightarrow \overline{\mathbb{R}}$ be convex and $x \in \text{dom } F \cap \text{dom } G$. Then,*

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

with equality if there exists an $x_0 \in (\text{dom } F)^o \cap \text{dom } G$.

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

$$(4.4) \quad \langle \xi, \tilde{x} - x \rangle_X \leq (F(\tilde{x}) + G(\tilde{x})) - (F(x) + G(x)) \quad \text{for all } \tilde{x} \in X.$$

Our goal is now to use (as in the proof of [Lemma 3.6](#)) the characterization of convex functionals via their epigraph together with the Hahn–Banach separation theorem to construct a bounded linear functional $\zeta \in \partial G(x) \subset X^*$ with $\xi - \zeta \in \partial F(x)$, i.e.,

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

For that purpose, we define the sets

$$\begin{aligned} C_1 &:= \{(\tilde{x}, t - (F(x) - \langle \xi, x \rangle_X)) : F(\tilde{x}) - \langle \xi, \tilde{x} \rangle_X \leq t\}, \\ C_2 &:= \{(\tilde{x}, G(x) - t) : G(\tilde{x}) \leq t\}, \end{aligned}$$

i.e.,

$$C_1 = \text{epi}(F - \xi) - (0, F(x) - \langle \xi, x \rangle_X), \quad C_2 = -(\text{epi } G - (0, G(x))).$$

Since these are merely translations and, for C_2 , reflections of epigraphs of proper convex functionals, these sets are nonempty and convex. Furthermore, since x_0 is an interior point of $\text{dom } F = \text{dom}(F - \xi)$, the point (x_0, α) for α sufficiently large is an interior point of C_1 . Hence, $(C_1)^\circ$ is nonempty. It remains to show that $(C_1)^\circ$ and C_2 are disjoint. But this holds since any $(\tilde{x}, \alpha) \in (C_1)^\circ \cap C_2$ satisfies by definition that

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

contradicting (4.4). Corollary 1.6 therefore yields a pair $(x^*, s) \in (X \times \mathbb{R})^* \setminus \{(0, 0)\}$ and a $\lambda \in \mathbb{R}$ with

$$(4.5) \quad \langle x^*, \tilde{x} \rangle_X + s(t - (F(x) - \langle \xi, x \rangle_X)) \leq \lambda, \quad \tilde{x} \in \text{dom } F, t \geq F(\tilde{x}) - \langle \xi, \tilde{x} \rangle_X,$$

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

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

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

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. If $s > 0$, choosing $t > F(x) - \langle \xi, x \rangle_X$ makes the term in parentheses in (4.5) strictly positive, and taking $t \rightarrow \infty$ with fixed \tilde{x} leads to a contradiction to the boundedness by λ .

Hence $s < 0$, and (4.5) with $t = F(\tilde{x}) - \langle \xi, \tilde{x} \rangle_X$ and (4.6) with $t = G(\tilde{x})$ imply that

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

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

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 4.9 (chain rule). *Let $A \in L(X, Y)$, $F : Y \rightarrow \overline{\mathbb{R}}$ be convex, and $x \in \text{dom}(F \circ A)$. Then,*

$$\partial(F \circ A)(x) \supset A^* \partial F(Ax) := \{A^* y^* : y^* \in \partial F(Ax)\}$$

with equality if there exists an $x_0 \in X$ with $Ax_0 \in (\text{dom } F)^{\circ}$.

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

$$F(A\tilde{x}) - F(Ax) \geq \langle \eta, A\tilde{x} - Ax \rangle_Y = \langle A^* \eta, \tilde{x} - x \rangle_X,$$

i.e., $\xi := A^* \eta \in \partial(F \circ A) \subset X^*$.

Let now $x \in \text{dom}(F \circ A)$ and $\xi \in \partial(F \circ A)(x)$, i.e.,

$$F(Ax) + \langle \xi, \tilde{x} - x \rangle_X \leq F(A\tilde{x}) \quad \text{for all } \tilde{x} \in X.$$

As in the proof of a sum rule, we could now construct an $\eta \in \partial F(Ax)$ with $\xi = A^* \eta$ by separating epi F and

$$\text{graph } A = \{(x, Ax) : x \in X\} \subset X \times Y.$$

For the sake of variety (and because the “lifting” technique we will use can be applied in other contexts as well), we will instead apply the sum rule to

$$H : X \times Y \rightarrow \overline{\mathbb{R}}, \quad (x, y) \mapsto F(y) + \delta_{\text{graph } A}(x, y).$$

Since A is linear, $\text{graph } A$ and hence $\delta_{\text{graph } A}$ are convex. Furthermore, $Ax \in \text{dom } F$ by assumption and thus $(x, Ax) \in \text{dom } H$.

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

$$\langle \xi, \tilde{x} - x \rangle_X + \langle 0, \tilde{y} - Ax \rangle_Y \leq F(\tilde{y}) - F(Ax) + \delta_{\text{graph } A}(\tilde{x}, \tilde{y}) - \delta_{\text{graph } A}(x, Ax).$$

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

$$\langle \xi, \tilde{x} - x \rangle_X \leq F(A\tilde{x}) - F(Ax) \quad \text{for all } \tilde{x} \in X,$$

i.e., $\xi \in \partial(F \circ A)(x)$. Conversely, let $\xi \in \partial(F \circ A)(x)$. Since $\delta_{\text{graph } A}(x, Ax) = 0$ and $\delta_{\text{graph } A}(\tilde{x}, \tilde{y}) \geq 0$, it then follows for all $\tilde{x} \in X$ and $\tilde{y} \in Y$ that

$$\begin{aligned} \langle \xi, \tilde{x} - x \rangle_X + \langle 0, \tilde{y} - Ax \rangle_Y &= \langle \xi, \tilde{x} - x \rangle_X \\ &\leq F(A\tilde{x}) - F(Ax) + \delta_{\text{graph } A}(\tilde{x}, \tilde{y}) - \delta_{\text{graph } A}(x, Ax) \\ &= F(\tilde{y}) - F(Ax) + \delta_{\text{graph } A}(\tilde{x}, \tilde{y}) - \delta_{\text{graph } A}(x, Ax), \end{aligned}$$

where we have used that last equality holds trivially as $\infty = \infty$ for $\tilde{y} \neq A\tilde{x}$. Hence, $(\xi, 0) \in \partial H(x, Ax)$.

We now consider $\tilde{F} : X \times Y \rightarrow \overline{\mathbb{R}}, (x, y) \mapsto F(y)$, and $(x_0, Ax_0) \in \text{graph } A = \text{dom } \delta_{\text{graph } A}$. Since $Ax_0 \in (\text{dom } F)^{\circ} Y$ by assumption, $(x_0, Ax_0) \in (\text{dom } \tilde{F})^{\circ}$ as well. We can thus apply [Theorem 4.8](#) to obtain

$$(\xi, 0) \in \partial H(x, Ax) = \partial \tilde{F}(Ax) + \partial \delta_{\text{graph } A}(x, Ax),$$

i.e., $(\xi, 0) = (x^*, y^*) + (w^*, z^*)$ for some $(x^*, y^*) \in \partial \tilde{F}(Ax)$ and $(w^*, z^*) \in \partial \delta_{\text{graph } A}(x, Ax)$.

Now we have $(x^*, y^*) \in \partial \tilde{F}(Ax)$ if and only if

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

Fixing $\tilde{x} = x$ and $\tilde{y} = Ax$ implies that $y^* \in \partial F(Ax)$ and $x^* = 0$, respectively. Furthermore, $(w^*, z^*) \in \partial \delta_{\text{graph } A}(x, Ax)$ if and only if

$$\langle w^*, \tilde{x} - x \rangle_X + \langle z^*, \tilde{y} - Ax \rangle_Y \leq 0 \quad \text{for all } (\tilde{x}, \tilde{y}) \in \text{graph } A,$$

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

$$\langle w^* + A^* z^*, \tilde{x} - x \rangle_X \leq 0 \quad \text{for all } \tilde{x} \in X$$

and hence $w^* = -A^* z^*$. Together we obtain

$$(\xi, 0) = (0, y^*) + (-A^* z^*, z^*),$$

which implies $y^* = -z^*$ and thus $\xi = -A^* z^* = A^* y^*$ with $y^* \in \partial F(Ax)$ as claimed. \square

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

Corollary 4.10. *Let $U \subset X$ be nonempty, convex, and closed, and let $F : X \rightarrow \overline{\mathbb{R}}$ be proper, convex, and lower semi-continuous. If there exists an $x_0 \in U^{\circ} \cap \text{dom } F$, then $\bar{x} \in U$ solves*

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

if and only if there exists a $\xi \in X^*$ with

$$(4.7) \quad \begin{cases} -\xi \in \partial F(\bar{x}), \\ \langle \xi, \tilde{x} - \bar{x} \rangle \leq 0 \quad \text{for all } \tilde{x} \in U. \end{cases}$$

Proof. Since F and U are convex, we can apply [Theorem 4.3](#) to $J := F + \delta_U$. Furthermore, since $x_0 \in U^{\circ} = (\text{dom } \delta_U)^{\circ}$, we can also apply [Theorem 4.8](#). 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). \square

If $F : X \rightarrow \mathbb{R}$ is Gâteaux differentiable (and hence finite-valued), (4.7) coincide with the classical *Karush–Kuhn–Tucker conditions*; the existence of an interior point $x_0 \in U^{\circ}$ is exactly the *Slater condition* needed to show existence of the Lagrange multiplier ξ .

5 FENCHEL DUALITY

A particularly useful calculus rule connects the convex subdifferential with the so-called Fenchel–Legendre transform. Let X be a normed vector space and $F : X \rightarrow \overline{\mathbb{R}}$ be proper but not necessarily convex. We then define the *Fenchel conjugate* of F as

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

(Since $\text{dom } F \neq \emptyset$ is excluded, we have that $F^*(x^*) > -\infty$ for all $x^* \in X^*$, and hence the definition is meaningful.) [Lemma 3.5](#) (v) and [Lemma 2.2](#) (v) immediately imply that F^* is always convex and lower semi-continuous (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 semi-continuous by [Lemma 3.6](#)), then F^* is proper as well. Finally, the definition directly yields the *Fenchel–Young inequality*

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

Intuitively, $F^*(x^*)$ is the (negative of the) affine part of the tangent to F (in the point x in which the supremum is attained) with slope x^* . Similarly, we define the Fenchel conjugate of $F : X^* \rightarrow \overline{\mathbb{R}}$ (i.e., if F is defined on some dual space) as

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

The point of this convention is that even in nonreflexive spaces, the *biconjugate* $F^{**} := (F^*)^*$ is again defined on X (rather than $X^{**} \supset X$). Intuitively, F^{**} is the convex hull of F , which by [Lemma 3.6](#) coincides with F itself if F is convex.

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

- (i) $F^{**} \leq F$;
- (ii) $F^{**} = F^\Gamma$;
- (iii) $F^{**} = F$ if and only if F is convex and lower semi-continuous.

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 semi-continuous by definition as a Fenchel conjugate as well as proper by (i). Hence, Lemma 3.6 yields that

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

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 3.6. \square

We again consider some relevant examples.

Example 5.2.

- (i) Let X be a Hilbert space and $F(x) = \frac{1}{2}\|x\|_X^2$. Using the Fréchet–Riesz Theorem 1.12, we identify X with its dual X^* and can express the duality pairing using the inner product. Since F is Fréchet differentiable with gradient $\nabla F(x) = x$, the solution $\bar{x} \in X$ to

$$\sup_{x \in X} \langle x^*, x \rangle_X - \frac{1}{2} \langle x, x \rangle_X$$

for given $x^* \in X$ has to satisfy the Fermat principle, i.e., $\bar{x} = x^*$. Inserting this into the definition and simplifying yields the Fenchel conjugate

$$F^* : X \rightarrow \mathbb{R}, \quad F^*(x^*) = \frac{1}{2}\|x^*\|_X^2.$$

- (ii) Let B_X be the unit ball in the normed vector space X and take $F = \delta_{B_X}$. Then we have for any $x^* \in X^*$ that

$$(\delta_{B_X})^*(x^*) = \sup_{x \in X} \langle x^*, x \rangle_X - \delta_{B_X}(x) = \sup_{\|x\|_X \leq 1} \langle x^*, x \rangle_X = \|x^*\|_{X^*}.$$

Similarly, one shows using the definition of the Fenchel conjugate in dual spaces and Corollary 1.7 that $(\delta_{B_{X^*}})^*(x) = \|x\|_X$.

- (iii) 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 $t \rightarrow \infty$ 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 $F^* = \delta_{B_{X^*}}$. As above, a similar argument shows that $(\|\cdot\|_{X^*})^* = \delta_{B_X}$.

Fenchel conjugates satisfy a number of useful calculus rules, which follow directly from the properties of the supremum.

Lemma 5.3. *Let $F : X \rightarrow \overline{\mathbb{R}}$ be proper. Then,*

- (i) $(\alpha F)^* = \alpha F^* \circ (\alpha^{-1} \text{Id})$ 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) $(F \circ A)^* = F^* \circ A^{-*}$ for continuously invertible $A \in L(Y, X)$ and $A^{-*} := (A^{-1})^*$.

Proof. *Ad (i):* For any $\alpha > 0$, we have that

$$(\alpha F)^*(x^*) = \sup_{x \in X} (\alpha \langle \alpha^{-1} x^*, x \rangle_X - \alpha F(x)) = \alpha \sup_{x \in X} (\langle \alpha^{-1} x^*, x \rangle_X - F(x)) = \alpha F^*(\alpha^{-1} x^*).$$

Ad (ii): Since $\{x + x_0 : x \in X\} = X$, we have that

$$\begin{aligned} (F(\cdot + x_0) + \langle x_0^*, \cdot \rangle_X)^*(x^*) &= \sup_{x \in X} \langle x^*, x \rangle_X - F(x^* + x_0) - \langle x_0^*, x_0 \rangle_X \\ &= \sup_{x \in X} (\langle x^* - x_0^*, x + x_0 \rangle_X - F(x^* + x_0)) - \langle x^* - x_0^*, x_0 \rangle_X \\ &= \sup_{\tilde{x} = x + x_0, x \in X} (\langle x^* - x_0^*, \tilde{x} \rangle_X - F(\tilde{x})) - \langle x^* - x_0^*, x_0 \rangle_X \\ &= F^*(x^* - x_0^*) - \langle x^* - x_0^*, x_0 \rangle_X. \end{aligned}$$

Ad (iii): Since $X = \text{ran } A$, we have that

$$\begin{aligned} (F \circ A)^*(y^*) &= \sup_{y \in Y} \langle y^*, A^{-1}Ay \rangle_Y - F(Ay) \\ &= \sup_{x = Ay, y \in Y} \langle A^{-*}y^*, x \rangle_X - F(x) = F^*(A^{-*}y^*). \end{aligned} \quad \square$$

As in [Lemma 4.6](#),¹ one can show that Fenchel conjugates of integral functionals can be computed pointwise; see, e.g., [[Rockafellar 1976](#), Theorem 3C].

Lemma 5.4. *Let $f : \mathbb{R} \rightarrow \overline{\mathbb{R}}$ be measurable, proper and lower semi-continuous, and let $F : L^p(\Omega) \rightarrow \overline{\mathbb{R}}$ with $1 \leq p < \infty$ be defined as in [Lemma 3.4](#). Then we have for $q = \frac{p}{p-1}$ that*

$$F^* : L^q(\Omega) \rightarrow \overline{\mathbb{R}}, \quad F^*(u^*) = \int_{\Omega} f^*(u^*(x)) \, dx.$$

There are some obvious similarities between the definitions of the Fenchel conjugate and of the subdifferential, which yield the following very useful property.

Lemma 5.5. *Let $F : X \rightarrow \overline{\mathbb{R}}$ be proper, convex, and lower semi-continuous. 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

$$(5.2) \quad \langle x^*, x \rangle_X - F(x) = F^*(x^*) \geq \langle x^*, \tilde{x} \rangle_X - F(\tilde{x}) \quad \text{for all } \tilde{x} \in X,$$

which again by definition is equivalent to $x^* \in \partial F(x)$. 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](#) yields that for all $\tilde{x}^* \in X^*$,

$$\langle x^*, x \rangle_X - F^*(x^*) = F(x) = F^{**}(x) \geq \langle \tilde{x}^*, x \rangle - F^*(\tilde{x}^*),$$

yielding as above the equivalence of (i) and (iii). □

Remark. If X is not reflexive, $x \in \partial F^*(x^*) \subset X^{**}$ in (iii) has to be understood via the canonical injection, i.e., as

$$\langle J(x), \tilde{x}^* - x^* \rangle_{X^*} = \langle \tilde{x}^* - x^*, x \rangle_X \leq F^*(\tilde{x}^*) - F^*(x^*) \quad \text{for all } \tilde{x}^* \in X^*.$$

Using (iii) to conclude equality in (i) or, equivalently, the subdifferential inclusion (ii) therefore requires the additional condition that $x \in X \subset X^{**}$. Conversely, if (i) or (ii) hold, (iii) also guarantees that the subgradient $x \in \partial F^*(x^*) \cap X$, which is a stronger fact. (Similar statements apply to $F : X^* \rightarrow \overline{\mathbb{R}}$ and $F^* : X \rightarrow \overline{\mathbb{R}}$.)

¹modulo some technical difficulties since measurability of $f \circ u$ does not immediately imply that of $f^* \circ u^*$

Lemma 5.5 plays the role of a “convex inverse function theorem”, and can be used to, e.g., 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

$$(5.3) \quad \inf_{x \in X} F(x) + G(Ax)$$

for $F : X \rightarrow \overline{\mathbb{R}}$ and $G : Y \rightarrow \overline{\mathbb{R}}$ proper, convex, and lower semi-continuous, and $A \in L(X, Y)$, we can use **Theorem 5.1** to replace G with the definition of G^{**} and obtain

$$\inf_{x \in X} \sup_{y^* \in Y^*} F(x) + \langle y^*, Ax \rangle_Y - G^*(y^*).$$

If(!) we were now able to exchange inf and sup, we could write (with $\inf F = -\sup(-F)$)

$$\begin{aligned} \inf_{x \in X} \sup_{y^* \in Y^*} F(x) + \langle y^*, Ax \rangle_Y - G^*(y^*) &= \sup_{y^* \in Y^*} \inf_{x \in X} F(x) + \langle y^*, Ax \rangle_Y - G^*(y^*) \\ &= \sup_{y^* \in Y^*} - \left(\sup_{x \in X} -F(x) + \langle -A^* y^*, x \rangle_X \right) - G^*(y^*). \end{aligned}$$

By definition of F^* , we thus obtain the *dual problem*

$$(5.4) \quad \sup_{y^* \in Y^*} -F^*(-A^* y^*) - G^*(y^*).$$

As a side effect, we have shifted the operator A from G to F^* without having to invert it.

The following theorem in an elegant way uses the Fermat principle, the sum and chain rules, and the Fenchel–Young equation to derive sufficient conditions for the exchangeability.

Theorem 5.6 (Fenchel–Rockafellar). *Let X and Y be normed vector spaces, $F : X \rightarrow \overline{\mathbb{R}}$ and $G : Y \rightarrow \overline{\mathbb{R}}$ be proper, convex, and lower semi-continuous, and $A \in L(X, Y)$. Assume furthermore that*

- (i) *the primal problem (5.3) admits a solution $\bar{x} \in X$;*
- (ii) *there exists an $x_0 \in \text{dom } F \cap \text{dom}(G \circ A)$ with $Ax_0 \in (\text{dom } G)^0$.*

Then, the dual problem (5.4) admits a solution $\bar{y}^ \in Y^*$ and*

$$(5.5) \quad \min_{x \in X} F(x) + G(Ax) = \max_{y^* \in Y^*} -F^*(-A^* y^*) - G^*(y^*).$$

Furthermore, \bar{x} and \bar{y}^ are solutions to (5.3) and (5.4), respectively, if and only if*

$$(5.6) \quad \begin{cases} -A^* \bar{y}^* \in \partial F(\bar{x}), \\ \bar{y}^* \in \partial G(A\bar{x}). \end{cases}$$

Proof. [Theorem 4.3](#) states that $\bar{x} \in X$ is a solution to (5.3) if and only if $0 \in \partial(F + G \circ A)(\bar{x})$. By assumption (ii), [Theorems 4.8](#) and [4.9](#) are applicable, and we thus obtain that

$$0 \in \partial(F + G \circ A)(\bar{x}) = \partial F(\bar{x}) + A^* \partial G(A\bar{x}),$$

which implies that there exists a $\bar{y}^* \in \partial G(A\bar{x})$ with $-A^* \bar{y}^* \in \partial F(\bar{x})$, i.e., satisfying (5.6).

The relations (5.6) together with [Lemma 5.5](#) further imply equality in the Fenchel–Young inequalities for F and G , i.e.,

$$(5.7) \quad \begin{cases} \langle -A^* \bar{y}^*, \bar{x} \rangle_X = F(\bar{x}) + F^*(-A^* \bar{y}^*), \\ \langle \bar{y}^*, A\bar{x} \rangle_Y = G(A\bar{x}) + G^*(\bar{y}^*). \end{cases}$$

Adding both equations now yields

$$(5.8) \quad F(\bar{x}) + G(A\bar{x}) = -F^*(-A^* \bar{y}^*) - G^*(\bar{y}^*).$$

It remains to show that \bar{y}^* is a solution to (5.4). For this purpose, we introduce

$$L : X \times Y^* \rightarrow \overline{\mathbb{R}}, \quad L(x, y^*) = F(x) + \langle y^*, Ax \rangle_Y - G^*(y^*).$$

For all $\tilde{x} \in X$ and $\tilde{y}^* \in Y^*$, we always have that

$$\sup_{y^* \in Y^*} L(\tilde{x}, y^*) \geq L(\tilde{x}, \tilde{y}^*) \geq \inf_{x \in X} L(x, \tilde{y}^*),$$

and hence (taking the infimum over all \tilde{x} in the first and the supremum over all \tilde{y}^* in the second inequality) that

$$\inf_{x \in X} \sup_{y^* \in Y^*} L(x, y^*) \geq \sup_{y^* \in Y^*} \inf_{x \in X} L(x, y^*).$$

We thus obtain that

$$\begin{aligned} F(\bar{x}) + G(A\bar{x}) &= \inf_{x \in X} \sup_{y^* \in Y^*} F(x) + \langle y^*, Ax \rangle_Y - G^*(y^*) \\ &\geq \sup_{y^* \in Y^*} \inf_{x \in X} F(x) + \langle y^*, Ax \rangle_Y - G^*(y^*) \\ &= \sup_{y^* \in Y^*} -F^*(-A^* y^*) - G^*(y^*). \end{aligned}$$

Combining this with (5.8) yields that

$$-F^*(-A^* \bar{y}^*) - G^*(\bar{y}^*) = F(\bar{x}) + G(A\bar{x}) \geq \sup_{y^* \in Y^*} -F^*(-A^* y^*) - G^*(y^*),$$

i.e., \bar{y}^* is a solution to (5.4), and hence (5.5) follows from (5.8).

Finally, if $\bar{x} \in X$ and $\bar{y}^* \in Y^*$ are solutions to (5.3) and (5.4), respectively, then (5.5) implies (5.8). Together with the productive zero, this implies that

$$0 = [F(\bar{x}) + F^*(-A^* \bar{y}^*) - \langle -A^* \bar{y}^*, \bar{x} \rangle_X] + [G(A\bar{x}) + G^*(\bar{y}^*) - \langle \bar{y}^*, A\bar{x} \rangle_Y].$$

Since both brackets have to be nonnegative due to the Fenchel–Young inequality, they each have to be zero. We therefore deduce that (5.7) holds, and hence [Lemma 5.5](#) implies (5.6). \square

The relations (5.6) are referred to as *Fenchel extremality conditions*; we can use [Lemma 5.5](#) to generate further, equivalent, optimality conditions by inverting one or the other sub-differential inclusion. We will later exploit this to derive implementable algorithms for solving optimization problems of the form (5.3).

6 MONOTONE OPERATORS AND PROXIMAL POINTS

Any minimizer $\bar{x} \in X$ of the convex functional $F : X \rightarrow \overline{\mathbb{R}}$ satisfies by [Theorem 4.3](#) the Fermat principle $0 \in \partial F(\bar{x})$. To obtain from this useful information about (and, later, implementable algorithms for the computation of) \bar{x} , we thus have to study the mapping $x \mapsto \partial F(x)$. To avoid technical difficulties – and since we will use the following results mainly for numerical algorithms, i.e., for $X = \mathbb{R}^N$ – we restrict the discussion in this and the next chapter to Hilbert spaces. This allows identifying X^* with X ; in particular, we will from now on identify the set $\partial F(x) \subset X^*$ of subgradients with the corresponding set in X of Riesz representations.

6.1 MONOTONE OPERATORS

For two normed vector spaces X and Y we consider a *set-valued mapping* $A : X \rightarrow \mathcal{P}(Y)$, also denoted by $A : X \rightrightarrows Y$, and define

- its *domain of definition* $\text{dom } A = \{x \in X : Ax \neq \emptyset\}$;
- its *range* $\text{ran } A = \bigcup_{x \in X} Ax$;
- its *graph* $\text{graph } A = \{(x, y) \in X \times Y : y \in Ax\}$;
- its *inverse* $A^{-1} : Y \rightrightarrows X$ via $A^{-1}(y) = \{x \in X : y \in Ax\}$ for all $y \in Y$.

(Note that $A^{-1}(y) = \emptyset$ is allowed by the definition; hence for set-valued mappings, their inverse and preimage – which always exists – coincide.)

For $A, B : X \rightrightarrows Y, C : Y \rightrightarrows Z$, and $\lambda \in \mathbb{R}$ we further define

- $\lambda A : X \rightrightarrows Y$ via $(\lambda A)(x) = \{\lambda y : y \in Ax\}$;
- $A + B : X \rightrightarrows Y$ via $(A + B)(x) = \{y + z : y \in Ax, z \in Bx\}$;
- $C \circ A : X \rightrightarrows Z$ via $(C \circ A)(x) = \{z : \text{there is } y \in Ax \text{ with } z \in Cy\}$.

A set-valued mapping $A : X \rightrightarrows X$ is called *monotone* if $\text{graph } A \neq \emptyset$ (to exclude trivial cases) and

$$(6.1) \quad (x_1^* - x_2^*, x_1 - x_2)_X \geq 0 \quad \text{for all } (x_1, x_1^*), (x_2, x_2^*) \in \text{graph } A.$$

Obviously, the identity mapping $\text{Id} : X \rightrightarrows X, x \mapsto \{x\}$, is monotone. If $F : X \rightarrow \overline{\mathbb{R}}$ is convex, then $\partial F : X \rightrightarrows X, x \mapsto \partial F(x)$, is monotone: 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

$$\begin{aligned} (x_1^*, \tilde{x} - x_1) &\leq F(\tilde{x}) - F(x_1) && \text{for all } \tilde{x} \in X, \\ (x_2^*, \tilde{x} - x_2) &\leq F(\tilde{x}) - F(x_2) && \text{for all } \tilde{x} \in X. \end{aligned}$$

Adding the first inequality for $\tilde{x} = x_2$ and the second for $\tilde{x} = x_1$ and rearranging the result yields (6.1). Furthermore, if $A, B : X \rightrightarrows X$ are monotone and $\lambda \geq 0$, then λA and $A + B$ are monotone as well.

In fact, we will need the following, stronger, property, which guarantees that A is closed: A monotone operator $A : X \rightrightarrows X$ is called *maximally monotone*, if for any $x \in X$ and $x^* \in X$ the condition

$$(6.2) \quad (x^* - \tilde{x}^*, x - \tilde{x})_X \geq 0 \quad \text{for all } (\tilde{x}, \tilde{x}^*) \in \text{graph } A$$

implies that $x^* \in Ax$. (In other words, (6.2) holds if and only if $(x, x^*) \in \text{graph } A$.) For fixed $x \in X$ and $x^* \in X$, the condition claims that if A is monotone, so is the extension

$$\tilde{A} : X \rightrightarrows X, \quad \tilde{x} \mapsto \begin{cases} Ax \cup \{x^*\} & \text{if } \tilde{x} = x, \\ A\tilde{x} & \text{if } \tilde{x} \neq x. \end{cases}$$

For A to be maximally monotone means that this is not a true extension, i.e., $\tilde{A} = A$. For example, the operator

$$A : \mathbb{R} \rightrightarrows \mathbb{R}, \quad t \mapsto \begin{cases} \{1\} & \text{if } t > 0, \\ \{0\} & \text{if } t = 0, \\ \{-1\} & \text{if } t < 0, \end{cases}$$

is monotone but not maximally monotone, since A is a proper subset of the monotone operator defined by $\tilde{A}t = \text{sign}(t) = \partial(|\cdot|)(t)$.

Several useful properties follow directly from the definition.

Lemma 6.1. *If $A : X \rightrightarrows X$ is maximally monotone and $\lambda > 0$, so is λA for all $\lambda > 0$.*

Proof. Let $x, x^* \in X$ and assume that

$$0 \leq (x^* - \tilde{x}^*, x - \tilde{x})_X = \lambda (\lambda^{-1}x^* - \lambda^{-1}\tilde{x}^*, x - \tilde{x})_X \quad \text{for all } (\tilde{x}, \tilde{x}^*) \in \text{graph } \lambda A.$$

Since $\tilde{x}^* \in \lambda Ax$ if and only if $\lambda^{-1}\tilde{x}^* \in Ax$ and A is maximally monotone, this implies that $\lambda^{-1}\tilde{x}^* \in A\tilde{x}$, i.e., $\tilde{x}^* \in (\lambda A)\tilde{x}$. Hence, λA is maximally monotone. \square

Lemma 6.2. *Let $A : X \rightrightarrows X$ be maximally monotone. Then A is weakly–strongly closed, i.e., $x_n \rightharpoonup x$ and $Ax_n \ni x_n^* \rightarrow x^*$ imply that $x^* \in Ax$.*

Proof. For arbitrary $\tilde{x} \in X$ and $\tilde{x}^* \in A\tilde{x}$, the monotonicity of A implies that

$$0 \leq (x_n^* - \tilde{x}^*, x_n - \tilde{x})_X \rightarrow (x^* - \tilde{x}^*, x - \tilde{x})_X$$

since the duality pairing and hence the inner product of weakly and strongly converging sequences is convergent. Since A is maximally monotone, we obtain that $x^* \in Ax$. \square

Of central importance to the theory of monotone operators is *Minty's theorem*, which states that a monotone operator A is maximally monotone if and only if $\text{Id} + A$ is surjective. As a preparation, we first prove an important partial result.

Lemma 6.3. *Let $F : X \rightarrow \overline{\mathbb{R}}$ be proper, convex and lower semi-continuous. Then $\text{Id} + \partial F$ is surjective.*

Proof. We consider for given $z \in X$ the functional

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

which is proper, (strictly) convex and lower semi-continuous by the assumptions on F . Furthermore, J is coercive by [Lemma 3.8](#). [Theorem 3.7](#) thus yields a (unique) $\bar{x} \in X$ with $J(\bar{x}) = \min_{x \in X} J(x)$, which by [Theorems 4.3](#), [4.4](#) and [4.8](#) satisfies that

$$0 \in \partial J(\bar{x}) = \{\bar{x} - z\} + \partial F(\bar{x}),$$

i.e., $z \in \{\bar{x}\} + \partial F(\bar{x}) = (\text{Id} + \partial F)(\bar{x})$. \square

We now turn to the general case.

Theorem 6.4 (Minty). *Let $A : X \rightrightarrows X$ be monotone. Then, A is maximally monotone if and only if $\text{Id} + A$ is surjective.*

Proof. First, assume that $\text{Id} + A$ is surjective, and consider $x \in X$ and $x^* \in X$ with

$$(6.3) \quad (x^* - \tilde{x}^*, x - \tilde{x})_X \geq 0 \quad \text{for all } (\tilde{x}, \tilde{x}^*) \in \text{graph } A.$$

The assumption now implies that for $x + x^* \in X$, there exist a $z \in X$ and a $z^* \in Az$ with

$$(6.4) \quad x + x^* = z + z^* \in (\text{Id} + A)z.$$

Inserting $(\tilde{x}, \tilde{x}^*) = (z, z^*)$ into (6.3) then yields that

$$0 \leq (x^* - z^*, x - z)_X = (z - x, x - z)_X = -\|x - z\|_X^2 \leq 0,$$

i.e., $x = z$. From (6.4) we further obtain $x^* = z^* \in Az = Ax$, and hence A is maximally monotone.

The proof of the converse implication is significantly more involved. The special case $A = \partial F$ for a convex functional F was already shown in Lemma 6.3; for the general case F_A we proceed similarly by constructing a functional that plays the same role for A as F does for ∂F . Specifically, we define for a maximally monotone operator $A : X \rightrightarrows X$ the *Fitzpatrick functional*

$$(6.5) \quad F_A : X \times X \rightarrow [-\infty, \infty], \quad (x, y) \mapsto \sup_{(z, w) \in \text{graph } A} ((z, y)_X + (x, w)_X - (z, w)_X),$$

which can be written equivalently as

$$(6.6) \quad F_A(x, y) = (x, y)_X - \inf_{(z, w) \in \text{graph } A} (x - z, y - w)_X.$$

Each characterization implies useful properties.

- (i) By maximal monotonicity of A , we have by definition that $(x - z, y - w)_X \geq 0$ for all $(z, w) \in \text{graph } A$ if and only if $(x, y) \in \text{graph } A$; in particular, $(x - z, y - w)_X < 0$ for all $(x, y) \notin \text{graph } A$. Hence, (6.6) implies that $F_A(x, y) \geq (x, y)_X$, with equality if and only if $(x, y) \in \text{graph } A$ (since in this case the infimum is attained in $(z, w) = (x, y)$). In particular, F_A is proper.
- (ii) On the other hand, the definition (6.5) yields that

$$F_A = (G_A)^* \quad \text{for } G_A(w, z) = (w, z)_X + \delta_{\text{graph } A^{-1}}(w, z)$$

(since $(z, w) \in \text{graph } A$ if and only if $(w, z) \in \text{graph } A^{-1}$). As part of the monotonicity of A , we have required that $\text{graph } A \neq \emptyset$; hence F_A is the Fenchel conjugate of a proper functional and therefore convex and lower semi-continuous.

As a first step, we now show the result for the special case $z = 0$, i.e., that $0 \in \text{ran}(\text{Id} + A)$. We now set $Z := X \times X$ as well as $\xi := (x, y)$ and consider the functional

$$J_A : Z \rightarrow \overline{\mathbb{R}}, \quad \xi \mapsto F_A(\xi) + \frac{1}{2} \|\xi\|_Z^2.$$

We first note that property (i) implies for all $\xi \in Z$ that

$$(6.7) \quad \begin{aligned} J_A(\xi) &= F_A(\xi) + \frac{1}{2} \|\xi\|_Z^2 = F_A(x, y) + \frac{1}{2} \|x\|_X^2 + \frac{1}{2} \|y\|_X^2 \\ &\geq (x, y)_X + \frac{1}{2} \|x\|_X^2 + \frac{1}{2} \|y\|_X^2 \\ &\geq 0. \end{aligned}$$

Furthermore, J_A is proper, (strictly) convex, lower semi-continuous, and (by Lemma 3.8) coercive. Theorem 3.7 thus yields a (unique) $\bar{\xi} := (\bar{x}, \bar{y}) \in Z$ with $J_A(\bar{\xi}) = \min_{\xi \in Z} J_A(\xi)$, which by Theorems 4.3, 4.4 and 4.8 satisfies that

$$0 \in \partial J_A(\bar{\xi}) = \{\bar{\xi}\} + \partial F_A(\bar{\xi}),$$

i.e., $-\bar{\xi} \in \partial F_A(\bar{\xi})$. By definition of the subdifferential, we thus have for all $\xi \in Z$ that

$$\begin{aligned} F_A(\xi) &\geq F_A(\bar{\xi}) + (-\bar{\xi}, \xi - \bar{\xi})_Z = J_A(\bar{\xi}) + \frac{1}{2} \|\bar{\xi}\|_Z^2 + (-\bar{\xi}, \xi)_Z \\ &\geq \frac{1}{2} \|\bar{\xi}\|_Z^2 + (-\bar{\xi}, \xi)_Z, \end{aligned}$$

where the last step follows from (6.7). For the sake of presentation, we will replace $\bar{\xi} \mapsto -\bar{\xi}$ from now on; property (i) then implies for all $(x, y) \in \text{graph } A$ that

$$\begin{aligned} (6.8) \quad (x, y)_X &= F_A(x, y) \geq \frac{1}{2} \|\bar{x}\|_X^2 + (\bar{x}, x)_X + \frac{1}{2} \|\bar{y}\|_X^2 + (\bar{y}, y)_X \\ &\geq -(\bar{x}, \bar{y})_X + (\bar{x}, x)_X + (\bar{y}, y)_X, \end{aligned}$$

and hence $(y - \bar{x}, x - \bar{y})_X \geq 0$. The maximal monotonicity of A thus yields that $\bar{x} \in A\bar{y}$, i.e., $(\bar{y}, \bar{x}) \in \text{graph } A$. Inserting this into the first inequality of (6.8) then implies that

$$(\bar{y}, \bar{x})_X \geq \frac{1}{2} \|\bar{x}\|_X^2 + (\bar{x}, \bar{y})_X + \frac{1}{2} \|\bar{y}\|_X^2 + (\bar{y}, \bar{x})_X = \frac{1}{2} \|\bar{x} + \bar{y}\|_X^2 + (\bar{y}, \bar{x})_X \geq (\bar{y}, \bar{x})_X$$

and hence $\|\bar{x} + \bar{y}\|_X = 0$, i.e., $0 = \bar{y} + \bar{x} \in (\text{Id} + A)(\bar{y})$.

Finally, let $z \in X$ be arbitrary and set $B : X \rightrightarrows X$, $x \mapsto \{-z\} + Ax$. Using the definition, it is straightforward to verify that B is maximally monotone as well. As we have just shown, there now exists a $\bar{y} \in X$ with $0 \in (\text{Id} + B)(\bar{y}) = \{\bar{y}\} + \{-z\} + A\bar{y}$, i.e., $z \in (\text{Id} + A)(\bar{y})$. Hence $\text{Id} + A$ is surjective. \square

Together with Lemma 6.3, this yields the maximal monotonicity of convex subdifferentials.

Corollary 6.5. *Let $F : X \rightarrow \overline{\mathbb{R}}$ be proper, convex, and lower semi-continuous. Then $\partial F : X \rightrightarrows X$ is maximally monotone.*

6.2 RESOLVENTS AND PROXIMAL POINTS

We know from Lemma 6.3 that $\text{Id} + \partial F$ is surjective for any proper, convex, and lower semi-continuous functional F ; the proof even shows that each preimage is unique. Hence $(\text{Id} + \partial F)^{-1}$ is single-valued even if ∂F is not. We can thus hope to use this object instead of a subdifferential – or, more generally, a maximally monotone operator – for algorithms.

We thus define for a maximally monotone operator $A : X \rightrightarrows X$ the *resolvent*

$$\mathcal{R}_A : X \rightrightarrows X, \quad \mathcal{R}_A(x) = (\text{Id} + A)^{-1}x,$$

as well as for a proper, convex, and lower semi-continuous functional $F : X \rightarrow \overline{\mathbb{R}}$ the *proximal point mapping*

$$(6.9) \quad \text{prox}_F : X \rightarrow X, \quad \text{prox}_F(x) = \arg \min_{z \in X} \frac{1}{2} \|z - x\|_X^2 + F(z).$$

Since $w \in \mathcal{R}_{\partial F}(x)$ are the necessary and sufficient conditions for the *proximal point* w to be a minimizer of the strictly convex functional in (6.9), we have that

$$(6.10) \quad \text{prox}_F = (\text{Id} + \partial F)^{-1} = \mathcal{R}_{\partial F}.$$

It remains to show that the resolvent of arbitrary maximally monotone operators is single-valued on X as well and we can thus write $\mathcal{R}_A : X \rightarrow X$.

Lemma 6.6. *Let $A : X \rightrightarrows X$ be maximally monotone. Then $\mathcal{R}_A : X \rightarrow X$.*

Proof. Since A is maximally monotone, $\text{Id} + A$ is surjective, which implies that $\text{dom } \mathcal{R}_A = X$. Let now $x, z \in X$ with $x^* \in \mathcal{R}_A(x)$ and $z^* \in \mathcal{R}_A(z)$, i.e., $x \in \{x^*\} + Ax^*$ and $z \in \{z^*\} + Az^*$. For $x - x^* \in Ax^*$ and $z - z^* \in Az^*$, the monotonicity of A implies that

$$(6.11) \quad \|x^* - z^*\|_X^2 \leq (x - z, x^* - z^*)_X.$$

Hence $x = z$ implies $x^* = z^*$, i.e., \mathcal{R}_A is single-valued. \square

The inequality (6.11) together with the Cauchy–Schwarz inequality shows that resolvents are Lipschitz continuous with constant $L = 1$; such mappings are called *nonexpansive*. Since (6.11) is in fact a stronger property, a mapping $T : X \rightarrow X$ is called *firmly nonexpansive* if it satisfies this inequality, i.e., if

$$\|Tx - Tz\|_X^2 \leq (Tx - Tz, x - z)_X \quad \text{for all } x, z \in X.$$

Firm nonexpansivity implies another useful inequality.

Lemma 6.7. *Let $A : X \rightrightarrows X$ be maximally monotone. Then $\mathcal{R}_A : X \rightarrow X$ is firmly nonexpansive and satisfies that*

$$\|\mathcal{R}_A x - \mathcal{R}_A z\|_X^2 + \|(\text{Id} - \mathcal{R}_A)x - (\text{Id} - \mathcal{R}_A)z\|_X^2 \leq \|x - z\|_X^2 \quad \text{for all } x, z \in X.$$

Proof. Firm nonexpansivity of \mathcal{R}_A was already shown in (6.11), which further implies that

$$\begin{aligned} \|(\text{Id} - \mathcal{R}_A)x - (\text{Id} - \mathcal{R}_A)z\|_X^2 &= \|x - z\|_X^2 - 2(x - z, \mathcal{R}_A x - \mathcal{R}_A z)_X + \|\mathcal{R}_A x - \mathcal{R}_A z\|_X^2 \\ &\leq \|x - z\|_X^2 - \|\mathcal{R}_A x - \mathcal{R}_A z\|_X^2. \end{aligned} \quad \square$$

Corollary 6.8. *Let $F : X \rightarrow \overline{\mathbb{R}}$ be proper, convex, and lower semi-continuous. Then $\text{prox}_F : X \rightarrow X$ is Lipschitz continuous with constant $L = 1$.*

The following useful result allows characterizing minimizers of convex functionals as proximal points.

Lemma 6.9. *Let $F : X \rightarrow \overline{\mathbb{R}}$ be proper, convex, and lower semi-continuous, and $x, x^* \in X$. Then for any $\gamma > 0$,*

$$x^* \in \partial F(x) \Leftrightarrow x = \text{prox}_{\gamma F}(x + \gamma x^*).$$

Proof. Multiplying both sides of the subdifferential inclusion by $\gamma > 0$ and adding x yields that

$$\begin{aligned} x^* \in \partial F(x) &\Leftrightarrow x + \gamma x^* \in (\text{Id} + \gamma \partial F)(x) \\ &\Leftrightarrow x \in (\text{Id} + \gamma \partial F)^{-1}(x + \gamma x^*) \\ &\Leftrightarrow x = \text{prox}_{\gamma F}(x + \gamma x^*), \end{aligned}$$

where in the last step we have used that $\gamma \partial F = \partial(\gamma F)$ by [Lemma 4.7](#) (ii) and hence that $\text{prox}_{\gamma F} = \mathcal{R}_{\gamma \partial F}$. \square

Corollary 6.10. *Let $F : X \rightarrow \overline{\mathbb{R}}$ be proper, convex and lower semi-continuous, and $\gamma > 0$ be arbitrary. Then $\bar{x} \in \text{dom } F$ is a minimizer of F if and only if*

$$\bar{x} = \text{prox}_{\gamma F}(\bar{x}).$$

Proof. Simply apply [Lemma 6.9](#) to the Fermat principle $0 \in \partial F(\bar{x})$. \square

This simple result should not be underestimated: It allows replacing (explicit) set inclusions by (implicit) Lipschitz continuous equations in optimality conditions, thus opening the door to fixed point iterations or Newton methods.

We can also derive a generalization of the orthogonal decomposition of vector spaces.

Theorem 6.11 (Moreau decomposition). *Let $F : X \rightarrow \overline{\mathbb{R}}$ be proper, convex, and lower semi-continuous. Then we have for all $x \in X$ that*

$$x = \text{prox}_F(x) + \text{prox}_{F^*}(x).$$

Proof. Setting $w = \text{prox}_F(x)$, [Lemmas 5.5](#) and [6.9](#) imply that

$$\begin{aligned} w = \text{prox}_F(x) = \text{prox}_F(w + (x - w)) &\Leftrightarrow x - w \in \partial F(w) \\ &\Leftrightarrow w \in \partial F^*(x - w) \\ &\Leftrightarrow x - w = \text{prox}_{F^*}((x - w) + w) = \text{prox}_{F^*}(x). \quad \square \end{aligned}$$

The following calculus rules will prove useful.

Lemma 6.12. *Let $F : X \rightarrow \overline{\mathbb{R}}$ be proper, convex, and lower semi-continuous. Then,*

(i) *for $\lambda \neq 0$ and $z \in X$ we have with $H(x) := F(\lambda x + z)$ that*

$$\text{prox}_H(x) = \lambda^{-1}(\text{prox}_{\lambda^2 F}(\lambda x + z) - z);$$

(ii) *for $\gamma > 0$ we have that*

$$\text{prox}_{\gamma F^*}(x) = x - \gamma \text{prox}_{\gamma^{-1} F}(\gamma^{-1} x);$$

(iii) *for proper, convex, lower semi-continuous $G : Y \rightarrow \overline{\mathbb{R}}$ and $\gamma > 0$ we have with $H(x, y) := F(x) + G(y)$ that*

$$\text{prox}_{\gamma H}(x, y) = \begin{pmatrix} \text{prox}_{\gamma F}(x) \\ \text{prox}_{\gamma G}(y) \end{pmatrix}.$$

Proof. Ad (i): By definition,

$$\text{prox}_H(x) = \arg \min_{w \in X} \frac{1}{2} \|w - x\|_X^2 + F(\lambda w + z) =: \bar{w}.$$

Substituting $v = \lambda w + z$ and using that $\min_{w \in X} G(w) = \min_{v \in X} G(v)$ for all G implies that

$$\begin{aligned} \bar{v} &= \arg \min_{v \in X} \frac{1}{2} \|\lambda^{-1}(v - z) - x\|_X^2 + F(v) \\ &= \arg \min_{v \in X} \frac{1}{2\lambda^2} \|v - (\lambda x + z)\|_X^2 + F(v) \\ &= \arg \min_{v \in X} \frac{1}{2} \|v - (\lambda x + z)\|_X^2 + \lambda^2 F(v) \\ &= \text{prox}_{\lambda^2 F}(\lambda x + z). \end{aligned}$$

Hence, $\bar{w} := \lambda^{-1}(\bar{v} - z)$ is the desired minimizer.

Ad (ii): [Theorem 6.11](#), [Lemma 5.3](#) (i), and (i) for $\lambda = \gamma^{-1}$ and $z = 0$ together imply that

$$\begin{aligned} \text{prox}_{\gamma F^*}(x) &= x - \text{prox}_{(\gamma F)^*}(x) \\ &= x - \text{prox}_{\gamma F^* \circ (\gamma^{-1} \text{Id})}(x) \\ &= x - \gamma \text{prox}_{\gamma(\gamma^{-2} F^*)}(\gamma^{-1} x). \end{aligned}$$

Applying this to F^* and using that $F^{**} = F$ now yields the claim.

Ad (iii): By definition of the norm on the product space $X \times Y$, we have that

$$\begin{aligned} \text{prox}_{\gamma H}(x, y) &= \arg \min_{(u, v) \in X \times Y} \frac{1}{2} \|(u, v) - (x, y)\|_{X \times Y}^2 + \gamma H(u, v) \\ &= \arg \min_{u \in X, v \in Y} \left(\frac{1}{2} \|u - x\|_X^2 + \gamma F(u) \right) + \left(\frac{1}{2} \|v - y\|_Y^2 + \gamma G(v) \right). \end{aligned}$$

Since there are no mixed terms in u and v , the two terms in parentheses can be minimized separately. Hence, $\text{prox}_{\gamma H}(x, y) = (\bar{u}, \bar{v})$ for

$$\begin{aligned}\bar{u} &= \arg \min_{u \in X} \frac{1}{2} \|u - x\|_X^2 + \gamma F(u) = \text{prox}_{\gamma F(x)}, \\ \bar{v} &= \arg \min_{v \in Y} \frac{1}{2} \|v - y\|_Y^2 + \gamma G(v) = \text{prox}_{\gamma G(x)}.\end{aligned}\quad \square$$

Computing proximal points is difficult in general since evaluating $\text{prox}F$ by its definition entails minimizing F . In some cases, however, it is possible to give an explicit formula for $\text{prox}F$.

Example 6.13. We first consider scalar functions $f : \mathbb{R} \rightarrow \overline{\mathbb{R}}$.

- (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 $\text{prox}_{\gamma f}(t) = (1 + \gamma)^{-1}t$.
- (ii) $f(t) = |t|$. By (4.3) we have that $\partial f(t) = \text{sign}(t)$; hence $s := \text{prox}_{\gamma f}(t) = (\text{Id} + \gamma \text{sign})^{-1}(t)$ if and only if $t \in \{s\} + \gamma \text{sign}(s)$. 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

$$\text{prox}_{\gamma f}(t) = \begin{cases} t - \gamma & \text{if } t > \gamma, \\ 0 & \text{if } t \in [-\gamma, \gamma], \\ t + \gamma & \text{if } t < -\gamma. \end{cases}$$

This mapping is also known as the *soft-shrinkage* or *soft-thresholding* operator.

- (iii) $f(t) = \delta_{[-1, 1]}(t)$. By Example 5.2 (iii) we have that $f^*(t) = |t|$, and hence Lemma 6.12 (ii) yields that

$$\begin{aligned}\text{prox}_{\gamma f}(t) &= t - \gamma \text{prox}_{\gamma^{-1}f^*}(\gamma^{-1}t) \\ &= \begin{cases} t - \gamma(\gamma^{-1}t - \gamma^{-1}) & \text{if } \gamma^{-1}t > \gamma^{-1}, \\ t - 0 & \text{if } \gamma^{-1}t \in [-\gamma^{-1}, \gamma^{-1}], \\ t - \gamma(\gamma^{-1}t + \gamma^{-1}) & \text{if } \gamma^{-1}t < -\gamma^{-1} \end{cases} \\ &= \begin{cases} 1 & \text{if } t > 1, \\ t & \text{if } t \in [-1, 1], \\ -1 & \text{if } t < -1. \end{cases}\end{aligned}$$

For every $\gamma > 0$, the proximal point of t is thus its projection onto $[-1, 1]$.

Example 6.14. We can generalize [Example 6.13](#) to $X = \mathbb{R}^N$ (endowed with the Euclidean inner product) by applying [Lemma 6.12](#) (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

$$[\text{prox}_{\gamma F}(x)]_i = \left(\frac{1}{1+\gamma}\right)x_i, \quad 1 \leq i \leq N;$$

(ii) for $F(x) = \|x\|_1 = \sum_{i=1}^N |x_i|$ that

$$[\text{prox}_{\gamma F}(x)]_i = (|x_i| - \gamma)^+ \text{sign}(x_i), \quad 1 \leq i \leq N;$$

(iii) for $F(x) = \delta_{B_\infty}(x) = \sum_{i=1}^N \delta_{[-1,1]}(x_i)$ that

$$[\text{prox}_{\gamma F}(x)]_i = x_i - (x_i - 1)^+ - (x_i + 1)^- = \frac{x_i}{\max\{1, |x_i|\}}, \quad 1 \leq i \leq N.$$

Here we have used the convenient notation $(t)^+ := \max\{t, 0\}$ and $(t)^- := \min\{t, 0\}$.

Many more examples can be found in [[Parikh & Boyd 2014](#), § 6.5].

Since the subdifferential of convex integral functionals can be evaluated pointwise by [Lemma 4.6](#), the same holds for the definition (6.10) of the proximal point mapping.

Corollary 6.15. Let $f : \mathbb{R} \rightarrow \overline{\mathbb{R}}$ be proper, convex, and lower semi-continuous, and $F : L^2(\Omega) \rightarrow \overline{\mathbb{R}}$ be defined as in [Lemma 3.4](#). Then we have for all $\gamma > 0$ and $u \in L^2(\Omega)$ that

$$[\text{prox}_{\gamma F}(u)](x) = \text{prox}_{\gamma f}(u(x)) \quad \text{for almost every } x \in \Omega.$$

Example 6.16. Let X be a Hilbert space. Similarly to [Example 6.13](#) one can show that

(i) for $F = \frac{1}{2}\|\cdot\|_X^2 = \frac{1}{2}(\cdot, \cdot)_X$, that

$$\text{prox}_{\gamma F}(x) = \left(\frac{1}{1+\gamma}\right)x;$$

(ii) for $F = \|\cdot\|_X$, using a case distinction as in [Theorem 4.5](#), that

$$\text{prox}_{\gamma F}(x) = \left(1 - \frac{\gamma}{\|x\|_X}\right)^+ x;$$

(iii) for $F = \delta_C$ with $C \subset X$ nonempty, convex, and closed, that by definition

$$\text{prox}_{\gamma F}(x) = \text{proj}_C(x) := \arg \min_{z \in C} \|z - x\|_X$$

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].

6.3 MOREAU–YOSIDA REGULARIZATION

Before we turn to algorithms for the minimization of convex functionals, we will 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 9](#).

Let $A : X \rightrightarrows X$ be a maximally monotone operator and $\gamma > 0$. Then we define the *Yosida approximation* of A as

$$A_\gamma := \frac{1}{\gamma} (\text{Id} - \mathcal{R}_{\gamma A}).$$

In particular, the Yosida approximation of the subdifferential of a proper, convex, and lower semi-continuous functional $F : X \rightarrow \overline{\mathbb{R}}$ is given by

$$(\partial F)_\gamma := \frac{1}{\gamma} (\text{Id} - \text{prox}_{\gamma F}),$$

which by [Corollary 6.8](#) is always Lipschitz continuous with constant $L = \gamma^{-1}$.

An alternative point of view is the following. For a proper, convex, and lower semi-continuous functional $F : X \rightarrow \overline{\mathbb{R}}$ and $\gamma > 0$, we define the *Moreau envelope*¹ as

$$F_\gamma : X \rightarrow \mathbb{R}, \quad x \mapsto \inf_{z \in X} \frac{1}{2\gamma} \|z - x\|_X^2 + F(z).$$

Comparing this with the definition (6.9) of the proximal point mapping of F , we see that

$$(6.12) \quad F_\gamma(x) = \frac{1}{2\gamma} \|\text{prox}_{\gamma F}(x) - x\|_X^2 + F(\text{prox}_{\gamma F}(x)).$$

(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 (6.12) that F_γ is convex as well.

Lemma 6.17. *Let $F : X \rightarrow \overline{\mathbb{R}}$ be proper, convex, and lower semi-continuous, and $\gamma > 0$. Then F_γ is convex.*

Proof. We first show that for any convex $G : X \rightarrow \overline{\mathbb{R}}$, the mapping

$$H : X \times X \rightarrow \overline{\mathbb{R}}, \quad (x, z) \mapsto F(z) + G(z - x)$$

is convex as well. Indeed, for any $(x_1, z_1), (x_2, z_2) \in X \times X$ and $\lambda \in [0, 1]$, convexity of F and G implies that

$$\begin{aligned} H(\lambda(x_1, z_1) + (1 - \lambda)(x_2, z_2)) &= F(\lambda z_1 + (1 - \lambda)z_2) + G(\lambda(z_1 - x_1) + (1 - \lambda)(z_2 - x_2)) \\ &\leq \lambda(F(z_1) + G(z_1 - x_1)) + (1 - \lambda)(F(z_2) + G(z_2 - x_2)) \\ &= \lambda H(x_1, z_1) + (1 - \lambda)H(x_2, z_2). \end{aligned}$$

¹not 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

$$H(x_1, z_n^1) \rightarrow F_\gamma(x_1), \quad H(x_2, z_n^2) \rightarrow F_\gamma(x_2).$$

From the definition of the infimum together with the convexity of H , we thus obtain for all $n \in \mathbb{N}$ that

$$\begin{aligned} F_\gamma(\lambda x_1 + (1 - \lambda)x_2) &\leq H(\lambda(x_1, z_n^1) + (1 - \lambda)(x_2, z_n^2)) \\ &\leq \lambda H(x_1, z_n^1) + (1 - \lambda)H(x_2, z_n^2), \end{aligned}$$

and passing to the limit $n \rightarrow \infty$ yields the desired convexity. \square

The next theorem links the two concepts and hence justifies the term *Moreau–Yosida regularization*.

Theorem 6.18. *Let $F : X \rightarrow \overline{\mathbb{R}}$ be proper, convex, and lower semi-continuous, and $\gamma > 0$. Then F_γ is Fréchet differentiable and satisfies that*

$$\nabla(F_\gamma) = (\partial F)_\gamma.$$

Proof. Let $x, y \in X$ be arbitrary and set $x^* = \text{prox}_{\gamma F}(x)$ and $y^* = \text{prox}_{\gamma F}(y)$. We first show that

$$(6.13) \quad \frac{1}{\gamma} (y^* - x^*, x - x^*)_X \leq F(y^*) - F(x^*).$$

(Note that for proper F , the definition of proximal points as minimizers necessarily implies that $x^*, y^* \in \text{dom } F$.) To this purpose, consider for $t \in (0, 1)$ the point $x_t^* := tx^* + (1 - t)x^*$. Using the minimizing property of the proximal point x^* together with the convexity of F and completing the square, we obtain that

$$\begin{aligned} F(x^*) &\leq F(x_t^*) + \frac{1}{2\gamma} \|x_t^* - x\|_X^2 - \frac{1}{2\gamma} \|x^* - x\|_X^2 \\ &\leq tF(y^*) + (1 - t)F(x^*) - \frac{t}{\gamma} (x - x^*, y^* - x^*)_X + \frac{t^2}{2\gamma} \|x^* - y^*\|_X^2. \end{aligned}$$

Rearranging the terms, dividing by $t > 0$ and passing to the limit $t \rightarrow 0$ then yields (6.13). Combining this with (6.12) implies that

$$\begin{aligned} F_\gamma(y) - F_\gamma(x) &= F(y^*) - F(x^*) + \frac{1}{2\gamma} (\|y - y^*\|_X^2 - \|x - x^*\|_X^2) \\ &\geq \frac{1}{2\gamma} (2(y^* - x^*, x - x^*)_X + \|y - y^*\|_X^2 - \|x - x^*\|_X^2) \\ &= \frac{1}{2\gamma} (2(y - x, x - x^*)_X + \|y - y^* - x + x^*\|_X^2) \\ &\geq \frac{1}{\gamma} (y - x, x - x^*)_X. \end{aligned}$$

By exchanging the roles of x^* and y^* in (6.13), we obtain that

$$F_Y(y) - F_Y(x) \leq \frac{1}{\gamma} (y - x, y - y^*)_X.$$

Together, these two inequalities yield that

$$\begin{aligned} 0 &\leq F_Y(y) - F_Y(x) - \frac{1}{\gamma} (y - x, x - x^*)_X \\ &\leq \frac{1}{\gamma} (y - x, (y - y^*) - (x - x^*))_X \\ &\leq \frac{1}{\gamma} (\|y - x\|_X^2 - \|y^* - x^*\|_X^2) \\ &\leq \frac{1}{\gamma} \|y - x\|_X^2, \end{aligned}$$

where the next-to-last inequality follows from the firm nonexpansivity of proximal point mappings (Lemma 6.7).

If we now set $y = x + h$ for arbitrary $h \in X$, we obtain that

$$0 \leq \frac{F_Y(x + h) - F_Y(x) - (\gamma^{-1}(x - x^*), h)_X}{\|h\|_X} \leq \frac{1}{\gamma} \|h\|_X \rightarrow 0 \quad \text{for } h \rightarrow 0,$$

i.e., F_Y is Fréchet differentiable with gradient $\frac{1}{\gamma}(x - x^*) = (\partial F)_Y$. \square

Since F_Y is convex by Lemma 6.17, this result together with Theorem 4.4 yields the catchy relation $\partial(F_Y) = (\partial F)_Y$.

Example 6.19. We consider again $X = \mathbb{R}^N$.

- (i) For $F(x) = \|x\|_1$, we have from Example 6.14 (ii) that the proximal point mapping is given by the component-wise soft-shrinkage operator. Inserting this into the definition yields that

$$[(\partial \|\cdot\|_1)_Y(x)]_i = \begin{cases} \frac{1}{\gamma}(x_i - (x_i - \gamma)) = 1 & \text{if } x_i > \gamma, \\ \frac{1}{\gamma}x_i & \text{if } x_i \in [-\gamma, \gamma], \\ \frac{1}{\gamma}(x_i - (x_i + \gamma)) = -1 & \text{if } x_i < -\gamma. \end{cases}$$

Comparing this to the corresponding subdifferential (4.3), 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 (6.12) shows that

$$F_Y(x) = \sum_{i=1}^N f_Y(x_i) \quad \text{for } f_Y(t) := \begin{cases} \frac{1}{2\gamma}|t - (t - \gamma)|^2 + |t - \gamma| = t - \frac{\gamma}{2} & \text{if } t > \gamma, \\ \frac{1}{2\gamma}|t|^2 & \text{if } t \in [-\gamma, \gamma], \\ \frac{1}{2\gamma}|t - (t + \gamma)|^2 + |t + \gamma| = -t + \frac{\gamma}{2} & \text{if } t < -\gamma. \end{cases}$$

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*.

- (ii) For $F(x) = \delta_{B_\infty}(x)$, we have from [Example 6.14](#) (iii) that the proximal mapping is given by the component-wise projection onto $[-1, 1]$ and hence that

$$\left[(\partial \delta_{B_\infty})_\gamma(x) \right]_i = \frac{1}{\gamma} \left(x_i - (x_i - (x_i - 1)^+ - (x_i + 1)^-) \right) = \frac{1}{\gamma} (x_i - 1)^+ + \frac{1}{\gamma} (x_i + 1)^-.$$

Similarly, inserting this and using that $\text{prox}_{\gamma F}(x) \in B_\infty$ and $((x + 1)^+, (x - 1)^-)_X = 0$ yields that

$$(\delta_{B_\infty})_\gamma(x) = \frac{1}{2\gamma} \|(x - 1)^+\|_2^2 + \frac{1}{2\gamma} \|(x + 1)^-\|_2^2,$$

which corresponds to the classical penalty functional for the inequality constraints $x - 1 \leq 0$ and $x + 1 \geq 0$ in nonlinear optimization.

A further connection exists between the Moreau envelope and the Fenchel conjugate.

Theorem 6.20. *Let $F : X \rightarrow \overline{\mathbb{R}}$ be proper, convex, and lower semi-continuous. 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

$$\begin{aligned} (F_\gamma)^*(x^*) &= \sup_{x \in X} \left((x^*, x)_X - \inf_{z \in X} \left(\frac{1}{2\gamma} \|x - z\|_X^2 + F(z) \right) \right) \\ &= \sup_{x \in X} \left((x^*, x)_X + \sup_{z \in X} \left(-\frac{1}{2\gamma} \|x - z\|_X^2 - F(z) \right) \right) \\ &= \sup_{z \in X} \left((x^*, z)_X - F(z) + \sup_{x \in X} \left((x^*, x - z)_X - \frac{1}{2\gamma} \|x - z\|_X^2 \right) \right) \\ &= F^*(x^*) + \left(\frac{1}{2\gamma} \|\cdot\|_X^2 \right)^*(x^*), \end{aligned}$$

since for any given $z \in X$, the inner supremum is always taken over the full space X . The claim now follows from [Example 5.2](#) (i) and [Lemma 5.3](#) (i). \square

We briefly sketch the relevance for nonsmooth optimization. For a convex functional $F : X \rightarrow \overline{\mathbb{R}}$, 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

$$x_\gamma = (\partial F^*)_\gamma(0) = -\frac{1}{\gamma} \text{prox}_{F^*}(0).$$

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

$$0 \in \partial(F_\gamma^*)^*(x_\gamma) = \partial \left(F^{**} + \frac{\gamma}{2} \|\cdot\|_X^2 \right) (x_\gamma) = \partial \left(F + \frac{\gamma}{2} \|\cdot\|_X^2 \right) (x_\gamma),$$

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 $x_\gamma \rightarrow \bar{x}$ for $\gamma \rightarrow 0$. In practice, this straightforward approach fails due to the difficulty of computing F^* and prox_{F^*} and is therefore usually combined with one of the splitting techniques which will be introduced in the next chapter.

7 PROXIMAL POINT AND SPLITTING METHODS

We now turn to algorithms for computing minimizers of functionals $J : X \rightarrow \overline{\mathbb{R}}$ of the form

$$J(x) := F(x) + G(x)$$

for $F, G : X \rightarrow \overline{\mathbb{R}}$ 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., [Ruszczynski 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 of the set-valued mapping $x \mapsto \partial J(x) \subset X^* \cong X$.

7.1 PROXIMAL POINT METHOD

We have seen in Corollary 6.10 that a root \bar{x} of $\partial F : X \rightrightarrows X$ can be characterized as a fixed point of $\text{prox}_{\gamma F}$ for any $\gamma > 0$. This suggests a fixed-point iteration: Choose $x^0 \in X$ and set for an appropriate sequence $\{\gamma_k\}_{k \in \mathbb{N}}$

$$(7.1) \quad x^{k+1} = \text{prox}_{\gamma_k F}(x^k).$$

To show convergence of this iteration, we have to show as usual that the fixed-point mapping is contracting in a suitable sense. As we will see, firm nonexpansivity will be sufficient, which by Corollary 6.8 is always the case for resolvents of maximally monotone operators (and hence in particular for proximal mappings of convex functionals). For later use, we treat the general version of (7.1) for arbitrary maximally monotone operators.

Theorem 7.1. *Let $A : X \rightrightarrows X$ be maximally monotone with root $x^* \in X$, and let $\{\gamma_k\}_{k \in \mathbb{N}} \subset (0, \infty)$ with $\sum_{k=0}^{\infty} \gamma_k^2 = \infty$. If $\{x^k\}_{k \in \mathbb{N}} \subset X$ is given by the iteration*

$$x^{k+1} = \mathcal{R}_{\gamma_k A} x^k,$$

then $x^k \rightarrow \bar{x}$ with $0 \in A\bar{x}$.

Proof. The iteration $x^{k+1} = \mathcal{R}_{\gamma_k A} x^k = (\text{Id} + \gamma_k A)^{-1} x^k$ implies that

$$w^k := \gamma_k^{-1}(x^k - x^{k+1}) \in Ax^{k+1}$$

and hence that $x^{k+1} - x^{k+2} = \gamma_{k+1} w^{k+1}$. (The vector w^k will play the role of a residual in the generalized equation $0 \in Ax$.) By monotonicity of A , we have for $\gamma_{k+1} > 0$ that

$$\begin{aligned} 0 &\leq \gamma_{k+1}^{-1} \left(w^k - w^{k+1}, x^{k+1} - x^{k+2} \right)_X \\ &= \left(w^k - w^{k+1}, w^{k+1} \right)_X \\ &= \left(w^k, w^{k+1} \right)_X - \|w^{k+1}\|_X^2 \\ &\leq \|w^{k+1}\|_X \left(\|w^k\|_X - \|w^{k+1}\|_X \right). \end{aligned}$$

Hence, the nonnegative sequence $\{\|w^k\|_X\}_{k \in \mathbb{N}} \subset \mathbb{R}$ is decreasing and hence convergent (as long as $w^{k+1} \neq 0$, but otherwise from $w^{k+1} \in Ax^{k+2}$ we immediately obtain that x^{k+2} is the desired root.)

Let now $x^* \in X$ be a root of A , i.e., $0 \in Ax^*$, which exists by assumption. As in the proof of [Corollary 6.10](#), we can then show that $x^* = \mathcal{R}_{\gamma A} x^*$ for all $\gamma > 0$. From [Lemma 6.7](#) together with $(\text{Id} - \mathcal{R}_{\gamma A})x^k = x^k - x^{k+1} = \gamma_k w^k$, we now obtain that

$$\begin{aligned} (7.2) \quad \|x^{k+1} - x^*\|_X^2 &= \|\mathcal{R}_{\gamma_k A} x^k - \mathcal{R}_{\gamma_k A} x^*\|_X^2 \\ &\leq \|x^k - x^*\|_X^2 - \|(\text{Id} - \mathcal{R}_{\gamma_k A})x^k - (\text{Id} - \mathcal{R}_{\gamma_k A})x^*\|_X^2 \\ &= \|x^k - x^*\|_X^2 - \gamma_k^2 \|w^k\|_X^2. \end{aligned}$$

Hence, $\{\|x^k - x^*\|_X\}_{k \in \mathbb{N}}$ is decreasing for *any* root x^* (such sequences are called *Féjer monotone*) and thus bounded. This implies that $\{x^k\}_{k \in \mathbb{N}} \subset X$ is bounded as well and thus contains a weakly convergent subsequence $x^{k_l} \rightharpoonup \bar{x}$.

Furthermore, recursive application of (7.2) yields that

$$0 \leq \|x^{k+1} - x^*\|_X^2 \leq \|x^0 - x^*\|_X^2 - \sum_{j=0}^k \gamma_j^2 \|w^j\|_X^2.$$

The (increasing) sequence of partial sums on the right-hand side is therefore bounded and hence $\sum_{k=0}^{\infty} \gamma_k^2 \|w^k\|_X^2$ converges. Since the sequence $\{\gamma_k^2\}_{k \in \mathbb{N}}$ is not summable by assumption, this requires that $\liminf_{k \rightarrow \infty} \|w^k\|_X^2 = 0$. This together with the convergence of $\{\|w^k\|_X\}_{k \in \mathbb{N}}$ implies that $w^k \rightarrow 0$. In particular, we have that $Ax^{k_l+1} \ni w^{k_l} \rightarrow 0$ for $x^{k_l+1} \rightharpoonup \bar{x}$, and the closedness of maximally monotone operators ([Lemma 6.2](#)) yields that $0 \in A\bar{x}$. Hence, every weak accumulation point of $\{x^k\}_{k \in \mathbb{N}}$ is a root of A .

We finally show convergence of the full sequence $\{x^k\}_{k \in \mathbb{N}}$. Let \bar{x} and \hat{x} be weak accumulation points and therefore roots of A . The Féjer monotonicity of $\{x^k\}_{k \in \mathbb{N}}$ 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 and therefore convergent. This implies that

$$\left(x^k, \bar{x} - \hat{x}\right)_X = \frac{1}{2} \left(\|x^k - \hat{x}\|_X^2 - \|x^k - \bar{x}\|_X^2 + \|\bar{x}\|_X^2 - \|\hat{x}\|_X^2 \right) \rightarrow c \in \mathbb{R}.$$

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 $x^{k_m} \rightarrow \hat{x}$. Hence,

$$(\bar{x}, \bar{x} - \hat{x})_X = \lim_{n \rightarrow \infty} \left(x^{k_n}, \bar{x} - \hat{x}\right)_X = c = \lim_{m \rightarrow \infty} \left(x^{k_m}, \bar{x} - \hat{x}\right)_X = (\hat{x}, \bar{x} - \hat{x})_X,$$

and therefore

$$0 = (\bar{x} - \hat{x}, \bar{x} - \hat{x})_X = \|\bar{x} - \hat{x}\|_X^2,$$

i.e., $\bar{x} = \hat{x}$. Every convergent subsequence thus has the same limit, which by a subsequence-subsequence argument must therefore be the limit of the full sequence $\{x^k\}_{k \in \mathbb{N}}$. \square

7.2 EXPLICIT SPLITTING

As we have repeatedly 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 (i.e., are *prox-simple*). We thus proceed differently: instead of applying the proximal point reformulation directly to $0 \in \partial J(\bar{x})$, we first apply the sum rule and obtain a $\bar{p} \in X$ with

$$(7.3) \quad \begin{cases} \bar{p} \in \partial F(\bar{x}), \\ -\bar{p} \in \partial G(\bar{x}). \end{cases}$$

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 6.9](#) only to the second inclusion in (7.3) to obtain

$$\begin{cases} \bar{p} \in \partial F(\bar{x}), \\ \bar{x} = \text{prox}_{\gamma G}(\bar{x} - \gamma \bar{p}). \end{cases}$$

The corresponding fixed-point iteration then consists in

1. choosing $p^k \in \partial F(x^k)$ (with minimal norm);
2. setting $x^{k+1} = \text{prox}_{\gamma_k G}(x^k - \gamma_k p^k)$.

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 $\partial F(x) = \{\nabla F(x)\}$. This leads to the *proximal gradient method*

$$(7.4) \quad x^{k+1} = \text{prox}_{\gamma_k G}(x^k - \gamma_k \nabla F(x^k)).$$

(The special case $G = \delta_C$ – i.e., $\text{prox}_{\gamma G}(x) = \text{proj}_C(x)$ – is also known as the *projected gradient method*).

Showing convergence of the proximal gradient method as for the proximal point method requires assuming Lipschitz continuity of the gradient (since we are not using a proximal point mapping for F which is always firmly nonexpansive and hence Lipschitz continuous). The following lemma may be familiar from nonlinear optimization.

Lemma 7.2. *Let $F : X \rightarrow \mathbb{R}$ be Gâteaux differentiable with Lipschitz continuous gradient. Then,*

$$F(y) \leq F(x) + (\nabla F(x), x - y)_X + \frac{L}{2} \|x - y\|_X^2 \quad \text{for all } x, y \in X.$$

Proof. The Gâteaux differentiability of F implies that

$$\frac{d}{dt} F(x + t(y - x)) = (\nabla F(x + t(y - x)), y - x)_X \quad \text{for all } x, y \in X,$$

and integration over all $t \in [0, 1]$ yields that

$$\int_0^1 (\nabla F(x + t(y - x)), y - x)_X dt = F(y) - F(x).$$

From this, we obtain together with the productive zero, the Cauchy–Schwarz inequality, and the Lipschitz continuity of the gradient that

$$\begin{aligned} F(y) &= F(x) + (\nabla F(x), y - x)_X + \int_0^1 (\nabla F(x + t(y - x)) - \nabla F(x), y - x)_X dt \\ &\leq F(x) + (\nabla F(x), y - x)_X + \int_0^1 \|\nabla F(x + t(y - x)) - \nabla F(x)\|_X \|x - y\|_X dt \\ &\leq F(x) + (\nabla F(x), y - x)_X + \int_0^1 Lt \|x - y\|_X^2 dt \\ &= F(x) + (\nabla F(x), y - x)_X + \frac{L}{2} \|x - y\|_X^2. \quad \square \end{aligned}$$

We can now show convergence of the proximal gradient method for sufficiently small step sizes.

Theorem 7.3. *Let $F : X \rightarrow \mathbb{R}$ and $G : X \rightarrow \overline{\mathbb{R}}$ be proper, convex, and lower semi-continuous. Furthermore, let F be Gâteaux differentiable with Lipschitz continuous gradient. If $0 < \gamma_{\min} \leq \gamma_k \leq L^{-1}$, the sequence generated by (7.4) converges weakly to a minimizer $\bar{x} \in X$ of J .*

Proof. We argue similarly as in the proof of [Theorem 7.1](#), replacing the monotonicity of the generalized residuals $w^k \in Ax^{k+1}$ with those of the functional values $J(x^k)$. For this purpose, we define the operator

$$T_\gamma : X \rightarrow X, \quad x \mapsto \gamma^{-1}(x - \text{prox}_{\gamma G}(x - \gamma \nabla F(x))),$$

which allows reformulating the iteration (7.4) as

$$x^{k+1} = \text{prox}_{\gamma_k G}(x^k - \gamma_k \nabla F(x^k)) = x^k - \gamma_k T_{\gamma_k}(x^k).$$

[Lemma 6.9](#) then implies that

$$(7.5) \quad T_{\gamma_k}(x^k) - \nabla F(x^k) \in \partial G(x^k - \gamma_k T_{\gamma_k}(x^k)).$$

[Lemma 7.2](#) with $x = x^k$, $y = x^{k+1} = x^k - \gamma_k T_{\gamma_k}(x^k)$, and $\gamma_k \leq L^{-1}$ further implies that

$$(7.6) \quad \begin{aligned} F(x^k - \gamma_k T_{\gamma_k}(x^k)) &\leq F(x^k) - \gamma_k \left(\nabla F(x^k), T_{\gamma_k}(x^k) \right)_X + \frac{\gamma_k^2 L}{2} \|T_{\gamma_k}(x^k)\|_X^2 \\ &\leq F(x^k) - \gamma_k \left(\nabla F(x^k), T_{\gamma_k}(x^k) \right)_X + \frac{\gamma_k}{2} \|T_{\gamma_k}(x^k)\|_X^2. \end{aligned}$$

Hence, using (7.5) and $\nabla F(x) \in \partial F(x)$, we obtain for all $z \in X$ that

$$(7.7) \quad \begin{aligned} J(x^{k+1}) &= F(x^k - \gamma_k T_{\gamma_k}(x^k)) + G(x^k - \gamma_k T_{\gamma_k}(x^k)) \\ &\leq F(x^k) - \gamma_k \left(\nabla F(x^k), T_{\gamma_k}(x^k) \right)_X + \frac{\gamma_k}{2} \|T_{\gamma_k}(x^k)\|_X^2 \\ &\quad + G(z) + \left(T_{\gamma_k}(x^k) - \nabla F(x^k), x^k - \gamma_k T_{\gamma_k}(x^k) - z \right)_X \\ &\leq F(z) + \left(\nabla F(x^k), x^k - z \right)_X - \gamma_k \left(\nabla F(x^k), T_{\gamma_k}(x^k) \right)_X + \frac{\gamma_k}{2} \|T_{\gamma_k}(x^k)\|_X^2 \\ &\quad + G(z) + \left(T_{\gamma_k}(x^k) - \nabla F(x^k), x^k - z - \gamma_k T_{\gamma_k}(x^k) \right)_X \\ &= J(z) + \left(T_{\gamma_k}(x^k), x^k - z \right)_X - \frac{\gamma_k}{2} \|T_{\gamma_k}(x^k)\|_X^2. \end{aligned}$$

For $z = x^k$ this implies that

$$J(x^{k+1}) \leq J(x^k) - \frac{\gamma_k}{2} \|T_{\gamma_k}(x^k)\|_X^2,$$

i.e., $\{J(x^k)\}_{k \in \mathbb{N}}$ is decreasing. (The proximal gradient method is thus a *descent method*.) Furthermore, by inserting $z = x^*$ with $J(x^*) = \min_{x \in X} J(x)$ in (7.7) and completing the square, we deduce that

$$(7.8) \quad \begin{aligned} 0 \leq J(x^{k+1}) - J(x^*) &\leq \left(T_{\gamma_k}(x^k), x^k - x^* \right)_X - \frac{\gamma_k}{2} \|T_{\gamma_k}(x^k)\|_X^2 \\ &= \frac{1}{2\gamma_k} \left(\|x^k - x^*\|_X^2 - \|x^k - x^* - \gamma_k T_{\gamma_k}(x^k)\|_X^2 \right) \\ &= \frac{1}{2\gamma_k} \left(\|x^k - x^*\|_X^2 - \|x^{k+1} - x^*\|_X^2 \right). \end{aligned}$$

In particular, $\{\|x^k - x^*\|_X\}_{k \in \mathbb{N}}$ is decreasing, and hence $\{x^k\}_{k \in \mathbb{N}}$ is Féjer monotone and therefore bounded. We can thus extract a weakly convergent subsequence $\{x^{k_l}\}_{l \in \mathbb{N}}$ with $x^{k_l} \rightharpoonup \bar{x}$.

We now sum (7.8) over $k = 1, \dots, n$ for arbitrary $n \in \mathbb{N}$ and obtain that

$$\begin{aligned} \sum_{k=1}^n (J(x^k) - J(x^*)) &\leq \frac{1}{2\gamma_{\min}} \sum_{k=1}^n \left(\|x^{k-1} - x^*\|_X^2 - \|x^k - x^*\|_X^2 \right) \\ &= \frac{1}{2\gamma_{\min}} \left(\|x^0 - x^*\|_X^2 - \|x^n - x^*\|_X^2 \right) \\ &\leq \frac{1}{2\gamma_{\min}} \|x^0 - x^*\|_X^2. \end{aligned}$$

Since $\{J(x^k)\}_{k \in \mathbb{N}}$ is decreasing, this implies that

$$(7.9) \quad J(x^n) - J(x^*) \leq \frac{1}{n} \sum_{k=1}^n (J(x^k) - J(x^*)) \leq \frac{1}{2n\gamma_{\min}} \|x^0 - x^*\|_X^2$$

and hence $J(x^n) \rightarrow J(x^*)$ for $n \rightarrow \infty$. The lower semi-continuity of F and G now yields that

$$J(\bar{x}) \leq \liminf_{l \rightarrow \infty} J(x^{k_l}) = J(x^*).$$

As in the proof of [Theorem 7.1](#), we can use the Féjer monotonicity of $\{x^k\}_{k \in \mathbb{N}}$ to show that $x^k \rightharpoonup \bar{x}$ for the full sequence. \square

In particular, we obtain from (7.9) that $J(x^k) = J(x^*) + \mathcal{O}(k^{-1})$. Ensuring $J(x^k) \leq J(x^*) + \varepsilon$ thus requires $\mathcal{O}(\varepsilon^{-1})$ iterations. By introducing a clever extrapolation, this can be reduced to $\mathcal{O}(\varepsilon^{-1/2})$ which is provably optimal; see [[Nesterov 1983](#)], [[Nesterov 2004](#), Theorem 2.1.7]. (However, the sequence of iterates is then no longer monotonically decreasing.) The corresponding iteration is given by

$$\begin{cases} x^{k+1} = \text{prox}_{\gamma_k G}(\bar{x}^k - \gamma_k \nabla F(\bar{x}^k)), \\ \bar{x}^{k+1} = x^{k+1} + \frac{1 - \tau_k}{\tau_{k+1}} (x^k - x^{k+1}), \end{cases}$$

for the (hardly intuitive) choice¹

$$\tau_0 = 1, \quad \tau_k = \frac{1 + \sqrt{1 + 4\tau_{k-1}^2}}{2} \quad (\rightarrow \infty),$$

see [[Beck & Teboulle 2009](#), § 4].

¹This choice satisfies the quadratic recursion $\tau_{k+1}^2 - \tau_{k+1} = \tau_k$, which cancels the $\mathcal{O}(k^{-1})$ terms in a key estimate.

One drawback of the explicit splitting is needing to know the Lipschitz constant L of ∇F in order to choose admissible step sizes γ_k . Looking at the proof of [Theorem 7.3](#), we can see that this is only used to obtain the estimate (7.6). Hence, if the Lipschitz constant is unknown, we can try to satisfy (7.6) by a line search in each iteration: Start with $\gamma^0 > 0$ and reduce γ_k (e.g., by halving) until

$$F(x^k - \gamma_k T_{\gamma_k}(x^k)) \leq F(x^k) - \gamma_k \left(\nabla F(x^k), T_{\gamma_k}(x^k) \right)_X + \frac{\gamma_k}{2} \|T_{\gamma_k}(x^k)\|_X^2$$

(which will be the case for $\gamma_k < L^{-1}$ at the latest). Of course, there's no free lunch: each step of the line search requires evaluating both F and $\text{prox}_{\gamma_k G}$ (although the latter can be avoided by exchanging gradient and proximal steps, i.e., *backward-forward splitting*).

7.3 IMPLICIT SPLITTING

Even with a line search, the restriction on the step sizes γ_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 (7.3), which yields the optimality system

$$\begin{cases} \bar{x} = \text{prox}_{\gamma F}(\bar{x} + \gamma \bar{p}), \\ \bar{x} = \text{prox}_{\gamma G}(\bar{x} - \gamma \bar{p}). \end{cases}$$

To eliminate \bar{p} from these equations, we set $\bar{z} := \bar{x} + \gamma \bar{p}$ and $\bar{w} := \bar{x} - \gamma \bar{p}$. This yields that $\bar{z} + \bar{w} = 2\bar{x}$, i.e.,

$$\bar{w} = 2\bar{x} - \bar{z}.$$

It remains to derive a recursion for \bar{z} , which we obtain from the productive zero

$$\bar{z} = \bar{z} + (\bar{x} - \bar{x}).$$

Applying a suitable fixed-point iteration to these equations yields the *Douglas-Rachford method*

$$(7.10) \quad \begin{cases} x^{k+1} = \text{prox}_{\gamma F}(z^k), \\ y^{k+1} = \text{prox}_{\gamma G}(2x^{k+1} - z^k), \\ z^{k+1} = z^k + y^{k+1} - x^{k+1}. \end{cases}$$

This iteration can be written as a proximal point iteration by introducing suitable block operators, which with some effort (in showing that these operators are maximally monotone) allows deducing the convergence from [Theorem 7.1](#); see, e.g., [[Eckstein & Bertsekas 1992](#)]. Here we will instead consider a variant which has proved extremely successful, in particular in mathematical imaging.

7.4 PRIMAL-DUAL EXTRAGRADIENT METHOD

Methods of this class were specifically developed to solve problems of the form

$$\min_{x \in X} F(x) + G(Ax)$$

for $F : X \rightarrow \overline{\mathbb{R}}$ and $G : Y \rightarrow \overline{\mathbb{R}}$ proper, convex, and lower semi-continuous, and $A \in L(X, Y)$. Applying [Theorem 5.6](#) and [Lemma 5.5](#) to such a problem yields the Fenchel extremality conditions

$$(7.11) \quad \begin{cases} -A^* \bar{y} \in \partial F(\bar{x}), \\ \bar{y} \in \partial G(A\bar{x}), \end{cases} \Leftrightarrow \begin{cases} -A^* \bar{y} \in \partial F(\bar{x}), \\ A\bar{x} \in \partial G^*(\bar{y}), \end{cases}$$

which can be reformulated using [Lemma 6.9](#) as

$$\begin{cases} \bar{x} = \text{prox}_{\tau F}(\bar{x} - \tau A^* \bar{y}), \\ \bar{y} = \text{prox}_{\sigma G^*}(\bar{y} + \sigma A \bar{x}), \end{cases}$$

for arbitrary $\sigma, \tau > 0$. Although this is already in the right form to obtain a fixed-point iteration for (\bar{x}, \bar{y}) , it is helpful from both a practical and a theoretical point of view to include an extrapolation step as in [\(7.10\)](#) to obtain the *primal-dual extragradient method*²

$$(7.12) \quad \begin{cases} x^{k+1} = \text{prox}_{\tau F}(x^k - \tau A^* y^k), \\ \bar{x}^{k+1} = 2x^{k+1} - x^k, \\ y^{k+1} = \text{prox}_{\sigma G^*}(y^k + \sigma A \bar{x}^{k+1}). \end{cases}$$

This iteration can now be written in the form of a proximal point method for (x, y) . For this, we rearrange [\(7.12\)](#) such that (x^k, y^k) and (x^{k+1}, y^{k+1}) appear on separate sides of each relation. For the first equation, we use $\text{prox}_{\tau F} = (\text{Id} + \tau \partial F)^{-1}$ to obtain that

$$\begin{aligned} x^{k+1} = \text{prox}_{\tau F}(x^k - \tau A^* y^k) &\Leftrightarrow x^k - \tau A^* y^k \in x^{k+1} + \tau \partial F(x^{k+1}) \\ &\Leftrightarrow \tau^{-1} x^k - A^* y^k \in \tau^{-1} x^{k+1} + \partial F(x^{k+1}). \end{aligned}$$

Similarly, we have for the second equation (after elimination of \bar{x}^{k+1}) that

$$y^{k+1} = \text{prox}_{\sigma G^*}(y^k + \sigma A(2x^{k+1} - x^k)) \Leftrightarrow \sigma^{-1} y^k - Ax^k \in \sigma^{-1} y^{k+1} + \partial G^*(y^{k+1}) - 2Ax^{k+1}.$$

If we now set $Z = X \times Y$, $z = (x, y)$,

$$M = \begin{pmatrix} \tau^{-1} \text{Id} & -A^* \\ -A & \sigma^{-1} \text{Id} \end{pmatrix}, \quad T = \begin{pmatrix} \partial F & A^* \\ -A & \partial G^* \end{pmatrix},$$

²This method was introduced in [[Chambolle & Pock 2011](#)], which is why it is frequently referred to as the *Chambolle–Pock method*. The relation to proximal point methods was first pointed out in [[He & Yuan 2012](#)].

we have shown that (7.12) is equivalent to

$$Mz^k \in (M + T)z^{k+1} \quad \Leftrightarrow \quad z^{k+1} \in (M + T)^{-1}Mz^k.$$

If M were invertible, we could use that $M = (M^{-1})^{-1}$ to obtain that $(M + T)^{-1}Mz^k = (\text{Id} + M^{-1}T)^{-1}z^k$; the iteration would indeed amount to a proximal point method for the operator $M^{-1}T$ (which hopefully is maximally monotone). To show invertibility of M , we first prove that – under suitable conditions on σ and τ – the operator M is self-adjoint and positive definite with respect to the inner product

$$(z_1, z_2)_Z = (x_1, x_2)_X + (y_1, y_2)_Y \quad \text{for all } z_1 = (x_1, y_1) \in Z, z_2 = (x_2, y_2) \in Z.$$

Lemma 7.4. *The operator M is bounded and self-adjoint. If $\sigma\tau\|A\|_{L(X,Y)}^2 < 1$, then M is positive definite.*

Proof. The definition of M directly implies boundedness (since $A \in L(X, Y)$ is bounded) and self-adjointness. Let now $z = (x, y) \in Z \setminus \{0\}$ be given. Then,

$$\begin{aligned} (Mz, z)_Z &= (\tau^{-1}x - A^*y, x)_X + (\sigma^{-1}y - Ax, y)_Y \\ &= \tau^{-1}\|x\|_X^2 - 2(x, A^*y)_X + \sigma^{-1}\|y\|_Y^2 \\ &\geq \tau^{-1}\|x\|_X^2 - 2\|A\|_{L(X,Y)}\|x\|_X\|y\|_Y + \sigma^{-1}\|y\|_Y^2 \\ &\geq \tau^{-1}\|x\|_X^2 - \|A\|_{L(X,Y)}\sqrt{\sigma\tau}(\tau^{-1}\|x\|_X^2 + \sigma^{-1}\|y\|_Y^2) + \sigma^{-1}\|y\|_Y^2 \\ &= (1 - \|A\|_{L(X,Y)}\sqrt{\sigma\tau})(\sqrt{\tau}^{-1}\|x\|_X^2 + \sqrt{\sigma}^{-1}\|y\|_Y^2) \\ &\geq C(\|x\|_X^2 + \|y\|_Y^2) \end{aligned}$$

for $C := (1 - \|A\|_{L(X,Y)}\sqrt{\sigma\tau}) \min\{\tau^{-1}, \sigma^{-1}\} > 0$. Hence, $(Mz, z)_Z > C\|z\|_Z^2$ for all $z \neq 0$, and therefore M is positive definite. \square

Under these conditions, the operator M induces an inner product $(z_1, z_2)_M := (Mz_1, z_2)_Z$ and, through it, a norm $\|z\|_M^2 = (z, z)_M$ that satisfies

$$(7.13) \quad c_1\|z\|_Z \leq \|z\|_M \leq c_2\|z\|_Z \quad \text{for all } z \in Z.$$

Corollary 7.5. *If $\sigma\tau\|A\|_{L(X,Y)}^2 < 1$, then M is continuously invertible, i.e., $M^{-1} \in L(Y, X)$.*

Proof. The following argument is standard in functional analysis. Let $z \in Z$ be given. Then (7.13) implies that the mapping $z \mapsto (z, v)_Z$ is a bounded (with respect to $\|\cdot\|_M$) linear functional. The [Fréchet–Riesz Theorem 1.12](#) thus yields a unique preimage $z^* \in Z$ with

$$(Mz^*, v)_Z = (z^*, v)_M = (z, v)_Z \quad \text{for all } v \in Z.$$

Furthermore, the mapping $M^{-1} : z \mapsto z^*$ is linear. Hence,

$$c_1^2\|z^*\|_Z^2 \leq \|z^*\|_M^2 = (Mz^*, z^*)_Z = (z, z^*)_Z \leq \|z\|_Z\|z^*\|_Z,$$

and dividing by $c_1^2\|z^*\|_Z$ yields the claimed boundedness of M^{-1} . \square

We have thus shown that $M^{-1}T$ is well-defined, i.e., $\text{graph } M^{-1}T \neq \emptyset$. To show that $M^{-1}T$ is also maximally monotone, we also need that the inner products $(\cdot, \cdot)_M$ and $(\cdot, \cdot)_Z$ are equivalent.

Corollary 7.6. *If $\sigma\tau\|A\|_{L(X,Y)}^2 < 1$, there exist $C_1, C_2 > 0$ with*

$$C_1(z, z')_Z \leq (z, z')_M \leq C_2(z, z')_Z \quad \text{for all } z, z' \in Z.$$

Proof. The parallelogram identity and (7.13) yield that

$$\begin{aligned} (z, z')_M &= \frac{1}{4} (\|z + z'\|_M^2 - \|z - z'\|_M^2) \\ &\geq \frac{1}{4} (c_1^2 \|z + z'\|_Z^2 - c_2^2 \|z - z'\|_Z^2) \\ &\geq \min\{c_1^2, c_2^2\} (z, z')_Z \end{aligned}$$

and thus the first inequality with $C_1 := \max\{c_1^2, c_2^2\}$. The second inequality is shown similarly. \square

Lemma 7.7. *If $\sigma\tau\|A\|_{L(X,Y)}^2 < 1$, then $M^{-1}T$ is maximally monotone.*

Proof. We first show the monotonicity of $M^{-1}T$. Let $z \in Z$ and $z^* \in M^{-1}Tz$, i.e., $Mz^* \in Tz$. By definition of T , we can thus find for any $z = (x, y)$ a $\xi \in \partial F(x)$ and an $\eta \in \partial G^*(y)$ with $Mz^* = (\xi + A^*y, \eta - Ax)$. Similarly, for given $\bar{z} = (\bar{x}, \bar{y}) \in Z$ and $\bar{z}^* \in M^{-1}T\bar{z}$ we can write $M\bar{z}^* = (\bar{\xi} + A^*\bar{y}, \bar{\eta} - A\bar{x})$ for a $\bar{\xi} \in \partial F(\bar{x})$ and an $\bar{\eta} \in \partial G^*(\bar{y})$. Then,

$$\begin{aligned} (M\bar{z}^* - Mz^*, \bar{z} - z)_Z &= ((\bar{\xi} + A^*\bar{y}) - (\xi + A^*y), \bar{x} - x)_X \\ &\quad + ((\bar{\eta} - A\bar{x}) - (\eta - Ax), \bar{y} - y)_Y \\ &= (\bar{\xi} - \xi, \bar{x} - x)_X + (A^*(\bar{y} - y), \bar{x} - x)_X \\ &\quad - (A(\bar{x} - x), \bar{y} - y)_Y + (\bar{\eta} - \eta, \bar{y} - y)_Y \\ &= (\bar{\xi} - \xi, \bar{x} - x)_X + (\bar{\eta} - \eta, \bar{y} - y)_Y \geq 0 \end{aligned}$$

by the monotonicity of subdifferentials. **Corollary 7.6** then implies that

$$(\bar{z}^* - z^*, \bar{z} - z)_Z \geq C_2^{-1} (M\bar{z}^* - Mz^*, \bar{z} - z)_Z \geq 0$$

and hence that $M^{-1}T$ is monotone.

To show maximal monotonicity, let $\bar{z}^*, \bar{z} \in Z$ with

$$(\bar{z}^* - z^*, \bar{z} - z)_Z \geq 0 \quad \text{for all } (z^*, z) \in \text{graph } M^{-1}T.$$

As above, we can write $Mz^* = (\xi + A^*y, \eta - Ax)$ for a $\xi \in \partial F(x)$ and an $\eta \in \partial G^*(y)$.

Corollary 7.6 then implies that

$$(7.14) \quad (M\bar{z}^* - Mz^*, \bar{z} - z)_Z \geq C_1 (\bar{z}^* - z^*, \bar{z} - z)_Z \geq 0 \quad \text{for all } z \in Z, Mz^* \in Tz.$$

We now set $\bar{\xi} := \bar{x}^* - A^*\bar{y}$ and $\bar{\eta} := \bar{y}^* + A\bar{x}$ for $M\bar{z}^* = (\bar{x}^*, \bar{y}^*)$ and $\bar{z} = (\bar{x}, \bar{y})$. Then $M\bar{z}^* = (\bar{\xi} + A^*\bar{y}, \bar{\eta} - A\bar{x})$, and (7.14) implies for all $(x, y) \in Z$ that

$$\begin{aligned} 0 &\leq ((\bar{\xi} + A^*\bar{y}) - (\xi + A^*y), \bar{x} - x)_X + ((\bar{\eta} - A\bar{x}) - (\eta - Ax), \bar{y} - y)_Y \\ &= (\bar{\xi} - \xi, \bar{x} - x)_X + (\bar{\eta} - \eta, \bar{y} - y)_Y. \end{aligned}$$

In particular, this holds for pairs (x, y) of the form (x, \bar{y}) for arbitrary $x \in X$ or (\bar{x}, y) for arbitrary $y \in Y$, which shows that each inner product on the right-hand side is nonnegative. The maximal monotonicity of subdifferentials now implies that $\bar{\xi} \in \partial F(\bar{x})$ and $\bar{\eta} \in \partial G^*(\bar{y})$. Hence

$$M\bar{z}^* = (\bar{\xi} + A^*\bar{y}, \bar{\eta} - A\bar{x}) \in T\bar{z},$$

i.e., $\bar{z}^* \in M^{-1}T\bar{z}$. We conclude that $M^{-1}T$ is maximally monotone as claimed. \square

To sum up, we have shown that the primal-dual extragradient method (7.12) is equivalent to the proximal point method $z^{k+1} = \mathcal{R}_{M^{-1}T}z^k$ for the maximally monotone operator $M^{-1}T$, and hence its convergence follows from Theorem 7.1.

Theorem 7.8. *If $F : X \rightarrow \overline{\mathbb{R}}$, $G : Y \rightarrow \overline{\mathbb{R}}$, and $A \in L(X, Y)$ satisfy the assumptions of Theorem 5.6 and if $\sigma\tau\|A\|_{L(X,Y)}^2 < 1$, then the sequence $\{(x^k, y^k)\}_{k \in \mathbb{N}}$ generated by (7.12) converges weakly in Z to a pair (\bar{x}, \bar{y}) satisfying (7.11).*

Note that although the iteration is implicit in F and G , it is still explicit in A ; it is therefore not surprising that step size restrictions based on A remain.³

Finally, we remark that by setting $A = \text{Id}$, $\tau = \gamma$, $\sigma = \gamma^{-1}$ and $z^k = x^k - \gamma y^k$ in (7.12) and applying Lemma 6.12 (ii), we recover the Douglas–Rachford method (7.10); however, since in this case $\sigma\tau\|A\|_{L(X,Y)}^2 = 1$, we cannot obtain its convergence from Theorem 7.8.

³Using a proximal point mapping for $G \circ A$ would lead to a fully implicit method but involve the inverse A^{-1} in the corresponding proximal point mapping. It is precisely the point of the primal-dual extragradient method to avoid having to invert A , which is often prohibitively expensive if not impossible (e.g., if A does not have closed range as in many inverse problems).

Part III

LIPSCHITZ ANALYSIS

8 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; these are the locally Lipschitz continuous functionals. Recall that $F : X \rightarrow \mathbb{R}$ is locally Lipschitz continuous in $x \in X$ 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

$$|F(x_1) - F(x_2)| \leq L\|x_1 - x_2\|_X \quad \text{for all } x_1, x_2 \in O_\delta(x).$$

We will refer to the $O_\delta(x)$ from the definition as the *Lipschitz neighborhood* of x . Note that in contrast to convexity, this is a purely local condition; on the other hand, we have to require that F is (locally) finite-valued.¹

We proceed as for the convex subdifferential and first define for $F : X \rightarrow \mathbb{R}$ the *generalized directional derivative* in $x \in X$ in direction $h \in X$ as

$$F^\circ(x; h) := \limsup_{\substack{y \rightarrow x \\ t \rightarrow 0^+}} \frac{F(y + th) - F(y)}{t}.$$

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 8.1. *Let $F : X \rightarrow \mathbb{R}$ be locally Lipschitz continuous in $x \in X$. 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., $F^\circ(x; h + g) \leq F^\circ(x; h) + F^\circ(x; g)$ for all $h, g \in X$;*
- (iii) *positive homogeneous, i.e., for all $\alpha > 0$ and $h \in X$ we have that $F^\circ(x; \alpha h) = (\alpha F)^\circ(x; h)$;*
- (iv) *reflective, i.e., $F^\circ(x; -h) = (-F)^\circ(x; h)$ for all $h \in X$.*

¹For $F : X \rightarrow \overline{\mathbb{R}}$, this is always the case in the interior of the effective domain. It is also possible to extend the generalized derivative introduced below to points on the boundary of the effective domain where F is finite using an equivalent, more geometrical, construction involving generalized normal cones to epigraphs; see [Clarke 1990, Definition 2.4.10].

Proof. *Ad (i):* Let $h, g \in X$ be arbitrary. The local Lipschitz continuity of F implies that

$$F(y + th) - F(y) \leq F(y + tg) - F(y) + tL\|h - g\|_X$$

for all y sufficiently close to x and t sufficiently small. Dividing by $t > 0$ and taking the lim sup then yields that

$$F^\circ(x; h) \leq F^\circ(x; g) + L\|h - g\|_X.$$

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.

Ad (ii): The definition of the lim sup and the productive zero immediately yield

$$\begin{aligned} F^\circ(x; h + g) &= \limsup_{\substack{y \rightarrow x \\ t \rightarrow 0^+}} \frac{F(y + th + tg) - F(y)}{t} \\ &\leq \limsup_{\substack{y \rightarrow x \\ t \rightarrow 0^+}} \frac{F(y + th + tg) - F(y + tg)}{t} + \limsup_{\substack{y \rightarrow x \\ t \rightarrow 0^+}} \frac{F(y + tg) - F(y)}{t} \\ &= F^\circ(x; h) + F^\circ(x; g), \end{aligned}$$

since $y \rightarrow x$ and $t \rightarrow 0$ implies that $y + tg \rightarrow x$ as well.

Ad (iii): Again from the definition we obtain for $\alpha > 0$ that

$$\begin{aligned} F^\circ(x; \alpha h) &= \limsup_{\substack{y \rightarrow x \\ t \rightarrow 0^+}} \frac{F(y - t(\alpha h)) - F(y)}{t} \\ &= \limsup_{\substack{y \rightarrow x \\ \alpha t \rightarrow 0^+}} \alpha \frac{F(y + (\alpha t)h) - F(y)}{\alpha t} = (\alpha F)^\circ(x; h). \end{aligned}$$

Ad (iv): Similarly, we have that

$$\begin{aligned} F^\circ(x; -h) &= \limsup_{\substack{y \rightarrow x \\ t \rightarrow 0^+}} \frac{F(y - th) - F(y)}{t} \\ &= \limsup_{\substack{w \rightarrow x \\ t \rightarrow 0^+}} \frac{-F(w + th) - (-F(w))}{t} = (-F)^\circ(x; h), \end{aligned}$$

since $y \rightarrow x$ and $t \rightarrow 0$ implies that $w := y - th \rightarrow x$ as well. □

In particular, [Lemma 8.1](#) (i–iii) implies that the mapping $h \mapsto F^\circ(x; h)$ is proper, convex, and lower semi-continuous.

We now define for a locally Lipschitz continuous functional $F : X \rightarrow \mathbb{R}$ the *Clarke subdifferential* in $x \in X$ as

$$(8.1) \quad \partial_C F(x) := \{x^* \in X^* : \langle x^*, h \rangle_X \leq F^\circ(x; h) \text{ for all } h \in X\}.$$

The definition together with [Lemma 8.1](#) (i) implies that $\partial_C F(x)$ is convex, weakly-* closed, and bounded (since $\|\xi\|_{X^*} \leq L$ for all $\xi \in \partial_C F(x)$ by definition of the operator norm). Again, the construction immediately yields a Fermat principle.²

Theorem 8.2. *If $F : X \rightarrow \mathbb{R}$ 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

$$\langle 0, h \rangle_X = 0 \leq \liminf_{t \rightarrow 0^+} \frac{F(\bar{x} + th) - F(\bar{x})}{t} \leq \limsup_{t \rightarrow 0^+} \frac{F(\bar{x} + th) - F(\bar{x})}{t} \leq F^\circ(x; h)$$

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., $f(t) = -|t|$). We will also need the following closedness property.

Lemma 8.3. *Let $F : X \rightarrow \mathbb{R}$ be locally Lipschitz continuous in $x \in X$. If $\{x_n\}_{n \in \mathbb{N}} \subset X$ is a sequence with $x_n \rightarrow x$ and if $x_n^* \in \partial_C F(x_n)$ for all $n \in \mathbb{N}$ with $x_n^* \rightharpoonup^* x^*$ in 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

$$\langle x^*, h \rangle_X = \lim_{n \rightarrow \infty} \langle x_n^*, h \rangle_X \leq \limsup_{n \rightarrow \infty} F^\circ(x_n; h).$$

Hence we are finished if we can show that $\limsup_{n \rightarrow \infty} F^\circ(x_n; h) \leq F^\circ(x; h)$ (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 $y_{n,m} \rightarrow x_n$ and $t_{n,m} \rightarrow 0$ for $m \rightarrow \infty$ realizing each lim sup. Hence, for all $n \in \mathbb{N}$ we can find a y_n and a t_n such that $\|y_n - x_n\|_X + t_n < n^{-1}$ (and hence in particular $y_n \rightarrow x$ and $t_n \rightarrow 0$) as well as

$$F^\circ(x_n; h) - \frac{1}{n} \leq \frac{F(y_n + t_n h) - F(y_n)}{t_n}$$

for n sufficiently large. Taking the lim sup for $n \rightarrow \infty$ on both sides yields the desired inequality. □

²Similarly to [Theorem 4.3](#), 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.

Next, we show that the Clarke subdifferential is indeed a generalization of the derivative concepts we've studied so far.

Theorem 8.4. *Let $F : X \rightarrow \mathbb{R}$ be continuously Fréchet differentiable in a neighborhood U of $x \in X$. Then, $\partial_C F(x) = \{F'(x)\}$.*

Proof. First we note that the assumption implies local Lipschitz continuity of F : Since F' is continuous in U , there exists a $\delta > 0$ with $\|F'(z) - F'(x)\|_{X^*} \leq 1$ and hence $\|F'(z)\|_{X^*} \leq 1 + \|F'(x)\|_{X^*}$ for all $z \in K_\delta(x) \subset U$. For any $x_1, x_2 \in K_\delta(x)$ we also have $x_2 + t(x_1 - x_2) \in K_\delta(x)$ for all $t \in [0, 1]$ (since balls in normed vector spaces are convex), and hence [Theorem 2.6](#) implies that

$$\begin{aligned} |F(x_1) - F(x_2)| &\leq \int_0^1 \|F'(x_2 + t(x_1 - x_2))\|_{X^*} t \|x_1 - x_2\|_X dt \\ &\leq \frac{1 + \|F'(x)\|_{X^*}}{2} \|x_1 - x_2\|_X. \end{aligned}$$

We now show that $F^\circ(x; h) = F'(x)h$ for all $h \in X$. Take again sequences $\{x_n\}_{n \in \mathbb{N}}$ and $\{t_n\}_{n \in \mathbb{N}}$ with $x_n \rightarrow x$ and $t_n \rightarrow 0^+$ realizing the lim sup. Applying again the mean value [Theorem 2.6](#) and using the continuity of F' yields for any $h \in X$ that

$$\begin{aligned} F^\circ(x; h) &= \lim_{n \rightarrow \infty} \frac{F(x_n + t_n h) - F(x_n)}{t_n} \\ &= \lim_{n \rightarrow \infty} \int_0^1 \frac{1}{t_n} \langle F'(x_n + t(t_n h)), t_n h \rangle_X dt \\ &= \langle F'(x), h \rangle_X \end{aligned}$$

since the integrand converges uniformly in $t \in [0, 1]$ 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 $x^* = F'(x)$. \square

Theorem 8.5. *Let $F : X \rightarrow \mathbb{R}$ be convex and lower semi-continuous. Then, $\partial_C F(x) = \partial F(x)$ for all $x \in X$.*

Proof. Since F is finite-valued, $(\text{dom } F)^\circ = X$, and hence F is local Lipschitz continuous in every $x \in X$ by [Theorem 3.11](#). We now show that $F^\circ(x; h) = F'(x; h)$ for all $h \in X$, which together with the definition (4.1) of the convex subdifferential (which is equivalent to definition (3.1) by [Lemma 4.2](#) with Definition (3.1)) yields the claim. First, we always have that

$$F'(x; h) = \lim_{t \rightarrow 0^+} \frac{F(x + th) - F(x)}{t} \leq \limsup_{\substack{y \rightarrow x \\ t \rightarrow 0^+}} \frac{F(y + th) - F(y)}{t} = F^\circ(x; h).$$

To show the reverse inequality, let $\delta > 0$ be arbitrary. Since the difference quotient of convex functionals is increasing by [Lemma 4.1 \(i\)](#), we obtain that

$$\begin{aligned} F^\circ(x; h) &= \lim_{\varepsilon \rightarrow 0^+} \sup_{y \in K_{\delta\varepsilon}(x)} \sup_{0 < t < \varepsilon} \frac{F(y + th) - F(y)}{t} \\ &\leq \lim_{\varepsilon \rightarrow 0^+} \sup_{y \in K_{\delta\varepsilon}(x)} \frac{F(y + \varepsilon h) - F(y)}{\varepsilon} \\ &\leq \lim_{\varepsilon \rightarrow 0^+} \frac{F(x + \varepsilon h) - F(x)}{\varepsilon} + 2L\delta \\ &= F'(x; h) + 2L\delta, \end{aligned}$$

where the last inequality follows by taking 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. \square

A locally Lipschitz continuous functional $F : X \rightarrow \mathbb{R}$ 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 semi-continuous functional is regular; intuitively, a function is thus regular in any points in which it is either differentiable or at least has a “convex kink”.

We now turn to calculus rules. The first one still follows directly from the definition.

Theorem 8.6. *Let $F : X \rightarrow \mathbb{R}$ be locally Lipschitz continuous in $x \in X$ and $\alpha \in \mathbb{R}$. Then,*

$$\partial_C(\alpha F)(x) = \alpha \partial_C(F)(x).$$

Proof. First, αF is clearly locally Lipschitz continuous for any $\alpha \in \mathbb{R}$. If $\alpha = 0$, both sides of the claimed equality are zero (which is easiest seen from [Theorem 8.4](#)). 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,

$$\begin{aligned} \alpha \partial_C F(x) &= \{\alpha x^* \in X^* : \langle x^*, h \rangle_X \leq F^\circ(x; h) \text{ for all } h \in X\} \\ &= \{\alpha x^* \in X^* : \langle \alpha x^*, h \rangle_X \leq \alpha F^\circ(x; h) \text{ for all } h \in X\} \\ &= \{y^* \in X^* : \langle y^*, h \rangle_X \leq (\alpha F)^\circ(x; h) \text{ for all } h \in X\} \\ &= \partial_C(\alpha F)(x). \end{aligned}$$

To conclude the proof, it suffices to show the claim for $\alpha = -1$. For that, we use [Lemma 8.1 \(iv\)](#) to obtain that

$$\begin{aligned} \partial_C(-F)(x) &= \{x^* \in X^* : \langle x^*, h \rangle_X \leq (-F)^\circ(x; h) \text{ for all } h \in X\} \\ &= \{x^* \in X^* : \langle -x^*, -h \rangle_X \leq F^\circ(x; -h) \text{ for all } h \in X\} \\ &= \{-y^* \in X^* : \langle y^*, g \rangle_X \leq F^\circ(x; g) \text{ for all } g \in X\} \\ &= -\partial_C(F)(x). \end{aligned} \quad \square$$

Corollary 8.7. *Let $F : X \rightarrow \mathbb{R}$ be locally Lipschitz continuous in $\bar{x} \in X$. 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, [Theorem 8.2](#) implies that

$$0 \in \partial_C(-F)(\bar{x}) = -\partial_C F(\bar{x}),$$

i.e., $0 = -0 \in \partial_C F(\bar{x})$. □

The remaining rules are again significantly more involved. As in the previous proofs, a key step is to relate different sets of the form (8.1), for which the following lemmas will be helpful.

Lemma 8.8. *Let $F : X \rightarrow \mathbb{R}$ be positive homogeneous, subadditive, and lower semi-continuous, and let*

$$A = \{x^* \in X^* : \langle x^*, x \rangle_X \leq F(x) \text{ for all } x \in X\}.$$

Then,

$$(8.2) \quad F(x) = \sup_{x^* \in A} \langle x^*, x \rangle_X \quad \text{for all } x \in X.$$

Proof. By definition of A , the inequality $\langle x^*, x \rangle_X - F(x) \leq 0$ holds for all $x \in X$ if and only if $x^* \in A$. Thus, a case distinction as in [Example 5.2](#) (iii) using the positive homogeneity of F shows that

$$F^*(x^*) = \sup_{x \in X} \langle x^*, x \rangle_X - F(x) = \begin{cases} 0 & x^* \in A, \\ \infty & x^* \notin A, \end{cases}$$

i.e., $F^* = \delta_A$. Further, F by assumption is also subadditive and hence convex as well as lower semi-continuous. [Theorem 5.1](#) thus implies that

$$F(x) = F^{**}(x) = (\delta_A)^*(x) = \sup_{x^* \in A} \langle x^*, x \rangle_X. \quad \square$$

The right-hand side of (8.2) is called the *support functional* of A . With its help, we can finally show the promised nonemptiness of the convex subdifferential.

Corollary 8.9. *Let $F : X \rightarrow \overline{\mathbb{R}}$ be proper, convex, and lower semi-continuous, and $x \in (\text{dom } F)^\circ$. Then, $\partial F(x)$ is non-empty and bounded.*

Proof. By [Theorem 3.11](#), F is locally Lipschitz in $x \in (\text{dom } F)^\circ$. Hence, [Theorem 8.5](#) together with [Lemma 8.8](#) shows that

$$\sup_{x^* \in \partial F(x)} \langle x^*, h \rangle = \sup_{x^* \in \partial_C F(x)} \langle x^*, h \rangle = F^\circ(x; h) = F'(x; h)$$

for all $h \in X$. But [Lemma 4.1](#) (iii) implies that $F'(x; h) \in \mathbb{R}$ for $x \in (\text{dom } F)^o$, and hence the supremum on the left-hand side cannot be over the empty set (for which any supremum is $-\infty$ by convention). Similarly, the boundedness follows from [Theorem 8.5](#) together with the boundedness of the Clarke subdifferential. \square

Lemma 8.10. *Let $F, G : X \rightarrow \mathbb{R}$ be positive homogeneous, subadditive, and lower semi-continuous, and let*

$$\begin{aligned} A &:= \{x^* \in X^* : \langle x^*, x \rangle \leq F(x) \text{ for all } x \in X\}, \\ B &:= \{x^* \in X^* : \langle x^*, x \rangle \leq G(x) \text{ for all } x \in X\}, \end{aligned}$$

be non-empty. Then, $F \leq G$ implies that $A \subset B$.

Proof. Assume that there exists an $x^* \in A$ with $x^* \notin B$. Since A and B are convex and weakly-* closed by construction, [Theorem 1.11](#) yields an $x \in X$ and a $\lambda \in \mathbb{R}$ with

$$\langle z^*, x \rangle_X \leq \lambda < \langle x^*, x \rangle_X \leq F(x) \quad \text{for all } z^* \in B.$$

We hence obtain from [Lemma 8.8](#) that

$$G(x) = \sup_{z^* \in B} \langle z^*, x \rangle_X < F(x),$$

in contradiction to $F \leq G$. \square

Lemma 8.11. *Let $A, B \subset X^*$ be convex and weakly-* closed. Then, $A = B$ if and only if*

$$(8.3) \quad \sup_{x^* \in A} \langle x^*, x \rangle_X = \sup_{x^* \in B} \langle x^*, x \rangle_X \quad \text{for all } x \in X.$$

Proof. The claim is obvious if $A = B$. Conversely, if (8.3) holds, we have for all $x \in X$ that

$$(\delta_A)^*(x) = \sup_{x^* \in A} \langle x^*, x \rangle_X = \sup_{x^* \in B} \langle x^*, x \rangle_X = (\delta_B)^*(x).$$

The assumptions on A and B now imply that the corresponding indicator functionals δ_A and δ_B are convex and weakly-* lower semi-continuous. Following the proofs of [Lemma 3.6](#) and [Theorem 5.1](#) shows that it is sufficient to take [Theorem 1.11](#) in place of [Theorem 1.5](#) (ii) to obtain that

$$\delta_A = ((\delta_A)^*)^* = ((\delta_B)^*)^* = \delta_B,$$

which by definition of the indicator functional holds if and only if $A = B$. \square

We now use these results to prove a sum rule.

Theorem 8.12. *Let $F, G : X \rightarrow \mathbb{R}$ be locally Lipschitz continuous in $x \in X$. Then,*

$$\partial_C(F + G)(x) \subset \partial_C F(x) + \partial_C G(x).$$

If F and G are regular in x , then $F + G$ is regular in x and equality holds.

Proof. It is clear that $F + G$ is locally Lipschitz continuous in x . Furthermore, from the properties of the lim sup we always have for all $h \in X$ that

$$(F + G)^\circ(x; h) \leq F^\circ(x; h) + G^\circ(x; h).$$

If F and G are regular in 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 $(F + G)^\circ(x; h) = (F + G)^\circ(x; h) = (F + G)'(x; h)$, i.e., $F + G$ is regular.

By [Lemma 8.10](#) we are thus finished if we can show that

$$\partial_C F(x) + \partial_C G(x) = \{x^* \in X^* : \langle x^*, x \rangle_X \leq F^\circ(x; h) + G^\circ(x; h) \text{ for all } h \in X\} =: A.$$

For this, we use that both sets are convex and weakly-* closed, and that generalized directional derivatives and hence their sums are positive homogeneous, convex, and lower semi-continuous by [Lemma 8.1](#). We thus obtain from [Lemma 8.8](#) 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 inclusion respectively equality of the sets now follows from [Lemma 8.11](#). \square

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).

To prove a chain rule, we need the following “nonsmooth” mean value theorem.

Theorem 8.13. *Let $F : X \rightarrow \mathbb{R}$ be locally Lipschitz continuous in $x \in X$ and y be in the Lipschitz neighborhood of x . Then there exists a $\lambda \in (0, 1)$ and an $x^* \in \partial_C F(x + \lambda(y - x))$ such that*

$$F(y) - F(x) = \langle x^*, y - x \rangle_X.$$

Proof. Define $\psi, \varphi : [0, 1] \rightarrow \mathbb{R}$ as

$$\psi(\lambda) := F(x + \lambda(y - x)), \quad \varphi(\lambda) := \psi(\lambda) + \lambda(F(x) - F(y)).$$

By the assumptions on F and y , both φ and ψ 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 [Theorem 8.2](#) or [Corollary 8.7](#), respectively, together with [Theorems 8.4](#) and [8.12](#) we thus obtain that

$$0 \in \partial_C \varphi(\bar{\lambda}) \subset \partial_C \psi(\bar{\lambda}) + \{F(x) - F(y)\}.$$

Hence we are finished if we can show for $x_{\bar{\lambda}} := x + \bar{\lambda}(y - x)$ that

$$(8.4) \quad \partial_C \psi(\bar{\lambda}) \subset \{ \langle x^*, y - x \rangle_X : x^* \in \partial_C F(x_{\bar{\lambda}}) \}.$$

For this purpose, consider for arbitrary $s \in \mathbb{R}$ the generalized derivative

$$\begin{aligned} \psi^\circ(\bar{\lambda}; s) &= \limsup_{\substack{\lambda \rightarrow \bar{\lambda} \\ t \rightarrow 0^+}} \frac{\psi(\bar{\lambda} + ts) - \psi(\bar{\lambda})}{t} \\ &= \limsup_{\substack{\lambda \rightarrow \bar{\lambda} \\ t \rightarrow 0^+}} \frac{F(x + (\lambda + ts)(y - x)) - F(x + \lambda(y - x))}{t} \\ &\leq \limsup_{\substack{z \rightarrow x_{\bar{\lambda}} \\ t \rightarrow 0^+}} \frac{F(z + ts(y - x)) - F(z)}{t} = F^\circ(x_{\bar{\lambda}}; s(y - x)), \end{aligned}$$

where the inequality follows from considering arbitrary sequences $z \rightarrow x_{\bar{\lambda}}$ (instead of special sequences of the form $z_n = x + \lambda_n(y - x)$) in the last lim sup. [Lemma 8.10](#) thus implies that

$$(8.5) \quad \partial_C \psi(\bar{\lambda}) \subset \{ t^* \in \mathbb{R} : t^* s \leq F^\circ(x_{\bar{\lambda}}; s(y - x)) \text{ for all } s \in \mathbb{R} \}.$$

It remains to show that the sets on the right-hand sides of (8.4) and (8.5) – which we will denote by A and B , respectively – coincide. But this follows again from [Lemmas 8.8](#) and [8.11](#), since for all $s \in \mathbb{R}$ we have that

$$\sup_{t^* \in A} t^* s = \sup_{x^* \in \partial_C F(x_{\bar{\lambda}})} \langle x^*, s(y - x) \rangle_X = F^\circ(x_{\bar{\lambda}}; s(y - x)) = \sup_{t^* \in B} t^* s. \quad \square$$

We now come to the chain rule, which in contrast to the convex case does not require the inner mapping to be linear; in our context, this is one of the main advantages of the Clarke subdifferential.

Theorem 8.14. *Let Y be a separable Banach space, $F : X \rightarrow Y$ be continuously Fréchet differentiable in $x \in X$, and $G : Y \rightarrow \mathbb{R}$ be locally Lipschitz continuous in $F(x)$. Then,*

$$\partial_C(G \circ F)(x) \subset F'(x)^* \partial_C G(F(x)) := \{ F'(x)^* y^* : y^* \in \partial_C G(F(x)) \}.$$

If G is regular in $F(x)$, then $G \circ F$ is regular in x , and equality holds.

Proof. The local Lipschitz continuity of $G \circ F$ follows from that of G and F . For the claimed inclusion respectively equality, we argue as above. First we show that for every $h \in X$ there exists a $y^* \in \partial_C G(F(x))$ with

$$(8.6) \quad (G \circ F)^\circ(x; h) = \langle y^*, F'(x)h \rangle_Y.$$

For this, 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 $x_n \rightarrow x$, $t_n \rightarrow 0$, and

$$(G \circ F)^\circ(x; h) = \lim_{n \rightarrow \infty} \frac{G(F(x_n + t_n h)) - G(F(x_n))}{t_n}.$$

Furthermore, by continuity of F , we can find an $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, both $F(x_n)$ and $F(x_n + t_n h)$ are in the Lipschitz neighborhood of $F(x)$. [Theorem 8.13](#) thus yields for all $n \geq n_0$ a $\lambda_n \in (0, 1)$ and a $y_n^* \in \partial_C G(y_n)$ for $y_n := F(x_n) + \lambda_n(F(x_n + t_n h) - F(x_n))$ with

$$(8.7) \quad \frac{G(F(x_n + t_n h)) - G(F(x_n))}{t_n} = \left\langle y_n^*, \frac{F(x_n + t_n h) - F(x_n)}{t_n} \right\rangle_Y.$$

Since $\lambda_n \in (0, 1)$ is uniformly bounded, we also have that $y_n \rightarrow F(x)$ for $n \rightarrow \infty$. Hence, for n large enough, y_n is in the Lipschitz neighborhood of $F(x)$ as well. The Lipschitz continuity thus implies that $y_n^* \in \partial_C G(y_n) \subset K_L(0)$ eventually. This implies that $\{y_n^*\}_{n \in \mathbb{N}} \subset Y^*$ is bounded, and the [Banach–Alaoglu Theorem 1.10](#) yields a weakly-* convergent subsequence with limit $y^* \in \partial_C G(F(x))$ by [Lemma 8.3](#). Finally, since F is Fréchet differentiable, the difference quotient on the right-hand side of (8.7) converges strongly in Y to $F'(x)h$. (This is not obvious, but can be shown using the mean value theorem.) Hence, the right-hand side is the duality pairing of weakly-* and strongly converging sequences and hence itself convergent. Passing to the limit in (8.7) therefore yields (8.6). By definition of the Clarke subdifferential, we thus have for a $y^* \in \partial_C G(F(x))$ that

$$(8.8) \quad (G \circ F)^\circ(x; h) = \langle y^*, F'(x)h \rangle_Y \leq G^\circ(F(x); F'(x)h).$$

If G is now regular in 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

$$\begin{aligned} G^\circ(F(x); F'(x)h) &= \lim_{t \rightarrow 0^+} \frac{G(F(x) + tF'(x)h) - G(F(x))}{t} \\ &= \lim_{t \rightarrow 0^+} \frac{G(F(x) + tF'(x)h) - G(F(x + th)) + G(F(x + th)) - G(F(x))}{t} \\ &\leq \lim_{t \rightarrow 0^+} \left(\frac{G(F(x + th)) - G(F(x))}{t} + L\|h\|_X \frac{\|F(x) + F'(x)th - F(x + th)\|_Y}{\|th\|_X} \right) \\ &= (G \circ F)'(x; h) \leq (G \circ F)^\circ(x; h). \end{aligned}$$

(Since both the sum and the second summand in the next-to-last line converge, this has to be the case for the first summand as well.) Together with (8.8), this implies that $(G \circ F)'(x; h) = (G \circ F)^\circ(x; h)$ (i.e., $G \circ F$ is regular in x) and that $(G \circ F)^\circ(x; h) = G^\circ(F(x); F'(x)h)$.

Proceeding as in the proof of [Theorem 8.13](#) now shows that

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

which yields the remaining claims. \square

Again, the generic inclusion is the reverse of the one in the convex chain rule. If G is not regular but $F'(x)$ is surjective, a similar argument shows that equality holds in the chain rule (but not the regularity of $G \circ F$); see [Clarke 2013, Theorem 10.19].

Example 8.15. As a simple example, we consider

$$f : \mathbb{R}^2 \rightarrow \mathbb{R}, \quad (x_1, x_2) \mapsto |x_1 x_2|,$$

which is not convex. To compute the Clarke subdifferential, we write $f = g \circ T$ for

$$g : \mathbb{R} \rightarrow \mathbb{R}, \quad t \mapsto |t|, \quad T : \mathbb{R}^2 \rightarrow \mathbb{R}, \quad (x_1, x_2) \mapsto x_1 x_2,$$

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 $T'(x) \in L(\mathbb{R}^2, \mathbb{R})$ given by

$$T'(x)h = x_2 h_1 + x_1 h_2.$$

Its adjoint is easily verified to be given by

$$T'(x)^* h = \begin{pmatrix} x_2 h \\ x_1 h \end{pmatrix}.$$

Hence, [Theorem 8.14](#) together with [Theorem 8.5](#) yields that f is regular at any $x \in \mathbb{R}^2$ and that

$$\partial_C f(x) = T'(x)^* \partial g(T(x)) = \begin{pmatrix} x_2 \\ x_1 \end{pmatrix} \text{sign}(x_1 x_2).$$

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 8.16 (Rademacher). *Let $U \subset \mathbb{R}^N$ be open and $F : U \rightarrow \mathbb{R}$ be Lipschitz continuous. Then F is Fréchet differential in 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 8.17. *Let $F : \mathbb{R}^N \rightarrow \mathbb{R}$ be locally Lipschitz continuous in $x \in \mathbb{R}^N$ and Fréchet differentiable on $\mathbb{R}^N \setminus E_F$ for a set $E_F \subset \mathbb{R}^N$ of Lebesgue measure 0. Then*

$$(8.9) \quad \partial_C F(x) = \text{co} \left\{ \lim_{n \rightarrow \infty} \nabla F(x_n) : x_n \rightarrow x, x_n \notin E_F \right\},$$

where $\text{co } A$ 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 – possibly after intersection with the Lipschitz neighborhood of x – has Lebesgue measure 0. Hence there indeed exist sequences $\{x_n\}_{n \in \mathbb{N}} \in \mathbb{R}^N \setminus E_F$ with $x_n \rightarrow x$. 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

$$|(\nabla F(x_n), h)| = \left| \lim_{t \rightarrow 0^+} \frac{F(x_n + th) - F(x_n)}{t} \right| \leq L \|h\|$$

and hence that $\|\nabla F(x_n)\| \leq L$. This implies that $\{\nabla F(x_n)\}_{n \in \mathbb{N}}$ is bounded and thus contains a convergent subsequence. The set on the right-hand side of (8.9) is therefore nonempty.

Let now $\{x_n\}_{n \in \mathbb{N}} \subset \mathbb{R}^N \setminus E_F$ be an arbitrary sequence with $x_n \rightarrow x$ and $\{\nabla F(x_n)\}_{n \in \mathbb{N}} \rightarrow x^*$ for some $x^* \in \mathbb{R}^N$. Since F is differentiable in every $x_n \notin E_F$ by definition, [Theorem 8.4](#) yields that $\nabla F(x_n) \in \partial_C F(x_n)$, and hence $x^* \in \partial_C F(x)$ by [Lemma 8.3](#). The convexity of $\partial_C F(x)$ now implies that any convex combination of such limits x^* is contained in $\partial_C F(x)$, which shows the inclusion “ \supset ” in (8.9).

For the other inclusion, we first show for all $h \in \mathbb{R}^N$ and $\varepsilon > 0$ that

$$(8.10) \quad F^\circ(x; h) - \varepsilon \leq \limsup_{E_F \ni y \rightarrow x} (\nabla F(y), h) =: M(h).$$

Indeed, by definition of $M(h)$ and of the lim sup, for every $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$(\nabla F(y), h) \leq M(h) + \varepsilon \quad \text{for all } y \in O_\delta(x) \setminus E_F.$$

Here, $\delta > 0$ can be chosen sufficiently small for F to be Lipschitz continuous on $O_\delta(x)$. In particular, $E_F \cap O_\delta(x)$ is a set of zero measure. Hence, F is differentiable in $y + th$ for almost all $y \in O_{\delta/2}(x)$ 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

$$(8.11) \quad F(y + th) - F(y) = \int_0^t (\nabla F(y + sh), h) ds \leq t(M(h) + \varepsilon)$$

since $y + sh \in O_\delta(x)$ for all $s \in (0, t)$ by the choice of t . The continuity of F implies that the full inequality (8.11) even holds for *all* $y \in O_{\delta/2}(x)$ and *all* $t \in (0, \frac{\delta}{2\|h\|})$. Dividing by $t > 0$ and taking the lim sup over all $y \rightarrow x$ and $t \rightarrow 0$ now yields (8.10). Since $\varepsilon > 0$ was arbitrary, we conclude that $F^\circ(x; h) \leq M(h)$ for all $h \in \mathbb{R}^N$.

As in [Lemma 8.1](#), one can show that the mapping $h \mapsto M(h)$ is positive homogeneous, subadditive, and lower semi-continuous. We are thus finished if we can show that the set on the right-hand side of (8.9) – hereafter denoted by $\text{co } A$ – can be written as

$$\text{co } A = \{x^* \in \mathbb{R}^N : (x^*, h) \leq M(h) \text{ for all } h \in \mathbb{R}^N\}.$$

For this, we once again appeal to [Lemma 8.11](#) (since both sets are closed and convex). First, we note that the definition of the convex hull implies for all $h \in \mathbb{R}^N$ that

$$\sup_{x^* \in \text{co } A} (x^*, h) = \sup_{\substack{x_i^* \in A \\ \sum_i t_i = 1, t_i \geq 0}} \sum_i t_i (x_i^*, h) = \sup_{\sum_i t_i = 1, t_i \geq 0} \sum_i t_i \sup_{x_i^* \in A} (x_i^*, h) = \sup_{x^* \in A} (x^*, h)$$

since the sum is maximal if and only if each summand is maximal. Now we have that

$$M(h) = \limsup_{E_F \ni y \rightarrow x} (\nabla F(y), h) = \sup_{E_F \ni x_n \rightarrow x} (\lim_{n \rightarrow \infty} \nabla F(x_n), h) = \sup_{x^* \in A} (x^*, h),$$

and hence the claim follows from [Lemma 8.8](#). □

Finally, similarly to [Lemma 4.6](#) one can show the following pointwise characterization of the Clarke subdifferential of integral functionals with Lipschitz continuous integrands; see, e.g., [[Clarke 1990](#), Theorem 2.7.3, 2.7.5].

Theorem 8.18. *Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be Lipschitz continuous and $F : L^p(\Omega) \rightarrow \overline{\mathbb{R}}$ with $1 \leq p \leq \infty$ as in [Lemma 3.4](#). Then we have for all $u \in L^p(\Omega)$ with $q = \frac{p}{p-1}$ (where $q = 1$ for $p = \infty$) that*

$$\partial_C F(u) \subset \{u^* \in L^q(\Omega) : u^*(x) \in \partial_C f(u(x)) \text{ for almost every } x \in \Omega\}.$$

If f is regular in $u(x)$ for all $x \in \Omega$, then F is regular in u and equality holds.

9 SEMISMOOTH NEWTON METHODS

The proximal point and splitting methods in [Chapter 7](#) are generalizations of gradient methods and in general have the same only linear convergence. In this chapter, we will therefore consider a generalization of Newton methods which admit (locally) superlinear convergence.

As a motivation, we first consider the most general form of a Newton-type method. Let X and Y be Banach spaces and $F : X \rightarrow Y$ 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 $M_k := M(x^k) \in L(X, Y)$;
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.,

$$(9.1) \quad \lim_{k \rightarrow \infty} \frac{\|x^{k+1} - \bar{x}\|_X}{\|x^k - \bar{x}\|_X} = 0.$$

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

$$\begin{aligned} \|x^{k+1} - \bar{x}\|_X &= \|x^k - M(x^k)^{-1}F(x^k) - \bar{x}\|_X \\ &= \|M(x^k)^{-1} [F(x^k) - F(\bar{x}) - M(x^k)(x^k - \bar{x})]\|_X \\ &= \|M(\bar{x} + e^k)^{-1} [F(\bar{x} + e^k) - F(\bar{x}) - M(\bar{x} + e^k)e^k]\|_X \\ &\leq \|M(\bar{x} + e^k)^{-1}\|_{L(Y, X)} \|F(\bar{x} + e^k) - F(\bar{x}) - M(\bar{x} + e^k)e^k\|_Y. \end{aligned}$$

Hence, (9.1) holds under

- (i) a *regularity condition*: there exists a $C > 0$ with

$$\|M(\bar{x} + e^k)^{-1}\|_{L(Y, X)} \leq C \quad \text{for all } k \in \mathbb{N};$$

(ii) an *approximation condition*:

$$\lim_{k \rightarrow \infty} \frac{\|F(\bar{x} + e^k) - F(\bar{x}) - M(\bar{x} + e^k)e^k\|_Y}{\|e^k\|_X} = 0.$$

This motivates the following definition: We call $F : X \rightarrow Y$ *Newton differentiable* in $x \in X$ with *Newton derivative* $D_N F(x) \in L(X, Y)$ if

$$\lim_{\|h\|_X \rightarrow 0} \frac{\|F(x + h) - F(x) - D_N F(x)h\|_Y}{\|h\|_X} = 0.$$

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.¹

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*

$$(9.2) \quad x^{k+1} = x^k - D_N F(x^k)^{-1} F(x^k).$$

Its local superlinear convergence follows directly from the construction.

Theorem 9.1. *Let $F : X \rightarrow Y$ be Newton differentiable in $\bar{x} \in X$ with $F(\bar{x}) = 0$ with Newton derivative $D_N F(\bar{x})$. Assume further that there exist $\delta > 0$ and $C > 0$ with $\|D_N F(x)^{-1}\|_{L(Y, X)} \leq C$ for all $x \in O_\delta(\bar{x})$. Then the semismooth Newton method (9.2) converges 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 $x^0 \in O_\delta(\bar{x})$,

$$(9.3) \quad \|e^1\|_X \leq C \|F(\bar{x} + e^0) - F(\bar{x}) - D_N F(\bar{x} + e^0)e^0\|_Y.$$

Let now $\varepsilon \in (0, 1)$ be arbitrary. The Newton differentiability of F then implies that there exists a $\rho > 0$ such that

$$\|F(\bar{x} + h) - F(\bar{x}) - D_N F(\bar{x} + h)h\|_Y \leq \frac{\varepsilon}{C} \|h\|_X \quad \text{for all } \|h\|_X \leq \rho.$$

¹Here we follow [Chen et al. 2000; Ito & Kunisch 2008; Schiela 2008] and only consider single-valued Newton derivatives (called *slanting functions* in the first-named work). Alternatively, one could fix for each $x \in X$ a set $\partial_N F(x)$, from which the linear operator $M(x)$ in the Newton step has to be taken. If the approximation condition together with a boundedness condition hold *uniformly* for all $M \in \partial_N F(x)$, the function F is called *semismooth* (explaining the title of this chapter). This approach is followed in, e.g., [Mifflin 1977; Kummer 1988; Ulbrich 2011].

Hence, if we choose x^0 such that $\|\bar{x} - x^0\|_X \leq \min\{\delta, \rho\}$, the estimate (9.3) implies that $\|\bar{x} - x^1\|_X \leq \varepsilon \|\bar{x} - x^0\|_X$. By induction, we obtain from this that $\|\bar{x} - x^k\|_X \leq \varepsilon^k \|\bar{x} - x^0\|_X \rightarrow 0$. Since $\varepsilon \in (0, 1)$ was arbitrary, we can take in each step k a different $\varepsilon_k \rightarrow 0$, which shows that the convergence is in fact superlinear. \square

The remainder of this chapter is dedicated to the construction of Newton derivatives (although it should be pointed out that the verification of the approximation condition is usually the much more involved step in practice). We begin with the obvious connection with the Fréchet derivative.

Theorem 9.2. *If $F : X \rightarrow Y$ is continuously differentiable in $x \in X$, then F is also Newton differentiable in x with Newton derivative $D_N F(x) = F'(x)$.*

Proof. We have for arbitrary $h \in X$ that

$$\begin{aligned} \|F(x+h) - F(x) - F'(x)h\|_Y &\leq \|F(x+h) - F(x) - F'(x)h\|_Y \\ &\quad + \|F'(x) - F'(x+h)\|_{L(X,Y)} \|h\|_X, \end{aligned}$$

where the first summand is $o(\|h\|_X)$ by definition of the Fréchet derivative and the second by the continuity of F' . \square

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 9.3. *Let X, Y , and Z be Banach spaces, and let $F : X \rightarrow Y$ be Newton differentiable in $x \in X$ with Newton derivative $D_N F(x)$ and $G : Y \rightarrow Z$ be Newton differentiable in $y := F(x) \in Y$ 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 in 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 2.5. 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

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

with $r_1(t)/t \rightarrow 0$ for $t \rightarrow 0$. The Newton differentiability of F further implies that

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

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

$$\|g\|_Y \leq \|D_N F(x+h)\|_{L(X,Y)} \|h\|_X + r_2(\|h\|_X).$$

The uniform boundedness of $D_N F$ now implies that $\|g\|_Y \rightarrow 0$ for $\|h\|_X \rightarrow 0$. Hence,

$$\begin{aligned} \|(G \circ F)(x+h) - (G \circ F)(x) - D_N G(F(x+h))D_N F(x+h)h\|_Z \\ \leq \|G(y+g) - G(y) - D_N G(y+g)g\|_Z \\ + \|D_N G(y+g)[g - D_N F(x+h)h]\|_Z \\ \leq r_1(\|g\|_Y) + \|D_N G(y+g)\|_{L(Y,Z)} r_2(\|h\|_X), \end{aligned}$$

and the claim thus follows from the uniform boundedness of $D_N G$. \square

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 9.4. *Let $F_i : X \rightarrow Y_i$ be Newton differentiable with Newton derivative $D_N F_i$ for $1 \leq i \leq m$. Then,*

$$F : X \rightarrow (Y_1 \times \cdots \times Y_m), \quad x \mapsto (F_1(x), \dots, F_m(x))^T$$

is also Newton differentiable with Newton derivative

$$D_N F(x) = (D_N F_1(x), \dots, D_N F_m(x))^T.$$

Since the definition does not include a constructive prescription of Newton derivatives, the question remains how to obtain a candidate for which the approximation condition can be verified. For two classes of functions, such an explicit construction is known.

LOCALLY LIPSCHITZ CONTINUOUS FUNCTIONS ON \mathbb{R}^N

If $F : \mathbb{R}^N \rightarrow \mathbb{R}$ is locally Lipschitz continuous, candidates can be taken from the Clarke subdifferential, which has an explicit characterization by [Theorem 8.17](#). Under some additional assumptions, each candidate is indeed a Newton derivative.²

A function $F : \mathbb{R}^N \rightarrow \mathbb{R}$ is called *piecewise (continuously) differentiable* or *PC¹ function*, if

- (i) F is continuous on \mathbb{R}^N , and
- (ii) for all $x \in \mathbb{R}^N$ there exists an open neighborhood $U \subset \mathbb{R}^N$ of x and a finite set $\{F_i : U \rightarrow \mathbb{R}\}_{i \in I}$ of continuously differentiable functions with

$$F(y) \in \{F_i(y)\}_{i \in I} \quad \text{for all } y \in U.$$

²This is the original derivation of semismooth Newton methods.

In this case, we call F a *measurable selection* of the F_i in U . The set

$$I(x) := \{i \in I : F(x) = F_i(x)\}$$

is called the *active index set* at x . Since the F_i are continuous, we have that $F(y) \neq F_j(y)$ for all $j \notin I(x)$ and y 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*

$$I_e(x) := \{i \in I : x \in \text{cl}(\{y \in U : F(y) = F_i(y)\}^o)\} \subset I(x).$$

An example of an active but not essentially active index set is as follows: Consider the function $f : \mathbb{R} \rightarrow \mathbb{R}$, $t \mapsto \max\{0, t, t/2\}$, i.e., $f_1(t) = 0$, $f_2(t) = t$ and $f_3(t) = t/2$. Then, $I(0) = \{1, 2, 3\}$ but $I_e(0) = \{1, 2\}$, since f_3 is active only in $t = 0$ and hence $\{t \in \mathbb{R} : f(t) = f_3(t)\}^o = \emptyset = \text{cl} \emptyset$.

PC¹ functions are always locally Lipschitz continuous; see [Scholtes 2012, Corollary 4.1.1].

Theorem 9.5. *Let $F : \mathbb{R}^N \rightarrow \mathbb{R}$ be piecewise differentiable. Then F is locally Lipschitz continuous in all $x \in \mathbb{R}^N$ with local constant $L(x) = \max_{i \in I(x)} L_i$.*

This yields the following explicit characterization of the Clarke subdifferential of PC¹ functions.

Theorem 9.6. *Let $F : \mathbb{R}^N \rightarrow \mathbb{R}$ be piecewise differentiable and $x \in \mathbb{R}^N$. Then*

$$\partial_C F(x) = \text{co} \{\nabla F_i(x) : i \in I_e(x)\}.$$

Proof. Let $x \in \mathbb{R}^N$ be arbitrary. By Theorem 8.17 it suffices to show that

$$\left\{ \lim_{n \rightarrow \infty} \nabla F(x_n) : x_n \rightarrow x, x_n \notin E_F \right\} = \{\nabla F_i(x) : i \in I_e(x)\}.$$

For this, let $\{x_n\}_{n \in \mathbb{N}} \subset \mathbb{R}^N$ be a sequence with $x_n \rightarrow x$, F is differentiable in x_n for all $n \in \mathbb{N}$, and $\nabla F(x_n) \rightarrow x^* \in \mathbb{R}^n$. Since F is differentiable in x_n , it must hold that $F(y) = F_{i_n}(y)$ for some $i_n \in I$ and all y 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 including another set of zero measure to E_F). 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

$$x^* = \lim_{k \rightarrow \infty} \nabla F(x_{n_k}) = \lim_{k \rightarrow \infty} \nabla F_i(x_{n_k}) \in \{\nabla F_i(x) : i \in I_e(x)\}.$$

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 $x_n \rightarrow x$ and $F = F_i$ in a sufficiently small neighborhood of each x_n . The continuous differentiability of the F_i thus implies that $\nabla F(x_n) = \nabla F_i(x_n)$ for all $n \in \mathbb{N}$ and hence that

$$\nabla F_i(x) = \lim_{n \rightarrow \infty} \nabla F_i(x_n) = \lim_{n \rightarrow \infty} \nabla F(x_n). \quad \square$$

From this, we obtain the Newton differentiability of PC^1 functions.

Theorem 9.7. *Let $F : \mathbb{R}^N \rightarrow \mathbb{R}$ 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 $x + h \in U$. By [Theorem 9.6](#), every $D_N F(x + h) \in \partial_C F(x + h)$ is of the form

$$D_N F(x + h) = \sum_{i \in I_e(x+h)} t_i \nabla F_i(x + h) \quad \text{for} \quad \sum_{i \in I_e(x+h)} t_i = 1, t_i \geq 0.$$

Since all F_i are continuous, we have for all $h \in \mathbb{R}^N$ sufficiently small that $I_e(x+h) \subset I(x+h) \subset I(x)$. Hence, $F(x + h) = F_i(x + h)$ and $F(x) = F_i(x)$ for all $i \in I_e(x + h)$. [Theorem 9.2](#) then yields that

$$|F(x + h) - F(x) - D_N F(x + h)h| = \sum_{i \in I_e(x+h)} t_i |F_i(x + h) - F_i(x) - \nabla F_i(x + h)h| = o(\|h\|),$$

since all F_i are continuously differentiable by assumption. \square

A natural application of the above are proximal point reformulations of optimality conditions for convex optimization problems.

Example 9.8. We consider the minimization of $F + G$ for a twice continuously differentiable functional $F : \mathbb{R}^N \rightarrow \mathbb{R}$ and $G = \|\cdot\|_1$. Proceeding as in the derivation of the forward-backward splitting [\(7.4\)](#), we write the necessary optimality condition $0 \in \partial(F + G)(\bar{x})$ equivalently as

$$\bar{x} - \text{prox}_{\gamma G}(\bar{x} - \gamma \nabla F(\bar{x})) = 0$$

for any $\gamma > 0$. By [Example 6.14](#) (ii), the proximal point mapping for G is given componentwise as

$$[\text{prox}_{\gamma G}(x)]_i = \begin{cases} x_i - \gamma & \text{if } x_i > \gamma, \\ 0 & \text{if } x_i \in [-\gamma, \gamma], \\ x_i + \gamma & \text{if } x_i < -\gamma, \end{cases}$$

which is clearly piecewise differentiable. [Theorem 9.6](#) thus yields (also componentwise) that

$$[\partial_C(\text{prox}_{\gamma G})(x)]_i = \begin{cases} \{1\} & \text{if } |x_i| > \gamma, \\ \{0\} & \text{if } |x_i| < \gamma, \\ [0, 1] & \text{if } |x_i| = \gamma. \end{cases}$$

By [Theorems 9.4](#) and [9.7](#), a possible Newton derivative is therefore given by

$$[D_N \text{prox}_{\gamma G}(x)h]_i = [\chi_{\{|x| \geq \gamma\}}h]_i := \begin{cases} h_i & \text{if } |x_i| \geq \gamma, \\ 0 & \text{if } |x_i| < \gamma. \end{cases}$$

(The choice of in which case to include the equality here is arbitrary.) Now, $D_N \text{prox}_{\gamma G}(x)$ and $D_N(\nabla F)(x) = \nabla^2 F(x)$ are locally uniformly bounded (obviously from the characterization and the continuous differentiability, respectively), and using the chain rule from [Theorem 9.3](#) and rearranging yields the semismooth Newton step

$$\left(\chi_{\mathcal{A}_k} + \gamma \chi_{\mathcal{I}_k} \nabla^2 F(x^k) \right) s^k = -x^k + \text{prox}_{\gamma G}(x^k - \gamma \nabla F(x^k)),$$

where we have defined the *active* and *inactive sets*, respectively, as

$$\mathcal{A}_k := \{i \in \{1, \dots, n\} : |x_i^k - \gamma[\nabla F(x^k)]_i| < \gamma\}, \quad \mathcal{I}_k := \{1, \dots, n\} \setminus \mathcal{A}_k.$$

If we now also partition s^k as well as the right-hand side in active and inactive components using the case distinction in the characterization of $\text{prox}_{\gamma G}$ (which follows the same partition), we can rearrange this linear system into blocks corresponding to active and inactive components to observe that the Newton step coincides with an *active set strategy* similar to those used for solving quadratic subproblems in sequential programming methods with inequality constraints; cf. [\[Ito & Kunisch 2008, Chapter 8.4\]](#).

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 again consider for an open and bounded domain $\Omega \subset \mathbb{R}^N$, a Carathéodory function $f : \Omega \times \mathbb{R} \rightarrow \mathbb{R}$ (i.e., f is measurable in x and continuous in z), and $1 \leq p, q \leq \infty$ the corresponding superposition operator

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

The goal is now to similarly obtain a Newton derivative $D_N F$ for F as a superposition operator defined by the Newton derivative $D_N f(x, z)$ for $z \mapsto f(x, z)$. 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 9.8](#)). 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.

Under certain growth conditions on f and $D_N f$,³ we can transfer the Newton differentiability of f to F , but we again have to take a two norm discrepancy into account.

³which can be significantly relaxed; see [\[Schiela 2008, Proposition A.1\]](#)

Theorem 9.9. *Let $f : \Omega \times \mathbb{R} \rightarrow \mathbb{R}$ be a Carathéodory function. Furthermore, assume that*

- (i) $z \mapsto f(x, z)$ is uniformly Lipschitz continuous for almost every $x \in \Omega$ and $f(x, 0)$ 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 $1 \leq q < p < \infty$, the corresponding superposition operator $F : L^p(\Omega) \rightarrow L^q(\Omega)$ is Newton differentiable with Newton derivative

$$D_N F : L^p(\Omega) \rightarrow L(L^p(\Omega), L^q(\Omega)), \quad [D_N F(u)h](x) = D_N f(x, u(x))h(x)$$

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

$$|f(x, z)| \leq |f(x, 0)| + L|z| \leq C + L|z|^{q/q} \quad \text{for almost every } x \in \Omega, z \in \mathbb{R},$$

and hence the growth condition (2.5) for all $1 \leq q < \infty$. Due to the continuous embedding $L^p(\Omega) \hookrightarrow L^q(\Omega)$ for all $1 \leq q < p < \infty$, the superposition operator $F : L^p(\Omega) \rightarrow L^q(\Omega)$ is therefore well-defined and continuous by [Theorem 2.8](#).

For any measurable $u : \Omega \rightarrow \mathbb{R}$, we have that $x \mapsto D_N f(x, u(x))$ is by definition 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 2.9](#), we deduce that the corresponding superposition operator $D_N F : L^p(\Omega) \rightarrow L^s(\Omega)$ is well-defined and continuous for $s := \frac{pq}{p-q}$, and that for any $u \in L^p(\Omega)$, the mapping $h \mapsto D_N F(u)h$ defines a bounded linear operator $D_N F(u) : L^p(\Omega) \rightarrow L^q(\Omega)$. (This time, we do not distinguish in notation between the linear operator and the function that defines this operator by pointwise multiplication.)

To show that $D_N F(u)$ is a Newton derivative for F in $u \in L^p(\Omega)$, we consider the pointwise residual

$$r : \Omega \times \mathbb{R} \rightarrow \mathbb{R}, \quad r(x, z) := \begin{cases} \frac{|f(x, z) - f(x, u(x)) - D_N f(x, z)(z - u(x))|}{|z - u(x)|} & \text{if } z \neq u(x), \\ 0 & \text{if } z = u(x). \end{cases}$$

Since f is a Carathéodory function and $D_N f$ is a Baire–Carathéodory function, the function $x \mapsto r(x, \tilde{u}(x)) =: R(\tilde{u})$ is measurable for any measurable $\tilde{u} : \Omega \rightarrow \mathbb{R}$ (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

$$(9.4) \quad |[R(\tilde{u})](x)| = \frac{|f(x, \tilde{u}(x)) - f(x, u(x)) - D_N f(x, \tilde{u}(x))(\tilde{u}(x) - u(x))|}{|\tilde{u}(x) - u(x)|} \leq L + C$$

and thus that $R(\tilde{u}) \in L^\infty(\Omega)$. Hence, the superposition operator $R : L^p(\Omega) \rightarrow L^s(\Omega)$ is well-defined.

Let now $\{u_n\}_{n \in \mathbb{N}} \subset L^p(\Omega)$ be a sequence with $u_n \rightarrow u \in L^p(\Omega)$. Then there exists a subsequence, again denoted by $\{u_n\}_{n \in \mathbb{N}}$, with $u_n(x) \rightarrow u(x)$ for almost every $x \in \Omega$. Since $z \mapsto f(x, z)$ is Newton differentiable almost everywhere, we have by definition that $r(x, u_n(x)) \rightarrow 0$ for almost every $x \in \Omega$. Together with the boundedness from (9.4), Lebesgue's dominated convergence theorem therefore yields that $R(u_n) \rightarrow 0$ in $L^s(\Omega)$ (and hence along the full sequence since the limit is unique).⁴ 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 $\|h\|_{L^p} \rightarrow 0$, we have that $\|R(u + h)\|_{L^s} \rightarrow 0$ and hence by definition the Newton differentiability of F in u with Newton derivative $h \mapsto D_N F(u)h$ as claimed. \square

For $p = q \in [1, \infty]$, however, the claim is false in general, as can be shown by counterexamples; see, e.g., [Ito & Kunisch 2008, Example 8.3].

Due to the two norm discrepancy, we can no longer apply the semismooth Newton method directly to proximal point reformulations in function spaces. We therefore have to fall back on the Moreau–Yosida regularization.

Example 9.10. We consider as in Example 9.8 the minimization of $F + G$ for a twice continuously differentiable functional $F : L^2(\Omega) \rightarrow \mathbb{R}$ and $G = \|\cdot\|_{L^1}$. The proximal point reformulation of $0 \in \partial(F + G)(\bar{u})$,

$$\bar{u} - \text{prox}_{\gamma G}(\bar{u} - \gamma \nabla F(\bar{u})) = 0,$$

now has to be considered as an equation in $L^2(\Omega)$; however, $\text{prox}_{\gamma G}$ is *not* Newton differentiable from $L^2(\Omega)$ to $L^2(\Omega)$. We therefore replace in the original optimality conditions

$$\begin{cases} -\bar{p} = \nabla F(\bar{u}), \\ \bar{u} \in \partial G^*(\bar{p}), \end{cases}$$

the subdifferential of G^* with its Moreau–Yosida regularization $H_\gamma := (\partial G^*)_\gamma$, which by Corollary 6.15 and Example 6.19 is given pointwise as $[H_\gamma(p)](x) = h_\gamma(p(x))$ for

$$h_\gamma : \mathbb{R} \rightarrow \mathbb{R}, \quad t \mapsto \begin{cases} \frac{1}{\gamma}(t - 1) & \text{if } t > 1, \\ 0 & \text{if } t \in [-1, 1], \\ \frac{1}{\gamma}(t + 1) & \text{if } t < -1. \end{cases}$$

⁴This step fails for $F : L^\infty(\Omega) \rightarrow L^\infty(\Omega)$ since pointwise convergence and boundedness together do not imply uniform convergence almost everywhere.

This function is clearly piecewise differentiable, and [Theorem 9.6](#) yields that

$$\partial_C h_\gamma(t) = \begin{cases} \left\{ \frac{1}{\gamma} \right\} & \text{if } |t| > 1, \\ \{0\} & \text{if } |t| < 1, \\ \left[0, \frac{1}{\gamma} \right] & \text{if } |t| = 1. \end{cases}$$

By [Theorems 9.4](#) and [9.7](#), a possible Newton derivative is therefore given by

$$D_N h_\gamma(t)h = \frac{1}{\gamma} \chi_{\{|t| \geq 1\}} h = \begin{cases} \frac{1}{\gamma} h & \text{if } |t| \geq 1, \\ 0 & \text{if } |t| < 1. \end{cases}$$

The function $D_N h_\gamma$ is now uniformly bounded (by $\frac{1}{\gamma}$) and can be approximated by the obvious pointwise limit of continuous functions. By [Theorem 9.9](#), the superposition operator $H_\gamma : L^p(\Omega) \rightarrow L^2(\Omega)$ is therefore Newton differentiable for all $p > 2$, and a possible Newton derivative is given by

$$[D_N H_\gamma(p)h](x) = \frac{1}{\gamma} \chi_{\{|p| \geq 1\}}(x)h(x).$$

Assume now that F is such that $\bar{p} = -\nabla F(\bar{u}) \in L^p(\Omega)$ for some $p > 2$. (This is the case, e.g., if F involves the solution operator to a partial differential equation.) Then, the reduced regularized optimality condition

$$u_\gamma - H_\gamma(-\nabla F(u_\gamma)) = 0$$

is Newton differentiable by [Theorems 9.2](#) and [9.3](#), and we arrive at the semismooth Newton step

$$\left(\text{Id} + \frac{1}{\gamma} \chi_{\{|\nabla F(u^k)| \geq 1\}} \nabla^2 F(u^k) \right) s^k = -u^k + H_\gamma(-\nabla F(u^k)).$$

In practice, the radius of convergence for semismooth Newtons applied to such a Moreau–Yosida regularization shrinks with $\gamma \rightarrow 0$. A possible way of dealing with this 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 (for which it 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).

BIBLIOGRAPHY

- H. W. ALT (2016), *Linear Functional Analysis, An application-oriented introduction*, Universitext, London: Springer, DOI: [10.1007/978-1-4471-7280-2](https://doi.org/10.1007/978-1-4471-7280-2).
- J. APPELL & P. ZABREIKO (1990), *Nonlinear Superposition Operators*, New York: Cambridge University Press.
- H. ATTOUCH, G. BUTTAZZO & G. MICHAILLE (2006), *Variational Analysis in Sobolev and BV Spaces*, vol. 6, MPS/SIAM Series on Optimization, Philadelphia, PA: Society for Industrial & Applied Mathematics (SIAM), DOI: [10.1137/1.9780898718782](https://doi.org/10.1137/1.9780898718782).
- H. H. BAUSCHKE & P. L. COMBETTES (2017), *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2nd ed., CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, New York: Springer, DOI: [10.1007/978-3-319-48311-5](https://doi.org/10.1007/978-3-319-48311-5).
- A. BECK & M. TEOULLE (2009), A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM J. Imaging Sci.* 2(1), 183–202, DOI: [10.1137/080716542](https://doi.org/10.1137/080716542).
- H. BREZIS (2010), *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, New York: Springer, DOI: [10.1007/978-0-387-70914-7](https://doi.org/10.1007/978-0-387-70914-7).
- M. BROKATE (2014), *Konvexe Analysis und Evolutionsprobleme*, Lecture notes, Zentrum Mathematik, TU München, URL: http://www-m6.ma.tum.de/~brokate/cev_ss14.pdf.
- A. CEGIELSKI (2012), *Iterative methods for fixed point problems in Hilbert spaces*, vol. 2057, Lecture Notes in Mathematics, Springer, Heidelberg, DOI: [10.1007/978-3-642-30901-4](https://doi.org/10.1007/978-3-642-30901-4).
- A. CHAMBOLLE & T. POCK (2011), A first-order primal-dual algorithm for convex problems with applications to imaging, *J Math Imaging Vis* 40(1), 120–145, DOI: [10.1007/s10851-010-0251-1](https://doi.org/10.1007/s10851-010-0251-1).
- X. CHEN, Z. NASHED & L. QI (2000), Smoothing methods and semismooth methods for nondifferentiable operator equations, *SIAM J. Numer. Anal.* 38(4), 1200–1216, DOI: [10.1137/s0036142999356719](https://doi.org/10.1137/s0036142999356719).
- F. CLARKE (2013), *Functional Analysis, Calculus of Variations and Optimal Control*, London: Springer, DOI: [10.1007/978-1-4471-4820-3](https://doi.org/10.1007/978-1-4471-4820-3).
- F. H. CLARKE (1990), *Optimization and Nonsmooth Analysis*, vol. 5, Classics Appl. Math. Philadelphia, PA: SIAM, DOI: [10.1137/1.9781611971309](https://doi.org/10.1137/1.9781611971309).
- E. DIBENEDETTO (2002), *Real analysis*, Birkhäuser Boston, Inc., Boston, MA, DOI: [10.1007/978-1-4612-0117-5](https://doi.org/10.1007/978-1-4612-0117-5).

- J. ECKSTEIN & D. P. BERTSEKAS (1992), On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators, *Mathematical Programming* 55(1-3), 293–318, DOI: [10.1007/bf01581204](https://doi.org/10.1007/bf01581204).
- B. HE & X. YUAN (2012), Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective, *SIAM J. Imag. Sci.* 5(1), 119–149, DOI: [10.1137/100814494](https://doi.org/10.1137/100814494).
- J. HEINONEN (2005), *Lectures on Lipschitz analysis*, vol. 100, Rep. Univ. Jyväskylä Dept. Math. Stat. University of Jyväskylä, URL: <http://www.math.jyu.fi/research/reports/rep100.pdf>.
- K. ITO & K. KUNISCH (2008), *Lagrange Multiplier Approach to Variational Problems and Applications*, vol. 15, Advances in Design and Control, Philadelphia, PA: SIAM, DOI: [10.1137/1.9780898718614](https://doi.org/10.1137/1.9780898718614).
- B. KUMMER (1988), Newton’s method for non-differentiable functions, *Mathematical Research* 45, 114–125.
- R. MIFFLIN (1977), Semismooth and semiconvex functions in constrained optimization, *SIAM J. Control Optimization* 15(6), 959–972, DOI: [10.1137/0315061](https://doi.org/10.1137/0315061).
- Y. E. NESTEROV (1983), A method for solving the convex programming problem with convergence rate $O(1/k^2)$, *Soviet Math. Doklad.* 27(2), 372–376.
- Y. NESTEROV (2004), *Introductory Lectures on Convex Optimization*, vol. 87, Applied Optimization, Kluwer Academic Publishers, Boston, MA, DOI: [10.1007/978-1-4419-8853-9](https://doi.org/10.1007/978-1-4419-8853-9).
- N. PARIKH & S. BOYD (2014), Proximal algorithms, *Foundations and Trends in Optimization* 1(3), 123–231, DOI: [10.1561/24000000003](https://doi.org/10.1561/24000000003).
- R. T. ROCKAFELLAR (1976), Integral functionals, normal integrands and measurable selections, in: *Nonlinear Operators and the Calculus of Variations (Summer School, Univ. Libre Bruxelles, Brussels, 1975)*, vol. 543, Lecture Notes in Math. Berlin: Springer, 157–207, DOI: [10.1007/bfb0079944](https://doi.org/10.1007/bfb0079944).
- W. RUDIN (1991), *Functional Analysis*, 2nd ed., New York: McGraw-Hill.
- A. RUSZCZYŃSKI (2006), *Nonlinear Optimization*, Princeton, NJ: Princeton University Press.
- A. SCHIELA (2008), A simplified approach to semismooth Newton methods in function space, *SIAM J. Opt.* 19(3), 1417–1432, DOI: [10.1137/060674375](https://doi.org/10.1137/060674375).
- W. SCHIROTZEK (2007), *Nonsmooth Analysis*, Universitext, Berlin: Springer, DOI: [10.1007/978-3-540-71333-3](https://doi.org/10.1007/978-3-540-71333-3).
- S. SCHOLTES (2012), *Introduction to piecewise differentiable equations*, Springer Briefs in Optimization, Springer, New York, DOI: [10.1007/978-1-4614-4340-7](https://doi.org/10.1007/978-1-4614-4340-7).
- M. ULBRICH (2002), Semismooth Newton methods for operator equations in function spaces, *SIAM J. Optim.* 13(3), 805–842 (2003), DOI: [10.1137/s1052623400371569](https://doi.org/10.1137/s1052623400371569).
- M. ULBRICH (2011), *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, vol. 11, MOS-SIAM Series on Optimization, Philadelphia, PA: SIAM, DOI: [10.1137/1.9781611970692](https://doi.org/10.1137/1.9781611970692).

K. YOSIDA (1995), *Functional Analysis*, Classics in Mathematics, Reprint of the sixth (1980) edition, Berlin: Springer-Verlag, DOI: [10.1007/978-3-642-61859-8](https://doi.org/10.1007/978-3-642-61859-8).