NONSMOOTH ANALYSIS AND OPTIMIZATION

LECTURE NOTES, WINTER 2021/22

Christian Clason

February 18, 2022

Institute of Mathematics and Scientific Computing
University of Graz

CONTENTS

I BACKGROUND

II CONVEX ANALYSIS

III LIPSCHITZ ANALYSIS

INTRODUCTION

Optimization is concerned with finding solutions to problems of the form

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

for a function $F : X \to \mathbb{R}$ 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
$$F(\bar{x}) \leq F(x) \qquad \text{for all } x \in U?$$
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
$$x^{k+1} = x^k - t_k F'(x^k), \qquad k = 0, \dots,$$
for suitable step sizes t_k, we have that $x^k \to \bar{x}$ for $k \to \infty$.

   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
$$x^{k+1} = x^k - F''(x^k)^{-1} F'(x^k), \qquad k = 0, \dots,$$
we have that $x^k \to \bar{x}$ for $k \to \infty$.

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 $F : X \to Y$, $\alpha \geq 0$ and $p, q \in [1,\infty)$ (in particular, $p = 1$ and/or $q = 1$). Such problems are ubiquitous in inverse problems, imaging, and optimal control of differential equations. Hence, we consider optimization in *infinite-dimensional* function spaces; i.e., we are looking for functions as minimizers. The main benefit (beyond the frequently cleaner notation) is that the developed algorithms become *discretization independent*: they can be applied to any (reasonable) finite-dimensional approximation, and the details – in particular, the fineness – of the approximation do not influence the convergence behavior of the 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 mainly based on [Brokate 2014; Schirotzek 2007; Attouch, Buttazzo & Michaille 2006; Bauschke & Combettes 2017; Clarke 2013; Ulbrich 2002; Schiela 2008]. All mistakes are of course 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] and to [Clason 2020].

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 $\|\cdot\| : X \to \mathbb{R}^+ := [0,\infty)$ is called a *norm* (on X), if for all $x \in X$ there holds

(i) $\|\lambda x\| = |\lambda|\|x\|$ for all $\lambda \in \mathbb{K}$,

(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 $X = \mathbb{R}^N$:

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

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

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

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

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

(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

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

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

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

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

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

- $\operatorname{dom} F := U$ the *domain of definition* of F;
- $\ker F := \{x \in U : F(x) = 0\}$ *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 $F : U \to Y$

- *continuous* in $x \in U$, if for all $\varepsilon > 0$ there exists a $\delta > 0$ with
$$\|F(x) - F(z)\|_Y \leq \varepsilon \qquad \text{for all } z \in U \text{ with } \|x - z\|_X \leq \delta;$$
- *Lipschitz continuous*, if there exists an $L > 0$ (called *Lipschitz constant*) with
$$\|F(x_1) - F(x_2)\|_Y \leq L\|x_1 - x_2\|_X \qquad \text{for all } x_1, x_2 \in U.$$
- *locally Lipschitz continuous* in $x \in U$, if there exists a $\delta > 0$ and a $L = L(x, \delta) > 0$ with
$$\|F(x) - F(z)\|_Y \leq L\|x - z\|_X \qquad \text{for all } z \in U \text{ with } \|x - z\|_X \leq \delta.$$

If $T : X \to Y$ is linear, continuity is equivalent to the existence of a constant $C > 0$ with
$$\|Tx\|_Y \leq C\|x\|_X \qquad \text{for all } x \in X.$$

For this reason, continuous linear mappings are called *bounded*; one speaks of a bounded linear *operator*. The space $L(X, Y)$ 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 $T^{-1} : Y \to X$ is continuous if and only if there exists a $c > 0$ with
$$c\|x\|_X \leq \|Tx\|_Y \qquad \text{for all } x \in X.$$
In this case, $\|T^{-1}\|_{L(Y,X)} = c^{-1}$ 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 $x_n \to x$, if
$$\lim_{n \to \infty} \|x_n - x\|_X = 0.$$

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 $F : X \to Y$ is continuous if and only if $x_n \to x$ implies $F(x_n) \to F(x)$, and *closed*, if $x_n \to x$ and $F(x_n) \to y$ imply $F(x) = y$ (i.e., 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 a 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*

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

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

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

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

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

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

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

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

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

(iii) Analogously, $L^p(\Omega)^* \cong L^q(\Omega)$ 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*

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

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

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

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

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

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

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's Theorem.*

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

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

Proof. Theorem 1.5 (i) yields the existence of x^* and λ satisfying the claim for all $x \in A^o$ (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 $x \in A \setminus A^o$. Since A^o is nonempty, there exists an $x_0 \in A^o$, 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 $x(t) := tx_0 + (1-t)x \in A^o$ for all $t \in (0,1)$.

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

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

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*

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

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

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

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

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

Weak convergence generalizes the concept of componentwise convergence in $\mathbb{R}^n$, which – as can be seen from the proof of the Heine–Borel Theorem – is the appropriate concept in the context of compactness. Strong convergence 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 $x_n \to x$ and $x_n^* \rightharpoonup^* x^*$ or $x_n \rightharpoonup x$ and $x_n^* \to x^*$, then $\langle x_n^*, x_n \rangle_X \to \langle x^*, x \rangle_X$. However, the duality pairing of weak(-∗) convergent sequences does not converge in general.

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

Theorem 1.8 (Eberlein–Šmulyan). *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

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

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

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

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

Theorem 1.10 (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 $1 \le p < \infty$ are separable; also, ℓ^p is separable for $1 \le p < \infty$. 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*

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

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

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

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 $(\cdot,\cdot) : X \times X \to \mathbb{R}$ 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*

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

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

Theorem 1.12 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

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

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* $T^\star \in L(Y, X)$ is defined via

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

If $T^\star = T$, the operator T is called *self-adjoint*. Both definitions of adjoints are related via $T^\star = J_X T^* J_Y^{-1}$. 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 $F : U \to \mathbb{R}$ for a subset U of a Banach space X. Answering such questions is one of the goals of the *calculus of variations*.

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

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

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

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

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

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

2.1 THE DIRECT METHOD

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 $\overline{F}$. Since we are only interested in minimizers, we only require a “one-sided” continuity: We call F *lower semicontinuous* in $x \in X$ if

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

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

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

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

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

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

Proof. The proof can be separated into three steps.

(i) *Pick a minimizing sequence.*

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

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

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

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

Assume to the contrary that $\{x_n\}_{n\in\mathbb{N}}$ is unbounded, i.e., that $\|x_n\|_X \to \infty$ for $n \to \infty$. The coercivity of F then implies that $F(x_n) \to \infty$ as well, in contradiction to $F(x_n) \to M < \infty$ by definition of the minimizing sequence. Hence, the sequence is bounded, i.e., there is an $M > 0$ with $\|x_n\|_X \leq M$ 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 $\{x_{n_k}\}_{k\in\mathbb{N}}$ with limit $\bar{x} \in X$. (This limit is a candidate for the minimizer.)

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

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

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

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

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

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 lim inf conditions). Since in the cases of interest to us no more than boundedness of a minimizing sequence can be expected, we cannot use a finer than the weak topology. We thus have to ask whether a sufficiently large class of (interesting) functionals are weakly lower semicontinuous.

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

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

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

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

(i) αF for all $\alpha \geq 0$;

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

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

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

(v) $x \mapsto \sup_{i\in I} F_i(x)$ with $F_i : X \to \overline{\mathbb{R}}$ weakly lower semicontinuous for 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.

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

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

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

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

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

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

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

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

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

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

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

Proof. Coercivity and $\operatorname{dom}\|\cdot\|_X = X$ follow directly from the definition. Weak lower semicontinuity follows from Lemma 2.2 (v) and Corollary 1.7 since

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

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 $F : X \to \overline{\mathbb{R}}$ 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 semicontinuous if U is convex and closed;

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

(iii) coercive if U is bounded.

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

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

Let now $x \notin U$. Since U is convex and closed and hence by Lemma 1.9 also weakly closed, there must be an $N \in \mathbb{N}$ with $x_n \notin U$ for all $n \geq N$ (otherwise we could – by passing to a subsequence if necessary – construct a sequence with $x_n \rightharpoonup x \in U$, in contradiction to the assumption). Thus, $\delta_U(x_n) = \infty$ for all $n \geq N$, and therefore

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

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

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

- If the one-sided limit

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

  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

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

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

- If additionally

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

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

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

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 $F : X \to Y$ be Fréchet differentiable in $x \in X$ and $G : Y \to Z$ 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 $x + h \in \operatorname{dom} F$ we have

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

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

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

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

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

with $r_2(t)/t \to 0$ for $t \to 0$. In particular, the reverse triangle inequality yields

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

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

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 $f : [a, b] \to X$ be continuous. Then the *Bochner integral* $\int_a^b f(t)\,dt \in X$ is well-defined and by construction satisfies

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

as well as

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

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

Theorem 2.6. *Let $F : U \to Y$ 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

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

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. □

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

[3] The *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.

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

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

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

$$0 \leq \frac{F(\bar{x} + th) - F(\bar{x})}{t} \to F'(\bar{x}; h) = \langle DF(\bar{x}), h\rangle_X \quad \text{for } t \to 0, \tag{2.4}$$

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

Of particular relevance in optimization is of course the special case $F : X \to \mathbb{R}$, where $DF(x) \in L(X; \mathbb{R}) = X^*$ (if the Gâteaux derivative exists). Note that since the Gâteaux derivative of $F : X \to \mathbb{R}$ is an element of X^*, it cannot be added to elements in X (as required for, e.g., a steepest descent method). However, in Hilbert spaces (and in particular in $\mathbb{R}^N$), we can use the Fréchet–Riesz Theorem 1.12 to identify $DF(x) \in X^*$ with an element $\nabla F(x) \in X$, called the *gradient* of F at x, in a canonical way via

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

We illustrate this with a simple example.

Example 2.8. Let $F(x) = \frac{1}{2}\|x\|_X^2 = \frac{1}{2}(x, x)_X$. Then we have for all $x, h \in X$ that

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

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

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

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

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

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

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

Example 2.9. Let $M \in L(X;X)$ be self-adjoint and positive definite (and thus continuously invertible). Then $(x,y)_Z := (Mx,y)_X$ also defines an inner product on the vector space X and induces an (equivalent) norm $\|x\|_Z := (x,x)_Z^{1/2}$ on X. Hence $(X,(\cdot,\cdot)_Z)$ is a Hilbert space as well, which we will denote by Z. Consider now the functional $\tilde F : Z \to \mathbb{R}$ with $\tilde F(x) := \frac{1}{2}\|x\|_X^2$ (which is well-defined since $\|\cdot\|_X$ is also an equivalent norm on Z). Then, the derivative $D\tilde F(x) \in Z^*$ is still given by $\langle D\tilde F(x),h\rangle_Z = (x,h)_X$ for all $h \in Z$ (or, equivalently, for all $h \in X$ since we defined Z via the same vector space). However, $\nabla\tilde F(x) \in Z$ is now characterized by

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

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

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

2.3 SUPERPOSITION OPERATORS

A special class of operators on function spaces arise from pointwise application of a real-valued function, e.g., $u(x) \mapsto \sin(u(x))$. We thus consider for $f : \Omega\times\mathbb{R} \to \mathbb{R}$ with $\Omega \subset \mathbb{R}^n$ open and bounded as well as $p,q \in [1,\infty]$ the corresponding *superposition* or *Nemytskii operator*

$$F : L^p(\Omega) \to L^q(\Omega), \qquad [F(u)](x) = f(x,u(x)) \quad \text{for almost every } x \in \Omega. \tag{2.5}$$

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

$$|f(x,z)| \leq a(x) + b(x)|z|^{p/q}. \tag{2.6}$$

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

Theorem 2.10. *If the Carathéodory function $f : \Omega\times\mathbb{R} \to \mathbb{R}$ satisfies the growth condition (2.6) for $p,q \in [1,\infty)$, then the superposition operator $F : L^p(\Omega) \to L^q(\Omega)$ defined via (2.5) 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 (2.6) and the triangle inequality that

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

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

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

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

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

In fact, the growth condition (2.6) 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 (2.6) (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 $F : L^p(\Omega) \to L^q(\Omega)$ and a direction $h \in L^p(\Omega)$, the pointwise(!) product $F'(u)h$ has to be in $L^q(\Omega)$. This leads to additional conditions on the superposition operator F' defined by f', which is known as *two norm discrepancy*.

Theorem 2.11. *Let $f : \Omega \times \mathbb{R} \to \mathbb{R}$ be a Carathéodory function that satisfies the growth condition (2.6) for $1 \leq q < p < \infty$. If the partial derivative f'_z is a Carathéodory function as well and satisfies (2.6) for $p' = p - q$, the superposition operator $F : L^p(\Omega) \to L^q(\Omega)$ is continuously Fréchet differentiable, and its derivative in $u \in L^p(\Omega)$ in direction $h \in L^p(\Omega)$ is given by*

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

Proof. Theorem 2.10 yields that for $r := \frac{pq}{p-q}$ (i.e., $\frac{r}{p} = \frac{p'}{q}$), the superposition operator

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

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

$$\|G(u)h\|_{L^q} \leq \|G(u)\|_{L^r} \|h\|_{L^p} \qquad \text{for all } h \in L^p(\Omega), \tag{2.7}$$

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

Let now $h \in L^p(\Omega)$ be arbitrary. Since $z \mapsto f(x, z)$ is continuously differentiable by assumption, the classical mean value theorem together with (2.3) and (2.7) implies that

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

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

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

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 $F : X \to \overline{\mathbb{R}}$ is called *convex* if

$$F(\lambda x + (1-\lambda)y) \leq \lambda F(x) + (1-\lambda)F(y) \quad \text{for all } x, y \in X \text{ and } \lambda \in [0,1] \tag{3.1}$$

(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 $F : X \to \overline{\mathbb{R}}$ is based on its *epigraph*

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

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

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

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

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

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

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

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

i.e., that

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

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

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

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

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

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

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

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

We now distinguish two cases.

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

b) $x\notin\operatorname{dom}F$: 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 semicontinuity and weak closedness follows in exactly the same way. □

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

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

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

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

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

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

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

(ii) F is (weakly) lower semicontinuous 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 $\lambda \in (0,1)$,

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

Further examples can be constructed as in Lemma 2.2 through the following operations.

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

(i) αF for all $\alpha \geq 0$;

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

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

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

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

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

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

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

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

Proof. Since affine functionals are convex and continuous, Lemma 3.4 (v) and Lemma 2.2 (v) imply that $F = F^\Gamma$ is always convex and lower semicontinuous.

Let now $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. It is 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 with $a \leq F$ but $a(x_0) > \lambda > F^\Gamma(x_0)$, which would contradict the definition of F^Γ. Since F is proper, convex, and lower semicontinuous, $\operatorname{epi} F$ is nonempty, convex, and closed by Lemma 3.1. Furthermore, $\{(x_0, \lambda)\}$ is compact and, as $\lambda < F(x_0)$, disjoint with $\operatorname{epi} F$. Theorem 1.5 (ii) hence yields a $z^* \in (X \times \mathbb{R})^*$ and an $\alpha \in \mathbb{R}$ with

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

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

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

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

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

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

(i) $s < 0$: We set

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

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

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

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

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

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

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

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

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

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

Lemma 3.6. *Let $f : \mathbb{R} \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. If $\Omega \subset \mathbb{R}^n$ is bounded and $1 \le p \le \infty$, this also holds for*

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

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

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

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

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

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

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

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

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

To show lower semicontinuity, we use Lemma 3.1. Let $\{(u_n, t_n)\}_{n\in\mathbb{N}} \subset \operatorname{epi} F$ with $u_n \to u$ in $L^p(\Omega)$ and $t_n \to t$ in $\mathbb{R}$. Then there exists a subsequence $\{u_{n_k}\}_{k\in\mathbb{N}}$ with $u_{n_k}(x) \to u(x)$ almost everywhere. Hence, the lower semicontinuity of f together with Fatou's Lemma implies that

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

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

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

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

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

(iii) U be bounded or F be coercive.

Then the problem

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

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

Proof. We consider the extended functional $\overline{F} = F + \delta_U : X \to \overline{\mathbb{R}}$. Assumption (i) together with Lemma 2.2 implies that δ_U is proper, convex, and weakly lower semicontinuous. From (ii) we obtain an $x_0 \in U$ with $\overline{F}(x_0) < \infty$, and hence $\overline{F}$ is proper, convex, and weakly lower semicontinuous. Finally, $\overline{F}$ 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 $\bar{x} \in \operatorname{dom} \overline{F} = U \cap \operatorname{dom} F$ of $\overline{F}$ with

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

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

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

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

while the strict convexity of F implies that

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

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

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

Lemma 3.8. *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $x_0 \in X$ be given. Then the functional*

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

is coercive.

Proof. Since F is proper, convex, and lower semicontinuous, it follows from Lemma 3.5 that F is bounded from below by 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

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

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

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

Lemma 3.9. *Let X be a normed vector space, $F : X \to \overline{\mathbb{R}}$ be convex, and $x \in X$. If there is a $\rho >$ 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 $L(x, \rho/2) := 4(M - F(x))/\rho$. □

It thus remains to show that convex functions are bounded from above. We start with the scalar case.

Lemma 3.10. *If $f : \mathbb{R} \to \overline{\mathbb{R}}$ is convex, then f is locally bounded from above on $(\operatorname{dom} f)^o$.*

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

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

Hence f is locally bounded from above in x. □

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

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

Lemma 3.11. *Let X be a Banach space. If $F : X \to \overline{\mathbb{R}}$ is convex and lower semicontinuous, then F is locally bounded from above on $(\operatorname{dom} F)^o$.*

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

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

It is straightforward to verify that f is convex and lower semicontinuous as well and satisfies $0 \in (\operatorname{dom} f)^o$. By Lemmas 3.9 and 3.10, f is thus locally Lipschitz continuous in 0; in particular, $|f(t) - f(0)| \leq Lt \leq 1$ for sufficiently small $t > 0$. The reverse triangle inequality therefore yields a $\delta > 0$ with

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

Hence 0 lies in the algebraic interior of the sublevel set 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 $z \in O_\rho(0)$.

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

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

Again, it is straightforward to verify convexity and lower semicontinuity of $\tilde{F}$ and that $0 \in (\operatorname{dom} \tilde{F})^o$. It follows from what we've shown before that $\tilde{F}$ is locally bounded from above on $O_\rho(0)$, which also implies that F is locally bounded from above on $O_\rho(x)$. □

Together with Lemma 3.9, we thus obtain the desired result.

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

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

(i) the function

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

is increasing;

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

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

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

Proof. (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 $x + sh = (1 - \frac{s}{t})x + \frac{s}{t}(x + th)$.

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

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

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

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

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

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

Unfortunately, this concept can't yet be what we are looking for, since the convex function $f : \mathbb{R} \to \mathbb{R}$, $f(t) = |t|$ 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 $F : X \to \overline{\mathbb{R}}$ and any $x \in \operatorname{dom} F$ the set

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

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

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

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

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

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

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

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

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

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

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

(Note that $\tilde{x} \notin \operatorname{dom} F$ is allowed since in this case the inequality is trivially satisfied.) For $x \notin \operatorname{dom} F$, we set $\partial F(x) = \emptyset$.[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 *subderivative*.[2]

Theorem 4.3 (Fermat principle). *Let $F : X \to \overline{\mathbb{R}}$ and $\bar{x} \in \operatorname{dom} F$. Then the following statements are equivalent:*

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

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

[1] We will later show that $\partial F(x)$ is nonempty and bounded for all $x \in (\operatorname{dom} F)^o$; see Corollary 8.14.

[2] Following the terminology for classical derivatives, we reserve the more common term *subgradient* for its Riesz representation $z_{x^*} \in X$ when X is a Hilbert space.

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

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

i.e., $F(\bar{x}) \leq F(\tilde{x})$ for all $\tilde{x} \in X$.[3] □

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; die condition $\xi = 0 \in \partial f(\tilde{x})$ thus means that f has a horizontal tangent in $\bar{x}$.

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 $F : X \to \overline{\mathbb{R}}$ 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 $x \in X$,*

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

[3] Note 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 $x \in \operatorname{dom} F$ implies that F is convex.) This leads to problems in particular for the proof of calculus rules, for which we will indeed have to assume convexity.

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

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

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

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 further 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. □

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

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

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

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

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.

Theorem 4.6. *Let $f : \mathbb{R} \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and let $F : L^p(\Omega) \to \overline{\mathbb{R}}$ with $1 \leq p < \infty$ be as in Lemma 3.6. Then we have for all $u \in \operatorname{dom} F$ with $q := \frac{p}{p-1}$ that*

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

Proof. Let $u, \tilde{u} \in \operatorname{dom} F$, i.e., $f \circ u, f \circ \tilde{u} \in L^1(\Omega)$ (otherwise there is nothing to show), and let $u^* \in L^q(\Omega)$ be arbitrary. If $u^* \in L^q(\Omega)$ 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

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

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

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

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

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

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 $F : X \to \overline{\mathbb{R}}$ be convex and $x \in \operatorname{dom} F$. 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* $x + x_0 \in \operatorname{dom} F$.

The sum rule is already considerably more delicate.

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

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

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

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

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

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

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

For that purpose, we define the sets

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

i.e.,

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

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

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

2. The critical point is of course the nonemptiness of C_1^o, for which we argue as follows. Since $x_0 \in (\operatorname{dom} F)^o$, we know from Lemma 3.11 that F is bounded in an open neighborhood $U \subset (\operatorname{dom} F)^o$ of x_0. We can thus find an open interval $I \subset \mathbb{R}$ such that $U \times I \subset C_1$. Since $U \times I$ is open by the definition of the product topology on $X \times \mathbb{R}$, any (x_0, α) with $\alpha \in I$ is an interior point of C_1.

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

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

contradicting (4.4). Hence C_1^o and C_2 are disjoint.

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

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

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

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

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

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 $t \to \infty$ 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

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

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

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 $A \in L(X, Y)$, $F : Y \to \overline{\mathbb{R}}$ be convex and lower semicontinuous, and $x \in \operatorname{dom}(F \circ A)$. 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 $Ax_0 \in (\operatorname{dom} F)^o$.

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

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

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

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

(This technique of getting rid of the operator composition by working in the graph space is called “lifting”.) Since A is linear, $\operatorname{graph} A$ and hence $\delta_{\operatorname{graph} A}$ are convex. Furthermore, $Ax \in \operatorname{dom} F$ by assumption and thus $(x, Ax) \in \operatorname{dom} H$.

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

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

In particular, this holds for all $\tilde{y} \in \operatorname{ran}(A) = \{A\tilde{x} : \tilde{x} \in X\}$. By $\delta_{\operatorname{graph} A}(\tilde{x}, A\tilde{x}) = 0$ 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 $\delta_{\operatorname{graph} A}(x, Ax) = 0$ and $\delta_{\operatorname{graph} A}(\tilde{x}, \tilde{y}) \geq 0$, it then follows for all $\tilde{x} \in X$ and $\tilde{y} \in Y$ that

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

where we have used that the 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 $\tilde{F} : X \times Y \to \overline{\mathbb{R}}$, $(x, y) \mapsto F(y)$, and $(x_0, Ax_0) \in \operatorname{graph} A = \operatorname{dom} \delta_{\operatorname{graph} A}$. Since $Ax_0 \in (\operatorname{dom} F)^o \subset Y$ by assumption, $(x_0, Ax_0) \in (\operatorname{dom} \tilde{F})^o = X \times (\operatorname{dom} F)^o \subset X \times Y$ as well. We can thus apply Theorem 4.8 to obtain

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

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

Finally, we "collapse" these subdifferentials back to the individual spaces to obtain the desired characterization. First, we have $(x^*, y^*) \in \partial \tilde{F}(x, Ax)$ 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 in turn $\tilde{x} = x$ and $\tilde{y} = Ax$ implies that $y^* \in \partial F(Ax)$ and $x^* = 0$, respectively. Second, $(w^*, z^*) \in \partial \delta_{\operatorname{graph} A}(x, Ax)$ if and only if

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

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 $w^* = -A^* z^* \in X^*$. Together we obtain

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

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

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 $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. If there exists an $x_0 \in U^o \cap \operatorname{dom} F$, then $\bar{x} \in U$ solves*

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

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

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

Proof. Due to the assumptions on F and U, we can apply Theorem 4.3 to $J := F + \delta_U$. Furthermore, since $x_0 \in U^o = (\operatorname{dom} \delta_U)^o$, 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). □

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

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 $F : X \to \overline{\mathbb{R}}$ be proper but not necessarily convex. We then define the *Fenchel conjugate* of F as

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

(Since $\operatorname{dom} F \neq \emptyset$ is excluded, we have that $F^*(x^*) > -\infty$ for all $x^* \in X^*$, and hence the definition is meaningful.) Lemma 3.4 (v) and Lemma 2.2 (v) immediately imply that F^* is always convex and lower semicontinuous (as long as F is indeed proper). If F is bounded from below by an affine functional (which is always the case if F is proper, convex, and lower semicontinuous by Lemma 3.5), then F^* is proper as well. Finally, the definition directly yields the *Fenchel–Young inequality*

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

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 $F : X^* \to \overline{\mathbb{R}}$ (i.e., if F is defined on some dual space) as

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

The point of this convention is that even in nonreflexive spaces, the *biconjugate* $F^{**} := (F^*)^*$ is again defined on X (rather than $X^{**} \supset X$). Intuitively, F^{**} is the convex hull of F, which by Lemma 3.5 coincides with F itself if F is convex.

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

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

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

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

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

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

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

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

We now show that we can replace F^{**} with F on the right-hand side. For this, let $a(x) = \langle x^*, x\rangle_X - \alpha$ with arbitrary $x^* \in X^*$ and $\alpha \in \mathbb{R}$. If $a \leq F^{**}$, then (i) implies that $a \leq F$. Conversely, if $a \leq F$, we have that $\langle x^*, x\rangle_X - F(x) \leq \alpha$ for all $x \in X$, and taking the supremum over all $x \in X$ yields that $\alpha \geq F^*(x^*)$. By definition of F^{**}, we thus obtain that

$$a(x) = \langle x^*, x\rangle_X - \alpha \leq \langle x^*, x\rangle_X - F^*(x^*) \leq F^{**}(x) \quad \text{for all } x \in X,$$

i.e., $a \leq F^{**}$.

Statement (iii) now directly follows from (ii) and Lemma 3.5. □

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

$$\sup_{x\in X} (x^*, x)_X - \tfrac{1}{2}(x, x)_X$$

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

$$F^* : X \to \mathbb{R}, \qquad F^*(x^*) = \tfrac{1}{2}\|x^*\|_X^2.$$

(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 $t \to \infty$ in

$$0 < t(\langle x^*, x_0\rangle_X - \|x_0\|_X) = \langle x^*, tx_0\rangle_X - \|tx_0\|_X \leq F^*(x^*)$$

yields $F^*(x^*) = \infty$.

Together we obtain that $F^* = \delta_{B_{X^*}}$. As above, a similar argument shows that $(\|\cdot\|_{X^*})^* = \delta_{B_X}$.

As for convex subdifferentials, Fenchel conjugates of integral functionals can be computed pointwise.

Theorem 5.3. *Let $f : \mathbb{R} \to \overline{\mathbb{R}}$ be measurable, proper and lower semicontinuous, and let $F : L^p(\Omega) \to \overline{\mathbb{R}}$ with $1 \leq p < \infty$ be defined as in Lemma 3.6. Then we have for $q = \frac{p}{p-1}$ that*

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

Proof. We argue similarly as in the proof of Theorem 4.6, with some changes that are needed since measurability of $f \circ u$ does not immediately imply that of $f^* \circ u^*$. Let $u^* \in L^q(\Omega)$ be arbitrary and consider for all $x \in \Omega$ the functions

$$\varphi(x) := \sup_{t\in\mathbb{R}} tu^*(x) - f(t) = f^*(u^*(x)),$$

as well as for $n \in \mathbb{N}$

$$\varphi_n(x) := \sup_{|t|\leq n} tu^*(x) - f(t) \leq f^*(u^*(x)).$$

By a measurable selection theorem ([Ekeland & Témam 1999, Theorem VIII.1.2]), the pointwise supremum in the definition of φ_n is attained at some t_x^* for almost every $x \in \Omega$ and defines a measurable mapping $x \mapsto u_n(x) := t_x^*$ with $\|u_n\|_{L^\infty} \leq n$. This also implies that $\varphi_n = u_n \cdot u^* - f \circ u_n$ is measurable. Furthermore, by assumption there exists a $t_0 \in \operatorname{dom} f$, and hence $u_0 := t_0u^*(x) - f(t_0)$ is measurable and satisfies $u_0 \leq \varphi_n(x)$ for all $n \geq |t_0|$. Finally, by construction, $\varphi_n(x)$ is monotonically increasing and converges to $\varphi(x)$ for all $x \in \Omega$. The sequence $\{\varphi_n - u_0\}_{n\in\mathbb{N}}$ of functions is thus measurable and nonnegative, and the monotone convergence theorem yields that

$$\int_\Omega \varphi(x) - u_0(x)\,dx = \int_\Omega \sup_{n\in\mathbb{N}} \varphi_n(x) - u_0(x)\,dx = \sup_{n\in\mathbb{N}} \int_\Omega \varphi_n(x) - u_0(x)\,dx.$$

Hence the pointwise limit $\varphi = f^* \circ u^*$ is measurable as well.

The measurable selection theorem also yields that

$$\begin{aligned}\int_\Omega f^*(u^*(x))\,dx &= \sup_{n\in\mathbb{N}} \int_\Omega \sup_{|t|\leq n} \{tu^*(x) - f(t)\}\,dx\\ &= \sup_{n\in\mathbb{N}} \int_\Omega u^*(x)u_n(x) - f(u_n(x))\,dx\\ &\leq \sup_{u\in L^p(\Omega)} \int_\Omega u^*(x)u(x) - f(u(x))\,dx = F^*(u^*),\end{aligned}$$

since $u_n \in L^\infty(\Omega) \subset L^p(\Omega)$ for all $n \in \mathbb{N}$.

For the converse inequality, we can now proceed as in the proof of Theorem 4.6. For any $u \in L^p(\Omega)$ and $u^* \in L^q(\Omega)$, we have by the Fenchel–Young inequality (5.1) applied to f and f^* that

$$f(u(x)) + f^*(u^*(x)) \geq u^*(x)u(x) \quad \text{for almost every } x \in \Omega.$$

Since both sides are measurable, this implies that

$$\int_\Omega f^*(u^*(x))\,dx \geq \int_\Omega u^*(x)u(x) - f(u(x))\,dx,$$

and taking the supremum over all $u \in L^p(\Omega)$ yields the claim. □

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

Lemma 5.4. *Let $F : X \to \overline{\mathbb{R}}$ be proper. Then,*

(i) $(\alpha F)^* = \alpha F^* \circ (\alpha^{-1}\mathrm{Id})$ *for any* $\alpha > 0$;

(ii) $(F(\cdot + x_0) + \langle x_0^*, \cdot\rangle_X)^* = F^*(\cdot - x_0^*) - \langle \cdot - x_0^*, x_0\rangle_X$ *for all* $x_0 \in X$, $x_0^* \in X^*$;

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

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

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

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

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

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

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

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 (Fenchel–Young). *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. Then the following statements are equivalent for any $x \in X$ and $x^* \in X^*$:*

(i) $\langle x^*, x\rangle_X = F(x) + F^*(x^*)$;

(ii) $x^* \in \partial F(x)$;

(iii) $x \in \partial F^*(x^*)$.

Proof. If (i) holds, the definition of F^* as a supremum immediately implies that

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

which again by definition is equivalent to (ii). Conversely, taking the supremum over all $\tilde{x} \in X$ in (5.2) yields

$$
\langle x^*, x\rangle_X \geq F(x) + F^*(x^*),
$$

which together with the Fenchel–Young inequality (5.1) leads to (i).

Similarly, (i) in combination with Theorem 5.1 implies that

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

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

Remark. If F is not convex, the above proof shows that we still have the equivalence (i) ⇔ (ii). Furthermore since always $F^{**} \leq F$ by Theorem 5.1 (i), it still holds that (ii) ⇒ (iii). However, we can only conclude from (iii) that (ii) holds for $F^{**} \neq F$ in place of F. Applying Lemma 5.5 to nonconvex functionals therefore inevitably introduces a *convexification* (by replacing the nonconvex F with its convex envelope F^{**}).

□

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 $F : X^* \to \overline{\mathbb{R}}$ and $F^* : X \to \overline{\mathbb{R}}$.)

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

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

for $F : X \to \overline{\mathbb{R}}$ and $G : Y \to \overline{\mathbb{R}}$ proper, convex, and lower semicontinuous, 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*

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

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

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

Theorem 5.6 (Fenchel–Rockafellar). *Let X and Y be normed vector spaces, $F : X \to \overline{\mathbb{R}}$ and $G : Y \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, 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 $x_0 \in \operatorname{dom} F \cap \operatorname{dom}(G \circ A)$ with $Ax_0 \in (\operatorname{dom} G)^o$.

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

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

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

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

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.,

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

Adding both equations now yields

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

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

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

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

$$\begin{aligned} F(\bar{x}) + G(A\bar{x}) &= \inf_{x\in X} \sup_{Y^*\in Y^*} F(x) + \langle y^*, Ax\rangle_Y - G^*(y^*) \\ &\geq \sup_{Y^*\in Y^*} \inf_{x\in X} F(x) + \langle y^*, Ax\rangle_Y - G^*(y^*) \\ &= \sup_{y^*\in Y^*} -F^*(-A^*y^*) - G^*(y^*). \end{aligned} \tag{5.9}$$

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). □

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 subdifferential 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 $F : X \to \overline{\mathbb{R}}$ 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 mechnical 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 subderivatives with the corresponding set in X of Riesz representations (*subgradients*).

6.1 MONOTONE OPERATORS

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

- its *domain of definition* $\operatorname{dom} A = \{x \in X : Ax \neq \emptyset\}$;
- its *range* $\operatorname{ran} A = \bigcup_{x\in X} Ax$;
- its *graph* $\operatorname{graph} A = \{(x, y) \in X \times Y : y \in Ax\}$;
- 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, the inverse always exists.) We say that set-valued mapping A is *surjective* if $\operatorname{ran} A = Y$. For $A, B : X \rightrightarrows Y$, $C : Y \rightrightarrows Z$, and $\lambda \in \mathbb{R}$ we further define

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

Let from now on X be a Hilbert space. A set-valued mapping $A : X \rightrightarrows X$ is called *monotone* if

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

Example 6.1. (i) It follows directly from the definition that the identity mapping $\mathrm{Id} : X \rightrightarrows X$, $x \mapsto \{x\}$, is monotone.

(ii) Similarly, if $A, B : X \rightrightarrows X$ are monotone and $\lambda \geq 0$, then λA and $A + B$ are monotone as well.

(iii) If $F : X \to \overline{\mathbb{R}}$ is proper, then $\partial F : X \rightrightarrows X$, $x \mapsto \partial F(x)$, is monotone: 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

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

Adding the first inequality for $\tilde{x} = x_2$ and the second for $\tilde{x} = x_1$ and rearranging the result yields (6.1).

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

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

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

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

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

Example 6.2. The operator

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

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

Several useful properties follow directly from the definition.

Lemma 6.3. *If $A : X \rightrightarrows X$ is maximally monotone, then so is λA for all $\lambda > 0$.*

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

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

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

Lemma 6.4. *Let $A : X \rightrightarrows X$ be maximally monotone. Then A is weakly–strongly closed, i.e., $x_n \rightharpoonup x$ and $Ax_n \ni x_n^* \to x^*$ 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

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

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 $x^* \in Ax$. □

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 $\operatorname{Id} + A$ is surjective. As a preparation, we first prove an important partial result.

Lemma 6.5. *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex and lower semicontinuous. Then* $\operatorname{Id} + \partial F$ *is surjective.*

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

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

which is proper, (strictly) convex and lower semicontinuous 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., $z \in \{\bar{x}\} + \partial F(\bar{x}) = (\operatorname{Id} + \partial F)(\bar{x})$. Hence $\operatorname{ran}(\operatorname{Id} + \partial F) = X$ as claimed. □

We now turn to the general case.

Theorem 6.6 (Minty). *Let $A : X \rightrightarrows X$ be monotone with* $\operatorname{graph} A \neq 0$. *Then A is maximally monotone if and only if* $\operatorname{Id} + A$ *is surjective.*

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

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

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

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

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 6.5; for the general case, we proceed similarly by constructing a functional F_A that plays the same role for A as F does for ∂F. Specifically, we define for a maximally monotone operator $A : X \rightrightarrows X$ with $\operatorname{graph} A \neq \emptyset$ the *Fitzpatrick functional*

$$F_A : X \times X \to [-\infty, \infty], \qquad (x, y) \mapsto \sup_{(z,w) \in \operatorname{graph} A} \left((z, y)_X + (x, w)_X - (z, w)_X\right), \tag{6.5}$$

which can be written equivalently as

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

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 $(z, w) \in \operatorname{graph} A$ if and only if $(x, y) \in \operatorname{graph} A$; in particular, $(x - z, y - w)_X < 0$ for all $(x, y) \notin \operatorname{graph} A$. Hence, (6.6) implies that $F_A(x, y) \geq (x, y)_X$ with equality for $(x, y) \in \operatorname{graph} A$ (since in this case the infimum is attained in $(z, w) = (x, y)$). Since $\operatorname{graph} A \neq \emptyset$, this shows that F_A is proper.

(ii) On the other hand, the definition (6.5) yields that

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

(since $(z, w) \in \operatorname{graph} A$ if and only if $(w, z) \in \operatorname{graph} A^{-1}$). Since we have required that $\operatorname{graph} A \neq \emptyset$, the Fitzpatrick functional F_A is the Fenchel conjugate of a proper functional and therefore convex and lower semicontinuous.

As a first step, we now show that $0 \in \operatorname{ran}(\mathrm{Id}+A)$. We set $Z := X\times X$ as well as $\xi := (x, y) \in Z$ and consider the functional

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

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

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

Furthermore, J_A is proper, (strictly) convex, lower semicontinuous, 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

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

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 $(x,y) \in \operatorname{graph} A$ that

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

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., $(\bar y, \bar x) \in \operatorname{graph} A$. 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., $0 = \bar y + \bar x \in (\mathrm{Id} + A)(\bar y)$.

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 with $\operatorname{graph} B \neq \emptyset$ as well. As we have just shown, there now exists a $\bar y \in X$ with $0 \in (\mathrm{Id} + B)(\bar y) = \{\bar y\} + \{-z\} + A\bar y$, i.e., $z \in (\mathrm{Id} + A)(\bar y)$. Hence $\mathrm{Id} + A$ is surjective. □

Together with Lemma 6.5 (which in particular implies $\operatorname{graph} \partial F \neq \emptyset$ for proper, convex, and lower semicontinuous F), this yields the maximal monotonicity of convex subdifferentials.

Corollary 6.7. *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. Then $\partial F : X \rightrightarrows X$ is maximally monotone.*

6.2 RESOLVENTS AND PROXIMAL POINTS

We know from Lemma 6.5 that $\mathrm{Id} + \partial F$ is surjective for any proper, convex, and lower semicontinuous functional F; the proof even shows that each preimage is unique. Hence $(\mathrm{Id} + \partial F)^{-1}$ 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 $A : X \rightrightarrows X$ with $\operatorname{graph} A \neq \emptyset$ the *resolvent*

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

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

$$\operatorname{prox}_F : X \to X, \qquad \operatorname{prox}_F(x) = \operatorname*{arg\,min}_{z\in X} \frac{1}{2}\|z - x\|_X^2 + F(z). \tag{6.9}$$

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

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

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

Lemma 6.8. *Let $A : X \rightrightarrows X$ be maximally monotone with* $\operatorname{graph} A \neq \emptyset$*. Then $\mathcal{R}_A : X \to X$.*

Proof. Since A is maximally monotone with $\operatorname{graph} A \neq \emptyset$, $\mathrm{Id}+A$ is surjective by Theorem 6.6, which implies that $\mathcal{R}_A(x) \neq \emptyset$ for all $x \in X$, i.e., $\operatorname{dom} \mathcal{R}_A = X$. Let now $x, z \in X$ 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

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

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

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 $T : X \to X$ is called *firmly nonexpansive* if it satisfies this inequality, i.e., if

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

Firm nonexpansivity implies another useful inequality.

Lemma 6.9. *Let $A : X \rightrightarrows X$ be maximally monotone with* $\operatorname{graph} A \neq \emptyset$*. Then $\mathcal{R}_A : X \to X$ is firmly nonexpansive and satisfies that*

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

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

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

Corollary 6.10. *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. Then* $\operatorname{prox}_F : X \to X$ *is Lipschitz continuous with constant $L = 1$.*

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

Lemma 6.11. *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $x, x^* \in X$. Then for any $\gamma > 0$,*

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

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

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

where in the last step we have used that $\gamma \partial F = \partial(\gamma F)$ by Lemma 4.7 (i) and hence that $\operatorname{prox}_{\gamma F} = \mathcal{R}_{\gamma \partial F}$. □

Corollary 6.12. *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex and lower semicontinuous, and $\gamma > 0$ be arbitrary. Then $\bar{x} \in \operatorname{dom} F$ is a minimizer of F if and only if*

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

Proof. Simply apply Lemma 6.11 to the Fermat principle $0 \in \partial F(\bar{x})$. □

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-type methods.

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

Theorem 6.13 (Moreau decomposition). *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. Then we have for all $x \in X$ that*

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

Proof. Setting $w = \operatorname{prox}_F(x)$, Lemmas 5.5 and 6.11 imply that

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

The following calculus rules will prove useful.

Lemma 6.14. *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. Then,*

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

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

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

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

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

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

Proof. *(i):* By definition,

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

Now note that since X is a vector space,

$$\min_{w \in X} \frac{1}{2}\|w - x\|_X^2 + F(\lambda w + z) = \min_{v \in X} \frac{1}{2}\|\lambda^{-1}(v - z) - x\|_X^2 + F(v),$$

and the respective minimizers $\bar{w}$ and $\bar{v}$ are related by $\bar{v} = \lambda \bar{w} + z$. The claim then follows from

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

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

(ii): Theorem 6.13, Lemma 5.4 (i), and (i) for $\lambda = \gamma^{-1}$ and $z = 0$ together imply that

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

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

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

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

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

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

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

Example 6.15. We first consider scalar functions $f : \mathbb{R} \to \overline{\mathbb{R}}$.

(i) $f(t) = \frac{1}{2}|t|^2$. Since f is differentiable, we can set the derivative of $\frac{1}{2}(s - t)^2 + \frac{\gamma}{2}s^2$ to zero and solve for s to obtain $\operatorname{prox}_{\gamma f}(t) = (1 + \gamma)^{-1}t$.

(ii) $f(t) = |t|$. By (4.3) we have that $\partial f(t) = \operatorname{sign}(t)$; hence $s := \operatorname{prox}_{\gamma f}(t) = (\mathrm{Id} + \gamma \operatorname{sign})^{-1}(t)$ if and only if $t \in \{s\} + \gamma \operatorname{sign}(s)$. Let t be given and assume this holds for some $\bar{s}$. We now proceed by case distinction.

Case 1: $\bar{s} > 0$. This implies that $t = \bar{s} + \gamma$, i.e., $\bar{s} = t - \gamma$, and hence that $t > \gamma$.

Case 2: $\bar{s} < 0$. This implies that $t = \bar{s} - \gamma$, i.e., $\bar{s} = t + \gamma$, and hence that $t < -\gamma$.

Case 3: $\bar{s} = 0$. This implies that $t \in \gamma[-1, 1] = [-\gamma, \gamma]$.

Since this yields a complete and disjoint case distinction for t, we can conclude that

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

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

(iii) $f(t) = \delta_{[-1,1]}(t)$. By Example 5.2 (iii) we have that $f^*(t) = |t|$. Hence Lemma 6.14 (ii) yields that

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

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

Example 6.16. We can generalize Example 6.15 to $X = \mathbb{R}^N$ (endowed with the Euclidean inner product) by applying Lemma 6.14 (iii) N times. We thus obtain componentwise

(i) for $F(x) = \frac{1}{2}\|x\|_2^2 = \sum_{i=1}^N \frac{1}{2} x_i^2$ that

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

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

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

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

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

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

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

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

Corollary 6.17. *Let $f : \mathbb{R} \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $F : L^2(\Omega) \to \overline{\mathbb{R}}$ be defined as in Lemma 3.6. Then we have for all $\gamma > 0$ and $u \in L^2(\Omega)$ that*

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

Example 6.18. Let X be a Hilbert space. Similarly to Example 6.15 one can show

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

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

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

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

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

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

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 introduced in Chapter 9.

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

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

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

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

which by Corollary 6.10 is always Lipschitz continuous with constant $L = \gamma^{-1}$.

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

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

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

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

(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 6.19. *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $\gamma > 0$. Then F_γ is convex.*

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

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

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

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

Let now $x_1, x_2 \in X$ and $\lambda \in [0, 1]$. Since $F_\gamma(x) = \inf_{z\in X} H(x, z)$ for $G(y) := \frac{1}{2\gamma}\|y\|_X^2$, there exist two minimizing sequences $\{z_n^1\}_{n\in\mathbb{N}}, \{z_n^2\}_{n\in\mathbb{N}} \subset X$ with

$$H(x_1, z_n^1) \to F_\gamma(x_1), \qquad H(x_2, z_n^2) \to F_\gamma(x_2).$$

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

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

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

[1] not to be confused with the *convex* envelope F^Γ!

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

Theorem 6.20. *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $\gamma > 0$. Then F_γ is Fréchet differentiable with*

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

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

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

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

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

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

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

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

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

Together, these two inequalities yield that

$$\begin{aligned} 0 &\le F_\gamma(y) - F_\gamma(x) - \frac{1}{\gamma}(y-x, x-x^*)_X \\ &\le \frac{1}{\gamma}(y-x, (y-y^*) - (x-x^*))_X \\ &\le \frac{1}{\gamma}\left(\|y-x\|_X^2 - \|y^*-x^*\|_X^2\right) \\ &\le \frac{1}{\gamma}\|y-x\|_X^2, \end{aligned}$$

where the next-to-last inequality follows from the firm nonexpansivity of proximal point mappings (Lemma 6.9).

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

$$0 \le \frac{F_\gamma(x+h) - F_\gamma(x) - \left(\gamma^{-1}(x-x^*), h\right)_X}{\|h\|_X} \le \frac{1}{\gamma}\|h\|_X \to 0 \qquad \text{for } h \to 0,$$

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

Since F_γ is convex by Lemma 6.19, this result together with Theorem 4.4 yields the catchy relation $\partial(F_\gamma) = (\partial F)_\gamma$.

Example 6.21. We consider again $X = \mathbb{R}^N$.

(i) For $F(x) = \|x\|_1$, we have from Example 6.16 (ii) that the proximal point mapping is given by the component-wise soft-shrinkage operator. Inserting this into the definition yields that

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

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

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

For small values, the absolute value is thus replaced by a quadratic function (which removes the nondifferentiability at 0). This modification is well-known under the name *Huber norm*.

(ii) For $F(x) = \delta_{B_\infty}(x)$, we have from Example 6.16 (iii) that the proximal mapping is given by the component-wise projection onto $[-1, 1]$ and hence that

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

Similarly, inserting this and using that $\operatorname{prox}_{\gamma F}(x) \in B_\infty$ 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 6.22. *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. Then we have for all $\gamma > 0$ that*

$$(F_\gamma)^* = F^* + \frac{\gamma}{2}\|\cdot\|_X^2.$$

Proof. We obtain directly from the definition of the Fenchel conjugate in Hilbert spaces and of the Moreau envelope that

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

since for any given $z \in X$, the inner supremum is always taken over the full space X. The claim now follows from Example 5.2 (i) and Lemma 5.4 (i). □

We briefly sketch the relevance for nonsmooth optimization. For a convex functional $F : X \to \overline{\mathbb{R}}$, every minimizer $\bar{x} \in X$ satisfies the Fermat principle $0 \in \partial F(\bar{x})$, which we can write equivalently as $\bar{x} \in \partial F^*(0)$. If we now replace ∂F^* with its Yosida approximation $(\partial F^*)_\gamma$, we obtain the regularized optimality condition

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

This is now an *explicit* and even Lipschitz continuous relation (which, among other things, can be used to derive stability properties for x_γ under perturbations). Although x_γ is no longer a minimizer of F, the convexity of F_γ implies that $x_\gamma \in (\partial F^*)_\gamma(0) = \partial(F_\gamma^*)(0)$ is equivalent to

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

i.e., x_γ is the (unique due to the strict convexity of the squared norm) minimizer of the functional $F + \frac{\gamma}{2}\|\cdot\|_X^2$. Hence, the regularization of ∂F^* has not made the original problem smooth but merely (more) strongly convex. The equivalence can also be used to show (similarly to the proof of Theorem 2.1) that $x_\gamma \rightharpoonup \bar{x}$ for $\gamma \to 0$. In practice, this straightforward approach fails due to the difficulty of computing F^* and $\operatorname{prox}_{F^*}$ and is therefore usually combined with one of the splitting techniques introduced in the next chapter.

7 PROXIMAL POINT AND SPLITTING METHODS

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

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

for $F, G : X \to \overline{\mathbb{R}}$ convex but not necessarily differentiable. One of the main difficulties compared to the differentiable setting is that the naive equivalent to steepest descent, the iteration

$$x^{k+1} \in x^k - \tau_k \partial J(x^k),$$

does not work since even in finite dimensions, arbitrary subgradients need not be descent directions – this can only be guaranteed for the subgradient of minimal norm; see, e.g., [Ruszczyński 2006, Example 7.1, Lemma 2.77]. Furthermore, the minimal norm subgradient of J cannot be computed easily from those of F and G. We thus follow a different approach and look for a root of the set-valued mapping $x \mapsto \partial J(x) \subset X^* \cong X$.

7.1 PROXIMAL POINT METHOD

We have seen in Corollary 6.12 that a root $\bar{x}$ of $\partial J : X \rightrightarrows X$ can be characterized as a fixed point of $\operatorname{prox}_{\gamma J}$ for any $\gamma > 0$. This suggests a fixed-point iteration: Choose $x^0 \in X$ and set for an appropriate sequence $\{\gamma_k\}_{k\in\mathbb{N}}$

$$x^{k+1} = \operatorname{prox}_{\gamma_k J}(x^k). \tag{7.1}$$

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 6.10 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 $x^k \rightharpoonup \bar{x}$ with $0 \in A\bar{x}$.

Proof. The iteration $x^{k+1} = \mathcal{R}_{\gamma_k A} x^k = (\mathrm{Id} + \gamma_k A)^{-1} x^k$ 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 6.12, this inclusion is equivalent to $x^* = \mathcal{R}_{\gamma A} x^*$ for all $\gamma > 0$. From Lemma 6.9 together with $(\mathrm{Id} - \mathcal{R}_{\gamma_k A})x^k = x^k - x^{k+1} = \gamma_k w^k$, we now obtain that

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

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 $\sum_{k=0}^{\infty} \gamma_k^2 \|w^k\|_X^2$ is finite. Since the sequence $\{\gamma_k^2\}_{k\in\mathbb{N}}$ is not summable by assumption, this requires that $\liminf_{k\to\infty} \|w^k\|_X^2 = 0$. This together with the convergence of $\{\|w^k\|_X\}_{k\in\mathbb{N}}$ implies that $w^k \to 0$. In particular, we have that $Ax^{k_l+1} \ni w^{k_l} \to 0$ for $x^{k_l+1} \rightharpoonup \bar{x}$, and the closedness of maximally monotone operators (Lemma 6.4) 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 $\{x^k\}_{k\in\mathbb{N}}$.[1] 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

[1] The following argument in a more general setting is known as *Opial's Lemma.*

both $\{\|x^k - \bar{x}\|_X\}_{k\in\mathbb{N}}$ and $\{\|x^k - \hat{x}\|_X\}_{k\in\mathbb{N}}$ are decreasing and bounded from below and therefore convergent. This implies that

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

Since $\bar{x}$ is a weak accumulation point, there exists a subsequence $\{x^{k_n}\}_{n\in\mathbb{N}}$ with $x^{k_n} \rightharpoonup \bar{x}$; similarly, there exists a subsequence $\{x^{k_m}\}_{m\in\mathbb{N}}$ with $x^{k_m} \rightharpoonup \hat{x}$. Hence,

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

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 $\{x^k\}_{k\in\mathbb{N}}$. □

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

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

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 apply Lemma 6.11 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 $\partial F(x) = \{\nabla F(x)\}$ is a singleton. This leads to the *proximal gradient method* or *forward-backward splitting method*

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

(The special case $G = \delta_C$ – i.e., $\operatorname{prox}_{\gamma G}(x) = \operatorname{proj}_C(x)$ – 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 $F : X \to \mathbb{R}$ be Gâteaux differentiable with Lipschitz continuous gradient. Then,*

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

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

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

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

Theorem 7.3. *Let $F : X \to \mathbb{R}$ and $G : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. 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

$$T_\gamma : X \to X, \qquad x \mapsto \gamma^{-1}(x - \operatorname{prox}_{\gamma G}(x - \gamma \nabla F(x))),$$

which allows reformulating the iteration (7.4) as

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

Applying Lemma 6.11 to the second equality then implies that

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

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

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

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

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

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

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

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

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

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

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

and hence $J(x^n) \to J(x^*)$ for $n \to \infty$. The lower semicontinuity of F and G now yields that

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

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. □

In particular, we obtain from (7.9) that $J(x^k) = J(x^*) + O(k^{-1})$. Ensuring $J(x^k) \le J(x^*) + \varepsilon$ thus requires $O(\varepsilon^{-1})$ iterations. By introducing a clever extrapolation, this can be reduced to $O(\varepsilon^{-1/2})$ 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

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

for the (hardly intuitive) choice[2]

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

see [Beck & Teboulle 2009, § 4].

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 $\operatorname{prox}_{\gamma_k G}$ (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 remains 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

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

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 implies 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 tautology

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

Replacing two (suitable) occurences of $\bar x$ by a new $\bar y$ in these four equations and then applying a fixed-point iteration yields the *Douglas–Rachford method*

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

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

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, inverse problems, and optimal control.

7.4 PRIMAL-DUAL SPLITTING

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

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

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

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

which can be reformulated using Lemma 6.11 as

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

for arbitrary $\sigma, \tau > 0$. This suggests the fixed-point iteration

$$\tag{7.12} \begin{cases} x^{k+1} = \operatorname{prox}_{\tau F}(x^k - \tau A^* y^k), \\ y^{k+1} = \operatorname{prox}_{\sigma G^*}(y^k + \sigma A x^{k+1}), \end{cases}$$

(where we have left the step sizes constant for simplicity). We now try to show convergence by interpreting it as a proximal point method. To that end, we rewrite (7.12) in fully explicit form to have (x^{k+1}, y^{k+1}) and (x^k, y^k) on different sides. For the first equation, we use $\operatorname{prox}_{\tau F} = (\mathrm{Id} + \tau\partial F)^{-1}$ to obtain that

$$\begin{aligned} x^{k+1} = \operatorname{prox}_{\tau F}(x^k - \tau A^* y^k) &\Leftrightarrow x^k - \tau A^* y^k \in \{x^{k+1}\} + \tau\partial F(x^{k+1}) \\ &\Leftrightarrow \tau^{-1}x^k - A^* y^k \in \{\tau^{-1}x^{k+1}\} + \partial F(x^{k+1}). \end{aligned}$$

Similarly, for the second equation we have

$$y^{k+1} = \operatorname{prox}_{\sigma G^*}(y^k + \sigma A x^{k+1}) \Leftrightarrow \sigma^{-1} y^k \in \{\sigma^{-1} y^{k+1} - A x^{k+1}\} + \partial G^*(y^{k+1}).$$

Setting $Z = X \times Y$, $z = (x, y)$, as well as

$$M = \begin{pmatrix} \tau^{-1}\mathrm{Id} & -A^* \\ 0 & \sigma^{-1}\mathrm{Id} \end{pmatrix}, \qquad T = \begin{pmatrix} \partial F & A^* \\ -A & \partial G^* \end{pmatrix},$$

we see that (7.12) is equivalent to

$$Mz^k \in (M+T)z^{k+1} \qquad \Leftrightarrow \qquad z^{k+1} \in (M+T)^{-1}Mz^k.$$

If M were invertible, we could use that $M = (M^{-1})^{-1}$ to obtain that $(M+T)^{-1}Mz^k = (\mathrm{Id} + M^{-1}T)^{-1}z^k$; the iteration would indeed amount to a proximal point method for the operator $M^{-1}T$ (which hopefully is maximally monotone).

Unfortunately, we cannot show the desired invertibility in this form. We therefore replace M by a self-adjoint operator for which we can show positive definiteness; i.e., we consider

$$M = \begin{pmatrix} \tau^{-1}\mathrm{Id} & -A^* \\ -A & \sigma^{-1}\mathrm{Id} \end{pmatrix},$$

so that the second step in the iteration becomes

$$\sigma^{-1}y^k - Ax^k \in \{\sigma^{-1}y^{k+1} - 2Ax^{k+1}\} + \partial G^*(y^{k+1}) \Leftrightarrow y^{k+1} = \mathrm{prox}_{\sigma G^*}(y^k + \sigma A(2x^{k+1} - x^k)).$$

This yields the *primal-dual extragradient method*[3]

$$\text{(7.13)} \qquad \begin{cases} x^{k+1} = \mathrm{prox}_{\tau F}(x^k - \tau A^* y^k), \\ \bar x^{k+1} = 2x^{k+1} - x^k, \\ y^{k+1} = \mathrm{prox}_{\sigma G^*}(y^k + \sigma A\bar x^{k+1}). \end{cases}$$

We now show that – under suitable conditions on σ and τ – the operator M is self-adjoint and positive definite with respect to the inner product

$$(z_1, z_2)_Z = (x_1, x_2)_X + (y_1, y_2)_Y \qquad \text{for all } z_1 = (x_1, y_1) \in Z, z_2 = (x_2, y_2) \in Z.$$

Lemma 7.4. *The operator $M : Z \to Z$ is bounded and self-adjoint. If $\sigma\tau\|A\|^2_{L(X,Y)} < 1$, then M is uniformly 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,

$$\begin{aligned} (Mz, z)_Z &= \left(\tau^{-1}x - A^*y, x\right)_X + \left(\sigma^{-1}y - Ax, y\right)_Y \\ &= \tau^{-1}\|x\|_X^2 - 2\,(x, A^*y)_X + \sigma^{-1}\|y\|_Y^2 \\ &\geq \tau^{-1}\|x\|_X^2 - 2\|A\|_{L(X,Y)}\|x\|_X\|y\|_Y + \sigma^{-1}\|y\|_Y^2 \\ &\geq \tau^{-1}\|x\|_X^2 - \|A\|_{L(X,Y)}\sqrt{\sigma\tau}(\tau^{-1}\|x\|_X^2 + \sigma^{-1}\|y\|_Y^2) + \sigma^{-1}\|y\|_Y^2 \\ &= (1 - \|A\|_{L(X,Y)}\sqrt{\sigma\tau})(\sqrt{\tau}^{-1}\|x\|_X^2 + \sqrt{\sigma}^{-1}\|y\|_Y^2) \\ &\geq C(\|x\|_X^2 + \|y\|_Y^2) \end{aligned}$$

[3] This 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].

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. □

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

$$c_1\|z\|_Z \leq \|z\|_M \leq c_2\|z\|_Z \qquad \text{for all } z \in Z, \tag{7.14}$$

where $c_1 = \sqrt{C} > 0$ from the proof of Lemma 7.4 and $c_2 := \|M\|_{L(Z,Z)} > 0$. From this, we can deduce continuous invertibility of M by a standard functional-analytic argument (a special case of the *Lax–Milgram Theorem*).

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. Let $z \in Z$ be given. Then (7.14) implies that the mapping $v \mapsto (z, v)_Z$ is a bounded (with respect to $\|\cdot\|_M$) linear functional. The Fréchet–Riesz Theorem 1.12 applied to the Hilbert space $(Z, (\cdot,\cdot)_M)$ thus yields a unique preimage $z^* \in Z$ with

$$(Mz^*, v)_Z = (z^*, v)_M = (z, v)_Z \qquad \text{for all } v \in Z.$$

Furthermore, the Riesz 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}. □

Hence $M^{-1}T$ is well-defined, i.e., $\operatorname{graph} M^{-1}T \neq \emptyset$. We now show maximal monotonicity with respect to the inner product $(\cdot,\cdot)_M$.

Lemma 7.6. *If $\sigma\tau\|A\|_{L(X,Y)}^2 < 1$, then $M^{-1}T$ is maximally monotone on $(Z, (\cdot,\cdot)_M)$.*

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 $\bar\eta \in \partial G^*(\bar y)$. Hence

$$\begin{aligned}(\bar z^* - z^*, \bar z - z)_M = (M\bar z^* - Mz^*, \bar z - z)_Z &= \big((\bar\xi + A^*\bar y) - (\xi + A^*y), \bar x - x\big)_X \\ &\quad + ((\bar\eta - A\bar x) - (\eta - Ax), \bar y - y)_Y \\ &= \big(\bar\xi - \xi, \bar x - x\big)_X + (A^*(\bar y - y), \bar x - x)_X \\ &\quad - (A(\bar x - x), \bar y - y)_Y + (\bar\eta - \eta, \bar y - y)_Y \\ &= \big(\bar\xi - \xi, \bar x - x\big)_X + (\bar\eta - \eta, \bar y - y)_Y \geq 0\end{aligned}$$

by the monotonicity of subdifferentials.

To show maximal monotonicity, let $\bar z^*, \bar z \in Z$ with

$$(Mz̄^* - Mz^*, \bar z - z)_Z = (\bar z^* - z^*, \bar z - z)_M \geq 0 \quad \text{for all } (z, z^*) \in \operatorname{graph} M^{-1}T, \tag{7.15}$$

i.e., for all $z \in Z$ and $Mz^* \in Tz$. As above, we can write $Mz^* = (\xi + A^*y, \eta - Ax)$ for some $\xi \in \partial F(x)$ and $\eta \in \partial G^*(y)$. 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 (7.15) implies for all $(x, y) \in Z$ that

$$\begin{aligned}0 &\leq \left((\bar\xi + A^*\bar y) - (\xi + A^*y), \bar x - x\right)_X + ((\bar\eta - A\bar x) - (\eta - Ax), \bar y - y)_Y\\ &= \left(\bar\xi - \xi, \bar x - x\right)_X + (\bar\eta - \eta, \bar y - y)_Y .\end{aligned}$$

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. □

In sum, we have shown that the primal-dual extragradient method (7.13) 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 together with the invertibility of M.

Theorem 7.7. *Let $F : X \to \overline{\mathbb{R}}$, $G : Y \to \overline{\mathbb{R}}$, and $A \in L(X, Y)$ satisfy the assumptions of Theorem 5.6. If $\sigma\tau\|A\|^2_{L(X,Y)} < 1$, then the sequence $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ generated by (7.13) converges weakly to some $(\bar x, \bar y) \in X \times Y$ satisfying (7.11).*

Proof. First, Theorem 5.6 yields the existence of a $\bar z := (\bar x, \bar y)$ satisfying the Fenchel extremality relations (7.11). By definition of T, this is equivalent to $0 \in T\bar z$, which by the invertibility from M due to Corollary 7.5 holds if and only if $0 \in M^{-1}T\bar z$. Hence there exists a root of $M^{-1}T$. By Lemma 7.6, $M^{-1}T$ is maximally monotone (with respect to $(\cdot, \cdot)_M$) and hence we can apply Theorem 7.1 to obtain that

$$\left(z^k, Mw\right)_Z = (z_k, w)_M \to (\bar z, w)_M = (\bar z, Mw)_Z \quad \text{for all } w \in Z,$$

where $\bar z \in Z$ satisfies $0 \in M^{-1}T\bar z$ and hence $0 \in T\bar z$. Since M is invertible and hence in particular surjective, this implies that $(z_k, \tilde w)_Z \to (\bar z, \tilde w)_Z$ for all $\tilde w := Mw \in Z$, which is the claimed weak convergence. □

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.[4]

[4] Using 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).

Finally, we remark that by setting $A = \mathrm{Id}$, $\tau = \gamma$, $\sigma = \gamma^{-1}$ and $z^k = x^k - \gamma y^k$ in (7.13) and applying Lemma 6.14 (ii), we recover the Douglas–Rachford method (7.10); however, since in this case $\sigma\tau\|A\|^2_{L(X,Y)} = 1$, we cannot obtain its convergence from Theorem 7.7.

7.5 STRONG CONVERGENCE AND RATES

The central idea of the convergence proofs we have seen so far is to use the iteration step together with the monotonicity of the set-valued mapping to obtain the boundedness of the error $\|x^k - \bar x\|_X$ and thus the weak convergence (at first of a subsequence) of the iterates. For *strong* convergence, however, we require additional properties that yield a more direct relation between the iteration step and the error, which will also allow deriving *convergence rates*.

One possibility is the following: A set-valued mapping $H : X \rightrightarrows X$ is called *strongly monotone* if there exists a $\gamma > 0$ such that

$$(7.16)\qquad \left(x_1^* - x_2^*, x_1 - x_2\right)_X \geq \gamma\|x_1 - x_2\|_X^2 \quad \text{for all } (x_1, x_1^*), (x_2, x_2^*) \in \operatorname{graph} H.$$

For example, $H = \partial F$ for $F(x) = \frac{1}{2}\|x\|_X^2$ is clearly strongly monotone with $\gamma = 1$; more generally, ∂F is strongly monotone if $F - \frac{\gamma}{2}\|\cdot\|_X^2$ is convex.

To illustrate the general approach, we show strong convergence for the proximal point method.

Theorem 7.8. *Let $H : X \rightrightarrows X$ be strongly monotone with $\gamma > 0$ and let $\bar x \in X$ satisfy $0 \in H(\bar x)$. Furthermore, let $\{x^k\}_{k\in\mathbb{N}}$ be generated via*

$$x^{k+1} = \mathcal{R}_{\tau_k H}(x^k)$$

for some $x^0 \in X$ and $\{\tau_k\}_{k\in\mathbb{N}} \subset (0,\infty)$.

(i) *If $\tau_k \equiv \tau$ is constant, then $x^k \to \bar x$ linearly, i.e., $\lim_{k\to\infty} \frac{\|x^{k+1}-\bar x\|_X}{\|x^k - \bar x\|_X} = \mu < 1$.*

(ii) *If $\tau_k \to \infty$, then $x^k \to \bar x$ superlinearly, i.e., $\lim_{k\to\infty} \frac{\|x^{k+1}-\bar x\|_X}{\|x^k - \bar x\|_X} = 0$.*

Proof. By definition of the resolvent, the iteration step is equivalent to

$$-\frac{1}{\tau_k}(x^{k+1} - x^k) \in H(x^{k+1}).$$

Together with $0 \in H(\bar x)$, it thus follows from (7.16) that

$$-\left(x^{k+1} - x^k, x^{k+1} - \bar x\right)_X \geq \tau_k\gamma\|x^{k+1} - \bar x\|_X^2.$$

We now apply to the left-hand side the (easily verified) *three-point identity*

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

for $x = x^{k+1}$, $y = x^k$, and $z = \bar{x}$. After rearranging, we obtain

$$\frac{1+2\tau_k\gamma}{2}\|x^{k+1}-\bar{x}\|_X^2 + \frac{1}{2}\|x^{k+1}-x^k\|_X^2 \le \frac{1}{2}\|x^k-\bar{x}\|_X^2.$$

In particular, it follows that

$$\frac{\|x^{k+1}-\bar{x}\|_X^2}{\|x^k-\bar{x}\|_X^2} \le \frac{1}{1+2\tau_k\gamma}.$$

We now make the case distinction:

(i) if $\tau_k \equiv \tau$, then $\mu := (1+2\tau\gamma)^{-1/2} < 1$ and hence $x^k \to \bar{x}$ linearly;

(ii) if $\tau_k \to \infty$, then $(1+2\tau_k\gamma)^{-1/2} \to 0$ and hence $x^k \to \bar{x}$ superlinearly. □

Similarly, one can show with a bit more effort that explicit splitting for strongly convex G converges linearly (but not superlinearly, since $\tau_k \le L^{-1}$ has to remain bounded). For the primal-dual extragradient method, this is possible as well (with significantly more effort) if the definition of strong monotonicity and the three-point identity are adapted to norms and inner products weighted with a “testing operator” that depends on the step sizes and the desired convergence rate; see [Valkonen 2020].

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 $F : X \to \mathbb{R}$ 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

$$|F(x_1) - F(x_2)| \le L\|x_1 - x_2\|_X \qquad \text{for all } x_1, x_2 \in O_\delta(x).$$

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]

8.1 DEFINITION AND BASIC PROPERTIES

We proceed as for the convex subdifferential and first define for $F : X \to \mathbb{R}$ the *generalized directional derivative* in $x \in X$ in direction $h \in X$ as

$$F^\circ(x; h) := \limsup_{\substack{y \to x \\ t \to 0^+}} \frac{F(y + th) - F(y)}{t}.$$

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 $F : X \to \mathbb{R}$ 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 $|F^\circ(x; h)| \le L\|h\|_X < \infty$;

(ii) subadditive, *i.e., $F^\circ(x; h + g) \le F^\circ(x; h) + F^\circ(x; g)$ for all $h, g \in X$;*

[1] For $F : X \to \overline{\mathbb{R}}$, 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].

(iii) positively homogeneous, *i.e., $F^\circ(x;\alpha h) = (\alpha F)^\circ(x;h)$ for all $\alpha \geq 0$ and $h \in X$;*

(iv) reflective, *i.e., $F^\circ(x;-h) = (-F)^\circ(x;h)$ for all $h \in X$.*

Proof. *(i):* Let $h, g \in X$ be arbitrary. The local Lipschitz continuity of F implies that

$$F(y+th) - F(y) \leq F(y+tg) - F(y) + tL\|h-g\|_X$$

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.

(ii): The definition of the lim sup and the productive zero immediately yield

$$\begin{aligned} F^\circ(x;h+g) &= \limsup_{\substack{y\to x\\ t\to 0^+}} \frac{F(y+th+tg) - F(y)}{t} \\ &\leq \limsup_{\substack{y\to x\\ t\to 0^+}} \frac{F(y+th+tg) - F(y+tg)}{t} + \limsup_{\substack{y\to x\\ t\to 0^+}} \frac{F(y+tg) - F(y)}{t} \\ &= F^\circ(x;h) + F^\circ(x;g), \end{aligned}$$

since $y \to x$ and $t \to 0$ implies that $y + tg \to x$ as well.

(iii): The claim is clear for $\alpha = 0$. For $\alpha > 0$, we obain again from the definition that

$$\begin{aligned} F^\circ(x;\alpha h) &= \limsup_{\substack{y\to x\\ t\to 0^+}} \frac{F(y - t(\alpha h)) - F(y)}{t} \\ &= \limsup_{\substack{y\to x\\ \alpha t\to 0^+}} \alpha\frac{F(y + (\alpha t)h) - F(y)}{\alpha t} = (\alpha F)^\circ(x;h). \end{aligned}$$

(iv): Similarly, we have that

$$\begin{aligned} F^\circ(x;-h) &= \limsup_{\substack{y\to x\\ t\to 0^+}} \frac{F(y - th) - F(y)}{t} \\ &= \limsup_{\substack{w\to x\\ t\to 0^+}} \frac{-F(w + th) - (-F(w))}{t} = (-F)^\circ(x;h), \end{aligned}$$

since $y \to x$ and $t \to 0$ implies that $w := y - th \to x$ as well. □

In particular, Lemma 8.1 (i–iii) implies that the mapping $h \mapsto F^\circ(x;h)$ is proper, convex, and lower semicontinuous.

We now define for a locally Lipschitz continuous functional $F : X \to \mathbb{R}$ the *Clarke subdifferential* in $x \in X$ as

$$\partial_C F(x) := \{x^* \in X^* : \langle x^*, h\rangle_X \leq F^\circ(x;h) \quad \text{for all } h \in X\}. \tag{8.1}$$

The definition together with Lemma 8.1 (i) directly implies the following properties.

Corollary 8.2. *Let $F : X \to \mathbb{R}$ be locally Lipschitz continuous and $x \in X$. Then $\partial_C F(x)$ is convex, weakly-∗ closed, and bounded. Specifically, if F is Lipschitz near x with constant L, then $\partial_C F(x) \subset K_L(0)$.*

Furthermore, we have the following useful closedness property.

Lemma 8.3. *Let $F : X \to \mathbb{R}$ be locally Lipschitz continuous in $x \in X$. If $\{x_n\}_{n\in\mathbb{N}} \subset X$ is a sequence with $x_n \to x$ 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

$$\langle x^*, h\rangle_X = \lim_{n\to\infty} \langle x_n^*, h\rangle_X \leq \limsup_{n\to\infty} F^\circ(x_n;h).$$

Hence we are finished if we can show that $\limsup_{n\to\infty} F^\circ(x_n;h) \leq F^\circ(x;h)$ (since then $x^* \in \partial_C F(x)$ by definition).

For this, we use that by definition of $F^\circ(x_n;h)$, there exist sequences $\{y_{n,m}\}_{m\in\mathbb{N}}$ and $\{t_{n,m}\}_{m\in\mathbb{N}}$ with $y_{n,m} \to x_n$ and $t_{n,m} \to 0$ for $m \to \infty$ realizing the lim sup for each x_n. Hence, for all $n \in \mathbb{N}$ we can find a $y_n := y_{n,m(n)}$ and a $t_n := t_{n,m(n)}$ such that $\|y_n - x_n\|_X + t_n < n^{-1}$ (and hence in particular $y_n \to x$ and $t_n \to 0$) as well as

$$F^\circ(x_n;h) - \tfrac{1}{n} \leq \frac{F(y_n + t_n h) - F(y_n)}{t_n}$$

for n sufficiently large. Taking the lim sup for $n \to \infty$ on both sides yields the desired inequality. □

Again, the construction immediately yields a Fermat principle.[2]

[2] Similarly 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.

Theorem 8.4. *If $F : X \to \mathbb{R}$ has a local minimum in $\bar{x}$, then $0 \in \partial_C F(\bar{x})$.*

Proof. If $\bar{x} \in X$ is a local minimizer of F, then $F(\bar{x}) \leq F(\bar{x} + th)$ for all $h \in X$ and $t > 0$ sufficiently small (since the topological interior is always included in the algebraic interior). But this implies that

$$\langle 0, h\rangle_X = 0 \leq \liminf_{t\to 0^+} \frac{F(\bar{x}+th) - F(\bar{x})}{t} \leq \limsup_{t\to 0^+} \frac{F(\bar{x}+th) - F(\bar{x})}{t} \leq F^\circ(x;h)$$

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., $f(t) = -|t|$).

Next, we show that the Clarke subdifferential is indeed a generalization of the derivative concepts we've studied so far.

Theorem 8.5. *Let $F : X \to \mathbb{R}$ be continuously Fréchet differentiable in a neighborhood U of $x \in X$. Then, $\partial_C F(x) = \{F'(x)\}$.*

Proof. First we note that the assumption implies local Lipschitz continuity of F: Since $F' : X \to X^*$ 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 $x_n \to x$ and $t_n \to 0^+$ 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

$$\begin{aligned} F^\circ(x;h) &= \lim_{n\to\infty} \frac{F(x_n + t_n h) - F(x_n)}{t_n} \\ &= \lim_{n\to\infty} \int_0^1 \frac{1}{t_n} \langle F'(x_n + t(t_n h)), t_n h\rangle_X \, dt \\ &= \langle F'(x), h\rangle_X \end{aligned}$$

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 $x^* = F'(x)$. □

However, if F is merely Fréchet differentiable, we only have that $F'(x) \in \partial_C F(x)$.

Theorem 8.6. *Let $F : X \to \mathbb{R}$ be convex and lower semicontinuous. Then, $\partial_C F(x) = \partial F(x)$ for all $x \in X$.*

Proof. Since F is finite-valued, $(\operatorname{dom} F)^o = X$, and hence F is local Lipschitz continuous in every $x \in X$ by Theorem 3.12. 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) yields the claim. First, we always have that

$$F'(x;h) = \lim_{t \to 0^+} \frac{F(x+th) - F(x)}{t} \le \limsup_{\substack{y \to x \\ t \to 0^+}} \frac{F(y+th) - F(y)}{t} = F^\circ(x;h).$$

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

$$\begin{aligned} F^\circ(x;h) &= \lim_{\varepsilon \to 0^+} \sup_{y \in K_{\delta\varepsilon}(x)} \sup_{0<t<\varepsilon} \frac{F(y+th) - F(y)}{t} \\ &\le \lim_{\varepsilon \to 0^+} \sup_{y \in K_{\delta\varepsilon}(x)} \frac{F(y+\varepsilon h) - F(y)}{\varepsilon} \\ &\le \lim_{\varepsilon \to 0^+} \frac{F(x+\varepsilon h) - F(x)}{\varepsilon} + 2L\delta \\ &= F'(x;h) + 2L\delta, \end{aligned}$$

where the last inequality follows by adding two productive zeros and using the local Lipschitz continuity in x. Since $\delta > 0$ was arbitrary, this implies that $F^\circ(x;h) \le F'(x;h)$, and the claim follows. □

A locally Lipschitz continuous functional $F : X \to \mathbb{R}$ with $F^\circ(x;h) = F'(x;h)$ for all $h \in X$ is called *regular* in $x \in X$. We have just shown that every continuously differentiable and every convex and lower semicontinuous functional is regular; intuitively, a function is thus regular in any points in which it is either differentiable or has at most a "convex kink".

Finally, similarly to Theorem 4.6 one can show the following pointwise characterization of the Clarke subdifferential of integral functionals with Lipschitz continuous integrands. We again assume that $\Omega \subset \mathbb{R}^d$ is open and bounded.

Theorem 8.7. *Let $f : \mathbb{R} \to \mathbb{R}$ be Lipschitz continuous and $F : L^p(\Omega) \to \overline{\mathbb{R}}$ with $1 \le p < \infty$ as in Lemma 3.6. Then we have for all $u \in L^p(\Omega)$ with $q = \frac{p}{p-1}$ (where $q = \infty$ for $p = 1$) that*

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

If f is regular at $u(x)$ for almost every $x \in \Omega$, then F is regular at u, and equality holds.

Proof. First, by the properties of the Lebesgue integral and the Lipschitz continuity of f, we have for any $u, v \in L^p(\Omega)$ that

$$|F(u) - F(v)| \leq \int_\Omega |f(u(x)) - f(v(x))| \, dx \leq L \int_\Omega |u(x) - v(x)| \, dx \leq LC_p \|u - v\|_{L^p},$$

where L is the Lipschitz constant of f and C_p the constant from the continuous embedding $L^p(\Omega) \hookrightarrow L^1(\Omega)$ for $1 \leq p \leq \infty$. Hence $F : L^p(\Omega) \to \mathbb{R}$ is Lipschitz continuous and therefore finite-valued as well.

Let now $\xi \in \partial_C F(u) \subset L^p(\Omega)^*$ be given and $h \in L^p(\Omega)$ be arbitrary. By definition, we thus have

$$\begin{aligned}
(8.2) \qquad \langle \xi, h \rangle_{L^p} \leq F^\circ(u; h) &= \limsup_{\substack{v \to u \\ t \to 0}} \frac{F(v + th) - F(v)}{t} \\
&\leq \int_\Omega \limsup_{\substack{v \to u \\ t \to 0}} \frac{f(v(x) + th(x)) - f(v(x))}{t} \, dx \\
&\leq \int_\Omega \limsup_{\substack{v_x \to u(x) \\ t_x \to 0}} \frac{f(v_x + t_x h(x)) - f(v_x)}{t_x} \, dx \\
&= \int_\Omega f^\circ(u(x); h(x)) \, dx,
\end{aligned}$$

where we were able to use the Reverse Fatou Lemma to exchange the lim sup with the integral in the first inequality since the integrand is bounded from above by the integrable function $L|h|$ due to Lemma 8.1 (i); the second inequality follows by bounding for almost every $x \in \Omega$ the (pointwise) limit over the sequences realizing the lim sup in the second line by the lim sup over all admissible sequences.

To interpret (8.2) pointwise, we define for $x \in \Omega$

$$g_x : \mathbb{R} \to \mathbb{R}, \qquad g_x(t) := f^\circ(u(x); t).$$

From Lemma 8.1 (ii)–(iii), it follows that g_x is convex; Lemma 8.1 (i) further implies that the function $x \mapsto g_x(h(x))$ is measurable for any $h \in L^p(\Omega)$. Since $g_x(0) = 0$, (8.2) implies that

$$\langle \xi, h - 0 \rangle_{L^p} \leq \int_\Omega g_x(h(x)) \, dx - \int_\Omega g_x(0) \, dx,$$

i.e., $\xi \in \partial G(0)$ for the superposition operator $G(h) := \int_\Omega g_x(h(x)) \, dx$. Arguing exactly as in the proof of Theorem 4.6 (using that the spatially-varying(!) integrand $g_x(t)$ is measurable in x), this implies that $\xi = u^* \in L^q(\Omega)$ with $u^*(x) \in \partial g_x(0)$ for almost every $x \in \Omega$, i.e.,

$$u^*(x)h(x) = u^*(x)(h(x) - 0) \leq g_x(h(x)) - g_x(0) = f^\circ(u(x); h(x))$$

for almost every $x \in \Omega$. Since $h \in L^p(\Omega)$ was arbitrary, this implies that $u^*(x) \in \partial_C f(u(x))$ almost everywhere as claimed.

It remains to show the remaining assertions when f is regular. In this case, it follows from (8.2) that for any $h \in L^p(\Omega)$,

$$\begin{aligned}(8.3)\qquad F^\circ(u;h) &\leq \int_\Omega f^\circ(u(x);h(x))\,dx = \int_\Omega f'(u(x);h(x))\,dx \\ &\leq \lim_{t\to 0}\frac{F(u+th)-F(u)}{t} = F'(u;h) \leq F^\circ(u;h),\end{aligned}$$

where the second inequality is obtained by applying Fatou's Lemma, this time appealing to the integrable lower bound $-L|h(x)|$. This shows that $F'(u;h) = F^\circ(u;h)$ and hence that F is regular. We further obtain for any $u^* \in L^q(\Omega)$ with $u^*(x) \in \partial_C f(u(x))$ almost everywhere and any $h \in L^p(\Omega)$, that

$$\langle u^*, h\rangle_{L^p} = \int_\Omega u^*(x)h(x)\,dx \leq \int_\Omega f^\circ(u(x);h(x))\,dx \leq F^\circ(u,h),$$

where we have used (8.3) in the last inequality. Since $h \in L^p(\Omega)$ was arbitrary, this implies that $u^* \in \partial_C F(u)$. □

Under additional assumptions similar to those of Theorem 2.10 and with more technical arguments, this result can be extended to spatially varying integrands $f : \Omega \times \mathbb{R} \to \mathbb{R}$; see, e.g., [Clarke 1990, Theorem 2.7.5].

8.2 CALCULUS RULES

We now turn to calculus rules. The first one still follows directly from the definition.

Theorem 8.8. *Let $F : X \to \mathbb{R}$ 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 8.5). If $\alpha > 0$, we have that $(\alpha F)^\circ(x;h) = \alpha F^\circ(x;h)$ for all $h \in X$ from the definition. Hence,

$$\begin{aligned}\alpha\partial_C F(x) &= \{\alpha x^* \in X^* : \langle x^*, h\rangle_X \leq F^\circ(x;h) \quad \text{for all } h \in X\} \\ &= \{\alpha x^* \in X^* : \langle \alpha x^*, h\rangle_X \leq \alpha F^\circ(x;h) \quad \text{for all } h \in X\} \\ &= \{y^* \in X^* : \langle y^*, h\rangle_X \leq (\alpha F)^\circ(x;h) \quad \text{for all } h \in X\} \\ &= \partial_C(\alpha F)(x).\end{aligned}$$

To conclude the proof, it suffices to show the claim for $\alpha = -1$. For that, we use Lemma 8.1 (iv) to obtain that

$$\begin{aligned}\partial_C(-F)(x) &= \{x^* \in X^* : \langle x^*, h\rangle_X \le (-F)^\circ(x;h) \quad \text{for all } h \in X\}\\ &= \{x^* \in X^* : \langle -x^*, -h\rangle_X \le F^\circ(x;-h) \quad \text{for all } h \in X\}\\ &= \{-y^* \in X^* : \langle y^*, g\rangle_X \le F^\circ(x;g) \quad \text{for all } g \in X\}\\ &= -\partial_C(F)(x).\end{aligned}$$

□

Corollary 8.9. *Let $F : X \to \mathbb{R}$ 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 8.4 implies that

$$0 \in \partial_C(-F)(\bar x) = -\partial_C F(\bar x),$$

i.e., $0 = -0 \in \partial_C F(\bar x)$. □

SUPPORT FUNCTIONALS

The remaining rules are 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 8.10. *Let $S : X \to \mathbb{R}$ be positively homogeneous, subadditive, and lower semicontinuous, and let*

$$A = \{x^* \in X^* : \langle x^*, x\rangle_X \le S(x) \quad \textit{for all } x \in X\}\,.$$

Then

$$S(x) = \sup_{x^* \in A} \langle x^*, x\rangle_X \qquad \textit{for all } x \in X. \tag{8.4}$$

Proof. By definition of A, the inequality $\langle x^*, x\rangle_X - S(x) \le 0$ 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 S (which in particular implies that $S(0) = 0$) shows that

$$S^*(x^*) = \sup_{x \in X} \langle x^*, x\rangle_X - S(x) = \begin{cases} 0 & x^* \in A,\\ \infty & x^* \notin A,\end{cases}$$

i.e., $S^* = \delta_A$. Furthermore, by assumption S is also subadditive and hence convex as well as lower semicontinuous. Theorem 5.1 thus yields

$$S(x) = S^{**}(x) = (\delta_A)^*(x) = \sup_{x^* \in A} \langle x^*, x\rangle_X.$$

□

The right-hand side of (8.4) is called the *support functional* of A.

Lemma 8.11. *Let $A, B \subset X^*$ be nonempty, convex, and weakly-$*$ closed. Then $A \subset B$ if and only if*

$$\sup_{x^*\in A} \langle x^*, x\rangle_X \leq \sup_{x^*\in B} \langle x^*, x\rangle_X \qquad \textit{for all } x \in X. \tag{8.5}$$

Proof. If $A \subset B$, then the right-hand side of (8.5) is obviously not less than the left-hand side. Conversely, assume that there exists an $x^* \in A$ with $x^* \notin B$. By the assumptions on A and B, we then obtain from Theorem 1.11 an $x \in X$ and a $\lambda \in \mathbb{R}$ with

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

Taking the supremum over all $z^* \in B$ and estimating the right-hand side by the supremum over all $x^* \in A$ then yields that

$$\sup_{z^*\in B} \langle z^*, x\rangle_X < \sup_{x^*\in A} \langle x^*, x\rangle_X.$$

Hence (8.5) is violated, and the claim follows by contraposition. □

Corollary 8.12. *Let $A, B \subset X^*$ be nonempty, convex, and weakly-$*$ closed. Then $A = B$ if and only if*

$$\sup_{x^*\in A} \langle x^*, x\rangle_X = \sup_{x^*\in B} \langle x^*, x\rangle_X \qquad \textit{for all } x \in X. \tag{8.6}$$

Proof. Again, the claim is obvious if $A = B$. Conversely, if (8.6) holds, then in particular (8.5) holds, and we obtain from Lemma 8.11 that $A \subset B$. Exchanging the roles of A and B now yields the claim. □

Lemma 8.10 together with Lemma 8.1 directly yields the following useful representation.

Corollary 8.13. *Let $F : X \to \mathbb{R}$ be locally Lipschitz continuous and $x \in X$. Then*

$$F^\circ(x; h) = \sup_{x^*\in\partial_C F(x)} \langle x^*, h\rangle_X \quad \textit{for all } h \in X.$$

With its help, we can finally show the promised nonemptiness of the convex subdifferential.

Corollary 8.14. *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $x \in (\operatorname{dom} F)^o$. Then, $\partial F(x)$ is nonempty, convex, weakly-$*$ closed, and bounded.*

Proof. Since $x \in (\operatorname{dom} F)^o$, Theorem 8.6 shows that $\partial_C F(x) = \partial F(x)$ and that F is regular in x. It thus follows from Corollary 8.13 and Lemma 4.1 (iii) that $\sup_{x^*\in\partial F(x)} \langle x^*, h\rangle_X = F'(x; h) \in \mathbb{R}$ for $x \in (\operatorname{dom} F)^o$, and hence the supremum cannot be over the empty set (for which any supremum is $-\infty$ by convention). The remaining properties follow from Corollary 8.2. □

SUM RULE

We now use these results to prove a sum rule.

Theorem 8.15. *Let $F, G : X \to \mathbb{R}$ 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

$$(F+G)^\circ(x;h) \le F^\circ(x;h) + G^\circ(x;h).$$

If F and G are regular in x, the calculus of limits yields that

$$F^\circ(x;h) + G^\circ(x;h) = F'(x;h) + G'(x;h) = (F+G)'(x;h) \le (F+G)^\circ(x;h),$$

which implies that $(F+G)^\circ(x;h) = (F+G)'(x;h)$, i.e., $F+G$ is regular.

By Lemma 8.11 we are thus finished if we can show that

$$\partial_C F(x) + \partial_C G(x) = \{x^* \in X^* : \langle x^*, h\rangle_X \le F^\circ(x;h) + G^\circ(x;h) \text{ for all } h \in X\} =: A.$$

For this, we use that $\partial_C F(x)$ and $\partial_C G(x)$ are convex and weakly-$*$ closed by Corollary 8.2, and hence so is their sum since both sets are bounded. Furthermore, as shown in Lemma 8.1, generalized directional derivatives and hence their sums are positively homogeneous, convex, and lower semicontinuous. We thus obtain from Lemma 8.10 for all $h \in X$ that

$$\begin{aligned}\sup_{x^* \in \partial_C F(x) + \partial_C G(x)} \langle x^*, h\rangle_X &= \sup_{x_1^* \in \partial_C F(x)} \langle x_1^*, h\rangle_X + \sup_{x_2^* \in \partial_C G(x)} \langle x_2^*, h\rangle_X \\ &= F^\circ(x;h) + G^\circ(x;h) = \sup_{x^* \in A} \langle x^*, h\rangle_X.\end{aligned}$$

The claimed equality of A and the sum of the subdifferentials now follows from Corollary 8.12. □

Note the differences to the convex sum rule: The generic inclusion is now in the other direction; furthermore, *both* functionals have to be regular, and in exactly the point where the sum rule is applied. By induction, one obtains from this sum rule for an arbitrary number of functionals (which all have to be regular).

CHAIN RULE

To prove a chain rule, we need the following "nonsmooth" mean value theorem due to Lebourg.

Theorem 8.16. *Let $F : X \to \mathbb{R}$ be locally Lipschitz continuous near $x \in X$ and $\tilde{x}$ be in the Lipschitz neighborhood of x. Then there exists an $x^* \in \partial_C F(x + \lambda(\tilde{x} - x))$ for some $\lambda \in (0,1)$ such that*

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

Proof. Define $\psi, \varphi : [0,1] \to \mathbb{R}$ as

$$\psi(\lambda) := F(x + \lambda(\tilde{x} - x)), \qquad \varphi(\lambda) := \psi(\lambda) + \lambda(F(x) - F(\tilde{x})).$$

By the assumptions on F and $\tilde{x}$, both ψ and φ are Lipschitz continuous. In addition, $\varphi(0) = F(x) = \varphi(1)$, and hence φ has a local minimum or maximum in an interior point $\bar{\lambda} \in (0,1)$. From the Fermat principle Theorem 8.4 or Corollary 8.9, respectively, together with the sum rule from Theorem 8.15 and the characterization of the subdifferential of the second term from Theorem 8.5, we thus obtain that

$$0 \in \partial_C \varphi(\bar{\lambda}) \subset \partial_C \psi(\bar{\lambda}) + \{F(x) - F(\tilde{x})\}.$$

Hence we are finished if we can show for $x_{\bar{\lambda}} := x + \bar{\lambda}(\tilde{x} - x)$ that

$$\partial_C \psi(\bar{\lambda}) \subset \left\{\langle x^*, \tilde{x} - x\rangle_X : x^* \in \partial_C F(x_{\bar{\lambda}})\right\} =: A. \tag{8.7}$$

For this purpose, consider for arbitrary $s \in \mathbb{R}$ the generalized directional derivative

$$\begin{aligned}
\psi^\circ(\bar{\lambda}; s) &= \limsup_{\substack{\lambda \to \bar{\lambda}\\ t \to 0}} \frac{\psi(\lambda + ts) - \psi(\lambda)}{t}\\
&= \limsup_{\substack{\lambda \to \bar{\lambda}\\ t \to 0}} \frac{F(x + (\lambda + ts)(\tilde{x} - x)) - F(x + \lambda(\tilde{x} - x))}{t}\\
&\leq \limsup_{\substack{z \to x_{\bar{\lambda}}\\ t \to 0}} \frac{F(z + ts(\tilde{x} - x)) - F(z)}{t} = F^\circ(x_{\bar{\lambda}}; s(\tilde{x} - x)),
\end{aligned}$$

where the inequality follows from considering arbitrary sequences $z \to x_{\bar{\lambda}}$ (instead of special sequences of the form $z_n = x + \lambda_n(\tilde{x} - x)$) in the last lim sup. Lemma 8.11 thus implies that

$$\partial_C \psi(\bar{\lambda}) \subset \left\{t^* \in \mathbb{R} : t^* s \leq F^\circ(x_{\bar{\lambda}}; s(\tilde{x} - x)) \text{ for all } s \in \mathbb{R}\right\} =: B. \tag{8.8}$$

It remains to show that the sets A and B from (8.7) and (8.8) coincide. But this follows again from Lemma 8.10 and Corollary 8.12, since for all $s \in \mathbb{R}$ we have that

$$\sup_{t^* \in A} t^* s = \sup_{x^* \in \partial_C F(x_{\bar{\lambda}})} \langle x^*, s(\tilde{x} - x)\rangle_X = F^\circ(x_{\bar{\lambda}}; s(\tilde{x} - x)) = \sup_{t^* \in B} t^* s. \qquad \square$$

We also need the following generalization of the argument in Theorem 8.5.

Lemma 8.17. *Let X, Y be Banach spaces and $F : X \to Y$ be continuously Fréchet differentiable at $x \in X$. Let $\{x_n\}_{n\in\mathbb{N}} \subset X$ be a sequence with $x_n \to x$ and $\{t_n\}_{n\in\mathbb{N}} \subset (0,\infty)$ be a sequence with $t_n \to 0$. Then for any $h \in X$,*

$$\lim_{n\to\infty} \frac{F(x_n + t_n h) - F(x_n)}{t_n} = F'(x)h.$$

Proof. Let $h \in X$ be arbitrary. By the Hahn–Banach extension Theorem 1.4, for every $n \in \mathbb{N}$ there exists a $y_n^* \in Y^*$ with $\|y_n^*\|_{Y^*} = 1$ and

$$\|t_n^{-1}(F(x_n + t_n h) - F(x_n)) - F'(x)h\|_Y = \langle y_n^*, t_n^{-1}(F(x_n + t_n h) - F(x_n)) - F'(x)h\rangle_Y.$$

Applying now the classical mean value theorem to the scalar functions

$$f_n : [0,1] \to \mathbb{R}, \qquad f_n(s) = \langle y_n^*, F(x_n + st_n h)\rangle_Y,$$

we obtain similarly to the proof of Theorem 2.6 for all $n \in \mathbb{N}$ that

$$\begin{aligned}\|t_n^{-1}(F(x_n + t_n h) - F(x_n)) - F'(x)h\|_Y &= t_n^{-1} \int_0^1 \langle y_n^*, F'(x_n + st_n h)t_n h\rangle_Y \, ds - \langle y_n^*, F'(x)h\rangle_Y \\ &= \int_0^1 \langle y_n^*, (F'(x_n + st_n h) - F'(x))h\rangle_Y \, ds \\ &\leq \int_0^1 \|F'(x_n + st_n h) - F'(x))\|_{L(X;Y)} \, ds \, \|h\|_X,\end{aligned}$$

where we have used (1.1) together with $\|y_n^*\|_{Y^*} = 1$ in the last step. Since F' is continuous by assumption, the integrand goes to zero as $n \to \infty$ uniformly in $s \in [0,1]$, and the claim follows. □

We now come to the chain rule, which in contrast to the convex case does not require the inner mapping to be linear; this is one of the main advantages of the Clarke subdifferential in the context of nonsmooth optimization.

Theorem 8.18. *Let Y be a separable Banach space, $F : X \to Y$ be continuously Fréchet differentiable at $x \in X$, and $G : Y \to \mathbb{R}$ be locally Lipschitz continuous near $F(x)$. Then*

$$\partial_C(G \circ F)(x) \subset F'(x)^* \partial_C G(F(x)) := \{F'(x)^* y^* : y^* \in \partial_C G(F(x))\}.$$

If G is regular at $F(x)$, then $G \circ F$ is regular at x, and equality holds.

Proof. The local Lipschitz continuity of $G \circ F$ follows from that of G and F (which follows from the assumption as in the proof of Theorem 8.5). For the claimed inclusion (or equality), we argue as before using the support functional calculus. First we show that for every $h \in X$ there exists a $y^* \in \partial_C G(F(x))$ with

$$(G \circ F)^\circ(x;h) = \langle y^*, F'(x)h\rangle_Y. \tag{8.9}$$

To this end, consider for given $h \in X$ sequences $\{x_n\}_{n\in\mathbb{N}} \subset X$ and $\{t_n\}_{n\in\mathbb{N}} \subset (0,\infty)$ with $x_n \to x$, $t_n \to 0$, and

$$(G \circ F)^\circ(x;h) = \lim_{n\to\infty} \frac{G(F(x_n + t_n h)) - G(F(x_n))}{t_n}.$$

Furthermore, by continuity of F, we can find $n_0 \in \mathbb{N}$ such that $F(x_n), F(x_n + t_n h)$ lie in the Lipschitz neighborhood of $F(x)$ for all $n \geq n_0$. Theorem 8.16 thus yields for all $n \geq n_0$ a $y_n^* \in \partial_C G(y_n)$ with $y_n := F(x_n) + \lambda_n(F(x_n + t_n h) - F(x_n))$ for some $\lambda_n \in (0,1)$ such that

$$\frac{G(F(x_n + t_n h)) - G(F(x_n))}{t_n} = \langle y_n^*, q_n\rangle_Y \quad \text{with} \quad q_n := \frac{F(x_n + t_n h) - F(x_n)}{t_n} \tag{8.10}$$

Since $\lambda_n \in (0,1)$ is uniformly bounded, we also have that $y_n \to F(x)$ for $n \to \infty$. Hence y_n is in the Lipschitz neighborhood of $F(x)$ for $n \in \mathbb{N}$ large enough, and Corollary 8.2 yields that $y_n^* \in \partial_C G(y_n) \subset K_L(0)$ for $n \in \mathbb{N}$ sufficiently large. This implies that $\{y_n^*\}_{n\in\mathbb{N}} \subset Y^*$ is bounded, and the Banach–Alaoglu Theorem 1.10 yields a weakly-$*$ convergent subsequence with limit $y^* \in \partial_C G(F(x))$ by Lemma 8.3. Finally, since F is continuously Fréchet differentiable, $q_n \to F'(x)h$ strongly in Y by Lemma 8.17. Hence, $\langle y_n^*, q_n\rangle_Y \to \langle y^*, F'(x)h\rangle$ as the duality pairing of weakly-$*$ and strongly converging sequences. Passing to the limit in (8.10) therefore yields (8.9) (first along the subsequence chosen above; by convergence of the left-hand side of (8.10) and the uniqueness of limits then for the full sequence as well). By definition of the Clarke subdifferential, we thus have for $y^* \in \partial_C G(F(x))$ that

$$(G \circ F)^\circ(x;h) = \langle y^*, F'(x)h\rangle_Y \leq G^\circ(F(x); F'(x)h). \tag{8.11}$$

If G is now regular at x, we have that $G^\circ(F(x); F'(x)h) = G'(F(x); F'(x)h)$ and hence by the local Lipschitz continuity of G and the Fréchet differentiability of F that

$$\begin{aligned}
&G^\circ(F(x); F'(x)h) \\
&\qquad = \lim_{t\to 0} \frac{G(F(x) + tF'(x)h) - G(F(x))}{t} \\
&\qquad = \lim_{t\to 0} \frac{G(F(x) + tF'(x)h) - G(F(x+th)) + G(F(x+th)) - G(F(x))}{t} \\
&\qquad \leq \lim_{t\to 0} \left(L\|h\|_X \frac{\|F(x) + F'(x)th - F(x+th)\|_Y}{\|th\|_X} + \frac{G(F(x+th)) - G(F(x))}{t} \right) \\
&\qquad = (G \circ F)'(x;h) \leq (G \circ F)^\circ(x;h).
\end{aligned}$$

Together with (8.11), this implies that $(G \circ F)'(x; h) = (G \circ F)^\circ(x; h)$ (i.e., $G \circ F$ is regular at x) and that

$$(G \circ F)^\circ(x; h) = G^\circ(F(x); F'(x)h). \tag{8.12}$$

As before, Lemma 8.10 now implies for all $h \in X$ that

$$\sup_{x^* \in F'(x)^* \partial_C G(F(x))} \langle x^*, h\rangle_X = \sup_{y^* \in \partial_C G(F(x))} \langle y^*, F'(x)h\rangle_Y = G^\circ(F(x); F'(x)h)$$

and hence by Lemma 8.11 that

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

Combined with (8.11) or (8.12) and the definition of the Clarke subdifferential in (8.1), this now yields the claimed inclusion or equality, respectively, for the Clarke subdifferential of the composition. □

Again, the generic inclusion is the reverse of the one in the convex chain rule. Note that equality in the chain rule also holds if $-G$ is regular, since we can then apply Theorem 8.18 to $-G \circ F$ and use that $\partial_C(-G)(F(x)) = -\partial_C G(F(x))$ by Theorem 8.8. Furthermore, if G is not regular but $F'(x)$ is surjective, a similar proof shows that equality (but not the regularity of $G \circ F$) holds in the chain rule; see [Clarke 2013, Theorem 10.19].

8.3 CHARACTERIZATION IN FINITE DIMENSIONS

A more explicit characterization of the Clarke subdifferential is possible in finite-dimensional spaces. The basis is the following theorem, which only holds in $\mathbb{R}^N$; a proof can be found in, e.g., [DiBenedetto 2002, Theorem 23.2] or [Heinonen 2005, Theorem 3.1].

Theorem 8.19 (Rademacher). *Let $U \subset \mathbb{R}^N$ be open and $F : U \to \mathbb{R}$ be Lipschitz continuous. Then F is Fréchet differentiable 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 8.20. *Let $F : \mathbb{R}^N \to \mathbb{R}$ 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*

$$\partial_C F(x) = \operatorname{co}\left\{\lim_{n\to\infty} \nabla F(x_n) : x_n \to x,\ x_n \notin E_F\right\}, \tag{8.13}$$

where $\operatorname{co} A$ denotes the convex hull of $A \subset \mathbb{R}^N$.

Proof. We first note that the Rademacher Theorem ensures that such a set E_F exists and has Lebesgue measure 0. Hence there indeed exist sequences $\{x_n\}_{n\in\mathbb{N}} \in \mathbb{R}^N \setminus E_F$ with $x_n \to x$. Furthermore, the local Lipschitz continuity of F yields that for any x_n in the Lipschitz neighborhood of x and any $h \in \mathbb{R}^N$, we have that

$$|\,(\nabla F(x_n), h)\,| = \left|\lim_{t\to 0^+} \frac{F(x_n+th) - F(x_n)}{t}\right| \leq L\|h\|$$

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 (8.13) is therefore nonempty.

Let now $\{x_n\}_{n\in\mathbb{N}} \subset \mathbb{R}^N \setminus E_F$ be an arbitrary sequence with $x_n \to x$ and $\{\nabla F(x_n)\}_{n\in\mathbb{N}} \to x^*$ for some $x^* \in \mathbb{R}^N$. Since F is differentiable in every $x_n \notin E_F$, we have that

$$\langle \nabla F(x_n), h\rangle = F'(x;h) \leq F^\circ(x;h)$$

and hence that $\nabla F(x_n) \in \partial_C F(x_n)$ by definition. Lemma 8.3 thus yields that $x^* \in \partial_C F(x)$. The convexity of $\partial_C F(x)$ from Corollary 8.2 now implies that any convex combination of such limits x^* is contained in $\partial_C F(x)$, which shows the inclusion "$\supset$" in (8.13).

For the other inclusion, we first show for all $h \in \mathbb{R}^N$ and $\varepsilon > 0$ that

$$F^\circ(x;h) - \varepsilon \leq \limsup_{E_F \not\ni y \to x} \,(\nabla F(y), h) =: M(h). \tag{8.14}$$

Indeed, by definition of $M(h)$ and of the lim sup, for every $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$(\nabla F(y), h) \leq M(h) + \varepsilon \qquad \text{for all } y \in O_\delta(x) \setminus E_F.$$

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

$$F(y+th) - F(y) = \int_0^t (\nabla F(y+sh), h)\, ds \leq t(M(h)+\varepsilon) \tag{8.15}$$

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 (8.15) 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 $y \to x$ and $t \to 0$ now yields (8.14). 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 positively homogeneous, subadditive, and lower semicontinuous. We are thus finished if we can show that the set on the right-hand side of (8.13) – hereafter denoted by co A – can be written as

$$\operatorname{co} A = \left\{x^* \in \mathbb{R}^N : (x^*, h) \leq M(h) \quad \text{for all } h \in \mathbb{R}^N\right\}.$$

For this, we once again appeal to Corollary 8.12 (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

$$\sup_{x^* \in \operatorname{co} A} \langle x^*, h\rangle = \sup_{\substack{x_i^* \in A \\ \sum_i t_i = 1, t_i \geq 0}} \sum_i t_i \left\langle x_i^*, h\right\rangle = \sup_{\sum_i t_i = 1, t_i \geq 0} \sum_i t_i \sup_{x_i^* \in A} \left\langle x_i^*, h\right\rangle = \sup_{x^* \in A} \langle x^*, h\rangle$$

since the sum is maximal if and only if each summand is maximal. Now we have that

$$M(h) = \limsup_{E_F \not\ni y \to x} \langle \nabla F(y), h\rangle = \sup_{E_F \not\ni x_n \to x} \langle \lim_{n\to\infty} \nabla F(x_n), h\rangle = \sup_{x^* \in A} \langle x^*, h\rangle ,$$

and hence the claim follows from Lemma 8.10. □

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.

9.1 CONVERGENCE OF GENERALIZED NEWTON METHODS

As a motivation, we first consider the most general form of a Newton-type method. Let X and Y be Banach spaces and $F : X \to Y$ be given and suppose we are looking for an $\bar{x} \in X$ with $F(\bar{x}) = 0$. A Newton-type method to find such an $\bar{x}$ then consists of repeating the following steps:

1. choose an invertible $M_k := M(x^k) \in L(X, Y)$;
2. solve the *Newton step* $M_k s^k = -F(x^k)$;
3. update $x^{k+1} = x^k + s^k$.

We can now ask under which conditions this method converges to $\bar{x}$, and in particular, when the convergence is *superlinear*, i.e.,

$$\lim_{k\to\infty} \frac{\|x^{k+1} - \bar{x}\|_X}{\|x^k - \bar{x}\|_X} = 0. \tag{9.1}$$

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

$$\|M(x^k)^{-1}\|_{L(Y,X)} \leq C \qquad \text{for all } k \in \mathbb{N};$$

(ii) an *approximation condition*:

$$\lim_{k\to\infty} \frac{\|F(\bar{x}+e^k) - F(\bar{x}) - M(\bar{x}+e^k)e^k\|_Y}{\|e^k\|_X} = 0.$$

This motivates the following definition: We call $F : X \to Y$ *Newton differentiable* in $x \in X$ if there exists a neighborhood $U \subset X$ of x and a mapping $D_N F : U \to L(X; Y)$ such that

$$\lim_{\|h\|_X\to 0} \frac{\|F(x+h) - F(x) - D_N F(x+h)h\|_Y}{\|h\|_X} = 0. \tag{9.2}$$

We then call $D_N F(x)$ a *Newton derivative* of F at x. 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*

$$x^{k+1} = x^k - D_N F(x^k)^{-1} F(x^k). \tag{9.3}$$

Its local superlinear convergence follows directly from the construction.

Theorem 9.1. *Let X, Y be Banach spaces and let $F : X \to Y$ 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* (9.3) *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 $x^0 \in O_\delta(\bar{x})$,

$$\|e^1\|_X \leq C\|F(\bar{x}+e^0) - F(\bar{x}) - D_N F(\bar{x}+e^0)e^0\|_Y. \tag{9.4}$$

[1] Here we follow [Chen, Nashed & Qi 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].

Let now $\varepsilon \in (0,1)$ be arbitrary. The Newton differentiability of F then implies that there exists a $\rho > 0$ such that

$$\|F(\bar{x}+h) - F(\bar{x}) - D_N F(\bar{x}+h)h\|_Y \leq \frac{\varepsilon}{C}\|h\|_X \qquad \text{for all } \|h\|_X \leq \rho.$$

Hence, if we choose x^0 such that $\|\bar{x} - x^0\|_X \leq \min\{\delta, \rho\}$, the estimate (9.4) implies that $\|\bar{x}-x^1\|_X \leq \varepsilon\|\bar{x}-x^0\|_X$. By induction, we obtain from this that $\|\bar{x}-x^k\|_X \leq \varepsilon^k\|\bar{x}-x^0\|_X \to 0$. Since $\varepsilon \in (0,1)$ was arbitrary, we can take in each step k a different $\varepsilon_k \to 0$, which shows that the convergence is in fact superlinear. □

9.2 NEWTON DERIVATIVES

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 $F : X \to Y$ 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

$$\begin{aligned}\|F(x+h) - F(x) - F'(x+h)h\|_Y \leq \|F(x+h) - F(x) - F'(x)h\|_Y \\ + \|F'(x) - F'(x+h)\|_{L(X,Y)}\|h\|_X,\end{aligned}$$

where the first summand is $o(\|h\|_X)$ by definition of the Fréchet derivative and the second by the continuity of F'. □

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 $F : X \to Y$ be Newton differentiable in $x \in X$ with Newton derivative $D_N F(x)$ and $G : Y \to Z$ 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 $r_1(t)/t \to 0$ for $t \to 0$. The Newton differentiability of F further implies that

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

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

$$\|g\|_Y \le \|D_N F(x+h)\|_{L(X,Y)}\|h\|_Y + r_2(\|h\|_X).$$

The uniform boundedness of $D_N F$ now implies that $\|g\|_Y \to 0$ for $\|h\|_X \to 0$. Hence,

$$\begin{aligned}\|(G \circ F)(x+h) - (G \circ F)(x) - D_N G(F(x+h))D_N F(x+h)h\|_Z \\ \le \|G(y+g) - G(y) - D_N G(y+g)g\|_Z \\ + \|D_N G(y+g)\,[g - D_N F(x+h)h]\,\|_Z \\ \le r_1(\|g\|_Y) + \|D_N G(y+g)\|_{L(Y,Z)} r_2(\|h\|_X),\end{aligned}$$

and the claim thus follows from the uniform boundedness of $D_N G$. □

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 X, Y_i be Banach spaces and let $F_i : X \to Y_i$ be Newton differentiable with Newton derivative $D_N F_i$ for $1 \le i \le m$. Then*

$$F : X \to (Y_1 \times \cdots \times Y_m), \qquad x \mapsto (F_1(x), \dots, F_m(x))^T,$$

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 $F : \mathbb{R}^N \to \mathbb{R}$ is locally Lipschitz continuous, candidates can be taken from the Clarke subdifferential, which has an explicit characterization by Theorem 8.20. Under some additional assumptions, each candidate is indeed a Newton derivative.[2]

A function $F : \mathbb{R}^N \to \mathbb{R}$ is called *piecewise (continuously) differentiable* or *PC^1 function*, if

(i) F is continuous on $\mathbb{R}^N$;

(ii) for all $x \in \mathbb{R}^N$ there exists an open neighborhood $U \subset \mathbb{R}^N$ of x and a finite set $\{F_i : U \to \mathbb{R}\}_{i\in I}$ of continuously differentiable functions with

$$F(\tilde{x}) \in \{F_i(\tilde{x})\}_{i\in I} \qquad \text{for all } \tilde{x} \in U.$$

In this case, we call F a *measurable selection* of the F_i in U. The set

$$I_a(x) := \{i \in I : F(x) = F_i(x)\}$$

is called the *active index set* at x. Since the F_i are continuous, we have that $F(\tilde{x}) \neq F_j(\tilde{x})$ for all $j \notin I_a(x)$ and $\tilde{x}$ sufficiently close to x. Hence, indices that are only active on sets of zero measure do not have to be considered in the following. We thus define the *essentially active index set*

$$I_e(x) := \{i \in I : x \in \operatorname{cl}\left(\{\tilde{x} \in U : F(\tilde{x}) = F_i(\tilde{x})\}^o\right)\} \subset I_a(x).$$

An example of an active but not essentially active index set is the following.

Example 9.5. Consider the function $f : \mathbb{R} \to \mathbb{R}$, $t \mapsto \max\{0, t, t/2\}$, i.e., $f_1(t) = 0$, $f_2(t) = t$ and $f_3(t) = t/2$. Then $I_a(0) = \{1, 2, 3\}$ but $I_e(0) = \{1, 2\}$, since f_3 is active only in $t = 0$ and hence $\{t \in \mathbb{R} : f(t) = f_3(t)\}^o = \emptyset = \operatorname{cl} \emptyset$.

Since any C^1 function $F_i : U_x \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant $L_i := \sup_{\tilde{x}\in U_x} |\nabla F(\tilde{x})|$, PC^1 functions are always locally Lipschitz continuous; see [Scholtes 2012, Corollary 4.1.1].

Theorem 9.6. *Let $F : \mathbb{R}^N \to \mathbb{R}$ be piecewise differentiable. Then F is locally Lipschitz continuous in all $x \in \mathbb{R}^N$ with local constant $L(x) = \max_{i\in I_a(x)} L_i$.*

This yields the following explicit characterization of the Clarke subdifferential of a PC^1 function.

[2] This is the original derivation of semismooth Newton methods.

Theorem 9.7. *Let $F : \mathbb{R}^N \to \mathbb{R}$ be piecewise differentiable and $x \in \mathbb{R}^N$. Then*

$$\partial_C F(x) = \operatorname{co}\{\nabla F_i(x) : i \in I_e(x)\}.$$

Proof. Let $x \in \mathbb{R}^N$ be arbitrary. By Theorem 8.20 it suffices to show that

$$\left\{\lim_{n\to\infty} \nabla F(x_n) : x_n \to x,\ x_n \notin E_F\right\} = \{\nabla F_i(x) : i \in I_e(x)\}.$$

For this, let $\{x_n\}_{n\in\mathbb{N}} \subset \mathbb{R}^N$ be a sequence with $x_n \to x$ such that F is differentiable in x_n for all $n \in \mathbb{N}$, and $\nabla F(x_n) \to x^* \in \mathbb{R}^N$. Since F is differentiable in x_n, it must hold that $F(\tilde{x}) = F_{i_n}(\tilde{x})$ for some $i_n \in I$ and all $\tilde{x}$ sufficiently close to x_n, which implies that $\nabla F(x_n) = \nabla F_{i_n}(x_n)$. For sufficiently large $n \in \mathbb{N}$, we can further assume that $i_n \in I_e(x)$ (if necessary, by adding x_n with $i_n \notin I_e(x)$ to E_F, which does not increase its Lebesgue measure). If we now consider subsequences $\{x_{n_k}\}_{k\in\mathbb{N}}$ with constant index $i_{n_k} =: i \in I_e(x)$ (which exist since $I_e(x)$ is finite), we obtain using the continuity of ∇F_i that

$$x^* = \lim_{k\to\infty} \nabla F(x_{n_k}) = \lim_{k\to\infty} \nabla F_i(x_{n_k}) \in \{\nabla F_i(x) : i \in I_e(x)\}.$$

Conversely, for every $\nabla F_i(x)$ with $i \in I_e(x)$ there exists by definition of the essentially active indices a sequence $\{x_n\}_{n\in\mathbb{N}}$ with $x_n \to x$ and $F = F_i$ in a sufficiently small neighborhood of each x_n for n large enough. The continuous differentiability of the F_i thus implies that $\nabla F(x_n) = \nabla F_i(x_n)$ for all $n \in \mathbb{N}$ large enough and hence that

$$\nabla F_i(x) = \lim_{n\to\infty} \nabla F_i(x_n) = \lim_{n\to\infty} \nabla F(x_n). \qquad \square$$

From this, we obtain the Newton differentiability of PC^1 functions.

Theorem 9.8. *Let $F : \mathbb{R}^N \to \mathbb{R}$ be piecewise differentiable. Then F is Newton differentiable for all $x \in \mathbb{R}^N$, and every $D_N F(x) \in \partial_C F(x)$ is a Newton derivative.*

Proof. Let $x \in \mathbb{R}^N$ be arbitrary and $h \in X$ with $x + h \in U$. By Theorem 9.7, every $D_N F(x+h) \in \partial_C F(x+h)$ is of the form

$$D_N F(x+h) = \sum_{i\in I_e(x+h)} \lambda_i \nabla F_i(x+h) \qquad \text{for} \sum_{i\in I_e(x+h)} \lambda_i = 1, \lambda_i \geq 0.$$

Since F is continuous, we have for all $h \in \mathbb{R}^N$ sufficiently small that $I_e(x+h) \subset I_a(x+h) \subset I_a(x)$, where the second inclusion follows from the fact that by continuity, $F(x) \neq F_i(x)$ implies that $F(x+h) \neq F_i(x+h)$. Hence, $F(x+h) = F_i(x+h)$ and $F(x) = F_i(x)$ for all $i \in I_e(x+h)$. Theorem 9.2 then yields that

$$|F(x+h) - F(x) - D_N F(x+h)h| \leq \sum_{i\in I_e(x+h)} \lambda_i |F_i(x+h) - F_i(x) - \nabla F_i(x+h)h| = o(\|h\|),$$

since all F_i are continuously differentiable by assumption. $\square$

A natural application of the above are proximal point reformulations of optimality conditions for convex optimization problems.

Example 9.9. We consider the minimization of $F + G$ for a twice continuously differentiable functional $F : \mathbb{R}^N \to \mathbb{R}$ and $G = \|\cdot\|_1$. Proceeding as in the derivation of the forward–backward splitting (7.4), we can use the regularity of F and G to write the necessary optimality condition $0 \in \partial_C(F+G)(\bar x)$ equivalently as

$$\bar x - \operatorname{prox}_{\gamma G}(\bar x - \gamma \nabla F(\bar x)) = 0$$

for any $\gamma > 0$. By Example 6.16 (ii), the proximal point mapping for G is given componentwise as

$$[\operatorname{prox}_{\gamma G}(x)]_i = \begin{cases} x_i - \gamma & \text{if } x_i > \gamma,\\ 0 & \text{if } x_i \in [-\gamma,\gamma],\\ x_i + \gamma & \text{if } x_i < -\gamma,\end{cases}$$

which is clearly piecewise differentiable. Theorem 9.7 thus yields (also componentwise) that

$$[\partial_C(\operatorname{prox}_{\gamma G})(x)]_i = \begin{cases} \{1\} & \text{if } |x_i| > \gamma,\\ \{0\} & \text{if } |x_i| < \gamma,\\ [0,1] & \text{if } |x_i| = \gamma.\end{cases}$$

By Theorems 9.4 and 9.8, a possible Newton derivative is therefore given by

$$[D_N\operatorname{prox}_{\gamma G}(x)h]_i = [\mathbb{1}_{\{|x|\geq\gamma\}}h]_i := \begin{cases} h_i & \text{if } |x_i| \geq \gamma,\\ 0 & \text{if } |x_i| < \gamma.\end{cases}$$

(The choice which case to include the equality in is arbitrary here.) Now, $D_N\operatorname{prox}_{\gamma G}(x)$ 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

$$\left(\mathbb{1}_{\mathcal{I}_k} + \gamma \mathbb{1}_{\mathcal{A}_k}\nabla^2 F(x^k)\right) s^k = -x^k + \operatorname{prox}_{\gamma G}(x^k - \gamma\nabla F(x^k)),$$

where we have defined the *active* and *inactive sets*, respectively, as

$$\mathcal{A}_k := \left\{i \in \{1,\dots,N\} : |x_i^k - \gamma[\nabla F(x^k)]_i| \geq \gamma\right\}, \qquad \mathcal{I}_k := \{1,\dots,N\}\setminus\mathcal{A}_k.$$

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 $\operatorname{prox}_{\gamma G}$ (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 $f : \Omega \times \mathbb{R} \to \mathbb{R}$ (i.e., f is measurable in x and continuous in z), and $1 \le p, q \le \infty$ the corresponding superposition operator

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

The goal is now to similarly obtain a Newton derivative $D_N F$ for F as a superposition operator defined by the Newton derivative $D_N f(x,z)$ of $z \mapsto f(x,z)$. Here, the assumption that $D_N f$ is also a Carathéodory function is too restrictive, since we want to allow discontinuous derivatives as well (see Example 9.9). Luckily, for our purpose, a weaker property is sufficient: A function is called *Baire–Carathéodory function* if it can be written as a pointwise limit of Carathéodory functions, i.e., if

$$f(x,z) = \lim_{n\to\infty} f_n(x,z) \qquad \text{for almost every } x \in \Omega \text{ and all } z \in \mathbb{R},$$

where f_n is a Carathéodory function for all $n \in \mathbb{N}$; see [Appell & Zabreiko 1990, Lemma 1.4].

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.

Theorem 9.10. *Let $f : \Omega \times \mathbb{R} \to \mathbb{R}$ be a Carathéodory function. Furthermore, assume that*

(i) $z \mapsto f(x,z)$ is uniformly Lipschitz continuous for almost every $x \in \Omega$ and $f(x,0)$ 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.

[3] which can be significantly relaxed; see [Schiela 2008, Proposition A.1]

Then for any $1 \le q < p < \infty$, the corresponding superposition operator $F : L^p(\Omega) \to L^q(\Omega)$ is Newton differentiable with Newton derivative

$$D_N F : L^p(\Omega) \to L(L^p(\Omega), L^q(\Omega)), \qquad [D_N F(u)h](x) = D_N f(x, u(x))h(x)$$

for almost every $x \in \Omega$ and all $h \in L^p(\Omega)$.

Proof. First, the uniform Lipschitz continuity together with the reverse triangle inequality yields that

$$|f(x,z)| \le |f(x,0)| + L|z| \le C + L|z|^{q/q} \quad \text{for almost every } x \in \Omega \text{ and all } z \in \mathbb{R},$$

and hence the growth condition (2.6) is satisfied for all $1 \le q < \infty$. Due to the continuous embedding $L^p(\Omega) \hookrightarrow L^q(\Omega)$ for all $1 \le q < p < \infty$, the superposition operator $F : L^p(\Omega) \to L^q(\Omega)$ is therefore well-defined and continuous by Theorem 2.10.

For any measurable $u : \Omega \to \mathbb{R}$, we have that $x \mapsto D_N f(x, u(x))$ is by assumption (iii) the pointwise limit of measurable functions and hence itself measurable. Furthermore, its uniform boundedness in particular implies the growth condition (2.6) for $p' := p$ and $q' := p - q > 0$. As in the proof of Theorem 2.11, we deduce that the corresponding superposition operator $D_N F : L^p(\Omega) \to L^s(\Omega)$ is well-defined and continuous for $s := \frac{pq}{p-q}$, and that for any $u \in L^p(\Omega)$, the mapping $h \mapsto D_N F(u)h$ defines a bounded linear operator $D_N F(u) : L^p(\Omega) \to L^q(\Omega)$. (This time, we do not distinguish in notation between the linear operator and the function defining this operator by pointwise multiplication.)

To show that $D_N F(u)$ is a Newton derivative for F in $u \in L^p(\Omega)$, we consider the pointwise residual

$$r : \Omega \times \mathbb{R} \to \mathbb{R}, \qquad r(x,z) := \begin{cases} \frac{|f(x,z) - f(x,u(x)) - D_N f(x,z)(z-u(x))|}{|z-u(x)|} & \text{if } z \neq u(x), \\ 0 & \text{if } z = u(x). \end{cases}$$

Since f is a Carathéodory function and $D_N f$ is a Baire–Carathéodory function, the function $x \mapsto r(x, \tilde u(x)) =: R(\tilde u)$ is measurable for any measurable $\tilde u : \Omega \to \mathbb{R}$ (since sums, products, and quotients of measurable functions are again measurable). Furthermore, for $\tilde u \in L^p(\Omega)$, the uniform Lipschitz continuity of f and the uniform boundedness of $D_N f$ imply that

$$|[R(\tilde u)](x)| = \frac{|f(x, \tilde u(x)) - f(x, u(x)) - D_N f(x, \tilde u(x))(\tilde u(x) - u(x))|}{|\tilde u(x) - u(x)|} \le L + C \tag{9.5}$$

and thus that $R(\tilde u) \in L^\infty(\Omega)$. Hence, the superposition operator $R : L^p(\Omega) \to L^s(\Omega)$ is well-defined.

Let now $\{u_n\}_{n\in\mathbb{N}} \subset L^p(\Omega)$ be a sequence with $u_n \to u \in L^p(\Omega)$. Then there exists a subsequence, again denoted by $\{u_n\}_{n\in\mathbb{N}}$, with $u_n(x) \to u(x)$ for almost every $x \in \Omega$. Since $z \mapsto f(x,z)$ is Newton differentiable almost everywhere, we have by definition that $r(x, u_n(x)) \to 0$ for almost every $x \in \Omega$. Together with the boundedness from (9.5), Lebesgue's dominated convergence theorem therefore yields that $R(u_n) \to 0$ in $L^s(\Omega)$ (and hence along the full sequence since the limit is unique).[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

$$\|F(\tilde u) - F(u) - D_N F(\tilde u)(\tilde u - u)\|_{L^q} = \|R(\tilde u)(\tilde u - u)\|_{L^q} \le \|R(\tilde u)\|_{L^s}\|\tilde u - u\|_{L^p}.$$

If we now set $\tilde u := u + h$ for $h \in L^p(\Omega)$ with $\|h\|_{L^p} \to 0$, we have that $\|R(u+h)\|_{L^s} \to 0$ and hence by definition the Newton differentiability of F in u with Newton derivative $h \mapsto D_N F(u)h$ as claimed. □

For $p = q \in [1, \infty]$, however, the claim is false in general, as can be shown by counterexamples.

Example 9.11. We take

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(z) = \max\{0, z\} := \begin{cases} 0 & \text{if } z \le 0, \\ z & \text{if } z \ge 0. \end{cases}$$

This is a piecewise differentiable function, and hence by Theorem 9.8 we can for any $\delta \in [0,1]$ take as Newton derivative

$$D_N f(z)h = \begin{cases} 0 & \text{if } z < 0, \\ \delta h & \text{if } z = 0, \\ h & \text{if } z > 0. \end{cases}$$

We now consider the corresponding superposition operators $F : L^p(\Omega) \to L^p(\Omega)$ and $D_N F(u) \in L(L^p(\Omega); L^p(\Omega))$ for any $p \in [1, \infty)$ and show that the approximation condition (9.2) is violated for $\Omega = (-1, 1)$, $u(x) = -|x|$, and

$$h_n(x) = \begin{cases} \frac{1}{n} & \text{if } |x| < \frac{1}{n}, \\ 0 & \text{if } |x| \ge \frac{1}{n}. \end{cases}$$

First, it is straightforward to compute $\|h_n\|_{L^p}^p = \frac{2}{n^{p+1}}$. Then since $[F(u)](x) = \max\{0, -|x|\} = 0$ almost everywhere, we have that

$$[F(u+h_n) - F(u) - D_N F(u+h_n)h_n](x) = \begin{cases} -|x| & \text{if } |x| < \frac{1}{n}, \\ 0 & \text{if } |x| > \frac{1}{n}, \\ -\frac{\delta}{n} & \text{if } |x| = \frac{1}{n}, \end{cases}$$

[4] This step fails for $F : L^\infty(\Omega) \to L^\infty(\Omega)$ since pointwise convergence and boundedness together do not imply uniform convergence almost everywhere.

and thus

$$\|F(u+h_n) - F(u) - D_N F(u+h_n)h_n\|_{L^p}^p = \int_{-\frac{1}{n}}^{\frac{1}{n}} |x|^p \, dx = \frac{2}{p+1}\left(\frac{1}{n}\right)^{p+1}.$$

This implies that

$$\lim_{n\to\infty} \frac{\|F(u+h_n) - F(u) - D_N F(u+h_n)h_n\|_{L^p}}{\|h_n\|_{L^p}} = \left(\frac{1}{p+1}\right)^{\frac{1}{p}} \neq 0$$

and hence that F is not Newton differentiable from $L^p(\Omega)$ to $L^p(\Omega)$ for any $p < \infty$.

For the case $p = q = \infty$, we take $\Omega = (0,1)$, $u(x) = x$, and

$$h_n(x) = \begin{cases} nx - 1 & \text{if } x \leq \frac{1}{n}, \\ 0 & \text{if } x \geq \frac{1}{n}, \end{cases}$$

such that $\|h_n\|_{L^\infty} = 1$ for all $n \in \mathbb{N}$. We also have that $x + h_n = (1+n)x - 1 \leq 0$ for $x \leq \frac{1}{n+1} \leq \frac{1}{n}$ and hence that

$$[F(u+h_n) - F(u) - D_N F(u+h_n)h_n](x) = \begin{cases} (1+n)x - 1 & \text{if } x \leq \frac{1}{n+1}, \\ 0 & \text{if } x \geq \frac{1}{n+1} \end{cases}$$

since either $h_n = 0$ or $F(u+h_n) = F(u) + D_N F(u)h_n$ in the second case. Now,

$$\sup_{x\in(0,\frac{1}{n+1}]} |(1+n)x - 1| = 1 \qquad \text{for all } n \in \mathbb{N},$$

which implies that

$$\lim_{n\to\infty} \frac{\|F(u+h_n) - F(u) - D_N F(u+h_n)h_n\|_{L^p}}{\|h_n\|_{L^p}} = 1 \neq 0$$

and hence that F is not Newton differentiable from $L^\infty(\Omega)$ to $L^\infty(\Omega)$ either.

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 9.12. We consider as in Example 9.9 the minimization of $F + G$ for a twice continuously differentiable functional $F : L^2(\Omega) \to \mathbb{R}$ and $G = \|\cdot\|_{L^1}$. The proximal point reformulation of $0 \in \partial(F+G)(\bar u)$,

$$\bar u - \operatorname{prox}_{\gamma G}(\bar u - \gamma\nabla F(\bar u)) = 0,$$

now has to be considered as an equation in $L^2(\Omega)$; however, $\operatorname{prox}_{\gamma G}$ 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 6.17 and Example 6.21 is given pointwise as $[H_\gamma(p)](x) = h_\gamma(p(x))$ for

$$h_\gamma : \mathbb{R} \to \mathbb{R}, \qquad t \mapsto \begin{cases} \frac{1}{\gamma}(t-1) & \text{if } t > 1, \\ 0 & \text{if } t \in [-1,1], \\ \frac{1}{\gamma}(t+1) & \text{if } t < -1. \end{cases}$$

This function is clearly piecewise differentiable, and Theorem 9.7 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 9.8, a possible Newton derivative is therefore given by

$$D_N h_\gamma(t)h = \frac{1}{\gamma}\mathbb{1}_{\{|t|\geq 1\}}h := \begin{cases} \frac{1}{\gamma}h & \text{if } |t| \geq 1, \\ 0 & \text{if } |t| < 1. \end{cases}$$

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 9.10, the superposition operator $H_\gamma : L^p(\Omega) \to L^2(\Omega)$ is therefore Newton differentiable for all $p > 2$, and a possible Newton derivative is given by

$$[D_N H_\gamma(p)h](x) = \frac{1}{\gamma}\mathbb{1}_{\{|p(x)|\geq 1\}}h(x),$$

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

$$\left(\mathrm{Id} + \tfrac{1}{\gamma}\mathbb{1}_{\{|\nabla F(u^k)|\geq 1\}}\nabla^2 F(u^k)\right) s^k = -u^k + H_\gamma(-\nabla F(u^k)),$$

where in a slight abuse of notation, $\mathbb{1}_{\{|p|\geq 1\}}$ denotes the function $x \mapsto \mathbb{1}_{\{|p(x)|\geq 1\}}$.

In practice, the radius of convergence for semismooth Newtons applied to such a Moreau–Yosida regularization shrinks with $\gamma \to 0$. 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).

10 LIMITING SUBDIFFERENTIALS

While the Clarke subdifferential is a suitable concept for nonsmooth but convex or nonconvex but smooth functionals, it has severe drawbacks for nonsmooth *and* nonconvex functionals: As shown in Corollary 8.9, its Fermat principle cannot distinguish minimizers from maximizers. The reason is that the Clarke subdifferential is always convex, which is a direct consequence of its construction (8.1) via polarity with respect to (generalized) directional derivatives. To obtain sharper results for such functionals, it is therefore necessary to construct *nonconvex* subdifferentials directly via a *dual* limiting process. On the other hand, deriving calculus rules for the previous subdifferentials crucially exploited their convexity by applying Hahn–Banach separation theorems, and calculus rules for nonconvex subdifferentials are thus significantly more difficult to obtain. As in Chapter 8, we will assume throughout this chapter that X is a Banach space unless stated otherwise.

10.1 BOULIGAND SUBDIFFERENTIALS

The first definition is motivated by Theorem 8.20: We *define* a subdifferential as a suitable limit of classical derivatives (without convexification). For $F : X \to \overline{\mathbb{R}}$, we first define the *set of Gâteaux points*

$$G_F := \{x \in X : F \text{ is Gâteaux differentiable at } x\} \subset \operatorname{dom} F$$

and then the *Bouligand subdifferential* of F at x as

$$\partial_B F(x) := \{x^* \in X^* : DF(x_n) \rightharpoonup^* x^* \text{ for some } G_F \ni x_n \to x\}\,. \tag{10.1}$$

For $F : \mathbb{R}^N \to \mathbb{R}$ locally Lipschitz, it follows from Theorem 8.20 that $\partial_C F(x) = \operatorname{co} \partial_B F(x)$. However, unless X is finite-dimensional, it is not clear a priori that the Bouligand subdifferential is nonempty even for $x \in \operatorname{dom} F$.[1] Furthermore, the subdifferential does not admit a satisfactory calculus; not even a Fermat principle holds.

[1] Although in special cases it is possible to give a full characterization in Hilbert spaces; see, e.g., [Christof et al. 2018].

Example 10.1. Let $F : \mathbb{R} \to \mathbb{R}$, $F(x) := |x|$. Then F is differentiable at every $x \neq 0$ with $F'(x) = \operatorname{sign}(x)$. Correspondingly,

$$0 \notin \{-1, 1\} = \partial_B F(0).$$

To make this approach work therefore requires a more delicate limiting process. The remainder of this chapter is devoted to one such approach, where we only give an overview and state important results following [Mordukhovich 2006]. For an alternative, more axiomatic, approach to generalized derivatives of nonconvex functionals, we refer to [Penot 2013; Ioffe 2017].

10.2 FRÉCHET SUBDIFFERENTIALS

We begin with the following limiting construction, which combines the characterizations of both the Fréchet derivative and the convex subdifferential. Let X be a Banach space and $F : X \to \overline{\mathbb{R}}$. The *Fréchet subdifferential* (or *regular subdifferential* or *presubdifferential*) of F at x is then defined as[2]

$$\partial_F F(x) := \left\{ x^* \in X^* : \liminf_{y \to x} \frac{F(y) - F(x) - \langle x^*, y - x \rangle_X}{\|y - x\|_X} \geq 0 \right\}. \tag{10.2}$$

Note how this "localizes" the definition of the convex subdifferential around the point of interest: the numerator does not need to be nonnegative for all y; it suffices if this holds for any y sufficiently close to x. By a similar argument as for Theorem 4.3, we thus obtain a Fermat principle for *local* minimizers.

Theorem 10.2. *Let $F : X \to \overline{\mathbb{R}}$ be proper and $\bar{x} \in \operatorname{dom} F$ be a local minimizer. Then $0 \in \partial_F F(\bar{x})$.*

Proof. Let $\bar{x} \in \operatorname{dom} F$ be a local minimizer. Then there exists an $\varepsilon > 0$ such that $F(\bar{x}) \leq F(y)$ for all $y \in O_\varepsilon(\bar{x})$, which is equivalent to

$$\frac{F(y) - F(\bar{x}) - \langle 0, y - \bar{x} \rangle_X}{\|y - \bar{x}\|_X} \geq 0 \quad \text{for all } y \in O_\varepsilon(\bar{x}) \setminus \{\bar{x}\}.$$

Now for any strongly convergent sequence $y_n \to \bar{x}$, we have that $y_n \in O_\varepsilon(\bar{x})$ for n large enough. Taking the lim inf in the above inequality thus yields $0 \in \partial_F(\bar{x})$. □

For convex functionals, of course, the numerator is always nonnegative by definition, and the Fréchet subdifferential reduces to the convex subdifferential.

[2] The equivalence of (10.2) with the usual definition based on corresponding normal cones follows from, e.g., [Mordukhovich 2006, Theorem 1.86].

Theorem 10.3. *Let $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous and $x \in \operatorname{dom} F$. Then $\partial_F F(x) = \partial F(x)$.*

Proof. By definition of the convex subdifferential, any $x^* \in \partial F(x)$ satisfies

$$F(y) - F(x) - \langle x^*, y - x\rangle_X \geq 0 \quad \text{for all } y \in X.$$

Dividing by $\|x-y\|_X > 0$ for $y \neq x$ and taking the lim inf as $y \to x$ thus yields $x^* \in \partial_F F(x)$.

Conversely, let $x^* \in \partial_F F(x)$ and $h \in X \setminus \{0\}$ be arbitrary. Then for any $\delta > 0$, there exists an $\varepsilon > 0$ such that

$$\frac{F(x+th) - F(x) - \langle x^*, th\rangle_X}{t\|h\|_X} \geq -\delta \quad \text{for all } t \in (0, \varepsilon).$$

Multiplying by $\|h\|_X > 0$ and letting $t \to 0$, we obtain from Lemma 4.1 that

$$\langle x^*, h\rangle_X \leq \frac{F(x+th) - F(x)}{t} + \delta \to F'(x;h) + \delta. \tag{10.3}$$

Since $\delta > 0$ was arbitrary, this implies by Lemma 4.2 that $x^* \in \partial F(x)$. □

Similarly, for Fréchet differentiable functionals, the limit in (10.2) is zero for all sequences.

Theorem 10.4. *Let $F : X \to \mathbb{R}$ be Fréchet differentiable at $x \in X$. Then $\partial_F F(x) = \{F'(x)\}$.*

Proof. The definition of the Fréchet derivative immediately yields

$$\lim_{y\to x} \frac{F(y) - F(x) - \langle F'(x), y - x\rangle_X}{\|x-y\|_X} = \lim_{\|h\|_X \to 0} \frac{F(x+h) - F(x) - F'(x)h}{\|h\|_X} = 0$$

and hence $F'(x) \in \partial_F F(x)$.

Conversely, let $x^* \in \partial_F F(x)$ and let again $h \in X \setminus \{0\}$ be arbitrary. As in the proof of Theorem 10.3, we then obtain that

$$\langle x^*, h\rangle_X \leq F'(x;h) = \langle F'(x), h\rangle_X. \tag{10.4}$$

Applying the same argument to $-h$ then yields $\langle x^*, h\rangle_X = \langle F'(x), h\rangle_X$ for all $h \in X$, i.e., $x^* = F'(x)$. □

For nonsmooth and nonconvex functionals, the Fréchet subdifferential can be strictly smaller than the Clarke subdifferential.

Example 10.5. Consider $F : \mathbb{R} \to \mathbb{R}$, $F(x) := -|x|$. For any $x \neq 0$, it follows from Theorem 10.4 that $\partial_F F(x) = \{-\operatorname{sign} x\}$. But for $x = 0$ and arbitrary $x^* \in \mathbb{R}$, we have that

$$\liminf_{y\to 0} \frac{F(y) - F(0) - \langle x^*, y - 0\rangle}{|y-0|} = \liminf_{y\to 0} (-1 - x^* \cdot \operatorname{sign}(y)) = -1 - |x^*| < 0$$

and hence that

$$\partial_F F(0) = \emptyset \subsetneq [-1,1] = \partial_C F(0).$$

Note that $0 \in \operatorname{dom} F$ in this example. Although the Fréchet subdifferential does not pick up a maximizer in contrast to the Clarke subdifferential, the fact that $\partial_F F(x)$ can be empty even for $x \in \operatorname{dom} F$ is a problem when trying to derive calculus rules that hold with equality. In fact, as Example 10.5 shows, the set-valued mapping $x \mapsto \partial_F F(x)$ fails to be closed, which is also not desirable. This leads to the next and final definition.

10.3 MORDUKHOVICH SUBDIFFERENTIALS

Let X be a reflexive Banach space and $F : X \to \overline{\mathbb{R}}$. The *Mordukhovich subdifferential* (or *basic subdifferential* or *limiting subdifferential*) of F at $x \in \operatorname{dom} F$ is then defined as the strong-to-weak* closure of $\partial_F F(x)$, i.e.,[3]

$$\begin{aligned}\partial_M F(x) &:= \text{w-}{*}\text{-}\limsup_{y \to x} \partial_F F(y) \\ &= \left\{x^* \in X^* : x_n^* \rightharpoonup^* x^* \text{ for } x_n^* \in \partial_F F(x_n) \text{ with } x_n \to x\right\},\end{aligned} \tag{10.5}$$

which can be seen as a generalization of the definition (10.1) of the Bouligand subdifferential. Note that in contrast to (10.1), this definition includes the constant sequence $x_n^* \equiv x^*$ even at nondifferentiable points, which makes this a more useful concept in general. This also implies that $\partial_F F(x) \subset \partial_M F(x)$ for any F, and Theorem 10.2 immediately yields a Fermat principle.

Corollary 10.6. *Let $F : X \to \overline{\mathbb{R}}$ be proper and $\bar x \in \operatorname{dom} F$ be a local minimizer. Then $0 \in \partial_M F(\bar x)$.*

As for the Fréchet subdifferential, maximizers do not satisfy the Fermat principle.

Example 10.7. Consider again $F : \mathbb{R} \to \mathbb{R}$, $F(x) := -|x|$. Using Example 10.5, we directly obtain from (10.5) that $\partial_M F(0) = \{-1, 1\} = \partial_B F(0)$.

Since the convex subdifferential is strong-to-weak* closed, the Mordukhovich subdifferential reduces to the convex subdifferential as well.

Theorem 10.8. *Let X be a reflexive Banach space, $F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous, and $x \in \operatorname{dom} F$. Then $\partial_M F(x) = \partial F(x)$.*

[3] The equivalence of this definition with the original geometric definition – which holds in *reflexive* Banach spaces – follows from [Mordukhovich 2006, Theorem 2.34].

Proof. From Theorem 10.3, it follows that $\partial F(x) = \partial_F F(x) \subset \partial_M F(x)$. Let therefore $x^* \in \partial_M F(x)$ be arbitrary. Then by definition there exists a sequence $\{x_n^*\}_{n\in\mathbb{N}} \subset X^*$ with $x_n^* \rightharpoonup^* x^*$ and $x_n^* \in \partial_F F(x_n) = \partial F(x_n)$ for $x_n \to x$. As in the proof of Corollary 6.7, it then follows that $x^* \in \partial F(x)$ as well. □

A similar result holds for *continuously* differentiable functionals.

Theorem 10.9. *Let X be a reflexive Banach space and $F : X \to \overline{\mathbb{R}}$ be continuously differentiable at $x \in X$. Then $\partial_M F(x) = \{F'(x)\}$.*

Proof. From Theorem 10.3, it follows that $\{F'(x)\} = \partial_F F(x) \subset \partial_M F(x)$. Let therefore $x^* \in \partial_M F(x)$ be arbitrary. Then by definition there exists a sequence $\{x_n^*\}_{n\in\mathbb{N}} \subset X^*$ with $x_n^* \rightharpoonup^* x^*$ and $x_n^* \in \partial_F F(x_n) = \{F'(x_n)\}$ for $x_n \to x$. The continuity of F' then immediately implies that $F'(x_n) \to F(x)$, and since strong limits are also weak-$*$ limits, we obtain $x^* = F'(x)$. □

We also have the following relation to Clarke subdifferentials, which should be compared to Theorem 8.20.

Theorem 10.10 ([Mordukhovich 2006, Theorem 3.57]). *Let X be a reflexive Banach space and $F : X \to \mathbb{R}$ be locally Lipschitz continuous around $x \in X$. Then $\partial_C F(x) = \operatorname{cl}^* \operatorname{co} \partial_M F(x)$, where $\operatorname{cl}^* A$ stands for the weak-$*$ closure of the set $A \subset X^*$.*[4]

The following example illustrates that the Mordukhovich subdifferential can be nonconvex.

Example 10.11. Let $F : \mathbb{R}^2 \to \mathbb{R}$, $F(x_1, x_2) = |x_1| - |x_2|$. Since F is continuously differentiable for any (x_1, x_2) where $x_1, x_2 \neq 0$ with

$$\nabla F(x_1, x_2) \in \{(1,1), (-1,1), (1,-1), (-1,-1)\},$$

[4] Of course, in reflexive Banach spaces the weak-$*$ closure coincides with the weak closure. The statement holds more general in so-called *Asplund spaces* which include some nonreflexive Banach spaces.

we obtain from (10.2) that

$$\partial_F F(x_1,x_2) = \begin{cases} \{(1,-1)\} & \text{if } x_1 > 0, x_2 > 0,\\ \{(-1,-1)\} & \text{if } x_1 < 0, x_2 > 0,\\ \{(-1,1)\} & \text{if } x_1 < 0, x_2 < 0,\\ \{(1,1)\} & \text{if } x_1 > 0, x_2 < 0,\\ \{(t,-1) : t \in [-1,1]\} & \text{if } x_1 = 0, x_2 > 0,\\ \{(t,1) : t \in [-1,1]\} & \text{if } x_1 = 0, x_2 < 0,\\ \emptyset & \text{if } x_2 = 0. \end{cases}$$

In particular, $\partial_F F(0,0) = \emptyset$. However, from (10.5) it follows that

$$\partial_M F(0,0) = \{(t,-1) : t \in [-1,1]\} \cup \{(t,1) : t \in [-1,1]\}.$$

In particular, $0 \notin \partial_M F(0,0)$. On the other hand, Theorem 10.10 then yields that

$$\partial_C F(0,0) = \{(t,s) : t,s \in [-1,1]\} = [-1,1]^2 \tag{10.6}$$

and hence $0 \in \partial_C F(0,0)$. (Note that F attains neither a minimum nor a maximum on $\mathbb{R}^2$, while $(0,0)$ is a nonsmooth saddle-point.)

In contrast to the Bouligand subdifferential, the Mordukhovich subdifferential admits a satisfying calculus, although the assumptions are understandably more restrictive than in the convex setting. The first rule follows as always straight from the definition.

Theorem 10.12. *Let X be a reflexive Banach space and $F : X \to \overline{\mathbb{R}}$. Then for any $\lambda \geq 0$ and $x \in X$,*

$$\partial_M(\lambda F)(x) = \lambda \partial_M F(x).$$

Full calculus in infinite-dimensional spaces holds only for a rather small class of mappings.

Theorem 10.13 ([Mordukhovich 2006, Proposition 1.107]). *Let X be a reflexive Banach space, $F : X \to \mathbb{R}$ be continuously differentiable, and $G : X \to \overline{\mathbb{R}}$ be arbitrary. Then for any $x \in \operatorname{dom} G$,*

$$\partial_M(F+G)(x) = \{F'(x)\} + \partial_M G(x).$$

While the previous two theorems also hold for the Fréchet subdifferential (the latter even for merely Fréchet differentiable F), the following chain rule is only valid for the Mordukhovich subdifferential. Compared to Theorem 8.18, it also allows for the outer functional to be extended-real valued.

Theorem 10.14 ([Mordukhovich 2006, Proposition 1.112]). *Let X be a reflexive Banach space, $F : X \to Y$ be continuously differentiable, and $G : Y \to \overline{\mathbb{R}}$ be arbitrary. Then for any $x \in X$ with $F(x) \in \operatorname{dom} G$ and $F'(x) : X \to Y$ surjective,*

$$\partial_M(G \circ F)(x) = F'(x)^* \partial_M G(F(x)).$$

More general calculus rules require X to be a reflexive Banach[5] space as well as additional, nontrivial, assumptions on F and G; see, e.g., [Mordukhovich 2006, Theorem 3.36 and Theorem 3.41].

[5] or Asplund

BIBLIOGRAPHY

H. W. Alt (2016), *Linear Functional Analysis, An application-oriented introduction*, Universitext, Springer, London, doi: 10.1007/978-1-4471-7280-2.

J. Appell & P. Zabreiko (1990), *Nonlinear Superposition Operators*, Cambridge University Press, New York.

H. Attouch, G. Buttazzo & G. Michaille (2006), *Variational Analysis in Sobolev and BV Spaces*, vol. 6, MPS/SIAM Series on Optimization, Society for Industrial & Applied Mathematics (SIAM), Philadelphia, PA, doi: 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, Springer, New York, doi: 10.1007/978-3-319-48311-5.

A. Beck & M. Teboulle (2009), A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM J. Imaging Sci.* 2(1), 183–202, doi: 10.1137/080716542.

H. Brezis (2010), *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer, New York, doi: 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.

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.

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.

C. Christof, C. Clason, C. Meyer & S. Walter (2018), Optimal control of a non-smooth semilinear elliptic equation, *Mathematical Control and Related Fields* 8(1), 247–276, doi: 10.3934/mcrf.2018011.

F. Clarke (2013), *Functional Analysis, Calculus of Variations and Optimal Control*, Springer, London, doi: 10.1007/978-1-4471-4820-3.

F. H. Clarke (1990), *Optimization and Nonsmooth Analysis*, vol. 5, Classics Appl. Math. SIAM, Philadelphia, PA, doi: 10.1137/1.9781611971309.

C. Clason (2020), *Introduction to Functional Analysis*, Compact Textbooks in Mathematics, Birkhäuser, Basel, doi: 10.1007/978-3-030-52784-6.

E. DiBenedetto (2002), *Real analysis*, Birkhäuser Boston, Inc., Boston, MA, doi: 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.

I. Ekeland & R. Témam (1999), *Convex Analysis and Variational Problems*, vol. 28, Classics Appl. Math. SIAM, Philadelphia, PA, doi: 10.1137/1.9781611971088.

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.

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.

A. D. Ioffe (2017), *Variational Analysis of Regular Mappings: Theory and Applications*, Springer Monographs in Mathematics, Springer International Publishing, doi: 10.1007/978-3-319-64277-2.

K. Ito & K. Kunisch (2008), *Lagrange Multiplier Approach to Variational Problems and Applications*, vol. 15, Advances in Design and Control, SIAM, Philadelphia, PA, doi: 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.

B. S. Mordukhovich (2006), *Variational Analysis and Generalized Differentiation I, Basic Theory*, vol. 330, Grundlehren der mathematischen Wissenschaften, Springer, doi: 10.1007/3-540-31247-1.

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.

N. Parikh & S. Boyd (2014), Proximal algorithms, *Foundations and Trends in Optimization* 1(3), 123–231, doi: 10.1561/2400000003.

J.-P. Penot (2013), *Calculus Without Derivatives*, vol. 266, Graduate Texts in Mathematics, Springer, New York, doi: 10.1007/978-1-4614-4538-8.

W. Rudin (1991), *Functional Analysis*, 2nd ed., McGraw-Hill, New York.

A. Ruszczyǹski (2006), *Nonlinear Optimization*, Princeton University Press, Princeton, NJ.

A. Schiela (2008), A simplified approach to semismooth Newton methods in function space, *SIAM J. Opt.* 19(3), 1417–1432, doi: 10.1137/060674375.

W. Schirotzek (2007), *Nonsmooth Analysis*, Universitext, Springer, Berlin, doi: 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.

M. Ulbrich (2002), Semismooth Newton methods for operator equations in function spaces, *SIAM J. Optim.* 13(3), 805–842 (2003), doi: 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, SIAM, Philadelphia, PA, DOI: 10.1137/1.9781611970692.

T. Valkonen (2020), Testing and non-linear preconditioning of the proximal point method, *Applied Mathematics & Optimization* 82(2), 591–636, DOI: 10.1007/s00245-018-9541-6.

K. Yosida (1995), *Functional Analysis*, Classics in Mathematics, Reprint of the sixth (1980) edition, Springer-Verlag, Berlin, DOI: 10.1007/978-3-642-61859-8.