A "crash course" in nonsmooth convex optimization

Zev Woodstock
woodstock(at)zib.de*

1 Introduction

These notes are supplementary material to a "crash course" I am teaching in May of 2023. The topic is *proximity operators and nonsmooth convex optimization*. These notes are not meant to be used as a standalone resource. Please cite peer-reviewed material. As a general reference text, I suggest *Convex Analysis and Monotone Operator Theory*, 2nd ed., by Bauschke and Combettes, published by Springer. Virtually all of the results in these notes also apply to real Hilbert spaces; for proofs in full-generality, read the book. If unspecified, $\mathcal{H}$ is a real finite-dimensional vector space (e.g., $\mathbb{R}^n$ is fine).

1.1 Optimization terminology and the extended real line

Notation 1.1 We will work with the **extended real line**, i.e., $[-\infty, +\infty] := \mathbb{R} \cup \{-\infty, +\infty\}$. Algebra on this field follows most "natural" rules one could expect (e.g., for $x \in \mathbb{R}$, $x + \infty = \infty$). However, the following quantities are **undefined**:

- Any subtraction of infinities: "$+\infty - (+\infty)$"
- Zero times infinity: "$0 \cdot (\pm\infty)$"
- Any quotient of infinities: "$\pm\infty / \pm\infty, \pm\infty / \mp\infty, \ldots$"

As a result, if we work with extended-real-valued functions, we must be sure to avoid anything which is undefined (e.g., the objective function $f(x) + g(x)$ could be undefined if there exists z such that $g(z) = -\infty$ and $f(z) = \infty$.)

*Please report typos/errors found in these notes. Homework solutions should be handed in to my office ZIB 3107.

Definition 1.2 Given a real vector space $\mathcal{H}$, a function $f\colon \mathcal{H} \to [-\infty, +\infty]$, and a set $C \subset \mathcal{H}$, consider the following optimization problem.

$$\underset{x \in C}{\text{minimize}} \quad f(x) \tag{1}$$

We call f the **objective function**. We call C a **constraint**. For any $x \in C$, we say x is **feasible**. Otherwise, for $x \in \mathbb{R}^n \setminus C$, x is infeasible. If a point $x^* \in C$ satisfies

$$(\forall x \in C) \quad f(x^*) \leq f(x), \tag{2}$$

we call x^* a **solution** to the optimization problem (1).

For this class, we consider minimization; to maximize f, just use the objective function $-f$.

Definition 1.3 For $I \subset [-\infty, +\infty]$, $a \in [-\infty, +\infty]$ is a **lower bound** (**upper bound**) if, for every $\xi \in I$, $a \leq \xi$ ($a \geq \xi$). The **greatest lower bound**, or **infimum**, of the set I is denoted $\inf I$. Analogously, the **least upper bound**, or **supremum**, of the set I is denoted $\sup I$. In general, $\inf I, \sup I \in [-\infty, +\infty]$. If, additionally, $\inf I \in I$ ($\sup I \in I$), we call it the **minimum** (**maximum**), and denote it $\min I$ ($\max I$). In these cases, we say the infimum (supremum) is *attained*.

A few things to mention:

(i) For $I \neq \varnothing$, we have $\inf I \leq \sup I$. For the empty set, $\inf \varnothing = +\infty$ and $\sup \varnothing = -\infty$.

(ii) While the $\inf$ and $\sup$ are always defined, $\max$ and $\min$ may not exist (e.g., consider $I = (0, 1)$ has $\inf I = 0$ and $\sup I = 1$. However, since $0, 1 \notin I$, neither $\max I$ nor $\min I$ exist.)

(iii) Let $f\colon \mathbb{R}^n \to [-\infty, +\infty]$. We adopt the notation that $\inf_{x \in C} f(x) = \inf\{f(x) \mid x \in C\}$.

(iv) It is common in optimization literature to abuse notation, and use

$$\min_{x \in C} f(x) \tag{3}$$

to describe the optimization problem (1). Technically, $\min_{x \in C} f(x)$ is not an optimization problem – it is the optimal value of the objective function at a solution, which may or may not exist.[1]

Definition 1.4 Let $f\colon \mathcal{H} \to [-\infty, +\infty]$. We will use the following terms.

(i) The **domain** of f is

$$\operatorname{dom} f = \left\{x \in \mathcal{H} \;\middle|\; f(x) < +\infty\right\} \tag{4}$$

[1]The *Weierstraß Theorem*, loosely stated, guarantees that a solution to (1) exists is if C is compact and f is lower-semicontinuous. For unconstrained functions, analytic notions of "coercivity" and "recession cones" can also yield existence results; however, they are not included in this class.

(ii) The **epigraph** of f is

$$\operatorname{epi} f = \big\{(x,\xi) \in \mathcal{H} \times \mathbb{R} \;\big|\; f(x) \leq \xi\big\} \tag{5}$$

(iii) The function f is **proper** if $\operatorname{dom} f \neq \varnothing$ and it never outputs the value $-\infty$ (i.e., $-\infty \notin f(\mathcal{H})$).

(iv) The function f is **lower semicontinuous** (sometimes abbreviated "lsc") at $x \in \mathcal{H}$ if, for every sequence $(x_n)_{n\in\mathbb{N}}$ satisfying $x_n \to x$, we have $f(x) \leq \liminf f(x_n)$

For this class, we will predominantly consider proper and lsc functions. A few things to note about the lsc assumption: (1) every continuous function is lsc, and (2) lower semicontinuity basically allows for a jump-discontinuity to occur at $x \in \mathcal{H}$, but requires that f takes the lowest possible limiting value at x (cf. the figures drawn in class, or here).

1.2 Inner product and norms

Definition 1.5 Let $\mathcal{H}$ be a real finite-dimensional vector space. A **scalar product** (sometimes called **inner product**) is a function $\langle \cdot \mid \cdot \rangle \colon \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ which satisfies the following properties.

(i) $(\forall x \in \mathcal{H} \setminus \{0\}) \quad \langle x \mid x \rangle > 0$

(ii) $(\forall x, y \in \mathcal{H}) \quad \langle x \mid y \rangle = \langle y \mid x \rangle$

(iii) $(\forall x, y, z \in \mathcal{H})(\forall \alpha \in \mathbb{R}) \quad \langle \alpha x + y \mid z \rangle = \alpha \langle x \mid z \rangle + \langle y \mid z \rangle$

Exercise 1.6 Let $\mathbf{0} \in \mathcal{H}$ be the zero element of $\mathcal{H}$. Show that, for every $x \in \mathcal{H}$, $\langle \mathbf{0} \mid x \rangle = 0$.

Exercise 1.7 Consider $\mathcal{H} = \mathbb{R}^n$. For two vectors $x, y \in \mathbb{R}^n$, the *dot product* is given by $\langle x \mid y \rangle = x^\top y$. Show that the dot product on $\mathbb{R}^n$ is a scalar product.

Exercise 1.8 Consider the vector space of matrices $\mathbb{R}^{n\times n}$. For two matrices $A = (a_{i,j})_{1\leq i,j\leq n}$ and $B = (b_{i,j})_{1\leq i,j\leq n}$, the *Frobenius inner product* is given by

$$\langle A \mid B \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{i,j} b_{i,j} \tag{6}$$

Show (6) is an inner product.

Proposition 1.9 (Cauchy-Schwarz) *For every $x, y \in \mathcal{H}$,*

$$\langle x \mid y \rangle^2 \leq \langle x \mid x \rangle \langle y \mid y \rangle. \tag{7}$$

Proof. If $y = 0$, (7) holds. Now suppose that $y \neq 0$. By Definition 1.5, $\langle y \mid y \rangle > 0$. Set $\alpha = \langle x \mid y \rangle / \langle y \mid y \rangle$. First, we find

$$\begin{aligned} 0 &\leq \langle x - \alpha y \mid x - \alpha y \rangle && (8)\\ &= \langle x \mid x \rangle - 2\alpha \langle x \mid y \rangle + \alpha^2 \langle y \mid y \rangle && (9)\\ &= \langle x \mid x \rangle - 2\alpha \langle x \mid y \rangle + \alpha \langle x \mid y \rangle && (10)\\ &= \langle x \mid x \rangle - \alpha \langle x \mid y \rangle. && (11) \end{aligned}$$

Rearranging the inequality, we find that

$$\begin{aligned} &\frac{\langle x \mid y \rangle^2}{\langle y \mid y \rangle} = \alpha \langle x \mid y \rangle \leq \langle x \mid x \rangle && (12)\\ \Leftrightarrow\quad &\langle x \mid y \rangle^2 \leq \langle y \mid y \rangle \langle x \mid x \rangle. && (13) \end{aligned}$$

□

Definition 1.10 Let $\mathcal{H}$ be a real finite-dimensional vector space. A function $\|\cdot\| \colon \mathcal{H} \to \mathbb{R}$ is a **norm** if the following hold.

(i) $(\forall x \in \mathcal{H}) \quad \|x\| = 0 \Rightarrow x = 0$

(ii) $(\forall x, y \in \mathcal{H}) \quad \|x + y\| \leq \|x\| + \|y\|$

(iii) $(\forall x \in \mathcal{H})(\forall \alpha \in \mathcal{H}) \quad \|\alpha x\| = |\alpha| \|x\|$

A norm is a way to measure magnitude of vectors, or the distance from one vector to another $\|x - y\|$.

Exercise 1.11 Let $\mathcal{H}$ be a real finite-dimensional vector space, and let $\langle \cdot \mid \cdot \rangle$ be a scalar product on $\mathcal{H}$. Show that the norm defined by

$$\|\cdot\| \colon \mathcal{H} \to \mathbb{R} \colon x \mapsto \sqrt{\langle x \mid x \rangle} \tag{14}$$

satisfies the properties in Definition 1.10.

The **Euclidean norm** on $\mathbb{R}^n$, given by $(\xi_1, \ldots, \xi_n) \mapsto \sqrt{\xi_1^2 + \cdots + \xi_n^2}$, arises from the dot product. Exercise 1.11 yields the following formulation of the Cauchy-Schwarz inequality

$$(\forall x, y \in \mathcal{H}) \quad \langle x \mid y \rangle \leq \|x\| \|y\|. \tag{C-S}$$

Exercise 1.12 Let $(x_1, x_2, x_3) \in \mathbb{R}^3$. Show that

$$2x_1 - x_2^4 + 6x_3 \leq 4\sqrt{x_1^2 + x_2^8 + 9x_3^2}. \tag{15}$$

Can the coefficient 4 in (15) be reduced?

The following theorem is referenced a few times in the notes, so I will provide its statement here. Regretfully, this class does not have enough time to detail the topics of compact/closed/lsc. The following theorem is often used as a tool to ensure that a solution to an optimization problem exists.

Theorem 1.13 (Weierstraß) *Let $f\colon \mathcal{H} \to [-\infty, +\infty]$ be lower semicontinuous and let C be a compact subset of $\mathcal{H}$. Suppose that $C \cap \operatorname{dom} f \neq \varnothing$. Then f achieves its infimum over C.*

2 Convexity

Definition 2.1 A set $C \subset \mathcal{H}$ is **convex** if, for every $x, y \in C$

$$(\forall \alpha \in \,]0,1[) \quad \alpha x + (1-\alpha)y \in C. \tag{16}$$

A function f is **convex** if $\operatorname{epi} f$ is convex.

Proposition 2.2 *$f\colon \mathcal{H} \to [-\infty, +\infty]$ is convex if and only if*

$$(\forall x, y \in \operatorname{dom} f) \quad (\forall \alpha \in \,]0,1[) \qquad f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y). \tag{17}$$

Proof. First, we note that if f is identically $+\infty$, then $\operatorname{dom} f = \varnothing$ if and only if $\operatorname{epi} f = \varnothing$, so (17) is vacuously true. Now assume that $\operatorname{dom} f \neq \varnothing$. Let (x, ξ) and (y, η) be in $\operatorname{epi} f$ and let $\alpha \in \,]0,1[$.
($\Rightarrow$) Assume that $\operatorname{epi} f$ is convex. Then

$$\alpha(x, \xi) + (1-\alpha)(y, \eta) = (\alpha x + (1-\alpha)y, \alpha\xi + (1-\alpha)\eta) \in \operatorname{epi} f. \tag{18}$$

Therefore, $f(\alpha x + (1-\alpha)y) \leq \alpha\xi + (1-\alpha\eta$. Taking the limit as $\xi \searrow f(x)$ and $\eta \searrow f(y)$ yields (17).
($\Leftarrow$) Assume that (17) holds. By definition, $f(x) \leq \xi$ and $f(y) \leq \eta$. So, by (17),

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y) \tag{19}$$
$$\leq \alpha\xi + (1-\alpha)\eta. \tag{20}$$

Therefore, $(\alpha x + (1-\alpha)y, \alpha\xi + (1-\alpha)\eta) \in \operatorname{epi} f$ which completes the proof. □

Definition 2.3 Let $\rho > 0$ and let $x \in \mathcal{H}$. A **closed ball** of radius ρ is $B(x;\rho) = \big\{z \in \mathcal{H} \,\big|\, \|x - z\| \leq \rho\big\}$.

Definition 2.4 Let $f\colon \mathcal{H} \to [-\infty, +\infty]$ and let $x \in \mathcal{H}$. x is a **local minimizer** of f if there exists $\rho > 0$ such that

$$(\forall z \in \mathcal{H} \cap B(x;\rho)) \quad f(x) \leq f(z). \tag{21}$$

x is a **global minimizer** of f if

$$(\forall z \in \mathcal{H}) \quad f(x) \leq f(z). \tag{22}$$

Fact 2.5 Let f be a convex and proper function. Then every local minimizer is a global minimizer.

Proof. This is left as an exercise (easier to prove after we learn about convex subdifferentials). □

Definition 2.6 Let $C \subset \mathcal{H}$ be nonempty.

(i) The **indicator function** of C is

$$\iota_C \colon \mathcal{H} \to [-\infty, +\infty] : x \mapsto \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{if } x \notin C. \end{cases} \tag{23}$$

(ii) Suppose that C is also closed. A **projection** of $x \in \mathcal{H}$ onto C is a solution to the minimization problem

$$\underset{z \in C}{\text{minimize}} \ \ \|x - z\|. \tag{24}$$

A solution to (24) is a "closest" point to x which resides in C.

Fact 2.7 Let $C \subset \mathcal{H}$ and let $x \in \mathcal{H}$.

(i) Without loss of generality, constrained optimization can be rephrased as unconstrained optimization via changing the objective function:

$$\inf_{x \in C} f(x) = \inf_{x \in \mathcal{H}} f(x) + \iota_C(x). \tag{25}$$

The objective function $f + \iota_C$ on the righthand side, although a bit fancier, allows us to rephrase the constraint on the lefthand side.

(ii) C is convex if and only if its indicator function ι_C is convex.

(iii) C is closed if and only if its indicator function ι_C is lsc.

(iv) Suppose that C is closed. Then a solution to (24) exists.

(v) Suppose that C is convex. If a solution to (24) exists, it is guaranteed to be unique.

The proofs of (ii) and (iii) follow from the fact that $\operatorname{epi} C = C \times [0, +\infty[$. Loosely speaking, the proof of (iv) follows from the Weierstraß theorem (compactness is achieved by intersecting C with $\{y \in \mathcal{H} \mid \|x - y\| \leq \eta\}$ for $\eta > 0$) and (v) follows from the fact that the norm is *strictly convex* – (a notion we have not yet defined, but the interested student could research).

Definition 2.8 Let $C \subset \mathcal{H}$ be nonempty, closed, and convex. In view of Fact 2.7(iv)–(v), for every $x \in \mathcal{H}$ there is a unique point, $\operatorname{Proj}_C(x) \in \mathcal{H}$, which solves (24). This implicitly defines the **projection operator** of C.

$$\operatorname{Proj}_C \colon \mathcal{H} \to \mathcal{H} \colon x \mapsto \operatorname{Proj}_C(x) \qquad \text{(solution to (24))} \tag{26}$$

Note: if $x \in C$, then $\operatorname{Proj}_C x = x$.

For all of the algorithms in this course, we will focus on functions from the following class

$$\Gamma_0(\mathcal{H}) = \big\{ f \colon \mathcal{H} \to \left]-\infty, +\infty\right] \;\big|\; f \text{ is proper, lower semicontinuous, and convex} \big\}. \tag{27}$$

The following functions live in $\Gamma_0(\mathcal{H})$:

(i) Exponentials: e^x

(ii) Log-barriers $f(x) = \begin{cases} -\ln(x) & \text{if } x > 0 \\ +\infty & \text{otherwise.} \end{cases}$

(iii) Any norm: $\|\cdot\|$ (e.g., $\|\cdot\|_1$ which promotes sparsity, $\|\cdot\|_{\text{nuclear}}$ which promotes low-rank)

(iv) Hinge-Loss, ReLU, KL-Divergence, ...

(v) Given a collection of functions $(f_i)_{i \in I}$ in $\Gamma_0(\mathcal{H})$, we can remain in $\Gamma_0(\mathcal{H})$ via the following operations.

   (a) $\max\{f_1, \ldots, f_m\}$

   (b) Positive linear combinations: $\lambda_1 f_1 + \cdots + \lambda_m f_m$, where $\{\lambda_i\}_{i=1}^m$ are positive.

   (c) Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two finite-dimensional real vector spaces. Let $b \in \mathcal{H}_2$ and let $A \colon \mathcal{H}_1 \to \mathcal{H}_2$ be a linear operator (e.g., a matrix from $\mathbb{R}^n$ to $\mathbb{R}^m$). If $f_1 \in \Gamma_0(\mathcal{H}_2)$, then $g(x) = f_1(Ax + b) \in \Gamma_0(\mathcal{H}_1)$.

Exercise 2.9 The **Minkowski sum** of two subsets A, B of $\mathcal{H}$ is given by

$$A + B = \big\{ a + b \;\big|\; a \in A \text{ and } b \in B \big\}. \tag{28}$$

Assume that A and B are convex. Prove that $A + B$ is convex.

Exercise 2.10 Show that the norm $\|\cdot\|$ is convex using Definition 1.10.

3 What is Differentiability?

There are a lot of ML engineers who brush off the mathematical details of what it means for a function to be differentiable. Algorithmic differentiation (sometimes misleadingly-called "automatic" differentiation) is only guaranteed to work when certain theoretical conditions about the *existence* of a gradient hold. This part of the class is dedicated to explaining that differentiability is not a freebie.

To start our discussion on differentiability, we will begin with a few preliminaries from analysis.

Definition 3.1 Let $A\colon \mathcal{H}_1 \to \mathcal{H}_2$. Then A is **linear** if, for every $\alpha \in \mathbb{R}$ and every $x, y \in \mathcal{H}_1$,

$$A(\lambda x) = \lambda A(x) \quad \text{and} \quad A(x+y) = A(x) + A(y). \tag{29}$$

Theorem 3.2 (Riesz-Fréchet representation) *Let $A\colon \mathcal{H} \to \mathbb{R}$ be linear. Then there exists a unique vector $u \in \mathcal{H}$ such that, for every $x \in \mathcal{H}$, $A(x) = \langle u \mid x \rangle$.*

Although at first-glance it looks unrelated, Theorem 3.2 is a central notion for defining the gradient. A necessary (albeit insufficient) condition for the existence of a gradient is the existence of a directional derivative, defined below.

Definition 3.3 Let $f\colon \mathcal{H} \to \left]-\infty, +\infty\right]$ be proper. The **directional derivative** of f at $x \in \operatorname{dom} f$ in the direction $y \in \mathcal{H}$ is

$$f'(x; y) = \lim_{\alpha \searrow 0} \frac{f(x + \alpha y) - f(x)}{\alpha}. \tag{30}$$

From Definition 3.3, we point out a few things.

(i) The limit in (30) might not exist.

(ii) If f is convex, then $f'(x; y) \in [-\infty, +\infty]$.

(iii) Even if a directional derivative exists, it might not exist in $\mathbb{R}$ (since it could be $+\infty$ or $-\infty$).

Definition 3.4 Let $x \in \operatorname{dom} f$. If $f'(x; \cdot)$ is linear, we say f is **differentiable at** x. In this case, the unique vector provided by Theorem 3.2 is called the **gradient** of f at x and denoted $\nabla f(x)$.

$$f'(x; \cdot) = \lim_{\alpha \searrow 0} \frac{f(x + \alpha\, \cdot) - f(x)}{\alpha} = \langle \nabla f(x) \mid \cdot \rangle \tag{31}$$

If f is differentiable at every $x \in \operatorname{dom} f$, we say that f is **differentiable**.

Exercise 3.5 Verify that $\nabla(\frac{1}{2}\|\cdot\|^2)(x) = x$.

All of the properties we know and love about differentiability (chain rule, product rule, etc.) have to be proven. Here is an example below.

Proposition 3.6 *Let $A\colon \mathcal{H}_1 \to \mathcal{H}_2$ be a linear operator (with adjoint denoted A^*), let $b \in \mathcal{H}_2$, and let $f\colon \mathcal{H} \to \mathbb{R}$ be proper and differentiable. Set $g = f(Ax+b)$. Then g is differentiable and*

$$\nabla g = A^*(\nabla f(A\cdot +b)). \tag{32}$$

Proof. Since $\operatorname{dom} f = \mathcal{H}_2$, $\operatorname{dom} g \neq \varnothing$ so we let $x \in \operatorname{dom} g$. By definition,

$$\begin{aligned}
g'(x;y) &= \lim_{\alpha\searrow 0} \frac{g(x+\alpha y) - g(x)}{\alpha} && (33)\\
&= \lim_{\alpha\searrow 0} \frac{f(A(x+\alpha y)+b) - f(Ax+b)}{\alpha} && (34)\\
&= \lim_{\alpha\searrow 0} \frac{f(Ax+b+\alpha Ay) - f(Ax+b)}{\alpha} && (35)\\
&= f'(Ax+b;Ay). && (36)
\end{aligned}$$

So the directional derivative of g exists. Now, since f is differentiable,

$$g'(x;y) = f'(Ax+b;Ay) = \langle \nabla f(Ax+b) \mid Ay\rangle = \langle A^*(\nabla f(Ax+b)) \mid y\rangle. \tag{37}$$

Hence the directional derivative of g is linear and g is differentiable. The specific form of the gradient is constructed in (37) □

Algorithmic differentiation tools use results like Proposition 3.6 to approximate a gradient of a function by reading its machine code. However, these subroutines do not check the theoretical conditions required for their theorems (e.g., *f must be differentiable*) – this must be done (and is oftentimes unjustly ignored) by the user.

Definition 3.7 Let f be proper and differentiable. f is **smooth** ("L-smooth") if there exists $L > 0$ such that

$$(\forall x,y \in \mathcal{H}) \quad \|\nabla f(x) - \nabla f(y) \leq L\|x-y\|. \tag{38}$$

Exercise 3.8 Construct a function which is differentiable and nonsmooth.

Proposition 3.9 *Let $f\colon \mathcal{H} \to \left]-\infty,+\infty\right]$ be proper and convex. Then,*

$$(\forall x \in \operatorname{dom} f)(\forall y \in \mathcal{H}) \qquad f'(x;y-x) + f(x) \leq f(y). \tag{39}$$

Proof. By Proposition 2.2, for every $\alpha \in \left]0,1\right[$,

$$\begin{aligned}
f(x+\alpha(y-x)) - f(x) &= f\left((1-\alpha)x + \alpha y\right) - f(x) && (40)\\
&\leq (1-\alpha)f(x) + \alpha f(y) - f(x) && (41)\\
&= \alpha(f(y) - f(x)). && (42)
\end{aligned}$$

Therefore,

$$\frac{f(x+\alpha(y-x)) - f(x)}{\alpha} \leq f(y) - f(x). \tag{43}$$

Taking the limit as $\alpha \searrow 0$ implies $f'(x;y) \leq f(y) - f(x)$, which in turn yields (39). □

Corollary 3.10 *Let $f\colon \mathcal{H} \to]-\infty, +\infty]$ be proper and convex. If f is differentiable at an interior point x of its domain* [2]*, then*

$$(\forall y \in \mathcal{H}) \quad \langle y - x \mid \nabla f(x) \rangle + f(x) \leq f(y). \tag{44}$$

When the lefthand side of (44) is viewed as a function of y, we see it is the first-order Taylor series approximation of f. Therefore, it follows from (39) that a convex differentiable function always remains above its first-order Taylor approximation! This is the motivating idea in defining a (convex) subgradient.

Definition 3.11 Let $f\colon \mathcal{H} \to]-\infty, +\infty]$. A vector g is a **subgradient** of f at $x \in \mathcal{H}$ if

$$(\forall y \in \mathcal{H}) \quad \langle y - x \mid g \rangle + f(x) \leq f(y). \tag{45}$$

The **subdifferential** of f at x is the set of all subgradients, denoted $\partial f(x)$.

This leads to the following fundamental theorem for optimization.

Theorem 3.12 (Fermat's Rule) *Let $f\colon \mathcal{H} \to]-\infty, +\infty]$ be proper. Then x is a minimizer of f if and only if $0 \in \partial f(x)$.*

Proof. By definition,

$$\begin{aligned} 0 \in \partial f(x) \Leftrightarrow (\forall y \in \mathcal{H}) &\qquad \langle 0 \mid y - x \rangle + f(x) \leq f(y) & (46)\\ \Leftrightarrow (\forall y \in \mathcal{H}) &\qquad f(x) \leq f(y). & (47) \end{aligned}$$

□

Unlike differentiable functions, there are technical conditions we must check in order to get the "standard" rules one would hope for.

Theorem 3.13 (Sum rule) *Let $f, g \in \Gamma_0(\mathcal{H})$ and suppose that one of the following holds:*

(i) *The interior of* $\operatorname{dom} g$ *intersects with* $\operatorname{dom} f$

(ii) $\operatorname{dom} g = \mathcal{H}$

(iii) *The relative interiors of* $\operatorname{dom} f$ *and* $\operatorname{dom} g$ *intersect.*

Then $\partial(f + g) = \partial f + \partial g$.

Note: If f is convex and differentiable at $x \in \mathcal{H}$, then $\partial f(x) = \{\nabla f(x)\}$.

[2] the *interior* of the domain