Advanced Topics in Machine Learning: Introduction to Convex Optimization

S. Salzo and M. Pontil

Overview: Convex optimization plays a key role in data sciences. The objective of this course is to provide basic tools and methods at the core of modern nonlinear convex optimization. Starting from the gradient descent method we will cover some state of the art algorithms, including proximal gradient methods, accelerated methods, stochastic subgradient method and randomized block-coordinate descent methods, as well as dual methods, which are nowadays very popular techniques to solve machine learning and inverse problems.

Duration. 15 h, in 10 lectures + 4.5 h in 3 lab sessions.

Pre-requisites. Basics of linear algebra and calculus of several variables.

Final exam. Classwork around the end of the lectures.

Program

1. *Basics on convex sets and functions* (1.5). Definition of convex sets and cones. Operations on convex sets (intersection, affine transformations, closure, interior). Convex hull, polyhedra, polytopes and simplexes. Extended real-valued functions, their effective domain and epigraph. Proper functions. Convex functions. Jensen inequality. Strongly convex functions. Examples: norms on $\mathbf{R}^n$, KL-divergence. Operations preserving convexity (sum, upper envelope, composition with a linear operator, composition with an increasing convex function). Closed functions. The class Γ_0. Continuity of convex functions in the interior of their domain. Coercivity. Existence and uniqueness of (global) minimizers.

2. *Smooth optimization* (1.5 h). Characterization of convexity, strict convexity and strong convexity in terms of the gradient. Characterizations of Lipschitz smoothness. The descent lemma. Fermat's rule. Fixed-point iterations. Banach-Caccioppoli theorem. Gradient descent algorithm for strongly convex functions. Convergence analysis and rate of convergence. Example: gradient descent for least squares problems.

3. *Differential theory for nonsmooth convex functions* (1.5h). Directional derivatives, subdifferential and subgradients. Subdifferential of convex functions of one real variable. Subdifferentials of norms. Normal cones. Subdifferential of a max of convex functions. Subdifferential of separable convex functions. Fermat's rule (nonsmooth case). Subdifferential calculus. The projected subgradient method.

4. *Duality theory I* (1.5h). Orthogonal projection onto closed convex sets. Definition and variational characterization. Affine hyperplanes and halfspaces. A separation theorem. Dual representation of convex sets. The Legendre-Fenchel transform and its properties. Geometrical meaning of the Fenchel conjugate. Examples: support functions and norms. Young-Fenchel inequality and duality in subgradients. Biconjugate theorem (of Fenchel-Moreau): representing functions in the dual form. Properties in duality: strict convexity and differentiability, and strong convexity and Lipschitz smoothness.

5. *The proximity operator* (45 min). Basic properties. Examples: orthogonal projection and soft-thresholding operator. Calculus with proximity operators. Moreau decomposition formula. Proximity operator of Group-Lasso regularization.

6. *The proximal gradient algorithm* (45 min). A general algorithm for composite minimization. Convergence of the iterates and rate of convergence in objective values. Example: solving the Lasso problem via iterative soft-thresholding algorithm.

7. *Accelerated proximal gradient algorithms* (45 min). Two acceleration techniques: Nesterov acceleration and heavy ball (momentum) method. Example: solving the Lasso problem via fast iterative soft-thresholding algorithm (FISTA). Rate of convergence.

8. *Elements of sparse estimation/recovery* (if time permit). Statistical properties of the least square estimator. The statistical analysis of the Lasso estimator in high dimension. Algorithms for Lasso, Elastic Net, and Group Lasso problems.

9. *Convex spectral functions* (45 min). Conjugacy formula. Subdifferential formula. Nuclear norm regularization. The proximal gradient method for matrix recovery.

10. *Stochastic optimization algorithms* (1.5h). The projected stochastic subgradient algorithm. Finite horizon and infininite horizon convergence rates. Examples: stochastic optimization and stochastic incremental methods. Randomized block-coordinate proximal gradient algorithm. Convergence theorem.

11. *Duality Theory II* (45 min). Definition of dual problem (in the sense of Fenchel-Rockafellar). Connection with Lagrange duality. Weak and strong duality. Duality gap. KKT conditions. Qualification conditions. Calculus rules for subdifferentials (general formula). Examples. Dealing with equality linear constraints. Optimal transport problem.

12. *Dual algorithms* (45 min) A general framework for dual algorithms. Relating dual and primal variables and dual and primal objective functions. Examples: dual proximal gradient algorithm, dual accelerated proximal gradient algorithm. Stochastic dual ascent methods.

13. *Applications in machine learning* (1.5 h). Statistical learning as a stochastic optimization problem. A worst-case statistical bound. Regularized empirical risk minimization (ERM): kernel methods, support vector machines, logistic regression. A dual algorithmic framework for ERM.

The 3 labs will cover: the gradient descent method in action, problems with sparsity constraints, stochastic gradient descent, and randomized coordinate methods.

References

[1] R. T. Boyd L. Vandenberghe, *Convex Optimization.* Cambridge University Press, Cambridge, 2004.

[2] J.B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis.* Springer, New York, 2004.

[3] Y. Nesterov, *Introductory Lectures on Convex Optimizations.* Kluwer, Dordrecht, 2004.

Lecture 1

Convex sets and functions

1.1 Basic notations

We set $\mathbb{R}_+ = \{\alpha \in \mathbb{R} \,|\, \alpha \geq 0\}$ and $\mathbb{R}_{++} = \{\alpha \in \mathbb{R} \,|\, \alpha > 0\}$. In these notes we consider *Euclidean spaces*, that is, finite dimensional real vector spaces endowed with a scalar product. We denote by X an Euclidean space and its associated *scalar product* and *norm* will be denoted by

$$\langle \cdot, \cdot \rangle \colon X \times X \to \mathbb{R}, \qquad \|\cdot\| : X \to \mathbb{R}.$$

Examples of Euclidean spaces are the space $\mathbb{R}^n$ endowed with Euclidean scalar product and norm

$$(\forall\, x, y \in \mathbb{R}^n) \qquad \langle x, y \rangle = x^\top y = \sum_{i=1}^n x_i y_i, \qquad \|x\|^2 = \sum_{i=1}^n x_i^2. \tag{1.1}$$

and the space of square symmetric matrices $\mathbb{S}^n \subset \mathbb{R}^{n \times n}$ endowed with the trace scalar product and Frobenius norm

$$(\forall\, x, y \in \mathbb{S}^n) \qquad \langle x, y \rangle = \operatorname{trace}(x^\top y) = \sum_{i=1,j=1}^n x_{i,j} y_{i,j}, \qquad \|x\|^2 = \sum_{i=1,j=1}^n x_{i,j}^2. \tag{1.2}$$

The scalar product and the norm satisfy the following properties:

(i) $(\forall\, x, y, z \in X)(\forall\, \alpha, \beta \in \mathbb{R})\ \langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$ (linearity)

(ii) $(\forall\, x, y \in X)\ \langle x, y \rangle = \langle y, x \rangle$ (symmetry)

(iii) $(\forall\, x \in X)\ \|x\|^2 = \langle x, x \rangle \geq 0$ and $x \neq 0 \ \Rightarrow\ \|x\| > 0$;

(iv) $(\forall\, x \in X)(\forall\, \lambda \in \mathbb{R})\, \|\lambda x\| = |\lambda|\, \|x\|$ (homogeinity)

(v) $(\forall\, x, y \in X)\, \|x + y\| \leq \|x\| + \|y\|$ (triangle inequality).

Salzo Saverio (saverio.salzo@iit.it) Istituto Italiano di Tecnologia. Via E. Melen, 16152, Genoa, Italy.

For every $x \in X$ and every $\delta > 0$ we denote by $B_\delta(x)$ the (closed) ball of center x and radius δ, that is $B_\delta(x) = \{y \in X \mid \|y - x\| \leq \delta\}$. A subset $C \subset X$ is said to be *open* if, for every $x \in C$ there exists $\delta > 0$ such that $B_\delta(x) \subset C$, and is said to be *closed* if its complement $X \setminus C$ is open. We recall that finite unions and possibly infinite intersections of closed sets are closed, while finite intersections and possibly infinite unions of open sets are open. The *interior* of C, which is denoted by $\operatorname{int}(C)$, is the largest open set that is contained in C. A point $x \in X$ belongs to $\operatorname{int}(C)$ if and only if there exists $\delta > 0$ such that $B_\delta(x) \subset C$. The *closure* of C, denoted by $\operatorname{cl}(C)$, is the smallest closed set that contains C. A point $x \in X$ belongs to $\operatorname{cl}(C)$ if and only if for every $\delta > 0$, $B_\delta(x) \cap C \neq \varnothing$. The *boundary* of C is $\operatorname{bdry}(C) = \operatorname{cl}(C) \setminus \operatorname{int}(C) = \operatorname{cl}(C) \cap \operatorname{cl}(X \setminus C)$.

Classically, in optimization, functions and constraints are treated separately. By introducing extended real-valued functions they can be treated in a unified way. Here with *extended real-valued functions*, we mean functions

$$f \colon X \to]-\infty, +\infty],$$

so that the value $-\infty$ will never be allowed. The *(effective) domain of* f is the set $\operatorname{dom} f := \{x \in X \mid f(x) < +\infty\}$ and the *epigraph of* f is the set

$$\boxed{\operatorname{epi}(f) = \{(x, t) \in X \times \mathbb{R} \mid f(x) \leq t\}.} \tag{1.3}$$

Note that $\operatorname{epi}(f) \subset X \times \mathbb{R}$. We also define the *sublevel sets* of f as

$$[f \leq t] := \big\{x \in X \mid f(x) \leq t\big\}, \qquad t \in \mathbb{R}, \tag{1.4}$$

and similarly we define the sets $[f > t]$. An extended real-valued function is called *proper* if $\operatorname{dom} f \neq \varnothing$, meaning that the function admits at least a finite value.

Extended real-valued functions allow to handle constraints, in optimization problems, as functions. Indeed let $C \subset X$ and define the *indicator function of* C as

$$\boxed{\iota_C \colon X \to]-\infty, +\infty] : x \mapsto \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{if } x \notin C. \end{cases}} \tag{1.5}$$

Then the constrained minimization problem

$$\min_{x \in C} h(x), \qquad h \colon X \to \mathbb{R}$$

can be equivalently written as

$$\min_{x \in X} f(x), \qquad f \colon X \to]-\infty, +\infty], \ f(x) = h(x) + \iota_C(x).$$

Remark 1.1.1. Indicator functions and epigraphs allow to establish a one to one correspondence between extended real-valued functions and sets. See Table 1.1.

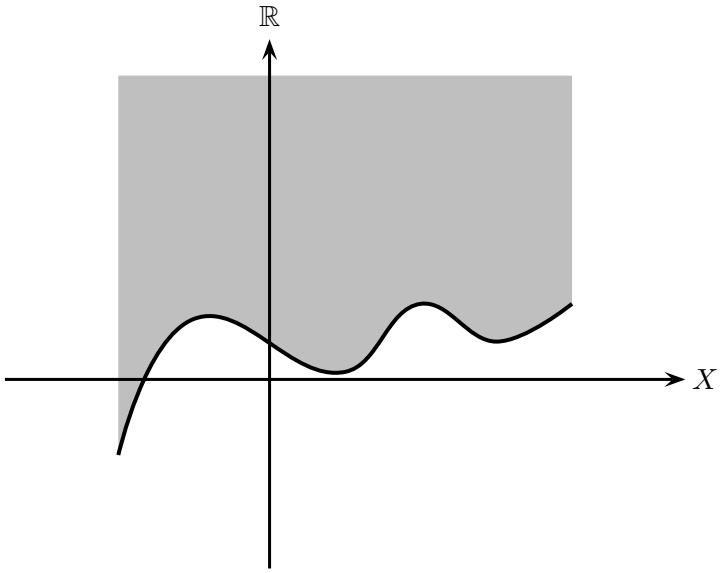


Figure 1.1: The epigraph of a function.

Sets	Functions
C	ι_C
$\mathrm{epi}(f)$	f

Table 1.1: Correspondence between sets and extended real-valued functions.

1.2 Convex sets

Definition 1.2.1. A subset $C \subset X$ is said to be *convex* if

$$(\forall\, x, y \in C)(\forall\, \lambda \in [0,1]) \quad (1-\lambda)x + \lambda y \in C, \tag{1.6}$$

meaning that for every $x, y \in C$, the *segment* $[x,y] = \big\{x + \lambda(y-x) \,\big|\, \lambda \in [0,1]\big\}$ is contained in C.

Definition 1.2.2. A *cone* of X is a subset $C \subset X$ such that

$$(\forall\, x \in C)(\forall\, \lambda \in \mathbb{R}_{++}) \quad \lambda x \in C, \tag{1.7}$$

meaning that, for every $x \in C$ the ray $\mathbb{R}_{++}x = \{\lambda x \,|\, \lambda \in \mathbb{R}_{++}\}$ is contained in C. A cone C is said to be *pointed* if $C \cap (-C) = \{0\}$.

Proposition 1.2.3 (operations on convex sets)**.** *Let X and Y be two Euclidean spaces.*

(i) *Let $A\colon X \to Y$ be a linear operator and let $b \in Y$. Then, for every convex set $C \subset X$, we have that $A(C) + b$ is a convex subset of Y and, for every convex set $D \subset Y$, we have that $A^{-1}(D)$ is a convex subset of X.*

(ii) *Let $(C_i)_{i\in I}$ be a family of convex subsets of X. Then $\bigcap_{i\in I} C_i$ is convex.*

(iii) *Let C be a convex subset of X. Then the closure* $\mathrm{cl}(C)$ *and the interior* $\mathrm{int}(C)$ *of C are convex subsets of X.*

Example 1.2.4. Here is a list of significant convex sets or cones that often occur as constrain sets in convex optimization.

- *Affine hyperplanes.* They are sets defined as solutions of a nontrivial linear equation, that is of type

$$H = \big\{x \in X \,\big|\, \langle u, x\rangle = \alpha\big\}, \quad (u, \alpha) \in (X \setminus \{0\}) \times \mathbb{R}. \tag{1.8}$$

- *Half-spaces.* They are sets defined as solutions of a nontrivial linear inequality, that is of type

$$H^- = \big\{x \in X \,\big|\, \langle u, x\rangle \leq \alpha\big\}, \quad (u, \alpha) \in (X \setminus \{0\}) \times \mathbb{R}. \tag{1.9}$$

- *Positive orthants.* When $X = \mathbb{R}^n$, the positive orthant and the strict positive orthant of $\mathbb{R}^n$

$$\mathbb{R}^n_+ = \big\{x \in \mathbb{R}^n \,|\, (\forall\, i = 1, \ldots, n)\ x_i \geq 0\big\},$$
$$\mathbb{R}^n_{++} = \big\{x \in \mathbb{R}^n \,|\, (\forall\, i = 1, \ldots, n)\ x_i > 0\big\}$$

  are cones.

- *Polyhedrals.* They are finite intersections of half-spaces, thus they are defined by means of a finite systems of linear inequalities and can be expressed as

$$C = \big\{x \in X \,\big|\, Ax \leq b\big\}, \tag{1.10}$$

  where $A\colon X \to \mathbb{R}^m$ is a linear operator, $b \in \mathbb{R}^m$, and "$\leq$" is the canonical ordering in $\mathbb{R}^m$, defined component-wise.

- *Polytopes.* They are bounded polyhedrals.

The intersection property (ii) of Proposition 1.2.3 justifies the following definition.

Definition 1.2.5. Let $A \subset X$. The intersection of the family of all convex subsets of X containing A (which is nonempty, since it contains X) is called the *convex hull of A* and it is denoted by $\mathrm{co}(A)$. In fact it is the smallest convex subset of X containing A.

Given k points $(x_i)_{1\leq i\leq k}$ in X, we call *convex combination of* $(x_i)_{1\leq i\leq k}$ any point $x = \sum_{i=1}^k \lambda_i x_i$, where $(\lambda_i)_{1\leq i\leq k} \in \mathbb{R}^k_+$ and $\sum_{i=1}^k \lambda_i = 1$. If $C \subset X$ is a convex set, then every convex combination of points of C belongs to C.

Proposition 1.2.6. *Let $(x_i)_{1\leq i\leq k}$ be k points in X. Then*

$$\mathrm{co}\big(\{x_1, \ldots, x_k\}\big) = \left\{x = \sum_{i=1}^k \lambda_i x_i \,\middle|\, (\lambda_i)_{1\leq i\leq k} \in \mathbb{R}^k_+, \sum_{i=1}^k \lambda_i = 1\right\}. \tag{1.11}$$

Proof. Hint: it is enough to prove that the set on the right hand side of (1.11) is convex. □

Definition 1.2.7. Let $k \in \mathbb{N}$, $k \geq 1$. A *k-simplex* is the convex hull of a $k+1$ affinely independent points $(x_i)_{0 \leq i \leq k}$, meaning that $(x_i - x_0)_{1 \leq i \leq k}$ are linear independent. The points x_i's are called *vertices* of the simplex.

Example 1.2.8.

(i) A 1-simplex is a segment $[x_0, x_1]$, a 2-simplex is a triangle, a 3-simplex is a tetrahedron.

(ii) When $X = \mathbb{R}^n$, with n an integer grater than 1, the standard simplex of $\mathbb{R}^n$ is $\Delta_n = \{x \in \mathbb{R}^n_+ \mid \sum_{i=1}^n x_i \leq 1\}$.

Proposition 1.2.9. *Polytopes are convex hulls of finite number of points.*

Lemma 1.2.10 (Accessibility)**.** *Let $C \subset X$ be a non empty convex set. Then*

$$x \in \operatorname{int}(C) \quad \textit{and} \quad y \in \operatorname{cl}(C) \implies [x, y[\subset \operatorname{int}(C).$$

As a consequence, if $\operatorname{int}(C) \neq \varnothing$*, then* $\operatorname{cl}(C) = \operatorname{cl}(\operatorname{int}(C))$ *and* $\operatorname{int}(C) = \operatorname{int}(\operatorname{cl}(C))$.

Proposition 1.2.11 (operations on convex sets – part 2)**.** *Let C be a convex subset of X. Then the closure* $\operatorname{cl}(C)$ *and the interior* $\operatorname{int}(C)$ *of C are convex subsets of X.*

Proof. The fact that $\operatorname{cl}(C)$ is convex is immediate. The convexity of $\operatorname{int}(C)$ follows from the accessibility Lemma 1.2.10. □

Remark 1.2.12. For a closed half-space H^- as defined in Example 1.2.4, we have

$$\operatorname{int}(H^-) = \big\{x \in X \,\big|\, \langle x, u\rangle < \alpha\big\}.$$

Indeed $x \in \operatorname{int}(H^-) \;\Leftrightarrow\; \exists \delta > 0$ such that $\sup_{v \in B_1(0)} \langle x + \delta v, u\rangle \leq \alpha \;\Leftrightarrow\; \exists \delta > 0$ such that $\langle x, u\rangle + \delta \|u\| \leq \alpha$.

1.3 Convex functions

Definition 1.3.1. An extended real-valued function $f\colon X \to]-\infty, +\infty]$ is *convex* if

$$(\forall\, x, y \in X)(\forall\, \lambda \in [0,1]) \quad f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y). \tag{1.12}$$

Moreover, f is *strictly convex* if in (1.12) the strict inequality holds when $x, y \in \operatorname{dom} f$, $x \neq y$ and $\lambda \in]0,1[$. Finally, $g\colon X \to [-\infty, +\infty[$ is *concave* (risp. *strictly concave*) if $-g$ is convex (risp. strictly convex).

Remark 1.3.2.

(i) We can actually get rid of infinite values expressing convexity as

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

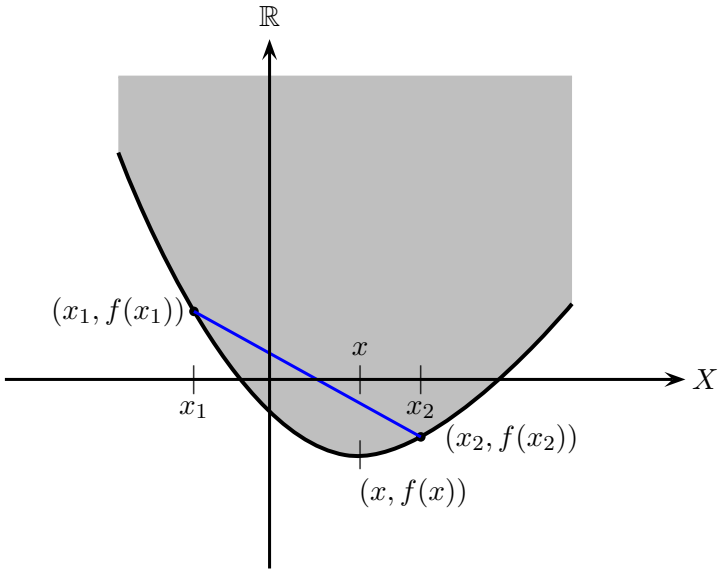


Figure 1.2: A convex function and its epigraph.

(ii) If f is convex, by induction Definition 1.3.1 yields *Jensen's inequality*, that is for every finite sequence $(x_i)_{1\leq i\leq m}$ in X and every $(\lambda_i)_{1\leq i\leq m} \in \mathbb{R}^n_+$ such that $\sum_{i=1}^m \lambda_i = 1$, we have

$$f\left(\sum_{i=1}^m \lambda_i x_i\right) \leq \sum_{i=1}^m \lambda_i f(x_i). \tag{1.13}$$

Example 1.3.3. $\|\cdot\|$ is convex and it not strictly convex, whereas $\|\cdot\|^2$ is strictly convex.

Proposition 1.3.4. *Let $f\colon X \to]-\infty, +\infty]$ be an extended real-valued function. Then f is convex if and only if* $\mathrm{epi}(f)$ *is convex. Moreover, if f is convex, then its sublevel sets are convex.*

Proof. First of all, we note that, by the definition of the epigraph (1.3) and Remark 1.3.2(i) the convexity of f is equivalent to

$$(\forall\, x, y \in \mathrm{dom} f)(\forall\, \lambda \in [0,1])\ (1-\lambda)(x, f(x)) + \lambda(y, f(y)) \in \mathrm{epi}(f). \tag{1.14}$$

Now suppose that f is convex. Let (x, s) and (y, t) be two points in $\mathrm{epi}(f)$. Then, since $f(x) \leq s$ and $f(y) \leq t$, we have, for every $\lambda \in [0, 1]$,

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y) \leq (1-\lambda)s + \lambda t,$$

hence $(1-\lambda)(x, s) + \lambda(y, t) \in \mathrm{epi}(f)$. Therefore $\mathrm{epi}(f)$ is convex. Vice versa, if $\mathrm{epi}(f)$ is convex, then (1.14) follows, and hence f is convex. The last part of the statement follows directly from the definition of convexity (1.12). □

Example 1.3.5. We provide a list of significant convex functions.

(i) Norms in $\mathbb{R}^n$: $\|x\|_p = \left(\sum_{i=1}^n x_i^p\right)^{1/p}$, $p \in [1, +\infty[$, and $\|x\|_\infty = \max_{1\leq i\leq n} |x_i|$.

(ii) Kullback–Leibler divergence. For every $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$, we define

$$f(x,y) = \begin{cases} \sum_{j=1}^n x_j \log(x_j/y_j) & \text{if } x \in \mathbb{R}^n_{++} \text{ and } y \in \mathbb{R}^n_{++}, \\ +\infty, & \text{otherwise.} \end{cases}$$

This function is (jointly) convex.

Proposition 1.3.6 (operations preserving convexity)**.** *The following hold.*

(i) *Let $f\colon X \to]-\infty, +\infty]$ and $g\colon X \to]-\infty, +\infty]$ be convex functions and let $\alpha \in \mathbb{R}_+$. Then $f + g\colon X \to]-\infty, +\infty]$ and $\alpha f\colon X \to]-\infty, +\infty]$ are convex.*

(ii) *Let $(f_i)_{i\in I}$ be a family of extended real-valued functions, $f_i\colon X \to]-\infty, +\infty]$. Then*

$$\sup_{i\in I} f_i\colon X \to]-\infty, +\infty], \quad (\forall\, x \in X)\ f(x) = \sup_{i\in I} f_i(x)$$

is convex.

(iii) *Let $f\colon X \to]-\infty, +\infty]$ be a convex function, let $A\colon Y \to X$ be a bounded linear operator, and let $\varphi\colon]-\infty, +\infty] \to]-\infty, +\infty]$ be an increasing convex function. Then $f \circ A\colon Y \to]-\infty, +\infty]$ and $\varphi \circ f\colon X \to]-\infty, +\infty]$ are convex.*

(iv) *Let $\varphi\colon X \times Y \to]-\infty, +\infty]$ be a convex function. Then, the* marginal function *$f\colon X \to [-\infty, +\infty]$, $f(x) = \inf_{y\in Y} \varphi(x, y)$, is convex*[1]

Proof. We prove (iv) only. Let $x_1, x_2 \in X$ with $f(x_1) < +\infty$ and $f(x_2) < +\infty$. Let $\xi_1 > f(x_1)$ and $\xi_2 > f(x_2)$. Then there exist $y_1, y_2 \in Y$ such that $\varphi(x_1, y_1) < \xi_1$ and $\varphi(x_2, y_2) < \xi_2$. Then, for every $\alpha \in]0, 1[$,

$$\begin{aligned} f((1-\alpha)x_1 + \alpha x_2) &\le \varphi((1-\alpha)x_1 + \alpha x_2, (1-\alpha)y_1 + \alpha y_2) \\ &\le (1-\alpha)\varphi(x_1, y_1) + \alpha\varphi(x_2, y_2) \\ &\le (1-\alpha)\xi_1 + \alpha\xi_2 \end{aligned}$$

Letting $\xi_1 \to f(x_1)$ and $\xi_2 \to f(x_2)$ we get $f((1-\alpha)x_1 + \alpha x_2) \le (1-\alpha)f(x_1) + \alpha f(x_2)$ and the statement follows. □

Definition 1.3.7. The function $f\colon X \to]-\infty, +\infty]$ is *closed* if $\operatorname{epi}(f)$ is a closed subset of $X \times \mathbb{R}$.

Proposition 1.3.8. *Let $f\colon X \to]-\infty, +\infty]$. Then the following are equivalent.*

(i) *For every sequence $(x_k)_{k\in\mathbb{N}}$ in X and every $x \in X$, $x_k \to x \ \Rightarrow\ f(x) \le \liminf_k f(x_k)$.*

(ii) *For every $\alpha \in \mathbb{R}$, the sublevel set $[f \le \alpha] := \{x \in X \mid f(x) \le \alpha\}$ is closed in X.*

(iii) *f is closed.*

[1]For functions taking values in $[-\infty, +\infty]$, this means that $\operatorname{epi}(f) = \{(x, t) \in X \times \mathbb{R} \mid f(x) \le t\}$ is convex, which is again equivalent to the inequality in Remark 1.3.2(i).

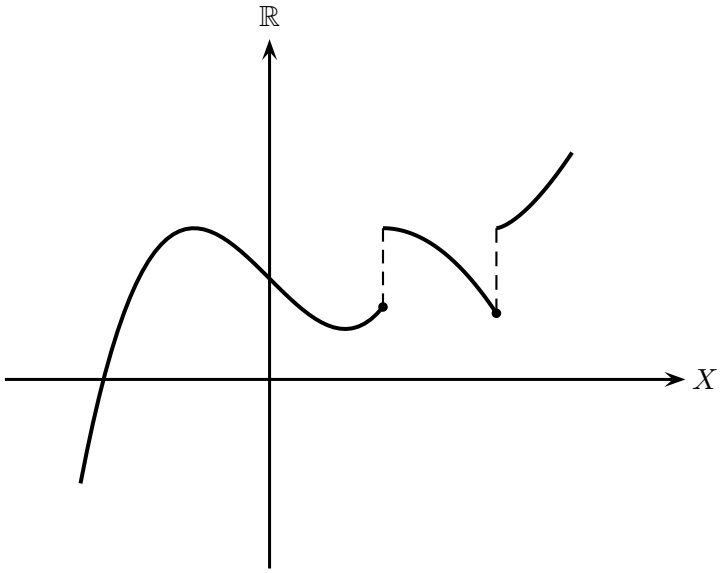


Figure 1.3: A closed (non-continuous) function.

Proof. (i) $\Leftrightarrow$ (ii): Suppose that $\operatorname{epi}(f)$ is closed in $X \times \mathbb{R}$ and let $\alpha \in \mathbb{R}$. Then, $\operatorname{epi}(f) \cap (X \times \{\alpha\}) = [f \leq \alpha] \times \{\alpha\}$ Therefore, $[f \leq \alpha]$ is closed in X. Vice versa suppose that, for every $\alpha \in \mathbb{R}$, $[f \leq \alpha]$ is closed, hence $[f > \alpha]$ is open. Then,

$$\begin{aligned}(\forall\,(x,\beta) \in X \times \mathbb{R}) \quad (x,\beta) \notin \operatorname{epi}(f) &\Leftrightarrow f(x) > \beta \\ &\Leftrightarrow (\exists \alpha \in \mathbb{R})\ f(x) > \alpha > \beta \\ &\Leftrightarrow (x,\beta) \in \bigcup_{\alpha \in \mathbb{R}} \big([f > \alpha] \times]-\infty, \alpha[\,\big).\end{aligned}$$

Therefore, the complement of $\operatorname{epi}(f)$ is open, since it is the union of open sets and hence $\operatorname{epi}(f)$ is closed.

(ii) $\Rightarrow$ (iii): Let $(x_k)_{k\in\mathbb{N}}$ be a sequence in X and $x \in X$ such that $x_k \to x$. Let $\alpha \in \mathbb{R}$ be such that $\alpha < f(x)$. We prove that necessarily $\liminf_k f(x_k) \geq \alpha$. Indeed if this was not the case, there would exist a subsequence $(x_{n_k})_{k\in\mathbb{N}}$ such that $x_{n_k} \in [f \leq \alpha]$ for all $k \in \mathbb{N}$. Since $x_{n_k} \to x$ and $[f \leq \alpha]$ is closed, we would have $x \in [f \leq \alpha]$ which gives a contradiction.

(iii) $\Rightarrow$ (ii): Let $(x_k)_{k\in\mathbb{N}}$ be a sequence in $[f \leq \alpha]$ and suppose that $x_k \to x$ for some $x \in X$. Then, for every $k \in \mathbb{N}$, $f(x_k) \leq \alpha$ and hence $f(x) \leq \liminf_k f(x_k) \leq \alpha$, that is, $x \in [f \leq \alpha]$. □

Definition 1.3.9. Let X be an Euclidean space. We denote by $\Gamma_0(X)$ the class of proper, convex, and closed extended real-valued functions defined on X.

Definition 1.3.10. Let $f\colon X \to]-\infty,+\infty]$. The function f is said to be *strongly convex* if there exists $\mu > 0$ such that, for every $x, y \in X$ and every $\lambda \in [0,1]$,

$$f\big((1-\lambda)x + \lambda y\big) \leq (1-\lambda)f(x) + \lambda f(y) - \frac{\mu}{2}(1-\lambda)\lambda \|x-y\|^2 . \tag{1.15}$$

In such case, μ is called the *modulus* of strong convexity of f and the function f is also said to be *μ-strongly convex*.

Lemma 1.3.11. *Let $x, y \in X$ and $\lambda \in \mathbb{R}$. Then*

$$\|(1-\lambda)x + \lambda y\|^2 = (1-\lambda)\|x\|^2 + \lambda\|y\|^2 - (1-\lambda)\lambda\|x-y\|^2 . \tag{1.16}$$

Proof. Indeed

$$\begin{aligned}\|(1-\lambda)x + \lambda y\|^2 &= (1-\lambda)^2\|x\|^2 + \lambda^2\|y\|^2 + 2(1-\lambda)\lambda\langle x, y\rangle \\ &= (1-\lambda)\|x\|^2 - \lambda(1-\lambda)\|x\|^2 \\ &\quad + \lambda\|y\|^2 - (1-\lambda)\lambda\|y\|^2 + 2(1-\lambda)\lambda\langle x, y\rangle \\ &= (1-\lambda)\|x\|^2 + \lambda\|y\|^2 - (1-\lambda)\lambda\big(\|x\|^2 + \|y\|^2 - 2\langle x, y\rangle\big)\end{aligned}$$

and the statement follows. □

The following result follows directly from (1.12) and Lemma 1.3.11.

Proposition 1.3.12. *Let $\mu > 0$. A function $f\colon X \to]-\infty,+\infty]$ is μ-strongly convex if and only if $f - (\mu/2)\|\cdot\|^2$ is convex.*

Proposition 1.3.13. *Let $\mu > 0$ and let $f\colon X \to]-\infty,+\infty]$ be a proper and μ-strongly convex function. Let $x^* \in X$ be a global minimizer of f. Then the following* quadratic growth *condition holds*

$$(\forall\, x \in X) \qquad f(x) - f(x^*) \geq \frac{\mu}{2}\|x - x^*\|^2 . \tag{1.17}$$

Proof. Let $x \in X$. It follows from Definition 1.3.10 that, for every $\lambda \in]0,1]$,

$$0 \leq \frac{f(x^* + \lambda(x-x^*)) - f(x^*)}{\lambda} \leq f(x) - f(x^*) - \frac{\mu}{2}(1-\lambda)\|x-x^*\|^2 . \tag{1.18}$$

Thus, we have

$$(\forall\, \lambda \in]0,1]) \quad \frac{\mu}{2}(1-\lambda)\|x - x_*\|^2 \leq f(x) - f(x_*). \tag{1.19}$$

Taking the supremum on the λ's on the left hand side of (1.19) we get the statement. □

Example 1.3.14. Let us consider the following generalization of the growth condition (1.17).

$$(\forall\, x \in X) \quad f(x) - f_* \geq \frac{\mu}{2}\mathrm{dist}(x, \operatorname{argmin} f)^2, \tag{1.20}$$

where $f_* = \inf_X f$. This condition can hold even for non-strongly convex functions. We provided a significant example. Let $A\colon X \to Y$ be a linear operator and let $b \in Y$. Set

$$f\colon X \to \mathbb{R} \quad f(x) = \frac{1}{2}\|Ax - b\|^2 . \tag{1.21}$$

Let y be the projection of b onto the range $R(A)$ of A. Then by Pytagoras' theorem, we have

$$(\forall\, x \in X) \quad f(x) = \frac{1}{2}\big(\,\|Ax - y\|^2 + \|y - b\|^2\,\big).$$

Thus, $f_* = \inf_X f = (1/2)\,\|y - b\|^2$. Now, let $x \in X$ and let $x_p = P_S x$, where $S = \operatorname{argmin} f = \big\{x \in X \,|\, Ax = y\big\}$, which is an affine set with direction $N(A)$. Therefore, since $y = Ax_p$, we have

$$f(x) - f_* = \frac{1}{2}\,\|Ax - y\|^2 = \frac{1}{2}\langle A^*A(x - x_p), x - x_p\rangle. \tag{1.22}$$

Moreover, since $x - x_p \in N(A)^\perp$, we have $\langle A^*A(x - x_p), x - x_p\rangle \geq \mu\,\|x - x_p\|^2 = \mu\,\operatorname{dist}(x, \operatorname{argmin} f)^2$, where μ is the minimum non-zero eigenvalue of A^*A.

1.4 Continuity of convex functions

In this section we prove that convex functions are always locally Lipschitz continuous (and hence continuous) on the interior of their domain (this occurs, without any additional condition, only in finite dimension).

Lemma 1.4.1. *Let $x \in \operatorname{dom} f$, let $\delta > 0$, and suppose that*

$$\sup_{y \in B_{2\delta}(x)} f(y) - f(x) \leq M < +\infty. \tag{1.23}$$

Then f is Lipschitz continuous on $B_\delta(x)$ with Lipschitz constant $2M/\delta$.

Proof. We first show that $\sup_{y \in B_{2\delta}(x)} f(x) - f(y) \leq M$. Indeed, let $y \in B_{2\delta}(x)$. Then $z := x + (x - y) \in B_{2\delta}(x)$ and $(y + z)/2 = x$. Therefore,

$$f(x) \leq \frac{1}{2} f(y) + \frac{1}{2} f(z) \tag{1.24}$$

and hence $f(x) - f(y) \leq f(z) - f(x) \leq M$. Now, let $x_1, x_2 \in B_\delta(x)$ with $x_1 \neq x_2$. Clearly $x_3 = x_2 + \delta(x_2 - x_1)/\,\|x_2 - x_1\| \in B_{2\delta}(x)$ and x_2 can be written as a convex combination of x_1 and x_3. Indeed, by the definition of x_3,

$$\begin{aligned} \|x_2 - x_1\|\, x_3 &= (\|x_2 - x_1\| + \delta)x_2 - \delta x_1 \\ \implies x_2 &= \frac{\delta}{\|x_2 - x_1\| + \delta} x_1 + \frac{\|x_2 - x_1\|}{\|x_2 - x_1\| + \delta} x_3 = (1 - \lambda)x_1 + \lambda x_3, \end{aligned} \tag{1.25}$$

where $\lambda = \|x_2 - x_1\|\,/(\|x_2 - x_1\| + \delta)$. Therefore $f(x_2) \leq (1 - \lambda) f(x_1) + \lambda f(x_3)$ and hence

$$\begin{aligned} f(x_2) - f(x_1) &\leq \lambda\big(f(x_3) - f(x) + f(x) - f(x_1)\big) \\ &\leq 2M \frac{\|x_2 - x_1\|}{\|x_2 - x_1\| + \delta} \leq \frac{2M}{\delta}\,\|x_2 - x_1\|\,. \end{aligned}$$

□

Theorem 1.4.2. *Let $f\colon X \to\,]-\infty, +\infty]$ be a proper convex function. Then f is locally Lipschitz continuous (and hence continuous) on* $\operatorname{int}(\operatorname{dom} f)$.

Proof. Let $n = \dim(X)$. We can identify X with $\mathbb{R}^n$. We first note that for every $\delta > 0$, we have

$$B_{\|\cdot\|_1}(\delta) = \{x \in X \mid \|x\|_1 \leq \delta\} = \operatorname{co}(\{\pm\delta e_i \mid i = 1 \dots n\}), \tag{1.26}$$

where $(e_i)_{1\leq i\leq n}$ is the canonical basis of $\mathbb{R}^n$. Indeed, clearly the inclusion $\supset$ is true, since $B_{\|\cdot\|_1}(\delta)$ is convex and $\pm\delta e_i \in B_{\|\cdot\|_1}(\delta)$. Now suppose that $0 < \|x\|_1 \leq \delta$. Then, setting $r = \|x\|_1$, we have

$$x = \sum_{i=1}^{n} \frac{x_i^+}{r} re_i + \sum_{i=1}^{n} \frac{x_i^-}{r}(-re_i), \tag{1.27}$$

where x_i^+ and x_i^- are the positive and negative parts of x_i respectively. Since $\sum_{i=1}^n (x_i^+/r) + \sum_{i=1}^n (x_i^-/r) = \|x\|_1 / r = 1$, it follows that $x \in \operatorname{co}(\{\pm re_i \mid i = 1 \dots n\})$. Now we note that $\pm re_i \in \operatorname{co}(\{\delta e_i, -\delta e_i\}) \subset \operatorname{co}(\{\pm\delta e_i \mid i = 1 \dots n\})$. Therefore $\operatorname{co}(\{\pm re_i \mid i = 1 \dots n\}) \subset \operatorname{co}(\{\pm\delta e_i \mid i = 1 \dots n\})$ and the statement follows. Now let $x_0 \in \operatorname{int}(\operatorname{dom} f)$. Then there exists $\delta > 0$ such that $x_0 + B_{\|\cdot\|_1}(\delta) \subset \operatorname{dom} f$. Let $x \in x_0 + B_{\|\cdot\|_1}(\delta)$. Then $x = x_0 + \sum_{i=1} \lambda_i^+ \delta e_i + \sum_{i=1}^n \lambda_i^-(-\delta e_i) = \sum_{i=1} \lambda_i^+ (x_0 + \delta e_i) + \sum_{i=1}^n \lambda_i^-(x_0 - \delta e_i)$ with $\sum_{i=1}^n \lambda_i^+ + \sum_{i=1}^n \lambda_i^- = 1$ and hence, by convexity, $f(x) \leq \sum_{i=1} \lambda_i^+ f(x_0 + \delta e_i) + \sum_{i=1}^n \lambda_i^- f(x_0 - \delta e_i) \leq \max_{i=1,\dots,n} f(x_0 \pm \delta e_i) < +\infty$. Thus, we proved that f is bounded on a neighborhood of x_0 and the statement follows from Lemma 1.4.1. □

Remark 1.4.3. The points of discontinuity of an extended real-valued convex function belong to the boundary of its domain.

Proposition 1.4.4. *Let $f\colon X \to]-\infty, +\infty]$ be a proper function. Then*

$$(x, s) \in \operatorname{int}(\operatorname{epi}(f)) \;\Leftrightarrow\; \limsup_{y\to x} f(y) < s. \tag{1.28}$$

Therefore, if f is continuous at $x \in \operatorname{dom} f$, then

$$(x, f(x)) \notin \operatorname{int}(\operatorname{epi}(f)) \neq \varnothing \qquad \textit{and} \qquad (x, f(x)) \in \operatorname{cl}\big(\operatorname{int}(\operatorname{epi}(f))\big). \tag{1.29}$$

As a consequence, if f is continuous on $\operatorname{dom} f$, $\operatorname{int}(\operatorname{epi}(f)) = \{(x, s) \in X \times \mathbb{R} \mid f(x) < s\}$.

Proof. Let $(x, s) \in X \times \mathbb{R}$. Then

$$\begin{aligned}
(x, s) \in \operatorname{int}(\operatorname{epi}(f)) &\;\Leftrightarrow\; \exists \delta > 0, \exists \varepsilon > 0,\ B_\delta(x) \times [s - \varepsilon, s + \varepsilon] \subset \operatorname{epi}(f) \\
&\;\Leftrightarrow\; \exists \delta > 0, \exists \varepsilon > 0,\ \forall\, y \in B_\delta(x),\ f(y) \leq s - \varepsilon \\
&\;\Leftrightarrow\; \exists \delta > 0, \exists \varepsilon > 0,\ \sup_{y\in B_\delta(x)} f(y) \leq s - \varepsilon \\
&\;\Leftrightarrow\; \exists \delta > 0,\ \sup_{y\in B_\delta(x)} f(y) < s \\
&\;\Leftrightarrow\; \inf_{\delta>0} \sup_{y\in B_\delta(x)} f(y) < s.
\end{aligned}$$

Suppose that f is continuous at x. Then $(x, s) \in \operatorname{int}(\operatorname{epi}(f)) \;\Leftrightarrow\; f(x) < s$. Therefore $(x, f(x)) \notin \operatorname{int}(\operatorname{epi}(f))$ and $\{x\} \times]f(x), +\infty[\subset \operatorname{int}(\operatorname{epi}(f))$. Hence (1.29) follows. □

Remark 1.4.5. Let $f\colon X \to]-\infty, +\infty]$ be a proper function.

(i) The equivalence $(x,s) \in \operatorname{int}(\operatorname{epi}(f)) \;\Leftrightarrow\; f(x) < s$ does not hold if f is not continuous at x as Figure 1.3 shows.

(ii) Let $x \in \operatorname{dom} f$. Even without requiring continuity at x, we have that, $(x, f(x)) \in \operatorname{bdry}(\operatorname{epi}(f))$. Indeed $(x, f(x)) \notin \operatorname{int}(\operatorname{epi}(f))$, since if it was not so, there would exists $\delta > 0$ and $\varepsilon > 0$ such that $B_\delta(x) \times [f(x) - \varepsilon, f(x) + \varepsilon] \subset \operatorname{epi}(f)$. Therefore, we would have $(x, f(x) - \varepsilon) \in \operatorname{epi}(f)$ which yields $f(x) - \varepsilon \geq f(x)$.

1.5 Minimizers of convex functions

Definition 1.5.1. Let $f\colon X \to]-\infty, +\infty]$ be a proper function. A (global) *minimizer* of f is a point $x_* \in X$ such that

$$(\forall\, x \in X) \quad f(x_*) \leq f(x).$$

A *local minimizer of* f is a point x_* such that

$$(\exists\, \delta > 0)\big(\forall\, x \in B_\delta(x_*)\big) \quad f(x_*) \leq f(x)$$

Proposition 1.5.2. *Let $f\colon X \to]-\infty, +\infty]$ be a proper convex function and let $x \in \operatorname{dom} f$. Then the following statements are equivalent:*

(i) *x is a global minimizer of f;*

(ii) *x is a local minimizer of f;*

Proof. (ii) $\Rightarrow$ (i): Let $\delta > 0$ be such that, $f \geq f(x)$ on $B_\delta(x)$. Let $y \in X$. Let $t \in]0, 1[$ be such that $t\,\|y - x\| \leq \delta$. Then, since f is convex, $f(x + t(y-x)) \leq f(x) + t\big(f(y) - f(x)\big)$ and hence $0 \leq \big(f(x + t(y-x)) - f(x)\big)/t \leq f(y) - f(x)$ and the statement follows. □

The set of minimizers of f is denoted by $\operatorname{argmin} f$. Since $\operatorname{argmin} f$ is a sublevel set ($\operatorname{argmin} f = [f \leq \min f]$), if f is convex, than $\operatorname{argmin} f$ is convex too and this implies that if $\operatorname{argmin} f$ is nonempty, then it is either a singleton or an infinite set (because if it contains two distinct points, it would contains the segment joining those points). When existence of minimizers is in order, the following definition is needed.

Definition 1.5.3. Let $f\colon X \to]-\infty, +\infty]$ be a proper function. Then f is *coercive* if

$$\lim_{\|x\| \to +\infty} f(x) = +\infty,$$

which is equivalent to say that, for every $\alpha \in \mathbb{R}$, $[f \leq \alpha]$ is bounded.

Theorem 1.5.4 (Existence of minimizers)**.** *Let $f\colon X \to]-\infty, +\infty]$ be a proper, closed, and coercive function. Then f admits a global minimizer.*

Theorem 1.5.5 (Uniqueness of minimizers)**.** *Let $f\colon X \to]-\infty, +\infty]$ be a proper and strictly convex function. Then there exists at most one minimizer.*

Proof. Let x and y be two different minimizers of f. Then $f_* := \inf_{x' \in X} f(x') = f(x) = f(y) \in \mathbb{R}$ and $f((x+y)/2) < (1/2)f(x) + (1/2)f(y) = f_*$ and this gives a contradiction. □

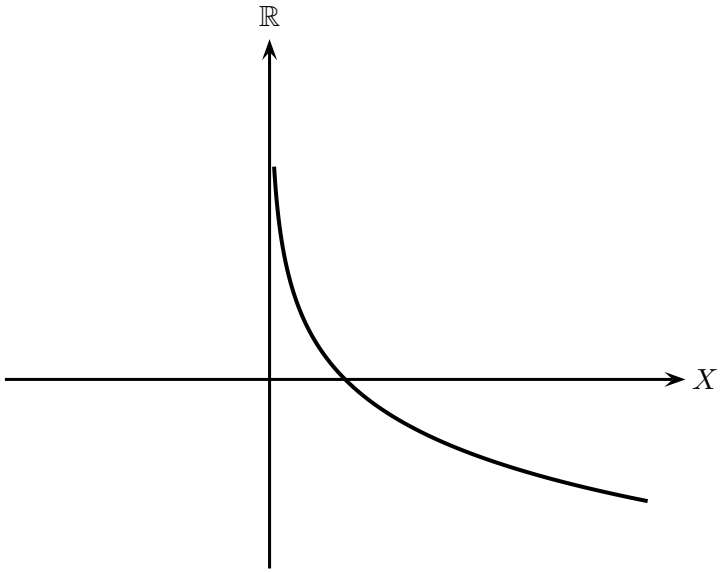


Figure 1.4: The function $-\log x$ is convex and unbounded from below.

Remark 1.5.6. The function $f(x) = -\log x$ if $x > 0$, and $f(x) = +\infty$, if $x \leq 0$, belongs to $\Gamma_0(X)$ and is not bounded from below. See Figure 1.4.

Lecture 2

Smooth Optimization: The gradient descent method

In this lecture we will talk about first order optimization methods that are gradient based. The aim is to solve the optimization problem

$$\min_{x \in X} f(x), \tag{2.1}$$

where f is differentiable. To this purpose we rely on *iterative methods*, that is methods that builds a sequence $(x_k)_{k \in \mathbb{N}}$ iteratively, that is by starting with an initial guess x_0 and then defining x_{k+1} by applying some explicit rule on the previous $x_k, \cdots, x_0$.

Among the several options, the simplest one is that of using the current point x_k and a *descent direction* at that point, that is, a unit vector u along with the derivative of f at x_k is negative

$$D_u f(x_k) := \lim_{t \to 0} \frac{f(x_k + tu) - f(x_k)}{t} < 0. \tag{2.2}$$

Indeed in such case, by the same definition of limit, we can find a sufficiently small $t_k > 0$ such that $f(x_k + t_k u) - f(x_k) < 0$; so defining

$$x_{k+1} = x_k + t_k u, \tag{2.3}$$

the objective function f is decreased. We note that for any unit vector u we have $D_u f(x) = \langle \nabla f(x_k), u \rangle$ and, by the Cauchy-Schwarz inequality,

$$- \|\nabla f(x_k)\| \leq \langle \nabla f(x_k), u \rangle \leq \|\nabla f(x_k)\| . \tag{2.4}$$

So, there exists the *steepest* descent direction which is

$$- \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}, \tag{2.5}$$

since

$$\min_{\|u\|=1} \langle \nabla f(x_k), u \rangle = - \|\nabla f(x_k)\| = \left\langle \nabla f(x_k), - \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|} \right\rangle.$$

Salzo Saverio (saverio.salzo@iit.it) Istituto Italiano di Tecnologia. Via E. Melen, 16152, Genoa, Italy.

The method that uses such direction is called the *steepest descent method* or the *gradient descent method*, and goes back to Cauchy: it has the form

$$x_{k+1} = x_k - \gamma \nabla f(x_k) \quad (\gamma > 0).$$

We note that if at some iteration, say k, this method does not make any progress, we have $x_k = x_k - \gamma \nabla f(x_k)$ and hence $\nabla f(x_k) = 0$, that is x_k is a minimizer of f.

2.1 Differentiability and convexity

We recall the definition of differentiable functions. Throughout the chapter X will be an Euclidean space.

Let $f\colon X \to]-\infty, +\infty]$ be a proper extended real-valued function and let $x_0 \in \operatorname{int}(\operatorname{dom} f)$. Then f *is (Gâteaux) differentiable at* x_0 if there exists a vector $\nabla f(x) \in X$ such that

$$(\forall\, v \in X) \qquad \lim_{t \to 0} \frac{f(x_0 + tv) - f(x_0)}{t} = \langle \nabla f(x_0), v \rangle. \tag{2.6}$$

In such case $\nabla f(x)$ is called the *gradient of* f *at* x_0. Thus, f admits *directional derivatives* at x_0 in every direction v and the directional derivatives depend linearly from v. When f is differentiable at every point of a subset $A \subset \operatorname{int}(\operatorname{dom} f)$ we say that f *is differentiable on* A.

Remark 2.1.1. In case $X = \mathbb{R}^n$, if we take $v = e_i$ (the canonical basis of $\mathbb{R}^n$), then we get $\langle \nabla f(x_0), e_i \rangle = \partial_i f(x_0)$, where $\partial_i f(x_0)$ is the partial derivative of f with respect to the i-th coordinate at the point x_0. Hence $\nabla f(x_0) = \big(\partial_1 f(x_0), \dots, \partial_n f(x_0)\big)$.

Proposition 2.1.2 (Characterizations of convexity)**.** *Let* $f\colon X \to]-\infty, +\infty]$ *be a proper extended real-valued function such that* $\operatorname{dom} f$ *is open and convex. Suppose that* f *is differentiable on* $\operatorname{dom} f$*. Then the following are equivalent statements.*

(i) f *is convex.*

(ii) *For every* $x, y \in \operatorname{dom} f$, $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.

(iii) *For every* $x, y \in \operatorname{dom} f$, $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$.

In case f *is twice differentiable on* $\operatorname{dom} f$*, the previous statements are equivalent to*

(iv) *for every* $x \in \operatorname{dom} f$ *and for every* $v \in X$, $\langle \nabla^2 f(x) v, v \rangle \geq 0$.

Proof. (i) $\Rightarrow$ (ii): Let $x, y \in \operatorname{dom} f$. Then, for every $\lambda \in]0, 1]$, we have $f((1-\lambda)x + \lambda y) \leq (1 - \lambda) f(x) + \lambda f(y)$ and hence

$$\frac{f(x + \lambda(y - x)) - f(x)}{\lambda} \leq f(y) - f(x). \tag{2.7}$$

Thus, by (2.6) and (2.7), we get $\langle \nabla f(x), y - x \rangle \leq f(y) - f(x)$.

(ii) $\Rightarrow$ (iii): Let $x, y \in \mathrm{dom} f$. Then we have

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \geq 0$$
$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq 0.$$

Summing, we get $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$ and hence the statement.

(iii) $\Rightarrow$ (i): Let $x, y \in \mathrm{dom} f$ and define $\phi\colon [0,1] \to \mathbb{R}$, such that, $\phi(\lambda) = f(x+\lambda(y-x))$. Then $\phi(0) = f(x)$, $\phi(1) = f(y)$, and ϕ is differentiable on $[0,1]$ and, for every $\lambda \in [0,1]$, $\phi'(\lambda) = \langle \nabla f(x + \lambda(y-x)), y - x \rangle$. Now, let $\lambda_1, \lambda_2 \in [0,1]$ be such that $\lambda_1 < \lambda_2$. Then

$$\langle \nabla f(x + \lambda_2(y-x)) - \nabla f(x + \lambda_1(y-x)), (\lambda_2 - \lambda_1)(y-x) \rangle \geq 0$$

and hence $(\lambda_2 - \lambda_1)(\phi'(\lambda_2) - \phi'(\lambda_1)) \geq 0$, which yields $\phi'(\lambda_1) \leq \phi'(\lambda_2)$. Therefore ϕ' is increasing. Now, let $\lambda \in]0,1[$. We show that

$$f(x + \lambda(y-x)) \leq f(x) + \lambda\big(f(y) - f(x)\big). \tag{2.8}$$

Indeed, it follows from Lagrange's theorem that there exist $\lambda_1 \in]0, \lambda[$ and $\lambda_2 \in]\lambda, 1[$ such that

$$\frac{\phi(\lambda) - \phi(0)}{\lambda} = \phi'(\lambda_1) \quad \text{and} \quad \frac{\phi(1) - \phi(\lambda)}{1 - \lambda} = \phi'(\lambda_2).$$

Thus, since $\phi'(\lambda_1) \leq \phi'(\lambda_2)$, we have $(1-\lambda)(\phi(\lambda) - \phi(0)) \leq \lambda(\phi(1) - \phi(\lambda))$. Rearranging this inequality (2.8) follows.

(iii) $\Rightarrow$ (iv): Let $x \in \mathrm{dom} f$ and $v \in X$. Since $\mathrm{dom} f$ is open, there exists $\delta > 0$ such that, for every $t \in]0, \delta]$, $x + tv \in \mathrm{dom} f$ and, because of (iii), $\langle \nabla f(x+tv) - \nabla f(x), tv \rangle \geq 0$, hence, dividing by t^2,

$$(\forall\, t \in]0, \delta]) \quad \left\langle \frac{\nabla f(x+tv) - \nabla f(x)}{t}, v \right\rangle \geq 0. \tag{2.9}$$

Since, by definition, $\nabla^2 f(x)v = \lim_{t \to 0} \big(\nabla f(x+tv) - \nabla f(x)\big)/t$, the statement follows.

(iv) $\Rightarrow$ (iii): Let $x, y \in \mathrm{dom} f$ and define $\phi\colon [0,1] \to \mathbb{R}$ as in the proof of (iii) $\Rightarrow$ (ii). Then, ϕ is twice differentiable and $\phi''(\lambda) = \langle \nabla^2 f(x + \lambda(y-x))(y-x), y - x \rangle \geq 0$. Therefore, ϕ' is increasing in $[0,1]$. Hence $\phi'(0) \leq \phi'(1)$, which means $\langle \nabla f(x), y - x \rangle \leq \langle \nabla f(y), y - x \rangle$. □

Remark 2.1.3. Strict convexity can be characterized by statements (ii) and (iii) of Proposition 2.1.2, where "$\geq$" is replaced by "$>$" and $x \neq y$.

Example 2.1.4. The function $f\colon \mathbb{R} \to]-\infty, +\infty]$ defined as $f(x) = -\log x$ if $x > 0$ and $f(x) = +\infty$ if $x \leq 0$ is strictly convex. Indeed if $x > 0$ and $y > 0$, with $x \neq y$, we have $(f'(x) - f'(y))(x-y) = (-1/x + 1/y)(x-y) = (x-y)^2/(xy) > 0$. Note, in passing, that by the convexity of $-\log$ one can derive the *arithmetic-geometric means inequality*, that is,

$$\frac{1}{n} \sum_{i=1}^{n} x_i \geq \prod_{i=1}^{n} x_i^{1/n}. \tag{2.10}$$

Indeed we have $-\log(\sum_{i=1}^n (1/n) x_i) \leq -\sum_{i=1}^n (1/n) \log x_i$. Changing the sign and taking the exponential we get $(1/n)\sum_{i=1}^n x_i \geq \exp(\sum_{i=1}^n \log x_i^{1/n}) = \prod_{i=1}^n x_i^{1/n}$. When the numbers x_i's are not all equal, then the inequality is strict (due to the strict convexity of $-\log$).

Theorem 2.1.5 (Fermat's rule)**.** *Let $f\colon X \to]-\infty,+\infty]$ be a proper convex function. Let $x \in \operatorname{int}(\operatorname{dom} f)$ and suppose that f is differentiable at x. Then the following statements are equivalent:*

(i) *x is a minimizer of f;*

(ii) $\nabla f(x) = 0$.

Proof. (ii) $\Rightarrow$ (i): It follows from Proposition 2.1.2(ii).

(i) $\Rightarrow$ (ii): It follows from (2.6) that,

$$(\forall\, y \in X) \quad \langle \nabla f(x), y - x\rangle = \lim_{t\to 0}(f(x + t(y - x)) - f(x))/t \geq 0.$$

Thus, taking $y = x - \nabla f(x)$, we get $-\|\nabla f(x)\|^2 \geq 0$ and hence $\nabla f(x) = 0$. □

From Proposition 2.1.2 and Proposition 1.3.12 we derive the following result.

Corollary 2.1.6. *Let $f\colon X \to]-\infty,+\infty]$ be a proper extended real-valued function such that $\operatorname{dom} f$ is open and convex and let $\mu > 0$. Suppose that f is differentiable on $\operatorname{dom} f$. Then the following statements are equivalent.*

(i) *f is μ-strongly convex.*

(ii) *For every $x, y \in \operatorname{dom} f$, $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + (\mu/2)\|y - x\|^2$.*

(iii) *For every $x, y \in \operatorname{dom} f$, $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \mu\|x - y\|^2$.*

In case f is twice differentiable on $\operatorname{dom} f$, the previous statements are equivalent to

(iv) *for every $x \in \operatorname{dom} f$ and for every $v \in X$, $\langle \nabla^2 f(x)v, v\rangle \geq \mu\|v\|^2$.*

Example 2.1.7. Let $A\colon X \to Y$ be a linear operator and let $b \in Y$. Set

$$f\colon X \to \mathbb{R} \quad f(x) = \frac{1}{2}\|Ax - b\|^2. \tag{2.11}$$

Suppose that A^*A is positive definite (i.e., the minimum eigenvalue of A^*A is strictly positive). Then, for every $x \in X$, $\nabla f(x) = A^*(Ax - b)$ and hence

$$(\forall\, x, y \in X) \quad \langle \nabla f(x) - \nabla f(y), x - y\rangle = \langle A^*A(x - y), x - y\rangle \geq \mu\|x - y\|^2,$$

where μ is the minimum eigenvalue of A^*A. Thus, by Corollary 2.1.6, f is μ-strongly convex.

Example 2.1.8. The function defined in Example 2.1.4 is not strongly convex. Indeed if it was so, then there would exists $\mu > 0$ such that, for every $x, y > 0$, $x \neq y$, we would have $(x - y)^2/(xy) = (f'(x) - f'(y))(x - y) \geq \mu(x - y)^2$, that is $1/(xy) > \mu$. But this last statement is false since $1/(xy) \to 0$ as $x \to +\infty$ and $y \to +\infty$.

Remark 2.1.9.

(i) Corollary 2.1.6 establishes that strongly convex functions can be bounded from below at each point by tangent quadratic functions.

(ii) Functions that are proper, strongly convex, and closed admits a unique minimizer. Indeed if f is such a function, then $f = g + (\mu/2)\|\cdot\|^2$. Since g is closed too, by Theorem 4.4.2, it has an affine minimizer, that is, there exists $u \in X$ and $\alpha \in \mathbb{R}$ such that $\alpha + \langle \cdot, u\rangle \leq g$. Thus $\lim_{\|x\|\to+\infty} f(x) \geq \lim_{\|x\|\to+\infty} \alpha + \langle x, u\rangle + (\mu/2)\|x\|^2 = +\infty$; hence f is coercive. So, since f is closed, Thereom 1.5.4 ensures that f has minimizers. Uniqueness comes from Theorem 1.5.5. Note that concerning the existence of minimizers, closedness is necessary even for strongly convex functions. Indeed the function

$$f\colon \mathbb{R} \to]-\infty, +\infty], \qquad f(x) = \begin{cases} x^2 & \text{if } x > 0 \\ +\infty & \text{if } x \leq 0 \end{cases}$$

does not have any minimizer. The problem is that, even though f is coercive, it is not closed.

(iii) Under the assumption of Corollary 2.1.6, suppose that x_* is a minimizer of f. Then, it follows from Corollary 2.1.6(iii) that, for every $x \in \mathrm{dom} f$, $\langle \nabla f(x), x - x_*\rangle \geq \mu\|x - x_*\|^2$ and hence

$$\boxed{(\forall\, x \in \mathrm{dom} f) \quad \mu\|x - x_*\| \leq \|\nabla f(x)\|\,.} \tag{2.12}$$

Proposition 2.1.10. *Let $f\colon X \to]-\infty, +\infty]$ be a proper extended real-valued function such that $\mathrm{dom} f$ is open and let $\mu > 0$. Suppose that f is differentiable on $\mathrm{dom} f$ and μ-strongly convex. Suppose that there exists the minimizer of f. Then,*

$$\boxed{(\forall\, x \in \mathrm{dom} f) \quad f(x) - \min_{x\in X} f(x) \leq \frac{1}{2\mu}\|\nabla f(x)\|^2\,.} \tag{2.13}$$

Proof. Let $x \in \mathrm{dom} f$. Then Corollary 2.1.6(ii) yields

$$\begin{aligned} \min_{y\in\mathrm{dom} f} f(y) &\geq \min_{y\in\mathrm{dom} f} \Big(f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2 \Big) \\ &= f(x) + \min_{y\in\mathrm{dom} f} \frac{1}{2\mu}\big(\underbrace{\|\mu(y - x) + \nabla f(x)\|^2}_{\geq 0} - \|\nabla f(x)\|^2 \big) \\ &\geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2\,. \end{aligned}$$

and the statement follows. □

Example 2.1.11. Condition (2.13) is called *Łojasiewicz inequality* and can hold even for non-strongly convex functions and very recently has been the objective of intense research which has unveiled its connection with quadratic growth conditions like (1.20) and ultimately its critical role in achieving linear convergence in optimization algorithms. Here we provide a significant example for that situation. Let f be as in Example 2.11 where now we do not assume A to be positive definite. Let b_* be the projection of b onto

the range $R(A)$ of A. Let $x \in X$ and $x_* \in \operatorname{argmin} f = \{x \in X \mid Ax = b_*\}$. Then by Pythagoras' theorem, we have

$$f(x) = \frac{1}{2}\big(\|Ax - b_*\|^2 + \|b_* - b\|^2 \big).$$

Hence $f_* = \inf_X f = (1/2)\,\|b_* - b\|^2$ and

$$f(x) - f_* = \frac{1}{2}\,\|Ax - b_*\|^2 = \frac{1}{2}\,\|A(x - x_*)\|^2 .$$

Moreover, $\nabla f(x) = A^*(Ax - b_*) = A^*A(x - x_*)$, and hence

$$\|\nabla f(x)\|^2 = \|A^*A(x - x_*)\|^2 .$$

Thus, inequality (2.13) in this case reduces to

$$(\forall\, x \in X) \quad \mu\,\|A(x - x_*)\|^2 \le \|A^*A(x - x_*)\|^2 ,$$

which is equivalent to

$$(\forall\, y \in R(A)) \quad \mu\,\|y\|^2 \le \|A^*y\|^2 = \langle AA^*y, y\rangle.$$

Now, we consider a singular value decomposition of A^*

$$A^*y = \sum_{i\in I} \sigma_i \langle y, b_i\rangle c_i,$$

where $(\sigma_i)_{i\in I}$ are the singular values of A^* ($(\sigma_i^2)_{i\in I}$ are the nonzero eigenvalues of AA^*), $(b_i)_{i\in I}$ is an orthonormal basis of $N(A^*)^\perp = R(A)$ and $(c_i)_{i\in I}$ is an orthonormal basis of $R(A^*)$. Then, for every $y \in N(A^*)^\perp$,

$$\|A^*y\|^2 = \sum_{i\in I} \sigma_i^2 |\langle y, b_i\rangle|^2 \ge \sigma_{\min}^2 \sum_{i\in I} |\langle y, b_i\rangle|^2 = \sigma_{\min}^2\,\|y\|^2 .$$

Therefore (2.13) holds, with $\mu = \sigma_{\min}^2$, the minimum nonzero eigenvalue of A^*A.

We now study the property of *L-smoothness*, which means that the gradient of the function is L-Lipschitz continuous. The following theorem provides several characterizations of L-smoothness that will be useful in analyzing the gradient descent method. The implication (i) $\Rightarrow$ (ii) is called the *descent lemma*, whereas the implication (i) $\Rightarrow$ (iv) is called the *Baillon-Haddad theorem*.

Theorem 2.1.12. *Let $f\colon X \to \mathbb{R}$ be a convex differentiable function and let $L \in \mathbb{R}_+$. The following statements are equivalent.*

(i) *For every x and y in X, $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$.*

(ii) *For every x and y in X,*

$$f(y) - f(x) - \langle \nabla f(x), y - x\rangle \leq \frac{L}{2}\|x - y\|^2. \tag{2.14}$$

(iii) *For every x and y in X,*

$$\frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x\rangle \tag{2.15}$$

(iv) *For every x and y in X,*

$$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y\rangle \tag{2.16}$$

(v) *For every x and y in X, $\langle \nabla f(x) - \nabla f(y), x - y\rangle \leq L\|x - y\|^2$.*

(vi) *$\frac{L}{2}\|\cdot\|^2 - f$ is convex.*

In case f is twice differentiable on X, the previous statements are equivalent to

(vii) *for every $x \in X$ and for every $v \in X$, $\langle \nabla^2 f(x)v, v\rangle \leq L\|v\|^2$.*

(viii) *for every $x \in X$, $\|\nabla^2 f(x)\| \leq L$.*

Proof. (i) $\Rightarrow$ (ii): Let $x, y \in X$ and set $\phi\colon [0,1] \to \mathbb{R}$, $\phi(\lambda) = f(x + \lambda(y - x))$. Then ϕ is continuously differentiable and, for every $\lambda \in [0,1]$, $\phi'(\lambda) = \langle \nabla f(x + \lambda(y - x)), y - x\rangle$. Thus,

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x\rangle &= \phi(1) - \phi(0) - \langle \nabla f(x), y - x\rangle \\ &= \int_0^1 \phi'(\lambda)\,\mathrm{d}\lambda - \langle \nabla f(x), y - x\rangle \\ &= \int_0^1 \langle \nabla f(x + \lambda(y - x)) - \nabla f(x), y - x\rangle\,\mathrm{d}\lambda \qquad (2.17) \\ &\leq \int_0^1 \|\nabla f(x + \lambda(y - x)) - \nabla f(x)\|\,\|y - x\|\,\mathrm{d}\lambda \\ &\leq L\|y - x\|^2 \int_0^1 \lambda\,\mathrm{d}\lambda \\ &= \frac{L}{2}\|y - x\|^2. \end{aligned}$$

(ii) $\Rightarrow$ (iii): Let $x \in X$ and let $g\colon X \to \mathbb{R}\colon y \mapsto f(y) - \langle \nabla f(x), y\rangle$. Then g is convex and differentiable and, for every $y \in X$, $\nabla g(y) = \nabla f(y) - \nabla f(x)$. Moreover, it is easy

to see that g satisfies (ii) too. Since $\nabla g(x) = 0$, x is a minimizer of g. Now let $y \in X$. Using implication(ii) applied to g,

$$\begin{aligned} g(x) = \min_{z\in X} g(z) &\leq \min_{z\in X} \Big(g(y) + \langle \nabla g(y), z - y\rangle + \frac{L}{2} \|z - y\|^2 \Big) \\ &= g(y) + \min_{z\in X} \frac{1}{2L} \big(\|L(z - y) + \nabla g(y)\|^2 - \|\nabla g(y)\|^2 \big) \\ &= g(y) - \frac{1}{2L} \|\nabla g(y)\|^2 . \end{aligned}$$

Substituting the expression of $g(x), g(y)$, and $\nabla g(y)$ into the above inequality, (2.15) follows.

(iii) $\Rightarrow$ (iv): The statement follows by swapping x and y in (2.15) and summing the resulting inequality with (2.15).

(iv) $\Rightarrow$ (i): Let x and y in X. By the Cauchy-Schwarz inequality we get

$$\frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y\rangle \leq \|\nabla f(x) - \nabla f(y)\| \, \|x - y\| .$$

Thus $\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$.

(ii) $\Rightarrow$ (v): It follows by swapping x and y in (2.14) and summing with (2.14).

(v) $\Rightarrow$ (ii): Let $x, y \in X$. If we define ϕ as in the proof of (i) $\Rightarrow$ (ii), we see that (v) implies that ϕ' is continuous. Therefore, it follows from (2.17) that

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x\rangle &= \int_0^1 \frac{1}{\lambda} \langle \nabla f(x + \lambda(y - x)) - \nabla f(x), \lambda(y - x)\rangle \, \mathrm{d}\lambda \\ &\leq \int_0^1 L \|y - x\|^2 \, \lambda \, \mathrm{d}\lambda \\ &= \frac{L}{2} \|y - x\|^2 . \end{aligned}$$

(v) $\Leftrightarrow$ (vi): Condition (v) can be equivalently written as

$$(\forall\, x, y \in X) \quad \langle Lx - \nabla f(x), x - y\rangle \geq 0.$$

Since $\nabla\big((L/2) \|\cdot\|^2 - f\big)(x) = Lx - \nabla f(x)$, the statement follows from Proposition 2.1.2. □

Proposition 2.1.13. *Let $f\colon X \to \mathbb{R}$ be a differentiable function. Then the following are equivalent*

(i) *f is μ-strongly convex and ∇f is Lipschitz continuous with constant L.*

(ii) $\forall\, x, y \in X$, $\dfrac{1}{L + \mu} \|\nabla f(x) - \nabla f(y)\|^2 + \dfrac{\mu L}{L + \mu} \|x - y\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y\rangle$.

Proof. Set $g = f - (\mu/2) \|\cdot\|^2$. Then $\frac{L-\mu}{2} \|\cdot\|^2 - g = (L/2) \|\cdot\|^2 - f$. Therefore It follows from Theorem 2.1.12(vi) that (i) is equivalent to g convex and with $(L - \mu)$-Lipschitz continuous gradient. Then, by Theorem 2.1.12 item (iv), this latter fact can be expressed as

$$(\forall\, x, y \in X) \quad \frac{1}{L - \mu} \|\nabla g(x) - \nabla g(y)\|^2 \leq \langle \nabla g(x) - \nabla g(y), x - y\rangle. \tag{2.18}$$

and hence, substituting the expressions of $\nabla g(x)$ and $\nabla g(y)$, as

$$(\forall\, x, y \in X)\ \frac{1}{L-\mu}\left\|\nabla f(x) - \nabla f(y) - \mu(x-y)\right\|^2 \leq \langle \nabla f(x) - \nabla f(y), x-y\rangle - \mu\left\|x-y\right\|^2.$$

The statement follows from the latter inequality, by multiplying by $L-\mu$, expanding the square norm on the left hand side and rearranging the terms. □

2.2 Nonexpansive and contractive operators

Definition 2.2.1. Let X be an Euclidean space and let $T\colon X \to X$. Then

(i) T is *nonexpansive* if

$$(\forall\, x, y \in X) \quad \left\|Tx - Ty\right\| \leq \left\|x-y\right\|.$$

(ii) T is a *contraction* if there exists $q \in\,]0,1[$ such that

$$(\forall\, x, y \in X) \quad \left\|Tx - Ty\right\| \leq q\left\|x-y\right\|,$$

The first important result concerns contractive mapping.

Theorem 2.2.2 (Banach-Caccioppoli)**.** *Let $T\colon X \to X$ be a q-contractive mapping for some $0 < q < 1$. Then there exists a unique fixed point of T, that is, a point $x_* \in X$ such that $Tx_* = x_*$. Moreover, let $x_0 \in X$ and define, iteratively*

$$x_{k+1} = Tx_k. \tag{2.19}$$

Then,

$$(\forall\, k \in \mathbb{N}) \quad \left\|x_k - x_*\right\| \leq q^k\left\|x_0 - x_*\right\| \quad \text{and} \qquad \left\|x_k - x_*\right\| \leq \frac{q^k}{1-q}\left\|x_0 - x_1\right\|. \tag{2.20}$$

Proof. We first note that

$$(\forall\, x, y \in X) \qquad \left\|x-y\right\| \leq \frac{1}{1-q}\big(\left\|x - Tx\right\| + \left\|y - Ty\right\|\big). \tag{2.21}$$

Indeed $\left\|x-y\right\| \leq \left\|x-Tx\right\| + \left\|Tx-Ty\right\| + \left\|Ty-y\right\| \leq \left\|x-Tx\right\| + q\left\|x-y\right\| + \left\|y-Ty\right\|$, hence $(1-q)\left\|x-y\right\| \leq \left\|x-Tx\right\| + \left\|Ty-y\right\|$ and (2.21) follows. Inequality (2.21) shows that there may exists at most one fixed point of T. Moreover, for every $k, h \in \mathbb{N}$,

$$\begin{aligned}
\left\|x_k - x_h\right\| &\leq \frac{1}{1-q}\big(\left\|x_k - x_{k+1}\right\| + \left\|x_h - x_{h+1}\right\|\big) \\
&\leq \frac{1}{1-q}\big(\left\|T^k x_0 - T^k x_1\right\| + \left\|T^h x_0 - T^h x_1\right\|\big) \\
&\leq \frac{1}{1-q}\big(q^k\left\|x_0 - x_1\right\| + q^h\left\|x_0 - x_1\right\|\big) \\
&\leq \frac{q^k + q^h}{1-q}\left\|x_0 - x_1\right\|.
\end{aligned} \tag{2.22}$$

where we used that T^k is q^k-contractive. Since $0 < q < 1$, q^k and q^h converge to zero as k and h go to $+\infty$. Therefore, $(x_k)_{k\in\mathbb{N}}$ is a Cauchy sequence and hence it converges, say to x_*. Then $Tx_k \to Tx_*$ and $Tx_k = x_{k+1} \to x_*$, so $Tx_* = x_*$, that is, x_* is a fixed point of T. The second inequality in (2.20) follows from (2.22) by letting $h \to +\infty$. The first equality in (2.20) follows from the following chain of inequalities

$$\|x_k - x_*\| = \|Tx_{k-1} - Tx_*\| \leq q \|x_{k-1} - x_*\| \leq \cdots \leq q^k \|x_0 - x_*\| .$$

□

Remark 2.2.3.

(i) Iterative methods of type (2.19) are called *fixed point iterations* or *Picard iterations*.

(ii) Nonexpansive operators, may have no fixed points. For instance, a translation $T = \mathrm{Id} + a$, with $a \neq 0$, does not have any fixed point.

(iii) For nonexpansive operators, even admitting fixed points, the Picard iteration may fail to converge. Indeed, this occurs if we take $T = -\mathrm{Id}$ and start with $x_0 \neq 0$. More generally rotations are nonexpansive operators admitting a fixed point, for which the Picard iterations do not converge.

Example 2.2.4.

(i) $\alpha\mathrm{Id}$, with $|\alpha| < 1$, is a contractive operator and its only fixed point is zero.

2.3 Convergence of the gradient descent method.

Now we define the gradient descent algorithm for minimizing smooth convex functions. In this section we assume that $f\colon X \to \mathbb{R}$ is convex differentiable with Lipschitz continuous gradient with constant L.

Algorithm 2.3.1. The *gradient descent algorithm* is defined as follows.

$$\begin{array}{l} \text{Let } \gamma > 0 \text{ and } x_0 \in X. \\ \text{For } k = 0, 1, \ldots \\ \quad \lfloor\ x_{k+1} = x_k - \gamma \nabla f(x_k). \end{array} \tag{2.23}$$

It is important to note that some restriction on the step-size γ should be required. Indeed if we do gradient descent to the function $f(x) = (L/2)\|x\|^2$, we have

$$x_{k+1} = (1 - \gamma L)x_k.$$

Thus if we take $\gamma = 2/L$, we have $x_{k+1} = -x_k$ and the sequence does not converge, unless $x_0 = 0$.

We have the following result.

Proposition 2.3.2. *Let $k \in \mathbb{N}$. Then*

$$\gamma\Big(1 - \frac{\gamma L}{2}\Big) \|\nabla f(x_k)\|^2 \leq f(x_k) - f(x_{k+1}). \tag{2.24}$$

Thus, if $\gamma \leq 2/L$, then $f(x_{k+1}) \leq f(x_k)$, that is, the algorithm is descending.

Proof. Since $x_{k+1} - x_k = -\gamma \nabla f(x_k)$, by Theorem 2.1.12(ii), we have

$$f(x_{k+1}) \leq f(x_k) - \gamma \langle \nabla f(x_k), \nabla f(x_k) \rangle + \frac{L}{2}\gamma^2 \|\nabla f(x_k)\|^2$$

and the statement follows. □

Theorem 2.3.3 (Convergence 1)**.** *Suppose that $f\colon X \to \mathbb{R}$ is Lipschitz smooth with constant $L > 0$ and strongly convex with modulus $\mu > 0$. Let x_* be the minimizer of f and suppose that $0 < \gamma < 2/L$. Then, for every $k \in \mathbb{N}$,*

$$f(x_k) - f(x_*) \leq q_1(\gamma)^{2k}(f(x_0) - f(x_*)) \text{ and } \|x_k - x_*\| \leq \sqrt{\frac{L}{\mu}} q_1(\gamma)^k \|x_0 - x_*\|, \quad (2.25)$$

where $q_1(\gamma) = \big(1 - \gamma\mu(2 - L\gamma)\big)^{1/2} < 1$. Moreover, the optimal value for the rate is achieved for $\gamma = 1/L$, for which $q_1(\gamma) = (1 - \mu/L)^{1/2}$.

Proof. First of all we note that $q_1(\gamma) < 1$. Indeed $0 < \gamma\mu(2 - L\gamma) \leq \gamma L(2 - \gamma L) = -(\gamma L - 1)^2 + 1 \leq 1$ (we used $\mu \leq L$ in the second inequality). Then, it follows from Proposition 2.3.2 and Proposition 2.1.10 that

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \gamma\Big(1 - \frac{L\gamma}{2}\Big) \|\nabla f(x_k)\|^2 \\ &\leq f(x_k) - \gamma\mu(2 - L\gamma)(f(x_k) - f(x_*)). \end{aligned}$$

Therefore

$$f(x_{k+1}) - f(x_*) \leq q_1(\gamma)^2\big(f(x_k) - f(x_*)\big)$$

and hence

$$f(x_k) - f(x_*) \leq q_1(\gamma)^{2k}(f(x_0) - f(x_*)). \quad (2.26)$$

In the end, recalling (1.17) and that $f(x_0) - f(x_*) \leq (L/2)\|x_0 - x_*\|^2$, equation (2.25) follows. □

Remark 2.3.4. In the derivation of (2.26) we did not use the strong convexity directly, but only inequality (2.13). So, we have linear convergence in function value, even for non-strongly convex functions, as long as inequality (2.13) holds. This is the case of the function considered in Example 2.1.11. Moreover in such situation, since (1.20) holds too, the second inequality in (2.25) is replace by

$$\operatorname{dist}(x_k, \operatorname{argmin} f) \leq \sqrt{\frac{2}{\mu}} q_1(\gamma)^k \sqrt{f(x_0) - \min f}.$$

We now address the analysis of convergence of the gradient descent method through the fixed-point Theorem 2.2.2, which will lead to improve the convergence rate given in Theorem 2.3.3. Set $T = \mathrm{Id} - \gamma\nabla f\colon X \to X$, where Id is the identity operator on X. Then the gradient descent algorithm (2.23) can be written as a *Picard iteration*

$$x_{k+1} = Tx_k \quad (2.27)$$

and the minimizers of f are nothing but the fixed points of T.

We first address the question of when the gradient descent operator T is a contraction. This will provide necessary conditions for applying the Banach fixed point theorem.

Proposition 2.3.5. *Let $f\colon X \to \mathbb{R}$ be a differentiable convex function. Suppose that for some $\gamma > 0$, the operator $T = \mathrm{Id} - \gamma\nabla f$ is a contraction. Then f is strongly convex and its gradient is Lipschitz continuous.*

Proof. Let $x, y \in X$. Then

$$\begin{aligned}
&\|Tx - Ty\|^2 \leq q^2 \|x - y\|^2 \\
\Leftrightarrow\ &\|x - y - \gamma(\nabla f(x) - \nabla f(y))\|^2 \leq q^2 \|x - y\|^2 \\
\Leftrightarrow\ &\|x - y\|^2 + \gamma^2 \|\nabla f(x) - \nabla f(y)\|^2 - 2\gamma\langle \nabla f(x) - \nabla f(y), x - y\rangle \leq q^2 \|x - y\|^2 \\
\Leftrightarrow\ &(1 - q^2)\|x - y\|^2 + \gamma^2 \|\nabla f(x) - \nabla f(y)\|^2 \leq 2\gamma\langle \nabla f(x) - \nabla f(y), x - y\rangle \\
\Rightarrow\ &\begin{cases} \dfrac{1 - q^2}{2\gamma}\|x - y\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y\rangle \\ \dfrac{\gamma}{2}\|\nabla f(x) - \nabla f(y)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y\rangle \end{cases}
\end{aligned}$$

So in virtue of Theorem 2.1.12(iv) and Corollary 2.1.6(iii), f is strongly convex and ∇f is Lipschitz continuous. □

Now we assume that f is strongly convex and with Lipschitz continuous gradient. Then we will prove that there exists an interval of values of γ for which T is a contraction.

Theorem 2.3.6 (Convergence 2)**.** *$f\colon X \to \mathbb{R}$ is Lipschitz smooth with constant $L > 0$ and strongly convex with modulus $\mu > 0$. Let x_* be the minimizer of f. Then, for every $\gamma \in]0, 2/(L + \mu)]$, $T = \mathrm{Id} - \gamma\nabla f$ is a contraction with constant*

$$q_2(\gamma) := \left(1 - \frac{2\gamma\mu L}{L + \mu}\right)^{1/2}, \tag{2.28}$$

hence, for every $k \in \mathbb{N}$,

$$\|x_k - x_*\| \leq \left(1 - \frac{2\gamma\mu L}{L + \mu}\right)^{k/2} \|x_0 - x_*\|, \tag{2.29}$$

Moreover, the optimal step size in (2.29) is $\gamma = 2/(L + \mu)$ and in such case

$$\|x_k - x_*\| \leq \left(\frac{L - \mu}{L + \mu}\right)^{k} \|x_0 - x_*\|, \tag{2.30}$$

$$f(x_k) - f(x_*) \leq \frac{L}{2}\left(\frac{L - \mu}{L + \mu}\right)^{2k} \|x_0 - x_*\|^2. \tag{2.31}$$

Proof. It follows from Proposition 2.1.13(ii) (multiplied by 2γ) that

$$\frac{2}{\gamma(L + \mu)}\|\gamma\nabla f(x) - \gamma\nabla f(y)\|^2 + \frac{2\gamma\mu L}{L + \mu}\|x - y\|^2 \leq 2\langle \gamma\nabla f(x) - \gamma\nabla f(y), x - y\rangle$$

Moreover,

$$\begin{aligned}
\|(x - y) - \gamma(\nabla f(x) - \nabla f(y))\|^2 = \|x - y\|^2 + \|\gamma\nabla f(x) - \gamma\nabla f(y)\|^2 \\
- 2\langle \gamma\nabla f(x) - \gamma\nabla f(y), x - y\rangle.
\end{aligned}$$

Hence

$$\|(x-y)-\gamma(\nabla f(x)-\nabla f(y))\|^2 \leq \left(1-\frac{2\gamma\mu L}{L+\mu}\right)\|x-y\|^2 - \left(\frac{2}{\gamma(L+\mu)}-1\right)\|\gamma\nabla f(x)-\gamma\nabla f(y)\|^2.$$

Now since $T = \mathrm{Id} - \gamma\nabla f$, the inequality above becomes

$$\|Tx-Ty\|^2 \leq \left(1-\frac{2\gamma\mu L}{\mu+L}\right)\|x-y\|^2 - \left(\frac{2}{\gamma(\mu+L)}-1\right)\|(\mathrm{Id}-T)x-(\mathrm{Id}-T)y\|^2.$$

Note that if $\gamma(L+\mu)/2 \leq 1$, then the above inequality yields

$$\|Tx-Ty\| \leq \left(1-\frac{2\gamma\mu L}{\mu+L}\right)^{1/2}\|x-y\|, \tag{2.32}$$

where

$$0 < \frac{2\gamma\mu L}{L+\mu} \leq \frac{4\mu L}{(L+\mu)^2} - 1 + 1 = 1 - \left(\frac{L-\mu}{L+\mu}\right)^2 < 1.$$

Therefore, for every $\gamma \in]0, 2/(L+\mu)]$, T is a *contraction* with the constant given in (2.32). Since x_* is a fixed point of T, (2.29) follows from Theorem 2.2.2. The minimum value of $\big(1-2\gamma\mu L/(L+\mu)\big)$ is reached for γ maximum, that is, $\gamma = 2/(L+\mu)$. In such case, the constant becomes $\big(1-4\gamma\mu L/(L+\mu)^2\big)^{1/2} = (L-\mu)/(L+\mu)$. Finally, it follows from Theorem2.1.12(ii), with $x = x_*$ and $y = x_k$ that

$$f(x_k) - f(x_*) \leq \frac{L}{2}\|x_k - x_*\|^2$$

and (2.31) follows. □

Remark 2.3.7. Comparing the rates given in Theorem 2.3.6 and Theorem 2.3.3, we have

$$1-\frac{2\gamma\mu L}{L+\mu} \leq 1-\gamma\mu(2-\gamma L) \;\Leftrightarrow\; \gamma \in \left]\frac{1}{L}\frac{2\mu}{L+\mu}, \frac{2}{L+\mu}\right] \supset \left[\frac{1}{L}, \frac{2}{L+\mu}\right].$$

So, Theorem 2.3.6 improves the rates given in Theorem 2.3.3 on the interval $[1/L, 2/(L+\mu)]$, which includes the optimal choices $2/(\mu+L)$ and $1/L$.

If we additionally assume that the function f is twice differentiable the results can further improved.

Theorem 2.3.8 (Convergence 3)**.** *Let $f\colon X \to \mathbb{R}$ be twice differentiable and suppose that f is μ-strongly convex and that ∇f is L-Lipschitz continuous. Then, for every $\gamma > 0$, $T = \mathrm{Id} - \gamma\nabla f$ is Lipschitz continuous with constant*

$$q_3(\gamma) = \max\{|1-\gamma\mu|, |1-\gamma L|\} = \begin{cases} 1-\gamma\mu & \text{if } \gamma \leq \dfrac{2}{L+\mu} \\ \gamma L - 1 & \text{if } \gamma \geq \dfrac{2}{L+\mu}. \end{cases} \tag{2.33}$$

So, T is a contraction if $\gamma \in]0, 2/L[$. Therefore the gradient descent algorithm features the following rate of convergence

$$\|x_k - x_*\| \leq q_3(\gamma)^k \|x_0 - x_*\| \qquad \text{and} \qquad f(x_k) - f(x_*) \leq \frac{L}{2} q_3(\gamma)^{2k} \|x_0 - x_*\|^2 .$$

For the optimal stepsize $\gamma = 2/(L+\mu)$, we have

$$\|x_k - x_*\| \leq \left(\frac{L-\mu}{L+\mu}\right)^k \|x_0 - x_*\| \qquad \text{and} \qquad f(x_k) - f(x_*) \leq \frac{L}{2}\left(\frac{L-\mu}{L+\mu}\right)^{2k} \|x_0 - x_*\|^2 .$$

Proof. The mapping T is differentiable and $T'(x) = \mathrm{Id} - \gamma\nabla^2 f(x)$. By the mean value theorem

$$\forall\, x, y \in X \quad \|Tx - Ty\| \leq q\|x - y\| \;\Leftrightarrow\; \forall\, x \in X \quad \|T'(x)\| \leq q.$$

Moreover, $\|T'(x)\| = \sup_{\lambda \in \sigma(\nabla^2 f(x))} |1 - \gamma\lambda|$. Since f is μ strongly convex and ∇f is L-Lipschitz continuous,

$$(\forall\, x \in X)(\forall\, u \in X) \quad \mu\|u\|^2 \leq \langle \nabla f(x)u, u\rangle \leq L\|u\|^2 .$$

So $\sigma(\nabla^2 f(x)) \subset [\mu, L]$ and hence $\|T'(x)\| \leq \max_{\lambda\in[\mu,L]} |1-\gamma\lambda| = q_3(\gamma)$. This last equality follows by noting that $\lambda \mapsto |1-\gamma\lambda|$ is a piecewise convex function and hence it achieves its maximum at the end points of the interval $[\mu, L]$. It follows from (2.33) that $q(\gamma) < 1 \;\Leftrightarrow\; \gamma \in]0, 2/L[$ (see Figure 2.1). The inequalities on the values follow from Theorem 2.1.12(ii) with $y = x$ and $x = x_*$. □

Remark 2.3.9. The constant $q_3(\gamma)$ given in Theorem 2.3.8 is always better than the constant $q_2(\gamma)$ given in Theorem 2.3.6. However, on the minimum value they agree.

Lecture 3

Nonsmooth Differential Theory

Here we address the question of optimization problems of the form

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

where f is convex, but possibly nondifferentiable.

3.1 Directional derivatives and subdifferentials

Proposition 3.1.1. *Let $f\colon X \to]-\infty,+\infty]$ be a proper convex function. Let $x \in \mathrm{dom} f$ and $v \in X$. Then the function*

$$t \in \mathbb{R}_{++} \mapsto \frac{f(x+tv)-f(x)}{t} \in]-\infty,+\infty]$$

is increasing.

Proof. Let $0 < t < s$ and $y = x+sv$. Then $x+tv = x+(t/s)(y-x)$. Therefore $f(x+tv) \leq f(x) + (t/s)\big(f(y)-f(x)\big)$ and hence $(f(x+tv)-f(x))/t \leq \big(f(x+sv)-f(x)\big)/s$. □

Definition 3.1.2. Let $f\colon X \to]-\infty,+\infty]$ be a proper convex function. Let $x \in \mathrm{dom} f$ and $v \in X$. We set

$$f'(x,v) := \lim_{t\to 0^+} \frac{f(x+tv)-f(x)}{t} = \inf_{t>0} \frac{f(x+tv)-f(x)}{t} \in [-\infty,+\infty]. \tag{3.1}$$

and we call it the *directional derivative of f at x in direction v*.

Remark 3.1.3.

(i) We have

$$(\forall\, x,y \in \mathrm{dom} f) \quad f'(x,y-x) \leq f(y)-f(x) < +\infty. \tag{3.2}$$

This follows from the fact that the map $t \in]0,1] \mapsto \big(f(x+t(y-x))\big)/t$ is increasing.

(ii) $f'(x,v)$ can be equal to $-\infty$. Indeed let $f\colon \mathbb{R} \to]-\infty,+\infty]$, be such that $f(x) = -\sqrt{x}$, if $x \geq 0$, and $f(x) = +\infty$, if $x < 0$. Then $0 \in \mathrm{dom} f$ and $f'(0,1) = \lim_{t\to 0^+} f(t)/t = \lim_{t\to 0^+} -1/\sqrt{t} = -\infty$.

Salzo Saverio (saverio.salzo@iit.it) Istituto Italiano di Tecnologia. Via E. Melen, 16152, Genoa, Italy.

(iii) Let $f\colon \mathbb{R} \to]-\infty,+\infty]$ be a proper convex function and let $x \in \mathrm{dom} f$. Then $f'(x,1) = f'_+(x)$ is the right derivative of f at x and $-f'(x,-1) = f'_-(x)$ is the left derivative of f at x. So, convex functions of one real variable admits left and right derivatives at every points of their domain.

Proposition 3.1.4. *Let $f\colon X \to]-\infty,+\infty]$ be a proper convex function and let $x \in \mathrm{dom} f$. Then $f'(x,\cdot)\colon X \to [-\infty,+\infty]$ is sublinear (and hence a fortiori convex), i.e.,*

(i) $(\forall\, v \in X)(\forall\, \lambda \in \mathbb{R}_+)\ f'(x,\lambda v) = \lambda f'(x,v)$;

(ii) $(\forall\, v \in X)\ -f'(x,-v) \le f'(x,v)$;

(iii) $(\forall\, u,v \in X)\ \max\{f'(x,u), f'(x,v)\} < +\infty \ \Rightarrow\ f'(x,u+v) \le f'(x,u) + f'(x,v)$.

(iv) *If $x \in \mathrm{int}(\mathrm{dom} f)$, then for every $v \in X$, $f'(x,v) \in \mathbb{R}$.*

Proof. (i): It is immediate.

(ii). Since $x = ((x-tv)+(x+tv))/2$, by convexity we have $2f(x) \le f(x-tv)+f(x+tv)$ and hence

$$-\frac{f(x-tv)-f(x)}{t} \le \frac{f(x+tv)-f(x)}{t}.$$

Since $(f(x+tv)-f(x))/t \to f'(x,v)$ and $(f(x-tv)-f(x))/t \to f'(x,-v)$, the statement follows.

(iii): Suppose that $\max\{f'(x,u), f'(x,v)\} < +\infty$. Then, it is easy to prove that, for every $\alpha \in [0,1]$, $f'(x,(1-\alpha)u+\alpha v) \le (1-\alpha)f'(x,u) + \alpha f'(x,v)$. Then, by (i), $f'(x,u+v) = f'(x,2[(1/2)u+(1/2)v]) = 2f'(x,(1/2)u+(1/2)v) \le f'(x,u)+f'(x,v)$.

(iv): Suppose that $x \in \mathrm{int}(\mathrm{dom} f)$ and let $\delta > 0$ such that $B_\delta(x) \subset \mathrm{dom} f$. Let $v \in X$, $v \neq 0$. Then we can find $t > 0$ such that $y = x + tv \in B_\delta(x)$. Hence $f'(x,v) \le (f(x+tv)-f(x))/t = (f(y)-f(x))/t < +\infty$. Therefore, picking $-v$ we have also $f'(x,-v) < +\infty$, and finally $-\infty < -f'(x,-v) \le f'(x,v) < +\infty$. □

The previous proposition establishes that the directional derivative of a convex function is sublinear. Note that if f is differentiable, for every $v \in X$, $f'(x,v) = \langle \nabla f(x), v\rangle$ and hence in this case the directional derivative is linear.

Definition 3.1.5. Let $f\colon X \to]-\infty,+\infty]$ be a proper convex function and $x \in \mathrm{dom} f$. We set

$$\partial f(x) := \big\{u \in X \,\big|\, (\forall\, y \in X)\ f(y) \ge f(x) + \langle u, y-x\rangle\big\} \tag{3.3}$$

and we call it the *subdifferential of f at x*. Every element of $\partial f(x)$ is called a *subgradient of f at x*. If $x \notin \mathrm{dom} f$, we set $\partial f(x) = \varnothing$. Finally, the *domain* of ∂f, denoted by $\mathrm{dom}\partial f$, is defined as the set of points at which the subdifferential is nonempty.

Remark 3.1.6. It is easy to see that the subdifferential ∂f is a *monotone* operator, that is, for every $x,y \in X$ and $u \in \partial f(x)$, $v \in \partial f(v)$ $\langle x-y, u-v\rangle \ge 0$.

The following result gives the geometrical meaning of subgradients.

Proposition 3.1.7. *Let $f\colon X \to]-\infty,+\infty]$ be a proper convex function and $x \in \mathrm{dom} f$. Then $u \in \partial f(x)$ if and only if* $\mathrm{epi}(f)$ *is containded in the negative half-space determined by the hyperplane of $X \times \mathbb{R}$ orthogonal to the vector $(u,-1)$ and passing through $(x,f(x))$. See Figure 3.1.*

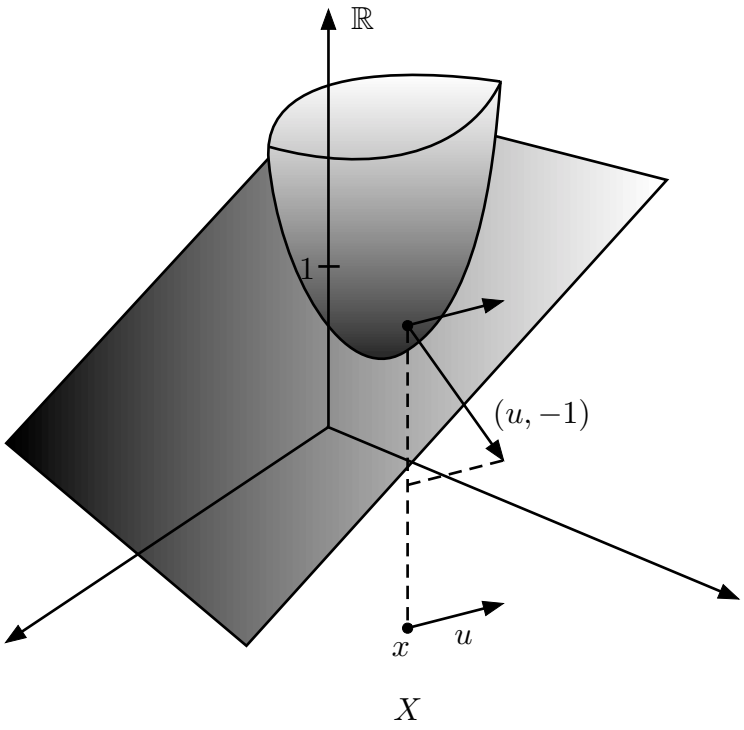


Figure 3.1: The subgradient u and the corresponding hyperplane.

Proof. Let $u \in X$. Then

$$H^-_{(u,-1)} = \big\{(y,t) \in X \times \mathbb{R} \,\big|\, \langle (u,-1), (y,t) - (x, f(x)) \rangle \leq 0\big\}$$

is the negative half-plane determined by the hyperplane of $X \times \mathbb{R}$ orthogonal to $(u,-1)$ and passing through $(x, f(x))$. Clearly $(y,t) \in H^-_{(u,-1)} \;\Leftrightarrow\; \langle u, y - x\rangle + f(x) \leq t$. Therefore,

$$\begin{aligned} u \in \partial f(x) &\;\Leftrightarrow\; (\forall\, y \in \mathrm{dom} f)\ \langle u, y - x\rangle + f(x) \leq f(y) \\ &\;\Leftrightarrow\; (\forall\, (y,t) \in \mathrm{epi}(f))\ \langle u, y - x\rangle + f(x) \leq t \\ &\;\Leftrightarrow\; \mathrm{epi}(f) \subset H^-_{(u,-1)}. \end{aligned}$$

□

The following result characterizes the subdifferential for functions of one real variable.

Proposition 3.1.8. *Let $f\colon \mathbb{R} \to \,]-\infty, +\infty]$ be a proper convex function and $x \in \mathrm{dom} f$. Then the following holds.*

(i) *For every $x \in \mathrm{dom} f$, $f'_-(x) \leq f'_+(x)$ and $\partial f(x) = [f'_-(x), f'_+(x)] \cap \mathbb{R}$.*

(ii) *For every x and y in $\mathrm{dom} f$ such that $x < y$, we have $f'_+(x) \leq f'_-(y)$.*

(iii) *The functions f'_- and f'_+ are increasing.*

(iv) *f is differentiable on $\mathrm{dom} f$ except for an at most countable number of points.*

Proof. (i): Let $x \in \mathrm{dom} f$. For every $t > 0$, since $x = \big((x-t)+(x+t)\big)/2$, by convexity, we have

$$f(x) \leq \frac{1}{2}\big(f(x-t) + f(x+t)\big)$$

and hence $-\big(f(x-t) - f(x)\big)/t \leq \big(f(x+t) - f(x)\big)/t$. Taking the limit as $t \to 0^+$ we get $f'_-(x) \leq f'_+(x)$. Now, by the true definition, it follows

$$\begin{aligned} u \in \partial f(x) &\Leftrightarrow (\forall\, t \in \mathbb{R})\ f(x+t) - f(x) \geq tu \\ &\Leftrightarrow \begin{cases} (f(x+t) - f(x))/t \geq u & \forall\, t > 0 \\ (f(x-t) - f(x))/(-t) \leq u & \forall\, t > 0. \end{cases} \end{aligned}$$

Now we note that by Proposition 3.1.1, $t \mapsto (f(x+t) - f(x))/t$ is increasing on $\mathbb{R}_{++}$ and $t \mapsto (f(x-t) - f(x))/(-t) = -(f(x+t(-1)) - f(x))/t$ is decreasing on $\mathbb{R}_{++}$. Therefore, the previous inequalities are equivalent to $f'_+(x) \geq u$ and $f'_-(x) \leq u$.

(ii): The statement follows from Proposition 3.1.4(ii) and Remark 3.1.3(iii).

(iii): It follows from (i) and (ii).

(iv): We note that $\mathrm{dom} f$ is a convex subset of $\mathbb{R}$ and hence it is an interval. Moreover, by Proposition 3.1.4(iv) and Remark 3.1.4(iii) we have that f'_- and f'_+ are finite on $\mathrm{int}(\mathrm{dom} f)$. We also note that if $x, y \in \mathrm{int}(\mathrm{dom} f)$ with $x < y$ then $f'_-(x) \leq f'_+(x) \leq f'_-(y)$. Therefore if f'_- is continuous at x, we have that $f'_-(x) = f'_+(x)$ and hence f is differentiable at x. So the set of points of nondifferentiability of f is contained in the set of points of discontinuity of f'_-. Now, since f'_- is increasing on $\mathrm{int}(\mathrm{dom} f)$ it has at most a countable number of discontinuities and the statement follows. □

Proposition 3.1.9. *Let $f\colon X \to]-\infty, +\infty]$ be a proper convex function and $x \in \mathrm{dom} f$. Then $\partial f(x)$ is a closed convex set.*

Proof. We have

$$\begin{aligned} u \in \partial f(x) &\Leftrightarrow (\forall\, y \in \mathrm{dom} f)\ \langle u, y-x\rangle \leq f(y) - f(x) \\ &\Leftrightarrow u \in \bigcap_{y \in \mathrm{dom} f} [\langle \cdot, y-x\rangle \leq f(y) - f(x)], \end{aligned}$$

which shows that $\partial f(x)$ is an intersection of (closed) affine half-spaces. □

Remark 3.1.10. It is easy to provide a function which is not subdifferentiable at one point of $\mathrm{dom} f$. For instance, for the function f defined in Remark 3.1.3(ii), we have $0 \in \mathrm{dom} f$, but $\partial f(0) = \varnothing$.

Proposition 3.1.11. *Let $f \in \Gamma_0(X)$. Then $\mathrm{dom}\partial f$ is dense in $\mathrm{dom} f$.*

Proof. Let $x \in \mathrm{dom} f$ and let $\xi < f(x)$. So, $\mathrm{epi}(f)$ is a closed convex set and $(x, \xi) \notin \mathrm{epi}(f)$. Let $(p, \pi) = P_{\mathrm{epi}(f)}(x, \xi)$ (see Definition 4.1.1). Then, it follows from Proposition 4.1.2 that

$$(\forall\, y \in \mathrm{dom} f)(\forall\, \eta \in [f(y), +\infty[) \quad \langle y-p, x-p\rangle + (\eta - \pi)(\xi - \pi) \leq 0. \tag{3.4}$$

Since η can be arbitrarily large, the above inequality necessarily implies that $\xi \leq \pi$. Now, since $(p, \pi) \in \operatorname{epi}(f)$, we have that $p \in \operatorname{dom} f$ and $f(p) \leq \pi$. Hence it follows from (3.4) with $y = p$ and $\eta = f(p)$, that

$$(\underbrace{f(p) - \pi}_{\leq 0})(\underbrace{\xi - \pi}_{\leq 0}) \leq 0.$$

Therefore, $f(p) = \pi$ or $\xi = \pi$. If it was $\xi = \pi$, then it would follows from (3.4), with $y = x$, that $\|x - p\|^2 = 0$, which gives $x = p$ and hence, in the end, $(p, \pi) = (x, \xi)$, which is a contradiction since $(x, \xi) \notin \operatorname{epi}(f)$ while $(p, \pi) \in \operatorname{epi}(f)$. Thus, necessarily $f(p) = \pi$. Finally, it follows from (3.4) with $\eta = f(y)$ that

$$(\forall\, y \in \operatorname{dom} f) \quad \langle y - p, x - p \rangle \leq (f(y) - f(p))(f(p) - \xi),$$

which implies, since $\xi < \pi = f(p)$, that $(f(p) - \xi)^{-1}(x - p) \in \partial f(p)$, so $p \in \operatorname{dom} \partial f$. Moreover, it follows from (3.4) with $y = x$ and $\eta = f(x)$, and the fact that $\xi < f(p)$, that

$$\begin{aligned} 0 \leq \|x - p\|^2 + (f(x) - f(p))^2 &\leq (f(x) - f(p))(\underbrace{f(x) - \xi}_{\geq 0}) \\ &\leq (f(x) - \xi)^2. \end{aligned}$$

Since ξ can be arbitrarily close to $f(x)$ the statement follows. □

Remark 3.1.12. Let $f \in \Gamma_0(X)$. Since, in virtue of Proposition 3.1.11, $\operatorname{cl}(\operatorname{dom} \partial f) \supset \operatorname{dom} f \neq \varnothing$, we have that $\operatorname{dom} \partial f \neq \varnothing$ and hence there exists $x \in \operatorname{dom} f$ and $u \in \partial f(x)$ such that $f \geq f(x) + \langle \cdot - x, u \rangle$. This shows that f admits an affine minimizer. We well reach the same conclusion later in Section 4.4 as a consequence of the Fenchel-Moreau Theorem 4.4.6.

Example 3.1.13. Let $|||\cdot|||$ be a norm on X. Then the dual norm is defined as

$$|||u|||_* = \sup_{|||x||| \leq 1} \langle x, u \rangle. \tag{3.5}$$

and, for every $x \in X$, we have

$$\partial\, |||\cdot|||\, (x) = \left\{ u \in X \;\middle|\; |||u|||_* \leq 1 \text{ and } \langle x, u \rangle = |||x||| \right\} \tag{3.6}$$

$$= \begin{cases} \left\{ u \in X \;\middle|\; |||u|||_* = 1 \text{ and } \langle x, u \rangle = |||x||| \right\} & \text{if } x \neq 0. \\ \left\{ u \in X \;\middle|\; |||u|||_* \leq 1 \right\} & \text{if } x = 0. \end{cases} \tag{3.7}$$

Indeed, by definition, $u \in \partial\, |||\cdot|||\, (x)$ is equivalent to

$$(\forall\, y \in X) \;\; |||y||| \geq |||x||| + \langle y - x, u \rangle, \tag{3.8}$$

which in turn is equivalent to

$$|||x||| = \langle x, u \rangle \;\text{ and }\; (\forall\, y \in X)\; \langle y, u \rangle \leq |||y|||\,. \tag{3.9}$$

This follows by taking $y = 0$ and $y = 2x$ in (3.8). Then the latter in (3.9) says that $\langle \cdot, u\rangle$ is continuous with respect to $|||\cdot|||$ and $|||u|||_* \leq 1$, so (3.6) follows. As regards (3.7), it is sufficient to note that by the first of (3.9) and (3.5) we have $|||x||| = \langle x, u\rangle \leq |||u|||_* \, |||x|||$. So, when $x \neq 0$, we have also $1 \leq |||u|||_*$.

We finally note that if $|||\cdot||| = \|\cdot\|$ (the Euclidean norm on X), and $u \in \partial \|\cdot\| (x)$ and $x \neq 0$, it follows from (3.7) that $\|u - x/\|x\|\|^2 = \|u\|^2 + 1 - 2\langle x, u\rangle/\|x\| = 0$. Therefore,

$$\partial \|\cdot\| (x) = \begin{cases} \dfrac{x}{\|x\|} & \text{if } x \neq 0. \\ B_1(0) & \text{if } x = 0. \end{cases} \tag{3.10}$$

Theorem 3.1.14 (Fermat's rule)**.** *Let $f\colon X \to]-\infty, +\infty]$ be a proper convex function and $x \in \mathrm{dom} f$. Then the following are equivalent*

(i) *x is a minimizer of f;*

(ii) $0 \in \partial f(x)$*;*

(iii) $(\forall\, y \in X)\ f'(x, y - x) \geq 0$*;*

(iv) $(\forall\, y \in \mathrm{dom} f)\ f'(x, y - x) \geq 0$.

Proof. (i) $\Leftrightarrow$ (ii): "x is minimizer of f" means that $(\forall\, y \in X)\ f(y) \geq f(x)$ and this is nothing but $0 \in \partial f(x)$.

(i) $\Rightarrow$ (iii): Let $y \in X$. Since x is a minimizer, for every $t > 0$, we have

$$0 \leq \frac{f(x + t(y - x)) - f(x)}{t}.$$

Then $f'(x, y - x) = \inf_{t>0} \big(f(x + t(y - x)) - f(x)\big)/t \geq 0$.

(iv) $\Rightarrow$ (i): It follows from (3.2) that, for every $y \in \mathrm{dom} f$, $0 \leq f(y) - f(x)$ and hence the statement. □

Definition 3.1.15. Let $C \subset X$ be a nonempty convex set and let $x \in C$. The set $\partial \iota_C(x)$ is called the *normal cone to C at x* and it is also denoted by $N_C(x)$, that is,

$$N_C(x) = \big\{u \in X \,|\, (\forall\, y \in C)\ \langle u, y - x\rangle \leq 0\big\}. \tag{3.11}$$

Remark 3.1.16.

(i) $N_C(x)$ is a closed convex cone and $0 \in N_C(x)$. Indeed, it is closed convex by Proposition 3.1.9 and $\lambda \in \mathbb{R}_+$ and $u \in N_C(x)$ implies $\lambda u \in N_C(x)$.

(ii) $u \in N_C(x) \ \Leftrightarrow\ C \subset H^-_{x,u}$, where $H^-_{x,u} = \big\{y \in X \,|\, \langle y - x, u\rangle \leq 0\big\}$ is the negative closed half-space determined by the closed hyperplane orthogonal to u and passing through x. See Figure 3.3.

(iii) In view of the previous point and Proposition 3.1.7, we have $u \in \partial f(x) \ \Leftrightarrow\ (u, -1) \in N_{\mathrm{epi}(f)}(x, f(x))$. Note, moreover, that if $(u, s) \in N_{\mathrm{epi}(f)}(x, f(x))$, then necessarily $s \leq 0$. Indeed, if $(u, s) \in N_{\mathrm{epi}(f)}(x, f(x))$, we have that, for every $(y, t) \in \mathrm{epi}(f))$, $\langle (y, t) - (x, f(x)), (u, s)\rangle \leq 0$. Hence, if we let $t > f(x)$, since $(x, t) \in \mathrm{epi}(f)$, we have $(t - f(x))s \leq 0$, which implies that $s \leq 0$.

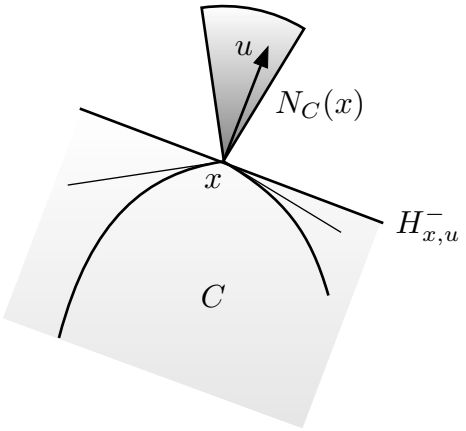


Figure 3.2: The normal cone to C at x and a normal vector u.

Proposition 3.1.17. *Let $f\colon X \to]-\infty,+\infty]$ be a proper convex function and $x \in \operatorname{dom} f$. Then*

$$\partial f(x) = \big\{u \in X \,\big|\, \langle u, \cdot\rangle \leq f'(x,\cdot)\big\}. \tag{3.12}$$

Proof. For every $u \in X$, we have

$$\begin{aligned} u \in \partial f(x) &\;\Leftrightarrow\; (\forall\, y \in X)\ \langle u, y-x\rangle \leq f(y) - f(x) \\ &\;\Leftrightarrow\; (\forall\, v \in X)(\forall\, t>0)\ t\langle u, v\rangle \leq f(x+tv) - f(x) \\ &\;\Leftrightarrow\; (\forall\, v \in X)\ \langle u, v\rangle \leq f'(x,v), \end{aligned}$$

since $f'(x,v) = \inf_{t>0}\big(f(x+tv)-f(x)\big)/t$. □

Remark 3.1.18. Since $f'(x,0) = 0$, it follows from (3.12) that

$$u \in \partial f(x) \;\Leftrightarrow\; (\forall\, v \in X)\ f'(x,v) \geq f'(x,0) + \langle u, v\rangle \;\Leftrightarrow\; u \in \partial f'(x,\cdot)(0).$$

Hence $\partial f(x) = \partial f'(x,\cdot)(0)$.

Proposition 3.1.19. *Let $f\colon X \to]-\infty,+\infty]$ be a proper convex function and let $x \in \operatorname{int}(\operatorname{dom} f)$. Suppose that f is Gâteaux differentiable at x. Then $\partial f(x) = \{\nabla f(x)\}$.*

Proof. Since f is Gâteaux differentiable we have $\langle \nabla f(x), \cdot\rangle = f'(x,\cdot)$. Therefore, it follows from Proposition 3.1.17 that $u \in \partial f(x) \;\Leftrightarrow\; \langle u, \cdot\rangle \leq \langle \nabla f(x), \cdot\rangle \;\Leftrightarrow\; \langle u - \nabla f(x), \cdot\rangle \leq 0 \;\Leftrightarrow\; u = \nabla f(x)$. □

Lemma 3.1.20. *Let $H = \{x \in X \,|\, \langle u, x\rangle = \alpha\} \times \mathbb{R}$ be a vertical hyperplane of $X \times \mathbb{R}$ and suppose that $\operatorname{epi}(f) \subset H^-$ for some proper function $f\colon X \to]-\infty,+\infty]$. Then $\{x \in X \,|\, \langle u, x\rangle = \alpha\} \cap \operatorname{int}(\operatorname{dom} f) = \varnothing$.*

Proof. Clearly $\operatorname{dom} f \subset \{x \in X \,|\, \langle u, x\rangle \leq \alpha\}$. Then, it follows from (4.10) that $\operatorname{int}(\operatorname{dom} f) \subset \{x \in X \,|\, \langle u, x\rangle < \alpha\}$ and the statement follows. □

3.2 Subdifferential calculus

In this section we discuss some calculus rules for the subdifferential of convex functions. In Section 8.1 additional results will be given.

Proposition 3.2.1. *Let $f\colon X \to]-\infty,+\infty]$ and $g\colon Y \to]-\infty,+\infty]$ be two proper and convex functions and let $A\colon X \to Y$ be a bounded operator between Euclidean spaces such that* $\operatorname{dom} g \cap A(\operatorname{dom} f) \neq \varnothing$. *Then*

$$(\forall\, x \in X) \qquad \partial f(x) + A^* \partial g(Ax) \subset \partial(f + g \circ A)(x). \tag{3.13}$$

Proof. Let $x \in X$, $u \in \partial f(x)$ and $v \in \partial g(Ax)$. Then, for all $y \in X$,

$$\begin{aligned} f(y) &\geq f(x) + \langle y - x, u\rangle \\ g(Ay) &\geq g(Ax) + \langle Ay - Ax, v\rangle \end{aligned}$$

and hence $f(y) + g(Ay) \geq f(x) + g(Ax) + \langle y - x, u + A^*v\rangle$. Therefore, $u + A^*v \in \partial(f + g \circ A)(x)$. □ □

In general, in (3.13) equality does not hold, as the following example shows.

Example 3.2.2. Let $f, g\colon \mathbb{R} \to]-\infty,+\infty]$ be given by

$$f(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ +\infty & \text{if } x > 0 \end{cases} \qquad \text{and} \qquad g(x) = \begin{cases} +\infty & \text{if } x < 0 \\ -\sqrt{x} & \text{if } x \geq 0. \end{cases}$$

Then, we have

$$\partial f(x) = \begin{cases} \{0\} & \text{if } x < 0 \\ [0,+\infty[& \text{if } x = 0 \\ \varnothing & \text{if } x > 0 \end{cases} \qquad \text{and} \qquad \partial g(x) = \begin{cases} \varnothing & \text{if } x \leq 0 \\ -\frac{1}{2\sqrt{x}} & \text{if } x > 0. \end{cases}$$

Therefore, $\partial f(x) + \partial g(x) = \varnothing$ for all $x \in \mathbb{R}$. On the other hand, $f + g = \iota_{\{0\}}$, which implies $\partial(f + g)(0) = \mathbb{R}$. We see that $\partial(f + g)(x)$ may differ from $\partial f(x) + \partial g(x)$.

The following proposition gives a first situation in which equality holds. A much more general result will be given in Section 8.1 (Corollary 8.1.9).

Proposition 3.2.3. *In Proposition 3.2.1, suppose that g is Gâteaux differentiable at* $Ax \in \operatorname{int}(\operatorname{dom} g)$. *Then* $\partial(f + g \circ A)(x) = \partial f(x) + \{A^* \nabla g(Ax)\}$.

Proof. In view of Proposition 3.2.1, we need only to prove that $\partial(f+g\circ A)(x) \subset \partial f(x)+ \{A^*\nabla g(Ax)\}$, that is, that $\partial(f + g \circ A)(x) - A^*\nabla g(Ax) \subset \partial f(x)$. Let $u \in \partial(f + g \circ A)(x)$. Let $y \in X$ and $\varepsilon > 0$. Then, since $\big(g(Ax + t(Ay - Ax)) - g(Ax)\big)/t \to \langle Ay - Ax, \nabla g(Ax)\rangle = \langle y - x, A^*\nabla g(Ax)\rangle$, there exists $t \in]0, 1[$, sufficiently small, such that

$$g(Ax + t(Ay - Ax)) - g(Ax) \leq \langle t(y - x), A^*\nabla g(Ax)\rangle + \varepsilon t. \tag{3.14}$$

Set $z = x + t(y-x)$. Since $u \in \partial(f + g \circ A)(x)$, we have

$$\begin{aligned} f(z) + g(Az) &\geq f(x) + g(Ax) + \langle z - x, u \rangle \\ &= f(x) + \langle z - x, u - A^* \nabla g(Ax) \rangle + g(Ax) + \langle t(y-x), A^* \nabla g(Ax) \rangle \\ &\geq f(x) + \langle z - x, u - A^* \nabla g(Ax) \rangle + g(Az) - \varepsilon t. \end{aligned}$$

Thus, using the convexity of f, we have

$$f(x) + t\big(f(y) - f(x)\big) \geq f(z) \geq f(x) + \langle t(y-x), u - A^* \nabla g(Ax) \rangle - \varepsilon t$$

and dividing by t, we obtain

$$f(y) - f(x) \geq \langle y - x, u - A^* \nabla g(Ax) \rangle - \varepsilon.$$

Since ε was taken arbitrarily, we get $f(y) - f(x) \geq \langle u - A^* \nabla g(Ax), y - x \rangle$. So we proved that $u - A^* \nabla g(Ax) \in \partial f(x)$. □ □

Corollary 3.2.4. *Let $f \colon X \to]-\infty, +\infty]$ be proper and convex. Then the following hold.*

(i) *For every $v \in X$, $\partial(f + \langle \cdot, v \rangle)(x) = \partial f(x) + v$;*

(ii) *If $g \colon X \to]-\infty, +\infty]$ is Gâteaux differentiable at $x \in \operatorname{int}(\operatorname{dom} g)$ and $\operatorname{dom} g \cap \operatorname{dom} f \neq \varnothing$, then $\partial(f+g)(x) = \partial f(x) + \{\nabla g(x)\}$.*

Proposition 3.2.5. *Let $(X_i)_{1 \leq i \leq m}$ be m Euclidean spaces and let $X = \bigoplus_{i=1}^m X_i$ be their direct product, endowed with the scalar product $\langle x, y \rangle = \sum_{i=1}^m \langle x_i, y_i \rangle$. Let $(f_i)_{1 \leq i \leq m}$ be a family of proper convex functions, $f_i \colon X_i \to]-\infty, +\infty]$ and define*

$$f \colon X \to]-\infty, +\infty], \qquad f(x) = f_1(x_1) + f_2(x_2) + \cdots, + f_m(x_m).$$

So the function f is separable. Then, for every $x \in \operatorname{dom} f = \prod_{i=1}^m \operatorname{dom} f_i$, we have

$$\partial f(x) = \partial f_1(x_1) \times \partial f_2(x_2) \times \cdots \times \partial f_m(x_m).$$

Proof. It follows directly from Definition 3.1.5. □ □

Example 3.2.6. Let us consider the case of the ℓ^1-norm on $\mathbb{R}^n$. Since $\|\cdot\|_1$ is clearly separable with components $|\cdot|$, it follows from Proposition 3.2.5 that

$$\partial \|\cdot\|_1 = \partial|\cdot|(x_1) \times \cdots \times |\cdot|(x_n).$$

3.3 Subdifferentials, strict convexity and strong convexity

Proposition 3.3.1. *Let $f \colon X \to]-\infty, +\infty]$ be proper convex. Then the following hold.*

(i) *Suppose that f is lower siemicontinuous and* supercoercive, *i.e., $\|f(x)\| / \|x\| \to +\infty$ as $\|x\| \to +\infty$. Then $\partial f(X) = X$.*

(ii) *Suppose that f is strictly convex. Then, for every $x, y \in \operatorname{dom} f$ such that $x \neq y$, we have $\partial f(x) \cap \partial f(y) = \varnothing$.*

Thus, if f is finite, strictly convex, supercoercive, and Gâteaux differentiable, then the gradient $\nabla f \colon X \to X$ is a bijection.

Proof. (i): Let $u \in X$. Then the function $f - \langle \cdot, u \rangle$ is proper, lower semicontinuous, and coercive, hence, in virtue of Theorem 1.5.4, it has a minimizer, say x. Therefore, it follows from Corollary 3.2.4(i) that $0 \in \partial(f - \langle \cdot, u \rangle)(x) = \partial f(x) - u$ and hence $u \in \partial f(x)$.

(ii): Let x and y in $\operatorname{dom} f$ and suppose that $u \in \partial f(x) \cap \partial f(y)$. Then, again from Corollary 3.2.4(i) we derive that $0 \in \partial(f - \langle \cdot u, \rangle)(x)$ and $0 \in \partial(f - \langle \cdot, u \rangle)(y)$. Therefore, x and y are minimizers of $f - \langle \cdot, u \rangle$. Since $f - \langle \cdot, u \rangle$ is strictly convex, Theorem 1.5.5 yields $x = y$. □

Proposition 3.3.2. *Let $f \colon X \to]-\infty, +\infty]$ be a proper function and suppose that it is μ-strongly convex, for some $\mu > 0$. Let $x \in \operatorname{dom} f$. Then*

$$u \in \partial f(x) \;\Leftrightarrow\; (\forall\, y \in X) \quad f(y) \geq f(x) + \langle y - x, u \rangle + \frac{\mu}{2} \|y - x\|^2 . \tag{3.15}$$

Moreover, we have

$$f(x) - \inf f \leq \frac{1}{2\mu} \|\partial f(x)\|_-^2 , \tag{3.16}$$

where, $\|\partial f(x)\|_- = \inf_{u \in \partial f(x)} \|u\|$.

Proof. Concerning the equivalence (3.15), it is sufficient to prove the implication "⇒". It follows from Proposition 1.3.12 that there exists a proper convex function $g \colon X \to]-\infty, +\infty]$ such that $f = (\mu/2) \|\cdot\|^2 + g$. Let $u \in \partial f(x)$. Then it follows from Corollary 3.2.4(ii) that $u - \mu x \in \partial g(x)$ and hence

$$(\forall\, y \in X) \quad f(y) - \frac{\mu}{2} \|y\|^2 \geq f(x) - \frac{\mu}{2} \|x\|^2 + \langle y - x, u - \mu x \rangle. \tag{3.17}$$

After simple algebraic computations, (3.15) follows. As regards the second part, if $\partial f(x) = \varnothing$ we have $\inf_{u \in \partial f(x)} \|u\| = +\infty$ and the statement follows. So, we assume $\partial f(x) \neq \varnothing$ and let $u \in \partial f(x)$. Then, it follows from (3.15) that

$$\begin{aligned} \inf_{y \in \operatorname{dom} f} f(y) &\geq \inf_{y \in \operatorname{dom} f} \Big(f(x) + \langle y - x, u \rangle + \frac{\mu}{2} \|y - x\|^2 \Big) \\ &= f(x) + \inf_{y \in \operatorname{dom} f} \frac{1}{2\mu} \big(\underbrace{\|\mu(y - x) + u\|^2}_{\geq 0} - \|u\|^2 \big) \\ &\geq f(x) - \frac{1}{2\mu} \|u\|^2 . \end{aligned}$$

Since u is arbitrary in $\partial f(x)$, the statement follows. □ □

Proposition 3.3.3. *Let $f \colon X \to]-\infty, +\infty]$ be a proper function and suppose that*

$$(\forall\, x, y \in \operatorname{dom} \partial f)(\forall\, u \in \partial f(x)) \quad f(y) \geq f(x) + \langle y - x, u \rangle + \frac{\mu}{2} \|y - x\|^2 . \tag{3.18}$$

for some $\mu > 0$. Then f is strongly convex with modulus μ on every convex subset of $\operatorname{dom} \partial f$.

Proof. Let $x, y \in \mathrm{dom}\partial f$ and let $\lambda \in [0,1]$ be such that $z := (1-\lambda)x + \lambda y \in \mathrm{dom}\partial f$. Let $w \in \partial f(z)$. Then we have

$$f(x) \geq f(z) + \langle x - z, w\rangle + \frac{\mu}{2}\|x - z\|^2$$
$$f(y) \geq f(z) + \langle y - z, w\rangle + \frac{\mu}{2}\|y - z\|^2.$$

Now since $x - z = \lambda(x - y)$ and $y - z = (1-\lambda)(y - x)$, it follows from the previous inequalities that

$$(1-\lambda)f(x) \geq (1-\lambda)f(z) + (1-\lambda)\lambda\langle x - y, w\rangle + \frac{\mu}{2}(1-\lambda)\lambda^2\|x - y\|^2$$
$$\lambda f(y) \geq \lambda f(z) + \lambda(1-\lambda)\langle y - x, w\rangle + \frac{\mu}{2}\lambda(1-\lambda)^2\|y - x\|^2.$$

Summing the inequalities above we get $(1-\lambda)f(x)+\lambda f(y) \geq f((1-\lambda)x+\lambda y)+(\mu/2)\lambda(1-\lambda)[\lambda+1-\lambda]\|x-y\|^2$. Thus, the statement follows by recalling Definition 1.3.10. □

3.4 The subgradient projection method

In this section we consider a real valued convex function $f\colon X \to \mathbb{R}$. Then by Theorem 1.4.2 and Theorem 4.5.1, we have that f is continuous and $\partial f(x) \neq \varnothing$ for every $x \in X$. The subgradient method was devised by Shor in 1960s (and then studied by Polyak, Bertsekas, Demyanov), and is the natural generalization of the gradient descent method to nonsmooth functions:

$$x_{k+1} = x_k - \gamma_k u_k, \quad \gamma_k > 0 \text{ and } u_k \in \partial f(x_k). \tag{3.19}$$

To implement the method the computation of the entire subdifferential is not needed. The analysis however does not work as in the smooth case. First, the subgradient method is not a descent method, as the following example shows.

Example 3.4.1. Let $f\colon \mathbb{R}^2 \to \mathbb{R}$, $f(x_1, x_2) = |x_1| + 2|x_2|$. Then

$$\partial f(1,0) = \{1\} \times [-2, 2],$$

and not all the opposte elements in $\partial f(0,1)$ are descent directions, for instance $-(1,2)$ is not.

Remark 3.4.2. One could think to generalize the method of steepest descent to nonsmooth functions, since in the subdifferential there are descent directions, e.g. the element of minimal norm. Indeed, let v be the steepest descent direction at x, then, by Theorem 4.5.1,

$$\min_{\|v\|=1} f'(x, v) = \min_{\|v\|=1} \max_{u \in \partial f(x)} \langle u, v\rangle = \max_{u \in \partial f(x)} -\|u\| = -\min_{u \in \partial f(x)} \|u\| < 0.$$

However, the implementation of the steepest descent method, namely

$$\begin{cases} \bar{u}_k \in \mathrm{argmin}_{u \in \partial f(x_k)} \|u\| \\ \gamma_k \in \mathrm{argmin}_{\gamma > 0} f(x_k - \gamma_k \bar{u}_k), \\ x_{k+1} = x_k - \gamma_k \bar{u}_k, \end{cases}$$

is not convergent, as the following example (due to Wolfe), shows.

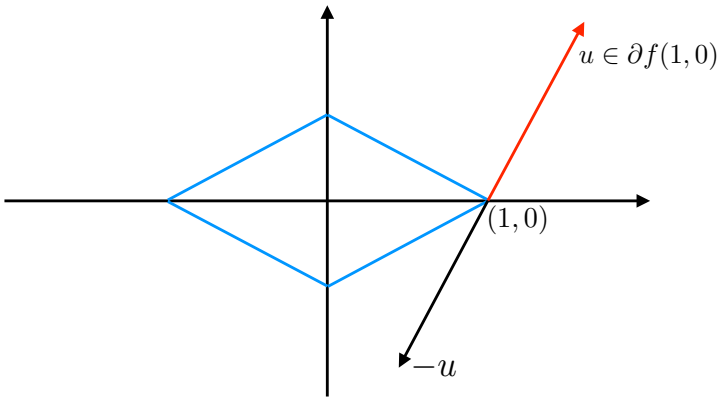


Figure 3.3: Level set and a subgradient which is not a descent direction for $f(x_1, x_2) = |x_1| + 2|x_2|$.

Example 3.4.3. Define

$$f(x_1, x_2) = \begin{cases} 5(9x_1^2 + 16x_2^2)^{1/2} & x_1 > |x_2| \\ 9x_1 + 16|x_2| & x_1 \leq |x_2|. \end{cases}$$

Then if (x_1^0, x_2^0) is such that $x_1^0 > |x_2^0| > (9/16)^2 |x_1^0|$, the steepest descent goes to 0, which is not a minimizer (and 0 is not the minimum).

Remark 3.4.4. (On the stepsize) The constant stepsize is not allowed. For example, if $f = \|\cdot\|$, $u \in \partial f(x)$, with $x \neq 0$ implies that $\|u\| = 1$. Therefore

$$\|x_{k+1} - x_k\| = \|x_k - \gamma u_k - x_k\| = \gamma \|u_k\| = \gamma,$$

and $(x_k)_{k\in\mathbb{N}}$ is not convergent.

The key point to ensure convergence of the method is the choice of the stepsize. There are several options that will be explained in the rest of the section.

In the following we consider a generalization of the subgradient method which is the subgradient projection method. We assume that $C \subset X$ is a nonempty, closed and convex set and $f\colon X \to \mathbb{R}$ is a convex and L-Lipschitz continuous function.

Algorithm 3.4.5 (Subgradient Projection)**.** Let $x_0 \in X$ and define, for every $k \in \mathbb{N}$,

$$x_{k+1} = P_C(x_k - \gamma_k u_k), \quad \gamma_k > 0 \text{ and } u_k \in \partial f(x_k).$$

Define, for every $k \in \mathbb{N}$,

$$f_k = \min_{0\leq i\leq k} f(x_i) \quad \text{and} \quad \bar{x}_k = \Big(\sum_{i=0}^{k} \gamma_i\Big)^{-1} \sum_{i=0}^{k} \gamma_i x_i.$$

Lemma 3.4.6. *Let $(a_k)_{k\in\mathbb{N}}$, $(\varepsilon_k)_{k\in\mathbb{N}}$ be sequences in $\mathbb{R}_+$ such that $\sum_{k\in\mathbb{N}} \varepsilon_k < +\infty$ and*

$$(\forall\, k \in \mathbb{N}) \quad a_{k+1} \leq a_k + \varepsilon_k. \tag{3.20}$$

Then $(a_k)_{k\in\mathbb{N}}$ is convergent.

Proof. Define $u_k = a_k + \sum_{i=k}^{+\infty} \varepsilon_i$. Then it follows from (3.20) that $u_{k+1} = a_{k+1} + \sum_{i=k+1}^{+\infty} \varepsilon_i \leq a_k + \sum_{i=k}^{+\infty} \varepsilon_i$, so that $(u_k)_{k\in\mathbb{N}}$ is decreasing and hence convergent. Then, by definition of u_k, $a_k = u_k - \sum_{i=k}^{+\infty} \varepsilon_i$ and hence $(a_k)_{k\in\mathbb{N}}$ is convergent too. □

Theorem 3.4.7. *Let $(x_k)_{k\in\mathbb{N}}$, $(f_k)_{k\in\mathbb{N}}$, and $(\bar{x}_k)_{k\in\mathbb{N}}$ be the sequences generated by Algorithm 3.4.5. The following hold.*

(i) *Suppose that $\gamma_k \to 0$ and $\sum_{k\in\mathbb{N}} \gamma_k = +\infty$. Then $\liminf_k f(x_k) = \lim_k f_k = \inf_C f$.*

(ii) *Suppose that $\operatorname{argmin}_C f \neq \varnothing$, that $\sum_{k\in\mathbb{N}} \gamma_k = +\infty$, and $\sum_{k\in\mathbb{N}} \gamma_k^2 < +\infty$. Then there exists $x_* \in \operatorname{argmin}_C f$, such that $x_k \to x_*$.*

(iii) *Let $x \in C$ and let $m, k \in \mathbb{N}$ with $m \leq k$. Then*

$$\sum_{j=m}^{k} \frac{\gamma_j}{\sum_{i=m}^{k} \gamma_i} f(x_j) - f(x) \leq \frac{\|x_m - x\|^2}{2} \frac{1}{\sum_{i=m}^{k} \gamma_i} + \frac{L^2}{2} \frac{\sum_{j=m}^{k} \gamma_i^2}{\sum_{i=m}^{k} \gamma_i}. \tag{3.21}$$

(iv) *Suppose that $\sum_{k\in\mathbb{N}} \gamma_k = +\infty$ and $\sum_{i=0}^{k} \gamma_i^2 / \sum_{i=0}^{k} \gamma_i \to 0$. Then $f_k \to \inf_C f$ and $f(\bar{x}_k) \to \inf_C f$. Moreover, if $\operatorname{argmin}_C f \neq \varnothing$, the right hand side of (7.6), with $m = 0$ and $x \in \operatorname{argmin}_C f$, yields a rate of convergence for both $f_k - \min_C f$ and $f(\bar{x}_k) - \min_C f$.*

Proof. We first prove the following inequality.

$$(\forall\, k \in \mathbb{N})(\forall\, x \in C) \quad \|x_{k+1} - x\|^2 \leq \|x_k - x\|^2 - 2\gamma_k(f(x_k) - f(x)) + \gamma_k^2 L^2. \tag{3.22}$$

Indeed let $k \in \mathbb{N}$, $y_k = x_k - \gamma_k u_k$, and $x \in C$. Then, since $u_k \in \partial f(x_k)$, we have $f(x) \geq f(x_k) + \langle u_k, x - x_k\rangle$, and hence, using the relation $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2$, we have

$$\begin{aligned} f(x_k) - f(x) \leq \langle x_k - x, u_k\rangle &= \frac{1}{\gamma_k}\langle x_k - x, x_k - y_k\rangle \\ &= \frac{1}{2\gamma_k}\big(\|x_k - x\|^2 + \|x_k - y_k\|^2 - \|y_k - x\|^2\big). \end{aligned} \tag{3.23}$$

Now, since P_C is nonexpansive, we have $\|x_{k+1} - x\| = \|P_C(y_k) - P_C(x)\| \leq \|y_k - x\|$. Therefore, it follows from (7.8) that

$$\begin{aligned} f(x_k) - f(x) &\leq \frac{1}{2\gamma_k}\big(\|x_k - x\|^2 - \|x_{k+1} - x\|^2 + \|x_k - y_k\|^2\big) \\ &= \frac{1}{2\gamma_k}\big(\|x_k - x\|^2 - \|x_{k+1} - x\|^2\big) + \frac{\gamma_k}{2}\|u_k\|^2. \end{aligned}$$

Multiplying the above inequality by $2\gamma_k$ and rearranging the terms and using the fact that $\|u_k\| \leq L$, (3.22) follows.

(i): Since $(f_k)_{k\in\mathbb{N}}$ is decreasing, we have $\lim_k f_k = \inf_k f_k = \inf_k f(x_k)$. Moreover, $\inf_C f \leq \inf_k f_k \leq \liminf_k f(x_k)$. The latter inequality follows from the fact that there exists a subsequence $(x_{n_k})_{k\in\mathbb{N}}$ such that $\lim_k f(x_{n_k}) = \liminf_k f(x_k)$ and $f(x_{n_k}) \geq f_{n_k} \to$

$\inf_k f_k$. Therefore it is sufficient to prove that $\liminf_k f(x_k) \leq \inf_C f$. Suppose that $x \in C$ be such that $f(x) < \liminf_k f(x_k) = \sup_n \inf_{k\geq n} f(x_k)$. Then there exists $n \in \mathbb{N}$ such that $f(x) < \inf_{k\geq n} f(x_k)$. Set $\rho = \inf_{k\geq n} f(x_k) - f(x) > 0$. Then, (3.22) yields

$$\begin{aligned}(\forall\, k \geq n) \quad \|x_{k+1} - x\|^2 &\leq \|x_k - x\|^2 - 2\gamma_k\rho + \gamma_k^2 L^2 \\ &= \|x_k - x\|^2 - \gamma_k\rho - \gamma_k\big(\rho - \gamma_k L^2\big).\end{aligned}$$

Now, since $\gamma_k \to 0$, there exists $m \in \mathbb{N}$ such that for every integer $k \geq m$ we have $\gamma_k L^2 \leq \rho$, and hence

$$(\forall\, k \geq \nu := \max\{n, m\}) \qquad \gamma_k\rho \leq \|x_k - x\|^2 - \|x_{k+1} - x\|^2 . \tag{3.24}$$

This yields $\rho \sum_{k\geq\nu} \gamma_k \leq \|x_\nu - x\|^2 < +\infty$, which contradicts the assumption $\sum_{k\in\mathbb{N}} \gamma_k = +\infty$. Therefore, we showed that there is no $x \in C$ such that $f(x) < \liminf_k f(x_k)$. This means that $\liminf_k f(x_k) \leq \inf_C f$.

(ii): It follows from (3.22) that

$$(\forall\, k \in \mathbb{N})(\forall\, x \in \operatorname{argmin}_C f) \quad \|x_{k+1} - x\|^2 \leq \|x_k - x\|^2 + \gamma_k^2 L^2. \tag{3.25}$$

Therefore, it follows from Lemma 3.4.6 that, for every $x \in \operatorname{argmin}_C f$, $(\|x_k - x\|)_{k\in\mathbb{N}}$ is convergent and hence bounded. Now, since $(x_k)_{k\in\mathbb{N}}$ is bounded and $\liminf_k f(x_k) = \inf_C f$, there exists a subsequence $(x_{i_k})_{k\in\mathbb{N}}$ such that

$$f(x_{i_k}) \to \inf_C f \quad \text{and} \quad x_{i_k} \to x_* \quad \text{for some } x_* \in C.$$

Since f is continuous at x_*, $f(x_{i_k}) \to f(x_*)$, hence $x_* \in \operatorname{argmin}_C f$. Now since $\|x_k - x_*\|$ is convergent, and $\|x_{i_k} - x_*\| \to 0$, necessarily we have $\|x_k - x_*\| \to 0$.

(ii): It follows from (3.22) that

$$(\forall\, i \in \mathbb{N}) \quad \gamma_i(f(x_i) - f(z)) \leq \frac{1}{2}\big(\|x_i - x\|^2 - \|x_{i+1} - x\|^2\big) + \frac{L^2}{2}\gamma_i^2. \tag{3.26}$$

So, summing form m to k, we have

$$\sum_{i=m}^{k} \gamma_i(f(x_i) - f(x)) \leq \frac{1}{2}\|x_m - x\|^2 + \frac{L^2}{2}\sum_{i=m}^{k}\gamma_i^2. \tag{3.27}$$

Dividing the above inequality by $\sum_{i=m}^k \gamma_i$ yields (7.6).

(iii): We first note that, since f is convex and $\bar{x}_k$ is a convex combination of the x_i's, with coefficients $\eta_i = \gamma_i / \sum_{j=0}^k \gamma_j$, with $0 \leq i \leq k$, we have $f(\bar{x}_k) \leq \sum_{i=0}^k \eta_i f(x_i)$. Moreover, $f_k = \sum_{i=0}^k \eta_i f_k \leq \sum_{i=0}^k \eta_i f(x_i)$. Therefore,

$$(\forall\, k \in \mathbb{N}) \quad \max\{f_k, f(\bar{x}_k)\} \leq \Big(\sum_{i=0}^{k}\gamma_i\Big)^{-1}\sum_{i=0}^{k}\gamma_i f(x_i). \tag{3.28}$$

Let $z \in C$. Then it follows from (7.6) and (7.12) that $\limsup_k \max\{f_k, f(\bar{x}_k)\} \leq f(z)$. Since z is arbitrary in C, we have $\limsup_k \max\{f_k, f(\bar{x}_k)\} \leq \inf_C f$. Moreover, clearly we have $\inf_C f \leq \liminf_k \max\{f_k, f(\bar{x}_k)\}$. Therefore, $\max\{f_k, f(\bar{x}_k)\} \to \inf_C f$. Since $\inf_C f \leq f_k \leq \max\{f_k, f(\bar{x}_k)\}$ and $\inf_C f \leq f(\bar{x}_k) \leq \max\{f_k, f(\bar{x}_k)\}$, the statement follows. □

Lemma 3.4.8. *Let $\phi\colon \mathbb{R}_{++} \to \mathbb{R}$ be a convex function and let Φ be a primitive of ϕ. Let $m, k \in \mathbb{N}$, with $1 \leq m < k$, and set $h_{m,k} = \sum_{i=m}^{k} \phi(i)$. Then $h_{m,n} \geq \Phi(k) - \Phi(m) + \frac{1}{2}(f(m) + f(k))$. Moreover, suppose in addition that ϕ is nonincreasing and that $m \geq 2$. Then $h_{m,k} \leq \Phi(k) - \Phi(m-1)$.*

Proof. Since f is convex, using the trapezoidal rule, we have

$$\begin{aligned}
\int_m^k \phi(x)dx &\leq \sum_{i=m}^{k-1} \frac{1}{2}\big(\phi(i) + \phi(i+1)\big) = \frac{1}{2}\bigg(\sum_{i=m}^{k-1} \phi(i) + \sum_{i=m+1}^{k} \phi(i)\bigg) \\
&= \frac{1}{2}\bigg(\phi(m) + 2\sum_{i=m+1}^{k-1} \phi(i) + \phi(k)\bigg) \\
&= \frac{1}{2}\big(\phi(m) + \phi(k)\big) + \sum_{i=m+1}^{k-1} \phi(i) \\
&= \frac{1}{2}\big(\phi(m) + \phi(k)\big) + h_{m,k} - \big(\phi(m) + \phi(k)\big) \\
&= h_{m,k} - \frac{1}{2}\big(\phi(m) + \phi(k)\big).
\end{aligned}$$

The first part of the statement follows. Regarding the second part, we have $h_{m,k} \leq \int_{m-1}^{k} \phi(x)dx = \Phi(k) - \Phi(m-1)$. □

Lemma 3.4.9. *Let $m, k \in \mathbb{N}$ with $2 \leq m < k$. Then, the following inequalities hold.*

(i) $\log\Big(\frac{k}{m}\Big) + \frac{1}{2}\Big(\frac{1}{m} + \frac{1}{k}\Big) \leq \sum_{i=m}^{k} \frac{1}{i} \leq \log\Big(\frac{k}{m-1}\Big)$

(ii) $2(\sqrt{k} - \sqrt{m}) + \frac{1}{2}\bigg(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{k}}\bigg) \leq \sum_{i=m}^{k} \frac{1}{\sqrt{i}}$.

(iii) $\sum_{i=0}^{+\infty} \frac{1}{i^2} = \frac{\pi}{6}$.

Proof. Inequalities in (i) and (ii) follows from Lemma 3.4.8. The last item is standard. □

Lemma 3.4.10. *Let $a \in \mathbb{R}^n_{++}$ and $\alpha, \beta \in \mathbb{R}_{++}$. Then*

$$\min_{\gamma \in \mathbb{R}^n_{++}} \frac{\alpha}{2a^\top\gamma} + \frac{\beta}{2}\frac{\|\gamma\|^2}{a^\top\gamma} = \sqrt{\frac{\alpha\beta}{\|a\|^2}} \tag{3.29}$$

and the minimum is achieved at $\gamma = \Big(\sqrt{\alpha/\beta\,\|a\|^2}\Big)a$.

Proof. Define $\varphi\colon \mathbb{R} \times \mathbb{R}^n \to \,]-\infty, +\infty]$ such that

$$\varphi(t, \gamma) = \begin{cases} \dfrac{\alpha + \beta\,\|\gamma\|^2}{2t} & \text{if } t > 0 \text{ and } \gamma \in \mathbb{R}^n_+ \\ +\infty & \text{otherwise.} \end{cases} \tag{3.30}$$

Clearly φ is closed, convex, and differentiable in $\mathbb{R}_{++}\times\mathbb{R}^n_{++}$, and, for all $(t,\gamma)\in\mathbb{R}_{++}\times\mathbb{R}^n_{++}$,

$$\nabla\varphi(t,\gamma)=\bigg(-\frac{\alpha+\beta\left\|\gamma\right\|^2}{2t^2},\frac{\beta}{t}\gamma\bigg). \tag{3.31}$$

Then,

$$\inf_{\gamma\in\mathbb{R}^n_{++}}\frac{\alpha}{2a^\top\gamma}+\frac{\beta}{2}\frac{\left\|\gamma\right\|^2}{a^\top\gamma}=\inf_{t>0}\inf_{\substack{\gamma\in\mathbb{R}^n_{++}\\ a^\top\gamma=t}}\frac{\alpha+\beta\left\|\gamma\right\|^2}{2t}=\inf_{\substack{(t,\gamma)\in\mathbb{R}\times\mathbb{R}^n\\ a^\top\gamma=t}}\varphi(t,\gamma), \tag{3.32}$$

and the right hand side can be written as

$$\inf_{(t,\gamma)\in\mathbb{R}\times\mathbb{R}^n}\varphi(t,\gamma)+\iota_{\{0\}}\big((-1,a)^\top(t,\gamma)\big). \tag{3.33}$$

So, Fermat's rule yields

$$0\in\nabla\varphi(t,\gamma)+A^*\partial\iota_{\{0\}}(A(t,\gamma)), \tag{3.34}$$

where $A\colon\mathbb{R}^{n+1}\to\mathbb{R}$ is the linear form $A=(-1,a)^\top\cdot$ and A^* is the map $s\mapsto s(-1,a)$. Therefore, we have

$$(-1,a)^\top(t,\gamma)=0\quad\text{and}\quad-\nabla\varphi(t,\gamma)\in\mathbb{R}(-1,a), \tag{3.35}$$

which, in view of (3.31), implies that there exists $s\in\mathbb{R}$ such that

$$\begin{cases}\dfrac{\alpha+\beta\left\|\gamma\right\|^2}{2t^2}=-s\\ -\dfrac{\beta}{t}\gamma=sa\\ a^\top\gamma=t\end{cases} \tag{3.36}$$

Now, it follows from the last two equations above that $-\beta=-\beta a^\top\gamma/t=s\left\|a\right\|^2$ and hence

$$\begin{cases}\dfrac{\alpha+\beta\left\|\gamma\right\|^2}{2t^2}=\dfrac{\beta}{\left\|a\right\|^2}\\ \gamma=\dfrac{t}{\left\|a\right\|^2}a\\ a^\top\gamma=t\end{cases} \tag{3.37}$$

It follows from the second equation above that $\left\|\gamma\right\|^2=t^2/\left\|a\right\|^2$ which, substituted into the first equation, gives $t_*=\sqrt{\alpha/\beta}\left\|a\right\|$. Therefore finally, we have

$$\gamma_*=\bigg(\sqrt{\frac{\alpha}{\beta\left\|a\right\|^2}}\bigg)a\quad\text{and}\quad\varphi(t_*,\gamma_*)=\sqrt{\frac{\alpha\beta}{\left\|a\right\|^2}}. \tag{3.38}$$

□

Corollary 3.4.11. *Under the same assumptions of Theorem 3.4.7, the following hold.*

(i) *Suppose that* $\operatorname{argmin}_C f\neq\varnothing$ *and let* $D\geq\operatorname{dist}(x_0,\operatorname{argmin}_C f)$ *and* $k\in\mathbb{N}$. *Then,*

$$\max\{f_k,f(\bar{x}_k)\}-\min_C f\leq\frac{D^2}{2}\frac{1}{\sum_{i=0}^k\gamma_i}+\frac{L^2}{2}\frac{\sum_{j=0}^k\gamma_i^2}{\sum_{i=0}^k\gamma_i}. \tag{3.38}$$

Moreover, the right hand side of (7.13) *is minimized when, for every* $i = 0, \dots, k$, $\gamma_i = D/(L\sqrt{k+1})$ *and in that case we have*

$$\max\left\{f_k, f(\bar{x}_k)\right\} - \min_C f \leq \frac{LD}{\sqrt{k+1}}. \tag{3.39}$$

(ii) *Let, for every* $k \in \mathbb{N}$, $\gamma_k = \bar{\gamma}/(k+1)$. *Then,* $f_k \to \inf_C f$ *and* $f(\bar{x}_k) \to \inf_C f$. *Moreover, if* $\operatorname{argmin}_C f \neq \varnothing$, *there exists* $x_* \in \operatorname{argmin}_C f$ *such that* $x_k \to x_*$ *and we have, for every* $k \in \mathbb{N}$,

$$\max\left\{f_k, f(\bar{x}_k)\right\} - \min_C f \leq \left(\frac{\operatorname{dist}(x_0, \operatorname{argmin}_C f)^2}{2\bar{\gamma}} + \frac{\pi\bar{\gamma}L^2}{12}\right)\frac{1}{\log(k+1)}. \tag{3.40}$$

(iii) *Let, for every* $k \in \mathbb{N}$, $\gamma_k = \bar{\gamma}/\sqrt{k+1}$. *Then,* $f_k \to \inf_C f$ *and* $f(\bar{x}_k) \to \inf_C f$. *Moreover, if* $\operatorname{argmin}_C f \neq \varnothing$, *for every integer* $k \geq 2$, *we have*

$$\max\left\{f_k, f(\bar{x}_k)\right\} - \min_C f \leq \frac{\operatorname{dist}(x_0, \operatorname{argmin}_C f)^2}{2\bar{\gamma}}\frac{1}{\sqrt{k+1}} + \bar{\gamma}L^2\frac{\log(k+1)}{\sqrt{k+1}}. \tag{3.41}$$

(iv) *Let, for every* $k \in \mathbb{N}$, $\gamma_k = \bar{\gamma}/\sqrt{k+1}$ *and suppose that* C *is bounded with diameter* $\bar{D} > 0$ *and that* $\operatorname{argmin}_C f \neq \varnothing$. *Set, for every* $k \in \mathbb{N}$, $\tilde{f}_k = \min_{\lfloor k/2 \rfloor \leq i \leq k} f(x_i)$ *and* $\tilde{x}_k = \left(\sum_{i=\lfloor k/2 \rfloor}^k \gamma_i\right)^{-1} \sum_{i=\lfloor k/2 \rfloor}^k \gamma_i x_i$. *Then, for every integer* $k \geq 2$,

$$\max\left\{\tilde{f}_k, f(\tilde{x}_k)\right\} - \min_C f \leq \left(\frac{3\bar{D}^2}{2\bar{\gamma}} + \frac{5\bar{\gamma}L^2}{2}\right)\frac{1}{\sqrt{k+1}}. \tag{3.42}$$

Proof. (i): Equation (7.13) follows from (7.12) and by minimizing the right hand side of (7.6), with $m = 0$, w.r.t. $x \in \operatorname{argmin}_C f$. Now, it follows from Lemma 3.4.10 that minimum of the right-hand side of (7.13) is achieved at $(\gamma_i)_{0 \leq i \leq k} \equiv D/(L\sqrt{k+1})$ and yield

$$\frac{1}{k+1}\sum_{i=0}^k f(x_i) - \min_C f \leq \frac{LD}{\sqrt{k+1}}. \tag{3.43}$$

The statement follows from the fact that $\max\{f_k, f(\bar{x}_k)\} \leq (1/(k+1))\sum_{i=0}^k f(x_i)$.

(ii): The first part follows from Theorem 3.4.7(iii). We derive from Lemma 3.4.9(i), with $m = 1$, that $\sum_{i=0}^k \gamma_i = \bar{\gamma}\sum_{i=1}^{k+1}(1/i) \geq \bar{\gamma}\log(k+1)$. Moreover, we have $\sum_{i=0}^k \gamma_i^2 = \bar{\gamma}\sum_{i=1}^{k+1} 1/i^2 \leq \bar{\gamma}\pi/6$. The statement follows from (7.13) with $D = \operatorname{dist}(x_0, \operatorname{argmin}_C f)$.

(iii): The first part follows from Theorem 3.4.7(iii). Lemma 3.4.9(ii), with $m = 1$, yields $\sum_{i=1}^k 1/\sqrt{i} \geq 2(\sqrt{k} - 1) + (1/2)(1 + 1/\sqrt{k}) \geq 2\sqrt{k} - 3/2$. Moreover, $2\sqrt{k} - 3/2 \geq \sqrt{k}$ for $k \geq 3$ and clearly for $k \leq 2$, $\sum_{i=1}^k 1/\sqrt{i} \geq \sqrt{k}$. Therefore, for every $k \in \mathbb{N}$, $\sum_{i=0}^k \gamma_i = \bar{\gamma}\sum_{i=1}^{k+1} 1/\sqrt{i} \geq \bar{\gamma}\sqrt{k+1}$. Moreover, by Lemma 3.4.9(i), we have $\sum_{i=1}^k 1/i = 1 + \sum_{i=2}^k 1/i \leq 1 + \log k \leq 2\log k$, for $k \geq 3$. Therefore, for every $k \in \mathbb{N}$, $k \geq 2$, we have $\sum_{i=0}^k \gamma_i^2 = \bar{\gamma}^2\sum_{i=1}^{k+1} 1/i \leq 2\bar{\gamma}^2\log(k+1)$. The statement follows from (7.13) with $D = \operatorname{dist}(x_0, \operatorname{argmin}_C f)$.

(iv): Let $k \in \mathbb{N}$, $k \geq 2$. It follows from Lemma 3.4.9(ii) that

$$\sum_{i=\lfloor k/2 \rfloor}^{k} \gamma_i^2 = \bar{\gamma}^2 \sum_{i=\lfloor k/2 \rfloor+1}^{k+1} \frac{1}{i} \leq \bar{\gamma}^2 \log\Big(\frac{k+1}{\lfloor k/2 \rfloor}\Big) \leq \bar{\gamma}^2 \log 4 \leq \bar{\gamma}^2 \frac{5}{3}.$$

Moreover, Lemma 3.4.9(ii) yields

$$\sum_{i=\lfloor k/2 \rfloor}^{k} \gamma_i = \bar{\gamma} \sum_{i=\lfloor k/2 \rfloor+1}^{k+1} \frac{1}{\sqrt{i}} \geq 2\bar{\gamma}(\sqrt{k+1} - \sqrt{\lfloor k/2 \rfloor + 1}) \geq 2\bar{\gamma}\sqrt{k+1}\Big(1 - \sqrt{\frac{\lfloor k/2 \rfloor + 1}{k+1}}\Big).$$

Now, since $(\lfloor k/2 \rfloor + 1)/(k+1) \leq 2/3$, we have

$$\sum_{i=\lfloor k/2 \rfloor}^{k} \gamma_i \geq 2\bar{\gamma}\Big(1 - \sqrt{\frac{2}{3}}\Big)\sqrt{k+1} \geq \frac{\bar{\gamma}}{3}\sqrt{k+1}$$

The statement follows from Theorem 3.4.7(ii), with $m = \lfloor k/2 \rfloor$ and $z \in \operatorname{argmin}_C f$, taking into account that, as in (7.12), $\max\{\tilde{f}_k, f(\tilde{x}_k)\} \leq \big(\sum_{i=\lfloor k/2 \rfloor}^{k} \gamma_i\big)^{-1} \sum_{i=\lfloor k/2 \rfloor}^{k} \gamma_i f(x_i)$. □

Lecture 4

Duality Theory – part I

4.1 The orthogonal Projection

Definition 4.1.1. Let C be a nonempty subset of X and let $x \in X$. Then $p \in C$ is a *projection of x onto C* (or *a best approximation to x from C*) if

$$(\forall\, y \in C) \quad \|p - x\| \leq \|y - x\|\,.$$

Note that projections onto C are nothing but minimizer of the function $\|\cdot - x\|^2 + \iota_C$. The set C is called *Chebyshev* if every point of X has exactly one projection onto C. In such case the *projection operator onto C* is the operator, denoted by P_C, that maps every point in X to its unique projection onto C.

Suppose that C is convex. Then $\|x - \cdot\|^2 + \iota_C$ is strongly convex, by Proposition 1.3.12. Hence it exists at most one projection of C onto x, in view of Theorem 1.5.5 If C is also closed, then, since $\|x - \cdot\|^2 + \iota_C$ is strongly convex and closed (the sum of two closed functions is closed), by Theorem 1.5.4, there exists a unique projection of x onto C. So if C is a nonempty closed and convex set, then C is a Chebyshev set and for every $x \in X$, it is defined the projection of x onto C which is denoted by $P_C(x)$. In the '30 it has been proved that Chebyshev sets in Euclidean spaces are necessarily closed and convex. It is still unknown if this result holds in infinite dimensional Hilbert spaces too.

Proposition 4.1.2 (Variational characterization of the projection)**.** *Let C be a nonempty convex set. Let $x \in X$ and $p \in C$. Then the following are equivalent:*

(i) *p is a projection of x onto C*

(ii) *$\forall\, y \in C$, $\langle x - p, y - p\rangle \leq 0$.*

Proof. Statement (i) means

$$(\forall\, y \in C) \quad \|p - x\|^2 \leq \|y - x\|^2\,. \tag{4.1}$$

Salzo Saverio (saverio.salzo@iit.it) Istituto Italiano di Tecnologia. Via E. Melen, 16152, Genoa, Italy.

Clearly, (4.1) is equivalent to

$$(\forall\, y \in C)(\forall\, t \in]0,1]) \quad \begin{aligned} \|p-x\|^2 &\leq \|p+t(y-p)-x\|^2 \\ &= \|p-x\|^2 + t^2\,\|y-p\| \\ &\qquad + 2t\langle p-x, y-p\rangle, \end{aligned}$$

which, in turn, is equivalent to

$$(\forall\, y \in C)(\forall\, t \in]0,1]) \quad 2\langle x-p, y-p\rangle \leq t\,\|y-p\|^2\,.$$

Since t is arbitrary in $]0,1]$ the statement follows. □

Remark 4.1.3. Proposition 4.1.2 can also be proved by using Fermat's rule for the function $f = (1/2)\,\|\cdot - x\|^2 + \iota_C$ and taking into account the calculus rule for subdifferentials (see in particular Corollary 3.2.4(ii)).

Remark 4.1.4. If p is a projection onto C of $x \notin C$, then $p \in \mathrm{bdry}(C)$, the boundary of C. Indeed if it was $p \in \mathrm{int}(C)$, then there would exist $\delta > 0$ such that $B_\delta(p) \subset C$ and hence $\inf_{z\in C} \|x-z\| < \|x-p\|$, which contradicts the definition of p.

Proposition 4.1.5. *Let C be a nonempty closed and convex set. Then*

$$\forall\, x, y \in X \qquad \|P_C(x) - P_C(y)\|^2 \leq \langle P_C(x) - P_C(y), x-y\rangle. \tag{4.2}$$

Then Cauchy-Schwarz inequality yields that $\|P_C(x) - P_C(y)\| \leq \|x-y\|$, which proves that P_C is a nonexpansive operator.

Proof. Let x and y in C. Then it follows from Proposition 4.1.2 that

$$\begin{aligned} \langle P_C(x) - P_C(y), y - P_C(y)\rangle &\leq 0 \\ \langle P_C(y) - P_C(x), x - P_C(x)\rangle &\leq 0. \end{aligned}$$

Summing the two inequalities above, we get $\langle P_C(x) - P_C(y), P_C(x) - P_C(y) - (x-y)\rangle \leq 0$ and hence (4.2). □

Remark 4.1.6. Property (4.2) is called firm nonexpansiveness.

4.2 Affine hyperplanes and half-spaces

An *affine hyperplane of X* is defined as the set

$$H = \big\{x \in X \,\big|\, \langle u, x\rangle = \alpha\big\}, \tag{4.3}$$

where $u \in X \setminus \{0\}$ and $\alpha \in \mathbb{R}$. Clearly $H \neq 0$, since $\alpha u/\|u\|^2 \in H$. Moreover,

$$x_0 \in H \implies H = \big\{x \in X \,\big|\, \langle u, x - x_0\rangle = 0\big\} = x_0 + \ker\langle u, \cdot\rangle. \tag{4.4}$$

We now take a special $x_0 \in H$. Consider the straight line $\mathbb{R}u = \{tu \,|\, t \in \mathbb{R}\}$, passing through the origin with direction u. Then we compute $\mathbb{R}u \cap H$. We have

$$(\forall\, t \in \mathbb{R}) \quad tu \in H \;\Leftrightarrow\; \langle u, tu\rangle = \alpha \;\Leftrightarrow\; t = \alpha/\|u\|^2\,.$$

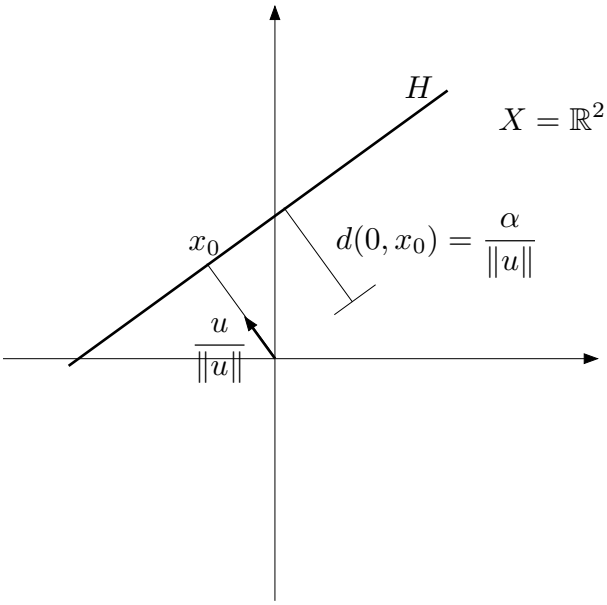


Figure 4.1: Parametrization of hyperplanes.

Therefore

$$x_0 = \frac{\alpha}{\|u\|}\frac{u}{\|u\|} \in H \cap \mathbb{R}u \quad \text{and} \quad \operatorname{dist}(0, H) = \|P_H(0) - 0\| = \|x_0\| = \frac{|\alpha|}{\|u\|}, \tag{4.5}$$

since $0 - x_0 \in H^\perp$, and hence $x_0 = P_H(0)$. Moreover, $\alpha/\|u\|$ is the signed (or oriented) distance of H from 0. Affine hyperplanes are parametrized by $(u, \alpha) \in (X \setminus \{0\}) \times \mathbb{R}$. See Figure 4.1.

Moreover, if for some $(v, \beta) \in (X \setminus \{0\}) \times \mathbb{R}$, $H = \big\{x \in X \,\big|\, \langle v, x\rangle = \beta\big\}$, then

$$\frac{\alpha}{\|u\|}\frac{u}{\|u\|} = \frac{\beta}{\|v\|}\frac{v}{\|v\|}, \tag{4.6}$$

which implies $u/\|u\| = \pm v/\|v\|$ and $\alpha/\|u\| \pm = \beta/\|v\|$, hence

$$(v, \beta) = \lambda(u, \alpha), \quad \text{with } \lambda = \frac{\|v\|}{\|u\|}. \tag{4.7}$$

So two parametrizations of an affine hyperplanes are multiples to each other.

Given $u \in X$, $u \neq 0$, and $\alpha \in \mathbb{R}$, define the *half-space*

$$H^- = \big\{x \in X \,\big|\, \langle u, x\rangle \leq \alpha\big\}. \tag{4.8}$$

We have

$$H^- = \big\{x \in X \,\big|\, \langle u, x - x_0\rangle \leq 0\big\}, \qquad x_0 = \frac{\alpha}{\|u\|^2}u. \tag{4.9}$$

We note that since $x_0 \in H$, then $x_0 + u \notin H^-$. Thus, H^- is the half-space opposite to that indicated by the direction u. We finally note that

$$\operatorname{int}(H^-) = \big\{x \in X \,\big|\, \langle u, x\rangle < \alpha\big\}. \tag{4.10}$$

Indeed it is easy to see that the following list of statements are equivalent:

(i) $x \in \operatorname{int}(H^-)$

(ii) $(\exists\, \delta > 0)$ such that $\forall\, v \in B_1(0)\ \langle u, x + \delta v\rangle \leq \alpha$

(iii) $(\exists\, \delta > 0)$ such that $\forall\, v \in B_1(0),\ \delta\langle u, v\rangle \leq \alpha - \langle u, x\rangle$

(iv) $(\exists\, \delta > 0)$ such that $\langle u, x\rangle + \delta\, \|u\| \leq \alpha$

(v) $\langle u, x\rangle < \alpha$.

We now clarify the forms of hyperplanes of $X \times \mathbb{R}$. According to the general definition, an hyperplane H of $X \times \mathbb{R}$ has the form $H = \{(x,t) \in X \times \mathbb{R} \,|\, \langle (u,s), (x,t)\rangle = \alpha\}$, with $(u,s) \neq (0,0)$ and $\alpha \in \mathbb{R}$. We have two cases:

(i) $s = 0$. Then $u \neq 0$ and $H = \{x \in X \,|\, \langle u, x\rangle = \alpha\} \times \mathbb{R}$, that is H is a *vertical* hyperplane of $X \times \mathbb{R}$.

(ii) $s \neq 0$. Then $H = \{(x,t) \in X \times \mathbb{R} \,|\, \langle -u/s, x\rangle + \alpha/s = t\}$, which is the graph of the affine functional $x \mapsto \langle -u/s, x\rangle + \alpha/s$.

4.3 Dual representation of convex sets

Theorem 4.3.1 (Separation of convex sets from points)**.** *Let $C \subset X$ be a nonempty, closed, and convex set with $C \neq X$. Let $x \in X \setminus C$. Then there exists a closed hyperplane H of X such that*

$$C \subset H^- \quad \text{and} \quad x \notin H^-. \tag{4.11}$$

Proof. We consider the projection operator onto C, $P_C \colon X \to X$. Then it follows from Proposition 4.1.2 that

$$(\forall\, y \in C) \qquad \langle x - P_C(x), y - P_C(x)\rangle \leq 0. \tag{4.12}$$

Then, setting $u = x - P_C(x) \neq 0$ and defining

$$H^- = \big\{y \in X \,\big|\, \langle u, y - P_C(x)\rangle \leq 0\big\},$$

we have $C \subset H^-$ and $x \notin H^-$, since $\langle u, x - P_C(x)\rangle = \|u\|^2 > 0$. □

Proposition 4.3.2. *Every proper, closed, and convex subset of X is the intersection of a nonempty family of half-spaces of X.*

Proof. Let $C \subset X$ be closed and convex, with $C \neq X$ and consider the family

$$\mathcal{F} = \big\{F \text{ halfspace of } X \,\big|\, C \subset F\big\}. \tag{4.13}$$

This family is non empty, since there exists $x \notin C$ and for such x, by Theorem 4.3.1, there exists an halfspace F containing C. Moreover, again Theorem 4.3.1 ensures that

$$x \notin C \ \Rightarrow\ \exists\, F \in \mathcal{F}\ x \notin F \ \Rightarrow\ x \notin \bigcap_{F \in \mathcal{F}} F$$

which means that $\bigcap_{F \in \mathcal{F}} F \subset C$. The other inclusion follows directly from (4.13). □

Definition 4.3.3. A *polyhedron* is the intersection of a finite family of half-spaces. A *polytope* is a bounded polyhedron.

Definition 4.3.4. Let C be a nonemtpy subset of X. Then a *supporting hyperplane* for C is an hyperplane H of X such that $C \subset H^-$ and $C \cap H \neq \varnothing$. If $x \in C \cap H$ we say that H is a *supporting hyperplane for C at x*.

Remark 4.3.5. If C admits a supporting hyperplane at x, then $x \in \mathrm{bdry}(C)$, the boundary of C. Indeed let H be a hyperplane of X such that $C \subset H^-$ and $x \in C \cap H$. Then it follows from (4.10) that $\mathrm{int}(C) \subset \mathrm{int}(H^-)$ and $\mathrm{int}(H^-) \cap H = \varnothing$. Therefore $x \notin \mathrm{int}(C)$.

Remark 4.3.6. In the proof of Theorem 4.3.1 we actually proved that, setting for every $x \notin C$,

$$H_{(x)} = \big\{y \in X \,\big|\, \langle x - P_C(x), y - P_C(x)\rangle = 0\big\},$$
$$H^-_{(x)} = \big\{y \in X \,\big|\, \langle x - P_C(x), y - P_C(x)\rangle \leq 0\big\},$$

we have $C = \bigcap_{x \notin C} H^-_{(x)}$. Moreover, for every $x \notin C$, $H_{(x)}$ is a supporting hyperplane for C, since $C \subset H^-_{(x)}$ and $P_C(x) \in C \cap H_{(x)}$. Therefore we can say that every proper closed and convex subset of X is an intersection of a nonempty family of half-spaces defined by supporting hyperplanes for C (at some point of C).

Definition 4.3.7. For any subset $C \subset X$, the function $\sigma_C \colon X \to]-\infty, +\infty]$ defined by

$$\sigma_C(u) = \sup_{x \in C} \langle u, x\rangle, \tag{4.14}$$

is called the *support function* of the set C. See Figure 4.2.

Theorem 4.3.8. *Let C be a nonempty closed convex subset of X. Then*

$$C = \bigcap_{u \in \mathrm{dom}\sigma_C, u \neq 0} H^-(u, \sigma_C(u)) = \big\{x \in X \,\big|\, \forall\, u \in X \ \langle u, x\rangle \leq \sigma_C(u)\big\}. \tag{4.15}$$

Proof. It follows from Proposition 4.3.2 that

$$C = \bigcap_{(u,\alpha) \in \Lambda(C)} H^-(u, \alpha), \quad H^-(u, \alpha) := \big\{x \in X \,\big|\, \langle x, u\rangle \leq \alpha\big\}, \tag{4.16}$$

where $\Lambda(C) = \{(u, \alpha) \in (X \setminus \{0\}) \times \mathbb{R} \,|\, C \subset H^-(u, \alpha)\}$. Equivalently we can write

$$C = \bigcap_{\substack{u \in X \setminus \{0\} \\ \exists\, \alpha \in \mathbb{R},\, C \subset H^-(u,\alpha)}} \ \bigcap_{\alpha \in \mathbb{R},\, C \subset H^-(u,\alpha)} H^-(u, \alpha). \tag{4.17}$$

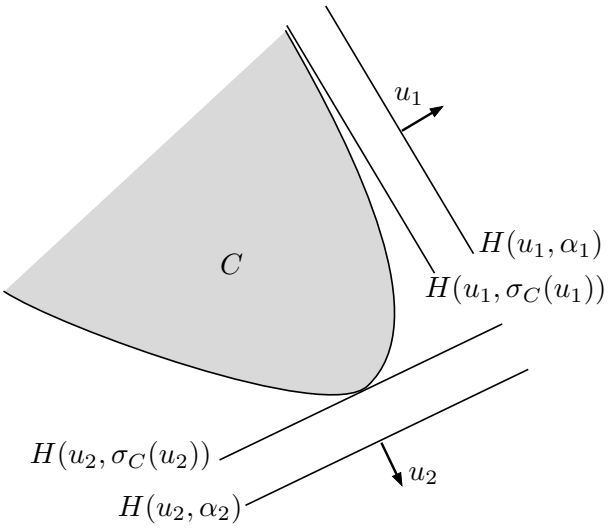


Figure 4.2: Support function of the set C.

Now, for every $u \in X$, such that $\exists \alpha \in \mathbb{R}, C \subset H^-(u, \alpha)$, we have

$$\bigcap_{\alpha \in \mathbb{R},\, C \subset H^-(u,\alpha)} H^-(u, \alpha) = H^-(u, \sigma_C(u)). \tag{4.18}$$

Indeed, for every $\alpha \in \mathbb{R}$, we have

$$\begin{aligned} C \subset H^-(u, \alpha) &\Leftrightarrow \forall\, x \in C\ \langle x, u \rangle \leq \alpha \\ &\Leftrightarrow \alpha \text{ is an upper bound of } (\langle x, u \rangle)_{x \in C}. \end{aligned}$$

Therefore $\left\{\alpha \in \mathbb{R} \,|\, C \subset H^-(u, \alpha)\right\}$ is the set of the upper bounds of $(\langle x, u \rangle)_{x \in C}$. This set has minimum equal to $\sup_{x \in C} \langle x, u \rangle = \sigma_C(u) < +\infty$. So,

$$C = \bigcap_{\substack{u \in X \setminus \{0\} \\ \exists\, \alpha \in \mathbb{R},\, C \subset H(u,\alpha)}} H^-(u, \sigma_C(u)). \tag{4.19}$$

Note that the statement "$\nexists\, \alpha \in \mathbb{R}$, $C \subset H(u, \alpha)$" is equivalent to $\sup_{x \in C} \langle x, u \rangle = +\infty$. □

Thus the theorem above asserts that C can be written in terms of the support function only. In the following we provide properties of σ_C.

Proposition 4.3.9. *Let C be a nonempty subset of X. Then σ_C is proper convex, closed and positively homogeneous.*

Proof. This follows from the fact that $\sigma_C = \sup_{x \in C} \langle \cdot, x \rangle$ is an upper envelope of bounded linear functions and that $\sigma_C(0) = 0$. □

Recalling (4.5) we have that $\sigma_C(u)/\left\|u\right\| = \sigma_C(u/\left\|u\right\|)$ is the oriented distance of $H(u, \sigma_C(u))$ from the origin.

An interesting question is whether it is possible to obtain a dual representation of closed convex sets by supporting hyperplanes.

Theorem 4.3.10. *Let $C \subset X$ be a nonempty, closed and convex set. Then for every $x \in \mathrm{bdry}(C)$ there exists a supporting hyperplane for C at x.*

Proof. Let $x \in \mathrm{bdry}(C)$ and let $(x_k)_{k\in\mathbb{N}}$ be a sequence in $X \setminus C$ such that $x_k \to x$. [1] Then it follows from Theorem 4.3.1 that, for every $k \in \mathbb{N}$ there exists a u_k (that we can take of norm 1) and a sequence of real number α_k such that

$$C \subset H_k^- = \big\{y \in X \,\big|\, \langle u_k, y\rangle \le \alpha_k\big\} \quad \text{and} \quad \langle u_k, x_k\rangle > \alpha_k.$$

Therefore, for every $k \in \mathbb{N}$

$$(\forall\, y \in C) \quad \langle u_k, y\rangle \le \alpha_k < \langle u_k, x_k\rangle$$

and hence

$$(\forall\, y \in C) \quad \langle u_k, y - x_k\rangle < 0. \tag{4.20}$$

Now, since $\|u_k\| = 1$, up to a subsequence we have that u_k converges to some u. Therefore, passing to the limit in (4.20) we have that, for every $y \in C$, $\langle u, y - x\rangle \le 0$. Moreover, since $u_k \to u$ and the norm $\|\cdot\|$ is continuous, we have $\|u\| = 1$ (since $\|u_k\| = 1$). So

$$C \subset H^- = \big\{y \in X \,\big|\, \langle u, y - x\rangle \le 0\big\} \quad \text{and} \quad x \in H. \qquad \square$$

Corollary 4.3.11. *Let C be a nonempty closed convex subset of X. Then for every $x \in \mathrm{bdry}(C)$, $N_C(x) \neq \{0\}$.*

Proof. Let $x \in \mathrm{bdry}(C)$. Then, it follows from Theorem 4.3.10 that there exists a supporting hyperplane of C at x, that is, there exists $u \in X \setminus \{0\}$ such that $C \subset \{y \in X \,|\, \langle u, y - x\rangle \le 0\}$. Hence, recalling Definition 3.1.15, we have $u \in N_C(x)$. $\square$

Example 4.3.12.

(i) Let $C = B_1(0)$ the closed unit ball of X. Then, for every u $\sigma_C(u) = \sup_{x\in C} \langle u, x\rangle = \|u\|$, that is $\sigma_C = \|\cdot\|$.

(ii) Let $(x_i)_{i\in I}$ be a finite family of points in X. Then $\sigma_{\{x_i,|\, i\in I\}} = \sigma_{\mathrm{co}\{x_i,|\, i\in I\}}$. Indeed this follows from Proposition 1.2.6 and Definition 4.3.7.

(iii) Let K be a nonempty cone, that is, $\lambda v \in K$ for every $v \in K$ and $\lambda > 0$. Then

$$\sigma_K(u) = \sup_{x\in K} \langle u, x\rangle = \begin{cases} 0 & \text{if } u \in K^\circ \\ +\infty & \text{otherwise} \end{cases} = \iota_{K^\circ}(u),$$

where the set

$$K^\circ = \big\{u \in X \,\big|\, \forall\, x \in K\ \langle u, x\rangle \le 0\big\}$$

is *the dual (or polar) cone of K*.

(iv) Let V be a subspace of X. Then $\sigma_V = \iota_{V^\perp}$.

[1] Since $\mathrm{bdry}(C) = \mathrm{cl}(C) \cap \mathrm{cl}(X \setminus C)$, we have $x \in \mathrm{cl}(X \setminus C)$, and hence there exists a sequence $(x_k)_{k\in\mathbb{N}}$ in $X \setminus C$ such that $x_k \to x$.

Exercise 4.3.13. Compute the support function of $C \subset \mathbb{R}$ in the following cases:

(i) $C = [0, +\infty[$;

(ii) $C =]-\infty, 0]$;

(iii) $C = \{x_0\}$, $x_0 \in \mathbb{R}$.

(iv) $C = [x_1, x_2]$, with $x_1, x_2 \in \mathbb{R}, x_1 < x_2$.

4.4 The Legendre-Fenchel Transform

Let $f\colon X \to]-\infty, +\infty]$ be a closed convex and proper function. We are going to introduce f^* the *Legendre-Fenchel* transform of f, by considering the set $C = \mathrm{epi}(f) \subset X \times \mathbb{R}$ and its dual representation, as given in Theorem 4.3.8. To that purpose we recall that

$$\begin{aligned} f \text{ is closed } &\Leftrightarrow \mathrm{epi}(f) \text{ is closed} \\ f \text{ is convex } &\Leftrightarrow \mathrm{epi}(f) \text{ is convex.} \end{aligned}$$

Moreover, if $(f_i)_{i\in I}$, is a family of functions $f_i\colon X \to \mathbb{R}$, then

$$\mathrm{epi}\left(\sup_{i\in I} f_i\right) = \bigcap_{i\in I} \mathrm{epi}(f_i). \tag{4.21}$$

We recall that *affine functions* on X are functions of type $\ell\colon X \to \mathbb{R}$, $\ell = \langle u, \cdot\rangle - \alpha$, with $u \in X$ and $\alpha \in \mathbb{R}$.

Lemma 4.4.1. *A closed half-space of $X \times \mathbb{R}$ containing the epigraph of a proper function is either vertical or the epigraph of an affine function.*

Proof. If the half space is not vertical, it can be written as

$$\big\{(x,t) \in X \times \mathbb{R} \,\big|\, t \geq \ell(x)\big\} \quad \text{or} \quad \big\{(x,t) \in X \times \mathbb{R} \,\big|\, t \leq \ell(x)\big\} \tag{4.22}$$

for some affine function $\ell\colon X \to \mathbb{R}$. However, since f is proper there exists an $x \in \mathrm{dom} f$ such that $\{x\} \times [f(x), +\infty[\subset \mathrm{epi}(f)$. Hence the second half space in (4.22) cannot contain $\mathrm{epi}(f)$. □

Concerning epigraphs, it is a striking and important fact that we can actually get rid of the vertical half-spaces in their dual representation.

Theorem 4.4.2. *Let $f\colon X \to]-\infty, +\infty]$ be a closed convex and proper function. Then f admits affine minorants and is equal to the supremum of all the affine functions which minorize f.*

Proof. We first observe that there exists an affine minorant of f. Indeed it follows from Proposition 4.3.2 that $\mathrm{epi}(f)$ is an intersection of a family of half-spaces of $X \times \mathbb{R}$. If in this family there were only vertical hyperplanes, then there would exist a family $(H_i)_{i\in I}$ of hyperplanes of X such that

$$\varnothing \neq \mathrm{epi}(f) = \bigcap_{i\in I} (H_i^- \times \mathbb{R}) = \Big(\bigcap_{i\in I} H_i^-\Big) \times \mathbb{R},$$

which implies that $f \equiv -\infty$ on the set $\bigcap_{i\in I} H_i^- \neq \varnothing$. So at least one hyperplane in the family have to be non-vertical. Then in view of Lemma 4.4.1 this must be the epigraph of an affine function ℓ. Thus, $\operatorname{epi}(f) \subset \operatorname{epi}(\ell)$ and hence $\ell \leq f$. Now, let $(\bar{x}, \bar{t}) \in X \times \mathbb{R}$ be such that $(\bar{x}, \bar{t}) \notin \operatorname{epi}(f)$. Then there exists a hyperplane of $X \times \mathbb{R}$ which separates $\operatorname{epi}(f)$ from $(\bar{x}, \bar{t})$. Now, soppose that this hyperplane is vertical. Then there exists a hyperplane $H = \big\{x \in X \,|\, \langle u, x\rangle = \alpha\big\}$ of X such that

$$\operatorname{epi}(f) \subset H^- \times \mathbb{R} \qquad \text{and} \qquad (\bar{x}, \bar{t}) \notin H^- \times \mathbb{R},$$

which yields

$$(\forall\, x \in \operatorname{dom} f) \qquad \langle u, x\rangle \leq \alpha \qquad \text{and} \qquad \langle u, \bar{x}\rangle > \alpha. \tag{4.23}$$

Moreover, according to what we proved at the beginning, there exists an affine function ℓ such that $\ell \leq f$. Let $\gamma > 0$ be such that $\ell(\bar{x}) + \gamma(\langle u, \bar{x}\rangle - \alpha) > \bar{t}$. Then,

$$g := \ell + \gamma(\langle u, \cdot\rangle - \alpha)$$

is an affine function and, for every $x \in \operatorname{dom} f$, $g(x) \leq \ell(x) \leq f(x)$ and $g(\bar{x}) > \bar{t}$. Therefore $\operatorname{epi}(f) \subset \operatorname{epi}(g)$ and $(\bar{x}, \bar{t}) \notin \operatorname{epi}(g)$. So, in any case, we can separate $\operatorname{epi}(f)$ from $(\bar{x}, \bar{t})$, by an epigraph of an affine function, which is a non-vertical half-space of $X \times \mathbb{R}$. This proves that $\operatorname{epi}(f)$ is the intersection of a family of epigraphs of affine functions. In view of (4.21) the statement follows. □

Corollary 4.4.3. *If f is proper lower semicontinuous and strongly convex, then f is coercive.*

Proof. By Proposition 1.3.12 we have $f = g + (\mu/2)\,\|\cdot\|^2$ for some $\mu > 0$ and $g \in \Gamma_0(X)$. Then it follows from Theorem 4.4.2 that g it has an affine minimizer, that is, there exists $u \in X$ and $\alpha \in \mathbb{R}$ such that $\alpha + \langle \cdot, u\rangle \leq g$. Thus $\lim_{\|x\|\to+\infty} f(x) \geq \lim_{\|x\|\to+\infty} \alpha + \langle x, u\rangle + (\mu/2)\,\|x\|^2 = +\infty$; hence f is coercive. □ □

Just like for convex sets, we are going to look for the simplest dual description of closed convex functions, i.e., using the simplest affine minorants. Let $f\colon X \to \,]-\infty, +\infty]$ be a closed convex and proper function. It follows from Theorem 4.4.2 that

$$(\forall\, x \in X) \quad f(x) = \sup\big\{\langle u, x\rangle - \alpha \,|\, (u, \alpha) \in X \times \mathbb{R} \text{ and } \langle u, \cdot\rangle - \alpha \leq f\big\}. \tag{4.24}$$

For a given $u \in X$, we have that, for every $\alpha \in \mathbb{R}$

$$\langle u, \cdot\rangle - \alpha \leq f \;\Leftrightarrow\; (\forall\, x \in X) \quad \langle u, x\rangle - f(x) \leq \alpha \Leftrightarrow \sup_{x\in X} \langle u, x\rangle - f(x) \leq \alpha. \tag{4.25}$$

Therefore, if f admits an affine minorant with slope u, the best α is obtaining by taking

$$\alpha = \sup_{x\in X} \langle u, x\rangle - f(x).$$

See Figure 4.3. Thus, the following definition make sense.

Definition 4.4.4. Let $f\colon X \to \,]-\infty, +\infty]$ be proper. The function

$$f^*\colon X \to \,]-\infty, +\infty]\,, \quad f^*(u) = \sup_{x\in X} \langle u, x\rangle - f(x)$$

is called the *Legendre-Fenchel* conjugate of f.[a]

[a] Since f is proper, there exists at least an $x \in X$ such that $f(x) \in \mathbb{R}$, hence $f^*(u) > -\infty$.

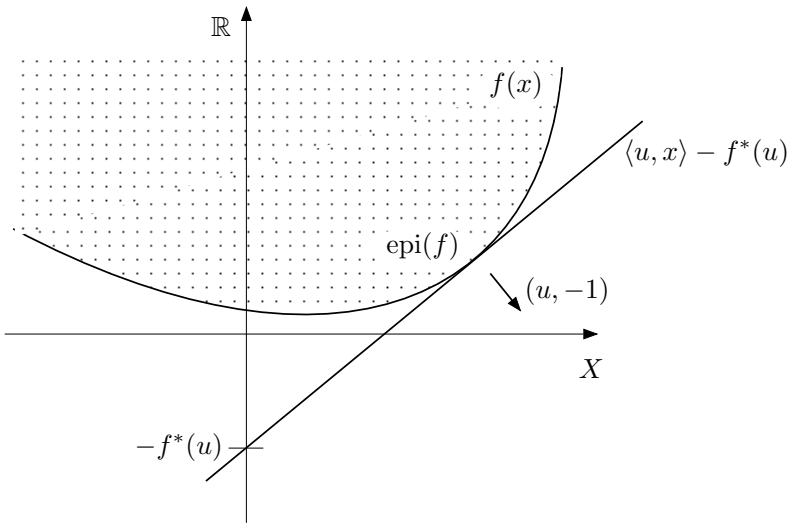


Figure 4.3: The meaning of the conjugate function f^* of f.

Remark 4.4.5. Let $u \in X$. It follows from Definition 4.4.4 that, for every $\alpha \in \mathbb{R}$,

$$\underbrace{-\alpha}_{\ell(0)} \leq -f^*(u) \;\Leftrightarrow\; f^*(u) \leq \alpha \;\Leftrightarrow\; \forall\, x \in X \;\; \langle u, x\rangle - f(x) \leq \alpha \;\Leftrightarrow\; \underbrace{\langle u, \cdot\rangle - \alpha}_{\text{affine function } \ell} \leq f.$$

Therefore, we see that

$$f^*(u) < +\infty \;\Leftrightarrow\; \exists \text{ an affine function } \ell\colon X \to \mathbb{R} \text{ with slope } u \text{ such that } \ell \leq f.$$

Moreover,

$$-f^*(u) = \sup\big\{\ell(0) \,\big|\, \ell \text{ affine function on } X \text{ with slope } u \text{ and } \ell \leq f\big\}.$$

Note that the last supremum is achieved (i.e., it is a maximum) if $f^*(u) < +\infty$. This explains the geometrical meaning of the Legendre-Fenchel conjugate, which is that for every slope u, $-f^*(u)$ is the maximum y-intercept of the affine functions with slope u below f. Also, we can say that $\langle u, \cdot\rangle - f^*(u)$ is the greatest affine function below f (in the sense of point-wise convergence). See Figure 4.3.

Then, in view of Definition 4.4.4 and Remark 4.4.5, equation (4.24) yields

$$\begin{aligned}(\forall\, x \in X) \quad f(x) &= \sup\big\{\langle u, x\rangle - \alpha \,\big|\, u \in \mathrm{dom} f^*, \alpha \in \mathbb{R}, f^*(u) \leq \alpha\big\} \\ &= \sup\big\{\langle u, x\rangle - f^*(u) \,\big|\, u \in \mathrm{dom} f^*\big\} \\ &= \sup\big\{\langle u, x\rangle - f^*(u) \,\big|\, u \in X\big\}.\end{aligned}$$

This proves the following important result.

Theorem 4.4.6 (Fenchel–Moreau)**.** *Let $f\colon X \to \,]-\infty, +\infty]$ be a closed, convex, and proper function. Then,*

$$(\forall\, x \in X) \qquad f(x) = \sup_{u \in X} \big(\langle u, x\rangle - f^*(u)\big),$$

that is, $f = f^{**}$.

Proposition 4.4.7 (Properties of the conjugate operation)**.** *Let $f\colon X \to]-\infty,+\infty]$ be a proper function. Then the following hold.*

(i) *f^* is closed and convex.*

(ii) *f has an affine minorant $\Leftrightarrow$ f^* is proper $\Leftrightarrow$ $f^* \in \Gamma_0(X)$ (recall Definition 1.3.9).*

(iii) $f^*(0) = -\inf_X f$

(iv) *Let $g\colon X \to]-\infty,+\infty]$ be a proper function. Then $f \leq g \;\Rightarrow\; f^* \geq g^*$.*

(v) *Let $(f_i)_{i\in I}$ be a family of proper extended real-valued functions. Then $\big(\inf_{i\in I} f_i\big)^* = \sup_{i\in I} f_i^*$.*

(vi) *Let $\gamma > 0$. Then, for every $u \in X$, $(\gamma f)^*(u) = \gamma f^*(u/\gamma)$.*

(vii) *(The conjugate of a separable function is separable). Suppose that $X = \prod_{i=1}^m X_i$ for some Euclidean spaces X_i, endowed with the scalar product $\langle x, y\rangle = \sum_{i=1}^m \langle x_i, y_i\rangle$. Suppose additionally that*

$$\forall\, x = (x_1, x_2, \ldots, x_m) \in X \qquad f(x) = f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m),$$

for some $f_i \in \Gamma_0(X_i)$, $i = 1, \ldots, m$. Then,

$$\forall\, u = (u_1, \ldots, u_m) \in X \qquad f^*(u) = f_1^*(u_1) + f_2^*(u_2) + \cdots + f_m^*(u_m).$$

(viii) *Let $x_0 \in X$ and suppose that $f(x) = g(x - x_0)$ for some proper function $g\colon X \to]-\infty,+\infty]$. Then $f^* = g^* + \langle \cdot, x_0\rangle$.*

Remark 4.4.8. The appropriate space to study the Fenchel conjugate is

$$\Gamma_0(X) = \big\{ f\colon X \to]-\infty,+\infty] \mid f \text{ proper, convex, and closed} \big\}.$$

Indeed

$$f \in \Gamma_0(X) \;\Rightarrow\; f^* \in \Gamma_0(X) \text{ and } f^{**} = f \text{ (Fenchel-Moreau theorem)}.$$

Thus,

$$\cdot^*\colon \Gamma_0(X) \to \Gamma_0(X)$$

is an *involution*, that is, bijective with inverse itself (like the complex conjugate $z \in \mathbb{C} \mapsto z^* \in \mathbb{C}$).

Example 4.4.9.

(i) Let $x_0 \in X$. Then, for every $u \in X$, $\iota_{x_0}^*(u) = \sup_x \big(\langle u, x\rangle - \iota_{x_0}(x)\big) = \langle u, x_0\rangle$.

(ii) The following holds:

- $\iota_C^* = \sigma_C$, where $C \subset X$;
- $\iota_K^* = \iota_{K^\circ}$, where K is a cone of X and $K^\circ = \big\{ v \in X \mid \forall\, x \in K,\ \langle v, x\rangle \leq 0 \big\}$ is its dual cone.

- $\iota_V^* = \iota_{V^\perp}$, where $V \subset X$ is a (vector) subspace of X and $V^\perp$ is its orthogonal.

(iii) Let $f(x) = e^x$. Then

$$f^*(u) = \begin{cases} u \log u - u & \text{if } u > 0 \\ 0 & \text{if } u = 0 \\ +\infty & \text{if } u < 0 \end{cases}$$

which is the (negative) *Bolzmann–Shannon Entropy*.

(iv) Suppose that f is differentiable and that $\nabla f \colon X \to X$ is a bijection. Then, for every $u \in X$, $f^*(u) = \sup_x \big(\langle u, x\rangle - f(x)\big) = \langle u, (\nabla f)^{-1}(u)\rangle - f\big((\nabla f)^{-1}(u)\big)$.

(v) For every $u \in X$, $\big((1/2)\,\|\cdot\|^2\big)^*(u) = \langle u, u\rangle - (1/2)\,\|u\|^2 = (1/2)\,\|u\|^2$.

(vi) Let $p > 1$ and set $\varphi(t) = (1/p)|t|^p$. Then the derivative $\varphi'(t) = |t|^{p-1}\operatorname{sign}(t)$ is a bijection and $(\varphi')^{-1}(s) = |s|^{1/(p-1)}\operatorname{sign}(s)$. Therefore

$$\varphi^*(s) = s|s|^{1/(p-1)}\operatorname{sign}(s) - (1/p)||s|^{1/(p-1)}\operatorname{sign}(s)|^p = (1/q)|s|^q$$

where $1/p + 1/q = 1$.

(vii) Let $\varphi \colon \mathbb{R} \to]-\infty, +\infty]$ be an even function and set $f(x) = \varphi(\|x\|)$. Then $f^*(u) = \varphi^*(\|u\|)$. Indeed

$$\begin{aligned} f^*(u) &= \sup_{x \in X} \big(\langle u, x\rangle - \varphi(\|x\|)\big) \\ &= \sup_{t \in \mathbb{R}_+} \sup_{\|x\|=1} \big(\langle u, tx\rangle - \varphi(\|tx\|)\big) \\ &= \sup_{t \in \mathbb{R}_+} \sup_{\|x\|=1} \big(t\langle u, x\rangle - \varphi(t)\big) \\ &= \sup_{t \in \mathbb{R}} \big(t\,\|u\| - \varphi(t)\big) \\ &= \varphi^*(\|u\|), \end{aligned}$$

where in the penultimate equality we used that $\varphi(-t) = \varphi(t)$. Then it follows from this example and (vi) that $\big((1/p)\,\|\cdot\|^p\big)^* = (1/q)\,\|\cdot\|^{1/q}$.

(viii) Let $f \colon X \to]-\infty, +\infty]$ be proper and positively homogeneous. Then, for every $u \in X$,

$$\begin{aligned} f^*(u) &= \sup_{x \in X} \big(\langle u, x\rangle - f(x)\big) \\ &= \sup_{t \in \mathbb{R}_+} \sup_{x \in X} t\big(\langle u, x\rangle - f(x)\big) \\ &= \begin{cases} 0 & \text{if } u \in \partial f(0) \\ +\infty & \text{if } u \notin \partial f(0). \end{cases} \end{aligned}$$

Therefore, $f^* = \iota_{\partial f(0)}$. Recall that $\partial f(0)$ is a closed convex cone.

(ix) It follows from Example 3.1.13 and (vii) that $(\|\cdot\|)^* = \iota_{B_1(0)}$ (this follows also from the simple fact that $\|\cdot\| = (\iota_{B_1(0)})^*$ and Theorem 4.4.6).

Proposition 4.4.10 (Important facts)**.** *Let $f\colon X \to]-\infty,+\infty]$ be proper and convex and let $x, u \in X$. Then, the following holds.*

(i) $\langle u, x\rangle \leq f(x) + f^*(u)$ (Young-Fenchel inequality).

(ii) $\langle u, x\rangle = f(x) + f^*(u) \;\Leftrightarrow\; u \in \partial f(x)$.

(iii) *If $f \in \Gamma_0(X)$, then $u \in \partial f(x) \Leftrightarrow\; x \in \partial f^*(u)$.*

Proof. (i): Let $(x, u) \in X \times X$. If either $x \notin \operatorname{dom} f$ or $u \notin \operatorname{dom} f^*$ we have $f(x) + f^*(u) = +\infty$ and the statement is trivial. Suppose that $x \in \operatorname{dom} f$ and $u \in \operatorname{dom} f^*$. Then, by definition of $f^*(u)$, we have $f^*(u) \geq \langle u, x\rangle - f(x)$ and the statement follows.

(ii): We have

$$\begin{aligned} u \in \partial f(x) &\;\Leftrightarrow\; f(x) \in \mathbb{R} \text{ and } \forall\, y \in X,\ \langle u, y\rangle - f(y) \leq \langle u, x\rangle - f(x) \\ &\;\Leftrightarrow\; f(x) \in \mathbb{R} \text{ and } f^*(u) \leq \langle u, x\rangle - f(x). \end{aligned}$$

Then, recalling (i), the statement follows.

(iii): The statement follows from the fact that $f^{**} = f$. □

Proposition 4.4.11. *Let $f\colon X \to]-\infty,+\infty]$ be a proper function. Then*

$$f = f^* \;\Leftrightarrow\; f = \frac{1}{2}\|\cdot\|^2. \tag{4.26}$$

Proof. The implication "⇐" was already proved in Example 4.4.9. Suppose that $f = f^*$. Then, by the Young-Fenchel inequality we have that, for every $x \in X$, $2f(x) \geq \langle x, x\rangle$. Thus, $f \geq (1/2)\|\cdot\|^2$. Therefore, by Proposition 4.4.7(iv), we have $f = f^* \leq (1/2)\|\cdot\|^2$. □

Remark 4.4.12. If $f \in \Gamma_0(X)$, it follows from Proposition 4.4.10 that,

$$u \in \operatorname{dom} \partial f^* \;\Leftrightarrow\; \partial f^*(u) \neq \varnothing \;\Leftrightarrow\; \exists\, x \in X \text{ such that } u \in \partial f(x).$$

Hence $\operatorname{dom}\partial f^* = \operatorname{ran} \partial f$.

4.5 Further results on subdifferentials and properties in duality

The following theorem is fundamental and provide a complete description of the subdifferential of a convex function.

Theorem 4.5.1. *Let $f\colon X \to]-\infty,+\infty]$ be a convex function and let $x \in \operatorname{dom} f$. Then the following hold.*

(i) *If $x \in \operatorname{int}(\operatorname{dom} f)$, then $\partial f(x)$ is nonempty and compact, $f'(x, \cdot)$ is finite and continuous, and*

$$(\forall\, v \in X) \quad f'(x, v) = \max\big\{\langle v, u\rangle \,\big|\, u \in \partial f(x)\big\}.$$

(ii) *If $x \in \operatorname{bdry}(\operatorname{dom} f)$, then $\partial f(x)$ is either unbounded or empty.*

Proof. (i): Suppose that $x \in \operatorname{int}(\operatorname{dom} f)$. By Proposition 3.1.9, $\partial f(x)$ is closed. We now prove that $f'(x,\cdot)$ is finite and continuous everywhere and that $\partial f(x)$ is nonempty and bounded. Since f is continuous at x, it is indeed Lipschitz continuous around x (see Lemma 1.4.1). Therefore, there exists $L \geq 0$ such that, for every y, y' in a ball $B_\delta(x)$ we have $|f(y) - f(y')| \leq L\|y - y'\|$. Thus, for every $v \in X$ and for sufficiently small $t > 0$ we have $|f(x+tv) - f(x)|/t \leq L\|v\|$. Letting $t \to 0$, we get

$$(\forall\, v \in X) \qquad |f'(x,v)| \leq L\|v\|\,. \tag{4.27}$$

Hence, $f'(x,\cdot)$ is finite everywhere. Then, in view of Proposition 3.1.4 and Theorem 1.4.2, $f'(x,\cdot)$ is continuous on X. Moreover, it follows from (4.27) and Proposition 3.1.17 that, for every $u \in \partial f(x)$, $\|u\| \leq L$. Finally, since $f'(x,\cdot)$ is positively homogeneous, it follows from Example 4.4.9(viii) and Remark 3.1.18 that $(f'(x,\cdot))^* = \iota_{\partial f(x)}$. Therefore, since $f'(x,\cdot) \in \Gamma_0(X)$, we derive from Theorem 4.4.6 that $(f'(x,\cdot))^*$ is proper and hence $\partial f(x) \neq \varnothing$ and moreover $f'(x,\cdot) = (\iota_{\partial f(x)})^*$. Thus, for every $v \in X$, $f'(x,v) = \sup_{u\in\partial f(x)} \langle u, v\rangle$. Now since $\partial f(x)$ is compact (being closed and bounded) and $\langle\cdot, v\rangle$ is continuous, the sup is indeed a max.

(ii): Let $x \in \operatorname{bdry}(\operatorname{dom} f)$. Then it follows from Corollary 4.3.11 that $N_{\operatorname{dom} f}(x) \neq \{0\}$. Hence, that there exists $u \in X$, $u \neq 0$, such that

$$(\forall\, y \in \operatorname{dom} f) \quad \langle y - x, u\rangle \leq 0.$$

Now, for every $v \in \partial f(x)$ and every $\lambda > 0$, we have

$$(\forall\, y \in \operatorname{dom} f) \quad f(x) + \langle y - x, v + \lambda u\rangle = f(x) + \langle y - x, v\rangle + \lambda \underbrace{\langle y - x, u\rangle}_{\leq 0} \leq f(y),$$

hence $v + \lambda u \in \partial f(x)$. Thus, since $\partial f(x)$ contains a ray, it is unbounded. □

Proposition 4.5.2. *Let $(f_i)_{i\in I}$ be a finite family of continuous affine functions on X, say $f_i = \langle\cdot, u_i\rangle + \alpha_i$, for some $u_i \in X$ and $\alpha_i \in \mathbb{R}$. Let $f = \max_{i\in I} f_i$, let $x \in X$ and set $I(x) = \{i \in I \mid f_i(x) = f(x)\}$. Then*

$$\partial f(x) = \operatorname{co}\{u_i \mid i \in I(x)\}. \tag{4.28}$$

Proof. By definition $\operatorname{dom} f = X$ and it follows from Proposition 1.3.6(ii), Theorem 1.4.2 Theorem 4.5.1 that f is convex, continuous, $\partial f(x) \neq \varnothing$, and $f'(x,v) = \sigma_{\partial f(x)}(v)$. Now, for every $y \in X$,

$$f(y) - f(x) = \max_{i\in I}\big(f_i(y) - f(x)\big) = \max_{i\in I}\big(\langle y - x, u_i\rangle + f_i(x) - f(x)\big).$$

Moreover, for every $i \in I(x)$, $f_i(x) - f(x) = 0$. Hence, for every $t > 0$ and $v \in X$,

$$\begin{aligned} f(x+tv) - f(x) &= \max\Big\{\max_{i\in I(x)} t\langle v, u_i\rangle, \max_{i\in I\setminus I(x)}\big(t\langle v, u_i\rangle + f_i(x) - f(x)\big)\Big\}. \\ &= t\max\Big\{\max_{i\in I(x)}\langle v, u_i\rangle, \max_{i\in I\setminus I(x)}\Big(\langle v, u_i\rangle + \underbrace{\frac{f_i(x) - f(x)}{t}}_{=:a_i(t)<0}\Big)\Big\}. \end{aligned}$$

Since $a_i(t) \to -\infty$ as $t \to 0$, for $t > 0$ sufficiently small, we have

$$\frac{f(x+tv)-f(x)}{t} = \max_{i\in I(x)} \langle v, u_i\rangle. \tag{4.29}$$

Therefore, $f'(x,v) = \max_{i\in I(x)} \langle v, u_i\rangle = \sigma_{\{u_i \,|\, i\in I(x)\}}(v) = \sigma_{\mathrm{co}\{u_i \,|\, i\in I(x)\}}(v)$ (recall Example 4.3.12(ii)) and hence $\sigma_{\partial f(x)} = \sigma_{\mathrm{co}\{u_i \,|\, i\in I(x)\}}$. The statement follows from Theorem 4.3.8. □

We now state the inverse of the result in Proposition 3.1.19(iii)

Proposition 4.5.3. *Let $f \in \Gamma_0(X)$ and $x \in \mathrm{dom} f$. Suppose that $\partial f(x) = \{u\}$. Then $x \in \mathrm{int}(\mathrm{dom} f)$, f is Gâteaux differentiable at x, and $\nabla f(x) = u$.*

Proof. Since $\partial f(x)$ is clearly nonempty and bounded, it follows from Theorem 4.5.1 that necessarily $x \in \mathrm{int}(\mathrm{dom} f)$ and $f'(x,\cdot) = \langle u, \cdot\rangle$. □

We now show that strict and strong convexity are dual properties of Gateâux differentiability and Lipschitz smoothness respectively.

Theorem 4.5.4. *Let $f \in \Gamma_0(X)$ be such that $\mathrm{dom}\partial f$ and $\mathrm{ran}\,\partial f$ are open*[2]*. Then the following statements are equivalent:*

(i) *f is strictly convex on $\mathrm{int}(\mathrm{dom} f)$;*

(ii) $\forall\, x, y \in X, \quad x \neq y \;\Rightarrow\; \partial f(x) \cap \partial f(y) = \varnothing$;

(iii) *f^* is differentiable on $\mathrm{dom}\,\partial f^* = \mathrm{int}(\mathrm{dom} f^*)$.*

Proof. (i) ⇒ (ii): Let $x, y \in \mathrm{int}(\mathrm{dom} f)$ with $x \neq y$. Let $\lambda \in]0,1[$ and set $z = (1-\lambda)x+\lambda y$. Then, for every $u \in X$, we have

$$\begin{aligned} 0 &\leq f(z) + f^*(u) - \langle z, u\rangle \\ &\leq (1-\lambda)f(x) + \lambda f(y) + f^*(u) - \langle z, u\rangle \\ &= (1-\lambda)\underbrace{\left[f(x) + f^*(u) - \langle x, u\rangle\right]}_{(a)} + \lambda \underbrace{\left[f(y) + f^*(u) - \langle y, u\rangle\right]}_{(b)}. \end{aligned} \tag{4.30}$$

Now suppose that there exists $u \in \partial f(x) \cap \partial f(y)$. Then it follows from Proposition 4.4.10(ii) that (a) and (b) above are equal to zero and also $f^*(u) < +\infty$. Therefore in (4.30) the inequalities become equalities and hence, $f((1-\lambda)x+\lambda y) = (1-\lambda)f(x)+\lambda f(y)$, which gives a contradiction.

(ii) ⇒ (i): Suppose, reasoning by contradiction, that there exist $x, y \in \mathrm{int}(\mathrm{dom} f)$ with $x \neq y$, and $\lambda \in]0,1[$, such that

$$f((1-\lambda)x + \lambda y) = (1-\lambda)f(x) + \lambda f(y).$$

Set $z = (1-\lambda)x + \lambda y$. Since $z \in \mathrm{int}(\mathrm{dom} f)$ it follows from Theorem 4.5.1 that there exists $u \in \partial f(z)$. Moreover in (4.30) equalities hold. Therefore, (a) and (b) are equal to zero. This implies, by Proposition 4.4.10(ii), that $u \in \partial f(x) \cap \partial f(y)$.

[2]Since $\mathrm{int}(\mathrm{dom} f) \subset \mathrm{dom}\partial f \subset \mathrm{dom} f$ and $\mathrm{int}(\mathrm{dom} f^*) \subset \mathrm{dom}\partial f^* = \mathrm{ran}\,\partial f \subset \mathrm{dom} f^*$, the assumptions implies that $\mathrm{dom}\partial f = \mathrm{int}(\mathrm{dom} f)$ and $\mathrm{ran}\,\partial f = \mathrm{dom}\partial f^* = \mathrm{int}(\mathrm{dom} f^*)$.

(ii) $\Leftrightarrow$ (iii): Statement (ii) is equivalent to

$$\forall\, x, y \in X \qquad \partial f(x) \cap \partial f(y) \neq \varnothing \;\Rightarrow\; x = y,$$

which is equivalent to

$$\forall\, x, y \in X, \forall\, u \in X \qquad u \in \partial f(x) \cap \partial f(y) \;\Rightarrow\; x = y,$$

and this statement, in virtue of Proposition 4.4.10(iii), is equivalent to

$$\forall\, u \in \mathrm{dom}\partial f^* \qquad \partial f^*(u) \text{ is a singleton.}$$

Then the statement follows from Proposition 4.5.3. □

Corollary 4.5.5. *Let $f \in \Gamma_0(X)$ be such that $\mathrm{dom}\partial f$ is open. Suppose that f is strictly convex and differentiable on $\mathrm{int}(\mathrm{dom} f)$ and that $\mathrm{ran}\,\nabla f$ is open. Then the gradient*

$$\nabla f\colon \mathrm{int}(\mathrm{dom} f) \to \mathrm{int}(\mathrm{dom} f^*) \text{ is a bijection}$$

and $(\nabla f)^{-1} = \nabla f^*$.

Proof. We know that $\mathrm{dom}\partial f = \mathrm{int}(\mathrm{dom} f)$ and $\mathrm{ran}(\partial f) = \mathrm{dom}\partial f^* = \mathrm{int}(\mathrm{dom} f^*)$. Then it follows from Proposition 4.4.10(iii) that for every $x \in \mathrm{int}(\mathrm{dom} f)$ and $u \in \mathrm{int}(\mathrm{dom} f^*)$,

$$u = \nabla f(x) \;\Leftrightarrow\; x = \nabla f^*(u).$$

The statement follows. □

Example 4.5.6. Let $f\colon X \to]-\infty, +\infty]$ be defined as

$$f(x) = \begin{cases} 1/x & \text{if } x > 0 \\ +\infty & \text{otherwise.} \end{cases}$$

Then $f \in \Gamma_0(\mathbb{R})$ and $\mathrm{dom} f = \mathrm{int}(\mathrm{dom} f) =]0, +\infty[$. Moreover f is strictly convex and differentiable on $x \in \mathrm{int}(\mathrm{dom} f)$ and, for every $x \in \mathrm{int}(\mathrm{dom} f)$, and $f'(x) = -1/x^2$. Therefore, $\mathrm{ran}\, f' = \mathrm{dom}(f^*)' =]-\infty, 0[$. So, it follows from Corollary 4.5.5 that $f'\colon]0, +\infty[\to]-\infty, 0[$ is a bijection, $(f^*)'\colon]-\infty, 0[\to]0, +\infty[$ and $(f^*)'(u) = (f')^{-1}(u) = 1/\sqrt{-u}$. Then we can find f^* by integration, that is,

$$f^*(u) = \int \frac{1}{\sqrt{-u}} du + \text{const.} = -2\sqrt{-u} + \text{const.}$$

Since $-f^*(0) = \inf f = 0$, we have that necessarily $f^*(u) = -2\sqrt{-u}$.

Proposition 4.5.7. *Let $f \in \Gamma_0(X)$ be strongly convex. Then f is supercoercive.*

Proof. Let $g \in \Gamma_0(X)$ be such that $f = (\mu/2)\,\|\cdot\|^2 + g$ for some $\mu > 0$. Since g admits an affine minimizer, there exists a continuous affine function $\ell = \langle \cdot, u\rangle + \beta$ such that $\ell \leq g$. Hence, for $x \neq 0$,

$$\frac{f(x)}{\|x\|} \geq \frac{\mu}{2}\,\|x\| + \left\langle \frac{x}{\|x\|}, u \right\rangle + \beta. \tag{4.31}$$

Since $\langle x/\|x\|, u\rangle$ is bounded from below, the statement follows. □

Proposition 4.5.8. *Let $f \in \Gamma_0(X)$, $x \in \mathrm{dom} f$ and $u \in \partial f(x)$. Then the following statements are equivalents*

(i) $(\forall\, y \in X)\ f(y) \geq f(x) + \langle y - x, u\rangle + \dfrac{\mu}{2}\|y - x\|^2$

(ii) $(\forall\, v \in X)\ f^*(v) \leq f^*(u) + \langle v - u, x\rangle + \dfrac{1}{2\mu}\|v - u\|^2$

Proof. Let $\varphi(y) = f(x) + \langle y - x, u\rangle + \dfrac{\mu}{2}\|y - x\|^2$. Then, recalling that $[(\mu/2)\|\cdot\|^2]^* = (1/2\mu)\|\cdot\|^2$ and that $\langle x, u\rangle = f(x) + f^*(u)$, we have

$$\begin{aligned}
\varphi^*(v) &= \sup_{y\in X} \langle y, v\rangle - f(x) - \langle y, u\rangle + \langle x, u\rangle - \frac{\mu}{2}\|y\|^2 - \frac{\mu}{2}\|x\|^2 + \mu\langle y, x\rangle \\
&= \sup_{y\in X} \langle y, v - u + \mu u\rangle - \frac{\mu}{2}\|y\|^2 - f(x) + \langle x, u\rangle - \frac{\mu}{2}\|x\|^2 \\
&= \frac{1}{2\mu}\|v - u + \mu u\|^2 + f^*(u) - \frac{\mu}{2}\|x\|^2 \\
&= f^*(u) + \langle v - u, x\rangle + \frac{1}{2\mu}\|v - u\|^2 .
\end{aligned}$$

Then the statement follows by recalling that, in virtue of Proposition 4.4.7(iv), we have $f \geq \varphi \ \Leftrightarrow\ f^* \leq \varphi^*$. □

Theorem 4.5.9. *Let $f \in \Gamma_0(X)$ and $\mu > 0$. Then, if f is μ-strongly convex, we have*

(a) $\mathrm{dom} f^* = X$, *f^* is differentiable on X and ∇f^* if $(1/\mu)$-Lipschitz continuous.*

Vice versa if (a) *holds, then f is μ-strongly convex on the convex subsets of* $\mathrm{dom}\partial f$.

Proof. Suppose that f is μ-strongly convex. It follows from Proposition 4.5.7 and Proposition 3.3.1(i) that $\mathrm{dom}\partial f^* = \mathrm{ran}\,\partial f = X$. Moreover, it follows from Proposition 3.3.1(ii) and Theorem 4.5.4 that f^* is Gâteaux differentiable on X. Then (a) follows from Proposition 3.3.2, Proposition 4.5.8, and Theorem 2.1.12(ii). So, the first part is proved. The second part follows from Theorem 2.1.12(ii), Proposition 4.5.8 and Proposition 3.3.3. □

We finish by introducing the following important notion.

Definition 4.5.10. Let $f \in \Gamma_0(X)$ and let $\lambda > 0$. The *Moreau envelope of f* with parameter λ is the function $f_\lambda\colon X \to \mathbb{R}$ such that, for every $u \in X$,

$$f_\lambda(u) = \inf_{x\in X}\{f(x) + \frac{1}{2\lambda}\|x - u\|^2\} \tag{4.32}$$

Remark 4.5.11. Since, for every $u \in X$, $f + (1/2\lambda)\|\cdot - u\|^2$ is proper strongly convex and lower semicontinuous, it follows that it has a unique minimizer. Therefore, $\inf(f + (1/2\lambda)\|\cdot - u\|^2) \in \mathbb{R}$. Moreover, by Proposition 1.3.6(iv) f_λ is convex.

Proposition 4.5.12. *Let $f \in \Gamma_0(X)$ and let $f_\lambda\colon X \to \mathbb{R}$ be its Moreau envelope with parameter $\lambda > 0$. Then the following hold.*

(i) $(\forall\, u \in X)\ \inf_X f \le f_\lambda(u) \le f(u)$.

(ii) $f_1 = (1/2)\,\|\cdot\|^2 - (f + (1/2)\,\|\cdot\|^2)^*$.

(iii) *f_λ is Frechet differentiable with $(1/\lambda)$-Lipschitz continuous derivative and*

$$\nabla f_\lambda(u) = \frac{1}{\lambda}\Big(u - \operatorname{argmin}_{x\in X}\Big\{f(x) + \frac{1}{2\lambda}\,\|x-u\|^2\Big\}\Big). \tag{4.33}$$

(iv) *For every $u \in X$, $f_\lambda(u) \uparrow f(u)$ as $\lambda \to 0$.*

Proof. (i): The first inequality is immediate since $f \le f + (1/(2\lambda))\,\|\cdot - u\|^2$. The second inequality follows by noting that in the definition of $f_\lambda(u)$, if $x = u$, then $f(x) + (1/(2\lambda))\,\|x-u\|^2 = f(x)$.

(ii): We have

$$\begin{aligned}(f + (1/2)\,\|\cdot\|^2)^*(u) &= \sup_{x\in X}\Big(\langle x, u\rangle - f(x) - \frac{1}{2}\,\|x\|^2\Big)\\ &= \sup_{x\in X}\Big(-\frac{1}{2}\,\|x-u\|^2 - f(x) + \frac{1}{2}\,\|u\|^2\Big)\\ &= \frac{1}{2}\,\|u\|^2 - \inf_{x\in X}\Big(\frac{1}{2}\,\|x-u\|^2 + f(x)\Big).\end{aligned}$$

(iii): We first note that $f_\lambda = (1/\lambda)g_1$, where $g = \lambda f$. Then, we prove that $(g + (1/2)\,\|\cdot\|^2)^*$ is differentiable. Indeed, since $g \in \Gamma_0(X)$, it follows from Proposition 4.4.10(iii) that $x \in \partial(g + (1/2)\,\|\cdot\|^2)^*(u) \iff u \in (g + (1/2)\,\|\cdot\|^2)(x) = \partial g(x) + x \iff x \in \operatorname{argmin}\{g(x) + (1/2)\,\|x-u\|^2\}$. Therefore, for every $u \in X$,

$$\partial\Big(g + \frac{1}{2}\,\|\cdot\|^2\Big)^*(u) = \operatorname{argmin}\Big\{g(x) + \frac{1}{2}\,\|x-u\|^2\Big\}, \tag{4.34}$$

which is a singleton, since $g + (1/2)\,\|\cdot - u\|^2$ is proper, strongly convex and lower semicontinuous. Thus, recalling Proposition 4.5.3 we have that $(g + (1/2)\,\|\cdot\|^2)^*$ is Gateâux differentiable and $\nabla(g + (1/2)\,\|\cdot\|^2)^*(u) = \operatorname{argmin}\{g(x) + (1/2)\,\|x-u\|^2\}$. Finally by formula (ii) we get that g_1 is Gateâux differentiable and $\nabla g_1(u) = u - \operatorname{argmin}\{g(x) + (1/2)\,\|x-u\|^2\} = u - \operatorname{argmin}\{f(x) + (1/(2\lambda))\,\|x-u\|^2\}$. It follows from Theorem 4.5.9 that $(g + (1/2)\,\|\cdot\|^2)^*$ is Lipschitz smooth with constant 1 and, by (ii), so is g_1. The statement follows.

(iv): Let $u \in X$ and set $J_\lambda u = \operatorname{argmin}_{x\in X}\{f(x) + (1/(2\lambda))\,\|x-u\|^2\}$. Clearly $(f_\lambda(u))_{\lambda\in\mathbb{R}_{++}}$ is decreasing in λ, so that $\lim_{\lambda\to 0} f_\lambda(u) = \sup_{\lambda\in\mathbb{R}_{++}} f_\lambda(u) =: \alpha$. It follows from (i) that $\lim_{\lambda\to 0} f_\lambda(u) \le f(u)$. Then the statement clearly follows if $\alpha = +\infty$. Therefore, suppose that $\alpha < +\infty$. Then, for every $\lambda \in\,]0,1]$,

$$f(J_\lambda(u)) + \frac{1}{2}\,\|J_\lambda(u) - u\|^2 \le f(J_\lambda(u)) + \frac{1}{2\lambda}\,\|J_\lambda(u) - u\|^2 = f_\lambda(u) \le \alpha.$$

Since $f + (1/2)\,\|\cdot - u\|^2$ is coercive, we have that $\beta := \sup_{\lambda\in]0,1]}\,\|J_\lambda(u)\| < +\infty$. Moreover, since f is minorized by a continuous affine function, say $f \ge \langle\cdot, v\rangle + \eta$, we have, for every $\lambda \in\,]0,1]$,

$$\frac{1}{2\lambda}\,\|J_\lambda(u) - u\|^2 \le \alpha - \langle J_\lambda(u), v\rangle - \eta \le \alpha + \|J_\lambda(u)\|\,\|v\| + |\eta|, \tag{4.35}$$

and hence $\|J_\lambda(u) - u\|^2 \leq \lambda(\alpha + \beta\|v\| + |\eta|)$. Thus, $J_\lambda(u) \to u$ as $\lambda \to 0$. Since f is lower semicontinuous, we have $f(x) \leq \liminf_{\lambda\to 0} f(J_\lambda(x)) \leq \limsup_{\lambda\to 0} f(J_\lambda(x)) \leq \lim_{\lambda\to 0} f_\lambda(u)$. The statement follows. □

Corollary 4.5.13. *Let $C \subset X$ be a nonempty closed and convex set. Then the square of the distance function d_C is convex, 2-Lispchitz smooth and $\nabla d_C^2(x) = 2(x - P_C(x))$.*

Proof. Let $f = \iota_C$ in Definition 4.5.10. Then $f_{1/2} = d_C^2$ and Proposition 4.5.12(iii) yields the statement. □

Lecture 5

Nonsmooth optimization: The proximal gradient method.

5.1 The proximity operator

Motivated by the use of non-smooth regularization techniques in inverse problems we introduce the proximity operator of a convex function.

Definition 5.1.1. Let $f \in \Gamma_0(X)$. Then, the *proximity operator of* f is

$$\mathrm{prox}_f \colon X \to X, \qquad \mathrm{prox}_{\gamma f}(x) = \mathrm{argmin}_{y \in X} \left\{ f(y) + \frac{1}{2}\|y - x\|^2 \right\}.$$

Note that the operator prox_f is well-defined, since the function $y \mapsto f(y) + (1/2)\,\|y - x\|^2$ is closed and strongly convex, hence it has a unique minimizer. Moreover, let us check that $\mathrm{prox}_f = (\mathrm{Id} + \partial f)^{-1}$. Using the sum rule for the subdifferential, which holds since the square norm is differentiable, we derive

$$\begin{aligned} z = \mathrm{prox}_f(x) &\iff 0 \in \partial f(z) + z - x \\ &\iff x \in (\mathrm{Id} + \partial f)(z) \\ &\iff z \in (\mathrm{Id} + \partial f)^{-1}(x). \end{aligned}$$

This shows that $(\mathrm{Id} + \gamma \partial f)^{-1}(x)$ is actually a singleton and its unique element is $\mathrm{prox}_{\gamma f}(x)$. Note that for every $x \in X$, $\mathrm{prox}_{\gamma f}(x) \in \mathrm{dom} f$, since the minimizer of $\gamma f + (1/2)\,\|\cdot\|^2$ is clearly in the domain of f.

Example 5.1.2. Let C be a closed and convex set. The proximity operator of ι_C is the projection on C. The projection is nonexpansive (and, indeed, firmly nonexpansive), but in general not a contraction, unless C is a singleton.

Proposition 5.1.3. *Let $f \in \Gamma_0(X)$. Then*

$$(\forall x, y \in X) \quad \|\mathrm{prox}_f(x) - \mathrm{prox}_f(y)\|^2 \leq \langle x - y, \mathrm{prox}_f(x) - \mathrm{prox}_f(y) \rangle. \tag{5.1}$$

So, in particular, prox_f *is nonexpansive. Property* (5.1) *is called* firm nonexpansiveness.

Salzo Saverio (saverio.salzo@iit.it) Istituto Italiano di Tecnologia. Via E. Melen, 16152, Genoa, Italy.

Proof. Let $x, y \in X$ and set $p_x = \operatorname{prox}_f(x)$ and $p_y = \operatorname{prox}_f(y)$. Then, by Fermat's rule, we have

$$x - p_x \in \partial f(p_x) \text{ and } y - p_y \in \partial f(p_y).$$

Therefore

$$\begin{aligned} f(p_y) &\geq f(p_x) + \langle x - p_x, p_y - p_x \rangle \\ f(p_x) &\geq f(p_y) + \langle y - p_y, p_x - p_y \rangle \end{aligned}$$

and summing $f(p_y) + f(p_x) \geq f(p_x) + f(p_y) + \langle y - p_y - x + p_x, p_x - p_y \rangle$. Then the statement follows. □

Remark 5.1.4. Proximity operators and Moreau envelops are strongly related. Indeed recalling Definition 4.5.10 and Proposition 4.5.12, it follows that the gradient of f_λ is given as

$$\nabla f_\lambda(u) = \frac{u - \operatorname{prox}_{\gamma f}(u)}{\lambda} \in \partial f(\operatorname{prox}_{\lambda f}(x)). \tag{5.2}$$

In the following we provide important properties of proximity operators.

Proposition 5.1.5 (Separable sum)**.** *Let $(X_i)_{1 \leq i \leq n}$ be Euclidean spaces and let, for every $i = 1, \ldots, n$, $f_i \in \Gamma_0(X_i)$. Set $X = \prod_{i=1}^n X_i$ and define $f \colon X \to]-\infty, +\infty]$ by $f(x) = \sum_{i=1}^n f_i(x_i)$, for every $x = (x_1, \ldots, x_n) \in X$. Then*

$$(\forall\, x = (x_1, \ldots, x_n) \in X) \qquad \operatorname{prox}_f(x) = (\operatorname{prox}_{f_1}(x_1), \ldots, \operatorname{prox}_{f_n}(x_n)). \tag{5.3}$$

Example 5.1.6 (Proximity operator of the ℓ_1 norm)**.** Let $X = \mathbb{R}^d$. The ℓ_1 norm on X is separable, thus the proximity operator can be computed componentwise, so it is enough to compute the proximity operator of the absolute value in $\mathbb{R}$. Let $\gamma > 0$. By definition, for every $t \in \mathbb{R}$, $\operatorname{prox}_{\gamma|\cdot|}(t) = (\mathrm{Id} + \gamma \partial |\cdot|)^{-1}(t)$. Thus, if we make the plot of the graph of $\mathrm{Id} + \gamma \partial |\cdot|$ and invert it, we discover that

$$\operatorname{prox}_{\gamma|\cdot|}(t) = \begin{cases} t - \gamma & \text{if } t > \gamma \\ 0 & \text{if } |t| \leq \gamma \\ t + \gamma & \text{if } t < -\gamma. \end{cases} \tag{5.4}$$

Thus, it follows from Proposition 5.1.5 that, for every $x \in \mathbb{R}^n$ and every $i = 1, \ldots, n$, $(\operatorname{prox}_{\gamma \|\cdot\|_1}(x))_i = \operatorname{prox}_{\gamma|\cdot|}(x_i)$.

Example 5.1.7 (Proximity operator of the $\ell_1 + \ell_2$ norm)**.**

$$f(x) = \|x\|_1 + \frac{\lambda}{2}\|x\|_2^2$$

$$\operatorname{prox}_{\gamma f}(x) = \operatorname{prox}_{(\gamma/(\gamma\lambda+1))\|\cdot\|_1}(x/(\gamma\lambda + 1))$$

$$(\operatorname{prox}_{\gamma f}(x))_i = \begin{cases} (x_i - \gamma)/(\gamma\lambda + 1) & \text{if } x_i > \gamma \\ 0 & \text{if } |x_i| \leq \gamma \\ (x_i + \gamma)/(\gamma\lambda + 1) & \text{if } x_i < -\gamma \end{cases}$$

Proposition 5.1.8 (Properties of the proximity operator)**.** *Let $f \in \Gamma_0(X)$ and let $\gamma > 0$. Then the following holds*

(i) *(linear perturbation) Let $g = f + \langle \cdot, u \rangle + a$, with $u \in X$ and $a \in \mathbb{R}$. Let $\gamma > 0$. Then*

$$\operatorname{prox}_{\gamma g}(x) = \operatorname{prox}_{\gamma f}(x - \gamma u).$$

(ii) *Let $g(x) = f(ax + b)$, with $a \in \mathbb{R}, a \neq 0$ and $b \in X$. Then*

$$\operatorname{prox}_{\gamma g}(x) = (\operatorname{prox}_{a^2 \gamma f}(ax + b) - b)/a.$$

(iii) *(composition with an orthogonal matrix) Let $g = f \circ L$, with $L \colon X \to X$ a bijective linear map such that $L^* = L^{-1}$. Then*

$$(\forall x \in X) \quad \operatorname{prox}_{\gamma g}(x) = L^* \operatorname{prox}_{\gamma f}(Lx).$$

Proof. (i): Just write the Fermat rule to derive that, if $p = \operatorname{prox}_{\gamma g}(x) = \operatorname{argmin}_{y \in X} \{\gamma f(y) + \gamma \langle u, y \rangle + a + \frac{1}{2}\|y - x\|^2\}$

$$\begin{aligned} 0 \in \gamma \partial f(p) + \gamma u + p - x &\iff x - \gamma u \in (\mathrm{Id} + \gamma \partial f)(p) \\ &\iff p = \operatorname{prox}_{\gamma f}(x - \gamma u). \end{aligned}$$

(ii): We have:

$$\begin{aligned} p = \operatorname{prox}_{\gamma g}(x) &\iff p = \operatorname{argmin}_{y \in X} \left\{ \gamma f(ay + b) + \frac{1}{2}\|y - x\|^2 \right\} \\ &\iff p = \operatorname{argmin}_{y \in X} \left\{ \gamma f(ay + b) + \frac{1}{2a^2} \|ay + b - (ax + b)\|^2 \right\} \\ &\iff p = \operatorname{argmin}_{y \in X} \left\{ \gamma a^2 f(ay + b) + \frac{1}{2} \|ay + b - (ax + b)\|^2 \right\} \\ &\iff ap + b = \operatorname{prox}_{a^2 \gamma f}(ax + b) \\ &\iff p = (\operatorname{prox}_{a^2 \gamma f}(ax + b) - b)/a. \end{aligned}$$

(iii) Let $x \in X$.

$$\begin{aligned} p = \operatorname{prox}_{\gamma g}(x) &\iff p = \operatorname{argmin}_{y \in X} \left\{ \gamma f(Ly) + \frac{1}{2}\|y - x\|^2 \right\} \\ &\iff 0 \in \gamma L^* \partial f(Lp) + p - x \\ &\iff x - p \in L^{-1} \partial f(Lp) \\ &\iff Lx \in \gamma \partial f(Lp) + Lp \\ &\iff p = L^* \operatorname{prox}_{\gamma f}(Lx) \end{aligned}$$

□

Remark 5.1.9. Regarding 5.1.8(iii), in general, if L is not orthogonal, we can apply a gradient descent on the dual of the minimization problem defining the prox to compute it approximately.

We now introduce an important identity, that is, the Moreau's decomposition formula. Let V be a linear subspace of X. Then we know that x can be uniquely decomposed in two orthogonal components, $P_V x$ and $P_{V^\perp} x$ such that:

$$x = x_V + x_{V^\perp} = P_V x + P_{V^\perp}(x). \tag{5.5}$$

If we set $f = \iota_V$, we first note that

$$(\iota_V)^*(u) = \sup_{x \in X} \langle u, x \rangle - \iota_V(x) = \iota_{V^\perp}(u).$$

Thus, we can rewrite (5.5) as

$$x = \mathrm{prox}_{\iota_V}(x) + \mathrm{prox}_{\iota_{V^\perp}}(x).$$

This last formula can be generalized to every convex function.

Theorem 5.1.10 (Moreau's decomposition)**.** *Let $f \in \Gamma_0(X)$ and let $x \in X$. Then*

$$x = \mathrm{prox}_f(x) + \mathrm{prox}_{f^*}(x).$$

More generally,

$$(\forall \gamma > 0) \quad x = \mathrm{prox}_{\gamma f}(x) + \gamma \mathrm{prox}_{f^*/\gamma}(x/\gamma).$$

Proof.

$$\begin{aligned} p = \mathrm{prox}_f(x) &\iff x - p \in \partial f(p) \\ &\iff p \in \partial f^*(x - p) \\ &\iff x - (x - p) \in \partial f^*(x - p) \\ &\iff x - p = \mathrm{prox}_{f^*}(x). \end{aligned}$$

□

Example 5.1.11 (The proximity operator of the Euclidean norm)**.** We want to compute the prox of the norm of X (which is an Euclidean space). First note that

$$\|x\| = \sup_{\|u\| \leq 1} \langle u, x \rangle = \sigma_{B_1(0)}(x).$$

Hence,

$$\|\cdot\| = \sigma_{B_1(0)} = (\iota_{B_1(0)})^*.$$

Therefore, it follows from Theorem 5.1.10 that

$$\mathrm{prox}_{\|\cdot\|}(x) = x - \mathrm{prox}_{\iota_{B_1(0)}}(x) = x - P_{B_1(0)}(x).$$

More explicitly:

$$\mathrm{prox}_{\|\cdot\|}(x) = \begin{cases} x - \dfrac{x}{\|x\|} & \text{if } \|x\| > 1 \\ 0 & \text{if } \|x\| \leq 1. \end{cases}$$

Note that this operation corresponds to a vector soft thresholding, which reduces to (5.4) for $\dim X = 1$ and $\gamma = 1$.

Example 5.1.12 (The proximity operator of the group lasso norm)**.** Let $\mathcal{J} = \{J_1, \dots, J_l\}$ be a partition of $\{1, \dots, n\}$. We define a norm on $\mathbb{R}^n$ by considering

$$\|x\|_{\mathcal{J}} = \sum_{i=1}^{l} \left(\sum_{j \in J_i} |x_j|^2 \right)^{1/2}.$$

For every $x \in \mathbb{R}^d$, let us call $x_{J_i} = (x_j)_{j \in J_i} \in \mathbb{R}^{J_i}$ and denote by $\|\cdot\|_{J_i}$ the Euclidean norm on $\mathbb{R}^{J_i}$. Then

$$\|x\|_{\mathcal{J}} = \sum_{i=1}^{l} \|x_{J_i}\|_{J_i}.$$

Make a picture with 3 variables and groups $J_1 = \{1, 2\}$ and $J_2 = \{3\}$.

We next compute the proximity operator of $\|\cdot\|_{\mathcal{J}}$. First note that $\|\cdot\|_{\mathcal{J}}$ is the sum of functions depending on groups of variables x_{J_i}. Therefore the prox can be computed group-wise (thanks to the decomposability property (5.3)). Thus

$$(\mathrm{prox}_{\|\cdot\|_{\mathcal{J}}}(x))_{J_i} = \mathrm{prox}_{\|\cdot\|_{J_i}}(x_{J_i}),$$

so we already know the formula, thanks to the example above:

$$(\mathrm{prox}_{\|\cdot\|_{\mathcal{J}}}(x))_{J_i} = \begin{cases} x_{J_i} - \dfrac{x_{J_i}}{\|x_{J_i}\|} & \text{if } \|x_{J_i}\|_{J_i} > 1 \\ 0 & \text{otherwise} \end{cases}$$

The resulting prox operator is called *block soft-thresholding operator*.

5.2 Averaged operators

Definition 5.2.1. Let $\alpha \in]0, 1[$. Then $T \colon X \to X$ is an *α-averaged operator* if $T = (1-\alpha)\mathrm{Id} + \alpha R$ for some non expansive operator R. 1/2-averaged operators are also called *firmly nonexpansive*.

Remark 5.2.2. Definition 5.2.1 says that an averaged operator is a convex combination of the identity and of a nonexpansive operator. Averaged operators are indeed nonexpansive. This follows by the following chain of inequalities

$$\begin{aligned} \|Tx - Ty\| = \|(1-\alpha)(x-y) + \alpha(Rx - Ry)\| &\leq (1-\alpha)\|x-y\| + \alpha\|Rx - Ry\| \\ &\leq (1-\alpha)\|x-y\| + \alpha\|x-y\| = \|x-y\|. \end{aligned}$$

In the following we give several characterizations of the property of being an averaged operators.

Proposition 5.2.3. *Let $T \colon X \to X$ and $\alpha \in]0, 1[$. Then the following statements are equivalent*

(i) *T is α-averaged*

(ii) $\left(1 - \dfrac{1}{\alpha}\right)\mathrm{Id} + \dfrac{1}{\alpha}T$ *is nonexpansive*

(iii) *For every* $x, y \in X$, $\|Tx - Ty\|^2 \leq \|x - y\|^2 - \left(\frac{1}{\alpha} - 1\right) \|(\mathrm{Id} - T)x - (\mathrm{Id} - T)y\|^2$.

(iv) *For every* $(x, y) \in X^2$

$$\|Tx - Ty\|^2 + (1 - 2\alpha) \|x - y\|^2 \leq 2(1 - \alpha)\langle x - y, Tx - Ty\rangle.$$

Proof. (i) $\Leftrightarrow$ (ii): It follows from the following equivalence

$$T = (1 - \alpha)\mathrm{Id} + \alpha R \;\Leftrightarrow\; R = \left(1 - \frac{1}{\alpha}\right)\mathrm{Id} + \frac{1}{\alpha}T.$$

(ii) $\Leftrightarrow$ (iii): Set $R = (1-\alpha^{-1})\mathrm{Id}+\alpha^{-1}T$ and let $x, y \in X$. It follows from Lemma 1.3.11 that

$$\begin{aligned}\|Rx - Ry\|^2 &= \left\|(1 - \alpha^{-1})(x - y) + \alpha^{-1}(Tx - Ty)\right\|^2 \\ &= (1 - \alpha^{-1}) \|x - y\|^2 + \alpha^{-1} \|Tx - Ty\|^2 \\ &\quad - \alpha^{-1}(1 - \alpha^{-1}) \|(\mathrm{Id} - T)x - (\mathrm{Id} - T)y\|^2\end{aligned}$$

and hence

$$\begin{aligned}&\|Rx - Ry\|^2 - \|x - y\|^2 \\ &\qquad = \frac{1}{\alpha}\Big(\|Tx - Ty\|^2 - \|x - y\|^2 + \left(\frac{1}{\alpha} - 1\right) \|(\mathrm{Id} - T)x - (\mathrm{Id} - T)y\|^2 \Big).\end{aligned}$$

So inequality $\|Rx - Ry\|^2 - \|x - y\|^2 \leq 0$ is equivalent to that in (iii).

(iii) $\Leftrightarrow$ (iv): It follows from the inequality

$$\|(\mathrm{Id} - T)x - (\mathrm{Id} - T)y\|^2 = \|x - y\|^2 + \|Tx - Ty\|^2 - 2\langle x - y, Tx - Ty\rangle. \qquad \square$$

Remark 5.2.4. The inequality in Proposition 5.2.3(iii) shows that If T is α-averaged, then it is also α'-averaged for every $\alpha' > \alpha$. So it makes sense to consider the best (smallest) constant of averagedness.

Remark 5.2.5. Contractions are α-averaged operators. More precisely, if T is a contraction with constant q then it is $(q + 1)/2$-averaged. By Proposition 5.2.3 it is enough to prove that $(1 - 2/(q + 1))\mathrm{Id} + 2/(q + 1)T$ is nonexpansive

$$\begin{aligned}(\forall x, y \in X) \quad &\left\|\frac{q-1}{q+1}x + \frac{2}{q+1}Tx - \frac{q-1}{q+1}y - \frac{2}{q+1}Ty\right\| \leq \\ &\leq \frac{1-q}{q+1}\|x - y\| + \frac{2q}{q+1}\|x - y\| \leq \|x - y\|.\end{aligned}$$

Remark 5.2.6. In view of Definition 5.2.1 and Proposition 5.2.3(iv), an operator T is firmly nonexpansive if and only if

$$(\forall\, x, y \in X) \quad \|Tx - Ty\|^2 \leq \langle x - y, Tx - Ty\rangle.$$

Averaged operators are important since, provided that they have fixed points, the *Picard iteration* always converges to some fixed point. In the rest of the section we will prove this result.

Lemma 5.2.7 (Opial)**.** *Let $F \subset X$ be a nonempty subset. Let $(x_k)_{k\in\mathbb{N}}$ be a sequence in X and suppose that for every $y \in F$, $(\|x_k - y\|)_{k\in\mathbb{N}}$ is convergent and that the cluster points of $(x_k)_{k\in\mathbb{N}}$, belongs to F. Then $(x_k)_{k\in\mathbb{N}}$ converges to a point in F.*

Proof. The assumptions ensure that $(x_k)_{k\in\mathbb{N}}$ is bounded. Therefore, the set of cluster points of $(x_k)_{k\in\mathbb{N}}$ is nonempty. Let $y_1, y_2 \in X$ and let $(x_k^1)_{k\in\mathbb{N}}$ and $(x_k^2)_{k\in\mathbb{N}}$ be subsequences of $(x_k)_{k\in\mathbb{N}}$ such that $x_k^1 \to y_1$ and $x_k^2 \to y_2$. Then, for every $k \in \mathbb{N}$,

$$\begin{aligned}\|x_k - y_1\|^2 - \|y_1\|^2 &= \|x_k\|^2 - 2\langle x_k, y_1\rangle\\ \|x_k - y_2\|^2 - \|y_2\|^2 &= \|x_k\|^2 - 2\langle x_k, y_2\rangle\end{aligned}$$

and hence (subtracting)

$$2\langle x_k, y_2 - y_1\rangle = \|x_k - y_1\|^2 - \|x_k - y_2\|^2 - \|y_1\|^2 + \|y_2\|^2 . \tag{5.6}$$

Since y_1 and y_2 are cluster points of $(x_k)_{k\in\mathbb{N}}$, by assumptions, $y_1, y_2 \in F$ and $(\|x_k - y_1\|)_{k\in\mathbb{N}}$ and $(\|x_k - y_2\|)_{k\in\mathbb{N}}$ are convergent. Therefore by (5.6), we obtain that there exists $\beta \in \mathbb{R}$ such that $\langle x_k, y_2 - y_1\rangle \to \beta$. Now, since $x_k^i \to y_i$, $i = 1, 2$, we have $\langle x_k^i, y_2 - y_1\rangle \to \langle y_i, y_2 - y_1\rangle$, which implies

$$\langle y_1, y_2 - y_1\rangle = \beta = \langle y_2, y_2 - y_1\rangle$$

and hence $\|y_2 - y_1\|^2 = 0$. This proves that the set of cluster points of the sequence $(x_k)_{k\in\mathbb{N}}$ is a singleton. Thus the sequence $(x_k)_{k\in\mathbb{N}}$ converges. □

Theorem 5.2.8. *Let $\alpha \in]0, 1[$ and let $T\colon X \to X$ be an α-averaged operator such that the set S of its fixed points is nonempty. Define $(x_k)_{k\in\mathbb{N}}$ recursively as*

$$x_{k+1} = Tx_k,$$

starting from any $x_0 \in X$. Then the following hold.

(i) *For every $k \in \mathbb{N}$, $\|x_{k+1} - x_k\| \leq \|x_k - x_{k-1}\|$*

(ii) $\displaystyle\sum_{k=0}^{+\infty} \|Tx_k - x_k\|^2 < \frac{\alpha}{1-\alpha}\operatorname{dist}(x_0, S)^2$

(iii) *$(x_k)_{k\in\mathbb{N}}$ converges to some fixed point of T.*

Proof. (i): Since T is nonexpansive, $\|x_{k+1} - x_k\| = \|Tx_k - Tx_{k-1}\| \leq \|x_k - x_{k-1}\|$.

(ii): Let $x_* \in S$. Then, it follows from Proposition 5.2.3(iii) (with $x = x_k$ and $y = x_*$) that

$$(\forall\, k \in \mathbb{N})\ \ \|x_{k+1} - x_*\|^2 \leq \|x_k - x_*\|^2 - \left(\frac{1}{\alpha} - 1\right)\|x_k - Tx_k\|^2 . \tag{5.7}$$

Therefore,

$$\frac{1-\alpha}{\alpha}\sum_{k=0}^{+\infty}\|x_k - Tx_k\|^2 \leq \sum_{k=0}^{+\infty}\big(\,\|x_k - x_*\|^2 - \|x_{k+1} - x_*\|^2\,\big) = \|x_0 - x_*\|^2\,.$$

(iii): It follows from (ii) that $\|Tx_k - x_k\| \to 0$. Let x_* be a cluster point of $(x_k)_{k\in\mathbb{N}}$ and let $(x'_k)_{k\in\mathbb{N}}$ be a subsequence of $(x_k)_{k\in\mathbb{N}}$ such that $x'_k \to x_*$. Then $Tx'_k - x'_k \to 0$. Moreover, since T is continuous and $x'_k \to x_*$, we have $Tx'_k - x'_k \to Tx_* - x_*$ and hence $Tx_* - x_* = 0$, that is $x_* \in S$. So, the cluster points of $(x_k)_{k\in\mathbb{N}}$ belong to S. Finally, for every $x_* \in S$, $\|x_{k+1} - x_*\| = \|Tx_k - Tx_*\| \leq \|x_k - x_*\|$, that is, $(\|x_k - x_*\|)_{k\in\mathbb{N}}$ is decreasing and hence convergent. Then the statement follows from Lemma 5.2.7. □

Example 5.2.9. Let $T\colon X \to X$ be a nonexpansive operator admitting fixed points and let $\lambda \in]0,1[$. The *Krasnosel'skiĭ-Mann iteration* is defined as follows

$$x_0 \in X, \qquad x_{k+1} = x_k + \lambda(Tx_k - x_k). \tag{5.8}$$

Then $(x_k)_{k\in\mathbb{N}}$ converges to some fixed point of T. Indeed, define $T_\lambda = (1-\lambda)\mathrm{Id} + \lambda T$. Then, by definition, T_λ is λ-averaged, the fixed points of T_λ are the same as those of T, and, by (5.8), $x_{k+1} = T_\lambda x_k$. Therefore the statement follows from Theorem 5.2.8.

The properties of being averaged is preserved by compositions, as the following result shows.

Proposition 5.2.10. *Let $T_1\colon X \to X$ and $T_2\colon X \to X$ be two averaged operators, with constants α_1 and α_2 respectively. Then $T_1 \circ T_2$ is averaged with constant*

$$\alpha = \frac{\alpha_1 + \alpha_2 - 2\alpha_1\alpha_2}{1 - \alpha_1\alpha_2}.$$

Proof. Since $\alpha_1(1-\alpha_2) < (1-\alpha_2)$, we have $\alpha_1+\alpha_2 < 1+\alpha_1\alpha_2$ and hence $\alpha < 1$. Moreover $\alpha_1 + \alpha_2 - 2\alpha_1\alpha_2 > \alpha_1^2 + \alpha_2^2 - 2\alpha_1\alpha_2 = (\alpha_1 - \alpha_2)^2 \geq 0$; hence $\alpha > 0$. Let $(x,y) \in D_2^2$. Then, by a twice application of Proposition 5.2.3(iii) to T_1 and T_2, we have

$$\begin{aligned}
\|T_1T_2x - T_1T_2y\|^2 &\leq \|T_2x - T_2y\|^2 - \frac{1-\alpha_1}{\alpha_1}\big\|(\mathrm{Id} - T_1)T_2x - (\mathrm{Id} - T_1)T_2y\big\|^2 \\
&\leq \|x - y\|^2 - \frac{1-\alpha_2}{\alpha_2}\big\|(\mathrm{Id} - T_2)x - (\mathrm{Id} - T_2)y\big\|^2 \\
&\quad - \frac{1-\alpha_1}{\alpha_1}\big\|(\mathrm{Id} - T_1)T_2x - (\mathrm{Id} - T_1)T_2y\big\|^2
\end{aligned} \tag{5.9}$$

Let $\tau = (1-\alpha_1)/\alpha_1 + (1-\alpha_2)/\alpha_2$. Then it follows from Lemma 1.3.11, that

$$\begin{aligned}
&\frac{1-\alpha_2}{\tau\alpha_2}\big\|(\mathrm{Id} - T_2)x - (\mathrm{Id} - T_2)y\big\|^2 + \frac{1-\alpha_1}{\tau\alpha_1}\big\|(\mathrm{Id} - T_1)T_2x - (\mathrm{Id} - T_1)T_2y\big\|^2 \\
&= \Big\|\frac{1-\alpha_2}{\tau\alpha_2}\big((\mathrm{Id} - T_2)x - (\mathrm{Id} - T_2)y\big) - \frac{1-\alpha_1}{\tau\alpha_1}\big((\mathrm{Id} - T_1)T_2x - (\mathrm{Id} - T_1)T_2y\big)\Big\|^2 \\
&\qquad + \frac{(1-\alpha_2)(1-\alpha_1)}{\tau^2\alpha_1\alpha_2}\,\|(\mathrm{Id} - T_1T_2)x - (\mathrm{Id} - T_1T_2)y\|^2 \\
&\geq \frac{(1-\alpha_2)(1-\alpha_1)}{\tau^2\alpha_1\alpha_2}\,\|(\mathrm{Id} - T_1T_2)x - (\mathrm{Id} - T_1T_2)y\|^2\,.
\end{aligned}$$

Thus, combining the above inequality with (5.9) we obtain

$$\|T_1T_2x - T_1T_2y\|^2 \leq \|x-y\|^2 \\ - \frac{(1-\alpha_2)(1-\alpha_1)}{\tau\alpha_1\alpha_2}\|(\mathrm{Id}-T_1T_2)x-(\mathrm{Id}-T_1T_2)y\|^2.$$

The statement follows from the fact that

$$\frac{(1-\alpha_2)(1-\alpha_1)}{\tau\alpha_1\alpha_2} = \frac{1-\alpha_1-\alpha_2+\alpha_1\alpha_2}{\alpha_1+\alpha_2-2\alpha_1\alpha_2} = \frac{1}{\alpha}-1 = \frac{1-\alpha}{\alpha}.$$

□

Remark 5.2.11. Note that

$$\frac{\alpha_1+\alpha_2-2\alpha_1\alpha_2}{1-\alpha_1\alpha_2} > \max\{\alpha_1,\alpha_2\}. \tag{5.10}$$

Example 5.2.12. Let $f\colon X \to \mathbb{R}$ be differentiable with Lipschitz continuous gradient with constant L and let $\gamma < 2/L$. We will see in Proposition 5.3.1 that $I-\gamma\nabla f$ is α_2-averaged with $\alpha_2 = \gamma L/2$. Moreover, if $g \in \Gamma_0(X)$ we will define the so called *proximity operator* with respect to γg, denoted by $\mathrm{prox}_{\gamma g}$. In Proposition 5.1.3 we will prove that $\mathrm{prox}_{\gamma g}$ is firmly nonexpansive, that is α_1-averaged with $\alpha_1 = 1/2$. Therefore $T = \mathrm{prox}_{\gamma g}\circ(I-\gamma\nabla f)$ is α-averaged with

$$\alpha = \frac{1/2+\gamma L/2-\gamma L/2}{1-(1/2)(\gamma L/2)} = \frac{2}{4-\gamma L}.$$

5.3 The forward-backward algorithm

Let $f\colon X \to \mathbb{R}$ be a convex differentiable function and let $g\colon X \to]-\infty,+\infty]$ be a closed, convex, and proper function. Let $F = f+g$. We aim at the following composite minimization problem

$$\min_{x\in X} f(x)+g(x). \tag{5.11}$$

The forward-backward algorithm is defined as follows. Let $x_0 \in X$ and $\gamma > 0$, then

$$(\forall\, k\in\mathbb{N}) \qquad x_{k+1} = \mathrm{prox}_{\gamma g}\big(x_k - \gamma\nabla f(x_k)\big). \tag{5.12}$$

The algorithm can be seen as a Picard iteration of the following operator

$$T = \mathrm{prox}_{\gamma g}\circ(\mathrm{Id}-\gamma\nabla f), \tag{5.13}$$

which is the composition of the proximity operator of γg and the operator $\mathrm{Id}-\gamma\nabla f$. We also note that the fixed points of T are the minimizers of $f+g$. Indeed

$$x = Tx \Leftrightarrow x = \mathrm{prox}_{\gamma g}(x-\gamma\nabla f(x)) \Leftrightarrow x-\gamma\nabla f(x)-x \in \partial\gamma g(x) \Leftrightarrow 0 \in \partial(f+g)(x).$$

So we need to study the operator T. We already know that $\mathrm{prox}_{\gamma g}$ is $(1/2)$-averaged. The following result concerns the operator $\mathrm{Id}-\gamma\nabla f$.

Proposition 5.3.1. *Let $f\colon X \to \mathbb{R}$ be differentiable and let $L > 0$. Let $\gamma > 0$ and set $T_1 = \mathrm{Id} - \gamma\nabla f$. Then, the L-Lipschitz continuity of ∇f is equivalent to the property*

$$(\forall\, x, y \in X)\ \ \|T_1 x - T_1 y\|^2 \leq \|x - y\|^2 - \left(\frac{2}{\gamma L} - 1\right) \|(\mathrm{Id} - T_1)x - (\mathrm{Id} - T_1)y\|^2 . \quad (5.14)$$

*In particular, if $\gamma < L/2$, T_1 is a α-*averaged operator*, with $\alpha = \gamma L/2 < 1$.*

Proof. Multiplying by $\gamma^2 L$ the inequality in Theorem 2.1.12(iv) and replacing $\gamma\nabla f$ with $\mathrm{Id} - T_1$, we obtain

$$\|(\mathrm{Id} - T_1)x - (\mathrm{Id} - T_1)y\|^2 \leq \gamma L \langle (\mathrm{Id} - T_1)x - (\mathrm{Id} - T_1)y, x - y\rangle.$$

Then, using the identity

$$\begin{aligned} 2\langle (\mathrm{Id} - T_1)x - (\mathrm{Id} - T_1)y, x - y\rangle & \\ = \|(\mathrm{Id} - T_1)x - (\mathrm{Id} - T_1)y\|^2 + \|x - y\|^2 - \|T_1 x - T_1 y\|^2 , & \quad (5.15) \end{aligned}$$

the statement follows. □

Then, T is a composition of two averaged operators and hence, in virtue of Proposition 5.2.10 T is still an α-averaged operator with $\alpha = 2/(4 - \gamma L)$ (see Example 5.2.12). In particular the sequence $(x_k)_{k\in\mathbb{N}}$ converges to some minimizer of $f + g$ (fixed point of T) and

$$\sum_{k\in\mathbb{N}} \|x_{k+1} - x_k\|^2 \leq \frac{2}{2 - \gamma L} \mathrm{dist}(x_0, S_*)^2, \quad (5.16)$$

where S^* is the set of minimizers of $f + g$ (equal to the set of fixed points of T).

Next we show that also the function values converge.

Proposition 5.3.2. *Let $k \in \mathbb{N}$. Then*

$$\left(\frac{1}{\gamma} - \frac{L}{2}\right) \|x_{k+1} - x_k\|^2 \leq (f + g)(x_k) - (f + g)(x_{k+1}). \quad (5.17)$$

Thus, if $\gamma \leq 2/L$, then $f(x_{k+1}) \leq f(x_k)$, that is, the algorithm is descending.

Proof. By Theorem 2.1.12(ii), we have

$$f(x_{k+1}) \leq f(x_k) + \langle x_{k+1} - x_k, \nabla f(x_k)\rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2 ,$$

and hence

$$f(x_k) - f(x_{k+1}) \geq \langle x_k - x_{k+1}, \nabla f(x_k)\rangle - \frac{L}{2} \|x_{k+1} - x_k\|^2 . \quad (5.18)$$

Moreover, since $(x_k - x_{k+1})/\gamma - \nabla f(x_k) \in \partial g(x_{k+1})$, we have

$$\begin{aligned} g(x_k) - g(x_{k+1}) &\geq \langle (x_k - x_{k+1})/\gamma - \nabla f(x_k), x_k - x_{k+1}\rangle \\ &= (1/\gamma)\langle x_k - x_{k+1}, x_k - x_{k+1}\rangle - \langle \nabla f(x_k), x_k - x_{k+1}\rangle. \quad (5.19) \end{aligned}$$

Therefore, summing (5.18) and (5.19) we obtain (5.17). □

Lemma 5.3.3. *For any $x, z \in X$, $y \in \mathrm{dom} g$ and for any $u \in \partial g(x)$. We have*

$$F(z) \geq F(x) + \langle z - x, \nabla f(y) + u \rangle - \frac{L}{2} \|x - y\|^2 .$$

Proof. Let $x, z \in X$ and let $y \in \mathrm{dom} g$. Then, it follows from Theorem 2.1.12 that

$$f(y) \geq f(x) - \langle \nabla f(y), x - y \rangle - \frac{L}{2} \|x - y\|^2 .$$

Hence, since f is convex,

$$f(z) \geq f(y) + \langle \nabla f(y), z - y \rangle \geq f(x) + \langle z - x, \nabla f(y) \rangle - \frac{L}{2} \|x - y\|^2 .$$

Now, since $u \in \partial g(x)$, $g(z) \geq g(x) + \langle u, z - x \rangle$. Summing the two previous inequality the statement follows. □

Lemma 5.3.4. *Let $(a_k)_{k \in \mathbb{N}}$ be a decreasing sequence in $\mathbb{R}_+$. If $\sum_{k=0}^{+\infty} a_k < +\infty$, then*

$$(\forall\, k \in \mathbb{N}) \quad a_k \leq \frac{1}{k+1} \sum_{k=0}^{+\infty} a_k, \ \text{and } a_k = o\Big(\frac{1}{k+1}\Big). \tag{5.20}$$

Proof. Let $k \in \mathbb{N}$. Since, for every $i = 0, 1, \ldots, k$, $a_k \leq a_i$, we have $\sum_{i=0}^{k} a_i \geq (k+1) a_k$, hence the first part of the statement. As regard the second part, we note that, for every integer $k \geq 2$, we have $\sum_{i=\lceil k/2 \rceil}^{+\infty} a_i \geq \sum_{i=\lceil k/2 \rceil}^{k} a_i \geq (k + 1 - \lceil k/2 \rceil) a_k \geq \frac{k+1}{2} a_k$. Therefore, $(k+1) a_k \leq 2 \sum_{i=\lceil k/2 \rceil}^{+\infty} a_i \to 0$ as $k \to +\infty$. □

The following theorem provide full convergence results concerning the forward-backward algorithm.

Theorem 5.3.5. *Suppose that $\gamma < 2/L$. Let S_* be the set of minimizers of $F := f + g$ and suppose that $S_* \neq \varnothing$. Then the following statements hold*

(i) $\displaystyle\sum_{k \in \mathbb{N}} \|x_{k+1} - x_k\|^2 \leq \frac{2}{2 - \gamma L} \mathrm{dist}(x_0, S_*)^2.$

(ii) *For every $k \in \mathbb{N}$ and for every $x \in X$,*

$$\|x_{k+1} - x\|^2 \leq \|x_k - x\|^2 + 2\gamma\big(F(x) - F(x_{k+1})\big) + (\gamma L - 1) \|x_{k+1} - x_k\|^2$$

(iii) *Let $F_* = \inf_{x \in X} (f+g)(x)$. Then $F(x_{k+1}) - F_* = o(1/(k+1))$ and, for every $k \in \mathbb{N}$,*

$$F(x_{k+1}) - F_* \leq \frac{\mathrm{dist}(x_0, S_*)^2}{k+1} \cdot \begin{cases} \dfrac{1}{2\gamma} & \text{if } \gamma \leq 1/L \\ \dfrac{L}{2} \dfrac{1}{2 - \gamma L} & \text{if } 1/L < \gamma < 2/L. \end{cases} \tag{5.21}$$

(iv) *The sequence $(x_k)_{k \in \mathbb{N}}$ converges to some $x_* \in S_*$.*

Proof. (i): We already saw this in (5.16).

(ii): Let $x \in X$ and let $k \in \mathbb{N}$. It follows from (5.12) that $u := (x_k - x_{k+1})/\gamma - \nabla f(x_k) \in \partial g(x_{k+1})$, hence

$$\frac{x_k - x_{k+1}}{\gamma} = \nabla f(x_k) + u, \qquad u \in \partial g(x_{k+1}). \tag{5.22}$$

Thus, by Lemma 5.3.3, we have that

$$\begin{aligned} F(x) &\geq F(x_{k+1}) + \langle x - x_{k+1}, \nabla f(x_k) + u\rangle - \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= F(x_{k+1}) + \frac{1}{\gamma}\langle x - x_{k+1}, x_k - x_{k+1}\rangle - \frac{L}{2}\|x_{k+1} - x_k\|^2; \end{aligned}$$

and using the identity $\|x_k - x\|^2 = \|x_k - x_{k+1}\|^2 + \|x_{k+1} - x\|^2 + 2\langle x_{k+1} - x_k, x - x_{k+1}\rangle$ we have

$$\begin{aligned} F(x) - F(x_{k+1}) &\geq \frac{1}{2\gamma}\Big[\|x_k - x_{k+1}\|^2 + \|x_{k+1} - x\|^2 - \|x_k - x\|^2\Big] - \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= \frac{1}{2\gamma}\big[(1 - \gamma L)\|x_k - x_{k+1}\|^2 + \|x_{k+1} - x\|^2 - \|x_k - x\|^2\big]. \end{aligned}$$

Therefore,

$$\|x_{k+1} - x\|^2 \leq \|x_k - x\|^2 + 2\gamma\big(F(x) - F(x_{k+1})\big) - (1 - \gamma L)\|x_k - x_{k+1}\|^2 \tag{5.23}$$

and the statement follows.

(iii): Let $x_* \in S_*$. Then, it follows from (ii) that, for every $k \in \mathbb{N}$,

$$0 \leq 2\gamma\big(F(x_{k+1}) - F(x_*)\big) \leq \|x_k - x_*\|^2 - \|x_{k+1} - x_*\|^2 + (\gamma L - 1)_+\|x_k - x_{k+1}\|^2.$$

Thus, summing and using (i), we have

$$\begin{aligned} 2\gamma\sum_{k=0}^{+\infty}\big(F(x_{k+1}) - F(x_*)\big) &\leq \|x_0 - x_*\|^2 + \frac{2\big(\gamma L - 1\big)_+}{2 - \gamma L}\|x_0 - x_*\|^2 \\ &= \|x_0 - x_*\|^2 \cdot \begin{cases} 1 & \text{if } \gamma \leq 1/L \\ \dfrac{\gamma L}{2 - \gamma L} & \text{if } 1/L < \gamma < 2/L. \end{cases} \end{aligned}$$

Then, since $\big(F(x_{k+1}) - F(x_*)\big)_{k\in\mathbb{N}}$ is decreasing and positive, the statement follows from Lemma 5.3.4.

(iv): This has been already discussed before equation (5.16). □

Remark 5.3.6. It follows from (5.21) that the best bound is achieved when $\gamma = 1/L$.

Remark 5.3.7 (Backtracking)**.** In many situation it is difficult, if not impossible, to compute the Lipschitz constant L. In such cases a backtracking line search procedure will overcome the issue. Suppose that x_k is given. Let $\delta \in]0,1[$, $\bar{\gamma} > 0$ and $\sigma \in]0,1[$. Then we determine

$$\gamma_k = \max\Big\{\gamma := \bar{\gamma}\sigma^i \;\Big|\; i \in \mathbb{N} \text{ and } f\big(x_k - \gamma\nabla f(x_k)\big) \leq f(x_k) - (1 - \delta)\gamma\|\nabla f(x_k)\|^2\Big\}.$$

It is easy to see that the above set is nonempty and γ_k can be determined by a finite procedure. Indeed one can determine $\gamma = \bar{\gamma}\sigma^i$ (for some $i \in \mathbb{N}$) such that $\gamma \leq 2\delta/L$, if i is large enough (but finite). Then $(1-\delta) \leq (1-\gamma L/2)$ and hence

$$\gamma(1-\delta)\left\|f(x_k)\right\|^2 \leq \gamma(1-\gamma L/2)\left\|\nabla f(x_k)\right\|^2 \leq f(x_k) - f\big(x_k - \gamma\nabla f(x_k)\big).$$

Moreover it follows from the definition of γ_k that

$$f\big(x_k - \gamma_k/\sigma\nabla f(x_k)\big) > f(x_k) - (1-\delta)\gamma_k/\sigma\left\|\nabla f(x_k)\right\|^2.$$

Thus, γ_k/σ can not be less than $2\delta/L$, that is, there must be $\gamma_k > 2\sigma\delta/L$. So $\inf_k \gamma_k \geq 2\sigma\delta/L > 0$. In this case we can still proceed as in Remark **??**. Indeed we have

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \gamma_k(1-\delta)\left\|\nabla f(x_k)\right\|^2 \\ &\leq f(x_k) - \gamma_k\mu 2(1-\delta)(f(x_k) - f(x_*)). \end{aligned}$$

Therefore

$$f(x_{k+1}) - f(x_*) \leq \big(1 - \gamma_k\mu 2(1-\delta)\big)\big(f(x_k) - f(x_*)\big). \tag{5.24}$$

Moreover, since $-\gamma_k \leq -2\sigma\delta/L$, we have

$$\big(1 - \gamma_k\mu 2(1-\delta)\big) \leq \big(1 - 4\sigma\mu\delta(1-\delta)/L\big) < 1,$$

since $0 < 4\sigma\mu\delta(1-\delta)/L \leq \sigma\mu/L < 1$ (we used that $\delta(1-\delta) \leq 1/4$). In the end (**??**) still holds with $q = \big(1 - 4\sigma\mu\delta(1-\delta)/L\big)^{1/2}$. Note that again the best value of q is for $\delta = 1/2$, which gives $q = (1-\sigma\mu/L)^{1/2}$.

Forward backward for LASSO We consider the minimization problem

$$\min_{x\in\mathbb{R}^d} \frac{1}{2}\|Ax - y\|^2 + \lambda\|x\|_1. \tag{5.25}$$

Then, the forward-backward algorithm yields the following algorithm.

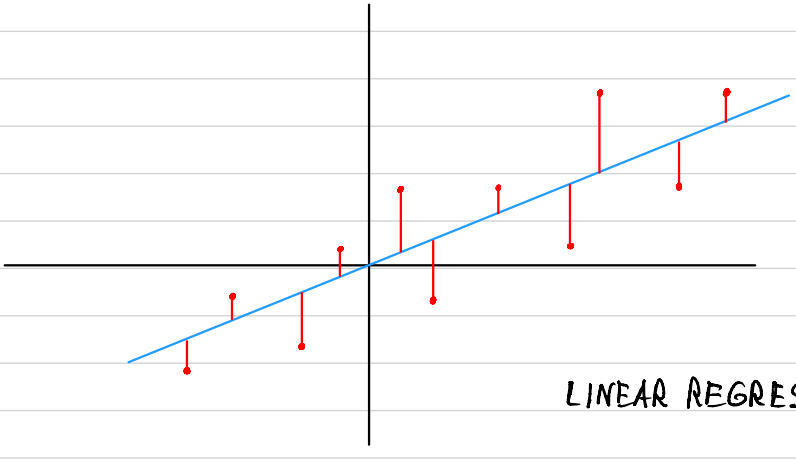
Algorithm 5.3.8 (*Forward-backward for ℓ_1 regularized least squares*)**.** Let $\gamma \in]0, 2/\|A^*A\|[$ and $x_0 \in X$.

$$\begin{array}{l} \text{For } k = 0, 1, \dots \\ \quad \lfloor\ x_{k+1} = \operatorname{prox}_{\gamma\lambda\|\cdot\|_1}(x_k - \gamma A^*(Ax_k - y)) \end{array} \tag{5.26}$$

Recall that the proximity operator of the ℓ_1 norm is the soft thresholding operator, see Example 5.1.6.

Lecture 6

The statistics of least squares and LASSO estimators.



$(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$

$f_\vartheta(X) = \langle \vartheta, X \rangle$ THE MODEL (LINEAR)

THE METHOD OF LEAST SQUARES

Legendre 1805, Gauss 1809 (1795)

$$\min_{\vartheta \in \mathbb{R}^d} \sum_{i=1}^{n} |f_\vartheta(X_i) - Y_i|^2.$$

Statistical assumptions

$x_1 \in \mathbb{R}^d$

x_2

$\vdots$

x_n

$\longrightarrow X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n \times d}$

no errors in the x_i's. (fixed design)

design matrix

design points

Y_1

Y_2

$\vdots$

Y_n

realization of

Y_1

Y_2

$\vdots$

Y_n

$Y_i = \langle \vartheta^*, x_i \rangle + \varepsilon_i$

ε_i are i.i.d and

$\varepsilon_i \in \mathcal{N}(0, \sigma^2)$.

$$Y = X\vartheta^* + \varepsilon \quad , \quad \varepsilon \in \mathcal{N}(0, \sigma^2 I_d)$$

$X : \mathbb{R}^d \to \mathbb{R}^n$, ϑ^* is unknown!

normal density $\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}$.

Given a generic estimator $\hat{\vartheta}$ (random variable dependent on the X_i's and Y_i's)

$$MSE(\hat{\vartheta}) = \frac{1}{n}\sum_{i=1}^{n} |\langle \hat{\vartheta}, X_i \rangle - \langle \vartheta^*, X_i \rangle|^2$$

THE MEAN SQUARED ERROR

$$= \frac{1}{n} \| X\hat{\vartheta} - X\vartheta^* \|^2$$

$$\hat{\vartheta}_{LS} \in \underset{\vartheta \in \mathbb{R}^d}{\operatorname{argmin}} \| X\vartheta - Y \|^2$$

$\hat{\vartheta}_{LS}$: random vector.

X: metrix (not random)

Y: random vector

Normal equations:

$$X^*(X\vartheta - Y) = 0 \iff X^*X\vartheta = X^*Y$$

$$\hat{\vartheta}_{LS} = (X^*X)^{\dagger} X^* Y$$

it is equal to $(X^*X)^{-1}$ if X^*X ($d \times d$) if full rank

$$\hat{\vartheta}_{LS} = \operatorname{argmin} \|\vartheta\|^2 \quad \text{(equivalent definition)}.$$
$$X^*X\vartheta = X^*Y.$$

$\hat{\vartheta}_{LS}$ it is a linear estimator since it depends linearly on the Y_i's.

$$MSE[\hat{\vartheta}_{LS}] := \frac{1}{n}\|X\hat{\vartheta}_{LS} - X\vartheta^*\|^2$$

$$\begin{aligned}
\mathbb{E}[X\hat{\vartheta}_{LS}] &= \mathbb{E}[X(X^*X)^\dagger X^*Y] \\
&= X(X^*X)^\dagger X^* \mathbb{E}[Y] \\
&= X\underbrace{(X^*X)^\dagger X^* X}\vartheta^* \\
&= X\vartheta^*
\end{aligned}$$

This shows that $X\hat{\vartheta}_{LS}$ is an unbiased estimator of $X\vartheta^*$.

[X^* stands for the transpose of X]

Theorem

$$\mathbb{E}[MSE[\hat{\vartheta}_{LS}]] \leq \frac{4r\sigma^2}{n}$$

where $r = \operatorname{rank}(X^*X)$

Proof:

$$MSE[\hat{\vartheta}_{LS}] = \frac{1}{n}\|X\hat{\vartheta}_{LS} - X\vartheta^*\|^2.$$

$$\|X\hat{\vartheta}_{LS} - Y\|^2 \leq \|X\vartheta^* - Y\|^2 = \|\varepsilon\|^2$$

$$\|X\hat{\vartheta}_{LS} - Y\|^2 = \|X\hat{\vartheta}_{LS} - X\vartheta^*\|^2 + \|\underbrace{X\vartheta^* - Y}_{-\varepsilon}\|^2$$

$$+ 2\langle X(\hat{\vartheta}_{LS} - \vartheta^*), -\varepsilon\rangle$$

$$\|X\hat{\vartheta}_{LS} - X\vartheta^*\|^2 \leq \cancel{\|\varepsilon\|^2} - \cancel{\|\varepsilon\|^2} + 2\langle \underbrace{X(\hat{\vartheta}_{LS} - \vartheta^*)}, \varepsilon\rangle$$

Let $v_1 \dots v_r$ be an orthonormal basis of $R(X)$.

$R(X) \cong N(X)^\perp = R(X^*) = R(X^*X)$

$X|_{N(X)^\perp} : N(X)^\perp \to R(X)$

$\dim R(X) = \dim R(X^*X) = \operatorname{rank}(X^*X)$

$$X(\hat{\vartheta}_{LS} - \vartheta^*) = \sum_{i=1}^{r} a_i v_i, \quad \text{for some } a \in \mathbb{R}^r.$$

$$\|X(\hat{\vartheta}_{LS} - \vartheta^*)\|^2 = \|\sum a_i v_i\|^2 = \sum_{i,j} a_i a_j \langle v_i, v_j \rangle$$

$$= \sum_{i=1}^{n} a_i^2 = \|a\|^2$$

$$V = [v_1 \quad v_r] \in \mathbb{R}^{n \times r} \qquad Va = \sum_{i=1}^{r} a_i v_i$$

$$\langle X(\hat{\vartheta}_{LS} - \vartheta^*), \varepsilon \rangle = \langle Va, \varepsilon \rangle = \langle a, V^*\varepsilon \rangle$$

$$\leq \|a\| \, \|V^*\varepsilon\| = \|X(\hat{\vartheta}_{LS} - \vartheta^*)\| \, \|V^*\varepsilon\|$$

$$V^*\varepsilon = \begin{bmatrix} \langle v_1, \varepsilon \rangle \\ \vdots \\ \langle v_r, \varepsilon \rangle \end{bmatrix} \qquad V : \mathbb{R}^r \to \mathbb{R}^n$$

$$V^* : \mathbb{R}^n \to \mathbb{R}^r.$$

$$\|X(\hat{\vartheta}_{LS} - X\vartheta^*\|^2 \leq 2 \|X(\hat{\vartheta}_{LS} - \vartheta^*)\| \, \|V^*\varepsilon\|$$

$$\frac{1}{n} \|X(\hat{\vartheta}_{LS} - \vartheta^*)\|^2 \leq \frac{4 \|V^*\varepsilon\|^2}{n} = \frac{4}{n} \sum_{i=1}^{r} \langle v_i, \varepsilon \rangle^2$$

$$\langle v_i, \varepsilon \rangle^2 = \langle \underbrace{(\varepsilon \otimes \varepsilon) v_i}_{(\varepsilon \varepsilon^T)}, v_i \rangle$$

$$\mathbb{E}\left[MSE[\hat{\theta}_{LS}]\right] \le \frac{4}{n} \sum_{i=1}^{r} \mathbb{E}\left[\langle v_i, \varepsilon \rangle^2\right]$$

$$= \frac{4}{n} \sum_{i=1}^{r} \langle \mathbb{E}[\varepsilon \otimes \varepsilon] v_i, v_i \rangle$$

$$= \frac{4}{n} \sum_{i=1}^{r} \langle \sigma^2 I_n v_i, v_i \rangle$$

$$= \frac{4}{n} \sigma^2 r. \qquad \square$$

Remark

If X^*X is full rank, then $r = d$, and hence

$$\mathbb{E}\left[MSE[\hat{\theta}_{LS}]\right] \le \frac{4\sigma^2 d}{n}$$

$4\sigma^2 d$ ← the number of unknowns

n ← the number of measurements.

In the classical setting (at Legendre's times) $n \gg d \Rightarrow \frac{4\sigma^2 d}{n}$ is small.

X^*X is full rank.

$\hat{\vartheta}_{LS} = (X^*X)^{-1} X^*Y.$

$$\frac{1}{n} \| X\hat{\vartheta}_{LS} - X\vartheta^* \|^2 = \frac{1}{n} \langle X^*X (\hat{\vartheta}_{LS} - \vartheta^*), \hat{\vartheta}_{LS} - \vartheta^* \rangle$$

$$\geq \frac{1}{n} \lambda_{min}(X^*X) \| \hat{\vartheta}_{LS} - \vartheta^* \|^2.$$

$MSE[X\hat{\vartheta}]$ $\qquad \hat{\mu} = X\hat{\vartheta}$, $\mu^* = X\vartheta^*$

$MSE[\hat{\mu}] = \frac{1}{n} \| \hat{\mu} - \mu^* \|^2.$

The LASSO ESTIMATOR

Prop 1

Let $X \sim \mathcal{N}(0, \sigma^2)$. Then

(i) $\forall s > 0 \quad \mathbb{E}[e^{sX}] = e^{\frac{s^2\sigma^2}{2}}$

(ii) $\forall t > 0 \quad \mathbb{P}(X > t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$.

Proof:

(i): Let $s > 0$. $\left[\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}\right]$

$$\mathbb{E}[e^{sX}] = \int e^{sx} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} dx \qquad \left(e^{-\frac{x^2}{2\sigma^2} - sx}\right)$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int e^{-\frac{1}{2\sigma^2}(x - s\sigma^2)^2} e^{\frac{s\sigma^2}{2}} dx$$

$$= e^{\frac{s\sigma^2}{2}} \underbrace{\frac{1}{\sigma\sqrt{2\pi}} \int e^{-\frac{1}{2\sigma^2}(x - s\sigma^2)^2} dx}_{\text{"1"}}$$

(ii) Let $s>0$.

Since $t \mapsto e^{st}$ is strictly increasing.

$$\mathbb{P}(X>t) = \mathbb{P}(e^{sX} > e^{st}) \underbrace{\leq}_{\text{Markov ineq}} \frac{\mathbb{E}[e^{sX}]}{e^{st}}$$

$$\overbrace{\leq}^{\text{(i)}} \exp\left(\frac{s^2\sigma^2}{2} - st\right) \quad (\forall s>0)$$

$$\mathbb{P}(X>t) \leq \exp\left(\inf_{s>0}\left(\frac{s^2\sigma^2}{2} - st\right)\right)$$

$$\inf_{s>0} \frac{s^2\sigma^2}{2} - st \overbrace{=}^{s\sigma^2 = t} \frac{t^2}{\sigma^4}\frac{\sigma^2}{2} - \frac{t^2}{\sigma^2} = -\frac{t^2}{2\sigma^2}.$$

□

Maximal inequalities

Theorem 1 Let $Z_i \in N(0, \sigma_i^2)$, $\sigma_i^2 \leq \sigma^2$

$Z_n = \max_{1\leq i\leq N} Z_i$. Then $\mathbb{E}[Z_n] \leq \sqrt{2\sigma^2 \log N}$

and $\mathbb{P}(Z_n > t) \leq N \exp\left(-\frac{t^2}{2\sigma^2}\right)$ ←

Proof:

I will prove only the second.

$$\{Z > t\} = \left\{ \max_{1 \le i \le N} Z_i > t \right\} = \bigcup_{i=1}^{N} \{Z_i > t\}$$

$$\mathbb{P}(\{Z > t\}) \le \sum_{i=1}^{N} \mathbb{P}(\{Z_i > t\})$$

$$\le \sum_{i=1}^{N} \exp\left(-\frac{t^2}{2\sigma_i^2}\right)$$

$$\underbrace{\sigma_i^2 \le \sigma^2}$$

$$\le N \exp\left(-\frac{t^2}{2\sigma^2}\right). \qquad \square$$

LASSO ESTIMATOR.

$$\hat{\vartheta}_{l_1} \in \operatorname*{argmin}_{\vartheta \in \mathbb{R}^d} \frac{1}{n} \|X\vartheta - Y\|^2 + \lambda \|\vartheta\|_1 \qquad (\lambda > 0)$$

Statistical analysis of the LASSO

1. $\|X\hat{\vartheta}_{\ell_1} - Y\|^2 + n\lambda\|\hat{\vartheta}_{\ell_1}\|_1 \leq \|X\vartheta^* - Y\|^2 + n\lambda\|\vartheta^*\|_1$.

$$= \|\varepsilon\|^2 + n\lambda\|\vartheta^*\|_1.$$

$$\|X\hat{\vartheta}_{\ell_1} - Y\|^2 = \|X\hat{\vartheta}_{\ell_1} - X\vartheta^*\|^2 + \|\underbrace{X\vartheta^* - Y}_{-\varepsilon}\|^2$$

$$+ 2\langle X(\hat{\vartheta}_{\ell_1} - \vartheta^*), -\varepsilon\rangle$$

$$\|X\hat{\vartheta}_{\ell_1} - X\vartheta^*\|^2 \leq \|\varepsilon\|^2 + n\lambda\|\vartheta^*\|_1 \overbrace{- n\lambda\|\hat{\vartheta}_{\ell_1}\|_1}^{} - \|\varepsilon\|^2 + 2\langle X(\hat{\vartheta}_{\ell_1} - \vartheta^*), \varepsilon\rangle$$

$$= n\lambda\left(\|\vartheta^*\|_1 - \|\hat{\vartheta}_{\ell_1}\|_1\right) + 2\langle \hat{\vartheta}_{\ell_1} - \vartheta^*, X^*\varepsilon\rangle$$

$$\overbrace{\leq}^{\text{Hölder ineq.}} n\lambda\left(\|\vartheta^*\|_1 - \|\hat{\vartheta}_{\ell_1}\|_1\right) + 2\left(\|\hat{\vartheta}_{\ell_1}\|_1 + \|\vartheta^*\|_1\right)\|X^*\varepsilon\|_\infty.$$

$$= \|\vartheta^*\|_1\left(2\|X^*\varepsilon\|_\infty + n\lambda\right) + \|\hat{\vartheta}_{\ell_1}\|_1\left(2\|X^*\varepsilon\|_\infty - n\lambda\right)$$

$X^* \varepsilon \in \mathbb{R}^d$.

$$(X^*\varepsilon)_i = \langle X^*\varepsilon, e_i \rangle = \langle \varepsilon, X e_i \rangle = \langle \varepsilon, x_i \rangle$$

where $(e_i)_{1 \le i \le d}$ the canonical basis of $\mathbb{R}^d$ and x_i is the i-th column of X.

$$\|X^*\varepsilon\|_\infty = \max_{1 \le i \le d} |\langle \varepsilon, x_i \rangle|$$

$$= \max_{1 \le i \le d} \langle \varepsilon, \pm x_i \rangle$$

$z_i = x_i \qquad z_{i+d} = -x_i \quad , \quad i = 1, \dots, d$

$$= \max_{1 \le i \le 2d} \underbrace{\langle \varepsilon, z_i \rangle}_{Z_i}$$

$$\|X^*\varepsilon\|_\infty = \max_{1 \le i \le 2d} Z_i$$

$$\underline{Z_i \in \mathcal{N}(0, \sigma^2 \|z_i\|^2)}$$

$$\mathbb{E}\left[\langle \varepsilon, z_i\rangle^2\right] = \mathbb{E}\left[\langle (\varepsilon\otimes\varepsilon) z_i, z_i\rangle\right]$$

$$= \langle \mathbb{E}[\varepsilon\otimes\varepsilon]\, z_i, z_i\rangle$$

$$= \sigma^2 \|z_i\|^2$$

Hypotheses: $\|X_i\| \le \beta\sqrt{n}$

This implies that $\|z_i\|^2\sigma^2 \le \beta^2\sigma^2 n$

$$\mathbb{P}\left(\|X^*\varepsilon\|_\infty > t\right) \le 2d\exp\left(\frac{-t^2}{2n\beta^2\sigma^2}\right).$$

$$t = \frac{\lambda n}{2}$$

$$\mathbb{P}\left(\|X^*\varepsilon\|_\infty > \frac{\lambda n}{2}\right) \le 2d\exp\left(-\frac{\lambda^2 n}{8\beta^2\sigma^2}\right) = \delta$$

$$\exp\left(-\frac{\lambda^2 n}{8\beta^2\sigma^2}\right) = \frac{\delta}{2d}$$

$$-\frac{\lambda^2 n}{8\beta^2\sigma^2} = \log\frac{\delta}{2d}$$

$$\frac{\lambda^2 n}{8\beta^2\sigma^2} = \log 2d + \log\frac{1}{\delta}$$

$$\lambda^2 = \frac{8\beta^2\sigma^2\left(\log 2d + \log\frac{1}{\delta}\right)}{n}$$

- $$\boxed{\lambda = 2\beta\sigma\sqrt{\frac{2\left(\log 2d + \log 1/\delta\right)}{n}}}$$

$$\mathbb{P}\left(\|X^*\varepsilon\|_\infty > \frac{\lambda n}{2}\right) \leq \delta$$

$$\mathbb{P}\left(\|X^*\varepsilon\|_\infty \leq \frac{\lambda n}{2}\right) \geq 1-\delta.$$

with $p. \geq 1-\delta$ we have $\|X^*\varepsilon\|_\infty \leq \frac{\lambda n}{2}$

and hence

$$\underbrace{\frac{1}{n}\|X\hat{\vartheta}_{\ell_1} - X\vartheta^*\|_2^2}_{MSE[\hat{\vartheta}_{\ell_1}]} \leq 2\lambda\|\vartheta^*\|_1 = 4B\sigma\sqrt{\frac{2(\log 2d + \log\frac{1}{\delta})}{n}}$$

Theorem Let $\delta \in]0,1[$. and set

$$\lambda = 2B\sigma\sqrt{\frac{2(\log 2d + \log\frac{1}{\delta})}{n}}.$$ Then

with probability $\geq 1-\delta$

$$MSE[\hat{\vartheta}_{\ell_1}] \leq \|\vartheta^*\|_1 B\sigma\sqrt{\frac{2(\log 2d + \log\frac{1}{\delta})}{n}}$$

Example $\delta = 0.05 \rightarrow 1-\delta \sim 95\%$

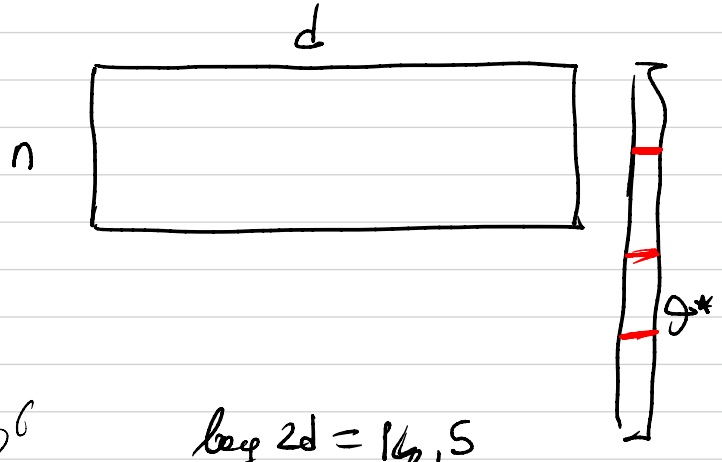
$$\log\frac{1}{\delta} = \log 20 \simeq 3.$$

$$MSE(\hat{\vartheta}_{LS}) \leq 4\sigma\beta \|\vartheta^*\|_1 \sqrt{\frac{2(\log 2d + 3)}{n}}$$

$d = 10^5$
$n = 10^2$
$\Rightarrow \log 2d \simeq 11{,}5$ and

$$\sqrt{\frac{2(\log 2d + 3)}{n}} \simeq 0{,}5.$$

$$MSE(\hat{\vartheta}_{LS}) \leq 2\sigma\beta \|\vartheta^*\|_1$$



$d = 10^6$ $\quad \log 2d = 14{,}5$

$n = 10^3$ $\quad \sqrt{\frac{2(\log 2d + 3)}{n}} \simeq 0{,}185$ $\square$

Lecture 7

Stochastic optimization algorithms

In this section we analyze stochastic versions of the algorithms previously presented. We will consider problems of type

$$\underset{x\in X}{\text{minimize}} \;\; f(x) + g(x), \tag{7.1}$$

where $f\colon X \to \mathbb{R}$ is a convex function and $g\colon X \to]-\infty, +\infty]$ is a proper convex and closed function, and depending on the hypotheses only a stochastic subgradient/gradient of f will be available. One of the main example for such situation is when f is given in the form of an expectation, that is,

$$f(x) = \mathsf{E}[\varphi(x, \zeta)], \tag{7.2}$$

which corresponds to the setting of *stochastic optimization*. In this case a stochastic subgradient/gradient of f is obtained through a subgradient/gradient of $\varphi(x, \xi)$. Finally, in general we will assume that the proximity operator of g is given explicitly. However, in the last section we will consider a situation in which the proximity operator of g is actually given through a stochastic oracle.

We start by recalling few facts on conditional expectation.

Fact 7.0.1. The following hold.

(i) Let ζ be a random variable with value in the measurable space $\mathcal{Z}$. Then the operator $\mathsf{E}[\cdot \,|\, \zeta]\colon L^1 \to L^1$ is linear and monotone increasing.

(ii) Let ξ be a real-valued summable random variable and ζ be a random variable with value in a measurable space $\mathcal{Z}$. Then, $\mathsf{E}[\mathsf{E}[\xi, \,|\, \zeta]] = \mathsf{E}[\xi]$.

(iii) Let ζ be a random variable with value in the measurable space $\mathcal{Z}$ and let $\varphi\colon \mathcal{Z} \to \mathbb{R}$ be a measurable real function such that $\mathsf{E}[|\varphi(\zeta)|] < +\infty$. Then $\mathsf{E}[\varphi(\zeta) \,|\, \zeta] = \varphi(\zeta)$.

(iv) Let X be a separable Hilbert space and let ζ_1 and ζ_2 be two X-valued random vectors such that $\mathsf{E}[|\langle \zeta_1, \zeta_2 \rangle|] < +\infty$ and $\mathsf{E}[\|\zeta_2\|] < +\infty$. Then $\mathsf{E}[\langle \zeta_1, \zeta_2 \rangle \,|\, \zeta_1] = \langle \zeta_1, \mathsf{E}[\zeta_2 \,|\, \zeta_1] \rangle$.

(v) Let ζ_1 and ζ_2 be two independent random variables with values in the measurable spaces $\mathcal{Z}_1$ and $\mathcal{Z}_2$ respectively. Let $\varphi\colon \mathcal{Z}_1 \times \mathcal{Z}_2 \to \mathbb{R}$ be measurable and such that $\mathsf{E}[|\varphi(\zeta_1, \zeta_2)|] < +\infty$. Then $\mathsf{E}[\varphi(\zeta_1, \zeta_2) \,|\, \zeta_1] = \psi(\zeta_1)$, where, for every $z_1 \in \mathcal{Z}_1$, $\psi(z_1) = \mathsf{E}[\varphi(z_1, \zeta_2)]$.

Salzo Saverio (saverio.salzo@iit.it) Istituto Italiano di Tecnologia. Via E. Melen, 16152, Genoa, Italy.

7.1 Stochastic subgradient method

We now consider a stochastic version of Algorithm 3.4.5. Here we take g in (7.1) as an indicator function of a closed convex set. Thus, we assume that $C \subset X$ is a nonempty closed and convex set and $f\colon X \to \mathbb{R}$ is a convex function and we want to solve the following problem

$$\operatorname*{minimize}_{x \in C} \ f(x), \tag{7.3}$$

where the projection onto C can be computed explicitly but, only a stochastic subgradient of f is available. The algorithm is detailed below.

Algorithm 7.1.1. Let $x_0 \in X$ and $(\gamma_k)_{k\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}_{++}$. Then,

$$\begin{array}{l} \text{for } k = 0, 1, \dots \\ \left\lfloor \begin{array}{l} \hat{u}_k \text{ is a summable } X\text{-valued random vector s.t. } \mathsf{E}[\hat{u}_k \,|\, x_k] \in \partial f(x_k), \\ x_{k+1} = P_C(x_k - \gamma_k \hat{u}_k). \end{array} \right. \end{array} \tag{7.4}$$

and

$$f_k = \min_{0 \le i \le k} \mathsf{E}[f(x_i)], \qquad \bar{x}_k = \left(\sum_{i=0}^{k} \gamma_i \right)^{-1} \sum_{i=0}^{k} \gamma_i x_i. \tag{7.5}$$

Remark 7.1.2. In addition to the sequence x_k, Algorithm 7.1.1 requires keeping track of the sequences $\Gamma_k := \sum_{i=0}^{k} \gamma_i$ and $\bar{x}_k$, which can updated recursively, as $\Gamma_{k+1} = \Gamma_k + \gamma_k$ and $\bar{x}_{k+1} = \Gamma_{k+1}^{-1}(\Gamma_k \bar{x}_k + \gamma_{k+1} x_{k+1})$.

The following theorem gives the main convergence results about the algorithm.

Theorem 7.1.3. *Let $C \subset X$ be a nonempty closed convex set and let $f\colon X \to \mathbb{R}$ be convex. Let $(x_k)_{k\in\mathbb{N}}$, $(f_k)_{k\in\mathbb{N}}$, and $(\bar{x}_k)_{k\in\mathbb{N}}$ be the sequences generated by Algorithm 7.1.1. We make the following additional assumption*

A1 *There exists $B \ge 0$, such that, for every $k \in \mathbb{N}$, $\mathsf{E}[\|\hat{u}_k\|^2] \le B^2$.*

Then, for every $k \in \mathbb{N}$, x_k is square summable in norm and $f(x_k)$ is summable and the following statements hold.

(i) *Suppose that $\gamma_k \to 0$ and $\sum_{k\in\mathbb{N}} \gamma_k = +\infty$. Then $\liminf_k \mathsf{E}[f(x_k)] = \lim_k f_k = \inf_C f$.*

(ii) *Let $x \in C$ and let $m, k \in \mathbb{N}$ with $m \le k$. Then*

$$\sum_{j=m}^{k} \frac{\gamma_j}{\sum_{i=m}^{k} \gamma_i} \mathsf{E}[f(x_j)] - f(x) \le \frac{\mathsf{E}[\|x_m - x\|^2]}{2} \frac{1}{\sum_{i=m}^{k} \gamma_i} + \frac{B^2}{2} \frac{\sum_{j=m}^{k} \gamma_i^2}{\sum_{i=m}^{k} \gamma_i}. \tag{7.6}$$

(iii) *Suppose that $\sum_{k\in\mathbb{N}} \gamma_k = +\infty$ and $\sum_{i=0}^{k} \gamma_i^2 / \sum_{i=0}^{k} \gamma_i \to 0$. Then $f_k \to \inf_C f$ and $\mathsf{E}[f(\bar{x}_k)] \to \inf_C f$. Moreover, if $\operatorname{argmin}_C f \neq \varnothing$, the right hand side of (7.6), with $m = 0$ and $x \in \operatorname{argmin}_C f$, yields a rate of convergence for both $f_k - \min_C f$ and $\mathsf{E}[f(\bar{x}_k)] - \min_C f$.*

Proof. Let $k \in \mathbb{N}$ and $x \in C$ and set $u_k = \mathsf{E}[\hat{u}_k \,|\, x_k]$. First of all, note that assumption A1 actually implies that $\|\hat{u}_k\|$ is square summable and hence summable. Then we prove the following inequality

$$2\gamma_k \langle x_k - x, \hat{u}_k \rangle \leq \|x_k - x\|^2 - \|x_{k+1} - x\|^2 + \gamma_k^2 \|\hat{u}_k\|^2 . \tag{7.7}$$

Indeed setting $y_k = x_k - \gamma_k \hat{u}_k$ and using the relation $2\langle a, b \rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2$, we have

$$\begin{aligned} 2\gamma_k \langle x_k - x, \hat{u}_k \rangle &= 2\langle x_k - x, x_k - y_k \rangle \\ &= \|x_k - x\|^2 + \|x_k - y_k\|^2 - \|y_k - x\|^2 . \end{aligned} \tag{7.8}$$

Now, since P_C is nonexpansive, we have $\|x_{k+1} - x\| = \|P_C(y_k) - P_C(x)\| \leq \|y_k - x\|$ and hence (7.7) follows.

We prove by induction that $\|x_k - x\|$ is square summable for every $k \in \mathbb{N}$. The statement is true for $k = 0$. Suppose that $\|x_k - x\|$ is square summable for some $k \in \mathbb{N}$. Then it follows from (7.7) that

$$\|x_{k+1} - x\|^2 \leq \|x_k - x\|^2 + 2\gamma_k \|x_k - x\| \, \|\hat{u}_k\| + \gamma_k^2 \|\hat{u}_k\|^2 .$$

The right-hand side is summable, and hence $\|x_{k+1} - x\|$ is square summable. So, all the terms in (7.7) are summable. Therefore, taking the conditional expectation given x_k of both terms of inequality (7.7) and using the fact that $u_k = \mathsf{E}[\hat{u}_k \,|\, x_k] \in \partial f(x_k)$ and the properties in Fact 7.0.1, we have almost surely

$$\begin{aligned} 2\gamma_k (f(x_k) - f(x)) &\leq 2\gamma_k \langle x_k - x, \mathsf{E}[\hat{u}_k \,|\, x_k] \rangle \\ &\leq \|x_k - x\|^2 - \mathsf{E}[\|x_{k+1} - x\|^2 \,|\, x_k] + \gamma_k^2 \mathsf{E}[\|\hat{u}_k\|^2 \,|\, x_k]. \end{aligned} \tag{7.9}$$

Now, being f subdifferentiable, there exists $(a, \beta) \in H \times \mathbb{R}$, $a \neq 0$, such that $\langle \cdot, a \rangle + \beta \leq f$, hence $\langle x_k, a \rangle + \beta \leq f(x_k)$. Therefore, we have $(f(x_k))_- \leq \|x_k\| \, \|a\| + |\beta|$, which together with (7.9) yields the summability of $f(x_k)$. Taking the expectation in (7.9) and recalling that $\mathsf{E}[\|\hat{u}_k\|^2] \leq B^2$, we get

$$2\gamma_k (\mathsf{E}[f(x_k)] - f(x)) \leq \mathsf{E}[\|x_k - x\|^2] - \mathsf{E}[\|x_{k+1} - x\|^2] + \gamma_k^2 B^2. \tag{7.10}$$

(i): Since $(f_k)_{k \in \mathbb{N}}$ is decreasing, we have $\inf_C f \leq \lim_k f_k = \inf_k f_k = \inf_k \mathsf{E}[f(x_k)] \leq \liminf_k \mathsf{E}[f(x_k)]$. Therefore it is sufficient to prove that $\liminf_k \mathsf{E}[f(x_k)] \leq \inf_C f$. Suppose that $x \in C$ is such that $f(x) < \liminf_k \mathsf{E}[f(x_k)] = \sup_n \inf_{k \geq n} \mathsf{E}[f(x_k)]$. Then there exists $n \in \mathbb{N}$ such that $f(x) < \inf_{k \geq n} \mathsf{E}[f(x_k)]$. Set $\rho = \inf_{k \geq n} \mathsf{E}[f(x_k)] - f(x) > 0$. Then, (7.10) yields

$$(\forall\, k \geq n) \quad \gamma_k \rho \leq \mathsf{E}[\|x_k - x\|^2] - \mathsf{E}[\|x_{k+1} - x\|^2] - \gamma_k \big(\rho - \gamma_k B^2\big).$$

Now, since $\gamma_k \to 0$, there exists $m \in \mathbb{N}$ such that for every integer $k \geq m$ we have $\rho - \gamma_k B^2 \geq 0$ and hence, setting $\nu := \max\{n, m\}$ we have

$$\rho \sum_{k \geq \nu} \gamma_k \leq \mathsf{E}[\|x_\nu - x\|^2] < +\infty.$$

This contradicts the assumption $\sum_{k\in\mathbb{N}} \gamma_k = +\infty$. Therefore, we showed that there is no $x \in C$ such that $f(x) < \liminf_k \mathsf{E}[f(x_k)]$, that is, $\liminf_k \mathsf{E}[f(x_k)] \leq \inf_C f$.

(ii): It follows from (7.10) that

$$(\forall\, i \in \mathbb{N}) \quad \gamma_i(\mathsf{E}[f(x_i)] - f(x)) \leq \frac{1}{2}\big(\mathsf{E}[\|x_i - x\|^2] - \mathsf{E}[\|x_{i+1} - x\|^2\,\big) + \frac{B^2}{2}\gamma_i^2. \tag{7.11}$$

So, summing form m to k, we have

$$\sum_{i=m}^{k} \gamma_i(\mathsf{E}[f(x_i)] - f(x)) \leq \frac{1}{2}\mathsf{E}[\|x_m - x\|^2] + \frac{B^2}{2}\sum_{i=m}^{k} \gamma_i^2.$$

Dividing the above inequality by $\sum_{i=m}^{k} \gamma_i$ yields (7.6).

(iii): We first note that, since f is convex and $\bar{x}_k$ is a convex combination of the x_i's, with coefficients $\eta_i = \gamma_i / \sum_{j=0}^{k} \gamma_j$, with $0 \leq i \leq k$, we have $\mathsf{E}[f(\bar{x}_k)] \leq \sum_{i=0}^{k} \eta_i \mathsf{E}[f(x_i)]$. Moreover, $f_k = \sum_{i=0}^{k} \eta_i f_k \leq \sum_{i=0}^{k} \eta_i \mathsf{E}[f(x_i)]$. Therefore,

$$(\forall\, k \in \mathbb{N}) \quad h_k := \max\{f_k, \mathsf{E}[f(\bar{x}_k)]\} \leq \Big(\sum_{i=0}^{k} \gamma_i\Big)^{-1} \sum_{i=0}^{k} \gamma_i \mathsf{E}[f(x_i)]. \tag{7.12}$$

Let $x \in C$. Then it follows from (7.6) and (7.12) that $\limsup_k h_k \leq f(x)$. Since x is arbitrary in C, we have $\limsup_k h_k \leq \inf_C f$. Moreover, clearly we have $\inf_C f \leq \liminf_k h_k$. Therefore, $h_k \to \inf_C f$. Since $\inf_C f \leq f_k \leq h_k$ and $\inf_C f \leq \mathsf{E}[f(\bar{x}_k)] \leq h_k$, the statement follows. □

Remark 7.1.4. In addition to the sequence x_k, Algorithm 7.1.1 requires keeping track of the sequences Γ_k and $\bar{x}_k$, which can updated recursively, as $\Gamma_{k+1} = \Gamma_k + \gamma_k$ and $\bar{x}_{k+1} = \Gamma_{k+1}^{-1}(\Gamma_k \bar{x}_k + \gamma_{k+1} x_{k+1})$.

Corollary 7.1.5. Under the same assumptions of Theorem 7.1.3, the following hold.

(i) Suppose that $\operatorname{argmin}_C f \neq \varnothing$ and let $D \geq \operatorname{dist}(x_0, \operatorname{argmin}_C f)$ and $k \in \mathbb{N}$. Then,

$$\max\{f_k, \mathsf{E}[f(\bar{x}_k)]\} - \min_C f \leq \frac{D^2}{2} \frac{1}{\sum_{i=0}^k \gamma_i} + \frac{B^2}{2} \frac{\sum_{j=0}^k \gamma_i^2}{\sum_{i=0}^k \gamma_i}. \tag{7.13}$$

Moreover, the right hand side of (7.13) is minimized when, for every $i = 0, \ldots, k$, $\gamma_i = D/(B\sqrt{k+1})$ and in that case we have

$$\max\left\{f_k, \mathsf{E}[f(\bar{x}_k)]\right\} - \min_C f \leq \frac{BD}{\sqrt{k+1}}.$$

(ii) Let, for every $k \in \mathbb{N}$, $\gamma_k = \bar{\gamma}/(k+1)$. Then, $f_k \to \inf_C f$ and $\mathsf{E}[f(\bar{x}_k)] \to \inf_C f$. Moreover, if $\operatorname{argmin}_C f \neq \varnothing$, we have, for every $k \in \mathbb{N}$,

$$\max\left\{f_k, \mathsf{E}[f(\bar{x}_k)]\right\} - \min_C f \leq \left(\frac{\operatorname{dist}(x_0, \operatorname{argmin}_C f)^2}{2\bar{\gamma}} + \frac{\pi \bar{\gamma} B^2}{12}\right) \frac{1}{\log(k+1)}. \tag{7.14}$$

(iii) Let, for every $k \in \mathbb{N}$, $\gamma_k = \bar{\gamma}/\sqrt{k+1}$. Then, $f_k \to \inf_C f$ and $\mathsf{E}[f(\bar{x}_k)] \to \inf_C f$. Moreover, if $\operatorname{argmin}_C f \neq \varnothing$, for every integer $k \geq 2$, we have

$$\max\left\{f_k, \mathsf{E}[f(\bar{x}_k)]\right\} - \min_C f \leq \frac{\operatorname{dist}(x_0, \operatorname{argmin}_C f)^2}{2\bar{\gamma}} \frac{1}{\sqrt{k+1}} + \bar{\gamma} B^2 \frac{\log(k+1)}{\sqrt{k+1}}. \tag{7.15}$$

(iv) Let, for every $k \in \mathbb{N}$, $\gamma_k = \bar{\gamma}/\sqrt{k+1}$ and suppose that C is bounded with diameter $\bar{D} > 0$ and that $\operatorname{argmin}_C f \neq \varnothing$. Set, for every $k \in \mathbb{N}$, $\tilde{f}_k = \min_{\lfloor k/2 \rfloor \leq i \leq k} f(x_i)$ and $\tilde{x}_k = \left(\sum_{i=\lfloor k/2 \rfloor}^k \gamma_i\right)^{-1} \sum_{i=\lfloor k/2 \rfloor}^k \gamma_i x_i$. Then, for every integer $k \geq 2$,

$$\max\left\{\tilde{f}_k, \mathsf{E}[f(\tilde{x}_k)]\right\} - \min_C f \leq \left(\frac{3\bar{D}^2}{2\bar{\gamma}} + \frac{5\bar{\gamma} B^2}{2}\right) \frac{1}{\sqrt{k+1}}. \tag{7.16}$$

Proof. (i): Equation (7.13) follows from (7.12) and by minimizing the right hand side of (7.6), with $m = 0$, w.r.t. $x \in \operatorname{argmin}_C f$. Now, it follows from Lemma 3.4.10 that the minimum of the right-hand side of (7.13) is $BD/\sqrt{k+1}$ and is achieved at $(\gamma_i)_{0 \leq i \leq k} \equiv D/(B\sqrt{k+1})$. Note that is this case $\bar{x}_k = (k+1)^{-1} \sum_{i=0}^k x_i$.

(ii): We derive from Lemma 3.4.9(i), with $m = 1$, that $\sum_{i=0}^k \gamma_i = \bar{\gamma} \sum_{i=1}^{k+1} (1/i) \geq \bar{\gamma} \log(k+1)$. Moreover, we have $\sum_{i=0}^k \gamma_i^2 = \bar{\gamma}^2 \sum_{i=1}^{k+1} 1/i^2 \leq \bar{\gamma}^2 \pi/6$. So, the first part follows from Theorem 7.1.3(iii), while the inequality in (7.14) follows from (7.13) with $D = \operatorname{dist}(x_0, \operatorname{argmin}_C f)$.

(iii): Lemma 3.4.9(ii), with $m = 1$, yields $\sum_{i=1}^k 1/\sqrt{i} \geq 2(\sqrt{k}-1) + (1/2)(1 + 1/\sqrt{k}) \geq 2\sqrt{k} - 3/2$. Moreover, $2\sqrt{k} - 3/2 \geq \sqrt{k}$ for $k \geq 3$ and clearly for $k \leq 2$, $\sum_{i=1}^k 1/\sqrt{i} \geq \sqrt{k}$. Therefore, for every $k \in \mathbb{N}$, $\sum_{i=0}^k \gamma_i = \bar{\gamma} \sum_{i=1}^{k+1} 1/\sqrt{i} \geq \bar{\gamma}\sqrt{k+1}$. Moreover, by Lemma 3.4.9(i), we have $\sum_{i=1}^k 1/i = 1 + \sum_{i=2}^k 1/i \leq 1 + \log k \leq 2 \log k$, for $k \geq 3$. Therefore, for every $k \in \mathbb{N}$, $k \geq 2$, we have $\sum_{i=0}^k \gamma_i^2 = \bar{\gamma}^2 \sum_{i=1}^{k+1} 1/i \leq 2\bar{\gamma}^2 \log(k+1)$. Again, the first part follows from Theorem 7.1.3(iii), while (7.15) follows from (7.13)

with $D = \operatorname{dist}(x_0, \operatorname{argmin}_C f)$.

(iv): Let $k \in \mathbb{N}$, $k \geq 2$. It follows from Lemma 3.4.9(i) that

$$\sum_{i=\lfloor k/2 \rfloor}^{k} \gamma_i^2 = \bar{\gamma}^2 \sum_{i=\lfloor k/2 \rfloor+1}^{k+1} \frac{1}{i} \leq \bar{\gamma}^2 \log\Big(\frac{k+1}{\lfloor k/2 \rfloor}\Big) \leq \bar{\gamma}^2 \log 4 \leq \bar{\gamma}^2 \frac{5}{3}.$$

Moreover, Lemma 3.4.9(ii) yields

$$\sum_{i=\lfloor k/2 \rfloor}^{k} \gamma_i = \bar{\gamma} \sum_{i=\lfloor k/2 \rfloor+1}^{k+1} \frac{1}{\sqrt{i}}$$

$$\geq 2\bar{\gamma}(\sqrt{k+1} - \sqrt{\lfloor k/2 \rfloor + 1}) \geq 2\bar{\gamma}\sqrt{k+1}\Big(1 - \sqrt{\frac{\lfloor k/2 \rfloor + 1}{k+1}}\Big).$$

Now, since $(\lfloor k/2 \rfloor + 1)/(k+1) \leq 2/3$, we have

$$\sum_{i=\lfloor k/2 \rfloor}^{k} \gamma_i \geq 2\bar{\gamma}\Big(1 - \sqrt{\frac{2}{3}}\Big)\sqrt{k+1} \geq \frac{\bar{\gamma}}{3}\sqrt{k+1}$$

The statement follows from Theorem 7.1.3(ii), with $m = \lfloor k/2 \rfloor$ and $x \in \operatorname{argmin}_C f$, taking into account that, as in (7.12), $\max\{\tilde{f}_k, f(\tilde{x}_k)\} \leq \big(\sum_{i=\lfloor k/2 \rfloor}^{k} \gamma_i\big)^{-1} \sum_{i=\lfloor k/2 \rfloor}^{k} \gamma_i f(x_i)$. □

Example 7.1.6. A case in which the above stochastic algorithm arises is in the incremental subgradient method. We aim at solving

$$\min_{x \in C} f(x) := \frac{1}{m} \sum_{j=1}^{m} f_j(x), \tag{7.17}$$

where every $f_j \colon X \to \mathbb{R}$ is convex and Lipschitz continuous with constant L_j. The projected incremental subgradient method is as follows. Let, for every j, $\tilde{\nabla} f_j \colon X \to X$ be a selection of ∂f_j. Let $x_0 \in X$ and define, for every $k \in \mathbb{N}$, x_{k+1} as follows.

$$\begin{cases} \text{chose an index } j_k \in \{1, \dots, m\} \text{ at random} \\ x_{k+1} = P_C(x_k - \gamma_k \underbrace{\tilde{\nabla} f_{j_k}(x_k)}_{\hat{u}_k}). \end{cases} \tag{7.18}$$

Since $\partial f = (1/m)\sum_{j=1}^{m} \partial f_j$, we have that $(1/m)\sum_{j=1}^{m} \tilde{\nabla} f_j(x) \in \partial f(x)$. Let $k \in \mathbb{N}$. Then, x_k is a random variable, depending on $j_0, \dots, j_{k-1}$. Hence, $\hat{u}_k := \tilde{\nabla} f_{j_k}(x_k)$ is a random variable, where x_k and j_k are independent random variables, and Fact 7.0.1 yields

$$u_k := \mathsf{E}\big[\tilde{\nabla} f_{j_k}(x_k) \,|\, x_k\big] = \frac{1}{m} \sum_{j=1}^{m} \tilde{\nabla} f_j(x_k) \in \partial f(x_k) \tag{7.19}$$

and

$$\mathsf{E}[\|\tilde{\nabla} f_{j_k}(x_k) \,|\, x_k\|^2] = \frac{1}{m} \sum_{j=1}^{m} \|\tilde{\nabla} f_j(x_k)\|^2 \leq \frac{1}{m} \sum_{j=1}^{m} L_j^2, \tag{7.20}$$

and hence $\mathsf{E}[\|\tilde{\nabla} f_{j_k}(x_k)\|^2] \leq (1/m)\sum_{j=1}^{m} L_j^2$. In the end assumptions of Theorem 7.1.3 are satisfied with $B^2 = (1/m)\sum_{j=1}^{m} L_j^2$.

Example 7.1.7 (Stochastic optimization)**.** We generalize the previous example. We consider the following optimization problem

$$\min_{x\in C} f(x), \qquad f(x) = \mathsf{E}[F(x,\zeta)], \tag{7.21}$$

where $f\colon X \to \mathbb{R}$, ζ is a random variable with values in a measurable space $\mathcal{Z}$ with distribution μ and $F\colon X \times \mathcal{Z} \to \mathbb{R}$ is such that

(SO_1) $\forall\, z \in \mathcal{Z}, F(\cdot, z)$ is convex and $L(z)$-Lipschitz continuous and $\int_{\mathcal{Z}} L(z)^2 d\mu < +\infty$.

(SO_2) $F(0,\cdot) \in L^1(\mathcal{Z},\mu)$.

Note that the above assumptions ensure that, for every $x \in X$, $F(x,\cdot) \in L^1(\mathcal{Z},\mu)$. Indeed, for every $z \in \mathcal{Z}$, $|F(x,z)| \le |F(x,z) - F(0,z)| + |F(0,z)| \le L(z)\,\|x\| + |F(0,z)|$. Hence $F(x,z) \in L^1(\mathcal{Z},\mu)$, since $L(z)$ and $F(0,z)$ are so. We let $\partial F\colon X \times \mathcal{Z} \to 2^X$ be such that $\partial F(x,z) = \partial F(\cdot,z)(x)$ and we make the following additional assumptions

(SO_3) there exists a measurable $\tilde{\nabla} F\colon X \times \mathcal{Z} \to X$, such that, for every $x \in X$ and for μ-a.e. $z \in \mathcal{Z}$, $\tilde{\nabla} F(x,z) \in \partial F(x,z)$.

(SO_4) $(\zeta_k)_{k\in\mathbb{N}}$ is a sequence of independent copies of ζ.

Then we consider the following algorithm

$$\begin{array}{l} x_0 \in X \\ \left\lfloor \begin{array}{l} \text{for } k = 0, 1, \dots \\ \left\lfloor \begin{array}{l} x_{k+1} = P_C\big(x_k - \gamma_k \underbrace{\tilde{\nabla} F(x_k, \zeta_k)}_{\hat{u_k}}\big). \end{array} \right. \end{array} \right. \end{array} \tag{7.22}$$

We have, for every $x_1, x_2 \in X$,

$$|f(x_1) - f(x_2)| \le \int_{\mathcal{Z}} |F(x_1,z) - F(x_2,z)| d\mu(z) \le \|x_1 - x_2\| \int_{\mathcal{Z}} L(z) d\mu(z). \tag{7.23}$$

Note that $\int_{\mathcal{Z}} L(z) d\mu(z) < +\infty$, since $L(z)$ is square μ-summable and $L^2(\mathcal{Z},\mu) \subset L^1(\mathcal{Z},\mu)$. So, f is Lipschitz continuous with constant $\int_{\mathcal{Z}} L(z) d\mu(z) \le \big(\int_{\mathcal{Z}} L(z)^2 d\mu(z)\big)^{1/2}$. Moreover, assumption ($\mathrm{SO}_3$) implies that

$$\text{for every } x, y \in X \text{ and for } \mu\text{-a.e. } z \in \mathcal{Z} \quad F(y,z) \ge F(x,z) + \langle y - x, \tilde{\nabla} F(x,z)\rangle. \tag{7.24}$$

Note that all terms of the above inequality are μ-summable, in particular, since $\|\tilde{\nabla} F(x,z)\| \le L(z)$ and $L(z)$ is μ-summable, $\tilde{\nabla} F(\cdot,z)$ is μ-summable. Hence, integrating (7.24) w.r.t. μ we get

$$(\forall\, x, y \in X) \qquad f(y) \ge f(x) + \langle y - x, \int_{\mathcal{Z}} \tilde{\nabla} F(x,z) d\mu(z)\rangle.$$

Therefore, for every $x \in X$, $\mathsf{E}\tilde{\nabla} F(x,\zeta) \in \partial f(x)$. Now, let $k \in \mathbb{N}$, $k \ge 1$. Then, it follows from (8.26) that

$$x_k = x_k(\zeta_0, \dots, \zeta_{k-1}), \tag{7.25}$$

hence x_k and ζ_k are independent random variables. Therefore, Fact 7.0.1(v) yields that $u_k := \mathsf{E}[\tilde{\nabla} F(x_k, \zeta_k) \,|\, x_k] = \int_{\mathcal{Z}} \tilde{\nabla} F(x_k, z) d\mu(z) \in \partial f(x_k)$ and

$$\mathsf{E}\big[\|\tilde{\nabla} F(x_k, \zeta_k)\|^2 \,|\, x_k\big] = \int_{\mathcal{Z}} \|\tilde{\nabla} F(x_k, z)\|^2 d\mu(z) \leq \int_{\mathcal{Z}} L(z)^2 d\mu(z) < +\infty, \tag{7.26}$$

and hence $\mathsf{E}[\|\tilde{\nabla} F(x_k, \zeta_k)\|^2] \leq \int_{\mathcal{Z}} L(z)^2 d\mu(z)$. In the end Theorem 7.1.3 applies with $B^2 = \int_{\mathcal{Z}} L(z)^2 d\mu(z)$, so that the stochastic algorithm (8.26) provides a solution to problem (7.21).

7.2 Randomized block coordinate gradient descent

In this section we address the following problem

$$\operatorname*{minimize}_{\boldsymbol{x} \in X} F(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x}), \qquad g(\boldsymbol{x}) = \sum_{i=1}^{m} g_i(x_i), \tag{7.27}$$

where X is the direct sum of m Euclidean spaces $(X_i)_{1 \leq i \leq m}$, i.e.,

$$X = \bigoplus_{i=1}^{m} X_i \quad \text{and} \quad (\forall\, \boldsymbol{x} = (x_1, \cdots, x_m), \boldsymbol{y} = (y_1, \cdots, y_m) \in X) \quad \langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^{m} \langle x_i, y_i \rangle$$

and the following assumptions hold

A1 $f \colon X \mapsto \mathbb{R}$ is convex and differentiable with Lipschitz continuous gradient.

A2 $(\forall i \in [m] := \{1, \ldots, m\})$, $g_i \in \Gamma_0(X_i)$.

We study the following algorithm.

Algorithm 7.2.1 (The randomized block-coordinate proximal-gradient method)**.** Let $\boldsymbol{x}^0 = (x_1^0, \ldots, x_m^0) \in X$ and $(\gamma_i)_{1 \leq i \leq m} \in \mathbb{R}^m_{++}$. Then,

$$\begin{array}{l} \text{for } k = 0, 1, \ldots \\ \quad \text{for } i = 0, 1, \ldots, m \\ \quad\quad x_i^{k+1} = \begin{cases} \operatorname{prox}_{\gamma_k g_{i_k}}\big(x_{i_k}^k - \gamma_{i_k} \nabla_{i_k} f(\boldsymbol{x}^k)\big) & \text{if } i = i_k \\ x_i^k & \text{if } i \neq i_k \end{cases} \end{array} \tag{7.28}$$

where $(i_k)_{k \in \mathbb{N}}$ are independent random variables uniformly distributed on $\{1, \ldots, m\}$.

In the following we denote by $J_i \colon X_i \to X$ the canonical embedding of X_i into X, that is, $J_i(x_i) = (0, \ldots, x_i, \ldots, 0)$, where x_i occurs in the i-th position. Thus, the algorithm can be equivalently written as

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + J_{i_k}\big(\operatorname{prox}_{\gamma_{i_k} g_{i_k}}(x_{i_k}^k - \gamma_{i_k} \nabla_{i_k} f(\boldsymbol{x}^k)) - x_{i_k}^k\big). \tag{7.29}$$

Moreover, we set

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\Gamma^{-1}} = \sum_{i=1}^{m} \frac{1}{\gamma_i} \langle x_i, y_i \rangle \tag{7.30}$$

The dual proximal gradient algorithm (8.27) writes down as follows

$$\begin{array}{l}\text{for } k=0,1,\dots\\ \left\lfloor\begin{array}{l} x^{(k)} = y - D^*\boldsymbol{u}^{(k)}\\ \boldsymbol{u}^{(k+1)} = P_{B_\lambda(0)^{n\times m}}(\boldsymbol{u}^{(k)} + \gamma D x^{(k)}),\end{array}\right.\end{array} \tag{9.10}$$

where $\gamma < 2/\left\|D\right\|^2 = 1/4$. Note also that the projection onto $B_\lambda(0)^{m\times n}$ is separable too and can be computed as

$$P_{B_\lambda(0)^{m\times n}}(\boldsymbol{u}) = \big(P_{B_\lambda(0)}(\boldsymbol{u}_{i,j})\big)_{\substack{1\le i\le m,\\ 1\le j\le n}} \qquad P_{B_\lambda(0)}(\boldsymbol{u}_{i,j}) = \begin{cases} \boldsymbol{u}_{i,j} & \text{if } \left\|\boldsymbol{u}_{i,j}\right\|_2 \le \lambda\\ \dfrac{\boldsymbol{u}_{i,j}}{\left\|\boldsymbol{u}_{i,j}\right\|_2} & \text{if } \left\|\boldsymbol{u}_{i,j}\right\|_2 > \lambda.\end{cases}$$

Then it follows from the theory given in Section **??** that the sequence $(x_k)_{k\in\mathbb{N}}$ converges to the minimizer of (9.8) as an $O(1/\sqrt{k})$.

We next specialize Algorithm 6.3.5 to problem (9.9). Let $\boldsymbol{u}_0 = \boldsymbol{v}_0 \in X$, $z_0 = y - D^*\boldsymbol{u}^{(0)}$, and $\gamma \in]0, 1/8[$. Define

$$\begin{array}{l}\text{for } k=0,1,\dots\\ \left\lfloor\begin{array}{l} x^{(k)} = y - D^*\boldsymbol{u}^{(k)}\\ \boldsymbol{u}^{(k+1)} = P_{B_\lambda(0)^{n\times m}}(\boldsymbol{v}^{(k)} + \gamma D z^{(k)}),\\ \boldsymbol{v}^{(k+1)} = \boldsymbol{u}^{(k+1)} + \beta_{k+1}(\boldsymbol{u}^{(k+1)} - \boldsymbol{u}^{(k)})\\ z^{(k+1)} = x^{(k+1)} + \beta_{k+1}(x^{(k+1)} - x^{(k)})\end{array}\right.\end{array} \tag{9.11}$$

With the choice of parameters as in Theorem 6.3.10, from the results in Section **??** we derive that the sequence $(x_k)_{k\in\mathbb{N}}$ converges to the minimizer of (9.8) as an $O(1/k)$.

Finally, we specialize the randomized proximal gradient Algorithm 7.2.1. Since ..., condition (ii) in Proposition 7.2.3 is satisfied with $L_{i,j} =$. Then, Algorithm 7.2.1 (assuming that each block is made of one $\mathbb{R}^2$ block only and (i_k, j_k) is uniformly distributed on $\{1,\dots,n\}\times\{1,\dots,m\}$) writes as

$$\begin{array}{l}\text{for } k=0,1,\dots\\ \left\lfloor\begin{array}{l} x^{(k)} = x^{k-1} + D^*(\boldsymbol{u}^{k-1} - \boldsymbol{u}^k)\\ \boldsymbol{u}^{(k+1)} = \boldsymbol{u}^{(k)} + J_{(i_k,j_k)}\big[P_{B_\lambda(0)}(\boldsymbol{u}^{(k)}_{i_k,j_k} + \gamma_{i_k,j_k}(Dx^{(k)})_{i_k,j_k}) - \boldsymbol{u}^{(k)}_{i_k,j_k}\big],\end{array}\right.\end{array} \tag{9.12}$$

where $\gamma_{i,j} < 2/\sqrt{17}$ and $J_{(i_k,j_k)}\colon \mathbb{R}^2 \to (\mathbb{R}^2)^{m\times n}$ is the canonical injection. Then, denoting by x_* the unique solution of (9.8), Theorem 7.2.11 and the results in Section **??** ensure that $\mathsf{E}[\|x^k - x_*\|^2] \le o(1/\sqrt{k})$.

9.3 Machine Learning

In statistical machine learning we are given two random variables ξ and η, with values in $\mathcal{X}$ and $\mathcal{Y}\subset\mathbb{R}$ respectively, with joint distribution μ. We let $\ell\colon \mathcal{X}\times\mathcal{Y}\times\mathbb{R}\to\mathbb{R}$ be a convex loss function and the goal is to find a function $\varphi\colon \mathcal{X}\to\mathcal{Y}$ in a given hypothesis function space which minimizes the averaged risk $R(\varphi) = \mathsf{E}[\ell(\xi,\eta,\varphi(\eta))]$ without knowing the distribution μ but based on some sequence $(\xi_k,\eta_k)_{k\in\mathbb{N}}$ of independent copies of (ξ,η).

In this problem, concerning the hypothesis function space one option is that of considering reproducing kernel Hilbert spaces (RKHS). They indeed are defined through kernel functions and are flexible enough to model even infinite dimensional function spaces. They are defined as follows. We let $\Lambda\colon \mathcal{X} \to H$ be a general map from the input space $\mathcal{X}$ to a separable Hilbert space H, endowed with a scalar product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$. Then the corresponding RKHS is defined as

$$\mathcal{H} = \{h \in \mathbb{R}^{\mathcal{X}} \,|\, \exists\, w \in H \text{ s.t. } h = \langle w, \Lambda(\cdot)\rangle\} \quad \|h\| = \inf\{\|w\| \,|\, \varphi = \langle w, \Lambda(\cdot)\rangle\}. \tag{9.13}$$

In this context the map Λ is called the *feature map* and the corresponding *kernel function* is defined as

$$K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \qquad K(x, x') = \langle \Lambda(x), \Lambda(x')\rangle. \tag{9.14}$$

In this way the above statistical learning problem becomes

$$\min_{w\in H} R(w) = \mathsf{E}[\ell(\xi, \eta, \langle w, \Lambda(\xi)\rangle)] = \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, y, \langle w, \Lambda(x)\rangle) d\mu(x, y), \tag{9.15}$$

which is supposed to be solved via some sequence $(\xi_k, \eta_k)_{k\in\mathbb{N}}$ of independent copies of (ξ, η).

In order to approach problem (9.15) we consider two strategies. The first one consists in considering the problem as an instance of a stochastic optimization problem as described in Example 7.1.7. The second one is to consider a regularized empirical version of (9.15) based on the available sample. In the following we describe these two approaches.

9.3.1 Statistical learning as stochastic optimization

We make the following assumptions.

(SL_1) For every $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\ell(x, y, \cdot)\colon \mathbb{R} \to \mathbb{R}$ is positive, convex and Lipschitz continuous with constant $\alpha > 0$ and $\mathsf{E}[\ell(\xi, \eta, 0)] < +\infty$.

(SL_2) The feature map Λ is measurable and $\mathsf{E}[\|\Lambda(\xi)\|^2] < +\infty$.

We show that problem (9.15) is an instance of Example 7.1.7. Indeed, we let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and, for every $w \in H$ and $z = (x, y) \in \mathcal{Z}$, $\varphi(w, z) = \ell(z, \langle w, \Lambda(x)\rangle)$. Then,

$$(\forall\, w_1, w_2 \in H)(\forall\, z = (x, y) \in \mathcal{X} \times \mathcal{Y}) \\ |\varphi(w_1, z) - \varphi(w_2, z)| \le \alpha |\langle w_1 - w_2, \Lambda(x)\rangle| \le \alpha \|\Lambda(x)\| \|w_1 - w_2\|.$$

Hence, items (SO_1)-(SO_2) in Example 7.1.7 hold with $L(z) = \alpha \|\Lambda(x)\|$. Moreover,

$$(\forall\, z \in \mathcal{Z})(\forall\, w \in H) \quad \partial\varphi(w, z) = \partial\ell(z, \langle w, \Lambda(x)\rangle)\Lambda(x), \tag{9.16}$$

where $\partial\varphi(w, z) = \partial\varphi(\cdot, z)(w)$. Now, let, for every $(z, t) \in \mathcal{Z} \times \mathbb{R}$, $\tilde{\ell}'(z, t)$ be a subgradient of $\ell(z, \cdot)$ at t and define

$$\tilde{\nabla}\varphi\colon H \times \mathcal{Z} \to H\colon (w, z) \mapsto \tilde{\ell}'(z, \langle w, \Lambda(x)\rangle)\Lambda(x) \in \partial\varphi(w, z).$$

Therefore, items (SO_3) and (SO_4) in Example 7.1.7 are satisfied and

$$\mathsf{E}[\tilde{\nabla}\varphi(w,\zeta)] = \int_{\mathcal{Z}} \tilde{\ell}'(x,y,\langle w,\Lambda(x)\rangle)\Lambda(x)d\mu(x,y) \in \partial R(w).$$

Then algorithm (8.26) becomes

$$w_{k+1} = w_k - \gamma_k \tilde{\ell}'(\xi_k,\eta_k,\langle w_k,\Lambda(\xi_k)\rangle)\Lambda(\xi_k). \tag{9.17}$$

If we define $h_k(x) = \langle w_k,\Lambda(x)\rangle$ and the kernel $K(x,x') = \langle\Lambda(x),\Lambda(x')\rangle$, then it follows from (9.17) that

$$h_{k+1}(x) = h_k(x) - \gamma_k \tilde{\ell}'(\xi_k,\eta_k,\varphi_k(\xi_k))K(x,\xi_k). \tag{9.18}$$

Moreover, set

$$\bar{w}_k = \big(\sum_{i=0}^{k}\gamma_i\big)^{-1}\sum_{i=0}^{k}\gamma_i w_i, \qquad \bar{h}_k(x) = \langle\bar{w}_k,\Lambda(x)\rangle = \big(\sum_{i=0}^{k}\gamma_i\big)^{-1}\sum_{i=0}^{k}\gamma_i h_i(x).$$

Then, the risk of $\bar{h}_k$ is $R(\bar{w}_k)$ and according to Theorem 7.1.3 we have that $R(\bar{w}_k) \to \inf_H R$, and if $S_* := \operatorname{argmin}_H R \neq \varnothing$, $D \geq \operatorname{dist}(x_0,S_*)$, and $\gamma_k = \bar{\gamma}/\sqrt{k+1}$, we have

$$(\forall\, k\in\mathbb{N}) \quad \mathsf{E}[R(\bar{w}_k)] - \min_H R \leq \frac{D^2}{2\bar{\gamma}}\frac{1}{\sqrt{k+1}} + \bar{\gamma}B^2\frac{\log(k+1)}{\sqrt{k+1}},$$

where $B^2 = \alpha^2\mathsf{E}[\|\Lambda(\xi)\|^2]$. Moreover, for every $k\in\mathbb{N}$, if $(\gamma_i)_{0\leq i\leq k} \equiv D/(B\sqrt{k+1})$, then

$$\mathsf{E}[R(\bar{w}_k)] - \min_H R \leq \frac{BD}{\sqrt{k+1}}. \tag{9.19}$$

Note that algorithm (9.18) is fully practicable, since it depends only on the kernel function K and on the data (ξ_k,η_k). In the following we provide a list of 1-Lipschitz continuous losses:

- the *hinge loss*: $\mathcal{Y} = \{-1,1\}$ and $\ell(x,y,t) = \max\{0,1-yt\}$;
- the *logistic loss for classification*: $\mathcal{Y} = \{-1,1\}$ and $\ell(x,y,t) = \log(1+e^{-yt})$;
- L^1*-loss*: $\mathcal{Y} = \mathbb{R}$ and $\ell(x,y,t) = |y-t|$;
- *logistic loss for regression*: $\mathcal{Y} = \mathbb{R}$ and $\ell(x,y,t) = -\log\dfrac{4e^{y-t}}{(1+e^{y-t})^2}$.
- ε*-insensitive loss*: $\mathcal{Y} = \mathbb{R}$ and $\ell(x,y,t) = \max\{0,|y-t|-\varepsilon\}$.

9.3.2 Regularized empirical risk minimization

Regularized empirical risk estimation solves the following optimization problem

$$\min_{w\in H}\frac{\lambda}{n}\sum_{i=1}^{n}\ell(y_i,\langle w,\Lambda(x_i)\rangle) + \frac{1}{2}\|w\|^2 =: \Phi(w), \tag{9.20}$$

Lecture 8

Duality Theory – part II

8.1 The Fenchel-Rockafellar duality

Let $A\colon X \to Y$ be a linear operator between Euclidean spaces, and let $f\colon X \to]-\infty, +\infty]$ and $g\colon Y \to]-\infty, +\infty]$ be proper, convex and closed functions. Consider the problem

$$\boxed{\min_{x\in X} f(x) + g(Ax) := \Phi(x).} \qquad (\mathscr{P})$$

Its *dual problem* is

$$\boxed{\min_{u\in Y} f^*(-A^*u) + g^*(u) := \Psi(u).} \qquad (\mathscr{D})$$

One can prove (see below) that

$$(\forall\, x \in X)(\forall u \in Y) \quad \Phi(x) \geq -\Psi(u), \tag{8.1}$$

hence

$$\inf_{x\in X} \Phi(x) \geq \sup_{u\in Y} -\Psi(u) = -\inf_{u\in Y} \Psi(u). \tag{8.2}$$

This means that the function Φ is (uniformly) above the function $-\Psi$ (which is concave). The difference between the infimum of Φ and the supremum of $-\Psi$, that is $\inf \Phi + \inf \Psi$, is called *the duality gap* and we say that *strong duality* holds if the duality gap is zero.[1] See figure 8.1.

The fundamental inequality below follows directly from the definition of Φ and Ψ, the Young-Fenchel inequality, and the fact that $\langle A^*u, x\rangle = \langle u, Ax\rangle$.

$$\Phi(x)+\Psi(u) = \big[\underbrace{f(x) + f^*(-A^*u) - \langle -A^*u, x\rangle}_{\geq 0}\big] + \big[\underbrace{g(Ax) + g^*(u) - \langle Ax, u\rangle}_{\geq 0}\big] \geq 0. \tag{8.3}$$

Thus, this proves (8.1).

Salzo Saverio (saverio.salzo@iit.it) Istituto Italiano di Tecnologia. Via E. Melen, 16152, Genoa, Italy.

[1] Note that if $\inf \Phi = -\infty$, it follows from (8.2) that $\inf \Phi = \sup(-\Psi) = -\inf \Psi = -\infty$. In this case $\Psi \equiv +\infty$ and $\inf \Phi + \inf \Psi = -\infty + \infty$ does not make sense. Anyway, since there is no gap between Φ and $-\Psi$, by convention, we set $\inf \Phi + \inf \Psi = 0$. The same situation occurs if $\inf \Psi = -\infty$.

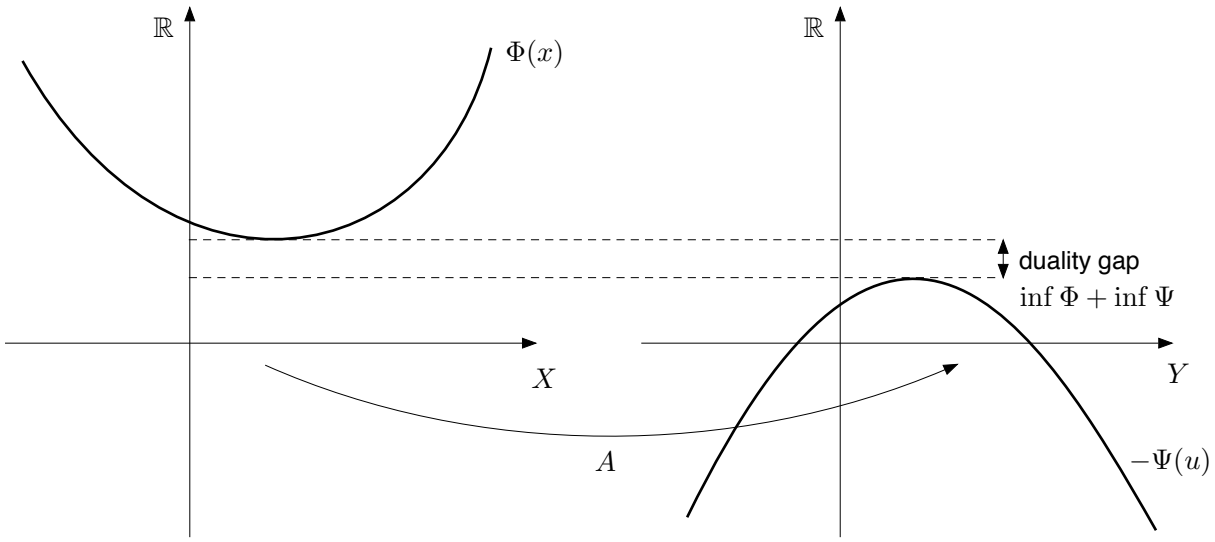


Figure 8.1: The meaning of duality.

Theorem 8.1.1. *Let $f \in \Gamma_0(X)$, $g \in \Gamma_0(Y)$, and $A\colon X \to Y$ be a linear operator. Let $\Phi\colon X \to]-\infty,+\infty]$ and $\Psi\colon Y \to]-\infty,+\infty]$ be defined as in $(\mathscr{P})$ and $(\mathscr{D})$ respectively. Let $\hat{x} \in X$ and $\hat{u} \in Y$. Let $S = \operatorname{argmin}\Phi$ and $S^* = \operatorname{argmin}\Psi$. Then the following statements are equivalent.*

(i) $\hat{x} \in S$, $\hat{u} \in S^*$, *and* $\inf_X \Phi + \inf_Y \Psi = 0$ (*duality gap is zero*);

(ii) $\Phi(\hat{x}) + \Psi(\hat{u}) = 0$ (*this implies that* $\hat{x} \in \operatorname{dom}\Phi$ *and* $\hat{u} \in \operatorname{dom}\Psi$);

(iii) $f(\hat{x}) + f^*(-A^*\hat{u}) = \langle -A^*\hat{u}, \hat{x}\rangle$ *and* $g(A\hat{x}) + g^*(\hat{u}) = \langle \hat{u}, A\hat{x}\rangle$

(iv) $\hat{x} \in \partial f^*(-A^*\hat{u})$ *and* $A\hat{x} \in \partial g^*(\hat{u})$

(v) $-A^*\hat{u} \in \partial f(\hat{x})$ *and* $\hat{u} \in \partial g(A\hat{x})$.

Proof. (i) $\Rightarrow$ (ii): Since Φ and Ψ never take value $-\infty$, we then have $-\infty < \Phi(\hat{x}) = \inf\Phi = \sup(-\Psi) = -\Psi(\hat{u}) < +\infty$.

(ii) $\Rightarrow$ (i): Clearly $\Phi(\hat{x}), \Psi(\hat{u}) \in \mathbb{R}$. Moreover, by (8.1)

$$(\forall, x \in X)(\forall\, u \in Y) \qquad \Phi(x) \geq -\Psi(\hat{u}) = \Phi(\hat{x}) \text{ and } -\Psi(\hat{u}) = \Phi(\hat{x}) \geq -\Psi(u).$$

Therefore $\inf\Phi = \Phi(\hat{x})$ and $\sup(-\Psi) = -\Psi(\hat{u})$.

(ii) $\Leftrightarrow$ (iii): It follows from (8.3) when we replace x and u with $\hat{x}$ and $\hat{u}$ respectively.

(iii) $\Leftrightarrow$ (iv) $\Leftrightarrow$ (v): It follows from Proposition 4.4.10(ii)-(iii). □

Remark 8.1.2. Conditions (iv)-(v) in Theorem 8.1.1 are called *KKT (Karush–Kuhn–Tucker) conditions*. Once one ensures that strong duality holds (that is, $\inf\Phi + \inf\Psi = 0$) they provide fully characterizations for a couple $(\hat{x}, \hat{u})$ to be a primal and dual solution.

Corollary 8.1.3. *Under the same assumptions of Theorem 8.1.1, suppose additionally that $\Phi = f + g \circ A$ is proper. Then, for every $\hat{x} \in X$, the following statements are equivalent.*

(i) $0 \in \partial(f + g \circ A)(\hat{x}), S^* \neq \varnothing$, *and* $\inf_X \Phi + \inf_Y \Psi = 0$.

(ii) $0 \in \partial f(\hat{x}) + A^* \partial g(A\hat{x})$.

Proof. We use the equivalence of conditions (i) and (v) in Theorem8.1.1. We have

$$\begin{aligned} 0 \in \partial f(\hat{x}) + A^* \partial g(A\hat{x}) &\Leftrightarrow \exists\, \hat{u} \in \partial g(A\hat{x}),\ 0 \in \partial f(\hat{x}) + A^* \hat{u} \\ &\Leftrightarrow \exists\, \hat{u} \in Y \text{ s.t. } \hat{u} \in \partial g(A\hat{x}) \text{ and } -A^* \hat{u} \in \partial f(\hat{x}) \\ &\Leftrightarrow \hat{x} \in S,\ \exists\, \hat{u} \in S_*, \text{ and } \inf_X \Phi + \inf_Y \Psi = 0. \end{aligned}$$

Then we note that, since Φ is proper, $\hat{x} \in S \Leftrightarrow 0 \in \partial(f + g \circ A)(\hat{x})$. □

Corollary 8.1.4. *Under the same assumptions of Corollary 8.1.3, suppose additionally that $S \neq \varnothing$ and that $\partial(f + g \circ A) = \partial f + A^* \partial g A$. Then $S^* \neq \varnothing$ and $\inf_X \Phi + \inf_Y \Psi = 0$.*

Proof. Since, the calculus rule $\partial(f + g \circ A) = \partial f + A^* \partial g A$ holds and there exists $\hat{x} \in S$, we have $0 \in \partial\Phi(\hat{x}) = \partial f(\hat{x}) + A^* \partial g(A\hat{x})$. Thus, the statement follows from Corollary 8.1.3. □

Remark 8.1.5. In the previous two corollaries we assumed that Φ is proper, that is, that $\mathrm{dom}\Phi \neq \varnothing$. This condition can be written also in another way. Indeed, since $\Phi(x) = f(x) + g(Ax)$, we have

$$\mathrm{dom}\Phi \neq \varnothing \Leftrightarrow \mathrm{dom}g - A(\mathrm{dom}f) \neq \varnothing \Leftrightarrow 0 \in \mathrm{dom}g - A(\mathrm{dom}f). \tag{8.4}$$

In the following we give a stronger condition that ensures that the same conclusions of Corollary 8.1.4 hold without any requirement about the primal problem and about the validity of a calculus rules for subdifferential (which is often difficult to check directly).

Theorem 8.1.6. *Under the same assumptions of Theorem 8.1.1, suppose additionally that $0 \in \mathrm{int}(\mathrm{dom}g - A(\mathrm{dom}f))$. Then Φ is proper and*

$$\inf_X \Phi = -\min_Y \Psi, \tag{8.5}$$

meaning that $S^* \neq \varnothing$ and $\inf_X \Phi + \inf_Y \Psi = 0$.

Proof. We already noted that $\mathrm{dom}\Phi \neq \varnothing \Leftrightarrow 0 \in \mathrm{dom}g - A(\mathrm{dom}f)$. Thus, Φ is a proper function. If $\inf_X \Phi = -\infty$, it follows from (8.2) that $\inf_Y \Psi = +\infty$ and hence $\Psi \equiv +\infty$; so, (8.5) holds. Suppose then that $\inf_X \Phi > -\infty$. Define

$$h\colon Y \to [-\infty, +\infty], \qquad h(v) = \inf_{x \in X} f(x) + g(Ax + v).$$

It is easy to verify that $\mathrm{dom}\, h = \{v \in Y \mid h(v) < +\infty\} = \mathrm{dom}g - A(\mathrm{dom}f)$. It follows from Proposition 1.3.6(iv) that h is convex. Moreover, for every $v \in Y$, $h(v) > 0$. Indeed if there were a $v \in Y$ such that $h(v) = -\infty$, then, since $0 \in \mathrm{int}(\mathrm{dom}h)$ there would exist $t > 0$ such that $-tv \in \mathrm{dom}\, h$. Then $0 = (1 - \alpha)v + \alpha(-tv)$ for some $\alpha \in]0, 1[$ and hence $h(0) \leq (1 - \alpha)h(v) + \alpha h(-tv) = -\infty$. Hence $\inf \Phi = h(0) = -\infty$ which contradicts our

assumption that $\inf \Phi > -\infty$. So, $h\colon Y \to]-\infty, +\infty]$ and $0 \in \operatorname{int}(\operatorname{dom} h)$. Therefore, it follows from Theorem 4.5.1 that $\partial h(0) \neq \varnothing$. Let $\hat{u} \in \partial h(0)$. Then,

$$\begin{aligned}(\forall\, v \in Y)(\forall\, x \in X) \qquad h(0) &\leq h(v) - \langle v, \hat{u}\rangle \\ &\leq f(x) - \langle v, \hat{u}\rangle + g(Ax + v) \\ &= f(x) - \langle x, -A^*\hat{u}\rangle + g(Ax + v) - \langle Ax + v, \hat{u}\rangle.\end{aligned}$$

Taking the infimum first in v (with x fixed) and then in x in the right hand side of the above inequality we get

$$\inf_X \Phi = h(0) \leq -f^*(-A^*\hat{u}) - g^*(\hat{u}) \leq \sup_Y(-\Psi) \leq \inf_X \Phi.$$

Since $\inf \Phi \in \mathbb{R}$, we have $\inf_X \Phi = -\Psi(\hat{u}) = \sup_Y -\Psi$ and the statement follows. □

Remark 8.1.7. Suppose that there exists $x \in \operatorname{dom} f$ such that $Ax \in \operatorname{int}(\operatorname{dom} g)$. [2] Then there exists $\delta > 0$ such that $Ax + B_\delta(0) \subset \operatorname{dom} g$. Therefore $B_\delta(0) \subset \operatorname{dom} g - Ax \subset \operatorname{dom} g - A(\operatorname{dom} f)$ and hence $0 \in \operatorname{int}(\operatorname{dom} g - A(\operatorname{dom} f))$.

Remark 8.1.8. Let $f \in \Gamma_0(X)$, $g \in \Gamma_0(Y)$, and $A\colon X \to Y$ be a linear operator and suppose that $0 \in \operatorname{int}(\operatorname{dom} g - A(\operatorname{dom} f))$. Then,

$$0 \in \partial(f + g \circ A)(\hat{x}) \;\Leftrightarrow\; 0 \in \partial f(\hat{x}) + A^*\partial g(A\hat{x}).$$

Indeed it just follows from Theorem 8.1.6 and Corollary 8.1.3.

Corollary 8.1.9 (Moreau-Rockafellar)**.** *Let $f \in \Gamma_0(X)$, $g \in \Gamma_0(Y)$, and $A\colon X \to Y$ be a linear operator and suppose that $0 \in \operatorname{int}(\operatorname{dom} g - A(\operatorname{dom} f))$. Then,*

$$(\forall\, x \in X) \qquad \partial(f + g \circ A)(x) = \partial f(x) + A^*\partial g(Ax). \tag{8.6}$$

Proof. Note that the inclusion "$\supset$" is always true. We use Remark 8.1.8 and Remark 3.2.4(i). For every $x \in X$ and $u \in X$, we have

$$\begin{aligned}u \in \partial(f + g \circ A)(x) &\;\Leftrightarrow\; 0 \in \partial(f + g \circ A)(x) - u = \partial(f - \langle u, \cdot\rangle + g \circ A)(x) \\ &\;\Leftrightarrow\; 0 \in \partial(f - \langle u, \cdot\rangle)(x) + A^*\partial g(Ax) \\ &\;\Leftrightarrow\; 0 \in \partial f(x) - u + A^*\partial g(Ax) \\ &\;\Leftrightarrow\; u \in \partial f(x) + A^*\partial g(Ax).\end{aligned}$$

Note that we could applied Remark 8.1.8 to $f - \langle u, \cdot\rangle$ since $\operatorname{dom}(f + \langle u, \cdot\rangle) = \operatorname{dom} f$. □

Corollary 8.1.10. *Let $f, g \in \Gamma_0(X)$ and suppose that $0 \in \operatorname{int}(\operatorname{dom} g - \operatorname{dom} f)$ (which is true if f is continuous at some point in $\operatorname{dom} g$). Then $\partial(f + g) = \partial f + \partial g$.*

Proposition 8.1.11. *Suppose that ∂f is single valued on its domain and denote by $\nabla f\colon \operatorname{dom} \partial f \to X$ its unique selection. Define the Bregman distance defined by f as*

$$D_f\colon X \times \operatorname{dom} \partial f \to]-\infty, +\infty], \quad D_f(x, y) = f(x) - f(y) - \langle \nabla f(x), x - y\rangle. \tag{8.7}$$

Let $\hat{x}$ and $\hat{u}$ be solution of problems $(\mathscr{P})$ and $(\mathscr{D})$ respectively and suppose that strong duality holds, that is, $\inf_X \Phi + \inf_Y \Psi = 0$. Let $x \in \operatorname{dom} \partial f$ and $u \in Y$ be such that $x \in \partial f^*(-A^*u)$. Then

$$\Psi(u) - \Psi(\hat{u}) \geq D_f(\hat{x}, x). \tag{8.8}$$

[2] This is equivalent to require that g is continuous at Ax.

Proof. It follows from Theorem 8.1.1(iii) and the definition of x that

$$f(\hat{x}) + f^*(-A^*\hat{u}) = \langle -A^*\hat{u}, \hat{x}\rangle \quad \text{and} \quad f(x) + f^*(-A^*u) = \langle -A^*u, x\rangle.$$

Thus, since $-A^*u = \nabla f(x)$,

$$\begin{aligned} f^*(-A^*u) - f^*(-A^*\hat{u}) &= f(\hat{x}) - f(x) + \langle A^*\hat{u}, \hat{x}\rangle + \langle -A^*u, x\rangle \\ &= D_f(\hat{x}, x) + \langle -A^*u, \hat{x} - x\rangle + \langle A^*\hat{u}, \hat{x}\rangle + \langle -A^*u, x\rangle \\ &= D_f(\hat{x}, x) + \langle \hat{u} - u, A\hat{x}\rangle. \end{aligned}$$

Now, since $A\hat{x} \in \partial g^*(\hat{u})$, we have

$$g^*(u) - g^*(\hat{u}) \geq \langle u - \hat{u}, A\hat{x}\rangle.$$

Summing the two inequalities above, we have

$$\big(f^*(-A^*u) + g^*(u)\big) - \big(f^*(-A^*\hat{u}) + g^*(\hat{u})\big) \geq D_f(\hat{x}, x)$$

and the statement follows. □

Example 8.1.12 (Equality constraints)**.** We consider the problem

$$\min_{Ax=b} f(x) \tag{8.9}$$

and we assume that a solution exists. We also assume that f is continuous at some x such that $Ax = b$. This problem can be equivalently formulated as

$$\min_{x\in X} f(x) + \iota_{\{b\}}(Ax), \tag{8.10}$$

which is in the form ($\mathscr{P}$). Then, the dual problem of (8.9) is

$$\min_{u\in Y} f^*(-A^*u) + \langle u, b\rangle.$$

In view of Corollary 8.1.4, to ensure the existence of dual solutions and a zero duality gap, we need to find conditions ensuring the validity of the calculus rule (8.6). We first prove that if $x \in X$ is such that $Ax = b$, then

$$\partial(\iota_{\{b\}} \circ A)(x) = R(A^*) = A^*\partial\iota_{\{b\}}(Ax). \tag{8.11}$$

Indeed, we note that $\iota_{\{b\}} \circ A = \iota_{A^{-1}(b)}$ and $A^{-1}(b) = x + N(A)$. Then,

$$\begin{aligned} u \in \partial(\iota_{\{b\}} \circ A)(x) &\iff (\forall y \in A^{-1}(b)) \quad \langle u, y - x\rangle \leq 0 \\ &\iff (\forall v \in N(A)) \quad \langle u, v\rangle \leq 0 \\ &\iff u \in N(A)^\perp = R(A^*). \end{aligned}$$

Therefore, $\partial(\iota_{\{b\}} \circ A)(x) = R(A^*)$. Moreover $A^*\partial\iota_{\{b\}}(Ax) = A^*\partial\iota_{\{b\}}(b)$ and the subdifferential of $\iota_{\{b\}}$ is

$$\partial\iota_{\{b\}} \colon Y \to Y \colon y \mapsto \begin{cases} Y & \text{if } y = b \\ \varnothing & \text{if } y \neq b, \end{cases} \tag{8.12}$$

hence $A^*\partial\iota_{\{b\}}(Ax) = R(A^*)$ and (8.11) holds. Finally, recalling the calculus rule for subdifferentials in Corollary 8.1.10 and that we assumed that f is continuous at some $x \in \operatorname{dom}(\iota_{\{b\}} \circ A)$, then, we have $\partial(f + \iota_{\{b\}} \circ A)(x) = \partial f(x) + \partial(\iota_{\{b\}} \circ A)(x) = \partial f(x) + A^*\partial\iota_{\{b\}}(Ax)$ and hence (8.6) holds. We note in passing that Fermat's rule for (8.10) is

$$\begin{aligned} 0 \in \partial(f + \iota_{\{b\}} \circ A)(\hat{x}) &\Leftrightarrow 0 \in \partial f(\hat{x}) + A^*\partial\iota_{\{b\}}(A\hat{x}) \\ &\Leftrightarrow 0 \in \partial f(\hat{x}) + R(A^*) \\ &\Leftrightarrow \exists\, \hat{u} \in Y \quad A^*\hat{u} \in \partial f(\hat{x}). \end{aligned}$$

In the differentiable case, this condition reduces to the classical Lagrange multiplier rule, that is, $\hat{x}$ is a solution of (8.9) if and only if there exists a multiplier $\hat{u}$ such that $A^*\hat{u} = \nabla f(\hat{x})$.

Example 8.1.13 (Linear programming)**.** The problem is

$$\min_{Ax \leq b} \langle c, x \rangle,$$

where $A\colon \mathbb{R}^d \to \mathbb{R}^m$ is linear and $c \in \mathbb{R}^d$. This can be equivalently formulated as

$$\min_{x \in X} \langle c, x \rangle + \iota_{\mathbb{R}^m_-}(Ax - b),$$

which is in the form $(\mathscr{P})$, with $f(x) = \langle c, x \rangle$ and $g = \iota_{\mathbb{R}^m_-}(\cdot - b)$. If we assume that there exists x such that $Ax < b$ (*Slater's condition*), then $Ax \in \operatorname{int}(\operatorname{dom} g)$ and the qualification condition $0 \in \operatorname{int}(\operatorname{dom} g - A(\operatorname{dom} f))$ holds, so that Theorem 8.1.1 and Theorem 8.1.6 fully apply. Therefore, since $f^* = \iota_{\{c\}}$ and $g^* = \iota^*_{\mathbb{R}^m_-} + \langle \cdot, b \rangle = \iota_{\mathbb{R}^m_+} + \langle \cdot, b \rangle$, the dual problem

$$\min_{\substack{A^*u + c = 0 \\ u \geq 0}} \langle u, b \rangle,$$

admits solutions and the duality gap is zero. In this case, the KKT conditions (v) in Theorem 8.1.1 becomes

$$-A^*\hat{u} = c \quad \text{and} \quad \hat{u} \in \partial\iota_{\mathbb{R}^m_-}(A\hat{x} - b).$$

The last condition above then is equivalent to

$$\hat{u}_i \in \partial\iota_{\mathbb{R}_-}\big((A\hat{x})_i - b_i\big) \;\Leftrightarrow\; \begin{cases} \hat{u}_i = 0 & \text{if } (A\hat{x})_i < b_i \\ \hat{u}_i \geq 0 & \text{if } (A\hat{x})_i = b_i. \end{cases}$$

Note that these conditions alltogether do not uniquely determine the primal solution $\hat{x}$ from the dual $\hat{u}$. We can only say that $\hat{u}$ is a solution of the linear system

$$A_I\hat{x} = b_I,$$

where A_I is the matrix composed by the rows of A indexed in the set $I = \big\{i \in \{1, \ldots, m\} \,|\, \hat{u}_i > 0\big\}$ and $b_I = (b_i)_{i \in I}$.

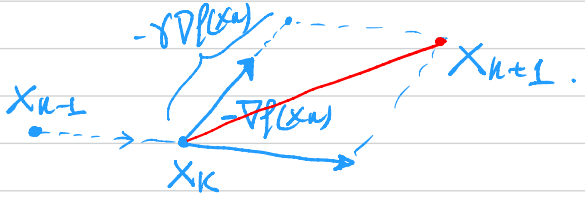
Lecture 9

ACCELERATED GRADIENT METHODS

1. Heavy-ball method (Polyak 64')
2. Nesterov's accelerated gradient algorithm. (83').

The heavy-ball method

$$x_{k+1} = x_k - \gamma \nabla f(x_k) + \underbrace{\beta (x_k - x_{k-1})}_{\text{momentum}} \quad , \gamma > 0, \beta \in [0,1[$$



Theorem 1

$f(x) = \langle Ax, x \rangle + \langle b, x \rangle + c$, A positive definite

μ and L are the minimum and maximum eigenvalues of A. ($L = \|A\|$, $\mu = \|A^{-1}\|^{-1}$).

1. $\gamma < \frac{2(1+\beta)}{L}$, $\beta \in]0,1[$. Then $\|x_k - x_*\| \to 0$ linearly.

2. If $\gamma = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ $\beta = \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2$.

Then $\|x_k - x_*\| = O(q^k)$ $\boxed{q = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}}$.

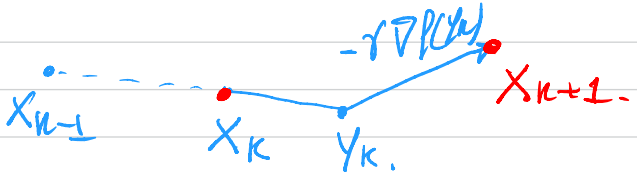
↑ rate of convergence

Nesterov's accelerated gradient algorithm (83')

$$\begin{cases} x_{k+1} = y_k - \gamma \nabla f(y_k) \\ y_{k+1} = x_{k+1} + \beta_{k+1}(x_{k+1} - x_k) \end{cases}$$

$$y_k = x_k + \beta_k (x_k - x_{k-1})$$

$$\rightarrow x_{k+1} = (x_k + \beta_k(x_k - x_{k-1})) - \gamma \nabla f(x_k + \beta_k(x_k - x_{k-1}))$$



FISTA

Let $(t_k)_{k\in\mathbb{N}}$, $t_0 = 1$, $t_k \geq 1$ and $t_k^2 - t_k \leq t_{k-1}^2$ $\forall k$.

Let $y_0 \in X$, $\gamma \leq 1/L$. Define

$$x_{k+1} = \text{prox}_{\gamma g}(y_k - \gamma \nabla f(y_k))$$

$$y_{k+1} = x_{k+1} + \underbrace{\frac{t_k - 1}{t_{k+1}}}_{\beta_{k+1}}(x_{k+1} - x_k)$$

?

[minimize $f(x) + g(x)$, f is L-Lipschitz smooth, $g \in \Gamma_0(X)$]

Convergence analysis.

Reminder: $F = f + g$. $x_{k+1} = \text{prox}_{\gamma g}(x_k - \gamma \nabla f(x_k))$. $\forall x \in X$:

$$\|x_{k+1} - x\|^2 \le \|x_k - x\|^2 + 2\gamma (F(x) - F(x_{k+1})) + (\gamma L - 1)\|x_k - x_{k+1}\|^2$$

Now, $x_{k+1} = \text{prox}_{\gamma g}(y_k - \gamma \nabla f(y_k))$. Therefore $\forall x \in X$:

$$\|x_{k+1} - x\|^2 \le \|y_k - x\|^2 + 2\gamma (F(x) - F(x_{k+1})) + \underbrace{(\gamma L - 1)}_{\le 0} \|y_k - x_{k+1}\|^2.$$

Since $\gamma L - 1 \le 0$, we have

$$F(x_{k+1}) + \frac{\|x_{k+1} - x\|^2}{2\gamma} \le F(x) + \frac{\|y_k - x\|^2}{2\gamma} \quad \forall x \in X.$$

Theorem $\forall n \in \mathbb{N}$, $k \ge 1$, $F(x_k) - \min F \le \dfrac{\|y_0 - x_*\|^2}{2\gamma t_{k-1}^2}$

Moreover,

1. If $t_k = \dfrac{1 + \sqrt{1 + 4t_{k-1}^2}}{2}$, then $t_k^2 - t_k = t_{k-1}^2$ and $\dfrac{1}{t_{k-1}} \le \dfrac{2}{k+1}$

2. If $t_k = \frac{k+a}{a}$ with $a \geq 2$, then $t_k^2 - t_k \leq t_{k+1}^2$

Hence in both cases $F(x_k) - \min F = O\left(\frac{1}{k^2}\right)$

Remark, in case 2 above if $a > 2$, then $x_k \to x_*$ with $x_* \in \operatorname{argmin} F$.

Proof:

$$y_{k+1} = x_{k+1} + \frac{t_k - 1}{t_{k+1}} (x_{k+1} - x_k)$$

$$= \left(1 - \frac{1}{t_{k+1}}\right) x_{k+1} + \frac{1}{t_{k+1}} \underbrace{\left[x_k + t_k (x_{k+1} - x_k)\right]}_{v_{k+1}}$$

- $y_k = \left(1 - \frac{1}{t_k}\right) x_k + \frac{v_k}{t_k}$. $\quad (v_0 = y_0)$

$$\frac{v_{k+1} - x_k}{t_k} = x_{k+1} - x_k \Rightarrow x_{k+1} = x_k + \frac{v_{k+1} - x_k}{t_k}.$$

- $x_{k+1} = \left(1 - \frac{1}{t_k}\right) x_k + \frac{v_{k+1}}{t_k}$

$$F(x_{k+1}) + \frac{\overbrace{\|x_{k+1}-x\|^2}^{a}}{2\gamma} \le F(x) + \frac{\overbrace{\|y_k-x\|^2}^{b}}{2\gamma} \quad \forall x \in X.$$

take $x = \left(1-\frac{1}{t_k}\right)x_k + \frac{x_*}{t_k}$ (convex combination of x_k and x_*)

$a = \frac{v_{k+1}-x_*}{t_k}$ and $b = \frac{v_k-x_*}{t_k}$.

$$F(x_{k+1}) + \frac{\|v_{k+1}-x_*\|^2}{2\gamma t_k^2} \le \left(1-\frac{1}{t_k}\right)F(x_k) + \frac{1}{t_k}F(x_*) + \frac{\|v_k-x_*\|^2}{2\gamma t_k^2}$$

$$F(x_{k+1}) - F(x_*) + \frac{\|v_{k+1}-x_*\|^2}{2\gamma t_k^2} \le \left(1-\frac{1}{t_k}\right)(F(x_k)-F(x_*)) + \frac{\|v_k-x_*\|^2}{2\gamma t_k^2}$$

$$t_k^2\left(F(x_{k+1})-F(x_*)\right) + \frac{\|v_{k+1}-x_*\|^2}{2\gamma} \overset{(1)}{\le} \underbrace{(t_k^2-t_k)}_{\le t_{k-1}^2}(F(x_k)-F(x_*)) + \frac{\|v_k-x_*\|^2}{2\gamma}.$$

$$\overset{(2)}{\leq} t_{k-1}^2 (F(x_k) - F(x_{*1})) + \frac{\|v_k - x_*\|^2}{2\gamma}$$

We got the recursive inequality (2). Thus, we have

$$t_{k-1}^2 (F(x_k) - F(x_{*1})) + \frac{\|v_k - x_*\|^2}{2\gamma}$$

$$\leq t_0^2 (F(x_1) - F(x_{*1})) + \frac{\|v_1 - x_*\|^2}{2\gamma}$$

$$\overset{(1)}{\leq} \frac{\|v_0 - x_*\|^2}{2\gamma}$$

$$\Rightarrow t_{k-1}^2 (F(x_k) - F(x_{*1})) \leq \frac{\|\underbrace{y_0}_{v_0 = y_0} - x_*\|^2}{2\gamma}$$

the first part of the statement follows.

1. $t_k^2 = t_k + t_{k-1}^2 > t_{k-1}^2 \Rightarrow t_k > t_{k-1}$

$$t_k = t_k^2 - t_{k-1}^2 = (t_k - t_{k-1})(t_k + t_{k-1}) \leq 2 t_k (t_k - t_{k-1})$$

$$t_k - t_{k-1} \geq \frac{1}{2} \Rightarrow t_k - t_0 = \sum_{i=1}^{k} (t_i - t_{i-1}) \geq \frac{k}{2}.$$

$$\Rightarrow t_k \geq \frac{k}{2} + 1 = \frac{k+2}{2}.$$

Application to LASSO.

$$\min \frac{1}{2}\|Ax-b\|^2 + \|x\|_1.$$

$$\begin{cases} x_{k+1} = \text{soft}_{\gamma\lambda}\left(y_k - \gamma A^*(Ay_k - b)\right) \\ y_{k+1} = x_{k+1} + \frac{t_k - 1}{t_{k+1}}(x_{k+1} - x_k) \end{cases}$$

Fast iterative soft thresholding algorithm (FISTA) 2009.

Lecture 10

- Dual algorithms
- Regularized empirical risk minimization (SVM's)

Dual algorithms

$$(P) \quad \min_{x \in X} f(x) + g(Ax) =: \phi(x) \qquad (D) \quad \min_{u \in Y} g^*(u) + f^*(-A^*u) =: \psi(u)$$

$f \in \Gamma_0(X)$, $g \in \Gamma_0(Y)$, $A : X \to Y$, $A^* : Y \to X$

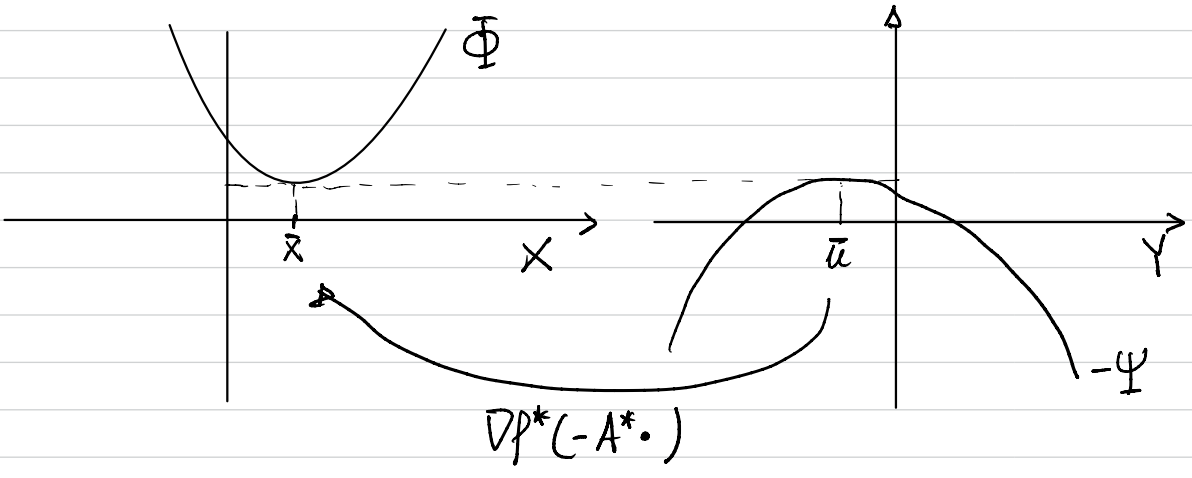
- $0 \in \text{int}(\text{dom } g - A(\text{dom } f))$ $\Rightarrow$ $\text{argmin } \Psi \neq \emptyset$ and $\inf \phi = \sup -\Psi$.

Assumption: f is μ-strongly convex ($\mu > 0$).

Consequences: 1) Problem P admits a unique solution ($\bar{x}$)

2) $\text{dom } f^* = X$ and f^* is $\frac{1}{\mu}$-Lipschitz smooth.

KKT : $\bar{x} = \nabla f^*(-A^*\bar{u})$ and $A\bar{x} \in \partial g^*(\bar{u})$



u_k in the dual $\longrightarrow$ $x_k = \nabla f^*(-A^* u_k)$

Prop. $\bar{x} = \nabla f^*(-A^*\bar{u})$, $x = \nabla f^*(-A^* u)$, with $u \in Y$.

Then $\boxed{\frac{\mu}{2}\|x - \bar{x}\|^2 \leq \Psi(u) - \Psi(\bar{u}).}$

Proof: $f(\bar{x}) + f^*(-A^*\bar{u}) = \langle \bar{x}, -A^*\bar{u} \rangle$

$f(x) + f^*(-A^* u) = \langle x, -A^* u \rangle$.

$$\psi(u) - \psi(\bar{u}) = f^*(-A^*u) - f^*(-A^*\bar{u}) + g^*(u) - g^*(\bar{u}).$$

$$= f(\bar{x}) + \langle \bar{x}, A^*\bar{u} \rangle - \langle x, A^*u \rangle - f(x) + g^*(u) - g^*(\bar{u})$$

[Since $A\bar{x} \in \partial g^*(\bar{u}) \Rightarrow g^*(u) - g^*(\bar{u}) \geq \langle u - \bar{u}, A\bar{x} \rangle$.]

$$\geq f(\bar{x}) - f(x) + \langle \bar{x}, A^*\bar{u} \rangle - \langle x, A^*u \rangle + \langle u - \bar{u}, A\bar{x} \rangle$$

$\langle \bar{x}, A^*u \rangle$

$= \langle \bar{x}, A^*u \rangle$

$$= f(\bar{x}) - f(x) + \langle \bar{x} - x, A^*u \rangle$$

Since $x = \nabla f^*(-A^*u) \iff -A^*u \in \partial f(x)$.

$$f(\bar{x}) - f(x) \geq \langle \bar{x} - x, -A^*u \rangle + \frac{\mu}{2}\|\bar{x} - x\|^2.$$

$$\geq \langle \bar{x} - x, -A^*u \rangle + \langle \bar{x} - x, A^*u \rangle + \frac{\mu}{2}\|\bar{x} - x\|^2$$

□.

Remark

if $(u_k)_{k\in\mathbb{N}}$ is such that $\psi(u_k) \to \min \psi$,

then $\|x_k - \bar{x}\| \to 0$ (where $x_k = \nabla f^*(-A^* u_k)$).

Example

assume that we can compute the prox of g^*.

$\nabla(f^*_\circ(-A^*)) = -A \nabla f^*(-A^*)$

$f^*_\circ(-A^*)$ is Lipschitz smooth with constant $\frac{\|A\|^2}{\mu}$

FISTA ON THE DUAL.

$$\begin{cases} u_{k+1} = \operatorname{prox}_{\gamma g^*}\big(v_k + \gamma A \overbrace{\nabla f^*(-A^* v_k)}^{y_k}\big) \\ v_{k+1} = u_{k+1} + \frac{t_k - 1}{t_{k+1}}(u_{k+1} - u_k) \end{cases}$$

$\Leftrightarrow$

$$\begin{cases} u_{k+1} = \operatorname{prox}_{\gamma g^*}(v_k + \gamma A y_k) \\ v_{k+1} = u_{k+1} + \frac{t_k - 1}{t_{k+1}}(u_{k+1} - u_k) \\ y_{k+1} = \nabla f^*(-A^* v_{k+1}) \end{cases}$$

Dual FISTA

$\gamma \leq \frac{\mu}{\|A\|^2}$

Convergence properties. $x_k = \nabla f^*(-A^* u_k)$.

$$\frac{2}{\mu}\|x_k - \bar{x}\|^2 \le \Psi(u_k) - \min \Psi = \mathcal{O}\left(\frac{1}{k^2}\right)$$

$$\|x_k - \bar{x}\| = \mathcal{O}\left(\frac{1}{k}\right)$$

Regularized empirical risk minimization (SVM's).

$(x_i, y_i)_{i \le n} \in (\mathcal{X} \times \mathcal{Y})^n$ the training set.

$$\min_{w \in H} \underbrace{\frac{1}{\lambda n} \sum_{i=1}^{n} \ell(y_i, \langle w, \Lambda(x_i) \rangle)}_{g(Aw)} + \underbrace{\frac{1}{2}\|w\|^2}_{f(w)} \qquad \lambda > 0$$

$\Lambda : \mathcal{X} \to H$ the feature map. H is abstract Hilbert space

$K(x, x') = \langle \Lambda(x), \Lambda(x') \rangle$

$$\mathcal{H}_K = \{ h : \mathcal{X} \to \mathbb{R} \mid h = \langle w, \Lambda(\cdot) \rangle, \ w \in H \}$$

$$\|h\|_K = \inf \{ \|w\| \mid h = \langle w, \Lambda(\cdot) \rangle \}$$

$\ell : Y \times \mathbb{R} \to \mathbb{R}_+$ which is convex in the second var.

$$\left[\min_u g^*(u) + f^*(-A^* u) \right]$$

$$A : H \to \mathbb{R}^n \qquad Aw = -\begin{bmatrix} \langle w, \Lambda(x_1) \rangle \\ \vdots \\ \langle w, \Lambda(x_n) \rangle \end{bmatrix}$$

$$g : \mathbb{R}^n \to \mathbb{R} \qquad g(z) = \frac{1}{\lambda n} \sum_{i=1}^n \ell(y_i, -z_i) = \sum_{i=1}^n g_i(z_i)$$

where $g_i(t) = \frac{1}{\lambda n} \ell(y_i, -t)$.

$$f(w) = \frac{1}{2} \|w\|^2.$$

- $A^* : \mathbb{R}^n \to H \qquad A^* u = -\sum_{i=1}^n u_i \Lambda(x_i)$.
- $f^*(w) = \frac{1}{2} \|w\|^2$.
- $g^*(u) = \sum_{i=1}^n g_i^*(u_i)$.

$$g_i^*(s) = \sup_t \; st - \frac{1}{\lambda n} \ell(y_i, -t)$$

$$= \frac{1}{\lambda n} \sup_t \left[(-\lambda n s)(-t) - \ell(y_i, -t) \right]$$

$$= \frac{1}{\lambda n} [\ell(y_i, \cdot)]^* (-\lambda n s)$$

$$= \frac{1}{\lambda n} \ell^*(y_i, -\lambda n s).$$

The dual of Regularized ERM is

$$\min_{u \in \mathbb{R}^n} \frac{1}{2} \left\| \sum_{i=1}^n u_i \Lambda(x_i) \right\|^2 + \frac{1}{\lambda n} \sum_{i=1}^n \ell^*(y_i, -\lambda n u_i)$$

$$\left\| \sum_{i=1}^n u_i \Lambda(x_i) \right\|^2 = \sum_i \sum_j u_i u_j \langle \Lambda(x_i), \Lambda(x_j) \rangle$$

$$= \sum_{ij} u_i u_j K(x_i, x_j).$$

$K \in \mathbb{R}^{n \times n}$ $K_{ij} = K(x_i, x_j)$ Gram matrix

$$= \langle Ku, u \rangle = u^T K u$$

$$\min_{u \in \mathbb{R}^n} \frac{1}{2} u^T K u + \frac{1}{\lambda n} \sum_{i=1}^n \ell^*(y_i, -\lambda n u_i)$$

KKT $\quad \bar{w} = \sum_{i=1}^{n} \bar{u}_i \Lambda(x_i)$ $\quad$ [Representer thm.]

$$A\bar{w} \in \partial g^*(\bar{u}) \iff (A\bar{w})_i \in \partial g_i^*(\bar{u}_i)$$

$$\iff \bar{u}_i \in \partial g_i(-\langle \bar{w}, \Lambda(x_i) \rangle)$$

$$\iff -\bar{u}_i \in \frac{1}{n} \partial \ell(y_i, \langle \bar{w}, \Lambda(x_i) \rangle) \quad \forall i.$$

$$\bar{h} = \langle \bar{w}, \Lambda(\cdot) \rangle$$

$$\bar{h}(x) = \sum_{i=1}^{n} \bar{u}_i K(x_i, x), \quad \bar{h} = \sum_{i=1}^{n} \bar{u}_i K(x_i, \cdot)$$

Let $(u^k)_{k \in \mathbb{N}}$ be such that $\psi(u^k) \to \min \psi$.

$$w^k = \sum_{i=1}^{n} u_i^k \Lambda(x_i)$$

$$\frac{1}{2} \| w^k - \bar{w} \|^2 \leq \psi(u^k) - \psi(\bar{u})$$

$$h_k = \langle w^k, \Lambda(\cdot) \rangle.$$

$$|h_k(x) - \bar{h}(x)| = |\langle w^k - \bar{w}, \Lambda(x)\rangle|$$

$$\leq \|w^k - \bar{w}\| \, \|\Lambda(x)\|$$

$$= \|w^k - \bar{w}\| \sqrt{K(x,x)}$$

$$\leq \underbrace{\left(\sup_x \sqrt{K(x,x)}\right)}_{\kappa} \|w^k - \bar{w}\|$$

$$\|h_k - \bar{h}\|_\infty \leq \kappa \|w^k - \bar{w}\| \leq \kappa \cdot 2 \cdot \sqrt{\varphi(w^k) - \min \varphi}.$$

Prox-gradient on the dual

$$\begin{cases} u^{k+1} = \operatorname{prox}_{\gamma g^*}(u^k - \gamma K u^k) \\ w^k = \sum_{i=1}^n u_i^k \Lambda(x_i) \end{cases}$$

Example SVM

l is the hinge loss. (for classification)

$$l(y_i, t) = (1 - y_i t)_+ \qquad y_i \in Y = \{-1, 1\}.$$
$$= \chi(r) \qquad \chi(r) = (1-r)_+$$

$$l^*(y_i, s) = \chi^*(y_i s) = y_i s + i_{[-1,0]}(y_i s).$$

The dual problem.

$$\min_u \frac{1}{2} u^T K u + \frac{1}{\lambda n}\left[\sum_{i=1}^n y_i(-\lambda n u_i) + \sum_{i=1}^n i_{[-1,0]}(-\lambda n u_i)\right]$$

$$\boxed{\min_u \frac{1}{2} u^T K u - \langle y, u\rangle + \frac{1}{\lambda n} i_{[0, \frac{1}{\lambda n}]^n}(y \odot u)}$$

Example ridge regression

$$l(y_i, t) = \frac{1}{2}(y_i - t)^2.$$

$$\min_{u} \frac{1}{2} u^T (K + \lambda n I_d) u - \langle y, u \rangle.$$

$$\bar{u} = (K + \lambda n I_d)^{-1} y.$$