Some elements on convex optimization

Maxime Sangnier

March 24, 2020

Contents

1 Basics of convex optimization **5**

1.1 Introduction . . . 6

1.1.1 Optimization in statistics and data science . . . 6

1.1.2 Canonical optimization problem . . . 9

1.1.3 Numerical optimization . . . 10

1.2 Convex analysis . . . 12

1.2.1 Convex sets . . . 12

1.2.2 Convex functions . . . 14

1.2.3 Properties of minimizers . . . 18

1.2.4 Optimality conditions . . . 21

1.2.5 Convex optimization problems . . . 23

1.3 Legendre-Fenchel transformation and duality . . . 27

1.3.1 The convex conjugate . . . 27

1.3.2 Duality . . . 29

1.3.3 Generalized inequality constraints . . . 33

1.3.4 Tikhonov, Ivanov and Morozov regularizations . . . 34

1.3.5 A relevant example . . . 35

2 Optimization algorithms **37**

2.1 Introduction . . . 38

2.2 Greedy methods . . . 38

2.2.1 Orthogonal matching pursuit . . . 38

2.2.2 Compressive sampling matching pursuit . . . 39

2.3 Linear programming . . . 40

2.3.1 Convex relaxation of compressed sensing and basis pursuit . . . 40

2.3.2 The simplex method . . . 41

2.3.3 Barrier methods . . . 42

2.4 Primal methods . . . 44

2.4.1 Gradient method . . . 44

2.4.2 Quasi-Newton method . . . 47

2.4.3 Subgradient method . . . 48

2.4.4 Proximal gradient method . . . 49

2.4.5 Accelerated proximal gradient method . . . 53

2.4.6 Douglas-Rachford method 53
2.5 Primal-dual methods 54
2.5.1 Lagrange multipliers 54
2.5.2 Augmented Lagrange multipliers 56
2.5.3 Alternating direction method of multipliers 57

References **61**

Chapter 1

Basics of convex optimization

[1]

Contents

1.1 Introduction 6

1.1.1 Optimization in statistics and data science 6

1.1.2 Canonical optimization problem 9

1.1.3 Numerical optimization 10

1.2 Convex analysis 12

1.2.1 Convex sets 12

1.2.2 Convex functions 14

1.2.3 Properties of minimizers 18

1.2.4 Optimality conditions 21

1.2.5 Convex optimization problems 23

1.3 Legendre-Fenchel transformation and duality 27

1.3.1 The convex conjugate 27

1.3.2 Duality 29

1.3.3 Generalized inequality constraints 33

1.3.4 Tikhonov, Ivanov and Morozov regularizations 34

1.3.5 A relevant example 35

[1] This chapter is deeply inspired from P. Bianchi, O. Fercoq and A. Sabourin's lecture notes on Optimization for machine learning (University Paris-Saclay, Télécom ParisTech) [1].

1.1 Introduction

1.1.1 Optimization in statistics and data science

In statistics and data science, many estimation procedures rely on minimizing an objective function. This class of techniques is referred to as M-estimation.

The objective function to minimize may be interpreted as an energy (physics, chemistry), a cost (finance), or a distance to the estimand. In the three forthcoming examples, we deal with estimating a finite-dimensional parameter $\theta^\star \in \mathbb{R}^d$ of a probability distribution. It can be seen that the functions to minimize are respectively a negative likelihood and (penalized) distances of projection.

Example 1.1 (Maximum likelihood estimator)**.** *Let $(P_\theta)_{\theta\in\Theta}$ be a statistical model dominated by a measure ν and denote, for all $\theta \in \Theta$, $g_\theta = \frac{dP_\theta}{d\nu}$ the probability density function of P_θ with respect to ν. Let $\theta^\star \in \Theta$. Given an observation $\mathbf{X}$ sampled from $P_{\theta^\star}$, a maximum likelihood estimator of $\theta^\star$ is any $\hat\theta$ such that:*

$$\hat\theta \in \arg\max_{\theta\in\Theta} g_\theta(\mathbf{X}).$$

Example 1.2 (Ordinary least squares estimate)**.** *Let $X \in \mathbb{R}^{n\times d}$ be a design matrix and $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$ a sample of real-valued random variables such that:*

- $\operatorname{rank}(X) = d$;
- $\exists \theta^\star \in \mathbb{R}^d : \mathbb{E}Y = X\theta^\star$;
- $\exists \sigma^2 \in \mathbb{R}_+^* : \mathbb{V}(Y) = \sigma^2 I_n$,

where I_n is the identity matrix of size $n \times n$. An ordinary least squares estimate of $\theta^\star$ is any $\hat\theta \in \mathbb{R}^d$ such that:

$$\hat\theta \in \arg\min_{\theta\in\mathbb{R}^d} \|Y - X\theta\|_2^2.$$

Example 1.3 (Ridge and Lasso regression)**.** *With the same notation as in Example 1.2 and with hypotheses:*

- $\exists \theta^\star \in \mathbb{R}^d : \mathbb{E}Y = X\theta^\star$;
- $\exists \sigma^2 \in \mathbb{R}_+^* : \mathbb{V}(Y) = \sigma^2 I_n$,

a ridge estimate of $\theta^\star$ is any $\hat\theta \in \mathbb{R}^d$ such that:

$$\hat\theta \in \arg\min_{\theta\in\mathbb{R}^d} \|Y - X\theta\|_2^2 + \lambda\|\theta\|_2^2,$$

for $\lambda \in \mathbb{R}_+^*$. Respectively, a lasso estimate verifies:

$$\hat\theta \in \arg\min_{\theta\in\mathbb{R}^d} \|Y - X\theta\|_2^2 + \lambda\|\theta\|_1.$$

It is quite important to note that even though optimization is a prevailing topic in modern statistics, many estimation procedures do not rely on optimization. For instance, the method of moments does not make use of any minimization technique.

Counterexample 1.4 (Method of moments)**.** *Let X be a real-valued random variable such that there exist a mapping $h\colon \mathbb{R} \to \mathbb{R}$ and a bijective function $\phi\colon \mathbb{R} \to \mathbb{R}$ such that:*

- $-\infty < \mathbb{E}[h(X)] < +\infty$*;*
- $\phi(\theta^\star) = \mathbb{E}[h(X)]$*.*

Let now $(X_1, \dots, X_n)$ be an i.i.d. sample drawn from the same distribution as X. Then, $\hat{\theta} = \phi^{-1}(\frac{1}{n}\sum_{i=1}^n h(X_i))$ is an estimate of $\theta^\star$.

It is tempting to say that expressing an estimate as a minimizer of an energy function is quite a week characterization, since:

- it provides few information on the behavior of the estimator;
- it does not provide an easy way to compute the estimate.

Thus, resorting to M-estimation conveys the incapability to state something stronger about the estimand of interest. As we will see, that is often the case for current problems in statistics and data science.

As a result, a general procedure to estimate a quantity $\theta^\star \in \mathbb{R}^d$ of a distribution is first to write it as a minimizer of a risk. Given a random variable X drawn from the distribution of interest, this risk is generally the expectation of a cost function $L\colon \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}_+$:

$$\theta^\star \in \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \mathbb{E}[L(X, \theta)].$$

Then, we proceed similarly to the method of moments: given and *i.i.d.* sample of observations $(X_1, \dots, X_n)$ drawn from the same distribution as X, an estimate of $\theta^\star$ is any $\hat{\theta} \in \mathbb{R}^d$ such that:

$$\hat{\theta} \in \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n L(X_i, \theta).$$

If $\theta \in \mathbb{R}^d \mapsto \sum_{i=1}^n L(X_i, \theta)$ is a convex function, then computing the estimate $\hat{\theta}$ is manageable. However, in general this empirical risk may be non-convex. In this case, computing $\hat{\theta}$ is *hard* and one may prefer to solve instead a convex problem. To this end, L may be *convexified* to a function $\varphi\colon \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ such that $\theta \in \mathbb{R}^d \mapsto \sum_{i=1}^n \varphi(X_i, \theta)$ is convex. φ is called a *convex surrogate* of L.

Remark 1.5. Non-convex optimization is currently a hot research topic. Even though one may see non-convexity as a pitfall, current advances tend to show that:

- in some practical situations, non-convexity can be overcome (for instance concerning optimization on manifolds);

- reaching a global minimum of the empirical risk is not essential to produce a suitable estimate (for instance with models based on artificial neural networks).

A last step of this procedure is to *regularize* the empirical risk with a convex function $\psi\colon \mathbb{R}^d \to \mathbb{R}$. This appears useful for assuring numerical stability and statistical guarantees concerning the deviation between $\mathbb{E}[\varphi(X, \theta)]$ and $\frac{1}{n}\sum_{i=1}^n \varphi(X_i, \theta)$. At the end of the day, an estimate of $\theta^\star$ is $\hat{\theta} \in \mathbb{R}^d$ such that:

$$\hat{\theta} \in \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \varphi(X_i, \theta) + \lambda\psi(\theta),$$

for some $\lambda > 0$. This general procedure is often referred to as *empirical* or *regularized risk minimization.*

Remark 1.6. With this procedure, the function to minimize is the sum of two components:

- a finite sum of convex functions;
- a convex regularized.

This special structure may be used to provide efficient algorithms.

Example 1.7 (Linear classification)**.** *Let* $\{(X_i, Y_i)\}_{i\in[n]}$ *be an i.i.d. sample of couples of random variables such that* $\forall i \in [n]\colon X_i \in \mathbb{R}^d, Y_i \in \{-1, 1\}$*. We aim at estimating the classification function:*

$$\eta : x \in \mathbb{R}^d \mapsto \operatorname{sign}\left(\frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = -1|X = x)} - 1\right),$$

agreeing that $\operatorname{sign}(0) = 1$*. One can note that:*

$$\eta \in \operatorname*{arg\,min}_{h\colon \mathbb{R}^d \to \{-1,1\}} \mathbb{P}\left(Y \neq h(X)\right) = \operatorname*{arg\,min}_{h\colon \mathbb{R}^d \to \{-1,1\}} \mathbb{E}\left[\mathbb{1}_{\mathbb{R}_-}(Yh(X))\right].$$

In this problem, the quantity to estimate is a function from $\mathbb{R}^d$ *to* $\{-1, 1\}$*. Following a parametric approach, we consider the linear model* $\{x \in \mathbb{R}^d \mapsto \operatorname{sign}(\theta^\top x) : \theta \in \mathbb{R}^d\}$*. Moreover, we remark that* $\forall \theta \in \mathbb{R}^d$*:* $\mathbb{1}_{\mathbb{R}_-}(Y \operatorname{sign}(\theta^\top X)) = \mathbb{1}_{\mathbb{R}_-}(Y(\theta^\top X))$*. Thus, the empirical risk to minimize for estimating a linear classifier is:*

$$\theta \in \mathbb{R}^d \mapsto \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\mathbb{R}_-}(Y_i(\theta^\top X_i)).$$

The difficulty we face now is the non-convexity of $\theta \in \mathbb{R}^d \mapsto \mathbb{1}_{\mathbb{R}_-}(Y(\theta^\top X))$*. Yet, this function may be convexified to* $\varphi\colon \theta \in \mathbb{R}^d \mapsto \max(0, 1 - Y(\theta^\top X))$*. In addition, we consider the usual squared norm as a regularizer, that is* $\psi\colon \theta \in \mathbb{R}^d \mapsto \|\theta\|_2^2$*.*

Finally, the regularized risk principle states that a linear estimate of the classification function η *is* $x \in \mathbb{R}^d \mapsto \operatorname{sign}(\hat{\theta}^\top x)$*, where:*

$$\hat{\theta} \in \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \max(0, 1 - Y(\theta^\top X)) + \lambda\|\theta\|_2^2,$$

where $\lambda > 0$*.*

This estimate is called linear support vector machine.

1.1.2 Canonical optimization problem

In this manuscript, we focus on finite-dimensional and single objective problems, which are the ones mainly encountered in practice. This means that:

- the objective function f is real-valued: $\operatorname{range}(f) \subset \mathbb{R}$;
- the domain of f, denoted $\mathcal{X}$, has finite dimension: $\mathcal{X} \subset \mathbb{R}^d$ ($d \in \mathbb{N}^*$);
- the optimization problem has a finite number of constraints. Especially, it has $p \in \mathbb{N}$ inequality constraints, defined by $g_j \colon \mathbb{R}^d \to \mathbb{R}$ ($\forall j \in [p]$), and $m \in \mathbb{N}$ equality constraints, defined by $h_j \colon \mathbb{R}^d \to \mathbb{R}$ ($\forall j \in [m]$).

The canonical formulation of an optimization problem is:

$$\begin{aligned} \underset{x \in \mathcal{X}}{\text{minimize}} \ & f(x) \\ \text{s.t.} \ & \begin{cases} \forall j \in [p] \colon g_j(x) \leqslant 0 \\ \forall j \in [m] \colon h_j(x) = 0. \end{cases} \end{aligned} \tag{P1}$$

Definition 1.8 (Feasibility)**.** *Let $\mathcal{C} = \{x \in \mathcal{X} : \forall j \in [p] \colon g_j(x) \leqslant 0, \forall j \in [m] \colon h_j(x) = 0\}$.*
$\mathcal{C}$ is called the feasible set *(or the set of feasible points) of Problem (P1).*
Respectively, a point $x \in \mathcal{X}$ is said feasible to Problem (P1) if $x \in \mathcal{C}$.
Finally, Problem (P1) is said feasible if $\mathcal{C} \neq \varnothing$.

Remark 1.9. An optimization problem is defined with the keyword minimize (or similarly with maximize), emphasizing that solving such a problem consists in determining:

- a minimizer $\hat{x} \in \arg\min_{x \in \mathcal{C}} f(x)$ and/or
- the optimal (infimum) objective value $\inf_{x \in \mathcal{C}} f(x)$,

according to the problem of interest. Respectively, we can use the contraction min. (or max.)

Problem (P1) is referred to as:

- a constrained optimization problem if $p + m \geqslant 1$;
- an unconstrained optimization problem if $p + m = 0$.

Definition 1.10 (Characteristic function)**.** *Let $\mathcal{A}$ be a set and let $\mathcal{B}$ be a subset of $\mathcal{A}$. The characteristic function of $\mathcal{B}$ is the function $\chi_{\mathcal{B}} \colon \mathcal{A} \to \mathbb{R} \cup \{+\infty\}$ such that:*

$$\forall x \in \mathcal{A} \colon \chi_{\mathcal{B}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{B} \\ \infty & \text{if } x \notin \mathcal{B}. \end{cases}$$

Remark 1.11. A constrained optimization problem of the form of Problem (P1) (with $p + m \geqslant 1$) can always be turned into an unconstrained problem. Indeed, agreeing that $f(x) = \infty$ when $x \notin \mathcal{X}$,

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\ f(x) + \chi_{\mathcal{C}}(x)$$

has same minimizers and minimal objective value than Problem (P1) (we may say that both problems are equivalent, even though the notion of equivalence should be well defined).

This remark highlights that a physical problem can be translated to many different optimization problems that lead to the same solutions.

1.1.3 Numerical optimization

Often, an optimization problem in the form of Problem (P1) cannot be solved analytically (that is we cannot exhibit the set of minimizers and the optimal objective value). This is consistent with the fact that we do not know much about the estimand of interest an that we decided to express an estimator thanks to an optimization problem and not with a closed-form formula.

However, numerical strategies can produce approximation solutions of Problem (P1). In practice, since an estimate is made to be numerically evaluated, approximate solutions of optimization problems are sufficient. Thus, according to the practitioner's interest, an ϵ-approximation (or ϵ-solution) to Problem (P1) may be a point $\tilde{x} \in \mathbb{R}^d$ such that $f(\tilde{x})$ is ϵ-close to $\inf_{x \in \mathcal{C}} f(x)$ or a value $\tilde{v}$, that is ϵ-close to $\inf_{x \in \mathcal{C}} f(x)$ (see Definition 1.12). Such an ϵ-approximation can be obtained thanks to a programming implementation of an algorithm.

Definition 1.12 (ϵ-solution)**.** *Let $\epsilon > 0$.*

A point *ϵ-solution to Problem (P1) is a point $\tilde{x} \in \mathcal{C}$ such that:*

$$f(\tilde{x}) - \inf_{x \in \mathcal{C}} f(x) \leqslant \epsilon.$$

A value *ϵ-solution to Problem (P1) is a value $\tilde{v} \in \mathbb{R}$ such that:*

$$\tilde{v} - \inf_{x \in \mathcal{C}} f(x) \leqslant \epsilon.$$

For a differentiable function f, a non-convex *ϵ-solution to Problem (P1) is a point $\tilde{x} \in \mathcal{C}$ such that:*

$$\|\nabla f(\tilde{x})\|_2 \leqslant \epsilon.$$

Definition 1.13 (Algorithm)**.** *Let Θ be a set and let $\theta \in \Theta$ be a given parameter. We consider the optimization problem depending on θ:*

$$\begin{array}{ll} \underset{x \in \mathbb{R}^d}{\text{minimize}} & f(x, \theta) \\ s.t. & \begin{cases} \forall j \in [p] \colon g_j(x, \theta) \leqslant 0 \\ \forall j \in [m] \colon h_j(x, \theta) = 0, \end{cases} \end{array}$$

where $f(\cdot, \theta) \colon \mathbb{R}^d \to \mathbb{R}$, $g_j(\cdot, \theta) \colon \mathbb{R}^d \to \mathbb{R}$ and $h_j(\cdot, \theta) \colon \mathbb{R}^d \to \mathbb{R}$ are some prescribed functions.

An algorithm is a mapping $\phi \colon \Theta \times \mathbb{R}^d \times \mathbb{R}_+^* \to \mathbb{R}^2$ such that when $(\tilde{x}, \tilde{v}) = \phi(\theta, x_0, \epsilon)$ (where x_0 is an initial point), $\tilde{x}$ or $\tilde{v}$ is an ϵ-solution to the previous optimization problem.

It is noteworthy to stress that implementations and algorithms are different objects. There can be different implementations of the same algorithm according to:

- the programming language used (R, Python, C++);
- the base routines (for linear algebra);
- the programming technique (loops, vectorization);
- the potential compiler (GCC, Borland C++, C++ builder, . . .);
- the programmer optimizations (code factorization).

As a consequence, we cannot compare *algorithms* with numerical simulations (but only implementations). However, algorithms can be assessed with an *oracle complexity*. Even though we will not cover that topic in this manuscript, we give the two basics definitions.

Definition 1.14 (Oracles)**.** *Let us consider Problem (P1).*
The zeroth-order oracle is $O\colon x \in \mathbb{R}^d \mapsto f(x)$*.*
If f *is differentiable, the first-order oracle is* $O\colon x \in \mathbb{R}^d \mapsto (f(x), \nabla f(x))$*.*
If f *is twice differentiable, the second-order oracle is* $O\colon x \in \mathbb{R}^d \mapsto (f(x), \nabla f(x), \nabla^2 f(x))$*.*

Definition 1.15 (Oracle complexity)**.** *Given a class of optimization problems defined by:*

- *a class of objective functions (with prescribed regularity conditions);*
- *a condition on the initial point* x_0*;*
- *an oracle* O*,*

the oracle complexity of an algorithm ϕ *is the minimal number of calls to the oracle* O *it has to perform in order to produce an* ϵ*-solution for any parameters* θ *from* Θ*, for all objective functions and any initial points* x_0 *(this is a worst-case complexity).*

Remark 1.16. The computation time of an implementation of an algorithm could be estimated with the *arithmetic complexity* of the algorithm. This one counts the total number of arithmetic operations to perform in order to produce an ϵ-solution, in the worst case. However, the arithmetic complexity of an algorithm is a biased estimate of the computation time of an implementation since it does not consider programming optimizations (parallelization, compiler, . . .). Moreover, it is much harder to prove bounds on the arithmetic complexity than on the oracle complexity.

1.2 Convex analysis

1.2.1 Convex sets

Definition 1.17 (Convex set)**.** *A set $K \subset \mathbb{R}^d$ is said convex if:*

$$\forall (x,y) \in K^2, \forall t \in (0,1) \colon tx + (1-t)y \in K.$$

Example 1.18. *Norm balls, vector spaces, affine subspaces, half spaces are convex sets. In addition, any intersection of convex sets is again a convex set.*

Exercise 1.19. *Show the previous statements.*

Proposition 1.20 (Finite combination for convex sets)**.** *A set $K \subset \mathbb{R}^d$ is convex if and only if: $\forall n \geqslant 2, \forall (x_i)_{1 \leqslant i \leqslant n} \in K^n, \forall (t_i)_{1 \leqslant i \leqslant n}$ such that $t_i \geqslant 0, \forall i \in [n]$ and $\sum_{i=1}^n t_i = 1$,*

$$\sum_{i=1}^n t_i x_i \in K.$$

Proof. By induction, remarking that when $\sum_{i=1}^n t_i \neq 0$:

$$\sum_{i=1}^{n+1} t_i x_i = \left(\sum_{i=1}^n t_i \right) \left(\sum_{i=1}^n \frac{t_i}{\sum_{i=1}^n t_i} x_i \right) + t_{n+1} x_{n+1}.$$

□

Definition 1.21 (Convex hull)**.** *The convex hull of a set $K \subset \mathbb{R}^d$, denoted* $\operatorname{conv}(K)$*, is the smallest convex set containing K.*

Theorem 1.22 (Finite combination for convex hull)**.** *Let $K \subset \mathbb{R}^d$.*

$$\operatorname{conv}(K) = \left\{ \sum_{i=1}^n t_i x_i : n \in \mathbb{N}, \forall (x_i)_{1 \leqslant i \leqslant n} \in K^n, t_i \geqslant 0, \forall i \in [n], \sum_{i=1}^n t_i = 1 \right\}$$

Proof. Since $\operatorname{conv}(K)$ is convex and $(x_i)_{1 \leqslant i \leqslant n} \in K^n \subset \operatorname{conv}(K)^n$, we have $\supset$.

In addition, the rhs is a convex set containing K. Since $\operatorname{conv}(K)$ is the smallest, convex set containing K, we have $\subset$. □

Definition 1.23 (Cone)**.** *A set $K \subset \mathbb{R}^d$ is a cone if:*

$$\forall x \in K, \forall t \in \mathbb{R}_+ \colon tx \in K.$$

If K is convex, then K is called a convex cone.

K is said proper if:

1. *K is closed;*

2. K is pointed: $K \cap (-K) = \{0\}$;

3. K has non-empty interior: $\mathrm{int}(K) \neq \varnothing$.

Proposition 1.24 (Characterization of a convex cone)**.** *A set $K \subset \mathbb{R}^d$ is a convex cone if and only if:*

$$\forall (x, y) \in K^2, \forall (s, t) \in \mathbb{R}_+^2 : sx + ty \in K.$$

Proof. By conicity, $2sx \in K$ and $2ty \in K$. So, by convexity, $\frac{1}{2}2sx + \frac{1}{2}2ty \in K$.

Conversely, the property is true for t $= 0$, and that is the definition of a cone. It is convex with $t = 1 - s$ and $s < 1$. □

Example 1.25. *Half spaces and the positive orthant $\mathbb{R}_+^d$ are convex cones.*

The set of positive semidefinite matrices on $\mathbb{R}^{d \times d}$ is a convex cone.

The second order cone,

$$\left\{(x, t) \in \mathbb{R}^d \times \mathbb{R} : \|x\|_2 \leqslant t\right\} \subset \mathbb{R}^{d+1},$$

is a convex cone.

Definition 1.26 (Dual cone)**.** *Let $K \subset \mathbb{R}^d$ be a cone. Its dual cone K^* is defined by*

$$K^* = \left\{y \in \mathbb{R}^d : x^\top y \geqslant 0, \forall x \in K\right\}.$$

Proposition 1.27. *Let $K \subset \mathbb{R}^d$ be a cone. Then*

- *K^* is a cone;*
- *K^* is closed and convex.*

Definition 1.28 (Extreme point)**.** *Let $K \subset \mathbb{R}^d$ be a convex set. A point $x \in K$ is called an extreme point of K if:*

$$\forall (y, z) \in K^2, \forall t \in (0, 1) : \quad x = ty + (1 - t)z \implies x = y = z.$$

Definition 1.29 (Compact space)**.** *A set $K \subset \mathbb{R}^d$ is compact if it is closed and bounded (Heine-Borel theorem).*

Theorem 1.30 (Compact convex set)**.** *A compact convex subset of $\mathbb{R}^d$ is the convex hull of its extreme points.*

Proof. $\subset$ is proved by induction on the dimension with the notion of relative boundary. □

1.2.2 Convex functions

Definition 1.31 (Extended-valued function). *A mapping $F\colon \mathbb{R}^d \to [-\infty, \infty] = \mathbb{R} \cup \{-\infty, \infty\}$ is called an extended-valued function on $\mathcal{X}$.*

The domain of such a function is defined as:

$$\operatorname{dom}(F) = \{x \in \mathbb{R}^d : F(x) < \infty\}.$$

A function $F\colon \mathbb{R}^d \to [-\infty, \infty]$ is said proper if $\operatorname{dom}(F) \neq \varnothing$ *and $F(x) > -\infty, \forall x \in \mathbb{R}^d$.*

Let $\mathcal{X} \subset \mathbb{R}^d$ be a set and $f\colon \mathcal{X} \to \mathbb{R}$ be a function. The canonical extension of f to an extended-valued function F is:

$$\forall x \in \mathbb{R}^d\colon F(x) = \begin{cases} f(x) & \textit{if } x \in \mathcal{X} \\ \infty & \textit{if } x \notin \mathcal{X}. \end{cases}$$

Then, by definition, $\operatorname{dom}(F) = \mathcal{X}$.

Definition 1.32 (Epigraph). *Let $f\colon \mathbb{R}^d \to [-\infty, \infty]$. The epigraph of f is defined as:*

$$\operatorname{epi}(f) = \{(x, \mu) \in \mathbb{R}^d \times \mathbb{R} : f(x) \leqslant \mu\} \subset \mathbb{R}^{d+1}.$$

Remark 1.33. Epigraph and domain of an extended-valued function f are related:

$$\operatorname{dom}(f) = \{x \in \mathbb{R}^d : \exists \mu \in \mathbb{R}, (x, \mu) \in \operatorname{epi}(f)\}.$$

Definition 1.34 (Convex function). *An extended-valued function is convex if its epigraph is a convex set of $\mathbb{R}^{d+1}$.*

Respectively, a function is convex if its canonical extension to an extended-valued function is convex.

Proposition 1.35 (Characterization of a convex function). *A function $f\colon \mathbb{R}^d \to (-\infty, \infty]$ is convex if and only if*

$$\forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d, \forall t \in (0,1)\colon \quad f(tx + (1-t)y) \leqslant tf(x) + (1-t)f(y).$$

Proof. First, let us assume that the epigraph of f is convex. For all $x, y \in \mathbb{R}^d$ and $t \in (0,1)$,

- if $(x, y) \in \operatorname{dom}(f)$, then $f(x), f(y) < \infty$, so $(x, f(x)) \in \operatorname{epi}(f)$ and $(y, f(y)) \in \operatorname{epi}(f)$. Since $\operatorname{epi}(f)$ is convex, $t(x, f(x)) + (1-t)(y, f(y)) = (tx + (1-t)y, tf(x) + (1-t)f(y)) \in \operatorname{epi}(f)$; in other words, $f(tx + (1-t)y) \leqslant tf(x) + (1-t)f(y)$.

- assume (without loss of generality) that $x \in \operatorname{dom}(f)$ and $y \notin \operatorname{dom}(f)$. Then $tf(x) + (1-t)f(y) = \infty$ (since $1 - t \neq 0$), so $f(tx + (1-t)y) \leqslant tf(x) + (1-t)f(y)$.

Second, let us assume that $f(tx + (1-t)y) \leqslant tf(x) + (1-t)f(y)$, $\forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d, \forall t \in (0,1)$. Then for any (x, μ) and (x', μ') in the epigraph of f and for any $t \in (0,1)$, $f(tx + (1-t)x') \leqslant tf(x) + (1-t)f(y) \leqslant t\mu + (1-t)\mu'$. Thus, $t(x, \mu) + (1-t)(x', \mu') = (tx + (1-t)x', t\mu + (1-t)\mu') \in \operatorname{epi}(f)$; that is, *epi*$(f)$ is convex. □

Remark 1.36. If one allows f to have values $-\infty$ (*i.e.* $f\colon \mathbb{R}^d \to [-\infty, \infty]$), then f is convex *if and only if*

$$\forall (x,y) \in \mathbb{R}^d \times \mathbb{R}^d, \forall t \in (0,1): \quad f(tx + (1-t)y) \leqslant t\alpha + (1-t)\beta,$$

$\forall (\alpha, \beta) \in [-\infty, \infty]^2 : f(x) < \alpha, f(y) < \beta.$

Remark 1.37. A function $f\colon \mathbb{R}^d \to (-\infty, \infty]$ is said strictly convex if

$$\forall (x,y) \in \mathbb{R}^d \times \mathbb{R}^d, x \neq y, \forall t \in (0,1): \quad f(tx + (1-t)y) < tf(x) + (1-t)f(y).$$

Proposition 1.38 (Elementary properties)**.** *We consider two convex functions* $f_1\colon \mathbb{R}^d \to (-\infty, \infty]$ *and* $f_2\colon \mathbb{R}^d \to (-\infty, \infty]$.

1. *The set* $\operatorname{dom}(f_1)$ *is convex.*

2. *For any non-negative* α *and* β, $\alpha f_1 + \beta f_2$ *is convex.*

3. $x \mapsto \max(f_1(x), f_2(x))$ *is convex.*

4. *Let* $(A, b) \in \mathbb{R}^{d \times d'} \times \mathbb{R}^d$, *then* $x \in \mathbb{R}^{d'} \mapsto f(Ax + b)$ *is convex.*

5. *For any* $(x, y) \in \operatorname{dom}(f_1)$ *and* $t \geqslant 1$, *denoting* $z_t = x + t(y - x)$, $f_1(z_t) \geqslant f_1(x) + t(f_1(y) - f_1(x))$.

6. *Let* $\varphi\colon \mathbb{R} \to \mathbb{R}$ *be convex and nondecreasing and* $f\colon \mathbb{R}^d \to \mathbb{R}$ *be a convex function, then* $\varphi \circ f$ *is convex.*

7. *The perspective function of* f_1*:*

$$g\colon (x,t) \in \mathbb{R}^d \times \mathbb{R} \mapsto \begin{cases} tf_1(\frac{1}{t}x) & \text{if } t > 0 \\ \infty & \text{otherwise,} \end{cases}$$

is convex.

Exercise 1.39. *Show the previous statements.*

Proof. For Point 5, let us remark that $y = \mu x + (1-\mu) z_t$, where $\mu = \frac{t-1}{t} \in [0,1]$.
For Point 7, make the coefficients inside f summable to 1. □

Theorem 1.40 (Jensen's inequality (finite form))**.** *A function* $f\colon \mathbb{R}^d \to (-\infty, \infty]$ *is convex if and only if:* $\forall n \geqslant 2, \forall (x_i)_{1 \leqslant i \leqslant n} \in (\mathbb{R}^d)^n, \forall (t_i)_{1 \leqslant i \leqslant n}$ *such that* $t_i \geqslant 0, \forall i \in [n]$ *and* $\sum_{i=1}^n t_i = 1$,

$$f\left(\sum_{i=1}^n t_i x_i\right) \leqslant \sum_{i=1}^n t_i f(x_i).$$

Proof. By induction for $\implies$ and by definition otherwise. □

Theorem 1.41 (Jensen's inequality (probabilistic form))**.** *If $f\colon \mathbb{R}^d \to (-\infty, \infty]$ is convex, then for any random vector $X \in \mathbb{R}^d$:*

$$f\left(\mathbb{E}X\right) \leqslant \mathbb{E}f(X).$$

Example 1.42 (Convex functions)**.**

1. *Every norm $\|\cdot\|$ on $\mathbb{R}^d$ is convex (this comes from the triangle inequality and homogeneity).*

2. *ℓ_p-norms for $1 < p < \infty$ are strictly convex and only convex for $p = 1$ and $p = \infty$.*

3. *For $\varphi\colon \mathbb{R} \to \mathbb{R}$ convex nondecreasing and every norm $\|\cdot\|$ on $\mathbb{R}^d$, $\varphi(\|\cdot\|)$ is convex. In particular, $\|\cdot\|^p$ is convex provided that $p \geqslant 1$.*

4. *For a positive semidefinite matrix $A \in \mathbb{R}^{d\times d}$, $f\colon x \in \mathbb{R}^d \to x^\top Ax$ is convex. If A is positive definite, f is strictly convex.*

5. *For any convex subset $\mathcal{A}$ of $\mathbb{R}^d$, $\chi_{\mathcal{A}}$ is convex.*

Exercise 1.43. *Show the last statement.*

Proposition 1.44 (Coordinate supremum)**.** *Let $\mathcal{Y} \subset \mathbb{R}^{d'}$ (potentially nonconvex set) and $F\colon (x, y) \in \mathbb{R}^d \times \mathcal{Y} \to (-\infty, \infty]$ be a function convex in x (that is, $\forall y \in \mathcal{Y}, F(\cdot, y)$ is convex). Then $f\colon x \in \mathbb{R}^d \mapsto \sup_{y\in\mathcal{Y}} F(x, y)$ is convex.*

Proof. Remark that $\operatorname{epi}(f) = \cap_{y\in\mathcal{Y}} \operatorname{epi}(F(\cdot, y))$, which is the intersection of convex sets. □

Remark 1.45. With the definitions of the previous proposition, f is sometimes called the upper hull of the family of convex functions $(F(\cdot, y))_{y\in\mathcal{Y}}$.

Proposition 1.46 (Coordinate infimum)**.** *Let $F\colon (x, y) \in \mathbb{R}^d \times \mathbb{R}^{d'} \to (-\infty, \infty]$ be a (jointly) convex function. Then $f\colon x \in \mathbb{R}^d \mapsto \inf_{y\in\mathbb{R}^{d'}} F(x, y)$ is convex.*

Proof.

$$\begin{aligned}
&\forall (x, x', t) \in \mathbb{R}^d \times \mathbb{R}^f \times (0, 1):\\
&\qquad f(tx + (1-t)x') = \inf_{y\in\mathbb{R}^{d'}} F(tx + (1-t)x', y)\\
&\qquad\qquad = \inf_{y\in\mathbb{R}^{d'}, y'\in\mathbb{R}^d} F(tx + (1-t)x', ty + (1-t)y')\\
&\qquad\qquad \leqslant \inf_{y\in\mathbb{R}^{d'}} tF(x, y) + \inf_{y\in\mathbb{R}^{d'}} (1-t)F(x', y')\\
&\qquad\qquad = tf(x) + (1-t)f(x').
\end{aligned}$$

□

Definition 1.47 (Strong convexity)**.** *Let $\mu \in \mathbb{R}_+^*$. A function $f\colon \mathbb{R}^d \to [-\infty, \infty]$ is μ-strongly convex if $f - \frac{\mu}{2}\|\cdot\|_2^2$ is convex.*

Proposition 1.48 (Characterization of a strongly convex function)**.** *Let $\mu \in \mathbb{R}_+^*$. A function $f\colon \mathbb{R}^d \to (-\infty, \infty]$ is μ-strongly convex if and only if*

$$\forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d, \forall t \in (0, 1)\colon \quad f(tx + (1-t)y) \leqslant tf(x) + (1-t)f(y) - \frac{\mu}{2}t(1-t)\|x - y\|_2^2.$$

Proposition 1.49 (Relation between convexities)**.** *Let $f\colon \mathbb{R}^d \to (-\infty, \infty]$ be a function.*

$$f \text{ strongly convex} \implies f \text{ strictly convex} \implies f \text{ convex.}$$

Proposition 1.50 (First-order conditions of convexity)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be a convex set and $f\colon \mathcal{X} \to \mathbb{R}$ be a differentiable function.*

f is convex if and only if:

$$\forall (x, y) \in \mathcal{X}^2\colon \quad f(y) \geqslant f(x) + \nabla f(x)^\top (y - x).$$

f is convex if and only if:

$$\forall (x, y) \in \mathcal{X}^2\colon \quad (\nabla f(y) - \nabla f(x))^\top (y - x) \geqslant 0.$$

f is strictly convex if and only if:

$$\forall (x, y) \in \mathcal{X}^2, x \neq y\colon \quad f(y) > f(x) + \nabla f(x)^\top (y - x).$$

Let $\mu \in \mathbb{R}_+^*$. f is μ-strongly convex if and only if:

$$\forall (x, y) \in \mathcal{X}^2\colon \quad f(y) \geqslant f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|_2^2.$$

Proof. For the first inequality, if f is convex, then:

$$\begin{aligned}
\nabla f(x)^\top (y - x) &= \lim_{t \to 0} \frac{f(x + t(y - x)) - f(x)}{t} \\
&= \lim_{t \to 0_+, t < 1} \frac{f((1-t)x + ty)) - f(x)}{t} \\
&\leqslant \lim_{t \to 0_+, t < 1} \frac{(1-t)f(x) + tf(y) - f(x)}{t} \\
&= f(y) - f(x).
\end{aligned}$$

Conversely, if $\forall (x, y) \in \mathcal{X}^2\colon f(y) \geqslant f(x) + \nabla f(x)^\top (y - x)$, then for any $t \in (0, 1)$, denoting $z = tx + (1-t)y$:

$$f(y) \geqslant f(z) + \nabla f(z)^\top (y - z)$$

and

$$f(x) \geqslant f(z) + \nabla f(z)^\top (x - z).$$

Combining the two inequalities gives:

$$tf(x) + (1-t)f(y) \geqslant f(z) + \nabla f(z)^\top (tx + (1-t)y - z) = f(z),$$

which proves convexity.

Concerning the second point, if f is convex, then $\forall(x,y) \in \mathcal{X}^2\colon f(y) \geqslant f(x)+\nabla f(x)^\top(y-x)$. So, for fixed (x,y), $f(y) \geqslant f(x) + \nabla f(x)^\top(y-x)$ and $f(x) \geqslant f(y) + \nabla f(y)^\top(x-y)$. Combining the two gives $(\nabla f(y)-\nabla f(x))^\top(y-x) \geqslant 0$. On the other hand, for any (x,y), let $g\colon t \in (0,1) \mapsto f(x+t(y-x))$. If ∇f is monotone, then $t(g'(t)-g'(0)) \geqslant 0$, that is $g'(t) \geqslant g'(0)$. In addition, $f(y) = g(1) = g(0) + \int_0^1 g'(t)\,dt \geqslant g(0) + g'(0) = f(x) + \nabla f(x)^\top(y-x)$. Thus f is convex.

Concerning the third point, if the inequality is satisfied, then f is strictly convex. On the other hand, if f is strictly convex, then $\forall(x,y) \in \mathcal{X}^2\colon f(y) \geqslant f(x) + \nabla f(x)^\top(y-x)$. Moreover, if $\exists(x,y) \in \mathcal{X}^2\colon f(y) = f(x) + \nabla f(x)^\top(y-x)$, then $\forall t \in (0,1)$

$$f(tx+(1-t)y) \leqslant tf(x) + (1-t)f(y) = f(x) + (1-t)\nabla f(x)^\top(y-x).$$

In addition,

$$f(tx+(1-t)y) \geqslant f(x) + \nabla f(x)^\top(tx+(1-t)y-x) = f(x) + (1-t)\nabla f(x)^\top(y-x).$$

So $f(tx+(1-t)y) = f(x) + (1-t)\nabla f(x)^\top(y-x)$, which is in contradiction with strict convexity. Thus, $\forall(x,y) \in \mathcal{X}^2\colon f(y) > f(x) + \nabla f(x)^\top(y-x)$. □

Remark 1.51. For convex functions,

$$\forall(x,y) \in \mathcal{X}^2\colon \quad f(y) \geqslant f(x) + \nabla f(x)^\top(y-x).$$

This is perhaps the most important property of convex functions since it shows that from a local information ($\nabla f(x)$), we can derive a global information concerning f (we have a global underestimator). In particular, if $\nabla f(x) = 0$, then x is a global minimizer.

Proposition 1.52 (Second-order conditions of convexity)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ and $f\colon \mathcal{X} \to \mathbb{R}$ be a twice differentiable function.*

f is convex if and only if:

$$\forall x \in \mathcal{X}\colon \quad \nabla^2 f(x) \succcurlyeq 0.$$

f is strictly convex if and only if:

$$\forall x \in \mathcal{X}\colon \quad \nabla^2 f(x) \succ 0.$$

Let $\mu \in \mathbb{R}_+^*$ and $I_d \in \mathbb{R}^{d\times d}$ be the identity matrix. f is μ-strongly convex if and only if:

$$\forall x \in \mathcal{X}\colon \quad \nabla^2 f(x) \succcurlyeq \mu I_d.$$

1.2.3 Properties of minimizers

In this section, we consider a proper extended-valued function $f\colon \mathbb{R}^d \to [-\infty,\infty]$ along with the optimization problem

$$\underset{x\in\mathbb{R}^d}{\operatorname{minimize}}\ f(x). \tag{P2}$$

Definition 1.53 (Minimizers)**.** *A point $x^\star \in \mathbb{R}^d$ is a global minimizer of Problem (P2) if $x^\star \in \operatorname{dom}(f)$ and*

$$\forall x \in \mathbb{R}^d\colon\, f(x^\star) \leqslant f(x).$$

A point $x^\star \in \mathbb{R}^d$ is a local minimizer of Problem (P2) if there exists a neighborhood $\mathcal{N}$ of $x^\star$ such that $x^\star$ is a global minimizer of Problem (P2) in this neighborhood. In other words, there exists $\epsilon > 0$, $\mathcal{N} = \{x \in \mathbb{R}^d : \|x^\star - x\| \leqslant \epsilon\}$ (for a given norm $\|\cdot\|$) such that $x^\star$ is a global minimizer of

$$\underset{x \in \mathbb{R}^d}{\operatorname{minimize}}\ f(x) + \chi_{\mathcal{N}}(x).$$

Remark 1.54. Global minimizers may not exist, as we can see for:

- $f(x) = \frac{1}{x^2}$ (agreeing that $f(0) = \infty$);
- $f(x) = \exp(-x^2)$;
- $f(x) = x + \chi_{\mathbb{R}_+^*}(x)$.

As we can see, the existence of minimizers of Problem (P2) is not guaranteed, even though f is proper (that is, there exists at least a point $x \in \mathbb{R}^d$ such that $-\infty < f(x) < \infty$. Consequently, the remaining of this section is devoted to characterize the existence of minimizers and their properties.

Definition 1.55 (Lower limit)**.** *The lower limit (or limit inferior) of a sequence $(u_n)_{n\in\mathbb{N}}$, where $u_n \in [-\infty, \infty]$ is*

$$\liminf_{n\to\infty} u_n = \lim_{n\to\infty} \inf_{k\geqslant n} u_k.$$

Let us remark that the sequence $(\inf_{k\geqslant n} u_k)_{n\in\mathbb{N}}$ is nondecreasing, thus the limit is well defined in $\mathbb{R} \cup \{-\infty, \infty\}$ and $\liminf_{n\to\infty} u_n = \sup_{n\in\mathbb{N}} \inf_{k\geqslant n} u_k$.

Definition 1.56 (Lower semi-continuity)**.** *Let $g\colon \mathbb{R}^d \to [-\infty, \infty]$. g is lower semi-continuous at $x \in \mathbb{R}^d$ if for every sequence $(x_n)_{n\in\mathbb{N}}$ converging to x,*

$$g(x) \leqslant \liminf_{n\to\infty} g(x_n).$$

g is lower semi-continuous if it is lower semi-continuous at every $x \in \mathbb{R}^d$.

Example 1.57 (Lower semi-continuous functions)**.**

1. *Every continuous function is lower semi-continuous.*
2. *$x \in \mathbb{R} \mapsto x^2 - \mathbb{1}_{R_-}(x)$ is lower semi-continuous.*
3. *$x \in \mathbb{R} \mapsto \mathbb{1}_{\mathbb{R}_+^*} - \mathbb{1}_{R_-^*}(x)$ is not lower semi-continuous.*

Proposition 1.58 (Epigraphs)**.** *A function $g\colon \mathbb{R}^d \to [-\infty, \infty]$ is lower semi-continuous if and only if its epigraph is closed.*

Proposition 1.59 (Lower level sets)**.** *A function $g\colon \mathbb{R}^d \to [-\infty, \infty]$ is lower semi-continuous if and only if for any $\alpha \in \mathbb{R}$, the lower level set $\{x \in \mathbb{R}^d : g(x) \leqslant \alpha\}$ is closed.*

Proposition 1.60 (Lower semi-continuity of upper hulls)**.** *Let $\mathcal{Y} \subset \mathbb{R}^{d'}$ (potentially non-convex set) and $G\colon (x, y) \in \mathbb{R}^d \times \mathcal{Y} \to (-\infty, \infty]$ be a function continuous in x (that is, $\forall y \in \mathcal{Y}, G(\cdot, y)$ is continuous). Then $g\colon x \in \mathbb{R}^d \mapsto \sup_{y\in\mathcal{Y}} G(x, y)$ is lower semi-continuous.*

This last property will be useful for min-max problems, which will appear in duality theory.

Definition 1.61 (Minimizing sequence)**.** *Assume that f is proper. A minimizing sequence for Problem (P2) is a sequence $(x_n)_{n\in\mathbb{N}}$, with $x_n \in \operatorname{dom}(f)$, such that*

$$\lim_{n\to\infty} f(x_n) = \inf_{x\in\mathbb{R}^d} f(x).$$

Remark 1.62. By definition of the infimum value of f, there always exists a minimizing sequence: $\forall n \in \mathbb{N}^*$, the set $\{x \in \operatorname{dom}(f) : f(x) - \inf_{y\in\operatorname{dom}(f)} f(y) \leqslant \frac{1}{n}\}$ is non-empty (otherwise $\inf_{y\in\operatorname{dom}(f)} f(y)$ is not an infimum), so we can set x_n to be any element of this set.

With this last definition, we have gathered the necessary components to claim the existence of a solution of a constrained optimization problem.

Theorem 1.63 (Existence of a solution for constrained problems)**.** *Let $\mathcal{C} \subset \mathbb{R}^d$ be a non-empty compact set and assume that f is proper, lower semi-continuous and of the form $f = g + \chi_{\mathcal{C}}$, where $g\colon \mathbb{R}^d \to [-\infty, \infty]$. Then Problem (P2) admits a global minimizer.*

Proof. We have that $\operatorname{dom}(f) \subset \mathcal{C}$. Let $(x_n)_n$ be a minimizing sequence for f. Because $\mathcal{C}$ is closed and bounded, it follows that the sequence $(x_n)_n$ admits a sub-sequence, say $(x'_n)_n$, with $x'_n \in \mathcal{C}$, converging to some point $x^\star \in \mathcal{C}$ (from Heine-Borel Theorem). Thus we have:

$$\inf_{x\in\mathbb{R}^d} f(x) = \lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} f(x'_n) = \liminf_{n\to\infty} f(x'_n) \geqslant f(x^\star),$$

which shows that $x^\star$ is a global minimizer. □

Let us highlight the similarity between this theorem and the Weierstrass extreme value theorem, which states that every continuous function on a compact set attains both a minimum and a maximum. Here, because we consider extended-valued functions and we are interested only in finding a minimum, we relax the assumption of continuity to the one of lower semi-continuity. Doing so, we lose the existence of a maximizer but provide a broader result for mathematical optimization.

Now, we state an existence theorem for unconstrained optimization problems. For this purpose, we introduce another definition first.

Definition 1.64 (Coercivity)**.** *A function $g\colon \mathbb{R}^d \to [-\infty, \infty]$ is coercive if for every sequence $(x_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} \|x_n\| = \infty$,*

$$\lim_{n\to\infty} g(x_n) = \infty.$$

Theorem 1.65 (Existence of a solution for unconstrained problems)**.** *Assume that f is proper, coercive and lower semi-continuous. Then Problem (P2) admits a global minimizer.*

Proof. The proof is similar to the one of Theorem 1.63, remarking that any minimizing sequence is necessarily bounded since $\inf_{x\in\mathbb{R}^d} f(x) < \infty$ (f proper) and f is coercive. □

Knowing that lower semi-continuous functions can attain their minima, let us go back to convex functions.

Proposition 1.66 (Minimizers of convex functions)**.** *Assume that f is convex. Then*

a) a local minimizer of f is a global one;

b) the set of minimizers of f is convex;

c) if f is strictly convex, then f has a unique minimizer.

Proof. a) For any point $x \in \mathbb{R}^d$ and a local minimizer $x^\star \in \mathbb{R}^d$, there exists $t \in (0, 1)$ such that $z = tx^\star + (1-t)x$ is in the neighborhood. So $f(x^\star) \leqslant f(z)$ and by convexity, $f(z) \leqslant t(f(x^\star) + (1-t)f(x)$. It follows that $(1-t)f(x^\star) \leqslant (1-t)f(x)$.

b) Trivial.

c) By contradiction.

□

Remark 1.67. When an estimator is built as a minimizer of an optimization problem, we are interested in a global minimizer. However, in order to verify that a point $x^\star$ is a global minimizer, one would have to compare $f(x^\star)$ to every other value $f(x)$, no matter how far from $x^\star$ x is. The fact that for convex functions, local minimizers are also global minimizers essentially explains our interest in convex optimization and the availability of efficient numerical methods. Indeed, local minimizers can be found by greedy approaches.

1.2.4 Optimality conditions

Differentiability plays a key role in optimization. First because it helps characterizing convexity (see Proposition 1.50), second (this is a consequence) because it is inherent in the mainly used optimality condition (the Fermat's rule). In this section, we introduce a generalization of the gradient to nondifferentiable functions.

Definition 1.68 (Subdifferential)**.** *Let $f\colon \mathbb{R}^d \to (-\infty, \infty]$ be a convex function. The subdifferential of f at $x \in \mathbb{R}^d$ is defined by*

$$\partial f(x) = \{v \in \mathbb{R}^d : \forall y \in \mathbb{R}^d, f(y) \geqslant f(x) + v^\top(y-x)\}.$$

The elements of $\partial f(x)$ are called the subgradients of f at x.

Example 1.69. *$f\colon x \in \mathbb{R} \mapsto |x|$ has a subdifferential for all x and*

$$\partial f(x) = \begin{cases} \{-1\} & \text{if } x < 0 \\ [-1,1] & \text{if } x = 0 \\ \{1\} & \text{if } x > 0. \end{cases}$$

Proposition 1.70 (Calculus of subgradients)**.** *Let $f\colon \mathbb{R}^d \to (-\infty, \infty]$ be a convex function and $x \in \mathbb{R}^d$.*

a) $\forall \alpha \geqslant 0, \partial(\alpha f)(x) = \alpha \partial f(x)$.

b) If $f = \sum_{i=1}^p f_i$, with f_i convex, $\mathrm{dom}(f_i) = \mathbb{R}^d$, then $\partial f(x) = \sum_{i=1}^p \partial f_i(x)$ (Minkowski sum).

c) If $f\colon y \mapsto \max_{1 \leqslant i \leqslant p} f_i(y)$, with f_i convex, then $\partial f(x) = \mathrm{conv}\left(\cup_{\substack{1\leqslant i \leqslant p \\ f_i(x) = f(x)}} \partial f_i(x)\right)$.

Example 1.71. *Consider $f = \|\cdot\|_1$ and remark that $\forall x \in \mathbb{R}^d\colon f(x) = \max\{s^\top x : s \in \{\pm 1\}^d\}$. Thus, for $x \in \mathbb{R}^d$, the max is achieved for $s \in \{\pm 1\}^d$ such that $s_i = 1$ if $x_i > 0$, $s_i = -1$ if $x_i < 0$ and $s_i = \pm 1$ for $x_i = 0$. As a consequence, $\partial f(x)$ is the convex hull of all such points s, that is:*

$$\begin{aligned} \partial f(x) &= \mathrm{conv}\left(\left\{s \in \{\pm 1\}^d : s^\top x = \|x\|_1\right\}\right) \\ &= \left\{ts + (1-t)s' : s, s' \in \{\pm 1\}^d, s^\top x = \|x\|_1, s'^\top x = \|x\|_1, t \in [0,1]\right\} \\ &= \left\{v \in \mathbb{R}^d : \|v\|_\infty \leqslant 1, v^\top x = \|x\|_1\right\}. \end{aligned}$$

Proposition 1.72 (Subgradient of differentiable functions)**.** *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a convex and differentiable function at $x \in \mathbb{R}^d$. Then $\partial f(x) = \{\nabla f(x)\}$.*

Proof. Choose a subgradient v and apply the inequality to $y = x + t(v - \nabla f(x))$. □

Proposition 1.73 (Subgradient of the sum of two functions)**.** *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a convex and differentiable function at $x \in \mathbb{R}^d$ and $g\colon \mathbb{R}^d \to \mathbb{R}$ be a convex function. Then $\partial(f+g)(x) = \{\nabla f(x)\} + \partial g(x)$.*

If the subdifferential of a differential function is a singleton, it is not obvious for a subdifferential (in general) to be non-empty. To analyze this property, we need the notion of relative interior of a set $\mathcal{C}$, which is the interior of $\mathcal{C}$ (the points that are not on the border of $\mathcal{C}$), relatively to the smallest subspace that contains $\mathcal{C}$ (the reader may think of the facet of a cube, for which the interior is empty but the relative interior is not). Here, we give a weak definition, restricted to convex sets.

Definition 1.74 (Relative interior)**.** *Let $\mathcal{C} \subset \mathbb{R}^d$ be a convex set. The relative interior of $\mathcal{C}$ is*

$$\mathrm{relint}(\mathcal{C}) = \{x \in \mathcal{C} : \forall y \in \mathcal{C}, \exists \lambda > 1 : y + \lambda(x - y) \in \mathcal{C}\}.$$

In other words, in any direction from $x \in$ relint$(\mathcal{C})$, there is always a point ahead of x which lies in relint$(\mathcal{C})$.

Remark 1.75. The relative interior of a convex set $\mathcal{C} \neq \varnothing$ is never empty. If $\mathcal{C}$ is a singleton, then $\operatorname{relint}(\mathcal{C}) = \mathcal{C}$.

Proposition 1.76. *Let $f\colon \mathbb{R}^d \to [-\infty, \infty]$ be a convex function and $x \in \operatorname{relint}(\operatorname{dom}(f))$ (which is well defined since $\operatorname{dom}(f)$ is convex). Then $\partial f(x)$ is non-empty.*

Exercise 1.77. *Find the subdifferentials of*

1. *$x \in \mathbb{R} \mapsto \chi_{[0,1]}(x)$ everywhere;*
2. *$x \in \mathbb{R}^2 \mapsto \chi_{\{y\in\mathbb{R}^2 : \|y\|_2 \leqslant 1\}}(x)$ at $\|x\|_2 = 1$ and $\|x\|_2 < 1$;*
3. *$x \in \mathbb{R}^d \mapsto \|x\|_2$ everywhere;*
4. *$x \in \mathbb{R} \mapsto x^3$ everywhere.*

With these elements, we can state the main optimality condition used in convex optimization. This one underlies efficient minimization algorithms such as proximal gradient descent.

Theorem 1.78 (Fermat's rule)**.** *Let $f\colon \mathbb{R}^d \to (-\infty, \infty]$ be a convex function. $x^\star \in \mathbb{R}^d$ is a global minimizer of f if and only if*

$$0 \in \partial f(x^\star).$$

Proof. See the definition of the subdifferential. □

1.2.5 Convex optimization problems

In the previous section, convex optimization has been presented with extended-valued function, and thus as always unconstrained optimization. Even though, extended-valued functions are useful tools to formalize optimization, it is often pleasant for the reader, as well as necessary for the numerical practitioner, to rewrite a problem into its canonical formulation (P1). As a reminder, this is

$$\begin{array}{ll} \underset{x\in\mathcal{X}}{\text{minimize}} & f(x) \\ \text{s.t.} & \begin{cases} \forall j \in [p]\colon g_j(x) \leqslant 0 \\ \forall j \in [m]\colon h_j(x) = 0, \end{cases} \end{array}$$

where we assume here that f, g_j and h_j are real-valued functions over $\mathcal{X}$. There is obviously an ambiguity in this formulation since it is not unique. However, the canonical formulation assumes that $\mathcal{X}$ is as large as possible (it defines the subset of $\mathbb{R}^d$ where f can be evaluated) and that g_j and h_j are defined according to the physical problem.

This writing underlines the notion of equivalence between two optimization problems. Without giving a formal definition (that would certainly not be accepted by every one), we say that two optimization problems are equivalent if from the minimizers of one, the minimizers of the other are readily found, and vice versa. For instance, the following operations provide equivalent optimization problems:

- change of variables;
- bijective transformation of objective and constraint functions;
- introduction of slack variables;
- optimizing over some variables;
- turning into the epigraph formulation;
- changing explicit constraints into implicit ones (that is moving constraints in $\mathcal{X}$), and vice versa.

Let us remind that the feasible set of Problem (P1) is $\mathcal{C} = \{x \in \mathcal{X} : \forall j \in [p] \colon g_j(x) \leqslant 0, \forall j \in [m] \colon h_j(x) = 0\}$.

Definition 1.79 (Convex optimization problem)**.** *Problem (P1) is a convex optimization problem if:*

1. *$f \colon \mathcal{X} \to \mathbb{R}$ is convex;*
2. *$(g_j)_{j \in [p]}$ are convex functions;*
3. *$(h_j)_{j \in [m]}$ are affine functions.*

Remark 1.80. Let us denote F the canonical extension of f to an extended-valued function. Stating that Problem (P1) is a convex optimization problem implies that $F + \chi_{\mathcal{C}}$ is a convex function. However, the converse is false. This is so because there are many ways to write a constrained optimization problem, while there is a single way to write an unconstrained one. Thus, a convex optimization problem cannot reduce to minimizing a convex real-valued function over a convex set of constraints.

Thus, a canonical convex optimization problem has the form:

$$\begin{aligned} &\underset{x \in \mathcal{X}}{\text{minimize}} \; f(x) \\ &\text{s.t.} \quad \begin{cases} \forall j \in [p] \colon g_j(x) \leqslant 0 \\ Ax = b, \end{cases} \end{aligned} \tag{P3}$$

where f and g_j are convex, $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$.

Local and global minimizers of an optimization problem are defined as the ones of its extended-valued objective function $F + \chi_{\mathcal{C}}$.

Proposition 1.81 (Optimality criterion)**.** *Assume that f is differentiable. Then $x^\star \in \mathcal{C}$ is a global minimizer of Problem (P3) if and only if*

$$\forall x \in \mathcal{C} \colon \nabla f(x^\star)^\top (x - x^\star) \geqslant 0.$$

Proof. Since $\forall x \in \mathcal{C}$: $f(x) \geqslant f(x^\star) + \nabla f(x^\star)(x - x^\star)$, if the equality is verified, then $x^\star$ is a global optimum.

Conversely, if $\exists x \in \mathcal{C} : \nabla f(x^\star)^\top(x - x^\star) < 0$, then f decreases strictly in the direction $x - x^\star$. As a consequence, we can find $t \in (0, 1)$ (close to 0) such that $f(x^\star) > f(tx^\star + (1-t)x)$ (and $tx^\star + (1 - t)x \in \mathcal{C}$ by convexity). □

Although quite powerful, this optimality criterion is hardly ever used (it should be verified for all x). However, the Lagrangian multiplier optimality condition comes from this property and is the one used in practice.

This section also provides examples of convex canonical optimization problems, that are *easily* solvable for medium-sized situations and that one should recognize. We present informally these special optimization problems, agreeing that all variables other than x are any vectors or matrices of appropriate dimensions.

Linear programs

A linear program (LP) is of the form:

$$\begin{array}{cl} \underset{x \in \mathbb{R}^d}{\text{minimize}} & c^\top x + d \\ \text{s.t.} & \begin{cases} Gx \preccurlyeq h \\ Ax = b. \end{cases} \end{array}$$

Here, the feasible set is a polyhedron.

A linear-fractional program

$$\begin{array}{cl} \underset{x \in \mathbb{R}^d : e^\top x + f > 0}{\text{minimize}} & \dfrac{c^\top x + d}{e^\top x + f} \\ \text{s.t.} & \begin{cases} Gx \preccurlyeq h \\ Ax = b \end{cases} \end{array}$$

can be turned into a linear program:

$$\begin{array}{cl} \underset{x \in \mathbb{R}^d}{\text{minimize}} & c^\top y + d^\top z \\ \text{s.t.} & \begin{cases} Gy - hz \preccurlyeq 0 \\ Ay - bz = 0 \\ e^\top y + f^\top z = 1 \\ z \geqslant 0. \end{cases} \end{array}$$

Quadratic programs

$$\begin{array}{cl} \underset{x \in \mathbb{R}^d}{\text{minimize}} & \frac{1}{2} x^\top P x + q^\top x + r \\ \text{s.t.} & \begin{cases} Gx \preccurlyeq h \\ Ax = b, \end{cases} \end{array}$$

where P is a positive semi-definite matrix. If there are quadratic constraints, we face a quadratically constrained quadratic program (QCQP):

$$\begin{aligned} &\underset{x\in\mathbb{R}^d}{\text{minimize}}\ \frac{1}{2}x^\top Px + q^\top x + r \\ &\text{s.t.} \quad \begin{cases} \forall j \in [p], \frac{1}{2}x^\top P_j x + q_j^\top x + r_j \leqslant 0 \\ Ax = b, \end{cases} \end{aligned}$$

where P_j are positive semi-definite.

LPs and QCQPs are special cases of second-order cone programs:

$$\begin{aligned} &\underset{x\in\mathbb{R}^d}{\text{minimize}}\ f^\top x \\ &\text{s.t.} \quad \begin{cases} \forall j \in [p], \|A_j x + b_j\|_2 \leqslant c_j^\top x + d_j \\ Fx = g. \end{cases} \end{aligned}$$

Geometric programs

Definition 1.82 (Posynomials). $\varphi\colon (\mathbb{R}_+^*)^d \to \mathbb{R}$ *is called a monomial if*

$$\exists \alpha \in \mathbb{R}^d, \exists \beta > 0\colon \quad \forall x \in (\mathbb{R}_+^*)^d, \varphi(x) = \beta \prod_{i=1}^d x_i^{\alpha_i}.$$

A posynomial is the sum of several monomials.

A geometric program has the form:

$$\begin{aligned} &\underset{x\in\mathcal{X}}{\text{minimize}}\ f(x) \\ &\text{s.t.} \quad \begin{cases} \forall j \in [p]\colon g_j(x) \leqslant 1 \\ \forall j \in [m]\colon h_j(x) = 1, \end{cases} \end{aligned}$$

where f and g_j are posynomials and h_j are monomials.

Geometric programs are not convex programs but can be turned into convex optimization programs with a simple change of variable.

Generalized inequality constraints

Let $\mathcal{K} \subset \mathbb{R}^d$ be a proper cone and denote $\leqslant_{\mathcal{K}}$ the relation defined by $y \leqslant_{\mathcal{K}} x \iff x - y \in \mathcal{K}$. A convex optimization problem with generalized inequality constraints is:

$$\begin{aligned} &\underset{x\in\mathbb{R}^d}{\text{minimize}}\ f(x) \\ &\text{s.t.} \quad \begin{cases} \forall j \in [p]\colon g_j(x) \leqslant_{\mathcal{K}_j} 0 \\ Ax = b, \end{cases} \end{aligned}$$

where f and g_j are convex. As special cases, a conic form problem is:

$$\begin{aligned} &\underset{x\in\mathbb{R}^d}{\text{minimize}}\ c^\top x \\ &\text{s.t.} \quad \begin{cases} Fx + g \preccurlyeq_{\mathcal{K}} 0 \\ Ax = b, \end{cases} \end{aligned}$$

while a semi-definite program (SDP) is:

$$\begin{aligned} &\underset{x\in\mathbb{R}^d}{\text{minimize}}\ c^\top x \\ &\text{s.t.} \quad \begin{cases} \sum_{i=1}^{d} x_i F_i + G \preccurlyeq 0 \\ Ax = b, \end{cases} \end{aligned}$$

where G and F_i are symmetric matrices.

LPs and SOCPs are special cases of SDPs. Yet, considering a tighter class of programs makes it possible to design fast numerical algorithms.

1.3 Legendre-Fenchel transformation and duality

1.3.1 The convex conjugate

Definition 1.83 (Legendre-Fenchel transformation)**.** *Let $f\colon \mathbb{R}^d \to [-\infty, \infty]$. The Legendre-Fenchel transformation (or convex conjugate) of f is:*

$$f^*\colon y \in \mathbb{R}^d \mapsto \sup_{x\in\mathbb{R}^d} \left\{ y^\top x - f(x) \right\}.$$

Remark 1.84. If $f\colon \mathbb{R}^d \to \mathbb{R}$, f is differentiable and the supremum is attained in x^*, then x^* is such that $y = \nabla f(x^*)$ and $f^*(y) = -(f(x^*) + \nabla f(x^*)^\top (0 - x^*))$, which is minus the linear approximation of f in 0 from $f(x^*)$.

Proposition 1.85 (Some properties of the convex conjugate)**.** *Let $f\colon \mathbb{R}^d \to [-\infty, \infty]$.*

1. *$f^*(0) = -\inf_{x\in\mathbb{R}^d} f(x)$.*

2. *f^* is convex and lower semi-continuous.*

3. *If $\operatorname{dom}(f) \neq \varnothing$, then $\forall y \in \mathbb{R}^d\colon f^*(y) > -\infty$.*

4. *If f is convex and proper, then f^* is proper.*

Proof.

1. Trivial.

2. sup of affine (thus convex) functions.

3. Trivial.

4. Since f is proper, f^* is proper except if $f = \infty$, which is not true (this comes from the existence of a subgradient of f).

□

Proposition 1.86 (Fenchel-Young inequality)**.** *Let $f \colon \mathbb{R}^d \to [-\infty, \infty]$. Then*

$$\forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d: \quad f(x) + f^*(y) \geqslant x^\top y.$$

Moreover, if f is convex, $\forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d$:

$$f(x) + f^*(y) = x^\top y \iff y \in \partial f(x).$$

Proof. Both statements come from the definition of the convex conjugate and of a subgradient. □

Example 1.87 (Remarkable convex conjugates)**.**

1. *If $f = \frac{1}{2}\| \cdot \|_2^2$, then $f^* = \frac{1}{2}\| \cdot \|_2^2$.*

2. *If $f = \exp$, then $f^*(y) = y(\log(y) - 1)$ if $y > 0$, $f(y) = \infty$ if $y < 0$ and $f(0) = 0$.*

3. *Let $\mathcal{K} \subset \mathbb{R}^d$. If $f = \chi_{\mathcal{K}}$, then $f^* \colon y \in \mathbb{R}^d \mapsto \sup_{x \in \mathcal{K}} y^\top x$.*

Exercise 1.88.

1. *Let $Q \in \mathbb{R}^{d \times d}$ be a positive definite matrix and $f \colon x \in \mathbb{R}^d \mapsto x^\top Q x$. Compute f^*.*

2. *Let $f \colon \mathbb{R}^d \to [-\infty, \infty]$. Show that $f = f^* \iff f = \frac{1}{2}\| \cdot \|_2^2$.*

Definition 1.89 (Dual norm)**.** *Let $\| \cdot \|$ be a norm on $\mathbb{R}^d$. Its dual norm $\| \cdot \|_*$ is defined by:*

$$\forall y \in \mathbb{R}^d: \quad \|y\|_* = \sup_{\|x\| \leqslant 1} y^\top x.$$

Proposition 1.90. *Let $\| \cdot \|$ be a norm on $\mathbb{R}^d$.*

1. *$\| \cdot \|_*$ is a norm.*

2. *$\forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d: \quad y^\top x \leqslant \|x\| \|y\|_*$.*

3. *The dual norm of a dual norm is the primal norm: $(\| \cdot \|_*)_* = \| \cdot \|$.*

Example 1.91 (Dual norms)**.**

1. *Let $p > 1$ and $q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Then $\| \cdot \|_p$ is dual to $\| \cdot \|_q$. We deduce Hölder's inequality: $\forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d: \quad y^\top x \leqslant \|x\|_p \|y\|_q$.*

2. *Particular cases are: ℓ_2 is self-dual, ℓ_1 and ℓ_∞ are dual.*

3. *For matrices, the Frobenius norm is sefl-dual, the spectral and the trace norms are dual.*

Proposition 1.92 (Convex conjugate of a norm)**.** *Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$. Then, $\|\cdot\|^* = \chi_{\{x\in\mathbb{R}^d : \|x\|_* \leqslant 1\}}$.*

In other words, the convex conjugate of a norm is the characteristic function of the dual norm ball.

Proposition 1.93 (Biconjugate, involution)**.** *Let $f\colon \mathbb{R}^d \to [-\infty, \infty]$ and $f^{**} = (f^*)^*$ the biconjugate of f. Then*

$$\forall x \in \mathbb{R}^d\colon\ f(x) \geqslant f^{**}(x).$$

In addition, if f is convex, proper and lower semi-continuous, then

$$f = f^{**}$$

and

$$y \in \partial f(x) \iff x \in \partial f^*(y).$$

Proof. The first and last statements come from Fenchel-Young (in)equality. The middle statement is admitted. □

Remark 1.94. The biconjugate f^{**} is sometimes called the convex relaxation of f.

Exercise 1.95. *Compute the convex conjugate of the pseudo-norm ℓ_0.*

1.3.2 Duality

All along this section, we will consider the canonical convex optimization problem:

$$\begin{aligned} &\underset{x\in\mathbb{R}^d}{\text{minimize}}\ f(x) \\ &\text{s.t.} \quad \begin{cases} g(x) \leqslant 0 \\ h(x) = 0, \end{cases} \end{aligned} \tag{P4}$$

where $f\colon \mathbb{R}^d \to \mathbb{R}$ is convex, $g\colon \mathbb{R}^d \to \mathbb{R}^p$ is component-wise convex and $h\colon \mathbb{R}^d \to \mathbb{R}^m$ is affine. Let us denote $\chi_{g\leqslant 0}$ and $\chi_{h=0}$ respectively the characteristic functions of $\{x \in \mathbb{R}^d : g(x) \leqslant 0\}$ and $\{x \in \mathbb{R}^d : h(x) = 0\}$, as well as the extended-valued function $F = f + \chi_{g\leqslant 0} + \chi_{h=0}$. We remark that F is convex and that Problem (P4) is equivalent to:

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\ F(x) = f(x) + \chi_{g\leqslant 0}(x) + \chi_{h=0}(x).$$

Since characterizing the solutions of F may be difficult, it is often useful to consider a dual problem. This one is deduced from the Lagrangian function.

Definition 1.96 (Lagrangian function)**.** *The Lagrangian function associated to Problem (P4) is:*

$$L\colon (x, \lambda, \nu) \in \mathbb{R}^d \times \mathbb{R}^p \times \mathbb{R}^m \mapsto f(x) + \lambda^\top g(x) + \nu^\top h(x) - \chi_{\mathbb{R}^p_+}(\lambda).$$

λ and ν are called Lagrange multipliers.

Proposition 1.97 (Supremum of the Lagrangian function)**.** *With the previous notation:*

$$\forall x \in \mathbb{R}^d : \quad F(x) = \sup_{(\lambda,\nu)\in\mathbb{R}^p\times\mathbb{R}^m} L(x, \lambda, \nu).$$

Proof. Immediate. □

From this we remark that Problem (P4) is equivalent to the saddle point problem:

$$\operatorname*{minimize}_{x\in\mathbb{R}^d} \sup_{(\lambda,\nu)\in\mathbb{R}^p\times\mathbb{R}^m} L(x, \lambda, \nu),$$

that has optimal value

$$\inf_{x\in\mathbb{R}^d} F(x) = \inf_{x\in\mathbb{R}^d} \sup_{(\lambda,\nu)\in\mathbb{R}^p\times\mathbb{R}^m} L(x, \lambda, \nu).$$

As a consequence, it is tempting to exchange inf and sup in order to get another (maximization) problem. That is exactly how we proceed (with caution) to get a dual problem.

Definition 1.98 (Dual function and dual problem)**.** *The Lagrange dual function of Problem (P4) is:*

$$G\colon (\lambda, \nu) \in \mathbb{R}^p \times \mathbb{R}^m \mapsto \inf_{x\in\mathbb{R}^d} L(x, \lambda, \nu).$$

The dual problem of Problem (P4) is:

$$\operatorname*{maximize}_{(\lambda,\nu)\in\mathbb{R}^p\times\mathbb{R}^m} G(\lambda, \nu). \tag{P5}$$

Example 1.99 (Link with convex conjugates)**.** *Assume that we are interested in an optimization problem with linear constraints:*

$$\begin{aligned} &\operatorname*{minimize}_{x\in\mathbb{R}^d} f(x) \\ &\quad s.t. \quad \begin{cases} Ax \preccurlyeq b \\ Cx = d, \end{cases} \end{aligned}$$

where A, C, b and d are any matrices and vectors. Then

$$\forall (\lambda, \nu) \in \mathbb{R}^p \times \mathbb{R}^m \colon G(\lambda, \nu) = -(\lambda^\top b + \nu^\top d) - f^*(-A^\top \lambda - C^\top \nu) - \chi_{\mathbb{R}^p_+}(\lambda).$$

Exercise 1.100. *Compute the dual functions of the following problems:*

1.

$$\begin{aligned} &\underset{x \in \mathbb{R}^d}{\text{minimize}}\ \|x\|_2^2 \\ &s.t. \qquad Cx = d. \end{aligned}$$

2.

$$\begin{aligned} &\underset{x \in \mathbb{R}^d}{\text{minimize}}\ c^\top x \\ &s.t. \qquad \begin{cases} Cx = d \\ x \geqslant 0. \end{cases} \end{aligned}$$

3.

$$\begin{aligned} &\underset{x \in \mathbb{R}^d}{\text{minimize}}\ \|x\| \\ &s.t. \qquad Ax \leqslant b. \end{aligned}$$

4.

$$\begin{aligned} &\underset{x \in \mathbb{R}^d}{\text{minimize}}\ \|x\|_1 \\ &s.t. \qquad Cx = d. \end{aligned}$$

Proposition 1.101. *$-G$ is convex and lower semi-continuous.*

Proof. Let $q\colon (\lambda, \nu) \in \mathbb{R}^p \times \mathbb{R}^m \mapsto \inf_{x \in \mathbb{R}^d} f(x) + \chi_{g \leqslant \lambda}(x) + \chi_{h=\nu}(x)$. Then $\forall (\lambda, \nu) \in \mathbb{R}^p \times \mathbb{R}^m\colon -G(\lambda, \nu) = q^*(-\lambda, -\nu)$. By convex conjugation, $-G$ is convex and lower semi-continuous. □

Proposition 1.102 (Weak duality)**.** *Let $p \in [-\infty, \infty]$ and $d \in [-\infty, \infty]$ be respectively the primal and dual optimal objective values:*

$$\begin{aligned} p &= \inf_{x \in \mathbb{R}^d} F(x) &&= \inf_{x \in \mathbb{R}^d} \sup_{(\lambda,\nu) \in \mathbb{R}^p \times \mathbb{R}^m} L(x, \lambda, \nu), \\ d &= \sup_{(\lambda,\nu) \in \mathbb{R}^p \times \mathbb{R}^m} G(\lambda, \nu) &&= \sup_{(\lambda,\nu) \in \mathbb{R}^p \times \mathbb{R}^m} \inf_{x \in \mathbb{R}^d} L(x, \lambda, \nu). \end{aligned}$$

Then

$$d \leqslant p.$$

Proof. Remark that

$$\sup_{(\lambda,\nu) \in \mathbb{R}^p \times \mathbb{R}^m} \inf_{x \in \mathbb{R}^d} L(x, \lambda, \nu) \leqslant \inf_{x \in \mathbb{R}^d} \sup_{(\lambda,\nu) \in \mathbb{R}^p \times \mathbb{R}^m} L(x, \lambda, \nu),$$

by definition of inf and sup. □

Remark 1.103. Let $(x', \lambda', \nu') \in \mathbb{R}^d \times \mathbb{R}^p \times \mathbb{R}^m$. In numerical optimization, the dual gap

$$F(x') - G(\lambda', \nu') \geqslant 0$$

is a certificate to get an ϵ-solution. Indeed, since $G(\lambda', \nu') \leqslant d \leqslant p \leqslant F(x')$, if $F(x') - G(\lambda', \nu') \leqslant \epsilon$, then

$$0 \leqslant F(x') - \inf_{x \in \mathbb{R}^d} F(x) \leqslant F(x') - G(\lambda', \nu') \leqslant \epsilon.$$

Proposition 1.104 (Saddle point)**.** *$(x^\star, \lambda^\star, \nu^\star) \in \mathbb{R}^d \times \mathbb{R}^p \times \mathbb{R}^m$ is a saddle point of L, that is*

$$\forall (x, \lambda, \nu) \in \mathbb{R}^d \times \mathbb{R}^p \times \mathbb{R}^m : \quad L(x^\star, \lambda, \nu) \leqslant L(x^\star, \lambda^\star, \nu^\star) \leqslant L(x, \lambda^\star, \nu^\star),$$

if and only if

$$\begin{aligned} \sup_{(\lambda,\nu)\in\mathbb{R}^p\times\mathbb{R}^m} \inf_{x\in\mathbb{R}^d} L(x, \lambda, \nu) &= \sup_{(\lambda,\nu)\in\mathbb{R}^p\times\mathbb{R}^m} L(x^\star, \lambda, \nu) \\ &= L(x^\star, \lambda^\star, \nu^\star) \\ &= \inf_{x\in\mathbb{R}^d} L(x, \lambda^\star, \nu^\star) \\ &= \inf_{x\in\mathbb{R}^d} \sup_{(\lambda,\nu)\in\mathbb{R}^p\times\mathbb{R}^m} L(x, \lambda, \nu). \end{aligned}$$

Proof. Suppose that $(x^\star, \lambda^\star, \nu^\star)$ is a saddle point. Then $\sup_{(\lambda,\nu)\in\mathbb{R}^p\times\mathbb{R}^m} L(x^\star, \lambda, \nu) \leqslant L(x^\star, \lambda^\star, \nu^\star)$ and $L(x^\star, \lambda^\star, \nu^\star) \leqslant \inf_{x\in\mathbb{R}^d} L(x, \lambda^\star, \nu^\star)$. Then

$$\begin{aligned} \sup_{(\lambda,\nu)\in\mathbb{R}^p\times\mathbb{R}^m} \inf_{x\in\mathbb{R}^d} L(x, \lambda, \nu) &\leqslant \inf_{x\in\mathbb{R}^d} \sup_{(\lambda,\nu)\in\mathbb{R}^p\times\mathbb{R}^m} L(x, \lambda, \nu) \\ &\leqslant \sup_{(\lambda,\nu)\in\mathbb{R}^p\times\mathbb{R}^m} L(x^\star, \lambda, \nu) \\ &\leqslant L(x^\star, \lambda^\star, \nu^\star) \\ &\leqslant \inf_{x\in\mathbb{R}^d} L(x, \lambda^\star, \nu^\star) \\ &\leqslant \sup_{(\lambda,\nu)\in\mathbb{R}^p\times\mathbb{R}^m} \inf_{x\in\mathbb{R}^d} L(x, \lambda, \nu). \end{aligned}$$

Thus, all inequalities are in fact equalities. In addition, the converse is straightforward. □

Theorem 1.105 (Strong duality)**.** *If Problem (P4) is strictly feasible (Slater's constraint qualification):*

$$\exists x \in \mathbb{R}^d : g(x) \prec 0 \text{ and } h(x) = 0,$$

where $\prec$ means component-wise strict inequality, then (with the same notation as previously):

1. *$\inf_{x\in\mathbb{R}^d} F(x) < \infty$ (the problem is feasible);*
2. *$d = p$ (zero duality gap);*
3. *$\exists (\lambda^\star, \nu^\star) \in \mathbb{R}^p_+ \times \mathbb{R}^m : d = G(\lambda^\star, \nu^\star)$ (dual is attained).*

Proof. Admitted. □

Remark 1.106. If we assume that the domains of f, g and h are not $\mathbb{R}^d$, constraint qualification is:

1. $0 \in \operatorname{relint}(h(\operatorname{dom} f))$;
2. $\exists x \in \operatorname{dom} f : g(x) \prec 0$.

Theorem 1.107 (Karush-Kuhn-Tucker conditions). *Assume that f, g and h are differentiable functions and that Slater's qualification holds for Problem (P4). Then, $(x^\star, \lambda^\star, \nu^\star) \in \mathbb{R}^d \times \mathbb{R}^p \times \mathbb{R}^m$ is a saddle point of L if and only if:*

1. *primal feasibility: $g(x^\star) \preccurlyeq 0$ and $h(x^\star) = 0$;*
2. *dual feasibility: $\lambda^\star \succcurlyeq 0$;*
3. *complementary slackness: $\forall j \in [p], \lambda_j^\star g_j(x^\star) = 0$;*
4. *stationarity: $\nabla_x L(x^\star, \lambda^\star, \nu^\star) = 0$.*

Remark 1.108. Lagrange multipliers λ and ν give strong information about the sensitivity of the optimal value with respect to perturbations to the constraints. To illustrate this, let $(u, v) \in \mathbb{R}^p \times \mathbb{R}^m$ and a perturbed problem be defined by:

$$\begin{aligned} &\underset{x\in\mathbb{R}^d}{\text{minimize}}\ f(x) \\ &\text{s.t.} \quad \begin{cases} g(x) \preccurlyeq u \\ h(x) = v. \end{cases} \end{aligned}$$

Let $q\colon (u, v) \in \mathbb{R}^p \times \mathbb{R}^m \mapsto \inf_{x\in\mathbb{R}^d} f(x) + \chi_{g\preccurlyeq u}(x) + \chi_{h=v}(x)$ be the optimal value of the perturbed problem and remark that $q(0, 0) = \inf_{x\in\mathbb{R}^d} F(x)$ (the optimal value of the original problem).

The dual of the perturbed problem is:

$$\underset{(\lambda,\nu)\in\mathbb{R}^p\times\mathbb{R}^m}{\text{maximize}}\ G(\lambda, \nu) - u^\top \lambda - v^\top \nu.$$

Let $(\lambda^\star, \nu^\star) \in \mathbb{R}^p \times \mathbb{R}^m$ be a solution of the dual original problem and assume that strong duality holds for this problem. Then we have (weak duality applied to the perturbed problem):

$$q(u, v) - q(0, 0) = q(u, v) - G(\lambda^\star, \nu^\star) \geqslant -(u^\top \lambda^\star + v^\top \nu^\star).$$

As a consequence, if q is differentiable in $(0, 0)$, we get:

$$\nabla_u q(0, 0) = -\lambda \quad \text{and} \quad \nabla_v q(0, 0) = -\nu.$$

1.3.3 Generalized inequality constraints

In Problem (P4), we remark that the inequality constraint $g(x) \preccurlyeq 0$ can be interpreted as $-g(x) \in K$, where $K = \mathbb{R}^p_+$ is the positive orthant, which is a cone. Thus, the inequality constraint is equivalent to $g(x) \preccurlyeq_K 0$. This motivates the extension of duality to generalized inequality constraints.

As a consequence, let us consider

$$\begin{aligned} &\underset{x\in\mathbb{R}^d}{\text{minimize}}\ f(x) \\ &\text{s.t.} \quad \begin{cases} g(x) \preccurlyeq_K 0 \\ h(x) = 0, \end{cases} \end{aligned}$$

where $K \subset \mathbb{R}^p$ is any proper convex cone.

Similarly to usual convex problems, the Lagrangian function can be defined by:

$$L\colon (x, \lambda, \nu) \in \mathbb{R}^d \times \mathbb{R}^p \times \mathbb{R}^m \mapsto f(x) + \lambda^\top g(x) + \nu^\top h(x) - \chi_{K^*}(\lambda),$$

where the dual cone K^* appeared. Let us remark that this is consistent with the previous definition of the Lagrangian since $\mathbb{R}^p_+$ is a self-dual cone.

Then, the dual function is defined similarly as before. Weak duality follows from these definitions, as well as strong duality with Slater's constraint qualification:

$$\exists x \in \mathbb{R}^d : g(x) \prec_K 0 \text{ and } h(x) = 0,$$

where $g(x) \prec_K 0$ means that $-g(x) \in \mathrm{int}(K)$.

Finally, KKT optimality conditions and the perturbation analysis are readily extended to generalized inequality constraints.

1.3.4 Tikhonov, Ivanov and Morozov regularizations

For many reasons (including numerical stability and sparsity), we often encounter optimization problems of the form:

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\ f(x) + \lambda \|x\|_p^p.$$

where $f\colon \mathbb{R}^d \to \mathbb{R}$ is convex (it embodies a data fitting term), $p \geqslant 1$ and $\lambda > 0$. The additional term on the right hand side is often referred to as Tikhonov regularization. Historically, Tikhonov regularization came out because of ill-posed problems.

It has to be known that this formulation is equivalent to two others. The first one is Ivanov regularization (or quasi-solution method):

$$\begin{aligned} &\underset{x \in \mathbb{R}^d}{\text{minimize}} && f(x) \\ &\text{s.t.} && \|x\|_p \leqslant \tau, \end{aligned}$$

where $\tau > 0$, and the second is Morozov regularization (or residual method):

$$\begin{aligned} &\underset{x \in \mathbb{R}^d}{\text{minimize}} && \|x\|_p \\ &\text{s.t.} && f(x) \leqslant \delta, \end{aligned}$$

where $\delta > 0$. These last two formulations can respectively be interpreted as fitting the data with not too rough parameter x, and finding x with minimal norm that fits the data up to a δ accuracy.

The following theorem makes the equivalence clear.

Theorem 1.109 (Equivalence between regularizations). *Let $\varphi\colon \mathbb{R}^d \to \mathbb{R}$ and $\psi\colon \mathbb{R}^d \to \mathbb{R}$ be to convex functions such that $\psi \geqslant 0$ and $0 \in \psi(\mathbb{R}^d)$.*

Let $\lambda \geqslant 0$. If $x^\star \in \mathbb{R}^d$ is a minimizer of:

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\ \varphi(x) + \lambda \psi(x), \tag{P6}$$

then there exists $\tau \geqslant 0$ such that $x^\star$ is a minimizer of

$$\begin{aligned} &\underset{x \in \mathbb{R}^d}{\text{minimize}} \; \varphi(x) \\ &\quad \textit{s.t.} \qquad \psi(x) \leqslant \tau. \end{aligned} \tag{P7}$$

Conversely, for $\tau > 0$, if $x^\star \in \mathbb{R}^d$ is a minimizer of Problem (P7), then there exists $\lambda \geqslant 0$ such that $x^\star$ is a minimizer of Problem (P6).

Proof. Let $\lambda \geqslant 0$ and $x^\star \in \mathbb{R}^d$ be a minimizer of Problem (P6). Let $\tau = \psi(x^\star) \geqslant 0$. Then $\psi(x^\star) \leqslant \tau$ and $\forall x \in \mathbb{R}^d : \psi(x) \leqslant \tau$,

$$\varphi(x^\star) = \varphi(x^\star) + \lambda\psi(x^\star) - \lambda\tau \leqslant \varphi(x) + \lambda\psi(x) - \lambda\tau \leqslant \varphi(x),$$

so $x^\star$ is a minimizer of Problem (P7).

Conversely, let $\tau > 0$ and $x^\star \in \mathbb{R}^d$ be a minimizer of Problem (P7). Since $0 \in \psi(\mathbb{R}^d)$ and $\tau > 0$, Slater's constraint qualification hold and there is strong duality. Consequently, the dual is attained:

$$\exists \lambda \in \mathbb{R}_+ : \varphi(x^\star) + \lambda(\psi(x^\star) - \tau) = \sup_{\lambda' \in \mathbb{R}_+} \varphi(x^\star) + \lambda'(\psi(x^\star) - \tau).$$

Therefore, $(x^\star, \lambda)$ is a saddle point of the Lagrangian. Hence, $\forall x \in \mathbb{R}^d$:

$$\varphi(x^\star) + \lambda\psi(x^\star) = \varphi(x^\star) + \lambda(\psi(x^\star) - \tau) + \lambda\tau \leqslant \varphi(x) + \lambda(\psi(x) - \tau) + \lambda\tau = \varphi(x) + \lambda\psi(x).$$

Thus, $x^\star$ is a minimizer of Problem (P6). □

1.3.5 A relevant example

Let us consider the optimization problem:

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \; f(x) + g(Ax), \tag{P8}$$

where $f : \mathbb{R}^d \to (-\infty, \infty]$, $g : \mathbb{R}^p \to (-\infty, \infty]$ are proper convex and $A \in \mathbb{R}^{p \times d}$. This problem is equivalent to

$$\begin{aligned} &\underset{x \in \mathbb{R}^d, y \in \mathbb{R}^p}{\text{minimize}} \; f(x) + g(y) \\ &\quad \text{s.t.} \qquad Ax = y. \end{aligned}$$

The dual function to this last problem is:

$$\begin{aligned} \forall \nu \in \mathbb{R}^p : G(\nu) &= \inf_{x \in \mathbb{R}^d, y \in \mathbb{R}^p} \left\{ f(x) + g(y) + \nu^\top Ax - \nu^\top y \right\} \\ &= -\sup_{y \in \mathbb{R}^p} \left\{ \nu^\top y - g(y) \right\} - \sup_{x \in \mathbb{R}^d} \left\{ -\nu^\top Ax - f(x) \right\} \\ &= -g^*(\nu) - f^*(-A^\top \nu). \end{aligned}$$

Therefore, the dual problem of interest is:

$$\underset{\nu \in \mathbb{R}^p}{\text{maximize}} \; -g^*(\nu) - f^*(-A^\top \nu). \tag{P9}$$

	Constraint	Set $\mathcal{C}$	Conjugate
Equality	$Ax = b$	$\{0\}$	0
Ball	$\|Ax - b\| \leqslant 1$	unit $\|\cdot\|$-ball	$\|\cdot\|_*$
Conic inequality	$Ax \preccurlyeq_{\mathcal{K}} b$	$-\mathcal{K}$	$\chi_{\mathcal{K}^*}$

Table 1.1 – Examples of constraints.

Theorem 1.110. *Let $f \colon \mathbb{R}^d \to (-\infty, \infty]$, $g \colon \mathbb{R}^p \to (-\infty, \infty]$ be proper convex functions and $A \in \mathbb{R}^{p \times d}$. Assume that either $\operatorname{dom}(f) = \mathbb{R}^p$ or $\operatorname{dom}(g) = \mathbb{R}^d$ and that $\exists x \in \mathbb{R}^d : Ax \in \operatorname{dom}(g)$. If the optima are attained in Problem (P8) and Problem (P9), then strong duality holds:*

$$\min_{x \in \mathbb{R}^d} f(x) + g(Ax) = \max_{\nu \in \mathbb{R}^p} -g^*(\nu) - f^*(-A^\top \nu).$$

Moreover, a primal-dual optimum is a solution to the saddle-point problem:

$$\underset{x \in \mathbb{R}^d}{\operatorname{minimize}}\, \underset{\nu \in \mathbb{R}^p}{\operatorname{maximize}}\ f(x) + \nu^\top A x - g^*(\nu).$$

The forthcoming paragraphs provide examples of such problems.

Set constraint

Here, we are in the case where $g = \chi_{\mathcal{C}}$, where $\mathcal{C} \subset \mathbb{R}^d$ is a convex set, and we aim at solving:

$$\underset{x \in \mathbb{R}^d}{\operatorname{minimize}}\ f(x) + \chi_{\mathcal{C}}(Ax - b),$$

whose dual is

$$\underset{\nu \in \mathbb{R}^p}{\operatorname{maximize}}\ -b^\top \nu - \chi_{\mathcal{C}}^*(\nu) - f^*(-A^\top \nu).$$

Table 1.1 provide some examples of sets of constraint, where $\mathcal{K} \subset \mathbb{R}^p$ is a proper convex cone.

Norm regularization

In this case, $g(y) = \|y - b\|$ and we want to minimize $f(x) + \|Ax - b\|$. Since, $g^*(\nu) = b^\top \nu + \chi_{\mathcal{B}}(\nu)$, where $\mathcal{B} = \{y \in \mathbb{R}^p : \|y\|_* \leqslant 1\}$, the dual reads:

$$\begin{aligned} \underset{\nu \in \mathbb{R}^p}{\operatorname{maximize}}\ & -b^\top \nu - f^*(-A^\top \nu) \\ \text{s.t.}\quad & \|\nu\|_* \leqslant 1. \end{aligned}$$

Chapter 2

Optimization algorithms

[1]

Contents

2.1 Introduction 38

2.2 Greedy methods 38

2.2.1 Orthogonal matching pursuit 38

2.2.2 Compressive sampling matching pursuit 39

2.3 Linear programming 40

2.3.1 Convex relaxation of compressed sensing and basis pursuit 40

2.3.2 The simplex method 41

2.3.3 Barrier methods 42

2.4 Primal methods 44

2.4.1 Gradient method 44

2.4.2 Quasi-Newton method 47

2.4.3 Subgradient method 48

2.4.4 Proximal gradient method 49

2.4.5 Accelerated proximal gradient method 53

2.4.6 Douglas-Rachford method 53

2.5 Primal-dual methods 54

2.5.1 Lagrange multipliers 54

2.5.2 Augmented Lagrange multipliers 56

2.5.3 Alternating direction method of multipliers 57

[1] This chapter is inspired from P. Bianchi, O. Fercoq and A. Sabourin's lecture notes on Optimization for machine learning (University Paris-Saclay, Télécom ParisTech) [1] as well as Vandenberghe's lecture notes on Optimization Methods for Large-Scale Systems (Spring 2016) [5].

2.1 Introduction

In this chapter, we aim at providing several concrete methods to produce a sequence $(x_k)_{k\in\mathbb{N}}$ that minimizes a function $f\colon \mathbb{R}^d \to [-\infty, \infty]$, that is such that $\lim_{k\to\infty} x_k$ exists in $\mathbb{R}^d$ and $\lim_{k\to\infty} f(x_k) = \inf_{x\in\mathbb{R}^d} f(x)$.

2.2 Greedy methods

2.2.1 Orthogonal matching pursuit

The problem of interest in compressed sensing is to find $x \in \mathbb{R}^d$ such that $Ax = y$, where $A \in \mathbb{R}^{p\times d}$ is a sensing matrix and $y \in \mathbb{R}^p$ the vector of measurements. Compressed sensing promotes two special features:

1. the number of measurements is much smaller than the dimension of the signal ($p \ll d$), so the problem of finding x such that $y = Ax$ is under-determined;
2. the signal to recover is supposed s-sparse ($s \in \mathbb{N}^*$).

Thus, compressed sensing can be summed up in the following manner: given a sensing matrix $A \in \mathbb{R}^{p\times d}$ and a vector of measurements $y \in \mathbb{R}^p$, solve the optimization problem

$$\begin{aligned} &\underset{x\in\mathbb{R}^d}{\text{minimize}} \ \|x\|_0 \\ &\text{s.t.} \quad\ \ Ax = y. \end{aligned}$$

A roughly equivalent formulation to the compressed sensing problem is:

$$\begin{aligned} &\underset{x\in\mathbb{R}^d}{\text{minimize}} \ f(x) \\ &\text{s.t.} \quad\ \ \|x\|_0 \leqslant s, \end{aligned}$$

where $f\colon x \in \mathbb{R}^d \mapsto \|Ax - y\|_2$ and $s \in \mathbb{N}$ is a prescribed sparsity level. Let us remark that if the signal to recover $x^\star \in \mathbb{R}^d$ is s-sparse, then it is a solution the previous optimization problem and $f(x^\star) = 0$.

Starting with an initial point $x_0 = 0 \in \mathbb{R}^d$ and $S_0 = \operatorname{supp}(x_0) = \varnothing$ its support, the orthogonal matching pursuit algorithm reads as follow and aims at providing a local solution to the previous optimization problem.

Algorithm 1 Orthogonal matching pursuit

```
$(j_k, \alpha_k) \in \operatorname*{arg\,min}_{j\in[d],\alpha\in\mathbb{R}} f(x_k + \alpha e_j)$
$S_{k+1} = S_k \cup \{j_k\}$
$x_{k+1} \in \operatorname*{arg\,min}_{x\in\mathbb{R}^d:\operatorname{supp}(x)\subset S_{k+1}} f(x),$
```

where e_j is the j^{th} canonical basis vector of $\mathbb{R}^d$.

We note that:

- When the columns of A norm to 1, Step 1 boils down to finding $j \in [d]$, that maximizes $|(A_j)^\top (Ax - y)|$, where A_j is the j^{th} column of A. In other words, we look for the atom of the dictionary A, that is the most correlated to the residue $Ax - y$.

- Step 2 potentially increments the sparsity of the current iteration x_{k+1}: $\|x_{k+1}\|_0 \leqslant k+1$, at each iteration k.

- Step 3 is an orthogonal projection, hence the name *orthogonal* matching pursuit.

Remark 2.1. Under some conditions, the orthogonal matching pursuit can recover any s-sparse signal $x^\star$ with at most s iterations. However, the weakness of orthogonal matching pursuit is that, once an incorrect index j has been selected, it remains in the support of the proposed solution. In this case, s iterations are not enough to recover an s-sparse signal.

2.2.2 Compressive sampling matching pursuit

The compressive sampling matching pursuit algorithm proposes a strategy to overcome the weaknesses of orthogonal matching pursuit. To describe it, let $L_s \colon \mathbb{R}^d \to [d]$ be such that $L_s(x)$ is the index set of s largest absolute entries of x, and $H_s \colon \mathbb{R}^d \to \mathbb{R}^d$ be the hard-thresholding operator of order s. H_s is such that $H_s(x)$ has support $L_s(x)$ and equals x on its support (the other entries are 0).

Starting with an initial point $x_0 = 0 \in \mathbb{R}^d$, the Compressive sampling matching pursuit algorithm is defined by

Algorithm 2 Compressive sampling matching pursuit

$$\begin{aligned} S_{k+1} &= \operatorname{supp}(x_k) \cup L_{2s}(A^\top(Ax_k - y)) \\ u_{k+1} &\in \operatorname*{arg\,min}_{u \in \mathbb{R}^d : \operatorname{supp}(u) \subset S_{k+1}} \|Au - y\|_2 \\ x_{k+1} &= H_s(u_{k+1}). \end{aligned}$$

Remark 2.2. Orthogonal and compressive sampling matching pursuits require to estimate the sparsity of the signal $x^\star$ to recover. This is not an easy task.

2.3 Linear programming

2.3.1 Convex relaxation of compressed sensing and basis pursuit

Given a sensing matrix $A \in \mathbb{R}^{p \times d}$ and a vector of measurements $y \in \mathbb{R}^p$, compressed sensing aims at solving the optimization problem

$$\begin{aligned} \underset{x \in \mathbb{R}^d}{\text{minimize}} \ & \|x\|_0 \\ \text{s.t.} \quad & Ax = y. \end{aligned}$$

Since this problem is non-convex and even NP-hard in general, we would like to convexify it. For this purpose, let us remark that $\|\cdot\|_0$ is relatively well approximated by $\|\cdot\|_q^q$ when $q \to 0_+$. Yet $\|\cdot\|_q^q$ is not convex for $0 \leqslant q < 1$. The smallest value of q for which $\|\cdot\|_q^q$ is convex is $q = 1$. As a consequence, we can legitimately replace the original compressed sensing optimization problem by:

$$\begin{aligned} \underset{x \in \mathbb{R}^d}{\text{minimize}} \ & \|x\|_1 \\ \text{s.t.} \quad & Ax = y. \end{aligned} \tag{P10}$$

This problem is often referred to as *Basis pursuit*.

Proposition 2.3 (Sparsity of basis pursuit)**.** *Assume that Problem (P10) has a minimizer $x^\star \in \mathbb{R}^d$. Then $\|x^\star\|_0 \leqslant p$.*

Proof. By contradiction, assume that $\|x^\star\|_0 > p$ and let S be its support. Consider the set of p-dimensional vectors, columns of A index by S. This set is linearly dependent, that is $\exists u \in \mathbb{R}^d$, $u \neq 0$, with same support as $x^\star$, such that $Au = 0$. Now, let $t > 0$ such that $t < \min_{i \in S} \frac{|x_i^\star|}{\|u\|_\infty}$ and $z = x^\star - t \operatorname{sign}\left(\sum_{i \in S} \operatorname{sign}(x_i^\star) u_i\right) u$. Then we have $Az = y$ and

$$\begin{aligned} \|z\|_1 &= \sum_{i \in S} |z_i| \\ &= \sum_{i \in S} \operatorname{sign}(z_i) z_i \\ &= \sum_{i \in S} \operatorname{sign}(x_i^\star) z_i && (t \text{ is small enough}) \\ &= \sum_{i \in S} \operatorname{sign}(x_i^\star) x_i^\star - t \operatorname{sign}\left(\sum_{i \in S} \operatorname{sign}(x_i^\star) u_i\right) \sum_{i \in S} \operatorname{sign}(x_i^\star) u_i \\ &< \|x^\star\|_1, \end{aligned}$$

where the last inequality comes from the fact that $u \neq 0$, so *sign* $\left(\sum_{i \in S} \operatorname{sign}(x_i^\star) u_i\right) \sum_{i \in S} \operatorname{sign}(x_i^\star) u_i > 0$. Finally, $\|z\|_1 < \|x^\star\|_1$ is a contradiction. □

Next, we remark that the objective function of Problem (P10) is not differentiable, which makes its numerical solving difficult. As a consequence, one may propose to reformulate Problem (P10) to an "easier" problem.

Proposition 2.4 (Variational ℓ_1-norm)**.**

$$\forall x \in \mathbb{R}^d : \quad \|x\|_1 = \min\left\{ \sum_{i=1}^{d} \xi_i^+ + \xi_i^- : x = \xi^+ - \xi^-, (\xi^+, \xi^-) \in (\mathbb{R}_+^d)^2 \right\}.$$

Proof. Let $x \in \mathbb{R}^d$ and let us exhibit a minimizer of the set considered in the proposition. Let $(\xi^+, \xi^-) \in (\mathbb{R}_+^d)^2$ be such that $\xi_i^+ = \max(0, x_i)$ and $\xi_i^- = \max(0, -x_i)$ $(\forall i \in [d])$. Then, $x = \xi^+ - \xi^-$, so (ξ^+, ξ^-) is feasible, and we claim that (ξ^+, ξ^-) is a minimizer of the set considered in the proposition.

To show that, let us observe that $\forall i \in [d]$, if $x_i = 0$ then $\xi_i^+ = 0$ and $\xi_i^- = 0$, if $x_i > 0$ then $\xi_i^+ > 0$ and $\xi_i^- = 0$, and if $x_i < 0$ then $\xi_i^+ = 0$ and $\xi_i^- > 0$. Consequently, $|x_j| = \xi_j^+ + \xi_j^-$.

Now, let $(z^+, z^-) \in (\mathbb{R}_+^d)^2$ be a feasible point (that is such that $x = z^+ - z^-$). Then, $\forall j \in [d]$: $z_j^+ + z_j^- = z_j^+ - z_j^- + 2z_j^- \geqslant z_j^+ - z_j^-$. Conversely, $z_j^+ + z_j^- \geqslant z_j^- - z_j^+$, thus $z_j^+ + z_j^- \geqslant |z_j^+ - z_j^-| = |x_j| = \xi_j^+ + \xi_j^-$. As a consequence, $\sum_{i=1}^d z_i^+ + z_i^- \geqslant \sum_{i=1}^d \xi_i^+ + \xi_i^-$ and (ξ^+, ξ^-) is a minimizer.

To conclude, we remark that $\|x\|_1 = \sum_{i=1}^d \xi_i^+ + \xi_i^-$. □

As a consequence of the previous proposition, Problem (P10) can be reformulated in:

$$\begin{aligned} \underset{x \in \mathbb{R}^d}{\text{minimize}} \quad & \min_{(\xi^+, \xi^-) \in (\mathbb{R}^d)^2} \sum_{i=1}^{d} \xi_i^+ + \xi_i^- \\ & \quad \text{s.t.} \quad \begin{cases} x = \xi^+ - \xi^- \\ \xi^+ \geqslant 0 \\ \xi^- \geqslant 0, \end{cases} \\ \text{s.t.} \quad & Ax = y, \end{aligned}$$

which, combining minimization procedures and deleting the variable x, that appears to be totally free (so useless), becomes:

$$\begin{aligned} \underset{(\xi^+, \xi^-) \in (\mathbb{R}^d)}{\text{minimize}} \quad & \sum_{i=1}^{d} \xi_i^+ + \xi_i^- \\ \text{s.t.} \quad & \begin{cases} A(\xi^+ - \xi^-) = y \\ \xi^+ \geqslant 0 \\ \xi^- \geqslant 0. \end{cases} \end{aligned} \tag{P11}$$

In the forthcoming sections, we focus on algorithms for solving such a linear program.

2.3.2 The simplex method

We focus on an optimization problem of the form:

$$\begin{aligned} \underset{x \in \mathbb{R}^d}{\text{minimize}} \quad & c^\top x \\ \text{s.t.} \quad & \begin{cases} Ax = b \\ x \geqslant 0, \end{cases} \end{aligned} \tag{P12}$$

where $c \in \mathbb{R}^d$, $A \in \mathbb{R}^{p \times d}$ and $b \in \mathbb{R}^p$ are any matrices.

Remark 2.5. A linear program can always be written in the form of Problem (P12).

The feasibility set of Problem (P12) reads $\mathcal{C} = \{x \in \mathbb{R}^d : x \geqslant 0, Ax = b\}$. It is the intersection between the positive orthant and the affine space $\{x \in \mathbb{R}^d : Ax = b\}$. As a consequence, it is either:

1. empty, so Problem (P12) is not feasible;

2. not compact;

3. or the convex hull of a finite number of points.

In Situation 3, $\mathcal{C}$ is called a polytope or a simplex. This is the case of interest.

Proposition 2.6 (Solution of a linear program)**.** *Let us assume that $\mathcal{C}$ is non-empty and compact. Then, Problem (P12) has a solution, which is an extreme point of $\mathcal{C}$.*

Proof. Any point $x \in \mathcal{C}$ is a convex combination of the extreme points of $\mathcal{C}$, denoted $\{\kappa_i : i \in [n]\}$. In other words, $x = \sum_{i=1}^n t_i \kappa_i$ for some $t_i \geqslant 0$ such that $\sum_{i=1}^n t_i = 1$. Then we have $c^\top x = \sum_{i=1}^n t_i \kappa_i^\top c \geqslant \kappa_{i^\star}^\top c$, where $i^\star \in \arg\min_{i \in [n]} \kappa_i^\top c$. But $\kappa_{i^\star} \in \mathcal{C}$, so it is a minimizer. □

Exploring all extreme points of $\mathcal{C}$ would be very expensive. Therefore, the simplex algorithm finds a path in the set of extreme points of $\mathcal{C}$ such that the objective function does not increase at each iteration.

Generally, the simplex algorithms converges linearly in the number of constraints. However, the worst-case complexity is very bad. On the so-called Klee-Minty cube, the simplex algorithm exhibits poor performance (it visits all 2^p corners of the cube, where p is the number of constraints).

2.3.3 Barrier methods

Here, we focus on the optimization problem:

$$\begin{aligned} &\underset{x \in \mathbb{R}^d}{\text{minimize}}\ f(x) \\ &\text{s.t.} \quad \begin{cases} \forall j \in [p] \colon g_j(x) \leqslant 0 \\ Ax = b, \end{cases} \end{aligned}$$

where $A \in \mathbb{R}^{m \times d}$ is a rank m matrix, f and g_j are twice differentiable. We present barrier methods, also called interior point methods. They are particularly useful when f and g_j are linear functions.

The starting point of barrier methods is to rewrite the previous optimization problem in:

$$\begin{aligned} &\underset{x \in \mathbb{R}^d}{\text{minimize}}\ f(x) + \sum_{j=1}^m \chi_{\mathbb{R}_-}(g_j(x)) \\ &\text{s.t.} \quad Ax = b, \end{aligned}$$

and to remark that $\chi_{\mathbb{R}_-}$ can be approximated by a smooth barrier function. For instance, we consider here the logarithmic barrier

$$\phi\colon x \in \mathbb{R}^d \mapsto \begin{cases} -\sum_{j=1}^m \log(-g_j(x)) & \text{if } g_j(x) < 0, \forall j \in [m] \\ \infty & \text{otherwise.} \end{cases}$$

For $t > 0$, the function $x \in \mathbb{R}^d \mapsto \frac{1}{t}\phi(x)$ approximate $\chi_{\mathbb{R}^d_-}$ and the approximation improves as $t \to \infty$.

Proposition 2.7. *The barrier ϕ is convex and twice differentiable.*

Therefore, for $t > 0$, the problem of interest becomes:

$$\begin{aligned} \underset{x\in\mathbb{R}^d}{\text{minimize}} \;\; & tf(x) + \phi(x) \\ \text{s.t.} \quad\;\; & Ax = b, \end{aligned} \tag{P13}$$

Proposition 2.8. *Assume that strong duality holds for Problem (P13) and let $x^\star(t)$ be a minimizer Problem (P13) for $t > 0$. Then*

$$0 \leqslant f(x^\star(t)) - p^* \leqslant \frac{m}{t},$$

where $p^* = \inf_{x\in\mathbb{R}^d} f(x) + \sum_{j=1}^m \chi_{\mathbb{R}_-}(g_j(x)) + \chi_b(Ax)$ is the infimum of the original problem.

Proof. This comes from KKT conditions. □

Starting with a strictly feasible initial point $x_0 \in \mathbb{R}^d$, $t_0 > 0$ and $\mu > 1$, a barrier (or interior point method) method is given by:

Algorithm 3 Barrier (or interior point) method.

$$\begin{aligned} x_{k+1} &\in \underset{x\in\mathbb{R}^d : Ax=b}{\arg\min}\; t_k f(x) + \phi(x), \\ t_{k+1} &= \mu t_k. \end{aligned}$$

As a stopping criterion, one can use $\frac{m}{t_k} \leqslant \epsilon$ since this ratio bounds the difference $f(x_{k+1}) - p^*$.

Remark 2.9.

1. The first step of a barrier algorithm is generally performed thanks to Newton method.
2. x_k is used to initialize the algorithm for solving Step 1 (warm start). This makes the all story faster and explains why only several iterations are needed in barrier methods.

Interior point methods are very reliable on small scale problems but are not workable for very large problems. First order methods seem to be the only option.

2.4 Primal methods

2.4.1 Gradient method

In this section, we consider a function $f\colon \mathbb{R}^d \to \mathbb{R}$, that is differentiable and convex, and we tackle the problem:

$$\underset{x \in \mathbb{R}^d}{\operatorname{minimize}}\ f(x).$$

The gradient descent is a simple algorithm to reach a minimizer of f. Suppose we are provided with an initial point $x_0 \in \mathbb{R}^d$. Then the gradient descent algorithm is:

Algorithm 4 Gradient descent.

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k),$$

where $\gamma_k > 0$ is a step size to be tuned.

The interpretation of the gradient descent method is minimizing a local quadratic approximation of f:

$$\forall x \in \mathbb{R}^d\colon \quad f(x) \approx f(x_k) + \nabla f(x_k)^\top (x - x_k) + \frac{1}{2\gamma_k}\|x - x_k\|_2^2,$$

where γ_k is unknown a priori. However, there are several manners to choose the step size γ_k. The first one is to consider it constant. In this case, we require f to be gradient Lipschitz in order to ensure convergence.

Theorem 2.10 (Convergence of gradient descent)**.** *Assume that f has a minimizer $x^\star \in \mathbb{R}^d$ and that the gradient of f is Lipschitz continuous with Lipschitz constant $L > 0$:*

$$\forall (x, y) \in (\mathbb{R}^d)^2\colon \quad \|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\|.$$

For a constant step size $\gamma_k = \frac{1}{L}$ ($\forall k \in \mathbb{N}$):

$$f(x_k) - f(x^\star) \leqslant \frac{L}{2k}\|x_0 - x^\star\|_2^2.$$

Proof. The proof relies on the fact that Lipschitz continuity of ∇f implies convexity of $\frac{L}{2}\|\cdot\|_2^2 - f$, which implies a quadratic upper bound:

$$\forall (x, y) \in (\mathbb{R}^d)^2\colon \quad f(y) \leqslant f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|_2^2.$$

□

Remark 2.11. The convergence analysis shows that there is a progress at each iteration. In other words, the gradient method is a descent method.

In addition, if f is strongly convex, convergence of gradient descent is faster.

Theorem 2.12 (Convergence of gradient descent (strong convexity))**.** *Assume that f has a minimizer $x^\star \in \mathbb{R}^d$, is μ-strongly convex ($\mu > 0$) and that the gradient of f is Lipschitz continuous with Lipschitz constant $L > 0$. For a constant step size $\gamma_k = \frac{2}{\mu+L}$ ($\forall k \in \mathbb{N}$):*

$$f(x_k) - f(x^\star) \leqslant c^k \frac{L}{2} \|x_0 - x^\star\|_2^2$$

and

$$\|x_k - x^\star\|^2 \leqslant c^k \|x_0 - x^\star\|_2^2,$$

where $c = \left(\frac{L/\mu - 1}{L/\mu + 1}\right)^2 \in (0,1)$.

Proof. The proof relies on the definition of strong convexity, which implies a quadratic lower bound:

$$\forall (x,y) \in \mathbb{R}^d)^2 : \quad f(y) \geqslant f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2} \|y - x\|_2^2.$$

□

Remark 2.13. The convergence rate of the gradient descent is not optimal: *fast* gradient methods have better convergence rates.

Another way to choose the step size is to perform an adaptive and local computation, called a line search.

Backtracking line search

Backtracking line search is also know as Armijo's rule:

Algorithm 5 Armijo's rule

Choose $\gamma_0 > 0$, $\alpha \in (0,1)$ and $\beta \in (0,1)$. Set $\gamma = \gamma_0$ and update γ with $\gamma \leftarrow \beta\gamma$ until

$$f(x_k - \gamma \nabla f(x_k)) < f(x_k) - \alpha\gamma \|\nabla f(x_k)\|_2^2.$$

For simplicity, we often take $\alpha = \frac{1}{2}$.

Theorem 2.14 (Convergence of gradient descent (backtracking))**.** *Assume that f has a minimizer $x^\star \in \mathbb{R}^d$ and that the gradient of f is Lipschitz continuous with Lipschitz constant $L > 0$. For a backtracking line search with same $\gamma_0 > 0$ and $\alpha = \frac{1}{2}$:*

$$f(x_k) - f(x^\star) \leqslant \frac{1}{2k \min(\gamma_0, \beta/L)} \|x_0 - x^\star\|^2.$$

Exact line search

Algorithm 6 Exact line search

```
Choose $\gamma_k$ such that
    $\gamma_k \in \arg\min_{\gamma \geqslant 0} f(x_k - \gamma \nabla f(x_k)).$
```

Advantages and drawbacks

Advantages of gradient descent are:

1. every iteration is inexpensive;
2. it does not require second order information (Hessian of f).

However, gradient descent

1. is often slow (oscillation);
2. does not handle nondifferentiable functions.

Other first-order methods address one or both disadvantages.
Methods with improved convergence:

- quasi-Newton methods;
- conjugate gradient method;
- accelerated gradient method

Methods for nondifferentiable or constrained problems:

- subgradient method;
- proximal gradient method;
- smoothing methods;
- cutting-plane methods.

2.4.2 Quasi-Newton method

This section deals with including second order information in gradient descent. For this purpose, let us assume that f is twice differentiable.

Algorithm 7 Newton method.

$$x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k).$$

The Newton method comes from minimizing a second-order approximation of f around x_k:

$$f(y) \approx f(x_k) + \nabla f(x_k)^\top (y - x_k) + \frac{1}{2}(y - x_k)^\top \nabla^2 f(x_k)(y - x_k).$$

If the Newton method demonstrates fast convergence, it has the disadvantage to be expensive for large scale applications. To overcome that, the Hessian can be approximated by a metric $H \in \mathbb{R}^{d \times d}$, that is symmetric positive definite.

Algorithm 8 Quasi-Newton method.

Given an initial point $x_0 \in \mathbb{R}^d$ and an initial metric $H_0 \in \mathbb{R}^{d \times d}$, that is symmetric positive definite, iterate

$$x_{k+1} = x_k - \gamma_k H_k^{-1} \nabla f(x_k),$$
$$\text{set } H_{k+1} \text{ based on } H_k,$$

where $\gamma_k > 0$ are step sizes, which can be chosen by line search. The second step of a quasi-Newton method can be done in several manners.

Newton method

Setting $H_k = \nabla^2 f(x_k)$ makes the last algorithm boiling down to a Newton method with adaptive step size.

Broyden-Fletcher-Goldfarb-Shanno (BFGS)

Setting $\Delta_x = x_{k+1} - x_k$ and $\Delta_y = \nabla f(x_{k+1}) - \nabla f(x_k)$, the BFGS update rule is:

$$H_{k+1} = H_k + \frac{1}{\Delta_x^\top \Delta_y} \Delta_y \Delta_y^\top - \frac{1}{\Delta_x^\top H_k \Delta_x} H_k \Delta_x \Delta_x^\top H_k.$$

Let us remark can the inverse can be computed efficiently:

$$H_{k+1}^{-1} = \left(I_d - \frac{1}{\Delta_x^\top \Delta_y} \Delta_x \Delta_y^\top \right) H_k^{-1} \left(I_d - \frac{1}{\Delta_x^\top \Delta_y} \Delta_y^\top \Delta_x \right) + \frac{1}{\Delta_x^\top \Delta_y} \Delta_x^\top \Delta_x,$$

where I_d is the identity matrix of size $d \times d$.

BFGS method converges for strongly convex functions (in that case $\Delta_x^\top \Delta_y > 0$).

Square root BFGS

Same as previously but with $H_k = L_k L_k^\top$ (Cholesky decomposition). The updates rule is:

$$L_{k+1} = L_k \left(I_d + \frac{1}{\tilde{\Delta}_x^\top \tilde{\Delta}_x} (\alpha \tilde{\Delta}_y - \tilde{\Delta}_x) \tilde{\Delta}_x^\top \right),$$

where $\tilde{\Delta}_x = L_k \Delta_x$, $\tilde{\Delta}_y = L_k^{-1} \Delta_y$ and $\alpha = \frac{\tilde{\Delta}_x^\top \tilde{\Delta}_x}{\Delta_x^\top \Delta_y}$.

Limited-memory BFGS (L-BFGS)

Leveraging the recursive formula of H_k^{-1}, we can compute a direction of descent $H_k^{-1} \nabla f(x_k)$ with only recursive updates of vectors. L-BFGS goes beyond this remark by truncating the recursion to the last m (often $m \approx 30$) iterations. This requires nevertheless to store the m last values of Δ_x and Δ_y.

2.4.3 Subgradient method

From now on, we no longer require f to be differentiable (but f is still convex). Subgradient method is certainly the simplest method for minimizing f. It is similar to gradient descent but replacing gradients by subgradients:

Algorithm 9 Subradient descent.

$$x_{k+1} = x_k - \gamma_k v_k,$$

where $v_k \in \partial f(x_k)$ and $\gamma_k > 0$ is a step size.

Remark 2.15. Contrarily to a negative gradient $-\nabla f(x_k)$, a negative subgradient $-v$ ($v \in \partial f(x_k)$) is not a direction of descent in general. This means that the subgradient method is not a descent method ($f(x_{k+1}) > f(x_k)$ can occur).

Akin to gradient descent, several step size rules coexist:

- fixed step: γ_k is constant;
- fixed length: $\gamma_k \|v_k\|_2 = \|x_k - x_{k-1}\|_2$;
- diminishing step: $\gamma_k \to 0$, with $\sum_{k=1}^{\infty} \gamma_k = \infty$.

For fixed step sizes and fixed length, the subgradient method does not converge. However, two cases are of interest: diminishing step sizes and fixed length for a given number of steps.

Remark 2.16. The convergence rate of subgradient descent is optimal (we can construct an optimization problem for which convergence is in $O(1/\sqrt{k})$.

Theorem 2.17 (Convergence of subgradient method)**.** *Assume that f has a minimizer $x^\star \in \mathbb{R}^d$ and that f is Lipschitz continuous with Lipschitz constant $L > 0$. For a diminishing step sizes $\gamma_k \to 0$, with $\sum_{k=1}^{\infty} \gamma_k = \infty$:*

$$\min_{0 \leqslant \ell \leqslant k} f(x_\ell) - f(x^\star) \leqslant \frac{\|x_0 - x^\star\|_2^2 + L^2 \sum_{\ell=1}^{k} \gamma_\ell^2}{2\sum_{\ell=1}^{k} \gamma_\ell}.$$

Proof. The proof relies on the fact that subgradients are bounded by L. □

Since $\frac{\sum_{\ell=1}^{k} \gamma_\ell^2}{2\sum_{\ell=1}^{k} \gamma_\ell} \to 0$, $\min_{0 \leqslant \ell \leqslant k} f(x_\ell)$ converges to $f(x^\star)$.

Theorem 2.18 (Convergence of subgradient method (fixed number of iterations))**.** *Assume that f has a minimizer $x^\star \in \mathbb{R}^d$ and that f is Lipschitz continuous with Lipschitz constant $L > 0$. Let $x_0 \in \mathbb{R}^d$ be an initial point close to a minimizer: $\|x_0 - x^\star\|_2 \leqslant R$, for $R > 0$. For a fixed step length: $\gamma_k \|v_{k-1}\|_2 = \frac{R}{\sqrt{k}}$:*

$$\min_{0 \leqslant \ell \leqslant k} f(x_\ell) - f(x^\star) \leqslant \frac{LR}{\sqrt{k}}.$$

In addition, any other step length increases the bound.

Remark 2.19. This convergence rate is optimal (it cannot be improved).

To sum up, subgradient descent:

1. handles nondifferentiable convex problems;
2. is an algorithm as simple as gradient descent;
3. has slow convergence;
4. does not provide easy stopping criterion.

2.4.4 Proximal gradient method

We have seen at the beginning of this class that optimization problems in machine learning are often of the form:

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\ f(x) + g(x), \tag{P14}$$

where $f\colon \mathbb{R}^d \to \mathbb{R}$ is a differentiable and convex function and $g\colon \mathbb{R}^d \to (-\infty, \infty]$ is a convex function. In this section, we leverage the special structure of this problem to introduce fast algorithms (compared to subgradient methods) even though the global function $f + g$ is not differentiable.

Definition 2.20 (Proximal operator)**.** *Let $h\colon \mathbb{R}^d \to [-\infty, \infty]$ be a proper, lower semi-continuous convex function. The proximal operator of h is defined by:*

$$\forall x \in \mathbb{R}^d : \operatorname{prox}_h(x) = \operatorname*{arg\,min}_{u \in \mathbb{R}^d} h(u) + \frac{1}{2}\|u - x\|_2^2.$$

(By strong convexity, the arg min *exists and is a singleton, so* prox_g *is well defined.)*

Example 2.21.

- *For $h = 0$,* $\operatorname{prox}_h(x) = x, \forall x \in \mathbb{R}^d$.
- *Let $\mathcal{C} \subset \mathbb{R}^d$ be a closed convex set and $h = \chi_{\mathcal{C}}$. Then* prox_h *is the orthogonal projector on $\mathcal{C}$.*
- *For $h = \|\cdot\|_1$,* prox_h *is the* soft-thresholding *operator:*

$$\forall x \in \mathbb{R}^d, \forall i \in [d]: \qquad \operatorname{prox}_h(x)_i = \begin{cases} x_i - 1 & \text{if } x_i \geqslant 1 \\ 0 & \text{if } |x_i| \leqslant 1 \\ x_i + 1 & \text{if } x_i \leqslant -1. \end{cases}$$

Proposition 2.22. *Let $h\colon \mathbb{R}^d \to [-\infty, \infty]$ be a proper, lower semi-continuous convex function. Let $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$. Then,*

$$y = \operatorname{prox}_h(x) \iff x - y \in \partial h(y).$$

Proof.

$$\begin{aligned} y = \operatorname{prox}_h(x) &\iff y \in \operatorname*{arg\,min}_{u \in \mathbb{R}^d} h(u) + \frac{1}{2}\|u - x\|_2^2 \\ &\iff 0 \in \partial h(y) + (y - x) \\ &\iff x - y \in \partial h(y). \end{aligned}$$

□

Theorem 2.23 (Moreau decomposition)**.** *Let $h\colon \mathbb{R}^d \to [-\infty, \infty]$ be a proper, lower semi-continuous convex function. Then*

$$\forall x \in \mathbb{R}^d: \qquad x = \operatorname{prox}_h(x) + \operatorname{prox}_{h^*}(x).$$

Proof. For any x, let $u = \operatorname{prox}_h(x)$. By definition, $x - u \in \partial h(u)$, thus $u \in \partial h^*(x - u)$, that is $x - (x - u) \in \partial h^*(x - u)$, which means that $x - u = \operatorname{prox}_{h^*}(x)$. □

The Moreau decomposition generalizes the decomposition by orthogonal projection on subspaces.

Proposition 2.24 (Nonexpansiveness of the proximal operator)**.** *Let* $h\colon \mathbb{R}^d \to [-\infty, \infty]$ *be a proper, lower semi-continuous convex function. Then* prox_h *is firmly nonexpansive:*

$$\forall (x, y) \in (\mathbb{R}^d)^2\colon \quad (\operatorname{prox}_h(y) - \operatorname{prox}_h(x))^\top (y - x) \geqslant \| \operatorname{prox}_h(y) - \operatorname{prox}_h(x)\|_2^2.$$

In addition, prox_h *is Lipschitz-continuous with parameter* 1*:*

$$\forall (x, y) \in (\mathbb{R}^d)^2\colon \quad \| \operatorname{prox}_h(y) - \operatorname{prox}_h(x)\|_2 \leqslant \|y - x\|.$$

Proof. For $u = \operatorname{prox}_h(x)$ and $v = \operatorname{prox}_h(y)$, we have $x - u \in \partial h(u)$ and $y - v \in \partial h(v)$. Then, from the subdifferential definition, we have $(x - u - y + v)^\top (u - v) \geqslant 0$. The second property comes from Cauchy-Schwarz inequality. □

The following algorithm is workable in the setting described previously, that is when:

1. $f\colon \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable;
2. $g\colon \mathbb{R}^d \to (-\infty, \infty]$ is convex with an *easy-to-compute* proximal operator.

Algorithm 10 Proximal gradient method.

$$x_{k+1} = \operatorname{prox}_{\gamma_k g} (x_k - \gamma_k \nabla f(x_k)),$$

where $\gamma_k > 0$ is a step size.

The interpretation of the proximal gradient method is very similar to the one of the gradient descent. It consists in minimizing a local quadratic approximation of f plus the original non-differentiable function g:

$$\begin{aligned}\forall x \in \mathbb{R}^d\colon \quad f(x) + g(x) &\approx f(x_k) + \nabla f(x_k)^\top (x - x_k) + \frac{1}{2\gamma_k}\|x - x_k\|_2^2 + g(x)\\ &= g(x) + \frac{1}{2\gamma_k}\|x - (x_k - \gamma_k \nabla f(x_k))\|_2^2 - \frac{\gamma_k}{2}\|\nabla f(x_k)\|_2^2,\end{aligned}$$

where γ_k is unknown a priori.

Example 2.25 (Soft-thresholding)**.** *When* $g = \|\cdot\|_1$*, we obtain the soft-thresholding method, where we first perform a gradient step* $x^+ = x_k - \gamma_k \nabla f(x_k)$*, and then a soft-thresholding:*

$$\forall i \in [d]\colon \quad (x_{k+1})_i = \begin{cases} x_i^+ - \gamma_k & \text{if } x_i \geqslant \gamma_k \\ 0 & \text{if } -\gamma_k \leqslant x_i \leqslant \gamma_k \\ x_i^+ + \gamma_k & \text{if } x_i \leqslant -\gamma_k. \end{cases}$$

Theorem 2.26 (Convergence of the proximal method)**.** *Consider Problem (P14) with* $f\colon \mathbb{R}^d \to \mathbb{R}$ *being differentiable with L-Lipschitz gradient* ($L > 0$)*, and* $g\colon \mathbb{R}^d \to (-\infty, \infty]$ *being proper,* *lower-semicontinuous and convex. Let us assume that $F = f + g$ has a minimizer $x^\star \in \mathbb{R}^d$. For a constant step size $\gamma_k = \frac{1}{L}$ ($\forall k \in \mathbb{N}$):*

$$F(x_k) - F(x^\star) \leqslant \frac{L}{2k}\|x_0 - x^\star\|_2^2.$$

In addition, if f is μ-strongly convex ($\mu > 0$), then:

$$\|x_k - x^\star\|_2^2 \leqslant c^k \|x_0 - x^\star\|_2^2,$$

where $c = 1 - \frac{m}{L} \in (0, 1)$.

Proof. The analysis is similar to the one of gradient descent, considering instead the direction of descent $d_k = \frac{1}{\gamma_k}(x_k - \operatorname{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k)))$, that is $x_k - \gamma_k d_k = \operatorname{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k))$.

In addition, let us remark that

$$\begin{aligned} d_k = 0 &\iff 0 = \frac{1}{\gamma_k}(x_k - \operatorname{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k))) \\ &\iff x_k = \operatorname{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k)) \\ &\iff (x_k - \gamma_k \nabla f(x_k)) - x_k \in \nabla(\gamma_k g)(x_k) \\ &\iff -\gamma_k \nabla f(x_k) \in \gamma_k \nabla g(x_k) \\ &\iff 0 \in \nabla f(x_k) + \partial g(x_k) \\ &\iff 0 \in \partial F(x_k). \end{aligned}$$

In other words, x_k is a minimizer of F *if and only if* $d_k = 0$. □

Remark 2.27. The convergence analysis shows that each proximal gradient iteration is a descent step. As a consequence, the proximal gradient method is a descent method.

Algorithm 11 Backtracking line search

Choose $\gamma_0 > 0$ and $\beta \in (0, 1)$. Set $\gamma = \gamma_0$ and update γ with $\gamma \leftarrow \beta\gamma$ until

$$f(x_k - \gamma d_\gamma) \leqslant f(x_k) - \gamma \nabla f(x_k)^\top d_\gamma + \frac{\gamma}{2}\|d_\gamma\|_2^2,$$

where $d_\gamma = \frac{1}{\gamma}(x_k - \operatorname{prox}_{\gamma g}(x_k - \gamma \nabla f(x_k)))$ is the direction of descent.

Theorem 2.28 (Convergence of the proximal method (backtracking))**.** *For a backtracking line search with same $\gamma_0 > 0$, the previous theorem holds replacing $\frac{1}{L}$ by $\min(\gamma_0, \beta/L)$ in the convergence rates.*

We can derive three special cases of the proximal gradient method:

1. when $g = 0$, the proximal gradient method is a gradient descent;

2. when $g = \chi_{\mathcal{C}}$ for a set $\mathcal{C} \subset \mathbb{R}^d$, the proximal method is a projected gradient descent;

3. when $f = 0$, we get the proximal point method.

2.4.5 Accelerated proximal gradient method

In this section, we analyze a method, called Nesterov's method, to accelerate the proximal gradient descent. The main trick of this method is to add a momentum term.

Starting with an initial $\lambda_0 = 0$ and initial points $x_0 = y_0 \in \mathbb{R}^d$, the accelerated proximal gradient method is:

Algorithm 12 Accelerated proximal gradient method.

$$\begin{aligned} x_{k+1} &= \operatorname{prox}_{\gamma_k g}(y_k - \gamma_k \nabla f(y_k)) \\ \lambda_{k+1} &= \frac{1 + \sqrt{1 + 4\lambda_k^2}}{2} \\ y_{k+1} &= x_{k+1} + \frac{\lambda_k - 1}{\lambda_{k+1}}(x_{k+1} - x_k), \end{aligned}$$

where $\gamma_k > 0$ is a step size.

Remark 2.29. In image processing and compressed sensing, this method is often called FISTA, for fast iterative shrinkage-thresholding algorithm.

As always, the step size may be set to $\frac{1}{L}$ or chosen by line search. Moreover, y_k is an extrapolated point where the proximal gradient step is performed.

Theorem 2.30 (Convergence of the accelerated proximal method)**.** *Consider Problem (P14) with $f \colon \mathbb{R}^d \to \mathbb{R}$ being differentiable with L-Lipschitz gradient ($L > 0$), and $g \colon \mathbb{R}^d \to (-\infty, \infty]$ being proper, lower-semicontinuous and convex. Let us assume that $F = f + g$ has a minimizer $x^\star \in \mathbb{R}^d$. For a constant step size $\gamma_k = \frac{1}{L}$ ($\forall k \in \mathbb{N}$):*

$$F(x_k) - F(x^\star) \leqslant \frac{2L}{k^2} \|x_0 - x^\star\|_2^2.$$

2.4.6 Douglas-Rachford method

Here, we focus on the optimization problem:

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\ f(x) + g(x), \tag{P15}$$

where $f \colon \mathbb{R}^d \to (-\infty, \infty]$ and $g \colon \mathbb{R}^d \to (-\infty, \infty]$ are two convex functions. Contrarily to the proximal gradient method, the Douglas-Rachford does not require f to be differentiable.

Starting from an initial point $y_0 \in \mathbb{R}^d$, the Douglas-Rachford algorithm reads:

Algorithm 13 Douglas-Rachford method.

$$\begin{aligned} x_k &= \operatorname{prox}_{\gamma_k f}(y_k) \\ y_{k+1} &= y_k + \mu_k \left(\operatorname{prox}_{\gamma_k g}(2x_k - y_k) - x_k\right), \end{aligned}$$

where $\gamma_k > 0$ is a step size (without restriction) and $\mu_k \in (0, 2)$.

Remark 2.31. Douglas-Rachford iteration can be written as fixed-point iteration:

$$y_{k+1} = y_k + \mu_k \left(\operatorname{prox}_{\gamma_k g}\left(2\operatorname{prox}_{\gamma_k f}(y_k) - y_k\right) - \operatorname{prox}_{\gamma_k f}(y_k)\right).$$

Defining the auxiliary mapping $x \in \mathbb{R}^d \mapsto \operatorname{rprox}_h(x) = 2\operatorname{prox}_h(x) - x$ for any proximable function $h \colon \mathbb{R}^d \to (-\infty, \infty]$, the fixed-point iteration also reads:

$$y_{k+1} = \left(1 - \frac{\mu_k}{2}\right) y_k + \frac{\mu_k}{2} \operatorname{rprox}_{\gamma_k g}\left(\operatorname{rprox}_{\gamma_k f}(y_k)\right).$$

The case where $\mu_k = 1$ ($\forall k \in \mathbb{N}$) is the usual Douglas-Rachford algorithm. When $\mu_k > 1$, it is an over-relaxation while when $\mu_k < 1$, we talk about under-relaxation. In practice, we usually consider $\mu_k = \gamma_k = 1$ ($\forall k \in \mathbb{N}$).

Theorem 2.32 (Convergence of Douglas-Rachford method)**.** *Consider Problem (P15) with $f \colon \mathbb{R}^d \to (-\infty, \infty]$ and $g \colon \mathbb{R}^d \to (-\infty, \infty]$ being two proper, lower semi-continuous and convex functions. Let us assume that $f + g$ has a minimizer in $\mathbb{R}^d$.*

For a fixed step size $\gamma_k = \gamma > 0$ and relaxation parameter $\mu_k \in [\underline{\mu}, \overline{\mu}]$ ($\forall k \in \mathbb{N}$), where $0 < \underline{\mu} \leqslant \overline{\mu} < 2$, the sequence $(x_k)_{k \in \mathbb{N}}$ generated by the Douglas-Rachford method converges to a minimizer of $f + g$.

2.5 Primal-dual methods

We have seen previously that the proximal gradient method, used for minimizing a composite objective function, reduces to:

1. the gradient method when $g = 0$;

2. the proximal point method when $f = 0$.

In this section, we exploit these tow simple algorithms with the dual problem (P9) to devise primal-dual methods.

2.5.1 Lagrange multipliers

We consider Problem (P8) and its dual Problem (P9) when $g = \chi_{\{b\}}$ ($b \in \mathbb{R}^p$), which boils down to the following primal:

$$\begin{aligned} \underset{x \in \mathbb{R}^d}{\text{minimize}} \quad & f(x) \\ \text{s.t.} \quad & Ax = b, \end{aligned} \tag{P16}$$

and dual:

$$\underset{\nu\in\mathbb{R}^p}{\text{minimize}}\ b^\top \nu + f^*(-A^\top \nu). \tag{P17}$$

Starting with an initial primal-dual point $(x_0, \nu_0) \in \mathbb{R}^d \times \mathbb{R}^p$, the method of Lagrange multipliers is:

Algorithm 14 Method of Lagrange multipliers.

```
$x_{k+1} \in \underset{x\in\mathbb{R}^d}{\arg\min}\, L(x, \nu_k)$
$\nu_{k+1} = \nu_k + \gamma_k(Ax_{k+1} - b).$
```

where $\gamma_k > 0$ is a step size (without restriction).

Proposition 2.33. *Let $h\colon \mathbb{R}^d \to (-\infty, \infty]$ be a convex, proper and lower semi-continuous function and $\mu > 0$. Then f is μ-strongly convex if and only if f^* is differentiable and ∇f^* is μ^{-1}-Lipschitz continuous.*

Theorem 2.34 (Method of Lagrange multipliers)**.** *If f is μ-strongly convex ($\mu > 0$), proper and lower semi-continuous, then the method of Lagrange multipliers for Problem (P16) is the gradient method applied to Problem (P17).*

In addition, if the Lagrangian of Problem (P16) has a saddle point and if $\gamma_k \leqslant \frac{\mu}{\sigma_A^2}$, where $\sigma_A > 0$ is the largest singular value of A, then $((x_k, \nu_k))_{k\in\mathbb{N}}$ converges to a saddle point of the Lagrangian.

Proof. By the previous proposition, f^* is differentiable. The gradient method is:

$$\begin{aligned}
&\nu_{k+1} = \nu_k + \gamma_k A\nabla f^*(-A^\top \nu_k) - \gamma_k b \\
\iff &\exists x_{k+1} \in \mathbb{R}^d : x_{k+1} = \nabla f^*(-A^\top \nu_k), \nu_{k+1} = \nu_k + \gamma_k(Ax_{k+1} - b) \\
\iff &\exists x_{k+1} \in \mathbb{R}^d : -A^\top \nu_k \in \partial f(x_{k+1}), \nu_{k+1} = \nu_k + \gamma_k(Ax_{k+1} - b) \\
\iff &\exists x_{k+1} \in \mathbb{R}^d : 0 \in \partial f(x_{k+1}) + A^\top \nu_k, \nu_{k+1} = \nu_k + \gamma_k(Ax_{k+1} - b) \\
\iff &\exists x_{k+1} \in \underset{x\in\mathbb{R}^d}{\arg\min}\, f(x) + \nu_k^\top(Ax - b), \nu_{k+1} = \nu_k + \gamma_k(Ax_{k+1} - b) \\
\iff &\exists x_{k+1} \in \underset{x\in\mathbb{R}^d}{\arg\min}\, L(x, \nu_k), \nu_{k+1} = \nu_k + \gamma_k(Ax_{k+1} - b),
\end{aligned}$$

where L is the Lagrangian of Problem (P16).

In addition, f^* is μ^{-1}-gradient Lipschitz, so $\nu \mapsto b^\top \nu + f^*(-A^\top \nu)$ is $\frac{\sigma_A^2}{\mu}$-gradient Lipschitz, where σ_A is the largest singular value of A. Therefore, the gradient descent with $\gamma_k \leqslant \frac{\mu}{\sigma_A^2}$ converges to a dual solution $\nu^\star$. Then, by continuity of ∇f^*, $(x_k)_{k\in\mathbb{N}}$ converges to $x^\star = \nabla f^*(-A^\top \nu^\star)$, that is $x^\star \in \arg\min_{x\in\mathbb{R}^d} L(x, \nu^\star)$. Thus $x^\star$ is a primal solution and $x^\star, \nu^\star)$ a saddle point. □

2.5.2 Augmented Lagrange multipliers

The proximal point method is obtained from the proximal gradient descent with $f = 0$. In a general context, the proximal point method is defined for minimizing a proper convex lower semi-continuous function $h\colon \mathbb{R}^d \to (-\infty, \infty]$:

Algorithm 15 Proximal point method.

$$x_{k+1} = \operatorname{prox}_{\gamma_k h}(x_k),$$

where $\gamma_k > 0$ is a step size (without restriction).

It is mainly a conceptual algorithm. Let us remark that the step size γ_k affects both the number of iterations to reach an ϵ-solution and the cost of prox-evaluations.

Definition 2.35 (Augmented Lagrangian function)**.** *Let $\gamma > 0$ be a parameter. The augmented Lagrangian function associated to Problem (P16) is:*

$$L_\gamma\colon (x, \nu) \in \mathbb{R}^d \times \mathbb{R}^p \mapsto f(x) + \nu^\top (Ax - b) + \frac{\gamma}{2}\|Ax - b\|_2^2.$$

Starting with an initial primal-dual point $(x_0, \nu_0) \in \mathbb{R}^d \times \mathbb{R}^p$, the augmented Lagrangian method is:

Algorithm 16 Augmented Lagrangian method.

$$\begin{aligned} x_{k+1} &\in \operatorname*{arg\,min}_{x \in \mathbb{R}^d} L_{\gamma_k}(x, \nu_k) \\ \nu_{k+1} &= \nu_k + \gamma_k(Ax_{k+1} - b), \end{aligned}$$

where $\gamma_k > 0$ is a step size (without restriction).

Theorem 2.36 (Augmented Lagrangian method)**.** *If f is convex, proper and lower semi-continuous, then the augmented Lagrangian method for Problem (P16) is the proximal point method applied to Problem (P17).*

As a consequence, $(\nu_k)_{k \in \mathbb{N}}$ converges to a dual solution.

Proof. Defining $h \colon \nu \in \mathbb{R}^p \mapsto b^\top \nu + f^*(-A^\top \nu)$, the proximal point method reduces to:

$$\begin{aligned}
& \nu_{k+1} = \operatorname{prox}_{\gamma_k h}(\nu_k) \\
\iff & \nu_{k+1} = \underset{\nu \in \mathbb{R}^p}{\arg\min}\, \gamma_k h(\nu) + \frac{1}{2}\|\nu - \nu_k\|_2^2 \\
\iff & 0 \in \gamma_k \partial h(\nu_{k+1}) + \nu_{k+1} - \nu_k \\
\iff & 0 \in -\gamma_k A \partial f^*(-A^\top \nu_{k+1}) + \gamma_k b + \nu_{k+1} - \nu_k \\
\iff & \exists x_{k+1} \in \partial f^*(-A^\top \nu_{k+1}) : \nu_{k+1} = \nu_k + \gamma_k(Ax_{k+1} - b) \\
\iff & \exists x_{k+1} \in \mathbb{R}^d : -A^\top \nu_{k+1} \in \partial f(x_{k+1}), \nu_{k+1} = \nu_k + \gamma_k(Ax_{k+1} - b) \\
\iff & \exists x_{k+1} \in \mathbb{R}^d : 0 \in \partial f(x_{k+1}) + A^\top \nu_{k+1}, \nu_{k+1} = \nu_k + \gamma_k(Ax_{k+1} - b) \\
\iff & \exists x_{k+1} \in \mathbb{R}^d : 0 \in \partial f(x_{k+1}) + A^\top \nu_k + \gamma_k A^\top(Ax_{k+1} - b), \nu_{k+1} = \nu_k + \gamma_k(Ax_{k+1} - b) \\
\iff & \exists x_{k+1} \in \underset{x \in \mathbb{R}^d}{\arg\min}\, f(x) + \nu_k^\top(Ax - b) + \frac{\gamma_k}{2}\|Ax - b\|_2^2 : \nu_{k+1} = \nu_k + \gamma_k(Ax_{k+1} - b) \\
\iff & \exists x_{k+1} \in \underset{x \in \mathbb{R}^d}{\arg\min}\, L_{\gamma_k}(x, \nu_k) : \nu_{k+1} = \nu_k + \gamma_k(Ax_{k+1} - b).
\end{aligned}$$

The assumption that f is convex, proper and lower semi-continuous is necessary to state that $x_{k+1} \in \partial f^*(-A^\top \nu_{k+1}) \iff -A^\top \nu_{k+1} \in \partial f(x_{k+1})$. □

2.5.3 Alternating direction method of multipliers

We consider Problem (P8) and its dual Problem (P9) when f and g are only proximable functions, which leads to primal:

$$\begin{aligned}
\underset{x \in \mathbb{R}^d, y \in \mathbb{R}^p}{\text{minimize}} \quad & f(x) + g(y) \\
\text{s.t.} \quad & Ax = y.
\end{aligned} \tag{P18}$$

and dual:

$$\underset{\nu \in \mathbb{R}^p}{\text{maximize}}\ -g^*(\nu) - f^*(-A^\top \nu). \tag{P19}$$

As a reminder, the augmented Lagrangian for Problem (P18) is:

$$L_\gamma(x, y, \nu) = f(x) + g(y) + \nu^\top(Ax - y) + \frac{\gamma}{2}\|Ax - y\|_2^2.$$

Starting with an initial primal-dual point $(x_0, y_0, \nu_0) \in \mathbb{R}^d \times \mathbb{R}^p \times \mathbb{R}^p$, the alternating direction method of multipliers is:

Algorithm 17 Alternating direction method of multipliers.

$$
\begin{aligned}
x_{k+1} &\in \operatorname*{arg\,min}_{x\in\mathbb{R}^d} L_{\gamma_k}(x, y_k, \nu_k) = \operatorname*{arg\,min}_{x\in\mathbb{R}^d} \left(f(x) + \nu_k^\top Ax + \frac{\gamma}{2}\|Ax - y_k\|_2^2 \right) \\
y_{k+1} &\in \operatorname*{arg\,min}_{y\in\mathbb{R}^p} L_{\gamma_k}(x_{k+1}, y, \nu_k) = \operatorname*{arg\,min}_{y\in\mathbb{R}^p} \left(g(y) - \nu_k^\top y + \frac{\gamma}{2}\|Ax_{k+1} - y\|_2^2 \right) \\
\nu_{k+1} &= \nu_k + \gamma_k(Ax_{k+1} - y_{k+1}),
\end{aligned}
$$

where $\gamma_k > 0$ is a step size (without restriction).

Theorem 2.37 (Alternating direction method of multipliers)**.** *If f and g are convex, proper and lower semi-continuous, then the alternating direction method of multipliers for Problem (P18) is the Douglas-Rachford method applied to Problem (P19).*

As a consequence, $(\nu_k)_{k\in\mathbb{N}}$ converges to a dual solution.

Proof. Let $F\colon \nu \in \mathbb{R}^p \mapsto g^*(\nu)$ and $G\colon \nu \in \mathbb{R}^p \mapsto f^*(-A^\top\nu)$. The Douglas-Rachford method with $\mu_k = 1$ and $\gamma_k = \gamma > 0$ applied to $F + G$ reads:

1. $\nu_{k+1} = \operatorname{prox}_{\gamma F}(\lambda_k)$;
2. $\delta_{k+1} = \operatorname{prox}_{\gamma G}(2\nu_{k+1} - \lambda_k)$;
3. $\lambda_{k+1} = \lambda_k + \delta_{k+1} - \nu_{k+1}$.

Using the properties:

1. $x = \operatorname{prox}_f(y) \iff y - x \in \partial f(x)$;
2. $x \in \partial f^*(y) \iff y \in \partial f(x)$ (f is convex, proper and lower semi-continuous, this is also true for g);
3. $y \in \partial f(x^\star) \iff x^\star \in \operatorname{arg\,min}_x f(x) - y^\top x$,

we get:

1. $\exists y_{k+1} \in \operatorname{arg\,min}_{y\in\mathbb{R}^p} g(y) - \lambda_k^\top y + \frac{\gamma}{2}\|y\|_2^2 : \nu_{k+1} = \lambda_k - \gamma y_{k+1}$;
2. $\exists x_{k+1} \in \operatorname{arg\,min}_x f(x) + \nu_{k+1}^\top Ax + \frac{\gamma}{2}\|Ax - y_{k+1}\|_2^2 : \delta_{k+1} = \nu_{k+1} + \gamma(Ax_{k+1} - y_{k+1})$;
3. $\lambda_{k+1} = \nu_{k+1} + \gamma Ax_{k+1}$.

Combining 1. and 3., we get

1. $\exists y_{k+1} \in \operatorname{arg\,min}_{y\in\mathbb{R}^p} g(y) - \nu_k^\top y + \frac{\gamma}{2}\|Ax_k - y\|_2^2 : \nu_{k+1} = \nu_k + \gamma(Ax_k - y_{k+1})$;
2. $x_{k+1} \in \operatorname{arg\,min}_x f(x) + \nu_{k+1}^\top Ax + \frac{\gamma}{2}\|Ax - y_{k+1}\|_2^2$ (no change).

Therefore, the iteration becomes:

1. $y_{k+1} \in \arg\min_{y \in \mathbb{R}^p} g(y) - \nu_k^\top y + \frac{\gamma}{2}\|Ax_k - y\|_2^2$;
2. $\nu_{k+1} = \nu_k + \gamma(Ax_k - y_{k+1})$;
3. $x_{k+1} \in \arg\min_x f(x) + \nu_{k+1}^\top Ax + \frac{\gamma}{2}\|Ax - y_{k+1}\|_2^2$.

Defining $\tilde{y}_k = y_{k+1}$ and $\tilde{\nu}_k = \nu_{k+1}$, we obtain:

1. $x_{k+1} \in \arg\min_x f(x) + \tilde{\nu}_k^\top Ax + \frac{\gamma}{2}\|Ax - \tilde{y}_k\|_2^2$;
2. $\tilde{y}_{k+1} \in \arg\min_{y \in \mathbb{R}^p} g(y) - \tilde{\nu}_k^\top y + \frac{\gamma}{2}\|Ax_{k+1} - y\|_2^2$;
3. $\tilde{\nu}_{k+1} = \tilde{\nu}_k + \gamma(Ax_{k+1} - \tilde{y}_{k+1})$.

□

References

[1] P. Bianchi, O. Fercoq, and A. Sabourin. *Lecture notes on optimization for machine learning.* University Paris-Saclay, Télécom ParisTech, 2016.

[2] S.P. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[3] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing.* Springer Science & Business Media, 2013.

[4] M. Grasmair. Minimizers of optimization problems. Technical report, Department of Mathematics, Norwegian University of Science and Technology, 2015.

[5] L. Vandenberghe. *Lecture notes on Optimization Methods for Large-Scale Systems.* University of California, Los Angeles, 2016.