# Mathematical Foundations of Data Sciences

Gabriel Peyré
CNRS & DMA
École Normale Supérieure
gabriel.peyre@ens.fr
https://mathematical-tours.github.io
www.numerical-tours.com

November 15, 2020

# Chapter 16

# Convex Analysis

The main references for this chapter are [10, 3]. This chapters uses different notations than the previous one, and we denote $f(x)$ a typical function to minimized with respect to the variable $x$. We discus here some important concepts from convex analysis for non-smooth optimization.

## 16.1 Basics of Convex Analysis

We consider minimization problems of the form

$$\min_{x \in \mathcal{H}} f(x) \tag{16.1}$$

over the finite dimension (Hilbertian) space $\mathcal{H} \overset{\text{def.}}{=} \mathbb{R}^N$, with the canonical inner product $\langle \cdot, \cdot \rangle$. Most of the results of this chapter extends to possibly infinite dimensional Hilbert space.

Here $f : \mathcal{H} \longrightarrow \bar{\mathbb{R}} \overset{\text{def.}}{=} \mathbb{R} \cup \{+\infty\}$ is a convex function. Note that we allow here $f$ to take the value $+\infty$ to integrate constraints in the objective, and the constraint set is thus the "domain" of the function

$$\mathrm{dom}(f) \overset{\text{def.}}{=} \{x \,;\, f(x) < +\infty\}.$$

A useful notation is the indicator function of a set $\mathcal{C} \subset \mathcal{H}$

$$\iota_{\mathcal{C}}(x) \overset{\text{def.}}{=} \begin{cases} 0 & \text{if} \quad x \in \mathcal{C}, \\ +\infty & \text{otherwise.} \end{cases}$$

### 16.1.1 Convex Sets and Functions

A convex set $\Omega \subset \mathcal{H}$ is such that

$$\forall\, (x, y, t) \in \mathcal{H}^2 \times [0, 1], \quad (1-t)x + ty \in \Omega.$$

A convex function is such that

$$\forall\, (x, y, t) \in \mathcal{H}^2 \times [0, 1], \quad f((1-t)x + ty) \leqslant (1-t)f(x) + tf(y) \tag{16.2}$$

and this is equivalent to its epigraph $\{(x, r) \in \mathcal{H} \times \mathbb{R} \,;\, r \geqslant f(x)\}$ being a convex set. Note that here we use $\leqslant$ as a comparison over $\bar{\mathbb{R}}$. The function $f$ is strictly convex if equality in (16.2) only holds for $t \in \{0, 1\}$. A set $\Omega$ being convex is equivalent to $\iota_{\mathcal{C}}$ being a convex function.

In the remaining part of this chapter, we consider convex function $f$ which are proper, i.e. such that $\mathrm{dom}(f) \neq \emptyset$, and that should be lower-semi-continuous (lsc), i.e. such that for all $x \in \mathcal{H}$,

$$\liminf_{y \to x} f(y) \geqslant f(x).$$

It is equivalent to $\mathrm{epi}(f)$ being a closed convex set. We denote $\Gamma_0(\mathcal{H})$ the set of proper convex lsc functions.

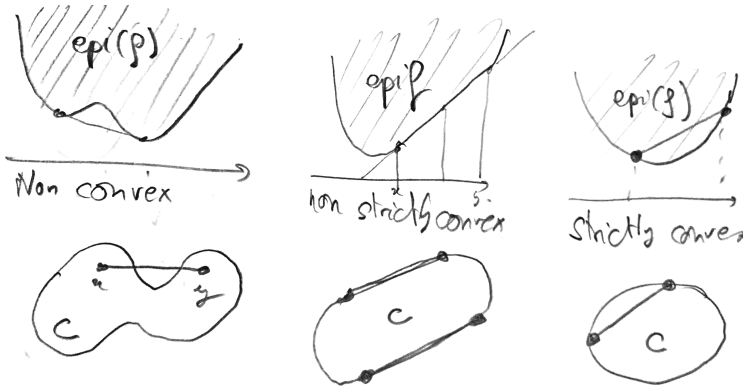


Figure 16.1: Convexity and strict convexity for function and sets.

### 16.1.2 First Order Conditions

**Existence of minimizers.** Before looking at first optimality conditions, one has to check that there exists minimizers, which is implied by the l.s.c. property and coercivity.

**Proposition 50.** *If $f$ is l.s.c. and coercive (i.e. $f(x) \to +\infty$ as $x \to +\infty$), then there exists a minimizer $x^\star$ of $f$.*

*Proof.* Since $f$ is coercive, it is bounded from below, one can consider a minimizing sequence $(x_n)_n$ such that $f(x_n) \to \min f$. Since $f$ is l.s.c., this implies that the sub-level set of $f$ are closed, and coercivity imply they are bounded, hence compact. One can thus extract from $(x_n)_n$ a converging sub-sequence $(x_{n(p)})_p$, $x_{n(p)} \to x^\star$. Lower semi-continuity implies that $\min f = \lim_p f(x_{n(p)}) \geqslant f(x^\star)$, and hence $x^\star$ is a minimizer. □

This existence proof is often called the "direct method of calculus of variation". Note that if the function $f$ is in $\Gamma_0(\mathcal{H})$, then the set of minimizer $\operatorname{argmin} f$ is a closed convex set, and all local minimizers (i.e. minimizer of the function restricted to an open ball) are global one. If it is furthermore strictly convex, then there is a single minimizer.

**Sub-differential.** The sub-differential at $x$ of such a $f$ is defined as

$$\partial f(x) \stackrel{\text{def.}}{=} \{u \in \mathcal{H}^* \,;\, \forall y, f(y) \geqslant f(x) + \langle u,\, y - x\rangle\}\,.$$

We denote here $\mathcal{H}^* = \mathbb{R}^N$ the set of "dual" vector. Although in finite dimensional Euclidean space, this distinction is not needed, it helps to distinguish primal from dual vectors, and recall that the duality pairing implicitly used depends on the choice of an inner product. The sub-differential $\partial f(x)$ is thus the set of "slopes" $u$ of tangent affine planes $f(x) + \langle u,\, z - x\rangle$ that fits below the graph of $f$.

Note that $f$ being differentiable at $x$ is equivalent to the sub-differential being reduced to a singleton (equal to the gradient vector)

$$\partial f(x) = \{\nabla f(x)\}.$$

Informally, the "size" of $\partial f(x)$ controls how smooth $f$ is at $x$.

Note that one can have $\partial f(x) = \emptyset$, for instance if $x \notin \operatorname{dom}(f)$. Note also that one can still have $x \in \operatorname{dom}(f)$ and $\partial f(x) = \emptyset$, for instance take $f(x) = -\sqrt{1-x^2} + \iota_{[-1,1]}(x)$ at $x = \pm 1$.

Since $\partial f(x) \subset \mathcal{H}^*$ is an intersection of half space, it is a closed convex set. The operator $\partial f : \mathcal{H} \mapsto 2^{\mathcal{H}^*}$ is thus "set-valued", and we often denote this as $\partial f : \mathcal{H} \hookrightarrow \mathcal{H}^*$.

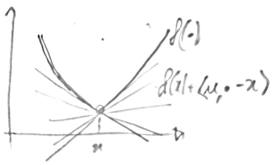

Figure 16.2: The subdifferential

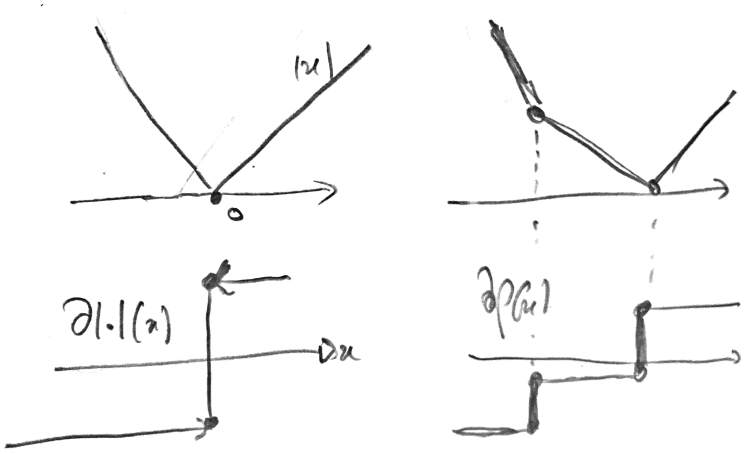


Figure 16.3: Subdifferential of the absolute value and a piecewise affine convex function.

*Remark* 8 (Maximally monotone operator). The operator $\partial f$ is particular instance of so-called monotone operator, since one can check that $U = \partial f$ satisfies

$$\forall\, (u,v) \in U(x) \times U(y), \quad \langle y-x,\, v-u \rangle \geqslant 0.$$

In the 1-D setting, being monotone is the same as being an increasing map. Sub-differential can also be shown to be maximally monotone, in the sense that such an operator is not striclty included in the graph of another monotone operator. Note that there exists monotone maps which are not subdifferential, for instance $(x,y) \mapsto (-y,x)$. Much of the theory of convex analysis and convex optimization can be extended to deal with arbitrary maximally monotone-maps in place of subdifferential, but we will not pursue this here.

A prototypical example is the absolute value $f(x) = |\cdot|$, and writing conveniently $\partial f(x) = \partial|\cdot|(x)$, one verifies that

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

**First Order Conditions.** The subdifferential is crucial for this simple but extremely important proposition.

**Proposition 51.** *$x^\star$ is a minimizer of $f$ is and only if $0 \in \partial f(x^\star)$.*

*Proof.* One has

$$x^\star \in \operatorname{argmin} f \quad \Leftrightarrow \quad (\forall y, f(x^\star) \leqslant f(y) + \langle 0,\, x^\star - y \rangle) \quad \Leftrightarrow \quad 0 \in \partial f(x^\star).$$

□

**Sub-differential calculus.** There is a large set of calculus rules that allows to simplify the computation of sub-differentials. For decomposable function $f(x_1,\ldots,x_K) = \sum_{k=1}^K f_k(x_k)$, the sub-differential is the product of the sub-differentials

$$\partial f(x_1,\ldots,x_K) = \partial f_1(x_1) \times \ldots \times \partial f_K(x_K).$$

This can be used to compute the sub-differential of the $\ell^1$ norm $\|x\|_1 = \sum_{k=1}^N |x_k|$

$$\partial\|\cdot\|_1(x) = \prod_{k=1}^N \partial|\cdot|(x_k)$$

which is thus an hyper rectangle. This means that, denoting $I = \mathrm{supp}(x)$, one has $u \in \partial\|\cdot\|_1(x)$ is equivalent to

$$u_I = \mathrm{sign}(x_I) \quad \text{and} \quad \|u_{I^c}\|_\infty \leqslant 1.$$

A tricky problem is to compute the sub-differential of the sum of two functions. If one of the two function is continuous at $x$ (i.e. it has a finite value), then

$$\partial(f+g)(x) = \partial f(x) \oplus \partial g(x) = \{u+v \;;\; (u,v) \in \partial f(x) \times \partial g(x)\}$$

where $\oplus$ thus denotes the Minkowski sum. For instance, if $f$ is differentiable at $x$, then

$$\partial(f+g)(x) = \nabla f(x) + \partial g(x) = \{\nabla f(x) + v \;;\; v \in \partial g(x)\}\,.$$

Positive linear scaling is simple to handle

$$\forall\, \lambda \in \mathbb{R}_+, \quad \partial(\lambda f)(x) = \lambda(\partial f(x)).$$

The chain rule for sub-differential is difficult since in general composition does not work so-well with convexity. The only simple case is composition with linear functions, which preserves convexity. Denoting $A \in \mathbb{R}^{P\times N}$ and $f \in \Gamma_0(\mathbb{R}^P)$, one has that $f \circ A \in \Gamma_0(\mathbb{R}^N)$ and

$$\partial(f\circ A)(x) = A^*(\partial f)(Ax) \overset{\text{def.}}{=} \{A^*u \;;\; u \in \partial f(Ax)\}\,.$$

**Normal cone.** The sud-differential of an indicator function is a convex cone, the so-called normal cone to the constraint

$$\forall\, x \in \mathcal{C}, \quad \partial\iota_{\mathcal{C}}(x) = \mathcal{N}_{\mathcal{C}}(x) \overset{\text{def.}}{=} \{v \;;\; \forall\, z \in \mathcal{C}, \langle z-x,\, v\rangle \leqslant 0\}\,.$$

Note that for $x \notin \mathcal{C}$, $\partial\iota_{\mathcal{C}}(x) = \emptyset$. For an affine space $\mathcal{C} = a + \mathcal{V}$ where $\mathcal{V} \subset \mathcal{H}$ is a linear space, then $\mathcal{N}_{\mathcal{C}}(x) = \mathcal{V}^\perp$ is the usual orthogonal for linear spaces. If $x \in \mathrm{int}(\mathcal{C})$ is in the interior of $\mathcal{C}$, then $\mathcal{N}_{\mathcal{C}}(x) = \{0\}$. In some sense, the more non-regular the boundary of $\mathcal{C}$ is at $x$, the larger is the normal cone.

The normal cone is a way to express first order condition for constrained problem

$$\min_{x\in\mathcal{C}} f(x)$$

which reads, if $f$ is continuous

$$0 \in \partial f(x) + \partial\iota_{\mathcal{C}}(x) \quad \Leftrightarrow \quad \exists \xi \in \partial f(x), -\xi \in \mathcal{N}_{\mathcal{C}}(x) \quad \Leftrightarrow \quad \partial f(x) \cap (-\mathcal{N}_{\mathcal{C}}(x)) \neq \emptyset.$$

If $f$ is differentiable, it reads $-\nabla f(x) \in \mathcal{N}_{\mathcal{C}}(x)$.

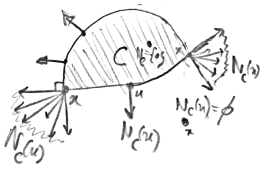

Figure 16.4: Normal cones

# 16.2 Convex Duality

Duality is associated to a particular formulation of the optimization problem, so that for instance making change of variables results in a different duality.

## 16.2.1 Lagrange Duality

We consider a minimization of the form

$$p^\star = \min_{x\in\mathbb{R}^N} \{f(x) \;;\; Ax = y \quad \text{and} \quad g(x) \leqslant 0\} \tag{16.3}$$

for a continuous convex functions $f : \mathcal{H} \to 0$, a matrix $A \in \mathbb{R}^{P\times N}$ and a function $g : \mathcal{H} \to \mathbb{R}^Q$ such that each of its coordinates $g_i : \mathcal{H} \to \mathbb{R}$ are continuous and convex. Note that it is always the case that equality in convex program corresponds to affine ones. One can always write a convex minimization problem with

positivity constraints in the form (16.3), although there exists infinite way of doing so (each one giving a different duality formula).

Here we have assumed for simplicity that $f$ is continuous, i.e. $\mathrm{dom}(f) = \mathbb{R}^N$. The following exposition can be generalized to $\mathrm{dom}(f)$ being arbitrary, but this is more technical. For the sake of simplicity, we thus assume all the constraint defining the domain are encoded in $Ax = y$ and $g(x) \leqslant 0$

Note that it is possible to generalized the previous Lagrange duality results by replacing "$x \geqslant 0$" by "$X \succeq 0$" where $X$ is a matrix (and in fact even more generally using convex cones).

We use the following fact

$$\sup_{u \in \mathbb{R}^P} \langle r,\, u \rangle = \begin{cases} 0 & \text{if} \quad r = 0, \\ +\infty & \text{if} \quad r \neq 0, \end{cases} \qquad \text{and} \qquad \sup_{v \in \mathbb{R}^Q_+} \langle s,\, v \rangle = \begin{cases} 0 & \text{if} \quad s \leqslant 0, \\ +\infty & \text{otherwise,} \end{cases}$$

to encode the constraints $r = Ax - y = 0$ and $s = g(x) \leqslant 0$.

One can represent the constraints appearing in (16.3) conveniently using a maximization over so-called Lagrange multipliers

$$p^\star = \inf_x \max_{u \in \mathbb{R}^P, v \in \mathbb{R}^Q_+} \mathcal{L}(x,u,v) \overset{\text{def.}}{=} f(x) + \langle Ax - y,\, u \rangle + \langle g(x),\, v \rangle.$$

It is tempting to inverse the inf and the sup, and study

$$d^\star = \sup_{(u,v) \in \mathbb{R}^P \times \mathbb{R}^Q_+} F(u,v) \overset{\text{def.}}{=} \inf_x f(x) + \langle Ax - y,\, u \rangle + \langle g(x),\, v \rangle. \tag{16.4}$$

One remarks that $F$ is a concave function (as being the minimum of linear forms), and this "dual" problem is thus a maximization of a concave function.

The following proposition is the so-called weak duality, which assert that values of the dual problems always lower bounds values of the primal one

**Proposition 52.** *One always has, for all $(u,v) \in \mathbb{R}^P \times \mathbb{R}^Q_+$,*

$$F(u,v) \leqslant p^\star \quad \Longrightarrow \quad d^\star \leqslant p^\star.$$

*Proof.* Since $g(x) \leqslant 0$ and $v \geqslant 0$, one has $\langle g(x),\, v \rangle \leqslant 0$, and since $Ax = y$, one has $\langle Ax - y,\, u \rangle = 0$, so that

$$\mathcal{L}(x,u,v) \leqslant f(x) \quad \Longrightarrow \quad F(u,v) = \inf_x \mathcal{L}(x,u,v) \leqslant \inf_x f(x) = p^\star.$$

□

The following fundamental theorem, more difficult to prove, gives a sufficient condition (so-called qualification of the constraints) such that one actually has equality.

**Theorem 26.** *If*

$$\exists x_0 \in \mathbb{R}^N, \quad Ax_0 = y \quad \textit{and} \quad g(x_0) < 0, \tag{16.5}$$

*then $p^\star = d^\star$. Furthermore, $x^\star$ and $(u^\star, v^\star)$ are solutions of respectively (16.3) and (16.4) if and only if*

$$Ax^\star = y, \quad g(x^\star) \leqslant 0, \quad u^\star \geqslant 0 \tag{16.6}$$

$$0 \in \partial f(x^\star) + A^* u^\star + \sum_i v_i^\star \partial g_i(x^\star) \tag{16.7}$$

$$\forall\, i, \quad u_i^\star g_i(x^\star) = 0 \tag{16.8}$$

The existence of such an $x_0$ is called "constraint qualification", and as written here, this corresponds to the so-called "Slater" qualification condition (many other weaker sufficient conditions exist).

Condition (16.6) is simply the primal and dual constraints. Condition (16.7) is the first order condition for the minimization of $\mathcal{L}(x,u,v)$ over $x$. Condition (16.8) is the first order condition for the maximization of $\mathcal{L}(x,u,v)$ over $(u,v)$. These three conditions are often referred to as "Karush-Kuhn-Tucker" (KKT) conditions, and under a constraint qualification condition, they are necessary and sufficient condition for optimality.

The last condition $u_i^\star g_i(x^\star)=0$ (so called "complementary slackness") states that if $g_i(x^\star)<0$ (the constraints is not saturated) then $u_i=0$, and also that if $u_i>0$ then $g_i(x^\star)=0$.

Note that it is possible to weaken the hypotheses of this theorem, for the linear constraints of the form $g_i(x)=\langle x,\,h_i\rangle-c_i\leqslant 0$, by replacing the $g_i(x_0)<0$ by the weaker condition $\langle x_0,\,h_i\rangle\leqslant c_i$.

One can generalize this theorem to the setting where $\mathrm{dom}(f)$ is not equal to $\mathbb{R}^N$ (i.e. it is not continuous, and thus integrates extra constraint beside the $\leqslant$). In this case, one has to add the extra constraint $x_0\in\mathrm{relint}(\mathrm{dom}(f))$.

Theorem 26 generalizes the necessary conditions provided by Lagrange multipliers for equality constrained optimization. The setting is both more complex because one can deal with inequalities that might be saturated (so this introduce positivity constraints on the multipliers $v$) but also simpler because of convexity (which thus gives also necessary conditions).

As a simple example, we now derive the dual for a simple linear projection problem. A more complicated computation is carried over in Section 10.1.5 for the Lasso. We consider

$$p^\star=\min_{Ax=y}\frac{1}{2}\|x-z\|^2=\min_x\max_u\frac{1}{2}\|x-z\|^2+\langle Ax-y,\,u\rangle=\max_u F(u)=\min_x\frac{1}{2}\|x-z\|^2+\langle Ax-y,\,u\rangle,$$

where we used the fact that strong duality holds because only linear constraints are involved. For each $u$, the optimal $x$ satisfies $x-z+A^*u$, i.e. $x=z-A^*u$, so that

$$F(u)=\frac{1}{2}\|A^*u\|^2+\langle A(z-A^*u)-y,\,u\rangle=-\frac{1}{2}\|A^*u\|^2+\langle u,\,Az-y\rangle.$$

Weak duality states $p^\star\geqslant F(u)$ for any $u$, and $p^\star=F(u^\star)$ where the optimal $u^\star$ satisfies $AA^*u=Az-y$. If $y\in\mathrm{Im}(A)$, then such a $u^\star$ exists and can be chosen as $u^\star=u=(AA^*)^{-1}(Az-y)$, and the (unique) primal solution reads

$$x^\star=\mathrm{Proj}_{A\cdot=y}(z)(\mathrm{Id}-A^+A)z-A^+y. \tag{16.9}$$

### 16.2.2 Legendre-Fenchel Transform

In order to simplify and accelerate computation involving Lagrange duality, it is very convenient to introduce a particular transformation of convex functions the Legendre-Fenchel transform. In some sense, it is the canonical "isomorphisms" (pairing) between convex functions. In spirit, is plays a similar role for convex function as the Fourier transform for signal or images.

For $f\in\Gamma_0(\mathcal{H})$, we define its Legendre-Fenchel transform as

$$f^*(u)\overset{\text{def.}}{=}\sup_x\langle x,\,u\rangle-f(x). \tag{16.10}$$

Being the maximum of affine functional, one obtains that $f^*$ is itself a convex function, and that in fact $f^\star\in\Gamma_0(\mathcal{H}^*)$. One can prove the following fundamental bi-duality result.

**Theorem 27.** *One has*

$$\forall\, f\in\Gamma_0(\mathcal{H}),\quad (f^*)^*=f.$$

In fact, $f^*$ is convex even in the case where $f$ is not, and $f^{**}$ is the convex envelop of $f$ (i.e. the largest convex function smaller than $f$). **[ToDo: drawing]**

One has the following basic property relating the sub-differentials of $f$ and $f^*$.

**Proposition 53.** *One has $\partial f^*=(\partial f)^{-1}$, where the inverse of a set valued map is defined in (17.10), and*

$$\forall\,(x,y),\quad \langle x,\,y\rangle\leqslant f(x)+f^*(y)\quad\textit{and}\quad\langle x,\,y\rangle=f(x)+f^*(y)\quad\Leftrightarrow\quad x\in\partial f^*(y)\quad\Leftrightarrow\quad y\in\partial f(x).$$

**Proposition 54.** *For $1/p + 1/q = 1$,*

$$(\iota_{\|\cdot\|_p \leqslant 1})^* = \|\cdot\|_q \quad \text{and} \quad (\|\cdot\|_q)^* = \iota_{\|\cdot\|_p \leqslant 1}$$

Let us now give some example of Legendre transform.

**Proposition 55.** *For $f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$ with $A$ inversible, then $f^*(u) = \frac{1}{2}\langle A^{-1}u, u\rangle - \frac{1}{2}\langle A^{-1}b, b\rangle$. In particular, for $f = \|\cdot\|^2/2$, then $f^* = f$. One has* ***[ToDo: check]***

$$f(\cdot - z)^* = f + \langle z, \cdot\rangle, \quad (f + \langle z, \cdot\rangle)^* = f(\cdot - z), \quad (\lambda f)^* = \lambda f^*(\cdot/\lambda).$$

*Proof.* One has $f^*(u) = \langle Ax^\star, x^\star\rangle - \langle b, x^\star\rangle$ where $x^\star$ solves

$$u = Ax^\star - b \quad \Longrightarrow \quad x^\star = A^{-1}u + A^{-1}b.$$

Hence

$$f^*(u) = \frac{1}{2}\langle AA^{-1}(u+b),\, A^{-1}(u+b)\rangle - \langle b,\, A^{-1}(u+b)\rangle = \frac{1}{2}\langle A^{-1}u, u\rangle - \frac{1}{2}\langle A^{-1}b, b\rangle$$

□

**Legendre transform and smoothness.** While the Fourier transform is a pairing between smoothness and decay (see Section **??**), the Legendre-Fenchel is really a pairing between smoothness and strong convexity. This can be intuitively seen by the fact that the Legendre-Fenchel inverts the sub-differentials (**??**) and hence when the functions involved are $\mathcal{C}^2$, it inverse the Hessians

$$\partial^2 f(x) = (\partial^2 f^*(y))^{-1} \quad \text{at} \quad y = \nabla f(x).$$

This relation between Hessian can be seen as implying the exchange of strong convexity and uniform bound on the Hessian, as detailed in Proposition 43.

**Proposition 56.** *One has*

$$\nabla f \text{ is } L\text{-Lipschitz} \quad \Longleftrightarrow \quad \nabla f^* \text{ is } \mu\text{-strongly convex.}$$

This results suggests a way to smooth any function $f$. Instead of doing a convolution, one can use the infimal convolution

$$(f \otimes g)(x) \stackrel{\text{def.}}{=} \sup_{y+y'=x} f(y) + g(y').$$

One can check that if $(f, g)$ are convex, so is $f \otimes g$, and that the Legendre transform actually exchanges sum and inf-convolution

$$(f+g)^* = f \otimes g \quad \text{and} \quad (f \otimes g)^* = f + g.$$

The Moreau-Yosida regularization of $f$ is corresponds to a $\mu$-strict-convexification of $f^*$, i.e.

$$f_\mu \stackrel{\text{def.}}{=} f \otimes (\frac{1}{2\mu}\|\cdot\|^2) = (f^* + \frac{\mu}{2}\|\cdot\|^2)^*. \tag{16.11}$$

Since $f^* + \frac{\mu}{2}\|\cdot\|^2$ is at least $\mu$-strongly convex, then $f_\mu$ as a $1/\mu$-Lipchitz gradient.

As an example, the Moreau-Yosida regularization of the absolute value reads

$$(|\cdot|_\mu)(x) = \begin{cases} \frac{1}{2\mu}x^2 & \text{if} \quad |x| \leqslant \mu, \\ |x| - \frac{\mu}{2} & \text{if} \quad |x| > \mu. \end{cases}$$

This should be compared with the regularization $\sqrt{x^2 + \mu^2}$ (which is above the curve) that we used previously. **[ToDo: add drawing]**

### 16.2.3 Fenchel-Rockafellar Duality

Very often the Lagrange dual can be expressed using the conjugate of the function $f$. We give here a particularly important example, which is often called Fenchel-Rockafellar Duality.

We consider the following structured minimization problem

$$p^\star = \inf_x \; f(x) + g(Ax). \tag{16.12}$$

Re-writing it as

$$\inf_{y=Ax} \; f(x) + g(y),$$

we can form the primal-dual problem

$$\inf_{(x,y)} \sup_u \; f(x) + g(y) + \langle Ax - y,\, u\rangle.$$

If sufficient condition on the domain of $(f, g)$ holds (such as those stated in Theorem **??**), one one can exchange the min and the max and obtains the dual problem

$$d^\star = \sup_u \min_{(x,y)} \; f(x) + g(y) + \langle Ax - y,\, u\rangle \tag{16.13}$$

$$= \sup_u \; \Big(\min_x \; \langle x,\, A^*u\rangle + f(x)\Big) + \Big(\min_y \; -\langle y,\, u\rangle + g(y)\Big) \tag{16.14}$$

which leads to the celebrated Fenchel-Rockafellar, which we summarize together with qualification sufficient condition ensuring strong duality.

**Theorem 28** (Fenchel-Rockafellar)**.** *If*

$$0 \in \operatorname{relint}(\operatorname{dom}(g)) - A \operatorname{relint}(\operatorname{dom}(f)) \tag{16.15}$$

*the one has the following strong duality*

$$\inf_x \; f(x) + g(Ax) = \inf_x \sup_u \; \mathcal{L}(x,u) = \sup_u \inf_x \; \mathcal{L}(x,u) = \sup_u \; -f^*(-A^*u) - g^*(u) \tag{16.16}$$

$$\text{where} \quad \mathcal{L}(x,u) \overset{\text{def.}}{=} f(x) + \langle Ax,\, u\rangle - g^*(u).$$

*Furthermore one has that $(x^\star, u^\star)$ is a pair of optimal primal-dual solutions if and only if*

$$-A^*u^\star \in \partial f(x^\star) \quad \text{and} \quad Ax^\star \in \partial g^*(u^\star). \tag{16.17}$$

Condition (16.15) is the constraint qualification ensuring that one can inverse the inf and the sup in (16.16). It can be recovered from Slater's qualification condition (16.5) when deriving the dual problem as in (16.13). The primal-dual relations (16.17) are the first order condition along the $x$ and the $u$ variables in minimization and maximization of $\mathcal{L}$. They are sometimes summarised in "matrix" form

$$0 \in \begin{pmatrix} \partial f & A^* \\ -A & \partial g^* \end{pmatrix} \begin{pmatrix} x^\star \\ u^\star \end{pmatrix}.$$

# Bibliography


[1] Amir Beck. *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*. SIAM, 2014.

[2] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[3] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[4] E. Candès and D. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise $C^2$ singularities. *Commun. on Pure and Appl. Math.*, 57(2):219–266, 2004.

[5] E. J. Candès, L. Demanet, D. L. Donoho, and L. Ying. Fast discrete curvelet transforms. *SIAM Multiscale Modeling and Simulation*, 5:861–899, 2005.

[6] A. Chambolle. An algorithm for total variation minimization and applications. *J. Math. Imaging Vis.*, 20:89–97, 2004.

[7] Antonin Chambolle, Vicent Caselles, Daniel Cremers, Matteo Novaga, and Thomas Pock. An introduction to total variation for image analysis. *Theoretical foundations and numerical methods for sparse recovery*, 9(263-340):227, 2010.

[8] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.

[9] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1999.

[10] Philippe G Ciarlet. Introduction à l'analyse numérique matricielle et à l'optimisation. 1982.

[11] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *SIAM Multiscale Modeling and Simulation*, 4(4), 2005.

[12] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. on Pure and Appl. Math.*, 57:1413–1541, 2004.

[13] D. Donoho and I. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, Dec 1994.

[14] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[15] M. Figueiredo and R. Nowak. An EM Algorithm for Wavelet-Based Image Restoration. *IEEE Trans. Image Proc.*, 12(8):906–916, 2003.

[16] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*, volume 1. Birkhäuser Basel, 2013.

[17] Stephane Mallat. *A wavelet tour of signal processing: the sparse way*. Academic press, 2008.

[18] D. Mumford and J. Shah. Optimal approximation by piecewise smooth functions and associated variational problems. *Commun. on Pure and Appl. Math.*, 42:577–685, 1989.

[19] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[20] Gabriel Peyré. *L'algèbre discrète de la transformée de Fourier*. Ellipses, 2004.

[21] J. Portilla, V. Strela, M.J. Wainwright, and Simoncelli E.P. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Trans. Image Proc.*, 12(11):1338–1351, November 2003.

[22] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, 1992.

[23] Otmar Scherzer, Markus Grasmair, Harald Grossauer, Markus Haltmeier, Frank Lenzen, and L Sirovich. *Variational methods in imaging*. Springer, 2009.

[24] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[25] Jean-Luc Starck, Fionn Murtagh, and Jalal Fadili. *Sparse image and signal processing: Wavelets and related geometric multiscale analysis*. Cambridge university press, 2015.