Topics in Statistical Learning Theory*

Lecture 6: Introduction to convex optimization

Contents

Convex optimization is the problem of finding (or rather approximating), through algorithmic procedures, minimizers of a convex function $F : \Theta \to \mathbb{R}$ defined on a convex set Θ. This lecture is the first, in this course, addressing this topic. While we will present algorithms and methods applicable in a broad range of applications, it is worth keeping in mind that the typical function we want to minimize in statistical learning is of the form

$$F(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f_\theta(X_i)) + \Omega(\theta),$$

where $\{(X_i, Y_i)\}_{i=1}^n$ is our learning sample of $\mathbb{R}^d \times \mathbb{R}$-valued labeled observations, where $\ell : \mathbb{R}^2 \to \mathbb{R}$ is a chosen loss function, where f_θ is a function $f_\theta : \mathbb{R}^d \to \mathbb{R}$ indexed by θ in some parameter space Θ and where $\Omega_n(\theta)$ denotes a regularization term. In this lecture, we review a few concepts from convex analysis that will be useful in the following lectures.

1 Differential calculus (survival guide)

1.1 Differentiable functions

Let U be an open[1] subset of $\mathbb{R}^k$. A function $f : U \to \mathbb{R}^\ell$ is said to be differentiable if, for every $x \in U$, there exists a linear function $Df(x) : \mathbb{R}^k \to \mathbb{R}^\ell$ such that, for all $h \in \mathbb{R}^k$ such that $x + h \in U$,

$$f(x+h) = f(x) + Df(x)(h) + \|h\|_2 \varepsilon(h),$$

where $\varepsilon : \mathbb{R}^k \to \mathbb{R}^\ell$ satisfies

$$\lim_{h \to 0} \|\varepsilon(h)\|_2 = 0.$$

If it exists, the linear function $Df(x) : \mathbb{R}^k \to \mathbb{R}^\ell$ is unique and called the differential of f at x.

*Teaching material can be found at https://www.qparis-math.com/teaching.

[1] Recall that a subset U of $\mathbb{R}^k$ is said to be open if, for all $x \in U$, there exists $r > 0$ such that $B(x, r) := \{y \in \mathbb{R}^k | \|x - y\|_2 < r\} \subset U$.

Remark 1.1. *Note that, for all $x \in U$, $Df(x)(h)$ is indeed defined for all $h \in \mathbb{R}^k$. To see this, note that since U is open, then for all $h \in \mathbb{R}^k$ there exists $t > 0$ small enough such that $x + th \in U$. Then, we have by definition of $Df(x)$ that*

$$\begin{aligned} f(x+th) &= f(x) + Df(x)(th) + t\|h\|_2 \varepsilon(th) \\ &= f(x) + tDf(x)(h) + t\|h\|_2 \varepsilon(th), \end{aligned}$$

where the last line follows from linearity of $Df(x)$. In particular, we deduce that for all $h \in \mathbb{R}^k$, and small enough $t > 0$,

$$Df(x)(h) = \frac{f(x+th) - f(x)}{t} - \|h\|_2 \varepsilon(th),$$

from which it follows that, for all $x \in U$ and all $h \in \mathbb{R}^k$,

$$Df(x)(h) = \lim_{t \to 0, t \neq 0} \frac{f(x+th) - f(x)}{t}. \tag{1.1}$$

Remark 1.2. *The above remark shows why it is important for the domain U of function f to be open. Whenever we consider a function $f : \Theta \to \mathbb{R}^\ell$ defined on a non open set $\Theta \subset \mathbb{R}^k$, it isn't clear a priori what it means for f to be differentiable. The convention in this case is very simple: we say that $f : \Theta \to \mathbb{R}^\ell$ is differentiable iff there exists an open set $U \supset \Theta$ such that f can in fact be defined on U and such that $f : U \to \mathbb{R}^\ell$ is differentiable as defined above. The differential $Df(x)$ of f at every $x \in \Theta$ is then defined, without ambiguity, as in (1.1)*

Example 1.3. *Whenever $k = 1$, then for any $x \in U$ and any $h \in \mathbb{R}$, we recover the more familiar formula*

$$Df(x)(h) = hf'(x),$$

where

$$f'(x) = \lim_{t \to 0, t \neq 0} \frac{f(x+t) - f(x)}{t} \in \mathbb{R}^\ell.$$

Example 1.4. *Suppose that, for all $x \in \mathbb{R}^k$, $f(x) = Ax + b$ for a matrix $A \in \mathbb{R}^{\ell \times k}$ and a vector $b \in \mathbb{R}^\ell$. Then $f : \mathbb{R}^k \to \mathbb{R}^\ell$ is differentiable and, for all $x, h \in \mathbb{R}^k$,*

$$Df(x)(h) = Ah.$$

1.2 Sum and composition of differentiable functions

Suppose $f : U \to \mathbb{R}^\ell$ is of the form

$$f = \sum_{i=1}^{n} \alpha_i f_i,$$

where $\{\alpha_i\}_{i=1}^n$ are real numbers and each $f_i : U \to \mathbb{R}^\ell$ is a differentiable function. Then, for every $x \in U$ and $h \in \mathbb{R}^k$, one easily checks that

$$Df(x)(h) = \sum_{i=1}^{n} \alpha_i Df_i(x)(h).$$

Now suppose that $h : U \subset \mathbb{R}^k \to \mathbb{R}^m$ and $g : V \subset \mathbb{R}^m \to \mathbb{R}^\ell$ are two differentiable functions such that $h(U) \subset V$. Then the function $f = g \circ h : U \to \mathbb{R}^\ell$ is differentiable and, for all $x \in U$,

$$Df(x) = Dg(h(x)) \circ Dh(x).$$

This formula is known as the *chain rule*.

Example 1.5. *Let $A \in \mathbb{R}^{m\times k}$, $b \in \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^\ell$ be differentiable. Define $f : \mathbb{R}^k \to \mathbb{R}^\ell$ by $f(x) = g(Ax+b)$. Then f is differentiable and, for all $x, h \in \mathbb{R}^k$,*

$$Df(x)(h) = Dg(Ax + b)(Ah).$$

Example 1.6. *As a particular case of Example 1.5 (case $m = \ell = 1$), consider $a \in \mathbb{R}^k$, $b \in \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ differentiable. Define $f : \mathbb{R}^k \to \mathbb{R}$ by $f(x) = g(a^\top x + b)$. Then f is differentiable and, for all $x, h \in \mathbb{R}$,*

$$Df(x)(h) = g'(a^\top x + b)a^\top h.$$

Example 1.7. *As another particular case of Example 1.5 (case $k = \ell = 1$), consider $a \in \mathbb{R}^m$, $b \in \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}$ differentiable. Define $f : \mathbb{R} \to \mathbb{R}$ by $f(t) = g(at + b)$. Then f is differentiable and, for all $t, h \in \mathbb{R}$,*

$$Df(t)(h) = hf'(t) = Dg(at + b)(ah) = hDg(at + b)(a),$$

and in particular

$$f'(t) = Dg(at + b)(a).$$

Example 1.8. *In the context of least-squares regression, consider the function $F : \mathbb{R}^k \to \mathbb{R}$ defined, for all $\theta = (\theta_1, \dots, \theta_k) \in \mathbb{R}^k$, by*

$$F(\theta) = \frac{1}{n}\sum_{i=1}^n (Y_i - \sum_{j=1}^k \theta_j f_j(X_i))^2 + \Omega(\theta),$$

where $\{(X_i, Y_i)\}_{i=1}^n$ is our learning sample of $\mathbb{R}^d \times \mathbb{R}$-valued labeled observations, where $f_1, \dots, f_k : \mathbb{R}^d \to \mathbb{R}$ are fixed references functions and where $\Omega(\theta)$ denotes a regularization term. Supposing that Ω is differentiable, and denoting for all $x \in \mathbb{R}^d$

$$\mathbf{f}(x) := (f_1(x), \dots, f_k(x))^\top \in \mathbb{R}^k,$$

F is differentiable and, for all $\theta, h \in \mathbb{R}^k$,

$$DF(\theta)(h) = -\frac{2}{n}\sum_{i=1}^n (Y_i - \theta^\top \mathbf{f}(X_i))\mathbf{f}(X_i)^\top h + D\Omega(\theta)(h).$$

Example 1.9. *Similarly, in the context of the convex approach to binary classification, consider the function $F : \mathbb{R}^k \to \mathbb{R}$ defined, for all $\theta = (\theta_1, \dots, \theta_k) \in \mathbb{R}^k$, by*

$$F(\theta) = \frac{1}{n}\sum_{i=1}^n \varphi(-Y_i \sum_{j=1}^k \theta_j f_j(X_i)) + \Omega(\theta),$$

where $\{(X_i, Y_i)\}_{i=1}^n$ is our learning sample of $\mathbb{R}^d \times \{-1, +1\}$-valued labeled observations, where $f_1, \dots, f_k : \mathbb{R}^d \to \{-1, +1\}$ are fixed hard classifiers and where $\Omega(\theta)$ denotes a regularization term. Supposing that $\varphi : \mathbb{R} \to \mathbb{R}$ and Ω are differentiable, and denoting for all $x \in \mathbb{R}^d$

$$\mathbf{f}(x) := (f_1(x), \dots, f_k(x))^\top \in \{-1, 1\}^k,$$

F is differentiable and, for all $\theta, h \in \mathbb{R}^k$,

$$DF(\theta)(h) = -\frac{1}{n}\sum_{i=1}^n \varphi'(-Y_i\theta^\top \mathbf{f}(X_i))Y_i\mathbf{f}(X_i)^\top h + D\Omega(\theta)(h).$$

1.3 Gradient

Let $U \subset \mathbb{R}^k$ be open and $f : U \to \mathbb{R}$ be differentiable. Introduce the canonical basis $e_1, \dots, e_k$ of $\mathbb{R}^k$ so that any $h = (h_1, \dots, h_k) \in \mathbb{R}^k$ writes (in a unique way)

$$h = \sum_{j=1}^k h_j e_j.$$

Then, for any $x \in U$, we deduce by linearity of $Df(x)$ that

$$Df(x)(h) = \sum_{j=1}^k h_j Df(x)(e_j). \tag{1.2}$$

It is classical to denote

$$\frac{\partial f}{\partial x_j}(x) := Df(x)(e_j),$$

which is called the partial derivative of f at x with respect to the j-th coordinate. It follows from (1.1) that

$$\frac{\partial f}{\partial x_j}(x) = \lim_{t\to 0, t\neq 0} \frac{f(x + te_j) - f(x)}{t}.$$

The gradient $\nabla f(x)$ of f at x is the vector of all partial derivatives of f at x, i.e.

$$\nabla f(x) := \left(\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_k}(x)\right)^\top \in \mathbb{R}^k.$$

Note finally that, for all $x \in U$ and all $h \in \mathbb{R}^k$, equation (1.2) reads

$$Df(x)(h) = h^\top \nabla f(x).$$

Example 1.10. *In the context of Example 1.5, when ever $\ell = 1$,*

$$\nabla f(x) = A^\top \nabla g(Ax + b).$$

Example 1.11. *In the context of Example 1.6,*

$$\nabla f(x) = g'(a^\top x + b)a.$$

Example 1.12. *In the context of Example 1.7,*

$$f'(t) = a^\top \nabla g(at + b).$$

Example 1.13. *In the context of Example 1.8,*

$$\nabla F(\theta) = -\frac{2}{n}\sum_{i=1}^n (Y_i - \theta^\top \mathbf{f}(X_i))\mathbf{f}(X_i) + \nabla\Omega(\theta).$$

Example 1.14. *In the context of Example 1.9,*

$$\nabla F(\theta) = -\frac{2}{n}\sum_{i=1}^n \varphi'(-Y_i\theta^\top \mathbf{f}(X_i))Y_i\mathbf{f}(X_i) + \nabla\Omega(\theta).$$

1.4 Interpretation of the gradient

The gradient of a differentiable function $f : U \subset \mathbb{R}^k \to \mathbb{R}$ benefits from a fundamental physical interpretation, quite basic to many optimization algorithms. The next result formalizes the following fact:

"The vector $-\nabla f(x)$ points in the direction of fastest immediate decrease of f at x."

Theorem 1.15. *Let $U \subset \mathbb{R}^k$ be open, $f : U \to \mathbb{R}$ be differentiable and $x \in U$. For any $v \in \mathbb{R}^k$, with $\|v\|_2 = 1$, set*

$$f_v(t) := f(x + tv),$$

which is well define for $t \in \mathbb{R}$ close enough to 0. Then, if $\nabla f(x) \neq 0$, $f_v'(0)$ is minimized for

$$v^* = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}.$$

Proof. We known from example 1.12 that

$$f_v'(0) = v^\top \nabla f(x).$$

In particular, it follows from Cauchy-Schwarz's inequality that, for every $v \in \mathbb{R}^k$ with $\|v\|_2 = 1$,

$$f_v'(0) \geq -\|\nabla f(x)\|_2.$$

Note finally that this lower bound is achieved for $v = v^*$. □

1.5 Taylor's formula and consequences

We'll often use the following version of Taylor's formula.

Theorem 1.16. *Let $U \subset \mathbb{R}^k$ be open and $f : U \to \mathbb{R}$ be differentiable. Let $x, y \in U$. Then,*

$$f(y) = f(x) + \int_0^1 (y - x)^\top \nabla f((1 - t)x + ty)\, \mathrm{d}t.$$

We may deduce from this formula the following. Recall that a function $f : U \to \mathbb{R}$ is called L-Lipschitz if, for all $x, y \in U$,

$$|f(x) - f(y)| \leq L\|x - y\|_2.$$

Theorem 1.17. *Let $U \subset \mathbb{R}^k$ be open and $f : U \to \mathbb{R}$ be differentiable. Then f is L-Lipschitz iff, for all $x \in U$,*

$$\|\nabla f(x)\|_2 \leq L.$$

Proof. Suppose that, for all $x \in U$,

$$\|\nabla f(x)\|_2 \leq L.$$

Then it follows from Taylor's formula that, for all $x, y \in U$,

$$\begin{aligned} |f(x) - f(y)| &\leq \Big(\sup_{t \in [0,1]} \|\nabla f((1 - t)x + ty)\|_2\Big) \|x - y\|_2 \\ &\leq L\|x - y\|_2. \end{aligned}$$

Conversely, suppose that $f : U \to \mathbb{R}$ is differentiable and L-Lipschitz. Then, since for all $x \in U$ and $h \in \mathbb{R}^k$ we have

$$h^\top \nabla f(x) = \lim_{t \to 0, t \neq 0} \frac{f(x + th) - f(x)}{t},$$

we get that

$$|h^\top \nabla f(x)| \leq L\|h\|_2.$$

Taking $h = \nabla f(x)$, we get

$$\|\nabla f(x)\|_2 \leq L.$$

□

2 Convex sets

A set $\Theta \subset \mathbb{R}^k$ is said to be convex if, for all $x, y \in \Theta$ and all $\lambda \in [0, 1]$,

$$(1 - \lambda)x + \lambda y \in \Theta.$$

This section lists some basic properties of convex sets.

Theorem 2.1 (Separation)**.** *Let $\Theta \subset \mathbb{R}^k$ be a closed convex set and $x_0 \in \mathbb{R}^k \setminus \Theta$. Then there exists $u \in \mathbb{R}^k$ and $t \in \mathbb{R}$ such that $u^\top x_0 < t$ and $\forall x \in \Theta, u^\top x \geq t$.*

The previous result means that point $x_0 \notin \Theta$ is separated from Θ by the affine hyperplane $\{x \in \mathbb{R}^k : u^\top x = t\}$. If Θ is not closed, we can only guarantee the existence of $u \in \mathbb{R}^k$ such that, $u^\top x_0 \leq u^\top x$ for all $x \in \Theta$. The next result follows from the separation theorem.

Theorem 2.2 (Supporting hyperplane)**.** *Let $\Theta \subset \mathbb{R}^k$ be a convex set and $x_0 \in \partial\Theta$ be a point on its boundary. Then, there exists $u \in \mathbb{R}^k$, $u \neq 0$, such that for all $x \in \Theta$, $u^\top x_0 \leq u^\top x$.*

For most of what we'll see next, an important notion is that of the projection onto a closed and convex set.

Theorem 2.3. *Let $\Theta \subset \mathbb{R}^k$ be a closed and convex set. Then, for all $x \in \mathbb{R}^k$, there exists a unique point $\Pi_\Theta(x) \in \Theta$ solving*

$$\|\Pi_\Theta(x) - x\|_2 = \min_{y \in \Theta} \|y - x\|_2.$$

The point $\Pi_\Theta(x)$ is called the projection of x onto Θ. In addition, $\Pi_\Theta(x)$ is the only point in Θ such that,

$$\forall y \in \Theta, \quad (x - \Pi_\Theta(x))^\top (y - \Pi_\Theta(x)) \leq 0.$$

3 Convex functions

Given a convex set $\Theta \subset \mathbb{R}^k$, a function $f : \Theta \to \mathbb{R}$ is convex if, for all $x, y \in \Theta$ and for all $\lambda \in [0, 1]$,

$$f((1 - \lambda)x + \lambda y) \leq (1 - \lambda)f(x) + \lambda f(y).$$

One checks that the function $f : \Theta \to \mathbb{R}$ is convex if and only if the epigraph of f, i.e. the set

$$\mathrm{epi}(f) = \{(x, t) \in \Theta \times \mathbb{R} : f(x) \leq t\},$$

is a convex subset of $\mathbb{R}^k \times \mathbb{R}$.

Definition 3.1 (Subgradients)**.** *Given a set $\Theta \subset \mathbb{R}^k$ and a function $f : \Theta \to \mathbb{R}$, a vector $g \in \mathbb{R}^k$ is called a subgradient of f at $x \in \Theta$ if,*

$$\forall y \in \Theta, \quad f(y) - f(x) \geq g^\top (y - x).$$

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

Theorem 3.2. *Let $\Theta \subset \mathbb{R}^k$ be a convex set and $f : \Theta \to \mathbb{R}$ be a function.*
(1) The function f is convex if, for all $x \in \Theta$, $\partial f(x) \neq \emptyset$.
(2) If f is convex then, for all $x \in \mathrm{int}(\Theta)$, $\partial f(x) \neq \emptyset$.
(3) If f is convex and differentiable, then for all $x \in \mathrm{int}(\Theta)$, $\partial f(x) = \{\nabla f(x)\}$.
(4) If f is convex, then for all $x, y \in \mathrm{int}(\Theta)$, all $g_x \in \partial f(x)$ and all $g_y \in \partial f(y)$,

$$(g_x - g_y)^\top (x - y) \geq 0.$$

Proof. (1) Let $x, y \in \Theta$ and $\lambda \in [0,1]$. Since there exists $g \in \partial f((1-\lambda)x+\lambda y)$, it follows by definition of a subgradient that

$$f(x) - f((1-\lambda)x + \lambda y) \geq \lambda g^\top (y - x),$$

and

$$f(y) - f((1-\lambda)x + \lambda y) \geq (1-\lambda) g^\top (x - y).$$

Multiplying the first inequality by $(1-\lambda)$, the second by λ and summing the obtained inequalities, we obtain that $f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$. Since this holds for all $x, y \in \Theta$ and all $\lambda \in [0,1]$, we deduce that f is convex.
(2) Let $x \in \Theta$. The point $(x, f(x))$ belongs to $\partial \mathrm{epi}(f)$. Since $\mathrm{epi}(f)$ is a convex set, we deduce from Theorem 2.2 that there exists $(a, b) \in \mathbb{R}^k \times \mathbb{R}$, $(a, b) \neq (0, 0)$, such that

$$\forall (y, t) \in \mathrm{epi}(f), \quad a^\top x + b f(x) \geq a^\top y + bt. \tag{3.1}$$

Observe that $(y, t) \in \mathrm{epi}(f)$ implies that $(y, t') \in \mathrm{epi}(f)$ for all $t' \geq t$. Hence, for any $y \in \Theta$ the above inequality should hold true for any $t \geq f(y)$ and in particular when $t \to +\infty$ which imposes that $b \leq 0$. Now suppose that $x \in \mathrm{int}(\Theta)$. Then, for $\varepsilon > 0$ small enough, the point $z = x + \varepsilon a$ belongs to Θ so that, for all $t \geq f(z)$,

$$a^\top x + b f(x) \geq a^\top z + bt \Leftrightarrow b f(x) \geq \varepsilon \|a\|^2 + bt.$$

If $b = 0$ we deduce that $a = 0$ which is a contradiction. Hence $b < 0$. Now for any $y \in \Theta$, writing (3.1) for $t = f(y)$ implies that

$$f(y) - f(x) \geq \frac{a^\top (y - y)}{|b|},$$

which shows that $a/|b| \in \partial f(x)$.
(3) Suppose that f is convex, differentiable and take $x \in \mathrm{int}(\Theta)$. For any $h \in \mathbb{R}^k$ and $t \in \mathbb{R}$ small enough so that both $x \pm th \in C$, a Taylor expansion of f around x reveals that

$$f(x) = f(x) \pm t \nabla f(x)^\top h + \mathrm{o}(t).$$

Now for any $g \in \partial f(x)$, we have by definition of a subgradient that

$$f(x \pm th) \geq f(x) \pm t g^\top h.$$

In particular, we deduce that

$$\pm t \nabla f(x)^\top h + \mathrm{o}(t) \geq \pm t g^\top h.$$

This imposes finally that, for all $h \in \mathbb{R}^k$, $\nabla f(x)^\top h = g^\top h$ which implies that $g = \nabla f(x)$.
(4) For all $x, y \in \mathrm{int}(\Theta)$, all $g_x \in \partial f(x)$ and all $g_y \in \partial f(y)$, summing the inequalities $f(x) - f(y) \geq g_y^\top (x - y)$ and $f(y) - f(x) \geq g_x^\top (y - x)$ easily provides the last property. □

Theorem 3.3 (First order optimality condition)**.** *Let $\Theta \subset \mathbb{R}^k$ be a convex set and $f : \Theta \to \mathbb{R}$ be a convex function. Then*

$$x^* \in \arg\min_{x \in \Theta} f(x) \quad \Leftrightarrow \quad 0 \in \partial f(x^*).$$

Proof. Both conditions are equivalent to the fact that $f(x) \geq f(x^*) + 0^\top (x - x^*)$, for all $x \in \Theta$. □

Theorem 3.4. *Let $\Theta \subset \mathbb{R}^k$ be an open convex set and $f : \Theta \to \mathbb{R}$ be a convex function. Then f is L-Lipschitz if and only if, for all $x \in \Theta$ and all $g \in \partial f(x)$, $\|g\|_2 \leq L$.*