6th Traditional Summer School for Young Researchers: "Control. Information. Optimization"

Giuseppe Carlo Calafiore

Dipartimento di Automatica e Informatica
Politecnico di Torino – ITALY

Moscow, June 22-29

LECTURE 1

Subgradients

The gradient does not exist, implying that the function may have kinks or corner points, and thus cannot be approximated locally by a tangent hyperplane, or by a quadratic approximation. Directional derivatives still exist because of the convexity property.

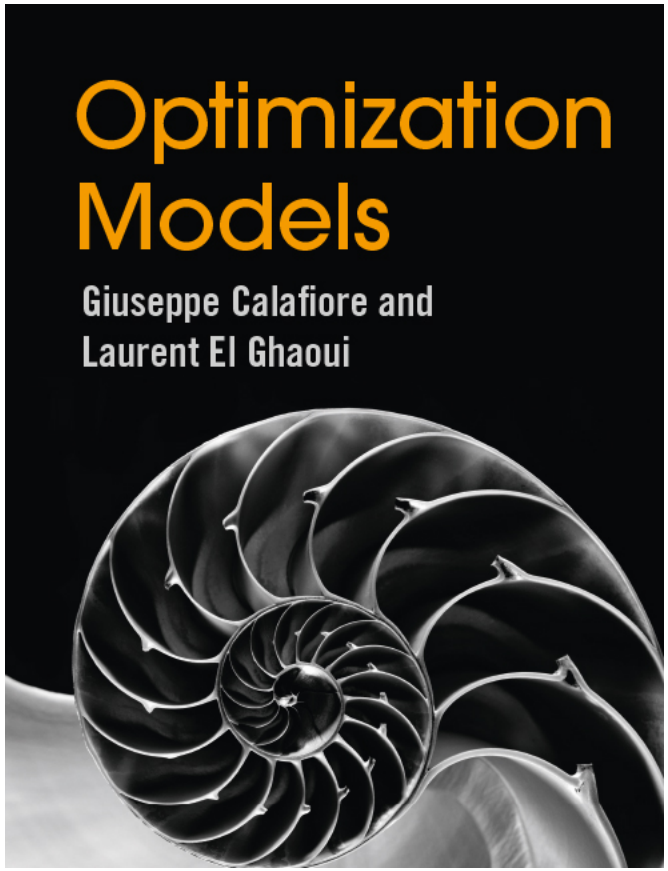
Jean-Louis Goffin

Outline

∼ Advertisement ∼

The material in these slides is taken from:

G.C. Calafiore and L. El Ghaoui
Optimization Models
Cambridge University Press, 2014



Gradients

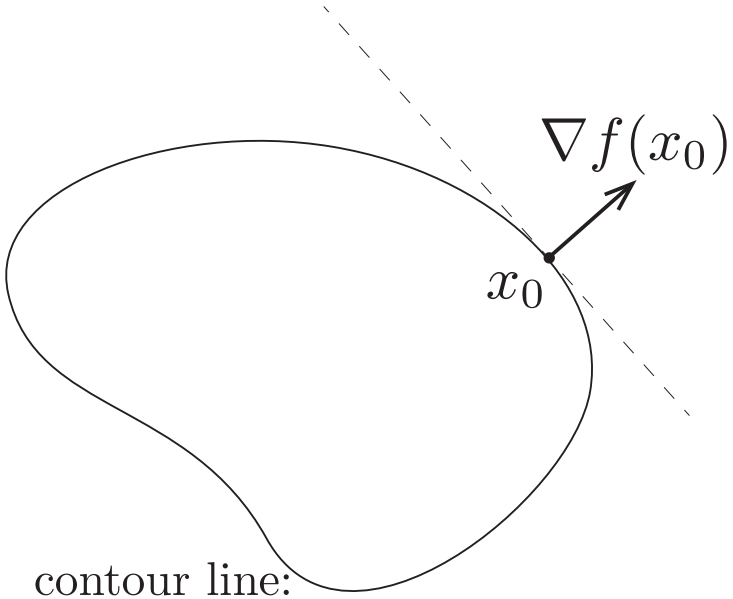
- Given a function $f : \mathbb{R}^n \to \mathbb{R}$ differentiable on its domain $\mathrm{dom}\, f$, and a point $x_0 \in \mathrm{dom}\, f$, the *gradient* of f at x_0 is defined as the vector of partial derivatives:

$$\nabla f(x_0) \doteq \left[\begin{array}{ccc} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{array} \right]^\top_{x=x_0}.$$

- The first-order Taylor series expansion of f at x states that

$$f(x) = f(x_0) + \nabla f(x_0)^\top (x - x_0) + o(\|x - x_0\|_2).$$

- The gradient is *orthogonal to the level set* $L(x_0) = \{x : f(x) = f(x_0)\}$.



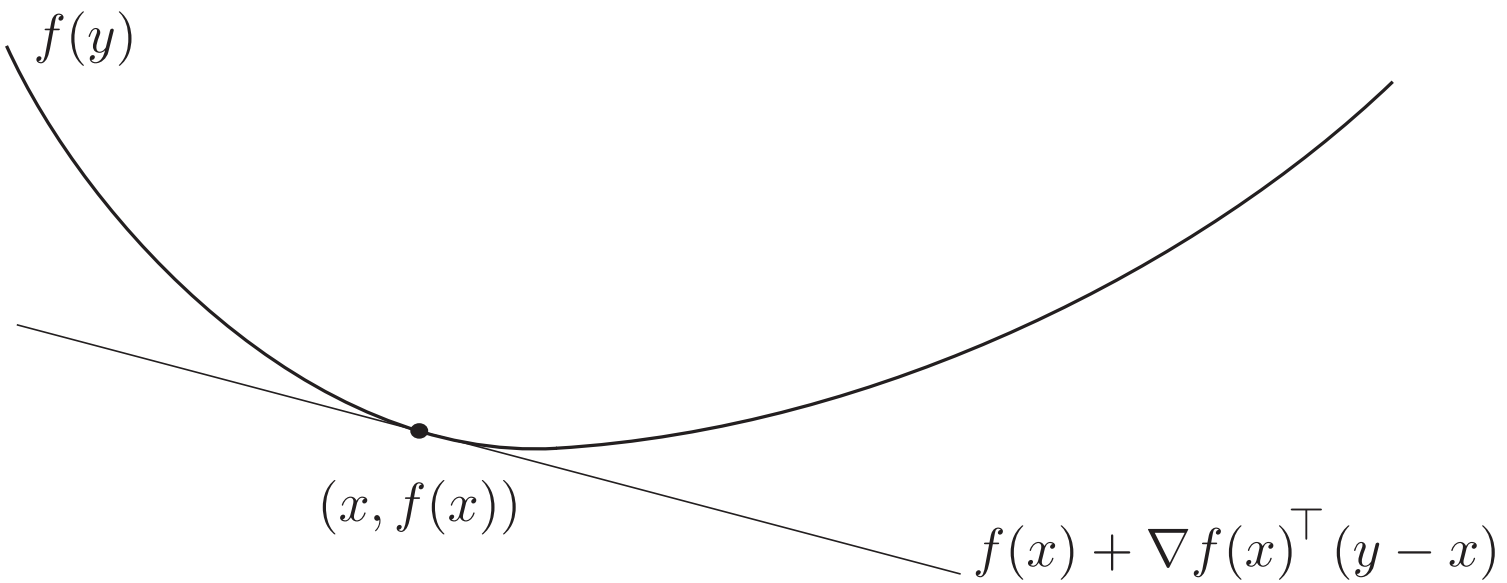
Gradients and convexity

- A key characterization of convexity for a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ states that f is convex if and only if

$$f(y) \geq f(x) + g_x^\top (y - x), \quad \forall y \in \operatorname{dom} f, \tag{1}$$

  where $g_x = \nabla f(x)$ (the gradient of f at x).

- The geometric interpretation of this condition is that the graph of f is bounded below everywhere by any one of its tangent hyperplanes or, equivalently, that any tangent hyperplane is a supporting hyperplane for the epigraph of f.



Gradients and convexity

From Eq. (1) we draw the following important observations:

- The gradient of a convex function at a point $x \in \mathbb{R}^n$ (if it is nonzero) divides the whole space in two half-spaces:

$$\begin{aligned} \mathcal{H}_{++}(x) &= \{y : \nabla f(x)^\top (y-x) > 0\}, \\ \mathcal{H}_{-}(x) &= \{y : \nabla f(x)^\top (y-x) \leq 0\} \end{aligned}$$

- For any point $y \in \mathcal{H}_{++}(x)$, we have that

$$f(y) \geq f(x) + \nabla f(x)^\top (y-x) > f(x)$$

  Thus, no point in $\mathcal{H}_{++}(x)$ can provide a value of f smaller than $f(x)$.

- Any point $y \in \mathcal{H}_{-}(x)$ defines a direction $v = \frac{y-x}{\|y-x\|_2}$ such that

$$\nabla f(x)^\top v \leq 0.$$

  By Taylor series expansion of f around x, we have

$$f(x+\epsilon v) \simeq f(x) + \epsilon \nabla f(x)^\top v \leq f(x) \text{ for small } \epsilon \geq 0.$$

Gradients and minimization

If f is convex and differentiable, and we aim at minimizing f, then

- The gradient $\nabla f(x)$ of f at a point x defines a set of descent directions

$$v : \ \nabla f(x)^\top v < 0$$

- If we do a sufficiently small step away from x in the direction of v (i.e., we move to $x_+ = x + \alpha v$), then $f(x_+) < f(x)$, i.e., we decrease the function value, if $\alpha > 0$ is small enough.
- This simple idea is the basis of first-order iterative methods for unconstrained minimization: we start from an initial point $x_0 \in \mathrm{dom}\, f$, and then iteratively update it according to

$$x_{k+1} \leftarrow x_k + \alpha_k v_k,$$

where v_k is a descent direction for f at x_k, and α_k is a suitable stepsize (much more on this later).

Subgradients and subdifferentials

What if f is not differentiable at x?

- If f is convex then relation (1) may still hold for some suitable vectors g_x.
- More precisely, if $x \in \mathrm{dom}\, f$ and (1) holds for some vector $g_x \in \mathbb{R}^n$, then g_x is called a *subgradient* of f at x.
- The set of all subgradients of f at x is called the *subdifferential*, and it is denoted by $\partial f(x)$.
- A subgradient is a "surrogate" of the gradient: it coincides with the gradient, whenever a gradient exists, and it generalizes the notion of gradient at points where f is non-differentiable.

Subgradients and subdifferentials

The following theorem holds.

Theorem 1

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and let $x \in \operatorname{relint} \operatorname{dom} f$. Then

1. *the subdifferential $\partial f(x)$ is a closed, convex, nonempty and bounded set;*
2. *if f is differentiable at x, then $\partial f(x)$ contains only one element: the gradient of f at x, that is, $\partial f(x) = \{\nabla f(x)\}$;*
3. *for any $v \in \mathbb{R}^n$ it holds that*

$$f'_v(x) \doteq \lim_{t \to 0^+} \frac{f(x+tv) - f(x)}{t} = \max_{g \in \partial f(x)} v^\top g,$$

where $f'_v(x)$ is the directional derivative *of f at x along the direction v.*

Subgradients and subdifferentials

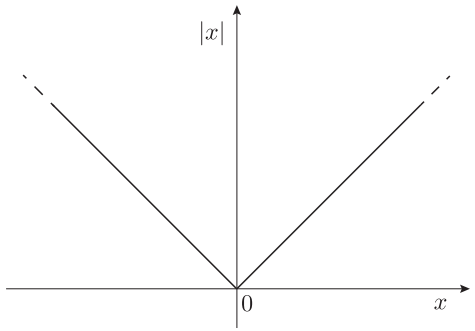
- In words, the previous theorem states that, for a convex f, a subgradient always exists at all points in the relative interior of the domain.
- Moreover, f is *directionally* differentiable at all such points.
- For a convex function f it thus holds that, for all $x \in \operatorname{relint} \operatorname{dom} f$,

$$f(y) \geq f(x) + g_x^\top (y - x), \quad \forall y \in \operatorname{dom} f, \forall g_x \in \partial f(x).$$

- We next give some examples of subgradient and subdifferential calculus.

Example: the absolute value function

- Consider the absolute value function $f(x) = |x|$, $x \in \mathbb{R}$.



- For $x > 0$, f is differentiable, hence $\partial f(x) = \{\nabla f(x)\} = \{1\}$. For $x < 0$, f is also differentiable, and $\partial f(x) = \{\nabla f(x)\} = \{-1\}$. On the contrary, f is non-differentiable at $x = 0$. However, for all $y \in \mathbb{R}$ we can write

$$f(y) = |y| = \max_{|g| \leq 1} gy \geq gy, \quad \forall g : |g| \leq 1,$$

  hence, for all $y \in \mathbb{R}$, we have $f(y) \geq f(0) + g(y - 0)$, $\forall g : |g| \leq 1$, which, compared to (1), shows that $[-1, 1]$ is the subdifferential of f at zero.

- Thus, we have that

$$\partial |x| = \begin{cases} \operatorname{sgn}(x) & \text{if } x \neq 0, \\ [-1, 1] & \text{if } x = 0. \end{cases}$$

Example: the ℓ_1 norm

- Consider the ℓ_1 norm function

$$f(x) = \|x\|_1, \quad x \in \mathbb{R}^n,$$

- We can write

$$\begin{aligned} f(y) = \|y\|_1 &= \sum_{i=1}^{n} |y_i| = \sum_{i=1}^{n} \max_{|g_i| \leq 1} g_i y_i \\ &\leq \sum_{i=1}^{n} g_i y_i, \quad \forall g : \|g\|_\infty \leq 1. \end{aligned}$$

  Hence, for all $y \in \mathbb{R}^n$ it holds that

$$f(y) \geq f(0) + g^\top (y - 0), \quad \forall g : \|g\|_\infty \leq 1.$$

- All vectors $g \in \mathbb{R}^n$ such that $\|g\|_\infty \leq 1$ are thus subgradients of f at zero, and indeed it holds that $\partial f(0) = \{g : \|g\|_\infty \leq 1\}$.

Subgradient calculus

Chain rule.

- Let $q : \mathbb{R}^n \to \mathbb{R}^m$ and $h : \mathbb{R}^m \to \mathbb{R}$ be such that the composite function $f = h \circ q : \mathbb{R}^n \to \mathbb{R}$, with values $f(x) = h(q(x))$, is convex. Then:
  1. if q is differentiable at x and $q(x) \in \operatorname{dom} h$, then

$$\partial f(x) = J_q(x)^\top \partial_q h(q(x)),$$

     where $J_q(x)$ is the Jacobian of q at x;
  2. if $m = 1$ and h is differentiable at $q(x) \in \operatorname{dom} h$, then

$$\partial f(x) = \frac{\mathrm{d}h(q(x))}{\mathrm{d}q(x)} \partial q(x).$$

Subgradient calculus

Affine variable transformation.

- As a particular case of the first of the previous chain rules, let $h : \mathbb{R}^m \to \mathbb{R}$ be convex and let $q(x) = Ax + b$, where $A \in \mathbb{R}^{m,n}$, $b \in \mathbb{R}^m$.
- Then the function from $\mathbb{R}^n$ to $\mathbb{R}$

$$f(x) = h(q(x)) = h(Ax + b)$$

has subdifferential $\partial f(x) = A^\top \partial_q h(q(x))$, for all x such that $q(x) \in \operatorname{dom} h$.

- Example: $f(x) = |a^\top x - b|$ is the composition of $h(x) = |x|$ with the affine function $q(x) = a^\top x - b$. Thus

$$\partial |a^\top x - b| = a \cdot \partial h(a^\top x - b) = \begin{cases} a \cdot \operatorname{sgn}(a^\top x - b) & \text{if } a^\top x - b \neq 0, \\ a \cdot [-1,\ 1] & \text{if } a^\top x - b = 0. \end{cases}$$

Subgradient calculus

Sum or linear combination.

- Let $h : \mathbb{R}^n \to \mathbb{R}$, $q : \mathbb{R}^n \to \mathbb{R}$ be convex functions, let $\alpha, \beta \geq 0$, and let

$$f(x) = \alpha h(x) + \beta q(x).$$

- Then, for any $x \in \text{relint dom } h \cap \text{relint dom } q$ it holds that

$$\partial f(x) = \alpha \partial h(x) + \beta \partial q(x).$$

- Example: for $f(x) = \sum_{i=1}^{m} |a_i^\top x - b_i|$ we have

$$\partial f(x) = \sum_{i=1}^{m} \partial |a_i^\top x - b_i| = \sum_{i=1}^{m} \begin{cases} a_i \cdot \text{sgn}(a_i^\top x - b_i) & \text{if } a_i^\top x - b_i \neq 0, \\ a_i \cdot [-1,\ 1] & \text{if } a_i^\top x - b_i = 0. \end{cases}$$

Subgradient calculus

Pointwise maximum.

- Let $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$, be convex functions, and let

$$f(x) = \max_{i=1,\ldots,m} f_i(x).$$

- Then, for $x \in \operatorname{dom} f$ it holds that

$$\partial f(x) = \operatorname{co}\{\partial f_i(x) : \ i \in a(x)\},$$

where $a(x)$ is the set of indices of the functions f_i that are "active" at x, that is the ones that attain the maximum in the definition of f, hence $f(x) = f_i(x)$, for $i \in a(x)$.

- Extension to pointwise maxima of arbitrary (possibly uncountable) families of convex functions, under some additional technical assumptions. More precisely, let $f(x) = \sup_{\alpha \in \mathcal{A}} f_\alpha(x)$, where f_α are convex and closed functions and $\mathcal{A}$ is compact. Then, for any $x \in \operatorname{dom} f$ it holds that

$$\partial f(x) = \operatorname{co}\{\partial f_\alpha(x) : \ \alpha \in a(x)\}.$$

where $a(x) = \{\alpha \in \mathcal{A} : f(x) = f_\alpha(x)\}$.

Subgradient calculus

Pointwise maximum: examples

- Subdifferentials of typical ℓ_p norms can be obtained by expressing the norm in the form of a supremum of linear functions over a suitable set:

$$\begin{aligned} \|x\|_1 &= \max_{\|v\|_\infty \leq 1} v^\top x, \\ \|x\|_2 &= \max_{\|v\|_2 \leq 1} v^\top x, \\ \|x\|_\infty &= \max_{\|v\|_1 \leq 1} v^\top x. \end{aligned}$$

- For the ℓ_2 norm case, we have, for $f_v \doteq v^\top x$,

$$\partial \|x\|_2 = \mathrm{co}\{\partial f_v(x) : \ v \in a(x)\} = \mathrm{co}\{v : \ v \in a(x)\} = \mathrm{co}\{a(x)\},$$

where $a(x) = \{v : \ \|x\|_2 = v^\top x, \ \|v\|_2 \leq 1\}$.

- For $x = 0$, we have $\|x\|_2 = 0$ which is attained for all feasible v, hence $a(x) = \{v : \ \|v\|_2 \leq 1\}$, and $\partial \|x\|_2 = \{v : \ \|v\|_2 \leq 1\}$:

$$\partial \|x\|_2 = \begin{cases} \frac{x}{\|x\|_2} & \text{if } x \neq 0, \\ \{g \in \mathbb{R}^n : \ \|g\|_2 \leq 1\} & \text{if } x = 0. \end{cases}$$

Subgradient calculus

Largest eigenvalue of a symmetric matrix (1/2).

- Consider a symmetric matrix $A(x)$ whose entries are affine functions of a vector of variables $x \in \mathbb{R}^n$:

$$A(x) = A_0 + x_1 A_1 + \cdots + x_n A_n,$$

where $A_i \in \mathbb{S}^m$, $i = 0, \ldots, n$, and let $f(x) = \lambda_{\max}(A(x))$.

- To determine the subdifferential of f at x we exploit Rayleigh's variational characterization, which states that

$$\begin{aligned} f(x) = \lambda_{\max}(A(x)) &= \max_{z:\, \|z\|_2=1} z^\top A(x) z \\ &= \max_{z:\, \|z\|_2=1} z^\top A_0 z + \sum_{i=1}^{n} x_i z^\top A_i z. \end{aligned}$$

$f(x)$ is thus expressed as the max over z (on the unit sphere) of functions $f_z(x) = z^\top A(x) z$ which are affine in x (hence, f is indeed convex).

Subgradient calculus

Largest eigenvalue of a symmetric matrix (2/2).

- The active set

$$a(x) = \{z : \|z\|_2 = 1,\ f_z(x) = f(x)\}$$

is composed of the eigenvectors of $A(x)$ associated with the largest eigenvalue (and normalized with unit norm). We hence have that

$$\partial f(x) = \mathrm{co}\{\nabla f_z(x) : A(x)z = \lambda_{\max}(A(x))z,\ \|z\|_2 = 1\},$$

where

$$\nabla f_z(x) = [z^\top A_1 z \ \cdots \ z^\top A_n z]^\top.$$

- In particular, f is differentiable at x whenever the eigenspace associated with $\lambda_{\max}(A(x))$ has dimension one, in which case $\partial f(x) = \{\nabla f_z(x)\}$, where z is the unique (up to a sign, which is, however, irrelevant) normalized eigenvector of $A(x)$ associated with $\lambda_{\max}(A(x))$.

Optimality conditions for unconstrained minimization

- Consider the unconstrained minimization problem

$$\min_{x \in \mathbb{R}^n} \; f(x), \quad f \text{ convex.}$$

- If f is differentiable then x is an optimal point if and only if

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

Indeed, if $\nabla f(x) = 0$ then from (1) we have that $f(y) \geq f(x)$ for all $y \in \operatorname{dom} f$, hence x is optimal. Conversely, suppose x is optimal and, by the purpose of contradiction, that $\nabla f(x) \neq 0$: then there would exist a descent direction at x, which would contradict optimality.

- If f is not differentiable then x is an optimal point if and only if

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