Lecture 1: Convex smooth optimisation

Luca Calatroni
CR CNRS, Laboratoire I3S
CNRS, UCA, Inria SAM, France

MIVA ERASMUS BIP PhD winter school

Advanced methods for mathematical image analysis

University of Bologna, IT

January 18-20 2022

Table of contents

Schedule

	WEDNESDAY 18/01	THURSDAY 19/01	FRIDAY 20/01
08:00			
09:00			
10:00			
11:00			
12:00			
13:00	Lunch	Lunch	Lunch
14:30	Comp. Imaging Lab	EXERCISES	Comp.Imaging LAB
15:30	LAB	LAB	EXERCISES
16:30	SEMINAR **Automotive**	SEMINAR **Industrial**	SEMINAR **Health**
17:30			
	Prof. L. Calatroni	**Social Dinner**	
	Prof. O. Öktem		

Introduction

Motivation

Goal: providing theoretical & practical tools (i.e. algorithms) for solving

$$\min_{x\in\mathbb{R}^n} F(x)$$

for a functional $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ with suitable properties.

- F is smooth $\to$ gradient descent & variants (**this lecture**)
- $F := f + g$, f smooth & g non-smooth $\to$ proximal-gradient algorithms & variants (**next lecture**)
- $F := f + \|x\|_0$ with f smooth $\to$ which algorithms? (**last lecture**)

Such minimisation problems often appears in many contexts:

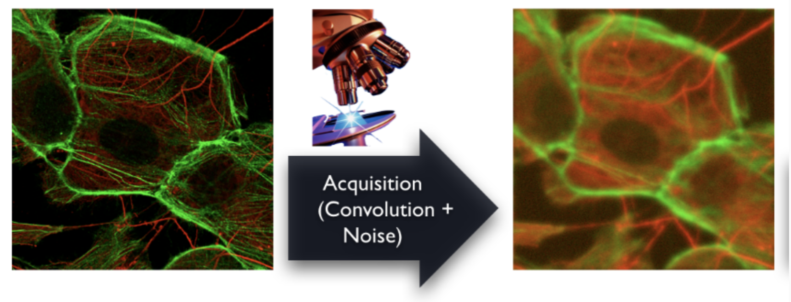
- **Inverse problems** in signal/image processing: image reconstruction, variable/parameter selection, compressed sensing. . . .
- **Statistical/machine learning**: empirical risk minimisation, regression. . .
- **Optimisation per se**: analysis/implementation of fast algorithms for solving large-scale problems. . .

Framework: optimisation for inverse problems in imaging

$$\text{Given } y \in \mathbb{R}^m, A \in \mathbb{R}^{m\times n} \quad \text{find} \quad x \in \mathbb{R}^n \quad \text{s.t.} \quad y = \mathcal{T}(Ax)$$

where $m \leq n$ and $\mathcal{T} : \mathbb{R}^m \to \mathbb{R}^m$ models noise degradation.

- **Image restoration** (denoising, deconvolution, super-resolution)

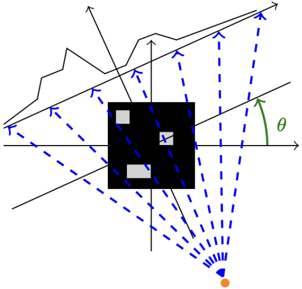


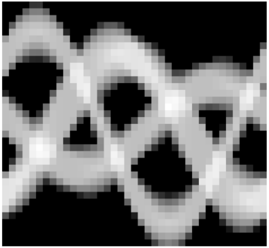
Framework: optimisation for inverse problems in imaging

$$\text{Given } y \in \mathbb{R}^m, A \in \mathbb{R}^{m\times n} \quad \text{find} \quad x \in \mathbb{R}^n \quad \text{s.t.} \quad y = \mathcal{T}(Ax)$$

where $m \leq n$ and $\mathcal{T} : \mathbb{R}^m \to \mathbb{R}^m$ models noise degradation.

- **Image restoration** (denoising, deconvolution, super-resolution)
- **Image reconstruction** (e.g., medical imaging)



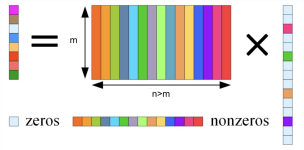


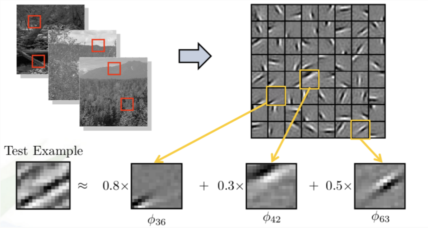
Framework: optimisation for inverse problems in imaging

$$\text{Given } y \in \mathbb{R}^m, A \in \mathbb{R}^{m\times n} \quad \text{find} \quad x \in \mathbb{R}^n \quad \text{s.t.} \quad y = \mathcal{T}(Ax)$$

where $m \leq n$ and $\mathcal{T} : \mathbb{R}^m \to \mathbb{R}^m$ models noise degradation.

- **Image restoration** (denoising, deconvolution, super-resolution)
- **Image reconstruction** (e.g., medical imaging)
- **Dictionary representation** (data analysis, vision)





Framework: optimisation for inverse problems in imaging

$$\text{Given } y \in \mathbb{R}^m, A \in \mathbb{R}^{m\times n} \quad \text{find} \quad x \in \mathbb{R}^n \quad \text{s.t.} \quad y = \mathcal{T}(Ax)$$

where $m \leq n$ and $\mathcal{T} : \mathbb{R}^m \to \mathbb{R}^m$ models noise degradation.

- **Image restoration** (denoising, deconvolution, super-resolution)
- **Image reconstruction** (e.g., medical imaging)
- **Dictionary representation** (data analysis, vision)

. . . "naive inversion" not possible for $y = Ax + n$, $n \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$:

$$\cancel{x = A^{-1}(y - n)}$$

Bad positioning of inverse filtering

$$y = Ax + n$$

Inverse filtering approach:

$$x = A^{-1}y = A^{-1}(Ax + n) = x + A^{-1}n$$

Amplification of the noise if A^{-1} is bad conditioned! Need of regularisation!

Find an estimate $\mathbb{R}^n \ni x^* \approx x$ by solving

$$x^* \in \underset{x \in \mathbb{R}^n}{\arg\min} \; F(x) := f(x) + g(x)$$

- f is the data fidelity term, it relates to noise statistics
- g is the regularisation term, it encodes *a priori* information expected on the desired solution

Variational regularisation: Bayesian motivation

Following a Bayesian/MAP approach consider:

$$P(y|Ax;\theta_f) \quad \text{(likelihood)}, \qquad P(x;\theta_g) \quad \text{(prior)}$$

with $\theta_f, \theta_g > 0$ hyperparameters of the distributions. By Bayes' theorem:

$$x^* \in \arg\max_x \; P(x|y) = \arg\max_x \; \frac{P(y|Ax;\theta_f)P(x;\theta_g)}{P(y)}$$

$$\Leftrightarrow x^* \in \arg\min_x \; -\ln(P(x|y)) = \arg\min_x \; -\ln(P(y|Ax;\theta_f)) - \ln(P(x;\theta_g)) + \cancel{\ln(P(y))}$$

Now, if $P(x;\theta_g) = e^{-\theta_g g(x)}$ and $P(y|Ax;\theta_f) = e^{-\theta_f f(x)}$, then:

$$x^* \in \arg\min_{x\in\mathbb{R}^n} \; f(x) + \lambda g(x), \qquad \lambda := \theta_g/\theta_f$$

Note: incorporate the parameter α in either of the two functions, e.g. $g(x) := \lambda g(x)$.

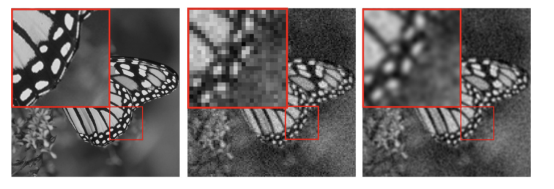
Exemplar problems: smooth optimisation

$$y = Ax + b$$

- **Generalised Tikhonov** $n \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$ (Gaussian noise) and assume x is smooth in some sense (e.g., in terms of an operator $L \in \mathbb{R}^{N \times n}$)

$$x^* \in \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - y\|^2 + \lambda \|Lx\|^2$$

Examples: $L = \mathrm{Id} \in \mathbb{R}^{n \times n}$, $L = D \in \mathbb{R}^{2n \times n}$ (discrete gradient) . . .



Parameter selection for ℓ_2-ℓ_2 single-image super-resolution, $A = SH$, where S is a decimation operator (Pragliola, Calatroni, Lanza, Sgallari, '21-'22)

Exemplar problems: non-smooth optimisation

Assume for simplicity additive white Gaussian noise → $f(x) = \frac{1}{2}\|Ax - y\|^2$

- **Sparsity** (Donoho et al., Candès, Romberg, Tao, '06): sparse recovery:

$$x^* \in \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - y\|^2 + \lambda\|x\|_1$$

Analysis approach: sparse representation of x in some overcomplete basis (e.g., wavelets, Mallat, '89) represented by $W \in \mathbb{R}^{N \times n}$

$$x^* \in \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - y\|^2 + \lambda\|Wx\|_1$$

- **Total variation reconstruction**: "few gradients" for removing noise oscillations and preserving edges (Rudin, Osher, Fatemi, '92):

$$x^* \in \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - y\|^2 + \lambda\|Dx\|_{2,1}$$

with $\|Dx\|_{2,1} = \sum_{i=1}^{n} \sqrt{(D_h x)_i^2 + (D_v x)_i^2}$ and Dx is the discrete image gradient.

Exemplar problems: non-smooth optimisation (continuation)

It helps in dealing with admissibility constraints:

$$x^* \in \arg\min_{x \in C} \frac{1}{2}\|Ax - y\|^2$$

with $C := \bigcap_{m=1}^{M} C_m$ and $C_m \subset \mathbb{R}^n$.

- **Non-negativity constraint**: $x \geq 0$, $C := \{x \geq 0\}$.
- **Box constraint**: $x \in [a, b] =: C$
- ...

How to encode it into a variational formulation?

Using the indicator function $\iota : \mathbb{R}^n \to \{0, +\infty\}$

$$\iota_{C_m}(x) := \begin{cases} 0 & \text{if } x \in C_m \\ +\infty & \text{if } x \notin C_m \end{cases}$$

$$x^* \in \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - y\|^2 + \sum_{m=1}^{M} \iota_{C_m}(x)$$

Exemplar problems: ℓ_2-ℓ_0 optimisation

Arising, e.g., in sparse dictionary representation problems

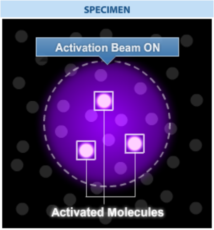
$$y = Ax + n$$

where $y \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$ and $x \in \mathbb{R}^n$ and $m \ll n$. Undetermined system!

To minimise the number of entries of solutions, the natural choice is to consider:

$$x^* \in \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-y\|^2 + \lambda\|x\|_0 \quad \text{or} \quad x^* \in \arg\min_{x:\|x\|_0\leq K} \frac{1}{2}\|Ax-y\|^2$$

$$\|x\|_0 := \#\{x_i, i=1,\dots,N : x_i \neq 0\}$$



Molecule localisation in super-resolution microscopy.

References

Some standard reference books/surveys:

- R. Tyller Rockafeller, *Convex Analysis*, Princeton University Press, 1970.
- S. Boyd, L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.
- N. Parikh, S. Boyd, *Proximal Algorithms*, Foundations and Trends in Optimization, 2013.
- A. Beck, *First-order methods in optimization*, Volume 25, MOS-SIAM series on Optimization, 2017.
- A. Chambolle, T. Pock, *An introduction to continuous optimization for imaging*, Acta Numerica, 2016
- S. Salzo, S. Villa, *Proximal Gradient Methods for Machine Learning and Imaging*, Handbook on Harmonic and Applied Analysis, Applied and Numerical Harmonic Analysis, 2021.

Back to the abstract problem

$$x^* \in \arg\min_{x \in \mathbb{R}^n} \; F(x) := f(x) + g(x)$$

Often the solution x^* cannot be expressed in closed form. We consider efficient iterative solvers for its computation (especially in large scale context!)

- Avoid inversion A^{-1} ($1 \ll m \leq n$)
- How to exploit the mathematical structure of the functions involved?
- How to handle constraints?
- How to speed up the efficiency of a first-order algorithm?
- What can be said in the non-convex case?

Notation, preliminaries & basic notions

Notation

- $(X, \langle v, w \rangle) = (\mathbb{R}^n, v^T w)$ with Euclidean norm $\|\cdot\|$ as reference Hilbert space. Extensions to general Hilbert setting straightforward.
- $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$, $\mathbb{R}_+ := \{\alpha \in \mathbb{R} : \alpha \geq 0\}$, $\mathbb{R}_{++} := \{\alpha \in \mathbb{R} : \alpha > 0\}$
- Closed ball of radius $\delta > 0$ in $x \in X$:

$$B_\delta(x) = \{y \in X : \|y - x\| \leq \delta\}$$

- Convex set $C \subset X$

$$(\forall x, y \in C) \quad \forall \alpha \in [0, 1] \quad \alpha x + (1 - \alpha) y \in C$$

- Epigraph of a function $f : \mathbb{R} \to \overline{\mathbb{R}}$:

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

Proper functions

Minimal property to have well-defined minimisation problems.

Definition (proper function)

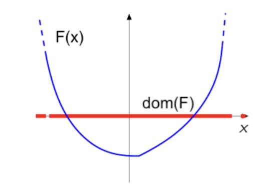
A function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ is said *proper* iff

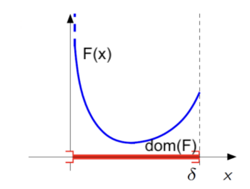
$$\exists x \in \mathbb{R}^n \quad \text{such that} \quad F(x) \neq +\infty.$$

We define $\mathcal{P} := \left\{ F : \mathbb{R}^n \to \overline{\mathbb{R}} : F \text{ is proper} \right\}$ and

$$\mathrm{dom}(F) := \{x \in \mathbb{R}^n : F(x) < +\infty\}$$

Clearly, $F \in \mathcal{P} \Leftrightarrow \mathrm{dom}(F) \neq \emptyset$.

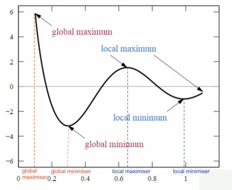




Global/local minimisers

For $F \in \mathcal{P}$, recall:

- **global minimiser**: $x^* \in \mathbb{R}^n$: $F(x^*) \leq F(x)$ for <u>every $x \in \mathbb{R}^n$</u>.
- **local minimiser**: $x^* \in \mathbb{R}^n$: there exists $\delta > 0$ and a neighbourhood $B_\delta(x^*)$ such that $F(x^*) \leq F(x)$ for <u>every $x \in B_\delta(x^*)$</u>.

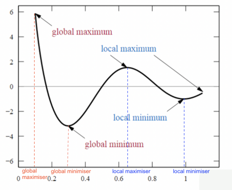


$$\min_{x \in \mathbb{R}^n} F(x) \quad \text{VS} \quad \arg\min_{x \in \mathbb{R}^n} F(x)$$

Global/local minimisers

For $F \in \mathcal{P}$, recall:

- **global minimiser**: $x^* \in \mathbb{R}^n$: $F(x^*) \leq F(x)$ for <u>every $x \in \mathbb{R}^n$</u>.
- **local minimiser**: $x^* \in \mathbb{R}^n$: there exists $\delta > 0$ and a neighbourhood $B_\delta(x^*)$ such that $F(x^*) \leq F(x)$ for <u>every $x \in B_\delta(x^*)$</u>.



$$\min_{x \in \mathbb{R}^n} F(x) \quad \text{VS} \quad \arg\min_{x \in \mathbb{R}^n} F(x)$$

Definition (set of minimisers)

The **set** of (local, global) minimisers of F is denoted by:

$$\arg\min F = \{x^* \in \mathbb{R}^n : x^* \text{ is a minimiser of } F\} \subset \mathbb{R}^n$$

Empty? Singleton? (it depends on F)

Notation, preliminaries & basic notions

Convexity, strong convexity

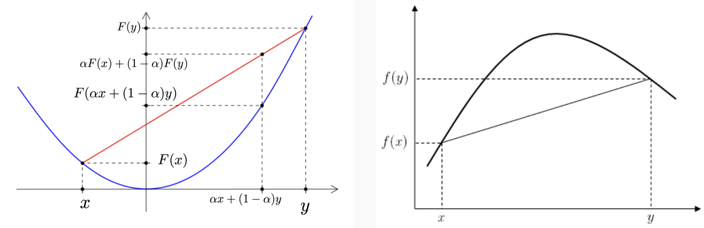
Convex functions

Definition (convex function)

$F \in \mathcal{P}$ is said to be *convex* if:

$$(\forall x, y \in \mathbb{R}^n), \quad (\forall \alpha \in [0,1]), \quad F(\alpha x + (1-\alpha)y) \leq \alpha F(x) + (1-\alpha)F(y).$$

Moreover, F is *strictly convex* if the inequality holds when $x, y \in \mathrm{dom}(F)$, $x \neq y$ and $\alpha \in (0,1)$. We say that $G : \mathbb{R}^n \to [-\infty, +\infty)$ is *concave* is $F = -G$ is convex. If a function is not convex nor concave we say that is *non-convex*.



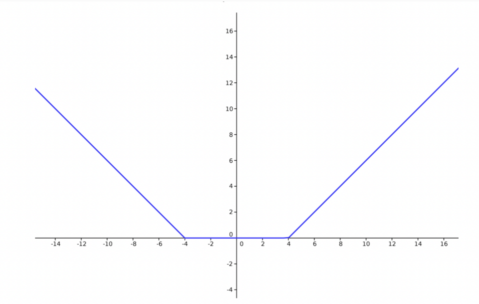
Convex/concave function

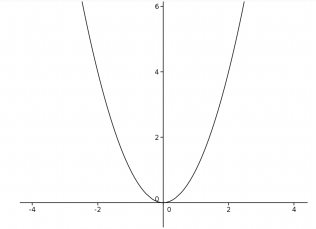
Definition (convex function)

$F \in \mathcal{P}$ is said to be *convex* if:

$$(\forall x, y \in \mathbb{R}^n), \quad (\forall \alpha \in [0,1]), \quad F(\alpha x + (1-\alpha)y) \leq \alpha F(x) + (1-\alpha)F(y).$$

Moreover, F is *strictly convex* if the inequality holds when $x, y \in \mathrm{dom}(F)$, $x \neq y$ and $\alpha \in (0,1)$. We say that $G : \mathbb{R}^n \to [-\infty, +\infty)$ is *concave* is $F = -G$ is convex. If a function is not convex nor concave we say that is *non-convex*.





Convex VS. strictly convex functions

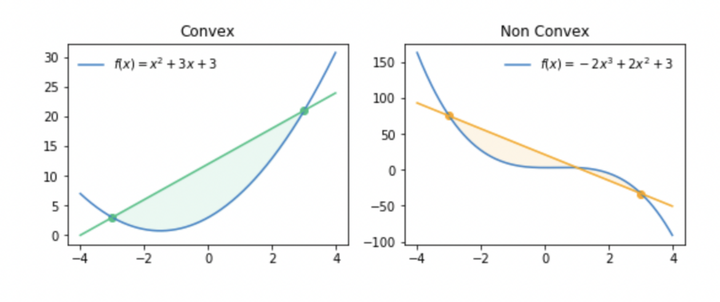
Convex functions

Definition (convex function)

$F \in \mathcal{P}$ is said to be *convex* if:

$$(\forall x, y \in \mathbb{R}^n), \quad (\forall \alpha \in [0,1]), \quad F(\alpha x + (1-\alpha)y) \leq \alpha F(x) + (1-\alpha)F(y).$$

Moreover, F is *strictly convex* if the inequality holds when $x, y \in \mathrm{dom}(F)$, $x \neq y$ and $\alpha \in (0,1)$. We say that $G : \mathbb{R}^n \to [-\infty, +\infty)$ is *concave* is $F = -G$ is convex. If a function is not convex nor concave we say that is *non-convex*.



Convex VS. non-convex function

Convex functions

Definition (convex function)

$F \in \mathcal{P}$ is said to be *convex* if:

$$(\forall x, y \in \mathbb{R}^n), \quad (\forall \alpha \in [0,1]), \quad F(\alpha x + (1-\alpha)y) \leq \alpha F(x) + (1-\alpha)F(y).$$

Moreover, F is *strictly convex* if the inequality holds when $x, y \in \mathrm{dom}(F)$, $x \neq y$ and $\alpha \in (0,1)$. We say that $G : \mathbb{R}^n \to [-\infty, +\infty)$ is *concave* is $F = -G$ is convex. If a function is not convex nor concave we say that is *non-convex*.

Examples:

- $F(x) = \|x\|$ is convex

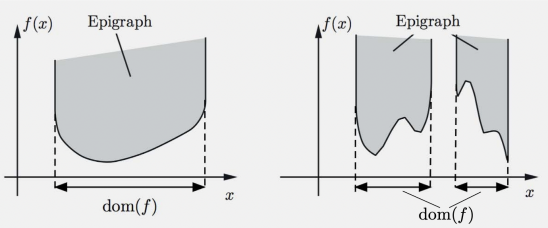
$$\|\alpha x + (1-\alpha)y\| \leq \|\alpha x\| + \|(1-\alpha)y\| = \alpha\|x\| + (1-\alpha)\|y\| \qquad \forall x, y \in \mathbb{R}^n$$

- $F(x) = \|x\|^2$ is strictly convex
- $F(x) = \|x\|_p$, $p \in [1, +\infty)$ are convex

Useful properties

Proposition (epigraph of convex functions is convex set)

Let $F \in \mathcal{P}$. Then F is convex if and only if $\text{epi}(F)$ is a convex set.



Proposition (operations with convex functions)

Let f and g be two convex functions and let $\beta \in \mathbb{R}_{++}$. Then, the sum $f + g$ is a convex function and the function βf is a convex function.

Strong convexity

Definition (strongly convex function)

$F \in \mathcal{P}$ is said to be *strongly convex* of parameter $\mu > 0$ iff $\forall x, y \in \mathbb{R}^n$ and $\forall \alpha \in [0, 1]$:

$$F(\alpha x + (1-\alpha)y) \leq \alpha F(x) + (1-\alpha)F(y) - \frac{\mu}{2}(1-\alpha)\alpha\|x-y\|^2$$

Proposition (characteristion of strongly convex functions)

$F \in \mathcal{P}$ is μ-*strongly convex* if and only if $G(\cdot) := F(\cdot) - \frac{\mu}{2}\|\cdot\|^2$ is convex.

Proposition (growth condition around minimisers)

If $F \in \mathcal{P}$ is μ-*strongly convex* and $x^* \in \arg\min_x F(x)$, then:

$$F(x) - F(x^*) \geq \frac{\mu}{2}\|x - x^*\|^2, \quad \forall x \in X.$$

Strong convexity

Definition (strongly convex function)

$F \in \mathcal{P}$ is said to be *strongly convex* of parameter $\mu > 0$ iff $\forall x, y \in \mathbb{R}^n$ and $\forall \alpha \in [0, 1]$:

$$F(\alpha x + (1-\alpha)y) \leq \alpha F(x) + (1-\alpha)F(y) - \frac{\mu}{2}(1-\alpha)\alpha\|x-y\|^2$$

Proposition (characteristion of strongly convex functions)

$F \in \mathcal{P}$ is μ-*strongly convex* if and only if $G(\cdot) := F(\cdot) - \frac{\mu}{2}\|\cdot\|^2$ is convex.

Proposition (growth condition around minimisers)

If $F \in \mathcal{P}$ is μ-*strongly convex* and $x^* \in \arg\min_x F(x)$, then:

$$F(x) - F(x^*) \geq \frac{\mu}{2}\|x - x^*\|^2, \quad \forall x \in X.$$

strong convexity $\Rightarrow$ strict convexity $\Rightarrow$ convexity

Counterexample (strict convexity $\not\Rightarrow$ strong convexity): $F : \mathbb{R} \to \overline{\mathbb{R}}$, $F(x) = e^x$.

Notation, preliminaries & basic notions

Lower semi-continuity & coercivity

Lower semi-continuity

Definition (lower semi-continuity)

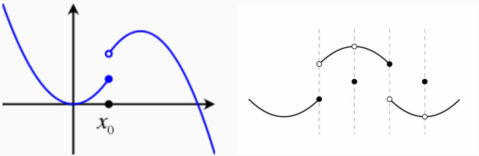
Let $F \in \mathcal{P}$. We say that F is *lower semi-continuous (l.s.c.) at the point* $x \in \mathbb{R}^n$ iff

$$F(x) \leq \liminf_{y \to x} F(y).$$

Equivalently, for every sequence $(x_k)_{k \in \mathbb{N}}$ with $x_k \to x$:

$$F(x) \leq \liminf_{k \to +\infty} F(x_k) \left(= \lim_{k \to +\infty} \inf \left\{ F(x_j) : j \geq k \right\} \right).$$

If F is l.s.c. at every $x \in \mathbb{R}^n$, we say that the function is l.s.c.



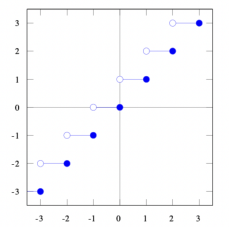
Left: lower l.s.c. **Right**: where the function is lower l.s.c.?

Examples of l.s.c. functions

- The functions

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}, \qquad F(x) = \lceil x \rceil = \min\{k \in \mathbb{Z} : x \leq k\}$$

are l.s.c. (but not continuous).



$F(x) = \lceil x \rceil$

- All continuous functions (l.s.c + u.s.c.).

How to ensure that the minimum is not attained at "extreme points" of the domain?

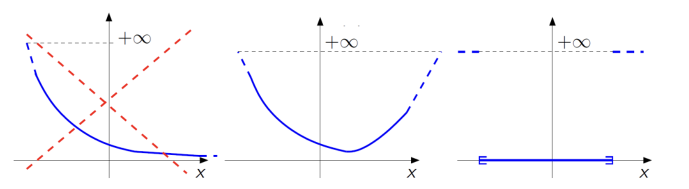
Definition (coercivity)

Let $F \in \mathcal{P}$. We say that F is *coercive* iff

$$\lim_{\|x\| \to +\infty} F(x) = +\infty.$$

Examples:

- $F : \mathbb{R} \to \mathbb{R}_+$, $F(x) = e^x$ is **not** coercive, but $F : \mathbb{R} \to \mathbb{R}_+$, $F(x) = e^{|x|}$ is.
- $F : \mathbb{R}^2 \to \mathbb{R}_+$, $F(x, y) = x^2 + y^2$ is coercive.
- $F : \mathbb{R}^2 \to \mathbb{R}_+$, $F(x, y) = x^2 - 2xy + x^2 = (x - y)^2$ is **not** coercive. Why?



Existence of minimisers

Theorem (existence of minimisers)

If F is proper, l.s.c. and coercive, then the set of minimisers of F is non-empty and compact.

Note: generalises the Bolzano-Weirestrass theorem holding for problems

$$\min_{x \in C} F(x)$$

for compact $C \subset \mathbb{R}^n$ s.t. $C \cap \mathrm{dom}(F) \neq \emptyset$ and *continuous* F.

Existence of minimisers

Theorem (existence of minimisers)

If F is proper, l.s.c. and coercive, then the set of minimisers of F is non-empty and compact.

Note: generalises the Bolzano-Weirestrass theorem holding for problems

$$\min_{x \in C} F(x)$$

for compact $C \subset \mathbb{R}^n$ s.t. $C \cap \text{dom}(F) \neq \emptyset$ and *continuous* F.

Theorem (convex case)

If F is proper, coercive and convex, then every local minimiser is a global minimiser.

Definition ($\Gamma_0(\mathbb{R}^n)$)

$$\Gamma_0(X) := \left\{ F : X \to \overline{\mathbb{R}} : F \text{ is proper, convex and l.s.c.} \right\}$$

Remark: $F \in \Gamma_0(X) \not\Rightarrow F$ admits a minimiser. Take e.g. $F(x) = -\log x, x > 0$ and $F(x) = +\infty, x \leq 0 \ldots$ no coercivity guaranteed!

Uniqueness of minimisers

So far, only existence of minimisers. How to guarantee uniqueness?

Theorem (existence+uniqueness of minimisers)

If F is proper, l.s.c., coercive and strictly convex, then F admits a unique minimiser.

Equivalently, $\arg\min F = \{x^*\}$, a singleton.

Remark: as strong convexity implies strict convexity, the same holds.

Notation, preliminaries & basic notions

Differentiability and L-smoothness

How to provide a characterisation of the minimisers of a function f in terms of a suitable notion of "∇f"?

Definition (Gâteaux differentiability)

Let $f \in \mathcal{P}$ and let $x \in \text{dom}(f)$. For $v \in \mathbb{R}^n$, we denote the *directional derivative* in x along the direction v as the limit

$$f'(x; v) = f'(x)[v] := \lim_{t \to 0^+} \frac{f(x + tv) - f(x)}{t},$$

when it exists. If there exists $w \in \mathbb{R}^n$ such that:

$$(\forall v \in \mathbb{R}^n) \quad f'(x)[v] = \langle w, v \rangle,$$

then we say that f is *Gâteaux differentiable* in x and denote by $\nabla f(x) = w$ the *Gâteaux derivative* (or, simply, the *gradient*) of f at x.

Optimality conditions and relations with convexity

Theorem (Fermat's rule)

Let $f \in \Gamma_0(\mathbb{R}^n)$ be differentiable at point x^*. Then:

$$x^* \in \underset{x \in \mathbb{R}^n}{\arg\min} f(x) \iff \nabla f(x^*) = 0.$$

Theorem (Fermat's rule)

Let $f \in \Gamma_0(\mathbb{R}^n)$ be differentiable at point x^*. Then:

$$x^* \in \arg\min_{x \in \mathbb{R}^n} f(x) \quad \Longleftrightarrow \quad \nabla f(x^*) = 0.$$

Proposition (Differentiability and convexity)

Let $f \in \Gamma_0(\mathbb{R}^n)$. Suppose that f is differentiable on $\mathrm{dom}(f)$. Then the following statements are equivalent:

1. f is convex;
2. $\forall x, y \in \mathrm{dom}(f)$, $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$;
3. $\forall x, y \in \mathrm{dom}(f)$, $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$.

Corollary (Differentiability and strong convexity)

Let $f \in \Gamma_0(\mathbb{R}^n)$ and $\mu > 0$. Suppose that f is differentiable on $\mathrm{dom}(f)$. Then the following statements are equivalent:

1. f is μ-strongly convex;
2. $\forall x, y \in \mathrm{dom}(f)$, $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2$;
3. $\forall x, y \in \mathrm{dom}(f)$, $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$.

Example: let $f(x) = \frac{1}{2}\|Ax - y\|^2$, for $A \in \mathbb{R}^{m \times n}$ positive definite, $y \in \mathbb{R}^m$. Then:

$$\nabla f(x) = A^T(Ax - y).$$

Since $A^T A$ is positive definite (i.e., $\lambda_{\min} := \lambda_{\min}(A^T A) > 0$), then:

$$(\forall x, y \in \mathbb{R}^n) \quad \langle \nabla f(x) - \nabla f(y), x - y \rangle = \langle A^T A(x - y), x - y \rangle \geq \lambda_{\min}\|x - y\|^2,$$

hence f is $\lambda_{\min}$-strongly convex.

Remark: from condition 3., if $x^* \in \arg\min f(x)$, then for all $x \in \mathrm{dom}(f)$:

$$\langle \nabla f(x) - 0, x - x^* \rangle \geq \mu \|x - x^*\|^2 \quad \Rightarrow \quad \boxed{\mu \|x - x^*\| \leq \|\nabla f(x)\|}$$

Polyak-Łojasiewicz condition

Proposition (Polyak-Łojasiewicz condition)

Let $f \in \Gamma_0(\mathbb{R}^n)$ and let $\mu > 0$. Suppose that f is differentiable on $\mathrm{dom}(f)$, that f is μ-strongly convex and that there exists $x^* \in \arg\min f(x)$. Then:

$$(\forall x \in \mathrm{dom}(f)) \quad \boxed{f(x) - \min_x \; f(x) \leq \frac{1}{2\mu}\|\nabla f(x)\|^2} \qquad (*)$$

Proof.

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

- "Gradient grows as a quadratic function as we increase f". Important condition for achieving fast convergence rates!
- $(*)$ holds also for non-strongly convex functions (e.g., $\frac{1}{2}\|Ax - y\|^2$ for A not positive definite)

Lipschitz smoothness (L-smoothness)

In the framework of first-order optimisation methods, it's important to provide conditions on the growth of functions considered.

Definition (L-smoothness)

Let $f \in \Gamma_0(\mathbb{R}^n)$ be differentiable. We say that f is an *L-smooth function* with constant $L \geq 0$ iff:

$$\exists L \geq 0 : \forall x, y \in \mathbb{R}^n \quad \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

Remark: For $f(x) = \frac{1}{2}\|Ax - y\|_2^2$, you can check $L = \|A^T A\| \leq \|A\|^2$.

Characterisations of L-smoothness

Theorem (characterisation of L-smooth functions)

Let $f : \mathbb{R}^n \to \mathbb{R}$ a convex differentiable function and let $L > 0$. The following statements are equivalent:

1. f is L-smooth
2. (descent lemma)
$$(\forall x, y \in \mathbb{R}^n) \quad f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2}\|x - y\|^2$$
3. $$(\forall x, y \in \mathbb{R}^n) \quad \frac{1}{2L}\|f(x) - f(y)\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle$$
4. $$(\forall x, y \in \mathbb{R}^n) \quad \frac{1}{L}\|f(x) - f(y)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle$$
5. $$(\forall x, y \in \mathbb{R}^n) \quad \langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L\|x - y\|^2$$
6. $\frac{L}{2}\| \cdot \|^2 - f(\cdot)$ is convex.

Comparing smoothness and strong convexity

- f is L-smooth if and only if:

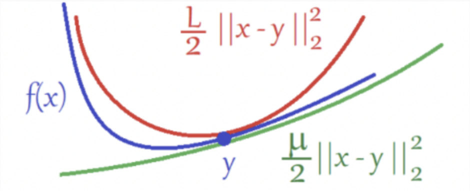
$$(\forall x, y \in \mathbb{R}^n) \quad f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|^2$$

- f is μ-strongly convex if and only if:

$$\forall x, y \in \mathrm{dom}(f), \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2$$

It can be proved that if f is a C^2 function there holds:

$$\mu \mathrm{Id} \preceq \nabla^2 f(x) \preceq L\mathrm{Id}, \qquad \text{for all } x$$



Smooth optimisation algorithms

Smooth optimisation algorithms

Gradient descent

Gradient descent

Gradient descent (GD) algorithm: ubiquitous in many applications for minimising (non-)convex, differentiable and proper functions $f : \mathbb{R}^n \to \overline{\mathbb{R}}$

Algorithm: Gradient Descent (GD) algorithm

Input: $\tau \in \left(0, \frac{2}{L}\right), x^0 \in \mathbb{R}^n$.
for $k \geq 0$ **do**

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

end for

- Choice of τ: important to guarantee convergence (need to be sufficiently small), it relates to L ($\sim$growth of f).
  Example: minimise $f(x) = x^2/2$. GD iteration: $x_{k+1} = (1 - \tau)x_k$, convergence for...?
- Convexity assumption: no dependence on x_0.
- Stopping criterion: relative error $\|x_{k+1} - x_k\| \leq \texttt{tol}$ or gradient check $\|\nabla f(x_{k+1})\| \leq \texttt{tol}$ (approaching 0).

Understanding the step-size upper bound

Lemma

For all $k \geq 0$, there holds:

$$\tau\left(1-\frac{\tau L}{2}\right)\|f(x_k)\|^2 \leq f(x_k)-f(x_{k+1}).$$

Thus, if $\tau < \frac{2}{L}$, then $f(x_{k+1}) \leq f(x_k)$, i.e. the GD algorithm is descending.

Proof. Since $x_{k+1}-x_k = -\tau\nabla f(x_k)$, then by the characterisation 2. of L-smoothness we have:

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

so the thesis follows.

Convergence of GD algorithm

Theorem (convergence of GD)

Let $(x_k)_k$ the sequence of iterates generated by GD. Then, if $\tau \in (0, 2/L)$ there holds:

$$f(x_k) - f(x^*) \leq \frac{\|x^0 - x^*\|^2}{2\tau k} = O\left(\frac{1}{k}\right)$$

Lemma (progress bounds)

For GD iterations with $\tau = 1/L$ there holds:

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

Proof. Using $x_{k+1} - x_k = -\frac{1}{L}\nabla f(x_k)$ we can apply the characterisation 2. to get:

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

We can use this progress bound to show improved rates under Polyak-Łojasiewicz condition (in particular, strongly convex functions).

Smooth optimisation algorithms

Convergence proof under PL condition

Linear convergence of GD under PL condition

Theorem (linear convergence of GD under PL)

Let $(x_k)_k$ the sequence of iterates generated by GD. Then, if $\tau = 1/L$ there holds:

$$f(x_k) - f(x^*) \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f(x^*)),$$

where, notice, $0 < \mu \leq L$.

Proof. Use the Lemma (progress bound) and the PL inequality:

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 \leq f(x_k) - \frac{\mu}{L}(f(x_k) - f(x^*)).$$

Subtracting $f(x^*)$ from both sides we get:

$$f(x_{k+1}) - f(x^*) \leq \left(1 - \frac{\mu}{L}\right)(f(x_k) - f(x^*)).$$

Applying this recursively gives the thesis since:

$$f(x_{k+1}) - f(x^*) \leq \left(1 - \frac{\mu}{L}\right)(f(x_k) - f(x^*)) \leq \left(1 - \frac{\mu}{L}\right)^2 (f(x_{k-1}) - f(x^*))$$
$$\leq \ldots \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f(x^*)).$$

To show $0 < \mu \leq L$, since by descent lemma we have that for all $v \in \mathbb{R}^n$:

$$f(x^*) \leq f(v) - \frac{1}{2L}\|\nabla f(v)\|^2.$$

Combining PL with this inequality we get:

$$\frac{1}{2\mu}\|\nabla f(v)\|^2 \geq f(v) - f(x^*) \geq \frac{1}{2L}\|\nabla f(v)\|^2 \quad \forall v \in \mathbb{R}^n \Rightarrow \mu \leq L$$

Do we practically see this gain in known problems?

$$f(x) = \frac{1}{2}\|Ax - y\|^2 + \frac{\lambda}{2}\|x\|^2, \quad \lambda > 0$$

f is λ-strongly convex. Convergence factor of the theorem:

$$\frac{\mu}{L} = \frac{\min\left\{\text{eig}(A^T A)\right\} + \lambda}{\max\left\{\text{eig}(A^T A)\right\} + \lambda}$$

- If $\lambda \gg 1$, then $\left(1 - \frac{\mu}{L}\right) \to 0$ hence faster convergence
- If $L \gg \mu$ ("small" PL), then this rate is not very informative, so in practice we observe the rate $O(1/k)$.
- The quantity L/μ is called the *condition number* of f (relates with the condition number of matrix $\nabla^2 f$ when f is C^2).

Smooth optimisation algorithms

Motivation for accelerated algorithms

Lower bounds for smooth optimisation

...back to standard GD iteration and $O(1/k)$ convergence rate.

[1] Nesterov, 2004, adapted from Chambolle-Pock, 2016

...back to standard GD iteration and $O(1/k)$ convergence rate.

Theorem (worst-case bounds[1])

For $x_0 \in \mathbb{R}^n$, $L > 0$ and $1 < k \leq \frac{1}{2}(n-1)$, there exists a convex, L-smooth function f s.t. for any first-order algorithm:

$$f(x_k) - f(x^*) \geq \frac{3L\|x_0 - x^*\|^2}{32(k+1)^2} = O\left(\frac{1}{(k+1)^2}\right).$$

It would be somehow 'optimal' finding convergence rates close to such lower (inevitable) bound...

How to fill the gap between $O(1/k)$ and $O(1/(k+1)^2)$ for convex functions?

[1] Nesterov, 2004, adapted from Chambolle-Pock, 2016

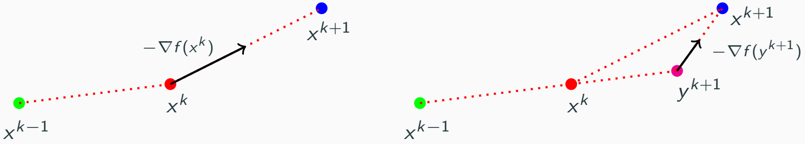
Accelerated smooth optimisation algorithms

Accelerated smooth optimisation algorithms

Nesterov acceleration of GD

Accelerated gradient descent

Idea: add inertia to "shift" the sequence of iterates.



Algorithm: Accelerated Gradient Descent (AGD) algorithm [2]

Input: $x_0 = x^{-1} \in \mathbb{R}^n$, $\tau \in \left(0, \frac{1}{L}\right]$, $t_0 = 0$.
for $k \geq 0$ **do**

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}$$

$$y_{k+1} = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1})$$

$$x_{k+1} = y_{k+1} - \tau \nabla f(y_{k+1})$$

end for

[2] Nesterov, 1983

A note on the sequence

Lemma (behaviour of the sequence (t_k))

Let t_0 and the sequence t_k be defined by:

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}.$$

Then $t_k \geq \frac{k+2}{2}$ for all $k \geq 0$. In particular, $t_k \to \infty$.

Proof: by induction. For $k = 0$ we have $t_0 \geq 1$. Suppose that the claim holds for some k, meaning that $t_k \geq \frac{k+2}{2}$. Want to show:

$$t_{k+1} \geq \frac{k+1+2}{2} = \frac{k+3}{2}.$$

Using recursion and $2t_k \geq k + 2$ (induction)

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2} \geq \frac{1 + \sqrt{1 + (k+2)^2}}{2} \geq \frac{1 + \sqrt{(k+2)^2}}{2} = \frac{k+3}{2}.$$

Remark: any sequence $(t_k)_k$ satisfying $t_{k+1}^2 - t_{k+1} \leq t_k^2$, $k \geq 0$ works (Chambolle, Dossal, 2015).

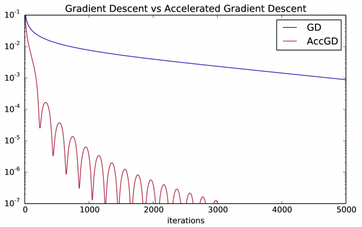
Accelerated convergence result

Theorem (convergence of AGD)[3]

Let $(x_k)_k$ the sequence of iterates generated by AGD. Then, there holds:

$$f(x_k) - f(x^*) \leq \frac{2\|x^0 - x^*\|^2}{\tau(k+1)^2}.$$

Get *faster*, at $O\left(\frac{1}{(k+1)^2}\right)$ to a reasonably accurate approximation of x^*.



. . . proof is quite technical. You'll see this in the case of non-smooth problems tomorrow.

[3] Nesterov, 2004, Chambolle-Pock, 2016

Accuracy view point

How many iterations are needed for such algorithms to achieve ε-accuracy, i.e.

$$\boxed{f(x_k) - f(x^*) \leq \varepsilon}$$

- GD: for all $k \geq 0$ such that $k \geq \lceil C/\varepsilon \rceil$
- AGD: for all $k \geq 0$ such that $k \geq \lceil C/\sqrt{\varepsilon} - 1 \rceil$
- GD + PL: for all $k \geq 0$ such that $k \geq \lceil C \log(1/\varepsilon) \rceil$

Conclusions

We focus on convex, smooth optimisation problems arising in applications (e.g., imaging inverse problems).

- We revised basic notions for having well-posedness of the underlying problem
- We considered GD as a reference first-order algorithm
- We commented on the improved speed achieved by GD whenever the underlying function enjoys *further regularity* (PL + strong convexity)
- We discussed Nesterov acceleration for improving convergence speed in convex cases

How to explore analogous ideas in the structured smooth+non-smooth setting?

Questions?

calatroni@i3s.unice.fr