Lecture 2: Convex smooth optimisation, Tikhonov regularisation

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

Inverse problems in biological imaging

MSc Data Science and Artificial Intelligence

January 23 2023

Table of contents

Introduction

Motivation

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

$$\min_{x \in \mathbb{R}^n} F(x), \quad x \in \mathbb{R}^n \text{ is a vectorised image of size } n_1 \times n_2 = n$$

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

- F is smooth $\to$ gradient descent algorithms
- $F := f + g$, f smooth & g non-smooth $\to$ proximal-gradient algorithms
- $F := f + \|x\|_0$ with f smooth and

$$\|x\|_0 := \# \{i : x_i \neq 0\} .$$

Such problems often appear in:

- **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. . .

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.

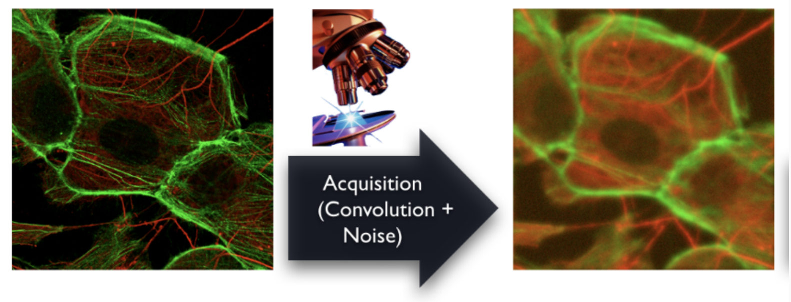
Optimisation for inverse problems in imaging

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

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

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

$$A \text{ is a convolution matrix} \quad Ax \Leftrightarrow h * X$$



Optimisation for inverse problems in imaging

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

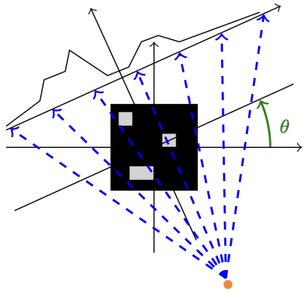
where $m \leq n$ and $\mathcal{T} : \mathbb{R}^m \to \mathbb{R}^m$ models noise degradation.

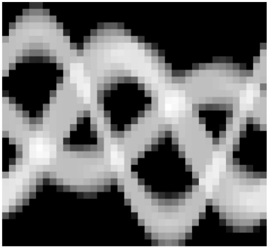
- **Image restoration** (denoising, deconvolution, super-resolution)

  $$A \text{ is a convolution matrix} \quad Ax \Leftrightarrow h * X$$

- **Image reconstruction** (e.g., medical imaging)

  $$A \text{ represents line integrals at a certain angle } \theta \quad Ax = R_\theta x$$





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.

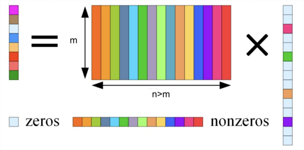
- **Image restoration** (denoising, deconvolution, super-resolution)

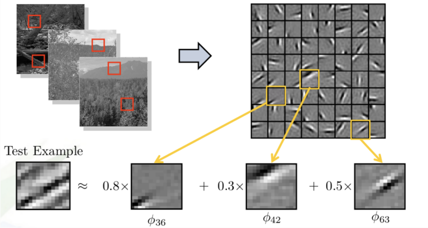
$$A \text{ is a convolution matrix} \quad Ax \Leftrightarrow h * X$$

- **Image reconstruction** (e.g., medical imaging)

$$A \text{ represents line integrals at a certain angle } \theta \quad Ax = R_\theta x$$

- **Dictionary representation** (data analysis, vision): $x = Dw$





Bad positioning of inverse filtering, Max. Likelihood approach

$$y = Ax + n$$

Naive approach: inverse filtering approach:

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

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

Bad positioning of inverse filtering, Max. Likelihood approach

$$y = Ax + n$$

Naive approach: inverse filtering approach:

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

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

Maximum-likelihood approach: find estimate $\mathbb{R}^n \ni x^* \approx x$ by solving

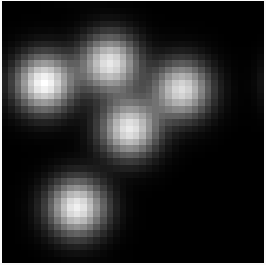
$$\boxed{x^* \in \arg\min_{x \in \mathbb{R}^n} \; F(x) = \frac{1}{2}\|Ax - y\|^2}$$

Regularisation idea

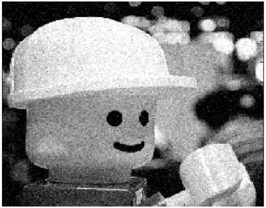
Consider instead:

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

- where, e.g., $f(x) = \frac{1}{2}\|Ax - y\|^2$ is the data fidelity term whose form relates to noise statistics (Gaussian, Poisson...)
- g is a regularisation term which encodes *a priori* information on the desired solution...



few non-zeros



piecewise constant

piecewise linear

Regularisation: Bayesian motivation

Following a Bayesian/MAP approach consider:

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

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

$$x^* \in \arg\max_x \ \pi_{X|Y}(x|y) = \arg\max_x \ \frac{\pi_{Y|X}(y|Ax;\theta_f)\pi_X(x;\theta_g)}{\pi_Y(y)}$$

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

Now, if $\pi_X(x;\theta_g) = e^{-\theta_g g(x)}, \theta_g > 0$ and $\pi_{Y|X}(y|Ax;\theta_f) = e^{-\theta_f f(x)}, \theta_f > 0$, then:

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

$\lambda > 0$ is the **regularisation parameter**: it weights the amount of regularisation against the trust in the data.

Some examples of regularisation

- **Modelling sparsity**: the prior is "the image has *few* non-zero entries". Natural choice is $g(x) = \|x\|_0$ (complex problem), so an alternative is:

$$g(x) = \|x\|_1 = \sum_{i=1}^{n} |x_i|$$

Some examples of regularisation

- **Modelling sparsity**: the prior is "the image has *few* non-zero entries". Natural choice is $g(x) = \|x\|_0$ (complex problem), so an alternative is:

$$g(x) = \|x\|_1 = \sum_{i=1}^{n} |x_i|$$

- **Modelling sparsity in some basis**: let $W \in \mathbb{R}^{n \times p}$ a dictionary. Assume that $x = Wa$, with $a \in \mathbb{R}^p$ (synthesis view point) and consider:

$$g(x) = g(a) = \|Wa\|_1 \to \min_{a \in \mathbb{R}^p} \frac{1}{2}\|AWa - y\|^2 + \lambda \|Wa\|_1$$

Some examples of regularisation

- **Modelling sparsity**: the prior is "the image has *few* non-zero entries". Natural choice is $g(x) = \|x\|_0$ (complex problem), so an alternative is:

$$g(x) = \|x\|_1 = \sum_{i=1}^{n} |x_i|$$

- **Modelling sparsity in some basis**: let $W \in \mathbb{R}^{n \times p}$ a dictionary. Assume that $x = Wa$, with $a \in \mathbb{R}^p$ (synthesis view point) and consider:

$$g(x) = g(a) = \|Wa\|_1 \to \min_{a \in \mathbb{R}^p} \frac{1}{2}\|AWa - y\|^2 + \lambda \|Wa\|_1$$

- **Modelling piece-wise constancy**: constant regions in images = regions with little variations = regions with small gradients. Hence natural choice is:

$$g(x) = \frac{1}{2}\|Dx\|_{2,2}^2 = \frac{1}{2}\sum_{i=1}^{n}\left((D_h x)_i^2 + (D_v x)_i^2\right), \quad g(x) = \|Dx\|_{2,1} = \sum_{i=1}^{n}\sqrt{(D_h x)_i^2 + (D_v x)_i^2}$$

Some examples of regularisation

- **Modelling sparsity**: the prior is "the image has *few* non-zero entries". Natural choice is $g(x) = \|x\|_0$ (complex problem), so an alternative is:

$$g(x) = \|x\|_1 = \sum_{i=1}^{n} |x_i|$$

- **Modelling sparsity in some basis**: let $W \in \mathbb{R}^{n \times p}$ a dictionary. Assume that $x = Wa$, with $a \in \mathbb{R}^p$ (synthesis view point) and consider:

$$g(x) = g(a) = \|Wa\|_1 \to \min_{a \in \mathbb{R}^p} \frac{1}{2}\|AWa - y\|^2 + \lambda \|Wa\|_1$$

- **Modelling piece-wise constancy**: constant regions in images = regions with little variations = regions with small gradients. Hence natural choice is:

$$g(x) = \frac{1}{2}\|Dx\|_{2,2}^2 = \frac{1}{2}\sum_{i=1}^{n}\left((D_h x)_i^2 + (D_v x)_i^2\right), \quad g(x) = \|Dx\|_{2,1} = \sum_{i=1}^{n}\sqrt{(D_h x)_i^2 + (D_v x)_i^2}$$

- **Modelling piece-wise linearity**: piece-wise linear regions in images = regions with small Hessian...

$$g(x) = \frac{1}{2}\|D^2 x\|^2$$

...how to choose a good prior? Open question!

A smooth example: Tikhonov regularisation

Idea: smooth regularisation of the image in some basis.

$$g(x) = \frac{1}{2}\|Bx\|_2^2, \quad B \in \mathbb{R}^{N \times n}$$

- **Ridge regularization**: $N = n$ and $B = \mathrm{Id}$:

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

Reduces high-values of the image x.

A smooth example: Tikhonov regularisation

Idea: smooth regularisation of the image in some basis.

$$g(x) = \frac{1}{2}\|Bx\|_2^2, \quad B \in \mathbb{R}^{N\times n}$$

- **Ridge regularization**: $N = n$ and $B = \mathrm{Id}$:

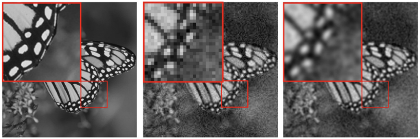
$$\boxed{x^* \in \arg\min_{x\in\mathbb{R}^n} \; \frac{1}{2}\|Ax-y\|^2 + \lambda\|x\|^2}$$

  Reduces high-values of the image x.

- **Sobolev regularization** (Tikhonov, Arsenin, '83): $N = 2n$ and $B = D = \begin{pmatrix} D_h \\ D_v \end{pmatrix}$:

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

  Reduces high-values of the finite-difference image **gradient** Dx (hence oscillations, but also edge sharpness).



Notation, preliminaries & basic notions

Notation and basic tools

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

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

- Convex set $C \subset X$

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

Proper functions

Minimal property to have well-defined minimisation problems.

Definition (proper function)

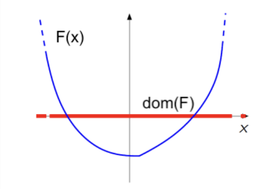
A function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ is said *proper* iff

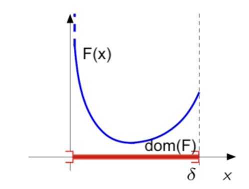
$$\exists x \in \mathbb{R}^n \quad \text{such that} \quad F(x) \neq +\infty.$$

We define

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

Clearly, F is proper $\Leftrightarrow \mathrm{dom}(F) \neq \emptyset$.

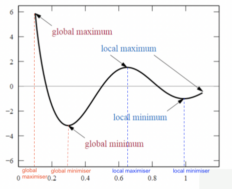




Global/local minimisers

Given a proper function:

- **global minimiser**: $x^* \in \mathbb{R}^n$: $F(x^*) \leq F(x)$ for $\underline{\text{every } x \in \mathbb{R}^n}$.
- **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 $\underline{\text{every } x \in B_\delta(x^*)}$.

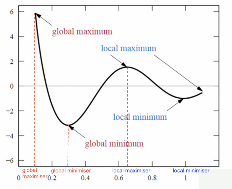


$$\min_{x \in \mathbb{R}^n} F(x) \quad \text{VS} \quad \arg\min_{x \in \mathbb{R}^n} F(x)$$

Global/local minimisers

Given a proper function:

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



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

Definition (set of minimisers)

The **set** of 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

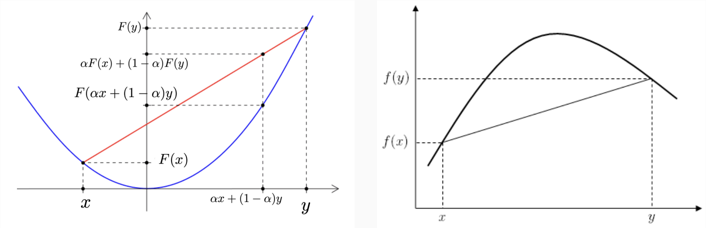
Convex functions

Definition (convex function)

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



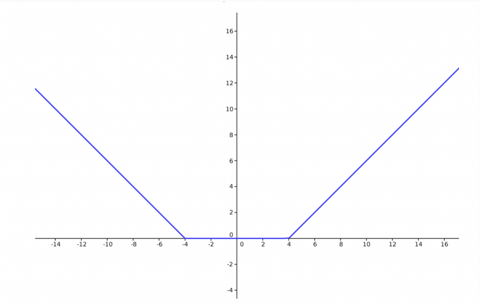
Convex/concave function

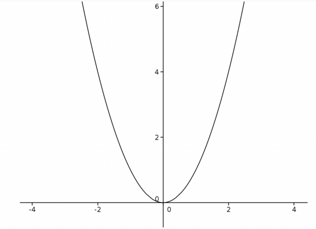
Definition (convex function)

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

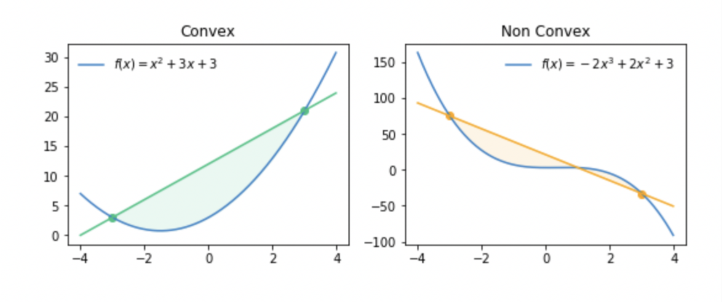
Convex functions

Definition (convex function)

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

Definition (convex function)

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

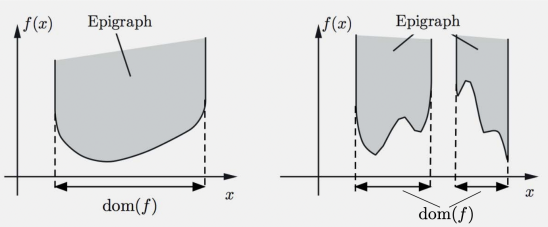
Useful properties

Proposition (epigraph of convex functions is convex set)

Let F be proper. Then F is convex if and only if

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

is convex.



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.

Notation, preliminaries & basic notions

Lower semi-continuity & coercivity

Lower semi-continuity

Definition (lower semi-continuity)

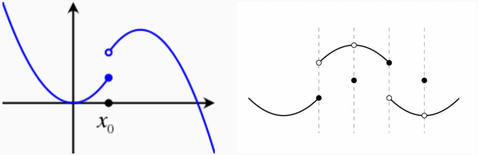
Let F be a proper function. F is *lower semi-continuous (l.s.c.) at* $x \in \mathbb{R}^n$ iff

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

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.



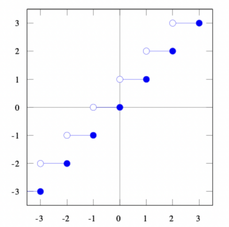
Left: lower l.s.c. **Right**: where the function is lower l.s.c.?

Examples of l.s.c. functions

- The functions $F : \mathbb{R} \to \mathbb{R}$

$$F(x) = |x|_0 = \begin{cases} 0 & \text{if } x = 0 \\ 1 & \text{if } x \neq 0 \end{cases}, \qquad F(x) = \lceil x \rceil = \min \{k \in \mathbb{Z} : x \leq k\}$$

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



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

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

Coercivity

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

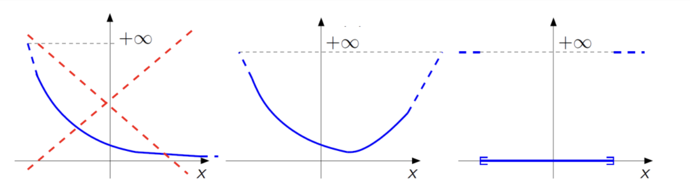
Definition (coercivity)

Let F be proper. 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.
- $F : \mathbb{R}^2 \to \mathbb{R}_+$, $F(x_1, x_1) = x_1^2 + x_2^2$ is coercive.
- $F : \mathbb{R}^2 \to \mathbb{R}_+$, $F(x_1, x_2) = x_1^2 - 2x_1x_2 + x_2^2 = (x_1 - x_2)^2$ is **not** coercive. Why?



Existence of minimisers

Theorem (existence of minimisers)

If F is proper, l.s.c. and coercive, then $\operatorname{argmin} F \neq \emptyset$.

Note: generalises the Bolzano-Weirestrass theorem holding for problems

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

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

Existence of minimisers

Theorem (existence of minimisers)

If F is proper, l.s.c. and coercive, then $\operatorname{argmin} F \neq \emptyset$.

Note: generalises the Bolzano-Weirestrass theorem holding for problems

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

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

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 (importance of coercivity): $F \in \Gamma_0(X) \not\Rightarrow F$ 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

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.

Notation, preliminaries & basic notions

Differentiability and L-smoothness

Gâteaux differentiability

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

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 $x \in \mathrm{dom}(f)$. 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 short, *differentiable*) in x and denote by $\nabla f(x) = w$ 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:

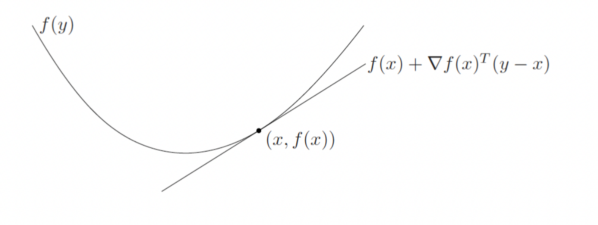
$$x^* \text{ is a minimiser of } f(\cdot) \quad \Longleftrightarrow \quad \nabla f(x^*) = 0.$$

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. $\forall x, y \in \mathrm{dom}(f)$, $f(y) \geq \overbrace{f(x) + \langle \nabla f(x), y - x \rangle}^{\phi(y;x):=}$;
3. $\forall x, y \in \mathrm{dom}(f)$, $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$.

- the function $\phi(y; x)$ is an affine lower bound/estimator of f
- the tangent to f is below f at all points.



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 ∇f is L-Lipschitz continuous:

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

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

Smoothness VS strong convexity

- f is L-smooth if and only if:

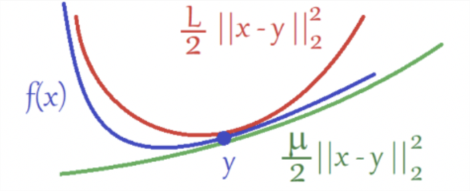
$$(\forall w, z \in \mathrm{dom}(f)) \quad f(z) \leq f(w) + \langle \nabla f(w), z - w \rangle + \frac{L}{2}\|w - z\|^2$$

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

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

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

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



Strong convexity entails better convergence properties.

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 $\|x_{k+1} - x_k\| \leq$ `tol` or gradient check $\|\nabla f(x_{k+1})\| \leq$ `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 $f(x_{k+1}) < f(x_k)$, i.e. the GD algorithm is descending.

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

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \tau\langle\nabla f(x_k), \nabla f(x_k)\rangle + \frac{L}{2}\tau^2\|\nabla f(x_k)\|^2, \\ &= f(x_k) + \left(\frac{L\tau}{2}-1\right)\tau\|\nabla f(x_k)\|^2 \end{aligned}$$

so the thesis follows.

Convergence of GD algorithm

Theorem (convergence of GD)

Let $(x_k)_k$ the GD sequence and $x^* \in \arg\min f$. Then, if $\tau \in (0, 2/L)$, there holds:

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

Remarks:

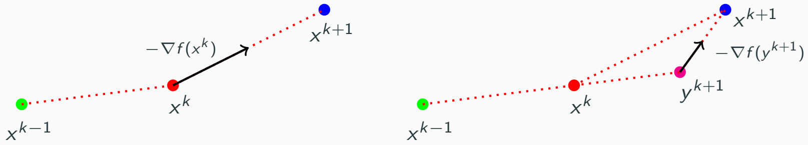
- Convergence in function values with speed $O(1/k)$.
- Note that the constant is unknown, as x^* is, but it is finite.

Smooth optimisation algorithms

Accelerated GD

Accelerated gradient descent

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



Algorithm: Accelerated Gradient Descent (AGD) algorithm [1]

Input: $x_0 = x_{-1} \in \mathbb{R}^n$, $\tau \in \left(0, \frac{1}{L}\right]$, $t_0 = 1$.

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

Note: For the use of AGD in inverse problems, the sequence (y_k) does not have to be related with the data y, which typically appears in the expression of ∇f.

[1] Nesterov, 1983

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)

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

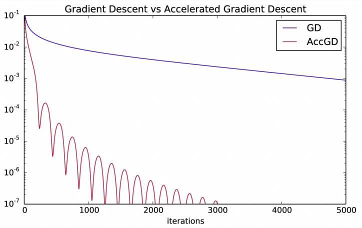
Accelerated convergence result

Theorem (convergence of AGD)[2]

Let $(x_k)_k$ be the AGD sequence and $x^* \in \arg\min f$. Then, there holds:

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

Get *faster* to a reasonably accurate approximation of x^*.



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

Accuracy viewpoint

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

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

- GD: all $k \geq 0$ such that $k \geq \lceil C/\varepsilon \rceil$
- AGD: all $k \geq 0$ such that $k \geq \lceil C/\sqrt{\varepsilon} - 1 \rceil$

First order optimisation or more?

This quadratic rate matches the *worst-case* lower bound for first-order optimisation methods (i.e., methods using only gradient information for optimisation the function).

Other possibility: **Newton's method**, but more involved:

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

where $H_f(x_k)$ is the Hessian of f evaluated in x_k (could be hard to invert).

Application to inverse problems

Maximum-likelihood approach

For $A \in \mathbb{R}^{m \times n}$ and $n \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$, observe noisy/blurred image y through:

$$\boxed{y = Ax + n}$$

Consider maximum-likelihood functional:

$$\boxed{x^* \in \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - y\|^2 = f(x)}$$

How to set-up GD iteration?

- Compute the (Gateaux) gradient $\nabla f(x)$
- Compute its Lipschitz constant L
- Choose $\tau < 2/L$ (as big as possible) and $x_0 \in \mathbb{R}^n$ (often, $x_0 = y$ in convex problems) to launch the algorithm
- Choose stopping criterion:

$$\|x_{k+1} - x_k\| \leq \epsilon, \quad \|f(x_{k+1}) - f(x_k)\| \leq \epsilon, \quad \|\nabla f(x_k)\| \leq \epsilon$$

Tikhonov regularisation

$$x^* \in \underset{x\in\mathbb{R}^n}{\arg\min}\ \frac{1}{2}\|Ax-y\|^2 + \lambda\|Bx\|^2 = f(x)$$

- Compute gradient $\nabla f = \nabla f_1 + \nabla f_2$
- Estimate L observing that

$$\|\nabla f(w) - \nabla f(z)\| \leq \|\nabla f_1(w) - \nabla f_1(z)\| + \|\nabla f_2(w) - \nabla f_2(z)\| \leq \underbrace{(L_1 + L_2)}_{L}\|w - z\|$$

- Choose $\tau < 2/L$ (as big as possible) and $x_0 \in \mathbb{R}^n$ to launch the algorithm

Conclusions

We focused on convex, smooth optimisation problems arising in imaging inverse problems.

- We revised basic notions for having well-posedness of the underlying problem and basic optimisation tools
- We considered GD as a reference first-order algorithm
- We discussed Nesterov acceleration for improving convergence speed

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

Questions?

calatroni@i3s.unice.fr