Multiresolution in image restoration

2- Basics in optimisation

Elisa Riccietti, Nelly Pustelnik

Hilbert spaces

A (real) **Hilbert space** $\mathcal{H}$ is a complete real vector space endowed with an inner product $\langle \cdot \mid \cdot \rangle$. The associated norm is

$$(\forall x \in \mathcal{H}) \qquad \|x\| = \sqrt{\langle x \mid x \rangle}.$$

- Particular case: $\mathcal{H} = \mathbb{R}^N$ (Euclidean space with dimension N).
- Course dedicated to finite dimension.

Optimization problems

Let

$$\begin{array}{rccc} f: & \mathcal{H} & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & f(x) \end{array}$$

We consider *unconstrained optimization problems*, that is problems of the form:

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

or equivalently

$$\text{Find} \quad \widehat{x} \in \operatorname*{Argmin}_{x \in \mathcal{H}} f(x)$$

Example: quadratic functions

A quadratic function is a function of the form

$$\begin{array}{rccll} q: & \mathbb{R}^N & \longrightarrow & \mathbb{R} & \\ & x & \longmapsto & q(x) & = \frac{1}{2} x^T A x - b^T x \\ & & & & = \frac{1}{2} \langle x, Ax \rangle - \langle b, x \rangle \\ & & & & = \frac{1}{2} \sum_{i,j=1}^{N} x_i a_{ij} x_j - \sum_{i=1}^{N} b_i x_i, \end{array}$$

with $A \in \mathbb{R}^{N \times N}$ symmetric, $b \in \mathbb{R}^N$.

Reminder about norms in finite dimension

- **Vectors**
  - Let $x = (x_i)_{1 \leq i \leq N} \in \mathbb{R}^N$.
  - ℓ_1-**norm**: $\|x\|_1 = \sum_i |x_i|$.
  - ℓ_2-**norm**: $\|x\|_2 = \sqrt{\sum_i x_i^2}$.

- **Matrices**
  - Let A be a real symmetric $N \times N$ matrix.
  - **Spectral/eigen decomposition**: A can be factored as

    $$A = Q \Lambda Q^\top$$

    where $Q \in \mathbb{R}^{N \times N}$ is orthogonal (i.e. $Q^\top Q = \mathrm{Id}$) and $\Lambda = \mathrm{diag}(\lambda_1, \dots \lambda_k)$ (where the real numbers λ_i are the eigenvalues of A).
    The column of Q form an orthonormal set of eigenvectors of A.
  - **Spectral norm**: $\|A\|_2 = \max_i |\lambda_i|$.
  - **Frobenius norm**: $\|A\|_F = \sqrt{\sum_i \lambda_i^2}$.

Linear operators

Let $\mathcal{H}$ and $\mathcal{G}$ be two Hilbert spaces.

An operator $L\colon \mathcal{H} \to \mathcal{G}$ is **linear** if for all $a, b \in \mathbb{R}$ and $x, y \in \mathcal{H}$

$$L(ax + by) = aL(x) + bL(y)$$

Linear operators

Let $\mathcal{H}$ and $\mathcal{G}$ be two Hilbert spaces.

A linear operator $L\colon \mathcal{H} \to \mathcal{G}$ is **bounded** (or continuous) if

$$\|L\| = \sup_{\|x\|_{\mathcal{H}} \leq 1} \|L(x)\|_{\mathcal{G}} < +\infty$$

Linear operators

Let $\mathcal{H}$ and $\mathcal{G}$ be two Hilbert spaces.

A linear operator $L\colon \mathcal{H} \to \mathcal{G}$ is **bounded** (or continuous) if

$$\|L\| = \sup_{\|x\| \leq 1} \|L(x)\| < +\infty$$

Linear operators

Let $\mathcal{H}$ and $\mathcal{G}$ be two Hilbert spaces.

A linear operator $L\colon \mathcal{H} \to \mathcal{G}$ is **bounded** (or continuous) if

$$\|L\| = \sup_{\|x\| \leq 1} \|L(x)\| < +\infty$$

- In finite dimension, every linear operator is bounded.

Linear operators

Let $\mathcal{H}$ and $\mathcal{G}$ be two Hilbert spaces.

A linear operator $L\colon \mathcal{H} \to \mathcal{G}$ is **bounded** (or continuous) if

$$\|L\| = \sup_{\|x\| \leq 1} \|L(x)\| < +\infty$$

- In finite dimension, every linear operator is bounded.

$\mathcal{B}(\mathcal{H}, \mathcal{G})$**:** Banach space of bounded linear operators from $\mathcal{H}$ to $\mathcal{G}$.

Norm and adjoint

Let $\mathcal{H}$ and $\mathcal{G}$ be two Hilbert spaces.
Let $L \in \mathcal{B}(\mathcal{H}, \mathcal{G})$. Its **adjoint** L^* is the operator in $\mathcal{B}(\mathcal{G}, \mathcal{H})$ defined as

$$(\forall (x, y) \in \mathcal{H} \times \mathcal{G}) \qquad \langle y \mid L(x) \rangle_{\mathcal{G}} = \langle L^*(y) \mid x \rangle_{\mathcal{H}} .$$

Example:

$$\text{If} \qquad L \colon \mathcal{H} \to \mathcal{H}^n \colon x \mapsto (x, \dots, x)$$

$$\text{then} \qquad L^* \colon \mathcal{H}^n \to \mathcal{H} \colon y = (y_1, \dots, y_n) \mapsto \sum_{i=1}^{n} y_i$$

Proof: $\langle L(x) \mid y \rangle = \langle (x, \dots, x) \mid (y_1, \dots, y_n) \rangle = \sum_{i=1}^{n} \langle x \mid y_i \rangle = \left\langle x \mid \sum_{i=1}^{n} y_i \right\rangle$

Norm and adjoint

Let $\mathcal{H}$ and $\mathcal{G}$ be two Hilbert spaces.
Let $L \in \mathcal{B}(\mathcal{H}, \mathcal{G})$. Its **adjoint** L^* is the operator in $\mathcal{B}(\mathcal{G}, \mathcal{H})$ defined as

$$(\forall (x, y) \in \mathcal{H} \times \mathcal{G}) \qquad \langle y \mid L(x) \rangle = \langle L^*(y) \mid x \rangle .$$

Example:

$$\text{If} \qquad L \colon \mathcal{H} \to \mathcal{H}^n \colon x \mapsto (x, \ldots, x)$$

$$\text{then} \qquad L^* \colon \mathcal{H}^n \to \mathcal{H} \colon y = (y_1, \ldots, y_n) \mapsto \sum_{i=1}^{n} y_i$$

Proof: $\langle L(x) \mid y \rangle = \langle (x, \ldots, x) \mid (y_1, \ldots, y_n) \rangle = \sum_{i=1}^{n} \langle x \mid y_i \rangle = \left\langle x \mid \sum_{i=1}^{n} y_i \right\rangle$

Norm and adjoint

- **About L^*:**
  - ⊙ **Compute gradient and proximity operator** operations
  - ⊙ Dual formulation
  - ⊙ Finite dimensions: If $L \in \mathcal{B}(\mathbb{R}^N, \mathbb{R}^K)$ then $L^*(y) = L^\top y$.
  - ⊙ Check the correct implementation by using its definition

$$(\forall (x, y) \in \mathbb{R}^N \times \mathbb{R}^K) \qquad \langle L(x) \mid y \rangle = \langle x \mid L^*(y) \rangle$$

- **About $\|L\|$:**
  - ⊙ Required for **gradient-based** algorithms;
  - ⊙ We have $\|L^*\| = \|L\|$;

Approximate the function using derivatives: Taylor's theorem

Most optimization methods employs approximations of the nonlinear objective functions built with polynomials. The theoretical foundation comes from the Taylor's theorem.

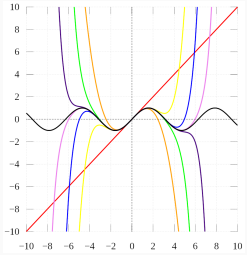


Figure 1: Taylor approximation of sin(x) for P(x) of degree 1, 3, 5, 7, 9, 11 and 12

Important quantities: gradient

If $f\colon \mathbb{R}^N \to \mathbb{R}$ is differentiable function in $x \in \mathbb{R}^N$, the **gradient of f at x** is $\nabla f(x) \in \mathbb{R}^N$ and its components are the partial derivatives of f:

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_j} \right)_{1 \le i \le N}$$

- Example: Let $x \in \mathbb{R}^N$, $z \in \mathbb{R}^K$ and $A \in \mathbb{R}^{K \times N}$ and $f(x) = \frac{1}{2}\langle x, Ax \rangle - \langle z, x \rangle$, then
$$\nabla f(x) = Ax - z$$
- Example: Let $x \in \mathbb{R}^N$, $z \in \mathbb{R}^K$ and $A \in \mathbb{R}^{K \times N}$ and $f(x) = \frac{1}{2}\|Ax - z\|^2$, then
$$\nabla f(x) = A^\top (Ax - z)$$

Important quantities: Hessian matrix

The second derivative of a real-valued function $f\colon \mathbb{R}^N \to \mathbb{R}$ or the Hessian matrix of f at x , denoted $H_f(x) \in \mathbb{R}^{N\times N}$, is given by

$$H_f(x) = \left(\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right)_{1\le i\le N, 1\le j\le N}$$

- Example: Let $x \in \mathbb{R}^N$, $z \in \mathbb{R}^K$ and $A \in \mathbb{R}^{K\times N}$ and $f(x) = \frac{1}{2}\langle x, Ax\rangle - \langle z, x\rangle$, then
$$H_f(x) = A$$
- Example: Let $x \in \mathbb{R}^N$, $z \in \mathbb{R}^K$ and $A \in \mathbb{R}^{K\times N}$ and $f(x) = \frac{1}{2}\|Ax - z\|^2$, then
$$H_f(x) = A^\top A$$

Scalar function $f : \mathbb{R} \to \mathbb{R}$.

Taylor's theorem (scalar version) Let $\ell \geq 1$ be an integer and let the function $f : \mathbb{R} \to \mathbb{R}$ be ℓ times differentiable at the point $a \in \mathbb{R}$. Then there exists a function $h_\ell : \mathbb{R} \to \mathbb{R}$ such that

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(\ell)}(a)}{\ell!}(x-a)^\ell + h_\ell(x)(x-a)^\ell,$$

and

$$\lim_{x \to a} h_\ell(x) = 0.$$

- This is called the **Peano form** of the remainder.

Taylor polynomial

The polynomial appearing in Taylor's theorem is the ℓ-th order Taylor polynomial:

$$P_\ell(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(\ell)}(a)}{\ell!}(x-a)^\ell$$
$$= \sum_{i=0}^{\ell} \frac{f^{(i)}(a)}{i!}(x-a)^i$$

of the function f at the point a.

The Taylor polynomial is the **unique "asymptotic best fit" polynomial** in the sense that if there exists a function h_ℓ and a ℓ-th order polynomial p such that

$$f(x) = p(x) + h_\ell(x)(x-a)^\ell, \quad \lim_{x \to a} h_\ell(x) = 0,$$

then $p = P_\ell$.

Taylor's theorem in N dimensions

Genralisation to functions defined in $\mathbb{R}^N$. Assume $f : \mathbb{R}^N \to \mathbb{R}$.

Taylor polynomials in $\mathbb{R}^N$:

1. **First-order Taylor polynomial**: Best linear approximation of f in a neighbourhood of x. Let f be differentiable and let $h \neq 0$, we define

$$T_1(x, h) = f(x) + \nabla f(x)^T h. \tag{1}$$

2. **Second-order Taylor polynomial**: Best quadratic approximation of f in a neighbourhood of x. Let f be twice continuously differentiable and $h \neq 0$, we define

$$T_2(x, h) = f(x) + \nabla f(x)^T h + \frac{1}{2} h^T H_f(x) h. \tag{2}$$

Remark: Usually in $\mathbb{R}^N$ we stop at order two: higher order derivatives are difficult or expensive to evaluate and are not used in practice.

Class of functions

$$\text{Find} \quad \widehat{x} \in \underset{x \in \mathcal{H}}{\operatorname{Argmin}} \quad f(x)$$

Class of functions $f \in C_{\beta}^{\ell,p}(\mathbb{R}^N)$:

- Continuously differentiable function
- Lipschitz-continuous derivatives

Class of functions $f \in \Gamma_0(\mathcal{H})$:

- Proper function
- Lower semi-continuous function
- Convex function

Continuously differentiable functions

Definition (Nesterov 2004) $f \in \mathcal{C}^{\ell}(\mathbb{R}^N)$ denotes the class of ℓ times continuously differentiable on $\mathbb{R}^N$.

Definition (Nesterov 2004) $f \in \mathcal{C}_{\beta}^{\ell,p}(\mathbb{R}^N)$ denotes the class of ℓ times continuously differentiable on $\mathbb{R}^N$ such that its p-th derivative is Lipschitz continuous on $\mathbb{R}^N$ with the constant β.

Definition (Nesterov 2004) $\mathcal{C}_{\beta}^{1,1}(\mathbb{R}^N)$ denotes the class of functions with a Lipschitz continuous gradient: $\|\nabla f(y) - \nabla f(x)\| \leq \beta \|y - x\|$.

Definition (Nesterov 2004) f belongs to $\mathcal{C}_{\beta}^{2,1}(S)$ with $S \subset \mathbb{R}^N$ if and only if $(\forall x \in S) \qquad \|\nabla^2 f(x)\| \leq \beta$.

L-smooth

A function $f \in \mathcal{C}^{1,1}_{\beta}$ is said to be Lipschitz smooth with constant β or β-smooth, that is **its gradient is Lipschitz continuous** with constant β:

$$(\forall (x, y) \in \mathbb{R}^N \times \mathbb{R}^N) \quad \|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$$

Remark:

- If f is twice differentiable, a function is β smooth if

$$(\forall x \in \mathbb{R}^N) \quad \nabla^2 f(x) \leq \beta \cdot \mathrm{Id}$$

- Particular case: $f = \|A \cdot -z\|_2^2$ is β-smooth with $\beta = \sigma_{\mathsf{max}}(A^\top A) = \|A\|^2$.

Functional analysis: definitions

Let $f : \mathcal{H} \to]-\infty, +\infty]$ where $\mathcal{H}$ is a Hilbert space.

- The **domain** of f is $\operatorname{dom} f = \{x \in \mathcal{H} \,|\, f(x) < +\infty\}$.
- The function f is **proper** if $\operatorname{dom} f \neq \varnothing$.

Let $f : \mathcal{H} \to]-\infty, +\infty]$. The **epigraph** of f is

$$\operatorname{epi} f = \left\{(x, \zeta) \in \operatorname{dom} f \times \mathbb{R} \;\middle|\; f(x) \leq \zeta\right\}$$

Lower semi-continuity

Let $f : \mathcal{H} \to]-\infty, +\infty]$.
f is a **lower semi-continuous** function on $\mathcal{H}$ if and only if $\operatorname{epi} f$ is closed

Lower semi-continuity

Let $f : \mathcal{H} \to]-\infty, +\infty]$.
f is a **lower semi-continuous** function on $\mathcal{H}$ if and only if $\operatorname{epi} f$ is closed

- Examples:
  - ⊙ **Do not allow for strict constraints** e.g. $Ax < b$ or $x > 0$;
  - ⊙ Allow for inequality or equality constraints e.g. $Ax = b$, $Ax \leq b$ or $x > 0$;

Lower semi-continuity

Let $f : \mathcal{H} \to]-\infty, +\infty]$.
f is a **lower semi-continuous** function on $\mathcal{H}$ if and only if $\operatorname{epi} f$ is closed

- Examples:
  - ⊙ **Do not allow for strict constraints** e.g. $Ax < b$ or $x > 0$;
  - ⊙ Allow for inequality or equality constraints e.g. $Ax = b$, $Ax \leq b$ or $x > 0$;

- Properties:
  - ⊙ Every continuous function on $\mathcal{H}$ is l.s.c.
  - ⊙ **Every finite sum of l.s.c. functions is l.s.c.**

Fejér monotonicity

Let $S \subseteq \mathbb{R}^N$ be a nonempty set. A sequence $\left(x^{[k]}\right)_{k\geq 0}$ is said to be *Fejér monotone with respect to* S if

$$\|x^{[k+1]} - x^\star\| \leq \|x^{[k]} - x^\star\|, \qquad \forall x^\star \in S, \quad \forall k \geq 0.$$

Equivalently,

$$\|x^{[k+1]} - x^\star\|^2$$

is nonincreasing in k $\forall x^\star \in S$.

Convex set

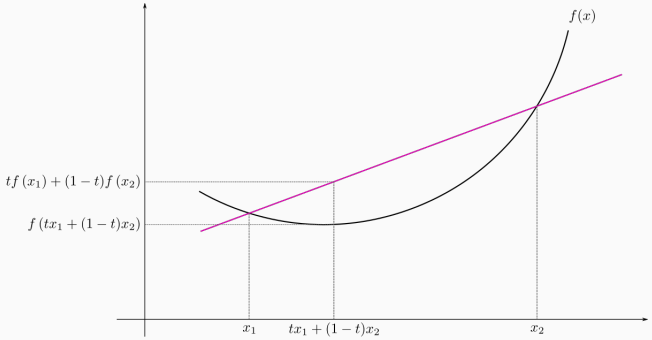
$C \subset \mathcal{H}$ is a **convex set** if

$$(\forall (x, y) \in C^2)(\forall \alpha \in]0, 1[) \qquad \alpha x + (1 - \alpha) y \in C$$

Convex function: definitions

$f : \mathcal{H} \to]-\infty, +\infty]$ is a **convex function** if

$$\big(\forall (x, y) \in \mathcal{H}^2\big)(\forall t \in (0, 1)) \qquad f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$



Convex functions: definition

$f : \mathcal{H} \to]-\infty, +\infty]$ is convex $\Leftrightarrow$ its epigraph is convex.

Convex functions: definition

$f : \mathcal{H} \to]-\infty, +\infty]$ is convex $\Leftrightarrow$ its epigraph is convex.

- **Properties**:
  - $\odot$ Composition of an increasing convex funct. and a convex funct. is convex.
  - $\odot$ If $f : \mathcal{H} \to]-\infty, +\infty]$ is convex, then $\operatorname{dom} f$ is convex.
  - $\odot$ $f : \mathcal{H} \to [-\infty, +\infty[$ is concave if $-f$ is convex.
  - $\odot$ Every finite **sum of convex functions is convex**.

Example: indicator function

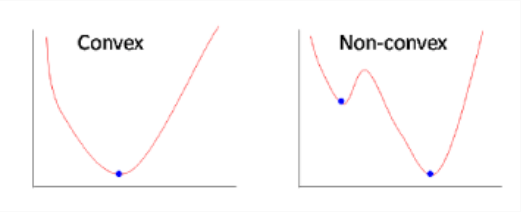
Let $C \subset \mathcal{H}$.

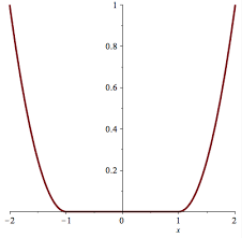
The **indicator function of** C is

$$(\forall x \in \mathcal{H}) \qquad \iota_C(x) = \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{otherwise.} \end{cases}$$

Remark: $\iota_C \in \Gamma_0(\mathcal{H}) \Leftrightarrow C$ is a nonempty closed convex set.

Convex vs non-convex





Strictly convex functions

Let $\mathcal{H}$ be a Hilbert space. Let $f\colon \mathcal{H} \to]-\infty, +\infty]$.
f is **strictly convex** if

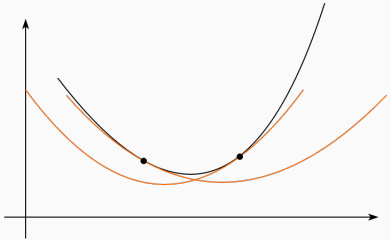
$$(\forall x \in \operatorname{dom} f)(\forall y \in \operatorname{dom} f)(\forall \alpha \in]0,1[)$$
$$x \neq y \quad \Rightarrow \quad f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha) f(y).$$

Strongly convex functions

- f is *strongly convex* with parameter $\rho > 0$ if one of the following statement holds:
  (i) $(\forall x, y \in \mathcal{H}) \qquad f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\rho}{2}\|y - x\|^2$.
  (ii) $f - \frac{\rho}{2}\|\cdot\|_2^2$ is convex.

- A strongly convex function is also strictly convex, but not vice versa.

- Strong convexity means that there exists a quadratic lower bound on the growth of the function.



Characterising convexity

Let C be a convex subset of $\mathbb{R}^N$. Suppose the function $f : C \longrightarrow \mathbb{R}$ is twice differentiable over C. Then, the followings are equivalent:

(i) f is convex.

(ii) $f(y) \geq f(x) + \nabla f(x)^T (y - x)$, for all $x, y \in C$.

(iii) $\nabla^2 f(x) \succeq 0$, for all $x \in C$.

Let C be a convex subset of $\mathbb{R}^N$. Suppose the function $f : C \longrightarrow \mathbb{R}$ is twice differentiable over C. Then:

(i) f is strictly convex if and only if $f(y) > f(x) + \nabla f(x)^T (y - x)$, for all $x, y \in C$, $x \neq y$.

(ii) $\nabla^2 f(x) \succ 0$, for all $x \in C$ implies f strictly convex.

Quadratic functions

Example: Quadratic function of the form

$$\begin{array}{rccl} q: & \mathbb{R}^N & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & q(x) = \frac{1}{2} x^T A x - b^T x \\ & & & \quad = \frac{1}{2} \sum_{i,j=1}^{N} x_i a_{ij} x_j - \sum_{i=1}^{N} b_i x_i, \end{array}$$

with $A \in \mathbb{R}^{N \times N}$ symmetric, $b \in \mathbb{R}^N$.

Remark:

- Gradient: $\nabla q(x) = Ax - b$.
- Hessian: $H(x) = A$.
- $q(x)$ is positive definite if A is positive definite.
- A quadratic function that is positive definite is a strictly convex function.

Functional analysis: minimizers

$$\text{Find} \quad \widehat{x} \in \operatorname*{Argmin}_{x \in C} \quad f(x)$$

- **Minimizers**
  - ⊙ Local versus global minimizers
  - ⊙ Coercivity and existence
  - ⊙ Convex function

Critical or stationary points

Let C be a nonempty set of a Hilbert space $\mathcal{H}$.
Let $f : C \to]-\infty, +\infty]$ be a proper function and let $\widehat{x} \in C$.

- $\widehat{x} \in \operatorname{dom} f$ is a **local minimizer** of f if there exists an open neigborhood O of $\widehat{x}$ such that

$$(\forall x \in O \cap C) \qquad f(\widehat{x}) \leq f(x).$$

- $\widehat{x}$ is a **(global) minimizer** of f if

$$(\forall x \in C) \qquad f(\widehat{x}) \leq f(x).$$

Let C be a nonempty set of a Hilbert space $\mathcal{H}$.
Let $f : C \to]-\infty, +\infty]$ be a proper function and let $\widehat{x} \in C$.

- $\widehat{x} \in \operatorname{dom} f$ is a **stationary point** of f if

$$\nabla f(\widehat{x}) = 0.$$

Critical or stationary points

Let C be a nonempty set of a Hilbert space $\mathcal{H}$.
Let $f : C \to]-\infty, +\infty]$ be a proper function and let $\widehat{x} \in C$.

- $\widehat{x}$ is a **strict local minimizer** of f if there exists an open neigborhood O of $\widehat{x}$ such that

$$(\forall x \in (O \cap C) \setminus \{\widehat{x}\}) \qquad f(\widehat{x}) < f(x).$$

- $\widehat{x}$ is a **strict (global) minimizer** of f if

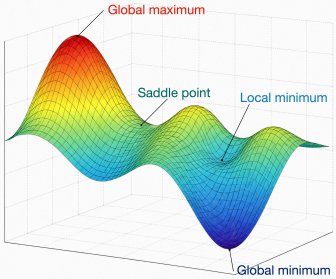
$$(\forall x \in C \setminus \{\widehat{x}\}) \qquad f(\widehat{x}) < f(x).$$

Let C be a nonempty set of a Hilbert space $\mathcal{H}$.
Let $f : C \to]-\infty, +\infty]$ be a proper function and let $\widehat{x} \in C$.

- $\widehat{x} \in \operatorname{dom} f$ is a **stationary point** of f if

$$\nabla f(\widehat{x}) = 0.$$

Critical or stationary points



How do we recognise a minimum?

- Necessary conditions for a point to be a minimum point: first order conditions (they use only first order information, coming from first order derivatives, i.e. the gradient)
- Sufficient conditions for a point to be a minimum point: second order ones (use first and second order information, coming from gradient and second order derivatives, i.e. the Hessian matrix).

Necessary conditions

Necessary conditions are conditions that need to be satisfied in order to have a minimum.

First order necessary condition

Let $f \in C^1(\Omega)$ in a neighbourhood Ω of $x^\star$.

If $x^\star$ is a minimizer for f (in Ω), then $\nabla f(x^\star) = 0$, i.e. $x^\star$ is a stationary point for f.

Remark:

• Necessary condition: All minimizers are stationary points but not all stationary points are minimizers, they may be maximizers or saddle points.

Proof (I)

Since $x^\star$ is a local minimizer, for every $d \in \mathbb{R}^N$ and for all sufficiently small $t \in \mathbb{R}$,

$$f(x^\star + td) - f(x^\star) \geq 0.$$

Since f is differentiable,

$$f(x^\star + td) - f(x^\star) = t\langle \nabla f(x^\star), d\rangle + o(t) \geq 0.$$

For $t > 0$, dividing by t and letting $t \to 0^+$ gives

$$\langle \nabla f(x^\star), d\rangle \geq 0.$$

For $t < 0$, dividing by t and letting $t \to 0^-$ gives

$$\langle \nabla f(x^\star), d\rangle \leq 0.$$

Hence

$$\langle \nabla f(x^\star), d\rangle = 0, \qquad \forall d \in \mathbb{R}^N.$$

Therefore $\boxed{\nabla f(x^\star) = 0}$.

Second order necessary condition

Let $f \in \mathcal{C}^2(\Omega)$ for a neighbourhood Ω of $x^\star$.

$x^\star$ is a minimum point for f (in Ω) $\implies$ $H_f(x^\star)$ is positive semidefinite.

Proof (I)

Let $d \in \mathbb{R}^N$. Since $x^\star$ is a local minimizer, for all sufficiently small $t \in \mathbb{R}$,

$$f(x^\star + td) - f(x^\star) \geq 0.$$

By the second-order Taylor expansion,

$$f(x^\star + td) = f(x^\star) + t\langle \nabla f(x^\star), d \rangle + \frac{t^2}{2} d^\top \nabla^2 f(x^\star) d + o(t^2).$$

By the first-order necessary condition,

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

Therefore,

$$\frac{t^2}{2} d^\top \nabla^2 f(x^\star) d + o(t^2) \geq 0.$$

Proof (II)

Dividing by $t^2 > 0$ and letting $t \to 0$ gives

$$d^\top \nabla^2 f(x^\star) d \geq 0.$$

Since $d \in \mathbb{R}^N$ is arbitrary,

$$\boxed{\nabla^2 f(x^\star) \succeq 0}.$$

Sufficient conditions

Let $f \in C^2(\Omega)$ for a neighbourhood Ω of $x^\star$.

$$\begin{cases} \nabla f(x^\star) = 0 \\ H_f(x^\star) \text{ is positive definite} \end{cases} \implies x^\star \text{ is a minimum point for } f.$$

Remark

It is in general expensive to establish if $H_f(x^\star)$ is positive definite, as this requires the computation of the eigenvalues of the matrix. This condition is then usually not employed in practice.

Proof.
Since $f \in \mathcal{C}^2$, Taylor's theorem gives, for x sufficiently close to $x^\star$,

$$f(x) = f(x^\star) + \nabla f(x^\star)^T (x - x^\star) + \frac{1}{2}(x - x^\star)^T \nabla^2 f(\xi)(x - x^\star),$$

for some ξ on the segment joining $x^\star$ and x.

Since $\nabla f(x^\star) = 0$, and since $\nabla^2 f(x^\star) \succ 0$, by continuity of the Hessian there exist a neighbourhood B of $x^\star$ and $m > 0$ such that

$$\nabla^2 f(x) \succeq mI, \qquad \forall x \in B. \tag{3}$$

Therefore,

$$f(x) - f(x^\star) \geq \frac{m}{2}\|x - x^\star\|^2 > 0$$

for every $x \neq x^\star$ sufficiently close to $x^\star$. Hence, $x^\star$ is a strict local minimum. □

Proof of (3)

Since $\nabla^2 f(x^\star) \succ 0$, its smallest eigenvalue satisfies $\lambda_{\min}\big(\nabla^2 f(x^\star)\big) > 0$. Let

$$m = \frac{1}{2}\lambda_{\min}\big(\nabla^2 f(x^\star)\big) > 0.$$

Since $\nabla^2 f$ is continuous at $x^\star$, there exists a neighbourhood B of $x^\star$ such that, for every $x \in B$,

$$\big\|\nabla^2 f(x) - \nabla^2 f(x^\star)\big\| < \frac{1}{2}\lambda_{\min}\big(\nabla^2 f(x^\star)\big).$$

Then, for every $d \in \mathbb{R}^N$,

$$\begin{aligned} d^T \nabla^2 f(x) d &= d^T \nabla^2 f(x^\star) d + d^T \big(\nabla^2 f(x) - \nabla^2 f(x^\star)\big) d \\ &\geq \lambda_{\min}\big(\nabla^2 f(x^\star)\big)\|d\|^2 - \big\|\nabla^2 f(x) - \nabla^2 f(x^\star)\big\|\,\|d\|^2 \\ &> \frac{1}{2}\lambda_{\min}\big(\nabla^2 f(x^\star)\big)\|d\|^2 = m\|d\|^2. \end{aligned}$$

Therefore,

$$\nabla^2 f(x) \succeq mI, \qquad \forall x \in B.$$

Minimizers of a convex function

Theorem: Let $\mathcal{H}$ be a Hilbert space. Let $f\colon \mathcal{H} \to]-\infty, +\infty]$ be a **proper convex** function such that $\mu = \inf f > -\infty$.

- $\left\{x \in \mathcal{H} \;\middle|\; f(x) = \mu\right\}$ is convex.
- Every local minimizer of f is a global minimizer.
- If f is strictly convex, then there exists at most one minimizer.

Existence of a minimizer

Let $\mathcal{H}$ be a Hilbert space. Let $f : \mathcal{H} \to]-\infty, +\infty]$.
f is **coercive** if $\lim_{\|x\| \to +\infty} f(x) = +\infty$.

Existence and uniqueness of a minimizer

Theorem: Let $\mathcal{H}$ be a Hilbert space and C a **closed convex** subset of $\mathcal{H}$. Let $f \in \Gamma_0(\mathcal{H})$ such that $\operatorname{dom} f \cap C \neq \varnothing$.

If f **is coercive** or C **is bounded**, then there exists $\widehat{x} \in C$ such that

$$f(\widehat{x}) = \inf_{x \in C} f(x).$$

If, moreover, f is strictly convex, this minimizer $\widehat{x}$ is unique.

Optimality condition

1st order necessary and sufficient condition (P. Fermat)

Let f in $\Gamma_0(\mathbb{R}^N)$ and $\mathcal{C}^1(\mathbb{R}^N)$.

$\widehat{x}$ is a global minimizer of f i.e.,

$$\widehat{x} \in \operatorname{Argmin}_{x \in \mathbb{R}^N} f(x) \qquad \Leftrightarrow \qquad \nabla f(\widehat{x}) = 0$$

- Limitations :
  - ⊙ Lead to a N equations - N unknown problem.
  - ⊙ Closed form expression for only few cases.
  - ⊙ If no closed form expression exists, an iterative procedure is required.

Example: quadratic functions

Theorem
A positive definite quadratic function $q(x)$ has a unique minimizer $\widehat{x}$, that is the unique solution of the problem

$$Ax = b.$$

Proof.
As $q(x)$ is strictly convex, $q(x)$ has at most one minimizer. The Hessian matrix of $q(x)$ is A and it is positive definite, so from the conditions it holds

$$\widehat{x} \text{ is a minimizer of } q(x) \quad \Longleftrightarrow \quad \widehat{x} \text{ is a solution of } \nabla q(x) = 0,$$

i.e.

$$\widehat{x} \text{ is a minimizer of } q(x) \quad \Longleftrightarrow \quad \widehat{x} \text{ is a solution of } Ax = b.$$

A is positive definite and therefore invertible, so the problem $Ax = b$ has a unique solution. □

Optimality condition for convex functions

Solving least squares

Find $\widehat{x} = \mathrm{Argmin}_{x \in \mathbb{R}^N} \|Ax - y\|_2^2$ with $\begin{cases} A \in \mathbb{R}^{N \times N} & \text{full rank} \\ y \in \mathbb{R}^M \end{cases}$

- Optimality condition:

$$\nabla f(\widehat{x}) = 0 \quad \Leftrightarrow \quad A^\top (A\widehat{x} - y) = 0$$

$$\boxed{\widehat{x} = (A^\top A)^{-1}(A^\top y)}$$

- **Closed form expression** but sometimes difficult to invert $A^\top A$.

Optimality condition

Logistic based criterion:

$$\text{Find} \quad \widehat{x} \in \operatorname{Argmin}_{x \in \mathbb{R}} \log\left(1 + \exp(-yx)\right) \qquad \text{with} \qquad y \in \mathbb{R}$$

- Optimality condition:

$$\nabla f(\widehat{x}) = 0 \quad \Leftrightarrow \quad \boxed{\frac{-y \exp(-y\widehat{x})}{1 + \exp(-y\widehat{x})} = 0}$$

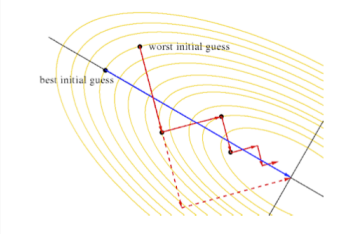
- **No closed form expression**. An iterative procedure is required.

Iterative methods

Iterative scheme

Let $f\colon \mathbb{R}^N \to \mathbb{R}$ be continuously differentiable on $\mathbb{R}^N$. Let $x^{[0]} \in \mathbb{R}^N$ and

$$(\forall k \in \mathbb{N}) \quad x^{[k+1]} = x^{[k]} + \gamma_k d^{[k]}, \quad \gamma_k > 0.$$

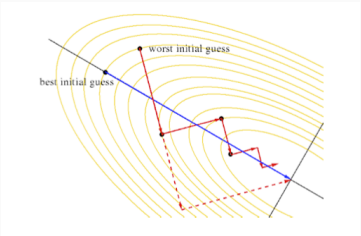


Iterative methods

Iterative scheme

Let $f\colon \mathbb{R}^N \to \mathbb{R}$ be continuously differentiable on $\mathbb{R}^N$. Let $x^{[0]} \in \mathbb{R}^N$ and

$$(\forall k \in \mathbb{N}) \quad x^{[k+1]} = x^{[k]} + \gamma_k d^{[k]}, \quad \gamma_k > 0.$$



- An iterative method consists to build a sequence $(x^{[k]})_{n\in\mathbb{N}}$ such that, at each iteration n

$$f(x^{[k+1]}) = f(x^{[k]} + \gamma_k d^{[k]}) < f(x^{[k]})$$

- How to choose $d^{[k]}$?
- How to prove convergence of the sequence $(x^{[k]})_{n\in\mathbb{N}}$ to $\widehat{x} \in \operatorname{Argmin} f(x)$?
- Choose γ_k for convergence ? For faster convergence ?

Descent directions

Let $f\colon \mathbb{R}^N \to \,]-\infty, +\infty]$ be a proper differentiable function in the neighborhood of $x \in \mathbb{R}^N$.

The **directional derivative** of f at x with respect to the direction $d \in \mathbb{R}^N$ is defined as:

$$\langle \nabla f(x) \mid d \rangle = \lim_{\gamma \to 0} \frac{f(x + \gamma d) - f(x)}{\gamma}.$$

A direction $d \in \mathbb{R}^N$ is a descent direction for f in x if

$$\langle \nabla f(x) \mid d \rangle < 0,$$

i.e., if the angle ϑ between d and $\nabla f(x)$ is such that $\vartheta \in \left(\frac{\pi}{2}, \pi\right]$.

If p is a descent direction, then it exists $\bar{\gamma} > 0$ such that

$$f(x + \gamma d) - f(x) < 0 \qquad \forall \gamma \in (0, \bar{\gamma}).$$

Examples of descent directions: steepest descent

If $\nabla f(x) \neq 0$, we can always find a descent direction: that of the antigradient:

$$d^{[k]} = -\nabla f(x^{[k]}).$$

The direction of steepest descent minimizes the directional derivative of f in $x^{[k]}$:

$$\nabla f(x^{[k]})^T d \underbrace{=}_{\substack{\vartheta \in [0,\pi] \text{ is the angle} \\ \text{between } \nabla f(x^{[k]}) \text{ and } d}} \|\nabla f(x^{[k]})\| \|d\| \cos\vartheta,$$

If we assume, without loss of generality, that $\|d\| = 1$, the direction that maximises the decrease is the one that minimizes $\cos\vartheta$, so it must be such that $\vartheta = \pi$ and so

$$d^{[k]} = \frac{-\nabla f(x^{[k]})}{\|\nabla f(x^{[k]})\|}.$$

What do we expect from an optimization method?

- Convergence to a stationary point:

$$\lim_{k\to\infty} \|\nabla f(x^{[k]})\| = 0. \tag{4}$$

- *Global convergence*: (4) is guaranteed for any initial guess $x^{[0]}$, independently of the proximity of $x^{[0]}$ to $x^\star$.
- Crucial for non-convex problems.
- Depends on the choice of the step-length: globalization strategies.

Remark: **Very important!**

Global convergence does not mean convergence to a global minimum!!

Convergence of the sequence

- Stronger: convergence of the generated sequence:

$$\lim_{k \to \infty} x^{[k]} = x^\star$$

- Usually difficult to prove and it is a result established only for initial guesses close enough to the true solution (*local convergence*).
- Crucial to ensure convergence to local minima and not saddle points

Rates of convergence

Let $\{x^{[k]}\}_{k\in\mathbb{N}}^{N} \to x^\star$.

- The sequence is said to converge *linearly* if it exists $r \in (0,1)$ such that

$$\lim_{k\to+\infty} \frac{\|x^{[k+1]} - x^\star\|}{\|x^{[k]} - x^\star\|} = r.$$

- The sequence is said to converge *superlinearly* (faster than linearly) if

$$\lim_{k\to+\infty} \frac{\|x^{[k+1]} - x^\star\|}{\|x^{[k]} - x^\star\|} = 0.$$

- The sequence is said to converge *sublinearly* (slower than linearly) if

$$\lim_{k\to+\infty} \frac{\|x^{[k+1]} - x^\star\|}{\|x^{[k]} - x^\star\|} = 1.$$

- The convergence is said to be *quadratic* if it exists $M > 0$ such that

$$\lim_{k\to+\infty} \frac{\|x^{[k+1]} - x^\star\|}{\|x^{[k]} - x^\star\|^2} < M.$$

Classification of methods

Based on the information they use, the optimization methods can be divided into two large classes:

1. **First order methods**:

   Only information coming from first order derivatives, i.e. the gradient.

   Usually cheap, but slow in reaching a stationary point (require many iterations).

2. **Second order methods**:

   Use the gradient information and also the second-order information contained in the Hessian matrix.

   Rather expensive and memory demanding, but can reach a stationary point much faster.

Rate vs cost

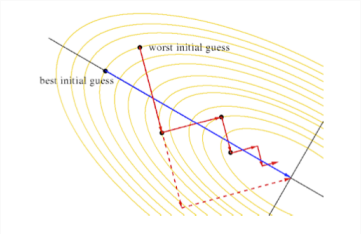
- Usually the cost of a method is directly proportional to the speed of convergence: generally an expensive method (for which a single iteration is expensive to compute) has a higher rate of convergence and requires less iterations to converge.
- Gradient methods enjoy linear convergence rate
- Second order methods based on Newton's directions enjoy a quadratic local rate of convergence.

Gradient descent for smooth functions

Gradient descent

Let $f\colon \mathbb{R}^N \to \mathbb{R}$ be continuously differentiable on $\mathbb{R}^N$. Let $x_0 \in \mathbb{R}^N$ and

$$(\forall n \in \mathbb{N}) \quad x^{[k+1]} = x^{[k]} - \gamma_k \nabla f(x^{[k]}).$$



Remark:

Can be interpreted as approximating the function locally with a first order Taylor model:

$$f(x^{[k]} + d) \approx f(x^{[k]}) + \nabla f(x^{[k]})^\top d := T_1(x^{[k]}, d)$$

and d is found minimizing the regularized model for $\lambda_k > 0$:

$$\min_d T_1(x^{[k]}, d) + \frac{1}{2}\lambda_k \|d\|^2 \quad \Leftrightarrow \quad d = -\frac{\nabla f(x^{[k]})}{\lambda_k}$$

Gradient method - algorithm

Gradient descent algorithm

Given $x^{[0]}$, $\varepsilon > 0$, set $k = 0$.

While $\|\nabla f(x^{[k]})\| > \varepsilon$

1. Compute the search direction $d^{[k]} = -\nabla f(x^{[k]})$.
2. Choose γ_k.
3. Set $x^{[k+1]} = x^{[k]} + \gamma_k d^{[k]}$
4. Set $k = k + 1$

- Cheap (it requires just the computation of the gradient of f at each iteration)
- but slow (it requires a large number of iterations to reach a stationary point)

Convergence of gradient method

- Convex + smooth: gradient descent iterates converge to a minimizer (assuming minimizers exist).
- Strongly convex + smooth: they converge linearly to the unique minimizer.
- Nonconvex + smooth: $\|\nabla f(x^{[k]})\| \to 0$ does not by itself imply that $x^{[k]}$ converges.

Convergence results

Assumptions on f	**Step size**	**Result on $f(x^{[k]})$**	**Result on the iterates $x^{[k]}$**
f differentiable, bounded below, ∇f L-Lipschitz	$0 < \gamma < 2/\beta$	$f(x^{[k]})$ decreases and $\|\nabla f(x^{[k]})\| \to 0$	**Not necessarily convergent** without additional assumptions
f convex and β-smooth, $\arg\min f \neq \varnothing$	$0 < \gamma < 2/\beta$	$f(x^{[k]}) \to f^\star$, with rate $O(1/k)$	$x^{[k]} \to x^\star \in \arg\min f$ (in finite dimensions)
f μ-strongly convex and β-smooth	$0 < \gamma < 2/\beta$	**Linear convergence**	$x^{[k]} \to x^\star$ **linearly**
f satisfies the PL inequality and is β-smooth	$0 < \gamma \leq 1/\beta$	**Linear convergence** to $f^\star$	Distance to the solution set converges linearly under additional conditions
f nonconvex, β-smooth, and satisfies the KL property	Suitable descent step	$f(x^{[k]}) \to f^\star$	$x^{[k]}$ converges to a critical point under standard boundedness assumptions

Table 1: Convergence properties of gradient descent under different assumptions.

- [Polyak–Łojasiewicz (PL) inequality] $f : \mathbb{R}^N \to \mathbb{R}$, $f \in \mathcal{C}^1$ satisfies the *Polyak–Łojasiewicz (PL) inequality* with constant $\mu > 0$ if

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu\big(f(x) - f^\star\big), \qquad \forall x \in \mathbb{R}^n,$$

where $f^\star = \inf_{x \in \mathbb{R}^n} f(x)$.

- [Kurdyka–Łojasiewicz (KL) property] Let $f : \mathbb{R}^N \to \mathbb{R}$ be differentiable and let $\bar{x}$ be a critical point of f. We say that f satisfies the *Kurdyka–Łojasiewicz (KL) property* at $\bar{x}$ if there exist a neighborhood U of $\bar{x}$, $\eta > 0$, and a concave function

$$\varphi : [0, \eta) \to \mathbb{R}_+,$$

such that $\varphi(0) = 0$, φ is continuous at 0, continuously differentiable on $(0, \eta)$, and $\varphi'(s) > 0$ for all $s \in (0, \eta)$, satisfying

$$\varphi'\big(f(x) - f(\bar{x})\big)\|\nabla f(x)\| \geq 1,$$

for all $x \in U$ such that

$$f(\bar{x}) < f(x) < f(\bar{x}) + \eta.$$

Convergence of gradient descent for smooth functions

Theorem

Let $f : \mathbb{R}^N \to \mathbb{R}$ be bounded from below and β-smooth, and consider the gradient descent iteration $x^{[k+1]} = x^{[k]} - \gamma \nabla f(x^{[k]})$, with $0 < \gamma < \frac{2}{\beta}$. Then $f(x^{[k]})$ is decreasing and converges to a finite value $f_\infty \geq f^\star$. Moreover,

$$\sum_{k=0}^{\infty} \|\nabla f(x^{[k]})\|^2 < +\infty,$$

and consequently

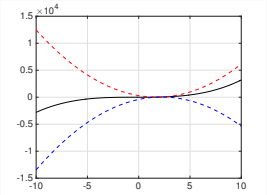
$$\boxed{\|\nabla f(x^{[k]})\| \longrightarrow 0.}$$

Convergence of gradient descent for smooth functions - key lemma

Lemma (Descent lemma - Nesterov 2004) Let $f \in C^{1,1}_{\beta}(\mathbb{R}^N)$. Then,

$$(\forall x, y \in \mathbb{R}^N) \qquad |f(y) - f(x) - \langle \nabla f(x) \mid y - x \rangle| \leq \frac{\beta}{2}\|y - x\|^2$$

- (black) $f(x) = 3x^3 + 2x^2$ with $x \in [-10, 10]$
- $x_0 = 2$, $\beta = 180$.
- (red) $\phi(x) = f(x_0) + \langle \nabla f(x_0) \mid x - x_0 \rangle| + \frac{\beta}{2}\|x - x_0\|^2$
- (blue) $\phi(x) = f(x_0) + \langle \nabla f(x_0) \mid x - x_0 \rangle| - \frac{\beta}{2}\|x - x_0\|^2$



Convergence of gradient descent for smooth functions - proof

- Function decrease for the gradient method i.e., $x^{[k+1]} = x^{[k]} - \gamma \nabla f(x^{[k]})$

$$\begin{aligned} f(x^{[k+1]}) &\leq f(x^{[k]}) + \left\langle \nabla f(x^{[k]}) \mid x^{[k+1]} - x^{[k]} \right\rangle + \frac{\beta}{2} \|x^{[k+1]} - x^{[k]}\|^2 \\ &= f(x^{[k]}) - \gamma \|\nabla f(x^{[k]})\|^2 + \frac{\gamma^2 \beta}{2} \|\nabla f(x^{[k]})\|^2 \\ &= f(x^{[k]}) - \gamma \left(1 - \frac{\gamma \beta}{2}\right) \|\nabla f(x^{[k]})\|^2 \end{aligned}$$

  $\rightarrow$ relaxation (i.e. $f(x^{[k+1]}) \leq f(x^{[k]})$) if $\gamma \in]0, \frac{2}{\beta}]$

  $\rightarrow$ best estimate for possible decrease $\gamma^* = \frac{1}{\beta}$.

- Summing up:

$$f(x^{[0]}) - \inf f \geq \sum_k f(x^{[k]}) - f(x^{[k+1]}) \geq \sum_k \gamma \left(1 - \frac{\gamma \beta}{2}\right) \|\nabla f(x^{[k]})\|^2$$

  $\rightarrow \|\nabla f(x^{[k]})\| \rightarrow 0$

Linear convergence of gradient descent - strongly convex case

Theorem

Let $f : \mathbb{R}^N \to \mathbb{R}$ be μ-strongly convex and β-smooth, with $0 < \mu \leq \beta$. Consider the gradient descent iteration

$$x^{[k+1]} = x^{[k]} - \gamma \nabla f(x^{[k]}),$$

with

$$0 < \gamma < \frac{2}{\beta}.$$

Then $x^{[k]}$ converges linearly to the unique minimizer $x^\star$:

$$\|x^{[k]} - x^\star\| \leq \rho(\gamma)^k \|x^{[0]} - x^\star\|,$$

where

$$\rho(\gamma) = \max\{|1 - \gamma\mu|, |1 - \gamma\beta|\} < 1.$$

Proof (I)

Since f is μ-strongly convex and β-smooth,

$$\mu I \preceq \nabla^2 f(x) \preceq \beta I.$$

Moreover, since $x^\star$ is the unique minimizer,

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

Using the mean-value representation,

$$\nabla f(x^{[k]}) - \nabla f(x^\star) = H_k\big(x^{[k]} - x^\star\big),$$

where

$$H_k = \int_0^1 \nabla^2 f\left(x^\star + t\big(x^{[k]} - x^\star\big)\right) dt.$$

Consequently,

$$\mu I \preceq H_k \preceq \beta I.$$

Proof (II)

Since $\nabla f(x^\star) = 0$, the gradient descent iteration gives

$$\begin{aligned} x^{[k+1]} - x^\star &= x^{[k]} - x^\star - \gamma \left(\nabla f(x^{[k]}) - \nabla f(x^\star) \right) \\ &= (I - \gamma H_k) \left(x^{[k]} - x^\star \right). \end{aligned}$$

The eigenvalues of H_k belong to $[\mu, \beta]$. Hence

$$\|I - \gamma H_k\| = \max_{\lambda \in [\mu, \beta]} |1 - \gamma\lambda| = \max \left\{ |1 - \gamma\mu|, |1 - \gamma\beta| \right\}.$$

Therefore,

$$\|x^{[k+1]} - x^\star\| \leq \rho(\gamma) \|x^{[k]} - x^\star\|,$$

where

$$\rho(\gamma) = \max \left\{ |1 - \gamma\mu|, |1 - \gamma\beta| \right\}.$$

Proof (III)

For $0 < \gamma < 2/\beta$, we have

$$|1 - \gamma\mu| < 1, \qquad |1 - \gamma\beta| < 1,$$

and therefore

$$\rho(\gamma) < 1.$$

Iterating the inequality yields

$$\boxed{\|x^{[k]} - x^\star\| \leq \rho(\gamma)^k \|x^{[0]} - x^\star\|}$$

and hence

$$x^{[k]} \longrightarrow x^\star \qquad \text{linearly.}$$

Optimal convergence rate

The step-size-dependent rate can also be written explicitly as

$$\rho(\gamma) = \begin{cases} 1 - \gamma\mu, & 0 < \gamma \leq \dfrac{2}{\beta + \mu}, \\ \gamma\beta - 1, & \dfrac{2}{\beta + \mu} \leq \gamma < \dfrac{2}{\beta} \end{cases}$$

with the optimal choice

$$\gamma^\star = \frac{2}{\beta + \mu}, \qquad \rho^\star = \frac{\beta - \mu}{\beta + \mu}.$$

Convergence of gradient descent for convex smooth functions

Theorem

Let $f : \mathbb{R}^N \to \mathbb{R}$ be convex and β-smooth, and assume that f admits a minimizer $x^\star$. Consider the gradient descent iteration $x^{[k+1]} = x^{[k]} - \gamma \nabla f(x^{[k]})$, with $0 < \gamma < \frac{2}{\beta}$. Then

$$x^{[k]} \longrightarrow x^\star \in \operatorname*{arg\,min}_{x \in \mathbb{R}^N} f(x).$$

Moreover,

$$f(x^{[k]}) - f^\star \leq \frac{\|x^{[0]} - x^\star\|^2}{2\gamma \left(1 - \frac{\gamma\beta}{2}\right) k}, \qquad k \geq 1,$$

and therefore

$$f(x^{[k]}) - f^\star = O\left(\frac{1}{k}\right).$$

Proof (I)

Since f is convex and β-smooth, its gradient is $1/\beta$-cocoercive:

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{\beta} \|\nabla f(x) - \nabla f(y)\|^2.$$

Since $x^\star$ is a minimizer,

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

Taking $y = x^\star$, we obtain

$$\langle \nabla f(x^{[k]}), x^{[k]} - x^\star \rangle \geq \frac{1}{\beta} \|\nabla f(x^{[k]})\|^2.$$

Proof (II)

Now, using the gradient descent iteration,

$$\begin{aligned}\|x^{[k+1]} - x^\star\|^2 &= \|x^{[k]} - x^\star - \gamma \nabla f(x^{[k]})\|^2 \\ &= \|x^{[k]} - x^\star\|^2 - 2\gamma \langle \nabla f(x^{[k]}), x^{[k]} - x^\star \rangle \\ &\quad + \gamma^2 \|\nabla f(x^{[k]})\|^2 \\ &\leq \|x^{[k]} - x^\star\|^2 - \gamma \left(\frac{2}{\beta} - \gamma\right) \|\nabla f(x^{[k]})\|^2.\end{aligned}$$

Since $0 < \gamma < 2/\beta$, $\gamma \left(\frac{2}{\beta} - \gamma\right) > 0$. Hence

$$\boxed{\|x^{[k+1]} - x^\star\| \leq \|x^{[k]} - x^\star\|.}$$

Thus the sequence $\left(x^{[k]}\right)_k$ is Fejér monotone with respect to the set of minimizers and is therefore bounded.

Proof (III)

Furthermore, summing the previous inequality from $k = 0$ to K gives

$$\gamma \left(\frac{2}{\beta} - \gamma \right) \sum_{k=0}^{K} \|\nabla f(x^{[k]})\|^2 \leq \|x^{[0]} - x^\star\|^2.$$

Therefore,

$$\sum_{k=0}^{\infty} \|\nabla f(x^{[k]})\|^2 < +\infty,$$

and consequently

$$\|\nabla f(x^{[k]})\| \longrightarrow 0.$$

Since $\left(x^{[k]}\right)_k$ is bounded, it has at least one accumulation point. Let

$$x^{[k_j]} \longrightarrow \bar{x}.$$

By continuity of the gradient,

$$\nabla f(\bar{x}) = 0.$$

Proof (IV)

Finally, Fejér monotonicity with respect to the nonempty closed convex set $\arg\min f$, together with the fact that every accumulation point belongs to this set, implies that the whole sequence converges to a minimizer:

$$\boxed{x^{[k]} \longrightarrow x^\star \in \arg\min f.}$$

Proof (V)

Since f is β-smooth, the descent lemma gives

$$f(x^{[k+1]}) \leq f(x^{[k]}) - \gamma \left(1 - \frac{\gamma\beta}{2}\right) \|\nabla f(x^{[k]})\|^2.$$

Define

$$q := 1 - \frac{\gamma\beta}{2} > 0.$$

Then

$$\gamma q \|\nabla f(x^{[k]})\|^2 \leq f(x^{[k]}) - f(x^{[k+1]}). \tag{5}$$

By convexity,

$$f(x^{[k]}) - f^\star \leq \left\langle \nabla f(x^{[k]}), x^{[k]} - x^\star \right\rangle.$$

Proof (VI)

Using the gradient descent iteration,

$$\left\langle \nabla f(x^{[k]}), x^{[k]} - x^\star \right\rangle = \frac{1}{2\gamma}\left(\|x^{[k]} - x^\star\|^2 - \|x^{[k+1]} - x^\star\|^2\right) + \frac{\gamma}{2}\|\nabla f(x^{[k]})\|^2.$$

Consequently,

$$f(x^{[k]}) - f^\star \leq \frac{1}{2\gamma}\left(\|x^{[k]} - x^\star\|^2 - \|x^{[k+1]} - x^\star\|^2\right) + \frac{\gamma}{2}\|\nabla f(x^{[k]})\|^2. \tag{6}$$

Proof (VII)

From (5),

$$\frac{\gamma}{2}\|\nabla f(x^{[k]})\|^2 \leq \frac{1}{2q}\left(f(x^{[k]}) - f(x^{[k+1]})\right).$$

Substituting into (6) gives

$$f(x^{[k]}) - f^\star \leq \frac{1}{2\gamma}\left(\|x^{[k]} - x^\star\|^2 - \|x^{[k+1]} - x^\star\|^2\right) + \frac{1}{2q}\left(f(x^{[k]}) - f(x^{[k+1]})\right). \tag{7}$$

Proof (VIII)

Summing (7) from $k = 0$ to $K - 1$ yields

$$\sum_{k=0}^{K-1} \left(f(x^{[k]}) - f^\star \right) \leq \frac{\|x^{[0]} - x^\star\|^2}{2\gamma} + \frac{f(x^{[0]}) - f^\star}{2q}.$$

Since $f(x^{[k]})$ is decreasing,

$$K \left(f(x^{[K]}) - f^\star \right) \leq \sum_{k=0}^{K-1} \left(f(x^{[k]}) - f^\star \right).$$

Moreover, β-smoothness and $\nabla f(x^\star) = 0$ imply

$$f(x^{[0]}) - f^\star \leq \frac{\beta}{2} \|x^{[0]} - x^\star\|^2.$$

Proof (IX)

Therefore,

$$f(x^{[K]}) - f^\star \leq \frac{\|x^{[0]} - x^\star\|^2}{K} \left(\frac{1}{2\gamma} + \frac{\beta}{4q} \right) = \frac{\|x^{[0]} - x^\star\|^2}{2\gamma q K}.$$

Since $q = 1 - \gamma\beta/2$, we obtain

$$\boxed{f(x^{[K]}) - f^\star \leq \frac{\|x^{[0]} - x^\star\|^2}{2\gamma \left(1 - \frac{\gamma\beta}{2}\right) K}.}$$

Hence

$$f(x^{[K]}) - f^\star = O\left(\frac{1}{K}\right).$$

Exercises

1. Prove that if f is convex and β-smooth, its gradient is $1/\beta$-cocoercive:

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{\beta} \|\nabla f(x) - \nabla f(y)\|^2.$$

2. Prove that if the sequence $\left(x^{[k]}\right)_k$ is Fejér monotone with respect to the set $S \neq \emptyset$ then it is bounded.
3. Consider $f(x) = \|Ax - b\|^2 + \|Cx - d\|^2$ for $A, C \in \mathbb{R}^{N \times N}$ and $b, d \in \mathbb{R}^N$. Determine the Lipschitz constant of the gradient. What can you deduce on the convergence of the gradient method applied to this function?

Correction 1

Define $h(x) = f(x) - \langle \nabla f(y), x \rangle$.

Since f is convex and β-smooth, h is also convex and β-smooth. Moreover,

$$\nabla h(y) = \nabla f(y) - \nabla f(y) = 0,$$

so y is a minimizer of h.

By the descent lemma, for any $x, z \in \mathbb{R}^N$,

$$h(z) \leq h(x) + \langle \nabla h(x), z - x \rangle + \frac{\beta}{2} \|z - x\|^2.$$

Correction 1

Choosing

$$z = x - \frac{1}{\beta}\nabla h(x),$$

we obtain

$$h(z) \leq h(x) - \frac{1}{2\beta}\|\nabla h(x)\|^2.$$

Since y minimizes h, $h(y) \leq h(z)$ and therefore

$$h(x) - h(y) \geq \frac{1}{2\beta}\|\nabla h(x)\|^2.$$

That is,

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq \frac{1}{2\beta} \|\nabla f(x) - \nabla f(y)\|^2.$$

Interchanging x and y gives

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \geq \frac{1}{2\beta} \|\nabla f(x) - \nabla f(y)\|^2.$$

Adding the two results, the function-value terms cancel, yielding

$$\boxed{\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{\beta} \|\nabla f(x) - \nabla f(y)\|^2.}$$

Hence, ∇f is $1/\beta$-cocoercive.

Proof.
Since $S \neq \emptyset$, let $\bar{x} \in S$. By Fejér monotonicity, we have

$$\|x^{[k+1]} - \bar{x}\| \leq \|x^{[k]} - \bar{x}\| \leq \cdots \leq \|x^{[0]} - \bar{x}\|, \qquad \forall k \in \mathbb{N}.$$

Using the triangle inequality, we obtain

$$\|x^{[k]}\| \leq \|x^{[k]} - \bar{x}\| + \|\bar{x}\| \leq \|x^{[0]} - \bar{x}\| + \|\bar{x}\|, \qquad \forall k \in \mathbb{N}.$$

Hence, $\left(x^{[k]}\right)_{k\in\mathbb{N}}$ is bounded. □

Correction 3

$$f(x) = \|Ax - b\|_2^2 + \|Cx - d\|_2^2.$$

We first compute the gradient:

$$\nabla f(x) = 2A^T(Ax - b) + 2C^T(Cx - d) = 2(A^TA + C^TC)x - 2(A^Tb + C^Td).$$

Hence, for all $x, y \in \mathbb{R}^N$,

$$\|\nabla f(x) - \nabla f(y)\| = 2\|(A^TA + C^TC)(x - y)\| \leq L\|x - y\|.$$

Therefore,

$$\boxed{L = \max \frac{2\|(A^TA + C^TC)(x - y)\|}{\|x - y\|} = 2\lambda_{\max}(A^TA + C^TC).}$$

Correction 3

Moreover, since $A^TA + C^TC \succeq 0$, the function f is convex. Therefore, the gradient method

$$x^{[k+1]} = x^{[k]} - \gamma \nabla f(x^{[k]})$$

converges to a minimizer of f for any

$$0 < \gamma < \frac{2}{L}.$$

If, in addition, $A^TA + C^TC \succ 0$, then f is strongly convex and admits a unique minimizer. In this case, the gradient method converges linearly to the unique minimizer.