Optimization for data processing at a large scale

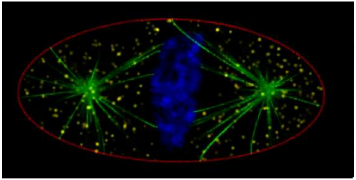
Sparsity4PSL Summer School

Emilie Chouzenoux
Center for Visual Computing
CentraleSupélec, INRIA Saclay

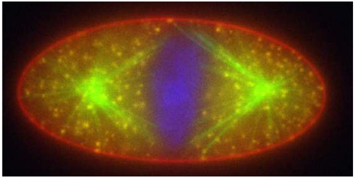
24 June 2019

Inverse problems and large scale optimization

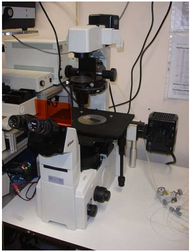
[Microscopy, ISBI Challenge 2013, F. Soulez]



Original image

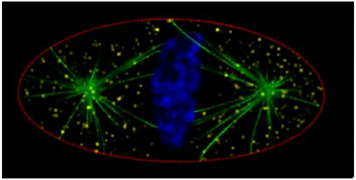


Degraded image



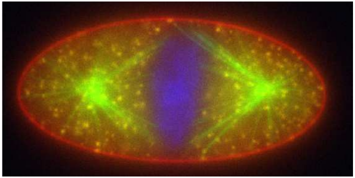
Inverse problems and large scale optimization

[Microscopy, ISBI Challenge 2013, F. Soulez]



Original image

$\overline{x} \in \mathbb{R}^N$



Degraded image

$z = \mathcal{D}(H\overline{x}) \in \mathbb{R}^M$

- $H \in \mathbb{R}^{M \times N}$: matrix associated with the degradation operator.
- $\mathcal{D} \colon \mathbb{R}^M \to \mathbb{R}^M$: noise degradation.

Inverse problem:

Find a good estimate of $\overline{x}$ from the observations z, using some a priori knowledge on $\overline{x}$ and on the noise characteristics.

Inverse problems and large scale optimization

Inverse problem:
Find an estimate $\widehat{x}$ close to $\overline{x}$ from the observations $z = \mathcal{D}(H\overline{x})$.

- Inverse filtering (if $M = N$ and H is invertible)

$$\begin{aligned}\widehat{x} &= H^{-1}z \\ &= H^{-1}(H\overline{x} + b) \quad \leftarrow \quad \text{if } b \in \mathbb{R}^M \text{ is an additive noise} \\ &= \overline{x} + H^{-1}b\end{aligned}$$

$\rightarrow$ Closed form expression, but amplification of the noise if H is ill-conditioned (*ill-posed problem*).

Inverse problems and large scale optimization

Inverse problem:
Find an estimate $\widehat{x}$ close to $\overline{x}$ from the observations $z = \mathcal{D}(H\overline{x})$.

- ~~Inverse filtering~~
- Variational approach

$$\widehat{x} \in \underset{x \in \mathbb{R}^N}{\text{Argmin}} \; \underbrace{f_1(x)}_{\text{Data fidelity term}} + \underbrace{f_2(x)}_{\text{Regularization term}}$$

Inverse problems and large scale optimization

Inverse problem:
Find an estimate $\widehat{x}$ close to $\overline{x}$ from the observations $z = \mathcal{D}(H\overline{x})$.

- ~~Inverse filtering~~
- Variational approach

$$\widehat{x} \in \operatorname*{Argmin}_{x\in\mathbb{R}^N} \underbrace{f_1(x)}_{\text{Data fidelity term}} + \underbrace{f_2(x)}_{\text{Regularization term}}$$

Examples of data fidelity term

- Gaussian noise

$$(\forall x \in \mathbb{R}^N) \qquad f_1(x) = \frac{1}{\sigma^2}\|Hx - z\|^2$$

- Poisson noise

$$(\forall x \in \mathbb{R}^N) \qquad f_1(x) = \sum_{m=1}^{M} \left([Hx]^{(m)} - z^{(m)} \log([Hx]^{(m)})\right)$$

Examples of regularization terms (1)

- Admissibility constraints

$$\text{Find} \qquad x \in C = \bigcap_{m=1}^{M} C_m$$

where $(\forall m \in \{1, \ldots, M\})\ C_m \subset \mathbb{R}^N$.

Examples of regularization terms (1)

- Admissibility constraints

$$\text{Find} \qquad x \in C = \bigcap_{m=1}^{M} C_m$$

where $(\forall m \in \{1, \ldots, M\})$ $C_m \subset \mathbb{R}^N$.

- Variational formulation

$$(\forall x \in \mathbb{R}^N) \qquad f_2(x) = \sum_{m=1}^{M} \iota_{C_m}(x)$$

where, for all $m \in \{1, \ldots, M\}$, ι_{C_m} is the indicator function of C_m:

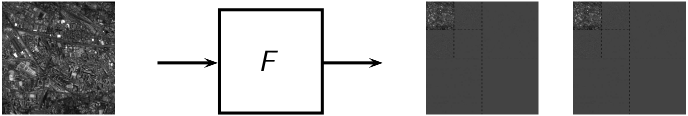
$$(\forall x \in \mathbb{R}^N) \qquad \iota_{C_m}(x) = \begin{cases} 0 & \text{if } x \in C_m \\ +\infty & \text{otherwise.} \end{cases}$$

Examples of regularization terms (2)

- ℓ_1 norm (analysis approach)

$$(\forall x \in \mathbb{R}^N) \qquad f_2(x) = \sum_{k=1}^{K} \left| [Fx]^{(k)} \right| = \|Fx\|_1$$

$F \in \mathbb{R}^{K \times N}$: Frame decomposition operator ($K \geq N$)



signal x frame coefficients

Examples of regularization terms (2)

- ℓ_1 norm (analysis approach)

$$(\forall x \in \mathbb{R}^N) \qquad f_2(x) = \sum_{k=1}^{K} \left| [Fx]^{(k)} \right| = \|Fx\|_1$$

- Total variation

$$(\forall x = (x^{(i_1,i_2)})_{1 \le i_1 \le N_1, 1 \le i_2 \le N_2} \in \mathbb{R}^{N_1 \times N_2})$$

$$f_2(x) = \mathrm{tv}(x) = \sum_{i_1=1}^{N_1} \sum_{i_2=1}^{N_2} \|\nabla x^{(i_1,i_2)}\|_2$$

$\nabla x^{(i_1,i_2)}$: discrete gradient at pixel (i_1, i_2).

Inverse problems and large scale optimization

Inverse problem:
Find an estimate $\widehat{x}$ close to $\overline{x}$ from the observations $z = \mathcal{D}(H\overline{x})$.

- ~~Inverse filtering~~
- Variational approach (more general context)

$$\widehat{x} \in \underset{x \in \mathbb{R}^N}{\text{Argmin}} \sum_{i=1}^{m} f_i(x)$$

where f_i may denote a data fidelity term / a (hybrid) regularization term / constraint.

Inverse problems and large scale optimization

Inverse problem:

Find an estimate $\widehat{x}$ close to $\overline{x}$ from the observations $z = \mathcal{D}(H\overline{x})$.

- ~~Inverse filtering~~
- Variational approach (more general context)

$$\widehat{x} \in \underset{x \in \mathbb{R}^N}{\operatorname{Argmin}} \sum_{i=1}^{m} f_i(x)$$

where f_i may denote a data fidelity term / a (hybrid) regularization term / constraint.

$\rightarrow$ Often no closed form expression or solution expensive to compute (especially in large scale context).

- **Need for an efficient iterative minimization strategy !**

Main challenges

- How to exploit the mathematical properties of each term involved in f? How to handle constraints efficiently ? How to deal with non differentiable terms in f ? Which convergence result can be expected if f is non convex?
- How to reduce the memory requirements of an optimization algorithm? How to avoid large-size matrix inversion?
- What are the benefits of block alternating strategies? What are their convergence guaranties?
- How to accelerate the convergence speed of a first-order (gradient-like) optimization method?

Outline

1. Introduction to optimization
   - Notation/definitions
   - Existence and unicity of minimizers
   - Differential/subdifferential
   - Optimality conditions

2. Majoration-Minimization approaches
   - Majorization-Minimization principle
   - Majorization techniques
   - MM quadratic methods
   - Forward-backward algorithm
   - Block-coordinate MM algorithms

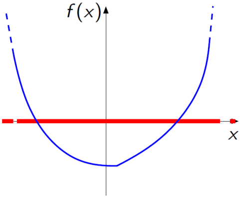
Introduction to optimization

Domain of a function

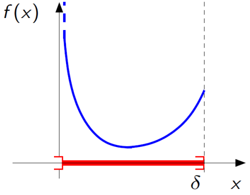
Let $f : \mathbb{R}^N \to \mathbb{R} \cup +\infty$.

- The domain of f is $\operatorname{dom} f = \{x \in \mathbb{R}^N \mid f(x) < +\infty\}$.
- The function f is proper if $\operatorname{dom} f \neq \varnothing$.

Domains of the functions ?



$\operatorname{dom} f = \mathbb{R}$
(proper)



$\operatorname{dom} f =]0, \delta]$
(proper)

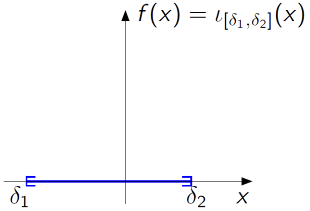
Indicator function

Let $C \subset \mathbb{R}^N$.
The indicator function of C is

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

Example:

$$C = [\delta_1, \delta_2]$$

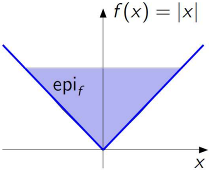


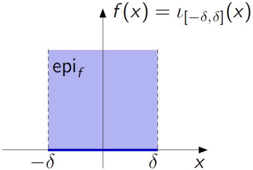
Epigraph

Let $f : \mathbb{R}^N \to \mathbb{R} \cup +\infty$. The epigraph of f is

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

Examples:



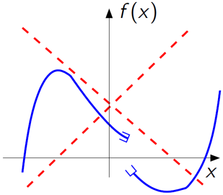


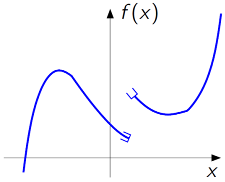
Lower semi-continuous function

Let $f : \mathbb{R}^N \to \mathbb{R} \cup +\infty$.
f is a lower semi-continuous function on $\mathbb{R}^N$ if and only if epi f is closed

Examples:

- l.s.c. functions ?



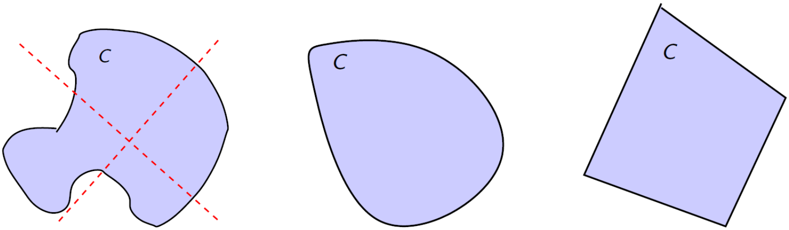


Convex set

$C \subset \mathbb{R}^N$ is a **convex set** if

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

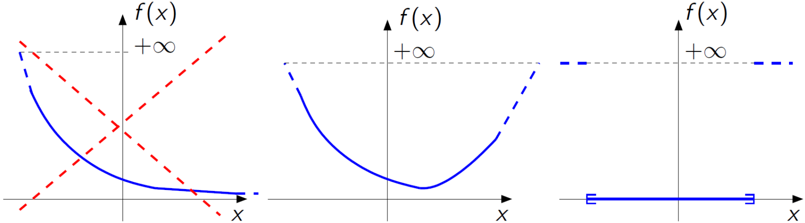
Convex sets ?



Coercive function

Let $f : \mathbb{R}^N \to \mathbb{R} \cup +\infty$.
f is coercive if $\lim_{\|x\| \to +\infty} f(x) = +\infty$.

Coercive functions ?



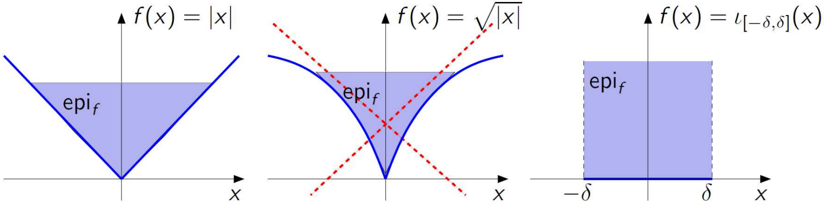
Convex function

$f : \mathbb{R}^N \to \mathbb{R} \cup +\infty$ is a **convex function** if

$$\big(\forall (x,y) \in (\mathbb{R}^N)^2\big)(\forall \alpha \in]0,1[)$$
$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$$

- f is convex $\Leftrightarrow$ its epigraph is convex.

Examples:

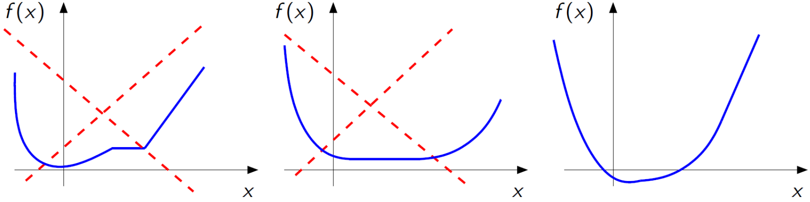


Strictly convex function

$f : \mathbb{R}^N \to \mathbb{R} \cup +\infty$ is **strictly convex** if

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

Strictly convex functions ?



Existence/unicity of minimizers

Theorem

Let $f : \mathbb{R}^N \to \mathbb{R} \cup +\infty$ be a proper l.s.c. coercive function.
Then, the set of minimizers of f is a nonempty compact set.

Convex case

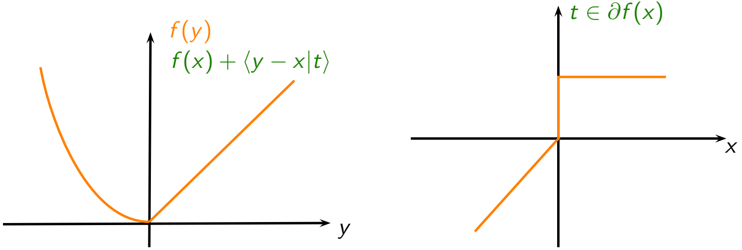
• Let $f : \mathbb{R}^N \to \mathbb{R} \cup +\infty$ be a proper convex function such that $\mu = \inf f > -\infty$. Then, every local minimizer of f is a global minimizer. Moreover, if f is strictly convex, then there exists at most one minimizer.

• Let C a closed convex subset of $\mathbb{R}^N$. Let $f : \mathbb{R}^N \to \mathbb{R} \cup +\infty$ proper, convex, lsc such that $\text{dom} f \cap C \neq \emptyset$. 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.

Subdifferential

Let $f : \mathbb{R}^N \to \mathbb{R} \cup +\infty$ be a proper function. The (Moreau) **subdifferential of f**, denoted by ∂f is such that

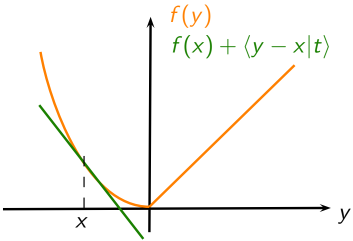
$$\partial f : \mathbb{R}^N \to 2^{\mathbb{R}^N}$$
$$x \to \{u \in \mathbb{R}^N \mid (\forall y \in \mathbb{R}^N)\ \langle y - x | u \rangle + f(x) \leq f(y)\}$$

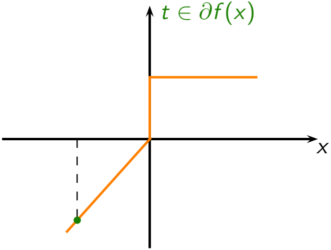


Subdifferential

Let $f : \mathbb{R}^N \to \mathbb{R} \cup +\infty$ be a proper function. The (Moreau) subdifferential of f, denoted by ∂f is such that

$$\begin{aligned}\partial f : \mathbb{R}^N &\to 2^{\mathbb{R}^N} \\ x &\to \{u \in \mathbb{R}^N \mid (\forall y \in \mathbb{R}^N)\ \langle y - x | u \rangle + f(x) \leq f(y)\}\end{aligned}$$

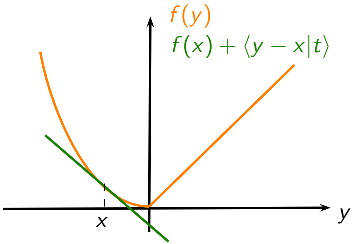


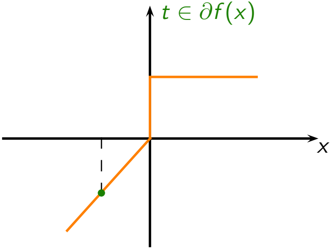


Subdifferential

Let $f : \mathbb{R}^N \to \mathbb{R} \cup +\infty$ be a proper function. The (Moreau) **subdifferential of f**, denoted by ∂f is such that

$$\partial f : \mathbb{R}^N \to 2^{\mathbb{R}^N}$$
$$x \to \{u \in \mathbb{R}^N \,|\, (\forall y \in \mathbb{R}^N)\ \langle y - x | u \rangle + f(x) \leq f(y)\}$$

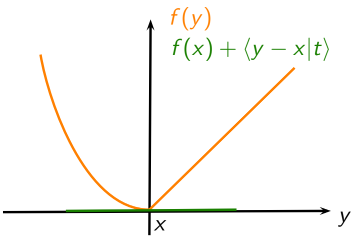


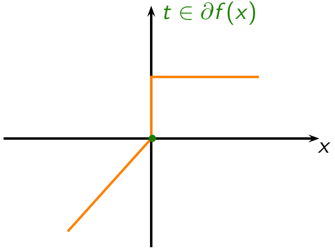


Subdifferential

Let $f : \mathbb{R}^N \to \mathbb{R} \cup +\infty$ be a proper function. The (Moreau) subdifferential of f, denoted by ∂f is such that

$$\begin{aligned}\partial f : \mathbb{R}^N &\to 2^{\mathbb{R}^N} \\ x &\to \{u \in \mathbb{R}^N \mid (\forall y \in \mathbb{R}^N)\ \langle y - x | u\rangle + f(x) \leq f(y)\}\end{aligned}$$

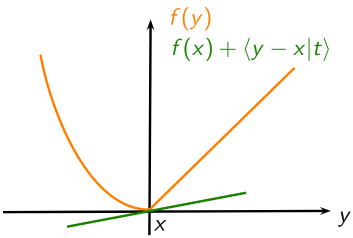


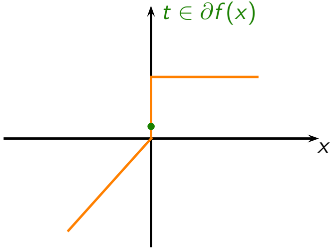


Subdifferential

Let $f : \mathbb{R}^N \to \mathbb{R} \cup +\infty$ be a proper function. The (Moreau) subdifferential of f, denoted by ∂f is such that

$$\partial f : \mathbb{R}^N \to 2^{\mathbb{R}^N}$$
$$x \to \{u \in \mathbb{R}^N \mid (\forall y \in \mathbb{R}^N)\ \langle y - x | u \rangle + f(x) \leq f(y)\}$$

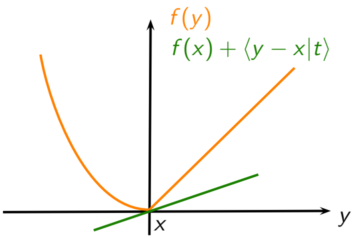


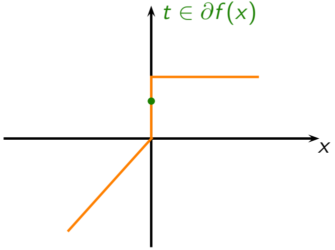


Subdifferential

Let $f : \mathbb{R}^N \to \mathbb{R} \cup +\infty$ be a proper function. The (Moreau) **subdifferential of f**, denoted by ∂f is such that

$$\begin{aligned}\partial f : \mathbb{R}^N &\to 2^{\mathbb{R}^N}\\ x &\to \{u \in \mathbb{R}^N \mid (\forall y \in \mathbb{R}^N)\ \langle y - x | u \rangle + f(x) \leq f(y)\}\end{aligned}$$

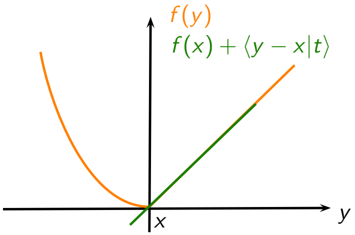


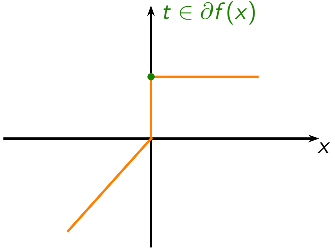


Subdifferential

Let $f : \mathbb{R}^N \to \mathbb{R} \cup +\infty$ be a proper function. The (Moreau) subdifferential of f, denoted by ∂f is such that

$$\begin{aligned}\partial f : \mathbb{R}^N &\to 2^{\mathbb{R}^N} \\ x &\to \{u \in \mathbb{R}^N \mid (\forall y \in \mathbb{R}^N)\ \langle y - x | u \rangle + f(x) \leq f(y)\}\end{aligned}$$

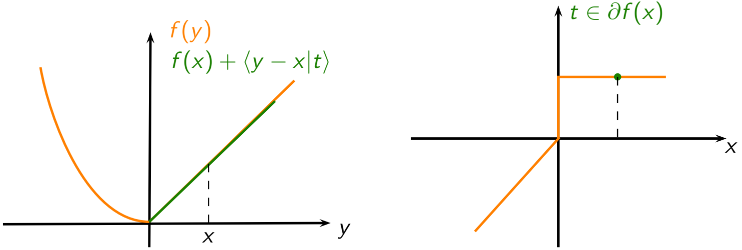




Subdifferential

Let $f : \mathbb{R}^N \to \mathbb{R} \cup +\infty$ be a proper function. The (Moreau) subdifferential of f, denoted by ∂f is such that

$$\partial f : \mathbb{R}^N \to 2^{\mathbb{R}^N}$$
$$x \to \{u \in \mathbb{R}^N \mid (\forall y \in \mathbb{R}^N)\ \langle y - x | u \rangle + f(x) \le f(y)\}$$



Optimality conditions

Fermat's rule: $0 \in \partial f(x) \Leftrightarrow x \in \operatorname{Argmin} f$

Differentiable case

Let C be a nonempty convex subset of $\mathbb{R}^N$. Let $f\colon \mathbb{R}^N \to \mathbb{R} \cup +\infty$ be Gâteaux differentiable at $\widehat{x} \in C$. If $\widehat{x}$ is a local minimizer of f over C, then

$$(\forall y \in C) \quad \nabla f(\widehat{x})^\top (y - \widehat{x}) \geq 0.$$

If $\widehat{x} \in \operatorname{int}(C)$, then the condition reduces to

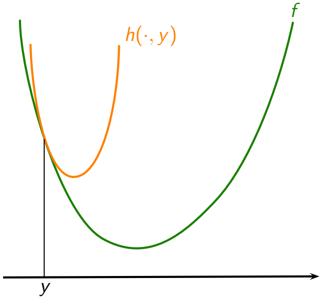
$$\nabla f(\widehat{x}) = 0.$$

Majoration-Minimization approaches

Majorant function

Let $f : \mathbb{R}^N \to \mathbb{R}$. Let $y \in \mathbb{R}^N$.
$h(\cdot, y) : \mathbb{R}^N \to \mathbb{R}$ is **a majorant function of f at y** if:

$$\begin{cases} (\forall x \in \mathbb{R}^N) \qquad f(x) \leq h(x, y), \\ f(y) = h(y, y). \end{cases}$$



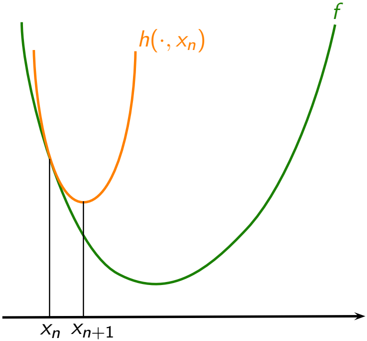
Majorization-Minimization algorithm

Problem: Minimization of function $f : \mathbb{R}^N \to \mathbb{R}$.

MM Algorithm

$$x_{n+1} \in \underset{x \in \mathbb{R}^N}{\text{Argmin}}\ h(x, x_n)$$

where $h(\cdot, x_n)$ is a majorant function for f at x_n.



$\Rightarrow$ The sequence $(f(x_n))_{n\in\mathbb{N}}$ is decreasing:

$$(\forall n \in \mathbb{N}) \quad f(x_{n+1}) \underset{\text{M}}{\leq} h(x_{n+1}, x_n) \underset{\text{M}}{\leq} h(x_n, x_n) = f(x_n)$$

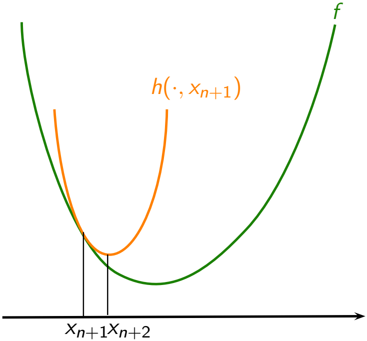
Majorization-Minimization algorithm

Problem: Minimization of function $f : \mathbb{R}^N \to \mathbb{R}$.

MM Algorithm

$$x_{n+1} \in \underset{x \in \mathbb{R}^N}{\text{Argmin}}\ h(x, x_n)$$

where $h(\cdot, x_n)$ is a majorant function for f at x_n.



$\Rightarrow$ The sequence $(f(x_n))_{n \in \mathbb{N}}$ is decreasing:

$$(\forall n \in \mathbb{N}) \quad f(x_{n+1}) \underset{\text{M}}{\leq} h(x_{n+1}, x_n) \underset{\text{M}}{\leq} h(x_n, x_n) = f(x_n)$$

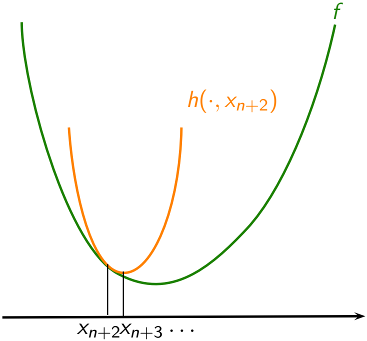
Majorization-Minimization algorithm

Problem: Minimization of function $f : \mathbb{R}^N \to \mathbb{R}$.

MM Algorithm

$$x_{n+1} \in \underset{x \in \mathbb{R}^N}{\text{Argmin}}\ h(x, x_n)$$

where $h(\cdot, x_n)$ is a majorant function for f at x_n.



$\Rightarrow$ The sequence $(f(x_n))_{n\in\mathbb{N}}$ is decreasing:

$$(\forall n \in \mathbb{N}) \quad f(x_{n+1}) \underset{\text{M}}{\leq} h(x_{n+1}, x_n) \underset{\text{M}}{\leq} h(x_n, x_n) = f(x_n)$$

Majorization techniques

- Subdifferential inequality
- Descent lemma
- Proximity operator
- Even smooth functions
- Jensen's inequality

Majorization techniques

Even differentiable function

Let f be defined as

$$(\forall x \in \mathbb{R}) \qquad f(x) = \psi(|x|)$$

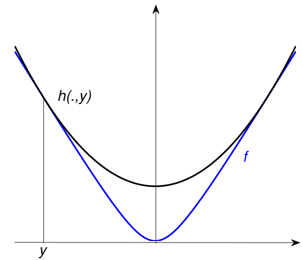
where

(i) ψ is differentiable on $]0, +\infty[$,

(ii) $\psi(\sqrt{\cdot})$ is concave on $]0, +\infty[$,

(iii) $(\forall x \in [0, +\infty[) \quad \dot{\psi}(x) \geq 0$,

(iv) $\lim_{\substack{x \to 0 \\ x > 0}} \left(\omega(x) := \frac{\dot{\psi}(x)}{x} \right) \in \mathbb{R}$.



Then, for all $y \in \mathbb{R}$,

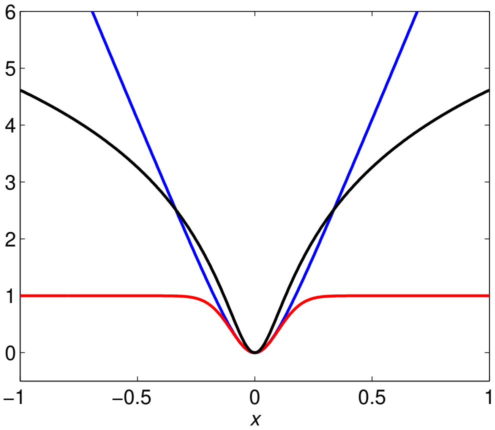
$$(\forall x \in \mathbb{R}) \quad f(x) \leq f(y) + \dot{f}(y)(x-y) + \frac{1}{2}\omega(|y|)(x-y)^2.$$

Examples of functions f

	$f(x)$	$\omega(x)$
Convex	$\|x\| - \delta \log(\|x\|/\delta + 1)$	$(\|x\| + \delta)^{-1}$
	$\begin{cases} x^2 & \text{if } \|x\| < \delta \\ 2\delta\|x\| - \delta^2 & \text{otherwise} \end{cases}$	$\begin{cases} 2 & \text{if } \|x\| < \delta \\ 2\delta/\|x\| & \text{otherwise} \end{cases}$
	$\log(\cosh(x))$	$\tanh(x)/x$
	$(1 + x^2/\delta^2)^{\kappa/2} - 1$	$(\kappa/\delta^2)(1 + x^2/\delta^2)^{\kappa/2-1}$
Nonconvex	$1 - \exp(-x^2/(2\delta^2))$	$(1/\delta^2)\exp(-x^2/(2\delta^2))$
	$x^2/(2\delta^2 + x^2)$	$4\delta^2/(2\delta^2 + x^2)^2$
	$\begin{cases} 1 - (1 - x^2/(6\delta^2))^3 & \text{if } \|x\| \le \sqrt{6}\delta \\ 1 & \text{otherwise} \end{cases}$	$\begin{cases} (1/\delta^2)(1 - x^2/(6\delta^2))^2 & \text{if } \|x\| \le \sqrt{6}\delta \\ 0 & \text{otherwise} \end{cases}$
	$\tanh(x^2/(2\delta^2))$	$(1/\delta^2)(\cosh(x^2/(2\delta^2)))^{-2}$
	$\log(1 + x^2/\delta^2)$	$2/(\delta^2 + x^2)$

$(\lambda, \delta) \in]0, +\infty[^2, \; \kappa \in [1, 2]$

Examples of functions f



$f(x) = (1 + \frac{x^2}{\delta^2})^{1/2} - 1$, $f(x) = \log\left(1 + \frac{x^2}{\delta^2}\right)$, $f(x) = 1 - \exp(-\frac{x^2}{2\delta^2})$.

Majorization techniques

Consequences of Jensen's inequality

Let $\psi : \mathbb{R}^N \to \mathbb{R}$ be a convex function.

- $(\forall (x,y,c) \in (]0,+\infty[^N)^3) \quad \psi\left(c^\top x\right) \leq \sum_{i=1}^{N} \frac{c^{(i)} y^{(i)}}{c^\top y} \psi\left(\frac{c^\top y}{y^{(i)}} x^{(i)}\right).$

- Let $\omega \in [0,+\infty[^N$ such that $\sum_{i=1}^N \omega^{(i)} = 1$ and $\omega^{(i)} = 0$ iff $c^{(i)} = 0$.

$$(\forall (x,y,c) \in (]-\infty,+\infty[^N)^3)$$
$$\psi\left(c^\top x\right) \leq \sum_{i=1}^{N} \omega^{(i)} \psi\left(\frac{c^{(i)}}{\omega^{(i)}}(x^{(i)} - y^{(i)}) + c^\top y\right).$$

MM algorithms

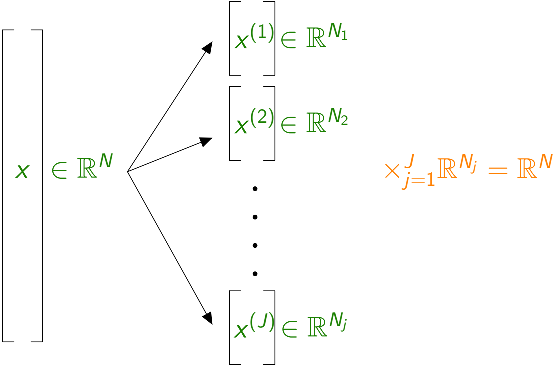
- Separable MM approach
- MM quadratic algorithm
- 3MG algorithm
- Forward-backward algorithm
- Block-alternating MM schemes

Acceleration via block-alternation

Problem: Minimization of $f : \mathbb{R}^N \to \mathbb{R}$.

Acceleration via block-alternation

Problem: Minimization of $f : \mathbb{R}^N \to \mathbb{R}$.



Acceleration via block-alternation

Problem: Minimization of $f : \mathbb{R}^N \to \mathbb{R}$.

$$f\left(\begin{bmatrix} \\ x \\ \\ \end{bmatrix}\right) = f\left(\begin{matrix} \begin{bmatrix} x^{(1)} \end{bmatrix} \\ \begin{bmatrix} x^{(2)} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} x^{(J)} \end{bmatrix} \end{matrix}\right)$$

Acceleration via block-alternation

Problem: Minimization of $f : \mathbb{R}^N \to \mathbb{R}$.

$$f\left(\begin{bmatrix} x \end{bmatrix}\right) = f\left(\begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(J)} \end{bmatrix}\right)$$

$\Rightarrow$ **Block-coordinate strategy:** Instead of updating the whole vector x at iteration $n \in \mathbb{N}$, restrict the update to a block $j_n \in \{1, \dots, J\}$.

Concluding remarks

- In large scale optimization, we search for the best possible tradeoff in terms of computational complexity and convergence rate.
- Availability of theoretical convergence results is important, to assess the reliability of an optimization scheme.
- There is rarely a single technique available for the resolution of an optimization problem.
- It is thus always recommended to test and compare different strategies, for a given application.

Not treated in this course: stochastic optimization, distributed algorithms, primal-dual strategies, etc.