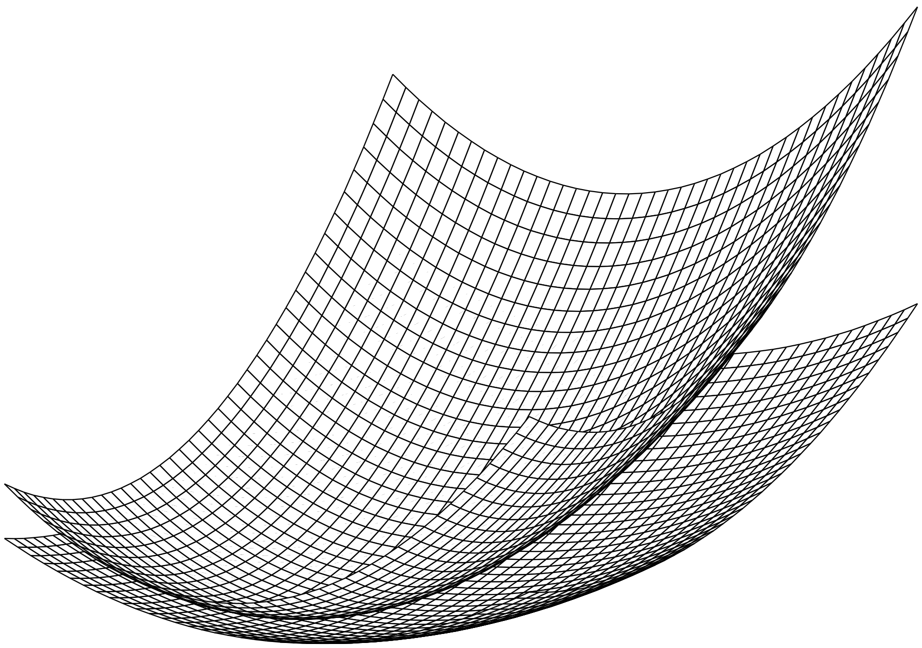
Data Sciences – ECP
Large Scale and Distributed Optimization
Part VI: Majorization-Minimization approaches

Emilie Chouzenoux and Jean-Christophe Pesquet

LIGM – CNRS UMR 8049
Univ. Paris-Est – ESIEE/ENPC
{emilie.chouzenoux,jean-christophe.pesquet}@univ-paris-est.fr

Majorization-Minimization principle

When it is successful, the MM algorithm substitutes a simple optimization problem for a difficult optimization problem. - K. Lange



Majorization-Minimization principle

Majorization - **M**inimization (MM)
(= optimization transfer = iterative majorization
= auxiliary function method = surrogate minimization)

The MM principle consists of solving a minimization problem by alternating between two steps:

1. **M**ajorize the criterion at current iterate with a majorant function ,
2. **M**inimize the majorant function to define the next iterate.

Majorization-Minimization principle

Majorization - **M**inimization (MM)
(= optimization transfer = iterative majorization
= auxiliary function method = surrogate minimization)

The MM principle consists of solving a minimization problem by alternating between two steps:

1. **M**ajorize the criterion at current iterate with a majorant function,
2. **M**inimize the majorant function to define the next iterate.

$\Rightarrow$ The construction of an MM algorithm thus requires to define

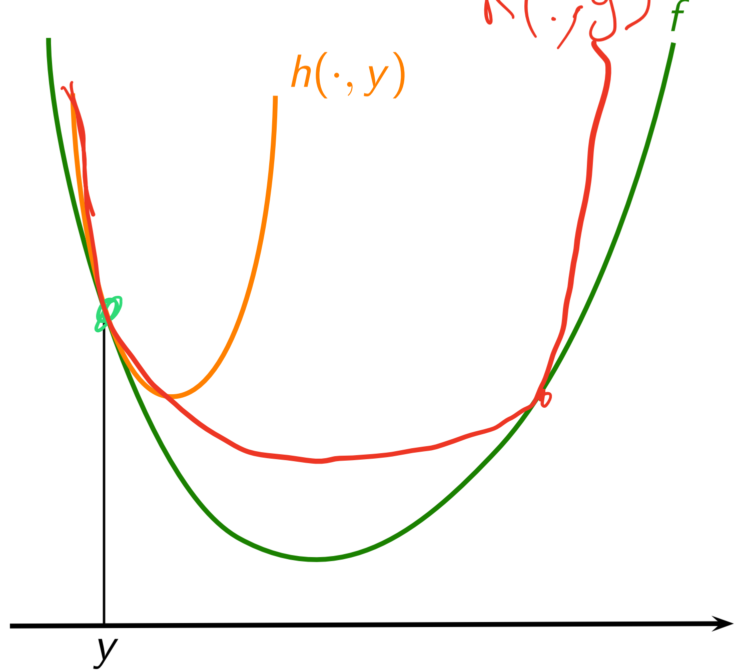
(i) a strategy for building majorant functions

(ii) a strategy for minimizing them.

Majorant function

Let $f : \mathcal{H} \to]-\infty, +\infty]$ where $\mathcal{H}$ is a Hilbert space. Let $y \in \mathcal{H}$.
$h(\cdot, y) : \mathcal{H} \to]-\infty, +\infty]$ is **a majorant function of f at y** if:

$$\begin{cases} (\forall x \in \mathcal{H}) & f(x) \leq h(x, y), \\ f(y) = h(y, y). \end{cases}$$



Majorant function

Properties

Let $f_1 : \mathcal{H} \to]-\infty, +\infty]$ and $f_2 : \mathcal{H} \to]-\infty, +\infty]$. Let $y \in \mathcal{H}$.
Let $h_1(\cdot, y) : \mathcal{H} \to]-\infty, +\infty]$ be a majorant function of f_1 at y,
and let $h_2(\cdot, y) : \mathcal{H} \to]-\infty, +\infty]$ be a majorant function of f_2 at y.

Sum

$h_1(\cdot, y) + h_2(\cdot, y)$ is a majorant function of $f_1 + f_2$ at y.

Product

If, for all $x \in \mathcal{H}$, $f_1(x) \geq 0$ and $f_2(x) \geq 0$, then $h_1(\cdot, y)h_2(\cdot, y)$ is a majorant function of $f_1 f_2$ at y.

Composition

If $\phi : \mathbb{R} \to]-\infty, +\infty]$ is an increasing function, then $\phi(h_1(\cdot, y))$ is a majorant function of $\phi(f_1)$ at y.

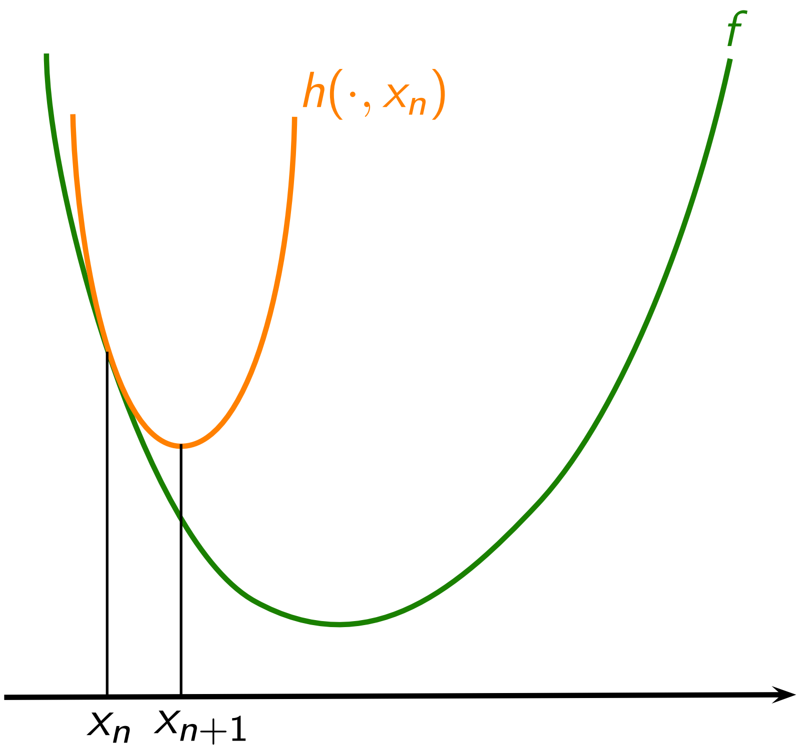
Majorization-Minimization algorithm

Problem: Minimization of function $f : \mathcal{H} \to \,]-\infty, +\infty]$.

MM Algorithm

$$x_{n+1} \in \underset{x \in \mathcal{H}}{\text{Argmin}} \; h(x, x_n)$$

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:

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

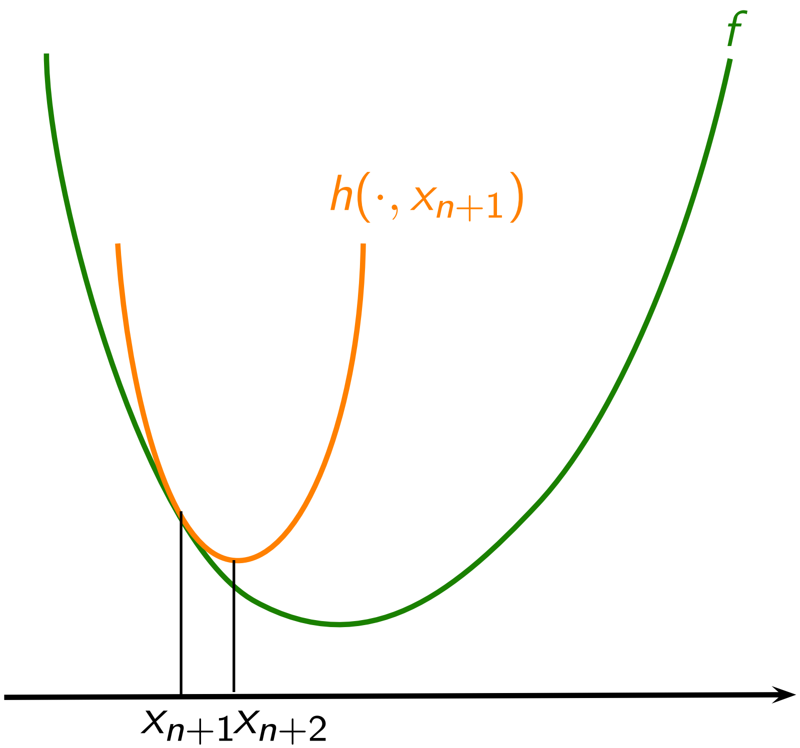
Majorization-Minimization algorithm

Problem: Minimization of function $f : \mathcal{H} \to]-\infty, +\infty]$.

MM Algorithm

$$x_{n+1} \in \underset{x \in \mathcal{H}}{\text{Argmin}} \; h(x, x_n)$$

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:

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

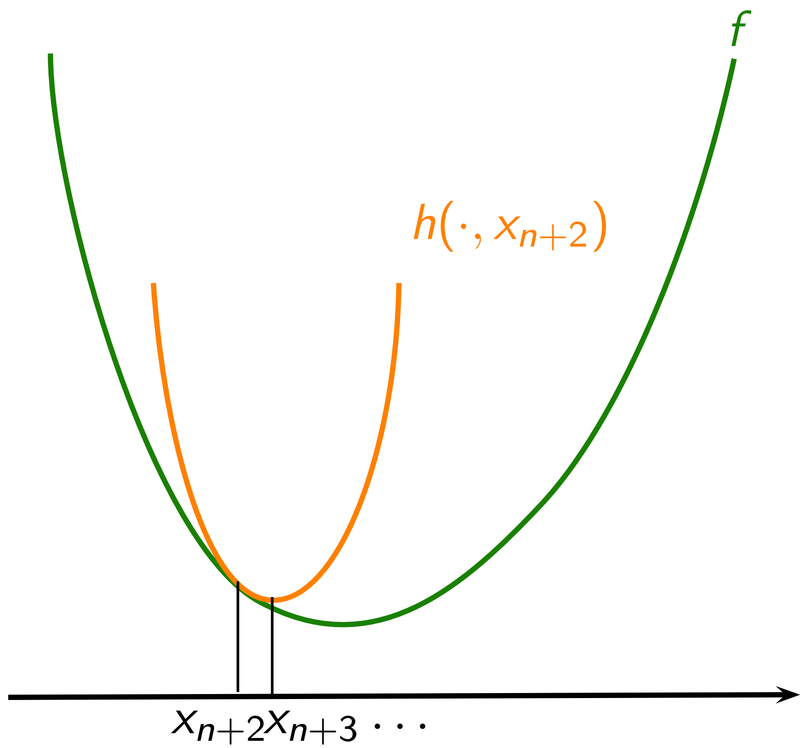
Majorization-Minimization algorithm

Problem: Minimization of function $f : \mathcal{H} \to \,]-\infty, +\infty]$.

MM Algorithm

$$x_{n+1} \in \underset{x \in \mathcal{H}}{\text{Argmin}}\; h(x, x_n)$$

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:

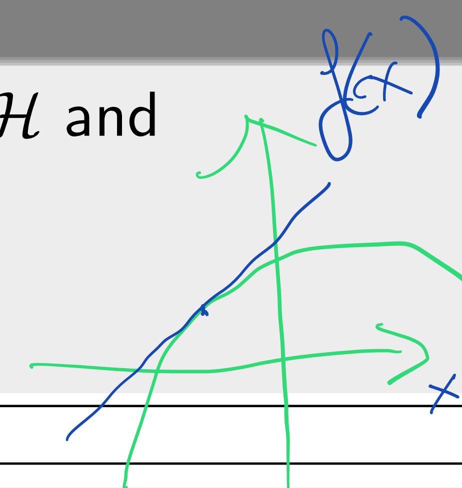
$$(\forall n \in \mathbb{N}) \quad f(x_{n+1}) \underset{\text{M}}{\leq} h(x_{n+1}, x_n) \underset{\text{M}}{\leq} h(x_n, x_n) = f(x_n)$$

Majorization techniques

Concave function

Let $f : \mathcal{H} \to [-\infty, +\infty[$ be a concave function. Let $y \in \mathcal{H}$ and $(-t) \in \partial(-f)(y)$. A majorant function for f at $y \in \mathcal{H}$ is

$$(\forall x \in \mathcal{H}) \quad h(x, y) = f(y) + \langle t | x - y \rangle .$$



Lipschitz differentiable function

Let $f : \mathcal{H} \to]-\infty, +\infty]$ a β-Lipschitz differentiable function on $\mathcal{H}$. Then, for every $y \in \mathcal{H}$ and for every $\mu \in [\beta, +\infty[$, a majorant function for f at $y \in \mathcal{H}$ is

$$(\forall x \in \mathcal{H}) \quad h(x, y) = f(y) + \langle \nabla f(y) | x - y \rangle + \frac{\mu}{2} \|x - y\|^2 .$$

Majorization techniques

Twice-differentiable function

Let $f : \mathbb{R}^N \to]-\infty, +\infty]$ be a twice differentiable function on $\mathbb{R}^N$ with Hessian $\nabla^2 f$. Let $A \in \mathbb{R}^{N\times N}$ a positive semidefinite matrix such that, for every $x \in \mathbb{R}^N$, $A - \nabla^2 f(x)$ is positive semidefinite. Then, for every $y \in \mathbb{R}^N$, a majorant function for f at $y \in \mathbb{R}^N$ is

$$(\forall x \in \mathbb{R}^N) \quad h(x, y) = f(y) + \langle \nabla f(y) | x - y\rangle + \frac{1}{2} \underbrace{\langle x - y \mid A(x - y)\rangle}_{\|x-y\|_A^2}.$$

Jensen's inequality

Let $\psi : \mathbb{R} \to]-\infty, +\infty]$ be a convex function and let $\omega = (\omega^{(i)})_{1\le i\le N} \in [0, +\infty[^N$ be such that $\sum_{i=1}^N \omega^{(i)} = 1$. Then,

$$(\forall (x^{(1)}, \ldots, x^{(N)}) \in \mathcal{H}^N) \quad \psi\left(\sum_{i=1}^N \omega^{(i)} x^{(i)}\right) \le \sum_{i=1}^N \omega^{(i)} \psi\left(x^{(i)}\right).$$

Whiteboard

Whiteboard

Exercises

Prove the following majorizing properties:

1. $(\forall (x,y) \in (\mathbb{R}^+)^2)(\forall q \in]0,1[) \quad x^q \leq q y^{q-1} x + (1-q) y^q$
2. $(\forall (x,y) \in (\mathbb{R}^{+*})^2) \quad \log x \leq \frac{x}{y} + \log y - 1$
3. $(\forall x \in \mathbb{R}^N) \quad \exp\left(\frac{1}{N}\sum_{i=1}^N x^{(i)}\right) \leq \frac{1}{N}\sum_{i=1}^N e^{x^{(i)}}$
4. $(\forall x \in \mathbb{R}^N)(\forall y \in \mathbb{R}^N \setminus \{0\}) \quad -\|x\| \leq -\frac{\langle x | y \rangle}{\|y\|}$
5. $(\forall (x,z) \in \mathbb{R}^2)(\forall (y,t) \in (\mathbb{R}^{+*})^2) \quad 2xz \leq \frac{x^2 t}{y} + \frac{z^2 y}{t}$

Exercises

Solutions :

1. Use concavity of $x \mapsto x^q$ on $\mathbb{R}^+$, for $q \in]0, 1[$.
2. Use concavity of $x \mapsto \log x$ on $\mathbb{R}^{+*}$.
3. Apply Jensen's inequality on the convex function exp.
4. Use concavity of $x \mapsto -\|x\|$.
5. Develop the inequality $(x/y - z/t)^2 \geq 0$, for $(x, z) \in \mathbb{R}^2$ and $(y, t) \in (\mathbb{R}^{+*})^2$.

Majorization techniques

Even differentiable function

Let f be defined as

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

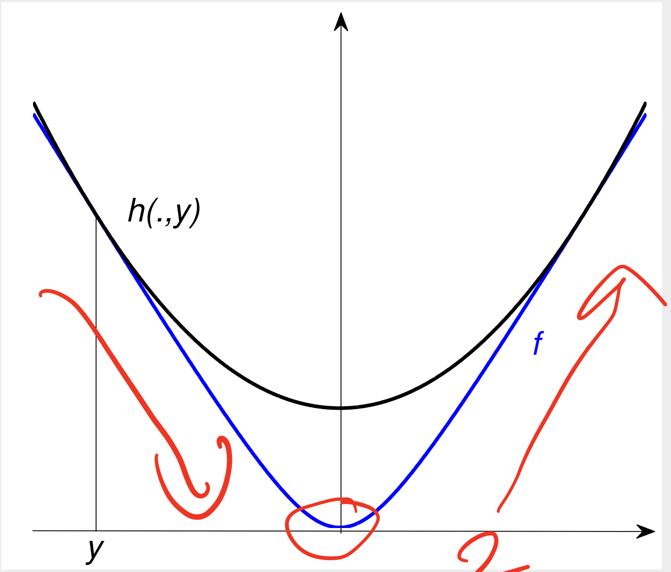
where

(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 $y \in \mathbb{R}$,

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

Proof

According to Assumption (ii), $\varphi = \psi(\sqrt{\cdot})$ is concave on $]0, +\infty[$. Thus, using Assumption (i), for all $(u, v) \in]0, +\infty[^2$, $\varphi(u) \leq \varphi(v) + (u - v)\dot{\varphi}(v)$, with $\dot{\varphi}(v) = \frac{\dot{\psi}(\sqrt{v})}{2\sqrt{v}}$ (which is positive by Assumption (iii)). Then, for every $(x, y) \in (\mathbb{R}^*)^2$,

$$\varphi(x^2) \leq \varphi(y^2) + (x^2 - y^2)\omega(|y|).$$

Using the equality $x^2 - y^2 = (x - y)^2 + 2y(x - y)$, we deduce that

$$\varphi(x^2) \leq \varphi(y^2) + \operatorname{sign}(y)\dot{\psi}(|y|)(x - y) + \frac{1}{2}(x - y)^2\omega(|y|),$$

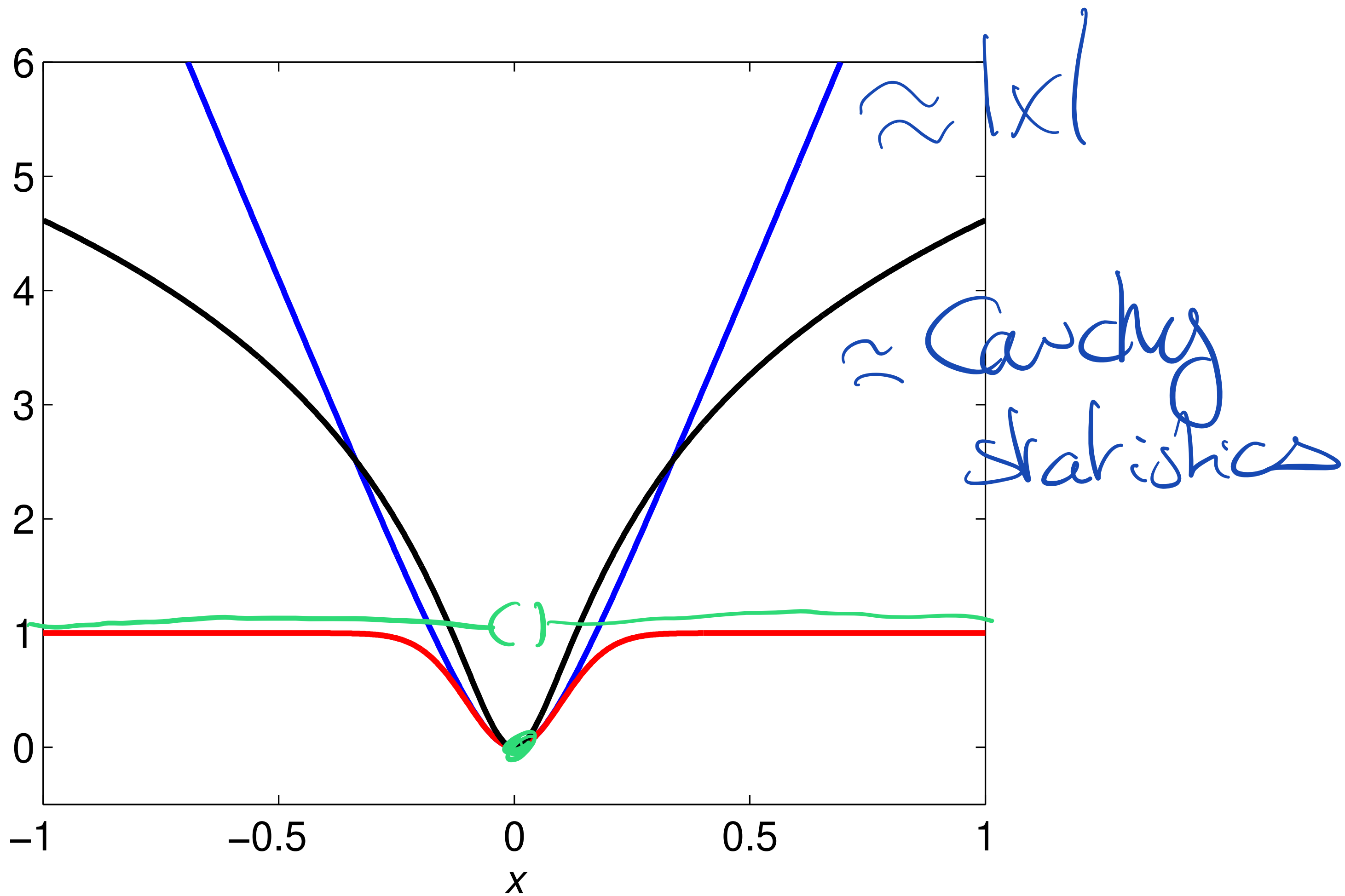
hence the result (by continuity, for $x = 0$ and/or $y = 0$ using Assumption (iv)).

Examples of functions f

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

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

Examples of functions f



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

MM quadratic algorithm

Problem: Minimization of a differentiable function $f : \mathcal{H} \to \mathbb{R}$.

Assumption: For every $y \in \mathcal{H}$, there exists a strongly positive self-adjoint operator $A(y)$ such that the quadratic function

$$(\forall x \in \mathcal{H}) \quad h(x, y) = f(y) + \langle \nabla f(y) | x - y \rangle + \frac{1}{2} \| x - y \|^2_{A(y)}$$

is a majorant function of f at y.

MM quadratic algorithm

$$x_{n+1} = x_n - \theta_n A(x_n)^{-1} \nabla f(x_n), \qquad \theta_n \in (0, 2).$$

⇒ $(\theta_n)_n$ acts as a stepsize parameter.
For $\theta_n \equiv 1$, we recover the basic MM algorithm.

Convergence properties

Assumptions

1. $f : \mathcal{H} \to \mathbb{R}$ is a coercive, differentiable function.
2. There exists $(\underline{\nu}, \overline{\nu}) \in]0, +\infty[^2$ such that $(\forall n \in \mathbb{N})$ $\underline{\nu}\mathrm{Id} \preceq A(x_n) \preceq \overline{\nu}\mathrm{Id}$,
3. There exists $(\underline{\theta}, \overline{\theta}) \in]0, +\infty[^2$ such that, $(\forall n \in \mathbb{N})$ $\underline{\theta} \leq \theta_n \leq 2 - \overline{\theta}$.

Sufficient descent property

There exists $(\mu_1, \mu_2) \in]0, +\infty[^2$ such that

$$(\forall n \in \mathbb{N}) \quad f(x_n) - f(x_{n+1}) \geq \mu_1 \|x_{n+1} - x_n\|^2 \geq \mu_2 \|\nabla f(x_n)\|^2.$$

Convergence theorem (in finite dimension)

1. $\nabla f(x_n) \to 0$ and $f(x_n) \searrow f(\widetilde{x})$ for some $\widetilde{x} \in \mathcal{H}$.
2. If f is continuously differentiable, any sequential cluster point of $(x_n)_{n\in\mathbb{N}}$ is a stationnary point of f.
3. If f is convex, any sequential cluster point of $(x_n)_{n\in\mathbb{N}}$ is a minimizer of f.
4. If f is strictly convex, then $x_n \to \widehat{x}$ where $\widehat{x}$ is the unique minimizer of f.

Proof

Let $n \in \mathbb{N}$. According to the majoration property, $f(x_{n+1}) \leq h(x_{n+1}, x_n)$, with $h(x_{n+1}, x_n) = f(x_n) + \langle \nabla f(x_n) | x_{n+1} - x_n \rangle + \frac{1}{2} \|x_{n+1} - x_n\|^2_{A(x_n)}$.
Moreover, we have $\nabla f(x_n) + \theta_n^{-1} A(x_n)(x_{n+1} - x_n) = 0$. Therefore, on the one hand,

$$\begin{aligned} f(x_{n+1}) &\leq f(x_n) - \left(\theta_n^{-1} - \frac{1}{2}\right) \|x_{n+1} - x_n\|^2_{A(x_n)}, \\ &\leq f(x_n) - \underbrace{\left(\frac{1}{2 - \overline{\theta}} - \frac{1}{2}\right) \underline{\nu}}_{\mu_1} \|x_{n+1} - x_n\|^2. \end{aligned}$$

On the other hand,

$$\begin{aligned} \|\nabla f(x_n)\| &= \theta_n^{-1} \|A(x_n)(x_{n+1} - x_n)\|, \\ &\leq \underbrace{\underline{\theta}^{-1} \overline{\nu}}_{\sqrt{\mu_1/\mu_2}} \|x_{n+1} - x_n\| \end{aligned}$$

Proof

Since f is coercive, $(f(x_n))_n$ is a decreasing bounded sequence so $(x_n)_n$ belongs to a compact subset of $\mathcal{H}$. Then, there exists a subsequence $(x_{n_k})_k$ which converges to some $\widetilde{x} \in \mathcal{H}$.

By continuity of f, $f(x_{n_k}) \longrightarrow f(\widetilde{x})$ so that $f(x_n) \searrow f(\widetilde{x})$.

According to the descent properties, $\|\nabla f(x_n)\| \longrightarrow 0$ and $\|x_{n+1} - x_n\| \longrightarrow 0$.

If f is continuously differentiable, $\nabla f(\widetilde{x}) = 0$, so that $\widetilde{x}$ is a critical point.

If f is convex, every critical point is a minimizer of f.

If f is strongly convex, the set of critical point is reduced to a singleton, equals to the unique minimizer of f.

Acceleration via subspace strategy

Problem: Minimization of differentiable function $f : \mathcal{H} \to \mathbb{R}$.

MM quadratic algorithm: $$x_{n+1} \in \underset{x \in \mathcal{H}}{\operatorname{Argmin}} \; h(x, x_n)$$

Difficulty: In the context of large scale optimization, the minimization of h over $\mathcal{H}$ may become untractable.

Acceleration via subspace strategy

Problem: Minimization of differentiable function $f : \mathcal{H} \to \mathbb{R}$.

MM quadratic algorithm: $$x_{n+1} \in \underset{x \in \mathcal{H}}{\operatorname{Argmin}} \; h(x, x_n)$$

Difficulty: In the context of large scale optimization, the minimization of h over $\mathcal{H}$ may become untractable.

⇒ **Subspace strategy:** Instead of minimizing h over the whole set $\mathcal{H}$, restrict the minimization space to a subspace spanned by a small number of vectors.

MM quadratic subspace algorithm: $$x_{n+1} \in \underset{x \in \operatorname{span}\left(d_n^1, d_n^2, \ldots, d_n^{M_n}\right)}{\operatorname{Argmin}} h(x, x_n),$$

where, for every $n \in \mathbb{N}$, $M_n \geq 1$, and $D_n = \left[d_n^1 \,|\, d_n^2 \,|\, \ldots \,|\, d_n^{M_n}\right] \in \mathcal{H}^{M_n}$.

Choices for the subspace

Subspace name	**Set of directions D_n**
Memory gradient	$[-\nabla f(x_n) \,|\, d_{n-1}]$
Supermemory gradient	$[-\nabla f(x_n) \,|\, d_{n-1} \,|\, \ldots \,|\, d_{n-m}]$
Gradient subspace	$[-\nabla f(x_n) \,|\, -\nabla f(x_{n-1}) \,|\, \ldots \,|\, -\nabla f(x_{n-m})]$
Nemirovski subspace	$\left[-\nabla f(x_n) \,|\, x_n - x_0 \,|\, \sum_{i=0}^{n} \omega_i \nabla f(x_i)\right]$
Sequential subspace	$\left[-\nabla f(x_n) \,|\, x_n - x_0 \,|\, \sum_{i=0}^{n} \omega_i \nabla f(x_i) \,|\, d_{n-1} \,|\, \ldots \,|\, d_{n-m}\right]$
Quasi-Newton subspace	$[-\nabla f(x_n) \,|\, \delta_{n-1} \,|\, \ldots \,|\delta_{n-m} \,|\, d_{n-1} \,|\, \ldots \,|\, d_{n-m}]$

where, for all $n \geq 0$, $(\omega_i)_{1 \leq i \leq n} \in \mathbb{R}^n$, $d_n = x_{n+1} - x_n$ and $\delta_n = \nabla f(x_{n+1}) - \nabla f(x_n)$.

MM quadratic subspace algorithm

Problem: Minimization of $f : \mathcal{H} \to \mathbb{R}$ where f is differentiable.

Assumption: For all $y \in \mathcal{H}$, there exists a strongly positive self-adjoint operator $A(y)$ such that the quadratic function

$$(\forall x \in \mathcal{H}) \quad h(x, y) = f(y) + \langle \nabla f(y) | x - y \rangle + \frac{1}{2} \|x - y\|^2_{A(y)}$$

is a majorant function of f at y.

MM quadratic subspace algorithm

Choose $D_n \in \mathcal{H}^{M_n}$,

$$u_n \in \underset{u \in \mathbb{R}^{M_n}}{\operatorname{Argmin}}\, h\left(x_n + \sum_{m=1}^{M_n} u^{(m)} d_n^m, x_n \right),$$

$$x_{n+1} = x_n + \sum_{m=1}^{M_n} u_n^{(m)} d_n^m.$$

$\Rightarrow$ **3MG algorithm** obtained when $(D_n)_n$ is the memory gradient subspace.

Convergence properties

Assumptions

1. $f : \mathcal{H} \to \mathbb{R}$ is a coercive, differentiable function.
2. There exists $(\underline{\nu}, \overline{\nu}) \in]0, +\infty[^2$ such that $(\forall n \in \mathbb{N})$ $\underline{\nu}\mathrm{Id} \preceq A(x_n) \preceq \overline{\nu}\mathrm{Id}$,
3. There exists $(\gamma_0, \gamma_1) \in]0, +\infty[^2$ such that

$$(\forall n \in \mathbb{N}) \quad \left\langle \nabla f(x_n) | d_n^1 \right\rangle \leq -\gamma_0 \|\nabla f(x_n)\|^2 \text{ and } \|d_n^1\| \leq \gamma_1 \|\nabla f(x_n)\|.$$

Sufficient descent property

There exists $(\mu_1, \mu_2) \in]0, +\infty[^2$ such that

$$(\forall n \in \mathbb{N}) \quad f(x_n) - f(x_{n+1}) \geq \mu_1 \|x_{n+1} - x_n\|^2 \geq \mu_2 \|\nabla f(x_n)\|^2.$$

Convergence theorem (in finite dimension)

1. $\nabla f(x_n) \to 0$ and $f(x_n) \searrow f(\widetilde{x})$ where $\widetilde{x}$ is a critical point of f.
2. If f is convex, any sequential cluster point of $(x_n)_{n \in \mathbb{N}}$ is a minimizer of f.
3. If f is strictly convex, then $x_n \to \widehat{x}$ where $\widehat{x}$ is the unique minimizer of f.

Whiteboard

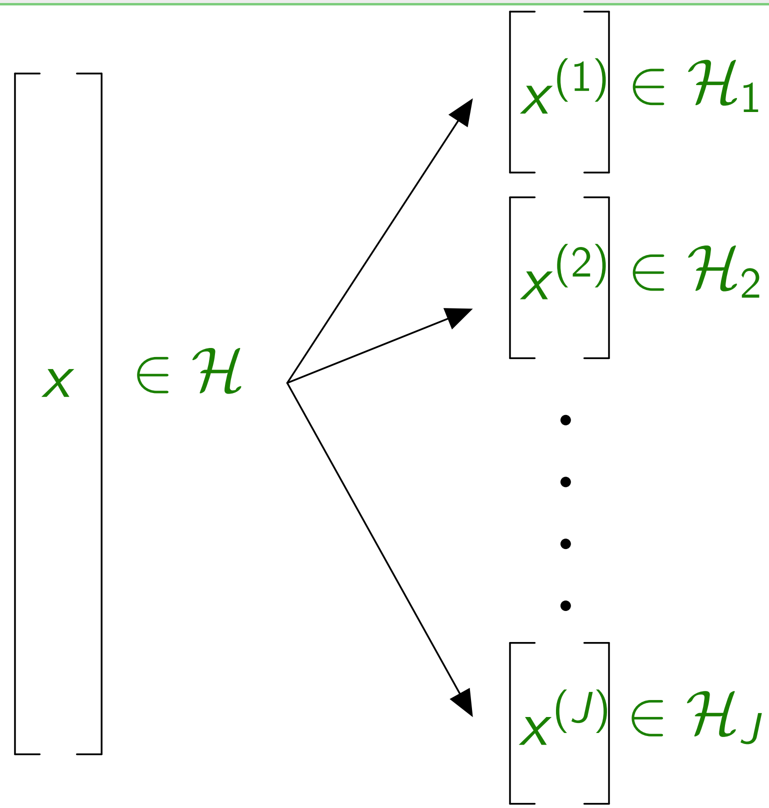
Whiteboard

Acceleration via block-alternation

Problem: Minimization of $f : \mathcal{H} \to]-\infty, +\infty]$.

Acceleration via block-alternation

Problem: Minimization of $f : \mathcal{H} \to \,]-\infty, +\infty]$.



$\times_{j=1}^{J} \mathcal{H}_j = \mathcal{H}$

where $\mathcal{H}_1, \ldots, \mathcal{H}_J$ separable real Hilbert spaces.

Acceleration via block-alternation

Problem: Minimization of $f : \mathcal{H} \to]-\infty, +\infty]$.

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

Acceleration via block-alternation

Problem: Minimization of $f : \mathcal{H} \to]-\infty, +\infty]$.

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

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

Block-coordinate MM quadratic algorithm

Problem: Minimization of $f : \mathcal{H} \to \mathbb{R}$ where f is differentiable.

Assumption: For every $y \in \mathcal{H}$, for every $j \in \{1, \dots, J\}$, there exists a strongly positive self-adjoint $A_j(y)$ such that the quadratic function

$$(\forall x^{(j)} \in \mathcal{H}_j)\; h_j(x^{(j)}, y^{(j)}; y) = f(y) + \langle \nabla_j f(y) | x^{(j)} - y^{(j)} \rangle + \frac{1}{2} \| x^{(j)} - y^{(j)} \|^2_{A_j(y)}$$

is a majorant function at $y^{(j)}$ of the restriction of f to its j-th block.

Block-coordinate MM quadratic algorithm

$$\begin{aligned}
&\text{Select } j_n \in \{1, \dots, J\}, \\
&x_{n+1}^{(j_n)} = x_n^{(j_n)} - \theta_n A_{j_n}(x_n)^{-1} \nabla_{j_n} f(x_n), \\
&x_{n+1}^{(\bar{\jmath}_n)} = x_n^{(\bar{\jmath}_n)},
\end{aligned}$$

where $\bar{\jmath}_n = \{1, \dots, J\} \setminus \{j_n\}$.

Selection of blocks

At each iteration $n \in \mathbb{N}$, $j_n \in \{1, \ldots, J\}$ can be chosen according to:

- the cyclic rule:

$$(\forall n \in \mathbb{N}) \quad j_n - 1 = n \bmod (J).$$

- a quasi-cyclic rule:
  There exists a constant $K \geq J$ such that, for every $n \in \mathbb{N}$,

$$\{1, \ldots, J\} \subset \{j_n, \ldots, j_{n+K-1}\}.$$

- a random rule:
  For every $n \in \mathbb{N}$, j_n is a realization of a random variable.

$\Rightarrow$ The convergence properties of the algorithm may depend on the block selection rule.

Convergence properties

Assumptions

1. $f : \mathbb{R}^N \to \mathbb{R}$ is a coercive, differentiable function.
2. There exists $(\underline{\nu}, \overline{\nu}) \in]0, +\infty[^2$ such that $(\forall n \in \mathbb{N})\ \underline{\nu}\mathrm{Id} \preceq A_{j_n}(x_n) \preceq \overline{\nu}\mathrm{Id}$
3. There exists $(\underline{\theta}, \overline{\theta}) \in]0, +\infty[^2$ such that $(\forall n \in \mathbb{N})\ \underline{\theta} \leq \theta_n \leq 2 - \overline{\theta}$.

Sufficient descent property

There exists $(\mu_1, \mu_2) \in]0, +\infty[^2$ such that

$$(\forall n \in \mathbb{N}) \quad f(x_n) - f(x_{n+1}) \geq \mu_1 \|x_{n+1} - x_n\|^2 \geq \mu_2 \|\nabla_{j_n} f(x_n)\|^2.$$

Convergence theorem (in finite dimension and (quasi-)cyclic rule)

1. $\nabla f(x_n) \to 0$ and $f(x_n) \searrow f(\widetilde{x})$ where $\widetilde{x}$ is a critical point of f.
2. If f is convex, any sequential cluster point of $(x_n)_{n \in \mathbb{N}}$ is a minimizer of f.
3. If f is strictly convex, then $x_n \to \widehat{x}$ where $\widehat{x}$ is the unique minimizer of f.

Whiteboard

Whiteboard

Case of non differentiable function

Problem: Minimization of $f : \mathcal{H} \to]-\infty, +\infty]$ where $f = f_1 + f_2$ with f_1 differentiable and f_2 non necessarily differentiable.

MM Algorithm: $$x_{n+1} \in \underset{x \in \mathcal{H}}{\text{Argmin}}\ h(x, x_n)$$

Difficulty: How to majorize the non-differentiable function f, so that the majorants remain easy to minimize?

Case of non differentiable function

Problem: Minimization of $f : \mathcal{H} \to \,]-\infty, +\infty]$ where $f = f_1 + f_2$ with f_1 differentiable and f_2 non necessarily differentiable.

MM Algorithm: $$x_{n+1} \in \underset{x \in \mathcal{H}}{\text{Argmin}}\; h(x, x_n)$$

Difficulty: How to majorize the non-differentiable function f, so that the majorants remain easy to minimize?

$\Rightarrow$ Two main approaches:

1. Use quadratic majorant functions for f (but, numerical issues at non differentiability points)
   $\rightsquigarrow$ Iterative Reweighted Least Squares algorithms (e.g. Weiszfeld, FOCUSS, IRLS, ...)

Case of non differentiable function

Problem: Minimization of $f : \mathcal{H} \to]-\infty, +\infty]$ where $f = f_1 + f_2$ with f_1 differentiable and f_2 non necessarily differentiable.

MM Algorithm: $$x_{n+1} \in \underset{x \in \mathcal{H}}{\text{Argmin}}\ h(x, x_n) + f_2(x)$$

Difficulty: How to majorize the non-differentiable function f, so that the majorants remain easy to minimize?

⇒ Two main approaches:

1. Use quadratic majorant functions for f (but, numerical issues at non differentiability points)
   $\rightsquigarrow$ Iterative Reweighted Least Squares algorithms (e.g. Weiszfeld, FOCUSS, IRLS, ...)
2. Use quadratic majorant function for f_1, and keep f_2 untouched
   $\rightsquigarrow$Variable metric forward-backward algorithm

Proximity operator within a metric

Definition

Let $f \in \Gamma_0(\mathcal{H})$. Let $A : \mathcal{H} \to \mathcal{H}$ be a strongly positive self-adjoint operator. For all $x \in \mathcal{H}$, $\mathrm{prox}_{A,f}(x)$ is the proximity operator of f in $(\mathcal{H}, \|\cdot\|_A)$, i.e. the unique minimizer of

$$y \mapsto f(y) + \frac{1}{2}\|x - y\|_A^2.$$

Remarks:

- If $A = \alpha^{-1}\mathrm{Id}$, with $\alpha > 0$, then $\mathrm{prox}_{\alpha^{-1}\mathrm{Id},f} \equiv \mathrm{prox}_{\alpha f}$ corresponds to the usual proximity operator.
- We have

$$(\forall x \in \mathbb{R}^N) \quad \mathrm{prox}_{A,f}(x) = A^{-1/2}\mathrm{prox}_{f \circ A^{-1/2}}(A^{1/2}x).$$

Property

Let $f \in \Gamma_0(\mathbb{R}^N)$. Assume that f is separable, i.e.

$$(\forall x = (x^{(i)})_{1\le i\le N} \in \mathbb{R}^N) \quad f(x) = \sum_{i=1}^{N} f_i(x^{(i)})$$

and A is diagonal with (strictly) positive diagonal elements $(a_i)_{1\le i\le N}$.
Then, for every $x \in \mathbb{R}^N$, $\mathrm{prox}_{A,f}(x) = p$ where $p = (p^{(i)})_{1\le i\le N} \in \mathbb{R}^N$ is given by

$$(\forall i \in \{1,\ldots,N\}) \quad p^{(i)} = \mathrm{prox}_{a_i^{-1} f_i}(x^{(i)}).$$

Variable metric forward-backward algorithm

Problem: Minimization of $f : \mathcal{H} \to]-\infty, +\infty]$ where $f = f_1 + f_2$ with f_1 differentiable and f_2 convex non necessarily differentiable.

Assumption: For every $y \in \mathcal{H}$, there exists a strongly positive self-adjoint operator $A(y) : \mathcal{H} \to \mathcal{H}$ such that the quadratic function

$$(\forall x \in \mathcal{H}) \quad h(x, y) = f_1(y) + \langle \nabla f_1(y) | x - y \rangle + \frac{1}{2} \| x - y \|^2_{A(y)}$$

is a majorant function of f_1 at y.

VMFB algorithm

$$x_{n+1} = \operatorname{prox}_{\theta_n^{-1} A(x_n),\, f_2} \left(x_n - \theta_n A(x_n)^{-1} \nabla f_1(x_n) \right).$$

$\Rightarrow$ $(\theta_n)_n$ acts as a stepsize parameter.

Variable metric forward-backward algorithm

Link between MM and VMFB algorithms

Let $\theta_n \equiv 1$. According to the definition of the proximity operator,

$$
\begin{aligned}
x_{n+1} &= \underset{x\in\mathcal{H}}{\operatorname{argmin}} \ \frac{1}{2}\|x - x_n + A(x_n)^{-1}\nabla f_1(x_n)\|^2_{A(x_n)} + f_2(x) \\
&= \underset{x\in\mathcal{H}}{\operatorname{argmin}} \ \left\langle x - x_n | A(x_n)^{-1}\nabla f_1(x_n)\right\rangle_{A(x_n)} + \frac{1}{2}\|x - x_n\|^2_{A(x_n)} + f_2(x) \\
&= \underset{x\in\mathcal{H}}{\operatorname{argmin}} \ \langle x - x_n | \nabla f_1(x_n)\rangle + \frac{1}{2}\|x - x_n\|^2_{A(x_n)} + f_2(x) \\
&= \underset{x\in\mathcal{H}}{\operatorname{argmin}} \ h(x, x_n) + f_2(x).
\end{aligned}
$$

Particular case: Assume that f_1 is β-Lipschitz differentiable. According to the descent lemma, a possible choice for the metric is $A(x_n) \equiv \beta^{-1}\mathrm{Id}$. Then, VMFB algorithm becomes equivalent to the usual forward-backward algorithm.

Convergence properties

Assumptions

1. $f_1 : \mathcal{H} \to \mathbb{R}$ is a coercive, differentiable function.
   $f_2 \in \Gamma_0(\mathcal{H})$ is continuous on its domain.
2. There exists $(\underline{\nu}, \overline{\nu}) \in]0, +\infty[^2$ such that, $(\forall n \in \mathbb{N})\ \underline{\nu}\mathrm{Id} \preceq A(x_n) \preceq \overline{\nu}\mathrm{Id}$,
3. There exists $(\underline{\theta}, \overline{\theta}) \in]0, +\infty[^2$ such that, $(\forall n \in \mathbb{N})\ \underline{\theta} \leq \theta_n \leq 2 - \overline{\theta}$.

Sufficient descent property

There exists $(\mu_1, \mu_2) \in]0, +\infty[^2$ such that

$$(\forall n \in \mathbb{N})\ f(x_n) - f(x_{n+1}) \geq \mu_1 \|x_{n+1} - x_n\|^2 \geq \mu_2 \|\nabla f_1(x_n) + r_n\|^2, \text{ with } r_n \in \partial f_2(x_n).$$

Convergence theorem (in finite dimension)

1. $\nabla f_1(x_n) + r_n \to 0$ and $f(x_n) \searrow f(\widetilde{x})$ where $\widetilde{x}$ is a critical point of f.
2. If f_1 is convex, any sequential cluster point of $(x_n)_{n\in\mathbb{N}}$ is a minimizer of f.
3. If f is strictly convex, then $(x_n)_n \to \widehat{x}$ where $\widehat{x}$ is a minimizer of f.

Whiteboard

Whiteboard