Gradient à pas optimal

⊳ <u>Refs</u> : FGN Analyse 4 p39-41 (thm), Gourdon Analyse p 365-366 (lemme)

Définition : Soit $f : \mathbb{R}^n \to \mathbb{R}$ et $\alpha > 0$. On dit que f est α-convexe si $\forall x, y \in \mathbb{R}^n$, $\forall t \in [0,1]$, $f(tx + (1-t)y) \leq t f(x) + (1-t) f(y) - \frac{\alpha}{2} t(1-t)\|x-y\|^2$

Lemme : Soit $f \in C^1(\mathbb{R}^n, \mathbb{R})$. Il y a équivalence entre :
1) f est α-convexe
2) $\forall x, y \in \mathbb{R}^n$, $f(y) \underset{(*)}{\geq} f(x) + \langle \nabla f(x), y-x \rangle + \frac{\alpha}{2}\|y-x\|^2$

Théorème : Soit $f \in C^1(\mathbb{R}^n, \mathbb{R})$ une fonction α-convexe. Alors f admet un unique minimum global atteint en x^*, et la suite définie par $x_0 \in \mathbb{R}^n$ et pour $k \in \mathbb{N}$, $x_{k+1} = x_k + \lambda_k \nabla f(x_k)$ où $\lambda_k = \begin{cases} \underset{t \in \mathbb{R}}{\operatorname{arg\,min}}\, f(x_k + t\nabla f(x_k)) & \text{si } x_k \neq x^* \\ 0 & \text{sinon} \end{cases}$
est bien définie et converge vers x^*.

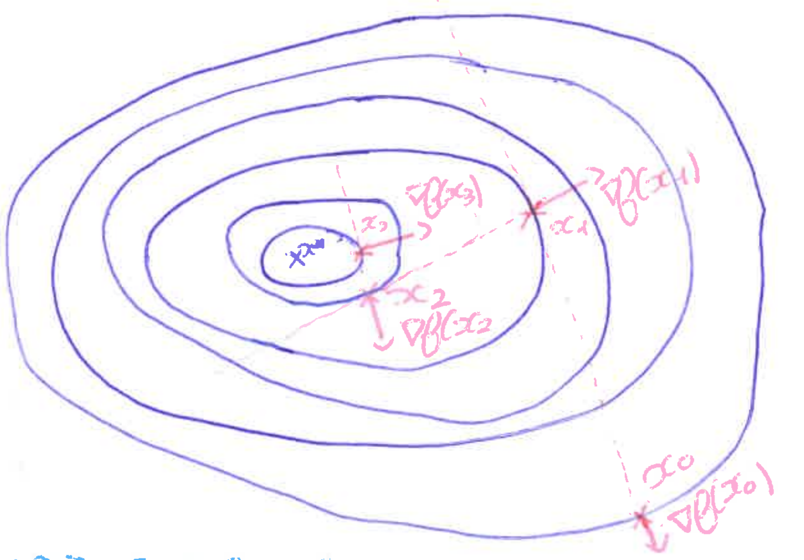
Preuve du théorème : <u>Etape 1</u> : Montrons l'existence et l'unicité de x^*.
L'équation $(*)$ du lemme pour $x = 0$ et $y \in \mathbb{R}^n$ nous donne :
$$f(y) \geq f(0) + \langle \nabla f(0), y \rangle + \frac{\alpha}{2}\|y\|^2 = \frac{\alpha}{2}\|y\|^2 + \underset{\|y\| \to +\infty}{O}(\|y\|) \xrightarrow[\|y\| \to +\infty]{} +\infty.$$
Ainsi $\lim_{\|y\| \to +\infty} f(y) = +\infty$ donc f est coercive.

Et une fonction coercive admet un minimum global : puisque f est coercive, il existe $M > 0$ tel que pour $\|x\| > M$, $f(x) > f(0)$. La fonction f est continue sur le compact $B(0, M)$ (dimension finie) donc elle est bornée et atteint son minimum en un point x^*. Et pour $\|x\| > M$, on a $f(x) > f(0) \geq f(x^*)$, donc x^* est un minimum global de f.
En particulier x^* est un point critique, donc $\nabla f(a) = 0$. Ainsi pour $y \neq x^*$, on a $f(y) \geq f(x^*) + \frac{\alpha}{2}\|y - x^*\|^2 > f(x^*)$. Donc x^* est l'unique minimum de f.

Avant de passer à l'étape 2, et d'expliquer la bonne définition de l'algorithme, et sa convergence, essayons de comprendre d'où vient l'idée :

Soit $a \in \mathbb{R}^n$ tel que $\nabla f(a) \neq 0$. Alors pour tout $x \in \mathbb{R}^n$ et $\lambda \in \mathbb{R}$ petit, on a
$$f(a + \lambda x) = f(a) + \lambda \underbrace{\langle \nabla f(a), x \rangle}_{} + o(\lambda)$$
$$-\|x\|\|\nabla f(a)\| \leq \langle \nabla f(a), x \rangle \leq \|x\|\|\nabla f(a)\|$$
Donc pour minimiser f, il faut prendre la direction de $\nabla f(a)$ (+ précisément dans le sens de $-\nabla f(a)$, mais sa n'impacte que le signe de λ) ⊳ et ensuite on va minimiser pour obtenir le meilleur λ
(on va montrer que 2 direct° successives sont $\perp$)

Représentation sur un exemple :
(les courbes représentent les lignes de niveau ($\{x, f(x)=c\}$)
(il faut imaginer une fonction du style



(gradient $\perp$ à la ligne de niveau :
$C=\{x, f(x)=c\}$, $a \in C$, si $\gamma :]-\varepsilon, \varepsilon[\to C$ ← TFI
telle que $\gamma(0)=a$, alors $g = f \circ \gamma :]-\varepsilon,\varepsilon[\to \mathbb{R}$ cste égale à c, donc de dérivée nulle or $0 = g'(0) = df(\gamma(0)) \cdot \gamma'(0) = \langle \nabla f(x), \gamma'(0) \rangle$.

Revenons en à la preuve !

Etape 2 : Justifions la bonne définition de la suite, en montrant que pour tout $x \in \mathbb{R}^n$ tel que $\nabla f(x) \neq 0$, $t \mapsto f(x + t\nabla f(x))$ admet un unique minimum.

Soit $x \in \mathbb{R}^n$ tel que $\nabla f(x) \neq 0$. On définit $\varphi_x : t \mapsto f(x + t\nabla f(x))$.
Alors φ est de classe C^1 et pour $t, t' \in \mathbb{R}$ et $\lambda \in [0,1]$, on a

$$\varphi_x(\lambda t + (1-\lambda)t') = f(\lambda(x + t\nabla f(x)) + (1-\lambda)(x + t'\nabla f(x))$$
$$\leq \lambda \varphi_x(t) + (1-\lambda)\varphi_x(t') - \frac{\alpha}{2}\lambda(1-\lambda) \times \|\nabla f(x)\|^2 \times |t'-t|^2$$

donc φ_x est $\alpha\|\nabla f(x)\|^2$ (>0) convexe donc d'après l'étape 1, on sait que φ_x admet un unique minimum global a_x. La suite est donc bien définie. De plus, en dérivant on obtient

$$0 = \varphi_x'(a_x) = df(x + a_x \nabla f(x)) \cdot \nabla f(x) = \langle \nabla f(x + a_x\nabla f(x)), \nabla f(x) \rangle$$

Donc $\nabla f(x) \perp \nabla f(x + a_x \nabla f(x))$. (*)

Etape 3 : Montrons que la suite converge vers x^*.
Si la suite atteint x^*, elle stationne donc le résultat est clair.
Supposons que ce ne soit pas le cas. La suite $(f(x_k))_{k\in\mathbb{N}}$ est décroissante par construction. Comme f est minorée, cette suite l'est également et donc elle converge vers une certaine limite l. En particulier, $f(x_{k+1}) - f(x_k) \xrightarrow[k\to\infty]{} 0$
Or nous avons vu à l'étape 2 ((*)) que pour tout $k \in \mathbb{N}$,
$\nabla f(x_k) \perp \nabla f(x_{k+1})$. Ainsi, on obtient pour tout $k \in \mathbb{N}$,

$$f(x_k) - f(x_{k+1}) \geq \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle + \frac{\alpha}{2}\|x_{k+1} - x_k\|^2$$
$$= -\lambda_k \langle \nabla f(x_k), \nabla f(x_{k+1}) \rangle + \frac{\alpha}{2}\|x_{k+1} - x_k\|^2 = \frac{\alpha}{2}\|x_{k+1} - x_k\|^2$$

2

Ainsi, $\|x_{k+1}-x_k\| \xrightarrow[k\to+\infty]{} 0$ par encadrement (donc $x_{k+1}-x_k \xrightarrow[k\to+\infty]{} 0$).

Puisque f est coercive, il existe $A>0$ tel que si $\|x\|>A$, $f(x)>f(x_0)$. Par décroissance de la suite $(f(x_k))_{k\in\mathbb{N}}$, on en déduit que pour tout $k\in\mathbb{N}$, $x_k\in B(0,A)$ compact. Ainsi d'après le théorème de Bolzano-Weierstrass, la suite $(x_k)_{k\in\mathbb{N}}$ admet une valeur d'adhérence : il existe $\varphi : \mathbb{N}\to\mathbb{N}$ extractrice et $x\in\mathbb{R}^n$ tels que $x_{\varphi(k)} \xrightarrow[k\to+\infty]{} x$. Par continuité de ∇f, on a $\nabla f(x_{\varphi(k)}) \xrightarrow[k\to+\infty]{} \nabla f(x)$. De plus, $x_{\varphi(k)+1} \xrightarrow[k\to+\infty]{} x$ car $\|x_{\varphi(k)+1}-x_k\| \xrightarrow[k\to+\infty]{} 0$. Ainsi, on a également $\nabla f(x_{\varphi(k)+1}) \xrightarrow[k\to+\infty]{} \nabla f(x)$.

De tout cela, on déduit :

$$0 = \langle \nabla f(x_{\varphi(k)+1}), \nabla f(x_{\varphi(k)})\rangle \xrightarrow[k\to+\infty]{} \|\nabla f(x)\|^2$$

Donc $\nabla f(x)=0$, donc $x=x^*$ (car f est α-convexe donc en particulier convexe donc puisqu'elle est C^1, x minimum $\Leftrightarrow \nabla f(x)=0$).
La suite est à valeurs dans le compact $B(0,A)$ et possède une unique valeur d'adhérence, donc la suite $(x_k)_{k\in\mathbb{N}}$ converge vers x^*.

(Bonus) Preuve du lemme : Montrons tout d'abord 1) $\Rightarrow$ 2).
Si f est α-convexe, alors $\forall (x,y)\in\mathbb{R}^n$ et $t\in]0,1]$, on a

$$\frac{f(y+t(x-y))-f(y)}{t} \leq f(x)-f(y)-\frac{\alpha}{2}(1-S)\|x-y\|^2,$$

et on obtient le résultat en faisant tendre $t\to 0$. (les rôles de x et y sont échangés p/r à la formule de l'énoncé).

Montrons maintenant 2) $\Rightarrow$ 1). On écrit 2) pour $(x+t(y-x), y)$ et pour $(x+t(y-x), x)$ avec $t\in[0,1]$, alors

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

On multiplie la 1ère inégalité par t, la seconde par $1-t$, on additionne et on obtient $tf(y)+(1-t)f(x)\geq f(x+t(y-x))+\frac{\alpha}{2}t(1-t)\|y-x\|^2$, donc f est α-c

Avant de passer aux remarques, quelques rappels / compléments sur la convexité :

- Si f est C^1, f est α-convexe $\Leftrightarrow \forall x,y$, $\langle\nabla f(y)-\nabla f(x), y-x\rangle \geq \alpha\|y-x\|^2$
- Si f est C^2, f est α-convexe $\Leftrightarrow \forall x,y$, $\langle \mathrm{Hess}(f)(x)y, y\rangle \geq \alpha\|y\|^2$

⚠ dim finie

- Si $U\subset\mathbb{R}^n$ ouvert convexe et $f: U\to\mathbb{R}^n$ convexe, alors f est continue. ("dessine moi une fct°CVX"

* Rappels sur convexité et minimum :

en dim finie :

	Existence	Unicité	locale -> globale
CVX	X $x \mapsto x$	X $x \mapsto 1$	✓
strict. CVX	X $x \mapsto e^x$	✓	✓
forte CVX	✓	✓	✓

(en dim infinie, on peut obtenir des résultats avec la topologie faible).

$f : U \to \mathbb{R}$ C^1 (ou juste différentiable) et convexe. Alors x_0 min de $f \iff \nabla f(x_0) = 0$

"$\Rightarrow$" ok

"$\Leftarrow$" $\forall y \in U$,

$f(y) \geq f(x_0) + \langle \nabla f(x_0), y - x_0 \rangle$
$\geq f(x_0)$.

Remarques : * Cas particulier classique de cette méthode : si $A \in S_n^{++}(\mathbb{R})$ et $b \in \mathbb{R}^p$, on définit $f : \mathbb{R}^n \to \mathbb{R}$, $x \mapsto \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$. Alors $\nabla f(x) = Ax - b$.

Donc $\forall x, y$, $\langle \nabla f(x) - \nabla f(y), x - y\rangle = \langle A(x-y), x-y\rangle \geq \alpha \|x-y\|^2$ où $\alpha = \min(Sp(A)) > 0$

Ainsi f vérifie les conditions du théorème, et on peut approcher x^* le min de f via la méthode du gradient optimal, et x^* vérifie $\nabla f(x^*) = 0$ donc $Ax^* = b$! En plus, on peut calculer λ_k explicitement et facilement :

$\lambda_k = -\dfrac{\|\nabla f(x_k)\|^2}{\langle A \nabla f(x_k), \nabla f(x_k)\rangle}$. (et la vitesse de CV dépend alors de cond(A), plus c'est proche de 1, plus ça converge vite)

(proche de 1 ◎, éloigné de 1 ◎)

* Intérêt de la méthode : se ramener à des problèmes de minimisation de fonctions à 1 variable que l'on sait mieux gérer* pour optimiser la méthode du gradient à pas fixé (mais tout ça peut devenir très complexe numériquement...)

↳ condit° de CV : ∇f lipschitz$^{(L)}$, pas assez petit, $\alpha < \frac{2}{L}$

* méthode du triplet → de la sect° dorée, travail sur f', ou expression explicite

* Autre algo proche du gradient à pas optimal dans le cadre de la première remarque : le gradient à pas conjugué. On a vu que notre algorithme, on a $\forall k \in \mathbb{N}$, $\nabla f(x_k) \perp \nabla f(x_{k+1})$, ici on va chercher λ_k tel que $\forall k \in \mathbb{N}$, $\forall j \in [\![0, k-1]\!]$, $\nabla f(x_k) \perp \nabla f(x_j)$. De plus, x_k n'appartient plus à $x_{k-1} + \mathrm{Vect}(\nabla f(x_{k-1}))$, mais à $x_0 + \mathrm{Vect}(\nabla f(x_0), \dots \nabla f(x_{k-1}))$

La famille $(\nabla f(x_0), \dots \nabla f(x_k))$ est libre à chaque étape, en particulier, l'algo stationne au bout de n étapes, où elle devient égale à x^*.

* Exemple simple de fonction fortement convexe : la norme euclidienne au carré !

Si $f : \mathbb{R}^n \to \mathbb{R}$, $x \mapsto \|x\|^2$, alors $\nabla f(x) = 2x$ et $\langle \nabla f(y) - \nabla f(x), y - x\rangle = 2\|y-x\|^2$

→ fortement convexe avec $\alpha = 2$!