Méthode de gradient à pas optimal

Références : Hirriart-Urruty, *Optimisation et analyse convexe*, p 17-19 et p 53-56

Lemme (Inégalité de Kantorovitch)**.**

Soit $A \in \mathcal{S}_n^{++}$ et $\lambda_1 \geqslant ... \geqslant \lambda_n$ ses valeurs propres, alors pour tout $x \in \mathbb{R}^n$,

$$\|x\|^4 \leqslant (Ax,x)(A^{-1}x,x) \leqslant \frac{1}{4}\left(\sqrt{\frac{\lambda_1}{\lambda_n}} + \sqrt{\frac{\lambda_n}{\lambda_1}}\right)^2 \|x\|^4 .$$

Démonstration. Il suffit de démontrer l'inégalité pour $\|x\| = 1$. Puis comme $A \in \mathcal{S}_n^{++}$, il existe $P \in \mathrm{O}_n(\mathbb{R})$ telle que $A = {}^tP\Delta P$ avec $\Delta = \mathrm{Diag}(\lambda_1, ..., \lambda_n)$. Par le changement de variable $y = Px$, il suffit de démontrer

$$1 \leqslant (\Delta y, y)(\Delta^{-1}y, y) \leqslant \frac{1}{4}\left(\sqrt{\frac{\lambda_1}{\lambda_n}} + \sqrt{\frac{\lambda_n}{\lambda_1}}\right)^2 .$$

On fixe y. On note ensuite M_i le point de coordonnées $\left(\lambda_i, \frac{1}{\lambda_i}\right)$, et M le barycentre des M_i avec les coefficients y_i^2. Alors M a pour coordonnées $\big((\Delta y, y), (\Delta^{-1}y, y)\big)$.
Tous les M_i sont dans l'intersection de l'épigraphe de $x \mapsto \frac{1}{x}$ et du demi-plan inférieur donné par la droite liant M_1 et M_n, donc M est dans ce domaine.
On peut alors majorer et minorer l'ordonnée de M :

$$\frac{1}{(\Delta y, y)} \leqslant (\Delta^{-1}y, y) \leqslant -\frac{(\Delta y, y)}{\lambda_1\lambda_n} + \frac{1}{\lambda_1} + \frac{1}{\lambda_n}.$$

D'où il vient

$$1 \leqslant (\Delta y, y)(\Delta^{-1}y, y) \leqslant \frac{(\Delta y, y)(\lambda_1 + \lambda_n - (\Delta y, y))}{\lambda_1\lambda_n}.$$

En trouvant le maximum de $u \mapsto \frac{u(\lambda_1 + \lambda_n - u)}{\lambda_1\lambda_n}$, on a exactement l'inégalité voulue. □

Le but de ce développement est d'appliquer l'algorithme de gradient optimal à la fonction $f : x \in \mathbb{R}^n \mapsto \frac{1}{2}(Ax, x) + (b, x) + c$ où $A \in \mathcal{S}_n^{++}$, $b \in \mathbb{R}^n$ et $c \in \mathbb{R}$. Celle-ci est sympathique car l'algorithme pourra être calculé explicitement avec elle, ce qui n'est quasiment jamais le cas.
On gardera la notation $\lambda_1 \geqslant ... \geqslant \lambda_n$ pour les valeurs propres de A.

L'algorithme de gradient à pas optimal donne une méthode numérique permettant de minimiser des fonctions de $\mathbb{R}^n$. Il est défini comme suit :
On se donne $x_0 \in \mathbb{R}^n$, puis pour passer de l'étape k à l'étape $k+1$, on calcule $x_{k+1} = x_k + t_k d_k$, où $d_k = -\nabla f(x_k)$ et t_k est l'unique réel positif minimisant $t \mapsto f(x_k + td_k)$ (si $\nabla f(x_k) \neq 0$). On s'arrête lorsqu'on est assez proche du minimum, c'est à dire lorsque $\|\nabla f(x_k)\| < \epsilon$ pour une tolérance ϵ que l'on s'est fixé.

Théorème.

La fonction f définie auparavant a un unique minimum, noté $\overline{x}$. On appelle $\overline{f} = f(\overline{x})$ et $c(A) = \|A\|\left\|A^{-1}\right\| = \frac{\lambda_1}{\lambda_n}$ le conditionnement de A, alors la suite $(x_k)_k$ donnée par la méthode de gradient à pas optimal vérifie

$$f(x_k) - \overline{f} \leqslant (f(x_0) - \overline{f})\left(\frac{c(A)-1}{c(A)+1}\right)^{2k}$$

et

$$\|x_k - \overline{x}\| \leqslant \left(\frac{2(f(x_0) - \overline{f})}{\lambda_n}\right)^{\frac{1}{2}} \left(\frac{c(A)-1}{c(A)+1}\right)^{k}$$

Démonstration. • Existence et unicité du minimum :
f est coercive ($\lim_{x\to\infty} f(x) = \infty$), donc le minimum de f n'est pas à l'infini. En se restreignant à un compact assez grand, comme f est continue sur ce compact, elle atteint son minimum sur celui-ci. On a donc existence du minimum.
Puis comme $D^2 f(x) = A$ est définie positive, f est fortement convexe, donc le minimum est unique.
Le minimum est caractérisé par $\nabla f(\overline{x}) = 0$, soit $\overline{x} = -A^{-1}b$. D'où $\overline{f} = -\frac{1}{2}(A^{-1}b, b) + c$.

• Précisions sur d_k et t_k :
On suppose que pour les k étudiés $\nabla f(x_k)$ ne s'annule pas, sinon l'algorithme est terminé.
On a, pour $t \in \mathbb{R}$,

$$\begin{aligned} f(x_k + td_k) &= \frac{1}{2}(Ax_k, x_k) + \frac{t}{2}(Ax_k, d_k) + \frac{t}{2}(Ad_k, x_k) + \frac{t^2}{2}(Ad_k, d_k) + (b, x_k) + t(b, d_k) + c \\ &= f(x_k) + t(Ax_k + b, d_k) + \frac{t^2}{2}(Ad_k, d_k) \end{aligned}$$

Ce polynôme est minimisé lorsque $t = t_k := -\dfrac{(Ax_k + b, d_k)}{(Ad_k, d_k)} = \dfrac{\|d_k\|^2}{(Ad_k, d_k)}$ (car $d_k = -\nabla f(x_k) = -(Ax_k + b)$).[1]

• On peut maintenant calculer $f(x_{k+1})$:

$$\begin{aligned} f(x_{k+1}) = f(x_k + t_k d_k) &= f(x_k) + \frac{\|d_k\|^2}{(Ad_k, d_k)}(Ax_k + b, d_k) + \frac{1}{2}\frac{\|d_k\|^4}{(Ad_k, d_k)^2}(Ad_k, d_k) \\ &= f(x_k) - \frac{\|d_k\|^2}{(Ad_k, d_k)}\|d_k\|^2 + \frac{\|d_k\|^4}{2(Ad_k, d_k)} \\ &= f(x_k) - \frac{\|d_k\|^4}{2(Ad_k, d_k)}. \end{aligned}$$

Pour se ramener au lemme de Kantorovitch, on calcule $(A^{-1}d_k, d_k)$:

$$\begin{aligned} (A^{-1}d_k, d_k) &= (A^{-1}(Ax_k + b), Ax_k + b) \\ &= (Ax_k, x_k) + (b, x_k) + (Ax_k, A^{-1}b) + (A^{-1}b, b) \\ &= 2\left(\frac{1}{2}(Ax_k, x_k) + (b, x_k) + c + \frac{1}{2}(A^{-1}b, b) - c\right) \\ &= 2(f(x_k) - \overline{f}). \end{aligned}$$

On a donc

$$\begin{aligned} f(x_{k+1}) - \overline{f} &= (f(x_k) - \overline{f}) - \frac{\|d_k\|^4}{2(Ad_k, d_k)} \\ &= (f(x_k) - \overline{f})\left(1 - \frac{\|d_k\|^4}{2(f(x_k) - \overline{f})(Ad_k, d_k)}\right) \\ &= (f(x_k) - \overline{f})\left(1 - \frac{\|d_k\|^4}{(Ad_k, d_k)(A^{-1}d_k, d_k)}\right). \end{aligned}$$

• On applique Kantorovitch :
On a

$$(Ad_k, d_k)(A^{-1}d_k, d_k) \leqslant \frac{1}{4}\left(\sqrt{\frac{\lambda_1}{\lambda_n}} + \sqrt{\frac{\lambda_n}{\lambda_1}}\right)^2 \|d\|_k^4.$$

1. On rappelle que (Ad_k, d_k) est non nul par définition de la définie positivité de A.

Donc, comme $f(x_k) - \overline{f} \geqslant 0$, on a

$$\begin{aligned} f(x_{k+1}) - \overline{f} &= (f(x_k) - \overline{f}) \left(1 - \frac{\|d_k\|^4}{(Ad_k, d_k)(A^{-1}d_k, d_k)}\right) \\ &\leqslant (f(x_k) - \overline{f}) \left(1 - 4\left(\sqrt{\frac{\lambda_1}{\lambda_n}} + \sqrt{\frac{\lambda_n}{\lambda_1}}\right)^{-2}\right) \\ &\leqslant (f(x_k) - \overline{f}) \left(1 - 4\frac{1}{c(A) + c(A)^{-1} + 2}\right) \\ &\leqslant (f(x_k) - \overline{f}) \left(1 - 4\frac{c(A)}{(c(A)+1)^2}\right) \\ &\leqslant (f(x_k) - \overline{f}) \left(\frac{c(A)-1}{c(A)+1}\right)^2. \end{aligned}$$

On a ainsi la première inégalité voulue par récurrence.
Pour la deuxième, on la déduit de la précédente en remarquant que

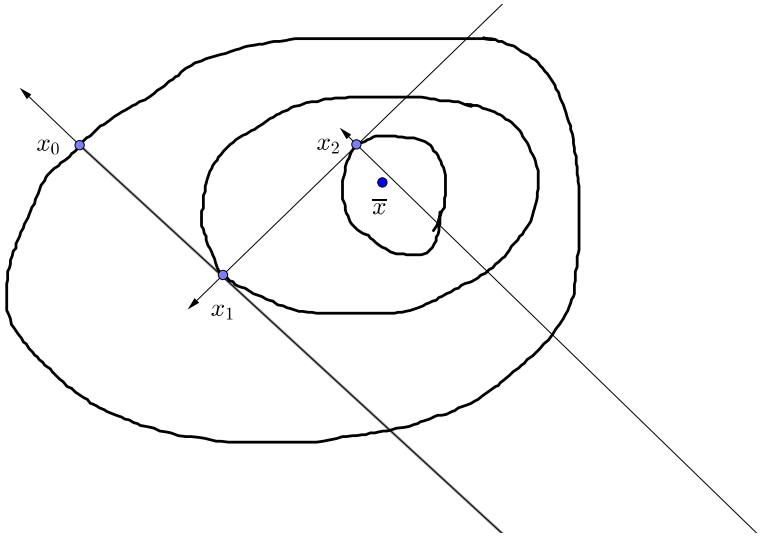
$$\begin{aligned} f(x_k) - \overline{f} &= \frac{1}{2}(Ax_k, x_k) + (b, x_k) + c + \frac{1}{2}(A^{-1}b, b) - c \\ &= \frac{1}{2}(Ax_k, x_k) - (A\overline{x}, x_k) + \frac{1}{2}(A\overline{x}, \overline{x}) \\ &= \frac{1}{2}(A(x_k - \overline{x}), (x_k - \overline{x})) \\ &\geqslant \frac{\lambda_n}{2} \|x_k - \overline{x}\|^2. \end{aligned}$$

□

On voit sur le dessin ci-dessous l'algorithme appliqué. On part de x_0, on cherche le gradient de f en ce point et on part chercher le prochain point sur la droite donné par $\nabla f(x_0)$. Les courbes tracées représentent certaines lignes de niveau de f.

On a $d_{k+1} = -Ax_{k+1} - b = -Ax_k - b - t_k Ad_k = d_k - t_k Ad_k$, donc $(d_{k+1}, d_k) = \|d_k\|^2 - t_k(Ad_k, d_k) = 0$. Les directions de descente sont donc consécutivement orthogonales entre elles.

D'autre part, x_{k+1} est le point d'une courbe de niveau tangente à $\nabla f(x_k)$ en x_{k+1}. En effet, si ce n'était pas le cas, la droite donnée par $t \mapsto x_k + td_k$ couperait la ligne de niveau contenant x_{k+1} en deux points distincts. Le segment reliant ceux-ci contiendrait des valeurs de f plus faibles que $f(x_{k+1})$ car f est fortement convexe ; cela est absurde. Par régularité de f, donc des courbes de niveau, la droite est tangente à la ligne de niveau en l'unique point d'intersection.



On voit que si $c(A)$ est proche de 1 - c'est à dire que l'on a un faible conditionnement/que les valeurs propres de A sont proches les unes des autres - alors l'algorithme converge rapidement. Sur le dessin, cela correspond au cas où les lignes de niveaux sont quasiment des cercles.
Au contraire, si le conditionnement est fort, la convergence est beaucoup plus lente. Cela correspond à des sortes d'ellipses très aplaties.
Trouver le minimum revient à inverser la matrice A, car $\overline{x} = -A^{-1}b$, et on se rend compte que plus la matrice est difficile à inverser, plus l'algorithme met du temps à converger.

Remarques : • En pratique, on admet que $(A^{-1}d_k, d_k) = 2(f(x_k) - \overline{f})$ et on insiste sur la convexité de f, l'existence et l'unicité du minimum et la dépendance au conditionnement.
• Si on présente la leçon convexité, on présente en détail le lemme de Kantorovitch et on va très vite sur le gradient à pas optimal. Sinon, on admet Kantorovitch.
• Dans le cas général, on peut montrer que si f est fortement convexe et $\mathcal{C}^1$, alors la suite donnée par cette algorithme converge vers le minimum.
• Cet algorithme est dit de descente. C'est un algorithme d'optimisation différentiable, destiné à minimiser une fonction réelle différentiable définie sur un espace euclidien ou, plus généralement, sur un espace hilbertien. Au point courant, un déplacement est effectué le long d'une direction de descente, de manière à faire décroître la fonction.
On peut en citer trois autres : celle de Newton-Raphson appliqué à ∇f, celle de gradient à pas constant (qui converge pour un pas assez petit si f est fortement convexe $\mathcal{C}^1$ et si ∇f est localement lipschitzien) et celle de gradient conjugué.
La méthode de gradient conjugué s'applique à la minimisation de fonctions de la forme $x \in \mathbb{R}^n \mapsto \frac{1}{2}(Ax, x) + (b, x)$. A chaque itération, au lieu de minimiser la fonction sur l'espace vectoriel $\mathrm{Vect}(\nabla f(x_k))$, on la minimise sur $\mathrm{Vect}(\nabla f(x_0), ..., \nabla f(x_k))$. L'algorithme converge donc exactement vers la solution en au plus n itérations.