Deux méthodes de gradient

Introduction générale.

Il s'agit de résoudre un système linéaire du type :

$$Ax = b \tag{1}$$

où, *a priori*, $A \in GL_n(\mathbf{R})$ et $b \in \mathbf{R}^n$. En fait, on prendra $A \in S_n^{++}(\mathbf{R})$. Un problème équivalent consiste à trouver le point qui minimise la fonctionnelle :

$$\Phi(y) = \frac{1}{2} y^T A y - y^T b.$$

En effet, il est facile de voir que

$$\nabla\Phi(y) = \frac{1}{2}(A^T + A)y - b = Ay - b. \tag{2}$$

Et si x est solution du système linéaire, alors

$$\Phi(y) = \Phi(x + (y - x)) = \Phi(x) + \frac{1}{2}(y - x)^T A(y - x), \quad \text{i.e} \quad \frac{1}{2}\|y - x\|_A^2 = \Phi(y) - \Phi(x)$$

où $\|z\|_A^2 = z^T A z$ est la norme d'énergie que l'on utilisera toujours par la suite. Une *méthode de gradient* consiste à partir d'un point $x_0 \in \mathbf{R}^n$ et à construire la suite

$$x_{k+1} = x_k + \alpha_k d_k \tag{3}$$

où $d_k \in \mathbf{R}^n$ est une direction à choisir et $\alpha_k \in \mathbf{R}$. Une idée naturelle est de choisir α_k de sorte à optimiser $\Phi(x_{k+1})$ dans la direction d_k, c'est à dire tel que :

$$\frac{d}{d\alpha_k}\Phi(x_k + \alpha_k d_k) = -d_k^T r_k + \alpha_k d_k^T A d_k = 0$$

où $-r_k := \nabla\Phi(x_k) = Ax_k - b$. On trouve :

$$\alpha_k = \frac{\langle d_k, r_k \rangle}{\|d_k\|_A^2} \tag{4}$$

(c'est bien défini lorsque $d_k \neq 0$ car $A \in S_n^{++}(\mathbf{R})$).

Théorème. *Soit x la solution du système* (1) *ou de façon équivalente, la solution du problème de minimisation* (2)*. Si α_k est choisi comme dans* (4)*, alors la suite* (3) *vérifie :*

$$\|x_{k+1} - x\|_A^2 = (1 - \sigma_k)\|x_k - x\|_A^2$$

où

$$\sigma_k = \frac{\langle d_k, r_k \rangle^2}{\|d_k\|_A^2 \|r_k\|_{A^{-1}}^2} \in (0, 1].$$

PREUVE. Il suffit de calculer :

$$\begin{aligned}\|x_{k+1}-x\|_A^2 &= \|x_k - x + \alpha_k d_k\|_A^2 \\ &= \|x_k-x\|_A^2 + \alpha_k^2\|d_k\|_A^2 + 2\alpha_k\langle d_k, A(x_k-x)\rangle \\ &= \|x_k-x\|_A^2 + \alpha_k^2\|d_k\|_A^2 - 2\alpha_k\langle d_k, r_k\rangle\end{aligned}$$

car $A(x_k-x) = Ax_k - b = -r_k$ et $\|x_k-x\|_A^2 = \|r_k\|_{A^{-1}}^2$. Et en remplaçant α_k par son expression :

$$\|x_{k+1}-x\|_A^2 = \left(1 - \frac{\langle d_k, r_k\rangle^2}{\|d_k\|_A^2\|r_k\|_{A^{-1}}^2}\right)\|x_k-x\|_A^2.$$

□

Méthode de gradient à pas optimal.

On choisit pour direction la "plus grande pente", autrement dit :

$$d_k = -\nabla\Phi(x_k) = -Ax_k + b = r_k.$$

Dans ce cas, $d_k \neq 0$ tant qu'on a pas atteint la solution et la convergence découle du théorème et de inégalité de Kantorovich[1] :

Lemme (Inégalité de Kantorovich)**.** *En notant $0 < \lambda_1 \leq \ldots \leq \lambda_n$ les valeurs propres de A, on a pour tout $y \in \mathbf{R}^n$,*

$$\frac{\|y\|^4}{\|y\|_A^2\|y\|_{A^{-1}}^2} \geq \frac{4\lambda_n\lambda_1}{(\lambda_n+\lambda_1)^2}.$$

PREUVE. On va montrer l'inégalité équivalente :

$$\forall y \in \mathbf{R}^n,\ \ \|y\|^4 \leq \frac{1}{4}\left(\sqrt{\frac{\lambda_n}{\lambda_1}} + \sqrt{\frac{\lambda_1}{\lambda_n}}\right)^2.$$

On va même supposer que $\|y\| = 1$ et commencer par remarquer :

$$1 = \|y\|^2 = \langle y, AA^{-1}y\rangle \leq \|y\|_A\|A^{-1}y\|_A = \|y\|_A\|y\|_{A^{-1}}.$$

Et dans une base orthonormale de vecteurs propres :

$$\begin{aligned}\|y\|_A\|y\|_{A^{-1}} &= \sqrt{\left(\sum_{i=1}^n \lambda_i y_i^2\right)\left(\sum_{i=1}^n \frac{1}{\lambda_i}y_i^2\right)} = \sqrt{\frac{\lambda_1}{\lambda_n}\left(\sum_{i=1}^n \frac{\lambda_i}{\lambda_1}y_i^2\right)\left(\sum_{i=1}^n \frac{\lambda_n}{\lambda_i}y_i^2\right)} \\ &\leq \frac{1}{2}\sqrt{\frac{\lambda_1}{\lambda_n}}\left(\left(\sum_{i=1}^n \frac{\lambda_i}{\lambda_1}y_i^2\right) + \left(\sum_{i=1}^n \frac{\lambda_n}{\lambda_i}y_i^2\right)\right) \\ &\leq \frac{1}{2}\sqrt{\frac{\lambda_1}{\lambda_n}}\left(\sum_{i=1}^n \left(\frac{\lambda_i}{\lambda_1} + \frac{\lambda_n}{\lambda_i}\right)y_i^2\right)\end{aligned}$$

[1] Comme c'est une inégalité de convexité, on peut la développer dans les leçons qui leur sont dévolues mais en fait, on n'en a pas besoin pour conclure : on peut obtenir une majoration de l'erreur (un peu différente mais pas pire) beaucoup plus rapidement et beaucoup plus simplement comme le fait P. D. Lax dans son *Linear Algebra* (dans la deuxième édition) !

La fonction $x \mapsto \frac{x}{\lambda_1} + \frac{\lambda_n}{x}$ admet un maximum en λ_1 ou en λ_n et il vaut dans les deux cas : $1 + \frac{\lambda_n}{\lambda_1}$. Ainsi,

$$\|y\|_A \|y\|_{A^{-1}} \leq \frac{1}{2}\sqrt{\frac{\lambda_1}{\lambda_n}} \left(\sum_{i=1}^{n} \left(1 + \frac{\lambda_n}{\lambda_1}\right) y_i^2 \right) \leq \frac{1}{2} \left(\sqrt{\frac{\lambda_n}{\lambda_1}} + \sqrt{\frac{\lambda_1}{\lambda_n}} \right)$$

et le résultat suit en élevant au carré. □

Et sachant que $\mathrm{cond}(A) = \lambda_n/\lambda_1$, on obtient le :

Théorème. *Avec les choix précédents et $d_k = r_k$, la suite* (3) *converge vers x avec :*

$$\|x_k - x\|_A \leq \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} \|x_k - x\|_A.$$

Un calcul supplémentaire donne :

$$\|x_k - x\| \leq \sqrt{\mathrm{cond}(A)} \left(\frac{\mathrm{cond}(A) - 1}{\mathrm{cond}(A) + 1} \right)^k \|x_0 - x\|.$$

PREUVE. La première inégalité découle directement de l'inégalité de Kantorovich. Pour la seconde, il s'agit de voir que pour tout $y \in \mathbf{R}^n$,

$$\lambda_1 \|y\|^2 \leq \|y\|_A^2 \leq \lambda_n \|y\|^2.$$

□

De la dernière inégalité, on voit que la convergente peut être lente lorsque la matrice est mal conditionnée.

Méthode de gradient conjugué.

Remarquons que pour tout $k \in \mathbf{N}$:

$$r_{k+1} = r_k - \alpha_k A d_k \tag{5}$$

et α_k est choisi de sorte à ce que

$$\langle r_{k+1}, d_k \rangle = 0. \tag{6}$$

Idée. Construire des directions (d_k) deux à deux A-orthogonales, comme ça r_{k+1} sera orthogonal à $\mathrm{Vect}(d_0, \ldots, d_k)$.

Notations. Pour $x, y \in \mathbf{R}^n$, on note $x \perp y$ lorsque x et y sont orthogonaux pour le produit scalaire euclidien et $x \perp_A y$ lorsque x et y sont orthogonaux pour le produit scalaire donné par A. On étend naturellement cette notation à des sous-espaces de $\mathbf{R}^n$.

On pose $d_0 = r_0$ et pour $k \in \mathbf{N}$, on construit d_{k+1} comme l'orthogonalisé de Gram-Schmidt pour le produit scalaire donné par A de r_{k+1} relativement à $\mathrm{Vect}(d_k)$:

$$d_{k+1} = r_{k+1} - \beta_k d_k \tag{7}$$

où

$$\beta_k = \frac{\langle r_{k+1}, A d_k \rangle}{\|d_k\|_A^2} \quad \text{si } d_k \neq 0, \quad \beta_k = 0 \text{ sinon.} \tag{8}$$

Remarquons que si $d_k = 0$ alors r_k et d_{k-1} sont colinéaires et comme ils sont aussi orthogonaux par (6), $r_k = 0$.

Lemme. *Avec le choix* (8)*, les directions* (7) *vérifient pour tout* $k \in \mathbf{N}$ *la propriété suivante : si* $r_0, \dots, r_k$ *ne sont pas nuls alors,*

(i) $\mathrm{Vect}(r_0, \dots, r_k) = \mathrm{Vect}(d_0, \dots, d_k)$

(ii) $r_{k+1} \perp \mathrm{Vect}(d_0, \dots, d_k)$

(iii) $d_{k+1} \perp_A \mathrm{Vect}(d_0, \dots, d_k)$

Preuve. On procède par récurrence sur $k \in \mathbf{N}$. Lorsque $k = 0$, (i), (ii) et (iii) sont vrais grâce aux relations $r_0 = d_0$, (6) et (7) et bien sûr $r_0 \neq 0$ sinon il n'y a rien à faire. Supposons donc le résultat vrai au rang $k-1$, $k \in \mathbf{N}^*$.

(i) Par (7), on a : $d_k = r_k - \beta_{k-1} d_{k-1}$.

(ii) Par (6), on a déjà $r_{k+1} \perp d_k$ et si $j \in \{0, \dots, k-1\}$, la relation (5) couplée à l'hypothèse de récurrence (ii) et (iii) donne $r_{k+1} \perp d_j$.

(iii) Par (7), on a déjà $d_{k+1} \perp_A d_k$ (c'est la définition) et si $j \in \{0, \dots, k-1\}$, la relation (7) couplée à l'hypothèse de récurrence (iii) donne :

$$\langle d_{k+1}, Ad_j \rangle = \langle r_{k+1}, Ad_j \rangle.$$

Montrons que $Ad_j \in \mathrm{Vect}(r_0, \dots, r_k)$, ce qui conclura grâce aux relations (i) et (ii) que l'on vient de prouver. Grâce à la relation (5) avec $k = j$, il suffit de montrer que $\alpha_j \neq 0$, ce qui est la cas car :

$$\alpha_j = 0 \overset{(4)}{\iff} \langle r_j, d_j \rangle = 0 \overset{(7)}{\iff} r_j = 0$$

et on a justement supposé le contraire.

□

Théorème. *La méthode de gradient associée aux directions* (7) *avec le choix* (8) *converge vers la solution* x *du problème* (1) *en au plus* n *itérations.*

Preuve. Les conditions (i) et (ii) du lemme précédent assurent que la famille $(r_k)_k$ est une famille orthogonale donc libre. On est en dimension n. □

Référence. A. Quarteroni, R. Sacco, F. Saleri, *Numerical Mathematics*

158 Matrices symétriques réelles, matrices hermitiennes.
162 Systèmes d'équations linéaires ; opérations élémentaires, aspects algorithmiques et conséquences théoriques.
219 Extrema : existence, caractérisation, recherche. Exemples et applications.
226 Suites vectorielles et réelles définies par une relation de récurrence $u_{n+1} = f(u_n)$. Exemples. Applications à la résolution approchée d'équations.
233 Méthodes itératives en analyse numérique matricielle.