Minimisation des fonctions quadratiques

1. Introduction (systèmes linéaires, moindres carrés)
2. Méthodes de descente à pas optimal : cadre général
3. Gradient à pas optimal
4. Gradient conjugué

Introduction

La minimisation de fonctions quadratiques apparaît notamment dans le contexte de la méthode des moindres carrés.

Exemple : modèle approché pour la relation force-extension (f-δ) d'un ressort non linéaire

$$f = x_1 \, \delta + x_2 \, \delta^2$$

$x = (x_1, x_2)^T \in \mathbb{R}^2$ est à déterminer à partir de mesures (f_i, δ_i) $(i = 1, 2, \ldots, m)$. On cherche x qui minimise $\|M \, x - g\|_2^2$ avec :

$$M = \begin{pmatrix} \delta_1 & \delta_1^2 \\ \delta_2 & \delta_2^2 \\ \vdots & \vdots \\ \delta_m & \delta_m^2 \end{pmatrix} \in M_{m,2}(\mathbb{R}) \quad , \quad g = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{pmatrix} \in \mathbb{R}^m$$

Introduction

- Cas général : soient $M \in M_{m,n}(\mathbb{R})$ avec $m \geqslant n$ et $g \in \mathbb{R}^m$.
- On suppose $\mathrm{Ker}\, M = \{0\}$ (ou de manière équivalente $\dim(\mathrm{Im}\, M) = n$).
- Si $m > n = 2$, le problème $M\,x = g$ n'a pas de solution $x \in \mathbb{R}^n$ en général (i.e. lorsque $g \notin \mathrm{Im}\, M$).
- Solution au sens des moindres carrés : on cherche $x \in \mathbb{R}^n$ tel que $\|M\,x - g\|_2^2 = \min_{y \in \mathbb{R}^n} \|M\,y - g\|_2^2$
- Ce problème admet une solution x unique qui vérifie les *équations normales* $A\,x = b$, avec :

$$A = M^T\,M \ \in M_n(\mathbb{R}) \text{ sym. déf. positive}, \qquad b = M^T\,g \ \in \mathbb{R}^n,$$

car $\forall\, y \in \mathbb{R}^n$, $\|M\,y - g\|_2^2 = \|M\,(y-x)\|_2^2 - \|M\,x\|_2^2 + \|g\|_2^2$

Introduction

- Minimiser $\|Mx - g\|_2^2$ équivaut à minimiser la fonction quadratique :
$$f(x) = \frac{1}{2} x^T A x - b^T x$$
où $A = M^T M$ sym. déf. positive, $b = M^T g$

$$\begin{aligned}
\frac{1}{2}\left(\|Mx - g\|_2^2 - \|g\|_2^2\right) &= \frac{1}{2}\left((Mx-g)^T (Mx-g) - \|g\|_2^2\right) \\
&= \frac{1}{2}\left((Mx)^T (Mx) - 2 g^T M x\right) \\
&= \frac{1}{2} x^T M^T M x - (M^T g)^T x
\end{aligned}$$

- En particulier, résoudre $Mx = g$ lorsque $M \in M_n(\mathbb{R})$ est inversible revient à minimiser f

Introduction

- De même, pour A matrice symétrique définie positive arbitraire, résoudre $Ax = b$ équivaut à minimiser la fonction quadratique :

$$f(x) = \frac{1}{2} x^T A x - b^T x$$

car en notant $\|y\|_A = (y^T A y)^{1/2}$ et $x = A^{-1} b$, il vient $\forall y \in \mathbb{R}^n$:

$$\begin{aligned}
\frac{1}{2}\left(\|y-x\|_A^2 - \|x\|_A^2\right) &= \frac{1}{2}\left((y-x)^T A (y-x) - x^T A x\right) \\
&= \frac{1}{2}\left(y^T A y - 2y^T A x\right) \\
&= \frac{1}{2} y^T A y - b^T y
\end{aligned}$$

- La minimisation de f fournit donc une approche supplémentaire pour la résolution des systèmes linéaires

Méthodes de descente à pas optimal

- On cherche $x \in \mathbb{R}^n$ minimisant
$$f(x) = \frac{1}{2} x^T A x - b^T x$$
où $A \in M_n(\mathbb{R})$ est symétrique définie positive et $b \in \mathbb{R}^n$.
- Minimum atteint en la solution unique du système $Ax = b$.
- une méthode de descente à pas optimal est une méthode itérative de la forme :
$$x_{k+1} = x_k - \rho_k w_k$$
où le choix des directions de descente $(w_k)_{k \geqslant 0}$ caractérise la méthode et le pas optimal ρ_k vérifie
$$f(x_k - \rho_k w_k) = \min_{\rho \in \mathbb{R}} f(x_k - \rho w_k).$$
- Nous allons calculer ρ_k (calcul explicite possible pour f quadratique)

Méthodes de descente à pas optimal

- Si $w_k \neq 0$, la restriction de f à la droite $x_k + \mathbb{R} w_k$:

$$\varphi(\rho) = f(x_k - \rho w_k)$$

est un polynôme de degré 2 avec $\lim_{\rho \to \pm\infty} \varphi(\rho) = +\infty$:

$$\varphi(\rho) = \frac{1}{2} \langle x_k - \rho w_k, x_k - \rho w_k \rangle_A - b^T (x_k - \rho w_k)$$

où on note $\langle y, z \rangle_A = y^T A z$.

- $\rho = \rho_k$ est donc l'unique solution de $\varphi'(\rho){=}0$, où

$$\begin{aligned} \varphi'(\rho) &= \langle x_k - \rho w_k, -w_k \rangle_A + b^T w_k \\ &= \rho \langle w_k, w_k \rangle_A - \langle x_k, w_k \rangle_A + b^T w_k \\ &= \rho \langle w_k, w_k \rangle_A - (A x_k)^T w_k + b^T w_k \\ &= \rho \langle w_k, w_k \rangle_A + (b - A x_k)^T w_k \end{aligned}$$

Méthodes de descente à pas optimal

En notant $r_k = A\,x_k - b$ le résidu, on obtient donc :

$$\varphi'(\rho) = \rho\,\langle w_k, w_k\rangle_A - r_k^T\, w_k$$

Donc la condition d'optimalité $\varphi'(\rho_k)$=0 donne

$$\rho_k = \frac{r_k^T\, w_k}{w_k^T\, A\, w_k}$$

La méthode de descente à pas optimal s'écrit donc à chaque itération :

$$x_{k+1} = x_k - \rho_k w_k, \quad \rho_k = \frac{(A\,x_k - b)^T\, w_k}{w_k^T\, A\, w_k}.$$

Nous détaillerons plus loin deux choix de directions de descente $(w_k)_{k\geqslant 0}$, correspondant aux méthodes du gradient à pas optimal et du gradient conjugué.

Méthodes de descente à pas optimal

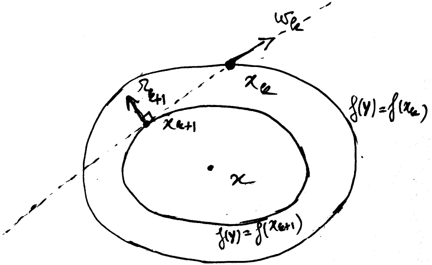
Interprétation géométrique du pas optimal :

nous allons montrer qu'en $y = x_{k+1}$, la droite $x_k + \mathbb{R}\, w_k$ est tangente à l'ellipsoide :

$$E = \{\, y \in \mathbb{R}^n, f(y) = f(x_{k+1}) \,\}$$

où on rappelle que

$$f(y) = \frac{1}{2}\left(\|y - x\|_A^2 - \|x\|_A^2 \right), \quad A\,x = b.$$



Méthodes de descente à pas optimal

- On a $r_{k+1}^T w_k = 0$ (avec $r_{k+1} = A x_{k+1} - b$) car

$$
\begin{aligned}
\varphi'(\rho_k) &= \langle x_k - \rho_k w_k, -w_k \rangle_A + b^T w_k \\
&= -\langle x_{k+1}, w_k \rangle_A + b^T w_k \\
&= -(A x_{k+1} - b)^T w_k = 0
\end{aligned}
$$

- Puisque $f(y) = \frac{1}{2} y^T A y - b^T y$ avec $A \in M_n(\mathbb{R})$ symétrique :

$$\nabla f(y) = A y - b$$

  Rappel de calcul différentiel : $\nabla f(x_{k+1}) = r_{k+1}$ est orthogonal à l'ensemble de niveau $E = \{ y \in \mathbb{R}^n, f(y) = f(x_{k+1}) \}$ en $y = x_{k+1}$, avec ∇f orienté dans le sens des valeurs de f croissantes

- Puisque $\nabla f(x_{k+1})^T w_k = 0$, la droite $x_k + \mathbb{R} w_k$ est donc tangente en x_{k+1} à l'ellipsoide E

Méthode du gradient à pas optimal

ou *Méthode de la plus grande pente*

- On choisit $w_k = \nabla f(x_k) = A x_k - b = r_k$
- Motivation : $\varphi(\rho) = f(x_k - \rho w_k)$ vérifie

$$\varphi'(\rho) = -\nabla f(x_k - \rho w_k)^T \, w_k$$

$$\varphi'(0) = -r_k^T \, w_k$$

En fixant $w_k = r_k$ on a donc :

$$\varphi'(0) = -\|r_k\|_2^2 \leqslant 0$$

De plus, le choix $w_k = r_k$ maximise $|\varphi'(0)|$ (minimise $\varphi'(0) \leqslant 0$) sur les vecteurs w_k de même norme que r_k :

$$|\varphi'(0)| = |r_k^T \, w_k| \leqslant \|r_k\|_2 \, \|w_k\|_2 = \|r_k\|_2^2$$

Méthode du gradient à pas optimal

On rappelle l'expression du pas optimal dans la direction $-w_k$

$$\rho_k = \frac{r_k^T\, w_k}{w_k^T\, A\, w_k}.$$

La méthode du gradient à pas optimal s'écrit donc à chaque itération $x_k \to x_{k+1}$:

$$r_k = A\, x_k - b$$

$$\rho_k = \frac{\|r_k\|_2^2}{r_k^T\, A\, r_k}$$

$$x_{k+1} = x_k - \rho_k r_k$$

Méthode du gradient à pas optimal

On étudie maintenant la convergence de la méthode.

- $\kappa_2 = \lambda_{\max}/\lambda_{\min}$ désigne le conditionnement euclidien de A sym. déf. positive ($\lambda_{\max} = \max \mathrm{Sp}(A)$, $\lambda_{\min} = \min \mathrm{Sp}(A)$)
- x désigne la solution du système $A\,x = b$, ou de manière équivalente le minimum de $f(y) = \frac{1}{2}\, y^T A\, y - b^T y$.

Théorème

Pour la méthode du gradient à pas optimal on a

$$\|x_k - x\|_A \leqslant \left(\frac{\kappa_2 - 1}{\kappa_2 + 1}\right)^k \|x_0 - x\|_A$$

- si $\kappa_2 = 1$ (i.e. $A = \lambda\, I$ avec $\lambda > 0$), convergence en une itération (ensembles de niveau E = sphères centrées en x).
- convergence lente si $\kappa_2 \gg 1$ (ellipsoides E aplatis).

Méthode du gradient conjugué

Idée : choisir w_k combinaison linéaire des vecteurs orthogonaux $r_k = \nabla f(x_k)$ et w_{k-1} pour améliorer la direction de descente.

- Initialisation : on se donne $x_0 \in \mathbb{R}^n$ et on fixe $w_0 = r_0$
- Pour tout $k \geqslant 1$:

$$w_k = r_k + \theta_k \, w_{k-1} \quad , \quad \theta_k = \frac{\|r_k\|_2^2}{\|r_{k-1}\|_2^2}$$

Le pas optimal dans la direction $-w_k$ vaut :

$$\rho_k = \frac{r_k^T \, w_k}{w_k^T \, A \, w_k} = \frac{\|r_k\|_2^2}{w_k^T \, A \, w_k}$$

(puisque $r_k^T \, w_{k-1} = 0$) et l'itération $x_k \to x_{k+1}$ s'écrit :

$$x_{k+1} = x_k - \rho_k w_k$$

Méthode du gradient conjugué

- On peut montrer que le choix $\theta = \theta_k$ dans $w_k = r_k + \theta_k\, w_{k-1}$ s'écrit aussi $\theta_k = -\langle r_k, w_{k-1}\rangle_A / \| w_{k-1} \|_A^2$, de sorte que

$$w_k^T\, A w_{k-1} = 0 \qquad (1)$$

w_k et w_{k-1} sont dits A-conjugués (orthogonaux pour le produit scalaire $\langle\, , \,\rangle_A$). On peut montrer que ce choix réalise :

$$\min_{\theta\in\mathbb{R}} \min_{\rho\in\mathbb{R}} f(x_k - \rho(r_k + \theta w_{k-1}))$$

- Puisque $r_k = A\, x_k - b = A\,(x_k - x)$ et $r_k^T\, w_{k-1} = 0$ on a donc $(x_k - x)^T\, A w_{k-1} = 0$
$\Rightarrow x_k - x$ (direction de descente idéale à l'itération $k+1$) est orthogonal à w_{k-1} pour le produit scalaire $\langle\, , \,\rangle_A$, ce qui donne une justification supplémentaire au choix (1) pour w_k

Méthode du gradient conjugué

On peut reformuler la méthode de manière à effectuer un seul produit matrice-vecteur $A\,w_k$ par itération, puisque $r_{k+1} = A\,x_{k+1} - b = A\,(x_k - \rho_k w_k) - b = r_k - \rho_k\,A\,w_k$

Initialisation $(k = 0)$: $x_0 \in \mathbb{R}^n$ donné, $w_0 = r_0 = A\,x_0 - b$

Tant que $\|r_k\|_2 >$ tolérance, faire :

$$\rho_k = \frac{\|r_k\|_2^2}{w_k^T\,A\,w_k}$$

$$x_{k+1} = x_k - \rho_k w_k$$

$$r_{k+1} = r_k - \rho_k\,A\,w_k$$

$$w_{k+1} = r_{k+1} + \frac{\|r_{k+1}\|_2^2}{\|r_k\|_2^2}\,w_k$$

$$k \leftarrow k+1$$

Méthode du gradient conjugué

On étudie maintenant la convergence de la méthode.

x désigne la solution du système $A\,x = b$, ou de manière équivalente le minimum de $f(y) = \frac{1}{2}\,y^T\,A\,y - b^T\,y$.

Théorème

Il existe $p \leqslant n$ tel que $x_p = x$.

- Preuve pour $n = 2$: $(x_1 - x)^T\,A\,w_0 = 0$ et $w_1^T\,A\,w_0 = 0$ impliquent w_1 colinéaire à $x_1 - x$, d'où $x_2 = x_1 - \rho_1\,w_1 = x$.
- La convergence en au plus n itérations est liée à un résultat d'orthogonalité des directions $w_0, w_1, \ldots, w_{n-1}$ pour $\langle\,,\,\rangle_A$ et au fait que $r_n^T\,w_i = 0\,\forall i$, qui donne $r_n = 0$ et $x_n = x$
- Ces propriétés sont vraies en arithmétique exacte, mais ce n'est plus le cas en arithmétique flottante. La méthode du gradient conjugué est donc utilisée en pratique comme une méthode itérative.

Méthode du gradient conjugué

Le résultat suivant est utile en pratique pour estimer la vitesse de convergence de la méthode suivant le conditionnement euclidien κ_2 de A :

Théorème

Pour la méthode du gradient conjugué on a

$$\|x_k - x\|_A \leqslant 2 \left(\frac{\sqrt{\kappa_2} - 1}{\sqrt{\kappa_2} + 1} \right)^k \|x_0 - x\|_A$$

Cette estimation d'erreur est meilleure que celle obtenue pour le gradient à pas optimal car on a :

$$\frac{\sqrt{\kappa_2} - 1}{\sqrt{\kappa_2} + 1} \leqslant \frac{\kappa_2 - 1}{\kappa_2 + 1}$$

puisque $1 \leqslant \sqrt{\kappa_2} \leqslant \kappa_2$

Méthode du gradient conjugué

On étudie maintenant pour $n \gg 1$ les ordres de grandeur du nombre d'itérations et du nombre d'opérations arithmétiques à effectuer pour résoudre $A\,x = b$.

On a $\|x_k - x\|_A \leqslant C \left(\frac{\sqrt{\kappa_2}-1}{\sqrt{\kappa_2}+1}\right)^k$ avec $C = 2\,\|x_0 - x\|_A$.

Le taux de décroissance de l'erreur vérifie :

$$\frac{\sqrt{\kappa_2}-1}{\sqrt{\kappa_2}+1} = 1 - \frac{2}{\sqrt{\kappa_2}+1} \leqslant e^{-\frac{2}{\sqrt{\kappa_2}+1}}$$

(par convexité de l'exponentielle), et puisque $\sqrt{\kappa_2} \geqslant 1$ on a :

$$\frac{\sqrt{\kappa_2}-1}{\sqrt{\kappa_2}+1} \leqslant e^{-\frac{1}{\sqrt{\kappa_2}}}, \quad \left(\frac{\sqrt{\kappa_2}-1}{\sqrt{\kappa_2}+1}\right)^k \leqslant e^{-\frac{k}{\sqrt{\kappa_2}}}$$

Méthode du gradient conjugué

Etant donné une tolérance d'erreur $\epsilon < 1$, pour avoir

$$\left(\frac{\sqrt{\kappa_2} - 1}{\sqrt{\kappa_2} + 1}\right)^k \leqslant \epsilon$$

il suffit donc de choisir :

$$k \geqslant \sqrt{\kappa_2}\,(-\ln \epsilon)$$

- Pour des matrices $A \in S_n^{++}(\mathbb{R})$ telles que $\kappa_2 \sim M\,n^2$, le nombre d'itérations est donc $O(n)$.
- Lorsque $\kappa_2 \ll n^2$, le nombre d'itérations est $\ll n$.
- Exemple : schéma aux différences finies pour l'équation de Poisson $-\Delta\, u = f$ dans $\Omega = (]0,1[)^d$ avec $u = 0$ sur $\partial\Omega$ $\Rightarrow n \sim h^{-d}$, où h est le pas de discrétisation, $\kappa_2 \sim (4/\pi^2)\, n^{2/d} = O(h^{-2})$ et nombre d'itérations $O(n^{1/d})$

Méthode du gradient conjugué

On prend maintenant en compte le nombre d'opérations arithmétiques élémentaires à chaque itération :

- Pour une matrice A pleine, le nombre d'opérations à chaque itération est $\sim 2\,n^2$ (avec $A\,w_k$). Par exemple si on réalise n itérations, le coût de la résolution de $A\,x = b$ est $\sim 2\,n^3$, plus coûteux que Cholesky $\sim (1/3)\,n^3$.
- Si la matrice A est creuse, avec au plus c coefficients non nuls par ligne (c fixé), le nombre d'opérations à chaque itération est $2\,c\,n + O(n) = O(n)$. Le coût de la résolution de $A\,x = b$ est donc $O(n\sqrt{\kappa_2})$ ($O(n^2)$ si le nombre d'itérations est $O(n)$)
- Exemple : différences finies pour $-\Delta\,u = f$ dans $\Omega = (]0,1[)^d$ avec $u = 0$ sur $\partial\Omega$
  $\Rightarrow c = 2d + 1$, $\kappa_2 = O(n^{2/d})$, le calcul de u coûte donc $O(n^{1+1/d})$ opérations (i.e. $O((1/h)^{d+1})$).