Méthodes numériques de base

Cours de première année - ENSIMAG

Cours : Hahmann S.
James G.
LaTeX: Poupin P.

Table des matières

Chapitre 1

Approximation de problèmes aux limites par différences finies

1.1 Introduction

On considère ici des problèmes aux limites en une dimension, linéaires, du second ordre :

$$\left\{ \begin{array}{ll} -u''(x) + p(x)u'(x) + q(x)u(x) = f(x) & x \in]a,b[\quad (1.1a) \\ u(a) = \alpha \\ u(b) = \beta \end{array} \right\} \quad (1.1b)$$

où $p, q, f \in \mathcal{C}^0([a,b])$ et $\alpha, \beta \in R$. On a une équation différentielle linéaire du 2^{nd} ordre, à coefficients a priori variables, augmentée des conditions aux limites (1.1b). Ces conditions fixent la valeur de u au bord du domaine : on parle de **conditions de Dirichlet**.

Lorsque les conditions aux limites fixent la valeur de u' au bord ($u'(a) = \alpha, u'(b) = \beta$), on parle de conditions de **Neumann**.

Exemples issus de la physique

a)
$$\left\{ \begin{array}{l} \dfrac{d^2\phi}{dr^2} + \dfrac{1}{r}\dfrac{d\phi}{dr} = \lambda\phi \quad , \quad r \in]r_a, R[\\ \phi'(r_a) = \alpha \\ \phi'(R) = 0 \end{array} \right.$$

Equations de Debye-Hückel donnant le potentiel électrique ϕ autour d'un cylindre chargé (de rayon r_a) polongé dans une solution ionique ($r_a < r < R$).

b)
$$\left\{ \begin{array}{l} -\dfrac{d}{dx}\left(k(x)\dfrac{du}{dx}\right) = f(x) \quad , \quad x \in]0, L[\\ u(0) = 0 \\ u(L) = 0 \end{array} \right.$$

Equation de la chaleur stationnaire donnant la température u dans une barre de longueur L chauffée et de conductivité $k(x)$ variable. ($f(x)$ = quantité de chaleur fournie par unité de temps et de longueur). La température aux extrémités de la barre est maintenue à 0.

Lorsque $q \geq 0$, le résultat suivant assure que (1.1a) - (1.1b) possède une solution unique. Mais on n'a pas en général d'expression explicite de u quand p ou q dépendent de x.

Théorème 1 *Si* $q(x) \geq 0$ *sur* $[a,b]$ *alors* (1.1a) - (1.1b) ***admet une unique solution*** $u \in \mathcal{C}^2([a,b])$.

De plus, si $p, q, f \in \mathcal{C}^k([a,b])$ *alors* $\boxed{u \in \mathcal{C}^{k+2}([a,b])}$.

Preuve 1 (de l'unicité) Si u_1 et u_2 vérifient (1.1a) - (1.1b), alors $v = u_1 - u_2$ est solution du problème homogène :

$$\begin{cases} -v'' + pv' + qv = 0 \\ v(a) = v(b) = 0 \end{cases}$$

Notons $r(x) = e^{-\int p\,dx}$ et $s(x) = q(x)e^{-\int p\,dx}$. En multipliant l'équation par r on a :

$$-\big(r(x)v'(x)\big)' + s(x)v(x) = 0$$

On multiplie maintenant cette équation par v et on intègre sur $[a,b]$, en utilisant les conditions aux limites :

$$\int_a^b r(x)v'^2\,dx + \int_a^b s(x)v^2\,dx = 0$$

avec $r > 0$ et $s \geq 0$

Donc $\int_a^b r(x)v'^2\,dx = 0 \implies v' = 0$ (puisque $r > 0$) $\implies v = 0$ à cause des conditions aux limites.

L'existence d'une solution $u \in \mathcal{C}^2([0,1])$ peut être obtenue de différentes façons (méthode de variation de la constante, méthodes d'analyse fonctionnelle, par exemple formulation variationnelle). La régularité $\mathcal{C}^k$ de u s'obtient simplement par récurrence sur k.

1.2 Discrétisation par différences finies

Nous allons étudier comment calculer u numériquement.

$$(p)\ \left\{ \begin{array}{c} -u'' + p(x)u' + q(x)u = f(x) \\ u(a) = \alpha \\ u(b) = \beta \end{array} \right. \quad \left| \begin{array}{c} x \in]a,b[\\ p, q, f \in \mathcal{C}^0([a,b]) \\ q \geq 0 \text{ sur } [a,b] \end{array} \right.$$

On discrétise $[a,b]$ suivant les points $x_i = a+ih$ $(i=0,\ldots,N+1)$ avec $h=\frac{b-a}{N+1}$. On notera u_i la valeur approchée de $u(x_i)$ à calculer et $U = {}^t(u_1,\ldots,u_N)$, $(u_0=\alpha, u_{N+1}=\beta)$.

On approche $u''(x_i)$ et $u'(x_i)$ en utilisant un développement de Taylor de u en x_i.

Si $u \in \mathcal{C}^4([a,b])$:

$$u(x_{i+1}) = u(x_i+h) = u(x_i) + h\,u'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{6}u^{(3)}(x_i) + \frac{h^4}{24}u^{(4)}(\theta_i^+)$$

$$\text{avec } \theta_i^+ \in]x_i, x_{i+1}[$$

$$u(x_{i-1}) = u(x_i+h) = u(x_i) - h\,u'(x_i) + \frac{h^2}{2}u''(x_i) - \frac{h^3}{6}u^{(3)}(x_i) + \frac{h^4}{24}u^{(4)}(\theta_i^-)$$

$$\text{avec } \theta_i^- \in]x_{i-1}, x_i[$$

Donc :

$$\frac{u(x_{i+1}) - u(x_{i-1})}{2h} = u'(x_i) + \mathcal{O}(h^2)$$

$$\frac{u(x_{i+1}) - u(x_i) + u(x_{i-1})}{h^2} = u''(x_i) + \mathcal{O}(h^2)$$

Notons $p_i = p(x_i)$, $q_i = q(x_i)$, $f_i = f(x_i)$. On a donc :

$$\frac{2u(x_i) - u(x_{i+1}) - u(x_{i-1})}{h^2} + p_i\frac{u(x_{i+1}) - u(x_{i-1})}{2h} + q_i u(x_i) - f_i = e_i$$

$$\text{pour tout } i = 1,\ldots,N, \text{ avec } e_i = \mathcal{O}(h^2)$$

L'approximation du problème (p) par différences finies consiste à résoudre :

$$(S)\ \begin{cases} \dfrac{2u_i - u_{i+1} - u_{i-1}}{h^2} + p_i\dfrac{u_{i+1} - u_{i-1}}{2h} + q_i u_i - f_i = 0 & i = 1,\ldots,N \\ \\ u_0 = \alpha,\ u_{N+1} = \beta \end{cases}$$

La solution de (p) vérifie donc (S) à une erreur e_i près qui est $\mathcal{O}(h^2)$ lorsque $h \longrightarrow 0$. On dit que le schéma (S) est **consistant**.

On appelle e_i l'erreur de troncature (ou erreur de consistance) du schéma (S) au point x_i.

Celle-ci étant $\mathcal{O}(h^2)$ (pour u assez régulière) on dit que le schéma (S) est **d'ordre 2**. Le schéma serait d'ordre p avec une erreur de troncature en $\mathcal{O}(h^p)$.

Remarque 1 On a ici

$$\max_{i\le i\le N}|e_i| \le \frac{h^2}{6}\left(\frac{1}{2}\left\|u^{(4)}\right\|_\infty + \|p\|_\infty \left\|u^{(3)}\right\|_\infty\right)$$

Le problème (S) consiste en un système de N équations à N inconnues. Il s'écrit sous forme matricielle

$$AU = B$$

avec $A \in M_N(\mathbb{R})$ et $B \in \mathbb{R}^N$ donnés par :

$$A = \begin{pmatrix} 2+h^2q_1 & -1+\frac{h}{2}p_1 & & 0 \\ -1-\frac{h}{2}p_2 & 2+h^2q_2 & -1+\frac{h}{2}p_2 & \\ & \ddots & \ddots & \\ & & 2+h^2q_{N-1} & -1+\frac{h}{2}p_{N-1} \\ 0 & & -1-\frac{h}{2}p_N & 2+h^2q_N \end{pmatrix} \quad B = \begin{pmatrix} h^2f_1+\alpha(1+\frac{h}{2}p_1) \\ h^2f_2 \\ \dots \\ h^2f_{N-1} \\ h^2f_n+\beta(1-\frac{h}{2}p_N) \end{pmatrix}$$

Le système (S) possède une unique solution pour h assez petit.

Théorème 2 *Si $q \geq 0$ sur $[a,b]$ et $h < \dfrac{2}{\|p\|_\infty}$ alors A est inversible.*

On va montrer ce résultat dans le cas où $q > 0$ sur $]a,b[$.

Définition 1 $A \in M_N(\mathbb{C})$ est à diagonale strictement dominante si

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \text{ pour tout } i = 1,..,N$$

Théorème 3 *Une matrice à diagonale strictement dominante est inversible.*

On vérifie que A est à diagonale strictement dominante sous les hypothèses du théorème 2 et pour $q_i > 0$. En effet :

$$\begin{cases} a_{i,i+1} \leq -1 + \frac{h}{2}\|p\|_\infty < 0 \\ a_{i,i-1} \leq -1 + \frac{h}{2}\|p\|_\infty < 0 \end{cases}$$

D'où :

* $|a_{11}| - |a_{12}| = 2 + h^2q_1 - (1-\frac{h}{2}p_1) \geq \underbrace{1 - \frac{h}{2}\|p\|_\infty}_{>0} + \underbrace{h^2q_1}_{>0} > 0$

* $|a_{N,N}| - |a_{N,N-1}| = 2 + h^2q_N - (1-\frac{h}{2}p_N) \geq \underbrace{1 - \frac{h}{2}\|p\|_\infty}_{>0} + \underbrace{h^2q_N}_{>0} > 0$

* Pour $i = 2, \dots, N-1$:

$$\begin{aligned} |a_{ii}| - |a_{i,i+1}| - |a_{i,i-1}| &= 2 + h^2q_i - (1 - \frac{h}{2}p_i) - (1 + \frac{h}{2}p_i) \\ &= h^2q_i > 0 \end{aligned}$$

A est donc à diagonale strictement dominante, et donc inversible d'après le théorème 3. Cela prouve le théorème 2 dans le cas $q > 0$.

1.3 Étude de la convergence dans un cas simplifié

Nous allons montrer que $u_i \underset{h\to 0}{\longrightarrow} u(x_i)$

$$(P)\ \begin{cases} -u''(x)+c(x)u(x)=f(x) & \text{sur }]0,1[\\ u(0)=u(1)=0 & c(x)\geq 0 \text{ est } \mathcal{C}^2 \text{ sur } [0,1] \end{cases}$$

Le problème discrétisé s'écrit : $(x_i = ih, c_i = c(x_i), f_i = f(x_i))$

$$(S)\ \begin{cases} -\frac{u_{i+1}-2u_i+u_{i-1}}{h^2}+c_iu_i=f_i & i=1,..,N \\ u_0=u_{N+1}=0 \end{cases}$$

Soit encore pour $U = {}^t(u_1,\cdots,u_N)$, $F = {}^t(f_1,\cdots,f_N)$

$$\frac{1}{h^2}A\,U=F, \qquad A=\begin{pmatrix} c_1h^2+2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & c_Nh^2+2 \end{pmatrix}$$

On a $A = M + h^2c$ avec

$$M=\begin{pmatrix} 2 & -1 & & 0 \\ -1 & \ddots & & -1 \\ & & \ddots & \\ 0 & & -1 & -2 \end{pmatrix} \text{ et } C=\begin{pmatrix} c_1 & & 0 \\ & \ddots & \\ 0 & & c_N \end{pmatrix}$$

Nous allons établir quelques propriétés de M qui seront utiles par la suite.

Lemme 1 Les valeurs propres de M sont $\lambda_k = 4\sin^2\left[\frac{k\pi}{2(N+1)}\right]$, $(k=1,\cdots,N)$, et les vecteurs propres associés $V^k = {}^t(v_1^k,\cdots,v_N^k)$ vérifient

$$v_i^k=\alpha\sin\left(\frac{ik\pi}{N+1}\right) \qquad i=1,\cdots,N$$

($\alpha\in\mathbb{R}$ constante arbitraire)

Preuve 2 M est réelle symétrique donc ses valeurs propres λ sont réelles. L'équation $MV = \lambda V$ s'écrit :

$$\begin{cases} 2v_j-v_{j+1}-v_{j-1}=\lambda v_j & j=1,\cdots,N \\ v_0=v_{n+1}=0 \end{cases}$$

Pour $1\leq j\leq N$, v_j vérifie donc une relation de récurrence linéaire du second ordre. Les suites solutions de cette relation de récurrence forment un espace vectoriel de dimension 2.

En posant $v_j = ar^j$ on botient l'équation caractéristique $r^2+(\lambda-2)r+1=0$, de discriminant $\Delta=\lambda(\lambda-4)$

- Plaçons-nous d'abord dans le cas $\Delta < 0$, c'est-à-dire $\lambda \in]0,4[$. r est complexe et de module 1, on pose $r = e^{iq}$ et on obtient $4\sin^2\left(\frac{q}{2}\right) = \lambda$. Les solutions de la relation de récurrence s'écrivent $v_j = ae^{iqj} + \overline{a}e^{-iqj} \quad (a \in \mathbb{C}, q \in [0,\pi])$

  Soit encore $v_j = A\sin(qj) + B\cos(qj)$.

  On cherche maintenant q, A, B de manière à satisfaire les conditions aux limites. $v_0 = 0$ donne $B = 0$ et $v_{N+1} = 0$ donne $q = \dfrac{k\pi}{N+1}, k = 1, \cdots, N$. D'où :

$$\lambda = 4\sin^2\left(\frac{k\pi}{2(N+1)}\right), \qquad v_j = A\sin\left(\frac{k\pi}{N+1}j\right)$$

- L'étude précédente donne N valeurs propres distinctes de M. On a donc trouvé toutes ses valeurs propres et il n'est pas nécessaire de considérer le cas $\Delta \geq 0$ (on n'obtiendrait alors que la solution $V = 0$).

Lemme 2 (Inégalité de Poincaré discrète)

$$\frac{1}{h^2}\,{}^tX\,M\,X \geq 4\,\|X\|_2^2 \qquad \forall X \in \mathbb{R}^N, h = \frac{1}{N+1}$$

Preuve 3 On a $M = P \wedge P^{-1}$ où $\wedge = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{pmatrix}$ et $P = \begin{pmatrix} V1 & \cdots & V^N \end{pmatrix}$

(V^k calculés dans le lemme 1). M est symétrique et les valeurs propres λ_k sont distinctes donc les vecteurs propres V^k sont deux à deux orthogonaux.

En fixant $\alpha = \sqrt{\frac{2}{N+1}}$ on a $\left\|V^k\right\|_2 = 1$ donc ${}^tPP = I$.

En posant $Y = {}^tPX$ on obtient

$$\frac{1}{h^2}\,{}^tX\,M\,X = \frac{1}{h^2}\,{}^tY \wedge Y = \sum_{k=1}^{N} \lambda_k\, y_k^2 \;\geq\; \frac{\lambda_1}{h^2}\,\|Y\|_2^2 = \frac{\lambda_1}{h^2}\,\|X\|_2^2$$

Donc :

$$\frac{1}{h^2}\,\frac{{}^tX\,M\,X}{\|X\|_2^2} \geq \frac{\lambda_1}{h^2} = \frac{1}{h^2} 4\sin^2\left(\frac{\pi k}{2}\right) \geq 4$$

$$\left(\sin x \geq \frac{2}{\pi}x \text{ si } 0 \leq x \leq \frac{\pi}{2}, \text{ et } \frac{\pi h}{2} \leq \frac{\pi}{4}\right)$$

On en déduit l'estimation suivante sur la norme euclidienne de la solution du système linéaire :

Lemme 3 La solution de $\dfrac{1}{h^2}\,A\,W = B$ vérifie :

$$\|W\|_2 \leq \frac{1}{4}\,\|B\|_2$$

Preuve 4

$$\frac{1}{h^2}\,{}^tW\,A\,W = \frac{1}{h^2}\,{}^tW\,M\,W + \underbrace{{}^tW\,C\,W}_{\geq 0} \geq 4\,\|W\|_2^2$$

d'après le lemme 2

Par ailleurs :

$$\frac{1}{h^2}\,{}^tW\,A\,W = {}^tW\,B \leq \|W\|_2\,\|B\|_2 \qquad \text{(inégalité de Cauchy-Schwartz)}$$

Donc $4\,\|W\|_2^2 \leq \|W\|_2\,\|B\|_2$

Ce résultat correspond à la **"stabilité L^2 du schéma numérique"**.

En effet, définissons la norme : $\|F\|_h = \sqrt{h}\,\|F\|_2$.

Si $f \in \mathcal{C}^0([0,1])$,

$$\|F\|_h^2 = \sum_{i=1}^{N} f(ih)^2 \times h \underset{h\to 0}{\longrightarrow} \int_0^1 f^2\,dx \qquad \text{(somme de Riemann)}$$

C'est-à-dire $\|F\|_h \underset{h\to 0}{\longrightarrow} \|f\|_{L^2(0,1)}$

Le lemme 3 donne $\|U\|_h \leq \frac{1}{4}\,\|F\|_h \leq \frac{1}{2}\,\|f\|_{L^2}$ pour h assez petit.

Soit U_h la "solution numérique" définie sur [0,1] par interpolation linéaire des u_i (donc $U_h(x_i) = u_i$). On vérifie que $\|U_h\|_{L^2} = \|U\|_h$.

On a donc

$$\|U_h\|_{L^2} \leq \frac{1}{2}\,\|f\|_{L^2} \qquad \text{pour } f \text{ fixé et } h \text{ assez petit}$$

La norme L^2 de la solution numérique reste bornée quand $h \longrightarrow 0$ et on dit alors que le schéma est stable dans L^2.

Notons maintenant $\tilde{u}_i = u(x_i)$ et $\tilde{U} = {}^t(\tilde{u}_1, \cdots, \tilde{u}_N)$

À l'aide d'un développement de Taylor en x_i, on obtient :

$$\begin{aligned}
&\frac{\tilde{u}_{i+1} - 2\tilde{u}_i + \tilde{u}_{i-1}}{h^2} + c_i\tilde{u}_i - f_i \\
&= -\,u''(x_i) + c(x_i)u(x_i) - f(x_i) + o(1) \qquad \text{quand } h \to 0 \\
&= o(1)
\end{aligned}$$

Si $u \in \mathcal{C}^4([0,1])$ (càd pour $f \in \mathcal{C}^2([0,1])$) on a mieux :

$$-\frac{\tilde{u}_{i+1} - 2\tilde{u}_i + \tilde{u}_{i-1}}{h^2} + c_i\tilde{u}_i - f_i = -\frac{h^2}{24}u^{(4)}(\theta_i^+) - \frac{h^2}{24}u^{(4)}(\theta_i^-)$$

$$\left(\theta_i^+ \in]x_i, x_{i+1}[,\ \text{et } \theta_i^- \in]x_{i-1}, x_i[\right)$$

D'où :

$$\left| -\frac{\tilde{u}_{i+1} - 2\tilde{u}_i + \tilde{u}_{i-1}}{h^2} + c_i\tilde{u}_i - f_i \right| \leq \frac{h^2}{12}\left\|U^{(4)}\right\|_\infty$$

(converge en h^2)

Donc la solution de (P) vérifie (S) avec une erreur qui tend vers 0 quand $h \to 0$. Le schéma (S) est donc **consistant**. Lorsque u est suffisamment régulière, cette erreur est en $\mathcal{O}(h^2)$; le schéma (S) est donc **d'ordre 2**.

Les calculs précédents montrent que :

$$\frac{1}{h^2} A\tilde{U} = F + E$$

$$\text{avec } \|E\|_\infty \leq \tfrac{h^2}{12} \left\|U^{(4)}\right\|_\infty \text{ si } u \in \mathcal{C}^4([0,1])$$

Puisque $\|.\|_h \leq \|.\|_\infty$ on a donc lorsque $h \to 0$:

$$\|E\|_h = \mathcal{O}(h^2) \qquad \text{si } u \text{ est } \mathcal{C}^4 \text{ (càd } f \text{ est } \mathcal{C}^2)$$

Après avoir montré la stabilité et la consistance du schéma (S), on peut maintenant conclure sur sa convergence.

En effet :

$$\frac{1}{h^2} A(\tilde{U} - U) = E$$

amène grâce au lemme 3 :

$$\left\|\tilde{U} - U\right\|_h \leq \frac{1}{4} \|E\|_h = \mathcal{O}(h^2) \qquad \text{si } f \text{ est } \mathcal{C}^2$$

Remarque 2

- Si on a simplement $f \in \mathcal{C}^0([0,1])$, on peut montrer que $\left\|\tilde{U} - U\right\|_h \longrightarrow 0$ lorsque $h \to 0$ à l'aide d'un argument de densité (cf TD)
- Si f est $\mathcal{C}^2$, $\left\|\tilde{U} - U\right\|_\infty \leq \left\|\tilde{U} - U\right\|_2 = \frac{1}{\sqrt{h}} \left\|\tilde{U} - U\right\|_h$ donne $\left\|\tilde{U} - U\right\|_\infty = \mathcal{O}(h^{3/2}))$

  Nous verrons en TD comment obtenir un résultat de convergence plus fort : $\left\|\tilde{U} - U\right\|_\infty = \mathcal{O}(h^2)$. Le principe est le même (stabilité + consistance en $\|.\|_\infty \implies$ convergence) mais la stabilité s'obtient d'une autre manière (en utilisant le "principe du maximum").

1.4 Conclusion

La méthode des différences finies conduit à la résolution de **systèmes linéaires de grande taille** pour lesquels nous introduirons différents types d'outils. Bien sûr nous n'avons vu qu'une introduction aux différences finies. Cette méthode est beaucoup utilisée pour la résolution d'équation aux dérivées partielles, comme l'équation de la chaleur $\dfrac{\partial u}{\partial t} = \dfrac{\partial^2 u}{\partial x^2}$ ou l'équation de Poisson $-\left(\dfrac{\partial^2 \phi}{\partial x^2} + \dfrac{\partial^2 \phi}{\partial y^2}\right) = f(x,y)$

Exemples de schémas pour ces deux équations

- Équation de la chaleur :

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = \frac{U_{j+1}^{n+1} - 2U_j^{n+1} + U_{j-1}^{n+1}}{\Delta x^2}$$

$U_j^n \simeq U(j\Delta x, n\Delta t) \qquad U_j^0 = U(j\Delta x, 0)$

Schéma **implicite** : $(U_j^{n+1})_j$ s'obtient à partir de $(U_j^n)_j$ en résolvant un système linéaire.

- Équation de Poisson :

$$\frac{-\phi_{i+1,k} - \phi_{i,k+1} + 4\phi_{i,k} - \phi_{i,k-1} - \phi_{i-1,?}}{h^2} = f_{i,k}$$

$\phi_{i,k} \simeq \phi(ih, kh)$

$(\phi_{i,k})_{i,k}$ s'obtient en fonction de $(f_{i,k})_{i,k}$ en résolvant un système linéaire.

Chapitre 2

Méthodes itératives pour des systèmes linéaires

On sait aujourd'hui résoudre numériquement des systèmes linéaires de l'ordre du million d'inconnues (et d'équations). Pour des **systèmes creux**, c'est-à-dire lorsque la matrice du système possède beaucoup de coefficients nuls, on arrive à une centaine de millions d'inconnues. Les **systèmes pleins** font appel à des **méthodes directes**, qui donnent la solution exacte (aux erreurs d'arrondi près) en un nombre fini d'itérations, et seront décrites dans un chapitre ultérieur.

Pour les **très grands systèmes creux** [1] on utilise des **méthodes itératives**, où on construit une suite de vecteurs qui convergent vers la solution.

L'intérêt est que **ces méthodes ne manipulent pas la matrice**, mais seulement une fonction qui définit une suite par récurrence.

Définition 2 Soit $A \in M_n(\mathbb{R})$ inversible et $b \in \mathbb{R}^n$. On appelle méthode itérative de résolution du système linéaire $Ax = b$, ($x \in \mathbb{R}^n$) une méthode qui construit une suite récurrente $(x_k)_{k\geq 0}$ telle que

$$(x_k \underset{k\to+\infty}{\longrightarrow} x) \Rightarrow Ax = b$$

Une méthode itérative est convergente si $x_k \underset{k\to+\infty}{\longrightarrow} x$ pour toute condition initiale $x_0 \in \mathbb{R}^n$

Tests d'arrêt typiques :

- $\frac{\|Ax_k - b\|}{\|b\|} < \varepsilon$ (norme du "résidu" / norme de b).
  Noter que :

$$\begin{aligned}\frac{\|x_k - x\|}{\|x\|} = \frac{\|A^{-1}(Ax_k - b)\|}{\|x\|} &\leq \|A^{-1}\|\frac{\|b\|}{\|x\|}\varepsilon \\ &\leq \|A^{-1}\|\|A\|\varepsilon, \text{ peut-être grand !}\end{aligned}$$

1. Exemple : discrétisation par différences finies de problèmes aux limites pour des équations aux dérivées partielles ... + schéma

Nous allons décrire ici des méthodes itératives avec "splitting" de A.

2.1 Description générale :

On considère ici le système

$$Ax = b \tag{2.1}$$

où $A \in M_n(\mathbb{R})$, $x \in \mathbb{R}^n$ et $b \in \mathbb{R}^n$. On suppose que la matrice A est inversible.

On considère une décomposition de A ("splitting") $A = M - N$ avec M inversible et on considère l'itération :

$$\begin{cases} Mx_{k+1} = Nx_k + b \\ \\ x_0 \in \mathbb{R}^n \end{cases} \tag{2.2}$$

Si $x_k \to x$ quand $k \to +\infty$ alors $Mx = Nx + b$, càd x est solution de (2.1).

Le choix du splitting est très important pour la performance de la méthode :

- Bien sûr la méthode doit être convergente (voir plus loin)
- On doit choisir M de telle sorte que le système (2.2) soit beaucoup plus facile à résoudre que (2.1) (il faut résoudre (2.2) à chaque étape de l'itération).

  Exemples : M diagonale ou triangulaire, diagonale ou triangulaire par blocs.

Étudions les conditions de convergence de (2.2).

Définition 3 Étant donné $A \in M_n(\mathbb{C})$, on note $S_p(A)$ l'ensemble des valeurs propres de A (ou "spectre de A"). On appelle rayon spectral de A et on note $\rho(A)$:

$$\rho(A) = \max_{\lambda \in S_p(A)} |\lambda|$$

Théorème 4 *La méthode (2.2) converge si et seulement si*

$$\rho(M^{-1}N) < 1$$

La preuve complète de ce résultat sera étudiée en TD. Ici nous allons simplement montrer que $\rho(M^{-1}N) < 1 \Rightarrow$ convergence de (2.2), en admettant pour cela deux résultats.

Théorème 5 (de l'application contractante (dans $\mathbb{R}^n$)) *Soit E un sous-ensemble de $\mathbb{R}^n$ fermé (et non vide). On considère une norme $\|\ \|$ sur $\mathbb{R}^n$.*

Soit $F : E \to E$ une application contractante, càd pour laquelle il existe $\alpha \in [0, 1[$ tel que :

$$\|F(x) - F(y)\| \leq \alpha \|x - y\|,\ \forall x, y \in E$$

Alors il existe un unique $x^* \in E$ tel que $F(x^*) = x^*$ (càd F admet un unique point fixe dans E). De plus, pour tout $x_0 \in E$, la suite définie par :

$$x_{k+1} = F(x_k)$$

converge vers x^*, avec

$$\|x^* - x_k\| \leq \frac{\alpha^k}{1-\alpha} \|x_1 - x_0\| \tag{2.3}$$

Le système (2.2) s'écrit :

$$\begin{cases} \boxed{x_{k+1} = M^{-1}Nx_k + M^{-1}b} \\ \\ x_0 \in \mathbb{R}^n \end{cases} \tag{2.4}$$

Remarque 3 Théorème encore appelé "Théorème du point fixe de Banach". Le théorème reste vrai lorsque E est un espace métrique complet.

Théorème 6 (cf TD pour la démonstration) *Soit $A \in M_n(\mathbb{C})$ et $\varepsilon > 0$. Il existe une norme $\|\ \|$ sur $\mathbb{C}^n$ telle que :*

$$\underbrace{\|A\|}_{\substack{\text{norme sur} \\ M_n(\mathbb{C}) \text{ induite} \\ \text{par la norme} \\ \|\ \| \text{ de } C^n}} := \sup_{\|x\|=1} \|Ax\| \leq \rho(A) + \varepsilon$$

Si $\rho(M^{-1}N) < 1$, il existe donc une norme matricielle induite telle que $\|M^{-1}N\| \leq \rho(M^{-1}N) + \varepsilon < 1$. Alors :

$$\|F(x) - F(y)\| = \|M^{-1}N(x - y)\| \leq \underbrace{\|M^{-1}N\|}_{<1} \|x - y\|$$

Donc $F : \mathbb{R}^n \to \mathbb{R}^n$ est une contraction.

Donc $\forall x_0 \in R^n$, la suite définie par (2.4) converge vers une limite $x \in \mathbb{R}^n$ unique, solution de $x = M^{-1}Nx + M^{-1}b$, c'est-à-dire $Ax = b$.

Vitesse de convergence : Plus $\rho(M^{-1}N)$ est petit, plus $\|M^{-1}N\|$ peut être choisie petite et plus la convergence est rapide. En effet, d'après (2.3) :

$$\|x - x_k\| \leq \frac{\|M^{-1}N\|^k}{1 - \|M^{-1}N\|}\|x_1 - x_0\|$$

Exemple de splitting : (peu utilisé)

$$M = \frac{1}{\alpha}I, \qquad N = \frac{1}{\alpha}I - A \qquad \Longrightarrow \qquad x_{k+1} = (I - \alpha A)x_k + \alpha b$$

(méthode de Richardson stationnaire, ou du gradient à pas fixe)

Elle converge si et seulement si $\forall \lambda \in Sp(A), |1 - \alpha\lambda| < 1$, c'est-à-dire toutes les valeurs propres de A se trouvent dans le disque (ouvert) de centre ($\frac{1}{\alpha}$ et rayon $\frac{1}{\alpha}$).

2.2 Méthode de Jacobi

On pose dans schéma (2.2) :

$$M = D \text{ avec } D \text{ diagonale et } d_{ii} = a_{ii}, \qquad N = D - A$$

Remarque 4 Cela suppose $a_{ii} \neq 0 \ \forall i$ (si cette condition n'est pas vérifiée on peut permuter des lignes de A).

Théorème 7 *Si A est à diagonale strictement dominante ($\rightarrow a_{ii} > 0$ et D inversible) alors la méthode de Jacobi converge.*

La démonstration sera vue en TD. On montre que le rayon spectral de la matrice $J = D^{-1}(D - A) = I - DA$ est < 1.

Nous avons rencontré ce type de matrices pour la discrétisation de problèmes aux limites dans le 1^{er} chapitre du cours.

2.3 Méthodes de Gauss-Seidel et SOR

On pose $A = D + L + U$ avec :

$$D = \begin{pmatrix} a_{00} & \cdots & 0 \\ \vdots & a_{ii} & \vdots \\ 0 & \cdots & a_{nn} \end{pmatrix}, L = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ a_{ij}(i > j) & \cdots & 0 \end{pmatrix}, U = \begin{pmatrix} 0 & \cdots & a_{ij}(j > i) \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix}$$

Dans la méthode de Gauss-Seidel, on fixe :

$$M = D + L, \ N = -U$$

La méthode s'écrit donc :

$$Dx_{k+1} = -Lx_{k+1} - Ux_k + b$$

En notant $x_k = (x_1^{(k)}, \ldots, x_n^{(k)})$ on obtient pour $i = 1..n$:

$$a_{ii}x_i^{(k+1)} = -\sum_{j<i} a_{ij}x_j^{(k+1)} - \sum_{(j>i)} a_{ij}x_j^{(k)} + b_i$$

Les méthodes de Jacobi et Gauss-Seidel ne sont guère utilisées. On leur préfère la méthode de relaxation.

La méthode SOR ("successive over-relaxation", ou "méthode de relaxation", environ 1950) généralise Gauss-Seidel en introduisant un paramètre de relaxation $\omega \neq 0$, que l'on ajuste afin d'accélérer la convergence de la méthode (avec un gain généralement très important si ω est bien choisi).

Pour $i = 1, \ldots, n$

$$\left\{ \begin{array}{rcl} a_{ii}\tilde{x}_i^{(k+1)} & = & -\sum_{j<i} a_{ij}x_j^{(k+1)} - \sum_{j>i} a_{ij}x_j^{(k)} + b_i \\ x_i^{(k+1)} & = & \omega\tilde{x}_i^{(k+1)} + (1-\omega)x_i^{(k)} \end{array} \right. \tag{2.5}$$

(Gauss-Seidel correspond à $\omega = 1$)

La méthode s'écrit (multiplier la seconde ligne par a_{ii}, et remplacer $a_{ii}\tilde{x}_i^{(k+1)}$ par son expression en fonction de x_{k+1} et x_k)

$$Dx_{k+1} = (1-\omega)Dx_k - \omega Lx_{k+1} - \omega Ux_k + \omega b$$

soit

$$(D + \omega L)x_{k+1} = [(1-\omega)D - \omega U]x_k + \omega b$$

On a donc :

$$M = \frac{1}{\omega}D + L, N = \frac{1-\omega}{\omega}D - U, M - N = D + L + U = A$$

On note :

$$\mathcal{L}_\omega := (\frac{1}{\omega}D + L)^{-1}(\frac{1-\omega}{\omega}D - U)$$

SOR converge si $\rho(\mathcal{L}_\omega < 1)$

Théorème 8 (demo en TD) *1. Soit $A \in M_n(\mathbb{R})$ inversible, avec $\forall i, a_{ii} \neq 0$ Une condition nécessaire pour que SOR converge est que $\omega \in]0, 2[$.*

2. Si A est symétrique définie positive, alors $\forall \omega \in]0, 2[$, SOR converge

Rappel 1 A est symétrique définie positive si A est symétrique, ${}^txAx = 0 \Rightarrow x = 0$, et $\forall x \in \mathbb{R}^n, {}^txAx \geq 0$

Corollaire 1 Si A est symétrique définie positive, alors la méthode de Gauss-Seidel converge. Nous avons rencontré ce type de matrices pour la discrétisation des problèmes aux limites dns le 1^{er} chapitre du cours (paragraphe 2), cas où la fonction p est identiquement nulle.

Remarque 5

1. Il y a des exemples où A est symétrique définie positive et où la méthode de Jacobi n'est pas convergente.

2. Si A est tridiagonale ($a_{ij} = 0$ si $|i-j| > 1$) et D inversible, on peut montrer que $\rho(\mathcal{L}_1) = \rho(J)^2$. Donc la méthode de Gauss-Seidel converge si et seulement si celle de Jacobi converge (et Gauss-Seidel converge plus vite).

3. Pour quelques types de matrices, on connaît la valeur de ω qui minimise $\rho(\mathcal{L}_\omega)$. Pour A tridiagonale (avec D inversible), et si les valeurs propres de J sont réelles, alors le paramètre de relaxation optimal dans SOR (c'est-à-dire la valeur de ω qui minimise $\rho(\mathcal{L}_\omega)$) est > 1 (donc Gauss-Seidel ne donne pas la vitesse optimale de convergence).

4. Pour optimiser empiriquement le choix de ω dans SOR, on peut évaluer le facteur de contractivité $\frac{\|x_{k+1}-x_k\|}{\|x_k-x_{k-1}\|}$ à partir du moment où $\|x_{k+1} - x_k\|$ décroît vers 0.

Méthode		Itérations	Temps CPU
Jacobi		34 900	902
Gauss-Seidel		20 450	1121
Relaxation	$\omega = 1,8$	3 270	180
Relaxation	$\omega = 1,93$	1 200	66
Relaxation	$\omega = 1,98$	530	24,7
Gradient conjugué		539	24,6
Gradient conjugué. Tridiagonale		443	44
Gradient conjugué SSOR	$\omega = 1,0$	152	15
Gradient conjugué SSOR	$\omega = 1,8$	57	5,8
Gradient conjugué SSOR	$\omega = 1,93$	40	4,1

Exemples pratiques

Chapitre 3

Méthode de Gauss pour les systèmes linéaires et factorisation LU

Soit $A \in M_n(\mathbb{R})$ inversible et $b \in \mathbb{R}^n$. La méthode de Gauss permet de résoudre le système $Ax = b$, $x \in \mathbb{R}^n$ en se ramenant à la résolution d'un système triangulaire. Nous allons commencer par rappeler cette méthode classique de résolution des systèmes linéaires. Il s'agit d'une méthode directe, càd qui donne la solution exacte un nombre fini d'opérations arithmétiques élémentaires. Nous verrons ensuite que l'élimination de Gauss fournit une factorisation $A = LU$ (ou $PA = LU$, P étant une matrice de permutation, dépendant du choix des pivots) avec L triangulaire inférieure et U triangulaire supérieure. Résoudre $Ax = b \Leftrightarrow PAx = Pb \Leftrightarrow LUx = Pb$ revient donc à :

1. Factoriser PA
2. Résoudre $Lc = Pb$ (étape de descente) $c_1 \to c_2 \to \cdots \to c_n$
3. Résoudre $Ux = c$ (étape de remontée) : $x_n \to x_{n-1} \to \cdots \to x_1$

Si on doit résoudre de nombreuses fois avec la même matrice :

$$Ax^{(k)} = b^{(k)}$$

Schémas "implicites" pour des EDP, schémas itératifs pour des systèmes linéaires ou non linéaires, ...) alors l'étape 1) qui est la plus coûteuse est effectuée une seule fois.

3.1 Rappel de l'élimination de Gauss :

Soit $A \in M_n(\mathbb{R})$ inversible et $b \in \mathbb{R}^n$. On cherche $x \in \mathbb{R}^n$ tel que $Ax = b$, soit :

$$\left\{ \begin{array}{rcl} a_{11}X_1 + a_{12}X_2 + \cdots + a_{1n}X_n & = & b_1 \\ \vdots & & \\ a_{n1}X_1 + a_{n2}X_2 + \cdots + a_{nn}X_n & = & b_n \end{array} \right. \tag{3.1}$$

En notant $L_i = (a_{i1}, \cdots, a_{in})$ la i$^{\text{ème}}$ ligne de A, on a

$$\left\{ \begin{array}{l} L_1X = b_1 \\ \vdots \\ L_nX = b_n \end{array} \right. \tag{S}$$

Si $a_{11} \neq 0$, on peut éliminer la variable x_1 dans les lignes 2 à n. On dit qu'on choisit a_{11} comme pivot. (S) équivaut à :

$$\left\{ \begin{array}{rcll} L_1 X & = & b_1 & \\ (L_i - \frac{a_{i1}}{a_{11}} L_1) X & = & b_i - \frac{a_{i1}}{a_{11}} b_1 & i = 2..n \end{array} \right.$$

Le nouveau système s'écrit $A^{(2)}X = b^{(2)}$ avec

$$A = \begin{pmatrix} a_{11}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & & \\ \vdots & & \\ 0 & & \end{pmatrix}, (a_{ij}^{(1)} = a_{ij})$$

Ligne $i = L_i - l_{i1} L_1$ avec $l_{i1} = \frac{a_{i1}}{a_{11}}$

$b_i^{(2)} = b_i - l_{i1} b_1$

Si a_{11} on permute la 1ère ligne de (S) avec une autre ou $a_{i1} \neq 0$. Cela est toujours possible puisque A est inversible. On effectue la même procédure que précédemment expliqué.

Le système $A^{(2)}X = b^{(2)}$ contient un sous-système de dimension $n-1$ pour $x_2 \dots x_n$ On répète la même procédure sur le sous-système pour éliminer X_2 des lignes 3 à n.

On continue ainsi et on déduit

$$A^{(3)}X = b^{(3)}, A^{(4)}X = b^{(4)}, \dots, A^{(n)}X = b^{(n)}$$

Le dernier système obtenu est triangulaire. Notons $A^{(n)} = U$, $b^{(n)} = C$.

$$\left\{ \begin{array}{rcccccl} u_{11}x_1 & + & \dots & + & u_{1n}x_n & = & c_1 \\ & u_{22}x_2 + & \dots & + & u_{2n}x_n & = & c_2 \\ & & & \ddots & & & \\ & & & & u_{nn}x_n & = & c_n \end{array} \right. \quad \text{(S')}$$

(S') est facile à résoudre : "étape de remontée".

$$x_{nn} = \frac{c_n}{x_{nn}}, x_i = \frac{1}{u_{ii}}(c_i - \sum_{j=i+1}^{n} u_{ij}x_j)$$

pour $i = n-1, \dots, 1$

Remarque 6 On appelle "factorisation" le calcul de U.

3.2 Factorisation LU

Nous avons vu que l'élimination de Gauss peut conduire à permuter des lignes de A puisqu'on a besoin de "pivots" non nuls $a_{11}^{(1)}, a_{22}^{(2)}$ etc ...Les cas où ces pivots sont voisins de 0 conduisent à des problèmes numériques (voir plus loin). Il est donc fréquent d'effectuer des permutations des lignes de A lors de l'élimination de Gauss. Nous allons tout d'abord voir que ces permutations sont une traduction matricielle simple.

Notons $p_1, p_2, \ldots, p_n$ une permutation des entiers $1, 2, \ldots, n$ et $(e_1, \ldots, e_n)$ la base canonique de $\mathbb{R}^n$. On appelle matrice de permutation une matrice de la forme :

$$P = \left(e_{p_1} \mid e_{p_2} \mid \ldots \mid e_{p_n}\right)$$

On a $p_{l_i} = e_{p_i}$ et :

$$P \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \vdots \\ \vdots \\ x_j \\ \vdots \end{pmatrix} \leftarrow \text{ligne } p_j, \qquad \underbrace{P \begin{pmatrix} L_1 \\ L_2 \\ \vdots \\ L_n \end{pmatrix}}_{\text{A}} = \begin{pmatrix} \vdots \\ \vdots \\ L_j \\ \vdots \end{pmatrix} \leftarrow \text{ligne } p_j$$

Soit $A \in M_n(\mathbb{R})$. Notons P la matrice correspondant aux permutations effectuées sur les lignes de A dans l'algorithme du paragraphe **1**. On a donc $PA = \tilde{A}$, où l'élimination de Gauss sur $\tilde{A}$ se fait sans permutation. Le passage de $A^{(j-1)}$ à $A^{(j)}$ s'écrit (même notations qu'auparavant) :

$$\text{Ligne } i = L_i - L_j \times \left(\frac{a_{ij}^{(j-1)}}{a_{jj}^{(j-1)}}\right), \quad i = j+1, \ldots, n$$

Soit matriciellement :

$$A^{(j)} = T_j A^{(j-1)}, \quad T_j = \left(\begin{array}{c|c|c} & I_j & 0 \\ \hline & -l_{j+1,j} & \\ 0 & \vdots & I_{n_j} \\ & -l_{n,j} & \end{array}\right)$$

$$l_{ij} = \frac{a_{ij}^{(j-1)}}{a_{jj}^{(j-1)}}, \quad I_k = \text{matrice identité de taille } k$$

En effet :

$$\begin{array}{l} \text{ligne } i \rightarrow \\ \\ \\ \\ \\ \\ \uparrow \text{colonne } j \end{array} \left(\begin{array}{c|c|c} & 0 & \\ \hline & 0 & \\ \hline 0 \ldots 0 & 1 & 0 \ldots 0 \\ \hline & 0 & \\ \hline & 0 & \end{array}\right) \times \left(\begin{array}{c} L_1 \\ \hline L_2 \\ \hline \vdots \\ \hline L_n \end{array}\right) = \left(\begin{array}{c} 0 \\ \hline 0 \\ \hline L_j \\ \hline 0 \\ \hline 0 \end{array}\right) \leftarrow \text{ligne } j$$

On a donc :

$$U = A^{(n-1)} = T_{n-1} \times T_{n-2} \times \cdots \times T_1 \times PA = TPA$$

Avec U triangulaire supérieure et T triangulaire inférieure (produit de matrices triangulaires inférieures).

Donc on a $PA = LU$ avec $L = T^{-1}$ triangulaire inférieure (l'inverse d'une matrice triangulaire inférieure l'est aussi).

Remarque 7 Dans S' on a $C = TPb$ et donc $LC = Pb$.

Soit maintenant :

$$\tilde{L} = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ l_{ij} & & 1 \end{pmatrix}$$

On a :

$$T_1\tilde{L} = \begin{pmatrix} 1 & & & 0 \\ 0 & 1 & & \\ \vdots & & \ddots & \\ 0 & l_{ij} & & 1 \end{pmatrix}, \ldots, T_{n-1} \times T_{n-2} \times T_1\tilde{L} = I$$

Donc $L = \tilde{L}$. Nous avons donc montré le résultat suivant :

Théorème 9 (Factorisation LU d'une matrice inversible) *Soit $A \in M_n(\mathbb{R})$ inversible. Il existe une matrice de permutation P et deux matrices triangulaires L (triangulaire inférieure de diagonale unité) et U (triangulaire supérieure inversible) telles que $PA = LU$.*

Cette décomposition est donnée explicitement par l'élimination de Gauss, avec (les coefficients l_{ij} sont ceux de l'élimination de Gauss sur $\tilde{A}$) :

$$L = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ l_{ij} & & 1 \end{pmatrix}, U = \begin{pmatrix} \ddots & & U_{ij} \\ & \ddots & \\ 0 & & \end{pmatrix}$$

Cette factorisation est unique lorsque l'on fixe P.

Remarque 8 L'unicité s'obtient simplement : PA est inversible (puisque P et A le sont).

Si $PA = L_1U_1 = L_2U_2$ alors L_i et U_i sont inversibles et donc $L_2^{-1}L_1 = U_2U_1^{-1}$. Le membre de droite est triangulaire supérieur, celui de gauche et triangulaire inférieur de diagonale unité.

Donc $L_2^{-1}L_1 = U_2U_1^{-1} = I$, i.e. $L1 = L2$ et $U_1 = U_2$.

Les permutations effectuées lors de l'élimination de Gauss sont très importantes d'un point de vue numérique (voir plus loin). Bien sûr, si l'on raisonne en arithmétique exacte (sans tenir compte des erreurs d'arrondi) on voit dans l'algorithme de Gauss que les cas nécessiteant une permutation sont exceptionnels (cela se produit lorsque a_{11} ou $a^{(i)}_{i+1,i+1} = 0$ pour certaines valeurs de i).

On a plus précisément le résultat suivant :

Théorème 10 *Dans le théorème 9, si les n sous-matrices*

$$\Delta = \begin{pmatrix} a_{11} & \dots & a_{1i} \\ \vdots & & \vdots \\ a_{i1} & \dots & a_{ii} \end{pmatrix}, (1 \leq i \leq n)$$

sont inversibles, alors on peut fixer $P = I$.

Preuve 5 $a_{11} \neq 0$ donc la 1ère de l'élimination de Gauss ne nécessite pas de permutation. Supposons qu'on ait j étapes sans permutation :

$$A^{(j)} = \left(\begin{array}{c|c} U^{(j)} & X \\ \hline 0 & X \end{array}\right) = T_j T_{j-1} \times \cdots \times T_1 \times A = \left(\begin{array}{ccc|c} 1 & & 0 & \\ & \ddots & & 0 \\ X & & 1 & \\ \hline & X & & I \end{array}\right) \left(\begin{array}{c|c} \Delta_j & X \\ \hline X & X \end{array}\right)$$

avec $U^{(j)} \in M_j(\mathbb{R})$ de la forme $\begin{pmatrix} a_{11} & & & X \\ & a_{22}^{(1)} & & \\ & & \ddots & \\ 0 & & & a_{jj}^{(j)} \end{pmatrix}$ (les coefficients diagonaux sont les pivots).

Alors on a aussi :

$$A^{(j)} = \left(\begin{array}{cc|c} & U^{(j+1)} & X \\ \hline 0 & X & X \end{array}\right) = \left(\begin{array}{ccc|c} 1 & & 0 & \\ & \ddots & & 0 \\ X & & 1 & \\ \hline & X & & I \end{array}\right) \left(\begin{array}{c|c} \Delta_{j+1} & X \\ \hline X & X \end{array}\right)$$

avec $U^{(j+1)} = \begin{pmatrix} a_{11} & & & X \\ & a_{22}^{(1)} & & \\ & & \ddots & \\ 0 & & & a_{j+1,j+1}^{(j)} \end{pmatrix}$. Donc $U^{(j+1)} = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ X & & 1 \end{pmatrix} \times \Delta_{j+1}$

D'où : $a_{11} \times a_{22}^{(1)} \times \cdots \times a_{(j+1),(j+1)}^{(j)} = \operatorname{Det} U^{(j+1)} = \operatorname{Det} \Delta_{j+1} \neq 0$

Donc $a_{j+1,j+1}^{(j)} \neq 0$ et on peut choisir ce coefficient comme pivot pour l'étape $j+1$.

Par récurrence, on peut donc choisir les coefficients $a_{11}, a_{i+1,i+1}^{(i)}$ comme pivots, puisque tous ces coefficients sont $\neq 0$. On obtient donc le résultat du théorème 9 avec $P = I$.

Remarque 9 Les coefficients l_{ij} du théorème 9 sont ceux qui apparaissent dans l'élimination de Gauss faite sur $\tilde{A}$. Cependant, ils peuvent aussi se calculer directement à partir

de l'élimination de Gauss faite sur A. Pour cela, quand on permute deux lignes de A, on réalise la même permutation sur les coefficients l_{ij} calculés précédemment (cf TD pour un exemple).

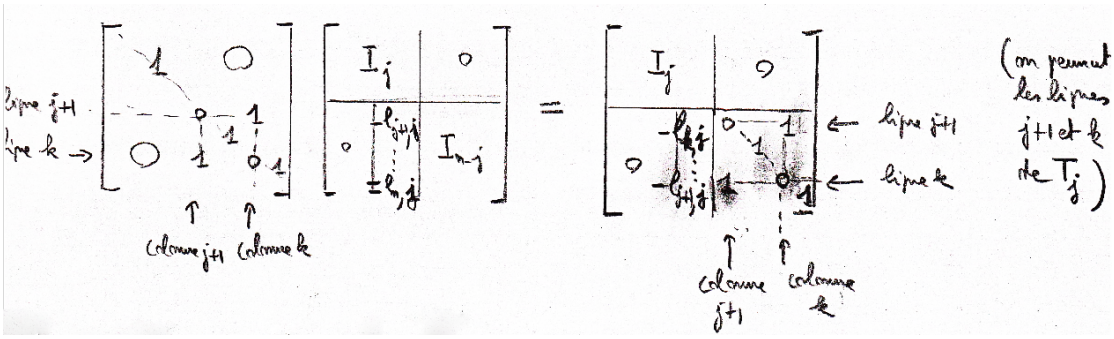
La remarque suivante détaille pourquoi ce procédé fonctionne.

Remarque 10 (Permutation des coefficients l_{ij} lors de l'élimination de Gauss) Soit P la matrice de permutation telle que $P_{l_i} = e_{P_i}$. Alors :

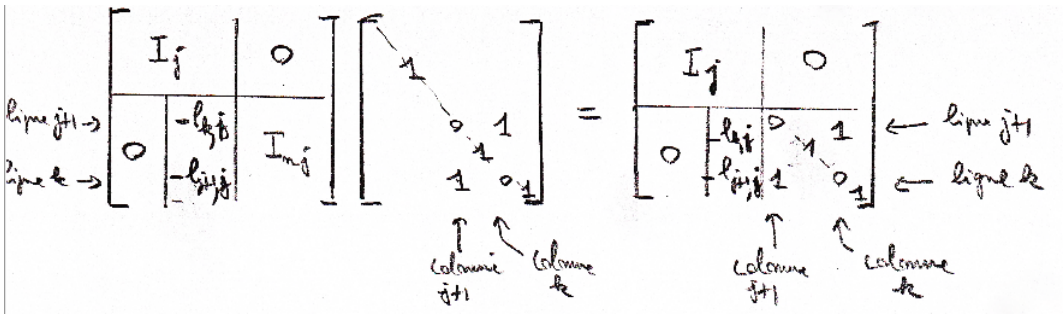
$$(c_1 \mid c_2 \mid \dots \mid c_n).P = (\quad \mid c_{p_i} \mid \quad) \leftarrow \text{colonne } i$$

(prendre la transposée du membre de droite et appliquer le résultat donné précédemment)

Dans l'élimination de Gauss, lorsqu'on effectue sur $A^{(j-1)}$ une combinaison linéaire de lignes, puis une permutation des lignes $j+1$ et k, on multiplie $A^{(j-1)}$ par :



Cela revient au même de faire d'abord la permutation des lignes de $A^{(j-1)}$ puis la combinaison linéaire où l'on permute les coefficients $l_{k,j}$ et $l_{j+1,j}$ (cf la propriété de P donnée plus loin : on permute les colonnes $j+1$ et k de T_j).



Donc les coefficients l_{ij} du théorème 1 (coefficients de l'élimination de Gauss faite sous permutation sur PA) s'obtiennent par permutation des coefficients l_{ij} du paragraphe 1) correspondant à l'élimination de Gauss sur la matrice A.

3.3 Techiques de choix du pivot

Le choix d'un pivot non nul mais très petit peut conduire à des erreurs numériques importantes. Par exemple, le système :

$$\begin{cases} \varepsilon\, x_1 + x_2 = 1 \\ x_1 + x_2 = 0 \end{cases}$$

a pour $\varepsilon \neq 1$ une solution unique $x_1 = -x_2 = \dfrac{1}{\varepsilon - 1}$

- Si on résout (S) par la méthode de Gauss en utilisant ε comme pivot, on obtient le système équivalent :

$$\varepsilon\, x_1 = 1 - x_2 \tag{3.2}$$

$$x_2(\frac{1}{\varepsilon} - 1) = \frac{1}{\varepsilon} \tag{3.3}$$

Par exemple , fixons $\varepsilon = 10^{-6}$. On simule un calcul en virgule flottante avec 5 chiffres significatifs. Alors $\frac{1}{\varepsilon} = 0,1.10^7$ (valeur exacte : $0,999999.10^6$) d'où $x_2 = 1$ (valeur exacte : $x_2 = 10 \times \frac{100000}{999999} = 1,000001000001\ldots$).

Mais alors $x_1 = 0$, ce qui est complètement faux (valeur exacte : $x_1 = -x_2$).

Dans le membre de droite de (3.2), on effectue une soustraction qui est très mal conditionnée car $x_2 \approx 1$. Dans le calcul à virgule flottante, on a $1 - x_2 = 0$, alors que la valeur exacte est $1 - x_2 \approx -10^6$; on commet donc une erreur relative de 100%, alors que x_2 est connu avec une erreur relative de 10^{-6}.

Si on choisit 1 comme pivot, on obtient :

$$\begin{cases} x_1 + x_2 = & 0 \\ (1-\varepsilon)x_2 = & 1 \end{cases}$$

Alors $-\varepsilon + 1 = +0,1.10^1$ en virgule flottante (valeur exacte $+0,999999$ d'où $x_2 = +1$ (précision $\sim 10^{-6}$) et $x_1 = -1$ (précision $\sim 10^{-6}$).

Cela motive la :

Méthode de Gauss avec pivot partiel

Même lorsque $a_{11} \neq 0$, on permute la 1ère ligne de (S) avec la ligne où $|a_{i1}|$ est le plus grand. La même stratégie est répétée pour tous les sous-systèmes apparaissant dans l'élimination de Gauss.

Remarque 11 Il existe aussi une méthode de Gauss avec pivot total, où on choisit comme pivot a_{i_0,j_0} avec $|a_{i_0,j_0}| = \mathrm{Max}\,|a_{ij}|$.

On permute alors la 1ère ligne de (S) avec la ligne i_0, et on permute les inconnues x_1 et x_{j_0}. On répète ce procédé pour tous les sous-systèmes qui apparaissent ensuite dans l'élimination de Gauss.

La méthode avec pivot partiel est la plus employée. Elle marche bien en pratique. La méthode avec pivot total est plus coûteuse en temps de calcul ; elle est donc assez peu employée.

3.4 Le coût de la méthode de Gauss

$$\begin{aligned} Ax = b \Leftrightarrow \quad & PAX = Pb \\ \Leftrightarrow \quad & LUX = Pb \end{aligned} \tag{3.4}$$

Résoudre 3.4 en 3 étapes :

1. Factoriser A ($PA = LU$)
2. Résoudre $LC = Pb$ (méthode de remontée)
3. Résoudre $UX = c$ (méthode de descente)

Remarque 12 1. L'étape 1. est la plus coûteuse

2. Utiliser Fact-LU quand on a plusieurs systèmes linéiaires. Avec la même matrice à résoudre : $AX^{(i)} = b^{(i)},\ i = 1, 2 \ldots$

* Factorisation $A = LU \quad (P = I)$, à l'étape 1 de la factorisation : passage $A \to A^{(2)}$

1. $(n-1)$ divisions (calcul de $l_{21}, \ldots, l_{n1}$
2. $2n(n-1)$ multitplications et additions (calcul $a_{ij} - l_{i1}a_{1j}, 2 \leq i \leq n, 1 \leq j \leq n$
3. Passage b à $b^{(2)}$, $2(n-1)$ multiplications et additions ($b_i - l_{i1}b_1, 2 \leq i \leq n$

Il faut renouveler cette procédure pour les sous-systèmes de taille $n-1, n-2, \ldots, 2$

Au total

$$\left.\begin{array}{ll} \sim 2\sum_{i=1}^{n}(n-i)^2 = 2.\frac{1}{3}n(n-\frac{1}{2})(n-1) & +, * \\ \sim 3\sum_{i=1}^{n}(n-i) = 3.\frac{1}{2}n(n-1) & +, *, \div \end{array}\right\} \Rightarrow \frac{2}{3}n^3 + O(n^2)$$

* Réoslution d'un système triangulaire (étape de remontée) (schéma pas pris)

$$X_n = \frac{b_n}{U_{nn}} x_i = \frac{1}{U_{2i}}(b_i - \sum_{k=i+1}^{n} a_{iK}X_K$$

$$O(n^2) \begin{cases} 1 + 2 + 3 + \cdots + n - 1 = \sum_{i=1}^{n-1} i = \frac{1}{2}(n-1)(n-2) \\ 1 + 2 + 3 + \cdots + n - 1 = \sum_{i=1}^{n-1} i = \frac{1}{2}(n-1)(n-2) \\ \text{n divisions} \end{cases}$$

$\Rightarrow$ Donc la méthode de Gauss nécessite $\frac{2}{3}n^3 + O(n^2)$ opérations.

Remarque 13 1. Le pivot partiel a un coût en $O(n^2)$.

$\to$ Trouver le max parmi $n, n-1, \ldots, 2, 1$ coeff.

2. Le pivot toal a un coût en $O(n^3)$

$\to$ Trouver le max parmi $n^2, (n-1)^2, \ldots, 2^2, 1^2$ coeff.

Chapitre 4

Factorisation de Cholesky

La méthode de Cholesky (ou Choleski) permet de résoudre des systèmes linéaires dont la matrice est symétrique définie positive. On rappelle que $A \in M_n(\mathbb{R})$ est symétrique définie positive si et seulement si :

1. ${}^tA = A$
2. ${}^tXAX \geq 0, \forall x \in \mathbb{R}$
3. ${}^tXAX = 0 \Rightarrow X = 0$

Pour une matrice $A \in M_n(\mathbb{R})$ symétrique, les propriétés 2) et 3) reviennent à supposer que toutes les valeurs propres de A sont > 0.

On a également :

1. A définie positive $\Rightarrow$ A inversible
2. A symétrique $\Rightarrow$ A diagonalisable. $\exists P / A = P^{-1}DP$
3. A symétrique et définie positive $\Rightarrow$ toutes les valeurs propres de A sont positives et réelles.

Nous allons voir que les matrices pevuent se factoriser sous la forme $A = T\,{}^tT$ où T est triangulaire inférieure.

Pour résoudre $Ax = b$ $(x, b \in \mathbb{R}^n)$, càd $T\,{}^tTx = b$ on résout alors successivement :

$$\begin{cases} Ty = b & \text{étape de “descente”} \\ {}^tTx = y & \text{étape de “remontée”} \end{cases}$$

Cette méthode est bien sûr intéressante lorsqu'on doit résoudre plusieurs systèmes linéaires avec des seconds membres b différents, mais la même matrice A (la factorisation est faite une seule fois, et les étapes de descente et remontée pour chaque second membre).

Théorème 11 *Soit $A \in M_n(\mathbb{R})$ symétrique définie positive. Il existe (au moins) une matrice réelle triangulaire inférieure T telle que :*

$$A = T\,{}^tT$$

De plus, si on impose que les éléments diagonaux de T soient tous positifs, alors la factorisation $A = T\,{}^tT$ est unique.

Preuve 6 Les n sous-matrices $\Delta_i = \begin{pmatrix} a_{11} & \dots & a_{1i} \\ \vdots & & \vdots \\ a_{i1} & \dots & a_{ii} \end{pmatrix}$ sont inversibles car elles sont symétriques définies positives. En effet, si $x = \begin{pmatrix} x_1 \\ \vdots \\ x_i \end{pmatrix}$ et $X = \begin{pmatrix} x_1 \\ \vdots \\ x_i \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ on a ${}^tx\Delta_i x = {}^tXAX$

Donc, d'après le théorème 2 du chapitre précédent, on a $A = LU$ avec L triangulaire inférieure de diagonale unité, et U triangulaire supérieure inversible ($\rightarrow u_{ii} \neq 0 \forall i$). Notons $D = \operatorname{diag} u_{ii}$

On a : $A = {}^tA = {}^tU\,{}^tL = \underbrace{{}^tU D^{-1}}_{\text{triang. inf de diagonale unité}} \times \underbrace{D\,{}^tL}_{\text{triangulaire supérieure}}$

Par unicité de la décomposition LU il vient $D\,{}^tL = U$ et donc $A = LD\,{}^tL$. De plus, si ${}^tLV_i = e_i$

$$0 < {}^tV_i\,A\,V_i = {}^tV_i\,L\,D\,{}^tL\,V_i = {}^te_i\,D\,e_i = u_{ii}$$

Notons maintenant $\sqrt{D} = \operatorname{diag}\sqrt{u_{ii}}$. Alors $A = L\sqrt{D}\,{}^t\sqrt{D}\,{}^tL$, c'est-à-dire $A = T\,{}^tT$ avec $T = L\sqrt{D}$.

L'unicité de la factorisation $A = T\,{}^tT$ vient de l'algorithme de calcul de T décrit plus loin.

Calcul de T :

T peut être calculé à partir d'une factorisation LU, mais on expose ici une méthode moins coûteuse en temps de calcul.

Puisque $A = T\,{}^tT$ avec T triangulaire inférieure :

$$a_{ij} = \sum_{1 \leq k \leq \operatorname{Min}(i,j)} t_{ik}t_{jk}$$

- Calculons la 1ère colonne de T à partir de celle de A :

$$a_{11} = t_{11}^2 \implies t_{11} = \sqrt{a_{11}} \qquad (a_{11} > 0 \text{ car } A \text{ symétrique définie positive})$$

Pour $i \geq 2$:

$$a_{i1} = t_{i1}t_{11} \implies t_{i1} = \frac{a_{i1}}{t_{11}}$$

- Supposons connues les colonnes 1 à p de T, et calculons la colonne $p+1$:

$$a_{p+1,p+1} = t^2_{p+1,p+1} + \sum_{1\leq k\leq p} (t_{p+1,k})^2 \rightarrow \text{calculés précédemment (colonnes 1 à } p)$$

$$\Longrightarrow t^2_{p+1,p+1} = a_{p+1,p+1} - \sum_{1\leq k\leq p} (t_{p+1,k})^2 > 0 \quad \underline{\text{puisque } T \text{ existe}}$$

$$\Longrightarrow t_{p+1,p+1} = \left(a_{p+1,p+1} - \sum_{1\leq k\leq p} (t_{p+1,k})^2) \right)^{\frac{1}{2}}$$

Pour $i \geq p+2$:

$$a_{i,p+1} = t_{i,p+1}t_{p+1,p+1} + \sum_{1\leq k\leq p} (t_{ik}t_{p+1,k}) \qquad \rightarrow \text{(calculés précédemment)}$$

$$\Longrightarrow t_{i,p+1} = (a_{i,p+1} - \sum_{1\leq k\leq p} t_{ik}t_{p+1,r}\frac{1}{t_{p+1,p+1}})$$

Coût du calcul de T :

On assimile l'extraction de racine carrée à une opération arithmétique élémentaire (ce qui est une approximation, car cette opération est plus compliquée que les opérations élémentaires $\times, +, -\ldots$).

Calcul de la $1^{\text{ère}}$ colonne de T : n opérations.

Calcul de la $(p+1)^{\text{ème}}$ de T :

- Calcul de $t_{p+1,p+1}$:

$$\left.\begin{array}{l} * \ p \text{ multiplications} \\ * \ p \text{ soustractions} \\ * \ 1 \text{ racine} \end{array}\right\} 2p+1 \text{ opérations}$$

- Calcul de $t_{i,p+1} (p+2 \leq i \leq n)$:

$$\left.\begin{array}{l} * \ p \text{ multiplications} \\ * \ p \text{ soustractions} \\ * \ 1 \text{ divison} \end{array}\right\} 2p+1 \text{ opérations}$$

$\rightarrow (n-p) \times (2p+1)$ opérations pour le calcul de la colonne $p+1$.

$$\begin{aligned} \text{Coût total} &= \sum_{p=0}^{n-1} (n-p)(2p+1) \\ &= (2n-1)\left(\sum_{p=1}^{n-1} p\right) + n^2 - 2\sum_{p=1}^{n-1} p^2 \\ &\sim_{n\rightarrow+\infty} n^3 - \frac{2}{3}n^3 \end{aligned}$$

Donc nombre d'opérations élémentaires $\sim_{n\to+\infty} \frac{1}{3}n^3$. ($\sim \frac{1}{6}n^3$ multiplications et $\frac{1}{6}n^3$ soustractions)

Coût de la résolution de $Ax = b$

Le coût des étapes des descente et de remontée est $\mathcal{O}(n^2)$.

L'étape la plus coûteuse est donc la factorisation $A = T\,{}^tT$, et le coût total est $\sim \frac{1}{3}n^3$. C'est donc mieux que Gauss (coût $\sim \frac{2}{3}n^3$).

Chapitre 5

Résolution numérique d'équations non linéaires

De nombreux problèmes issus notamment de la physique conduisent à la résolution d'équations non linéaires,

$$f(x) = 0,\ f \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R}^n)$$

5.1 Méthode des approximations successives

À partir d'une équation $f(x) = 0$ ($f : \mathbb{R}^n \longrightarrow R^n$ de classe $\mathcal{C}^1$), on peut se ramener à un problème de point fixe :

$$x = \Phi(x) \tag{5.1}$$

avec $\Phi : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ de classe $\mathcal{C}^1$. On peut poser par exemple :

$$\Phi(x) = x - Bf(x)$$

avec $B \in M_n(\mathbb{R})$ inversible.

Pour résoudre (5.1), on se donne une condition initiale $x_0 \in \mathbb{R}^n$ (la plus proche possible d'une solution de (5.1)) et on considère la méthode itérative :

$$x_{k+1} = \Phi(x_k) \tag{5.2}$$

Nous allons étudier la convergence de ce type de méthodes itératives.

Définition 4 Soit a un point fixe de Φ ($\Phi(a) = a$).

i) a est stable au sens de Lyapunov si

$$\forall \varepsilon > 0, \exists \eta \ /\ \|x_0 - a\| < \eta \implies \|x_k - a\| < \varepsilon \qquad \forall k \geq 0$$

ii) a est instable s'il n'est pas stable au sens de Lyapunov.

iii) a est asymptotiquement stable s'il est stable au sens de Lyapunov et

$$\exists r \ /\ \|x_0 - a\| < r \implies x_k \xrightarrow{k \to +\infty} a$$

Lorsque a est asymptotiquement stable, la méthode (5.2) permet de calculer numériquement a à partir d'une condition initiale x_0 "suffisamment proche" de a.

Théorème 12 *Soit Ω un ouvert de $\mathbb{R}^n$ et $\Phi : \Omega \longrightarrow \mathbb{R}^n$ de classe $\mathcal{C}^1$. Soit $a \in \Omega$ un point fixe de Φ, i.e. $\Phi(a) = a$. Alors :*

a) Si $\rho\Big(D\Phi(a)\Big) < 1$ alors a est asymptotiquement stable.

b) Si $\rho\Big(D\Phi(a)\Big) > 1$ alors a est instable.

Rappel 2 $D\Phi(a) \in M_n(\mathbb{R})$ est définie par :

$$D\Phi(a) = \left(\frac{\partial \Phi_i}{\partial x_j}(a)\right)_{1\leq i,j\leq n} \left.\begin{array}{c} \\ \\ \end{array}\right\} \begin{array}{c} \text{différentielle de } \Phi \text{ au point } a, \\ \text{matrice Jacobienne de } \Phi \text{ au point } a \end{array}$$

Preuve 7 (du a)) Notons $x_k = a + e_k$.

$$\left\{\begin{array}{c} x_{k+1} = \Phi(x_k) \\ a = \Phi(a) \end{array}\right. \quad \Longrightarrow \quad e_{k+1} = \Phi(a + e_k) - \Phi(a)$$

On utilise un développement de Taylor à l'ordre 1 :

$$\Phi(a + e_k) = \Phi(a) + D\Phi(a)e_k + \|e_k\|\,\varepsilon(e_k)$$

avec $\|\varepsilon(e_k)\| \to 0$ quand $e_k \to 0$.

Donc $e_{k+1} = D\Phi(a)e_k + o(\|e_k\|)$.

Si $\rho\big(D\Phi(a)\big) < 1$, il exsite une norme matricielle induite pour laquelle $\|D\Phi(a)\| < 1$.

Donc $\exists \eta > 0$ et $\alpha < 1$ tels que si $\|e_k\| < \eta$:

$$\|e_{k+1}\| \leq \alpha\,\|e_k\|$$

Donc si $\|e_0\| < \eta$, $\|e_k\| \leq \alpha^k\,\|e_0\| \xrightarrow{k\to+\infty} 0$

Remarque 14

- Ce résultat donne la convergence <u>locale</u> de la méthode : convergence de (x_k) vers un point fixe a de Φ si $\rho\big(D\Phi(a)\big) < 1$ et $\|x_0 - a\|$ assez petit.
- La solution de (5.1) n'est pas forcément unique.
- a) $\Longrightarrow$ $\|x_{k+1} - a\| \leq \alpha\,\|x_k - a\|$ avec $\alpha < 1$, et plus $\rho\big(D\Phi(a)\big)$ est petit, plus α est petit. On dit que la convergence est (au moins) <u>linéaire</u>.
- Sous l'effet des termes non linéaires, dans certains cas la méthode numérique (5.2) peut être localement convergente avec $\rho\big(D\Phi(a)\big) = 1$. <u>Exemple :</u> $x_{k+1} = x_k - x_k^3$, point fixe 0 asymptotiquement stable.

Critères d'arrêt :

a) On se donne une tolérance absolue *tol* (on pourrait aussi travailler en relatif)

$$\|x_k - x_{k-1}\| < tol$$

Cela indique également que $\|\Phi(x_{k-1} - x_{k-1})\| < tol$, c'est-à-dire que x_{k-1} est "presque" solution de $\Phi(x) = x$.

b) Lorsque Φ est une contraction sur un sous-ensemble fermé E de de $\mathbb{R}^n$, on sait que Φ admet un unique point fixe a dans E. Si $\alpha \in]0,1[$ désigne le facteur de contraction de Φ on montre que si $x_{k-1} \in E$ alors $\|x_k - a\| \leq \frac{\alpha}{1-\alpha}\|x_k - x_{k-1}\|$.

Fixer le critère d'arrêt $\|x_k - x_{k-1}\| < tol \times \left(\frac{1}{\alpha} - 1\right)$ et $x_{k-1} \in E$ garantit que $\|x_k - a\| < tol$.

c) Un critère intéressant peut être obtenu lorsque :

$$\frac{\|x_k - x_{k-1}\|}{\|x_{k-1} - x_{k-2}\|} \xrightarrow{k\to+\infty} \lambda \in]0,1[$$

Cette propriété est vérifiée avec $\lambda = \rho\big(D\Phi(a)\big)$ et pour presque toute condition initiale $x_0 \approx 0$ si λ <u>ou</u> $-\lambda$ est une valeur propre réelle simple de $D\Phi(a)$, avec toutes les autres valeurs propres de module $< \lambda$.

(alors $x_k = a + V.(\pm\lambda)^k + o(\lambda^k)$, V vecteur propre associé à $\pm\lambda$)

Alors pour k assez grand et $p \geq k$

$$\|x_k - x_p\| \leq \|x_k - x_{k+1}\| + \|x_{k+1} - x_{k+2}\| + \cdots + \|x_{p-1} - x_p\|$$

$$\implies \|x_k - a\| \leq \sum_{j\geq k} \|x_j - x_{j+1}\| \qquad \text{(on fait tendre } p \text{ vers } +\infty)$$

On fait maintenant l'approximation :

$$\begin{aligned}\sum_{j\geq k} \|x_j - x_{j+1}\| &\simeq \|x_k - x_{k+1}\| \times \sum_{j\geq 0} \lambda^j \simeq \frac{\lambda}{1-\lambda}\|x_k - x_{k-1}\| \\ &\simeq \frac{\|x_k - x_{k-1}\|}{\|x_{k-1} - x_{k-2}\|} \times \frac{1}{1 - \frac{\|x_k - x_{k-1}\|}{\|x_{k-1} - x_{k-2}\|}}\|x_k - x_{k-1}\| \\ &= \frac{\|x_k - x_{k-1}\|^2}{\|x_{k-1} - x_{k-2}\| - \|x_k - x_{k-1}\|}\end{aligned}$$

On en déduit le critère d'arrêt :

$$\begin{cases} \dfrac{\|x_k - x_{k-1}\|^2}{\|x_{k-1} - x_{k-2}\| - \|x_k - x_{k-1}\|} < tol \\ \\ \|x_k - x_{k-1}\| < \|x_{k-1} - x_{k-2}\| \end{cases} \qquad \text{(c)}$$

Le théorème de convergence de la méthode des approximations successives suppose que $\rho\big(D\Phi(a)\big) < 1$. Un choix tel que $\Phi(x) = x - Bf(x)$ ($B \in M_n(\mathbb{R})$ inversible) ne garantit pas que cette hypothèse soit respectée, et que le rayon spectral soit petit (condition pour que la convergence soit rapide).

Nous allons définir un choix astucieux de fonction Φ à partir de f, pour lequel $D\Phi(a) = 0$. Il s'agit de la méthode de Newton.

5.2 Méthode de Newton

Soit $f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ de classe $\mathcal{C}^2$. On veut calculer numériquement une solution de l'équation :

$$f(x) = 0 \tag{5.3}$$

Le principe de la méthode de Newton est le suivant. Si $x_0 \in \mathbb{R}^n$ est une approximation de la solution x recherchée, on linéarise f autour de x_0 :

$$f(x) \simeq f(x_0) + Df(x_0)(x - x_0)$$

On calcule alors la solution x_1 de :

$$f(x_0) + Df(x_0)(x_1 - x_0) = 0$$

Si $Df(x_0)$ est inversible, on obtient :

$$x_1 = x_0 - Df(x_0)^{-1} f(x_0)$$

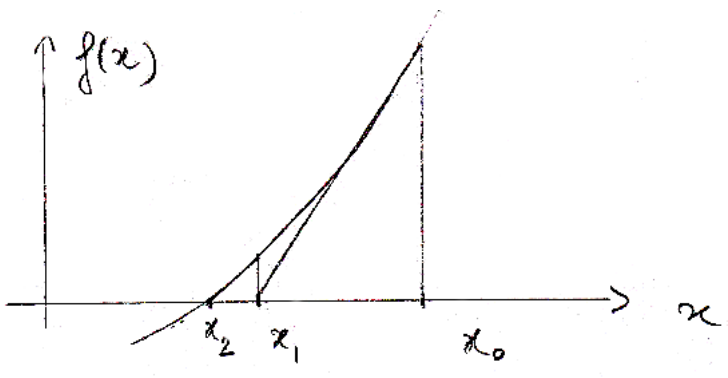
Puis on prend x_1 comme nouvelle approximation de la solution et on recommence l'opération. Cela définit la méthode itérative :

$$x_{k+1} = x_k - Df(x_k)^{-1} f(x_k) \underset{\text{déf}}{=} \Phi(x_k) \tag{5.4}$$

Remarque 15 Numériquement on ne calcule pas $Df(x_k)^{-1}$ mais on résout à chaque étape le système linéaire donnant x_{k+1} :

$$Df(x_k)(x_{k+1} - x_k) = -f(x_k)$$

Interprétation géométrique en dimension 1 :



Nous sommes dans le cadre de la méthode des approximations successives : on cherche une solution de $\Phi(x) = x$ avec $\Phi(x) = x - Df(x)^{-1}f(x)$. La méthode itérative (5.4) s'écrit $x_{k+1} = \Phi(x_k)$.

Théorème 13 *Soit $f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ de classe $\mathcal{C}^2$ au voisinage de $a \in \mathbb{R}^n$, avec $f(a) = 0$. On suppose que $Df(a)$ est inversible. Alors la fonction Φ de l'itération (5.4) est $\mathcal{C}^1$ au voisinage de a, et a est asymptotiquement stable. De plus, il existe $\eta > 0$ et $\alpha > 0$ tels que si $\|x_0 - a\| < \eta$ alors :*

$$\|x_{k+1} - a\| \leq \alpha \|x_k - a\|^2 \qquad \forall k \geq 0$$

Remarque 16 On dit qu ela convergence de la méthode de Newton est en moyenne quadratique. On obtient par récurrence :

$$\|x_k - a\| \leq \frac{1}{\alpha}(\alpha \|x_0 - a\|^{2^k}$$

Par exemple, si $\alpha = 1$ et $\|x_0 - a\| = 10^{-1}$, $\|x_4 - a\| \leq 10^{-16}$.

Preuve 8 La fonction $x \longmapsto \operatorname{Det} Df(x)$ est continue sur $\mathbb{R}^n$, et $\operatorname{Det} Df(a) \neq 0$, donc $\exists r > 0 \ / \ \|x - a\| < r \implies \operatorname{Det} Df(x) \neq 0$, c'est-à-dire que $Df(x)$ est inversible. La fonction Φ définie par $\Phi(x) = x - Df(x)^{-1}f(x)$ est donc $\mathcal{C}^1$ au voisinage de $x = a$. On peut donc appliquer le théorème 1.

Calculons $D\Phi(a)$.

$f(a+h) = Df(a)h + \mathcal{O}(\|h\|^2)$ amène :

$$\begin{aligned}
\Phi(a+h) &= a + h - Df(a+h)^{-1}\Big(Df(a)h + \mathcal{O}(\|h\|^2)\Big) \\
&= a + h - \Big(Df(a)\mathcal{O}(\|h\|\Big)\Big(Df(a)h + \mathcal{O}(\|h\|^2)\Big) \\
&= a + h - \Big(I + \mathcal{O}(\|h\|)^{-1}\Big)Df(a)^{-1}\Big(Df(a)h + \mathcal{O}(\|h\|^2)\Big) \\
&= a + h - \Big(I + \mathcal{O}(\|h\|\Big)\Big(h + \mathcal{O}(\|h\|^2)\Big) \\
\Phi(a+h) &= \Phi(a) + \mathcal{O}(\|h\|^2)
\end{aligned}$$

Donc :

$D\Phi(a) = 0 \implies a$ est un point fixe de Φ asymptotiquement stable

Avec $x_k = a + e_k$ on obtient $e_{k+1} = \Phi(a + e_k) - \Phi(a) = \mathcal{O}(\|e_k\|^2)$

Remarque 17 - Lorsque la forme analytique de $Df(x)$ est inconnue, on approche $\dfrac{\partial f_i}{\partial x_j}$ par $\dfrac{f_i(x_1, \ldots, x_{j-1}, x_{j+\delta}, x_{j+1}, \ldots, x_n) - f_i(x_1, \ldots, x_n)}{\delta}$ avec $\delta \approx 0$.

- Comme précédemment, le théorème 2 donne la convergence locale de la méthode de Newton, i.e. pour une condition suffisamment proche d'un point fixe a.

- Avantage de Newton : convergence très rapide (quadratique).

- Inconvénient de Newton : coût très élevé à chaque étape, car il faut calculer à chaque fois $A_k = Df(x_k)$ et résoudre un système linéaire $A_k(x_{k+1} - x_k) = -f(x_k)$ (coût en $\mathcal{O}(n^3)$, cf méthode de Gauss).

$\Longrightarrow$ plusieurs modifications de la méthode ont été proposées. Nous verrons par exemple en TD la méthode de Broyden très employée.

Voici une autre modification (plus simple mais efficace) de Newton :

$$\begin{cases} x_{k+1} = \Phi(x_k) & \Phi(x_k) = x_k - A^{-1}f(x_k) \\ x_0 \in \mathbb{R}^n & A = Df(x_0) \end{cases}$$

On calcule une seule fois la factorisation LU de la matrice A, et on l'utilise à chaque étape pour résoudre $A(x_{k+1} - x_k) = -f(x_k)$ (coût $\mathcal{O}(n^2)$ pour $k \geq 2$).

L'inconvénient est bien sûr qu'on perd la convergence quadratique pour une convergence uniquement linéaire. En effet, si $f(a) = 0$, $D\Phi(a) = I - A^{-1}Df(a) \approx 0$ si $x_0 \approx a$, mais $\rho\Big(D\Phi(a)\Big) \neq 0$ en général.

Ce schéma se généralise en remplaçant A par $Df(x_k)$ toutes les "quelques itérations".

Chapitre 6

Équations différentielles à condition initiale

6.1 Problème de Cauchy

$f : [a, b] \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ de classe $\mathcal{C}^0$

$$\boxed{\begin{array}{rl} y'(t) = & f(t, y(t)) \\ y(t_0) = & y_0 \\ t \in [a, b] & \end{array}}$$

Problème différentiel de condition initiale (problème de Cauchy)

où $[a, b] \subset \mathbb{R}$

$y : t \in [a, b] \longrightarrow y(t) \in \mathbb{R}^n$ application dérivable inconnue

$f : (t, \theta) \in [a, b] \times \mathbb{R}^n \longrightarrow f(t, \theta)$ application donnée

y_0 : valeur initiale donnée

Résoudre le problème c'est donc déterminer une application y, si elle existe, qui est solution de l'équation différentielle $y'(t) = f(t, y(t))$ et qui prend la valeur numérique donnée y_0 à l'instant initial $t = t_0$

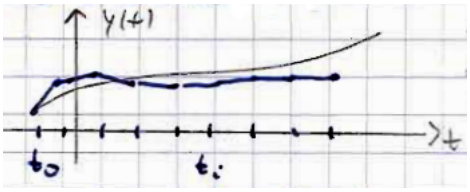
Domaines d'application :

- mécanique
- système solaire ($\rightarrow$ modélisation par des lois de Newton, $n \geq 3$ pas de solution analyitique)
- cinétique chimique ($\rightarrow$ réaction chaotiques)
- météo ($\rightarrow$ évolution des champs de pression à la surface de la Terre, turbulences)
- Animation d'objets 3D par ordinateur

Résolutions numériques des EDO

* Les erreurs de troncature ne sont pas négligeables quand on a beaucoup d'itérations

* Convergence



Calcul de n valeurs approchées

* Stabilité : petites perturbations à l'entrée restent bornées à la sortie ?

* Consistance : erreur locale doit tendre vers 0 pour $h \to 0$.

* Chaos : une faible perturbation sur les CI peut entraîner une divergence de la solution.

Théorème 14 (Cauchy-Lipschitz : Existence + Unicité) *Soit* $f : [a,b] \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ *de classe* $\mathcal{C}^0$ *et vérifiant la propriété suivante :*

Il existe une constante $L \in \mathbb{R}$ *telle que :*

$$\forall t \in [a,b], \forall y_1, y_2 \in \mathbb{R}^n, \qquad \|f(t,y_1) - f(t,y_2)\| \leq L \|y_1 - y_2\|$$

Alors quelque soit $t_0 \in [a,b]$ *et* $y_0 \in \mathbb{R}^n$, *il existe une unique fonction* $y : [a,b] \longrightarrow \mathbb{R}^n$ *avec*

(i) $y(t)$ *est de classe* $\mathcal{C}^1$ *sur* $[a,b]$

(ii) $y'(t) = f(t, y(t))$ *pour* $t \in [a,b]$

(iii) $y(t_0) = y_0$

Dans la suite on se restreint au cas d'**une** équation ($n = 1$) différentielle.

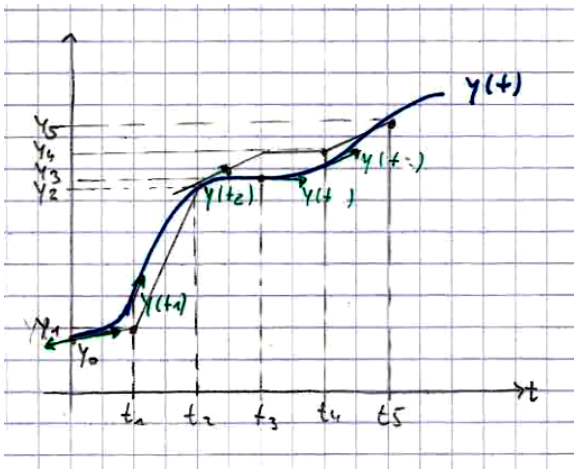
Exemple 1 (Méthode d'Euler) On subdivise $[a,b]$ en n intervalles de longueur $h = \dfrac{b-a}{N}$, $t_i = a + ih$, $i = 0, \ldots, N$

La méthode d'Euler consiste à calculer par récurrence des valeurs approchées $y_1, \ldots, y_N$ de $y(t_1), \ldots, y(t_N)$ respectivement au moyen de la formule suivante :

$$y_{k+1} = y_k + hf(t_k, y_k) \qquad k = 0, \ldots, N-1$$

L'idée de la méthode est alors de considérer que sur le petit intervalle $[t_0, t_{0+h}]$ la courbe n'est pas très éloignée de sa tangente en t_0

$$\begin{aligned} \frac{y(t_0+h) - y(t_0)}{h} &\approx f(t_0, y(t_0)) \\ y(t_0+h) &\approx y(t_0) + hf(t_0, y(t_0)) \\ y_1 &= y_0 + hf(t_0, y_0) \end{aligned}$$



Partant de $t_0 = a$, on connaît $y_0 = y(t_0)$ donc aussi la dérivée en ce point $f(t_0, y_0)$

Convergence :

On fixe un **point** $\tilde{t}$ et on regarde le comportement de l'erreur $e_k = y_k - y(t_k)$ en diminuant $h_j = \dfrac{b-a}{j}$ ($h \to 0$ pour $j \to +\infty$). On voudrait que les valeurs approximatives y_k tendent vers $y(t_k)$ si h tend vers 0 : $|y_k - y(t_k)| \leq C.h$

Théorème 15 (Convergence d'Euler) *On suppose vérifiées les hypothèses "Cauchy-Lipschitz" et que la solution y du problème de Cauchy appartient à $\mathcal{C}^2[a,b]$*

On pose $M_2 = \underset{t\in[a,b]}{\text{Max}}|y''(t)|$

Alors si on note $e_k = y_k - y(t_k)$ l'erreur au point t_k, on a la majoration

$$|e_k| \leq \underbrace{\frac{1}{L}(e^{L(b-a)} - 1)\frac{M_2}{2}h}_{= c \text{ ne dépend pas de } t_k}$$

$$|e_k| \leq C.h$$

Preuve 9 Comme $y \in \mathcal{C}^2[a,b]$ on a en particulier la formule de Taylor :

$$\begin{aligned} & y(t_k + h) = y_k(t_k) + hy'(t_k) + \frac{1}{2}h^2y''(\xi) \\ \iff & y(t_{k+1}) = y(t_k) + h + f(t_k, y(t_k)) + \frac{1}{2}h^2y''(\xi) \end{aligned} \qquad (*)$$

Soustrayons donc :

$$\begin{aligned} y_{k+1} =& y_k + hf(t_k, y_k) \\ & -(*) \end{aligned}$$

On obtient :

$$e_{k+1} = e_k + h\Big[f(t_k, y_k) - f(t_k, y(t_k))\Big] - \frac{1}{2}h^2y''(\xi)$$

En appliquant “Cauchy-Lipschitz” :

$$|e_{k+1}| \leq |e_k|(1+Lh) + \frac{1}{2}h^2 M_2 \qquad 0 \leq k \leq N-1 \qquad (**)$$

Lemme 4 Soit $(\varepsilon_k), k = 0, \ldots, N$ une suite de nombres positifs vérifiant :

$$\varepsilon_{k+1} \leq \varepsilon_k a + b \qquad k = 0, \ldots, N-1$$

Alors :

$$\varepsilon_k \leq \varepsilon_0 \; a^k + b\frac{a^k - 1}{a-1}$$

Preuve 10 Immédiate par récurrence.

Lemme 5 Pour tout $k \in \mathbb{N}$ et tout réel $u \geq u$ on a :

$$(1+u)^k \leq e^{ku}$$

Preuve 11 Il suffit de montrer que $1 + u \leq e^u$.

Posons $z(u) = 1 + u - e^u$, on a $z'(u) = 1 - e^u$

z' est négatif et donc z est décroissant sur $[0, +\infty[$

Or $z(0) = 0 \implies 1 + u - e^u \leq 0 \implies$ d'où le résultat.

Lemme 5 appliqué à (**) donne (puisque $e_0 = 0$, CI) :

$$|e_k| \leq 0.(1+Lh)^k + \frac{h^2 M}{2}\frac{(1+Lh)^k - 1}{Lh} = \frac{4M2}{2L}$$

On applique alors 11 :

$$|c_k| \leq \frac{e^{Lhk} - 1}{L}.\frac{M_2}{2}h$$

et $kh = t_k - a \leq b - a$ donc :

$$|e_k| \leq \frac{e^{L(b-a)} - 1}{L}.\frac{M_2}{2}h$$

On a donc la convergence, mais qu'en est-il de la stabilité et de la consistance ?

2 classes de méthodes :

(1) À pas séparé **MPS** : y_{k+1} approximation de $y(t_{k+1})$ est calculé à partir de y_k

(2) À pas multiple **MPM** : y_{k+1} est calculé à partir de plusieurs points précédents $y_k, y_{k-1}, \ldots, y_{k-p}$.

6.2 Méthodes à pas séparé

"Schéma à un pas"

6.2.1 Définition

$$\begin{cases} y'(t) = f(t, y(t)) \\ y(a) \text{ donné} \end{cases}$$

f continue, lipschitzienne par rapport à $y \implies \overline{y}$ solution uique au problème de Cauchy.

But :

Approcher $\overline{y}(t_k)$ aux points $t_k = a + kh$

$h = \dfrac{b-a}{N}$ pas constant, $t_0 = a, t_n = b$, $k = 0, \ldots, N$

Définition 5 (MPS) Une **MPS** est un schéma itératif de la forme :

$$y_{k+1} = y_k + h\ \Phi(t_k, y_k, h)$$

y_0 donné : $y_0 = \overline{y}(a)$

$t_{k+1} = t_k + h$

Exemple 2 Euler $\Phi(t, y, h) = f(t, y)$

Ici Φ est indépendant de k.

On dira que Φ définit la MPS. On appelle **erreur** au point t_k : $e_k = y_k - \overline{y}(t_k)$

Le but est de construire des MPS (i.e. Φ) telles que

$$\text{Max}\,|e_k| \xrightarrow{h \to 0} 0$$

i.e.

$$\text{Max}\,|e_k| = \mathcal{O}(h^p) \qquad p \in \mathbb{N}$$

Plus p est grand, plus la méthode converge vite.

6.2.2 Consistance, stabilité et convergence

Définition 6 On dit que la **MPS** est **convergente** si :

$$\forall y_0 \in \mathbb{R}, \quad \underbrace{\lim_{h \to 0}}_{\text{erreur en un pt } t_k \to 0} \quad \text{Max}_{k \in \{1,..,N\}} |y_k - \overline{y}(t_k)| = 0$$

Remarque 18 On peut même aller plus loin dans la définition de la convergence en ne supposant pas que le schéma part de la condition initiale exacte.

Autrement dit, la méthode doit converger même s'il y a une erreur (de troncature ...) sur la condition initiale.

Ce qui donne la :

Définition 7 La MPS est **convergente** si :

$$\lim_{y_0 \to \overline{y}(a), h \to 0} \text{Max}_k |y_k - \overline{y}(t_k)| = 0$$

On verra maintenant que la convergence résulte de deux propriétés : **stabilité** et **consistance**.

→ La **stabilité** est une propriété propre au schéma. Elle assure que le schéma n'amplifie pas trop les erreurs (numériques) que l'on commet à chaque pas.

Schéma exactement calculé : $y_{k+1} = y_k + h\ \Phi(t_k, y_k, h)$

$\neq$ schéma calculé par ordinateur :

$$\begin{cases} z_0 = y_0 + \varepsilon_0 \\ z_{k+1} = z_k + h\big[\Phi(t_k, z_k, h) + \varepsilon_k\big] \end{cases}$$

Définition 8 La méthode **MPS** est <u>stable</u> si : $\exists M > 0, \exists \overline{\varepsilon} > 0$ t.q.

$$\forall h, \forall \varepsilon_i < \overline{\varepsilon} : \underset{k}{\text{Max}} |y_k - z_k| < M. \underset{i \in \{0,..,M-1\}}{\text{Max}} |\varepsilon_i|$$

→ La **consistance** définit une relation entre le schéma et l'équation différentielle. Elle implique que le schéma s'écarte peu localement de la solution.

Définition 9 Une **MPS** est dite **consistance** avec l'équation différentielle si

$$\left| \underbrace{\frac{\overline{y}(t+h)-\overline{y}(t)}{h}}_{\Delta(t,\overline{y}(t),h)} - \Phi(t,\overline{y}(t),h) \right| \underset{N\to+\infty}{\overset{h\to 0}{\longrightarrow}} 0$$

Autrement dit, si l'on veut que le schéma marche, il faut au minimum qu'il soit à peu près vérifié par la solution formelle $\overline{y}$ quand h est assez petit $\iff \overline{y}(t+h) - \overline{y}(t) - h\ \Phi(t,\overline{y}(t),h) = \mathcal{O}(h) \leq L.h$

Théorème 16 (Th. Fondamental)

Stabilité + Consistance $\implies$ *Convergence.*

Preuve 12

$$y_{k+1} = y_k + h\ \Phi(t_k, y_k, h)$$

Idée : considérer la solution exacte $\overline{y}$ comme une perturbation de la solution numérique !

$$\begin{aligned}\overline{y}(t_{k+1}) &= \overline{y}(t_k) + h\ \Delta(t_k, \overline{y}(t_k), h) \\ &= \overline{y}(t_k) + h\ \Phi(t_k, \overline{y}(t_k), h) + \varepsilon_k\end{aligned}$$

On pose $\varepsilon_k = [\Delta - \Phi]_k$

$$\begin{aligned} z_k &= \overline{y}(t_k) & y_0 &= z_0 \\ z_{k+1} &= \overline{y}(t_{k+1}) \end{aligned}$$

$$\implies \varepsilon_k = \frac{\overline{y}(t_{k+1} - \overline{y}(t_k)}{h} - \Phi(t_k, \overline{y}(t_k), h)$$

Hypothèse de consistance $\implies |\varepsilon_k| \overset{h\to 0}{\longrightarrow} 0$

Hypothèse de stabilité $\implies \exists M, \exists \overline{\varepsilon} > 0$ t.q $\forall \varepsilon_i < \overline{\varepsilon}$:

$$\text{Max}\,|y_k - \overline{y}(t_k)| < M.\underset{k}{\text{Max}}\,|\varepsilon_k| \overset{h\to 0}{\longrightarrow} 0$$

Donc $\text{Max}\,|y_k - \overline{y}(t_k)| \overset{h\to 0}{\longrightarrow} 0$

Donc **MPS est convergente**.

Remarque 19 Réduire la démonstration de la convergence à la vérification de la consistance et de la stabilité a un double avantage :

– Un schéma stable qui n'est pas consistant calcule bien quelque chose, mais pas ce que l'on cherche.

– Un schéma instable mais consistant calcule une solution qui peut être proche initialemet de ce que l'on cherche, mais qui s'éloigne rapidement (souvent de façon oscillante).

6.2.3 Caractérisation de la consistance et de la stabilité

On suppose que Φ est continue en $t \in [a,b], y \in \mathbb{R}$, et h, en $h = 0$.

Proposition 1 MPS est consistante $\iff \Phi(t,y,0) = f(t,y)$.

Preuve 13 " $\Longrightarrow$ " MPS consistante

$$\Longrightarrow \left| \underbrace{\frac{\overline{y}(t+h) - \overline{y}(t)}{h}}_{\overline{y}'(t)+o(t)} - \Phi(t, \overline{y}(t), h) \right| \overset{h\to 0}{\longrightarrow} 0 \qquad (*)$$

Soit $\varepsilon > 0$:

$$\begin{aligned} \overline{y} \in \mathcal{C}^1 &\Longrightarrow \exists h_0 : \forall h < h_0 \left| \frac{\overline{y}(t+h) - \overline{y}(t)}{h} - \overline{y}'(t) \right| < \frac{\varepsilon}{2} \\ &\Longrightarrow \frac{\overline{y}(t+h) - \overline{y}(t)}{h} - \frac{\varepsilon}{2} < f(t, y(t)) < \frac{\overline{y}(t+h) - \overline{y}}{h} + \frac{\varepsilon}{2} \qquad (*2) \end{aligned}$$

$$\begin{aligned} (*) &\Longrightarrow \exists h_1 : \forall h < h_1 \left| \frac{\overline{y}(t+h) - \overline{y}(t)}{h} - \Phi(t, \overline{y}(t), h) \right| < \frac{\varepsilon}{2} \\ &\Longrightarrow \frac{\overline{y}(t+h) - \overline{y}(t)}{h} - \frac{\varepsilon}{2} < \Phi(t, \overline{y}(t), h) < \frac{\overline{y}(t+h) - \overline{y}(t)}{h} + \frac{\varepsilon}{2} \qquad (*3) \end{aligned}$$

$h_2 := \text{Min}(h_0, h_1),\ \forall h < h_2$

$(*2) - (*3) \Longrightarrow |f(t, \overline{y}(t) - \Phi(t, \overline{y}(t), h| < \varepsilon$

$$\Longrightarrow \text{Pour } h \to 0 : f(t, y(t)) = \Phi(t, \overline{y}(t), 0), \text{ car } \Phi \text{ continue.} \qquad (*4)$$

Maintenant, il faut montrer que l'égalité est vraie $\forall y$. (*4) est vraie pour tout y qui sont solution du problème de Cauchy. On peut donc appliquer (*4) à l'unique solution du problème de Cauchy

$$\begin{cases} y(t_0) = y_0 \\ y'(t) = f(t, y(t)) \end{cases}$$

$\Longrightarrow$ On trouve $f(t_0, y_0, 0) = f(t_0, y_0),\ \forall t_0, y_0$

"$\Longleftarrow$" : $\Phi(t,y,0) = f(t,y)$

Comme Φ est continue en h :

$$\Phi(t,y,h) \overset{h\to 0}{\longrightarrow} \Phi(t,y,0) = f(t,y) \tag{6.1}$$

Comme $\overline{y} \in \mathcal{C}^1$:

$$\frac{\overline{y}(t+h) - \overline{y}(t)}{h} \overset{h\to 0}{\longrightarrow} \overline{y}'(t) = f(t,y) \tag{6.2}$$

$$(6.1) - (6.2) \implies \frac{\overline{y}(t+h) - \overline{y}(t)}{h} - \Phi(t,\overline{y},h) \overset{h\to 0}{\longrightarrow} 0$$

Donc la **MPS est consistante**.

Proposition 2 Si Φ est continue et lipschitzienne par rapport à y, alors

MPS est stable.

Lemme 6 Soit (a_n) la suite vérifiant :

$$a_{n+1} \leq (1+A)a_n + B \qquad (A, B > 0)$$

Alors

$$\forall n : a_n \leq a_0\, e^{nA} + \frac{e^{nA} - 1}{A} B$$

Preuve 14 Soit $y_{k+1} = y_k + h\, \Phi(t, y_k, h)$, y_0 donné.

k fixé $\in [0, .., N-1]$, N fixé.

(z_k) le schéma perturbé par (ε_k)

$$\begin{aligned} |y_{k+1} - z_{k+1}| &= \left|y_k + z_k - h\, \left[\Phi(t, y_k, h) - \Phi(t, z_k, h)\right] - h\, \varepsilon_k\right| \\ &\leq (1 + hL)\, |y_k - z_k| + h\, |\varepsilon_k| \\ &\leq (1 + hL)\, |y_k - z_k| + h \underset{j\in[0,..,N-1]}{\text{Max}} |\varepsilon_j| \end{aligned}$$

On applique le lemme avec $A = hL$, $B = h \underset{j\in[0,..,N-1]}{\text{Max}} |\varepsilon_j|$

$$\leq e^{khL}\, |y_0 - z_0| + \frac{e^{khL} - 1}{hL} h \,\text{Max}\, |\varepsilon_j|$$

Or $kh \leq N.h = |b - a|$

$$\leq \underbrace{(e^{L(b-a)} + \frac{e^{L(b-a)} - 1}{L})}_{\text{constante } M > 0} \operatorname{Max}_{j} |\varepsilon_j|$$

$$\Longrightarrow |z_{k+1} - z_{k+1}| < M \operatorname*{Max}_{j \in [0,..,N-1]} |\varepsilon_j| \text{ indép de k}$$

$$\Longrightarrow |y_k - z_k| < \operatorname*{Max}_{j \in [0,..,N-1]} |\varepsilon_j|$$

$$\Longrightarrow \text{MPS stable}$$

Théorème 17 *Si Φ est continue, lipschitzienne par rapport à y et vérifie $\Phi(t, y, 0) = f(t, y)$ alors :*

(i) $\forall$ CI, il y a une solution unique.

(ii) La MPS_Φ converge.

Preuve 15 (i) Φ consistante $\Longleftrightarrow$ $\Phi(t, y, 0) = f(t, y)$. Donc comme Φ est continue et lipschitzienne, f l'est aussi $\Longrightarrow$ conditions de Cauchy sur f.

(ii) On a :

$$\left.\begin{array}{c}\text{Consistance (Prop. 1)} \\ \text{Stabilité (Prop. 2)}\end{array}\right\} \overset{\text{Th.2}}{\Longrightarrow} \text{Convergence}$$

6.2.4 Ordre d'un schéma à un pas

Il ne suffit pas qu'un schéma converge, il faut aussi qu'il converge suffisamment vite pour être intéressant en pratique.

Définition 10 MPS est dite **d'ordre p** ($p \in \mathbb{N}$) si et seulement si :

$$\left| \frac{\overline{y}(t+h) - \overline{y}(t)}{h} - \Phi(t, \overline{y}(t), h) \right| = \mathcal{O}(h^p)$$

Théorème 18 *Si f vérifie les conditions de Cauchy et si MPS_Φ est d'ordre p et stable, alors :*

$$\operatorname{Max} |y_k - \overline{y}(t_k) < c.h^p|$$

La MPS_Φ est convergente d'ordre p.

6.2.5 Exemples de MPS

Méthode d'Euler :

$\Phi(t, y, h) = f(t, y)$

$\Phi(t, y, 0) = f(t, y) \implies$ consistance

Comme f est lipschitzienne par rapport à $y \implies \Phi$ lipsch/y $\implies$ stabilité

$\implies$ convergence de la méthode d'Euler

Convergence d'ordre 1.

Méthode d'Euler-Cauchy

Dans cette méthode, on introduit un "étage" supplémentaire, en effectuant 2 évaluations de y' en 2 pas de taille $\frac{h}{2}$.

Un pas d'itération (? ? ?) normal h avec Euler donne la valeur :

$$y_{k+1}^{(1)} = y_k + h\ f(t_k, y_k)$$

2 pas avec $\frac{h}{2}$ donnent les 2 valeurs successives :

$$y_{k+\frac{1}{2}}^{(2)} = y_k + \frac{h}{2} f(t_k, y_k)$$
$$y_{k+1}^{(2)} = y_{k+\frac{1}{2}}^{(2)} + \frac{h}{2} f(t_k + \frac{h}{2}, y_{k+\frac{1}{2}}^{(2)})$$

Et on obtient la valeur finale :

$$\begin{aligned} y_{k+1} &= 2y_{k+1}^{(2)} - y_{k+1}^{(1)} \\ &= y_k + h\ f(t_k + \frac{h}{2}, y_k + \frac{h}{2} f(t_k, y_k)) \end{aligned}$$

par l'extrapolation de Richardson.

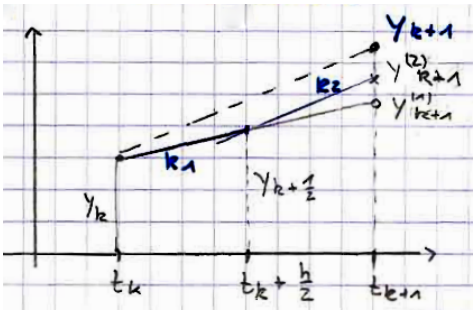
Algorithme

$$\boxed{\begin{aligned} k_1 &= f(t_k, y_k) \\ k_2 &= f(t_k + \frac{h}{2}, y_k + \frac{h}{2} k_1) \\ y_{k+1} &= y_k + h\ k_2 \end{aligned}}$$

Ce schéma, aussi appelé Euler modifié exige pour 1 pas d'itération (? ? ?) 2 évaluations de $f(t, y)$ en 2 points différents.

- k_1 : détermine la pente au départ en t_k pour trouver le point auxiliaire $(t_k + \frac{h}{2}, y_{k+\frac{1}{2}}^{(2)})$
- k_2 : est la pente en $(t_k + \frac{h}{2}, y_{k+\frac{1}{2}}^{(2)})$ et permet de trouver le point (t_{k+1}, y_{k+1}) en corrigeant la trajectoire.

La méthode d'Euler-Cauchy est un **schéma d'ordre 2**.



COURS INCOMPLET.

Chapitre 7

Optimisation sans contrainte

Étant donné $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ suffisamment régulière (typiquement $\mathcal{C}^2$), nous allons étudier d'un point de vue analytique et numérique l'existence d'extrema de f (un extremum = un minimum puis un maximum). On parle d'optimisation **sans contrainte** (un problème d'optimisation avec contrainte étant posé sur un sous-ensemble de $\mathbb{R}^n$). Il s'agira pour nous de trouver un **minimum** d'une fonction f, sous perte de généralité car le maximum d'une fonction $\tilde{f}$ revient à chercher un minimum de $f = -\tilde{f}$.

Exemples :

- Résolution d'un système linéaire $Ax = b$, avec A symétrique définie positive. Exemple d'application : résolution de l'équation de Poisson par différences finies. Nous verrons que la solution du système minimise $f(x) = \frac{1}{2}\,{}^t xAx - {}^t bx$. Cette propriété permet d'introduire de nouvelles méthodes de résolution du système.

- Exemple de minimisation d'une fonction non quadratique :

$$f(x) = \frac{1}{2} \sum_{1 \leq i,j \leq N, i \neq j} V(\|x_i - x_j\|)$$

Cristal constitué de N atomes, de positions $x_i \in \mathbb{R}^3$, interagissent par paires via un potentiel V

f représente l'énergie potentielle totale du cristal. La forme d'équilibre du cristal à $T = 0$ Kelvin est donnée par un minimum de l'énergie potentielle f.

Sous certaines hypothèses sur v et en deux dimensions d'espace, il a été démontré très récemment (F. Theil, 2005) que ce minimum est atteint pour un arrangement périodique des atomes.

7.1 Quelques résultats de base en calcul différentiel et optimisation

7.1.1 Étude locale des fonctions à n variables

Définition 11 f est différentiable en $x \in \Omega$ s'il existe une application linéaire $T : \mathbb{R}^n \longrightarrow \mathbb{R}$ telle que pour $h \approx 0$:

$$f(x+h) = f(x) + T\ h + o(\|h\|)$$

L'application T est alors unique et on note $T = Df(x)$.

T est appelée différentielle de f au point x.

Remarque 20 Si f est différentiable en x, alors elle est continue en x.

Lemme 7 Si f est différentiable en x, alors $\dfrac{\partial f}{\partial x_1}(x), \ldots, \dfrac{\partial f}{\partial x_n}(x)$ existent et $Df(x)h = \Big(\dfrac{\partial f}{\partial x_i}, \ldots \dfrac{\partial f}{\partial x_n}(x)\Big)h$

Remarque 21

1. Par abus de lanage, on confond souvent l'application linéaire $Df(x)$ et sa matrice $\Big(\dfrac{\partial f}{\partial x_1}(x), \ldots \dfrac{\partial f}{\partial x_n}(x)\Big)$ appelée matrice Jacobienne de f au point x.
2. Le fait que $\dfrac{\partial f}{\partial x_1}(x), \ldots, \dfrac{\partial f}{\partial x_n}(x)$ existent n'implique pas que f est différentiable en x.
3. On appelle $\nabla f(x) = {}^t\Big(\dfrac{\partial f}{\partial x_1}(x), \ldots, \dfrac{\partial f}{\partial x_n}(x)\Big)$. Alors $Df(x)h = \nabla f(x).h$ où . est le produit scalaire usuel sur $\mathbb{R}^n$.

Définition 12 f est différentiable sur un ouvert Ω si elle est différentiable en tout point de Ω.

Définition 13 $f : \Omega \longrightarrow \mathbb{R}$ est $\mathcal{C}^1$ si f est différentiable sur Ω et si l'applictaion $x \mapsto \nabla f(x)$ est continue.

On peut montrer le résultat suivant :

Lemme 8 $f : \Omega \longrightarrow \mathbb{R}$ est $\mathcal{C}^1$ si est seulement si ses dérivées partielles $\dfrac{\partial f}{\partial x_i}(i = 1..n)$ existent et sont continues sur Ω.

Définition 14 $f : \Omega \longrightarrow \mathbb{R}$ est $\mathcal{C}^2$ si f est $\mathcal{C}^1$ sur Ω et ses dérivées partielles $\dfrac{\partial f}{\partial x_i}(i = 1, \ldots, n)$ sont $\mathcal{C}^1$ sur Ω.

Lemme 9 (de Schwarz) Soit $f : \Omega \longrightarrow \mathbb{R}$ de classe $\mathcal{C}^2$. Alors pour tout $x \in \Omega, \forall i, j = 1..n$:

$$\frac{\partial}{\partial x_i}\Big(\frac{\partial f}{\partial x_j}\Big)(x) = \frac{\partial}{\partial x_j}\Big(\frac{\partial f}{\partial x_i}\Big)(x)$$

Remarque 22 On notera ces dérivées $\dfrac{\partial^2 f}{\partial x_i \partial x_j}(x)$.

Théorème 19 (formule de Taylor à l'ordre 2) *Soit $f : \Omega \longrightarrow \mathbb{R}$ de classe $\mathcal{C}^2$. Pour tout $x \in \Omega$ et $h \approx 0$:*

$$f(x+h) = f(x) + \nabla f(x).h + \frac{1}{2}\,{}^t h.Hf(x).h + o(\|h\|^2)$$

avec $Hf(x) \in M_n(\mathbb{R})$ définie par :

$$\Big(Hf(x)\Big)_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x)$$

$Hf(x)$ est appelée matrice hessienne de f en x (autre notation : $H_f(x)$)

Remarque 23 - $Hf(x)$ est symétrique d'après le lemme de Schwarz.

- On appelle $D^2 f(x)$ (différentielle seconde de f en x) la forme bilinéaire symétrique $\mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$ définie par $D^2 f(x)(h, y) = {}^t h H f(x) y$.

Définition 15 - $f : \Omega \longrightarrow \mathbb{R}$ admet un minimum local en $x \in \Omega$ s'il existe un voisinage ouvert u de x tel que $f(x) \leq f(y), \forall y \in u$.

- f admet un maximum local en $x \in \Omega$ si $f(x) \geq f(y), \forall y \in u$

Remarque 24

- Supposons $f : \mathbb{R}^n \longrightarrow \mathbb{R}$. On parle de minimum ou maximum **global** lorsque $u = \mathbb{R}^n$.
- Lorsque les inégalités sont strictes (pour $y \neq x$- on parle de minimum ou de maximum **strict**.

Lemme 10 Soit $f : \Omega \longrightarrow \mathbb{R}$ de classe $\mathcal{C}^1$ (Ω ouvert de $\mathbb{R}^n$).

Si f admet un extremum local en $x \in \Omega$, alors $\nabla f(x) = 0$.

Remarque 25
- Faux en général si Ω n'est pas un ouvert (un extremum peut être atteint sur le bord de Ω sans que ∇f s'y annule.
- $\nabla f(x) = 0$ peut être résolu par exemple par la méthode de Newton.
- On peut avoir $\nabla f(x) = 0$ sans que f admette un extremum en x. Exemple : $f(x, y) = x^2 - y^2$ en $(x, y) = (0, 0)$.

Lemme 11 Soit $f : \Omega \longrightarrow \mathbb{R}$ de classe $\mathcal{C}^2$. On suppose qu'il existe $x \in \Omega$ te que $\nabla f(x) = 0$. Alors :

- Si les valeurs propres de $Hf(x)$ sont > 0, f admet un minimum local strict en x.
- Si les valeurs propres de $Hf(x)$ sont < 0, f admet un maximum local strict en x.
- Si les valeurs propres de $Hf(x)$ sont $\neq 0$ et pas toutes de même signe, f n'admet pas d'extremum au point x (x est appelé un "point selle").

Remarque 26 Si $Hf(x)$ n'est pas inversible, la nature du point x (extremum de f ou non) dépend des termes d'ordre supérieure donc le développement de Taylor de f en x. Exemple : $f(x, y) = x^2 \pm y^2$.

Lemme 12 Soit $f : \Omega \longrightarrow \mathbb{R}$ de classe $\mathcal{C}^1$.

Soit $x_0 \in \Omega$ tel que $\nabla f(x_0) \neq 0$. L'équation $f(x) = f(x_0)$ définit localement (pour $x \approx x_0$) une hypersurface S (de dimension $n-1$), qui admet un plan tangent en tout point $x \approx x_0$.

Le plan tangent à S en x_0 est orthogonal à $\nabla f(x_0)$.

Lemme 13 Sous les hypothèses précédentes, $\nabla f(x_0)$ est orienté dans le sens des valeurs de f croissantes. Plus précisément :

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f(x_0 + \varepsilon \, \nabla f(x_0))_{|\varepsilon=0} = \|\nabla f(x_0)\|_2^2 > 0$$

7.1.2 Conditions suffisantes pour l'existence et l'unicité d'un minimum

Voyons d'abord une condition suffisante pour l'existence d'un minimum.

Théorème 20 *Soit* $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ *continue et telle que* $\lim_{\|x\| \to +\infty} f(x) = +\infty$

Alors il existe $x \in \mathbb{R}^n$ *tel que* $f(x) \leq f(y) \forall y \in \mathbb{R}^n$.

Remarque 27 On dit que f admet un minimum global en x.

Preuve 16 Si $\|y\| \geq R$ avec R assez grand, $f(x) \geq f(0)$. Donc $\inf_{y \in \mathbb{R}^n} f(y) = \inf_{\|y\| \leq R} f(y) = f(x)$ avec $\|x\| \leq \mathbb{R}$, puisque la boule $\|y\| \leq R$ est compacte.

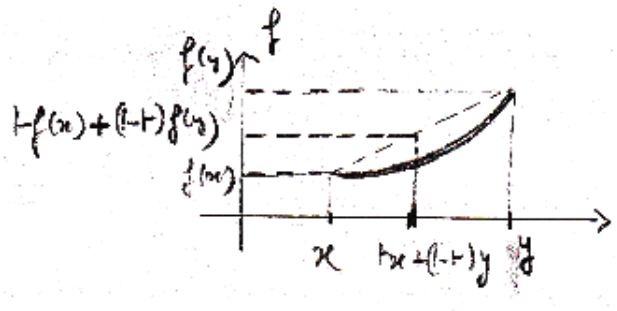
Le minimum de f peut ne pas être unique. Nous allons donner maintenant une condition suffisante d'unicité.

Définition 16 $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ est convexe si $\forall x, y \in \mathbb{R}^n, \forall t \in [0, 1]$

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

Remarque 28 Si $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ est convexe, alors elle est continue sur $\mathbb{R}^n$.

Interprétation en dimension 1 :



Définition 17 $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ est strictement convexe si $\forall x, y \in \mathbb{R}^n$ tels que $x \neq y, \forall t \in]0, 1[$,

$$f(tx + (1-t)y) < tf(x) + (1-t)f(y)$$

Théorème 21 *Si $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ est strictement convexe, il existe au plus un $x \in \mathbb{R}^n$ tel que $f(x) = \underset{y\in\mathbb{R}^n}{\text{Min}}\, f(y)$.*

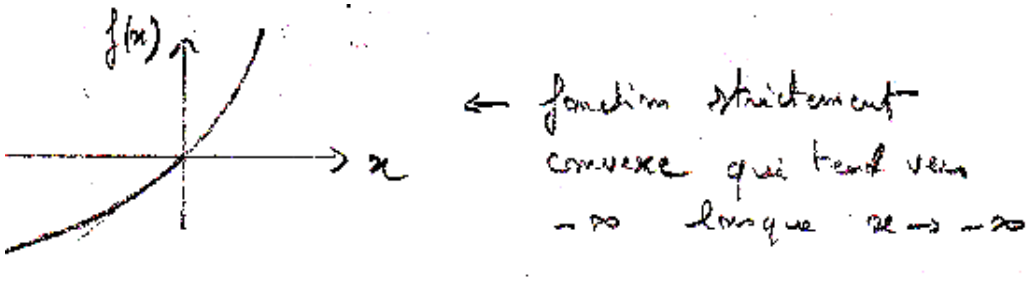
Preuve 17 Supposons l'existene de deux minima en x_1 et x_2. Alors

$$f(tx_1 + (1-t)x_2) < tf(x_1) + (1-t)f(x_2) = f(x_1) = \underset{x\in\mathbb{R}^n}{\text{Min}}\, f(x)$$

On arrive alors à une contradiction.

Remarque 29 Ce théorème ne donne pas l'existence d'un minimum.

Exemple :



Théorème 22 *Soit $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ strictement convexe et telle que $\lim\limits_{\|x\|\longrightarrow+\infty} f(x) = +\infty$. Alors il existe un unique $x \in \mathbb{R}^n$ tel que*

$$f(x) = \underset{y\in\mathbb{R}^n}{\text{Min}}\, f(y)$$

Nous allons maintenant relier les notions de point critique ($\nabla f(x) = 0$) et minimum pour les fonctions convexes.

Le résultat suivant fournit une caractérisation utile de la convexité pour les fonctions $\mathcal{C}^1$.

Lemme 14 Soit $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ de classe $\mathcal{C}^1$.

- f est convexe si et seulement si

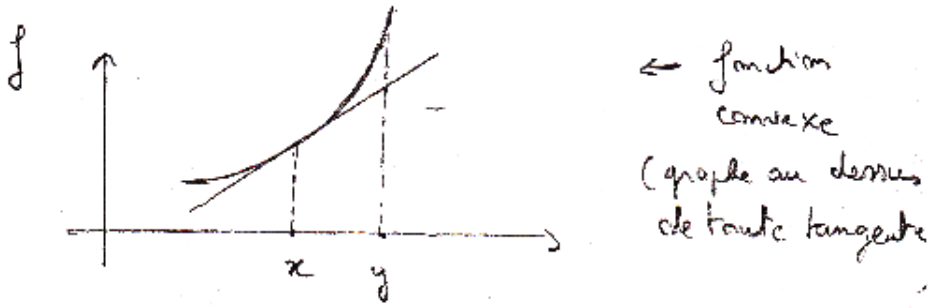
$$\forall x, y \in \mathbb{R}^n, \quad f(y) \geq f(x) + Df(x)(y-x)$$

- f est strictement convexe si et seulement si

$$\forall x, y \in \mathbb{R}^n \text{ avec } x \neq y, \quad f(y) > f(x) + Df(x)(y-x)$$

(résultat admis)

Interprétation en dimension 1 :



Théorème 23 *Soit $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ de classe $\mathcal{C}^1$ et convexe. Alors :*

$$f(x) = \operatorname*{Min}_{y\in\mathbb{R}^n} f(y) \iff \nabla f(x) = 0$$

Preuve 18 • " $\Longrightarrow$ " voir §**1.1**

• "$\Longleftarrow$" $f(y) \geq f(x) + Df(x)\underbrace{(y-x)}_{=0} \quad \forall y \in \mathbb{R}^n$

Remarque 30 On peut donc calculer numériquement les minima de fonctions convexes en recherchant les zéros de $x \mapsto \nabla f(x)$ (par exemple par la méthode de Newton).

On admettra la caractérisation suivante de la convexité pour des fonctions $\mathcal{C}^2$:

Lemme 15 Soit $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ de classe $\mathcal{C}^2$. Alors :

- f est convexe $\iff$ ${}^t y\, Hf(x)\, y \geq 0 \qquad \forall x \in \mathbb{R}^n, \forall y \in \mathbb{R}^n$
- Si $Hf(x)$ est symétrique définie positive $\forall x \in \mathbb{R}^n$ alors f est strictement convexe [1].

Remarque 31 Pour $f(x) = \frac{x^4}{12}$ (strictement convexe), $H_f(x) = x^2$.

$\Longrightarrow$ $H_f(0) = 0$ n'est pas symétrique définie positive.

Application

Soit $A \in M_n(\mathbb{R})$ avec A symétrique définie positive et $f(x) = \frac{1}{2}\,{}^t x\, A\, x - {}^t b\, x$

1. À vérifier, ce n'est pas lisible sur le kiosk

$$f(x) = \frac{1}{2}\sum_{i=1}^{n}\left(\sum_{j=1}^{n} a_{ij}x_j x_i\right) - \sum_{i=1}^{n} b_i x_i$$

$$\Big(Hf\Big)_{(i,j)} = \frac{1}{2}(a_{ij} + a_{ji}) \implies Hf = \frac{1}{2}(A + {}^tA) = A \text{ (symétrique définie positive)}$$

$$\implies f \text{ est strictement convexe.}$$

De plus, $f(x) \longrightarrow +\infty$ quand $\|x\| \longrightarrow +\infty$ car (λ désigne la plus petite valeur propre de A, qui est positive) :

$$f(x) \geq \frac{\lambda}{2}\|x\|_2^2 - \|b\|_2 \|x\|_2 \longrightarrow +\infty \text{ quand } \|x\|_2 \longrightarrow +\infty$$

Donc il existe un unique $x \in \mathbb{R}^n$ / $\underset{\mathbb{R}^n}{\text{Min}}\, f = f(x)$. Cette propriété est équivalente à $\nabla f(x) = 0$.

$$\frac{\partial f}{\partial x_i} = \frac{1}{2}\sum_{j=1}^{n}(a_{ij} + a_{ji})x_j - b_i \implies \nabla f(x) = \frac{1}{2}(A + {}^tA)x - b$$

Puisque A est symétrique, $\nabla f(x) = Ax - b$. Donc :

$$Ax = b \iff f(x) = \underset{\mathbb{R}^n}{\text{Min}}\, f, \text{ avec } f(x) = \frac{1}{2}\,{}^tx\,A\,x - {}^tbx$$

Cela permet de reformuler la résolution du système $Ax = b$ comme un problème de minimisation.

7.2 Quelques méthodes numériques pour l'optimisation sous contraintes :

Nous abordons maintenant le calcul numérique d'un minimum de $f : \mathbb{R}^n \longrightarrow \mathbb{R}$de classe $\mathcal{C}^1$. On suppose que $\lim\limits_{\|x\| \longrightarrow +\infty} f(x) = +\infty$, de sorte que ce minimum existe.

Nous allons d'abord voir des **méthodes de gradient**, qui sont des algorithmes itératifs utilisant uniquement f et ∇f.

L'exemple le plus simple d'une telle méthode est l'algorithme du **gradient à pas constant** :

$$\begin{cases} x_{k+1} = x_k - \rho\nabla f(x_k) & \rho > 0 \text{ fixé} \\ x_0 \in \mathbb{R}^n \text{ donné} \end{cases}$$

Cette méthode est motivée par la propriété que $-\nabla f$ est orienté dans le sens des valeurs de f décroissantes.

On dit alors que $-\nabla f(x_k)$ est une **direction de descente** en x_k.

La méthode du gradient à pas constant est assez peu utilisée en pratique car elle conduit facilement à des instabilités numériques. Par exemple, pour $f(x) = x^4$ (fonction

strictement convexe) on obtient $x_{k+1} = x_k(1 - 4\rho x_k^2)$. Si $x_0^2 \geq \frac{1}{\rho}$, on montre par récurrence que $|x_{k+1}| \geq 3|x_k|$ (car $1 - 4\rho x_k^2 \leq -3$) et donc $|x_k| \longrightarrow_{k\to+\infty} +\infty$.

Pour éviter ce type de phénomène, on peut considérer la **méthode de la plus grande pente** (ou steepest descent method) dans laquelle ρ est adapté à chaque itération de manière **optimale** :

$$\begin{cases} x_0 \in \mathbb{R}^n \text{ donné} \\ x_{k+1} = x_k - \rho_k \nabla f(x_k) \end{cases} \qquad f\left(x_k - \rho_k \nabla f(x_k)\right) = \underset{\rho \geq 0}{\text{Min}}\, f\left(x_k - \rho \nabla f(x_k)\right)$$

À chaque étape de l'itération, il faut donc résoudre un problème de minimisation en une dimension ; plus précisément minimser la fonction $\phi : \mathbb{R}^+ \longrightarrow \mathbb{R}$:

$$\rho \mapsto f(x_k - \rho \nabla f(x_k)) := \phi(\rho)$$

un minimum étant atteint en $\rho = \rho_k$ (le minimum existe sans être nécessairement unique) puisque $\lim\limits_{\|x\|\to+\infty} f(x) = +\infty$.

Calcul de ρ_k :

Il y a plusieurs possibilités.

- Méthode de Newton ou méthode de la sécante pour résoudre $\phi'(\rho) = 0$. Noter que c'est une condition nécessaire mais en général non suffisante pour obtenir un minimum.

  Cependant, si f est convexe alors ϕ est aussi convexe (c'est la restriction de f à une droite passant par x_k).

  Dans ce cas $\phi'(\rho) = 0 \iff \phi(\rho) = \underset{y\in]0,+\infty[}{\text{Min}}\, \phi(y)$

- Posons $a = 0$.

  On suppose $\phi : [a, b] \longrightarrow \mathbb{R}$ unimodale, c.à.d.

  $$\exists \rho^{\in}]a, b[\text{ tel que } \phi' < 0 \text{ sur }]a, \rho^*[\text{ et } \phi' > 0 \text{ sur }]\rho^*, b[.$$

  On pose $\delta = \dfrac{b-a}{4}, x_i = a + i\delta$.

  Selon la position relative des $f(x_i)$ $(i = 1, 2, 3)$ on peut choisir $a' < b'$ tels que f est unimodale sur $[a', b'] \subset [a, b]$ et $b' - a' = \frac{1}{2}(b - a)$. On recommence l'opération sur $[a', b']$ jusqu'à atteindre la précision souhaitée.

- **Cas particulier d'une fonction quadratique :**

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

  $A \in M_n(\mathbb{R})$ symétrique définie positive, $b \in \mathbb{R}^n$.

  Notons $r_k = \nabla f(x_k) = Ax_k - b \neq 0$ (sinon le min est déjà atteint !)

  $\phi'(\rho_k) = 0 \iff r_{k+1}.r_k = 0 \iff \underbrace{Ax_k - \rho_k\, A\, r_k}_{Ax_{k+1}} - b).r_k = 0$

On obtient donc explicitement :

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

avec ${}^t r_k \, A \, r_k \neq 0$ puisque A est symétrique définie positive.

Remarque 32 En pratique le calcul de p_k n'a pas besoin d'être réalisé avec une très grande précision.

On peut montrer que la méthode de la plus grande pente converge pour toute condition initiale x_0 si x est strictement convexe. La convergence est linéaire et peut donc être assez lente.

Pour avoir ue convergence plus rapide, on peut utiliser la méthode de Newton pour résoudre $\nabla f(x) = 0$. En particulier, si f est convexe on obtient ainsi forcément un minimum de f. Il existe par ailleurs des variantes moins coûteuses que Newton et efficaces, comme la méthode de Broyden.

Une autre méthode beaucoup utilisée est la **méthode du gradient conjugué**.

Soit $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ de classe $\mathcal{C}^2$, avec $f(x) \xrightarrow{\|x\| \to +\infty} +\infty$ et $H_f(x)$ symétrique définie positive $\forall x \in \mathbb{R}^n$.

f possède alors un minimum global strict $\overline{x} \in \mathbb{R}^n$. La méthode du gradient conjugué utilise une direction de descente plus efficace que $\nabla f(x_k)$, qui fait également appel à $\nabla f(x_{k-1})$. Nous allons étudier cette méthode lorsque f est une fonction quadratique mais elle s'applique dans un cadre plus général.

7.3 Méthode du gradient conjugé pour une fonction quadratique

On considère $f(x) = \frac{1}{2} {}^t x \, A \, x - {}^t b \, x$ avec $A \in M_n(\mathbb{R})$ symétrique définie positive et $b \in \mathbb{R}^n$. Nous avons vu que $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ admet un minimum global strict en $x = \overline{x}$ avec $A\overline{x} = b$.

La méthode du gradient conjugué définit une suite $(x_k)_{k \geq 0}$ qui converge vers $\overline{x}$. Nous allons voir que la convergence se fait en **un nombre fini d'itérations** $\leq n$; de ce point de vue, la méthode du gradient conjugué est donc à classer parmi les méthodes directes. Cependant, à cause des erreurs d'arrondis, cette propriété n'est pas vérifiée en pratique (plus particulièrement pour de grands systèmes) et la méthode est plutôt considérée comme itérative. On contrôlera donc cet algorithme par un nombre maximal d'itérations et par un test d'arrêt.

7.3.1 Description de la méthode :

On notera par la suite $r_k = \nabla f(x_k) = Ax_k - b$. Si $r_k = 0$ alors l'algorithme s'arrête ($x_k =$? illisible).

i) **Initialisation :** On fixe $x_0 \in \mathbb{R}^n$.

- Si $x_0 = 0$ alors l'algorithme s'arrête car $x_0 = \overline{x}$.

- Si $x_0 \neq 0$, on calcule x_1 par la méthode de plus grande pente.
  On pose $\omega_0 = \nabla f(x_0)$.
  $-\omega_0 =$ direction de pente pour calculer x_1.
  $x_1 = x_0 - \rho_0\omega_0, \qquad f(x_0 - \rho_0\omega_0) = \underset{\rho\geq 0}{\text{Min}}\ f(x_0 - \rho\omega_0)$

  Remarque 33 Minimum explicite car minimse un polynôme de degré 2 en ρ.

ii) **Itération :** On suppose connus x_k et ω_{k-1} ($-\omega_{k-1}$ est la direction de la pente utilisée pour calculer x_k).

- Si $r_k = 0$ alors l'algorithme s'arrête car $x_k = \overline{x}$.
- Si $r_k \neq 0$: on pose

$$\begin{aligned} \omega_k &= r_k + \theta_k\omega_{k-1} \\ \theta_k &= \frac{{}^t r_k(r_k - r_{k-1}}{\|r_{k-1}\|_2^2} \end{aligned} \tag{7.1}$$

($-\omega_k =$ direction de la descente pour calculer x_{k+1})

$x_{k+1} = x_k - \rho_k\omega_k, \qquad f(x_k - \rho_k\omega_k) = \underset{\rho\geq 0}{\text{Min}}\ f(x_k - \rho\omega_k)$

Dans le cas présent où f est quadratique, la valeur de ρ_k est connue explicitement (voir le lemme qui suit).

Nous allons montrer les résultats suivants : (en patriculier, $r_k \neq 0$ implique $\omega_k \neq 0$ puisque $r_k \perp \omega_{k-1}$

Lemme 16 *i)* $f(x_{k+1}) = \underset{\theta\in\mathbb{R}}{\text{Min}}\ \underset{\rho\geq 0}{\text{Min}}\ f\Big[x_k - \rho(r_k + \theta\omega_{k-1}\Big]$

ii) ${}^t r_k w_{k-1} = 0, \ \ \rho_k = \dfrac{\|r_k\|_2^2}{{}^t\omega_k A\omega_k}$

iii) ${}^t\omega_k\, A\, \omega_{k-1} = 0$ (w_k et ω_{k-1} sont dits "A-conjugés")

Lemme 17 ${}^t r_k\, r_{k-1} = 0$ et (7.1) se transforme en :

$$\theta_k = \frac{\|r_k\|_2^2}{\|r_{k-1}\|_2^2} \tag{7.2}$$

Remarque 34 Les formules (7.1) et (7.2) sont équivalentes pour une fonction f quadratique. Pour f plus générale, (7.1) correspond à la méthode de Polak-Ribière et (7.2) à celle de Fletcher-Reeves. La méthode du gradient conjugué dans le cas quadratique est dûe à Hestenes et Steifel (1952).

7.3.2 Preuve du lemme 16

Nous allons montrer successivement *ii*), *i*) et *iii*).

Tout d'abord, puisque $f(x_{k-1} - \rho_{k-1}\omega_{k-1}) = \underset{\rho \geq 0}{\text{Min}}\ f(x_{k-1} - \rho\omega_{k-1})$

On a $\nabla f(x_{k-1} - \rho_{k-1}\omega_{k-1}) - \omega_{k-1} = 0$, soit $r_k \omega_{k-1} = 0 \implies$ on a montré *ii*) 1ère égalité.

Pour $\omega = r_k + \theta\omega_{k-1}$ on a (polnyôme du second degré en ρ.

$$\begin{aligned} f(x_k - \rho\omega = f(x_k) - \rho\nabla f(x_k)\,\omega + \frac{1}{2}\rho^2\, {}^t\omega\, H_f(x_k)\,\omega \\ = f(x_k) - \rho\, r_k\,\omega + \frac{1}{2}\rho^2\, {}^t\omega\, A\,\omega \end{aligned}$$

Puisque $r_k\,\omega_{k-1} = 0$, $r_k\,\omega$ est indépendant de θ et on obtient :

$$f(x_k - \rho\omega) = f(x_k) - \rho\,\|r_k\|^2 + \frac{1}{2}\rho^2\, {}^t\omega\, A\,\omega \tag{7.3}$$

Le minimum de ce polynôme de degré 2 est atteint en :

$$\rho_\theta = \frac{\|r_k\|_2^2}{{}^t\omega\, A\,\omega} \qquad 2^{\text{ème}} \text{ égalité de } ii)$$

et vaut

$$f(x_k - \rho_\theta\omega) = f(x_k) - \frac{1}{2}\frac{\|r_k\|_2^4}{{}^t\omega\, A\,\omega}$$

Pour minimiser $f(x_k - \rho_\theta\omega)$ suivant θ il faut minimiser ${}^t\omega\, A\,\omega$, c'est à dire $|\omega|$. Il faut choisir pour cela $\omega = \omega_k$ tel que ${}^t\omega_k\, A\,\omega_{k-1} = 0$, ce qu'on notera $\omega_k \perp \omega_{k-1}$:

$$< r_k + \theta\omega_{k-1}, r_k + \theta\omega_{k-1} = |r_k|^2 + 2\theta < r_k, \omega_{k-1} > + \theta^2|\omega_{k-1}|^2$$

Minimum pour :

$$\theta = \theta_k = -\frac{< r_k, \omega_{k-1} >}{|\omega_{k-1}|^2} \tag{7.4}$$

Donc :

$$\omega_k = r_k - \omega_{k-1}\frac{< r_k, \omega_{k-1} >}{|w_{k-1}|^2} \tag{7.5}$$

D'où $\omega_k \perp \omega_{k-1}$. Afin de montrer le lemme 16, il reste à montrer que (7.4) correspond bien à (7.1). D'une part :

$$\begin{aligned} r_k - r_{k-1} = A\,(x_k - x_{k-1}) = -\rho_{k-1}\, A\,\omega_{k-1} \qquad \text{donc :} \\ {}^t r_k(r_k - r_{k-1}) = -\rho_{k-1} < r_k, \omega_{k-1} > \end{aligned} \tag{7.6}$$

D'autre part :

$$\begin{aligned}
|\omega_{k-1}|^2 = (A\omega_{k-1}, \omega_{k-1}) &= -\frac{1}{\rho_{k-1}}(A(x_k - x_{k-1}), \omega_{k-1}) \\
&= -\frac{1}{\rho_{k-1}}(r_k - r_{k-1}, \omega_{k-1}) \\
&= \frac{1}{\rho_{k-1}}(r_{k-1}, \omega_{k-1}) \qquad (\text{car } (r_k, \omega_{k-1}) = 0) \\
&= \frac{1}{\rho_{k-1}}(r_{k-1}, r_{k-1} - \theta_{k-1}\omega_{k-2}) \\
&= \frac{1}{\rho_{k-1}} \|r_k\|^2 \qquad (\text{car } (r_{k-1}, \omega_{k-2}) = 0)
\end{aligned}$$

Donc :

$$\|r_k\|^2 = \rho_{k-1}|\omega_{k-1}|^2 \tag{7.7}$$

Avec (7.4), (7.6) et (7.7) on obtient donc :

$$\frac{{}^t r_k(r_k - r_{k-1}}{\left\|r_{k-1}^2\right\|} - \frac{< r_k, \omega_{k-1} >}{|\omega_{k-1}|^2} = \theta_k$$

On obtient donc la formule (7.2) plus simple pour le calcul de θ_k.

7.3.3 Convergence de la méthode du gradient conjugué et preuve du 17

Supposons $r_k \neq 0$ pour $k = 0, \ldots, n-1$ (si r_k s'annule l'algorithme converge). Cela implique $\rho_k \neq 0$ pour $k = 0, \ldots, n-1$.

Lemme 18 Pour tout $k = 1, \ldots, n$ on a :

$$(P_k) \quad \begin{cases} r_k\, \omega_q = 0 & \text{pour } q = 0, \ldots, k-1 \\ {}^t\omega_k\, A\omega_q = 0 & \text{pour } q = 0, \ldots, k-1 \\ r_k\, r_q = 0 & \text{pour } q = 0, \ldots, k-1 \end{cases}$$

Preuve 19 Par récurrence. On considère les produits scalaires

$$\begin{cases} (x, y) & = & {}^t xy = x.y \\ & \text{et} & \\ < x, y > & = & {}^t x\, A\, y \end{cases}$$

- P_1 est vraie : $r_1\, r_0 = r_1\, \omega_0 = 0$ (condition d'optimalité de ρ_0)

$$< \omega_1, \omega_0 >= 0 \qquad \text{d'après le lemme 1}$$

- Supposons P_k vraie et montrons $P_{k+1}(k \leq n-1)$. On a :

$$r_{k+1}.\omega_k = 0 \qquad \text{(condition d'optimalité de } \rho_k)$$

$$\begin{aligned} r_{k+1}.\omega_q &= (Ax_{k+1} - b, \omega_q) = (A(x_{k+1} - x_k) + Ax_k - b, \omega_q) \\ &= -\rho_k < \omega_k, \omega_q > + r_k.\omega_q \\ &= 0 \quad \text{pour } q = 0, .., k-1 \text{ (hyp de récurrence } P_k) \end{aligned}$$

Donc $r_{k+1}.\omega_q = 0$ pour $q = 0, .., k$.
Par ailleurs, $r_{k+1}.r_q = r_{k+1}.(\omega_q - \theta_q\omega_{q-1})$ (avec $\theta_0 := 0$ car $r_0 = \omega_0$)

Donc $r_{k+1}.r_q = 0$ pour $q = 0, .., k$

Ensuite $< \omega_{k+1}, \omega_k >= 0$ d'après le lemme 16, et pour $q = 0, \ldots, k-1$:

$$< \omega_{k+1}, \omega_q >=< r_{k+1}, \omega_q > + \theta_{k+1} < \omega_k, \omega_q >=< r_{k+1}, \omega_q >$$

(par l'hypothèse de récurrence P_k)

$\rho_{q+1} - r_q = A(x_{q+1} - x_q) = -\rho_q\, A\, \omega_q$ amène alors :

$$\begin{aligned} < \omega_{k+1}, \omega_q > =< r_{k+1}, \omega_q >= \frac{1}{\rho_q}(r_{k+1}, -r_{q+1} + r_q) \\ = 0 \end{aligned}$$

car $0 \leq q \leq k-1$ et $(r_{k+1}, r_p) = 0$ pour $p = 0, \ldots, k$
Cela prouve P_k par récurrence.

En conclusion, la famille $(\omega_0, \ldots, \omega_{n-1})$ est libre car les ω_i sont deux à deux orthogonaux pour le produit scalaire $< x, y >= {}^t x\, A\, y$. C'est donc une base de $\mathbb{R}^n$. Puisque r_n est orthogonal à $\omega_0, \ldots, \omega_{n-1}$, on a donc $r_n = 0$. Nous avons donc montré que $A\, x_n = b$, i.e. $x_n =$??

Théorème 24 *Soit $A \in M_n(\mathbb{R})$ symétrique définie positive, $b \in \mathbb{R}^n$ et $f(x) = \frac{1}{2}\, {}^t x\, A\, x - {}^t b\, x$. Alors l'algorithme du gradient conjugué définit une suite $(x_k)_{k=0,\ldots,p}$ avec $p \leq n$ et $A\, x_p = b$. On a $f(x_p) = \underset{x \in \mathbb{R}^n}{Min}\, f(x)$.*

Remarque 35 Le cas $p < n$ est exceptionnel.

Enfin, nous avons montré dans le lemme 18 que $r_k.r_{k-1} = 0$, ce qui prouve le lemme 17.