Briques de base de l'option B: calcul scientifique

Clara Genes, Adrienne Le Meur et Matthias Hostein

6 juin 2024

Table des matières

Question 1 : Dans la résolution d'un système linéaire, quelles difficultés peuvent se poser ? Quelles sont les stratégies envisageables ?

Référence de la réponse :
- J-E Rombaldi, Analyse matricielle.
- M. Schatzman, Analyse numérique
- P. Ciarlet, Introduction à l'analyse numérique

1 Introduction aux problèmes

On s'interesse à la résolution de systèmes linéaires $Ax = b$ d'inconnu $x \in \mathbb{K}^n$ et de données $b \in \mathbb{K}^n$,$A \in \mathcal{M}_n(\mathbb{K})$. On peut déjà faire plusieurs remarques :

- Le système étant linéaire, quitte à en résoudre deux, on peut se contenter d'étudier les systèmes réels.
- Si $b \notin Im(A)$ alors il n'y a aucune solution exacte.
- Si $A \in \mathcal{GL}_n(\mathbb{K})$ alors il y a une unique solution exacte qui est $A^{-1}b$. Le problème réside dans le calcul de A^{-1}. On supposera toujours cette hypothèse vérifiée, quitte à extraire une matrices plus petite : les lignes qui s'écrivent comme combinaisons linéaires d'autres lignes étant des redondances augmentant la complexité temporelle et spatiale des algorithmes.

On distingue deux catégories de méthodes de résolution d'un système linéaire, les méthodes directes et les méthodes itératives.

- Les méthodes directes conduisent à la solution exacte en un nombre fini d'étapes. Ces métodes modifient la matrice de départ, ainsi, pour les grands systèmes, la propagation d'erreurs d'arrondi en diminue l'efficacité. Ainsi, ces méthodes sont plutôt adaptés aux "petits systèmes" ($n < 50$). Elles sont notamment utilisées pour la résolution des problèmes d'interpolation ou d'approximation.
- Les méthodes itératives donnent une approximation de la solution et tendent vers la solution exacte en un nombre potentiellement infini d'étapes. Elles sont bien adaptées aux cas des matrices creuses car ces méthodes ne modifient pas la matrice de départ. Les méthodes itératives sont utilisées en relation avec les méthodes de différences finies et d'éléments finis pour la résolution d'équations aux dérivées partielles.

Un algorithme de résolution numérique n'a pas de précision infinie, ainsi on distingue trois types d'erreur numérique :

- les erreurs sur les données (pour des données expérimentales). Dans la pratique, A et b ne sont connues que de façon approximative : il faut donc savoir si de petites variations des coefficients peuvent entraîner de grosses variations sur la solution.
- les erreurs d'arrondis (calculs en flottant). Elles s'accumulent au cours des calculs : il faut connaître le nombre d'opérations que nécessite un algorithme.
- les erreur de troncature. Un système est dégénéré si $det(A) = 0$, mais numériquement il faut une précision infinie pour vérifier cette condition : il faut donc définir un concept de dégénérescence numérique d'un système linéaire.

Pour le permier point, le conditionnement de la matrice nous donnera une idée de la réponse. En effet, on a le théorème suivant.

Théorème 1. Soit $A \in \mathcal{GL}_(\mathbb{R})$ et $b \in \mathbb{R}^n$. On considère les deux équations

$$Ax = b, \qquad A(x + \delta x) = b + \delta b.$$

Alors, en notant $Cond(A) \coloneqq \|A\|\|A^{-1}\|$, on a

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\delta b}{\|b\|}, \qquad \frac{\|\delta x\|}{\|x + \delta x\|} \leq Cond(A)\frac{\|\delta A\|}{\|A\|}.$$

Démonstration. Les inégalités venant du fait que $\|b\| \leq \|A\|\|x\|$, et $\|\delta x\| \leq \|A^{-1}\|\|\delta b\|$. □

Théorème 2. Si $A \in S_n(\mathbb{R})$, alors $Cond_{\|\cdot\|_2}(A) = \frac{|\lambda_{max}|}{|\lambda_{min}|}$.

2 Méthodes directes

Si la matrice A est triangulaire, pour résoudre le système on procède simplement par "remontée" qui se fait en $O(n^2)$. Mieux : si la matrice est tridiagonale, on peut résoudre le problème en $O(n)$. Mais ce sont des cas particuliers...

Pour le cas plus général, on pourrait penser à deux choses :

- la méthode de Cramer (résolution avec les déterminants) : bien qu'historiquement intéressante, elle est à proscrire du point de vue numérique. Elle est super instable et se fait en $O(n!n^2)$.
- trigonaliser les matrices : c'est également à proscrire. Cela peut faire apparaître des complexes (et donc des arrondis en plus et de la complexité spatiale en plus) et de toute façon, chercher des éléments propres est très coûteux (cf Question 4).

Ainsi, on veut se ramener au problème simplifié de la matrice trangulaire sans augmenter trop fortement la complexité. Ce principe d'élimination des inconnus est exactement le principe des méthodes directes et c'est là qu'interviennent les matrices d'opérations élémentaires et la méthode du pivot de Gauss ainsi que ses cas particuliers. On a alors les méthodes suivantes :

- Sur une matrice A inversible, il est toujours possible d'effectuer le pivot de Gauss en $\frac{2n^3}{3} + O(n^2)$ opérations. Les difficultés à prévoir sont les cas où un pivot est trop petit, auquel cas des arrondis s'accumulent. Pour résoudre ce problème on peut chercher un pivot partiel (à chaque étape on choisit le plus grand pivot de la ligne ou colonne) ou chercher un pivot total (à chaque étape on prend le plus grand pivot de la matrice). Ceci se faire par un tri sur n ou n^2 éléments donc rajoute de la complexité.
- Si les mineurs principaux de A sont non nuls (ex : matrices à diagonale strictement dominante), on n'a pas besoin de faire d'opérations sur les lignes dans la méthode de Gauss et on peut simplifier la méthode en la décomposition LU mais elle reste en $\frac{2n^3}{3} + O(n^2)$ opérations. L'intérêt étant surtout dans la résolution de plusieurs systèmes avec la même matrice.

- Si A est symétrique définie positive, alors les matrices L et U peuvent être liées sous la forme $L = U^T$ *i.e.* $A = LL^T$, c'est la décomposition de Cholesky et elle s'effectue en $\frac{n^3}{3} + O(n^2)$ opérations. Donc on fait la moitié des opérations mais la condition sur A est très restrictive.

- On peut aussi envisager d'utiliser la décomposition QR qui consiste à écrire une matrice inversible sous la forme d'un produit d'une matrice orthogonale et d'une matrice triangulaire supérieure, et qu'on peut obtenir grâce à la méthode de Householder. On pourrait penser que cette décomposition est une bonne stratégie pour gérer des problèmes de conditionnemnet mais le conditionnement de la matrice R est le même que celui de notre matrice de base. Mais on peut également traiter grâce à cette méthode le cas où notre matrice A n'est pas inversible, auquel cas on fait une résolution au sens des moindres carrés. Cette méthode a cependant le désavantage d'être plus coûteuse que les précédentes car elle s'effectue en $\frac{4n^3}{3} + O(n^2)$ opérations.

3 Méthodes itératives

Les méthodes itératives consistent à écrire $A = M - N$ avec M "facile à inverser" et à construire une suite $(x_k)_{k\in\mathbb{N}}$ vérifiant $x_{k+1} = M^{-1}Nx_k + M^{-1}b$ convergente. À noter que si la suite ainsi définie converge, c'est nécessairement vers la solution x du système. Les principaux problèmes des méthodes itératives sont les suivants :

- Comment choisir M (que veut dire "facilement inversible") ?
- À quelles conditions on a convergence de la méthode ?
- Quelle est la vitesse de convergence ?

La deuxième question est répondue grâce au théorème suivant.

Théorème 3. La méthode itérative associée au couple (M,N) tel que $A = M-N$ converge si et seulement si $\rho(M^{-1}N) < 1$. De plus, si $A \in S_n(\mathbb{R})$, alors $M^t + N \in S_n(\mathbb{R})$ et si $M^t + N \in S_n^{++}(\mathbb{R})$, alors la méthode converge.

On considère les méthodes suivantes.

- La méthode de Jacobi : $M = \text{diag}(a_{11}, \ldots, a_{nn})$ matrice diagonale extraite de A. Si A est à diagonale strictement dominante, la méthode converge.

- La méthode de Gauss-Seidel : $M = (a_{ij}$ si $i \geq j$ et 0 sinon), matrice triangulaire inférieure extraite de A. Si A est symétrique définie positive ou à diagonale strictement dominante, la méthode converge. Lorsque les deux méthodes converge, il est préférable de choisir celle de Gauss-Seidel, qui converge plus vite. Notamment, dans le cas des matrices tridiagonales les deux méthodes convergent simultanément. Cependant, il est possible que la méthode de Jacobi converge alors que celle de Gauss-Seidel diverge.

- La méthode de relaxation : $M := \frac{1}{\omega}(\text{diag}(a_{11}, \ldots, a_{nn}) - \omega(a_{i,j}$ si $i > j$ et 0 sinon)) et $N := A - M$, où $\omega \in \mathbb{R}^*$. Une condition nécessaire à la convergence de cette méthode est que $\omega \in]0,2[$. Cette condition est également suffisante si $A \in \mathcal{H}_n^{++}(\mathbb{C})$.

4 Méthodes de descente de gradient

La solution x du problème $Ax = b$ peut également être vu comme le vecteur réalisant le minimum de la fonctionnelle quadratique $\varphi(x) = \frac{1}{2}\langle Ax,x\rangle - \langle b,x\rangle$ puisque le gradient de cette fonction est exactement $Ax - b$ et la fonctionnelle admet un unique minimum (coercivité, continuité et *stricte* convexité) vérifiant $\nabla\varphi(x) = 0$. On se ramène donc à un problème d'optimisation étudié en Question 5.

5 Cas d'une matrice rectangle

Lorsque la matrice A n'est plus carrée, cas qui nous intéresse par exemple si on veut faire une régression linéaire, on peut tout de même trouver une solution au problème à partir du moment que b est dans l'image de A. Pour ce faire, on peut résoudre par les moindres carrés l'équation $Ax = b$ où $A \in \mathcal{M}_{m,n}(\mathbb{K})$ et $b \in \mathbb{K}^n$. L'ensemble des solutions de cette équation est le sous espace affine de direction $Ker(A)$ et passant par un point x_0 tel que Ax_0 est le projeté orthogonal de b sur $Im(A)$.
Par ailleurs tout vecteur x est solution au sens des moindres carrés ssi il est solution de l'équation dite normale $A^*Ay = A^*b$ (car $Ker(A) = Ker(A^*A)$). On peut alors résoudre cette équation grâce à la décomposition en valeurs singulières de A : si $A = U\Sigma V^*$, avec $\Sigma = \text{Diag}(\sigma_1,\dots,\sigma_r,0,\dots,0) \in \mathcal{M}_{m,n}(\mathbb{R})$, on peut définir le *pseudo-inverse* de A, noté $A^\dagger$ ainsi :

$$A^\dagger := V\Sigma^\dagger U^*, \quad \text{où} \quad \Sigma^\dagger := \text{Diag}\left(\sigma_1^{-1},\dots,\sigma_r^{-1},0,\dots,0\right) \in \mathcal{M}_{n,m}(\mathbb{R})$$

et alors le vecteur $A^\dagger b$ est l'unique solution du système $Ax = b$ au sens des moindres carrés qui soit de norme euclidienne minimale.

Question 2 : Comment approcher une intégrale ? Quel est le lien avec l'interpolation polynomiale ?

Référence de la réponse : Calcul scientifique et symbolique, A. Yger

6 Introduction aux problèmes

Notre objectif ici est de calculer pour une fonction f réelle et continue définie sur un domaine borné $[a,b]$ l'intégrale

$$\int_a^b f(x)\mathrm{d}x.$$

Dans des cas très limités, cette intégrale peut être calculée analytiquement. Le plus souvent on a au moins un des problèmes suivant :

- Le calcul analytique est long et compliqué,
- Le résultat de l'intégrale est une fonction compliquée qui fait appel à d'autres fonctions elles-même longues à évaluer
- Cette intégrale n'a pas d'expression analytique.

On veut alors trouver des méthodes pouvant nous permettre de calculer numériquement une valeur approchée de l'intégrale. Presque toutes les méthodes sont basées sur l'évaluation de la fonction f en des point distincts. Pour comparer ces méthodes, on regardera deux critères :

- l'erreur de la méthode : ne pouvant pas être calculée exactement puisqu'en général, on ne connaît pas l'intégrale que l'on cherche à calculer, une majoration peut souvent être estimée en étudiant le développement en série de Taylor de la fonction.
- la vitesse de la méthode.

Définition 1. L'ordre d'une méthode d'approximation d'intégrale est donné par le plus grand entier n tel que la méthode rend exact les monômes de degré n.

7 Méthodes de Newton-Cotes composites

Le principe général de ces méthodes est d'approcher la fonction f par un polynôme P. L'avantage est que l'on sait analytiquement calculer les intégrales de polynômes. On choisit préalablement un degré p et on choisit le polynôme interpolateur de degré p coïncident avec f en $p+1$ points. On divise régulièrement n fois notre intervalle et les $p+1$ points seront choisit indépendamment dans chaque sous-intervalle $[a_k,b_k]$:

$$\{x_i = a_k + ih, i \in [0,p]\}, \text{ avec } h = \frac{b_k - a_k}{p}.$$

Nous noterons également $h = b_k - a_k$ si $p = 0$. Les méthodes les plus courantes sont les méthodes de degré le plus bas.

Méthode	degré	principe sur chacun des sous-intervalle	vitesse	erreur
Rectangles	$p=0$	polynôme constant égal à la borne droite ou gauche.	$O(n)$	$O(h)$
Points milieux	$p=0$	polynôme constant égal au milieu du sous-intervalle	$O(n)$	$O(h^2)$
Trapèzes	$p=1$	droite passant par les extermités de l'intervalle	$O(2n)$	$O(h^2)$
Simpson simple	$p=2$	parabole passant par les extrémités et le milieu	$O(3n)$	$O(h^4)$

Erreur de la méthode des rectangles. Soit $f \in \mathcal{C}^1$, on a

$$\varepsilon_k = \left| \int_a^b f(t)\mathrm{d}t - \sum_{k=0}^{n} f(x_k)h \right| \leqslant \sum_{k=0}^{n} \int_{x_k}^{x_{k+1}} |f(t) - f(x_k)|\,\mathrm{d}t \leqslant \sum_{k=0}^{n} |f'(x_k)| \frac{h^2}{2} \leqslant Kh.$$

La dérivé étant continue sur un compacte, elle y est bornée. □

Erreur de la méthode des trapèzes. Soit $f \in \mathcal{C}^2$, on a l'approximation

$$g(x) = \frac{f(x_{k+1}) - f(x_k)}{h}(x - x_k) + f(x_k)$$

avec

$$\int_{x_k}^{x_{k+1}} g(t)\mathrm{d}t = h\frac{f(x_{k+1}) + f(x_k)}{2}$$

Ainsi, puisque f est deux fois dérivable et que la fonction g est un polynôme d'interpolation, pour tout $x \in [x_k, x_{k+1}]$, il existe un point $\zeta_x \in]\, x_k, x_{k+1}\,[$ tel que

$$f(x) - g(x) = \frac{1}{2}(x - x_k)(x_{k+1} - x)f''(\zeta_x).$$

Donc,

$$\begin{aligned}
\varepsilon_k &\leqslant \sum_{k=0}^{n} \int_{x_k}^{x_{k+1}} |f(t) - g(t)|\mathrm{d}t \\
&\leqslant \sum_{k=0}^{n} \frac{f''(\zeta_x)}{2} \int_{x_k}^{x_{k+1}} (t - x_k)(x_{k+1} - t)\mathrm{d}t \\
&\leqslant M \sum_{k=0}^{n} \int_{x_k}^{x_{k+1}} (x_{k+1} - x_k)(x_{k+1} - t)\mathrm{d}t \leqslant M \sum_{k=0}^{n} \frac{h^3}{3} = Kh^2.
\end{aligned}$$

□

Des méthodes d'ordre plus élevé existent. On peut construire des approximations en utilisant les polynômes de Lagrange. Evidemment, plus le degré est élevé, plus la méthode est précise ($O(h^{p+2})$) mais puisque plus le degré est élevé, plus la mthode est lente, on ne cherche généralement pas à obtenir de degré trop élevé. On peut retrouver des méthodes classiques d'ordre 3 (Simpson 3/8) et d'ordre 4 (Boole) mais au delà de $p = 7$ elles ne sont jamais utilisées.

8 Méthode de Gauss

Le principe reste le même, mais cette fois ci, les sous-intervalles ne sont plus régulier. Ceci nous donne des degrés de liberté en plus (choix de la position des a_k,b_k) et donc réduit l'erreur sans augmenter le coût. Ce sont donc des méthodes plus générales que celles de Newton-Cotes car permettent d'approcher l'intégrale même si la fonction comprends une singularité intégrable ($x \mapsto 1/\sqrt{x}$ en 0, par exemple). La méthode consiste à calculer l'intégrale pondérée

$$\int_{[a,b]} f(x)\omega(x)\mathrm{d}x.$$

Pour déterminer les bons points d'interpolation, on cherche à rendre exactes les méthodes sur les monômes. Les points d'interpolations seront les racines de polynômes orthogonaux pour le produit scalaire associé au poids ω. Les méthodes de Gauss peuvent être utilisées avec un nombre de points assez grand, même si en général, un nombre de points de 4 à 6 est largement suffisant pour la plupart des applications. Les méthodes les plus populaires sont les suivantes :

Méthode de Gauss-	poids	intervalle
Legendre	$\omega(x) = 1$	$[-1,1]$
Tchebychev	$\omega(x) = 1/\sqrt{1-x^2}$	$]-1,1[$
Laguerre	$\omega(x) = \exp(-x)$	$]0,+\infty[$
Hermite	$\omega(x) = \exp(-x^2)$	$\mathbb{R}$

9 Méthode de Monte-Carlo

L'intégrale d'une fonction correspond à l'aire algébrique sous sa courbe. Estimer l'intégrale d'une fonction revient donc à estimer l'aire d'une surface et on peut donc appliquer une méthode de Monte-Carlo. L'intégrale peut être vue comme une espérence. En effet, si on regarde $a = 0$ et $b = 1$ et soit U une variable aléatoire de loi uniforme sur [0,1] alors

$$\mathbb{E}[f(U)] = \int_0^1 f(u)\mathrm{d}u.$$

On génère alors n variables aléatoires *i.i.d.* $U_1, \ldots, U_n$ uniformément dans [0,1], alors

$$\mathbb{E}[f(U)] \sim \frac{1}{n}\sum_{k=1}^{n} f(U_k).$$

Ce procédé peut s'appliquer à une intégrale en dimension quelconque et pas nécessairement en dimension une comme ici. D'ailleurs en grande dimension, les méthodes de Monte-Carlo sont bien plus efficaces que d'autres méthodes classiques comme celle des sommes de Riemann par exemple.

La vitesse de convergence est de $\sqrt{n}$ par le TCL, et l'erreur peut être quantifiée grâce à des intervalles de confiance non asymptotiques obtenus par des inégalités de concentration telles que Tchebychev ou Hoeffding, ou encore par intervalles de confiance asymptotiques encore une fois grâce au TCL et à la connaissance de la fonction de répartition de la loi normale.

Question 3 : Pour résoudre un système d'équations non linéaires, quelles méthodes employer ? Quelles sont leurs propriétés ?

Référence pour la réponse :
- A. Yger, Calcul scientifique et symbolique.
- Quarteroni, Sacco, Méthodes numériques.

10 Introduction aux problèmes

Notre objectif est de résoudre l'équation

$$f(x) = 0$$

avec $f : [a,b] \mapsto \mathbb{R}$ continue et possédant une unique racine α dans $]a,b[$. Quitte à réduire l'intervalle, on suppose que $f(a)f(b) < 0$. On sait d'après le théorème des valeurs intermédiaires que α est bien dans $]a,b[$. On va présenter plusieurs méthodes itératives résolvant ce problème.

- La méthode de dicotomie : l'intervalle $[a,b]$ pouvant être grand, cette méthode consiste en la réduction de l'intervalle autour de α pour avoir une idée plus précise d'où il se trouve. À précision donnée, on sait par avance combien d'itérations il faudra pour approcher α selon la condition.
- Les méthodes de points fixes : on transforme le problème $f(x) = 0$ en $g(x) = x$ et donc on transforme l'intersection de f avec l'axe des abscisse en l'intersection de g avec la droite $y = x$.

Puisque ce sont des méthodes itératives, on crée des suites de points $(x_k)_{\mathbb{N}}$. Donc on a plusieurs problèmes :

- la suite (x_n) converge ? Vers α ?
- à précision fixée, combien d'itérations il faudra ?

Le premier problème est purement mathématique et est résolu grâce à la continuité de f et à l'unicité de α. On rappelle le principe du point fixe de Banach : soit $x_0 \in [a,b]$ et $x_{n+1} = g(x_n)$. Alors si g est une application strictement contractante de $[a,b]$ dans $[a,b]$, la suite ainsi définie converge géométriquement vers α. On rappelle également que l'ordre d'une méthode de point fixe est le plus grand entier p tel que $\lim_{n\to\infty} \frac{|x_{n+1}-\alpha|}{|x_n-\alpha|} \neq 0$

11 Méthode de la dichotomie

On initialise avec $x_0 = a$ et $y_0 = b$ et on calcule

$$c = \frac{x_0 + y_0}{2}.$$

Si $f(c) = 0$ on a trouvé notre racine, sinon si $f(x)f(c) < 0$ on pose $x_1 = x_0$ et $y_1 = c$ si $f(x)f(c) > 0$ on pose $x_1 = c$ et $y_1 = y_0$ et on continue avec le segment $[x_1,y_1]$. On construit ainsi une suite de segments emboîtés $[x_k,y_k]$ dont le diamètre $(b-a)/2^k$ tend vers 0. Puisque $\mathbb{R}$ vérifie la propriété des segments emboîtés, les suites $(x_k)_{\mathbb{N}}$ et $(y_k)_{\mathbb{N}}$ tendent vers la même limite, α. C'est une méthode d'ordre 1.

12 Méthode de la sécante

Supposons que f soit dérivable et que sa dérivée ne s'annule pas sur l'intervalle. Le principe est le suivant : soit Δ_0 la droite passant par $(a,f(a)),(b,f(b))$ et x_0 sont intersection avec les abscisses. Si $f(x_0)=0$ on a fini, sinon si $f(x_0)f(a)<0$ on considère Δ_1 la droite passant par $(a,f(a)),(x_0,f(x_0))$ ou sinon Δ_1 la droite passant par $(x_0,f(x_0)),(b,f(b))$ et on réitère. Dans le cas général, on obtient ainsi

$$\begin{array}{c} x_0 = a \\ x_{n+1} = x_n + u(x_n,x_{n-1})f(x_n) \end{array} \qquad \text{avec } u(x,y) = \frac{x-y}{f(x)-f(y)}$$

Si f est convexe (resp. concave), on ne modifie qu'à gauche (resp. droite) et on obtient

$$u(x) = \frac{b-x}{f(b)-f(x)} \qquad \left(\text{resp. } \frac{x-a}{f(x)-f(a)}\right).$$

La méthode est d'ordre 1 et dans le cas général est de pas 2 (car utilise x_n et x_{n-1})

13 Méthode de Newton

La méthode de Newton consiste à remplacer la courbe par sa tangente en une de ses deux extrémités. Le point x_1 est l'intersection de cette tangente avec l'axe des abscisses. Par exemple en b, l'équation de la tangente donne comme point $x_1 = b - f(b)/f'(b)$. Le point d'intersection pourrait sortie de l'intervalle $[a,b]$ donc il faut faire attention. Pour éviter ce souci, en pratique, on utilise souvent la méthode de dichotomie pour trouver un x_0 assez proche de la racine. On a l'itération

$$\begin{array}{c} x_0 = a \\ x_{n+1} = x_n + u(x_n)f(x_n) \end{array} \qquad \text{avec } u(x) = \frac{-1}{f'(x)}$$

La méthode de Newton est encore valable en dimension supérieure et on a le théorème suivant.

Théorème 4. Soit Ω un ouvert de $\mathbb{R}^n$. On suppose $f \in \mathcal{C}^2(\Omega,\mathbb{R}^n)$ et $x^* \in \Omega$ tel que $f(x^*)=0$ et $df_{x^*} \in \mathcal{GL}_n(\mathbb{R})$. Alors il existe $r>0$ tel que $B(x^*,r) \subset \Omega$ et pour tout $x \in B(x^*,r)$ la suite définie pour tout $k \in \mathbb{N}$ par

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

est bien définie et converge vers x^* avec

$$\|x_{k+1}-x^*\| \leqslant C\,\|x_k-x^*\|^2.$$

Démonstration. Pour la bonne définition, $f \in \mathcal{C}^2$ implique que sa différentielle est continue localement lipschitzienne et est inversible sur une boule autour de x^* d'inverse continue sur cette boule. Pour la convergence :

$$\begin{aligned} \|x_{k+1}-x^*\| &= \|x_k - df_{x_k}^{-1}(f(x_k)-f(x^*)) - x^*\| \\ &= \|df_{x_k}^{-1}\left(df_{x_k}(x_k-x^*) - f(x_k) - f(x^*)\right)\| \\ &\leqslant \|df_{x_k}^{-1}\|\|df_{x_k}(x_k-x^*) - f(x_k) - f(x^*)\| \leqslant LK\|x_k-x^*\|^2 \end{aligned}$$

où la dernière inégalité est dûe d'une part aux formules de Taylor grâce au fait que f soit deux fois continûement différentiable et d'autre part au caractère continue de la différentielle sur une boule autour de x^*. □

Question 4 : Comment résoudre un problème aux valeurs propres et quelles difficultés peuvent alors survenir ?

Références pour la réponse :
- G. Allaire, Analyse numérique et optimisation
- A. Quarteroni, R. Sacco, F. Saleri Méthodes numériques
- J-E. Rombaldi, Analyse matricielle
- P. Ciarlet, Introduction à l'analyse numérique

14 Introduction aux problèmes

On appelle problème aux valeurs propres tout problème de la forme : trouver

$$(x,\lambda) \in V \times \mathbb{K}, \text{ tels que } Ax = \lambda x$$

où A est une matrice. On pourrait penser qu'il "suffit" de factoriser le polynôme caractéristique de la matrice associée. Mais on sait depuis Galois et Abel qu'on ne peut pas calculer par opérations élémentaires les racines d'un polynôme quelconque de degré supérieur à 5. Ainsi, il ne peut pas exister de méthode directe (donnant le résultat en un nombre fini d'opérations). On ne verra donc que des méthodes itératives de calcul de valeurs propres. Il existe deux classes de méthodes numériques pour traiter ce problème :

- les méthodes partielles : calcul approché des valeurs propres extrêmes de A (c'est-à-dire λ_1 (resp. λ_2) de plus grand (resp. de plus petit) module). La résolution d'un tel problème présente un grand intérêt dans beaucoup d'applications concrètes (sismique, étude des vibrations des structures et des machines, analyse de réseaux électriques, mécanique quantique,...) dans lesquelles λ_n et le vecteur propre associé x_n permettent la détermination de la fréquence propre et du mode fondamental d'un système physique donné. Il peut être aussi utile de disposer d'approximations de λ_1 et λ_n pour analyser des méthodes numériques.

- les méthodes globales : fournissent des approximations de tout le spectre de A.

Certaines méthodes permettent le calcul simultané des vecteurs propres (*c.f.* méthode de la puissance) alors que d'autres ne permettent que le calcul de valeurs propres (*c.f.* méthode QR) nécéssitant alors des calculs supplémentaires.
Les méthodes étant toutes itératives, il est utile de connaître leur localisation dans le plan pour accélérer la convergence. Nous savons déjà que pour toute norme d'algèbre,

$$\rho(A) \leqslant \|A\|.$$

Ainsi, même si en général cette inégalité est assez grossière, ont sait que les valeurs propres sont contenues dans un disque. Le théorème des disques de Gershgörin donne une deuxième estimation

$$\rho(A) \subset \bigcup_{i=1}^{n} \left\{ z \in \mathbb{C} : |z - a_{ii}| \leqslant \sum_{j \neq i} |a_{ij}| \right\}.$$

Et puisque A et A^T ont les mêmes valeurs propres,

$$\rho(A) \subset \bigcup_{j=1}^{n} \left\{ z \in \mathbb{C} : |z - a_{jj}| \leqslant \sum_{i \neq j} |a_{ij}| \right\}.$$

Il existe un autre théorème de Gershgörin, disant qu'il y a autant de valeur propre (avec multiplicité algébrique) que de disque dans chaque composante connexe.

15 Méthode de la puissance

La méthode de la puissance est une méthode partielle. C'est une très bonne approximation des valeurs propres extrémales d'une matrice et des vecteurs propres associés. Supposons A diagonalisable, λ_1 de multiplicité algébrique égale à 1 et de module strictement plus grand que celui des autres valeurs propres et posons $q_0 \in \mathbb{C}^n$ unitaire. Le principe de la méthode est le suivant

$$\left\{ \begin{array}{rcl} z_k = & Aq_{k-1}, \\ q_k = & z_k/\|z_k\|_2. \end{array} \right.$$

On peut vérifier par récurrence que

$$q_k = \frac{A^k q_0}{\|A^k q_0\|_2}.$$

Le fait que A soit diagonalisable, implique que l'on peut décomposer q_0 dans une base de vecteurs propres. En supposant que le scalaire α_1 selon λ_1 soit non nul,

$$A^k q_0 = \alpha_1 \lambda_1^k \left(x_1 + \sum_{i=2}^{n} \frac{\alpha_i}{\alpha_1} \left(\frac{\lambda_i}{\lambda_1} \right)^k x_i \right) \text{ ainsi } \lim_{k\to\infty} \|q_k\| = \lambda_1 \text{ et } \lim_{k\to\infty} q_k = \pm x_1$$

avec une convergence d'autant plus rapide que le quotient λ_2/λ_1 est petit en module. De cette méthode, en découle deux applications.

- Soit un nombre $\mu \in \mathbb{C}$ donnée, en appliquant la méthode de la puissance à la matrice $(A - \mu I)^{-1}$ on obtient la valeur propre de A la plus proche de μ, notamment pour $\mu = 0$ et si A est inversible, on obtient λ_n.
- Si les valeurs propres sont distinctes en module, en appliquant la méthode de la puissance à la matrice $A - \lambda_1 e_1 e_1^T$, on obtient λ_2, etc.

16 Méthode de Givens-Householder

Soit A une matrice symétrique réelle. La méthode de Givens-Householder se compose de deux étapes successives :

- l'algorithme de Householder réduit une matrice A symétrique en une matrice tridiagonale (en un nombre fini d'étapes),
- l'algorithme de Givens fournit (de manière itérative) les valeurs propres d'une matrice tridiagonale.

Principe de la méthode de Householder :

$$\left\{ \begin{array}{rcl} A_1 & = & A \\ A_k & = & \begin{pmatrix} T_k & E_k^* \\ E_k & M_k \end{pmatrix} \\ A_{k+1} & = & H_k^* A_k H_k \end{array} \right. \qquad \left\{ \begin{array}{rcl} H_0 & = & I_n \\ H_k & = & \begin{pmatrix} I_k & 0 \\ 0 & \widetilde{H}_k \end{pmatrix} \end{array} \right.$$

où

$$\widetilde{H}_k = I_{n-k} - 2\frac{v_k v_k^*}{\|v_k\|^2} \text{ et } v_k = a_k + \|a_k\| e_1$$

Où T_k est tridiagonale et E_k est nulle partout sauf sa dernière colonne étant a_k. La matrice $(H_1 H_2 \cdots H_{n-2})^* A (H_1 H_2 \cdots H_{n-2})$ est tridiagonale symétrique.

Principe de la méthode de Givens :
Pour une matrice tridiagonale , on définit par récurrence une suite de polynômes caractéristique (p_j) comme les polynômes caractéristiques des mineurs principaux. Pour calculer la i-ième valeur propre λ_i on choisit un intervalle $[a_0,b_0]$ où on est sûr qu'elle s'y trouve et on procède par dicotomie en choisissant le côté droit (resp. gauche) selon que la suite $\left(p_j\left(\frac{a_0+b_0}{2}\right)\right)_j$ change de signe moins (resp. plus) de i-fois.

17 Méthode de Jacobi

Ajouter (p. 111 Ciarlet)

18 Méthode QR

Naïvement on pourrait penser que pour trouver la décomposition QR d'une matrice il faudrait faire un algorithme d'orthonormalisation de Gram Schmidt. Mais numériquement, plus la taille de la matrice considérée est grande et plus les erreurs d'arrondis résultant de projections s'accumulent. Alors la matrice Q qu'on obtient peut s'avérer ne même pas être orthogonale.

A la place, on utilise l'algorithme dit de Householder. On doit d'abord introduire un certain type de matrices appelées matrices de Householder. Soit $v \in \mathbb{R}^n \setminus \{0\}$. On appelle matrice de Householder associée au vecteur v la matrice $H(v) := I - 2\frac{v^t v}{{}^t vv}$. Cette matrice est symétrique et orthogonale et c'est la symétrie orthogonale par rapport à $\{v\}^\perp$. En particulier on remarque que pour e un vecteur unitaire, alors $H(v - \|v\|e)v = \|v\|e$. C'est cette observation qui pousse à prendre pour v le premier vecteur colonne de la matrice A et pour e le premier vecteur de la base canonique, de sorte à ce que $H(v - \|v\|e)A$ ait comme premier vecteur un vecteur colinéaire à e. On propage ensuite l'échelonnement de proche en proche.

Si les mineurs principaux de la matrice considérée sont tous non nuls, alors on va converger vers la matrice diagonalisée avec les valeurs propres rangées par ordre de modules décroissants, avec une vitesse géométrique de raison $\max\left\{\frac{|\lambda_{i+1}|}{|\lambda_i|}\right\}$.

Question 5 : Quelles méthodes numériques peuvent permettre de résoudre un problème d'optimisation ? Un problème d'optimisation sous contraintes d'égalité ?

Référence de la réponse :
- P. Ciarlet Analyse numérique et optimisation
- Allaire, Analyse numérique et optimisation

19 Introduction aux problèmes

Après l'étape de modélisation d'un phénomène physique et l'étape de simulation numérique, on veut souvent agir sur le phénomène ou sur le système afin d'en améliorer cetaines performances. Il s'agit de l'étape de l'optimisation : la minimisation ou maximisation d'une fonction dépendant du modèle. Notre objectif est donc de déterminer

$$a \in V, \qquad J(a) = \inf_{x \in V} J(x),$$

avec J une fonction $\alpha-$convexe différentiable sur l'espace de Hilbert réel V. Les méthodes sont toutes itératives.

20 Méthode du gradient sans contrainte

L'algorithme du gradient consiste à se "déplacer" d'une itérée u_n en suivant la ligne de plus grande pente associée à la fonction coût J issue de u_n. La direction de cette ligne est donnée par le gradient $\nabla J(u_n)$. Le principe de la méthode est le suivant.

Théorème 5. Soit $J \in \mathcal{C}^1(\mathbb{R}^n,\mathbb{R})$ et $\alpha > 0$ tel que J soit α-convexe, *i.e.* pour tout $x,y \in \mathbb{R}^n$,

$$J(x) - J(y) \geqslant \langle \nabla J(y) \ , \ x - y \rangle + \alpha \|x - y\|^2.$$

Alors, J admet un unique minimum (global) x^* et pour tout $x_0 \in \mathbb{R}^n$, la suite définie par $x_{n+1} = x_n + \mu_n \nabla J(x_n)$ où

$$\begin{cases} \mu_n &= \operatorname{argmin}_{\mu \in \mathbb{R}} J(x_n + \mu \nabla J(x_n)), & \text{si } x_n \neq x^* \\ \mu_n &= 0, & \text{sinon.} \end{cases}$$

est bien définie et converge géométriquement vers x^*.

Démonstration. Preuve de la convergence faite dans le cas de la fonctionnelle quadratique définie par

$$J : x \in \mathbb{R}^n \mapsto \frac{1}{2}\langle Ax, x\rangle + \langle b,x\rangle \in \mathbb{R}.$$

Le gradien de cette fonction est $\nabla J(x) = Ax - b$ et donc l'unique minimum global x^* de cette fonction vérifie $Ax^* = b$

$$\begin{aligned} \|x_{n+1} - x^*\| &= \|x_n + \mu_n(Ax_n - b) - x^*\| \\ &= \|x_n + \mu_n Ax_n - \mu_n Ax^* - x^*\| \\ &= \|(\mu_n A + I_n)(x_n - x^*)\| \\ &\leqslant \|\mu_n A + I_n\| \, \|x_n - x^*\| \end{aligned}$$

□

Il existe des variantes à cette méthodes :

- Gradient à pas constant : on fixe le pas μ_n à une constante μ. Il est convergant sous la condition $0 < \mu < 2\alpha/C^2$ où ∇J est C-lipschitzienne. Vitesse de convergence géométrique
- Gradient à pas conjugué : la direction de pente dépend du gradient $\nabla J(u_n)$ mais également des directions de descentes des itérations précédentes.

21 Méthode du gradient avec contraintes

Ici, on veut que l'inf soit atteint sur une contrainte K convexe fermé non vide

$$a \in K, J(a) = \inf_{x \in K \subset V} J(x),$$

l'inégalité d'Euler assurant l'existence. L'algorithme du gradient à pas constant s'adapte au cas avec contraintes à partir du fait suivant. Pour tout réel $\mu > 0$,

$$\langle \nabla J(u), v - u \rangle \geqslant 0, \quad \forall v \in K$$

se traduit par le fait que u est la projection orthogonale de $u - \mu \nabla J(u)$ sur K, *i.e.*

$$u = P_K(u - \mu \nabla J(u)), \quad \forall \mu > 0.$$

Le principe du gradient à pas fixe avec projection est donc

$$\left\{ \begin{array}{ccc} u_0 & \in & K \\ u_{n+1} & = & P_K(u_n - \mu \nabla J(u_n)) \\ \mu > 0 & \text{fixé} & \end{array} \right.$$

Cette méthode, applicable à une large classe de problèmes est en fait un leurre. En effet, la projection n'est pas connue explicitement en général et peut-être très difficile à déterminer. Une exception importante concerne, en dimension finie, les pavés $K = \Pi_i [a_i, b_i]$ avec éventuellement $a_i = -\infty$ ou $b_i = +\infty$. En effet, le projeté de $x = (x_1, \cdots, x_n)$ est le vecteur de compostante $\min(\max(a_i, x_i), b_i)$. (Ceci est une observation importante pour l'algorithme d'Uzawa).

22 Méthode de pénalisation

à rajouter

23 Méthode de Newton

On se place en dimension finie et on considère ici une fonction $F \in \mathcal{C}^2(\mathbb{R}^n)$. Soit u un zéro régulier de F *i.e.* $F(u) = 0$ et dF_u est inversible. Une formule de Taylor nous donne au voisinage de u

$$F(u) = F(v) + dF_v(u - v) + O(\|u - v\|) \implies u = v - (dF_y)^{-1} F(v) + O(\|u - v\|).$$

Le principe de la méthode est le suivant

$$\left\{ \begin{array}{ccc} u_0 & \in & \mathbb{R}^n, \\ u_{n+1} & = & u_n - (dF_{u_n})^{-1}[F(u_n)]. \end{array} \right.$$

Rappelons que l'on ne calcule pas l'inverse de la jacobienne mais que l'on résout un système linéaire par l'une des méthodes de la question 1. Le nom de la méthode fait écho à la méthode vue question 3 puisque chercher un minimum de F revient à chercher un zéro de dF. Cette méthode a des avantages et des inconvénients :

- La convergence est bien plus rapide : elle est quadratique. Attention, cette convergence n'a lieu que si F est de classe $\mathcal{C}^2$, si u_0 est assez proche de u et si u est régulier (hypothèses plus restrictives qu'avant).
- Elle nécessite la résolution d'un système linéaire, qui est couteux.
- Même dans les cas les plus simples de $\mathbb{R}$ la méthode peut diverger si u_0 n'est pas assez proche de u.
- La méthode peut converger mais vers un maximum car elle ne fait que chercher les zéros de la différentielle.

Question 6 : Qu'est-ce qu'un point d'équilibre pour un système différentiel autonome ? Sa stabilité ? Comment l'étudier ?

Référence de la réponse :
- M. Schatzman, Analyse numérique
- Zuily, Queffelec, Elements d'analyse

24 Introduction aux problèmes

Notre objectif ici est d'étudier les points d'équilibres d'un système différentiel. Faisons déjà deux observations :

- On peut le supposer du premier ordre puisqu'on peut transformer un système d'ordre p donné par $u^{(p)}(t) = f(t,u(t),u'(t),\dots,u^{(p-1)}(t))$ en un système du premier ordre $U'(t) = F(t,U(t))$ où

$$U(t) \;=\; \begin{pmatrix} u(t) \\ \vdots \\ u^{(p-1)}(t) \end{pmatrix} \quad \text{et} \quad F(t,X(t)) \;=\; \begin{pmatrix} x_1(t) \\ \vdots \\ x_{p-2}(t) \\ f(t,x_1(t),\dots,x_{p-1}(t)) \end{pmatrix}$$

- On peut le supposer autonome mais cela peut croître la dimension d'une unité et peut faire perdre le caractère linéaire du système. Soit le système du premier ordre $u'(t) = f(t,u(t))$ on peut le transformer en posant $s(t) = t$ en $U' = F(U)$ où

$$U(t) \;=\; \begin{pmatrix} u(t) \\ s(t) \end{pmatrix} \quad \text{et} \quad F(X) \;=\; \begin{pmatrix} f(x_1,x_2) \\ 1 \end{pmatrix} \quad \text{avec} \quad X \;=\; \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^d \times \mathbb{R}$$

  Dans ce cas, l'équation linéaire simple $x'(t) = a(t)x(t)$ a été remplacée par le système $X_1'(t) = a(X_2(t))X_1(t); X_2'(t) = 1$.

Une fois ces observations faites, on ne considère plus q'un système autonome du premier ordre

$$u'(t) = f(u(t)); \;\; u(0) = x_0$$

avec $f : \Omega \to \mathbb{R}^d$ supposée assez régulière pour posséder une unique solution maximale qu'on notera $\varphi(t,x_0)$.

25 Notions de points d'équilibre et de stabilité

Un point d'équilibre est un point $a \in \Omega$ tel que $f(a) = 0$. Par unicité de la solution, à chaque point d'équilibre correspond la solution constante $[t \in \mathbb{R} \mapsto a \in \Omega]$. Il existe différentes notions de stabilité

- a est un point d'équilibre stable si c'est un point d'équilibre et si

$$\forall \epsilon > 0, \exists \delta > 0, \forall x \in \Omega, \forall t > 0, \|x - a\| < \delta \implies \|\varphi(t,x_0) - a\| < \epsilon.$$

- a est un point d'équilibre instable si c'est un point d'équilibre et si la condition ci-dessus n'est pas vérifiée.

- a est un point d'équilibre localement asymptotiquement stable si c'est un point d'équilibre stable et s'il existe un voisinage $U_a \in \mathcal{V}(a) \subset \Omega$ tel que

$$\forall x \in U_a, \lim_{t\to+\infty} \varphi(t,x_0) = a.$$

- a est un point d'équilibre asymptotiquement stable si c'est un point d'équilibre stable et si

$$\forall x \in \Omega, \lim_{t\to+\infty} \varphi(t,x_0) = a.$$

26 Cas d'un système linéaire

Ici, on a le système

$$u' = Au + b$$

où $A \in \mathcal{M}_n(\mathbb{R})$ et $b \in \mathbb{R}^d$. Un point d'équilibre du système linéaire est un point u_0 tel que

$$Au_0 = -b$$

(*c.f.* Question 1). Quitte à poser $v(t) = u(t) - u_0$, on peut toujours supposer que le point d'équilibre à étudier est 0. À nouveau, quitte à poser $u(t) = v(t) + u_0$ et en utilisant la linéarité, on remarque que u est solution du système linéaire ssi v est solution du système homogène associé. Ainsi, on peut toujours se reconduire au système homogène. Dans la suite, on considère le système

$$u' = Au$$

dont on remarque que 0 est toujours un point d'équilibre et c'est le seul ssi A est inversible. En notant λ_i les valeurs propres de A, on a les résultats suivants :

- Le point 0 est asymptotiquement stable si et seulement si $\Re(\lambda_i) < 0, \forall i \in \{1,\dots,d\}$.
- Le point 0 est stable si et seulement si
  - soit $\forall i, \Re(\lambda_i) < 0$
  - soit $\forall i, \Re(\lambda_i) \leqslant 0$ et il existe un λ_i tel que $\Re(\lambda_i) = 0$ avec λ_i une valeur propre non-déféctive (*i.e.* multiplicité algébrique = multiplicité géométrique = 1)
- Le point 0 est instable si et seulement si
  - soit il existe $i \in \{1,\dots,d\}$ tel que $\Re(\lambda_i) > 0$
  - soit $\forall i, \Re(\lambda_i) \leqslant 0$ et il existe un λ_i tel que $\Re(\lambda_i) = 0$ avec λ_i une valeur propre déféctive.

Et on a la classification suivante pour un point critique :

- Il est hyperbolique si $\Re(\lambda_i) \neq 0, \forall i$,
- Il est un puits si $\Re(\lambda_i) < 0, \forall i$
- Il est une source si $\Re(\lambda_i) > 0, \forall i$
- Il est un point-selle s'il existe i et j tels que $\Re(\lambda_i)\Re(\lambda_j) \leqslant 0$

27 Cas d'un système non linéaire

Soit 0 un point d'équilibre du système

$$u'(t) = f(u(t)),$$

où f est $\mathcal{C}^1$ différentiable en 0. On note df_0 sa différentielle en ce point. On a les résultat suivant

- Si toute valeur propre de df_0 est de partie réelle strictement négative, alors 0 est asymptotiquement stable
- S'il existe une valeur propre de df_0 de partie réelle strictement positive alors 0 est instable
- Si pour tout valeur propre de df_0 est de partie réelle négative dont une qui est imaginaire pure alors on ne sait pas conclure.

Pour étudier la stabilité dans le cas non linéaire, une condition suffisante de stabilité du point 0 est donnée par l'existence d'une fonction de Lyapunov : une fonction différentiable sauf peut-être en 0, qui s'annule en 0 et est strictement positive ailleurs et telle que $\langle \nabla \mathcal{L}(x), f(x) \rangle \leqslant 0$ partout.

Question 7 : Comment approcher la solution d'une EDO et analyser la convergence d'une telle approximation ?

Référence de la réponse :
- L. Dumas, Modélisation à l'oral de l'agrégation
- M. Schatzman, Analyse numérique
- A. Quarteroni, R. Sacco F. Saleri, Méthodes numériques
- A. Yger, Calcul scientifique et symbolique

28 Introduction aux problèmes

Toutes les équations différentielles ordinaires, et , à plus forte raison, tous les systèmes différentiels, n'admettent pas de solution explicite, même en faisant intervenir des fonctions spéciales très compliquées : il y a énormément plus d'équations qu'on ne sait pas intégrer explicitement que d'équations qu'on sait intégrer. On souhaite donc, grâce à l'analyse numérique, avoir des informations qu'on ne pourrait pas obtenir autrement : informations de nature qualitative en temps long comme la périodicité, monotonie,... Il y a plusieurs soucis :

- les erreurs de troncature sont un vrai souci ici : elles ne sont plus négligeables si on a beaucoup d'itérations. Une façon de les modéliser serait de les considérer comme des perturbations aléatoires indépendantes (mais on est *a priori* dans un cadre déterministe...)
- le choix d'un schéma d'approximation d'un système différentiel dépend de l'analyse qualitative qu'on fait sur le système : si on s'attend à observer une propriété comme par exemple le fait d'avoir des solutions bornées uniformément en temps, on veut trouver une méthode conservant assez bien cette propriété
- En général, dans les estimations de convergence des approximations numériques, il apparaît une constante qui croit exponentiellement en temps. Ainis, on ne sait pas montrer la convergence en temps infini : montrer que le comportement qualitatif des approximations discrètes d'un système donne une information pertinente sur le comportement du système est en soi déjà compliqué.

Avant de pourvoir simuler numériquement un système différentiel, il faut s'assurer qu'il soit bien posé :

- Un système de Cauchy est mathématiquement bien posé s'il existe une unique solution maximale dépendant continûment de la donnée initiale.
- Un système est numériquement bien posé s'il est mathématiquement bien posé et si la dépendance en la donnée initiale est numériquement contrôlable. Par exemple,

  $\begin{cases} y'(t) = 3y(t) - 1 \\ y(0) = \frac{1}{3} \end{cases}$ a pour solution $y(t) = \frac{1}{3}$

  $\begin{cases} y'(t) = 3y(t) - 1 \\ y(0) = \frac{1}{3} + \varepsilon \end{cases}$ a pour solution $\widetilde{y}(t) = \frac{1}{3} + \varepsilon \mathrm{e}^{3t}$

  Donc avec $\varepsilon = 10^{-17}$ qui est la précision machine, $y(30) - \widetilde{y}(30) = 10^{22}$. Ce problème, bien que mathématiquement bien posé, ne l'est pas numériquement.
- Un système est bien conditionné s'il est numériquement bien posé et si la méthode numérique envisagée approche de manière satisfaisante la solution exacte avec un pas de temps raisonnable. (Nous verrons plus tard un exemple de problème mal conditionné).

29 Méthodes à un pas

On cherche à approcher la solution du problème bien posé suivant :

$$\left\{ \begin{array}{lcl} y'(t) & = & f(t,y(t)),\ t \in [t_0,t_0+T] \\ y(t_0) & = & y_0 \in \mathbb{R}^d. \end{array} \right. \tag{0.1}$$

Une méthode de résolution approchée consiste d'abord à effectuer une N-subdivision de l'intervalle avec $N \in \mathbb{N}^*$:

$$[t_0,t_0+T] = \bigcup_{n=0}^{N-1} [t_n,t_{n+1}]$$

puis à construire une suite $(y_n)_{n\in\mathbb{N}}$ telle que y_n approche $y(t_n)$. On se limite au pas de temps uniforme, *i.e.* $t_n = t_0 + nh$ avec $h = T/N$. On définit tout d'abord un schéma numérique à un pas par

$$y_{k+1} = y_k + hF(t_k,y_{k+1},y_k;h)$$

avec y_0 la donnée initiale, $h > 0$ fini, et $F : [t_0,t_0+T] \times \mathbb{R} \times [0,T] \to \mathbb{R}$ continue. La méthode est dite explicite si F est indépendante de y_{k+1} et implicite sinon.
Pour savoir si une méthode numérique est satisfaisante, il y a plusieurs notion.

Définition 2. La consistance :

$$\lim_{N\to+\infty} \sum_{n=0}^{N-1} |y(t_{n+1}) - y(t_n) - hF(t_n,y(t_n),h)| = 0$$

nous assure que la méthode considéré est cohérente avec le problème initial.

Définition 3. La stabilité : si on perturbe de ε_n la méthode alors on peut numériquement contrôler la différence des solutions apportées :

$$\max_{0\leqslant n\leqslant N} |y_n - z_n| \leqslant M \left(|y_0 - z_0| + \sum_{k=0}^{N-1} |\varepsilon_k| \right)$$

nous assure que les éventuelles erreurs numériques commises dans l'évaluation des termes de la suite d'approximation sont contrôlables.

Définition 4. La convergence :

$$\lim_{N\to+\infty} \max_{0\leqslant n\leqslant N} |y(t_n) - y_n| = 0$$

nous assure que si on affine notre pas de temps, la méthode tend vers la solution exacte.

Définition 5. L'ordre : le plus grand $p \in \mathbb{N}^*$ tel que

$$\exists C > 0, \quad \forall h > 0, \quad \max_{0\leqslant n\leqslant N} |y(t_n) - y_n| \leqslant Ch^p.$$

nous informe sur la vitesse de convergence.

Il existe encore la notion de stabilité inconditionnelle (ou A-stabilité) décrivant la façon dont les erreurs faussent le comportement de la solution approchée pour $t \to +\infty$.

Notamment, si une méthode explicite à un pas est stable et consistante alors elle est convergente.

Récapitulatif des méthodes classiques pour le problème de Cauchy bien posé 0.1 :

Méthode	principe : $U_{n+1} = U_n +$	ordre	A-stab.
Euler expl. (rect. gauche)	$hf(t_n,U_n)$	1	non
Euler impl. (rect. droite)	$hf(t_{n+1},U_{n+1})$	1	oui
Crank-Nicolson (trapèzes)	$\frac{h}{2}\left[f(t_n,U_n) + f(t_{n+1},U_{n+1})\right]$	$f \in \mathcal{C}^2 : 2$	oui
Heun (CN explicite)	$\frac{h}{2}\left[f(t_n,U_n) + f(t_{n+1},U_n + hf(t_n,U_n))\right]$	$f \in \mathcal{C}^2 : 2$	non

Etude des schémas d'Euler. Les schémas d'Euler viennent de l'écriture de la solution d'une équation différentielle sous la forme

$$y(t) = \int_0^t f(s,y(s))\mathrm{d}s.$$

Pour calculer l'intégrale, on l'approche par la méthode des rectangles à gauche (Euler explicite) ou à droite (Euler implicite). La méthode d'Euler explicite est :

- consistante, puisque f est supposée $\mathcal{C}^1$, y est $\mathcal{C}^2$ et on applique Taylor-Lagrange.
- stable, donné par la global lipschitzianité de f et l'inégalité $(1 + Lh)^N \leqslant \mathrm{e}^{LhN}$.
- convergente d'ordre 1 car la suite $\{y(t_n)\}_n$ vérifie le schéma modulo une erreur correspondant à l'erreur de consistance donc puisque le schéma est stable et consistant, il converge.

L'exponentielle de la stabilité (apparaissant également dans la convergence) peut poser soucis. En effet, avec le problème

$$\begin{cases} y' &= \lambda y, \\ y(0) &= 1, \end{cases}$$

l'itération d'Euler explicite donne pour solution $y_n = (1+\lambda h)^n$ alors que Euler implicite donne $y_n = \frac{1}{(1-\lambda h)^n}$ l'une étant divergente avec λ, l'autre non. □

Il existe des schéma plus complexes d'ordre plus élevé (Runge-Kutta d'ordre 4) mais les coefficients deviennent vite compliqués et les ordres trop élevés ne sont pratiquement pas utilisés.

Question 8 : Quand et comment la série de Fourier d'une fonction converge-t-elle ?

Référence de la réponse :
El Amrani, Analyse de Fourier dans les espaces fonctionnels

30 Introduction aux problèmes

Notre objectif ici est de répondre aux problèmes suivant. Soit f une fonction 2π-périodique et $S_N(f) : x \mapsto \sum_{n=-N}^{N} c_n(f)e^{ixn}$, alors

- Pour quelles fonctions y a-t-il convergence de $S_N(f)$?
- S'il y a convergence, la limite est-elle f ?
- De quel type de convergence s'agit-il ?
  - pour la norme $\|\cdot\|_2$?
  - convergence simple ? si oui, en quels points ?
  - convergence uniforme ? si oui, sur quels ensembles ?
  - au sens de Cesàro ?

En effet, dans un cas général, la série de Fourier de f peut converger ou non et si elle converge, elle peut converger ou non vers f. On note S_N la série de Fourier, σ_N la somme de Cesàro, D_N le noyau de Dirichlet, K_N le noyau de Fejér.

31 Cas des fonctions $\mathcal{C}_{2\pi}$

Grâce au théorème de Banach-Steinhaus on peut montrer qu'il existe un ensemble dense de fonctions continues 2π-périodiques dont la série de Fourier diverge au moins en un point. Cependant,

Théorème 6 (Théorème de Fejér)**.** Si f est continue 2π-périodique alors **au sens de Cesàro**

$$\|\sigma_N(f)\|_\infty \leqslant \|f\|_\infty, \qquad \lim_{N\to\infty} \|\sigma_N(f) - f\|_\infty = 0.$$

Démonstration. Argument clé : Heine pour récupérer l'uniforme continuité + découpage avec le fait que l'intégrale du noyau de Féjer vaut 2π. □

Théorème 7. Si f est continue 2π-périodique et $x_0 \in \mathbb{R}$ alors **ponctuellement**

$$\left(\lim_{N\to+\infty} S_N(f)(x_0) = l\right) \implies f(x_0) = l.$$

Théorème 8. Si f est continue 2π-périodique et si la série de Fourier converge simplement sur $\mathbb{R}$ alors **ponctuellement** pour tout $x \in \mathbb{R}$,

$$f(x) = \sum_{n=+\infty}^{+\infty} c_n(f) e^{inx}.$$

Théorème 9. Si f est continue 2π-périodique et si $\sum_{n\in\mathbb{Z}} |c_n(f)|$ converge alors **ponctuellement**

$$f(x) = \sum_{n=+\infty}^{+\infty} c_n(f) e^{inx}$$

la série converge **absolument et uniformément** sur $\mathbb{R}$.

Théorème 10. Si f est continue 2π-périodique et $\mathcal{C}^1$ par morceaux , alors sa série de Fourier converge **normalement** sur $\mathbb{R}$ et on a

$$f = \sum_{n=+\infty}^{+\infty} c_n(f) e_n.$$

32 Cas des fonctions $L^1_{2\pi}$ ou $L^p_{2\pi}$

- (Dirichlet) Soit $f \in L^1_{2\pi}$ et $0 \leq x \leq 2\pi$ et si f admet des limites à gauche et à droite en x et $t \mapsto \frac{f(x+t)-f(x^+)}{t}$ et $t \mapsto \frac{f(x-t)-f(x^-)}{t}$ sont bornées au voisinage de $t = 0^+$, alors

  $$S_n(f)(x) \xrightarrow[n\to\infty]{} \frac{f(x^+) + f(x^-)}{2}.$$

  (parité de D_n pour introduire l'intégrale de $f(x-\cdot) + f(x+\cdot)$ contre D_n, puis découper en deux bouts et utiliser Riemann-Lebesgue sur $\frac{f(x\pm\cdot)-f(x^\pm)}{\sin(\cdot/2)}$ qui est L^1 par inégalité de concavité de sin)

- Si $f \in L^1_{2\pi}$ et $(u_n)_\mathbb{Z}$ une suite de nombres complexes. Si la suite $S_N = \sum_{-N}^{N} u_n e^{inx}$ converge vers f dans $L^1_{2\pi}$ (en particulier si elle converge **uniformément**) alors c'est la série de Fourier de f (*i.e.* $u_n = c_n(f) \forall n$).

- Si $S_N(f)$ converge vers f dans L^p alors il en va de même de $\sigma_N(f)$, donc convergence **au sens de Cesàro**.

- Si f,g sont $L^1_{2\pi}$ et si elles sont égale sur un voisinage ouvert de x_0 alors $\lim_{n\to\infty}[S_n(f)(x_0) - S_n(g)(x_0)] = 0$.

- (Fejér) Si $f \in L^p_{2\pi}$,$p \in [1, +\infty[$ alors **convergence de Cesàro dans L^p**

  $$\|\sigma_N(f)\|_p \leqslant \|f\|_p, \qquad \lim_{N\to\infty} \|\sigma_N(f) - f\|_p = 0.$$

- Si $f \in L^2_{2\pi}$, alors **dans L^2**

  $$\lim_{N\to\infty} \|S_N(f) - f\|_2 = 0, \qquad \|f\|_2^2 = \sum_{n=-\infty}^{+\infty} |c_n(f)|^2.$$

- Si $f \in L^1_{2\pi}$ et $x_0 \in \mathbb{R}$ et si $\lim_{x \to x_0^-} f(x) = f(x_0^-)$ et $\lim_{x \to x_0^+} f(x) = f(x_0^+)$ existent et f possède une dérivée à gauche et à droite en x_0, alors **ponctuellement**

$$\lim_{N \to \infty} S_N(f)(x_0) = \frac{f(x_0^+) + f(x_0^-)}{2}.$$

  Si f est de plus continue, on a donc convergence **ponctuelle**.

- Si f est L^2 et si $\sum_{n \in \mathbb{Z}} |c_n(f)|$ converge alors f est continue et la série de Fourier converge **normalement** vers f

On peut encore signaler quelques remarques :

- Il existe encore d'autres critères utilisant les fonctions à variations bornées, les fonctions hölderiennes, ou le critère de convergence d'Abel.

- La convergence en norme 2 ne signifie pas qu'il y a convergence ponctuelle.

- Plus la fonction est régulière, plus les coefficients de Fourier décroissent rapidement.
  - $L^1 \implies c_n(f) \to 0$
  - $L^2 \implies \sum |c_n(f)|^2 < +\infty$
  - $\mathcal{C}^1 \implies \sum |c_n(f)| < +\infty$
  - $\mathcal{C}^2 \implies |n^2 c_n(f)| \to 0$
  - $\mathcal{C}^\infty \implies |n^k c_n(f)| \to 0 \quad \forall k \in \mathbb{N}$

Question 9 : Le transport linéaire 1D. Qu'est-ce que la méthode des caractéristiques ? Quelles sont les propriétés de régularité des solutions du transport ?

Référence de la réponse :
- L. Dumas, Modélisation à l'oral de l'agrégation.
- cours de F. Bolley

33 Introduction aux problèmes

On s'intéresse ici à l'équation

$$\begin{cases} \dfrac{\partial u}{\partial t}(x,t) + a(x,t)\dfrac{\partial u}{\partial x}(x,t) = f \\ u(0,x) = u_0(x). \end{cases}$$

sur $\Omega \times I$ avec $\Omega \subset \mathbb{R}$ et $I \subset \mathbb{R}^+$. Cette équation est linéaire et est du premier ordre. C'est l'équation de transport : une équation dépendante du temps, modélisant une propagation à vitesse finie. C'est une équation hyperbolique, comme l'équation des ondes, par exemple.

On cherche à répondre à plusieurs questions :

- Quelles conditions aux bords mettre pour donner une existence/unicité de la solution ? S'il n'y a pas unicité, quelle solution choisir ?
- Quelle est la solution explicite ? (question rarement répondue)
- Quel est le cadre mathématique permettant de
  - montrer l'existence, l'unicité
  - vérifier la well-poseness de l'équation (si la valeur initiale, la condition au bord ou un coefficient de l'équation change un peu, est-ce que la solution va beaucoup varier ?)
- quelles sont les propriétés de la solution (périodicité, monotonie, régularité : continuité, dérivabilité, t petit, t grand, x grand,...) ?
- trouver des schémas numériques permettant d'approcher correctement les solutions.

34 Coefficient constant et sans second membre

On a l'équation

$$\begin{cases} \dfrac{\partial u}{\partial t}(x,t) + a\dfrac{\partial u}{\partial x}(x,t) &= 0 \\ u(x,0) &= u_0(x). \end{cases}$$

Si on pose $v(x,t) = u(x+at,t)$ alors $\frac{\partial v}{\partial t} = a\frac{\partial u}{\partial x} + \frac{\partial u}{\partial t}$ c'est à dire v est solution de

$$\begin{cases} \dfrac{\partial v}{\partial t}(x,t) &= 0 \\ v(x,0) &= u_0(x). \end{cases}$$

C'est à dire, u est de classe $\mathcal{C}^1$ et solution du problème si et seulement si $v(x,t) = u_0(x)$ pour tout $(x,t) \in \mathbb{R} \times [0,T]$. Ainsi,

$$u(x,t) = u_0(x - ta)$$

est l'unique solution de classe $\mathcal{C}^1$ (c'est bien une solution).

35 Coefficient global Lipschitz et sans second membre

On a l'équation

$$\left\{ \begin{array}{rcl} \dfrac{\partial u}{\partial t}(x,t) + a(x,t)\dfrac{\partial u}{\partial x}(x,t) & = & 0 \\ u(x,0) & = & u_0(x). \end{array} \right. \tag{0.2}$$

et on suppose ici que a est $\mathcal{C}^1(\mathbb{R} \times [0,T])$ et qu'il existe $L > 0$ telle que pour tous $x,x' \in \mathbb{R}$ et $t \in [0,T]$, a vérifie

$$|a(x,t) - a(x',t)| \leqslant L|x - x'|.$$

Définition 6. On définit l'équation caractéristique associée à 0.2 par l'équation de paramètres (x,t) et d'inconnue $X(s) \in \mathbb{R}$, $s \in [0,T]$

$$\left\{ \begin{array}{rcl} X'(s) & = & a(X(s),s) \\ X(t) & = & x. \end{array} \right.$$

Le théorème de Cauchy-Lipschitz assure que sous les hypothèses faites sur a il existe une unique solution globale à l'équation caractéristique. Cette solution est notée $X(s;x,t)$. La courbe $\{X(s;x,t); s \in [0,T]\}$ est appelée courbe caractéristique passant par (x,t).

Ces courbes vérifient plusieurs propriétés :

- elles ne se coupent pas : pour tout couple de paramètres (x,t) il existe une unique courbe passant par (x,t) (c'est l'unicité du théorème de Cauchy-Lipschitz). La solution u est constante le long des caractéristiques.
- $X(t;x,t) = x$
- $X(t_3;X(t_2;x,t_1),t_2) = X(t_3;x,t_1)$
- Si a est $\mathcal{C}^k$, alors $(x,t) \mapsto X(t_0;x,t)$ est $\mathcal{C}^k$ (application du théorème de point fixe de Banach à paramètres). Ici vu nos hypothèses, la fonction $X : (s;x,t) \mapsto X(s;x,t)$ est $\mathcal{C}^1$ et la fonction $x \mapsto X(s;x,t)$ est un $\mathcal{C}^1$-difféomorphisme d'inverse $x \mapsto X(t;x,s)$ par propriété de semi-groupe.
- si u_0 est de classe $\mathcal{C}^1$ et X solution de l'équation caractéristique, alors le problème admet une unique solution $\mathcal{C}^1$ donnée par $u(x,t) = u_0(X(0;x,t))$. En particulier, pusique $x \mapsto X(0;x,t)$ est une bijection, $\|u(\cdot,t)\|_\infty = \|u_0\|_\infty$.

On peut vérifier que c'est cohérent avec le cas où a est une constante. De plus, si a est constante, les courbes caractéristiques sont des droites parallèles de pente a : $X'(s) = a$ signifie que $X(s)$ se propage à vitesse constante finie a.

On a les propriétés suivantes.

- On a déjà vu qu'une donnée initiale bornée et $\mathcal{C}^1$ donne une solution bornée et $\mathcal{C}^1$.
- Si de plus, $|a(0,t)| \leqslant M$ et si u_0 est à support compact, alors $u(\cdot,t)$ est encore à support compact pour tout $t \geqslant 0$. Plus précisément, si $u_0(x) = 0$ pour $|x| \geqslant R$, alors pour tout $t \geqslant 0$,

$$u(x,t) = 0 \text{ pour } |x| \geqslant e^{Lt}\left(\frac{M}{L} + R\right).$$

- Si u_0 est positive, $\mathcal{C}^1$ et à support compact, alors $u(x,t)$ est positive, $\mathcal{C}^1$, à support compact : $u(\cdot,t) \in L^p(\mathbb{R})$.

36 Coefficients constants et avec second membre

On a l'équation

$$\left\{ \begin{array}{rcl} \dfrac{\partial u}{\partial t}(x,t) + a\dfrac{\partial u}{\partial x}(x,t) + bu(x,t) & = & f(x,t) \\ u(x,0) & = & u_0(x). \end{array} \right. \tag{0.3}$$

On a vu le cas $b = f = 0$, supposons maintenant $b = cst$ et f une fonction. En reprenant notre $v(x,t) = u(x + at,t)$ on a maintenant u solution de 0.3 si et seulement si

$$\left\{ \begin{array}{rcl} \dfrac{\partial v}{\partial t}(x,t) + bv(x,t) & = & f(x + at,t) \\ v(x,0) & = & u_0(x) \end{array} \right.$$

qui a pour solution (équation différentielle du premier ordre en t) :

$$v(x,t) = u_0(x)e^{-bt} + \int_0^t f(x + as,s)e^{b(s-t)}\mathrm{d}s.$$

D'où la solution

$$u(x,t) = u_0(x - at)e^{-bt} + \int_0^t f(x + a(s - t),s)e^{b(s-t)}\mathrm{d}s.$$

37 Le cas général

On peut à nouveau utiliser l'équation caractéristique pour trouver une solution du problème

$$\left\{ \begin{array}{rcl} \dfrac{\partial u}{\partial t}(x,t) + a(x,t)\dfrac{\partial u}{\partial x}(x,t) + b(x,t)u(x,t) & = & f(x,t) \\ u(x,0) & = & u_0(x). \end{array} \right.$$

qui sera $\mathcal{C}^1$.

On peut également préciser que tout ce qui a été fait est encore valable en dimension d finie quelconque. On pourrait également étudier l'équation non linéaire (par exemple l'équation de Burger où on remplace $a(x,t)$ par $u(x,t)$ avec $b = f = 0$) mais cela fait intervenir les espaces de Sobolev... [Je crois que dans le PGCD on montre l'existence et l'unicité de l'équation de Burgers par la méthode des caractéristiques : on peut résoudre l'équation des caractéristiques grâce au théorème des fonctions implicites mais je ne sais plus comment... (Matou)]

38 Étude d'un schéma décentré pour le transport linéaire à vitesse contante

On veut un schéma pour approcher la solution du problème suivant :

$$\left\{ \begin{array}{rcll} \partial_t u(x,t) + a\partial_x u(x,t) & = & 0, & x \in [0,L], \quad t \in [0,T] \\ u(0,x) & = & u_0(x), & x \in [0,L]. \end{array} \right.$$

On introduit les discrétisations régulières

$$(t_n)_{n\in[\![0,N]\!]}, \qquad (x_j)_{j\in[\![0,J]\!]},$$

telles qu'on ait des pas de discrétisation Δx et Δt et

$$\forall n \in [\![0,N]\!], \quad t_n = n\Delta t, \qquad \forall j \in [\![0,J]\!], \quad x_j = j\Delta x.$$

On souhaite approcher $u(x_j,t_n)$ par une valeur u_j^n. On initialise le procédé avec $u_j^0 = u_0(x_j)$ et pour $a > 0$, on considère le schéma décentré gauche (ou amont) :

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + a\frac{u_j^n - u_{j-1}^n}{\Delta x} = 0.$$

On considère les conditions périodiques au bord $u_0^n = u_J^n$. En notant $R^n = (u_1^n, \ldots, u_J^n)$ on obtient le schéma sous forme matricielle :

$$R^{n+1} = AR^n,$$

$$A = \begin{pmatrix} 1-\lambda & 0 & 0 & \cdots & \lambda \\ \lambda & 1-\lambda & 0 & \cdots & 0 \\ 0 & \lambda & 1-\lambda & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \lambda & 1-\lambda \end{pmatrix} \in \mathcal{M}_J(\mathbb{R})$$

avec $\lambda = a\frac{\Delta t}{\Delta x}$.

Le schéma est stable si la matrice A a un rayon spectral plus petit que 1. Or ses valeurs propres sont les $\{1 - \lambda(1-\omega^l), l = 0, \ldots J-1\}$ où $\omega = \mathrm{e}^{2i\pi/J}$. Ainsi le schéma est stable si $\lambda \leqslant 1$. De plus, on a qu'il est consistant. Donc sous la condition $\lambda \leqslant 1$, le schéma est convergent d'ordre 1. On procède de même si $a < 0$ avec le schéma décentré droit (ou aval) :

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + a\frac{u_{j+1}^n - u_j^n}{\Delta x} = 0$$

qui converge si $-1 \leqslant \lambda < 0$.

Question 10 : L'équation de Poisson. Comment utiliser le théorème de Lax-Milgram ? Quelles stratégies numériques pour approcher la solution ?

Référence de la réponse :
- Cours de F. Bolley
- Cours de T. Gerdahoui
- L. Dumas, Modélisation à l'oral de l'agrégation

39 Introduction aux problèmes

On s'intéresse ici à l'équation

$$\left\{ -u'' = f \right.$$

sur $\mathbb{R}^d$. Cette équation est linéaire et est d'ordre d. C'est l'équation de Laplace ou de Poisson (idk) : une équation indépendante du temps, modélisant un phénomène stationnaire. C'est une équation elliptique. On cherche à répondre aux même questions que pour l'équation de transport (Question 9).

Une solution de l'équation n'est pas toujours unique (par exemple, si $f = 0$ tout polynôme de degré 1 ou 0 est solution. Pour avoir unicité, on doit poser des conditions aux bords (semblables aux conditions aux limites des EDO) en imposant la valeur de $u(0)$,$u'(0)$ ou autre. On s'intéresse aux conditions de Dirichlet : On s'intéresse ici à l'équation

$$\left\{ \begin{array}{lcl} -u''(x) & = & f \\ u(a) = u(b) & = & 0 \end{array} \right.$$

sur $[a,b]$. On pourrait en fait étudier un problème plus général

$$\left\{ \begin{array}{lcl} -(pu')' + ru' + qu & = & f \\ u(a) = u(b) & = & 0 \end{array} \right.$$

Ces sont des problèmes qui ne sont pas de Cauchy puisque les conditions ne sont pas données en terme de dérivées mais en terme de conditions aux limites. On appelle ces problème le problèmes de Dirichlet. Il y a deux méthodes pour faire face à ces problèmes.

- La méthode variationnelle :
  - donner la définition d'une solution faible du problème et montrer qu'une solution classique est nécessairement une solution faible. Cette étape est fondée sur les espaces de Sobolev,
  - montrer l'existence et l'unicité de la solution faible, étape fondée sur les théorèmes de Riesz et Lax-Milgram,
  - monter que si f est continue, la solution faible est une solution classique $\mathcal{C}^2$.
- la méthode de tir :
  - montrer l'existence d'une solution à partir de deux problèmes de Cauchy en choisissant un paramètre adapté au problème,
  - montrer l'unicité de la solution via le principe du maximum.

Pour avoir au final une solution (forte), il est nécessaire d'avoir f continue. Cette solution est $\mathcal{C}^2$.

40 Méthode variationnelle

ÉTAPE 1 : définition d'une solution faible du problème et montrer qu'une solution classique est nécessairement une solution faible.
On pose

$$a(u,v) \coloneqq \int_0^1 pu'v' + \int_0^1 quv \quad \text{et} \quad l(v) \coloneqq \int_0^1 fv.$$

Dans le cas de l'équation de Laplace :

$$a(u,v) \coloneqq \int_0^1 u'v' \quad \text{et} \quad l(v) \coloneqq \int_0^1 fv.$$

Une fonction $u \in H_0^1(0,1)$ est une **solution faible** du problème de Dirichlet si pour tout $v \in H_0^1(0,1)$, $a(u,v) = l(v)$.
Soit $u \in \mathcal{C}^2$ une solution forte. Etant continues, u et u' sont de carré intégrable sur l'intervalle et puisque u vérifie les conditions aux limites on a $u \in H_0^1(0,1)$. De plus, en multipliant par $v \in \mathcal{C}_c^1(0,1)$ et en intégrant on peut faire une intégration par partie et utiliser le fait que $\mathcal{C}_c^1(0,1)$ est dense dans $H_0^1(0,1)$. Donc u est bien une solutoin faible.

ÉTAPE 2 : montrer l'existence et l'unicité de la solution faible.
On choisit $f \in L^2(0,1)$. On se place sur $(H_0^1(0,1), \|\cdot\|)$ avec $\|v\| \coloneqq \|v'\|_2$. C'est un espace de Hilbert. La fonction a est bilinéaire, continue, coercive et l est une forme linéaire continue donc on peut appliquer le théorème de Lax-Milgram : d'où l'existence d'une unique solution faible. (En fait, pour l'équation de Laplace a est même symétrique et est le produit scalaire donc on pourrait plus simplement utiliser le théorème de représentation de Riesz).

ÉTAPE 3 : si f est continue, la solution faible est une solution classique $\mathcal{C}^2$.
La solution faible est $\mathcal{C}^2$ car f est continue (utilise un peu les distributions et la notion de dérivée faible) et donc la dérivée seconde de la solution faible est égale à sa dérivée faible qui vérifie l'équation. Ainsi, ponctuellement u vérifie l'équation, et étant dans $H_0^1(0,1)$, elle vérifie les conditions aux limites.

41 Méthode de tir

On s'interesse au problème un poil différent suivant :

$$\begin{cases} -(pu')' + qu &= f \\ u(a) &= u_a \\ u(b) &= u_b \end{cases}$$

ÉTAPE 1 : existence d'une solution.
Soient les deux problèmes de Cauchy suivant :

$$\begin{cases} -(pu_1')' + qu_1 &= f \\ u_1(a) &= u_a \\ u_1'(a) &= 0 \end{cases} \quad \text{et} \quad \begin{cases} -(pu_2')' + qu_2 &= 0 \\ u_2(a) &= 0 \\ u_2'(a) &= 1 \end{cases}$$

Pour l'équation de Laplace, on aurait

$$\begin{cases} -u_1'' &= f \\ u_1(a) &= u_a \\ u_1'(a) &= 0 \end{cases} \quad \text{et} \quad \begin{cases} -u_2'' &= 0 \\ u_2(a) &= 0 \\ u_2'(a) &= 1 \end{cases}$$

Il existe une unique solution gloable à chaque problème et on considère $u := u_1 + ku_2$ qui est bien solution de l'équation voulue. On doit simplement choisir k tel que u vérifie les conditions aux limites. Une ipp montre que

$$k = \frac{u_b - u_1(b)}{u_2(b)}$$

a un sens et convient.

Dans le cas de l'équation de Laplace, on a simplement $u_2(x) = x - a$ et

$$u(x) = u_1(x) + (u_b - u_1(b))\frac{x-a}{b-a}$$

.

ÉTAPE 2 : unicité de la solution.
On définit l'opérateur

$$\mathcal{L} : u \in \mathcal{C}^2([a,b],\mathbb{R}) \mapsto -(pu')' + qu \in \mathcal{C}^0([a,b],\mathbb{R}).$$

qui est simplement $-\Delta$ dans le cas de l'équation de Laplace. Le principe du maximum nous donne que si $u \in \mathcal{C}^2$ et $\mathcal{L}(u) \leqslant 0$ alors

$$\max_{x\in[a,b]} u(x) \leqslant \max(0,u(a),u(b)).$$

Et on a un résultat similaire avec des min dans le cas où l'opérateur est positif. Dans le cas de l'équation de Laplace, on regarde simplement si u est convexe ou concave
En supposant qu'il existe u et v solutions du problème, on pose $w := u - v$ et puisque les fonctions sont solutions de la même équation, $\mathcal{L}(w) = 0$ or $w(a) = w(b) = 0$ donc w est la fonction nulle. Dans le cas de l'équation de Laplace, c'est évident (w est un polynôme de degré 1 avec deux racines).

42 Méthodes des différences finies

On veut approcher le problème suivant

$$\begin{cases} -u'' + qu &= f \\ u(a) &= 0 \\ u(b) &= 0 \end{cases}$$

Puisque le problème est indépendant du temps, on peut faire quelque chose de très similaire aux EDO : on approche la dérivée seconde par une différence centrée

$$u''(x) \simeq \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}$$

pour h petit et on introduit une subdivision régulière à $N+1$ points (x_i) de l'intervalle $I = [0,1]$. On obtient le schéma

$$\left(\frac{1}{h^2}A + Q(X)\right)Y = F(X)$$

où A est la matricien du laplacien, Q est la matrice diagonale de termes $(q(x_i))_{1\leqslant i\leqslant N-1}$, Y est un vecteur colonne et F est le vecteur des points $f(x_i)$. Dans le cas de l'équation de Laplace, on obtient un système linéaire

$$\frac{1}{h^2}AY = f(X).$$

On rappelle que la matrice A a les propriétés suivantes

- elle est symétrique définie positive (en particulier inversible),
- on peut expliciter ses valeurs propres $(4\sin^2(k\pi/2N))$,
- elle est mal conditionnée $cond_2(A) \sim 4N^2/\pi^2$.
- monotone

Dans le cas de l'équation de Laplace avec $f \in \mathcal{C}^2$, le schéma est

- consistant,
- d'ordre 2,
- stable en norme infinie,
- donc : convergent d'ordre 2.
- puisque A est tridiagonale symétrique définie positive, la résolution du système linéaire se fait en $O(N)$ mais la matrice étant mal conditionnée, les petites perturbations influent grandement sur la solution.

On peut également résoudre ce système linéaire avec une méthode de périodisation du Laplacien puis d'utilisation de la FFT qui se fait en $O(n \log n)$. Le principe est le suivant : à partir d'une équation $Ax = b$ d'inconnue $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ où A est la matrice du Laplacien, on en déduit une autre équation en transformant les vecteurs de la façon suivante : on considère maintenant le vecteur $X := (\underbrace{x_0}_{:=0}, x_1, \ldots, x_n, \underbrace{x_{n+1}}_{:=0}, -x_1, \ldots, -x_n)$ et la matrice modifiée de taille $N := 2n + 2$

$$A_{\text{per}} := \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & -1 \\ -1 & 2 & \ddots & 0 & \cdots & 0 \\ 0 & \ddots & & & \ddots & \vdots \\ \vdots & \ddots & & \ddots & -1 & 0 \\ 0 & \cdots & 0 & -1 & 2 & -1 \\ -1 & 0 & \cdots & 0 & -1 & 2 \end{pmatrix}.$$

On remarque qu'elle est circulante c'est-à-dire que si on pose $a := (a_0, \ldots, a_{n-1}) := (2, -1, 0, \ldots, 0, -1)$, alors on a

$$(A_{\text{per}})_{i,j} = a_{|j-i|}.$$

Ainsi on obtient le calcul suivant :

$$(AX)_k = \sum_{l=0}^{N-1} a_{|k-l|} x_l$$

qui est une convolution discrète périodique. Un calcul permet ensuite d'établir que $(\mathcal{F}(AX))_n = \mathcal{F}(a)_n \cdot \mathcal{F}(X)_n$, où $\mathcal{F}$ est la transformation de Fourier discrète $\mathcal{F}(x) = \left(\sum_{k=0}^{N-1} x_k e^{-2i\pi \frac{k}{N} n}\right)_n$. Ainsi si on considère la division de vecteurs au sens de la division coefficient par coefficient, alors

$$X = \mathcal{F}^{-1}(\mathcal{F}(b)/\mathcal{F}(a)), \text{ où } \mathcal{F}^{-1}(x) = \left(\frac{1}{N} \sum_{k=0}^{N-1} x_k e^{2ni\pi \frac{k}{N}}\right)_n.$$

Si on résolvait l'équation initiale sans périodisation mais en utilisant le module sparse de python et la structure tridiagonale de la matrice du Laplacien, alors on a une complexité linéaire, ce qui est donc mieux. Mais on perd en précision de la résolution quand on augmente N i.e quand on affine le pas de discrétisation, à cause du mauvais conditionnement (en $O(N^2)$). Ici en périodisant en revanche, l'erreur d'approximation est seulement liée à l'erreur d'approximation de notre solution par sa série de Fourier tronquée. L'erreur au contraire diminue donc à mesure que N augmente.

Il existe d'autres méthode d'approximations, par exemple la méthode des éléments finis. Elle uilise des sous-espaces vectoriels de dimension finie de $H_0^1(0,1)$, qui reste donc des espaces de Hilbert. L'idée est de prendre l'unique solution sur chacun de ces sous-espaces et de se dire que plus la dimension augmente, plus on espère se rapprocher de la vraie solution. Cette méthode fait apparaître également la matricie du Laplacien et en fait, dans le cas de l'équation de Laplace, la méthode des éléments finis et celle des différences finies coïncident.

Question 11 : Les équations des ondes 1D et de la chaleur 1D. Quelles sont les propriétés qualitatives de leurs solutions ? Comment utiliser la théorie de Fourier pour les analyser ?

Référence de la réponse :
- L. Dumas, Modélisation pour l'oral de l'agrégation.

43 Introduction aux problèmes

Notre objectif ici est d'étudier les deux problèmes suivant

- L'équation de la chaleur
$$\begin{cases} \dfrac{\partial u}{\partial t} - c\dfrac{\partial^2 u}{\partial x^2} &= 0 \\ u(x,0) &= u_0(x) \end{cases}$$
Cette équation est linéaire et est d'ordre 2. Elle es dépendante du temps, modélisant un phénomène de propagation à vitesse infinie : elle est irréversible. C'est une équation parabolique.

- L'équation des ondes
$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} - c^2\dfrac{\partial^2 u}{\partial^2 x} &= 0 \\ u(x,0) &= u_0(x) \\ \frac{\partial u}{\partial t}(x,0) &= u_1(x) \end{cases}$$
Cette équation est linéaire et est d'ordre 2. Elle fait partie de la même famille que l'équation de transport (Question 9) : une équation hyperbolique, dépendante du temps, modélisant une propagation à vitesse finie : elle est réversible.

On cherche à répondre aux même questions que pour l'équation de transport (Question 9).

44 Résolution des équations

On pourrait considérer plusieurs contextes :

- sur un cercle : $u_0 \in \mathcal{C}^0_{2\pi}(\cap \mathcal{C}^1_{pm})$
- sur une barre : $u_0 \in \mathcal{C}^1([a,b])$
- sur la droite réelle : $u_0 \in \mathcal{S}(\mathbb{R})$

Tout d'abord, sur un cercle : on procède par analyse synthèse :

- pour l'équation de la chaleur :
  - Analyse : on utilise la régularité attendue de la solution pour l'écrire comme sa série de Fourier. Après des ipp, on peut écrire la "solution" en fonction de la donnée initiale et du noyau gaussien, donc si elle existe, elle est unique
  - Synthèse : la fonction trouvée est effectivement solution, d'où l'existence
- pour l'équation des ondes : ça existe ? ?

Ensuite, pour le cas borné : on procède aussi par analyse synthèse :

- pour l'équation de la chaleur
  - Analyse : on procède par séparation des variables. On utilise la régularité de la donnée initiale pour l'écrire comme sa série de Fourier et après calculs, on peut écrire la solution attendue en fonction de la donnée initiale et de fonctions usuelles, donc si elle existe, elle est unique
  - Synthèse : par la théorie des séries de Fourier, on montre que la fonction trouvée est effectivement solution, d'où l'existence
- pour l'équation des ondes
  - Analyse : on procède par séparation des variables. On utilise la régularité des données initiales pour l'écrire comme leur série de Fourier et après calculs, on peut écrire la solution attendue en fonction des données initiales, donc si elle existe, elle est unique
  - Synthèse : par la théorie des séries de Fourier, on montre que la fonction trouvée est effectivement solution, d'où l'existence

Enfin, dans le cas de la droite réelle, on procède un peu différemment :

- pour l'équation de la chaleur :
  - Analyse : on utilise la régularité attendue de la solution pour prendre sa transformée de Fourier. Après des ipp, on peut écrire la "solution" en fonction de la donnée initiale et du noyau gaussien, donc si elle existe, elle est unique
  - Synthèse : la fonction trouvée est effectivement solution, d'où l'existence
- pour l'équation des ondes :
  - on peut procéder de façon identique si les données initiales sont dans la classe de Schwartz.
  - sinon, on fait comme suit. Le changement de variable $(x,t) \to (y,s) = (x-ct, x+ct)$ transforme notre problème en
    $$\frac{\partial^2 u}{\partial y \partial s} = 0$$
    et donc notre solution s'écrit comme somme de deux ondes progressives de vitesses respectives $-c$ et $+c$:
    $$u(x,t) = u_+(x-ct) + u_-(x+ct).$$
    La formule de Duhamel, déjà utilisée de l'équation de transport, nous permet d'affirmer que pour $u_0 \in \mathcal{C}^2$ et $u_1 \in \mathcal{C}^1$ il existe une unique solution de classe $\mathcal{C}^2$ au problème de Cauchy qu'est l'équation des ondes. Cette solutin est donnée par
    $$u(x,t) = \frac{1}{2}\left[u_0(x+ct) + u_0(x-ct)\right] + \frac{1}{2c}\int_{x-ct}^{x+ct} u_1(s)\mathrm{d}s.$$
    Pour obtenir cette solution, on peut poser $v = \partial_t u + c\partial_x u$.

On remarque que dans le tout dernier point, il y a du déjà-vu. En effet, nous avons déjà étudié l'équation du transport, qui fait partie de la même famille d'EDP que l'équation des ondes, donc les idées se croisent.

45 Propriétés qualitatives des solutions

On peut remarquer trois choses.

- Pour l'équation de la chaleur, la solution est au moins autant régulière que la donnée initiale. Par exemple, dans le cas du cercle, à partir d'une donnée initiale continue, on obtient une solution infiniment dérivable. Ou encore, sur la droite, la solution reste dans la classe de Schwartz si la donnée initiale y était.

- Pour l'équation des ondes, la dépendance continue en les données initiale nous donne une stabilité de la solution. Mais contrairement à l'équation de la chaleur, il n'y a pas d'effet régularisant : la solution n'est pas plus (ni moins) régulière que sa donnée initiale.

- De plus, l'équation de la chaleur propage la donnée initiale "à vitesse infinie" : si u_0 est une fonction à support compact, alors $u(t,x)$ solution de l'équation de la chaleur est à support dans $\mathbb{R}$ pour tout $t > 0$. L'équation des ondes, quant à elle, propage la donnée initiale à vitesse finie c : si u_0 est à support dans $[-R,R]$, alors $u(t,\cdot)$ solution de l'équation des ondes est à support dans $[-R - ct,R + ct]$.

Question 12 : Approximation numérique par différences finies. Présenter la démarche et les idées d'analyse pour approcher la solution d'une EDP d'évolution modèle.

Référence de la réponse :
- L. Dumas, Modélisation pour l'oral de l'agrégation.

46 Équations elliptiques : l'équation de Laplace

On veut approcher le problème suivant

$$\left\{ \begin{array}{rcl} -u'' & = & f \\ u(1) & = & 0 \\ u(0) & = & 0 \end{array} \right.$$

avec $f \in \mathcal{C}^2([0,1],\mathbb{R})$. Puisque le problème est indépendant du temps, on peut faire quelque chose de très similaire aux EDO : on utilise les formules de Taylor pour approcher la dérivée seconde par une différence centrée

$$u''(x) \simeq \frac{u(x+\Delta x) - 2u(x) + u(x-\Delta x)}{\Delta x^2}.$$

Ainsi on approche $u(x_j)$ par une valeur u_j calculée par le schéma aux différences finies ci-dessus avec les données initiales

$$u_0 = u_J = 0.$$

On introduit une subdivision régulière à $J+1$ points $(x_j)_{0 \leqslant j \leqslant J}$ espacés de manière régulière dans l'intervalle $I = [0,1]$, c'est à dire : $x_j = j\Delta x$, avec $\Delta x = 1/J$. On obtient le schéma

$$\frac{1}{\Delta x^2} AU = F(X).$$

où A est la matrice du laplacien, U est un vecteur colonne $(u_j)_{1 \leqslant j \leqslant J-1}$ et $F(X)$ est le vecteur des points $f(x_j)_{1 \leqslant j \leqslant J-1}$. Le schéma est

- consistant d'ordre 2,
- stable en norme infinie,
- convergent d'ordre 2.
- puisque A est tridiagonale symétrique définie positive, la résolution du système linéaire se fait en $O(J)$ mais la matrice étant mal conditionnée, les petites perturbations influent grandement sur la solution.

Il existe d'autres méthode d'approximations, par exemple la méthode des éléments finis. Elle uilise des sous-espaces vectoriels de dimension finie de $H_0^1(0,1)$, qui reste donc des espaces de Hilbert. L'idée est de prendre l'unique solution sur chacun de ces sous-espaces et de se dire que plus la dimension augmente, plus on espère se rapprocher de la vraie solution. Cette méthode fait apparaître également la matricie du Laplacien et en fait, dans le cas de l'équation de Laplace, la méthode des éléments finis et celle des différences finies coïncident.

47 Équations hyperboliques : l'équation de transport, des ondes

On veut approcher le problème suivant

$$\left\{\begin{array}{lcl} \dfrac{\partial^2 u}{\partial t^2} - c^2 \dfrac{\partial^2 u}{\partial x^2} & = & 0 \\ u(x,0) & = & g(x) \\ \dfrac{\partial u}{\partial t}(x,0) & = & h(x) \end{array}\right.$$

avec $g \in \mathcal{C}^2([0,1],\mathbb{R})$, $h \in \mathcal{C}^1([0,1],\mathbb{R})$ admettant des prolongements 1-périodiques de classe $\mathcal{C}^2$ et $\mathcal{C}^1$ respectivement. On suppose que la solution u est 1-périodique en x. Ici à nouveau on s'inspire des développements de Taylor pour approcher la solution en les points

$$x_j = j\Delta x, j = 0, \ldots, J+1 \quad \text{avec } \Delta x = \frac{1}{J+1}$$

et aux instants

$$t_n = n\Delta t, n \in \mathbb{N}.$$

Ainsi on approche $u(x_j,t_n)$ par une valeur u_j^n calculée par le schéma aux différences finies suivant.

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial u}{\partial x} \simeq \frac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{\Delta t^2} - c^2 \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2}$$

avec les données initiales

$$u_j^0 = g(x_j),\ u_j^1 = u_j^0 + \Delta t\, h(x_j), \quad \forall 0 \leqslant j \leqslant J$$

et la condition de périodicité

$$u_{J+1}^n = j_0^n, \quad \forall n \in \mathbb{N}^*.$$

C'est un schéma explicite puisque l'on peut exprimer le temps $n+1$ grâce aux approximations au temps n. Le schéma est de plus

- consistant
- d'ordre 2 en temps et en espace
- sous la condition $\Delta x > c\Delta t$, le schéma est convergent.

48 Équations paraboliques : l'équation de la chaleur

On veut approcher le problème suivant

$$\left\{\begin{array}{lcl} \dfrac{\partial u}{\partial t} - c \dfrac{\partial^2 u}{\partial x^2} & = & 0 \\ u(x,0) & = & u_0(x) \end{array}\right.$$

avec $u_0 \in \mathcal{C}_L^0(\mathbb{R},\mathbb{R})$ sur $[0,T] \times [0,L]$. Ici à nouveau on s'inspire des développements de Taylor pour approcher la solution en les points

$$x_j = j\Delta x, j = 0, \ldots, J \quad \text{avec } \Delta x = \frac{1}{J}$$

et aux instants

$$t_n = n\Delta t, n \in \mathbb{N}.$$

Ainsi on approche $u(x_j,t_n)$ par une valeur u_j^n calculée par le schéma aux différences finies suivant.

$$\frac{\partial u}{\partial t} - c\frac{\partial^2 u}{\partial x^2} \simeq \frac{u_j^{n+1} - u_j^n}{\Delta t} - c^2\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2}$$

avec les données initiales

$$u_j^0 = u_0(x_j), \quad \forall 0 \leqslant j \leqslant J$$

et les conditions de périodicité

$$u_0^n = u_{J+1} = 0.$$

C'est un schéma explicite puisque l'on peut exprimer le temps $n+1$ grâce aux approximations au temps n. Le schéma est de plus

- consistant
- d'ordre 1 en temps et 2 en espace
- sous la condition CFL : $\Delta x^2 \geqslant 2c\Delta t$, le schéma est convergent.

On peut améliorer les schémas présentés ici en introduisant les schémas explicites ou encore un mix : les $\theta-$schéma. En effet, les schémas implicites sont inconditionnellement stables (donc convergents) et la conditions CFL étant assez restrictive on préfère en général(malgré le coût de calcul en plus) s'attaquer aux schémas implicites. De plus, le schéma de Crank-Nicholson (moité moité entre explicite et implicite) nous fait gagner un ordre en temps.