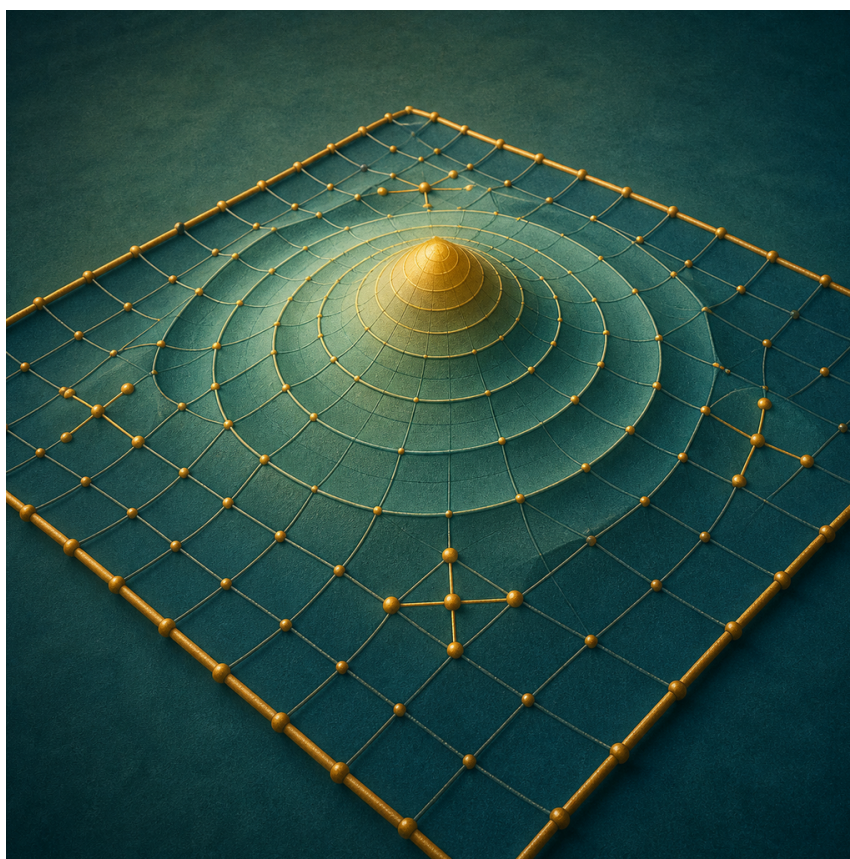

Bases d'analyse numérique

Première partie

XMS1ME090

Master 1 MACS – Nantes Université
Option Mathématiques et applications – École Centrale Nantes



Samuel TRÉTON

Polycopié établi sur la base des notes de cours de [Françoise FOUCHER](#)

Année universitaire 2026–2027

Table des matières

	Présentation du cours	2
1	Introduction : EDP et équations modèles	3
1.1	Généralités	3
1.2	Classification des EDP linéaires d'ordre 2	5
1.3	Équations modèles (linéaires et d'ordre 1 ou 2)	6
2	Équation de la chaleur stationnaire (Poisson)	11
2.1	Remarques sur les solutions de (11)	11
2.2	Schéma aux différences finies pour l'équation de Poisson	14
2.3	Consistance du schéma aux différences finies (21)	18
2.4	Stabilité du schéma aux différences finies (21)	23
2.5	Convergence du schéma aux différences finies (21)	27
3	Équation de la chaleur	28
3.1	Remarques sur les solutions de (54)	28
3.2	Construction de schémas aux différences finies pour l'équation de chaleur	32
3.3	Un premier schéma : Euler explicite	34
3.4	Un deuxième schéma : Euler implicite	39
3.5	Résultats généraux sur la consistance, la stabilité et la convergence	41
4	Équation de transport	56
5	Équation des ondes	56
	Liste de questions de cours	57

Présentation du cours

Ce document présente les bases de l'analyse numérique pour le Master 1 MACS à Nantes Université et l'option Mathématiques et applications à l'École Centrale Nantes. Il a été réalisé sur la base des notes de cours de **Françoise Foucher**, maîtresse de conférences à l'École Centrale Nantes.

Objectifs du cours

À l'issue de cet enseignement, les étudiant.e.s devront être capables de :

- connaître les différentes classes d'équations aux dérivées partielles linéaires et les équations modèles associées ;
- construire un schéma aux différences finies et étudier ses propriétés de stabilité et de convergence ;
- mettre en œuvre une résolution numérique sur ordinateur, notamment lors des travaux pratiques avec Python ;
- analyser les résultats obtenus et quantifier les erreurs numériques.

Plan de l'enseignement

- EDP et équations modèles ;
- équation de la chaleur stationnaire : méthode des différences finies, consistance, stabilité et convergence ;
- équation de la chaleur non stationnaire : schémas explicites et implicites de différences finies ;
- équation de transport : courbes caractéristiques et schémas de différences finies ;
- équation des ondes.

Bibliographie

Les références sont classées par ordre de proximité avec le contenu du cours.

- Brigitte LUCQUIN, *Équations aux dérivées partielles et leurs approximations*, Ellipses, 2004.
- Grégoire ALLAIRE, *Analyse numérique et optimisation*, Ellipses, 2005.
- Daniel EUVRARD, *Résolution numérique des équations aux dérivées partielles*, Masson, 3^e édition, 1994.
- Lionel SAINSAULIEU, *Calcul scientifique*, Dunod, Sciences Sup, 2000.
- Bijan MOHAMMADI et Jacques-Hervé SAIAC, *Pratique de la simulation numérique*, Dunod, 2003.
- Ionut DANAILA, Pascal JOLY, Sidi Mahmoud KABER et Marie POSTEL, *Introduction au calcul scientifique par la pratique*, Dunod, Sciences Sup, 2005.
- Mark H. HOLMES, *Introduction to Numerical Methods in Differential Equations*, Springer, 2007.
- Randall J. LEVEQUE, *Finite Difference Methods for Ordinary and Partial Differential Equations*, SIAM, 2007.
- Eleuterio F. TORO, *Riemann Solvers and Numerical Methods for Fluid Dynamics*, Springer, 3^e édition, 2010.

1 Introduction : EDP et équations modèles

1.1 Généralités

Une Équation aux Dérivées Partielles (EDP) est une relation entre une fonction inconnue, plusieurs variables indépendantes, et ses dérivées partielles. Par exemple, pour une inconnue $u = u(t, x)$ dépendant des variables indépendantes t et x , l'équation de Korteweg–de Vries s'écrit :

$$\partial_t u + 6u\partial_x u + \partial_x^3 u = 0. \quad (1)$$

Définition 1.1 (Ordre d'une EDP). *L'ordre d'une EDP est le plus grand ordre des dérivées partielles apparaissant dans l'équation.*

Par exemple, dans l'équation de Korteweg–de Vries (1), l'ordre est trois, puisque la plus haute dérivée partielle est $\partial_x^3 u$.

Dans ce qui suit, les dérivées partielles pourront être signalées de différentes manières, plus ou moins abrégées :

$$\frac{\partial u}{\partial t} = \partial_t u = u_t, \quad \text{ou encore,} \quad \frac{\partial^3 u}{\partial x^3} = \partial_x^3 u = u_{xxx}.$$

Point de vigilance 1.2. *Si u dépend de plusieurs variables indépendantes, la notation u' , u'' , etc., n'a pas de sens puisqu'elle ne précise pas par rapport à quelle variable la dérivée est prise.*

Une EDP est définie sur un **domaine**, qui est un ensemble de points dans l'espace des variables indépendantes où la fonction inconnue et ses dérivées partielles sont considérées. On le note généralement Ω , par exemple,

$$(t, x) \in \Omega = (0, T) \times (0, L) \subset \mathbb{R}^2.$$

En général, on associera la variable t au temps, et les variables x, y, z , aux coordonnées spatiales. On dira alors que l'EDP est **d'évolution** si son inconnue dépend du temps, et qu'elle est **stationnaire** si elle n'en dépend pas. Pour les équations d'évolution, la notation Ω désigne généralement uniquement le domaine spatial, c'est-à-dire que l'EDP est considérée sur $(0, T) \times \Omega$ avec Ω le domaine spatial—qui est un ouvert de $\mathbb{R}^d$, pas forcément rectangulaire, ni borné.

Comme dans le cas des équations différentielles ordinaires, une EDP nécessite des **conditions aux limites** pour être bien posée. En effet, pour rappel, pour que l'EDO $X' = f(t, X)$ soit bien posée, il est nécessaire de fournir une condition initiale $X(t_0) = X_0$ ¹. Cette nécessité provient essentiellement du fait qu'il faut intégrer l'équation pour obtenir la solution et qu'une condition supplémentaire est nécessaire pour fixer la constante d'intégration qui apparaît lors de cette intégration.

De manière analogue, pour une EDP, il est nécessaire de fournir des conditions initiales et/ou des conditions aux limites appropriées pour que le problème soit bien posé. Le nombre et la nature de ces conditions dépendent de l'EDP considérée, de son ordre et de son type—elliptique, parabolique ou hyperbolique (voir la [Section 1.2](#)).

Dans les domaines de dimension 1 en espace, les conditions aux limites les plus couramment rencontrées sont :

1. Voir le théorème de Cauchy–Lipschitz pour les EDO.

- la **condition initiale** si l'EDP est d'évolution (dépend du temps)

$$u(t=0, x) = u_0(x), \quad \text{et/ou} \quad u_t(t=0, x) = u_1(x), \quad \text{pour tout } x \in (0, L);$$

- la condition de bords de **Dirichlet** (valeur de l'inconnue imposée sur le bord)

$$u(t, x=0) = u_g, \quad \text{et/ou} \quad u(t, x=L) = u_d, \quad \text{pour tout } t \in (0, T); \quad (2)$$

- la condition de bords de **Neumann** (valeur de la dérivée normale sur le bord)

$$\partial_x u(t, x=0) = v_g, \quad \text{et/ou} \quad \partial_x u(t, x=L) = v_d, \quad \text{pour tout } t \in (0, T). \quad (3)$$

On dit que les conditions aux limites de Dirichlet et de Neumann sont **homogènes** si les valeurs imposées pour la solution ou pour sa dérivée normale sont nulles, *i.e.*

$$u_g = u_d = 0, \quad v_g = v_d = 0.$$

La solution d'une EDP peut représenter diverses quantités physiques telles que la température, la pression, la concentration d'une substance chimique, la vitesse d'un fluide, le déplacement d'une structure ou encore une densité de population. Dans ce cours on s'intéressera principalement aux aspects mathématiques et numériques de ces équations et moins à la modélisation qui leur est associée.

Définition 1.3 (EDP linéaire/non-linéaire). Une EDP est dite **linéaire** si elle peut s'écrire sous la forme

$$L(u) = f,$$

où L est une combinaison linéaire de u et de ses dérivées, et f une fonction donnée.

Définition 1.4 (EDP à coefficient constant/variable). Une EDP est dite à **coefficient constant** si les coefficients des dérivées de l'inconnue sont des constantes. Elle est dite à **coefficient variable** si ces coefficients dépendent des variables indépendantes.

Définition 1.5 (EDP homogène/non-homogène). Une EDP est dite **non-homogène** si elle contient des termes indépendants de l'inconnue u . Sinon, elle est dite **homogène**. Dans les cas non-homogènes, on a pour habitude d'appeler **termes source** les termes indépendants de l'inconnue. S'il n'y a qu'un seul terme source, on le note généralement f .

Par exemple (essayez de vous en convaincre),

$$\underbrace{314}_{\text{coef. constant}} u_t(t, x, y) - \underbrace{\cos(x)}_{\text{coef. variable}} u_{xx}(t, x, y) - \underbrace{ty}_{\text{coef. variable}} u(t, x, y) = \underbrace{f(t, x, y)}_{\text{terme source}}$$

est linéaire et homogène si et seulement si $f(t, x, y) \equiv 0$. Un autre exemple :

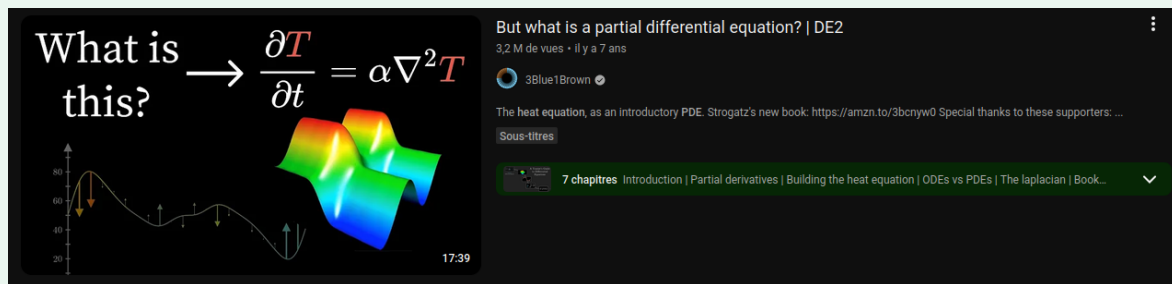
$$u_t = u_{xx}$$

est une EDP linéaire à coefficient constant et homogène, tandis que

$$u_t = (u^2)_{xx} = ((u^2)_x)_x = (2uu_x)_x = 2(u_x)^2 + 2uu_{xx}$$

est une EDP non-linéaire à coefficient constant et homogène.

Conseil 1.6. Allez voir la vidéo de 3blue1brown sur les équations aux dérivées partielles pour une explication visuelle et intuitive : <https://www.youtube.com/watch?v=ly4S0oi3Yz8>



1.2 Classification des EDP linéaires d'ordre 2

Une EDP linéaire d'ordre 2 à deux variables indépendantes t et x sur $\Omega \subset \mathbb{R}^2$, s'écrit toujours sous la forme

$$au_{tt} + 2bu_{tx} + cu_{xx} + du_t + eu_x + ru = f, \quad (4)$$

où a, b, c, d, e, r et f dépendent éventuellement des variables indépendantes t et x .

En utilisant les notations de divergence et de gradient, respectivement,

$$\text{pour } V(t, x) = \begin{pmatrix} V_1(t, x) \\ V_2(t, x) \end{pmatrix} \in \mathbb{R}^2, \quad \text{div}(V) = \frac{\partial V_1}{\partial t} + \frac{\partial V_2}{\partial x},$$

et

$$\text{pour } u(t, x) \in \mathbb{R}, \quad \nabla u = \begin{pmatrix} u_t \\ u_x \end{pmatrix},$$

si a, b et c sont de classe $\mathcal{C}^1$ et si u est de classe $\mathcal{C}^2$, la règle du produit donne

$$\begin{aligned} \text{div} \left(\begin{pmatrix} a & b \\ b & c \end{pmatrix} \nabla u \right) &= \text{div} \begin{pmatrix} au_t + bu_x \\ bu_t + cu_x \end{pmatrix} \\ &= au_{tt} + 2bu_{tx} + cu_{xx} + (\partial_t a + \partial_x b)u_t + (\partial_t b + \partial_x c)u_x. \end{aligned}$$

Ainsi, l'équation linéaire (4) peut prendre la forme divergencielle

$$\text{div}(A \nabla u) + B \cdot \nabla u + ru = f, \quad (5)$$

avec

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}, \quad B = \begin{pmatrix} d - \partial_t a - \partial_x b \\ e - \partial_t b - \partial_x c \end{pmatrix}.$$

En particulier, si a, b et c sont constants, alors $B = (d, e)^T$.

Exemple : pour l'équation

$$u_{tt} + 2u_{tx} + x^2 u_{xx} + u_x + u = 0,$$

on a

$$A = \begin{pmatrix} 1 & 1 \\ 1 & x^2 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1 - 2x \end{pmatrix}.$$

Pour classifier l'équation, on ne retient que sa partie principale

$$au_{tt} + 2bu_{tx} + cu_{xx},$$

à laquelle on associe la forme quadratique

$$q_A(\xi) = \xi^T A \xi, \quad \xi \in \mathbb{R}^2.$$

On associe ainsi l'équation formelle

$$\xi^T A \xi + B \cdot \xi + r = 0$$

qui représente effectivement une conique dans le plan (ξ_1, ξ_2) . Cette analogie permet de classifier les équations linéaires d'ordre 2 en fonction de la nature de la conique associée : ellipse, parabole, hyperbole. Pour chacun de ces trois types, les solutions des EDP associées présentent des comportements caractéristiques.

Plus précisément, la nature de la conique est déterminée par le discriminant du polynôme caractéristique $b^2 - ac$ de la matrice A :

Type d'EDP	Conique	Valeurs propres de A	Critère	Comportement caractéristique
Elliptique	Ellipse	même signe strict	$b^2 - ac < 0$	équilibre spatial, effet régularisant
Parabolique	Parabole	une valeur propre nulle	$b^2 - ac = 0$	diffusion, effet régularisant, évolution irréversible
Hyperbolique	Hyperbole	signes opposés	$b^2 - ac > 0$	propagation à vitesse finie

Remarque 1.7. L'hypothèse $(t, x) \in \mathbb{R}^2$ est éclairante pour une première compréhension de la classification des équations linéaires d'ordre 2 ; cependant, on peut facilement généraliser cette classification à $(t, x) \in \mathbb{R} \times \mathbb{R}^d$ ou $x \in \mathbb{R}^d$. La matrice A devient alors une matrice symétrique $(d+1) \times (d+1)$ ou $d \times d$ selon le cas et on s'intéresse au signe de ses valeurs propres pour déterminer la nature de l'équation.

Pour rappel, le terme “signe des valeurs propres” est bien défini ici puisque la matrice A est symétrique, ce qui garantit (par le théorème spectral) que toutes ses valeurs propres sont réelles ; cela n'aurait pas eu de sens pour une matrice non symétrique dont certaines valeurs propres pourraient être complexes.

À retenir 1.8. La matrice A contient les coefficients des dérivées secondes dans l'équation linéaire d'ordre 2. Pour déterminer la nature de l'équation (elliptique, parabolique ou hyperbolique), il suffit d'examiner le signe de ses valeurs propres :

- spectre de A a un signe strict : elliptique
- spectre de A contenant exactement une valeur propre nulle et les autres de même signe : parabolique
- spectre de A contenant des valeurs propres de signes opposés : hyperbolique

1.3 Équations modèles (linéaires et d'ordre 1 ou 2)

1.3.1 Équation de la chaleur stationnaire

Dans le chapitre 2, on étudiera l'équation de la chaleur stationnaire sur un domaine $\Omega \subset \mathbb{R}^d$:

$$-\Delta u = f \quad \text{dans } \Omega, \tag{6}$$

où $u = u(x)$ et Δ désigne l'opérateur **Laplacien** :

$$\Delta u = \sum_{i=1}^d \frac{\partial^2 u}{\partial x_i^2},$$

c'est-à-dire

$$\begin{aligned} \Delta u &= u_{x_1 x_1} = u'' & \text{si } d = 1, \\ \Delta u &= u_{x_1 x_1} + u_{x_2 x_2} & \text{si } d = 2 \dots \end{aligned}$$

On remarquera que (6) est en fait une simple EDO lorsque $d = 1$. Cette équation est linéaire, elliptique et du second ordre. On l'appelle également **équation de Poisson** dans le cas inhomogène où $f \neq 0$ et **équation de Laplace** lorsque $f = 0$.

La quantité u modélise la température d'un milieu à l'équilibre thermique, sous l'effet de la source de chaleur f .

Lorsqu'on considère (6) sous sa forme divergencielle (5), on trouve

$$A = -I_d, \quad B = 0, \quad r = 0,$$

de sorte que les valeurs propres de A sont toutes égales à -1 , ce qui confirme le caractère elliptique de l'équation.

L'équation de la chaleur stationnaire sera étudiée dans le chapitre 2, la fiche de TD 1, et la fiche de TP 1.

1.3.2 Équation de la chaleur

Dans le chapitre 3, on s'intéressera à l'équation de la chaleur (instationnaire) sur un domaine $\Omega \subset \mathbb{R}^d$:

$$\partial_t u - \sigma \Delta u = f \quad \text{dans } (0, T) \times \Omega, \quad (7)$$

où $u = u(t, x)$, $\sigma > 0$ est le **coefficient de diffusion**, et $T > 0$ est communément appelé **l'horizon temporel**.

Cette équation est linéaire, parabolique et du second ordre. En effet, en dimension $d = 2$ pour rester simple, l'équation (7) prend la forme divergencielle (5), avec

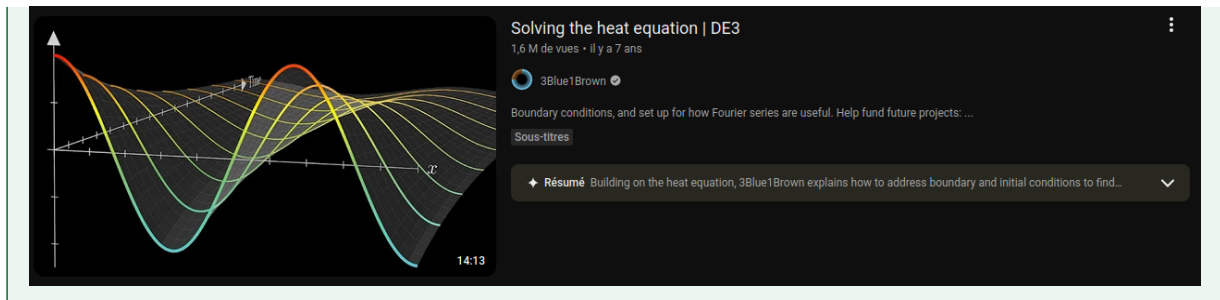
$$A = \begin{pmatrix} -\sigma & 0 & 0 \\ 0 & -\sigma & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad B = 0, \quad r = 0.$$

de sorte que le spectre de la matrice A est donné par $\{-\sigma, -\sigma, 0\}$, ce qui confirme le caractère parabolique de l'équation.

Les solutions de (7) évoluent typiquement vers un état d'équilibre thermique, c'est-à-dire une solution stationnaire u_∞ qui satisfait l'équation de la chaleur stationnaire (6).

En dimension $d = 1$, avec $f = 0$, et conditions de Neumann (c'est-à-dire $\partial_x u = 0$ aux frontières signifiant une isolation thermique du milieu) la solution de l'équation de la chaleur évolue vers une solution constante correspondant à l'équilibre thermique.

Conseil 1.9. Vous pouvez aller visionner la vidéo de 3blue1brown pour plus d'intuition sur l'équation de la chaleur : <https://www.youtube.com/watch?v=ToIXSwZ1pJU>



L'équation de la chaleur sera étudiée dans le [chapitre 3](#), la [fiche de TD 2](#), et la [fiche de TP 2](#).

1.3.3 Équation de transport

Dans le chapitre 4, on s'intéressera à l'équation de transport (ou convection) sur un domaine $\Omega \subset \mathbb{R}^d$:

$$\partial_t u + \mathbf{c} \cdot \nabla u = f \quad \text{dans } (0, T) \times \Omega, \quad (8)$$

où $u = u(t, x)$, $\mathbf{c} \in \mathbb{R}^d$ est le **champ de vitesse**.

Bien qu'elle soit d'ordre 1 et non d'ordre 2, on la classe tout de même dans la catégorie des EDP hyperboliques, en raison de son comportement caractéristique : l'information se propage à vitesse finie le long de courbes appelées caractéristiques. Dans le cas où le champ de vitesse $\mathbf{c}$ est constant, celles-ci sont données par

$$x(t) = x_0 + t\mathbf{c}.$$

L'équation de transport (8) modélise le déplacement d'une quantité u sous l'effet d'un champ de vitesse $\mathbf{c}$, avec éventuellement une source f , par exemple la concentration d'un polluant dans une rivière. En dimension $d = 1$, lorsque $\mathbf{c} = c \in \mathbb{R}$ est constant et que $f = 0$, la solution évolue simplement par translation : $u(t, x) = u_0(x - ct)$, où u_0 est la condition initiale.

Conseil 1.10. Vous pouvez aller visionner la vidéo de Chris Rycroft pour plus de détails sur l'équation de transport : <https://www.youtube.com/watch?v=s09AkBVWOPc>



L'équation de transport sera étudiée dans le [chapitre 4](#), la [fiche de TD 3](#), et la [fiche de TP 3](#).

1.3.4 Équation des ondes

Dans le chapitre 5, on s'intéressera à l'équation des ondes sur un domaine $\Omega \subset \mathbb{R}^d$:

$$\partial_{tt} u - c^2 \Delta u = f \quad \text{dans } (0, T) \times \Omega, \quad (9)$$

où $u = u(t, x)$ et $c > 0$ est la **vitesse de propagation** des ondes. Cette équation d'ordre 2 peut modéliser, par exemple, les vibrations d'une corde ou la propagation des ondes acoustiques, l'inconnue u représentant le déplacement de la corde ou la pression acoustique.

Pour un $\Omega \subset \mathbb{R}^2$, l'équation des ondes (9) peut se mettre sous la forme divergencielle (5), avec

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -c^2 & 0 \\ 0 & 0 & -c^2 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad r = 0.$$

Les valeurs propres de A sont 1, $-c^2$ et $-c^2$, ce qui rend l'équation des ondes hyperbolique.

En dimension $d = 1$, avec $f = 0$ et des conditions de Dirichlet homogènes ($u = 0$ aux frontières), une perturbation initiale se propage à vitesse finie dans le domaine et se réfléchit lorsqu'elle atteint les frontières. Contrairement à l'équation de la chaleur, il n'y a pas de phénomène de diffusion : les oscillations persistent au cours du temps.

Conseil 1.11. Vous pouvez aller visionner la vidéo de Parth G. pour plus d'intuition sur l'équation des ondes : <https://www.youtube.com/watch?v=ub7lok-JQJE>



L'équation des ondes sera étudiée dans le [chapitre 5](#), la [fiche de TD 3](#), et la [fiche de TP 4](#).

À retenir 1.12. Les équations de Poisson, chaleur, transport et ondes sont les prototypes d'EDP linéaires elliptique, parabolique, hyperbolique et hyperbolique étudiés dans ce cours, respectivement.

1.3.5 Un dernier exemple (non-linéaire) : advection-réaction-diffusion

Dans ce dernier exemple on se propose de regarder l'équation

$$\partial_t u + \mathbf{c} \cdot \nabla u - \sigma \Delta u = f(t, x, u), \quad (10)$$

sur un domaine $\Omega \subset \mathbb{R}^d$ avec des conditions initiales et aux frontières appropriées. Cette équation combine plusieurs phénomènes :

- advection : transport de la quantité u par le champ de vitesse $\mathbf{c}$,
- diffusion : dissipation de u par le terme $-\sigma \Delta u$,
- réaction : croissance ou décroissance locale de u par la fonction $f(t, x, u)$.

Notez que ce dernier terme dans le membre de droite n'est plus une simple source $f(t, x)$ comme c'était le cas pour les équations linéaires précédentes, mais dépend maintenant de la solution u elle-même. Ceci rend l'équation non-linéaire et complique son analyse mathématique ainsi que sa résolution numérique. Ce terme est souvent appelé terme de **réaction** et est analogue à la fonction de croissance F pour l'EDO $X'(t) = F(t, X)$.

Cette équation peut modéliser une population végétale se dispersant dans son environnement. Par exemple, $u(t, x)$ pourrait représenter la densité de cette population sur une île, délimitée

par le domaine Ω , à l'instant t et en position x . La population serait alors soumise à une dispersion diffusive non dirigée, modélisée par $-\sigma\Delta u$, résultant de phénomènes déplacement individuels aléatoires ; à un transport dirigé, modélisé par $\mathbf{c} \cdot \nabla u$, pouvant par exemple représenter la dispersion des graines sous l'effet du vent ; et à une dynamique locale de croissance ou de décroissance, modélisée par $f(t, x, u)$, représentant, par exemple, la reproduction, la mortalité, la consommation par la faune, ou encore compétition intraspécifique. Des conditions aux limites appropriées pourraient enfin traduire la mort des individus à la frontière de l'île, dans le cas de conditions de Dirichlet homogènes, ou son confinement à l'intérieur de l'île, dans le cas de conditions de Neumann homogènes.

L'équation (10) est parabolique si le coefficient de diffusion σ est strictement positif, hyperbolique si $\sigma = 0$. Ce point est laissé en exercice.

2 Équation de la chaleur stationnaire (Poisson)

On étudie dans ce chapitre l'équation de la chaleur stationnaire en dimension $d = 1$, également connue sous le nom d'équation de Poisson avec des conditions aux limites de Dirichlet :

$$\begin{cases} -u'' = f & \text{sur } (0, 1), \\ u(0) = u_g, \\ u(1) = u_d, \end{cases} \quad (11)$$

Remarque 2.1. Voir le *fiche de TD 1* pour des conditions aux limites mixtes :

$$\begin{cases} u(0) = u_g, \\ u'(1) = v_d, \end{cases}$$

et pour le problème lorsque le domaine Ω est un carré de dimension $d = 2$:

$$\begin{cases} -\Delta u = f & \text{dans } \Omega = (0, 1) \times (0, 1), \\ u = 0 & \text{sur } \partial\Omega. \end{cases}$$

2.1 Remarques sur les solutions de (11)

Proposition 2.2 (Caractère bien posé pour l'équation de Poisson avec source régulière). *Si f est continue par morceaux sur $[0, 1]$, alors l'équation (11) admet une unique solution u qui est de classe C^1 sur $[0, 1]$ et de classe C^2 sur chaque sous-intervalle où f est continue.*

Par ailleurs, on a l'expression explicite

$$u(x) = u_g + (u_d - u_g)x + x \int_{y=0}^1 (1-y)f(y) dy - \int_{y=0}^x (x-y)f(y) dy. \quad (12)$$

Preuve de la Proposition 2.2. Pour l'existence, on pourrait simplement vérifier que la fonction u définie par (12) satisfait bien l'équation (11). Par souci de constructivité, regardons plutôt comment la construire à partir de l'équation. Pour cela, on intègre deux fois l'équation $-u'' = f$, en notant C et D les constantes d'intégration qui seront déterminées par les deux conditions aux limites.

À partir de $u'' = -f$, on obtient une première fois

$$u'(x) = - \underbrace{\int_{y=0}^x f(y) dy}_{=: F(x)} + C,$$

puis une seconde fois

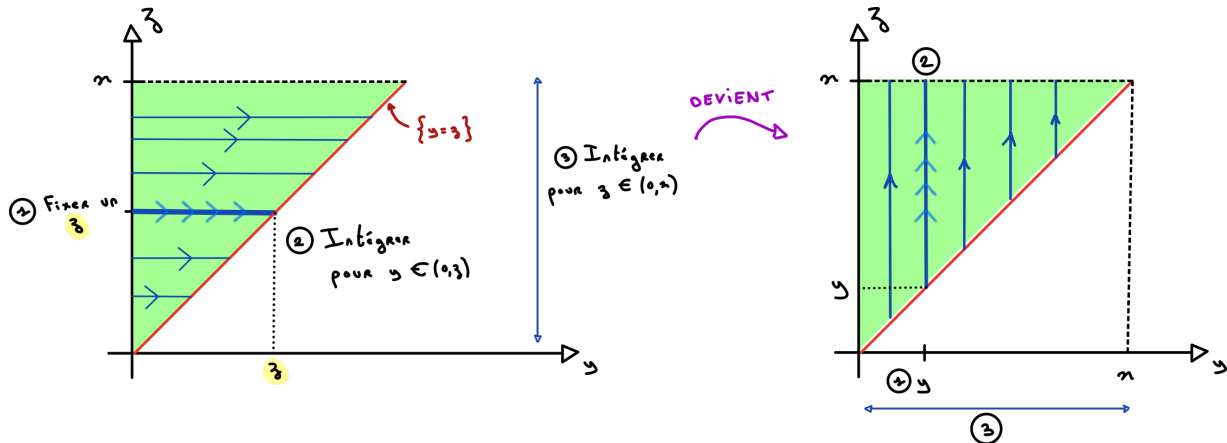
$$\begin{aligned} u(x) &= - \int_{z=0}^x F(z) dz + Cx + D \\ &= - \int_{z=0}^x \left(\int_{y=0}^z f(y) dy \right) dz + Cx + D. \end{aligned} \quad (13)$$

Puisque f est continue par morceaux sur $[0, 1]$, le théorème fondamental de l'analyse donne

$$\left[f \in C_{\text{PW}}^0([0, 1]) \right] \Rightarrow \left[F \in C^0 \cap C_{\text{PW}}^1([0, 1]) \right]$$

$$\Rightarrow \left[x \mapsto \int_{z=0}^x F(z) dz \in C^1 \cap C_{PW}^2([0, 1]) \right].$$

En particulier, l'intégrale double apparaissant dans (13) est bien définie ce qui permet d'utiliser le théorème de Fubini :



si bien qu'on trouve,

$$\begin{aligned} \int_{z=0}^x \int_{y=0}^z f(y) dy dz &= \int_{y=0}^x \int_{z=y}^x f(y) dz dy \\ &= \int_{y=0}^x f(y) \int_{z=y}^x dz dy \\ &= \int_{y=0}^x f(y)(x - y) dy. \end{aligned}$$

Ainsi, (13) devient

$$u(x) = - \int_{y=0}^x f(y)(x - y) dy + Cx + D. \quad (14)$$

La condition $u(0) = u_g$ donne immédiatement

$$D = u_g.$$

La condition $u(1) = u_d$ donne ensuite

$$u_d = - \int_{y=0}^1 (1 - y)f(y) dy + C + u_g,$$

d'où

$$C = u_d - u_g + \int_{y=0}^1 (1 - y)f(y) dy.$$

En reportant les valeurs de C et D dans l'expression de u (14), on obtient

$$u(x) = u_g + (u_d - u_g)x + x \int_{y=0}^1 (1 - y)f(y) dy - \int_{y=0}^x (x - y)f(y) dy,$$

ce qui est bien (12). La régularité énoncée découle également de la construction précédente.

Passons à l'unicité : supposons que u_1 et u_2 soient deux solutions de (11) avec les mêmes conditions aux limites. Alors $w = u_1 - u_2$ satisfait

$$\begin{cases} w''(x) = 0, & x \in (0, 1), \\ w(0) = 0, \\ w(1) = 0. \end{cases}$$

L'intégration de l'équation différentielle donne $w(x) = Ax + B$. Les conditions aux limites imposent $B = 0$ et $A = 0$, donc $w \equiv 0$. Ainsi, $u_1 = u_2$, ce qui prouve l'unicité. $\square$

Remarque 2.3 (Sur l'équation de Laplace). *Dans le cas sans terme source, c'est-à-dire lorsque $f \equiv 0$, on retrouve l'équation de Laplace*

$$-u'' = 0.$$

La solution de (11) est alors simplement la fonction affine reliant les deux valeurs imposées au bord :

$$u(x) = u_g + (u_d - u_g)x.$$

En dimension supérieure, les solutions de l'équation de Laplace ne sont plus, en général, des fonctions affines. Elles dépendent à la fois des données imposées sur le bord et de la géométrie du domaine.

À retenir 2.4.

En dimension 1 : $-\Delta u = -u'' = 0 \implies u$ affine.

En dimension $d = 2$, le problème de Laplace $-\Delta u = 0$ possède des solutions à variables séparées, c'est-à-dire de la forme $u(x, y) = v(x)w(y)$, mais il existe également des solutions plus générales qui ne se séparent pas de cette manière. Ce point est abordé en détail dans la [fiche de TD 1](#).

Pour des termes sources moins réguliers que continus par morceaux, il est nécessaire de recourir à des solutions faibles et à des espaces de Sobolev pour formuler correctement le problème et garantir l'existence et l'unicité de la solution :

Proposition 2.5 (Caractère bien posé pour l'équation de Poisson avec source dans L^2). *Si $f \in L^2(0, 1)$, alors le problème de Poisson avec conditions de Dirichlet homogènes, à savoir*

$$\begin{cases} -u'' = f & \text{sur } (0, 1), \\ u(0) = 0, \\ u(1) = 0, \end{cases} \quad (15)$$

admet une unique solution faible $u \in H_0^1(0, 1)$, où

$$H_0^1(0, 1) := \{v \in L^2(0, 1) \text{ tels que } v' \in L^2(0, 1) \text{ et } v(0) = v(1) = 0\}. \quad (16)$$

Point de vigilance 2.6. *La proposition précédente est énoncée de manière très formelle. Nous n'entrerons pas dans les détails ici, mais il est important de noter que la notion de **solution faible** reste à définir rigoureusement, et que l'élément v' dans la définition de l'espace de Sobolev $H_0^1(0, 1)$ dans (16) doit être compris au sens des dérivées faibles—qui est une généralisation de la dérivée classique aux fonctions qui ne sont pas nécessairement différentiables au sens classique.*

À retenir 2.7. *En général, on ne sait pas calculer la solution du problème de Poisson de manière explicite lorsque la source f est dans L^2 . C'est pourquoi il est nécessaire de recourir à des méthodes numériques d'approximation (comme la **méthode des différences finies**, des **éléments finis**, ou les **volumes finis**, etc.) pour approcher la solution. Dans la première*

partie de ce cours, nous nous concentrerons sur la **méthode des différences finies** pour résoudre numériquement les EDP présentées dans l'introduction.

2.2 Schéma aux différences finies pour l'équation de Poisson

On considère donc l'équation de Poisson sur l'intervalle $(0, 1)$ munie de conditions de Dirichlet non-homogènes (11) qu'on rappelle ici :

$$\begin{cases} -u'' = f & \text{sur } (0, 1), \\ u(0) = u_g, \\ u(1) = u_d. \end{cases} \quad (17)$$

Pour résoudre ce problème numériquement, on introduit pour $N \in \mathbb{N}^*$ un vecteur inconnu

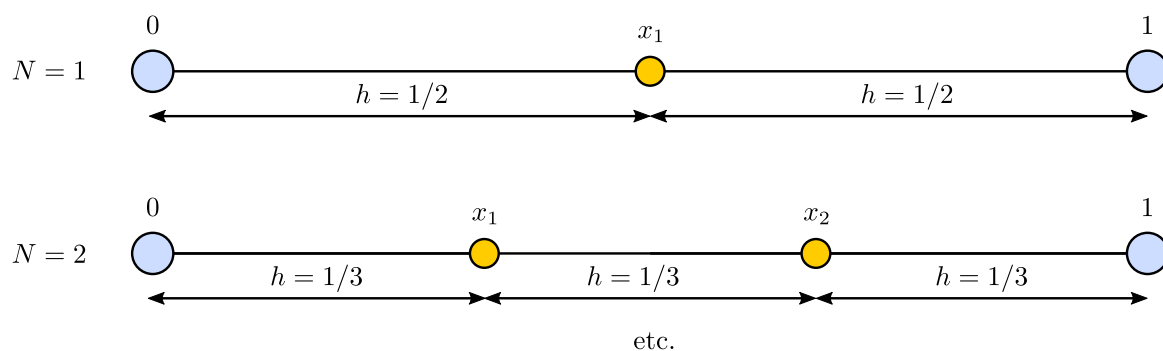
$$u_h = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{pmatrix} \in \mathbb{R}^N \quad (18)$$

dont le rôle sera d'approximer la solution u . Pour cela, on introduit un maillage uniforme $\{x_i\}_{i=1}^N$ de l'intervalle $(0, 1)$, avec

$$x_i := ih, \quad i = \llbracket 1, N \rrbracket,$$

où h désigne le **pas de discrétisation** du schéma

$$h := \frac{1}{N+1}.$$



Remarque 2.8. On peut poser $u_0 = u_g$ et $u_{N+1} = u_d$ puisqu'elles correspondent aux conditions de Dirichlet imposées aux bornes de l'intervalle ; ce ne sont pas des inconnues.

Notre objectif est de construire une méthode permettant d'approximer

$$u(x_i) \approx u_i, \quad i = \llbracket 1, N \rrbracket,$$

et de quantifier la qualité de cette approximation.

Pour cela, on va utiliser des développements de Taylor autour des points du maillage pour approximer la dérivée seconde $u''(x_i)$ en fonction des valeurs u_i et des points voisins.

Point de vigilance 2.9. On a maintenant deux objets distincts : la fonction continue $u(x)$ que l'on cherche à approximer, et le vecteur discret u_h qui contient les valeurs approchées aux points du maillage. Il est important de garder cette distinction à l'esprit lors de l'analyse des

schémas numériques.

En supposant que la fonction u est suffisamment régulière pour permettre un développement de Taylor à l'ordre 2 autour du point x_i , on obtient, à gauche

$$u(x_{i-1}) = u(x_i) - hu'(x_i) + \frac{h^2}{2}u''(x_i) + o(h^2), \quad (19)$$

et, à droite,

$$u(x_{i+1}) = u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + o(h^2). \quad (20)$$

En sommant (19) et (20), on obtient

$$u(x_{i-1}) + u(x_{i+1}) = 2u(x_i) + h^2u''(x_i) + o(h^2),$$

ce qui conduit à l'approximation centrée de la dérivée seconde :

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

Ceci nous pousse à proposer le problème discret suivant sur u_h pour approximer la solution u de $-u'' = f$:

$$-\frac{1}{h^2}(u_{i-1} - 2u_i + u_{i+1}) = f(x_i), \quad i = \llbracket 1, N \rrbracket, \quad (21)$$

où $u_0 = u_g$ et $u_{N+1} = u_d$ correspondent aux conditions de Dirichlet aux bornes de l'intervalle.

Les équations de (21) (il y en a N) forment un grand système d'équations, qui est linéaire en les inconnues $u_1, \dots, u_N$, si bien que l'on peut écrire ce système sous la forme matricielle

$$A_h u_h = b_h, \quad (22)$$

où A_h est la matrice carrée et tridiagonale de taille $N \times N$, et b_h le vecteur de dimension N , donnés par

$$A_h = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ 0 & 0 & -1 & 2 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & -1 \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{pmatrix} \quad \text{et} \quad b_h = \begin{pmatrix} f(x_1) + \frac{u_g}{h^2} \\ f(x_2) \\ f(x_3) \\ \vdots \\ f(x_{N-1}) \\ f(x_N) + \frac{u_d}{h^2} \end{pmatrix}. \quad (23)$$

Note pour plus tard : La matrice A_h est la matrice du **moins Laplacien** : $A_h \approx -\Delta$.

Conseil 2.10. *Soyez convaincu que la matrice A_h s'écrit bien sous la forme tridiagonale donnée ci-dessus. La compréhension de ce point est essentielle pour bien appréhender la structure du système linéaire obtenu.*

On l'obtient ligne par ligne en s'appuyant sur le schéma aux différences finies (21).

On commence par les lignes intérieures (i.e. les lignes correspondant à $i \in \llbracket 2, N-1 \rrbracket$) qui sont les plus simples.

On traite ensuite les lignes correspondant aux indices $i = 1$ et $i = N$, qui prennent en compte les conditions aux bords du domaine, à savoir $u_0 = u_g$ et $u_{N+1} = u_d$. Pour ces deux lignes, il faut écrire explicitement (21) en remplaçant i par 1 puis par N . Vous verrez que les termes $u_0 = u_g$ et $u_{N+1} = u_d$ apparaissent dans le membre de gauche et sont transférés dans le

membre de droite, ce qui explique la présence des termes $\frac{u_g}{h^2}$ et $\frac{u_d}{h^2}$ dans le vecteur b_h dans (23).

Un premier point est de savoir si le problème linéaire (22) est bien posé.

Proposition 2.11 (Caractère bien posé du schéma aux différences finies (22)). *Pour tout $N \geq 1$, la matrice A_h définie en (23) est définie positive, donc inversible. En conséquence, le problème linéaire (22) admet une unique solution.*

Preuve de la Proposition 2.11. Pour rappel, une matrice symétrique M est définie positive si, pour tout vecteur non nul x , on a $x^T M x > 0$. Les matrices symétriques définies positives se caractérisent également par le fait que toutes leurs valeurs propres sont réelles (c'est une conséquence du théorème spectral) et strictement positives. En effet, si λ est une valeur propre de M associée à un vecteur propre non nul v , on a

$$Mv = \lambda v,$$

et donc

$$0 < v^T M v = \lambda v^T v = \lambda \|v\|^2,$$

ce qui implique que $\lambda > 0$.

Ainsi, si M est une matrice symétrique définie positive, toutes ses valeurs propres sont strictement positives. Cela implique que M est inversible. En effet, par contraposée, si M n'était pas inversible, alors il existerait un vecteur non nul v dans $\text{Ker}(M) \setminus \{0\} \neq \emptyset$, ce qui impliquerait que $Mv = 0 = \lambda v$ avec $\lambda = 0$, et nous dirait donc que 0 est une valeur propre de M . Ainsi

$$[M \text{ non inversible}] \implies [0 \text{ est une valeur propre de } M]$$

d'où, par contraposée,

$$[0 \text{ n'est pas une valeur propre de } M] \implies [M \text{ inversible}].$$

Reste enfin à montrer que A_h est définie positive. Pour tout vecteur non nul $x \in \mathbb{R}^N$, on a

$$A_h x = \frac{1}{h^2} \begin{pmatrix} 2x_1 - x_2 \\ -x_1 + 2x_2 - x_3 \\ \vdots \\ -x_{N-2} + 2x_{N-1} - x_N \\ -x_{N-1} + 2x_N \end{pmatrix},$$

d'où

$$\begin{aligned} x^T A_h x &= \frac{1}{h^2} \begin{pmatrix} x_1 & x_2 & \cdots & x_{N-1} & x_N \end{pmatrix} \begin{pmatrix} 2x_1 - x_2 \\ -x_1 + 2x_2 - x_3 \\ \vdots \\ -x_{N-2} + 2x_{N-1} - x_N \\ -x_{N-1} + 2x_N \end{pmatrix} \\ &= \frac{1}{h^2} \left(\underbrace{\sum_{i=2}^{N-1} (2x_i^2 - x_{i-1}x_i - x_i x_{i+1})}_{\text{cas } i \in [2, N-1]} + \underbrace{2x_1^2 - x_1 x_2}_{\text{cas } i=1} + \underbrace{2x_N^2 - x_{N-1} x_N}_{\text{cas } i=N} \right) \end{aligned}$$

$$\begin{aligned}
x^T A_h x &= \frac{1}{h^2} \left(x_1^2 + \overbrace{x_1^2 - 2x_1x_2 + x_2^2}^{=(x_1-x_2)^2} + \overbrace{x_2^2 - 2x_2x_3 + x_3^2}^{=(x_2-x_3)^2} + \cdots + \overbrace{x_{N-1}^2 - 2x_{N-1}x_N + x_N^2}^{=(x_{N-1}-x_N)^2} + x_N^2 \right) \\
&= \frac{1}{h^2} \left(x_1^2 + \sum_{i=1}^{N-1} (x_i - x_{i+1})^2 + x_N^2 \right).
\end{aligned}$$

Ainsi, on a bien

$$x^T A_h x \geq 0 \quad \text{pour tout } x \in \mathbb{R}^N \quad (24)$$

Par ailleurs, si $x^T A_h x = 0$, alors on doit avoir

$$x_1 = x_N = 0 \quad \text{et} \quad x_i - x_{i+1} = 0 \quad \text{pour tout } i \in \llbracket 1, N-1 \rrbracket,$$

ce qui implique par une récurrence immédiate que

$$x_i = 0 \quad \text{pour tout } i \in \llbracket 1, N \rrbracket.$$

Par conséquent, on a

$$[x^T A_h x = 0] \implies [x = 0], \quad (25)$$

et donc en combinant (24) et la contraposée de (25), on obtient

$$[x^T A_h x > 0] \quad \text{pour tout } x \neq 0.$$

□

À présent qu'on a montré que notre schéma aux différences finies mène à une unique solution discrète, on peut se poser les questions qui suivent :

- **(Convergence)** Est-ce que la solution discrète se rapproche de la solution continue lorsque le pas de discrétisation h tend vers 0 ? Autrement dit, a-t-on

$$\lim_{h \rightarrow 0} \left\| \begin{pmatrix} u(x_1) - u_1 \\ u(x_2) - u_2 \\ \vdots \\ u(x_N) - u_N \end{pmatrix} \right\| = 0,$$

pour une norme $\|\cdot\|$ appropriée ?

- **(Stabilité)** Comment varie la solution discrète lorsque les données f , u_g ou u_d sont légèrement perturbées ? En particulier, de petites perturbations des données entraînent-elles seulement de petites perturbations de la solution ?
- **(Consistance)** Si l'on remplace dans le schéma discret les inconnues u_i par les valeurs $u(x_i)$ de la solution continue aux points de discrétisation, l'équation discrète est-elle approximativement satisfaite ? Et cette approximation devient-elle de plus en plus précise lorsque h tend vers 0 ?

À retenir 2.12. On verra par la suite que

$$\boxed{\text{“stabilité + consistance} \iff \text{convergence”}}$$

Point de vigilance 2.13. *Il est important de noter que les limites considérées lorsque $h \rightarrow 0$ dépendent de la norme choisie pour mesurer les erreurs. En effet, pour chaque $h > 0$ fixé, les objets considérés (par exemple l'erreur de consistance ou l'erreur de convergence qu'on définira plus bas) sont bien des vecteurs d'un espace de dimension finie $\mathbb{R}^N$, où $N = N(h)$. Cependant, lorsque le pas de discrétisation h varie, le nombre de points de discrétisation $N(h)$ varie également, et tend généralement vers l'infini lorsque $h \rightarrow 0$.*

Un exemple simple permet d'illustrer ce phénomène. Soit

$$N_1 := \|\cdot\|$$

une norme sur $\mathbb{R}^N$. Pour tout $h > 0$, l'application

$$N_h := \frac{1}{h} \|\cdot\|$$

définit également une norme sur $\mathbb{R}^N$, équivalente à N_1 pour chaque h fixé. Cependant, la constante d'équivalence dépend ici de h et diverge lorsque $h \rightarrow 0$. Ainsi, si une famille $(u_h)_h$ vérifie

$$\|u_h\| = h \rightarrow 0 \quad \text{lorsque } h \rightarrow 0,$$

alors

$$N_h(u_h) = \frac{1}{h} \|u_h\| = 1,$$

et donc

$$N_h(u_h) \not\rightarrow 0.$$

Cet exemple montre que l'équivalence des normes pour chaque h fixé ne permet pas, à elle seule, de conclure que les notions de convergence lorsque $h \rightarrow 0$ sont équivalentes : il faut également contrôler uniformément en h les constantes d'équivalence.

On ne travaille donc pas dans un espace vectoriel de dimension finie fixé. En particulier, on ne peut pas invoquer directement l'équivalence des normes en dimension finie pour conclure que le choix de la norme est indifférent : les constantes intervenant dans les inégalités d'équivalence peuvent dépendre de $N = N(h)$, et donc de h , et peuvent même dégénérer lorsque $h \rightarrow 0$.

Ainsi, le choix de la norme dans laquelle on mesure l'erreur joue un rôle essentiel dans l'étude de la convergence lorsque $h \rightarrow 0$.

2.3 Consistance du schéma aux différences finies (21)

Définition 2.14 (Erreur de consistance). *L'erreur de consistance mesure à quel point le schéma discret s'écarte de l'équation continue lorsqu'on remplace les inconnues discrètes par les valeurs exactes de la solution continue aux points de discrétisation.*

Autrement dit, si u est la solution exacte de l'équation continue et $\pi_h : C(\Omega) \rightarrow \mathbb{R}^N$ est l'opérateur de projection défini par

$$\pi_h(u) = \begin{pmatrix} u(x_1) \\ u(x_2) \\ \vdots \\ u(x_N) \end{pmatrix} \in \mathbb{R}^N, \quad (26)$$

l'erreur de consistance $\varepsilon_h(u)$ est définie par

$$\varepsilon_h(u) := A_h \pi_h(u) - b_h.$$

qui est un vecteur de $\mathbb{R}^N$.

À retenir 2.15. Pour l'erreur de consistance :

1. On prend u la solution **exacte de l'équation continue**. Attention à ne pas confondre u avec u_h ...
2. Notre schéma discret s'écrit $A_h u_h - b_h = 0$.
3. Pour savoir à quel point notre schéma est fidèle à l'équation continue on met u (la solution exacte) dans le schéma discret : on calcule $A_h \pi_h(u) - b_h =: \varepsilon_h(u)$. (π_h sert à transformer u qui est une fonction continue en un vecteur de valeurs discrètes pour pouvoir l'insérer dans le schéma discret.)
4. Forcément, $\pi_h(u)$ n'est a priori pas la solution exacte du schéma discret, donc on a très peu de chance que $\varepsilon_h(u) = A_h \pi_h(u) - b_h$ soit nul. En revanche, on peut espérer que $\varepsilon_h(u)$ soit petit lorsque le pas de discrétisation h tend vers zéro.

Définition 2.16 (Consistance d'un schéma discret). Pour une norme $\|\cdot\|$ donnée, on dit que le schéma discret $A_h u_h - b_h = 0$ est **consistant** avec l'équation continue si, pour toute solution exacte u de l'équation continue, l'erreur de consistance $\varepsilon_h(u) = A_h \pi_h(u) - b_h$ tend vers zéro lorsque le pas de discrétisation h tend vers zéro :

$$[\text{schéma discret consistant}] \stackrel{\text{déf}}{\iff} [\text{pour toute solution exacte } u, \lim_{h \rightarrow 0} \|\varepsilon_h(u)\| = 0].$$

Pour être plus précis, on peut dire que le schéma discret est **consistant d'ordre p pour une norme $\|\cdot\|$ donnée** si pour toute solution exacte u de l'équation continue, l'erreur de consistance satisfait

$$\exists C > 0, \exists h_0 > 0 \quad \text{tels que} \quad \forall h \leq h_0, \quad \|\varepsilon_h(u)\| \leq Ch^p.$$

Autrement dit,

$$\|\varepsilon_h(u)\| = \mathcal{O}(h^p) \quad \text{lorsque } h \rightarrow 0.$$

Point de vigilance 2.17. L'ordre de consistance peut dépendre de la norme choisie. Par exemple, pour les normes ℓ^1 et ℓ^∞ , on a pour $x \in \mathbb{R}^N$:

$$\|x\|_1 = \sum_{i=1}^N |x_i| \leq N \max_i |x_i| = N \|x\|_\infty \leq (N+1) \|x\|_\infty = \frac{1}{h} \|x\|_\infty.$$

Ainsi, si

$$\|x\|_\infty \leq Ch^p,$$

on a

$$\|x\|_1 \leq \frac{1}{h} \|x\|_\infty \leq \frac{C}{h} h^p = Ch^{p-1}.$$

En conclusion, si un schéma est consistant d'ordre p pour la norme ℓ^∞ , il sera consistant d'ordre $p - 1$ pour la norme ℓ^1 .

Rappel 2.18. Les normes vectorielles importantes dans $\mathbb{R}^N$:

- la norme ℓ^1 : $\|x\|_1 = \sum_{i=1}^N |x_i|$;
- la norme ℓ^2 : $\|x\|_2 = \left(\sum_{i=1}^N |x_i|^2 \right)^{1/2}$;
- la norme ℓ^∞ : $\|x\|_\infty = \max_i |x_i|$.

Proposition 2.19 (Consistance du schéma aux différences finies (21)). *Si la solution exacte u de l'équation de Dirichlet (11) est de classe $\mathcal{C}^4$ sur $[0, 1]$, alors le schéma aux différences finies (21) est consistant d'ordre 2 pour la norme $\|\cdot\|_\infty$.*

Preuve de la Proposition 2.19. On a

$$\varepsilon_h(u) = A_h \pi_h(u) - b_h = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & -1 & 2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} u(x_1) \\ u(x_2) \\ u(x_3) \\ \vdots \\ u(x_N) \end{pmatrix} - \begin{pmatrix} f(x_1) + \frac{u_g}{h^2} \\ f(x_2) \\ f(x_3) \\ \vdots \\ f(x_N) + \frac{u_d}{h^2} \end{pmatrix}.$$

Puisque u vérifie les conditions de Dirichlet aux points $x_0 = 0$ et $x_{N+1} = 1$, on a

$$u(x_0) = u_g, \quad u(x_{N+1}) = u_d.$$

Ainsi, on peut écrire, pour toute composante de l'erreur de consistance,

$$\varepsilon_h(u)_i = -\frac{u(x_{i-1}) - 2u(x_i) + u(x_{i+1}))}{h^2} - f(x_i), \quad \text{pour tout } i \in \llbracket 1, N \rrbracket. \quad (27)$$

En utilisant les expansions de Taylor Lagrange de $u(x_{i-1})$ et $u(x_{i+1})$ autour de x_i à l'ordre 4 (licite, puisque u est de classe $\mathcal{C}^4$), on obtient

$$u(x_{i-1}) = u(x_i) - hu'(x_i) + \frac{h^2}{2}u''(x_i) - \frac{h^3}{6}u^{(3)}(x_i) + \frac{h^4}{24}u^{(4)}(\alpha_{i-1}), \quad \text{où } \alpha_{i-1} \in (x_{i-1}, x_i).$$

$$u(x_{i+1}) = u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{6}u^{(3)}(x_i) + \frac{h^4}{24}u^{(4)}(\beta_{i+1}), \quad \text{où } \beta_{i+1} \in (x_i, x_{i+1}).$$

En sommant ces deux expansions, il vient

$$u(x_{i-1}) - 2u(x_i) + u(x_{i+1}) = h^2u''(x_i) + \frac{h^4}{24} \left(u^{(4)}(\alpha_{i-1}) + u^{(4)}(\beta_{i+1}) \right). \quad (28)$$

On a donc, en injectant (28) dans (27),

$$\varepsilon_h(u)_i = -\left(u''(x_i) + \frac{h^2}{24} \left(u^{(4)}(\alpha_{i-1}) + u^{(4)}(\beta_{i+1}) \right) \right) - f(x_i).$$

À ce moment de l'analyse, il faut s'appuyer sur le fait que u satisfait l'équation de Poisson (11), en particulier $-u''(x_i) = f(x_i)$, ce qui permet de simplifier l'expression de l'erreur de consistance :

$$\varepsilon_h(u)_i = -\frac{h^2}{24} \left(u^{(4)}(\alpha_{i-1}) + u^{(4)}(\beta_{i+1}) \right).$$

À présent, on majore le module de $\varepsilon_h(u)_i$ par inégalité triangulaire :

$$|\varepsilon_h(u)_i| \leq \frac{h^2}{24} \left(|u^{(4)}(\alpha_{i-1})| + |u^{(4)}(\beta_{i+1})| \right) \leq \frac{h^2}{12} \sup_{x \in [0,1]} |u^{(4)}(x)|. \quad (29)$$

Notons que la quantité $\sup |u^{(4)}(x)|$ est finie puisque u est de classe C^4 (donc $u^{(4)}$ est continue) sur le compact $[0, 1]$.

En prenant alors le maximum sur tous les indices $i \in \llbracket 1, N \rrbracket$ dans (29), on obtient

$$\|\varepsilon_h(u)\|_\infty \leq \frac{h^2}{12} \sup_{x \in [0,1]} |u^{(4)}(x)|. \quad (30)$$

On a ainsi montré qu'il existe une constante $C = \frac{1}{12} \sup_{x \in [0,1]} |u^{(4)}(x)|$ (indépendante de h) telle que

$$\|\varepsilon_h(u)\|_\infty \leq Ch^2.$$

Le schéma est donc consistant d'ordre 2 pour la norme infinie. $\square$

Remarque 2.20. Si la solution exacte $u^{(4)} \equiv 0$ (c'est-à-dire² si $f'' \equiv 0$, i.e. f est affine), alors on peut se rendre compte en regardant (29) que l'erreur de consistance est nulle.

Dans ce cas, le schéma est exactement consistant : $A_h \pi_h(u) = b_h$ et on a, dans ce cas idéal, coïncidence parfaite entre la solution numérique u_h et la solution exacte évaluée aux points du maillage.

Le contrôle de l'erreur de consistance est la première étape pour établir la convergence d'un schéma numérique.

Dans ce qui précède, on a défini l'erreur de consistance $\varepsilon_h(u)$ comme

$$\varepsilon_h(u) = A_h \pi_h(u) - b_h,$$

où $\pi_h(u)$ désigne le vecteur $(u(x_i))_{i \in \llbracket 1, N \rrbracket}$ des valeurs de la solution exacte aux points du maillage intérieur.

À présent, on définit l'erreur de convergence :

Définition 2.21 (Erreur de convergence). L'erreur de convergence $e_h(u)$ est définie comme la différence entre la solution exacte évaluée aux points du maillage intérieur et la solution numérique :

$$e_h(u) = \pi_h(u) - u_h,$$

où $\pi_h(u)$ est le vecteur de l'évaluation de la solution $(u(x_i))_{i \in \llbracket 1, N \rrbracket}$ et u_h est la solution numérique, c'est-à-dire la solution du système linéaire $A_h u_h = b_h$.

Comme dans le cas de la consistance, on dira que **le schéma est convergent pour une norme $\|\cdot\|$ donnée** si l'erreur de convergence tend vers zéro lorsque le pas de maillage h tend vers zéro. On dira également que le schéma est **convergent d'ordre p pour une norme $\|\cdot\|$ donnée** si

$$\exists C > 0, \exists h_0 > 0, \quad \forall h \leq h_0, \quad \|e_h(u)\| \leq Ch^p.$$

2. En dérivant deux fois l'équation $-u'' = f$.

À noter, dans la continuité de la [remarque 2.20](#) : si le schéma est exactement consistant et que la solution exacte est suffisamment régulière, alors le schéma est exactement convergent.

Pour revenir au lien entre la consistance et la convergence, observons qu'on a

$$A_h u_h = b_h \quad (31)$$

d'une part pour la solution numérique u_h , et

$$A_h \pi_h(u) = b_h + \varepsilon_h(u) \quad (32)$$

d'autre part pour la solution exacte. Par linéarité de l'opérateur A_h , on obtient, en soustrayant (31) de (32) :

$$A_h(\pi_h(u) - u_h) = \varepsilon_h(u),$$

et donc

$$e_h(u) = A_h^{-1} \varepsilon_h(u). \quad (33)$$

Aparté 2.22. Pour une norme $\|\cdot\|$ donnée sur $\mathbb{R}^N$, on utilise la **norme matricielle subordonnée** $\|\cdot\|$ pour les matrices de $M_N(\mathbb{R})$, qui est une norme adaptée à la norme vectorielle choisie. Cette norme se définit par

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\| \quad (\text{définitions équivalentes}).$$

Ainsi, par définition de la norme subordonnée, on obtient immédiatement l'inégalité

$$\frac{\|Ax\|}{\|x\|} \leq \|A\| \quad \forall x \neq 0,$$

autrement dit, pour tout vecteur non nul $x \in \mathbb{R}^N$, on a

$$\|Ax\| \leq \|A\| \|x\|. \quad (34)$$

Les normes subordonnées associées aux normes vectorielles usuelles ℓ^1 , ℓ^2 et ℓ^∞ ont également des expressions explicites sympathiques :

$$\|A\|_1 = \max_{1 \leq j \leq N} \sum_{i=1}^N |a_{ij}|, \quad \|A\|_2 = \sqrt{\rho(A^T A)}, \quad \|A\|_\infty = \max_{1 \leq i \leq N} \sum_{j=1}^N |a_{ij}|,$$

où $\rho(A^T A)$ désigne le rayon spectral de la matrice $A^T A$.

On peut maintenant contrôler l'erreur de convergence $e_h(u)$ en utilisant la norme subordonnée :

$$\underbrace{\|e_h(u)\|}_{\text{erreur de convergence}} = \|A_h^{-1} \varepsilon_h(u)\| \leq \|A_h^{-1}\| \cdot \underbrace{\|\varepsilon_h(u)\|}_{\text{erreur de consistance}}. \quad (35)$$

Il est donc clair que l'erreur de convergence est contrôlée par l'erreur de consistance, avec un facteur de proportionnalité donné par la norme subordonnée de l'inverse de l'opérateur discret A_h . Dans la suite, on va chercher une majoration de $\|A_h^{-1}\|$ pour obtenir une estimation pratique de l'erreur de convergence. Ce sera le résultat de stabilité.

2.4 Stabilité du schéma aux différences finies (21)

Proposition 2.23 (Stabilité du schéma aux différences finies (21)). *Les coefficients de la matrice A_h^{-1} sont tous positifs ou nuls. Par ailleurs, cette matrice vérifie l'inégalité de stabilité en norme infinie suivante :*

$$\|A_h^{-1}\|_\infty \leq \frac{1}{8}. \quad (36)$$

Par ailleurs, le schéma (21) est stable en norme infinie :

$$\|u_h\|_\infty \leq \frac{1}{8} \|b_h\|_\infty, \quad (37)$$

autrement dit, la solution discrète u_h est contrôlée par le second membre $b_h = b_h(f, u_g, u_d)$.

Remarque 2.24. *La stabilité exprimée dans (36) entraîne la stabilité au sens général :*

une légère variation du second membre (f) et/ou des conditions de bord (u_g, u_d) entraîne une variation proportionnellement contrôlée de la solution discrète.

En effet, si δb_h désigne une petite perturbation de b_h , alors la solution discrète subira une variation δu_h telle que

$$A_h(u_h + \delta u_h) = b_h + \delta b_h.$$

En conséquence, puisque $A_h u_h = b_h$, il vient

$$A_h \delta u_h = \delta b_h,$$

et donc

$$\delta u_h = A_h^{-1} \delta b_h.$$

Ainsi, en s'appuyant sur (36), il vient

$$\|\delta u_h\|_\infty \leq \|A_h^{-1}\|_\infty \|\delta b_h\|_\infty \stackrel{(36)}{\leq} \frac{1}{8} \|\delta b_h\|_\infty.$$

L'erreur commise sur la solution discrète est donc proportionnellement contrôlée par la perturbation du second membre.

Preuve de l'estimation (37) de la Proposition 2.23. Le contrôle (37) s'obtient directement du fait que $A_h u_h = b_h$ et du contrôle de la norme infinie de l'inverse de A_h (36) :

$$\|u_h\|_\infty = \|A_h^{-1} b_h\|_\infty \leq \|A_h^{-1}\|_\infty \|b_h\|_\infty \stackrel{(36)}{\leq} \frac{1}{8} \|b_h\|_\infty. \quad (38)$$

□

Pour obtenir (36)—qui constitue finalement la partie principale de la Proposition 2.23, on va avoir besoin du lemme suivant :

Lemme 2.25 (Principe du maximum discret). *La matrice A_h associée au schéma aux différences finies satisfait le principe du maximum discret :*

$$\boxed{\forall x \in \mathbb{R}^N, \quad [A_h x \succeq 0] \implies [x \succeq 0]}, \quad (39)$$

où, pour tout vecteur $y \in \mathbb{R}^N$, la notation $y \succeq 0$ signifie que toutes les composantes y_i de y sont positives ou nulles.

Aparté 2.26. En règle générale, on parle de principe du maximum lorsqu'un encadrement des données du problème permet d'en déduire un encadrement de la solution. Ici, en posant $y := A_h x \iff x = A_h^{-1} y$, (39) devient

$$[y \succeq 0] \implies [A_h^{-1} y \succeq 0],$$

où y joue typiquement le rôle du second membre b_h .

Preuve du Lemme 2.25. Soit $x \in \mathbb{R}^N$ tel que $A_h x \succeq 0$. Notons $y := A_h x$, alors $y \succeq 0$. Il s'agit de montrer que $x_i \geq 0$ pour tout $i \in \llbracket 1, N \rrbracket$.

Pour cela, on va chercher l'indice $k \in \llbracket 1, N \rrbracket$ pour lequel x_k est minimal : posons

$$k \in \operatorname{argmin}_{i \in \llbracket 1, N \rrbracket} \{x_i\},$$

et montrons que $x_k \geq 0$ ce qui permettra de conclure la preuve.

Point de vigilance 2.27. Noter que $\operatorname{argmin}_{i \in \llbracket 1, N \rrbracket} \{x_i\}$ n'est pas vide et peut contenir plusieurs indices ; auquel cas on peut choisir n'importe lequel de ces indices comme k .

On procède par dichotomie : on va distinguer deux cas selon que k est sur un bord (c'est-à-dire $k = 1$ ou $k = N$) ou à l'intérieur (c'est-à-dire $1 < k < N$).

- Si $k = 1$, en regardant la première ligne de $A_h x = y$, on a

$$2x_1 - x_2 = h^2 y_1. \quad (40)$$

Puisque $y \succeq 0$, on a $y_1 \geq 0$, donc

$$h^2 y_1 \geq 0. \quad (41)$$

Par ailleurs, comme x_1 est minimal, on a $x_1 \leq x_2$, d'où

$$2x_1 - x_2 \leq 2x_1 - x_1 = x_1. \quad (42)$$

En combinant (42)-(40)-(41), il vient

$$x_1 \stackrel{(42)}{\geq} 2x_1 - x_2 \stackrel{(40)}{=} h^2 y_1 \stackrel{(41)}{\geq} 0,$$

ce qui conclut le cas $k = 1$.

- Pour $k = N$, la méthode est identique : on trouve

$$x_N \geq 2x_N - x_{N-1} = h^2 y_N \geq 0.$$

- Enfin, si $1 < k < N$, la k -ième ligne de $A_h x = y$ nous donne

$$-x_{k-1} + 2x_k - x_{k+1} = h^2 y_k \geq 0. \quad (43)$$

Par ailleurs, comme x_k est minimal, on a $x_k \leq x_{k-1}$ et $x_k \leq x_{k+1}$, donc

$$-x_{k-1} + 2x_k - x_{k+1} \leq -x_k + 2x_k - x_k = 0. \quad (44)$$

En combinant (43) et (44), il vient

$$0 \stackrel{(44)}{\geq} -x_{k-1} + 2x_k - x_{k+1} \stackrel{(43)}{=} h^2 y_k \geq 0,$$

ce qui permet d'affirmer que

$$-x_{k-1} + 2x_k - x_{k+1} = 0, \quad (45)$$

ou, dit autrement,

$$x_k - x_{k-1} = x_{k+1} - x_k.$$

Maintenant, puisque $x_k \leq x_{k-1}$ et $x_k \leq x_{k+1}$, ce qui implique que

$$0 \geq x_k - x_{k-1} = x_{k+1} - x_k \geq 0$$

et force à avoir

$$x_k - x_{k-1} = x_{k+1} - x_k = 0,$$

si bien que si x est minimal à l'indice k , alors x est également minimal aux indices $k-1$ et $k+1$:

$$x_k = x_{k-1} = x_{k+1}.$$

En répétant ce raisonnement de proche en proche, on se ramène au cas $k=1$ ou $k=N$, pour lesquels on a déjà montré que $x_k \geq 0$.

En conclusion, x est systématiquement minimal et positif ou nul sur l'un de ses deux indices de bord. On a donc bien $x \succeq 0$. $\square$

On passe maintenant à la seconde partie de la preuve de la [Proposition 2.23](#) sur la stabilité du schéma.

Preuve de l'estimation (36) de la [Proposition 2.23](#). On commence par montrer que les coefficients de la matrice inverse A_h^{-1} sont tous positifs ou nuls :

$$\alpha_{ij} := (A_h^{-1})_{ij} \geq 0, \quad \forall (i, j) \in \llbracket 1, N \rrbracket^2. \quad (46)$$

Pour cela, on note, pour $j \in \llbracket 1, N \rrbracket$, c_j les vecteurs colonnes de A_h^{-1} :

$$A_h^{-1} = \left(\begin{array}{c|c|c|c} c_1 & c_2 & \cdots & c_N \end{array} \right).$$

En partant de la relation fondamentale $A_h A_h^{-1} = I_N$, on a

$$A_h \left(\begin{array}{c|c|c|c} c_1 & c_2 & \cdots & c_N \end{array} \right) = \left(\begin{array}{c|c|c|c} e_1 & e_2 & \cdots & e_N \end{array} \right), \quad (47)$$

où e_j est le j -ième vecteur de la base canonique de $\mathbb{R}^N$. La lecture de (47) colonne par colonne nous donne alors

$$A_h c_j = e_j \succeq 0, \quad \forall j \in \llbracket 1, N \rrbracket, \quad (48)$$

ce qui permet d'appliquer N fois le principe du maximum discret ([Lemme 2.25](#)) sur (48) :

$$\forall j \in \llbracket 1, N \rrbracket, \quad c_j \succeq 0, \quad (49)$$

et donc tous les coefficients de A_h^{-1} sont positifs ou nuls.

On passe maintenant à la majoration (36) de la norme de A_h^{-1} . On utilise la formule explicite dont on dispose pour la norme subordonnée ℓ^∞ :

$$\begin{aligned} \|A_h^{-1}\|_\infty &= \max_{1 \leq i \leq N} \sum_{j=1}^N |(A_h^{-1})_{ij}| \\ &= \max_{1 \leq i \leq N} \left| \sum_{j=1}^N (A_h^{-1})_{ij} \right| \quad \text{puisque tous les coefficients de } A_h^{-1} \text{ sont positifs ou nuls,} \\ &= \max_{1 \leq i \leq N} |A_h^{-1} \mathbf{1}_h|, \quad \text{avec } \mathbf{1}_h = (1, \dots, 1)^T \in \mathbb{R}^N, \\ &= \|A_h^{-1} \mathbf{1}_h\|_\infty. \end{aligned}$$

On va alors interpréter $A_h^{-1} \mathbf{1}_h$ comme la solution u_h du schéma numérique (22) $A_h u_h = b_h$ pour $b_h = \mathbf{1}_h$. En effet, on a bien

$$A_h(A_h^{-1} \mathbf{1}_h) = \mathbf{1}_h.$$

Maintenant, $b_h = \mathbf{1}_h$ se traduit coordonnée par coordonnée par

$$\begin{cases} f(x_1) + \frac{u_g}{h^2} = 1, \\ f(x_2) = 1, \\ \vdots \\ f(x_{N-1}) = 1, \\ f(x_N) + \frac{u_d}{h^2} = 1. \end{cases} \quad (50)$$

Une condition suffisante pour vérifier (50) pour tout h est de prendre $u_g = u_d = 0$ et $f \equiv 1$.

On est ainsi confronté à la solution numérique $A_h^{-1} \mathbf{1}_h$ correspondant au problème

$$\begin{cases} -u'' = 1 & \text{sur } (0, 1), \\ u(0) = 0, \\ u(1) = 0. \end{cases} \quad (51)$$

Dans ce cas précis, on a notamment $u^{(4)} = 0$, ce qui rend le schéma numérique exactement consistant :

$$A_h^{-1} \mathbf{1}_h = \pi_h u, \quad \text{pour tout } h > 0,$$

où u est la solution du problème (51).

Par ailleurs, on sait que cette solution u est régulière (cf. Proposition 2.2) sur le compact $[0, 1]$. La solution u est donc bornée, et majore à plus forte raison $A_h^{-1} \mathbf{1}_h$ puisque ses coordonnées sont des échantillons de u sur le maillage. On trouve ainsi

$$\|A_h^{-1}\|_\infty = \|A_h^{-1} \mathbf{1}_h\|_\infty \leq \|u\|_{L^\infty(0,1)}.$$

Il ne reste plus qu'à rendre explicite la borne supérieure $\|u\|_{L^\infty(0,1)}$ pour obtenir (36). Pour cela, remarquons qu'on a une solution explicite du problème (51) :

$$u(x) = \frac{1}{2}x(1-x), \quad x \in [0, 1],$$

et qu'elle est maximale en $x = \frac{1}{2}$, avec $u\left(\frac{1}{2}\right) = \frac{1}{8}$. Ce dernier point est laissé en exercice aux lectrices et lecteurs.

On en déduit finalement que

$$\|u\|_{L^\infty(0,1)} = \frac{1}{8},$$

ce qui recouvre la borne supérieure (36) attendue et conclut la démonstration. $\square$

2.5 Convergence du schéma aux différences finies (21)

Proposition 2.28 (Convergence du schéma aux différences finies (21)). *Si la solution u du problème de Poisson (11) est de classe $\mathcal{C}^4$ sur $[0, 1]$, alors le schéma aux différences finies (21) est convergent d'ordre 2 pour la norme $\|\cdot\|_\infty$.*

Preuve de la Proposition 2.28. Comme annoncé précédemment, la stabilité et la consistance du schéma aux différences finies (21) permettent d'assurer sa convergence. Plus précisément, en s'appuyant sur (35), à savoir

$$\underbrace{\|e_h(u)\|}_{\text{erreur de convergence}} \stackrel{\text{déf}}{=} \|\pi_h(u) - u_h\| \leq \|A_h^{-1}\| \cdot \underbrace{\|\varepsilon_h(u)\|}_{\text{erreur de consistance}}, \quad (52)$$

on trouve, en utilisant les résultats de stabilité (cf. (36)) et de consistance du schéma (cf. (30)),

$$\|e_h(u)\|_\infty \leq \frac{h^2}{96} \sup_{x \in [0,1]} |u^{(4)}(x)|. \quad (53)$$

La convergence d'ordre 2 annoncée s'ensuit immédiatement. $\square$

Remarque de fin de chapitre :

En TD et TP on s'intéressera au problème de Poisson avec conditions mixtes :

$$\begin{cases} -u'' = f & \text{sur } (0, 1), \\ u(0) = u_g, \\ u'(1) = 0. \end{cases}$$

Pour une analyse du problème elliptique plus général

$$-\operatorname{div}(A \nabla u) + \mathbf{b} \cdot \nabla u + cu = f,$$

on pourra consulter le livre de Grégoire Allaire, *Analyse et optimisation*.

3 Équation de la chaleur

Dans ce chapitre, on s'intéresse à l'équation de la chaleur en dimension $d = 1$. Plus précisément, on va travailler sur le problème de Cauchy (*i.e.* problème avec condition initiale) avec conditions de Dirichlet suivant :

$$\begin{cases} \partial_t u(t, x) - \partial_{xx} u(t, x) = f(t, x) & t \in (0, T), \quad x \in (0, 1), \quad (\text{équation intérieure}), \\ u(t, 0) = u_g(t), & t \in (0, T), \quad x = 0, \quad (\text{condition de bord à gauche}), \\ u(t, 1) = u_d(t), & t \in (0, T), \quad x = 1, \quad (\text{condition de bord à droite}), \\ u(0, x) = u_0(x) & x \in (0, 1), \quad (\text{condition initiale}). \end{cases} \quad (54)$$

Lorsque la donnée initiale u_0 possède des traces aux extrémités, par exemple si $u_0 \in H^1(0, 1)$, on dit qu'elle est **compatible** avec les conditions de bord lorsque

$$\begin{cases} u_0(0) = u_g(0), \\ u_0(1) = u_d(0). \end{cases}$$

Cette compatibilité n'est pas nécessaire au caractère bien posé du problème dans le cadre faible, mais elle améliore la régularité de la solution au voisinage de l'instant initial.

Dans le cadre considéré ci-dessous, on suppose seulement $u_0 \in L^2(0, 1)$. Les valeurs ponctuelles $u_0(0)$ et $u_0(1)$ ne sont alors pas définies en général, et aucune condition de compatibilité n'est imposée.

3.1 Remarques sur les solutions de (54)

Supposons que

$$u_g, u_d \in C^1([0, T]), \quad (\text{i.e. continûment dérivables sur } [0, T]), \quad (55)$$

$$u_0 \in L^2(0, 1), \quad (\text{i.e. de carré intégrable sur } (0, 1)), \quad (56)$$

et

$$f \in L^2(0, T; L^2(0, 1)) \quad (57)$$

$$\begin{aligned} &:= \{f : (0, T) \rightarrow L^2(0, 1) \text{ mesurable et } \int_{t=0}^T \|f(t, \cdot)\|_{L^2(0, 1)}^2 dt < \infty\} \\ &\simeq L^2((0, T) \times (0, 1)). \end{aligned}$$

Définition 3.1 (Notion de solution pour (54)). On appelle **solution** de (54) toute fonction

$$u \in \underbrace{C([0, T]; L^2(0, 1))}_{\text{condition initiale}} \cap \underbrace{L^2(0, T; H^1(0, 1))}_{\substack{\text{existence de } \partial_x u \text{ dans } L^2 \\ \text{et des valeurs de } u \text{ aux bords}}},$$

vérifiant les trois conditions suivantes :

1. **l'équation intérieure au sens variationnel** : pour toute fonction test

$$\varphi \in C_c^\infty(0, 1),$$

l'application

$$t \longmapsto \int_0^1 u(t, x) \varphi(x) dx$$

est absolument continue^a sur $[0, T]$ et, pour presque tout $t \in (0, T)$,

$$\frac{d}{dt} \int_0^1 u(t, x) \varphi(x) dx + \int_0^1 \partial_x u(t, x) \partial_x \varphi(x) dx = \int_0^1 f(t, x) \varphi(x) dx; \quad (58)$$

2. les conditions de Dirichlet :

$$u(t, 0) = u_g(t), \quad u(t, 1) = u_d(t), \quad \text{pour presque tout } t \in (0, T),$$

où les valeurs au bord sont comprises au sens des traces ;

3. la condition initiale :

$$u(0, \cdot) = u_0, \quad \text{dans } L^2(0, 1).$$

a. Une fonction absolument continue est, en particulier, dérivable presque partout. Cela permet de donner un sens à la dérivée temporelle $\frac{d}{dt} \int_0^1 u(t, x) \varphi(x) dx$ pour presque tout $t \in (0, T)$.

Remarque 3.2 (Sur l'espace $H^1(0, 1)$). L'espace $H^1(0, 1)$ est un espace de Sobolev, sous-espace de $L^2(0, 1)$, dans lequel la dérivée spatiale $\partial_x u$ peut être définie au sens faible et appartient elle aussi à $L^2(0, 1)$. Les fonctions de $H^1(0, 1)$ possèdent également des valeurs bien définies aux extrémités $x = 0$ et $x = 1$, appelées traces. Ces deux propriétés ne sont pas disponibles pour une fonction générale de $L^2(0, 1)$, irrégulière a priori et définie seulement presque partout.

Remarque 3.3 (Sur la formulation faible (58)). La formulation faible (58) est obtenue en multipliant la première ligne de (54) par une fonction test $\varphi \in C_c^\infty(0, 1)$ et en intégrant sur $(0, 1)$, puis en utilisant l'intégration par parties pour le terme en $\partial_x^2 u$.

On obtient ainsi une formulation qui ne nécessite pas que u soit deux fois dérivable par rapport à x , mais seulement qu'il existe une dérivée spatiale faible $\partial_x u$ appartenant à $L^2(0, 1)$. On parle de **formulation faible** car on rend plus gros l'ensemble des solutions admissibles—ce qui est parfois nécessaire pour attraper une solution lorsque les données sont irrégulières, quitte à montrer que la solution trouvée est en fait régulière a posteriori.

Proposition 3.4 (Stabilité, décroissance de l'énergie et unicité pour (54)). Sous les hypothèses (55)-(56)-(57) sur les données, toute solution de (54) vérifie les propriétés suivantes :

1. (Stabilité) on a l'estimation de stabilité

$$\sup_{t \in (0, T)} \|u(t, \cdot)\|_{L^2(0, 1)} \leq c \left(\|f\|_{L^2((0, T) \times (0, 1))} + \|u_g\|_{L^\infty(0, T)} + \|u_d\|_{L^\infty(0, T)} + \|u_0\|_{L^2(0, 1)} \right), \quad (59)$$

où $c > 0$ est une constante indépendante des données^a ;

2. (Décroissance de l'énergie) si $f \equiv 0$ et $u_g \equiv u_d \equiv 0$, alors l'énergie

$$E(t) := \int_0^1 |u(t, x)|^2 dx = \|u(t, \cdot)\|_{L^2(0, 1)}^2$$

est décroissante sur $(0, T)$;

3. (Unicité) le problème (54) admet au plus une solution.

a. Cette constante peut dépendre de T , mais pas des données f , u_g , u_d et u_0 .

Preuve de la Proposition 3.4, (1) stabilité. Le résultat de stabilité (59) est admis—voir le livre de Brigitte Lucquin, *Équations aux dérivées partielles et leurs approximations*, pour des éléments de la preuve, dont on donne ici une esquisse.

Par linéarité, l'idée consiste à décomposer la solution sous la forme

$$u = u^{(1)} + u^{(2)},$$

de façon que $u^{(1)}$ porte la source et la donnée initiale avec des conditions de Dirichlet homogènes, tandis que $u^{(2)}$ porte la contribution des données de bord.

Une estimation d'énergie appliquée à $u^{(1)}$ permet alors de contrôler

$$\sup_{t \in (0, T)} \|u^{(1)}(t, \cdot)\|_{L^2(0,1)}$$

par les normes de f et de u_0 . D'autre part, le principe du maximum permet de contrôler la contribution $u^{(2)}$ à l'aide des normes des données de bord u_g et u_d .

Enfin, l'inégalité triangulaire

$$\|u(t, \cdot)\|_{L^2(0,1)} \leq \|u^{(1)}(t, \cdot)\|_{L^2(0,1)} + \|u^{(2)}(t, \cdot)\|_{L^2(0,1)}$$

conduit à l'estimation (59). □

Preuve de la Proposition 3.4, (2–3) décroissance de l'énergie et unicité. Supposons tout d'abord que

$$f \equiv 0, \quad u_g \equiv u_d \equiv 0.$$

On introduit l'énergie

$$E(t) := \|u(t, \cdot)\|_{L^2(0,1)}^2 = \int_0^1 |u(t, x)|^2 dx.$$

En dérivant formellement par rapport au temps, on obtient

$$E'(t) = 2 \int_0^1 u(t, x) \partial_t u(t, x) dx.$$

Dans le cas considéré, l'équation de la chaleur s'écrit

$$\partial_t u = \partial_{xx} u,$$

d'où

$$E'(t) = 2 \int_0^1 u(t, x) \partial_{xx} u(t, x) dx.$$

Une intégration par parties donne

$$E'(t) = 2 [u(t, x) \partial_x u(t, x)]_{x=0}^{x=1} - 2 \int_0^1 |\partial_x u(t, x)|^2 dx.$$

Les conditions de Dirichlet homogènes

$$u(t, 0) = u(t, 1) = 0$$

annulent le terme de bord. Ainsi,

$$E'(t) = -2 \|\partial_x u(t, \cdot)\|_{L^2(0,1)}^2 \leq 0.$$

L'énergie E est donc décroissante sur $(0, T)$.

Cette propriété permet également d'obtenir l'unicité. Soient u et v deux solutions de (54) associées aux mêmes données. Par linéarité, leur différence

$$w := u - v$$

est solution du problème homogène, avec conditions de Dirichlet homogènes et donnée initiale nulle :

$$f \equiv 0, \quad w(t, 0) = w(t, 1) = 0, \quad w(0, \cdot) = 0 \quad \text{dans } L^2(0, 1).$$

L'énergie associée à w ,

$$E_w(t) := \|w(t, \cdot)\|_{L^2(0,1)}^2,$$

est donc décroissante. Or

$$E_w(0) = 0$$

et, par définition,

$$E_w(t) \geq 0.$$

Par conséquent,

$$0 \leq E_w(t) \leq E_w(0) = 0,$$

et donc

$$E_w(t) = 0, \quad \forall t \in [0, T].$$

Ainsi,

$$w \equiv 0,$$

c'est-à-dire

$$u \equiv v.$$

Le problème (54) admet donc au plus une solution. $\square$

Proposition 3.5 (Représentation en série de Fourier de la solution homogène). *Dans le cas où $f \equiv 0$, $u_g \equiv u_d \equiv 0$, la solution u de (54) s'écrit*

$$u(t, x) = \sum_{k=1}^{\infty} b_k e^{-k^2 \pi^2 t} \sin(k\pi x), \quad \text{dans } L^2(0, 1),$$

où

$$b_k = 2 \int_0^1 u_0(x) \sin(k\pi x) dx.$$

Preuve de la Proposition 3.5. On décompose la solution en série de Fourier adaptée aux conditions de Dirichlet homogènes (Fourier sinus) : puisque pour tout $t \in [0, T]$, $u(t, \cdot)$ est un élément de $L^2(0, 1)$, il existe une suite $(\beta_k(t))_{k \geq 1}$ telle que

$$u(t, x) = \sum_{k=1}^{\infty} \beta_k(t) \sin(k\pi x), \quad \text{dans } L^2(0, 1) \text{ pour tout } t \in [0, T]. \quad (60)$$

En injectant formellement (60) dans l'équation

$$\partial_t u - \partial_{xx} u = 0,$$

et en dérivant formellement terme à terme, on obtient

$$\sum_{k=1}^{\infty} \left(\beta'_k(t) + k^2 \pi^2 \beta_k(t) \right) \sin(k\pi x) = 0, \quad \text{dans } L^2(0, 1).$$

Par unicité du développement en série de Fourier sinus, on en déduit que, pour tout $k \geq 1$, β_k satisfait l'EDO linéaire

$$\beta'_k(t) + k^2 \pi^2 \beta_k(t) = 0, \quad t \in (0, T), \quad (61)$$

munie de la condition initiale

$$\beta_k(0) = 2 \int_0^1 u_0(x) \sin(k\pi x) dx =: b_k. \quad (62)$$

La résolution de cette EDO est immédiate et donne

$$\beta_k(t) = \beta_k(0)e^{-k^2\pi^2 t} = b_k e^{-k^2\pi^2 t}.$$

Ainsi, on trouve bien que

$$u(t, x) = \sum_{k=1}^{\infty} b_k e^{-k^2\pi^2 t} \sin(k\pi x), \quad \text{dans } L^2(0, 1),$$

ce qui conclut la preuve. $\square$

3.2 Construction de schémas aux différences finies pour l'équation de chaleur

On se propose ici de construire des schémas numériques aux différences finies pour résoudre le problème de Cauchy (54).

Discretisation. On commence par construire un **maillage espace-temps** uniforme de $[0, 1] \times [0, T]$: pour deux entiers N (espace) et M (temps) de $\mathbb{N}^*$, on définit les pas de discrétisation

$$h = \Delta x := \frac{1}{N+1}, \quad \text{et} \quad \tau = \Delta t := \frac{T}{M+1}.$$

On note x_i et t_j les points et instants discrets, respectivement,

$$\begin{aligned} x_i &= i\Delta x, & i &\in \llbracket 0, N+1 \rrbracket, \\ t_j &= j\Delta t, & j &\in \llbracket 0, M+1 \rrbracket. \end{aligned}$$

Ainsi, les solutions discrètes u^j sont définies pour $j \in \llbracket 0, M+1 \rrbracket$, tandis que les itérations $j \mapsto j+1$ sont effectuées pour $j \in \llbracket 0, M \rrbracket$. Les termes sources et les erreurs de consistance associés à une itération sont donc indexés par $j \in \llbracket 0, M \rrbracket$.

L'objectif est d'approcher la solution u du problème continu aux points de discrétisation intérieurs (t_j, x_i) , pour $i \in \llbracket 1, N \rrbracket$ et $j \in \llbracket 1, M+1 \rrbracket$ par l'intermédiaire de l'inconnue u_i^j .

Pour les bords, on pose

- $u_0^j = u(0, t_j) = u_g(t_j)$, car $u(t, 0) = u_g(t)$, sur le bord $\{x = 0\}$ (○),
- $u_{N+1}^j = u(1, t_j) = u_d(t_j)$, car $u(t, 1) = u_d(t)$, sur le bord $\{x = 1\}$ (○),
- $u_i^0 = u(0, x_i) = u_0(x_i)$, car $u(0, x) = u_0(x)$, au temps initial $\{t = 0\}$ (●).

Pour $j \in \llbracket 1, M+1 \rrbracket$

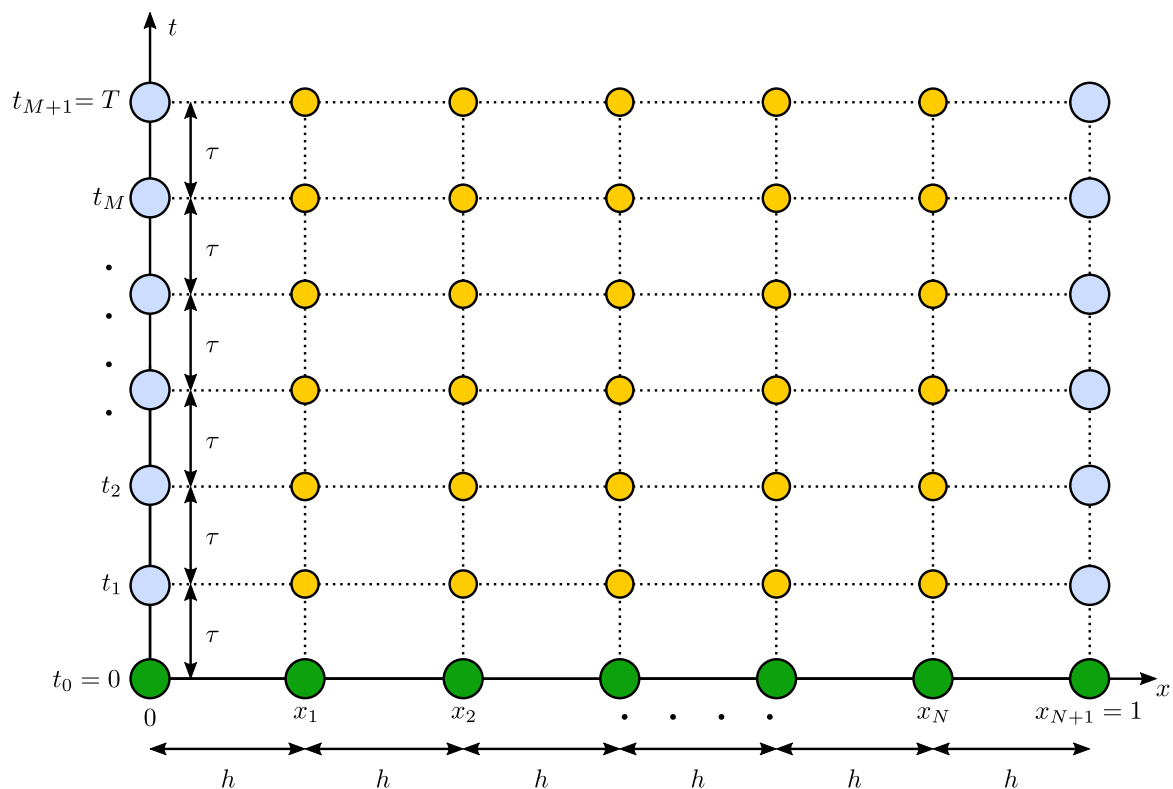
$$u^j = \begin{pmatrix} u_1^j \\ u_2^j \\ \vdots \\ u_N^j \end{pmatrix} \in \mathbb{R}^N,$$

le vecteur des inconnues u_i^j à l'instant t_j . Le vecteur initial u^0 est quant à lui connu :

$$u^0 = \begin{pmatrix} u_1^0 \\ u_2^0 \\ \vdots \\ u_N^0 \end{pmatrix} \in \mathbb{R}^N.$$

Point de vigilance 3.6. *Noter qu'en tant que solution numérique, u^j dépend des pas de discrétisation h et τ et devrait donc être notée $u_{h,\tau}^j$. On fera souvent l'abstraction de cette dépendance dans la notation pour alléger l'écriture. Il convient toutefois de toujours bien faire la différence entre la solution continue u et la solution approchée u^j .*

De même, certains indices h et τ pourront être également effacés des notations par souci de lisibilité. Normalement, la nature de chaque objet et sa dépendance aux pas de discrétisation sont claires dans le contexte.



Approximation des dérivées partielles. On utilise des différences finies centrées pour approcher les dérivées spatiales et des différences finies avant pour les dérivées temporelles :

$$\frac{\partial^2 u}{\partial x^2}(t_j, x_i) \approx \frac{u_{i+1}^j - 2u_i^j + u_{i-1}^j}{h^2}, \quad (\text{comme dans le chapitre précédent})$$

$$\frac{\partial u}{\partial t}(t_j, x_i) \approx \frac{u_i^{j+1} - u_i^j}{\tau}, \quad \text{ou bien} \quad \frac{u_i^j - u_i^{j-1}}{\tau}, \quad \text{ou bien} \quad \frac{u_i^{j+1} - u_i^{j-1}}{2\tau}. \quad (63)$$

On peut choisir différents taux d'accroissement pour approcher la dérivée temporelle : avant, arrière ou centrée. Cela mène à différents schémas aux différences finies, chacun ayant ses propres avantages et inconvénients.

Point de vigilance 3.7. *Sur la notion de schéma **EXPLICITE**/**IMPLICITE**.*

*Un schéma est dit **explicite** si les valeurs au nouvel instant t_{j+1} se calculent directement à partir de valeurs déjà connues.*

*Il est dit **implicite** si les inconnues au temps t_{j+1} apparaissent elles-mêmes dans le schéma, de sorte qu'il faut résoudre un système^a pour les déterminer.*

En pratique : si u^{j+1} est obtenu directement, le schéma est explicite ; s'il faut résoudre un système en u^{j+1} , il est implicite.

On peut donner un exemple simple, même dans le cadre des EDO : pour l'EDO $X' = X^2$:

$$\text{le schéma } X^{j+1} = X^j + \tau(X^j)^2 \text{ est } \mathbf{explicite} ; \quad (64)$$

$$\text{le schéma } X^{j+1} = X^j + \tau(X^{j+1})^2 \text{ est } \mathbf{implicite} ; \quad (65)$$

$$\text{le schéma } X^{j+1} = X^j + \frac{\tau}{2} \left((X^j)^2 + (X^{j+1})^2 \right) \text{ est } \mathbf{implicite}. \quad (66)$$

On voit bien que dans le premier cas (64), calculer X^{j+1} à partir de X^j ne pose pas de problème (il suffit d'appliquer la formule) ; tandis que dans les deux derniers cas (65) et (66), isoler X^{j+1} nécessite de résoudre une équation non linéaire (ce qui est déjà moins trivial).

a. i.e. faire des opérations algébriques pour trouver u^{j+1} si le système est linéaire, ou faire une méthode de Newton si le système est non-linéaire.

3.3 Un premier schéma : Euler explicite

3.3.1 Le schéma explicite

On choisit l'approximation avant pour la dérivée temporelle dans (63) :

$$\frac{\partial u}{\partial t}(t_j, x_i) \approx \frac{u_i^{j+1} - u_i^j}{\tau},$$

et la différence finie centrée pour le laplacien en espace (comme dans le chapitre précédent) :

$$\frac{\partial^2 u}{\partial x^2}(t_j, x_i) \approx \frac{u_{i+1}^j - 2u_i^j + u_{i-1}^j}{h^2}.$$

On obtient ainsi le schéma aux différences finies explicite suivant :

$$\frac{u_i^{j+1} - u_i^j}{\tau} - \frac{u_{i+1}^j - 2u_i^j + u_{i-1}^j}{h^2} = f_i^j, \quad \text{pour tout } i \in \llbracket 1, N \rrbracket \text{ et } j \in \llbracket 0, M \rrbracket, \quad (67)$$

où $f_i^j = f(t_j, x_i)$ est la source discrétisée. On peut réécrire (67) sous la forme du schéma itératif suivant :

$$u_i^{j+1} = u_i^j + \frac{\tau}{h^2} \left(u_{i+1}^j - 2u_i^j + u_{i-1}^j \right) + \tau f_i^j, \quad \text{pour tout } i \in \llbracket 1, N \rrbracket \text{ et } j \in \llbracket 0, M \rrbracket. \quad (68)$$

À noter que ce schéma est bien explicite : les inconnues u_i^{j+1} sont directement calculables à partir des valeurs connues u_i^j .

À nouveau, comme dans le chapitre précédent, on peut réécrire le schéma (67) sous forme matricielle :

$$u^{j+1} = u^j - \tau A_h u^j + \tau b_h^j, \quad \text{pour tout } j \in \llbracket 0, M \rrbracket, \quad (69)$$

où A_h est la matrice de discrétisation spatiale du (moins) Laplacien et b_h^j est l'analogue (temporel) du second membre b_h défini dans le chapitre précédent—cf. (23) :

$$A_h = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ 0 & 0 & -1 & 2 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & -1 \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{pmatrix} \quad \text{et} \quad b_h^j = \begin{pmatrix} f(t_j, x_1) + \frac{u_g(t_j)}{h^2} \\ f(t_j, x_2) \\ f(t_j, x_3) \\ \vdots \\ f(t_j, x_{N-1}) \\ f(t_j, x_N) + \frac{u_d(t_j)}{h^2} \end{pmatrix}. \quad (70)$$

Point de vigilance 3.8. Noter que A_h contient déjà la multiplication par $1/h^2$, d'où son nom : A_h .

3.3.2 Consistance du schéma explicite (68)

Comme dans le chapitre précédent (cf. (26)), on définit la projection π_h , agissant sur une fonction $u(t, \cdot)$ (qu'on va simplement noter $u(t)$ pour faire court), définie sur $[0, 1]$ par

$$\pi_h u(t, \cdot) = \pi_h u(t) = \begin{pmatrix} u(t, x_1) \\ u(t, x_2) \\ \vdots \\ u(t, x_N) \end{pmatrix} \in \mathbb{R}^N. \quad (71)$$

Définition 3.9 (Erreur de consistance du schéma explicite). Étant donné une solution exacte u de (54), on définit l'erreur de consistance $\varepsilon_{h,\tau}^j$ du schéma explicite (67) par

$$\varepsilon_{h,\tau}^j(u) = \frac{\pi_h u(t_{j+1}) - \pi_h u(t_j)}{\tau} + A_h \pi_h u(t_j) - b_h^j, \quad \text{pour tout } j \in \llbracket 0, M \rrbracket. \quad (72)$$

(On met la solution exacte u dans le schéma numérique et on regarde ce qui reste...)

(Noter que $\varepsilon_{h,\tau}^j$ est un vecteur de $\mathbb{R}^N$ pour tout $j \in \llbracket 0, M \rrbracket$.)

On dit que le schéma est **consistant** si

$$\lim_{(h,\tau) \rightarrow (0,0)} \max_{j \in \llbracket 0, M \rrbracket} \|\varepsilon_{h,\tau}^j(u)\|_\infty = 0,$$

c'est-à-dire si l'élément maximal de $((\varepsilon_{h,\tau}^j(u))_{i,j})_{(i,j) \in \llbracket 1, N \rrbracket \times \llbracket 0, M \rrbracket}$ tend vers 0 lorsque les pas h et τ tendent vers 0.

Proposition 3.10 (Consistance du schéma explicite). Si $u \in C^2([0, T]; C^4([0, 1]))$ est une solution de (54), alors le schéma (67) est consistant, d'ordre 1 en temps et d'ordre 2 en espace pour la norme $\|\cdot\|_\infty$.

Plus précisément, on a l'estimation

$$\max_{j \in \llbracket 0, M \rrbracket} \|\varepsilon_{h,\tau}^j(u)\|_\infty \leq \frac{h^2}{12} \sup_{[0,T] \times [0,1]} \left| \frac{\partial^4 u}{\partial x^4} \right| + \frac{\tau}{2} \sup_{[0,T] \times [0,1]} \left| \frac{\partial^2 u}{\partial t^2} \right|. \quad (73)$$

La preuve de cette proposition sera faite lors du TD 2. Dans l'exercice en question, on montre ensuite directement la convergence du schéma à partir de la consistance, sans passer par la stabilité.

On peut également montrer la stabilité du schéma puis en déduire la convergence en utilisant à la fois la consistance et la stabilité. C'est ce qu'on va faire maintenant.

3.3.3 Stabilité du schéma explicite (68)

Proposition 3.11 (Stabilité du schéma explicite). *Sous la condition de CFL^a*

$$\boxed{\frac{\tau}{h^2} \leq \frac{1}{2}}, \quad (74)$$

le schéma (68) est stable pour la norme $\|\cdot\|_\infty$. Plus précisément,

$$\max_{j \in \llbracket 0, M+1 \rrbracket} \|u^j\|_\infty \leq \max \left\{ \|u^0\|_\infty, \max_{j \in \llbracket 0, M \rrbracket} |u_g(t_j)|, \max_{j \in \llbracket 0, M \rrbracket} |u_d(t_j)| \right\} + T \max_{j \in \llbracket 0, M \rrbracket} \|f^j\|_\infty. \quad (75)$$

Ainsi, la solution discrète reste uniformément contrôlée par les données initiales, les conditions aux limites et le terme source, indépendamment des pas de discrétisation h et τ .

^a. CFL pour Courant-Friedrichs-Lewy, du nom des trois mathématiciens qui ont introduit cette condition dans les années 1920.

Remarque 3.12. Le contrôle (75) est un analogue discret de l'estimation de stabilité (59) pour le problème continu. On retrouve la même structure : la norme de la solution est contrôlée par les normes des données.

À noter que lorsque $f \equiv 0$, l'estimation (75) donne directement le principe du maximum discret

$$\max_{i,j} |u_i^j| \leq \max \left\{ \max_i |u_i^0|, \max_j |u_g(t_j)|, \max_j |u_d(t_j)| \right\}.$$

Preuve de la Proposition 3.11. Posons

$$r := \frac{\tau}{h^2} \quad (\leq \frac{1}{2} \text{ par hypothèse CFL (74)}).$$

Le schéma (68) s'écrit sous la forme

$$u_i^{j+1} = r u_{i-1}^j + (1 - 2r) u_i^j + r u_{i+1}^j + \tau f_i^j, \quad \text{pour tout } i \in \llbracket 1, N \rrbracket, \quad (76)$$

où l'on prolonge le schéma (76) jusqu'aux indices de bord $i = 0$ et $i = N + 1$ en posant

$$u_0^j := u_g(t_j) \quad \text{et} \quad u_{N+1}^j := u_d(t_j).$$

Sous la condition CFL $r \leq 1/2$, les coefficients

$$r, \quad 1 - 2r, \quad r$$

sont positifs et leur somme vaut 1. Par conséquent, la combinaison dans (76),

$$r u_{i-1}^j + (1 - 2r) u_i^j + r u_{i+1}^j$$

est convexe, et donc bornée par le maximum des trois termes :

$$|u_i^{j+1}| \leq \max \left\{ |u_{i-1}^j|, |u_i^j|, |u_{i+1}^j| \right\} + \tau |f_i^j|. \quad (77)$$

En posant

$$B := \max \left\{ \max_j |u_g(t_j)|, \max_j |u_d(t_j)| \right\}, \quad F := \max_j \|f^j\|_\infty,$$

on obtient donc, en passant au maximum sur $i \in \llbracket 1, N \rrbracket$ dans (77) :

$$\|u^{j+1}\|_\infty \leq \max \left\{ \|u^j\|_\infty, B \right\} + \tau F. \quad (78)$$

Posons maintenant

$$K := \max \left\{ \|u^0\|_\infty, B \right\}.$$

En utilisant (78) pour $j = 0$, il vient

$$\begin{aligned} \|u^1\|_\infty &\leq \max \left\{ \|u^0\|_\infty, B \right\} + \tau F \\ &= K + \tau F. \end{aligned}$$

En utilisant (78) pour $j = 1$, il vient

$$\begin{aligned} \|u^2\|_\infty &\leq \max \left\{ \|u^1\|_\infty, B \right\} + \tau F \\ &\leq \max \left\{ K + \tau F, B \right\} + \tau F \\ &= K + \tau F + \tau F \\ &= K + 2\tau F. \end{aligned}$$

Une récurrence sur j donne ainsi

$$\begin{aligned} \|u^j\|_\infty &\leq K + j\tau F \\ &= K + t_j F \\ &\leq K + TF \quad (\text{indépendant de } j). \end{aligned}$$

On trouve ainsi en passant au maximum sur $j \in \llbracket 0, M+1 \rrbracket$:

$$\max_j \|u^j\|_\infty \leq \underbrace{\max \left\{ \|u^0\|_\infty, \underbrace{\max_j |u_g(t_j)|, \max_j |u_d(t_j)|}_{=B} \right\}}_{=K} + \underbrace{T \max_j \|f^j\|_\infty}_{=TF},$$

ce qui redonne bien l'estimation (75) annoncée. $\square$

3.3.4 Convergence du schéma explicite (68)

Comme annoncé plus haut, on peut montrer la convergence du schéma à partir de la stabilité uniquement.

Proposition 3.13 (Convergence du schéma explicite). *Si la solution u du problème continu (54) est C^4 en espace et C^2 en temps, et si la condition de CFL (74) $\frac{\tau}{h^2} \leq \frac{1}{2}$ est satisfaite, alors le schéma explicite (68) est convergent pour la norme $\|\cdot\|_\infty$, d'ordre 1 en temps et d'ordre 2 en espace. Plus précisément, on a l'estimation*

$$\max_{j \in \llbracket 0, M+1 \rrbracket} \|u^j - \pi_h u(t_j)\|_\infty \leq T \left(\frac{h^2}{12} \sup_{[0,T] \times [0,1]} \left| \frac{\partial^4 u}{\partial x^4} \right| + \frac{\tau}{2} \sup_{[0,T] \times [0,1]} \left| \frac{\partial^2 u}{\partial t^2} \right| \right). \quad (79)$$

Preuve de la Proposition 3.13. On peut faire la preuve en utilisant uniquement la consistance (cf. le TD 2).

Ici, on va utiliser la consistance (Proposition 3.10) et la stabilité (Proposition 3.11) pour montrer la convergence.

On a notre schéma explicite (67) :

$$u^{j+1} = u^j - \tau A_h u^j + \tau b_h^j, \quad \text{pour tout } j \in \llbracket 0, M \rrbracket,$$

et la solution exacte u du problème continu (54) satisfait, par définition de l'erreur de consistance (72),

$$\pi_h u(t_{j+1}) = \pi_h u(t_j) - \tau A_h \pi_h u(t_j) + \tau b_h^j + \tau \varepsilon_{h,\tau}^j, \quad \text{pour tout } j \in \llbracket 0, M \rrbracket. \quad (80)$$

Maintenant en notant l'erreur de convergence

$$e^j := u^j - \pi_h u(t_j), \quad \text{pour tout } j \in \llbracket 0, M+1 \rrbracket,$$

on trouve que

$$\begin{aligned} e^{j+1} &= u^{j+1} - \pi_h u(t_{j+1}) \\ &\stackrel{(69)(80)}{=} \left(u^j - \tau A_h u^j + \tau b_h^j \right) - \left(\pi_h u(t_j) - \tau A_h \pi_h u(t_j) + \tau b_h^j + \tau \varepsilon_{h,\tau}^j \right) \\ &= \left(u^j - \pi_h u(t_j) \right) - \tau A_h \left(u^j - \pi_h u(t_j) \right) - \tau \varepsilon_{h,\tau}^j \\ &= e^j - \tau A_h e^j - \tau \varepsilon_{h,\tau}^j. \end{aligned}$$

Ainsi, l'erreur de convergence e^j satisfait la relation de récurrence suivante :

$$e^{j+1} = e^j - \tau A_h e^j - \tau \varepsilon_{h,\tau}^j, \quad \text{pour tout } j \in \llbracket 0, M \rrbracket. \quad (81)$$

On reconnaît alors dans (81) le schéma explicite (69) appliqué à l'erreur e^j , avec :

- pour second membre,

$$b_h^j = -\varepsilon_{h,\tau}^j;$$

- pour condition initiale,

$$e^0 = u^0 - \pi_h u(0) = 0_{\mathbb{R}^N},$$

car le schéma est initialisé exactement sur la donnée initiale : $u^0 = \pi_h u(0)$;

- pour condition de bord à gauche, en prolongeant l'erreur aux indices de bord,

$$e_0^j := u_0^j - u(t_j, 0) = u_g(t_j) - u_g(t_j) = 0;$$

- de même, pour la condition de bord à droite,

$$e_{N+1}^j := u_{N+1}^j - u(t_j, 1) = u_d(t_j) - u_d(t_j) = 0.$$

On peut par conséquent appliquer au schéma (81) le résultat de stabilité de la Proposition 3.11 pour obtenir, sous la condition CFL. Plus précisément, (75) donne

$$\begin{aligned} \max_{j \in \llbracket 0, M+1 \rrbracket} \|e^j\|_\infty &\leq \max \left\{ \underbrace{\|e^0\|_\infty}_{=0}, \max_{j \in \llbracket 0, M+1 \rrbracket} \underbrace{|e_0^j|}_{=0}, \max_{j \in \llbracket 0, M+1 \rrbracket} \underbrace{|e_{N+1}^j|}_{=0} \right\} + T \max_{j \in \llbracket 0, M \rrbracket} \|\varepsilon_{h,\tau}^j\|_\infty \\ &= T \max_{j \in \llbracket 0, M \rrbracket} \|\varepsilon_{h,\tau}^j\|_\infty. \end{aligned}$$

Ainsi, en utilisant l'estimation de consistance (73) de la Proposition 3.10, on obtient bien le résultat de convergence (79) annoncé :

$$\max_{j \in \llbracket 0, M+1 \rrbracket} \|e^j\|_\infty \leq T \left(\frac{h^2}{12} \|u_{xxx}\|_{L_{t,x}^\infty} + \frac{\tau}{2} \|u_{tt}\|_{L_{t,x}^\infty} \right).$$

□

Remarque 3.14 (Sur la condition CFL). La condition CFL (74) impose en particulier que

$$\tau \leq \frac{h^2}{2}.$$

Donc si on souhaite augmenter la précision du schéma numérique en espace (i.e. en diminuant h), il faudra diminuer τ de manière significativement plus forte (comme le carré de h).

Ainsi, si on veut diviser par 10 le pas d'espace h , il faudra diviser par 100 le pas de temps τ . C'est un inconvénient majeur des schémas explicites pour l'équation de la chaleur : raffiner le maillage en espace entraîne une diminution drastique du pas de temps et donc beaucoup plus d'itérations pour atteindre l'horizon temporel T .

L'utilité du schéma implicite est justement de ne pas avoir cette contrainte CFL sur le pas de temps τ et de pouvoir choisir τ indépendamment de h tout en conservant la stabilité du schéma.

3.4 Un deuxième schéma : Euler implicite

3.4.1 Le schéma implicite

Au lieu d'utiliser le schéma explicite pour approcher la dérivée temporelle

$$\partial_t u(t_j, x_i) \approx \frac{u(t_{j+1}, x_i) - u(t_j, x_i)}{\tau} \quad \text{(schéma explicite),}$$

on va maintenant utiliser le schéma implicite :

$$\partial_t u(t_{j+1}, x_i) \approx \frac{u(t_{j+1}, x_i) - u(t_j, x_i)}{\tau} \quad \text{(schéma implicite).} \quad (82)$$

On obtient ainsi le schéma itératif implicite pour l'équation de la chaleur (54) :

$$\frac{u_i^{j+1} - u_i^j}{\tau} - \frac{u_{i+1}^{j+1} - 2u_i^{j+1} + u_{i-1}^{j+1}}{h^2} = f_i^{j+1}, \quad \text{pour tout } i \in \llbracket 1, N \rrbracket \text{ et } j \in \llbracket 0, M \rrbracket, \quad (83)$$

que l'on pourra comparer avec le schéma explicite (67), et que l'on écrit également sous la forme matricielle :

$$u^{j+1} = u^j - \tau A_h u^{j+1} + \tau b_h^{j+1}, \quad \text{pour tout } j \in \llbracket 0, M \rrbracket, \quad (84)$$

où, comme pour le schéma explicite, A_h est la même matrice de discrétisation spatiale du (moins) Laplacien, et b_h^{j+1} est l'analogue (temporel) du second membre b_h défini dans le chapitre précédent ; on les rappelle ici :

$$A_h = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ 0 & 0 & -1 & 2 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & -1 \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{pmatrix} \quad \text{et} \quad b_h^{j+1} = \begin{pmatrix} f(t_{j+1}, x_1) + \frac{u_g(t_{j+1})}{h^2} \\ f(t_{j+1}, x_2) \\ f(t_{j+1}, x_3) \\ \vdots \\ f(t_{j+1}, x_{N-1}) \\ f(t_{j+1}, x_N) + \frac{u_d(t_{j+1})}{h^2} \end{pmatrix}. \quad (85)$$

Notons que la résolution du schéma implicite (84) n'est plus directe, comme c'était le cas pour le schéma explicite où on trouvait directement u^{j+1} en fonction de u^j par simple application du schéma. Ici, on pourra voir que (84) équivaut à **résoudre** le système linéaire suivant :

$$\underbrace{(I_N + \tau A_h)}_{=: M_{\tau, h}} u^{j+1} = u^j + \tau b_h^{j+1}, \quad \text{pour tout } j \in \llbracket 0, M \rrbracket. \quad (86)$$

Dans la suite, on notera simplement $M = M_{\tau, h} = I_N + \tau A_h$ pour alléger les notations. La matrice M est ainsi tridiagonale de taille $N \times N$, avec pour coefficients :

- $1 + 2\frac{\tau}{h^2}$ sur la diagonale,
- $-\frac{\tau}{h^2}$ sur les sous-diagonales et les sur-diagonales,
- 0 ailleurs.

On peut voir immédiatement que la matrice M est symétrique et définie positive en s'appuyant sur le fait que A_h est symétrique et définie positive (cf. la preuve de la [Proposition 2.11](#) au chapitre sur l'équation de Poisson) :

Proposition 3.15 (Caractère bien posé du schéma implicite). *La matrice $M = I_N + \tau A_h$ est symétrique et définie positive, ce qui garantit que le système linéaire (86) admet une unique solution.*

Preuve de la Proposition 3.15. En utilisant la notation du produit scalaire usuel sur $\mathbb{R}^N$, à savoir

$$(x, y) = y^T x,$$

on a pour tout vecteur non nul $x \in \mathbb{R}^N \setminus \{0\}$,

$$\begin{aligned} (Mx, x) &= ((I_N + \tau A_h)x, x) \\ &= (x, x) + \tau(A_h x, x) \\ &= \underbrace{\|x\|^2}_{>0} + \tau \underbrace{(A_h x, x)}_{>0} > 0, \end{aligned}$$

puisque A_h est définie positive. Il en résulte que M est définie positive. □

3.4.2 Consistance du schéma implicite (83)

Définition 3.16 (Erreur de consistance du schéma implicite). *On définit l'erreur de consistance $\varepsilon_{h,\tau}^{j+1}$ du schéma implicite (83) à l'instant t_{j+1} par*

$$\varepsilon_{h,\tau}^{j+1}(u) = \frac{\pi_h u(t_{j+1}) - \pi_h u(t_j)}{\tau} + A_h \pi_h u(t_{j+1}) - b_h^{j+1}, \quad \text{pour tout } j \in \llbracket 0, M \rrbracket. \quad (87)$$

(On met la solution exacte u dans le schéma numérique et on regarde ce qui reste...)

On pourra comparer avec la [Définition 3.9](#) pour l'erreur de consistance du schéma explicite. Noter également que la notation de projection π_h est la même que celle utilisée pour le schéma explicite et est définie dans (71).

Proposition 3.17 (Consistance du schéma implicite). *Si $u \in C^2([0, T]; C^4([0, 1]))$ est une solution de (54), alors le schéma (83) est consistant, d'ordre 1 en temps et d'ordre 2 en espace pour la norme $\|\cdot\|_\infty$.*

Plus précisément, on a l'estimation

$$\max_{j \in \llbracket 0, M \rrbracket} \|\varepsilon_{h,\tau}^{j+1}(u)\|_\infty \leq \frac{h^2}{12} \sup_{[0,T] \times [0,1]} \left| \frac{\partial^4 u}{\partial x^4} \right| + \frac{\tau}{2} \sup_{[0,T] \times [0,1]} \left| \frac{\partial^2 u}{\partial t^2} \right|. \quad (88)$$

Preuve de la Proposition 3.17. La preuve de la [Proposition 3.17](#) est analogue à celle de la [Proposition 3.10](#) pour le schéma explicite et est laissée en exercice à la lectrice ou au lecteur. □

3.4.3 Stabilité et convergence du schéma implicite

Le schéma implicite (83) est inconditionnellement stable et convergent, contrairement au schéma explicite dont la stabilité impose une condition liant le pas de temps τ au pas d'espace h .

Ces propriétés seront établies dans la section suivante à l'aide de résultats généraux, notamment du théorème de Lax–Richtmyer.

3.5 Résultats généraux sur la consistance, la stabilité et la convergence

Les schémas explicite et implicite (69) et (86) sont deux exemples de schémas numériques pour approcher la solution du problème de la chaleur avec conditions de Dirichlet (54).

Ces schémas prennent la forme matricielle itérative générale :

$$\mathcal{M}_h u^{j+1} = \mathcal{N}_h u^j + \tau b_{h,\tau}^j, \quad (89)$$

avec

- Dans le cas explicite (69)
 - $\mathcal{M}_h = I_N$,
 - $\mathcal{N}_h = I_N - \tau A_h$,
 - $b_{h,\tau}^j$ défini dans (70) ;
- Dans le cas implicite (86)
 - $\mathcal{M}_h = I_N + \tau A_h$,
 - $\mathcal{N}_h = I_N$,
 - $b_{h,\tau}^j$ défini dans (85).

Lorsque $\mathcal{M}_h$ est inversible, on peut réécrire le schéma sous la forme itérative classique

$$u^{j+1} = \underbrace{\mathcal{M}_h^{-1} \mathcal{N}_h}_{=: \mathcal{G}_{h,\tau}} u^j + \tau \underbrace{\mathcal{M}_h^{-1} b_{h,\tau}^j}_{=: \mathcal{B}_{h,\tau}^j}. \quad (90)$$

On travaille donc dans cette section sur le schéma itératif général

$$u^{j+1} = \mathcal{G}_{h,\tau} u^j + \tau \mathcal{B}_{h,\tau}^j, \quad j \in \llbracket 0, M \rrbracket. \quad (91)$$

Définition 3.18 (Stabilité). Le schéma (91) est dit stable pour une norme $\|\cdot\|$ sur $[0, T]$ s'il existe deux constantes $C_1, C_2 > 0$, indépendantes de h, τ et M , telles que

$$\max_{0 \leq j \leq M+1} \|u^j\| \leq C_1 \|u^0\| + C_2 \max_{0 \leq j \leq M} \|\mathcal{B}_{h,\tau}^j\|, \quad (92)$$

où u^j désigne le vecteur de la solution approchée au temps t_j .

Théorème 3.19 (Théorème de Lax–Richtmyer (1956)). Dans le cadre d'un problème linéaire bien posé et d'un schéma linéaire bien défini, le schéma itératif (91) est convergent si et seulement s'il est stable et consistant.

À retenir 3.20. Dans le cadre d'un problème linéaire bien posé et d'un schéma linéaire bien défini,

$$\text{"stabilité + consistance"} \iff \text{convergence"}$$

On admet l'implication

$$\text{convergence} \implies \text{stabilité + consistance},$$

et on prouve l'implication inverse

$$\text{stabilité + consistance} \implies \text{convergence},$$

la plus pratique à vérifier dans les applications numériques.

Preuve du Théorème 3.19 : implication $\Leftarrow$. Supposons que le schéma itératif (91) est stable pour une norme $\|\cdot\|$ donnée. D'après la Définition 3.18, cette stabilité s'exprime de la manière suivante : il existe des constantes $C_1, C_2 > 0$ pouvant dépendre de l'horizon temporel T , mais pas des pas de discrétisation τ et h , telles que

$$\max_{j \in \llbracket 0, M+1 \rrbracket} \|u^j\| \leq C_1 \|u^0\| + C_2 \max_{j \in \llbracket 0, M \rrbracket} \|\mathcal{B}_{h,\tau}^j\|, \quad (\text{stabilité}), \quad (93)$$

et ceci, pour toute donnée initiale u^0 et tout terme source $\mathcal{B}_{h,\tau}^j$.

La consistance, quant à elle, s'écrit

$$\max_{j \in \llbracket 0, M \rrbracket} \|\mathcal{E}_{h,\tau}^j(u)\| \rightarrow 0 \quad \text{lorsque} \quad h \rightarrow 0 \text{ et } \tau \rightarrow 0, \quad (\text{consistance}), \quad (94)$$

où l'erreur de consistance $\mathcal{E}_{h,\tau}^j(u)$ est définie à partir du schéma itératif (91) par

$$\mathcal{E}_{h,\tau}^j(u) := \frac{\pi_h u(t_{j+1}) - \mathcal{G}_{h,\tau} \pi_h u(t_j)}{\tau} - \mathcal{B}_{h,\tau}^j. \quad (95)$$

À présent qu'on a éclairci nos hypothèses de travail, on note l'erreur de convergence à l'instant t_j par

$$e_{h,\tau}^j := \pi_h u(t_j) - u^j. \quad (96)$$

On va maintenant utiliser la linéarité du schéma itératif (91) : notons qu'on a

- $u^{j+1} = \mathcal{G}_{h,\tau} u^j + \tau \mathcal{B}_{h,\tau}^j$ pour la solution numérique u^j , et
- $\pi_h u(t_{j+1}) = \mathcal{G}_{h,\tau} \pi_h u(t_j) + \tau \mathcal{B}_{h,\tau}^j + \tau \mathcal{E}_{h,\tau}^j(u)$ pour la solution projetée sur le maillage $\pi_h u(t_j)$.

En soustrayant la première ligne à la deuxième, on obtient immédiatement l'équation de l'erreur de convergence :

$$e_{h,\tau}^{j+1} = \mathcal{G}_{h,\tau} e_{h,\tau}^j + \tau \mathcal{E}_{h,\tau}^j(u). \quad (97)$$

Remarquons à ce point de la démonstration, en observant (97) que l'erreur de convergence $e_{h,\tau}^{j+1}$ satisfait un schéma itératif linéaire du même type que (91), où

- $\mathcal{G}_{h,\tau}$ est identique à celui du schéma itératif original (91),
- $\tau \mathcal{E}_{h,\tau}^j(u)$ joue le rôle du terme source $\tau \mathcal{B}_{h,\tau}^j$ dans le schéma itératif original (91).

On peut à présent utiliser notre résultat de stabilité (93) (valide pour tout terme source et toute donnée initiale) sur l'équation de l'erreur de convergence (97). Il vient

$$\max_{j \in \llbracket 0, M+1 \rrbracket} \|e_{h,\tau}^j\| \leq C_1 \underbrace{\|e_{h,\tau}^0\|}_{=0} + C_2 \max_{j \in \llbracket 0, M \rrbracket} \|\mathcal{E}_{h,\tau}^j(u)\|, \quad (\text{stabilité pour } e_{h,\tau}^j). \quad (98)$$

On remarquera que l'erreur initiale $e_{h,\tau}^0$ est nulle si la donnée initiale numérique coïncide avec la projection de la donnée initiale continue, c'est-à-dire si $u_h^0 = \pi_h u^0$. C'est ce que l'on fait en pratique : on choisit initialement $u_h^0 = \pi_h u_0$, ce qui annule l'erreur initiale.

Ainsi, on peut éliminer la contribution de l'erreur initiale nulle dans (98), et la consistance du schéma nous permet de conclure :

$$\max_{j \in \llbracket 0, M+1 \rrbracket} \|e_{h,\tau}^j\| \leq C_2 \max_{j \in \llbracket 0, M \rrbracket} \|\mathcal{E}_{h,\tau}^j(u)\| \rightarrow 0, \quad \text{lorsque } h \rightarrow 0 \text{ et } \tau \rightarrow 0, \quad (99)$$

□

Remarque 3.21. Il convient de noter que l'on a utilisé dans cette preuve la consistance au sens où l'erreur de consistance $\mathcal{E}_{h,\tau}^j(u)$ tend vers zéro lorsque $h \rightarrow 0$ et $\tau \rightarrow 0$ —cf. (94).

Si l'on savait a priori que l'erreur de consistance était d'ordre $\mathcal{O}(h^p + \tau^q)$ pour certains $p, q > 0$, alors on pourrait en déduire une estimation plus précise de l'erreur de convergence (seule la dernière ligne de la preuve (99) serait modifiée) :

$$\max_{j \in \llbracket 0, M+1 \rrbracket} \|e_{h,\tau}^j\| \leq C_2 \mathcal{O}(h^p + \tau^q). \quad (100)$$

On voit alors que l'erreur de convergence hérite des ordres de l'erreur de consistance ; autrement dit,

$$[\text{erreur de consistance d'ordre } \mathcal{O}(h^p + \tau^q)] \implies [\text{erreur de convergence d'ordre } \mathcal{O}(h^p + \tau^q)].$$

Proposition 3.22 (Conditions nécessaires et suffisantes pour la stabilité). Un schéma numérique de la forme générale (91) est stable pour une norme $\|\cdot\|$ donnée (cf. (93)) si et seulement s'il existe une constante $C > 0$ indépendante de h et τ telle que

$$\max_{j \in \llbracket 0, M+1 \rrbracket} \|(\mathcal{G}_{h,\tau})^j\| \leq C, \quad (101)$$

où $\|\cdot\|$ désigne la norme subordonnée à la norme $\|\cdot\|$ sur les opérateurs linéaires ($\mathcal{G}_{h,\tau}$ est une matrice !). Voir l'aparté 2.22.

Remarque 3.23. Il convient de noter que la matrice $\mathcal{G}_{h,\tau}$ dépend en général des deux pas de discrétisation h et τ . Dans (101), la notation $(\mathcal{G}_{h,\tau})^j$ désigne la matrice $\mathcal{G}_{h,\tau}$ élevée à la puissance j .

Preuve de la Proposition 3.22 (stabilité $\implies$ (101)). Supposons que le schéma général (91) soit stable, c'est-à-dire (93) pour toute donnée initiale u^0 et tout terme source $\mathcal{B}_{h,\tau}^j$: il existe $C_1, C_2 > 0$ tels que

$$\max_{j \in \llbracket 0, M+1 \rrbracket} \|u^j\| \leq C_1 \|u^0\| + C_2 \max_{j \in \llbracket 0, M \rrbracket} \|\mathcal{B}_{h,\tau}^j\|, \quad (102)$$

et ceci, pour toute donnée initiale u^0 et tout terme source $\mathcal{B}_{h,\tau}^j$.

Lorsque le terme source $\mathcal{B}_{h,\tau}^j$ vaut $0_{\mathbb{R}^N}$, le schéma itératif (91) se réduit à

$$u^{j+1} = \mathcal{G}_{h,\tau} u^j, \quad \forall j \in \llbracket 0, M \rrbracket, \quad (103)$$

et donc

$$u^j = (\mathcal{G}_{h,\tau})^j u^0, \quad \forall j \in \llbracket 0, M+1 \rrbracket. \quad (104)$$

Noter dans l'équation précédente que le j du second membre est une puissance de la matrice $\mathcal{G}_{h,\tau}$.

En utilisant maintenant l'inégalité de stabilité (102) pour le terme source nul, on obtient

$$\max_{j \in \llbracket 0, M+1 \rrbracket} \|u^j\| \leq C_1 \|u^0\|, \quad \forall u^0 \in \mathbb{R}^N,$$

et donc, en particulier,

$$\|u^j\| \leq C_1 \|u^0\|, \quad \forall j \in \llbracket 0, M+1 \rrbracket, \quad \forall u^0 \in \mathbb{R}^N.$$

En remplaçant u^j par $(\mathcal{G}_{h,\tau})^j u^0$ d'après (104), on obtient

$$\|(\mathcal{G}_{h,\tau})^j u^0\| \leq C_1 \|u^0\|, \quad \forall j \in \llbracket 0, M+1 \rrbracket, \quad \forall u^0 \in \mathbb{R}^N,$$

ce qui montre, en passant au suprémum sur $u^0 \in \mathbb{R}^N \setminus \{0\}$, que

$$\|(\mathcal{G}_{h,\tau})^j\| \stackrel{\text{d\'ef}}{=} \sup_{u^0 \in \mathbb{R}^N \setminus \{0\}} \frac{\|(\mathcal{G}_{h,\tau})^j u^0\|}{\|u^0\|} \leq C_1, \quad \forall j \in \llbracket 0, M+1 \rrbracket. \quad (105)$$

L'inégalité (101) apparaît en passant au maximum dans (105) sur $j \in \llbracket 0, M+1 \rrbracket$. $\square$

Preuve de la Proposition 3.22 ((101) $\Rightarrow$ stabilité). Réciproquement, si l'on suppose que l'on a (101), alors on a (on lâche les indices un instant pour simplifier la lecture...)

$$u^1 = \mathcal{G}u^0 + \tau \mathcal{B}^0.$$

$$\begin{aligned} u^2 &= \mathcal{G}u^1 + \tau \mathcal{B}^1 \\ &= \mathcal{G}(\mathcal{G}u^0 + \tau \mathcal{B}^0) + \tau \mathcal{B}^1 \\ &= (\mathcal{G})^2 u^0 + \tau(\mathcal{G}\mathcal{B}^0 + \mathcal{B}^1). \end{aligned}$$

$$\begin{aligned} u^3 &= \mathcal{G}u^2 + \tau \mathcal{B}^2 \\ &= \mathcal{G}\left((\mathcal{G})^2 u^0 + \tau(\mathcal{G}\mathcal{B}^0 + \mathcal{B}^1)\right) + \tau \mathcal{B}^2 \\ &= (\mathcal{G})^3 u^0 + \tau(\mathcal{G}^2 \mathcal{B}^0 + \mathcal{G}\mathcal{B}^1 + \mathcal{B}^2). \end{aligned}$$

$$\begin{aligned} u^4 &= \mathcal{G}u^3 + \tau \mathcal{B}^3 \\ &= \mathcal{G}\left((\mathcal{G})^3 u^0 + \tau(\mathcal{G}^2 \mathcal{B}^0 + \mathcal{G}\mathcal{B}^1 + \mathcal{B}^2)\right) + \tau \mathcal{B}^3 \\ &= (\mathcal{G})^4 u^0 + \tau(\mathcal{G}^3 \mathcal{B}^0 + \mathcal{G}^2 \mathcal{B}^1 + \mathcal{G}\mathcal{B}^2 + \mathcal{B}^3). \end{aligned}$$

Par une récurrence évidente, on obtient pour tout $j \in \llbracket 1, M+1 \rrbracket$,

$$u^j = (\mathcal{G})^j u^0 + \tau \sum_{k=0}^{j-1} (\mathcal{G})^{j-1-k} \mathcal{B}^k.$$

En passant à la norme, on obtient, par inégalité triangulaire, puis sous-multiplicativité de la norme (cf. (34)), et enfin l'hypothèse de contrôle sur les puissances de $\mathcal{G}$ (101) :

$$\|u^j\| \stackrel{\text{i.t.}}{\leq} \|(\mathcal{G})^j u^0\| + \tau \sum_{k=0}^{j-1} \|(\mathcal{G})^{j-1-k} \mathcal{B}^k\|$$

$$\begin{aligned}
&\stackrel{(34)}{\leq} \left\| \mathcal{G}^j \right\| \|u^0\| + \tau \sum_{k=0}^{j-1} \left\| (\mathcal{G})^{j-1-k} \right\| \|\mathcal{B}^k\| \\
&\stackrel{(101)}{\leq} C \|u^0\| + \tau \sum_{k=0}^{j-1} C \|\mathcal{B}^k\| \\
&\leq C \left(\|u^0\| + \tau \sum_{k=0}^M \|\mathcal{B}^k\| \right).
\end{aligned}$$

On s'approche peu à peu de l'estimation de stabilité (102). Pour avancer, on majore la somme par $M + 1$ fois le maximum des termes :

$$\tau \sum_{k=0}^M \|\mathcal{B}^k\| \leq \tau(M+1) \max_{k \in \llbracket 0, M \rrbracket} \|\mathcal{B}^k\|.$$

Puis, en se rappelant que, par définition, $\tau(M+1) = T$, il vient

$$\|u^j\| \leq C \left(\|u^0\| + T \max_{k \in \llbracket 0, M \rrbracket} \|\mathcal{B}^k\| \right),$$

ce qui recouvre l'estimation de stabilité (102), à savoir

$$\max_{j \in \llbracket 0, M+1 \rrbracket} \|u^j\| \leq C_1 \|u^0\| + C_2 \max_{k \in \llbracket 0, M \rrbracket} \|\mathcal{B}_{h,\tau}^k\|,$$

avec $C_1 = C$, et $C_2 = CT$. □

Remarque 3.24. *Le comportement des puissances de la matrice itérative $\mathcal{G}_h$ dépend de son rayon spectral*

$$\rho(\mathcal{G}_h) = \max_{\lambda \in \text{Sp}(\mathcal{G}_h)} |\lambda|,$$

où $\text{Sp}(\mathcal{G}_h)$ est appelé le spectre de $\mathcal{G}_h$ et désigne l'ensemble de ses valeurs propres.

On sait en particulier que

$$\lim_{j \rightarrow \infty} (\mathcal{G}_h)^j = 0_{\mathbb{R}^N \times N} \iff \rho(\mathcal{G}_h) < 1.$$

On va dans la suite montrer que la condition de stabilité sur les puissances de la matrice itérative $\mathcal{G}_h$ (101) est équivalente à la condition sur son rayon spectral, à savoir $\rho(\mathcal{G}_h) \leq 1$ pour certaines matrices $\mathcal{G}_h$ avec certaines normes $\|\cdot\|$.

Choix de la norme sur $\mathbb{R}^N$. Depuis le début du cours, on a notamment utilisé la norme $\|\cdot\|_\infty$ sur $\mathbb{R}^N$, définie par

$$\|u_h\|_\infty \stackrel{\text{déf}}{=} \max_{i \in \llbracket 1, N \rrbracket} |u_i|.$$

Cette norme permet de contrôler *uniformément* les composantes d'un vecteur : une estimation sur $\|u_h\|_\infty$ fournit, en particulier, une estimation valable simultanément pour chacune de ses composantes u_i . Dans le contexte d'une discrétisation spatiale, elle mesure ainsi l'erreur maximale commise sur l'ensemble des points du maillage.

En particulier, si une famille $(u_h)_h$ vérifie

$$\|u_h\|_\infty \leq Ch^p,$$

alors

$$|(u_h)_i| \leq Ch^p, \quad \forall i \in \llbracket 1, N \rrbracket,$$

c'est-à-dire que toutes les composantes convergent vers 0 *uniformément* par rapport à l'indice du maillage.

Au vu de la [remarque 3.24](#), on souhaite toutefois introduire une norme mieux adaptée aux “manipulations spectrales”, c'est-à-dire permettant d'exploiter plus directement les propriétés des valeurs propres des matrices intervenant dans les schémas. La norme $\|\cdot\|_\infty$ n'est pas la plus commode de ce point de vue.

Dans la suite, pour tout $h > 0$, nous allons donc utiliser la norme $\|\cdot\|_h$ sur $\mathbb{R}^N$, définie par

$$\|u_h\|_h := \left(h \sum_{i=1}^N |u_i|^2 \right)^{1/2} = \sqrt{h} \|u_h\|_2. \quad (106)$$

Le facteur $\sqrt{h}$ permet d'interpréter cette norme comme l'analogue discret de la norme $L^2(0, 1)$.

En effet, considérons un maillage uniforme de $[0, 1]$,

$$x_i = ih, \quad i \in \llbracket 1, N \rrbracket, \quad h = \frac{1}{N+1},$$

et supposons que les composantes de u_h soient données par

$$(u_h)_i = u(x_i), \quad i \in \llbracket 1, N \rrbracket,$$

pour une fonction u suffisamment régulière. Lorsque $h \rightarrow 0$, la somme de Riemann associée donne

$$h \sum_{i=1}^N |u(x_i)|^2 \xrightarrow{h \rightarrow 0} \int_0^1 |u(x)|^2 dx.$$

Ainsi, pour h au voisinage de 0, on a

$$\|u_h\|_h^2 = h \sum_{i=1}^N |(u_h)_i|^2 \approx \|u\|_{L^2(0,1)}^2,$$

et donc

$$\|u_h\|_h \approx \|u\|_{L^2(0,1)}.$$

Pour se convaincre d'utiliser la norme $\|\cdot\|_h$ plutôt que $\|\cdot\|_2$, supposons que pour tout pas de discrétisation $h > 0$, on ait

$$u_i = 1, \quad i \in \llbracket 1, N \rrbracket.$$

Dans ce cas, on aimerait que la famille $(u_h)_h$ soit associée à la fonction $u \equiv 1$ sur $[0, 1]$. Calculons :

$$\begin{aligned} \|u_h\|_h &= \left(h \sum_{i=1}^N |u_i|^2 \right)^{1/2} = \left(h \sum_{i=1}^N 1 \right)^{1/2} = (hN)^{1/2} = \left(\frac{N}{N+1} \right)^{1/2} \rightarrow 1. \\ \|u_h\|_2 &= \left(\sum_{i=1}^N |u_i|^2 \right)^{1/2} = \left(\sum_{i=1}^N 1 \right)^{1/2} = \sqrt{N} \rightarrow \infty. \end{aligned}$$

Bien que les normes $\|\cdot\|_h$ et $\|\cdot\|_\infty$ soient équivalentes **pour tout h fixé**, on peut perdre cette équivalence lorsque $h \rightarrow 0$ —cf. le [point de vigilance 2.13](#). On laissera le lecteur vérifier qu'on a les inégalités d'équivalence suivantes :

$$\sqrt{h} \|u_h\|_\infty \leq \|u_h\|_h \leq \|u_h\|_\infty.$$

En conséquence, la norme $\|\cdot\|_\infty$ “voit mieux la convergence” que la norme $\|\cdot\|_h$:

$$\begin{aligned} \left[\|u_h\|_\infty \leq Ch^p \right] &\implies \left[\|u_h\|_h \leq Ch^p \right], \\ \left[\|u_h\|_h \leq Ch^p \right] &\implies \left[\|u_h\|_\infty \leq Ch^{p-1/2} \right]. \end{aligned}$$

Autrement dit

$$\begin{aligned} \left[\text{ordre } p \text{ pour } \|u_h\|_\infty \right] &\implies \left[\text{ordre } p \text{ pour } \|u_h\|_h \right], \\ \left[\text{ordre } p \text{ pour } \|u_h\|_h \right] &\implies \left[\text{ordre } p - 1/2 \text{ pour } \|u_h\|_\infty \right]. \end{aligned}$$

La raison sous-jacente de ce point est que la norme $\|\cdot\|_h$ est une moyenne pondérée par h des valeurs absolues au carré des composantes, tandis que la norme $\|\cdot\|_\infty$ ne retient que la plus grande composante. Ainsi, une perturbation locale (sur une marginale par exemple) aura un impact direct sur la norme $\|\cdot\|_\infty$, mais sera atténuée par un facteur $\sqrt{h}$ dans la norme $\|\cdot\|_h$. Imaginez par exemple que l’on ait

$$u_h = (1, 0, 0, \dots, 0, 0)^T = e_1 \in \mathbb{R}^N, \quad \text{pour tout } h.$$

Alors

$$\|u_h\|_\infty = 1, \quad \|u_h\|_h = \sqrt{h}, \quad \text{pour tout } h,$$

et donc $\|u_h\|_\infty \rightarrow 1$, tandis que $\|u_h\|_h \rightarrow 0$ lorsque $h \rightarrow 0$.

Proposition 3.25 (Norme matricielle subordonnée à la norme $\|\cdot\|_h$). *Pour toute matrice $M \in \mathbb{R}^{N \times N}$, la norme subordonnée à la norme $\|\cdot\|_h$ est donnée par*

$$\|M\|_h = \|M\|_2 = \sqrt{\rho(M^*M)}, \quad (107)$$

où $\rho(\cdot)$ désigne le rayon spectral d’une matrice (i.e. le plus grand module des valeurs propres)

$$\rho(A) := \max\{|\text{Sp}(A)|\} = \max\{|\lambda| \text{ tels que } \lambda \in \text{Sp}(A) \subset \mathbb{C}\},$$

et M^* est l’adjointe de M : $M^* = \overline{M}^T$.

Preuve de la Proposition 3.25. La première égalité $\|M\|_h = \|M\|_2$ découle du fait que

$$\frac{\|Mx\|_h}{\|x\|_h} = \frac{\sqrt{h}\|Mx\|_2}{\sqrt{h}\|x\|_2} = \frac{\|Mx\|_2}{\|x\|_2},$$

et donc

$$\|M\|_h = \sup_{x \neq 0} \frac{\|Mx\|_h}{\|x\|_h} = \sup_{x \neq 0} \frac{\|Mx\|_2}{\|x\|_2} = \|M\|_2.$$

La seconde partie de (107) est plus difficile à obtenir. Notons d’abord que l’on peut écrire

$$\begin{aligned} \|M\|_2 &= \sup_{x \neq 0} \frac{\|Mx\|_2}{\|x\|_2} \\ &= \sup_{x \neq 0} \left\| \frac{1}{\|x\|_2} Mx \right\|_2 \\ &= \sup_{\|y\|_2=1} \|My\|_2 \end{aligned}$$

$$\begin{aligned}
&= \sup_{\|y\|_2=1} \sqrt{(My, My)} \\
&= \sqrt{\sup_{\|y\|_2=1} (My, My)},
\end{aligned}$$

par croissance de $x \mapsto \sqrt{x}$. En élevant le tout au carré, on obtient

$$\begin{aligned}
\|M\|_2^2 &= \sup_{\|y\|_2=1} (My, My) \\
&= \sup_{\|y\|_2=1} (M^*My, y).
\end{aligned}$$

Remarquons à présent que la matrice M^*M est hermitienne ($M^*M = (M^*M)^*$) et positive semi-définie :

$$(M^*My, y) = (My, My) = \|My\|_2^2 \geq 0, \quad \forall y \in \mathbb{C}^N.$$

Elle est définie positive si et seulement si M est injective. Par le théorème spectral, toutes les valeurs propres de M^*M sont donc réelles et positives ou nulles, et il existe une base orthonormée de vecteurs propres $\mathcal{B} = (e_i)_{i \in \llbracket 1, N \rrbracket}$, une matrice diagonale $D = \text{diag}(\lambda_1, \dots, \lambda_N)$, et une matrice unitaire U (i.e. $U^*U = I$) telle que

$$M^*M = UDU^*.$$

Pour $\|y\|_2 = 1$, on écrit

$$y = \sum_{i=1}^N \alpha_i e_i.$$

Par orthonormalité de la base $\mathcal{B}$, on a (Pythagore) :

$$1 = \|y\|_2^2 = \sum_{i=1}^N \|\alpha_i e_i\|_2^2 = \sum_{i=1}^N |\alpha_i|^2 = 1. \quad (108)$$

Maintenant, comme on a

$$M^*Me_i = \lambda_i e_i, \quad \forall i \in \llbracket 1, N \rrbracket,$$

on a

$$M^*My = M^*M \sum_{i=1}^N \alpha_i e_i = \sum_{i=1}^N \alpha_i M^*Me_i = \sum_{i=1}^N \alpha_i \lambda_i e_i.$$

Par conséquent,

$$(M^*My, y) = \left(\sum_{i=1}^N \alpha_i \lambda_i e_i, \sum_{j=1}^N \alpha_j e_j \right) = \sum_{i=1}^N \sum_{j=1}^N \alpha_i \lambda_i \overline{\alpha_j} \underbrace{(e_i, e_j)}_{=\delta_{ij}} = \sum_{i=1}^N \lambda_i |\alpha_i|^2. \quad (109)$$

Ainsi, à la vue de (109), on peut voir que (M^*My, y) est une combinaison convexe des valeurs propres λ_i , avec les coefficients $|\alpha_i|^2$ qui sont positifs et dont la somme vaut 1 d'après (108). Il s'ensuit que

$$\min_i \lambda_i \leq (M^*My, y) \leq \max_i \lambda_i,$$

avec égalité si et seulement si y est un vecteur propre correspondant à la valeur propre minimale ou maximale.

En somme, on trouve bien que

$$\|M\|_2 = \sqrt{\max_i \lambda_i} = \sqrt{\rho(M^*M)}.$$

□

Remarque 3.26 (Norme $\|\cdot\|_h$ d'une matrice normale). Pour une matrice normale M , i.e. si $M^*M = MM^*$, on a $\|M\|_h = \rho(M)$.

En effet, l'équivalence

$$[M \text{ normale}] \iff [\exists U \text{ unitaire}, \exists D \text{ diagonale telle que } M = UDU^*]$$

permet d'écrire

$$\begin{aligned} M^*M &= (UDU^*)^*(UDU^*) \\ &= (UD^*U^*)(UDU^*) \\ &= UD^*(U^*U)DU^* \\ &= UD^*DU^*, \end{aligned}$$

donc M^*M est semblable à D^*D , et

$$D^*D = \text{diag}(\overline{\lambda_1}, \dots, \overline{\lambda_N}) \text{diag}(\lambda_1, \dots, \lambda_N) = \text{diag}(|\lambda_1|^2, \dots, |\lambda_N|^2).$$

Ainsi,

$$\rho(M^*M) = \rho(D^*D) = \max_i |\lambda_i|^2,$$

et donc

$$\|M\|_h = \sqrt{\rho(M^*M)} = \max_i |\lambda_i| = \rho(M).$$

Rappel 3.27. Faisons un point sur les matrices :

- $M^* := \overline{M}^T$ est la matrice adjointe de M .
- on dit que M est auto-adjointe ou hermitienne si $M = M^*$.
- noter que si M est à valeurs réelles alors $[M \text{ hermitienne}] \iff [M \text{ symétrique}]$.
- on dit que M est normale si $M^*M = MM^*$.
- toute matrice hermitienne est normale.
- on dit que M est unitaire si $M^*M = I$, ce qui implique aussi que $MM^* = I$ et donc que M est normale. On retiendra les implications

$$[M \text{ hermitienne}] \implies [M \text{ normale}], \quad [M \text{ unitaire}] \implies [M \text{ normale}].$$

En général, il n'y a aucune implication entre les propriétés "hermitienne" et "unitaire".

Proposition 3.28 (Stabilité des schémas pour les matrices d'itération normales).

Si la matrice $\mathcal{G}_{h,\tau}$ est normale, alors un schéma numérique de la forme générale (91), à savoir

$$u^{j+1} = \mathcal{G}_{h,\tau} u^j + \tau \mathcal{B}_{h,\tau}^j,$$

est stable^a pour la norme $\|\cdot\|_h$ si et seulement s'il existe une constante $C_0 > 0$ (dépendant de l'horizon T mais indépendante de h et τ) telle que

$$\rho(\mathcal{G}_{h,\tau}) \leq 1 + C_0 \tau, \tag{110}$$

pour tout $h, \tau > 0$ admissibles.

^a. Voir la Définition 3.18 pour la définition de la stabilité.

Preuve de la Proposition 3.28. On commence par l'implication réciproque : (110) implique la stabilité. Pour cela, on va s'appuyer sur la caractérisation de la stabilité donnée par la Proposition 3.22.

Supposons qu'il existe une constante $C_0 > 0$ telle que (110) soit vérifiée, c'est-à-dire que

$$\rho(\mathcal{G}_{h,\tau}) \leq 1 + C_0\tau, \quad \forall h, \tau > 0 \text{ admissibles.}$$

Comme $\mathcal{G}_{h,\tau}$ est normale, ses puissances itérées le sont également. En effet, cela est trivial pour $j = 1$, et si $\mathcal{G}_{h,\tau}^j$ est normale, alors, en utilisant le fait que

$$(AB)^* = B^*A^*, \quad \forall A, B \in \mathbb{C}^{n \times n}. \quad (111)$$

et en simplifiant la notation $G = \mathcal{G}_{h,\tau}$,

$$\begin{aligned} (\mathcal{G}^{j+1})^* \mathcal{G}^{j+1} &= (\mathcal{G}^j \mathcal{G})^* (\mathcal{G}^j \mathcal{G}) \\ &= \mathcal{G}^* (\mathcal{G}^j)^* \mathcal{G}^j \mathcal{G} \quad (111) \\ &= \mathcal{G}^* \mathcal{G}^j (\mathcal{G}^j)^* \mathcal{G} \quad (\text{HR}) \\ &= \mathcal{G}^* \mathcal{G}^j (\mathcal{G}^*)^j \mathcal{G} \quad (111) \\ &= \mathcal{G} \mathcal{G}^* \mathcal{G}^{j-1} (\mathcal{G}^*)^{j-1} \mathcal{G} \mathcal{G}^* \quad (111) \quad (\text{HR pour } j = 1) \\ &= \mathcal{G}^2 \mathcal{G}^* \mathcal{G}^{j-2} (\mathcal{G}^*)^{j-2} \mathcal{G} (\mathcal{G}^*)^2 \quad (111) \quad (\text{HR pour } j = 1) \\ &= \dots \\ &= \mathcal{G}^j \mathcal{G}^* \mathcal{G} (\mathcal{G}^*)^j \quad (\text{HR pour } j = 1) \\ &= \mathcal{G}^j \mathcal{G} \mathcal{G}^* (\mathcal{G}^*)^j \quad (\text{HR pour } j = 1) \\ &= \mathcal{G}^j \mathcal{G} \mathcal{G}^* (\mathcal{G}^j)^* \quad (111) \\ &= \mathcal{G}^j \mathcal{G} (\mathcal{G}^j \mathcal{G})^* \quad (111) \\ &= \mathcal{G}^{j+1} (\mathcal{G}^{j+1})^*. \end{aligned}$$

Regardons maintenant la norme spectrale de $\mathcal{G}_{h,\tau}^j$. Puisque les puissances itérées de $\mathcal{G}_{h,\tau}$ sont normales, on a :

$$\begin{aligned} \|\mathcal{G}_{h,\tau}^j\|_h &= \rho(\mathcal{G}_{h,\tau}^j) \\ &= \max(|\lambda| : \lambda \in \text{Sp}(\mathcal{G}_{h,\tau}^j)) \\ &= \max(|\lambda^j| : \lambda \in \text{Sp}(\mathcal{G}_{h,\tau})) \\ &= \max(|\lambda|^j : \lambda \in \text{Sp}(\mathcal{G}_{h,\tau})) \\ &= (\max(|\lambda| : \lambda \in \text{Sp}(\mathcal{G}_{h,\tau})))^j \\ &= \rho(\mathcal{G}_{h,\tau})^j \\ &\leq (1 + C_0\tau)^j \quad \text{par l'hypothèse (110)} \\ &\leq (e^{C_0\tau})^j \quad \text{par l'inégalité } 1 + x \leq e^x \text{ pour tout } x \in \mathbb{R} \\ &= e^{C_0\tau j} \\ &= e^{C_0 t_j} \\ &\leq e^{C_0 T} =: C \quad (\text{celui de la Proposition 3.22!}). \end{aligned}$$

On a donc la stabilité.

Montrons maintenant l'implication réciproque, c'est-à-dire que la stabilité implique l'inégalité (110) sur le rayon spectral de la matrice d'itération $\mathcal{G}_{h,\tau}$. Supposons que le schéma soit stable, c'est-à-dire, à la vue de la Proposition 3.22, qu'il existe une constante C telle que

$$\|\mathcal{G}_{h,\tau}^j\|_h \leq C, \quad \text{pour tout } j \in \llbracket 1, M+1 \rrbracket.$$

En supposant que T est un multiple de τ , c'est-à-dire $T = (M+1)\tau$ (autrement on peut toujours ajuster T pour que ce soit le cas), on a $M+1 = T/\tau$, et $\|\mathcal{G}_{h,\tau}^{M+1}\|_h \leq C$.

Dans la fin de la preuve de l'implication réciproque, on a déjà montré que $\|\mathcal{G}_{h,\tau}^{M+1}\|_h = \rho(\mathcal{G}_{h,\tau})^{M+1}$. Par conséquent, on a

$$\rho(\mathcal{G}_{h,\tau})^{T/\tau} \leq C,$$

et donc

$$\rho(\mathcal{G}_{h,\tau}) \leq C^{\tau/T} =: g(\tau), \quad (112)$$

avec

$$g : \begin{cases} [0, T] \rightarrow \mathbb{R}, \\ t \mapsto C^{t/T} = \exp\left(\frac{t}{T} \ln C\right). \end{cases}$$

En calculant directement, on peut voir que $g'' > 0$ sur $[0, T]$ (donc g est convexe sur $[0, T]$), et que $g(0) = 1$ et $g(T) = C$. Par conséquent, par convexité, le graphe de g est situé sous la droite reliant les points $(0, 1)$ et (T, C) , autrement dit, on a

$$g(t) \leq 1 + \frac{C-1}{T}t, \quad \text{pour tout } t \in [0, T].$$

En particulier, pour $t = \tau$, on obtient

$$g(\tau) \leq 1 + \underbrace{\frac{C-1}{T}}_{=: C_0 > 0} \tau. \quad (113)$$

Noter qu'on peut toujours prendre $C > 1$, quitte à redéfinir $C_{\text{new}} = \max(C, 1)$ si nécessaire.

Ainsi, en réunissant (112) et (113), on obtient la majoration du rayon spectral de $\mathcal{G}_{h,\tau}$ suivante :

$$\rho(\mathcal{G}_{h,\tau}) \leq g(\tau) \leq 1 + C_0\tau,$$

c'est-à-dire, exactement (110). □

Corollaire 3.29 (Critère spectral suffisant de stabilité). *En conséquence de la Proposition 3.28,*

$$[\rho(\mathcal{G}_{h,\tau}) \leq 1] \implies [\text{le schéma itératif (91) avec } \mathcal{G}_{h,\tau} \text{ normale est stable pour la norme } \|\cdot\|_h].$$

Le corollaire précédent fournit une condition suffisante particulièrement commode :

$$\rho(\mathcal{G}_{h,\tau}) \leq 1 \implies \text{stabilité pour la norme } \|\cdot\|_h.$$

Dans les applications qui suivent, cette condition suffira pour conclure. On prendra garde à ne pas la présenter comme une condition nécessaire en général : le critère nécessaire et suffisant reste celui de la Proposition 3.28, à savoir

$$\rho(\mathcal{G}_{h,\tau}) \leq 1 + C_0\tau.$$

Lemme 3.30 (Éléments propres du moins-Laplacien discret). *Les éléments propres $(\lambda_k, v_k)_{k=1}^N$ de la matrice du (moins) Laplacien A_h , définie par (23) sont donnés, pour leur valeurs propres $\lambda_k \in \mathbb{R}_+^*$, par*

$$\lambda_k = \frac{4}{h^2} \sin^2 \left(\frac{k\pi h}{2} \right),$$

et pour leurs vecteurs propres $v_k \in \mathbb{R}^N$, par

$$v_{k,\ell} = \sqrt{\frac{2}{N+1}} \sin(k\pi h \ell) = \sqrt{2h} \sin(k\pi h \ell), \quad \text{pour tout } \ell \in \llbracket 1, N \rrbracket.$$

Remarque 3.31. Ce lemme donne une preuve directe du caractère défini positif de la matrice A_h dont on a donné une preuve plus algébrique dans le premier chapitre sur l'équation de Poisson—cf. la Proposition 2.11.

La preuve du Lemme 3.30 consiste simplement à vérifier que les vecteurs propres v_k satisfont l'équation aux valeurs propres $A_h v_k = \lambda_k v_k$ pour les valeurs propres λ_k données. Ce point est laissé en exercice à la lectrice ou au lecteur.

Application au schéma explicite (69). Rappelons le schéma explicite pour l'équation de la chaleur (69) :

$$u^{j+1} = u^j - \tau A_h u^j + \tau b_h^j, \quad \text{pour tout } j \in \llbracket 0, M \rrbracket, \quad (114)$$

équivalent à

$$u^{j+1} = \underbrace{(I - \tau A_h)}_{=: \mathcal{G}_{h,\tau}} u^j + \tau \underbrace{b_h^j}_{=: \mathcal{B}_{h,\tau}^j}, \quad \text{pour tout } j \in \llbracket 0, M \rrbracket, \quad (115)$$

à imaginer sous la forme du schéma itératif général (91), à savoir

$$u^{j+1} = \mathcal{G}_{h,\tau} u^j + \tau \mathcal{B}_{h,\tau}^j.$$

Ici la matrice d'itération $\mathcal{G}_{h,\tau}$ est donnée par

$$\mathcal{G}_{h,\tau} = I - \tau A_h = \begin{pmatrix} 1 - 2\frac{\tau}{h^2} & \frac{\tau}{h^2} & 0 & 0 & \cdots & 0 \\ \frac{\tau}{h^2} & 1 - 2\frac{\tau}{h^2} & \frac{\tau}{h^2} & 0 & \cdots & 0 \\ 0 & \frac{\tau}{h^2} & 1 - 2\frac{\tau}{h^2} & \frac{\tau}{h^2} & \cdots & 0 \\ 0 & 0 & \frac{\tau}{h^2} & 1 - 2\frac{\tau}{h^2} & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \frac{\tau}{h^2} \\ 0 & 0 & 0 & \cdots & \frac{\tau}{h^2} & 1 - 2\frac{\tau}{h^2} \end{pmatrix}$$

est réelle et symétrique, donc normale—cf. les cinq premiers points de Rappel 3.27. On est donc dans le bon cadre pour appliquer les résultats établis plus haut.

Comme $\mathcal{G}_{h,\tau} = I - \tau A_h$, cette matrice admet les mêmes vecteurs propres que A_h , et ses valeurs propres sont

$$\mu_k = 1 - \tau \lambda_k, \quad \text{pour tout } k \in \llbracket 1, N \rrbracket, \quad (116)$$

où les λ_k sont les valeurs propres de A_h données dans le Lemme 3.30.

D'après la Proposition 3.28, une condition suffisante commode de stabilité est

$$\rho(\mathcal{G}_{h,\tau}) \leq 1.$$

Comme $\mathcal{G}_{h,\tau}$ est symétrique, cette condition équivaut à

$$\begin{aligned}\rho(\mathcal{G}_{h,\tau}) \leq 1 &\iff |\mu_k| \leq 1, \quad \text{pour tout } k \in \llbracket 1, N \rrbracket \\ &\iff -1 \leq 1 - \tau \lambda_k \leq 1, \quad \text{pour tout } k \in \llbracket 1, N \rrbracket \\ &\iff 0 \leq \tau \lambda_k \leq 2, \quad \text{pour tout } k \in \llbracket 1, N \rrbracket \\ &\iff 0 \leq \tau \leq \frac{2}{\lambda_{\max}(A_h)}.\end{aligned}$$

Or

$$\lambda_{\max}(A_h) = \lambda_N = \frac{4}{h^2} \cos^2\left(\frac{\pi h}{2}\right).$$

La condition spectrale exacte $\rho(\mathcal{G}_{h,\tau}) \leq 1$ s'écrit donc

$$0 \leq \tau \leq \frac{h^2}{2 \cos^2\left(\frac{\pi h}{2}\right)}.$$

En particulier, la condition CFL plus simple

$$\boxed{\tau \leq \frac{h^2}{2}}$$

est suffisante pour tout N et redonne la condition obtenue précédemment en norme $\|\cdot\|_\infty$.

Application au schéma implicite (84). Rappelons le schéma implicite pour l'équation de la chaleur (84) :

$$u^{j+1} = u^j - \tau A_h u^{j+1} + \tau b_h^{j+1}, \quad \text{pour tout } j \in \llbracket 0, M \rrbracket, \quad (117)$$

équivalent à

$$u^{j+1} = \underbrace{(I + \tau A_h)^{-1}}_{=: \mathcal{G}_{h,\tau}} u^j + \tau \underbrace{(I + \tau A_h)^{-1} b_h^{j+1}}_{=: \mathcal{B}_{h,\tau}^j}, \quad \text{pour tout } j \in \llbracket 0, M \rrbracket, \quad (118)$$

à imaginer sous la forme du schéma itératif général (91), à savoir

$$u^{j+1} = \mathcal{G}_{h,\tau} u^j + \tau \mathcal{B}_{h,\tau}^j.$$

Proposition 3.32 (Stabilité inconditionnelle du schéma implicite (118)). *Le schéma d'Euler implicite (118) pour l'équation de la chaleur avec conditions aux limites de Dirichlet est stable, pour tout pas d'espace $h \in (0, 1)$ et tout pas de temps $\tau \in (0, T)$.*

Preuve de la Proposition 3.32. On a déjà montré précédemment que la matrice $I + \tau A_h$ est réelle, symétrique et définie positive, donc inversible. Comme $A_h v_k = \lambda_k v_k$, on a

$$(I + \tau A_h) v_k = (1 + \tau \lambda_k) v_k.$$

Les valeurs propres de son inverse $\mathcal{G}_{h,\tau} = (I + \tau A_h)^{-1}$ sont donc

$$\eta_k = \frac{1}{1 + \tau \lambda_k} = \frac{1}{1 + 4 \frac{\tau}{h^2} \sin^2\left(\frac{k\pi}{2(N+1)}\right)}, \quad \text{pour tout } k \in \llbracket 1, N \rrbracket.$$

Comme $\lambda_k > 0$, on a

$$0 < \eta_k < 1, \quad \text{pour tout } k \in \llbracket 1, N \rrbracket.$$

Ainsi,

$$\rho(\mathcal{G}_{h,\tau}) < 1$$

sans aucune condition liant les pas h et τ . Le [Corollaire 3.29](#) permet donc de conclure à la stabilité, sans condition de type CFL : le schéma est inconditionnellement stable. $\square$

Remarque 3.33 (Consistance en norme $\|\cdot\|_h$). On pourra remarquer que l'on a

$$\begin{aligned} \|x\|_h^2 &\stackrel{\text{déf}}{=} h\|x\|_2^2 \\ &= h \sum_{i=1}^N x_i^2 \\ &\leq h \sum_{i=1}^N \|x\|_\infty^2 \\ &= hN\|x\|_\infty^2 \\ &= \frac{N}{N+1}\|x\|_\infty^2 \\ &\leq \|x\|_\infty^2. \end{aligned}$$

En conséquence,

$$\|x\|_h \leq \|x\|_\infty$$

pour tout vecteur $x \in \mathbb{R}^N$, et donc la consistance en norme $\|\cdot\|_\infty$ implique la consistance en norme $\|\cdot\|_h$, et l'ordre de consistance en norme $\|\cdot\|_h$ est au moins celui en norme $\|\cdot\|_\infty$ —cf. la [Proposition 3.17](#) sur la consistance en norme $\|\cdot\|_\infty$ du schéma implicite.

La précédente remarque aboutit à la proposition suivante, qui est une réplique quasiment identique de la [Proposition 3.17](#), mais pour la norme $\|\cdot\|_h$ cette fois.

Proposition 3.34 (Consistance du schéma implicite en norme $\|\cdot\|_h$). Si $u \in C^2([0, T]; C^4([0, 1]))$ est une solution de (54), alors le schéma (83) est consistant, d'ordre 1 en temps et d'ordre 2 en espace pour la norme $\|\cdot\|_h$.

Plus précisément, on a l'estimation

$$\max_{j \in \llbracket 0, M \rrbracket} \|\varepsilon_{h,\tau}^{j+1}(u)\|_h \leq \frac{h^2}{12} \sup_{[0,T] \times [0,1]} \left| \frac{\partial^4 u}{\partial x^4} \right| + \frac{\tau}{2} \sup_{[0,T] \times [0,1]} \left| \frac{\partial^2 u}{\partial t^2} \right|. \quad (119)$$

On peut maintenant appliquer le [Théorème de Lax-Richtmyer](#) :

Proposition 3.35 (Convergence du schéma implicite en norme $\|\cdot\|_h$). Si $u \in C^2([0, T]; C^4([0, 1]))$ est une solution de (54), alors le schéma (83) est convergent pour la norme $\|\cdot\|_h$, d'ordre 2 en espace et d'ordre 1 en temps :

$$\begin{aligned} \max_{j \in \llbracket 0, M+1 \rrbracket} \|u^j - \pi_h u(t_j)\|_h &\leq C \max_{j \in \llbracket 1, M+1 \rrbracket} \|\varepsilon_{h,\tau}^j(u)\|_h && (\text{stabilité}^3) \\ &\leq C \left(\frac{h^2}{12} \sup_{[0,T] \times [0,1]} \left| \frac{\partial^4 u}{\partial x^4} \right| + \frac{\tau}{2} \sup_{[0,T] \times [0,1]} \left| \frac{\partial^2 u}{\partial t^2} \right| \right) && (\text{consistance—cf. (119)}). \end{aligned}$$

³voir (98)-(99) dans la preuve du [Théorème de Lax-Richtmyer](#).

Qu'en est-il de la convergence en norme $\|\cdot\|_\infty$? On pourra remarquer que l'on a,

$$\begin{aligned}\|x\|_\infty^2 &\stackrel{\text{déf}}{=} \left(\max_{i \in \llbracket 1, N \rrbracket} |x_i| \right)^2 \\ &= \max_{i \in \llbracket 1, N \rrbracket} |x_i|^2 \\ &\leq \sum_{i=1}^N |x_i|^2 \\ &= \|x\|_2^2 \\ &= \frac{1}{h} \|x\|_h^2.\end{aligned}$$

Ainsi, on obtient l'inégalité

$$\|x\|_\infty \leq \frac{1}{\sqrt{h}} \|x\|_h, \quad (120)$$

pour tout vecteur $x \in \mathbb{R}^N$.

En conséquence, on trouve

$$\begin{aligned}\max_{j \in \llbracket 0, M+1 \rrbracket} \|u^j - \pi_h u(t_j)\|_\infty &\leq \frac{1}{\sqrt{h}} \max_{j \in \llbracket 0, M+1 \rrbracket} \|u^j - \pi_h u(t_j)\|_h \\ &\leq \frac{C}{\sqrt{h}} \left(\frac{h^2}{12} \sup_{[0,T] \times [0,1]} \left| \frac{\partial^4 u}{\partial x^4} \right| + \frac{\tau}{2} \sup_{[0,T] \times [0,1]} \left| \frac{\partial^2 u}{\partial t^2} \right| \right) \\ &= C \left(\frac{h^{3/2}}{12} \sup_{[0,T] \times [0,1]} \left| \frac{\partial^4 u}{\partial x^4} \right| + \frac{\tau}{2\sqrt{h}} \sup_{[0,T] \times [0,1]} \left| \frac{\partial^2 u}{\partial t^2} \right| \right).\end{aligned}$$

On voit notamment que la convergence en norme $\|\cdot\|_\infty$ est plus lente que celle en norme $\|\cdot\|_h$, avec un facteur $\frac{1}{\sqrt{h}}$. Pire encore : si h tend vers 0 plus vite que τ , le contrôle sur l'erreur en norme $\|\cdot\|_\infty$ peut devenir très mauvais.

Cependant, bien que ce contrôle en norme $\|\cdot\|_\infty$ soit *a priori* plus mauvais que celui en norme $\|\cdot\|_h$, la régularité des problèmes discrets et continus fait que les valeurs ponctuelles des u_i^j restent du même ordre de grandeur, de sorte que l'erreur en norme $\|\cdot\|_\infty$ se comporte finalement de manière analogue à celle en norme $\|\cdot\|_h$. Ce point sera illustré dans la [fiche de TP 2](#).

Remarque 3.36 (Constater rapidement un ordre de convergence numériquement). *Si, pour un schéma numérique, on a, formellement,*

$$\text{Erreur}(h, \tau) \approx Ch^2 + D\tau,$$

alors,

- pour τ fixé proche de 0, si h est divisé par 10, l'erreur est supposée être divisée par 100 (puisque l'erreur est d'ordre h^2);
- de même, pour h fixé proche de 0, si τ est divisé par 10, l'erreur est supposée être divisée par 10 (puisque l'erreur est d'ordre τ).

Pour mettre ce phénomène pleinement en évidence, on fixe proche de zéro un des deux paramètres (h ou τ) et on fait varier l'autre sur une plage de valeurs suffisamment large—eg. h fixé et $\tau \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$, ou τ fixé et $h \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$.

On peut alors tracer le graphe $\log(\text{Erreur})$ en fonction de $\log(h)$ (si on fait varier h) ou de $\log(\tau)$ (si on fait varier τ). En effet, on a :

$$\begin{aligned}\overbrace{\log(\text{Erreur})}^{=Y} &= \log(Ch^p + D\tau^q) \\ &\approx \log(Ch^p) \quad \text{si } \tau \text{ est fixé petit—continuité du logarithme} \\ &= \log(C) + p \underbrace{\log(h)}_{=X}.\end{aligned}$$

On trouve que le graphe $\log(\text{Erreur})$ en fonction de $\log(h)$ est approximativement une droite de pente p : $Y = \log(C) + pX$.

Il suffit alors de faire une régression linéaire sur cette droite pour estimer la pente p , ce qui permet de déterminer l'ordre de convergence par rapport à h .

4 Équation de transport

5 Équation des ondes

Liste de questions de cours

Cette liste rassemble les principales questions de cours à maîtriser pour le contrôle. Une question pourra être posée telle quelle, partiellement, ou sous une forme légèrement reformulée. **Les questions portant sur des points qui n'auront pas été abordés en cours seront comptabilisées comme points bonus.**

1. Équation de Poisson et différences finies

- P1.** Énoncer le problème de Poisson avec conditions de Dirichlet (11).
- P2.** Énoncer la Proposition 2.2 sur le caractère bien posé du problème de Poisson avec source continue par morceaux et conditions de Dirichlet. Retrouver la formule explicite de la solution en intégrant deux fois l'équation. **Démonstration à connaître.**
- P3.** À partir d'un maillage uniforme de $[0, 1]$, définir le pas h , les points x_i et le vecteur inconnu u_h . Retrouver par développements de Taylor l'approximation centrée de $u''(x_i)$, puis en déduire le schéma scalaire et le système matriciel

$$A_h u_h = b_h$$

approximant l'équation de Poisson avec conditions de Dirichlet non homogènes. Expliciter le traitement de la première et de la dernière lignes du système, et distinguer les objets u , $\pi_h(u)$ et u_h —voir la Section 2.2, jusqu'à (23).

- P4.** Énoncer et démontrer la Proposition 2.11 : la matrice A_h est symétrique définie positive, donc inversible. Il faut notamment savoir établir

$$x^T A_h x = \frac{1}{h^2} \left(x_1^2 + \sum_{i=1}^{N-1} (x_i - x_{i+1})^2 + x_N^2 \right).$$

Démonstration à connaître.

- P5.** Définir les erreurs de consistance et de convergence du schéma de Poisson, ainsi que les notions d'ordre de consistance et d'ordre de convergence—voir les Définition 2.14 et Définition 2.21.
- P6.** Énoncer et démontrer la Proposition 2.19. Retrouver, à l'aide de développements de Taylor–Lagrange à l'ordre quatre,

$$\|\varepsilon_h(u)\|_\infty \leq \frac{h^2}{12} \sup_{x \in [0,1]} |u^{(4)}(x)|.$$

Préciser dans quel cas le schéma est exactement consistant—voir la Remarque 2.20 pour ce point. **Démonstration à connaître.**

- P7.** Définir la norme matricielle subordonnée à une norme vectorielle et établir la relation

$$e_h(u) = A_h^{-1} \varepsilon_h(u).$$

Expliquer comment la stabilité du schéma et sa consistance permettent de contrôler l'erreur de convergence—voir l'Aparté 2.22 et (31)–(35).

- P8.** Énoncer et démontrer le principe du maximum discret (Lemme 2.25). En déduire que tous les coefficients de A_h^{-1} sont positifs ou nuls—voir (46)–(49) pour ce point. **Démonstration à connaître.**

2. Équation de la chaleur : problème continu et schémas

- C1.** Écrire intégralement le problème de la chaleur avec donnée initiale, conditions de Dirichlet non homogènes et terme source (54).
- C2.** Dans le cas homogène $f \equiv 0$ et $u_g \equiv u_d \equiv 0$, établir la décroissance de l'énergie

$$E(t) = \|u(t, \cdot)\|_{L^2(0,1)}^2.$$

En déduire l'unicité de la solution du problème de la chaleur en considérant la différence de deux solutions—voir la [Proposition 3.4](#), (2–3). **Démonstration à connaître.**

- C3.** Définir le maillage espace-temps et les indices correspondant aux points intérieurs, aux valeurs de bord, à la donnée initiale et aux itérations en temps. Retrouver les schémas d'Euler explicite et implicite, sous forme scalaire puis matricielle, et expliquer la différence entre un schéma explicite et un schéma implicite—voir la [Section 3.2](#).
- C4.** Énoncer et démontrer la [Proposition 3.11](#) sur la stabilité du schéma explicite. Faire apparaître la condition CFL

$$\frac{\tau}{h^2} \leq \frac{1}{2}$$

en écrivant le schéma sous la forme d'une combinaison convexe. Expliquer enfin quelle contrainte un raffinement $h \mapsto h/10$ impose sur le pas de temps τ . **Démonstration à connaître.**

- C5.** Donner les avantages et inconvénients respectifs des schémas d'Euler explicite et implicite, en termes de coût à chaque itération, de stabilité et de contraintes sur les pas h et τ —voir le [Point de vigilance 3.7](#), la [Proposition 3.11](#) et la [Proposition 3.32](#).
- C6.** Interpréter concrètement le fait que l'erreur de convergence pour la norme $\|\cdot\|_h$ soit d'ordre

$$\mathcal{O}(h^2 + \tau),$$

c'est-à-dire qu'elle satisfasse une estimation de la forme

$$\max_{0 \leq j \leq M+1} \|u^j - \pi_h u(t_j)\|_h \leq C(h^2 + \tau),$$

où C est indépendante de h et de τ .

Expliquer l'effet attendu sur l'erreur :

- lorsque h est divisé par 10, dans un régime où $\tau \ll h^2$;
- lorsque τ est divisé par 10, dans un régime où $h^2 \ll \tau$.

Expliquer enfin comment ces ordres peuvent être mis en évidence numériquement. Voir à partir de la [Proposition 3.35](#) jusqu'à la [Remarque 3.36](#).

3. Théorème de Lax–Richtmyer et applications

- S1.** Pour un schéma linéaire itératif de la forme

$$u^{j+1} = \mathcal{G}_{h,\tau} u^j + \tau \mathcal{B}_{h,\tau}^j,$$

énoncer la définition de la stabilité sur $[0, T]$. Préciser que les constantes de stabilité peuvent dépendre de T , mais doivent être indépendantes de h , de τ et du nombre d'itérations M —voir (91) et la [Définition 3.18](#).

S2. Énoncer le [Théorème 3.19 de Lax–Richtmyer](#), avec ses hypothèses. Démontrer l'implication utile

$$\text{stabilité} + \text{consistance} \implies \text{convergence}$$

en introduisant l'erreur de convergence et en écrivant le schéma itératif qu'elle satisfait.

Démonstration à connaître.

S3. Pour la norme discrète (106) :

$$\|v\|_h = \left(h \sum_{i=1}^N |v_i|^2 \right)^{1/2},$$

énoncer le critère spectral suffisant

$$\rho(\mathcal{G}_{h,\tau}) \leq 1$$

de stabilité pour le schéma itératif général (91) lorsque la matrice d'itération $\mathcal{G}_{h,\tau}$ est normale—cf. le [Corollaire 3.29](#). En admettant que les valeurs propres du moins-Laplacien Dirichlet discret sur $(0, 1)$ sont

$$\lambda_k = \frac{4}{h^2} \sin^2 \left(\frac{k\pi h}{2} \right) \quad (\text{seront données si cette question tombe}),$$

appliquer ce critère aux schémas d'Euler explicites et implicites pour l'équation de la chaleur afin de :

— retrouver la condition suffisante (CFL)

$$\frac{\tau}{h^2} \leq \frac{1}{2}$$

pour la stabilité du schéma explicite ;

— montrer que le schéma implicite est inconditionnellement stable.

Voir à partir du [Lemme 3.30](#), jusqu'à la [Proposition 3.32](#).