Introduction aux différences finies et aux éléments finis

François Dubois

Conservatoire National des Arts et Métiers
Paris

IDEF

Introduction aux Différences finies et aux Eléments Finis

François Dubois
Professeur des Universités
Mathématiques, 1995-96.
Dactylographie assurée par Pierrette Foulon

Conservatoire National des Arts et Métiers

Introduction aux Différences Finies et aux Éléments Finis

COURS 1995/1996

François DUBOIS

IDEF

Introduction aux Différences Finies et aux Éléments Finis

PLAN

I) Motivation : équations différentielles ordinaires et équations aux dérivées partielles

1) Quelques modèles conduisant à des équations différentielles ordinaires
   - Placement financier
   - Système masse-ressort
   - Système masse-ressort avec amortissement fluide
   - Circuit électrique RLC
   - Mouvement d'un satellite
   - Cinétique chimique
   - Météorologie : attracteur de Lorenz
   - Définition

2) Quelques modèles conduisant à des équations aux dérivées partielles
   - Préambule
   - Diffusion de la chaleur à une dimension d'espace
   - Résistance des matériaux : élasticité tridimensionnelle
   - Résistance des matériaux : flexion des poutres
   - Membrane vibrante (tambour)
   - Mécanique des fluides : équations de Navier-Stokes des fluides incompressibles
   - Mécanique des fluides : équations d'Euler de la dynamique des gaz
   - Equation des ondes acoustiques
   - Electromagnétisme : équations de Maxwell
   - Equation d'advection : diffusion à une dimension d'espace
   - Physique moléculaire : équation de Schrödinger

II) Différences finies pour les équations différentielles ordinaires

1) Schémas aux différences : première approche

2) Test des schémas d'Euler explicite, Euler implicite et Crank-Nicolson pour le modèle $\frac{du}{dt} + \lambda u = 0$ $(\lambda > 0)$.

   - Définition : méthode à un pas

3) Exemples de méthodes multipas

   - Schéma instable
   - Schéma d'Adams-Bashford

4) Schémas à un pas explicites précis

   - Schéma d'Euler modifié
   - Schéma de Heun
   - Schéma de Runge-Kutta

5) Ordre d'un schéma aux différences

6) Mémorandum

III) Différences finies pour l'équation d'advection à une dimension d'espace

1) Equation d'advection à une dimension d'espace.
2) Discrétisation en espace et en temps
3) Analyse de stabilité par Fourier
4) Quelques propriétés de deux schémas classiques
5) Problème à valeur initiale et à la limite

IV) Différences finies pour l'équation de la chaleur à une dimension d'espace

1) Modèle physico-mathématique
2) Schéma aux différences explicite à une dimension
3) θ–Schéma en temps
4) Schémas à trois niveaux en temps

V) Différences finies pour l'équation de Poisson à deux dimensions d'espace

1) Discrétisation par différences finies

2) Formation du système linéaire

3) Résolution du système linéaire

VI) Introduction à l'écriture variationnelle des problèmes elliptiques

1) Motivation

2) Problème de Dirichlet homogène pour l'équation de Poisson

3) Problème de Dirichlet non homogène pour l'équation de Poisson

4) Problème mixte pour l'équation de Poisson

5) Problème de Neumann pour l'équation de Poisson

VII) Introduction à la méthode des éléments finis

1) Introduction

2) Problème de Dirichlet à une dimension d'espace

3) Problème de Dirichlet à deux dimensions d'espace

VIII) Mise en oeuvre informatique de la méthode des éléments finis

1) Introduction

2) Mise en oeuvre d'un problème monodimensionnel

3) Mise en oeuvre de l'élément P1 dans $\mathbb{R}^2$

4) Condition de Dirichlet non homogène

5) Etide d'un exemple, dit "Hadhri"

IX) Compléments

(i) Formule d'intégration par parties

(ii) Algorithme du gradient conjugué

(iii) Etude du schéma de Newmark

X) Travaux pratiques

(i) Etude du schéma de Newmark

I) Calcul d'une intégrale par la méthode des trapèzes

II) Schéma de Newmark

III) Précision du schéma de Newmark

(ii) Régularisation d'une fonction par résolution d'un opérateur elliptique

I) Principe de la méthode numérique

II) Organisation du code de calcul (programme)

III) Précision du schéma de Newmark

I. MOTIVATION : Équations différentielles ordinaires et Équations aux dérivées partielles

1) Quelques modèles conduisant à des équations différentielles ordinaires

- **Placement financier**

Disposant d'une masse d'argent x_0 au temps t = 0, on la place à la banque à un taux r. Ceci signifie que la somme $x(t)$ disponible à l'instant t ultérieur est solution de l'équation différentielle suivante :

$$(1) \qquad \begin{cases} \dfrac{dx}{dt} = rx & (t > 0) \\ x(0) = x_0 & \text{(condition initiale)} \end{cases}$$

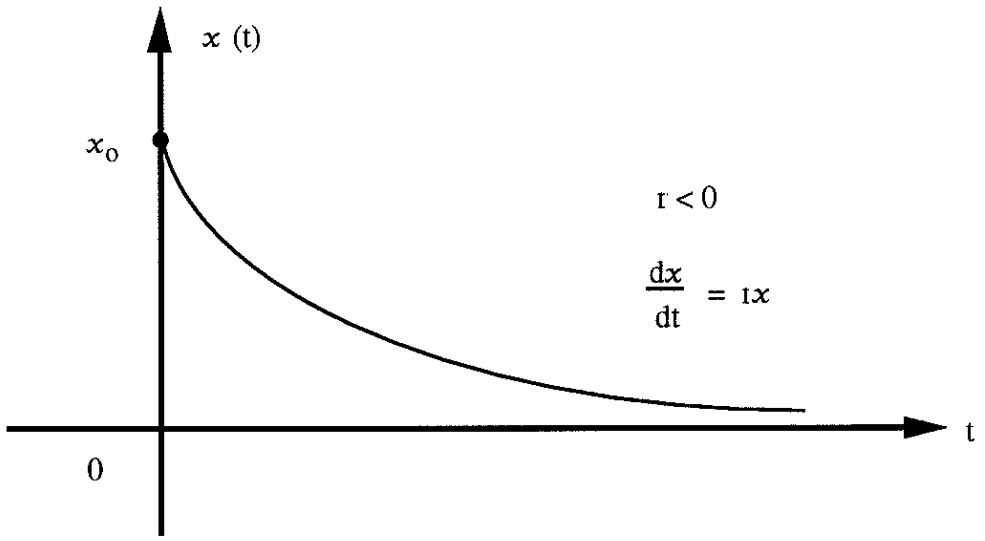
La solution du modèle (1) est bien connue, puisqu'elle fait intervenir la **fonction exponentielle** :

$$(2) \qquad x(t) = (\exp(rt))\ x_0$$

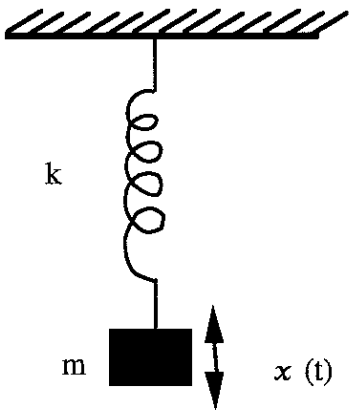
et il n'est aucunement besoin de méthode numérique pour calculer la valeur $x(t)$ donnée à la relation (2) ; une simple calculette suffit.

Toutefois, le modèle exponentiel est fondamental à plusieurs titres :

l'équation (1) est très simple, est linéaire [ce qui signifie que la fonction $x \to rx$ au second membre de la première équation de (1) est une fonction linéaire de la variable x], et pour $r < 0$ (ce qui constitue un cas purement mathématique de placement financier qu'aucun client d'une banque n'accepterait) est un bon modèle de système convergeant, pour t tendant vers $+\infty$, vers un **point fixe**, puisque $x(t)$ tend vers 0 si $t \to +\infty$ lorsque $r < 0$.



- **Système masse-ressort**



Il s'agit d'un modèle de mécanique tout à fait fondamental. Un ressort de raideur k est d'abord allongé pour assurer l'équilibre d'une masse m. A l'instant initial t = 0, on l'écarte de sa position d'équilibre d'une abscisse x_o avec une vitesse nulle (ou initiale v_o en toute généralité).

(3) $\quad x\,(o) = x_o \quad ; \quad \frac{dx}{dt}\,(o) = 0 \;\; [v_o]$.

L'évolution en temps écrit que la force de réaction du ressort [−kx] est exactement égale au produit de la masse par l'accélération $\left[m \frac{d^2x}{dt^2} \right]$:

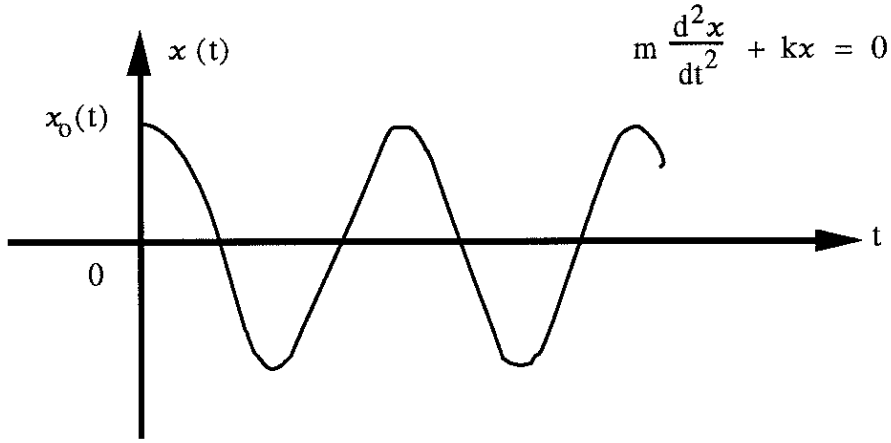
(4) $\quad m \frac{d^2x}{dt^2} + kx = 0 \quad (t > 0)$.

La solution de l'équation différentielle du second ordre (4) est bien connue et est de forme sinusoïdale :

(5) $\quad x\,(t) = A \sin(\omega_o t) + B \cos(\omega_o t)$

avec une pulsation ω_0 donnée par la relation classique :

(6) $$\omega_0 = \sqrt{\frac{k}{m}}$$



Cette fois, la variable x oscille indéfiniment entre deux valeurs sans s'amortir. Ce type de comportement porte le nom de **cycle limite**.

Si le système (3) (4) est présenté plus haut sous la forme d'une équation du second ordre avec deux conditions initiales, on peut très facilement se ramener à un système différentiel du premier ordre, mais relatif à un vecteur d'inconnues X(t) à deux composantes :

(7) $$X = \begin{pmatrix} x \\ y \end{pmatrix}$$

avec $y = \frac{dx}{dt}$. En dérivant y par rapport au temps, et en reportant l'expression obtenue à l'aide de la relation (4), on obtient :

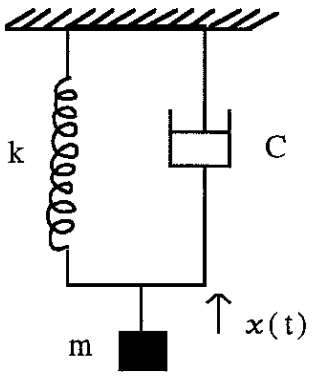
(8) $$\frac{dX}{dt} = \begin{pmatrix} 0 & 1 \\ -\frac{k}{m} & 0 \end{pmatrix} X \equiv \begin{pmatrix} y \\ -\frac{k}{m}x \end{pmatrix}$$

L'équation (8) d'inconnue le vecteur X est du premier ordre en temps, et la conditions initiale (3) peut s'écrire facilement sous la forme :

(9) $$X(o) = X_0 \equiv \begin{pmatrix} x_0 \\ v_0 \end{pmatrix}$$

On note également que la présence de la matrice au second membre de la relation (8) est caractéristique du caractère **linéaire** du modèle d'évolution.

- **Système masse-ressort avec amortissement fluide**



Si on complique un peu le système mécanique en adjoignant une force de frottement fluide $-C\frac{dx}{dt}$ à la force de rappel du ressort, l'équation d'évolution (4) devient dans ce cas :

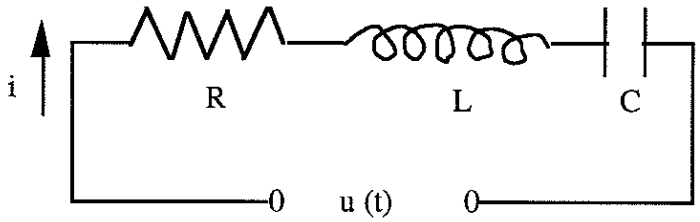
$$(10) \qquad m\frac{d^2x}{dt^2} + C\frac{dx}{dt} + kx = 0.$$

C'est encore un modèle linéaire, qui s'écrit, grâce à la variable X introduite à la relation (7).

$$(11) \qquad \frac{dX}{dt} = \begin{pmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{C}{m} \end{pmatrix} X .$$

Si $C > 0$, ce qui signifie physiquement que l'on a bien à faire à un amortisseur (et non un amplificateur si $C < 0$), alors X (t) tend vers 0 si t tend vers $+\infty$ et l'on retrouve encore dans ce cas un comportement de type **point fixe** [ie X(t) tend vers un état constant pour $t \to +\infty$].

- **Circuit électrique RLC**



Ce modèle électrique fondamental s'obtient en écrivant que la différence de potentiel u_1 aux bornes de la résistance R vérifie la loi d'Ohm.

$$(12) \qquad u_1 = Ri ,$$

la différence de potentiel u_2 aux bornes de la self L dérive du courant i.

$$(13) \qquad u_2 = L\frac{di}{dt} ,$$

la différence de potentiel u_3 aux bornes du condensateur C induit une charge q selon la relation :

(14) $\quad q = C\, u_3$

et la dérivée en temps de la charge électrique q est exactement égale au courant i :

(15) $\quad i = \frac{dq}{dt}$

De plus, la différence de potentiel u (t) est la source des trois ddp précédentes :

(16) $\quad u_1 + u_2 + u_3 = u\,(t)$

on dérive par rapport au temps les relations (12), (13), (14) en injectant les valeurs obtenues dans la relation issue de (16) par dérivation, en tenant compte de la relation (15). Il vient :

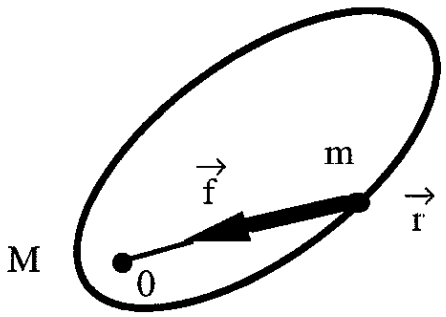
(17) $\quad L \frac{d^2 i}{dt^2} + R \frac{di}{dt} + \frac{1}{C} i = \frac{du}{dt}$

Le modèle obtenu est mathématiquement analogue au modèle (10) décrit au point précédent. On peut l'écrire matriciellement :

(18) $\quad \frac{dX}{dt} = \begin{pmatrix} 0 & 1 \\ -\frac{1}{LC} & -\frac{R}{L} \end{pmatrix} X + \begin{pmatrix} 0 \\ \frac{du}{dt} \end{pmatrix}$

ce qui prend le nom de modèle affine ou "**linéaire avec second membre**".

- **Mouvement d'un satellite**



La masse m placée en $\vec{r}$ est soumise à la force $\vec{f}$ d'attraction de la terre de masse M placée en 0, donnée par la loi de l'attraction universelle de Newton :

(19) $\quad \vec{f} = -\frac{kMm}{|\vec{r}|^2} \frac{\vec{r}}{|\vec{r}|}$

qui équilibre le produit de la masse par l'accélération (loi f = mγ, due au même Newton)

$$(20) \qquad \vec{f} = m \frac{d^2\vec{r}}{dt^2}$$

Le rapprochement de (19) et (20) permet d'écrire l'équation d'évolution du vecteur (à trois composantes) $\vec{r}(t)$.

$$(21) \qquad \frac{d^2\vec{r}}{dt^2} + \frac{kM}{|\vec{r}|^3}\vec{r} = 0$$

On notera qu'historiquement, Newton a également dû inventer la notion de dérivée (!). On peut ici faire deux remarques sur ce modèle :

* Le vecteur $\vec{r}$ est soumis à une équation différentielle **non linéaire** puisque le facteur $\frac{kM}{|\vec{r}|^3}$ devant le terme en $\vec{r}$ dans la relation (21) n'est pas une constante, mais dépend explicitement de $\vec{r}$ (par l'inverse du cube de son module).

* Si la condition initiale :

$$(22) \qquad \vec{r}(o) = \vec{r}_o \ , \ \frac{d\vec{r}}{dt}(o) = \vec{v}_o$$

fait apparaître une vitesse $\vec{v}_o$ de module assez petit, le mouvement est elliptique (lois de Kepler) et est, comme pour le ressort, un comportement de **cycle limite** : r (t) suit indéfiniment une trajectoire fermée de période T.

- **Cinétique chimique**

La réaction chimique de réactants A, B et de produits C, D est paramétrée par les coefficients stoechiométriques ν_A, ν_B, ν_C, ν_D qui sont des nombres entiers, et les constantes de réaction directe k et rétrograde k' :

$$(23) \qquad \nu_A A + \nu_B B \underset{k}{\overset{k'}{\leftrightarrows}} \nu_C C + \nu_D D$$

Étant donné des concentrations molaires initiales $[A]_o$, $[B]_o$, $[C]_o$ et $[D]_o$, quelles sont les concentrations molaires à un instant ultérieur t ? Le vecteur des inconnues à cette fois quatre composantes :

(24) $X = ([A], [B], [C], [D])^T$

[$( )^T$ signifie qu'on transpose la ligne ( ) ; X est donc un vecteur colonne] et la réaction (23) entraîne que les quatre concentrations constituant X sont liées par les trois relations suivantes :

(25) $$-\frac{1}{\nu_A}\frac{d[A]}{dt} = -\frac{1}{\nu_B}\frac{d[B]}{dt} = \frac{1}{\nu_C}\frac{d[C]}{dt} = \frac{1}{\nu_D}\frac{d[D]}{dt} \ (= q)$$

qui définissent le **taux d'avancement** q de la réaction, lequel est donné par la relation :

(26) $$q = k\,[A]^{\nu_A}\,[B]^{\nu_B} - k'\,[C]^{\nu_C}\,[D]^{\nu_D}$$

Cette fonction q est clairement une fonction **non linéaire** du vecteur X introduit en (24) et l'équation d'évolution de X issue des relations (25) et (26) est donnée par l'équation différentielle ordinaire :

(27) $$\frac{dX}{dt} = q(X)\begin{pmatrix} -\nu_A \\ -\nu_B \\ \nu_C \\ \nu_D \end{pmatrix}.$$

L'évolution décrite par la relation (27) peut modéliser des explosions ($\frac{dX}{dt}$ de module très grand devant une grandeur de référence) ou des réactions équilibrées.

- **Météorologie : attracteur de Lorenz**

Dans les années 1960, Lorenz a proposé le modèle **non linéaire** suivant :

(28) $$\begin{cases} \frac{dx}{dt} = -Px + Py \\ \frac{dy}{dt} = -y + rx - xz \\ \frac{dz}{dt} = -bz + xy \end{cases}$$

qui décrit l'évolution du vecteur $X = (x, y, z)^T$ à trois dimensions. Pour des valeurs bien choisies des paramètres P, r, b, l'évolution de X(t), qui reste borné dans une région de l'espace, ne tend pas vers un point fixe ni ne se rapproche d'un cycle limite, mais suit une trajectoire complexe, **difficilement prédictible** (X(t) est délicat à estimer numériquement si on se donne

une condition initiale X(o) et un temps t relativement grand), qui se rapproche de plus en plus d'un objet "proche" d'une surface, et qui est appelée pour toutes ces raisons **attracteur étrange**. La simplicité des équations (28) et les mathématiques vivantes (ie objet de recherches scientifiques en cours actuellement) qu'elles suscitent, montrent que les équations différentielles ordinaires peuvent également constituer un modèle qualitatif pour le comportement de l'atmosphère, difficilement prédictible à long terme malgré la nature déterministe des équations qui le régissent.

- **Définition**

Une équation différentielle ordinaire est une équation d'évolution du type :

(29) $$\frac{dX}{dt} = f(X)$$

où l'inconnue X est caractéristique d'un modèle discret, c'est-à-dire contient un nombre **fini** de composantes :

(30) $$X(t) \in \mathbb{R}^n \text{ , } n \text{ entier ,}$$

la fonction f est une fonction (assez régulière) de la variable X, à valeurs dans $\mathbb{R}^n$:

(31) $$\mathbb{R}^n \ni X \rightarrow f(X) \in \mathbb{R}^n .$$

L'équation différentielle ordinaire (29) est associée à une condition initiale :

(32) $$X(o) = X_o$$

qui est indispensable pour **fermer mathématiquement** le problème, c'est-à-dire assurer qu'en général le problème (29) (32) admet une solution unique X(t), définie sur un intervalle [O,T[($T = \infty$ est le cas le plus courant), et qui vérifie :

(33) $$\frac{d}{dt} X(t) = f(X(t)) \qquad 0 \leq t < T .$$

Lorsque la fonction f peut se mettre sous la forme :

(34) $$f(X) = A X$$

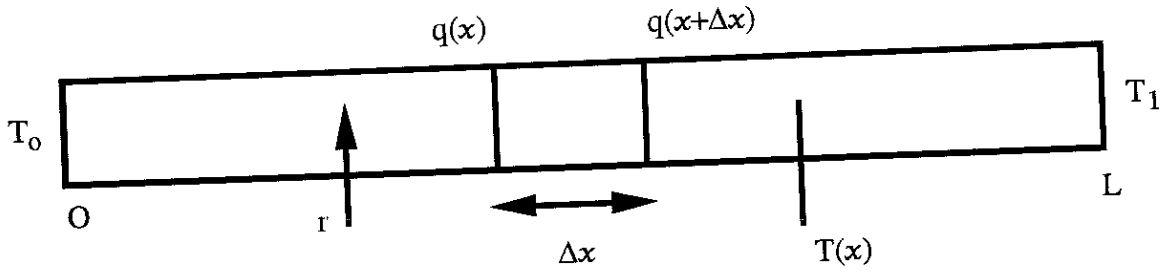
où A est une **matrice** fixée n×n, on dit que l'équation différentielle ordinaire (29) (34) est **linéaire**.

2) Quelques modèles conduisant à des équations aux dérivées partielles.

- **Préambule**

Si une équation différentielle ordinaire se caractérise par une inconnue vivant dans un espace de dimension finie, une équation aux dérivées partielles est caractéristique d'un milieu continu, c'est-à-dire un modèle comportant un nombre infini de degrés de liberté.

- **Diffusion de la chaleur à une dimension d'espace**



Un barreau de longueur L est chauffé par une source r et est maintenu à ses extrémités à une température T_0 (en $x = 0$) et T_L (en $x = L$). On cherche à calculer le champ de température $T(x)$, c'est-à-dire la fonction :

(35) $\quad T : [O,L] \ni x \rightarrow T(x) \in \mathbb{R}$.

L'inconnue est formée maintenant de l'ensemble de toutes les valeurs $T(x)$ pour x parcourant l'intervalle [O,L]. On se donne d'abord les **conditions aux limites**

(36) $\quad T(o) = T_0, \quad T(L) = T_1$

et on indique dans ce qui suit l'équation vérifiée par la température dans l'intervalle]O,L[. On effectue d'abord un **bilan** d'énergie qui indique que le flux de chaleur q entre les abscisses x et $x + \Delta x$ est exactement compensé par les sources $r\,\Delta x$:

(37) $\quad q(x + \Delta x) - q(x) + r\,\Delta x = 0$

on divise par Δx, que l'on fait ensuite tendre vers 0. Il vient simplement :

(38) $\quad \dfrac{dq}{dx} + r(x) = 0.$

Le flux de chaleur est relié au champ de température à l'aide de la **loi de Fourier** qui est une loi phénoménologique caractéristique du matériau et peut ne pas être vérifiée pour certains matériaux, ce qui n'est pas le cas pour la loi du bilan d'énergie :

(39) $$q = -k\frac{dT}{dx}$$

paramétrée par le coefficient de conduction thermique k.

Si on injecte la représentation (39) dans le bilan (38), il vient facilement :

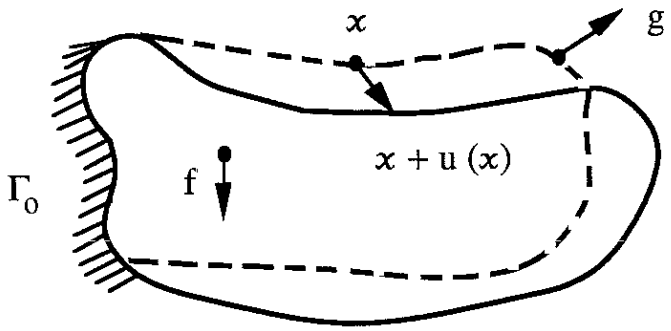
(40) $$-\frac{d}{dx}\left(k\frac{dT}{dx}\right)+r=0, \quad 0<x<L.$$

Le système (36) (40) est caractéristique d'un problème de diffusion de la chaleur avec température imposée aux deux bouts (on parle alors de condition limite de Dirichlet). On peut remplacer l'une des conditions (36) par une condition sur le flux de chaleur en A :

(41) $$-k\frac{dT}{dx}(A) = q_A \qquad A=0 \text{ ou } L$$

qui devient une nouvelle condition limite, dite de **Neumann**. On remarque que le problème (40) est du **second ordre** (il fait intervenir des dérivées partielles d'ordre ou plus égal à deux) et à chaque extrémité, on dispose **d'une** condition limite.

- **Résistance des matériaux : élasticité tridimensionnelle**



Sous l'action des forces volumiques f et de forces de surfaces g, un solide élastique Ω ($\Omega \subset \mathbb{R}^3$ pour signifier que l'on s'intéresse à un modèle tridimensionnel) se déforme. On s'intéresse au **déplacement** u(x) du point x appartenant à Ω (x devient x + u(x) sous l'action des forces f et g), qui est le champ de vecteur des inconnues :

(42) $$\mathbb{R}^3 \supset \Omega \ni x \rightarrow u(x) \in \mathbb{R}^3.$$

L'écriture des équations de l'équilibre mécanique dans le domaine Ω demande d'introduire le tenseur des **déformations** ε_{ij} et le tenseur des **contraintes** σ_{ij} (ε_{ij} et σ_{ij} sont des

matrices 3x3 qui dépendent bien entendu du point x ; $1 \leq i, j \leq 3$). Pour le tenseur des déformations (linéarisées), on pose :

$$(43) \qquad \varepsilon_{ij}(x) = \frac{1}{2}\left(\frac{\partial i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), \quad x \in \Omega.$$

Le tenseur des contraintes σ_{ij} lui est relié par la **loi de comportement** de Hooke, qui spécifie que le solide est élastique isotrope (un métal par exemple, tant que les déformations ne sont pas trop importantes) :

$$(44) \qquad \sigma_{ij} \equiv \lambda\left(\sum_{k=1}^{3} \varepsilon_{kk}\right)\delta_{ij} + 2\mu\ \varepsilon_{ij}.$$

Dans l'expression (44), δ_{ij} est le symbole de Kroncker, qui vaut 1 si $i = j$ et 0 si $i \neq j$ et (λ, μ) forme le couple des **coefficients de Lamé**.

On remarque que la loi de Hooke est une relation linéaire liant ε et σ (la fonction $\{\varepsilon_{kl}\} \rightarrow \sigma_{ij}$ est linéaire) qui n'introduit pas de dérivation en espace (la loi de Hooke est **locale**). Comme ε_{ij}, compte tenu de la relation (43), dépend des dérivées premières du champ inconnu u, il en est de même des contraintes : le tenseur des contraintes σ_{ij} dépend linéairement (et symétriquement) du gradient $\partial_k u_l \equiv \dfrac{\partial u_l}{\partial x_k}$ du champ de déplacement.
L'écriture de l'**équilibre des efforts** au point x (courant) de Ω prend la forme classique

$$(45) \qquad \operatorname{div} \sigma + f = 0 \qquad \text{dans } \Omega$$

soit sous forme développée :

$$(46) \qquad \sum_{j=1}^{3} \frac{\partial}{\partial x_j}\sigma_{ij}(x) + f_i(x) = 0, \quad x \in \Omega, \quad i = 1, 2, 3.$$

Il s'agit clairement [exercice] d'une équation du **second ordre**, c'est-à-dire faisant intervenir les dérivées secondes $\partial_k \partial_l u_m$ du champ de vecteur u.

A l'équation (45), il convient d'ajouter les conditions aux limites qui sont pour cet exemple de deux types **déplacement imposé** (nul par exemple) sur Γ_0 et **force de surface** donnée (égale à g) sur la partie complémentaire Γ_1 (sur le bord $\partial\Omega$ du domaine Ω).

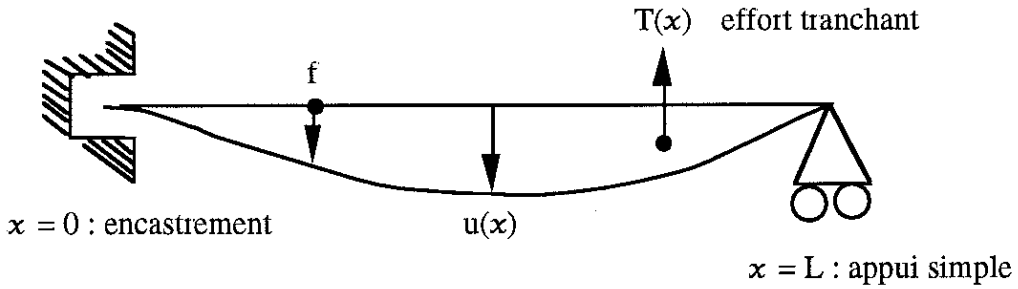
$$(47) \qquad u = u_0(x) \qquad\qquad x \in \Gamma_0,$$

(48) $$\left(\sum_j \sigma_{ij}\, n_j\right)(x) = g_i \qquad x \in \Gamma_1, \qquad i = 1, 2, 3.$$

Dans la relation (48), n_j est la j^{e} composante de la normale unitaire extérieure au bord Γ_1.

Ce modèle de matériau élastique est un modèle **cher** puisque les inconnues $u(x)$ sont vectorielles ($u(x)$ est un vecteur à trois composantes) et le domaine Ω tridimensionnel (x décrit un domaine à trois dimensions). Aussi lui préfère-t-on lorsque la géométrie n'est pas réellement tridimensionnelle (poutres, plaques, coques, etc...), des modèles plus simples. Nous en décrivons deux dans la suite.

- **Résistance des matériaux : flexion des poutres**



Le matériau n'est considéré dans ce modèle que sous un aspect monodimensionnel : le détail d'une section x = cste n'est pas modélisé. L'inconnue $u(x)$, qui s'appelle maintenant la flèche, est un **scalaire** qui ne dépend plus que du point x ($0 \le x \le L$).

Pour établir les équations régissant l'équilibre de la poutre, il faut introduire l'effort tranchant $T(x)$ et le moment fléchissant $M(x)$ qui sont respectivement la moyenne des contraintes verticales dans la section de poutre $x = 0$ et le moment moyen de cette contrainte verticale. L'équation monodimensionnel donnant l'équilibre mécanique est très analogue à la relation (45) :

(49) $$\frac{dT}{dx} + f(x) = 0 \qquad 0 \le x \le L$$

et il faut adjoindre une identité de nature géométrique liant l'effort tranchant et le moment fléchissant :

(50) $$\frac{dM}{dx} + T(x) = 0 \qquad 0 \le x \le L.$$

Par ailleurs, la loi de comportement de solide élastique s'écrit pour le moment fléchissant sous la forme :

$$(51) \qquad M(x) = EI(x) \frac{d^2u}{dx^2} \qquad 0 \le x \le L$$

où E est le module d'Young du matériau, I(x) l'inertie de la section x = cste de la poutre et $\frac{d^2u}{dx^2}$ la courbure linéarisée de la déformée. En introduisant (50) et (51) dans la loi d'équilibre (49), on trouve l'équation vérifiée par la flèche u(x) :

$$(52) \qquad \frac{d^2}{dx^2}\left(EI(x) \frac{d^2u}{dx^2}\right) = f(x), \qquad 0 \le x \le L .$$

Pour cette équation du **quatrième ordre,** deux conditions aux limites sont naturellement à prendre en compte à chaque extrémité. Pour l'exemple décrit sur la figure, nous avons choisi un **encastrement** en $x = 0$ et un **appui simple** en $x = L$. L'encastrement exprime que le déplacement est nul en ce point et que la tangente à la déformée est horizontale, c'est-à-dire :

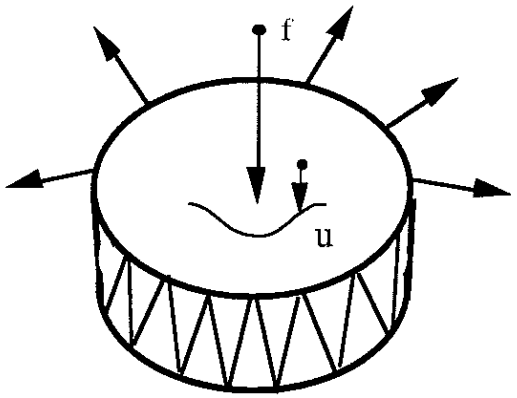
$$(53) \qquad u(o) = 0 \ ; \ \frac{du}{dx}(o) = 0, \quad \text{encastrement en } x = 0.$$

Pour l'appui simple, le déplacement reste nul en $x = L$, mais c'est le moment fléchissant qui est également nul. Compte tenu de la relation (51), on peut écrire :

$$(54) \qquad u(L) = 0 \ ; \ \frac{d^2u}{dx^2}(L) = 0, \quad \text{appui simple en } x = L.$$

L'ensemble (52) (53) (54) constitué de l'équation aux dérivées partielles et des conditions aux limites est facile à résoudre si l'inertie I (x) et les forces f (x) sont constantes ou font l'objet de variations polynomiales. Dans ce cas, la résolution "analytique" de ce modèle est à conseiller, l'emploi d'une méthode d'approximation aux différences finies ne pouvant au mieux fournir qu'une solution approchée, ce qui est inadapté si on connaît une solution exacte. On notera que du point de vue géométrique, le modèle des poutres est très simple et c'est ce qui permet une éventuelle résolution explicite.

- **Membrane vibrante (tambour)**



p : tension du tambour

u = flèche

Une membrane vibrante est modélisée par le déplacement $u(x)$ longitudinal du point x décrivant le domaine bidimensionnel Ω :

$$(55) \qquad \mathbb{R}^3 \supset \Omega \ni x \ \rightarrow u(x) \in \mathbb{R}$$

L'équation donnant l'évolution en temps de la flèche u est paramétrée par la densité ρ de la membrane, l'inertie I, le module d'Young E, le coefficient de poisson ν, l'épaisseur ε et la tension p. Nous ne donnons pas de détail ici sur son obtention ; elle s'écrit :

$$(56) \qquad \rho I \frac{\partial^2 u}{\partial t^2} + \frac{2}{3} \frac{E \varepsilon^3}{1-\nu^2} \Delta^2 u - p \Delta u = f \qquad \text{dans } \Omega$$

où :

$$(57) \qquad \Delta u \equiv \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}$$

est le laplacien et $\Delta^2 u = \Delta (\Delta u)$ le bilaplacien.

Les conditions aux limites sur le bord $\partial\Omega$ expriment la nullité du déplacement u et l'encastrement par la nullité de la dérivée normale $\frac{\partial u}{\partial n}$.

$$(58) \qquad \frac{\partial u}{\partial n} \equiv \nabla u \,.\, n = \sum_{j=1}^{2} \frac{\partial u}{\partial x_j} \, n_j$$

pour un mauvais tambour, c'est-à-dire :

$$(59) \qquad u(x) = 0, \quad \frac{\partial u}{\partial n}(x) = 0, \quad x \in \partial\Omega$$

ou bien la nullité du déplacement et de son laplacien sur le bord, pour un bon tambour :

(60) $\quad u(x) = 0, \quad \Delta u(x) = 0, \quad x \in \partial\Omega$.

• Nous pouvons rapidement, pour le cas de la statique ($\frac{\partial^2}{\partial t^2}$ devient 0 dans l'équation (56)) considérer les deux cas limites où ε est très petit d'une part, et où la tension p est nulle d'autre part. Dans le cas où ε est très petit, on néglige le terme en bilaplacien devant le terme en laplacien, et on obtient l'équation des **membranes prétendues**

$$(61) \qquad \begin{cases} -p\,\Delta u = f & \Omega \\ u = 0 & \partial\Omega \end{cases}$$

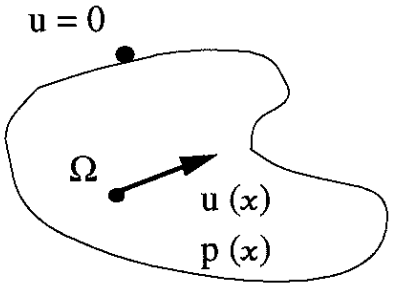
qui constitue le problème de Dirichlet pour l'équation de **Poisson**. On remarque que la condition limite du problème (61) exprime simplement la nullité du déplacement sur le bord, alors qu'on devait prendre deux conditions limites (par exemple (60)) pour le problème initial qui comportait un opérateur du quatrième ordre.

Dans le cas où il n'y a pas de prétension de la membrane, on obtient **l'équation des plaques** :

$$(62) \qquad \begin{cases} \frac{2}{3}\frac{E\varepsilon^3}{1-\nu^2}\,\Delta^2 u = f & \Omega \\ u = 0 \text{ et } \frac{\partial u}{\partial n} = 0 & \partial\Omega \end{cases}$$

qui est l'analogue à dimensions d'espaces de la flexion des poutres à une dimension d'espace.

- **Mécanique des fluides : équations de Navier-Stokes des fluides incompressibles**



Nous abordons maintenant une équation aux dérivées partielles **non linéaire** issue du modèle classique de l'hydrodynamique : les équations de Navier-Stokes d'un fluide incompressible. Le domaine d'étude Ω appartenant à $\mathbb{R}^2$ ou $\mathbb{R}^3$, on cherche un champ de

vitesse u(x) (vectoriel) et un champ de pression p(x) (scalaire) de façon à garantir l'équilibre du bilan d'impulsion et l'incompressibilité :

$$(63) \qquad \frac{\partial u}{\partial t} + u \,.\, \nabla u - \nu \,\Delta u + \nabla p = f \qquad \Omega$$

$$(64) \qquad \text{div } u = 0 \qquad \Omega$$

Dans la relation (63), u . ∇u est un vecteur dont la $i^{ème}$ composante est donnée par la relation :

$$(65) \qquad (u \,.\, \nabla u)_i = \sum_j u_j \left(\frac{\partial}{\partial x_j} u_i \right)$$

A cette équation, il convient de se donner une condition à la limite et le choix le plus classique consiste à annuler la vitesse au bord :

$$(66) \qquad u\,(x, t) = 0 \qquad x \in \partial\Omega\,,\ t > 0$$

De plus, une **condition initiale** est indispensable, puisque l'équation (63) indique que les champs dépendent du temps :

$$(67) \qquad u\,(x, o) = u_o\,(x) \qquad x \in \Omega$$

On remarque qu'aucune condition initiale n'est demandée sur la pression, qui, compte tenu de la condition limite (66), n'est définie qu'à une constante additive près puisque l'ajout d'une constante à p ne change pas le terme ∇p dans l'équation (63).

Si on néglige le terme non linéaire u . ∇u, ce qui est légitime pour une toute petite vitesse qui évolue doucement en espace, et qu'on suppose le champ de vitesse stationnaire, on obtient le **problème de Stokes** des fluides incompressibles :

$$(68) \qquad \begin{cases} -\nu\,\Delta u + \nabla p = f & \Omega \\ \text{div } u = 0 & \Omega \end{cases}$$

La condition limite (66) est inchangée. On remarque qu'il n'y a pas explicitement d'équation pour la pression, ce qui constitue la difficulté mathématique essentielle dans l'étude de ce problème.

- **Mécanique des fluides : équations d'Euler de la dynamique des gaz**

Pour un gaz de densité ρ, d'énergie interne e, la pression p est donnée par la **loi d'état** thermodynamique. Si le gaz est de chaleurs spécifiques constantes (de rapport $\gamma = Cp/Cv$, $\gamma = 1,4$ pour l'air), le gaz est parfait et l'on a :

(69) $p = (\gamma - 1)\ \rho e$ gaz parfait polytropique,

ce qui constitue une différence essentielle par rapport au modèle précédent du fluide incompressible. Les équations d'Euler consistent simplement à écrire la conservation de la masse, de l'impulsion ρu ($u(x)$ est la vitesse du fluide) et de l'énergie totale ρE, avec E qui représente l'énergie totale par unité de masse :

(70) $E = e + \frac{1}{2}\ |u|^2$ énergie totale.

Ce sont des équations non linéaires ; nous les écrivons dans le cas de deux dimensions d'espace :

(71) $$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) + \frac{\partial}{\partial y}(\rho v) = 0$$ masse

(72) $$\frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) + \frac{\partial}{\partial y}(\rho uv) = 0$$ impulsion en x

(73) $$\frac{\partial}{\partial t}(\rho v) + \frac{\partial}{\partial x}(\rho uv) + \frac{\partial}{\partial y}(\rho v^2 + p) = 0$$ impulsion en y

(74) $$\frac{\partial}{\partial t}(\rho E) + \frac{\partial}{\partial x}(\rho uE + pu) + \frac{\partial}{\partial y}(\rho vE + pv) = 0$$ énergie

où la vitesse est maintenant notée (u, v).

Les conditions limites sur une **paroi solide** s'écrivent simplement en imposant la non-pénétration du fluide dans la paroi :

(75) $u \cdot n = 0$ paroi solide

alors que la condition (66) imposait que toutes les composantes de la vitesse soient nulles au bord. Cette différence est due au fait que la modèle d'Euler néglige le flux de chaleur et la diffusion visqueuse, présentes pour les équations de Navier-Stokes.

Il faut aussi imposer des conditions initiales pour prendre en compte l'aspect instationnaire des équations (71) - (74). Notons que ces équations d'Euler sont **fortement non**

linéaires, n'ont de sens physique sous cette forme que pendant un temps limité même si la donnée initiale est très régulière puisqu'en général des **discontinuités** (ondes de choc, surfaces de glissement) peuvent apparaître dans l'écoulement. Toutefois, tant que l'on manipule une solution régulière du modèle, on peut, au terme d'un calcul classique, écrire la **forme non-conservative** des équations (71)-(74) :

$$(76) \qquad \begin{cases} \dfrac{\partial \rho}{\partial t} + \operatorname{div}(\rho u) = 0 \\ \dfrac{\partial u}{\partial t} + u \cdot \nabla u + \dfrac{1}{\rho} \nabla p = 0 \\ \dfrac{\partial s}{\partial t} + u \cdot \nabla u = 0 \end{cases}$$

où s représente l'entropie par unité de masse, qui pour le gaz polytropique d'équation d'état (69), est calculée grâce à la relation :

$$(77) \qquad s = C_v \log\left(\frac{\rho}{\rho^{\gamma}}\right) \qquad \text{entropie spécifique}$$

en fonction de la chaleur spécifique à volume constant C_v et des variables déjà introduites.

- **Équation des ondes acoustiques**

Un état constant est toujours solution des équations (76) de la dynamique des gaz. On peut donc étudier l'évolution d'une petite perturbation autour de cet état constant, la perturbation étant si petite qu'il est alors raisonnable de négliger les termes non-linéaires. On introduit donc un champ perturbant (avec des primes) un état constant de vitesse nulle (avec l'indice zéro) :

$$(78) \qquad \rho = \rho_o + \rho', \quad u = 0 + u', \quad p = p_o + p', \quad s = s_o + s'$$

dans les équations (76). On développe au premier ordre par rapport à la perturbation et on obtient ainsi les équations fondamentales de l'**acoustique** :

$$(79) \qquad \frac{\partial p'}{\partial t} + \rho_o \operatorname{div} u' = 0$$

$$(80) \qquad \rho_o \frac{\partial u'}{\partial t} + \nabla p' = 0$$

$$(81) \qquad \frac{\partial s'}{\partial t} = 0.$$

La relation (81) entraîne l'isentropie de la perturbation, donc compte tenu de la relation (77), on obtient :

$$(82) \qquad \frac{p'}{p_o} = \gamma \frac{\rho'}{\rho_o}$$

qui permet de définir une célérité c_o par la relation :

$$(83) \qquad c_o = \sqrt{\frac{\gamma p_o}{\rho_o}}$$

En prenant la dérivée en temps de la première équation (79) et la divergence de la seconde (80), on obtient par différence, compte tenu de (82) et (83), la relation suivante :

$$(84) \qquad \frac{1}{c_o^2} \frac{\partial^2 p'}{\partial t^2} - \Delta p' = 0 \qquad \Omega \times]O,T[$$

qui est l'**équation des ondes sonores**. On doit lui adjoindre des conditions aux limites sur le bord du domaine d'étude Ω, par exemple :

$$(85) \qquad p' = p'_o \text{ sur } \Gamma_o , \quad \frac{\partial p'}{\partial n} = g \text{ sur } \Gamma_1$$

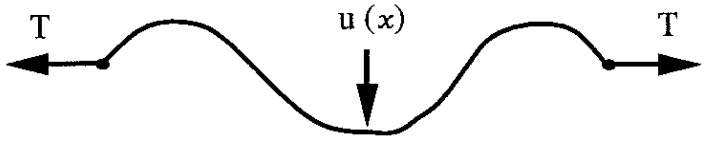
ainsi que des conditions initiales :

$$(86) \qquad p'(O,X) = q_o \qquad \frac{\partial p'}{\partial t}(O,X) = q'_o \qquad \text{dans } \Omega$$

pour assurer que le problème (84) (85) (86) admet une unique solution. Dans le cas où les conditions aux limites sont rejetées à l'infini, on sait depuis d'Alembert que toute combinaison du type :

$$(87) \qquad p'(x,t) = \varphi(\eta . x - c_o t) + \psi(\eta . x + c_o t)$$

où η est un vecteur unitaire donné, est solution de l'équation (84). Le terme en φ est relatif à un profil qui, lorsque le temps croît, se déplace dans la direction η, alors que le terme en ψ se déplace dans la direction opposée au vecteur η. Ici encore, le fait de disposer de solutions de l'équation (87) permet de résoudre de nombreux problèmes pour les ondes sans avoir recours à une méthode numérique de différences finies. Mais lorsque le domaine Ω est de géométrie quelconque, une méthode numérique s'impose (il faut prendre une combinaison infinie de solutions de type (87)) et les éléments finis, aptes à prendre en compte les détails géométriques, constituent un choix possible.



Notons qu'à partir de l'équation (56), en négligeant le terme en ε et en se restreignant par exemple à une dimension d'espace, la perturbation u d'une corde vibrante obéit à une équation identique à (84) :

$$(88) \qquad \frac{1}{c_0^2}\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0 \qquad 0 \le x \le L, \qquad t > 0$$

mais la célérité des ondes est alors donnée par la relation :

$$(89) \qquad c_0^2 = \sqrt{\frac{T}{\rho_L}}$$

où ρ_L est la densité linéique de la corde tendue.

- **Électromagnétisme : équations de Maxwell**

Les équations de Maxwell du champ électromagnétique sont les équations locales qui unifient l'électricité, l'optique et le magnétisme. L'inconnue est le couple (E, B) formé du champ électrique E et du champ magnétique B qui sont tous deux fonctions du temps et de l'espace x (x dans un domaine tridimensionnel Ω pour fixer les idées) :

$$(90) \qquad \begin{cases} \mathbb{R}^3 \times]0,+\infty[\ni (x, t) \to E(x, t) \in \mathbb{R}^3 \\ \mathbb{R}^3 \times]0,+\infty[\ni (x, t) \to B(x, t) \in \mathbb{R}^3 \end{cases}$$

on s'attend donc, comme l'inconnue (E, B) à six composantes, à six équations à résoudre. Les deux équations d'évolution en temps, de Maxwell Ampère pour le champ électrique et de Maxwell Faraday pour le champ magnétique.

$$(91) \qquad -\frac{\partial E}{\partial t} + \operatorname{rot} B = j \qquad \text{Maxwell Ampère}$$

$$(92) \qquad \frac{\partial B}{\partial t} + \operatorname{rot} E = 0 \qquad \text{Maxwell Faraday}$$

donnent l'évolution du champ (E, B) si on se donne la source j du champ, c'est-à-dire la densité du courant électrique, qui est elle-même un champ de vecteurs supposé connu :

(93) $\mathbb{R}^3 \times]0,+\infty[\supset \Omega \times]0,+\infty[\ni (x, t) \rightarrow j(x, t) \in \mathbb{R}^3$

Mais en pratique, on introduit également la densité de charge ρ, qui est un champ scalaire :

(94) $\mathbb{R}^3 \times]0,+\infty[\supset \Omega \times]0,+\infty[\ni (x, t) \rightarrow \rho(x, t) \in \mathbb{R}$

qui vérifie la relation locale dite de **conservation de la charge** :

(95) $$\frac{\partial \rho}{\partial t} + \operatorname{div} j = 0$$

Le champ électrique et le champ magnétique vérifient alors deux équations de contraintes de Gauss :

(96) $\operatorname{div} E = \rho(x, t)$

(97) $\operatorname{div} B = 0$

Si les relations (96) (97) sont satisfaites à l'instant initial $t = 0$, on voit facilement [en prenant la divergence de la relation (91), en dérivant (96) par rapport au temps et en tenant compte de la relation (95) pour le champ électrique ; en prenant la divergence de (92) et en dérivant (97) par rapport au temps pour le champ magnétique] qu'il en est de même à tout instant. Les équations de Gauss (96) et (97) sont donc à considérer comme des **contraintes** à assurer pour la **condition initiale** :

(98) $E(x,0) = E_0(x)$; $B(x, 0) = B_0(x)$, $x \in \Omega$

Par ailleurs, au bord du domaine Ω, on peut se donner une condition limite de conducteur parfait :

(99) $E \times n = 0$ sur $\partial\Omega$, $t > 0$

Signalons que nous avons adopté un système d'unité où la vitesse de la lumière est égale à 1, ainsi que la constante μ_0 de perméabilité du vide. Dans une région de l'espace vide de charge ($\rho = 0$, $j = 0$), on peut dériver (91) par rapport au temps, prendre le rotationnel de (92), tenir compte de l'identité :

(100) $\operatorname{rot}(\operatorname{rot} \psi) = \nabla(\operatorname{div} \psi) - \Delta\psi$

$\psi(x, t)$ champ de vecteurs arbitraire

puis de la relation (96) pour conclure que le champ électrique E satisfait à l'**équation des ondes**.

$$(101) \qquad \frac{\partial^2 E}{\partial t^2} - \Delta E = 0.$$

Il en est de même pour le champ magnétique :

$$(102) \qquad \frac{\partial^2 B}{\partial t^2} - \Delta B = 0.$$

On retrouve la même équation que pour l'acoustique ou les cordes vibrantes, mais dans un modèle physique tout à fait différent. Toute méthode directe ou approchée de résolution de l'équation des ondes pourra s'appliquer indifféremment en électromagnétisme, en acoustique ou en élastodynamique.

- **Équation d'advection : diffusion à une dimension d'espace**

La résolution approchée des modèles fluides réalistes (équations de Navier-Stokes ou d'Euler) ou des équations de Maxwell demandent d'utiliser des logiciels très performants qui demandent d'employer des calculateurs très puissants (vectoriels, parallèles). Dans la suite de ce cours, un modèle très simple mais à la physique réaliste servira de support à l'analyse de plusieurs méthodes numériques : il s'agit de l'advection-diffusion d'un scalaire u dans le cas d'une dimension d'espace.

On se donne un intervalle d'étude]O,L[et on cherche une fonction inconnue $u(x,t)$ qui est scalaire. On se donne également une vitesse d'advection $a > 0$ (éventuellement fonction de x) et une viscosité $\nu > 0$ (également fonction de x si nécessaire). L'équation d'advection-diffusion décrit l'évolution en temps du champ u :

$$(103) \qquad \frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} - \nu \frac{\partial^2 u}{\partial x^2} = 0, \quad 0 \le x \le L, \quad t > 0.$$

C'est un modèle linéaire déjà riche : si $\nu \equiv 0$, on a une simple **advection** et si $u_0(x)$ est la donnée initiale :

$$(104) \qquad u(x,o) = u_0(x) \qquad 0 \le x \le L$$

il est facile de voir que la solution u de (103) [avec $\nu \equiv 0$] est fournie par la relation :

$$(105) \qquad u(x,t) = u_0(x - at)$$

si la vitesse d'advection a est constante. Si réciproquement on suppose a ≡ o mais $\nu \neq 0$, alors (103) est une équation de **diffusion** (évolution en temps de la chaleur). Lorsque $\nu \neq 0$, les conditions aux limites peuvent être de type Dirichlet ou Neumann :

(106) Dirichlet : $u(x,t) = v_0(t) \qquad x = 0 \text{ ou } x = L$

(107) Neumann : $\dfrac{\partial u}{\partial X}(x,t) = w_0(t) \quad x = 0 \text{ ou } x = L$

alors que pour $\nu = 0$, on peut se donner une condition de Dirichlet (106) en $x = 0$, mais aucune condition limite n'est nécessaire en $x = L$!

Notons que le calcul du prix d'une option en mathématiques financières peut être décrit par le modèle dit de Black-Scholès, qui est de type advection-diffusion :

(108)
$$\begin{cases} \dfrac{\partial u}{\partial t} + r x \dfrac{\partial u}{\partial x} + \dfrac{1}{2}\sigma^2 x^2 \dfrac{\partial^2 u}{\partial x^2} - ru = 0 \qquad t > 0\,,\ x > 0 \\ u(T,x) = (x - K) + \left[\equiv x - K \text{ si } x \geq K,\ 0 \text{ sinon}\right] \\ u(t,o) = \varphi(t) \end{cases}$$

- **Physique moléculaire : équation de Schrödinger**

Nous terminons ce panorama de quelques exemples d'équations aux dérivées partielles (qui n'a pas la prétention d'être complet mais cause de nombreux problèmes modèles pour l'ingénieur) par l'équation de schrödinger. Pour x appartenant à $\mathbb{R}^n$ (n=1, 2 ou 3 selon les applications), on cherche une fonction inconnue ψ à valeurs **complexes** :

(109) $\mathbb{R}^n \times]0,+\infty[\ni (x,t) \to \psi(x,t) \in \mathbb{C}$

qui s'interprète comme l'amplitude de probabilité relative à une particule quantique (un électron par exemple) : la probabilité de trouver, l'instant t, la particule dans la région Ω de $\mathbb{R}^n$ est donnée par l'intégrale suivante :

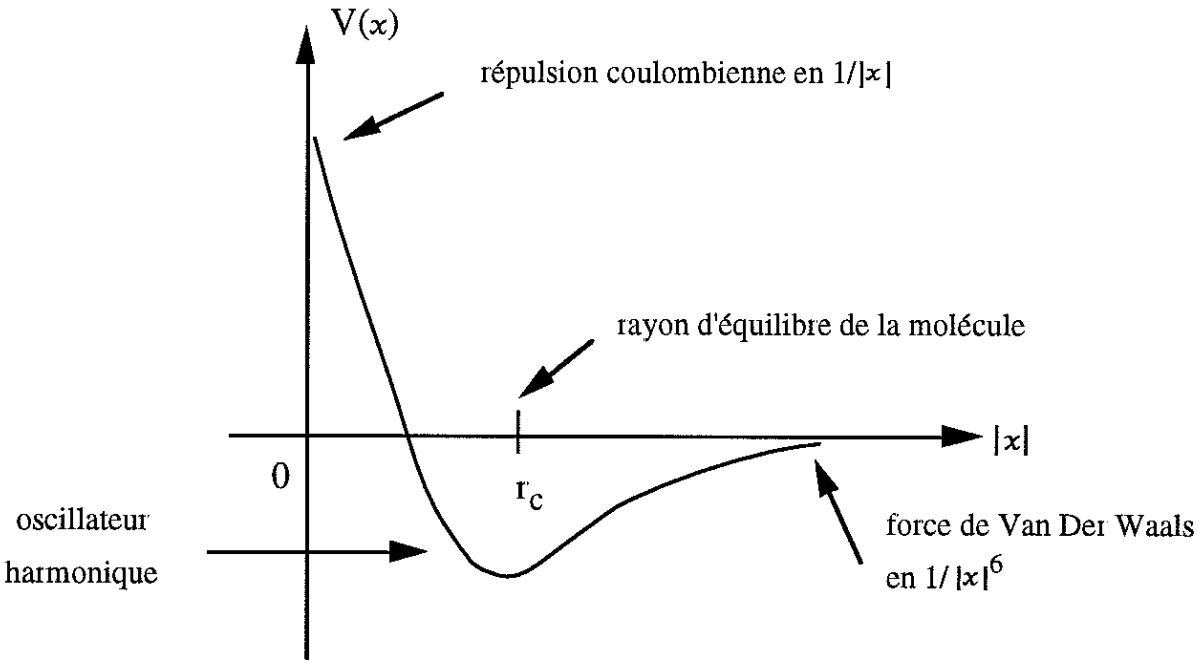
(110) $\displaystyle\int_\Omega |\psi(x,t)|^2 \, dx$ = probabilité de présence dans Ω

et par suite, on cherche ψ de carré intégrable

(111) $\displaystyle\int_{\mathbb{R}^n} |\psi(x,t)|^2 \, dx < \infty$

Si on suppose la particule de masse m soumise au potentiel $V(x)$, l'équation d'évolution due à Schrödinger s'écrit :

(112) $$i \frac{\partial \psi}{\partial t} = -\frac{1}{2m} \Delta\psi + V(x)\,\psi.$$



Potentiel moléculaire typique

Il faut bien entendu se donner une condition initiale, la relation (111) permettant de fixer le comportement à l'infini en espace.

Bien que décrivant une réalité physique microscopique, l'équation (112) ressemble à l'équation de la chaleur [le coefficient i devant le terme $\frac{\partial \psi}{\partial t}$ étant la seule différence, mathématiquement essentielle, si on suppose $V(x) \equiv$] et utilise un opérateur de dérivation en espace parmi les plus courants : le laplacien.

II. DIFFÉRENCES FINIES POUR LES ÉQUATIONS DIFFÉRENTIELLES ORDINAIRES

1) Schémas aux différences : première approche

• On s'intéresse à un modèle différentiel très général, qui peut toujours s'écrire sous la forme :

(1) $\frac{du}{dt} = f(u)$

(2) $u(o) = u_o$

où u(•) est une fonction inconnue qui prend ses valeurs dans $\mathbb{R}^m$ (m allant de 1 à plusieurs millions en pratique) et f une fonction donnée, en général non linéaire de u :

(3) $\mathbb{R}^m \ni u \rightarrow f(u) \in \mathbb{R}^m$

assez régulière pour que les conditions (1) (2) garantissent existence et unicité d'une solution u(•) pour t appartenant à un intervalle [O,T[, avec T égal à $+\infty$ dans de nombreux cas pratiques.

On suppose qu'on ne dispose **pas** de solution "analytique" du modèle (1) (2) et la méthode des différences finies propose une approche pour évaluer en pratique des valeurs voisines des valeurs exactes, représentant une réalité mathématique hors d'atteinte du calcul numérique.

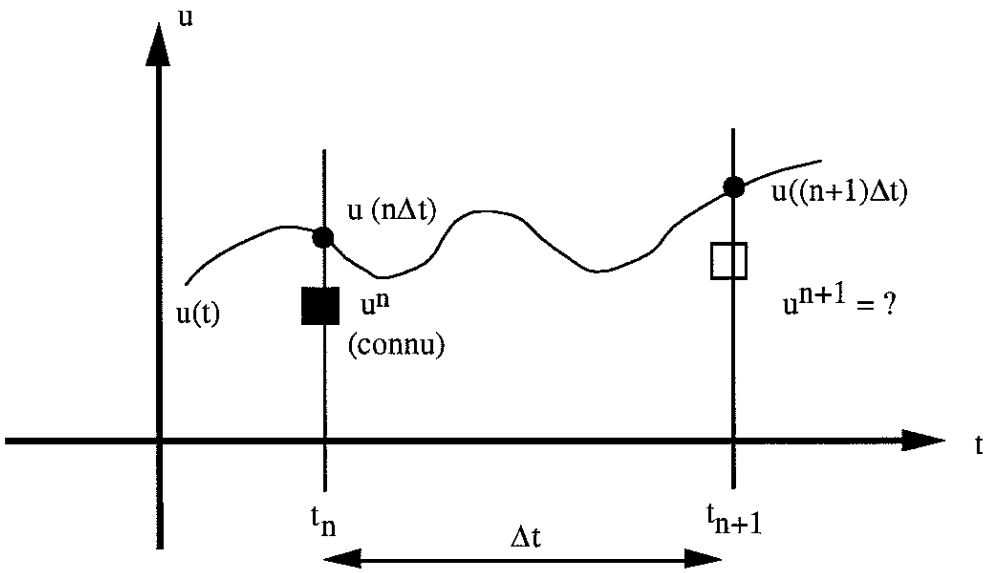
• On se donne un pas de temps $\Delta t > 0$ et la méthode des différences finies consiste à rechercher une solution approchée de u(•) aux **seuls** instants $t_n = n\Delta t$:

(4) $t_m - t_n = \Delta t \qquad n \geq 0 ; \Delta t > 0$

on introduit la notation classique :

(5) $u^n \simeq u(n\Delta t)$

Le problème qui se pose est le suivant : connaissant une approximation u^n de $u(t_n)$, comment calculer une nouvelle approximation au pas de temps suivant, pour u^{n+1} ?



• L'idée la plus simple consiste à écrire la formule de Taylor autour de $u(t_n)$ pour développer $u(t_{n+1})$ autour du point précédent :

(6) $$u(t_{n+1}) = u(t_n) + \Delta t \frac{du}{dt}(t_n) + 0(\Delta t^2).$$

Compte tenu de l'équation différentielle (1), la formule de Taylor peut encore s'écrire :

(7) $$u(t_{n+1}) = u(t_n) + \Delta t\, f(u(t_n)) + 0(\Delta t^2).$$

Le schéma d'**Euler explicite** (forward Euler en anglais) consiste à remplacer la relation (7), exacte pour la solution du problème différentiel (1) (2), par la relation obtenue en tronquant le développement (7) au premier ordre en temps. On pose :

(8) $$u^{n+1} = u^n + \Delta t\, f(u^n)$$ Euler explicite.

Nous disons (et justifierons plus loin cette expression) que le schéma numérique est du premier ordre en temps. Par ailleurs, il est **explicite**, ce qui signifie que si u^n est connu (ce qui est le cas à l'instant initial n = 0 ; il suffit de prendre $u^0 = u_0$), alors u^{n+1} est calculé très simplement à partir de celui-ci et de la fonction f.

On peut aussi écrire la formule de Taylor "à l'envers" (backward), c'est-à-dire essayer de retrouver la valeur (supposée connue) u^n à partir de la valeur (supposée inconnue) u^{n+1} :

(9) $u(t_n) = u(t_{n+1}) - \Delta t\, f(u(t_{n+1})) + 0\,(\Delta t^2)$

Si on tronque ce développement, on définit un nouveau schéma aux différences pour calculer u^{n+1} à partir de la valeur u^n :

(10) $u^{n+1} - \Delta t\, f(u^{n+1}) = u^n$ Euler implicite.

Ce schéma, encore du premier ordre de précision, est un schéma dit **implicite** ; en effet, il ne fournit qu'une **équation** (qu'il faut ensuite résoudre numériquement à l'aide de techniques classiques comme l'algorithme de Newton par exemple) relative à l'inconnue u^{n+1}, et non une "formule" comme à la relation (8). Nous verrons que ce défaut pratique est en fait compensé par d'autres propriétés (stabilité).

- Si on essaie de construire un schéma numérique plus précis, l'idée naturelle est de pousser plus loin les développements de Taylor (7) et (9) :

(11) $u(t_{n+1}) = u(t_n) + \Delta t\, f(u(t_n)) + \frac{1}{2}\,\Delta t^2\,\frac{d^2u}{dt^2}\,(t_n) + 0\,(\Delta t^3)$

(12) $u(t_n) = u(t_{n+1}) - \Delta t\, f(u(t_{n+1})) + \frac{1}{2}\,\Delta t^2\,\frac{d^2u}{dt^2}\,(t_{n+1}) + 0\,(\Delta t^3)$

On remarque alors que pour la dérivée seconde, on a le développement très simple :

(13) $\frac{d^2u}{dt^2}\,(t_{n+1}) = \frac{d^2u}{dt^2}\,(t_n) + 0\,(\Delta t)$

qu'on reporte dans la relation (12) pour obtenir :

(14) $u(t_n) = u(t_{n+1}) - \Delta t\, f(u\,(t_{n+1})) + \frac{1}{2}\,\Delta t^2\,\frac{d^2u}{dt^2}\,(t_n) + 0\,(\Delta t^3)$

on retranche alors la relation (14) de la relation (12), ce qui nous donne, au troisième ordre près :

(15) $2\left[u(t_{n+1}) - u(t_n)\right] = \Delta t\,\left[f(u(t_n)) + f(u(t_{n+1}))\right]$.

Si on décide de tronquer ce développement de Taylor, on obtient un schéma numérique, dit de **Crank-Nicolson**, précis à l'**ordre deux**.

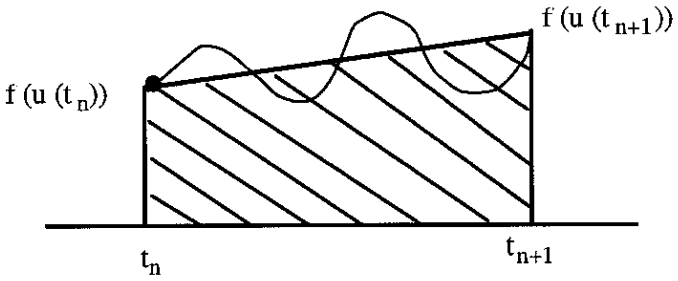
(16) $$u^{n+1} - \frac{1}{2}\,\Delta t\, f(u^{n+1}) = u^n + \frac{1}{2}\,\Delta t\, f(u^n)$$ Crank-Nicolson.

La relation (16) est une équation d'inconnue u^{n+1} qui, comme pour le schéma d'Euler implicite, ne fournit pas une formule directement calculable pour la valeur approchée à l'instant intérieur. Le schéma de Crank-Nicolson est **implicite**.

• Une autre façon, plus directe, pour trouver facilement la relation (16), consiste à écrire la différence finie $u(t_{n+1}) - u(t_n)$ sous forme intégrale, compte tenu de la relation (1) :

(17) $$u(t_{n+1}) = u(t_n) + \Delta t \left[\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(u(\theta))\, d\theta\right]$$

et de calculer l'intégrale au second membre de (17) par la formule du trapèze (on remplace $f(u(\theta))$ par son interpolé affine entre t_n et t_{n+1}) :



(18) $$\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(u(\theta))\, d\theta = \frac{1}{2}\left[f(u(t_n)) + f(u(t_{n+1}))\right] + 0\,(\Delta t^2)$$

Le schéma de Crank-Nicolson consiste alors simplement à tronquer la relation (18) et à injecter la formule approchée du trapèze au second membre de la relation (17). On retrouve alors la relation (16) qui définit le schéma numérique.

2) Test des schémas d'Euler explicite, Euler implicite et Crank-Nicolson pour le modèle $\frac{du}{dt} + \lambda u = 0$ ($\lambda > 0$)

Comme toujours devant une nouvelle approche, il est bon de critiquer les résultats obtenus sur une base de connaissances bien assises sur l'expérience. On considère donc le modèle différentiel très simple (mais fondamental !).

(19) $$\frac{du}{dt} + \lambda u = 0 \qquad (\lambda > 0)$$

dont la solution au temps $(n+1)\,\Delta t$ est bien connue en fonction de la solution à l'instant $n\Delta t$:

(20) $$u(t_{n+1}) = u(t_n)\, e^{-\lambda \Delta t}$$

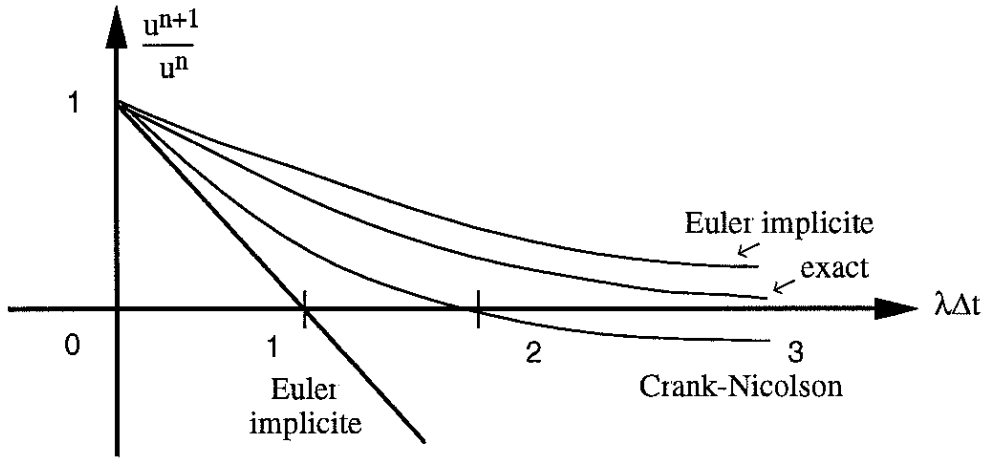
En remplaçant f(u) par sa valeur $-\lambda u$ dans les relations (8), (10) et (16), on calcule les approximations fournies par les trois schémas introduits jusqu'ici :

(21) $u^{n+1} = u^n (1 - \lambda \Delta t)$ Euler explicite

(22) $u^{n+1} = u^n \dfrac{1}{1 + \lambda \Delta t}$ Euler implicite

(23) $u^{n+1} = u^n \dfrac{1 - \frac{\lambda \Delta t}{2}}{1 + \frac{\lambda \Delta t}{2}}$ Crank-Nicolson

Il est intéressant de représenter sur une même graphe le rapport $\dfrac{u^{n+1}}{u^n}$ en fonction de la variable $x = \lambda \Delta t$.



- Il saute aux yeux que le quotient $\dfrac{u^{n+1}}{u^n}$ reste **positif** pour la solution exacte (20), quelle que soit la valeur de $\lambda \Delta t$, ce qui n'est **pas** le cas pour le schéma d'Euler explicite si $\lambda \Delta t > 1$ ou le schéma de Crank-Nicolson pour $\lambda \Delta t > 2$; nous voyons un champ u(t) "physiquement" toujours positif qui devient négatif, ce qui n'est pas admissible ! Nous devons donc imposer une **condition de stabilité** sur le pas de temps, pour garantir cette cohérence :

(24)	Euler explicite	$\lambda\Delta t \leq 1$
(25)	Euler implicite	rien, pas de condition
(26)	Crank-Nicolson	$\lambda\Delta t \leq 2$

Le schéma d'Euler implicite reste toujours utilisable quelle que soit la valeur du pas de temps selon le critère de maintien de la positivité qui nous a servi ici ; toutefois la propriété d'être implicite n'est pas un gage de stabilité pour tout pas de temps, ainsi qu'en témoigne le schéma de Crank-Nicolson à la relation (26).

- **Définition - Méthode à un pas**

Une méthode à un pas est une méthode de calcul de la solution approchée u^{n+1} au temps $t_{n+1} = t_n + \Delta t$ telle que **seule** la donnée u^n, approximation de la solution au temps t_n, suffit pour définir la méthode. Dans le cas contraire, on parle d'une méthode multipas.

3) Exemples de méthodes multipas

- **Schéma instable**

On expose d'abord un mauvais choix qui conduit à une **instabilité numérique**, c'est-à-dire un comportement de la méthode d'approximation incompatible avec le comportement attendu de l'équation différentielle ordinaire $\frac{du}{dt} = f(u)$.

L'idée de la construction de ce schéma est identique à l'approche déjà vue plus haut ; on développe par la formule de Taylor u autour de t_n jusqu'aux instants t_{n-1} et t_{n+1} :

(27) $$u(t_{n+1}) = u(t_n) + \Delta t\, f(u(t_n)) + \frac{1}{2}\Delta t^2 \frac{d^2u}{dt^2}(t_n) + 0\,(\Delta t^3)$$

(28) $$u(t_{n-1}) = u(t_n) - \Delta t\, f(u(t_n)) + \frac{1}{2}\Delta t^2 \frac{d^2u}{dt^2}(t_n) + 0\,(\Delta t^3).$$

On retranche ensuite la relation (28) de (27), ce qui donne le développement limité suivant :

(29) $$u\,(t_{n+1}) = u(t_{n-1}) + 2\Delta t\, f(u(t_n)) + 0\,(\Delta t^3).$$

Le schéma numérique résulte alors de cette dernière relation : on néglige le reste, quitte à remplacer les valeurs $u(t_k)$ de la solution exacte par des valeurs approchées u^k :

(30) $\quad u^{n+1} = u^{n-1} + 2\Delta t\, f(u^n)$.

Ce schéma est à **deux pas** puisqu'il faut connaître à la fois u^{n-1} et u^n pour calculer u^{n+1}, il est explicite [la relation (30) est une formule de calcul de u^{n+1} en fonction des données u^{n-1} et u^n] et il est d'ordre deux de précision.

- **Schéma d'Adams-Bashford**

Le choix qui conduit à une méthode stable a été proposé par **Adams-Bashford.** On part encore du développement de Taylor (27) mais on essaie alors d'exprimer la dérivée seconde $\frac{d^2u}{dt^2}(t_n)$ à l'aide de dérivées premières :

(31) $$\frac{d^2u}{dt^2}(t_n) = \frac{1}{\Delta t}\left\{\frac{du}{dt}(t_n) - \frac{du}{dt}(t_{n-1})\right\} + 0\,(\Delta t)$$

(32) $$\frac{d^2u}{dt^2}(t_n) = \frac{1}{\Delta t}\left(f(u(t_n)) - f(u(t_{n-1}))\right) + 0\,(\Delta t)$$

Quand on reporte la relation (32) dans le développement (27), on obtient la relation suivante :

(33) $$u(t_{n+1}) = u(t_n) + \frac{3}{2}\,\Delta t\, f(u(t_n)) - \frac{\Delta t}{2} f(u(t_{n-1})) + 0\,(\Delta t^3)$$

qui, une fois tronquée, définit un schéma qui a exactement les mêmes caractéristiques que le schéma en plus haut à la relation (30), à deux pas, du second ordre de précision en temps, explicite ; c'est le schéma d'**Adams-Bashford d'ordre 2** (AB2 en abrégé) :

(34) $$u^{n+1} = u^n + \frac{3}{2}\,\Delta t\, f(u^n) - \frac{1}{2}\Delta t\, f(u^{n-1})$$ Adams-Bashford d'ordre 2.

Pour étudier de façon plus approfondie ces deux schémas, on les teste sur le modèle différentiel exponentiel très simple : $\frac{du}{dt} + \lambda u = 0$, avec λ strictement positif. Les relations (30) et (34) prennent alors la forme de deux **suites récurrentes linéaires à coefficients constants** :

(35) $$u^{n+1} + 2\lambda\Delta t\, u^n - u^{n-1} = 0$$

(36) $$u^{n+1} - \left(1 - \frac{3}{2}\lambda\Delta t\right) u^n - \frac{1}{2}\,\lambda\Delta t\, u^{n-1} = 0$$ (AB2)

on sait alors que l'ensemble des solutions de (35) [respectivement (36)] est un espace vectoriel de dimension deux, engendré par les solutions exponentielles de la forme :

(37) $u^n = s^n$

où s est un nombre solution de l'équation (38) [respectivement (39)] :

(38) $s^2 + 2(\lambda\Delta t)\, s - 1 = 0$

(39) $s^2 - \left(1 - \frac{3}{2}\lambda\Delta t\right) s - \frac{1}{2}\lambda\Delta t = 0$ (AB2)

et dans le membre de droite de la relation (37), s^n désigne effectivement s, à la puissance n.

- Le calcul des racines de la relation (38) est élémentaire :

(40) $s_+ = \sqrt{1 + (\lambda\Delta t)^2} - \lambda\Delta t = 1 - \lambda\Delta t + 0(\Delta t^2)$

(41) $s_- = -\sqrt{1 + (\lambda\Delta t)^2} - \lambda\Delta t$

et l'on remarque que le module de s_- est toujours strictement supérieur à 1 dès que $\lambda\Delta t > 0$:

(42) $|s_-| > 1 \qquad \forall \Delta t > 0\ ,\quad \text{si } \lambda > 0.$

La solution générale du schéma (30) s'écrit donc :

(43) $u^n = \alpha\,(s_+)^n + \beta\,(s_-)^n$

où α et β sont des nombres réels fixés. Pour $\lambda\Delta t$ assez petit, le module de s_+ reste inférieur à 1 (cf la relation (40)) alors que $(s_-)^n$ converge vers l'infini ($+\infty$ **et** $-\infty$ puisque $s_- < 0$), ce qui entraîne que la suite u^n définie par (43) est **non bornée** et prend des valeurs alternativement positives et négatives si n devient assez grand. Ceci contredit bien sûr le comportement exponentiellement décroissant de u^n au fur et à mesure que le temps $t_n = n\Delta t$ croît. Même si on choisit le couple (u^o, u^1) de sorte que le coefficient β soit nul, les **erreurs d'arrondis** qu'effectue la calculatrice électronique imposent de voir β comme une variable aléatoire, ce qui montre que, quitte à attendre assez longtemps, β est non nul et le terme $3(s_-)^n$ dans la relation (43) devient finalement dominant, et a le comportement décrit plus haut. Nous concluons que **le schéma (30) est instable** lorsqu'on essaie de l'utiliser sur l'équation modèle $\frac{du}{dt} + \lambda u = 0$, ce quelle que soit la valeur du pas de temps $\Delta t > 0$!

• On procède de façon analogue pour l'équation du second degré (39). Les racines de cette équation sont données par les relations :

$$(44) \qquad s_+ = \frac{1}{2}\left(1 - \frac{3}{2}\lambda\Delta t\right) + \frac{1}{2}\sqrt{1 - \lambda\Delta t + \frac{9}{4}(\lambda\Delta t)^2} = 1 - \lambda\Delta t + 0(\Delta t^2)$$

$$(45) \qquad s_- = \frac{1}{2}\left(1 - \frac{3}{2}\lambda\Delta t\right) + \frac{1}{2}\sqrt{1 - \lambda\Delta t + \frac{9}{4}(\lambda\Delta t)^2} = -\frac{1}{2}\lambda\Delta t + 0(\Delta t^2)$$

Le mode s_+ correspond au "mode physique" : si on prend $\beta = 0$ dans la relation (43), le rapport $\frac{u^{n+1}}{u^n}$ vaut s_+ qui est une approximation au premier ordre de exp $(-\lambda\Delta t)$, compte tenu du développement limité proposé à la relation (44). Le mode s_- donné par la relation (45) est un **mode numérique parasite**, qui s'amortit tant que s_- reste de module strictement inférieur à 1, ce qui, si on se contente du développement donné à la relation (45), prend la forme :

$$(46) \qquad \lambda\Delta t < 2.$$

Cette condition est en fait générale : il est clair, en prenant $s = 0$ dans la relation (39), que 0 est entre les deux racines s_- et s_+, puis en prenant $s = -1$, que -1 est extérieur aux deux racines tant que $1 + \left(1 - \frac{3}{2}\lambda\Delta t\right) - \frac{1}{2}\lambda\Delta t > 0$, ce qui exprime la condition de stabilité générale.

$$(47) \qquad \lambda\Delta t < 1.$$

Le schéma d'Adams Bashford est **conditionnellement stable** et une condition de type (47) relie le pas de temps à la dynamique du système différentiel étudié (constante de temps $1/\lambda$).

• Les schémas multipas souffrent des deux défauts suivants : d'une part l'**initialisation** demande, pour un schéma à deux pas, de connaître u^o et u^1, c'est-à-dire plus d'information que la seule condition initiale (2). D'autre part, il n'est pas possible, pour garder la précision du schéma, d'introduire un pas de temps variable Δt_n, avec :

$$(48) \qquad t_{n+1} = t_n + \Delta t_n$$

et le pas de temps Δt **reste fixe** pour toute la simulation numérique. Ces défauts nous incitent à approfondir l'étude des schémas à un pas, en construisant des schémas à un pas, explicites et du second ordre de précision.

4) Schémas à un pas explicites précis

- **Schéma d'Euler modifié**

On part de la forme intégrale (17) de l'équation différentielle (1) entre les instants t_n et t_{n+1}, puis on calcule l'intégrale par une formule de quadrature précise au second ordre :

$$(49) \qquad \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(u(\theta))\, d\theta = f\left(u(t_{n+\frac{1}{2}})\right) + 0(\Delta t^2)$$

La valeur de u à l'instant intermédiaire $n+\frac{1}{2}$ est prédite par un développement de Taylor :

$$(50) \qquad u(t_{n+\frac{1}{2}}) = u(t_n) + \frac{1}{2}\,\Delta t\; f(u(t_n)) + 0(\Delta t^2)$$

Si on tronque maintenant les relations (49) et (50), on définit le schéma d'Euler modifié :

$$(51) \qquad \tilde{u}^{n+\frac{1}{2}} = u^n + \frac{1}{2}\,\Delta t\, f(u^n)$$

$$(52) \qquad u^{n+1} = u^n + \Delta t\, f\left(\tilde{u}^{n+\frac{1}{2}}\right)$$

Ce schéma est **explicite, d'ordre deux**, et **à un pas** : en remplaçant la relation (51) au sein de la relation (52), on a :

$$(53) \qquad u^{n+1} = u^n + \Delta t\, f\left[u^n + \frac{1}{2}\Delta t\, f(u^n)\right]$$

ce qui montre que seule la connaissance de u^n est nécessaire pour calculer u^{n+1}.

- **Schéma de Heun**

L'idée est très voisine de celle qui a conduit au schéma de Crank-Nicolson. On développe l'intégrale du membre de droite de la relation (17) à l'aide de la formule des trapèzes (18) mais on remplace dans cette dernière relation u^{n+1} par la relation prédite à l'aide du schéma d'Euler explicite. Avec des formules algébriques, on a donc :

$$(54) \qquad \tilde{u}^{n+1} = u^n + \Delta t\, f(u^n)$$

(55) $$u^{n+1} = u^n + \frac{\Delta t}{2}\left(f(u^n) + f(\tilde{u}^{n+1})\right)$$

soit sous forme équivalente :

(56) $$u^{n+1} = u^n + \frac{1}{2}\,\Delta t\left[f(u^n) + f(u^n + \Delta t\, f(u^n))\right]$$

Le schéma de Heun (56) est **explicite**, à **un pas** et **d'ordre deux**.

- **Schéma de Runge-Kutta**

Nous terminons cette introduction aux méthodes à un pas précises par le schéma de **Runge et Kutta d'ordre 4**, très populaire, malgré des défauts que le lecteur ne pourra juger qu'avec sa propre pratique.

On commence par développer l'intégrale au membre de droite de la relation (17) à l'aide de la formule de Simpson, qui consiste à remplacer la fonction à intégrer par un interpolé polynomial de degré deux :

(57) $$\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(u(\theta))\,d\theta = \frac{1}{6} f(u(t_n)) + \frac{2}{3} f\left(u(t_{n+\frac{1}{2}})\right) + \frac{1}{6} f(u(t_{n+1})) + 0\,(\Delta t^4)$$

Comme pour le schéma de Heun (qui est en fait un schéma de Runge et Kutta d'ordre deux), on remplace aux temps $t_{n+\frac{1}{2}}$ et t_{n+1} la fonction u par des approximations calculées à l'aide du schéma d'Euler explicite et Euler modifié. Nous donnons les relations de définition de ces états intermédiaires sans commentaire particulier.

(58) $$\tilde{u}^{n+\frac{1}{2}} = u^n + \frac{1}{2}\,\Delta t\; f(u^n)$$

(59) $$\tilde{\tilde{u}}^{n+\frac{1}{2}} = n^n + \frac{1}{2}\,\Delta t\; f(\tilde{u}^{n+\frac{1}{2}})$$

(60) $$\tilde{u}^{n+1} = u^n + \Delta t\; f(\tilde{\tilde{u}}^{n+\frac{1}{2}})$$

(61) $$u^{n+1} = u^n + \frac{1}{6}\,\Delta t\left[f(u^n) + 2\,f(\tilde{u}^{n+\frac{1}{2}}) + 2\,f(\tilde{\tilde{u}}^{n+\frac{1}{2}}) + f(\tilde{u}^{n+1})\right]$$

Le schéma défini par les relations (58)-(61) est explicite, d'ordre 4 et conditionnellement stable (le pas de temps Δt doit être limité par une condition de type (47) pour garantir que les valeurs numériques calculées à l'aide de l'approximation gardent quel que soit $n \geq 0$ un sens physique.

5) Ordre d'un schéma aux différences

La notion d'ordre d'un schéma aux différences finies, employée jusque là de façon intuitive, admet une définition précise qui n'est pas très simple à exprimer ; c'est la raison pour laquelle elle intervient si tard dans ce chapitre.

On envisage une forme très générale de schéma pour approcher numériquement les solutions de l'équation (1), forme qui contient tous les exemples vus jusqu'ici.

$$(62) \qquad \frac{1}{\Delta t}\left(u^{n+1} - n^{n}\right) = \phi_f\left(u^{n+1}, u^{n}, u^{n-1}, \ldots, u^{n-p}\right)$$

Les points de suspension indiquent que le schéma utilise les p instants précédents (schéma multipas d'ordre p+1, à un pas si p = 0). De plus, sous la forme (62) le schéma est implicite. A titre d'exercice, nous invitons le lecteur à expliciter la fonction ϕ_f pour le schéma de Runge-Kutta (58)-(61).

L'erreur de troncature $\tau(\Delta t)$ est obtenue en remplaçant dans la relation (62) les valeurs approchées u^{n+j} par les valeurs exactes $u(t_{n+j})$ de la solution de (1) aux différents instants.

Comme la solution exacte ne vérifie pas a priori la relation (62), celle-ci n'est plus vérifiée dans le remplacement proposé. Aussi on définit l'erreur de troncature $\tau(\Delta t)$ par la relation suivante :

$$(63) \qquad \tau(\Delta t) \equiv \frac{1}{\Delta t}\left[u(t_{n+1}) - u(t_n)\right] - \phi_f\left(u(t_{n+1}), u(t_n), u(t_{n-1}), \ldots, u(t_{n-p})\right)$$

On remarque que, sauf cas exceptionnel où l'équation (1) possède une solution analytique, l'erreur de troncature (63) ne peut pas se calculer par une formule analytique. Toutefois, on peut faire un développement limité de $\tau(\Delta t)$ quand Δt tend vers zéro. Lorsque $\tau(\Delta t)$ **est de l'ordre de 0** (Δt^k) pour un certain entier k, ie :

$$(64) \qquad \tau(\Delta t) = 0(\Delta t^k) \qquad \text{ordre k,}$$

on dit que le schéma numérique (62) est **d'ordre k** de précision.

A titre d'exercice, nous invitons le lecteur à vérifier que les différents schémas introduits dans ce chapitre ont bien l'ordre (au sens de la définition ci-dessus) annoncé dans le cours du texte.

6) Mémorandum

schéma	formule (1)	nombre de pas	explicite ou implicite	ordre	condition de stabilité
Euler explicite	$u^{n+1} = u^n + \Delta t\, f(u^n)$	1	explicite	1	$\lambda\Delta t < 1$
Euler implicite	$u^{n+1} - \Delta t\, f(u^{n+1}) = u^n$	1	implicite	1	–
Crank-Nicolson	$u^{n+1} - \frac{\Delta t}{2} f(u^{n+1}) = u^n + \frac{\Delta t}{2} f(u^n)$	1	implicite	2	$\lambda\Delta t < 2$
Adams-Bashford à 2 pas	$u^{n+1} = u^n + \frac{3}{2}\Delta t\, f(u^n) - \frac{1}{2}\Delta t\, f(u^{n+1})$	2	explicite	2	$\lambda\Delta < 1$
Euler modifié	$\tilde{u}^{n+\frac{1}{2}} = n^n + \frac{\Delta t}{2} f(u^n)$ $u^{n+1} = u^n + \Delta t\, f(\tilde{u}^{n+\frac{1}{2}})$	1	explicite	2	$\lambda\Delta t < 2$
Heun	$\tilde{u}^{n+1} = u^n + \Delta t\, f(u^n)$ $u^{n+1} = u^n + \frac{\Delta t}{2}\left[f(u^n) + f(\tilde{u}^{n+1})\right]$	1	explicite	2	$\lambda\Delta t < 2$
Runge-Kutta	$k_1 = f(u^n)$ $k_2 = f(u^n + \frac{\Delta t}{2} k_1)$ $k_3 = f(u^n + \frac{\Delta t}{2} k_2)$ $k_4 = f(u^n + \Delta t\, k_3)$ $u^{n+1} = u^n + \frac{\Delta t}{6}\{k_1+2k_2+2k_3+k_4\}$	1	explicite	4	$\lambda\Delta t < 2{,}8$

III. DIFFÉRENCES FINIES POUR L'ÉQUATION D'ADVECTION À UNE DIMENSION D'ESPACE

1) Équation d'advection à une dimension d'espace

• L'équation d'advection à une dimension d'espace est un modèle mathématique ultérieurement simplifié qui représente une partie des phénomènes de transport présents par exemple pour la dynamique des gaz. Ainsi, considérons un instant le modèle des équations d'Euler des fluides parfaits compressibles (pas de viscosité ni de diffusion de la chaleur) que nous écrivons sous forme non conservative. La densité ρ, la vitesse $\vec{v}$ et l'entropie spécifique s évoluent en temps selon les relations :

(1) $$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \vec{v}) = 0$$

(2) $$\frac{d\vec{v}}{dt} + \vec{v}.\nabla\vec{v} + \frac{1}{\rho}\nabla p = 0$$

(3) $$\frac{ds}{dt} + \vec{v}.\nabla s = 0$$

Supposons le champ de vitesses uniforme que nous notons a dans la suite et alignons l'axe des x le long de ce champ, ce qui annule du coup les termes en $\frac{\partial}{\partial y}$ dans les équations (1) et (3). La densité ρ et l'entropie s vérifient la **même** équation d'advection, nous l'écrivons :

(4) $$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0, \qquad t > 0, \quad x \in \mathbb{R}$$

la lettre u désignant l'une quelconque des deux variables scalaires précédentes.

Nous nous intéressons dans la suite au **problème de Cauchy** pour l'équation d'advection (4), c'est-à-dire à l'équation aux dérivées partielles, jointe à la **condition initiale**.

(5) $u(o,x) = u_o(x) \qquad x \in \mathbb{R}$.

• Le problème de Cauchy (4) (5) est mathématiquement trivial, puisque nous avons la proposition suivante.

Proposition : La solution u (t,x) du problème (4) (5) est donnée par l'expression suivante :

(6) $u(t,x) = u_o(x - at)$.

Preuve

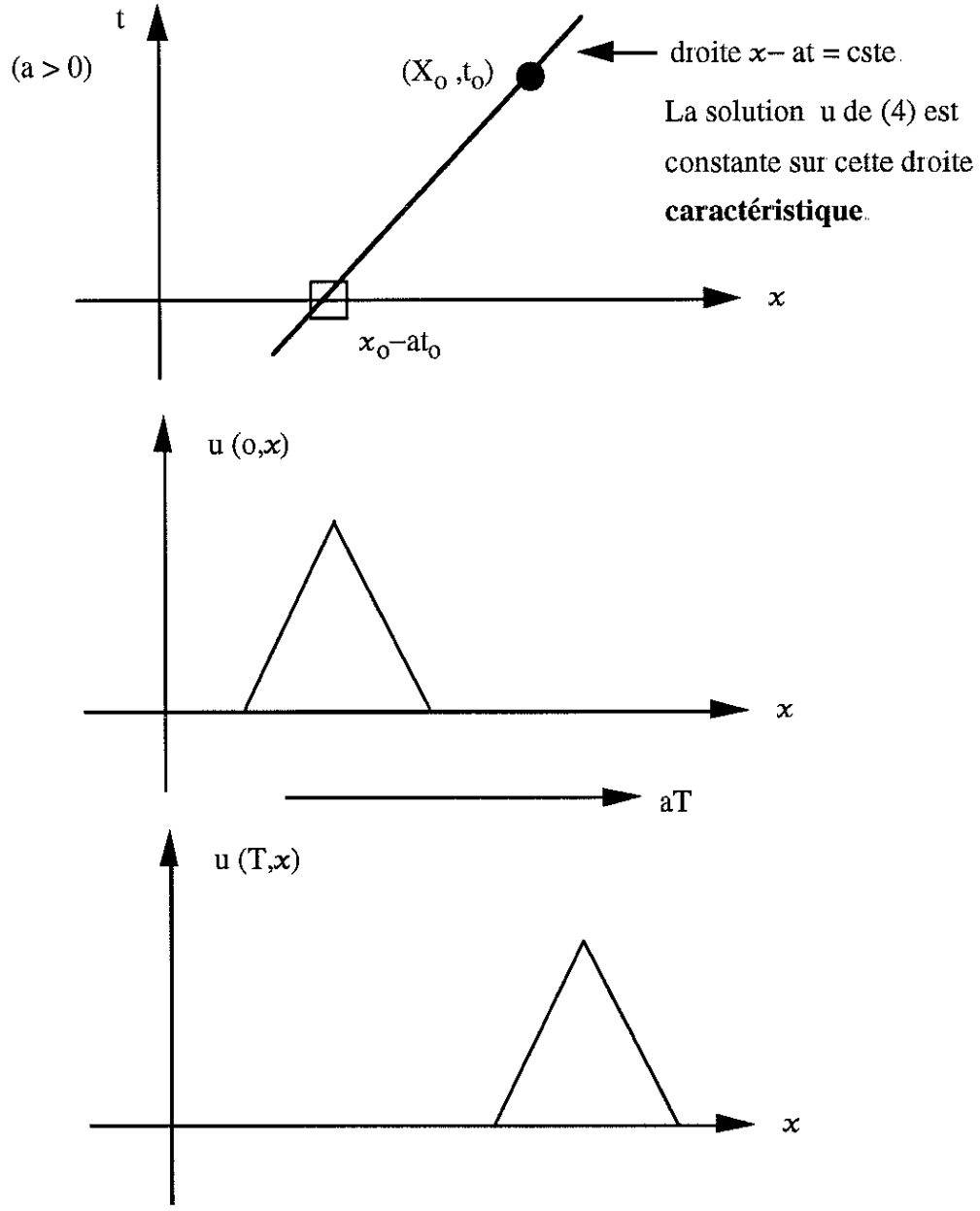
Prenant t = 0 dans la relation (6), la condition (5) est clairement vérifiée. On dérive ensuite en temps la relation (6) :

(7) $$\frac{\partial u}{\partial t}(t,x) = -a\, u'_0 (x - at)$$

puis en espace :

(8) $$\frac{\partial u}{\partial x}(t,x) = u'_0 (x - at)$$

ce qui montre la propriété ■



La solution du problème de Cauchy (4) (5) consiste à déplacer le profil u_0 (•) le long de l'axe des x à la vitesse a (advection à la vitesse a).

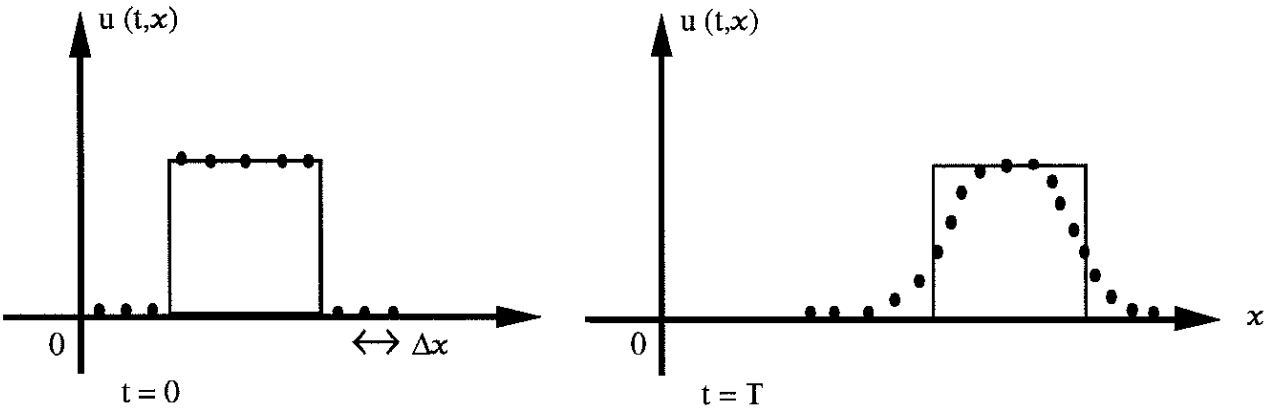
Il importe surtout de commenter la propostion précédente : l'argument de u_0 au membre de droite de la relation (6) est constant le long d'une droite de l'espace-temps d'équation.

(9) $x - at = \text{constante}$ (droite caractéristique)

qui est appelée **droite caractéristique**. L'interprétation "physique" de la relation (6) consiste à dire que pour résoudre le problème de Cauchy, c'est-à-dire calculer u (t_0, x_0) en un point (t_0, x_0) bien défini, on "remonte" la droite caractéristique d'équation (9), c'est-à-dire ici $x - at = x_0 - at_0$, jusqu'à l'origine du temps (t = 0), donc au point $x = x_0 - at_0$ (voir la figure). Le point x_0 étant fixé, on lui retranche at_0 pour arriver à l'origine du temps donc si on raisonne avec un temps croissant, la valeur initiale u_0 (X) se retrouve à l'instant t_0 au point X + at_0, ce qui correspond à une translation at_0 du profil u_0 (voir les figures). Cette interprétation étant faite, la notion d'advection (ou de transport, ou de convection) à la vitesse a justifie pleinement le nom donné à l'équation (4).

- Le fait que le problème (4) (5) soit d'une part relié à des phénomènes présents dans les fluides (transport par le courant) et d'autre part soluble exactement en fait un bon modèle mathématique pour introduire, comprendre, tester, valider les schémas aux différences finies qui sont ensuite utilisés pour des modèles fluides plus complexes dont on ne connait pas de "solution analytique".

- Notons que dès qu'une discontinuité est présente dans le profil u_0, les solutions numériques approchées sont des approximations peu satisfaisantes puisqu'il faut en général au moins cinq à six mailles pour "capturer" la discontinuité, ce avec les meilleures méthodes numériques.



Étalement d'une discontinuité lors de la résolution numérique de l'équation d'advection.

• Nous terminons cette introduction par deux propriétés des solutions de l'équation d'advection.

Proposition : Stabilité L^2 et L^∞

– Pour tout instant $t > 0$, la norme L^2 de la solution $u(t,\bullet)$, c'est-à-dire :

$$\|u(t)\|_{L^2} = \left\{ \int_{\mathbb{R}} |u(t,x)|^2 \, dx \right\}^{\frac{1}{2}} \tag{10}$$

est une constante égale à sa valeur initiale :

$$\|u(t)\|_{L^2} = \|u_0\|_{L^2} \qquad \forall t > 0 \tag{11}$$

– Si u_0 vérifie une relation de minoration et majoration du type :

$$u_* \le u_0(X) \le u^* \qquad \forall X \in \mathbb{R}, \tag{12}$$

alors il en est de même à tout instant ultérieur

$$u_* \le u(t,x) \le u^* \qquad \forall t > 0, \quad \forall X \in \mathbb{R} \tag{13}$$

La preuve de cette propriété, fondée sur des calculs élémentaires à partir de la formule de représentation (6), est laissée au lecteur.

2) Discrétisation en espace et en temps

• Nous commençons dans ce paragraphe par décrire la **méthode des lignes**, qui consiste à discrétiser l'équation (4) en espace seulement, gardant une représentation continue en temps.

$u_j(t) \simeq u(t, j\Delta x)$

Semi-Discrétisation en espace

On se contente de chercher des valeurs approchées de la solution u au point $x_j = j\Delta x$, ayant fixé au préalable un pas d'espace $\Delta x > 0$:

$$u_j(t) \simeq u(t, j\Delta x) \qquad j \in \mathbb{Z}, \quad t \ge 0 \tag{14}$$

Le problème est alors de **trouver l'équation différentielle ordinaire** qui permet de décrire l'évolution en temps de la variable discrète u_j :

$$(15) \qquad \frac{d}{dt} u_j(t) = - a \; " \frac{\partial}{\partial x} \{u_l\} " \; (x_j)$$

Le calcul du membre de droite de la relation (15) n'est pas naturel et résulte d'un **choix** du concepteur de la méthode. En effet, une dérivée en espace peut être définie à l'aide des trois limites suivantes :

$$(16) \qquad \left(\frac{\partial u}{\partial x}\right)(x_j) = \lim_{h \to 0,\, h>0} \frac{1}{h} \left[u(x_j) - u(x_j - h)\right]$$

$$(17) \qquad \left(\frac{\partial u}{\partial x}\right)(x_j) = \lim_{h \to 0,\, h>0} \frac{1}{h} \left[u(x_j + h) - u(x_j)\right]$$

$$(18) \qquad \left(\frac{\partial u}{\partial x}\right)(x_j) = \lim_{h \to 0,\, h>0} \frac{1}{2h} \left[u(x_j + h) - u(x_j - h)\right] .$$

La relation (16) introduit une différence finie **décentrée à gauche**, la relation (17) une différence finie **décentrée à droite** et la relation (18) une différence finie **centrée** autour du point x_j. Lorsqu'on regarde l'axe des x à l'échelle $h = \Delta x$, les trois expressions dont on prend la limite aux relations (16) à (18) ont un sens et sont les "meilleures" approximations des limites considérées à l'échelle Δx du maillage. Elles conduisent aux trois schémas d'approximation suivants :

$$(19) \qquad (D_- u)_j = \frac{1}{\Delta x} \left(u_j - u_{j-1}\right) \qquad \text{décentré à gauche}$$

$$(20) \qquad (D_+ u)_j = \frac{1}{\Delta x} \left(u_{j+1} - u_j\right) \qquad \text{décentré à droite}$$

$$(21) \qquad (D_o u)_j = \frac{1}{2\Delta x} \left(u_{j+1} - u_{j-1}\right) \qquad \text{centré}$$

Quand, dans le second membre de la relation (15), on remplace la dérivée partielle par rapport à x par l'un des trois schémas (19), (20) ou (21), on obtient alors un système différentiel du type :

$$(22) \qquad \frac{d}{dt} U(t) = f(U)$$

où $U = \{ u_j, j \in \mathbb{Z} \}$ est l'ensemble des variables discrètes et $(f(U))_j = -a (Du)_j$ ($D = D_+, D_-$ ou D_0 selon le choix de discrétisation en espace). Bien que le système (22) soit posé avec un nombre infini de variables, il s'agit d'un artifice mathématique pour faciliter les analyses qui vont suivre et on doit le considérer comme une équation différentielle ordinaire à un nombre fini de paramètres.

• On est alors en mesure d'introduire une échelle de temps $\Delta t > 0$ et de discrétiser en temps le système (22) à l'aide des méthodes générales introduites au chapitre 2 de ce cours. Nous nous restreignons dans ce paragraphe au cas du schéma d'**Euler explicite en temps** ; on note avec un indice supérieur (qui n'est pas un exposant !) l'approximation de u_j au temps $n\Delta t$:

(23) $$u_j^n \simeq u_j (n\Delta t) \simeq u (n\Delta t, j\Delta x)$$

et le schéma d'Euler explicite relatif à l'équation différentielle (22)

(24) $$\frac{1}{\Delta t}\left(U^{n+1} - U^n\right) = f\left(U^n\right)$$

prend l'une des trois formes suivantes, selon qu'on discrétise en espace avec l'une des trois relations (19), (20) ou (21) :

(25) $$\frac{1}{\Delta t}\left(u_j^{n+1} - u_j^n\right) + \frac{a}{\Delta x}\left(u_j^n - u_{j-1}^n\right) = 0$$ décentré à gauche

(26) $$\frac{1}{\Delta t}\left(u_j^{n+1} - u_j^n\right) + \frac{a}{\Delta x}\left(u_{j+1}^n - u_j^n\right) = 0$$ décentré à droite

(27) $$\frac{1}{\Delta t}\left(u_j^{n+1} - u_j^n\right) + \frac{a}{2\Delta x}\left(u_{j+1}^n - u_{j-1}^n\right) = 0$$ centré.

• **L'ordre de précision** de ces schémas est défini de la façon suivante. On considère l'expression $\tau (\Delta t, \Delta x)$ obtenue en replaçant dans l'une des écritures (25) à (27) u_k^m par la valeur de la solution **exacte** au point $(k\Delta t, m\Delta x)$, c'est-à-dire $u (k\Delta t, m\Delta x)$. Comme la solution **exacte** de l'équation aux dérivées partielles (4) n'a aucune raison de vérifier le schéma d'approximation, $\tau (\Delta t, \Delta x)$ est une expression non nulle dont on peut faire le développement limité pour Δt et Δx tendant vers zéro. Si $\tau (\Delta t, \Delta x)$ tend vers zéro dans ces conditions, on dit que le schéma est **consistant** avec l'équation (4). Si, de plus, on a le développement asymptotique du type :

(28) $$\tau (\Delta t, \Delta x) = 0 (\Delta t^p) + 0 (\Delta x^q) \quad (p, q \text{ entiers})$$

on dit que le schéma est d'ordre p en temps et d'ordre q en espace. On a la propriété suivante, pour les schémas (25) à (27), dont la preuve est laissée au lecteur.

Proposition

- Les schémas décentrés à gauche (25) et décentré à droite (26) sont **d'ordre 1** en espace **et** en temps.
- Le schéma centré (27) est d'ordre 2 en espace et d'ordre 1 en temps.

Afin de pouvoir critiquer les choix conduisant aux trois schémas (25) à (27), nous effectuons au paragraphe suivant une **analyse de stabilité**.

3) Analyse de stabilité par Fourier

• Cette méthode, proposée par J. Von Neumann en 1950, consiste à envisager comme condition initiale une **onde sinusoïdale** et regarder son comportement quand on lui applique le schéma numérique et quand on regarde son évolution en temps exacte à l'aide de l'équation aux dérivées partielles (4).

On considère donc une onde sinusoïdale à l'instant t_n, de la forme :

(29) $\quad u_j^n = \hat{u}(k) \exp(ikj\Delta x), \qquad j \in \mathbb{Z}$

où k est le vecteur d'onde de l'onde, $k\Delta x$ étant défini à un facteur 2Π près, compte tenu de l'expression de l'exponentielle complexe au membre de droite de la relation (29)

(30) $\quad 0 \leq k\Delta x < 2\,\Pi.$

• Lorsqu'on itère l'un des trois schémas en temps (25) (26) ou (27), avec la condition (29) au temps t_n, l'expression u_j^{n+1} est de la forme :

(31) $\quad u_j^{n+j} = g(k\Delta x, \sigma) \quad \hat{u}(k) \exp(ikj\Delta x)$

et il est donc toujours possible, lorsque le schéma numérique est **linéaire**, d'écrire le schéma numérique sous la forme :

(32) $\quad u_j^{n+1} = g(k\Delta x, \sigma) \quad u_j^n, \quad$ avec $\sigma = \dfrac{a\Delta t}{\Delta x}$

lorsque u_j^n est donné par la relaiton (29). Le coefficient g ($k\Delta x$, σ) est appelé **coefficient d'amplification** du schéma numérique.

Définition. On dit qu'un schéma numérique est **stable** (au sens de Von Neumann) lorsque son coefficient d'amplificaiton g, défini aux relations (29) et (32), est pour tout k vérifiant la relation (30), de module inférieur ou égal à 1 :

(33) $$|g(k\Delta x, \sigma)| \leq 1 \qquad \forall k,\ 0 \leq k\Delta x < 2\Pi$$

- La condition (33) exprime que l'onde (29) n'est pas amplifiée par le schéma. Si c'est le cas, la multiplication des itérations en temps conduit à la manipulation d'une onde dont l'amplitude croît exponentiellement, c'est-à-dire finalement à la manipulation dans l'ordinateur de nombres dont le module devient plus grand que le plus grand nombre représentable en machine (overflow) : c'est l'instabilité numérique, qu'il faut absolument prévenir.

La notion de stabilité numérique est purement mathématique : elle exprime que les manipulations algébriques caractéristiques du schéma ne contiennent pas de suite géométrique divergente. Cette notion est complètement indépendante de la précision du schéma, qui est plus intuitive et se ramène à des développements de Taylor. Nous pouvons donc énoncer la :

Proposition

- Le schéma décentré à gauche (25) est **stable** au sens de Von Neumann sous la condition (dite de Courant Friedrichs Lewy, ou CFL).

(34) $$\sigma \equiv a\,\frac{\Delta t}{\Delta x} \leq 1 \quad (a > 0)$$

- Les schémas décentré à droite (26) et centré (27) sont **instables** quel que soit le pas de temps Δt strictement positif.

Preuve de la proposition

- On calcule les coefficients d'amplification des trois schémas. Lorsqu'on change j en j+1 dans l'expression (29), on multiplie u_j^n par le coefficient exp ($ik\,\Delta x$) ; cette remarque permet de calculer très rapidement les coefficients d'amplification des schémas.

- Pour le schéma décentré à gauche, on a :

(35) $g(k\Delta x,\sigma) = 1 - \sigma\left(1 - e^{-ik\Delta x}\right)$

Alors $|g|^2 = \left(1 - 2\sigma \sin^2 \frac{k\Delta x}{2}\right)^2 + \sigma^2 \sin^2 k\Delta x$

$$= 1 - 4\sigma \sin^2 \frac{k\Delta x}{2} + 4\sigma^2 \sin^4 \frac{k\Delta x}{2} + 4\sigma^2 \sin^2 \frac{k\Delta x}{2} \cos^2 \frac{k\Delta x}{2}$$

(36) $$|g|^2 = 1 - 4\sigma (1-\sigma) \sin^2 \frac{k\Delta x}{2}$$

et l'expression au membre de droite de (36) n'est inférieure à 1 que si $\sigma (1-\sigma) \geq 0$, c'est-à-dire sous la condition (34) dès que $a > 0$ et $\Delta t > 0$. Ceci prouve le premier point de la proposition.

- Pour le schéma décentré à droite, on a :

$$g = 1 - \sigma\left(e^{ik\Delta x} - 1\right)$$

et en prenant un nombre d'onde k tel que $k\Delta x = \Pi$, on a dans ce cas :

$$g\ (k\Delta x = \Pi) = 1 + 2\sigma > 1 \quad \text{dès que} \quad \sigma > 0$$

ce qui montre que ce schéma est instable.

- Pour le schéma centré en espace (27), le coefficient d'amplificaiton prend la forme suivante :

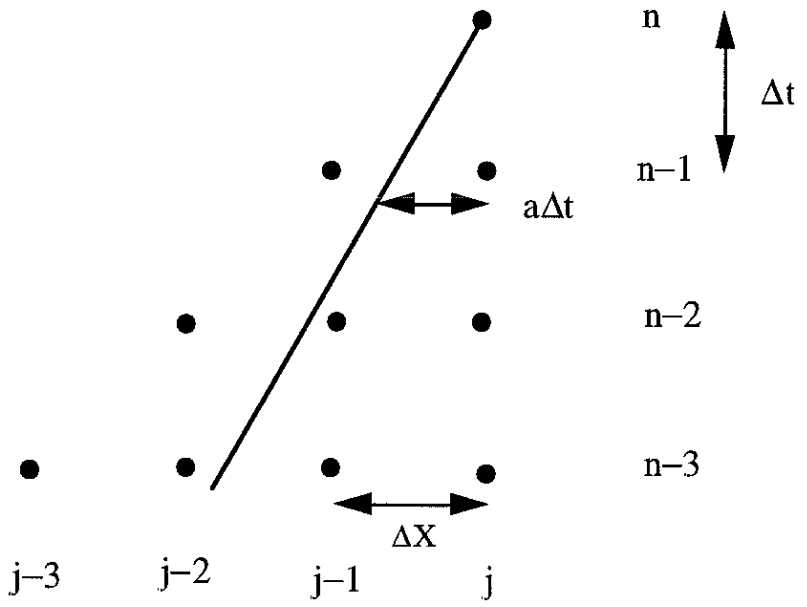
(37) $$g\ (k\Delta x,\sigma) = 1 - i\sigma \sin (k\Delta x)$$

et pour tout k, cette expression est un nombre complexe de module strictement supérieur à 1 si σ n'est pas nul.

- **Cône de dépendance numérique**

Afin de se donner une idée intuitive de la condition de stabilité (34), on construit, pour le schéma décentré (25) le cône de dépendance numérique entre les instants n et n–m (m entier ≥ 0), c'est-à-dire l'ensemble des noeuds $\left((j+\ell)\,\Delta x,\ (n-k)\,\Delta t\right)\ \left(0 \le k \le m\right)$ dont la valeur est nécessaire pour calculer u_j^n.

Lorsque m = 3 par exemple le cône de dépendance numérique du schéma (25) est donné par le graphe suivant :



Cône de dépendance numérique du schéma décentré à gauche (25).

On peut superposer à ce graphe la droite caractéristique (9) passant par le point (jΔx, nΔt). La condition de stabilité (34) exprime alors que pour l'instant (n–1) Δt, cette droite passe "entre" les noeuds (j–1) Δx et jΔx du maillage en espace. De façon plus générale, la droite caractéristique est située à l'intérieur du cône de dépendance numérique relatif au schéma décentré à gauche lorsque la relation (34) est vérifiée.

Nous retiendrons la condition nécessaire (pratique) suivante.

Caractéristique et cône de dépendance numérique

Pour qu'un schéma soit stable, il est nécessaire que le cône de dépendance numérique contienne la droite caractéristique.

Cette condition permet très rapidement d'illustrer la non stabilité du schéma décentré à droite : le schéma a un cône de dépendance numérique formé d'états tous situés "à droite" du noeud ($j\Delta x$, $n\Delta t$), c'est-à-dire de la forme $\big((j+k)\Delta x, (n-m)\Delta t\big)$ $k \geq 0$ alors que la droite caractéristique, qui porte l'information mathématiquement (physiquement ?) nécessaire pour connaître la valeur de la solution au point ($j\Delta x$, $n\Delta t$), vient de la gauche ! Il n'est pas possible d'interpoler les données de la grille partielle de dépendance pour évaluer de façon approchée la valeur du champ inconnu sur la droite caractéristique.

- La condition précédente n'est que **nécessaire**. En effet, dans le cas du schéma centré instable (27) (voir son coefficient d'amplification en (37)), la condition de stabilité (34) permet de placer la droite caractéristique aboutissant au noeud ($j\Delta x$,$n\Delta t$) à l'intérieur du cône de dépendance numérique de ce schéma (le dessin est laissé au lecteur), alors que le schéma demeure instable que que soit $\Delta t > 0$! On mesure bien sûr cet exemple, combien la notion de stabilité numérique est une notion purement mathématique dont une représentation intuitive comme la condition géométrique de cône de dépendance ne donne qu'une image imparfaite.

4) Quelques propriétés de deux schémas classiques

Dans la suite de ce chapitre, nous ne conservons que le schéma décentré à gauche (25), vu que les schémas (26) et (27) sont instables pour tout pas de temps $\Delta t > 0$. Nous remarquons toutefois que si $a < 0$, il convient de prendre le schéma décentré à droite et ne pas utiliser le schéma décentré à gauche (exercice laissé au lecteur !). Le choix d'un schéma décentré impose de connaître le sens de l'écoulement.

- Nous introduisons maintenant le schéma de Lax Wendroff (1960), qui s'obtient par un procédé dual de celui proposé au second paragraphe de ce chapitre : on discrétise d'abord le temps (l'espace restant continu) puis on forme les opérateurs aux différences centrées les plus simples pour calculer de façon approchée les dérivées partielles en espace.

On part du développement de Taylor suivant :

(37) $$u^{n+1} = u^n + \Delta t\left(\frac{\partial u}{\partial t}\right)^n + \frac{1}{2}(\Delta t)^2\left(\frac{\partial^2 u}{\partial t^2}\right)^n + 0\,(\Delta t^3)$$

que l'on va tronquer pour définir le schéma, et on remplace les dérivées en temps par des dérivées en espace à l'aide de l'équation (4) :

(38) $$\frac{\partial u}{\partial t} = -a\frac{\partial u}{\partial x}$$

(39) $$\frac{\partial^2 u}{\partial t^2} = a^2\frac{\partial^2 u}{\partial x^2}$$

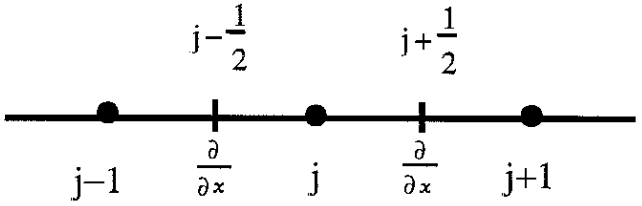
(40) $$u^{n+1} = u^n - a\Delta t\left(\frac{\partial u}{\partial x}\right)^n + \frac{1}{2}(a\Delta t)^2\left(\frac{\partial^2 u}{\partial x^2}\right)^n$$

On remplace ensuite les dérivées partielles en espace où à la relation (40) par des différences finies **centrées.**

(41) $$\left(\frac{\partial u}{\partial x}\right)_j^n \simeq \frac{1}{2\Delta x}\left(u_{j+1}^u - u_{j-1}^u\right)$$

(42) $$\left(\frac{\partial^2 u}{\partial x^2}\right)_j^n \simeq \frac{1}{(\Delta x)^2}\left(u_{j+1}^u - 2u_j^u + u_{j-1}^n\right)$$

On remarque que ces deux approximations sont précises au **second ordre en espace**, et l'expression (42) donnant la dérivée seconde au noeud $j\Delta x$ s'obtient en introduisant d'abord une approximation de la dérivée première entre les deux noeuds j et (j+1) :



(43) $$\left(\frac{\partial u}{\partial x}\right)_{j+\frac{1}{2}} \simeq \frac{1}{\Delta x}\left(u_{j+1} - u_j\right)$$

et en utilisant ensuite cette expression du gradient de part et d'autre du sommet j :

(44) $$\left(\frac{\partial^2 u}{\partial x^2}\right)_j \simeq \frac{1}{\Delta x}\left\{\left(\frac{\partial u}{\partial x}\right)_{j+\frac{1}{2}} - \left(\frac{\partial u}{\partial x}\right)_{j-\frac{1}{2}}\right\}$$

Ceci correspond finalement à l'expression (42)

Le schéma de Lax-Wendroff est donc défini par la relation :

$$(45) \qquad u_j^{n+1} = u_j^n - \frac{a\Delta t}{2\Delta x}\left(u_{j+1}^n - u_{j-1}^n\right) + \frac{1}{2}\left(\frac{a\Delta t}{\Delta x}\right)^2\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right).$$

• L'intérêt essentiel du schéma de Lax-Wendroff réside en sa grande **précision** spatio temporelle :

Proposition. Le schéma (45) est précis au second ordre en temps et en espace.

De plus, il possède de bonnes propriétés de **stabilité**.

Proposition. Le schéma de Lax-Wendroff (45) est stable au sens de Von Neumann sous la condition CFL.

$$(46) \qquad \sigma^2 \le 1 \text{ ; où } \sigma = \frac{a\Delta t}{\Delta x}.$$

• Pour démontrer cette dernière proposition (la preuve de la proposition précédente est laissée au lecteur), on calcule le coefficient d'amplification, en adoptant la notation classique $\xi = k\Delta x$.

$$(47) \qquad g(\xi,\sigma) = 1 - i\,\sigma \sin\xi + \sigma^2(\cos\xi - 1).$$

Le calcul du carré du module de g est alors facile :

$$|g|^2 = \left[1 - \sigma^2(1 - \cos\xi)\right]^2 + \sigma^2 \sin^2\xi$$

$$= 1 - \sigma^2\left(2 - 2\cos\xi - \sin^2\xi\right) + \sigma^4(1 - \cos\xi)^2$$

$$= 1 - \sigma^2\left(1 - 2\cos\xi + \cos^2\xi\right) + \sigma^4(1 - \cos\xi)^2$$

$$(48) \qquad |g|^2 = 1 - \sigma^2(1 - \sigma^2)(1 - \cos\xi)^2$$

Lorsque ξ varie de 0 à 2Π, le coefficient $(1 - \cos\xi)^2$ varie de 0 à 4, et $|g|^2$ ne reste plus petit que 1 que si $(1 - \sigma^2)$ est positif, ce qu'exprime exactement la condition (46).

Proposition : Stabilité L^∞

Le schéma décentré (25) est, sous la condition CFL (34), stable L^∞, alors que même sous la condition analogue (46), le schéma de Lax-Wendroff (45) ne l'est pas.

- La stabilité L^∞, définie aux relations (12) et (13) est simple à établir pour le schéma décentré, qui peut s'écrire sous la forme :

(49) $$u_j^{n+1} = (1-\sigma)\, u_j^n + \sigma\, u_{j-1}^n$$

c'est-à-dire comme **combinaison convexe** ($\sigma \leq 0$, $1-\sigma \geq 0$) des points du cône de dépendance numérique. Si $u_* \leq u_k^n \leq u^*$ pour tout $k \in \mathbb{Z}$, alors u_j^{n+1} satisfait clairement cette propriété compte tenu de l'expression (49) qui met en jeu des coefficients positifs pour u_j^n et u_{j-1}^n, et ceci est vrai quel que soit j.

- Dans le cas du schéma de Lax-Wendroff, on a :

(50) $$u_j^{n+1} = (1-\sigma^2)\, u_j^n - \frac{1}{2}\,\sigma\,(1-\sigma)\, u_{j+1}^n + \frac{\sigma}{2}\,(1+\sigma)\, u_{j-1}^n$$

Pour $0 \leq \sigma \leq 1$, le coefficient de u_{j+1}^n est négatif, et pour $-1 \leq \sigma \leq 1$, celui de u_{j-1}^n est négatif. Ceci ne constitue pas une preuve de la non stabilité du schéma de Lax Wendroff, qui résulte d'un résultat général de V. Thomée (non démontré ici) qu'un schéma stable L^∞ est d'ordre impair. ■

- **Erreur de phase**

Le coefficient d'amplification g mesure la façon dont le schéma modifie une onde sinusoïdale. Si on a une condition initiale du type :

(51) $$u_0(x) = \exp(ikx)$$

la solution de l'équation d'advection après un temps Δt s'écrit :

(52) $$u^{ex}(\Delta t) = e^{-ika\Delta t}\, u_0(x)$$

ce qui conduit à introduire le coefficient d'amplification exact relatif à **l'équation** (4) et non plus au schéma numérique. En introduisant le nombre de courant σ, on a donc :

(53) $$g^{ex} = e^{-ik\sigma\Delta x}$$

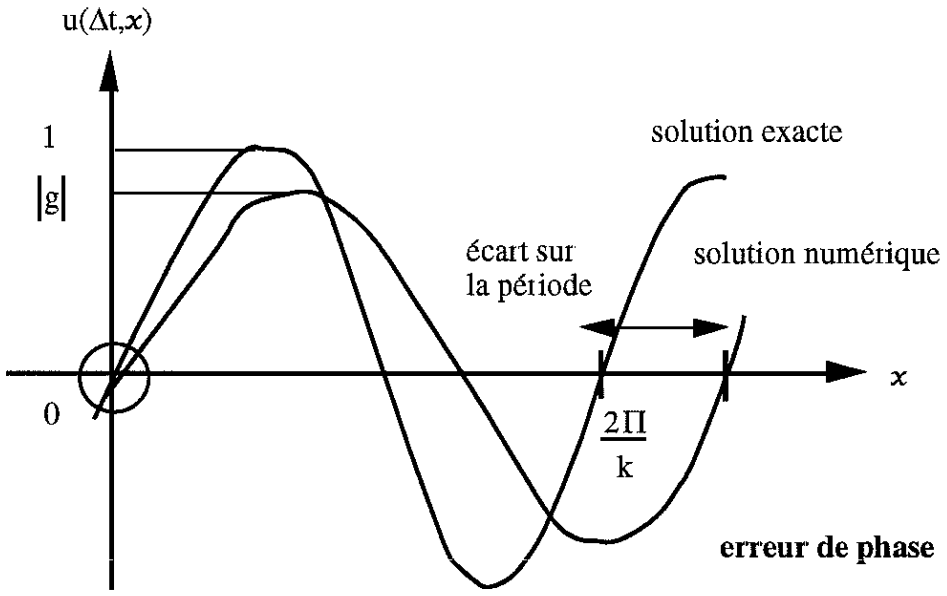
Nous écrivons par ailleurs le coefficient d'amplification du schéma numérique en le décomposant en module et argument.

(54) $g = |g| \exp -i\theta$.

La condition de stabilité $|g| \leq 1$ est "automatiquement" satisfaite pour le coefficient exact introduit en (53) qui est de module toujours égal à 1.

Le rapport entre la solution approchée calculée par le schéma numérique et la solution exacte (52) est donc, pour une onde sinusoïdale, donnée par l'expression suivante :

(55) $$\frac{u^{sh}(\Delta t)}{u^{ex}(\Delta t)} = |g| \exp -i\,(\theta - k\sigma\Delta x).$$



L'argument $(\theta - k\sigma\Delta x)$ de l'exponentielle complexe au second membre de (55) est appelé **erreur de phase** du schéma numérique. Il mesure l'écart entre la période du signal introduit comme condition initiale (51) et le signal reçu grâce au schéma numérique. Si $\theta - k\sigma\Delta x$ est positif, l'onde numérique "va plus vite" que l'onde physique et elle va moins vite si $\theta - k\sigma\Delta x < 0$, ce qui est par exemple le cas de la figure ci-dessus.

Proposition : Erreur de phase pour $k\Delta x$ petit

Pour le schéma décentré amont (25), l'erreur de phase admet le développement limité suivant :

$$(56) \qquad \theta - k\sigma\Delta x = -\frac{1}{6}\sigma(1-\sigma)(1-2\sigma)(k\Delta x)^3 + 0\left((k\Delta x)^5\right)$$

et pour le schéma de Lax-Wendroff (45), nous avons :

$$(57) \qquad \theta - k\sigma\Delta x = -\frac{1}{6}\left(1+\frac{1}{2}\sigma^2\right)(k\Delta x)^3 + 0\left((k\Delta x)^5\right).$$

Exercice. Le calcul de u_j^{n+1} avec le schéma décentré (respectivement de Lax-Wendroff) revient à interpoler la valeur au pied de la caractéristique :

$$(58) \qquad x = x_j - a\Delta t \qquad (a\Delta t \leq \Delta x)$$

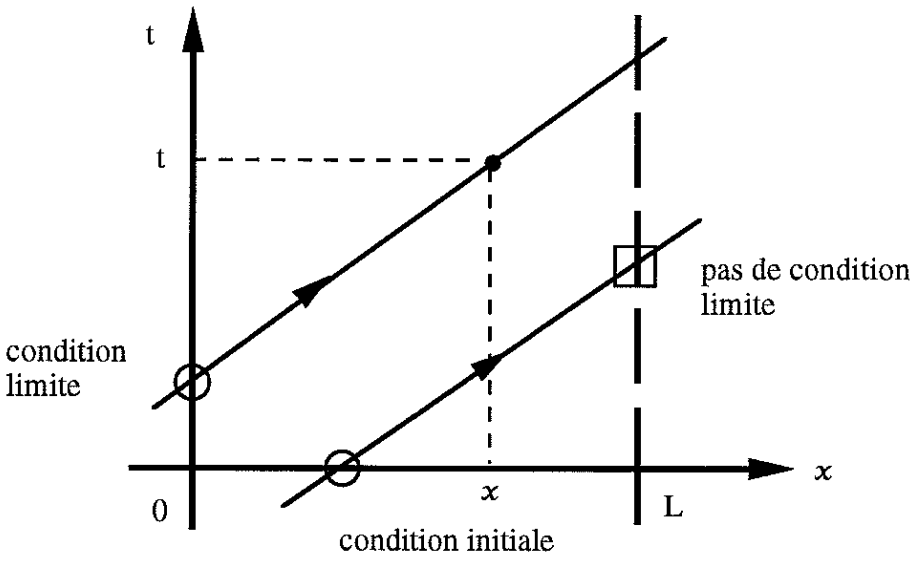
à l'aide d'une interpolation affine entre les valeurs u_{j-1}^n et u_{j+1}^n (respectivement une interpolation parabolique à partir des trois valeurs u_{j-1}^n, u_j^n, u_{j+1}^n).

5) Problème à valeur initiale et à la limite

• Nous avons étudié dans ce chapitre le problème de Cauchy pour l'équation d'advection, c'est-à-dire l'équation (4) avec une variable d'espace x décrivant $\mathbb{R}$ pour entier. Dans la pratique, la variable d'espace est bornée, et nous devons donc nous intéresser au problème dit "IBVP" (Initial Boundary Value Problem) c'est-à-dire avec valeur à la **limite** et valeur **initiale** :

$$(59) \qquad \begin{cases} \dfrac{\partial u}{\partial t} + a\dfrac{\partial u}{\partial x} = 0 \qquad t>0 \qquad 0<x<L \\ u(o,x) = u_o(x) \qquad 0<x<L \\ \text{condition limite en } x=0 \text{ et/ou } x=L \end{cases}$$

Le jeu de condition limite à imposer en $x = 0$ et $x = L$ se devine facilement à partir de la connaissance des caractéristiques : la solution u est constante le long d'une droite caractéristique du type $x - at = cste$.



"IBVP" pour l'équation d'advection

Partant de $(t > 0,\ x \in]O,L[)$, on "remonte" vers le passé la droite caractéristique passant par ce point et on finit par rencontrer l'axe des x (t=0) ou l'axe des temps (x=0). Dans le premier cas, u(t,x) est égal à la valeur initiale u_0 (relation (6)) et dans le second cas, u(t,x) est égal à la valeur limite v_0 en x=0 pour le temps $t - \frac{x}{a}$:

$$(60) \qquad u(t,x) = v_0\left(t - \frac{x}{a}\right) \qquad 0 < at < x < L.$$

• Le problème IBVP (59) doit donc être mathématiquement (nous ne l'avons pas démontré ici !) et physiquement (nous venons de l'illustrer) posé sous la forme suivante :

$$(61) \quad \begin{cases} \dfrac{\partial u}{\partial t} + a\dfrac{\partial u}{\partial x} = 0 & t > 0,\ \ 0 < x < L \\ u(o,x) = u_o(x) & 0 < x < L \quad \text{condition initiale (t=0)} \\ u(t,o) = v_o(t) & t > 0 \qquad \text{condition limite en } x = 0 \end{cases}$$

Nous remarquons qu'**aucune condition limite** n'est nécessaire **en x = L**, partie de la frontière où la droite caractéristique x = at = cste "sort" du domaine de calcul (si on la suit en l'orientant vers les temps croissants !).

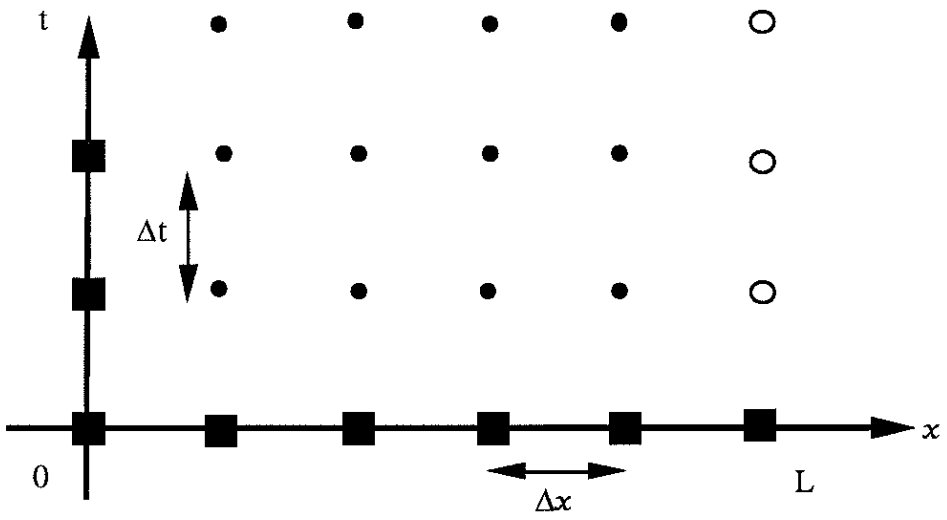
Nous nous attendons à ce que tout schéma numérique doive prendre en compte cette particularité de l'équation d'advection.

• Nous discrétisons l'espace temps $0 < x < L$, $t > 0$ à l'aide d'un pas d'espace compatible avec la géométrie :

(62) $\Delta x = \frac{L}{J}$ J entier

et un pas de temps compatible avec la condition de Courant-Friedrichs-Lewy :

(63) $a\Delta t \leq \Delta x$



En t = 0, la valeur aux points de grille est donnée à l'aide de la condition initiale :

(64) $u_j^o = u_o(j\Delta x)$ $0 \leq j \leq J$

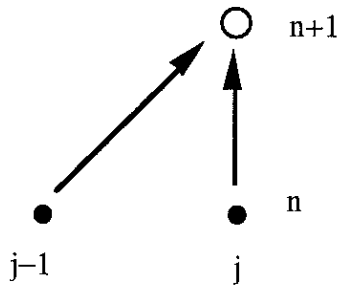
et en $x = 0$, on utilise la condition limite entrante :

(65) $u_o^n = v_o(n\Delta t)$ $n \geq 0$

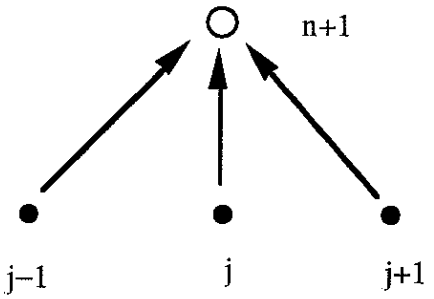
[on remarque que la condition de continuité $u_o(0) = v_o(0)$ est nécesaire pour donner un sens non ambigu aux relations (64) et (65) pour j = 0 et n = 0 respectivement].

Pour les points de grille "intérieurs" au domaine de calcul (avec un point noir sur le graphe précédent), le schéma numérique permet de calculer un pas de temps dès que toutes les valeurs sont calculées au pas de temps antérieur. Par contre, en $x = L$ (ou j = J), l'abscence de point de grille "à droite" de ces points limites impose d'analyser avec soin ce qui peut arriver.

• Dans le cas du schéma décentré (25), seuls les points j et j−1 sont nécessaires à l'instant n pour itérer le schéma. Donc en j = J (au bord du domaine), il n'y **a pas de problème** particulier. Le schéma décentré "suit" la physique en allant chercher l'information numérique dans la bonne direction.



• Dans le cas du schéma de Lax Wendroff (45), les trois points j−1, j, j+1 sont utilisés pour itérer le schéma. **On ne peut pas l'utiliser sans modification à la limite.**



On peut ajouter une "condition limite numérique" en x=L, qui enrichit la physique sans raison particulière, en "prétraitant" une valeur en j = J+1, pour calculer ensuite avec le schéma de Lax Wendroff entre les instants n et n+1. On peut aussi **changer de schéma** pour le calcul du seul point limite j = J ; le "schéma à la limite" (par exemple le schéma décentré à gauche vu ce qui a été dit plus haut) doit être compatible avec le "schéma intérieur" (ici Lax Wendroff). L'analyse de la **stabilité** pour le problème à condition limite et initiale est **délicate** et récente (Gustafsson-Kreiss-Sundström, 1972) ; les critères de stabilité du schéma global sont encore du domaine des spécialistes !

IV. DIFFÉRENCES FINIES POUR L'ÉQUATION DE LA CHALEUR À UNE DIMENSION D'ESPACE

1) Modèle physico-mathématique

Dans ce chapitre, nous abordons l'étude d'un modèle de diffusion. Le laplacien du champ décrit directement l'évolution en temps du processus. De façon plus générale, étant donné un domaine Ω de $\mathbb{R}^n$ ($n \geq 1$ au moins au tout début de ce paragraphe, $n = 1$ ensuite), on cherche un champ scalaire u (la température) fonction de $x \in \Omega$ et du temps t.

(1) $\quad [O,T[\times \Omega \quad \ni (t,x) \rightarrow u(t,x) \in \mathbb{R}$

solution de l'équation "de la chaleur" :

(2) $\quad \dfrac{\partial u}{\partial t} - \operatorname{div}(k \nabla u) = f \qquad (t,x) \in]O,T[\times \Omega$

Le second membre f est la source volumique de chaleur (flamme, radiateur) et k le coefficient de diffusion thermique (qui peut être une matrice pour un matériau anisotrope) qui relie le gradient de la température au flux de chaleur q :

(3) $\quad q = -k \nabla u, \quad (t,x) \in]O,T[\times \Omega$

On pose le problème sur un intervalle de temps]O,T[, mais en pratique T peut être arbitrairement grand lorsqu'on recherche un régime permanent ou entretenu régulièrement.

En plus de l'équation aux dérivées partielles (2), on doit aussi se donner une condition initiale u_0 :

(4) $\quad u(o,x) = u_0(x) \qquad x \in \Omega$

ainsi qu'une condition limite en tout point du bord $\partial\Omega$; on se bornera ici à une condition de Dirichlet sur une portion Γ_1 de la frontière :

(5) $\quad u(t,x) = v_0(t) \qquad t \in]O,T[, \quad x \in \Gamma_1$

et à une condition de Neumann de flux imposé sur la partie complémentaire Γ_2 :

(6) $\quad -k \dfrac{\partial u}{\partial n}(t,x) = g(t), \qquad t \in]O,T[, \quad x \in \Gamma_2$

(7) $\quad \partial\Omega = \Gamma_1 \cup \Gamma_2 \ ; \ \Gamma_1 \cap \Gamma_2 = \varnothing$

• Dans le cas d'un barreau infini ($\Omega = \mathbb{R}$), le bord de Ω est rejeté à l'infini (pas de condition limite) et le problème de Cauchy (2) (4) a une solution unique $u(t,x)$ qui peut se représenter à l'aide d'une intégrale :

$$(8) \qquad u(t,x) = \int_{\mathbb{R}} \frac{1}{\sqrt{2\Pi t}} \exp - \left(\frac{|x-x'|^2}{4t} \right) u_o(x')\, dx'$$

Le noyau de Green $G(t,x)$,

$$(9) \qquad G(t,x) = \frac{1}{\sqrt{2\Pi t}} \exp - \left(\frac{|x|^2}{4t} \right)$$

est la solution élémentaire de l'équation de la chaleur avec un coefficient de diffusion $k \equiv 1$, c'est-à-dire est solution du problème suivant :

$$(10) \qquad \frac{\partial u}{\partial t} - \Delta u = 0$$

$$(11) \qquad u(o,x) = \delta(x)$$

où la "fonction" δ, ou plus justement la "masse de Dirac" δ est une distribution qui opère sur les fonctions $\varphi : \mathbb{R} \to \mathbb{R}$ de la façon suivante :

$$(12) \qquad \langle \delta, \varphi \rangle = \varphi(o)$$

On peut dire aussi que $\delta(x)$ est un "pic" valant 0 sauf en $x=0$ où il prend une valeur infinie de sorte que la relation (12) soit vérifiée, en identifiant $\langle \delta, \varphi \rangle$ à l' "intégrale" $\int_{\mathbb{R}} \delta(x)\, \varphi(x)\, dx$.

Il convient de retenir de ce qui précède que si on se contente d'une source de chaleur de "masse" unité en $t = 0$, placée en $x = 0$, alors la solution du problème (10 (11) est donnée par la relation (9). De plus, si on remplace la condition particulière (11) par la condition générale (4), le problème (10) (4) a pour solution la température u donnée par la relation (8), qui est une convolution du noyau de Green (9) par la donnée initiale (4) :

$$(13) \qquad u = G * u_o$$

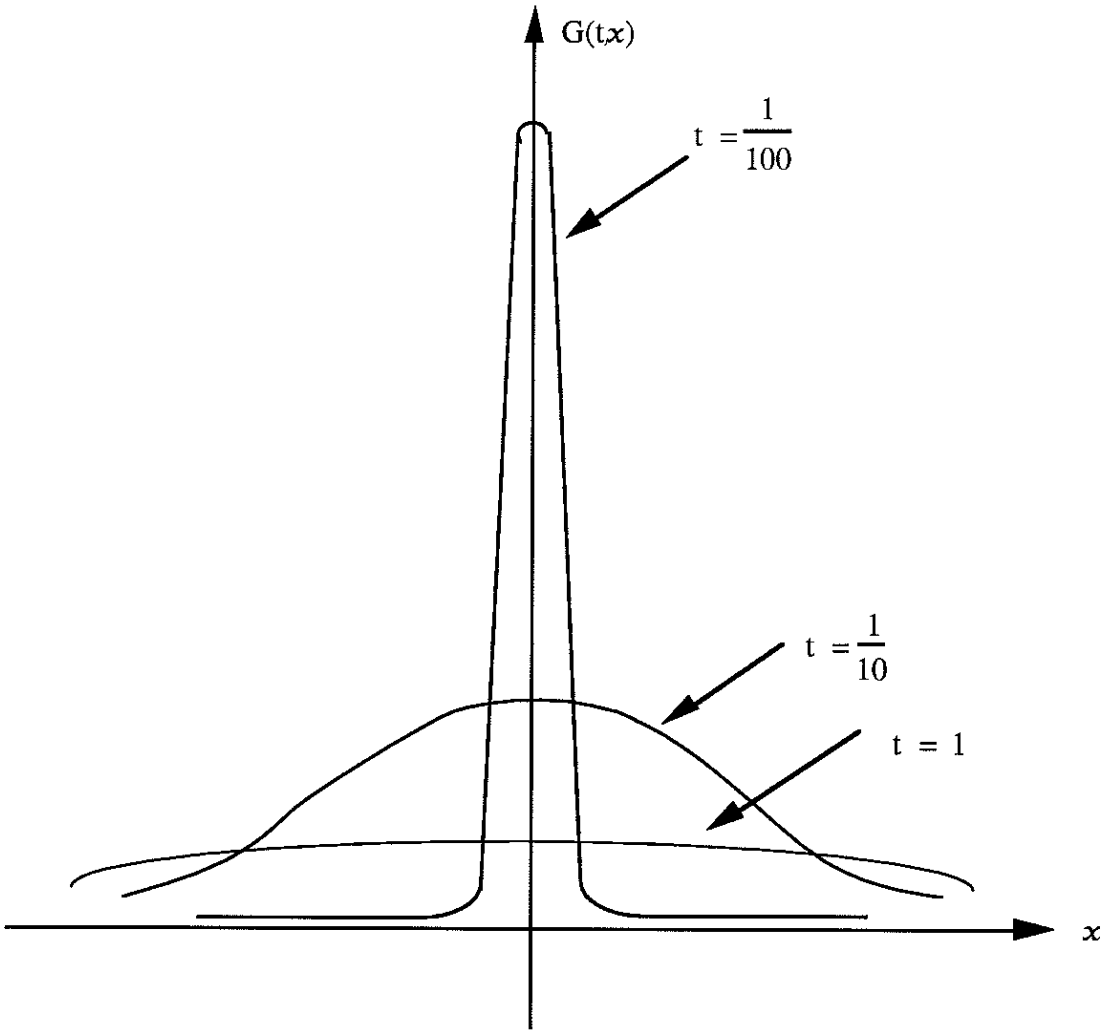


Figure 1 : Noyau de Green de l'équation de la chaleur à différents instants.

La représentation graphique du noyau de Green Gaussien (figure 1) montre que dès que $t > 0$, la fonction $G(t,\bullet)$ n'est pas nulle, ce qui indique que la donnée initiale (11), s'est propagée infiniment loin en un temps arbitrairement petit. Nous retenons que la **vitesse de propagation** de l'information est **infinie** pour l'équation de la chaleur (10).

Notons qu'avec le modèle d'advection étudié au chapitre 3, la propagation s'effectuait à vitesse finie, ce qui constitue la principale différence mathématique entre les modèles. L'équation d'advection est le prototype de ce que les mathématiciens appellent un problème hyperbolique alors que l'équation de la chaleur est un modèle dit **parabolique**.

A partir de la représentation intégrale (8), il est facile de tirer plusieurs propriétés de la solution de l'équation de la chaleur, qu'on peut ensuite chercher à retrouver lors de la résolution approchée.

Proposition : Principe du maximum.

On a les implications suivantes :

(14) $u_0 \geq 0 \Rightarrow u(t,x) \geq 0 \qquad \forall t, \forall x$

(15) $u_* \leq u_0 \leq u^* \Rightarrow u_* \leq u(t,x) \leq u^*, \qquad \forall t, \forall X$

Proposition : Stabilité L^2

Nous notons $\| \ \|_0$ la norme L^2 d'une fonction réelle :

(16) $$\|w\|_0 = \left(\int_{\mathbb{R}} w^2(x)\, dx \right)^{\frac{1}{2}}$$

La solution du problème (10) (4) vérifie :

(17) $\|u(\bullet,t)\|_0 \leq \|u_0\|_0$

2) Schéma aux différences explicite à une dimension

Nous prenons un second membre f identiquement nul, nous plaçons sur une barre infinie ($\Omega = \mathbb{R}$), et cherchons donc à approcher numériquement l'équation de diffusion suivante :

(18) $$\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left(k \frac{\partial u}{\partial x} \right) = 0 \qquad t > 0, \quad x \in \mathbb{R}$$

Nous approchons la relation (18) par un schéma à deux niveaux en temps :

(19) $$\frac{\partial u}{\partial t} \simeq \frac{1}{\Delta t}\left(u_j^{n+1} - u_j^n \right) \qquad n \geq 0, \quad j \in \mathbb{Z}$$

où u_j^n désigne l'approximation classique

(20) $u_j^n \simeq u(n\Delta t, j\Delta x)$

De plus, nous utilisons d'abord un schéma explicite, c'est-à-dire :

(21) $$-\frac{\partial}{\partial x}\left(k \frac{\partial u}{\partial x} \right) \simeq -\frac{\partial}{\partial x}\left(k \frac{\partial u}{\partial x} \right)(t^n)$$

• La discrétisation de l'opérateur (21) par un schéma aux différences est très classique (figure 2).

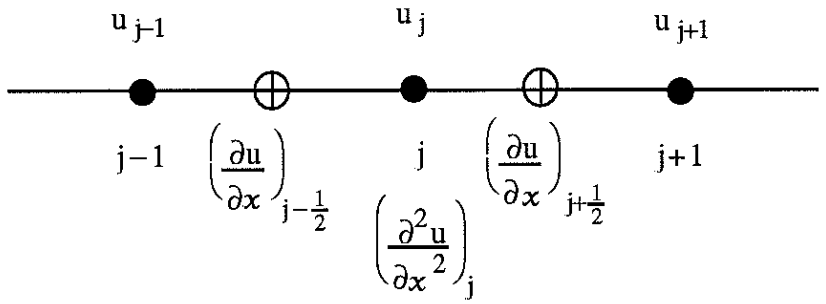


Figure 2 : Différences finies centrées pour l'approximation d'une dérivée seconde.

On a d'abord une approximation centrée de la dérivée première entre les deux points de grille j et j+1.

$$k\left(\frac{\partial u}{\partial x}\right)_{j+\frac{1}{2}} = \frac{k}{\Delta x}\left(u_{j+1} - u_j\right) + 0\,(\Delta x^2) \tag{22}$$

avec une relation analogue en $\left(j-\frac{1}{2}\right)$:

$$k\left(\frac{\partial u}{\partial x}\right)_{j-\frac{1}{2}} = \frac{k}{\Delta x}\left(u_j - u_{j-1}\right) + 0\,(\Delta x^2) \tag{23}$$

On effectue ensuite la différence entre les relations (22) et (23) pour approcher l'opérateur du second ordre (21) :

$$-\frac{\partial}{\partial x}\left(k\frac{\partial u}{\partial x}\right)_j (x_j) = \frac{1}{\Delta x}\left\{\frac{k}{\Delta x}\left(u_j - u_{j-1}\right) - \frac{k}{\Delta x}\left(u_{j+1} - u_j\right)\right\} + 0\,(\Delta x^2) \tag{24}$$

$$-\frac{\partial}{\partial x}\left(k\frac{\partial u}{\partial x}\right)_j (x_j) \simeq \frac{k}{\Delta x^2}\left(-u_{j+1} + 2u_j - u_{j-1}\right) \tag{25}$$

On aboutit, compte tenu des choix antérieurs (19) et (21), au schéma explicite suivant :

$$\frac{1}{\Delta t}\left(u_j^{n+1} - u_j^n\right) + \frac{k}{\Delta x^2}\left(-u_{j+1}^n + 2\,u_j^n - u_{j-1}^n\right) = 0 \tag{26}$$

dont le graphe de dépendance est illustré figure 3. Connaissant l'ensemble des états à l'instant $n\Delta t$, le calcul de chacun des états à l'instant ultérieur $(n+1)\,\Delta t$ est directement fourni par la relation (26).

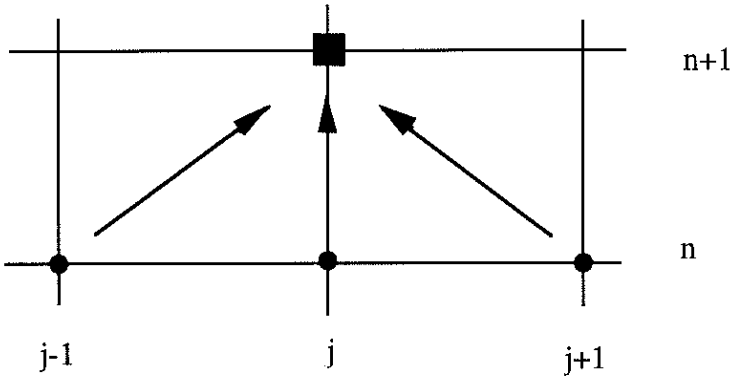


Figure 3 : Graphe de dépendance du schéma explicite (26).

Proposition

Le schéma (26) approche l'équation de la chaleur (18) au **premier ordre en temps** et au **second ordre en espace.**

La preuve résulte d'un développement de Taylor de la relation au membre de gauche de (26), quand on l'applique aux valeurs de grille d'une solution de l'équation (18). Elle est laissée au lecteur.

Proposition : Stabilité

On introduit deux coefficients sans dimension α et ξ, définis pour une onde exp (i p x) par les relations suivantes :

(27) $$\alpha = k\frac{\Delta t}{\Delta x^2}$$

(28) $$\xi = p\Delta x.$$

Le coefficient d'amplification $g(\alpha, \xi)$ du schéma (26) s'écrit :

(29) $$g(\alpha,\xi) = 1 - 4\,\alpha \sin^2 \frac{\xi}{2},$$

la condition de stabilité de Von Neumann :

(30) $$\forall\, \xi \in [0, 2\Pi]\,,\ |g(\alpha,\xi)| \leq 1$$

s'écrit dans ce cas :

(31) $$\alpha \leq \frac{1}{2}.$$

- **Développement d'une instabilité**

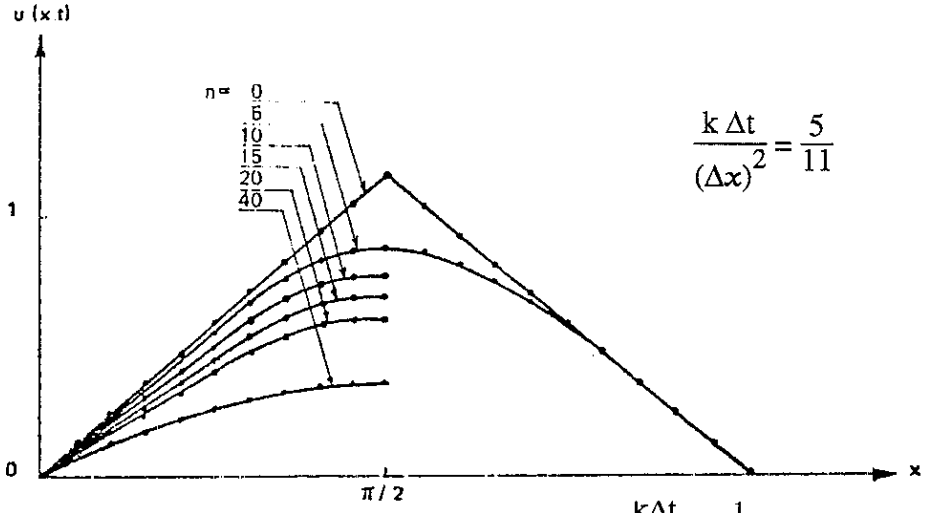


Figure 3 : Stabilité du schéma pour $\frac{k\Delta t}{\Delta x^2} \leq \frac{1}{2}$.

Choisissons maintenant Δt tel que $\frac{\Delta t}{\Delta x^2} = \frac{5}{9} \geq \frac{1}{2}$. On représente les solutions approchées pour n = 5, 10, 15.

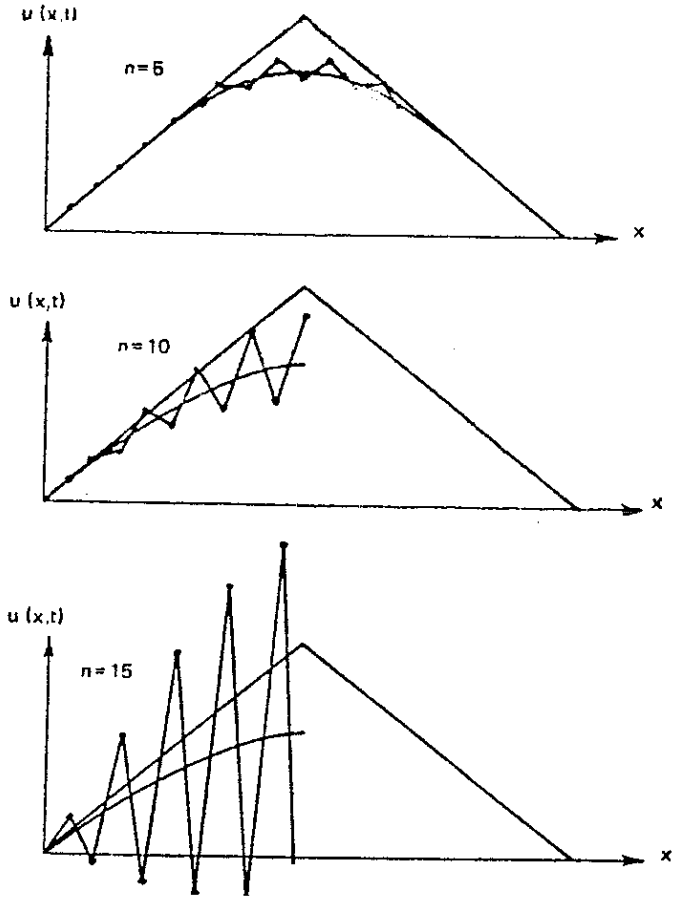


Figure 4 : Instabilité du schéma pour $\frac{\Delta t}{\Delta x^2} > \frac{1}{2}$.

3) θ-schéma en temps

L'idée est de passer du schéma explicite en temps, défini à la relation (21), à un schéma implicite paramétré par θ, $0 \le \theta \le 1$. L'opérateur de diffusion $-\frac{\partial}{\partial x}\left(k \frac{\partial u}{\partial x}\right)$ est interpolé en temps entre les instants nΔt (avec le poids (1−θ)) et (n+1) Δt (avec le poids θ). On a donc la relation suivante :

$$(32)\quad \frac{1}{\Delta t}\left(u_j^{n+1} - u_j^n\right) - \frac{k}{\Delta x^2}\left\{\theta\left(u_{j+1}^{n+1} - 2\,u_j^{n+1} + u_{j-1}^{n+1}\right) + (1-\theta)\left(u_{j+1}^{n} - 2\,u_j^{n} + u_{j-1}^{n}\right)\right\} = 0.$$

On remarque que pour θ = 0, on retrouve le schéma explicite déja vu au second paragraphe, alors que les choix θ = 1 et $\theta = \frac{1}{2}$ définissent un schéma d'Euler implicite du premier ordre et le schéma de Crank-Nicolson du second ordre en temps. Quelle que soit la valeur de θ, le schéma (32) reste du second ordre de précision en espace.

Proposition : Stabilité

Aves les notations introduites en (27)-(29), le θ-schéma (32) pour l'équation de la chaleur a un coefficient d'amplification g (α,ξ) donné par la relation :

$$(33)\qquad g(\alpha,\xi) = \frac{1 - 4\,\alpha\,(1-\theta)\sin^2\frac{\xi}{2}}{1 + 4\,\alpha\theta\sin^2\frac{\xi}{2}}.$$

Ce schéma est donc inconditionnellement stable si $\theta \ge \frac{1}{2}$:

$$(34)\qquad \theta \ge \frac{1}{2} \Rightarrow \forall\xi,\ \forall\alpha \quad |g(\alpha,\beta)| \le 1$$

et stable sous la conditoin suivante si $\theta < \frac{1}{2}$:

$$(35)\qquad \theta < \frac{1}{2} \text{ et } \alpha \le \frac{1}{2-4\theta} \Rightarrow \forall\xi, \quad |g(\alpha,\beta)| \le 1. \qquad \blacksquare$$

On gagne donc en stabilité avec le schéma (32) puisque si θ est plus grand que $\frac{1}{2}$, il n'y a plus de restriction sur le pas de temps pour garantir que le schéma est stable, c'est-à-dire utilisable en pratique sur une machine qui fait des erreurs d'arrondis. Par contre, la construction du schéma est plus complexe. La relation (26) définissait une **formule** de calcul pour évaluer u à l'intant (n+1) Δt à partir de l'ensemble des valeurs u_j à l'instant nΔt. La relation (32) ne définit qu'un **système linéaire** pour calculer l'ensemble des valeurs u_k à l'instant (n+1) Δt à partir de l'ensemble des valeurs u_j connues à l'instant précédent. Nous précisons un peu ce point dans ce qui suit.

• On dispose d'un domaine unidimensionnel Ω =]O,L[discrétisé avec un pas constant de façon à placer N+1 cellules.

(36) $$\Delta x = \frac{L}{N+1}$$

Le point $x_o = 0$ est situé sur la frontière de gauche ; on se donne par exemple une condition de Dirichlet :

(37) $$u_o^n = v_o^n$$

alors que le point $x_{N+1} = L$ est situé sur la frontière de droite du domaine Ω ; on se donne encore (pour simplifier l'exposé) une condition de Dirichlet.

(38) $$u_{N+1}^n = v_L^n$$

Rappelons que l'indice n au second membre des relations (37) et (38) indique que la donnée au bord est variable en espace.

• Les inconnues sont les nombres u_j^{n+1} à évaluer à l'instant "suivant", pour j = 1, 2, ..., N, ce à partir des valeurs u_k^n pour k = 0, ..., N+1, en tenant compte des relations (37) et (38). Notons également que la condition limite à l'instant (n+1) est connue, puisque c'est une donnée du problème global à résoudre.

On lit pour les différentes valeurs de j la relation (32) et on fait apparaître des matrices. On range d'abord les inconnues u_j^{n+1} dans un vecteur U^{n+1} :

(39 $$U^{n+1} = \begin{pmatrix} u_1^{n+1} \\ \vdots \\ u_j^{n+1} \\ \vdots \\ u_N^{n+1} \end{pmatrix}$$

On regarde la "ligne courante" pour la relation (32), c'est-à-dire des valeurs de j qui ne vont pas faire apparaître les conditions aux limites (37) et (38). En pratique, $2 \leq j \leq N-1$, on range à gauche du signe d'égalité tout ce qui est inconnu (n+1) et à droite tout ce qui est connu, après multiplication par Δt. Il vient :

(40) $$-\theta\,\alpha\,u_{j-1}^{n+1} + (1+2\theta\alpha)\,u_j^{n+1} - \theta\,\alpha\,u_{j+1}^{n+1} =$$

$$= (1-\theta)\,\alpha\,u_{j-1}^{n} + (1-2\,(1-\theta)\,\alpha)\,u_j^{n} + (1-\theta)\,\alpha\,u_{j+1}^{n}$$

• On s'intéresse ensuite aux valeurs extrêmes de j (j = 1 et j = N) où les conditions aux limites (37) (38) jouent un rôle actif. Pour j = 1, on obtient :

(41) $$(1+2\theta\alpha)\,u_1^{n+1} - \theta\,\alpha\,u_2^{n+1} = (1-2\,(1-\theta)\,\alpha)\,u_1^{n} + (1-\theta)\,\alpha\,u_2^{n}$$

$$+ \theta\,\alpha\,v_o^{n+1} + (1-\theta)\,\alpha\,v_o^{n}$$

et pour j = N, on a de même :

(42) $$-\theta\,\alpha\,u_{N-1}^{n+1} + (1+2\theta\alpha)\,u_N^{n+1} = (1-\theta)\,\alpha\,u_{N-1}^{n} + (1-2\,(1-\theta)\,\alpha)\,u_N^{n}$$

$$+ \theta\,\alpha\,v_L^{n+1} + (1-\theta)\,\alpha\,v_L^{n}$$

• On écrit de façon synthétique les relations (41), (40) et (42) à l'aide du vecteur d'inconnues (39) sous la forme :

(43) $$A\,U^{n+1} = BU^{n} + V$$

où A, B sont des matrices NxN données par les relations suivantes :

(44) $$A = \begin{bmatrix} 1+2\theta\alpha & -\theta\alpha & & & 0 \\ -\theta\alpha & 1+2\theta\alpha & -\theta\alpha & & \\ & \ddots & \ddots & \ddots & \\ & & -\theta\alpha & 1-2\theta\alpha & -\theta\alpha \\ 0 & & & -\theta\alpha & 1+2\theta\alpha \end{bmatrix}$$

(45) $$B = \begin{bmatrix} 1-2(1-\theta)\alpha & (1-\theta)\alpha & & & 0 \\ (1-\theta)\alpha & 1-2(1-\theta)\alpha & \ddots & & \\ & \ddots & \ddots & (1-\theta)\alpha & \\ & & (1-\theta)\alpha & 1-2(1-\theta)\alpha & (1-\theta)\alpha \\ 0 & & & (1-\theta)\alpha & 1-2(1-\theta)\alpha \end{bmatrix}$$

et V est un vecteur d'ordre N associé aux valeurs au bord du domaine

(46) $$V = \begin{pmatrix} \left(\theta\, v_o^{n+1} + (1-\theta)\, v_o^n\right)\alpha \\ o \\ \vdots \\ \vdots \\ o \\ \left(\theta\, v_L^{n+1} + (1-\theta)\, v_L^n\right)\alpha \end{pmatrix}$$

- Disposant d'un schéma inconditionnellement stable $\left(\text{pour } \theta \geq \frac{1}{2}\right)$, ou peut être plus exigeant quant à sa qualité, c'est-à-dire se donner une propriété supplémentaire vérifiée par l'équation et se demander si elle est vérifiée par le schéma. Le principe du maximum (que nous ne détaillons pas dans ces notes) entraîne que si la température initiale est positive, cette propriété demeure vraie à tout instant. Cette notion ayant un sens physique évident, nous énonçons le résultat qui suit :

Proposition

Le θ-schéma (32) est de **type positif**, ie :

(47) $$\left(\forall j \in \mathbb{Z},\ u_j^n \geq 0\right) \Rightarrow \left(\forall k \in \mathbb{Z},\ u_k^{n+1} \geq 0\right)$$

sous la condition suivante :

(48) $$\alpha \leq \frac{2-\theta}{4\,(1-\theta)^2}$$

où α est défini à la relation (27).

- Pour le schéma de Crank-Nicolson par exemple $\left(\theta = \frac{1}{2}\right)$, on trouve une condition de type $\alpha \leq \frac{3}{2}$, qui est moins restrictive que la condition (31) de stabilité pour le schéma explicite, mais donne toujours une condition où le pas de temps est limité par le carré du pas d'espace, à un facteur multiplicatif près.

4) Schémas à trois niveaux en temps

Afin d'enrichir le panel de schémas possibles, où jusqu'ici seul le schéma de Crank-Nicolson est du second ordre en temps mais est en limite de stabilité inconditionnelle, nous présentons deux schémas à deux niveaux en temps, dont l'un a un pur intérêt historique et pédagogique et l'autre est plus récent.

- **Schéma de Richardson**

Ce schéma s'écrit sous forme d'un saute-mouton au-dessus du $n^{ème}$ temps entre les instants n−1 et n+1, l'opérateur en espace étant discrétisé à l'instant n :

(49) $$\frac{1}{\Delta t}\left(u_j^{n+1} - u_j^{n-1}\right) - \frac{k}{\Delta x^2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right) = 0 \quad .$$

On a la proposition suivante :

Proposition : Le schéma de Richardson (49) est du second ordre en espace et en temps. Il est **instable** quel que soit le pas de temps Δt.

Cette idée, proposée à la fin du $XIX^{ème}$ siècle, ne peut être mise en oeuvre à cause de l'instabiblité du schéma explicite centré.

- **Schéma de Gear**

Ce schéma utilise une différentiation rétrograde implicite du second ordre à trois points :

(50) $$\frac{\partial \varphi}{\partial t}(t^{n+1}) = \frac{1}{2}\left[\frac{3}{2}\varphi^{n+1} - 2\varphi^n + \frac{1}{2}\varphi^{n-1}\right] + 0\,(\Delta t^2)$$

et dans le contexte de l'équation de la chaleur, s'écrit :

$$(51) \qquad \frac{1}{\Delta t}\left(\frac{3}{2}u_j^{n+1} - 2u_j^n + \frac{1}{2}u_j^{n-1}\right) - \frac{k}{\Delta x^2}\left(u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}\right) = 0 .$$

Proposition

Le schéma de Gear (51) est implicite, du second ordre de précision en espace et en temps, **inconditionnellement stable**.

La preuve de cette proposition est laissée en exercice au lecteur.

V. DIFFÉRENCES FINIES POUR L'ÉQUATION DE POISSON À DEUX DIMENSIONS D'ESPACE

1) Discrétisation par différences finies

On se propose de résoudre de façon approchée le problème suivant

(1) $-\Delta u = f \ \Omega$

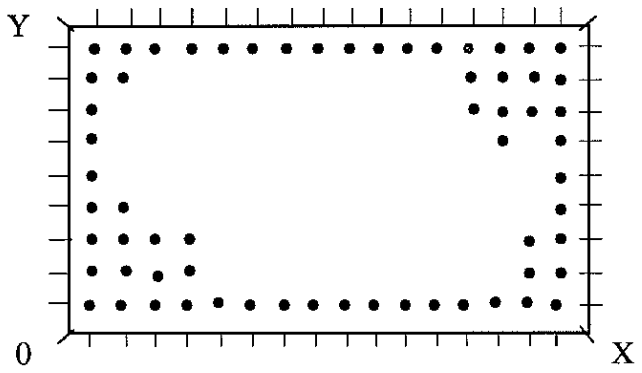
(2) $u = 0 \ \partial\Omega$

où Ω est le rectangle [0,X] x [0,Y]. Pour cela, on introduit un pas d'espace en X noté h et un pas en ordonnée k, définis par :

(3) $$h = \frac{X}{n+1}$$

(4) $$k = \frac{Y}{m+1}$$

où n (respectivement m) désigne le nombre de points intérieurs au domaine Ω dans la direction x (respectivement la direction y).



Avant toute définition de schéma numérique, il faut avoir conscience que la variable y ne joue **pas** le rôle d'un temps (et il en est de même pour x en échangeant les rôles de x et y). Plus précisément, les équations d'évolution étudiées aux chapitres précédents, advection et chaleur :

(5) $$\frac{\partial u}{\partial y} + a\frac{\partial u}{\partial x} = 0$$

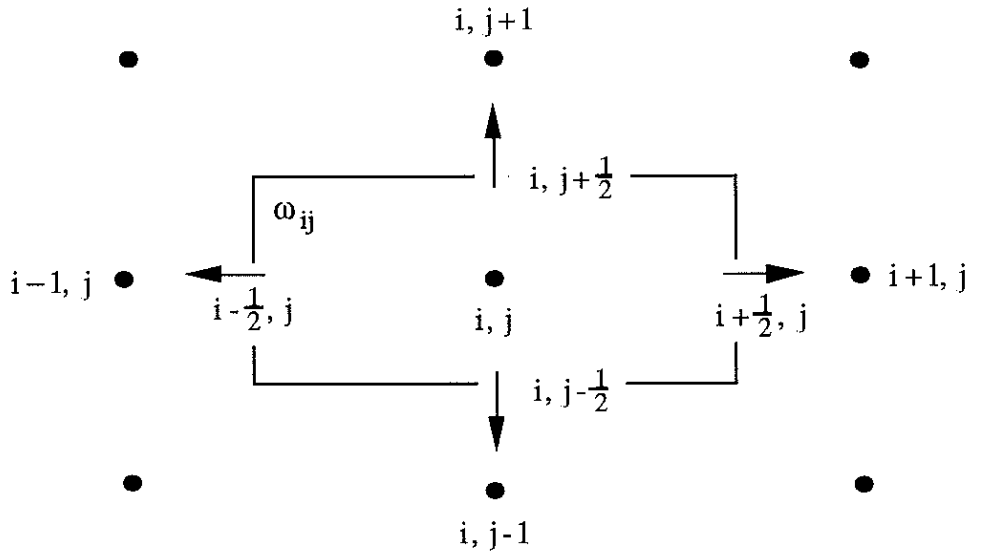
(6) $$\frac{\partial u}{\partial y} - \frac{\partial}{\partial x}\left(K\frac{\partial u}{\partial x}\right) = 0$$

ont été résolues par une méthode numérique mettant en avant le rôle particulier du temps (noté y à dessein dans les équations (5) et (6)). Une condition initiale étant donnée, on calcule **de proche en proche u** aux instants ultérieurs (on balaie pour y fixé à une nouvelle valeur l'ensemble des valeurs de x, ce avec un schéma explicite ou avec un schéma implicite) et on obtient finalement la valeur finale u (•,Y) à l' "instant" Y.

Une telle procédure n'est **pas** possible pour le problème (1) (2). En effet, si on applique le programme précédent avec un hypothétique schéma numérique, il est nécessaire que la valeur finale obtenue u (•,Y) vérifie la condition limite sur le bord]O,X[x {Y} du domaine Ω, c'est-à-dire soit nulle ! On ne peut en pratique imposer une telle contrainte sur un schéma numérique.

Nous devons donc traiter ensemble **tous** les points de discrétisation (i,j) ($1 \leq i \leq n$, $1 \leq j \leq m$) ce qui constitue une difficulté pratique nouvelle sur laquelle nous reviendrons.

La raison profonde concernant le bon choix d'une méthode numérique pour le problème (1) (2) est due à l'analyse mathématique de l'équation de Poisson (2). Sans pouvoir entrer dans les "détails mathématiques" (nous renvoyons par exemple au traité de Courant-Hilbert [1948], ou à la synthèse plus récente de Dautray-Lions [1984]), nous retiendrons ici que l'équation de Poisson est de **type elliptique** (donc doit être résolue par une méthode numérique "globale" qui prend en compte l'ensemble des points de discrétisation), alors que l'équation d'advection (5) est de type **hyperbolique** (rôle particulier donné au temps et vitesse finie de propagation de l'information) et l'équation (6) est de type **parabolique** (rôle particulier pour la variable temporelle, mais vitesse infinie de propagation des informations par l'équation aux dérivées partielles).



Volume de contrôle ω_{ij}

Nous construisons le schéma numérique en cherchant une relation entre les variables discrètes $u_{i,j}$.

(7) $\quad u_{i,j} \simeq u\,(ih, jk)$

de façon à approcher la relation (1) à un ordre prêt, calculé par développement de Taylor.

Nous adoptons dans cet exposé (mais on peut procéder autrement) une approche de type "volumes finis" : on intègre l'équation (1) dans le petit volume :

$$\omega_{ij} \equiv \left]\left(i-\frac{1}{2}\right)h,\ \left(i+\frac{1}{2}\right)h\right[\times \left]\left(j-\frac{1}{2}\right)k,\ \left(j+\frac{1}{2}\right)k\right[$$

en remarquant que l'on a l'identité suivante :

(8) $\quad -\Delta u \equiv -\,\mathrm{div}\,(\nabla u)$

qui, jointe à la formule de Green :

(9) $$\int_{\omega} \mathrm{div}\,\varphi\ dx\,dy = \int_{\partial\omega} \varphi \bullet n\ d\gamma$$

donne la relation :

(10) $$\int_{\omega_{ij}} (-\Delta u)\ dx\ dy = \int_{\partial\omega_{ij}} -\frac{\partial u}{\partial n}\ d\gamma$$

où $\dfrac{\partial}{\partial n}$ désigne la dérivée dans la direction de la normale extérieure au bord de ω_{ij}. Le bord de ω_{ij} est composé de quatre segments :

(11) $$\partial\omega_{ij} = \left\{\left(\left(i+\frac{1}{2}\right)h,\ jk\right)\right\} \times \left]\left(j-\frac{1}{2}\right)k,\ \left(j+\frac{1}{2}\right)k\right[$$

$$\cup\ \left]\left(j-\frac{1}{2}\right)h,\ \left(j+\frac{1}{2}\right)h\right[\times \left\{\left(ih,\left(j+\frac{1}{2}\right)k\right)\right\}$$

$$\cup\ \left\{\left(\left(i-\frac{1}{2}\right)h,\ jk\right)\right\} \times \left]\left(j-\frac{1}{2}\right)k,\ \left(j+\frac{1}{2}\right)k\right[$$

$$\cup\ \left\{\left]\left(i-\frac{1}{2}\right)h,\left(i+\frac{1}{2}\right)h\right[\right\} \times \left\{\left(ih,\left(j-\frac{1}{2}\right)k\right)\right\}$$

et la normale extérieure, en suivant l'ordre adopté à la relation (11), est donnée par :

(12) directions normales à $\partial\omega_{ij}$: $\dfrac{\partial}{\partial x}, \dfrac{\partial}{\partial y}, -\dfrac{\partial}{\partial x}, -\dfrac{\partial}{\partial y}$.

La relation (10) est donc approchée en utilisant un point d'intégration par segment composant le bord ω_{ij} :

$$(13) \qquad -\int_{\omega_{ij}} \Delta u \, dx \, dy \simeq k\left[\left(\frac{\partial u}{\partial x}\right)_{i-\frac{1}{2},j} - \left(\frac{\partial u}{\partial x}\right)_{i+\frac{1}{2},j}\right] + h\left[\left(\frac{\partial u}{\partial y}\right)_{i,j-\frac{1}{2}} - \left(\frac{\partial u}{\partial y}\right)_{i,j+\frac{1}{2}}\right]$$

Il suffit d'utiliser des différences finies **centrées** pour le calcul approché des gradients aux points intermédiaires :

$$(14) \qquad \left(\frac{\partial u}{\partial x}\right)_{i+\frac{1}{2},j} = \frac{1}{h}\left(u_{i+1,j} - u_{i-1,j}\right) + 0\,(h^2)$$

$$(15) \qquad \left(\frac{\partial u}{\partial y}\right)_{i,j+\frac{1}{2}} = \frac{1}{k}\left(u_{i,j+1} - u_{i,j-1}\right) + 0\,(k^2)$$

pour obtenir une expression approchée du laplacien sur la grille (h, k).

$$(16) \qquad -\left(\Delta_{h,k}\, u\right)_{ij} \simeq \frac{1}{|\omega_{ij}|}\int_{\omega_{ij}} (-\Delta u)\, dx\, dy \ ,$$

sachant que la surface de ω_{ij} vaut hxk, on a donc :

$$(17) \qquad -\left(\Delta_{hk}\, u\right)_{ij} \equiv -\frac{1}{h^2}u_{i-1,j} - \frac{1}{k^2}u_{i,j-1} + 2\left(\frac{1}{h^2}+\frac{1}{k^2}\right)u_{i,j} - \frac{1}{k^2}u_{i,j+1} - \frac{1}{h^2}u_{i+1,j}$$

et la version discrétisée par différences finies du problème (1) (2) s'écrit :

$$(18) \qquad -\left(\Delta_{hk}\, u\right)_{ij} = f_{ij} \qquad 1 \le i \le n, \quad 1 \le j \le m$$

$$(19) \qquad u_{ij} = 0 \qquad i = 0 \text{ ou } n+1, \quad j = 0 \text{ ou } n+1$$

Dans la relation (18), f_{ij} désigne bien sûr une moyenne de f dans le petit volume ω_{ij} :

$$(20) \qquad f_{ij} = \frac{1}{|\omega_{ij}|}\int_{\omega_{ij}} f\, dx\, dy$$

ou bien, ce qui revient à faire une formule de quadrature à un point pour le membre de droite de la relation (20), la valeur ponctuelle au point (ih, jk) :

(21) $f_{ij} = f(ih, jk)$.

2) Formation du système linéaire

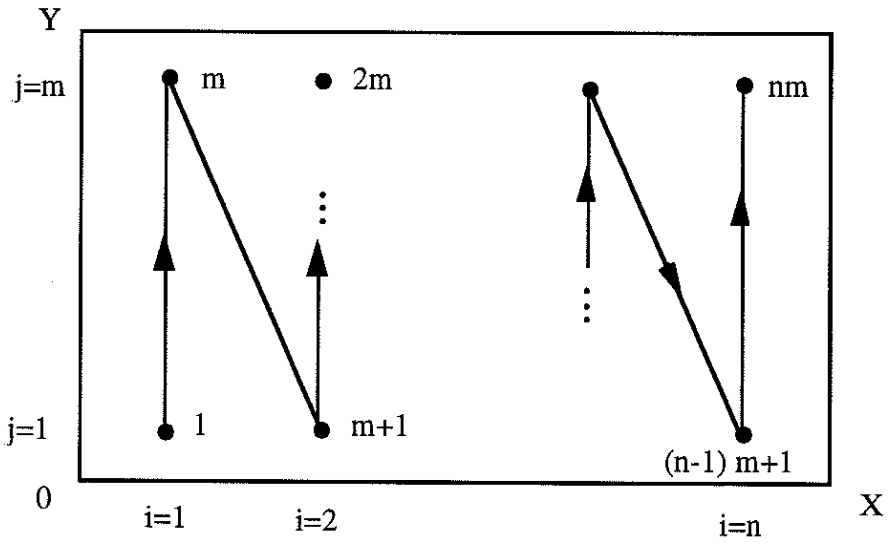
Nous avons, lors de l'étape de discrétisation vue au premier paragraphe, remplacé le problème continu (1) (2) par le problème discret (18) (19) mais nous devons garder à l'esprit que les inconnues de ce problème sont les valeurs ou points de grille $u_{i,j}$ (voir la relation (7)) et qu'on en compte nxm. L'équation (18) est un **ensemble d'équations scalaires linéaires** (il y en a exactement nxm) qu'on peut donc écrire sous une forme matricielle du type :

(22) $AU = F$.

La difficulté présente est que le vecteur U n'est pas encore défini et qu'il n'y a pas de façon naturelle de le définir. En effet, U est un vecteur (unicolonne) contenant nxm éléments alors que la notation "géométrique" u_{ij} indique une matrice rectangulaire paramétrée par les éléments géométriques du problème. Nous devons donc numéroter le double indice (i,j) par un **monoindice** ℓ. Nous choisissons ici (mais de nombreux autres choix sont possibles, il y en a (nm) !) la relation :

(23) $\ell(i,j) = (i-1)\,m + j$

qui revient à numéroter les points du domaine de calcul colonne par colonne, en partant de la colonne la plus à gauche.



Numérotation des sommets proposée à la relation (23)

Le vecteur U_k des inconnues est donc donné par la colonne suivante :

$$(24 \qquad U = \begin{pmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1m} \\ u_{21} \\ \vdots \\ u_{2m} \\ \vdots \\ u_{i1} \\ \vdots \\ u_{im} \\ \vdots \\ u_{n1} \\ \vdots \\ u_{nm} \end{pmatrix}$$

et nous devons maintenant découvrir la matrice A, d'ordre nmxnm, cachée derrière les relations (17). Il suffit par exemple de remplacer l'indice double ij par le monoindice ℓ proposé à la relation (23). On a donc les correspondances suivantes :

$$(25) \qquad (i, j) \longleftrightarrow \ell$$

$$(26) \qquad (i, j+1) \longleftrightarrow \ell + 1$$

$$(27) \qquad (i, j-1) \longleftrightarrow \ell - 1$$

$$(28) \qquad (i+1, j) \longleftrightarrow \ell + m$$

$$(29) \qquad (i-1, j) \longleftrightarrow \ell - m$$

et l'équation (18) relative au "point courant" (i,j) donc à l'indice courant ℓ, prend la forme :

$$(30) \qquad -\left(\Delta_{hk}\, u\right)_\ell = f_\ell$$

avec :

$$(31) \qquad -\left(\Delta_{hk}\, u\right)_\ell \equiv -\frac{1}{h^2}\, u_{\ell-m} +$$

$$- \frac{1}{k^2}\, u_{\ell-1} + 2\left(\frac{1}{h^2}+\frac{1}{k^2}\right) u_\ell - \frac{1}{k^2}\, u_{\ell+1}$$

$$- \frac{1}{h^2}\, u_{\ell+m}$$

La matrice A est donc naturellement décomposée en nxn blocs de taille mxm. De plus seuls les blocs diagonaux ainsi que les blocs directement supérieurs et directement inférieurs aux blocs diagonaux sont non nuls : A est **tridiagonale par blocs**. Comme les coefficients du second membre de la relation (31) ne dépendent pas de ℓ, on a à mettre en évidence **un** bloc diagonal, **un** bloc de la diagonale supérieure et **un** bloc de la diagonale inférieure.

De plus, compte tenu de l'égalité des coefficients de $u_{\ell+m}$ et de $u_{\ell-m}$ dans la relation (31), les deux blocs extradiagonaux sont clairement égaux. La matrice A se décompose donc sous la forme :

$$(32) \qquad A = \begin{pmatrix} C & J & 0 & \cdots & 0 \\ J & C & \ddots & \ddots & \vdots \\ 0 & \ddots & C & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & J \\ 0 & \cdots & 0 & J & C \end{pmatrix}$$

c'est-à-dire sous forme tridiagonale de nxn blocs, chacun des blocs C et J étant lui-même une matrice mxm. Il résulte alors simplement de la relation (31) qu'on a :

$$(33) \qquad J = -\frac{1}{h^2}\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}$$

et :

$$(34) \qquad C = \begin{pmatrix} 2\left(\frac{1}{h^2}+\frac{1}{k^2}\right) & -\frac{1}{k^2} & 0 & \cdots & 0 \\ -\frac{1}{k^2} & 2\left(\frac{1}{h^2}+\frac{1}{k^2}\right) & \ddots & \ddots & \vdots \\ 0 & -\frac{1}{k^2} & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -\frac{1}{k^2} \\ 0 & \cdots & 0 & -\frac{1}{k^2} & 2\left(\frac{1}{h^2}+\frac{1}{k^2}\right) \end{pmatrix}$$

La matrice A est **creuse**. Elle a a priori cinq éléments non nuls par ligne (sauf pour les points du bord où il y en a moins), ce qui donne un maximum de 5 nm termes non nuls au lieu de $(nm)^2$ pour une matrice de cet ordre. Cette remarque doit absolument être prise en compte lors de l'utilisation pratique de A sur un ordinateur (la résolution numérique du système (22)) sous peine de saturer très vite la mémoire et les capacités de calcul des plus gros ordinateurs du moment, avec une programmation ne prenant pas en compte les spécificités du problème posé. Nous avons aussi :

Proposition

La matrice A définie en (32) (33) (34) est symétrique, définie positive.

La symétrie est claire ; nous admettons le caractère défini positif qui apparaîtra à nouveau plus clairement, lors de la méthode des éléments finis. Toutefois, la proposition précédente fournit "naturellement" une méthode de résolution numérique, l'algorithme du gradient conjugué.

3) Résolution du système linéaire par la méthode du gradient conjugué.

Proposé sous forme d'exercice au chapitre 9.

VI. INTRODUCTION À L'ÉCRITURE VARIATIONNELLE DES PROBLÈMES ELLIPTIQUES

1) Motivation

La méthode des éléments finis est une technique numérique de discrétisation qui est très naturelle si le problème mathématique posé peut s'écrire sous la forme d'un problème de minimisation ou sous une forme variationnelle. La difficulté essentielle pour comprendre les fondements de la méthode des éléments finis consiste à savoir écrire les équations aux dérivées partielles de type elliptique (équation de Poisson, élasticité, problème de Stokes) sous forme fonctionnelle où les opérateurs différentiels sont cachés derrière les fonctions tests. C'est cette difficulté que nous abordons et détaillons dans ce chapitre.

2) Problème de Dirichlet homogène pour l'équation de Poisson

• Nous reprenons le problème aux limites posé au chapitre 5. Étant donné un domaine borné Ω dans $\mathbb{R}$, $\mathbb{R}^2$ ou $\mathbb{R}^3$, on cherche $u(x)$, solution du problème de Poisson :

(1) $\quad -\Delta u = f \qquad \Omega$

avec condition limite homogène :

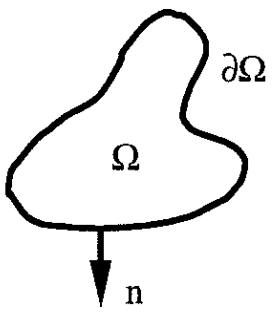
(2) $\quad u = 0 \qquad \partial\Omega$

où $\partial\Omega$ désigne le bord du domaine Ω. On appelle dans la suite "formulation EDP" (pour Équation aux Dérivées Partielles) le problème (1) (2).

Avant de poursuivre et d'introduire la "formulation variationnelle" FV du problème (1) (2), nous avons besoin d'un outil de calcul, la formule de Green d'intégration par parties.

Proposition

Ω désigne un ouvert borné de $\mathbb{R}^n$, $n(x)$ est la normale extérieure au domaine Ω, x décrivant le bord $\partial\Omega$ de Ω, u et v sont des fonctions $\Omega \to \mathbb{R}$ prolongeables continuement sur l'adhérence $\overline{\Omega}$ (donc le bord $\partial\Omega$), dont la dérivée (le gradient) est une fonction (de carré) intégrable. On a la relation d'intégration par parties suivante :



(3) $$\int_\Omega u \frac{\partial v}{\partial x_j}\, dx = -\int_\Omega \frac{\partial u}{\partial x_j} v\, dx + \int_{\partial\Omega} uv\, n_j\, d\gamma.$$

- La preuve de cette proposition est donnée en annexe dans un cas particulier.

- Si $\Omega =]a,b[$, la relation (3) prend la forme élémentaire suivante :

$$\int_a^b u \frac{dv}{dx} dx = -\int_a^b \frac{du}{dx} v \, dx + [uv]_a^b$$

qui est conséquence du fait que la normale extérieure en a (respectivement en b) est égale à -1 (respectivement $+1$).

On peut également dériver toute une série de formules plus ou moins savantes de la relation (3) qui est, seule, fondamentale à retenir. Nous en donnons une, utile dans ce chapitre.

(4) $$\int_\Omega (\Delta u)\, v \, dx = -\int_\Omega \nabla u . \nabla v \, dx + \int_{\partial\Omega} \frac{\partial u}{\partial n} \, v \, d\gamma$$

où :

(5) $$\Delta u = \sum_{j=1}^{n} \frac{\partial^2 u}{\partial x_j^{\,2}}$$

(6) ∇u est le vecteur de j° composante $\dfrac{\partial u}{\partial x_j}$

(7) $$\nabla u . \nabla v = \sum_{j=1}^{n} \frac{\partial u}{\partial x_j} \cdot \frac{\partial v}{\partial x_j}$$

(8) $$\frac{\partial u}{\partial n} = \nabla u \cdot n = \sum_{j=1}^{n} \frac{\partial u}{\partial x_j} \cdot n_j \quad \text{(sur } \partial\Omega \text{ seulement)}$$

- **Formulation variationnelle**

L'écriture variationnelle d'un problème aux limites elliptiques prend toujours une forme du type :

(9) $u \in V$

(10) $a(u,v) = L(v) \qquad \forall\, v \in V$

Nous notons que la formulation variationnelle est composée **à la fois** de (9) et de (10). Écrire un problème EDP sous forme variationnelle FV, c'est suivre le **programme de travail** suivant :

(i) * Trouver l'espace V où l'on cherche la solution.

(ii) * Trouver une forme **bilinéaire** a (u,v)

(11) $V \times V \ni (u,v) \rightarrow a(u,v) \in \mathbb{R}$

(12) $$\begin{cases} a(\lambda u_1 + \mu u_2, v) = \lambda\, a(u_1, v) + \mu\, a(u_2, v) \\ a(u, \lambda v_1 + \mu v_2) = \lambda\, a(u, v_1) + \mu\, a(u, v_2) \end{cases}$$

(iii) * Trouver une forme linéaire L(v) :

(13) $V \ni v \rightarrow L(v) \in \mathbb{R}$

(14) $L(\lambda v + \mu w) = \lambda L(v) + \mu L(w)$.

On notera aussi que c'est le **même** espace V qui figure aux relations (9) **et** (10) ; V est un **espace vectoriel** :

(15) $u, v \in V \Rightarrow \lambda u + \mu V \in V \qquad \forall (\lambda, \mu)$.

Pour écrire le problème (1) (2) sous forme variationnelle, on choisit d'introduire la condition de Dirichlet homogène u **dans** l'espace V, c'est-à-dire de poser :

(16) $V = \{ u : \Omega \rightarrow \mathbb{R},\ u = 0 \ \text{ sur le bord } \partial\Omega \}$,

la condition (9), jointe au choix (16), exprimant alors exactement la condition de Dirichlet (2). On trouve alors la forme bilinéaire a(•, •) et la forme linéaire L(•) à l'aide de manipulations algébriques fondées sur la formule de Green, et plus précisément dans le cas qui nous intéresse ici, la formule (4). On multiplie l'équation (1) par une fonction test v et on intègre sur le domaine d'étude Ω. Nous avons :

(17) $$\int_\Omega (-\Delta u)\, v\, dx = \int_\Omega f v\, dx \qquad \forall v \in V$$

Nous venons ainsi de trouver **une** formulation variationnelle pour le problème EDP, mais poser $a(u,v) = -\int_\Omega (\Delta u)\, v\, dx$ n'est pas mathématiquement satisfaisant. On cherche (pour des raisons théoriques qui assurent ensuite facilement que le problème a une solution unique) une forme bilinéaire **elliptique**, c'est-à-dire telle que il existe $\alpha > 0$ de sorte que :

(18) $a(v,v) \geq \alpha \|v\|_V^2 \qquad \forall v \in V$.

En pratique, le nombre a (v,v) doit être **clairement** positif (cela doit sauter aux yeux : on intègre une fonction positive par exemple), ce qui n'est pas le cas si on reste à la relation (17). Nous transformons l'intégrale du membre de gauche de (17) à l'aide de la relation (4) :

$$(19) \qquad \int_\Omega (-\Delta u)\, v\, dx = \int_\Omega \nabla u.\nabla v\, dx - \int_{\partial\Omega} \frac{\partial u}{\partial n}\, v\, d\gamma$$

Le terme de bord dans le membre de droite de la relation (19) est toujours nul dès que v appartient à l'espace V proposé à la relation (16) [souvenez-vous que c'est le **même** espace V aux relations (9) et (10)]. On peut donc réécrire la relation (17) sous la forme :

$$(20) \qquad \int_\Omega \nabla u.\nabla v\, dx = \int_\Omega fv\, dx \qquad \forall\, v \in V$$

et cette fois, on a clairement :

$$(21) \qquad \int_\Omega |\nabla v|^2\, dx \geq 0$$

ce qui ne montre pas que la relation (18) est satisfaite (on n'a même pas défini de norme sur l'espace V, et la relation (16) ne définit même pas V de façon mathématiquement précise !) mais indique qu'elle peut l'être.

- La formulation FV du problème de Poisson (1) (2) est donc donnée par les deux relations (9) **et** (10), le choix de l'espace de fonctions proposé à la relation (16), la forme bilinéaire a :

$$(22) \qquad a(u,v) = \int_\Omega \nabla u.\nabla v\, dx$$

et la forme linéaire L :

$$(23) \qquad L(v) = \int_\Omega fv\, dx.$$

Nous venons donc de démontrer la :

Proposition

Si u est solution du problème EDP (1) (2), il est aussi solution du problème sous forme variationnelle FV (9) (10) (16) (22) (23).

Nous avons aussi la proposition réciproque, qui montre que la formulation variationnelle et l'écriture sous forme d'équations aux dérivées partielles sont **équivalentes**.

Proposition

Si u est solution de la formulation variationnelle FV [(9) (10) (16) (22) (23)], alors u est solution de l'équation aux dérivées partielles (1) associée à la condition à la limite (2).

Preuve

Le choix (16) pour espace des fonctions tests V montre que la solution du problème variationnel FV est une fonction nulle sur le bord du domaine Ω, donc que la condition limite (2) de Dirichlet homogène est automatiquement satisfaite. La fonction u solution de FV vérifie (20). On intègre par parties le terme bilinéaire $\int_\Omega \nabla u.\nabla v \, dx$ à l'aide de la relation (19), en tenant compte du fait que l'intégrale de bord est nulle puisque la fonction test v est nulle sur le bord $\partial\Omega$. Il vient alors :

$$(24) \qquad \int_\Omega \left(-\Delta u - f\right) v \, dx = 0 \qquad \forall\, v \in V.$$

Cette relation est vraie pour toute fonction v, donc le coefficient $-\Delta u - f$ est nécessairement nul.

- **Formulation énergie**

Au lieu d'adopter une écriture variationnelle (9) (10), on peut aussi chercher $u \in V$ comme **fonction minimisant une énergie**. On pose :

$$(25) \qquad J(v) = \frac{1}{2} \int_\Omega |\nabla v|^2 \, dx - \int_\Omega fv \, dx$$

pour $v \in V$. La formulation énergie revient à chercher u appartenant à V (relation (9)) tel que :

$$(26) \qquad J(u) \leq J(v) \qquad \forall\, v \in V$$

ce qu'on exprime aussi en disant que u est solution du problème d'optimisation

$$(27) \qquad \inf_{v \in V} J(v)$$

Nous avons la :

Proposition

Une fonction u est solution de la formulation variationnelle FV [(9) (10) (16) (22) (23)] si et seulement si u est solution de la formulation énergie ENE (9) (26).

Preuve

• On développe J (u + θv) où θ est un paramètre réel destiné à tendre vers zéro et v une fonction test appartenant à V ; on suppose que u est solution du problème ENE. On a :

$$(28) \qquad |\nabla (u+\theta v)|^2 = |\nabla u|^2 + 2\theta\, \nabla u.\nabla v + \theta^2 |\nabla v|^2$$

et par intégration sur le domaine Ω, il vient :

$$(29) \qquad J(u+\theta v) = J(u) + \frac{\theta^2}{2}\int_\Omega |\nabla v|^2 dx + \theta \left\{ a(u,v) - L(v) \right\}$$

On écrit maintenant l'hypothèse sous la forme suivante :

$$(30) \qquad J(u) \leq J(u+\theta v)$$

En tenant compte de la relation (29) il vient :

$$(31) \qquad \left[a(u,v) - L(v) \right] \theta + \frac{1}{2} \left[\int_\Omega |\nabla v|^2 dx \right] \theta^2 \geq 0$$

et cette relation est vraie quel que soit θ appartenant à ℝ. Le polynôme du second degré au membre de gauche de la relation (31) est nul en zéro et reste toujours positif, donc sa dérivée en zéro est nulle, ce qui exprime très exactement la relation (10). L'implication ENE PV est donc établie.

• Réciproquement, si u est solution de la formulation variationnelle, on a :

$$(32) \qquad J(v) - J(u) = \left\{ a(u, v-u) - L(v-u) \right\} + \frac{1}{2}\int_\Omega |\nabla (v-u)|^2 dx$$

Le terme entre accolades est nul car u est solution du problème FV (la fonction test s'appelle v−u au lieu de v à la relation (10)) et le terme complémentaire est l'intégrale d'un carré, donc est positif. La relation (26) est donc satisfaite, ce qui achève la démonstration.

3) Problème de Dirichlet non homogène pour l'équation de Poisson

Si on remplace la condition limite homogène (2) par une condition de Dirichlet **non homogène**.

$$(33) \qquad u = u_0 \quad \text{sur } \partial\Omega,$$

la mise sous forme variationnelle n'est pas immédiate. On suppose d'abord que u_0 est la restriction au bord $\partial\Omega$ d'une fonction (encore notée u_0) définie sur le domaine Ω tout entier, ainsi que sur le bord $\partial\Omega$:

(34) $\quad u_0 : \overline{\Omega} \rightarrow \mathbb{R}$ $\quad$ donnée.

Bien entendu, seules les valeurs de u_0 sur le bord (que nous noterons parfois γu_0) sont utiles pour établir la condition limite (33), mais disposer de (34) permet d'écrire la relation (33) sous la forme :

(35) $\quad u = u_0 + w \qquad w \in V$

avec V défini à la relation (16) : si u vaut u_0 au bord, la différence $(u-u_0)$ est une fonction nulle au bord du domaine Ω. L'écriture (35) permet de réintroduire l'espace **vectoriel** V, alors que l'écriture (35) montre que u appartient naturellement à l'espace **affine** passant par u_0 et dirigé par l'espace vectoriel V.

Pour établir la formulation variationnelle, on change de fonction inconnue, remplaçant u par w introduite à la relation (35). On a donc :

(36) $\quad -\Delta w = f + \Delta u_0 \quad \Omega$

(37) $\quad w \in V.$

Nous avons donc un problème de Dirichlet **homogène** pour la fonction inconnue w, et sommes de ce fait dans les conditons du paragraphe précédent. Une formulation variationnelle s'écrit donc sous la forme (9) (10), avec a $(\bullet, \bullet)$ toujours donné par la relation (23) mais L(v) prenant maintenant la forme :

(38) $\quad L_0(v) = \int_\Omega fv\, dx + \int_\Omega \Delta u_0\, v\, dx.$

Une intégration par parties du type (19) et la prise en compte du fait que la fonction test v est nulle sur le bord permet une autre écriture de la forme linéaire $L(\bullet)$:

(39) $\quad L_0(v) = \int_\Omega fv\, dx + \int_\Omega \nabla u_0 \cdot \nabla v\, dx, \qquad v \in V.$

Nous avons donc établi la :

Proposition

Le problème de Dirichlet non homogène (1) (33) s'écrit sous forme variationnelle (37) jointe à la relation :

(40) $\quad a(w,v) = L_0(v) \quad \forall\, v \in V$

où $a(\bullet, \bullet)$ est donné par (33) et $L(\bullet)$ par (39). La solution u s'obtient à partir de la solution variationnelle w à l'aide du changement de fonction inconnue (35).

Bien entendu, cette formulation est équivalente à la formulation EDP initiale. Nous laissons la preuve détaillée en exercice au lecteur.

• Il peut être utile aussi d'utiliser une formulation "énergie". Pour cela, on introduit explicitement l'espace affine V_0 :

(41) $\quad V_0 = \left\{ v : \overline{\Omega} \to \mathbb{R},\ v = \gamma u_0 \quad \text{sur } \partial\Omega \right\}$

et on utilise toujours la fonctionnelle (25). Nous posons la recherche d'un minimum d'énergie sous la forme :

(42) $\quad u \in V_0$

(43) $\quad J(u) \leq J(z) \qquad \forall_z \in V_0$

Noter que l'espace de travail est maintenant l'espace affine V_0 qui prend explicitement en compte la condition limite non homogène, et non plus l'espace vectoriel V. Nous avons la :

Proposition

La formulation énergétique ENE (42) (43) du problème de Dirichlet non homogène est équivalente à la formulation variationnelle (37) (40).

Preuve

• Nous établissons d'abord que $u = u_0 + w$ est minimum d'énergie sur V_0, dès que w est solution de (37) (40). Nous écrivons $z \in V_0$ sous la forme générale :

(44) $\quad z = u_0 + v, \qquad v \in V$

et nous avons alors :

(45) $\quad z = u + (v-w)$

Nous prenons $\theta = 1$ dans la relation (29) et remplaçons v par (v–w). Nous obtenons :

(46) $$J(z) = J(u) + \frac{1}{2}\int_\Omega |\nabla (v-w)|^2 \, dx + \left[a(u, v-w) - L(v-w)\right]$$

Nous remarquons que l'écriture (40) est équivalente à :

(47) $$a(u,v) = L(v) \qquad \forall\, v \in V$$

avec $L(\cdot)$ toujours donné à la relation (23) ; donc en changeant v en w dans la fonction test pour la relation (47), le terme entre crochets au membre de droite de l'égalité (46) est nul, ce qui prouve facilement que :

(48) $$J(z) \geq J(u) \qquad \forall_z \in V_0$$

et on établit l'écriture énergétique.

- Réciproquement, si u est minimum d'énergie sur V_0, nous écrivons $z \in V_0$ sous la forme :

(49) $$z = u + \theta v \qquad v \in V.$$

Le développement (29) est inchangé ainsi que la fin du raisonnement relatif à la relation (31). La fonction u est donc solution de :

(50) $$u \in V_0 = u_0 + V$$

(51) $$a(u,v) = L(v) \qquad \forall\, v \in V$$

qui, au changement de variable (35) près, est une réécriture de la formulation variationnelle (37) (40).

4) Problème mixte pour l'équation de Poisson

Nous progressons peu à peu vers la généralité. Nous supposons la frontière $\partial\Omega$ décomposée en deux parties disjointes Γ_1 et Γ_2 :

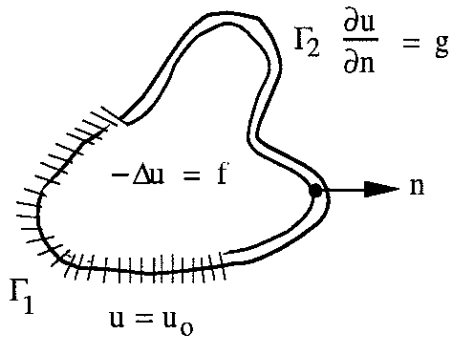
(52) $$\partial\Omega = \Gamma_1 \cup \Gamma_2,$$

et imposons une condition de Dirichlet sur Γ_1

(53) $$u = u_0 \qquad \Gamma_1$$

et une condition sur la dérivée normale, de Neumann, sur Γ_2 :

(54) $$\frac{\partial u}{\partial n} = g \qquad \Gamma_2$$



Problème mixte de Dirichlet Neumann

tout en supposant que u vérifie une équation de Poisson (1) dans le domaine Ω. Comme au paragraphe précédent, nous faisons l'hypothèse (34) que u_0 est définie sur $\overline{\Omega}$ tout entier, mais nous introduisons un **nouvel** espace de fonctions tests :

$$(55) \qquad W = \left\{ v : \overline{\Omega} \to \mathbb{R}, \quad v = 0 \text{ sur } \Gamma_1 \right\}$$

de façon à écrire la condition de Dirichlet non homogène (52) sous la forme équivalente :

$$(56) \qquad u = u_0 + w \qquad\qquad w \in W.$$

Pour établir la formulation variationnelle, nous multiplions (1) par $v \in W$ (attention : $v \in V$ dans les paragraphes précédents, v n'est maintenant nulle **que** sur Γ_1 et non plus sur le bord tout entier) et intégrons sur Ω. Compte tenu de la relation (19), il vient :

$$(57) \qquad \int_\Omega \nabla u . \nabla v \, dx = \int_\Omega fv \, dx + \int_\Omega \frac{\partial u}{\partial n} v \, d\gamma .$$

Nous détaillons le terme de bord de la relation précédente :

$$\int_{\partial\Omega} \frac{\partial u}{\partial n} v \, d\gamma = \int_{\Gamma_1} \frac{\partial u}{\partial n} v \, d\gamma + \int_{\Gamma_2} \frac{\partial u}{\partial n} v \, d\gamma$$

$$= \int_{\Gamma_1} \frac{\partial u}{\partial n} . 0 \, d\gamma + \int_{\Gamma_2} gv \, d\gamma$$

compte tenu de la condition de Neumann (53). Nous avons donc :

$$(58) \qquad \int_{\partial\Omega} \frac{\partial u}{\partial n} v \, d\gamma = \int_{\Gamma_2} gv \, d\gamma$$

et u est solution du problème suivant :

$$(59) \qquad \int_\Omega \nabla u.\nabla v \, dx = \int_\Omega fv \, dx + \int_{\Gamma_2} gv \, d\gamma \qquad \forall\, v \in W$$

ou, ce qui est équivalent, w est solution de :

$$(60) \qquad \int_\Omega \nabla w.\nabla v \, dx = \int_\Omega fv \, dx - \int_\Omega \nabla u_0 \cdot \nabla v \, dx + \int_{\Gamma_2} gv \, d\gamma \qquad \forall\, v \in W$$

Nous avons donc :

(i) changé d'espace de fonctions tests pour que celles-ci soient **nulles dans la portion** de la frontière **où u est donné** par une condition de Dirichlet,

(ii) choisi u sous la forme (56), ce qui revient à introduire la condition de Dirichlet dans l'espace (affine) de recherche de la solution,

(iii) changé le second membre dans la formulation variationnelle (relation (57)) ; on constate qu'il prend maintenant en compte à la fois le "chargement" f dans le domaine Ω et la condition de Neumann g sur la portion ad hoc de la frontière.

Proposition

Si u est solution de la formulation variationnelle (55) (56) (59), alors u est solution de l'équation aux dérivées partielles (1) (53) (54).

Preuve

Nous calculons le membre de gauche de la relation (59) par intégration par parties (relation (19)), en prenant en compte le fait que v est nulle sur Γ_1 dans le développement du terme de bord. Il vient :

$$(61) \qquad \int_\Omega \left(-\Delta u - f\right) v \, dx + \int_{\Gamma_2} \left(\frac{\partial u}{\partial n} - g\right) v \, d\gamma = 0, \qquad \forall\, v \in W.$$

Nous prenons d'abord une fonction v nulle sur l'ensemble du bord ($v \in V$), ce qui annule automatiquement le terme de bord sur Γ_2 à la relation (61). Nous en déduisons (24), ce qui établit (1) car v est arbitraire dans le domaine Ω, même si elle est nulle sur tout le bord ! Regardons à nouveau la relation (61), compte tenu de ce que nous venons d'établir. Il ne reste que le terme de bord cette fois puisque (1) est vraie. Mais $v \in W$ n'est pas nulle sur Γ_2 où elle peut prendre des valeurs arbitraires (sauf peut être à l'interface entre Γ_1 et Γ_2), donc le coefficient multiplicateur de v est nécessairement partout nul sur Γ_2, ce qui établit la condition de Neumann (54). Comme la condition de Dirichlet est directement exprimée par la relation

(56), nous venons de vérifier que u satisfait à la fois l'équation aux dérivées partielles (1) et les deux conditions aux limites (53) et (54).

■

• La formulation énergie du problème de Dirichlet Neumann est, compte tenu de ce qui précède, simple à deviner. On pose $W_o = u_o + W$, c'est-à-dire :

(62) $$W_o = \left\{ v : \overline{\Omega} \to \mathbb{R},\ v = u_o \quad \text{sur } \Gamma_1 \right\}$$

et l'énergie J(v) prend la forme :

(63) $$J(v) = \frac{1}{2}\int_\Omega |\nabla v|^2 \, dx - \int_\Omega fv \, dx - \int_{\Gamma_2} gv \, d\gamma$$

Une fonction u solution du problème énergétique ENE est par définition solution de :

(64) $$u \in W_o$$

(65) $$J(u) \leq J(v) \qquad \forall\, v \in W_o$$

Nous laissons au lecteur le soin de vérifier que les problèmes ENE ((64) (65)) et FV ((55) (56) (57)) sont équivalents.

5) Problème de Neumann pour l'équation de Poisson

Nous particularisons le cas précédent en étudiant de façon détaillée ce qui se passe lorsque Γ_1 est vide, c'est-à-dire u solution du problème de Neumann pour (1), c'est-à-dire satisfaisant à la condition limite.

(66) $$\frac{\partial u}{\partial n} = g \qquad \partial\Omega$$

On constate d'abord que le problème de Neumann (1) (66) **ne saurait avoir de solution unique**. En effet, si u est remplacé par u + c, où c est une fonction constante, u + c est solution du problème Neumann si u l'est, puisque $-\Delta c = 0$ dans Ω et $\nabla c_o n = 0$ sur $\partial\Omega$. Il convient donc de chercher une (éventuelle) solution de (1) (66) "à une constante près", c'est-à-dire de se placer dans la classe (d'équivalence) de fonctions suivante :

(67) $$X = \left\{ v : \Omega \to \mathbb{R} \right\} / \mathbb{R}$$

où deux fonctions sont dites égales si elles diffèrent d'une constante. De plus, nous supposons que Ω est **connexe** (d'un seul tenant), quitte à dupliquer le raisonnement autant de fois que nécessaire si Ω est fait de plusieurs morceaux.

Intégrons maintenant l'équation (1) dans le domaine Ω, ce qui revient à la multiplier par la constante "1" puis à intégrer. Compte tenu du fait que le gradient d'une constante est nulle, il ne reste que le terme de bord qui, compte tenu de la relation (66), vaut l'intégrale de g dans le domaine Ω. Nous venons d'établir la **relation de compatibilité** entre les données.

$$\int_\Omega f\,dx + \int_{\partial\Omega} g\,dx = 0. \tag{68}$$

Si les données (f, g) ne satisfont pas la relation (68), il n'y a pas de solution pour le problème de Neumann (66) ! Nous supposons donc dans la suite cette condition satisfaite.

Quand on multiplie (1) par une fonction test v et qu'on intègre, on peut remplacer v par v + cste puisque le terme dû à la constante va donner zéro à cause de la relation de compatibilité. Nous pouvons donc considérer que v est défini à une constante additive près, c'est-à-dire supposer $v \in X$. Nous cherchons donc u "à une constante près".

$$u \in X. \tag{69}$$

De plus, multipliant la relation (1) par $v \in X$ et intégrant par parties, il vient :

$$\int_\Omega \nabla u.\nabla v\,dx = \int_\Omega fv\,dx + \int_{\partial\Omega} gv\,d\gamma, \qquad \forall\, v \in X \tag{70}$$

La formulation variationnelle du problème de Neumann est donnée par les relations (69) (70). La difficulté est que l'espace X introduit en (67) est un espace de "fonctions à une constante additive près", qui a un sens mathématique précis, mais d'emploi moins simple que les espaces V ou W introduits dans les relations précédentes. On notera en particulier que si g = 0 (problème de Neumann homogène), la **seule** différence entre les formulations variationnelles du problème de Dirichlet homogène ((9) (20)) et du problème de Neumann homogène ((69) (70)) est le choix de l'espace de travail ! C'est l'espace V des fonctions nulles au bord pour le problème de Dirichlet, c'est l'espace X des fonctions "à une constante près" pour le problème de Neumann. Dans le cas du problème de Dirichlet, l'espace V contient en son sein l'expression de la condition à la limite, alors que pour le problème de Neumann, cette dernière est exprimée grâce au second membre de la formulation variationnelle.

- La formulation en énergie du problème de Neumann est simple. On pose :

(71) $$J(v) = \frac{1}{2} \int_\Omega |\nabla v|^2 \, dx - \int_\Omega fv \, dx - \int_{\partial\Omega} gv \, d\gamma$$

Cette fonctionnelle est définie pour toute fonction v et en particulier pour toute classe de fonctions de X dès que la relation de compatibilité (68) est satisfaite. Le problème d'optimisation :

(72) $$\inf_{v \in X} J(v)$$

a une unique solution $u \in X$, caractérisée par la nullité de la dérivée de J en u le long de toute direction v :

(73) $$dJ(u) \cdot v = 0 \qquad \forall\, v \in X$$

et comme :

(74) $$dJ(u) \cdot v = \int_\Omega \nabla u . \nabla v \, dx - \int_\Omega fv \, dx - \int_{\partial\Omega} gv \, d\gamma$$

u est caractérisée par le fait d'être solution de la formulation variationnelle.

VII. INTRODUCTION À LA MÉTHODE DES ÉLÉMENTS FINIS

1) Introduction

Au chapitre précédent, nous avons vu que l'équation de Poisson :

(1) $\quad -\Delta u = f \qquad \Omega$

associée à divers jeux de conditions aux limites peut être formulée sous forme variationnelle sous la forme :

(2) $\quad u \in V$

(3) $\quad a(u,v) = L(v) \qquad \forall\, v \in V$

où le choix de l'espace de fonctions V, la forme bilinéaire (coercice) $a(\bullet, \bullet)$, ie telle que :

(4) $\quad \exists\, \alpha > 0,\ a(v,v) \geq \alpha \|v\|_V^2 \qquad \forall\, v \in V$

et la forme linéaire $L(\bullet)$ doivent être choisis avec soin pour prendre en compte l'équation de Poisson (1)$\left[a(\bullet, \bullet) \text{ est du type } \int_\Omega \nabla\bullet\nabla\, dx \right]$, les conditions de Dirichlet [l'espace V est formé de fonctions qui s'annulent là où u est explicitement donné sur le bord], le changement f et les conditions de Neumann [via la forme linéaire $L(\bullet)$].

La méthode des éléments finis consiste à se placer dans des sous-espaces de **dimension finie,** poser un problème approché à la place du problème continu (2) (3) pour se ramener in fine à la résolution numérique d'un système linéaire.

2) Problème de Dirichlet à une dimension d'espace

• Dans ce paragraphe, on se place dans le cas très simple d'une dimension d'espace. Plus précisément, nous prenons :

(5) $\quad \Omega =]0,1[\ ;$

alors l'équation de Poisson prend dans ce cas la forme suivante :

(6) $\quad -\dfrac{d^2u}{dx^2} = f \qquad x \in]0,1[$

et nous choisissons des conditions de Dirichlet homogènes

(7) $\quad u(0) = u(1) = 0.$

L'espace V adapté aux conditions (7) est le suivant :

$$(8) \qquad V = \left\{ v : [0,1] \to \mathbb{R}, \quad v(o) = v(1) = 0 \right\}$$

avec une forme bilinéaire $a(\bullet, \bullet)$ définie sur l'espace V :

$$(9) \qquad a(u,v) = \int_o^1 \frac{du}{dx} \frac{dv}{dx} \, dx \qquad u, v \in V$$

ainsi qu'une forme linéaire $L(\cdot)$ définie sur V :

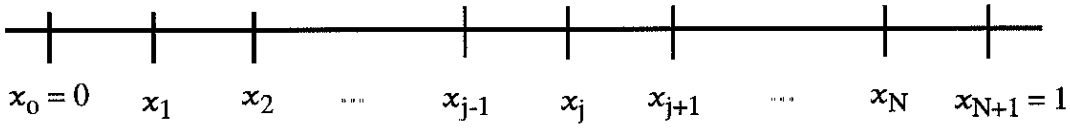
$$(10) \qquad L(v) = \int_o^1 fv \, dx \qquad v \in V.$$

• Nous construisons maintenant un sous-espace V_h de l'espace V donné en (8) de la façon suivante : nous fixons un entier $N \geq 1$, posons :

$$(11) \qquad h = \frac{1}{N+1}$$

et introduisons un ensemble de points sur une grille de pas h :

$$(12) \qquad x_j = jh \qquad j = 0, \ldots, N+1.$$



Grille régulière pour l'équation de Poisson à une dimension d'espace

Nous définissons l'espace V_h comme le sous-espace de V formé de fonctions **continues sur [0,1] et affines dans chaque intervalle** $]x_j, x_{j+1}[$:

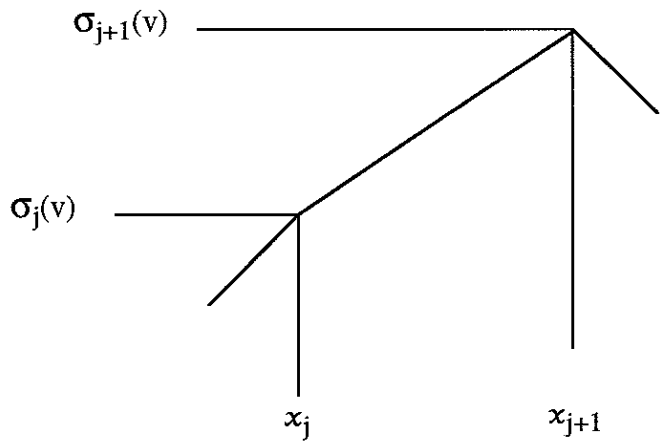
$$(13) \qquad \begin{cases} V_h = \left\{ v : [0,1] \to \mathbb{R}, \ v \text{ continue sur } [0,1], \right. \\ \qquad \forall j = 0, \ldots, N \ v_{]x_j, x_{j+1}[} \text{ affine}, \qquad v(0) = v(1) = 0 \left. \right\}. \end{cases}$$

Une fonction v de V_h est affine dans chaque intervalle donc dépend uniquement des deux valeurs limites aux bornes de celui-ci, c'est-à-dire de $V(x_j+0)$ et $V(x_{j+1}-0)$. Mais v étant globalement continue sur [0,1], les valeurs $v(x_j-0)$ et $v(x_j+0)$ de part et d'autre du sommet x_j sont égales à leur valeur commune $V(x_j)$. Une fonction de v de V_h dépend donc uniquement des valeurs $v(x_o)$, $v(x_1)$, ..., $v(x_j)$, ..., $v(x_N)$, $v(x_{N+1})$ aux sommets des intervalles.

Ces valeurs nécessaires pour calculer les autres valeurs $V(x)$ pour $x \in]0,1[$ de la fonction V sont appelés **degrés de liberté** de l'espace V_h et notés σ_j (V) :

(14) $\sigma_j(v) = v(x_j) \quad j = 0, \ldots, N+1$

De plus, $v \in V_h$ est nulle en $x=0$ et $x=1$, c'est-à-dire aux sommets x_0 et x_{N+1}. Une fonction v dans V_h dépend uniquement de N valeurs $\sigma_1(V), \ldots, \sigma_N(v)$.

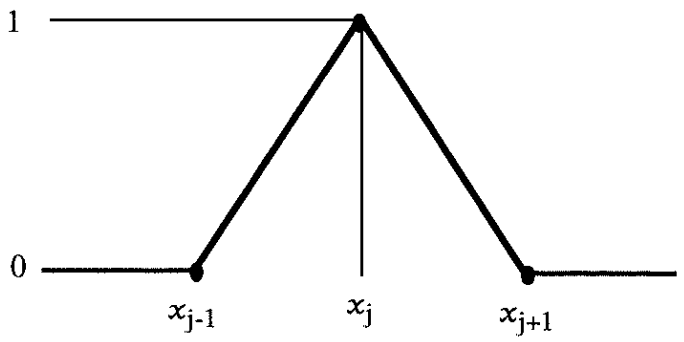


Interpolation affine à l'aide des degrés de liberté de la fonction $v \in V_h$

Nous venons de démontrer que toute fonction v de V_h est entièrement connue dès que les N degrés de liberté $v(x_1), \ldots, v(x_N)$ sont connus, ce qui montre que l'espace **V_h est (au plus) de dimension N**. Pour montrer que V_h est exactement de dimension N, nous explicitons maintenant une base de cet espace. Pour cela, nous utilisons les degrés de liberté introduits à la relation (14) : on fixe j (entre 1 et N) et on cherche une fonction φ_j satisfaisant aux deux relations suivantes :

(15) $\varphi_j \in V_h$

(16) $\sigma_i(\varphi_j) = \begin{cases} 0 & \text{si } i \neq j \\ 1 & \text{si } i = j \end{cases}$



Fonction de base φ_j de l'espace V_h

La construction explicite de φ_j est élémentaire : il suffit d'interpoler par une fonction affine entre les valeurs $\varphi_j(x_i)$, et l'on obtient de cette manière :

(17) $\varphi_j(x) = 0 \quad$ si $\quad x \leq x_{j-1} \quad$ ou $\quad x \geq x_{j+1}$

(18) $\varphi_j(x) = \frac{1}{h}\left(x - x_{j-1}\right)$ si $x_{j-1} \leq x \leq x_j$

(19) $\varphi_j(x) = \frac{1}{h}\left(x_{j+1} - x\right)$ si $x_j \leq x \leq x_{j+1}$

Les fonctions φ_j ainsi obtenues (j = 1, ..., N) appartiennent à V_h et forment une base de cet espace. En effet, si une combinaison linéaire :

(20) $$v = \sum_{j=1}^{N} v_j \varphi_j(x)$$

est identiquement nulle, il en est de même de $\sigma_1(v)$, et comme $\sigma_i(v) = v_i$ compte tenu de la relation (16), le coefficient v_i est nul pour tout i. Réciproquement, si $v \in V_h$ est une fonction arbitraire, alors la fonction $w = v - \sum_{j=1}^{N} v(x_j)\,\varphi_j$ appartient à V_h, vérifie :

(21) $\sigma_j(w) = \sigma_j(v) - v(x_j) = 0 \qquad \forall j = 1, ..., N$

donc est identiquement nulle. Toute fonction V de V_h s'écrit donc de façon unique sous la forme :

(22) $$v(x) = \sum_{j=1}^{N} v(x_j)\ \varphi_j(x)$$

• Ayant construit un sous espace V_h de V qui se manipule avec les N nombres $v(x_j)$ (j = 1, ..., N), nous posons maintenant un **nouveau problème variationnel** dans l'espace V_h, en remplaçant simplement V par V_h dans les relations (2) et (3). Nous cherchons donc u_h tel que :

(23) $u_h \in V_h$

(24) $a(u_h, v) = L(v) \qquad \forall v \in V_h$

Proposition

Le problème variationnel discret (23) (24) admet une unique solution u_h, $a(\bullet, \bullet)$ et $L(\bullet)$ étant donnés par les relations (9) et (10).

Preuve

Nous cherchons u_h sous la forme d'une combinaison des fonctions de base φ_j :

$$u_h = \sum_{j=1}^{N} \underbrace{u_h(x_j)}_{w_j} \varphi_j \qquad (25)$$

Si la relation (24) est vraie pour toute fonction v de V_h, elle est en particulier vraie pour $V = \varphi_i$, ce qui conduit aux N équations suivantes :

$$\sum_{j=1}^{N} a(\varphi_i, \varphi_j)\, w_j = L(\varphi_j) \quad ; \; i = 1, \ldots, N \,. \qquad (26)$$

Le système (26) est formé de N équations et possède N inconnues w_j. Il suffit de montrer que la matrice symétrique $a_{ij} = a(\varphi_i, \varphi_j)$

$$a_{ij} = a(\varphi_i, \varphi_j) \qquad (27)$$

est définie positive pour être certain que si $L(\varphi_i) = 0$ pour tout i, alors $w_j = 0$ pour tout j, ce qui montre qu'alors (26) possède une solution unique. Si $\{\xi_j\}$ désigne un ensemble de N coefficients, on a :

$$\sum_{i,j} a_{i,j}\, \xi_i \xi_j = \int_0^1 \left(\sum_i \xi_i \frac{d\varphi_i}{dx} \right) \left(\sum_j \xi_j \frac{d\varphi_j}{dx} \right) dx \qquad (28)$$

et la forme quadratique (28) est positive. De plus, si $\sum a_{ij}\, \xi_i \xi_j$ est nul, alors :

$$\sum_i \xi_i \frac{d\varphi_i}{dx} = 0 \qquad (29)$$

la fonction $\sum_i \xi_i \varphi_i$ est donc constante et cette constante est nulle car toutes les fonctions φ_i sont nulles en $x=0$ et $x=1$. Donc les ξ_i sont tous nuls car les φ_i forment une base de V_h. Ceci termine la preuve de l'unicité de u_h solution du problème variationnel discret (23) (24).

Il suffit de vérifier que u_h, défini par les relations (25) (26) est effectivement solution du problème (23) (24). D'une part, u_h appartient à V_h puisque c'est une combinaison linéaire de fonctions de base. D'autre part, si la relation (26) est vraie pour tout i, il en est de même après multiplication par un scalaire arbitraire ξ_i et sommation sur i. Nous obtenons de cette façon :

(30) $$a\left(\sum_{j=1}^{N} w_j \varphi_j\,,\ \sum_{i=1}^{N} \xi_i \varphi_i\right) = L\left(\sum_{i=1}^{N} \xi_i \varphi_i\right)$$

compte tenu de bilinéarité de $a(\bullet, \bullet)$ et de la linéarité de $L(\bullet)$. Si ξ_i décrit $\mathbb{R}$ tout entier pour toute valeur de i, $\sum \xi_i \varphi_i$ est une fonction v_h arbitraire dans V_h, et (30) se réécrit sous la forme:

(31) $$a(u_h, v_h) = L(v_h) \qquad \forall v_h \in V_h$$

qui constitue une réplique exacte de la relation (24), au nom de la variable muette près !

La démonstration est donc achevée. ■

• La résolution pratique du problème variationnel (23) (24), qui est une approximation dans V_h du problème continu (2) (3), demande donc la résolution du système linéaire (26). Nous explicitons d'abord les éléments de matrice (27) puis indiquons une méthode approchée pour calculer le second membre $L(\varphi_i)$.

Compte tenu de (27), (9), (17), (18) et (19), l'élément de matrice a_{ij} est nul dès que $|i-j|$ est supérieur (ou égal) à deux :

(32) $$a_{ij} = 0 \qquad \text{si } |i-j| \geq 2.$$

En effet, le support de φ_i, c'est-à-dire l'ensemble des points où φ_i n'est pas nulle est l'intervalle $[x_{i-1}, x_{i+1}]$, donc $\dfrac{d\varphi_i}{dx}$ est nulle hors de cet intervalle :

(33) $$x \notin [x_{i-1}, x_{i+1}] \Rightarrow \varphi_i(x) = 0$$

si $|i-j| \geq 2$, prenons $i \leq j$ pour fixer les idées.

Pour calculer a_{ij}, on décompose l'intervalle [0,1] en cinq parties :

(34) $$[0,1] = \bigcup_{k=1}^{5} I_k$$

avec :

(35) $$I_1 = [0, x_{i-1}]$$

(36) $I_2 = [x_{i-1}, x_{i+1}]$

(37) $I_3 = [x_{i+1}, x_{j-1}]$

(38) $I_4 = [x_{j-1}, x_{j+1}]$

(39) $I_5 = [x_{j+1}, 1]$

et il suffit de vérifier que sur chacun de ces intervalles, l'intégrale :

(40) $$a_{ij}(I_k) = \int_{I_k} \frac{d\varphi_i}{dx} \frac{d\varphi_j}{dx}\, dx$$

est nulle pour établir (32). Or, compte tenu de la relation (33), φ_i **et** φ_j sont nulles sur I_1, I_3 et I_5, φ_i est nulle sur I_4 et φ_j est nulle sur I_2 car $i+1 \leq j-1$. Donc les cinq intégrales introduites en (40) sont nulles et a_{ij} est nul dès que $|i-j| \geq 2$.

Il reste à évaluer $a_{j,\,j+1}$ et a_{jj} pour j quelconque. Compte tenu de (33), le produit $\frac{d\varphi_j}{dx} \frac{d\varphi_{i+1}}{dx}$ est nul sur [0,1] sauf sur l'intervalle $[x_j, x_{j+1}]$ où il vaut $-\frac{1}{h} \times \frac{1}{h}$. Comme cet intervalle est de longueur h, ceci montre que :

(41) $$a_{j,\,j+1} = -\frac{1}{h}.$$

De même le produit $\left(\frac{d\varphi_j}{dx}\right)^2$ est non nul sur $[x_{j-1}, x_{j+1}]$ où il vaut $\frac{1}{h^2}$. Cet intervalle étant de mesure 2h, on en déduit :

(42) $$a_{jj} = \frac{2}{h}.$$

Nous venons de prouver la :

Proposition

L'espace V_h étant défini en (13) sur le maillage régulier (11) (12), le problème variationnel (23) (24) conduit à la résolution numérique du système linéaire (26) dont la matrice a_{ij} est tridiagonale, donnée aux relations (32) (41) (42), c'est-à-dire :

$$(43) \qquad \left[a_{ij}\right] = \frac{1}{h}\begin{pmatrix} 2 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & -1 & 0 & & \vdots \\ 0 & -1 & 2 & -1 & \ddots & \vdots \\ \vdots & 0 & -1 & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & \cdots & 0 & -1 & 2 \end{pmatrix}$$

• On remarque que sur l'intervalle [0,1] avec un maillage régulier, on retrouve une matrice déjà rencontrée au chapitre 4 dans l'étude de l'approximation par différences finies de l'équation de la chaleur. En effet, sur un tel maillage, on a :

$$(44) \qquad -\left(\frac{d^2u}{dx^2}\right)(x_j) \simeq \frac{1}{h^2}\left(-u_{j-1} + 2u_j - u_{j+1}\right)$$

ce qui correspond, à un facteur h près, à la $j^{ème}$ ligne de la matrice $\left[a_{ij}\right]$ écrite en (43).

• L'évaluation du second membre demande le calcul des N intégrales :

$$(45) \qquad b_j = \int_0^1 f(x)\ \varphi_j(x)\ dx\ ;\ j = 1, \ldots, N$$

qui est sans difficulté si f est simple mais doit être lui-même mené de façon approchée si f(•) n'a pas de forme algébrique élémentaire, ce qui est en général le cas. Pour cela, on introduit un espace d'interpolation W_h qui est "un tout petit peu plus grand" que V_h, puisqu'on ne suppose plus que les fonctions de cet espace sont nulles en $x = 0$ et en $x = 1$ (la fonction f(•) n'a aucune raison d'être nulle en $x = 0$ et $x = 1$) :

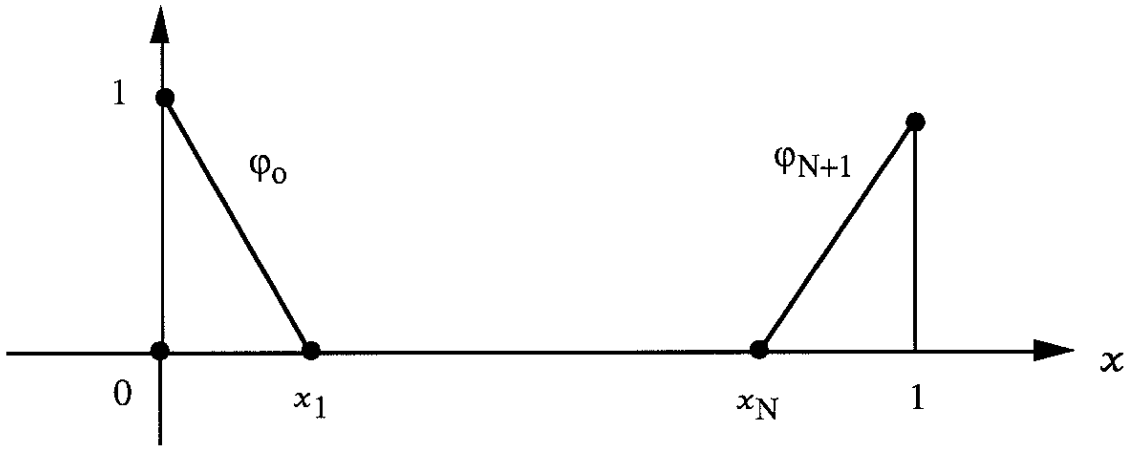
$$(46) \qquad W_h = \left\{ (v : [0,1] \to \mathbb{R},\ v \text{ continue sur } [0,1]\ ,\right.$$

$$\left. v_{|\,]x_j,\, x_{j+1}[} \text{ affine}\ ,\ \forall_j = 0, \ldots, N \right\}.$$

Il est facile de voir que W_h est un espace de dimension finie N+2, contenant V_h mais non inclus dans V. Une base de W_h est donnée par $\{\varphi_o, \varphi_1, \ldots, \varphi_{N+1}\}$, avec φ_j $(1 \le j \le N)$ déjà trouvées plus haut et φ_o et φ_{N+1} satisfaisant à :

$$(47) \qquad \sigma_j(\varphi_o) = 0 \quad \text{si } j \neq 0, \qquad \sigma_j(\varphi_o) = 1 \quad \text{si } j = 0$$

$$(48) \qquad \sigma_j(\varphi_{N+1}) = 0 \quad \text{si } j \neq N+1, \qquad \sigma_j(\varphi_{N+1}) = 1 \quad \text{si } j = N+1$$

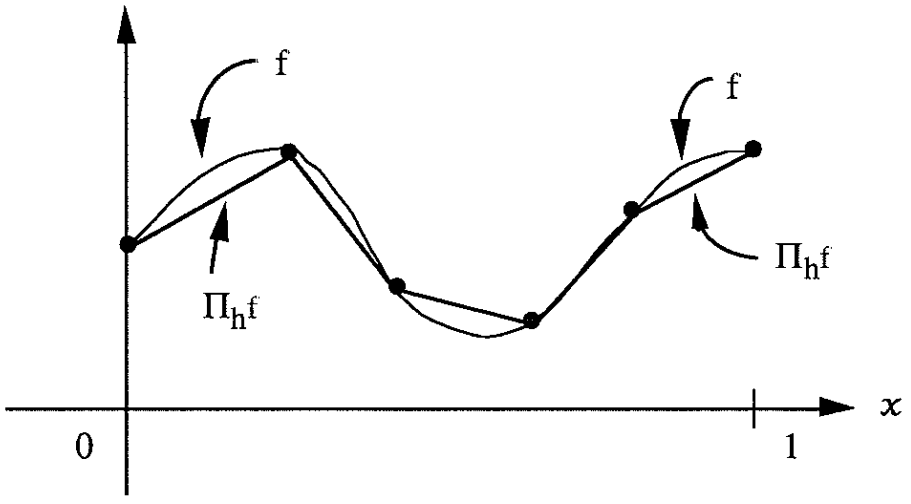


Fonctions de base φ_o et φ_{N+1} pour l'espace W_h

L'espace W_h nous permet d'**interpoler** f par une fonction affine dans chaque intervale $[x_j, x_{j+1}]$, valant f (x_j) pour tous les sommets $x_o, x_1, \ldots, x_{N+1}$. Nous posons donc :

$$(49) \qquad (\Pi_h f)(x) = \sum_{j=0}^{N+1} f(x_j)\ \varphi_j(x)$$

qui n'est pas égale à f, mais définit une fonction de W_h qui est facile à manipuler pratiquement.



Fonction f et interpolé $\Pi_h f$ dans W_h

L'intérêt de la représentation (49) est de remplacer l'intégrale (45) par une valeur approchée facile à calculer numériquement en fonction des données. On pose :

$$(50) \qquad c_j = \int_0^1 \left(\Pi_h f\right)(x)\ \varphi_j(x)\ dx \ ; \quad j = 1, \ldots, N$$

Au lieu de résoudre le système linéaire :

$$(51) \qquad A\,W = B$$

avec A donnée à la relation (43), W vecteur des inconnues w_j introduites à la relation (25), B second membre du terme générique calculé en (45) (qui n'est qu'une réécriture matricielle de (26)), on résout le système linéaire :

$$(52) \qquad A\,W = C$$

où le second membre B est remplacé par C, de terme générique (50). Le calcul de (50) peut être poursuivi plus avant ; nous introduisons le vecteur F (à N+2 composantes ; il faut faire attention à l'introduction de matrices rectangulaires à ce niveau) d'élément générique :

$$(53) \qquad F_k = f(x_k) \ ; \quad k = 0, \ldots, N+1$$

ainsi que la matrice de masse M d'élément générique

$$(54) \qquad M_{jk} = \int_0^1 \varphi_j(x)\ \varphi_k(x)\ dx \ ; \ k = 0, \ldots, N+1 \ ; \ j = 1, \ldots, N$$

ce qui permet d'exprimer C sous la forme :

$$(55) \qquad C = M\,F$$

obtenue en injectant la représentation (49) dans la formule (50). L'avantage de cette écriture est qu'il n'y a plus de calcul numérique laissé en suspens : le vecteur F est connu par les valeurs aux sommets du second membre et la matrice M se calcule de façon explicite. Nous avons :

$$(56) \qquad M_{oo} = M_{N+1,N+1} = \frac{k}{3}$$

$$(57) \qquad M_{j,\,j+1} = \frac{h}{6} \qquad\qquad j = 0, 1, \ldots, N$$

$$(58) \qquad M_{jj} = 2\frac{h}{3} \qquad\qquad j = 1, \ldots, N$$

le calcul, facile à partir des relations (17) à (19), étant laissé au lecteur.

• Etant donné une fonction f : [0,1] $\rightarrow \mathbb{R}$, la résolution approchée du problème (6) (7) par la **méthode des éléments finis** consiste donc à :

* choisir un maillage tel que celui proposé en (12) (mais ce n'est pas obligatoire qu'il soit uniforme),

* calculer les matrices de rigidité A (relation (43)) et de masse M (relations (56) à (58)),

* résoudre le système linéaire :

(59) $\quad A\,W = CF$

dont les inconnues w_j représentent une approximation au $j^{ème}$ sommet de la solution u.

Le résultat est donc une fonction u_h, caractérisée en tout point par la relation (25), donc par les degrés de liberté w_j calculés par résolution de (25).

• Nous terminons ce paragraphe par quelques éléments de **vocabulaire**, général à la méthode des éléments finis. L'ensemble des sommets $\{x_j\}$ définit le **maillage** du domaine d'étude Ω ; chaque intervalle du type $[x_j, x_{j+1}]$ est un **élément fini géométrique** ; leur réunion permet de recouvrir Ω et leur intersection deux à deux est soit vide, soit formée d'un sommet de maillage, soit un élément lui-même. Les **degrés de liberté** σ_j définis en (14) sont des fonctionnelles qui permettent, pour toute fonction v de l'espace d'interpolation V_h, de calculer les nombres σ_j (v) qui la caractérisent complètement. Sur l'ensemble que nous avons étudié ici, les degrés de liberté sont formés par la valeur de la fonction au $j^{ème}$ sommet, qui sont dans ce cas appelés **noeuds** du maillage.

De façon plus générale, **les noeuds sont les supports géométriques des degrés de liberté.** Par exemple, dans certaines applications, on a besoin de définir comme degré de liberté la valeur moyenne d'une fonction dans un élément ; on pose alors :

(60) $$\sigma_{j+\frac{1}{2}}(v) = \frac{1}{h}\int_{x_j}^{x_{j+1}} v(x)\,dx$$

et l'élément $[x_j, x_{j+1}]$ est lui-même un "noeud" du maillage au sens adopté par les praticiens des éléments finis.

3) Problème de Dirichlet à deux dimensions d'espace

• Nous supposons maintenant que Ω est un ouvert borné du plan à frontière polygonale, nous cherchons à approcher par la méthode des éléments finis l'équation de Poisson (1) avec conditions limite de Dirichlet

(61) $u = 0 \quad$ sur $\partial\Omega$.

Pour cela, nous avons vu au chapitre précédent qu'une formulation variationnelle possible est de la forme (2) (3), avec :

(62) $V = \left\{ v : \overline{\Omega} \to \mathbb{R},\ v = 0 \text{ sur le bord } \partial\Omega \right\}$

(63) $a(u,v) = \int_\Omega \nabla u \bullet \nabla v \, dx$

(64) $L(v) = \int_\Omega fv \, dx$.

• La construction d'un sous-espace V_h de V est plus compliquée à deux dimensions d'espace à cause de la **géométrie** du domaine Ω. Nous commençons par réaliser un **maillage** de Ω à l'aide de triangles K, c'est-à-dire découpons Ω en "éléments finis" K qui le recouvrent complètement :

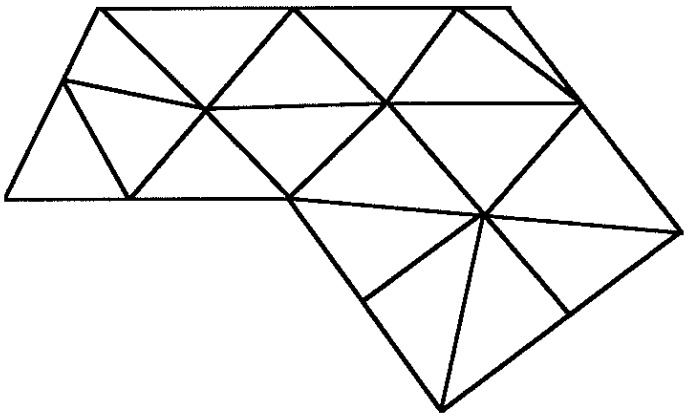
(65) $\overline{\Omega} = \bigcup_{K \in \mathcal{C}} \overline{K}$.

Nous imposons de plus à l'intersection à deux triangles K et L de satisfaire la condition suivante :

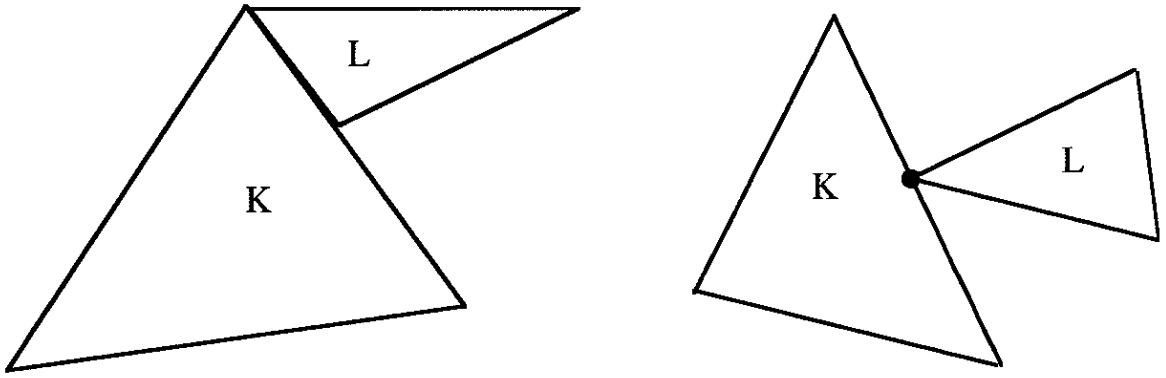
(66) $$\overline{K} \cap \overline{L} = \begin{cases} \varnothing \\ \text{un sommet de K et L} \\ \text{une arête de K et L} \\ K = L \end{cases}$$

qui interdit les deux configurations relatives proposées à la figure de la page suivante, où $\overline{K} \cap \overline{L}$ est une arête de L mais pas de K ou un sommet de L mais pas de K.

Le maillage $\mathcal{T}$ ainsi obtenu permet de définir un ensemble fini d'entités géométriques telles que les éléments du maillage, les arêtes du maillage (communes à deux



Maillage de l'ouvert polygonal Ω par des triangles



Configurations interdites pour deux triangles du maillage $\mathcal{T}$

éléments en touchant le bord $\partial\Omega$) et les sommets du maillage (communs à un nombre indéterminé d'éléments, touchant éventuellement le bord).

Nous définissons l'espace V_h associé à la triangulation $\mathcal{T}$ en choisissant des fonctions de V (donc nulles au bord), continues globalement sur $\overline{\Omega}$ et affines dans chaque triangle :

$$(67) \qquad \begin{cases} V_h = \{ v : \overline{\Omega} \to \quad , \text{ v continue sur } \overline{\Omega} , \\ \qquad\qquad v_{|K} \text{ est affine } \forall K \in \mathcal{T} , \ v_{|\partial\Omega} \equiv 0 \} . \end{cases}$$

Le point clef est de remarquer qu'une fonction de V_h est entièrement déterminée par ses valeurs sur les **sommets du maillage intérieurs** au domaine Ω. On appelle $\mathcal{T}_o$ l'ensemble de ces sommets :

(68) $\mathcal{T}_o = \{ A \in \mathcal{C},\ A \text{ sommet},\ A \in \Omega,\ A \notin \partial\Omega \}$

et pour $A \in \mathcal{T}_o$, on introduit le degré de liberté σ_A associé à ce sommet :

(69) $\sigma_A(v) = v(A) \qquad A \in \mathcal{T}_o, \quad v \in V_h$

On note également N_h le cardinal de l'ensemble $\mathcal{T}_o$, c'est à dire le nombre de sommets du maillage qui ne sont pas sur le bord. On a la

Proposition

Soit $(\alpha_1, \cdots, \alpha_A, \cdots, \alpha_{N_h})$ une collection de N_h réels fixés. Il existe une unique fonction w nulle sur $\partial\Omega$ telle que :

(70) $\sigma_A(w) = \alpha_A \qquad \forall A \in \mathcal{T}_o$

(71) $\omega|_K$ est une fonction affine $\overline{K} \to \mathbb{R}$ continue.

De plus, w est globalement continue sur $\overline{\Omega}$, ce qui montre que w appartient à V_h.

Pour des raisons de commodités, nous appelons $\mathcal{T}$ l'ensemble de **tous** les sommets du maillage.

Preuve de la proposition

Si A désigne un sommet quelconque du maillage, ou bien A appartient à $\mathcal{T}_o$ et la valeur w(A) vaut α_A compte tenu de la relation (70), ou bien A est sur le bord du domaine Ω et la valeur w(A) est nulle. Donc les valeurs de w aux sommets du maillage sont toutes bien définies. De même qu'il faut trois points pour définir un plan dans l'espace, il existe, pour chaque triangle K de la triangulation $\mathcal{T}$, une unique fonction affine φ_K continue sur $\overline{K}$ et de valeurs données aux trois sommets du triangle. La seule chose à montrer est que si on définit w en recollant les fonctions φ_K précédentes, c'est à dire en posant :

(72) $w(x) = \varphi_K(x) \qquad x \in K, \quad K \in \mathcal{T}$

on définit bien une unique fonction continue $\overline{\Omega} \to \mathbb{R}$ de cette façon. Le seul point non banal est d'établir la continuité sur les arêtes a du maillage à l'interface entre deux éléments K et L du maillage, car en un point $x \in a$, $w(x)$ peut avoir a priori **deux** valeurs limites selon que l'on tend vers x depuis K ou depuis L. Mais pour $x \in a$, la valeur $\varphi_K(x)$ est l'interpolé affine entre les deux valeurs $\sigma_A(\varphi_K)$ et $\sigma_B(\varphi_K)$ (où l'on a posé $a = [A, B]$, lesquelles sont

indépendantes de $K \in \mathcal{T}$ compte tenu de (70). La valeur $\varphi_K(x)$, pour $x \in$ a, est définie de façon unique par interpolation linéaire à l'aide des valeurs $\sigma_A(W)$ et $\sigma_B(W)$, donc ne dépend pas du choix de K ou L, ce qui établit la propriété.

■

Réciproquement, étant donné une fonction w de V_h, la collection $\{v(A),\ A \in \mathcal{T}_o\}$ de ses valeurs aux sommets intérieurs du maillage est bien définie car w est continue sur $\overline{\Omega}$. Nous avons donc la

Proposition

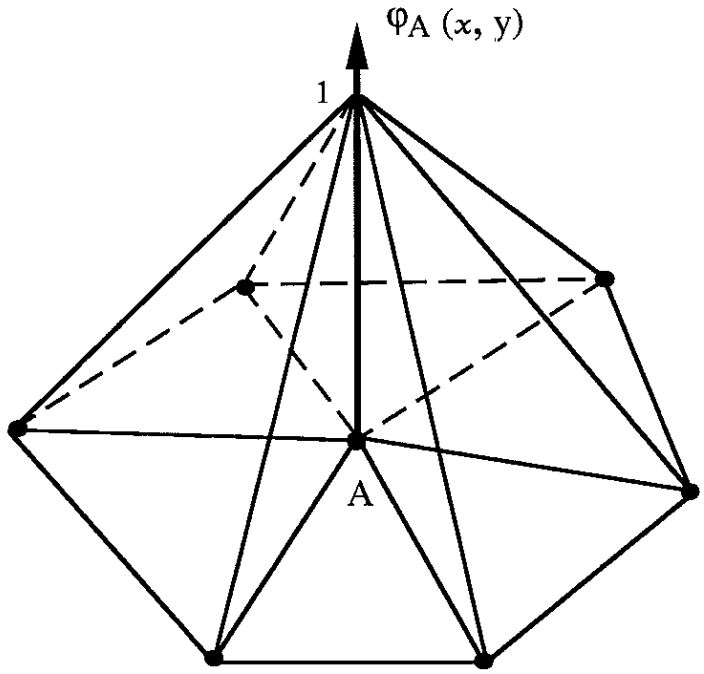
L'espace V_h défini à la relation (67) est de dimension N_h ; il existe une famille unique de fonctions $\{v(A),\ A \in \mathcal{T}_o\}$ formant une **base** de V_h et telle que :

(73) $\sigma_B(\varphi_A) = 0$ si $B \neq A$

(74) $\sigma_A(\varphi_A) = 1$;

toute fonction $w \in V_h$ admet une décomposition unique sous la forme :

(75) $$w(x) = \sum_{A \in \mathcal{T}_o} w(A)\ \varphi_A(x)\ , \quad x \in \Omega .$$



Graphe d'une fonction de base φ_A relative à un sommet A du maillage commun à six triangles

Preuve

On a vu à la proposition précédente que pour toute collection $\{\alpha_A,\ A \in \tau_o\}$ de réels, il existe une unique fonction $\varphi \in V_h$ de sorte que $\sigma_A(\varphi) = \alpha_A$ pour tout sommet A de τ_o En prenant le cas particulier proposé aux relations (73) (74), c'est à dire $\alpha_B = 0$ sauf si B = A (A désignant un sommet fixé), on construit une **unique** fonction φ_A appartenant à V_h et satisfaisant aux deux relations (73) et (74). La fonction φ_A a pour valeur zéro sur tous les sommets, sauf pour le sommet A où elle vaut 1 ; son graphe à l'allure proposée à la page précédente.

Montrons que la famille $\{\varphi_A,\ A \in \tau_o\}$ forme effectivement une base de V_h, c'est à dire que la relation (75) a nécessairement lieu pour toute fonction $w \in V_h$. La différence

$$(76) \qquad z(x) = w(x) - \sum_{A \in \tau_o} w(A)\ \varphi_A(x)\ , \quad x \in \Omega$$

est une fonction appartenant à V_h, nulle en tout sommet B du maillage qui n'est pas sur le bord :

$$(77) \qquad z(B) = w(B) - \sum_{A \in \tau_o} w(A)\ \varphi_A(B) = w(B) - w(B) = 0$$

et nulle également sur le bord de Ω ; la fonction $z(\bullet)$ est nulle sur tout sommet du maillage, donc en tout point de Ω par interpolation affine, donc elle est nulle. Ce raisonnement vaut pour toute fonction $w \in V_h$ donc la famille de fonctions $\{\varphi_A,\ A \in \tau_o\}$ engendre bien l'espace V_h. Cette famille est de plus libre : si une collection $\{\xi_A,\ A \in \tau_o\}$ est telle que :

$$(78) \qquad \sum_{A \in \tau_o} \xi_A\ \varphi_A \equiv 0$$

alors cette fonction est nulle pour tout sommet B du maillage, ce qui implique :

$$(79) \qquad \xi_B\ \varphi_B(B) = 0$$

compte tenu des relations (73) et (74), donc tous les coefficients ξ_B sont nuls, ce qui montre la propriété.

■

• Une fonction $w \in V_h$ se manipule en pratique très facilement ; il suffit de se donner les N_h valeurs aux sommets du maillage (les N_h degrés de liberté), ce qui constitue une famille de N_h nombres à manipuler.

Nous pouvons donc poser un problème variationnel **approché** pour résoudre approximativement le problème variationnel (2) (3) associé à (62) (63) (64). On remplace simplement l'espace V par l'espace V_h de dimension finie construit au-dessus. Il vient :

(80) $\quad u_h \in V_h$

(81) $\quad \int_\Omega \nabla u_h \bullet \nabla v \; dx = \int_\Omega f\, v \, dx \,, \quad \forall\, v \in V_h$

Nous avons la

Proposition

Le problème variationnel discret (80) (81) admet une unique solution u_h. Les coefficients w_A qui caractérisent u_h dans la base $\{\varphi_A\}$:

(82) $\quad W_h = \sum_{A \in \mathcal{T}_0} w_A \varphi_A$

sont solutions d'un système linéaire

(83) $\quad \mathcal{A}\, W = \mathcal{B}$

d'ordre N_h, la matrice $\mathcal{A}$ et le second membre $\mathcal{B}$ ayant des éléments respectifs $\mathcal{A}_{IJ}$ et $\mathcal{B}_I$ donnés par les relations :

(84) $\quad \mathcal{A}_{IJ} = \int_\Omega \nabla\varphi_I \bullet \nabla\varphi_J \; dx$

(85) $\quad \mathcal{B}_I = \int_\Omega f \; \varphi_I \; dx$

Preuve

Elle reprend dans ses grandes lignes l'argumentation donnée dans le cas monodimensionnel. On cherche u_h sous la forme (82) et on prend pour fonction test $v = \varphi_I$. Il vient alors nécessairement :

(86) $\quad \sum_J \mathcal{A}_{IJ}\, W_J = \mathcal{B}_I$

qui est un système symétrique (clair sur la forme (84)) et défini positif puisque :

(87) $$\sum_{IJ} \mathcal{A}_{IJ}\ \xi_I\ \xi_J = \int_\Omega \left[\nabla\left(\sum_J \xi_J\ \varphi_J\right)\right]^2 dx$$

est toujours positif et n'est nul que si $\sum_J \xi_J\ \varphi_J$ est constante, laquelle est nécessairement nulle en regardant la valeur de cette fonction sur un sommet du bord de Ω. Donc u_h est nécessairement égale à la solution (unique) du système (83) (ou (86), ce qui constitue une écriture équivalente).

Réciproquement, la fonction u_h définie ci-dessus est bien solution de (80) (81) car d'une part (82) est une réécriture de (80) et d'autre part si u_h est solution de (81) dans le cas particulier où $v = \varphi_I$, c'est encore vrai dans le cas général par combinaison linéaire des N_h équations, compte tenu de la linéarité par rapport à v des deux membres de l'égalité (81). Ceci montre la propriété.

■

Le calcul explicite des éléments de matrice $\mathcal{A}_{IJ}$ et du second membre $\mathcal{B}_I$ est traité dans un chapitre particulier, mettant en évidence la possibilité de traitement automatique.

Comme pour le cas monodimensionnel, on peut introduire l'interpolé $\Pi_h f$ dans l'espace W_h de toutes les fonctions continues et affines dans chaque triangle. Le système (83) s'écrit alors :

(88) $$\mathcal{A}\ W = \mathcal{M}\ F$$

en introduisant la matrice de masse $\mathcal{M}$ d'élément générique $\mathcal{M}_{IJ}$:

(89) $$\mathcal{M}_{IJ} = \int_\Omega \varphi_I\ \varphi_J\ dx.$$

Les détails sont laissés au lecteur.

VIII. MISE EN OEUVRE INFORMATIQUE DE LA MÉTHODE DES ÉLÉMENTS FINIS

1) Introduction

Dans ce chapitre, nous montrons comment les principes généraux de la méthode des éléments finis sont susceptibles d'un traitement algorithmique qui est ensuite programmable dans un langage évolué (Fortran, C) en vue d'un traitement automatique sur ordinateur. Dans le cas monodimensionnel, très élémentaire, nous introduisons les notions essentielles. Celles-ci sont ensuite utilisées sans modification pour le cas de deux dimensions d'espace, représentatif des problèmes que l'on rencontre dans les applications. Nous développons en détail la méthode d'élimination des noeuds bloqués qui permet de prendre en compte les conditions de Dirichlet non homogènes, et traitons en détail un cas bidimensionnel modèle mais prototype des vérifications à faire lors du développement de logiciels complexes.

2) Mise en oeuvre d'un problème monodimensionnel

Nous approchons, comme au paragraphe précédent, le problème de Dirichlet pour l'équation de Poisson sur l'ouvert Ω [0,1[:

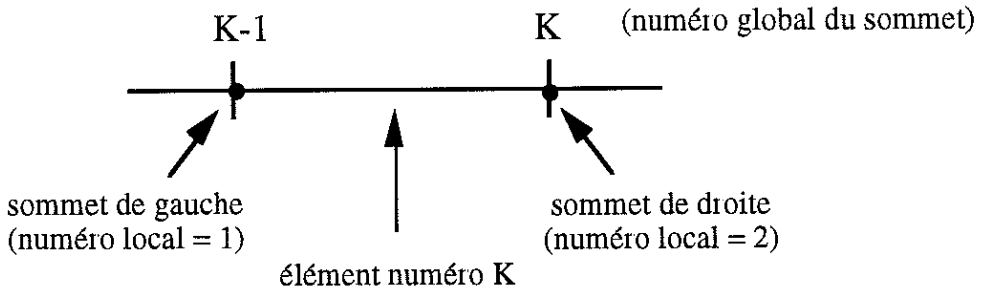
$$(1) \qquad -\frac{d^2u}{dx^2} = f \qquad \text{sur }]0,1[$$

$$(2) \qquad u(o) = u(1) = 0.$$

Nous découpons Ω en (N+1) intervalles $]x_j, x_{j+1}[$ définis par les points x_j.

$$(3) \qquad x_j = jh \; ; \quad h = \frac{1}{N+1} \; .$$

La grille $\{x_0, x_1, ..., x_N, x_{N+1}\}$ s'appelle le **maillage** du domaine Ω ; les intervalles $]x_j, x_{j+1}[$ sont les **éléments** du maillage et les sommets x_j sont porteurs d'un **degré de liberté** (ou **noeud**). Nous avons, avec la relation (3), adopté une numérotation croissante avec l'abscisse, ce qui est un choix simple et naturel compte tenu du caractère monodimensionnel du problème (1) (2). Toutefois, nous allons dans ce qui suit faire abstraction de cette propriété pour laisser émerger les notions générales relatives à la mise en oeuvre.



Numérotations locale et globale

- Un élément "courant" K ($1 \leq K \leq N+1$) est un intervalle $]x_{K-1}, x_K[$ contenant deux noeuds aux extrémités : le noeud à la gauche de K et le noeud à la droite de K ; sans changer la numérotation (3), le noeud de gauche est x_{K-1} et le noeud de droite est x_K. Nous pouvons changer de vocabulaire de la façon suivante : un élément K est porteur de deux noeuds qu'on peut **numéroter localement.** Le noeud de numéro local égal à 1 est **par convention** celui de gauche et celui de numéro local égal à 2 est celui de droite. D'autre part, ces deux sommets ont un numéro relatif à l'ensemble du maillage. Avec le choix (3), on a clairement :

  * numéro du sommet de gauche = K−1
  * numéro du sommet de droite = K.

On appelle **numéro global** le numéro d'un noeud relativement à l'ensemble du maillage. Il convient donc de savoir effectuer la correspondance entre un sommet considéré comme appartenant à un élément K (repéré avec son numéro local relatif à l'élément) et ce même sommet associé à une fonction de base (globale) ; celle-ci s'effectue avec un tableau appelé traditionnellement NELT :

(4) $\quad II = NELT(K, I)$.

La relation (4) signifie que le Ième sommet de l'élément K (le sommet de numéro local I) a pour numéro global II. On a bien entendu $1 \leq I \leq 2$ dans ce cas et pour suivre le choix (3) : $0 \leq II \leq N+1$; $1 \leq K \leq N+1$.

Le tableau NELT est un tableau d'entiers paramétré par les entiers K et I ; il rappelle les éléments liés à la **topologie** de chaque élément (le bord de K, qui porte les noeuds, est constitué de deux sommets).

• En complément du tableau NELT de topologie, il est nécessaire de connaître les coordonnées des sommets du maillage (cf. relation (3)). On introduit le fichier de **géométrie** du maillage :

(5) $X = XX(II)$,

qui, à tout numéro global de noeud II, associe son abscisse X, grâce au tableau XX à valeurs réelles.

• Nous abordons maintenant le calcul de l'élément de matrice $A(II, JJ)$. Compte tenu des choix faits au chapitre précédent, on a :

(6) $$A(II, JJ) = \int_\Omega \nabla\varphi_{II} \bullet \nabla\varphi_{JJ}\, dx$$

où φ_{II} est la fonction de base (globale) relative au noeud II, dont le support (l'ensemble des x où φ_{II} est non nulle) est constitué des deux éléments contenant le sommet II dans leur bord (on parle parfois pour cette raison du **cobord** du sommet II). On calcule l'intégrale au membre de droite de (6) en la découpant sur les éléments du maillage :

(7) $$A(II, JJ) = \sum_{k \in \tau} \int_K \nabla\varphi_{II} \bullet \nabla\varphi_{JJ}\, dx$$

on a alors une remarque essentielle qui fonde la notion même de **matrice élémentaire** :

Proposition

L'intégrale $\int_K \nabla\varphi_{II} \bullet \nabla\varphi_{JJ}\, dx$ ne peut être non nulle que si les sommets II **et** JJ sont tous deux des sommets appartenant au bord de l'élément K.

Preuve

On montre que si l'un des sommets II ou JJ n'appartient pas au bord de K, alors l'intégrale :

(8) $$AB(K, II, JJ) = \int_K \nabla\varphi_{II} \bullet \nabla\varphi_{JJ}\, dx$$

est nulle, ce qui montre la propriété. Si, pour fixer les idées, le noeud II n'appartient pas au bord de l'élément K, alors le support de φ_{II} est constitué par les deux éléments L et M contenant II comme sommet et aucun des deux n'est égal à K (sinon II appartiendrait au bord de K). Donc $\nabla\varphi_{II}$ est non seulement sur les éléments L et M, donc $\nabla\varphi_{II}$ est identiquement nul sur K, ce qui montre que $AB(K, II, JJ)$ est nul.

■

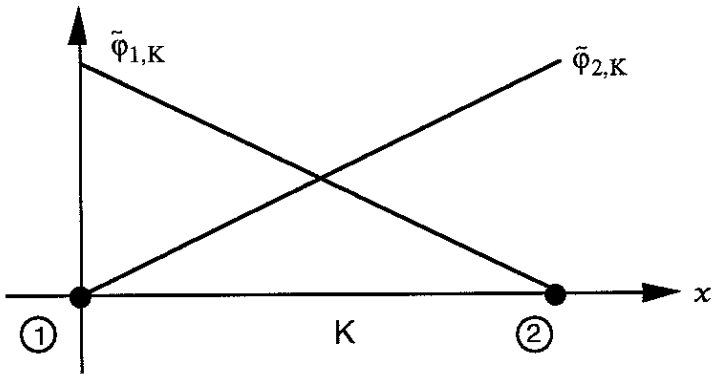
L'intégrale AB (K, İİ, JJ) n'étant non nulle que dans le cas où les sommets II et JJ sont eux-mêmes des sommets de l'élément K, on peut adopter une numérotation locale pour ces sommets. Nous introduisons d'abord la notion de **fonction de base locale**.

Pour I numéro local d'un noeud relativement à l'élément K, on appelle fonction de base locale et on note $\tilde{\varphi}_{I,K}$ la restriction à K de la fonction de base φ_{II}, où II est le numéro global du noeud II, c'est à dire II = NELT (K, I).

(9) $\tilde{\varphi}_{I,K} : K \rightarrow \mathbb{R}$

(10) $\tilde{\varphi}_{I,K}(x) = \varphi_{NELT(K,I)}(x) \quad \forall x \in K.$

Dans le cas de l'élément P1 à une dimension d'espace, les fonctions de base $\tilde{\varphi}_{1,K}$ et $\tilde{\varphi}_{2,K}$ ont un graphe très simple représenté ci-desous :



Fonctions de base locales relatives à un élément K

• La notion de **matrice élémentaire** est alors naturelle : l'intégrale AB (K, İİ, JJ) est évaluée en utilisant les numéros locaux dans l'élément K, donc les fonctions de base locales puisque l'intégrale AB (K, •, •) est à prendre dans l'élément K.

Pour I, J numéros locaux dans l'élément K ($1 \leq I, J \leq 2$ dans le cas présent), on pose :

(11) $AELT(K, I, J) = \int_K \nabla\tilde{\varphi}_{I,K} \bullet \nabla\tilde{\varphi}_{J,K}\, dx$

et on appelle matrice élémentaire relative à l'élément K la matrice obtenue en fixant K et en faisant varier les indices I et J du tableau AELT (K, I, J). Compte tenu de la définition (9) (10) d'une fonction de base locale, on a bien évidemment :

(12) $$\begin{cases} AELT\ (K, I, J) = AB\ (K, II, JJ) \\ \qquad\qquad \text{avec } II = NELT\ (K,I),\ JJ = NELT\ (K,J). \end{cases}$$

Lorsque K est un segment de longueur h, un calcul élémentaire laissé au lecteur montre que, avec la numérotation locale choisie, on a :

(13) $$AELT\ (K, \bullet, \bullet) = \begin{pmatrix} \frac{1}{h} & -\frac{1}{h} \\ -\frac{1}{h} & \frac{1}{h} \end{pmatrix}.$$

• Le calcul d'un élément de matrice A(II,JJ) (relation (6)), appelé parfois élément de la **matrice globale** pour la distinguer des matrices élémentaires, s'obtient par l'algorithme **d'assemblage**, qui constitue le point clef de la mise en oeuvre automatique de la méthode des éléments finis. La mauvaise méthode consiste à faire un double parcours (double boucle) sur tous les sommets du maillage :

(14)
```
boucle sur II
    boucle sur JJ
        calcul de A (II,JJ)
```

En effet, l'étape de calcul de A(II,JJ) s'effectue (si on a recours à un procédé automatisé) à l'aide des matrices élémentaires AELT (K, I, J), et on doit faire successivement les opérations suivantes :

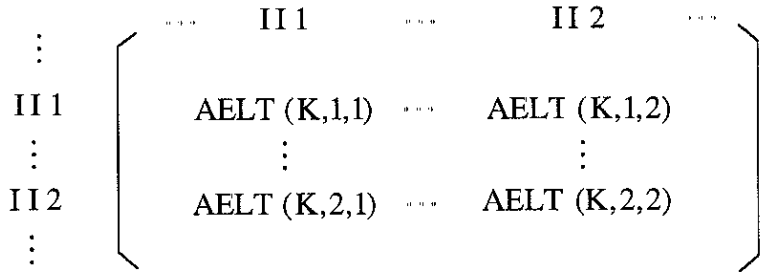
* parcourir les éléments K du maillage,
* parcourir les numéros locaux I et J de K,
* tester si l'on a II = NELT (K, I) et JJ = NELT (K, J).

* Dans l'affirmative, incrémenter A(II,JJ) de la valeur AELT (K, I, J) :

(15) $$A(II,JJ) \leftarrow A(II,JJ) + AELT\ (K, I, J).$$

On passe son temps à rechercher les deux numéros locaux I et J du couple de sommets II et JJ, dans le cas favorable où ceux-ci appartiennent au même élément K !

• L'idée de l'algorithme d'assemblage est de calculer **en même temps tous** les éléments de matrice A(II,JJ), toujours à l'aide de la relation (15), mais en "éclatant" la matrice élémentaire AELT(K, I, J) au sein de la matrice assemblée A (•,•) au cours de l'étape de base.



"Éclatement" de la matrice élémentaire AELT (K,•,•) au sein de la matrice assemblée A (•,•).
(On a II1 = NELT (K,1) et II2 = NELT (K,2)) .

L'algorithme d'assemblage prend alors la forme :

* Initialiser A (•,•) à zéro

(16)
- boucle sur les éléments K du maillage
  - boucle sur les degrés de libertés locaux I et J
  - II = NELT (K,I), JJ = NELT (K,J)
  - Incrémentation de A(II, JJ) à l'aide de la relation (15) .

• Pour le calcul du second membre du système linéaire à résoudre,

(17) $$B(II) = \int_\Omega f(x)\ \varphi_{II}(x)\ dx,$$

on procède exactement de la même façon : on découpe l'intégrale (17) en autant d'intégrales sur chacun des éléments du maillage.

(18) $$B(II) = \sum_{K \in \tau} \int_K f(x)\ \varphi_{II}(x)\ dx$$

et on remarque que chaque intégrale du membre de droite de (18) est non nulle seulement si II est un sommet de l'élément K, compte tenu de la forme particulière des fonctions de base P1, on introduit donc les **seconds membres élémentaires.**

(19) $\quad BELT\,(K,I) = \int_{\Omega} f(x)\ \tilde{\varphi}_I\, dx$

qu'on évalue éventuellement avec une **formule de quadrature** approchée (voir le cours d'analyse numérique élémentaire). L'assemblage du second membre (17) est exactement analogue à l'algorithme (16) pour l'assemblage de la matrice :

* Initialiser B(•) à zéro

(20)
```
⌈boucle sur les éléments du maillage
|⌈boucle sur les numéros I des degrés de liberté locaux
|| II = NELT (K,I)
|| B(II) ← B(II) + BELT (K,I)
```

• Nous terminons ce paragraphe en reprenant l'ordre global des opérations à effectuer :

- Lecture du maillage (tableaux NELT et XX des relations (4) et (5)).
- Calcul des matrices élémentaires et du second membre élémentaire (relations (11) et (19)).
- Assemblage du second membre (algorithme (20) et de la matrice (algorithme (16)).
- Résolution du système linéaire (algorithme de Cholesky ou du gradient conjugué).

3) Mise en oeuvre de l'élément P1 dans $\mathbb{R}^2$

• Les méthodes introduites au paragraphe précédent s'étendent très naturellement au cas bidimensionnel. On sait que le domaine (polygonal) Ω est recouvert par une famille τ d'éléments triangulaires K dont l'intersection deux à deux est soit vide, soit un sommet commun, soit une arête commune, soit pleine. Les fonctions de base de l'espace V_h d'approximations sont les fonctions continues sur $\overline{\Omega}$, affines dans chaque triangle K et nulles sur le bord $\partial\Omega$ puisque nous étudions le cas particulier du problème de Dirichlet homogène.

(21) $\quad -\Delta u = f \qquad \Omega$

(22) $\quad u = 0 \qquad \partial\Omega.$

La dimension de V_h est égale au nombre de fonctions φ_{II} formant une base, c'est à dire le cardinal de τ_o, ensemble des sommets du maillage n'appartenant pas au bord. Cette vision globale du maillage τ, développée au chapitre précédent est soulignée en appelant **fonctions de base globales** les fonctions φ_{ii}, définies de $\overline{\Omega}$ à valeurs réelles. Rappelons enfin que la résolution numérique du problème (21) (22) par éléments finis P1 sur le maillage τ consiste à rechercher la valeur

(23) $\quad w_{II} = u(A_{II})$

d'une fonction $u \in V_h$ sur le sommet A_{II} de numéro (global) II par résolution d'un système linéaire d'ordre N_h.

(24) $\quad \mathcal{A} W = \mathcal{B}$

où la matrice $\mathcal{A}$ a pour terme générique.

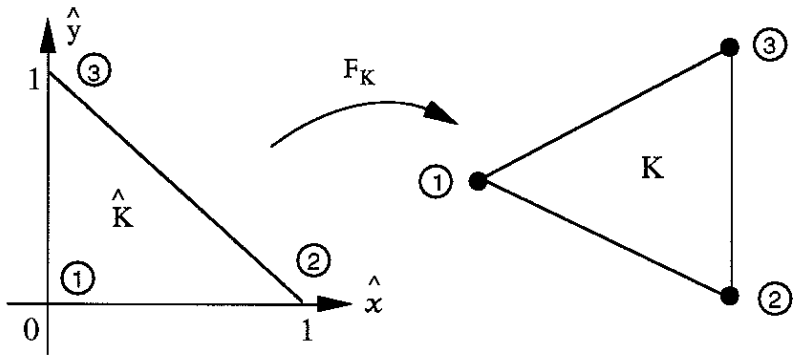
(25) $\quad \mathcal{A}(II,JJ) = \int_\Omega \nabla\varphi_{II} \bullet \nabla\varphi_{JJ}\, dx$

et le second membre $\mathcal{B}$ a pour II^{ème} coefficient :

(26) $\quad \mathcal{B}(II) = \int_\Omega f(x)\, \varphi_I\, dx$

• Un maillage triangulaire de $\mathbb{R}^2$ est une collection de triangles. En pratique, le "mailleur" (logiciel qui permet de générer le maillage, utilisé par l'ingénieur qui a à effectuer un calcul, ou conçu par cet ingénieur dans un cas simple) fournit deux fichiers informatiques : le tableau NELT des numéros globaux de sommets et le tableau XY des coordonnées des sommets.

Le tableau de topologie NELT associé à tout couple (K,I) formé d'un élément du maillage et d'un numéro I compris entre 1 et 3 (un triangle contient trois sommets) le numéro



global (c'est à dire du point de vue de l'ensemble τ du maillage) du sommet correspondant :

(27) $\quad II = NELT(K,I) \qquad 1 \le I \le 3.$

La numérotation locale (27) permet de définir une transformation géométrique "usuelle" F entre le "triangle de référence" $\hat{K}$

(28) $\quad \hat{K} = \left\{ (\hat{x}, \hat{y}) \, , \; 0 \le \hat{x}, \hat{y} \le 1 \; ; \; \hat{x} + \hat{y} \le 1 \right\}$

et le triangle couvrant K à l'aide des numéros locaux introduits à la relation (27). On pose :

(29) $\quad II1 = NELT(K,1) \; ; \; II2 = NELT(K,2), \; II3 = NELT(K,3)$

puis on introduit l'unique application affine F_K qui envoie $\hat{K}$ sur le triangle K de sorte que le sommet numéro I de $\hat{K}$ s'envoie par F_K sur le sommet numéro I de K. De façon précise (et ceci constitue en fait simplement une **convention** !), on a :

(30) $\quad F_K(0,0) = A_{II1} \; ; \; F_K(1,0) = A_{II2} \; ; \; F_K(0,1) = A_{II3}$.

L'application F(•) s'explicite facilement du point de vue algébrique ; on introduit les **coordonnées barycentriques** d'un point $\hat{M} = (\hat{x}, \hat{y})$ de $\hat{K}$ relativement aux sommets $\hat{A}_1 = (0,0), \; \hat{A}_2 = (1,0), \; \hat{A}_3 = (0,1)$:

(31) $\quad \hat{\lambda}_1(\hat{x}, \hat{y}) = (1 - \hat{x} - \hat{y})$

(32) $\quad \hat{\lambda}_2(\hat{x}, \hat{y}) = \hat{x}$

(33) $\quad \hat{\lambda}_3(\hat{x}, \hat{y}) = \hat{y}$

on sait qu'on a alors dans l'élément de référence :

(34) $\quad \hat{M} = \sum_{j=1}^{3} \hat{\lambda}_j(\hat{M}) \; \hat{\lambda}_j \qquad \forall \, \hat{M} \in \hat{K}$

(35) $\quad \sum_{j=1}^{3} \hat{\lambda}_j(\hat{M}) \equiv 1 \qquad \forall \, \hat{M} \in \hat{K}$

et le transport vers l'élément courant K prend la forme suivante :

(36) $\quad K \ni M = F_K(\hat{M}) = \sum_{j=1}^{3} \hat{\lambda}_j(\hat{M}) \, A_{IIj} \, , \qquad \hat{M} \in K$

qui exprime que dans une transformation affine, les barycentres se conservent.

- Nous introduisons alors le tableau XY des coordonnées des sommets du maillage :

(37) $X = XY(II, 1) \qquad Y = XY(II, 2)$

expriment que les coordonnées X,Y du sommet II ont pour adresses (II, 1) et (II, 2) dans le tableau XY. Notons que pour éviter de dupliquer l'information, on fait appel aux numéros globaux des sommets, mais on peut aussi introduire les coordonnées X(K,I) et Y(K,I) du noeud de numéro local I dans l'élément K, à travers la double dépendance suivante :

(38) $X(K,I) = XY(NELT(K,I), 1)$

(39) $Y(K,I) = XY(NELT(K,I), 2)$.

L'explicitation, coordonnées par coordonnées, de la relation (36) prend alors la forme suivante : un point M = (X,Y) de K est paramétré par un point $\hat{M} = (\hat{x}, \hat{y})$ de $\hat{K}$, grâce aux coordonnées baricentriques (31) (32) (33) et aux relations :

(40) $$X = \sum_{j=1}^{3} \hat{\lambda}_j \left(\hat{M}\right) X(K, j)$$

(41) $$Y = \sum_{j=1}^{3} \hat{\lambda}_j \left(\hat{M}\right) Y(K, j).$$

On le voit, tous les calculs avec des nombres réels se font "élément par élément" dans la méthode des éléments finis. C'est cette remarque qui motive la notion de **fonction de base locale** puis de **matrice élémentaire**.

- On se donne un élément K du maillage τ ainsi qu'un numéro local de sommet (ou de degré de liberté puisqu'on s'intéresse à l'élément P1 où les noeuds sont biunivoquement associés aux sommets du maillage) I. On appelle **fonction de base locale** et on note $\tilde{\varphi}_{I,K}$ la restriction à l'élément K de la fonction de base globale φ_{II}, où II = NELT (K,I) et le numéro global du degré de liberté associé. Les relations de définition (9) et (10) sont inchangées ; nous ne les réécrirons pas.

L'expression algébrique de $\tilde{\varphi}_{I,K}$, (fonction de base locale) est très simple ; nous avons la :

Proposition

La fonction de base locale $\tilde{\varphi}_{I,K}$, est identiquement égale à la j° coordonnée barycentrique λ_j relativement à l'élément K :

(42) $\tilde{\varphi}_{j,K}(M) = \lambda_j(M) \qquad \forall M \in K$

Démonstration

Un point $M \in K$ admet un unique triplet de coordonnées barycentriques $\lambda_1, \lambda_2, \lambda_3$ qui par définition sont les poids qu'il faut donner aux sommets A_1, A_2, A_3 pour que le barycentre pondéré soit exactement égal à M. Par définition, la somme de ces poids vaut 1. Nous avons donc :

(43) $$\sum_{j=1}^{3} X(K,j)\ \lambda_j(M) = X$$

(44) $$\sum_{j=1}^{3} Y(K,j)\ \lambda_j(M) = Y$$

(45) $$\sum_{j=1}^{3} \lambda_j(M) = 1$$

La fonction $M \mapsto \lambda_j(M)$ est clairement fonction linéaire du triplet (X,Y,1), c'est-à-dire fonction affine de (X,Y). De plus, prenant dans le second membre la valeur particulière X = X(K,i), Y = Y(K,i) (coordonnées au sommet A_i), il est évident qu'alors λ_j (j° coordonnée barycentrique) vaut 1 pour j = i et 0 si $j \neq i$ (le **vérifier** sur les relations (43)-(45) si vous n'êtes pas convaincus !) :

(46) $\lambda_j(A_i) = \delta_{ij}$

La fonction λ_j est donc affine, vaut 1 pour le sommet de numéro local j et 0 pour les deux autres sommets ; elle est donc identiquement égale à la restriction dans l'élément K, de la fonction de base associée à ce sommet. C'est exactement ce qu'exprime la relation (42) ! ■

L'expression algébrique de $\tilde{\varphi}_{j,K}$ s'obtient alors facilement en fonction de X,Y, par résolution du système (43) (45) de trois équations à trois inconnues. Nous explicitons ce calcul élémentaire, en adoptant une notation simplifiée le temps du calcul algébrique :

(47) $x_j \equiv X(K,j), \quad x \equiv X, \quad y \equiv Y$

On a :

(48) $x_1 \lambda_1 + x_2 \lambda_2 + x_3 \lambda_3 = x$

(49) $y_1 \lambda_1 + y_2 \lambda_2 + y_3 \lambda_3 = y$

(50) $\lambda_1 + \lambda_2 + \lambda_3 = 1$

on retranche x_3 de (48), en l'écrivant $x_3 (\lambda_1 + \lambda_2 + \lambda_3)$ et on procède de façon analogue pour y_3 avec la relation (49). Il vient :

(51) $(x_1-x_3) \lambda_1 + (x_2-x_3) \lambda_2 = x-x_3$

(52) $(y_1-y_3) \lambda_1 + (y_2-y_3) \lambda_2 = y-y_3$

et, en remarquant que :

(53)
$$\begin{cases} (y_2-y_3)(x_1-x_3) - (x_2-x_3)(y_1-y_3) = \\ = (y_2-y_1)(x_1-x_3) - (x_2-x_1)(y_1-y_3) \\ = (x_2-x_1)(y_3-y_1) - (y_2-y_1)(x_3-x_1) \end{cases}$$

on trouve l'expression suivante pour λ_1 :

(54) $$\lambda_1 (x,y) = \frac{(y_2-y_3)(x-x_3) - (x_2-x_3)(y-y_3)}{(x_2-x_1)(y_3-y_1) - (y_2-y_1)(x_3-x_1)}$$

Les autres coordonnées barycentriques se déduisent de (54) par permutation circulaire des indices 1, 2, 3. On remarque que le dénominateur est le double de la surface du triangle K, puisque égal au produit mixte suivant :

(55) $$S_K = \frac{1}{2} \left(\overrightarrow{A_1A_2} \times \overrightarrow{A_1A_3}, k \right)$$

ceci dans l'hypothèse où **l'orientation des sommets A_1, A_2, A_3 est "directe"**, ce qui revient à dire que le déterminant de F_K est positif. Cette remarque constitue en fait une **règle de numérotation locale**. Il n'y a donc pas, même avec cette convention, une seule numérotation locale possible ; toute permutation paire de 1,2,3 permet de retrouver une nouvelle numérotation locale cohérente des noeuds de l'élément K.

• La notion de matrice élémentaire est fondée, comme dans le cas monodimensionnel, sur la propriété suivante :

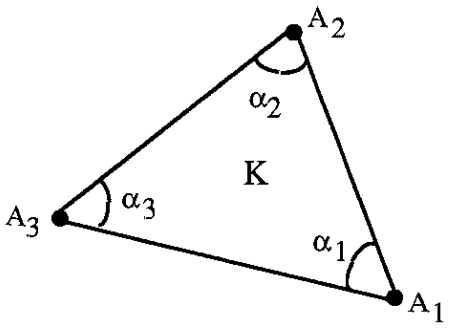
Propositton :

L'intégrale $\int_K \nabla\varphi_{II} \bullet \nabla\varphi_{JJ}\, dx$ ne peut être non nulle que si les sommets II et JJ sont tous deux des sommets appartenant au bord de l'élément K.

La preuve donnée pour le cas monodimensionnel est inchangée. La définition des matrices élémentaires AELT (K,I,J) est encore obtenue par la relation (11), mais I et J varient cette fois entre 1 et 3. Le calcul de cette matrice élémentaire est un exercice laissé au lecteur (compte tenu de l'explicitation de la relation (54) donnant les fonctions de base $\tilde{\varphi}_{j,K}$). Nous pouvons de plus en donner un résultat sous forme géométrique.

Proposition :

Avec les notations introduites à la figure ci-contre, on a :



(56) $$\int_K \nabla\tilde{\varphi}_{1,K} \bullet \nabla\tilde{\varphi}_{2,K}\, dx = -\frac{1}{2\,\mathrm{tg}\alpha_3}$$

(57) $$\int_K \left|\nabla\tilde{\varphi}_{1,K}\right|^2 dx = \frac{\sin\alpha_1}{2\sin\alpha_2 \sin\alpha_3}$$

Proposition

Les matrices élémentaires sont assemblées pour former la matrice globale A(II,JJ). L'algorithme (16) qui consiste à "éclater" la matrice élémentaire AELT (K,•,•) au sein de la matrice assemblée est inchangé. De même, on peut définir un second membre élémentaire BELT (K,I) à l'aide de la relation (19). L'assemblage du second membre s'obtient encore à l'aide de la relation (20).

4) Condition de Dirichlet non homogène

• Nous avons jusqu'ici supposé que la condition de Dirichlet à prendre en compte est homogène : pour les relations (2) et (22), la donnée de u(•) sur le bord de Ω est la fonction nulle. Nous supposons dans ce paragraphe que ce n'est plus le cas, c'est-à-dire que u est donnée non nulle sur une partie Γ_1 du bord.

(58) $\quad u = \overline{u_o} \quad$ sur Γ_1

Sur la partie Γ_2 complémentaire de Γ_1, nous adoptons une condition de Neumann.

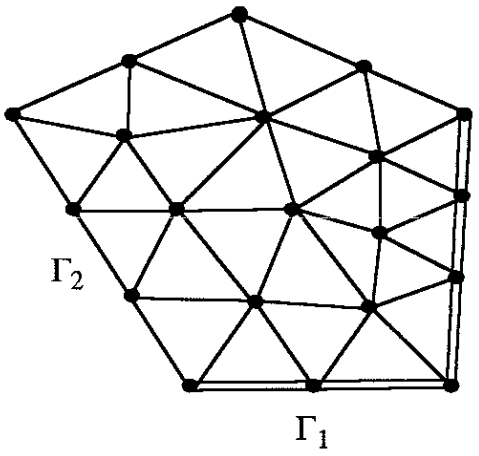
(59) $\quad \dfrac{\partial u}{\partial n} = g \quad$ sur Γ_2

et nous continuons à résoudre un problème de Poisson dans le domaine d'étude.

(60) $\quad -\Delta u = f \quad$ dans Ω.

Les degrés de liberté du problème discrétisé sont les sommets A_j du maillage τ. On numérote en premier les sommets qui n'appartiennent pas à Γ_1.

(61) $\quad$ sommets n'appartenant pas à Γ_1 : $A_1, A_2, \ldots, A_N$



Choix d'une numérotation pour les noeuds ; on numérote en dernier les sommets de Γ_1

puis ensuite ceux qui appartiennent à Γ_1.

(62) $\quad$ sommets de τ sur Γ_1 : $A_{N+1}, \ldots, A_{N+M}$.

On dispose de ce fait de deux espaces de dimension finie V_h et W_h : les fonctions de ces deux espaces sont continues sur $\overline{\Omega}$, affines dans chaque triangle K ; les fonctions de V_h sont

supposées nulles sur la partie Γ_1 du bord, alors que nous ne faisons pas cette hypothèse pour les fonctions de W_h. Avec les notations usuelles pour les fonctions de base, nous avons donc :

(63) $V_h = \text{Vect} < \varphi_1, \ldots, \varphi_N >$

(64) $W_h = \text{Vect} < \varphi_1, \ldots, \varphi_N, \varphi_{N+1}, \ldots, \varphi_{N+M} >$

et bien entendu, les dimensions de ces deux espaces s'en déduisent :

(65) $\dim V_h = N, \qquad \dim W_h = N+M.$

• Ces préliminaires étant posés, nous allons développer deux approches pour la prise en compte de la condition de Dirichlet non homogène (58) : l'une théorique qui consiste à se ramener au cas où $\bar{u}_o$ est nulle grâce à un relèvement de cette condition non homogène, l'autre pratique où l'**on "élimine les noeuds bloqués"** qui est mise en oeuvre au sein des algorithmes de calcul.'

Le **relèvement** dans l'espace W_h de la condition de Dirichlet non homogène $\bar{u}_o$ consiste à chercher une fonction $\tilde{u}_o$ telle que

(66) $\tilde{u}_o \in W_h$

(67) $\tilde{u}_o (A_k) = \bar{u}_o (A_k) \quad \forall k = N+1, \ldots, N+M :$

la fonction $\tilde{u}_o$ est définie par ses valeurs aux sommets du maillage, est donnée égale à $\tilde{u}_o$ pour les sommets de Γ_1 (de numéros N+1 à N+M) et n'est pas déterminée a priori pour les autres sommets. Nous pouvons par exemple choisir :

(68) $\tilde{u}_o (A_j) = 0 \qquad \forall j = 1, \ldots, N$

mais tout autre choix est possible aussi. Nous faisons le choix (68) car il simplifie un peu les calculs. La fonction $\tilde{u}_o$ peut donc se calculer sans difficulté compte tenu des conditions (66) (67) (68) et la condition de Dirichlet (58) peut se réécrire de façon équivalente :

(69) $u = \tilde{u}_o + z, \qquad z \in V_h$

ce qui ne fait qu'exprimer que u s'obtient à partir de $\tilde{u}_o$ grâce à une fonction z nulle sur Γ_1. La formulation variationnelle s'obtient en multipliant la relation (60) par une fonction test v appartenant à V_h et en intégrant par parties, ce qui permet de prendre en compte la condition de Neumann (59). Il vient :

(70) $$\int_\Omega \nabla u \bullet \nabla v \, dx = \int_\Omega fv \, dx + \int_{\Gamma_2} gv \, d\gamma , \quad \forall\, v \in V_h$$

En prenant en compte la relation (69), on obtient une formulation variationnelle discrète pour la fonction z.

(71) $$z \in V_h$$

(72) $$\int_\Omega \nabla z \bullet \nabla v \, dx = \int_\Omega fv \, dx + \int_{\Gamma_2} gv \, d\gamma - \int_\Omega \nabla \tilde{u}_o \bullet \nabla v \, dx, \quad \forall\, v \in V_h$$

Le système linéaire décrit par (71) (72) prend la forme classique.

(73) $$A_1 Z = F + G - B \bar{u}_o$$

avec des notations naturelles rappelées ci-dessous :

(74) $$(A_1)_{ij} = \int_\Omega \nabla \varphi_i \bullet \nabla \varphi_j \, dx \qquad 1 \le i, j \le N$$

(75) $$F_i = \int_\Omega f \, \varphi_i \, dx$$

(76) $$G_i = \int_{\Gamma_2} g \, \varphi_i \, d\gamma$$

(77) $$B_{ik} = \int_\Omega \nabla \varphi_i \bullet \nabla \varphi_k \, dx, \quad \begin{cases} 1 \le i \le N \\ N+1 \le k \le N+M \end{cases}$$

• Tout est bien établi du point de vue théorique ; du point de vue pratique, le calcul explicite de la fonction $\tilde{u}_o$ n'est jamais réalisé. On adopte une autre démarche, fondée sur l'écriture de la relation (70) non pas pour $v \in V_h$, mais pour $w \in W_h$. L'écriture des termes de bord est un peu différente et nous avons en toute rigueur

(78) $$\int_\Omega \nabla u \bullet \nabla w \, dx = \int_\Omega fw \, dx + \int_{\Gamma_2} gw \, d\gamma + \int_{\Gamma_1} \frac{\partial u}{\partial n} w \, d\gamma, \quad \forall\, w \in W_h$$

expression qui est identique à (70) si nous particularisons le choix (70). La matrice de ce système (78), qui suppose $u \in W_h$ mais oublie (pour un moment) la relation (69), est donnée par :

(79) $$A_{ij} = \int_\Omega \nabla\varphi_i \bullet \nabla\varphi_j \; dx \qquad 1 \leq i, j \leq N+M$$

et l'on notera que la seule différence entre (79) et (74) est la borne supérieure de variation du double indice (i, j). De façon plus précise, la matrice A se décompose en blocs sous la forme :

(80) $$A = \begin{pmatrix} A_1 & B_1 \\ {}^tB_1 & A_2 \end{pmatrix}.$$

Quand on explicite la relation (78), en recherchant l'inconnue u sous la forme :

(81) $$u = z + \sum_{k=N+1}^{N+M} \lambda_k \varphi_k$$

(sans savoir pour l'instant que $\lambda_k = \bar{u}_o\ (A_k)$!), on obtient un système qui s'écrit par blocs :

(82) $$\begin{pmatrix} A_1 & B_1 \\ {}^tB_1 & A_2 \end{pmatrix} \begin{pmatrix} Z \\ \Lambda \end{pmatrix} = \begin{pmatrix} F+G \\ * \end{pmatrix}.$$

On reconnait la relation (70) pour les N premières lignes de (82) alors qu'on n'a pas explicité le second membre des M dernières.

• **L'élimination des noeuds bloqués** consiste à transformer le système (82) en écrivant brutalement la relation (69) à la place de la seconde ligne, c'est-à-dire ici

(83) $$\Lambda = \bar{u}_o,$$

en éliminant d'une part les matrices blocs ${}^tB_1$ et A_2 et en les remplaçant par 0 et la matrice Id respectivement, en modifiant le second membre pour "forcer" la relation (83) d'autre part. Il vient :

(84) $$\begin{pmatrix} A_1 & B_1 \\ 0 & Id \end{pmatrix} \begin{pmatrix} Z \\ \Lambda \end{pmatrix} = \begin{pmatrix} F+G \\ \bar{u}_o \end{pmatrix}$$

qui a l'inconvénient de faire apparaître une matrice non symétrique. Ce défaut peut être corrigé en éliminant Λ de la première équation et nous déduisons :

$$(85) \qquad \begin{pmatrix} A_1 & 0 \\ 0 & Id \end{pmatrix} \begin{pmatrix} Z \\ \Lambda \end{pmatrix} = \begin{pmatrix} F+G - B_1\,\bar{u}_o \\ \bar{u}_o \end{pmatrix}$$

qui est parfaitement équivalent à la relation (73).

L'intérêt de cette dernière approche est qu'elle est généralisable facilement au cas où l'on n'a pas adopté de numérotation particulière pour les degrés de liberté sur Γ_1 (cf. relations (61) et (62)). On assemble la matrice A et le second membre F+G grâce aux algorithmes (16) et (20). Puis on élimine les noeuds bloqués, opération qui consiste à passer de (82) à (85). En pratique, on a donc l'algorithme suivant :

boucle sur les noeuds : $1 \leq i \leq N+M$

i est-il un noeud sur Γ_1 ?

non → passer à i+1

oui

  boucle sur les noeuds : $1 \leq j \leq N+M$

  j est-il un noeud sur Γ_1 ?

  non :

    $B_j \leftarrow B_j - A_{ji}\ \bar{u}_o\ (A_i)$

    $A_{ji} \leftarrow 0$

  oui → passer à j+1

  boucle sur les noeuds : $1 \leq j \leq N+M$

  j = i ?

  non :

    $A_{ij} \leftarrow 0$

  oui :

    $A_{ii} \leftarrow 1$

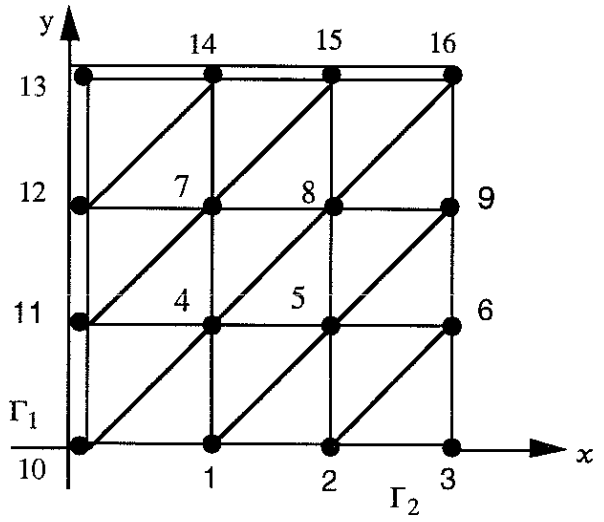
Notons enfinque l'adressage matriciel est rarement à double indice comme écrit à l'algorithme (86), ce qui en pratique en rend la lecture plus délicate ....

5) Étude d'un exemple, dit "Hadhri"

On considère le domaine $\Omega =]0,3[\times]0,3[$, sa frontière $\partial\Omega$ est décomposée en $\Gamma_1 = \{0\} \times]0,3[\cup]0,3[\times \{3\}$, et $\Gamma_2 =]0,3[\times \{0\} \cup \{3\} \times]0,3[$ (voir figure ci-dessous). On veut résoudre le problème.

$$(87) \qquad \begin{cases} -\Delta u = f & \Omega \\ u = 0 & \Gamma_1 \\ \dfrac{\partial u}{\partial n} = g & \Gamma_2 \end{cases}$$

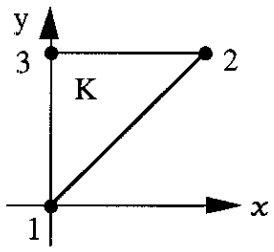
de façon approchée, à l'aide de la méthode des éléments finis, avec le maillage (de 18 éléments triangulaires) proposé à la figure ci-dessous. Les sommets de numéros 1 à 9 engendrent l'espace V_h alors que les sommets 10 à 16 sont situés sur Γ_1.



maillage en triangles de :

$\Omega =]0,3[\times]0,3[$

• Compte tenu de la simplicité du maillage, le calcul des matrices élémentaires se réduit aux deux cas de figure suivants :

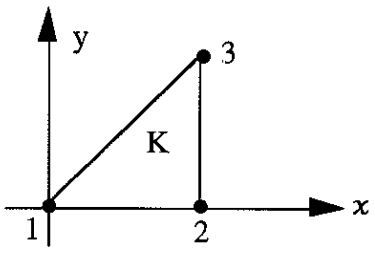
1) 

$\varphi_1(x,y) = 1-y$

$\varphi_2(x,y) = x$

$\varphi_3(x,y) = y-x$

matrice élémentaire AELT = $\begin{pmatrix} \frac{1}{2} & 0 & -\frac{1}{2} \\ 0 & \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & 1 \end{pmatrix}$

2) 

$\varphi_1(x,y) = 1-x$

$\varphi_2(x,y) = x-y$

$\varphi_3(x,y) = y$

matrice élémentaire AELT = $\begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & 0 \\ -\frac{1}{2} & 1 & -\frac{1}{2} \\ 0 & -\frac{1}{2} & \frac{1}{2} \end{pmatrix}$.

• Il est utile aussi de calculer la matrice élémentaire de masse, ie :

(88) $\quad MELT(K, I, J) = \int_K \tilde{\varphi}_{I,K}\ \tilde{\varphi}_{J,K}\, dx$.

Dans les deux cas précédents, on obtient :

(89) $\quad MELT = \frac{1}{24} \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$.

• Si on interpole les fonctions f et g dans l'espace W_h (avec les notations précédentes), on doit calculer les seconds membres :

$$(90) \qquad (MF)_i = \sum_{j=1}^{16} \left(\int_\Omega \varphi_i \varphi_j \, dx \right) f_j \, , \; 1 \le i \le 9$$

et

$$(91) \qquad (\Gamma G)_i = \sum_{j=1}^{16} \left(\int_{\Gamma_2} \varphi_i \varphi_j \, d\gamma \right) g_j \, , \; 1 \le i \le 9$$

Le calcul (89) de la matrice élémentaire et la relation (90) conduisent à une matrice de masse M d'ordre 9 x 16 (9 lignes et 16 colonnes) dont nous donnons ci-dessous les éléments non nuls.

$$(92) \quad M = \frac{1}{24} \begin{pmatrix} 6 & 1 & 0 & 2 & 2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 6 & 1 & 0 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 12 & 2 & 0 & 2 & 2 & 0 & 2 & 2 & 0 & 0 & 0 & 0 & 0 \\ 2 & 2 & 0 & 2 & 12 & 2 & 0 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 & 2 & 6 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 24 & 2 & 0 & 0 & 2 & 2 & 0 & 2 & 2 & 0 \\ 0 & 0 & 0 & 2 & 2 & 0 & 2 & 24 & 2 & 0 & 0 & 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 0 & 2 & 1 & 0 & 2 & 6 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

La relation (91) est non triviale pour i = 1,2,3,6,9 (ce sont les numéros des sommets sur le bord Γ_2) et nous avons finalement :

$$(93) \qquad \begin{cases} (\Gamma G)_1 = \frac{2}{3} g_1 + \frac{1}{6} g_2 + \frac{1}{6} g_{10} \\ (\Gamma G)_2 = \frac{1}{6} g_1 + \frac{2}{3} g_2 + \frac{1}{6} g_3 \\ (\Gamma G)_3 = \frac{1}{6} g_2 + \frac{2}{3} g_3 + \frac{1}{6} g_6 \\ (\Gamma G)_6 = \frac{1}{6} g_3 + \frac{2}{3} g_6 + \frac{1}{6} g_9 \\ (\Gamma G)_9 = \frac{1}{6} g_6 + \frac{2}{3} g_9 + \frac{1}{6} g_{16} \end{cases}$$

puisque nous avons :

(94) $$\int_{\text{côté}} \varphi_1^2 \, d\gamma = \frac{1}{3} \quad \text{et} \quad \int_{\text{côté}} \varphi_1 \, \varphi_2 \, d\gamma = \frac{1}{6}$$

- Le système d'équations à résoudre s'écrit :

(95) $$AU = MF + \Gamma G$$

avec A obtenu également par assemblage des matrices élémentaires de rigidité, ce qui conduit à :

$$A = \begin{bmatrix}
2 & -\frac{1}{2} & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
-\frac{1}{2} & 2 & -\frac{1}{2} & 0 & -1 & 0 & 0 & 0 & 0 \\
0 & -\frac{1}{2} & 1 & 0 & 0 & -\frac{1}{2} & 0 & 0 & 0 \\
-1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0 \\
0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\
0 & 0 & -\frac{1}{2} & 0 & -1 & 2 & 0 & 0 & -\frac{1}{2} \\
0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 & 0 \\
0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 \\
0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & 0 & -1 & 2
\end{bmatrix}$$

Nous laissons au lecteur le soin d'expliciter la matrice B_1 (cf. relation (73)) pour transformer le système (95) lorsqu'une condition de Dirichlet non homogène est à prendre en compte sur Γ_1.

IX - COMPLÉMENTS

i) Formule d'intégration par parties

Nous établissons la formule d'intégration par parties :

(1) $$\int_\Omega u \frac{\partial v}{\partial x_j}\, dx = -\int_\Omega \frac{\partial u}{\partial x_j} v\, dx + \int_{\partial\Omega} uv\, n_j\, d\gamma$$

dans le cas où Ω est un ouvert de $\mathbb{R}^2$ de frontière polygonale.

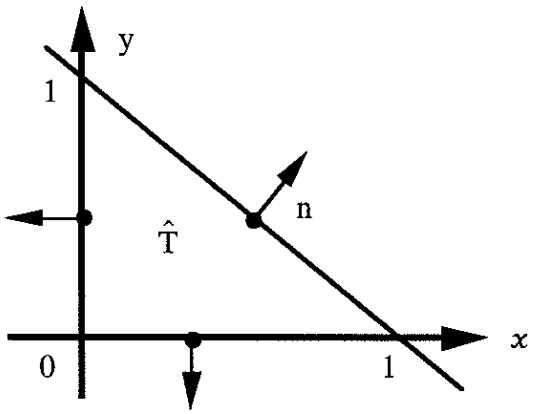
- Nous commençons par écrire la relation (1) sous la forme équivalente suivante :

(2) $$\int_\Omega \frac{\partial w}{\partial x_j}\, dx = \int_{\partial\Omega} w\, n_j\, d\gamma.$$

- En effet, prenant $u \equiv 1$, $v \equiv w$ dans la relation (1), nous établissons (2) et réciproquement, si (2) est vrai nous posons $w = uv$, ce qui donne (1) par un simple calcul de la dérivée d'un produit de deux fonctions.

Un ouvert de $\mathbb{R}^2$ à frontière polygonale est réunion de triangles donc nous commençons par le cas où Ω est le **triangle de référence** : $\hat{T}$

(3) $$\hat{T} \left\{ (x,y)\, x \geq 0,\ y \geq 0,\ x + y \leq 1 \right\}.$$



Triangle de référence

La normale extérieure sur le bord de $\hat{T}$ est donnée par :

(4) $\quad n = (0, -1) \qquad$ sur $[0,1] \times \{0\}$

(5) $\quad n = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) \qquad$ sur $\{(x,y),\ x \geq 0,\ y \geq 0,\ x + y = 1\}$

(6) $\quad n = (-1, 0) \qquad$ sur $\{0\} \times [0,1]$.

Nous établissons la relation (2) pour j = 1, la propriété s'en déduisant pour j = 2 en échangeant les rôles de x et y. Nous montrons donc :

(7) $$\int\int_{\hat{T}} \frac{\partial w}{\partial x}\, dx\, dy = \int_{\partial \hat{T}} w\, n_x\, d\gamma$$

avec :

(8) $$\int_{\partial \hat{T}} w\, n_x\, d\gamma = \int_0^1 w\,(1-t, t)\, \frac{1}{\sqrt{2}}\,(\sqrt{2}dt) - \int_0^1 w\,(o,t)\, dt$$

Nous calculons l'intégrale double du membre de gauche de (7) par la formule de Fubini, en commençant par l'intégrale en y :

(9) $$\int\int_{\hat{T}} \frac{\partial w}{\partial x}\, dx\, dy = \int_0^1 dy \int_0^{1-y} dx\, \frac{\partial w}{\partial x}\,(x,y)$$

(10) $$\int\int_{\hat{T}} \frac{\partial w}{\partial x}\, dx\, dy = \int_0^1 dy\, \left[w\,(1-y, y) - w\,(0,y)\right]$$

qui est, à la notation près, identique au second membre de (8), ce qui établit la propriété dans ce cas.

• Nous établissons la relation (2) dans le cas où Ω est un **triangle quelconque** T = (A,B,C). Un tel triangle est image de $\hat{T}$ par la transformation suivante :

(11) $\quad M = (1 - \hat{x} - \hat{y})\ A + \hat{x}\ B + \hat{y}\ C, \qquad \hat{M} = (\hat{x}, \hat{y}) \in \hat{T}$

dont la jacolienne $J = \frac{\partial M}{\partial \hat{M}}$ est une constante égale au double de la surface du triangle T :

(12) $$J \equiv \frac{\partial M}{\partial \hat{M}} = 2\,|T|$$

Nous calculons une intégrale sur T par changement de variable en la ramenant sur $\hat{T}$:

$$(13) \qquad \int_T u(M)\, dx = J \int_{\hat{T}} u\left(M(\hat{M})\right) d\hat{x}$$

donc :

$$(14) \qquad \int_T \frac{\partial W}{\partial x_j}\, dx = J \int_{\hat{T}} \sum_{k=1}^{2} \frac{\partial w}{\partial \hat{x}_k} \frac{\partial \hat{x}_k}{\partial x_j}\, d\hat{x}$$

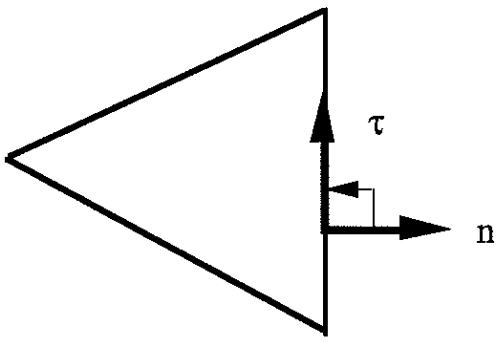
$$(15) \qquad \int_T \frac{\partial w}{\partial x_j} dx = J \sum_{k=1}^{2} \frac{\partial \hat{x}_k}{\partial x_j} \int_{\partial \hat{T}} w\, \hat{n}_k\, d\hat{\gamma}$$

compte tenu du fait que la formule de Green est vraie dans le triangle $\hat{T}$. Nous devons maintenant transformer l'expression de $\hat{n}_k$ et $d\hat{\gamma}$ afin de faire apparaître une intégrale sur le bord du triangle T. Les directions tangentielles de $\partial\hat{T}$ sont images de celles de ∂T dans la transformation $T \rightarrow \hat{T}$ inverse de (11). Nous avons donc :

$$(16) \qquad \hat{\tau}_k = \alpha \sum_{k=1}^{2} \frac{\partial \hat{x}_k}{\partial x_m}\, \tau_m$$

à un coefficient de proportionalité près qui assure que $\hat{\tau}$ et τ sont unitaires. Cette relation montre que l'élément de longueur $d\hat{\gamma} = \sqrt{d\hat{M}^2}$ peut s'écrire en fonction de $d\gamma = \sqrt{dM^2}$ le long de ∂T selon le changement de variables

$$(17) \qquad d\hat{\gamma} = \left\{ \sum_{k,\ell,m} \frac{\partial \hat{x}_k}{\partial x_\ell} \frac{\partial \hat{x}_k}{\partial x_m}\, \tau_\ell\, \tau_m \right\}^{\frac{1}{2}} d\gamma$$



Vecteur tangent τ le long de ∂T et vecteur normal n vers l'extérieur

ou en échangeant les rôles de T et $\hat{T}$:

$$(18) \qquad d\hat{\gamma} = \left\{ \sum_{k,\ell,m} \frac{\partial x_k}{\partial \hat{x}_\ell} \frac{\partial x_k}{\partial \hat{x}_m} \hat{\tau}_k \hat{\tau}_m \right\}^{-\frac{1}{2}} d\gamma$$

De plus, puisque le produit scalaire $\hat{n} \bullet \hat{\tau}$ est nul, nous avons :

$$(19) \qquad \sum_{m=1}^{2} \left\{ \sum_{k=1}^{2} \frac{\partial \hat{x}_k}{\partial x_m} \hat{n}_k \right\} \tau_m = 0$$

ce qui montre que la $m^{ième}$ composante n_m du vecteur normal n est proportionnelle à l'expression entre parenthèses dans la relation (19), puisque $n \bullet \tau$ est nul :

$$(20) \qquad n_m = \beta \sum_k \frac{\partial \hat{x}_k}{\partial x_m} \hat{n}_k$$

avec

$$(21) \qquad \beta^2 \left\{ \sum_m \left(\sum_k \frac{\partial \hat{x}_k}{\partial x_m} \hat{n}_k \right)^2 \right\} = 1$$

puisque n est un vecteur unitaire. Nous supposons pour simplifier les calculs (et ne pas écrire de signe qui in fine arriveront en nombre pair !) que la transformation $\hat{M} \rightarrow M$ est **directe**, ie conserve l'orientation. Une direction extérieure à $\hat{T}$ se transforme donc en une direction extérieure à T puisque $\hat{\tau}$ se transforme en τ et $(\hat{n}, \hat{\tau})$ d'une part, (n, τ) d'autre part, sont deux angles droits de $+\frac{\Pi}{2}$. Donc β est positif. Nous obtenons donc, en utilisant la relation (20) au sein du second membre de la relation (15) :

$$(22) \qquad \int_T \frac{\partial w}{\partial x_j} dx = \int_{\partial T} J \frac{1}{\beta} n_j w \frac{d\gamma}{\left\{ \sum_{k,\ell,m} \frac{\partial x_k}{\partial \hat{x}_\ell} \frac{\partial x_k}{\partial \hat{x}_m} \hat{\tau}_k \hat{\tau}_m \right\}^{\frac{1}{2}}}$$

compte tenu de la relation (18). La fin de la preuve résulte d'un calcul élémentaire sur l'inverse d'une matrice 2x2. Nous notons provisoirement la jacobienne de la transformation (11) sous la forme :

$$(23) \qquad \frac{\partial M}{\partial \hat{M}} = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$$

et détaillons le calcul de $\frac{1}{\beta^2}$ à l'aide de (21), en tenant compte du fait que :

(24) $\quad \hat{n}_1 = -\hat{\tau}_2 \ , \ \hat{n}_2 = \hat{\tau}_1$

et de la propriété classique :

(25) $$\frac{\partial \hat{M}}{\partial M} = \frac{1}{J} \begin{pmatrix} d & -c \\ -b & a \end{pmatrix}$$

Nous avons :

$$\frac{1}{\beta^2} = \frac{1}{J^2} \left[\left(d\,(-\hat{\tau}_2) - c\,\hat{\tau}_1 \right)^2 + \left((-b)\,(-\hat{\tau}_2) + a\,\hat{\tau}_1 \right) \right]$$

$$= \frac{1}{J^2} \left(\left(a\,\hat{\tau}_1 + b\,\hat{\tau}_2 \right)^2 + \left(c\,\hat{\tau}_1 + d\,\hat{\tau}_2 \right)^2 \right)$$

c'est-à-dire :

(26) $$\frac{1}{\beta^2} = \frac{1}{J^2} \sum_{k\,\ell\,m} \frac{\partial x_k}{\partial \hat{x}_\ell} \frac{\partial x_k}{\partial \hat{x}_m} \hat{\tau}_k \hat{\tau}_m$$

La relation (22) se simplifie alors en :

(27) $$\int_T \frac{\partial w}{\partial x_j} dx = \int_{\partial T} w\, n_j\, d\gamma$$

ce qui prouve la propriété dans le cas où Ω est un triangle quelconque du plan.

• Nous abordons maintenant le cas général. L'ouvert Ω étant polygonal, on peut le recouvrir par un maillage de type éléments finis, formé de triangles T :

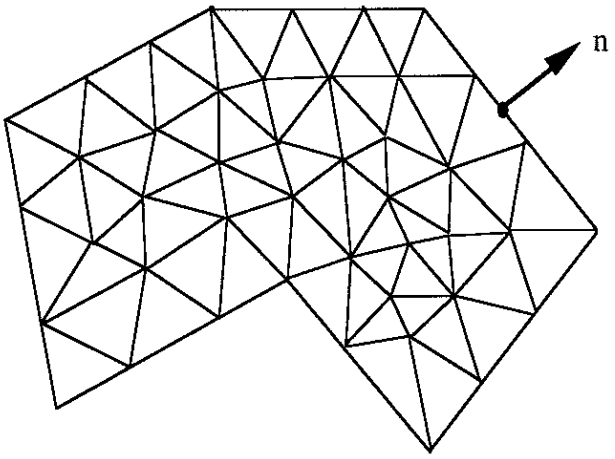
(28) $$\overline{\Omega} = \bigcup_{T \in \tau} \overline{T}$$

où l'intersecton $\overline{T} \cap \overline{T'}$ de deux triangles de la triangulation τ est formée de l'ensemble vide, d'un sommet de T et T', d'une arête commune à T et T' ou est égale à T et T'. On découpe alors l'intégrale (2) en autant d'intégrales sur les triangles du maillage.

(29) $$\int_\Omega \frac{\partial w}{\partial x_j} dx = \sum_{T \in \tau} \int_T \frac{\partial w}{\partial x_j} dx$$

pour lesquelles la relation (2) est applicable :

$$(30) \qquad \int_{\Omega} \frac{\partial w}{\partial x_j} \, dx = \sum_{T \in \tau} \int_{\partial T} w \, n_j \, d\gamma$$

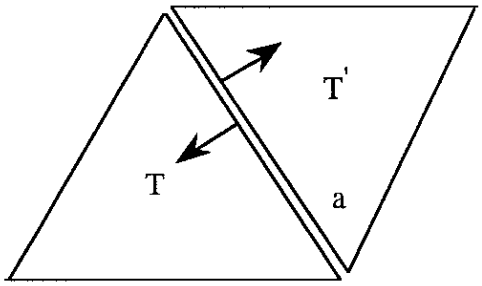


Triangulaire d'un ouvert Ω polygonal du plan

Le membre de droite de la relation (30) est une somme d'intégrales sur les arêtes a du maillage. Si l'arête a est intérieure au domaine, elle appartient au bord de deux triangles T et T' donc contribue par les deux intégrales suivantes :

$$(31) \qquad \int_{a \cap \partial T} w \, n_j \, d\gamma + \int_{a \cap \partial T'} w \, n_j \, d\gamma = 0$$

dont la somme est nulle car les normales extérieures à T et T' sont opposées.



Arête a intérieure au domaine Ω

Si l'arête a est au bord du domaine, la normale extérieure à T coïncide avec la normale extérieure à ∂Ω et l'arête n'apparaît qu'une seule fois dans la somme (30) puisqu'elle n'appartient qu'à un seul triangle du maillage. Nous avons donc :

(32) $$\int_{\Omega} \frac{\partial w}{\partial x_j}\, dx = \sum_{\substack{a \in \tau \\ a \subset \partial\Omega}} \int_a w\ n_j\ d\gamma$$

qui peut s'écrire plus simplement (puisque la réunion des arêtes du bord de Ω constitue le bord lui-même) :

(33) $$\int_{\Omega} \frac{\partial w}{\partial x_j}\, dx = \int_{\partial\Omega} w\, n_j\, d\gamma\,,$$

ce qui montre la propriété.

ii) Algorithme du gradient conjugué

A désigne une matrice carrée d'ordre n symétrique définie positive et b un vecteur de $\mathbb{R}^n$.

1) Montrer que la résolution du système linéaire

(1) $A\ x = b$

équivaut à rechercher le minimum, pour x appartenant à $\mathbb{R}^n$, de la fonctionnelle.

(2) $J(x) = \frac{1}{2}\ (x,\ A\,x) - (b, x)$

2) Description de l'algorithme du gradient conjugué

- Initialisation : $x^o \in \mathbb{R}^n$, $w^o = g^o = A\,x^o - b$.
- Itération de l'algorithme.

On suppose connus l'état x^{k-1} après la kième itération ainsi que la direction de descente w^{k-1}. L'état suivant x^k de l'algorithme est issu de x^{k-1} via un incrément proportionnel à la direction de descente :

(3) $x^k = x^{k-1} + \rho^{k-1} w^{k-1}$

de façon à minimiser la fonctionnelle J :

(4) $J(x^k) \leq J(x^{k-1} + \rho\, w^{k-1}) \quad \forall\ \rho \in \mathbb{R}$

On introduit également le gradient de J au point x^k.

(5) $g^k = A\,x^k - b$.

2.1) Calculer le coefficient ρ^{k-1} en fonction de g^{k-1}, w^{k-1} et A.

2.2) Montrer que l'on a :

(6) $(g^k, w^{k-1}) = 0$.

La nouvelle direction de descentre w^k est recherchée sous la forme :

(7) $\quad w^k = g^k + \alpha^k w^{k-1}$

de sorte que w^k soit orthogonal à la direction w^{k-1} pour le produit scalaire associé à la matrice A, c'est-à-dire :

(8) $\quad (w^k, A\ w^{k-1}) = 0$

2.3) Calculer la valeur de α^k en fonction de g^k, w^{k-1} et A.

On notera que le choix plus simple $\alpha^k = 0$ conduit à la méthode du gradient simple, où la direction de descente de l'algorithme w^k correspond à la ligne de plus grande pente g^k.

3) L'algorithme converge en au plus n étapes

3.1) Remarquer que si les gradients successifs $g^o, g^1, \ldots, g^{k-1}$ sont non nuls mais que g^k est nul, alors l'état x^k est la solution du système linéaire (1).

3.2) On suppose dans cette question que tous les gradients g^i sont non nuls jusqu'à l'étape m incluse. Montrer qu'on a alors les relations d'orthogonalité suivantes :

(9) $\quad (g^k, w^j) = 0, \qquad 0 \le j < k \le m$

(10) $\quad \rho^j \neq 0 \text{ et } w^j \neq 0, \qquad 0 \le j < k \le m$

(11) $\quad (g^k, g^j) = 0, \qquad 0 \le j < k \le m$

(12) $\quad (w^k, A\,w^j) = 0, \qquad 0 \le j < k \le m.$

On pourra raisonner par récurrence sur k pour l'ensemble des quatre propriétés et démontrer les relations (9) à (12) dans l'ordre où elles apparaissent.

3.3) Montrer que l'algorithme converge en au plus n itérations.

4) Propriétés auxiliaires

4.1) Montrer que l'on a :

$$(13) \quad \alpha^k = \frac{\|g^k\|^2}{\|g^{k-1}\|^2}$$

En déduire que le calcul très précis des produits scalaires est crucial pour assurer le succès pratique (informatique) de la convergence de la méthode.

4.2) Montrer que l'état x^k réalise le minimum de la fonctionnelle J sur le sous-espace affine $x^o + < w^o, \ldots, w^{k-1} >$. Quel commentaire pouvez-vous faire ?

Solution

1) La fonctionnelle a bien un minimum car la matrice A est symétrique définie positive. L'équation d'Euler au minimum s'écrit ici :

$$J'(x) \equiv A\,x - b = 0,$$

ce qui montre la propriété.

2.1) $$(14) \quad \rho^{k-1} = -\frac{(g^{k-1},\, w^{k-1})}{(w^{k-1},\, A\,w^{k-1})}$$

2.2) La condition précédente a été obtenue en écrivant qu'au point x^k, la fonctionnelle J est minimale si on la restreint à la droite affine passant par x^{k-1} et dirigée par le vecteur w^{k-1}, ce qui s'exprime en annulant le gradient de J au point x^k dans la direction w^{k-1}, ce qu'exprime exactement la condition (6).

2.3) $$(15) \quad \alpha^k = -\frac{(g^k,\, A\,w^{k-1})}{(w^{k-1},\, A\,w^{k-1})}$$

3.1) Dès que l'un des gradients g^m est nul, on est en un point x^m qui est solution du système (1) et l'algorithme ne présente plus alors aucun intérêt puisque l'on a résolu le problème posé.

3.2) • La propriété est vraie pour k = 1.

On remarque que $w^o = g^o$ est non nul en vertu de la question précédente. On a ensuite $(g^1, w^o) = 0$ compte tenu de la relation (6). Si on considère la relation (7) pour k = 1 et qu'on la multiplie scalairement par g^1, on obtient $(g^1, w^1) = \|g^1\|^2$ ce qui montre que w^1 est non nul puisque g^1 est non nul et le numérateur comme le dénominateur de l'expression (14) qui permet de calculer ρ^o sont non nuls, donc ρ^o est non nul et la relation (10) est vraie à l'ordre j = 0. La relation (11) est une conséquence simple du choix de la direction de descente initiale : $(g^1, g^o) = (g^1, w^o) = 0$ compte tenu de la relation (9). Enfin, la relation (12) exprime simplement la relation (8) pour k = 1.

• On suppose les relations (9) à (12) vérifiées jusqu'à l'ordre k inclus et on les étend pour l'indice k+1, en supposant g^{k+1} non nul.

On remarque d'abord que la relation (3) entraîne clairement :

$$(16) \quad g^{k+1} = g^k + \rho^k A w^k$$

On a d'une part $(g^{k+1}, w^k) = 0$ compte tenu de la relation (6) et pour $j < k$, on a d'autre part :

$$\begin{aligned}(g^{k+1}, w^j) &= (g^{k+1} - g^k, w^j) + (g^k, w^j) \\ &= \rho^k (A w^k, w^j)\end{aligned}$$

compte tenu de (16) et de l'hypothèse de récurrence (9). Or cette dernière expression est nulle en vertu de l'hypothèse de récurrence (12), donc la relation (9) est établie.

Si on multiplie scalairement l'identité (7) écrite au rang k+1 par g^{k+1}, la relation (9) que nous venons de démontrer entraîne que l'on a :

$$(17) \quad (g^{k+1}, w^{k+1}) = \|g^k\|^2$$

et la relation (10) est alors une conséquence simple de la relation qui permet de calculer ρ^k.

Exprimons g^j grâce à la relation (7) (considérée avec k = j). Il vient :

$$(g^{k+1}, g^j) = (g^{k+1}, w^j) - \alpha^j (g^{k+1}, w^{j-1})$$

et cette expression est nulle compte tenu de la relation (9) prise à l'ordre k+1. Ceci montre la relation (11).

La relation (12) est la conséquence de la relation (8) pour j = k et du calcul suivant j < k :

$$(w^{k+1},\ A\,w^{j}) = (g^{k+1},\ A\,w^{j}) \text{ compte tenu de (7) et de (12),}$$

$$= (g^{k+1},\ \frac{1}{\rho^{j}}\,(g^{j+1} - g^{j})) \quad \text{en vertu de (16),}$$

$$= 0 \quad \text{compte tenu de (11).}$$

La propriété est donc démontrée par récurrence.

3.3) Compte tenu de la relation (11), la suite de gradients $g^{o}, g^{1}, \ldots, g^{k}$ est composée de vecteurs orthogonaux deux à deux jusqu'à un ordre m et ceci a lieu dans l'espace $\mathbb{R}^{n}$. La conclusion est alors claire.

4.1) Dans la relation (15), on exprime $A\,w^{k-1}$ au numérateur et au dénominateur à l'aide de la relation (16) et on utilise la relation (11) pour le numérateur. Il vient :

$$\alpha^{k} = \frac{\|g^{k}\|^{2}}{(g^{k-1},\ w^{k-1})}$$

et la relation (13) est alors conséquence directe de la relation (17).

4.2) Si on écrit l'inéquation d'Euler qui caractérise le minimum de la fonction convexe dérivable J sur le convexe fermé $x^{o} + \langle w^{o}, \ldots, w^{k-1} \rangle$, atteint au point y, il vient $J'(y) \bullet w^{j} = 0$ pour tout j = 1, 2, ... k−1. Mais le choix $y = x^{k}$ vérifie ces relations puisque d'une part x^{k} appartient au convexe $x^{o} + \langle w^{o}, \ldots, w^{k-1} \rangle$ compte tenu de l'initialisation de l'algorithme et de la relation constitutive (3), et d'autre part $J'(x^{k}) \bullet w^{j} = (g^{k}, w^{j})$ lequel est nul en vertu de la relation (9).

Cette propriété montre que le point x^{k}, conçu initialement à la relation (4) pour minimiser la fonctionnelle J le long de la droite affine passant par x^{k-1} et dirigée selon le vecteur w^{k-1} la minimise en fait sur tout un sous-espace de dimension k ! On touche là au génie de Hestenes et Stiefel qui ont inventé la méthode du gradient conjugué en 1952.

iii) Étude du schéma de Newmark

On se propose de définir une famille à deux paramètres pour approcher numériquement la solution du système différentiel

$$(D) \quad M \frac{d^2q}{dt^2} + C \frac{dq}{dt} + K\,q = p(t)$$

où q(t) est un vecteur d'état appartenant à $\mathbb{R}^m$ (modèle de structure à m degrés de liberté). On suppose connu à l'instant $t_n = n\Delta t$ le vecteur d'état $U = \begin{pmatrix} \dot{q} \\ q \end{pmatrix}$ et on cherche à calculer ce même vecteur à l'instant $t_{n+1} = (n+1)\,\Delta t$.

1) Écrire la formule de Taylor avec reste intégral pour $\dot{q}_{n+1}$ et q_{n+1}, en faisant apparaître la dérivée seconde $\ddot{q}(\theta)$ $(t_n < \theta < t_{n+1})$.

2) Quel schéma obtient-on si on suppose que dans la relation précédente, on remplace $\ddot{q}(\theta)$ par la moyenne, ie $\ddot{q}(\theta) \simeq \frac{1}{2}\left(\ddot{q}_n + \ddot{q}_{n+1}\right)$?

3) Même question, mais avec $\ddot{q}(\theta)$ remplacé par l'interpolé affine entre les valeurs $\ddot{q}_n$ et $\ddot{q}_{n+1}$:

$$\ddot{q}(\theta) = \ddot{q}_n + \frac{1}{\Delta t}(\theta - t_n)\left(\ddot{q}_{n+1} + \ddot{q}_n\right).$$

Le schéma de Newmark consiste à écrire :

$$(N) \quad \begin{cases} \dot{q}_{n+1} = \dot{q}_n + (1-\gamma)\,\Delta t\,\ddot{q}_n + \gamma\,\Delta t\,\ddot{q}_{n+1} \\ q_{n+1} = q_n + \Delta t\,\dot{q}_n + \Delta t^2\left(\frac{1}{2} - \beta\right)\ddot{q}_n + \Delta t^2\,\beta\,\ddot{q}_{n+1} \end{cases}$$

4) A quelles valeurs de (β, γ) correspondent les schémas calculés aux questions 2 et 3 ?

5) Montrer que si $\gamma \neq \frac{1}{2}$, le schéma est d'ordre 1 quelque soit β et pour $\gamma = \frac{1}{2}$, le schéma est d'ordre 2 si $\beta \neq \frac{1}{6}$ et d'ordre 3 si $\beta = \frac{1}{6}$.

6) Proposer un algorithme de résolution du schéma (N) qui prenne en compte la loi de la dynamique (D). (On pourra résoudre un système linéaire du type $S\ \ddot{q}_{n+1} = f$ avec $S = M + \gamma \Delta t\, C + \beta \Delta t^2 K$ et f un second membre qu'on précisera).

7) Pour étudier la stabilité du schéma, on se restreint dans la suite au cas où M = 1, C = 0, $K = \omega^2$, p = 0 dans le modèle (D). Montrer qu'alors le schéma de Newmark (N) s'écrit :

$$U_{n+1} = A\ U_n$$

où A est une matrice 2x2 qu'on précisera, on pourra introduire ξ tel que :

$$\xi^2 = \frac{\omega^2 \Delta t^2}{1+\beta\, \omega^2 \Delta t^2}.$$

8) Montrer, en étudiant le module des racines complexes conjuguées de l'équation caractéristique relative à la matrice A que :

* le schéma est instable pour $\gamma < \frac{1}{2}$,

* le schéma est stable inconditionnellement pour $\gamma \geq \frac{1}{2}$ et $\beta \geq \frac{1}{4}\left(\gamma+\frac{1}{2}\right)^2$,

* le schéma est stable sous la condition

$$(5) \qquad \left(\gamma+\frac{1}{2}\right)^2 - 4\beta \leq \frac{4}{\omega^2 \Delta t^2}$$

lorsque $\gamma \geq \frac{1}{2}$ et $\beta < \frac{1}{4}\left(\gamma+\frac{1}{2}\right)^2$

On tracera dans le plan (γ,β) ces domaines de stabilité.

9) Quel schéma "optimum" proposez-vous de choisir ?

10) Montrer que l'erreur relative d'amplitude des ondes d'une part calculées par le schéma de Newmark et d'autre part exacte est donnée par le développement.

$$\rho - 1 = -\frac{1}{2}\left(\gamma - \frac{1}{2}\right)\ \omega^2\ \Delta t^2\ +\ 0\left(\Delta t^4\right)$$

et que l'erreur de relative de périodicité, dans le cas $\gamma = \frac{1}{2}$, est donnée par :

$$\frac{\Delta T}{T} = \frac{\omega \Delta t}{\varphi} - 1 = \frac{1}{2}\left(\beta - \frac{1}{2}\right)\omega^2\,\Delta t^2 + 0\left(\Delta t^3\right)$$

où $\lambda = \rho\ \exp\left(\pm\, i\varphi\right)$ sont les racines (complexes) de l'équation caractéristique de la matrice A.

{plus difficile}.

Corrigé

1) $$\dot{q}_{n+1} = \dot{q}_n + \int_{t_n}^{t_{n+1}} \ddot{q}\,(\theta)\ d\theta$$

$$q_{n+1} = q_n + \Delta t\ \dot{q}_n + \int_{t_n}^{t_{n+1}} \left(t_{n+1} - \theta\right)\ \ddot{q}\,(\theta)\ d\theta$$

2) On remplace $\ddot{q}\,(\theta)$ par $\frac{1}{2}\left(\ddot{q}_n + \ddot{q}_{n+1}\right)$ dans les relations précédentes. On obtient :

$$\dot{q}_{n+1} = \dot{q}_n + \frac{\Delta t}{2}\left(\ddot{q}_n + \ddot{q}_{n+1}\right)$$

$$q_{n+1} = q_n + \Delta t\ \dot{q}_n + \Delta t^2\left(\frac{1}{4}\,\ddot{q}_n + \frac{1}{4}\,\ddot{q}_{n+1}\right)$$

3) Même méthode mais pour $\ddot{q}\,(\theta) = \ddot{q}_n + \frac{1}{\Delta t}\left(\theta - t_n\right)\left(\ddot{q}_{n+1} + \ddot{q}_n\right)$. On trouve :

$$\dot{q}_{n+1} = \dot{q}_n + \frac{\Delta t}{2}\left(\ddot{q}_n + \ddot{q}_{n+1}\right)$$

$$q_{n+1} = q_n + \Delta t\ \dot{q}_n + \Delta t^2\left(\frac{1}{3}\,\ddot{q}_n + \frac{1}{6}\,\ddot{q}_{n+1}\right)$$

4) Question 2 : $\gamma = \frac{1}{2}$, $\beta = \frac{1}{4}$; question 3 : $\gamma = \frac{1}{2}$, $\beta = \frac{1}{6}$.

5) On suppose que q(t) vérifie l'équation différentielle (D), c'est-à-dire le système différentiel suivant pour U (t) :

$$\frac{dU}{dt} = \begin{bmatrix} -M^{-1}C & -M^{-1}K \\ 1 & 0 \end{bmatrix} U + \begin{pmatrix} M^{-1}\, p(t) \\ 0 \end{pmatrix}$$

On commence donc par mettre le schéma (N) sous une forme analogue, en remplaçant la dérivée $\frac{dU}{dt}$ par le quotient des différences finies.

$$\frac{1}{\Delta t}\left(U_{n+1} - U_n\right) = \begin{pmatrix} (1-\gamma)\ \ddot{q}_n + \gamma\, \ddot{q}_{n+1} \\ \dot{q}_n + \Delta t\ (\frac{1}{2} - \beta)\ \ddot{q}_n + \Delta t\, \beta\, \ddot{q}_{n+1} \end{pmatrix}$$

Puis l'erreur de troncature est ici vectorielle :

$$\tau = \frac{1}{\Delta t}\left(U_{n+1} - U_n\right) - \begin{pmatrix} (1-\gamma)\ \ddot{q}_n + \gamma\, \ddot{q}_{n+1} \\ \dot{q}_n + \Delta t\ (\frac{1}{2} - \beta)\ \ddot{q}_n + \Delta t\, \beta\, \ddot{q}_{n+1} \end{pmatrix}.$$

Compte tenu des formules de Taylor établies à la question 1, il suffit d'évaluer les restes dans les formules de quadrature.

(1) $$\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} \ddot{q}(\theta)\, d\theta \simeq (1-\gamma)\ \ddot{q}_n + \gamma\, \ddot{q}_{n+1}$$

(2) $$\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} (t_{n+1} - \theta)\ \ddot{q}(\theta)\, d\theta \simeq \Delta t\left[(\frac{1}{2} - \beta)\ \ddot{q}_n + \beta\, \ddot{q}_{n+1}\right]$$

pour trouver l'ordre demandé.

* La formule (1) est a priori du premier ordre, sauf si $\gamma = \frac{1}{2}$ et elle est alors du 2ème ordre (formule des trapèzes). L'ordre global du schéma est donc 1 si $\gamma \neq \frac{1}{2}$.

Si $\gamma = \frac{1}{2}$, la formule des trapèzes est d'ordre 2, et un surcroît de précision peut être donné par la formule de quadrature (2). Celle-ci est d'ordre deux si $\ddot{q}(\theta)$ est approchée par n'importe quelle constante compte tenu de l'infiniment petit $(t_{n+1} - \theta)$ à intégrer, et elle est d'ordre trois si $\ddot{q}(\theta)$ est approchée par l'interpolation affine proposée à la question 3, ie pour $\beta = \frac{1}{6}$ (cf question 4).

6) Algorithme

Dans les relations (N), les dérivées secondes sont calculées à l'aide de l'équation (D). Mais si $\beta \neq 0$, la présence de $\ddot{q}_{n+1}$ au second membre de la seconde équation rend le schéma implicite. Il convient donc de calculer d'abord ce terme, en tenant compte du schéma (N) et de l'équation (D) écrite à l'instant t_{n+1}.

$$M \ddot{q}_{n+1} + C \dot{q}_{n+1} + K q_{n+1} = p_{n+1}$$

On introduit dans cette relation $\dot{q}_{n+1}$ et q_{n+1} calculés par le schéma :

$$M \ddot{q}_{n+1} + C \left\{ \dot{q}_n + (1-\gamma) \Delta t \ddot{q}_n + \gamma \Delta t \ddot{q}_{n+1} \right\}$$

$$+ K \left\{ q_n + \Delta t \dot{q}_n + \Delta t^2 (\frac{1}{2} - \beta) \ddot{q}_n + \Delta t^2 \beta \ddot{q}_{n+1} \right\} = p_{n+1}$$

D'où le système à résoudre :

$$S \ddot{q}_{n+1} = f$$

où $S = M + \gamma \Delta t C + K \beta \Delta t^2$

$$f = p_{n+1} - C \left(\dot{q}_n + (1-\gamma) \Delta t \ddot{q}_n \right) - K \left(q_n + \Delta t \dot{q}_n + (\frac{1}{2} - \beta) \Delta t^2 \ddot{q}_n \right)$$

* Algorithme

- $q_o, \dot{q}_o$ connus.

  Calculer $\ddot{q}_o$ par résolution du système (D).

- q_n, $\dot{q}_n$, $\ddot{q}_n$ connus

(i) $\dot{q}^* = \dot{q}_n + (1-\gamma)\,\Delta t\,\ddot{q}_n$

$q^* = q_n + \Delta t\,\dot{q}_n + (\frac{1}{2} - \beta)\,\Delta t^2\,\ddot{q}_n$ (prédiction)

(ii) $S\,\ddot{q}_{n+1} = p_{n+1} - C\dot{q}^* - K q^*$ (résolution)

(iii) $\dot{q}_{n+1} = \dot{q}^* + \Delta t\,\gamma\,\ddot{q}_{n+1}$

$q_{n+1} = q^* + \Delta t^2\,\beta\,\ddot{q}_{n+1}$ (incrémentation)

7) Dans le cas simplifié d'un seul oscillateur non amorti, l'algorithme précédent s'explicite simplement :

$$\begin{cases} \dot{q}_{n+1} = \dot{q}_n + (1-\gamma)\,\Delta t\,(-\omega^2 q_n) + \gamma\,\Delta t\,(-\omega^2 q_{n+1}) \\ q_{n+1} = q_n + \Delta t\,\dot{q}_n + \Delta t^2\,(\frac{1}{2} - \beta)\,(-\omega^2 q_n) + \Delta t^2\,\beta\,(-\omega^2 q_{n+1}) \end{cases}$$

c'est-à-dire :

$$\begin{pmatrix} 1 & \gamma\,\Delta t\,\omega^2 \\ 0 & 1+\beta\,\Delta t^2\,\omega^2 \end{pmatrix} \begin{pmatrix} \dot{q} \\ q \end{pmatrix}_{n+1} = \begin{pmatrix} 1 & -(1-\gamma)\,\Delta t\,\omega^2 \\ \Delta t & 1-(\frac{1}{2}-\beta)\,\Delta t^2\,\omega^2 \end{pmatrix} \begin{pmatrix} \dot{q} \\ q \end{pmatrix}_n$$

D'où le résultat, avec :

$$A = \begin{pmatrix} 1-\gamma\,\xi^2 & -\omega^2\,\Delta t\,(1-\frac{\gamma}{2}\,\xi^2) \\ \dfrac{\xi^2}{\omega\,\Delta t} & 1-\frac{1}{2}\,\xi^2 \end{pmatrix}$$

et $$\xi^2 = \frac{\Delta t^2\,\omega^2}{1+\beta\,\omega^2\,\Delta t^2}$$

8) On a $\det(A-\lambda I) = \lambda^2 - \lambda\left(2-(\gamma+\frac{1}{2})\ \xi^2\right) + 1-(\gamma-\frac{1}{2})\ \xi^2$.

Les racines sont complexes si le discriminant réduit Δ' est négatif. Or :

$$\Delta' = \left(1-\frac{1}{2}\ (\gamma+\frac{1}{2})\ \xi^2\right)^2 - \left(1-(\gamma-\frac{1}{2})\ \xi^2\right)$$

$$= \xi^2\left(\frac{1}{4}\ (\gamma+\frac{1}{2})^2\ \xi^2 - 1\right).$$

$\Delta' \leq 0$ équivaut à :

$$(\gamma+\frac{1}{2})^2\ \omega^2\ \Delta t^2 \leq 4\ \left(1+\beta\ \omega^2\ \Delta t^2\right)$$

c'est-à-dire :

$$(\gamma+\frac{1}{2})^2 - 4\beta \leq \frac{4}{\omega^2\ \Delta t^2} \quad (S)$$

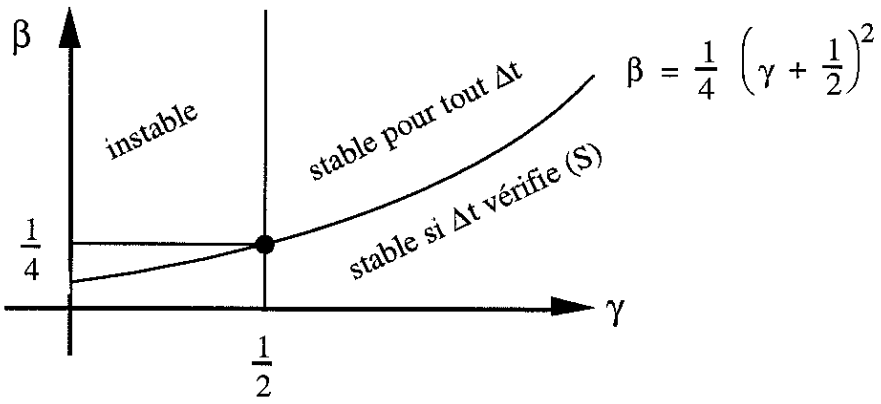
Cette condition limite le pas de temps, sauf si $(\gamma+\frac{1}{2})^2 - 4\ \beta \leq 0$, auquel cas elle ne donne aucune condition. Par ailleurs, le produit des deux racines est égal au carré de leur module commun, c'est-à-dire $1-(\gamma-\frac{1}{2})\ \xi^2$. La condition $|\lambda|^2 \leq 1$ équivaut à $(\gamma-\frac{1}{2})\ \xi^2 \geq 0$ ie $\gamma \geq \frac{1}{2}$.

Les deux racines sont réelles si $\Delta' \geq 0$. La condition :

$$(\gamma+\frac{1}{2})^2 - 4\ \beta > \frac{4}{\omega^2 \Delta t^2}$$

donne une valeur minimale du pas de temps, ce qui indique que pour les valeurs toutes petites de Δt, le schéma est instable. Ce type de comportement est en général rejeté par les utilisateurs, et nous ne considérons pas ce cas dans la discussion qui suit.

- Si $\gamma \leq \frac{1}{2}$, le schéma est instable si $\gamma \geq \frac{1}{2}$ et $(\gamma+\frac{1}{2})^2 < 4\beta$, le schéma est stable lorsque le pas de temps Δt satisfait à la condition (S).



9) Le schéma optimal correspond à la limite de stabilité inconditionnelle, c'est-à-dire $\gamma = \frac{1}{2}$, $\beta = \frac{1}{4}$, qui donne un schéma d'ordre **deux**.

10) Dans le cas de deux valeurs propres complexes pour l'équation caractéristique, on a :

$$\lambda = \rho \exp(i\varphi)$$

avec : $\rho^2 = 1 - (\gamma - \frac{1}{2})\ \xi^2$

$$\text{tg}\ \varphi = \frac{\xi \sqrt{1 - \frac{1}{4}(\gamma + \frac{1}{2})^2 \xi^2}}{1 - \frac{1}{2}(\gamma + \frac{1}{2})\ \xi^2}$$

L'erreur relative de phase et d'amplitude s'obtient en comparant la solution exacte de l'équation $\ddot{q} + \omega^2 q = 0$ à la solution numérique calculée par le schéma, c'est-à-dire λ^n à une constante près. On doit donc comparer $\left(e^{i\omega \frac{\Delta t}{n}}\right)^n$ à λ^n, avec λ donné plus haut.

L'erreur d'amplitude s'obtient en comparant les modules, or :

$$\rho - 1 = \frac{1}{\rho + 1}(\rho^2 - 1) = \frac{1}{2}\left(-(\gamma - \frac{1}{2})\ \xi^2\right) + 0(\xi^2)$$

$$= -\frac{1}{2}(\gamma - \frac{1}{2})\ \Delta t^2 \omega^2 + 0(\Delta t^4)$$

ce qui montre que le choix $\gamma = \frac{1}{2}$ conduit à une erreur d'ordre 4 en Δt. Pour l'erreur de phase, on compare l'écart de périodes entre solution exacte et numérique, ramenée à la période exacte.

$$\frac{\Delta T}{T} = \frac{\omega}{2\Pi}\left(\frac{2\Pi}{\frac{\varphi}{\Delta t}} - \frac{2\Pi}{\omega}\right)$$

ie $$\frac{\Delta T}{T} = \frac{\omega h}{\varphi} - 1$$

or
$$\begin{aligned}
\varphi &= \text{Arctg}\,(\text{tg}\,\varphi) \\
&= (\text{tg}\,\varphi) - \frac{1}{3}\,(\text{tg}\,\varphi)^3 + 0\,(\text{tg}\,\varphi^5) \\
&= \xi\left(1 + \frac{1}{2}\,(\gamma + \frac{1}{2})\,\xi^2\right)\left(1 - \frac{1}{8}(\gamma + \frac{1}{2})\,\xi^2\right) - \frac{1}{3}\,\xi^2 + 0\,(\xi^5) \\
&= \xi\left(1 + \left[\frac{1}{2}(\gamma + \frac{1}{2}) - \frac{1}{8}(\gamma + \frac{1}{2})^2 - \frac{1}{3}\right]\xi^2\right) + 0\,(\xi^5) \\
&= \omega\,\Delta t\left(1 - \frac{\beta}{2}\,\omega^2\,\Delta t^2\right)\left(1 + \frac{1}{24}\,\omega^2\,\Delta t^2\right) + 0\,(\Delta t^4) \quad \text{si } \gamma = \frac{1}{2} \\
&= \omega\,\Delta t\left(1 - \frac{1}{2}\,(\beta - \frac{1}{12})\,\omega^2\,\Delta t^2\right) + 0\,(\Delta t^4) \\
\frac{1}{\varphi} &= \frac{1}{\omega \Delta t}\left(1 + \frac{1}{2}\,(\beta - \frac{1}{12})\,\omega^2\,\Delta t^2\right) + 0\,(\Delta t^3) \\
\frac{\omega\,\Delta t}{\varphi} &= 1 + \frac{1}{2}\,(\beta - \frac{1}{12})\,\omega^2\,\Delta t^2 + 0\,(\Delta t^3)
\end{aligned}$$

ce qui montre le résultat proposé.

■