

# Analyse numérique des équations aux dérivées partielles

Édition 2016

Philippe Saucez

Faculté Polytechnique de Mons

## **Première partie**

### **La méthode des différences finies**

## **TABLE DES MATIERES**

### **PREMIERE PARTIE : LA METHODE DES DIFFERENCES FINIES**

#### **Chapitre I. Généralités**

I.1 Définition et notation	1
I.2 Classification	1
I.3 Equations elliptiques	2
I.4 Equations paraboliques	3
I.5 Equations hyperboliques	3
I.6 Equations dérivées de principes de conservation	4
I.7 Exemples divers	6
I.8 Conditions initiales et conditions aux limites	6

#### **Chapitre II. Concepts de base de la méthode des différences finies**

II.1 Différences finies	9
II.2 Plan général de la méthode des différences finies	14

#### **Chapitre III. Equations elliptiques**

III.1 Généralités	15
III.2 Résolution de l'équation de Poisson dans un rectangle	15
III.3 Discréétisation des conditions aux limites	18
III.4 Application dans le domaine de la thermique	22
III.5 Compléments	24
III.6 Résolution itérative des systèmes linéaires	27
III.7 Méthodes relaxées	30
III.8 Exemples	32
III.9 Méthodes multigrilles	37

#### **Chapitre IV. Equations hyperboliques**

IV.1 Généralités	45
IV.2 Résolution de l'équation d'advection	46
IV.3 Exemple	47

#### **Chapitre V. Equations paraboliques**

V.1 Généralités	50
V.2 Résolution de l'équation de la diffusion	50
V.3 Exemple	51

#### **Chapitre VI. Analyse des performances numériques**

VI.1 Erreur de troncature	53
VI.2 Consistance	53

VI.3 Stabilité : analyse de Von Neumann	54
VI.4 Analyse spectrale des erreurs	58
VI.5 Convergence	60

## **Chapitre VII. Catalogue de quelques méthodes de résolution d'équations hyperboliques**

VII.1 Méthodes d'Euler explicites instables	61
VII.2 Condition de Courant-Friedrichs-Lowy : méthode d'Euler stabilisée	62
VII.3 Méthode de Lax	68
VII.4 Méthode leap frog	70
VII.5 comparaison des méthodes	72

## **Chapitre VIII. Catalogue de quelques méthodes de résolution d'équations paraboliques**

VIII.1 Méthode explicite simple	74
VIII.2 Méthode implicite simple	78
VIII.3 Méthode de Cranck-Nicholson	79

## **Chapitre IX. Analyse de la stabilité par la méthode de l'équation différentielle équivalente**

IX.1 Equation différentielle équivalente – Equation différentielle équivalente modifiée	84
IX.2 Formulation générale de l'équation différentielle équivalente modifiée	91
IX.3 Génération de nouveaux algorithmes avec un ordre de précision donné	100

## **Chapitre X. Extensions de la méthode de Von Neumann pour l'analyse de la stabilité**

X.1 Résolution d'une équation multidimensionnelle linéaire	105
X.2 Résolution d'un système monodimensionnel linéaire	110
X.3 Résolution d'un système multidimensionnel linéaire	112
X.4 Problèmes non linéaires	115
X.5 Globalisation des conditions de Von Neumann par famille de schémas	116

## **Chapitre XI. Méthode matricielle pour l'analyse de la stabilité – Introduction à la méthode des lignes**

XI.1 Stabilité matricielle	119
XI.2 Méthode des lignes – Introduction – Interaction avec la stabilité matricielle	125
XI.3 Exemples	132

## **DEUXIEME PARTIE : LA METHODE DES LIGNES**

### **Chapitre XII. Généralités**

XII.1 Introduction	137
XII.2 Implémentation de la méthode des lignes sous matlab	138

## **Chapitre XIII. Méthode des lignes et différences finies**

XIII.1 Les différences finies en maillage uniforme	146
XIII.2 Les différences finies en maillage non uniforme	147
XIII.3 Influence du choix des schémas de différences finies sur l'apparition d'oscillations parasites	153
XIII.4 Les limiteurs de pente	157
XIII.5 Implémentation des conditions aux limites	172

## **Chapitre XIV. Stabilité et intégration temporelle**

XIV.1 Introduction	180
XIV.2 Exploitation de la condition de stabilité spatiale	181
XIV.3 Problèmes stiff	183
XIV.4 Les intégrateurs à un pas : stabilité et nouvelles méthodes	187
XIV.5 Stabilité temporelle des intégrateurs à pas multiples	194
XIV.6 Nouvelles conditions de stabilité des intégrateurs à pas multiples	196
XIV.7 Nouvelles méthodes à pas multiples	198
XIV.8 Stiffness et méthode des lignes	202
XIV.9 Stabilité dans le cas des problèmes non linéaires	203
XIV.10 Stabilité générale des méthodes à un pas : théorie de la B-stabilité	204
XIV.11 Stabilité générale des méthodes à pas multiple : méthode one-leg et G-stabilité	205
XIV.12 Les intégrateurs de matlab	208

## **Chapitre XV. Problèmes à deux dimensions spatiales. Résolution à l'aide des différences finies**

XV.1 Equation de la chaleur dans un rectangle	219
XV.2 Problème de Graetz avec température de paroi constante	224
XV.3 Equation de la chaleur dans un quadrilatère convexe	229
XV.4 Problème test de convection – diffusion	238
XV.5 Equation de Burgers à deux dimensions	246

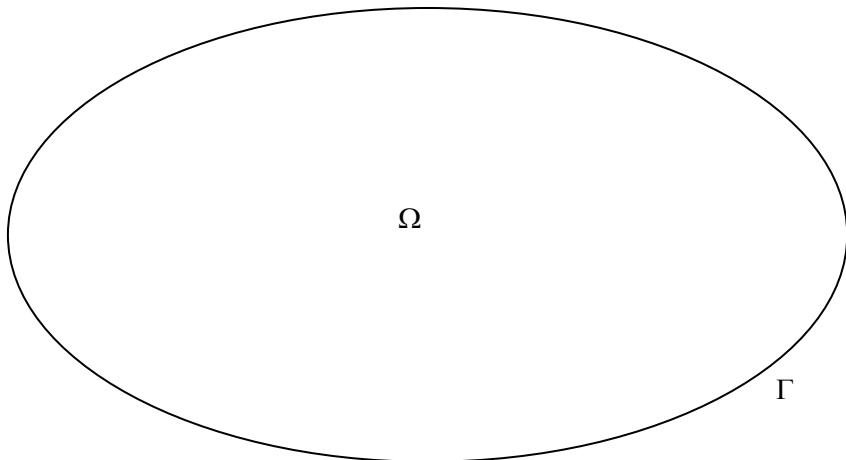
## Chapitre I. Généralités

### I.1 Définition et notation

Les équations aux dérivées partielles (EDP) sont un des modèles mathématiques les plus utilisés par les ingénieurs et les mathématiciens pour décrire des systèmes physiques. Les grandeurs qui caractérisent ces systèmes (par exemple la densité, la vitesse, l'énergie,...) dépendent d'une manière générale de 4 variables indépendantes : les 3 variables spatiales ( $x, y, z$ ) et le temps  $t$ . Cette dépendance est exprimée par le biais de l' EDP qui se présente comme une relation reliant la grandeur physique étudiée  $u$  à ses dérivées de tous ordres par rapport aux variables indépendantes  $x, y, z$  et  $t$  : une telle équation prend donc la forme générale :

$$F(u, u_x, u_y, \dots, x, y, z, t) = 0 \quad I-1$$

Cette équation a un domaine de définition noté conventionnellement  $\Omega$  et sa frontière est  $\Gamma$  :



### I.2 Classification

Quoiqu'il n'existe aucune limitation théorique sur la forme de l'équation I-1, les ingénieurs s'intéressent en général à des formes particulières de cette équation. Pour une très large gamme de problèmes, une classification est possible ; elle est basée sur l'équation suivante appelée équation linéaire du second ordre à deux variables indépendantes :

$$au_{xx} + bu_{xy} + cu_{yy} + du_x + eu_y + fu = g \quad \text{avec } a > 0 \quad I-2$$

La classification dépend du signe du discriminant

$$\Delta = b^2 - 4ac \quad I-3$$

I-2 est dite elliptique si  $\Delta < 0$ , parabolique si  $\Delta = 0$  et hyperbolique si  $\Delta > 0$ .

Les exemples les plus courants de ces trois catégories sont

pour le type elliptique : l'équation de Poisson  $u_{xx} + u_{yy} = g(x, y)$  I-4

pour le type parabolique : l'équation de la chaleur  $u_t = \alpha u_{xx}$  ( $\alpha > 0$ ) I-5

pour le type hyperbolique : l'équation d'onde  $u_{tt} = c^2 u_{xx}$  I-6

pour les types parabolique et hyperbolique la variable  $t$  est utilisée à la place de  $y$  car ces équations décrivent des problèmes dépendant du temps. A l'inverse, les équations elliptiques modélisent des phénomènes d'état stationnaire ou d'équilibre. I-4 à I-6 sont appelées formes canoniques de ces trois catégories.

Les paragraphes qui suivent montrent comment des généralisations sont possibles.

### I.3 Equations elliptiques

D'une manière générale, toute équation elliptique linéaire s'écrit

$$Lu = g \quad I-7$$

où  $L$  est un opérateur elliptique. Si on se limite aux opérateurs du second ordre à coefficients constants, et si  $N$  est le nombre de dimensions spatiales du domaine  $\Omega$  (en principe  $N = 1, 2$  ou  $3$ ), on a

$$L = \sum_{j,k=1}^N R_{jk} \frac{\partial^2}{\partial x_j \partial x_k} + \sum_{j=1}^N P_j \frac{\partial}{\partial x_j} + Q \quad I-8$$

où  $R_{jk}$ ,  $P_j$  et  $Q$  sont des nombres réels. Désignant par  $R$  la matrice  $(R_{jk})$ , I-7 est elliptique si  $R$  est symétrique définie positive (c'est-à-dire si toutes ses valeurs propres sont positives).

Vérifions cette affirmation sur I-2 : la matrice  $R$  de cette équation vaut (dans I-8, la première somme distingue les termes  $\frac{\partial^2}{\partial x \partial y}$  et  $\frac{\partial^2}{\partial y \partial x}$ )

$$R = \begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix} \quad I-9$$

Ses valeurs propres sont solutions de

$$\lambda^2 - (a+c)\lambda + ac - \frac{b^2}{4} = 0 \quad I-10$$

$$\Rightarrow \lambda = \frac{(a+c) \pm \sqrt{(a+c)^2 + b^2 - 4ac}}{2} = \frac{(a+c) \pm \sqrt{(a-c)^2 + b^2}}{2} \quad I-11$$

Ces deux racines sont réelles ; établissons la condition qui les rend positives :

a) $(a+c) - \sqrt{(a-c)^2 + b^2} > 0$	si $\sqrt{(a-c)^2 + b^2} < (a+c)$	
c.à.d. si $(a-c)^2 + b^2 < (a+c)^2$		
c.à.d. si $b^2 - 4ac < 0$		I-12

b) la deuxième racine est alors forcément positive car elle est égale à la précédente augmentée de  $\sqrt{(a - c)^2 + b^2}$

On retrouve donc bien la condition  $b^2 - 4ac < 0$ .

#### I.4 Equations paraboliques

Si  $L$  est un opérateur elliptique, l'équation

$$u_t = Lu - g \quad I-13$$

est parabolique. Dans le cas où  $L$  est le laplacien  $\nabla^2 = \sum_{k=1}^N \frac{\partial^2}{\partial x_k^2}$ , I-13 est la généralisation de I-5.

#### I.5 Equations hyperboliques

On a coutume de remplacer l'équation I-6 par l'équation dite d'advection

$$u_t + au_x = 0 \quad I-14$$

comme équation canonique de cette catégorie d'EDP. Strictement parlant, I-14 ne répond pas à la classification proposée par I-2 et I-3. La justification de cet écart provient de l'habitude qu'on a de traiter I-6 en la transformant en un système d'équations du premier ordre pour la dérivation par rapport au temps : il est coutumier de poser dans I-6

$$v = u_t \quad \text{et} \quad w = cu_x \quad I-15$$

Il vient alors aisément

$$\begin{pmatrix} v_t \\ w_t \end{pmatrix} = \begin{pmatrix} cw_x \\ cv_x \end{pmatrix} = \begin{pmatrix} 0 & c \\ c & 0 \end{pmatrix} \begin{pmatrix} v \\ w \end{pmatrix} \quad I-16$$

ou encore

$$\bar{u}_t + A\bar{u}_x = \bar{0} \quad I-17$$

avec

$$\bar{u} = \begin{bmatrix} v \\ w \end{bmatrix} \quad \text{et} \quad A = \begin{bmatrix} 0 & -c \\ -c & 0 \end{bmatrix} \quad I-18$$

Observons que les valeurs propres de  $A$  sont réelles :  $\lambda_1 = +c$  et  $\lambda_2 = -c$ . Ceci est à la base de la définition générale d'un problème hyperbolique : I-17 est appelé système hyperbolique du premier ordre si la matrice  $A$ , supposée de dimension  $N$ , possède des valeurs propres réelles et si  $A$  est diagonalisable, c'est-à-dire si  $A$  possède  $N$  vecteurs propres linéairement indépendants.  
Observons pour terminer que I-14, qui servira par la suite, est la forme scalaire de I-17.

## I.6 Equations dérivées de principes de conservation

Le principe de la conservation d'une grandeur est à la base de nombreuses EDP ou systèmes d'EDP. Imaginons par exemple que la grandeur physique  $u$  étudiée ne dépend que d'une variable spatiale  $x$  et du temps  $t$ , et que  $u(x, t)$  représente, pour fixer les idées, la densité d'un polluant dans une canalisation ;  $x$  représente la position le long de celle-ci. Il est alors intuitif d'obtenir la masse totale  $M(t)$  de polluant comprise à l'instant  $t$  entre deux sections d'abscisses  $x_1$  et  $x_2$  en intégrant  $u(x, t)$  entre ces deux sections :

$$M(t) = \int_{x_1}^{x_2} u(x, t) dx \quad I-19$$

Si en outre, il n'y a ni création ni destruction de polluant entre ces deux sections, la seule modification éventuelle de cette masse totale ne peut provenir que du flux de polluant traversant les sections d'abscisses  $x_1$  et  $x_2$ . Ce flux, en toute généralité, peut dépendre de la densité  $u(x, t)$  elle-même, et on supposera par simplicité qu'il ne dépend que d'elle.

### Cas de l'advection

Si le polluant est simplement transporté dans un flux à vitesse constante  $a$ , ce flux est alors égal à la densité du polluant multiplié par sa vitesse de déplacement :

$$f(u) = au \quad I-20$$

Si on convient d'affecter du signe + le flux se déplaçant de la gauche vers la droite, la variation de masse de polluant à l'instant  $t$  vaut alors

$$\frac{dM(t)}{dt} = \frac{d}{dt} \int_{x_1}^{x_2} u(x, t) dx = f(u(x_1, t)) - f(u(x_2, t)) \quad I-21$$

qui peut aussi s'écrire, pour autant que  $f$  et  $u$  soient continus

$$\int_{x_1}^{x_2} \frac{\partial}{\partial t} u(x, t) dx = - \int_{x_1}^{x_2} \frac{\partial}{\partial x} f(u(x, t)) dx \quad I-22$$

ou encore

$$\int_{x_1}^{x_2} \left[ \frac{\partial}{\partial t} u(x, t) + \frac{\partial}{\partial x} f(u(x, t)) \right] dx = 0 \quad I-23$$

Comme cette intégrale doit être nulle pour toutes les valeurs de  $x_1$  et  $x_2$ , il résulte que l'intégrant doit être identiquement nul :

$$\frac{\partial}{\partial t} u(x, t) + \frac{\partial}{\partial x} f(u(x, t)) = 0 \quad I-24$$

I-24 est une équation de conservation. Si on remplace maintenant  $f(u)$  par I-20, on retrouve l'équation d'advection I-14.

### **Cas de la diffusion**

Imaginons maintenant que le fluide dans la canalisation ne s'écoule plus :  $f(u) = 0$  car  $a = 0$ . Il en découle  $u_t = 0$  : le profil de la densité du polluant reste constant dans la canalisation ; ceci n'est pas en accord avec ce à quoi on s'attend, si cette densité n'est pas uniforme tout au long de la canalisation : de la même manière qu'une goutte d'encre se dilue dans l'eau, des molécules de polluant vont migrer par diffusion des régions de haute densité vers les régions de basse densité. L'étude précise de ce phénomène indique que le flux de diffusion qui en résulte à l'abscisse  $x$  ne dépend que de la dérivée spatiale de  $u$  à cet endroit, et plus précisément qu'il lui est proportionnel :

$$f(u_x) = -\kappa u_x \quad \text{I-25}$$

$\kappa$  est appelé coefficient de diffusion ; c'est un nombre positif, ce qui explique le signe de la relation I-25 : on a convenu que le flux est positif lorsqu'il va de la gauche vers la droite. Comme le flux de diffusion va des régions de haute densité vers celles de faible densité, I-25 est positif si la pente  $u_x$  est négative. I-25 est appelée loi de Fick ; combinée avec I-24, elle donne la loi de la diffusion (appelée aussi équation de la chaleur) :

$$u_t = \kappa u_{xx} \quad \text{I-26}$$

Et lorsque  $\kappa$  dépend de  $x$ , I-26 devient

$$u_t = (\kappa(x)u_x)_x \quad \text{I-27}$$

Notons que la diffusion et le déplacement de polluant par advection peuvent coexister : il en résulte alors l'équation d'advection-diffusion

$$u_t + au_x = \kappa u_{xx} \quad \text{I-28}$$

### **Présence d'un terme de source**

Une possibilité supplémentaire de variation de la masse totale de polluant peut être la présence d'une source de cette substance entre les sections d'abscisses  $x_1$  et  $x_2$  ; si  $\psi(x, t)$  représente la densité de cette source, la relation I-28 devient

$$u_t + au_x = \kappa u_{xx} + \psi \quad \text{I-29}$$

### **Equation de réaction-diffusion**

Les réacteurs chimiques présentent en général un ou des termes de source ; et si  $\bar{u}(x, t)$  supposé de dimension  $S$  représente les concentrations des  $S$  composants présents dans le réacteur,  $\bar{\psi}$  dépend en général de  $\bar{u}$ . A priori, ce type de système n'est le siège d'aucun mouvement d'ensemble des matières incluses dans le réacteur, et dans ce cas I-29 est réduit à un système d'équations de réaction-diffusion

$$\bar{u}_t = \kappa \bar{u}_{xx} + \bar{\psi}(\bar{u}) \quad \text{I-30}$$

mais si la réaction a lieu dans un fluide qui se déplace, un terme d'advection s'ajoute pour donner

$$\bar{u}_t + \bar{f}(\bar{u})_x = \kappa \bar{u}_{xx} + \bar{\psi}(\bar{u}) \quad \text{I-31}$$

## I.7 Exemples divers

Les équations ci-dessous, qui n'entrent pas forcément dans la classification qui précède, ont toutes un point commun : elles ont été – et souvent sont toujours – un banc d'essai idéal de tests pour les méthodes numériques de résolution d'EDP :

### *Equation de Burgers*

$$u_t + uu_x = \varepsilon u_{xx} \quad \text{I-32}$$

Cette équation est du même type que I-28 si on y remplace le terme d'advection par le flux

$$f(u) = \frac{u^2}{2} \quad \text{I-33}$$

C'est un modèle scalaire simple des effets observés dans les écoulements de fluides compressibles. Ce modèle a été largement étudié dans la littérature.

### *Equation de Korteweg-de Vries (KdV)*

$$u_t + uu_x = vu_{xxx} \quad \text{I-34}$$

Très semblable à I-32, elle modélise cependant des phénomènes très différents dont le plus courant est l'existence de solutions de type solitons.

### *Equation de Tricomi*

$$yu_{xx} + u_{yy} = 0 \quad \text{I-35}$$

Typique des écoulements transsoniques non visqueux, cette équation inclut un passage du type elliptique au type hyperbolique selon le signe de  $y$ .

### *Equation de Helmholtz*

$$u_{xx} + u_{yy} + k^2 u = 0 \quad \text{I-36}$$

Cette équation est utilisée notamment dans la description de la propagation d'ondes acoustiques.

### *Equation biharmonique*

$$u_{xxxx} + u_{yyyy} = 0 \quad \text{I-37}$$

Cette équation est une relation de base de la théorie de l'élasticité.

## I.8 Conditions initiales et conditions aux limites

La résolution d'une équation aux dérivées partielles exige la connaissance a priori de conditions supplémentaires dites « conditions initiales » et « conditions aux limites ».

Les conditions initiales sont propres aux problèmes où la variable  $u$  fait l'objet d'une dérivation par rapport au temps : c'est le cas de I-5, I-6, I-13, I-14, ... Ce ne sera donc pas le cas des équations

elliptiques. Le nombre de conditions initiales nécessaires dépend de l'ordre de la plus haute dérivation en temps : si  $m$  est l'ordre de la dérivée temporelle la plus élevée, le nombre de conditions initiales sera égal à  $m$ .

La forme de ces conditions sera la suivante, si on suppose que l'instant à partir duquel la résolution est demandée est  $t_0$  (en général on aura  $t_0 = 0$ ) :

- $u(\bar{x}_\Omega, t_0)$  est connu en tous les points  $\bar{x}_\Omega$  du domaine spatial
- $u_t(\bar{x}_\Omega, t_0)$  est connu en tous les points  $\bar{x}_\Omega$  du domaine spatial
- ...
- $u_{kt}(\bar{x}_\Omega, t_0)$  est connu en tous les points  $\bar{x}_\Omega$  du domaine spatial

où l'ordre  $k$  de la plus haute dérivée vaut  $m - 1$ . Ainsi, par exemple pour I-6,  $u(\bar{x}_\Omega, t_0)$  et  $u_t(\bar{x}_\Omega, t_0)$  sont donnés :

$$\begin{aligned} u(\bar{x}_\Omega, t_0) &= f_1^0(\bar{x}_\Omega) \forall \bar{x}_\Omega \\ u_t(\bar{x}_\Omega, t_0) &= f_2^0(\bar{x}_\Omega) \forall \bar{x}_\Omega \end{aligned}$$

Les conditions aux limites appellent de plus amples commentaires. Signalons d'abord qu'elles ne sont pas toujours forcément présentes : prenons par exemple le cas de la résolution de I-5, forme canonique des équations paraboliques :

$$u_t = \alpha u_{xx} \quad \text{I-5}$$

Le domaine d'étude de cette équation peut prendre trois formes :

- le demi-plan infini  $\{(x, t) : [-\infty < x < +\infty] \times [0 \leq t]\}$  : dans ce cas I-5 doit seulement être complété par la connaissance de la condition initiale  $u(x, 0)$

- le quart de plan infini  $\{(x, t) : [x_m \leq x < +\infty] \times [0 \leq t]\}$  : I-5 doit être complété par la condition initiale et une condition à la limite  $x = x_m$  du domaine spatial : par exemple connaître  $u(x_m, t)$ .

- le rectangle ouvert  $\{(x, t) : [x_m \leq x < x_M] \times [0 \leq t]\}$  : I-5 doit être complété par la condition initiale et deux conditions aux limites  $x = x_m$  et  $x = x_M$  du domaine spatial : par exemple connaître  $u(x_m, t)$  et  $\frac{\partial u}{\partial x}(x_M, t)$ .

Présentes dans les deux derniers cas, les conditions aux limites ne peuvent être quelconques : elles doivent respecter une règle analogue à celle des conditions initiales : fournir des informations sur  $u$  et/ou d'éventuelles dérivées spatiales  $u_{kx}$  avec  $k$  au maximum égal  $n - 1$  si  $n$  est l'ordre de la plus haute dérivée spatiale présente dans l'EDP à résoudre. Ainsi, pour I-5,  $n = 2$  : les conditions aux limites porteront donc sur  $u$  et/ou sa dérivée spatiale première.

Lorsque l'EDP à résoudre présente des dérivées spatiales au maximum d'ordre deux (ce qui est le cas des trois grandes catégories d'EDP étudiées), les conditions aux limites (CL) peuvent donc être de trois types :

CL de type Dirichlet :  $u$  est donné (par exemple en  $x = x_M$ )

$$u(x_M, t) = g_D(t) \quad I-38$$

CL de type Neumann :  $\frac{\partial u}{\partial x}$  est donné (par exemple en  $x = x_m$ )

$$\frac{\partial u}{\partial x}(x_m, t) = g_N(t) \quad I-39$$

CL de type Robin :  $k_1 \frac{\partial u}{\partial x} + k_0 u$  est donné (par exemple en  $x = x_m$ )

$$k_1 \frac{\partial u}{\partial x}(x_m, t) + k_0 u(x_m, t) = g_R(t) \quad I-40$$

Ces conditions se généralisent sans peine lorsque le domaine spatial est de dimensions deux ou trois : si  $\bar{x}_\Gamma$  est l'ensemble des points frontières (à distance finie) du domaine spatial d'étude, les conditions précédentes deviennent

$$CL de type Dirichlet : \quad u(\bar{x}_\Gamma, t) = g_{\Gamma,D}(t) \quad I-41$$

$$CL de type Neumann : \quad \frac{\partial u}{\partial n}(\bar{x}_\Gamma, t) = g_{\Gamma,N}(t) \quad I-42$$

$$CL de type Robin : \quad k_1 \frac{\partial u}{\partial n}(\bar{x}_\Gamma, t) + k_0 u(\bar{x}_\Gamma, t) = g_{\Gamma,R}(t) \quad I-43$$

On notera que dans les deux derniers types de CL,  $n$  est le vecteur normal à  $\Gamma$  dirigé vers l'extérieur du domaine spatial.

Signalons encore que dans un même problème des CL de types différents peuvent coexister sur des portions différentes de  $\Gamma$ .

## **Chapitre II. Concepts de base de la méthode des différences finies**

### **II.1 Différences finies**

La méthode des différences finies est l'outil le plus populaire d'évaluation numérique des dérivées d'une fonction d'une ou plusieurs variables. Considérons d'abord le cas simple d'une fonction d'une seule variable  $u : x \rightarrow u(x)$  définie sur  $[x_0, x_N]$  et connue seulement en un ensemble fini de valeurs de  $x$  :

$x$	$u$
$x_0$	$u_0$
$\vdots$	$\vdots$
$x_{i-1}$	$u_{i-1}$
$x_i$	$u_i$
$x_{i+1}$	$u_{i+1}$
$\vdots$	$\vdots$
$x_N$	$u_N$

L'estimation numérique d'une dérivée d'ordre  $n$  quelconque de  $u$  en une abscisse  $x_i$  par la méthode des différences finies, soit  $\frac{d^n u(x_i)}{dx^n}$ , est basée sur les développements en série de Taylor de  $u(x)$  en des points  $x_k$  voisins de  $x_i$ .

Exemples simples : calculs de la dérivée 1<sup>ère</sup> et de la dérivée 2<sup>de</sup> de  $u$  en  $x_i$  :

Ecrivons les développements de Taylor de  $u(x)$  en  $x_{i+1}$  et  $x_{i-1}$ , autour de  $x_i$  :

$$u_{i+1} = u_i + \frac{x_{i+1} - x_i}{1!} u^{(I)}(x_i) + \frac{(x_{i+1} - x_i)^2}{2!} u^{(II)}(x_i) + \frac{(x_{i+1} - x_i)^3}{3!} u^{(III)}(x_i) + \frac{(x_{i+1} - x_i)^4}{4!} u^{(IV)}(x_i) + \dots \quad \text{II-1}$$

$$u_{i-1} = u_i - \frac{x_i - x_{i-1}}{1!} u^{(I)}(x_i) + \frac{(x_i - x_{i-1})^2}{2!} u^{(II)}(x_i) - \frac{(x_i - x_{i-1})^3}{3!} u^{(III)}(x_i) + \frac{(x_i - x_{i-1})^4}{4!} u^{(IV)}(x_i) + \dots \quad \text{II-2}$$

si on convient d'appeler  $u_i^{(k)}$  la valeur (inconnue) de la dérivée  $k^{\text{ème}}$  de  $u$  en  $x_i$ .

Supposons en outre pour simplifier que les abscisses  $x_k$  soient uniformément réparties sur  $[x_0, x_N]$  :

$$\left. \begin{array}{l} x_k = x_0 + kh \\ h = c^{\text{te}} \end{array} \right\} \quad \text{II-3}$$

II-1 à II-3 donnent

$$u_{i+1} = u_i + \frac{h}{1!} u_i^{(I)} + \frac{h^2}{2!} u_i^{(II)} + \frac{h^3}{3!} u_i^{(III)} + \frac{h^4}{4!} u_i^{(IV)} + \dots \quad \text{II-4}$$

$$u_{i-1} = u_i - \frac{h}{1!} u_i^{(I)} + \frac{h^2}{2!} u_i^{(II)} - \frac{h^3}{3!} u_i^{(III)} + \frac{h^4}{4!} u_i^{(IV)} - \dots , \quad \text{II-5}$$

### *Estimations de la dérivée première :*

Une estimation courante de cette dérivée est obtenue de la manière suivante : soustrayons II-5 de II-4 :

$$u_{i+1} - u_{i-1} = 2 \frac{h}{1!} u_i^{(I)} + 2 \frac{h^3}{3!} u_i^{(III)} + \dots \quad \text{II-6}$$

c'est-à-dire

$$u_i^{(I)} = \frac{u_{i+1} - u_{i-1}}{2h} - \frac{h^2}{3!} u_i^{(III)} \dots \quad \text{II-7}$$

et quand  $h$  devient très petit :

$$u_i^{(I)} \cong \frac{u_{i+1} - u_{i-1}}{2h} + O(h^2) \quad \text{II-8}$$

Autres estimations courantes de la dérivée première : en isolant  $u_i^{(I)}$  dans II-4 et II-5 respectivement, on trouve deux autres estimations de cette dérivée :

$$u_i^{(I)} = \frac{u_{i+1} - u_i}{h} - \frac{h}{2!} u_i^{(II)} + \dots \cong \frac{u_{i+1} - u_i}{h} + O(h) \quad \text{II-9}$$

et

$$u_i^{(I)} = \frac{u_i - u_{i-1}}{h} + \frac{h}{2!} u_i^{(II)} + \dots \cong \frac{u_i - u_{i-1}}{h} + O(h) \quad \text{II-10}$$

### *Estimation de la dérivée seconde :*

additionnons II-4 et II-5 :

$$u_{i+1} + u_{i-1} = 2u_i + 2 \frac{h^2}{2!} u_i^{(II)} + 2 \frac{h^4}{4!} u_i^{(IV)} + \dots$$

c'est-à-dire :

$$u_i^{(II)} = \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} - \frac{h^2}{12} u_i^{(IV)} - \dots \quad \text{II-11}$$

et quand  $h$  devient très petit :

$$u_i^{(II)} \cong \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} + O(h^2) \quad \text{II-12}$$

II-8 et II-12 sont deux formules de différences finies permettant d'estimer les dérivées 1<sup>ère</sup> et 2<sup>de</sup> de  $u$  au point  $i$  à partir de la connaissance de  $u$  en les points  $i-1$ ,  $i$  et  $i+1$ .

### *Commentaires :*

- Ces formules sont des approximations d'autant meilleures que  $h$  est petit : l'erreur pour ces deux formules est proportionnelle à  $h^2$  et tend donc vers zéro avec  $h$ .
- Ces formules sont dites centrées dans la mesure où le « pattern » de points sélectionnés ( $i-1$ ,  $i$ ,  $i+1$ ) est centré sur le point en lequel on veut estimer les dérivées. Des formules non centrées peuvent être déduites en sélectionnant d'autres « pattern ». C'est le cas de II-9 et II-10 ; c'est

aussi ce qu'on obtiendrait en calculant  $u_i^{(I)}$  et  $u_i^{(II)}$  à partir des développements de Taylor de  $u$ , calculés en  $i+1$  et  $i+2$  :  $u_i^{(I)}$  et  $u_i^{(II)}$  seraient alors des combinaisons linéaires des valeurs  $u_i, u_{i+1}, u_{i+2}$ .

Par ailleurs, en procédant au calcul de ces nouvelles formules, on s'aperçoit que les termes d'erreur obtenus sont proportionnels respectivement à  $h^2$  et à  $h$ . Ceci signifie que les formules non centrées sont moins précises que les formules centrées, car  $h^2$  tend vers zéro plus vite que  $h$ , quand  $h$  tend vers zéro.

On utilisera donc en général des formules centrées, en réservant les autres à des applications particulières.

En voici un exemple : si on veut calculer  $u_0^{(I)}$ , à partir du tableau initial, il est impossible d'utiliser une formule centrée car il faudrait alors disposer de  $u_1$  et de  $u_{-1}$  qui n'existe pas :

$$u_0^{(I)} = \frac{u_1 - u_{-1}}{2h}$$

- Les formules fournies sont les plus simples : on peut à volonté utiliser un « pattern » de points plus étendu : par exemple, calculer  $u_i^{(II)}$  à partir des points  $u_{i-2}, u_{i-1}, u_i, u_{i+1}, u_{i+2}$ . Il suffit pour cela d'écrire les développements de Taylor de  $u$  en les abscisses  $i-2, i-1, i+1, i+2$  et de combiner judicieusement les relations obtenues pour en déduire  $u_i^{(II)}$ . A cet égard, l'écriture des développements de Taylor peut être interprétée comme la création d'un système d'équations linéaires en les inconnues  $u_i^{(I)}, u_i^{(II)}, \dots$ . Ainsi, les équations II-4 et II-5 peuvent être considérées comme un système de deux équations linéaires en les inconnues  $u_i^{(I)}$  et  $u_i^{(II)}$  :

$$\begin{cases} \frac{h}{1!} u_i^{(I)} + \frac{h^2}{2!} u_i^{(II)} = u_{i+1} - u_i - \frac{h^3}{3!} u_i^{(III)} - \frac{h^4}{4!} u_i^{(IV)} \dots \\ -\frac{h}{1!} u_i^{(I)} + \frac{h^2}{2!} u_i^{(II)} = u_{i-1} - u_i + \frac{h^3}{3!} u_i^{(III)} - \frac{h^4}{4!} u_i^{(IV)} \dots \end{cases} \quad \text{II-13}$$

Cette interprétation a un double avantage :

1. Le nombre de développements écrits fixe le nombre de dérivées calculables : le système précédent compte 2 équations, on ne peut donc estimer que  $u_i^{(I)}$  et  $u_i^{(II)}$  ; si on écrit les développements de Taylor en  $i-2, i-1, i+1$  et  $i+2$ , on pourra en tirer des estimations de 4 inconnues :  $u_i^{(I)}, u_i^{(II)}, u_i^{(III)}, u_i^{(IV)}$ .
2. Il est aisément de déterminer la puissance de  $h$  de l'erreur qui affectera la formule de différence finie obtenue :

pour s'en apercevoir, reprenons le système II-13 réécrit selon

$$\begin{aligned} \begin{bmatrix} +\frac{h}{1!} & \frac{h^2}{2!} \\ -\frac{h}{1!} & \frac{h^2}{2!} \end{bmatrix} \begin{bmatrix} u_i^{(I)} \\ u_i^{(II)} \end{bmatrix} &= \begin{bmatrix} u_{i+1} - u_i \\ u_{i-1} - u_i \end{bmatrix} + h^3 \begin{bmatrix} -\frac{u_i^{(III)}}{3!} - \frac{h}{4!} u_i^{(IV)} \dots \\ +\frac{u_i^{(III)}}{3!} - \frac{h}{4!} u_i^{(IV)} \dots \end{bmatrix} \\ &= \begin{bmatrix} \text{terme en } \Delta u \\ \text{terme en } \Delta u \end{bmatrix} + h^3 \begin{bmatrix} \text{erreur} \\ \text{erreur} \end{bmatrix} \end{aligned}$$

La résolution par la méthode de Cramer de ce système donne une solution dont la structure est

$$u_i^{(I)} = \frac{\Delta u_i^{(I)}}{\Delta} \quad \text{et} \quad u_i^{(II)} = \frac{\Delta u_i^{(II)}}{\Delta}$$

avec  $\Delta \div h^3$

$$\begin{aligned}\Delta u_i^{(I)} &\div h^2 [\text{terme en } \Delta u] + h^5 [\text{erreur}] \\ \Delta u_i^{(II)} &\div h [\text{terme en } \Delta u] + h^4 [\text{erreur}]\end{aligned}$$

c'est-à-dire :

$$u_i^{(I)} \div \frac{\text{terme en } \Delta u}{h} + h^2 * (\text{erreur}) \quad \text{II-14}$$

$$u_i^{(II)} \div \frac{\text{terme en } \Delta u}{h^2} + h * (\text{erreur}) \quad \text{II-15}$$

La comparaison de II-14 et de II-8 montre clairement que l'interprétation qui précède est correcte. C'est un peu moins évident avec II-15 et II-12 : on trouve en pratique un terme d'erreur proportionnel à  $h^2$  et non à  $h$  : ceci provient de la structure centrée du schéma : il y a, lors de la résolution du système, « disparition » heureuse du terme d'erreur en  $h$  ; si on avait pris un pattern de points non centrés, cette disparition n'aurait pas eu lieu et on aurait trouvé une formule de type II-15. Il en résulte que pour un nombre de développements de Taylor fixé à priori, les estimations des dérivées seront de moins en moins précises au fur et à mesure que l'ordre de la dérivée augmente : Ainsi, pour des développements de Taylor écrits en  $i+1, i+2, i+3$  et  $i+4$ , autour de  $i$ , on aura :

Pour  $u_i^{(I)}$  : erreur  $\div h^4$

$u_i^{(II)}$  : erreur  $\div h^3$

$u_i^{(III)}$  : erreur  $\div h^2$

$u_i^{(IV)}$  : erreur  $\div h$

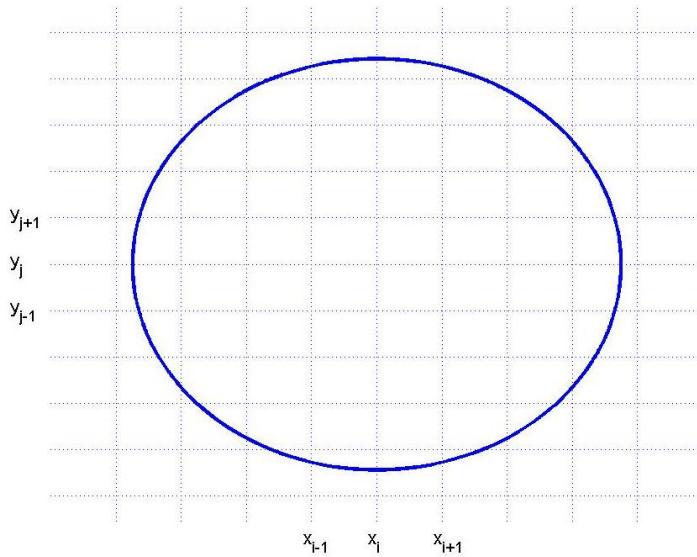
Ceci a pour conséquence que, si on doit estimer simultanément deux dérivées d'ordres différents avec la même précision, le nombre de développements de Taylor à prendre en compte pour ces 2 dérivées doit être différent, ou, ce qui revient au même, les pattern de points à prendre en compte pour le calcul de ces dérivées devront comporter des nombres de points différents.

### **Extrapolation aux fonctions de plusieurs variables :**

Prenons l'exemple d'une fonction de 2 variables  $u : (x, y) \rightarrow u(xy)$  et imaginons devoir estimer

$\frac{\partial^2 u}{\partial x \partial y}$  : les étapes à suivre sont une généralisation de ce qu'on fait dans le cas d'une fonction d'une variable.

- a) Il faut définir un tableau de valeurs discrètes de  $u$  en des points  $(x_i, y_j)$ . Ceci se fait en général en superposant au domaine de définition de  $u$  un « maillage » ou « grille ».



b) Les dérivées sont alors traitées en appliquant les formules de différences finies comme si la fonction ne dépendait que d'une seule variable :

$$\begin{aligned}
 \left( \frac{\partial^2 u}{\partial x \partial y} \right)_{(x_i, y_j)} &= \frac{\partial}{\partial x} \left[ \frac{\partial u}{\partial y} \right]_{x_i, y_j} = \frac{\partial}{\partial x} \left[ \frac{u(x_i, y_{j+1}) - u(x_i, y_{j-1})}{2\Delta y} \right]_{x_i} \\
 &= \frac{1}{2\Delta y} \left[ \frac{\partial}{\partial x} (u(x_i, y_{j+1})) - \frac{\partial}{\partial x} (u(x_i, y_{j-1})) \right]_{x_i} \\
 &= \frac{1}{2\Delta y} \left[ \frac{u(x_{i+1}, y_{j+1}) - u(x_{i-1}, y_{j+1})}{2\Delta x} - \frac{u(x_{i+1}, y_{j-1}) - u(x_{i-1}, y_{j-1})}{2\Delta x} \right]
 \end{aligned} \tag{II-16}$$

et si on convient de noter, pour alléger l'écriture,

$$u(x_i, y_j) = u_{ij}, \tag{II-17}$$

$$\left( \frac{\partial^2 u}{\partial x \partial y} \right)_{x_i, y_j} = \frac{1}{4\Delta x \Delta y} (u_{i+1,j+1} - u_{i-1,j+1} - u_{i+1,j-1} + u_{i-1,j-1}) \tag{II-18}$$

Semblablement, on montre par exemple aisément que

$$\left( \frac{\partial^2 u}{\partial x^2} \right)_{ij} = \frac{1}{\Delta x^2} [u_{i+1,j} - 2u_{ij} + u_{i-1,j}] \tag{II-19}$$

$$\left( \frac{\partial^2 u}{\partial y^2} \right)_{ij} = \frac{1}{\Delta y^2} [u_{ij+1} - 2u_{ij} + u_{ij-1}] \tag{II-20}$$

et, dans le cas où  $\Delta x = \Delta y = h$ ,

$$\left( \frac{\partial^2 u}{\partial y^2} \right)_{ij} + \left( \frac{\partial^2 u}{\partial x^2} \right)_{ij} = \frac{1}{h^2} [u_{ij+1} + u_{i+1,j} - 4u_{ij} + u_{ij-1} + u_{i-1,j}] \tag{II-21}$$

Les termes d'erreurs se calculent de la même manière que pour les fonctions d'une variable.

## II.2 Plan général de la méthode des différences finies

Ainsi qu'on l'a mentionné au chapitre I, la résolution d'une équation aux dérivées partielles inclut en général la prise en compte de conditions initiales et aux limites. Ces conditions vont donc intervenir également dans la résolution par la méthode des différences finies. Convenons donc de compléter l'EDP initiale I-1

$$F(u, u_x, u_y, \dots, x, y, z, t) = 0 \quad \text{I-1}$$

par la relation suivante symbolisant globalement ces conditions :

$$G(u, u_x, u_y, \dots, x, y, z, t) = 0 \quad \text{II-22}$$

La résolution de I-1 complété par II-22 par la méthode des différences finies comprend les étapes générales suivantes :

1. Discrétisation du domaine de définition  $\Omega(x, y, z, t)$  de l'équation et de sa frontière  $\Gamma$ . Cette discrétisation consiste à se donner un ensemble de valeurs discrètes  $(x_i, y_j, z_l, t^k)$  des variables indépendantes.
2. Ecriture de I-1 et de II-22 en chacun des nœuds du maillage défini par la discrétisation précédente : on obtient ainsi un système d'équations en les inconnues  $u_{ijl}^k$  et en les estimations des dérivées de ces inconnues en les points  $(x_i, y_j, z_l, t^k)$ .
3. Remplacement des dérivées par les estimations qu'en donnent les formules de différences finies. Cette opération transforme le système précédent en un système en les seules inconnues  $u_{ijl}^k$ .
4. La résolution de ce système fournit la solution de l'équation sous la forme d'un tableau des valeurs de  $u$  en les nœuds  $(x_i, y_j, z_l, t^k)$ .

$$\begin{cases} F(u, u_x, u_y, \dots, x, y, z, t)_{ijlk} = 0 & \forall i, j, l, k : (x_i, y_j, z_l, t^k) \in \Omega \\ G(u, u_x, u_y, \dots, x, y, z, t)_{ijlk} = 0 & \forall i, j, l, k : (x_i, y_j, z_l, t^k) \in \Gamma \end{cases} \quad \text{II-23}$$

## Chapitre III. Equations elliptiques

### **III.1 Généralités**

L'archétype de cette catégorie d'EDP est l'équation de Poisson

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \equiv \nabla^2 u = f(x, y) \quad \text{III-1}$$

définie dans un domaine  $\Omega(x, y)$ . En toute généralité, les conditions aux limites sont du type Dirichlet, Neumann ou Robin, appliquées sur l'entièreté ou une partie du contour  $\Gamma$  de  $\Omega$ . L'absence de dépendance de  $u$  par rapport au temps exclut la présence de conditions initiales.

Rappelons que III-1 n'est qu'un cas particulier de l'équation générale du second ordre

$$a(x, y)u_{xx} + b(x, y)u_{xy} + c(x, y)u_{yy} + d(x, y)u_x + e(x, y)u_y + f(x, y)u = g(x, y) \quad \text{III-2}$$

avec  $a > 0$  et  $b^2 - 4ac < 0$

De même, il existe des équations elliptiques d'ordre supérieur, telle l'équation biharmonique :

$$\nabla^4 u = \nabla^2(\nabla^2 u) = u_{xxxx} + 2u_{xxyy} + u_{yyyy} = f(x, y) \quad \text{III-3}$$

Consacrés à la résolution d'exemples, les paragraphes qui suivent vont mettre en exergue les éléments essentiels de la résolution de ce type d'équations.

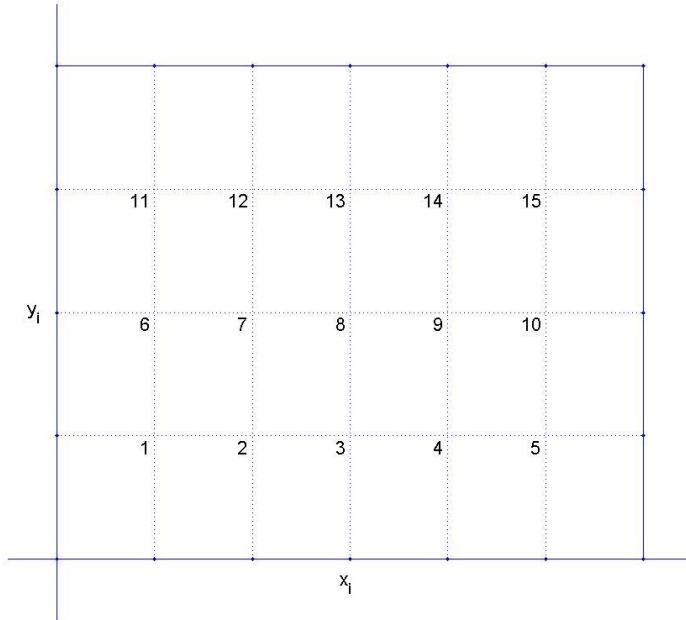
### **III.2 Résolution de l'équation de Poisson dans un rectangle avec des conditions aux limites de type Dirichlet**

Comme annoncé au chapitre précédent, la première étape de la résolution consiste à discréteriser les variables dépendantes : convenons de définir le rectangle sur lequel on résout III-1 par

$$\Omega = \{(x, y) : 0 \leq x \leq x_M, 0 \leq y \leq y_M\} \quad \text{III-4}$$

et choisissons de discréteriser ces variables de manière à ce que le maillage qui en résulte soit carré de largeur  $h$  :

$$\begin{aligned} x_i &= ih \quad i : 0 \rightarrow N_x & N_x &= \frac{x_M}{h} \\ y_j &= ih \quad j : 0 \rightarrow N_y & N_y &= \frac{y_M}{h} \end{aligned} \quad \text{III-5}$$



La deuxième étape consiste à écrire l'équation à résoudre et les conditions aux limites en tous les nœuds du maillage ; dans le cas de conditions de type Dirichlet, la fonction  $u$  étant connue en tous les nœuds situés sur  $\Gamma$  (nœuds « renforcés » sur la figure), il n'y a qu'en les nœuds intérieurs à  $\Omega$  que l'équation III-1 doit être écrite : pour ces nœuds, les variables  $i$  et  $j$  de III-5 varient selon

$$i:1 \rightarrow N_x - 1 \quad j:1 \rightarrow N_y - 1 \quad \text{III-6}$$

On écrit donc

$$\left( \frac{\partial^2 u}{\partial x^2} \right)_{ij} + \left( \frac{\partial^2 u}{\partial y^2} \right)_{ij} = f(x_i, y_j) \equiv f_{ij} \quad i:1 \rightarrow N_x - 1 \quad j:1 \rightarrow N_y - 1 \quad \text{III-7}$$

Si les conditions aux limites sont

$$u = g(x, y) \quad \forall (x, y) \in \Gamma \quad \text{III-8}$$

leur écriture en les nœuds frontières est

$$u_{ij} = g_{ij} \quad \text{avec}$$

$$i \times j = \{(0, \dots, N_x) \times (0), (0, \dots, N_x) \times (N_y), (0) \times (1, \dots, N_y - 1), (N_x) \times (1, \dots, N_y - 1)\} \quad \text{III-9}$$

La troisième étape de la méthode consiste à remplacer dans III-7 et III-8 les dérivées par des différences finies. La formule de différences finies la plus couramment utilisée pour évaluer le laplacien de III-7 est la relation II-21 :

$$\left( \frac{\partial^2 u}{\partial y^2} \right)_{ij} + \left( \frac{\partial^2 u}{\partial x^2} \right)_{ij} = \frac{1}{h^2} [u_{ij+1} + u_{i+1j} - 4u_{ij} + u_{ij-1} + u_{i-1j}] \quad \text{II-21}$$

En se référant à la figure précédente où, ayant choisi pour fixer les idées  $N_x = 6$  et  $N_y = 4$ , la numérotation des nœuds intérieurs va de 1 à 15 et la relation II-21 appliquée à un nœud intérieur non voisin des bords (par exemple le nœud 7) s'écrit

$$\frac{u_{12} + u_8 + u_2 + u_6 - 4u_7}{h^2} = f_7 \quad \text{III-10}$$

En un nœud voisin des bords (par exemple le nœud 14), il est nécessaire de faire intervenir les conditions aux limites :

$$\frac{u_{13} + g_{44} + u_{15} + u_9 - 4u_{14}}{h^2} = f_{14} \quad \text{III-11}$$

L'écriture de formules analogues à III-10 et III-11 pour tous les nœuds 1,...15 fournit un système linéaire de 15 équations en les 15 inconnues  $u_1, \dots, u_{15}$  ; si on convient de noter ce système  $A\bar{u} = \bar{b}$ , la matrice A s'écrit :

$$A = \left( \begin{array}{cc|ccc|c} -4 & 1 & & 1 & & \\ 1 & -4 & 1 & & 1 & \\ & 1 & -4 & 1 & & \\ & & 1 & -4 & 1 & \\ & & & 1 & -4 & \\ & & & & 1 & | & 1 \\ \hline 1 & & -4 & 1 & & 1 & \\ & 1 & & 1 & -4 & 1 & \\ & & 1 & & 1 & -4 & \\ & & & 1 & & 1 & \\ & & & & 1 & -4 & \\ & & & & & 1 & | & 1 \\ \hline & & 1 & & -4 & 1 & \\ & & & 1 & & 1 & \\ & & & & 1 & -4 & \\ & & & & & 1 & -4 & \\ & & & & & & 1 & -4 \end{array} \right) \quad \text{III-12}$$

Les éléments non nuls de cette matrice sont situés dans 3 diagonales de blocs diagonaux carrés de dimension égale à 5 ; la diagonale centrale compte 3 blocs et les autres 2 : cette structure est directement liée au découpage de  $\Omega$  par le maillage et à la numérotation des nœuds : à cet égard, si on avait numéroté les nœuds « verticalement », on aurait obtenu des blocs de dimension 3, au nombre de 5 pour la diagonale centrale et de 4 pour les 2 autres diagonales, les blocs centraux étant du type :

$$\begin{pmatrix} -4 & 1 & \\ 1 & -4 & 1 \\ & 1 & -4 \end{pmatrix}$$

Dans les deux cas on a affaire à une matrice dite tridiagonale par blocs. Les systèmes linéaires présentant ce type de matrice sont généralement résolus par des méthodes itératives (voir plus loin).

Le vecteur des inconnues est  $\bar{u} = (u_1 \ u_2 \ \dots \ u_{15})^T$  et le vecteur  $\bar{b}$  des termes indépendants vaut

$$\begin{pmatrix} h^2 f_1 - (g_{10} + g_{01}) \\ h^2 f_2 - g_{20} \\ h^2 f_3 - g_{30} \\ h^2 f_4 - g_{40} \\ h^2 f_5 - (g_{50} + g_{61}) \\ h^2 f_6 - g_{02} \\ h^2 f_7 \\ h^2 f_8 \\ h^2 f_9 \\ h^2 f_{10} - g_{62} \\ h^2 f_{11} - (g_{03} + g_{14}) \\ h^2 f_{12} - g_{24} \\ h^2 f_{13} - g_{34} \\ h^2 f_{14} - g_{44} \\ h^2 f_{15} - (g_{54} + g_{63}) \end{pmatrix}$$

Comme on a pu l'observer, l'introduction des conditions aux limites s'est faite naturellement. Ceci résulte de la nature des conditions (elles sont de type Dirichlet) et de la coïncidence des bords du domaine  $\Omega$  avec des lignes du maillage. Lorsqu'une des deux au moins de ces caractéristiques n'est pas rencontrée, la prise en compte des conditions aux limites est plus délicate et mérite qu'on s'y attarde plus longuement.

### III.3 Discréétisation des conditions aux limites

Cette discréétisation est une opération où l'ingéniosité du numéricien est souvent requise. Il n'existe pas de solution-miracle universelle et les procédés fournis ci-après quoique généraux peuvent être mal adaptés à des problèmes spécifiques. Ajoutons que la qualité de cette discréétisation est essentielle au point de faire échouer une résolution, ou de fournir des résultats complètement erronés si les choix opérés sont malheureux. Comme annoncé, les procédés dépendent de la nature des conditions et de la forme du domaine spatial. Remarquons enfin que ces procédés, détaillés ci-après, seront valables pour tout type d'EDP, et non pas seulement pour les EDP elliptiques.

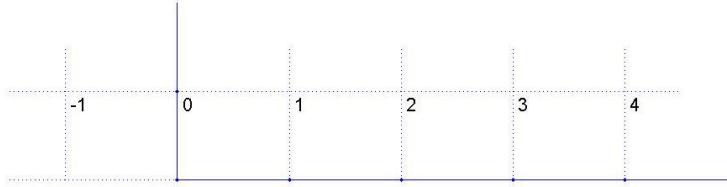
#### *Conditions de type Dirichlet*

a) Le maillage coïncide avec le contour  $\Gamma$  :

C'est ce qui a été traité au paragraphe précédent : on a vu que dans ce cas, l'équation à résoudre fait l'objet de la résolution par les différences finies seulement en les nœuds intérieurs à  $\Omega$  et que les conditions aux limites remplacent les valeurs de  $u$  en les nœuds appartenant à  $\Gamma$ .

Il existe néanmoins des cas où ce remplacement demande un aménagement de la procédure : imaginons pour illustrer que l'EDP à résoudre contient le terme  $u_{xx}$ . Si le pattern de points retenu pour évaluer cette dérivée au point  $i$  est  $(i-1, i, i+1)$ , on sait que la formule de différences finies à utiliser est II-12 :

$$(u_{xx})_i = \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} \quad \text{II-12}$$

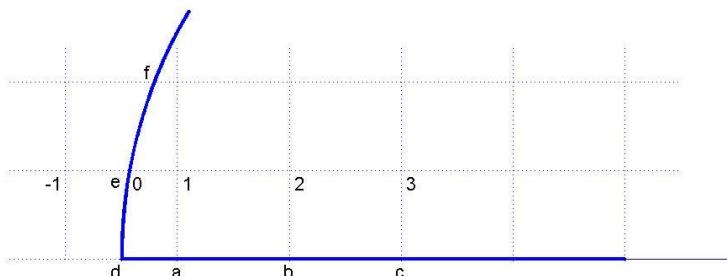


Appliquée au point 1 voisin de  $\Gamma$ , II-12 donne

$$(u_{xx})_1 = \frac{u_2 - 2u_1 + u_0}{h^2} \quad \text{III-13}$$

qui ne pose aucune difficulté d'évaluation, puisque  $u_0$  est fixé par les conditions aux limites. Si maintenant on souhaite évaluer cette dérivée par un schéma déduit du pattern  $(i-2, i-1, i, i+1, i+2)$ ,  $(u_{xx})_1$  sera évalué par une combinaison linéaire des valeurs  $u_{-1}, u_0, u_1, u_2$  et  $u_3$  où  $u_{-1}$  ne peut clairement pas être déduit des conditions aux limites. Dans ce cas, le seul recours consiste à décaler le pattern utilisé au point 1 : on utilisera la formule de différences finies construite sur les valeurs  $u_0, u_1, u_2, u_3$  et  $u_4$ .

b) le maillage n'épouse pas le contour  $\Gamma$  :



Grâce aux conditions aux limites, la valeur de  $u$  est connue en les noeuds du maillage coïncidant avec  $\Gamma$  (noeuds a, b, c), mais aussi en les intersections du maillage avec ce même contour (noeuds d, e  $\equiv 0$  et f).

Comme dans le cas précédent, l'équation à résoudre est à traiter par les différences finies en les noeuds intérieurs à  $\Omega$  et non sur  $\Gamma$  : si on utilise à nouveau II-12 pour évaluer  $u_{xx}$ , aucune difficulté n'apparaît pour les noeuds non-adjacents à  $\Gamma$ , pas plus que pour les noeuds voisins des portions de  $\Gamma$  qui coïncident avec le maillage (noeuds 2 et 3). C'est quand un noeud intérieur est voisin d'une portion de  $\Gamma$  non superposée au maillage que les difficultés apparaissent (noeud 1). Pour de tels noeuds, la procédure de calcul de  $u_{xx}$  consiste à estimer  $u$  en les noeuds 2 et 0 à partir de développements de Taylor de  $u$  autour du noeud 1 :

$$u_2 = u_1 + hu_{x1} + \frac{h^2}{2} u_{xx1} \quad III-14$$

$$u_0 = u_1 - (\theta h)u_{x1} + \frac{(\theta h)^2}{2} u_{xx1} \quad III-15$$

où  $\theta h$  est une fraction connue de  $h$  dépendant de la position du nœud 0. La combinaison de ces deux expressions fournit  $u_{xx1}$  :

$$u_{xx1} = \frac{2[\theta u_2 - (\theta + 1)u_1 + u_0]}{h^2 \theta (\theta + 1)} \quad III-16$$

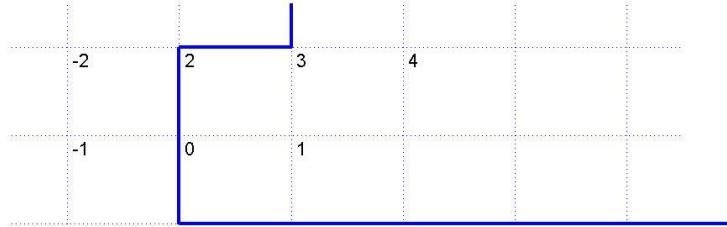
Ce procédé se généralise sans peine au calcul de toute dérivée.

### **Conditions de type Neumann**

Rappelons d'abord la forme de ces conditions :

$$\frac{\partial u}{\partial n}(\bar{x}_\Gamma, t) = g_{\Gamma, N}(t) \quad I-42$$

a) Le maillage coïncide avec le contour  $\Gamma$  :



Avec de telles conditions aux limites,  $u$  est inconnu le long de  $\Gamma$  et doit donc y être calculé, de la même manière qu'en les nœuds intérieurs à  $\Omega$ . Cela signifie que l'EDP à résoudre doit être discrétisée en les nœuds 0, 2 et 3 de la même manière qu'en les nœuds 1 et 4. Si on suppose toujours devoir calculer  $u_{xx}$  avec la formule II-12, des points délicats de la figure sont 0, 2 et 3 : on écrira

$$\begin{aligned} u_{xx0} &= \frac{u_1 - 2u_0 + u_{-1}}{h^2} \\ u_{xx2} &= \frac{u_3 - 2u_2 + u_{-2}}{h^2} \\ u_{xx3} &= \frac{u_4 - 2u_3 + u_2}{h^2} \end{aligned} \quad III-17$$

Le point 3 n'offrant pas de difficulté, il reste à évaluer les valeurs  $u_{-1}$  et  $u_{-2}$  en les nœuds fictifs -1 et -2. Ces valeurs découlent de la discrétisation de I-42 :

$$- I-42 \text{ s'écrit en } 0 : \left( \frac{\partial u}{\partial n} \right)_0 = - \left( \frac{\partial u}{\partial x} \right)_0 = \frac{u_{-1} - u_1}{2h} = (g_{\Gamma, N})_0 \quad III-18$$

III-17 et III-18 donnent alors

$$u_{xx0} = \frac{2}{h^2} (u_1 - u_0 + h(g_{\Gamma,N})_0) \quad \text{III-19}$$

- l'écriture de I-42 en 2 est plus délicate : la direction normale à  $\Gamma$  en ce point n'étant a priori pas définie, on peut imaginer par raison de continuité qu'elle fait un angle de  $135^\circ$  avec l'axe des x positifs ; ceci permet alors de projeter sur cet axe la condition I-42 selon

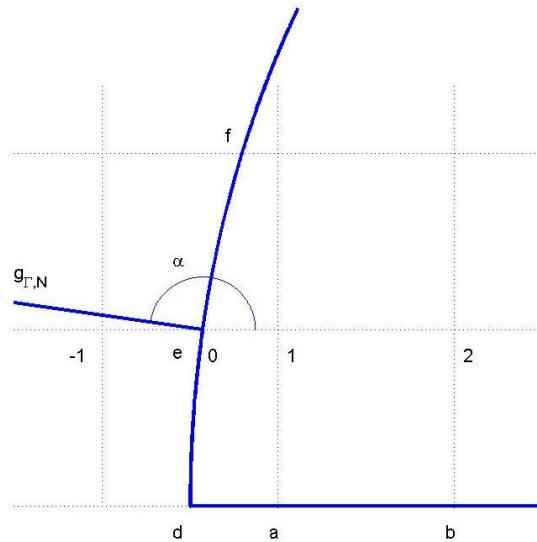
$$\left( \frac{\partial u}{\partial x} \right)_2 = - \left( \frac{\partial u}{\partial n} \right)_2 \frac{\sqrt{2}}{2} = - \frac{\sqrt{2}}{2} (g_{\Gamma,N})_2 = \frac{u_3 - u_{-2}}{2h} \quad \text{III-20}$$

qui donne alors avec III-17

$$u_{xx2} = \frac{2}{h^2} \left( u_3 - u_2 + h \frac{\sqrt{2}}{2} (g_{\Gamma,N})_2 \right) \quad \text{III-21}$$

b) le maillage n'épouse pas le contour  $\Gamma$  :

C'est le cas le plus compliqué à traiter.



A priori, compte tenu du maillage choisi, la solution de l'EDP est recherchée seulement en les nœuds du maillage intérieurs au domaine (nœuds tels 1 et 2) ou sur sa frontière (nœuds tels a et b) : la solution n'est donc pas à calculer en d, e et f. Voyons donc comment évaluer  $u_{xx}$  en 1 : le plus simple est d'écrire

$$u_{xx1} = \frac{u_2 - 2u_1 + u_{-1}}{h^2} \quad \text{III-22}$$

et de déduire  $u_{-1}$  de la condition aux limites : on a d'abord

$$\left( \frac{\partial u}{\partial x} \right)_0 = - \left( \frac{\partial u}{\partial n} \right)_0 \cos(\pi - \alpha) = (g_{\Gamma,N})_0 \cos(\alpha) \quad \text{III-23}$$

et ensuite, approximativement :

$$u_{-1} \cong u_1 - h \left( \frac{\partial u}{\partial x} \right)_0 = u_1 - h (g_{\Gamma,N})_0 \cos(\alpha) \quad III-24$$

Finalement

$$u_{xx1} = \frac{u_2 - u_1 - h (g_{\Gamma,N})_0 \cos(\alpha)}{h^2} \quad III-25$$

### **Conditions de type Robin**

Elles seront implémentées de la même manière que les conditions de type Neumann.

## **III.4 Application dans le domaine de la thermique**

Une plaque métallique mince de 4 cm sur 8 cm est le siège d'un dégagement de chaleur Q uniforme sur toute la surface. Si k désigne le coefficient de conductibilité thermique du matériau, le profil de température en régime dans la plaque obéit à l'équation de Poisson :

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{Q}{k} = 0 \quad III-26$$

avec  $Q = 5$      $k = 0.16$

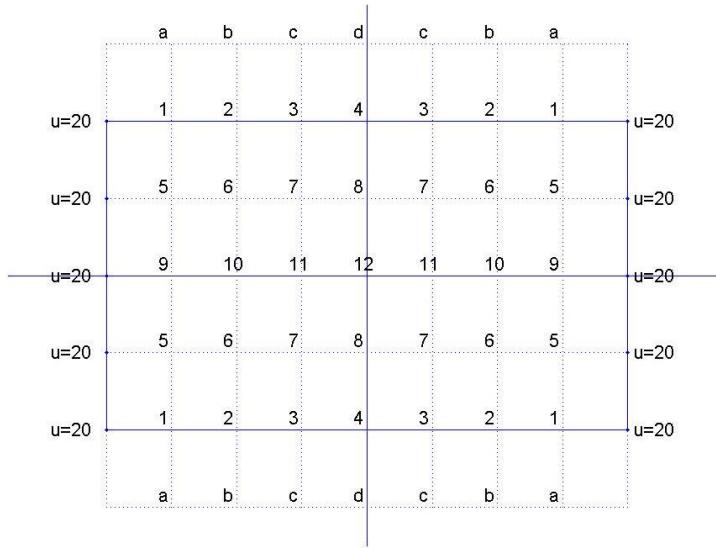
Les CL imposées sont les suivantes : la température est maintenue à 20 °C le long des petits côtés et le flux de chaleur s'échappant par les grands côtés vaut 15 °C/cm.

Ecrivons tout d'abord les CL dans le système d'axes oxy :

$$u(0, y) = 20 ; u(8, y) = 20 ; \left( \frac{\partial u}{\partial y} \right)_{(x, 0)} = +15 ; \left( \frac{\partial u}{\partial y} \right)_{(x, 4)} = -15 \quad III-27$$

Les signes des CL de type flux résultent du raisonnement suivant : la plaque étant le siège d'un dégagement de chaleur, le flux de calories qui s'en dégage est bien un flux sortant : ceci n'est possible que si la température de la plaque est supérieure à celle de l'ambiance. Cela signifie que  $\frac{\partial u}{\partial y}$  est positif en  $y = 0$  et est négatif en  $y = 4$ .

Il est élémentaire ici de faire coïncider le maillage avec  $\Gamma$ , ce qui facilite l'implémentation des conditions aux limites. On prendra donc par exemple  $h = 1\text{cm}$ . Le long des petits côtés du domaine, les conditions sont de type Dirichlet, et seront donc introduites naturellement dans les équations. Le long des grands côtés, la technique des nœuds fictifs sera appliquée. On supposera en outre que les « coins » du domaine d'étude font partie des petits côtés : les conditions aux limites y seront de type Dirichlet. Enfin, les symétries d'axes horizontal et vertical du problème permettent de limiter l'étude à un quart du domaine : la numérotation des nœuds proposée tient compte de cette limitation.



Détaillons la discréétisation de III-26 et de III-27 pour quelques nœuds typiques :

$$\text{III-26 au nœud 1 : } \frac{20 + u_a + u_2 + u_5 - 4u_1}{h^2} + \frac{5}{0.16} = 0 \quad \text{III-28}$$

$$\text{III-26 au nœud 7 : } \frac{u_6 + u_3 + u_8 + u_{11} - 4u_7}{h^2} + \frac{5}{0.16} = 0 \quad \text{III-29}$$

$$\text{III-26 au nœud 12 : } \frac{u_{11} + u_8 + u_{11} + u_8 - 4u_{12}}{h^2} + \frac{5}{0.16} = 0 \quad \text{III-30}$$

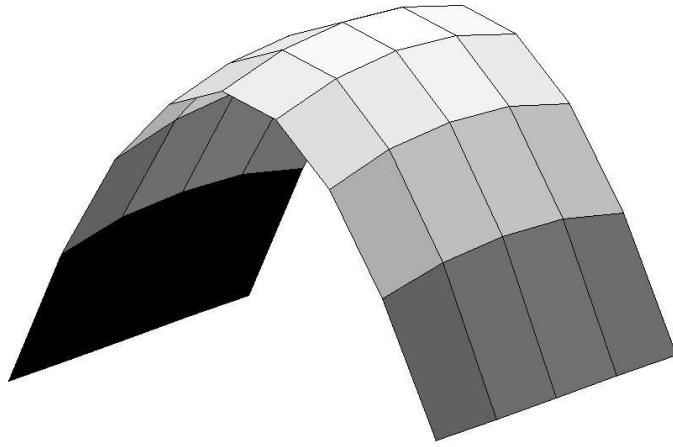
$$\text{III-27 au nœud 3 : } \frac{u_c - u_7}{2h} = -15 \quad \text{III-31}$$

L'extension des relations précédentes aux autres nœuds débouche finalement sur le système linéaire en les inconnues  $u_1$  à  $u_{12}$  suivant :

$$\left( \begin{array}{ccccccccc|c} -4 & 1 & & 2 & & & & & \\ 1 & -4 & 1 & & 2 & & & & \\ & 1 & -4 & 1 & & 2 & & & \\ & & 2 & -4 & & & 2 & & \\ 1 & & & -4 & 1 & & 1 & & \\ & 1 & & 1 & -4 & 1 & & 1 & \\ & & 1 & & 1 & -4 & 1 & & 1 \\ & & & 2 & -4 & & & 1 & \\ & & & 2 & & -4 & 1 & & \\ & & & & 2 & & 1 & -4 & 1 \\ & & & & & 2 & 1 & -4 & 1 \\ & & & & & & 2 & -4 & 1 \end{array} \right) \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{12} \end{pmatrix} = \begin{pmatrix} -5/.16 + 10 \\ -5/.16 + 30 \\ -5/.16 + 30 \\ -5/.16 + 30 \\ -5/.16 - 20 \\ -5/.16 \\ -5/.16 \\ -5/.16 \\ -5/.16 - 20 \\ -5/.16 \\ -5/.16 \\ -5/.16 \end{pmatrix} \quad \text{III-32}$$

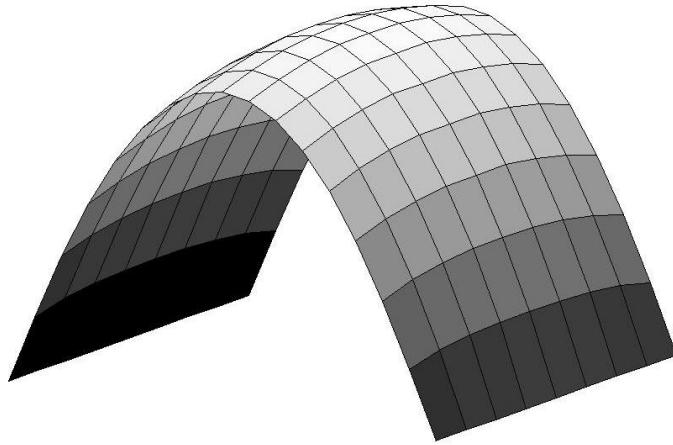
La matrice est tridiagonale par blocs, comme lors des conditions aux limites de type Dirichlet, mais la régularité des coefficients est affectée par les conditions aux limites mixtes et la symétrie du problème.

La résolution de ce système (par une méthode itérative : voir plus loin) débouche sur la solution suivante



Comme on pouvait s'y attendre, c'est au nœud 12 que la température calculée est la plus élevée : 215.55°.

On peut se faire une idée de la qualité de cette solution en la comparant par exemple à celle qu'on obtient avec une maille de largeur  $\frac{h}{2}$ , c'est-à-dire quatre fois plus petite :

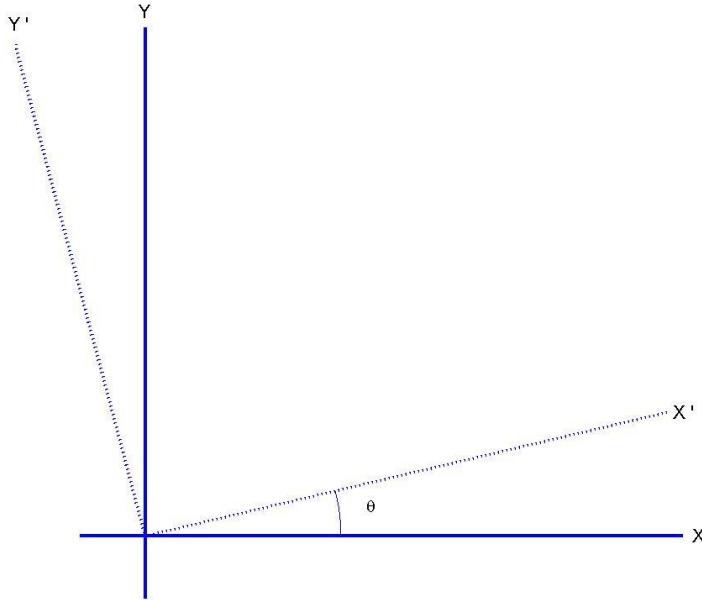


La température au même endroit du domaine y est cette fois de 215.12°, ce qui donne une idée de la qualité de la première solution.

### III.5 Compléments

Il est parfois intéressant d'utiliser un maillage du domaine  $\Omega$  autre que le maillage rectangulaire. Deux maillages présentent un intérêt particulier.

### maillage triangulaire



Le calcul de  $\nabla^2 u$  dans un tel système d'axes est basé sur la relation liant la dérivée seconde spatiale suivant un axe  $X'$  aux dérivées secondes spatiales dans le système OXY : si  $\theta$  est l'angle fixant la position de l'axe  $X'$  dans le repère OXY, on peut montrer qu'on a

$$\frac{\partial u}{\partial x'} = \frac{\partial u}{\partial x} \cos \theta + \frac{\partial u}{\partial y} \sin \theta \quad \text{III-33}$$

$$\frac{\partial^2 u}{\partial x'^2} = \frac{\partial^2 u}{\partial x^2} \cos^2 \theta + 2 \frac{\partial^2 u}{\partial x \partial y} \sin \theta \cos \theta + \frac{\partial^2 u}{\partial y^2} \sin^2 \theta \quad \text{III-34}$$

Dans le cas (le plus courant) d'un maillage en triangles équilatéraux, les directions A, B et C faisant un angle de 0, 60 et 120 degrés avec OX (figure suivante), on obtient aisément, à partir de III-34 :

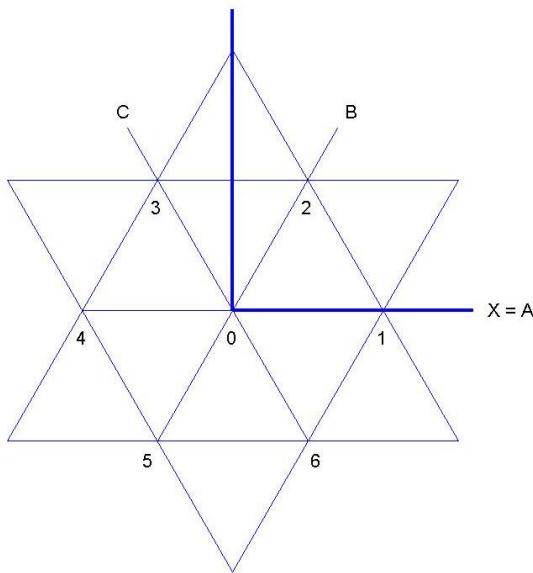
$$\frac{\partial^2 u}{\partial a^2} = \frac{\partial^2 u}{\partial x^2} \quad \text{III-35}$$

$$\frac{\partial^2 u}{\partial b^2} = \frac{1}{4} \frac{\partial^2 u}{\partial x^2} + \frac{\sqrt{3}}{2} \frac{\partial^2 u}{\partial x \partial y} + \frac{3}{4} \frac{\partial^2 u}{\partial y^2} \quad \text{III-37}$$

$$\frac{\partial^2 u}{\partial c^2} = \frac{1}{4} \frac{\partial^2 u}{\partial x^2} - \frac{\sqrt{3}}{2} \frac{\partial^2 u}{\partial x \partial y} + \frac{3}{4} \frac{\partial^2 u}{\partial y^2} \quad \text{III-37}$$

c'est-à-dire

$$\frac{\partial^2 u}{\partial a^2} + \frac{\partial^2 u}{\partial b^2} + \frac{\partial^2 u}{\partial c^2} = \frac{3}{2} \nabla^2 u \quad \text{III-38}$$



Et si chacune des dérivées secondes est évaluée par le schéma centré à trois points habituel, il vient

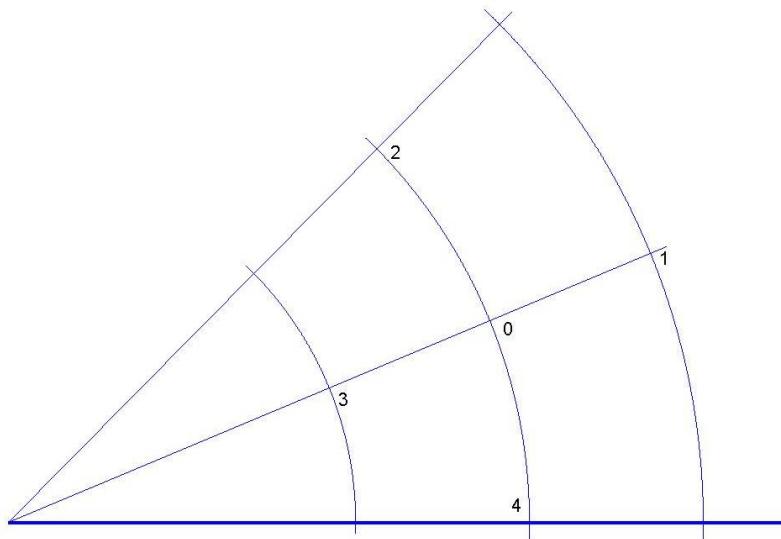
$$(\nabla^2 u)_0 = \frac{2}{3h^2} (u_1 + u_2 + u_3 + u_4 + u_5 + u_6 - 6u_0) \quad \text{III-39}$$

### *maillage en coordonnées polaires*

si  $\Delta\theta$  et  $\Delta r$  sont les incrémentations en angle et en rayon du maillage (figure ci-dessous), alors on peut montrer qu'on a

$$(\nabla^2 u)_0 = \frac{1}{\Delta r^2} \left[ (1 + \frac{\Delta r}{2r_0})u_1 + (1 - \frac{\Delta r}{2r_0})u_3 + (\frac{\Delta r}{r_0 \Delta \theta})^2 u_2 + (\frac{\Delta r}{r_0 \Delta \theta})^2 u_4 - [2 + 2(\frac{\Delta r}{r_0 \Delta \theta})^2]u_0 \right] \quad \text{III-40}$$

où  $r_0$  est la coordonnée radiale du point 0.



### III.6 Résolution itérative des systèmes linéaires

La résolution par voie itérative des systèmes d'équations linéaires est particulièrement bien adaptée lorsque ces systèmes possèdent deux caractéristiques : ils sont de grande taille et leur matrice comporte de nombreux éléments nuls. C'est exactement dans ce cas de figure que l'on se trouve lorsqu'on est amené à résoudre des équations elliptiques par la méthode des différences finies : le problème traité au paragraphe III.4 en est un exemple d'autant plus parlant si on imagine travailler avec un maillage spatial de plus en plus fin. A côté des méthodes traditionnelles telles que celles de Jacobi et de Gauss-Seidel existent d'autres méthodes dont les paragraphes suivants donnent un aperçu. Mais exposons d'abord brièvement les méthodes citées.

#### *Méthodes de Jacobi et de Gauss-Seidel*

Si le système à résoudre est mis sous la forme

$$A\bar{u} = \bar{b} \quad \text{III-41}$$

la méthode de Jacobi construit une suite d'approximations  $\bar{u}^k$  de la solution à partir d'une estimation initiale  $\bar{u}^0$  à l'aide de la relation suivante :

$$u_i^{k+1} = \frac{1}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij} u_j^k - \sum_{j=i+1}^n a_{ij} u_j^k \right) \quad i = 1, \dots, n \quad \text{III-42}$$

Cette relation possède un équivalent matriciel obtenu en décomposant la matrice A en

$$A = L + D + U \quad \text{III-43}$$

où

$$L = \begin{pmatrix} 0 & \cdots & \cdots & 0 \\ a_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn-1} & 0 \end{pmatrix} \quad D = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & a_{nn} \end{pmatrix} \quad U = \begin{pmatrix} 0 & a_{12} & \cdots & a_{1n} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & a_{n-1n} \\ 0 & \cdots & \cdots & 0 \end{pmatrix} \quad \text{III-44}$$

Il est en effet facile de transformer III-42 en

$$D\bar{u}^{k+1} = -(L + U)\bar{u}^k + \bar{b} \quad \text{III-45}$$

La méthode de Gauss-Seidel construit une suite d'approximations  $\bar{u}^k$  de la solution à partir d'une estimation initiale  $\bar{u}^0$  à l'aide de la relation suivante :

$$u_i^{k+1} = \frac{1}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij} u_j^{k+1} - \sum_{j=i+1}^n a_{ij} u_j^k \right) \quad i = 1, \dots, n \quad \text{III-46}$$

Son équivalent matriciel est alors

$$(D + L)\bar{u}^{k+1} = -U\bar{u}^k + \bar{b} \quad \text{III-47}$$

### Convergence

Observons d'abord que les deux méthodes qui précèdent correspondent à une même démarche de partitionnement de la matrice initiale A :

$$A = A_L + A_R \quad \text{III-48}$$

avec pour Jacobi :  $A_L = D$        $A_R = L + U$       III-49

et pour Gauss-Seidel :  $A_L = D + L$        $A_R = U$       III-50

Avec ces notations, III-45 et III-47 deviennent

$$A_L \bar{u}^{k+1} = -A_R \bar{u}^k + \bar{b} \quad \text{III-51}$$

C'est sur ce formalisme général que portera l'étude de la convergence. Celle-ci peut être évaluée par exemple par l'erreur qui affecte l'itéré de rang k ou le résidu de rang k.

*Erreur sur l'itéré de rang k* : si  $\bar{U}$  représente la solution exacte de III-41, cette erreur est définie par

$$\bar{e}^k = \bar{U} - \bar{u}^k \quad \text{III-52}$$

*Résidu de rang k* : il est défini par

$$\bar{r}^k = A \bar{u}^k - \bar{b} \quad \text{III-53}$$

Les méthodes précédentes seront convergentes si ces grandeurs tendent vers zéro quand k tend vers l'infini.

Intéressons-nous plus particulièrement à l'erreur : puisque  $\bar{U}$  est la solution exacte, elle vérifie III-51 :

$$A_L \bar{U} = -A_R \bar{U} + \bar{b} \quad \text{III-54}$$

Soustrayons III-51 à III-54

$$\begin{aligned} & A_L \bar{U} - A_L \bar{u}^{k+1} = -A_R \bar{U} + \bar{b} - (-A_R \bar{u}^k + \bar{b}) \\ \Leftrightarrow & A_L \bar{e}^{k+1} = -A_R \bar{e}^k \\ \Leftrightarrow & A_L \bar{e}^{k+1} = (A_L - A) \bar{e}^k \\ \Rightarrow & \bar{e}^{k+1} = (I - A_L^{-1} A) \bar{e}^k = (I - A_L^{-1} A)^{k+1} \bar{e}^0 \end{aligned} \quad \text{III-55}$$

La matrice  $I - A_L^{-1} A$ , notée G, est appelée matrice d'amplification. III-51 sera donc une méthode convergente si la norme de cette matrice est inférieure à un, ou encore si son rayon spectral  $\rho(G)$  est inférieur à un. Une autre manière d'exprimer cette condition est de dire que toutes les valeurs propres de G doivent être en module inférieures à un :

$$\lambda_j(G) < 1 \quad \forall j \quad \text{III-56}$$

Comme choisir une méthode itérative revient finalement à choisir un partitionnement de type III-48, il est clair que la recherche d'une méthode itérative efficace peut se résumer à partitionner A de telle sorte que III-56 soit vérifié et que III-51 puisse être implémenté avec une bonne protection contre les accumulations d'erreurs d'arrondi (il faut pour cela que  $A_L$  soit facilement inversible).

Pour les méthodes de Jacobi et de Gauss-Seidel, la matrice d'amplification vaut

$$\text{Jacobi : } G_J = I - D^{-1}A \quad \text{III-57}$$

$$\text{Gauss-Seidel : } G_{GS} = I - (D + L)^{-1}A \quad \text{III-58}$$

On peut montrer que ces matrices ont un rayon spectral inférieur à un dans deux cas rencontrés dans la majorité des systèmes linéaires déduits des équations elliptiques :

1° A est symétrique définie positive

2° A est irréductible et à diagonale principale dominante : ceci est le cas le plus souvent rencontré :

A est à diagonale principale dominante signifie

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| \quad \forall i \quad \text{et} \quad \exists k : |a_{kk}| > \sum_{\substack{j=1 \\ j \neq k}}^n |a_{kj}|$$

ou bien

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ji}| \quad \forall i \quad \text{et} \quad \exists k : |a_{kk}| > \sum_{\substack{j=1 \\ j \neq k}}^n |a_{jk}|$$

A est irréductible signifie il n'existe aucune matrice P dit de permutation telle qu'on ait

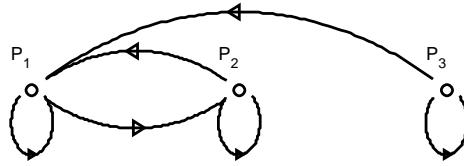
$$P^T A P = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} \quad \text{où } A_{11} \text{ et } A_{22} \text{ sont des matrices carrées de dimension p et q .}$$

En pratique, l'irréductibilité d'une matrice est détectée au moyen du graphe  $G(A)$  qu'on lui associe ; celui-ci est obtenu de la manière suivante : si A est de dimension n, on lui associe n éléments  $P_1, \dots, P_n$ . Un arc orienté  $P_i \rightarrow P_j$  relie  $P_i$  à  $P_j$  si  $a_{ij} \neq 0$ . A est irréductible ssi G est connecté c'est-à-dire si pour chaque paire  $(P_i, P_j)$  il existe un chemin orienté de  $P_i$  à  $P_j$ .

Par exemple, pour la matrice suivante :

$$A = \begin{pmatrix} 1 & 2 & 0 \\ -1 & 1 & 0 \\ 3 & 0 & 5 \end{pmatrix}$$

le graphe est



A n'est donc pas irréductible (pas de chemin de 1 vers 3 ni de 2 vers 3)  
Notons encore les propriétés de convergence suivantes :

1° la convergence des méthodes de Jacobi et de Gauss-Seidel est indépendante du vecteur  $\bar{b}$  : cette propriété sera mise à profit lors de la relaxation (voir plus loin).

2° les méthodes de Jacobi et de Gauss-Seidel sont des méthodes de point fixe linéaire : elles s'écrivent

$$\bar{u}^{k+1} = (I - A_L^{-1}A)\bar{u}^k + A_L^{-1}\bar{b}$$

Ceci a deux conséquences, heureuse et malheureuse : la convergence est linéaire, c'est-à-dire lente ( $\alpha \neq 0$ ) mais la valeur de  $\bar{u}^0$  peut être quelconque.

### III.7 Méthodes relaxées

Rappelons que si  $u_{iR}^k$  désigne l'itéré  $k$  relaxé de l'inconnue  $u_i$  et  $u_{iNR}^k$  la même grandeur non relaxée (c'est celle qui est calculée au paragraphe précédent), le procédé de relaxation consiste à remplacer le calcul de  $u_{iNR}^{k+1}$  par celui de

$$u_{iR}^{k+1} = u_{iR}^k + \omega(u_{iNR}^{k+1} - u_{iR}^k) \quad \text{III-59}$$

où  $\omega$  réel est le paramètre de relaxation.

#### *Méthode de Jacobi relaxée*

On y remplace III-42 par

$$u_i^{k+1} = u_i^k + \omega \left( \frac{1}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij} u_j^k - \sum_{j=i+1}^n a_{ij} u_j^k \right) - u_i^k \right) \quad \text{III-60}$$

et III-45 par

$$D\bar{u}^{k+1} = -\omega A\bar{u}^k + D\bar{u}^k + \omega\bar{b} \quad \text{III-61}$$

$$\Leftrightarrow \frac{1}{\omega} D\bar{u}^{k+1} = -A\bar{u}^k + \frac{1}{\omega} D\bar{u}^k + \bar{b} \quad \text{III-62}$$

La matrice d'amplification vaut donc

$$G_{J,R} = I - \omega D^{-1}A = \omega I - \omega I + I - \omega D^{-1}A = \omega G_J + (1-\omega)I \quad \text{III-63}$$

La convergence sera assurée si le rayon spectral de III-63 est inférieur à un, c'est-à-dire si

$$\begin{aligned}\rho(G_{J,R}) &\leq \omega\rho(G_J) + |1 - \omega| < 1 \\ \Leftrightarrow -1 + \omega\rho(G_J) &< 1 - \omega < 1 - \omega\rho(G_J) \\ \Leftrightarrow 2 - \omega\rho(G_J) &> \omega > \omega\rho(G_J)\end{aligned}$$

Comme  $\rho(G_J)$  est strictement positif, la double inégalité précédente est vérifiée si on a  $\rho(G_J) < 1$  et

$$0 < \omega < \frac{2}{1 + \rho(G_J)} \quad \text{III-64}$$

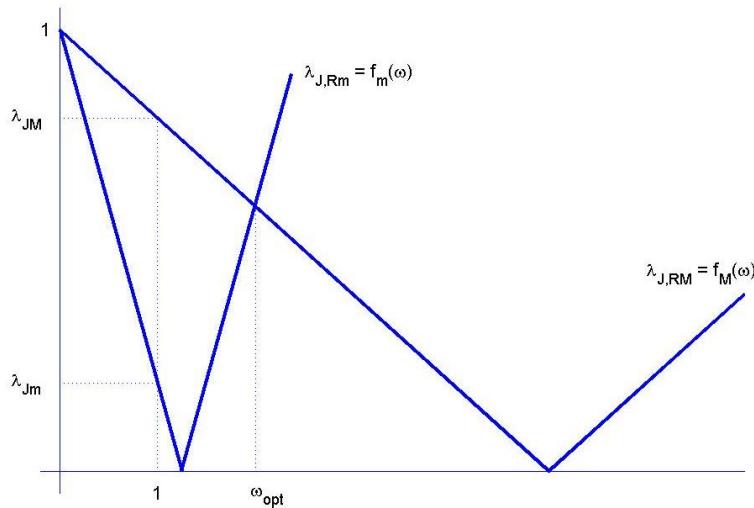
La méthode de Jacobi relaxée ne peut donc converger que si la méthode non relaxée converge elle-même. En termes de valeurs propres, III-63 donne également

$$\lambda_{J,R} = \omega\lambda_J + 1 - \omega \quad \text{III-65}$$

Soient  $\lambda_{Jm}$  et  $\lambda_{JM}$  la plus petite et la plus grande (en valeur absolue) des valeurs propres de  $G_J$ , toutes les deux inférieures à un car il faut  $\rho(G_J) < 1$ . Les valeurs propres correspondantes de la méthode relaxée sont des fonctions de  $\omega$  :

$$\begin{aligned}\lambda_{J,Rm} &= 1 - \omega(1 - \lambda_{Jm}) = f_m(\omega) \\ \lambda_{J,RM} &= 1 - \omega(1 - \lambda_{JM}) = f_M(\omega)\end{aligned} \quad \text{III-66}$$

Ces fonctions sont représentées ci-dessous en valeur absolue, pour deux valeurs arbitraires quelconques (mais comprises entre 0 et 1 à cause de  $\rho(G_J) < 1$ ) de  $\lambda_{Jm}$  et  $\lambda_{JM}$  :



C'est évidemment pour  $\lambda_{J,Rm} = \lambda_{J,RM}$  que la relaxation sera optimale ; il lui correspond la valeur suivante :

$$\omega_{opt} = \frac{1}{1 - \frac{\lambda_{Jm} + \lambda_{JM}}{2}} \quad \text{III-67}$$

### Méthode de Gauss-Seidel relaxée

On y remplace III-46 par

$$u_i^{k+1} = u_i^k + \omega \left( \frac{1}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij} u_j^{k+1} - \sum_{j=i+1}^n a_{ij} u_j^k \right) - u_i^k \right) \quad \text{III-68}$$

et III-47 par

$$\left( L + \frac{1}{\omega} D \right) \bar{u}^{k+1} = - \left( U + \left( 1 - \frac{1}{\omega} \right) D \right) \bar{u}^k + \bar{b} \quad \text{III-69}$$

La matrice d'amplification vaut donc

$$G_{GS,R} = I - \left( L + \frac{1}{\omega} D \right)^{-1} A = I - \omega ( \omega L + D )^{-1} A \quad \text{III-70}$$

Une étude détaillée de  $G$  débouche sur les résultats suivants :

- le rayon spectral  $\rho(G_{GS,R})$  de la méthode est inférieur à un si  $0 < \omega < 2$ .
- si  $A$  est symétrique,  $\rho(G_{GS,R})$  est minimal pour

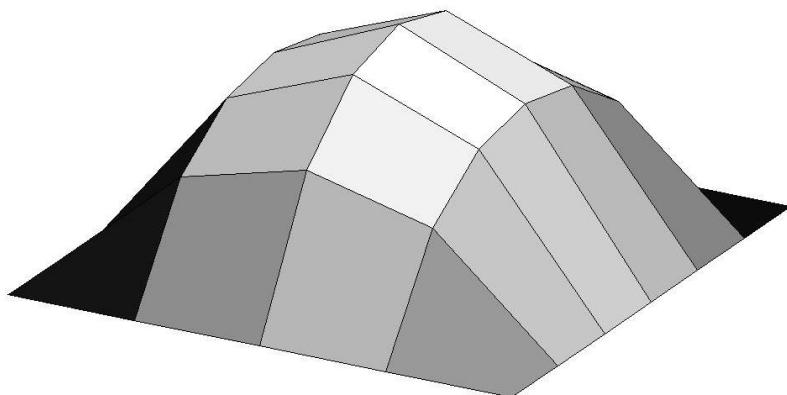
$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - \rho(G_J)^2}} \quad \text{III-71}$$

### III.8 Exemples

#### *Exemple 1*

Reprendons le problème traité au paragraphe III-2 en le complétant par les valeurs numériques suivantes :

$$f(x, y) = \text{cte} = \frac{-5}{0.16} \quad u = g(x, y) = x + y \quad \forall (x, y) \in \Gamma$$



La solution est représentée ci-dessus ; sa valeur maximale est atteinte au point de coordonnées  $(x, y) = (3,2)$  et vaut  $u(3,2) = 54.0151$ .

a) méthode de Jacobi non relaxée : pour la matrice III-12, on a (cf. III-57)

$$G_J = I + .25 * A \quad \text{III-72}$$

où  $A$  est la matrice III-12. Les valeurs propres de  $G_J$  sont comprises entre  $\lambda_{Jm} = -0.7866$  et  $\lambda_{JM} = 0.7866$ . La méthode est donc convergente.

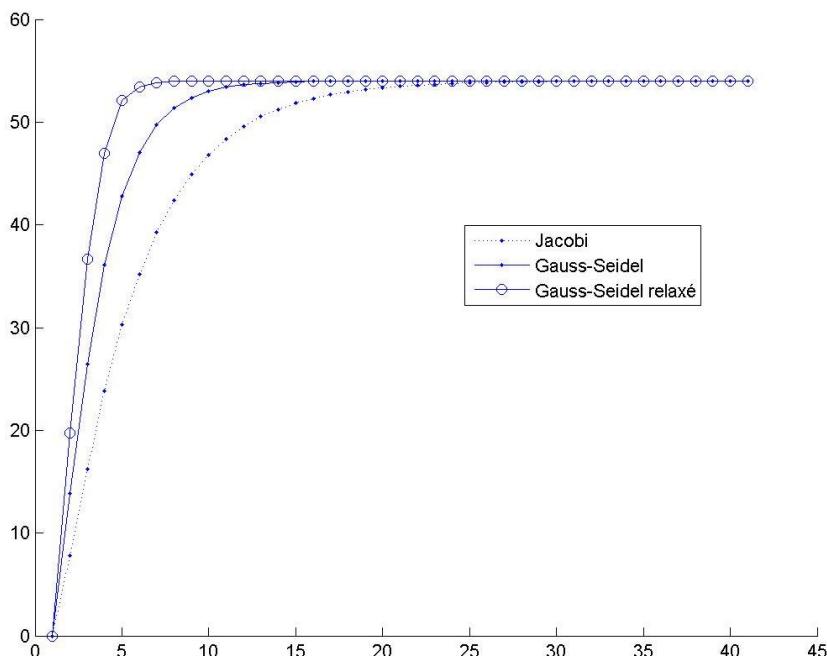
b) méthode de Jacobi relaxée : comme  $\lambda_{Jm} = -\lambda_{JM}$ ,  $\omega_{opt} = 1$ . La relaxation n'apporte donc rien.

c) méthode de Gauss-Seidel non relaxée : les valeurs propres de  $G_{GS}$  pour la matrice III-12 sont comprises entre 0.6187 et 0. La méthode converge également, en principe plus vite que la méthode de Jacobi puisque le rayon spectral de la méthode est plus petit.

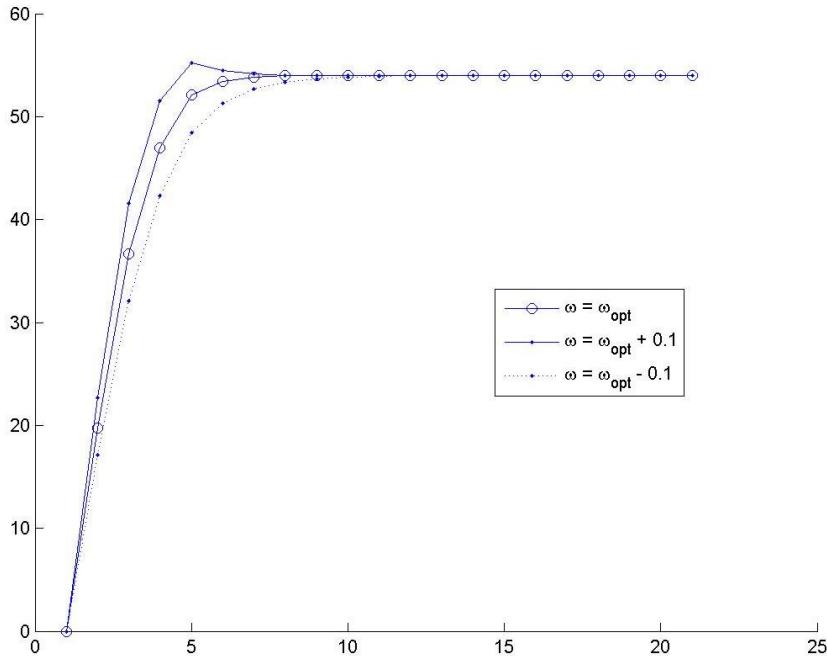
d) méthode de Gauss-Seidel relaxée : la matrice III-12 est symétrique et  $\rho(G_J) = 0.7866$  ; il en résulte par III-71

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - (0.7866)^2}} = 1.2365 \quad \text{III-73}$$

Les deux figures qui suivent permettent d'évaluer l'impact des possibilités qui précèdent ; la première représente la suite des valeurs de  $u(3,2)$  calculée à partir de  $\bar{u}^0 = \bar{0}$  par la méthode de Jacobi, de Gauss-Seidel et de Gauss-Seidel relaxée avec  $\omega = \omega_{opt}$  :



La deuxième compare les résultats obtenus par la méthode de Gauss-Seidel relaxée pour trois valeurs de  $\omega$ , donnant ainsi une idée de la sensibilité du résultat à ce paramètre



Notons que quand la matrice n'est pas symétrique, la recherche d'une valeur optimale du paramètre de relaxation est beaucoup plus difficile et a fait l'objet de nombreux développements théoriques. Ce qui suit propose une procédure simple de recherche par tâtonnements : la valeur optimale de  $\omega$  est celle qui rend la norme de la matrice d'amplification III-70 la plus petite possible. Cette matrice étant indépendante de  $\bar{b}$  dans la résolution de  $A\bar{u} = \bar{b}$ , c'est elle aussi qui détermine – via la relation III-55 – l'évolution des erreurs dans la résolution de  $A\bar{u} = \bar{0}$ . Comme la solution – connue – de ce problème est  $\bar{u} = \bar{0}$ , les itérés  $\bar{u}^k$  de cette résolution par toute méthode itérative sont identiques aux erreurs :

$$\bar{e}^k = \bar{0} - \bar{u}^k = -\bar{u}^k \quad \text{III-74}$$

La procédure est alors la suivante :

1° on choisit une valeur arbitraire  $\omega_0$  du paramètre de relaxation (en général  $\omega_0 = 1$ ) et on procède au calcul de quelques itérés  $\bar{u}^k = \bar{e}^k$  de la résolution de  $A\bar{u} = \bar{0}$  en démarrant avec  $\bar{u}^0 \neq \bar{0}$ . Ces itérés permettent de se faire une idée du module de contraction de la suite ainsi calculée :

$$\alpha = \lim_{k \rightarrow \infty} \frac{\|\bar{e}^{k+1}\|}{\|\bar{e}^k\|} \Rightarrow \text{on calcule } \frac{\|\bar{e}^1\|}{\|\bar{e}^0\|}, \frac{\|\bar{e}^2\|}{\|\bar{e}^1\|}, \frac{\|\bar{e}^3\|}{\|\bar{e}^2\|}, \dots$$

qui fournit une estimation  $\alpha_0 = \alpha(\omega_0)$ .

2° on modifie le paramètre de relaxation : soit  $\omega_1 = \omega_0 + \Delta\omega$  et on reprend la procédure précédente qui fournit  $\alpha_1 = \alpha(\omega_1)$ . Si  $\alpha_1 < \alpha_0$ , c'est qu'on a modifié le paramètre de relaxation « dans le bon sens » et on poursuit avec  $\omega_2 = \omega_1 + \Delta\omega$ , sinon il faut passer à  $\omega_1 = \omega_0 - \Delta\omega$ .

3° le procédé est poursuivi jusqu'à obtention d'une valeur satisfaisante de  $\omega$ , soit  $\omega_{\text{opt,est}}$

4° on résout alors  $A\bar{u} = \bar{b}$  avec cette valeur de  $\omega$ . Notons encore qu'il n'est pas nécessaire de rechercher un  $\omega_{opt,est}$  avec grande précision : la dernière figure montre que travailler avec  $\Delta\omega = 0.1$  est suffisamment précis ; en d'autres termes, si on démarre la procédure avec  $\omega_0 = 1$ , une petite dizaine de valeurs de  $\omega$  au maximum devront être testées (rappelons que la convergence n'est possible que si  $0 < \omega < 2$ ).

### **Exemple 2**

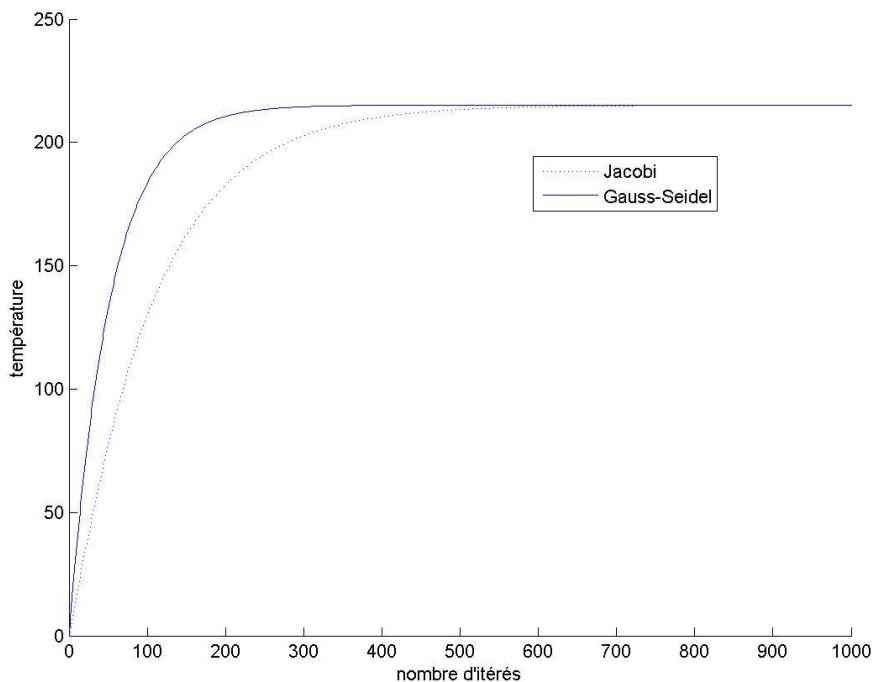
Le traitement du problème du paragraphe III-4 (avec le pas de maillage égal à  $\frac{h}{2}$ ) par ces mêmes méthodes donne les résultats suivants

a) méthodes de Jacobi non relaxée et relaxée : appliquée à la généralisation de la matrice III-32, la méthode se caractérise par une matrice d'amplification III-57 identique à celle du cas précédent :

$$G_J = I + .25 * A \quad \text{III-75}$$

Ses valeurs propres sont comprises entre  $\lambda_{Jm} = -0.9904$  et  $\lambda_{JM} = 0.9904$ . La méthode est convergente, mais comme dans l'exemple précédent, la relaxation n'apporte rien. C'est ce qui faire dire que lorsqu'on veut résoudre l'équation de Poisson par la méthode de Jacobi, le procédé de relaxation est inutile.

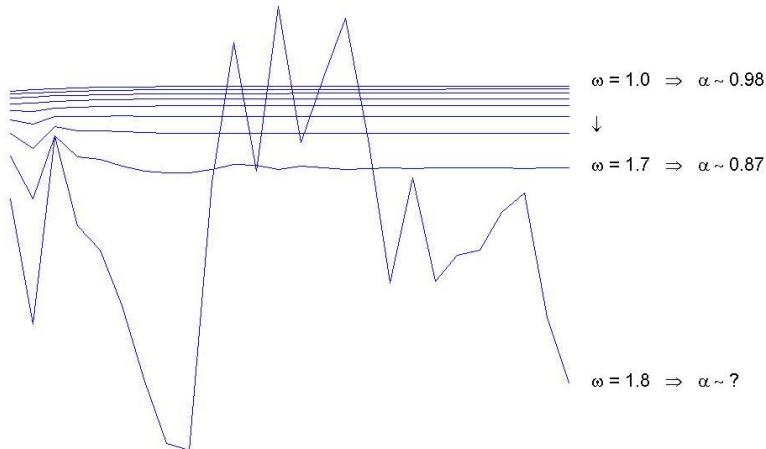
b) méthode de Gauss-Seidel : la figure suivante représente l'évolution de la température au nœud où elle est la plus élevée ( $215.12^\circ$ ) pour les méthodes de Jacobi et de Gauss-Seidel : cette dernière est donc à nouveau plus efficace.



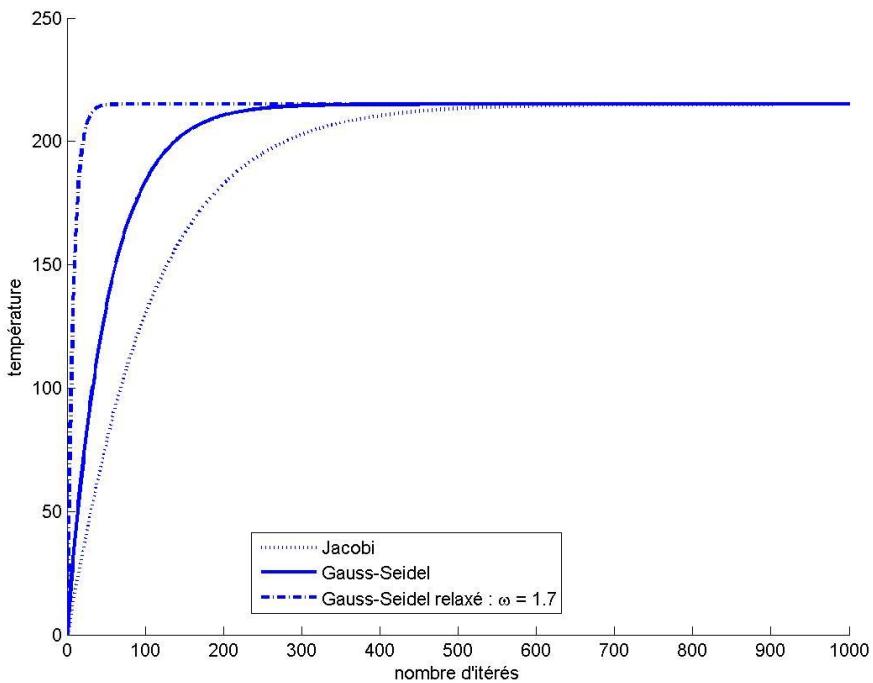
c) méthode de Gauss-Seidel relaxée : la matrice du système III-32 n'étant pas symétrique, le paramètre de relaxation ne peut être évalué que par tâtonnements comme expliqué plus haut. La figure suivante

représente les itérés de rang  $k = 5 \rightarrow 30$  de la suite  $\{\alpha_k\} = \left\{ \frac{\|\bar{e}^{k+1}\|}{\|\bar{e}^k\|} \right\}$  pour  $\omega = 1, 1.1, \dots, 1.8$  (on n'a pas

représenté  $\alpha_0, \dots, \alpha_4$  car les premières valeurs de cette suite ne représentent pas correctement le module de contraction).



On observe que la convergence s'améliore quand  $\omega$  passe de 1.0 à 1.7 pour se détériorer rapidement quand  $\omega = 1.8$ . Ceci indique que la meilleure valeur de  $\omega$  est de l'ordre de 1.7. La figure suivante donne une idée du gain d'accélération obtenu.



Signalons pour terminer que la méthode de Gauss-Seidel relaxée présente de nombreuses variantes dont le point commun est de modifier dans III-68 l'ordre naturel  $i : 1, \dots, n$  de calcul des inconnues  $u_i^{k+1}$ . Par exemple en alternant cet ordre naturel un pas sur deux avec l'ordre inverse.

Cela donne une méthode itérative à deux (demi-)pas qu'on peut écrire

$$\begin{cases} u_i^{k+1/2} = u_i^k + \omega \left( \frac{1}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij} u_j^{k+1/2} - \sum_{j=i+1}^n a_{ij} u_j^k \right) - u_i^k \right) & i : 1, \dots, n \\ u_i^{k+1} = u_i^{k+1/2} + \omega \left( \frac{1}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij} u_j^{k+1/2} - \sum_{j=i+1}^n a_{ij} u_j^{k+1} \right) - u_i^{k+1/2} \right) & i : n, \dots, 1 \end{cases} \quad \text{III-76}$$

On se référera à la littérature spécialisée pour plus de détails.

### III.9 Méthodes multigrilles

Considérées actuellement comme les méthodes itératives générales les plus efficaces pour la résolution des systèmes linéaires résultant de l'application de la méthode des différences aux équations aux dérivées partielles, ces méthodes exploitent l'idée suivante : si pour fixer les idées on suppose devoir résoudre l'équation de Laplace

$$\nabla^2 u = 0 \quad \text{III-77}$$

avec des conditions aux limites de type Dirichlet, il est intuitif d'admettre que l'influence de ces dernières va se propager d'autant plus lentement à travers tout le domaine spatial que la discréétisation spatiale est fine, simplement parce que la discréétisation du laplacien II-21 est compacte (la valeur de l'inconnue  $u_{ij}$  est seulement liée à celles de ses quatre voisines les plus proches). On a d'ailleurs pu observer l'influence de la finesse du maillage sur le nombre d'itérations utiles dans les deux exemples précédents : dans le premier, le maillage génère un système de dimension égale à 15, et dans le second on passe à 135 inconnues (avec des symétries exploitées pour résoudre plus rapidement) : une quinzaine d'itérations suffit dans le premier cas, et environ 650 sont nécessaires dans le second.

Les méthodes multigrille exploitent l'idée suivante : l'utilisation d'une grille grossière va permettre de propager rapidement les conditions aux limites à travers tout le domaine, et l'adjonction d'une grille fine permettra de préserver la précision. Ces méthodes seront applicables a priori aux méthodes qui précèdent, mais on se contentera dans ce texte introductif de l'appliquer à la méthode de Gauss-Seidel pour la résolution de III-77. Fixons plus précisément l'exemple traité : le domaine spatial est le carré

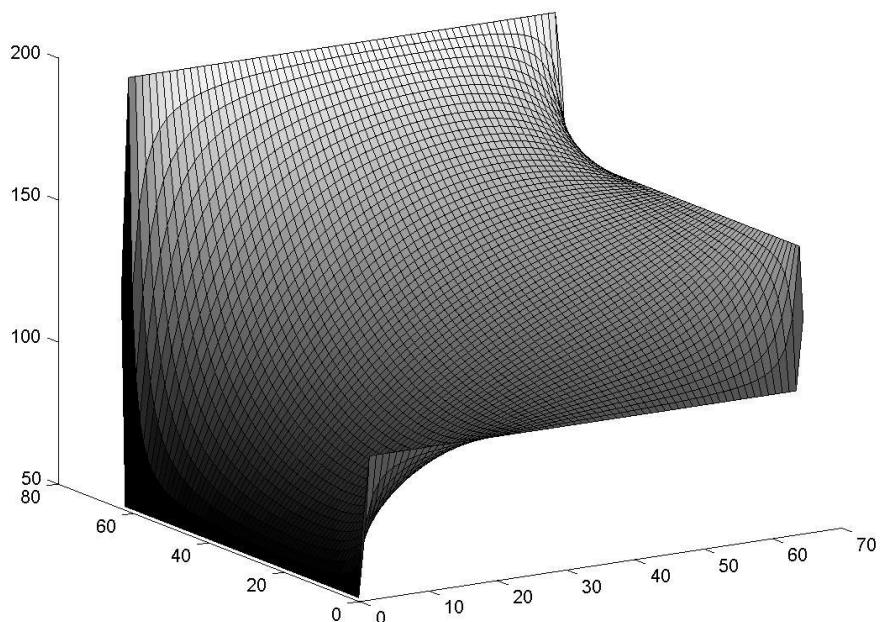
$$\Omega = \{(x, y) : 0 \leq x \leq 1 \text{ et } 0 \leq y \leq 1\}$$

Les conditions aux limites sont

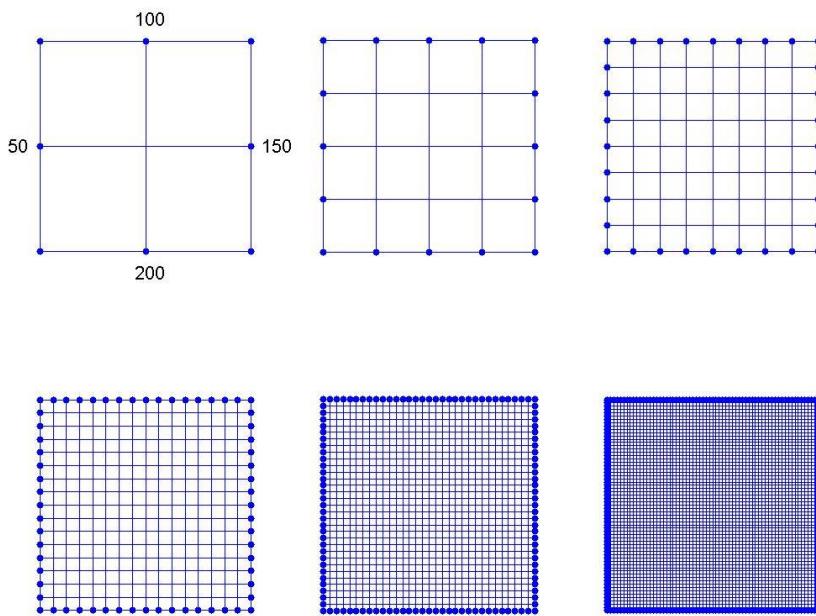
$$\begin{aligned} u = u_b &= 200 & \text{en } 0 < x < 1 \text{ et } y = 0 \\ u = u_h &= 100 & \text{en } 0 < x < 1 \text{ et } y = 1 \\ u = u_g &= 50 & \text{en } 0 < y < 1 \text{ et } x = 0 \\ u = u_d &= 150 & \text{en } 0 < y < 1 \text{ et } x = 1 \end{aligned}$$

On notera que ces conditions ignorent les sommets du carré : on peut supposer que  $u$  y est égal à la moyenne arithmétique des valeurs qu'il prend le long des côtés adjacents à chaque sommet ; cette précision est néanmoins superflue dans la mesure où la valeur de  $u$  en ces points n'apparaît nulle part dans la résolution du problème par les différences finies.

Avec de telles conditions, la solution de III-77 est représentée à la figure suivante. Ce résultat a été obtenu tant par les méthodes qui précèdent que par les méthodes multigrille décrites ci-après.



Les grilles spatiales utilisées ci-dessous dans les méthodes multigrille résultent du découpage du côté du carré en  $2^n$  divisions avec  $n = 1, 2, \dots, 6$ , chaque valeur de  $n$  correspondant à une grille différente (d'autant plus fine que  $n$  est grand).



Comme les conditions aux limites sont de type Dirichlet, il n'y a qu'en les nœuds intérieurs à  $\Omega$  (non renforcés sur la figure) que III-77 doit être discréteisé. Cela implique que le nombre d'inconnues  $n_{ic}$  (et donc la taille du système à résoudre) sera lié à  $n$  par la relation suivante :

$$n_{ic} = (2^n - 1)^2$$

III-78

ce qui donne pour les valeurs de n retenues

n	n <sub>ic</sub>
1	1
2	9
3	49
4	225
5	961
6	3969

La majeure partie des idées mises en œuvre dans les méthodes multigrille peuvent être décrites dans le cas simple de deux grilles, appelées ci-dessous « fine » et « grossière » par commodité. Supposons donc pour fixer les idées travailler avec les grilles relatives à  $n = 2$  et  $n = 3$  et affectons des indices « f » et « g » l'inconnue  $u$  selon qu'on se trouve sur la grille fine ou grossière. Rappelons encore la définition des grandeurs qui permettent de mesurer la convergence de la méthode itérative choisie :

$$\text{L'erreur sur l'itéré de rang } k : \quad \bar{e}^k = \bar{U} - \bar{u}^k \quad \text{III-52}$$

$$\text{Le résidu de rang } k : \quad \bar{r}^k = A\bar{u}^k - \bar{b} \quad \text{III-53}$$

Ces grandeurs sont liées : il vient facilement

$$A\bar{e}^k = A\bar{U} - A\bar{u}^k = \bar{b} - (\bar{r}^k + \bar{b}) = -\bar{r}^k \quad \text{III-79}$$

III-79 est importante : cette relation exprime qu'après avoir calculé  $\bar{u}^k$  par k itérations de la méthode retenue pour résoudre le système initial III-41, il suffit de calculer le résidu correspondant  $\bar{r}^k$  et de résoudre le système III-79 pour déduire la correction  $\bar{e}^k$  à apporter à  $\bar{u}^k$  pour obtenir la solution exacte  $\bar{U}$ .

### Méthode à deux grilles :

1° on procède à n itérations de la résolution de III-41 sur la grille fine (n de l'ordre de 3 ou 4 en général) ; la méthode utilisée est celle de Gauss-Seidel non relaxée (on observe que la relaxation n'apporte rien dans le contexte des méthodes multigrille) : soit  $\bar{u}_f^n$  la solution obtenue à partir par exemple de  $\bar{u}_f^0 = \bar{0}$ . Un test de convergence sur la solution permet à ce stade d'arrêter ou de poursuivre la procédure. Ce test consistera par exemple à vérifier si

$$\left\| \bar{u}_f^n - \bar{u}_f^{n-1} \right\| \leq \delta \quad \text{III-80}$$

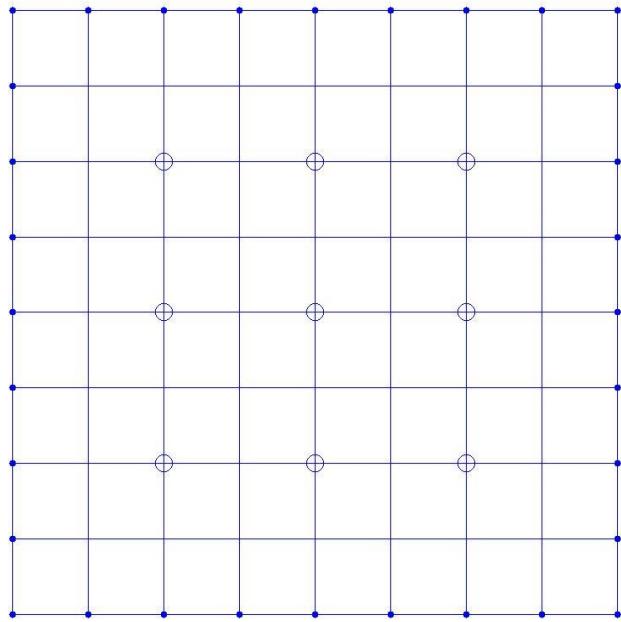
2° on calcule le résidu correspondant sur la grille fine :

$$\bar{r}_f^n = A\bar{u}_f^n - \bar{b} \quad \text{III-81}$$

3° on transfère  $\bar{r}_f^n$  sur la grille grossière

$$\bar{r}_g = \text{Trf}(\bar{r}_f^n) \quad \text{III-82}$$

L'opérateur de transfert  $\text{Tr}_f$  consiste à ne retenir que les éléments de  $\bar{r}_f^n$  relatifs aux nœuds de la grille grossière : la figure ci-dessous superpose les grilles relatives à  $n = 2$  et  $n = 3$  :



tous les nœuds intérieurs appartiennent à la grille fine, tandis que les nœuds marqués « o » sont ceux de la grille grossière ; c'est donc les valeurs prises par  $\bar{r}_f^n$  en ces nœuds qui constituent  $\bar{r}_g$

4° on résout III-79 sur la grille grossière :

$$A_g \bar{e}_g = -\bar{r}_g \quad \text{III-83}$$

par la méthode choisie (Gauss-Seidel ici) jusqu'à la convergence en démarrant les itérés avec  $\bar{e}_g^0 = \bar{0}$ .

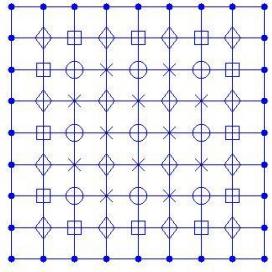
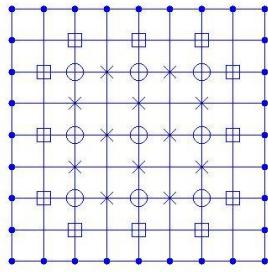
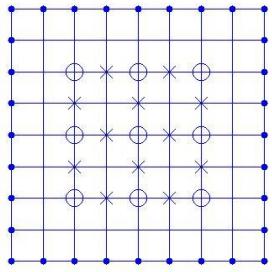
Soit  $\bar{e}_g^k$  la valeur trouvée.

5° on rapatrie  $\bar{e}_g^k$  sur la grille fine :

$$\bar{e}_f = \text{Rap}(\bar{e}_g^k) \quad \text{III-84}$$

L'opération est plus compliquée que celle de l'étape 3 et se fait en plusieurs passes :

- a) en les points de la grille fine compris entre deux points de grille grossière (nœuds x sur la figure ci-dessous), on prend la moyenne arithmétique des valeurs de  $\bar{e}_g^k$  aux nœuds voisins
- b) la même règle est ensuite appliquée en les nœuds (marqués □ sur la figure) compris entre un nœud de grille grossière et un nœud de bord pour lequel, à cause des conditions aux limites de Dirichlet on a  $\bar{e} = 0$
- c) en les points restants, la même règle de moyenne arithmétique s'applique en utilisant les valeurs produites dans les étapes a) et b) (nœuds ○).



6°  $\bar{e}_f$  est ajouté à la solution trouvée en 1° pour générer une nouvelle valeur initiale  $\bar{u}_{f,N}^0$  de la solution

$$\bar{u}_{f,N}^0 = \bar{u}_f^n + \bar{e}_f$$

III-85

A ce stade, un cycle complet a été effectué et on reprend à l'étape 1°.

#### **Méthode multigrille :**

1° identique à l'étape 1° de la procédure à 2 grilles

2° identique à l'étape 2° de la procédure à 2 grilles

3° on transfère  $\bar{r}_f^n$  sur la première grille intermédiaire (grille 1) de la même manière que ce qui est fait dans la méthode à 2 grilles :

$$\bar{r}_1 = \text{Trf}(\bar{r}_f^n)$$

III-86

4° on procède à n itérations de résolution de III-79 sur la grille 1 en démarrant les itérés avec  $\bar{e}_1^0 = \bar{0}$  :

$$A_1 \bar{e}_1 = -\bar{r}_1$$

III-87

Soit  $\bar{e}_1^n$  la correction obtenue. A ce stade on garde en mémoire  $\bar{r}_1$  et  $\bar{e}_1^n$  pour l'étape ultérieure de rapatriement (voir plus loin). On calcule une version corrigée du résidu sur la grille 1 :

$$\bar{r}_1^c = \bar{r}_1 + A_1 \bar{e}_1^n$$

III-88

5° on répète les étapes 3° et 4° ci-dessus pour toutes les grilles intermédiaires : si m est le rang de la dernière grille intermédiaire, les grandeurs mémorisées sont  $\bar{r}_m$  et  $\bar{e}_m^n$ , et la grandeur à transférer est  $\bar{r}_m^c$ .

6° on transfère  $\bar{r}_m^c$  sur la grille grossière

$$\bar{r}_g = \text{Trf}(\bar{r}_m^c) \quad \text{III-89}$$

et on procède de manière identique à ce qui est fait dans la méthode à 2 grilles : résolution de III-79 sur la grille grossière :

$$A_g \bar{e}_g = -\bar{r}_g \quad \text{III-90}$$

par la méthode choisie (Gauss-Seidel ici) jusqu'à la convergence et en démarrant les itérés avec  $\bar{e}_g^0 = \bar{0}$ . Soit  $\bar{e}_g^k$  la valeur trouvée.

7° on rapatrie  $\bar{e}_g^k$  sur la  $m^{\text{ième}}$  grille intermédiaire

$$\bar{e}_m = \text{Rap}(\bar{e}_g^k) \quad \text{III-91}$$

et on ajoute  $\bar{e}_m$  à  $\bar{e}_m^n$  pour donner un nouvel itéré de démarrage

$$\bar{e}_m^0 = \bar{e}_m + \bar{e}_m^n \quad \text{III-92}$$

à partir duquel on procède à n nouvelles itérations de résolution de III-79 sur la grille m

$$A_m \bar{e}_m = -\bar{r}_m \quad \text{III-93}$$

Appelons  $\bar{e}_m^{2n}$  la correction obtenue.

8° on rapatrie  $\bar{e}_m^{2n}$  sur la grille intermédiaire de rang m-1 et de proche en proche on répète l'étape 7° sur toutes les grilles intermédiaires jusqu'au calcul de  $\bar{e}_1^{2n}$  sur la première grille intermédiaire

9° on rapatrie  $\bar{e}_1^{2n}$  sur la grille fine

$$\bar{e}_f = \text{Rap}(\bar{e}_1^{2n}) \quad \text{III-94}$$

et on ajoute  $\bar{e}_f$  à la solution  $\bar{u}_f^n$  obtenue sur la grille fine pour donner une nouvelle valeur initiale  $\bar{u}_{f,N}^0$

$$\bar{u}_{f,N}^0 = \bar{u}_f^n + \bar{e}_f \quad \text{III-95}$$

pour ensuite reprendre à l'étape 1°.

### **Comparaison avec les méthodes de Gauss-Seidel simple et relaxée :**

Elle sera effectuée de la manière suivante : comme la méthode multigrille manipule des systèmes de taille variable, le moyen le plus simple de contrôle de l'efficacité des méthodes est de comptabiliser le nombre d'opérations arithmétiques à faire pour obtenir la même solution. Celle-ci, puisqu'il s'agit de méthodes itératives, résulte de la réalisation d'un test de convergence défini de la manière suivante : on arrête à l'itération  $\bar{u}_f^k$  dès que

$$\max_j |u_{f,j}^k - u_{f,j}^{k-1}| \leq \delta \quad \text{III-96}$$

où  $\delta$  a été arbitrairement fixé à 0.002 .

Le comptage des opérations arithmétiques a été réalisé de la manière suivante : il est aisément de calculer le nombre d'opérations  $n_{op}$  requises pour effectuer une itération de type Gauss-Seidel (relaxée ou non) : si  $n_{ic}$  désigne la taille du système à résoudre, on trouve

$$n_{op} = 8n_{ic} - 2\sqrt{n_{ic}} \quad \text{III-97}$$

Dans le cas des méthodes de Gauss-Seidel simple et relaxée,  $n_{ic}$  reste constant, égal à la taille du système à résoudre sur la grille fine, tandis qu'il est variable dans la méthode multigrille. Le tableau suivant compare en fonction de  $n_{ic}$  les performances des méthodes de Gauss-Seidel simple (GS), relaxée avec paramètre de relaxation optimal donné par III-71 (GSR), multigrille à 2 grilles (MG2) et multigrille complète (MGmax) ; la colonne GS donne le nombre d'opérations arithmétiques requises par cette méthode, tandis que les colonnes relatives aux autres méthodes donnent ces mêmes nombres en pourcents du nombre correspondant de la méthode GS.

n	$n_{ic}$	GS	GSR (%)	MG 2(%)	MG max(%)
2	9	1188	55.6	67.0	67.0
3	49	23814	33.3	34.6	32.1
4	225	382320	18.1	18.2	10.4
5	961	5475468	10.4	12.2	3.3
6	3969	72360288	6.2	10.5	1.1

Plusieurs constatations s'imposent :

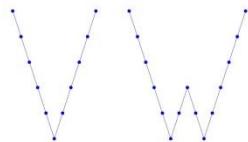
- l'efficacité de la relaxation (optimale ici) dans la méthode de Gauss-Seidel grandit avec la taille du système à résoudre ;
- la méthode multigrille à 2 grilles est moins performante que la méthode de Gauss-Seidel relaxée
- la méthode multigrille complète surpasse d'autant plus la méthode de Gauss-Seidel relaxée que la taille du système à résoudre est grande.

On peut mesurer l'impact du nombre de grilles prises en compte en modifiant à grille fine inchangée le nombre de grilles plus grossières : si on démarre à chaque fois de la grille relative à  $n_{ic} = 3969$ , les performances de la méthode multigrille en fonction du nombre total de grilles  $n_{TG}$  est décrite par la table ci-dessous :

$n_{TG}$	MG (%)
2	10.4
3	1.9
4	1.1
5	1.1
6	1.1

Il n'est donc pas forcément nécessaire de « descendre » dans l'algorithme jusqu'à la grille la plus grossière (qui correspond ici à un système linéaire de dimension égale à un) pour bénéficier de toute la puissance de la méthode.

Enfin, diverses stratégies de passages d'une grille à l'autre peuvent aussi être élaborées : dans ce qui précède, on s'est contenté parcourir toutes les grilles à chaque cycle, de la plus fine à la plus grossière et vice versa : on parle dans ce cas de cycle en V. On peut aussi imaginer des cycles en W tels que représentés sur la figure ci-dessous.



## Chapitre IV. Equations hyperboliques

### IV.1 Généralités

L'archétype de cette catégorie d'EDP est l'équation d'advection I-14

$$u_t + au_x = 0 \quad \text{avec } a = \text{cte} \quad \text{IV-1}$$

$$\text{avec la condition initiale} \quad u(x,0) = u^0(x) \quad \text{IV-2}$$

$$\text{et la condition aux limites} \quad u(0,t) = f(t) \quad \text{IV-3}$$

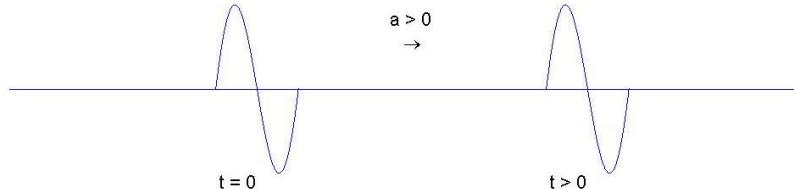
(remarquons au passage que ces conditions doivent être compatibles : il faut  $u^0(0) = f(0)$  ).

Il est aisément vérifier que la solution analytique de ce problème est

$$u(x,t) = u^0(x - at) \quad \text{IV-4}$$

Ce résultat nous apporte deux renseignements fondamentaux.

Premièrement, la solution à tout instant  $t$  est une copie de la condition initiale décalée vers la droite, si  $a$  est positif ou vers la gauche si  $a$  est négatif



Une autre manière d'énoncer cette propriété est de dire que la solution ne dépend que de la valeur de  $\xi = x - at$ . Les lignes dans le plan  $(x,t)$  le long desquelles  $x - at$  est constant sont appelées lignes caractéristiques. Le paramètre  $a$ , dont la dimension est celle d'une distance divisée par un temps est appelé vitesse de propagation le long de la caractéristique. La solution IV-4 peut donc être interprétée comme une onde se propageant à la vitesse  $a$  sans changer de forme.

Deuxièmement, alors que IV-1 ne semble avoir de sens que si  $u$  est différentiable, la solution IV-4 ne requiert aucunement cette propriété. Ceci explique qu'on accepte des solutions discontinues pour les problèmes hyperboliques telles les ondes de choc.

Le concept de lignes caractéristiques, fondamental pour les équations hyperboliques s'étend aisément à toute équation de la forme

$$u_t + au_x = f(x,t,u) \quad \text{IV-5}$$

et même à tout système

$$\bar{u}_t + A(x, t)\bar{u}_x + B(x, t)\bar{u} = \bar{F}(x, t) \quad IV-6$$

Un tel système n'est que la généralisation de I-17

$$\bar{u}_t + A\bar{u}_x = \bar{0} \quad (\text{où } A = \begin{bmatrix} 0 & -c \\ -c & 0 \end{bmatrix}) \quad I-17$$

On a vu au chapitre I que I-17 était un système hyperbolique pour autant que les valeurs propres de A soient réelles. Cette condition vaut aussi pour IV-6, et si les valeurs propres de  $A(x, t)$  sont notées  $a_1(x, t), \dots, a_N(x, t)$ , on montre alors que les lignes caractéristiques sont les solutions des équations différentielles

$$\frac{dx^i}{dt} = a_i(x, t) \quad \text{avec} \quad x^i(0) \text{ donné} \quad IV-7$$

Il en résulte par exemple que les lignes caractéristiques de I-17 ont pour équation

$$\frac{dx}{dt} = \pm c \quad IV-8$$

c'est-à-dire

$$\begin{aligned} x - x_0 &= +c(t - t_0) \\ x - x_0 &= -c(t - t_0) \end{aligned} \quad IV-9$$

Ce sont donc deux familles de droites parallèles de pentes égales à  $+c$  et  $-c$ .

## IV.2 Résolution numérique de l'équation d'advection

$$\text{Reprendons} \quad u_t + au_x = 0 \quad IV-1$$

$$\text{avec la condition initiale} \quad u(x, 0) = u^0(x) \quad IV-2$$

$$\text{et la condition aux limites} \quad u(0, t) = f(t) \quad IV-3$$

Appliquons les différentes étapes de la méthode des différences finies :

1. Discrétisation des variables dépendantes : comme  $\Omega = \{(x, t) : 0 < x \text{ et } 0 < t\}$ , il vient

$$x_i = i\Delta x \quad \text{et} \quad t^k = k\Delta t \quad IV-10$$

2. Ecriture de IV-1, 2 et 3 en chacun des nœuds du maillage défini par la discrétisation précédente :

$$\begin{aligned} (u_t)_i^k + a(u_x)_i^k &= 0 \\ u(x_i, 0) &= u_i^0 \\ u(0, t^k) &= f^k = u_0^k \end{aligned} \quad IV-11$$

3. Remplacement des dérivées par les estimations qu'en donnent les formules de différences finies,  
Pour fixer les idées sur la structure de la résolution, utilisons respectivement les formules II-9 et II-10  
pour discréteriser les dérivées temporelle et spatiale :

$$\text{II-9 : } u_i^{(I)} = \frac{u_{i+1} - u_i}{h} \Rightarrow (u_t)_i^k = \frac{u_i^{k+1} - u_i^k}{\Delta t} \quad \text{IV-12}$$

$$\text{II-10 : } u_i^{(I)} = \frac{u_i - u_{i-1}}{h} \Rightarrow (u_x)_i^k = \frac{u_i^k - u_{i-1}^k}{\Delta x} \quad \text{IV-13}$$

4. Résolution du système linéaire qui en résulte : IV-12 et IV-13 dans IV-11 conduisent à exprimer la solution à l'instant  $k+1$  à partir de la solution à l'instant  $k$  :

$$u_i^{k+1} = u_i^k - a \frac{\Delta t}{\Delta x} (u_i^k - u_{i-1}^k) \quad \text{IV-14}$$

IV-14 et les deux dernières relations de IV-11 nous permettent de résoudre totalement le problème posé : le vecteur des inconnues  $\bar{u}^{k+1} = [u_1^{k+1} \ u_2^{k+1} \ \dots \ u_N^{k+1}]^T$  est calculable à partir de la relation matricielle

$$\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_N \end{bmatrix}^{k+1} = \begin{bmatrix} 1-v & 0 & & & 0 \\ v & 1-v & 0 & & 0 \\ 0 & v & 1-v & 0 & 0 \\ & & \ddots & & \vdots \\ & & & \ddots & \vdots \\ 0 & & & 0 & v \\ & & & v & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_N \end{bmatrix}^k + \begin{bmatrix} vu_0^k \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix} \quad \text{IV-15}$$

si on pose  $v = \frac{a\Delta t}{\Delta x}$ . La résolution est donc itérative :  $\bar{u}^0$  est connu (deuxième relation de IV-11) ; son injection dans IV-15 permet de calculer  $\bar{u}^1$  qui, réinjecté ensuite dans IV-15 donnera  $\bar{u}^2$  et ainsi de suite. Ce caractère itératif qui était absent de la résolution des équations elliptiques, se retrouvera aussi dans la résolution des équations paraboliques.

Remarquons que la numérotation « spatiale » vaut pour  $i = 1, 2, \dots, N$  et non pour  $i = 0, 1, \dots, N$  : la solution en le nœud 0 à l'extrême gauche du domaine spatial est connue par la condition aux limites (troisième relation de IV-11) et intervient à chaque utilisation de IV-15.

### IV.3 Exemple

On se propose de résoudre le problème IV-1, 2, 3 dans les hypothèses suivantes :

$$0 \leq x \leq 5 \quad 0 \leq t \leq 3 \quad a = 1 \quad \text{IV-16}$$

$$u^0(x) = 100 \quad x \in [0, 0.25[ \cup ]0.5, 5] \quad \text{IV-17}$$

$$110 \quad x \in [0.25, 0.5] \quad \text{IV-17}$$

$$u_0^k = 100 \quad \text{IV-18}$$

La figure suivante représente la condition initiale IV-17 ainsi que la solution obtenue aux instants

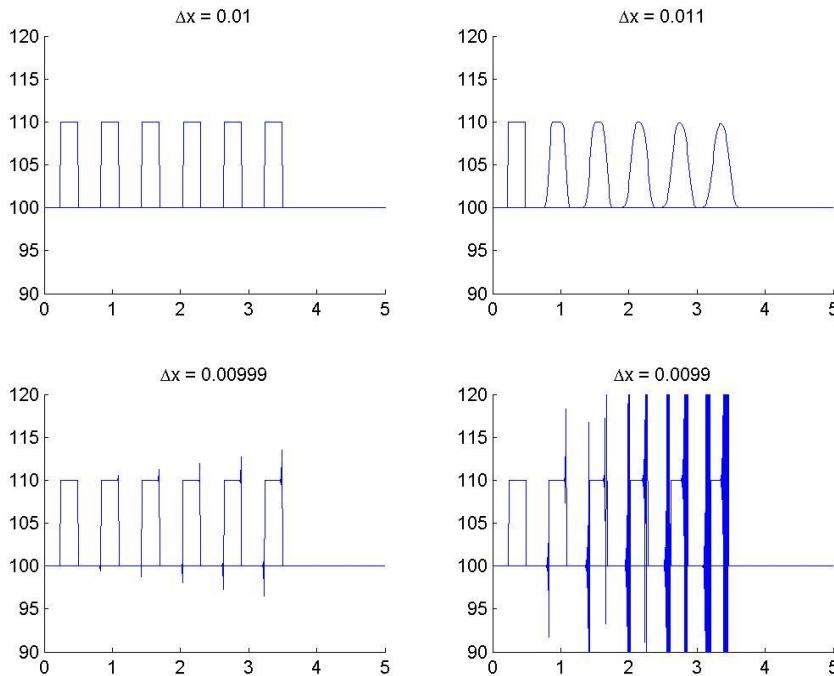
$$t_j = j \Delta T \quad j=1, \dots, 5 \quad \Delta T = 0.6$$

IV-19

pour quatre maillages :

$$\Delta t = 0.01 \quad \text{et} \quad \Delta x = 0.01, 0.011, 0.00999, 0.0099$$

IV-20



Commentaires :

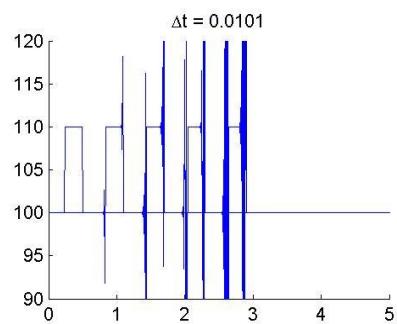
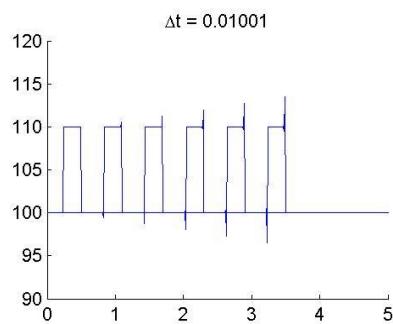
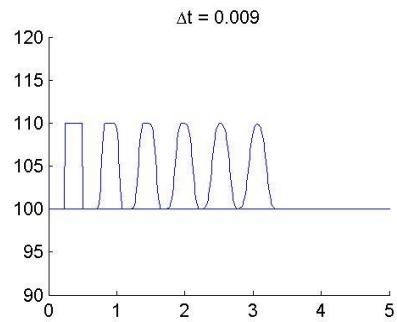
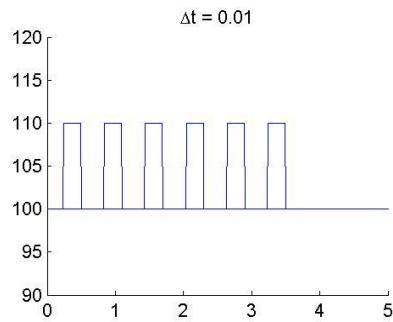
Sachant que la solution théorique de IV-1 est la condition initiale qui se déplace vers la droite quand  $t$  grandit, il apparaît que la seule solution conforme à la théorie est obtenue avec  $\Delta x = 0.01$ .

Quand  $\Delta x$  grandit ( $\Delta x = 0.011$ ), la solution se dégrade par estompements des points anguleux et élargissement de la zone occupée. Il fallait s'attendre à cette dégradation : le maillage étant moins fin, l'erreur de troncature sur le schéma de différences finies II-10 augmente.

Quand  $\Delta x$  diminue ( $\Delta x = 0.0099$  et  $0.00999$ ), la solution se dégrade aussi mais d'une tout autre manière : des oscillations apparaissent au droit des points anguleux et elles sont d'autant plus importantes qu'on réduit  $\Delta x$ . Ce résultat est plus surprenant, si on observe que dans ce cas, l'erreur de troncature sur II-10 diminue. Observons encore que cette dégradation est extrêmement sensible à la réduction de  $\Delta x$ .

De la même manière (figure suivante), de légères modifications de  $\Delta t$  autour de la valeur initiale 0.01 engendre des dégradations de la solution inverses de celles du cas précédent.

La compréhension de ces phénomènes sera établie plus loin.



## Chapitre V. Equations paraboliques

### V.1 Généralités

L'archétype de cette catégorie d'équations est la 2<sup>e</sup> loi de Fourier appelée aussi équation de la diffusion :

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2} \quad \text{avec} \quad 0 < x < 1. \quad \text{V-1}$$

Il est possible en précisant judicieusement la condition initiale et les conditions aux limites de résoudre analytiquement cette équation : en prenant comme

$$\text{condition initiale : } u(x, 0) = 0 \quad \text{V-2}$$

$$\text{conditions aux limites : } u(0, t) = 0 \quad \text{et} \quad u(1, t) = U_0, \quad \text{V-3}$$

la solution est

$$u(x, t) = U_0 x + \sum_{n=1}^{\infty} \frac{2U_0(-1)^n}{n\pi} \exp(-n^2\pi^2\alpha t) \sin(n\pi x) \quad \text{V-4}$$

### V.2 Résolution numérique de l'équation de la diffusion

La structure de la résolution est identique à celle des équations hyperboliques ; prenons par exemple II-9 pour évaluer la dérivée temporelle et II-12 pour la dérivée spatiale. Cela donne

$$\frac{u_i^{k+1} - u_i^k}{\Delta t} = \alpha \frac{u_{i+1}^k - 2u_i^k + u_{i-1}^k}{(\Delta x)^2} \quad \text{V-5}$$

On déduit alors la solution à l'instant k+1 à partir de la solution à l'instant k :

$$u_i^{k+1} = u_i^k + \alpha \frac{\Delta t}{(\Delta x)^2} (u_{i+1}^k - 2u_i^k + u_{i-1}^k) \quad \text{V-6}$$

ou encore sous forme matricielle :

$$\begin{bmatrix} u_1 \\ u_2 \\ .. \\ .. \\ u_{N-1} \end{bmatrix}^{k+1} = \begin{bmatrix} 1-2r & r & & & \\ r & 1-2r & r & & \\ & r & 1-2r & r & \\ & & r & 1-2r & \\ \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ .. \\ .. \\ u_{N-1} \end{bmatrix}^k + r \begin{bmatrix} u_0^k \\ 0 \\ .. \\ 0 \\ u_N^k \end{bmatrix} \quad \text{V-7}$$

$$\text{avec (cf. V-2 et V-3) : } u_i^0 = 0 \quad \forall i, \quad u_0^k = 0 \quad \forall k, \quad u_N^k = U_0 \quad \forall k \quad \text{et} \quad r = \alpha \frac{\Delta t}{(\Delta x)^2} \quad \text{V-8}$$

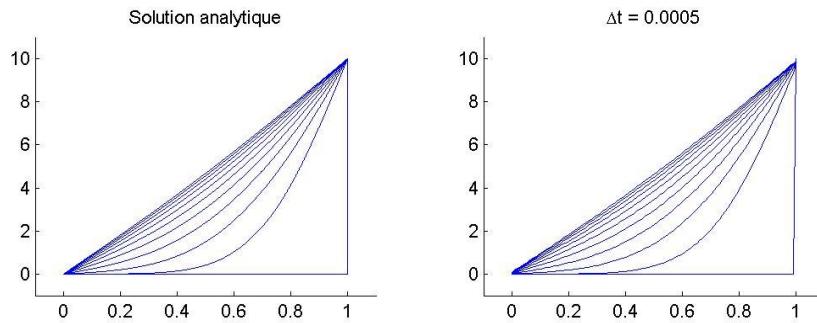
### V.3 Exemple

Les relations V-1 à V-3 décrivent le problème suivant : une paroi d'épaisseur égale à 1 et de dimensions infinies est initialement à une température de 0 degré. Cette paroi sépare deux milieux : celui de gauche ( $x \leq 0$ ) est à zéro degré et celui de droite à la température  $U_0$ . On demande d'évaluer en fonction du temps le profil spatial de la température dans la paroi. La figure suivante compare la solution analytique V-4 aux résultats numériques donnés par V-6 dans les conditions suivantes :

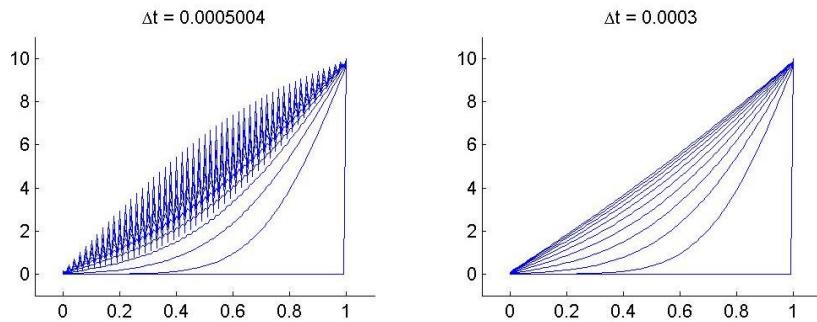
$$\Delta x = 0.01 \quad \Delta t = 0.0005 \quad U_0 = 10 \quad \alpha = 0.1 \quad \text{V-9}$$

La solution est visualisée aux instants

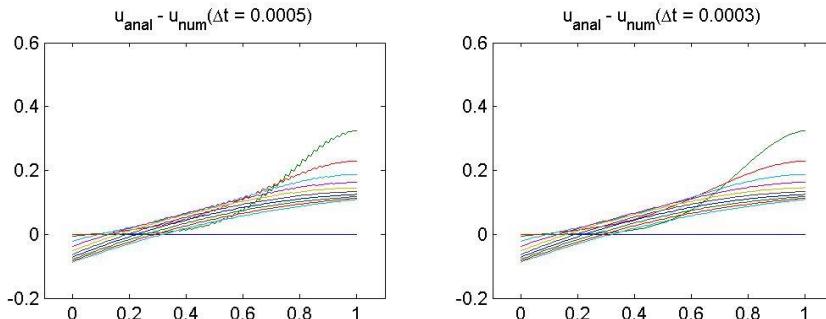
$$t^k = k\Delta T \quad \Delta T = 0.3 \quad k = 0, 1, \dots, 10 \quad \text{V-10}$$



Visuellement, les deux solutions sont identiques, validant ainsi la résolution numérique. Néanmoins, si  $\Delta t$  subit des modifications, la qualité du résultat peut être très dégradée :



Une comparaison plus fine des résultats est possible en visualisant l'écart entre la solution analytique et les solutions relatives à  $\Delta t = 0.0005$  et  $\Delta t = 0.0003$  :



Elle montre que les erreurs commises sont comparables avec cependant de légères oscillations sur le graphe relatif à  $\Delta t = 0.0005$  , préfigurant ainsi ce qu'on observe quand  $\Delta t = 0.0005004$  . À nouveau, ces résultats seront expliqués plus loin.

## Chapitre VI. Analyse des performances numériques

Comme on a pu l'observer dans les chapitres précédents, la résolution numérique des équations paraboliques et hyperboliques peut être extrêmement sensible aux valeurs numériques des pas de discréttisation  $\Delta t$  et  $\Delta x$ . Les développements qui suivent ont pour but d'apporter une réponse aux phénomènes d'instabilité précédemment observés, et plus généralement d'étudier les performances des méthodes de résolution des équations hyperboliques et paraboliques. Ces développements utilisent des concepts classiques de l'analyse numérique.

### VI.1 Erreur de troncature

Cette erreur provient du remplacement des dérivées par les schémas de différences finies. Prenons par exemple la 2<sup>e</sup> loi de Fourier I-5 :

$$\frac{\partial u}{\partial t} - \alpha \frac{\partial^2 u}{\partial x^2} = 0 \quad \text{I-5}$$

remplaçons-y la dérivée temporelle par II-9 et la dérivée spatiale par II-11 : il vient

$$\frac{\partial u}{\partial t} - \alpha \frac{\partial^2 u}{\partial x^2} = \left[ \frac{u_i^{k+1} - u_i^k}{\Delta t} - \alpha \frac{u_{i+1}^k - 2u_i^k + u_{i-1}^k}{(\Delta x)^2} \right] + \left[ -\frac{\Delta t}{2!} \left( \frac{\partial^2 u}{\partial t^2} \right)_i^k - \dots + \alpha \frac{(\Delta x)^2}{12} \left( \frac{\partial^4 u}{\partial x^4} \right)_i^k + \dots \right] = 0 \quad \text{VI-1}$$

Le dernier terme du second membre de VI-1 est l'erreur de troncature  $\varepsilon_T$  ; elle représente l'écart entre l'EDP et l'équation dite aux différences finies EDF :  $\varepsilon_T = \text{EDP} - \text{EDF}$  :

$$\varepsilon_T = \left[ -\frac{\Delta t}{2!} \left( \frac{\partial^2 u}{\partial t^2} \right)_i^k - \dots + \alpha \frac{(\Delta x)^2}{12} \left( \frac{\partial^4 u}{\partial x^4} \right)_i^k + \dots \right] = \left[ \frac{\partial u}{\partial t} - \alpha \frac{\partial^2 u}{\partial x^2} \right] - \left[ \frac{u_i^{k+1} - u_i^k}{\Delta t} - \alpha \frac{u_{i+1}^k - 2u_i^k + u_{i-1}^k}{(\Delta x)^2} \right]$$

L'EDF sera d'autant plus précise que les puissances de  $\Delta t$  et  $\Delta x$  qui sont présentes dans  $\varepsilon_T$  seront élevées. On parle alors d'ordre de la méthode qui ici est  $O(\Delta t) + O[(\Delta x)^2]$ , plus fréquemment écrit  $O[\Delta t, (\Delta x)^2]$ .

### VI.2 Consistance

Un schéma de différences finies est consistant avec l'EDP qu'il représente si on a

$$\lim_{\text{grille} \rightarrow 0} (\varepsilon_T) = 0 \quad \text{VI-2}$$

C'est le cas de VI-1 ; on verra des contre-exemples plus loin.

### VI.3 Stabilité : analyse de Von Neumann

Dans le cas des problèmes évolutifs (EDP paraboliques et hyperboliques), un schéma numérique est stable si toute erreur d'arrondi ne croît pas avec le calcul des pas successifs induits par le caractère

évolutif du problème. Cette propriété indispensable est beaucoup plus difficile à établir que la consistance. Reprenons l'exemple de la 2<sup>e</sup> loi de Fourier et intéressons-nous à l'EDF:

$$\frac{u_i^{k+1} - u_i^k}{\Delta t} - \alpha \frac{u_{i+1}^k - 2u_i^k + u_{i-1}^k}{(\Delta x)^2} = 0 \quad VI-3$$

Appelons A la solution analytique de l'équation I-5, D la solution exacte de VI-3 (c'est celle qu'on obtiendrait avec un ordinateur de précision infinie), et N la solution réellement obtenue avec une machine de précision finie. On a

$$\varepsilon_T = A - D$$

et, si  $\varepsilon_A$  désigne l'erreur d'arrondi,

$$\varepsilon_A = D - N$$

Le concept de stabilité se résume alors à examiner deux situations :

a) l'erreur d'arrondi globale (c'est-à-dire celle qu'on a accumulée) au pas  $k\Delta t$

grandit avec  $k \rightarrow$  instabilité forte  
ne grandit pas avec  $k \rightarrow$  stabilité forte

b) l'erreur d'arrondi commise au pas  $k\Delta t$

grandit avec  $k \rightarrow$  instabilité faible  
ne grandit pas avec  $k \rightarrow$  stabilité faible

La stabilité forte est évidemment le but à atteindre. Malheureusement il est impossible d'en établir les conditions. On se contentera donc de la recherche des conditions de la stabilité faible : ceci est l'objet de l'analyse de Von Neumann. On admet alors communément que la preuve de la stabilité faible implique la stabilité forte.

### **Décomposition en série de Fourier de l'erreur**

Exprimons que D est la solution exacte de VI-3 :

$$\frac{D_i^{k+1} - D_i^k}{\Delta t} = \alpha \frac{D_{i+1}^k - 2D_i^k + D_{i-1}^k}{(\Delta x)^2} \quad VI-4$$

Mais N étant la solution réellement calculée, N obéit aussi à VI-4 ; il en résulte, puisque  $\varepsilon_A = D - N$ , que  $\varepsilon_A$  lui aussi vérifie la même relation (dans la suite on notera  $\varepsilon$  à la place de  $\varepsilon_A$  pour alléger les notations) :

$$\frac{\varepsilon_i^{k+1} - \varepsilon_i^k}{\Delta t} = \alpha \frac{\varepsilon_{i+1}^k - 2\varepsilon_i^k + \varepsilon_{i-1}^k}{(\Delta x)^2} \quad VI-5$$

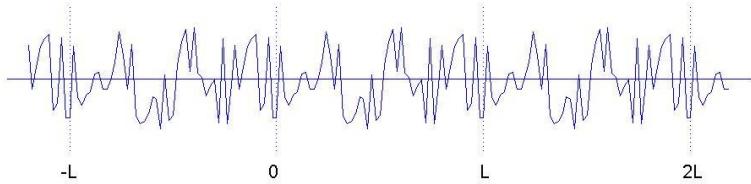
La solution D et l'erreur  $\varepsilon$  satisfont la même équation : ceci nous permettra de déterminer à la fois les conditions de stabilité d'une méthode et, celle-ci étant acquise, les performances de cette même méthode. Voyons d'abord la stabilité.

L'étude de Von Neumann repose sur les considérations suivantes :

- le domaine spatial est monodimensionnel :  $\Omega(x) = [0, L]$
- les conditions aux limites du problème étudié sont périodiques :

$$u(0, t) = u(L, t) \Rightarrow u_0^k = u_N^k \Rightarrow \varepsilon_0^k = \varepsilon_N^k \quad VI-6$$

- à un instant  $k\Delta t$  quelconque, on imagine que le signal  $\varepsilon$  est une fonction continue sur  $\Omega$  (en réalité,  $\varepsilon$  est constitué d'un ensemble de valeurs discrètes localisées en  $i\Delta x$ ,  $i = 0, \dots, N$  avec  $N = \frac{L}{\Delta x}$ ) qu'on reproduit infiniment vers la gauche et vers la droite de  $[0, L]$  créant ainsi un signal périodique continu grâce à VI-6



Un tel signal admet un développement en série de Fourier :

$$\varepsilon(x) = \sum_{m=-\infty}^{+\infty} A_m \exp(j\omega_m x) \quad VI-7$$

qu'il faut corriger pour tenir compte du caractère discret de  $\varepsilon$  : la correction essentielle consiste à limiter le nombre de pulsations  $\omega_m$  présentes dans VI-7 : on peut montrer qu'on a

$$\varepsilon(x) = \sum_{m=-N}^{+N} A_m \exp(j\omega_m x) \quad \text{avec} \quad \omega_m = m \frac{\pi}{L} = m \frac{\pi}{N\Delta x} \quad VI-8$$

Précisons encore que VI-8 est le développement de Fourier de  $\varepsilon$  à l'instant  $k\Delta t$  et que sa valeur en  $i\Delta x$  est  $\varepsilon_i^k = \varepsilon^k(i\Delta x)$  :

$$\varepsilon_i^k = \sum_{m=-N}^{+N} A_m^k \exp(jm \frac{\pi}{N\Delta x} i\Delta x) = \sum_{m=-N}^{+N} A_m^k \exp(jm \frac{i\pi}{N}) \quad VI-9$$

posons  $\phi_m = m \frac{\pi}{N}$  et retenons pour la suite que  $m$  variant de  $-N$  à  $+N$ ,  $\phi_m$  est compris entre  $-\pi$  et  $+\pi$ . Il vient enfin

$$\varepsilon_i^k = \sum_{m=-N}^{+N} A_m^k \exp[j(i\phi_m)] \quad VI-10$$

### Facteur d'amplification

VI-10 va nous permettre de décrire comment se comporte  $\varepsilon_i^k = \varepsilon(i\Delta x, k\Delta t)$  lorsque  $k$  grandit : ce comportement est fixé par VI-5 ; c'est donc aussi celui de tout terme  $A_m^k \exp[j(i\phi_m)]$  :

$$\frac{A_m^{k+1} e^{[j(i\phi_m)]} - A_m^k e^{[j(i\phi_m)]}}{\Delta t} = \frac{\alpha}{(\Delta x)^2} [A_m^k e^{[j((i+1)\phi_m)]} - 2A_m^k e^{[j(i\phi_m)]} + A_m^k e^{[j((i-1)\phi_m)]}] \quad VI-11$$

ou encore, en posant  $r = \alpha \frac{\Delta t}{(\Delta x)^2}$  et en regroupant ce qui a trait aux termes définis en  $k\Delta t$  et en  $(k+1)\Delta t$ ,

$$A_m^{k+1} e^{[j(i\phi_m)]} = [1 + r(e^{j\phi_m} - 2 + e^{-j\phi_m})] A_m^k e^{[j(i\phi_m)]} = [1 + 2r(\cos \phi_m - 1)] A_m^k e^{[j(i\phi_m)]}$$

Ceci montre que lorsqu'on passe de l'instant  $k\Delta t$  à l'instant  $(k+1)\Delta t$ , le  $m^{\text{ième}}$  terme de VI-10 est multiplié par

$$G = \frac{A_m^{k+1} e^{[j(i\phi_m)]}}{A_m^k e^{[j(i\phi_m)]}} = 1 + 2r(\cos \phi_m - 1) = 1 - 4r \sin^2 \frac{\phi_m}{2} \quad VI-12$$

VI-12 est appelé facteur d'amplification : c'est lui qui conditionne la manière avec laquelle la valeur en  $i\Delta x$  de l'harmonique de rang  $m$  de l'erreur se propage au fil du temps. Remarquons que l'influence de l'abscisse  $i\Delta x$  est contenue dans le facteur  $\exp[j(i\phi_m)]$  ; pour l'exemple ici traité, cette influence disparaît par la simplification haut et bas de ce facteur ; ce ne sera pas toujours le cas. Il en résulte ici que  $G$  est réel et égal au rapport des amplitudes de l'harmonique de rang  $m$  du signal  $\varepsilon$  aux instants  $(k+1)\Delta t$  et  $k\Delta t$ . La condition de stabilité est donc que ce facteur soit en valeur absolue inférieur ou égal à 1, pour tout harmonique  $m$  :

$$|G| \leq 1 \quad VI-13$$

c'est-à-dire, pour l'exemple traité ici :

$$\left| 1 - 4r \sin^2 \frac{\phi_m}{2} \right| \leq 1 \quad \text{c.a.d.} \quad -1 \leq 1 - 4r \sin^2 \frac{\phi_m}{2} \leq 1$$

ou encore

$$0 \leq 4r \sin^2 \frac{\phi_m}{2} \leq 2.$$

Comme  $-\pi < \phi_m < +\pi$ , il en résulte que VI-3 est vérifié  $\forall m$  si

$$0 \leq r \leq \frac{1}{2} \quad VI-14$$

VI-14 est la condition de stabilité du schéma numérique VI-3 qui est dit conditionnellement stable. Cette notion de stabilité est fondamentale car elle montre qu'on ne peut choisir  $\Delta x$  et  $\Delta t$  de manière indépendante : on doit respecter

$$0 \leq \alpha \frac{\Delta t}{(\Delta x)^2} \leq \frac{1}{2}$$

La forme même de cette relation mérite qu'on s'y attarde : elle montre que l'idée intuitive qui consiste à dire « si je travaille avec une discrétisation spatiale plus fine, je peux probablement me contenter de prendre des pas de temps plus grands » est erronée : plus  $\Delta x$  est petit, plus  $\Delta t$  doit être petit lui aussi ! On verra que ce type de limitation est général pour toute méthode conditionnellement stable.

Le caractère impératif de cette limitation peut aussi être apprécié en reprenant l'exemple traité au chapitre V : on y avait

$$\Delta x = 0.01 \quad \Delta t = 0.0005 \quad \alpha = 0.1$$

c'est-à-dire

$$r = \alpha \frac{\Delta t}{(\Delta x)^2} = 0.1 \frac{0.0005}{0.0001} = \frac{1}{2}$$

On se trouve dans ce cas à la limite du comportement stable ; un léger dépassement de cette limite ( $\Delta t = 0.0005004 \Rightarrow r = 0.5004$ ) génère de l'instabilité

### ***Calcul pratique du facteur d'amplification***

La technique est simple : il suffit de comparer le schéma numérique à traiter VI-3 avec VI-11 qui donne naissance au facteur d'amplification : en oubliant l'indice « m » pour simplifier les notations, on observe que le passage de l'indice temporel  $k$  à  $k+1$  dans VI-3 génère le passage de  $A^k$  à  $A^{k+1}$  tandis celui de l'indice spatial  $i$  à  $i+1$  ou  $i-1$  génère l'intervention de l'opérateur  $\exp[j\phi]$  ou  $\exp[-j\phi]$ . Notons qu'il est a priori impossible à ce stade de traiter des schémas où l'indice  $k$  varie de plus d'une unité. On verra comment s'y prendre sur des exemples précis.

L'exemple traité ci-avant est celui d'un schéma conditionnellement stable. A cet égard, deux autres situations sont possibles : un schéma peut être inconditionnellement stable, ou inconditionnellement instable. En voici des exemples.

- méthode d'Euler implicite avec discrétisateur spatial centré appliquée à l'équation d'advection  $u_t + au_x = 0$  : cette méthode consiste à utiliser les différences finies suivantes :

$$(u_t)_i^k = \frac{u_{i+1}^{k+1} - u_i^k}{\Delta t} \quad \text{et} \quad (u_x)_i^k = \frac{u_{i+1}^{k+1} - u_{i-1}^{k+1}}{2\Delta x} \quad \text{VI-15}$$

Appliquons les règles de calcul du facteur d'amplification :

$$\frac{A^{k+1}e^{j\phi} - A^k e^{j\phi}}{\Delta t} + a \frac{A^{k+1}e^{j(i+1)\phi} - A^{k+1}e^{j(i-1)\phi}}{2\Delta x} = 0 \quad \text{VI-16}$$

$$\Rightarrow A^{k+1} \left( 1 + \frac{v}{2} (e^{j\phi} - e^{-j\phi}) \right) = A^k \quad \text{VI-16}$$

$$\text{si} \quad v = \frac{a\Delta t}{\Delta x} \quad \text{VI-17}$$

$$\Rightarrow G = \frac{1}{1 + \frac{v}{2} (e^{j\phi} - e^{-j\phi})} = \frac{1}{1 + jv \sin(\phi)} \quad VI-18$$

$G$  est cette fois complexe, et le schéma est inconditionnellement stable car

$$|G| = \sqrt{\frac{1}{1 + v^2 \sin^2(\phi)}} \quad VI-19$$

qui est toujours inférieur ou égal à un quel que soit  $\phi$ .

- méthode d'Euler explicite avec discréteur spatial centré appliquée à l'équation d'advection  $u_t + au_x = 0$  : la seule différence avec ce qui précède consiste à remplacer VI-15 par

$$(u_t)_i^k = \frac{u_i^{k+1} - u_i^k}{\Delta t} \quad \text{et} \quad (u_x)_i^k = \frac{u_{i+1}^k - u_{i-1}^k}{2\Delta x} \quad VI-20$$

Un calcul analogue à ce qui précède donne cette fois

$$|G| = \sqrt{1 + v^2 \sin^2(\phi)} \quad VI-21$$

qui est toujours supérieur ou égal à un quel que soit  $\phi$  : le schéma est inconditionnellement instable.

#### VI.4 Analyse spectrale des erreurs

Outre les renseignements qu'il fournit sur la stabilité du schéma numérique, le facteur d'amplification  $G$  fournit des renseignements importants sur la qualité du schéma, quand il est stable. L'idée utilisée est la suivante : partant du développement en série de Fourier de la CI :

$$u(x,0) = \sum_{m=-N}^N E_m^0 \exp(j\omega_m x), \quad VI-22$$

on compare l'évolution subie au cours du temps par un harmonique quelconque  $E_m^0 \exp(j\omega_m x)$  quand il « traverse deux filtres mathématiques » :

a) le schéma numérique utilisé.

b) l'EDP proprement dite,

La comparaison des « sorties » de ces deux filtres devrait nous permettre d'évaluer la qualité du schéma numérique.

La sortie du a) est connue : par VI-12, on a après un pas de temps  $\Delta t$

$$E_m^1 \exp[j(i\phi_m)] = G E_m^0 \exp[j(i\phi_m)] \quad VI-23$$

où en toute généralité (voir plus haut)  $G$  est complexe

$$G = |G| \exp(j\theta) \quad VI-24$$

et convenant d'indiquer « num » la « sortie » du schéma numérique, il vient

$$(E_m^1)_{\text{num}} = [G | \exp(j\theta)] E_m^0 \quad \text{VI-25}$$

La sortie du b) va dépendre de l'EDP : détaillons le calcul pour les équations de base parabolique et hyperbolique :

**Modèle parabolique** :  $u_t = \alpha u_{xx}$

Introduisons-y  $E_m^0 \exp(j\omega_m x)$  ou, plus exactement  $E_m(t) \exp(j\omega_m x)$  : on a

$$\begin{aligned} u_t &= (E_m)_t \exp(j\omega_m x) \\ u_{xx} &= E_m (j\omega_m)^2 \exp(j\omega_m x) \end{aligned}$$

c'est-à-dire en remplaçant dans l'EDP :

$$(E_m)_t = -\alpha(\omega_m)^2 E_m$$

dont la solution est

$$E_m(t) = E_m(0) \exp[-\alpha(\omega_m)^2 t] = E_m^0 \exp[-\alpha(\omega_m)^2 t] \quad \text{VI-26}$$

Après un pas de temps  $\Delta t$ , cela donne  $E_m^1 = E_m^0 \exp[-\alpha(m \frac{\pi}{N\Delta x})^2 \Delta t]$  VI-27

Et, en se rappelant que  $r = \alpha \frac{\Delta t}{(\Delta x)^2}$  et que  $\phi = m \frac{\pi}{N}$  (en oubliant l'indice  $m$  pour  $\phi$ ), on trouve pour la « sortie analytique » :

$$(E_m^1)_{\text{anal}} = E_m^0 \exp(-\phi^2 r) \quad \text{VI-28}$$

VI-25 et VI-28 fournissent la comparaison cherchée : le schéma numérique est responsable de deux erreurs :

- a) l'erreur en amplitude  $\varepsilon_D$ , appelée *erreur de diffusion*, est égale au rapport de l'amplitude numérique à l'amplitude analytique :

$$\varepsilon_D = \frac{|G|}{\exp(-\phi^2 r)} \quad \text{VI-29}$$

- b) l'erreur de phase  $\varepsilon_\Phi$ , appelée *erreur de dispersion*, est définie, pour les problèmes paraboliques par l'écart

$$\varepsilon_\Phi = \arg[(E_m^1)_{\text{num}}] - \arg[(E_m^1)_{\text{anal}}]$$

qui, au vu de VI-28, se réduit ici à

$$\varepsilon_\Phi = \arg[(E_m^1)_{\text{num}}] \quad \text{VI-30}$$

Ces deux erreurs seront utilisées pour évaluer les performances d'un schéma numérique stable dédiacé au modèle parabolique ;  $G$  dépendant de  $\phi$ , c'est en fonction de ce dernier paramètre que ces erreurs seront évaluées. Un schéma numérique sera d'autant meilleur que  $\varepsilon_D$  sera proche de un et que  $\varepsilon_\phi$  sera proche de zéro.

**Modèle hyperbolique :**  $u_t = -au_x$

Un raisonnement analogue au cas du modèle parabolique donne

$$(E_m^1)_{\text{anal}} = E_m^0 \exp(-j\phi v) \quad \text{VI-31}$$

où  $v = \frac{a\Delta t}{\Delta x}$ . On en déduit

a) erreur de diffusion :

$$\varepsilon_D = |G| \quad \text{VI-32}$$

b) erreur de dispersion : définie pour les problèmes hyperboliques par le rapport

$$\varepsilon_\phi = \frac{\arg[(E_m^1)_{\text{num}}]}{\arg[(E_m^1)_{\text{anal}}]}$$

c'est-à-dire (cf. VI-25 et VI-31)

$$\varepsilon_\phi = \frac{\theta}{-\phi v} \quad \text{VI-33}$$

A nouveau, ces erreurs seront évaluées en fonction de  $\phi$ . Un schéma numérique sera cette fois d'autant meilleur que  $\varepsilon_D$  et  $\varepsilon_\phi$  seront proches de un.

## VI.5 Convergence

Un schéma numérique de différences finies est convergent si la solution qu'il fournit approche la véritable solution de l'EDP.

Théorème d'équivalence de Lax : pour tout problème évolutif décrit par un système d'EDP linéaires, la consistance et la stabilité d'un schéma numérique assurent la convergence.

Ce théorème n'a été démontré que pour des systèmes d'EDP linéaires mais on s'accorde à l'étendre à tout les types d'EDP.

## Chapitre VII. Catalogue de quelques méthodes de résolution d'équations hyperboliques

L'équation utilisée pour évaluer ces méthodes reste l'équation d'advection

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = 0 \quad \text{VII-1}$$

L'étude des performances numériques des divers schémas comporte les étapes suivantes : détermination de l'ordre de la méthode via le calcul de l'erreur de troncature, calcul du facteur d'amplification et détermination des éventuelles conditions de stabilité, calcul et exploitation des erreurs de diffusion et de dispersion.

### **VII.1 Méthodes d'Euler explicites instables**

Citées pour mémoire, elles montrent qu'un choix malheureux des formules de différences finies peut générer des schémas inconditionnellement instables : ainsi qu'on l'a vu, c'est le cas pour le choix

$$(u_t)_i^k = \frac{u_i^{k+1} - u_i^k}{\Delta t} \quad \text{et} \quad (u_x)_i^k = \frac{u_{i+1}^k - u_{i-1}^k}{2\Delta x} \quad \text{VII-2}$$

qui génère le facteur d'amplification

$$|G| = \sqrt{1 + v^2 \sin^2(\phi)} \quad \text{VII-3}$$

toujours supérieur ou égal à un.

C'est aussi le cas pour le choix

$$(u_t)_i^k = \frac{u_i^{k+1} - u_i^k}{\Delta t} \quad \text{et} \quad (u_x)_i^k = \frac{u_{i+1}^k - u_i^k}{\Delta x} \quad \text{VII-4}$$

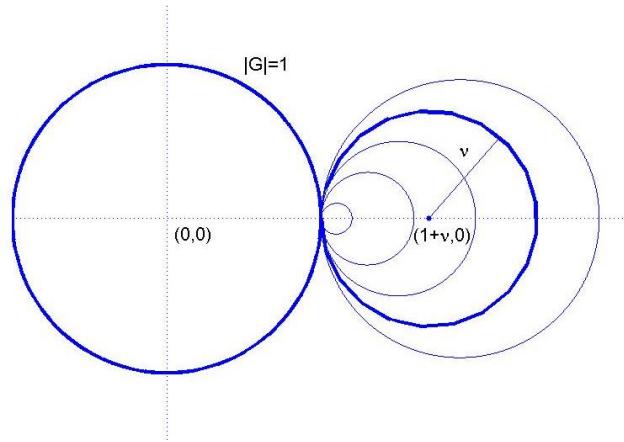
Le calcul de son facteur d'amplification est en effet le suivant :

$$\begin{aligned} A^{k+1} - A^k &= -vA^k (e^{j\phi} - 1) \\ \Rightarrow G &= 1 - v(e^{j\phi} - 1) \\ \Rightarrow |G| &= \sqrt{1 + 2(v + v^2)(1 - \cos \phi)} \end{aligned} \quad \text{VII-5}$$

qui est toujours supérieur ou égal à un car  $v = \frac{a\Delta t}{\Delta x}$  est positif et  $1 - \cos \phi$  est compris entre zéro et deux. Remarquons qu'il n'est pas indispensable de calculer VII-5 pour conclure à l'instabilité inconditionnelle : le facteur d'amplification s'écrit aussi

$$G = 1 + v - ve^{j\phi} \quad \text{VII-6}$$

Dans le plan complexe, quand  $\phi$  varie de  $-\pi$  à  $\pi$ ,  $G$  décrit le cercle de centre  $(1+v, 0)$  et de rayon  $v$



## VII.2 Condition de Courant-Friedrichs-Lowy : méthode d'Euler stabilisée

Une simple modification des choix précédents de discréétisation spatiale permet d'obtenir une méthode stable :

$$\left(\frac{\partial u}{\partial x}\right)_i^k = \frac{u_i^k - u_{i-1}^k}{\Delta x} \quad \text{VII-7}$$

il vient

$$\frac{u_i^{k+1} - u_i^k}{\Delta t} = \frac{-a}{\Delta x} (u_i^k - u_{i-1}^k) \quad \text{VII-8}$$

erreur de troncature : elle est de l'ordre de  $O(\Delta t, \Delta x)$ .

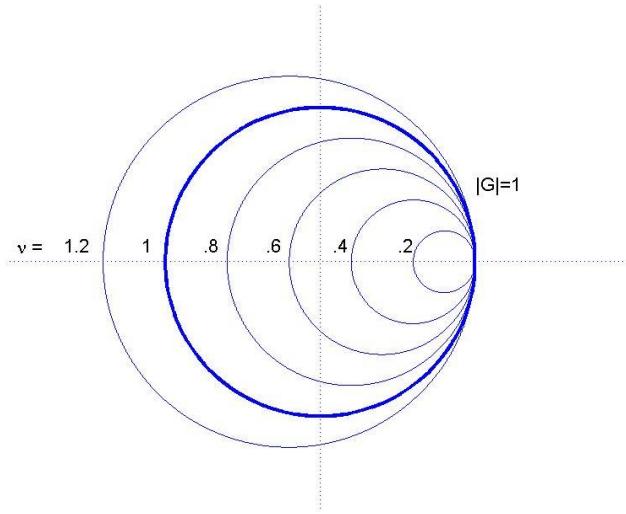
calcul du facteur d'amplification : on trouve

$$G = 1 - v(1 - e^{-j\phi}) = 1 - v + ve^{-j\phi} \quad \text{VII-9}$$

qui est le cercle de centre  $(1-v, 0)$  et de rayon  $v$  dans le plan complexe : la méthode est stable si ce lieu est complètement inclus dans le cercle unitaire : la figure suivante nous montre la position relative de ces deux cercles pour  $v = 0.2, 0.4, \dots, 1.2$ . Il est clair que la condition de stabilité est

$$0 \leq v \leq 1 \quad \text{VII-10}$$

C'est la condition de Courant-Friedrichs-Lowy (condition CFL) ; elle exprime que le schéma numérique VII-8 est conditionnellement stable.  $v = \frac{a\Delta t}{\Delta x}$  est appelé nombre de Courant ; VII-10 fixe la dépendance qui lie les choix de  $\Delta t$  et de  $\Delta x$  et montre que VII-8 est instable pour  $a < 0$ .



erreur de diffusion :

$$\varepsilon_D = |G| = \dots = \sqrt{1 - 4v(1-v) \sin^2 \frac{\phi}{2}} \quad \text{VII-11}$$

erreur de dispersion :

$$\varepsilon_\phi = \frac{\theta}{-\phi v} = \frac{\arctg \left[ \frac{-v \sin \phi}{1 - v + v \cos \phi} \right]}{-\phi v} \quad \text{VII-12}$$

Ces deux erreurs sont des fonctions de  $\phi$ , ou plutôt (voir chap. VI) de  $\phi_m = m \frac{\pi}{N}$ , si on se souvient

que l'indice a été omis par raison de simplicité ; en outre, comme (cf. VI-8)  $\omega_m = m \frac{\pi}{L} = m \frac{\pi}{N \Delta x}$ , il vient

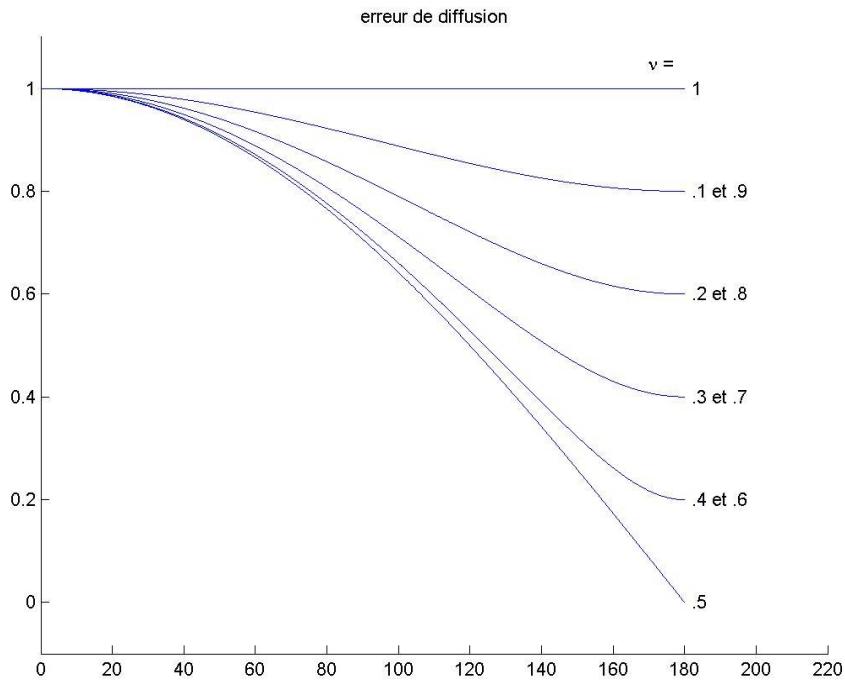
$$\phi_m = \Delta x \omega_m \quad \text{VII-13}$$

Représenter les erreurs en fonction de  $\phi$  revient à les représenter en fonction de la pulsation  $\omega$  des composantes de la série de Fourier de la condition initiale. Les figures suivantes représentent ces deux erreurs pour  $v = 0.1, 0.2, \dots, 1$ .

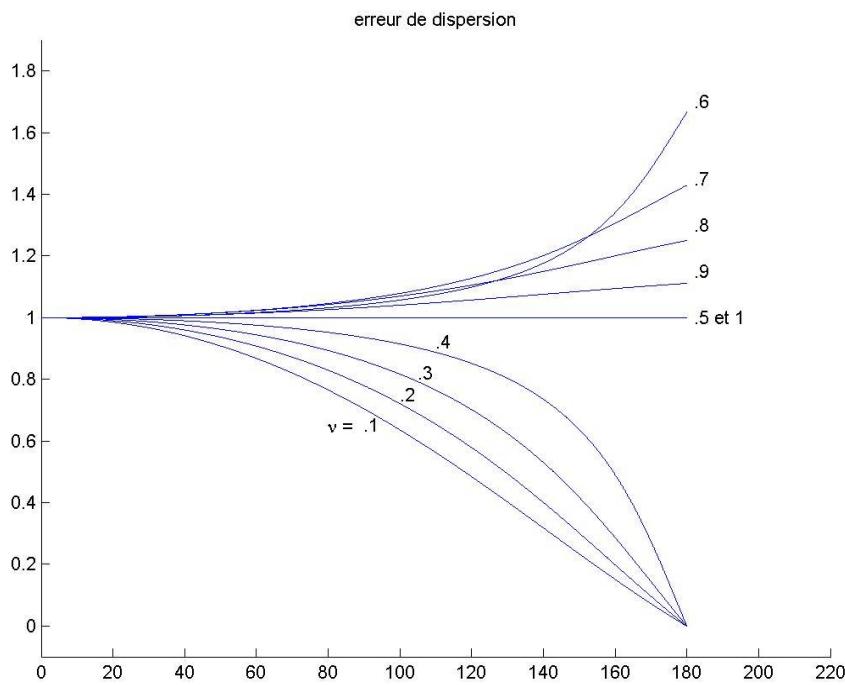
Rappelons (chapitre VI) qu'un schéma est d'autant meilleur que ces deux erreurs sont proches de un.

Du point de vue de l'erreur de diffusion, la seule valeur idéale est  $v = 1$ , et pour toute autre valeur l'erreur de diffusion devient plus importante quand  $\phi$  grandit, ce qui correspond à une atténuation de l'amplitude des signaux.

L'identité de comportement pour deux valeurs également distantes de 0.5 (par exemple 0.8 et 0.2) n'est qu'apparente, comme on le verra plus loin.



L'erreur de dispersion fait apparaître deux valeurs idéales de  $v$  : 0.5 et 1. Hormis ces valeurs particulières, le comportement du schéma est radicalement différent selon que  $v$  est inférieur ou supérieur à 0.5 : pour  $v < 0.5$ , le schéma génère du retard de phase, tandis que pour  $v > 0.5$ , il génère de l'avance de phase.



### Exemple 1

Ce premier exemple a pour but d'illustrer le caractère impératif de conditions telles que la condition CFL VII-10 : au chapitre IV, on y avait résolu

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = 0$$

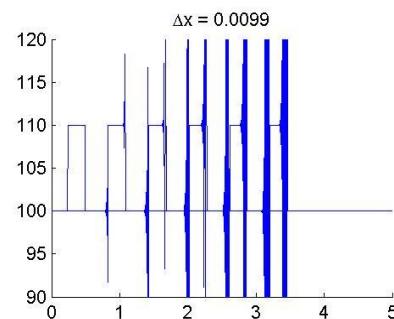
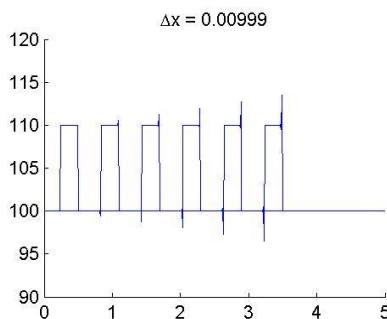
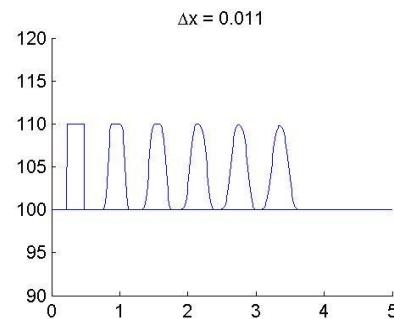
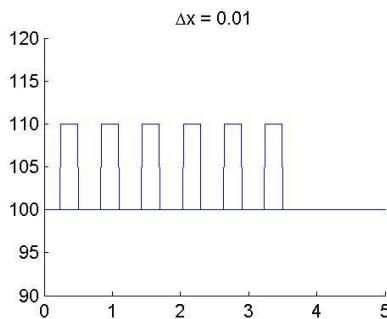
VII-14

dans les hypothèses suivantes :

$$\begin{aligned} 0 \leq x \leq 5 & \quad 0 \leq t \leq 3 & a = 1 \\ u^0(x) = 100 & \quad x \in [0, 0.25[ \cup ]0.5, 5] \\ 110 & \quad x \in [0.25, 0.5] \\ u_0^k = 100 \end{aligned}$$

Le schéma numérique utilisé était VII-8 et deux ensembles de valeurs de  $\Delta x$  et  $\Delta t$  avaient été testés :

1°  $\Delta t = 0.01$       et       $\Delta x = 0.01, 0.011, 0.00999, 0.0099$  qui engendrait les solutions suivantes :



Les valeurs de  $v$  correspondant à ces quatre essais sont

$\Delta x$	$\Delta t$	$v$
0.01	0.01	1
0.011	0.01	0.909
0.00999	0.01	1.001
0.0099	0.01	1.01

Le premier essai confirme que  $\nu = 1$  est la valeur idéale fournissant une solution numérique identique à la solution analytique. Pour  $\nu > 1$ , même d'infimes dépassements du seuil de stabilité génèrent des instabilités dans la solution. Et  $\nu = 0.909$  confirme l'atténuation attendue, plus importante pour les hautes fréquences du spectre de Fourier de la condition initiale (les points anguleux de l'échelon) que pour les basses fréquences (l'amplitude globale du signal est peu affectée).

$$2^\circ \quad \Delta x = 0.01 \quad \text{et} \quad \Delta t = 0.01, 0.009, 0.01001, 0.0101$$

Les constatations sont identiques,  $\nu$  variant cette fois par action sur  $\Delta t$

### **Exemple 2**

Les essais qui suivent illustrent les informations qu'on peut tirer des erreurs de diffusion et de dispersion : résolvons à nouveau

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = 0 \quad \text{avec} \quad a = 0.8 \quad \text{VII-15}$$

sur  $0 \leq x \leq 1$

$$\text{avec } u(x,0) = \sin(6\pi x) \quad \text{VII-16}$$

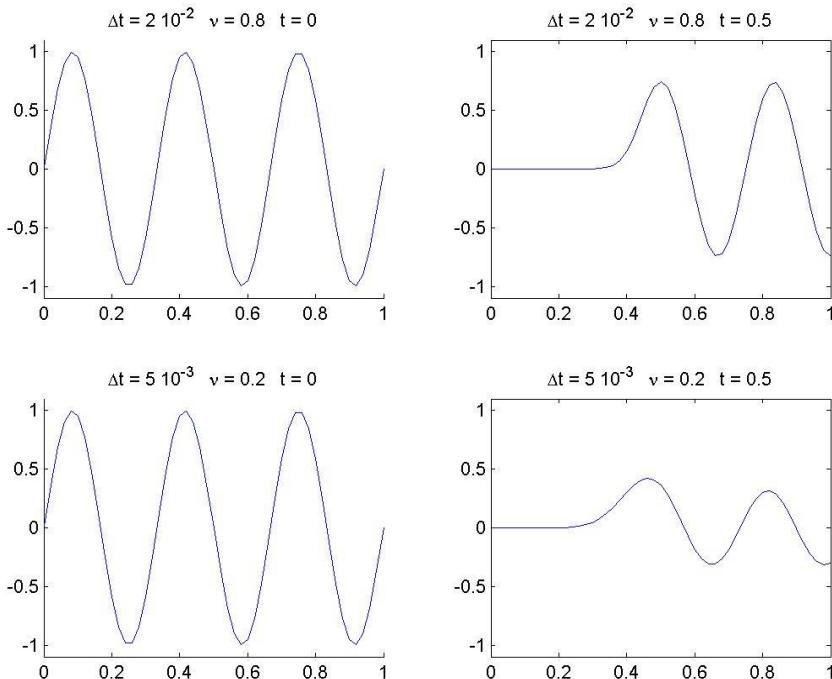
$$\text{et } u(0,t) = 0$$

La figure ci-dessous représente la condition initiale VII-16 et la solution numérique obtenue grâce à VII-8 en  $t = 0.5$  avec

$$\Delta x = 0.02 \text{ et } \Delta t = 0.02 \Rightarrow \nu = 0.8$$

et

$$\Delta x = 0.02 \text{ et } \Delta t = 0.005 \Rightarrow \nu = 0.2$$



En désaccord avec ce que le graphe de l'erreur de diffusion suggère, l'atténuation est plus importante avec  $v = 0.2$  qu'avec  $v = 0.8$  : cela provient de ce qu'il est nécessaire de calculer un nombre de pas quatre fois plus grand avec  $v = 0.2$  pour obtenir  $u(x, 0.5)$  qu'avec  $v = 0.8$  : le facteur d'amplification, et l'atténuation qu'il engendre, intervient donc quatre fois plus également.

Les valeurs numériques suivantes peuvent être calculées :

Le spectre de Fourier de la condition initiale ne comporte qu'une seule pulsation :  $\omega = 6\pi$ . Il lui correspond la phase suivante :  $\phi = \omega\Delta x = 0.12\pi$

Pour  $v = 0.8$ , on a donc

$$\varepsilon_D = |G| = \sqrt{1 - 4v(1-v)\sin^2 \frac{\phi}{2}} = \dots \cong 0.9887 \quad \text{VII-17}$$

La solution en  $t = 0.5$  est obtenue au bout de 25 itérations VII-8, ce qui correspond à une atténuation de l'ordre de 0.7527, conforme à l'observation visuelle.

Pour  $v = 0.2$ ,  $\varepsilon_D$  est inchangé mais 100 itérations ont été nécessaires, ce qui engendre une atténuation de l'ordre de 0.3210, également conforme à l'observation.

Voyons maintenant comment utiliser l'erreur de dispersion. Dans le cas de l'équation

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = 0 \quad \text{VII-18}$$

on sait que la solution analytique est la condition initiale se déplaçant à la vitesse  $a$  :

$$u(x, t) = u(x - at, 0) = u^0(x - at) \quad \text{VII-19}$$

On a choisi ici

$$u^0(x) = \sin(6\pi x) = \sin(\omega x) \quad \text{VII-20}$$

À l'instant  $\Delta t$  on a donc

$$u(x, \Delta t) = \sin(\omega(x - a\Delta t)) = \sin(\omega x - a\omega\Delta t) \quad \text{VII-21}$$

Le déphasage de la solution analytique vaut donc  $-a\omega\Delta t$  : il est proportionnel à la vitesse de déplacement de la solution.

L'erreur de dispersion vaut (VII-12)

$$\varepsilon_\Phi = \frac{\theta}{-\phi v} = \frac{\arg(G)}{-\omega\Delta x \frac{a\Delta t}{\Delta x}} = \frac{\arg(G)}{-a\omega\Delta t} \quad \text{VII-22}$$

$\varepsilon_\Phi$  est un rapport de déphasage, mais comme le déphasage est proportionnel à la vitesse de déplacement, il peut être aussi interprété comme un rapport de vitesse de déplacement en observant que la vitesse déplacement de la solution numérique  $a_{\text{num}}$  est définie par

$$a_{\text{num}} = \frac{\arg(G)}{-\omega \Delta t}$$

VII-23

Quand  $\varepsilon_\phi$  est supérieur à un, l'avance de phase observée correspond à une solution numérique se déplaçant plus vite que l'analytique, et quand il est inférieur à un, le retard de phase équivaut à une vitesse de déplacement inférieure à la vitesse analytique.

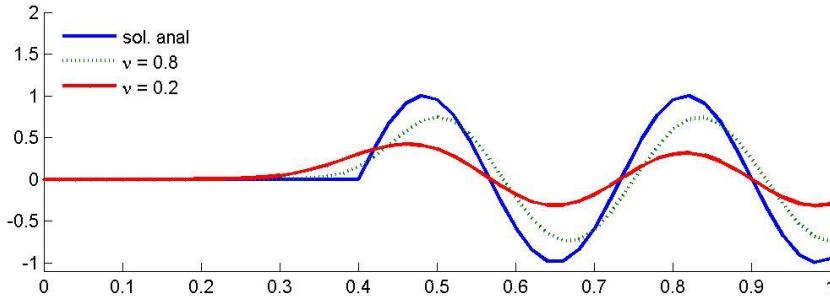
Pour  $v = 0.8$ , on a

$$\varepsilon_\phi = \frac{\arctg\left[\frac{-v \sin \phi}{1 - v + v \cos \phi}\right]}{-\phi v} = \frac{\arctg\left[\frac{-0.8 \sin(0.12\pi)}{1 - 0.8 + 0.8 \cos(0.12\pi)}\right]}{-(0.12\pi)0.8} = 1.0029$$

et pour  $v = 0.2$ , on trouve

$$\varepsilon_\phi = \frac{\arctg\left[\frac{-0.2 \sin(0.12\pi)}{1 - 0.2 + 0.2 \cos(0.12\pi)}\right]}{-(0.12\pi)0.2} = 0.9886$$

En illustration, la figure ci-dessous superpose la solution analytique en  $t = 0.5$  aux deux solutions numériques.



### VII.3 Méthode de Lax

Cette méthode utilise les discréétisations suivantes :

$$\left(\frac{\partial u}{\partial t}\right)_i^k = \frac{u_i^{k+1} - (u_{i+1}^k + u_{i-1}^k)/2}{\Delta t}$$

et

$$\left(\frac{\partial u}{\partial x}\right)_i^k = \frac{u_{i+1}^k - u_{i-1}^k}{2\Delta x}$$

VII-24

erreur de troncature : on a

$$u_i^{k+1} = u_i^k + \frac{\Delta t}{1!} u_{it}^k + \frac{\Delta t^2}{2!} u_{itt}^k + \dots$$

$$u_{i+1}^k = u_i^k + \frac{\Delta x}{1!} u_{ix}^k + \frac{\Delta x^2}{2!} u_{ixx}^k + \dots$$

$$u_{i-1}^k = u_i^k - \frac{\Delta x}{1!} u_{ix}^k + \frac{\Delta x^2}{2!} u_{ixx}^k + \dots$$

d'où

$$\left(\frac{\partial u}{\partial t}\right)_i^k + a\left(\frac{\partial u}{\partial x}\right)_i^k = u_{it}^k + \frac{\Delta t}{2!} u_{itt}^k + \dots + \frac{\Delta x^2}{2\Delta t} u_{ixx}^k + \dots + a u_{ix}^k + a \frac{\Delta x^2}{3!} u_{ixxx}^k + \dots$$

l'erreur de troncature est donc de l'ordre de  $O(\Delta t, \frac{\Delta x^2}{\Delta t})$  : la méthode n'est consistante que si

$$\lim_{grille \rightarrow 0} \frac{\Delta x^2}{\Delta t} = 0$$

Calcul du facteur d'amplification : un calcul analogue à ceux des cas précédents fournit

$$G = \cos \phi - j v \sin \phi$$

VII-25

dont le module vaut  $\sqrt{\cos^2 \phi + v^2 \sin^2 \phi}$  : la méthode est conditionnellement stable : la condition de stabilité est

$$-1 \leq v \leq 1$$

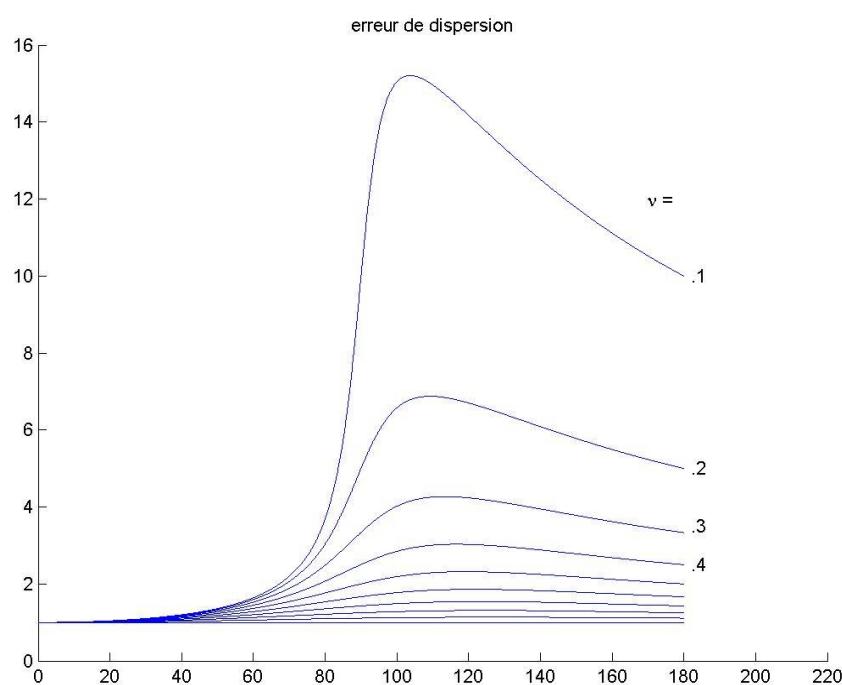
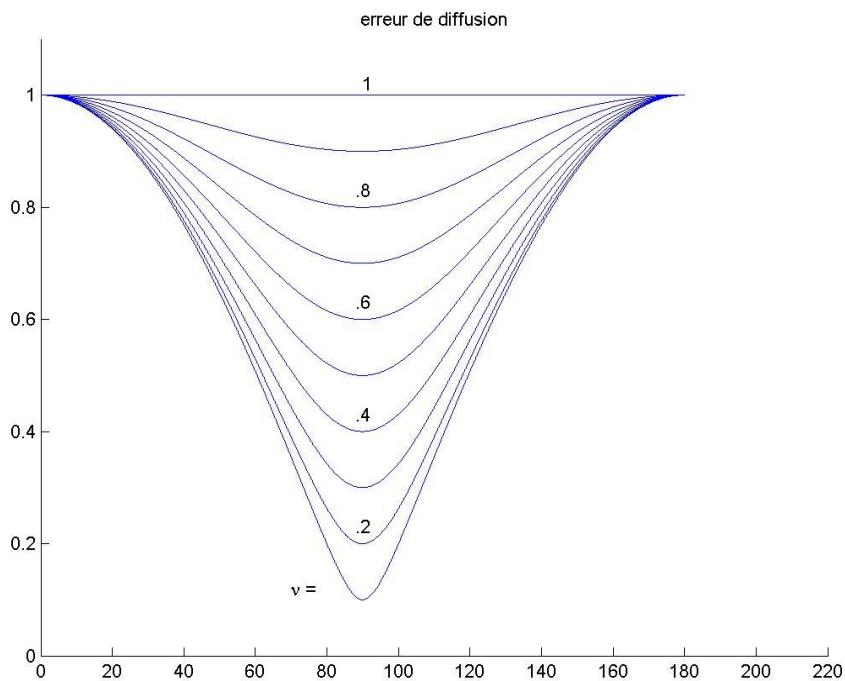
Ce schéma permet donc de traiter les cas  $a > 0$  et  $a < 0$ .

erreurs de diffusion et de dispersion :

$$\varepsilon_D = |G| = \sqrt{(\cos \phi)^2 + (v \sin \phi)^2} \quad VII-26$$

$$\varepsilon_\Phi = \frac{\theta}{-\phi v} = \frac{\operatorname{arctg} \left[ \frac{-v \sin \phi}{\cos \phi} \right]}{-\phi v} \quad VII-27$$

Représentées aux figures suivantes, ces erreurs montrent qu'à nouveau la seule valeur de  $v$  assurant une simulation parfaite est  $v = 1$ . Toute autre valeur introduit de l'atténuation et de l'avance de phase qui peut être très importante quand  $v$  est très petit.



#### VII.4 Méthode leap frog

Cette méthode utilise les discréétisations suivantes :

$$\left( \frac{\partial u}{\partial t} \right)_i^k = \frac{u_i^{k+1} - u_i^{k-1}}{2\Delta t}$$

VII-28

et

$$\left(\frac{\partial u}{\partial x}\right)_i^k = \frac{u_{i+1}^k - u_{i-1}^k}{2\Delta x}$$

VII-29

il en résulte :

$$\frac{u_i^{k+1} - u_i^{k-1}}{2\Delta t} = -a \frac{u_{i+1}^k - u_{i-1}^k}{2\Delta x}$$

VII-30

erreur de troncature : elle est de l'ordre de  $O(\Delta t^2, \Delta x^2)$ .

Ce schéma utilise 3 niveaux de temps : le démarrage requiert donc de connaître  $u_i^0$  et  $u_i^1 \forall i$  ; la CI fournit  $u_i^0$  ;  $u_i^1$  est calculé avec un schéma à 2 niveaux de temps analogue à ceux qui précédent.

Facteur d'amplification : la présence des 3 niveaux demande de pratiquer le calcul de la manière suivante : on introduit d'abord de manière classique les opérateurs  $A^k$  et  $\exp(j\phi)$  dans VII-30 :

$$A^{k+1} - A^{k-1} = -v(A^k e^{j\phi} - A^k e^{-j\phi}).$$

Compte tenu de la définition :

$$G = \frac{A^{k+1}}{A^k},$$

on écrit dans le schéma de calcul  $A^{k+1} = GA^k$  et  $A^{k-1} = G^{-1}A^k$  : il vient

$$(G - \frac{1}{G})A^k = -vA^k(e^{j\phi} - e^{-j\phi})$$

c'est-à-dire

$$G = -jv \sin \phi \pm \sqrt{1 - v^2 \sin^2 \phi}$$

VII-31

pour  $v > \left| \frac{1}{\sin \phi} \right|$ , le radicant est négatif et  $G$  est imaginaire pur. Son amplitude vaut

$v \sin \phi \pm \sqrt{v^2 \sin^2 \phi - 1}$ . Une de ces valeurs est plus grande que l'unité ce qui fait dire que le schéma est instable. En outre,  $\phi$  étant la phase d'une composante quelconque du développement en série de

Fourier du signal d'erreur, il est prudent d'envisager le cas le plus défavorable : celui où  $\phi = \frac{\pi}{2}$ . C'est

cette prudence qui fait dire que le schéma est instable pour  $v > 1$ .

Pour  $v \leq 1$ , le radicant est positif,  $G$  est complexe et son amplitude vaut 1 quel que soit  $\phi$  : on dit que le schéma est neutralement stable.

Analyse spectrale des erreurs : pour  $v \leq 1$  :

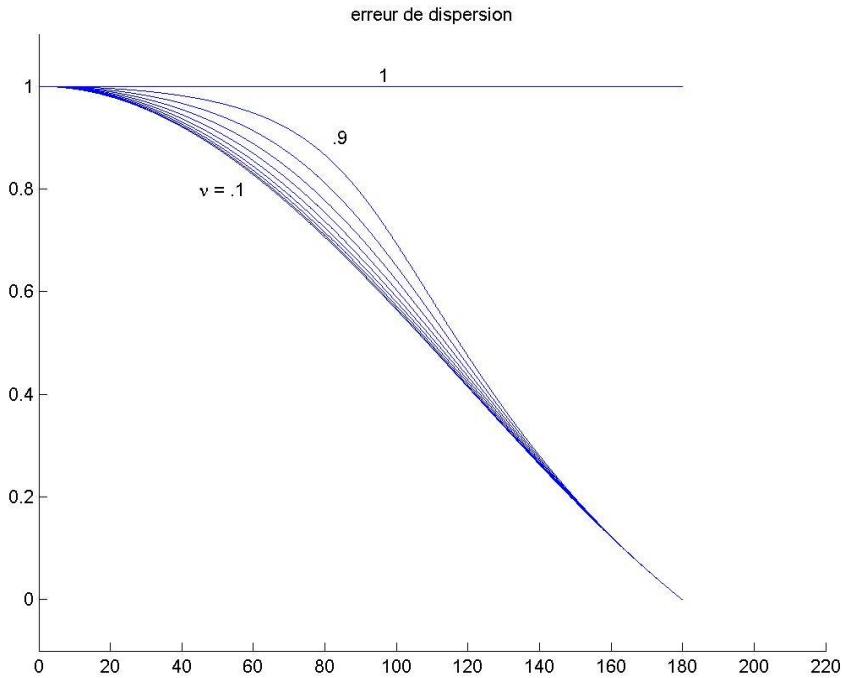
$$\varepsilon_D = |G| = 1$$

VII-32

$$\varepsilon_\phi = \frac{\theta}{-\phi v} = \frac{\arctg \frac{-v \sin \phi}{\sqrt{1 - (v \sin \phi)^2}}}{-\phi v}$$

VII-33

$\varepsilon_\phi$  est représenté à la figure suivante. A nouveau, seule la valeur  $v = 1$  fournit une simulation parfaite.



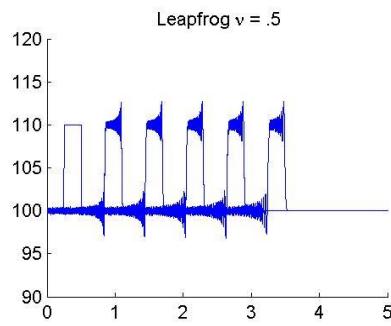
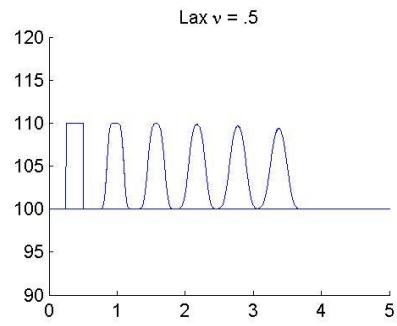
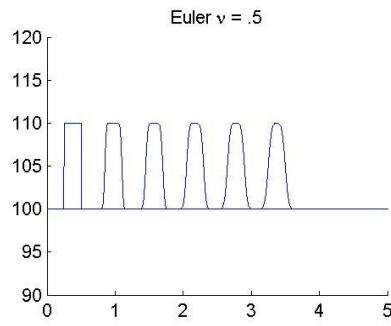
## VII.5 comparaison des méthodes

Les trois méthodes précédentes sont comparées globalement en reprenant l'exemple 1 du paragraphe VII-2 en utilisant la même valeur de  $v$ . Cette comparaison, comme l'illustre la figure suivante met en évidence deux types de comportements très différents :

Les méthodes d'Euler et de Lax présentent un comportement diffusif : c'est l'atténuation qu'impose l'erreur de diffusion qui apparaît dans la forme de la solution ; cette atténuation affecte surtout les composantes de fréquences élevées du spectre de Fourier de la condition initiale. Un tel comportement est aussi appelé dissipatif.

La méthode leapfrog présente un comportement dispersif, avec apparition d'oscillations aux points anguleux de la solution.

Une explication plus détaillée de ces comportements sera abordée dans le chapitre relatif à l'équation différentielle équivalente modifiée. On y verra que l'examen de l'erreur de troncature permet de mieux comprendre pourquoi des schémas numériques sont dissipatifs ou dispersifs.



## Chapitre VIII. Catalogue de quelques méthodes de résolution d'équations paraboliques

L'équation utilisée pour évaluer ces méthodes reste la deuxième loi de Fourier I-5

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2} \quad \text{VIII-1}$$

### **VIII.1 Méthode explicite simple**

Il s'agit de la méthode déjà rencontrée au chapitre VI : rappelons qu'elle utilise :

$$\left( \frac{\partial u}{\partial t} \right)_i^k = \frac{u_i^{k+1} - u_i^k}{\Delta t} \quad \text{VIII-2}$$

et

$$\left( \frac{\partial^2 u}{\partial x^2} \right)_i^k = \frac{u_{i+1}^k - 2u_i^k + u_{i-1}^k}{(\Delta x)^2} \quad \text{VIII-3}$$

erreur de troncature : on a vu qu'elle est de l'ordre de  $O(\Delta t, \Delta x^2)$

facteur d'amplification : en posant  $r = \alpha \frac{\Delta t}{(\Delta x)^2}$  on a trouvé

$$G = 1 + 2r(\cos \phi - 1) \quad \text{VIII-4}$$

et la condition de stabilité

$$0 \leq r \leq \frac{1}{2} \quad \text{VIII-5}$$

erreur de diffusion : on a vu qu'elle valait (cf. VI-29)

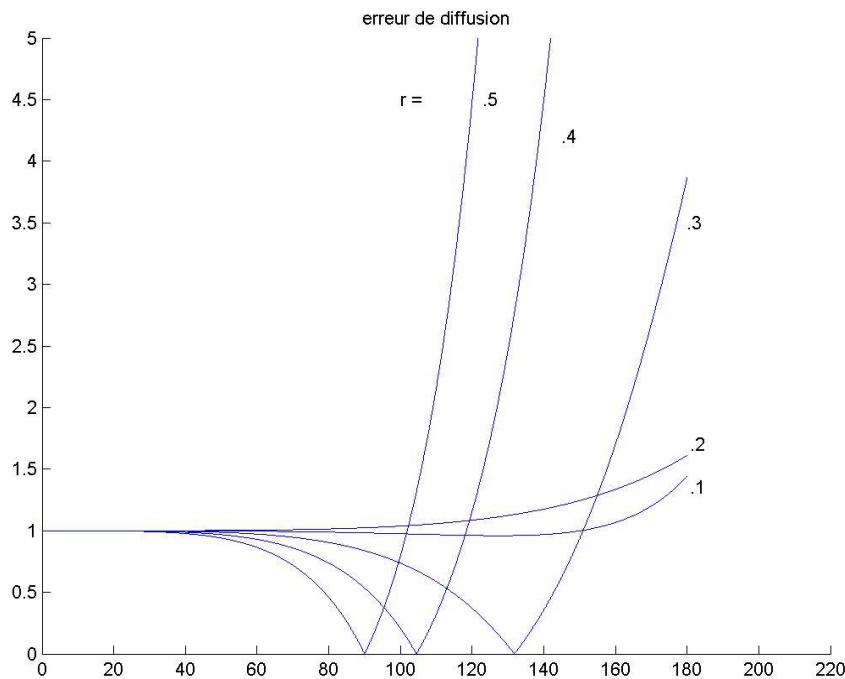
$$\varepsilon_D = \frac{|G|}{\exp(-\phi^2 r)} = \frac{|1 + 2r(\cos \phi - 1)|}{e^{-\phi^2 r}} \quad \text{VIII-6}$$

erreur de dispersion : elle vaut (cf. VI-30)

$$\varepsilon_\Phi = \arg[(E_m^1)_{\text{num}}]$$

cette erreur est nulle puisque G est réel.

$\varepsilon_D$  est représenté pour  $r = 0.1, 0.2, \dots, 0.5$  à la figure suivante :

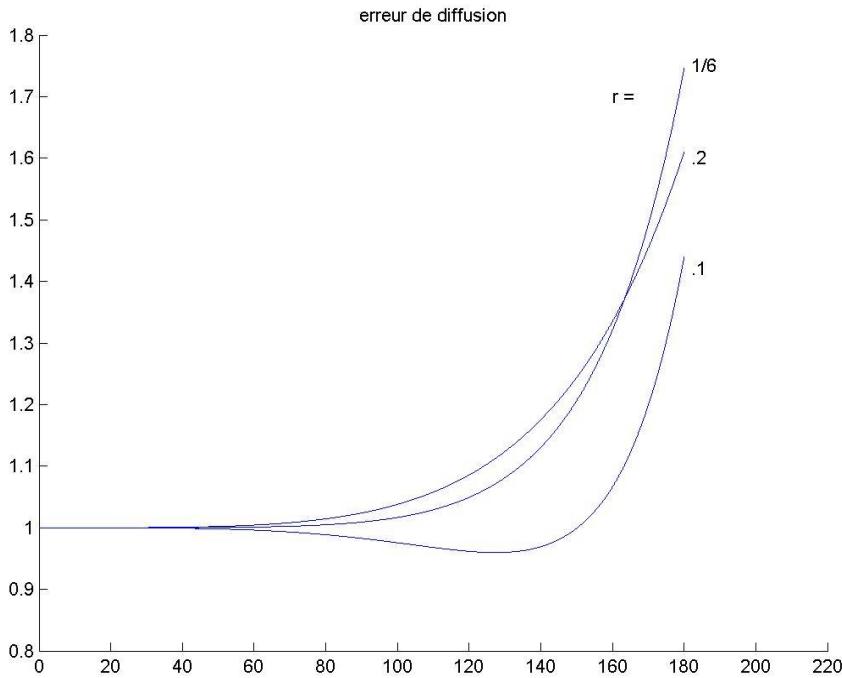


On observe que les meilleures valeurs sont 0.1 et 0.2. Une valeur optimale de  $r$  peut être dégagée en développant  $\varepsilon_D$  en série de puissance de  $\phi$  : connaissant les développements de  $\cos \phi$  et de l'exponentielle, on obtient aisément la relation

$$\varepsilon_D \approx 1 + \left( \frac{r}{12} - \frac{r^2}{2} \right) \phi^4 + \dots \quad \text{VIII-7}$$

Il est aisément d'annuler le premier terme d'erreur de cette expression : il suffit de prendre  $r = \frac{1}{6}$  : la figure suivante représente  $\varepsilon_D$  pour  $r = 0.1, 1/6$  et  $0.2$ .

C'est bien pour  $1/6$  que  $\varepsilon_D$  reste proche de 1 pour la plus grande plage de valeurs de  $\phi$ .



Néanmoins, contrairement à ce qu'on a obtenu avec les équations hyperboliques, il n'existe pas de valeur de  $r$  permettant d'avoir une résolution numérique parfaite. On peut d'ailleurs craindre que les résultats de simulation soient médiocres si on observe que pour  $r \geq 0.3$  par exemple, l'erreur de diffusion est rapidement variable entre 0 et 3 à 5. Cette crainte n'est pas fondée, comme on peut le voir en reprenant l'exemple développé au chapitre V :

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2} \quad \text{avec} \quad 0 < x < 1. \quad \text{VIII-8}$$

avec la condition initiale :  $u(x, 0) = 0$

VIII-9

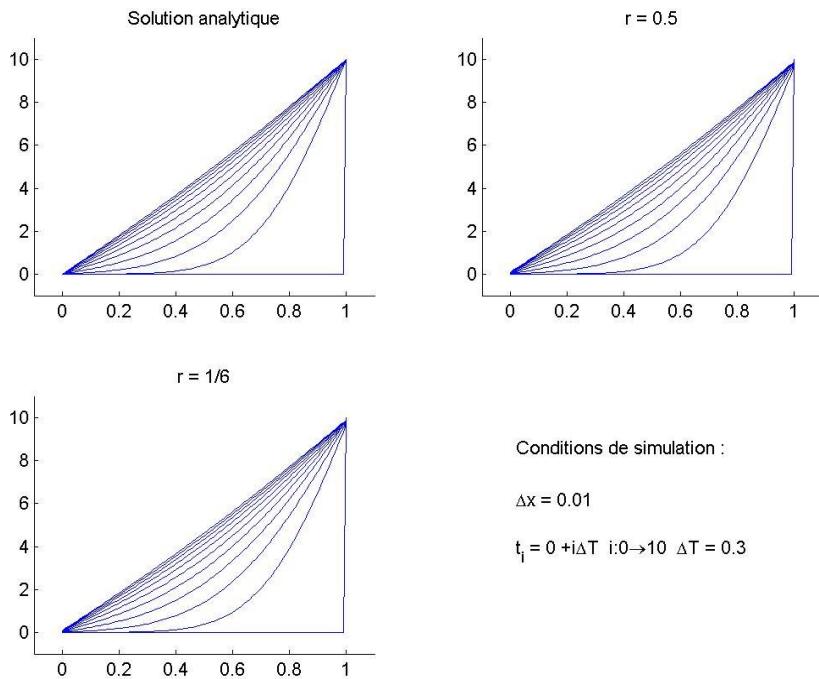
et les conditions aux limites :  $u(0, t) = 0$  et  $u(1, t) = U_0$ ,

VIII-10

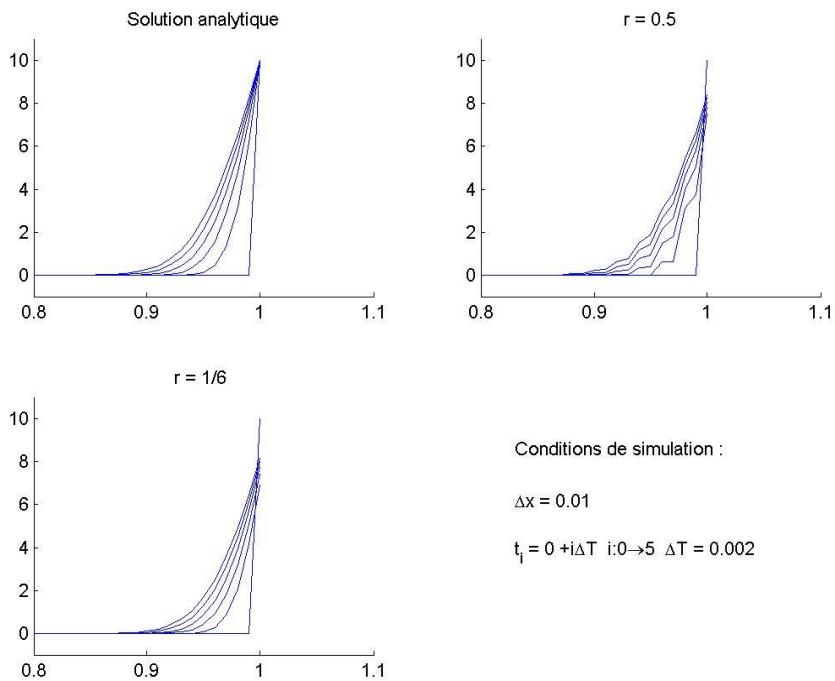
dont la solution est

$$u(x, t) = U_0 x + \sum_{n=1}^{\infty} \frac{2U_0(-1)^n}{n\pi} \exp(-n^2\pi^2\alpha t) \sin(n\pi x) \quad \text{VIII-11}$$

La figure suivante compare la solution analytique aux solutions numériques obtenues avec  $r = 0.5$  et  $r = \frac{1}{6}$ . S'il semble logique que la solution relative à  $r = \frac{1}{6}$  soit proche de la solution analytique, il est plus étonnant que la solution relative à  $r = 0.5$  soit semblable aux deux autres. L'explication est la suivante : l'atténuation « analytique » subie par les composantes du spectre de Fourier de la condition initiale en  $k\Delta t$  vaut (cf. par exemple VI-28)  $(\exp(-\phi^2 r))^k$ . Cette atténuation se marque très fort pour les hautes fréquences de ce spectre qui rapidement ne comporte quasi plus que des fréquences relatives aux très petites valeurs de  $\phi$ . Pour celles-ci ( $\phi < 40^\circ$  sur le graphe de l'erreur de diffusion),  $\varepsilon_D$  vaut quasi un pour toutes les valeurs de  $r$ .



Pour évaluer l'impact du choix de la valeur de  $r$ , il est nécessaire d'observer les solutions numériques dans les tous premiers instants de calcul, quand l'amortissement des composantes à fréquences élevées n'a pas encore eu lieu :



Il est clair que dans ce cas,  $r = \frac{1}{6}$  fournit de bien meilleurs résultats.

## VIII.2 Méthode implicite simple

Elle utilise

$$\left(\frac{\partial u}{\partial t}\right)_i^k = \frac{u_i^{k+1} - u_i^k}{\Delta t} \quad \text{VIII-12}$$

et

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_i^k = \frac{u_{i+1}^{k+1} - 2u_i^{k+1} + u_{i-1}^{k+1}}{(\Delta x)^2} \quad \text{VIII-13}$$

erreur de troncature : sa détermination est moins immédiate que dans les méthodes vues jusqu'à présent : en reprenant le raisonnement du paragraphe VI.1 on écrit :

$$\begin{aligned} u_i^{k+1} &= u_i^k + \Delta t(u_t)_i^k + \frac{\Delta t^2}{2}(u_{tt})_i^k + \dots \\ u_{i+1}^{k+1} &= u_i^{k+1} + \Delta x(u_x)_i^{k+1} + \frac{\Delta x^2}{2}(u_{xx})_i^{k+1} + \dots \\ u_{i-1}^{k+1} &= u_i^{k+1} - \Delta x(u_x)_i^{k+1} + \frac{\Delta x^2}{2}(u_{xx})_i^{k+1} + \dots \end{aligned}$$

on rapporte alors toutes les expressions au point (i,k) en complétant par (et en omettant les indices  $(\cdot)_i^k$  par simplicité d'écriture) :

$$\begin{aligned} (u_x)_i^{k+1} &= u_x + \Delta t(u_{xt}) + \frac{\Delta t^2}{2}(u_{xxt}) \dots \\ (u_{xx})_i^{k+1} &= u_{xx} + \Delta t(u_{xxt}) + \frac{\Delta t^2}{2}(u_{xxx}) \dots \end{aligned}$$

VIII-12 et VIII-13 donnent alors

$$\left(\frac{\partial u}{\partial t}\right) = u_t + \frac{\Delta t}{2} u_{tt} + \frac{\Delta t^2}{6} u_{ttt} + \dots \quad \text{VIII-14}$$

$$\left(\frac{\partial^2 u}{\partial x^2}\right) = \frac{u_{i+1}^{k+1} - 2u_i^{k+1} + u_{i-1}^{k+1}}{(\Delta x)^2} = (u_{xx}) + \Delta t(u_{xxt}) + \frac{\Delta t^2}{2}(u_{xxx}) \dots + \frac{\Delta x^2}{12} u_{xxxx} \dots \quad \text{VIII-15}$$

VIII-14 et VIII-15 donnent alors

$$\begin{aligned} \frac{\partial u}{\partial t} - \alpha \frac{\partial^2 u}{\partial x^2} &= \\ u_t + \frac{\Delta t}{2} u_{tt} + \frac{\Delta t^2}{6} u_{ttt} + \dots - \alpha[(u_{xx}) + \Delta t(u_{xxt}) + \frac{\Delta t^2}{2}(u_{xxx}) \dots + \frac{\Delta x^2}{12} u_{xxxx} \dots] &= 0 \end{aligned} \quad \text{VIII-16}$$

l'erreur de troncature vaut donc

$$ET = -\frac{\Delta t}{2} u_{tt} - \frac{\Delta t^2}{6} u_{ttt} - \dots + \alpha[\Delta t(u_{xxt}) + \frac{\Delta t^2}{2} u_{xxx} + \dots + \frac{\Delta x^2}{12} u_{xxxx} \dots]$$

ET est donc de l'ordre de  $O(\Delta t, \Delta x^2)$ .

Facteur d'amplification : on trouve

$$G = \frac{1}{1 + 2r(1 - \cos \phi)}$$

VIII-17

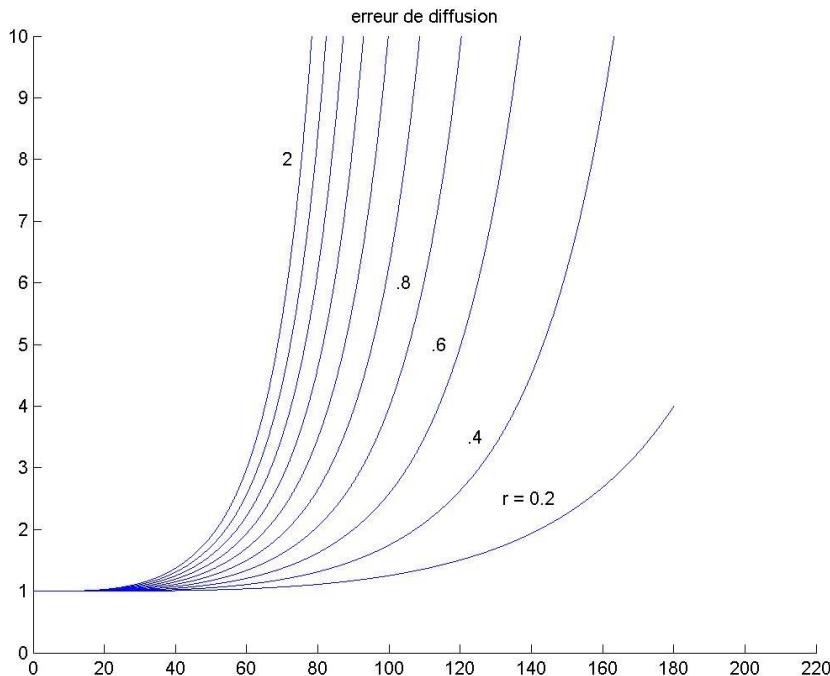
$\phi$  étant compris entre zéro et  $\pi$ , G est toujours inférieur ou égal à un : la méthode est inconditionnellement stable

analyse spectrale des erreurs :

$$\text{erreur de diffusion : } \varepsilon_D = \frac{|G|}{\exp(-\phi^2 r)} = \frac{1}{|1 + 2r(1 - \cos \phi)| e^{-\phi^2 r}} \quad \text{VIII-18}$$

erreur de dispersion : cette erreur est nulle puisque G est réel.

$\varepsilon_D$  est représenté à la figure suivante pour  $r = 0.2, 0.4, \dots, 2$ . On observe que  $\varepsilon_D$  reste d'autant plus proche de un que  $r$  est petit ; seules les CI dont le contenu spectral correspond à de petites valeurs de  $\phi$  sont susceptibles d'être traitées correctement avec  $r$  grand. A ce cas particulier près, il en résulte que pour autant qu'elle soit stable, la méthode explicite simple donne généralement de meilleurs résultats que la méthode implicite simple, dont l'intérêt est alors d'introduire la méthode suivante.



### VIII.3 Méthode de Cranck-Nicholson

C'est une combinaison des deux précédentes : elle utilise

$$\left( \frac{\partial u}{\partial t} \right)_i^k = \frac{u_i^{k+1} - u_i^k}{\Delta t}$$

VIII-19

et

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_i^k = \frac{1}{2} \frac{u_{i+1}^k - 2u_i^k + u_{i-1}^k}{(\Delta x)^2} + \frac{1}{2} \frac{u_{i+1}^{k+1} - 2u_i^{k+1} + u_{i-1}^{k+1}}{(\Delta x)^2} \quad \text{VIII-20}$$

erreur de troncature : un raisonnement analogue à celui du paragraphe précédent fournit

$$\begin{aligned} \frac{\partial u}{\partial t} - \alpha \frac{\partial^2 u}{\partial x^2} &= \\ u_t + \frac{\Delta t}{2} u_{tt} + \frac{\Delta t^2}{6} u_{ttt} + \dots - \frac{\alpha}{2} [ &(u_{xx}) + \Delta t(u_{xxt}) + \frac{\Delta t^2}{2}(u_{xxtt}) \dots + \frac{\Delta x^2}{12} u_{xxxx} \dots + u_{xx} + \frac{\Delta x^2}{12} u_{xxxx} + \dots ] = 0 \end{aligned} \quad \text{VIII-21}$$

c'est-à-dire

$$ET = -\frac{\Delta t}{2} u_{tt} - \frac{\Delta t^2}{6} u_{ttt} - \dots + \frac{\alpha}{2} [\Delta t(u_{xxt}) + \frac{\Delta t^2}{2} u_{xxtt} + \dots + \frac{\Delta x^2}{6} u_{xxxx} \dots] \quad \text{VIII-22}$$

En regroupant les termes en  $\Delta t$  de cette expression, on observe qu'ils se mettent sous la forme

$$\frac{\Delta t}{2} u_{tt} - \frac{\alpha}{2} \Delta t u_{xxt} = \frac{\Delta t}{2} \frac{\partial}{\partial t} (u_t - \alpha u_{xx}) \quad \text{VIII-23}$$

VIII-21 s'écrit alors

$$u_t - \alpha u_{xx} = -\frac{\Delta t}{2} \frac{\partial}{\partial t} (u_t - \alpha u_{xx}) + O(\Delta t^2, \Delta x^2) \quad \text{VIII-24}$$

en dérivant cette relation par rapport à  $t$  et en la multipliant ensuite par  $\frac{\Delta t}{2}$  on obtient

$$\frac{\Delta t}{2} \frac{\partial}{\partial t} (u_t - \alpha u_{xx}) = -\frac{\Delta t^2}{4} \frac{\partial^2}{\partial t^2} (u_t - \alpha u_{xx}) + O(\Delta t^2, \Delta x^2) \quad \text{VIII-25}$$

c'est-à-dire, en remplaçant dans VIII-24

$$u_t - \alpha u_{xx} = O(\Delta t^2, \Delta x^2) \quad \text{VIII-26}$$

ce qui nous indique que l'erreur de troncature est de l'ordre de  $O(\Delta t^2, \Delta x^2)$ .

Facteur d'amplification : on trouve

$$G = \frac{1 - r(1 - \cos \phi)}{1 + r(1 - \cos \phi)}$$

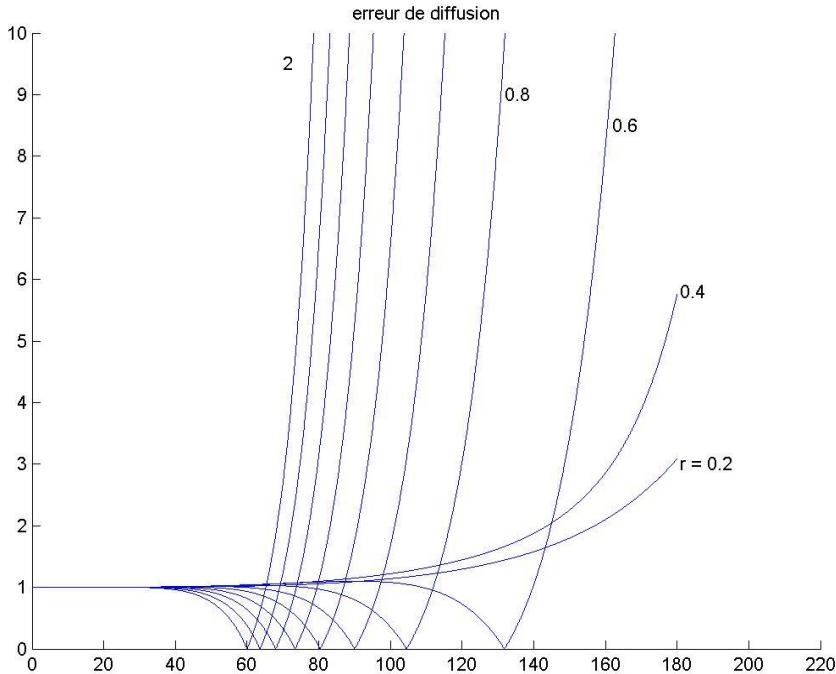
on trouve aisément que la méthode est inconditionnellement stable.

analyse spectrale des erreurs :

$$\text{erreur de diffusion : } \varepsilon_D = \frac{|G|}{\exp(-\phi^2 r)} = \frac{|1 - r(1 - \cos \phi)|}{|1 + r(1 - \cos \phi)| e^{-\phi^2 r}}$$

erreur de dispersion : cette erreur est nulle puisque G est réel.

$\varepsilon_D$  est représenté à la figure suivante pour  $r = 0.2, 0.4, \dots, 2$  :

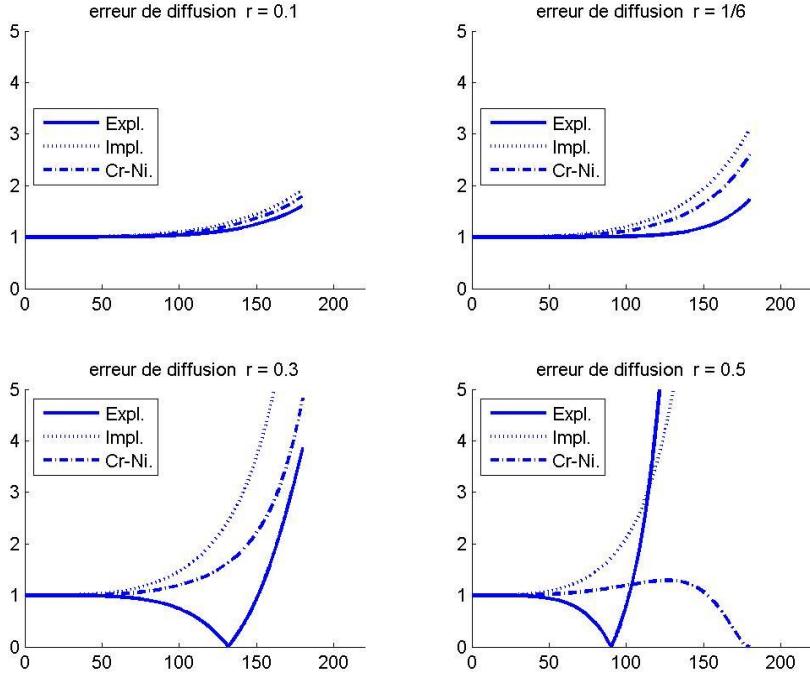


Il apparaît que c'est pour  $r$  égal à environ 0.5 que  $\varepsilon_D$  reste proche de un pour la plus grande plage de valeurs de  $\phi$ . Un développement de  $\varepsilon_D$  en puissances de  $\phi$  comme on l'a fait en VIII-1 ne permet pas de trouver une valeur optimale de  $r$  : on trouve en effet

$$\varepsilon_D \approx 1 + \frac{r}{12} \phi^4 + \left( \frac{r}{360} + \frac{r^3}{12} \right) \phi^6 + \dots$$

Remarquons cependant que cette méthode réunit dans tous les domaines de bonnes performances : la figure suivante compare les erreurs de diffusion des trois méthodes précédentes pour  $r = 0.1, 1/6, 0.3$  et 0.5 (la comparaison n'a pas de sens pour  $r > 0.5$  puisque la méthode explicite devient instable). Pour  $r = 0.1$ , les trois méthodes sont quasi équivalentes. Pour  $r = 0.5$ , la meilleure méthode est clairement celle de Cranck-Nicholson. Pour les valeurs intermédiaires, la méthode implicite est toujours la moins bonne ; la valeur  $r = 1/6$  est particulière et favorise la méthode explicite ; néanmoins, les petites valeurs de l'erreur de diffusion ne permettent pas d'exclure la méthode de Cranck-Nicholson. Enfin, pour  $r = 0.3$ , cette dernière méthode est celle pour laquelle l'erreur de diffusion reste proche de un pour la plus grande plage de valeurs de  $\phi$ . Tout ceci justifie que cette méthode est la plus couramment utilisée.

Ces avantages « se paient » toutefois par une résolution plus délicate : l'introduction de VIII-19 et VIII-20 dans VIII-1 donne



$$\frac{u_i^{k+1} - u_i^k}{\Delta t} = \frac{\alpha}{2(\Delta x)^2} [u_{i+1}^{k+1} - 2u_i^{k+1} + u_{i-1}^{k+1} + u_{i+1}^k - 2u_i^k + u_{i-1}^k] \quad \text{VIII-28}$$

c'est-à-dire,

$$-\frac{r}{2}u_{i-1}^{k+1} + (1+r)u_i^{k+1} - \frac{r}{2}u_{i+1}^{k+1} = \frac{r}{2}u_{i-1}^k + (1-r)u_i^k + \frac{r}{2}u_{i+1}^k \quad \text{VIII-29}$$

Complétons le problème en y ajoutant les condition initiale et aux limites :

$$\text{CI : } u(x,0) = 0 \quad \text{VIII-30}$$

$$\text{CL : } u(0,t) = u_L(t) \quad \text{et} \quad u(1,t) = u_R(t), \quad \text{VIII-31}$$

Si la numérotation spatiale est décrite par  $i : 0, \dots, N$ , les CL donnent à tout instant les valeurs  $u_0^k$  et  $u_N^k \forall k$  et VIII-31 est à écrire pour  $i : 1, \dots, N-1$  et VIII-29, VIII-30 et VIII-31 donnent finalement

$$\begin{pmatrix} 1+r & -r/2 & & & \\ -r/2 & 1+r & -r/2 & & \\ & \ddots & \ddots & \ddots & \\ & & -r/2 & 1+r & -r/2 \\ & & & -r/2 & 1+r \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-2} \\ u_{N-1} \end{pmatrix}^{k+1} = \begin{pmatrix} 1-r & r/2 & & & \\ r/2 & 1-r & r/2 & & \\ & \ddots & \ddots & \ddots & \\ & & r/2 & 1-r & r/2 \\ & & & r/2 & 1-r \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-2} \\ u_{N-1} \end{pmatrix}^k + \begin{pmatrix} ru_L^k \\ 0 \\ \vdots \\ 0 \\ ru_R^k \end{pmatrix} \quad \text{VIII-32}$$

qui est un système tridiagonal à résoudre à chaque pas de temps. Le temps de calcul pour passer de l'itéré  $k$  à l'itéré  $k+1$  est donc plus long qu'avec une méthode explicite mais la stabilité

inconditionnelle permet en principe de compenser ce ralentissement en choisissant  $\Delta t$  aussi grand qu'on veut ; on est toutefois limité dans ce sens par l'augmentation de l'erreur de troncature.

## **Chapitre IX. Analyse de la stabilité par la méthode de l'équation différentielle équivalente**

Etroitement lié à la notion d'erreur de troncature (voir paragraphe VI.1), le concept d'équation différentielle équivalente vient compléter l'étude de la stabilité de Von Neumann : tout en présentant les résultats de cette étude sous un éclairage nouveau, elle permet aussi dans le cas des équations hyperboliques, de développer des algorithmes nouveaux dont l'ordre de précision soit imposé a priori, tout en préservant la notion de stabilité.

### **IX.1 Equation différentielle équivalente – Equation différentielle équivalente modifiée**

#### *Cas des problèmes paraboliques*

Reprenons l'exemple ayant servi à définir l'erreur de troncature :

$$\frac{\partial u}{\partial t} - \alpha \frac{\partial^2 u}{\partial x^2} = 0 \quad \text{IX-1}$$

et introduisons-y les schémas numériques déjà utilisés

$$\frac{\partial u}{\partial t} \rightarrow \frac{u_i^{k+1} - u_i^k}{\Delta t} \quad \text{IX-2}$$

$$\frac{\partial^2 u}{\partial x^2} \rightarrow \frac{u_{i+1}^k - 2u_i^k + u_{i-1}^k}{\Delta x^2} \quad \text{IX-3}$$

Compte tenu de ce que

$$u_i^{k+1} = u_i^k + \Delta t u_{ti} + \frac{\Delta t^2}{2} u_{tti} + \frac{\Delta t^3}{6} u_{ttti} + \dots \quad \text{IX-4}$$

$$u_{i\pm 1}^k = u_i^k \pm \Delta u_{xi} + \frac{\Delta x^2}{2} u_{xxi} \pm \frac{\Delta x^3}{6} u_{xxxxi} + \dots \quad \text{IX-5}$$

on obtient

$$\frac{u_i^{k+1} - u_i^k}{\Delta t} = u_{ti} + \frac{\Delta t}{2} u_{tti} + \frac{\Delta t^2}{6} u_{ttti} + \dots \quad \text{IX-6}$$

$$\frac{u_{i+1}^k - 2u_i^k + u_{i-1}^k}{\Delta x^2} = u_{xxi} + \frac{\Delta x^2}{12} u_{xxxxi} + \frac{\Delta x^4}{360} u_{xxxxxi} + \dots \quad \text{IX-7}$$

qui, introduits dans IX-1, donnent (en omettant l'indice i par simplicité)

$$u_t + \frac{\Delta t}{2} u_{tt} + \frac{\Delta t^2}{6} u_{ttt} + \dots - \alpha \left( u_{xx} + \frac{\Delta x^2}{12} u_{xxxx} + \frac{\Delta x^4}{360} u_{xxxxx} + \dots \right) = 0 \quad \text{IX-8}$$

ou encore

$$u_t - \alpha u_{xx} = -\frac{\Delta t}{2} u_{tt} - \frac{\Delta t^2}{6} u_{ttt} - \dots + \alpha \left( \frac{\Delta x^2}{12} u_{xxxx} + \frac{\Delta x^4}{360} u_{xxxxx} + \dots \right) \quad \text{IX-9}$$

Le second membre de IX-9 est l'erreur de troncature déjà rencontrée ; IX-8 est l'équation différentielle équivalente : c'est l'équation que vérifie exactement la solution numérique. Cette équation est habituellement modifiée de manière à éliminer de l'ET les termes de dérivées temporelles  $u_{tt}, u_{ttt}, \dots$ . Ceci est obtenu par le procédé suivant, toujours le même, quelle que soit la dérivée à éliminer : une dérivation judicieuse de IX-8 permet de générer une expression de cette dérivée. Cette expression est alors introduite dans IX-8 et on passe à l'élimination du terme suivant. Toutefois, il faut remarquer que toute dérivation de IX-8 élaborée en vue d'une élimination génère en général des dérivées nouvelles indésirables qu'il faudra aussi éliminer. Ces diverses éliminations sont menées à bien en suivant la procédure suivante.

a) on choisit a priori l'ordre de la plus grande dérivée spatiale jusqu'à laquelle on souhaite exprimer l'erreur de troncature (pour fixer les idées, choisissons  $u_{xxxxx}$  dans l'exemple traité ci-dessus et convenons d'écrire  $u_{xxxxx} = u_{6x}$  pour alléger la notation).

b) on construit alors ligne par ligne le tableau suivant.

La première ligne reprend toutes les dérivées temporelles, mixtes et spatiales possibles, rangées par ordre croissant de dérivation, depuis les dérivées premières jusqu'aux dérivées de même ordre que celui choisi en a) et, pour un même ordre, rangées en ordre décroissant pour la dérivation temporelle et croissant pour la dérivation spatiale. Pour l'exemple étudié, cette première ligne s'écrit donc :

$$u_t \ u_x \ u_{2t} \ u_{tx} \ u_{2x} \ u_{3t} \ u_{2tx} \ u_{t2x} \ u_{3x} \dots u_{6t} \dots u_{6x} \quad \text{IX-10}$$

La deuxième ligne reprend les coefficients des dérivées présentes dans l'équation différentielle équivalente parmi celles de la première ligne. Pour l'exemple étudié cela génère, si on omet les dérivées absentes de cette équation :

$$\begin{array}{cccccccccc} u_t & u_{2t} & u_{2x} & u_{3t} & u_{4t} & u_{4x} & u_{5t} & u_{6t} & u_{6x} \\ 1 & \frac{\Delta t}{2} & -\alpha & \frac{\Delta t^2}{6} & \frac{\Delta t^3}{24} & -\alpha \frac{\Delta x^2}{12} & \frac{\Delta t^4}{120} & \frac{\Delta t^5}{720} & -\alpha \frac{\Delta x^4}{360} \end{array} \quad \text{IX-11}$$

La troisième ligne est celle qui génère l'élimination de la première dérivée gênante. Cette élimination est obtenue en « multipliant » la ligne précédente par l'opérateur  $k \frac{\partial^{n_1+n_2}}{\partial t^{n_1} \partial x^{n_2}}$  où  $k, n_1$  et  $n_2$  sont choisis de manière à générer l'élimination souhaitée par application de l'opérateur au terme en  $u_t$  de la deuxième ligne. Pour l'exemple étudié, la première dérivée à éliminer est le terme  $\frac{\Delta t}{2} u_{2t}$ . L'opérateur à appliquer à  $u_t$  est donc  $-\frac{\Delta t}{2} \frac{\partial}{\partial t}$ .

Le tableau IX-11 devient donc (en ne retenant que les colonnes utiles de la première ligne)

$$\begin{array}{ccccccc} u_{2t} & u_{3t} & u_{t2x} & u_{4t} & u_{5t} & u_{6t} \\ -\frac{\Delta t}{2} & -\frac{\Delta t^2}{4} & \alpha \frac{\Delta t}{2} & -\frac{\Delta t^3}{12} & -\frac{\Delta t^4}{48} & -\frac{\Delta t^5}{240} \end{array} \quad \text{IX-12}$$

A ce stade, deux remarques s'imposent :

- on voit apparaître des termes n'existant pas jusque là ( $\alpha \frac{\Delta t}{2} u_{t2x}$ ) et qu'il faudra aussi éliminer,
- on s'aperçoit qu'il n'est pas nécessaire lors de l'écriture de la première ligne du tableau de prévoir toutes les combinaisons possibles de dérivées temporelles, mixtes et spatiales : pour l'exemple étudié, on n'aura jamais de termes avec un ordre impair de dérivée spatiale puisque le terme en  $u_x$  est absent de l'équation initiale.

La quatrième ligne génère l'élimination du deuxième terme gênant, soit pour l'exemple traité  $(\frac{\Delta t^2}{6} - \frac{\Delta t^2}{4})u_{3t}$  qui provient de l'équation de départ et de l'élimination précédente. L'opérateur à utiliser est donc  $\frac{\Delta t^2}{12} \frac{\partial^2}{\partial t^2}$ .

Le tableau complet pour l'exemple étudié est repris ci-après, en ne retenant que les colonnes utiles : celles où la dérivation spatiale est paire. Remarquons encore qu'on ajoute une colonne supplémentaire à la gauche du tableau où on fait figurer les divers opérateurs utilisés.

Finalement, l'équation différentielle équivalente modifiée s'écrit

$$u_t - \alpha u_{xx} = -(\alpha^2 \frac{\Delta t}{2} - \alpha \frac{\Delta x^2}{12})u_{4x} - (-\alpha \frac{\Delta x^4}{360} + \alpha^2 \frac{\Delta x^2 \Delta t}{24} + \alpha^2 \frac{\Delta x^2 \Delta t}{24} - \alpha^3 \frac{\Delta t^2}{3})u_{6x} + \dots \quad \text{IX-13}$$

ou encore, avec  $r = \alpha \frac{\Delta t}{\Delta x^2}$ ,

$$u_t - \alpha u_{xx} = -\alpha \frac{\Delta x^2}{2} (r - \frac{1}{6})u_{4x} + \alpha \frac{\Delta x^4}{3} (r^2 - \frac{r}{4} + \frac{1}{120})u_{6x} + \dots \quad \text{IX-14}$$

On retrouve la valeur optimale  $r = \frac{1}{6}$  déjà rencontrée au paragraphe VIII.1 qui permet ici d'annuler le premier terme de l'erreur de troncature. Remarquons aussi que cette dernière ne contient que des termes de dérivées spatiales paires : ainsi qu'on le verra plus en détail lors de l'étude des équations hyperboliques, cette caractéristique traduit le caractère purement diffusif, et non dispersif, des erreurs introduites par le schéma étudié.

	$u_t$	$u_{2t}$	$u_{2x}$	$u_{3t}$	$u_{t2x}$	$u_{4t}$	$u_{2t2x}$	$u_{4x}$	$u_{5t}$	$u_{3t2x}$	$u_{t4x}$	$u_{6t}$	$u_{4t2x}$	$u_{2t4x}$	$u_{6x}$
	$1$	$\frac{\Delta t}{2}$	$-\alpha$	$\frac{\Delta t^2}{6}$		$\frac{\Delta t^3}{24}$		$-\alpha \frac{\Delta x^2}{12}$	$\frac{\Delta t^4}{120}$			$\frac{\Delta t^5}{720}$			$-\alpha \frac{\Delta x^4}{360}$
$-\frac{\Delta t}{2} \frac{\partial}{\partial t}$		$-\frac{\Delta t}{2}$		$-\frac{\Delta t^2}{4}$	$\alpha \frac{\Delta t}{2}$	$-\frac{\Delta t^3}{12}$			$-\frac{\Delta t^4}{48}$		$\alpha \frac{\Delta t \Delta x^2}{24}$	$-\frac{\Delta t^5}{240}$			
$\frac{\Delta t^2}{12} \frac{\partial^2}{\partial t^2}$				$\frac{\Delta t^2}{12}$		$\frac{\Delta t^3}{24}$	$-\alpha \frac{\Delta t^2}{12}$		$\frac{\Delta t^4}{72}$			$\frac{\Delta t^5}{288}$		$-\frac{\alpha \Delta t^2 \Delta x^2}{144}$	
$-\alpha \frac{\Delta t}{2} \frac{\partial^2}{\partial x^2}$					$-\alpha \frac{\Delta t}{2}$		$-\alpha \frac{\Delta t^2}{4}$	$\alpha^2 \frac{\Delta t}{2}$		$-\alpha \frac{\Delta t^3}{12}$			$-\alpha \frac{\Delta t^4}{48}$		$\alpha^2 \frac{\Delta t \Delta x^2}{24}$
$\alpha \frac{\Delta t^2}{3} \frac{\partial^3}{\partial x^2 \partial t}$						$\alpha \frac{\Delta t^2}{3}$			$\alpha \frac{\Delta t^3}{6}$	$-\alpha^2 \frac{\Delta t^2}{3}$			$\alpha \frac{\Delta t^4}{18}$		
$-\frac{\Delta t^4}{720} \frac{\partial^4}{\partial t^4}$									$-\frac{\Delta t^4}{720}$			$-\frac{\Delta t^5}{720}$	$\alpha \frac{\Delta t^4}{720}$		
$-\alpha \frac{\Delta t^3}{12} \frac{\partial^4}{\partial x^2 \partial t^2}$									$-\alpha \frac{\Delta t^3}{12}$				$-\alpha \frac{\Delta t^4}{24}$	$\alpha^2 \frac{\Delta t^3}{12}$	
$-(\alpha \frac{\Delta t \Delta x^2}{24} - \alpha^2 \frac{\Delta t^2}{3}) \frac{\partial^4}{\partial x^4}$										$-(\alpha \frac{\Delta t \Delta x^2}{24} - \alpha^2 \frac{\Delta t^2}{3})$			$-(\alpha \frac{\Delta t \Delta x^2}{24} - \alpha^2 \frac{\Delta t^2}{3}) \frac{\Delta t}{2}$	$\alpha(\alpha \frac{\Delta t \Delta x^2}{24} - \alpha^2 \frac{\Delta t^2}{3})$	
$\Delta t^5 (\frac{1}{240} - \frac{1}{288}) \frac{\partial^5}{\partial t^5}$												$\Delta t^5 (\frac{1}{240} - \frac{1}{288})$			

### Cas des problèmes hyperboliques

Considérons l'équation d'advection

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = 0 \quad \text{IX-15}$$

qu'on résout par la méthode d'Euler stabilisée (voir paragraphe VII.2) :

$$\frac{u_i^{k+1} - u_i^k}{\Delta t} + a \frac{u_i^k - u_{i-1}^k}{\Delta x} = 0 \quad \text{IX-16}$$

Un raisonnement analogue à celui qui est fait au paragraphe précédent nous fournit l'équation différentielle équivalente :

$$u_t + au_x - a \frac{\Delta x}{2} u_{xx} + a \frac{\Delta x^2}{6} u_{xxx} - \dots + \frac{\Delta t}{2} u_{tt} + \frac{\Delta t^2}{6} u_{ttt} + \dots = 0 \quad \text{IX-17}$$

De la même manière, l'équation différentielle modifiée résulte du tableau repris plus loin ; cette équation s'écrit :

$$u_t + au_x + (a^2 \frac{\Delta t}{2} - a \frac{\Delta x}{2}) u_{xx} + (a^3 \frac{\Delta t^2}{3} - a^2 \frac{\Delta t \Delta x}{2} + a \frac{\Delta x^2}{6}) u_{xxx} + \dots = 0 \quad \text{IX-18}$$

c'est-à-dire, avec  $v = \frac{a \Delta t}{\Delta x}$  :

$$u_t + au_x + \frac{a \Delta x}{2} [v - 1] u_{xx} + \frac{a \Delta x^2}{6} [2v^2 - 3v + 1] u_{xxx} + \dots = 0 \quad \text{IX-19}$$

L'erreur de troncature vaut donc

$$ET = - \frac{a \Delta x}{2} [v - 1] u_{xx} - \frac{a \Delta x^2}{6} [2v^2 - 3v + 1] u_{xxx} \dots \quad \text{IX-20}$$

Si le processus était poursuivi, on s'apercevrait de ce que tous les termes de cette erreur s'annulent pour  $v = 1$  : on retrouve ici la conclusion essentielle déduite de l'étude menée au paragraphe VII.2 : pour  $v = 1$ , la résolution de l'équation d'advection est parfaite : le schéma numérique IX-16 génère des itérés vérifiant strictement  $u_t + au_x = 0$ . Le schéma lui-même se réduit à

$$u_i^{k+1} - u_i^k + \frac{a \Delta t}{\Delta x} (u_i^k - u_{i-1}^k) = 0 \quad \text{IX-21}$$

c'est-à-dire

$$u_i^{k+1} = u_{i-1}^k \quad \text{IX-22}$$

On dit des schémas de différences finies qui vérifient IX-22 qu'ils satisfont la « shift condition ».

Quand  $v$  est différent de 1, le premier terme de l'erreur de troncature est proportionnel à  $u_{xx}$ . Un tel terme en mécanique des fluides est représentatif de la présence d'une certaine viscosité. C'est la raison pour laquelle on dit que le schéma IX-16 introduit de la viscosité artificielle dans la solution. Cette viscosité artificielle tend à réduire les gradients présents dans la solution, que ces gradients correspondent à une réalité physique des phénomènes décrits par l'équation, ou qu'ils soient numériquement introduits. Cet effet, qui est le résultat direct de la présence de dérivées paires dans l'erreur de troncature modifiée a déjà été rencontré : il est appelé diffusion.

L'effet de cette viscosité artificielle, dite aussi viscosité numérique peut être observé et comparé sur divers schémas :

a) pour la méthode d'Euler, on vient de voir qu'elle était égale à

$$v_{\text{num},E} = -a \frac{\Delta x}{2} (v - 1) = \frac{1}{2} \frac{a \Delta x^2}{\Delta t} \frac{\Delta t}{\Delta x} (1 - v) = \frac{1}{2} \frac{\Delta x^2}{\Delta t} v (1 - v)$$

Cette viscosité génère une stabilisation de la simulation numérique pour peu qu'elle soit positive, c'est-à-dire pour peu qu'on ait  $0 \leq v \leq 1$ , ce qui est conforme à l'analyse de Von Neumann.

b) pour la méthode de Lax, le calcul de l'équation différentielle équivalente modifiée conduit à une viscosité numérique égale à

$$v_{\text{num},L} = \frac{1}{2} \frac{\Delta x^2}{\Delta t} (1 - v^2)$$

On a donc

$$\frac{v_{\text{num},L}}{v_{\text{num},E}} = \frac{1 + v}{v} > 1 \text{ car } 0 \leq v \leq 1$$

On doit donc s'attendre à ce que la diffusion du schéma de Lax soit plus importante que celle du schéma d'Euler.

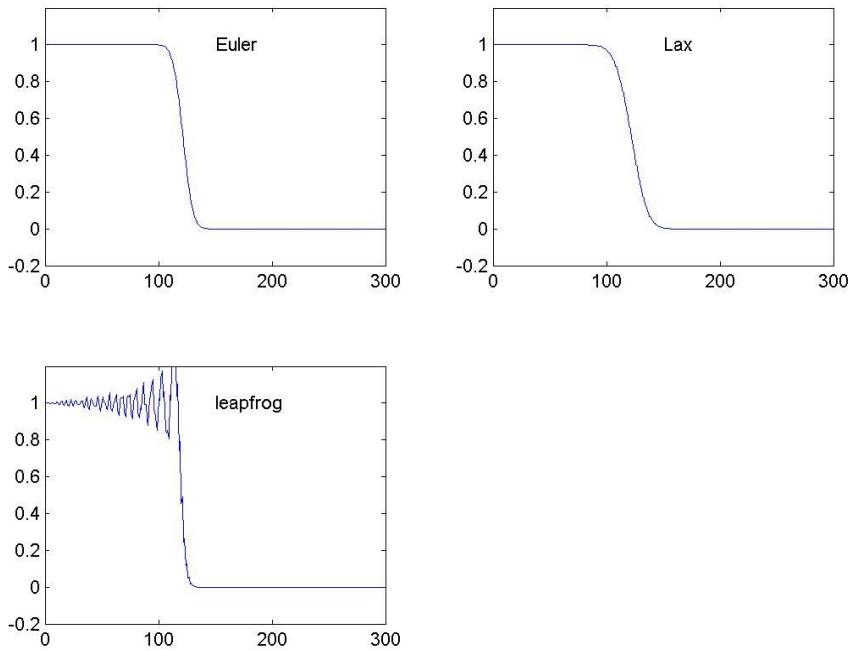
Un autre effet quasi physique des schémas numériques est généré par la présence de termes de dérivées impaires dans l'erreur de troncature modifiée : c'est la dispersion également déjà rencontrée. Elle a pour effet de distordre le spectre des phases des ondes présentes dans le spectre de la solution. Quand la CI comporte une discontinuité, la diffusion a tendance à amollir le profil, tandis que la dispersion génère des oscillations.

À cet égard, le calcul de l'équation différentielle équivalente modifiée du schéma leapfrog débouche sur la relation

$$u_t + au_x = -\frac{\Delta x^3}{6\Delta t} v (1 - v^2) u_{3x} \dots$$

Le premier terme de l'erreur de troncature étant proportionnel à une dérivée impaire, il faut s'attendre à des oscillations quand  $v$  est différent de un.

La figure suivante confirme les conclusions annoncées pour ces trois schémas : on y a représenté les solutions de l'équation d'advection avec une condition initiale présentant un échelon négatif (les simulations sont faites avec la valeur commune  $v = 0.6$ ) :



Ajoutons encore que d'une manière générale, l'erreur de troncature modifiée comportant des termes de dérivées paires et impaires, un schéma n'est jamais purement diffusif ou dispersif : c'est la parité de la dérivée du premier terme de l'erreur de troncature modifiée qui fixe le comportement prédominant du schéma numérique : il sera plutôt diffusif ou plutôt dispersif selon cette parité.

	$u_t$	$u_x$	$u_{tt}$	$u_{tx}$	$u_{xx}$	$u_{ttt}$	$u_{txx}$	$u_{txx}$	$u_{xxx}$
	1	$a$	$\frac{\Delta t}{2}$		$-\frac{a\Delta x}{2}$	$\frac{\Delta t^2}{6}$			$\frac{a\Delta x^2}{6}$
$-\frac{\Delta t}{2} \frac{\partial}{\partial t}$			$-\frac{\Delta t}{2}$	$-a \frac{\Delta t}{2}$		$-\frac{\Delta t^2}{4}$		$a \frac{\Delta x \Delta t}{4}$	
$a \frac{\Delta t}{2} \frac{\partial}{\partial x}$				$a \frac{\Delta t}{2}$	$a^2 \frac{\Delta t}{2}$		$a \frac{\Delta t^2}{4}$		$-a^2 \frac{\Delta t \Delta x}{4}$
$\frac{\Delta t^2}{12} \frac{\partial^2}{\partial t^2}$						$\frac{\Delta t^2}{12}$	$a \frac{\Delta t^2}{12}$		
$-\frac{a\Delta t^2}{3} \frac{\partial^2}{\partial t \partial x}$							$-a \frac{\Delta t^2}{3}$	$-a^2 \frac{\Delta t^2}{3}$	
$(\frac{a^2 \Delta t^2}{3} - \frac{a\Delta t \Delta x}{4}) \frac{\partial^2}{\partial x^2}$							$\frac{a^2 \Delta t^2}{3} - \frac{a\Delta t \Delta x}{4}$		$(\frac{a^2 \Delta t^2}{3} - \frac{a\Delta t \Delta x}{4})a$

## IX.2 Formulation générale de l'équation différentielle équivalente modifiée

Si on observe que l'ordre de précision en  $\Delta x$  de l'erreur de troncature est inchangé quand on élimine les termes en  $u_{mt}$ , ceci permet a priori d'écrire l'équation différentielle équivalente modifiée sous la forme

$$u_t + au_x = \sum_{l=p}^{\infty} a_{l+1} \Delta x^l u_{(l+1)x} \quad \text{IX-23}$$

Les coefficients  $a_{l+1}$  de cette expression peuvent être obtenus plus simplement qu'en dressant les tableaux qui précèdent.

### Cas des équations hyperboliques

Attirons l'attention sur le fait que les détails de calculs qui suivent ne valent que pour un schéma numérique explicite à deux étages temporels, tout autre schéma (implicite et/ou à plus de 2 étages) pouvant être traité selon la même démarche.

L'équation aux différences finies de tout schéma explicite à deux niveaux de temps s'écrit

$$u_i^{k+1} = \sum_j b_j u_{i+j}^k \quad \text{IX-24}$$

où la somme sur  $j$  définit les points de grille utilisés dans le schéma (par exemple, pour IX-16,  $j = -1$  et  $0$ ).

L'obtention de l'équation différentielle équivalente modifiée résulte de la généralisation des étapes évoquées lors de l'établissement de cette formule dans les deux exemples traités ci-avant : on remplace d'abord  $u_i^{k+1}$  et  $u_{i+j}^k$  par leurs développements de Taylor :

$$u_i^{k+1} = u_i^k + \sum_{m=1}^{\infty} \frac{\Delta t^m}{m!} u_{mt} \quad \text{IX-25}$$

$$u_{i+j}^k = u_i^k + \sum_{m=1}^{\infty} \frac{(j\Delta x)^m}{m!} u_{mx} \quad \text{IX-26}$$

ce qui donne avec IX-24

$$u_i^k + \sum_{m=1}^{\infty} \frac{\Delta t^m}{m!} u_{mt} = \sum_j b_j \left[ u_i^k + \sum_{m=1}^{\infty} \frac{(j\Delta x)^m}{m!} u_{mx} \right] \quad \text{IX-27}$$

Remarquons au passage que l'obtention de l'équation différentielle équivalente génère des conditions, dites de consistance, que les  $b_j$  doivent satisfaire : tout d'abord, les coefficients de  $u_i^k$  dans les deux membres de IX-27 doivent être identiques : il faut

$$\sum_j b_j = 1 \quad \text{IX-28}$$

IX-28 étant acquis, il faut encore que  $u_x$  et  $u_t$  soient liés par l'équation à résoudre :

$$u_t + au_x = 0$$

IX-29

ceci exige d'avoir

$$\frac{1}{\Delta t} \sum_j b_j \left[ \frac{j \Delta x}{1!} \right] = -a \quad \text{IX-30}$$

c'est-à-dire

$$\sum_j b_j j = -v \quad \text{IX-31}$$

IX-28 et IX-31 étant acquis, IX-27 devient

$$u_t + au_x = \frac{1}{\Delta t} \sum_j b_j \left[ \sum_{m=2}^{\infty} \frac{(j \Delta x)^m}{m!} u_{nx} \right] - \sum_{m=2}^{\infty} \frac{(\Delta t)^{m-1}}{m!} u_{nt} \quad \text{IX-32}$$

L'équation différentielle équivalente modifiée est alors obtenue en éliminant les termes en  $u_{nt}$ . Cette élimination sera obtenue en utilisant IX-23 réécrit selon

$$u_t = -au_x + \sum_{l_i=p}^{\infty} a_{l_i+1} \Delta x^{l_i} u_{(l_i+1)x} \quad \text{IX-33}$$

qui s'écrit aussi symboliquement

$$u_t = \left[ -a \frac{\partial}{\partial x} + \sum_{l_i=p}^{\infty} a_{l_i+1} \Delta x^{l_i} \frac{\partial^{l_i+1}}{\partial x^{l_i+1}} \right] u \quad \text{IX-34}$$

ce qui fournit

$$u_{nt} = \left[ -a \frac{\partial}{\partial x} + \sum_{l_i=p}^{\infty} a_{l_i+1} \Delta x^{l_i} \frac{\partial^{l_i+1}}{\partial x^{l_i+1}} \right]^m u \quad \text{IX-35}$$

$$u_{nt} = [C_m^0 (-a)^m \frac{\partial^m}{\partial x^m} + C_m^1 (-a)^{m-1} \frac{\partial^{m-1}}{\partial x^{m-1}} \sum_{l_i=p}^{\infty} a_{l_i+1} \Delta x^{l_i} \frac{\partial^{l_i+1}}{\partial x^{l_i+1}} + \\ C_m^2 (-a)^{m-2} \frac{\partial^{m-2}}{\partial x^{m-2}} \sum_{l_1, l_2=p}^{\infty} a_{l_1+1} a_{l_2+1} \Delta x^{l_1+l_2} \frac{\partial^{l_1+l_2+2}}{\partial x^{l_1+l_2+2}} + \dots] u$$

ou encore

$$u_{nt} = [C_m^0 (-a)^m \frac{\partial^m}{\partial x^m} + C_m^1 (-a)^{m-1} \sum_{l_i=p}^{\infty} a_{l_i+1} \Delta x^{l_i} \frac{\partial^{l_i+m}}{\partial x^{l_i+m}} + \\ C_m^2 (-a)^{m-2} \sum_{l_1, l_2=p}^{\infty} a_{l_1+1} a_{l_2+1} \Delta x^{l_1+l_2} \frac{\partial^{l_1+l_2+m}}{\partial x^{l_1+l_2+m}} + C_m^3 (-a)^{m-3} \sum_{l_1, l_2, l_3=p}^{\infty} a_{l_1+1} a_{l_2+1} a_{l_3+1} \Delta x^{l_1+l_2+l_3} \frac{\partial^{l_1+l_2+l_3+m}}{\partial x^{l_1+l_2+l_3+m}} + \dots] u$$

IX-36

Le remplacement des dérivées temporelles dans IX-32 fournit alors

$$\begin{aligned}
 u_t + au_x &= \frac{\Delta x}{\Delta t} \sum_{m=2}^{\infty} \left[ \left[ \sum_j b_j j^m - (-v)^m \right] \frac{\Delta x^{m-1}}{m!} \right] u_{mx} - \sum_{m=2}^{\infty} \frac{(-v)^{m-1}}{l!(m-1)!} \sum_{l_1=p}^{\infty} a_{l_1+1} \Delta x^{l_1+m-1} u_{(l_1+m)x} \\
 &\quad - \frac{\Delta t}{\Delta x} \sum_{m=2}^{\infty} \frac{(-v)^{m-2}}{2!(m-2)!} \sum_{l_1, l_2=p}^{\infty} a_{l_1+1} a_{l_2+1} \Delta x^{l_1+l_2+m-1} u_{(l_1+l_2+m)x} \\
 &\quad - \left( \frac{\Delta t}{\Delta x} \right)^2 \sum_{m=3}^{\infty} \frac{(-v)^{m-3}}{3!(m-3)!} \sum_{l_1, l_2, l_3=p}^{\infty} a_{l_1+1} a_{l_2+1} a_{l_3+1} \Delta x^{l_1+l_2+l_3+m-1} u_{(l_1+l_2+l_3+m)x} \\
 &\quad - \dots
 \end{aligned}$$

IX - 37

La comparaison avec IX-23 nous fournit des relations de consistance supplémentaires : la méthode étant d'ordre  $p$ , tous les termes en  $\Delta x, \dots, \Delta x^{p-1}$  doivent être nuls :

$$\sum_j b_j j^m - (-v)^m = 0 \quad \text{pour } m = 2, 3, \dots, p \quad \text{IX-38}$$

Associées à IX-28 et à IX-31, ces relations forment les  $m+1$  conditions de consistance qui se mettent sous la forme générale

$$\sum_j b_j j^m = (-v)^m \quad \text{pour } m = 0, 1, \dots, p \quad \text{IX-39}$$

Celles-ci étant acquises, la première sommation de IX-37 démarre à  $m = p + 1$  et la comparaison avec IX-23 permet de calculer les coefficients  $a_{l+1}$  de l'erreur de troncature. Ce calcul consistant à comparer les coefficients des termes de mêmes puissances en  $\Delta x$  dans les deux formules, on observe immédiatement que les résultats vont dépendre de l'ordre  $p$  : selon la valeur de ce paramètre, les quatre sommes de IX-37 vont contribuer de manière différente aux termes  $\Delta x^k u_{(k+1)x}$ . Plus précisément, le calcul détaillé montre que les deux premiers termes de l'erreur de troncature sont fournis par des expressions indépendantes de  $p$  et qu'au-delà, les expressions trouvées en dépendent. On trouve ainsi :

premier terme :

$$\forall p : a_{p+1} = \frac{\Delta x}{\Delta t} \left[ \sum_j b_j j^{p+1} - (-v)^{p+1} \right] \frac{1}{(p+1)!} \quad \text{IX-40}$$

deuxième terme :

$$\forall p : a_{p+2} = \frac{\Delta x}{\Delta t} \left[ \sum_j b_j j^{p+2} - (-v)^{p+2} \right] \frac{1}{(p+2)!} - \frac{(-v)^1}{1!!} a_{p+1} \quad \text{IX-41}$$

troisième terme :

$$p=1 : \quad a_{p+3} = \frac{\Delta x}{\Delta t} \left[ \sum_j b_j j^{p+3} - (-v)^{p+3} \right] \frac{1}{(p+3)!} - \left[ \frac{(-v)^2}{1!2!} a_{p+1} + \frac{(-v)^1}{1!1!} a_{p+2} \right] - \frac{\Delta t}{\Delta x} \frac{(-v)^0}{2!0!} a_{p+1}^2$$

IX-42

$$p>1 : \quad a_{p+3} = \frac{\Delta x}{\Delta t} \left[ \sum_j b_j j^{p+3} - (-v)^{p+3} \right] \frac{1}{(p+3)!} - \left[ \frac{(-v)^2}{1!2!} a_{p+1} + \frac{(-v)^1}{1!1!} a_{p+2} \right] \quad \text{IX-43}$$

quatrième terme :

$p=1$  :

$$\begin{aligned} a_{p+4} &= \frac{\Delta x}{\Delta t} \left[ \sum_j b_j j^{p+4} - (-v)^{p+4} \right] \frac{1}{(p+4)!} - \left[ \frac{(-v)^3}{1!3!} a_{p+1} + \frac{(-v)^2}{1!2!} a_{p+2} + \frac{(-v)^1}{1!1!} a_{p+3} \right] \\ &\quad - \frac{\Delta t}{\Delta x} \left[ \frac{(-v)^1}{2!1!} a_{p+1}^2 + \frac{(-v)^0}{2!0!} 2a_{p+2}a_{p+1} \right] \end{aligned} \quad \text{IX-44}$$

$p=2$  :

$$\begin{aligned} a_{p+4} &= \frac{\Delta x}{\Delta t} \left[ \sum_j b_j j^{p+4} - (-v)^{p+4} \right] \frac{1}{(p+4)!} - \left[ \frac{(-v)^3}{1!3!} a_{p+1} + \frac{(-v)^2}{1!2!} a_{p+2} + \frac{(-v)^1}{1!1!} a_{p+3} \right] \\ &\quad - \frac{\Delta t}{\Delta x} \left[ \frac{(-v)^0}{2!0!} 2a_{p+2}a_{p+1} \right] \end{aligned} \quad \text{IX-45}$$

$p>2$  :

$$a_{p+4} = \frac{\Delta x}{\Delta t} \left[ \sum_j b_j j^{p+4} - (-v)^{p+4} \right] \frac{1}{(p+4)!} - \left[ \frac{(-v)^3}{1!3!} a_{p+1} + \frac{(-v)^2}{1!2!} a_{p+2} + \frac{(-v)^1}{1!1!} a_{p+3} \right] \quad \text{IX-46}$$

cinquième terme :

$p=1$  :

$$\begin{aligned} a_{p+5} &= \frac{\Delta x}{\Delta t} \left[ \sum_j b_j j^{p+5} - (-v)^{p+5} \right] \frac{1}{(p+5)!} - \left[ \frac{(-v)^4}{1!4!} a_{p+1} + \frac{(-v)^3}{1!3!} a_{p+2} + \frac{(-v)^2}{1!2!} a_{p+3} + \frac{(-v)^1}{1!1!} a_{p+4} \right] \\ &\quad - \frac{\Delta t}{\Delta x} \left[ \frac{(-v)^2}{2!2!} a_{p+1}^2 + \frac{(-v)^1}{2!1!} 2a_{p+2}a_{p+1} + \frac{(-v)^0}{2!0!} (2a_{p+3}a_{p+1} + a_{p+2}^2) \right] - \left( \frac{\Delta t}{\Delta x} \right)^2 \frac{(-v)^0}{3!0!} a_{p+1}^3 \end{aligned} \quad \text{IX-47}$$

$p=2$  :

$$\begin{aligned} a_{p+5} &= \frac{\Delta x}{\Delta t} \left[ \sum_j b_j j^{p+5} - (-v)^{p+5} \right] \frac{1}{(p+5)!} - \left[ \frac{(-v)^4}{1!4!} a_{p+1} + \frac{(-v)^3}{1!3!} a_{p+2} + \frac{(-v)^2}{1!2!} a_{p+3} + \frac{(-v)^1}{1!1!} a_{p+4} \right] \\ &\quad - \frac{\Delta t}{\Delta x} \left[ \frac{(-v)^1}{2!1!} a_{p+1}^2 + \frac{(-v)^0}{2!0!} 2a_{p+2}a_{p+1} \right] \end{aligned} \quad \text{IX-48}$$

p=3 :

$$a_{p+5} = \frac{\Delta x}{\Delta t} \left[ \sum_j b_j j^{p+5} - (-v)^{p+5} \right] \frac{1}{(p+5)!} - \left[ \frac{(-v)^4}{1!4!} a_{p+1} + \frac{(-v)^3}{1!3!} a_{p+2} + \frac{(-v)^2}{1!2!} a_{p+3} + \frac{(-v)^1}{1!1!} a_{p+4} \right] - \frac{\Delta t}{\Delta x} \left[ \frac{(-v)^0}{2!0!} a_{p+1}^2 \right] \quad \text{IX-49}$$

p>3 :

$$a_{p+5} = \frac{\Delta x}{\Delta t} \left[ \sum_j b_j j^{p+5} - (-v)^{p+5} \right] \frac{1}{(p+5)!} - \left[ \frac{(-v)^4}{1!4!} a_{p+1} + \frac{(-v)^3}{1!3!} a_{p+2} + \frac{(-v)^2}{1!2!} a_{p+3} + \frac{(-v)^1}{1!1!} a_{p+4} \right] \quad \text{IX-50}$$

Remarquons que c'est lors du calcul de ce cinquième terme que la troisième somme de IX-37 intervient pour la première fois ; la quatrième somme (non mentionnée) de cette formule interviendra pour la première fois lors du calcul de  $a_{p+7}$ , avec p=1.

$$\text{Exemple : reprenons IX-16 : } u_i^{k+1} = u_i^k - v(u_i^k - u_{i-1}^k) \quad \text{IX-51}$$

$$\text{IX-24 s'écrit ici } u_i^{k+1} = vu_{i-1}^k + (1-v)u_i^k \quad \text{IX-52}$$

ce qui donne

$$b_{-1} = v \quad b_0 = 1 - v$$

Le schéma étant d'ordre 1, on vérifie aisément que les conditions de consistance IX-39 sont satisfaites. L'application des formules IX-40 à IX-42 conduit rapidement à

$$u_t + au_x = \frac{a}{2}(1-v)\Delta xu_{2x} + \frac{a}{6}(1-v)(2v-1)\Delta x^2 u_{3x} + \frac{a}{24}(1-v)(1-6v+6v^2)\Delta x^3 u_{4x} \quad \text{IX-53}$$

qui confirme IX-20.

### Cas des équations paraboliques

Les relations qui étaient à la base des développements relatifs aux équations hyperboliques sont IX-23 et IX-24. Dans le cas des équations paraboliques, le formalisme IX-23 pour l'équation différentielle équivalente modifiée doit être corrigé : l'équation utilisée est

$$u_t - \alpha u_{xx} = 0 \quad \text{IX-54}$$

En outre, tout schéma de différences finies permettant l'estimation de  $u_{xx}$  a une erreur de troncature dont le premier terme est « au pire » proportionnel à  $u_{3x}\Delta x$ . Ceci nous indique que si la méthode est d'ordre p en  $\Delta x$ , IX-23 devient, pour IX-54,

$$u_t - \alpha u_{2x} = \sum_{l=p}^{\infty} a_{l+2} \Delta x^l u_{(l+2)x} \quad \text{IX-55}$$

Quant à IX-24, il est inchangé si on se limite, comme on l'a fait pour les équations hyperboliques, aux schémas explicites à deux niveaux de temps :

$$u_i^{k+1} = \sum_j b_j u_{i+j}^k \quad \text{IX-56}$$

Il en résulte que IX-27 reste d'application :

$$u + \Delta t u_t + \frac{\Delta t^2}{2} u_{2t} + \frac{\Delta t^3}{6} u_{3t} + \dots = \sum_j b_j \left[ u + (j\Delta x)u_x + \frac{(j\Delta x)^2}{2} u_{2x} + \frac{(j\Delta x)^3}{6} u_{3x} + \dots \right] \quad \text{IX-57}$$

Les premières conditions de consistance sont différentes de celles des équations hyperboliques : si l'égalité des termes en  $u$  impose toujours

$$\sum_j b_j = 1 \quad \text{IX-58}$$

l'absence de terme en  $u_x$  dans l'ET implique

$$\sum_j b_j j = 0 \quad \text{IX-59}$$

et l'identification des coefficients des termes en  $u_t$  et en  $u_{xx}$  avec IX-54 impose

$$\frac{1}{\Delta t} \sum_j b_j \frac{(j\Delta x)^2}{2!} = \alpha$$

c'est-à-dire, avec  $r = \frac{\alpha \Delta t}{\Delta x^2}$ ,

$$\sum_j b_j j^2 = 2!r \quad \text{IX-60}$$

IX-58 à IX-60 étant acquis, l'équivalent de IX-32 s'écrit

$$u_t - \alpha u_{xx} = \frac{1}{\Delta t} \sum_j b_j \left[ \sum_{m=3}^{\infty} \frac{(j\Delta x)^m}{m!} u_{mx} \right] - \sum_{m=2}^{\infty} \frac{(\Delta t)^{m-1}}{m!} u_{mt} \quad \text{IX-61}$$

De IX-55 on tire alors

$$u_{mt} = \left[ \alpha \frac{\partial^2}{\partial x^2} + \sum_{l_i=p}^{\infty} a_{l_i+2} \Delta x^{l_i} \frac{\partial^{l_i+2}}{\partial x^{l_i+2}} \right]^m u \quad \text{IX-62}$$

Des développements analogues à ceux qui ont conduit à IX-37 donnent alors

$$\begin{aligned}
u_t - \alpha u_{xx} &= \frac{\Delta x^2}{\Delta t} \sum_{m=1}^{\infty} \left[ \sum_j b_j j^{2m+1} \right] \frac{\Delta x^{2m-1}}{(2m+1)!} u_{(2m+1)x} + \frac{\Delta x^2}{\Delta t} \sum_{m=2}^{\infty} \left[ \frac{1}{(2m)!} \sum_j b_j j^{2m} - \frac{r^m}{m!} \right] \Delta x^{2m-2} u_{2mx} \\
&\quad - \sum_{m=2}^{\infty} \frac{r^{m-1}}{l!(m-1)!} \sum_{l_1=p}^{\infty} a_{l_1+2} \Delta x^{l_1+2m-2} u_{(l_1+2m)x} \\
&\quad - \left( \frac{\Delta t}{\Delta x^2} \right)^2 \sum_{m=3}^{\infty} \frac{r^{m-3}}{3!(m-3)!} \sum_{l_1,l_2,l_3=p}^{\infty} a_{l_1+2} a_{l_2+2} a_{l_3+2} \Delta x^{l_1+l_2+l_3+2m-2} u_{(l_1+l_2+l_3+2m)x} \\
&\quad - \dots
\end{aligned}
\tag{IX-63}$$

En tenant compte de ce que la méthode est d'ordre p, on obtient, en incluant IX-58 à IX-60

Méthode d'ordre p pair :

$$\sum_j b_j j^{2m+1} = 0 \quad m = 0, 1, \dots, \frac{p}{2}
\tag{IX-64}$$

$$\sum_j b_j j^{2m} = \frac{(2m)!}{m!} r^m \quad m = 0, 1, \dots, \frac{p}{2} - 1
\tag{IX-65}$$

Méthode d'ordre p impair :

$$\sum_j b_j j^{2m+1} = 0 \quad m = 0, 1, \dots, \frac{p+1}{2}
\tag{IX-66}$$

$$\sum_j b_j j^{2m} = \frac{(2m)!}{m!} r^m \quad m = 0, 1, \dots, \frac{p+1}{2}
\tag{IX-67}$$

L'introduction de IX-64 à IX-67 dans IX-63 permet alors de déduire les relations de calcul des termes de l'erreur de troncature. Celles-ci dépendent de la parité de l'ordre de la méthode et, pour en faciliter la détermination, il est commode de détailler IX-63 en explicitant les sommes qui y interviennent et, en particulier, de regrouper les termes de mêmes puissances en  $\Delta x$  des 3<sup>ème</sup> et 4<sup>ème</sup> sommes (on a négligé la 5<sup>ème</sup> somme : elle intervient au plus tôt (c'est-à-dire pour  $p=1$ ) quand

$l_1 + l_2 + l_3 + 2m = 3p + 2m = 9$ , c'est-à-dire pour générer un terme en  $\Delta x^7$ , c'est-à-dire lors du calcul du 7<sup>ème</sup> terme de l'erreur de troncature !). On obtient donc :

p pair :

$$\begin{aligned}
 u_t - \alpha u_{xx} = & \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+3)!} \sum_j b_j j^{p+3} \right] \Delta x^{p+1} u_{(p+3)x} + \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+5)!} \sum_j b_j j^{p+5} \right] \Delta x^{p+3} u_{(p+5)x} + \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+7)!} \sum_j b_j j^{p+7} \right] \Delta x^{p+5} u_{(p+7)x} + \dots \\
 & + \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+2)!} \sum_j b_j j^{p+2} - \frac{r^2}{(\frac{p+2}{2})!} \right] \Delta x^p u_{(p+2)x} + \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+4)!} \sum_j b_j j^{p+4} - \frac{r^2}{(\frac{p+4}{2})!} \right] \Delta x^{p+2} u_{(p+4)x} + \dots \\
 & - \left\{ \left[ \frac{r^1}{1!} a_{p+2} \right] \Delta x^{p+2} u_{(p+4)x} + \left[ \frac{r^1}{1!} a_{p+3} \right] \Delta x^{p+3} u_{(p+5)x} + \left[ \frac{r^1}{1!} a_{p+4} + \frac{r^2}{2!} a_{p+2} \right] \Delta x^{p+4} u_{(p+6)x} + \left[ \frac{r^1}{1!} a_{p+5} + \frac{r^2}{2!} a_{p+3} \right] \Delta x^{p+5} u_{(p+7)x} \right\} \\
 & - \left\{ + \left[ \frac{r^1}{1!} a_{p+6} + \frac{r^2}{2!} a_{p+4} + \frac{r^3}{3!} a_{p+2} \right] \Delta x^{p+6} u_{(p+7)x} + \dots \right. \\
 & - \left. \left[ \frac{r^0}{2!0!} a_{p+2}^2 \right] \Delta x^{2p+2} u_{(2p+4)x} + \left[ \frac{r^0}{2!0!} 2a_{p+2} a_{p+3} \right] \Delta x^{2p+3} u_{(2p+5)x} + \left[ 2! \frac{r^1}{1!} a_{p+2}^2 + \frac{r^0}{2!0!} (a_{p+3}^2 + 2a_{p+2} a_{p+4}) \right] \Delta x^{2p+4} u_{(2p+6)x} \right\} \\
 & - \frac{\Delta t}{\Delta x^2} \left\{ + \left[ \frac{r^1}{2!1!} 2a_{p+2} a_{p+3} + \frac{r^0}{2!0!} (2a_{p+2} a_{p+5} + 2a_{p+3} a_{p+4}) \right] \Delta x^{2p+5} u_{(2p+7)x} \right. \\
 & \left. + \left[ \frac{r^2}{2!2!} a_{p+2}^2 + \frac{r^1}{2!1!} (a_{p+3}^2 + 2a_{p+2} a_{p+4}) + \frac{r^0}{2!0!} (a_{p+4}^2 + 2a_{p+3} a_{p+5} + 2a_{p+2} a_{p+6}) \right] \Delta x^{2p+6} u_{(2p+7)x} + \dots \right\}
 \end{aligned}$$

-....

IX-68

p impair :

$$\begin{aligned}
 u_t - \alpha u_{xx} = & \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+2)!} \sum_j b_j j^{p+2} \right] \Delta x^p u_{(p+2)x} + \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+4)!} \sum_j b_j j^{p+4} \right] \Delta x^{p+2} u_{(p+4)x} + \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+6)!} \sum_j b_j j^{p+6} \right] \Delta x^{p+4} u_{(p+6)x} + \dots \\
 & + \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+3)!} \sum_j b_j j^{p+3} - \frac{r^2}{(\frac{p+3}{2})!} \right] \Delta x^{p+1} u_{(p+3)x} + \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+5)!} \sum_j b_j j^{p+5} - \frac{r^2}{(\frac{p+5}{2})!} \right] \Delta x^{p+3} u_{(p+5)x} + \dots \\
 & - \left\{ \left[ \frac{r^1}{1!} a_{p+2} \right] \Delta x^{p+2} u_{(p+4)x} + \left[ \frac{r^1}{1!} a_{p+3} \right] \Delta x^{p+3} u_{(p+5)x} + \left[ \frac{r^1}{1!} a_{p+4} + \frac{r^2}{2!} a_{p+2} \right] \Delta x^{p+4} u_{(p+6)x} + \left[ \frac{r^1}{1!} a_{p+5} + \frac{r^2}{2!} a_{p+3} \right] \Delta x^{p+5} u_{(p+7)x} \right\} \\
 & - \left\{ + \left[ \frac{r^1}{1!} a_{p+6} + \frac{r^2}{2!} a_{p+4} + \frac{r^3}{3!} a_{p+2} \right] \Delta x^{p+6} u_{(p+7)x} + \dots \right. \\
 & - \left. \left[ \frac{r^0}{2!0!} a_{p+2}^2 \right] \Delta x^{2p+2} u_{(2p+4)x} + \left[ \frac{r^0}{2!0!} 2a_{p+2} a_{p+3} \right] \Delta x^{2p+3} u_{(2p+5)x} + \left[ 2! \frac{r^1}{1!} a_{p+2}^2 + \frac{r^0}{2!0!} (a_{p+3}^2 + 2a_{p+2} a_{p+4}) \right] \Delta x^{2p+4} u_{(2p+6)x} \right\} \\
 & - \frac{\Delta t}{\Delta x^2} \left\{ + \left[ \frac{r^1}{2!1!} 2a_{p+2} a_{p+3} + \frac{r^0}{2!0!} (2a_{p+2} a_{p+5} + 2a_{p+3} a_{p+4}) \right] \Delta x^{2p+5} u_{(2p+7)x} \right. \\
 & \left. + \left[ \frac{r^2}{2!2!} a_{p+2}^2 + \frac{r^1}{2!1!} (a_{p+3}^2 + 2a_{p+2} a_{p+4}) + \frac{r^0}{2!0!} (a_{p+4}^2 + 2a_{p+3} a_{p+5} + 2a_{p+2} a_{p+6}) \right] \Delta x^{2p+6} u_{(2p+7)x} + \dots \right\}
 \end{aligned}$$

-....

IX-69

Il est alors ais  de d uire les coefficients de l'erreur de troncature :

premier terme :

$$p \text{ pair : } a_{p+2} = \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+2)!} \sum_j b_j j^{p+2} - \frac{r^{\frac{p+2}{2}}}{(\frac{p+2}{2})!} \right] \quad \text{IX-70}$$

$$p \text{ impair : } a_{p+2} = \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+2)!} \sum_j b_j j^{p+2} \right] \quad \text{IX-71}$$

deuxième terme :

$$p \text{ pair : } a_{p+3} = \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+3)!} \sum_j b_j j^{p+3} \right] \quad \text{IX-72}$$

$$p \text{ impair : } a_{p+3} = \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+3)!} \sum_j b_j j^{p+3} - \frac{r^{\frac{p+3}{2}}}{(\frac{p+3}{2})!} \right] \quad \text{IX-73}$$

troisième terme :

$$p \text{ pair : } a_{p+4} = \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+4)!} \sum_j b_j j^{p+4} - \frac{r^{\frac{p+4}{2}}}{(\frac{p+4}{2})!} \right] - \frac{r^1}{1!} a_{p+2} \quad \text{IX-74}$$

$$p \text{ impair : } a_{p+4} = \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+4)!} \sum_j b_j j^{p+4} \right] - \frac{r^1}{1!} a_{p+2} \quad \text{IX-75}$$

quatrième terme :

$$p=1 : a_{p+5} = \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+5)!} \sum_j b_j j^{p+5} - \frac{r^{\frac{p+5}{2}}}{(\frac{p+5}{2})!} \right] - \frac{r^1}{1!} a_{p+3} - \frac{\Delta t}{\Delta x^2} \frac{r^0}{2!0!} a_{p+2}^2 \quad \text{IX-76}$$

$$p>1 : p \text{ pair : } a_{p+5} = \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+5)!} \sum_j b_j j^{p+5} \right] - \frac{r^1}{1!} a_{p+3} \quad \text{IX-77}$$

$$p \text{ impair : } a_{p+5} = \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+5)!} \sum_j b_j j^{p+5} - \frac{r^{\frac{p+5}{2}}}{(\frac{p+5}{2})!} \right] - \frac{r^1}{1!} a_{p+3} \quad \text{IX-78}$$

cinquième terme :

$$p=1 : a_{p+6} = \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+6)!} \sum_j b_j j^{p+6} \right] - \left[ \frac{r^1}{1!} a_{p+4} + \frac{r^2}{2!} a_{p+2} \right] - \frac{\Delta t}{\Delta x^2} \left[ \frac{r^0}{2!0!} 2a_{p+2} a_{p+3} \right] \quad \text{IX-79}$$

$$p=2 : \quad a_{p+6} = \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+6)!} \sum_j b_j j^{p+6} - \frac{r^{\frac{p+6}{2}}}{(\frac{p+6}{2})!} \right] - \left[ \frac{r^1}{1!} a_{p+4} + \frac{r^2}{2!} a_{p+2} \right] - \frac{\Delta t}{\Delta x^2} \left[ \frac{r^0}{2!0!} a_{p+2}^2 \right]$$

IX-80

$$p>2 : \quad p \text{ pair :} \quad a_{p+6} = \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+6)!} \sum_j b_j j^{p+6} - \frac{r^{\frac{p+6}{2}}}{(\frac{p+6}{2})!} \right] - \left[ \frac{r^1}{1!} a_{p+4} + \frac{r^2}{2!} a_{p+2} \right]$$

IX-81

$$p \text{ impair :} \quad a_{p+6} = \frac{\Delta x^2}{\Delta t} \left[ \frac{1}{(p+6)!} \sum_j b_j j^{p+6} \right] - \left[ \frac{r^1}{1!} a_{p+4} + \frac{r^2}{2!} a_{p+2} \right]$$

IX-82

*Exemple :* reprenons IX-1 à IX-3 :

$$\frac{u_i^{k+1} - u_i^j}{\Delta t} = \alpha \frac{u_{i+1}^k - 2u_i^k + u_{i-1}^k}{\Delta x^2}$$

IX-83

qui s'écrit ici

$$u_i^{k+1} = ru_{i-1}^k + (1-2r)u_i^k + ru_{i+1}^k$$

IX-84

c'est-à-dire

$$b_{-1} = b_1 = r \quad b_0 = 1 - 2r$$

IX-85

la méthode étant d'ordre 2, les conditions de consistante sont IX-64 et IX-65. Celles-ci donnent rapidement

$$\sum_j b_j = 0$$

$$\sum_j b_j j^3 = 0$$

$$\sum_j b_j = 1$$

qui sont vérifiées par IX-85. L'application des formules utiles parmi IX-70 à IX-82 donnent directement l'erreur de troncature de l'équation différentielle équivalente modifiée :

$$u_t - \alpha u_{xx} = -\frac{\alpha}{2}(r - \frac{1}{6})\Delta x^2 u_{4x} + \frac{\alpha}{3}(r^2 - \frac{r}{4} + \frac{1}{120})\Delta x^4 u_{6x} - \frac{3\alpha}{8}(r^3 - \frac{r^2}{3} + \frac{7r}{270} - \frac{1}{756})\Delta x^6 u_{8x}$$

IX-86

identique à IX-14.

### IX.3 Génération de nouveaux algorithmes avec un ordre de précision imposé

Les problèmes de stabilité étant généralement plus importants pour les équations hyperboliques que pour les paraboliques, on limitera le champ d'application de l'équation différentielle équivalente modifiée au seul cas des équations hyperboliques. Dans cet ordre d'idée, une des conséquences

importantes des développements précédents est la possibilité de générer des familles d'algorithmes ayant un support spatial et un ordre de précision donnés. Les conditions de consistance IX-39 définissent  $p+1$  relations pour les coefficients  $b_j$ . Si le support spatial couvre  $M$  points et si  $M$  est égal à  $p+1$ , alors il existe un seul schéma à  $M$  points d'ordre  $p$ ; mais si  $M$  est supérieur à  $p+1$ , des familles de schémas peuvent être générées. Celles-ci sont obtenues en ajoutant à une solution particulière du système IX-39 un multiple arbitraire de la solution du même système homogène. Par exemple, pour le support à trois points ( $M = 3 : j = -1, 0, 1$ ), le schéma du premier ordre doit satisfaire IX-39 avec  $m = 0$  et 1, c'est-à-dire

$$\begin{aligned} b_{-1} + b_0 + b_1 &= 1 \\ -b_{-1} + b_1 &= -v \end{aligned}$$

Ce système sous forme homogène s'écrit

$$\begin{aligned} b_{-1} + b_0 + b_1 &= 0 \\ -b_{-1} + b_1 &= 0 \end{aligned}$$

Sa solution générale est

$$\gamma \begin{bmatrix} +1 \\ -2 \\ +1 \end{bmatrix}$$

Si on prend comme solution particulière du système non homogène IX-52, la famille ( $M = 3 : j = -1, 0, 1 ; p = 1$ ) s'écrit

$$u_i^{k+1} = (v + \gamma)u_{i-1}^k + (1 - v - 2\gamma)u_i^k + \gamma u_{i+1}^k \quad \text{IX-87}$$

qui fournit tout schéma à trois points d'ordre 1. Il est cependant à remarquer que n'importe quelle valeur de  $\gamma$  ne conduit pas à un schéma véritablement utilisable : les seules valeurs utiles de  $\gamma$  sont celles qui génèrent des schémas stables.

Ceux-ci sont trouvés en repartant de IX-23 :

$$u_t + au_x = \sum_{l=p}^{\infty} a_{l+1} \Delta x^l u_{(l+1)x} \quad \text{IX-88}$$

La stabilité de ce schéma répond au même raisonnement que celui mené au paragraphe VI.3d : en « alimentant » cette équation par un signal

$$E_m(t)e^{j\omega_m x}$$

on peut déduire le facteur d'amplification de IX-88 exprimé en fonction des coefficients de l'erreur de troncature  $a_{l+1}$  : posant

$$u(x, t) = E_m(t)e^{j\omega_m x}$$

il vient aisément

$$u_t = \frac{dE_m}{dt} e^{j\omega_m t} \quad \text{et} \quad u_{kx} = E_m (j\omega_m)^k e^{j\omega_m t} \quad \text{IX-89}$$

Introduites dans IX-88, ces expressions génèrent l'équation différentielle ordinaire

$$\frac{dE_m}{dt} = E_m \left[ -aj\omega_m + \sum_{l=p}^{\infty} a_{l+1} \Delta x^l (j\omega_m)^{l+1} \right] \quad \text{IX-90}$$

c'est-à-dire

$$E_m(t) = E_m(0) e^{\left[ -aj\omega_m + \sum_{l=p}^{\infty} a_{l+1} \Delta x^l (j\omega_m)^{l+1} \right] t} \quad \text{IX-91}$$

dont on déduit

$$G = e^{\left[ -aj\omega_m + \sum_{l=p}^{\infty} a_{l+1} \Delta x^l (j\omega_m)^{l+1} \right] \Delta t} \quad \text{IX-92}$$

Cette relation nous montre que selon leur parité, les coefficients  $a_{l+1}$  de l'erreur de troncature contribuent à introduire une variation d'amplitude ou un déphasage ou, ce qui revient au même, contribuent à l'erreur de diffusion ou de dispersion. En se rappelant que pour les équations hyperboliques, ces erreurs ont été définies par

$$\varepsilon_D = |G| \quad \text{et} \quad \varepsilon_\phi = \frac{\arg(G)}{-v\phi}$$

et en utilisant la relation  $\phi = \omega \Delta x$ , on obtient aisément

méthode d'ordre p pair :

$$\varepsilon_D = \exp \left[ \left( a_{p+2} (j\phi)^{p+2} + a_{p+4} (j\phi)^{p+4} + \dots \right) \frac{\Delta t}{\Delta x} \right] \quad \text{IX-93}$$

$$\varepsilon_\phi = 1 - \frac{1}{a} \left[ a_{p+1} (j\phi)^p + a_{p+3} (j\phi)^{p+2} + \dots \right] \quad \text{IX-94}$$

méthode d'ordre p impair :

$$\varepsilon_D = \exp \left[ \left( a_{p+1} (j\phi)^{p+1} + a_{p+3} (j\phi)^{p+3} + \dots \right) \frac{\Delta t}{\Delta x} \right] \quad \text{IX-95}$$

$$\varepsilon_\phi = 1 - \frac{1}{a} \left[ a_{p+2} (j\phi)^{p+1} + a_{p+4} (j\phi)^{p+3} + \dots \right] \quad \text{IX-96}$$

Dans la pratique, et pour peu qu'on se limite aux petites valeurs de  $\phi$ , on ne retient que les premiers termes des développements en série de ces expressions et on utilise

$$p \text{ pair : } \varepsilon_D = 1 + a_{p+2} (j\phi)^{p+2} \frac{\Delta t}{\Delta x} + O(\phi^{p+4}) \quad IX-97$$

$$\varepsilon_\Phi = 1 - \frac{a_{p+1}}{a} (j\phi)^p + O(\phi^{p+2}) \quad IX-98$$

$$p \text{ impair : } \varepsilon_D = 1 + a_{p+1} (j\phi)^{p+1} \frac{\Delta t}{\Delta x} + O(\phi^{p+3}) \quad IX-99$$

$$\varepsilon_\Phi = 1 - \frac{a_{p+2}}{a} (j\phi)^{p+1} + O(\phi^{p+3}) \quad IX-100$$

La condition générale de stabilité étant  $|G| < 1$ , on en déduit les conditions nécessaires de stabilité suivantes (en utilisant  $j = \sqrt{-1}$ ) :

$$p \text{ pair : } a_{p+2} (-1)^{\frac{p}{2}} \frac{\Delta t}{\Delta x} > 0 \quad IX-101$$

$$p \text{ impair : } a_{p+1} (-1)^{\frac{p+1}{2}} \frac{\Delta t}{\Delta x} < 0 \quad IX-102$$

exemple : pour les algorithmes d'ordre 1, IX-102 donne

$$a_2 \frac{\Delta t}{\Delta x} > 0 \quad IX-103$$

c'est-à-dire (cf. IX-40)

$$\left( \sum_j b_j j^2 - (-v)^2 \right) \frac{1}{2!} > 0 \quad IX-104$$

ou encore pour un schéma à trois points (-1, 0, 1)

$$(b_{-1} + b_1 - v^2) \frac{1}{2!} > 0 \quad IX-105$$

c'est-à-dire, avec IX-87,

$$v + \gamma + \gamma - v^2 > 0$$

ou encore

$$\gamma > \frac{1}{2} (v^2 - v) \quad IX-106$$

Application à la méthode de Lax : calculons d'abord la valeur de  $\gamma$  à introduire dans IX-87 pour retrouver cette méthode : VII-14 se met facilement sous la forme

$$u_i^{k+1} = \frac{1+v}{2} u_{i-1}^k + \frac{1-v}{2} u_{i+1}^k \quad IX-107$$

Comparé à IX-87, on a immédiatement

$$\gamma = \frac{1-v}{2}$$

IX-108

la méthode de Lax est donc stable si IX-108 vérifie IX-106 :

$$\frac{1-v}{2} > \frac{v^2 - v}{2} ?$$

ceci est vérifié si

$$-1 < v < 1$$

IX-109

qui reproduit rigoureusement les résultats du paragraphe VIII.3.

## Chapitre X. Extensions de la méthode de Von Neumann pour l'analyse de la stabilité

Le but de ce chapitre est de proposer les pistes à emprunter lorsqu'il s'agit d'appliquer la méthode de Von Neumann à diverses généralisations :

- la résolution d'une équation linéaire multidimensionnelle
- les systèmes monodimensionnels linéaires
- les systèmes multidimensionnels linéaires
- les problèmes non linéaires
- globalisation des conditions de Von Neumann par famille de schémas

### X.1 Résolution d'une équation multidimensionnelle linéaire

Dans une large mesure, les notions développées au chapitre VI sont généralisables. Voyons-en un exemple classique :

*Deuxième loi de Fourier dans le cas bidimensionnel*

Il s'agit de résoudre

$$\frac{\partial u}{\partial t} = \alpha \left[ \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right] \quad X-1$$

Toutes les méthodes applicables au cas monodimensionnel sont généralisables. En particulier, la méthode explicite simple s'écrit

$$\frac{u_{i,j}^{k+1} - u_{i,j}^k}{\Delta t} = \alpha \left[ \frac{u_{i+1,j}^k - 2u_{i,j}^k + u_{i-1,j}^k}{(\Delta x)^2} + \frac{u_{i,j+1}^k - 2u_{i,j}^k + u_{i,j-1}^k}{(\Delta y)^2} \right] \quad X-2$$

Facteur d'amplification : en posant  $r_x = \alpha \frac{\Delta t}{(\Delta x)^2}$  et  $r_y = \alpha \frac{\Delta t}{(\Delta y)^2}$  et en utilisant les arguments  $\phi_x$  et  $\phi_y$  on a par un raisonnement analogue à celui du paragraphe VI.3

$$\frac{A^{k+1} - A^k}{\Delta t} = \alpha \left[ \frac{A^k (e^{i\phi_x} - 2 + e^{-i\phi_x})}{(\Delta x)^2} + \frac{A^k (e^{i\phi_y} - 2 + e^{-i\phi_y})}{(\Delta y)^2} \right] \quad X-3$$

c'est-à-dire

$$G = 1 - 4r_x \sin^2 \frac{\phi_x}{2} - 4r_y \sin^2 \frac{\phi_y}{2} \quad X-4$$

$\phi_x$  et  $\phi_y$  pouvant être compris entre 0 et  $\pi$ , la condition de stabilité devient

$$0 \leq r_x + r_y \leq \frac{1}{2} \quad X-5$$

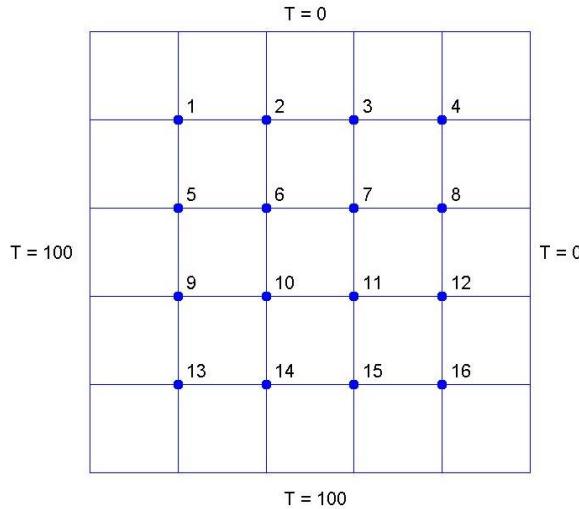
remarquons que si  $\Delta x = \Delta y$ , la condition de stabilité devient

$$0 \leq r \leq \frac{1}{4} \quad X-6$$

ce qui est plus sévère que dans le cas monodimensionnel et rend cette méthode encore moins praticable.

La généralisation de la méthode de Cranck-Nicholson ne pose pas de difficultés particulières ; elle est inconditionnellement stable, comme dans le cas monodimensionnel mais elle conduit à un système linéaire qui n'est plus tridiagonal. Voyons le détail des calculs sur un exemple :

On considère une plaque d'acier carrée de 15 cm de côté. Initialement, la plaque est à zéro °C. On demande de déterminer l'évolution temporelle de la température dans la plaque si à l'instant  $t = 0$  on maintient deux côtés adjacents à 0° C et si on porte les deux autres instantanément à 100° C (on supposera qu'il n'y a pas de pertes par les faces planes de la plaque). On choisit  $\Delta x = \Delta y = 3$  cm :



Posons  $r = \alpha \frac{\Delta t}{(\Delta x)^2} = \alpha \frac{\Delta t}{(\Delta y)^2}$  ; la méthode de Cranck-Nicholson donne

$$u_{i,j}^{k+1} - u_{i,j}^k = \frac{r}{2} [u_{i+1,j}^{k+1} - 2u_{i,j}^{k+1} + u_{i-1,j}^{k+1} + u_{i+1,j}^k - 2u_{i,j}^k + u_{i-1,j}^k + u_{i,j+1}^{k+1} - 2u_{i,j}^{k+1} + u_{i,j-1}^{k+1} + u_{i,j+1}^k - 2u_{i,j}^k + u_{i,j-1}^k] \quad X-7$$

c'est-à-dire, en divisant tout par  $\frac{r}{2}$ ,

$$-u_{i,j-1}^{k+1} - u_{i,j+1}^{k+1} + \left(\frac{2}{r} + 4\right)u_{i,j}^{k+1} - u_{i-1,j}^{k+1} - u_{i+1,j}^{k+1} = u_{i,j-1}^k + u_{i,j+1}^k + \left(\frac{2}{r} - 4\right)u_{i,j}^k + u_{i-1,j}^k + u_{i+1,j}^k \quad X-8$$

En y ajoutant les conditions aux limites et en tenant compte du maillage choisi, le calcul de la solution requiert de résoudre à chaque pas de temps un système linéaire  $A\bar{u}^{k+1} = B\bar{u}^k + \bar{c}^k$  de taille égale à 16 et dont la matrice est tridiagonale par bloc :

Un tel système est généralement résolu par une méthode itérative (voir chapitre III), ce qui peut générer un nombre très important de calculs, si le nombre de pas de temps est élevé.

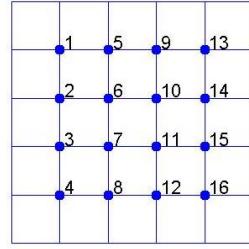
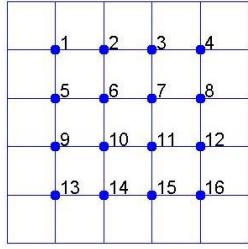
C'est pour répondre à cette situation que des méthodes directes restaurant le caractère tridiagonal pur ont été imaginées : les méthodes aux directions alternées (ADI) et leurs dérivées. Ces méthodes utilisent le procédé suivant : on estime tout d'abord dans  $X-1$   $\frac{\partial^2 u}{\partial x^2}$  par une approximation prise au début de l'intervalle  $[k, k+1]$  et  $\frac{\partial^2 u}{\partial y^2}$  par une approximation prise à la fin de cet intervalle : il vient, pour autant que  $\Delta x = \Delta y$

$$u_{i,j}^{k+1} - u_{i,j}^k = r[u_{i+1,j}^k - 2u_{i,j}^k + u_{i-1,j}^k + u_{i,j+1}^{k+1} - 2u_{i,j}^{k+1} + u_{i,j-1}^{k+1}] \quad X-9$$

Le biais introduit est alors corrigé lors du calcul de l'itéré suivant, en inversant les instants d'évaluation de  $\frac{\partial^2 u}{\partial x^2}$  et de  $\frac{\partial^2 u}{\partial y^2}$  :

$$u_{i,i}^{k+2} - u_{i,i}^{k+1} = r[u_{i+1,i}^{k+2} - 2u_{i,i}^{k+2} + u_{i-1,i}^{k+2} + u_{i,i+1}^{k+1} - 2u_{i,i}^{k+1} + u_{i,i-1}^{k+1}] \quad X-10$$

En prenant la peine de numérotter les nœuds du maillage dans le sens « des x » pour X-10 et dans celui « des y » pour X-9 ,



on obtient deux systèmes tridiagonaux à résoudre (en divisant tout par  $r$ ) :

$$X-9 : -u_{i,j-1}^{k+1} + \left(2 + \frac{1}{r}\right)u_{i,j}^{k+1} - u_{i,j+1}^{k+1} = u_{i-1,j}^k - \left(2 - \frac{1}{r}\right)u_{i,j}^k + u_{i+1,j}^k \quad X-11$$

$$X-10 : -u_{i-1,j}^{k+2} + \left(2 + \frac{1}{r}\right)u_{i,j}^{k+2} - u_{i+1,j}^{k+2} = u_{i,j-1}^{k+1} - \left(2 - \frac{1}{r}\right)u_{i,j}^{k+1} + u_{i,j+1}^{k+1} \quad X-12$$

Si on convient de désigner par la lettre  $u$  la valeur de la fonction inconnue quand la numérotation est dans le sens des  $x$  et par la lettre  $v$  quand la numérotation est dans le sens des  $y$ , les deux systèmes précédents s'écrivent (en tenant compte des CL) en posant  $a = 2 + \frac{1}{r}$  et  $b = 2 - \frac{1}{r}$  :

$$X-11 : Av^{k+1} = Bu^k + C_1$$

$$X-12 : Au^{k+2} = Bv^{k+1} + C_2 \text{ avec}$$

$$A = \left( \begin{array}{cccc|cccc|cccc} a & -1 & & & & & & & & & & \\ -1 & a & -1 & & & & & & & & & \\ & -1 & a & -1 & & & & & & & & \\ & & -1 & a & & & & & & & & \\ \hline & & & a & -1 & & & & & & & \\ & & & -1 & a & -1 & & & & & & \\ & & & & -1 & a & -1 & & & & & \\ & & & & & -1 & a & & & & & \\ \hline & & & & & & a & -1 & & & & \\ & & & & & & -1 & a & -1 & & & \\ & & & & & & & -1 & a & -1 & & \\ & & & & & & & & -1 & a & & \\ \hline & & & & & & & & & a & -1 & \\ & & & & & & & & & -1 & a & -1 \\ & & & & & & & & & & -1 & a & -1 \\ & & & & & & & & & & & -1 & a \end{array} \right)$$

$$B = \begin{pmatrix} -b & 1 & & & & \\ & -b & 1 & & & \\ & & -b & 1 & & \\ & & & -b & 1 & \\ \hline 1 & -b & 1 & & & \\ & 1 & -b & 1 & & \\ & & 1 & -b & 1 & \\ & & & 1 & -b & 1 \\ \hline & 1 & -b & 1 & & \\ & & 1 & -b & 1 & \\ & & & 1 & -b & \\ & & & & 1 & -b \end{pmatrix}$$

$$C_1 = (100 \quad 100 \quad 100 \quad 200 \quad 0 \quad 0 \quad 0 \quad 100 \quad 0 \quad 0 \quad 0 \quad 100 \quad 0 \quad 0 \quad 0 \quad 100)^T$$

$$C_2 = (100 \quad 0 \quad 0 \quad 0 \quad 100 \quad 0 \quad 0 \quad 0 \quad 100 \quad 0 \quad 0 \quad 0 \quad 200 \quad 100 \quad 100 \quad 100 \quad 100)^T$$

Solution obtenue : la thermique nous enseigne que pour ce problème  $\alpha = \frac{k}{\rho c}$  où k est la conductibilité

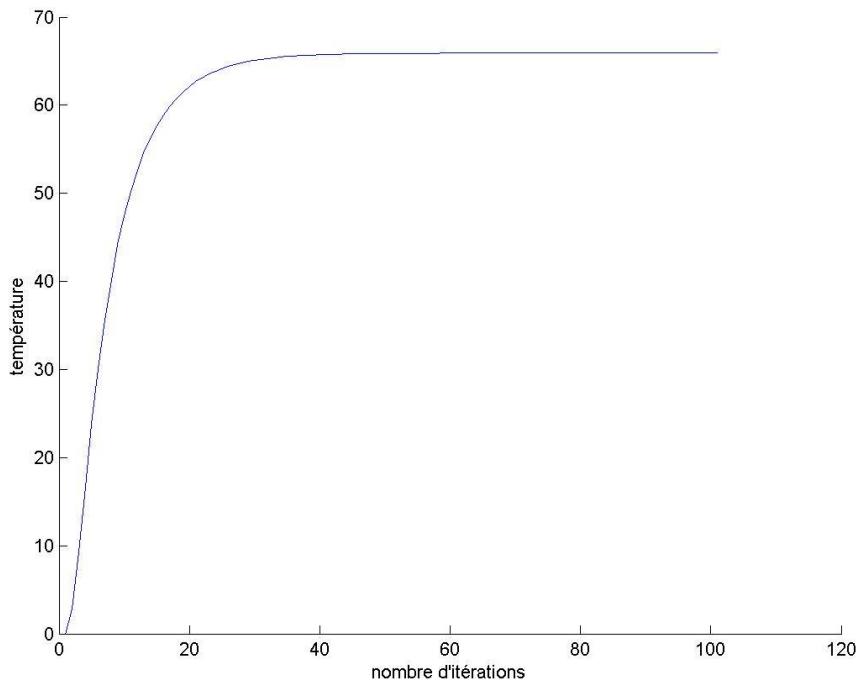
thermique du matériau, c sa capacité calorifique et  $\rho$  sa masse volumique . Pour l'exemple considéré, on prend

$$k = 0.13$$

$$c = 0.11$$

$$\rho = 7.8$$

La figure suivante représente l'évolution temporelle de u en le point 10 de la numérotation horizontale. Le paramètre r a été arbitrairement fixé à 0.1, et comme  $\Delta x = \Delta y = 3$ , cela revient à imposer  $\Delta t = 5.94$



La valeur de régime est donc atteinte en plus ou moins 50 (doubles) itérations, c'est-à-dire après 590 secondes environ.

## X.2 Résolution d'un système monodimensionnel linéaire

La différence essentielle avec le cas d'une équation monodimensionnelle linéaire est le remplacement de la notion de facteur d'amplification par celle de matrice d'amplification. Voyons un exemple.

Le système d'équations suivant modélise - en le linéarisant – un écoulement monodimensionnel de faible profondeur et à surface libre :  $h$  représente la profondeur de fluide et  $v$  la vitesse de l'écoulement (supposé monodimensionnel selon l'axe  $x$ ) :

$$\begin{cases} \frac{\partial h}{\partial t} + v_0 \frac{\partial h}{\partial x} + h_0 \frac{\partial v}{\partial x} = 0 \\ \frac{\partial v}{\partial t} + v_0 \frac{\partial v}{\partial x} + g \frac{\partial h}{\partial x} = 0 \end{cases} \quad X-13$$

Définissons le vecteur  $\bar{u}$  et la matrice  $M$  :

$$\bar{u} = \begin{pmatrix} h \\ v \end{pmatrix} \quad M = \begin{pmatrix} v_0 & h_0 \\ g & v_0 \end{pmatrix} \quad X-14$$

Il vient avec X-13 :

$$\frac{\partial \bar{u}}{\partial t} + M \frac{\partial \bar{u}}{\partial x} = 0 \quad X-15$$

Discrétisons X-15 à l'aide du schéma centré II-8 pour la dérivée spatiale de chaque composante de  $\bar{u}$  et du schéma décentré II-9 pour la dérivée temporelle : remarquons d'abord pour cela que la discrétisation spatiale va générer le vecteur

$$\bar{u} = \begin{pmatrix} \vdots \\ u_{i-1} \\ u_i \\ u_{i+1} \\ \vdots \end{pmatrix} \quad \text{avec} \quad u_i = \begin{pmatrix} h_i \\ v_i \end{pmatrix} \quad X-16$$

Les schémas de différences finies choisis induisent la relation matricielle

$$\frac{u_i^{k+1} - u_i^k}{\Delta t} = -M \frac{u_{i+1}^k - u_{i-1}^k}{2\Delta x} \quad X-17$$

c'est-à-dire

$$u_i^{k+1} = u_i^k - \frac{M\Delta t}{2\Delta x} (u_{i+1}^k - u_{i-1}^k) \quad X-18$$

Il est nécessaire à ce stade de tenir compte de la nature vectorielle de  $u_i$  pour poursuivre : X-18 s'écrit aussi

$$\begin{pmatrix} h_i^{k+1} \\ v_i^{k+1} \end{pmatrix} = \begin{pmatrix} h_i^k \\ v_i^k \end{pmatrix} - \begin{pmatrix} v_0 & h_0 \\ g & v_0 \end{pmatrix} \frac{\Delta t}{2\Delta x} \begin{pmatrix} h_{i+1}^k - h_{i-1}^k \\ v_{i+1}^k - v_{i-1}^k \end{pmatrix} \quad X-19$$

En vue de faire apparaître la matrice d'amplification, introduisons l'opérateur de décalage spatial  $e^{j\phi}$  :

$$\begin{pmatrix} h_i^{k+1} \\ v_i^{k+1} \end{pmatrix} = \begin{pmatrix} h_i^k \\ v_i^k \end{pmatrix} - \begin{pmatrix} v_0 & h_0 \\ g & v_0 \end{pmatrix} \frac{\Delta t}{2\Delta x} \begin{pmatrix} (e^{j\phi} - e^{-j\phi})h_i^k \\ (e^{j\phi} - e^{-j\phi})v_i^k \end{pmatrix} \quad X-20$$

qui s'écrit encore

$$\begin{pmatrix} h_i^{k+1} \\ v_i^{k+1} \end{pmatrix} = \begin{pmatrix} h_i^k \\ v_i^k \end{pmatrix} - \begin{pmatrix} v_0 & h_0 \\ g & v_0 \end{pmatrix} \frac{\Delta t}{\Delta x} \begin{pmatrix} j \sin \phi h_i^k \\ j \sin \phi v_i^k \end{pmatrix}$$

ou

$$\begin{pmatrix} h_i^{k+1} \\ v_i^{k+1} \end{pmatrix} = \begin{pmatrix} 1 - j v_0 \frac{\Delta t}{\Delta x} \sin \phi & - j h_0 \frac{\Delta t}{\Delta x} \sin \phi \\ - j g \frac{\Delta t}{\Delta x} \sin \phi & 1 - j v_0 \frac{\Delta t}{\Delta x} \sin \phi \end{pmatrix} \begin{pmatrix} h_i^k \\ v_i^k \end{pmatrix} \quad X-21$$

La condition de stabilité de Von Neumann dans le cas d'une équation consistait à exprimer que tout harmonique de l'erreur d'arrondi ne doit pas croître avec  $k$  : il apparaît que la stabilité sera ici acquise si la norme de la matrice de X-21 est inférieure ou égale à un. Cette matrice est appelée matrice d'amplification :

$$G = \begin{pmatrix} 1 - j \frac{v_0 \Delta t}{\Delta x} \sin \phi & -j \frac{h_0 \Delta t}{\Delta x} \sin \phi \\ -j \frac{g \Delta t}{\Delta x} \sin \phi & 1 - j \frac{v_0 \Delta t}{\Delta x} \sin \phi \end{pmatrix} \quad X-22$$

La condition de stabilité est donc

$$\|G\| \leq 1 \quad X-23$$

En pratique, on préfère manipuler les valeurs propres de G : si  $\lambda_i$  sont ces valeurs propres, la condition de stabilité s'écrit

$$\rho(G) \equiv \max_i |\lambda_i| \leq 1 \quad X-24$$

Pour l'exemple traité X-22, en posant

$$v_0 = \frac{v_0 \Delta t}{\Delta x} \quad \text{et} \quad v = \frac{\sqrt{gh_0} \Delta t}{\Delta x} \quad X-25$$

les valeurs propres valent

$$\lambda_{1,2} = 1 - j(v_0 \pm v) \sin \phi \quad X-26$$

On a donc

$$\rho(G) = \sqrt{1 + \frac{\Delta t^2}{\Delta x^2} (v_0 + \sqrt{gh_0})^2 \sin^2 \phi} \quad X-27$$

qui est toujours supérieur ou égal à un. La méthode est donc instable, comme on pouvait s'y attendre en vertu des résultats du paragraphe VII.1.

D'un point de vue pratique, observons que la matrice d'amplification peut être rapidement déduite sans passer par le détail des calculs X-19 à X-21 : il est en effet commode, et rigoureusement correct, d'écrire à partir de X-18

$$G = I - \frac{M \Delta t}{2 \Delta x} (e^{j\phi} - e^{-j\phi}) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - j \frac{\Delta t}{\Delta x} \sin \phi \begin{pmatrix} v_0 & h_0 \\ g & v_0 \end{pmatrix} \quad X-28$$

identique à X-22.

### X.3 Résolution d'un système multidimensionnel linéaire

Généralisons le problème précédent :

$$\frac{\partial \bar{u}}{\partial t} + M \frac{\partial \bar{u}}{\partial x} + N \frac{\partial \bar{u}}{\partial y} = 0 \quad X-29$$

avec  $MN = NM$  et appliquons-lui la généralisation de la méthode de Lax (cf. VII-24) :

$$\left(\frac{\partial u}{\partial t}\right)_{ij}^k = \frac{u_{i+1j}^{k+1} - (u_{i+1j}^k + u_{i-1j}^k + u_{ij-1}^k + u_{ij+1}^k)/4}{\Delta t}$$

et

$$\left(\frac{\partial u}{\partial x}\right)_{ij}^k = \frac{u_{i+1j}^k - u_{i-1j}^k}{2\Delta x} \quad \left(\frac{\partial u}{\partial y}\right)_{ij}^k = \frac{u_{ij+1}^k - u_{ij-1}^k}{2\Delta y}$$

X-30

Il vient donc

$$u_i^{k+1} = \frac{(u_{i+1j}^k + u_{i-1j}^k + u_{ij-1}^k + u_{ij+1}^k)}{4} - \frac{\Delta t}{2\Delta x} M(u_{i+1j}^k - u_{i-1j}^k) - \frac{\Delta t}{2\Delta y} N(u_{ij+1}^k - u_{ij-1}^k) \quad X-31$$

La matrice d'amplification découle de

$$u_i^{k+1} = \frac{(e^{j\phi_x} + e^{-j\phi_x} + e^{j\phi_y} + e^{-j\phi_y})}{4} u_i^k - \frac{\Delta t}{2\Delta x} M(e^{j\phi_x} - e^{-j\phi_x}) - \frac{\Delta t}{2\Delta y} N(e^{j\phi_y} - e^{-j\phi_y}) u_i^k \quad X-32$$

$$\Rightarrow G = \frac{(\cos \phi_x + \cos \phi_y)}{2} I - j \frac{\Delta t}{\Delta x} M \sin \phi_x - j \frac{\Delta t}{\Delta y} N \sin \phi_y \quad X-33$$

G est complexe : imposer  $\|G\| \leq 1$  est atteint plus facilement en remplaçant cette condition par

$$\|G\|^2 = \|GG^*\| \leq 1 \quad X-34$$

On trouve aisément

$$GG^* = \frac{1}{4} (\cos \phi_x + \cos \phi_y)^2 I + \left( \frac{\Delta t}{\Delta x} \sin \phi_x M + \frac{\Delta t}{\Delta y} \sin \phi_y N \right)^2 \quad X-35$$

et successivement

$$\begin{aligned} \|GG^*\| &\leq \frac{1}{4} (\cos \phi_x + \cos \phi_y)^2 \|I\| + \left\| \left( \frac{\Delta t}{\Delta x} \sin \phi_x M + \frac{\Delta t}{\Delta y} \sin \phi_y N \right)^2 \right\| \\ \|GG^*\| &\leq \frac{1}{4} (\cos \phi_x + \cos \phi_y)^2 + \left\| \left( \frac{\Delta t}{\Delta x} \sin \phi_x M + \frac{\Delta t}{\Delta y} \sin \phi_y N \right) \right\|^2 \end{aligned}$$

Par ailleurs

$$\left\| \left( \frac{\Delta t}{\Delta x} \sin \phi_x M + \frac{\Delta t}{\Delta y} \sin \phi_y N \right) \right\| \leq \frac{\Delta t}{\Delta x} |\sin \phi_x| \|M\| + \frac{\Delta t}{\Delta y} |\sin \phi_y| \|N\|$$

engendre

$$\|GG^*\| \leq \frac{1}{4} (\cos \phi_x + \cos \phi_y)^2 + \left( \frac{\Delta t}{\Delta x} |\sin \phi_x| \|M\| + \frac{\Delta t}{\Delta y} |\sin \phi_y| \|N\| \right)^2 \quad X-36$$

En termes de rayon spectral, si on pose

$$v_x = \frac{\Delta t}{\Delta x} \rho(M) \quad \text{et} \quad v_y = \frac{\Delta t}{\Delta y} \rho(N)$$

X-37

on obtient finalement

$$\rho(GG^*) \leq \frac{1}{4} (\cos \phi_x + \cos \phi_y)^2 + (\sin \phi_x |v_x| + \sin \phi_y |v_y|)^2 \quad X-38$$

La stabilité est alors acquise si

$$\frac{1}{4} (\cos \phi_x + \cos \phi_y)^2 + (\sin \phi_x |v_x| + \sin \phi_y |v_y|)^2 \leq 1 \quad X-39$$

quels que soient  $\phi_x$  et  $\phi_y$ . Les seuls moyens d'action qu'on ait pour essayer d'atteindre X-39 est de trouver des valeurs maximales admissibles pour  $v_x$  et  $v_y$ . La situation la plus drastique est atteinte quand  $\phi_x$  et  $\phi_y$  sont proches de zéro : en effet, en ne conservant que le premier terme des développements de Taylor des fonctions sinus et cosinus :

$$\cos \phi = 1 - \frac{\phi^2}{2!} + \frac{\phi^4}{4!} - \dots \quad \text{et} \quad \sin \phi = \phi - \frac{\phi^3}{3!} + \dots$$

X-39 devient

$$(\phi_x v_x + \phi_y v_y)^2 \leq 0 \quad X-40$$

qui est impossible à atteindre. A l'inverse, quand  $\phi_x$  et  $\phi_y$  sont proches de  $\frac{\pi}{2}$ , X-39 s'écrit

$$(v_x + v_y)^2 \leq 1 \quad X-41$$

qui est beaucoup moins sévère que X-40. C'est donc le cas  $\phi_x$  et  $\phi_y$  proches de zéro qui sera retenu pour établir une condition nécessaire de stabilité. Celle-ci peut être trouvée de manière plus probante que X-40 en développant dans X-39 les fonctions sinus et cosinus jusqu'au deuxième terme : tous calculs faits, X-39 peut se mettre sous la forme

$$\frac{1}{4} (\cos \phi_x + \cos \phi_y)^2 + (\sin \phi_x |v_x| + \sin \phi_y |v_y|)^2 = 1 - \left[ \left( \frac{1}{2} - v_x^2 \right) \phi_x^2 + \left( \frac{1}{2} - v_y^2 \right) \phi_y^2 - 2 v_x v_y \phi_x \phi_y \right] + O(\phi_x^4, \phi_y^4)$$

La stabilité est donc assurée si la forme quadratique entre [ ] est positive : ceci se produit si le discriminant

$$(v_x v_y)^2 - \left( \frac{1}{2} - v_x^2 \right) \left( \frac{1}{2} - v_y^2 \right) \leq 0$$

c'est-à-dire si

$$v_x^2 + v_y^2 \leq \frac{1}{2} \quad X-41$$

C'est cette condition qui sera conservée pour assurer la stabilité du schéma.

## X.4 Problèmes non linéaires

Voyons d'abord le cas des équations linéaires à coefficients non constants : les équations canoniques des deux catégories habituelles (parabolique et hyperbolique) deviennent

$$\frac{\partial u}{\partial t} - \frac{\partial}{\partial x} \left[ \alpha(x) \frac{\partial u}{\partial x} \right] = 0 \quad X-39$$

$$\frac{\partial u}{\partial t} + a(x) \frac{\partial u}{\partial x} = 0 \quad X-40$$

*modèle parabolique :*

$u_t$  est estimé par une formule de différences finies classique : par exemple

$$(u_t)_i^n = \frac{u_i^{n+1} - u_i^n}{\Delta t} \quad X-41$$

En accord avec ce qui a été fait à de multiples reprises pour le cas linéaire,  $\left( \alpha(x) \frac{\partial u}{\partial x} \right)_i^n$  sera estimé

par une formule centrée : si, par exemple, le souhait est de ne faire apparaître que les abscisses  $x_{i-1}$ ,  $x_i$ ,  $x_{i+1}$  dans le résultat final, il est nécessaire d'écrire

$$1^o \left( \frac{\partial}{\partial x} \left[ \alpha(x) \frac{\partial u}{\partial x} \right] \right)_i^n = \frac{1}{\Delta x} \left[ \alpha(x_{i+1/2}) \left( \frac{\partial u}{\partial x} \right)_{x_{i+1/2}} - \alpha(x_{i-1/2}) \left( \frac{\partial u}{\partial x} \right)_{x_{i-1/2}} \right]$$

$$2^o \left( \frac{\partial u}{\partial x} \right)_{x_{i+1/2}} = \frac{1}{\Delta x} [u(x_{i+1}) - u(x_i)]$$

et donc

$$\left( \frac{\partial}{\partial x} \left[ \alpha(x) \frac{\partial u}{\partial x} \right] \right)_i^n = \frac{1}{\Delta x^2} [\alpha_{i+1/2} (u_{i+1}^n - u_i^n) - \alpha_{i-1/2} (u_i^n - u_{i-1}^n)] \quad X-42$$

X-41 et X-42 dans X-39 permettent de calculer le facteur d'amplification :

$$\frac{A^{n+1} - A^n}{\Delta t} = \frac{1}{\Delta x^2} [\alpha_{i+1/2} (e^{j\phi} - 1) - \alpha_{i-1/2} (1 - e^{-j\phi})] A^n$$

$$\Leftrightarrow \frac{A^{n+1}}{A^n} = 1 + \frac{\Delta t}{\Delta x^2} [\alpha_{i+1/2} (e^{j\phi} - 1) - \alpha_{i-1/2} (1 - e^{-j\phi})]$$

G dépend de  $\phi$  mais aussi de  $x_{i+1/2}$  et de  $x_{i-1/2}$ . Cette dernière dépendance a pour effet de limiter l'analyse de Von Neumann à la recherche d'une stabilité locale : le facteur d'amplification correspondant devient dépendant de l'endroit où on se trouve dans le domaine spatial, c'est-à-dire de  $x$  : il est obtenu en imaginant que  $x_{i+1/2}$  et  $x_{i-1/2}$  sont très proches de  $x$  :

$$G(x, \phi) = 1 + \frac{\Delta t}{\Delta x^2} [\alpha(x) (e^{j\phi} - 1) - \alpha(x) (1 - e^{-j\phi})]$$

$$\Leftrightarrow G(x, \phi) = 1 + \frac{\Delta t}{\Delta x^2} \alpha(x) [2 \cos \phi - 2] = 1 - 4 \frac{\Delta t}{\Delta x^2} \alpha(x) \sin^2(\phi / 2)$$

**modèle hyperbolique XI-2 :**

prenant à nouveau

$$(u_t)_i^n = \frac{u_i^{n+1} - u_i^n}{\Delta t}$$

et

$$\left( a(x) \frac{\partial u}{\partial x} \right)_i^n = a(x_{i-1/2}) \frac{u_i^n - u_{i-1}^n}{\Delta x},$$

il est aisément d'établir que le facteur d'amplification vaut

$$G(x, \phi) = 1 - \frac{\Delta t}{\Delta x} a(x) [1 - e^{-j\phi}]$$

**commentaires :**

on peut montrer que pour des équations linéaires à coefficients non constants telles que X-39 et X-40, l'étude locale, c'est-à-dire à  $x$  fixé, du facteur d'amplification fournit une condition nécessaire de stabilité. Des conditions suffisantes peuvent théoriquement être obtenues : elles résultent de la nécessité de maîtriser les harmoniques à hautes fréquences générées par le comportement non linéaire de l'équation à résoudre. Si ceci est particulièrement crucial pour les équations hyperboliques non linéaires - car elles décrivent essentiellement des phénomènes de propagation d'ondes sans atténuation physique -, c'est aussi vrai pour les équations paraboliques – où une telle atténuation physique existe – qui requièrent aussi des conditions de stabilité supplémentaires. Cette maîtrise peut être atteinte grâce à l'utilisation de schémas particuliers dits dissipatifs.

Pour ce qui concerne les équations non linéaires, peu d'informations sont disponibles quant à l'étude de la stabilité des schémas appliqués à de tels problèmes. La linéarisation des équations, avec des coefficients « gelés » se prête à une étude classique de la stabilité. Les critères qui en résultent sont nécessaires à la stabilité de la forme non linéaire mais sont pas forcément suffisants.

## X.5 Globalisation des conditions de Von Neumann par famille de schémas

Indépendant de l'équation traitée, ce type de généralisation permet de formuler les conditions nécessaires et suffisantes de stabilité pour une famille donnée de schémas : à titre d'exemple, on traitera le cas des schémas monodimensionnels à deux niveaux de temps et spatialement à trois points centrés. Un tel schéma s'écrit

$$b_3 u_{i+1}^{k+1} + b_2 u_i^{k+1} + b_1 u_{i-1}^{k+1} = a_3 u_{i+1}^k + a_2 u_i^k + a_1 u_{i-1}^k \quad X-43$$

Les coefficients  $b_i$  et  $a_i$  ne peuvent être quelconques : ils doivent vérifier une condition de consistance : toute grandeur  $u(x, t) = \text{cte}$  est solution de X-43. Cela implique d'avoir

$$b_3 + b_2 + b_1 = a_3 + a_2 + a_1$$

On y ajoute une normalisation arbitraire à un :

$$b_3 + b_2 + b_1 = a_3 + a_2 + a_1 = 1$$

X-44

Calculons le facteur d'amplification de X-43 :

$$(b_3 e^{j\phi} + b_2 + b_1 e^{-j\phi}) A^{k+1} = (a_3 e^{j\phi} + a_2 + a_1 e^{-j\phi}) A^k$$

En éliminant  $b_2$  et  $a_2$  grâce à X-44, il vient

$$G = \frac{1 - (a_3 + a_1)(1 - \cos \phi) + j(a_3 - a_1)\sin \phi}{1 - (b_3 + b_1)(1 - \cos \phi) + j(b_3 - b_1)\sin \phi} \quad X-45$$

dont on déduit

$$|G|^2 = \frac{|1 - (a_3 + a_1)(1 - \cos \phi) + j(a_3 - a_1)\sin \phi|^2}{|1 - (b_3 + b_1)(1 - \cos \phi) + j(b_3 - b_1)\sin \phi|^2} = \frac{1 + A_2 \left( \sin \frac{\phi}{2} \right)^2 + A_1 \left( \sin \frac{\phi}{2} \right)^4}{1 + B_2 \left( \sin \frac{\phi}{2} \right)^2 + B_1 \left( \sin \frac{\phi}{2} \right)^4} \quad X-46$$

si

$$\begin{aligned} A_1 &= 16a_3a_1 & A_2 &= 4[(a_3 - a_1)^2 - (a_3 + a_1)] \\ B_1 &= 16b_3b_1 & B_2 &= 4[(b_3 - b_1)^2 - (b_3 + b_1)] \end{aligned} \quad X-47$$

La condition de stabilité  $|G|^2 \leq 1$  s'écrit donc

$$(A_1 - B_1) \left( \sin \frac{\phi}{2} \right)^4 + (A_2 - B_2) \left( \sin \frac{\phi}{2} \right)^2 \leq 0$$

où  $0 \leq \left( \sin \frac{\phi}{2} \right)^2 \leq 1$ . Cela implique les conditions nécessaires et suffisantes de stabilité pour cette famille de schémas :

$$\begin{aligned} A_2 - B_2 &\leq 0 \\ A_1 - B_1 + A_2 - B_2 &\leq 0 \end{aligned} \quad X-48$$

### *Exemple : équation de la diffusion*

Le schéma V-6 fait partie de la famille traitée ci-dessus :

$$u_i^{k+1} = u_i^k + r(u_{i+1}^k - 2u_i^k + u_{i-1}^k) \quad X-49$$

$$\text{avec } r = \alpha \frac{\Delta t}{\Delta x^2}.$$

Pour X-49, on a  $b_3 = b_1 = 0$       et       $a_3 = a_1 = r$  ; il vient donc

$$\begin{array}{ll} A_1 = 16r^2 & A_2 = -8r \\ B_1 = 0 & B_2 = 0 \end{array}$$

et avec X-48 :

$$\begin{aligned} r &\geq 0 \\ 2r^2 - r &\leq 0 \end{aligned}$$

c'est-à-dire finalement

$$0 \leq r \leq \frac{1}{2}$$

confirmant le résultat déjà obtenu.

## Chapitre XI. Méthode matricielle pour l'analyse de la stabilité - Introduction à la méthode des lignes

L'objectif de la méthode matricielle est de pouvoir prendre en compte les conditions aux limites réelles dans l'étude de la stabilité : celles-ci n'interviennent en effet pas dans les méthodes de Von Neumann et de l'équation différentielle équivalente.

### XI.1 Stabilité matricielle

Afin de dégager les principes de cette méthode matricielle, considérons l'exemple suivant d'EDP traité au chapitre VIII :

$$u_t = \alpha u_{xx} \quad \text{XI-1}$$

avec les conditions aux limites

$$u(x_{\min}, t) = u_0(t) \quad u(x_{\max}, t) = u_N(t) \quad \text{XI-2}$$

Les discréétisations de la méthode explicite :

$$(u_t)_i^k = \frac{u_i^{k+1} - u_i^k}{\Delta t} \quad \text{XI-3}$$

et

$$(u_{xx})_i^k = \frac{u_{i+1}^k - 2u_i^k + u_{i-1}^k}{(\Delta x)^2} \quad \text{XI-4}$$

débouchent sur le schéma de calcul suivant, si on pose  $r = \frac{\alpha \Delta t}{\Delta x^2}$

$$u_i^{k+1} = ru_{i-1}^k + (1-2r)u_i^k + ru_{i+1}^k \quad \text{XI-5}$$

c'est-à-dire, sous forme matricielle et en tenant compte des conditions aux limites,

$$\begin{pmatrix} u_1^{k+1} \\ \vdots \\ u_i^{k+1} \\ \vdots \\ u_{N-1}^{k+1} \end{pmatrix} = \begin{pmatrix} 1-2r & r & & & \\ r & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & r \\ & & & r & 1-2r \end{pmatrix} \begin{pmatrix} u_1^k \\ \vdots \\ u_i^k \\ \vdots \\ u_{N-1}^k \end{pmatrix} + \begin{pmatrix} ru_0 \\ 0 \\ \vdots \\ 0 \\ ru_N \end{pmatrix} \quad \text{XI-6}$$

La méthode de la stabilité matricielle consiste à exprimer que les erreurs d'arrondi générées lors du calcul de chaque nouvel itéré  $u^k$  ne croissent pas indéfiniment quand  $k$  grandit : si on convient de noter  $e^k$  ces erreurs et  $\tilde{u}^k$  la solution de XI-6 débarrassée des erreurs, on a

$$u^k = \tilde{u}^k + e^k \quad \text{XI-7}$$

Ecrivons XI-6 sous la forme

$$u^{k+1} = Au^k + b$$

XI-8

Bien évidemment, on a aussi

$$\tilde{u}^{k+1} = A\tilde{u}^k + b$$

XI-9

de sorte que, par soustraction,  $e^k$  vérifie

$$e^{k+1} = Ae^k = A^2e^{k-1} = \dots = A^{k+1}e^0$$

XI-10

La stabilité matricielle impose la condition

$$\lim_{k \rightarrow \infty} \|e^k\| \text{ borné.}$$

XI-11

Ceci n'est possible que si la norme de la matrice A est inférieure ou égale à l'unité, ou encore si son rayon spectral vérifie

$$\rho(A) = \max_k |\lambda_k| \leq 1$$

XI-12

où  $\lambda_k$  sont ses valeurs propres. A est tridiagonale : on peut montrer que les valeurs propres d'une telle matrice écrite sous la forme

$$A = \begin{pmatrix} b & c & & & \\ a & \ddots & \ddots & & \\ & \ddots & \ddots & c & \\ & & a & b & \end{pmatrix}$$

XI-13

valent

$$\lambda_k = b + 2\sqrt{ac} \cos\left(\frac{k\pi}{N+1}\right) \quad k = 1, \dots, N$$

XI-14

où N est la dimension de la matrice. Pour la matrice de XI-6, on a donc

$$\lambda_k = 1 - 2r + 2r \cos\left(\frac{k\pi}{N}\right) \quad k = 1, \dots, N-1$$

XI-15

Les valeurs extrêmes de XI-15 sont atteintes pour  $N \rightarrow \infty$  et valent 1 et  $1 - 4r$ . XI-12 donne donc

$$|1 - 4r| \leq 1$$

$$\Leftrightarrow -1 \leq 1 - 4r \leq 1$$

$$\Leftrightarrow 0 \leq r \leq \frac{1}{2}$$

XI-15

condition déjà établie lors de l'étude de la stabilité par la méthode de Von Neumann.

On peut mesurer l'influence des conditions aux limites en modifiant XI-2 de diverses manières en en implémentant ces conditions de plusieurs façons : par exemple

Conditions aux limites	implémentation	Equations modifiées
Mixtes : $u_x(x_{\min}, t) = k_g$ $u(x_{\max}, t) = u_N(t)$	$(u_1 - u_0)/\Delta x = k_g$	en 1 : $u_1^{k+1} = (1-r)u_1^k + ru_2^k - rk_g \Delta x$
Neumann 1 : $u_x(x_{\min}, t) = k_g$ $u_x(x_{\max}, t) = k_d$	$(u_1 - u_0)/\Delta x = k_g$ $(u_N - u_{N-1})/\Delta x = k_d$	en 1 : $u_1^{k+1} = (1-r)u_1^k + ru_2^k - rk_g \Delta x$ en N-1 : $u_{N-1}^{k+1} = ru_{N-2}^k + (1-r)u_N^k + rk_d \Delta x$
Neumann 2 : $u_x(x_{\min}, t) = k_g$ $u_x(x_{\max}, t) = k_d$	$(u_1 - u_{-1})/2\Delta x = k_g$ $(u_{N+1} - u_{N-1})/2\Delta x = k_d$	en 0 : $u_0^{k+1} = (1-2r)u_0^k + 2ru_1^k - 2rk_g \Delta x$ en N : $u_N^{k+1} = 2ru_{N-1}^k + (1-2r)u_N^k + 2rk_d \Delta x$

Pour ces trois cas, XI-6 devient :

$$\text{Mixtes : } \begin{pmatrix} u_1^{k+1} \\ \vdots \\ u_i^{k+1} \\ \vdots \\ u_{N-1}^{k+1} \end{pmatrix} = \begin{pmatrix} 1-r & r & & & \\ r & 1-2r & r & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & r \\ & & & r & 1-2r \end{pmatrix} \begin{pmatrix} u_1^k \\ \vdots \\ u_i^k \\ \vdots \\ u_{N-1}^k \end{pmatrix} + \begin{pmatrix} -rk_g \Delta x \\ 0 \\ \vdots \\ 0 \\ ru_N \end{pmatrix} \quad \text{XI-18}$$

$$\text{Neumann 1 : } \begin{pmatrix} u_1^{k+1} \\ \vdots \\ u_i^{k+1} \\ \vdots \\ u_{N-1}^{k+1} \end{pmatrix} = \begin{pmatrix} 1-r & r & & & \\ r & 1-2r & r & & \\ & \ddots & \ddots & \ddots & \\ & & r & 1-2r & r \\ & & r & 1-r & \end{pmatrix} \begin{pmatrix} u_1^k \\ \vdots \\ u_i^k \\ \vdots \\ u_{N-1}^k \end{pmatrix} + \begin{pmatrix} -rk_g \Delta x \\ 0 \\ \vdots \\ 0 \\ rk_d \Delta x \end{pmatrix} \quad \text{XI-19}$$

$$\text{Neumann 2 : } \begin{pmatrix} u_0^{k+1} \\ \vdots \\ u_i^{k+1} \\ \vdots \\ u_N^{k+1} \end{pmatrix} = \begin{pmatrix} 1-2r & 2r & & & \\ r & 1-2r & r & & \\ & \ddots & \ddots & \ddots & \\ & & r & 1-2r & r \\ & & 2r & 1-2r & \end{pmatrix} \begin{pmatrix} u_0^k \\ \vdots \\ u_i^k \\ \vdots \\ u_N^k \end{pmatrix} + \begin{pmatrix} -2rk_g \Delta x \\ 0 \\ \vdots \\ 0 \\ 2rk_d \Delta x \end{pmatrix} \quad \text{XI-20}$$

Il n'existe pas d'expression analytique des valeurs propres de telles matrices ; tout au plus peuvent-elles être calculées numériquement pour une valeur de  $r$  donnée. Une telle étude montre que ces

matrices ont toutes un rayon spectral inférieur ou égal à un pour  $0 \leq r \leq \frac{1}{2}$ , comme pour XI-6.

La situation n'est pas toujours aussi simple, comme le montre l'exemple suivant : soit à résoudre

$$u_t = -au_x$$

XI-21

avec la condition aux limites

$$u(x_{\min}, t) = u_0(t)$$

XI-22

Les discrétisations de la méthode d'Euler :

$$(u_t)_i^k = \frac{u_i^{k+1} - u_i^k}{\Delta t}$$

XI-23

et

$$(u_x)_i^k = \frac{u_i^k - u_{i-1}^k}{\Delta x}$$

XI-24

débouchent sur le schéma de calcul suivant, si on pose  $v = \frac{a\Delta t}{\Delta x}$  :

$$u_i^{k+1} = vu_{i-1}^k + (1-v)u_i^k$$

XI-25

c'est-à-dire, sous forme matricielle et en tenant compte de la condition aux limites,

$$\begin{pmatrix} u_1^{k+1} \\ \vdots \\ u_i^{k+1} \\ \vdots \\ u_N^{k+1} \end{pmatrix} = \begin{pmatrix} 1-v & & & & \\ v & \ddots & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & v & 1-v \end{pmatrix} \begin{pmatrix} u_1^k \\ \vdots \\ u_i^k \\ \vdots \\ u_N^k \end{pmatrix} + \begin{pmatrix} vu_0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

XI-26

L'application de XI-14 donne immédiatement

$$\lambda_k = 1 - v \quad \forall k$$

XI-27

et XI-12 s'écrit donc

$$0 \leq |1 - v| \leq 1$$

XI-28

c'est-à-dire

$$0 \leq v \leq 2$$

XI-29

en désaccord avec la condition VII-10, dite condition CFL

$$0 \leq v \leq 1$$

XI-30

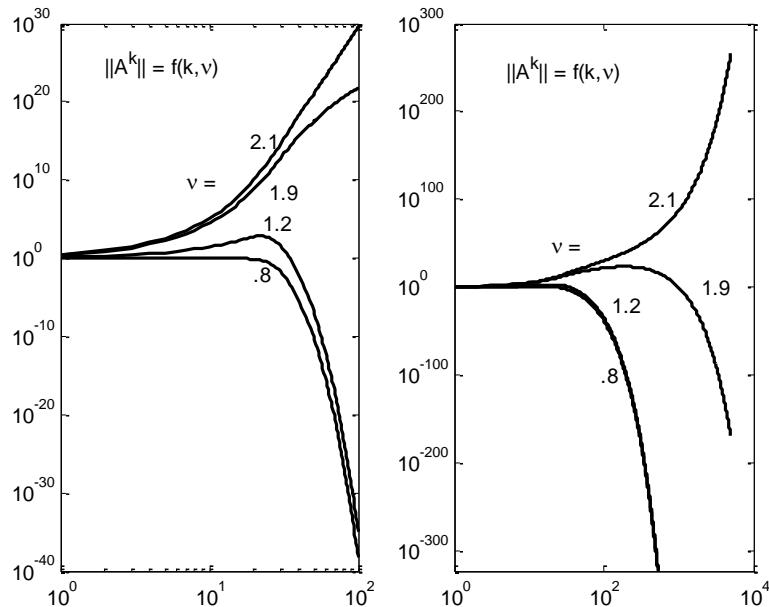
Ce désaccord est explicable : XI-11, dont découle XI-29 assure une stabilité strictement asymptotique : quand  $k$  tend vers l'infini,  $\|e^k\|$  tend bien vers zéro, mais rien ne prouve que pour  $k$  fini,  $\|e^k\|$  ne puisse pas prendre des valeurs intempestivement très élevées. Ceci est mis en évidence de la

manière suivante : on a vu que les valeurs prises par  $e^k$  sont liées à  $e^0$  (qui représente les erreurs d'arrondi à l'introduction de la condition initiale) par la relation

$$e^k = A^k e^0$$

XI-31

où  $A^k$  est la matrice  $A$  à la puissance  $k$ . La portée de XI-29 et celle de XI-30 sur la stabilité sont visualisables en représentant  $\|A^k\|$  en fonction de l'exposant  $k$ , pour diverses valeurs de  $v$  :



il y apparaît clairement le caractère asymptotique de la limite de stabilité  $v = 2$ , ainsi que le caractère strict de la stabilité assurée par la condition  $v = 1$ . Signalons que les graphes proposés correspondent à  $N=20$  ; l'allure reste qualitativement la même quand on modifie ce dernier paramètre, le retour à la stabilité, à valeur de  $v$  identique (inférieure ou égale à 2) apparaissant pour des valeurs de  $k$  d'autant plus grandes que  $N$  augmente. Pour la suite, la condition de stabilité à retenir doit être celle qui découle de la méthode de Von Neumann. Cela ne condamne pas pour autant la méthode matricielle : il y a moyen d'éviter les problèmes de contradiction qu'elle soulève vis-à-vis de la méthode de Von Neumann. En effet, il suffit pour cela de se rappeler que le point de départ de cette dernière méthode est l'hypothèse qui consiste à considérer que la solution  $u^k = (u_0^k \ u_1^k \ \dots \ u_N^k)^T$  à l'instant  $t = k\Delta t$  représente une période spatiale d'un signal se prolongeant infiniment vers la gauche et vers la droite. Ceci revient à écrire

$$u_0^k = u_N^k ; \quad u_1^k = u_{N+1}^k ; \quad \dots$$

XI-32

Introduit dans XI-26, XI-32 donne

$$\begin{bmatrix} u_1 \\ \dots \\ u_N \end{bmatrix}^{k+1} = \begin{bmatrix} 1-v & & v \\ v & \ddots & \\ & \ddots & \ddots \\ & & v & 1-v \end{bmatrix} \begin{bmatrix} u_1 \\ \dots \\ u_N \end{bmatrix}^k \quad XI-33$$

Le calcul des valeurs propres de la matrice de XI-33 peut être mené en remarquant que cette matrice est un cas particulier de la matrice générale suivante

$$C = \begin{bmatrix} c_1 & c_2 & c_3 & \dots & c_{N-1} & c_N \\ c_N & c_1 & c_2 & c_3 & \dots & c_{N-1} \\ c_{N-1} & c_N & c_1 & c_2 & c_3 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ c_3 & \dots & c_{N-1} & c_N & c_1 & c_2 \\ c_2 & c_3 & \dots & c_{N-1} & c_N & c_1 \end{bmatrix} \quad \text{XI-34}$$

dont les valeurs propres sont connues :

$$\Omega_k = \sum_{i=0}^{N-1} c_{i+1} \exp\left(j(k-1)\frac{i2\pi}{N}\right) \quad k = 1, \dots, N \quad \text{XI-35}$$

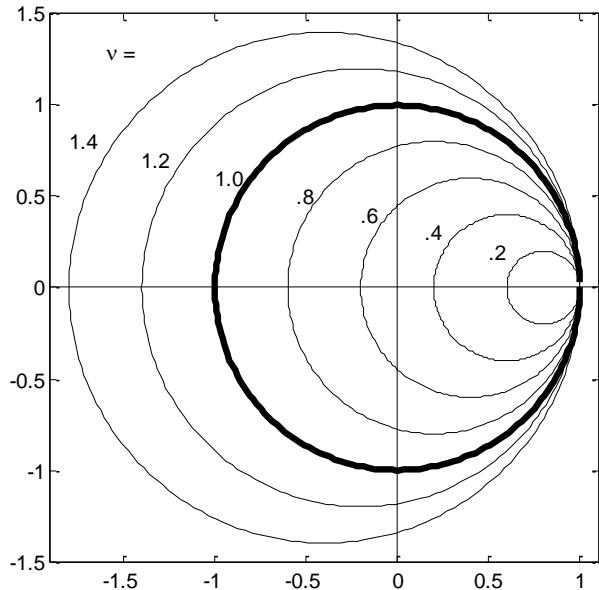
Pour XI-33 on a donc

$$\Omega_k = (1 - v) + ve^{\frac{j(k-1)(N-1)2\pi}{N}} \quad k = 1, \dots, N \quad \text{XI-36}$$

Ces valeurs propres sont situées dans le plan complexe le long d'un cercle de centre  $(1 - v, 0)$  et de rayon  $v$ . La figure suivante qui visualise ces cercles pour  $v = 0.2, 0.4, \dots, 1.4$  montre clairement que la limite de stabilité est cette fois

$$v \leq 1 \quad \text{XI-37}$$

conforme à la condition CFL.



Remarquons pour terminer que le passage de la matrice de XI-26 à celle de XI-33 est formellement facile à exécuter : il résulte de l'introduction de conditions aux limites particulières, dites périodiques : les conditions XI-32, auxquelles on fera appel par la suite lorsqu'on s'intéressera à la stabilité d'un schéma numérique. La méthode de Von Neumann et celle de la stabilité matricielle étant maintenant

équivalentes, il est intéressant de constater que la méthode de la stabilité matricielle présente un avantage par rapport à l'autre : elle offre la possibilité de séparer les influences des schémas de différences finies choisis pour calculer les dérivées spatiales d'une part et temporelle d'autre part. Cet avantage de la méthode matricielle est largement exploité dans la méthode des lignes.

## XI.2 Méthode des lignes – Introduction – Interaction avec la stabilité matricielle

Reprendons l'exemple déjà utilisé

$$u_t = -au_x$$

XI-21

avec la condition aux limites

$$u(x_{\min}, t) = u_0(t)$$

XI-22

et complété par la condition initiale

$$u(x, 0) = u^0(x)$$

XI-38

La méthode des lignes comporte pour l'essentiel les étapes suivantes :

1° on définit un maillage spatial sur le domaine d'étude : les seules valeurs de  $x$  en lesquelles  $u$  sera évalué sont

$$x_i = x_{\min} + i\Delta x ; \quad i:0 \rightarrow N ; \quad \Delta x = \frac{x_{\max} - x_{\min}}{N}$$

XI-39

2° l'équation XI-21 est écrite pour chaque valeur de  $x_i$  où  $u$  est inconnu :

$$(u_t)_i = -a(u_x)_i \quad i:1, \dots, N$$

XI-40

3° la dérivée spatiale est remplacée par une formule de différences finies, par exemple XI-24 :

$$(u_x)_i^k = \frac{u_i^k - u_{i-1}^k}{\Delta x}$$

XI-24

Ce remplacement transforme XI-40 en un système d'équations différentielles ordinaires en les fonctions du temps inconnues  $u_i(t)$  :

$$\begin{bmatrix} u_1 \\ \vdots \\ u_i \\ \vdots \\ u_N \end{bmatrix}_t = \begin{bmatrix} -\frac{a}{\Delta x}(u_1) \\ \vdots \\ -\frac{a}{\Delta x}(u_i - u_{i-1}) \\ \vdots \\ -\frac{a}{\Delta x}(u_N - u_{N-1}) \end{bmatrix} + \begin{bmatrix} \frac{a}{\Delta x}u_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

XI-41

4° la condition initiale XI-38 fournit la valeur de chacune des fonctions inconnues  $u_1, \dots, u_N$  à l'instant initial et XI-41 peut alors être intégré par une méthode classique : Euler, Runge-Kutta, prédiction-correction, ...

Les caractéristiques importantes de la méthode des lignes feront l'objet d'une étude séparée ; voyons seulement comment la méthode matricielle y sépare les influences sur la stabilité des schémas de différences finies choisis pour calculer les dérivées spatiales et temporelle. Comme on l'a vu dans les exemples précédents, la méthode matricielle requiert d'utiliser des conditions aux limites périodiques : ceci revient à remplacer XI-41 par

$$\begin{pmatrix} u_{1t} \\ \vdots \\ u_{it} \\ \vdots \\ u_{Nt} \end{pmatrix} = \frac{-a}{\Delta x} \begin{pmatrix} 1 & & & & -1 \\ -1 & 1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_i \\ \vdots \\ u_N \end{pmatrix} \quad \text{XI-42}$$

ou encore symboliquement

$$\frac{du}{dt} = Su \quad \text{XI-43}$$

où

$$u = [u_1 \dots u_N]^T \quad \text{XI-44}$$

### *stabilité de la discréétisation spatiale*

L'analyse mathématique nous enseigne que la solution exacte de systèmes tels que XI-43 dépend directement des valeurs propres et des vecteurs propres de la matrice S : si  $v_j$  désigne le vecteur propre associé à la valeur propre  $\Omega_j$ , on a

$$Sv_j = \Omega_j v_j \quad \text{XI-45}$$

Si le rang de la matrice S vaut N, les vecteurs  $v_j$  sont linéairement indépendants et constituent une base de  $R_N$  dans laquelle la solution exacte de XI-43 peut s'exprimer :

$$u(t) = \sum_{j=1}^N r_j(t) v_j \quad \text{XI-46}$$

XI-46 dans XI-43 donne

$$\frac{d}{dt} \left[ \sum_{j=1}^N r_j(t) v_j \right] = S \left[ \sum_{j=1}^N r_j(t) v_j \right] = \sum_{j=1}^N r_j(t) Sv_j, \quad \text{XI-47}$$

c'est-à-dire, grâce à XI-45

$$\frac{d}{dt} \left[ \sum_{j=1}^N r_j(t) v_j \right] = \sum_{j=1}^N r_j(t) \Omega_j v_j \quad XI-48$$

ou encore, pour une composante  $r_j$  quelconque

$$\frac{dr_j}{dt} = \Omega_j r_j \quad XI-49$$

XI-49 est l'équation différentielle ordinaire à laquelle satisfait  $r_j$ , appelé *mode*. Sa solution est

$$r_j(t) = r_j(0)e^{\Omega_j t} \quad XI-50$$

XI-50 fournit un renseignement capital sur la stabilité de la solution : si on imagine résoudre XI-49 en partant de l'état d'équilibre ( $r_j(t) = 0$  pour  $t < 0$ ) et en perturbant cette équation de manière extrêmement brève ( $r_j(0) = \delta(t)$ ), on conçoit que XI-50 est stable si  $r_j(t)$  reste borné quand  $t$  tend vers l'infini, c'est-à-dire si

$$\lim_{t \rightarrow \infty} r_j(t) \text{ fini}. \quad XI-51$$

Ceci exige d'avoir

$$\Re(\Omega_j) \leq 0 \quad \forall j \quad XI-52$$

XI-52 est la condition de stabilité cherchée. Cette condition ne dépend que de la discréétisation spatiale et des conditions aux limites, périodiques en l'occurrence. Elle est une condition nécessaire à la stabilité et est un préalable indispensable au choix d'un intégrateur temporel.

### *stabilité de l'intégration temporelle*

Repartons de XI-49 :

$$\frac{dr}{dt} = \Omega r \quad XI-53$$

où  $r$  et  $\Omega$  sont un mode et une valeur propre quelconques de  $u$ . En pratique, les intégrateurs utilisés sont de deux types : à pas multiples et à pas simple.

1° intégrateurs à pas multiples : appliqués à la résolution de  $u' = f(t, u)$  ils sont décrits par la récurrence linéaire

$$\alpha_k u_{n+k} + \dots + \alpha_0 u_n = \Delta t (\beta_k f_{n+k} + \dots + \beta_0 f_n). \quad XI-54$$

Appliqué à XI-53, cela donne

$$\alpha_k r_{n+k} + \dots + \alpha_0 r_n = \Delta t (\beta_k \Omega r_{n+k} + \dots + \beta_0 \Omega r_n) \quad XI-55$$

ou encore

$$(\alpha_k - \beta_k \Omega \Delta t) r_{n+k} + \dots + (\alpha_0 - \beta_0 \Omega \Delta t) r_n = 0$$

XI-56

Le polynôme caractéristique  $\rho(\xi)$  de cette récurrence s'écrit

$$\rho(\xi) = (\alpha_k - \beta_k \Omega \Delta t) \xi^k + \dots + (\alpha_0 - \beta_0 \Omega \Delta t) \xi^0$$

XI-57

La condition de stabilité temporelle est donc la suivante : il faut que les racines  $\xi_i$  de ce polynôme soient en module inférieures ou égales à un, et que les racines de module unitaire soient simples :

$$|\xi_i| \leq 1 \quad \forall i$$

XI-58

Dans le plan complexe de la variable  $z = \Omega \Delta t$ , le domaine de stabilité est donc défini par l'ensemble des  $z$  tel que les racines simples  $\xi_{si}$  et multiples  $\xi_{mi}$  de XI-57 où on a posé  $z = \Omega \Delta t$  vérifient

$$|\xi_{si}| \leq 1 \text{ et } |\xi_{mi}| < 1 \quad \forall i$$

Il est utile à ce stade de se rappeler que la famille des intégrateurs à pas multiples regroupe les méthodes BDF, les méthodes de Adams et de Milne-Simpson, et les méthodes de prédiction-correction.

a) les méthodes de prédiction-correction : soit par exemple le couple prédiction-correction :

$$\text{prédiction : } y_i^{(0)} = y_{i-1} + \frac{\Delta t}{24} [-9f_{i-4} + 37f_{i-3} - 59f_{i-2} + 55f_{i-1}] \quad \text{XI-59}$$

$$\text{correction : } y_i = y_{i-1} + \frac{\Delta t}{24} [f_{i-3} - 5f_{i-2} + 19f_{i-1} + 9f(t_i, y_i^{(0)})] \quad \text{XI-60}$$

Introduisons XI-53 dans XI-59 et 60 :

$$\text{prédiction : } r_i^{(0)} = r_{i-1} + \frac{\Delta t}{24} [-9\Omega r_{i-4} + 37\Omega r_{i-3} - 59\Omega r_{i-2} + 55\Omega r_{i-1}] \quad \text{XI-61}$$

$$\text{correction : } r_i = r_{i-1} + \frac{\Delta t}{24} [\Omega r_{i-3} - 5\Omega r_{i-2} + 19\Omega r_{i-1} + 9\Omega r_i^{(0)}] \quad \text{XI-62}$$

c'est-à-dire, en éliminant  $r_i^{(0)}$

$$\begin{aligned} r_i &= r_{i-1} + \frac{\Delta t}{24} \left[ \Omega r_{i-3} - 5\Omega r_{i-2} + 19\Omega r_{i-1} + 9\Omega \left( r_{i-1} + \frac{\Delta t}{24} [-9\Omega r_{i-4} + 37\Omega r_{i-3} - 59\Omega r_{i-2} + 55\Omega r_{i-1}] \right) \right] \\ &\Leftrightarrow \\ r_i &= r_{i-1} \left[ 1 + \frac{28}{24} \Omega \Delta t + \frac{495}{576} (\Omega \Delta t)^2 \right] - r_{i-2} \left[ \frac{5}{24} \Omega \Delta t + \frac{531}{576} (\Omega \Delta t)^2 \right] + r_{i-3} \left[ \frac{1}{24} \Omega \Delta t + \frac{333}{576} (\Omega \Delta t)^2 \right] - r_{i-4} \left[ \frac{81}{576} (\Omega \Delta t)^2 \right] \end{aligned} \quad \text{XI-63}$$

dont il découle le polynôme XI-57 suivant (avec  $z = \Omega \Delta t$ ) :

$$\rho(\xi) = \xi^4 - \left[ 1 + \frac{28}{24} z + \frac{495}{576} z^2 \right] \xi^3 + \left[ \frac{5}{24} z + \frac{531}{576} z^2 \right] \xi^2 - \left[ \frac{1}{24} z + \frac{333}{576} z^2 \right] \xi + \left[ \frac{81}{576} z^2 \right] \quad \text{XI-64}$$

La formule de correction pouvant être utilisée seule, appliquons le raisonnement précédent à XI-60 seule : on obtient le polynôme suivant :

$$\rho(\xi) = \left[1 - \frac{9}{24}z\right]\xi^3 - \left[1 + \frac{19}{24}z\right]\xi^2 + \left[\frac{5}{24}z\right]\xi - \left[\frac{1}{24}z\right] \quad \text{XI-65}$$

La frontière des domaines de stabilité de XI-64 et XI-65 est obtenue en remplaçant  $\xi$  par  $\exp(j\theta)$   $0 \leq \theta \leq 2\pi$  dans ces relations et en y recherchant les valeurs de  $z$  qui annulent  $\rho(\xi)$  :

pour la prédition-correction :

$$\left\{ z : \left[ -\frac{495}{576}e^{3j\theta} + \frac{531}{576}e^{2j\theta} - \frac{333}{576}e^{j\theta} + \frac{81}{576} \right]z^2 + \left[ -\frac{28}{24}e^{3j\theta} + \frac{5}{24}e^{2j\theta} - \frac{1}{24}e^{j\theta} \right]z + [e^{4j\theta} - e^{3j\theta}] = 0 \right\} \quad \text{XI-66}$$

pour la correction seule :

$$\left\{ z : \left[ -\frac{9}{24}e^{3j\theta} - \frac{19}{24}e^{2j\theta} + \frac{5}{24}e^{j\theta} - \frac{1}{24} \right]z + [e^{3j\theta} - e^{2j\theta}] = 0 \right\} \quad \text{XI-67}$$

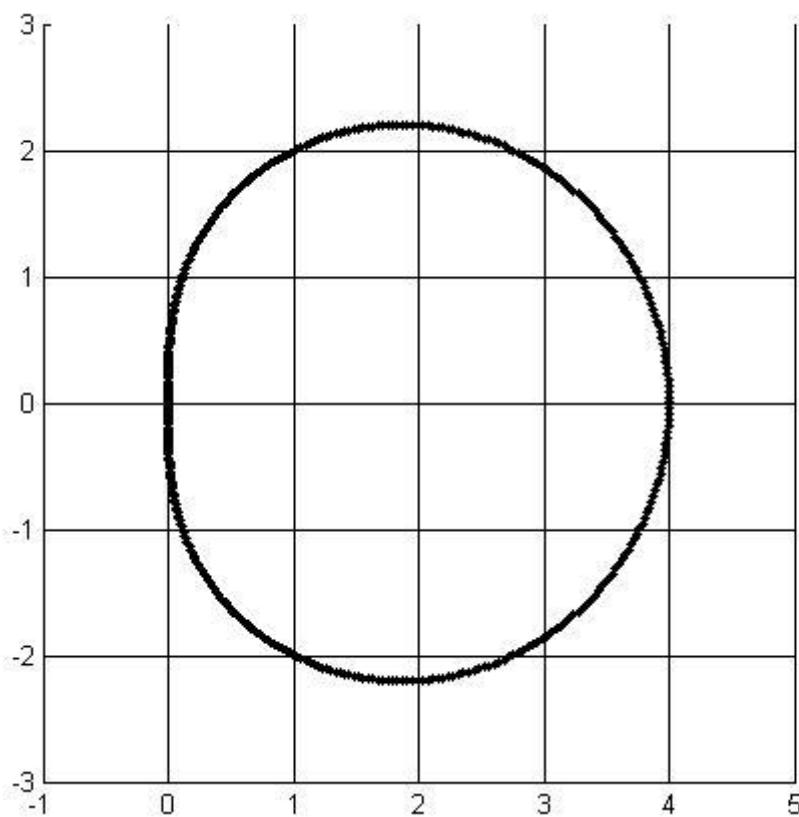
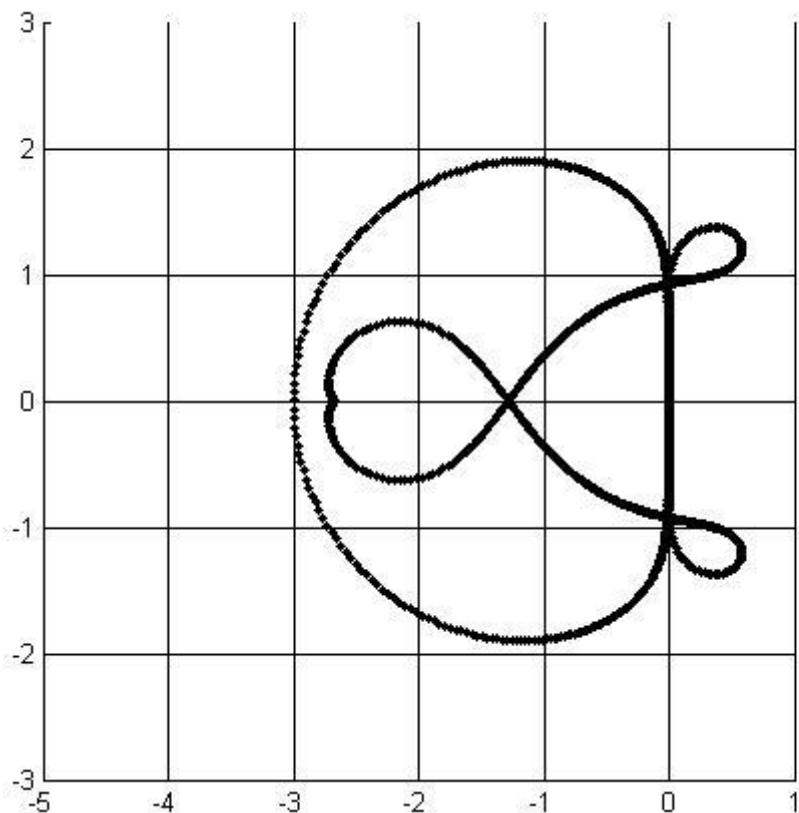
b) les méthodes BDF : soit la méthode décrite par

$$y_{n+1} = -\frac{1}{3}y_{n-1} + \frac{4}{3}y_n + \frac{2h}{3}f_{n+1} \quad \text{XI-68}$$

Un raisonnement identique débouche sur la frontière du domaine de stabilité suivant :

$$\left\{ z : \left[ -\frac{2}{3}e^{2j\theta} \right]z + \left[ e^{2j\theta} - \frac{4}{3}e^{j\theta} + \frac{1}{3} \right] = 0 \right\} \quad \text{XI-69}$$

XI-66 et XI-67 d'une part, et XI-69 d'autre part sont représentés sur les deux figures suivantes : la première montre clairement que le domaine de stabilité de la prédition-correction (courbe avec des points multiples) est plus petit que celui de la correction seule et la seconde que XI-69 est inconditionnellement stable (le domaine de stabilité est l'extérieur du contour) eu égard à la condition de stabilité XI-52.



2° intégrateurs à pas simple : les méthodes RK : rappelons que ces méthodes proposent le schéma de calcul suivant :

$$\begin{aligned}
k_1 &= f(t^0, u^0) \\
k_2 &= f(t^0 + c_2 \Delta t, u^0 + \Delta t a_{21} k_1) \\
k_3 &= f\left[t^0 + c_3 \Delta t, u^0 + \Delta t(a_{31} k_1 + a_{32} k_2)\right] \\
&\dots \\
k_s &= f\left[t^0 + c_s \Delta t, u^0 + \Delta t(a_{s1} k_1 + \dots + a_{ss-1} k_{s-1})\right] \\
u^1 &= u^0 + \Delta t [b_1 k_1 + b_2 k_2 + \dots + b_s k_s]
\end{aligned} \tag{XI-70}$$

Il est ais  de d uire ce que devient XI-70 si on l'applique 脿 XI-53 : par exemple, pour les m thodes 脿 un pas, on obtient

$$r^1 = r^0 + \Delta t b_1 f(t^0, r^0) = r^0 + \Delta t b_1 r^0 \Omega$$

Parmi toutes les m thodes 脿 un pas, la seule qui soit utile est la m thode d'Euler d j  rencontr e et pour laquelle on a

$$b_1 = 1$$

et donc

$$r^1 = r^0 (1 + \Omega \Delta t)$$

$r(t)$  restera born  pour autant que

$$|1 + \Omega \Delta t| \leq 1$$

Le domaine de stabilit  de la m thode d'Euler est donc la r gion du plan de la variable complexe  $z = \Omega \Delta t$  脿 l'int rieur de laquelle on a

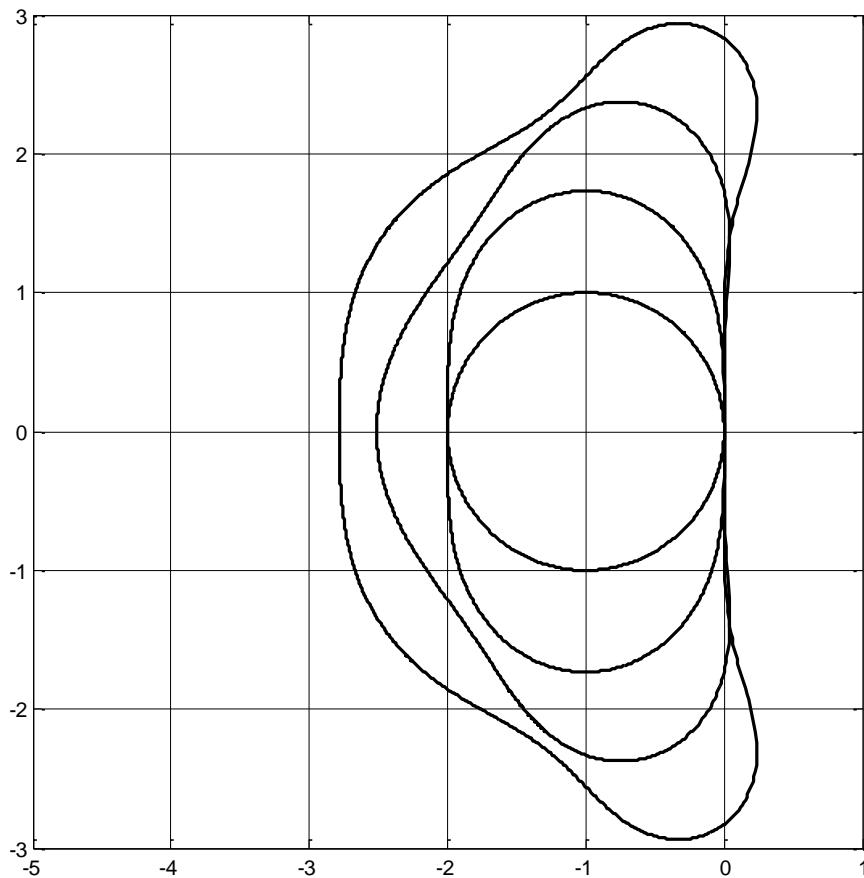
$$|1 + z| \leq 1$$

Il s'agit de l'int rieur du cercle de centre  $(-1, 0)$  et de rayon unitaire.

La g n ralisation de ce qui pr c de d bouche sur les domaines de stabilit  suivants pour les m thodes 脿 1, 2, 3 et 4  tages, respectivement d'ordre 1, 2, 3 et 4 :

m�thode 脿 un �tage :	$\left\{ z :  1 + z  \leq 1 \right\}$
m�thodes 脿 deux �tages :	$\left\{ z : \left  1 + z + \frac{z^2}{2!} \right  \leq 1 \right\}$
m�thodes 脿 trois �tages :	$\left\{ z : \left  1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} \right  \leq 1 \right\}$
m�thodes 脿 quatre �tages :	$\left\{ z : \left  1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \frac{z^4}{4!} \right  \leq 1 \right\}$

La figure suivante montre la forme de ces domaines dans le plan complexe  $z = \Omega \Delta t$  :



### **stabilité globale**

les conditions à réunir sont donc les suivantes :

1° il faut que les valeurs propres  $\Omega_j$  de la matrice de discrétisation spatiale, couplée à des conditions aux limites périodiques, soient à partie réelle négative ou nulle

2° il faut que les produits  $\Omega_j \Delta t$  soient situés à l'intérieur du domaine de stabilité de l'intégrateur choisi.

### **XI.3 Exemples**

Passons en revue quelques exemples vus dans les chapitres précédents

**Méthode leapfrog appliquée à XI-21 :  $u_t = -au_x$**

Cette méthode utilise les discrétisations suivantes :

$$\left( \frac{\partial u}{\partial t} \right)_i^k = \frac{u_i^{k+1} - u_i^{k-1}}{2\Delta t} \quad \left( \frac{\partial u}{\partial x} \right)_i^k = \frac{u_{i+1}^k - u_{i-1}^k}{2\Delta x} \quad \text{XI-71}$$

a) Matrice de discréétisation spatiale :

$$A = \frac{-a}{2\Delta x} \begin{pmatrix} 0 & 1 & & -1 \\ -1 & 0 & 1 & \\ & \ddots & \ddots & \ddots \\ & & -1 & 0 & 1 \\ 1 & & -1 & 0 & \end{pmatrix} \quad \text{XI-72}$$

Par XI-35, ses valeurs propres sont

$$\Omega_k = \frac{-a}{2\Delta x} \left( e^{j(k-1)\frac{2\pi}{N}} - e^{-j(k-1)\frac{(N-1)2\pi}{N}} \right) = \frac{-a}{\Delta x} j \sin \left( (k-1) \frac{2\pi}{N} \right) \quad \text{XI-73}$$

$\Omega_k$  est imaginaire pure : la stabilité de la discréétisation spatiale est assurée.

b) Intégrateur temporel : appliqué à XI-53, il s'écrit

$$\frac{r^{k+1} - r^{k-1}}{2\Delta t} = \Omega r^k \quad \text{XI-74}$$

Il est donc à pas multiples et son polynôme caractéristique vaut

$$\rho(\xi) = \xi^2 - 2\Omega\Delta t\xi - 1 = 0 \quad \text{XI-75}$$

La stabilité temporelle est donc assurée pour autant que les racines

$$\xi_{1,2} = \Omega\Delta t \pm \sqrt{1 + (\Omega\Delta t)^2} \quad \text{XI-76}$$

soient en valeur absolue inférieures ou égales à un.  $\Omega\Delta t$  est imaginaire pur :

$$\Omega\Delta t = \frac{-a\Delta t}{\Delta x} jy = -jvy \quad \text{avec } -1 \leq y \leq 1$$

$$\Rightarrow \xi_{1,2} = -jvy \pm \sqrt{1 - (vy)^2} \quad \text{XI-77}$$

La situation est la plus critique quand  $|y|=1$  :

$$\Rightarrow \xi_{1,2} = \pm jv \pm \sqrt{1 - (v)^2}$$

Quand  $v > 1$ ,  $\xi_{1,2}$  est imaginaire pur :

$$\xi_{1,2} = \pm j \left( v \pm \sqrt{(v)^2 - 1} \right)$$

et une des deux racines est plus grande que un : l'intégration temporelle est instable. Quand  $v \leq 1$ ,  $\xi_{1,2}$  est complexe et son module vaut

$$|\xi_{1,2}| = 1$$

et l'intégration est stable. La condition de stabilité est donc

$$v \leq 1$$

conforme à l'étude de Von Neumann.

### **Méthode explicite simple appliquée à XI-1 : $u_t = \alpha u_{xx}$**

Cette méthode utilise les discrétisations suivantes

$$\left( \frac{\partial u}{\partial t} \right)_i^k = \frac{u_i^{k+1} - u_i^k}{\Delta t} \quad \left( \frac{\partial^2 u}{\partial x^2} \right)_i^k = \frac{u_{i+1}^k - 2u_i^k + u_{i-1}^k}{(\Delta x)^2} \quad \text{XI-78}$$

a) Matrice de discrétisation spatiale :

$$A = \frac{\alpha}{\Delta x^2} \begin{pmatrix} -2 & 1 & & & 1 \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ 1 & & 1 & -2 & \end{pmatrix} \quad \text{XI-79}$$

Par XI-35, ses valeurs propres sont

$$\Omega_k = \frac{\alpha}{\Delta x^2} \left( -2 + e^{j(k-1)\frac{2\pi}{N}} + e^{j(k-1)\frac{(N-1)2\pi}{N}} \right) = \frac{-4\alpha}{\Delta x^2} \sin^2 \left( \frac{(k-1)\pi}{N} \right) \quad \text{XI-80}$$

$\Omega_k$  est réel négatif ou nul : la stabilité de la discrétisation spatiale est assurée.

b) Intégrateur temporel : appliqué à XI-53, il s'écrit

$$\frac{r^{k+1} - r^k}{\Delta t} = \Omega r^k \quad \text{XI-81}$$

Il est à un pas et son domaine de stabilité est donc  $\{z : |1+z| \leq 1\}$  : c'est le cercle de centre  $(-1,0)$  et de rayon unitaire.  $\Omega_k \Delta t$  sera inclus à ce domaine  $\forall k$  si

$$\frac{4\alpha \Delta t}{\Delta x^2} \leq 2 \quad \text{c'est-à-dire si} \quad r = \frac{\alpha \Delta t}{\Delta x^2} \leq \frac{1}{2} \quad \text{XI-82}$$

qui est conforme à l'étude de Von Neumann.

### Contre-exemples

Les associations qui suivent conduisent à des simulations instables ; l'explication provient à chaque fois du non respect des conditions précédentes :

a) résolutions de  $u_t = -au_x$  :

1)  $(\frac{\partial u}{\partial t})_i^k = \frac{u_i^{k+1} - u_i^k}{\Delta t}$  et  $(\frac{\partial u}{\partial x})_i^k = \frac{u_{i+1}^k - u_{i-1}^k}{2\Delta x}$  : les valeurs propres  $\Omega_k$  de la matrice de discréétisation spatiale sont imaginaires pures et la stabilité temporelle est assurée si  $\Omega_k \Delta t$  se trouve dans le cercle de centre  $(-1,0)$  et de rayon unitaire : impossible à satisfaire.

2)  $(\frac{\partial u}{\partial t})_i^k = \frac{u_i^{k+1} - u_i^{k-1}}{2\Delta t}$  et  $(\frac{\partial u}{\partial x})_i^k = \frac{u_i^k - u_{i-1}^k}{\Delta x}$  : la matrice de discréétisation spatiale vaut

$$A = \frac{-a}{\Delta x} \begin{pmatrix} 1 & & & & -1 \\ -1 & 1 & & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 1 & \\ & & & -1 & 1 \end{pmatrix} \Rightarrow \Omega_k = \frac{-a}{\Delta x} \left( 1 - e^{j(k-1)\frac{(N-1)2\pi}{N}} \right)$$

Les valeurs propres sont situées le long du cercle de centre  $\left(\frac{-a}{\Delta x}, 0\right)$  et de rayon  $\frac{a}{\Delta x}$  : elles sont à partie réelle négative ou nulle alors que la stabilité de l'intégration temporelle requiert (voir XI-81) d'avoir  $\Omega_k \Delta t$  sur la portion d'axe imaginaire  $[-j, j]$  : à nouveau c'est impossible à satisfaire.

b) résolution de  $u_t = \alpha u_{xx}$  :

$(\frac{\partial u}{\partial t})_i^k = \frac{u_i^{k+1} - u_i^{k-1}}{2\Delta t}$  et  $(\frac{\partial^2 u}{\partial x^2})_i^k = \frac{u_{i+1}^k - 2u_i^k + u_{i-1}^k}{(\Delta x)^2}$  : incompatible car  $\Omega_k$  est réel négatif ou nul et  $\Omega \Delta t$  doit être imaginaire pur.

### Cas où la séparation des influences des schémas temporel et spatial est impossible

C'est par exemple le cas de la méthode de Lax pour  $u_t = -au_x$  : le discréétisateur temporel ne fait pas partie des intégrateurs classiques ; c'est aussi le cas de la méthode de Cranck-Nicholson pour  $u_t = \alpha u_{xx}$  : ici, c'est le discréétisateur spatial qui fait défaut. Pour de tels schémas, outre l'analyse de Von Neumann toujours praticable, l'analyse matricielle peut être menée sans découplage des discréétisations : ainsi, pour la méthode de Cranck-Nicholson, le schéma numérique est (voir chapitre VIII) :

$$-\frac{r}{2}u_{i-1}^{k+1} + (1+r)u_i^{k+1} - \frac{r}{2}u_{i+1}^{k+1} = \frac{r}{2}u_{i-1}^k + (1-r)u_i^k + \frac{r}{2}u_{i+1}^k \quad \text{XI-83}$$

c'est-à-dire matriciellement, avec des conditions aux limites périodiques

$$\begin{pmatrix} 1+r & -\frac{r}{2} & & -\frac{r}{2} \\ -\frac{r}{2} & 1+r & -\frac{r}{2} & \ddots \\ & \ddots & \ddots & \ddots \\ & & -\frac{r}{2} & 1+r & -\frac{r}{2} \\ -\frac{r}{2} & & -\frac{r}{2} & 1+r \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \end{pmatrix}^{k+1} = \begin{pmatrix} 1-r & \frac{r}{2} & & \frac{r}{2} \\ \frac{r}{2} & 1-r & -\frac{r}{2} & \ddots \\ \ddots & \ddots & \ddots & \ddots \\ & & \frac{r}{2} & 1-r & \frac{r}{2} \\ \frac{r}{2} & & \frac{r}{2} & 1-r \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \end{pmatrix}^k \quad \text{XI-84}$$

Il s'agit donc de calculer les valeurs propres de

$$A = \begin{pmatrix} 1+r & -\frac{r}{2} & & -\frac{r}{2} \\ -\frac{r}{2} & 1+r & -\frac{r}{2} & \ddots \\ & \ddots & \ddots & \ddots \\ & & -\frac{r}{2} & 1+r & -\frac{r}{2} \\ -\frac{r}{2} & & -\frac{r}{2} & 1+r \end{pmatrix}^{-1} \begin{pmatrix} 1-r & \frac{r}{2} & & \frac{r}{2} \\ \frac{r}{2} & 1-r & -\frac{r}{2} & \ddots \\ \ddots & \ddots & \ddots & \ddots \\ & & \frac{r}{2} & 1-r & \frac{r}{2} \\ \frac{r}{2} & & \frac{r}{2} & 1-r \end{pmatrix} \quad \text{XI-85}$$

Un calcul numérique de ces valeurs propres montre qu'elles sont réelles, toujours comprises dans l'intervalle  $[-1, 1]$  quelle que soit la valeur de  $r$ , ce qui confirme la stabilité inconditionnelle déduite de l'analyse de Von Neumann.

## Chapitre XII. Généralités sur la méthodes des lignes

### XII.1 Introduction

L'étude générale des méthodes mises en œuvre pour résoudre les équations aux dérivées partielles par la méthode des différences finies confronte l'utilisateur à la difficulté de choisir la méthode la mieux adaptée au problème qu'il doit résoudre : il est habituel à cet égard de s'intéresser au caractère parabolique, elliptique ou hyperbolique de l'équation. C'est en effet en fonction de l'appartenance de l'équation étudiée à une de ces catégories que la méthode de résolution sera choisie.

Cette démarche présente plusieurs désavantages :

- la définition précise de ces catégories peut varier d'un auteur à l'autre,
- il est difficile pour un utilisateur non averti de déterminer l'appartenance d'une équation à l'une ou l'autre de ces familles,
- certaines équations peuvent appartenir à plusieurs familles selon la portion du domaine sur laquelle on les étudie,
- certaines équations n'appartiennent à aucune de ces catégories.

La méthode des lignes a le mérite d'éliminer ces difficultés par la généralité de son propos : rappelons-en les étapes : si l'équation à résoudre s'écrit (dans le cas du domaine spatial monodimensionnel)

$$L(u, u_t, u_x, u_{xx}, \dots, t) = 0 \quad \text{XII-1}$$

les étapes de la méthode sont les suivantes :

1° définition d'un maillage spatial sur le domaine d'étude : les seules valeurs de  $x$  en lesquelles  $u$  sera évalué sont

$$x_i = x_{\min} + i\Delta x ; \quad i: 0 \rightarrow N ; \quad \Delta x = \frac{x_{\max} - x_{\min}}{N} \quad \text{XII-2}$$

2° écriture de l'équation à résoudre en chaque valeur de  $x_i$  où  $u$  est inconnu :

$$L(u_i, (u_t)_i, (u_x)_i, (u_{xx})_i, \dots, t) = 0 \quad i = 1, \dots, N \quad \text{XII-3}$$

3° remplacement des dérivées spatiales par des approximations algébriques, par exemple des formules de différences finies telles que

$$(u_x)_i = \frac{u_i - u_{i-1}}{\Delta x} \quad (u_{xx})_i = \frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2} \quad \text{XII-4}$$

Ce remplacement transforme XII-3 en un système d'équations différentielles ordinaires en les fonctions du temps inconnues  $u_i(t)$

4° la condition initiale qui accompagne l'équation aux dérivées partielles fournit la valeur de chacune des fonctions inconnues  $u_1, \dots, u_N$  à l'instant initial et le système précédent peut alors être intégré par une méthode classique : Euler, Runge-Kutta, prédition-correction, ...

Rappelons encore l'exemple déjà évoqué :

$$u_t = -au_x$$

XII-5

complété de la condition initiale

$$u(x,0) = u^0(x)$$

XII-6

et de la condition aux limites

$$u(x_{\min}, t) = u_0(t)$$

XII-7

L'application de la méthode conduit, moyennant l'intervention de la formule de différences finies

$$(u_x)_i^k = \frac{u_i^k - u_{i-1}^k}{\Delta x}$$

XII-8

à la résolution du système d'équations différentielles ordinaires (ODE)

$$\begin{pmatrix} u_1 \\ \vdots \\ u_i \\ \vdots \\ u_N \end{pmatrix}_t = \frac{-a}{\Delta x} \begin{pmatrix} 1 & & & & \\ -1 & 1 & & & \\ & \ddots & \ddots & & \\ & & -1 & 1 & \\ & & & -1 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_i \\ \vdots \\ u_N \end{pmatrix} + \frac{-a}{\Delta x} \begin{pmatrix} u_0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

XII-9

Les étapes de la résolution d'une équation aux dérivées partielles par la méthode des lignes requièrent une connaissance avertie de plusieurs mécanismes typiques de l'analyse numérique :

- a) les techniques de maillage du domaine spatial,
- b) le remplacement des dérivées spatiales par des approximations algébriques,
- c) la discréétisation des conditions aux limites,
- d) l'intégration temporelle d'un système d'équations différentielles ordinaires.

Chacun de ces mécanismes sera envisagé plus en détail par la suite.

## XII.2 Implémentation de la méthode des lignes sous matlab

On se propose de présenter l'essentiel de la programmation en matlab de la résolution d'une équation aux dérivées partielles. C'est également au travers de cet exemple que les mécanismes précédents seront étudiés.

Soit donc à résoudre l'équation suivante, dite équation de Burgers :

$$u_t = -uu_x + \mu u_{xx}$$

XII-10

définie sur le domaine spatial  $0 \leq x \leq 1$  et avec  $t \geq 0$ .

Cette équation possède une solution analytique :

$$u(x, t) = \frac{0.1 \exp(a) + 0.5 \exp(b) + \exp(c)}{\exp(a) + \exp(b) + \exp(c)} \quad \text{XII-11}$$

avec

$$\begin{aligned} a &= -\frac{0.05}{\mu}(x - 0.5 + 4.95t) \\ b &= -\frac{0.25}{\mu}(x - 0.5 + 0.75t) \\ c &= -\frac{0.5}{\mu}(x - 0.375) \end{aligned} \quad \text{XII-12}$$

La connaissance de cette solution va nous permettre d'évaluer par comparaison la qualité de toute résolution numérique ; elle nous permet aussi de fixer la condition initiale et les conditions aux limites à utiliser : la condition initiale est

$$u(x, 0) = \frac{0.1 \exp(a) + 0.5 \exp(b) + \exp(c)}{\exp(a) + \exp(b) + \exp(c)} \quad \text{XII-13}$$

avec

$$\begin{aligned} a &= -\frac{0.05}{\mu}(x - 0.5) \\ b &= -\frac{0.25}{\mu}(x - 0.5) \\ c &= -\frac{0.5}{\mu}(x - 0.375) \end{aligned} \quad \text{XII-14}$$

et les conditions aux limites sont

$$u(0, t) = \frac{0.1 \exp(a) + 0.5 \exp(b) + \exp(c)}{\exp(a) + \exp(b) + \exp(c)} \quad \text{XII-15}$$

avec

$$\begin{aligned} a &= -\frac{0.05}{\mu}(-0.5 + 4.95t) \\ b &= -\frac{0.25}{\mu}(-0.5 + 0.75t) \\ c &= -\frac{0.5}{\mu}(-0.375) \end{aligned} \quad \text{XII-16}$$

et

$$u(1, t) = \frac{0.1 \exp(a) + 0.5 \exp(b) + \exp(c)}{\exp(a) + \exp(b) + \exp(c)} \quad \text{XII-17}$$

avec

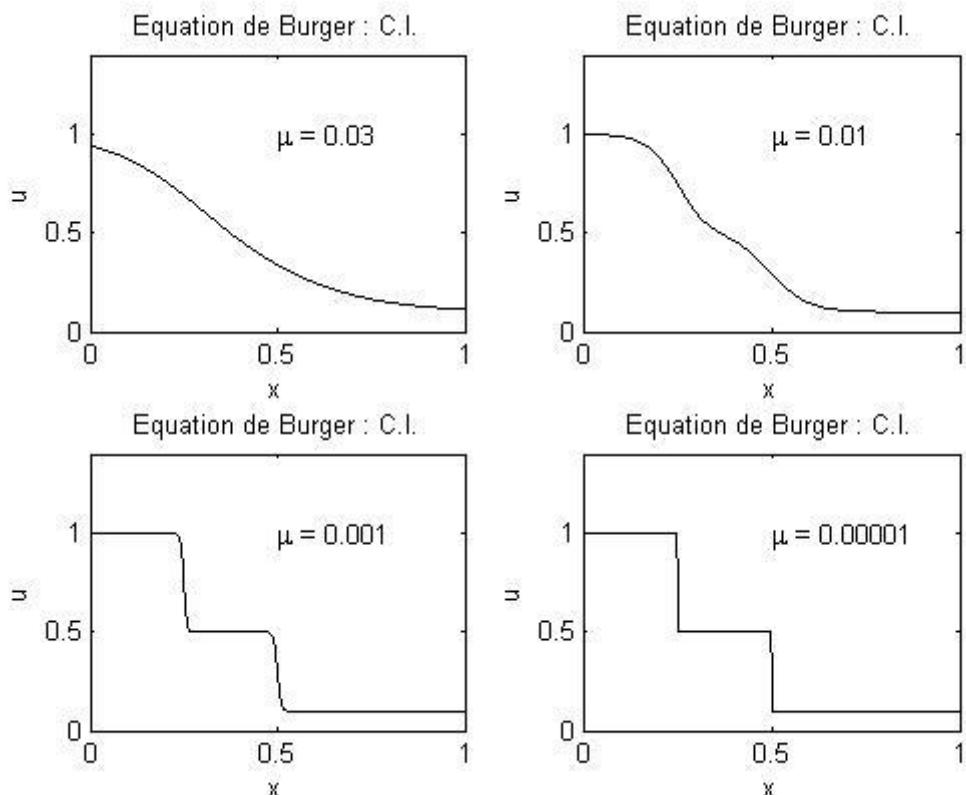
$$a = -\frac{0.05}{\mu} (0.5 + 4.95t)$$

$$b = -\frac{0.25}{\mu} (0.5 + 0.75t)$$

$$c = -\frac{0.5}{\mu} (0.625)$$

XII-18

Dans XII-10,  $\mu$  est un paramètre variable ; on trouvera ci-dessous le graphe de la condition initiale pour  $\mu$  égal à 0.03, 0.01, 0.001, 0.00001.



La résolution sous matlab requiert essentiellement la création des deux fichiers suivants :

1° un programme principal :

```
%...
%... the following code computes a solution to Burgers' equation
%...
close all
clear all
%...
%... start a stopwatch timer
tic
%...
%... set global variables
global mu;
global x0 xL n D1 D2;
%...
%... spatial grid
x0=0.0;
xL=1.0;
n=2001;
dx=(xL-x0)/(n-1);
x=[x0:dx:xL]';
%...
%... model parameter
mu=0.001;
%...
%... initial conditions
for i=1:n,
    u(i) = burgers_exact(x(i),0);
end;
%...
%... select finite difference (FD) approximations of the spatial
%... derivatives
v = 1;
D1=five_point_biased_upwind_uni_D1(x0,xL,n,v);
%...
D2=three_point_centered_uni_D2(x0,xL,n);
%...
%... call to ODE solver
%...
t=[0:0.1:1];

options = odeset('RelTol',1e-5,'AbsTol',1e-5,'stats','on');
%...
[tout, yout] = ode15s(@burgers_flux_pde,t,u,options);
plot(x,yout,'.-k');
xlabel('x');
ylabel('u(x,t)');
title('Equation de Burger')
hold on
%...
for k=1:length(tout),
    for i=1:n,
        yexact(i) = burgers_exact(x(i),tout(k));
    end;
    plot(x,yexact,'r')
end;
%...
%... read the stopwatch timer
tcpui=toc
```

On y trouve successivement :

- la fermeture de toute fenêtre matlab préalablement ouverte (`close all`) et l'annulation de toute définition préalable de variable (`clear all`),
- le démarrage d'un chronomètre (`tic`) en vue de mesurer le cpu,
- la déclaration en global d'une série de variables qui seront communes au programme principal et aux sous-programmes qui les reprendront également sous ce mode de déclaration (`global mu; global x0 xL n D1 D2`),
- la définition de la grille spatiale sous la forme d'un vecteur-colonne (`x=[x0:dx:xL]'`),
- la fixation de la valeur du paramètre  $\mu$  (`mu=0.001`),
- la création du vecteur des conditions initiales,
- le choix des schémas de différences finies utilisés pour évaluer numériquement les dérivées première et seconde apparaissant dans l'équation : il s'agit ici de deux matrices carrées D1 et D2 pour lesquelles des explications plus détaillées seront fournies ultérieurement,
- la définition des instants auxquels on souhaite visualiser la solution (`t=[0:0.1:1]`),
- la définition d'options utilisées par l'intégrateur temporel (`options = odeset('RelTol',1e-5, 'AbsTol',1e-5, 'stats','on')`) : dans le cas présent, il s'agit des tolérances relative et absolue, et de la volonté d'afficher des statistiques de calcul ; on verra plus loin que de nombreuses autres options sont possibles,
- le calcul proprement dit (`[tout,yout]=ode15s(@burgers_flux_pde,t,u,options)`) de la solution : `tout` est un vecteur reprenant les instants de visualisation ; `yout` est une matrice qui contient la solution : si `nprint` désigne le nombre d'instants de visualisation (`nprint` est la dimension de `tout`) et si `n` est le nombre de points de grille, `yout` est une matrice à `nprint` lignes et `n` colonnes ; `ode15s` est le nom de l'intégrateur temporel choisi (voir plus loin pour plus de détails sur les intégrateurs) ; `burgers_flux_pde` est le nom du sous-programme où seront implémentés les membres de droite des équations différentielles ordinaires résultant de l'application de la méthode des lignes ; `t` est le vecteur des instants de visualisation de la solution et `u` contient la solution initiale ; enfin, la liste d'appel de `ode15s` se termine par les options souhaitées,
- `plot(x,yout,'.-k')` qui dessine la solution obtenue : `yout` en fonction de `x`, avec des choix graphiques : `'.-k'` : voir dans le « help » de matlab,
- les légendes des axes et le titre du graphe ; l'instruction `hold on` permet la superposition de plusieurs courbes sur un même graphe,
- la boucle `for .... end` qui superpose la solution analytique au graphe précédent, permettant de juger de la qualité du résultat obtenu,
- `tcpu=toc` qui arrête le chronomètre et affiche le cpu.

2° un sous-programme :

```
function ut = burgers_flux_pde(t,u)
%...
%... set global variables
    global mu;
    global x0 xL n D1 D2;
    t
%...
%... boundary conditions at x = x0
    u(1) = burgers_exact(x0,t);
%...
%... boundary conditions at x = xL
    u(n) = burgers_exact(xL,t);
%...
%... second-order spatial derivative
%...
    ux = D2*u;
%...
%... first-order spatial derivative
    fx = ux.*u;
%...
%... temporal derivatives
%...
    ut = -fx + mu*ux;
    ut(1) = 0;
    ut(n) = 0;
```

On y trouve

- `function ut = burgers_flux_pde(t,u)` : la première ligne indique qu'il s'agit d'un sous-programme : `ut` en désigne la sortie (on verra qu'il s'agit des membres de droite des ODE) et `burgers_flux_pde` en est le nom : c'est par lui que le sous-programme sera appelé par le programme principal ; `t` est l'instant courant et `u` est le vecteur des variables dépendantes à l'instant courant,

- la définition des variables globales qui sont communes au programme principal,
- l'instruction `t` : elle permet lors de l'exécution d'afficher le déroulement du temps : on vérifie ainsi que le programme se déroule correctement,  
`u(1) = burgers_exact(x0,t)` et `u(n) = burgers_exact(xL,t)` : on impose les conditions aux limites, à partir de la solution analytique connue,
- le calcul des termes de dérivées spatiales,
- le calcul des membres de droites des ODE, stockés dans le vecteur de sortie du sous-programme :  
`ut = -fx + mu*uxx`,
- `ut(1) = 0` et `ut(n) = 0` : les dérivées temporelles de `u(1)` et de `u(n)` sont annulées : ces deux variables sont connues par les conditions aux limites et ne font donc pas l'objet d'une intégration temporelle.

3° programme de calcul de la solution analytique : outre les programmes précédents, on a également besoin d'un sous-programme de calcul de la solution analytique

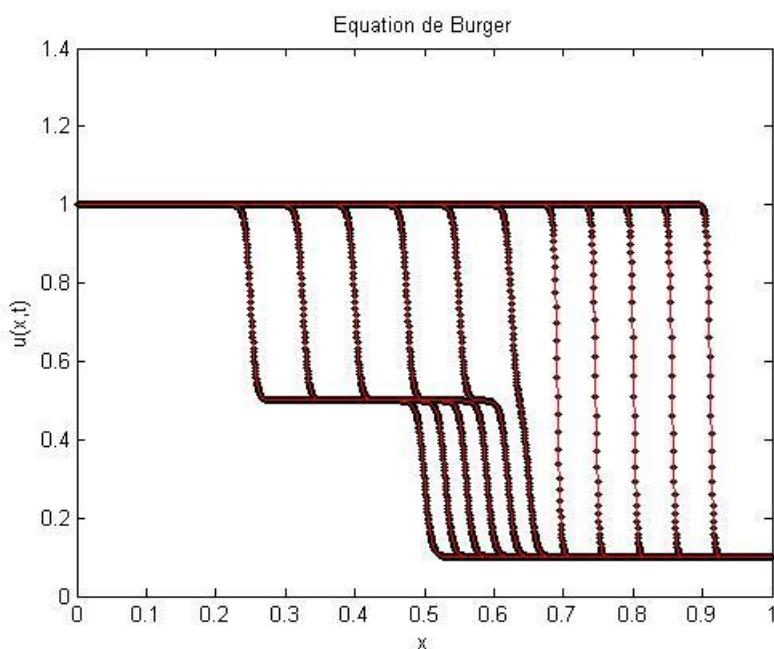
```

function [u] = burgers_exact(x,t)
%...
%... this function computes an exact solution to Burgers' equation
%...
global mu
%...
a(1) = -(0.05/mu)*(x-0.5+4.95*t);
a(2) = -(0.25/mu)*(x-0.5+0.75*t);
a(3) = -(0.5/mu)*(x-0.375);
%...
ea = 0;
eb = 0;
ec = 0;
temp = max(a);
if a(1)-temp >= -35
    ea = exp(a(1)-temp);
end
if a(2)-temp >= -35
    eb = exp(a(2)-temp);
end
if a(3)-temp >= -35
    ec = exp(a(3)-temp);
end
%...
ux = (0.1*ea+0.5*eb+ec)/(ea+eb+ec);

```

Outre l'implémentation des formules XII-11 et XII-12, on trouve une série d'instructions protégeant d'un overflow le calcul des exponentielles quand  $\mu$  devient trop petit.

L'exécution des programmes précédents débouche sur la figure suivante qui superpose les solutions analytique (trait continu) et numérique (point gras) aux instants de visualisation  $t = [0:0.1:1]$ .



L'excellente concordance entre ces deux solutions montre la qualité de la résolution numérique. Elle résulte de la mise en œuvre d'outils numériques dont le choix n'est pas toujours simple. Le propos des chapitres qui suivent est d'éclairer les utilisateurs sur ces choix.

## **Chapitre XIII. Méthode des lignes et différences finies**

Les différences finies constituent le procédé de calcul des dérivées spatiales le plus classique. Dans le cas de maillage régulier, ce procédé est particulièrement maniable, mais il existe toutefois actuellement des programmes permettant de calculer ces différences sur des maillages irréguliers, principalement dans le cas de fonctions ne dépendant que d'une seule variable spatiale :

$$\frac{\partial^n u}{\partial x^n} \quad \text{mais pas} \quad \frac{\partial^n u}{\partial x^r \partial y^s} \quad \text{avec } n = r + s \quad \text{XIII-1}$$

Ces programmes fournissent les coefficients  $k_i$  de la relation :

$$\frac{\partial^n u_i}{\partial x^n} = k_{-j} u_{i-j} + k_{-j+1} u_{i-j+1} + \dots + k_0 u_i + k_1 u_{i+1} + \dots + k_m u_{i+m} \quad \text{XIII-2}$$

permettant ainsi de calculer les équivalents sur maillage non uniforme des formules centrées et non centrées à nombre de points quelconque des maillages uniformes. Notons que le calcul de  $\frac{\partial^n u}{\partial x^n}$  est soit direct, soit déduit de l'application répétée d'un opérateur de dérivation élémentaire : par exemple

$$\frac{\partial^n u}{\partial x^n} = \frac{\partial}{\partial x} \left( \frac{\partial}{\partial x} \left( \frac{\partial}{\partial x} \left( \dots \frac{\partial u}{\partial x} \right) \right) \right) \quad \text{XIII-3}$$

Cette technique inhabituelle peut parfois se révéler plus efficace sur le plan de la limitation d'oscillations numériques.

L'ensemble des ces procédés fait l'objet de développements actuellement en cours ; les sous-programmes de calcul correspondants sont regroupés au sein du toolbox Matmol développé par le service Mathro et le service d'Automatique.

### **XIII.1 Les différences finies en maillage uniforme**

Matmol contient une série d'opérateurs de dérivation par formules de différences finies permettant d'évaluer principalement des dérivées première et seconde de fonctions d'une seule variable. Ces opérateurs sont des sous-programmes fonction à appeler à la demande et qui proposent tous le même type de produit : par exemple, le sous-programme de calcul d'une dérivée première par un schéma à deux points décentré vers l'amont en grille uniforme :

$$(u_i)_x = \frac{u_i - u_{i-1}}{\Delta x} \quad \text{XIII-4}$$

pour une convection de la gauche vers la droite et

$$(u_i)_x = \frac{u_{i+1} - u_i}{\Delta x} \quad \text{XIII-5}$$

pour une convection de la droite vers la gauche,

construit et a pour grandeur de sortie la matrice (pour l'implémentation de XIII-4)

$$D = \frac{1}{\Delta x} \begin{pmatrix} -1 & 1 & & & \\ -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix} \quad \text{XIII-6}$$

ou la matrice (pour l'implémentation de XIII-5)

$$D = \frac{1}{\Delta x} \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \\ & & & -1 & 1 \end{pmatrix} \quad \text{XIII-7}$$

Le calcul de la dérivée de  $u$  en tous les points de la grille est obtenu par simple multiplication :

$$u_x = Du \quad \text{XIII-8}$$

Le sous-programme est conçu de manière à pouvoir fonctionner avec les deux directions possibles : convection vers la droite et vers la gauche, via un paramètre  $v$  de la liste d'appel à fixer à 1 ou -1 (voir plus loin).

On remarquera que les calculs proposés en première ligne de XIII-6 et dernière ligne de XIII-7 correspondent à des formules décentrées dans le sens opposé à celui des autres lignes de ces matrices ; la raison en est la volonté de n'utiliser que des points intérieurs au domaine spatial (la volonté de garder le même schéma en tous les points de la grille conduirait à utiliser des nœuds fictifs extérieurs au domaine spatial).

## XIII.2 Les différences finies en maillage non uniforme

En grille non uniforme, le calcul découle de la propriété suivante : si l'ensemble des données connues d'une fonction  $u(x)$  est la table des valeurs

$$\begin{array}{ll} x_0 & u_0 \\ x_1 & u_1 \\ \vdots & \vdots \\ x_i & u_i , \\ x_{i+1} & u_{i+1} \\ \vdots & \vdots \\ x_N & u_N \end{array} \quad \text{XIII-9}$$

la formule des différences finies utilisant toutes ces données pour calculer la dérivée d'ordre  $m$  de  $u(x)$  au point d'indice  $i$ , soit  $(u_{mx})_i$ , est identique à la valeur prise par la dérivée  $m$ ième - calculée en  $x_i$  - du polynôme d'interpolation de  $u(x)$  construit à partir de cette table.

Fornberg propose une élégante méthode de calcul de  $(u_{mx})_i$  ou, plus précisément, des coefficients de la combinaison linéaire

$$(u_{nx})_i = \sum_{k=0}^N c_{Nk}^m u_k$$

XIII-10

Appelons  $P_{0...N}(x)$  le polynôme d'interpolation de  $u(x)$  construit à partir de XIII-9. On a donc

$$(u_{nx})_i = \left( \frac{d^m P_{0...N}(x)}{dx^m} \right)_{x_i} \quad \text{XIII-11}$$

$P_{0...N}(x)$  est le polynôme d'interpolation de Lagrange ; il est donné par la formule

$$P_{0...N}(x) = \sum_{k=0}^N \phi_{Nk}(x) u_k \quad \text{XIII-12}$$

où  $\phi_{Nk}(x)$  est appelé coefficient de Lagrange :

$$\phi_{Nk}(x) = \frac{(x - x_0) \dots (x - x_{k-1})(x - x_{k+1}) \dots (x - x_N)}{(x_k - x_0) \dots (x_k - x_{k-1})(x_k - x_{k+1}) \dots (x_k - x_N)} \quad \text{XIII-13}$$

XIII-11, 12 et 13 donnent alors

$$(u_{nx})_i = \sum_{k=0}^N \left( \frac{d^m \phi_{Nk}(x)}{dx^m} \right)_{x_i} u_k \quad \text{XIII-14}$$

A leur tour, XIII-10 et 14 fournissent

$$c_{Nk}^m = \left( \frac{d^m \phi_{Nk}(x)}{dx^m} \right)_{x_i} \quad \text{XIII-15}$$

C'est en partant de ce résultat théorique que Fornberg imagine une procédure itérative du calcul des coefficients  $c_{Nk}^1, c_{Nk}^2, \dots, c_{Nk}^m$  intervenant dans les formules de type XIII-10 et donnant en bloc toutes les dérivées d'ordre inférieur ou égal à  $m$  de  $u$  en  $x_i$ , soient

$$(u_x)_i = \sum_{k=0}^N c_{Nk}^1 u_k$$

$$(u_{2x})_i = \sum_{k=0}^N c_{Nk}^2 u_k \quad \text{XIII-16}$$

...

$$(u_{mx})_i = \sum_{k=0}^N c_{Nk}^m u_k$$

La procédure est basée sur les développements suivants : on déduit aisément de XIII-13 que  $\phi_{Nk}(x)$  est lié à  $\phi_{N-1k}(x)$  par la formule

$$\phi_{Nk}(x) = \phi_{N-1k}(x) \frac{x - x_N}{x_k - x_N}$$

XIII-17

utilisable quand  $k$  est différent de  $N$ .

Quand  $k = N$ ,  $\phi_{N-1N}$  n'existant pas, on doit se rabattre sur la formule liant  $\phi_{NN}$  à  $\phi_{N-1N-1}$  : on a

$$\phi_{NN}(x) = \frac{(x - x_0) \dots (x - x_{N-1})}{(x_N - x_0) \dots (x_N - x_{N-1})}$$

XIII-18

et

$$\phi_{N-1N-1}(x) = \frac{(x - x_0) \dots (x - x_{N-2})}{(x_{N-1} - x_0) \dots (x_{N-1} - x_{N-2})}$$

XIII-19

donc,

$$\phi_{NN}(x) = \phi_{N-1N-1}(x) \frac{(x - x_{N-1})(x_{N-1} - x_0) \dots (x_{N-1} - x_{N-2})}{(x_N - x_0) \dots (x_N - x_{N-1})}$$

XIII-20

Cette relation est formellement simplifiable en définissant

$$\omega_N(x) = (x - x_0) \dots (x - x_N)$$

XIII-21

Il vient effet

$$\phi_{NN}(x) = \phi_{N-1N-1}(x) \frac{\omega_{N-2}(x_{N-1})}{\omega_{N-1}(x_N)} (x - x_{N-1})$$

XIII-22

Les formules XIII-17 et 22 étant établies, explicitons  $\phi_{Nk}(x)$  : il s'agit d'un polynôme de degré  $N$  qui peut être exprimé sous la forme d'une somme de puissances de  $(x - x_i)$  :

$$\phi_{Nk}(x) = \alpha_0 + \alpha_1(x - x_i) + \dots + \alpha_N(x - x_i)^N$$

XIII-23

dont il est aisément de tirer

$$\alpha_m = \frac{1}{m!} \left( \frac{d^m \phi_{Nk}}{dx^m} \right)_{x_i}$$

XIII-24

On a donc

$$\phi_{Nk}(x) = \sum_{m=0}^N \frac{1}{m!} \left( \frac{d^m \phi_{Nk}}{dx^m} \right)_{x_i} (x - x_i)^m = \sum_{m=0}^N \frac{c_{Nk}^m}{m!} (x - x_i)^m$$

XIII-25

L'introduction de XIII-25 dans XIII-17 et 22 va finalement nous donner les formules itératives du calcul des coefficients  $c_{Nk}^m$  : dans XIII-17 :

$$\sum_{m=0}^N \frac{c_{Nk}^m}{m!} (x - x_i)^m = \sum_{m=0}^{N-1} \frac{c_{N-1k}^m}{m!} (x - x_i)^m \frac{x - x_N}{x_k - x_N} = \sum_{m=0}^{N-1} \frac{c_{N-1k}^m}{m!} (x - x_i)^m \frac{x - x_i + x_i - x_N}{x_k - x_N}$$

L'égalité des termes de même puissance en  $(x - x_i)$  donne la première formule de Fornberg :

$$\frac{c_{Nk}^m}{m!} = \frac{c_{N-1k}^{m-1}}{(m-1)!} \frac{1}{x_k - x_N} + \frac{c_{N-1k}^m}{m!} \frac{x_i - x_N}{x_k - x_N}$$

c'est-à-dire

$$c_{Nk}^m = \frac{1}{x_k - x_N} \left[ m c_{N-1k}^{m-1} + (x_i - x_N) c_{N-1k}^m \right]. \quad \text{XIII-26}$$

Le cas particulier  $k = N$  est déduit de XIII-22

$$\sum_{m=0}^N \frac{c_{NN}^m}{m!} (x - x_i)^m = \sum_{m=0}^{N-1} \frac{c_{N-1N-1}^m}{m!} (x - x_i)^m \frac{\omega_{N-2}(x_{N-1})}{\omega_{N-1}(x_N)} (x - x_i + x_i - x_{N-1})$$

L'égalité des termes de même puissance en  $(x - x_i)$  donne la deuxième formule de Fornberg :

$$\frac{c_{NN}^m}{m!} = \frac{\omega_{N-2}(x_{N-1})}{\omega_{N-1}(x_N)} \left[ \frac{c_{N-1N-1}^m}{m!} (x_i - x_{N-1}) + \frac{c_{N-1N-1}^{m-1}}{(m-1)!} \right]$$

c'est-à-dire

$$c_{NN}^m = \frac{\omega_{N-2}(x_{N-1})}{\omega_{N-1}(x_N)} \left[ c_{N-1N-1}^m (x_i - x_{N-1}) + m c_{N-1N-1}^{m-1} \right] \quad \text{XIII-27}$$

XIII-26 et XIII-27 sont les deux formules programmées par Fornberg ; on y observe clairement le caractère itératif du calcul, à la fois quant à l'ordre croissant des dérivées calculables (indice  $m$ ) qu'au nombre de points pris en compte pour calculer (indice  $N$ ). Ces deux formules sont programmées au sein d'un sous-programme function (attention aux notations : les abscisses des points de grille sont ici représentés par la lettre  $z$  et non  $x$ ) :

```
function [w]=weights(zd,zs,ns,m)
%...
%... weighting coefficients for finite difference approximations, computed
%... by an algorithm of B. Fornberg (1,2), are used in the following
%... approximations.
%...
%... (1) Fornberg, B., fast generation of weights in finite difference
%... formulas, in recent developments in numerical methods and
%... software for odes/daes/pdes, G. Byrne et al (eds), World
%... Scientific, River Edge, NJ, 1992
%...
%... (2) Fornberg, B., calculation of weights in finite difference
%... formulas, Siam Review, vol. 40, no. 3, pp 685-691, September,
%... 1999
%...
%...
%... slight adaptations by: W.E. Schiesser, P. Saucez and A. Vande Wouwer
```

```

%...
%... this function computes the weights of a finite difference scheme
%... on a nonuniform grid
%...
%... input Parameters
%...
%... zd           location where the derivative is to be computed
%...
%... ns           number of points in the stencil
%...
%... zs(ns)       stencil of the finite difference scheme
%...
%... m            highest derivative for which weights are sought
%...
%... output Parameter
%...
%... w(1:ns,1:m+1)  weights at grid locations z(1:ns) for
%...                  derivatives of order 0:m, found in w(1:ns,1:m+1)
%...
%...
c1 = 1.0;
c4 = zs(1)-zd;
for k=0:m
    for j=0:ns-1
        w(j+1,k+1) = 0.0;
    end
end
w(1,1) = 1.0;
for i=1:ns-1
    mn = min(i,m);
    c2 = 1.0;
    c5 = c4;
    c4 = zs(i+1)-zd;
    for j=0:i-1
        c3 = zs(i+1)-zs(j+1);
        c2 = c2*c3;
        if (j==i-1)
            for k=mn:-1:1
                w(i+1,k+1) = c1*(k*w(i,k)-c5*w(i,k+1))/c2;
            end
            w(i+1,1) = -c1*c5*w(i,1)/c2;
        end
        for k=mn:-1:1
            w(j+1,k+1) = (c4*w(j+1,k+1)-k*w(j+1,k))/c3;
        end
        w(j+1,1) = c4*w(j+1,1)/c3;
    end
    c1 = c2;
end

```

Ce sous-programme est lui-même appelé par les sous-programmes de calcul de dérivées, par exemple `function [D]=two_point_upwind_D1(x,v)` (où  $x$  est le vecteur des abscisses des points de la grille et  $v$  le paramètre indiquant le sens de la convection valant 1 ou -1, cf. paragraphe XIII.1). Ce dernier programme est utilisable aussi bien en grille uniforme que non uniforme. Son listing est le suivant (attention aux notations : les abscisses des points de grille sont ici représentés par la lettre  $z$  et non  $x$ ) :

```

function [D] = two_point_upwind_D1(z,v)
%...
%... The MatMol Group (2009)

```

```

%...
%... function two_point_upwind_D1 returns the differentiation matrix
%... for computing the first derivative, xz, of a variable x over a
%... nonuniform grid z from upwind two-point, first-order finite
%... difference approximations
%...
%... the following parameters are used in the code:
%...
%... z           spatial grid
%...
%... v           fluid velocity (positive from left to right - only
the sign is used)
%...
%... n           number of grid points
%...
%... zs(ns)      stencil of the finite difference scheme
%...
%... ns          number of points in the stencil
%...
%... zd          location where the derivative is to be computed
%...
%... m           highest derivative for which weights are sought
%...
m=1;
ns=2;
%...
%... sparse discretization matrix
n = length(z);
D = sparse(n,n);
%...
%... (1) finite difference approximation for positive v
if v > 0
%...
%... boundary point
zs=z(1:ns);
zd=z(1);
[w]=weights(zd,zs,ns,m);
D(1,1:2)=w(1:ns,m+1)';
%...
%... interior points
for i=2:n,
zs=z(i-1:i);
zd=z(i);
[w]=weights(zd,zs,ns,m);
D(i,i-1:i)=w(1:ns,m+1)';
end;
end;
%...
%... (2) finite difference approximation for negative v
if v < 0
%...
%... interior points
for i=1:n-1,
zs=z(i:i+1);
zd=z(i);
[w]=weights(zd,zs,ns,m);
D(i,i:i+1)=w(1:ns,m+1)';
end;
%...
%... boundary point
zs=z(n-1:n);

```

```

zd=z(n);
[w]=weights(zd,zs,ns,m);
D(n,n-1:n)=w(1:ns,m+1)';
%
end;

```

On observera que le nom du sous-programme renseigne l'utilisateur sur le type de dérivée calculée : d'abord le nombre de points utilisés dans la formule de différences finies (`two_point`), ensuite le décentrage éventuel (`upwind`), et l'opérateur de dérivation (`D1`) donnant l'ordre de la dérivée (ici première). Enfin, la liste d'appel (voir plus haut). La grandeur de sortie (`[D]`) contient la matrice XIII-6 ou XIII-7, selon le signe de `v`, ou l'équivalente de cette matrice en grille non uniforme, si `x` est lui-même non uniforme.

### XIII.3 Influence du choix des schémas de différences finies sur l'apparition d'oscillations parasites

Devant le large choix (en principe infini) des schémas de différences finies disponibles pour calculer des dérivées de tous ordres, il est naturel de s'interroger sur l'éventualité de privilégier certains de ces schémas ; le but est évidemment d'obtenir les résultats numériques les plus précis possibles et, à cet égard, diverses remarques peuvent être formulées :

- a priori, choisir des schémas à grand nombre de points réduit l'erreur de troncature,
- les schémas centrés sont plus précis que les schémas décentrés,
- on a vu, lors de l'exposé de la théorie relative à la stabilité matricielle, qu'il existait un lien entre schémas utilisables et choix de l'intégrateur temporel : on reviendra d'ailleurs sur ce point au chapitre intitulé **Stabilité et intégration temporelle**,

Les développements qui suivent ont pour but de mettre en évidence une situation fréquemment rencontrée : lorsque la solution numérique présente des oscillations parasites (c'est-à-dire non explicables par la physique des phénomènes décrits par les équations qu'on est en train de résoudre) importantes, une manière commode de s'en débarrasser est de privilégier l'usage de schémas upwind pour la simulation des phénomènes d'advection (c'est-à-dire pour le calcul des termes de dérivées spatiales premières). Voyons en détails sur un exemple simple l'impact de ce choix.

Soit à résoudre l'équation

$$-u \frac{d\phi}{dx} + k \frac{d^2\phi}{dx^2} = 0 \quad \text{avec} \quad 0 \leq x \leq L = 1 \quad \text{XIII-28}$$

et les conditions aux limites

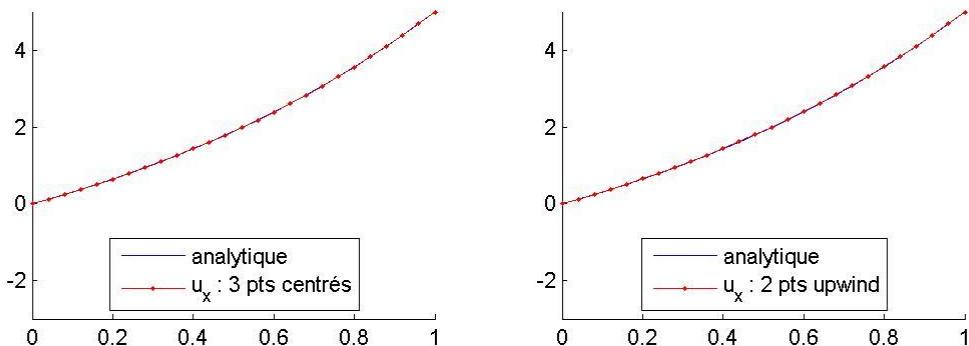
$$\phi(0) = 0 \quad \phi(1) = 5 \quad \text{XIII-29}$$

La solution analytique de ce problème est

$$\phi(x) = \frac{5 \left( \exp\left(\frac{u x}{k}\right) - 1 \right)}{\exp\left(\frac{u L}{k}\right) - 1} \quad \text{XIII-30}$$

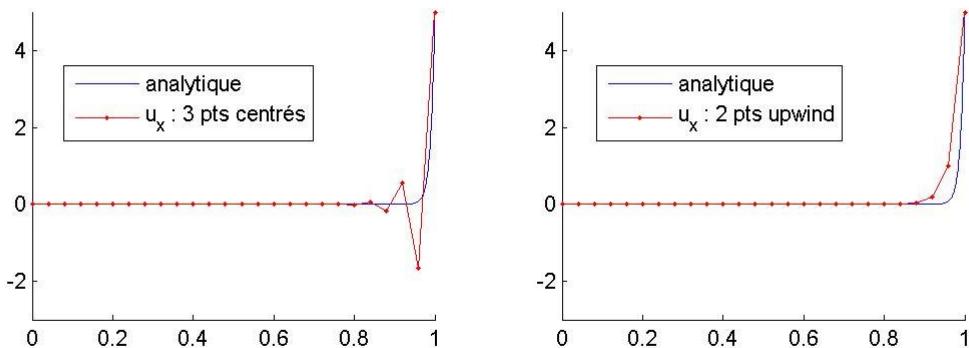
On se propose de résoudre cette équation avec différents schémas de différences finies, pour diverses valeurs de `u` et `k`. La grille spatiale utilisée est décrite par  $x_i = 0 + i\Delta x$  avec  $\Delta x = 0.04$ . Dans tous les cas proposés, la dérivée seconde est calculée par le schéma centré à trois points.

1°  $u = k = 1$  : la figure suivante superpose la solution analytique avec les solutions numériques obtenues avec la dérivée première calculée par le schéma centré à trois points et le schéma à deux points upwind :

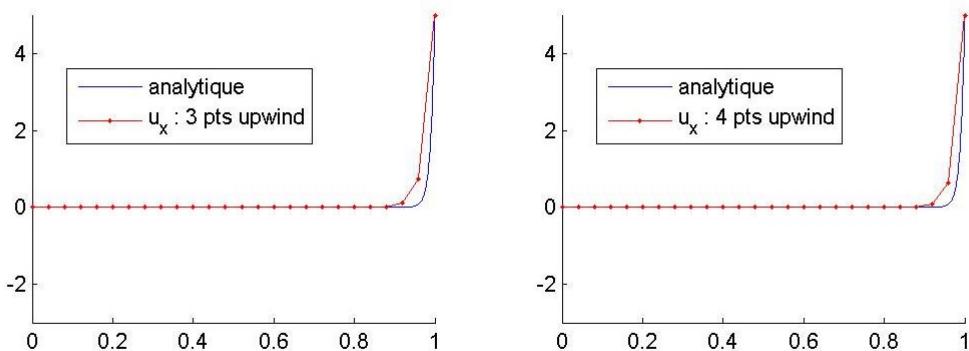


Pour ce cas, les deux schémas conduisent à une simulation quasi identique et de bonne qualité.

2°  $u = 100$  et  $k = 1$  : les mêmes schémas donnent cette fois :



Le schéma centré, quoique plus précis sur le plan de l'erreur de troncature, conduit à l'apparition d'oscillations éliminées si on utilise le schéma upwind à deux points. Celui-ci présente cependant un amortissement qu'il est possible de résorber, au moins en partie, en utilisant des schémas upwind plus sophistiqués : la figure suivante présente les résultats obtenus avec les schémas upwind à trois et quatre points :



Il est possible d'expliquer pourquoi le remplacement d'un schéma centré par un schéma upwind pour le calcul de la dérivée première atténue les oscillations : lorsque les deux dérivées sont calculées par des schémas à trois points centrés, l'équation aux différences finies s'écrit :

$$-u \frac{\phi_{i+1} - \phi_{i-1}}{2\Delta x} + k \frac{\phi_{i+1} - 2\phi_i + \phi_{i-1}}{\Delta x^2} = 0 \quad \text{XIII-31}$$

Lorsqu'on calcule la dérivée première par le schéma à deux points upwind, cela devient

$$-u \frac{\phi_i - \phi_{i-1}}{\Delta x} + k \frac{\phi_{i+1} - 2\phi_i + \phi_{i-1}}{\Delta x^2} = 0 \quad \text{XIII-32}$$

Il y a moyen de transformer cette dernière relation en y faisant apparaître le schéma à trois points pour le calcul de la dérivée première : une rapide manipulation algébrique permet d'écrire XIII-32 sous la forme

$$-u \frac{\phi_{i+1} - \phi_{i-1}}{2\Delta x} + \left( k + \frac{u\Delta x}{2} \right) \frac{\phi_{i+1} - 2\phi_i + \phi_{i-1}}{\Delta x^2} = 0 \quad \text{XIII-33}$$

Tout se passe donc comme si l'utilisation du schéma upwind avait introduit une diffusion supplémentaire de coefficient  $\tilde{k} = \frac{u\Delta x}{2}$  : c'est donc comme si on résolvait une équation dont la solution était forcément plus amortie.

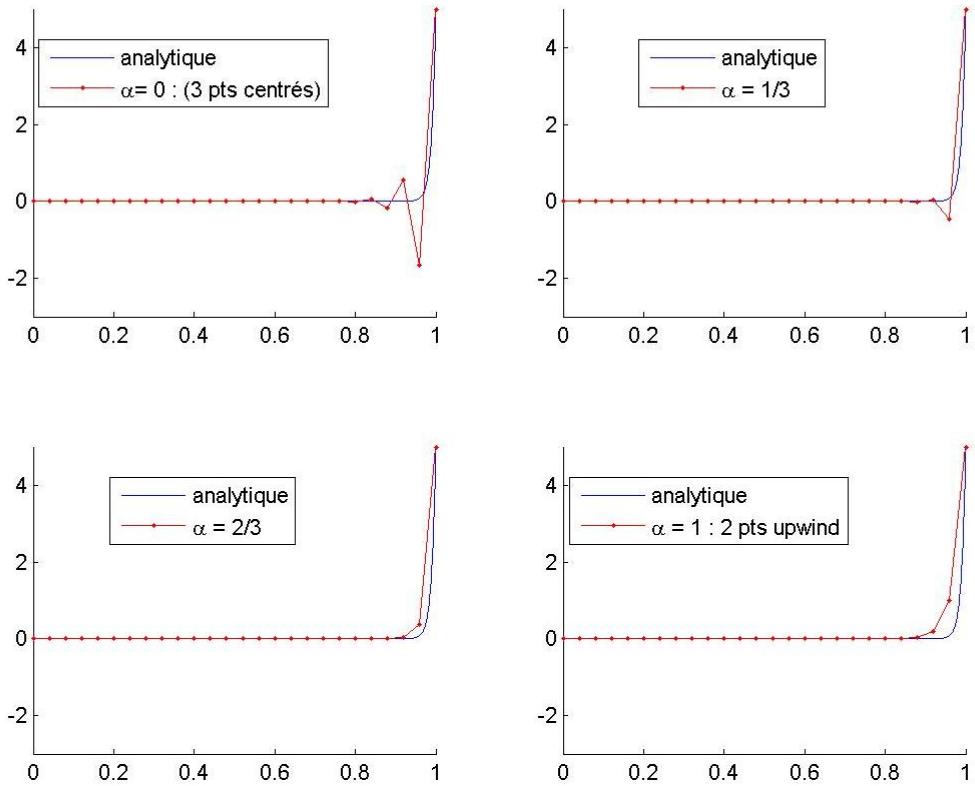
L'idée d'ajouter de la diffusion numérique peut être améliorée pour conduire à un schéma idéal : avec cette diffusion supplémentaire, XIII-33 s'écrit

$$-u \frac{\phi_{i+1} - \phi_{i-1}}{2\Delta x} + k \left( 1 + \frac{\tilde{k}}{k} \right) \frac{\phi_{i+1} - 2\phi_i + \phi_{i-1}}{\Delta x^2} = 0$$

En observant sur les figures précédentes que l'apport de cette diffusion numérique amortit trop les oscillations du schéma à trois points centrés, il vient naturellement l'idée de remplacer cette dernière relation par

$$-u \frac{\phi_{i+1} - \phi_{i-1}}{2\Delta x} + k \left( 1 + \alpha \frac{\tilde{k}}{k} \right) \frac{\phi_{i+1} - 2\phi_i + \phi_{i-1}}{\Delta x^2} = 0 \quad \text{avec } 0 \leq \alpha \leq 1 \quad \text{XIII-34}$$

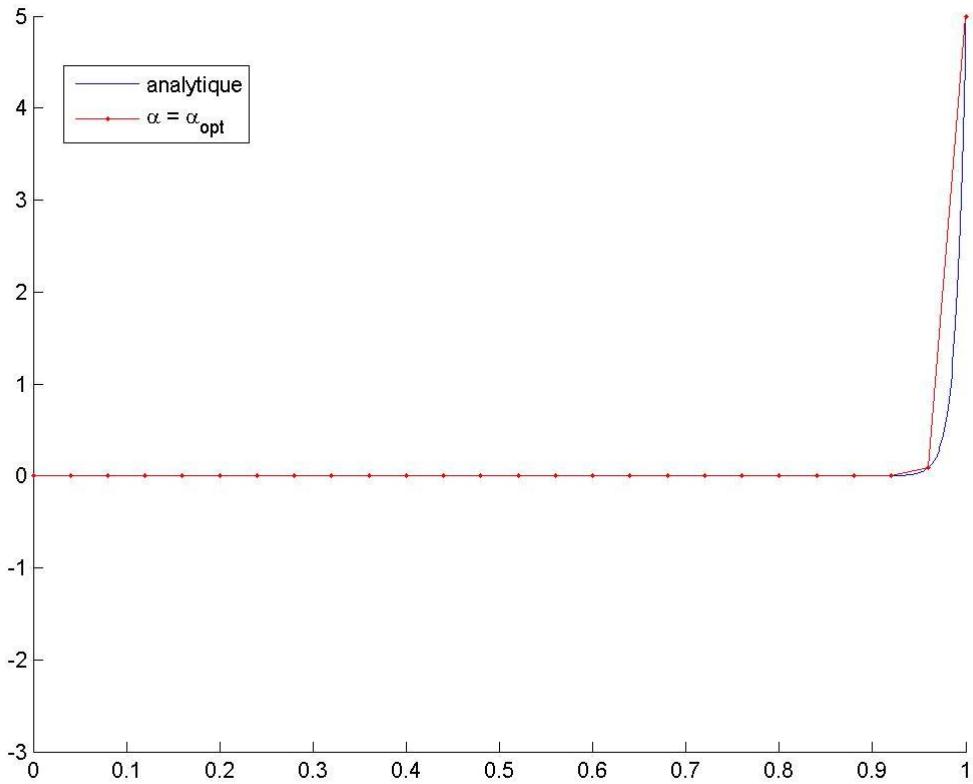
Les figures suivantes montrent les solutions obtenues pour diverses valeurs de  $\alpha$



En pratique, on peut montrer qu'il existe une valeur optimale de  $\alpha$  :

$$\alpha_{\text{opt}} = \frac{e^{Pe} + 1}{e^{Pe} - 1} - \frac{1}{Pe}$$

où  $Pe$  est le nombre de Péclet local :  $Pe = \frac{u\Delta x}{2k}$ . Pour cette valeur de  $\alpha$ , la solution numérique coïncide avec la solution analytique aux nœuds du maillage :



Cet ajustement optimal de la valeur de  $\alpha$  n'est malheureusement possible que pour quelques équations simples telles XIII-28 et on se contente en général du remplacement pur et simple des schémas centrés par des schémas upwind. Ce remplacement permet d'améliorer substantiellement la qualité des simulations quand les oscillations ne sont pas trop importantes. Néanmoins, dans les cas les plus difficiles il faudra recourir à une technique non linéaire : l'emploi des limiteurs de pente.

#### XIII.4 Les limiteurs de pente

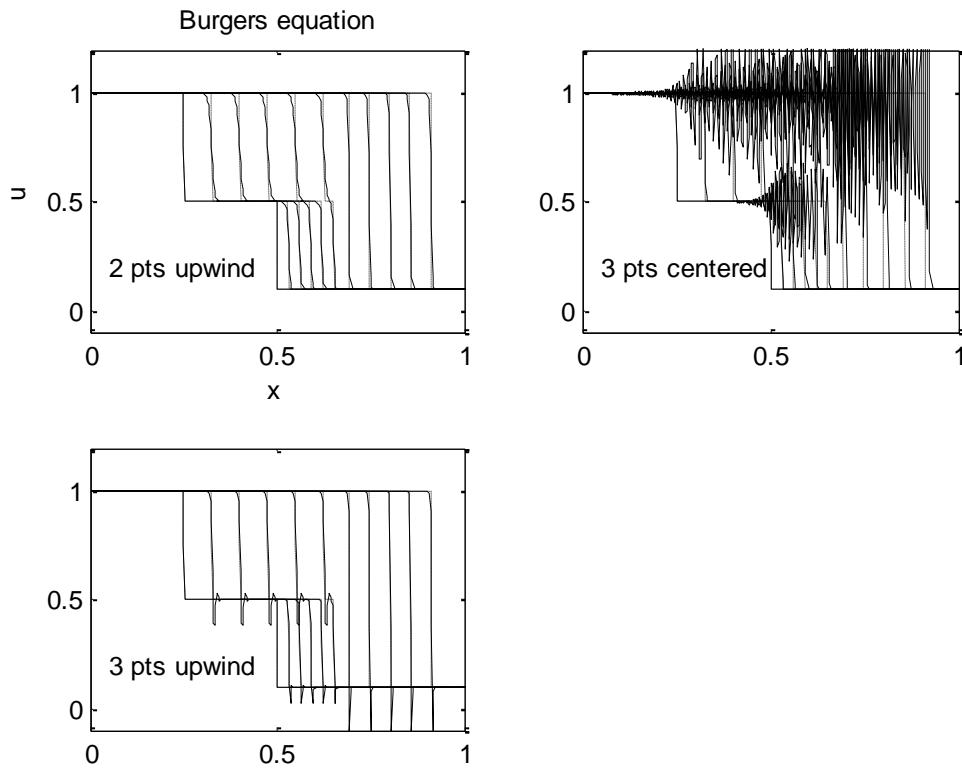
La résolution d'équations aux dérivées partielles de type hyperbolique présente des difficultés importantes quand la solution présente des fronts raides délimités par des points anguleux. Reprenons pour illustrer l'équation de Burgers utilisée au chapitre I :

$$u_t = -uu_x + \mu u_{xx} \quad \text{XIII-35}$$

qu'on réécrit selon

$$u_t = -\left(\frac{u^2}{2}\right)_x + \mu u_{xx} \quad \text{XIII-36}$$

définie sur  $0 \leq x \leq 1$ ,  $0 \leq t \leq 1$  et avec  $\mu = 0.00001$ . Pour cette très petite valeur de  $\mu$ , la simulation à l'aide de formules de différences finies est peu satisfaisante, ainsi que le montre la figure suivante où le nombre de points de grille vaut 401 :



et où les dérivées ont été estimées par les schémas de différences finies suivants :

dérivée première : deux points upwind, trois points centrés et trois points upwind  
dérivée seconde : trois points centrés.

Plusieurs constatations s'imposent :

1° la solution analytique (en trait interrompu sur la figure) est une fonction monotone décroissante de l'espace avec des fronts raides mobiles délimités par des points anguleux marqués.

2° le schéma d'ordre un (deux points upwind) préserve le caractère monotone de la solution mais il est incapable de reproduire correctement les points anguleux.

3° les schémas d'ordre supérieur (trois points centré et upwind) améliorent la reproduction des points anguleux mais présentent des oscillations parasites et perdent donc le caractère monotone de la solution.

De très nombreuses études ont été menées pour essayer de concilier ces exigences incompatibles en apparence : préserver le caractère monotone de la solution tout en conservant un ordre élevé, gage de précision dans le calcul de la solution. Le point de départ de ces études est un théorème célèbre établi par Lax en 1973 : pour toute loi de conservation scalaire

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} [f(u)] = 0$$

XIII-37

la variation totale de toute solution physique possible ne croît pas dans le temps, cette variation totale étant donnée par

$$TV = \int_{\text{dom}} \left| \frac{\partial u}{\partial x} \right| dx$$

XIII-38

Dans le cas d'une fonction discrète, XIII-38 devient

$$TV(u) = \sum_i |u_{i+1} - u_i|$$

XIII-39

Une méthode numérique est dite « total variation diminishing » (TVD) si

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

XIII-40

Si on convient d'appeler « schéma monotone » tout schéma préservant la monotonie de la solution, on peut alors montrer (Harten – 1983) que

1° tout schéma monotone est TVD

2° tout schéma TVD préserve la monotonie.

Ainsi que le montre la simulation de la figure précédente, le schéma de calcul de la dérivée première par la formule upwind deux points est un schéma TVD d'ordre faible ; il utilise l'approximation suivante de la dérivée première

$$\left( \frac{\partial f(u)}{\partial x} \right)_i = \frac{f(u_i) - f(u_{i-1})}{\Delta x}$$

XIII-41

qui est du premier ordre. L'obtention de schémas TVD d'ordre élevé reposera sur l'idée de base suivante : dans le voisinage des points anguleux, on s'arrangera pour que le schéma soit d'ordre un, et partout où la solution est « smooth », on imposera qu'il soit d'ordre supérieur à un. La transition entre schéma d'ordre un et schéma d'ordre supérieur à un peut se faire de deux manières : dans la première, si  $FD_{low}$  et  $FD_{high}$  sont des schémas de calcul de  $\left( \frac{\partial f(u)}{\partial x} \right)_i$  d'ordre un et d'ordre supérieur à un, le

schéma TVD final s'écrira

$$\left( \frac{\partial f(u)}{\partial x} \right)_i = FD_{low} + \phi(FD_{high} - FD_{low}).$$

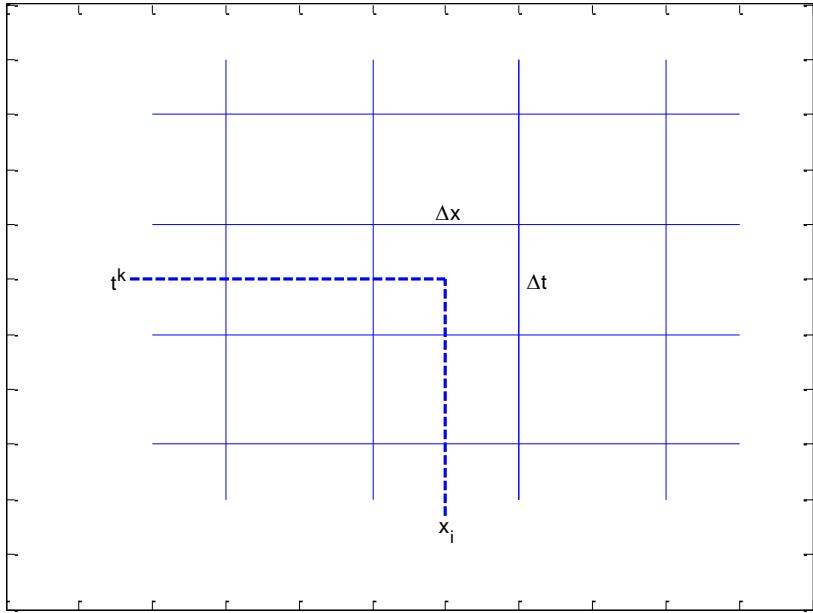
$\phi$  est appelé limiteur de flux (car  $f(u)$  est appelée « fonction de flux ») et c'est une fonction qui vaut zéro dans le voisinage des points anguleux et un là où la solution est smooth.

La seconde porte son attention sur la solution  $u$  proprement dite : elle impose que l'estimation  $u_i$  qui en est faite soit d'ordre un au voisinage des points anguleux et d'ordre supérieur ailleurs. C'est cette estimation qui est alors introduite dans la formule XIII-41. La transition de l'estimation d'ordre un à celle d'ordre supérieur est réalisée à l'aide de limiteurs dits de pente. C'est cette dernière technique qui est retenue pour la suite car on peut montrer qu'elle s'adapte mieux que celle des limiteurs de flux à la méthode des lignes.

Cette idée de base étant posée, le calcul proprement dit trouve son origine dans la notion de volume fini qui interprète XIII-37 de la manière suivante :  $\frac{\partial u}{\partial t}$  représente la variation temporelle de la grandeur  $u$  dans un petit volume de dimensions  $\Delta t$  et  $\Delta x$  à l'intérieur duquel  $u$  est supposé uniforme égal à

$$u_i^k = \frac{1}{\Delta t \Delta x} \int_{x_i - \frac{\Delta x}{2}}^{x_i + \frac{\Delta x}{2}} \int_{t^k - \frac{\Delta t}{2}}^{t^k + \frac{\Delta t}{2}} u(t, x) dt dx \quad \text{XIII-42}$$

Ceci invite à définir un maillage en  $x$  et en  $t$  tel que le point de coordonnées  $(x_i, t^k)$  se trouve au centre du petit volume  $(\Delta x, \Delta t)$  :



Si on oublie alors provisoirement le principe de la méthode des lignes (qui conserve les dérivées temporelles) pour remplacer toutes les dérivées par des différences finies, la discréétisation de  $\frac{\partial u}{\partial t}$  ne pose pas de problème particulier : quand on passe de  $t^k$  à  $t^{k+1}$ ,  $u$  passe de  $u^k$  à  $u^{k+1}$  et on écrit

$$\left( \frac{\partial u}{\partial t} \right)_i^k = \frac{u_i^{k+1} - u_i^k}{\Delta t}. \quad \text{XIII-43}$$

C'est moins simple pour  $\frac{\partial}{\partial x}[f(u)]$  : pour comprendre la discréétisation qui suit, il faut imaginer que  $f(u)$  est la grandeur responsable de la variation de « quantité de  $u$  » dans le petit volume : en cela,  $f(u)$  est bien un flux qui « traverse » les parois gauche et droite du petit volume ; c'est donc en évaluant ce flux lorsqu'il traverse ces parois qu'on traduit son action d'apport (ou de retrait) de quantité de  $u$  :

$$\left( \frac{\partial f(u)}{\partial x} \right)_i^k = \frac{f(u_{i+1/2}^k) - f(u_{i-1/2}^k)}{\Delta x} \quad \text{XIII-44}$$

$$\text{avec } f(u_{i\pm 1/2}^k) = f\left[u\left(t^k, x_i \pm \frac{\Delta x}{2}\right)\right] \quad \text{XIII-45}$$

La forme discrétisée complète de XIII-37 est donc

$$\frac{u_i^{k+1} - u_i^k}{\Delta t} = -\frac{1}{\Delta x} [f(u_{i+1/2}^k) - f(u_{i-1/2}^k)] \quad \text{XIII-46}$$

Ceci étant établi, on revient à la méthode des lignes en rétablissant la dérivation temporelle :

$$\frac{du_i}{dt} = -\frac{1}{\Delta x} [f(u_{i+1/2}) - f(u_{i-1/2})] \quad \text{XIII-47}$$

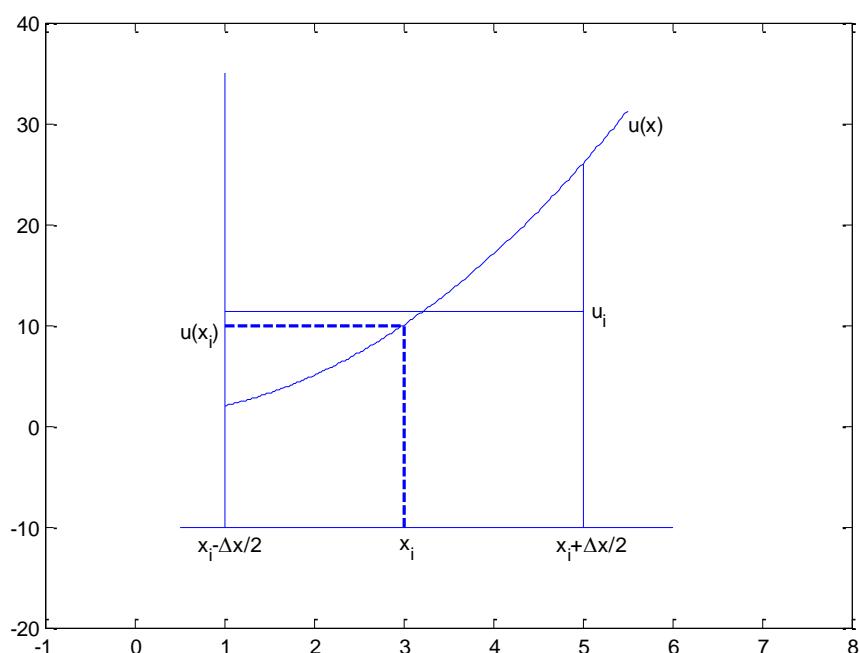
A ce stade, la définition des valeurs frontières  $u_{i\pm 1/2}$ , d'ordre un ou supérieur à un selon le caractère plus ou moins lisse de la solution, est délicate et dépend des considérations suivantes.

Rappelons que, en vertu de XIII-42,  $u_i$  désigne maintenant la valeur moyenne de  $u$  dans l'intervalle de largeur  $\Delta x$  :

$$u_i = \frac{1}{\Delta x} \int_{x_i - \frac{\Delta x}{2}}^{x_i + \frac{\Delta x}{2}} u(x) dx \quad \text{XIII-48}$$

(l'indice temporel  $k$  a disparu puisque la discrétisation temporelle a été supprimée).

Le calcul qui suit permet de comprendre comment élaborer des estimations d'ordre de plus en plus élevé de la valeur de  $u$  en tout point de l'intervalle  $[x_i - \frac{\Delta x}{2}, x_i + \frac{\Delta x}{2}]$ . Remarquons d'abord que  $u_i$  n'est pas forcément égal à  $u(x_i)$  :



Partons maintenant du développement de Taylor de  $u(x)$  autour de  $u(x_i)$  :

$$u(x) = u(x_i) + (x - x_i) \left( \frac{du}{dx} \right)_{x_i} + \frac{(x - x_i)^2}{2!} \left( \frac{d^2u}{dx^2} \right)_{x_i} + \dots \quad \text{XIII-49}$$

Le but des développements qui suivent est d'établir une relation analogue à XIII-49 qui fasse intervenir les valeurs moyennes XIII-48. Pour cela, remplaçons d'abord  $u(x)$  par XIII-49 dans XIII-48 :

$$u_i = \frac{1}{\Delta x} \int_{x_i - \frac{\Delta x}{2}}^{x_i + \frac{\Delta x}{2}} \left[ u(x_i) + (x - x_i) \left( \frac{du}{dx} \right)_{x_i} + \frac{(x - x_i)^2}{2!} \left( \frac{d^2u}{dx^2} \right)_{x_i} + \dots \right] dx \quad \text{XIII-50}$$

dont on déduit aisément

$$u_i = u(x_i) + \frac{\Delta x^2}{24} \left( \frac{d^2u}{dx^2} \right)_{x_i} + \dots \quad \text{XIII-51}$$

Modifions ensuite légèrement XIII-49 :

$$u(x) = u_i - u_i + u(x_i) + (x - x_i) \left( \frac{du}{dx} \right)_{x_i} + \frac{(x - x_i)^2}{2!} \left( \frac{d^2u}{dx^2} \right)_{x_i} + \dots \quad \text{XIII-52}$$

et introduisons XIII-51 dans XIII-52 :

$$u(x) = u_i + (x - x_i) \left( \frac{du}{dx} \right)_{x_i} + \left( \frac{(x - x_i)^2}{2!} - \frac{\Delta x^2}{24} \right) \left( \frac{d^2u}{dx^2} \right)_{x_i} + \dots \quad \text{XIII-53}$$

Les dérivées  $\left( \frac{du}{dx} \right)_{x_i}$  et  $\left( \frac{d^2u}{dx^2} \right)_{x_i}$  doivent également être évaluées à partir des valeurs moyennes ...,  $u_{i-1}, u_i, u_{i+1}, \dots$ . Pour cela, repartons de XIII-51 appliqué aux intervalles d'indices  $i-1$  et  $i+1$  :

$$u_{i-1} = u(x_{i-1}) + \frac{\Delta x^2}{24} \left( \frac{d^2u}{dx^2} \right)_{x_{i-1}} + \dots \quad \text{XIII-54}$$

$$u_{i+1} = u(x_{i+1}) + \frac{\Delta x^2}{24} \left( \frac{d^2u}{dx^2} \right)_{x_{i+1}} + \dots \quad \text{XIII-55}$$

et tirons-en les équivalents en valeurs moyennes des formules de différences finies classiques de dérivées première et seconde :

$$\frac{u_{i+1} - u_{i-1}}{2\Delta x} = \frac{u(x_{i+1}) - u(x_{i-1})}{2\Delta x} + \frac{\Delta x}{48} \left[ \left( \frac{d^2u}{dx^2} \right)_{x_{i+1}} - \left( \frac{d^2u}{dx^2} \right)_{x_{i-1}} \right] + \dots \quad \text{XIII-56}$$

et

$$\frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2} = \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{\Delta x^2} + \frac{1}{24} \left[ \left( \frac{d^2 u}{dx^2} \right)_{x_{i+1}} - 2 \left( \frac{d^2 u}{dx^2} \right)_{x_i} + \left( \frac{d^2 u}{dx^2} \right)_{x_{i-1}} \right] + \dots$$

XIII-57

Les formules de différences finies classiques sont, elles, bien connues :

$$\frac{u(x_{i+1}) - u(x_{i-1})}{2\Delta x} = \left( \frac{du}{dx} \right)_{x_i} + \frac{\Delta x^2}{3!} \left( \frac{d^3 u}{dx^3} \right)_{x_i} + \dots$$

XIII-58

et

$$\frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{\Delta x^2} = \left( \frac{d^2 u}{dx^2} \right)_{x_i} + \frac{\Delta x^2}{12} \left( \frac{d^4 u}{dx^4} \right)_{x_i} + \dots$$

XIII-59

XIII-58 dans XIII-56 et XIII-59 dans XIII-57 donnent

$$\frac{u_{i+1} - u_{i-1}}{2\Delta x} = \left( \frac{du}{dx} \right)_{x_i} + \frac{\Delta x^2}{3!} \left( \frac{d^3 u}{dx^3} \right)_{x_i} + \frac{\Delta x}{48} \left[ \left( \frac{d^2 u}{dx^2} \right)_{x_{i+1}} - \left( \frac{d^2 u}{dx^2} \right)_{x_{i-1}} \right] + \dots$$

XIII-60

et

$$\frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2} = \left( \frac{d^2 u}{dx^2} \right)_{x_i} + \frac{\Delta x^2}{12} \left( \frac{d^4 u}{dx^4} \right)_{x_i} + \frac{1}{24} \left[ \left( \frac{d^2 u}{dx^2} \right)_{x_{i+1}} - 2 \left( \frac{d^2 u}{dx^2} \right)_{x_i} + \left( \frac{d^2 u}{dx^2} \right)_{x_{i-1}} \right] + \dots$$

XIII-61

Pour être complet, il ne reste plus qu'à évaluer  $\left( \frac{d^2 u}{dx^2} \right)_{x_{i\pm 1}}$  : à nouveau, par Taylor, on a

$$\left( \frac{d^2 u}{dx^2} \right)_{x_{i-1}} = \left( \frac{d^2 u}{dx^2} \right)_{x_i} - \Delta x \left( \frac{d^3 u}{dx^3} \right)_{x_i} + \frac{\Delta x^2}{2!} \left( \frac{d^4 u}{dx^4} \right)_{x_i} - \dots$$

XIII-62

et

$$\left( \frac{d^2 u}{dx^2} \right)_{x_{i+1}} = \left( \frac{d^2 u}{dx^2} \right)_{x_i} + \Delta x \left( \frac{d^3 u}{dx^3} \right)_{x_i} + \frac{\Delta x^2}{2!} \left( \frac{d^4 u}{dx^4} \right)_{x_i} + \dots$$

XIII-63

XIII-62 et XIII-63 dans XIII-60 et XIII-61 donnent

$$\frac{u_{i+1} - u_{i-1}}{2\Delta x} = \left( \frac{du}{dx} \right)_{x_i} + \frac{\Delta x^2}{3!} \left( \frac{d^3 u}{dx^3} \right)_{x_i} + \frac{\Delta x}{48} \left[ 2\Delta x \left( \frac{d^3 u}{dx^3} \right)_{x_i} + \dots \right] + \dots = \left( \frac{du}{dx} \right)_{x_i} + \frac{5\Delta x^2}{24} \left( \frac{d^3 u}{dx^3} \right)_{x_i} + \dots$$

XIII-64

et

$$\frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2} = \left( \frac{d^2 u}{dx^2} \right)_{x_i} + \frac{\Delta x^2}{12} \left( \frac{d^4 u}{dx^4} \right)_{x_i} + \frac{1}{24} \left[ \Delta x^2 \left( \frac{d^2 u}{dx^2} \right)_{x_i} + \dots \right] + \dots = \left( \frac{d^2 u}{dx^2} \right)_{x_i} + \frac{\Delta x^2}{8} \left( \frac{d^4 u}{dx^4} \right)_{x_i} + \dots$$

XIII-65

XIII-64 et XIII-65 dans XIII-53 donnent le résultat cherché :

$$u(x) = u_i + (x - x_i) \left( \frac{u_{i+1} - u_{i-1}}{2\Delta x} - \frac{5\Delta x^2}{24} \left( \frac{d^3 u}{dx^3} \right)_{x_i} + \dots \right) + \left( \frac{(x - x_i)^2}{2!} - \frac{\Delta x^2}{24} \right) \left( \frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2} - \frac{\Delta x^2}{8} \left( \frac{d^4 u}{dx^4} \right)_{x_i} + \dots \right)$$

ou encore

$$u(x) = u_i + (x - x_i) \frac{u_{i+1} - u_{i-1}}{2\Delta x} + \left( \frac{(x - x_i)^2}{2!} - \frac{\Delta x^2}{24} \right) \frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2} - \left[ \frac{5\Delta x^2}{24} (x - x_i) \left( \frac{d^3 u}{dx^3} \right)_{x_i} + \left( \frac{(x - x_i)^2}{2!} - \frac{\Delta x^2}{24} \right) \frac{\Delta x^2}{8} \left( \frac{d^4 u}{dx^4} \right)_{x_i} + \dots \right]$$
XIII-66

XIII-66 nous donne les estimations cherchées de  $u(x)$  dans  $\left[ x_i - \frac{\Delta x}{2}, x_i + \frac{\Delta x}{2} \right]$  :

estimation d'ordre un :

$$u(x) = u_i \quad \text{XIII-67}$$

estimation d'ordre deux :

$$u(x) = u_i + (x - x_i) \frac{u_{i+1} - u_{i-1}}{2\Delta x} \quad \text{XIII-68}$$

estimation d'ordre trois :

$$u(x) = u_i + (x - x_i) \frac{u_{i+1} - u_{i-1}}{2\Delta x} + \left[ \frac{(x - x_i)^2}{2!} - \frac{\Delta x^2}{24} \right] \frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2} \quad \text{XIII-69}$$

On se contentera pour la suite d'estimations d'ordre un et deux, en utilisant toutefois XIII-69 de la manière suivante : toute expression

$$u(x) = u_i + (x - x_i) \frac{u_{i+1} - u_{i-1}}{2\Delta x} + \kappa \left[ \frac{(x - x_i)^2}{2!} - \frac{\Delta x^2}{24} \right] \frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2} \quad \text{XIII-70}$$

est une estimation du second ordre (sauf pour la valeur  $\kappa = 1$  qui fournit une estimation du troisième ordre) ; en faisant varier  $\kappa$ , on génère une famille d'estimations de  $u(x)$  du second ordre utilisables

partout dans  $\left[ x_i - \frac{\Delta x}{2}, x_i + \frac{\Delta x}{2} \right]$ . En particulier, les valeurs  $u_{i\pm 1/2}$  apparaissant dans XIII-47 sont déductibles de XIII-70 en y remplaçant  $x$  par  $x_i \pm \frac{\Delta x}{2}$  :

$$u_{i+1/2} = u_i + \frac{1}{4}(u_{i+1} - u_{i-1}) + \frac{\kappa}{12}(u_{i+1} - 2u_i + u_{i-1})$$

$$u_{i-1/2} = u_i - \frac{1}{4}(u_{i+1} - u_{i-1}) + \frac{\kappa}{12}(u_{i+1} - 2u_i + u_{i-1})$$

qu'on écrit plutôt en faisant apparaître les écarts  $u_i - u_{i-1}$  et  $u_{i+1} - u_i$  :

$$u_{i+1/2} = u_i + \left( \frac{1}{4} - \frac{\kappa}{12} \right)(u_i - u_{i-1}) + \left( \frac{1}{4} + \frac{\kappa}{12} \right)(u_{i+1} - u_i) \quad \text{XIII-71}$$

$$u_{i-1/2} = u_i - \left( \frac{1}{4} + \frac{\kappa}{12} \right)(u_i - u_{i-1}) - \left( \frac{1}{4} - \frac{\kappa}{12} \right)(u_{i+1} - u_i) \quad \text{XIII-72}$$

Trois commentaires importants accompagnent XIII-71 et 72 :

1° ces estimations sont valables pour autant qu'on reste à l'intérieur de l'intervalle  $\left[ x_i - \frac{\Delta x}{2}, x_i + \frac{\Delta x}{2} \right]$ ; on écrira donc  $u_{i+1/2,g}$  plutôt que  $u_{i+1/2}$  et  $u_{i-1/2,d}$  plutôt que  $u_{i-1/2}$ . Des estimations  $u_{i+1/2,d}$  et  $u_{i-1/2,g}$  peuvent également être écrites : elles proviendront de l'adaptation de XIII-71 et 72 aux intervalles  $\left[ x_i + \frac{\Delta x}{2}, x_i + \frac{3\Delta x}{2} \right]$  et  $\left[ x_i - \frac{3\Delta x}{2}, x_i - \frac{\Delta x}{2} \right]$ . On aura finalement en tout

$$u_{i-1/2,g} = u_{i-1} + \left( \frac{1}{4} - \frac{\kappa}{12} \right)(u_{i-1} - u_{i-2}) + \left( \frac{1}{4} + \frac{\kappa}{12} \right)(u_i - u_{i-1}) \quad \text{XIII-73}$$

$$u_{i-1/2,d} = u_i - \left( \frac{1}{4} + \frac{\kappa}{12} \right)(u_i - u_{i-1}) - \left( \frac{1}{4} - \frac{\kappa}{12} \right)(u_{i+1} - u_i) \quad \text{XIII-74}$$

$$u_{i+1/2,g} = u_i + \left( \frac{1}{4} - \frac{\kappa}{12} \right)(u_i - u_{i-1}) + \left( \frac{1}{4} + \frac{\kappa}{12} \right)(u_{i+1} - u_i) \quad \text{XIII-75}$$

$$u_{i+1/2,d} = u_{i+1} - \left( \frac{1}{4} + \frac{\kappa}{12} \right)(u_{i+1} - u_i) - \left( \frac{1}{4} - \frac{\kappa}{12} \right)(u_{i+2} - u_{i+1}) \quad \text{XIII-76}$$

2° au sein d'un intervalle  $\left[ x_i - \frac{\Delta x}{2}, x_i + \frac{\Delta x}{2} \right]$ , le caractère linéaire de la validité de XIII-70 implique que  $u$  varie linéairement.

3° en ajustant  $\kappa$ , on peut conférer aux schémas XIII-73 à 76 un caractère plus ou moins centré. Par exemple, pour  $\kappa = -3$ , les schémas sont entièrement upwind :

$$u_{i-1/2,g} = u_{i-1} + \frac{1}{2}(u_{i-1} - u_{i-2})$$

$$u_{i-1/2,d} = u_i - \frac{1}{2}(u_{i+1} - u_i)$$

$$u_{i+1/2,g} = u_i + \frac{1}{2}(u_i - u_{i-1})$$

$$u_{i+1/2,d} = u_{i+1} - \frac{1}{2}(u_{i+2} - u_{i+1})$$

Ce sont ces quatre dernières expressions qui seront utilisées dans XIII-47. La raison en est que le caractère upwind est celui qui est le plus favorable à l'absence d'oscillations parasites. Ce choix étant fait, il reste qu'une incertitude subsiste : faut-il prendre

$$u_{i+1/2} = u_{i+1/2,g} \quad \text{ou} \quad u_{i+1/2,d}$$

$$u_{i-1/2} = u_{i-1/2,g} \quad \text{ou} \quad u_{i-1/2,d}.$$

La réponse à cette question est relativement simple : un schéma donné est upwind relativement au sens de déplacement de la solution  $u(x)$  dans le temps. Ce sens peut être évalué de la manière suivante : si on se souvient que dans l'équation d'advection déjà rencontrée

$$\frac{\partial u}{\partial t} = -v \frac{\partial u}{\partial x}$$

Le paramètre  $v$  représente en grandeur et en sens la vitesse avec laquelle la solution se déplace le long de l'axe  $x$ , on en déduit que, transformant XIII-37 en

$$\frac{\partial u}{\partial t} = -\frac{\partial f(u)}{\partial x} = -\frac{df}{du} \frac{\partial u}{\partial x}, \quad \text{XIII-77}$$

Le signe de  $\frac{df}{du}$  nous donne le sens de déplacement : la solution se déplace de la gauche vers la droite

(respectivement de la droite vers la gauche) quand  $\frac{df}{du}$  est positif (respectivement négatif). Par conséquent, les choix suivants seront faits :

$$\frac{df}{du} > 0 : \quad u_{i+1/2} = u_{i+1/2,g} = u_i + \frac{1}{2}(u_i - u_{i-1}) \quad \text{XIII-78a}$$

$$u_{i-1/2} = u_{i-1/2,d} = u_{i-1} + \frac{1}{2}(u_{i-1} - u_{i-2}) \quad \text{XIII-78b}$$

$$\frac{df}{du} < 0 : \quad u_{i+1/2} = u_{i+1/2,d} = u_{i+1} - \frac{1}{2}(u_{i+2} - u_{i+1}) \quad \text{XIII-79a}$$

$$u_{i-1/2} = u_{i-1/2,g} = u_i - \frac{1}{2}(u_{i+1} - u_i) \quad \text{XIII-79b}$$

Traitons le cas  $\frac{df}{du} > 0$ , étant entendu que  $\frac{df}{du} < 0$  se traite de manière identique. Tout d'abord, il est intéressant de souligner que le schéma upwind à deux points appliqué avec succès – quant à la

préservation de la monotonicité - à la résolution de l'équation de Burgers, découle du choix XIII-78, limité au premier ordre :

$$u_{i+1/2} = u_{i+1/2,g} = u_i \quad \text{XIII-80a}$$

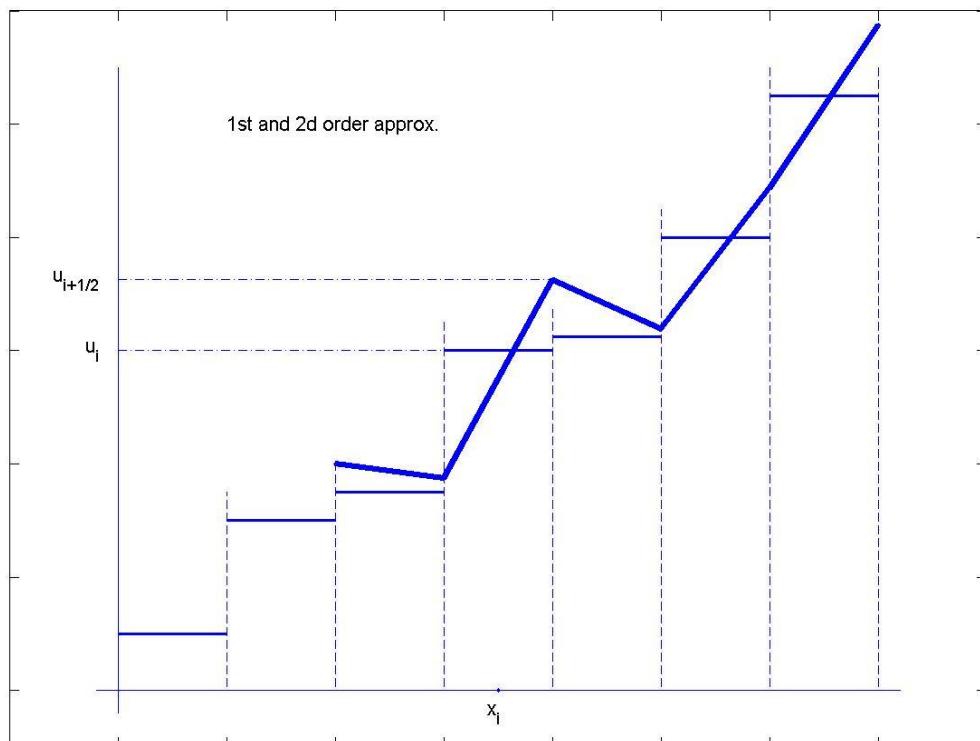
$$u_{i-1/2} = u_{i-1/2,g} = u_{i-1} \quad \text{XIII-80b}$$

XIII-80 appliqué à XIII-47 donne

$$\frac{du_i}{dt} = -\frac{1}{\Delta x} [f(u_i) - f(u_{i-1})]$$

qui est le schéma upwind à deux points.

Montrons maintenant que les schémas du second ordre XIII-78 peuvent introduire des oscillations parasites : la figure suivante montre comment la représentation  $\dots, u_{i-1}, u_i, u_{i+1}, \dots$  du premier ordre d'une fonction monotone (croissante dans la figure), représentation elle-même monotone, se transforme lors de la représentation du second ordre générée par XIII-78 :



Ces oscillations parasites sont rabotées par l'introduction de limiteurs de pente  $\phi$  qui transforment XIII-78 en

$$u_{i+1/2} = u_i + \frac{\phi_i}{2}(u_i - u_{i-1}) \quad \text{XIII-81a}$$

$$u_{i-1/2} = u_{i-1} + \frac{\phi_{i-1}}{2}(u_{i-1} - u_{i-2}) \quad \text{XIII-81b}$$

Ces fonctions  $\phi$  vont dépendre de la rapidité de variation du gradient de la solution, mesurée par le rapport

$$r_i = \frac{\frac{u_{i+1} - u_i}{x_{i+1} - x_i}}{\frac{u_i - u_{i-1}}{x_i - x_{i-1}}} = \frac{u_{i+1} - u_i}{u_i - u_{i-1}} \text{ car le maillage est uniforme.}$$

On aura donc

$$u_{i+1/2} = u_i + \frac{\phi(r_i)}{2}(u_i - u_{i-1}) \quad \text{XIII-82a}$$

$$u_{i-1/2} = u_{i-1} + \frac{\phi(r_{i-1})}{2}(u_{i-1} - u_{i-2}) \quad \text{XIII-82b}$$

La définition précise des fonctions  $\phi(r)$  résulte de développements mathématiques traduisant la volonté de rendre TVD le schéma final. Citons quelques conditions importantes à vérifier par tout limiteur de pente :

1°  $\phi(r) \geq 0$  si  $r \geq 0$  :  $r \geq 0$  traduit le caractère monotone de la solution. Ce caractère serait détruit avec  $\phi(r) < 0$ .

2°  $\phi(r) = 0$  si  $r < 0$  :  $r < 0$  est la traduction d'une oscillation (point anguleux ou de rebroussement avec variation du signe de la pente) qu'on refuse en restreignant alors XIII-75 au schéma du premier ordre.

3° il faut qu'un limiteur donné puisse traiter indifféremment les cas  $\frac{df}{du} > 0$  et  $\frac{df}{du} < 0$ .

On trouvera ci-dessous une liste non exhaustive de tels limiteurs ainsi que leur graphe en fonction de  $r$

koren :  $\phi(r) = \max\left(0, \min\left(2r, \frac{1+2r}{3}, 2\right)\right)$

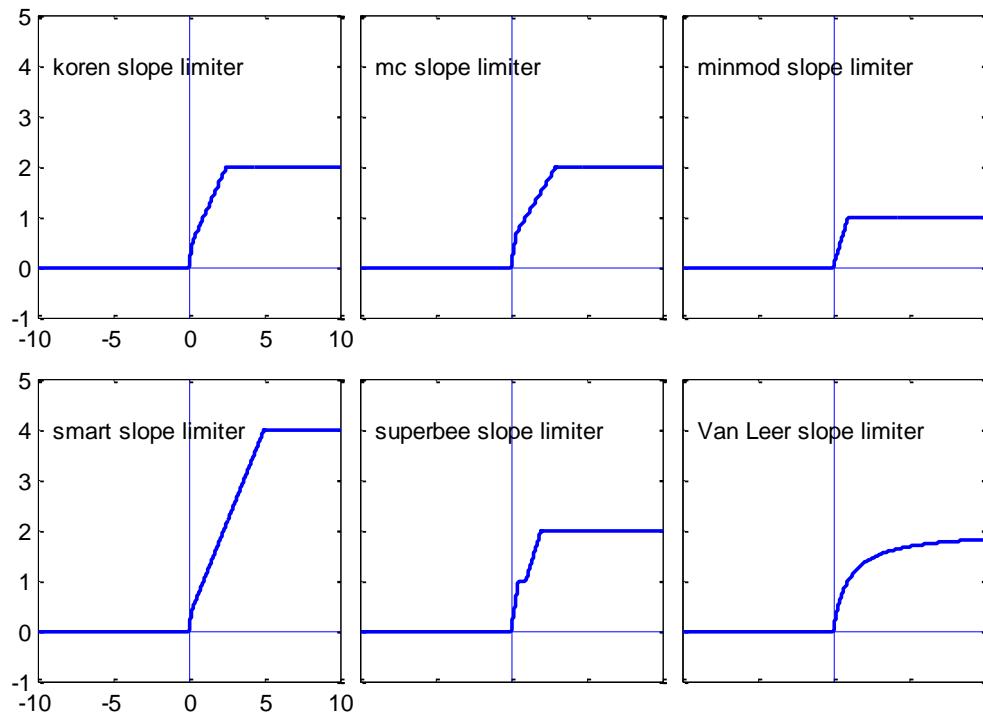
mc :  $\phi(r) = \max\left(0, \min\left(2r, \frac{1+r}{2}, 2\right)\right)$

minmod :  $\phi(r) = \max(0, \min(1, r))$

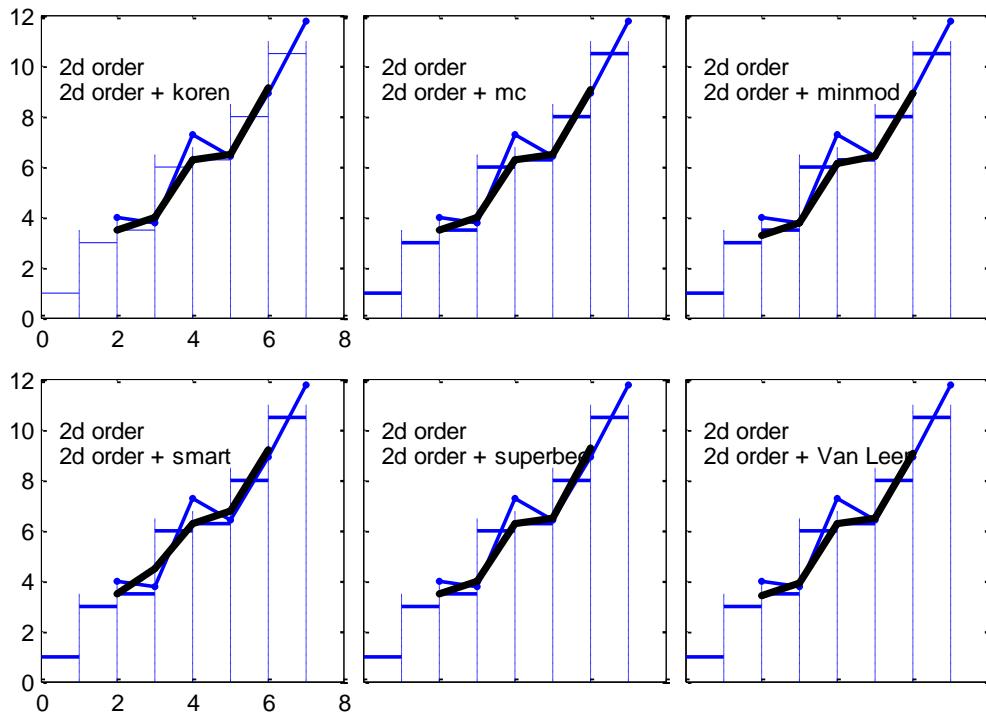
smart :  $\phi(r) = \max\left(0, \min\left(4, \frac{1+3r}{4}, 2r\right)\right)$

superbee :  $\phi(r) = \max(0, \min(2r, 1), \min(r, 2))$

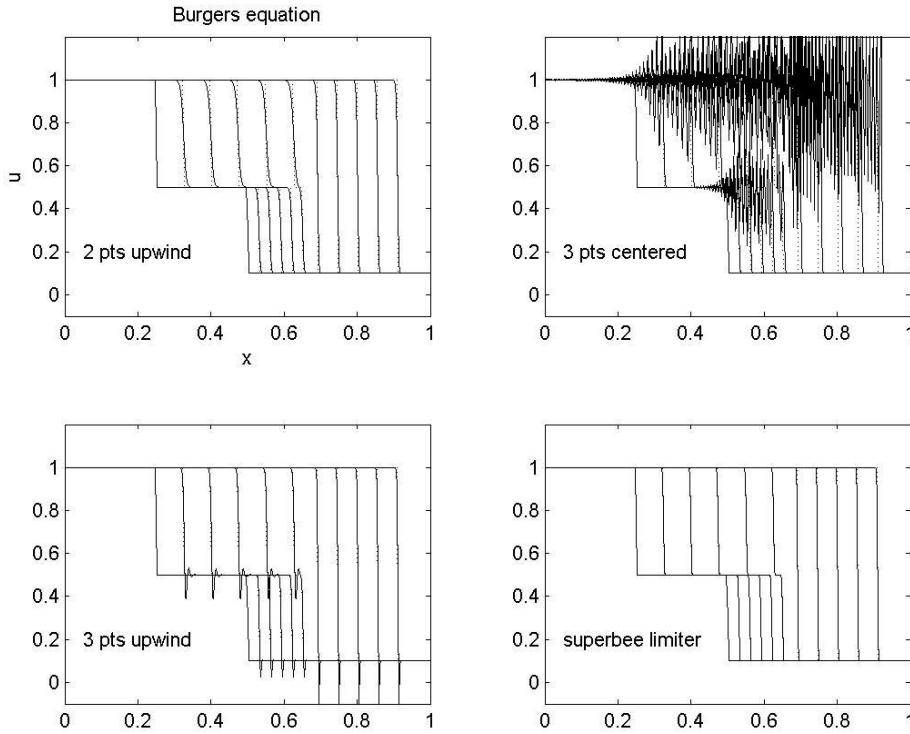
Van Leer :  $\phi(r) = \frac{r + |r|}{1 + |r|}$



La figure suivante montre l'effet de l'application de ces différents limiteurs sur les représentations du 2d ordre évoquées plus haut :



Enfin, l'apport réel des limiteurs de pente peut être mesuré en comparant les simulations de l'équation de Burger proposées plus haut à ce que l'utilisation d'un slope limiter (ici le superbee) permet d'obtenir :



Le caractère monotone de la solution est préservé tout en garantissant une meilleure simulation de la solution au voisinage des points les plus anguleux.

La bibliothèque Matmol propose une série de sous-programmes fonction permettant le calcul de  $\frac{\partial f(u)}{\partial x}$  par ces limiteurs de pente. On notera que la programmation y a été menée dans le cas du maillage spatial non uniforme, ce qui généralise l'exposé qui précède. Ces limiteurs sont utilisables dans le cas de problèmes scalaires où de systèmes d'équations aux dérivées partielles à condition qu'il n'y ait pas de couplage des variables dépendantes dans le terme de flux : on peut traiter

$$\begin{cases} u_{1t} = -f_{1x}(u_1) + \dots \\ \vdots \\ u_{Nt} = -f_{Nx}(u_N) + \dots \end{cases}$$

mais pas

$$\begin{cases} u_{1t} = -f_{1x}(u_1, \dots, u_N) + \dots \\ \vdots \\ u_{Nt} = -f_{Nx}(u_1, \dots, u_N) + \dots \end{cases}$$

Pour ce dernier cas, Matmol contient un sous-programme proposant un limiteur de pente vectoriel dû à Kurganov et Tadmor utilisant des schémas centrés en maillage non-uniforme. Ce limiteur, contrairement aux précédents qui sont de type upwind par rapport au sens de déplacement de la solution, est indépendant du spectre des valeurs propres de la matrice jacobienne de  $f(u)$ . Inspiré du schéma centré proposé en 1954

$$u_i^{k+1} = \frac{u_{i+1}^k + u_{i-1}^k}{2} - \frac{\Delta t}{2\Delta x} (f(u_{i+1}^k) - f(u_{i-1}^k))$$

XIII-83

par Lax et Friedrichs, il s'en différencie par d'importantes améliorations.

a) le schéma XIII-83 repose sur une représentation spatiale de  $u(x, t)$  constante par morceaux (à l'instar de XIII-67). Cette représentation sommaire a pour effet de conférer à XIII-83 une importante dissipation numérique excluant la résolution de problèmes dont les solutions présentent des fronts raides. Le schéma de Kurganov et Tadmor utilise une représentation quadratique par morceaux réduisant la dissipation numérique et permettant une représentation améliorée des fronts raides.

b) tel quel ce schéma a le défaut d'une viscosité numérique dont l'amplitude, inversement proportionnelle à  $\Delta t$ , est une entrave à l'implémentation du schéma dans le contexte de la méthode des lignes : ceci requiert en effet de faire tendre  $\Delta t$  vers zéro pour préserver la dérivation temporelle. Kurganov et Tadmor résolvent cette difficulté en faisant appel à l'estimation de la vitesse de

propagation locale de la solution aux frontières  $x_i \pm \frac{\Delta x}{2}$  :

$$a_{i \pm \frac{1}{2}}^k = \max \left\{ \rho \left[ \frac{\partial f}{\partial u} \left( u_{i \pm \frac{1}{2}}^- \right) \right], \rho \left[ \frac{\partial f}{\partial u} \left( u_{i \pm \frac{1}{2}}^+ \right) \right] \right\} \quad \text{XIII-84}$$

Dans XIII-84,  $\rho$  désigne le rayon spectral (c'est-à-dire la plus grande des valeurs propres en valeur absolue) de la matrice  $\frac{\partial f}{\partial u}$  évaluée (par exemple pour  $a_{i+1/2}^k$ ) en

$$u_{i+1/2}^+ = u_{i+1}^k - \frac{\Delta x}{2} (u_x)_{i+1}^k$$

$$u_{i+1/2}^- = u_i^k + \frac{\Delta x}{2} (u_x)_i^k$$

Tous calculs faits, on obtient

$$\frac{du_i}{dt} = -\frac{1}{2\Delta x} \left\{ \left[ f \left( u_{i+1/2}^+ \right) + f \left( u_{i+1/2}^- \right) \right] - \left[ f \left( u_{i-1/2}^+ \right) + f \left( u_{i-1/2}^- \right) \right] - a_{i+1/2} \left[ u_{i+1/2}^+ - u_{i+1/2}^- \right] + a_{i-1/2} \left[ u_{i-1/2}^+ - u_{i-1/2}^- \right] \right\}$$

avec

$$u_{i+1/2}^+ = u_{i+1} - \frac{\Delta x}{2} (u_x)_{i+1}$$

$$u_{i+1/2}^- = u_i + \frac{\Delta x}{2} (u_x)_i$$

$$u_{i-1/2}^+ = u_i - \frac{\Delta x}{2} (u_x)_i$$

$$u_{i-1/2}^- = u_{i-1} + \frac{\Delta x}{2} (u_x)_{i-1}$$

$$a_{i+1/2} = \max \left\{ \rho \left[ \frac{\partial f}{\partial u} (u_i) \right], \rho \left[ \frac{\partial f}{\partial u} (u_{i+1}) \right] \right\}$$

$$a_{i-1/2} = \max \left\{ \rho \left[ \frac{\partial f}{\partial u} (u_{i-1}) \right], \rho \left[ \frac{\partial f}{\partial u} (u_i) \right] \right\}$$

et

$$(u_x)_i = \min \mod \left( \theta \frac{u_{i+1} - u_i}{x_{i+1} - x_i}, \frac{u_{i+1} - u_{i-1}}{x_{i+1} - x_{i-1}}, \theta \frac{u_i - u_{i-1}}{x_i - x_{i-1}} \right) \quad 1 \leq \theta \leq 2$$

$\theta = 1$  garantit une solution non-oscillante, tandis que  $\theta = 2$  correspond au limiteur le moins dissipatif.

$$\min \text{ mod}(z_1, z_2, \dots) = \begin{cases} \min_j(z_j) & \text{si } z_j > 0 \ \forall j \\ \max_j(z_j) & \text{si } z_j < 0 \ \forall j \\ 0 & \text{dans les autres cas} \end{cases}$$

Deux implémentations de ce limiteur, d'ordre un et d'ordre deux sont proposées dans la bibliothèque Matmol : kurg\_centred\_slope\_limiter\_fz\_order1 et kurg\_centred\_slope\_limiter\_fz.

### XIII.5 Implémentation des conditions aux limites

Etape importante dans la résolution des équations aux dérivées partielles (EDP) par la méthode des lignes, l'implémentation des conditions aux limites est une opération délicate pouvant parfois conduire à l'échec de la simulation. Il n'existe pas à cet égard de recommandations universelles mais plutôt une série de procédés numériques plus ou moins approchés, dont il est utile d'avoir une connaissance détaillée et qu'on utilisera en ayant à l'esprit qu'un échec de la simulation peut trouver là son origine.

Dans la mesure où la très grande majorité des problèmes de l'ingénieur sont décrits par des EDP où les dérivées spatiales présentes sont d'ordre un et/ou deux, on limitera les cas étudiés aux 3 types de conditions aux limites suivantes :

Conditions de type Dirichlet : elles fixent la valeur de la variable dépendante sur les bords du domaine spatial (supposé monodimensionnel pour simplifier) :

$$\begin{aligned} u(t, x_{\min}) &= g_{\min}(t) \\ u(t, x_{\max}) &= g_{\max}(t) \end{aligned} \quad \text{XIII-85}$$

Conditions de type Neumann : elles fixent la valeur de la dérivée normale au bord du domaine spatial :

$$\begin{aligned} \frac{\partial u}{\partial n}(t, x_{\min}) &= h_{\min}(t) \\ \frac{\partial u}{\partial n}(t, x_{\max}) &= h_{\max}(t) \end{aligned} \quad \text{XIII-86}$$

Conditions de type Robin : il s'agit d'une combinaison linéaire des précédentes :

$$\begin{aligned} \frac{\partial u}{\partial n}(t, x_{\min}) + \alpha_{\min} u(t, x_{\min}) &= k_{\min}(t) \\ \frac{\partial u}{\partial n}(t, x_{\max}) + \alpha_{\max} u(t, x_{\max}) &= k_{\max}(t) \end{aligned} \quad \text{XIII-87}$$

On se propose dans la suite de l'exposé de passer en revue les techniques d'implémentation de ces conditions aux limites au travers de l'exemple suivant :

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \quad 0 < x < 1 \quad 0 < t \quad \text{XIII-88}$$

avec la CI :  $u(x, 0) = 1$  XIII-89

et des conditions aux limites diverses.

### XIII.5a Conditions aux limites de type Dirichlet

Ce sont les plus simples à implémenter : numérotons de 1 à N les nœuds du maillage spatial et les variables dépendantes et estimons la dérivée spatiale de XIII-88 par le schéma centré à trois points ; les conditions aux limites s'écrivent

$$\begin{aligned} u_1(t) &= g_1(t) \\ u_N(t) &= g_N(t) \end{aligned} \quad \text{XIII-90}$$

et le système différentiel généré par la méthode des lignes s'écrit

$$\begin{bmatrix} u_{2t} \\ u_{2t} \\ \vdots \\ u_{N-2t} \\ u_{N-1t} \end{bmatrix} = \frac{1}{\Delta x^2} \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ \ddots & \ddots & \ddots & \ddots & \\ & 1 & -2 & 1 & \\ & & 1 & -2 & \end{bmatrix} \begin{bmatrix} u_2 \\ u_2 \\ \vdots \\ u_{N-2} \\ u_{N-1} \end{bmatrix} + \frac{1}{\Delta x^2} \begin{bmatrix} g_1(t) \\ 0 \\ \vdots \\ 0 \\ g_N(t) \end{bmatrix} \quad \text{XIII-91}$$

XIII-90 supprime deux fonctions inconnues et le système différentiel résultant n'est plus que de dimension N-2. La programmation dans le contexte de Matmol est toutefois différente : comme on l'a vu, les programmes de calcul des dérivées spatiales utilisent des schémas sans apport de nœud extérieur : ainsi, le fichier de calcul `three_point_centered_D2` fournit la matrice suivante :

$$D_2 = \frac{1}{\Delta x^2} \begin{bmatrix} 1 & -2 & 1 & & \\ 1 & -2 & 1 & & \\ \ddots & \ddots & \ddots & \ddots & \\ & 1 & -2 & 1 & \\ & 1 & -2 & 1 & \end{bmatrix}$$

La programmation des deuxièmes membres de XIII-91 sera la suivante :

```
function ut = diffusion_pde(t,u)
%...
%... set global variables
%...
global n D2
%...
%... boundary conditions
%...
u(1) = g1(t);
u(n) = gn(t);
%...
%... second-order spatial derivative
%...
uxx = D2*u;
%...
%... temporal derivatives
%...
ut = uxx;
```

$$\begin{aligned} \text{ut}(1) &= 0; \\ \text{ut}(n) &= 0; \end{aligned}$$

Les conditions aux limites étant fixées, le vecteur  $\mathbf{uxx}$  va contenir

$$\begin{pmatrix} u_{xx}(1) \\ u_{xx}(2) \\ \vdots \\ u_{xx}(N-1) \\ u_{xx}(N) \end{pmatrix} = \frac{1}{\Delta x^2} \begin{pmatrix} 1 & -2 & 1 & & \\ 1 & -2 & 1 & & \\ \ddots & \ddots & \ddots & \ddots & \\ & 1 & -2 & 1 & \\ & 1 & -2 & 1 & \end{pmatrix} \begin{pmatrix} g_1(t) \\ u(2) \\ \vdots \\ u(N-1) \\ g_n(t) \end{pmatrix}$$

A ce stade,  $u_{xx}(2), \dots, u_{xx}(N-1)$  contiennent les estimations souhaitées tandis que  $u_{xx}(1)$  et  $u_{xx}(N)$  sont inutiles. Les instructions suivantes stockent alors  $\mathbf{uxx}$  dans  $\text{ut}$  et annulent  $\text{ut}(1)$  et  $\text{ut}(n)$  :  $\text{ut}(2), \dots, \text{ut}(n-1)$  contiennent les estimations correctes et seules les fonctions  $u(2)$  à  $u(n-1)$  font l'objet de l'intégration temporelle.

Cette méthode est particulièrement bien adaptée aux conditions aux limites de type Dirichlet ; elle est néanmoins également applicable aux autres conditions aux limites.

### XIII.5b Conditions aux limites de type Neumann et Robin

Remplaçons XIII-90 par

$$\begin{aligned} \frac{\partial u}{\partial x}(0, t) &= 0 \\ \frac{\partial u}{\partial x}(1, t) &= k_1 u(1, t) \end{aligned} \quad \text{XIII-92}$$

Il est possible d'utiliser une programmation identique à la précédente en remplaçant les dérivées spatiales de XIII-92 par des schémas de différences finies : par exemple

$$\frac{u_1 - u_0}{\Delta x} = 0 \quad \Leftrightarrow \quad u_0 = u_1 \quad \text{XIII-93}$$

$$\frac{u_N - u_{N-1}}{\Delta x} = k_1 u_N \quad \Leftrightarrow \quad u_N = u_{N-1} \frac{1}{1 - k_1 \Delta x} \quad \text{XIII-94}$$

La programmation précédente est remplacée par

```
function ut = diffusion_pde(t, u)
%...
%... set global variables
%...
global n D2 dx
%...
%... boundary conditions
%...
u(1) = u(2);
u(n) = u(n-1) / (1-k1*dx);
%...
%... second-order spatial derivative
```

```
%...
uxx = D2*u;
%...
%... temporal derivatives
%...
ut = uxx;
ut(1) = 0;
ut(n) = 0;
```

Remarquons qu'il y a moyen d'utiliser des schémas de différences finies plus sophistiqués : un schéma à trois points décentrés

$$\left( \frac{du}{dx} \right)_1 = \frac{-3u_1 + 4u_2 - u_3}{2\Delta x} \quad \text{XIII-95}$$

peut être utilisé pour donner :

$$\frac{-3u_1 + 4u_2 - u_3}{2\Delta x} = 0 \Leftrightarrow u_1 = \frac{1}{3}(4u_2 - u_3) \quad \text{XIII-96}$$

$$\frac{3u_N - 4u_{N-1} + u_{N-2}}{2\Delta x} = k_1 u_N \Leftrightarrow u_N = \frac{4u_{N-1} - u_{N-2}}{3 - 2k_1 \Delta x} \quad \text{XIII-97}$$

### **XIII.5c Rapatriement des conditions aux limites**

Les instructions

$$ut(1) = 0 \quad \text{XIII-98}$$

$$ut(n) = 0 \quad \text{XIII-99}$$

présentes dans les programmations précédentes sont responsables d'une erreur qu'il faut corriger dans le programme principal : dans ce dernier, les fonctions  $u_1(t)$  et  $u_N(t)$  restent égales à leurs valeurs initiales  $u_1(0)$  et  $u_N(0)$  puisque, alors qu'il faut imposer XIII-90 (ou XIII-93,94), XIII-98,99 annulent toute variation dans le temps de  $u_1$  et  $u_N$ . La correction à cette erreur consiste à programmer, après intégration numérique, les conditions aux limites dans le programme principal, par exemple en écrivant

```
% intégration temporelle
%
[timeout,yout]= ode15s(@diffusion_pde,time,u);
%
rapatriement des conditions aux limites
%
for i = 1:length(timeout)
    yout(i,1) = g1(timeout(i));
    yout(i,n) = gn(timeout(i));
end
```

### **XIII.5d Introduction d'équations algébriques**

Plus élégante que les procédés précédents, cette méthode consiste à interpréter les conditions aux limites comme des équations différentielles particulières : une condition de type Dirichlet

$$u_1(t) = g_1(t)$$

XIII-100

devient

$$0 \frac{du_1}{dt} = u_1(t) - g_1(t)$$

XIII-101

et une condition de type Robin

$$\frac{\partial u}{\partial x}(1, t) = k_1 u(1, t)$$

XIII-102

devient

$$0 \frac{du_N}{dt} = \frac{du_N}{dx} - k_1 u_N$$

XIII-103

XIII-101 et XIII-103 rendent différentiel-algébrique le système final qui s'écrit

$$\begin{pmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \\ & & & 0 \end{pmatrix} \begin{pmatrix} u_{1t} \\ u_{2t} \\ \vdots \\ u_{N-1t} \\ u_{Nt} \end{pmatrix} = \begin{cases} u_1 - g_1 \\ (u_1 - 2u_2 + u_3)/\Delta x^2 \\ \vdots \\ (u_{n-2} - 2u_{n-1} + u_n)/\Delta x^2 \\ (u_N - u_{N-1})/\Delta x - k_1 u_N \end{cases}$$

XIII-104

Plusieurs remarques peuvent être faites :

1. le vecteur des dérivées temporelles est multiplié par une matrice dite de masse qu'il faut programmer et qui requiert l'utilisation d'intégrateurs temporels particuliers (voir chapitre IV) : dans le cas présent, le choix s'est porté sur ode15s.

```
options = odeset('RelTol', 1e-5, 'AbsTol', 1e-5, 'stats', 'on', 'Mass', mass(n));
tout, yout] = ode15s(@diffusion_pde, t, x, options);
```

et la matrice de masse est calculée par exemple dans un sous-programme

```
function M = mass(n)
Mg = eye(n);
Mg(1,1) = 0;
Mg(n,n) = 0;
M = sparse(Mg);
```

2. le programme de calcul des membres de droite de XIII-88 est alors

```
function ut = diffusion_pde(t, u)
%...
%... set global variables
%...
global n D2 dx
```

```
%...
%... second-order spatial derivative
%...
uxx = D2*u;
%...
%... temporal derivatives
%...
ut = uxx;
ut(1) = u(1) - g1(t);
ut(n) = (u(n)-u(n-1))/dx - k1*u(n);
```

3. Si on souhaite une meilleure implémentation de la condition XIII-103, on peut par exemple songer à

```
function ut = diffusion_pde(t,u)
%...
%... set global variables
%...
global n D2 D1 dx
%...
%... spatial derivatives
%...
ux = D1*u;
uxx = D2*u;
%...
%... temporal derivatives
%...
ut = uxx;
ut(1) = u(1) - g1(t);
ut(n) = ux(1) - k1*u(n);
```

où  $D1$  provient du programme principal et est un schéma de différences finies plus précis que le schéma à deux points upwind proposé juste avant.

### XIII.5e Retour aux équations différentielles

Dans certains cas, la présence d'équations algébriques est une entrave au bon déroulement de l'intégration temporelle. La modification suivante permet d'éliminer cette difficulté : elle consiste à remplacer toute équation algébrique

$$0 \frac{du_k}{dt} = f_k(\bar{u}) \quad \text{XIII-105}$$

par l'équation différentielle

$$\frac{du_k}{dt} = Gf_k(\bar{u}) \quad \text{XIII-106}$$

avec  $G$  très grand (à la limite, faire tendre  $G$  vers l'infini restitue la forme algébrique). La valeur à conférer à  $G$  dépend de la rapidité d'évolution des composantes de  $\bar{u}$  :  $G$  doit être grand assez pour que la vitesse d'évolution de  $u_k$ , c'est-à-dire  $\frac{du_k}{dt}$ , soit d'un ordre de grandeur supérieur à celle des autres composantes de  $\bar{u}$ .

Outre sa valeur absolue, il faut aussi choisir correctement le signe de  $G$  : ceci est facilement compréhensible : supposons pour fixer les idées devoir implémenter une condition de type Dirichlet :

$$u_1(t) = U_0 \left( 1 - e^{-\frac{t}{T}} \right) \quad U_0 > 0$$

XIII-107

avec  $u_1(0) = 0$

XIII-108

La technique proposée ci-dessus suggère d'implémenter l'instruction

$$\frac{du_1}{dt} = G \left( u_1 - U_0 \left( 1 - e^{-\frac{t}{T}} \right) \right)$$

XIII-109

Il est clair que le signe de  $G$  n'est pas indifférent : en  $t = 0^+$ ,  $u_1 - U_0 \left( 1 - e^{-\frac{t}{T}} \right)$  est négatif et si on veut que  $u_1(t)$  s'identifie à XIII-107, il s'agit donc que  $\frac{du_1}{dt}$  soit positif : il faut donc imposer  $G < 0$ .

L'implémentation de ce procédé conduit au code suivant (on suppose ne pas avoir modifié la programmation de la deuxième condition aux limites) :

```
function ut = diffusion_pde(t,u)
%...
%... set global variables
%...
global n D2 dx G U0 T
%...
%... second-order spatial derivative
%...
uxx = D2*u;
%...
%... temporal derivatives
%...
ut = uxx;
ut(1) = G*(u(1) - U0*(1-exp(-t/T)));
ut(n) = (u(n)-u(n-1))/dx - k1*u(n);
```

et il faut aussi modifier la programmation de la matrice de masse qui devient

```
function M = mass(n)
Mg = eye(n);
Mg(n,n) = 0;
M = sparse(Mg);
```

### XIII.5f Procédé « Stagewise »

Un procédé parfois intéressant lorsque des conditions aux limites de type Neumann ou Robin sont présentes permet d'intégrer celles-ci plus étroitement dans la programmation : il consiste à estimer le calcul de la dérivée seconde par la mise en cascade d'un opérateur de dérivée première :

Supposons que les conditions aux limites souhaitées soient XIII-92 ; leur programmation par le procédé stagewise est décrit par le listing suivant :

```
function ut = diffusion_pde(t,u)
%...
```

```

%... set global variables
%...
    global n D1 dx
%...
%... spatial derivatives
%...
ux = D1*u;
ux(1) = 0 ;
ux(n) = k1*u(n) ;
uxx = D1*ux;
%...
%... temporal derivatives
%...
ut = uxx;

```

Le point délicat de ce procédé est le remplacement d'un vrai opérateur de calcul de dérivée seconde par la mise en cascade d'un opérateur de calcul de dérivée première : on veillera à utiliser un schéma D1 centré de sorte que la mise en cascade `uxx = D1*ux` le soit aussi.

## Chapitre XIV. Stabilité et intégration temporelle

### XIV.1 Introduction

Etablies lors de l'étude de la résolution des équations aux dérivées partielles par la méthode des différences finies, les conditions de stabilité matricielle de la résolution d'un système d'équations différentielles ordinaires (cf. première partie, chapitre X, relation X-48)

$$\frac{d\bar{u}}{dt} = S\bar{u} \quad \text{XIV-1}$$

sont les suivantes :

- a) stabilité spatiale : les valeurs propres  $\Omega_j$  de la matrice de discréétisation spatiale  $S$  doivent être à partie réelle négative ou nulle ;  $S$  dépend des schémas de discréétisation choisis pour évaluer les dérivées spatiales et suppose que les conditions aux limites sont périodiques.
- b) stabilité temporelle : le produit  $z = \Omega_j \Delta t$  où  $\Delta t$  est le pas d'intégration temporelle doit se situer à l'intérieur du domaine de stabilité de l'intégrateur temporel. Celui-ci résulte de la théorie de la stabilité absolue (A-stabilité) qui fournit une condition nécessaire et suffisante de stabilité lorsque le système des équations différentielles à résoudre est linéaire, c'est-à-dire lorsqu'il se met sous la forme suivante :

$$\bar{y}' = A\bar{y} + \bar{b} \quad \text{où } A \text{ est une matrice de constante, ce qui est le cas de XIV-1}$$

La détermination de ce domaine dépend de la nature de l'intégrateur utilisé : on a vu par exemple que pour les méthodes RK explicites, ce domaine vérifie

méthode à un étage :	$\left\{ z :  1+z  \leq 1 \right\}$
méthodes à deux étages :	$\left\{ z : \left  1+z + \frac{z^2}{2!} \right  \leq 1 \right\}$
méthodes à trois étages :	$\left\{ z : \left  1+z + \frac{z^2}{2!} + \frac{z^3}{3!} \right  \leq 1 \right\}$
méthodes à quatre étages :	$\left\{ z : \left  1+z + \frac{z^2}{2!} + \frac{z^3}{3!} + \frac{z^4}{4!} \right  \leq 1 \right\}$

Pour les intégrateurs à pas multiples obéissant à

$$\alpha_k y_{n+k} + \dots + \alpha_0 y_n = \Delta t (\beta_k f_{n+k} + \dots + \beta_0 f_n), \quad \text{XIV-2}$$

la condition de stabilité temporelle est la suivante : il faut que les racines  $\xi_i$  du polynôme caractéristique  $\rho_{\Omega \Delta t}(\xi)$

$$\rho_{\Omega \Delta t}(\xi) = (\alpha_k - \beta_k \Omega \Delta t) \xi^k + \dots + (\alpha_0 - \beta_0 \Omega \Delta t) \quad \text{XIV-3}$$

soient en module inférieures ou égales à un, et que les racines de module unitaire soient simples. Dans le plan complexe de la variable  $z = \Omega\Delta t$ , le domaine de stabilité est donc défini par l'ensemble des  $z$  tel que les racines simples  $\xi_{si}$  et multiples  $\xi_{mi}$  de XIV-3 vérifient

$$|\xi_{si}| \leq 1 \text{ et } |\xi_{mi}| < 1 \quad \forall i \quad \text{XIV-4}$$

## XIV.2 Exploitation de la condition de stabilité spatiale

Plusieurs informations supplémentaires à ce qui a été exposé dans la théorie de la stabilité matricielle peuvent être apportées.

### *Sélection de schémas préférentiels*

On y a vu (première partie, paragraphe X-3) que le choix des schémas de différences finies utilisables pour une équation donnée dépendait de l'intégrateur temporel utilisé. Il est important de souligner que certains schémas sont inutilisables quel que soit l'intégrateur temporel utilisé : ceux qui conduisent à une matrice de discrétisation spatiale dont les valeurs propres sont à partie réelle positive. C'est le cas par exemple du schéma de calcul de la dérivée première à deux points downwind

$$u_{xi} = \frac{u_{i+1} - u_i}{\Delta x} \quad \text{XIV-5}$$

appliqué à la résolution de l'équation d'advection

$$u_t = -au_x \quad a > 0 \quad \text{XIV-6}$$

Il est alors pertinent de se poser la question suivante : le calcul d'une dérivée première par XIV-5 conduit-il toujours à l'instabilité spatiale et doit-il donc être définitivement rejeté ? La réponse à cette question est non : cela dépend de l'équation à résoudre. Illustrons par un exemple.

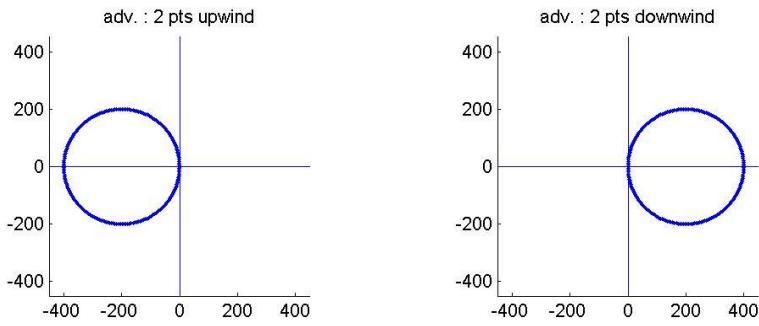
La résolution de l'équation d'advection en estimant la dérivée spatiale par un schéma à deux points upwind et par un schéma à deux points downwind engendrent des matrices de discrétisation spatiale dont les valeurs propres valent

$$\text{schéma upwind : } \Omega_k = \frac{a}{\Delta x} (\cos \alpha_k - 1) - j \frac{a}{\Delta x} \sin \alpha_k \quad \text{XIV-7}$$

$$\text{schéma downwind : } \Omega_k = \frac{a}{\Delta x} (-\cos \alpha_k + 1) - j \frac{a}{\Delta x} \sin \alpha_k \quad \text{XIV-8}$$

$$\text{avec } \alpha_k = (k-1) \frac{2\pi}{N} \quad k = 1, \dots, N \quad \text{XIV-9}$$

XIV-7 est toujours à partie réelle négative et XIV-8 toujours à partie réelle positive. La figure suivante représente le lieu des ces valeurs propres pour  $a = 1$  et  $\Delta x = 0.005$ .



Par contre, la résolution de l'équation de diffusion-convection

$$u_t = -au_x + bu_{xx} \quad a \text{ et } b > 0 \quad \text{XIV-10}$$

avec une estimation de la dérivée seconde par un schéma centré à trois points conduit aux valeurs propres suivantes si la dérivée première est à nouveau évaluée par les schémas upwind et downwind à deux points

$$\text{schéma upwind : } \Omega_k = \left( \frac{a}{\Delta x} + \frac{2b}{\Delta x^2} \right) (\cos \alpha_k - 1) - j \frac{a}{\Delta x} \sin \alpha_k \quad \text{XIV-11}$$

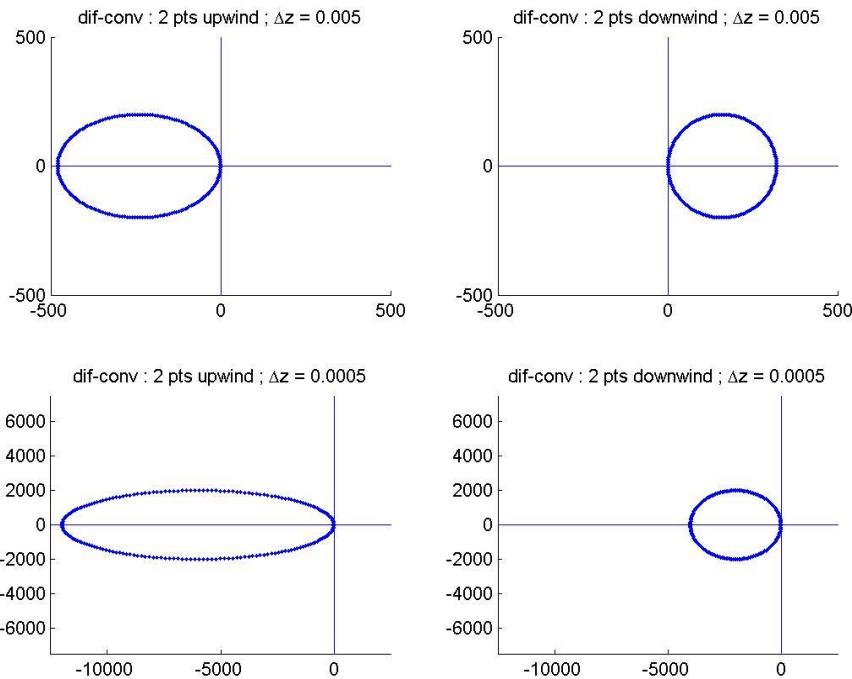
$$\text{schéma downwind : } \Omega_k = \left( \frac{a}{\Delta x} - \frac{2b}{\Delta x^2} \right) (-\cos \alpha_k + 1) - j \frac{a}{\Delta x} \sin \alpha_k \quad \text{XIV-12}$$

$$\text{avec } \alpha_k = (k-1) \frac{2\pi}{N} \quad k = 1, \dots, N \quad \text{XIV-13}$$

Les valeurs propres correspondant au schéma upwind restent à partie réelle négative, mais pour le schéma downwind, la position du lieu des valeurs propres dépend cette fois de la valeur de  $\Delta x$  : il sera à gauche de l'axe imaginaire du plan complexe quand

$$\Delta x < \frac{2b}{a} \quad \text{XIV-14}$$

et à droite dans le cas contraire. La figure ci-dessous illustre cette situation pour les deux schémas, pour  $a=1$  et  $b=0.0005$  :



On peut donc affirmer qu'il est impossible d'exclure a priori tout schéma de calcul d'une dérivée par différences finies.

### XIV.3 Problèmes stiff

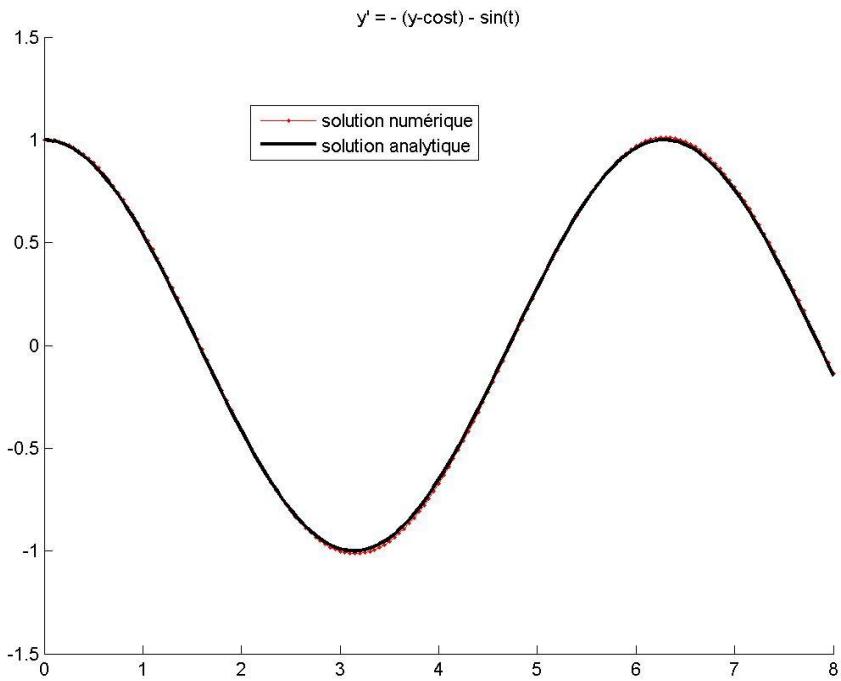
Concept lié aux équations différentielles ordinaires, la stiffness est difficile à définir mathématiquement. Historiquement, la première définition (1952) est aussi la plus parlante : les problèmes stiff sont décrits par des équations différentielles ordinaires pour lesquels la résolution par certaines méthodes implicites est remarquablement meilleure que celle fournie par les méthodes explicites. Une deuxième définition également parlante est la suivante : une équation différentielle ordinaire sera dite stiff lorsque, intégrée par une méthode explicite, la valeur maximale admissible du pas d'intégration est conditionnée par la stabilité de la solution et non par le maintien de son erreur de troncature sous un seuil donné. L'exemple suivant illustre cette deuxième définition :

$$y' = -k(y - \cos t) - \sin t \quad \text{XIV-15}$$

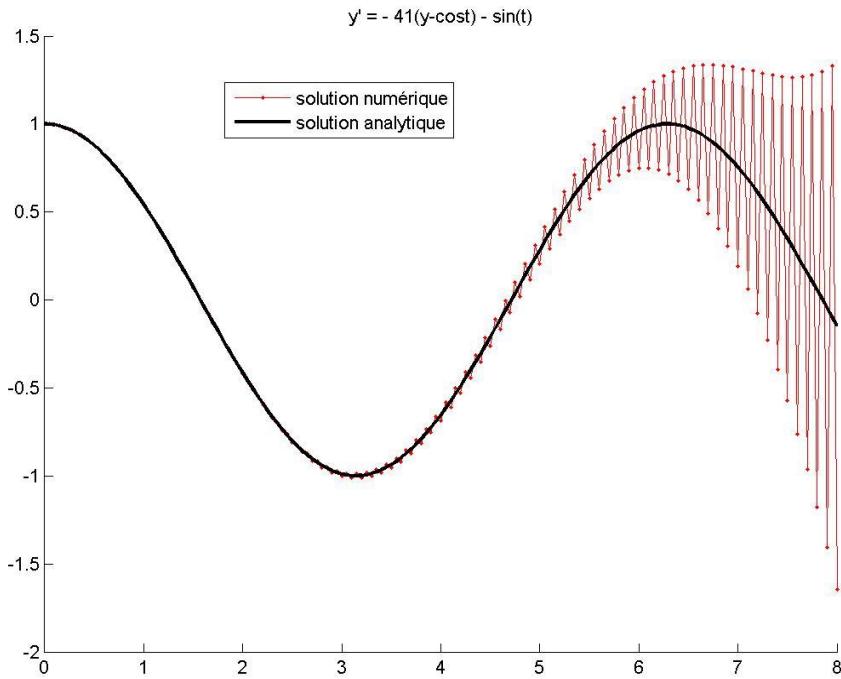
avec  $t \geq 0$ ,  $y(0) = 1$  et  $k$  réel

la solution analytique, indépendante de  $k$  vaut  $y(t) = \cos t$ . XIV-16

La figure suivante superpose la solution analytique et la solution numérique obtenue par la méthode d'Euler avec le pas  $h = 0.05$  quand  $k = 1$  :



Ce pas est suffisamment petit pour assurer une solution numérique de bonne qualité, c'est-à-dire avec une erreur de troncature petite. Voici maintenant le résultat obtenu en gardant le même pas mais en imposant dans l'équation  $k = 41$  :



Ce n'est qu'au prix d'une réduction du pas  $h$  qu'il y a moyen de retrouver une solution numérique correcte. Et si on impose  $k = 1000$ , il est nécessaire de diminuer le pas à la valeur 0.002, alors que la solution analytique est inchangée. On est alors clairement dans le cas d'un problème où le pas maximum admissible est conditionné par la stabilité et non plus par la volonté de réduire l'erreur de troncature.

Une troisième définition de la stiffness peut être déduite de cet exemple : rappelons d'abord qu'une équation différentielle  $\bar{y}' = \bar{f}(t, \bar{y})$  est stable si les valeurs propres de son jacobien  $\frac{\partial \bar{f}}{\partial \bar{y}}$  sont à partie réelle négative, et est d'autant plus stable que ces parties réelles sont grandes en valeur absolue. Pour l'exemple traité, le jacobien s'identifie ici à la dérivée première  $\frac{df}{dy} = -k$  : or on vient de voir que le problème est d'autant plus stiff que  $k$  est grand. Cela signifie que les problèmes stiff sont des systèmes hyper-stables.

Revenons à la première définition : maintenant que l'on sait que les problèmes stiff sont décrits par des équations différentielles ordinaires dont les valeurs propres du jacobien, ont leurs parties réelles « très » négatives, il est normal de comprendre, puisque aucune méthode explicite n'est A-stable, que seules certaines méthodes implicites permettent de résoudre efficacement ces équations.

En 1976, Hall et Watt proposent une autre définition calquée sur le raisonnement mathématique qui conduit à la notion de système stable : soit le système d'équations différentielles ordinaires

$$\bar{y}' = \bar{f}(t, \bar{y}) \quad \text{et} \quad \bar{y}(t_0) = \bar{y}_0 \quad \text{XIV-17}$$

Désignons par  $\bar{\varphi}(t)$  une solution smooth de ce système et par  $\bar{y}_{ST}(t)$  une solution stiff qui soit dans le voisinage de  $\bar{\varphi}(t)$  :  $\bar{\varphi}(t)$  et  $\bar{y}_{ST}(t)$  étant solutions de XIV-17, on a

$$\bar{\varphi}' = \bar{f}(t, \bar{\varphi}) \quad \text{et} \quad \bar{y}_{ST}' = \bar{f}(t, \bar{y}_{ST}) \quad \text{XIV-18}$$

La partie stiff de  $\bar{y}_{ST}(t)$  est égale à  $\Delta\bar{y} = \bar{y}_{ST} - \bar{\varphi}$  et elle est solution de, en limitant le développement de Taylor au premier terme,

$$\Delta\bar{y}' = \bar{f}(t, \bar{y}_{ST}) - \bar{f}(t, \bar{\varphi}) \approx [J](\bar{y}_{ST} - \bar{\varphi}) = [J]\Delta\bar{y} \quad \text{XIV-19}$$

où  $[J]$  est la matrice jacobienne  $\left[ \frac{\partial f_i}{\partial y_j} \right]_{(t, \bar{\varphi})}$ . Pour autant que  $[J]$  soit diagonalisable, on peut admettre

que les solutions de XIV-19 sont une combinaison linéaire d'exponentielles  $\exp(\lambda_k t)$  où  $\lambda_k$  sont les valeurs propres de  $[J]$ . Il en résulte que toute solution stiff dans le voisinage de  $\bar{\varphi}(t)$  se met sous la forme

$$\bar{y}_{ST}(t) = \bar{\varphi}(t) + \sum_k c_k \exp(\lambda_k t) \bar{\xi}_k \quad \text{XIV-20}$$

où les  $c_k$  sont des constantes et les  $\bar{\xi}_k$  sont les vecteurs propres de  $[J]$ . Le problème étant stable, on a

$$\Re(\lambda_k) < 0 \quad \forall k \quad \text{XIV-21}$$

Cela signifie que la partie stiff  $\sum_k c_k \exp(\lambda_k t) \bar{\xi}_k$  va peu à peu disparaître après une durée dépendant des constantes de temps  $\tau_k = \frac{1}{\Re(-\lambda_k)}$

Le système XIV-17 est alors dit stiff si

$$S = \frac{\max_k \Re(-\lambda_k)}{\min_k \Re(-\lambda_k)} \gg 0$$

XIV-23

$S$  est appelé « ratio de stiffness » ; une valeur de l'ordre de 10 à 100 correspond à un problème modérément stiff, tandis qu'une valeur de l'ordre de  $10^6$  correspond à une stiffness sévère. Comme les nombres  $\Re(-\lambda_k)$  représentent les inverses des constantes de temps qui conditionnent la disparition de la partie stiff de XIV-20, on comprend qu'il est courant dans la littérature de parler de problèmes stiff comme étant caractérisés par des constantes de temps d'ordres de grandeur très différents. Ceci permet de comprendre pourquoi la résolution de ces problèmes par les méthodes explicites requiert des temps de calculs extrêmement longs : voyons par exemple la résolution par la méthode d'Euler d'un problème dont le rapport de stiffness vaut  $10^6$ . On a donc

$$\frac{\max_k \Re(-\lambda_k)}{\min_k \Re(-\lambda_k)} = \frac{\tau_{\max}}{\tau_{\min}} = 10^6$$

La plus grande des constantes de temps fixe le temps de simulation nécessaire à l'obtention de la solution complète :

$$T_{\text{simul}} \approx 5\tau_{\max}$$

XIV-24

La plus grande des valeurs propres fixe le pas d'intégration maximum admissible, via la théorie de la stabilité absolue : si on admet pour simplifier que toutes les valeurs propres  $\lambda_k$  sont réelles, le rayon de stabilité de la méthode d'Euler impose

$$|\lambda_{\max}| \Delta t_{\max} = 2$$

c'est-à-dire

$$\Delta t_{\max} = \frac{2}{|\lambda_{\max}|} = 2\tau_{\min}$$

XIV-25

Le nombre total de pas d'intégration vaut alors

$$N = \frac{T_{\text{simul}}}{\Delta t_{\max}} = \frac{5\tau_{\max}}{2\tau_{\min}} = \frac{5}{2} 10^6$$

XIV-26

ce qui peut être prohibitif. Ceci permet de comprendre l'intérêt des intégrateurs implicites pour qui la condition de stabilité temporelle peut ne pas exister. Comme on l'a vu, la notion essentielle à cet égard est celle du domaine de stabilité.

Les développements qui suivent proposent l'étude approfondie de la notion de stabilité et quelques méthodes nouvelles répondant à des impératifs d'efficacité en termes de résolution des problèmes stiff. On distinguera les intégrateurs à un pas des intégrateurs à pas multiples.

## XIV.4 Les intégrateurs à un pas : stabilité et nouvelles méthodes

### A-stabilité

Rappelons qu'une méthode est dite A-stable si elle ne souffre d'aucune restriction de stabilité lorsqu'on l'applique à la résolution de

$$y' = \lambda y \quad \text{XIV-27}$$

avec  $\Re(\lambda) < 0$  et  $\Delta t$  quelconque.

Dahlquist précise mathématiquement cette propriété en déclarant qu'une méthode est A-stable si son domaine de stabilité comprend

$$C^- = \{z : \Re(z) \leq 0\} \quad \text{XIV-28}$$

Rappelons également à titre d'exemple comment établir la forme du domaine de stabilité de la méthode d'Euler implicite. Cet intégrateur est décrit par le schéma de calcul

$$y_{i+1} = y_i + hf(t_{i+1}, y_{i+1}) \quad \text{XIV-29}$$

Appliqué à XIV-28, cela donne

$$y_{i+1} = y_i + h\lambda y_{i+1} \quad \text{XIV-30}$$

ou encore

$$y_{i+1} = \frac{y_i}{1 - h\lambda} \quad \text{XIV-31}$$

La condition de stabilité de cet intégrateur est donc, en posant  $z = h\lambda$

$$\left| \frac{1}{1 - z} \right| \leq 1 \quad \text{XIV-32}$$

Il s'agit de l'extérieur du cercle de centre  $(1,0)$  et de rayon unitaire. Ce domaine inclut le demi-plan  $\Re(z) \leq 0$  : la méthode est donc A-stable. Notons que la fonction

$$F(z) = \frac{1}{1 - z} \quad \text{XIV-33}$$

est appelée fonction de stabilité.

Le calcul de la fonction de stabilité de toute méthode RK implicite à  $s$  étages peut être établi ; si on convient pour toute méthode décrite par

$$\begin{aligned} k_1 &= f(t_0 + c_1 h, y_0 + ha_{11}k_1 + \dots + ha_{1s}k_s) \\ &\vdots \\ k_s &= f(t_0 + c_s h, y_0 + ha_{s1}k_1 + \dots + ha_{ss}k_s) \\ y_1 &= y_0 + h(b_1k_1 + \dots + b_sk_s) \end{aligned} \quad \text{XIV-34}$$

de poser

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1s} \\ \vdots & & \vdots \\ a_{s1} & \cdots & a_{ss} \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_s \end{pmatrix}, \quad [1] = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \text{XIV-35}$$

on peut montrer que la fonction de stabilité de XIV-34 vaut

$$F(z) = \frac{\det(I - zA + z[1]b^T)}{\det(I - zA)} \quad \text{XIV-36}$$

qui est une fraction rationnelle en  $z$ .

Exemples :

1° la méthode dite des trapèzes est décrite par le tableau

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 0.5 & 0.5 \\ \hline & 0.5 & 0.5 \end{array}$$

On a donc pour cette méthode

$$F(z) = \frac{\det\left(\begin{pmatrix} 1 & 0 \\ -0.5z & 1-0.5z \end{pmatrix} + z\begin{pmatrix} 1 \\ 1 \end{pmatrix}(0.5 \quad 0.5)\right)}{\det\left(\begin{pmatrix} 1 & 0 \\ -0.5z & 1-0.5z \end{pmatrix}\right)} = \frac{1 + \frac{z}{2}}{1 - \frac{z}{2}} \quad \text{XIV-37}$$

La frontière du domaine de stabilité est donc

$$\left\{ z : \frac{|1+0.5z|}{|1-0.5z|} = 1 \right\} \quad \text{XIV-38}$$

Il s'agit de l'axe imaginaire. Il est immédiat de vérifier que la zone de stabilité est  $\Re(z) \leq 0$  : la méthode est A-stable.

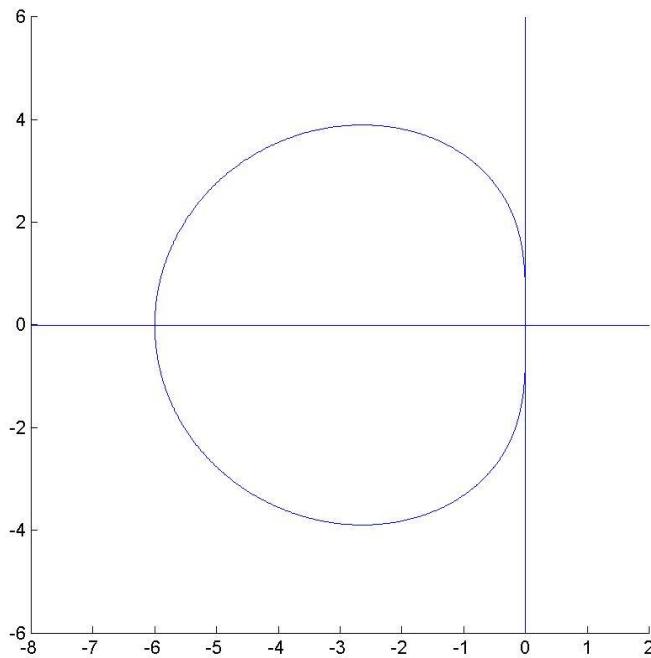
2° méthode de Hammer-Hollingsworth : elle est décrite par le tableau

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 2/3 & 1/3 & 1/3 \\ \hline & 1/4 & 3/4 \end{array}$$

On trouve alors

$$F(z) = \frac{1 + \frac{2z}{3} + \frac{z^2}{6}}{1 - \frac{z}{3}} \quad \text{XIV-39}$$

Le domaine de stabilité correspondant est l'intérieur du contour suivant :



La méthode n'est donc pas A-stable : toutes les méthodes implicites ne sont donc pas A-stable.

Il n'est pas indispensable de représenter le domaine de stabilité d'une méthode RK implicite pour savoir si elle est A-stable : on peut montrer (grâce au principe du maximum) qu'une méthode est A-stable si sa fonction de stabilité  $F(z)$  vérifie

$$|F(iy)| \leq 1 \quad \forall y \in \mathbb{R} \quad \text{XIV-40a}$$

$$\text{et } F(z) \text{ est analytique dans } \Re(z) < 0 \quad \text{XIV-40b}$$

XIV-40a implique la stabilité le long de l'axe imaginaire et est appelé I-stabilité. Tenons compte de ce que  $F(z)$  est une fraction rationnelle en  $z$  (voir plus haut) :

$$F(z) = \frac{P(z)}{Q(z)} \quad \text{XIV-41}$$

XIV-40a donne

$$\frac{|P(iy)|}{|Q(iy)|} \leq 1 \quad \Rightarrow \quad \frac{|P(iy)|^2}{|Q(iy)|^2} \leq 1$$

$$\Leftrightarrow |Q(iy)|^2 - |P(iy)|^2 = Q(iy)Q(-iy) - P(iy)P(-iy) \geq 0 \quad \text{XIV-42}$$

On pose alors  $E(y) = Q(iy)Q(-iy) - P(iy)P(-iy)$  et la méthode est I-stable si  $E(y) \geq 0 \quad \forall y \in \mathbb{R}$ .

Exemple : soit la méthode SDIRK

$$\begin{array}{c|cc} \gamma & \gamma & 0 \\ \hline 1-\gamma & 1-2\gamma & \gamma \\ \hline & 0.5 & 0.5 \end{array} \quad \text{avec } \gamma > 0$$

La fonction de stabilité de cette méthode vaut

$$F(z) = \frac{1 + (1 - 2\gamma)z + (0.5 - 2\gamma + \gamma^2)z^2}{(1 - \gamma z)^2}$$

XIV-40b :  $F(z)$  est analytique dans  $\Re(z) < 0$

XIV-40a : après calculs on trouve

$$E(y) = y^4 \left( \gamma - \frac{1}{4} \right) (2\gamma - 1)^2$$

$E(y)$  est toujours positif pour  $\gamma \geq \frac{1}{4}$  : la méthode est donc A-stable pour  $\gamma \geq \frac{1}{4}$ .

### L-stabilité

En 1977, R. Alexander affirmait que « la A-stabilité n'est pas la réponse entière au problème de la stiffness » : du point de vue de la A-stabilité, la méthode d'Euler implicite et la méthode des trapèzes sont strictement équivalentes : toutes les deux contiennent le demi-plan  $\Re(z) < 0$  dans leur domaine de stabilité. Elles présentent toutefois des comportements très différents face au problème de la stiffness, comme le montre l'exemple suivant : soit à résoudre

$$y' = -k(y - t) \quad \text{XIV-43}$$

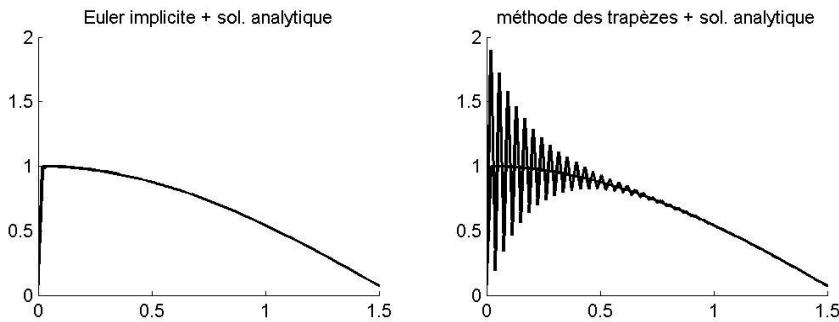
avec  $0 \leq t \leq 1.5$  et  $y(0) = 0$ .

La solution analytique de XIV-43 vaut

$$y_{\text{anal}}(t) = \left( y(0) - \frac{k^2}{1+k^2} \right) \exp(-kt) + \frac{k^2}{1+k^2} (k \cos t + \sin t) \quad \text{XIV-44}$$

Les composantes stiff de la solution se réduisent ici au seul terme exponentiel de XIV-44, dont la valeur propre correspondante  $\lambda = -k$  est donc réelle négative.

La figure suivante représente les solutions obtenues avec ces deux méthodes avec le même pas d'intégration  $h=0.01875$  pour  $k = 2000$  :



Cette différence inattendue des solutions s'explique de la manière suivante : les fonctions de stabilité de ces deux méthodes valent :

$$\text{Euler implicite : } F(z) = \frac{1}{1-z} \quad \text{XIV-45}$$

$$\text{Méthode des trapèzes : } F(z) = \frac{1+0.5z}{1-0.5z} \quad \text{XIV-46}$$

Ce sont des fractions rationnelles en  $z$  pour lesquelles la propriété suivante est établie (principe du maximum) :

$$\lim_{z \rightarrow -\infty} F(z) = \lim_{z \rightarrow +\infty} F(z) = \lim_{\substack{z=iy \\ y \rightarrow \pm\infty}} F(z) \quad \text{XIV-47}$$

Or on sait que le long de l'axe imaginaire, XIV-46 en module vaut un ; XIV-47 signifie alors que pour  $z$  proche de l'axe réel négatif avec  $\Re(z)$  de grande amplitude, XIV-46 en module est inférieur ou égal à un, mais très proche de un. La composante stiff de XIV-44 est donc, certes atténuée, mais seulement très lentement. Pour XIV-45, on a au contraire

$$\lim_{z \rightarrow -\infty} F(z) = 0 \quad \text{XIV-48}$$

ce qui engendre une réduction beaucoup plus rapide de cette composante. La A-stabilité n'est donc pas une condition suffisante pour assurer une bonne résolution des problèmes stiff : on a besoin pour cela de la L-stabilité, que l'on définit de la manière suivante :

une méthode est L-stable si elle est A-stable et si

$$\lim_{z \rightarrow -\infty} F(z) = 0 \quad \text{XIV-49}$$

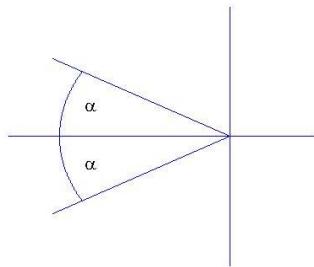
### *A( $\alpha$ )-stabilité*

On vient de voir que la A-stabilité est une propriété insuffisante pour l'intégration des problèmes stiff, mais par ailleurs elle peut être trop sévère dans la mesure où de nombreuses méthodes traitant correctement la stiffness ne sont pas A-stable. C'est par exemple le cas des méthodes ayant l'axe réel négatif dans leur domaine de stabilité : ces méthodes sont en effet capables de résoudre tout problème stiff scalaire sans aucune restriction sur la valeur  $h$ .

C'est la raison de la définition qui suit, utile pour les méthodes à pas multiples : une méthode est dite  $A(\alpha)$ -stable (Widlund, 1967) si le secteur

$$S_\alpha = \{ z : |\arg(-z)| \leq \alpha \}$$

XIV-50



est contenu dans son domaine de stabilité.

### ***méthodes de Rosenbrock***

Inspirées des méthodes RK diagonales implicites, les méthodes de Rosenbrock évitent le recours à l’itération newtonienne pour le calcul des étages  $k_i$ . Les étapes principales de la construction de ces méthodes sont les suivantes : soit à résoudre

$$y' = f(t, y) \quad \text{avec} \quad y_0 = y(t_0) \quad \text{donné.}$$

XIV-51

La première étape consiste à mettre le problème sous la forme suivante dite autonome : on pose

$$z = \begin{pmatrix} t \\ y \end{pmatrix}$$

XIV-52

dont il découle

$$z' = \begin{pmatrix} 1 \\ y' \end{pmatrix}$$

XIV-53

En posant alors

$$g(z) = \begin{pmatrix} 1 \\ f(z) \end{pmatrix}$$

XIV-54

on remplace le problème à résoudre par

$$z' = g(z) \quad \text{avec} \quad z_0 = \begin{pmatrix} 1 \\ y_0 \end{pmatrix} \quad \text{donné.}$$

XIV-55

La résolution de XIV-55 par une méthode RK diagonale implicite peut se mettre sous la forme

$$k_1 = g(z_0 + \alpha_{11}h k_1)$$

$$\vdots$$

$$k_s = g(z_0 + h \sum_{j=1}^s \alpha_{sj} k_j)$$

$$z_1 = z_0 + h \sum_{j=1}^s b_j k_j$$

XIV-56

Rosenbrock propose les modifications suivantes au calcul des étages :

1° les étages  $k_1, \dots, k_{i-1}$  ayant été calculés, le calcul de l'étage suivant est linéarisé :

$$k_i = g(z_0 + h \sum_{j=1}^{i-1} \alpha_{ij} k_j) + g'(z_0 + h \sum_{j=1}^{i-1} \alpha_{ij} k_j) h \alpha_{ii} k_i \quad \text{XIV-57}$$

où  $g'$  est le jacobien de  $g$ .

2°  $g'$  devant être évalué à chaque étage, il élimine cette lourdeur en calculant  $g'$  une fois pour tous les étages : cela donne

$$k_i = g(z_0 + h \sum_{j=1}^{i-1} \alpha_{ij} k_j) + g'(z_0) h \alpha_{ii} k_i \quad \text{XIV-58}$$

3° il « compense » cette simplification en multipliant le jacobien par une combinaison linéaire de tous les étages jusqu'au rang  $i$  :

$$k_i = g(z_0 + h \sum_{j=1}^{i-1} \alpha_{ij} k_j) + g'(z_0) h \sum_{j=1}^i \gamma_{ij} k_j \quad \text{XIV-59}$$

Il y a moyen de déduire ensuite la formulation de cette méthode pour XIV-51 : tous calculs faits, on trouve

$$k_1 = f(t_0 + c_1 h, y_0) + \left( h \left( \frac{\partial f}{\partial y} \right)_{t_0, y_0} \right) \gamma_{11} k_1 + h \gamma_1 \left( \frac{\partial f}{\partial t} \right)_{t_0, y_0}$$

$$\vdots$$

$$k_s = f(t_0 + c_s h, y_0 + \sum_{j=1}^{s-1} \alpha_{sj} k_j) + \left( h \left( \frac{\partial f}{\partial y} \right)_{t_0, y_0} \right) \sum_{j=1}^s \gamma_{sj} k_j + h \gamma_s \left( \frac{\partial f}{\partial t} \right)_{t_0, y_0} \quad \text{XIV-60}$$

$$y_1 = y_0 + h \sum_{j=1}^s b_j k_j$$

$$\text{avec } c_i = \sum_{j=1}^{i-1} \alpha_{ij} \quad \text{et} \quad \gamma_i = \sum_{j=1}^i \gamma_{ij} \quad \text{XIV-61}$$

Les paramètres  $\alpha_{ij}, \gamma_{ij}$  et  $b_i$  sont alors déterminés par des conditions d'ordre selon une démarche identique à celle des méthodes RK classiques ; une hypothèse couramment retenue consiste à imposer

$$\gamma_{11} = \gamma_{22} = \dots = \gamma_{ss} = \gamma$$

XIV-62

Cette condition permet d'économiser de nombreuses opérations dans le calcul des  $k_i$  : chacun d'eux est solution d'un système linéaire qui s'écrit

$$\left( I - h\gamma \left( \frac{\partial f}{\partial y} \right)_{t_0, y_0} \right) k_i = f(t_0 + c_i h, y_0 + h \sum_{j=1}^{i-1} \alpha_{ij} k_j) + \left( h \left( \frac{\partial f}{\partial y} \right)_{t_0, y_0} \right) \sum_{j=1}^{i-1} \gamma_{ij} k_j + h\gamma_i \left( \frac{\partial f}{\partial t} \right)_{t_0, y_0} \quad \text{XIV-63}$$

et dont la matrice est la même pour tous les étages.

Il est à noter que cette importante famille de méthodes constitue un excellent outil de résolution des problèmes stiff

## XIV.5 Stabilité temporelle des intégrateurs à pas multiples

### *A-stabilité et deuxième barrière de Dahlquist*

La A-stabilité est définie de la même manière que pour les intégrateurs à un pas : il faut que  $\{z : \Re(z) \leq 0\}$  fasse partie du domaine de stabilité de l'intégrateur.

Voyons quelques exemples :

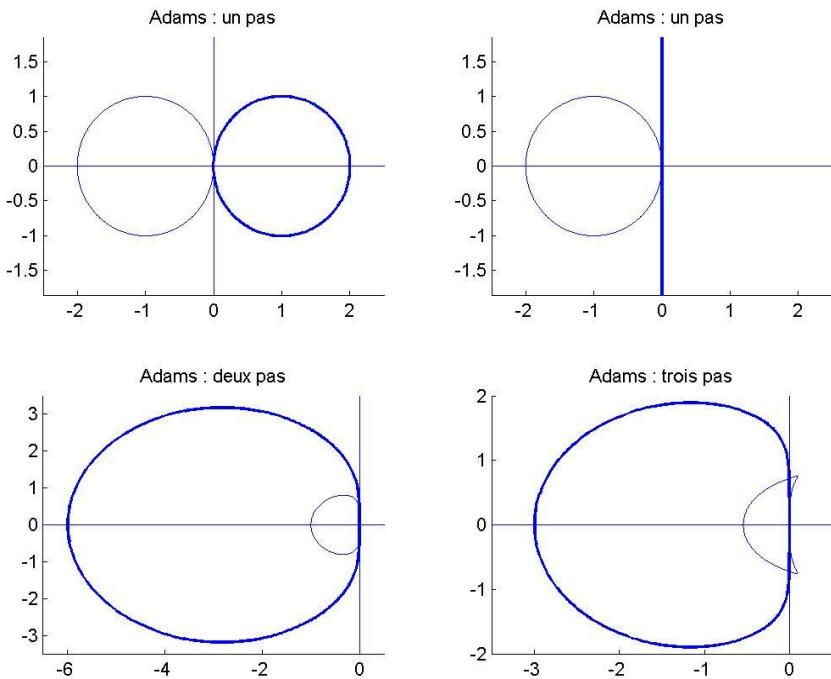
1° méthodes de Adams : les trois méthodes explicites les plus simples sont

un pas :	$u_{n+1} - u_n = \Delta t f_n$	
deux pas :	$u_{n+2} - u_{n+1} = \Delta t \left( \frac{3}{2} f_{n+1} - \frac{1}{2} f_n \right)$	XIV-64
trois pas :	$u_{n+3} - u_{n+2} = \Delta t \left( \frac{23}{12} f_{n+2} - \frac{16}{12} f_{n+1} + \frac{5}{12} f_n \right)$	

et les méthodes implicites correspondantes sont

un pas :	$u_{n+1} - u_n = \Delta t f_{n+1}$	et	$u_{n+1} - u_n = \Delta t \left( \frac{1}{2} f_{n+1} + \frac{1}{2} f_n \right)$	
deux pas :	$u_{n+2} - u_{n+1} = \Delta t \left( \frac{5}{12} f_{n+2} + \frac{8}{12} f_{n+1} - \frac{1}{12} f_n \right)$			XIV-65
trois pas :	$u_{n+3} - u_{n+2} = \Delta t \left( \frac{9}{24} f_{n+3} + \frac{19}{24} f_{n+2} - \frac{5}{24} f_{n+1} + \frac{1}{24} f_n \right)$			

Les domaines de stabilité correspondants (trait fin : méthodes explicites, traits gras : méthodes implicites)



montrent que

- 1° les méthodes explicites ne sont pas A-stables,
- 2° parmi les méthodes implicites seules celles à un pas sont A-stables.
- 3° même quand elles ne sont pas A-stables, les méthodes implicites ont un domaine de stabilité plus grand que les méthodes explicites à même nombre de pas.

Signalons que les deux méthodes implicites à un pas se distinguent par leur ordre :

$$u_{n+1} - u_n = \Delta t f_{n+1} \text{ est d'ordre un}$$

$$u_{n+1} - u_n = \Delta t \left( \frac{1}{2} f_{n+1} + \frac{1}{2} f_n \right) \text{ est d'ordre deux}$$

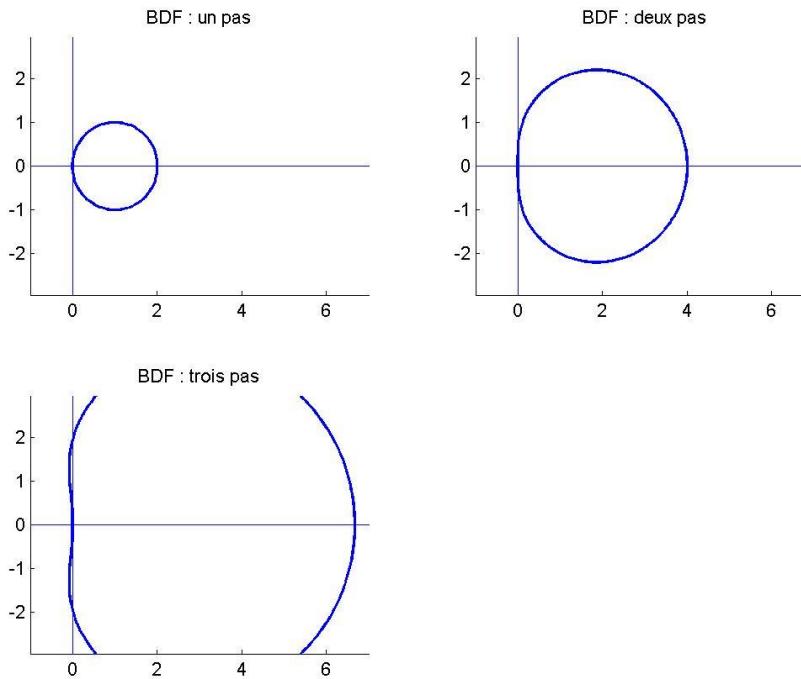
2° méthodes BDF : les trois méthodes les plus simples sont

$$\text{un pas : } u_{n+1} - u_n = \Delta t f_{n+1}$$

$$\text{deux pas : } \frac{3}{2} u_{n+2} - 2u_{n+1} + \frac{1}{2} u_n = \Delta t f_{n+2} \quad \text{XIV-66}$$

$$\text{trois pas : } \frac{11}{6} u_{n+3} - 2u_{n+2} + \frac{3}{2} u_{n+1} - \frac{1}{3} u_n = \Delta t f_{n+3}$$

Comme le montre la figure ci-après, seules les méthodes à un et deux pas sont A-stables. Ces méthodes sont respectivement d'ordre un et deux.



La volonté d'obtenir des méthodes à pas multiples A-stables d'ordre élevé se heurte à la « deuxième barrière de Dahlquist » : en 1963, ce dernier démontre le célèbre théorème suivant : les seules méthodes à pas multiples A-stables sont d'ordre un ou deux.

Cette sévère limitation à l'utilisation pratique des méthodes à pas multiples a été à l'origine de nombreuses recherches dans des sens divers : définir des conditions de stabilité sous forme faible d'une part, élaborer de nouveaux concepts de méthodes à pas multiples d'autre part.

## XIV.6 Nouvelles conditions de stabilité des intégrateurs à pas multiples

### *A( $\alpha$ )-stabilité et stiff-stabilité*

Rappelons qu'une méthode est dite  $A(\alpha)$ -stable si le secteur

$$S_\alpha = \{ z : |\arg(-z)| \leq \alpha \} \quad \text{XIV-67}$$

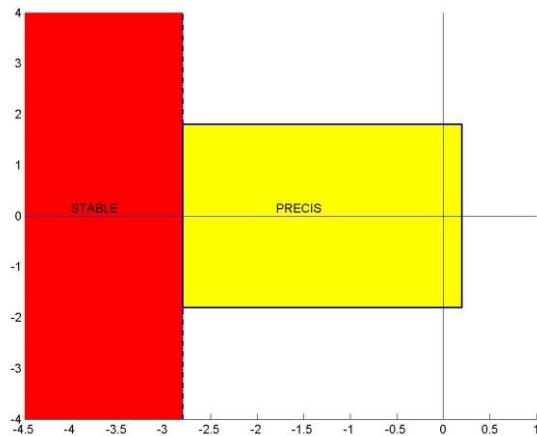
est contenu dans son domaine de stabilité.

Une méthode est stiff-stable (Gear, 1971) si

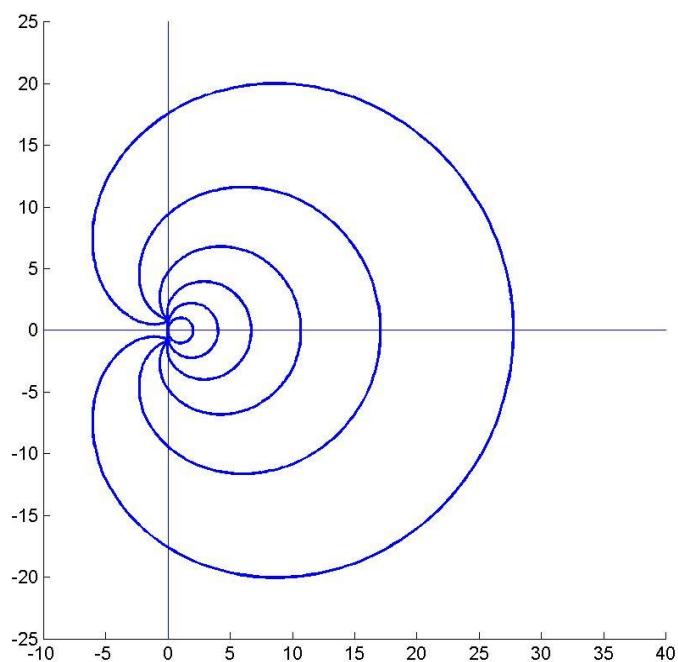
$$S_D = \{ z : \Re(z) < -D \} \quad \text{avec} \quad D > 0 \quad \text{XIV-68}$$

est contenu dans son domaine de stabilité et si la méthode est précise dans le rectangle

$$R = \{ z : -D \leq \Re(z) \leq a \quad \text{et} \quad -\vartheta \leq \Im(z) \leq \vartheta \} \quad \text{avec} \quad a > 0 \quad \text{et} \quad \vartheta \approx \frac{\pi}{5} \quad \text{XIV-69}$$



Cette définition se justifie lorsqu'on examine la forme du domaine de stabilité des six premières méthodes BDF :



Il y correspond le tableau suivant

méthode	$\alpha$	D
un pas	$90^\circ$	0
deux pas	$90^\circ$	0
trois pas	$86.03^\circ$	0.083
quatre pas	$73.35^\circ$	0.667
cinq pas	$51.84^\circ$	2.327
six pas	$17.84^\circ$	6.075

Les deux premières méthodes sont A-stables ; les suivantes sont stiff-stables.

De nombreuses recherches ont alors permis de démontrer le théorème suivant :

$$\forall \alpha < \frac{\pi}{2} \text{ et } \forall k \text{ entier}, \exists \text{ une méthode linéaire à } k \text{ pas A}(\alpha)\text{-stable d'ordre } k.$$

Malheureusement ces méthodes présentent peu d'intérêt en pratique car elles sont caractérisées par d'énormes constantes d'erreur (pour rappel, l'erreur globale d'une méthode à pas multiple d'ordre p vaut

$$e_n = Ch^p \quad \text{avec } C = \text{constante d'erreur}).$$

XIV-70

## XIV.7 Nouvelles méthodes à pas multiples

### *Méthodes à pas multiples avec dérivée seconde*

Cette catégorie de méthodes est une première réponse apportée à l'échec précédent : on y complète la forme classique

$$\alpha_k u_{n+k} + \dots + \alpha_0 u_n = \Delta t (\beta_k f_{n+k} + \dots + \beta_0 f_n) \quad \text{XIV-71}$$

par une combinaison linéaire de valeurs de la dérivée première de f (c'est-à-dire de la dérivée seconde de y) : si on pose

$$g(t, y) \equiv y'' = f' = f_t + f_y f, \quad \text{XIV-72}$$

XIV-71 est remplacé par

$$\alpha_k u_{n+k} + \dots + \alpha_0 u_n = \Delta t (\beta_k f_{n+k} + \dots + \beta_0 f_n) + \Delta t^2 (\gamma_k g_{n+k} + \dots + \gamma_0 g_n) \quad \text{XIV-73}$$

Les paramètres  $\alpha_i, \beta_i, \gamma_i$  sont évidemment à déterminer. Enright (1974) a élaboré une famille de méthodes, qui peuvent être considérées comme la généralisation des méthodes implicites de Adams et dont les trois plus simples sont :

un pas :	$u_{n+1} - u_n = \Delta t \left( \frac{2}{3} f_{n+1} + \frac{1}{3} f_n \right) - \frac{1}{6} \Delta t^2 g_{n+1}$
deux pas :	$u_{n+2} - u_{n+1} = \Delta t \left( \frac{29}{48} f_{n+2} + \frac{20}{48} f_{n+1} - \frac{1}{48} f_n \right) - \frac{1}{8} \Delta t^2 g_{n+2}$
trois pas :	$u_{n+3} - u_{n+2} = \Delta t \left( \frac{614}{1080} f_{n+3} + \frac{513}{1080} f_{n+2} - \frac{54}{1080} f_{n+1} + \frac{7}{1080} f_n \right) - \frac{19}{180} \Delta t^2 g_{n+3}$

XIV-74

Ces méthodes présentent d'excellentes performances en termes de stabilité, d'ordre et de précision. On peut ainsi montrer que l'ordre p et le nombre de pas k sont liés par la relation

$$p = k + 2$$

XIV-75

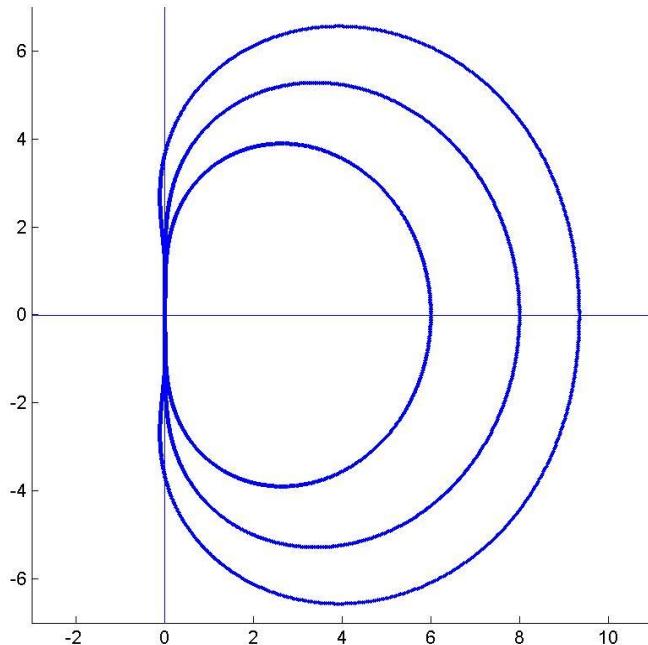
et que les constantes d'erreur de ces méthodes sont très petites. Les domaines de stabilité sont trouvés en généralisant la technique déjà utilisée : on s'intéresse à la résolution de  $y' = \lambda y$ , ce qui implique  $y'' = \lambda^2 y$  ; introduites dans XIV-73, ces relations donnent

$$\alpha_k u_{n+k} + \dots + \alpha_0 u_n = \lambda \Delta t (\beta_k u_{n+k} + \dots + \beta_0 u_n) + \lambda^2 \Delta t^2 (\gamma_k u_{n+k} + \dots + \gamma_0 u_n) \quad \text{XIV-76}$$

Le polynôme dont il faut vérifier que les racines simples (respectivement multiples) sont en modules inférieures ou égales (respectivement inférieures) à un est

$$\rho_z(\xi) = (\alpha_k - z\beta_k - z^2\gamma_k)\xi^k + \dots + (\alpha_0 - z\beta_0 - z^2\gamma_0) \quad \text{XIV-77}$$

Les domaines relatifs à XIV-74 sont les suivants :



Pour ces trois méthodes, la constante d'erreur et les paramètres du domaine de stabilité sont donnés par la table

Méthode	$\alpha$	D	C
un pas	$90^\circ$	0	0.01389
deux pas	$90^\circ$	0	0.00486
trois pas	$87.88^\circ$	0.103	0.00236

Signalons l'existence de méthodes BDF complétées comme les méthodes de Adams par un terme de dérivée seconde. Elles présentent des constantes d'erreur plus grandes que celles des méthodes d'Enright mais sont stiff-stables jusqu'à un ordre plus élevé que ces dernières.

### **Méthodes à pas multiples étendues**

Ces méthodes sont une deuxième réponse apportée à la seconde barrière de Dahlquist. L'idée de base est d'ajouter des pas « dans le futur » à la méthode à pas multiples. Cash, en 1980, développe l'idée pour les méthodes BDF : si on se souvient que la forme générale de ces dernières est

$$\hat{\alpha}_k u_{n+k} + \dots + \hat{\alpha}_0 u_n = \Delta t f_{n+k} \quad \text{XIV-78}$$

en convenant de « chapeauter » les variables  $\alpha$  pour les différencier des variables de la formule suivante), les méthodes étendues obéissent à

$$\alpha_k u_{n+k} + \dots + \alpha_0 u_n = \Delta t \beta_k f_{n+k} + \Delta t \beta_{k+1} f_{n+k+1} \quad \text{XIV-79}$$

Les coefficients de cette formule sont trouvés en résolvant

$$\sum_{i=0}^k \alpha_i i^q = q \sum_{i=0}^{k+1} \beta_i i^{q-1} \quad \text{pour } q = 0, 1, \dots, k+1 \quad \text{XIV-80}$$

et en imposant  $\alpha_k = 1$  XIV-81

Ces coefficients étant trouvés (voir des exemples plus loin), la présence du pas futur  $f_{n+k+1}$  demande de préciser la technique de calcul : la solution étant supposée connue jusqu'à la valeur  $u_{n+k-1}$ , le calcul de  $u_{n+k}$  demande quatre étapes :

1° on calcule une estimation  $\hat{u}_{n+k}$  par la méthode BDF classique à même nombre de pas,

2° on calcule une estimation  $\hat{u}_{n+k+1}$  par la même méthode BDF classique en décalant d'un pas et en utilisant l'estimation  $\hat{u}_{n+k}$  précédente,

3° on calcule une estimation  $\hat{f}_{n+k+1} = f(t_{n+k+1}, \hat{u}_{n+k+1})$ ,

4° on calcule  $u_{n+k}$  par XIV-70 en y remplaçant  $f_{n+k+1}$  par  $\hat{f}_{n+k+1}$ .

On obtient ainsi une procédure de type prédition-correction à trois étages ; par exemple pour les procédures à un, deux et trois pas on trouve

un pas :

$$\text{calcul de } \hat{u}_{n+k} : \quad \hat{u}_{n+1} - u_n = \Delta t f(t_{n+1}, \hat{u}_{n+1}) \quad \text{XIV-82}$$

$$\text{calcul de } \hat{u}_{n+k+1} : \quad \hat{u}_{n+2} - \hat{u}_{n+1} = \Delta t f(t_{n+2}, \hat{u}_{n+2}) \quad \text{XIV-83}$$

$$\text{calcul de } \hat{f}_{n+k+1} : \quad \hat{f}_{n+2} = f(t_{n+2}, \hat{u}_{n+2}) \quad \text{XIV-84}$$

$$\text{calcul de } u_{n+k} : \quad u_{n+1} - u_n = \Delta t \left( \frac{3}{2} f_{n+1} - \frac{1}{2} \hat{f}_{n+2} \right) \quad \text{XIV-85}$$

deux pas :

$$\text{calcul de } \hat{u}_{n+k} : \quad \frac{3}{2} \hat{u}_{n+2} - 2u_{n+1} + \frac{1}{2} u_n = \Delta t f(t_{n+2}, \hat{u}_{n+2}) \quad \text{XIV-86}$$

$$\text{calcul de } \hat{\hat{u}}_{n+k+1} : \quad \frac{3}{2} \hat{\hat{u}}_{n+3} - 2\hat{u}_{n+2} + \frac{1}{2} u_{n+1} = \Delta t f(t_{n+3}, \hat{\hat{u}}_{n+3}) \quad \text{XIV-87}$$

$$\text{calcul de } \hat{f}_{n+k+1} : \quad \hat{f}_{n+3} = f(t_{n+3}, \hat{\hat{u}}_{n+3}) \quad \text{XIV-88}$$

$$\text{calcul de } u_{n+k} : \quad u_{n+2} - \frac{28}{23} u_{n+1} + \frac{5}{23} u_n = \Delta t \left( \frac{22}{23} f_{n+2} - \frac{4}{23} \hat{f}_{n+3} \right) \quad \text{XIV-89}$$

trois pas :

$$\text{calcul de } \hat{u}_{n+k} : \quad \frac{11}{6} \hat{u}_{n+3} - 3u_{n+2} + \frac{3}{2} u_{n+1} - \frac{1}{3} u_n = \Delta t f(t_{n+3}, \hat{u}_{n+3}) \quad \text{XIV-90}$$

$$\text{calcul de } \hat{\hat{u}}_{n+k+1} : \quad \frac{11}{6} \hat{\hat{u}}_{n+4} - 3\hat{u}_{n+3} + \frac{3}{2} u_{n+2} - \frac{1}{3} u_{n+1} = \Delta t f(t_{n+4}, \hat{\hat{u}}_{n+4}) \quad \text{XIV-91}$$

$$\text{calcul de } \hat{f}_{n+k+1} : \quad \hat{f}_{n+4} = f(t_{n+4}, \hat{\hat{u}}_{n+4}) \quad \text{XIV-92}$$

$$\text{calcul de } u_{n+k} : \quad u_{n+3} - \frac{279}{197} u_{n+2} + \frac{99}{197} u_{n+1} - \frac{17}{197} u_n = \Delta t \left( \frac{150}{197} f_{n+3} - \frac{18}{197} \hat{f}_{n+4} \right) \quad \text{XIV-93}$$

Le calcul de  $u_{n+k}$  requiert donc la résolution de trois équations non linéaires (XIV-82, 83 et 85 dans la méthode à un pas) ou de trois systèmes non linéaires dans  $\Re_N$ .

Le calcul du domaine de stabilité résulte d'une démarche analogue à ce qu'on a déjà vu pour les méthodes de prédition-correction classique ; il en résulte un lieu de la frontière du domaine de stabilité qui est une équation du troisième degré en  $z$ . On trouve

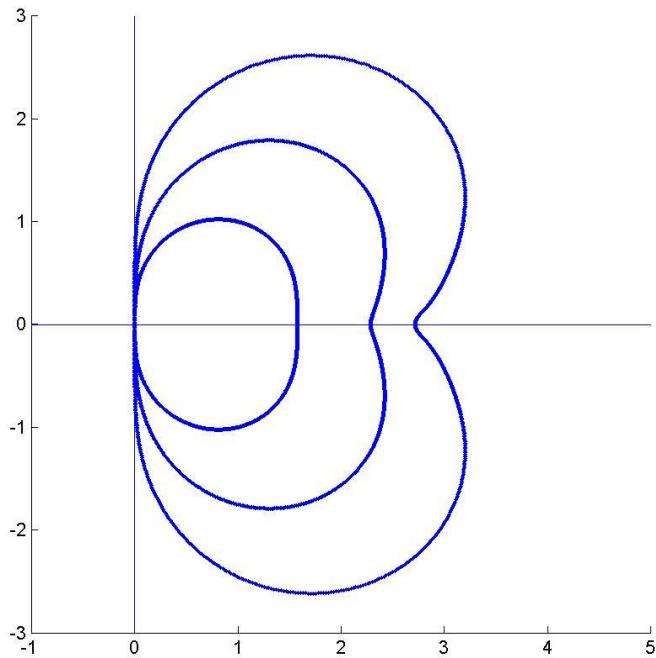
$$Az^3 + Bz^2 + cz + D = 0 \quad \text{XIV-94}$$

avec

$$\begin{aligned} A &= \beta_k \exp(kj\theta) \\ B &= -2\hat{\alpha}_k \beta_k \exp(kj\theta) - T + \beta_{k+1} S \\ C &= \hat{\alpha}_k^2 \beta_k \exp(kj\theta) + 2\hat{\alpha}_k T - \hat{\alpha}_k \beta_{k+1} S + \hat{\alpha}_{k-1} \beta_{k+1} R \\ D &= -\hat{\alpha}_k^2 T \end{aligned} \quad \text{XIV-95}$$

$$R = \sum_{i=0}^{k-1} \hat{\alpha}_i \exp(ij\theta), \quad S = \sum_{i=0}^{k-2} \hat{\alpha}_i \exp((i+1)j\theta), \quad T = \sum_{i=0}^k \alpha_i \exp(ij\theta)$$

Pour les méthodes à un, deux et trois pas, cela donne les domaines de la figure suivante. Ces trois méthodes sont A-stables ; les méthodes d'ordre 5 à 9 (c'est-à-dire de 4 à 8 pas) sont stiff-stables.



#### XIV.8 Stiffness et méthode des lignes

Tout ce qui précède concerne la stiffness des équations différentielles ordinaires  $\bar{y}' = \bar{f}(t, \bar{y})$  sans préjuger de la structure de  $\bar{f}(t, \bar{y})$ . Il est donc utile de savoir si la méthode des lignes génère des systèmes différentiels stiff ou non. Pour répondre à cette question, on se contentera de reprendre l'exemple simple de l'équation de diffusion-convection

$$u_t = -au_x + bu_{xx} \quad a \text{ et } b > 0 \quad \text{XIV-10}$$

traitée en utilisant quelques schémas de différences finies pour le calcul des dérivées spatiales. Il est aisé de calculer ce que vaut le spectre des valeurs propres de la matrice de discrétisation spatiale pour ces différents schémas. La table suivante donne le maximum  $\max_k \Re(-\lambda_k)$  de ce spectre pour diverses combinaisons de schémas de calcul de dérivées première et seconde :

$\max_k \Re(-\lambda_k)$		$u_{xx}$	
		3 pts centrés	5 pts centrés
$u_x$	2 pts upwind	$\frac{2a}{\Delta x} + \frac{4b}{\Delta x^2}$	$\frac{2a}{\Delta x} + \frac{8b}{\Delta x^2}$
	3 pts centrés	$\frac{4b}{\Delta x^2}$	$\frac{8b}{\Delta x^2}$
	3 pts upwind	$\frac{4a}{\Delta x} + \frac{4b}{\Delta x^2}$	$\frac{4a}{\Delta x} + \frac{8b}{\Delta x^2}$

Plusieurs remarques s'imposent :

1° dans tous les cas le rapport de stiffness vaut l'infini car tous les spectres obtenus contiennent la valeur propre  $\lambda = 0$ . Ce rapport ne pouvant donc pas être utilisé, on se contente d'évaluer  $\max_k \Re(-\lambda_k)$  pour avoir une idée de la stiffness.

2° d'une manière générale la stiffness augmente si

- on réduit le pas de discrétisation spatiale  $\Delta x$
- on prend des schémas upwind plutôt que centrés pour le calcul de la dérivée première
- on augmente le nombre de points des schémas de discrétisation

3° lorsqu'on réduit  $\Delta x$ , le terme de diffusion engendre une augmentation de la stiffness plus marquée que le terme de convection.

#### XIV.9 Stabilité dans le cas des problèmes non linéaires

Comme on l'a mentionné à de multiples reprises, la théorie de la stabilité absolue fournit la condition nécessaire et suffisante de stabilité d'une méthode numérique lorsque l'équation différentielle ordinaire à résoudre est linéaire, c'est-à-dire lorsqu'elle s'écrit

$$\bar{y}' = A\bar{y} + \bar{b} \quad \text{XIV-96}$$

Dans le cas général

$$\bar{y}' = \bar{f}(t, \bar{y}) \quad \text{XIV-97}$$

elle fournit seulement une condition nécessaire. C'est en 1975 que Dahlquist proposa un élargissement de cette théorie. Il définit d'abord les équations sur lesquelles bâtir la généralisation de la stabilité absolue : il proposa de s'intéresser aux équations différentielles vérifiant une condition de Lipschitz dite unilatérale pour la norme euclidienne : XIV-97 vérifie une telle condition si on a

$$\langle \bar{f}(t, \bar{y}) - \bar{f}(t, \bar{z}), \bar{y} - \bar{z} \rangle \leq v \|\bar{y} - \bar{z}\|^2 \quad \text{XIV-98}$$

Dans cette relation,  $\langle \bar{p}, \bar{q} \rangle$  est le produit scalaire :

$$\langle \bar{p}, \bar{q} \rangle = \sum_{i=1}^n p_i q_i \quad \text{XIV-99}$$

et  $v$  est dite constante de Lipschitz unilatérale de  $\bar{f}$ . L'intérêt de cette catégorie d'équations différentielles apparaît grâce au théorème suivant.

**Théorème** : si  $\bar{f}(t, \bar{y})$  est continue et vérifie une condition de Lipschitz unilatérale, alors deux solutions quelconques  $\bar{y}(t)$  et  $\bar{z}(t)$  de XIV-97 vérifient

$$\|\bar{y}(t) - \bar{z}(t)\| \leq \|\bar{y}(t_0) - \bar{z}(t_0)\| e^{v(t-t_0)} \quad \forall t \geq t_0 \quad \text{XIV-100}$$

Quand  $v$  est négatif ou nul, XIV-100 montre que la distance entre deux solutions ne grandit pas avec  $t$  : l'équation différentielle est stable ; il en résulte alors que XIV-98 s'écrit

$$\langle \bar{f}(t, \bar{y}) - \bar{f}(t, \bar{z}), \bar{y} - \bar{z} \rangle \leq 0 \quad \text{XIV-101}$$

Cette condition, dite de contractivité, est la condition de stabilité générale du problème non linéaire XIV-97.

Dahlquist s'est alors intéressé aux critères que devait remplir une méthode de résolution d'équations différentielles ordinaires pour obtenir également la contractivité des solutions numériques. Comme on va le voir, il est nécessaire de distinguer les méthodes à un pas et les méthodes à pas multiples.

#### XIV.10 Stabilité générale des méthodes à un pas : théorie de la B-stabilité

##### B-stabilité

Une méthode d'intégration d'équations différentielles ordinaires à un pas est dite B-stable si pour toute équation différentielle vérifiant la condition de contractivité, on a

$$\|\bar{y}_1 - \bar{z}_1\| \leq \|\bar{y}_0 - \bar{z}_0\| \quad \forall h \quad \text{XIV-102}$$

$\bar{y}_1$  et  $\bar{z}_1$  représentant les solutions numériques obtenues après un pas d'intégration en démarrant avec les valeurs initiales  $\bar{y}_0$  et  $\bar{z}_0$ .

**Théorème :** la B-stabilité implique la A-stabilité.

**Critère algébrique de B-stabilité :** si les coefficients d'une méthode RK à  $s$  étages vérifient

- a)  $b_i \geq 0$  pour  $i = 1, \dots, s$
- b)  $[M] \equiv [b_i a_{ij} + b_j a_{ji} - b_i b_j]_{i,j=1,\dots,s}$  est une matrice définie non négative

alors la méthode est B-stable. Toute méthode vérifiant ces deux conditions est dite algébriquement stable. Notons que ce théorème établit la relation

stabilité algébrique  $\Rightarrow$  B-stabilité

mais pas la relation inverse.

**Exemple :** soit la méthode IRK à deux étages

$$\begin{array}{c|cc} \gamma & \gamma \\ \hline 1-\gamma & 1-2\gamma & \gamma \\ \hline & 0.5 & 0.5 \end{array} \quad \text{XIV-103}$$

La matrice  $[M]$  ci-dessus vaut

$$\begin{bmatrix} 2b_1 a_{11} - b_1^2 & b_1 a_{12} + b_2 a_{21} - b_1 b_2 \\ b_1 a_{12} + b_2 a_{21} - b_1 b_2 & 2b_2 a_{22} - b_2^2 \end{bmatrix} = \dots = \left( \gamma - \frac{1}{4} \right) \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

Ses valeurs propres valent

$$\lambda_{1,2} = \left\{ 0, 2\left(\gamma - \frac{1}{4}\right) \right\}$$

La condition de non négativité est donc  $\gamma \geq \frac{1}{4}$  (qui est aussi – cf. paragraphe XIV.4 – la condition de A-stabilité ).

La condition de non négativité étant peu pratique à manipuler, il est plus commode d'utiliser le théorème suivant.

**Théorème** : rappelons d'abord les conditions d'ordre générales des méthodes IRK : une méthode IRK est d'ordre  $p$  si les conditions suivantes sont réunies :

$$\sum_{i=1}^s b_i c_i^{q-1} = \frac{1}{q} \quad q : 1, \dots, p \quad (1)$$

$$\sum_{j=1}^s a_{ij} c_j^{q-1} = \frac{c_i^q}{q} \quad i : 1, \dots, s \quad q : 1, \dots, \eta \quad (2)$$

$$\sum_{i=1}^s b_i c_i^{q-1} a_{ij} = \frac{b_j}{q} \left(1 - c_j^q\right) \quad j : 1, \dots, s \quad q : 1, \dots, \xi \quad (3)$$

pour  $p \leq \eta + \xi + 1$  et  $p \leq 2\eta + 2$

ces relations servent aussi à établir la B-stabilité : la méthode est B-stable si

- (1) est vérifié pour  $p = 2s - 2$
- (2) est vérifié pour  $\eta = s - 1$
- (3) est vérifié pour  $\xi = s - 1$

et si la fonction de stabilité  $F(z)$  de la méthode vérifie  $|F(\infty)| \leq 1$ . On vérifiera aisément que la

méthode XIV-111 répond à ces critères si  $\gamma \geq \frac{1}{4}$ .

Signalons pour terminer que la supériorité de la B-stabilité sur la A-stabilité n'était a priori pas établie : rien ne prouvait donc que la capacité de traiter des équations différentielles vérifiant la condition de contractivité conférait une stabilité supérieure à celle requise pour résoudre les équations différentielles linéaires. Cette supériorité a néanmoins pu être établie en 1986 par E. Hairer dans « A- and B-Stability for Runge-Kutta Methods - Characterization and Equivalence » paru dans Numerische Mathematik, 48, 383-389, Springer Verlag.

## XIV.11 Stabilité générale des méthodes à pas multiple : méthode one-leg et G-stabilité

### Méthode one-leg

Soit la méthode linéaire à  $k$  pas

$$\alpha_k y_{m+k} + \dots + \alpha_0 y_m = h(\beta_k f_{m+k} + \dots + \beta_0 f_m)$$

XIV-104

$$\text{avec } f_{m+i} = f(t_{m+i}, y_{m+i}) \quad i : 0, \dots, k$$

et ses polynômes génératrices

$$\rho(\xi) = \sum_{i=0}^k \alpha_i \xi^i \quad \text{et} \quad \sigma(\xi) = \sum_{i=0}^k \beta_i \xi^i \quad \text{XIV-105}$$

Si ces polynômes sont à coefficients réels et n'ont pas de diviseurs communs et si on fait l'hypothèse de la normalisation

$$\sigma(1) = 1 \quad \text{XIV-106}$$

alors, la méthode one-leg associée à XIV-104 est définie par la relation

$$\alpha_k y_{m+k} + \dots + \alpha_0 y_m = h f \left( \sum_{i=0}^k \beta_i t_{m+i}, \sum_{i=0}^k \beta_i y_{m+i} \right) \quad \text{XIV-107}$$

Observons d'abord que si l'équation différentielle à résoudre est linéaire :

$$y' = Ay \quad \text{XIV-108}$$

alors XIV-104 et 107 sont identiques : le membre de droite de XIV-107 s'écrit en effet

$$h f \left( \sum_{i=0}^k \beta_i t_{m+i}, \sum_{i=0}^k \beta_i y_{m+i} \right) = h A \left( \sum_{i=0}^k \beta_i y_{m+i} \right) = h \sum_{i=0}^k \beta_i A y_{m+i}$$

et celui de XIV-104 vaut

$$h (\beta_k f_{m+k} + \dots + \beta_0 f_m) = h (\beta_k A y_{m+k} + \dots + \beta_0 A y_m)$$

Observons ensuite que XIV-104 et XIV-107 sont aussi identiques pour les méthodes BDF, quelle que soit l'équation différentielle à résoudre, : pour ces méthodes, XIV-104 s'écrit en effet

$$\alpha_k y_{m+k} + \dots + \alpha_0 y_m = h \beta_k f_{m+k}$$

qui est aussi la forme de XIV-107 puisque le polynôme  $\sigma(\xi)$  ne compte qu'un seul terme.

Dans tous les autres cas, les formules sont différentes, mais les solutions qu'elles génèrent sont reliées par certaines transformations : soit par exemple la méthode à deux pas

$$y_{m+1} - y_m = \frac{h}{2} (f(t_m, y_m) + f(t_{m+1}, y_{m+1})) \quad \text{XIV-109}$$

et sa méthode one-leg correspondante :

$$y_{m+1} - y_m = \frac{h}{2} f \left( \frac{t_m + t_{m+1}}{2}, \frac{y_m + y_{m+1}}{2} \right) \quad \text{XIV-110}$$

Il est aisément de montrer que si  $\{y_m\}$  est solution de XIV-110, alors

$$\hat{y}_m = \frac{1}{2}(y_m + y_{m+1}), \quad \hat{t}_m = \frac{1}{2}(t_m + t_{m+1})$$

est solution de XIV-109. Réciproquement, si  $\{\hat{y}_m, \hat{t}_m\}$  est solution de XIV-109, alors

$$y_m = \hat{y}_m - \frac{1}{2}f(\hat{t}_m, \hat{y}_m) \quad t_m = \hat{t}_m - \frac{h}{2}$$

est solution de XIV-110.

### G-stabilité

Le point de départ est le même que pour les méthodes à un pas : on souhaite que la solution numérique à toute équation différentielle vérifiant la condition de contractivité XIV-101 soit également contractive. Une difficulté apparaît à ce stade : la condition de contractivité des solutions numériques ne peut pas prendre la forme XIV-102 car pour les méthodes à pas multiples,  $y_{m+k}$  dépend des  $k$  valeurs antérieures  $y_{m+k-1}, \dots, y_m$ . C'est la raison pour laquelle on est amené à manipuler les vecteurs

$$Y_m = (y_{m+k-1}, \dots, y_m)^T \quad \text{XIV-111}$$

et la norme

$$\|Y_m\|_G^2 = \sum_{i=1}^k \sum_{j=1}^k g_{ij} < y_{m+i-1}, y_{m+j-1} > \quad \text{XIV-112}$$

où  $<., .>$  est le produit scalaire XIV-99 et où  $G = (g_{i,j})_{i,j=1,\dots,k}$  est une matrice réelle, symétrique et définie positive.

Dahlquist propose alors en 1975 la définition suivante : la méthode one-leg XIV-107 est dite G-stable si il existe une matrice  $G$  réelle, symétrique et définie positive telle que pour deux solutions numériques  $\{y_m\}$  et  $\{z_m\}$  on ait

$$\|Y_{m+1} - Z_{m+1}\|_G \leq \|Y_m - Z_m\|_G \quad \forall h \quad \text{XIV-113}$$

**Théorème :** La G-stabilité implique la A-stabilité.

**Critère algébrique de G-stabilité :** soit la méthode à pas multiple de polynômes générateurs  $\rho$  et  $\sigma$  ; s'il existe une matrice réelle, symétrique et définie positive  $G$  et des nombres réels  $a_0, \dots, a_k$  tels que

$$\frac{1}{2}(\rho(\xi)\sigma(\omega) + \rho(\omega)\sigma(\xi)) = (\xi\omega - 1) \sum_{i,j=1}^k g_{ij} \xi^{i-1} \omega^{j-1} + \left( \sum_{i=0}^k a_i \xi^i \right) \left( \sum_{j=0}^k a_j \omega^j \right) \quad \text{XIV-114}$$

alors la méthode one-leg correspondante est G-stable.

Ce critère algébrique appelle évidemment une question : pour quelles méthodes de polynômes générateurs  $\rho$  et  $\sigma$  XIV-114 est-il vérifié ? La réponse à cette question fut également fournie par Dahlquist en 1978 dans un théorème célèbre :

**Théorème :** la méthode one-leg correspondant à la méthode à pas multiple  $(\rho, \sigma)$  est G-stable si et seulement si  $\rho$  et  $\sigma$  n'ont pas de facteur commun et si cette méthode  $(\rho, \sigma)$  est A-stable.

## XIV.12 Les intégrateurs de matlab

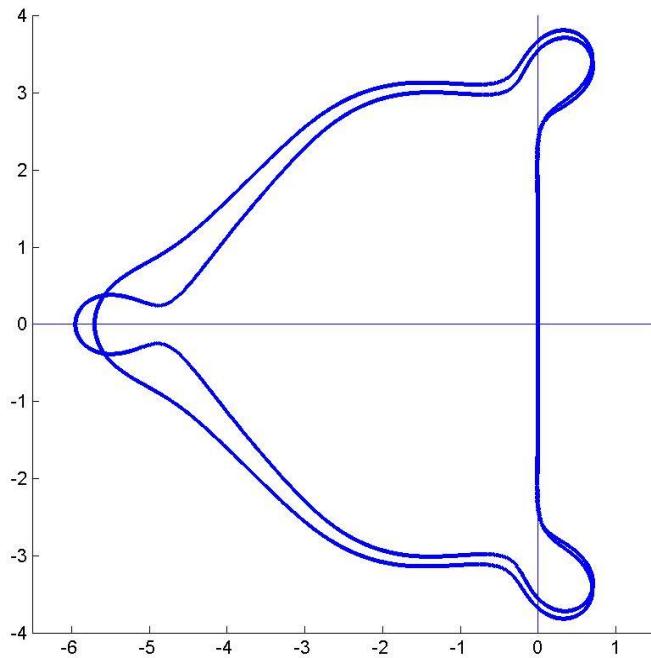
### Ode45

Elaboré par Dormand et Prince (1980) et bâti sur un couple RK imbriqué constitué d'une formule d'ordre quatre et d'une formule d'ordre cinq, cet intégrateur est décrit par le tableau suivant :

0						
$\frac{2}{9}$	$\frac{2}{9}$					
$\frac{1}{3}$	$\frac{1}{12}$	$\frac{1}{4}$				
$\frac{5}{9}$	$\frac{55}{324}$	$-\frac{25}{108}$	$\frac{50}{81}$			
$\frac{2}{3}$	$\frac{83}{330}$	$-\frac{13}{22}$	$\frac{61}{66}$	$\frac{9}{110}$		
1	$-\frac{19}{28}$	$\frac{9}{4}$	$\frac{1}{7}$	$-\frac{27}{7}$	$\frac{22}{7}$	
1	$\frac{19}{200}$	0	$\frac{3}{5}$	$-\frac{243}{400}$	$\frac{33}{40}$	$\frac{7}{80}$
	$\frac{19}{200}$	0	$\frac{3}{5}$	$-\frac{243}{400}$	$\frac{33}{40}$	$\frac{7}{80}$
	$\frac{431}{5000}$	0	$\frac{333}{500}$	$-\frac{7857}{10000}$	$\frac{957}{1000}$	$\frac{193}{2000}$
						$-\frac{1}{50}$

Les coefficients de ce tableau ont été déterminés afin de rencontrer deux objectifs :

- 1° réduire le coefficient du premier terme de l'erreur de troncature dans la relation du cinquième ordre
- 2° étendre au maximum le domaine de stabilité ; pour les deux formules ci-dessus, ces domaines sont les suivants :



Signalons qu'un objectif important à atteindre dans l'élaboration d'une paire RK imbriquée efficace est l'obtention de domaines de stabilité les plus proches possibles pour les deux formules. En effet, le contrôle du pas résulte du calcul de l'écart qui sépare les résultats fournis par ces deux formules lorsqu'un pas de calcul est effectué : il est clair que si le pas utilisé conduit à des valeurs de  $\Omega\Delta t$  intérieures seulement à un des deux domaines de stabilité, l'écart calculé n'a aucun sens.

L'ajustement automatique et permanent du pas d'intégration a un effet indésirable : l'utilisateur souhaite en général obtenir la solution à des instants  $t_k$  uniformément répartis sur l'intervalle d'intégration qui, sauf par chance, ne coïncident pas avec les pas générés par l'intégration. Il est donc indispensable d'adoindre une procédure d'interpolation ; dans ode45, celle-ci est d'ordre 4 et présente l'avantage de ne nécessiter aucune évaluation de  $f(t, y)$  supplémentaire à celles requises pour le calcul des étages.

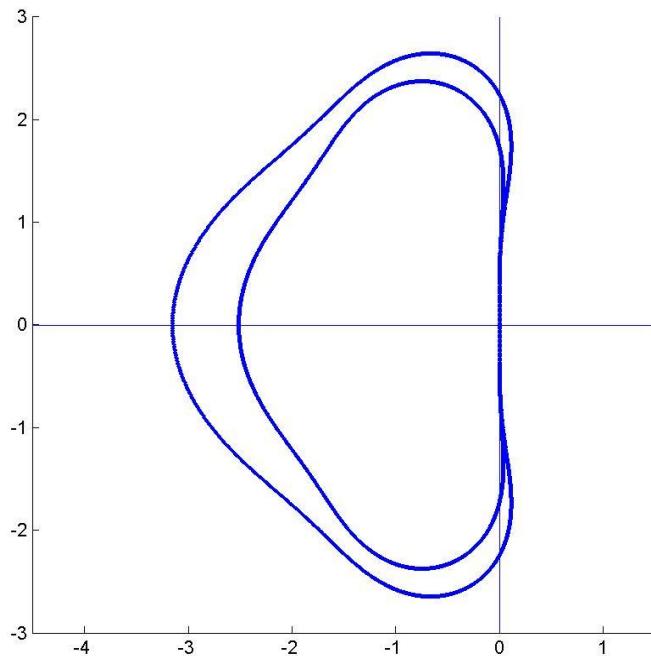
### *Ode23*

Il s'agit également d'un couple RK imbriqué regroupant une formule d'ordre trois et une formule d'ordre deux.

Conformément à la théorie des méthodes explicites, la formule d'ordre 3 est à trois étages. Une telle formule a deux degrés de liberté : deux de ses coefficients peuvent être choisis arbitrairement ; on sait aussi que son domaine de stabilité est indépendant de ces choix, tout comme la valeur du premier terme de son erreur de troncature. Les deux coefficients ont donc été choisis de manière à annuler les deuxième et troisième termes de cette erreur. Il en résulte le tableau suivant :

	0		
$\frac{1}{2}$	$\frac{1}{2}$		
$\frac{3}{4}$	0	$\frac{3}{4}$	
	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{4}{9}$

La formule d'ordre deux a été obtenue en utilisant les coefficients  $b_k$  de la formule précédente pour générer un quatrième étage, sans calcul supplémentaire : il s'agit donc d'une formule à quatre étages d'ordre deux. Elle dispose de quatre degrés de liberté (les coefficients  $\hat{b}_k$ ) dont deux sont mis à profit pour conférer l'ordre deux à la méthode, et dont les deux autres ont été fixés de manière à obtenir un domaine de stabilité comparable à celui de la formule d'ordre trois (en réalité, il est un peu plus grand) :



La table des coefficients de la formule d'ordre deux est la suivante :

	0		
$\frac{1}{2}$	$\frac{1}{2}$		
$\frac{3}{4}$	0	$\frac{3}{4}$	
1	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{4}{9}$
	$\frac{7}{24}$	$\frac{1}{4}$	$\frac{1}{3}$
			$\frac{1}{8}$

On peut légitimement s'interroger sur la pertinence d'une telle paire, d'ordre peu élevé quand on la compare à la paire de ode45 : sa justification réside dans une constatation communément faite : lorsque les exigences de tolérance sont peu élevées, les paires d'ordre plus petit se montrent généralement plus efficaces.

A titre d'exemple, la résolution de l'équation de Burgers

$$u_t - \left( \frac{u^2}{2} \right)_x + \mu u_{xx} \quad \text{avec} \quad \mu = 0.0001$$

sur une grille à 1001 points conduit aux temps de calcul suivants

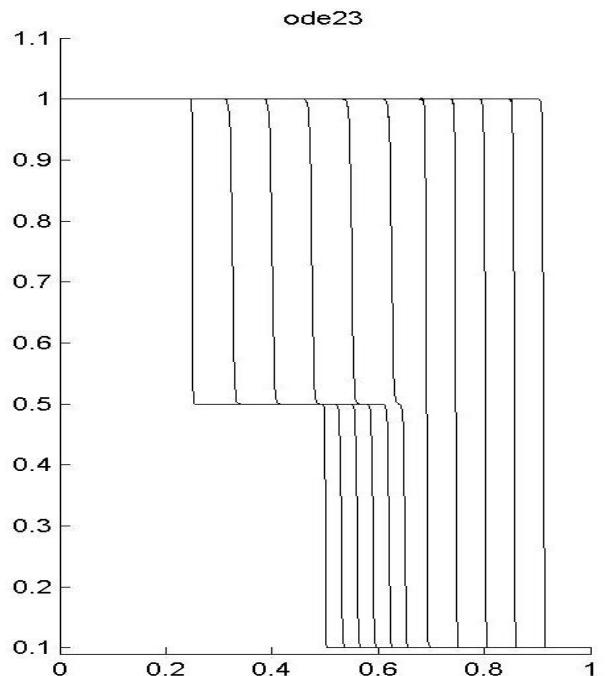
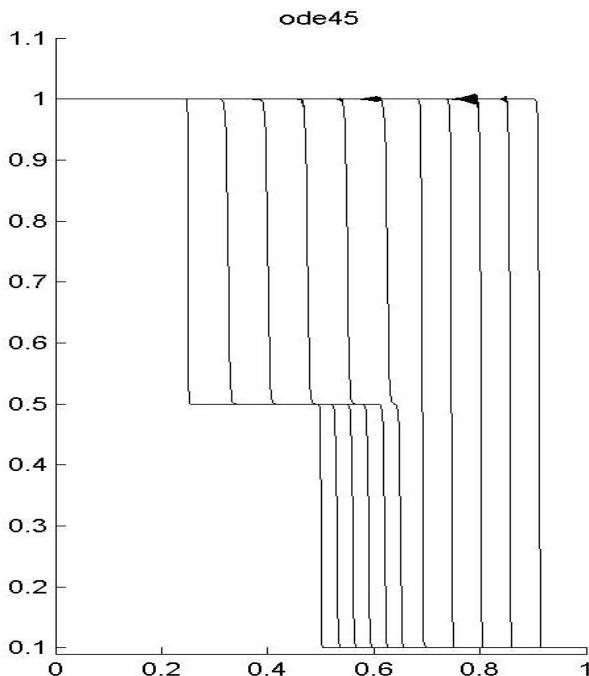
Intégrateur	CPU
Ode45	4.24
Ode23	3.60

si les tolérances sont les valeurs standards de Matlab. Les graphes obtenus sont alors identiques.

Si on passe à Reltol = Abstol = 0.01 on obtient

Intégrateur	CPU
Ode45	4.45
Ode23	2.80

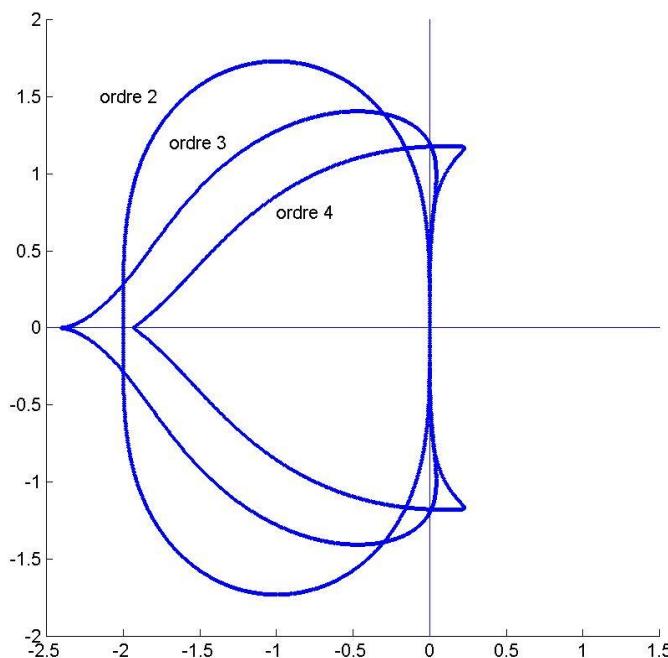
En outre, la solution obtenue avec ode45 est moins bonne :



Comme pour ode45, ode23 est muni d'une procédure d'interpolation n'exigeant aucune évaluation supplémentaire de  $f(t, y)$ .

### Ode 113

Il s'agit d'une méthode de prédiction-correction construite sur les méthodes d'Adams : la formule de prédiction est une formule d'Adams explicite et celle de correction, d'Adams implicite. Les formules sont accolées en respectant le bon accord entre les ordres des deux formules : la prédiction est d'ordre  $k$  et la correction d'ordre  $k + 1$ . Le code contient les couples prédiction-correction jusqu'à  $k = 12$ . Les domaines de stabilité des trois premiers couples sont les suivants :



Sans entrer dans le détail des algorithmes mis en œuvres, signalons qu'ode113 est muni de plusieurs procédures indispensables à un fonctionnement correct :

- une procédure d'acceptation ou de rejet de la valeur obtenue après chaque pas calculé,
- une procédure de sélection de l'ordre de la méthode,
- une procédure d'adaptation du pas,
- une procédure de démarrage.

Ode113 est particulièrement efficace quand les exigences de tolérance sont très sévères. A nouveau, l'équation de Burgers traitée par ode45 et ode113 débouche sur les temps de calcul suivants quand on modifie le niveau des tolérances :

Tolérances	Ode45	Ode113
Valeurs standards de Matlab	4.25	5.17
Re $\text{ltol} = \text{Abstol} = 1e - 10$	119	15.9

Cette efficacité provient de ce que l'intégrateur peut utiliser un schéma de prédiction-correction d'ordre très élevé (jusqu'à 13) quand les tolérances deviennent strictes.

### Ode15s

Ode15s est basé sur une version améliorée des méthodes BDF classiques. Il est aisément de montrer que ces dernières peuvent être mises sous la forme générale

$$\sum_{m=1}^k \frac{1}{m} \nabla^m u_{n+1} = \Delta t f(t_{n+1}, u_{n+1}) \quad \text{XIV-115}$$

où  $k$  désigne l'ordre de la méthode et avec

$$\nabla u_{n+1} = u_{n+1} - u_n \quad \text{et} \quad \nabla^{k+1} u_{n+1} = \nabla^k u_{n+1} - \nabla^k u_n \quad \text{XIV-116}$$

Le caractère implicite de la méthode est traité par une itération de Newton appliquée en général à la valeur de prédiction suivante

$$u_{n+1}^{(0)} = \sum_{m=0}^k \nabla^m u_n \quad \text{XIV-117}$$

Par exemple, pour  $k = 2$ , XIV-115, 116 et 117 donnent

$$\frac{3}{2} u_{n+1} - 2u_n + \frac{1}{2} u_{n-1} = \Delta t f_{n+1} \quad \text{XIV-118}$$

$$\text{et} \quad u_{n+1}^{(0)} = 3u_n - 3u_{n-1} + u_{n-2} \quad \text{XIV-119}$$

Sachant également que le premier terme de l'erreur de troncature de la méthode d'ordre  $k$  peut être approximé par la formule suivante :

$$\frac{1}{k+1} \Delta t^{k+1} \frac{d^{k+1} y}{dt^{k+1}} \approx \frac{1}{k+1} \nabla^{k+1} y_{n+1}, \quad \text{XIV-120}$$

Klopfenstein (1971) propose d'exploiter la dépendance de  $u_{n+1}^{(0)}$  vis-à-vis de valeurs antérieures absentes du calcul de  $u_{n+1}$  pour obtenir des méthodes plus performantes que les BDF classiques. Il propose les méthodes NDF (« numerical differentiation formulas ») décrites par

$$\sum_{m=1}^k \frac{1}{m} \nabla^m u_{n+1} - \kappa \gamma_k (u_{n+1} - u_{n+1}^{(0)}) = \Delta t f(t_{n+1}, y_{n+1}) \quad \text{XIV-121}$$

$$\text{avec} \quad \gamma_k = \sum_{j=1}^k \frac{1}{j} \quad \text{XIV-122}$$

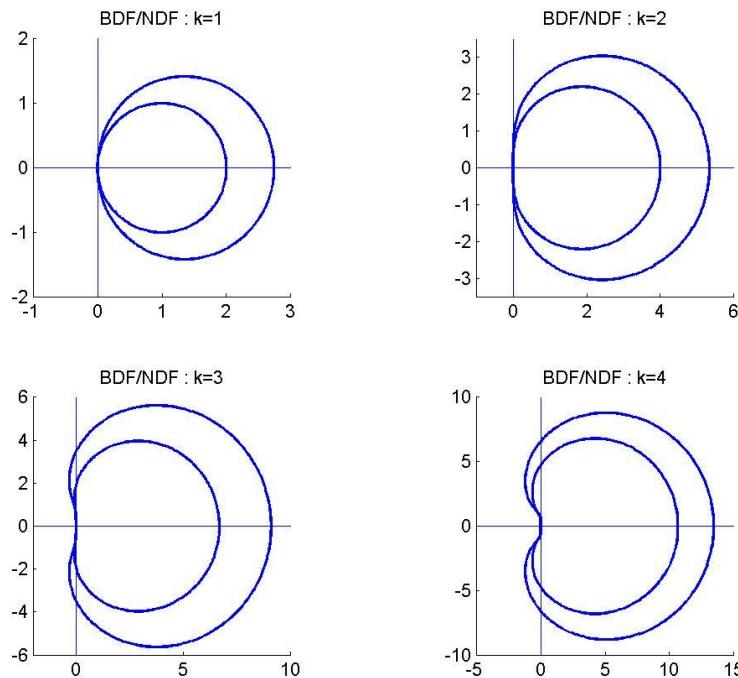
et où  $\kappa$  est un scalaire permettant soit d'augmenter la taille du domaine de stabilité, soit de réduire le coefficient du premier terme de l'erreur de troncature : celui-ci devient

$$\left( \kappa \gamma_k + \frac{1}{k+1} \right) \Delta t^{k+1} \frac{d^{k+1} y}{dt^{k+1}} \quad \text{XIV-123}$$

Ode15s utilise XIV-121 pour  $k$  variant de 1 à 5 avec les valeurs de  $\kappa$  données par la table qui suit. Ces valeurs ont été ajustées par tâtonnements de manière à préserver les propriétés de stabilité de la méthode BDF correspondante : A-stabilité quand elle est acquise, ou quasi même A( $\alpha$ )-stabilité. Pour

$k < 5$ , l'angle  $\alpha$  de la méthode NDF est légèrement plus petit que celui de la méthode BDF correspondante, mais la réduction du premier terme de l'erreur de troncature permet d'atteindre la même précision que la méthode BDF avec une augmentation du pas  $\Delta t$ . Pour  $k = 5$ , l'angle  $\alpha$  étant assez petit, on a préféré en conserver la valeur et revenir donc à la méthode BDF classique.

ordre k	$\kappa$	$\Delta t_{NDF} / \Delta t_{BDF}$	$A(\alpha)_{BDF}$	$A(\alpha)_{NDF}$
1	-0.1850	1.26	90°	90°
2	-1/9	1.26	90°	90°
3	-0.0823	1.26	86°	80°
4	-0.0415	1.12	73°	66°
5	0	1	51°	51°



Ode15s est capable de résoudre les systèmes différentiels-algébriques.

### Ode23s

Cet intégrateur est basé sur une méthode de Rosenbrock à deux étages et munie d'une estimation d'erreur utilisant trois évaluations de  $f(t, y)$ . Rappelons la formule de calcul des étages d'une telle formule :

$$\left( I - h\gamma \left( \frac{\partial f}{\partial y} \right)_{t_0, y_0} \right) k_i = f(t_0 + c_i h, y_0 + h \sum_{j=1}^{i-1} \alpha_{ij} k_j) + \left( h \left( \frac{\partial f}{\partial y} \right)_{t_0, y_0} \right) \sum_{j=1}^{i-1} \gamma_{ij} k_j + h\gamma_i \left( \frac{\partial f}{\partial t} \right)_{t_0, y_0} \quad \text{XIV-124}$$

Posant  $J = \left( \frac{\partial f}{\partial y} \right)_{t_0, y_0}$  et  $W = (I - h\gamma J)$ , la méthode à deux étages s'écrit

$$W k_1 = f(t_0, y_0) + h\gamma_1 \left( \frac{\partial f}{\partial t} \right)_{t_0, y_0} \quad \text{XIV-125}$$

$$W k_2 = f(t_0 + c_2 h, y_0 + h\alpha_{21} k_1) + hJ\gamma_{21} k_1 + h\gamma_2 \left( \frac{\partial f}{\partial t} \right)_{t_0, y_0} \quad \text{XIV-126}$$

$$y_1 = y_0 + h(b_1 k_1 + b_2 k_2) \quad \text{XIV-127}$$

où (cf. XIV-61,62) les paramètres de la méthode doivent vérifier

$$\begin{aligned} \gamma_1 &= \gamma \\ \gamma_2 &= \gamma + \gamma_{21} \\ c_2 &= \alpha_{21} \end{aligned} \quad \text{XIV-128}$$

On y ajoute le calcul d'un étage supplémentaire utilisé seulement pour le calcul de l'erreur :

$$W k_3 = f[t_0 + c_3 h, y_0 + h(\alpha_{31} k_1 + \alpha_{32} k_2)] + hJ(\gamma_{31} k_1 + \gamma_{32} k_2) + h\gamma_3 \left( \frac{\partial f}{\partial t} \right)_{t_0, y_0} \quad \text{XIV-129}$$

avec

$$\begin{aligned} \gamma_3 &= \gamma + \gamma_{31} + \gamma_{32} \\ c_3 &= \alpha_{31} + \alpha_{32} \end{aligned}$$

Les valeurs numériques retenues sont les suivantes :

$$\begin{aligned} \gamma &= \frac{1}{2 + \sqrt{2}} & \gamma_{21} &= -\gamma & c_2 &= \frac{1}{2} & b_1 &= \alpha_{31} = 0 & b_2 &= \alpha_{32} = 1 \\ \gamma_{31} &= 3 - \sqrt{2} & \gamma_{32} &= -5 + 2\sqrt{2} & \gamma_3 &= -1 + \frac{\sqrt{2}}{2} \end{aligned}$$

de telle sorte que la méthode s'écrit

$$W k_1 = f(t_0, y_0) + h\gamma \left( \frac{\partial f}{\partial t} \right)_{t_0, y_0} \quad \text{XIV-130}$$

$$W k_2 = f(t_0 + \frac{h}{2}, y_0 + \frac{h}{2}k_1) - hJ\gamma k_1 \quad \text{XIV-131}$$

$$y_1 = y_0 + hk_2 \quad \text{XIV-132}$$

$$W k_3 = f(t_0 + h, y_0 + hk_2) + hJ[(3 - \sqrt{2})k_1 - (5 - 2\sqrt{2})k_2] - h \left( 1 - \frac{\sqrt{2}}{2} \right) \left( \frac{\partial f}{\partial t} \right)_{t_0, y_0} \quad \text{XIV-133}$$

L'estimation d'erreur utilisée est

$$\text{err} = \frac{h}{6} (k_1 - 2k_2 + k_3) \quad \text{XIV-134}$$

La détermination du domaine de stabilité, dans le cas d'une méthode Rosenbrock, est calculé de la manière suivante : de la relation  $y' = \lambda y$  on déduit  $J = \lambda$ . Reporté dans XIV-130 à 132 cela donne (avec  $z = \lambda h$ )

$$(1 - \gamma z)k_1 = \lambda y_0 \quad \text{XIV-135}$$

$$(1 - \gamma z)k_2 = \lambda \left( y_0 + \frac{h}{2} k_1 \right) - z\gamma k_1 \quad \text{XIV-136}$$

et après calculs

$$k_2 = \lambda y_0 \frac{2 + (1 - 4\gamma)z}{2(1 - \gamma z)^2} \quad \text{XIV-137}$$

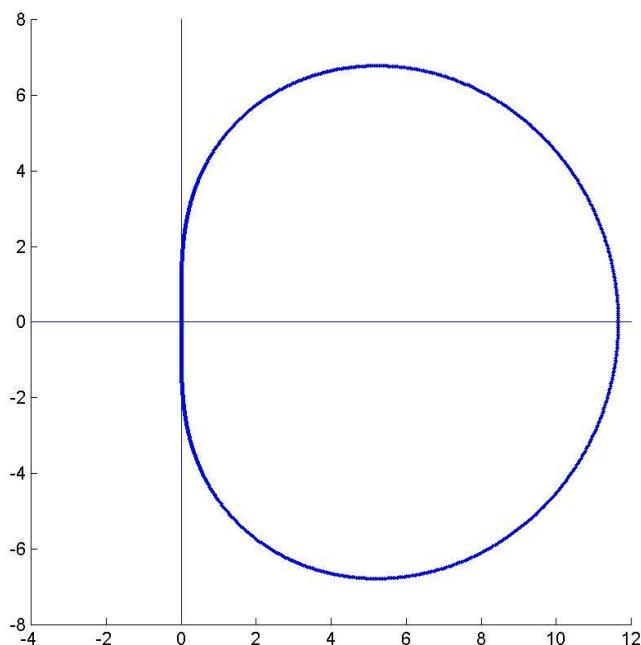
Finalement

$$y_1 = y_0 \frac{z^2(2\gamma^2 + 1 - 4\gamma) + (2 - 4\gamma)z + 2}{2(1 - \gamma z)^2} \quad \text{XIV-138}$$

et en tenant compte de la valeur de  $\gamma$

$$y_1 = y_0 \frac{1 + \frac{\sqrt{2}z}{2 + \sqrt{2}}}{\left(1 - \frac{z}{2 + \sqrt{2}}\right)^2} \quad \text{XIV-139}$$

ce qui donne



La méthode est donc A-stable, et même L-stable au vu de XIV-139. La méthode est donc bien adaptée aux problèmes stiff. Signalons pour terminer que Ode23s est muni d'une procédure d'interpolation permettant d'obtenir la solution à des instants régulièrement espacés.

L'ordre peu élevé de la méthode la rend particulièrement efficace lorsque les exigences de tolérances sont peu élevées.

### Ode23tb

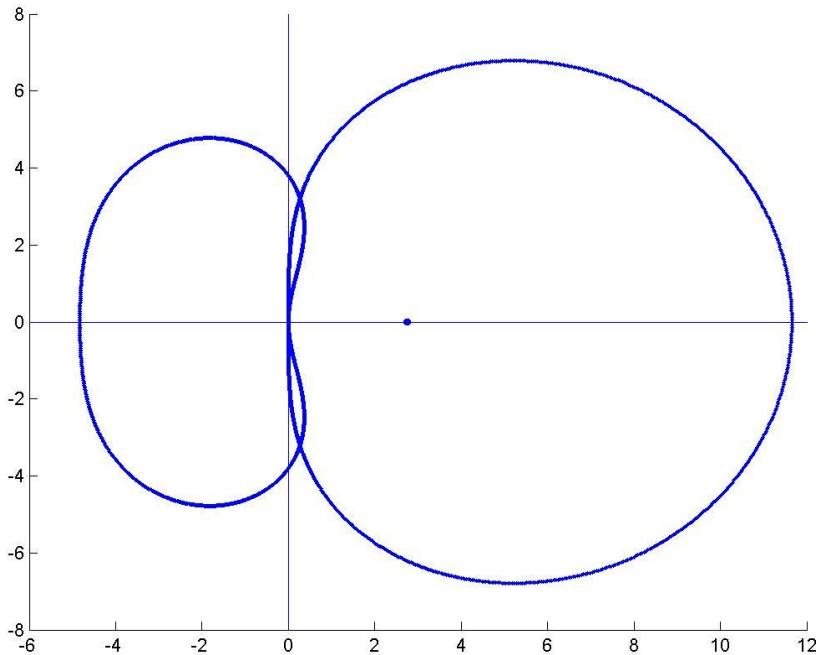
Basé sur la méthode TR-BDF2 (Bank, 1985), cette méthode à un pas peut être interprétée comme une méthode DIRK avec correction du pas d'ordre 2(3). Elle est décrite par le tableau

0	0	0	0
$\gamma$	d	d	0
1	w	w	d
	w	w	d
	$\frac{1-w}{3}$	$\frac{3w+1}{3}$	$\frac{d}{3}$

$$\text{avec } \gamma = 2 - \sqrt{2}, \quad d = \frac{\gamma}{2}, \quad w = \frac{\sqrt{2}}{4}$$

Les propriétés principales de cette méthode sont les suivantes :

Elle est A-stable, comme le montre son domaine de stabilité (formule d'ordre 2 relative à  $b^T = [w \ w \ d]$ ) : domaine situé à droite de l'axe imaginaire :



et L-stable : sa fonction de stabilité vaut

$$F(z) = \frac{1 + (\sqrt{2} - 1)z}{\left(1 + \left(\frac{\sqrt{2}}{2} - 1\right)z\right)^2} \quad \text{et donc} \quad \lim_{z \rightarrow -\infty} F(z) = 0 \quad \text{XIV-140}$$

A ce stade, le point faible de la méthode provient de la formule d'ordre 3 ( $\hat{b}^T = [\frac{1-w}{3} \ \frac{3w+1}{3} \ \frac{d}{3}]$ ) : son domaine de stabilité montre qu'elle n'est pas A-stable (domaine situé à gauche de l'axe imaginaire), ce qui induit une difficulté certaine à élaborer des stratégies efficaces de contrôle du pas quand le problème à traiter est stiff. Les auteurs ont corrigé cette faiblesse en proposant un contrôle du pas basé sur une estimation d'erreur modifiée : plutôt que d'utiliser l'écart classique entre les solutions des deux formules :

$$\text{err} = \hat{y}_{n+1} - y_{n+1}, \quad \text{XIV-141}$$

ils retiennent la grandeur Err solution du système suivant

$$(I - \Delta t dJ)Err = err \quad \text{XIV-142}$$

où J est la matrice jacobienne du problème à résoudre. Ils montrent que Err conserve les propriétés de err aux petites valeurs de  $\Delta t$  tout en améliorant le contrôle du pas pour les systèmes stiff.

Ode23tb est également muni d'une procédure d'interpolation utilisant les étages calculés par la méthode elle-même : à chaque pas, la méthode fournit  $y_{n+\gamma}$  et  $y_{n+1}$  à partir de  $y_n$ , et aussi  $f_{n+\gamma}$  et  $f_{n+1}$ . Ces valeurs sont utilisées pour élaborer un polynôme d'interpolation de Hermite permettant d'évaluer la solution à des instants régulièrement espacés.

Ode23tb convient pour les problèmes stiff lorsque les exigences de tolérance sont peu élevées : ceci provient, comme pour ode23, de l'ordre peu élevé de ses deux formules.

### *Ode23t*

Cet intégrateur propose une implémentation de la méthode des trapèzes. On a vu plus haut que cette méthode est A-stable mais pas L-stable. Tout comme ode15s, ode23t est capable de résoudre des systèmes différentiels-algébriques.

## **Chapitre XV. Problèmes à deux dimensions spatiales. Résolution à l'aide des différences finies**

La formulation générale des problèmes traités dans ce chapitre est la suivante : il s'agit de résoudre

$$u_t = f(u, u_x, u_y, u_{xx}, u_{yy}, \dots, x, y, t) \quad \text{XV-1}$$

où  $u(x, y, t)$  peut être scalaire ou vectoriel. Le domaine d'étude  $\Omega(x, y, t)$  est constitué d'une portion d'un seul tenant  $D$  du plan oxy, et  $t$  varie entre  $t_{\min}$  (en général  $t_{\min} = 0$ ) et  $t_{\max}$ . On supposera que  $D$  est invariant dans le temps, c'est-à-dire que sa frontière  $\Gamma$  ne se modifie pas avec  $t$ .

XV-1 est complété par une condition initiale

$$u(x, y, 0) = u^0(x, y) \quad \forall (x, y) \in D \quad \text{XV-2}$$

et par une ou plusieurs conditions aux limites le long de  $\Gamma$ . Rappelons que ces conditions peuvent être de trois types :

$$\text{Dirichlet : } u(x_\Gamma, y_\Gamma, t) = g_D(t) \quad \text{XV-3}$$

$$\text{Neumann : } \left( \frac{\partial u}{\partial n} \right)(x_\Gamma, y_\Gamma, t) = g_N(t) \quad \text{XV-4}$$

$$\text{Robin : } k_0 u(x_\Gamma, y_\Gamma, t) + k_1 \left( \frac{\partial u}{\partial n} \right)(x_\Gamma, y_\Gamma, t) = g_R(t) \quad \text{XV-5}$$

où  $(x_\Gamma, y_\Gamma)$  désigne un point quelconque de  $\Gamma$ .

Ces conditions valent pour une partie ou l'entièreté de  $\Gamma$  et peuvent donc coexister pour un même problème.

Cette formulation très générale est à préciser au cas par cas. C'est l'objectif des paragraphes qui suivent où l'on se propose d'étudier des exemples de complexité croissante.

### **XV.1 Equation de la chaleur dans un rectangle**

Soit à résoudre

$$T_t = k(T_{xx} + T_{yy}) \quad \text{avec} \quad k = 10 \quad \text{XV-6}$$

$$\text{sur } D = \{(x, y) : 0 \leq x \leq 1 \text{ et } 0 \leq y \leq 2\} \quad \text{avec} \quad 0 \leq t \leq 0.01 \quad \text{XV-7}$$

et avec les conditions suivantes :

$$\text{condition initiale : } T(x, y, 0) = 0 \quad \forall (x, y) \in D \quad \text{XV-8}$$

$$\text{conditions aux limites : } T(x, 0, t) = T(0, y, t) = 0 \quad \text{XV-9}$$

$$T(x, 2, t) = T(1, y, t) = T_M = 100 \quad \text{XV-10}$$

On supposera en outre arbitrairement qu'aux points (1,0) et (0,2) c'est XV-10 qui s'applique.

La résolution sous Matlab de ce problème est structurée de la même manière que pour les problèmes à une dimension spatiale : un problème principal fixe les données à utiliser (maillage, vecteur des conditions initiales, choix des opérateurs de dérivation et de l'intégrateur temporel), procède à l'intégration des équations différentielles ordinaires résultant du passage par la méthode des lignes, et assure la visualisation des résultats. Un sous-programme annexe évalue à tout instant les membres de droites de ces équations, pour le compte de l'intégrateur.

**Programme principal :**

```
close all
clear all
tic
global k TM
global nx ny D2X D2Y
%
%   constantes du problème
%
TM = 100;
k = 10;
%
%   maillage
%
absc = 0:0.005:1;
ordo = 0:0.01:2;
nx = length(absc);
ny = length(ordo);
nv = nx*ny;
%
%   conditions initiales
%
x0(1:nv,1) = zeros(nv,1);
%
%   opérateur de dérivation
%
D2X = five_point_centered_D2(absc);
D2Y = five_point_centered_D2(ordo);
%
%   intégration temporelle
%
time = 0:.0005:.01;
[timeout,xout] = ode45(@eqchal,time,x0);
%
%   visualisation
%
%   C.L.
%
for k = 1:length(timeout)
    xout(k,1:nx) = 0;
    xout(k,1:nx:ny*(ny-1)+1) = 0;
    xout(k,nx:ny:ny*ny) = TM;
    xout(k,ny*(ny-1)+1:ny*ny)= TM;
end

for j = 1:length(time)

    figure
    u(ny:-1:1,:) = reshape(xout(j,1:nv),nx,[])';

```

```

surf(absc,ordo(ny:-1:1),u,'EdgeColor','none');
axis([0 1 0 2 0 100])
camlight right ;% apparition du relief
lighting gouraud ; % lissage
end
toc

```

Commentaires :

1. Le programme débute de manière classique : fermeture des fenêtres ouvertes et effacement de toutes les variables préexistantes, démarrage du chronomètre et déclaration des variables globales utilisées par le sous-programme (voir plus loin) ; fixation des constantes caractéristiques du problème traité et définition du maillage. Le caractère rectangulaire du domaine spatial facilite cette définition : il suffit d'imposer des discrétisations d'axes ox (abscisse) et oy (ordonnée) pour générer automatiquement le maillage total :

```

absc = 0:0.005:1;
ordo = 0:0.01:2;

```

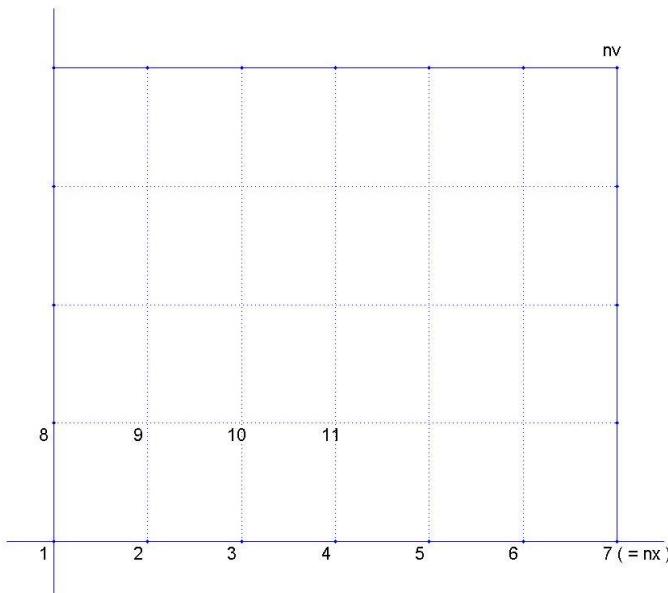
les longueurs nx et ny de absc et ordo permettent ensuite de connaître la taille du vecteur des fonctions inconnues, fixées pour commencer à la condition initiale :

```

nx = length(absc);
ny = length(ordo);
nv = nx*ny;
%
%   conditions initiales
%
x0(1:nv,1) = zeros(nv,1);

```

À ce stade il est important de préciser le système de numérotation adopté pour la transformation de la condition initiale : celle-ci est une matrice de valeurs (en l'occurrence ici la valeur unique zéro) de ny lignes et nx colonnes. Cette matrice doit être transformée en un vecteur (ici le vecteur x0) dont les éléments sont numérotés de 0 à nv. Il a été décidé de respecter la numérotation reprise à la figure suivante :



Cette manière de procéder a plusieurs avantages : la numérotation est naturelle et elle est croissante aussi bien en x qu'en y quand ces coordonnées augmentent ; lors du calcul de dérivées par différences finies, elle facilite aussi le choix de schémas non centrés quand c'est nécessaire : on a vu que ce décentrement est utile

lorsque le problème traité génère des solutions avec fronts mobiles et que, dans ce cas, le sens du déplacement des fronts conditionne le sens du décentrement (schémas upwind par rapport au sens du déplacement). En adoptant la numérotation ci-dessus, le déplacement de la gauche vers la droite correspond à une numérotation croissante, comme dans le cas des problèmes monodimensionnels ; le déplacement vers le haut (sens croissant des  $y$ ) correspond lui aussi à une numérotation croissante (malheureusement mais inévitablement discontinue).

Cette numérotation a néanmoins un inconvénient : on verra plus loin que la visualisation en 3D requiert de transformer le vecteur  $x_0$  en une matrice (baptisée ci-dessous arbitrairement  $X_0$ ) qui restitue à chaque élément  $x_{0_k}$  du vecteur sa position dans le domaine  $D$ . Cette position est naturellement donnée par les indices  $i$  de la ligne et  $j$  de la colonne de  $X_0$  qui contient  $x_{0_k}$  : si l'indice  $j$  épouse strictement la loi de variation de l'indice des points repris dans le vecteur  $absc$ , l'indice  $i$  varie dans le sens opposé à celui du vecteur  $ordo$  :  $i = 1$  correspond à  $ordo(ny)$  tandis que  $i = ny$  correspond à  $ordo(1)$ . Il faudra en tenir compte pour une visualisation correcte.

2. On procède ensuite au choix des schémas de différences finies utilisés :

```
D2X = five_point_centered_D2(absc);
D2Y = five_point_centered_D2(ordo);
```

Observons qu'il n'est pas indispensable d'utiliser le même schéma pour les discrétilisations suivant les deux directions : on aurait pu choisir par exemple

```
D2X = five_point_centered_D2(absc);
D2Y = three_point_centered_D2(ordo);
```

sans que la suite de la programmation ne soit affectée.

3. Viennent ensuite l'intégration temporelle et la visualisation. Cette dernière commence par le rapatriement des conditions aux limites

```
xout(k,1:nx) = 0;
xout(k,1:nx:ny*(ny-1)+1) = 0;
xout(k,nx:ny*nx:ny) = TM;
xout(k,ny*(ny-1)+1:ny*nx)= TM;
```

Observons que ce rapatriement doit tenir compte de la numérotation proposée plus haut.

La visualisation proprement dite utilise ici plusieurs options de confort visuel :

```
for j = 1:length(time)
figure
u(ny:-1:1,:) = reshape(xout(j,1:ny),nx,[])';
surf(absc,ordo(ny:-1:1),u,'EdgeColor','none');
axis([0 1 0 2 0 100])
camlight right;% apparition du relief
lighting gouraud;% lissage
end
```

La représentation 3D propose une figure par instant de visualisation (contrairement à l'habitude prise pour les problèmes monodimensionnels où les solutions aux instants de visualisation sont superposées sur une seule figure) ; l'instruction `reshape` transforme le vecteur `xout` en une matrice où l'indice de ligne est inversé. L'instruction `surf` utilise `absc` et `ordo` (également inversé) pour visualiser la solution. L'option `EdgeColor` et les instructions `camlight` `right` et `lighting` `gouraud` ont pour but d'améliorer la visualisation.

Une visualisation sous forme de courbes isothermes est également possible :

```

for j = 1:length(time)
figure
u(ny:-1:1,:) = reshape(xout(j,1:ny),nx,[])';
[C,h] = contour(absc,ordo(ny:-1:1),u);
set(h,'ShowText','on','TextStep',get(h,'LevelStep')*2)
axis([-1 1 -1 2])
end

```

On trouvera plus loin les figures relatives à ces deux représentations.

**Sous-programme de calcul des membres de droite des équations différentielles :**

```

function xt = eqchal(t,u)
global k TM
global nx ny D2X D2Y
t
%
% C.L.
%
u(1:nx,1) = 0;
u(1:nx:ny-1+1,1) = 0;
u(nx:ny,1) = TM;
u(nx*(ny-1)+1:nx*ny,1) = TM;
%
% dérivées spatiales
%
[uxx uyy] = second_deriv(u,nx,ny,D2X,D2Y);
%
% odes
%
xt = k*(uxx+uyy);

```

Sa structure est identique à ce qu'on a rencontré dans les problèmes monodimensionnels : après avoir affiché le temps  $t$  afin de contrôler le bon déroulement de l'intégration, on introduit les conditions aux limites, en respect du mode de numérotation choisi et on calcule (via le programme `second_deriv` ci-dessous) les dérivées utiles au problème traité. Ces dérivées sont ensuite utilisées pour calculer le vecteur  $u_t$ .

**Sous-programme `second_deriv`:**

```

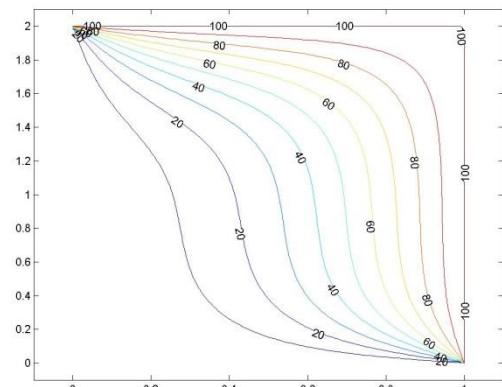
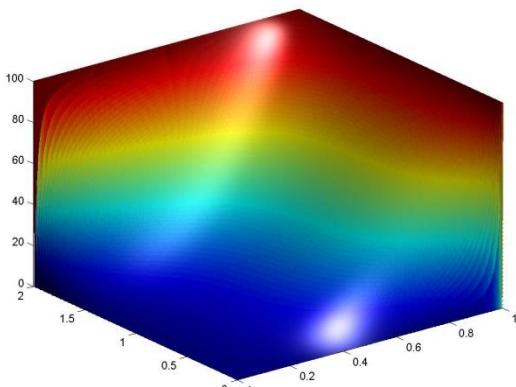
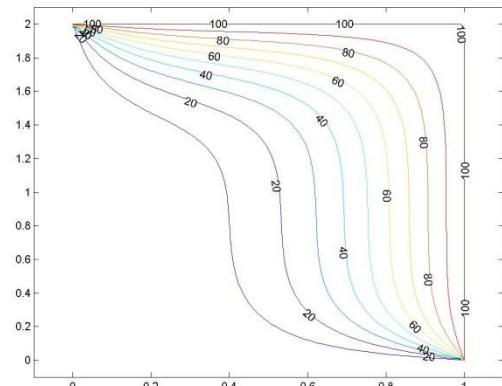
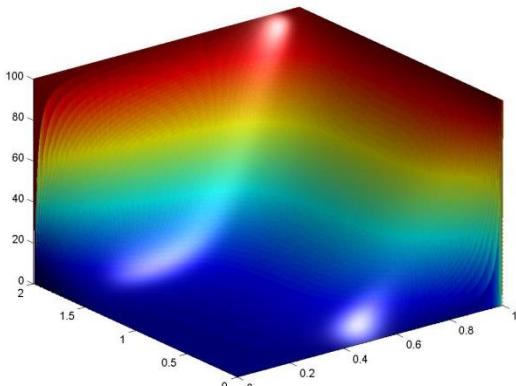
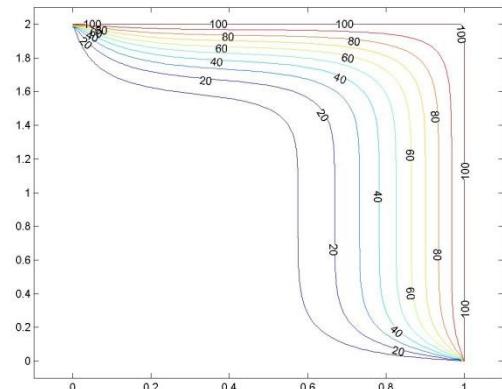
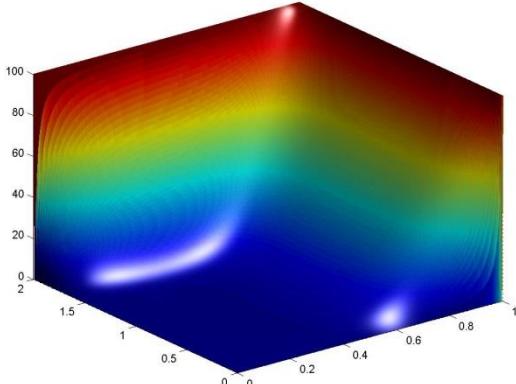
function [uxx uyy] = second_deriv(u,nx,ny,D2X,D2Y)
v = reshape(u,nx,[]);
uxx = reshape(D2X*v,nx*ny,1);
uyy = reshape(v*D2Y',nx*ny,1);

```

ce programme effectue les tâches suivantes :

- mise sous forme matricielle  $v$  du vecteur  $u$  à dériver,
- calcul du produit  $D2X \cdot v$  qui fournit la matrice de  $u_{xx}$  en tous les points de  $D$ ,
- remise sous forme vectorielle de  $u_{xx}$
- traitement identique pour le calcul du vecteur  $u_{yy}$ , les opérations de transposition et de produit matriciel étant agencées de manière à minimiser la quantité d'opérations à effectuer.

La figure ci-dessous reprend les résultats de simulation à 3 instants :  $t = \frac{t_{\max}}{3}, \frac{2t_{\max}}{3}, t_{\max}$  en 3D et sous forme d'isothermes :



## XV.2 Problème de Graetz avec température de paroi constante

L'exemple qui précède a permis d'illustrer la manière avec laquelle les programmes de calcul de dérivées spatiales développés initialement pour des problèmes à une dimension spatiale sont facilement utilisable dans un espace de dimension deux. L'exemple qui suit montre comment utiliser les mêmes sous-programmes lorsque le système d'axes coordonnés n'est plus le repère cartésien classique.

Les équations (en variables réduites) suivantes modélisent l'écoulement laminaire d'un fluide newtonien dans un tube cylindrique dont la paroi externe est maintenue à température constante :

$$T_t = -P_e v(r) T_z + T_{zz} + \frac{1}{r} T_r + T_{rr} \quad \text{XV-11}$$

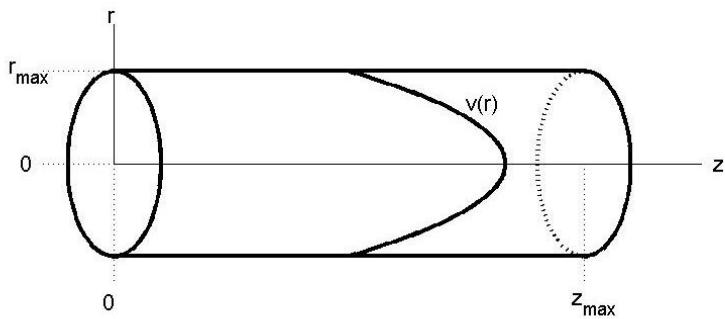
avec  $P_e$  = nombre de Peclet = 60

$v(r) = 2(1 - r^2)$  = vitesse de déplacement du fluide : la vitesse présente un profil parabolique dans la section du tube

$$0 \leq t \leq t_{\max} = 0.5$$

$$0 \leq r \leq r_{\max} = 1$$

$$0 \leq z \leq z_{\max} = 30$$



et avec les conditions suivantes :

$$\text{condition initiale : } T(r, z, 0) = 0 \quad \text{XV-12}$$

$$\text{conditions aux limites : } T_r(0, z, t) = 0 \quad \text{XV-13}$$

$$T(r_{\max}, z, t) = 1 \quad \text{XV-14}$$

$$T(r, 0, t) = 0 \quad \text{XV-15}$$

$$T_z(r, z_{\max}, t) = 0 \quad \text{XV-16}$$

D'emblée une remarque s'impose : le long de l'axe horizontal ( $r = 0$ ), la condition aux limites XV-13 exprime la symétrie centrale du problème traité ; c'est cette symétrie qui limite d'ailleurs le nombre de variables spatiales à deux : la coordonnée radiale  $r$  et la coordonnée longitudinale  $z$ . Il en résulte aussi qu'en ces points, le terme  $\frac{1}{r} T_r$  de XV-11 est indéterminé : il vaut  $\frac{0}{0}$ . Le recours à la règle de l'Hospital lève cette indétermination :

$$\lim_{r \rightarrow 0} \frac{1}{r} T_r = \lim_{r \rightarrow 0} T_{rr} \quad \text{XV-17}$$

Il en résulte que le long de l'axe  $z$ , l'équation à simuler n'est plus XV-11 mais

$$T_t = -P_e v(r) T_z + T_{zz} + 2T_{rr}$$

XV-18

Détaillons maintenant les programmes de simulation.

**Programme principal :**

```

close all
clear all
tic
global Pe v r_inv nv nabsc nrad absc rad dr D1x D1r D2x D2r
%
% grille spatiale
%
xmin = 0;
xmax = 30;
nabsc = 101;
dx = (xmax-xmin)/(nabsc-1);
absc = xmin:dx:xmax;

rmin = 0;
rmax = 1;
nrad = 101;
dr = (rmax-rmin)/(nrad-1);
rad = rmin:dr:rmax;
%
% constantes du problème
%
Pe = 60;
v = zeros(nabsc*nrad,1);
r_inv = zeros(nabsc*nrad,1);
for i = 1:nrad
    if i == 1
        v((i-1)*nabsc+1:i*nabsc,1) = 2*(1-rad(i)^2);
    else
        v((i-1)*nabsc+1:i*nabsc,1) = 2*(1-rad(i)^2);
        r_inv((i-1)*nabsc+1:i*nabsc,1) = 1/rad(i);
    end
end
%
% conditions initiales
%
x0 = zeros(nabsc*nrad,1);
nv = length(x0);
%
% opérateur de dérivation
%
D1x = four_point_biased_upwind_D1(absc,1);
D1r = five_point_centered_D1(rad);
D2x = five_point_centered_D2(absc);
D2r = five_point_centered_D2(rad);
%
% intégration temporelle
%
time = 0:.05:.5;
[timeout,xout] = ode45(@Graetz_2,time,x0);
%
% visualisation
%
% C.L.

```

```

%
for k = 1:length(timeout)
    xout(k,1:nabsc) = xout(k,nabsc+1:2*nabsc);
    xout(k,nv-nabsc+1:nv) = 1;
    xout(k,1:nabsc:nv-nabsc+1) = 0;
    xout(k,nabsc:nabsc:nv) = xout(k,nabsc-1:nabsc:nv-1);
end
[X Y] = meshgrid(absc,rad(nrad:-1:1));
[X1 Y1] = meshgrid(absc,-rad);

for j = 1:length(timeout)

figure

u(nrad:-1:1,:) = reshape(xout(j,1:nv),nabsc,[])';
surf(X,Y,u,'EdgeColor','none')
hold
surf(X1,Y1,u(nrad:-1:1,:),'EdgeColor','none')
axis([xmin xmax -rmax rmax 0 1])
view(-60,16)
camlight right; % apparition du relief
lighting gouraud; % lissage
pause
end
toc

```

Sa structure est analogue à celle de l'exemple précédent. Les points importants à signaler sont les suivants :

1° le caractère orthogonal des coordonnées radiale et longitudinale permet de travailler comme si on était en maillage cartésien, les variables spatiales correspondantes portant ici les noms `absc` et `rad` et la numérotation des nœuds du maillage obéit à la même règle que dans l'exemple précédent.

2° la vitesse  $v(r)$  et la grandeur  $\frac{1}{r}$  sont créées en tous les nœuds du maillage où on en aura besoin (en tous les nœuds pour  $v(r)$  et en tous les nœuds sauf sur l'axe horizontal pour  $\frac{1}{r}$ ).

3° les schémas de différences finies sont sélectionnés en fonction de la nature des phénomènes que les termes de dérivées spatiales représentent dans XV-11 (ou XV-18).

4° après intégration temporelle et rapatriement des conditions aux limites a lieu la visualisation. On y tient compte de la symétrie centrale du problème en élargissant l'axe radial aux valeurs  $[-1, 1]$  : c'est la raison des instructions

```
[X1 Y1] = meshgrid(absc,-rad);
et      surf(X1,Y1,u(nrad:-1:1,:),'EdgeColor','none')
```

**Sous-programme de calcul des membres de droite des équations différentielles :**

```

function xt = Graetz_2(t,u)
global Pe v r_inv nv nabsc nrad dr D1x D1r D2x D2r
t
%
%   C.L.
%
u(1:nabsc,1) = u(nabsc+1:2*nabsc,1);
```

```

u(nv-nabsc+1:nv,1) = 1;
u(1:nabsc:nv-nabsc+1,1) = 0;
u(nabsc:nabsc:nv,1) = u(nabsc-1:nabsc:nv-1,1);
%
%   dérivées spatiales
%
[Tx Tr] = first_deriv(u,nabsc,nrad,D1x,D1r);
[Txx Trr] = second_deriv(u,nabsc,nrad,D2x,D2r);
%
%   odes
%
xt = -Pe*v.*Tx + Txx + Trr;
xt(nabsc+1:nabsc*nrad,1) = xt(nabsc+1:nabsc*nrad,1) +
r_inv(nabsc+1:nabsc*nrad,1).*Tr(nabsc+1:nabsc*nrad,1);
xt(1:nabsc,1) = xt(1:nabsc,1) + Trr(1:nabsc,1);

```

Les points importants à signaler sont les suivants :

1° l'implémentation des conditions aux limites XV-13 et XV-16, de type Neumann, est proposée en les « transformant » en conditions de type Dirichlet grâce au recours à des schémas de différences finies simples : schéma à deux points décentré vers l'intérieur. On notera à nouveau la nécessité de bien tenir compte de la numérotation des nœuds du maillage pour une implémentation correcte de ces conditions.

2° les dérivées premières et secondes spatiales sont alors calculées :

```
[Tx Tr] = first_deriv(u,nabsc,nrad,D1x,D1r);
[Txx Trr] = second_deriv(u,nabsc,nrad,D2x,D2r);
```

`second_deriv` a déjà été présenté dans l'exemple précédent

`first_deriv` (voir ci-après) a une structure identique à `second_deriv` et peut aussi utiliser des schémas de différences finies différents selon les axes longitudinal et radial.

3° les membres de droites des équations différentielles sont alors programmés : on implémente d'abord les termes communs à XV-11 et XV-18 :

```
xt = -Pe*v.*Tx + Txx + Trr;
```

pour ensuite ajouter la quantité manquante :

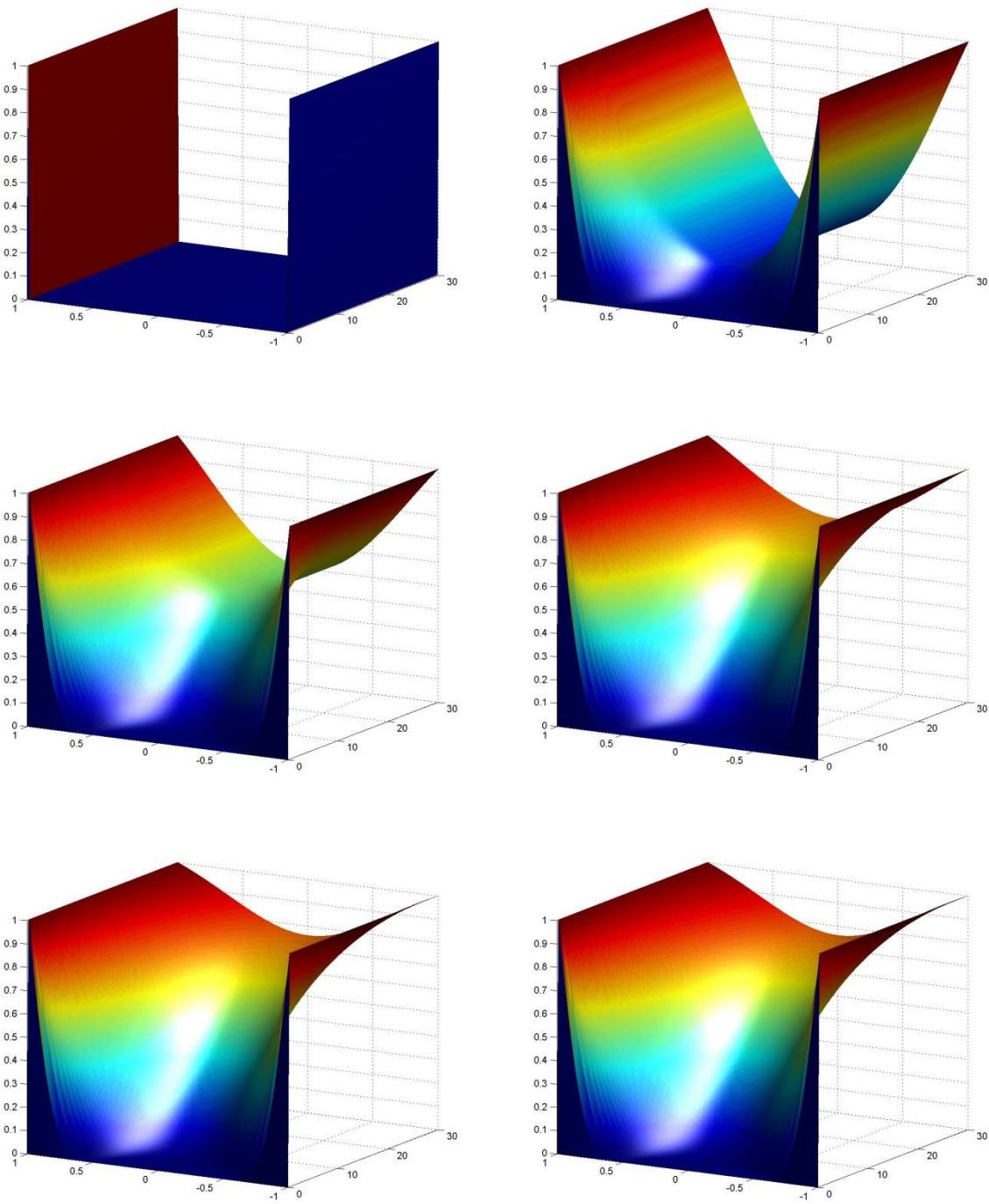
```
xt(nabsc+1:nabsc*nrad,1) = xt(nabsc+1:nabsc*nrad,1) +
r_inv(nabsc+1:nabsc*nrad,1).*Tr(nabsc+1:nabsc*nrad,1);

xt(1:nabsc,1) = xt(1:nabsc,1) + Trr(1:nabsc,1);
```

**Sous-programme `first_deriv`:**

```
function [ux uy] = first_deriv(u,nx,ny,D1X,D1Y)
v = reshape(u,nx,[]);
ux = reshape(D1X*v,nx*ny,1);
uy = reshape(v*D1Y',nx*ny,1);
```

les figures ci-après proposent les solutions obtenues aux instants  $t_k = k\Delta t$   $k : 0, \dots, 5$   $\Delta t = 0.1$



### XV.3 Equation de la chaleur dans un quadrilatère convexe

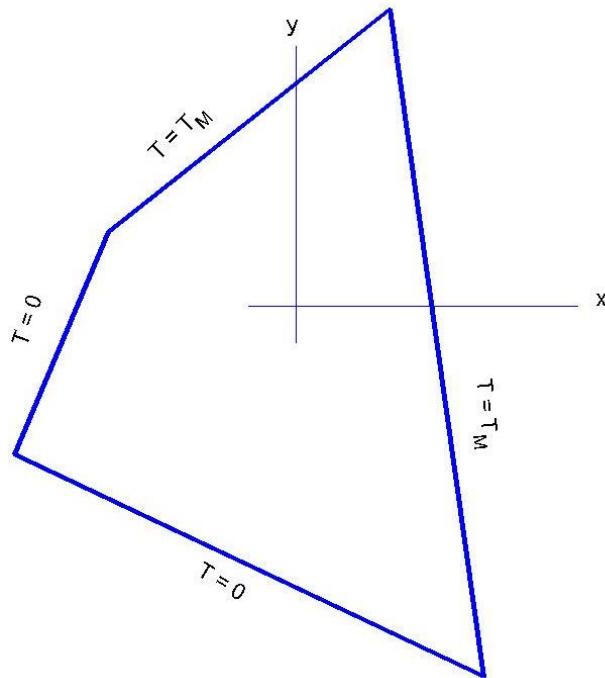
Les deux exemples précédents réunissent des conditions de confort de simulation liées à la géométrie du domaine spatial D. Le but du présent paragraphe est de fournir les outils permettant de traiter des problèmes définis sur tout quadrilatère convexe.

On se propose donc de résoudre le problème XV-6

$$T_t = k(T_{xx} + T_{yy}) \quad \text{avec} \quad k = 10$$

XV-6

sur le quadrilatère D de sommets  $(-3 -2), (2 -5), (1 4), (-2 1)$  :



avec les conditions suivantes :

$$\text{condition initiale : } T(x, y, 0) = 0 \quad \forall (x, y) \in D \quad \text{XV-19}$$

$$\text{conditions aux limites : } T = 0 \text{ le long de deux côtés adjacents} \quad \text{XV-20}$$

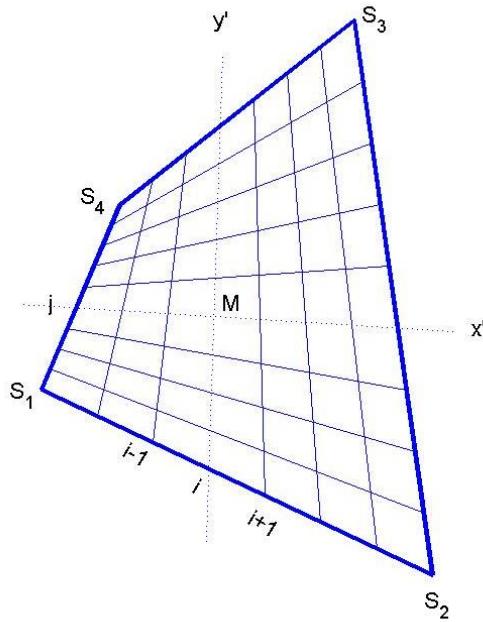
$$T = T_M = 100 \text{ le long des deux autres côtés} \quad \text{XV-21}$$

La première étape de la résolution d'un tel problème est la définition du maillage spatial. D'emblée il apparaît que le recours à un maillage rectangulaire classique va générer des difficultés le long des bords de D : il est impossible avec de telles géométries de faire coïncider l'entièreté des bords avec des lignes du maillage. Celui-ci ne sera donc plus rectangulaire mais sera constitué de deux familles de droites telles que dessinées à la figure suivante.

Avec un tel maillage, il est immédiat de pouvoir estimer les dérivées spatiales en un point M quelconque du maillage selon les directions des droites qui passent par M : par exemple, selon la direction  $x'$  on a

$$T_{x'x'} = \frac{T_{i-1,j} - 2T_{i,j} + T_{i+1,j}}{\Delta x'^2}$$

si  $\Delta x'$  est le pas de discréétisation (constant) le long de  $x'$  et si le schéma retenu est à trois points centrés. Les formules de différences finies classiques vont donc fournir à volonté des estimations de  $T_{x'x'}, T_{y'y'}, T_{x'x}, T_{y'}$  et même  $T_{x'y'}$  en tous les nœuds du maillage à partir desquelles il faudra calculer les dérivées utiles ( $T_{xx}$  et  $T_{yy}$  ici).



Ce calcul peut-être systématisé pour tous les points de la grille en ayant recours à un domaine de référence  $D_{\text{réf}}$  défini dans un système de coordonnées réduites  $\xi$  et  $\eta$  :

$$D_{\text{réf}} = \{(\xi, \eta) : 0 \leq \xi \leq 1, 0 \leq \eta \leq 1\}.$$

Tout point de coordonnées  $(x, y)$  de  $D$  sera associé à un point  $(\xi, \eta)$  de  $D_{\text{réf}}$  via les relations qui transforment  $D$  en  $D_{\text{réf}}$ . Ces relations établissent les liens suivants entre les deux systèmes de coordonnées :

$$\begin{aligned} x &= x(\xi, \eta) & y &= y(\xi, \eta) \\ \xi &= \xi(x, y) & \eta &= \eta(x, y) \end{aligned} \tag{XV-22}$$

Elles sont obtenues en identifiant chacun des sommets de  $D$  à un sommet de  $D_{\text{réf}}$ . On convient arbitrairement d'associer au sommet  $(0,0)$  de  $D_{\text{réf}}$  le sommet de  $D$  dont l'abscisse est la plus petite ( $S_1$  sur la figure) ; si deux sommets de  $D$  peuvent prétendre à jouer le rôle de  $S_1$ , on retient celui dont l'ordonnée est la plus petite. Les autres sommets sont alors numérotés en parcourant le périmètre de  $D$  dans le sens trigonométrique.

### ***Calcul du changement de variables***

Soit  $(x_i, y_i)$  les coordonnées du sommet  $S_i$ ,  $i = 1, \dots, 4$ . Le changement de variables est opéré grâce aux relations

$$\begin{aligned} x &= a_0 + a_1\xi + a_2\eta + a_3\xi\eta \\ y &= b_0 + b_1\xi + b_2\eta + b_3\xi\eta \end{aligned} \tag{XV-23}$$

Sachant que les sommets de  $D$  et de  $D_{\text{réf}}$  se correspondent selon

$$\begin{aligned} S_1 &\leftrightarrow (0, 0) \\ S_2 &\leftrightarrow (1, 0) \\ S_3 &\leftrightarrow (1, 1) \\ S_4 &\leftrightarrow (0, 1) \end{aligned}$$

il vient facilement

$$\begin{aligned} a_0 &= x_1 & a_1 &= x_2 - x_1 & a_2 &= x_4 - x_1 & a_3 &= (x_3 + x_1) - (x_4 + x_2) \\ b_0 &= y_1 & b_1 &= y_2 - y_1 & b_2 &= y_4 - y_1 & b_3 &= (y_3 + y_1) - (y_4 + y_2) \end{aligned} \quad \text{XV-24}$$

### **Calcul des dérivées**

Le changement de variables XV-23 modifie l'équation à résoudre : l'inconnue  $u(x, y)$  est remplacée par  $u(\xi, \eta) = u(\xi(x, y), \eta(x, y))$ . Il en résulte les remplacements suivants de dérivées dans l'équation (en se limitant aux dérivées premières et secondes)

dérivées premières :

$$\begin{aligned} u_x &= u_\xi \xi_x + u_\eta \eta_x \\ u_y &= u_\xi \xi_y + u_\eta \eta_y \end{aligned} \quad \text{XV-25}$$

dérivées secondes : après quelques calculs :

$$\begin{aligned} u_{xx} &= u_{\xi\xi} (\xi_x)^2 + u_{\eta\eta} (\eta_x)^2 + u_{\xi\eta} (2\xi_x \eta_x) + u_\xi (\xi_{xx}) + u_\eta (\eta_{xx}) \\ u_{yy} &= u_{\xi\xi} (\xi_y)^2 + u_{\eta\eta} (\eta_y)^2 + u_{\xi\eta} (2\xi_y \eta_y) + u_\xi (\xi_{yy}) + u_\eta (\eta_{yy}) \end{aligned} \quad \text{XV-26}$$

On a donc besoin de  $\xi_x, \xi_y, \xi_{xx}, \xi_{yy}$  et de  $\eta_x, \eta_y, \eta_{xx}, \eta_{yy}$ . Se rappelant que le changement de variables XV-23 a déjà été utilisé en éléments finis quadratiques, on dispose déjà des relations suivantes : posant

$$Den = x_\xi y_\eta - x_\eta y_\xi \quad \text{XV-27}$$

il vient

$$\xi_x = \frac{y_\eta}{Den} \quad \xi_y = -\frac{x_\eta}{Den} \quad \eta_x = -\frac{y_\xi}{Den} \quad \eta_y = \frac{x_\xi}{Den} \quad \text{XV-28}$$

Et grâce à XV-23 :

$$\begin{aligned} Den &= (a_1 b_2 - a_2 b_1) + (a_1 b_3 - a_3 b_1) \xi + (a_3 b_2 - a_2 b_3) \eta \\ y_\eta &= b_2 + b_3 \xi \\ x_\eta &= a_2 + a_3 \xi \\ y_\xi &= b_1 + b_3 \eta \\ x_\xi &= a_1 + a_3 \eta \end{aligned} \quad \text{XV-29}$$

Le calcul des dérivées secondes, un peu plus long, découle de relations semblables par exemple à

$$\xi_{xx} = \frac{\partial}{\partial x} \left( \frac{y_\eta}{Den} \right); \text{ tous calculs faits, on trouve}$$

$$\begin{aligned}\xi_{2x} &= \frac{2y_\eta y_\xi}{\text{Den}^3} (a_3 b_2 - a_2 b_3) \\ \xi_{2y} &= \frac{2x_\eta x_\xi}{\text{Den}^3} (a_3 b_2 - a_2 b_3) \\ \eta_{2x} &= \frac{2y_\eta y_\xi}{\text{Den}^3} (a_1 b_3 - a_3 b_1) \\ \eta_{2y} &= \frac{2x_\eta x_\xi}{\text{Den}^3} (a_1 b_3 - a_3 b_1)\end{aligned}\tag{XV-30}$$

A ce stade, on dispose de toutes les relations nécessaires : les étapes utiles à la résolution de

$$T_t = k(T_{xx} + T_{yy})$$

sur D sont les suivantes :

1° compte tenu de la géométrie de D, on calcule une fois pour toutes, pour tous les nœuds du maillage de D,  $\xi_x, \xi_y, \xi_{xx}, \xi_{yy}$  et  $\eta_x, \eta_y, \eta_{xx}, \eta_{yy}$  grâce à XV-28, XV-29 et XV-30.

2° on calcule  $T_{xx}$  et  $T_{yy}$  par XV-26 en y remplaçant  $T_{\xi\xi}, T_{\eta\eta}, T_\xi, T_\eta, T_{\xi\eta}$  par des formules de différences finies classiques.

### **Programme principal**

```
close all
clear all
tic
global k TM
global nv nksi neta absc_ksi ordo_eta D1_ksi D1_eta D2_ksi D2_eta ksi_x ksi_y
eta_x eta_y ksi_xx ksi_yy eta_xx eta_yy
%
% constantes du problème
%
TM = 100;
k = 10;
sommets = [ -3 -2 ; 2 -5 ; 1 4 ; -2 1 ];
%
% définition des variables réduites ksi et eta : carré [0 1]x[0 1] :
% les maillages sur ces axes sont absc_ksi et ordo_eta
%
nksi = 51;
neta = 71;
nv = nksi*neta;
dksi = 1/(nksi-1);
absc_ksi = (0:dksi:1)';
deta = 1/(neta-1);
ordo_eta = (0:deta:1)';
%
% grilles spatiales et coordonnées des points du maillage : sont définies
% à partir des sommets et des nombres de points placés sur les côtés
% S1-S2 (nksi points en tout) et S1-S4 (neta points en tout)
%
% signification des variables :
```

```

%
%      absc et ordo contiennent les coordonnées des points du domaine D :
%      utilisés pour la visualisation
%
%      ksi_x, ... , eta_yy : dérivée 1e de ksi par rapport à x, ... ,
%      dérivée 2e de eta par rapport à y
[absc ordo ksi_x ksi_y eta_x eta_y ksi_xx ksi_yy eta_xx eta_yy] =
grille_D1_D2(sommets,nksi,neta,absc_ksi,ordo_eta);
%
%      conditions initiales
%
x0(1:nv,1) = zeros(nv,1);
%
%      schéma de différences finies
%
D1_ksi = three_point_centered_D1(absc_ksi);
D1_eta = three_point_centered_D1(ordo_eta);
D2_ksi = five_point_centered_D2(absc_ksi);
D2_eta = five_point_centered_D2(ordo_eta);
%
%      intégration temporelle
%
time = 0:.0005:.01;
[timeout,xout] = ode45(@eqchal_quadri,time,x0);
%
%      visualisation
%
%
%      C.L.
%
for k = 1:length(timeout)
xout(k,1:nksi) = 0;
xout(k,1:nksi:nksi*(neta-1)+1) = 0;
xout(k,nksi:nksi:nksi*neta) = TM;
xout(k,nksi*(neta-1)+1:nksi*neta)= TM;
end
for j = 1:length(time)
figure
u(neta:-1:1,:) = reshape(xout(j,1:nv),nksi,[])';
surf(absc,ordo,u,'FaceColor',[1 0 0],'EdgeColor','none')
view([25 32]);
camlight right ;% apparition du relief
lighting gouraud ; % lissage
end
toc

```

A nouveau la structure de ce programme suit celle des exemples précédents. Les points principaux sont :

1° après la fixation des constantes du problème et des coordonnées des sommets du quadrilatère, on définit le maillage sur  $D_{\text{réf}}$  ;

2° l'appel au sous-programme `grille_D1_D2` (voir plus loin) fournit les matrices `absc` et `ordo` contenant les coordonnées des points du maillage de  $D$  et utilisées dans la visualisation. `grille_D1_D2` fournit aussi les dérivées  $\xi_x, \xi_y, \xi_{xx}, \xi_{yy}$  et  $\eta_x, \eta_y, \eta_{xx}, \eta_{yy}$  passées en variables globales et utilisées dans le calcul des membres de droite des équations différentielles ordinaires.

3° après fixation de la condition initiale et sélection des schémas de différences finies utilisés ont lieu l'intégration temporelle, le rapatriement des conditions aux limites et la visualisation.

### **Sous-programme de calcul des membres de droite des équations différentielles :**

```

function xt = eqchal_quadri(t,u)
global k TM
global nksi neta D1_ksi D1_eta D2_ksi D2_eta ksi_x ksi_y eta_x eta_y ksi_xx
ksi_yy eta_xx eta_yy
t
%
% C.L.
%
u(1:nksi,1) = 0;
u(1:nksi:nksi*(neta-1)+1,1) = 0;
u(nksi:nksi*neta,1) = TM;
u(nksi*(neta-1)+1:nksi*neta,1) = TM;
%
% dérivées spatiales
%
[uksi ueta] = first_deriv(u,nksi,neta,D1_ksi,D1_eta);
[uksiksi uetaeta] = second_deriv(u,nksi,neta,D2_ksi,D2_eta);
uksieta = second_mixed_deriv(u,nksi,neta,D1_ksi,D1_eta);
uxx = uksiksi.* (ksi_x.^2) + 2*uksieta.*ksi_x.*eta_x + uetaeta.* (eta_x.^2) +
uksi.*ksi_xx + ueta.*eta_xx;
uyy = uksiksi.* (ksi_y.^2) + 2*uksieta.*ksi_y.*eta_y + uetaeta.* (eta_y.^2) +
uksi.*ksi_yy + ueta.*eta_yy;
%
% odes
%
xt = k*(uxx+uyy);

```

Après fixation des conditions aux limites, on calcule  $T_\xi, T_\eta, \dots, T_{\eta\eta}$  sur  $D_{\text{réf}}$  grâce aux sous-programmes déjà utilisés (`first_deriv` et `second_deriv`).  $T_{\xi\eta}$  requiert la création d'un nouveau sous-programme : `second_mixed_deriv` (voir plus loin). Viennent ensuite le calcul de  $T_{xx}$  et  $T_{yy}$  par XV-26.

### **Sous-programme grille\_D1\_D2**

```

function [absc ordo ksi_x ksi_y eta_x eta_y ksi_xx ksi_yy eta_xx eta_yy] =
grille_D1_D2(sommets,nksi,neta,ksi,eta)
%
% programme de calcul des abscisses et ordonnées des points du maillage
% et des opérateurs de dérivation permettant le changement de variables
% (x,y) du domaine initial vers les variables (ksi,eta).
%
% input : sommets : matrice (4 x 2) contenant en 1e colonne les abscisses
% et en 2e colonne les ordonnées des 4 sommets du quadrilatère.
%
% nksi et neta : nombre de points de grille placés sur l'axe
% S1-S2 (voir figure) sommets inclus et sur l'axe S1-S4 sommets
% inclus.
%
% la numérotation des sommets sera réalisée par le programme de la manière
% suivante : S1 est le plus à gauche. Si 2 sommets peuvent prétendre à
% jouer le rôle de S1, on retient celui dont l'ordonnée est la plus
% petite. Les autres sont ordonnés d'après la pente de la droite joignant
% chacun d'eux à S1. Compte tenu du choix fait pour S1, cette pente
% correspond forcément à un angle compris entre -pi/2 et +pi/2.

```

```

% Par exemple :

%
%
% S4 o
%
%
% S1 o
%
%
% S3 o
%
%
% S2 o
%
%
ksi et eta : vecteurs de dimensions nksi et neta contenant les
points de grilles choisis sur l'axe ksi et l'axe eta

%
output : absc et ordo : matrices (neta x nksi) contenant les abscisses
et les ordonnées des nœuds du maillage ; ces matrices servent
à la visualisation de la solution. On notera que la
numérotation des nœuds est la suivante : on imagine
que S1-S2 est l'axe ox et S1-S4 l'axe oy. Cela étant, la
numérotation se fait de gauche à droite et de bas vers le haut
en démarrant en S1, et en finissant en S3.

%
ksi_x, ksi_y, eta_x et eta_y : vecteurs de dimension nksi*neta
donnant pour chaque point de la grille les opérateurs de
changement de système de coordonnées :

%
d(ksi)   d(ksi)   d(eta)   d(eta)
----- , ----- , ----- , ----- [1]
dx       dy       dx       dy

%
ksi_xx, ksi_yy, eta_xx et eta_yy : vecteurs de dimension nksi*neta
donnant pour chaque point de la grille les opérateurs de
changement de système de coordonnées :

%
d2(ksi)   d2(ksi)   d2(eta)   d2(eta)
----- , ----- , ----- , ----- [2]
dx2      dy2      dx2      dy2

%
xmin = min(sommets(:,1));
ind = find(sommets(:,1) == xmin);
if length(ind) == 2
    Sselect = sommets(ind,:);
    ymin = min(Sselect(:,2));
    indbis = find(Sselect(:,2) == ymin);
    S1 = Sselect(indbis,:);
else
    S1 = sommets(ind,:);
end
S = setdiff(sommets,S1,'rows');
test = sortrows([(S(:,2)-S1(2))./(S(:,1)-S1(1)) S]);
S2 = test(1,2:3);
S3 = test(2,2:3);
S4 = test(3,2:3);
%
% génération d'une table des coordonnées (x,y) des points du maillage :
%
% les coordonnées sont stockées dans coord(nksi*neta,2) chaque ligne
% est constituée des données suivantes :
%
% absc x   ordo y

```

```

%
coord = zeros(nksi*neta,2);
for k = 0:neta-1
    coordg = S1 + k*(S4-S1)/(neta-1);
    coordd = S2 + k*(S3-S2)/(neta-1);
    for j = 0:nksi-1
        coord(1+k*nksi+j,1:2) = coordg + j*(coordd-coordg)/(nksi-1);
    end
end
%
% absc et ordo sont des matrices contenant les coordonnées des points du
% maillage
%
absc = zeros(neta,nksi);
ordo = zeros(neta,nksi);
for i = 1:neta
    absc(neta+1-i,1:nksi) = coord((i-1)*nksi+1:i*nksi,1);
    ordo(neta+1-i,1:nksi) = coord((i-1)*nksi+1:i*nksi,2);
end
%
% le changement de système de coordonnées s'écrit
%
%     x = a(0) + a(1)*ksi + a(2)*eta + a(3)*ksi*eta
%
%     y = b(0) + b(1)*ksi + b(2)*eta + b(3)*ksi*eta
%
% dans ces relations, les a(i) et b(i) dépendent des coordonnées des
% sommets ; a(0) et b(0) ne sont pas calculés car ne sont pas utilisés par
% la suite ; a(1), a(2), a(3), b(1), b(2) et b(3) interviennent dans le
% calcul des dérivées [1]
%
a = zeros(1,3);
b = zeros(1,3);
a(1) = S2(1) - S1(1);
a(2) = S4(1) - S1(1);
a(3) = S3(1) + S1(1) - S2(1) - S4(1);
b(1) = S2(2) - S1(2);
b(2) = S4(2) - S1(2);
b(3) = S3(2) + S1(2) - S2(2) - S4(2);
denom = zeros(nksi*neta,1);
ksi_x = zeros(nksi*neta,1);
ksi_y = zeros(nksi*neta,1);
eta_x = zeros(nksi*neta,1);
eta_y = zeros(nksi*neta,1);
d0 = a(1)*b(2)-a(2)*b(1);
d1 = a(1)*b(3)-a(3)*b(1);
d2 = a(3)*b(2)-a(2)*b(3);
for i = 1:neta
    denom((i-1)*nksi+1:(i-1)*nksi+nksi,1) = d0 + d1*ksi + d2*eta(i);
    ksi_x((i-1)*nksi+1:(i-1)*nksi+nksi,1) = (b(2)+b(3)*ksi)./denom((i-
1)*nksi+1:(i-1)*nksi+nksi,1);
    ksi_y((i-1)*nksi+1:(i-1)*nksi+nksi,1) = -(a(2)+a(3)*ksi)./denom((i-
1)*nksi+1:(i-1)*nksi+nksi,1);
    ksi_xx((i-1)*nksi+1:(i-1)*nksi+nksi,1) =
2*(b(2)+b(3)*ksi)*(b(1)+b(3)*eta(i))*(a(3)*b(2)-a(2)*b(3))./
(denom((i-
1)*nksi+1:(i-1)*nksi+nksi,1).^3);
    ksi_yy((i-1)*nksi+1:(i-1)*nksi+nksi,1) =
2*(a(2)+a(3)*ksi)*(a(1)+a(3)*eta(i))*(a(3)*b(2)-a(2)*b(3))./
(denom((i-
1)*nksi+1:(i-1)*nksi+nksi,1).^3);
end
for i = 1:nksi

```

```

eta_x(i:nksi:(neta-1)*nksi+i,1) = -(b(1)+b(3)*eta)./denom(i:nksi:(neta-
1)*nksi+i,1);
eta_y(i:nksi:(neta-1)*nksi+i,1) = (a(1)+a(3)*eta)./denom(i:nksi:(neta-
1)*nksi+i,1);
eta_xx(i:nksi:(neta-1)*nksi+i,1) =
2*(b(1)+b(3)*eta)*(b(2)+b(3)*ksi(i))*(a(1)*b(3)-a(3)*b(1))./(denom(i:nksi:(neta-
1)*nksi+i,1).^3);
eta_yy(i:nksi:(neta-1)*nksi+i,1) =
2*(a(1)+a(3)*eta)*(a(2)+a(3)*ksi(i))*(a(1)*b(3)-a(3)*b(1))./(denom(i:nksi:(neta-
1)*nksi+i,1).^3);
end

```

Ce programme effectue les tâches suivantes :

- classement et numérotation des sommets du quadrilatère
- calcul des matrices `absc` et `ordo`
- calcul des dérivées  $\xi_x, \xi_y, \xi_{xx}, \xi_{yy}$  et  $\eta_x, \eta_y, \eta_{xx}, \eta_{yy}$ .

#### *Sous-programme `second_mixed_deriv`*

```

function uxy = second_mixed_deriv(u,nx,ny,D1X,D1Y)
v = reshape(u,nx,[]);
uxy = reshape((D1X*v)*D1Y',nx*ny,1);

```

Sa structure est identique à celles de `first_deriv` et `second_deriv`. On notera que les schémas de différences finies utilisés sont des schémas de calcul de dérivées premières.

On trouvera à la figure suivante donne les solutions obtenues en  $t_k = k\Delta t$   $k:0,\dots,5$   $\Delta t = 0.02$ .

## XV.4 Problème test de convection – diffusion

Proposé par M. Tabata (« A Theoretical and Computational Study of Upwind-Type Finite Element Methods.” Masahisa Tabata, Pattern and Waves - Qualitative Analysis of Nonlinear Differential Equations - pp319-356 (1986) ), ce problème test va permettre de mettre en évidence les difficultés qui peuvent apparaître lors du choix des schémas de différences finies les plus appropriés.

Le problème à résoudre est le suivant :

$$u_t = -b_1 u_x - b_2 u_y + v(u_{xx} + u_{yy}) + f \quad \text{XV-31}$$

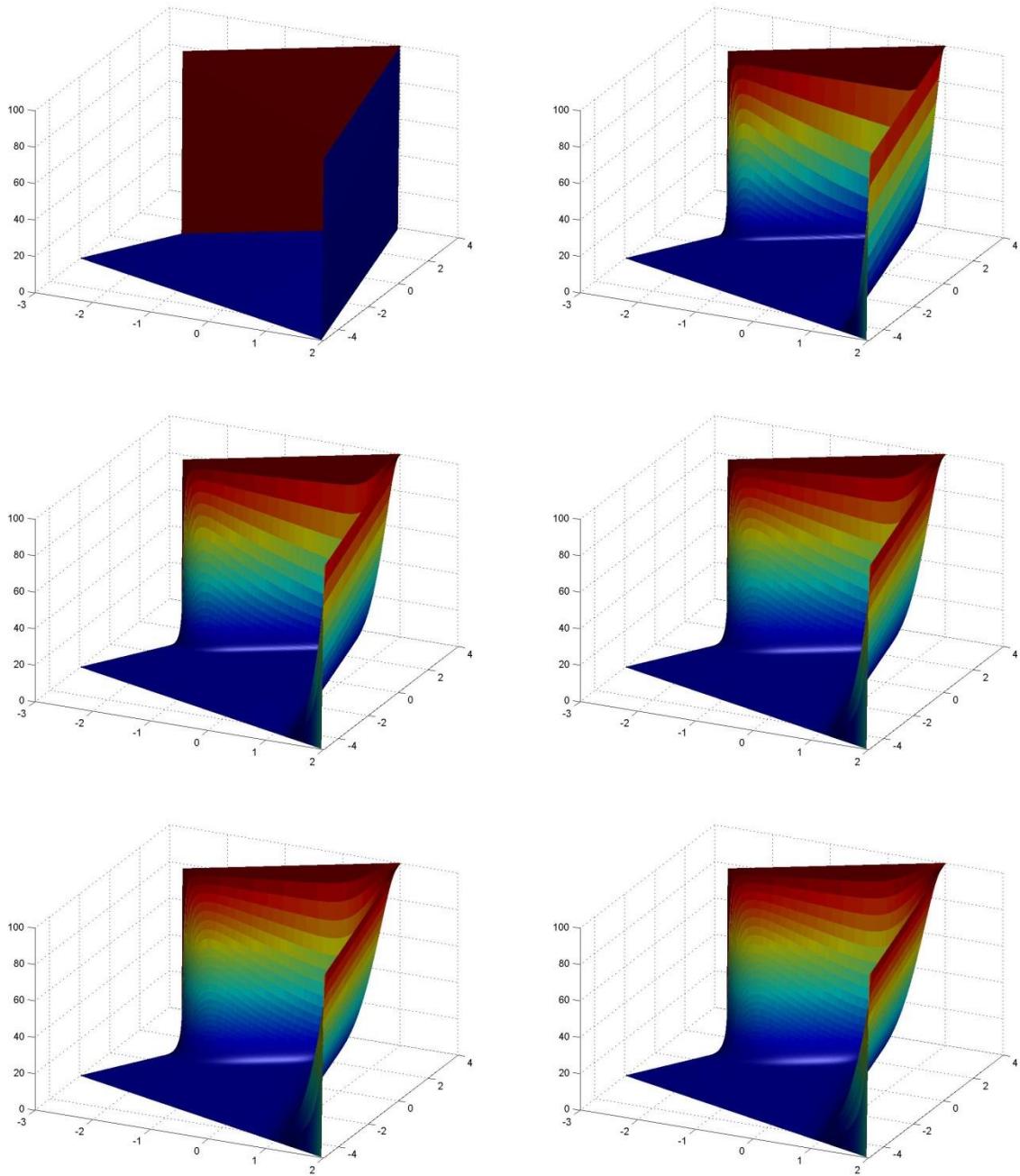
avec  $D = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$   $0 \leq t \leq 1$   
 $v = 0.001$

$$b_1 = 0.5 \cos(\theta) \quad b_2 = 0.5 \sin(\theta) \quad -\frac{\pi}{2} < \theta < \frac{\pi}{2}$$

$$f = 1$$

Condition initiale :  $u(x, y, 0) = 0 \quad \forall (x, y) \in D$

Conditions aux limites :  $u(x, y, t) = 0 \quad \forall (x, y) \in \Gamma$



Le programme repris ci-dessous va permettre de visualiser l'impact du choix de divers schémas de différences finies en liaison avec la valeur de  $\theta$ .

### **Programme principal**

```
close all
clear all
tic
global nu b1 b2 f D1NCM D1NCP D1C3 D1C5 D2C3 D2C5 ncoord nv
%
%   constantes du problème
```

```

%
nu = 1e-3;
theta = 0;
b1 = cos(theta)/2;
b2 = sin(theta)/2;
%
% grille spatiale
%
coordmin = 0;
coordmax = 1;
ncoord = 201;
dcoord = (coordmax-coordmin)/(ncoord-1);
coord = coordmin:dcoord:coordmax;
%
% conditions initiales et terme source
%
x0 = zeros(ncoord*ncoord,1);
nv = length(x0);
f = ones(nv,1);
%
% opérateur de dérivation
%
D1NCM = two_point_upwind_D1(coord,-1);
D1NCP = two_point_upwind_D1(coord,1);
D1C3 = three_point_centered_D1(coord);
D1C5 = five_point_centered_D1(coord);
D2C3 = three_point_centered_D2(coord);
D2C5 = five_point_centered_D2(coord);
%
% intégration temporelle
%
time = 0:.1:1;
[timeout,xout] = ode45(@tabata,time,x0);
%
% C.L.
%
xout(:,1:ncoord) = 0;
xout(:,nv-ncoord+1:nv) = 0;
xout(:,1:ncoord:nv-ncoord+1) = 0;
xout(:,ncoord:ncoord:nv) = 0;
%
% visualisation
%
[X Y] = meshgrid(coord,coord(ncoord:-1:1));
for j = 1:length(timeout)
    figure

        u(ncoord:-1:1,:) = reshape(xout(j,1:nv),ncoord,[]);
        surf(X,Y,u,'FaceColor',[.95 .95 .95],'EdgeColor','none')
        hold
        axis([0 1.1 0 1 -.1 1.3])
        view([32 18]);
        camlight right; % apparition du relief
        lighting gouraud; % lissage)
    end
toc

```

**Sous-programme de calcul des membres de droite des équations différentielles :**

```

function xt = tabata(t,u)
global nu b1 b2 f D1NCM D1NCP D1C3 D1C5 D2C3 D2C5 ncoord nv

```

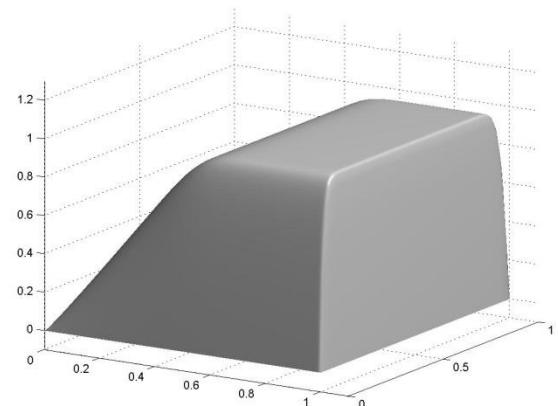
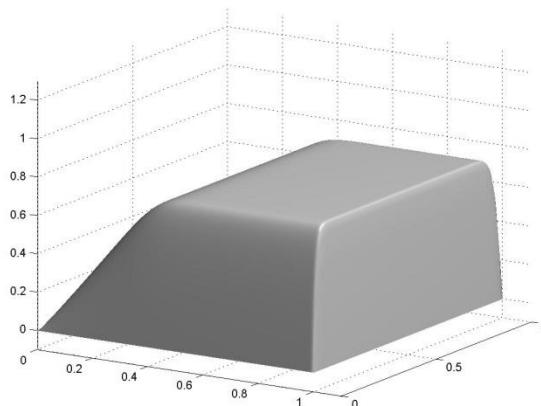
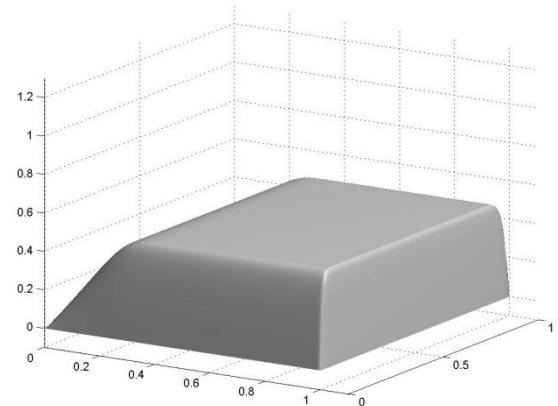
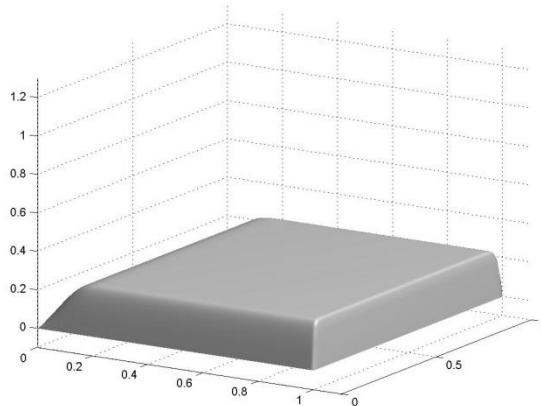
```

t
%
% C.L.
%
u(1:ncoord,1) = 0;
u(nv-ncoord+1: nv,1) = 0;
u(1:ncoord:nv-ncoord+1,1) = 0;
u(ncoord:ncoord:nv,1) = 0;
%
% dérivées spatiales
%
[ux uy] = first_deriv(u,ncoord,ncoord,D1NCP,D1NCP);
[uxx uyy] = second_deriv(u,ncoord,ncoord,D2C5,D2C5);
%
% odes
%
xt = - b1*ux - b2*uy + nu*(uxx + uyy) +f;

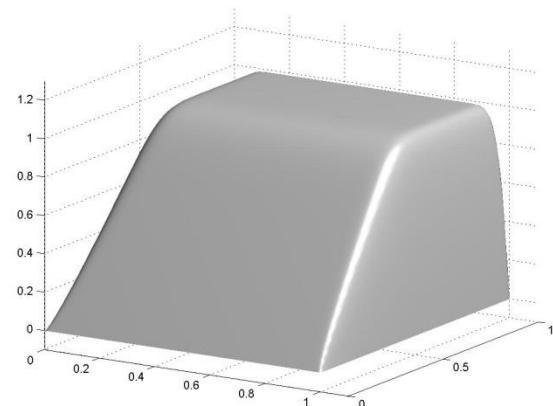
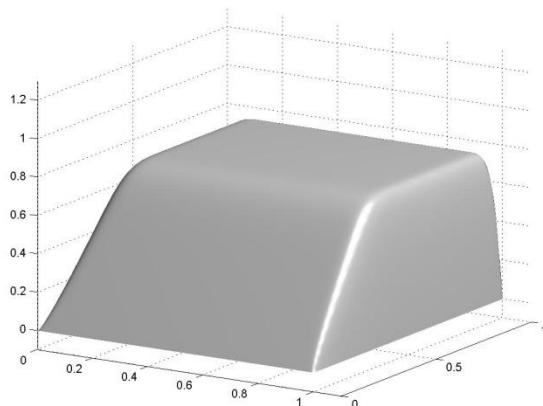
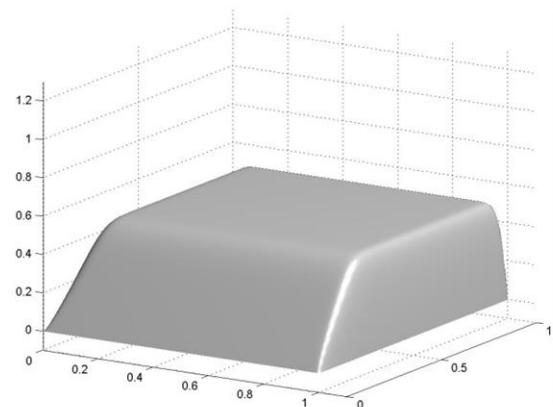
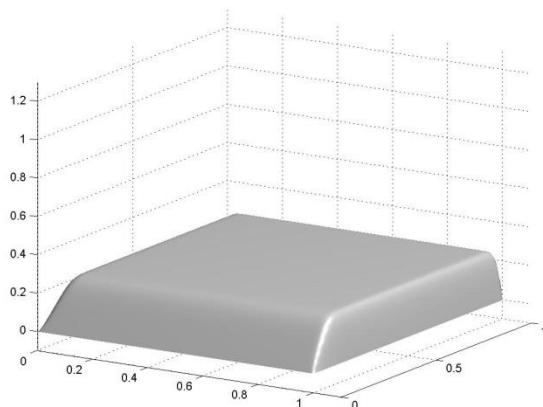
```

Les figures suivantes représentent la solution obtenue en  $t = 0.25, 0.50, 0.75, 1$ . Pour tous les essais,  $u_{xx}$  et  $u_{yy}$  sont calculés par un schéma centré à 5 points.

Essai 1 :  $\theta = 0$  ;  $u_x$  et  $u_y$  calculés par un schéma à 2 points upwind

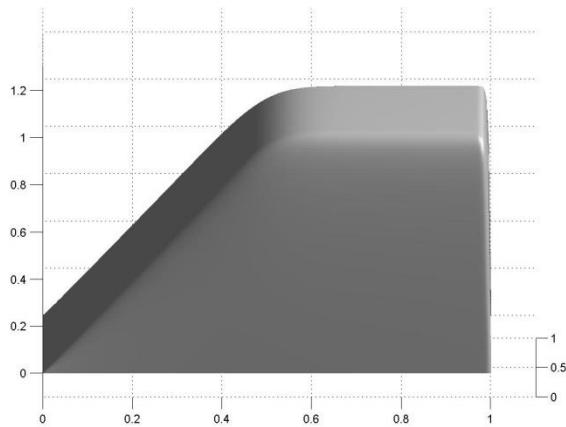


Essai 2 :  $\theta = \frac{\pi}{2}$ ;  $u_x$  et  $u_y$  calculés par un schéma à 2 points upwind

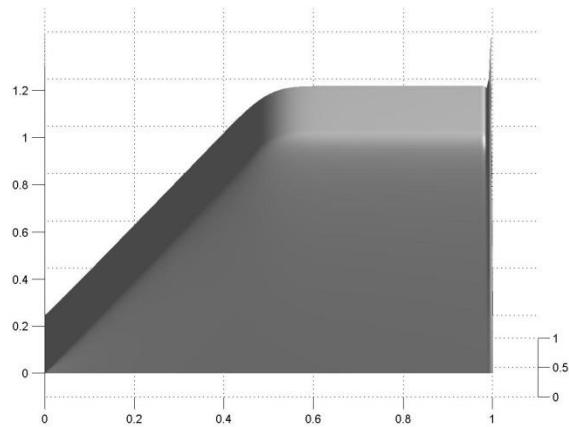


Clairement, on observe l'influence de la valeur de  $\theta$  sur le sens du déplacement de la solution.

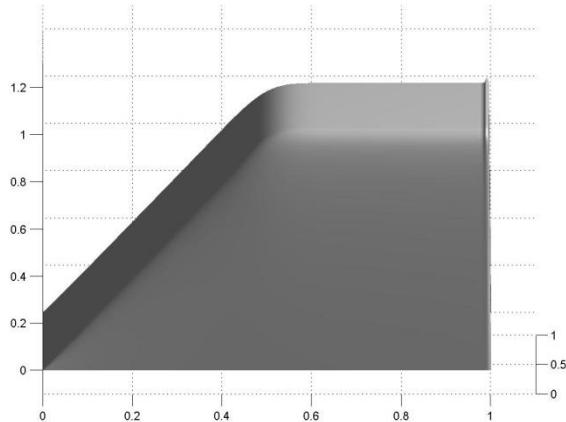
Essai 3 : on peut également apprécier la sensibilité de la solution au choix du type de schéma utilisé pour le calcul des termes de convection : la figure suivante représente, sous un angle de vue différent, les résultats obtenus en  $t = 1$  avec  $\theta = 0$  quand  $u_x$  et  $u_y$  sont calculés avec un schéma à deux points upwind, à trois points centré, à quatre points biased upwind et cinq points biased-upwind :



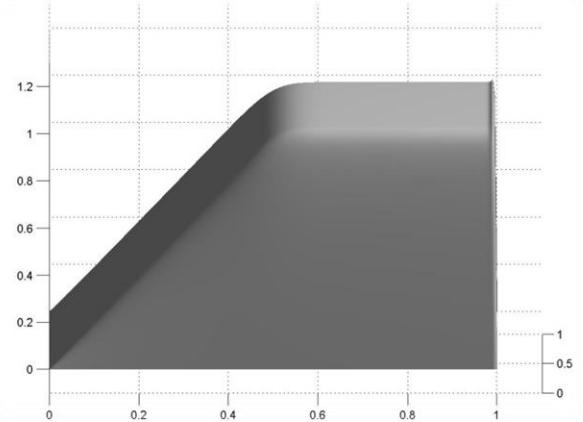
deux points upwind



trois points centré



quatre points biased-upwind



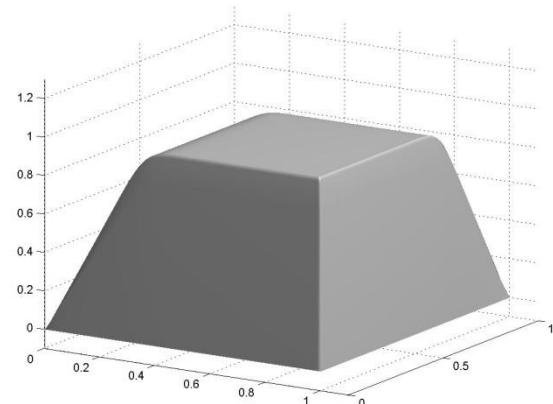
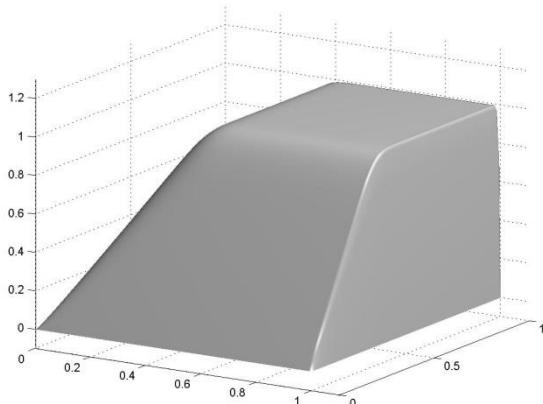
cinq points biased-upwind

les comportements sont identiques à ce qu'on observe dans les problèmes à une dimension spatiale : mauvaise restitution des fronts raides avec des schémas à deux points, oscillation générées par les schémas centrés et meilleur comportement pour les schémas décentrés à plus de deux points.

La seule difficulté afférant au choix des schémas décentrés dans le cas des problèmes à une dimension spatiale était le choix du sens de décentrement : celui-ci devait être upwind par rapport au sens de déplacement de la solution, sous peine d'oscillations importantes ou tout simplement d'échec de la simulation. C'est aussi ce qu'on observe ici, avec des nuances :

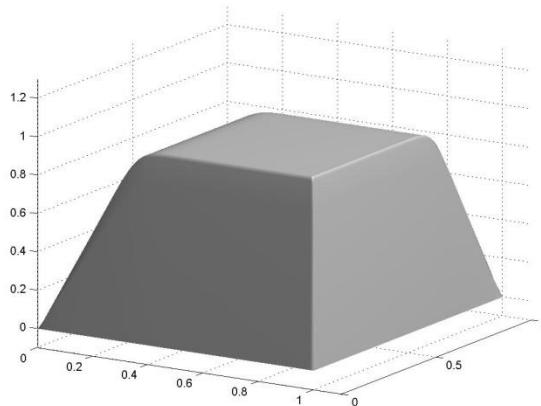
Essai 4 : comparons d'abord les solutions obtenues avec  $\theta = \frac{\pi}{4}$  et  $\theta = -\frac{\pi}{4}$  avec des schémas décentrés dans

le bon sens : upwind en x et en y pour  $\theta = \frac{\pi}{4}$ , upwind en x et downwind en y pour  $\theta = -\frac{\pi}{4}$  (solutions en  $t = 1$ ) :

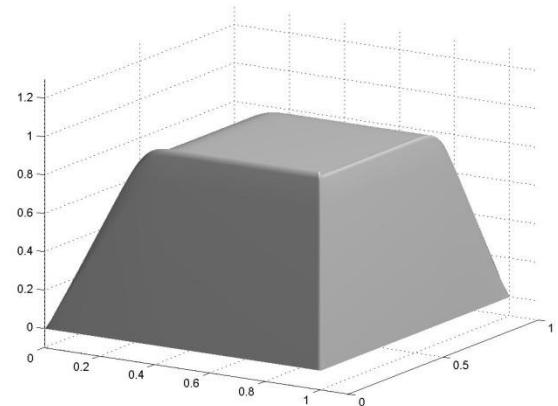


Essai 5 : plus inattendus sont les résultats suivants : on conserve  $\theta = -\frac{\pi}{4}$  et on compare la solution correcte

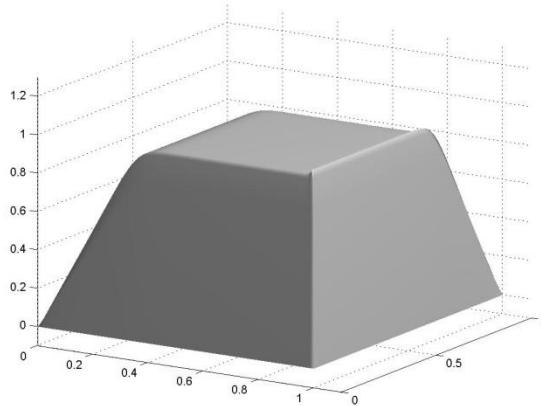
(celle de la figure précédente) aux résultats obtenus quand un des deux schémas de différences finies est mal choisi ( tous les deux upwind ou tous les deux downwind) et quand les deux schémas sont mal choisis (downwind en x et upwind en y) :



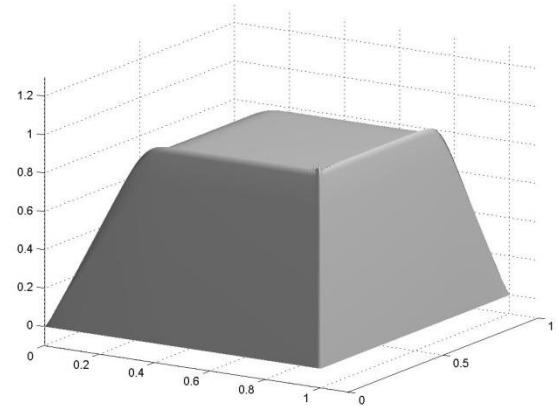
upwind en x et downwind en y



upwind en x et en y



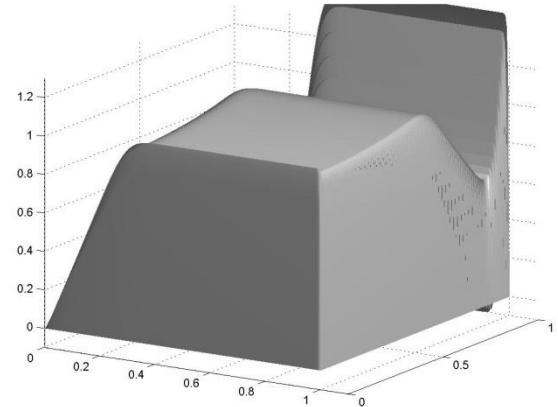
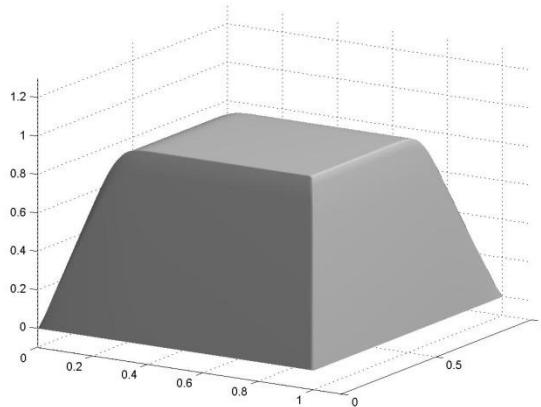
donwwind en x et en y



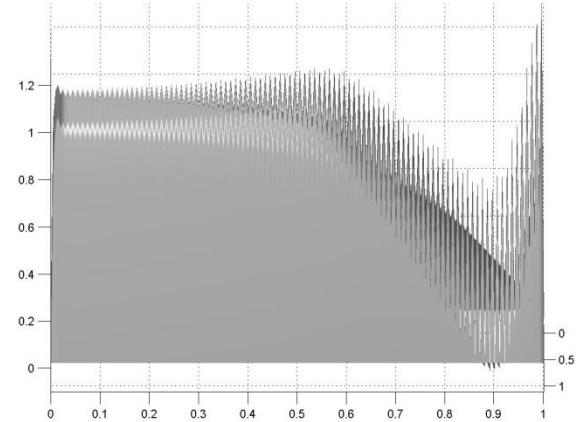
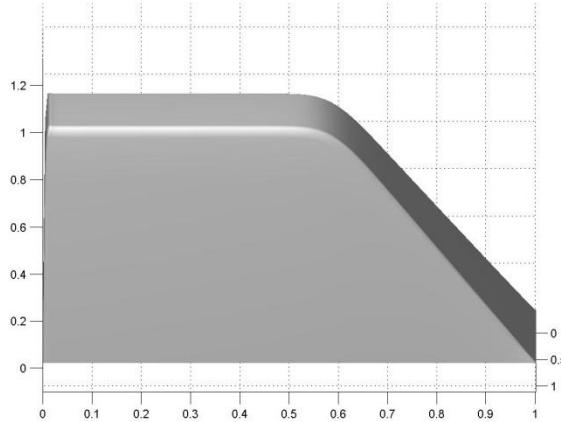
downwind en x et upwind en y

Il semblerait donc qu'un mauvais choix de schémas se solde seulement par de légères oscillations. Cette constatation est trompeuse ainsi que le montre l'essai suivant.

Essai 6 :  $\theta = -\frac{\pi}{3.25} (\approx -55^\circ)$  : la figure suivante compare la solution obtenue avec un choix correct des schémas de différences finies (upwind en x et downwind en y) au résultat obtenu avec un schéma upwind en x et en y :



Le caractère oscillant de la dernière solution apparaît mieux en modifiant l'angle de vue de la dernière figure :



Il semble donc à ce stade malaisé de choisir correctement les schémas de différences finies relatifs aux termes de convection. Cette difficulté est dans la pratique d'autant plus grande si la direction de déplacement de la solution est a priori impossible à prévoir : c'est ce qui se passe dans le cas général des équations de conservation

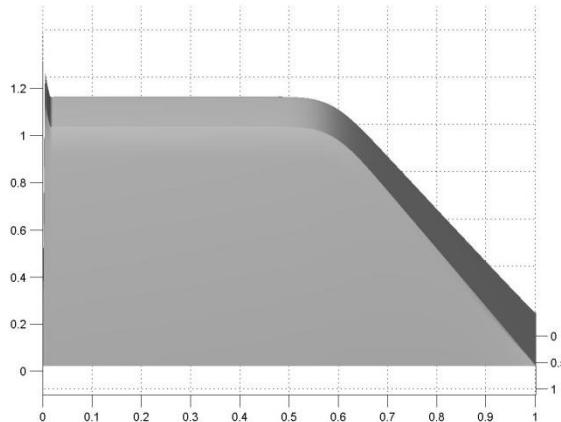
$$\bar{u}_t = \bar{f}(\bar{u})_x$$

X-32

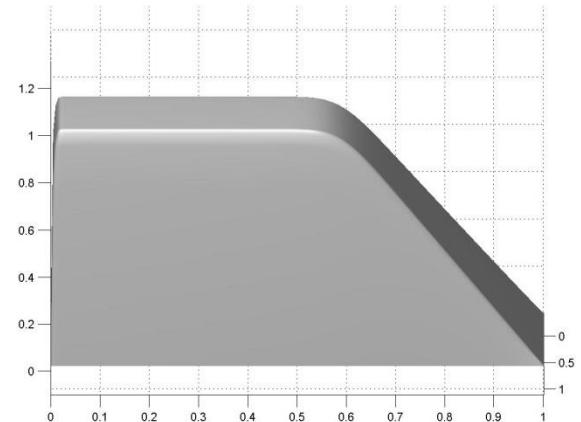
Cette direction est alors liée au spectre des valeurs propres de la matrice jacobienne  $\frac{\partial \bar{f}}{\partial \bar{u}}$  ; en toute généralité, ces valeurs propres dépendent elles-mêmes du temps, rendant ainsi malaisé le choix de schémas de différences finies.

Il serait donc préférable de pouvoir disposer de schémas de calcul des dérivées spatiales (premières) indépendant du sens de déplacement. A priori, les seuls schémas disposant de cette propriété sont les schémas centrés.

Essai 7 :  $\theta = -\frac{\pi}{3.25}$  ;  $u_x$  et  $u_y$  calculés par un schéma à 3points centré et à 5 points centré :



$u_x$  et  $u_y$  par 3points centré



$u_x$  et  $u_y$  par 5points centré

La conclusion de l'étude de cet exemple est donc d'éviter l'emploi des schémas décentrés dans le calcul des termes de convection quand le déplacement de la solution n'est pas connu. Les schémas centrés semblent alors être une piste qui mérite d'être investiguée.

## XV.5 Equation de Burgers à deux dimensions

L'exemple suivant est une généralisation de l'équation de Burgers :

$$u_t = -\left(u^2\right)_x - \left(u^2\right)_y + \epsilon[v(u)u_x]_x + \epsilon[v(u)u_y]_y \quad \text{XV-33}$$

avec  $D = \{(x, y) : -1.5 \leq x \leq 1.5, -1.5 \leq y \leq 1.5\}, 0 \leq t \leq 0.5, \epsilon = 0.5$

et  $v(u) = 0 \text{ si } |u| \leq 0.5$   
 $v(u) = 1 \text{ si } |u| > 0.5 \quad \text{XV-34}$

Condition initiale :

$$\begin{aligned} u(x, y, 0) &= 0 \text{ partout sauf} \\ u(x, y, 0) &= 1 \text{ dans le cercle de centre } (-0.5, -0.5) \text{ et de rayon égal à 0.4} \\ u(x, y, 0) &= -1 \text{ dans le cercle de centre } (+0.5, +0.5) \text{ et de rayon égal à 0.4} \end{aligned}$$

Conditions aux limites :

$$u(x, y, t) = 0 \text{ le long de } \Gamma.$$

On se propose avec cet exemple de poursuivre l'étude de l'impact du mode de calcul des dérivées spatiales sur la qualité de la simulation.

### Programme principal

```
close all
clear all
tic
global eps nu nv ncoord coord D1C3 D1C5 D1NC
%
%   constantes du problème
%
eps = 0.5;
%
%   grille spatiale
%
coordmin = -1.5;
coordmax = 1.5;
ncoord = 101;
dcoord = (coordmax-coordmin)/(ncoord-1);
coord = coordmin:dcoord:coordmax;
%
%   conditions initiales
%
x0 = zeros(ncoord*ncoord,1);
nv = length(x0);
for k = 1:nv
```

```

x = coordmin + (k - ncoord*fix(k/ncoord)-1)*dcoord;
y = coordmin + fix(k/ncoord)*dcoord;
test1 = (x+.5)^2+(y+.5)^2;
if test1 <= .16
    x0(k) = 1;
else
    test2 = (x-.5)^2+(y-.5)^2;
    if test2 <= .16
        x0(k) = -1;
    end
end
nu = zeros(nv,1);
%
%     opérateurs de dérivation
%
D1NC = two_point_upwind_D1(coord,1);%four_point_biased_upwind_D1(coord,1);%
D1C3 = three_point_centered_D1(coord);
D1C5 = five_point_centered_D1(coord);
%
%     intégration temporelle
%
time = 0:.2:0.6;
[timeout,xout] = ode45(@burgers2D,time,x0);
%
%     visualisation
%
[X Y] = meshgrid(coord,coord(ncoord:-1:1));
for j = 1:length(timeout)
    figure
    u(ncoord:-1:1,:) = reshape(xout(j,1:nv),ncoord,[])';
    %
    surf(X,Y,u,'FaceColor',[.95 .95 .95],'EdgeColor','none')
    hold
    axis([-1.5 1.5 -1.5 1.5 -1 1])
    view([32 18]);
    camlight right ; % apparition du relief
    lighting gouraud ; % lissage)
end
toc

```

Les caractéristiques essentielles sont les suivantes : la symétrie du problème traité et celle du domaine spatial invitent à utiliser la même discréétisation du domaine en x et en y :

```

coordmin = -1.5;
coordmax = 1.5;
ncoord = 101;
dcoord = (coordmax-coordmin)/(ncoord-1);
coord = coordmin:dcoord:coordmax;

```

et le calcul de la condition initiale est ensuite abordé ainsi que l'initialisation de  $v(u)$

```

x0 = zeros(ncoord*ncoord,1);
nv = length(x0);
for k = 1:nv
    x = coordmin + (k - ncoord*fix(k/ncoord)-1)*dcoord;
    y = coordmin + fix(k/ncoord)*dcoord;
    test1 = (x+.5)^2+(y+.5)^2;
    if test1 <= .16
        x0(k) = 1;
    end
end

```

```

    else
        test2 = (x-.5)^2+(y-.5)^2;
        if test2 <= .16
            x0(k) = -1;
        end
    end
nu = zeros(nv,1);

```

Plusieurs opérateurs de dérivation, centrés et non centrés, sont envisagés :

```

D1NC = two_point_upwind_D1(coord,1);%four_point_biased_upwind_D1(coord,1);%
D1C3 = three_point_centered_D1(coord);
D1C5 = five_point_centered_D1(coord);

```

Observons d'abord que les opérateurs utilisés sont tous des estimateurs de dérivées premières : en effet, les termes de diffusion interdisent tout calcul direct de dérivée seconde.

Afin de corroborer les résultats de l'exemple précédent, on a décidé de tester encore une fois des schémas décentrés sur les termes de convection.

La suite du programme comprend l'intégration temporelle et la visualisation.

**Sous-programme de calcul des membres de droite des équations différentielles :**

```

function xt = burgers2D(t,u)
global eps nu nv ncoord coord D1C3 D1C5 D1NC
t
%
% C.L.
%
u(1:ncoord,1) = 0;
u(nv-ncoord+1:nv,1) = 0;
u(1:ncoord:nv-ncoord+1,1) = 0;
u(ncoord:ncoord:nv,1) = 0;
%
% paramètre nu
%
for i = 1:nv
    if abs(u(i))> 0.5
        nu(i) = 1;
    else
        nu(i) = 0;
    end
end
%
% dérivées spatiales
%
% 1°calcul de (u^2)x et (u^2)y :
%
[fx fy] = first_deriv(u.*u,ncoord,ncoord,D1NC,D1NC);
%
% calcul de (nu*ux)x et (nu*uy)y :
%
[u1x u1y] = first_deriv(u,ncoord,ncoord,D1C3,D1C3);
nuux = nu.*u1x;
nuuy = nu.*u1y;
[nu_uxx nonutil] = first_deriv(nuux,ncoord,ncoord,D1C3,D1C3);
[nonutil nu_uyy] = first_deriv(nuuy,ncoord,ncoord,D1C3,D1C3);

```

```

%
% odes
%
xt = - fx - fy + eps*(nu_uxx+nu_uyy);

```

Observons que le calcul de  $(u^2)_x$  et de  $(u^2)_y$  est direct, tandis que celui de  $\varepsilon[v(u)u_x]_x + \varepsilon[v(u)u_y]_y$  fait appel à une mise en cascade de `first_deriv` qui génère des termes de type  $u_{xy}$  inutiles mais inévitables.

L'exécution du programme avec les choix indiqués de schémas de différences finies se solde par un échec : comme on s'y attendait, à cause de l'emploi de schémas décentrés pour les termes de convection ; observons d'ailleurs que les termes de convection  $(u^2)_x + (u^2)_y$  s'écrivent aussi  $2uu_x + 2uu_y$  ; la condition initiale comportant des valeurs positives ou négatives selon la région du domaine D où on se trouve, on comprend que la solution se déplace dans des sens opposés selon l'endroit où on se trouve dans D. Il est donc impossible ici de tenir compte de ce déplacement pour choisir l'opérateur de dérivation première. Une solution possible serait de découper le domaine initial en sous-domaines traités séparément et ensuite mis ensemble, mais cela soulève le problème de conditions aux limites supplémentaires aux frontières communes des sous-domaines.

La seule solution semble donc être le recours aux schémas centrés pour toutes les dérivées. La figure suivante présente les résultats de simulation aux instants  $t = 0, 0.2, 0.4, 0.6$  avec l'emploi de schémas centrés à 3 et 5 points pour le calcul des termes de convection :

schéma à 3 points :

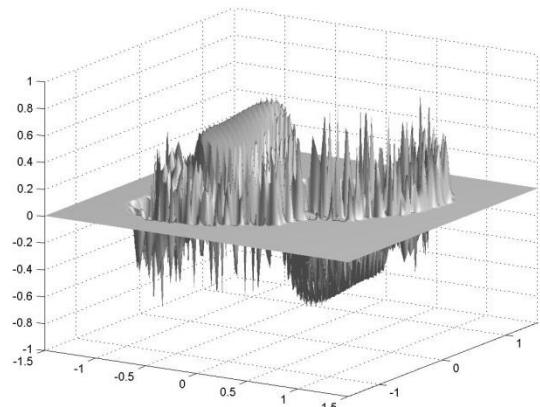
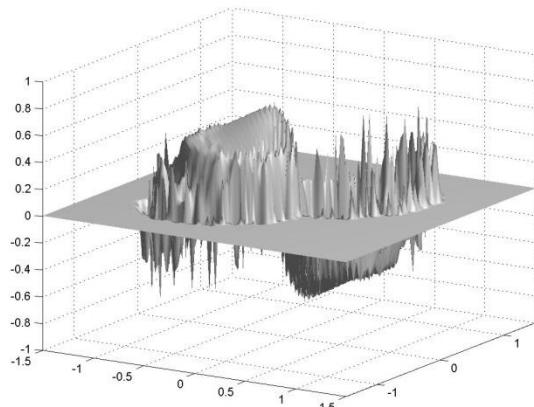
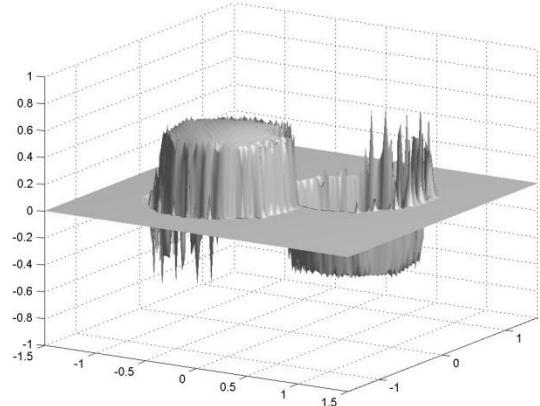
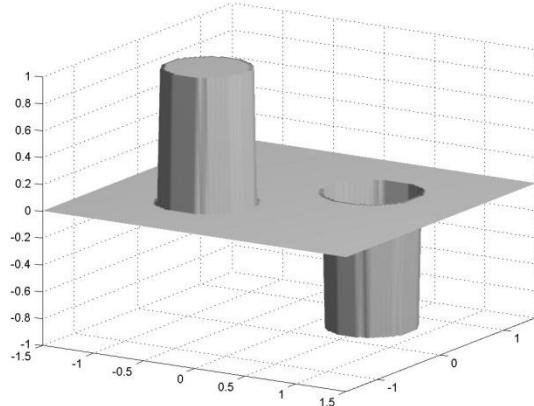
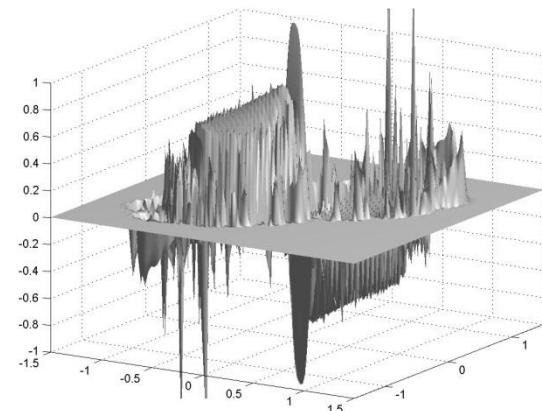
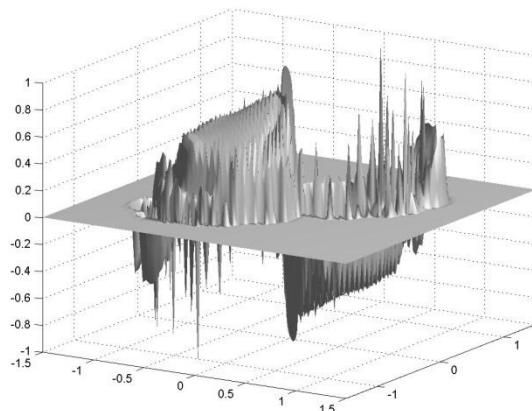
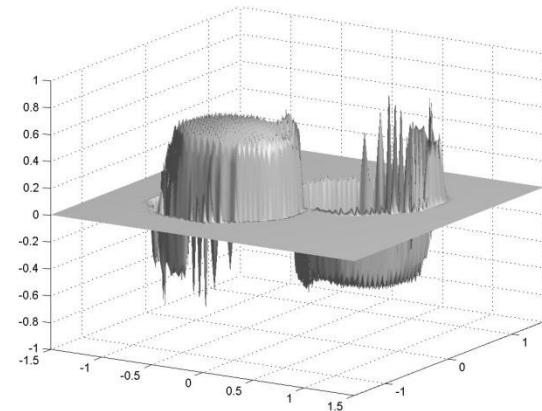
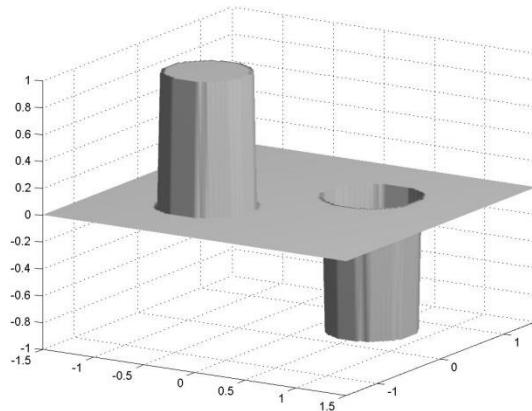


schéma à 5 points :



Ces échecs peuvent être vaincus à l'aide de slope limiters : la programmation 2D du limiteur de Kurganov a été menée sur un rectangle ; le calcul des dérivées spatiales dans le sous-programme précédent devient :

```
% dérivées spatiales
%
% 1° calcul de (u^2)x et (u^2)y :
%
[fx fy] = kurg2D(u,coord,coord,ncoord,ncoord,t);
%
% calcul de (nu*ux)x et (nu*uy)y :
%
[u1x u1y] = first_deriv(u,ncoord,ncoord,D1C5,D1C5);
nuux = nu.*u1x;
nuuy = nu.*u1y;
[nu_uxx nonutil] = first_deriv(nuux,ncoord,ncoord,D1C5,D1C5);
[nonutil nu_uyy] = first_deriv(nuuy,ncoord,ncoord,D1C5,D1C5);
```

#### *Sous-programme de calcul kurg2D :*

```
function [fx fy] = kurg2D(u,absc,ordo,nabsc,nordo,t)
fkurgx = zeros(nabsc,nordo);
fkurgy = zeros(nordo,nabsc);
```

```

usquarex = reshape(u,nabsc,[]);
for i = 1:nordo
    fkurgx(:,i) =
kurg_centred_slope_limiter_fz(1,nabsc,absc,t,usquarex(:,i),@fluabsc,@dfluabsc);
end
fx = reshape(fkurgx,nabsc*nordo,1);
usquarey = usquarex';
for i = 1:nabsc
    fkurgy(:,i) =
kurg_centred_slope_limiter_fz(1,nordo,ordo,t,usquarey(:,i),@fluordo,@dfluordo);
end
fy = reshape(fkurgy',nabsc*nordo,1);

```

Outre la mise sous forme rectangulaire du vecteur  $u$ , le programme utilise les « flux » de convection  $f(u)$  et  $g(u)$  (appelés ici  $\text{fluabsc}$  et  $\text{fluordo}$ ) et leurs dérivées par  $\frac{\partial f}{\partial u}$  et  $\frac{\partial g}{\partial u}$  (appelés ici  $d\text{fluabsc}$  et  $d\text{fluordo}$ ) définis à partir de XV-33 réécrit sous la forme

$$u_t = -(f(u))_x - (g(u))_y + \varepsilon [v(u)u_x]_x + \varepsilon [v(u)u_y]_y,$$

soit ici  $f(u) = g(u) = u^2$ . On a donc ainsi par exemple

*Sous-programme de calcul fluabsc :*

```

function out = fluabsc(n,ne,t,u)
out = u.*u;

```

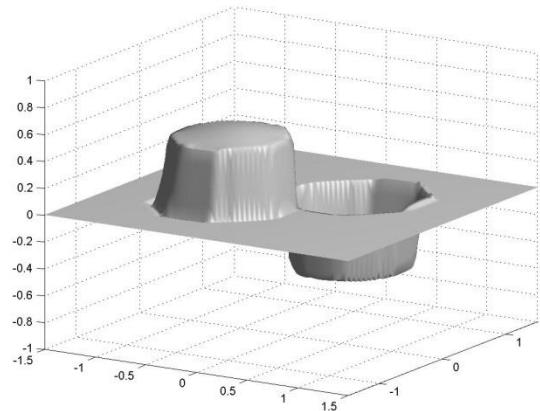
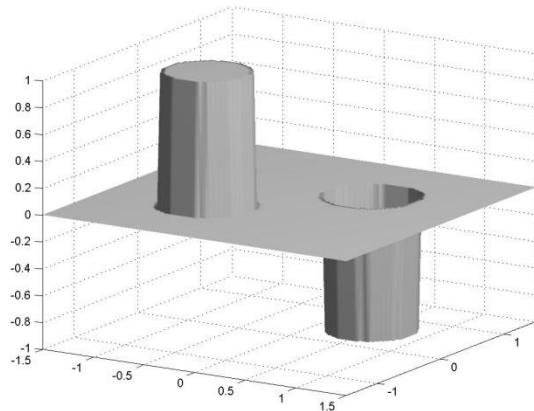
*Sous-programme de calcul dfluabsc :*

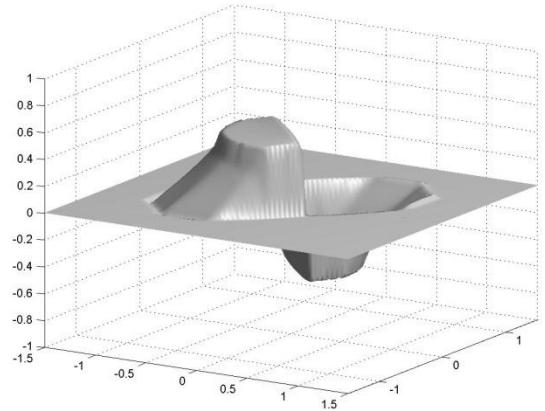
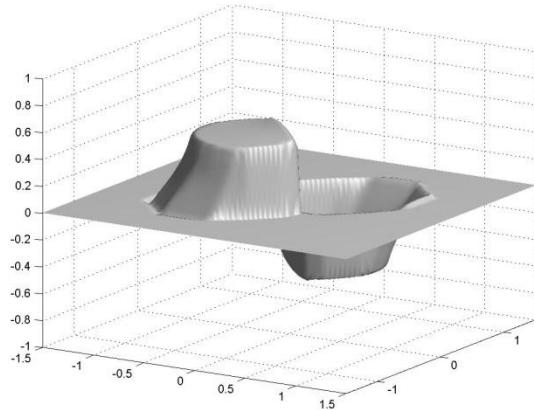
```

function out = dfluabsc(n,ne,u)
out = 2*u;

```

Les résultats de simulation sont cette fois corrects :





Cette très nette amélioration conduit cependant à une augmentation du temps de calcul qui est multiplié par environ 200 !